Data Viz: Homework 2

Reading:

a. Chapter 2 in Kieran Healy's book "Data Visualization". This should be very familar.
b. Chapter 3 in Kieran Healy's book "Data Visualization". Please pay attention to the illustration that shows how ggplot works:

```
(Tidy) Data -> Mapping (through `aes()`) -> Geoms -> Coordinates -> Labels
```

Also, after reading you should understand the difference between

`ggplot(data=mydata, aes(x=x,y=y, size=2, color="green")) + geom_point()`

and

`ggplot(data=mydata, aes(x=x,y=y)) + geom_point(size=2, color=green)`

This is the difference between MAPPING aesthetics and just SETTING aesthetics

c. Read Section 3 on my R handout on "Exploring the GSS", available on Canvas.
d. Read the "Manupulating Factors" R handout on Canvas

##1. Setup ### options Set up global options

libraries

Load in needed libraries

```r
library(tidyverse)
```

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

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ----------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(RColorBrewer)
library(haven)
```

2. File management

Create variables for directories

```r
project.dir <- getwd() #naeem
output.dir <- "/Output"
data.dir <- "C:/Users/Naeem Cho/Desktop/School Work/Data_Viz/Datasets"
setwd(project.dir)
getwd()
```

```
## [1] "C:/Users/Naeem Cho/Desktop/School Work/Data_Viz/Data_Viz/2020-03-26-HW2"
```

3. Importing Data

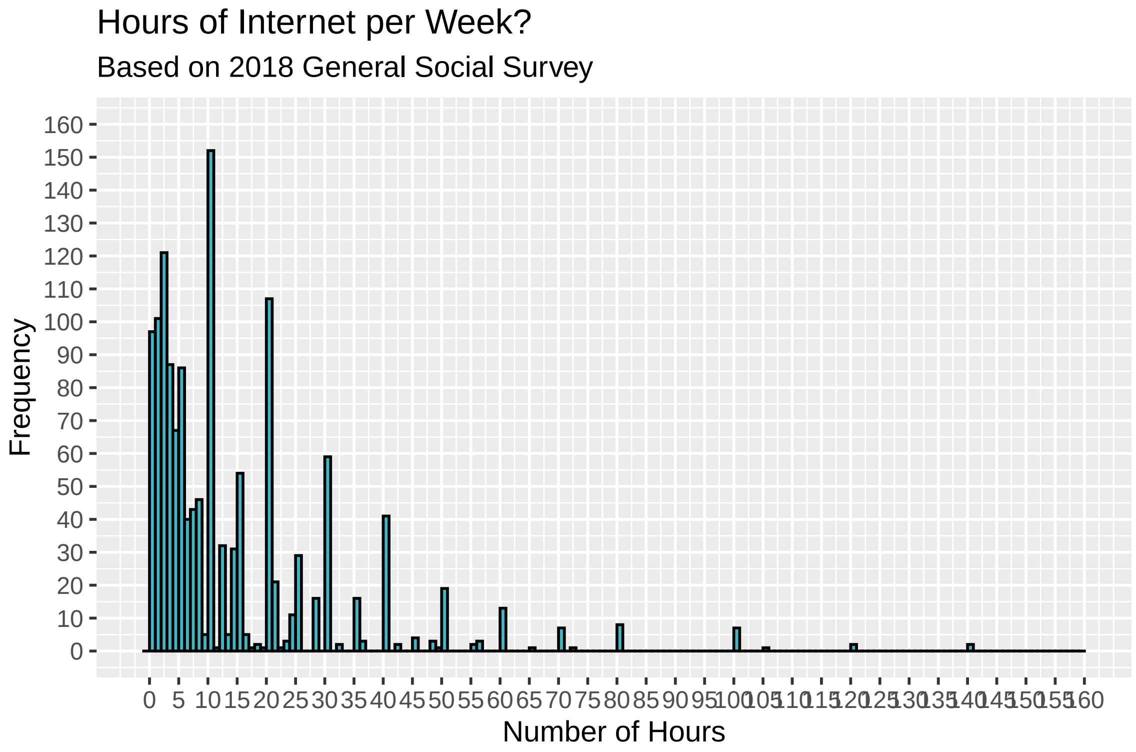
```
GSS_Data <- read_dta(file.path(data.dir, "GSS2018.dta"))
```

Projects:

1. Pick a **continuous** variable from the 2018 General Social Survey and visualize its distribution through two different graphs. Prepare three slides: the first two showing your two graphs (of the same data) and a third slide highlighting some of the R code you used to obtain one of your graphs.

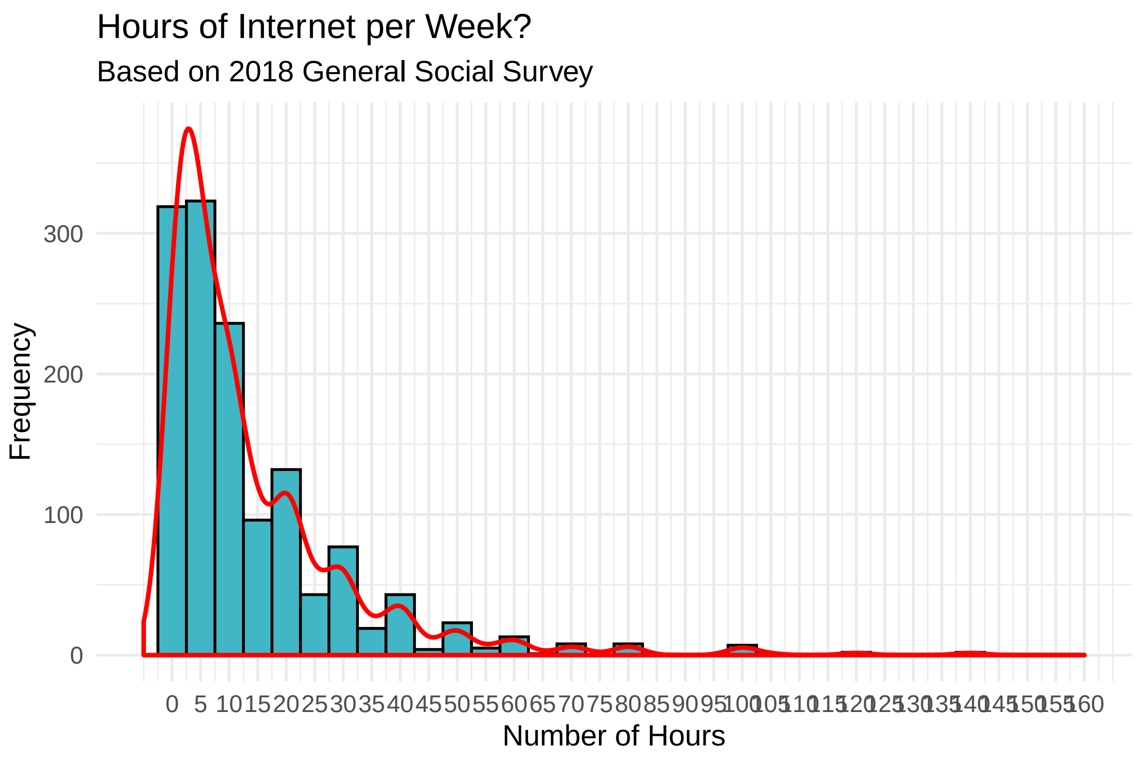
```
# Reading in dataset of number of hours of internet use per week
wwwhr <- GSS_Data %>%
  select(wwwhr) %>%
  filter(wwwhr != -1 & wwwhr != 998 & wwwhr != 999 ) %>%
  drop_na()
```

