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.

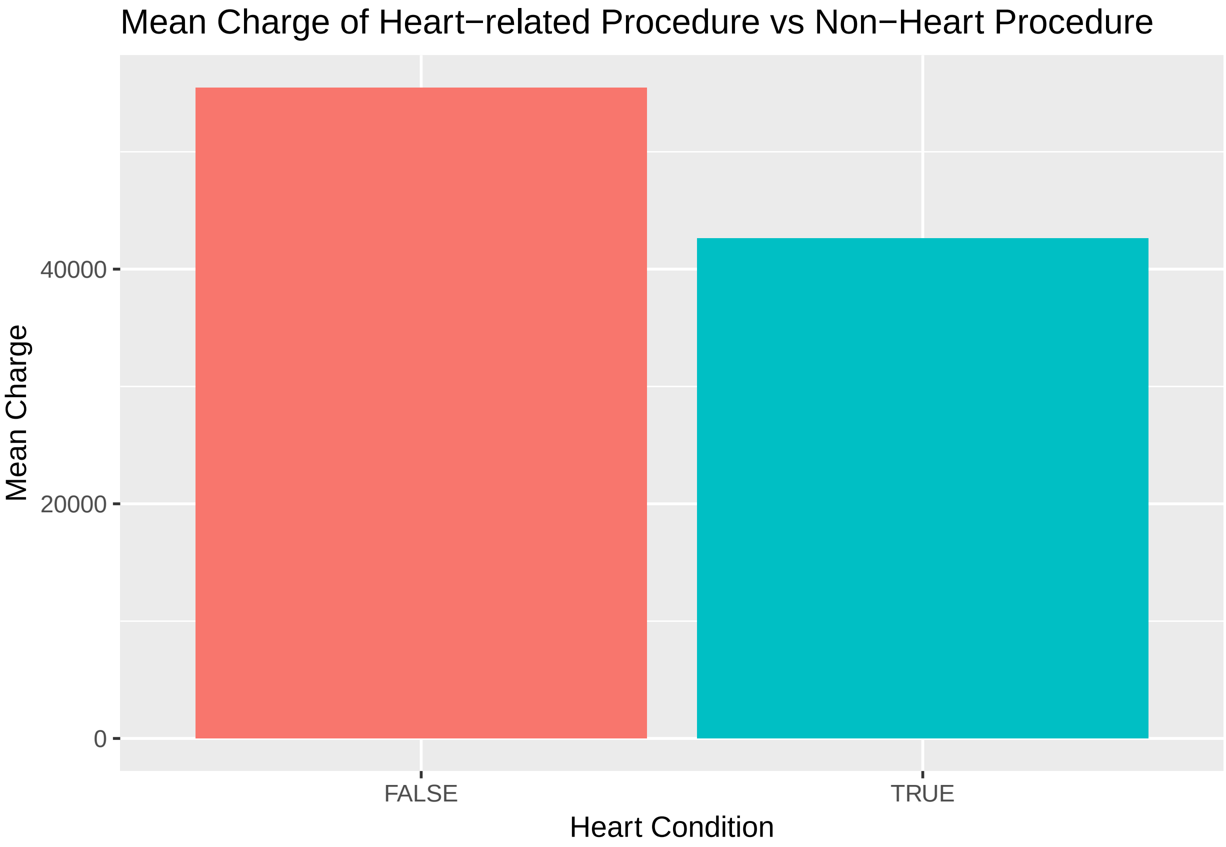
```
med_data <- med_data %>%
  separate(DRG, c("DRG_Code", "DRG_Descr"), sep = 3)

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

Since we are given the total number of discharges per hospital for each type of procedure, it may be beneficial to find the totals for each catefory. For example, it may work out better finding the total for the charges for a specific procedure and dividing by the total discharges, thereby giving a more accurate representation of the procedure's mean charge instead of taking the mean of average charges. Furthermore, seeing as the City and State in which the hospital is located, the HRR (Hospital Referral Region) seems to be redundant. The HRR columns seems to be just a string column combining the States and Cities.

```
med_data <- med_data %>% select(-HRR)

med_data <- med_data %>%
  mutate(TCharge = Discharges * AvgCharge, TotalPmts = Discharges * AvgTotalPmts, TMedPmts = Discharges
```

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 expensive than other types of procedures since they are typically very serious.

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


heart_only %>%
  group_by(Heart) %>%
  summarise(mean_charge = sum(TCharge) / sum(Discharges)) %>%
  ggplot(aes(Heart, mean_charge)) +
    geom_bar(stat = "identity", aes(fill = Heart), show.legend = FALSE) +
    labs(x = "Heart Condition", y = "Mean Charge", title = "Mean Charge of Heart-related Procedure vs N
```

