BIOL 3000: Project 1 In Class Worksheets

Group Members: Hannah M., Ainsley B., Ella C., Ben S., Hannah G., and Elissa O.

Topic: People on campus walk at different speeds

Worksheet 1: Hypothesis and Predictions (September 8th, 2025)

Group Members Present: Hannah M., Ainsley B., Ella C., Ben S., Hannah G., and Elissa O.

Observation: People walk at different speed on campus.

Hypothesis #1: Walking speed changes with peoples energy levels.

Hypothesis #2: Students walking speed is influenced by their distraction level.

Hypothesis # 3: The location someone is walking influences their speed.

Tests	Depends on day/hour	Depends on distracted walking	Depends on indoor/outdoor
Time people walk between 2 markers in morning vs evening indoor/outdoor	Walk faster in morning compared to evening	No difference between distracted and not distracted	Walk faster outdoors
Record speed of people walking, whether or not they are distracted by phone, group or alone, morning vs evening	No difference in time of day	People walk faster when not distracted or in a group	No difference inside vs outside

Methods:

- Time ppl walking from point A to B
- 1 person timing inside, 1 timing outside (at the same time, morning and evening for 30 minutes each)
- Ensure consistent distance of point A to B inside and outside
- Record the time it takes for each person
- Record if alone or distracted (in a group or on phone)

Literature Sources:

Hypothesis #1:

Intra-day variation in daily outdoor walking speed

https://link.springer.com/article/10.1186/s12877-021-02349-w#Sec2

Kawai, H., Obuchi, S., Hirayama, R. *et al.* Intra-day variation in daily outdoor walking speed among community-dwelling older adults. *BMC Geriatr* **21**, 417 (2021). https://doi.org/10.1186/s12877-021-02349-w

Hypothesis #2:

Walk speed decreases linearly with growing group size: https://pmc.ncbi.nlm.nih.gov/articles/PMC2850937/

How distraction affects pedestrian response: https://www.sciencedirect.com/science/article/pii/S1369847822002583

Hypothesis #3:

People walk faster outdoors: https://www.neurores.org/index.php/neurores/article/view/187/201

Walking speeds outdoors: https://link.springer.com/article/10.1007/s40279-020-01351-3

Summary:

What test(s) are you going to do for the project?

We are going to test people's walking speed in the morning and evening, and inside and outside. We will also record whether the person walking was distracted (on their phone or in a group) or not distracted. We will do these tests simultaneously for 20 mins in the morning, afternoon, and evening.

How will this test(s) allow you to answer the research question?

We will be gathering a lot of data telling us walking speed over a certain distance. We are also gathering many explanatory variables to plot against that walking speed data. The test will have a high sample size since there are currently a lot of people walking on campus. With the options we have of variables to compare, answering the original question should be feasible.

Are these tests feasible to do in the next couple weeks?

Yes, these tests are feasible to be completed within the next couple weeks. Classes are in session so there is a large sample size, and supplies are easy to acquire.

Worksheet 2: Experimental Design (September 10th, 2025)

Group members present: Hannah G, Hannah M, Ben, Elissa, Ella, Ainsley

What is the response (dependent) variable?

Walking speed

How could it be measured?

Time people walking a known distance and divide to find average speed in m/s.

What type of variable would it be in each case?

Numerical

What are the units/what format is it?

Meters/second (m/s)

What is the first explanatory (independent) variable?

Time of day

How will it vary?

Morning, afternoon, and evening (8:00, 12:00, and 16:00)

How could it be measured?

By time since 8am

What type of variable would it be?

Numerical (time in hours since 8am)

What are the units/what format is it?

Hours (24 hour clock)

What is the second explanatory variable?

Indoor versus outdoor location

How will it vary?

Inside or outside

How could it be measured?

The location of data collection

What type of variable would it be?

Categorical

What are the units/what format is it?

Either inside or outside (no units)

What is the third explanatory variable?

Distraction (via phone or walking with other people) or not distracted

How will it vary?

Either distracted or not distracted

How could it be measured?

Subjective assessment

What type of variable would it be?

Categorical (distracted = true and not distracted = false)

Could also be numerical (binomial 0, 1)

What are the units/what format is it?

Either distracted or not distracted (no units)

What factors could confound the results

- Classes time/events/reason for walking (going to/leaving class…)
- Weather (outside)
- How busy the path is
- Direction people are walking in (accounted for this by only including people walking in one specific direction)

What is being kept constant and how?

Time of observation is being constant by ensuring we have people collecting data inside and outside at the same times throughout the day.

We will collect data from 8:10-8:30, 12:10-12:30, and 4:10-4:30.

Walking direction: outside = people walking toward old main, inside = people walking into old main

What biases are possible?

Unconscious biases in picking people to time/observe their walk speed, and whether they are considered distracted or not.

How will you avoid/minimize these?

Have all group members pick people to time and have a large number of people timed (high sample size), and have clear criteria for what counts as “distracted.”

How are you ensuring data is randomly collected?

Consistency among individual data collectors timing technique. Start and stop timing when foot crosses the marker. No bias when choosing people to measure, choose whoever is first to cross the marker.

How many categorical variables do you have?

Two (inside/outside & distracted)

How many categories are in each categorical variable?

Each has two categories.

How many numerical variables are there?

One (time of day)

Total # categories = #categories(var1) X #categories(var2) X 3 per numerical var

= 2 x 2 x 3 = 12

Minimum sample size = Total # categories X 5 samples per group

= 12 x 5 = 60 (10 per location, per time slot)

What will data collection look like?

Pairs of data collectors inside and outside at 3 time points in the day for 20 minutes each. One person times the person walking, other person records categorical data.

How long with it take to get one sample?

Less than 1 minute.

Can this be made more efficient?

Be in pairs, 1 person time, 1 person record distracted/not distracted and what distraction (phone/group).

Write out your protocol:

1- Tuesday Sept 16th. Set up markers (lines of tape) inside old main in front of Starbucks and outside old main along straight pathway beside basketball courts at a morning (8:10-8:30 am), afternoon (12:10-12:30 pm) and evening time (4:10-4:30 pm). Measure walking distance in metres for calculations later.
2- Data collectors (2) outside observe pedestrians walking TOWARDS Old Main and data collectors (2) inside observe pedestrians walking INTO Old Main. One observer uses stop watches and records the seconds to walk from marker to marker (phone). Second observer records whether people being timed are on phone or not AND if alone or with 1+ person
3- Observe and collect data for 20 minutes, continually collecting samples.
4- Convert the walking speed vs travel distance into rate (m/s)
5- Data analysis

Time slots (2 people per slot, Tuesday the 16th);

8:10-8:30

Inside: Ella, Ben

Outside: Ainsley, Hannah M

12:10-12:30

Inside: Elissa, Hannah M

Outside: Ben, Hannah G

4:10-4:30

Inside: Ainsley, Hannah G

Outside: Elissa, Ella

Worksheet 3: Experimental Design Part B (September 12th, 2025)

Group members present: Hannah G, Ben, Elissa, Ella, Ainsley

All data collection forms should be created in google forms. (EXCEL)

What are the columns?

Time slot (8:10am, 12:10pm and 4:10pm), inside/outside, distance walked (m), walking time (s), distracted (yes or no), walking speed (m/s)

Will you be able to calculate your response and explanatory variables from these? If not, is there an alternative design that you could use?

Yes, we can calculate m/s from measured times and length between markers.

Test Protocol:

What worked?

Successfully able to find 2 markers and measure the distance between them

How long did it take?

20 minutes

Given sample size needed, is this feasible?

Yes, lots of people walking through campus every hour.

What didn't work?

Some people deviate from the path, so must not use their data unless they complete the full distance between markers.

What did you modify?

Group decision about exactly when to start and end the timer (when foot crosses the marker line).

Designate one data collector to timing and one to recording the categorical data.

Ensuring distance being measured is the same for each time period (set markers ahead of time).

Test the protocol a second time.

Does it work?

Yes

Is the data being input correctly?

Yes, having 2 people collecting data together allows one person to time and one person to record categorical data.

Worksheet 4: Summary Statistics and Univariate Graphs (September 24th, 2025)

Group members present: Hannah G, Ben, Elissa, Ella, Ainsley

Record who completed the following:

Response Variable (walking speed): Ella

Explanatory Variable 1 (time of day): Ainsley

Explanatory Variable 2 (distracted/not distracted): Ben

Explanatory Variable 3 (outside/inside): Elissa

Summary discussion: Hannah G.

