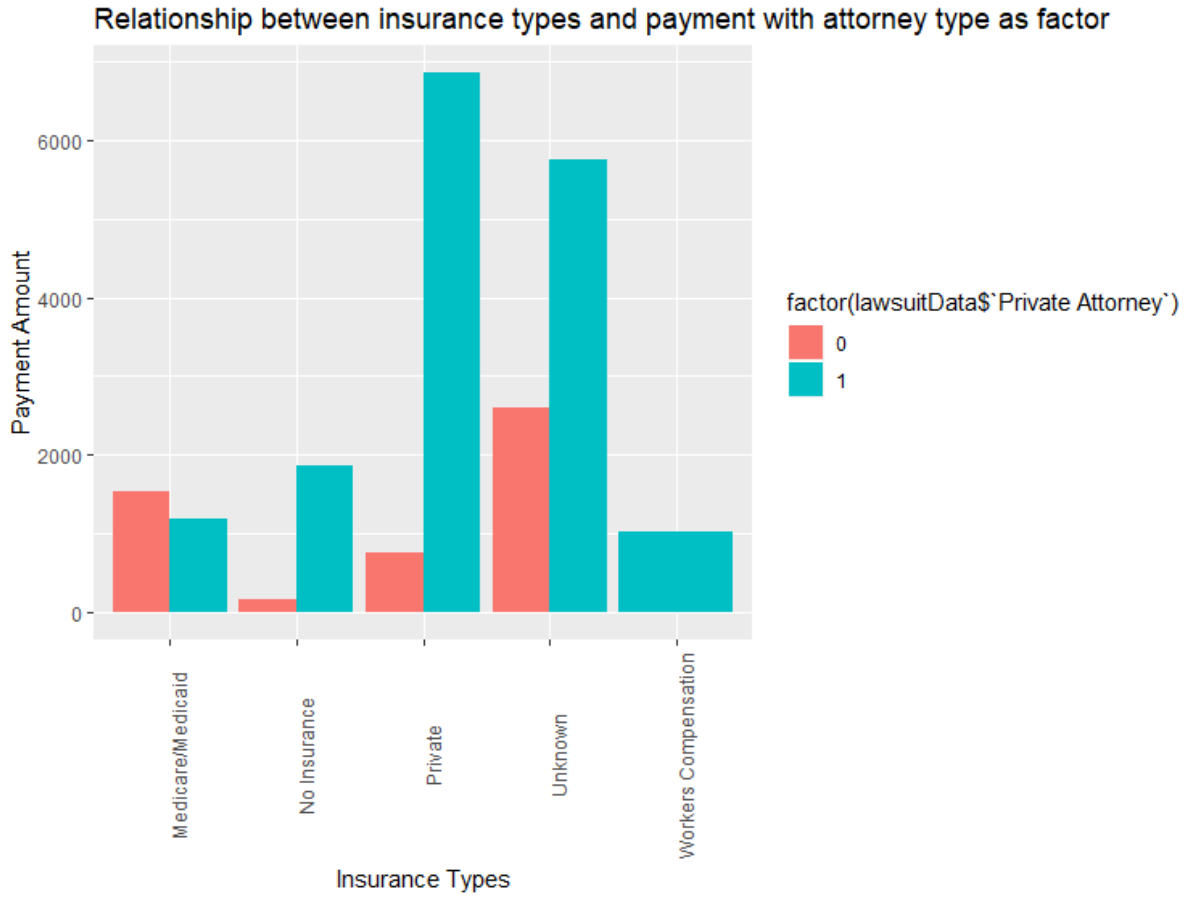
To: The manager of the insurance company
From: Vedant Dashora, Yogesh Selvaraj Narayanan, Chinmayi Suryakant Mahadik
Subject: Report on the amount of claims the company paid for medical malpractice lawsuits.
Date: 2/26/2020

A report to discuss different factors influencing payments for medical malpractice lawsuits and the insurance company.

