US Medicare and payment charges

```
chooseCRANmirror(graphics=FALSE, ind=1)
```

```
## Warning in download.file(url, destfile = f, quiet = TRUE): InternetOpenUrl
## failed: 'A connection with the server could not be established'
```

```
## Warning: failed to download mirrors file (cannot open URL 'https://cran.r-
## project.org/CRAN_mirrors.csv'); using local file 'C:/PROGRA~1/R/R-34~1.1/
## doc/CRAN_mirrors.csv'
```

```
knitr::opts_chunk$set(echo = TRUE)
```

```
#Acquire process
#Load the US inpatient hospital data
setwd("~/Guru/Pers/BIG DATA/assignments")
hospi = read.csv('inpatient hospital data.csv')
str(hospi)
```

```
## 'data.frame':    163065 obs. of  12 variables:
##  $ DRG.Definition                     : Factor w/ 100 levels "039 - EXTRACRANIAL PROCEDURES W/O CC/
##  $ Provider.Id                        : int  10001 10005 10006 10011 10016 10023 10029 10033 10039
##  $ Provider.Name                      : Factor w/ 3201 levels "ABBEVILLE GENERAL HOSPITAL",..: 2519
##  $ Provider.Street.Address            : Factor w/ 3326 levels "#1 MEDICAL PARK DRIVE",..: 319 1494
##  $ Provider.City                      : Factor w/ 1977 levels "ABBEVILLE","ABERDEEN",..: 455 178 583
##  $ Provider.State                     : Factor w/ 51 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2 2 2
##  $ Provider.Zip.Code                  : int  36301 35957 35631 35235 35007 36116 36801 35233 35801
##  $ Hospital.Referral.Region.Description: Factor w/ 306 levels "AK - Anchorage",..: 3 2 2 2 2 6 2 2 4
##  $ Total.Discharges                   : int  91 14 24 25 18 67 51 32 135 34 ...
##  $ Average.Covered.Charges            : Factor w/ 160236 levels "$10,000.36 ",..: 97819 29922 10741
##  $ Average.Total.Payments             : Factor w/ 147842 levels "$10,000.05 ",..: 93533 93742 87031
##  $ Average.Medicare.Payments          : Factor w/ 150328 levels "$1,148.90 ","$1,327.23 ",..: 85751
```

```
dim(hospi)
```

```
## [1] 163065     12
```

```
library(readr)
library(proto)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(stringr)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
#Refine process
#The column names have spaces. Let's just rename them all
names(hospi) <- c('drg_def', 'prov_id', 'prov_name', 'prov_address', 'prov_city', 'prov_state', 'prov_z
```

```
# We need to get rid of the dollar sings in the charges and payments columns and convert to numeric
hospi$mean_total_payments = hospi$mean_total_payments %>% str_replace("\\$", "")
hospi$mean_total_payments = hospi$mean_total_payments %>% str_replace("\\," , "")
str(hospi)
```

```
## 'data.frame':    163065 obs. of  12 variables:
##  $ drg_def               : Factor w/ 100 levels "039 - EXTRACRANIAL PROCEDURES W/O CC/MCC",..: 1 1 1
##  $ prov_id               : int  10001 10005 10006 10011 10016 10023 10029 10033 10039 10040 ...
##  $ prov_name             : Factor w/ 3201 levels "ABBEVILLE GENERAL HOSPITAL",..: 2519 1499 730 2714
##  $ prov_address          : Factor w/ 3326 levels "#1 MEDICAL PARK DRIVE",..: 319 1494 1230 2291 87 1
##  $ prov_city             : Factor w/ 1977 levels "ABBEVILLE","ABERDEEN",..: 455 178 583 163 14 1152
##  $ prov_state            : Factor w/ 51 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ prov_zip              : int  36301 35957 35631 35235 35007 36116 36801 35233 35801 35903 ...
