Inpatient Charge Data 2016

Antonio Avila

April 6, 2019

Begin by loading in the data

```
med_data = read_csv("medicare_data.csv", guess_max = 112000)
```

```
## Parsed with column specification:
## cols(
##   `DRG Definition` = col_character(),
##   `Provider Id` = col_double(),
##   `Provider Name` = col_character(),
##   `Provider Street Address` = col_character(),
##   `Provider City` = col_character(),
##   `Provider State` = col_character(),
##   `Provider Zip Code` = col_double(),
##   `Hospital Referral Region (HRR) Description` = col_character(),
##   `Total Discharges` = col_number(),
##   `Average Covered Charges` = col_character(),
##   `Average Total Payments` = col_character(),
##   `Average Medicare Payments` = col_character()
## )
```

```
real_names = names(med_data)
names(med_data) <- c("DRG", "ID", "Provider", "Address", "City", "State", "Zip", "HRR", "Discharges", "
```

There seems to be a problem parsing the data. The variable "Total Discharges" doesn't read in a few of the observations correctly because they're value is above 1,000. The commas seem to be affecting the parsing of those particular observations. In addition, The charges and payments variables are being parsed in as character types instead of numeric (or doubles) because of the dollar sign.

```
parse2num <- med_data %>%
    select("AvgCharge":"AvgMedPmts") %>%
    purrr::map(parse_number) %>%
    as_tibble()

med_data <- med_data %>%
  select(-("AvgCharge":"AvgMedPmts")) %>%
  bind_cols(parse2num)
```

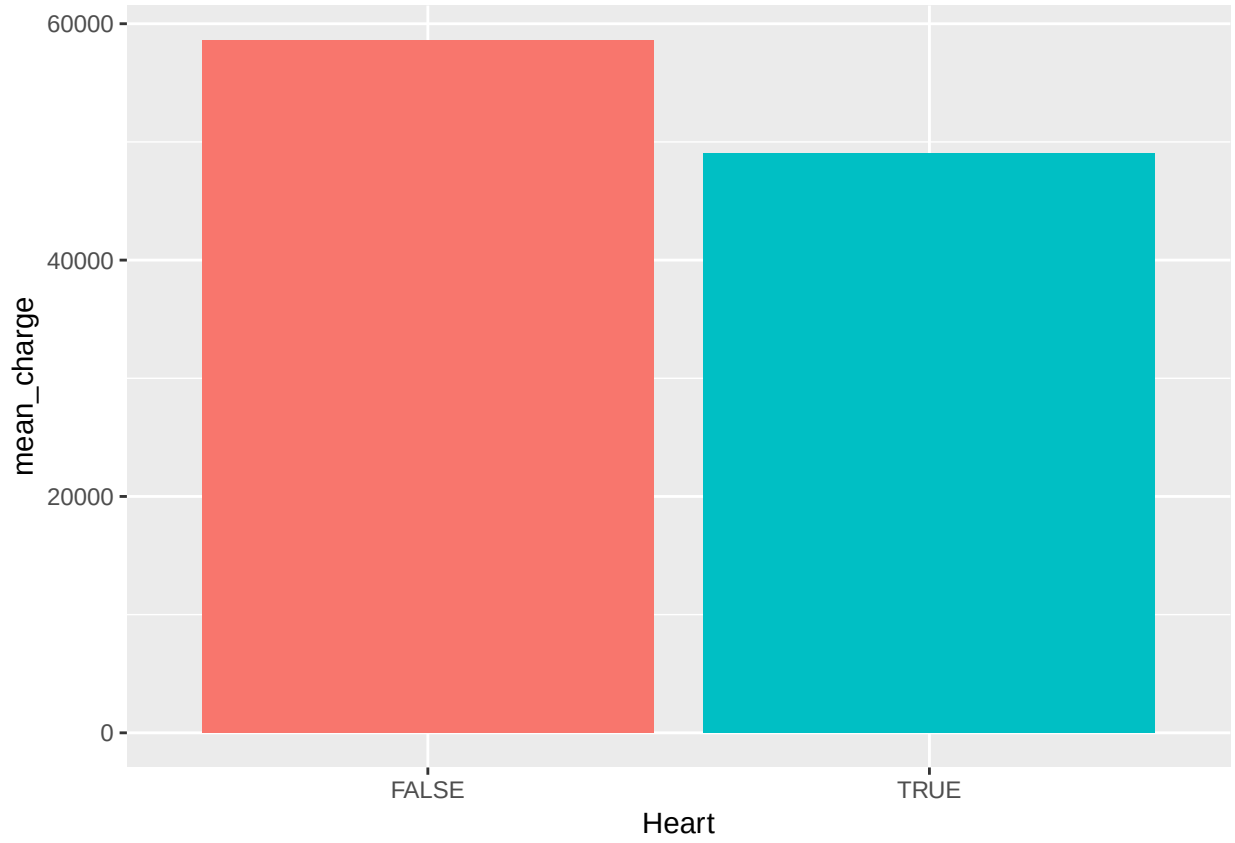
Fixed the parsing issue for the Total Discharges column by extending the number of rows the read_csv() function reads in to determine the type of column it is to 120,000 since the first occurrence of a value over 1,000 occurred at about the 117,00th row, thus fixing the problem. Secondly, converted the Average dollar payment columns into numeric columns, dropping the dollar symbol and ensuring the values are of the numeric type.