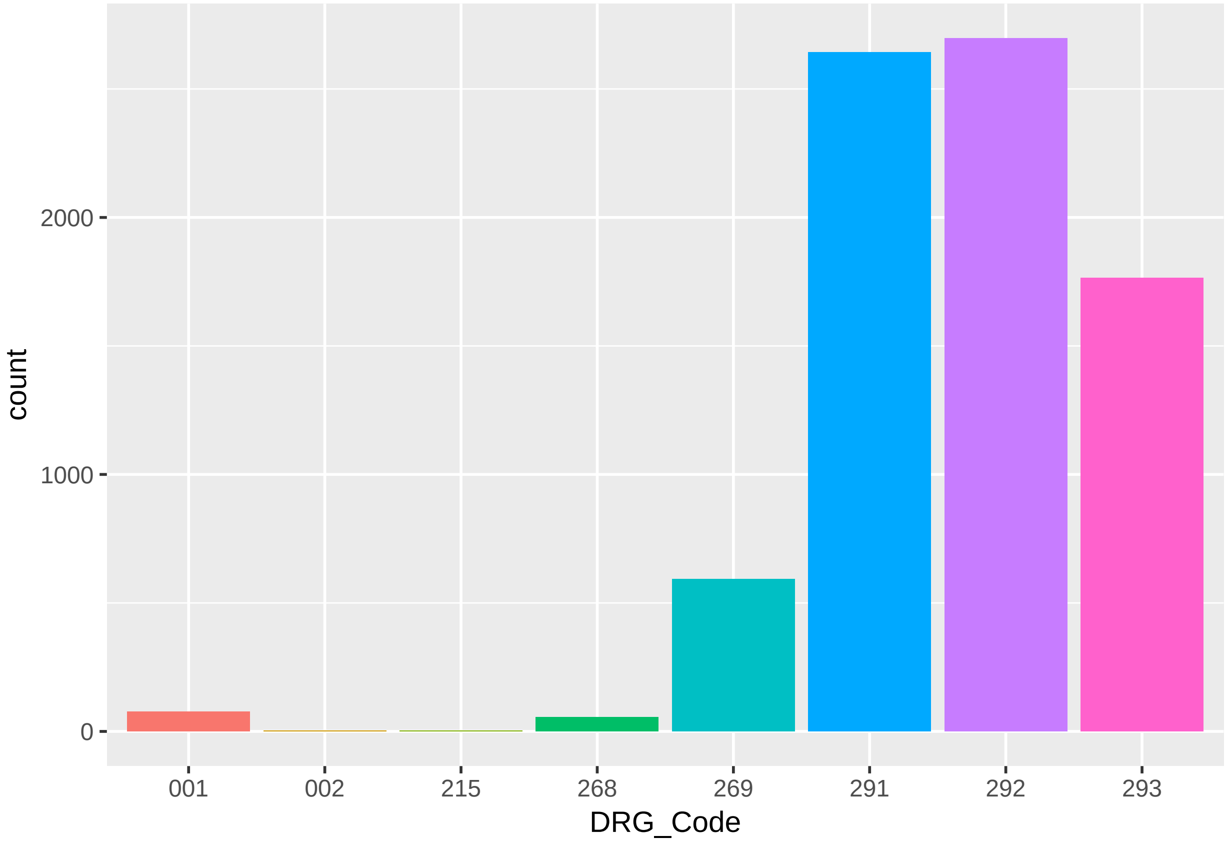


It turns out that Heart related procedures as a whole are not more expensive when compared to all others, which is a little unexpected. I would expect heart related procedures to be more expensive in general because it is a vital organ and any type of major procedures is sure to be invasive, causing 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.

```
heart_only <- heart_only %>%
  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
```

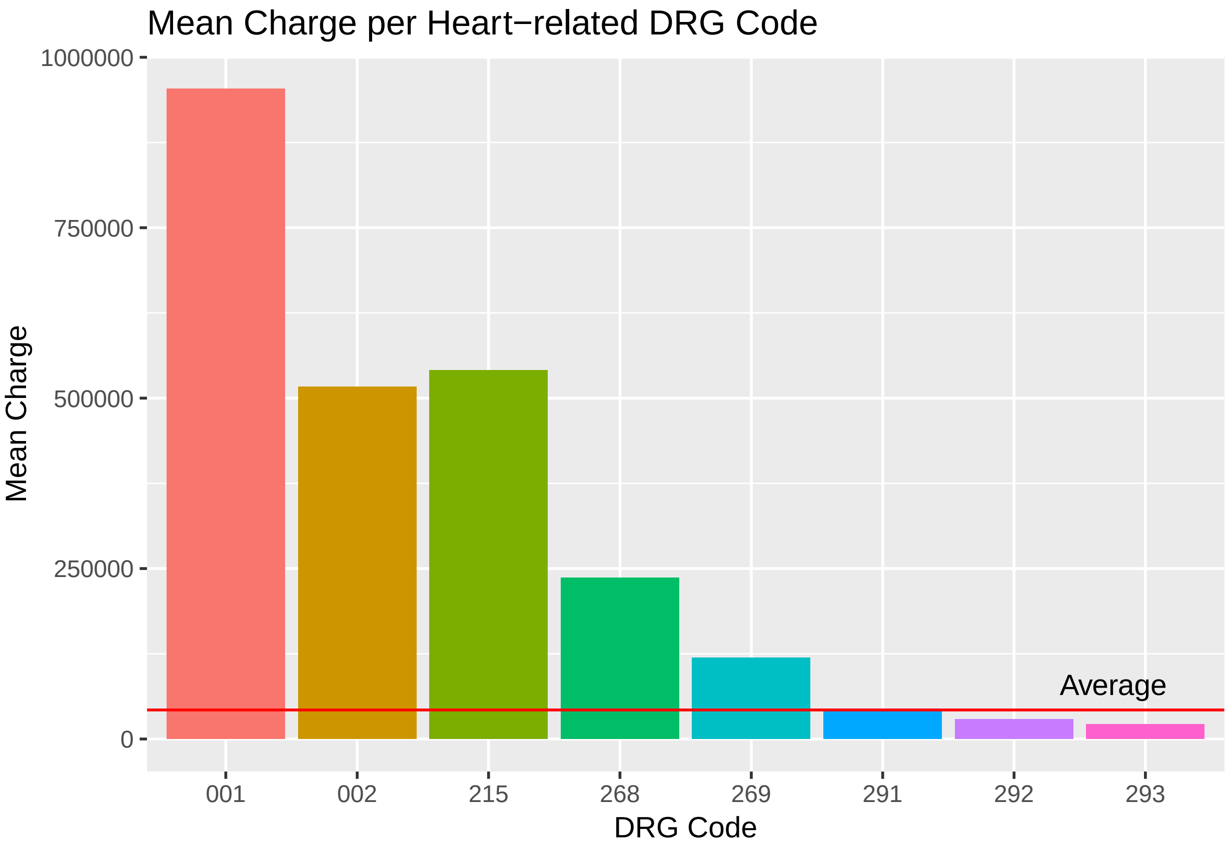
```
# Visualizing the counts of each heart related DRG designated procedure.
heart_only %>%
  ggplot(aes(DRG_Code)) +
  geom_bar(aes(fill = DRG_Code), show.legend = FALSE)
```