##  $ referral_reg          : Factor w/ 306 levels "AK - Anchorage",..: 3 2 2 2 2 6 2 2 4 2 ...
##  $ total_discharges      : int  91 14 24 25 18 67 51 32 135 34 ...
##  $ mean_covered_charges  : Factor w/ 160236 levels "$10,000.36 ",..: 97819 29922 107411 22768 94638 3
##  $ mean_total_payments   : chr  "5777.24 " "5787.57 " "5434.95 " "5417.56 " ...
##  $ mean_medicare_payments: Factor w/ 150328 levels "$1,148.90 ","$1,327.23 ",..: 85751 89901 79721 73
```

```
hospi$mean_total_payments = as.numeric(hospi$mean_total_payments)
hospi$mean_covered_charges = hospi$mean_covered_charges %>% str_replace("\\$", "")
hospi$mean_covered_charges = hospi$mean_covered_charges %>% str_replace("\\," , "")
hospi$mean_covered_charges = as.numeric(hospi$mean_covered_charges)
hospi$mean_medicare_payments = hospi$mean_medicare_payments %>% str_replace("\\$", "")
hospi$mean_medicare_payments= hospi$mean_medicare_payments %>% str_replace("\\," , "")
hospi$mean_medicare_payments = as.numeric(hospi$mean_medicare_payments)
str(hospi)
```

```
## 'data.frame':    163065 obs. of  12 variables:
##  $ drg_def               : Factor w/ 100 levels "039 - EXTRACRANIAL PROCEDURES W/O CC/MCC",..: 1 1 1
##  $ prov_id               : int  10001 10005 10006 10011 10016 10023 10029 10033 10039 10040 ...
##  $ prov_name             : Factor w/ 3201 levels "ABBEVILLE GENERAL HOSPITAL",..: 2519 1499 730 2714
##  $ prov_address          : Factor w/ 3326 levels "#1 MEDICAL PARK DRIVE",..: 319 1494 1230 2291 87 1
##  $ prov_city             : Factor w/ 1977 levels "ABBEVILLE","ABERDEEN",..: 455 178 583 163 14 1152
##  $ prov_state            : Factor w/ 51 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ prov_zip              : int  36301 35957 35631 35235 35007 36116 36801 35233 35801 35903 ...
##  $ referral_reg          : Factor w/ 306 levels "AK - Anchorage",..: 3 2 2 2 2 6 2 2 4 2 ...
