Final Project

Bin Lin

2016-12-16

1. Introduction:

Since 2014, US has been spending more than $3 trillion dollars annually on healthcare and the average health expenditure is about $9523 per capita. In addition, these numbers keep increasing as those baby boomers start hitting their retirement age. If america spends too much resources on healthcare, there will be less money and resources we can spend on elsewhere because the budget is always limited.

Medicare and Medicaid are two of the largest federal entitlement programs. People who enroll under Medicare are usually elderly who are over 65 years old. On the other hand, the medicaid is designed for people whose household incomes are under certain limit of federal poverty level.

2. Objectives For this final project, I want to create a markdown file that can show the breakdown of spending of medicare and medicaid. In addition to that, I want to compare the spendings between different states, to get some insights why some states have higher spending than the other and investigate if the differences are statistically significant.

3. Data Sources: I am using the datasets directly from https://www.data.gov/ (https://www.data.gov/). The datasets were recently updated, therefore the analysis will be considered current to reflect the utilization of Medicare and Medicaid. Also since I am conducting analysis that is corresponding to the entire country. Govenment data is the best shot for me to obtain.

4. Analysis

a. The first dataset I load to the RStudio is called “Medicare Hospital Spending by Claim”. The data shows average spending levels during hospitals episodes. An MSPB (Medicare Spending per Beneficiary (MSPB) episode includes all Medicare Part A and Part B claims paid during the period from 3 days prior to a hospital admission through 30 days after discharge. The payment amount have been adjusted based on geographic effetcs on payment.

```
#Loading necessary libraries to RStudio
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(jsonlite)
library(XML)
library(RCurl)
```

```
## Loading required package: bitops
```

```
##
## Attaching package: 'RCurl'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
library(RMySQL)
```

```
## Warning: package 'RMySQL' was built under R version 3.3.2
```

```
## Loading required package: DBI
```

```
## Warning: package 'DBI' was built under R version 3.3.2
```

```
library(ggmap)
```

```
## Warning: package 'ggmap' was built under R version 3.3.2
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
setwd("C:/Users/blin261/Desktop/DATA607/DATA607Final")

#Exploring the dataset
raw_data <- read.table("Medicare_Hospital_Spending_by_Claim.csv", sep = ",", stringsAsFactors =
FALSE, header = TRUE)
head(raw_data)
```

```
##           Hospital.Name Provider.Number State
## 1 HELEN KELLER HOSPITAL           10019    AL
## 2 HELEN KELLER HOSPITAL           10019    AL
## 3 HELEN KELLER HOSPITAL           10019    AL
## 4 HELEN KELLER HOSPITAL           10019    AL
## 5 HELEN KELLER HOSPITAL           10019    AL
## 6 HELEN KELLER HOSPITAL           10019    AL
##                                                             Period
## 1                                 During Index Hospital Admission
## 2                                 During Index Hospital Admission
## 3                                 During Index Hospital Admission
## 4 1 through 30 days After Discharge from Index Hospital Admission
## 5 1 through 30 days After Discharge from Index Hospital Admission
## 6 1 through 30 days After Discharge from Index Hospital Admission
##                 Claim.Type Avg.Spending.Per.Episode..Hospital.
## 1  Skilled Nursing Facility                                  $0
## 2 Durable Medical Equipment                                 $18
## 3                  Carrier                               $1062
## 4       Home Health Agency                                $917
## 5                  Hospice                                $172
## 6                Inpatient                               $2518
##   Avg.Spending.Per.Episode..State. Avg.Spending.Per.Episode..Nation.
## 1                               $0                                $0
## 2                              $31                               $24
## 3                            $1480                             $1540
## 4                             $948                              $816
## 5                             $154                              $122
## 6                            $2634                             $2702
##   Percent.of.Spending..Hospital. Percent.of.Spending..State.
## 1                             0%                          0%
## 2                           0.1%                       0.16%
## 3                          6.01%                       7.71%
## 4                          5.19%                       4.94%
## 5                          0.97%                        0.8%
## 6                         14.25%                      13.72%
##   Percent.of.Spending..Nation. Measure.Start.Date Measure.End.Date
## 1                           0%      01/01/1012015   01/01/12312015
## 2                        0.12%      01/01/1012015   01/01/12312015
## 3                        7.52%      01/01/1012015   01/01/12312015
## 4                        3.98%      01/01/1012015   01/01/12312015
## 5                         0.6%      01/01/1012015   01/01/12312015
## 6                       13.18%      01/01/1012015   01/01/12312015
```

```
str(raw_data)
```

```
## 'data.frame':    32971 obs. of  13 variables:
##  $ Hospital.Name                    : chr  "HELEN KELLER HOSPITAL" "HELEN KELLER HOSPITAL"
"HELEN KELLER HOSPITAL" "HELEN KELLER HOSPITAL" ...
##  $ Provider.Number                  : int  10019 10019 10019 10019 10019 10019 10019 10019
10019 10019 ...
##  $ State                            : chr  "AL" "AL" "AL" "AL" ...
##  $ Period                           : chr  "During Index Hospital Admission" "During Index
Hospital Admission" "During Index Hospital Admission" "1 through 30 days After Discharge from In
dex Hospital Admission" ...
##  $ Claim.Type                       : chr  "Skilled Nursing Facility" "Durable Medical Equi
pment" "Carrier" "Home Health Agency" ...
##  $ Avg.Spending.Per.Episode..Hospital.: chr  "$0" "$18" "$1062" "$917" ...
##  $ Avg.Spending.Per.Episode..State.   : chr  "$0" "$31" "$1480" "$948" ...
##  $ Avg.Spending.Per.Episode..Nation.  : chr  "$0" "$24" "$1540" "$816" ...
##  $ Percent.of.Spending..Hospital.     : chr  "0%" "0.1%" "6.01%" "5.19%" ...
##  $ Percent.of.Spending..State.        : chr  "0%" "0.16%" "7.71%" "4.94%" ...
##  $ Percent.of.Spending..Nation.       : chr  "0%" "0.12%" "7.52%" "3.98%" ...
##  $ Measure.Start.Date               : chr  "01/01/1012015" "01/01/1012015" "01/01/1012015"
"01/01/1012015" ...
##  $ Measure.End.Date                 : chr  "01/01/12312015" "01/01/12312015" "01/01/1231201
5" "01/01/12312015" ...
```