RESPONSE VARIABLE: walking speed (m/s)

For each answer record as much detail as possible.

1. Is there any missing data

a. Record what R function you used to do this

summary()

b. Is there any missing data? If so, are these real missing values or should they be 0's?

No missing data. If 0s were present it would represent a stationary person.

c. If they are real missing data, are they appearing as "NA" or blanks?

No missing data.

d. If your answers to any of these questions is that the data is not appearing as it should, use R to make clean this up.

No missing data, so R was not needed to clean data.

2. Are there strange (unrealistic) values?

a. Make a graph and use summarizing functions to assess. Record what functions you used to do this.

ggplot() and geom_histogram

summarize(max_speed(), mean_speed)

b. Are there any strange values?

No strange values were observed. All values were real numbers and there were no values for speed that were unreasonably fast or slow.

c. Are these values real or typos that you can tell what the real value is or typos that are not discernable?

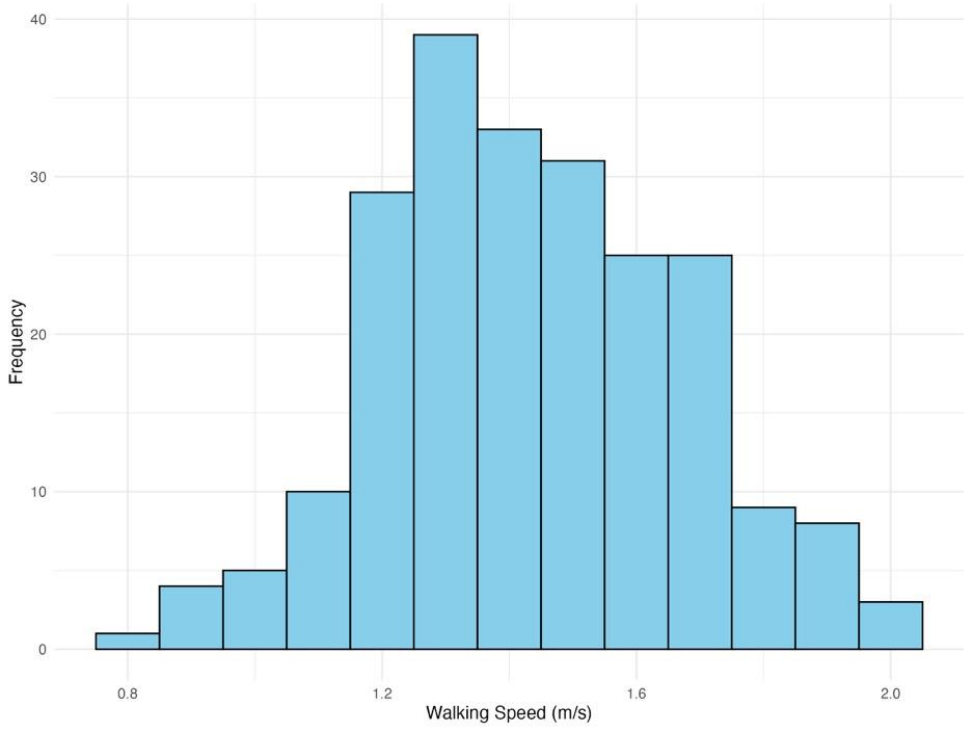
All values were real numbers and no typos were observed.

d. If there are strange values, use R to clean these as appropriate.

No strange values were observed, so R was not needed to clean these values.

3. Determine the distribution type

a. Plot the variable (copy a quick sketch here)



b. What is the distribution type?

The plot of walking speed above shows normal distribution. The data peaks are close to the middle of the plot and there is only one peak present, so the data is unimodal. There is a slightly longer 'tail' on the left side compared to the right which is why it could be argued the data is slightly left skewed, but a normal distribution seems to fit the curve better.

4. Assess the central tendency and spread

a. Calculate the appropriate summary statistic for the central tendency according to the distribution.

The distribution is normal so the appropriate summary statistic for central tendency would be the mean. Mean speed was calculated to be 1.431 m/s using mean () in R.

b. Calculate the appropriate summary statistic for the spread according to the distribution.

The distribution is normal so the appropriate summary statistic for spread is standard deviation. The standard deviation for speed was calculated to be 0.235 m/s using sd () in R.

c. Is there a lot variation in the variable? Thinking about how variable this variable could be, has the data captured most of that variability? If not, how is this data skewed or biased?

Yes, the standard deviation is moderate, so there is some variation in the data. This variable could be very variable depending on whose walking speed is being measured. For example, a child's walking speed will be different than a university student. Therefore, the data would be biased towards the walking speed of an average university student because data collection was done on a university campus.

5. What about the structure of this variable needs to be kept in mind as your group proceeds with analysis?

This variable represents the walking speed, which should not be confused with walking time. The units are different and by measuring the speed instead of time the data accounts for variation in the distance walked.

What about this variable needs to be kept in mind when interpreting results and what they mean for the real world from any analysis?

It needs to be kept in mind that the walking speed was calculated as the average time it took someone to walk a set distance. We tried to choose a distance small enough to make timing reasonable and large enough to account for any slight differences in speed over that distance; however, the speed likely varied over the distance.

EXPLANATORY VARIABLE 1: Time of day (numerical)

1. Is there any missing data

a. Record what R function you used to do this

summary*(tablename)*

b. Is there any missing data? If so, are these real missing values or should they be 0's?

There was no missing data. Any 0s present would represent 8:00 am and since sampling started at 8:10 no 0s were present in the data.

c. If they are real missing data, are they appearing as "NA" or blanks?

There was no missing data.

d. If your answers to any of these questions is that the data is not appearing as it should, use R to make clean this up.

No missing data, so R was not needed to clean.

2. Are there strange (unrealistic) values?

a. Make a graph and use summarizing functions to assess. Record what R functions you used to do this.

project1<-mutate(project1, format_time = hm(Hour.into.the.day))

project1<-mutate(project1, time_since_8am=format_time-hm("8:00"))

project1<-mutate(project1, decimal_hours=hour(time_since_8am) + minute(time_since_8am)/60)

ggplot(data=project1, aes(x=decimal_hours))+ geom_histogram(bindwidth=0.1)

b. Are there any strange values?

No strange values were observed.

c. Are these values real or typos that you can tell what the real value is or typos that are not discernable?

No strange values were present, and all values were real values.

d. If there are strange values, use R to clean these as appropriate.

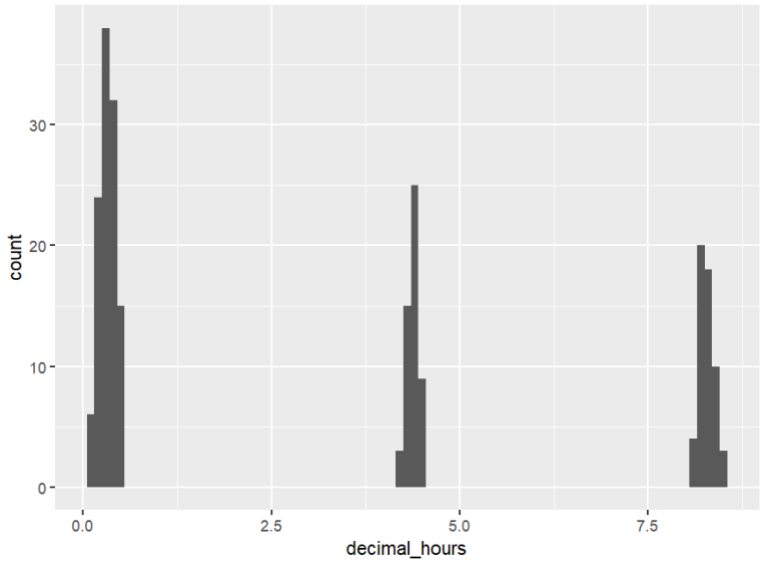
No strange values, so R was not needed to clean them.

What type of variable is this? If numerical, complete steps 3, 4, and 7. If categorical, complete steps 5, 6, and 7.

Time of day is a numerical variable.

3. Determine the distribution type

a. Plot the variable (copy a quick sketch here)



b. What is the distribution type?

