Midterm Project

Mallika Gupta

2020-10-09

Part 1

Exercise 1

```
elections %>%
 summarize(
   totalvotes_for_clinton = sum(votes_for_clinton)
 )
```

totalvotes_for_clinton
65844241

```
elections %>%
 summarize(
   totalvotes_for_trump = sum(votes_for_trump)
 )
```

totalvotes_for_trump
62979031

Based on the output, Clinton got the most votes.

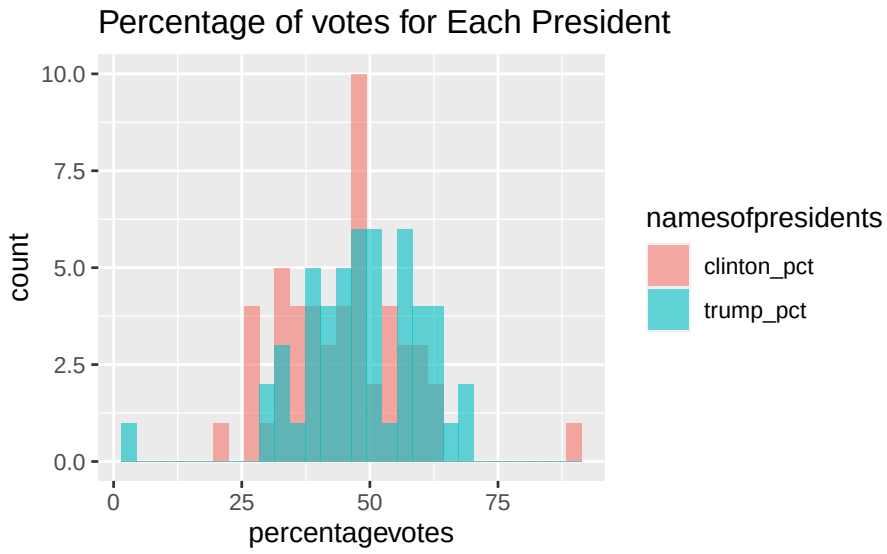
Exercise 2

```
elections2 <- mutate(elections, trump_pct = 100 * votes_for_trump / total_votes,
clinton_pct = 100 * votes_for_clinton/total_votes)
```

Exercise 3

```
elections2 %>%
gather(key = "namesofpresidents", value = "percentagevotes", trump_pct:clinton_pct) %>%
  ggplot() +
geom_histogram(mapping = aes(x = percentagevotes, fill = namesofpresidents),
        bins = 30,
        alpha = 0.6,
        position = "identity"
    ) +
  labs(
```

```
    title = "Percentage of votes for Each President"
  )
```



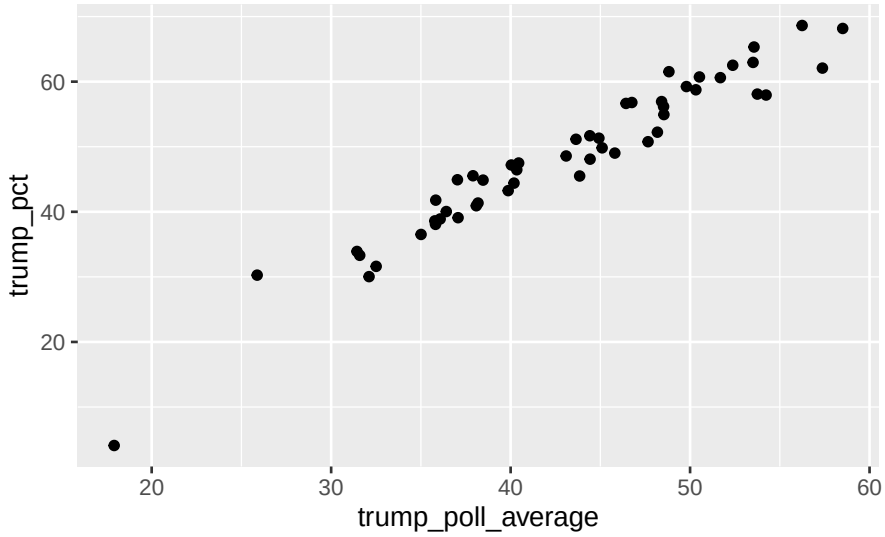
Exercise 4

```
elections3 <- mutate(elections, vote_diff = abs(votes_for_trump - votes_for_clinton)) %>%
  select(state, vote_diff) %>%
  arrange(vote_diff)%>%
  head()
```

Exercise 5

i.