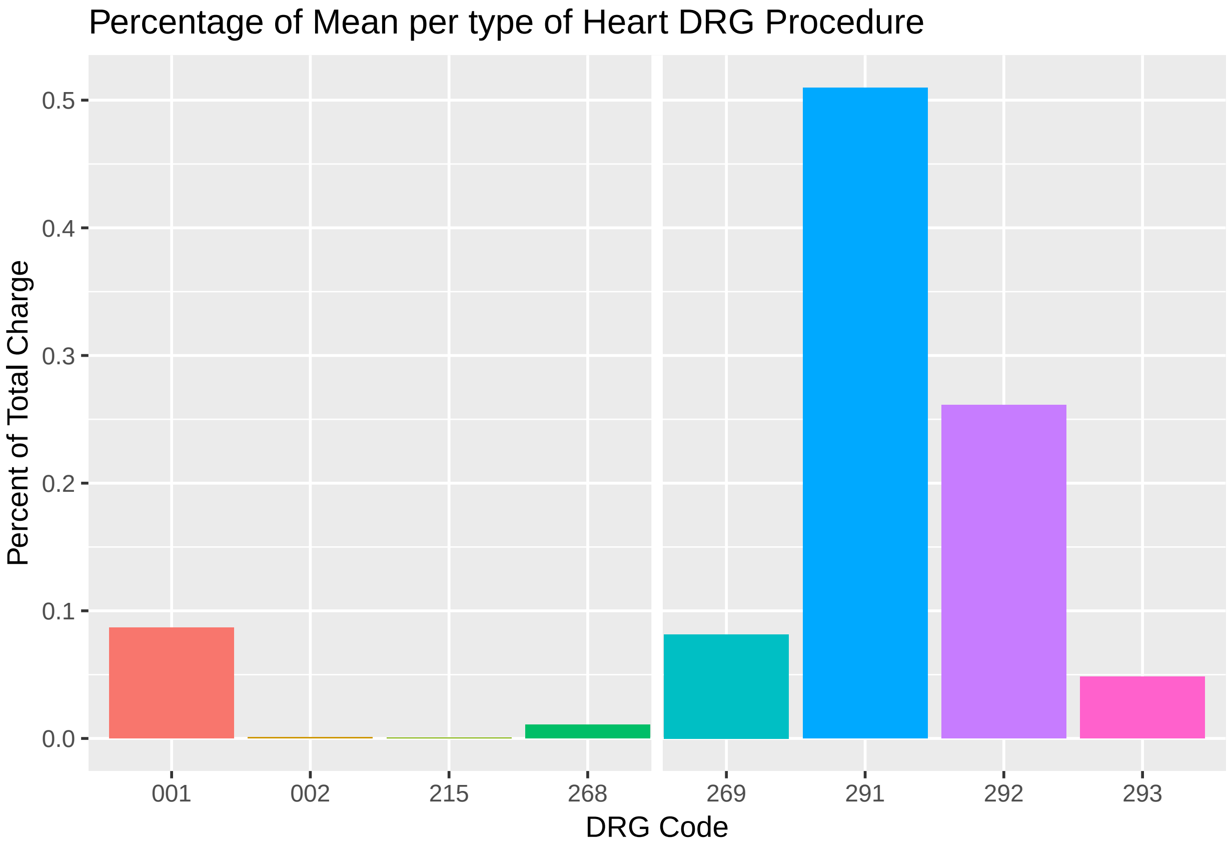
```
heart_only %>%
  group_by(DRG_Code) %>%
  summarise(mean_heart_charge = sum(TCharge) / sum(Discharges))
```

```
## # A tibble: 8 x 2
##   DRG_Code mean_heart_charge
##   <chr>                <dbl>
## 1 001                953994.
## 2 002                516430.
## 3 215                541144.
## 4 268                236304.
## 5 269                119000.
## 6 291                 44495.
## 7 292                 29247.
## 8 293                 21438.
```

```
# Visualizing the Mean Charge for a heart-related procedure by its DRG Code
heart_only %>%
  group_by(DRG_Code) %>%
  summarise(mean_heart_charge = sum(TCharge) / sum(Discharges)) %>%
  ungroup() %>%
  ggplot(aes(DRG_Code, mean_heart_charge)) +
    geom_bar(stat = "identity", aes(fill = DRG_Code), show.legend = FALSE) +
    geom_hline(yintercept = sum(heart_only$TCharge) / sum(heart_only$Discharges), color = "red") +
    labs(title = "Mean Charge per Heart-related DRG Code",x = "DRG Code", y = "Mean Charge") +
    annotate("text", max(heart_only$DRG_Code), mean(heart_only$AvgCharge), hjust = 0.8, vjust = -0.5, la
```



```
# Visualizing the proportion of payments. Skew towards the less expensive procedures.
total_charge = sum(heart_only$TCharge)
heart_only %>%
  group_by(DRG_Code) %>%
  summarize(group_charge = sum(TCharge), 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 Mean per type of Heart DRG Procedure", x = "DRG Code", y = "Percent of
```



```r
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
```

