Data Viz: Homework 3

Reading:

a. Chapter 4 in Kieran Healy's book "Data Visualization". A good review of what we have done so far.
b. Chapter 5 in Kieran Healy's book "Data Visualization". We have already seen the use of tidyverse commands, but there are some new graphs in there.

1. Lending Club Data

We already started working on this at the end of class, but now finish creatin a graph that compares the distribution (in terms of percentages) of one categorical varaible (with at least three categories) for given levels of another categorical variable (with also at least 3 categories) through

Data Read-in

```
library(tidyr)
library(tidyverse)
```

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

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

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

```
sample <- read_csv('C:\\Users\\Naeem Cho\\Desktop\\School Work\\Data_Viz\\Datasets\\LendingClubLoanData
```

```
## 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
## )
```

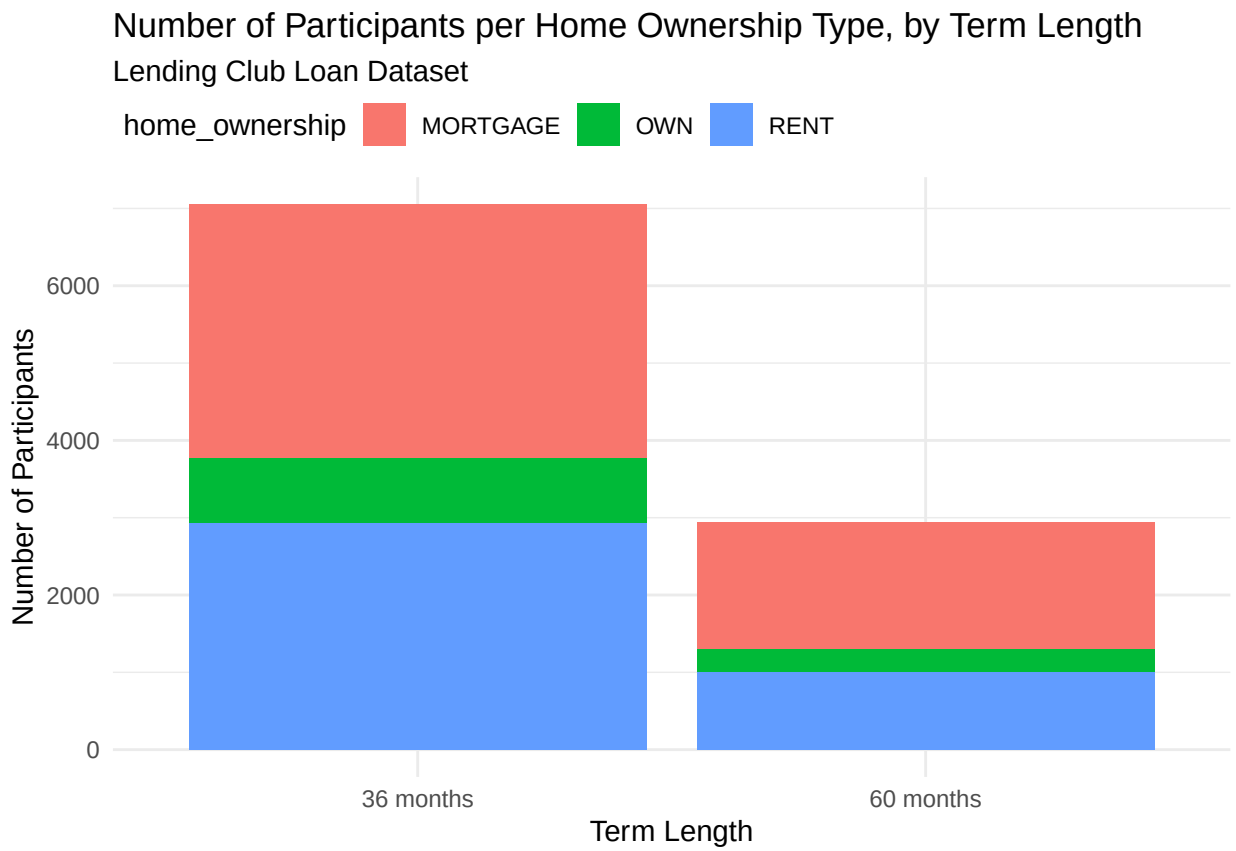
```
## See spec(...) for full column specifications.
```

```
## Warning: 1 parsing failure.
##  row  col           expected
## 1481 desc 1/0/T/F/TRUE/FALSE We knew that using our credit cards to finance an adoption would squeez
```

```
id.unique <- sample %>% distinct(id, .keep_all = TRUE)

sample.home <- sample %>% group_by(term) %>% count(home_ownership) %>% mutate(total = sum(n), prop = n/
```