This subset of original data frame contains the aggragate information about medicare expenses incurred from hospital visit. I performed neccessary cleaning and transformation of the data. To have it ready for further studies.

```
raw_complete <- subset(raw_data, raw_data$Period == "Complete Episode")

complete_episode <- raw_complete[, c("State", "Period", "Claim.Type", "Avg.Spending.Per.Episod
e..State.", "Avg.Spending.Per.Episode..Nation.")]

colnames(complete_episode) <- c("State", "Period", "Claim_Type", "Avg_Spending_Per_Episode_Stat
e", "Avg_Spending_Per_Episode_Nation")
head(complete_episode)
```

```
##     State           Period Claim_Type Avg_Spending_Per_Episode_State
## 11     AL Complete Episode      Total                         $19201
## 33     AL Complete Episode      Total                         $19201
## 55     AL Complete Episode      Total                         $19201
## 80     AL Complete Episode      Total                         $19201
## 102    AL Complete Episode      Total                         $19201
## 125    AL Complete Episode      Total                         $19201
##     Avg_Spending_Per_Episode_Nation
## 11                           $20497
## 33                           $20497
## 55                           $20497
## 80                           $20497
## 102                          $20497
## 125                          $20497
```

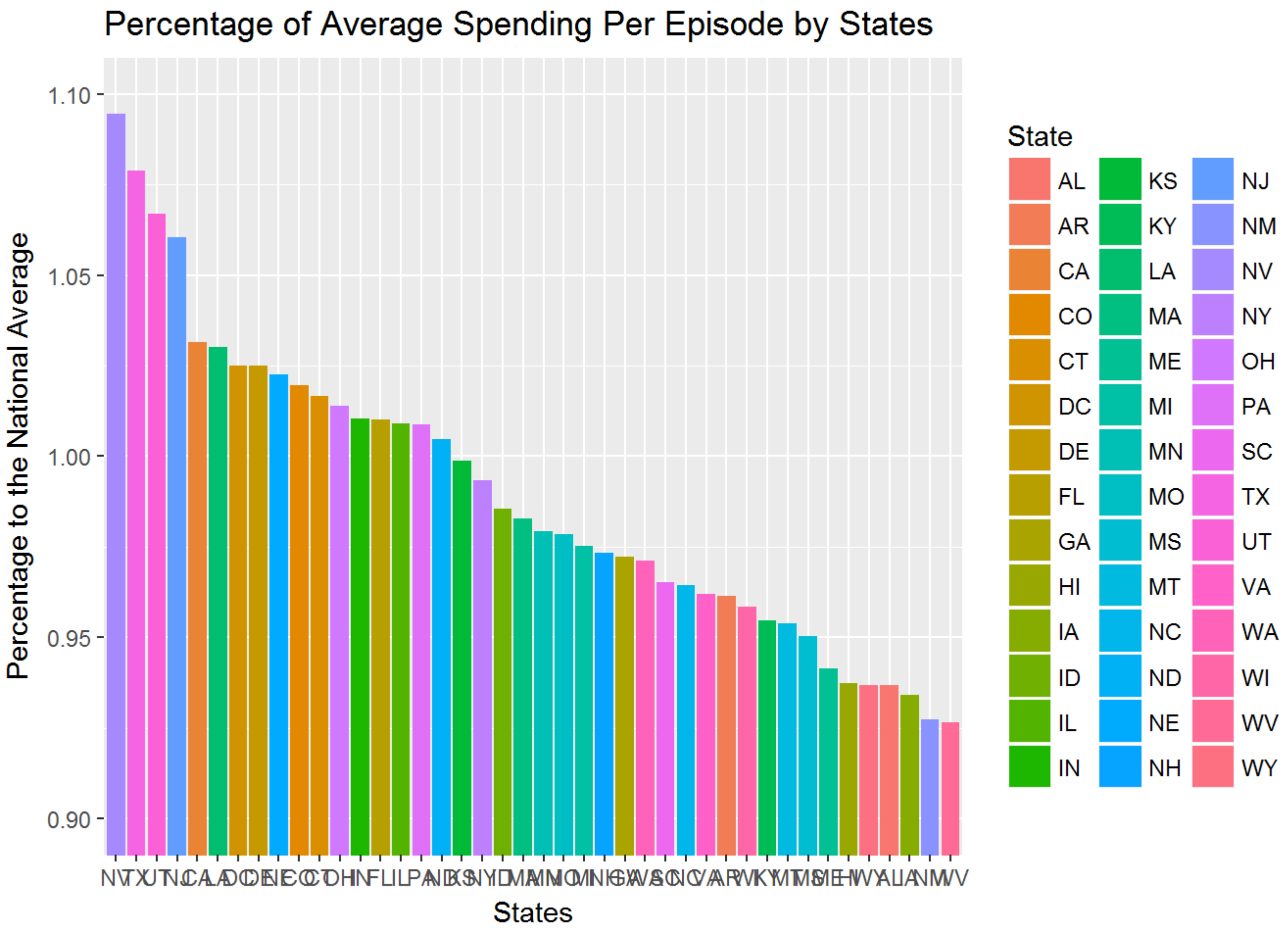
I grouped the data by state, therefore it is easier to make any comparison between states. I also order the list by descending order according to the average spending per episode in that state.

```
episode_cost <- complete_episode%>%
  group_by(State)%>%
  mutate(Percentage = as.numeric(sub("\\$", "", Avg_Spending_Per_Episode_State)) / as.numeric(su
b("\\$", "", Avg_Spending_Per_Episode_Nation)), count = n())%>%
  unique()%>%
  arrange(desc(Percentage))
head(episode_cost)
```

```
## Source: local data frame [6 x 7]
## Groups: State [6]
## 
##   State           Period Claim_Type Avg_Spending_Per_Episode_State
##   <chr>            <chr>      <chr>                          <chr>
## 1    NV Complete Episode      Total                         $22432
## 2    TX Complete Episode      Total                         $22110
## 3    UT Complete Episode      Total                         $21871
## 4    NJ Complete Episode      Total                         $21733
## 5    CA Complete Episode      Total                         $21141
## 6    LA Complete Episode      Total                         $21116
## # ... with 3 more variables: Avg_Spending_Per_Episode_Nation <chr>,
## #   Percentage <dbl>, count <int>
```