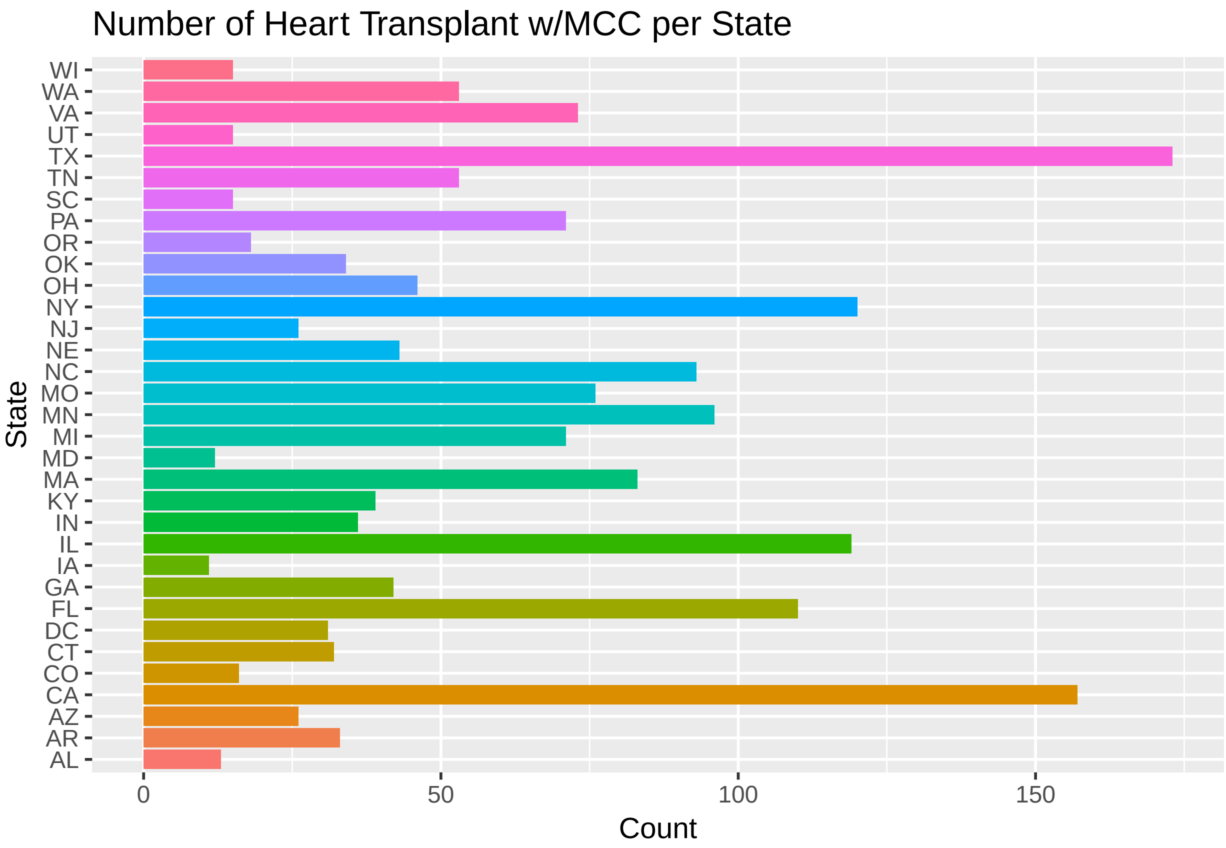
```r
filter(heart_only, !(DRG_Code %in% c("001", "002", "215", "268"))) %>% select(DRG_Descr) %>% unique()
```

```
## # 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
```

