Lab 8 Peoples Park

```
library(stat20data)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(infer)
```

```
Attaching package: 'infer'

The following object is masked from 'package:stat20data':

    rep_sample_n
```

```
data(ppk)
```

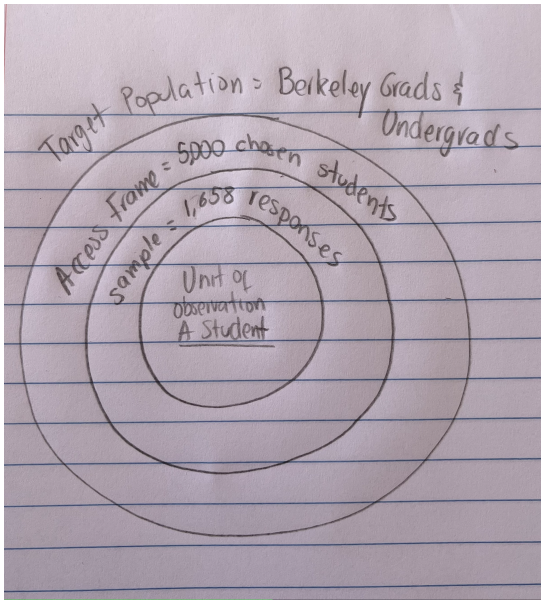
Based on your interpretation of these documents, address the following questions.

1. **What was the goal(s) of the Chancellor's office in commissioning this survey?**

The goal of the chancellor in commissioning this survey could me so many. Some of these include that she did it to gather the opinions of students regarding the peoples park project. She wanted to measure uninformed support for the project and also measure informed support (after being presented with the four main elements). She wanted to make sure she took into account the students perspectives when deciding to do the project, she wanted to see if students supported the project, additionally, she wanted to see how the students support or engagement with Berkeley would change regarding the project. She also wanted to use this as a tool to gain support, by showing that most students support this project it would be much easier for her to carry it through.

2. **Identify the target population, access frame, sample, and unit of observation. Draw a data scope diagram that shows the relationship between the target, frame, and sample.**

In this case, the target population are the Berkeley grads and undergrads (aka, all of the Berkeley student body). The access frame would be the 5,000 students the survey was sent to. The sample would be the 1,658 students that actually responded to the survey and the unit of observation would be a student.



3. **For each of the following types of bias, describe the precautions the Chancellor's office took to limit this kind of bias.**

coverage:

Coverage bias happens the the access frame does not include every unit in the target population. In this case, the chancellor made sure to limit this kind of bias by sending out the survey through a method that reaches all the students. The survey was sent to their berkeley email,

which all of the berkeley access frame (5,000 students) has access to, as well as everybody in the target population. All students already have a berkeley email, and are required to create it once enrolled in the school. This ensures that everyone in the target frame is included in the access frame.

selection:

A selection bias occurs wen the mechanism used selects some units more than others. To avoid this issue, the survey was carried out did it via random sampling. This was done to avoid oversampling students who are not representative of the Berkeley students.

non-response:

Unit non response bias happens when someone selected in the sample is unwilling to participate. Item non response bias happens when someone refuses to answer a particular survey question.

In order to avoid this bias, the survey was designed to collect a specific amount of responses to maintain a good confidence level, in this case they expected and designed the survey to collect 1,250 responses to meet a 25% respond rate and maintain the confidence level. Having this goal enabled them to avoid biased conclusions.

measurement:

Measurement bias happens when the instrument systematically misses the target in one direction. In this survey, this was avoided when the questions where very easy and direct, where everyone was comfortable and knew how to answer.

4. **Which single source of bias potentially creates the most serious problem for the generalizing from the sample to the population? How might this bias impact the findings, e.g., unduly inflate or reduce the measured support for the People's Park Project?**

I think that a source of bias that could create a serious problem when generalizing from the sample into the population would be that there are some groups of people who may be more or less likely to respond to the survey. For example, its more likely tht students who currently live on campus to respond, as well as students who have strong opinions to respond. There is a lot of people in between who have opinions, but not strong enough for them to want to fill out the survey. While strong minded people possibly answered the survey and skew the results or inflate the support for the park or against the park.

