EDA of Medicare Inpatient Charge Data

Anand Raman

12/9/2018

Loading Packages

```
library(tidyverse)
library(ggstance)
library(usmap)
```

Datasets

For this project I used data from the Centers for Medicare and Medicaid Services. Both datasets fall under the category of Medicare and Provider Utilization Payment Data: Inpatient. This means the charge data is entirely composed of discharges and billings for inpatient treatments and diagnoses (heart transplants, septicemia treatments, joint replacements etc.) Specifically, I use data from 2015 and 2016 – the most recently published data. I have merged the two datasets to enable comparison. Details of the merging process can be found in cleaning_merging.R. All details regarding variables contained in the dataset can be found in the Methodology document. This is supplemented by my own codebook which is a simple excel workbook.

Brief note on terminology: In this data, the abbreviations CC and MCC stand for "complication or comorbidity" and "major complication or comorbidity" respectively. Comorbidity is the presence of two or more chronic conditions in a patient.

Reading in Data

```
med <- read_csv("data/processed/medicare_inpatient_15_16.csv")
```

Analysis

Most Common Inpatient Charges

I will display the most commonly diagnosed issues across both years, taking the sum of total number of discharged patients across years 2015 and 2016.

```
med %>%
  gather(tot_discharges_15,
         tot_discharges_16,
         key = tot_by_yr,
         value = tot_discharges) %>%
  group_by(drg_definition) %>%
  summarise(tot_discharges = sum(tot_discharges)) %>%
  arrange(desc(tot_discharges)) %>%
  slice(1:10)
```

```
## # A tibble: 10 x 2
##    drg_definition                                              tot_discharges
##    <chr>                                                                <int>
##  1 470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER ~         860198
##  2 291 - HEART FAILURE & SHOCK W MCC                                401477
##  3 292 - HEART FAILURE & SHOCK W CC                                 329721
##  4 872 - SEPTICEMIA OR SEVERE SEPSIS W/O MV >96 HOURS W/O ~         266111
##  5 690 - KIDNEY & URINARY TRACT INFECTIONS W/O MCC                  260930
##  6 683 - RENAL FAILURE W CC                                         254155
##  7 193 - SIMPLE PNEUMONIA & PLEURISY W MCC                          248957
##  8 194 - SIMPLE PNEUMONIA & PLEURISY W CC                           242524
##  9 190 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE W MCC                241138
## 10 189 - PULMONARY EDEMA & RESPIRATORY FAILURE                      239947
```

The next tibble splits the counts of each charge across years.

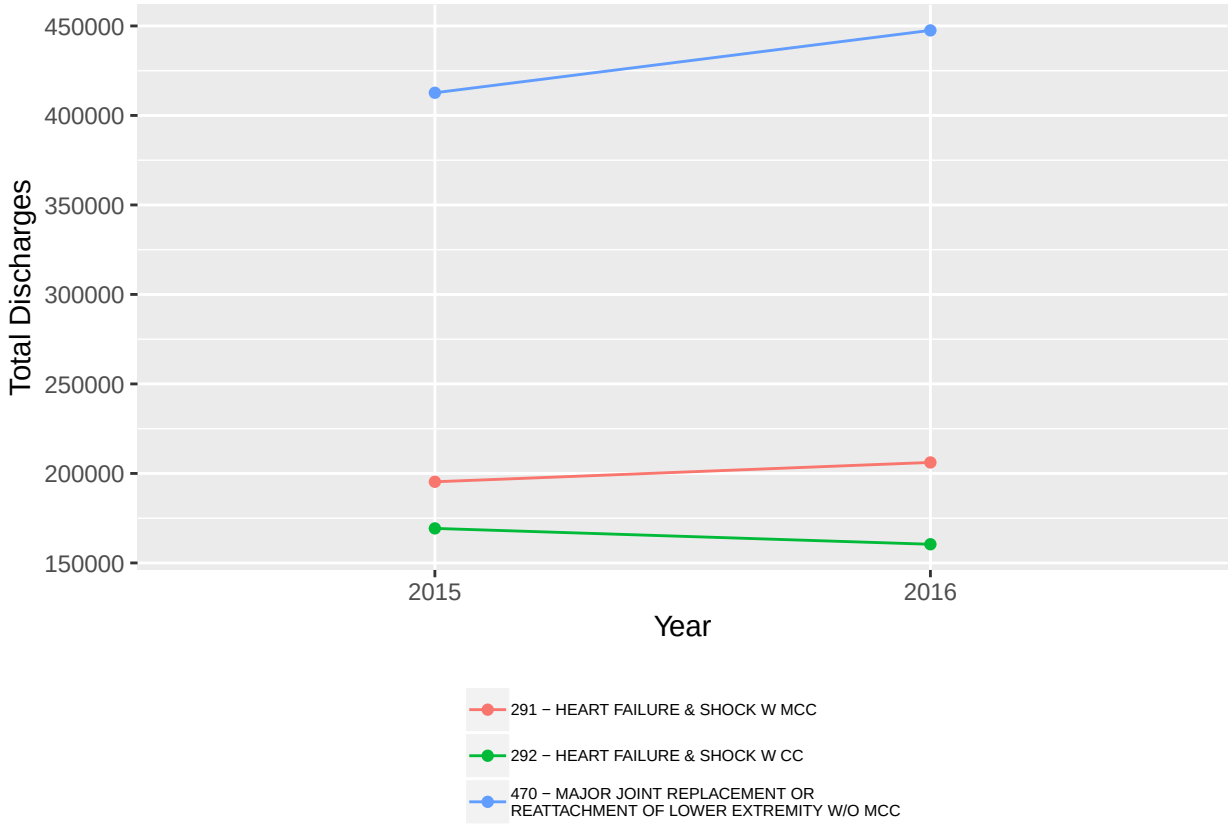
```
med %>%
  group_by(drg_definition) %>%
  summarise( discharges_2015 = sum(tot_discharges_15),
             discharges_2016 = sum(tot_discharges_16)) %>%
  arrange(desc(discharges_2016, discharges_2015))
```

```
## # A tibble: 418 x 3
##    drg_definition                           discharges_2015 discharges_2016
##    <chr>                                              <int>           <int>
##  1 470 - MAJOR JOINT REPLACEMENT OR REATT~          412675          447523
##  2 291 - HEART FAILURE & SHOCK W MCC                195331          206146
##  3 292 - HEART FAILURE & SHOCK W CC                 169292          160429
##  4 872 - SEPTICEMIA OR SEVERE SEPSIS W/O ~          132481          133630
##  5 690 - KIDNEY & URINARY TRACT INFECTION~          132831          128099
##  6 683 - RENAL FAILURE W CC                         129094          125061
##  7 189 - PULMONARY EDEMA & RESPIRATORY FA~          115492          124455
##  8 193 - SIMPLE PNEUMONIA & PLEURISY W MCC          127270          121687
##  9 190 - CHRONIC OBSTRUCTIVE PULMONARY DI~          124407          116731
## 10 194 - SIMPLE PNEUMONIA & PLEURISY W CC           131095          111429
## # ... with 408 more rows
```

The next analysis focuses on the three most commonly diagnosed issues: Major joint replacement or reattachment of lower extremity w/o MCC, heart failure and shock with MCC, and heart failure and shock with CC. First, I select the rows I am interested in and then I plot the trends in the top 3 most common procedures.

```
top_3 <- med %>% gather(tot_discharges_15,
                tot_discharges_16,
                key = tot_by_yr,
                value = tot_discharges) %>%
  mutate(year = case_when(
    tot_by_yr == "tot_discharges_16" ~ "2016",
    tot_by_yr == "tot_discharges_15" ~ "2015"
  )) %>%
  group_by(drg_definition, year) %>%
  summarise(tot_discharges = sum(tot_discharges)) %>%
  arrange(desc(tot_discharges))

top_3 <- top_3[1:6, ]
```

```
top_3 %>%
  ggplot(aes(x = year, y = tot_discharges, color = str_wrap(drg_definition, 40))) +
  geom_point() +
  geom_line(aes(group = drg_definition)) +
  theme(legend.position = "bottom", legend.direction = "vertical",
        legend.text=element_text(size=6), legend.title = element_blank()) +
  xlab("Year") +
  ylab("Total Discharges")
```



Between 2015 and 2016 there was an increase in the number of major joint replacements and heart failure with MCC, but a decrease in the number of heart failures with CC.