```
# Histogram of Hours of Internet per Week

wwwhist <- wwwhr %>%
  ggplot(aes(x = wwwhr)) +
  labs(title = "Hours of Internet per Week?",
       subtitle="Based on 2018 General Social Survey",
       x = "Number of Hours",
       y = "Frequency") +
  geom_histogram(color = "black",
                 fill = "#41B6C4",
                 boundary = 0,
                 binwidth = 1,
                 closed = "left") +
  scale_y_continuous(limits = c(0,160),
                     breaks = seq(0,160,10)) +
  scale_x_continuous(limits = c(-1, 160),
                     breaks = seq(0,160, 5))

wwwhist
```



```
# Histogram with density overlay
wwwdensity <- wwwhr %>%
  ggplot(aes(x = wwwhr)) +
  labs(title = "Hours of Internet per Week?",
       subtitle="Based on 2018 General Social Survey",
       x = "Number of Hours",
       y = "Frequency") +
  geom_histogram(aes(y = ..count..),
                 color = "black",
                 fill = "#41B6C4",

                 binwidth = 5,
                 closed = "left") +
  geom_density(aes(y = 5*(..count..)), adjust = 1, size = 0.8, color = "red") +
  theme_minimal() +
  scale_x_continuous(limits = c(-5, 160),
                     breaks = seq(0,160, 5))

wwwdensity
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Hours of Internet per Week?

Based on 2018 General Social Survey

2. Pick one *categorical* and one *continuous* variable from the LendingClub dataset and explore the distribution through a graph. For each graph, write one short paragraph describing the distribution of the variable. (For categorical variables, perhaps mention the number of categories and the most common (or two most common) categories. For continuous variables, mention (i) shape, (ii) center, (iii) variability, and (iv) unusual features, such as outliers.

Reading in the Data

```
lending <- read_csv(file.path(data.dir, "LendingClubLoanData.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   member_id = col_logical(),
##   term = col_character(),
##   grade = col_character(),
##   sub_grade = col_character(),
##   emp_title = col_character(),
##   emp_length = col_character(),
##   home_ownership = col_character(),
##   verification_status = col_character(),
##   issue_d = col_character(),
##   loan_status = col_character(),
##   pymnt_plan = col_character(),
##   url = col_character(),
##   desc = col_logical(),
##   purpose = col_character(),
##   title = col_character(),
##   zip_code = col_character(),
##   addr_state = col_character(),
##   earliest_cr_line = col_character(),
##   initial_list_status = col_character(),
##   last_pymnt_d = col_character()
##   # ... with 30 more columns
## )

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

## Warning: 1456203 parsing failures.
##   row  col           expected
##  1481 desc 1/0/T/F/TRUE/FALSE We knew that using our credit cards to finance an adoption would squeez
## 33066 desc 1/0/T/F/TRUE/FALSE I had a bad year two years ago, with some late and missed payments. I'm
## 37861 desc 1/0/T/F/TRUE/FALSE Lenders,  I have the ability to pay off my current debt but, would like
## 50495 desc 1/0/T/F/TRUE/FALSE I paid off my first Prosper loan, but had an emergency and took out a s
## 68328 desc 1/0/T/F/TRUE/FALSE I want to cut down on my credit  card debt now (while they are not wild
## ..... .... .................. ......................................................................
## See problems(...) for more details.
```

```
# Looking at variables to choose
# lending %>% colnames()

lending %>% select(int_rate, home_ownership) %>% drop_na()
```

```
## # A tibble: 2,260,668 x 2
##    int_rate home_ownership
##       <dbl> <chr>
##  1    14.0  MORTGAGE
##  2    12.0  MORTGAGE
##  3    10.8  MORTGAGE
##  4    14.8  MORTGAGE
##  5    22.4  MORTGAGE
##  6    13.4  RENT
##  7     9.17 MORTGAGE
##  8     8.49 MORTGAGE
##  9     6.49 RENT
## 10    11.5  MORTGAGE
## # ... with 2,260,658 more rows
```