The distribution is trimodal for the time-of-day variable which was expected because data collection occurred at three different times of day and not throughout the day.

c. Could this variable be converted to a categorical variable? Should it be?

The time-of-day variable should be converted to a categorical variable with three categories, Morning, Afternoon, and Evening. However, we required a numerical explanatory variable to fit project criteria so we cannot convert the variable to categorical.

4. Assess the central tendency and spread

a. Calculate the appropriate summary statistic for the central tendency according to the distribution.

The data has a non-normal distribution, it is trimodal, so the appropriate central tendency would be median. The median was calculated to be 0.4833 using median () in R. The mean was also calculated to be 3.2485 using mean () in R. However, the mean is not very helpful as the data does not have normal distribution.

b. Calculate the appropriate summary statistic for the spread according to the distribution.

Because the distribution is trimodal, the appropriate summary statistic for spread is the range. The minimum and maximum values were calculated using min () and max () in R to be 0.15 hrs and 8.48 hrs respectively. The range was calculated to be 8.33 hrs using range () in R. This means data collection spanned a total of 8.33 hours of the day.

c. Is there a lot of variation in the variable? Thinking about how variable this variable could be, has the data captured most of that variability? If not, how is this data skewed or biased?

No not a lot of variation is present. The data is biased to only three time points throughout the day and was not measured over time throughout the entire day.

7. What about the structure of this variable needs to be kept in mind as your group proceeds with analysis?

The variable is trimodal, so it can be difficult to treat as a single variable and should likely be converted to a categorical variable.

What about this variable needs to be kept in mind when interpreting results and what they mean for the real world from any analysis?

The variable should probably be categorical, but because we need a numerical variable the data is biased to only three time points.

EXPLANATORY VARIABLE 2: Distraction (categorical)

This variable measures if the person walking is or is not distracted.

1. Is there any missing data?

No missing data was observed.

a. Record what R function you used to do this

summarize(walkSpeed)

b. Is there any missing data? If so, are these real missing values or should they be 0's?

No missing data. If the data was a 0, it would indicate incorrect data input because data should be represented as either yes or no or true or false.

c. If they are real missing data, are they appearing as "NA" or blanks?

No missing data so no blanks present.

d. If your answers to any of these questions is that the data is not appearing as it should, use R to make clean this up.

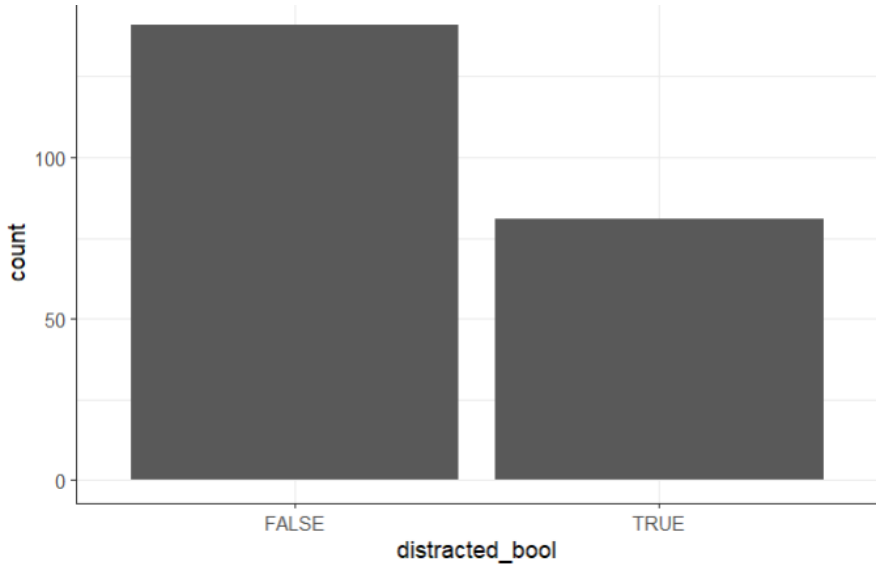
This line of code was used to fix the column to Boolean values instead of the yes/no strings:

walkSpeed <- mutate(walkSpeed, distracted_bool = Distracted=="yes" | Distracted=="Yes")

2. Are there strange (unrealistic) values?

No strange variables were observed all values were either true or false.

a. Make a graph and use summarizing functions to assess. Record what R functions you used to do this.



ggplot with geom_bar() was used to plot a bar chart of the two possible values.

Also used group_by and summarise to check the count of each value in both the columns that have yes/no strings and the column with Booleans to check they match.

b. Are there any strange values?

All the values were normal, either true or false.

c. Are these values real or typos that you can tell what the real value is or typos that are not discernable?

All the values are real values.

d. If there are strange values, use R to clean these as appropriate.

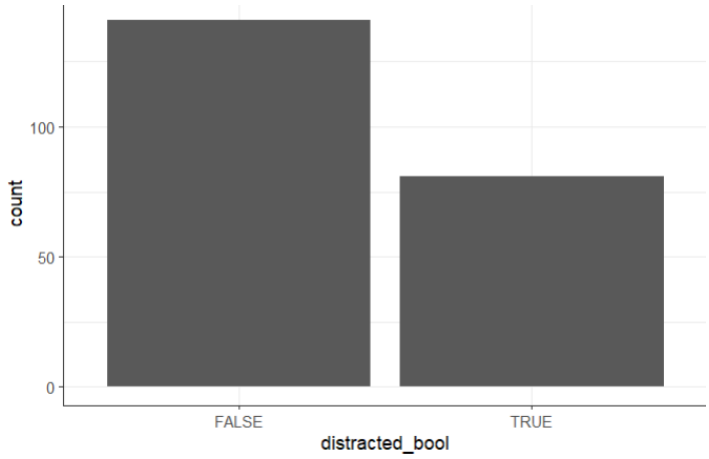
No strange values so R was not needed to clean data other than to change character strings to Boolean values.

What type of variable is this? If numerical, complete steps 3, 4, and 7. If categorical, complete steps 5, 6, and 7.

This variable is categorical with two possible categories.

5. Determine the distribution

a. Plot the variable (copy a quick sketch here)



b. How is the data distributed across categories?

There is almost double the number of False values than True, indicating most of the walking speeds collected were from non distracted people.

c. Could this variable have been collected as numerical variable? Should it have been?

No, it would have been impractical to do this as a numerical variable. It could technically be converted into a binomial variable, but it's unnecessary.

6. Assess the spread

a. Do any categories have low sample sizes?

No, the distracted category has a slightly smaller sample size, but it's large enough.

b. Are there categories that could logically be combined to increase sample sizes?

No

c. Use R to combine these categories together and then re-assess the distribution. Make a quick sketch here.

N/A

7. What about the structure of this variable needs to be kept in mind as your group proceeds with analysis?

The yes/no distinction of distracted or not could be ambiguous as to what distracted actually means.

What about this variable needs to be kept in mind when interpreting results and what they mean for hthe real world from any analysis?

Whether a person was distracted or not was determined by the use of their phone or by visibly interacting with someone else while walking. There are other factors that could contribute to someone being distracted, but only these were accounted for.

EXPLANATORY VARIABLE 3 (Inside/outside): categorical

1. Is there any missing data

a. Record what R function you used to do this

summary (rawdata)

b. Is there any missing data? If so, are these real missing values or should they be 0's?

No missing data

c. If they are real missing data, are they appearing as "NA" or blanks?

No missing data

d. If your answers to any of these questions is that the data is not appearing as it should, use R to make clean this up.

NA

2. Are there strange (unrealistic) values?

a. Make a graph and use summarizing functions to assess. Record what R functions you used to do this.

count <- rawdata %>% group_by (Location) %>% summarize (count=n())

```
ggplot (data=rawdata,
    aes (x=Location))+
     geom_bar()
```

b. Are there any strange values?

No all values are either inside or outside

c. Are these values real or typos that you can tell what the real value is or typos that are not discernable?

d. If there are strange values, use R to clean these as appropriate.

