

DATASCI 306 Week 1

Lecture 1

Jonathan Terhorst

September 1, 2026

Lecture 1

Administrative

- Prof. Jonathan Terhorst (jonth@umich.edu)
- Office hours: W 9-11am, F 9-10am, 252 West Hall
- Bring a laptop: we will write R together in every class.
- Canvas is the official course home.
- Office hours and GSI information will be posted there.
- There is no separate discussion section.

What this course is about

The process of data science

1. **PREDICT** what we expect after framing the question and naming the stakeholder.
2. **INVESTIGATE** by understanding, cleaning, transforming, visualizing, or modeling the data.
3. **INTERPRET** the result in context and state what the evidence does not show.
4. **COMMUNICATE** the conclusion, evidence, limitations, and next step.
5. **ITERATE** by checking the work, revising it, and beginning the next cycle.

Examples we will see in this class

- Use **our class survey** to ask how we should summarize data we collected ourselves.
- Use **Spotify** data to ask what “popular” means.
- Use **career** data to ask what makes an occupation promising.
- Use **youth wellbeing** data to ask what has changed and what might explain it.
- Use **election polls** to ask whether a two-point lead is really a lead.
- Use **college football** data to predict Michigan–Ohio State.
- Use **college outcomes** data to ask what makes a college or major worth it, and for whom.

Class structure

- During **lectures**, we work through Predict, Investigate, Interpret, Communicate, and Iterate together.
- On **problem sets**, you continue the same investigation in a short analysis.
- In **tutorials**, you present solutions, answer questions, and work through difficulties in a small group.
- On **exams**, you read, debug, and interpret code without a computer or AI.

How your grade is determined

- Final exam: **40%**
- Midterm: **25%**
- Tutorials: **25%**
- Homework: **10%**, graded for completion only

Your lowest homework score and lowest tutorial score are dropped automatically.

Exams (65%)

- The **midterm** is worth 25% and takes place in class on Thursday, October 15.
- The comprehensive **final** is worth 40% and takes place Friday, December 18, 4–6 p.m.
- Pen and paper exams. No devices or AI.
- Read and trace code, debug errors, interpret results, and use statistical judgment.

Tutorials (25%)

- Each week you and up to four other students will meet with a GSI for one hour.
- You will take turns presenting your solutions to questions from last week's problem set.
 - Explain what your code does and why it works; use of devices or AI is not allowed.
 - Discuss difficulties and fix them together.
- Your understanding will be assessed by the GSI, and you will be graded based on your participation and performance.
- Your lowest score is dropped.
- Complete the [Tutorial Time Preferences survey](#) to rank your preferred weekly meeting times.

Homework (10%)

- There are **11 problem sets**, normally one per instructional week. Each should take about 60–90 minutes.
- Most run from Wednesday to Wednesday and must be submitted before your tutorial.
- Homework is graded for completion. Make a genuine attempt at every task.
- You may use AI, but you must disclose how you used it and check its work.
- Your lowest homework score is dropped.

Extra credit

- Interactive questions will appear throughout most lectures.
- Each prompt can earn one credit for a **good-faith response**.
- Credit is based on participating, not on getting the answer right.
- Revising an answer does not earn a second credit for the same question.
- Every 10 credited responses add **0.1 percentage points** to your course grade, after any curve is applied.

How to respond

1. Open live.ds306.org on your phone or laptop and sign in with your U-M account.
2. Following along with the slides. Questions will become available to answer when I open them.
3. Discuss with those around you and then submit your own answer.

What happens after you submit?

- Your response is recorded for participation.
- The class display groups the reasons into common themes and shows how common each theme is.
- You may revise your response while the question remains open.
- A question submitted through **Help!** may appear on the projector for us to work through.



AI is good at problem solving

A capable AI can do every take-home assignment in this course.

It is very good at:

- solving a well-posed problem;
- writing and debugging R; and
- producing a polished first explanation.

... but it can't decide what problems are worth solving

You must decide:

- Which questions matter?
- Who is affected?
- What could go wrong?
- What evidence is enough?

A model that predicts the wrong thing

U-M wants to identify students who may need mental health support. It asks an AI to predict who will schedule a counseling appointment. The AI decides to build a predictive model using Canvas activity and previous appointments as predictors. The resulting model predicts appointments accurately, so the university plans to contact only students whom it flags.

Describe one student this system could miss. What does the model predict? What does U-M need to know? Are these the same thing?

Course AI policies

- You may use AI on **problem sets**. Disclose the tool, what it helped with, and how you checked it.
- During **tutorials**, you must present your work and answer questions without assistance from AI.
- **Exams** are closed-book: no AI, devices, or notes. You will read, debug, and interpret code using pen and paper.

Our first investigation

Course staff want to know whether students in data science are taller on average compared to typical U-M students.

How we begin every investigation

Before coding or touching the data, we ask:

- **Stakeholder:** Who needs the answer, and what decision will it inform?
- **Target population:** Whom should the data represent?
- **Observations:** What does one row of data represent?
- **Prediction:** What do we expect to find?

Where the data come from

Class survey

Before class, you (hopefully) completed a short anonymous survey on Canvas.

[View the survey](#)

How should we summarize the survey?

Before seeing the survey responses, predict the median reported height among the respondents. Give a number, its unit, and one reason for your prediction.

Pre-analysis questions

- **Stakeholder:** this class, describing itself.
- **Target population:** the students invited to take the survey.
- **Row unit:** one released survey response.
- **Prediction:** what do we expect the median reported height to be?

What do we need from R?

To evaluate our height prediction, we need to:

- load the survey responses;
- find the columns that record height;
- check whether every value represents what we think it does; and
- summarize the data and report back.

Meet R

- R is a programming language for giving a computer instructions about data.
- **WebR** runs R inside these slides, so nothing needs to be installed today. (Type ``` to open R.)
- Our first task is to load the class survey and find the height columns.

Load our data

```
library(tidyverse)

survey <- read_csv(
  "class_survey.csv",
  show_col_types = FALSE
)
```

What did that code do?

- `read_csv( . . . )` read the file and produced a table called a **data frame**.
- `survey <-` saved that table under the name `survey`.

What is a data frame?

- A **data frame** is a table of data with rows and columns.
- Each row is one observation, and each column is one variable.
- You can access a column by name using the \$ operator, e.g.,
survey\$height_value.

```
survey$height_value
```

Inspect our data frame

```
glimpse(survey)
```

- `glimpse( )` is a **function** that we called on (the data frame named) `survey`.
- It displays the table's dimensions, column names, types, and some values from each column.
- `height_value` contains numbers; `height_unit` tells us what those numbers mean.
- Each column is a **vector**: an ordered collection of values.

R is vectorized

Vector is the basic unit of data in R.

```
height_inches <- c(64, 68, 72) # create a vector  
height_inches * 2.54           # convert every entry to centimeters
```

What does this table represent?

```
glimpse(survey)
```

`glimpse()` reports a number of rows. What does one of those rows represent? Does each row represent a student, or a survey response?

Rows, columns, and values

- One row is one survey response.
- Each column is a recorded variable.
- Each cell is one value.
- `response_id` is the **primary key**: it should be present and unique in every row.

Looking around

```
nrow(survey)  
ncol(survey)  
names(survey)
```

- `glimpse()` gave a quick overview.
- `nrow()` and `ncol()` report the table's dimensions.
- `names()` reports its column names.

Study participation rate

What's the participation rate among the 306 students invited to take the survey? (There are 120 students enrolled.)

Verifying the primary key

How would you check that `response_id` is unique in every row?
(State in words if you do not know the R code.)

```
# verify primary key
```

Different kinds of variables

- **Numerical:** height_value
- **Categorical:** sex, commute_mode, height_unit, favorite_food, birth_continent
- **Numerical storage, categorical meaning:** umid_last_digit
- **Identifier:** response_id

Before calculating a typical height

What is a typical reported height among the survey respondents?

Before calculating a mean or median, what should we do first?