One way to further contextualize data on the change in the number of diagnoses of a certain condition is to describe the percent change in occurrences of conditions. In this, case I filter for conditions which have over 25000 diagnoses to improve the stability of estimates and to assess the most common of medical conditions in the elderly.

```
med %>%
  group_by(drg_definition) %>%
  filter(sum(tot_discharges_16) >= 25000) %>%
  summarise(percent_chng = (sum(tot_discharges_16)/sum(tot_discharges_15)-1) * 100,
            discharges_16 = sum(tot_discharges_16)) %>%
  arrange(desc(percent_chng))
```

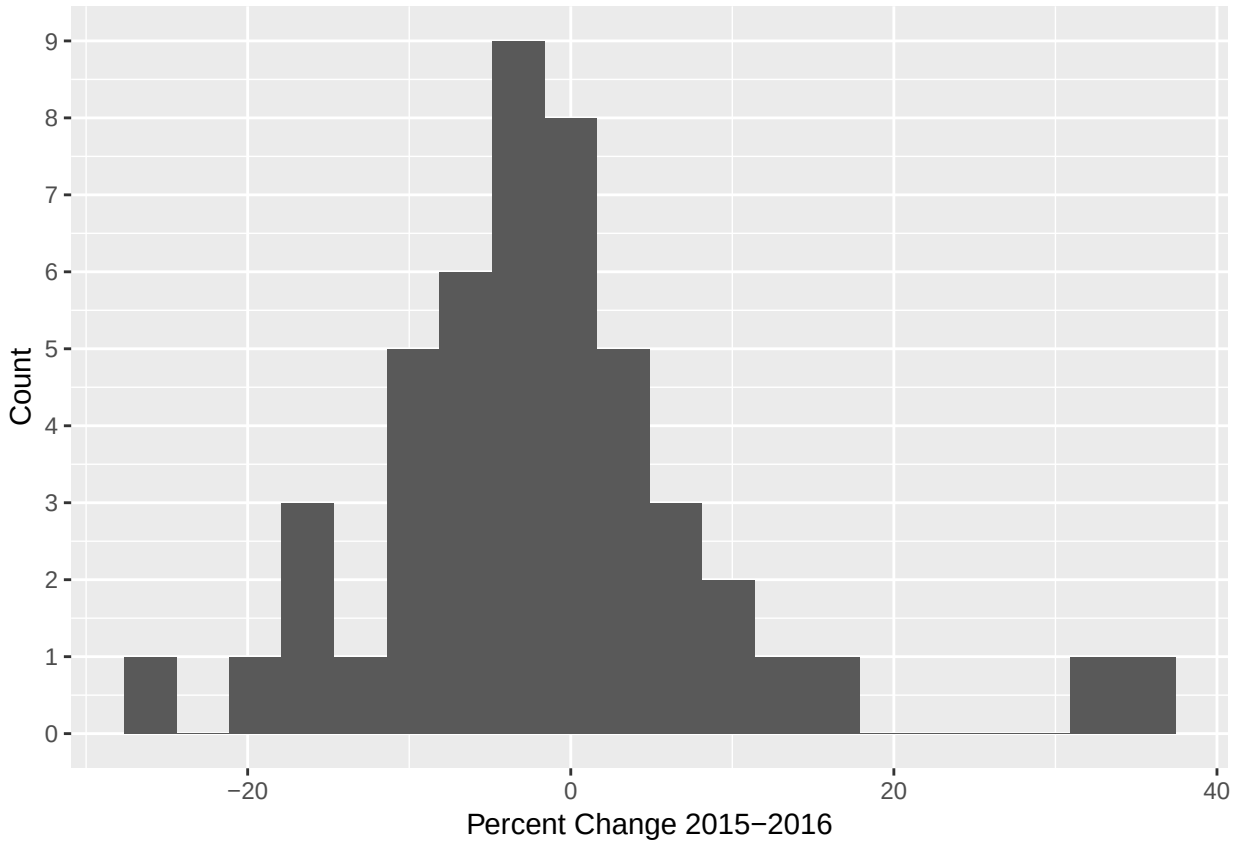
```
## # A tibble: 48 x 3
##    drg_definition                                percent_chng discharges_16
##    <chr>                                                <dbl>         <int>
##  1 853 - INFECTIOUS & PARASITIC DISEASES W O.R~         35.5          74014
```

```
##  2 252 - OTHER VASCULAR PROCEDURES W MCC               31.4          30047
##  3 483 - MAJOR JOINT/LIMB REATTACHMENT PROCEDU~        17.1          42693
##  4 330 - MAJOR SMALL & LARGE BOWEL PROCEDURES ~        12.8          40900
##  5 065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL I~         9.94         96024
##  6 470 - MAJOR JOINT REPLACEMENT OR REATTACHME~         8.44        447523
##  7 189 - PULMONARY EDEMA & RESPIRATORY FAILURE          7.76        124455
##  8 246 - PERC CARDIOVASC PROC W DRUG-ELUTING S~         7.75         28816
##  9 291 - HEART FAILURE & SHOCK W MCC                    5.54        206146
## 10 682 - RENAL FAILURE W MCC                            4.37        106412
## # ... with 38 more rows
```

```
med %>%
  group_by(drg_definition) %>%
  filter(sum(tot_discharges_16) >= 25000) %>%
  summarise(percent_chng = (sum(tot_discharges_16)/sum(tot_discharges_15)-1) * 100,
            discharges_16 = sum(tot_discharges_16)) %>%
  arrange(percent_chng)
```

```
## # A tibble: 48 x 3
##    drg_definition                                  percent_chng discharges_16
##    <chr>                                                  <dbl>         <int>
##  1 192 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE~         -26.4          33432
##  2 195 - SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC         -20.5          28666
##  3 191 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE~         -16.3          80565
##  4 194 - SIMPLE PNEUMONIA & PLEURISY W CC               -15.0         111429
##  5 066 - INTRACRANIAL HEMORRHAGE OR CEREBRAL I~         -14.9          30660
##  6 314 - OTHER CIRCULATORY SYSTEM DIAGNOSES W ~         -14.6          27974
##  7 293 - HEART FAILURE & SHOCK W/O CC/MCC               -10.3          39018
##  8 812 - RED BLOOD CELL DISORDERS W/O MCC                -9.26         48628
##  9 378 - G.I. HEMORRHAGE W CC                           -8.70         110005
## 10 313 - CHEST PAIN                                     -8.43          44683
## # ... with 38 more rows
```

```
med %>%
  group_by(drg_definition) %>%
  filter(sum(tot_discharges_16) >= 25000) %>%
  summarise(percent_chng = (sum(tot_discharges_16)/sum(tot_discharges_15) - 1) * 100) %>%
  ggplot(aes(x = percent_chng)) +
  geom_histogram(bins = 20) +
  scale_x_continuous("Percent Change 2015-2016", breaks = seq(-40, 40, 20)) +
  scale_y_continuous("Count", breaks = seq(0, 10, 1))
```