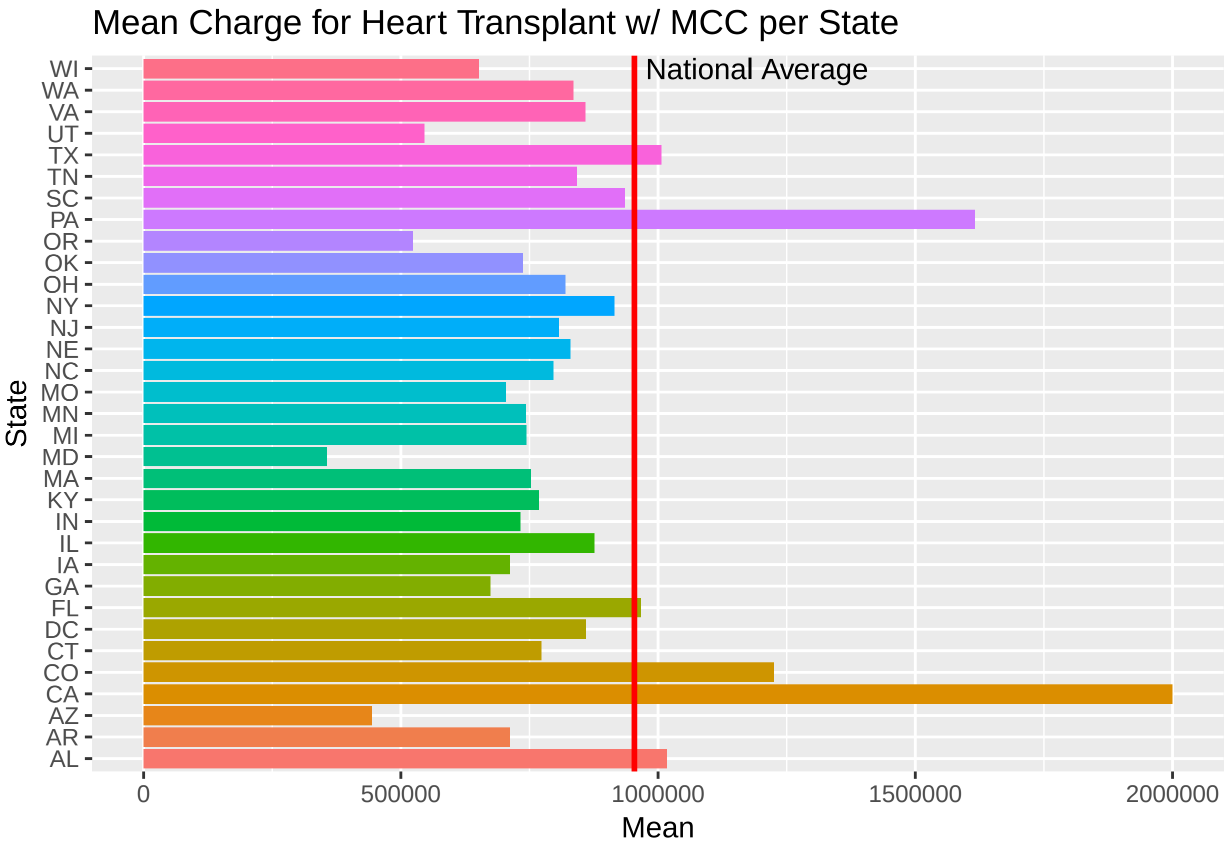
Looking at only the heart related procedures confirm my hypothesis. There are significantly more heart related procedures whose charges are less than the mean than there are expensice procedures/diagnosis. As a results, they account for a larger proportion of the overall average charge, driving it down 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. The most common diagnosis was DRG Code 291, which accounted for most about half of the total charge for heart-related diagnosis. This corresponded with having some form of Heart Failure or Shock with a major complication or comorbidity, including diagonses such as hypertensive heart diseases and systolic/diastolic heart failure. It makes sense that there are a lot more of these types of diagnoses as opposed to heart transplants since transplants are incredibly risky for the elderly and their hearts are more likely to begin failing given their age.

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. It is very likely this would turn out to be true, otherwise they wouldn't be called complications or treated separately from their non-MCC counterparts.

```
heart_only %>%
  filter(DRG_Code == "001") %>%
  group_by(State) %>%
  summarize(Disch = sum(Discharges)) %>%
  ggplot(aes(State, Disch, fill = State)) +
    geom_bar(stat = "Identity", show.legend = FALSE) +
    coord_flip() +
    labs(x = "State", y = "Count", title = "Number of Heart Transplant w/MCC per State")
```



```
heart_trans_mean = sum(heart_only[heart_only$DRG_Code=="001",]$TCharge)/sum(heart_only[heart_only$DRG_C
heart_only %>%
  filter(DRG_Code == "001") %>%
  group_by(State) %>%
  summarize(mean = sum(TCharge)/sum(Discharges)) %>%
  ggplot(aes(State, mean, fill = State)) +
    geom_bar(stat = "identity", show.legend = FALSE) +
    geom_ref_line(h = heart_trans_mean, colour = "red", size = 1) +
    labs(y = "Mean", title = "Mean Charge for Heart Transplant w/ MCC per State") +
    annotate("text", max(heart_only[heart_only$DRG_Code == "001",]$State), heart_trans_mean, hjust = -0
    coord_flip()
```



```
heart_only %>%
  filter(State == "TX", DRG_Code == "001") %>%
  group_by(City) %>%
  summarise(Discharges = sum(Discharges))
```

```
## # A tibble: 3 x 2
##   City        Discharges
##   <chr>            <dbl>
## 1 DALLAS              57
## 2 HOUSTON            101
## 3 SAN ANTONIO         15
```