- A. Calculate both and choose the smaller one
- B. Check the recorded units, missing values, and implausible values
- C. Divide the data by sex
- D. Remove the largest and smallest values

What if we calculate first anyway?

```
raw_median <- median(survey$height_value, na.rm = TRUE)
raw_median
```

The code runs, but does it answer our question?

Is the answer meaningful?

```
table(survey$height_unit, useNA = "ifany")
```

Which statement best describes `raw_median`?

- A. It is the median height in the most common unit
- B. It is the median height in inches because most values are inches
- C. It orders numerical entries recorded on different scales and has no coherent interpretation
- D. It is valid because medians are resistant to unusual values

What went wrong?

- The values mix inches, centimeters, and meters.
- The calculation ran, but its answer has no coherent unit.

Plot the recorded values

(Review from STATS 250) What does a histogram tell us?

```
hist(  
  survey$height_value,  
  main = "Reported height values",  
  xlab = "Value as entered"  
)
```

Some reported heights

```
filter(survey, height_unit == 'inches')
```

`filter()` keeps the rows where its condition is TRUE. It does not change the values in those rows.

Our first data pipeline

```
survey_copy <- survey |>  
  mutate(height_copy = height_value)
```

- `survey` is the table we start with.
- `|>` passes that table to the next function. This is called **pipe** and we'll discuss it more soon.
- `mutate()` adds or replaces columns while keeping the same rows.
- `height_copy = height_value` creates a column named `height_copy` by copying the values from `height_value`.

if_else() chooses between two results

The structure is `if_else(condition, result_if_true, result_if_false)`.

```
survey |>
  mutate(
    height_cm = if_else(
      height_unit == "inches", # condition
      height_value * 2.54,    # when TRUE
      height_value            # when FALSE
    )
  )
```

- For an inches row, the condition is TRUE, so R multiplies by 2.54.
- For a centimeters row, the condition is FALSE, so R keeps the value.
- The choice is made separately for each row.
- This two-result rule does not yet handle the rows recorded in meters.

Our first investigation

- We **PREDICTED** a value before seeing the data and named the group of interest.
- We **INVESTIGATED** the rows and discovered that height used several units.
- We **INTERPRETED** the raw median and found that it had no coherent height unit.
- We **COMMUNICATED** why that number should not be reported as a typical height.
- We **ITERATED** by beginning an explicit conversion rule and identifying the remaining cases.

PS1 asks you to transfer the idea

In lecture we investigate **height**. On PS1, you transfer the same process to a new measurement: **sleep**.

- **Predict** by identifying the stakeholder, target population, row unit, and expected result.
- **Investigate** using the key, variable types, recorded units, and NA versus zero.
- **Interpret** the result for the responses supported by the cleaning rule.
- **Communicate** the result, analyzed group, exclusions, and an important limitation.
- **Iterate** by checking the result, auditing plausible AI code, and revising the conclusion.

Apply the five steps to `sleep_value`. Due Wednesday at 8 p.m.

What can we conclude from this survey?

The file contains released survey responses, and the reported heights still mix units.

Which is the strongest conclusion currently supported?

- A. `raw_median` is the median height of all DATASCI 306 students
- B. `raw_median` is the median height of all U-M students
- C. The raw values use different units, so they must be standardized before we estimate a typical height among respondents with usable values
- D. The survey contains no useful information

What do you want to know?

On the problem set, you will be asked to choose two survey columns and compare them.

State a comparison that seems interesting to you. Name the two columns and describe what you would compare.

Week 1, Lecture 1

Closed September 1, 2026 at 4:06 PM

10 class questions · 808 anonymous responses · 0 student questions

The following slides preserve what students submitted during class. No names or login information are included.

Describe one student this system could miss. What does the model predict? What does U-M need to know? Are these the same thing?

99 anonymous responses

Model predicts appointments, not need

23

Responses highlighting that the model mainly identifies who will schedule (or repeat) appointments based on past behavior and may not equal the university's need to know who actually requires support; concerns about representativeness and generalization are included.

No prior appointments / first-time seekers

23

Responses noting the system will miss students who have never scheduled counseling before (including freshmen or new arrivals) because the model is trained on prior appointment history.

Response themes for question 1

Outreach limits: unaware, reluctant, external

19

Responses about students who are unaware of services, afraid or reluctant to seek help, rely on external resources, schedule only in-person, or otherwise fall outside the system's outreach methods.

Low or no Canvas activity

16

Comments pointing out students who rarely use Canvas, use other course platforms, or have sparse Canvas data — causing the model to miss them or producing poor signals.

Strong academics can hide distress

9

Responses that emphasize students who maintain coursework and grades or show normal Canvas activity while nonetheless struggling mentally, so academic indicators fail to signal need.

Response themes for question 1

New events or sudden crises

6

Notes about students who experience recent trauma or sudden life changes not reflected in historical Canvas or appointment data, so the model would not capture emergent needs.

Prediction roughly equals outreach target

3

A small set of responses asserting the model's prediction of who will schedule an appointment aligns with what U-M needs to know (i.e., focusing on who will make appointments is itself the university's target).

Student responses

1. A student that has no previous appointments within the UM counseling department. This system needs to know everyone that is suffering especially those who aren't actively seeking help. These aren't the same thing because some people don't have a good way of coping.
2. A student that they might miss are transfer students or first years who don't have a history in Canvas. The model uses Canvas history to predict their "mental health". It is not the same as who might actually need it.
3. A student who has never scheduled an appointment. The model predicts based on past activity so only current cases. UM needs to know current and future. No they are not the same thing
4. A student who is struggling with mental health outside of school reasons. The model is predicting mental health based on Canvas activity, but this particular case will probably not have activity on Canvas the model will flag. U-M would need to know other things in the students life

Student responses

1. A student who needs help for the first time or has an offline event that isn't documented by UM. The model predicts who the most likely students are to schedule an appointment, not the most likely students who need an appointment
2. I think it could leave out freshmen who have no prior appointment history at the University. This model predicts who will need mental health support at the University based on past history. U-M needs to know appointment history and the canvas information. No these are the not the same thing.
3. If UM want to identify all students who may be in need of help this system would miss the students who are first year students who have not had an appointment before or students who have not used this service. UM needs to know who needs help and that not what that model predicts.
4. it could miss a student who hasn't booked any appointments yet and is experiencing hardship for the first time. The model predicts which student might need additional support by looking at previous behavior. No because a history doesn't mean they currently need help

Student responses

1. It could miss students who are new to using mental health services and U-M needs to know the actual amount of students who need mental health services which would not be the same because it misses any potential new users of the service.
2. It might miss someone who wouldn't schedule an appointment but needs mental health support. U-M needs to know who needs an appointment, which is not the same as who will schedule one.
3. It would miss a student who doesn't show canvas activity indicating mental health needs and who hasn't had any previous appointments, The model predicts students who will likely schedule another counseling appointment, U-M needs to know individual students mental state, no.
4. One student that the system could miss would be someone like a freshman, because they haven't ever had any previous appointments, leading to the AI overlooking them. This model predicts who may need mental health support. U-M needs to know who does and doesn't need mental health support. These are not the same things.

Student responses

1. Students who fit characteristics that don't fit within the characteristics of the students who have previously scheduled appointments since this is what the model uses as predictors.
2. The model could miss a student that has made no prior appointments and has "typical" Canvas activity. The model predicts which students will schedule a counseling appointment. U-M needs to know which students actually need help, not just which ones are willing to ask for help. These are not the same thing.
3. The model will miss students who haven't previously made an appointment but are still in need of help. The model makes predictions from canvas activity and previous appointments, but U-M needs this to also include those who have not previously sought help.
4. The model will predict those that have already had previous appointments will need mental health support. U-m needs to know everyone.
5. The system could miss a student hasn't made an appointment before or is unlikely to speak up or seek help. U-M needs to know who needs mental health support not who is likely to make an appointment. These are not the same.