Having fixed the parsing issues, I can begin cleaning the data a little. I will begin by separating the code and descriptions from the DRG column to shorten it. The DRG Code are unique to their descriptions so I will separate the two. The code will be used for general analysis since it is compact, making it easier to display on graphics while the description will be kept in case I want to group and subset of the data based on a more general type of procedure, i.e. heart procedures, respiratory, etc. This type of grouping can be easily be done by looking for key words in their descriptions, whereas the code provides no clue on how to do this, making it more difficult to automate.

```r
med_data <- med_data %>%
  separate(DRG, c("DRG_Code", "DRG_Descr"), sep = 3)

med_data$DRG_Descr = str_sub(med_data$DRG_Descr, 4)
```

Even though procedure are already categorized into groups via the DRG classification system, it may be worth exploring whether certain groups of procedures are more expensive than others; for example, heart related procedures could be more or less expensive than other types of procedures even though not all heart procedures have the same level of severity.

```r
med_data <- med_data %>%
  mutate(Heart = str_detect(DRG_Descr, "HEART"))


med_data %>%
  group_by(Heart) %>%
  summarise(mean_charge = mean(AvgCharge)) %>%
  ggplot(aes(Heart, mean_charge)) +
    geom_bar(stat = "identity", aes(fill = Heart), show.legend = FALSE)
```

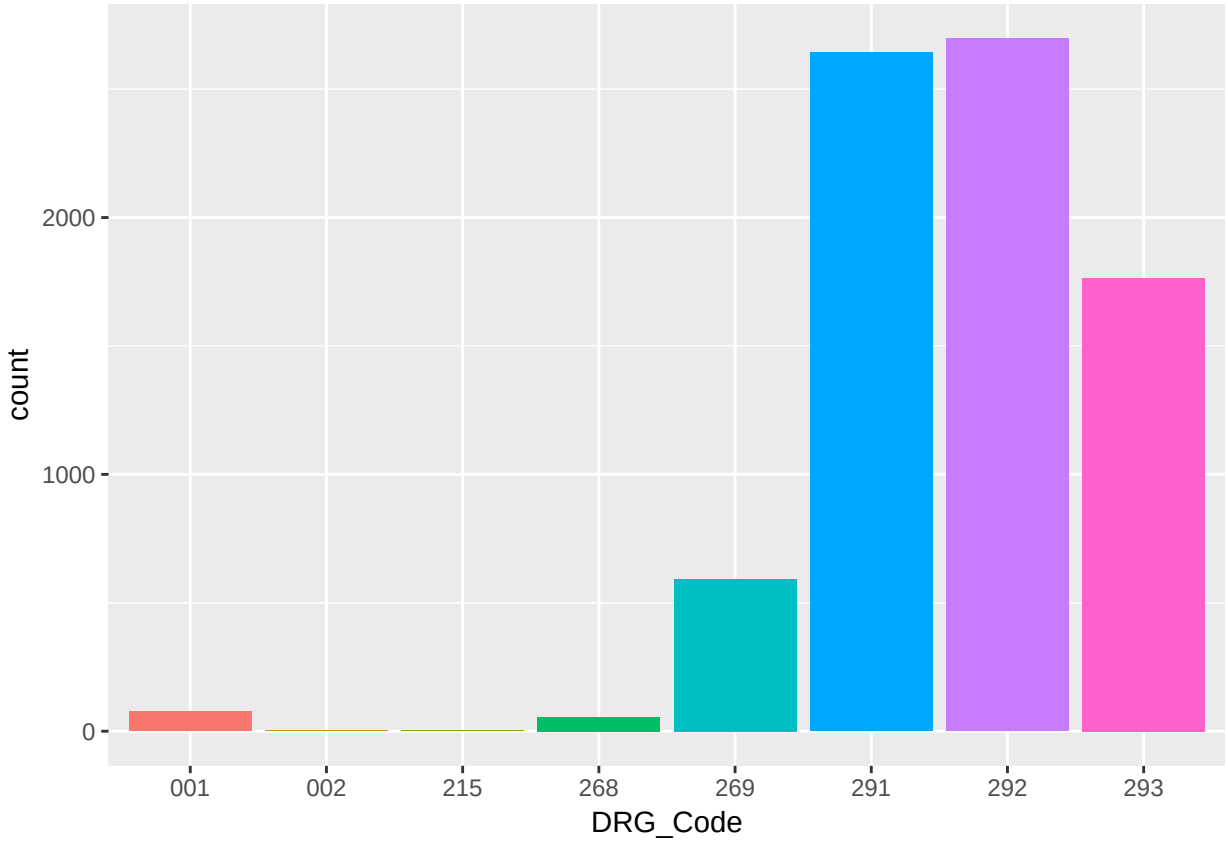


In this case, it turns out that Heart related procedures as a whole are not more expensive when compared to all others, which is a little unexpected. I would expect heart related procedures to be more expensive in general because it is a vital organ and any type of major procedures is sure to be invasive, causing the need to consult a specialist. The average charge being lower may be because there are many more procedures provided that may not be very severe nor expensive. May be worth it to take a look and confirm if this is correct.

```r
heart_only <- med_data %>%
  filter(Heart == TRUE)

heart_only %>%
  count(DRG_Code)
```

```
## # A tibble: 8 x 2
##   DRG_Code     n
##   <chr>    <int>
## 1 001         77
## 2 002          3
## 3 215          2
## 4 268         56
## 5 269        592
## 6 291       2643
## 7 292       2697
## 8 293       1765
```