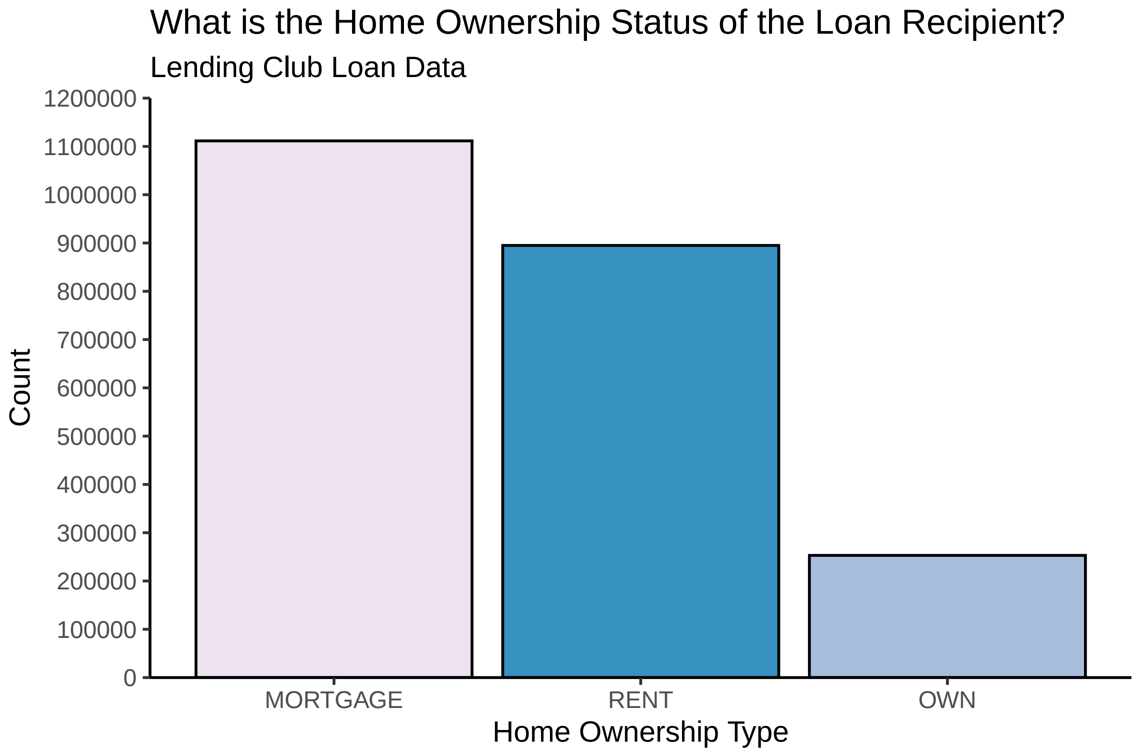
A Categorical Variable - Home Ownership

Below I've produced a bargraph, in descending order, of the home ownership type of loan recipients in the Lending Club data. Although the variable initially had 6 categories, I removed 3 of the categories due to their comparably low counts. What we are left with is Mortgage, Rent, and Own. It is clear that Mortgages are the most common type of ownership, followed closely by Rent. Home owners constitute a much smaller proportion of the loan recipients.

```
# Color Palette
mycols <- c(brewer.pal(9, "PuBuGn")[2], brewer.pal(9, "PuBuGn")[4], brewer.pal(9, "PuBuGn")[6])


home.contingency <- lending %>% select(home_ownership) %>%
  filter(home_ownership != "ANY" & home_ownership != "OTHER" & home_ownership != "NONE") %>%
  table()

home.contingency
```

```
## .
## MORTGAGE      OWN     RENT
##  1111450   253057   894929
```

```
home.table <- data.frame(home.contingency = c("MORTGAGE","OWN", "RENT"), count = as.vector(home.conting

home.bar <- home.table %>% ggplot( aes(x=reorder(home.contingency, -count),y=count, fill = home.conting
  geom_bar(stat="identity", color = "black") +
  scale_fill_manual(values=mycols) +
  theme_classic() +
  ylab("Count") +
  xlab("Home Ownership Type") +
  ggtitle(label="What is the Home Ownership Status of the Loan Recipient?", subtitle="Lending Club Loan
  scale_y_continuous(limit=c(0,1200000), expand=c(0,0), breaks = seq(0,1200000,100000)) +
#  annotate("text", x = c('Yes', 'No'), y = c(30.8, 69.2) -4, label = c("30.8%", "69.2%"), color = "whi
  theme(legend.position = 'none')

home.bar
```