EXECUTIVE SUMMARY

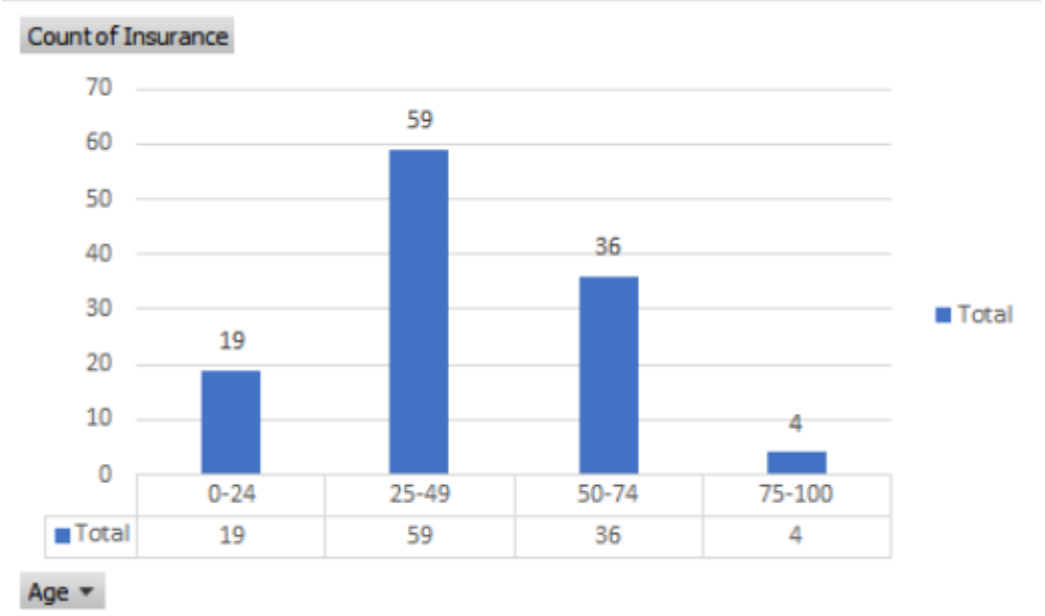
Major Findings:

- The patients' data tells us that the maximum of them have private medical insurance and their doctor was a family practitioner.
- The sum of the payment amount is maximum for male patients who are married and have private medical insurance.
- Severity level 7 (i.e. major permanent damage) has paid maximum total claim amount.
- The claim payment amounts which are more than $16000 have private attorney which is why it is possible that those claims have to pay such a huge amount.
- The age group of 25-49 has the maximum number of lawsuits filed compared to the other age groups.



Recommendations for Action:

- Having a private attorney turns out to be very expensive. Hence the insurance company should rely more on in-house attorney.
- The lawsuits which are claimed by patients without private attorney are comparatively less with private attorney, hence we should invest more resources to fight against lawsuit with private attorney. According to the above graph all type of insurance has paid more in private attorney except Medicare.
- The age group 25-49 has maximum lawsuits, so the company should make more analysis while issuing insurance policies to this age group.



Analytical Overview:

- The approach or method used to analyze and recommend is exploratory data analysis techniques using excel and R programming.
- For data visualization we used R programming as well as excel for interactive graphs which help to understand the data better
- We used EDA as we need to make some assumptions which are explained in the Appendix along with the detailed analysis carried out.

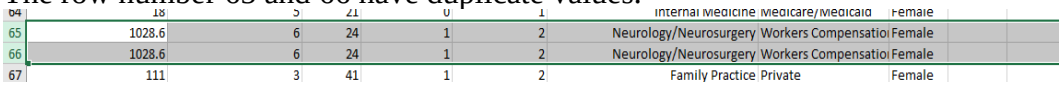
APPENDIX

Process used in Data Analysis

- Data Cleaning
- Data Summarizing
- Handling Outliers
- Inferences from univariant charts
- Inferences from bi-variant charts
- Inferences from multivariate charts

Data Cleaning

- The row number 65 and 66 have duplicate values.



64	18	5	21	0	1	Internal Medicine	Medicare/Medicaid	Female
65	1028.6	6	24	1	2	Neurology/Neurosurgery	Workers Compensatio	Female
66	1028.6	6	24	1	2	Neurology/Neurosurgery	Workers Compensatio	Female
67	111	3	41	1	2	Family Practice	Private	Female

- For this pre-processing, we will be using “Lawsuits.xlsx”. Primarily, we installed three packages namely, “tidyverse”, “funModelling” and “Hmic” for summarise function with detailed reports.
- To make the entire excel sheet not case sensitive, we cleaned the Lawsuits file by using the formula “=PROPER()” on columns “Speciality” and “Insurance” as they contained words that are inconsistent on cases.
- We made sure that there are no NULL values on all variables using the “describe” function in R.
- Since “0” in the column “Age” is not viable, cross checking with the Specialty category, which was Gynaecologist for a woman, we can conclude that age=0 could be a misinterpretation. Thus, woman and gynaecologist as filters, mean value was taken on Age column to fill in the misinterpreted value “0”.