5. **Describe two parameters that the Chancellor's office is trying to estimate using the survey data.**

Two parameters the chancellor tries to estimate could be the amount/rate of support from the students for the peoples park project (both before they were informed and after they were informed), this could also give them a rate of change of how much information changes the

support around this debate. Another parameter could be levels of awareness students have regarding this project.

6. **Consider the type of data collected in question 8, which is measured using the Likert Scale. Review the Wikipedia article on the Likert Scale (particularly the Scoring and Analysis section) to determine: Where does this type of data fall in the Data Taxonomy?**

Question 8 asks students about how likely they are to consider Berkeley housing if it was offered to them. They have to rank their answers from "very likely" to "very unlikely".

Answers based on this scale are often called "summative scales"

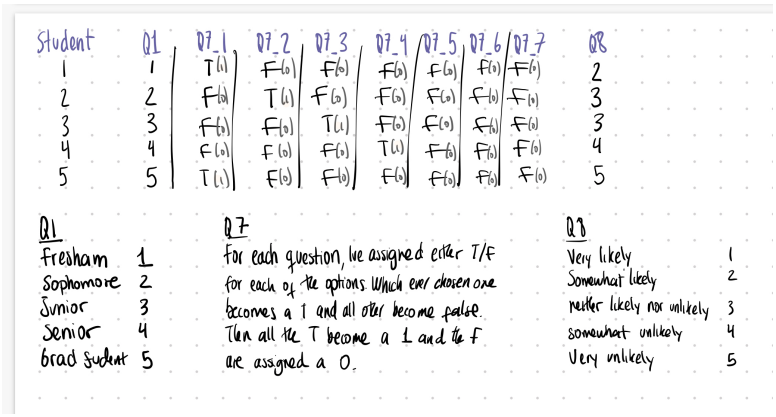
They are considered ordinal categorical because although the chosen answer has no objective numerical basis, a value is assigned to is, and then we can get a distance metric from these values assigned to each option. These are decided by the researcher. Although these are often sentences that we see and choose from, each have an assigned value and follow a specific order, which is why they are ordinal.

7. **Sketch a data frame of what the first 5 rows of the data frame might look like that contains the responses from the first 5 students. Include columns showing what the data might look like that comes out of questions 1, 7, and 8. Note that in the data set, the data values are all translated from words into numbers. Speculate as to how this translation is done.**

I attached what I think the data might look like. I assigned the first column student and took the first 5 rows, aka the first 5 student answers. Next I wrote out Q1, where the answers could be freshman, sophomore, junior, senior, or grad student. These values then are converted into values 1-5, accordingly.

For question 7, I assigned a T/F to each of the challenges students have experienced. These could also be assigned a 0 or 1 value. In this case I assigned a 1 to each T value and a 0 to each F value.

In all of these, even though the answers are in words, we assigned values to translate these to data values (numbers)



Part II: Computing on the Data

The ppk data set represents a subset of questions that were asked in the questionnaire and have had random noise added to them. The results, in aggregate, share similar statistical properties to the raw data, but a given row no longer reflects an individual student's response completely.

