To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

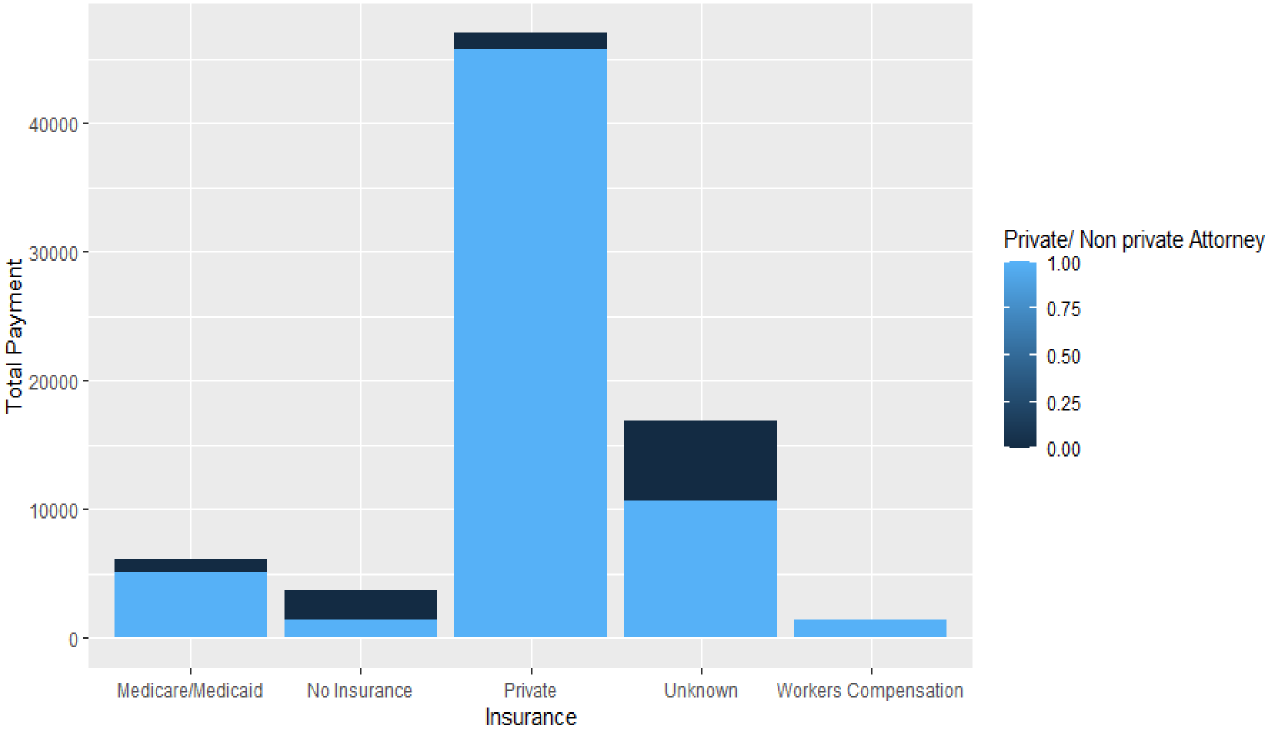
This summary reports on the elements that influence settlements for medical malpractice and the insurance company.

EXECUTIVE SUMMARY

Major Findings:

- Across all the payments, total sum of private attorney payments weighs 88.27% from the total payments. Also payments more than $800,000 are mostly private attorney that's why claims have to be paid more as compared to non-private attorney.
- Total of private insurance payments is highest by value $46,950,500 as compared to all other type of insurances, covering 62.53% from the total.
- Severity of 7 and 8 has the highest payments which is 38.69% of the total payments, where severity 7 has paid maximum total of payments of $14,606,200 and severity 8 of $14,446,100.
- Maximum number of lawsuits are filed by people whose age was 25 to 50 in between.
- Maximum of the payments have a private medical insurance, as total sum of private insurance is largest with the value of $46,950,500. And lowest was for Worker's Compensation worth $1,387,555.
- Maximum sum of payments are paid by married ones with private insurance.
- Medical payments are more than payments for surgeries by $21,308,100.

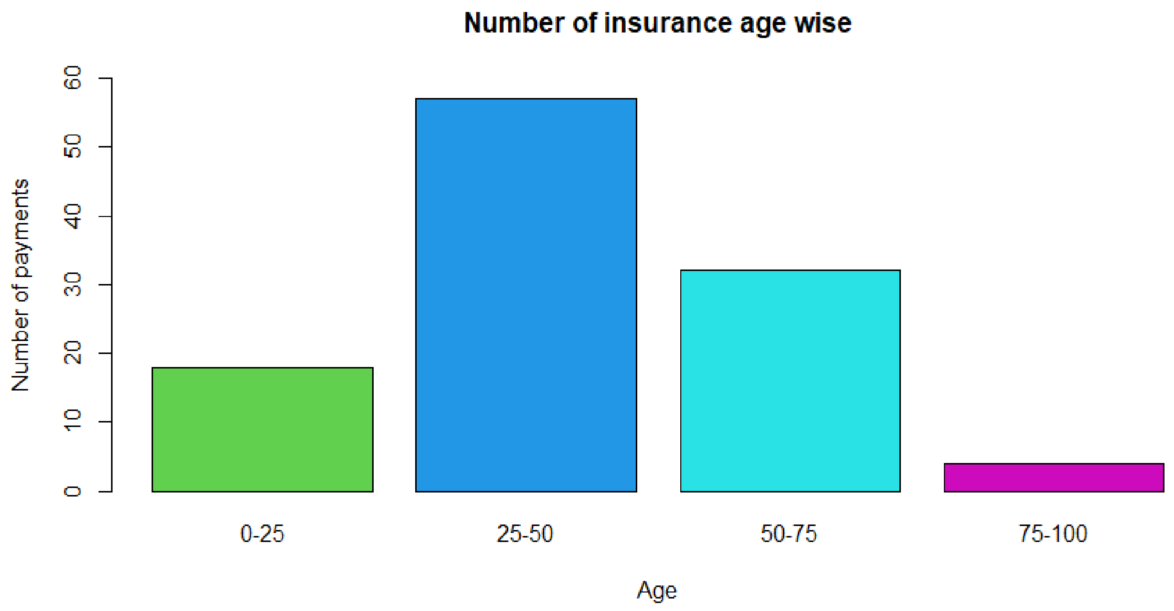
Relation between payments from various type of insurance for private / non private attorney



To: **Analytic Manager, United Health Care.**
From: **Urvish Patel**
Subject: **Insurance payment affected by various factors.**

Recommendations for Action

- Company should focus on people with age range from 25 to 50 and improve their policy terms.
- Private attorney are the ones who are costly, so company should hire more experienced lawyers to fight with them in court so that they could win it and compensate less to the patients.
- Should make a policy of compensating a small amount for medical bills as they are more than surgical ones.
- Should increase Insurance cost of doctors who are in family practice and OBGYN.



Analytical Overview

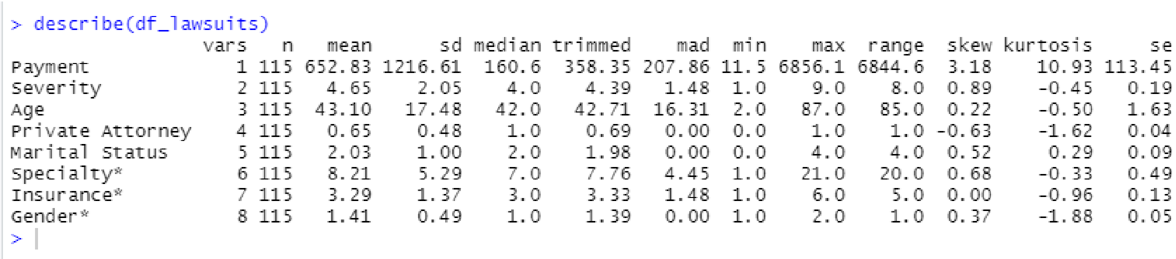
- Excel data was cleaned before analysing such as duplicates and where age was 0.
- For all graphs and data visualisation, only R programming was used.
- All major findings and recommendations are based on EDA which is explained in Documentation Page.