Evidently, most inpatient treatments did not change in occurence between 2015 and 2016, which makes the upward trend in lower extremity reattachment or replacement look more like noise than a signal. However, the vast change in infectious & parasitic diseases with O.R. procedure with MCC is surely something to which attention should be paid, as is the increase in the number of joint replacements. The number of occurrences is so large that an increase of 8.44% is very interesting. Furthermore, the number of occurences of pneumonia is declining, with and without major comorbidity.

Medicare Charges by State

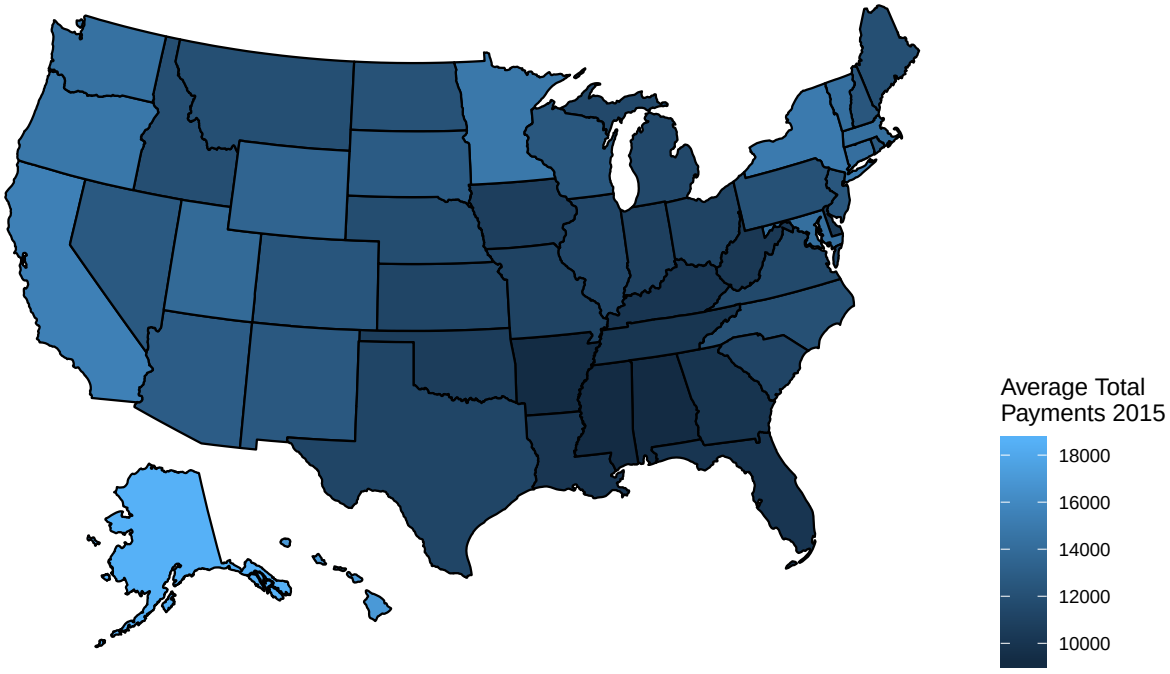
In this section, I analyze data related to inpatient charges across states in the US. Using the techniques specified in the Methodology document from the CMS, I recalculate totals to enable grouped summary statistics by state. The data in its original form contains total discharges and then averages for medicare payments, total payments and total covered charges. By multiplying these averages by the total discharges, I am able to recreate the totals that the CMS used to calculate the averages. This allows me to perform more accurate grouped summaries by state. Taking the average of averages weights observations unequally.

```
med <- med %>% mutate(
  tot_cov_charges_15 = avg_covered_charges_15 * tot_discharges_15,
  tot_cov_charges_16 = avg_covered_charges_16 * tot_discharges_16,
  tot_payments_15 = avg_tot_payments_15 * tot_discharges_15,
  tot_payments_16 = avg_tot_payments_16 * tot_discharges_16,
  tot_medicare_payments_15 = avg_medicare_payments_15 * tot_discharges_15,
  tot_medicare_payments_16 = avg_medicare_payments_16 * tot_discharges_16
)
```

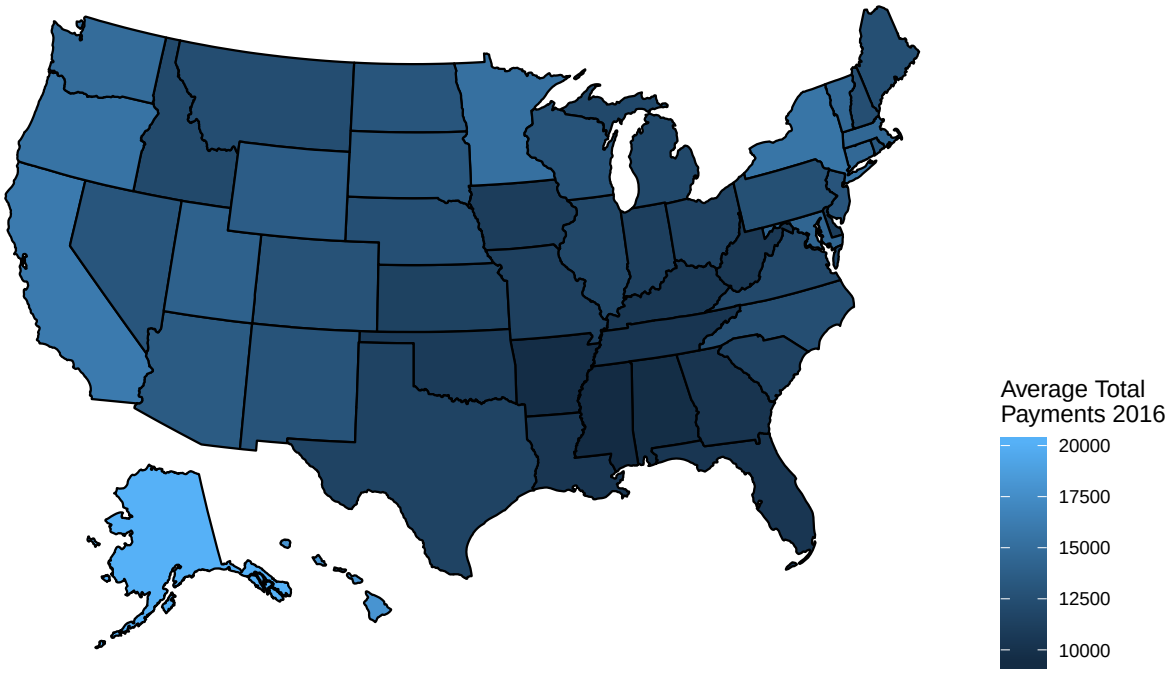
Highest Total Payments By State

The units for each graph are in dollars.

```
avg_total_payments <- med %>%
  rename("state" = "provider_state") %>%
  group_by(state) %>%
  summarise(avg_total_payments_15 = sum(tot_payments_15)/sum(tot_discharges_15),
            avg_total_payments_16 = sum(tot_payments_16)/sum(tot_discharges_16),
            diff = avg_total_payments_16 - avg_total_payments_15
            ) %>%
  arrange(desc(avg_total_payments_15, avg_total_payments_16))

plot_usmap(data = avg_total_payments, values = "avg_total_payments_15") +
  scale_fill_continuous(str_wrap("Average Total Payments 2015", 15)) +
  theme(legend.position = "right")
```