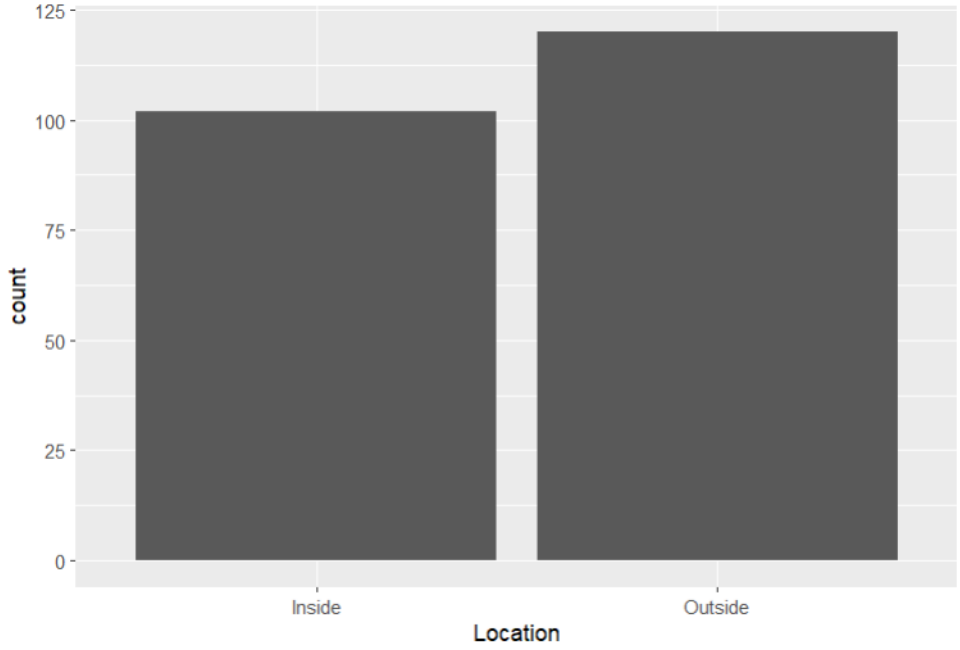
What type of variable is this? If numerical, complete steps 3, 4, and 7. If categorical, complete steps 5, 6, and 7.

Catagorical

5. Determine the distribution

There is none its categorical data with only two categories.

a. Plot the variable (copy a quick sketch here)



b. How is the data distributed across categories?

46 % inside and 54 % outside so more samples were collected outside than inside.

c. Could this variable have been collected as numerical variable? Should it have been?

It could have been collected as a binomial numerical variable but that wouldn't have made much sense because not clear which one would correlate to 0 and which would correlate to 1.

6. Assess the spread

a. Do any categories have low sample sizes?

No, inside is lower than outside but the data is not incredibly skewed one way or the other.

b. Are there categories that could logically be combined to increase sample sizes?
No the categories cannot be combined any further.

c. Use R to combine these categories together and then re-assess the distribution. Make a quick sketch here.
N/A

7. What about the structure of this variable needs to be kept in mind as your group proceeds with analysis?

Because there are only two categories the amount of analysis we can do on the data is very limited.

What about this variable needs to be kept in mind when interpreting results and what they mean for the real world from any analysis?

These were very specific locations observed for only one day so applications of these results should consider that what is seen "inside" doesn't apply to all indoor environments but is representative of a specific indoor environment on TRU campus.

SUMMARY DISSCUSION:

Discuss with your group what you have learnt about each variable.

We learned that "time of day" would've been better treated as a categorical variable rather than a numerical variable. The other two explanatories are both categorical with two options so they are pretty straightforward, but it was good to check that their columns have clean data.

What are the issues that still need to be solved?

The time since 8am needs to be converted into a decimal so we can use it in analyses, as it's more difficult to work with as a time variable.

What is your plan for solving these?

Data <- mutate(Data, decimal_hours = hour(time_since_8am) + minute(time_since_8am)/60)

Worksheet 5: Exploratory Graphs (September 26th, 2025)

Group members present: Hannah G, Hannah M, Ella C, Ainsley B, Elissa O, Ben S

Part 1: For each pair of explanatory variables: 1. Plot the variables against each other 2. Determine if there is a relationship 3. Assess what the implications of that relationship will be on future analyses

Part 2: For each explanatory variable: 1. Plot the variable against the response variable. 2. Determine the shape of the relationship 3. Determine if any of the other explanatory variables might confound this relationship 4. Assess

Group Member participation

Explanatory Variable 1 = Time of day

Explanatory Variable 2 = distracted

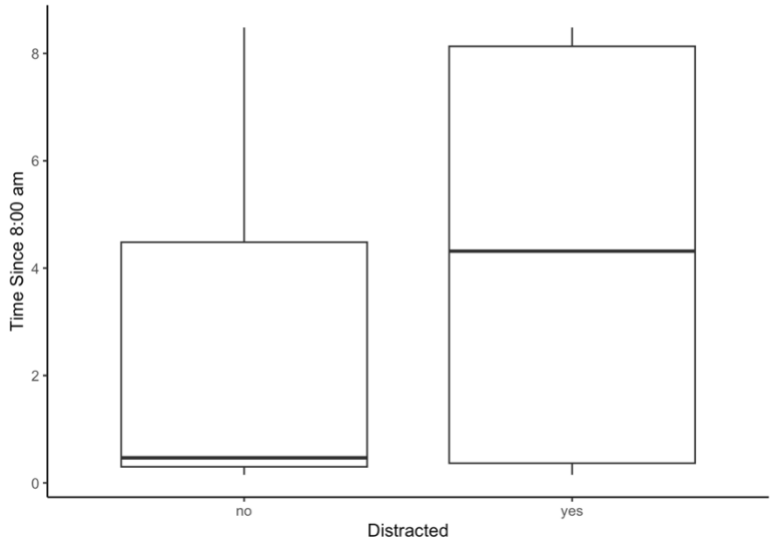
Explanatory Variable 3 = inside/outside

1.Split up the first section of these worksheets by pair of variables. Record who completed the following....
Explanatory Variable 1 vs Explanatory 2: Ben
Explanatory Variable 1 vs Explanatory 3: Hannah M
Explanatory Variable 2 vs Explanatory 3: Ella

Split up the second section of the worksheet by explanatory variable. Record who completed the following...
Explanatory Variable 1 vs Response: Elissa
Explanatory Variable 2 vs Response: Hannah G
Explanatory Variable 3 vs Response: Ainsley

Summary discussion:

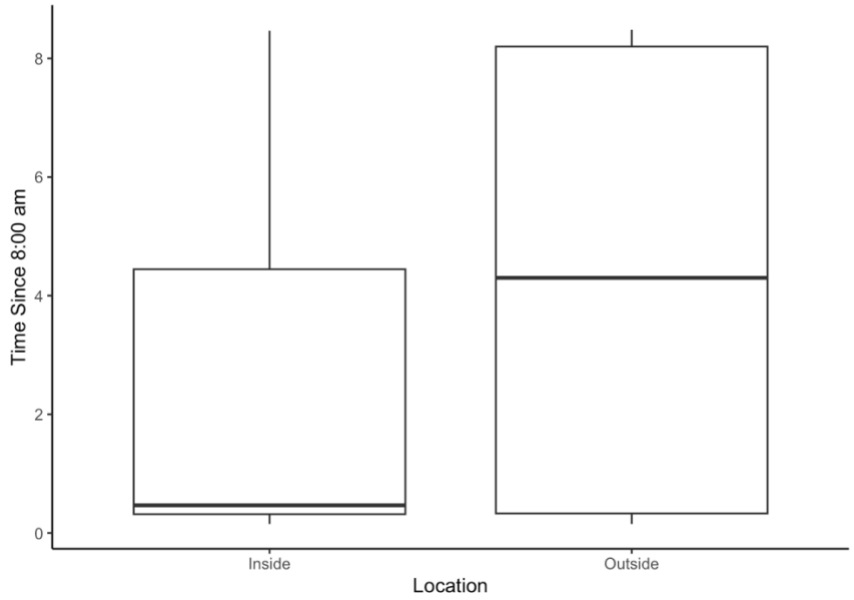
Section 1 – Assessing relationships between explanatory variables
Explanatory 1 is Time since 8 am, Explanatory 2 is distracted/not distracted. Create a plot in R to assess the relationship between these two variables. (make a rough sketch here)



What is the relationship between the variables?
People are more distracted later in the day than they are in the morning.