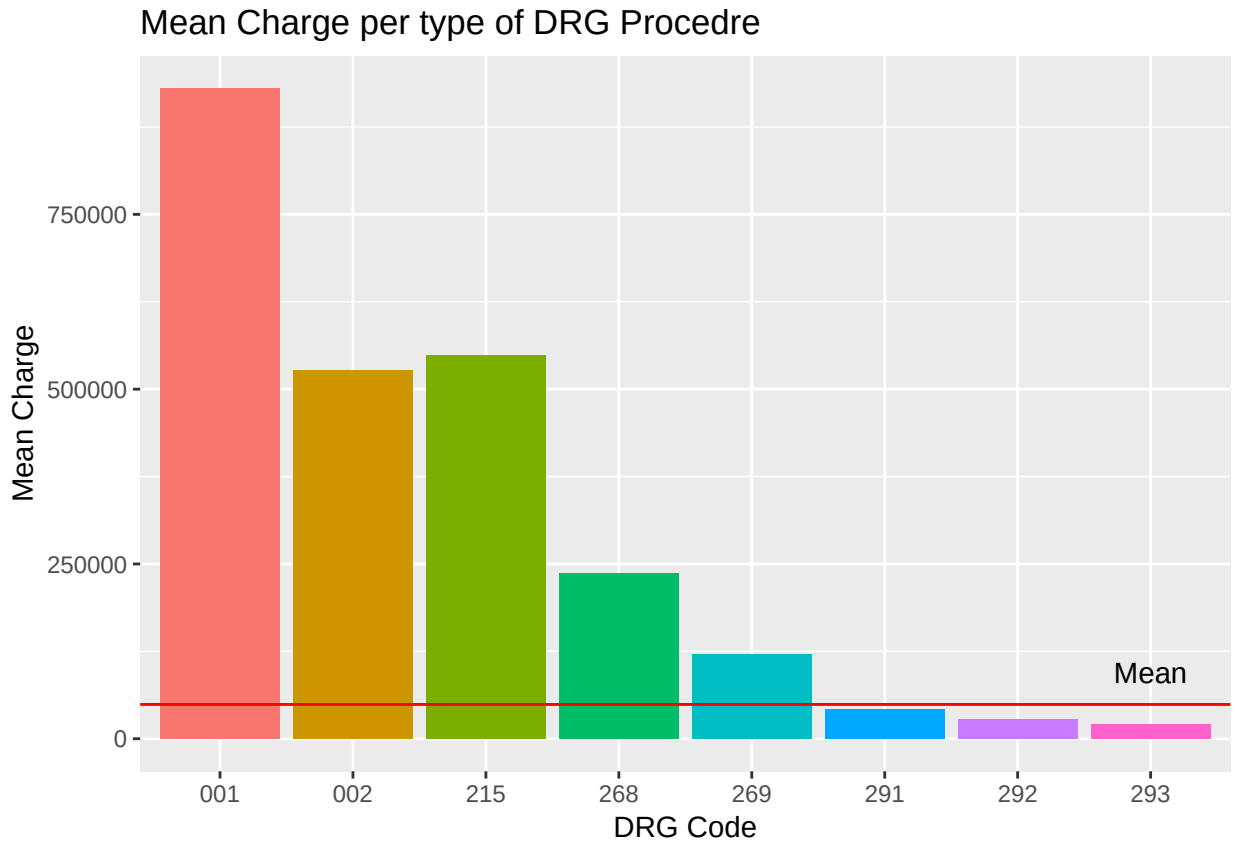
```r
# Visualizing the counts of each heart related DRG designated procedure.
heart_only %>%
  ggplot(aes(DRG_Code)) +
  geom_bar(aes(fill = DRG_Code), show.legend = FALSE)
```




```r
heart_only %>%
  group_by(DRG_Code) %>%
  summarise(mean_charge = mean(AvgCharge))
```