Data Summarizing

- With the help of the installed packages, we used different functions to derive multiple conclusions from the dataset provided.

Installed packages and the library functions used are shown below:

```
install.packages("tidyverse")
install.packages("funModeling")
install.packages("Hmisc")
library(tidyverse)
library(funModeling)
library(Hmisc)
```

Glimpse function is used to reveal the dimensions (Observations) and names of the variables in the dataset.

glimpse(Lawsuits)

```
Console ~/
> glimpse(Lawsuits)
Observations: 118
Variables: 8
$ Payment           <dbl> 3156.1, 473.6, 130.3, 14.7, 307.1, 249.4, 1313.5, 473.6, 38.9, 207.2, 70.3, 29.6, 281.9, 251.6, 159.1...
$ Severity          <dbl> 6, 4, 7, 3, 5, 7, 3, 4, 5, 5, 4, 3, 3, 3, 5, 3, 3, 9, 7, 3, 8, 2, 4, 7, 5, 3, 3, 3, 3, 4, 9, 5, 4, 6,...
$ Age               <dbl> 37, 66, 69, 56, 42, 69, 34, 45, 42, 73, 2, 31, 36, 24, 42, 29, 61, 49, 36, 28, 34, 60, 69, 80, 29, 48...
$ `Private Attorney` <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,...
$ `Marital Status`  <dbl> 2, 2, 2, 2, 2, 2, 2, 1, 2, 4, 1, 2, 2, 2, 0, 2, 2, 2, 1, 2, 2, 2, 2, 3, 2, 2, 4, 2, 4, 4, 2, 2, 1, 2,...
$ Specialty         <chr> "Pediatrics", "Plastic Surgeon", "Internal Medicine", "Urological Surgery", "General Surgery", "Gener...
$ Insurance         <chr> "Private", "Medicare/Medicaid", "Private", "Private", "Unknown", "Private", "Private", "Private", "Pr...
$ Gender            <chr> "Female", "Female", "Male", "Female", "Male", "Male", "Female", "Female", "Female", "Female", "Female...
>
```

- **df_status** function is used to identify the type of values in the attributes and the unique numbers present in it.

df_status(Lawsuits)

```
Console ~/
> df_status(Lawsuits)
          variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1          Payment       0    0.00    0    0     0     0   numeric     91
2         Severity       0    0.00    0    0     0     0   numeric      9
3              Age       0    0.00    0    0     0     0   numeric     54
4 Private Attorney      40   33.90    0    0     0     0   numeric      2
5   Marital Status       6    5.08    0    0     0     0   numeric      5
6        Specialty       0    0.00    0    0     0     0 character     20
7        Insurance       0    0.00    0    0     0     0 character      5
8           Gender       0    0.00    0    0     0     0 character      2
>
```

- **freq** function is used to give the number of times each value is repeated in the dataset. Also, along with frequency, it provides cumulative percentage of each variables with its graphical representation.

freq(Lawsuits)

```
Console ~/
> freq(Lawsuits)
                Specialty frequency percentage cumulative_perc
1          Family Practice        17      14.41            14.41
2          General Surgery        14      11.86            26.27
3           Anesthesiology        13      11.02            37.29
4                    Obgyn        13      11.02            48.31
5       Orthopedic Surgery        11       9.32            57.63
6        Internal Medicine         8       6.78            64.41
7       Emergency Medicine         7       5.93            70.34
8  Neurology/Neurosurgery         7       5.93            76.27
9             Ophthamology         5       4.24            80.51
10              Cardiology         4       3.39            83.90
11               Radiology         3       2.54            86.44
12                Resident         3       2.54            88.98
13     Urological Surgery         3       2.54            91.52
14            Dermatology         2       1.69            93.21
15             Pediatrics         2       1.69            94.90
16         Plastic Surgeon         2       1.69            96.59
17  Occupational Medicine         1       0.85            97.44
18               Pathology         1       0.85            98.29
19      Physical Medicine         1       0.85            99.14
20        Thoracic Surgery         1       0.85           100.00

             Insurance frequency percentage cumulative_perc
1              Private        51      43.22            43.22
2              Unknown        36      30.51            73.73
3    Medicare/Medicaid        16      13.56            87.29
4          No Insurance        12      10.17            97.46
5 Workers Compensation         3       2.54           100.00

  Gender frequency percentage cumulative_perc
1 Female        71      60.17            60.17
2   Male        47      39.83           100.00
```

