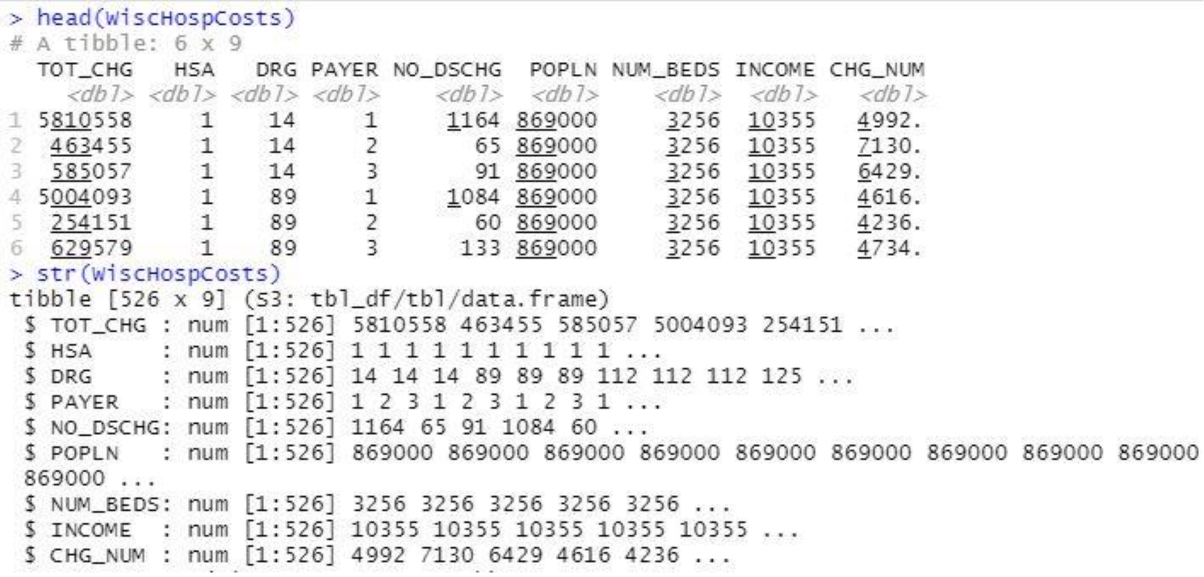
SMR - 3rd ESE - Hemanath Kumar J (21121015)

Outputs & Insights

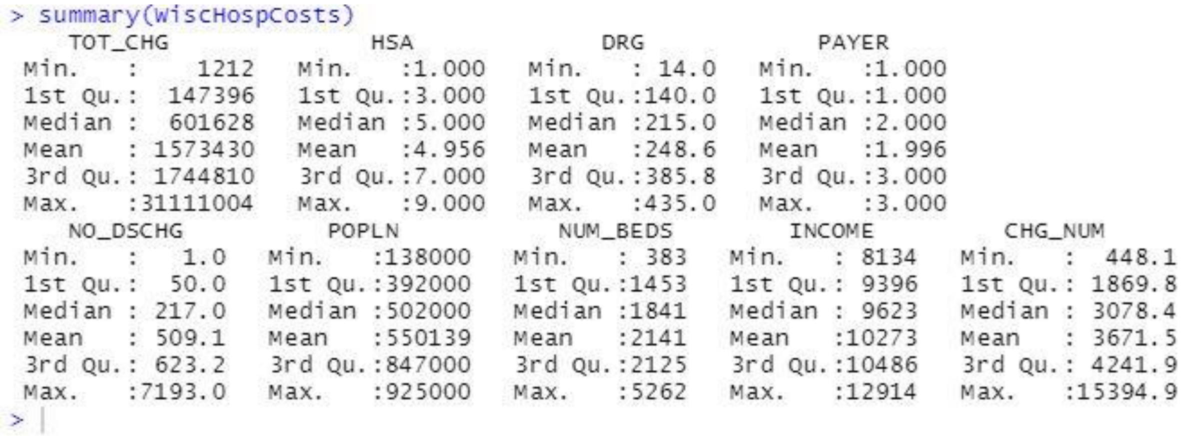
Q1) Descriptive analysis to gain basic insights into the variables



```
> head(WiscHospCosts)
# A tibble: 6 x 9
  TOT_CHG   HSA   DRG PAYER NO_DSCHG  POPLN NUM_BEDS INCOME CHG_NUM
    <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl>    <dbl>  <dbl>   <dbl>
1 5810558     1    14     1     1164 869000     3256  10355   4992.
2  463455     1    14     2       65 869000     3256  10355   7130.
3  585057     1    14     3       91 869000     3256  10355   6429.
4 5004093     1    89     1     1084 869000     3256  10355   4616.
5  254151     1    89     2       60 869000     3256  10355   4236.
6  629579     1    89     3      133 869000     3256  10355   4734.
> str(WiscHospCosts)
tibble [526 x 9] (S3: tbl_df/tbl/data.frame)
 $ TOT_CHG : num [1:526] 5810558 463455 585057 5004093 254151 ...
 $ HSA     : num [1:526] 1 1 1 1 1 1 1 1 1 1 ...
 $ DRG     : num [1:526] 14 14 14 89 89 89 112 112 112 125 ...
 $ PAYER   : num [1:526] 1 2 3 1 2 3 1 2 3 1 ...
 $ NO_DSCHG: num [1:526] 1164 65 91 1084 60 ...
 $ POPLN   : num [1:526] 869000 869000 869000 869000 869000 869000 869000 869000 869000
 869000 ...
 $ NUM_BEDS: num [1:526] 3256 3256 3256 3256 3256 ...
 $ INCOME  : num [1:526] 10355 10355 10355 10355 10355 ...
 $ CHG_NUM : num [1:526] 4992 7130 6429 4616 4236 ...
```

