Problem Set 1

Isabelle Gibson

1) Public Opinion in the US

We will be using public opinion data from 2019 from the Pew Research Center to conduct some basic analyses in R. The data is called `publicopinionUS.dta`. Remember that this is a Stata file when trying to load it into R.

Name	Description
`state`	State
`q2`	Do you approve or disapprove of the way Donald Trump is handling his job as President?
`q73`	On balance, do you think having an increasing number of people of many different races, ethnic groups and nationalities in the United States makes this country a better place to live, a worse place to live, or doesn't make much difference either way?
`smok1`	Have you smoked at least 100 cigarettes in your entire life
`pvote16a`	In the 2016 presidential election between Donald Trump and Hillary Clinton, did things come up that kept you from voting, or did you happen to vote?
`partyln` As o	f today do you lean more to the Republican Party or more to the Democratic Party?

Question 1.1

Load the data into R and check the dimensions of the data. How many observations are there? What is the name of the variables?

```
library(foreign)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
# set working directory
d <- read.dta("publicopinionUS.dta")
dim(d)
```

```
## [1] 1503    6
```

```
names(d)
```

```
## [1] "state"    "q2"       "q73"      "smok1"    "pvote16a" "partyln"
```

ANSWER: 1503, “state” “q2” “q73” “smok1” “pvote16a” “partyln”

Question 1.2

Generate a binary indicator of support for Donald Trump by using the question “do you approve or disapprove of the way Donald Trump is handling his job as President?” (**q2**). Use this binary indicator to show how many respondents approve how the president is doing his job.

Answer: 614 respondents approve of how the president is doing his job.

```
d <- d |>
  mutate(trump_support = ifelse(q2 == "Approve", 1, 0))
approve <- sum(d$trump_support, na.rm =T)
approve
```

```
## [1] 614
```

Question 1.3

What are the states with the highest and lowest support for Donald Trump?

```
state_support <- d |>
  group_by(state) |>
  summarize(
    total_responses = n(),
    approve_count = sum(q2 == "Approve"),
    approve_proportion = approve_count/total_responses
)
high_support <- state_support |>
  filter(approve_proportion == max(approve_proportion))
low_support <- state_support |>
  filter(approve_proportion == min(approve_proportion))

high_support
```

```
## # A tibble: 1 x 4
##   state total_responses approve_count approve_proportion
##   <fct>           <int>         <int>              <dbl>
## 1 MS                  8             7              0.875
```

```
low_support
```

```
## # A tibble: 2 x 4
##   state total_responses approve_count approve_proportion
##   <fct>           <int>         <int>              <dbl>
## 1 DC                  2             0                  0
## 2 NH                  4             0                  0
```

ANSWER: State with highest support: MS, 87.5%

State with lowest support: DC and NH, 0.0%

Question 1.4

What is the proportion of republicans that approve Donald Trump?

```
rep <- d |>
  filter(partyln == "Republican")
approve_gop <- mean(rep$q2 == "Approve", na.rm = TRUE)
approve_gop
```

```
## [1] 0.7631579
```

ANSWER: The proportion of Republicans that approve of Trump is 76.32%

Question 1.5

What is the proportion of democrats that approve Donald Trump?

```
dems <- d |>
  filter(partyln == "Democrat")
approve_dems <- mean(dems$q2 == "Approve", na.rm = TRUE)
approve_dems
```

```
## [1] 0.06741573
```

ANSWER: Proportion of Democrats that approve of Trump is 6.95%

Question 1.6

What is the state with the highest number of respondents that are both democrats and Trump supporters?

```
dem_support <- d |>
  filter(partyln == "Democrat" & trump_support == 1)
state_counts <- dem_support |>
  group_by(state) |>
  summarize(count = n())

high_dem_support <- state_counts |>
  filter(count == max(count))

high_dem_support
```

```
## # A tibble: 1 x 2
##   state count
##   <fct> <int>
## 1 TX        3
```

ANSWER: State with the highest number of Democrats who support Trump is TX with 3 respondents.

Question 1.7

Generate a new binary indicator called "support for diversity" using the following question from the survey: "on balance, do you think having an increasing number of people of many different races, ethnic groups and nationalities in the United States makes this country a better place to live, a worse place to live, or doesn't make much difference either way?" (**q73**). Answering a better place to live is considered support for diversity. How many people support diversity? What is the mean for support of diversity?

```
d <- d |>
  mutate(support_for_div = ifelse(q73 == "A better place to live", 1, 0))
div_support <- sum(d$support_for_div, na.rm = TRUE)
```

```r
div_support_mean <- mean(d$support_for_div, na.rm = TRUE)
div_support
```

```
## [1] 903
```

ANSWER:

Number of people who support diversity: 903

Mean support for diversity: 60.07

Question 1.8.