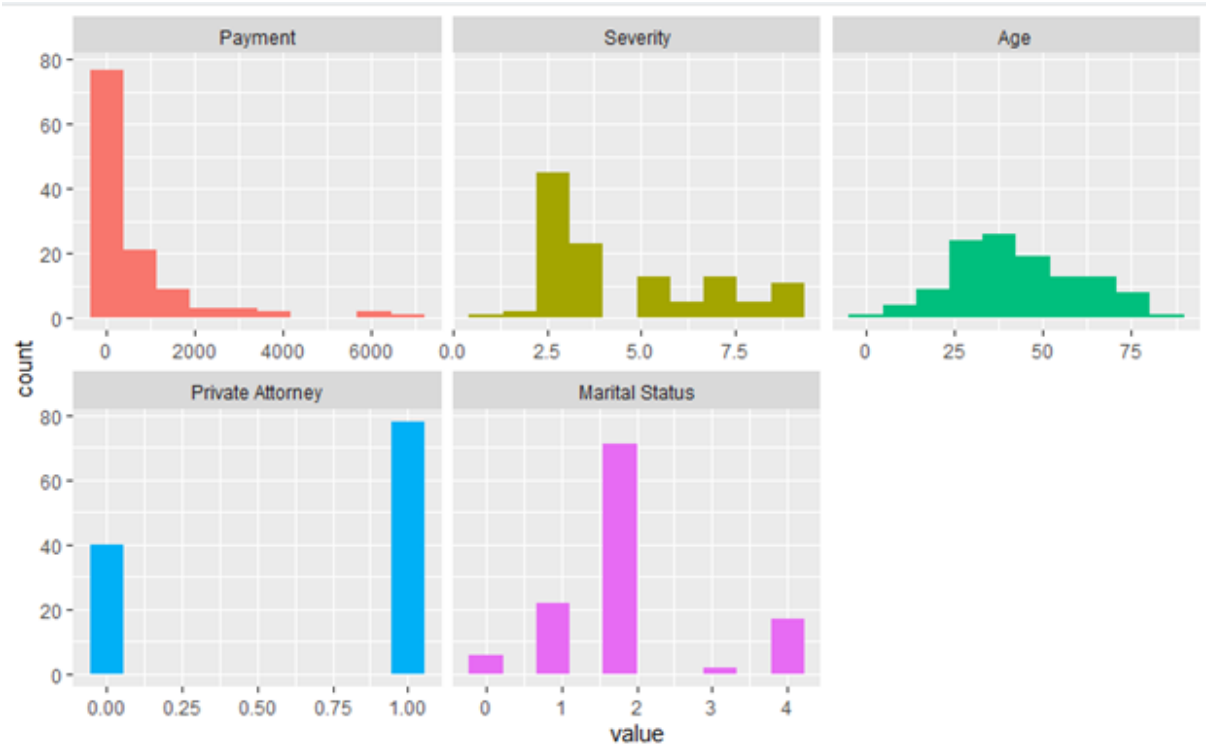
- **profiling_num** function is used to get more detailed information on statistical summary of the dataset. With this function, skewness, kurtosis values of each attribute can be obtained which will help in finding outliers in the dataset.

profiling_num(Lawsuits)

```
Console ~/
> profiling_num(Lawsuits)
         variable       mean      std_dev variation_coef  p_01   p_05  p_25 p_50  p_75   p_95    p_99   skewness  kurtosis
1         Payment 673.7330508 1217.3140599     1.8068196 11.60 14.505 56.425 168.4 667.85 2841.6 6206.75  3.1239192 13.636232
2        Severity   4.7203390    2.0791418     0.4404645  2.00  3.000  3.000   4.0   6.00    9.0    9.00  0.8346007  2.460768
3             Age  43.1355932   17.5722758     0.4073730  7.68 18.000 31.000  41.5  56.00   73.0   79.66  0.2424694  2.504582
4 Private Attorney   0.6610169    0.4753827     0.7191687  0.00  0.000  0.000   1.0   1.00    1.0    1.00 -0.6803091  1.462821
5   Marital Status   2.0169492    0.9955718     0.4936028  0.00  0.850  2.000   2.0   2.00    4.0    4.00  0.5402016  3.400887
     iqr       range_98      range_80
1 611.425 [11.6, 6206.75] [18.56, 1861.1]
2   3.000         [2, 9]         [3, 8]
3  25.000   [7.68, 79.66]      [21.7, 69]
4   1.000         [0, 1]         [0, 1]
5   0.000         [0, 4]         [1, 4]
>
```