Observations:

- From head() we can see the first 6 rows in the dataset and the variables name, which gives the basic idea about the dataset and the datatypes.
- Str() is used to know the structure of the dataset that is imported, which is displayed. Through this, we can know the class of each column.

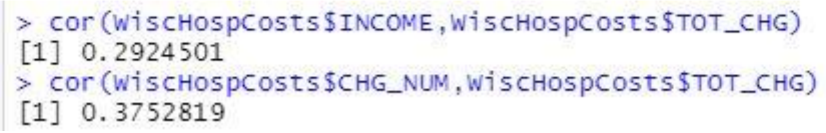


```
> summary(WiscHospCosts)
    TOT_CHG               HSA             DRG             PAYER      
 Min.   :    1212   Min.   :1.000   Min.   : 14.0   Min.   :1.000  
 1st Qu.:  147396   1st Qu.:3.000   1st Qu.:140.0   1st Qu.:1.000  
 Median :  601628   Median :5.000   Median :215.0   Median :2.000  
 Mean   : 1573430   Mean   :4.956   Mean   :248.6   Mean   :1.996  
 3rd Qu.: 1744810   3rd Qu.:7.000   3rd Qu.:385.8   3rd Qu.:3.000  
 Max.   :31111004   Max.   :9.000   Max.   :435.0   Max.   :3.000  
    NO_DSCHG          POPLN           NUM_BEDS        INCOME         CHG_NUM       
 Min.   :   1.0   Min.   :138000   Min.   : 383   Min.   : 8134   Min.   :  448.1  
 1st Qu.:  50.0   1st Qu.:392000   1st Qu.:1453   1st Qu.: 9396   1st Qu.: 1869.8  
 Median : 217.0   Median :502000   Median :1841   Median : 9623   Median : 3078.4  
 Mean   : 509.1   Mean   :550139   Mean   :2141   Mean   :10273   Mean   : 3671.5  
 3rd Qu.: 623.2   3rd Qu.:847000   3rd Qu.:2125   3rd Qu.:10486   3rd Qu.: 4241.9  
 Max.   :7193.0   Max.   :925000   Max.   :5262   Max.   :12914   Max.   :15394.9  
> 
```

Observations:

- Summary() is used to gain some basic insights of the dataset and use it for further analysis.
- Here, we can see the minimum & maximum value, Mean and median of all the variables that are present in the dataset. It also gives the values of 1st quartile and 3rd quartile.

