

# Problem Set 2

*Raj Thakkar*

*22 January 2019*

Note: I worked with Varun Miranda in the class

1. Draw ONE graph that clearly shows the differences in the shape of the distributions of feeling thermometer scores for Clinton, Sanders, Rubio, and Trump. (A faceted plot counts as one graph.) Describe what you see in a paragraph.

Importing the dataset

```
library(readr)
anes_pilot_2016 <- read_csv("anes_pilot_2016.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   version = col_character(),
##   weight = col_double(),
##   weight_spss = col_double(),
##   pid2d = col_character(),
##   pid2r = col_character(),
##   other10_open = col_character(),
##   race_other = col_character(),
##   employ_t = col_character(),
##   religpew_t = col_character(),
##   page_page_follow_timing = col_double(),
##   page_page_turnout12_timing = col_double(),
##   page_page_vote12_timing = col_double(),
##   page_page_percent16_timing = col_double(),
##   page_page_meet_timing = col_double(),
##   page_page_givefut_timing = col_double(),
##   page_page_info_timing = col_double(),
##   page_page_march_timing = col_double(),
##   page_page_sign_timing = col_double(),
##   page_page_give12mo_timing = col_double(),
##   page_page_compromise_timing = col_double()
##   # ... with 147 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
print("Summary of fttrump is: ")
```

```
## [1] "Summary of fttrump is: "
```

```
summary(anes_pilot_2016$fttrump)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   2.00   30.00  40.78   72.00  998.00
```

```
print("Summary of fthrc is: ")
```

```
## [1] "Summary of fthrc is: "
```

```
summary(anes_pilot_2016$fthrc)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   3.00   44.00   43.79   76.00   998.00
```

```
print("Summary of ftsanders is: ")
```

```
## [1] "Summary of ftsanders is: "
```

```
summary(anes_pilot_2016$ftsanders)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   19.00   51.00   56.73   82.00   998.00
```

```
print("Summary of ftrubio is: ")
```

```
## [1] "Summary of ftrubio is: "
```

```
summary(anes_pilot_2016$ftrubio)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.0    15.0    48.0    51.9    61.0    998.0
```

We can see that all the variables have a max value of 998. As per the document, any number above 100 in these variables is to be treated as missing value and has to be omitted. In order to draw the faceted graphs, we need to gather the variables in the dataset and then filter where Score  $\leq 100$ . The filtering is done after gathering to reduce the data loss.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages -----
```

```
## v ggplot2 3.1.0    v purrr  0.2.5
## v tibble  1.4.2    v dplyr  0.7.6
## v tidyr   0.8.2    v stringr 1.3.1
## v ggplot2 3.1.0    v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## -- Conflicts -----
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
```

```
anes_pilot_2016_feeling_gathered <- anes_pilot_2016 %>% gather(key = "Candidate",
  value = "Thermometer_score", c("fttrump", "fthrc", "ftsanders", "ftrubio"))
```

```
# Filtering on the variables and selecting only the desired variables for
# ease
```

```
anes_pilot_2016_feeling_therm_gathered <- anes_pilot_2016_feeling_gathered %>%
  filter(Thermometer_score <= 100) %>% select(immig_numb, Candidate, Thermometer_score)
```