What is the proportion of Trump supporters that approve diversity?

```r
trump_support1 <- d |>
  filter(q2 == "Approve")
div_trump_support <- mean(trump_support1$q73 == "A better place to live", na.rm = TRUE)
div_trump_support
```

```
## [1] 0.4446254
```

ANSWER: Proportion of Trump supporters who approve diversity: 0.4446254 or 44.5%

Question 1.9

What are the states with the highest and lowest support for diversity?

```r
state_div_support <- d |>
  group_by(state) |>
  summarize(
    total_responses = n(),
    support_count = sum(q73 == "A better place to live"),
    support_proportion = support_count/total_responses
  )
highest_div_support <- state_div_support |>
  filter(support_proportion == max(support_proportion))
lowest_div_support <- state_div_support |>
  filter(support_proportion == min(support_proportion))

highest_div_support
```

```
## # A tibble: 1 x 4
##   state total_responses support_count support_proportion
##   <fct>           <int>         <int>              <dbl>
## 1 DC                  2             2                  1
```

```r
lowest_div_support
```

```
## # A tibble: 1 x 4
##   state total_responses support_count support_proportion
##   <fct>           <int>         <int>              <dbl>
## 1 MS                  8             2               0.25
```

ANSWER: The state with the highest support for diversity is D.C. at 100%.

The state with the lowest support for diversity is Mississippi at 25%.

Question 1.10

Generate two binary indicators. One for states located in the midwest (North Dakota, South Dakota, Nebraska, Minnesota, Iowa, Missouri, Wisconsin, Illinois, Kansas, Michigan, Indiana, and Ohio). Other for states located in the south (Alabama, Kentucky, Mississippi,Tennessee,Arkansas, Louisiana, Oklahoma, and Texas). What is the proportion of respondents from the midwest and from the south?

```
d <- d |>
  mutate(
    midwest = ifelse(state %in% c("ND", "SD", "NE", "MN",
                                  "IA", "MO", "WI", "IL",
                                  "KS", "MI", "IN", "OH"), 1, 0),
    south = ifelse(state %in% c("AL", "KY", "MS", "TN",
                                "AR", "LA", "OK", "TX"), 1, 0)
  )
midwest <- mean(d$midwest == 1, na.rm = TRUE)


south <- mean(d$south == 1, na.rm = TRUE)

midwest
```

```
## [1] 0.2115768
```

```
south
```

```
## [1] 0.1736527
```

ANSWER: The proportion of respondents in the Midwest is 0.21158 or 21.2%.

The proportion of respondents in the South is 0.17365 or 17.4%.

Question 1.11

What is the proportion of respondents that support Trump in the midwest and in the south?

```
trump_support2 <- d |>
  filter(q2 == "Approve")
midwest_trump_support <- mean(trump_support2$midwest == 1, na.rm = TRUE)
south_trump_support <- mean(trump_support2$south == 1, na.rm = TRUE)
midwest_trump_support
```

```
## [1] 0.2166124
```

ANSWER: The proportion of respondents that support Trump in the Midwest is 0.2166 or 21.7%

The proportion of respondents that support Trump in the South is 0.2019 or 20.2%

Question 1.12

Generate a binary indicator for people that do NOT live in the south and that are democracts? What is the proportion of Trump supporters among that group?

```
d <- d |>
  mutate(non_south_democrat = ifelse(!(state %in% south) & partyln == "Democrat", 1, 0))
trump_support3 <- d |>
  filter(q2 == "Approve")
nonsouth_dem_trump <- mean(trump_support3$non_south_democrat == 1, na.rm = TRUE)

nonsouth_dem_trump
```

```
## [1] 0.02931596
```

ANSWER: The proportion of Trump supporters among people that don't live in the South and that are Democrats is 0.02932 or 2.93%

Question 1.13

Generate two new datasets: one with respondents from the midwest and another with respondents from the south.

```
midwest <- d |>
  filter(state %in% c("ND", "SD", "NE", "MN",
                      "IA", "MO", "WI", "IL",
                      "KS", "MI", "IN", "OH"))
midwest <- midwest |>
  dplyr::select(-midwest, -south)
write.csv(midwest, "publicopinionUS_midwest.csv")
write.dta(midwest, "publicopinionUS_midwest.dta")
south <- d |>
  filter(state %in% c("AL", "KY", "MS", "TN",
                      "AR", "LA", "OK", "TX"))
south <- south |>
  dplyr::select(-midwest, -south)
write.csv(south, "publicopinionUS_south.csv")
write.dta(south, "publicopinionUS_south.dta")

# Filter for respondents from the midwest
midwest <- d |>
  filter(state %in% c("ND", "SD", "NE", "MN",
                      "IA", "MO", "WI", "IL",
                      "KS", "MI", "IN", "OH"))
```