Could this confound interpretation of the results of an analysis? How?
People might walk slower because they are distracted not because it is later in the day

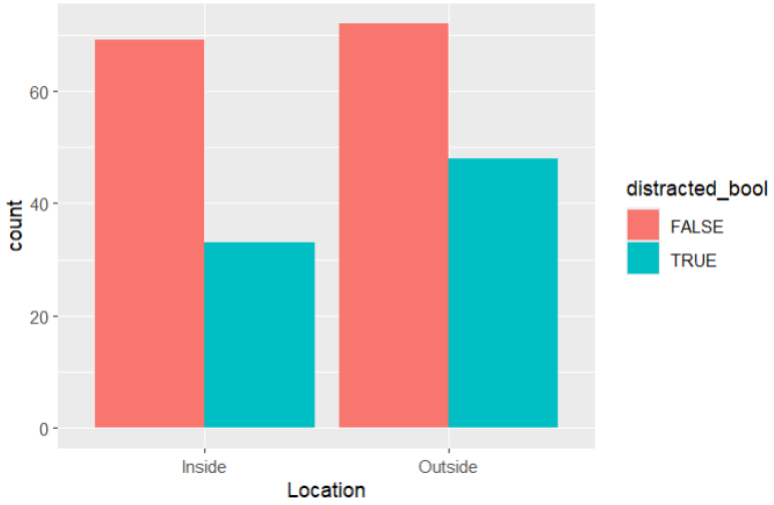
Explanatory 1 is Time since 8 am, Explanatory 3 is Location (inside/outside). Create a plot in R to assess the relationship between these two variables. (make a rough sketch here)



What is the relationship between the variables?
More people were walking outside later in the day than inside.

Could this confound interpretation of the results of an analysis? How?
People might've been walking slower because they are outside not because it is later in the day.

Explanatory 2 is Distracted/Not distracted, Explanatory 3 is Inside/Outside. Create a plot in R to assess the relationship between these two variables. (make a rough sketch here)



What is the relationship between the variables?
There are a bit more distracted people outside than inside but it doesn't look like a definite relationship
Could this confound interpretation of the results of an analysis? How?
This would probably not affect any analysis

What variables have no relationship between them?
Distracted and Location appear to have no relationship between them.

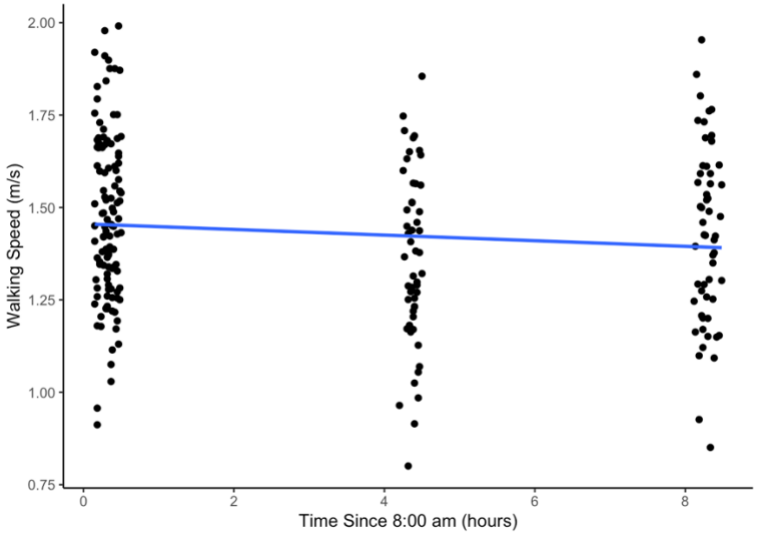
What variables have a relationship between them?
The data shows there are more people walking outside and distracted later in the day than earlier in the day.

What concerns do you have moving forward?
No major concerns since these relationships above aren't defined but we should keep it in mind that it could be confounding.

What is your plan for dealing with these concerns?
Keeping this in mind while doing our analysis and just being careful when drawing conclusions about walking speed at different times of day.

Section 2 – Assessing relationship between explanatory and response

1-Response is Walking speed, Explanatory 1 is Time of day since 8 am. Create a plot in R to assess the relationship between these two variables.



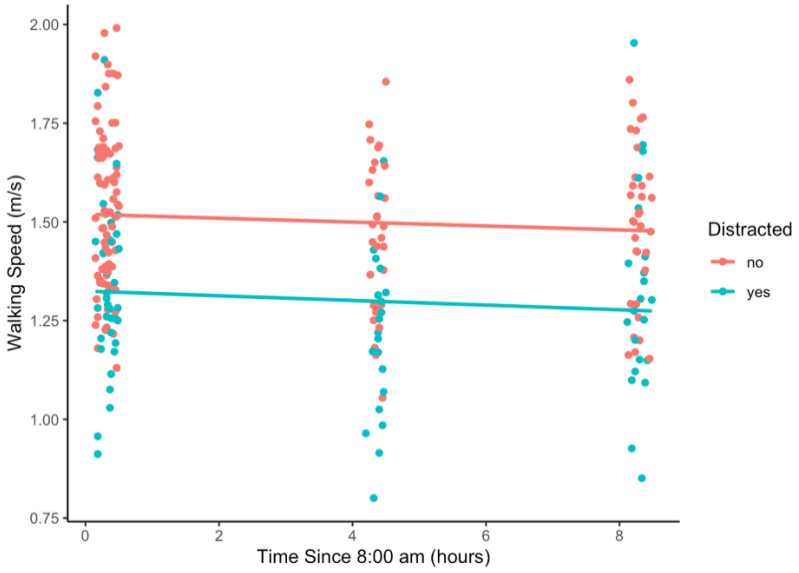
What is the relationship between the variables?

There is a slight decrease in walking speed as the time in hours since 8 am increases.

What shape is this relationship?

No relationship or slightly decreasing linear relationship.

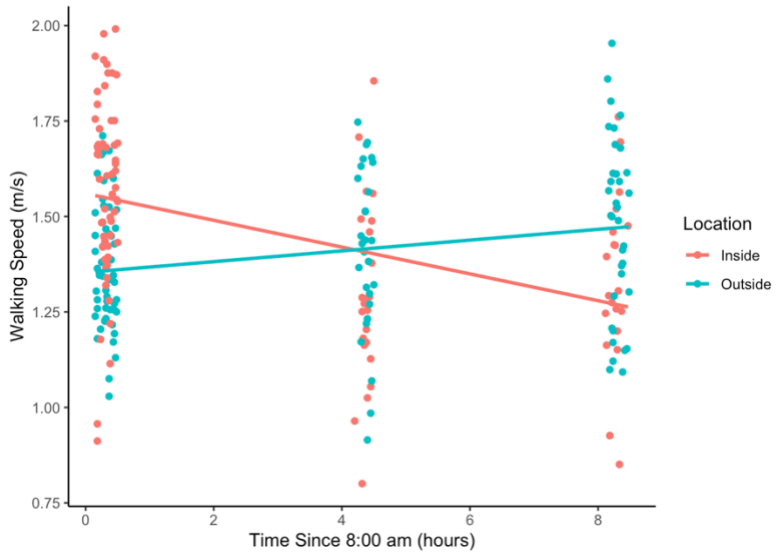
Adjust the graph so explanatory 2 (distracted/not distracted) is represented with colour.



Does considering this explanatory variable change the relationship? How?

No, the relationship still shows the slightest decrease in speed over time; however, it also shows that distracted people walked at a slower speed than non-distracted people, but the relationship over time was the same.

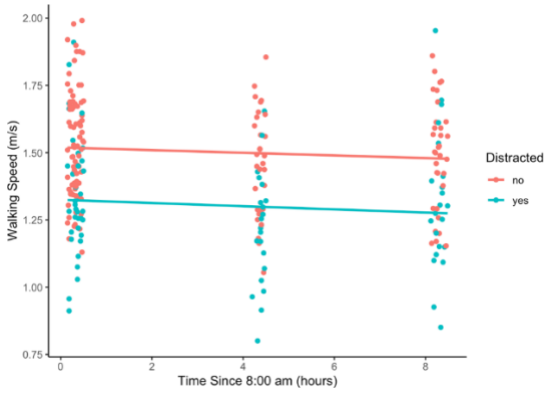
Adjust the graph so explanatory 3 (location) is now represented with colour.