##  $ total_discharges      : int  91 14 24 25 18 67 51 32 135 34 ...
##  $ mean_covered_charges  : num  32963 15132 37560 13998 31633 ...
##  $ mean_total_payments   : num  5777 5788 5435 5418 5658 ...
##  $ mean_medicare_payments: num  4764 4977 4454 4129 4851 ...
```

```
head(hospi)
```

```
##                                  drg_def prov_id
## 1 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC   10001
## 2 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC   10005
## 3 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC   10006
## 4 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC   10011
## 5 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC   10016
## 6 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC   10023
##                         prov_name              prov_address  prov_city
## 1 SOUTHEAST ALABAMA MEDICAL CENTER     1108 ROSS CLARK CIRCLE     DOTHAN
## 2    MARSHALL MEDICAL CENTER SOUTH 2505 U S HIGHWAY 431 NORTH       BOAZ
## 3   ELIZA COFFEE MEMORIAL HOSPITAL         205 MARENGO STREET   FLORENCE
```

```
## 4               ST VINCENT'S EAST 50 MEDICAL PARK EAST DRIVE BIRMINGHAM
## 5    SHELBY BAPTIST MEDICAL CENTER    1000 FIRST STREET NORTH  ALABASTER
## 6     BAPTIST MEDICAL CENTER SOUTH  2105 EAST SOUTH BOULEVARD MONTGOMERY
##   prov_state prov_zip    referral_reg total_discharges
## 1         AL    36301     AL - Dothan               91
## 2         AL    35957 AL - Birmingham               14
## 3         AL    35631 AL - Birmingham               24
## 4         AL    35235 AL - Birmingham               25
## 5         AL    35007 AL - Birmingham               18
## 6         AL    36116 AL - Montgomery               67
##   mean_covered_charges mean_total_payments mean_medicare_payments
## 1             32963.07             