```
# Recoding the candidate names for ease of understanding
```

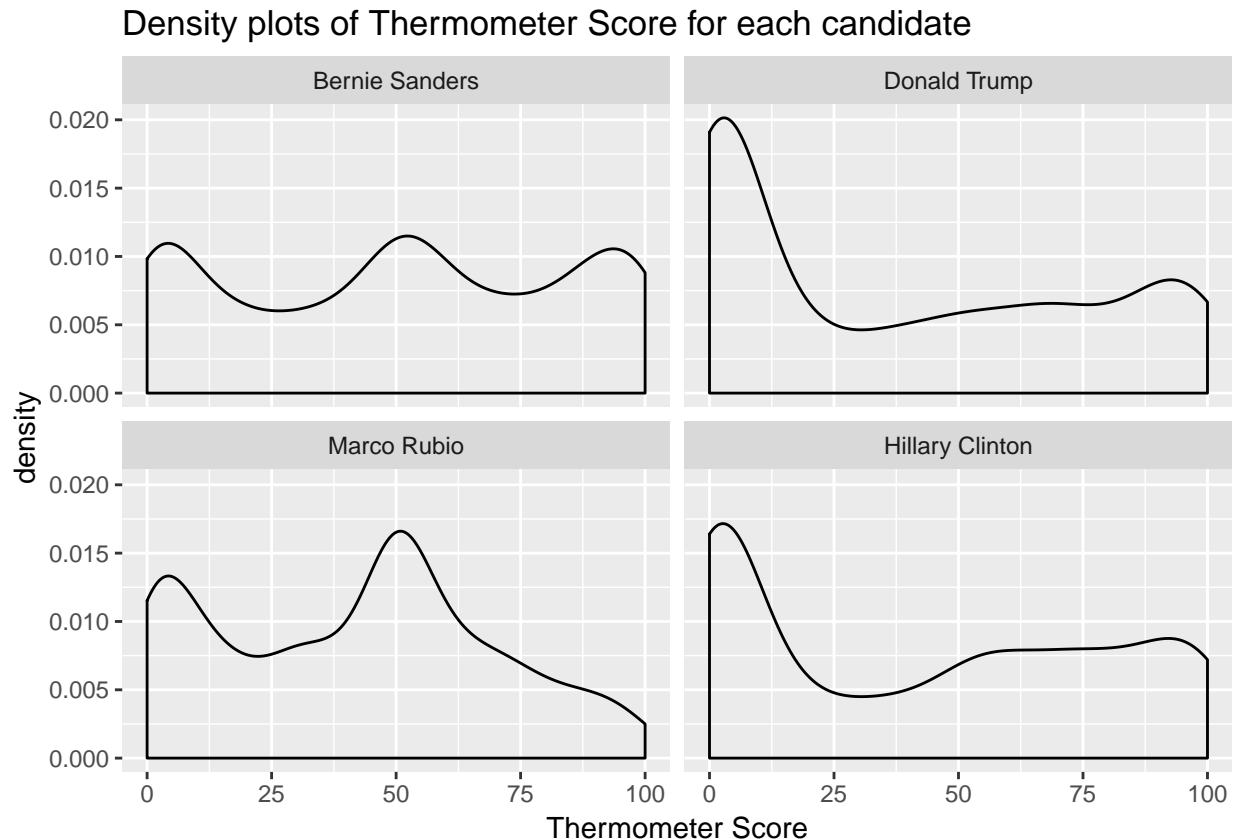
```
candidate = recode(anes_pilot_2016_feeling_therm_gathered$Candidate, fttrump = "Donald Trump",
  fthrc = "Hillary Clinton", ftsanders = "Bernie Sanders", ftrubio = "Marco Rubio")
```

```
# Changing the levels for easy comparison of distributions
```

```
anes_pilot_2016_feeling_therm_gathered$Candidate <- factor(candidate, levels = c("Bernie Sanders",  
"Donald Trump", "Marco Rubio", "Hillary Clinton"), labels = c("Bernie Sanders",  
"Donald Trump", "Marco Rubio", "Hillary Clinton"))
```

Now we will plot the graphs to compare the distributions