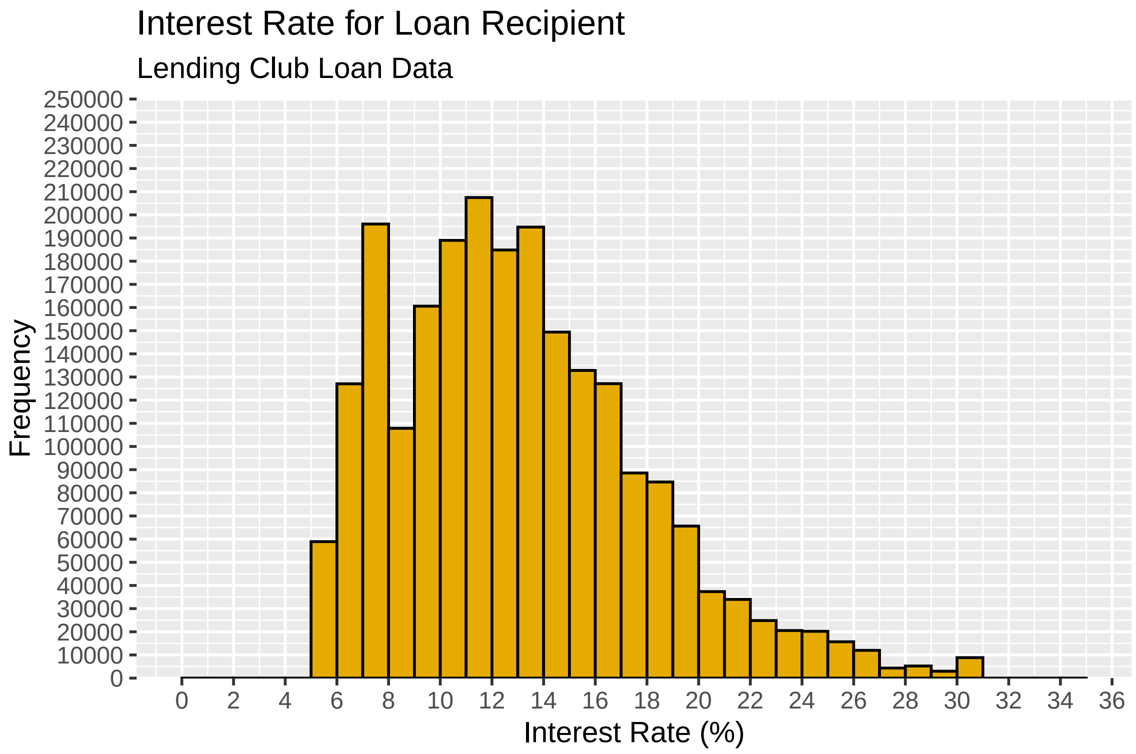
A Continuous Variable - Interest Rate

Below I've created two graphical representations of the distribution of interest rates for loans in the Lending Club Loan Data. The distribution is roughly bimodal and right-skewed. The data has a variance of 2.2, and a median of 12.62%. According to base R, the data has about 41 thousand statistical outliers.