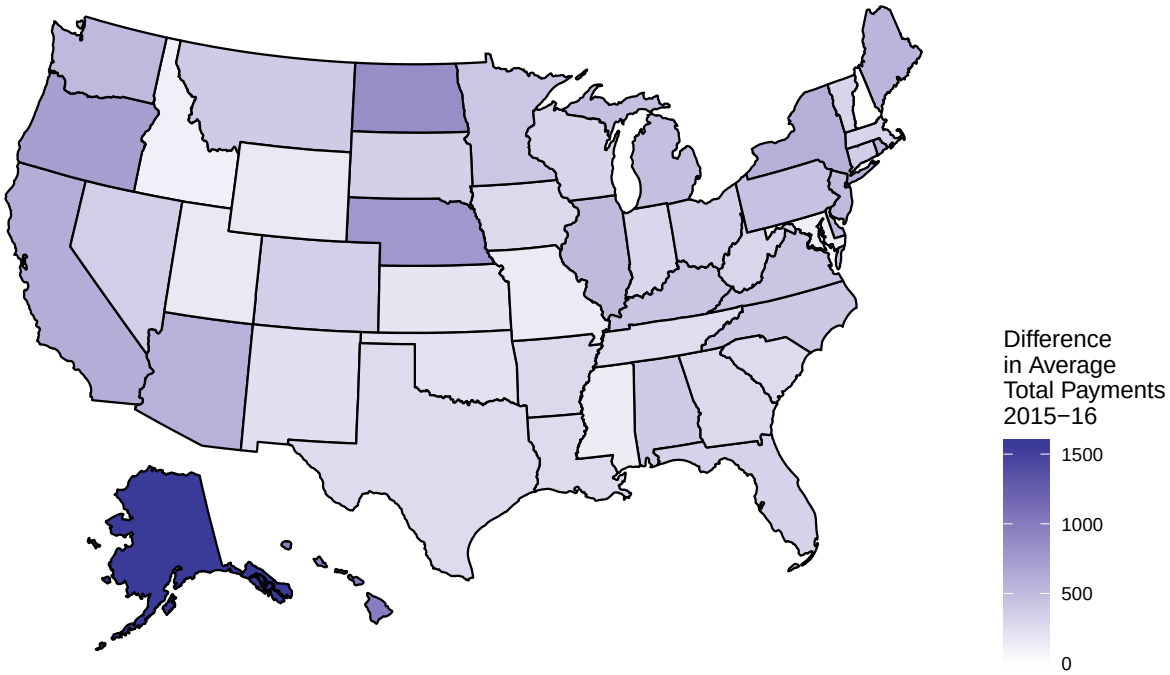



```
plot_usmap(data = avg_total_payments, values = "avg_total_payments_16") +
  scale_fill_continuous(str_wrap("Average Total Payments 2016", 15)) +
  theme(legend.position = "right")
```



Health care for the elderly was most costly in California, Hawaii, Alaska, Oregon, Washingon, Minnesota, New York and Vermont. The next graph represents the change in average total payments by state between 2015 and 2016.

```
plot_usmap(data = avg_total_payments, values = "diff") +
  scale_fill_gradient2(str_wrap("Difference in Average Total Payments 2015-16", 15)) +
  theme(legend.position = "right")
```



The key finding from these three graphs is that overall, average total payments are increasing. Almost every state saw an incrase in average total payments. This has many implications and could be related to the increasing number of elderly in the United States or changes to health care infrastructure and legislaton. In assessing the change in health care costs in this dataset, it is important to pay attention to the difference between individual contributions and medicare contributions.

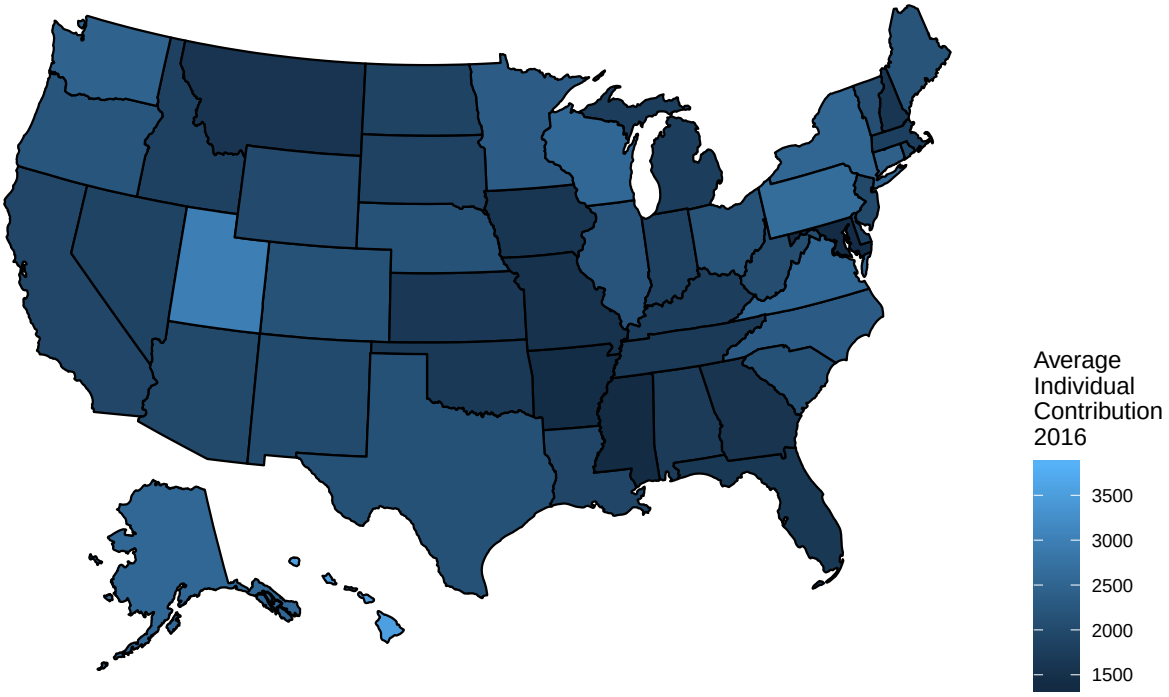
Individual Contributions by State

Using the methodology section of the CMS data release, I was able to calculate estimates of individuals contributions to their inpatient charges. This was calcultated by taking the difference between average total charges and average medicare payments.