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 in 2016: California, Florida, Illinois, Texas, and New York. Furthermore, Texas hospitals perform the most heart transplants across the nation, with California a close second. This makes sense since the Texas Medical Center in Houston is renown for its hospitals, including its Cardiology specialists. As such, I would expect most of the transplants in

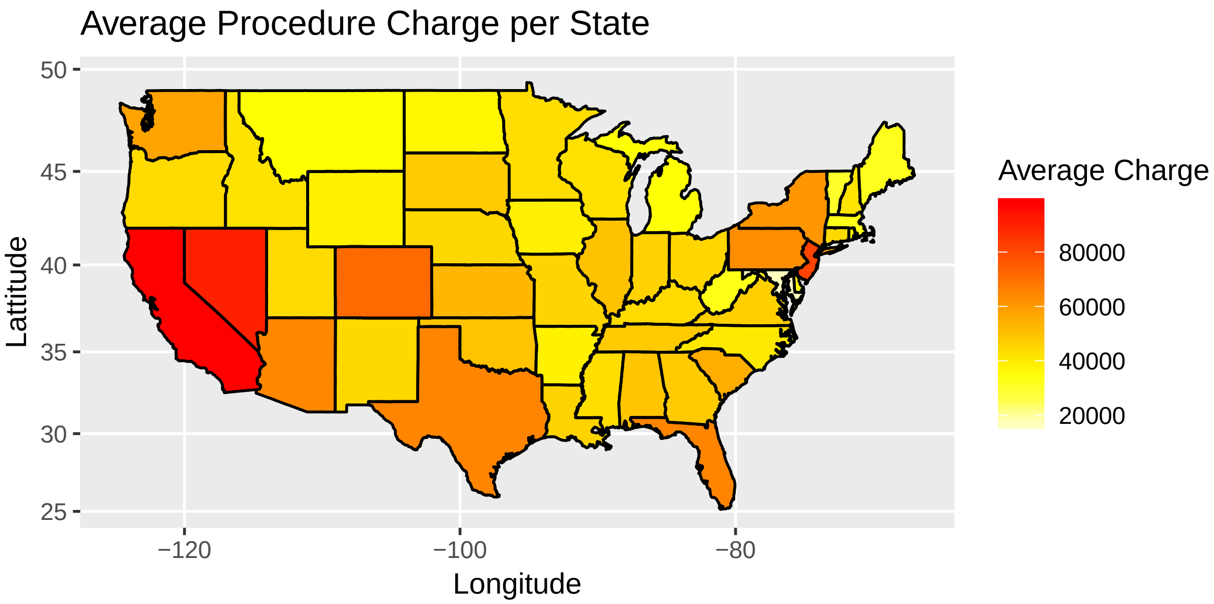
Texas to occur in Houston, which is confirmed by looking at the total number of patient discharges per city. While Texas has the most discharges and charge around the national average, California has a significantly higher cost for the procedure at nearly double the cost. I would think it is because California has a high cost of living compared to most other states, but New York has a similar cost of living and is below the national average. Wyoming, however, had no heart transplants with MCCs performed in any of its hospitals. 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.

Taking a step back from considering only heart-related diagonses, I would like to consider how all diagnoses are charged across the nation. To get a better sense if there is a geographical relationship with the average charge for a procedure, I want to plot a heat map of the United States. This will allow me to visualize if, for example, Medicare services and procedures are cheaper in the MidWestern states as opposed to Northeastern states.

The states whose hospitals charge the most per Medicare service on average seem to correspond with the most populous states, similar to what we saw in the heart related diagnoses, save for Nevada and DC. DC's high average charge makes sense because of its high cost of living, thanks to it being the home of our political institutions. On the other hand, Nevada's high average charge doesn't make much sense initially; I suspect it is a result of its population being concentrated near Las Vegas.