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

Documentation Page

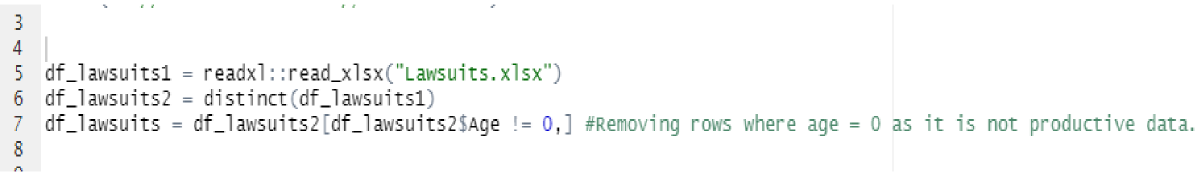
Cleaning data

- First by using “describe” function I got to know there was no “NULL” data in excel. I also got a brief summary of all the columns. For using “describe” function I installed a package and included a library (“install.packages("psych") ; library(psych)”).



```
> describe(df_lawsuits)
                  vars   n   mean      sd median trimmed    mad  min    max  range  skew kurtosis     se
Payment              1 115 652.83 1216.61  160.6  358.35 207.86 11.5 6856.1 6844.6  3.18    10.93 113.45
Severity             2 115   4.65    2.05    4.0    4.39   1.48  1.0    9.0    8.0  0.89    -0.45   0.19
Age                  3 115  43.10   17.48   42.0   42.71  16.31  2.0   87.0   85.0  0.22    -0.50   1.63
Private Attorney     4 115   0.65    0.48    1.0    0.69   0.00  0.0    1.0    1.0 -0.63    -1.62   0.04
Marital Status       5 115   2.03    1.00    2.0    1.98   0.00  0.0    4.0    4.0  0.52     0.29   0.09
Specialty*           6 115   8.21    5.29    7.0    7.76   4.45  1.0   21.0   20.0  0.68    -0.33   0.49
Insurance*           7 115   3.29    1.37    3.0    3.33   1.48  1.0    6.0    5.0  0.00    -0.96   0.13
Gender*              8 115   1.41    0.49    1.0    1.39   0.00  1.0    2.0    1.0  0.37    -1.88   0.05
> |
```

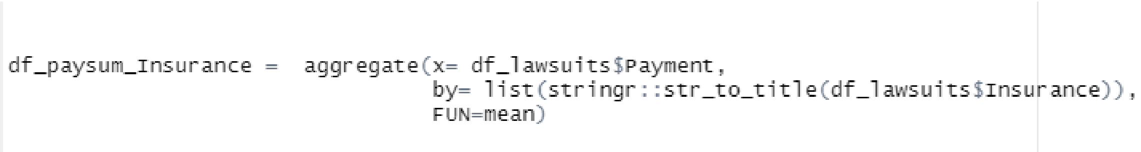
- Secondly, what I found was ambiguity in data i.e. I found a row which was repeating in data frame. So I omitted the repeating row through R code. I used ”distinct” function to do so.(Code line 6)
- Thirdly, I got to know that there was a value of “Age” which was 0. As it is not possible so I omitted through R code. (Code line 7)



```
3
4
5  df_lawsuits1 = readxl::read_xlsx("Lawsuits.xlsx")
6  df_lawsuits2 = distinct(df_lawsuits1)
7  df_lawsuits = df_lawsuits2[df_lawsuits2$Age != 0,] #Removing rows where age = 0 as it is not productive data.
8
```

  There could be many possibilities of “Why age was written 0”. May be it was “10” or may be “20” or even”30”. So I decided to omit the data so that I can evaluate the rest data precisely.

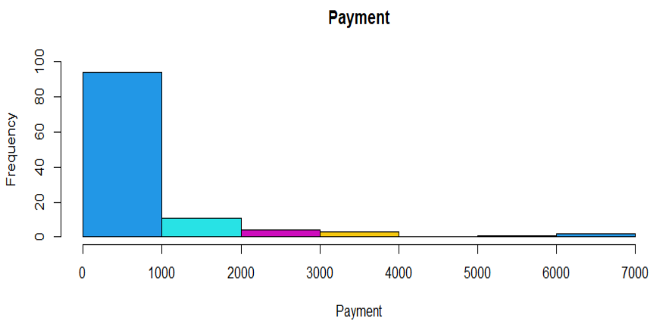
- Fourthly, I also came to know that some data was case sensitive, like ‘OBGYN’ and ‘ObGyn’, also ‘Unknown’ and ‘unknown’. So for that whenever I used text data I made sure I changed the case to “Title” i.e. First letter upper case and the rest in lower case. (Code line 2). I used a library “stringr” and function “str_to_title” to do the same.

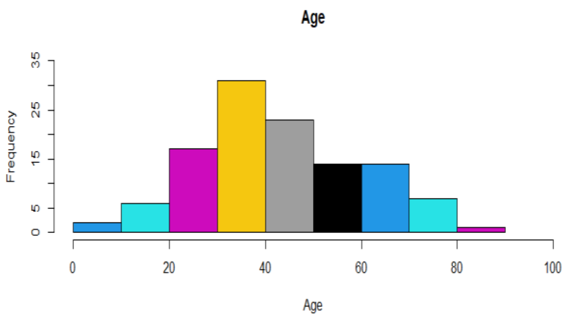


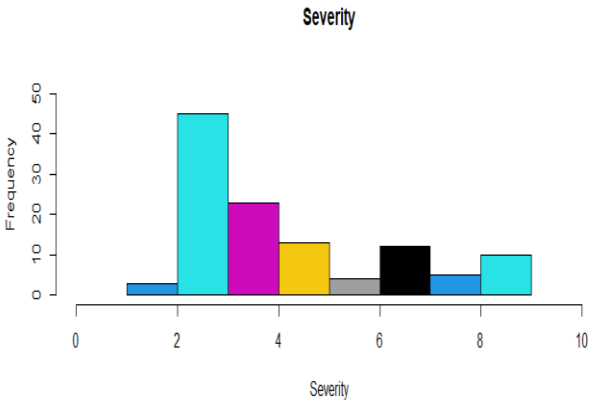
```
df_paysum_Insurance =   aggregate(x= df_lawsuits$Payment,
                                  by= list(stringr::str_to_title(df_lawsuits$Insurance)),
                                  FUN=mean)
```

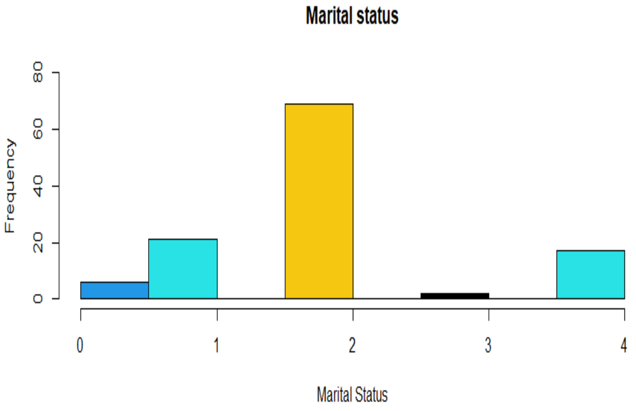
Histogram of various numeric columns

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.