Question 1.14

Generate a binary indicator of having smoked more than 100 cigarettes in your life (**smok1**). Respondents that answered "Yes" are the ones that smokes. Use the binary indicators for support for Trump, support for diversity, and smoking to answer the following questions: What is the average approval of Trump among people that smoke and support diversity? What is the average approval of the Trump among people that smoke and do not support diversity? What is the average approval of Trump among people that do not smoke and support diversity? What is the average approval of Trump among people that do not smoke and do not support diversity?

```
d <- d |>
  mutate(smoke = ifelse(smok1 == "Yes", 1, 0))
support_smoke_div <- d |>
  group_by(smoke, support_for_div) |>
  summarize(avg_trump_support = mean(trump_support, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'smoke'. You can override using the
## `.groups` argument.
```

```
support_smoke_div
```

```
## # A tibble: 4 x 3
## # Groups:   smoke [2]
##   smoke support_for_div avg_trump_support
##   <dbl>           <dbl>             <dbl>
```

```
## 1     0               0             0.548
## 2     0               1             0.276
## 3     1               0             0.598
## 4     1               1             0.341
```

ANSWER:

1. The average approval of Trump among people that smoke and support diversity is 0.34066 or 34.1%
2. The average approval of Trump among people that smoke and do not support diversity is 0.59756 or 59.8%
3. The average approval of Trump among people that do not smoke and support diversity is 0.27644 or 27.6%
4. The average approval of Trump among people that do not smoke and do not support diversity is 0.54802 or 54.8%

Question 1.15

Generate a binary indicator to identify the people that voted in the last election (**pvote16a**). People that answered "voted" are the ones that participated in that election. Check if is true or false that people that voted in the last election are more likely to smoke than people that did not vote.

```
d <- d |>
  mutate(voted = ifelse(pvote16a == "Voted", 1, 0))
smoke_likelihood <- d |>
  group_by(voted) |>
  summarize(smoke_rates = mean(smoke, na.rm = T))
smoke_likelihood
```

```
## # A tibble: 2 x 2
##   voted smoke_rates
##   <dbl>       <dbl>
## 1     0       0.394
## 2     1       0.410
```

ANSWER: It's true, people who voted in the last election are on average more likely to smoke (41.02%) than those who did not vote (39.40%)

Question 1.16

Is it true or false that people from Indiana were more likely to vote than people from Texas?

```
indiana_state <- d |>
  group_by(state) |>
  summarize(vote_likelihood = mean(voted, na.rm = TRUE)) |>
  filter(state %in% c("IN", "TX"))

indiana_state
```

```
## # A tibble: 2 x 2
##   state vote_likelihood
##   <fct>           <dbl>
## 1 IN              0.92
## 2 TX              0.624
```

ANSWER: It is true that on average, people in Indiana are more likely to vote (92%) than people from Texas (62.4%)

Question 1.17

Run a regression using support for Trump as the dependent variable and support for diversity as the independent variable (binary indicators). Interpret the results (constant, coefficients, and p-values). Compute and interpret the confidence intervals.

```
# logit regression -
logit_support_trump_div <- glm(trump_support ~ support_for_div, data = d, family = binomial)

summary(logit_support_trump_div)
```

```
##
## Call:
## glm(formula = trump_support ~ support_for_div, family = binomial,
##     data = d)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.27505    0.08242   3.337 0.000847 ***
## support_for_div -1.11130    0.10974 -10.126  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2033.0  on 1502  degrees of freedom
## Residual deviance: 1927.3  on 1501  degrees of freedom
## AIC: 1931.3
##
## Number of Fisher Scoring iterations: 4
```

```
# computing confidence intervals
confint(logit_support_trump_div)
```

```
## Waiting for profiling to be done...
```

```
##                      2.5 %     97.5 %
## (Intercept)      0.1140152  0.4372851
## support_for_div -1.3273710 -0.8970663
```

```
#double check CI computation
beta_1 <- -1.1113
se_beta_1 <- 0.10974

# Number of observations
n <- 1503

# Compute the critical t-value for a 95% confidence level
alpha <- 0.05
t_critical <- qt(1 - alpha / 2, df = n - 2)

# Calculate the confidence interval
lower_ci <- beta_1 - t_critical * se_beta_1
upper_ci <- beta_1 + t_critical * se_beta_1

# Display the results
cat("95% Confidence Interval for beta1:", "[", round(lower_ci, 4), ",", round(upper_ci, 4), "]\n")
```

```
## 95% Confidence Interval for beta1: [ -1.3266 , -0.896 ]
```

ANSWER: The results show support for diversity, it is significantly associated with a lower level of support for Trump, with a negative coefficient of -1.11130 ($p < 2e\text{-}16$).