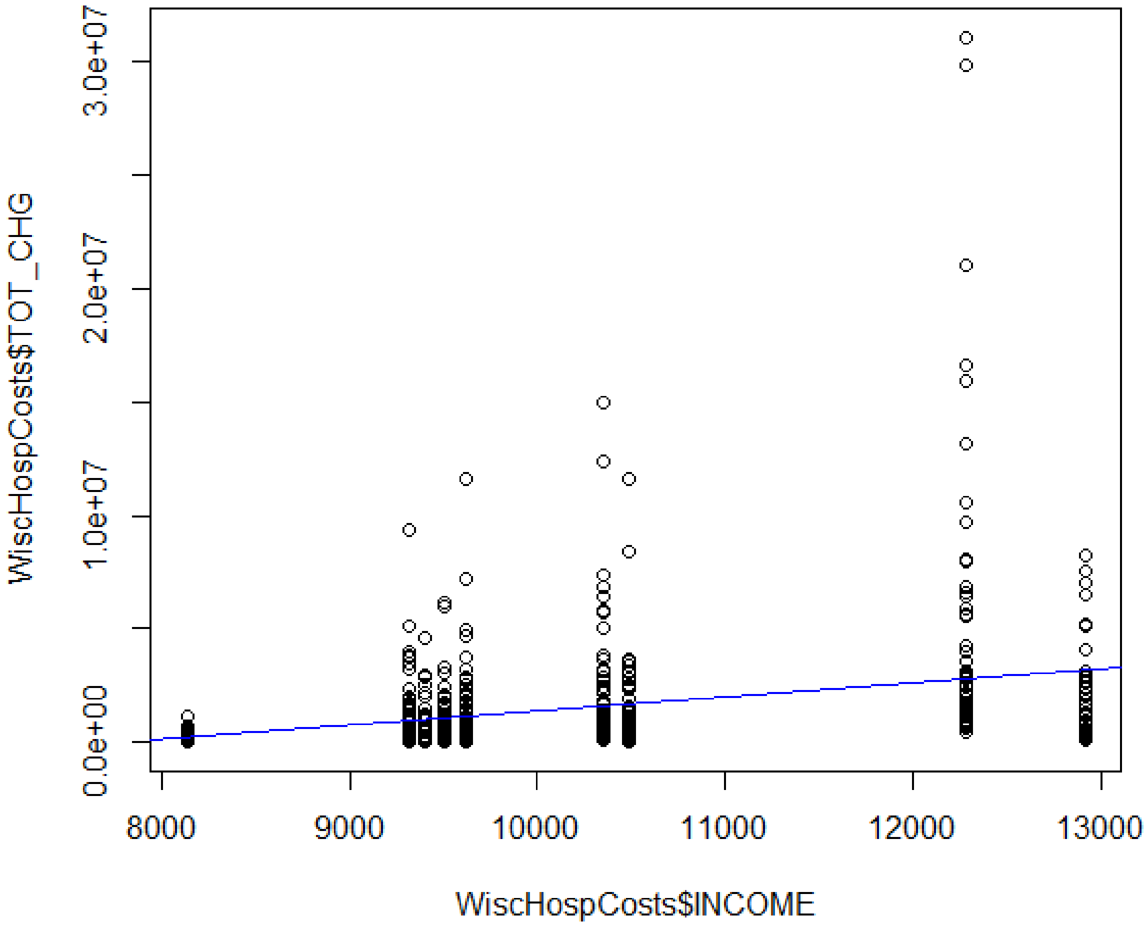
Q2) How hospital cost (TOT_CHG) is related to Income and Hospital discharge cost per patient (CHG_NUM)

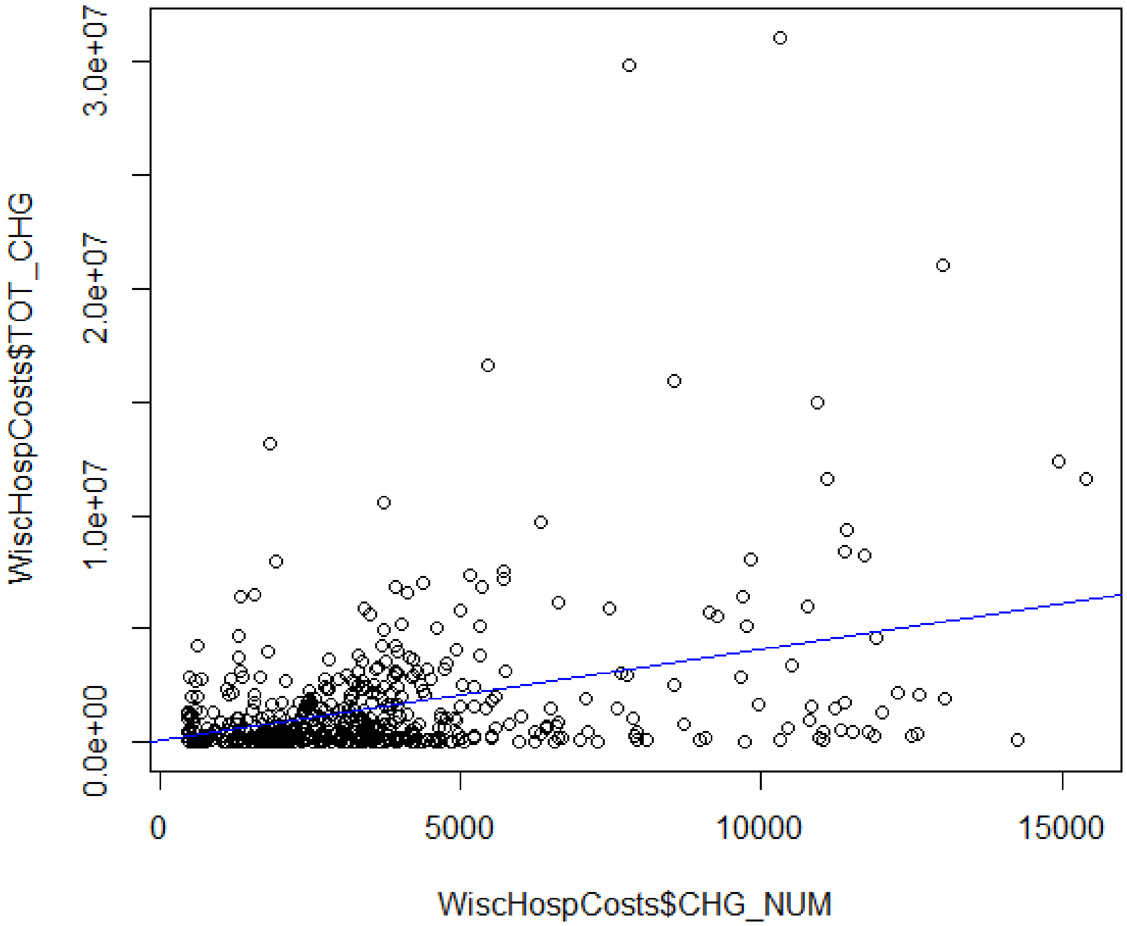


```
> cor(WiscHospCosts$INCOME,WiscHospCosts$TOT_CHG)
[1] 0.2924501
> cor(WiscHospCosts$CHG_NUM,WiscHospCosts$TOT_CHG)
[1] 0.3752819
```