Student responses

1. The system is looking at previous appointments as predictors, however it does not cover the rest of the students who have mental health issues that have not scheduled previous appointments. U-M needs to know students who are currently struggling that have never sought help. Also canvas activity is not the best proxy.
2. The system will be unable to flag students who have not scheduled appointments before. Additionally, the system may flag students who have sought help previously but may no longer need mental health support. Canvas activity by itself may not be the best measure for mental health.
3. The system will not flag students who have never had an appointment to target mental health. This system just predicts whether people who have made appointments will come back. The system needs to also know what are the signs that led to certain students making appointments, and how can we find those before anything bad may happen.

Student responses

1. There are some people who will not follow the pattern that as the previous ones, so these student will be missed. Also some people who need support, but the haven't make the apppointmen before.
2. This system could potentially miss new students, since they might not have used Canvas that much or they also might not have had an appointment before. This will mainly be used for students that are currently here since they will already have a lot of time in canvas and also will have experience in potentially going to appointments. U-M needs to get new student data in order to help with newer students, and these are not the same since it getsold students and new students
3. This system could potentially miss transfer students who have not attended U-M during prior semesters. U-M needs to know if the students making previous appointmens are representative of the entire student population.

Student responses

1. A student who has never made an appointment before. It predicts return patients or students who already use the CAPS system, not necessarily ones who need help.
2. A student who never have an appointment before
3. It could miss a person that has never had a previous appointment. The model predicts who might schedule another appointment after already having one.
4. It might miss students who haven't scheduled an appointment before and are trying to get help for the first time.
5. New student or Transfer student that are not familiar with the Canvas activity.
6. New students or transfer students who may have not previously booked an appointment
7. New students or transfer students, who have no previous Canvas activity or appointments due to not attending U-M, rather than not needing counseling.
8. One student that this system could miss is a student who has never went to a appointment before or students who don't use canvas to contact professors/others for help.
9. Someone who has not had a prev appointment. Also, you need to know what percentage captured

Student responses

1. Students who have never scheduled an appointment may not be accurately predicted and will not be contacted even if they need it.
2. The model could miss students who need mental health support but have never had an appointment before.
3. the model could potentially miss students that are new patients
4. The model is based on previous appointment so it would most likely miss anyone who has never previously scheduled an appointment.
5. The model might miss a student that needs mental health support but has never scheduled a previous CAPS appointment before.
6. The model predicts that students with previous appointments may need mental health support, thus it will miss students who have never booked appointments.
7. The student that this system could miss would be someone who has never once scheduled a counseling appointment.
8. The students who never did the counseling before. Cannot determine the exact index of the health parameter.

Student responses

1. The system could miss a student that develops a need for mental health support without having made a prior appointment. The model predicts when prior appointment making students will likely make another appointment.
2. The system could miss students that have not had any previous appointments before, or even are just beginning to need mental health support.
3. The system would miss students who have not scheduled an appointment in the past, who may be interested in using the services or are newer to the school/program and have not had the chance to do so yet.
4. This system could miss a student who is new to Michigan because they will have never had a previous appointment wt CAPS.
5. this system could miss someone who needs help but did not previously book an appointment. U-M also needs to know new students that need support and the model would mis this.
6. This system could miss students who have not previously made counseling appointments, but still could use mental health support.

Student responses

1. A student who has never taken initiative to schedule a counseling appointment because they are unaware of their mental health needs. U-M needs to acquire more information on student well-being apart from Canvas data.
2. A student who is unaware that CAPS exists or who doesn't feel safe revealing personal information on Canvas. U-M is targeting students who are already being helped by the system and not those who need it most, those who don't know about the resources.
3. First time student, may only schedule in person appointments and not online, students who may need support but don't recognize it
4. It misses people who haven't made any previous appointments, people afraid to reach out, etc. The model predicts who is more likely to make an appointment for mental health.

Student responses

1. One student the system could miss is a student that might have poor mental health that results in a lack of use of Canvas. Possibly they like laying in bed and sleeping due to poor mental health, which could affect Canvas activity. The model attempts to predict whether or not a student will schedule a counseling appointment. U of M needs to know who needs mental health support. This is not the same thing as there are different levels of how much mental health support a student needs beyond.
2. One student this system could miss is someone who is inactive on Canvas and also has not booked any previous appointments. There are probably many students who have never used the college's mental health resources before, either because they don't know it exists or they don't think they need help. U-M needs to know about the students who have never been to counseling before.

Student responses

1. One student this system could miss is students who are not proactively seeking counseling appointments, and who seem fine externally. While Canvas activity and previous appointments are predictive variables, it is impossible to capture the status of a person's mental health with them alone.
2. Students who do not want to go to counseling. The model predicts students who already had appointments and not students who might need it. UM needs to know who needs mental health support not who is getting help.
3. Students who have received mental health support from another service/institution
4. The AI may miss students who have not yet reached out for help but are struggling because the previous appointments encompass only students who are receiving help. Canvas is also not a place where people talk about their mental health struggles.

Student responses

1. The model is likely to miss the student that is afraid to seek help. Since the entire population of students in its model comes from those who have made appointments or had visual struggles on canvas, it is not trained to understand individuals who have never sought counselling. U-M needs to have a larger population to better account for this group.
2. The system can miss students who have never heard of the appointments page and/or students that use outside UofM support systems. The model predicts which students will need mental health support/an appointment. U-M needs to know students background information and their past history.
3. The system could miss a student who needs help but didn't know how to make counseling appointment. The model predict students who had mental issues and had tried to solve it. U-M need to know the students who need mental support. It's not the same thing.

Student responses

1. The system could miss students that may have previously scheduled appointments, but not through the university and through other resources. It may also miss students who use another site to complete coursework, as some departments do not heavily rely on Canvas.
2. The system may miss students who make appointments outside of campus. The model predicts students who've been to a counseling appointment before.
3. There may be students who haven't had an appointment yet and want to start, but they won't be contacted. Students who accessed other resources also won't be contacted.
4. This model misses students who have never reached out for help before, who may be the students that could use the support the most. By going based on canvas activity and previous appointments, students who have not yet made an appointment will not be flagged.

Student responses

1. This system could miss a student without the knowledge of the available U-M resources. This could mean the student still may need mental health support despite not having previous appointments. U-M needs to know students seeking help that have never had an appointment.
2. This system could miss students who will not seek to schedule a counseling appointment. Moreover, since the system is using previous appointments as predictors, the students who didn't schedule appointments previously could be missed

Student responses

1. a student that is depressed and seldom use Canvas / appointment. The model predict students who use Canvas heavily will depressed, which is not accurate
2. A student who doesn't use Canvas often and may haven't used appointments system before. The model predicts the type of students who have large chance to schedule a counseling appointment. U-M needs to know all the previous data and Canvas data.
3. a student who hasent been logging onto canvas or hasent scheduled any previous appointments.
4. EECS students that aren't on canvas much, students that are taking a gap semester, students not part of these systems or a student that has never taking a counseling appointment before
5. if the AI is sing Canvas activity and previous appointments as predictors, it may miss some student that never doing the counseling before, and also, Canvas activity is not actually a good predictor
6. If there is no strong correlation between the canvas activity and the students who need mental health support

Student responses

1. might miss people who don't participate in class/are on the verge of withdrawing, or having at home issues. the model predicts the likelihood of needing help based on canvas activity and how likely you were to schedule for help in the past
2. People who don't use canvas for some of their classes. FOr example, a lot of EECS classes aren't on canvas, but have their course material on another site.
3. Student who does not use Canvas but are suffering from depression. The model only predicts students who are using canvas in making appointments. U-M also needs to know students who need mental support not only in Canvas but also from different platforms.
4. students that dont check canvas because they dont care about school
5. the model predicts the likelyhood of needing a cousling apointment and then reaches out to the student. The university wants data about squedule future counseling apointments. I dont think having canvas is a good mental health indicator

