Lab 8

Vivian Yeh

Part I: Understanding the Context of the Data

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

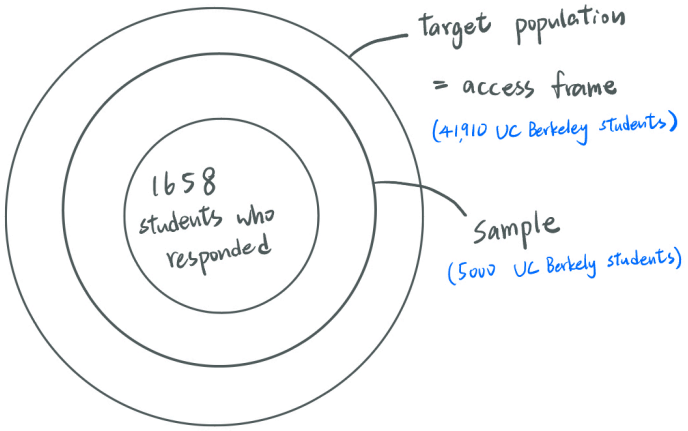
   To learn about the level of students' acknowledge about the People's Park Project and listen to the full range of representative student perspectives regarding housing issues in general, and plans for People's Park in particular.

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.**

   Target population = Access Frame: All UC Berkeley Students (41,910 students)

   Sample frame: Random selected 5000 students

   Unit of observation: A UC Berkeley student who answered the survey



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

a. **coverage:** They engaged a leading independent survey firm with extensive experience helping student voices to be heard on campuses across the country. They use the entire population and the random selection to prevent from the lost of coverage.

b. **selection:** They use scientific, random sample survey of the graduate and undergraduate student populations to avoid the selection bias.

c. **non-response:** They pre-assumed the response rate, and the result is the response rate is higher than the prediction. However, the number is still low compared to the whole population.

d. **measurement:** They keep track of the response rate, 99% confidence level, and 3% margin of error. The survey questions are well designed to avoid the measurement bias.

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?**

   Non-response creates the most serious problem for the generalizing from the sample to the population because compared to the all 41,910 UC Berkeley students, 1658 is only 3.96% of total students. The responded students may be someone who very concerns about this project and wants to express their thoughts, and this will affect the sample result. It is hard to say 3.96% of students can represent all the UC Berkeley students.

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

   The two parameters are students' awareness and attitudes toward the People's Park Project.

   Students' awareness is how the students are aware of the People's Park project and its various components. And, how they are aware of the current housing crisis.

   Students' attitudes is what they know and believe (attitudes) regarding new housing for students and supportive housing for the unhoused on the People's Park site.

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?**

   The Likert Scale falls in Ordinal Categorical Variable.

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.**

   For Question 1, 0 represents "I am not a student", 1 represents "Freshman (1st year)", 2 represents "Sophomore(2nd year)", 3 represents "Junior (3rd year)", 4 represents "Senior(4+ year)", and 5 represents "Graduate/Professional student.

   For Question 7, 0 represents "unselected" and 1 represents "selected"

   For Question8, 0 represents "Very likely", 1 represents "Somewhat likely", 2 represents "Neither likely nor unlikely", 3 represents "Somewhat unlikely", and 4 represents "Very unlikely".

student	Q1	Q7- too expensive	Q7- lack of availability	Q7 Process was difficult/complex	Q7- Unsure where/how to start the process	Q7- Disatisfying Accommodations	Q7- other	Q7- None	Q8
1	2	0	1	0	1	1	0	0	0
2	3	1	1	1	0	1	0	0	1
3	1	1	0	1	1	0	0	0	3
4	1	1	0	0	1	0	0	0	4
5	5	0	1	0	0	1	0	0	3

Part II: Computing on the Data

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.**

```
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(stat20data)
data(ppk)

#select(ppk, Q1, Q7_1, Q7_2, Q7_3, Q7_4, Q7_5, Q7_6, Q7_7, Q8)
ppk[0:5,0:9]
```