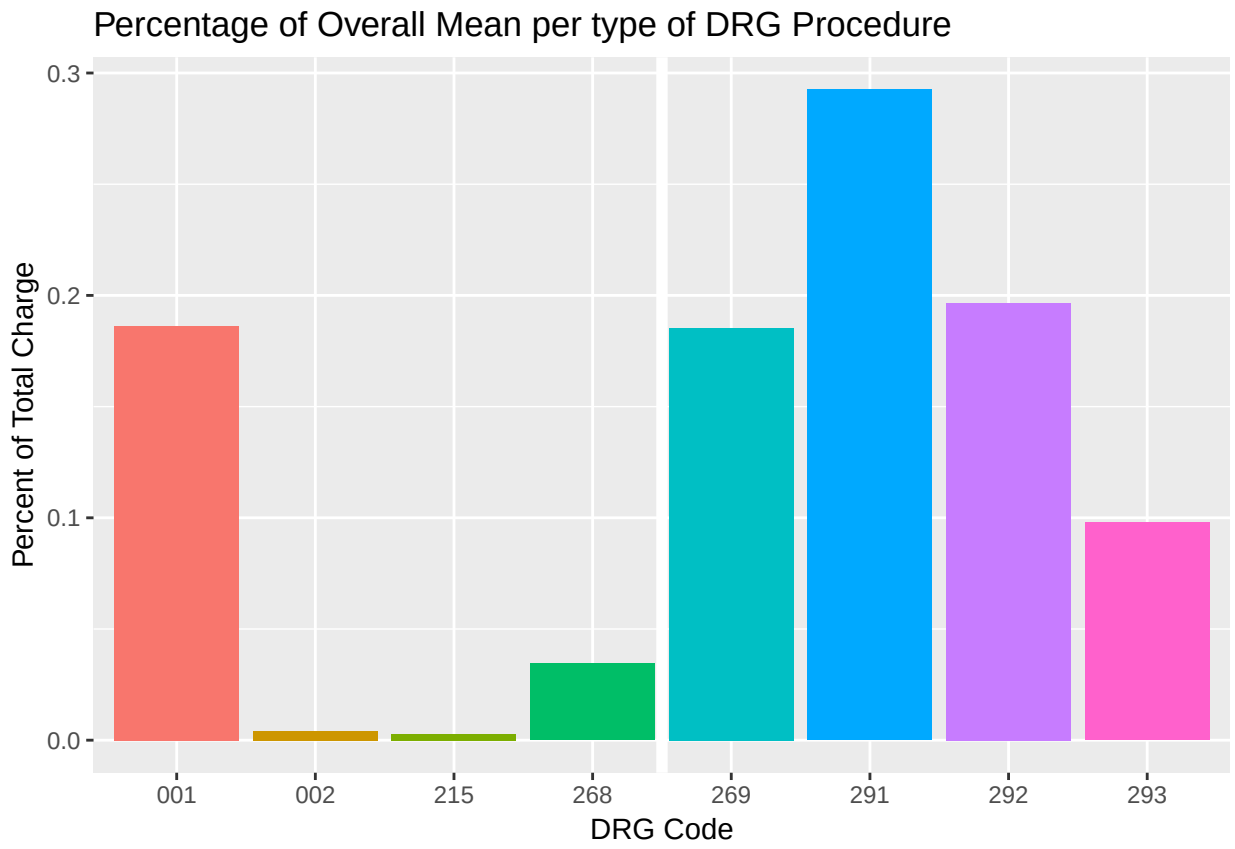
```
## # A tibble: 8 x 2
##   DRG_Code mean_charge
##   <chr>          <dbl>
## 1 001          930184.
## 2 002          526696.
## 3 215          549117.
## 4 268          237191.
## 5 269          120344.
## 6 291           42537.
## 7 292           28018.
## 8 293           21314.
```

```
# Visualizing the Mean Charge for a heart-related procedure by its DRG Code
heart_only %>%
  group_by(DRG_Code) %>%
  summarise(mean_charge = mean(AvgCharge)) %>%
  ggplot(aes(DRG_Code, mean_charge)) +
    geom_bar(stat = "identity", aes(fill = DRG_Code), show.legend = FALSE) +
    geom_hline(yintercept = mean(heart_only$AvgCharge), color = "red") +
    labs(title = "Mean Charge per type of DRG Procedre",x = "DRG Code", y = "Mean Charge") +
    annotate("text", max(heart_only$DRG_Code), mean(heart_only$AvgCharge), vjust = -1, label = "Mean")
```




```
# Visualizing the proportion of payments. Skew towards the less expensive procedures.
total_charge = sum(heart_only$AvgCharge)
heart_only %>%
  group_by(DRG_Code) %>%
  summarize(group_charge = sum(AvgCharge), perc_charge = group_charge / total_charge) %>%
```

```
  ggplot(aes(DRG_Code, perc_charge)) +
   geom_bar(stat = "identity", aes(fill = DRG_Code), show.legend = FALSE) +
   geom_ref_line(v = 4.5, size = 2)  +
   labs(title = "Percentage of Overall Mean per type of DRG Procedure", x = "DRG Code", y = "Percent o
```




```
filter(heart_only, DRG_Code %in% c("001", "002", "215", "268")) %>% select(DRG_Descr) %>% unique()
```

```
## # A tibble: 4 x 1
##   DRG_Descr
##   <chr>
## 1 HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W MCC
## 2 HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W/O MCC
## 3 OTHER HEART ASSIST SYSTEM IMPLANT
## 4 AORTIC AND HEART ASSIST PROCEDURES EXCEPT PULSATION BALLOON W MCC
```

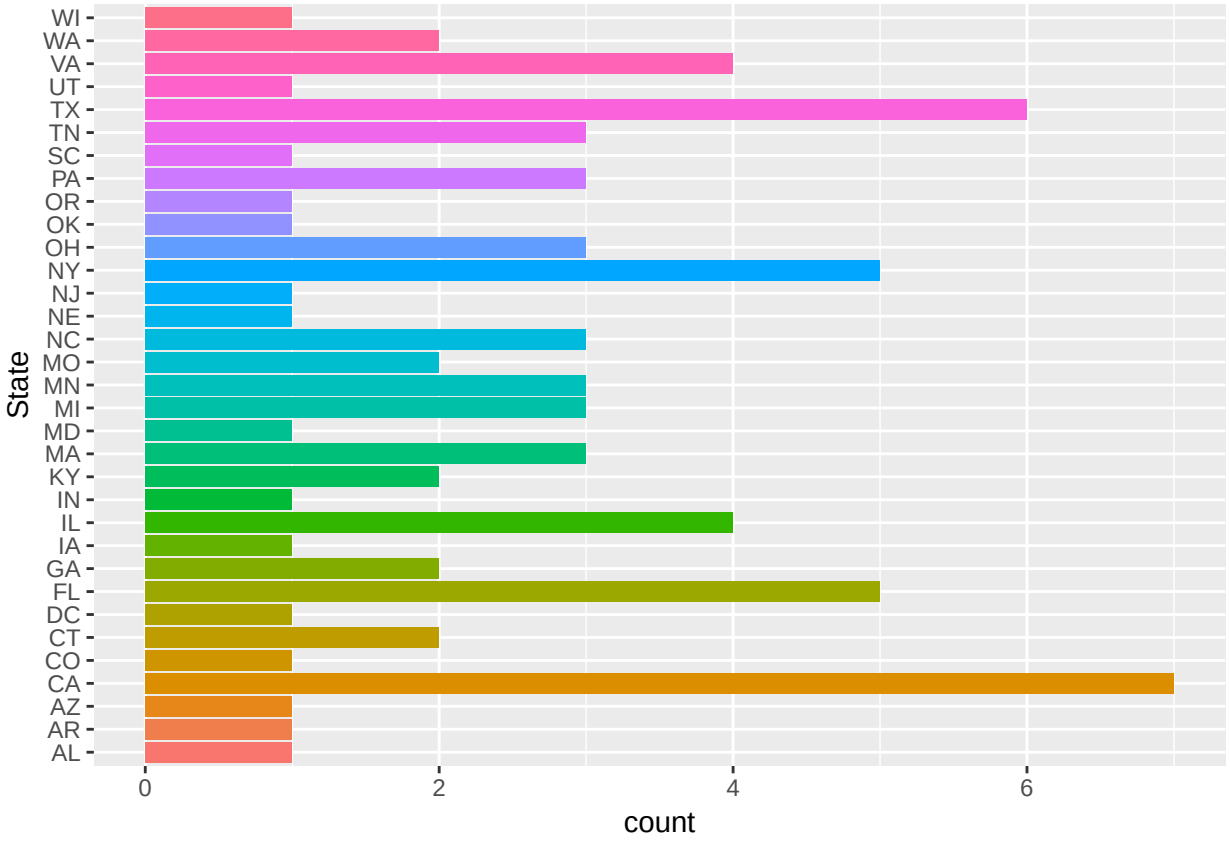
```
filter(heart_only, !(DRG_Code %in% c("001", "002", "215", "268"))) %>% select(DRG_Descr) %>% unique()
```

```
## # A tibble: 4 x 1
##   DRG_Descr
##   <chr>
## 1 AORTIC AND HEART ASSIST PROCEDURES EXCEPT PULSATION BALLOON W/O MCC
## 2 HEART FAILURE & SHOCK W MCC
## 3 HEART FAILURE & SHOCK W CC
## 4 HEART FAILURE & SHOCK W/O CC/MCC
```