This shows us that individuals who support diversity are, on average holding all else constant, less likely to support Trump vs. those who do not support Trump (intercept: 0.275, $p = 0.000847$). The p-values suggest statistical significance ($p < 0.001$) provide support the negative relationship between support for diversity and Trump support.

The confidence interval for those who support diversity ranges from (-1.3266, -0.896), which does not contain 0, we can conclude that there is a statistically significant negative association between support for diversity and not supporting Trump.

Question 1.18

Run a regression using support for Trump as the dependent variable and smoking as the independent variable (binary indicators). Interpret the results (constant, coefficients, and p-values). Compute and interpret the confidence intervals.

```
# Run the logistic regression
logit_smoking_trump <- glm(trump_support ~ smoke, data = d, family = binomial)

# Regression summary
summary(logit_smoking_trump)
```

```
##
## Call:
## glm(formula = trump_support ~ smoke, family = binomial, data = d)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4722     0.0688  -6.863 6.74e-12 ***
## smoke         0.2483     0.1066   2.328   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2033.0  on 1502  degrees of freedom
## Residual deviance: 2027.6  on 1501  degrees of freedom
## AIC: 2031.6
##
## Number of Fisher Scoring iterations: 4
```

```
# Compute 95% confidence intervals for smoking
confint(logit_smoking_trump)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) -0.60779230 -0.3379913
## smoke        0.03924279  0.4574012
```

```
# Double check computation of CIs
beta_smoke <- coef(logit_smoking_trump)["smoke"]
se_smoke <- summary(logit_smoking_trump)$coefficients["smoke", "Std. Error"]
```

```r
# Compute the critical t-value for a 95% confidence level
n <- nrow(d)
t_critical <- qt(1 - 0.05 / 2, df = n - 2)

# Calculate the confidence interval
lower_ci <- beta_smoke - t_critical * se_smoke
upper_ci <- beta_smoke + t_critical * se_smoke

# Display the results
cat("95% Confidence Interval for smoke coefficient:", "[", round(lower_ci, 4), ",", round(upper_ci, 4),
```

```
## 95% Confidence Interval for smoke coefficient: [ 0.0391 , 0.4575 ]
```

ANSWER: The results show support for a positive association between supporting trump and smoking (0.2483, $p < 0.0199$). The intercept (-.4722, $p < 0.001$) shows a negative association between support for Trump and non-smokers (smoke = 0). Both the intercept and smoke coefficient are statistically significant by at least the 0.05% level. This shows us that individuals who do smoke are, on average holding all else constant, more likely to support Trump vs. those who do not smoke.

The confidence interval for smoking ranges from (0.0391, 0.4575), which does not contain 0, we can conclude that there is a statistically significant positive association between those who smoke and support for Trump.

Question 1.19

Run a regression using support for Trump as the dependent variable and partisanship as the independent variable (nominal variable). Interpret the results (constant, coefficients, and p-values). Compute and interpret the confidence intervals.

```r
# Ensure partisanship is a factor with specified reference level
d$partyln <- factor(d$partyln, levels = c("Republican", "Independent", "Democrat"))

# Run the logistic regression
logit_party <- glm(trump_support ~ partyln, data = d, family = binomial)

# Display the regression summary
summary(logit_party)
```

```
##
## Call:
## glm(formula = trump_support ~ partyln, family = binomial, data = d)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.1701     0.1558   7.511 5.85e-14 ***
## partylnDemocrat  -3.7972     0.2895 -13.114  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.11  on 494  degrees of freedom
## Residual deviance: 381.47  on 493  degrees of freedom
##   (1008 observations deleted due to missingness)
## AIC: 385.47
##
## Number of Fisher Scoring iterations: 5
```

```r
# Compute confidence intervals
confint(logit_party)
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %    97.5 %
## (Intercept)      0.8723151  1.484237
## partylnDemocrat -4.3929209 -3.253389
```

ANSWER: The results show support for a negative association between being a Democrat (-4.3929, $p < 0.001$) and supporting Trump. The intercept (1.1701, $p < 0.001$) represents being a Republican, showing a positive association with being a Republican and supporting Trump. Both the intercept and partylnDemocrat coefficient are statistically significant by at least the 0.05% level. This shows us that individuals who are Democrats, on average holding all else constant, are more likely to not support Trump vs. those who are Republicans.