8. **Print the first few rows with the columns that correspond to the responses to survey questions 1, 7, and 8. Note: we have changed the data back from all numerical data, as suggested by lab question 7, to a mix of numerical and categorical data. Please comment on whether your encoding of the data from Q7 on the questionnaire matches the encoding in ppk.**

As seen in the code below, I selected the question 1,7 and 8 and used the code "head" in order to have the 5 responses on the rows and the questions on the columns. In this case, I decided to show the first 10 rows because I believe this represent "the first few rows"

My encoding from the data back in question 7 does match the encoding in ppk. In this one however, there is no T or F printed and the values of 1 and 0 are direclty assigned (T=1, F=0)

I however, did not think or consider the answers to include NA.

```
ppk%>%
  select(Q1,Q7_1,Q7_2,Q7_3,Q7_4,Q7_5,Q7_6,Q7_7,Q8)%>%
head(10)
```

```
# A tibble: 10 x 9
   Q1                          Q7_1  Q7_2  Q7_3  Q7_4  Q7_5  Q7_6  Q7_7 Q8
   <chr>                      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
```

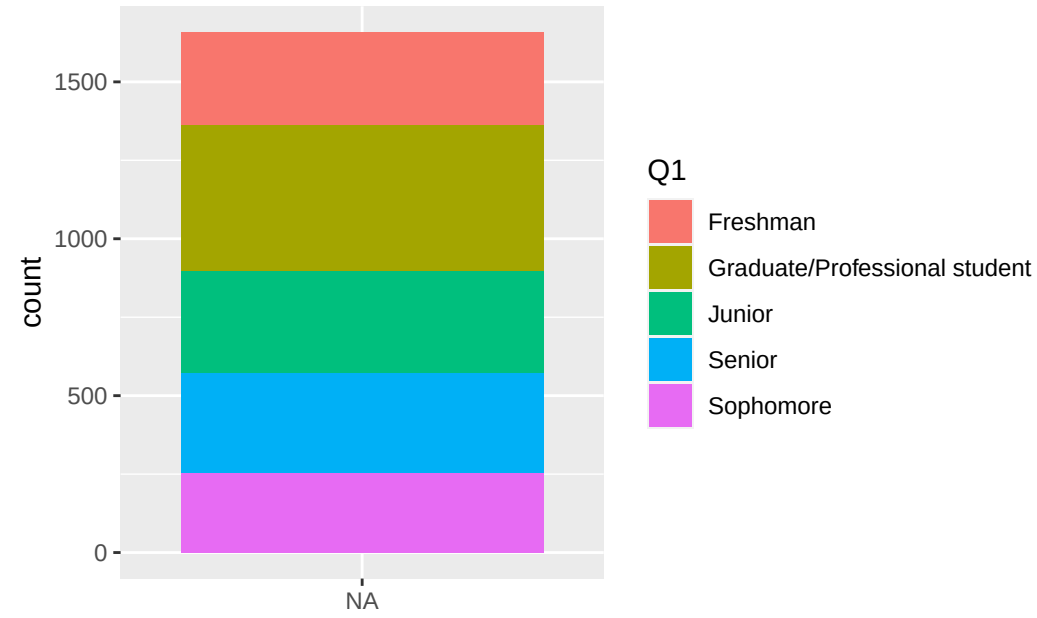
```
 1 Senior                          0     0     0     0     0     0     1 Very~
 2 Junior                          0     1     0     0     0     0     0 Very~
 3 Graduate/Professional student  NA    NA    NA    NA    NA    NA    NA Very~
 4 Junior                          1     0     1     0     1     0     0 Some~
 5 Graduate/Professional student  NA    NA    NA    NA    NA    NA    NA Some~
 6 Graduate/Professional student  NA    NA    NA    NA    NA    NA    NA Some~
 7 Graduate/Professional student  NA    NA    NA    NA    NA    NA    NA Very~
 8 Junior                          1     0     0     0     1     0     0 Very~
 9 Graduate/Professional student  NA    NA    NA    NA    NA    NA    NA Some~
10 Sophomore                      1     1     1     0     1     0     0 Very~
```

9. **Create visualizations for each of the following survey questions. For each, add a title and axis labels to make it clear what they are showing, and describe the distribution in words. If your visualization is of ordinal data, the bars should be ordered accordingly.**

a. Question 10

```
ppk%>%
  ggplot(aes(x=factor(Q10,
                      levels=c("very unimportant",
                               "somewhat unimportant",
                               "neither important nor unimportant",
                               "somewhat important",
                               "very important")),fill=Q1))+

  geom_bar()
```



Q10, levels = c("very unimportant", "somewhat unimportant", ...

b. Question 18 and 21, being sure to show the change of each individual respondent before and after the information.

