Problem Set 1

Week 2, GV481 Quantitative Analysis for Political Science

Before we start, let us load tidyverse and set the working directory

```
library("tidyverse")
setwd("/Users/kevinli/Documents/GitHub/notes/GV481/problems")
```

Let us also load the dataset for the questions

```
df <- read_csv("haber.csv")
```

As instructed, save only the data needed and rename column names

```
df <- df %>%
    select(cnamehabmen, year, Fiscal_Reliance, polity_2,
     ↪  Total_Oil_Income_PC, Regime)

colnames(df) <- c("country", "year", "fiscalreliance", "democracy",
 ↪  "oilincome", "autocracy")
```

Question 1

The first step of data analysis is to describe the data

1a) Let's describe the scope of the dataset. What is the range of the variable year? What is the number of countries included in the dataset?

Let us find the range of variable year:

```
summary(df$year)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1800    1902    1948    1937    1981    2008
```

```
2008 - 1800
```

```
[1] 208
```

The range of the variable year is 208

How many countries? Let us use function unique()

```
unique_countries <- unique(df$country)
# length function to find how long unique vector is

length(unique_countries)
```

```
[1] 169
```

Thus, 169 unique countries included.

b) Let's now turn to describing the dependent and independent variables. In this problem, you will use two different measures of oil reliance and two different measures of democracy. Generate a dummy (binary) variable called democraticcountry which is equal to 1 if the country is democratic and 0 otherwise. To do this, use the vari- able autocracy which is equal to 1 if the country is autocratic and 0 otherwise.

There are two ways to do this - a simple math way, and using the recode() function. I am going to be lazy.

```
df$democraticcountry <- 1 - df$autocracy
```

c) Put together a summary statistics table for the following variables: democracy, democraticcountry, fiscalreliance, oilincome. Your table should include the mean, the standard deviation, the number of observations, the minimum and the maximum value for each of these variables.

We could use a package psych and function describe()

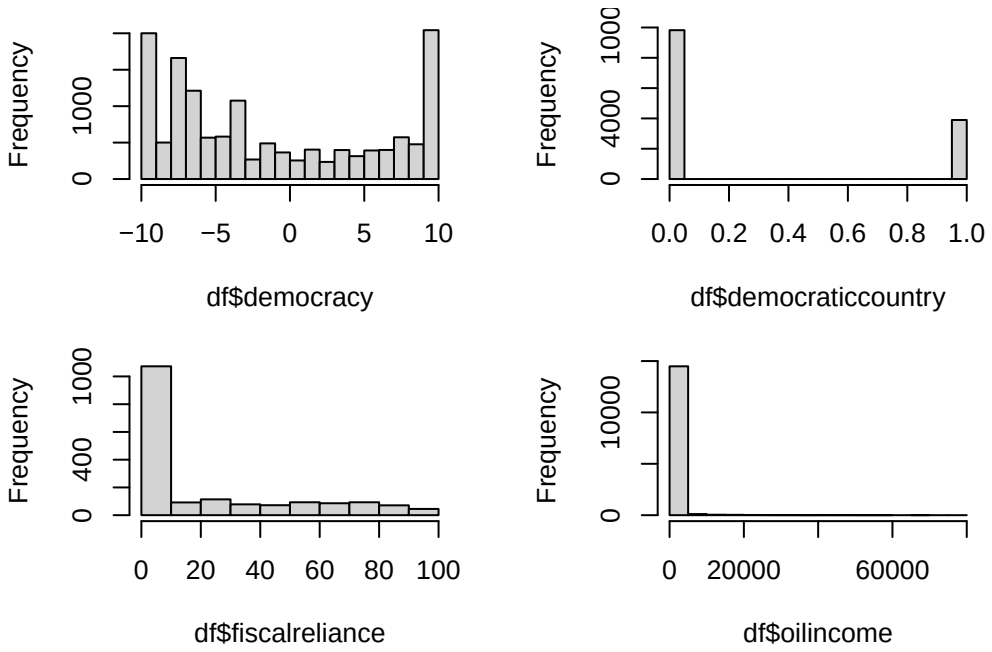
```
library(psych)
```

```
summary_variables <- df %>%
  select(democracy, democraticcountry, fiscalreliance, oilincome)

describe(summary_variables)
```

```
                  vars     n   mean      sd median trimmed  mad min
                  max
democracy            1 14213  -0.84    6.99  -3.00   -1.09 7.41 -10
10.00
democraticcountry    2 13720   0.28    0.45   0.00    0.23 0.00   0
1.00
fiscalreliance       3  1817  21.47   29.62   0.39   16.32 0.58   0
97.82
oilincome            4 14729 343.04 2645.72   0.00    6.05 0.00   0
78588.80
                     range  skew kurtosis    se
democracy            20.00  0.37    -1.36  0.06
democraticcountry     1.00  0.96    -1.08  0.00
fiscalreliance       97.82  1.11    -0.23  0.69
oilincome         78588.80 14.91   287.49 21.80
```

d) Provide an histogram or a bar graph for the variables democracy, democraticcountry, fiscalreliance, oilincome. What do you notice?