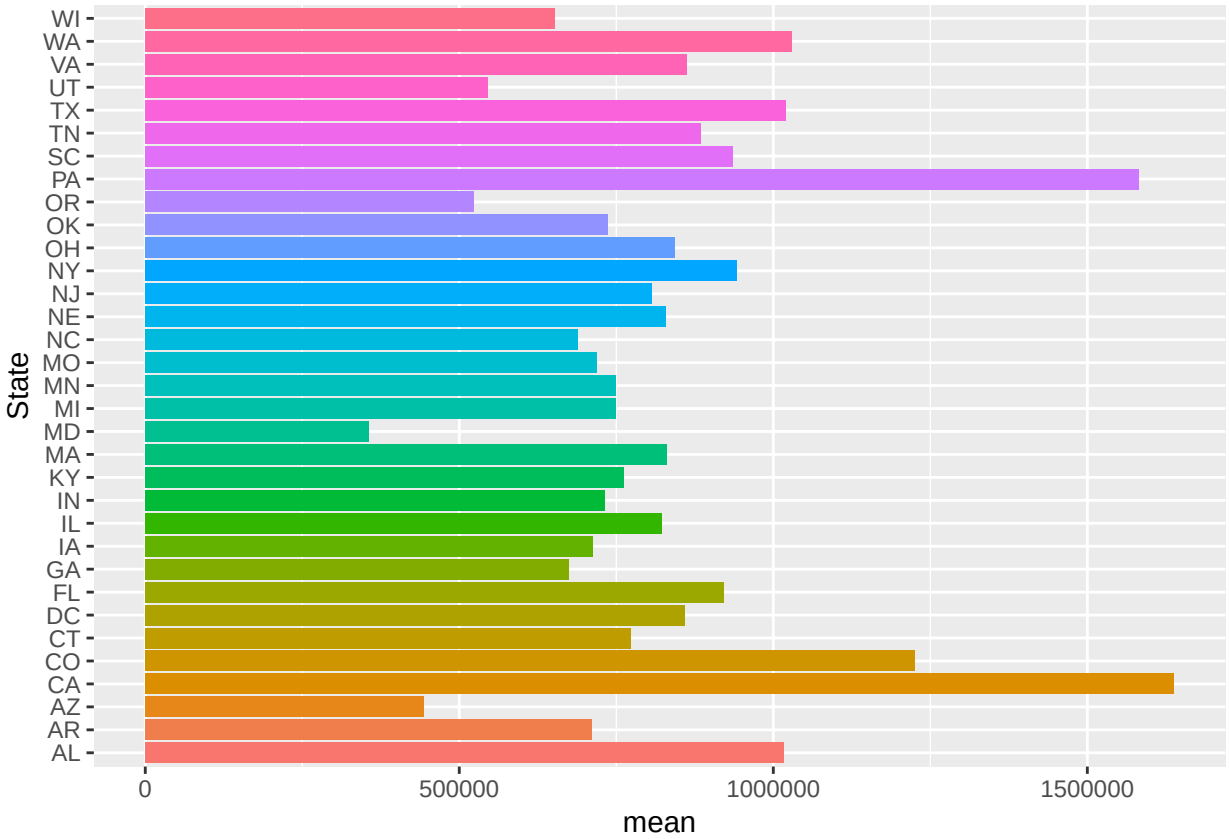
Looking at only the heart related procedures confirm my hypothesis. There are significantly more heart related procedures whose expense is significantly less than there are expensive ones. As a results, they

account for a larger proportion of the overall average charge, driving it down and explaining why it was unexpectedly low. The procedures designated by DRG Codes 001, 002, 215, and 268 correspond to an invasive procedure, be it a heart transplant, heart assist implant, or aortic assist procedure. The less expensive and more common procedures correspond to some variation of heart failure, thus not requiring surgery at the time of the initial diagnosis; meaning it could be some kind of initial consultation resulting in heart failure diagnosis and possibly needing an invasive procedure in the future depending on the gravity of the situation. One thing to note from this for future reference is that procedures designated as having a major complication or comorbidity (MCC) tend to be more expensive than their non-MCC counterpart, which makes sense. May be worth investigating this comparison for all types of procedures in the future as well as to which places in the country have more MCC procedures.

```
heart_only %>%
  filter(DRG_Code == "001") %>%
  ggplot(aes(State, fill = State)) +
    geom_bar(show.legend = FALSE) +
    coord_flip()
```




```
heart_only %>%
  filter(DRG_Code == "001") %>%
  group_by(State) %>%
  summarize(mean = mean(AvgCharge)) %>%
  ggplot(aes(State, mean, fill = State)) +
    geom_bar(stat = "identity", show.legend = FALSE) +
    coord_flip()
```