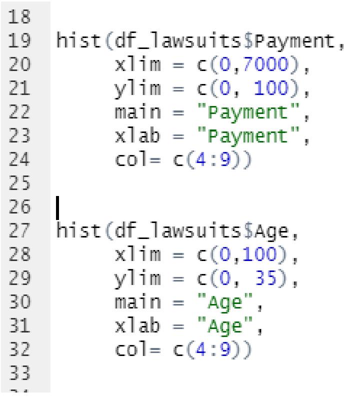




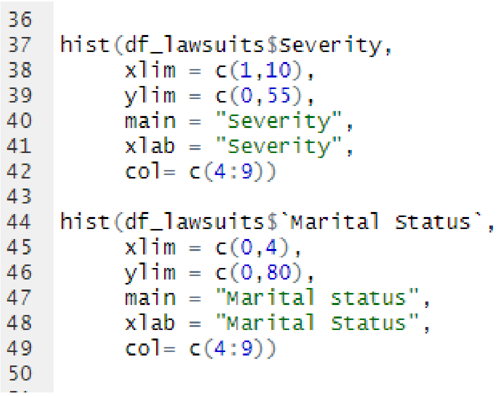
Following things I can conclude from the above graphs:

- Most of the claims were paid by the company was below $1000.
- Most of the payment was made for age in between 20 to 50.
- Most of the case was for severity 3 which is minor temporary damage.
- Most of the payments were made to married people.

We obtained the following graphs by the following R code:



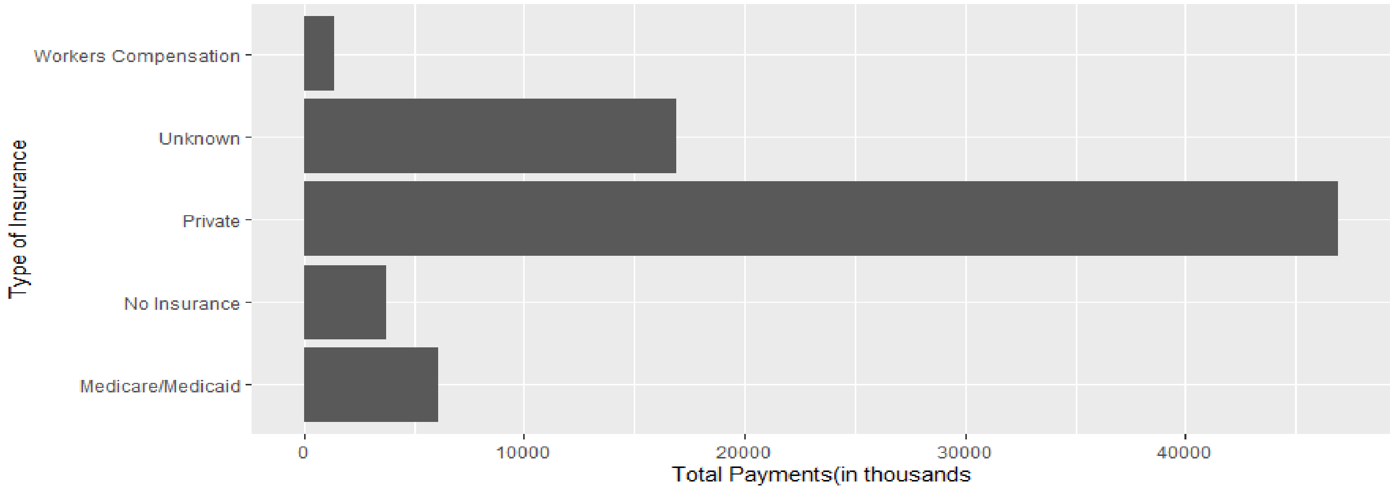
```
hist(df_lawsuits$Payment,
     xlim = c(0,7000),
     ylim = c(0, 100),
     main = "Payment",
     xlab = "Payment",
     col= c(4:9))


hist(df_lawsuits$Age,
     xlim = c(0,100),
     ylim = c(0, 35),
     main = "Age",
     xlab = "Age",
     col= c(4:9))
```



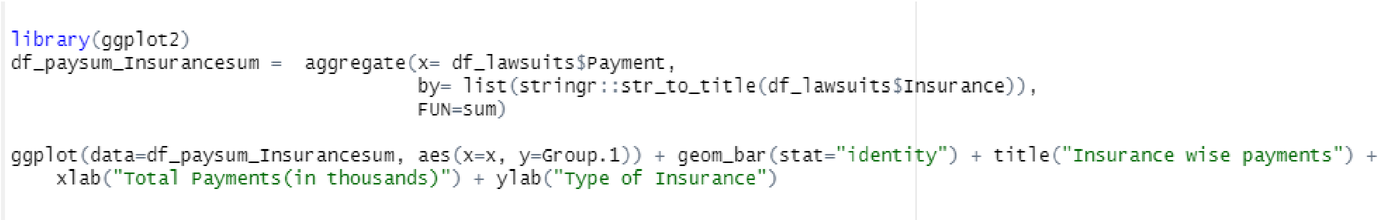
```
hist(df_lawsuits$Severity,
     xlim = c(1,10),
     ylim = c(0,55),
     main = "Severity",
     xlab = "Severity",
     col= c(4:9))

hist(df_lawsuits$`Marital Status`,
     xlim = c(0,4),
     ylim = c(0,80),
     main = "Marital status",
     xlab = "Marital Status",
     col= c(4:9))
```

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

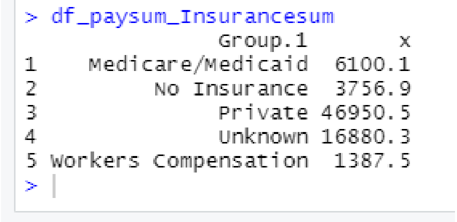
Insurance wise Payments



The above graph represents the total payments claimed by various insurance types. I can conclude that the highest is made by 'Private Insurance'. I achieved this graph by following R code:

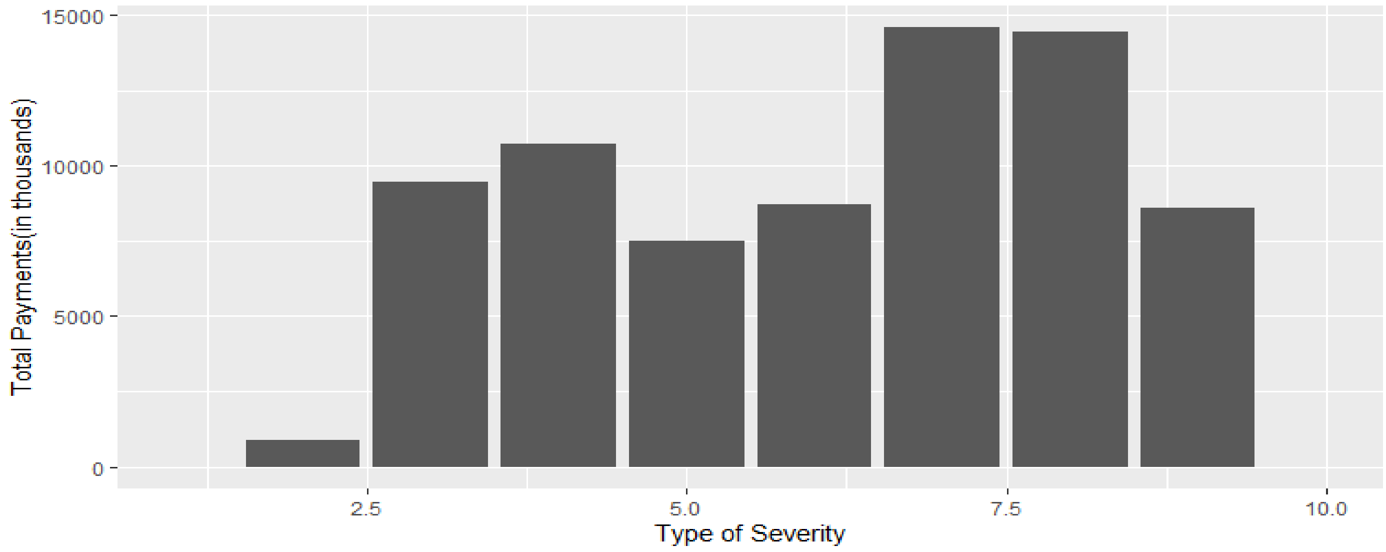


```
library(ggplot2)
df_paysum_Insurancesum =  aggregate(x= df_lawsuits$Payment,
                                    by= list(stringr::str_to_title(df_lawsuits$Insurance)),
                                    FUN=sum)

ggplot(data=df_paysum_Insurancesum, aes(x=x, y=Group.1)) + geom_bar(stat="identity") + title("Insurance wise payments") +
    xlab("Total Payments(in thousands)") + ylab("Type of Insurance")
```