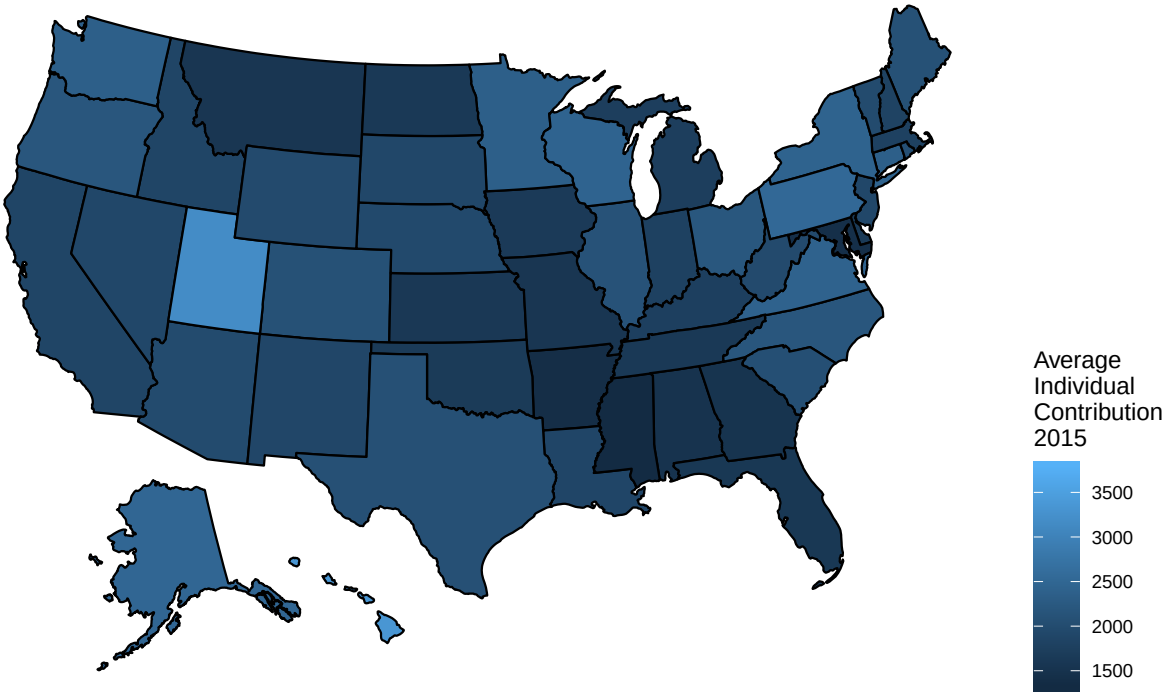
```
med <- med %>% mutate(tot_copay_deductible_16 = tot_payments_16 - tot_medicare_payments_16,
                      tot_copay_deductible_15 = tot_payments_15 - tot_medicare_payments_15)


avg_copay_deductible <- med %>%
  rename("state" = "provider_state") %>%
  group_by(state) %>%
  summarise(avg_copay_16 = sum(tot_copay_deductible_16)/sum(tot_discharges_16),
         avg_copay_15 = sum(tot_copay_deductible_15)/sum(tot_discharges_15),
         diff = avg_copay_16 - avg_copay_15)


avg_copay_deductible %>%
  arrange(desc(avg_copay_16, avg_copay_15))
```

```
## # A tibble: 51 x 4
##    state avg_copay_16 avg_copay_15   diff
##    <chr>        <dbl>        <dbl>  <dbl>
##  1 DC           3820.        3786.   33.6
##  2 HI           3519.        3331.  188.
##  3 UT           2970.        3162. -192.
##  4 PA           2627.        2527.  101.
##  5 AK           2549.        2491.   57.8
##  6 WI           2546.        2433.  113.
##  7 VA           2542.        2413.  129.
##  8 NY           2518.        2430.   88.1
##  9 WA           2451.        2358.   93.4
## 10 CT           2417.        2358.   59.7
## # ... with 41 more rows
```

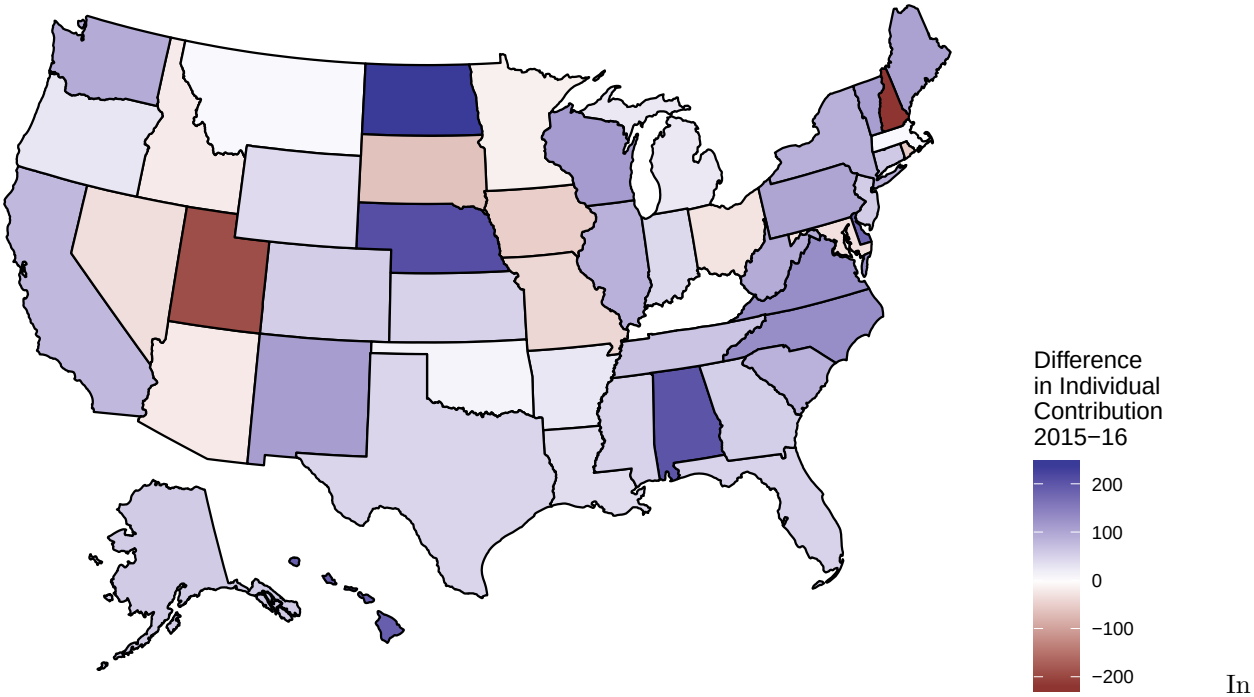
```
plot_usmap(data = avg_copay_deductible, values = "avg_copay_16") +
  scale_fill_continuous(str_wrap("Average Individual Contribution 2016", 15)) +
  theme(legend.position = "right")
```




```
plot_usmap(data = avg_copay_deductible, values = "avg_copay_15") +
  scale_fill_continuous(str_wrap("Average Individual Contribution 2015", 15)) +
  theme(legend.position = "right")
```