The result of the bar plot imply that states such as Nevada, Texas, Utah, and New Jersey has much higher percentage of medicare hospital spending than the national average. NY which is my home state has actually lower percentage. West Virginia is the state with the lowest medicare hospital spending per member in the entire country.

```
ggplot(data = episode_cost, aes(x = reorder(episode_cost$State, -episode_cost$Percentage), y = e
pisode_cost$Percentage, fill = State)) + geom_bar(stat = "identity") + coord_cartesian(ylim = c(
0.9, 1.1)) + ggtitle("Percentage of Average Spending Per Episode by States")+ xlab("States") + y
lab("Percentage to the National Average")
```



Then I pick California and New York to investigate what hospital claims to cause one state (CA) to have higher expenses than the other(NY)

```
#Get a new subset to contain variable about claim type.
claim <- raw_data[, c("Hospital.Name", "State", "Period", "Claim.Type", "Avg.Spending.Per.Episod
e..State.", "Avg.Spending.Per.Episode..Nation.")]

colnames(claim) <- c("Hosital_Name", "State", "Period", "Claim_Type", "Avg_Spending_Per_Episode_
State", "Avg_Spending_Per_Episode_Nation")

head(claim)
```

```
##            Hosital_Name State
## 1 HELEN KELLER HOSPITAL    AL
## 2 HELEN KELLER HOSPITAL    AL
## 3 HELEN KELLER HOSPITAL    AL
## 4 HELEN KELLER HOSPITAL    AL
## 5 HELEN KELLER HOSPITAL    AL
## 6 HELEN KELLER HOSPITAL    AL
##                                                             Period
## 1                                 During Index Hospital Admission
## 2                                 During Index Hospital Admission
## 3                                 During Index Hospital Admission
## 4 1 through 30 days After Discharge from Index Hospital Admission
## 5 1 through 30 days After Discharge from Index Hospital Admission
## 6 1 through 30 days After Discharge from Index Hospital Admission
##                  Claim_Type Avg_Spending_Per_Episode_State
## 1  Skilled Nursing Facility                             $0
## 2 Durable Medical Equipment                            $31
## 3                   Carrier                          $1480
## 4        Home Health Agency                           $948
## 5                   Hospice                           $154
## 6                 Inpatient                          $2634
##   Avg_Spending_Per_Episode_Nation
## 1                              $0
## 2                             $24
## 3                           $1540
## 4                            $816
## 5                            $122
## 6                           $2702
```

```
#I also add two new variables to the subset. One is the average cost per claim. It differs based
on the claim type. The second variable is the percentage compare to the national average about t
he medicare spending of each claim type.

claim_cost_NY <- claim%>%
  filter(State == "NY")%>%
  group_by(State, Claim_Type)%>%
  summarize(count = n(), ave_cost = sum(as.numeric(sub("\\$", "", Avg_Spending_Per_Episode_State
))) / count, percentage = ave_cost / (sum(as.numeric(sub("\\$", "", Avg_Spending_Per_Episode_Nat
ion)))/ count))
claim_cost_NY
```