In this case we can see that the blue Q21 represents the post-information while Q18 (Red) represents the uninformed support.

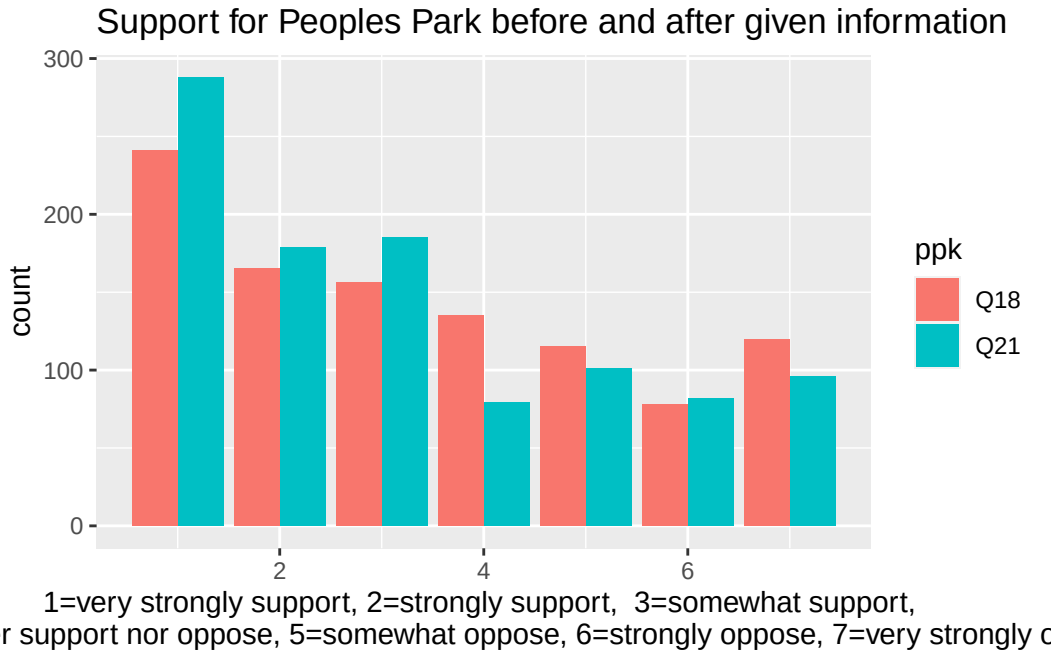
We can see very strongly support increased from around 240 responses to almost 300 responses.

Very little time did the information presented to the respondents had no positive effect, for example for neither support nor oppose, somewhat oppose and very strongly oppose, these still had more people chose these before information than after information.

```
Q18_Q21_PPK<-ppk%>%
  select(Q18,Q21)%>%
  drop_na()%>%
  gather(ppk,event,Q18:Q21)%>%
  group_by(event,ppk)
```

```
Warning: attributes are not identical across measure variables;
they will be dropped
```

```
ggplot(Q18_Q21_PPK,aes(event,fill=ppk))+
  geom_bar(stat="count",
           position="dodge")+
  labs(x="1=very strongly support, 2=strongly support,  3=somewhat support,
       4=neither support nor oppose, 5=somewhat oppose, 6=strongly oppose, 7=very strongly
  ggtitle("Support for Peoples Park before and after given information")
```