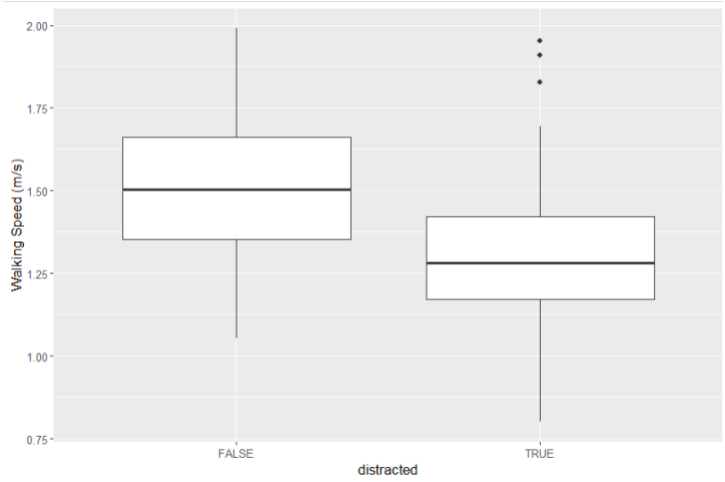
Does considering this explanatory variable change the relationship? How?

Yes, we can see that the speed inside decreased and the speed outside increased over time.

Create a rough sketch of the graph that best shows how your explanatory variable influences the response. (Include confounding other explanatory variables if they exist)



2-Response is Average Walking Speed Explanatory 2 is Distracted vs Not. Create a plot in R to assess the relationship between these two variables.

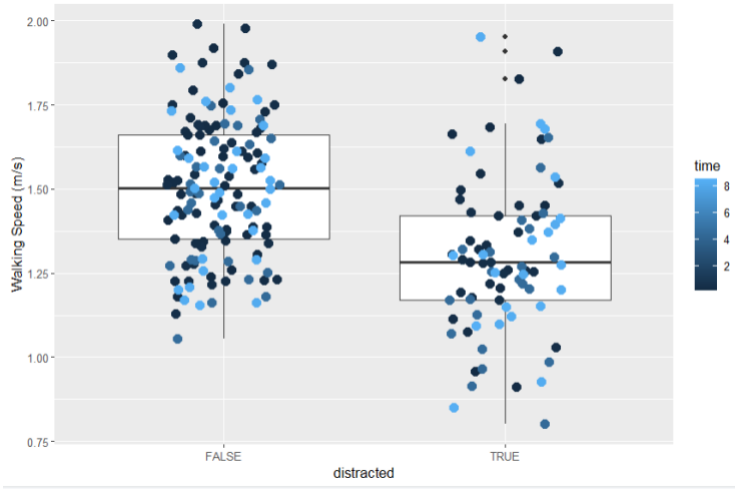



```
ggplot (data=walk,
      aes(x=distracted_bool, y=Walking.speed..m.s.))+
 geom_boxplot()+
 labs(y="Walking Speed (m/s)", x="distracted")
```

What is the relationship between the variables?
Average walking speed higher when not distracted

What shape is this relationship?
N/A

Adjust the graph so explanatory 1 (time of day) is represented with colour. Does considering this explanatory variable change the relationship? How?

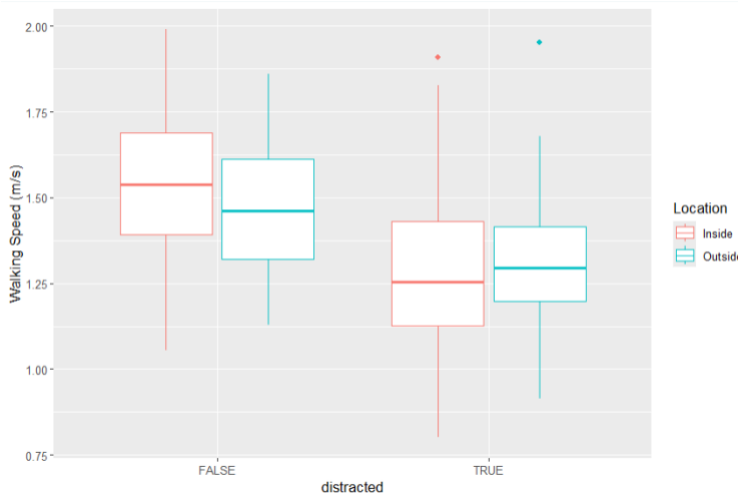


No, time didn't have a noticeable effect on distraction vs walking speed.

```
ggplot (data=walk,
      aes(x=distracted_bool, y=Walking.speed..m.s.))+
 geom_boxplot()+
```

```
geom_jitter(aes(color=time), width=0.2, size=3)+
labs(y="Walking Speed (m/s)", x="distracted")
```

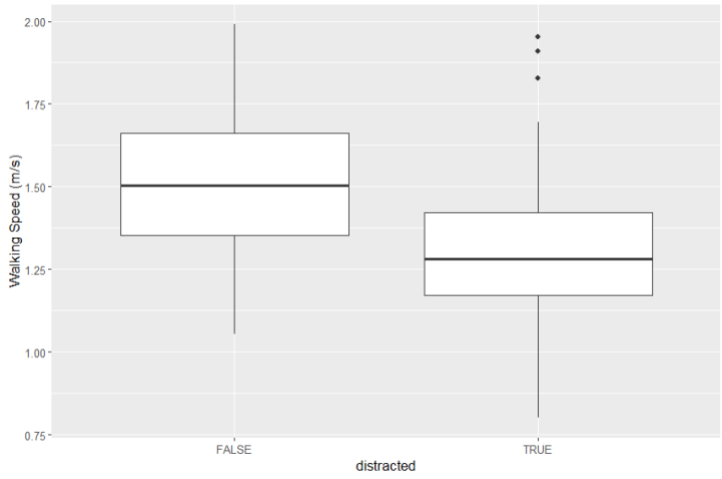
Adjust the graph so explanatory 3 (inside/outside) is now represented with colour. Does considering this explanatory variable change the relationship? How?



Location effected the walking speed more when not distracted (notice speed change between inside vs outside when no distraction). Although, the change is very slight.

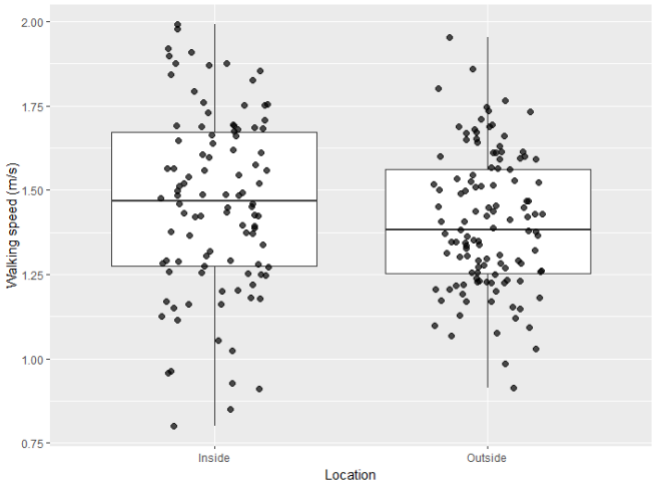
```
ggplot (data=walk,
     aes(x=distracted_bool, y=Walking.speed..m.s., col=Location))+
 geom_boxplot()+
 labs(y="Walking Speed (m/s)", x="distracted")
```

Create a rough sketch of the graph that best shows how your explanatory variable influences the response. (Include confounding other explanatory variables if they exist)



3-Response is **walking speed** Explanatory 3 is **location** **(inside/outside)**

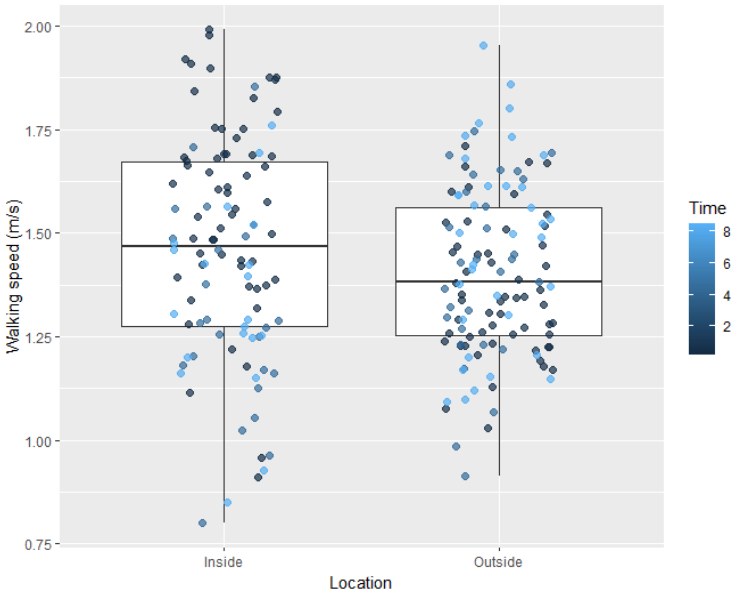
Create a plot in R to assess the relationship between these two variables.