- **plot_num** function is used to count the number of observations for a specific category of each variables in graphical representation.

plot_num(Lawsuits)

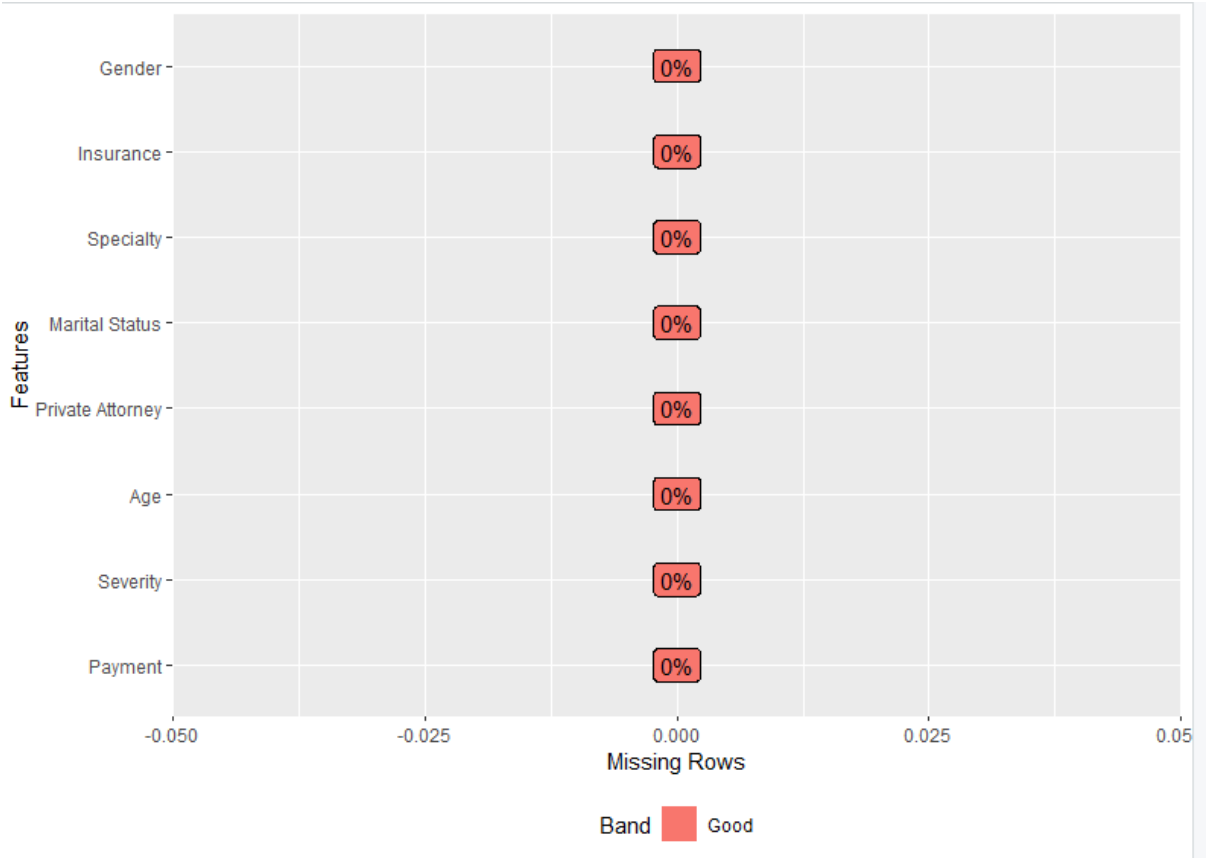


- **describe** function is used to give tabular information in missing and distinct values in the dataset with its proportion.
- describe(Lawsuits)

```
Console ~/
> describe(Lawsuits)
Lawsuits

 8  Variables      118  Observations
---------------------------------------------------------------------------------------------------------------------------
Payment
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     118        0       91        1    673.7      967    14.51    18.56    56.42   168.40   667.85  1861.10  2841.60

lowest :   11.5   11.6   12.2   13.4   14.7, highest: 3934.7 3970.1 5746.1 6301.1 6856.1
---------------------------------------------------------------------------------------------------------------------------
Severity
       n  missing distinct     Info     Mean      Gmd
     118        0        9    0.934     4.72    2.232

lowest : 1 2 3 4 5, highest: 5 6 7 8 9

Value          1     2     3     4     5     6     7     8     9
Frequency      1     2    45    23    13     5    13     5    11
Proportion 0.008 0.017 0.381 0.195 0.110 0.042 0.110 0.042 0.093
---------------------------------------------------------------------------------------------------------------------------
Age
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     118        0       54    0.999    43.14    20.01     18.0     21.7     31.0     41.5     56.0     69.0     73.0

lowest :  2  7 11 12 14, highest: 73 76 78 80 87
---------------------------------------------------------------------------------------------------------------------------
Private Attorney
       n  missing distinct     Info      Sum     Mean      Gmd
     118        0        2    0.672       78    0.661    0.452