Student responses

1. The system could miss students that have never scheduled an appointment before, or don't use canvas the way that it is intended in order to flag the AI.
2. There might not be any data on canvas activity that show a student needs mental health support. The system may flag students and contact those that need help, but it can miss a lot of students with limited data.
3. This doesn't generalize to all students, like ones who don't use Canvas. For example, some med students use other platforms.
4. This model is solely based on how Canvas activity and previous appointments is intertwined with how likely a student is going to seek mental health support. This system could miss a student who is not very active on Canvas.
5. This system could miss some students who don't use canvas frequently

Student responses

1. A student that needs mental health support but is still maintaining good activity in their classes. Not everyone who is struggling will necessarily be struggling academically.
2. One student this system could miss is someone in need of counseling that doesn't show extreme Canvas activity or has had previous appoints. There are other reasons, other than Canvas activity and appointment scheduling, that influence if a student needs a counseling appointment for mental health reasons (such as problems at home, problems with extracurriculars, etc.)
3. The AI could miss a student who still attends class regularly and is completing assignments, but is still struggling with mental health. The AI only knows Canvas activity and previous appointment data, so it has very little insight into a student's personal life.
4. The model predicts students who are struggling with classes and previously had An academically perfect student who is struggling with mental health issues and has no previous appointment history

Student responses

1. The model predicts that a student who doesn't perform well, maybe missing a lot of assignments might have a mental problem and needs support. It will miss the students who still well in classes who are suffering from mental problems. UM need to know if a student experiences anything unfail systematically
2. The system could miss out on student who are transfer students and students who performs well at school but are still suffering mentally. The model predicts students mental healthiness base on schoolwork performance. U-M need to know students' mood.
3. The system could miss students who are suffering from mental health issues that don't reflect what they do on canvas. E.g. A student with an eating disorder might not have a change in canvas or school activity.
4. The system uses only Canvas activity and someone who still might be dealing with things mentally could still be active and missing work or deadlines is not the only sign of mental health issues.

Student responses

1. This system could miss a student who has never before booked an appointment and is still maintaining normal canvas activity despite the fact that they are incredibly stressed but don't know how to stop continuing and how to seek out help.

Student responses

1. A one-off instance where a student has something going on in their life and they need mental health support. U-M would need to know more about the specific situation so they can best support the student. The model predicts student's only require support if they have a lot of work to do or have been to an appointment before.
2. A student that had recently had a mental health issue where they haven't been to support in the past. The student may also not use Canvas as frequently as other students. The model predicts frequent counseling sessions from the past, but U-M needs to know the holistic total of all of the students.
3. It could miss a student who has not previously visited counseling but recently had a major life event that negatively affected their mental health. Also, if they are relatively introverted/do not share their struggles, using only these factors to predict who needs help will likely lead to this student not getting the help they need. The model essentially just reaches students who have gone before. U-M needs to know who needs help, even if they haven't gotten it before.

Student responses

1. It wont flag someone who hasnt previously had a problem, but may develop one and wont know how to get help
2. It would miss students whose mental health needs are not just based on past data but new situations they encounter that are not in the data.
3. students who just had a sudden traumatic event but hadn't needed mental health support prior

Student responses

1. A freshman (no Canvas activity or previous appointments). The model predicts which students may want to schedule an appointment for mental health services. U-M needs to know which students would be interested in scheduling an appointment. These are the same thing.
2. The model could miss incoming students who don't have much canvas activity etc. The model predicts which students are most likely to need mental health support. UM needs to know what students are going to make mental health appointments
3. This student could miss a freshman who doesn't have much Canvas activity or any previous appointments. The model predicts the who is most likely to want to schedule an appointment. UM needs to know who is going to schedule appointments.

Before seeing the survey responses, predict the median reported height among the respondents. Give a number, its unit, and one reason for your prediction.

106 anonymous responses

Average of male and female heights

33

Predictions based on averaging known male and female averages or general U.S./population averages (mid-60s to high-60s inches), sometimes doing a weighted average between genders.

Male-dominated class skew

27

Predictions that explicitly cite a higher proportion of men in the class — students expect the median to be pulled up toward typical male heights.

Metric-unit predictions

17

Responses given in centimeters or meters (e.g., 160–176 cm or 1.7 m), sometimes with the same averaging or gender assumptions but reported in metric units.

Response themes for question 2

Unspecified or arbitrary guesses**11**

Short answers with little or no reasoning, unclear numeric entries, odd or unsupported justifications, or idiosyncratic stereotypes without clear calculation.

Visual or personal observation**10**

Estimates based on looking around the classroom or campus, comparing to the respondent's own height, or using perceived bounds of heights seen in the room.

Population or regional references**7**

Predictions that invoke specific populations or regions (U.S. adult averages, Michigan/University of Michigan population, or northern-region genetics) to justify the median.

Other responses**1**

Student responses

1. 5 foot 6 inches, or 66 inches, I predict this because its the average of the average height of a male and the average height of a female
2. 5 ft 8 in, I would make this prediciton because it is a number around the average height men as well as women.
3. 65 inches because I think 66 or 67 is the median male height and womens med height is lower so it brings it down a little bit.
4. 65 inches, the average height of an adult male in the United States is 69 inches, and I believe 64 for an adult female.
5. 65 inches. I think it will be between the average height for men and avrage height for women in the US.
6. 66 inches average height
7. 66 Inches, I am predicting this because I believe the average height worldwide for both males and females is somewhere around there and there is probably more males in this course than females so leaning closer the male average
8. 66 inches, it seems like a good median between the median height for women and men.
9. 66, inches, 68 is average for men, 64 is average for women, so I think it will be in the middle

Student responses

1. 66, inches, factoring in both men and women...tried to estimate the in between number between avg. height of man and woman
2. 66.5 We took a weighted average of our best estimate of the proportion of males/females and their average heights
3. 67 inches because it is representative of the height population of the United States.
4. 67 inches is my guess, as that would average out the average height of most college guys and girls.
5. 67 Inches, based on average heights and estimated gender distribution
6. 67 inches, somewhere in between average heights for men and women
7. 67 inches, taking into account that girls tend to be shorter ~64in and guys average height is around 70in, this is in between the two.
8. 67 inches; considering that both men and women are taking this class, and this is not a class where we would expect a disproportionate amount of athletes who might be taller than other students.
9. 67 inches. That number is around the average height of people in the U.S., and there is most likely not a correlation between if you are taking this class and your height.

Student responses

1. 67, inches, i averaged the average male height of 70 and the average female height of 64, but i am not totally sure about these numbers
2. 67in, because 69in is the average height of a male in the US, but it would be slightly lower than that because you need to account for both genders.
3. 68 inches in my guess because i feel like it is a happy medium between the average hieght of women and men
4. 68 inches. I don't think that it would be that different from the actual population height, to be honest.
5. 68 inches. I took the average of girls and guys and put it into inches.
6. 69, inches, median from males and females in emprical system
7. I predict that the median heigh reported among respondents is probably 65 inches, as that may be the upper value of the lower half of the data. On average, women residing in the US are about 5'4 to 5'5 and men are 5'7-5'9.

Student responses

1. I predict that the median reported height is 67. inches. This is based off the average male height of 69 inches, and average female height of 64 inches. There are likely more males in the class, so it should be skewed to their avg height. final answer - 67.5 inches
2. I predicted that the median reported height will be around 65 inches. This is because the average height of a 20 year old male in the U.S is around 69 inches, and I suspect the median height will be a bit lower.
3. I think that most people answered in inches, and I believe the median is 66 inches. The reason for the unit is that it is the most precise, and the reason for the number is that it is a mix of average male and female heights
4. I think the median height among the respondents is 5 feet 6 inches. I predict this because I think the median height for men is 5 feet 8 inches and women is 5 feet 2 inches.
5. I think the median height will be about 67 inches because the average height for girls is 64 inches and the average height for guys is 69 inches, and 67 is in the middle of that.