10. **Create a new column called `support_before` that takes the response data from question 18 and returns `TRUE` for answers of "Very strongly support", "Strongly support", and "Somewhat support" and FALSE otherwise. What proportion of the survey participants in each class (freshman, sophomore, etc) supported the People's Park Project?**

```
#Create Column
ppk%>%
  mutate(support_before=Q18_words %in% c("very strongly support","strongly support","somew
```

```
#Categorize by class
ppk%>%
group_by(Q1, support_before)%>%
  count(support_before)
```

```
# A tibble: 5 x 3
# Groups:   Q1, support_before [5]
  Q1                            support_before     n
  <chr>                         <lgl>          <int>
1 Freshman                      FALSE            295
2 Graduate/Professional student FALSE            467
3 Junior                        FALSE            324
4 Senior                        FALSE            320
5 Sophomore                     FALSE            252
```

```
#Calculate proportions
fresh_spt<-95/(200+95)
soph_spt<-107/(145+107)
jun_spt<-116/(208+116)
sen_spt<-100/(220+100)
grad_spt<-144/(323+144)
c(fresh_spt,soph_spt,jun_spt,sen_spt,grad_spt)
```

```
[1] 0.3220339 0.4246032 0.3580247 0.3125000 0.3083512
```

In this case, the proportion of freshmen is 0.322

The proportion of sophomore is 0.4246

The proportion of juniors is 0.3580

The proportion of seniors is 0.3125

The proportion of grad students is 0.3083

11. **What is the mean and median rating of the condition of People's Park (question 15 on the survey)?**

As seen in the code, the mean rating of the condition of peoples park in question 15 was 3.0472, and the median of the question was 2.

I also addressed the missing values in the data by adding na.rm=T

```
#Calculate mean
mean(ppk$Q15_1, na.rm=T)
```

```
[1] 3.047269
```

```
#Calculate median
median(ppk$Q15_1, na.rm=T)
```

```
[1] 2
```

12. **Create a new column called `change_in_support` that measures the change in support from question 18 to 21. What is the average change in support of the survey participants in each class (freshman, sophomore, etc) for the People's Park Project after being presented the information on page 14 of the questionnaire?**

Specifically for each class, we have that for freshmen it was = -0.4831

Graduates/professional = -0.1148

For juniors it was=-0.21658

For sophomores it was = -0.293785

We can see how the class that had the most change regarding perspective on the part after information was given to them was the freshman, while the ones that changed their perspectives the least where the graduates

```
#Change of support specific to each class
ppk<-ppk%>%
  mutate(change_in_support=Q21-Q18)
```

```
ppk%>%
group_by(Q1)%>%
  summarize(mean=mean(change_in_support,na.rm=T))
```

```
# A tibble: 5 x 2
  Q1                                mean
  <chr>                            <dbl>
1 Freshman                        -0.483
2 Graduate/Professional student -0.115
3 Junior                          -0.217
4 Senior                          -0.315
5 Sophomore                       -0.294
```

Part III: Making Inferences about Berkeley Students

13. **Create a 95% bootstrap confidence interval for the Berkeley student median rating of the condition of People's Park. Interpret the interval in the context of the problem.**

In this case we are 95% confident that the median in this question of the condition of People's Park is 2. This is clear because our interval is (2,2) meaning, it has to be 2. I also graphed the question and its answers, to have better visualization.