What is the relationship between the variables?
The walking speed, on average, is higher inside (n=102) rather than outside (n=120). However, there is a larger spread among speeds recorded inside than outside.

What shape is this relationship? N/A

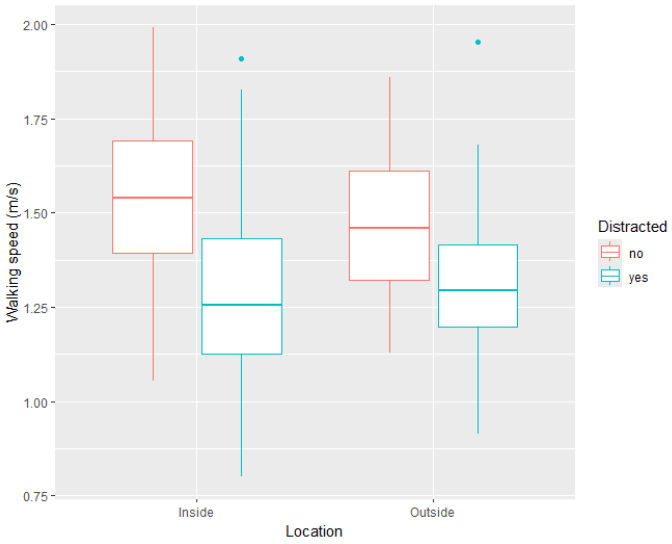
Adjust the graph so explanatory 1 (**time of day)** is represented with colour. Does considering this explanatory variable change the relationship? How?



ggplot(data=project1, aes(x=Location, y=Walking.speed..m.s.)) + labs(x="Location", y="Walking speed (m/s)", color="Time") + geom_boxplot() + geom_jitter(aes(color=decimal_hours), width=0.2, alpha=0.7, size=2)

Considering the time of day does not impact the relationship between walking speed and location as the colours are generally evenly distributed. Something that does stick out is the considerable number of points associated with indoor/morning walking speeds (darkest blue) being faster.

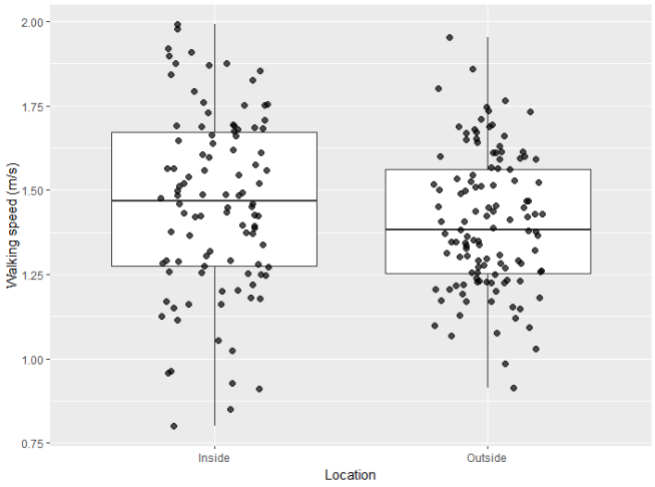
Adjust the graph so explanatory 2 (**distraction)** is now represented with colour. Does considering this explanatory variable change the relationship? How?



ggplot(data=project1, aes(x=Location, y=Walking.speed..m.s.)) + labs(x="Location", y="Walking speed (m/s)") + geom_boxplot(aes(color=Distracted))

The median walking speed of the indoor undistracted box is *slightly* greater than the outdoor undistracted box. This indicates that, undistracted people walk *slightly* faster inside rather than outside. Overall, the relationship between walking speed and location is not significantly different when considering distraction.

Create a rough sketch of the graph that best shows how your explanatory variable influences the response. (Include confounding other explanatory variables if they exist)



What are your expected results?
We expected that

What specific ways will you have to treat each explanatory variable during analysis?
Time of day:
Location: This is a categorical variable with two groups (inside or outside). Box plots should be used to represent this data. A two-sample t-test can be used to test whether walking speeds differ between the two groups.
Distraction: This is a binomial variable so specific statistics and specific plot types will have to be used during analysis. Box plots should be used when plotting this variable and a two sample t-test is likely suitable to determine if there is a difference between Distracted or not-Distracted.

Is there consensus about whether any explanatory variables have an interaction between them? Which and how?
Time since 8am and location interact - speed inside decreased and the speed outside increased over time.

Distraction and location did not interact significantly. The number of people distracted inside vs outside was random. (location of distraction didn't significantly effect speed)

Worksheet 6: T-tests and ANOVA (October 1st, 2025)

Group members present: Hannah G, Ben, Elissa, Ella, Ainsley, Hannah M

Record who is completing which analysis here: T tests: everyone did all the tests in their own programming and helped write summaries together

Step 1 – Final cleaning of data

1. Ensure that you have an R-script that completes all the cleaning on your data that is needed.

```
#Turning time into a decimal number of hours since 8am
walk<-mutate(walk, format_time=hm(Hour.into.the.day))
walk<-mutate(walk, time_since_8am=format_time-hm("8:00"))
walk<-mutate(walk, time=hour(time_since_8am)+minute(time_since_8am)/60)

##making time into a boolean value (true/false)
walk <- mutate(walk, distracted_bool = Distracted=="yes" | Distracted=="Yes")

#making a categorical time column
walkSpeed <- mutate(walkSpeed, time_categorical = if_else(time<1, "Morning",
                                    if_else(time<5, "Afternoon",
                                        "Evening")))
```

2. List here what steps must be taken to clean the data. (What alterations did your group decide needed to be done to the data prior to analysis.)

We converted the time of day into 'time since 8am' as a decimal, and the distracted column to "TRUE" and "FALSE" rather than "yes" and "no." We also converted the time-of-day variable into a categorical variable including "morning," "afternoon," and "evening." This allowed us to use an ANOVA to compare whether the time-of-day impacts walking speed.

Step 2- T-tests. Now that everyone has a script that ensures you are all working on the same data, commence analysis.

1. What variables require a t-test for analysis?

   Distracted vs not distracted
   Inside vs outside

2. Explain in plain language what are you would be testing with the t-test. (I.e., what will the results tell you about your data or hypothesis).

   With the t-test we will be looking at whether there is a statistical difference between walking speed averages of two groups. For the distracted vs. not distracted t-test, it will tell us whether people who are distracted walk significantly faster or slower than those not distracted. Similarly, for the inside vs outside t-test, it will tell us whether people inside walk

significantly faster or slower than those outside. This will tell us if our hypotheses, people walk slower when distracted and people walk slower inside, are supported by the data.

3. What type of t-test is most appropriate for this and why?

The two sample t-test is most suitable since there are two sample groups, for each of our categorical variables. Each group has their own mean and we want to test if the two sample means are different, hence the two sample t-test. Note the samples aren't paired.