Student responses

1. I would estimate the average height to be 5'7 feet because i think that is the midpoint between the average male and female height.
2. median likely mirrors U.S. population average around 66 inches. This prediction comes from averaging the U.S. population averages for men and women which are around 69 inches for men and 64 inches for women.
3. Probably 66. I believe the average height for US males is 69, and for women its 63 or 64 inches. 66 is kind of between those averages.

Student responses

1. 5 feet 8 inches because of the average height in the US and since there is more men than women in the class.
2. 5 foot 7 inches. The avergae height of males and females within DS shouldn't change, but there are more men than women, so i would assume its a little higher
3. 5 foot 8, average male height is 5 foot 9 and since there are more males than females in data science I believe it'll be lower than five foot 9 but not by that much
4. 5'6 because theres going to be more males than females in data science courses I think and thats going to draw up the average.
5. 65 inches because mostly male dominated field so above normal median.
6. 66 inches, in my experience there has been more males in data science classes than females, and the avg male height is probably around 69 inches
7. 66 inches, this is close to the average height for the area, leaning toward the median male height as the class seems predominately male.
8. 67 inches because I feel that there are a lot more guys in this class than girls, so I guessed the average to be a little bit taller than the average person.
9. 67 inches because the class has

Student responses

1. 67 inches, because it is around the average height of both men and women, but slightly higher because I see somewhat more men than women in class.
2. 67 inches, this class has more males than females and will be a higher number because that affects the median
3. 68 in, it seems like students in the data science field are majority male, pushing the median height higher.
4. 68 inches because I think data science is a male dominated field so the average height will be around the average male height.
5. 68 inches because I think it'll be closer to the average height of males since the class seems to be predominantly male.
6. 68 inches because this major is male dominated and the average height of men in the US is 5'8
7. 69 inches because I think a lot of people take this class after they are a declared datasci or stats major, and generally this is a more male dominant field, so I think the height average is a little above the average male height
8. 69 inches. Because there are more boys than girls and the median should be the height of a male
9. 69 inches. this is because looking around most of the people are male.

Student responses

1. i beleieve teh mean will be around 5 foot or 66 inches i think the men average for men is 66 inches and the mean for woman is 62 inches however men are the majority of the room and they look a little taller 6
2. I expect 5'8" to be the median height because the avg height for guys is 5'9" and for girls is 5'6" and data science tends to have more men than women.
3. I predict a median height of 67 inches because of the consideration of the average height of men versus women, and the class having more men than women.
4. I predict roughly 67 inches because there might be more men in this class and the average male height out of the general population is 69 inches, so accounting for the demographic 67 inches seems about the median
5. i predict that the median height of this class would be 5'6. this is because there are more men in the field so the median height would be closer to the
6. I think that the median reported height will be 67 inches. I predict this will be the result because it is around the average height for all people, skewed slightly higher given that there appear to be more men in the class.

Student responses

1. I think the median the median height among the respondents would be around 70 inches because there are a lot more guys in the class than girls
2. My prediction is 5ft 7inches because data science is definately a more male dominated field so I would tilt the scale a bit more in the male favor
3. The median height is likely 5'7". That is because there are more men than women in the classroom, roughly a 2-1 ratio, so the median would skew closer to the average male height of 5'9".

Student responses

1. 1.7m, its arguably the average height when men and women are averaged.
2. 160 cm. Just speculation.
3. 168cm, average from US adult height from both men and women
4. 170 cm
5. 170cm Based on my observation
6. 170cm I guss
7. 170cm, because i think a college student will have a height between 160 to 180
8. 173 cm, because women are shorter, and men are taller so when mixing up, my prediction of height is173cm
9. 174 cm. In general, the male students should be relatively taller, and female students should be within the range of 160-175cm. Assuming there's 50% of each gender among the respondents

Student responses

1. 174cm I would think we are slightly below average as data science students for average height , so I believe 174 is a suitable estimation
2. 175 c
3. 175 cm. I assume that the average height of female students are 165 and for males 185. There might be more males.
4. 176cm; more male than female for ds course, people tends to be taller who are living in northern
5. Around 170cm, most people fall between 160cm-180cm in height.
6. I will say the reported median height will be 168 cm. This is because we have boys and girls, and boys are usually taller than girls.
7. The median height of the population 18-24 is roughly 170 cm. I'm going to guess that the median reported height among the survey responses is a bit taller than that. Looking around the classroom, there seems to be more men than women, who are taller on average. My prediction is 172 cm.
8. the median reported height will be 175 cm. This is because I assume that girls and boys are evenly distributed and thus the middle number will be around the max of girl height and min of boy height.

Student responses

1. 5 ft 6 to ac
2. 65 in
3. 65 in compsci students sit for many hours at their comuter
4. 66 Inches
5. 66 inches, i think the median is shorter than the median of the entire school
6. 66inches
7. 67 inches
8. 67 inches
9. 68 inches

Student responses

1. 70 inches
2. I think 67 inches

Student responses

1. 165cm. cuz i'm 165cm and i look like a medium height in the room
2. 5.5 feet as my perceived median of students our age
3. 5'7 (feet + inches), most common "middle" height that I've seen in most people.
4. 65 in inches. I think most people did inches, and that seems like a fair guss based on the people in this class
5. 66 in. This Is similar to what I feel like I see
6. 66 inches and I think It would be this because there is a good diversity of tall and short people in the class so I think this would make sense as the best answer. Also its close to my height which I think is quite average.
7. 66 inches. I believe this is a reasonable prediction for the median height amount the respondents, because it takes into account taller than average and shorter than average students. Then, I used my own judgement to estimate 66 inches.
8. 67 inches. I feel like it's a good middle ground in the heights of people I have noticed around campus.
9. I think the median is around 67 inches. I think this is because usually the shortest is around 5'4" an the tallest is not more usually more than 6'0"

Student responses

1. Predicting the median reported height among respondents, I think it may be around 68 inches. One reason for this is that there are a decent amount of taller people in the class, but it also seems there are people of various heights.

Student responses

1. 5 ft 5 in. Googled the average height of Americans and assumed major is independent to height
2. 5'7". Although data science is a subpopulation, I have no data to support whether or not it diverges from the larger UM population. Going off of that, the approximate UM population average is 5 feet 7 inches tall.
3. 5'9, inches. I believe this beacuse I think its the U.S. median height.
4. 68 inches because of assumed backgrounds of students that might genetically effect height.
5. around 65 inches because thats around the average height for american adults
6. I think the median height is 72 inches because that is the median height of people who live in michigan and UofM is mainly people from michigan.
7. Median height that I would guess is 69 inches because that is the median height of adults in the US. I would not guess that data science students are taller on average than other UM students.

`glimpse()` reports a number of rows. What does one of those rows represent? Does each row represent a student, or a survey response?

105 anonymous responses

Row = survey response / observation

49

Student says each row is a single survey response or observational unit; often notes that this corresponds to one student's submitted answers (may not be unique if duplicates exist).

Row = variable / question (column-wise confusion)

23

Student interprets rows as variables/questions or says rows show all values for a particular variable, reflecting confusion between rows and columns.

Row = student (equating row to student)

15

Student explicitly states each row represents a student (sometimes implying one row per student).

Other responses

10

Responses the model could not place reliably.

Response themes for question 3

Unclear or fragmented responses

8

Very short, unclear, or incomplete answers that don't clearly state whether a row is a response, student, or variable.

Student responses

1. A student response
2. a survey response
3. a survey response
4. a survey response
5. A survey response
6. each represents one survey response
7. Each row corresponds to a student response.
8. Each row in this case represents a survey response, or the answer to one type of question
9. each row is a survey response

Student responses

1. Each row is a survey response
2. Each row represents a different category of data that we are collecting, so a survey response type.
3. Each row represents a response
4. Each row represents a student response, it contains all the data values submitted
5. each row represents a student's answers
6. each row represents a students responses to the survey questions
7. Each row represents a submitted survey response
8. each row represents a survey response
9. each row represents a survey response

Student responses

1. each row represents a survey response
2. Each row represents a survey response
3. Each row represents a survey response
4. Each row represents a survey response, because each row is one student's collected responses
5. Each row represents a survey response.
6. Each row represents a survey response.
7. Each row represents an observation/response of a student
8. Each row represents one observational unit
9. Each row represents one response to all the survey questions.