5777.24                4763.73
## 2             15131.85             5787.57                4976.71
## 3             37560.37             5434.95                4453.79
## 4             13998.28             5417.56                4129.16
## 5             31633.27             5658.33                4851.44
## 6             16920.79             6653.80                5374.14
```

```
#Transform process
#mean average provider coverage charges  by state
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
```

```
by_state <- hospi %>% group_by(prov_state) %>%
  summarise(mean=mean(mean_covered_charges)) %>% arrange(desc(mean))
head(by_state, 10)
```

```
## # A tibble: 10 x 2
##    prov_state     mean
##        <fctr>    <dbl>
##  1         CA 67508.62
##  2         NJ 66125.69
##  3         NV 61047.12
##  4         FL 46016.23
##  5         TX 41480.19
##  6         AZ 41200.06
##  7         CO 41095.14
##  8         AK 40348.74
##  9         DC 40116.66
## 10         PA 39633.96
```

```
## Cheapest Diagnosis in each state as per Total Payment
hospi %>% group_by(prov_state) %>% filter(mean_total_payments == min(mean_total_payments)) %>% do(head(
```

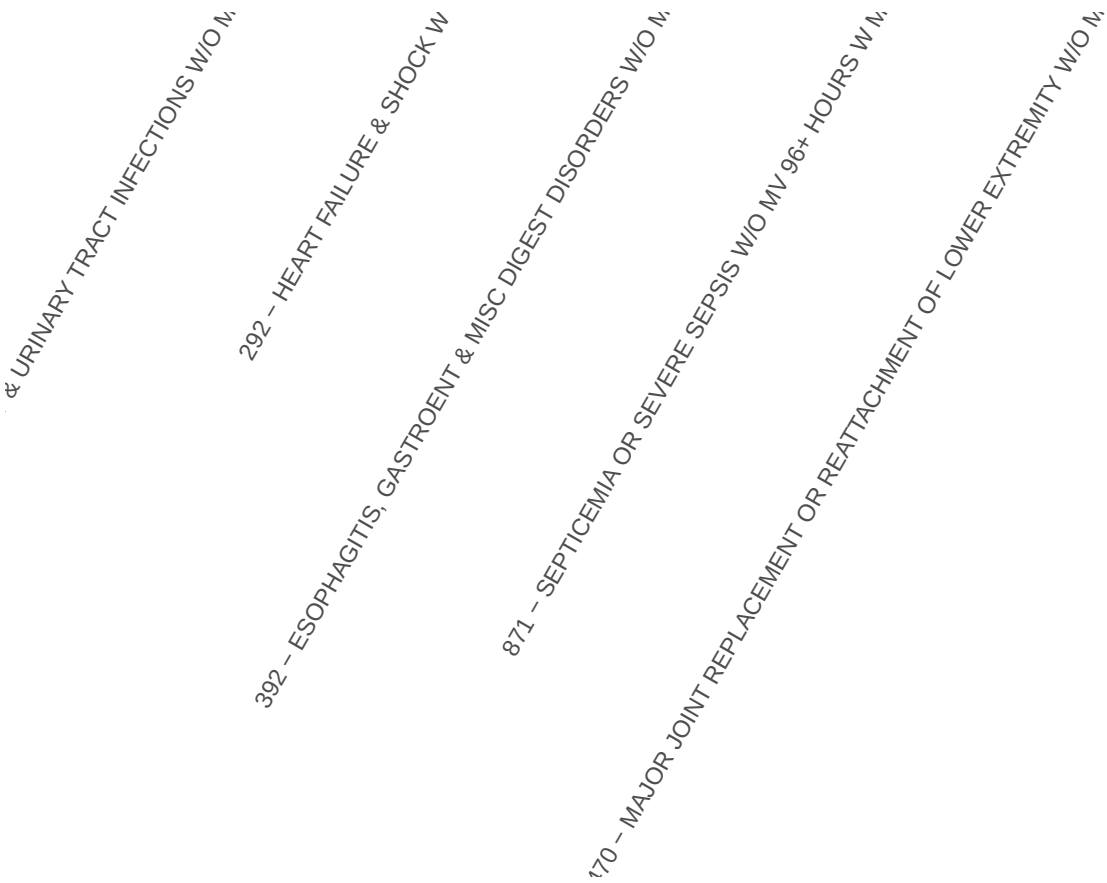
```
## # A tibble: 51 x 3
## # Groups:   prov_state [51]
##    prov_state          drg_def mean_total_payments
```

```
##        <fctr>           <fctr>              <dbl>
##  1         AK 313 - CHEST PAIN            4717.04
##  2         AL 313 - CHEST PAIN            2682.64
##  3         AR 313 - CHEST PAIN            2845.78
##  4         AZ 313 - CHEST PAIN            3002.00
##  5         CA 313 - CHEST PAIN            3465.00
##  6         CO 313 - CHEST PAIN            3171.91
##  7         CT 313 - CHEST PAIN            3605.00
##  8         DC 313 - CHEST PAIN            3247.18
##  9         DE 313 - CHEST PAIN            3562.13
## 10         FL 313 - CHEST PAIN            2769.00
## # ... with 41 more rows
```