Observations:

- From the output 1 we can see the correlation coefficient of income and total charge is 0.2924501. This denotes that the relation is positive but not as strong as hospital discharge costs and the relation is just around 29% which does not have much relation.
- From the output 2 the correlation coefficient of hospital discharge cost and total charge is 0.3752819, which is positive and has more relation compared to income, but it is around 37% which is less.

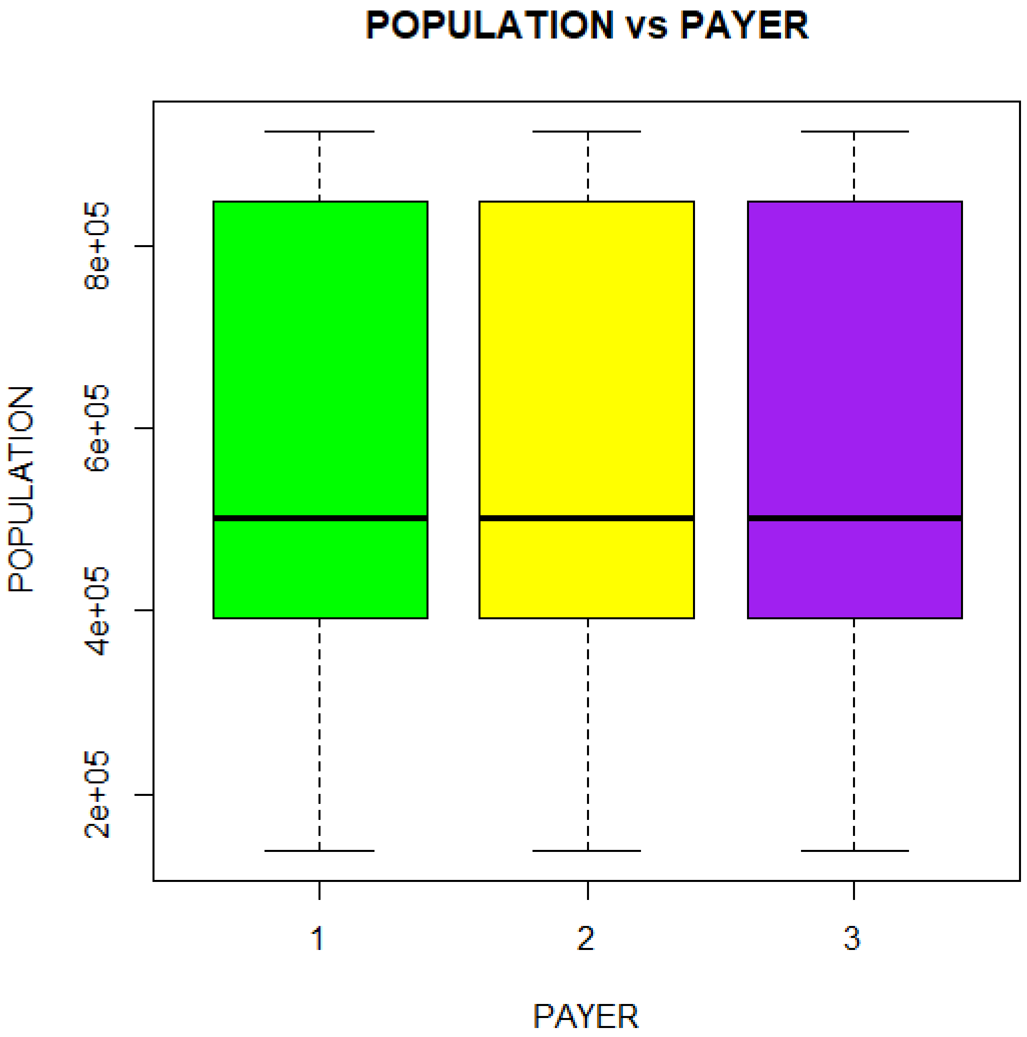

WiscHospCosts$TOT_CHG
0.0e+00
1.0e+07
2.0e+07
3.0e+07
8000