Student responses

1. Each row represents one survey response, which should have only been completed one time by each student
2. Each row response
3. Each survey response submitted
4. Each row represent a survey response of a student. The number of the rows represent the number of students who take the survey.
5. I think each row represent a survey response
6. It represent a data response.
7. It represents a survey response
8. It represents a survey response
9. It represents a survey response

Student responses

1. One of the rows represent a student's survey responses
2. One of the rows represents a survey response
3. one of those rows is a survey response
4. One row represent a response that was submitted by a student. It is structured in such a way that each row showcases the responses that the students filling it out gave.
5. row - observation, survey response
6. rows represent the number of students who answered the survey, and each row represents one student's response to each of the questions
7. student's survey response
8. survey response
9. survey response

Student responses

1. survey response
2. Survey response
3. The name of the category, the data type, and then the data. each row is a single released survey response (not necessarily a unique student).
4. We are analyzing the students of the class, but each row technically represents a single response. A student might accidentally respond multiple times

Student responses

1. A row represents to all the responses of the students from one of the questins
2. A survey question's responses
3. each row corrispodes with a varaiable or asurvey question. w can use rows to help group data and look for corrleation e.
4. each row represent a question asked to the student.
5. Each row represents a characteristic or question that someone answered like how they commute to school and then the rest of the data in the row are their other responses
6. Each row represents a different unit of measure. For example one of the rows represents sex while the row below represents communitime. Each row represents a surovery response
7. each row represents a individual variable
8. Each row represents a survey response for a specific category
9. Each row represents a variable in the survey together with all the responses of this variable.

Student responses

1. each row represents all entries of one response. In this survey, this means the response of one student.
2. Each row represents all the responses to a survey question.
3. each row represents all the values from a given question
4. Each row represents all the values of a particular column of the data frame
5. Each row represents the survey response of the corresponding column
6. Each row that glimpse() reports gives a certain variable such as height, sex, etc. and then it gives some of the values of the table for each variable
7. In glimpse, each row represents a collection of answers to one survey question. It's a list of survey responses.
8. it corresponds to every response within a certain data point/ question
9. It represents a variable

Student responses

1. it represents a variable and all of its data, such as heights or sex (it doesn't represent a student or a survey response for each row)
2. it represents all student responses for one field.
3. It represents all students' response on one question.
4. represents a variable and all its responses
5. the rows represent all the survey responses to one question.

Student responses

1. 'each row represents a student
2. A different student
3. Each of the rows represent a student.
4. Each row corresponds to a different student, as there are 98 rows and 98 responses. The 10 columns are the categories.
5. each row corresponds to a student
6. Each row represent a student, and column represent the question the student answered
7. Each row represents a student
8. Each row represents a student
9. Each row represents a student and their survey responses. Each row represents a student.

Student responses

1. Each row represents a student, the individual observation.
2. Each row represents an individual student
3. It represents a student
4. It represents a student.
5. Since there are 98 rows, each row represents one stude.
6. The observation unit is the student so each row represents a student's survey response

Student responses

1. Each row represents a student and contains all thier responses to the survey
2. each row represents te response of a student for that question
3. Each row represents the answer to each question that each student filled out on the survey
4. Each row represents the responses for each student
5. Each row responds to a student's responses.
6. it represents a students selections?
7. number of responses of survey
8. One of those rows represents a single student's responses to the 10 questions which are represented as columns
9. Rows represents a student who has completed the survey response.

Student responses

1. survey response

Student responses

1. A question
2. A row is the answers for multiple students for a particular category. The first row is for response id and each value is a student's response.
3. Each of those represents a student response. There may be duplicate student responses or missing student responses, thus it is not bijective to a list of students.
4. Each row responds to a student and survey response?
5. each row means a new survey
6. It represents a survey response, which if everyone filled it out once, would represent a student.
7. one of each row represents a response off one question from the student
8. one re

What's the participation rate among the 306 students invited to take the survey? (There are 120 students enrolled.)

103 anonymous responses

Percentage ≈81.7%	74
Responses giving the participation rate as about 81.6–81.7% or rounded to 82% (numeric percent formats and slight variants).	
Fraction form 98/120	13
Responses stating the participation rate as the fraction 98/120 or equivalent exact fraction expression.	
Percentage of all invited (~39.2%)	7
Responses that computed participation using invited students (98/306) yielding about 39.2% or similar decimals around 0.392.	
Other numeric errors	3
Responses with numeric values that don't match either correct calculations (e.g., 0.32026, 0.3921, 32%, 95, 116/120, 83).	

Response themes for question 4

Other responses	3
Responses the model could not place reliably.	
Other responses	3

Student responses

1. .81667
2. .81667
3. ~82%
4. 0.81
5. 0.816
6. 0.816
7. 0.816666667
8. 0.8166667
9. 0.8166667

Student responses

1. 0.8166667
2. 0.8167
3. 0.8167
4. 0.817
5. 0.817
6. 0.82
7. 80%
8. 81
9. 81.6%

Student responses

1. 81.6%
2. 81.6%
3. 81.6%
4. 81.6%
5. 81.6%
6. 81.6%, the 98 students that participated divided by the 120 students that could've participated
7. 81.6%.
8. 81.66
9. 81.66 percent

Student responses

1. 81.66%
2. 81.66%
3. 81.66% of students participated in the survey.
4. 81.666
5. 81.667%
6. 81.667%, as there are 98 rows (98 submissions) and 120 possible submissions.
7. 81.67
8. 81.67%
9. 81.67%

Student responses

1. 81.67%
2. 81.67%
3. 81.67%
4. 81.67% response rate. You need to know what the total number of students enrolled is and the number that filled out the survey, assuming no repeat
5. 81.7
6. 81.7
7. 81.7
8. 81.7%
9. 81.7%

Student responses

1. 81.7%
2. 81.7%
3. 81.7%
4. 81.7%
5. 81.7%
6. 81.7%
7. 81.7%
8. 81.7%
9. 81.7%

Student responses

1. 81%
2. 81%
3. 81% (98/120)
4. 82 percent response rate of the students enrolled in this clas.
5. 82 PERCNET
6. 82%
7. 82%
8. 82%
9. 82%

Student responses

1. 83
2. $89/120$
3. 98 student responses (assumed only 1 response per person) / 120 total students = 0.816666667
4. $98/120 = 0.816$
5. $98/120 = 81.7\%$
6. $98/120$, around 82%
7. About 81.67%
8. The participation rate is 81.67% since each student can only submit one response.
9. The participation rate is roughly 81.67%

Student responses

1. the rate is 81.7%
2. well we had 98 people do the survey and 120 enrolled so its about 98/120 or around 81 percent

Student responses

1. 98/120
2. 98/120
3. 98/120
4. 98/120
5. 98/120
6. 98/120
7. 98/120
8. 98/120
9. 98/120

Student responses

1. $98/120$
2. $98/120$
3. $98/120 = 0.82$
4. $98/120 = 0.816667$

Student responses

1. 0.392
2. 0.3921
3. 32.03%
4. 39.2 percent
5. 98/306
6. 98/306
7. The response rate among all students was 39.22%.

Student responses

1. 0.3202614379
2. 116/120
3. 95?

Student responses

1. 1/3
2. 98 students did it, or 81.6% of the 306
3. 98%

Student responses

1. $120/306$ is the participation rate
2. I guess it will be 78%.
3. $\text{nrow()}/120$

How would you check that `response\_id` is unique in every row? (State in words if you do not know the R code.)

93 anonymous responses

Vague or incorrect methods

31

Responses that are unclear, incomplete, or contain mistakes about which dimension to compare or which object to pass to unique.

Compare counts with unique

20

Suggest computing the number of unique response_id values and comparing it to the total number of rows or entries (e.g., `length(unique(...)) == nrow` or `length(...)`).