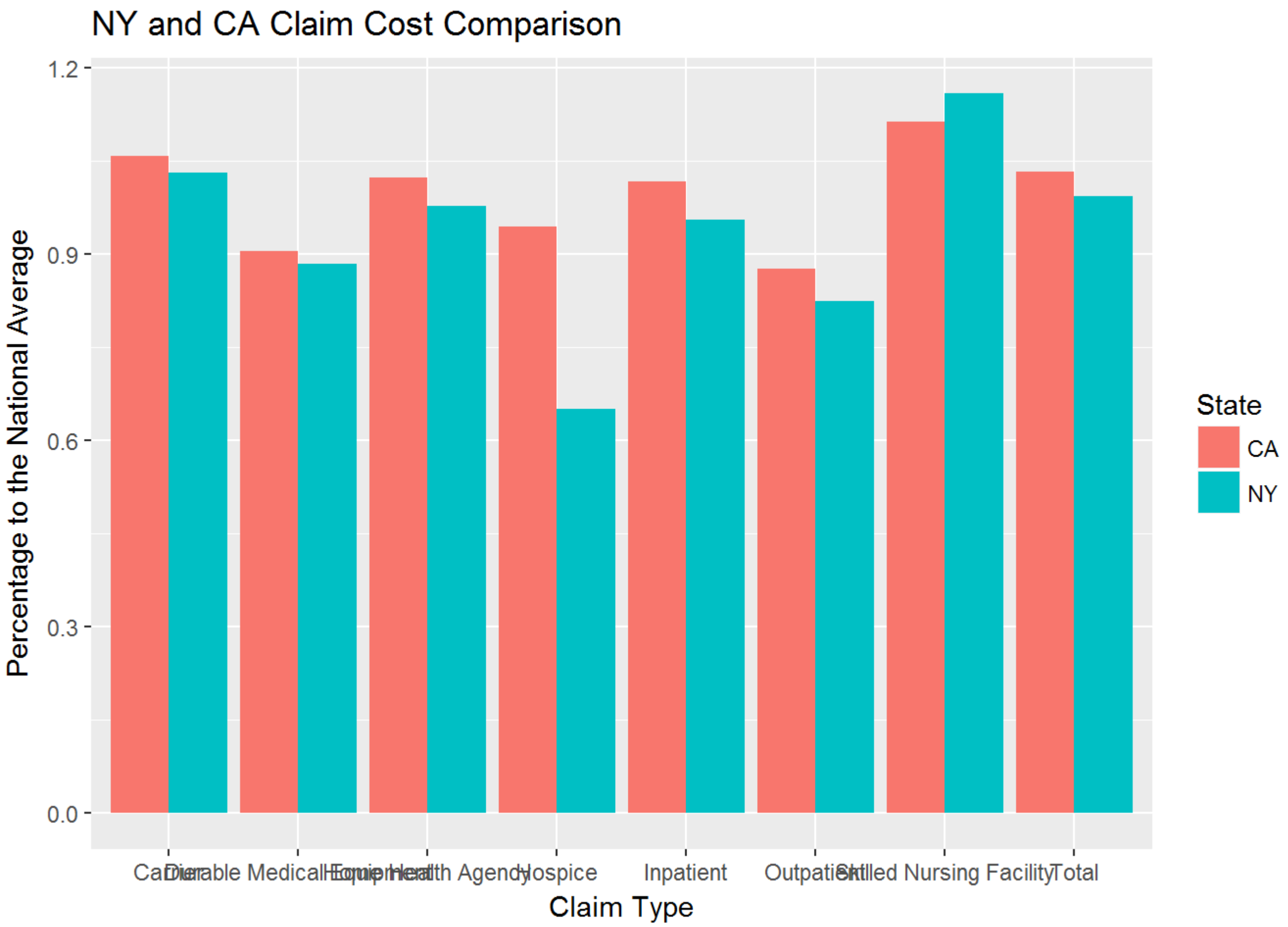
```
## Source: local data frame [8 x 5]
## Groups: State [?]
##
##   State                Claim_Type count    ave_cost percentage
##   <chr>                     <chr> <int>       <dbl>      <dbl>
## 1    NY                   Carrier   203  1100.55665  1.0313305
## 2    NY Durable Medical Equipment   208    40.38942  0.8837576
## 3    NY        Home Health Agency   205   268.90732  0.9772034
## 4    NY                   Hospice   204    27.45098  0.6504821
## 5    NY                 Inpatient   201  3844.52239  0.9554573
## 6    NY                Outpatient   207   241.86957  0.8240396
## 7    NY  Skilled Nursing Facility   200  1268.86500  1.1577869
## 8    NY                     Total    70 20363.00000  0.9934625
```

```
claim_cost_CA <- claim%>%
  filter(State == "CA")%>%
  group_by(State, Claim_Type)%>%
  summarize(count = n(), ave_cost = sum(as.numeric(sub("\\$", "", Avg_Spending_Per_Episode_State
))) / count, percentage = ave_cost / (sum(as.numeric(sub("\\$", "", Avg_Spending_Per_Episode_Nat
ion)))/ count))
claim_cost_CA
```

```
## Source: local data frame [8 x 5]
## Groups: State [?]
##
##   State                Claim_Type count    ave_cost percentage
##   <chr>                     <chr> <int>       <dbl>      <dbl>
## 1    CA                   Carrier   594  1128.34343  1.0575883
## 2    CA Durable Medical Equipment   592    40.94426  0.9044066
## 3    CA        Home Health Agency   593   283.38954  1.0229237
## 4    CA                   Hospice   587    38.93015  0.9432841
## 5    CA                 Inpatient   584  4107.59075  1.0161304
## 6    CA                Outpatient   582   252.45876  0.8752874
## 7    CA  Skilled Nursing Facility   589  1229.24788  1.1123168
## 8    CA                     Total   196 21141.00000  1.0314192
```

The following bar plot shows the Medicare program in California has higher spending percentage for almost all claim types except the claims for the nursing facility. While NY and CA's medicare hospital spending are about the same for most claim type, however, california's medicare program spend much more money on the hospice care of the elderly. It outspends NY by almost 30%.

```
claim_cost <- rbind(claim_cost_NY, claim_cost_CA)
ggplot(data = claim_cost, aes(x = Claim_Type, y = percentage, fill = State)) + geom_bar(stat =
"identity", position = "dodge") + ggtitle("NY and CA Claim Cost Comparison")+ xlab("Claim Type")
+ ylab("Percentage to the National Average")
```