```
ggplot(anes_pilot_2016_feeling_therm_gathered, aes(x = Thermometer_score)) +  
  geom_density() + facet_wrap(~Candidate, ncol = 2) + xlab("Thermometer Score") +  
  ggtitle("Density plots of Thermometer Score for each candidate")
```



From the above graph, we can say that the spread of thermometer scores for each candidate is between 0 to 100. The distribution of scores for Marco Rubio has the highest peak near 50 which shows that many respondents were unsure about their feelings towards Marco Rubio. The distribution of Bernie Sanders is tri-modal and has peaks near 0, 50 and 100. The distributions of scores for Hillary Clinton and Donald Trump are interesting because they have similar shapes and they have the highest peaks near 0 which shows that many respondents felt cold towards them.

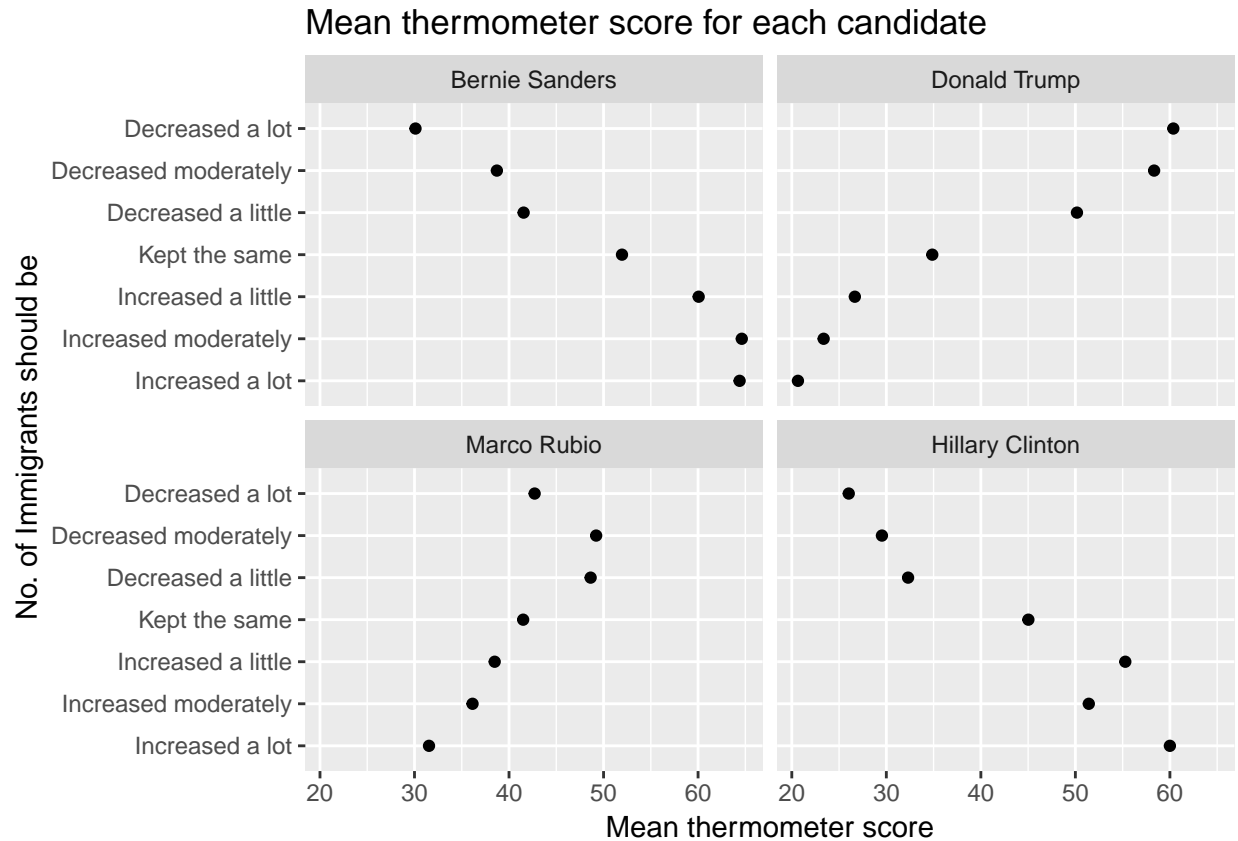
**2. Draw ONE graph that clearly shows the differences in the mean feeling thermometer score for each level of immigration attitude for each of the four candidates. (Again, a faceted plot counts as one graph.) Describe what you see in a paragraph.**

```
# Manipulating the data using group by to bring the data in desired format
```

```
anes_pilot_2016_feeling_therm_mean <- anes_pilot_2016_feeling_therm_gathered %>%  
  group_by(immig_numb, Candidate) %>% summarise(mean_ther_score = mean(Thermometer_score))
```

Plotting the means of each candidate by each group

```
anes_pilot_2016_feeling_therm_mean$immig_numb <- factor(anes_pilot_2016_feeling_therm_mean$immig_numb,
  levels = c(1, 2, 3, 4, 5, 6, 7), labels = c("Increased a lot", "Increased moderately",
    "Increased a little", "Kept the same", "Decreased a little", "Decreased moderately",
    "Decreased a lot"))
ggplot(anes_pilot_2016_feeling_therm_mean, aes(x = mean_ther_score, y = immig_numb)) +
  geom_point() + facet_wrap(~Candidate, ncol = 2) + xlab("Mean thermometer score") +
  ylab("No. of Immigrants should be") + ggtitle("Mean thermometer score for each candidate")
```



From the graph above, we can see that people who believe that number of immigrants should be “Decreased a lot” gave the highest mean score to Donald Trump while the mean score for Hillary Clinton was the lowest. Moreover, as we go from “Increased a lot” (immig\_numb = 0) to “Decreased a lot” (imig\_numb = 7) the mean thermometer score for Donald Trump increases, it decreases for Bernie Sanders and almost always decreases for Hillary Clinton. We don’t spot any particular trend for Marco Rubio though.