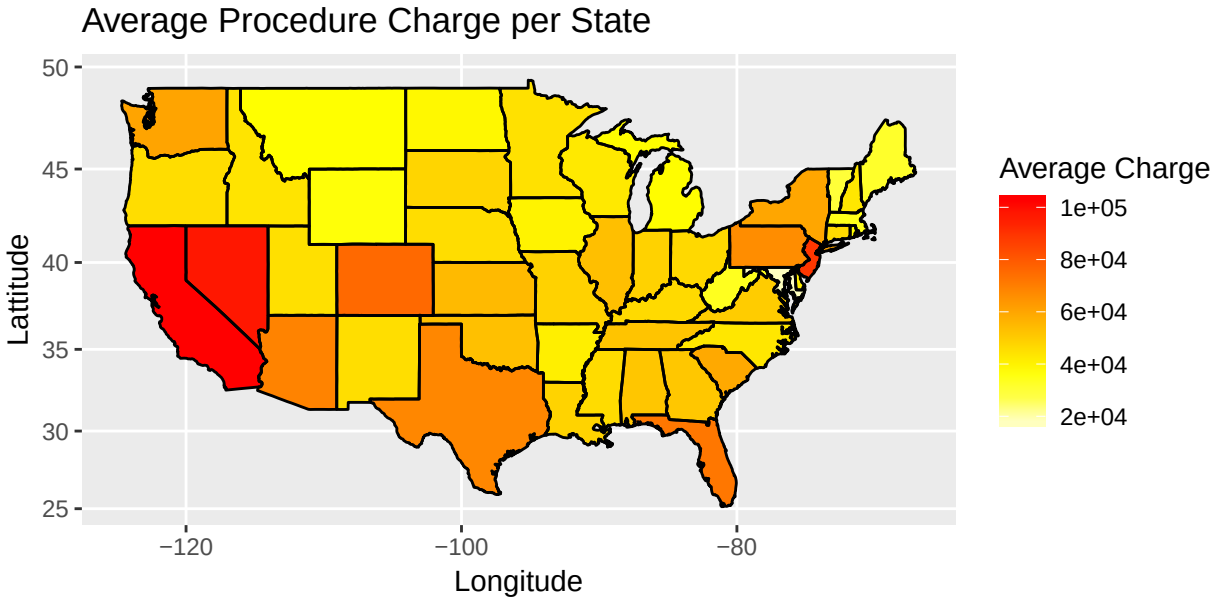


Out of curiosity, I decided to look into where most of the heart transplant with MCCs tend to occur. To no surprise, they are most commonly performed in the most populous states: California, Florida, Texas, and New York. Out of these four states, California has a significantly higher cost for that type of procedure whereas the other three states have a similar cost to other states, on average. I would think it is because California has a high cost of living compared to most other states, but New York has a similar cost of living. Surprisingly, Pennsylvania's average cost is comparable to California even though it is has about a quarter of the population and had about half of the procedure occur. The final thing to note, if you are in need in of a heart transplant or implant and have some sort of chronic disease that could complicate the surgery, go to Maryland. It still isn't cheap, but it is cheaper than mostly everywhere else. Arizona is comparable but it's too hot and dry out there and that is the last thing you need after a major operation.

```
# to create a map of the U.S, need the goegraphical coordinates. Loading in pre-built data
#with coordinates but run into a problem.the data frame containing the coordinates has the
#full state names whereas teh medicare data has abberviated state names so cant join directly

states <- map_data("state")

statenames.df <- bind_cols(tibble(state.abb), tibble(state.name))
statenames.df$state.name <- str_to_lower(statenames.df$state.name)

states <- statenames.df %>% left_join(states, by = c("state.name" = "region"))

charges <- med_data %>% group_by(State) %>% summarize(mean_charge = mean(AvgCharge))
map_df <- left_join(states, charges, by = c("state.abb" = "State"))


ggplot(map_df, aes(long, lat, group = group)) +
```