---------------------------------------------------------------------------------------------------------------------------
Marital Status
       n  missing distinct     Info     Mean      Gmd
     118        0        5    0.773    2.017   0.9836

lowest : 0 1 2 3 4, highest: 0 1 2 3 4

Value          0     1     2     3     4
Frequency      6    22    71     2    17
Proportion 0.051 0.186 0.602 0.017 0.144
---------------------------------------------------------------------------------------------------------------------------
Specialty
       n  missing distinct
     118        0       20

lowest : Anesthesiology     Cardiology         Dermatology        Emergency Medicine Family Practice
highest: Plastic Surgeon    Radiology          Resident           Thoracic Surgery   Urological Surgery
---------------------------------------------------------------------------------------------------------------------------
```

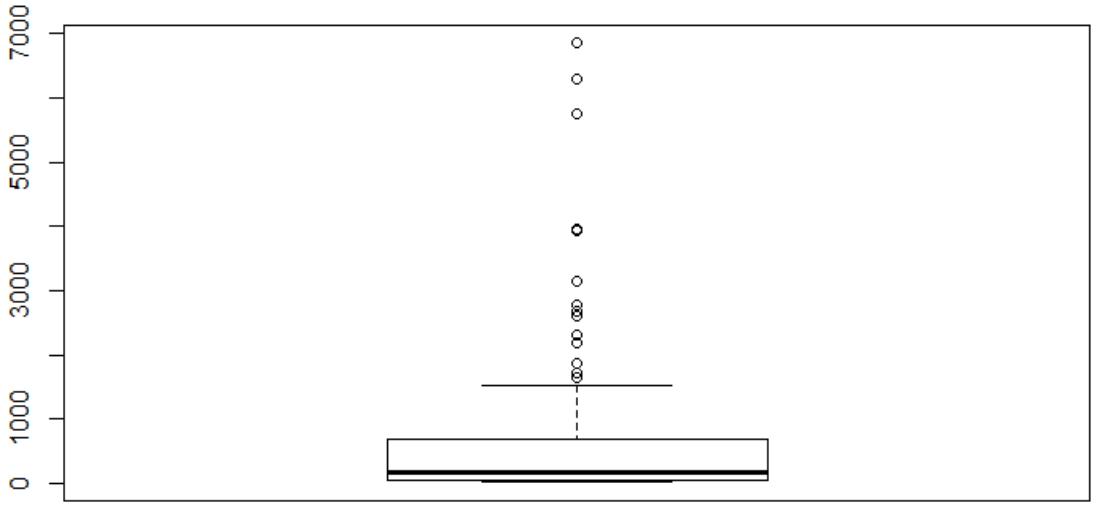
```
Insurance
       n  missing distinct
     118        0        5

lowest : Medicare/Medicaid    No Insurance         Private              Unknown              Workers Compensation
highest: Medicare/Medicaid    No Insurance         Private              Unknown              Workers Compensation

Value          Medicare/Medicaid         No Insurance              Private              Unknown Workers Compensation
Frequency                     16                   12                   51                   36                    3
Proportion                 0.136                0.102                0.432                0.305                0.025
-----------------------------------------------------------------------------------------------------------------------
Gender
       n  missing distinct
     118        0        2

Value      Female   Male
Frequency      71     47
Proportion  0.602  0.398
-----------------------------------------------------------------------------------------------------------------------
```