b. Second dataset is about amount of reimbursement that are paid by state Medicaid program for each prescription drugs. It contains variable such as drug name, number of units reimbursed, amount of reimbursement et cetera.

```
#Load and explore the data.
setwd("C:/Users/blin261/Desktop/DATA607/DATA607Final")
raw_data1 <- read.table("State_Drug_Utilization_Data_2016.csv", sep = ",", stringsAsFactors = FA
LSE, header = TRUE)

head(raw_data1)
```

```
##   Utilization.Type State Labeler.Code Product.Code Package.Size Year
## 1             FFSU    AK            2         1433           80 2016
## 2             FFSU    AK            2         1433           80 2016
## 3             FFSU    AK            2         1434           80 2016
## 4             FFSU    AK            2         1434           80 2016
## 5             FFSU    AK            2         1975           90 2016
## 6             FFSU    AK            2         3227           30 2016
##   Quarter Product.Name Suppression.Used Units.Reimbursed
## 1       1   TRULICITY              true               NA
## 2       2   TRULICITY              true               NA
## 3       1   TRULICITY             false               32
## 4       2   TRULICITY              true               NA
## 5       2   AXIRON                 true               NA
## 6       1   STRATTERA             false             1333
##   Number.of.Prescriptions Total.Amount.Reimbursed
## 1                      NA                      NA
## 2                      NA                      NA
## 3                      16                 8882.87
## 4                      NA                      NA
## 5                      NA                      NA
## 6                      40                14311.75
##   Medicaid.Amount.Reimbursed Non.Medicaid.Amount.Reimbursed Quarter.begin
## 1                         NA                             NA           1/1
## 2                         NA                             NA           4/1
## 3                    8882.87                           0.00           1/1
## 4                         NA                             NA           4/1
## 5                         NA                             NA           4/1
## 6                   13192.79                        1118.96           1/1
##       Quarter.Begin.Date X_latitude X_longitude             Location
## 1 01/01/2016 12:00:00 AM     61.385   -152.2683 (61.3850, -152.2683)
## 2 04/01/2016 12:00:00 AM     61.385   -152.2683 (61.3850, -152.2683)
## 3 01/01/2016 12:00:00 AM     61.385   -152.2683 (61.3850, -152.2683)
## 4 04/01/2016 12:00:00 AM     61.385   -152.2683 (61.3850, -152.2683)
## 5 04/01/2016 12:00:00 AM     61.385   -152.2683 (61.3850, -152.2683)
## 6 01/01/2016 12:00:00 AM     61.385   -152.2683 (61.3850, -152.2683)
##       NDC
## 1 2143380
## 2 2143380
## 3 2143480
## 4 2143480
## 5 2197590
## 6 2322730
```

```
str(raw_data1)
```

```
## 'data.frame':    1103372 obs. of  20 variables:
##  $ Utilization.Type             : chr  "FFSU" "FFSU" "FFSU" "FFSU" ...
##  $ State                        : chr  "AK" "AK" "AK" "AK" ...
##  $ Labeler.Code                 : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ Product.Code                 : int  1433 1433 1434 1434 1975 3227 3227 3228 3228 3229 ...
##  $ Package.Size                 : int  80 80 80 80 90 30 30 30 30 30 ...
##  $ Year                         : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ Quarter                      : int  1 2 1 2 2 1 2 1 2 1 ...
##  $ Product.Name                 : chr  "TRULICITY " "TRULICITY " "TRULICITY " "TRULICITY "
...
##  $ Suppression.Used             : chr  "true" "true" "false" "true" ...
##  $ Units.Reimbursed             : num  NA NA 32 NA NA ...
##  $ Number.of.Prescriptions      : int  NA NA 16 NA NA 40 30 93 77 122 ...
##  $ Total.Amount.Reimbursed      : num  NA NA 8883 NA NA ...
##  $ Medicaid.Amount.Reimbursed   : num  NA NA 8883 NA NA ...
##  $ Non.Medicaid.Amount.Reimbursed: num  NA NA 0 NA NA ...
##  $ Quarter.begin                : chr  "1/1" "4/1" "1/1" "4/1" ...
##  $ Quarter.Begin.Date           : chr  "01/01/2016 12:00:00 AM" "04/01/2016 12:00:00 AM" "0
1/01/2016 12:00:00 AM" "04/01/2016 12:00:00 AM" ...
##  $ X_latitude                   : num  61.4 61.4 61.4 61.4 61.4 ...
##  $ X_longitude                  : num  -152 -152 -152 -152 -152 ...
##  $ Location                     : chr  "(61.3850, -152.2683)" "(61.3850, -152.2683)" "(61.38
50, -152.2683)" "(61.3850, -152.2683)" ...
##  $ NDC                          : num  2143380 2143380 2143480 2143480 2197590 ...
```