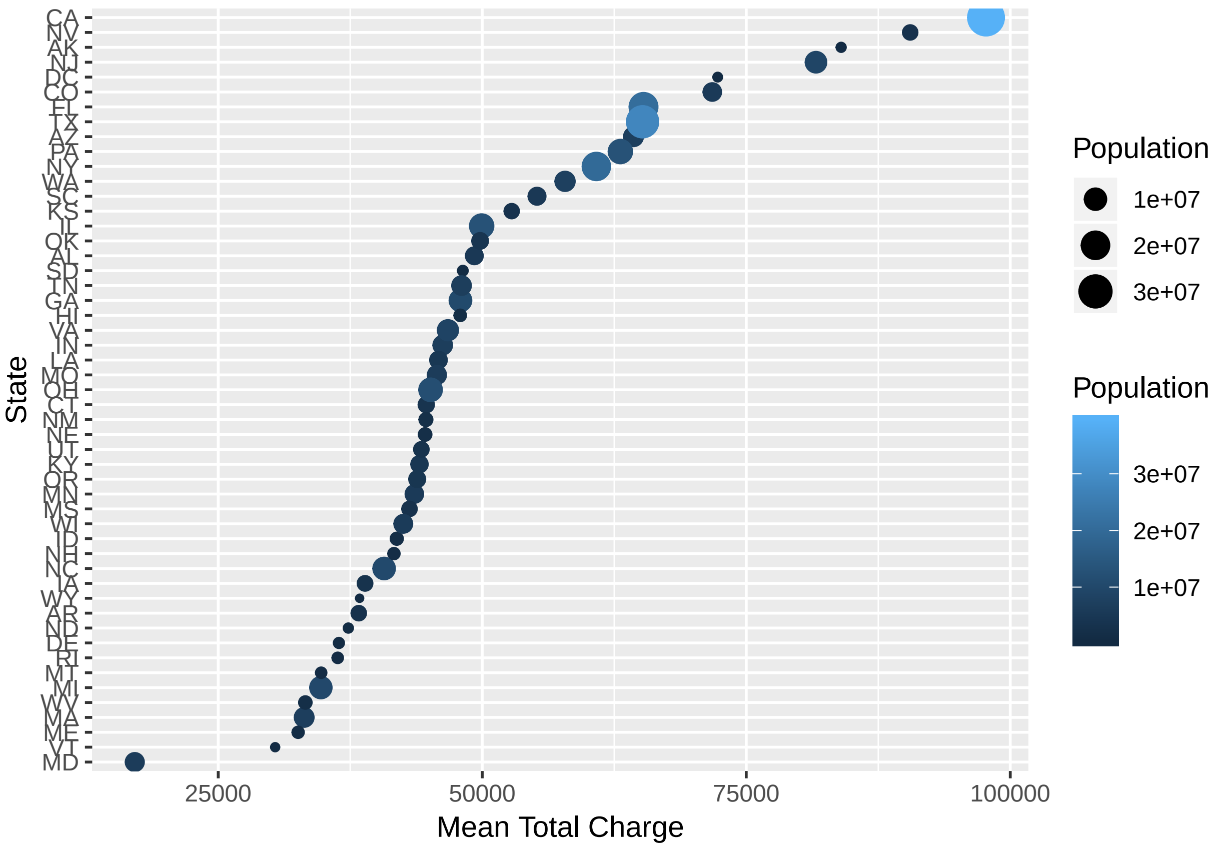
```r
# to create a map of the U.S, need the goegraphical coordinates. Loading in pre-built data
#with coordinates but run into a problem.the data frame containing the coordinates has the
#full state names whereas teh medicare data has abberviated state names so cant join directly

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 = sum(TCharge)/sum(Discharges))
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).
```



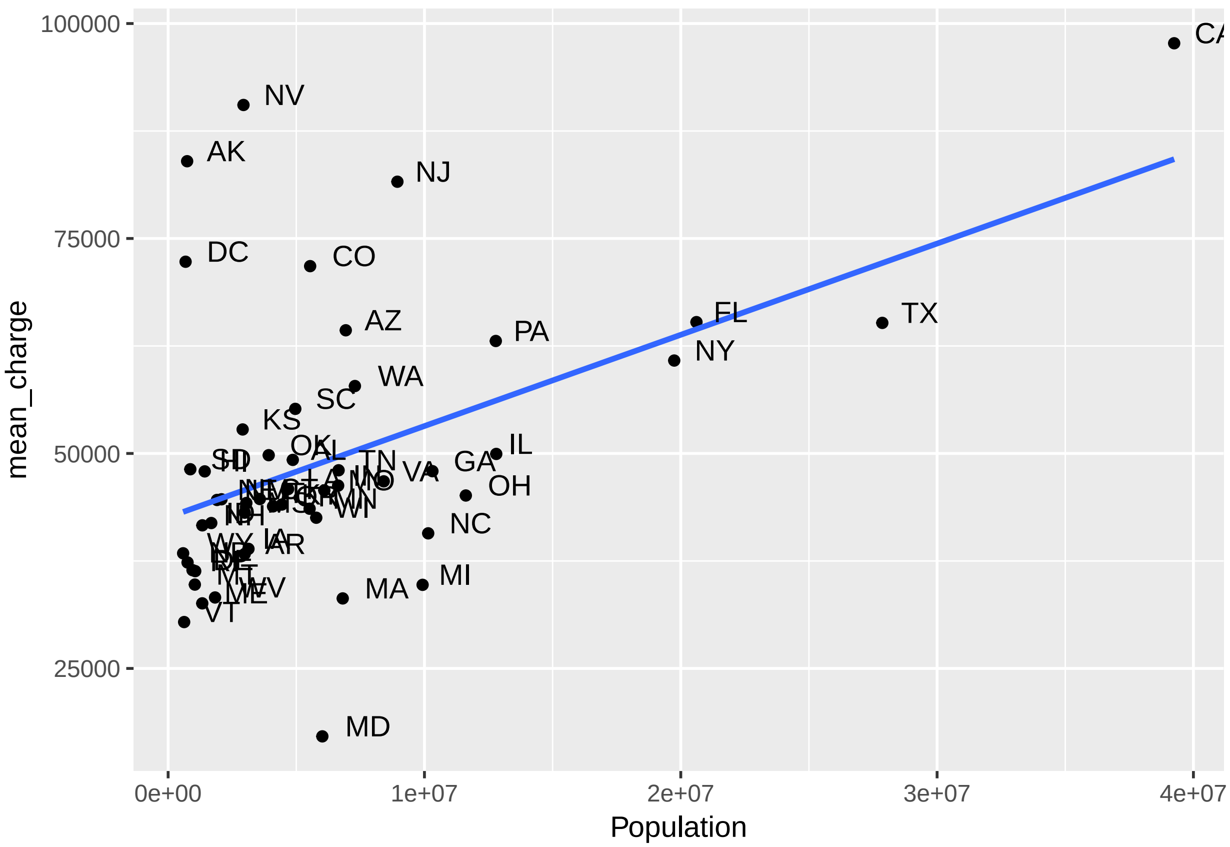
```r
# med_data %>% group_by(State) %>% summarise(sum = sum(Discharges)) %>% left_join(state_pop) %>% ggplot


# may have tp fix everything, ie taking the means of the average charge. instead find total charge per


# 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 = sum(TCharge)/sum(Discharges)) %>%
  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 = sum(TCharge)/sum(Discharges)) %>%
    left_join(state_pop) %>%
    ggplot(aes(Population, mean_charge)) +
    geom_point() +
    geom_smooth(se = FALSE, method = "lm") +
    geom_text(aes(label = State), hjust = - 0.5, vjust = 0)
```