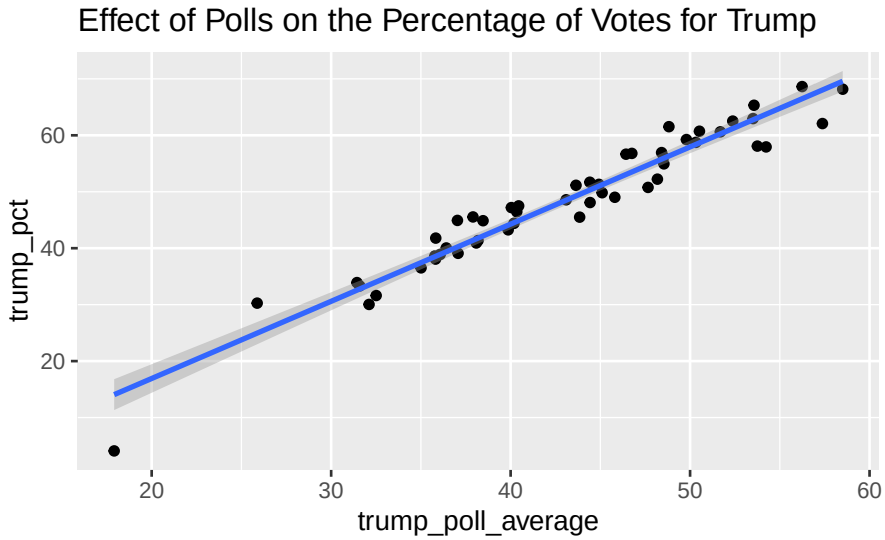
```
ggplot(data = elections2) +
  geom_point(
    mapping = aes(x = trump_poll_average, y = trump_pct)
)
```



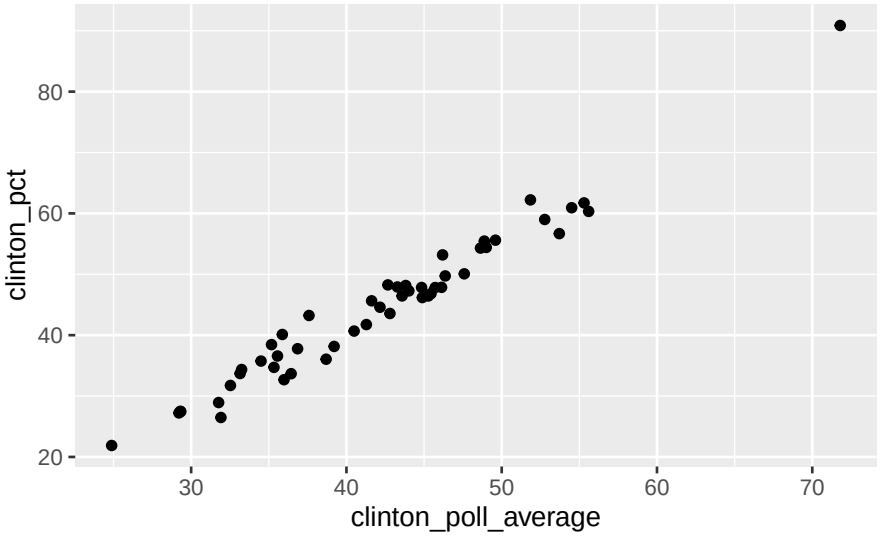

```
election2 <- lm(trump_pct ~ trump_poll_average, data = elections2)
```

```
ggplot(elections2) +
geom_point(mapping = aes(x = trump_poll_average, y = trump_pct)) +
geom_smooth(mapping = aes(x = trump_poll_average, y = trump_pct),
    method = "lm") +
  labs(
    title = "Effect of Polls on the Percentage of Votes for Trump"
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



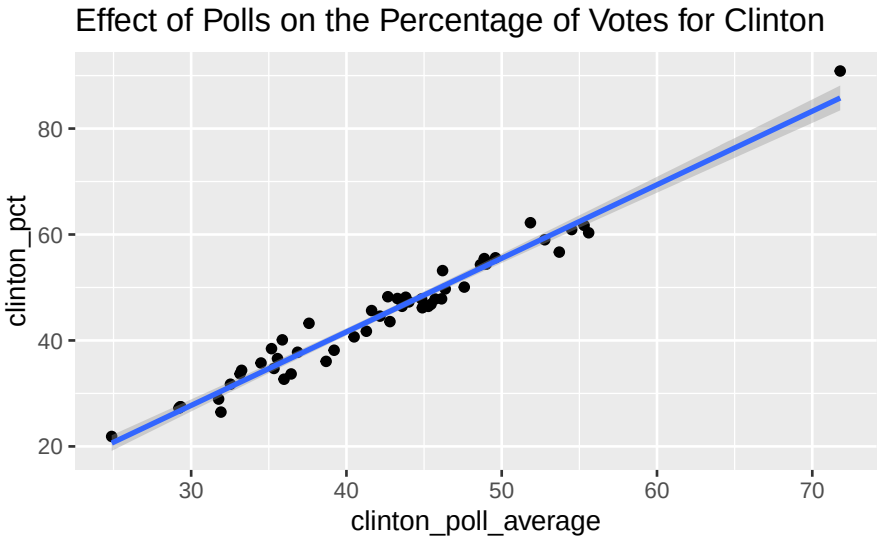