9000
10000
11000
12000
13000
WiscHospCosts$INCOME



Observations:

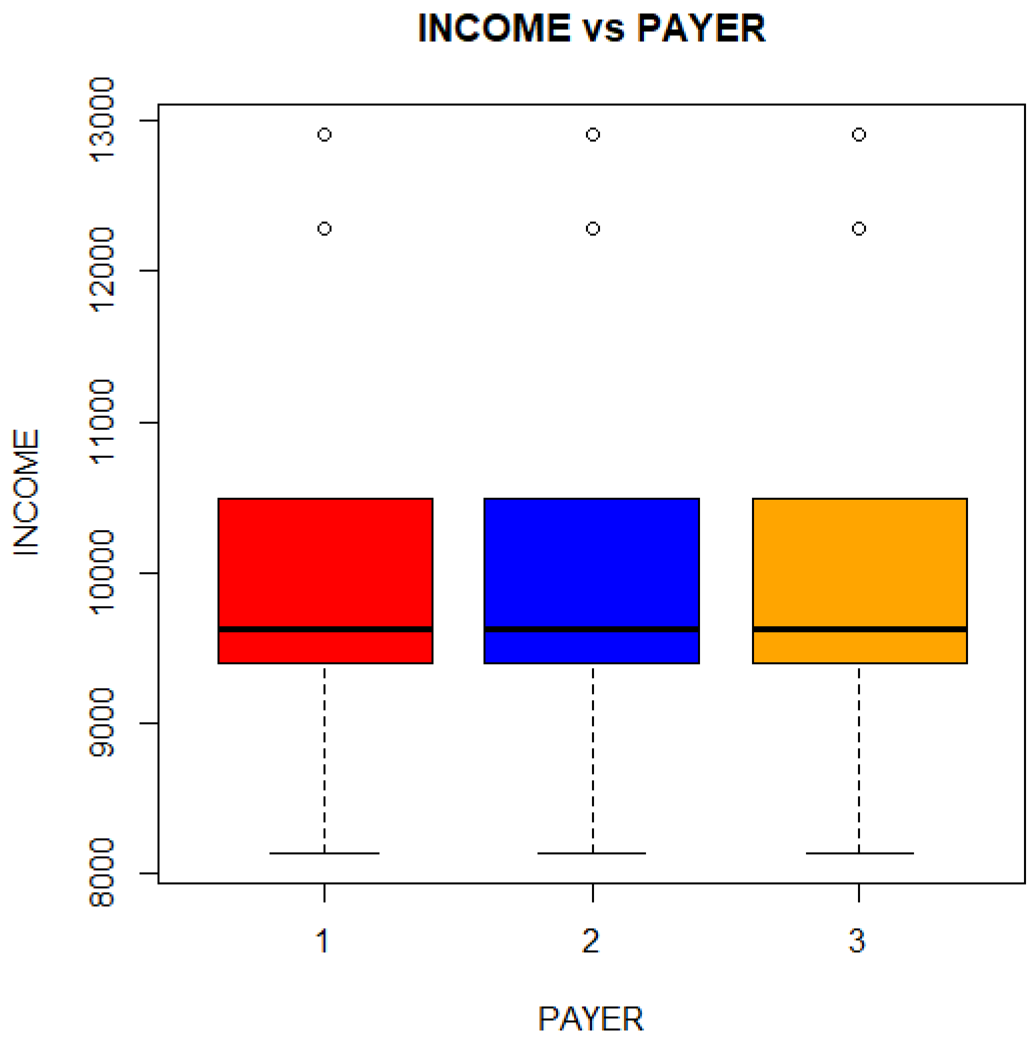
- From the scatter plot 1 we can see the data is very much dispersed between the income which denotes the frequency width is more in income.
- From the scatter plot 2 it is seen that the CHG_NUM and TOT_CHG has positive linear relationship between them.
- The trend line is drawn in both the graphs using abline() function. In the 2nd scatter plot, the trend line looks like a best fit and there are few outliers observed in both the scatter plots in higher range.