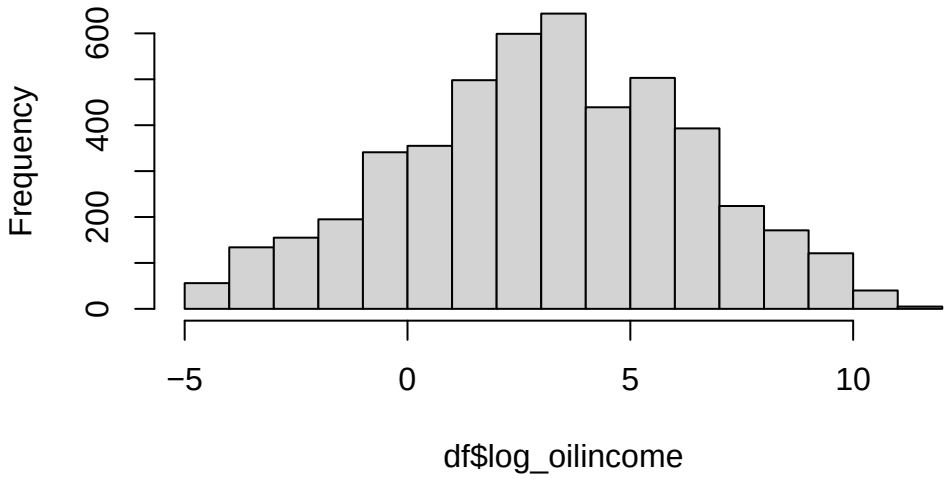
```
par.orig <- par(mfrow = c(2,2), mar = c(5, 6, 0.5, 0.5)) # arranging plots
hist(df$democracy, main = "")
hist(df$democraticcountry, main = "")
hist(df$fiscalreliance, main = "")
hist(df$oilincome, main = "")
```



I notice that democraticcountry is binary, fiscalreliance and oilincome are very right skewed.

e) For highly skewed variables, it is standard to transform the variable using the log transformation. Generate a new variable equal to the log of oilincome using the function log. What do you notice?

```
df$log_oilincome <- log(df$oilincome)

# reset par parameter from before
par.orig <- par(mfrow = c(1,1))
hist(df$log_oilincome, main = "")
```



f) Provide a graph showing the proportion of countries who are democratic over time, and another to show the average oil income per capita over time. What do you notice?

I will use ggplot for this

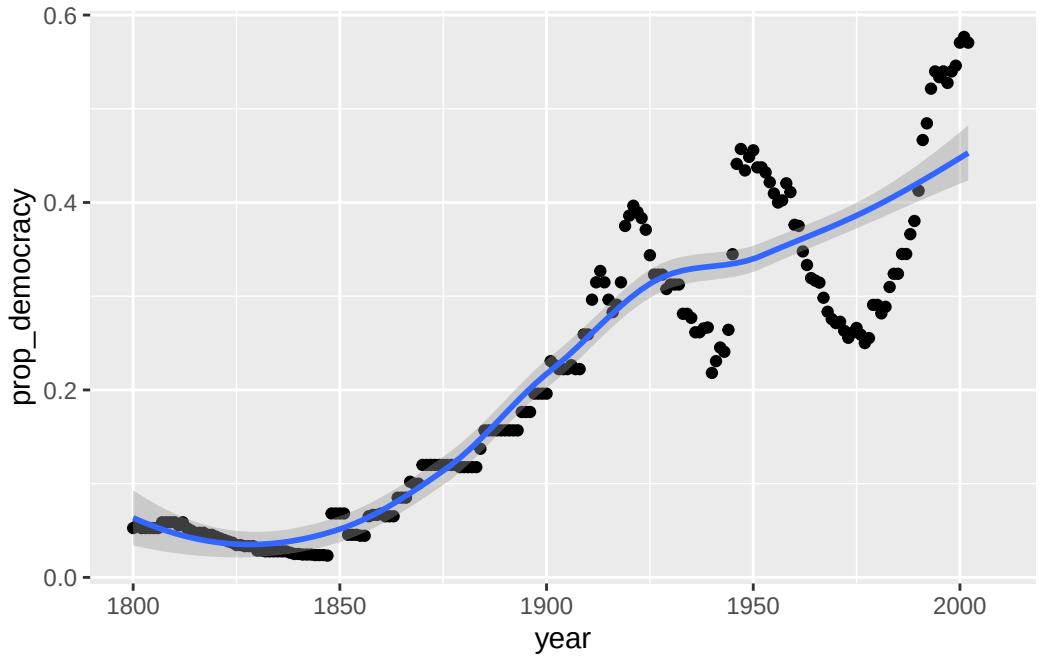
```
# graph of proportion of democracy over time
# Let us group_by() year to find mean
dem_mean <- df %>%
  group_by(year) %>%
  summarise(prop_democracy = mean(democraticcountry, na.rm = TRUE), .groups
   ↪  = 'drop')

# now ggplot

ggplot(dem_mean, aes(x = year, y = prop_democracy)) +
```

```
  geom_point() +
  geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning: Removed 6 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 6 rows containing missing values or values outside the
scale range
(`geom_point()`).
```