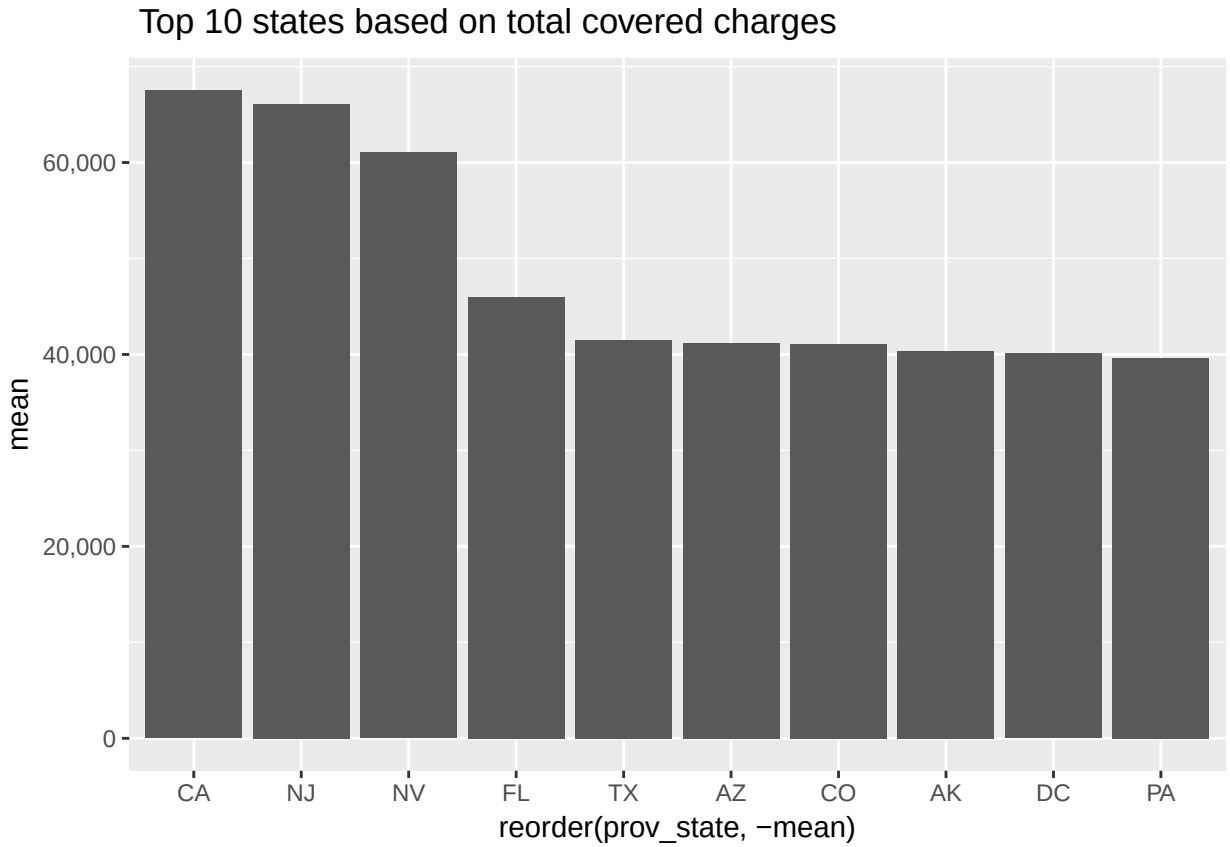
```
# Top 5 medical conditions by discharges
IPC <- hospi %>%
  select(drg_def, prov_id, total_discharges, mean_covered_charges,mean_total_payments, mean_medicare_pay
  group_by(drg_def) %>%
  summarise(total_discharges =sum(total_discharges), mean_covered_charges=mean(mean_covered_charges), me
  arrange(desc(total_discharges)) %>%
  top_n(5, total_discharges)
IPC
```

```
## # A tibble: 5 x 5
##                                                                  drg_def
##                                                                   <fctr>
## 1 470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC
## 2                 871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC
## 3             392 - ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC
## 4                                       292 - HEART FAILURE & SHOCK W CC
## 5                          690 - KIDNEY & URINARY TRACT INFECTIONS W/O MCC
## # ... with 4 more variables: total_discharges <int>,
## #   mean_covered_charges <dbl>, mean_medicare_payments <dbl>,
## #   mean_total_payments <dbl>
```

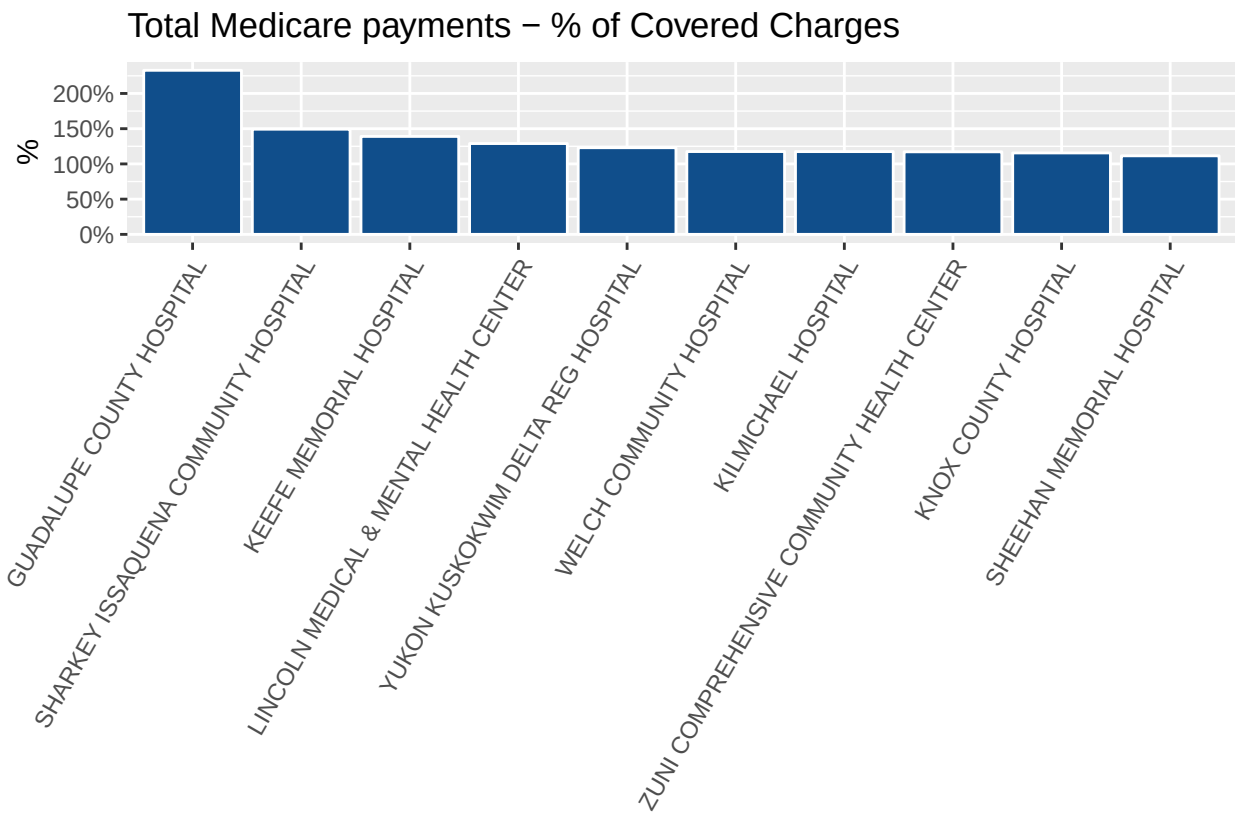
```
# Explore process
p1 = ggplot(IPC) +aes(reorder(drg_def, total_discharges), weight = total_discharges/1000) +geom_bar() +
print(p1)
```




```
# Top 10 states based on covered charges
ggplot(by_state[1:10,], aes(reorder(prov_state, -mean), mean)) + geom_bar(stat = "identity") +scale_y_c
```




```
# Which hospital have highest ratio of medicare payment to total charges
hospi %>%
    mutate(payments_to_charges = mean_medicare_payments / mean_covered_charges) %>%
    group_by(prov_name) %>%
    summarize(m = mean(payments_to_charges)) %>%
    arrange(-m) %>%
    head(10) %>%