Q4) Distribution of area population and Income (both separately) based on type of payer.



Observation:

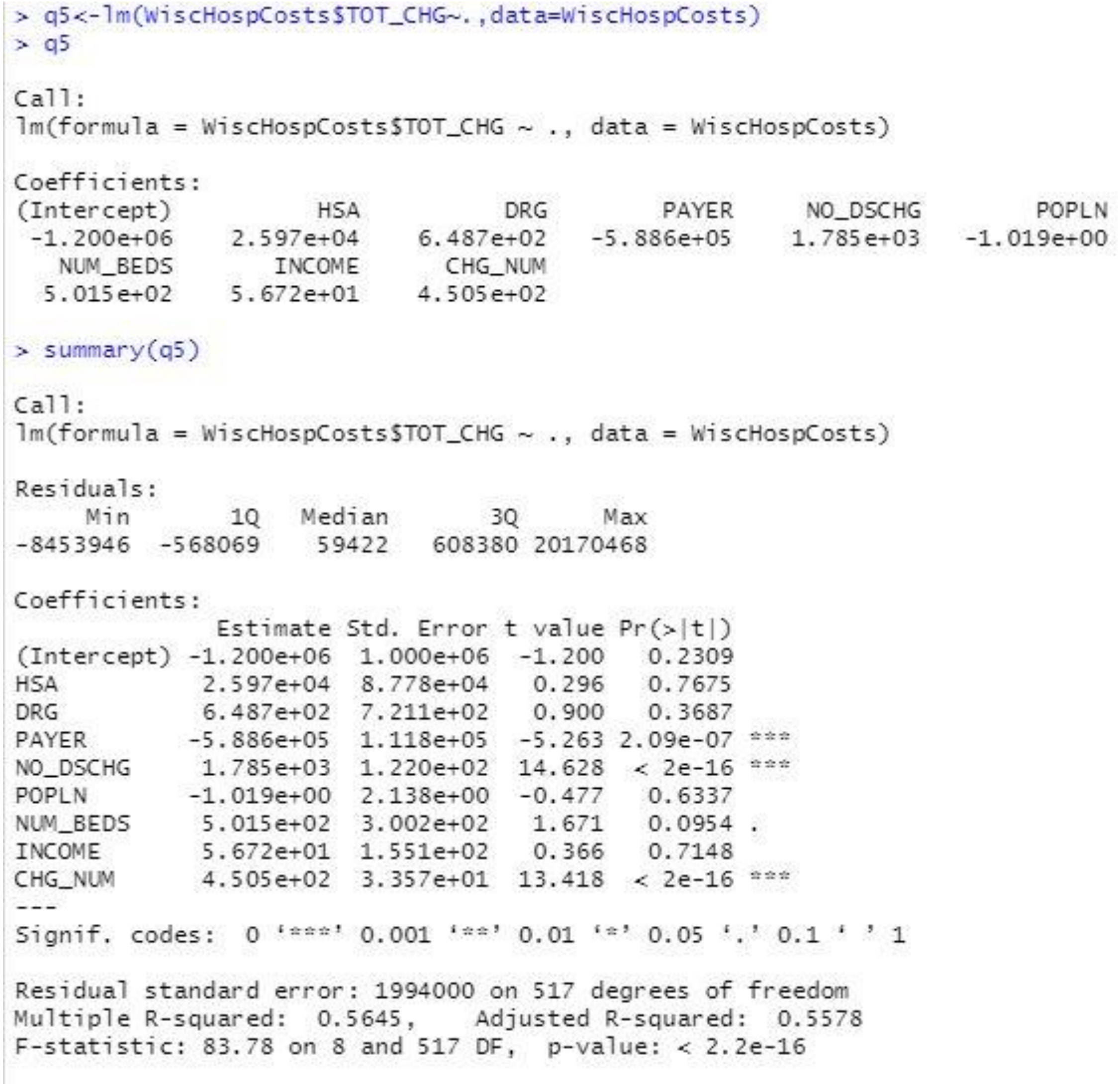
- Here, the box represents the 50% of the central data, with a line inside that represents the median. The median here is in between 4 and 6(which represents the amount of population). The data above the median is more dispersed as that is around 75%. The observations are similar to all the 3 payers. There is no much differences.
- There are no outliers found in all the 3 payer levels.



Observation:

- The median here is in between 9000 and 10000(which represents the income). The data above the median is more dispersed as that is around 75%. The observations are similar to all the 3 payers. There is no much differences.
- There are 2 outliers found in all the 3 payer levels which are present at the higher extreme.
- The data above the median is more dispersed compared to below.