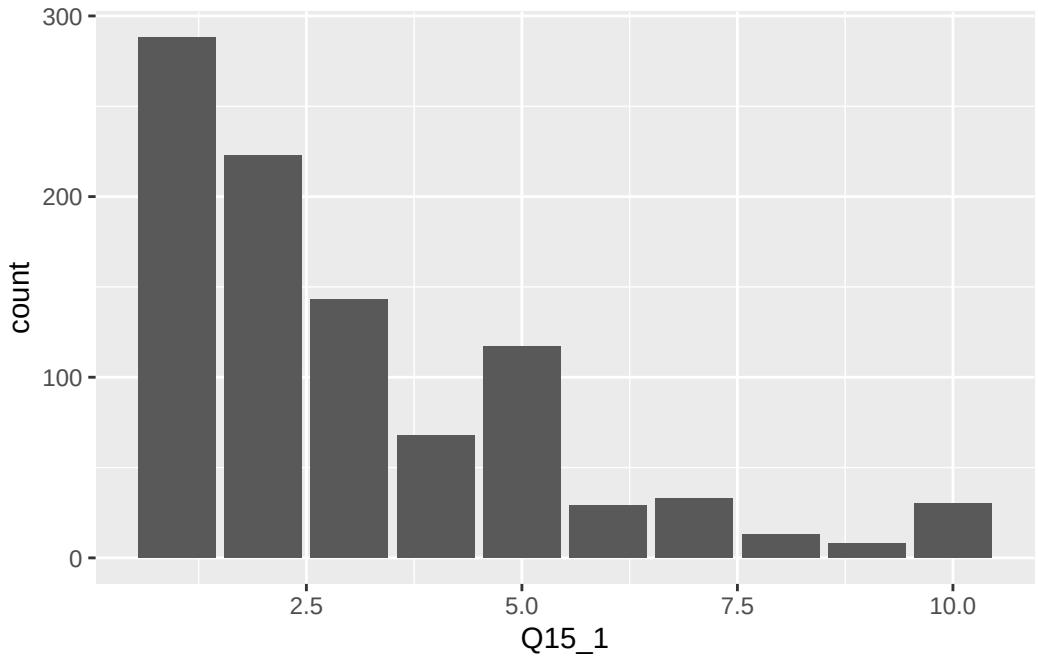
```
ppk%>%
  specify(response=Q15_1)%>%
  generate(reps=500,
           type="bootstrap")%>%
  calculate(stat="median")%>%
  get_ci(level=.95)
```

```
Warning: Removed 706 rows containing missing values.
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1        2        2
```

```
ppk%>%
  ggplot(aes(x=Q15_1))+
  geom_bar()
```

```
Warning: Removed 706 rows containing non-finite values (stat_count).
```



14. **Create a 95% confidence interval based on the normal curve for the proportion of Berkeley students who support the People's Park Project. Interpret the interval in the context of the problem.**

In this case we can be 95% confident that the actual proportion of Berkeley students who support the Peoples Park Project is between (0.4202004,0.5188858) or just 0.42 and 0.52 (approximately).

For this question, my values are:

mean of=0.4695431

sd of =0.4997061

se of=0.02517483

This is based off of before information, because the question does not specific when. (as in, we don't know to use Q18 or Q21, I used Q18)

```
ppk%>%
  drop_na()%>%
  mutate(support = Q18<4)%>%
  summarize(mean=mean(support),
            sd=sd(support),
            SE=sd/sqrt(n()))
```

```
# A tibble: 1 x 3
   mean    sd     SE
  <dbl> <dbl>  <dbl>
1 0.470 0.500 0.0252
```

```
lower_conf<-0.4695431 - 1.96*0.02517483
upper_conf<-0.4695431 + 1.96*0.02517483
c(lower_conf,upper_conf)
```

```
[1] 0.4202004 0.5188858
```

15. **Create a 95% bootstrap confidence interval for the average change in support for the Project among Berkeley students before and after being exposed to the information on page 14 of the questionnaire. Does the interval contain 0? What are the implications of that for those working in the Chancellor's Office on the People's Park Project?**

In this case, we are 95% confident that the average change in support for the project among the berkeley student. body after being exposed to more information on page 14 is between -0.3425 and -0.1945.

In this case, the interval does not include the number 0. The implications for those working in the Chancellors Office is that they are 95% confident that every person that got exposed to new information regarding the project, their support for the peoples park project increased. This helps them in the future also by letting them know that letting students know more information will make them gain or increase their support for the Peoples Park Project.

```
ppk%>%
  specify(response=change_in_support)%>%
  generate(reps=500,
           type="bootstrap")%>%
  calculate(stat="mean")%>%
  get_ci(level=.95)
```

```
Warning: Removed 648 rows containing missing values.
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1   -0.339   -0.200
```