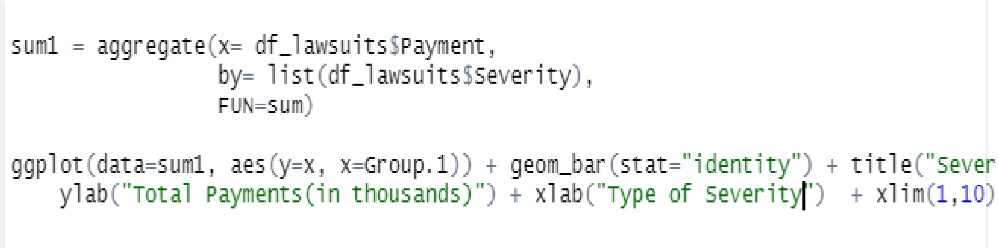
```
> df_paysum_Insurancesum
                Group.1       x
1     Medicare/Medicaid  6100.1
2          No Insurance  3756.9
3               Private 46950.5
4               Unknown 16880.3
5 Workers Compensation  1387.5
> 
```

Severity wise Payment

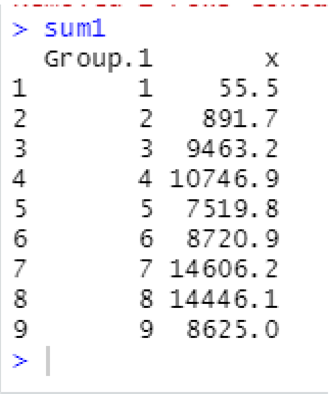


To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

The above plot indicates that Severity with 7 and 8 has caused this company a lot, as they has they had the highest payment. Whereas Severity 1 has the least total payment. I achieved this graph by the following R code: (Group.1 = Type of severity, x = Total payment of severity)

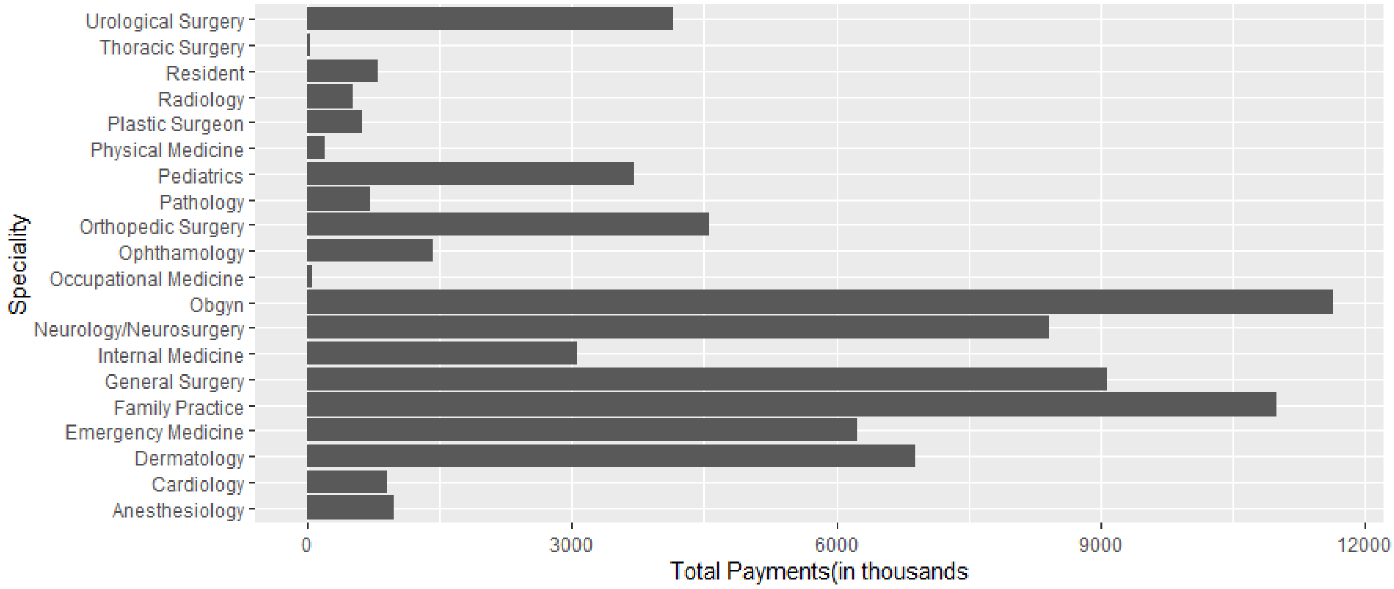


```
sum1 = aggregate(x= df_lawsuits$Payment,
                 by= list(df_lawsuits$Severity),
                 FUN=sum)

ggplot(data=sum1, aes(y=x, x=Group.1)) + geom_bar(stat="identity") + title("Severity wise payments") +
    ylab("Total Payments(in thousands)") + xlab("Type of Severity")  + xlim(1,10)
```



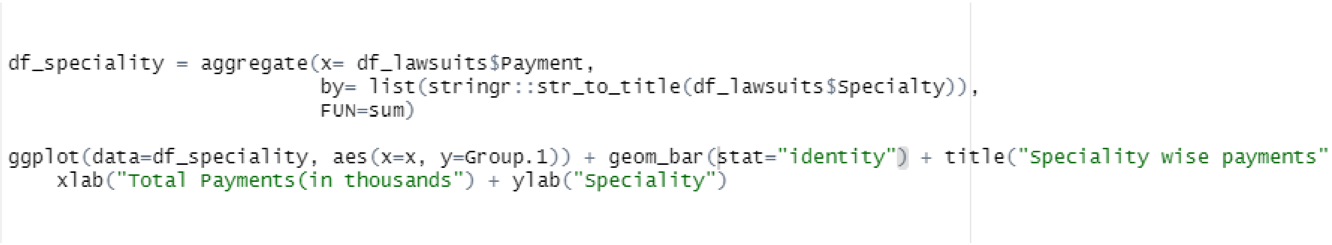
```
> sum1
  Group.1       x
1       1    55.5
2       2   891.7
3       3  9463.2
4       4 10746.9
5       5  7519.8
6       6  8720.9
7       7 14606.2
8       8 14446.1
9       9  8625.0
> 
```

Bar graph for Speciality

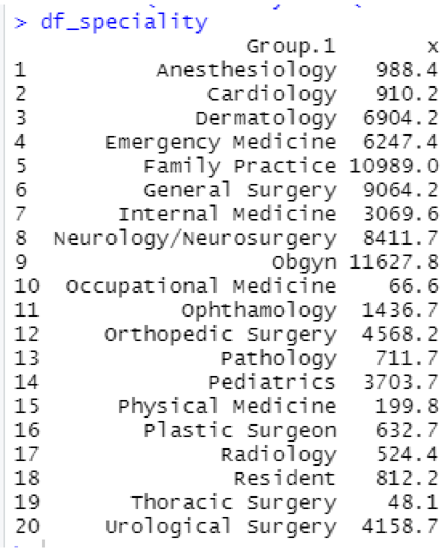


The above graph indicates that maximum payments were done for 'OBGYN' then 'Family Practice'. I obtained this result by the following R code where I used a code stringr::str_to_title which formats all the rows of particular column to title case (First alphabet Upper rest Lower) to avoid ambiguity.