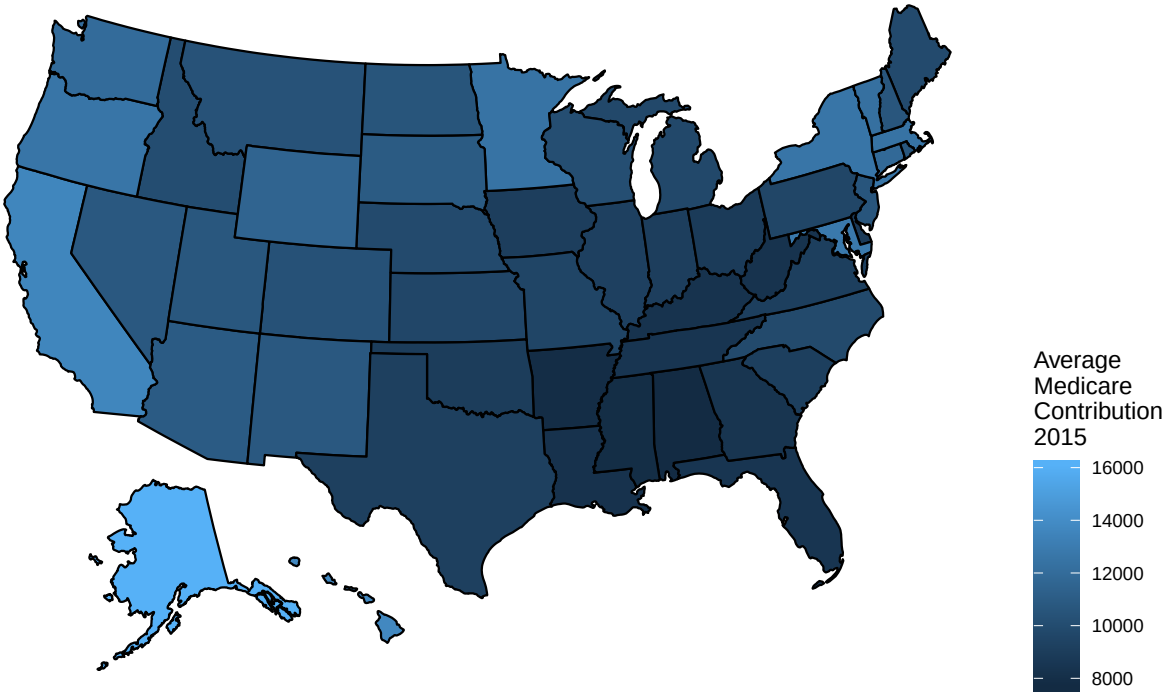

```
plot_usmap(data = avg_copay_deductible, values = "diff") +
  scale_fill_gradient2(str_wrap("Difference in Individual Contribution 2015-16", 15)) +
  theme(legend.position = "right")
```



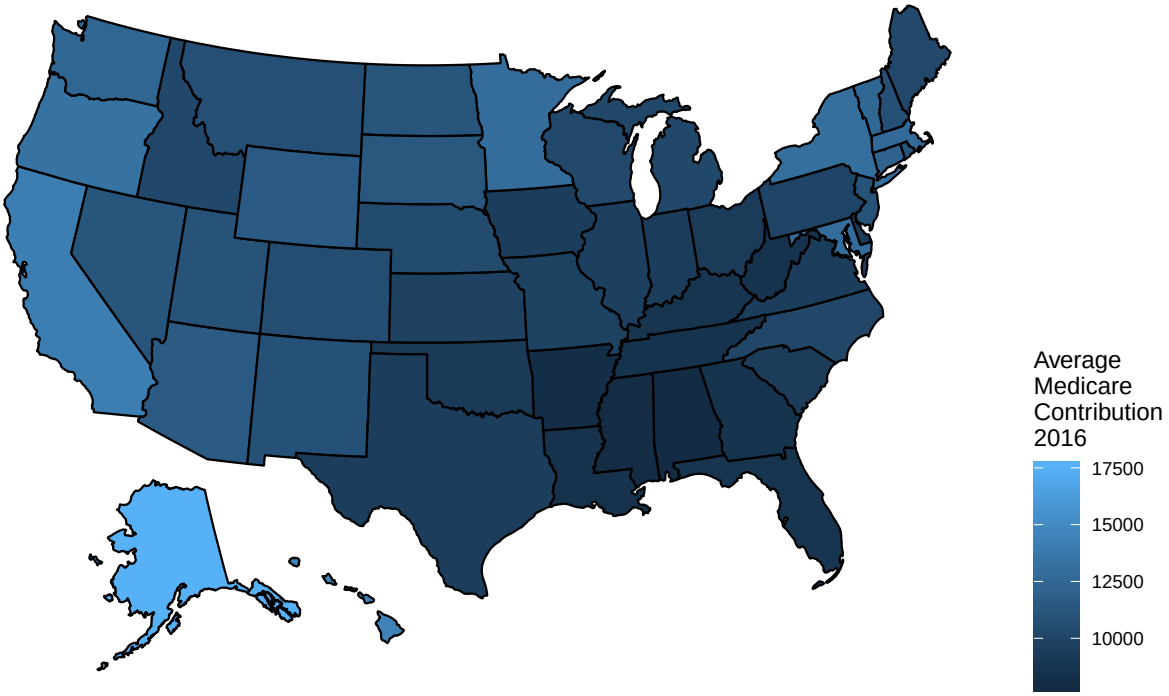
In 2016, average individual contributions were highest in Hawaii, Utah, Pennsylvania, Alaska, Wisconsin, Virginia, New York, Washington, and Connecticut. This is a very interesting mix of states to be in the top 10 as there does not seem to be a self evident factor connecting the states. Further research should analyze commonalities and look more deeply into what could be keeping individual contributions for inpatient procedures high for Medicare recipients. Furthermore, while most states saw either modest increases,or no change in individual contributions, average individual contributions in Utah fell by $192 and in Vermont by $218. The data on Vermont is particularly interesting because Vermont has relatively high average total costs. North Dakota, Nebraska, and Alabama all saw large increases in individual contributions to inpatient charges.