4. Complete the t-test. Record what code you used, and what the output of the t-test is.

T-test code for inside vs. outside explanatory variable:

t.test(Walking.speed..m.s.~Location, data=rawdata)

Output:

Welch Two Sample t-test

data: Walking.speed..m.s. by Location

t = 1.7789, df = 186.11, p-value = 0.07689

alternative hypothesis: true difference in means between group Inside and group Outside is not equal to 0

95 percent confidence interval: -0.006235589 0.120653401

sample estimates:

mean in group Inside 1.461945

mean in group Outside 1.404736

T test code for distracted vs not:

t.test(Walking.speed..m.s.~distracted_bool, data=walk)

Output:

Welch Two Sample t-test

data: Walking.speed..m.s. by distracted_bool

t = 6.4903, df = 149.82, p-value = 1.181e-09

alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0

95 percent confidence interval: 0.1397301 0.2620478

sample estimates:

mean in group FALSE (not distracted) 1.504318

mean in group TRUE (distracted) 1.303430 t.test(Walking.speed..m.s.~distracted_bool, data=walk)

Not sure if we need all this below since its done in the ANOVA section? Not sure -Ben

Numerical time variable converted to categorical for T-testing:

T-test code for time blocks:

```
WALKING_SPEED <- WALKING_SPEED %>% # Convert hour_into_the_day to proper time
mutate(hour_numeric = hour(hm(Hour.into.the.day))) %>% # Categorize into 3 blocks
mutate(time_block = case_when( hour_numeric >= 8 & hour_numeric < 12 ~ "8AM",
hour_numeric >= 12 & hour_numeric < 16 ~ "12PM", hour_numeric >= 16 ~ "4PM", TRUE ~
NA_character_
```

Subset data (pick two time blocks):

```
subset_data <- WALKING_SPEED %>% filter(time_block %in% c("12PM", "4PM"))
```

Run t-test using subset of time blocks:

```
t.test(Walking.speed..m.s. ~ time_block, data = subset_data)
```

12pm vs 8am t-test output:

data: Walking.speed..m.s. by time_block

t = -2.7895, df = 95.863, p-value = 0.006369

alternative hypothesis: true difference in means between group 12PM and group 8AM is not equal to 0

95 percent confidence interval:

-0.18505367 -0.03118052

sample estimates:

mean in group 12PM 1.359129

mean in group 8AM 1.467246

12pm-4pm t-test output:

Welch Two Sample t-test

data: Walking.speed..m.s. by time_block

t = -1.4011, df = 104.87, p-value = 0.1641

alternative hypothesis: true difference in means between group 12PM and group 4PM is not equal to 0

95 percent confidence interval:

-0.15486095 0.02662283

sample estimates:

mean in group 12PM 1.359129

mean in group 4PM 1.423248

8am vs 4pm output:

Welch Two Sample t-test

data: Walking.speed..m.s. by time_block

t = -1.1403, df = 101.63, p-value = 0.2568

alternative hypothesis: true difference in means between group 4PM and group 8AM is not equal to 0

95 percent confidence interval:

-0.12053410 - 0.03253803

sample estimates:

mean in group 4PM 1.423248

mean in group 8AM 1.467246

5. Record your interpretation of the results here. What does it tell you about your hypothesis and predictions?

Inside vs. Outside: The results from the T-test shows that there is no significant difference in walking speed between the two locations. The T value was 1.7789 which is < 1.96; therefore, the mean walking speed outside was not statistically different from the mean walking speed inside. The difference between the two means is between -0.006235589 and 0.120653401 with 95% confidence.

Distracted vs. Not Distracted: The results from the T-test show that there is a statistical difference in walking speed when distracted or not. The mean walking speed of distracted people is statistically different from the mean speed of those not distracted. The difference between the two means is between 0.1397301 and 0.2620478 with 95% confidence. The p-value is less than 0.05, which concludes that there is a statistically significant difference between the two groups. This supports our hypothesis that distracted walkers have a slower walking speed than non-distracted walkers.

Timeblocks:

8am vs. 12pm: the results from the t-test showed a significant difference between the mean speed of walkers at 8 am (m= 1.467246 m/s) vs walkers at 12pm (m= 1.359129). t = -2.7895, p-value = 0.006369 CI= -0.12053410 - 0.03253803

8am vs. 4pm: the results from the t-test showed no significant difference between the mean speed of walkers at 8 am (m= 1.467246 m/s) vs walkers at 4pm (m= 1.423248). t = -1.1403, df = 101.63, p-value = 0.2568. CI= -0.18505367 - 0.03118052.

12pm vs 4pm: the results from the t-test showed no significant difference between the mean speed of walkers at 12pm (m= 1.359129) vs walkers at 4pm (m= 1.423248). t = -1.4011, df = 104.87, p-value = 0.1641. CI= -0.12053410 - 0.03253803

.

Step 3- ANOVAs

1. What variables require an ANOVA for analysis?

The categorical version of the time-of-day variable requires an ANOVA as there are three categories, morning, afternoon, and evening.

2. Explain in plain language what are you would be testing with the ANOVA. (I.e., what will the results tell you about your data or hypothesis).

An ANOVA test of the three categories of morning, afternoon, and evening will determine if there is a statistical difference between the walking speeds (the response variable) at each time. We can conclude whether the time of day impacts how fast people walk.

3. Complete the ANOVA. Record what code you used, and what the output of the ANOVA is

```
time_anova <- lm(walking_speed~time_categorical, data=walkSpeed)

summary(time_anova)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.57256 -0.17549 -0.01481  0.17058  0.53027

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.35913    0.03214  42.289  < 2e-16 ***
time_categoricalEvening   0.06412    0.04483   1.430  0.15404
time_categoricalMorning   0.10812    0.03873   2.792  0.00571 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2318 on 219 degrees of freedom
Multiple R-squared:  0.03471,   Adjusted R-squared:  0.0259
F-statistic: 3.938 on 2 and 219 DF,  p-value: 0.02089
```

4. Complete a Tukey Post-hoc test and record the code and output here.

```
pairwise(time_anova)
```

```
── Tukey's Honestly Significant Differences ──────────────────────
Model: walking_speed ~ time_categorical

time_categorical
Levels: 3
Family-wise error-rate: 0.05
```

	group_1	group_2	diff	pooled_se	q	df	lower	upper	p_adj
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
1	Evening	Afternoon	0.064	0.032	2.023	219	-0.042	0.170	.3271
2	Morning	Afternoon	0.108	0.027	3.948	219	0.017	0.200	.0157
3	Morning	Evening	0.044	0.027	1.638	219	-0.046	0.134	.4796

5. Record your interpretation of the results here. What does it tell you about your hypothesis and predictions?

We are 95% certain that the true difference between the groups falls between the upper and lower values. Since zero falls between the upper and lower limits of “Evening vs Afternoon” and “Morning vs Evening”, these categories do not contain statistically different walking speeds. Their p-value is also greater than 0.05, so we can’t state that they are different. The “Morning vs Afternoon” categories; however, do not contain zero between their upper and lower limits, and their p-value is less than 0.05, confirming that these categories are statistically different. This means that walking speed in the morning statistically differs from walking speed in the afternoon.

Worksheet 7: Regressions (October 3, 2025)

Group members present: Hannah G, Ben, Elissa, Ella, Ainsley, Hannah M

Record who is completing which analysis here:

Now that everyone has a script that ensures you are all working on the same data, commence analysis.

1.What variables require a regression for analysis?

Time (numerical, "hours since 8am" - decimal form)

2.Explain in plain language what are you would be testing with the regression. (I.e., what will the results tell you about your data or hypothesis).

The sign of the slope of the line will provide whether walking speed increases, decreases, or stays the same with time. The R-squared value tells us how well the line fits the data and its linearity.

3. Complete the regression. Record what code you used, and what the output of the regression is.

time_regress<-lm(Walking.speed..m.s.~time, data=walk)

summary(time_regress)

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.62242 -0.17294 -0.00502  0.17337  0.56029

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.455732   0.021973  66.252   <2e-16 ***
time        -0.007607   0.004731  -1.608    0.109
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.234 on 220 degrees of freedom
Multiple R-squared:  0.01161,   Adjusted R-squared:  0.007122
F-statistic: 2.585 on 1 and 220 DF,  p-value: 0.1093
```

4. Record your interpretation of the results here. What does it tell you about your hypothesis and predictions?

Intercept tells us the presumed walking speed average at 8am.

Slope tells us walking speed decreases with time, but zero falls within the confidence interval. No significant slope.

The R-squared value of 0.01161 tells us that our data is very far from linear (no significant correlation between time and walking speed. 1.16% of walking speed data is explained by time alone.

The p value of 0.1093 tells us that we cannot reject the null hypothesis

The t-value of –1.608 tells us that time does not have a statistically significant impact on walking speed

Final handing in of project:

1. Good copy of worksheets + dragon kill points sheet (1 person)
    a. **Hannah M**

2. Presentation (2 people)
    a. **Ella**
    b. **Elissa**
3. Methods write up (1 person)
    a. **Ainsley**

4. Results + interpretations write up (1person)
    a. **Hannah G**

5. R-script + datafiles
    a. **Ben**

Upload finished copies of everything to a shared google drive folder so everyone can see.