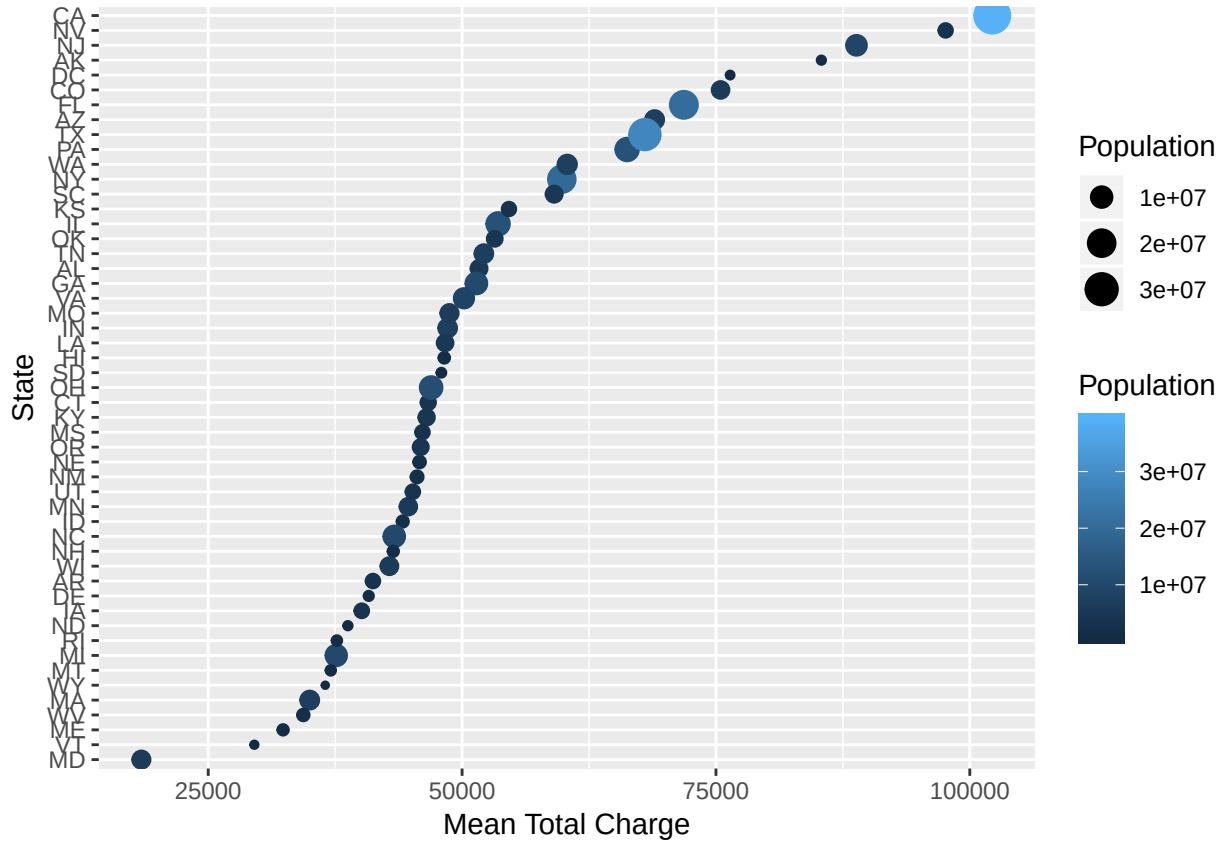
```
  geom_polygon(aes(fill = mean_charge)) +
  geom_path() +
  scale_fill_gradientn(colours=rev(heat.colors(10)),na.value="grey90")+
  coord_map() +
  labs(x = "Longitude", y = "Lattitude", title = "Average Procedure Charge per State", fill = "Average
```

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




```
# Plot ordering the states by their mean Total charge
# is there a correlation between mean total charge of a state and its population? (new jersey a possibl
# includes how population interacts with the average charge per state
med_data %>%
  group_by(State) %>%
  summarize(mean_charge = mean(AvgCharge)) %>%
  left_join(state_pop) %>%
  ggplot(aes(mean_charge, reorder(State, mean_charge))) +
    geom_point(aes(size = Population, color = Population)) +
    labs(x = "Mean Total Charge", y = "State")
```

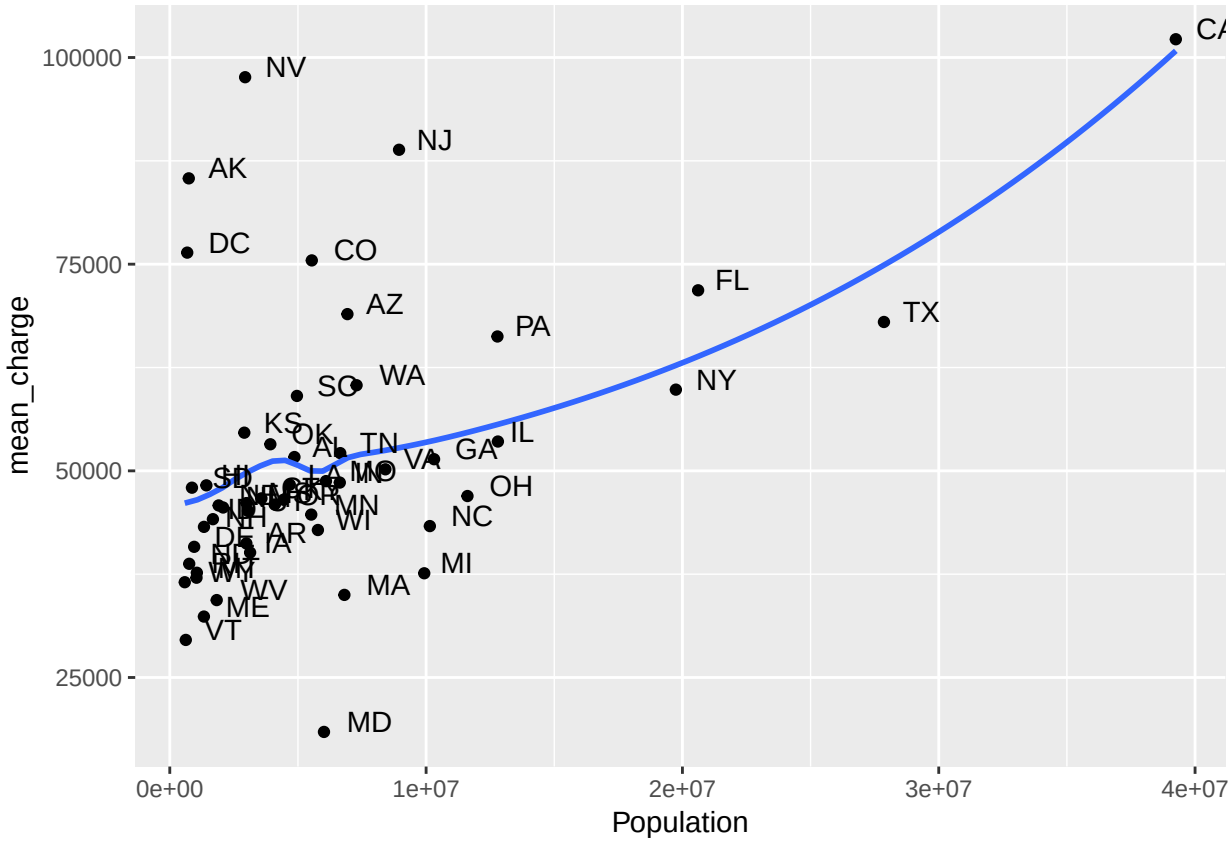
```
## Joining, by = "State"
```




```
# See if there is a trend/ relationship between charges and states population. Theres an increasing tre
med_data %>%
    group_by(State) %>%
    summarize(mean_charge = mean(AvgCharge)) %>%
    left_join(state_pop) %>%
    ggplot(aes(Population, mean_charge)) +
    geom_point() +
    geom_smooth(se = FALSE) +
    geom_text(aes(label = State), hjust = - 0.5, vjust = 0)
```

```
## Joining, by = "State"
```