Medicare contributions by state

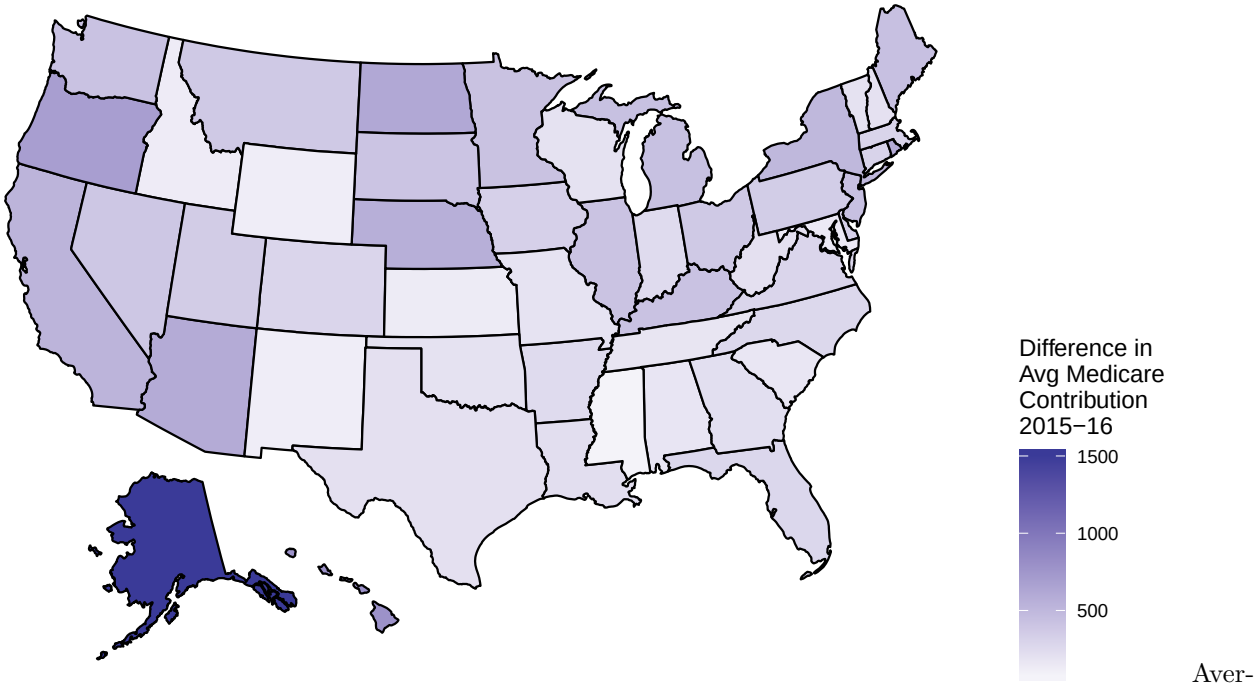
```
med_payments <- med %>%
  rename("state" = "provider_state") %>%
  group_by(state) %>%
  summarise(avg_med_payments_15 = sum(tot_medicare_payments_15)/sum(tot_discharges_15),
            avg_med_payments_16 = sum(tot_medicare_payments_16)/sum(tot_discharges_16),
            diff = avg_med_payments_16 - avg_med_payments_15)

plot_usmap(data = med_payments, values = "avg_med_payments_15") +
  scale_fill_continuous(str_wrap("Average Medicare Contribution 2015", 15)) +
  theme(legend.position = "right")
```




```
plot_usmap(data = med_payments, values = "avg_med_payments_16") +
  scale_fill_continuous(str_wrap("Average Medicare Contribution 2016", 15)) +
  theme(legend.position = "right")
```




```
plot_usmap(data = med_payments, values = "diff") +
  scale_fill_gradient2(str_wrap("Difference in Avg Medicare Contribution 2015-16", 15)) +
  theme(legend.position = "right")
```



Average medicare contributions are highest in states such as California, Oregon, Washington, Minnesota, New York, Vermont, Delaware, and Massachusetts. There appears to be a realtively consistent trend that in northern states, there is higher Medicare contribution. States with a lot of retirees that also have relatively low Medicare contribution (i.e. Florida) likely have low numbers because health care costs are lower in these states, as shown in the data related to average total costs. Additionally, across every state, medicare contributions increased between 2015 and 2016.

Most Expensive Conditions to Treat

```
avg_cost <- med %>%
  group_by(drg_definition) %>%
  summarise(avg_payments_15 = sum(tot_payments_15)/sum(tot_discharges_15),
            avg_payments_16 = sum(tot_payments_16)/sum(tot_discharges_16),
            avg_medicare_16 = sum(tot_medicare_payments_16)/sum(tot_discharges_16),
            avg_medicare_15 = sum(tot_medicare_payments_15)/sum(tot_discharges_15))

avg_cost %>% arrange(desc(avg_payments_16, avg_tot_payments_15))
```

```
## # A tibble: 418 x 5
##    drg_definition          avg_payments_15 avg_payments_16 avg_medicare_16
##    <chr>                             <dbl>           <dbl>           <dbl>
##  1 001 - HEART TRANSPLANT~         258349.         260137.         224075.
##  2 005 - LIVER TRANSPLANT~         127243.         121724.          95358.
##  3 014 - ALLOGENEIC BONE ~         118329.         119396.         104951.
##  4 007 - LUNG TRANSPLANT           112664.         111440.          89346.
##  5 453 - COMBINED ANTERIO~          91454.          92416.          85122.
##  6 020 - INTRACRANIAL VAS~          84323.          84259.          76394.
##  7 216 - CARDIAC VALVE & ~          76046.          77561.          69419.
##  8 008 - SIMULTANEOUS PAN~          88885.          75934.          37525.
##  9 957 - OTHER O.R. PROCE~          67854.          71720.          58199.
## 10 834 - ACUTE LEUKEMIA W~          67226.          70869.          56995.
## # ... with 408 more rows, and 1 more variable: avg_medicare_15 <dbl>
```

At the top of the list are very intensive procedures. This is to be expected. Heart assist systems or heart transplants with MCC tops the list, likely because the procedure is so complex. Liver transplants, bone marrow transplants and lung transplants follow, which also makes sense given the complexity of the treatment. However, it is also a good idea to look at the conditions that are best and worst covered by medicare.

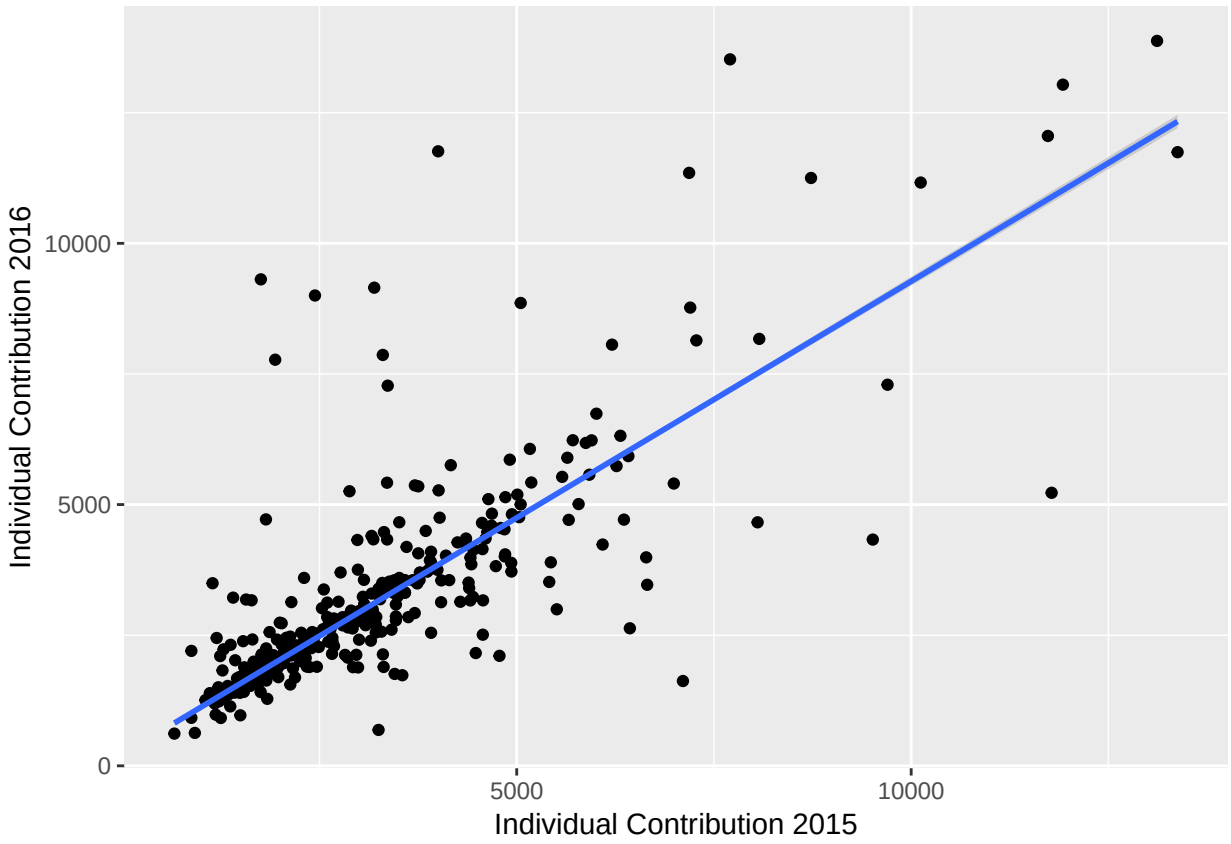
Copays and Deductibles by Condition

```
copay_deductible <- med %>%
  group_by(drg_definition) %>%
  summarise(avg_payments_15 = sum(tot_payments_15)/sum(tot_discharges_15),
            avg_payments_16 = sum(tot_payments_16)/sum(tot_discharges_16),
            avg_medicare_16 = sum(tot_medicare_payments_16)/sum(tot_discharges_16),
            avg_medicare_15 = sum(tot_medicare_payments_15)/sum(tot_discharges_15)) %>%
  transmute(drg_definition = drg_definition,
            copay_deductible_15 = avg_payments_15 - avg_medicare_15,
            copay_deductible_16 = avg_payments_16 - avg_medicare_16
            )
arrange(copay_deductible, desc(copay_deductible_16))
```

```
## # A tibble: 418 x 3
##    drg_definition                    copay_deductible~ copay_deductible~
##    <chr>                                         <dbl>             <dbl>
##  1 008 - SIMULTANEOUS PANCREAS/KIDNEY~           58177             38409.
##  2 001 - HEART TRANSPLANT OR IMPLANT ~           40011.            36062.
##  3 005 - LIVER TRANSPLANT W MCC OR IN~           30425.            26367.
##  4 007 - LUNG TRANSPLANT                         29404.            22094.
##  5 014 - ALLOGENEIC BONE MARROW TRANS~           19385.            14445.
##  6 834 - ACUTE LEUKEMIA W/O MAJOR O.R~           13117.            13874.
##  7 006 - LIVER TRANSPLANT W/O MCC                16937.            13818.
##  8 957 - OTHER O.R. PROCEDURES FOR MU~            7704.            13521.
##  9 823 - LYMPHOMA & NON-ACUTE LEUKEMI~           11923.            13037.
## 10 016 - AUTOLOGOUS BONE MARROW TRANS~           11734.            12055.
## # ... with 408 more rows
```