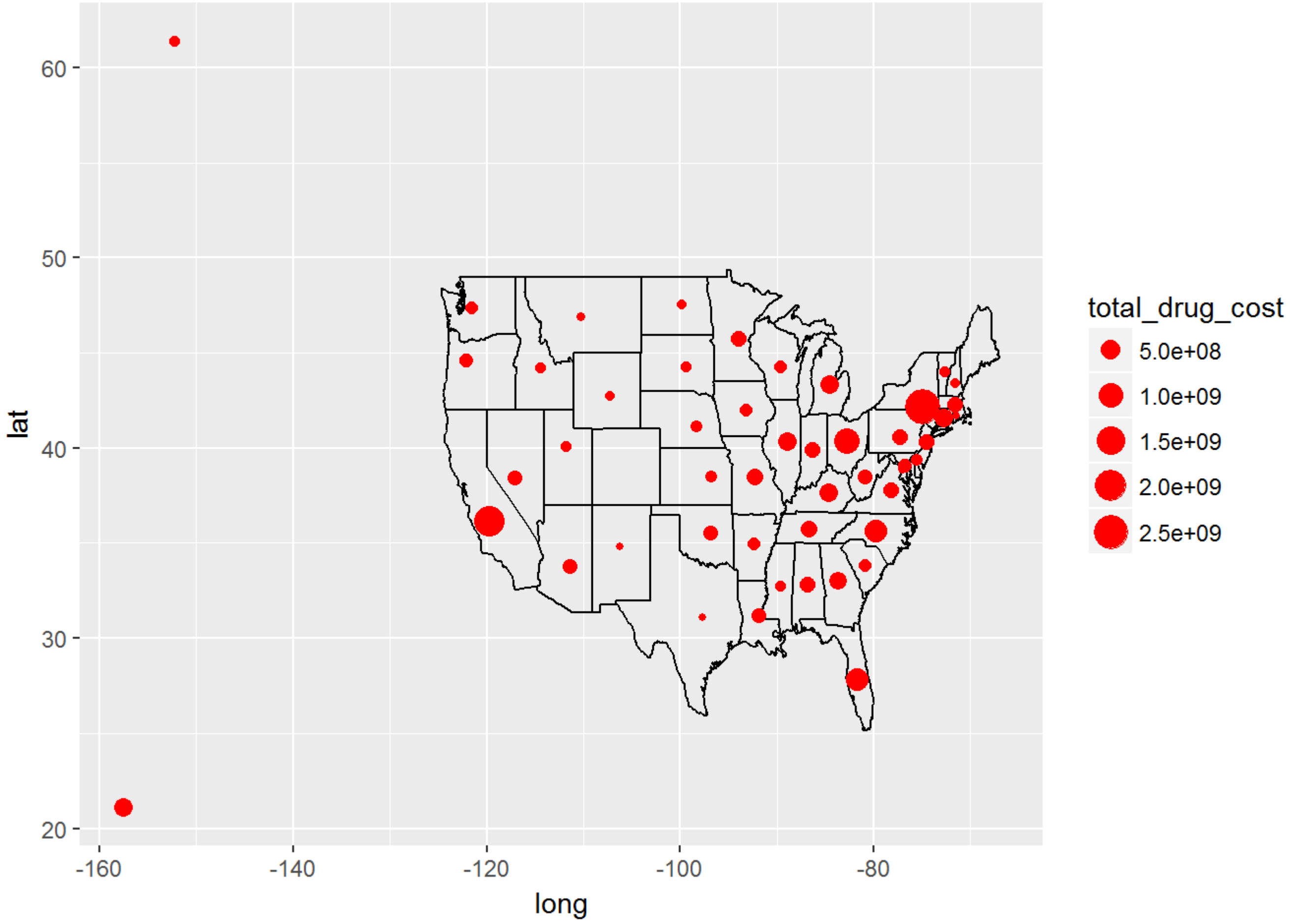
I subset the data and extract the geographic information of each individual state as well as the total drug cost for that state.

```
state_drug_cost <- raw_data1%>%
  group_by(State, X_latitude, X_longitude)%>%
  summarize(total_drug_cost = sum(Medicaid.Amount.Reimbursed, na.rm = TRUE))%>%
  filter(total_drug_cost != 0 & !is.na(X_latitude) & !is.na(X_longitude))
state_drug_cost
```

```
## Source: local data frame [51 x 4]
## Groups: State, X_latitude [51]
##
##    State X_latitude X_longitude total_drug_cost
##    <chr>      <dbl>       <dbl>           <dbl>
## 1     AK    61.3850   -152.2683        41488838
## 2     AL    32.7990    -86.8073       234814278
## 3     AR    34.9513    -92.3809       108603678
## 4     AZ    33.7712   -111.3877       159296822
## 5     CA    36.1700   -119.7462      1845314670
## 6     CT    41.5834    -72.7622       414550321
## 7     DC    38.8964    -77.0262         4551225
## 8     DE    39.3498    -75.5148        77059174
## 9     FL    27.8333    -81.7170       806509180
## 10    GA    32.9866    -83.6487       328901605
## # ... with 41 more rows
```

Created a visualization using the USA map. It gives us clear picture about prescription expenses from Medicaid program across the country. There are a few states stands out, such as NY, CA, and FL, which are the three states with the most population in america excluding TX. Just by eye balling this figure. NY's Medicaid seem to be the one with the highest prescription spending.

```
usa <- map_data("state")
ggplot() +
geom_path(data = usa, aes(x = long, y = lat, group = group)) +
geom_point(data = state_drug_cost, aes(x = X_longitude, y = X_latitude, size = total_drug_cost),
color = "red")
```



Next few line of codes is just about transformation of the data. I calculated the total product cost of each medication for each states(By the way, state "XX" means the entire country) and the number of that medications were dispense in that state. After we obtain these two numbers we can simply divide the two numbers to calculate the cost of the one unit of that medication.

For the sake of testing the difference in terms of each medication's cost across states level, Average cost for each unit of medication and its corresponding standard deviation for each states were also calculated.

```
total_cost <- raw_data1%>%
  group_by(State, Product.Name)%>%
  summarize(product_cost = sum(Medicaid.Amount.Reimbursed, na.rm = TRUE), count = sum(Units.Reim
bursed, na.rm = TRUE))%>%
  filter(product_cost != 0 & count != 0)%>%
  arrange(desc(product_cost))

total_cost <- total_cost%>%
  mutate(ave_drug_cost = product_cost / count, average = mean(ave_drug_cost), sd = sd(ave_drug_c
ost))
head(total_cost, 10)
```

```
## Source: local data frame [10 x 7]
## Groups: State [2]
##
##    State Product.Name product_cost      count ave_drug_cost   average
##    <chr>        <chr>        <dbl>      <dbl>         <dbl>     <dbl>
## 1     XX   HUMIRA 40     508154317   275517.9   1844.360648  79.93311
## 2     XX   LANTUS 100    370682730 14647604.3     25.306714  79.93311
## 3     XX   LANTUS 3ML    357558665 14406491.7     24.819274  79.93311
## 4     XX   SEROQUEL X    232025482 12703862.8     18.264168  79.93311
## 5     XX   SYMBICORT     219604001  8010996.3     27.412820  79.93311
## 6     XX   TRIUMEQ 50    198518884  2376716.0     83.526548  79.93311
## 7     NY   HARVONI  (    169696555   161171.0   1052.897578 136.76103
## 8     XX   INVEGASUST    162707518   117616.9   1383.368339  79.93311
## 9     XX   SUBOXONE 8    159894121 21730078.0      7.358194  79.93311
## 10    XX   ARIPIPRAZO    151011727 11117112.5     13.583719  79.93311
## # ... with 1 more variables: sd <dbl>
```