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

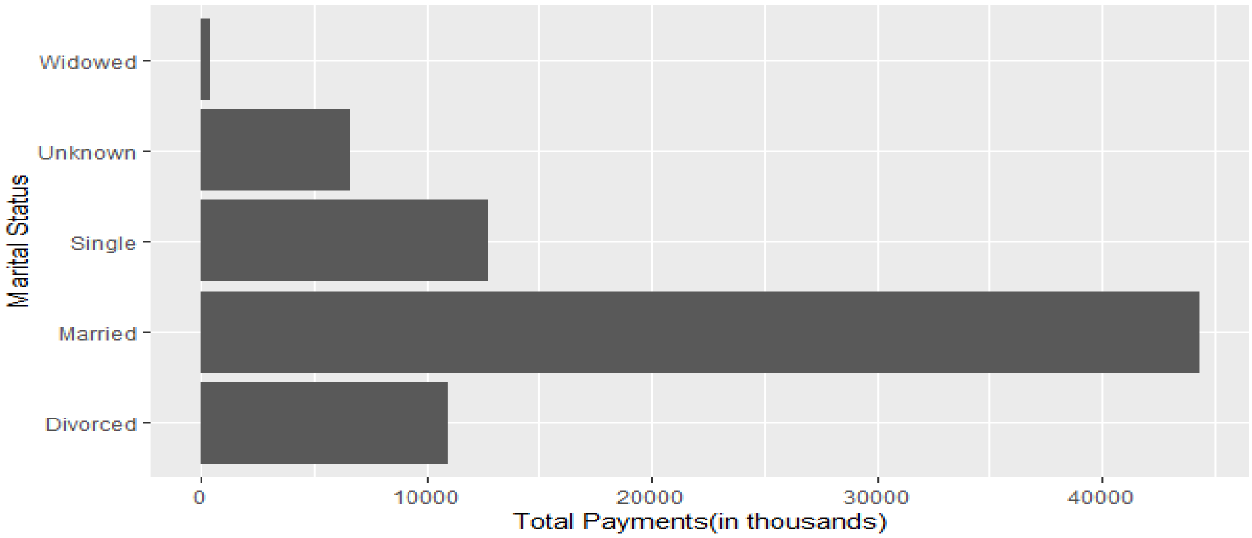


```
df_speciality = aggregate(x= df_lawsuits$Payment,
                          by= list(stringr::str_to_title(df_lawsuits$Specialty)),
                          FUN=sum)

ggplot(data=df_speciality, aes(x=x, y=Group.1)) + geom_bar(stat="identity") + title("Speciality wise payments"
    xlab("Total Payments(in thousands") + ylab("Speciality")
```

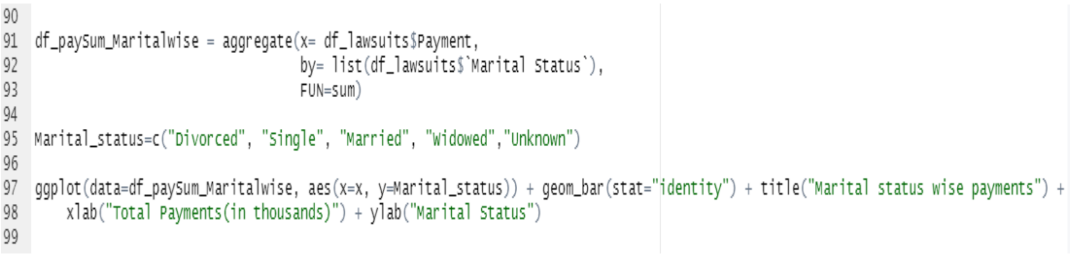


```
> df_speciality
                    Group.1       x
1             Anesthesiology   988.4
2                 Cardiology   910.2
3                Dermatology  6904.2
4         Emergency Medicine  6247.4
5            Family Practice 10989.0
6            General Surgery  9064.2
7          Internal Medicine  3069.6
8  Neurology/Neurosurgery  8411.7
9                      Obgyn 11627.8
10       Occupational Medicine    66.6
11              Ophthamology  1436.7
12         Orthopedic Surgery  4568.2
13                 Pathology   711.7
14                Pediatrics  3703.7
15          Physical Medicine   199.8
16            Plastic Surgeon   632.7
17                 Radiology   524.4
18                  Resident   812.2
19           Thoracic Surgery    48.1
20         Urological Surgery  4158.7
```

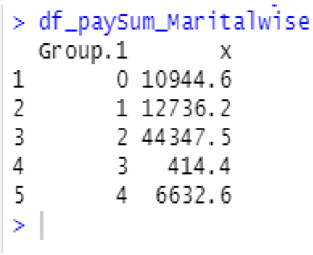
Marital Status wise payment



The above graph denotes that the married patients have paid the highest total payment as compared to all others. I obtained the following graph by the following R code:



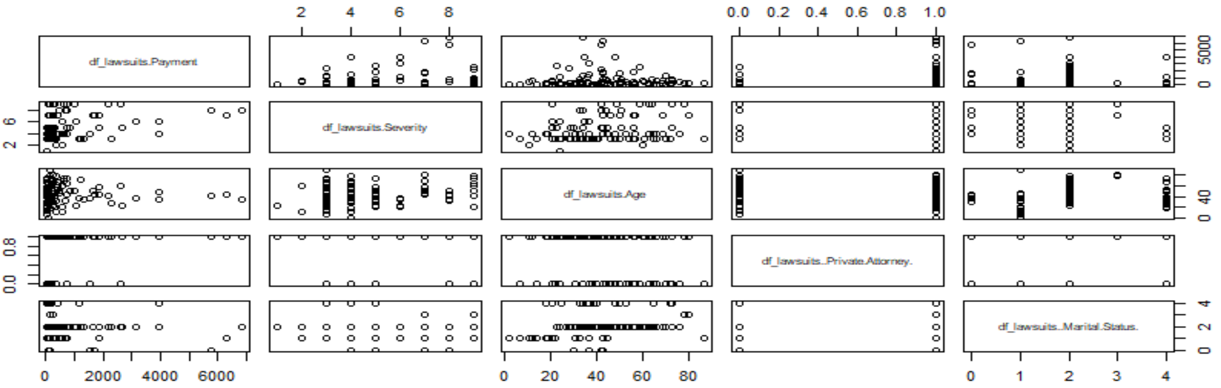
```
df_paySum_Maritalwise = aggregate(x= df_lawsuits$Payment,
                                  by= list(df_lawsuits$`Marital Status`),
                                  FUN=sum)

Marital_status=c("Divorced", "Single", "Married", "Widowed","Unknown")

ggplot(data=df_paySum_Maritalwise, aes(x=x, y=Marital_status)) + geom_bar(stat="identity") + title("Marital status wise payments") +
    xlab("Total Payments(in thousands)") + ylab("Marital Status")
```



```
> df_paySum_Maritalwise
  Group.1       x
1       0 10944.6
2       1 12736.2
3       2 44347.5
4       3   414.4
5       4  6632.6
> 
```

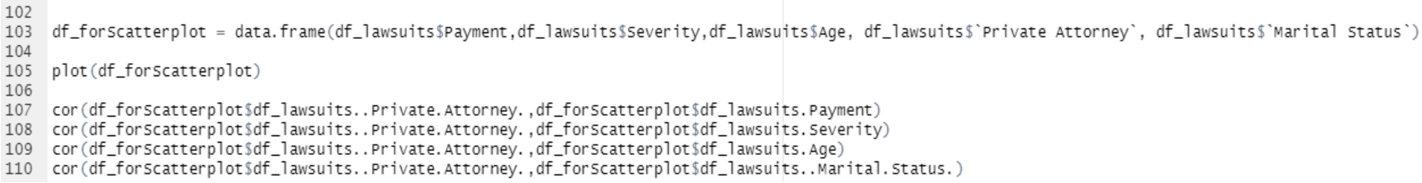
To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

Correlations between numeric columns:



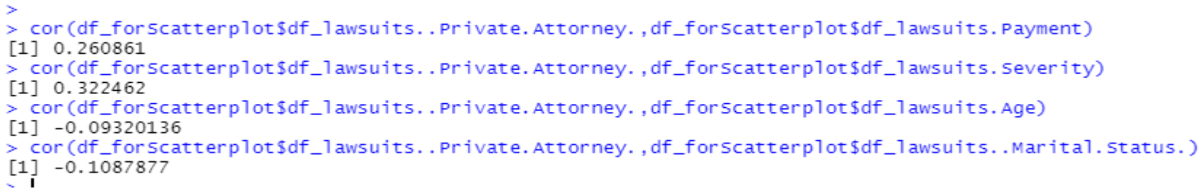
The above graph shows the correlations between all numeric columns. From this I can conclude that "Private Attorney and Payment" and "Private Attorney and Severity" has a positive correlation. Whereas "Private Attorney and Marital Status has a negative correlation. I also found the correlations through the R program as follows:

Input



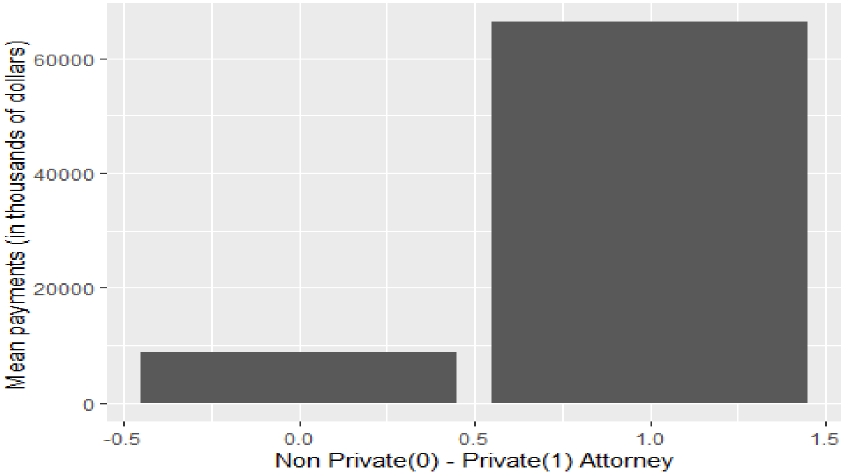
```
df_forScatterplot = data.frame(df_lawsuits$Payment,df_lawsuits$Severity,df_lawsuits$Age, df_lawsuits$`Private Attorney`, df_lawsuits$`Marital Status`)

plot(df_forScatterplot)

cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits.Payment)
cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits.Severity)
cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits.Age)
cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits..Marital.Status.)
```

Output



```
>
> cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits.Payment)
[1] 0.260861
> cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits.Severity)
[1] 0.322462
> cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits.Age)
[1] -0.09320136
> cor(df_forScatterplot$df_lawsuits..Private.Attorney.,df_forScatterplot$df_lawsuits..Marital.Status.)
[1] -0.1087877
>
```

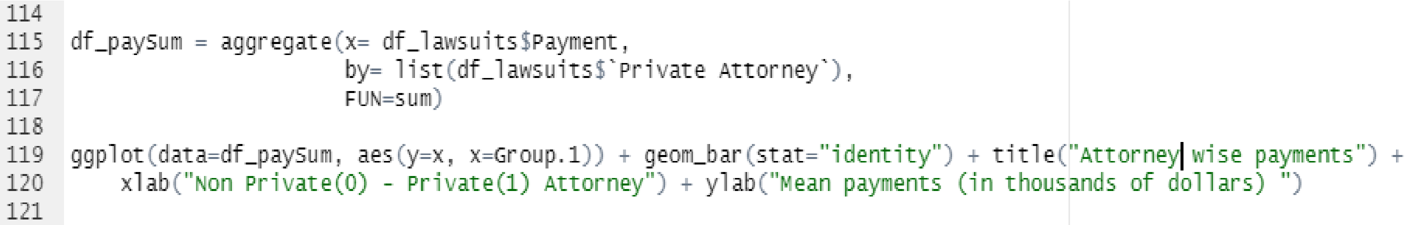
Comparing Private/ Non private attorney Payments



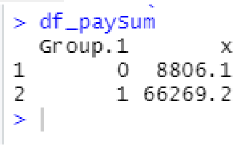
Beside graph indicates the difference between a Private attorney payments and a Non Private attorney payments. Private attorney Payments are lot higher than a Non private one.

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

We achieved this plot by the following R code:

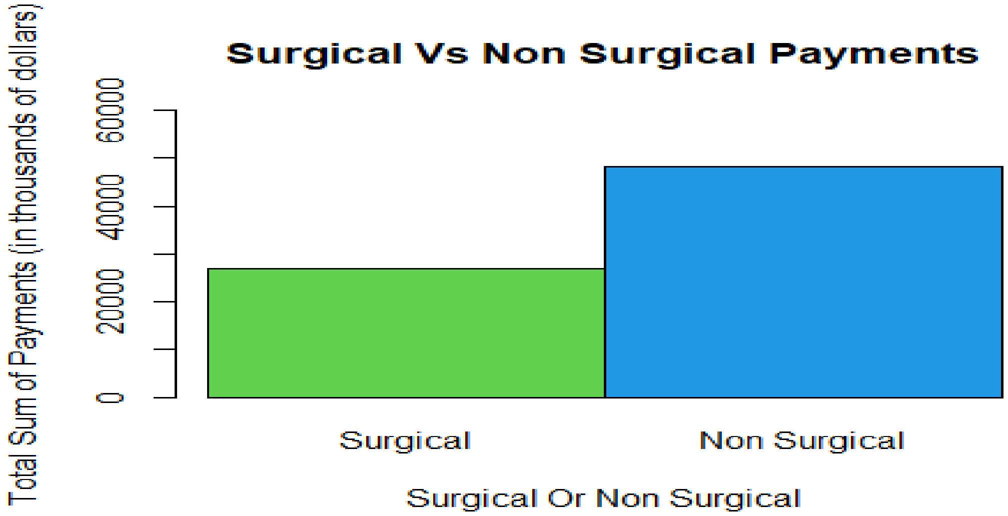


```
df_paySum = aggregate(x= df_lawsuits$Payment,
                      by= list(df_lawsuits$`Private Attorney`),
                      FUN=sum)

ggplot(data=df_paySum, aes(y=x, x=Group.1)) + geom_bar(stat="identity") + title("Attorney wise payments") +
    xlab("Non Private(0) - Private(1) Attorney") + ylab("Mean payments (in thousands of dollars) ")
```

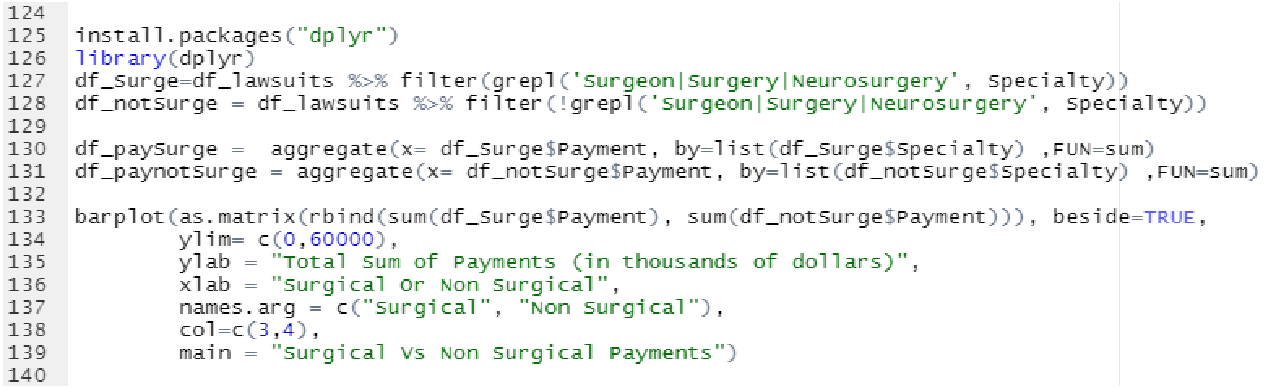


```
> df_paySum
  Group.1       x
1       0  8806.1
2       1 66269.2
> 
```