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

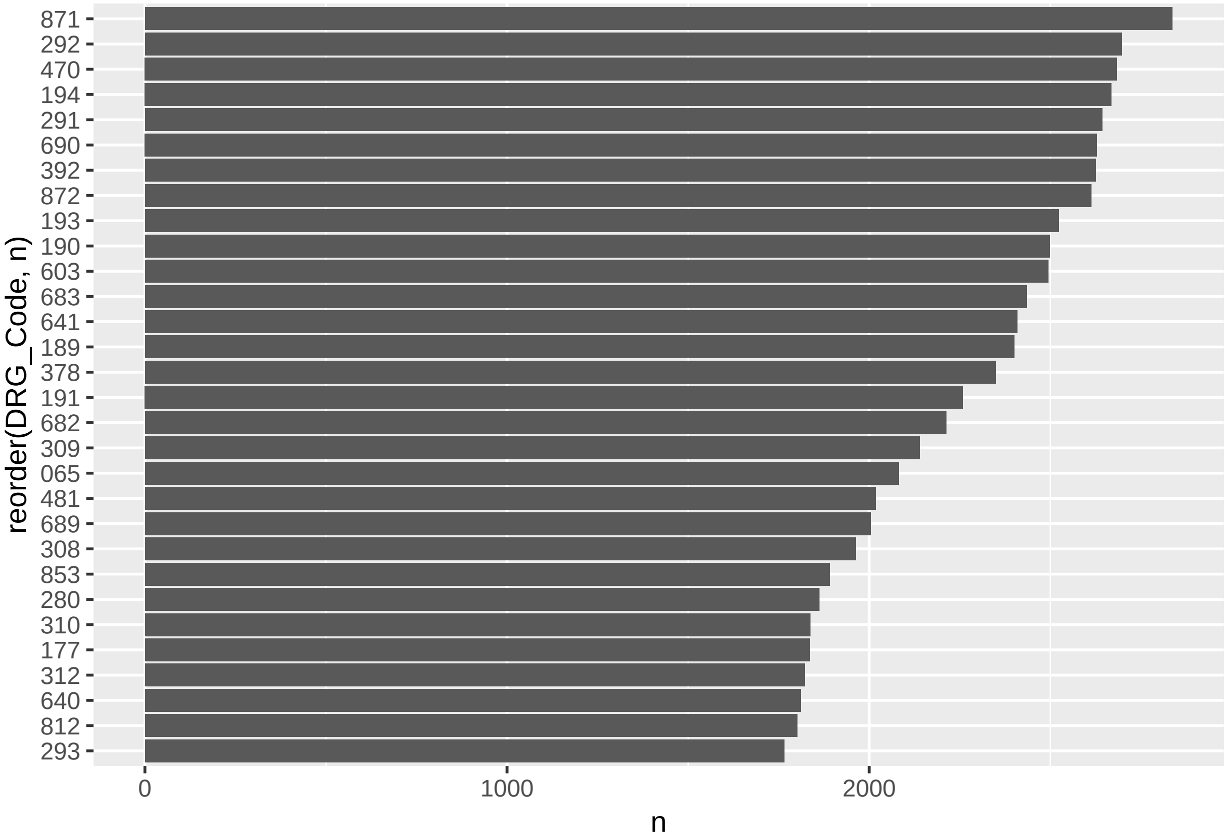


```
#would like to standardize the average charge based on the number of Medicare beneficiaries in each sta
```

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
```