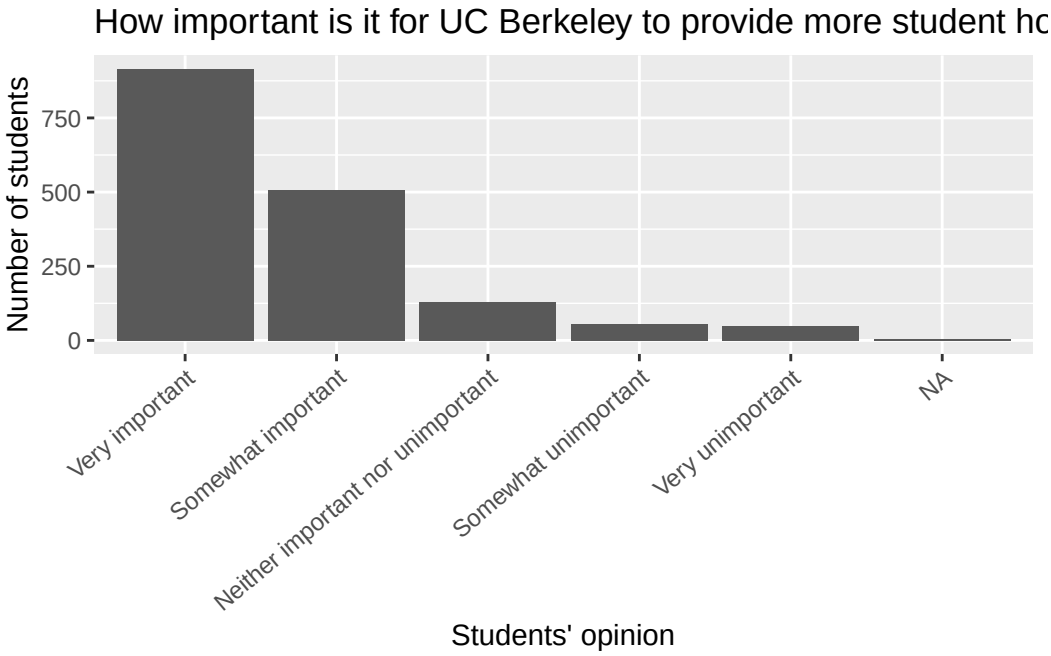
```
# A tibble: 5 x 9
  Q1                            Q7_1  Q7_2  Q7_3  Q7_4  Q7_5  Q7_6  Q7_7 Q8
  <chr>                        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 Senior                           0     0     0     0     0     0     1 Very ~
2 Junior                           0     1     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 Somew~
5 Graduate/Professional student   NA    NA    NA    NA    NA    NA    NA Somew~
```

Since the output of Q7 are composed of 0, 1, and N/A, my encoding of the data from Q7 on the questionnaire matches the encoding in ppk.

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

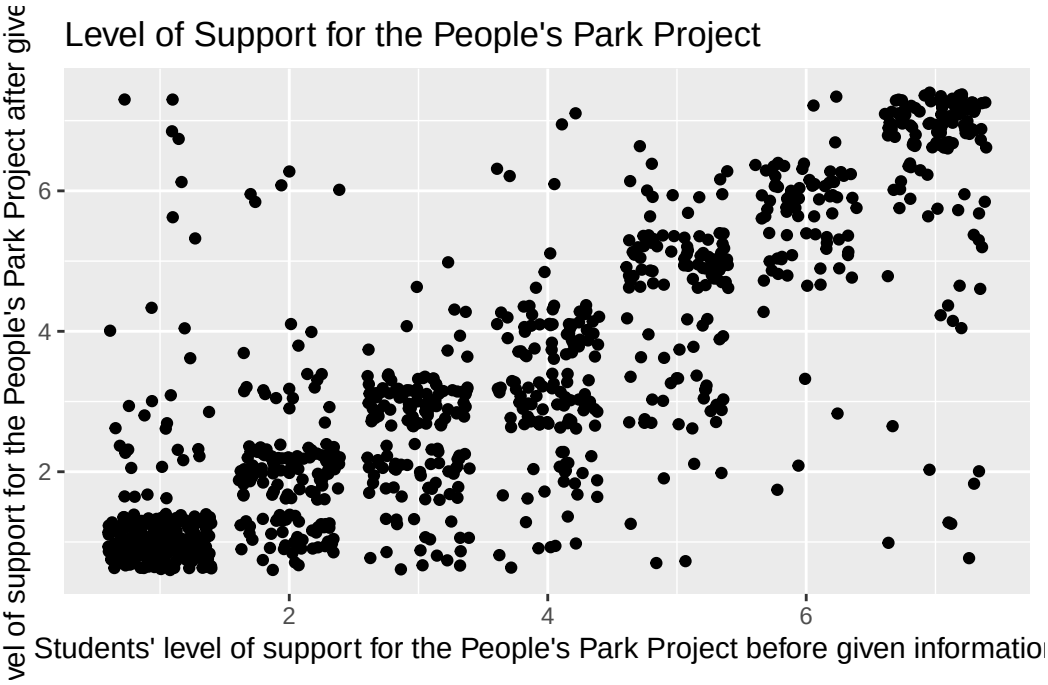
```
library(ggplot2)
ggplot(ppk, aes(x = factor(Q10, level = c('Very important', 'Somewhat important', 'Ne
  geom_bar()+
  theme(axis.text.x = element_text(angle = 40, vjust = 1, hjust = 1))+
  xlab("Students' opinion") +
  ylab("Number of students") +
  labs(title = "How important is it for UC Berkeley to provide more student housing?"
```