```
ggplot(data = elections2) +
  geom_point(
    mapping = aes(x = clinton_poll_average, y = clinton_pct)
)
```




```
election2 <- lm(clinton_pct ~ clinton_poll_average, data = elections2)
```

```
ggplot(elections2) +
geom_point(mapping = aes(x = clinton_poll_average, y = clinton_pct)) +
geom_smooth(mapping = aes(x = clinton_poll_average, y = clinton_pct),
    method = "lm") +
  labs(
    title = "Effect of Polls on the Percentage of Votes for Clinton"
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



ii.Yes, it looks like there is a linear relationship between the polling and actual vote data. In both the graphs for Trump and Clinton, the points are in a straight line. If the points are in a straight line, there is a relationship between the two variables. Therefore, there is a linear relationship.

The polls are unbiased for both Trump and Clinton. The line of best fit predicts the same number in the election for any poll value. For example, in the scatterolot that shows the information about

Trump, the polling value of 40% corresponds to an election percentage of 40%. In the scatterplot that shows the information about Clinton, the polling value of 20% corresponds to an election percentage of 20%. Yes, the smoothed best - fit lines in my graphs appear to follow this.

Part 2

Variables of Interest

The election - related variable that I have chosen is the votes for Trump. The demographic - related variable that I have chosen is the median household income in each state which is median_hh_income. I am trying to find how the percent of votes for Trump are related to the median income of the state.

Summary Statistics

```
elections2 %>%
summarize(
    mean = mean(votes_for_trump, na.rm = TRUE),
    median = median(votes_for_trump, na.rm = TRUE),
    sd = sd(votes_for_trump, na.rm = TRUE),
    iqr = IQR(votes_for_trump, na.rm = TRUE),
    min = min(votes_for_trump, na.rm = TRUE),
    max = max(votes_for_trump, na.rm = TRUE)
)
```

mean	median	sd	iqr	min	max
1234883	949136	1142243	1198476	12723	4685047

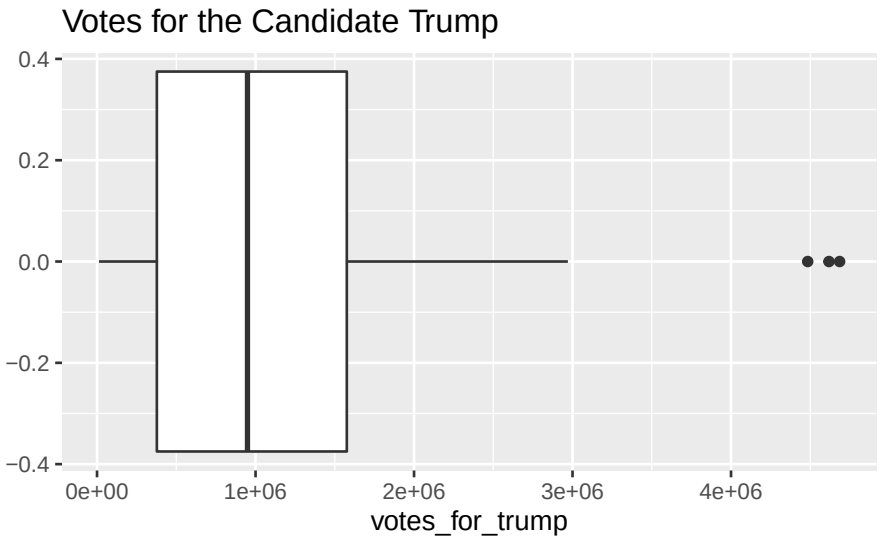
```
elections2 %>%
summarize(
    mean = mean(median_hh_income, na.rm = TRUE),
    median = median(median_hh_income, na.rm = TRUE),
    sd = sd(median_hh_income, na.rm = TRUE),
    iqr = IQR(median_hh_income, na.rm = TRUE),
    min = min(median_hh_income, na.rm = TRUE),
    max = max(median_hh_income, na.rm = TRUE)
)
```