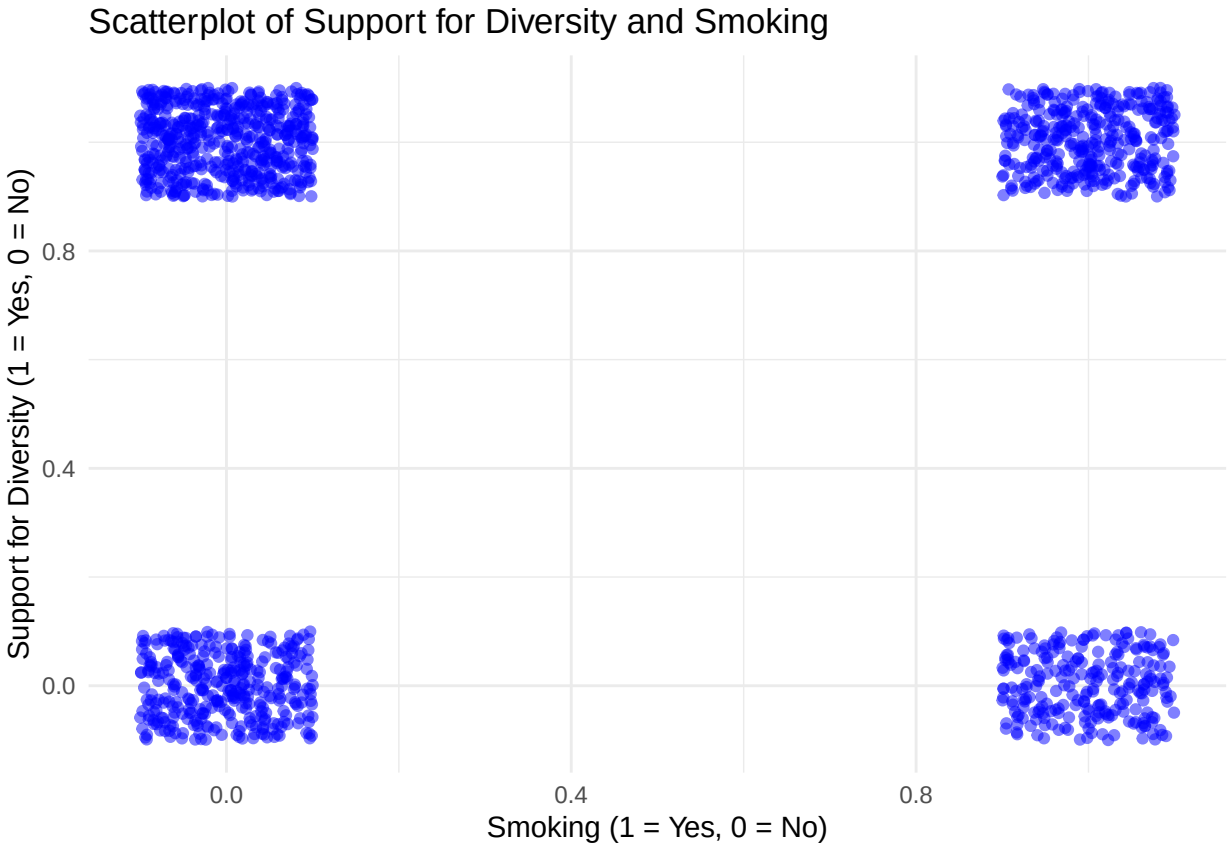
The confidence interval for being a Democrat ranges from (-4.3929209, -3.253389,) which does not contain 0, we can conclude that there is a statistically significant negative association between those who are Democrats and do support for Trump.

Question 1.20

Generate a scatterplot for diversity and smoking. Interpret the figure.

```r
library(ggplot2)

# Generate scatterplot
ggplot(d, aes(x = smoke, y = support_for_div)) +
  geom_jitter(width = 0.1, height = 0.1, alpha = 0.5, color = "blue") +
  labs(
    title = "Scatterplot of Support for Diversity and Smoking",
    x = "Smoking (1 = Yes, 0 = No)",
    y = "Support for Diversity (1 = Yes, 0 = No)"
  ) +
  theme_minimal()
```



ANSWER: The scatterplot shows that there are both smokers and non-smokers across both supporting and not supporting diversity. This suggests to us that regardless of their stance of supporting diversity, smoking is present across all respondent groups. This leads us to conclude that there is not a strong association between supporting diversity and smoking, and further investigation is needed to get more information on the relationship.