Use unique on the column

14

Recommend running unique on the response_id column to inspect or obtain the set of distinct ids (often without explicitly mentioning the length comparison).

Compare unique to full vector length

11

State comparing `length(unique(...))` to `length(survey$response_id)` or similar equalities as the check for uniqueness.

Response themes for question 5

Compare unique rows or data frames

5

Propose using `unique` on the whole data frame or comparing `nrow(unique(survey))` to `nrow(survey)` to detect duplicate rows/IDs.

Manual or iterative checking

5

Describe checking row-by-row, building a list, or visually scanning `unique` output to detect repeated ids.

Other responses

4

Responses the model could not place reliably.

Help or documentation requests

3

Indicate using help or consulting documentation for the `unique` function rather than giving a direct check.

Student responses

1. Call the number of unique vectors and compare it to the number of vectors in the dataframe
2. Can verify by matching the function length with the N response of survey IDs. can also use unique to check that it match the number of responses
3. check if there is repeated unique name
4. Check response_id is unique and present in each row.
5. Check to see if the number of rows is equal to the count of unique response_id values
6. compare counts of unique versus total
7. compare length and unique
8. count the number of unique responses versus the total number of responses and see if they are the same
9. find the last response id and see if it matches number of ids

Student responses

1. Find the length of the vector using unique and not using unique and compare the results.
2. length function combined with unique
3. `length(unique(survey$response_id, incomparables = FALSE))`
4. `length(unique(survey$response)) = ncol(survey)`
5. `ncol(unique(survey$response_id))`
6. Pass in the unique ones
7. put unique entries into vector, check length, then compare to length of original survey
8. Run the unique function on the variable survey
9. See if the number of unique responses is the same as the number of total responses.

Student responses

1. unique
2. unique and length function
3. unique(survey\$response_id) its length should be the same as the number of columns
4. Use a loop to prompt the unique() function to check for each.
5. Use the unique command in R, and check if the resulting data frame has the same number of entries using length.
6. use the unique function?
7. use unique to compare lengths
8. using a for loop
9. We would have to find a way to access response_id using the column types, and then run that through unique which should return all of the unique response ids

Student responses

1. would it be `length(unique(survey$response_id))`
2. You could run the unique function for the column
3. you use the unique function
4. You would run the unique function on the dataset and this would extract the unique elements in the data frame, and then you would compare that to the `response_id` value of 98

Student responses

1. Id run `length(unique(survey$response_id))`
2. `length(unique(survey$response_id))`
3. `length(unique(survey$response_id))`
4. `length(unique(survey$response_id))`
5. `length(unique(survey$response_id))`
6. `length(unique(survey$response_id))`
7. `length(unique(survey$response_id))`
8. `length(unique(survey$response_id))`
9. `length(unique(survey$response_id))`

Student responses

1. `length(unique(survey$response_id))`
2. `length(unique(survey$response_id))`
3. `length(unique(survey$response_id))`
4. `length(unique(survey$response_id))`
5. `length(unique(survey$response_id))`
6. `length(unique(survey$response_id))`
7. `length(unique(survey$response_id))`
8. `length(unique(survey$response_id)) == length(survey$response_id)`
9. `length(unique(survey$response_id)) == nrow(survey)`

Student responses

1. `length(unique(survey$response_id))`
2. `nrow(unique(survey$response_id))`

Student responses

1. Enter `unique(survey$response_id)` and check how many vectors it provides.
2. Find the number of unique response id's across the entire survey
3. I think what you would need to do is `unique(survey$response_id)` and you would be able to see manually if each idea is different or not
4. Id run "`unique(survey$response_id)`" and look to see if the total output measures the number o total rows from `glimpse()`.
5. Run the unique R code on the data set for the variable `response_id`
6. Run unique function to see if the number of entries has changed or not
7. unique function.
8. `unique(response_id, incomparables = FALSE)`
9. `unique(response_id)`

Student responses

1. `unique(survey)`
2. `unique(survey$response_id)`
3. `unique(x, incomparables = FALSE, ...)`
4. We would do `unique(survey$response_id)`
5. You would run a unique on the data (to return the elements that are unique)

Student responses

1. `$length(unique(survey$response_id))`
2. check if the length of the unique response ids is the same as the regular response ids
3. `length(survey$response_id)`
4. `length(survey$response_id)` compare to `unique(survey$response_id)`
5. `length(survey$response_id) length(unique(survey$response_id))`
6. `length(unique(survey$response_id)) == 98`
7. `return (length(survey$response_id) == length(unique(survey$response_id)))`
8. see if length of unique response id equals length of survey response id
9. Take `length(survey$response_id)` and `length(unique(survey$response_id))` and make sure they are the same

Student responses

1. To do this you must first just work with the response_id column. I would do `length(unique(survey$response_id))` as it would allow you to generate all the unique elements and then count how many.
2. You can use `length` to check the number of responses in the vector, and use `length(unique(survey$response_id))` to see that it is still 98 after removing duplicates.

Student responses

1. call the unique function the survey or use a for loop to check
2. compare the number of rows to the number of rows in the unique data frame
3. `nrow(unique(survey)) == nrow(survey)`
4. Running `unique(survey$response_id)` and seeing that the dimensions of the dataset are the same, meaning there were no duplicate rows to be removed, indicates that `response_ids` are unique
5. `unique(survey)`

Student responses

1. check through each row and add to a list and every time something is added chek if its in that list and then it should work
2. compare the id in every row and count if there is a very same id
3. I would
4. look through the keys
5. To check if the id is unique in every row, you can isolate to the response_id data and compare the number of different entries there are to the number of unique entries and see if they are the same.

Student responses

1. check if `length(unique(survey$response_id)) == length(survey$response_id)`
2. `length(unique(survey$response_id))=98`
3. use length and unique or use sum and duplicated `length(unique(survey$response_id)) sum(duplicated(survey$response_id))`
4. use the unique function

Student responses

1. help (unique)
2. help(unique)
3. you could ask using the help

Before calculating a mean or median, what should we do first? A. Calculate both and choose the smaller one B. Check the recorded units, missing values, and implausible values C. Divide the data by sex D. Remove the largest and smallest values

100 anonymous responses

Check and clean data first

98

Responses saying you should inspect and clean the data: verify recorded units, handle missing values, and flag implausible values before computing mean or median.

Other responses

2

Student responses

1. b
2. b
3. b
4. b
5. b
6. b
7. b
8. b
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B
8. B
9. B

Student responses

1. B
2. B
3. B
4. B
5. B
6. B
7. B, because the survey allowed for different units to be used, meaning they must all be the same unit first for accurate data to be made.
8. b, because the units could be different etc...
9. B, because you can't compare values of different units.

Student responses

1. B, check recorded units missing values etc. first
2. B, different responses have different values that cannot be calculated together
3. B; Check the recorded units, missing values, and implausible values
4. B: check recorded units, missing values, and implausible values
5. B.
6. B.
7. B. Check the recorded units, missing values, and implausible values
8. B. Check the recorded units, missing values, and implausible values
9. B. Check the recorded units, missing values, and implausible values

Student responses

1. B. Check the recorded units, missing values, and implausible values.
2. B. clean the data
3. B. Clean the nasty data and normalize it to the same units.
4. B. I need to understand the units and types of the vectors, because quantitative calculations may not function as needed for categorical.
5. check the recorded units, missing values, and implausible values
6. Checkthe recorded units, missing values, and implausible values.
7. I would say you need to do B, since some of them are in different units so that would make an inaccurate data check
8. We should first check the units, missing values, and implausible values.

Which statement best describes `raw\_median`?
A. It is the median height in the most common unit
B. It is the median height in inches because most values are inches
C. It orders numerical entries recorded on different scales and has no coherent interpretation
D. It is valid because medians are resistant to unusual values

95 anonymous responses

Median uninterpretable with mixed units

89

Students say raw_median merely ranks numbers recorded on different scales/units and thus has no coherent interpretation or unit.