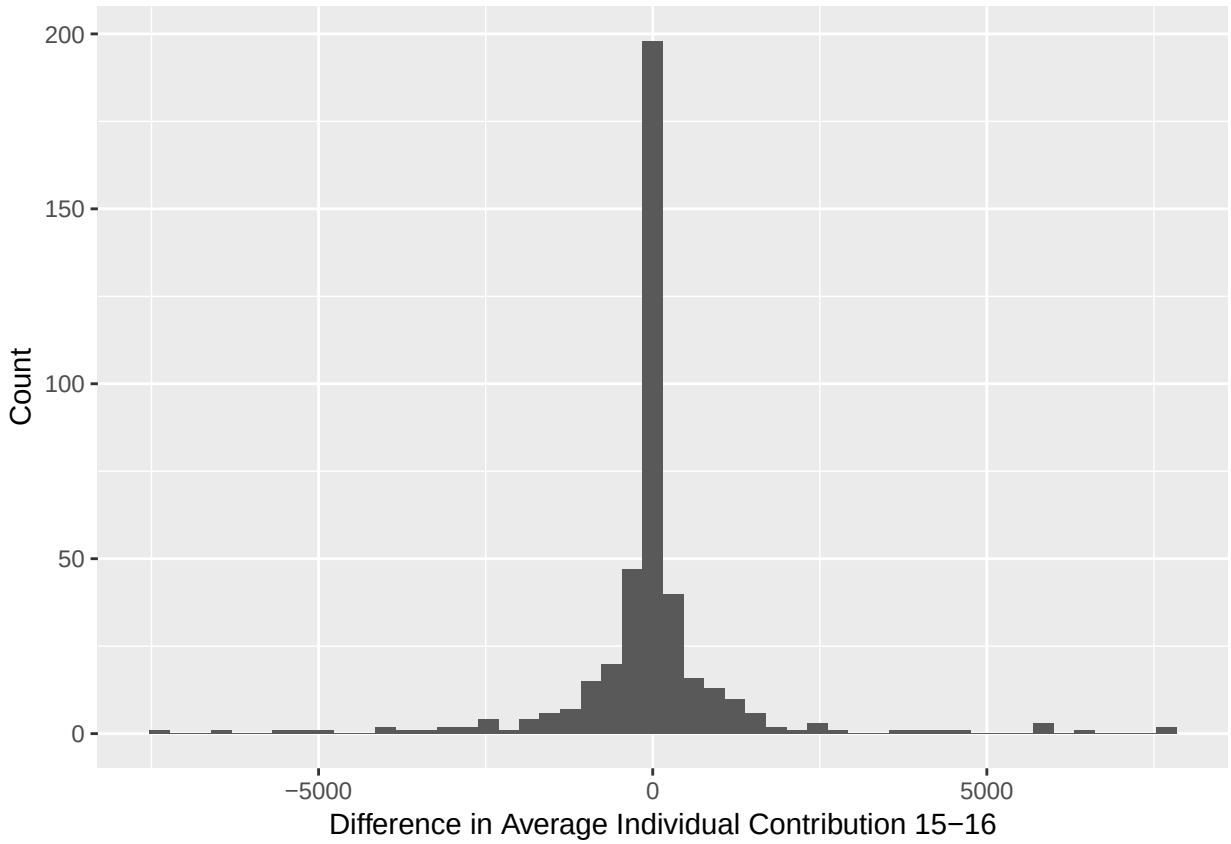
There is a considerable amount of overlap between the most expensive procedures and the ones with the highest out of pocket cost. Most data falls under $15000 for individual contributions so I will filter out observations greater than 15000. The following is a graph that represents the relationship between individuals contributions in 2015 and 2016.

```
copay_deductible %>% filter(copay_deductible_15 <= 15000) %>%
  ggplot(aes(x = copay_deductible_15, y = copay_deductible_16)) +
  geom_point() +
  geom_smooth(method = MASS::rlm) +
  xlab("Individual Contribution 2015") +
  ylab("Individual Contribution 2016")
```



It appears that even a robust linear model does not fit the data particularly well. There are still quite large numbers of outliers. I am going to plot a histogram of the difference in individual contributions by MS-DRG between 2015 and 2016.

```
copay_deductible <- copay_deductible %>%
  mutate(diff = copay_deductible_16 - copay_deductible_15)

copay_deductible %>%
  filter(diff > -10000) %>%
  ggplot(aes(x = diff)) +
  geom_histogram(bins = 50) +
  xlab("Difference in Average Individual Contribution 15-16") +
  ylab("Count")
```




```
var(copay_deductible$diff)
```

```
## [1] 2801467
```

The vast majority of copays/deductibles did not change with respect to condition. However, the distribution has an immense variance. This volatility in pricing should be further investigated in future analysis of Medicare data.