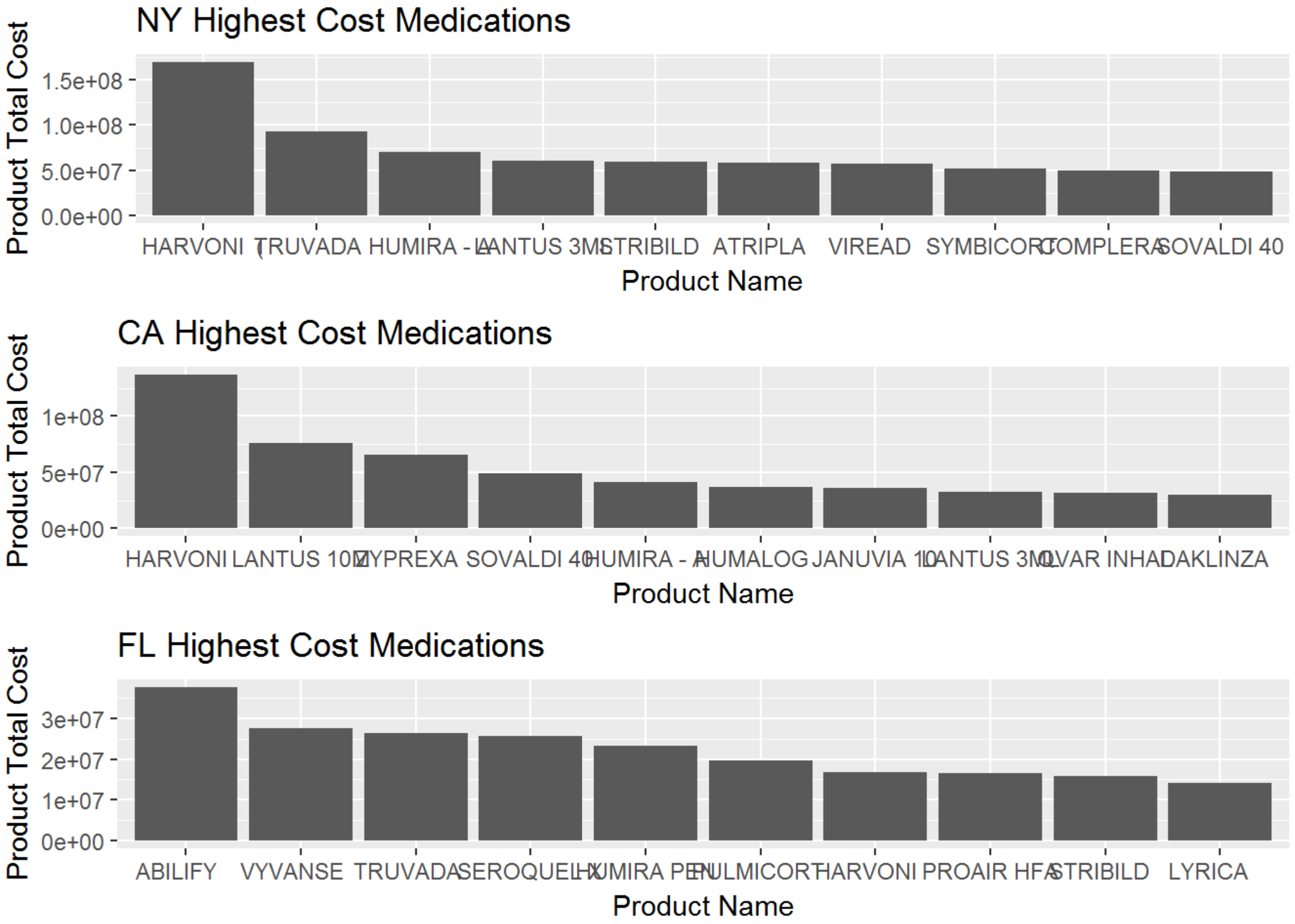
I just want to compare the drug cost for states where most americans live, I perform the test solely for CA, NY, and FL. The barplot shows that medications for HIV and HCV infections are usually most costly in state Medicaid agency's budget. It is reasonable because a lot of these medications have no generic available. In addition to that, they are life-saving medications. Therefore, even the cost is high, people will still have to pay for them. Medications for diabetes and respiratory disorders also have their spot on the highest cost medication list, probably because those medications are common, therefore, many of these prescriptions are filled nationwide.

```
NY_meds <- subset(total_cost, total_cost$State == "NY")
NY <- ggplot(head(NY_meds, 10), aes(x = reorder(Product.Name, -product_cost), y = product_cost))
+ geom_bar(stat = "identity") + ggtitle("NY Highest Cost Medications")+ xlab("Product Name") + y
lab("Product Total Cost")

CA_meds <- subset(total_cost, total_cost$State == "CA")
CA <- ggplot(head(CA_meds, 10), aes(x = reorder(Product.Name, -product_cost), y = product_cost))
+ geom_bar(stat = "identity") + ggtitle("CA Highest Cost Medications")+ xlab("Product Name") + y
lab("Product Total Cost")

FL_meds <- subset(total_cost, total_cost$State == "FL")
FL <- ggplot(head(FL_meds, 10), aes(x = reorder(Product.Name, -product_cost), y = product_cost))
+ geom_bar(stat = "identity") + ggtitle("FL Highest Cost Medications")+ xlab("Product Name") + y
lab("Product Total Cost")

grid.arrange(NY, CA, FL, nrow=3, ncol=1)
```