mean	median	sd	iqr	min	max
58142.92	56565	9819.789	14339	41754	78945

Variation of Each Variable

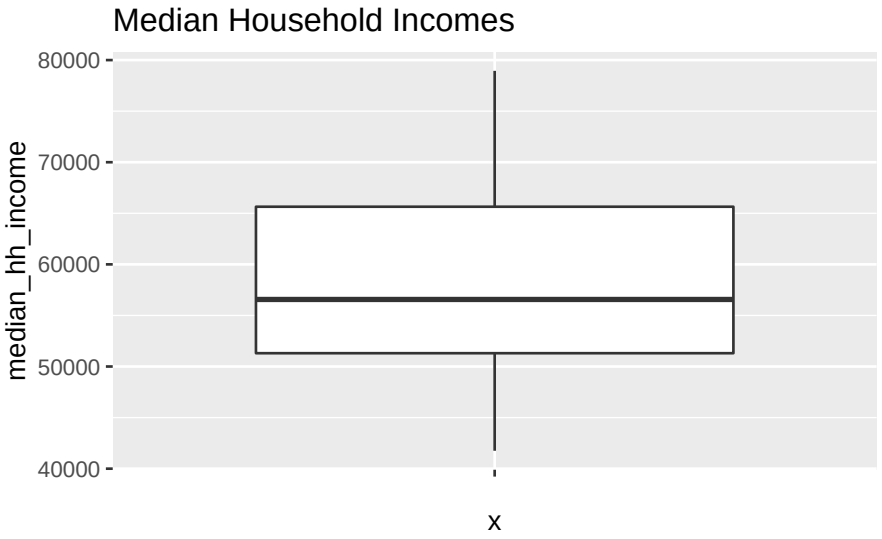
```
ggplot(data = elections2) +
  geom_boxplot(
    mapping = aes(x = votes_for_trump)
```

```
  ) +
labs(
    title = "Votes for the Candidate Trump"
  )
```



The center of this box plot is 1 * 10 ^ 6, which is the median.The shape of the box plot is roughly symmetric.There are three outliers in the boxplot. Most of the data is on the left tail of the boxplot.

```
ggplot(data = elections2) +
  geom_boxplot(
    mapping = aes(x = "", y = median_hh_income)
  ) +
  labs(
    title = "Median Household Incomes"
  )
```

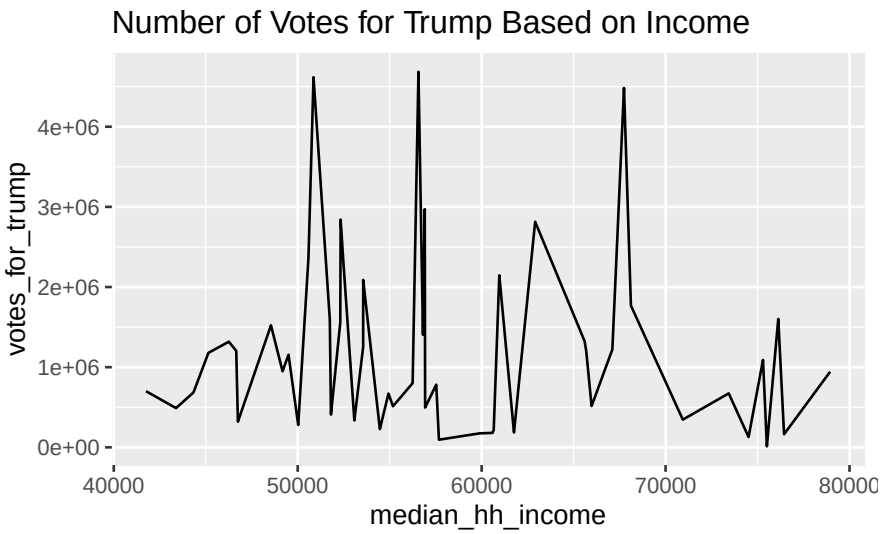


The center of this boxplot is 56,000 dollars which is the median for this boxplot. The 56,000 dollars is the median household income. The boxplot has no outliers. The minimum household income is

around 42,000 dollars. The maximum household income is 80,000 dollars.

Covariation Between Variables

```
ggplot(data=elections2) +
  geom_line(
    mapping = aes(x= median_hh_income, y = votes_for_trump)
  ) +
labs(
    title ="Number of Votes for Trump Based on Income"
  )
```



There is a relationship between the two variables that I have chosen which are median_hh_income and votes_for_trump. According to the line plot above, the median household income is shown to be decreasing and increasing over and over again. The three incomes at which the votes for Trump were high are around 50000 dollars, 56000 dollars, and 66000 dollars. Since it is a line plot, I can not interpret whether the relationship is linear or not linear. I can not also say whether it is strong or weak.

Interpretation

Yes, I have noticed a pattern in how people seem to vote. This can be seen in the box plot and line plot that are shown above. If the median income of a household is high, then there were more votes given to Trump. The households that had a low median come, there were less votes given to Trump. Yes, the exploratory data analysis has confirmed what I expected to find. I expected to find that Trump would get more votes from the households that had a high income.

There could be some confounding variables that could be creating or obscuring a relationship between the election and demographic variables. One of the confounding variables could be that some people might not have voted due to their age. Another confounding variable could be that some people might have voted for Clinton and not Trump. Also, the state in which people voted from could also be a confounding variable.