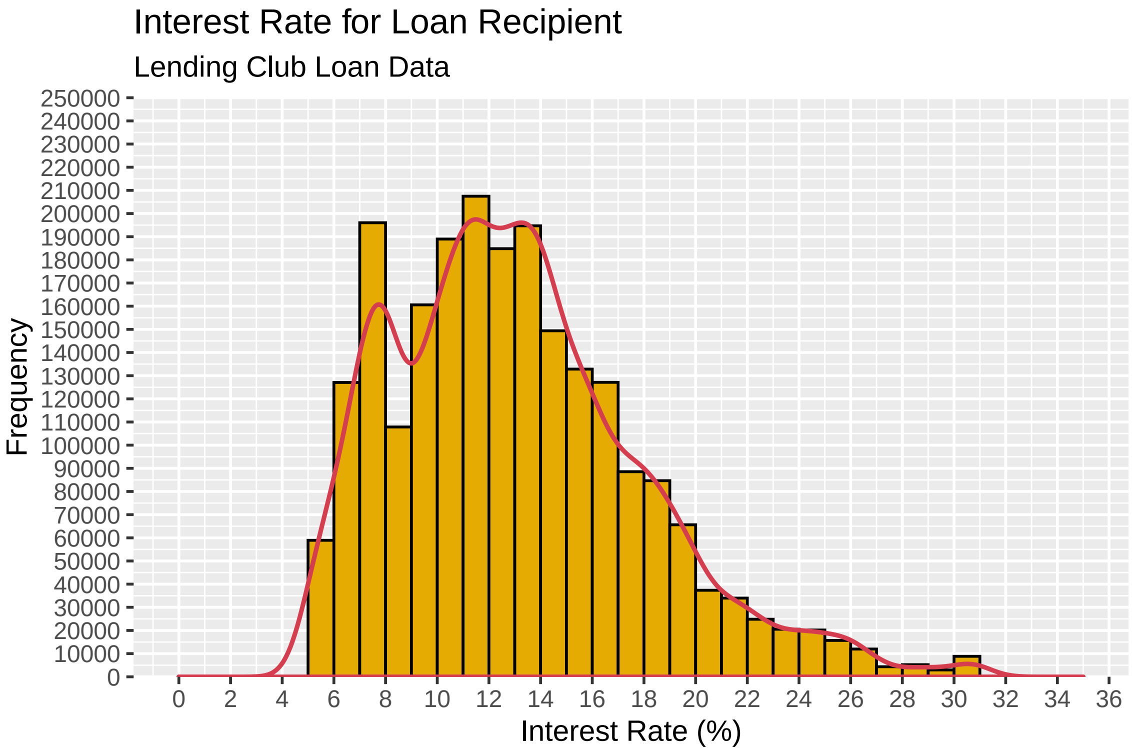
```
### Sample Statistics
int_rates <- lending %>% select(int_rate) %>% select_if(is.numeric) %>% drop_na()

sd <- int_rates %>% sapply(sd)
var <- sqrt(sd)
```

```
med <- int_rates %>% summarise(med = median(int_rate))

OutVals = boxplot(int_rates$int_rate, plot=FALSE)$out
```

```
### Producing a histogram.


intrate.hist <- int_rates %>%
  ggplot(aes(x = int_rate)) +
  labs(title = "Interest Rate for Loan Recipient",
       subtitle="Lending Club Loan Data",
       x = "Interest Rate (%)",
       y = "Frequency") +
  geom_histogram(color = "black",
                 fill = "#E6AB02",
                 boundary = 0,
                 binwidth = 1,
                 closed = "left") +
  scale_y_continuous(limit=c(0,250000),
                     expand=c(0,0),
                     breaks = seq(0,250000,10000)) +
  scale_x_continuous(limits = c(0, 35),
                     breaks = seq(0,36, 2))

intrate.hist
```



```
### Producing a density overlay.
intrate.overlay <- int_rates %>%
  ggplot(aes(x = int_rate)) +
  labs(title = "Interest Rate for Loan Recipient",
       subtitle="Lending Club Loan Data",
       x = "Interest Rate (%)",
       y = "Frequency") +
  geom_histogram(aes(y = ..count..),
                 color = "black",
                 fill = "#E6AB02",
                 boundary = 0,
                 binwidth = 1,
                 closed = "left") +
   geom_density(aes(y = ..count..), adjust = 3, size = 0.8, color = "#D53E4F")+
  scale_y_continuous(limit=c(0,250000),
                     expand=c(0,0),
                     breaks = seq(0,250000,10000)) +
  scale_x_continuous(limits = c(0, 35),
                     breaks = seq(0,36, 2))


intrate.overlay
```


Interest Rate for Loan Recipient
Lending Club Loan Data
Frequency
Interest Rate (%)