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

```
ggplot(ppk, aes(x = Q18, y = Q21)) +
  geom_jitter() +
  xlab("Students' level of support for the People's Park Project before given informa
  ylab("Students' level of support for the People's Park Project after given informat
  labs(title = "Level of Support for the People's Park Project")
```

```
Warning: Removed 648 rows containing missing values (geom_point).
```



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?**

```
ppk <- mutate(ppk, support_before = if_else(Q18 <= 3, TRUE, FALSE))

ppk %>% group_by(Q1) %>%
  summarize(proportion_parti = mean(support_before, na.rm = TRUE))
```

```
# A tibble: 5 x 2
  Q1                              proportion_parti
  <chr>                                      <dbl>
1 Freshman                                   0.534
2 Graduate/Professional student              0.613
3 Junior                                     0.535
4 Senior                                     0.493
5 Sophomore                                  0.605
```

In freshman, there are 53.37% of the survey participants supported the People's Park Project.

In Sophomore, there are 60.45% of the survey participants supported the People's Park Project.

In Junior, there are 53.46% of the survey participants supported the People's Park Project.

In Senior, there are 49.26% of the survey participants supported the People's Park Project.

In Graduate/Professional student, there are 61.28% of the survey participants supported the People's Park Project.

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

```
ppk %>%
  summarize(mean_Q15 = mean(Q15_1, na.rm = TRUE),
            median_Q15 = median(Q15_1, na.rm = TRUE))
```

```
# A tibble: 1 x 2
  mean_Q15 median_Q15
     <dbl>      <dbl>
1     3.05          2
```

The mean of the condition of People's Park is 3.05.

The median of the condition of People's Park is 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?**

```
ppk <- mutate(ppk, change_in_support = Q21 - Q18)
ppk %>% group_by(Q1) %>%
  summarize(average_change_class = mean(change_in_support, na.rm = TRUE))
```

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

```
5 Sophomore                                    -0.294
```

In Freshman, the average change in support of the survey participants is -0.4831461.

In Sophomore, the average change in support of the survey participants is -0.2937853.

In Junior, the average change in support of the survey participants is -0.2165899.

In Senior, the average change in support of the survey participants is -0.3152709.

In Graduate/Professional student, the average change in support of the survey participants is -0.1148936.

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**

```r
library(infer)
```

```
Attaching package: 'infer'

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

    rep_sample_n
```

```r
set.seed(805)
ppk %>%
  drop_na(Q15_1) %>%
  specify(response = Q15_1) %>%
  generate(reps = 500, type = "bootstrap") %>%
  calculate(stat = "median") %>%
  get_ci(.95)
```

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

We can say that we are 95% confident that the population parameter is between 2 and 2.

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**

```
ppk %>%
  summarize(mean_14 = mean(Q18 < 4, na.rm = TRUE),
            sd_14 = sd(Q18 < 4, na.rm = TRUE)/ sqrt(1658),
            lower_end = mean_14 - (1.96 * sd_14),
            upper_end = mean_14 + (1.96 * sd_14))
```

```
# A tibble: 1 x 4
  mean_14  sd_14 lower_end upper_end
    <dbl>  <dbl>     <dbl>     <dbl>
1   0.556 0.0122     0.533     0.580
```

We are 95% confident that the population parameter is between 0.54 to 0.58.

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?**

```
set.seed(531)
ppk %>%
  drop_na(Q10, Q18, Q21) %>%
  mutate(change_in_support_15 = Q21 - Q18) %>%
  specify(response = change_in_support_15) %>%
  generate(reps = 500, type = 'bootstrap') %>%
  calculate(stat = 'mean') %>%
  get_ci(0.95)
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1   -0.339   -0.205
```

We are 95% confident that the population parameter is between -0.34 and -0.21.

The interval does not contain 0. This means that people who have read the information in the survey turn to support the People's Park Project, and this is a encouragement to the Chancellor's Office on the People's Park Project.