    ggplot(aes(x=reorder(prov_name, -m), y = m)) +
        geom_bar(stat = 'identity', fill = 'dodgerblue4', color = 'white') +
        labs(x = '', y = '%', title = 'Total Medicare payments - % of Covered Charges') +
        scale_y_continuous(labels = scales::percent) +
theme(axis.text.x = element_text(angle = 60, hjust = 1) )
```

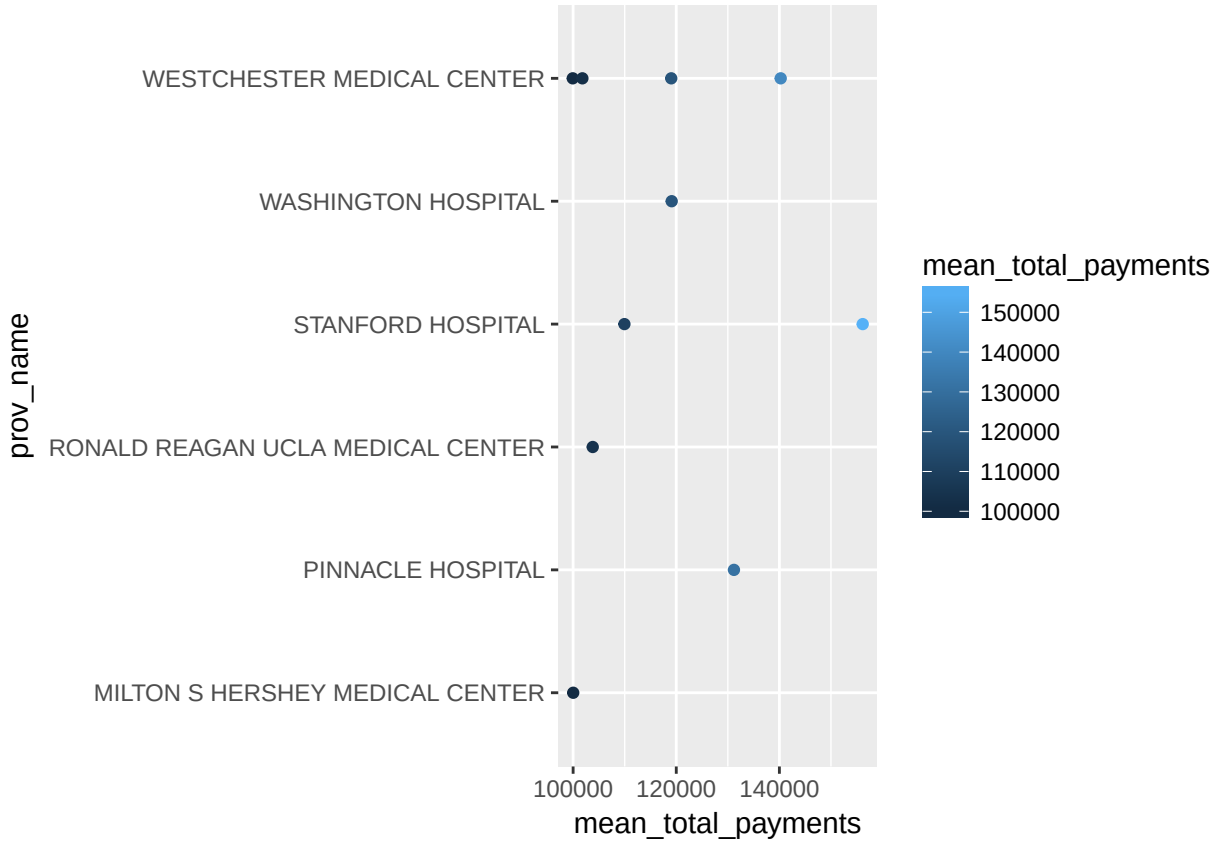



```
# Average total payment by state
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
hospi = as.data.table(hospi)
AVTPS = hospi[, mean_total_payments, by= prov_name] %>% top_n(10, mean_total_payments)
ggplot(data = AVTPS, mapping = aes(y = prov_name, x = mean_total_payments, colour = mean_total_payments
```




```
#Model process
#Develop linear model to understand statistics summary between medicare payment and total discharge
model <- lm(formula=mean_medicare_payments ~ total_discharges,data = hospi)
summary(model)
```

```
##
## Call:
## lm(formula = mean_medicare_payments ~ total_discharges, data = hospi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7442  -4310  -2330   1582 146030
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8624.8058    23.6002 365.455   <2e-16 ***
## total_discharges   -3.0464     0.3541  -8.603   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7308 on 163063 degrees of freedom
## Multiple R-squared:  0.0004536,  Adjusted R-squared:  0.0004475
## F-statistic: 74.01 on 1 and 163063 DF,  p-value: < 2.2e-16
```

```
d = data.frame(hospi$mean_covered_charges, hospi$prov_id)
aov(hospi$prov_id ~ hospi$mean_covered_charges, data = d) ->av
summary(av)
```

```
##                                Df    Sum Sq   Mean Sq F value Pr(>F)
## hospi$mean_covered_charges      1 4.740e+13 4.740e+13    2090 <2e-16 ***
## Residuals                  163063 3.698e+15 2.268e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```