Q5) Find the variable that mainly affects the hospital costs

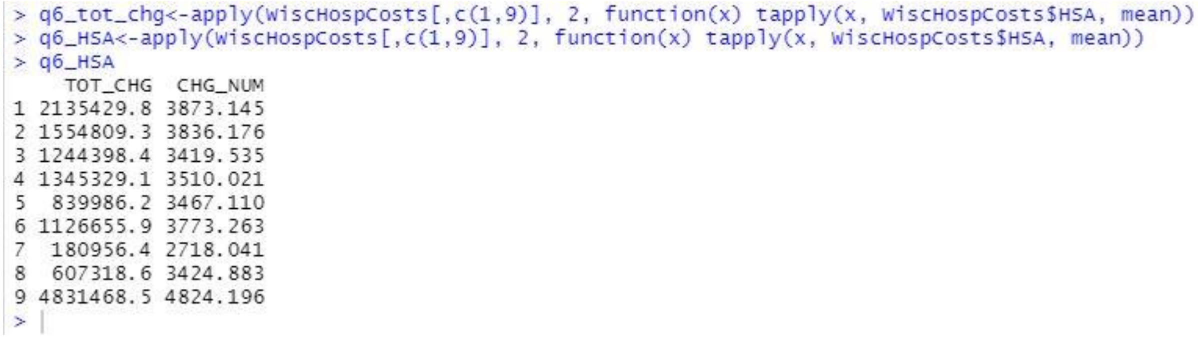


```
> q5<-lm(WiscHospCosts$TOT_CHG~.,data=WiscHospCosts)
> q5

Call:
lm(formula = WiscHospCosts$TOT_CHG ~ ., data = WiscHospCosts)

Coefficients:
(Intercept)          HSA          DRG        PAYER     NO_DSCHG        POPLN  
 -1.200e+06    2.597e+04    6.487e+02   -5.886e+05    1.785e+03   -1.019e+00  
   NUM_BEDS       INCOME      CHG_NUM  
  5.015e+02    5.672e+01    4.505e+02  

> summary(q5)

Call:
lm(formula = WiscHospCosts$TOT_CHG ~ ., data = WiscHospCosts)

Residuals:
     Min       1Q   Median       3Q      Max 
-8453946  -568069    59422   608380 20170468 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.200e+06  1.000e+06  -1.200   0.2309    
HSA          2.597e+04  8.778e+04   0.296   0.7675    
DRG          6.487e+02  7.211e+02   0.900   0.3687    
PAYER       -5.886e+05  1.118e+05  -5.263 2.09e-07 ***
NO_DSCHG     1.785e+03  1.220e+02  14.628  < 2e-16 ***
POPLN       -1.019e+00  2.138e+00  -0.477   0.6337    
NUM_BEDS     5.015e+02  3.002e+02   1.671   0.0954 .  
INCOME       5.672e+01  1.551e+02   0.366   0.7148    
CHG_NUM      4.505e+02  3.357e+01  13.418  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1994000 on 517 degrees of freedom
Multiple R-squared:  0.5645,	Adjusted R-squared:  0.5578 
F-statistic: 83.78 on 8 and 517 DF,  p-value: < 2.2e-16
```

Observation:

- The variables that mainly affects the hospital costs are NO_DSCHG and CHG_NUM which are number of patients discharged from the hospital and hospital discharged costs per patient.
- Whereas all other variables are just related and the above-mentioned variables are strongly affecting.