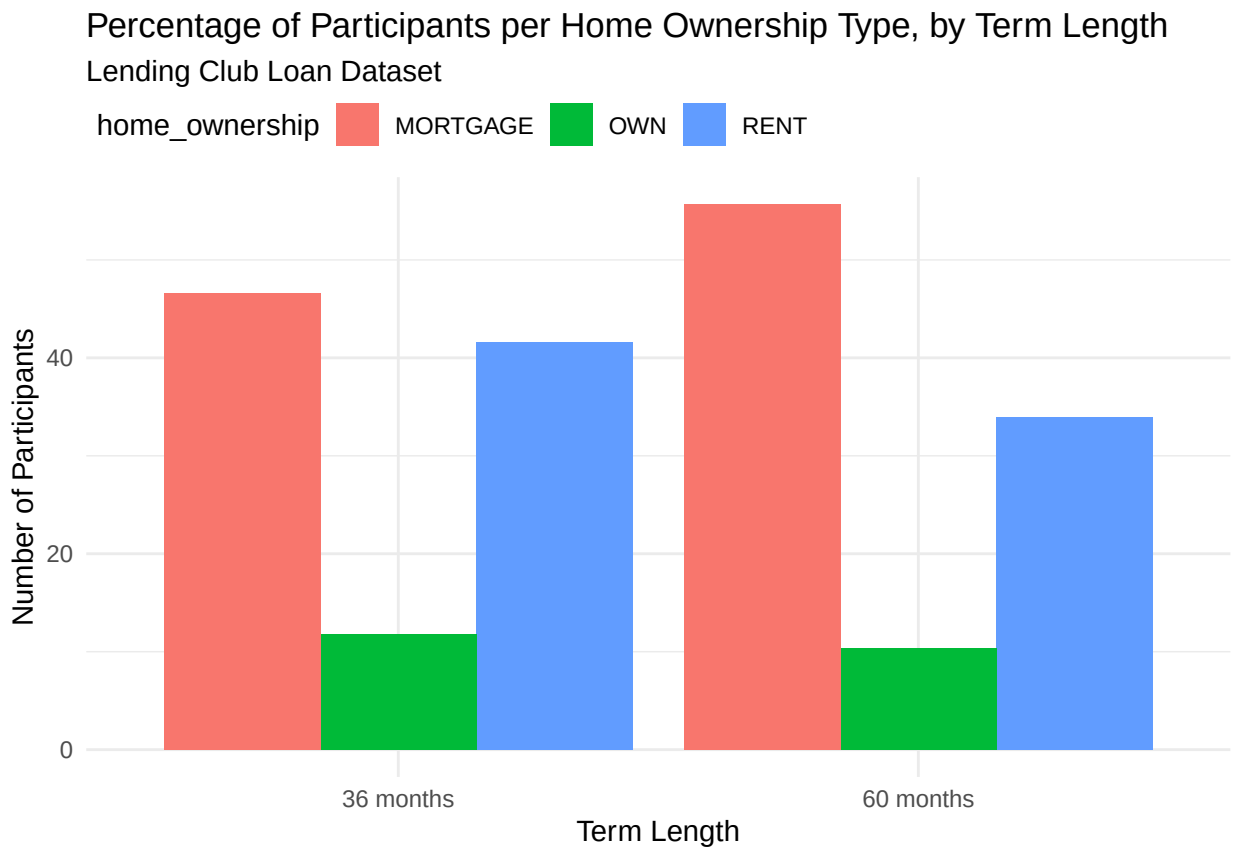
a) a stacked bar chart

```
pA <- ggplot(sample.home,
       aes (x = term, y = n)) +
  geom_bar(aes(fill = home_ownership), stat = "identity") +
  theme_minimal() +
  labs(title = "Number of Participants per Home Ownership Type, by Term Length",
       subtitle = 'Lending Club Loan Dataset',
       x = 'Term Length', y = 'Number of Participants') +
  theme(legend.position = 'top',
        legend.justification = 'left',
        legend.margin = margin(0,1,0,4),
        legend.key = element_rect(fill = NA, color = NA))


pA
```



b) a side-by-side bar chart

```
pB <- ggplot(sample.home,
             aes (x = term, y = 100*prop)) +
  geom_bar(aes(fill = home_ownership), stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Percentage of Participants per Home Ownership Type, by Term Length",
       subtitle = 'Lending Club Loan Dataset',
       x = 'Term Length', y = 'Number of Participants') +
  theme(legend.position = 'top',
        legend.justification = 'left',
        legend.margin = margin(0,1,0,4),
        legend.key = element_rect(fill = NA, color = NA))

pB
```



c) a separate bar chart through faceting

```
pC <- ggplot(sample.home,
             aes (x = term, y = 100*prop)) +
  geom_bar(aes(fill = home_ownership), stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Percentage of Participants per Home Ownership Type, by Term Length",
       subtitle = 'Lending Club Loan Dataset',
       x = 'Term Length', y = 'Number of Participants') +
  theme(legend.position = 'top',
        legend.justification = 'left',
        legend.margin = margin(0,1,0,4),
        legend.key = element_rect(fill = NA, color = NA))
```

Make each graph as close as possible to "production ready", with appropriate lables, legend, title, etc. Briefly comment on which of the three graphs you would prefer when you have to explain the data to others, and why.

2. Post to Social Media

For data of interest to you (you could use variables from the GSS, or World Bank data, any source listed under www.google.com/publicdata, or anything else), create a scatterplot that shows an interesting relationship between two variables, but also conveys information on a third (continuous or categorical) variable. Prepare a short statement about the story the graph is telling of no more than 280 characters. Save the graph in a useful format so that you can share it with others on social media. (We haven't talked about saving ggplots, but do research on your own, in particular about `ggsave`.) Use the graph and the statement to post it to your favorite social media account (this could be a tweet you are sending, a facebook post, an Instagram

post, a What's App message, etc., about "Look at my cool class project from Data Visualization"). If you are not on Social Media, create a Github repository where you showcase your work.

What to turn in: I need to see a screenshot of your post.

In class next week Thursday I'm going to ask some of you to share their posts or repositories!