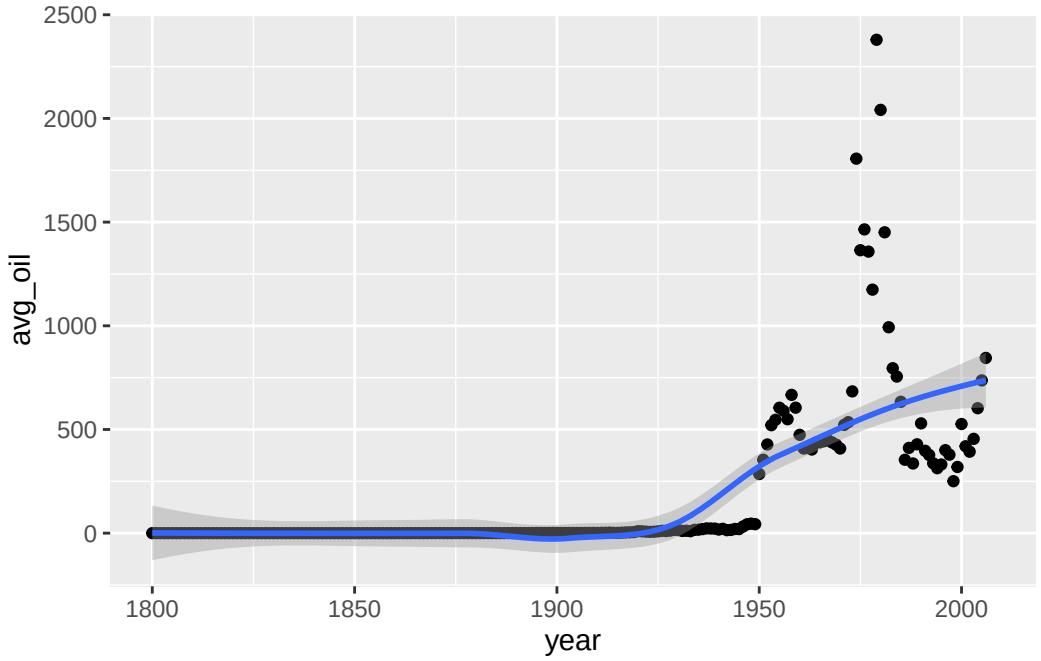

```
# graph of average oil income over time
# group_by() year to find mean

oil_mean <- df %>%
  group_by(year) %>%
  summarise(avg_oil = mean(oilincome, na.rm = TRUE), .groups = 'drop')

# now ggplot
```

```r
ggplot(oil_mean, aes(x = year, y = avg_oil)) +
  geom_point() +
  geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 2 rows containing missing values or values outside the
scale range
(`geom_point()`).
```



Question 2

Estimate the covariance, the correlation coefficient, and the slope of the line of best fit between democracy and oilincome. What do you conclude about the relationship between democracy and oilincome? Describe the one unit change in the dependent variable for a one unit change in the independent variable and describe the standard deviation change in the dependent variable for a one standard deviation change in the independent variable.

Let us do covariance first.

```
cov(df$democracy, df$oilincome, use = "complete.obs")
```

```
[1] -1662.304
```

Now, let us find correlation coefficient

```
cor(df$democracy, df$oilincome, use = "complete.obs")
```

```
[1] -0.09481115
```

Now, let us find linear line of best fit:

```
model1 <- lm(democracy ~ oilincome, data = df)
summary(model1)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-9.251 -6.250 -2.251  6.765 14.957

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.490e-01  5.920e-02  -12.65   <2e-16 ***
oilincome   -2.652e-04  2.348e-05  -11.30   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.972 on 14069 degrees of freedom
  (3135 observations deleted due to missingness)
Multiple R-squared:  0.008989,  Adjusted R-squared:  0.008919
F-statistic: 127.6 on 1 and 14069 DF,  p-value: < 2.2e-16
```