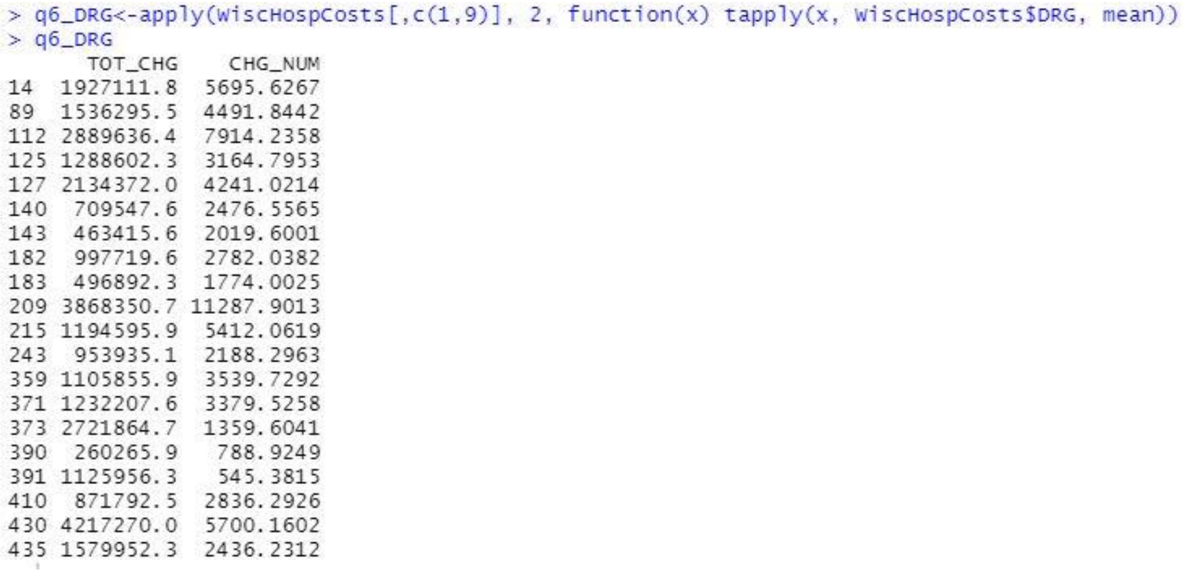
Q6) At which HSA, DRG and Payer level their TOT_CHG and CHG_NUM can get increased



```
> q6_tot_chg<-apply(WiscHospCosts[,c(1,9)], 2, function(x) tapply(x, WiscHospCosts$HSA, mean))
> q6_HSA<-apply(WiscHospCosts[,c(1,9)], 2, function(x) tapply(x, WiscHospCosts$HSA, mean))
> q6_HSA
    TOT_CHG  CHG_NUM
1 2135429.8 3873.145
2 1554809.3 3836.176
3 1244398.4 3419.535
4 1345329.1 3510.021
5  839986.2 3467.110
6 1126655.9 3773.263
7  180956.4 2718.041
8  607318.6 3424.883
9 4831468.5 4824.196
> |
```

Observations:

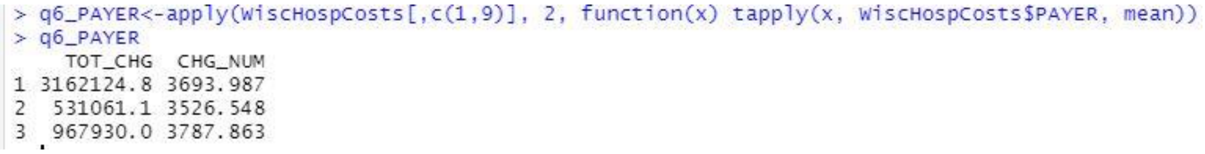
- HSA 9 has highest number of hospital discharged costs and hospital discharged costs per patient.
- So, at HSA 9 TOT_CHG and CHG_NUM can get increased.
- HSA 1, 2 and 4 comes next in the list compare to rest of the HSA’s.



```
> q6_DRG<-apply(WiscHospCosts[,c(1,9)], 2, function(x) tapply(x, WiscHospCosts$DRG, mean))
> q6_DRG
      TOT_CHG    CHG_NUM
14  1927111.8  5695.6267
89  1536295.5  4491.8442
112 2889636.4  7914.2358
125 1288602.3  3164.7953
127 2134372.0  4241.0214
140  709547.6  2476.5565
143  463415.6  2019.6001
182  997719.6  2782.0382
183  496892.3  1774.0025
209 3868350.7 11287.9013
215 1194595.9  5412.0619
243  953935.1  2188.2963
359 1105855.9  3539.7292
371 1232207.6  3379.5258
373 2721864.7  1359.6041
390  260265.9   788.9249
391 1125956.3   545.3815
410  871792.5  2836.2926
430 4217270.0  5700.1602
435 1579952.3  2436.2312
```

Observations:

- DRG 430 has highest number of hospital discharged costs and hospital discharged costs per patient.
- So, at DRG 430, TOT_CHG and CHG_NUM can get increased.
- DRG 209, 112 and 373 comes next in the list compare to rest of the DRG’s.
- Though the TOT_CHG is less for few DRG’s, the CHG_NUM are higher for those TOT_CHG.



```
> q6_PAYER<-apply(WiscHospCosts[,c(1,9)], 2, function(x) tapply(x, WiscHospCosts$PAYER, mean))
> q6_PAYER
    TOT_CHG  CHG_NUM
1 3162124.8 3693.987
2  531061.1 3526.548
3  967930.0 3787.863
```

Observations:

- Payer 1 has highest number of hospital discharged costs and hospital discharged costs per patient.
- So, at Payer 1, TOT_CHG and CHG_NUM can get increased.
- Payer level 3 and 2 comes next in the list.