Furthermore, I created a data frame that is suitable for conducting two-way ANOVA test. ANOVA test will pretty much tell people if there are statistically significant differences among the mean of the response variable. Explanatory variable in this case is states.

```
NY_meds <- NY_meds%>%
  mutate(ave_drug_cost = product_cost / count, mean = mean(ave_drug_cost), sd = sd(ave_drug_cost
))
head(NY_meds)
```

```
## Source: local data frame [6 x 8]
## Groups: State [1]
##
##   State Product.Name product_cost      count ave_drug_cost average
##   <chr>        <chr>        <dbl>      <dbl>         <dbl>   <dbl>
## 1    NY   HARVONI  (    169696555  161171.00    1052.89758 136.761
## 2    NY   TRUVADA        92900282 1906729.03      48.72233 136.761
## 3    NY   HUMIRA - A     70177869   39304.97    1785.47062 136.761
## 4    NY   LANTUS 3ML     60894786 2467884.40      24.67489 136.761
## 5    NY   STRIBILD       59398979  662728.00      89.62799 136.761
## 6    NY   ATRIPLA        58606706  739374.00      79.26531 136.761
## # ... with 2 more variables: sd <dbl>, mean <dbl>
```

```
CA_meds <- CA_meds%>%
  mutate(ave_drug_cost = product_cost / count, mean = mean(ave_drug_cost), sd = sd(ave_drug_cost
))
head(CA_meds)
```

```
## Source: local data frame [6 x 8]
## Groups: State [1]
##
##   State Product.Name product_cost      count ave_drug_cost  average
##   <chr>        <chr>        <dbl>      <dbl>         <dbl>    <dbl>
## 1    CA   HARVONI       137430073  123214.00    1115.37709 118.2146
## 2    CA   LANTUS 10M     75606284 3036489.90      24.89924 118.2146
## 3    CA   ZYPREXA        65444504 3199537.00      20.45437 118.2146
## 4    CA   SOVALDI 40     48943074   49476.00     989.22860 118.2146
## 5    CA   HUMIRA - A     41057726   22306.87    1840.58684 118.2146
## 6    CA   HUMALOG        36853379 1586498.60      23.22938 118.2146
## # ... with 2 more variables: sd <dbl>, mean <dbl>
```

```
FL_meds <- FL_meds%>%
  mutate(ave_drug_cost = product_cost / count, mean = mean(ave_drug_cost), sd = sd(ave_drug_cost
))
head(FL_meds)
```

```
## Source: local data frame [6 x 8]
## Groups: State [1]
##
##   State Product.Name product_cost     count ave_drug_cost  average
##   <chr>        <chr>        <dbl>     <dbl>         <dbl>    <dbl>
## 1    FL   ABILIFY        37695603 1207029.0     31.230073 160.1519
## 2    FL   VYVANSE        27624579 3399028.0      8.127199 160.1519
## 3    FL   TRUVADA        26376180  541912.0     48.672436 160.1519
## 4    FL   SEROQUEL X     25712900 1375687.0     18.690953 160.1519
## 5    FL   HUMIRA PEN     23334504   12591.5   1853.194932 160.1519
## 6    FL   PULMICORT      19577216 4007985.0      4.884553 160.1519
## # ... with 2 more variables: sd <dbl>, mean <dbl>
```

```
anova_df <- rbind(NY_meds[,c(1, 5)], CA_meds[,c(1, 5)], FL_meds[,c(1, 5)])
head(anova_df)
```

```
## Source: local data frame [6 x 2]
## Groups: State [1]
##
##   State ave_drug_cost
##   <chr>         <dbl>
## 1    NY    1052.89758
## 2    NY      48.72233
## 3    NY    1785.47062
## 4    NY      24.67489
## 5    NY      89.62799
## 6    NY      79.26531
```

The result of the test shows even though the average medication cost are varid among these three states. The differences are not statistically significant, because the p-value is 0.497 which is above 0.05 (significance level). Also, the confidence interval across 0.

```
drug_aov <- aov(anova_df$ave_drug_cost ~ anova_df$State)
summary(drug_aov)
```

```
##                  Df    Sum Sq Mean Sq F value Pr(>F)
## anova_df$State    2 1.513e+06  756391   0.699  0.497
## Residuals      6514 7.051e+09 1082431
```

```
TukeyHSD(drug_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = anova_df$ave_drug_cost ~ anova_df$State)
##
## $`anova_df$State`
##            diff        lwr       upr     p adj
## FL-CA  41.93726  -42.30162 126.17613 0.4729805
## NY-CA  18.54642  -48.62333  85.71616 0.7939182
## NY-FL -23.39084 -107.06951  60.28784 0.7893701
```

```
confint(drug_aov)
```

```
##                      2.5 %    97.5 %
## (Intercept)       78.08484 158.34438
## anova_df$StateFL -28.50605 112.38056
## anova_df$StateNY -37.62312  74.71596
```

5. Conclusion. In general, the expenses of hospital episode by Medicare differs quite a lot. Some states are around 9% above the national average, while the other state could be about 7% below national average. By

breaking down the expense by different claim type, we can usually detect where the discrepancies are. For example, from the NY and CA claim cost comparison, we know it is the hospice care that account for most of the differences between the two states. For the sake of time, I could not perform the similar analysis among other states. Another point that arise after analyzing the drug cost data from Medicaid program is that generally speaking, if a state has high proportion of HIV, HCV patients, the state medicaid program will have to reimburse more for the corresponding prescription. This is phenomenon is manifested in the national aggreagate data also the state specific data. Moreover, if we want to compare the prescription cost among different states, It is very difficult to establish statistically significant conclusion, even though the sample data seem to show differences in terms of average drug cost. With over 1 million observations, we still can not claim prescriptions sold in NY is cheaper that sold in FL, although the NY does have sample mean that is about 24 dollars cheaper. More investigation will be undergoing to gain more insight on this issue.