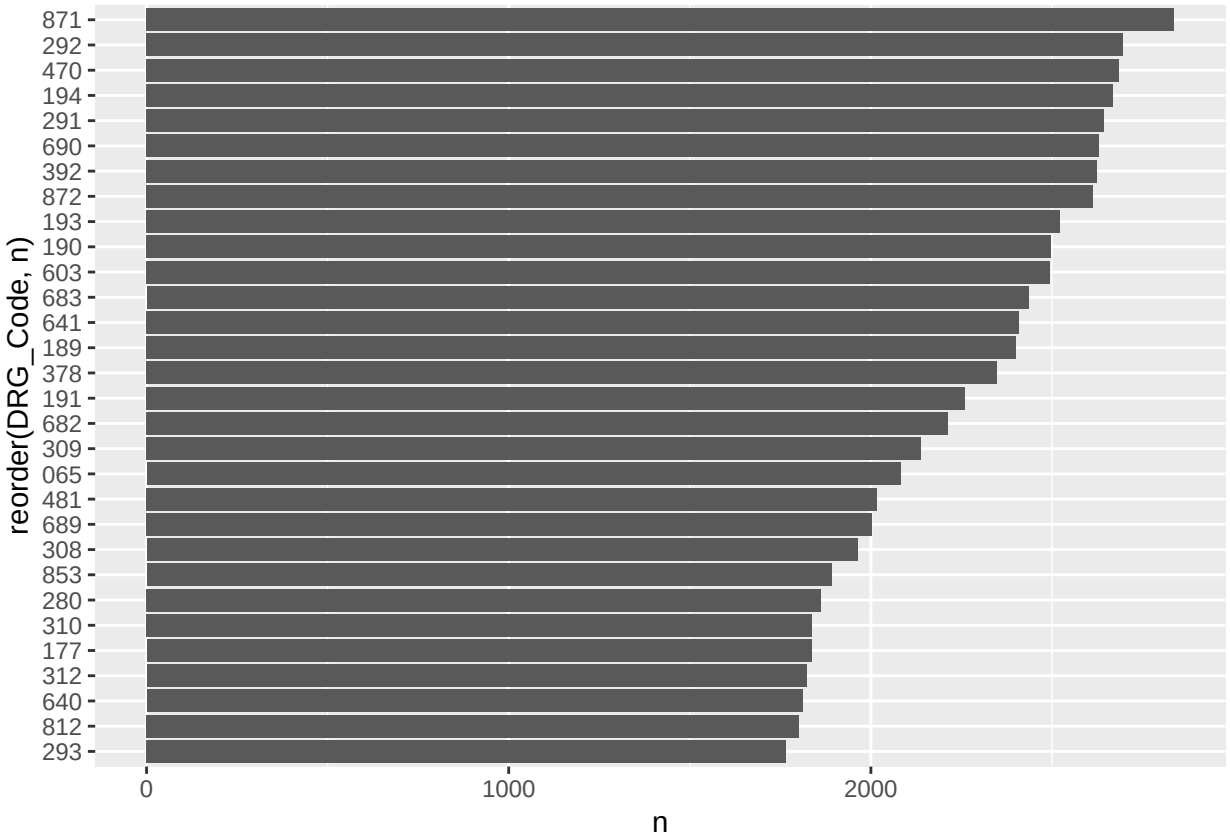
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#would like to standardize the average charge based on the number of Medicare beneficiaries in each sta
```

There may be a correlation between the average charge per procedure per state and its population, though it may not be strong. The most populous states have tend to charge higher per procedure while the least populous states charge less on average. After a little investigation, there is a positive trend between average charge and the population of a state. A better comparison would be between the average charge and the number of Medicare beneficiaries. It is noteworthy that Maryland charges the least on average by a significant margin compared to the other states. May be worth looking into why Maryland's Medicare procedures are so inexpensive compared to other states and see how it compares to the more expensive states, such as California. It may have to do with the services and procedures provided to the Medicare beneficiaries, meaning some of the more expensive procedures may be less prevalent in Maryland.

```
common_codes <- med_data %>%
  count(DRG_Code, DRG_Descr)


top_30codes <- common_codes %>%
  arrange(n) %>%
  tail(30)
top_30codes %>%
  ggplot(aes(reorder(DRG_Code, n), n)) +
    geom_bar(stat = "identity") +
    coord_flip()
```




```
med_data %>%
  mutate(MCC = str_detect(DRG_Descr, "MCC")) %>%
  filter(MCC == TRUE) %>%
  count(MCC)
```

```
## # A tibble: 1 x 2
##   MCC        n
##   <lgl>  <int>
## 1 TRUE  128380
```