The relationship is negative. For every one unit increase in oil income, there is a predicted 0.00002652 decrease in democracy

In terms of standard deviations:

```
(sd(df$oilincome, na.rm = TRUE) * -2.652e-04) / sd(df$democracy, na.rm =
 ↪  TRUE)
```

```
[1] -0.1003838
```

Thus, for every one standard deviation increase in oil income, there is a predicted 0.1 standard deviation decrease in democracy

Question 3

Next, you will explore correlation in oil and democracy over time.

a) Regress democracy on oilincome using data for year 1800 only. What is happening here? Which estimates do you recover here?

```
# let us first filter for 1800
df_1800 <- df %>%
  filter(year == 1800)

model2 <- lm(democracy ~ oilincome, data = df_1800)
summary(model1)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-9.251 -6.250 -2.251  6.765 14.957

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.490e-01  5.920e-02  -12.65   <2e-16 ***
oilincome   -2.652e-04  2.348e-05  -11.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.972 on 14069 degrees of freedom
  (3135 observations deleted due to missingness)
Multiple R-squared:  0.008989,  Adjusted R-squared:  0.008919
F-statistic: 127.6 on 1 and 14069 DF,  p-value: < 2.2e-16
```

Now, regress democracy on oilincome separately for the years 1900, 1950, 2000, 2006. What do you notice here? Interpret the coefficients

1900

```r
df_1900 <- df %>%
  filter(year == 1900)

model3 <- lm(democracy ~ oilincome, data = df_1900)
summary(model3)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df_1900)

Residuals:
    Min      1Q  Median      3Q     Max
-11.391  -4.037  -2.037   4.463  10.963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.9629     0.9044  -1.065    0.292
oilincome     0.1388     0.2038   0.681    0.499

Residual standard error: 6.24 on 49 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.009372,  Adjusted R-squared:  -0.01084
F-statistic: 0.4636 on 1 and 49 DF,  p-value: 0.4992
```

1950

```r
df_1950 <- df %>%
  filter(year == 1950)

model4 <- lm(democracy ~ oilincome, data = df_1950)
summary(model4)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df_1950)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.0822  -7.0643  -0.6678   8.1692  10.6699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.082216   0.870428   0.094    0.925
oilincome   -0.003903   0.004014  -0.972    0.334

Residual standard error: 7.426 on 74 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.01261,   Adjusted R-squared:  -0.0007281
F-statistic: 0.9454 on 1 and 74 DF,  p-value: 0.3341
```

2000

```r
df_2000 <- df %>%
  filter(year == 2000)

model5 <- lm(democracy ~ oilincome, data = df_2000)
summary(model5)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df_2000)

Residuals:
    Min      1Q  Median      3Q     Max
-12.517  -5.517   2.483   5.485  15.935

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5171429  0.5199774   6.764 2.55e-10 ***
oilincome   -0.0010391  0.0002608  -3.984 0.000104 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.291 on 156 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.09233,   Adjusted R-squared:  0.08651
F-statistic: 15.87 on 1 and 156 DF,  p-value: 0.0001039
```

2006

```
df_2006 <- df %>%
  filter(year == 2006)

model6 <- lm(democracy ~ oilincome, data = df_2006)
summary(model6)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df_2006)

Residuals:
    Min      1Q  Median      3Q     Max
-13.254  -5.451   2.538   4.757  17.558

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.2536244  0.5171698   8.225 7.40e-14 ***
oilincome   -0.0007443  0.0001766  -4.215 4.23e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.193 on 155 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1028,    Adjusted R-squared:  0.09705
F-statistic: 17.77 on 1 and 155 DF,  p-value: 4.226e-05
```

The coefficients go more and more negative, and more significant

I don't feel like interpreting every coefficient, fight me.

Question 4

Let's now turn to examine the correlation between democracy and oil within country.

a) Start by regressing democracy on oilincome for Burundi. What is happening? Which estimates do you recover?

burundi

```
df_burundi <- df %>%
  filter(country == "Burundi")

model7 <- lm(democracy ~ oilincome, data = df_burundi)
summary(model7)
```

```
Call:
lm(formula = democracy ~ oilincome, data = df_burundi)

Residuals:
   Min     1Q Median     3Q    Max
-2.844 -2.844 -2.844  3.156 10.156

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.1556     0.5996   -6.93 1.45e-08 ***
oilincome         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.022 on 44 degrees of freedom
  (1 observation deleted due to missingness)
```

No esimtates

b) We want to restrict our data to countries for which there is variation in both democracy and oilincome. Here is a strategy to do this. Run the following code and provide a scatter plot for democracy and oilincome for the remaining

countries (here we mean a different graph for each country, look up the graph combine function to combine several graphs into one). What do you conclude about the correlation between oil and democracy within country.

??