Other responses

6

Student responses

1. c
2. c
3. c
4. c
5. c
6. c
7. c
8. c
9. c

Student responses

1. c
2. c
3. c
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C
2. C
3. C
4. C
5. C
6. C
7. C
8. C
9. C

Student responses

1. C because no units
2. C.
3. C.
4. C.
5. c. because it just looks at the raw number
6. C. It orders numerical entries recorded on different units
7. C. There is no single units value
8. It orders numerical entries recorded on different scales and has no coherent interpretation

Student responses

1. A
2. A. There is no specified unit, but likely is inches.
3. b
4. b
5. c
6. It doesnt make any sense

(Review from STATS 250) What does a histogram tell us?

107 anonymous responses

Frequency of values or responses

40

States that a histogram shows how often values or responses occur (counts or frequency).

Distribution, shape, center, spread

27

Emphasizes the distribution and its properties (shape, center, spread, skewness, outliers, normality).

Bins or ranges (bucketing) explained

14

Mentions that histograms group data into intervals or bins and show counts per bin.

Other responses

8

Responses the model could not place reliably.

Response themes for question 8

Visual summary/helps spot problems

8

Notes that histograms are visual tools to summarize data and help identify issues or compare values.

Incorrect or vague responses

6

Responses that are unclear, incorrect, or too vague to classify as above (mentions median, historical data, dataframe info, or single words).

Specific dataset/unit observations

4

Gives dataset-specific interpretations (e.g., bimodality, mixed units like inches vs centimeters).

Student responses

1. A frequency plot to visualize the distribution of numerical data
2. A histogram can visualize the frequency of different responses
3. A histogram is a visual representation of a variable and the distribution of its frequency/how many times the variable occurred at a value.
4. A histogram is a visual table to show frequency of responses across different values
5. A histogram shows the frequency of each response or group of responses on a graph
6. a histogram shows the frequency of replies for a given value
7. a histogram tells us how many counts there are for each segment of the independent data
8. A histogram tells us how many datapoints are within "bins" of certain amounts (for example, a bin that covers values between 60 and 80).
9. A histogram tells us how often specific values (within a given interval) are observed

Student responses

1. A histogram tells us the amount of responses that fall under certain intervals
2. A histogram tells us the frequency of a particular response in a dataset.
3. a histogram tells us the frequency of different responses.
4. A histogram tells us the frequency of each unit that is reported
5. A histogram tells us the frequency of values recorded between specific value intervals
6. A histogram tells us the frequency of values recorded, displayed visually with bucket sizes
7. A histogram will create different ranges or "buckets" to put numerical data points into. It helps break up the data and tells us the concentration of the data in a discrete way.
8. A histogram shows how many responses gave the same value or value range for a certain variable
9. A visual graphical representation of the frequency of data among different ranges

Student responses

1. Histogram shows the frequency of different response.
2. Histograms help us plot values to see how frequent a value occurs compared to other values
3. Histograms shows how frequent an outcome is in a dataset, and sorts the values into buckets to make the data easier to visualize
4. Histograms tell us how many reponses are in a certain numerical range.
5. how many responses of each variable there are
6. It displays the frequency distribution of responses.
7. It is a chart that shows the frequency of the var x
8. It is a chart that shows the frequency of different responses
9. It shows the frequencies of the data.

Student responses

1. It shows the frequency of the results given from the survey
2. It shows us the frequency distribution.
3. It tells us how many responses occurred within a set bin or range of values.
4. It tells us the frequency of the values from responses.
5. responses and the frequency of each response.
6. tells you how many numerical values of the data appear within a certain range
7. The amount of responses for a interval of "answers" or values.
8. The frequency of different responses
9. The frequency of reported responses

Student responses

1. the frequency of responses
2. The frequency of the data selected
3. The frequency of values that fall within a certain range.
4. This shows the distribution of the heights with each bar representing a different value

Student responses

1. A histogram gives us an idea of how the data is spread, as well as an idea of what the median is.
2. data distribution
3. Distribution of a variable
4. Distribution of numerical data
5. High in middle
6. Histogram tells us the distribution of a dataset
7. It is a visual display of the distribution of a data set. It can help us identify skew
8. It shows the frequency of data to infer the distribuion of data
9. It tells us a distribution of a sample of observations.

Student responses

1. It tells us distribution of te data
2. It tells us how the data goes, the changing rate ,the average level
3. It tells us the distribution of a variable.
4. It tells us the distribution of values
5. It tells us the distributon of all values and the frequency of the values
6. It tells us the frequency of the response values for a certain variable in a range format, unlike a bar chart.
7. It tells us values like center, spread, and shape, which can be used to summarize and undestand data
8. It tells us what it looks like. shows shape spread and any outliers, and if it is normal or not
9. It tells you the spread of the data and the frequency of values

Student responses

1. show data distribution
2. Shows the distrubtion of data values in numerical groups
3. tells us the distribution of data points
4. Tells us the frequency regarding our data and the skewness
5. The distribution of a continuous variable
6. The distribution of the responses
7. The histogram tells us about the frequency that is realized in our data. It allows us to see where the largest concentrations are based on the value on the x-axis
8. Visual distribution of the entries
9. What the data looks like - is it normally distrubuted?

Student responses

1. A graph that shows a number of responses in a range of answers.
2. A histogram is a graph that displays the distribution of a data set while not specifying specific values. Instead, it provides ranges.
3. A histogram is a visual representation of a quantitative characteristic's distribution across a set of data.
4. A histogram is a way of visualizing continuous data.
5. Bucketing answers into equally spaced intervals to see where most answers lie.
6. Gives the information of counts for each category or range.
7. histogram is a type of bar graph that shows how often certain values appear
8. It is a graphical representation of the data that splits the data into bins of fixed sizes.
9. It tells us the frequency at which units lie on certain ranges of values.

Student responses

1. number of each corresponding variables, distribution of the data
2. plots ranges of data or categorical on x and compare with result on y axis
3. teh frequency of a certian value within are dataset reagardless of unit
4. the frequency of responses corresponding to columns of a certain range
5. visual way to see how much values differ from one another

Student responses

1. A histogram provides a distribution of the values in a dataset
2. frequency of data at different values
3. Histogram tells us the distribution of the selected column of the data set
4. Histograms tell us the distribution of values for a variable
5. It tells us the frequency and distribution of the variable
6. It tells you how frequently people responded to a certain category
7. shows frequency of responses across the range
8. shows the different frequencies of value next to each other for easiness to compare

Student responses

1. displays data in a convenient format to view
2. distribution and frequency
3. distribution of different responses and frequency as well
4. It gives a visual interpretation of a set of data in the form of a graph with an x and y axis display information.
5. It gives us a visualization of the data so we can see if there are any problems with the data
6. It is a visual graphic to help us understand data via different categories.
7. It will give us how much frequency the other variable has
8. visual plot displaying the frequencies of responses, does not include individual data values

Student responses

1. freqieucy
2. historical data
3. it shows a distribution of the responses from the survey question
4. tells median
5. The information included in a dataframe
6. using histogram we can compare different variables, and check frequency

Student responses

1. According to that histogram, the reported height values is biomodal; however, considering the unit is mixed, thhis shape is resonable
2. It shows us how most people are doing inches. The small bump at the end is the cm people
3. It visualizes the frequency of answers, so there is a huge frequency around 60 and 170 which highlights inches versus centimeters
4. We either have some NBA starts in this class, or, more likely, we have different scales for this numerical data responses.

Which is the strongest conclusion currently supported? A. ``raw_median`` is the median height of all DATASCI 306 students B. ``raw_median`` is the median height of all U-M students C. The raw values use different units, so they must be standardized before we estimate a typical height among respondents with usable values D. The survey contains no useful information

0 anonymous responses

No responses were submitted.

On the problem set, you will be asked to choose two survey columns and compare them. State a comparison that seems interesting to you. Name the two columns and describe what you would compare.

0 anonymous responses

No responses were submitted.