Surgical Payments VS Non-Surgical payments

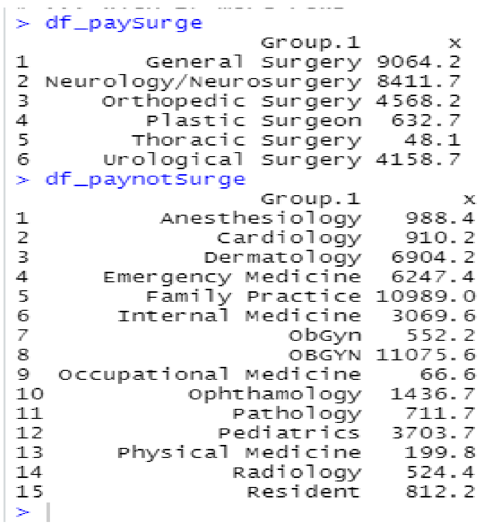


Beside graph denotes the total payment done for surgical and Non-surgical events. I can conclude that Non-Surgical expenses are more than the surgical expenses. I got this graph by the following R code.



```
install.packages("dplyr")
library(dplyr)
df_Surge=df_lawsuits %>% filter(grepl('Surgeon|Surgery|Neurosurgery', Specialty))
df_notSurge = df_lawsuits %>% filter(!grepl('Surgeon|Surgery|Neurosurgery', Specialty))

df_paySurge =  aggregate(x= df_Surge$Payment, by=list(df_Surge$Specialty) ,FUN=sum)
df_paynotSurge = aggregate(x= df_notSurge$Payment, by=list(df_notSurge$Specialty) ,FUN=sum)

barplot(as.matrix(rbind(sum(df_Surge$Payment), sum(df_notSurge$Payment))), beside=TRUE,
        ylim= c(0,60000),
        ylab = "Total Sum of Payments (in thousands of dollars)",
        xlab = "Surgical Or Non Surgical",
        names.arg = c("Surgical", "Non Surgical"),
        col=c(3,4),
        main = "Surgical Vs Non Surgical Payments")
```

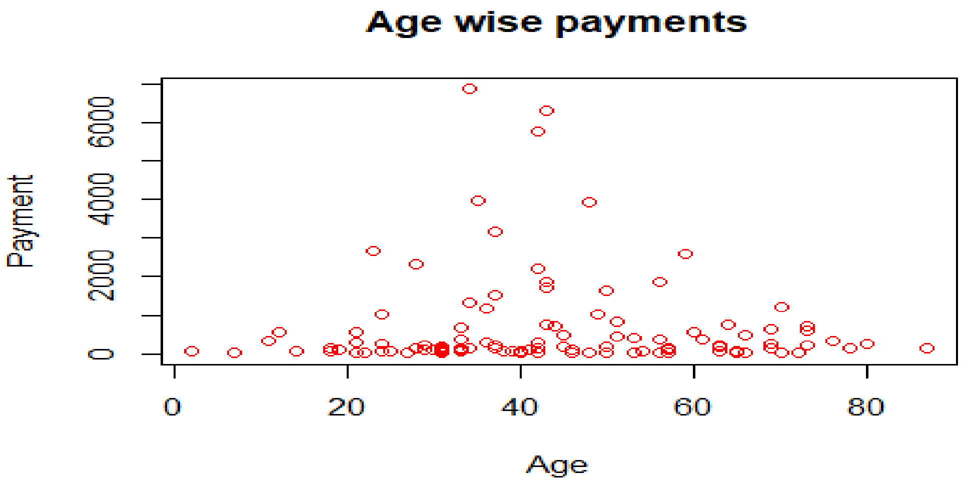
To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.



```
> df_paySurge
                   Group.1      x
1          General Surgery 9064.2
2 Neurology/Neurosurgery 8411.7
3       Orthopedic Surgery 4568.2
4          Plastic Surgeon  632.7
5         Thoracic Surgery   48.1
6       Urological Surgery 4158.7
> df_paynotSurge
                  Group.1       x
1          Anesthesiology   988.4
2              Cardiology   910.2
3             Dermatology  6904.2
4      Emergency Medicine  6247.4
5         Family Practice 10989.0
6       Internal Medicine  3069.6
7                   ObGyn   552.2
8                   OBGYN 11075.6
9   Occupational Medicine    66.6
10           Ophthamology  1436.7
11              Pathology   711.7
12             Pediatrics  3703.7
13       Physical Medicine   199.8
14              Radiology   524.4
15               Resident   812.2
> 
```

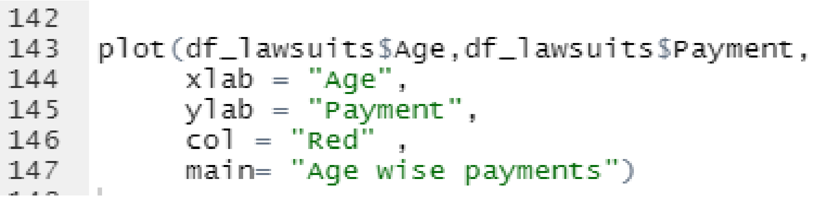
First I installed a library ‘dplyr’. Secondly I created two separate data’s for Surgical and non-surgical speciality, for that I used filter function (line 127) to get all rows which has terms like ‘Surgeon’, ‘Surgery’ and ’Neurosurgery’ and stored in df_Surge. Then I filtered for non-surgical data by just putting !grepl (line 128) and stored it into df_notSurge. Thereafter I found sum of payments (line 130 and 131). After that I simply plotted a bar graph using a bar plot.

Age wise Payments



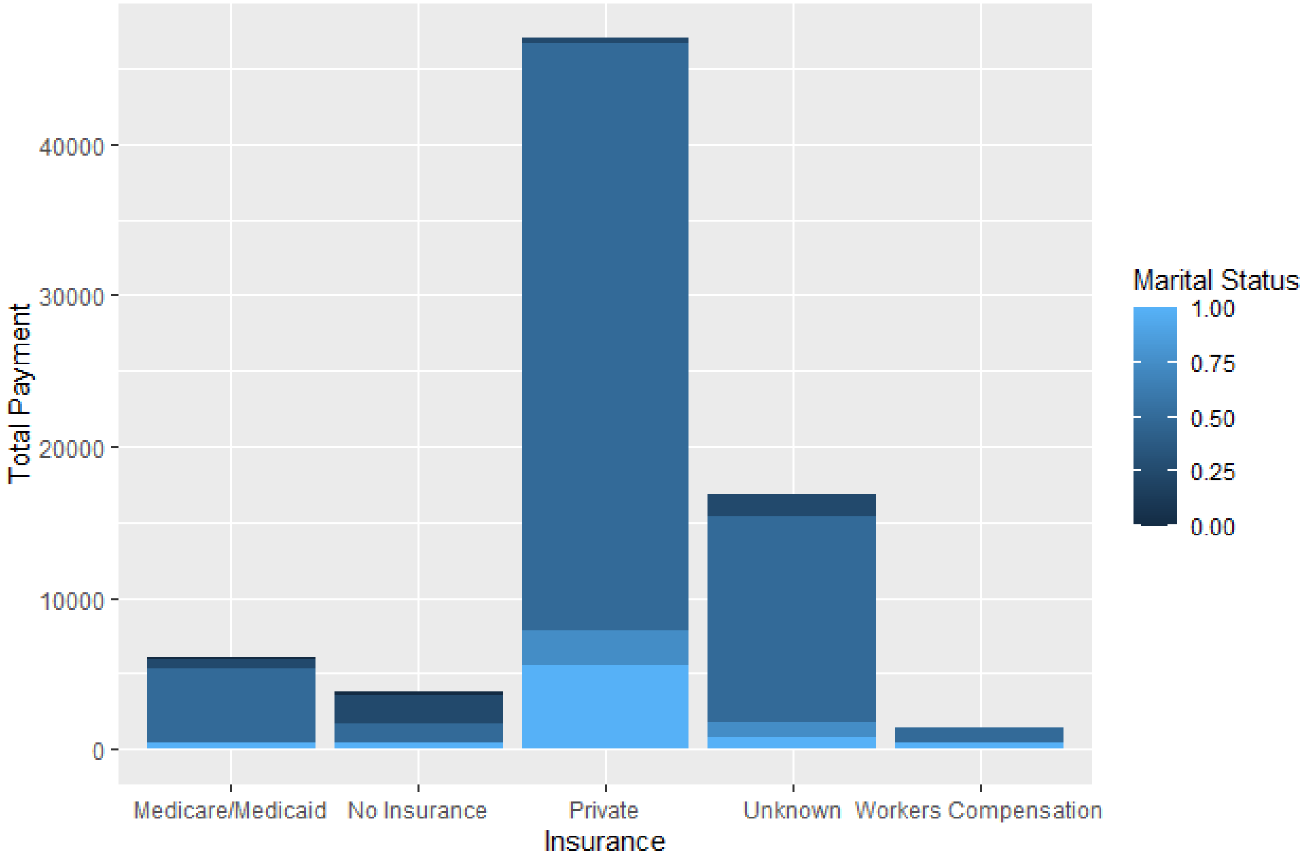
From this scatter plot I can conclude that from the age 25 to 50 highest number of payments were made and also the highest payment was claimed. I got this plot by following R code:

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

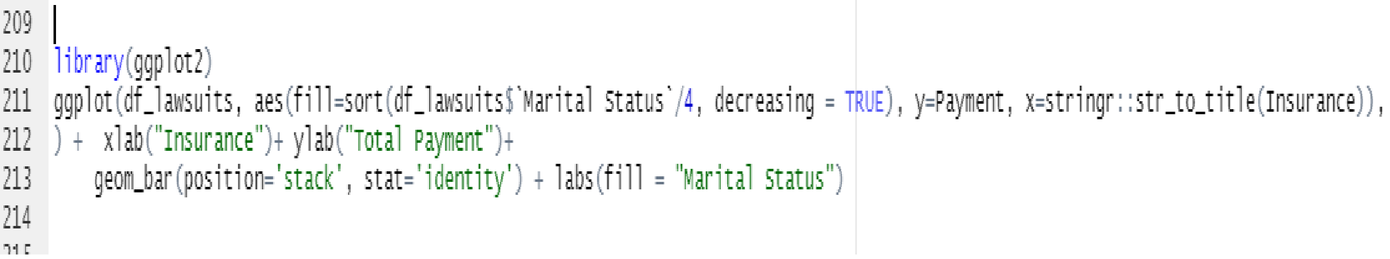


```
plot(df_lawsuits$Age,df_lawsuits$Payment,
     xlab = "Age",
     ylab = "Payment",
     col = "Red" ,
     main= "Age wise payments")
```

Relation between payments from various type of insurance according to Marital Status of patients.



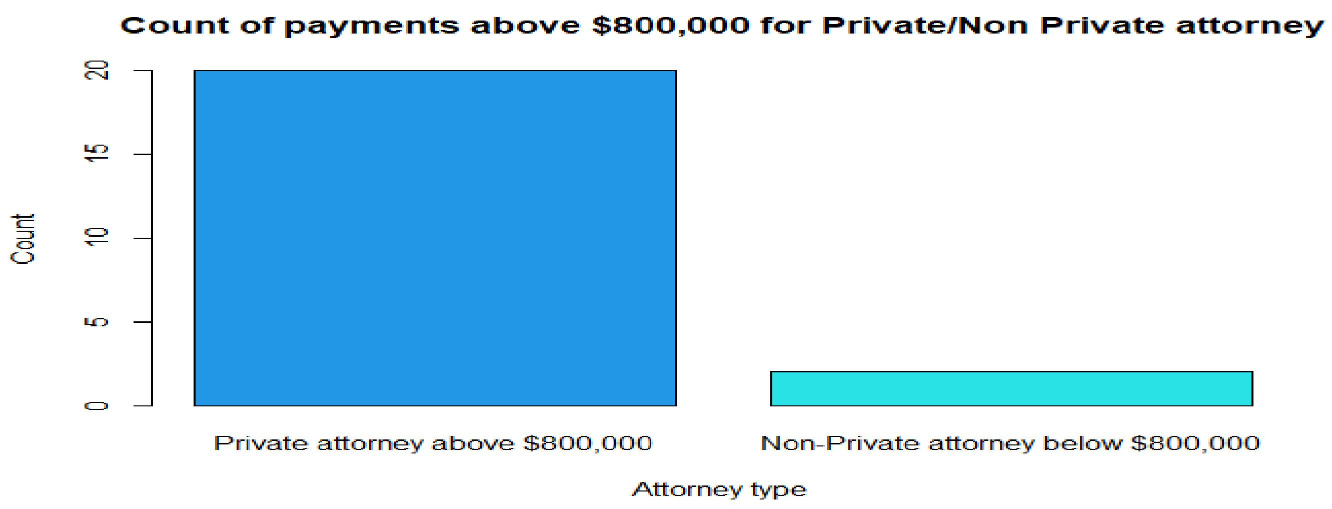
From this stacked bar graph I can conclude that marital status of 2 (Married) has paid the most number and sum of payments. I got the following plot by below R code in which I have divided the marital status column by 4, so 1.00 represents marital status 4 (unknown), 0.75 represents marital status 3 (widowed), 0.5 represents marital status 2 (married), 0.25 represents marital status 1 (single) and 0.00 represents marital status 0 (divorced).



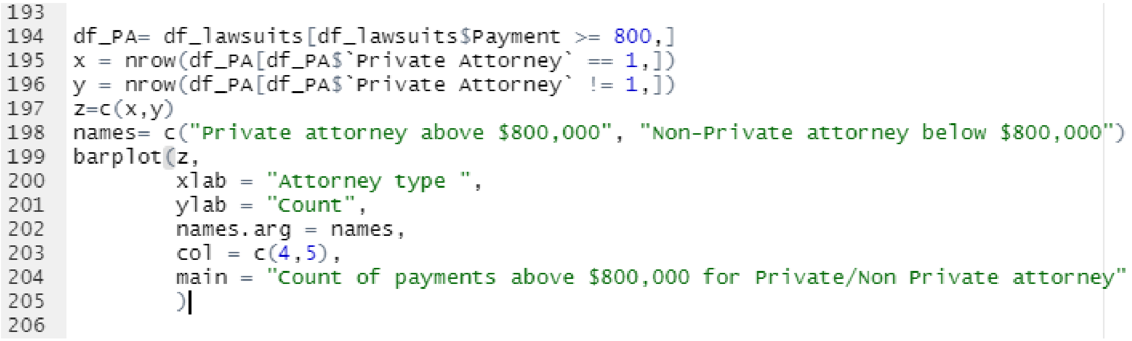
```

library(ggplot2)
ggplot(df_lawsuits, aes(fill=sort(df_lawsuits$`Marital Status`/4, decreasing = TRUE), y=Payment, x=stringr::str_to_title(Insurance)),
) +  xlab("Insurance")+ ylab("Total Payment")+
    geom_bar(position='stack', stat='identity') + labs(fill = "Marital Status")

```

To: Analytic Manager, United Health Care.
From: Urvish Patel
Subject: Insurance payment affected by various factors.

Relation between payments above $800,000 and Private/Non-Private attorney



The above graph concludes that the majority of payments above $800,000 is from Private attorney. I got the graph from the following R code.



```

df_PA= df_lawsuits[df_lawsuits$Payment >= 800,]
x = nrow(df_PA[df_PA$`Private Attorney` == 1,])
y = nrow(df_PA[df_PA$`Private Attorney` != 1,])
z=c(x,y)
names= c("Private attorney above $800,000", "Non-Private attorney below $800,000")
barplot(z,
        xlab = "Attorney type ",
        ylab = "Count",
        names.arg = names,
        col = c(4,5),
        main = "Count of payments above $800,000 for Private/Non Private attorney"
        )

```