- **Missing value graph** (Which determines there are no missing values in the dataset)



Finding outliers

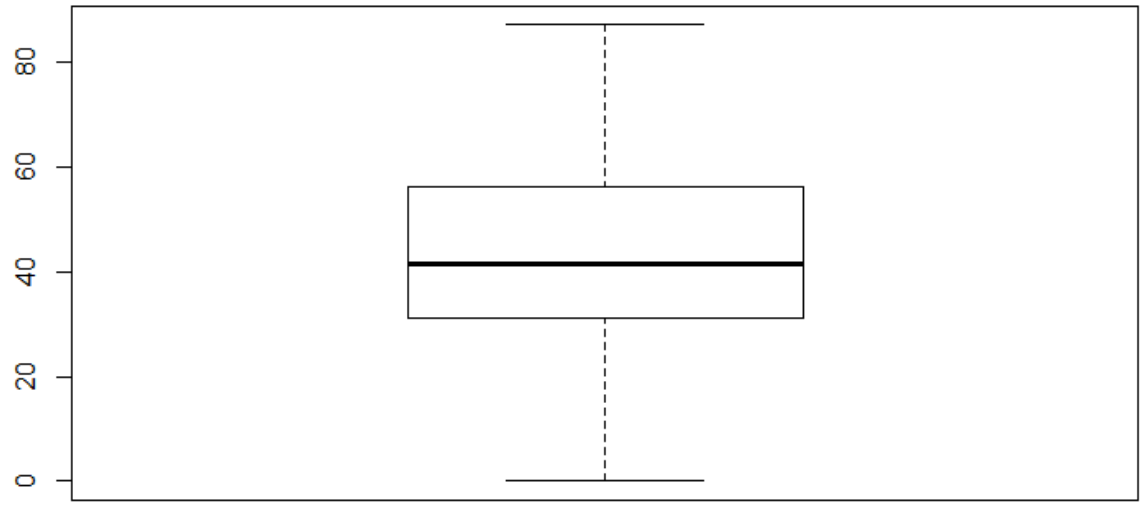
boxplot(Lawsuits$Payment)



- In order to find outliers, if we take on boxplots on each attribute like Payments, Age etc. We can see that only on payments, they are a few outliers above the maximum of interquartile range (Q3-Q1).

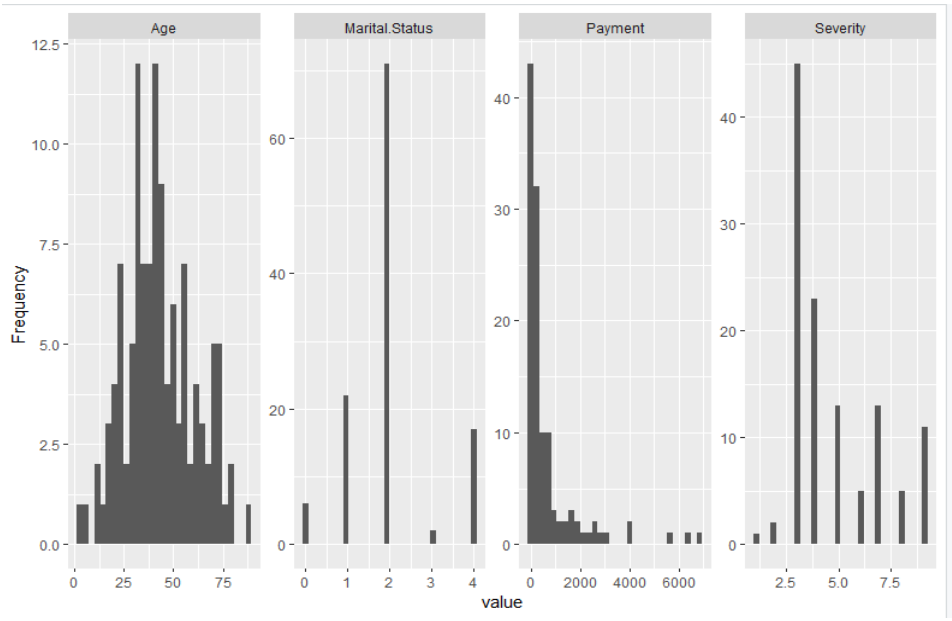
boxplot(Lawsuits$Age)

- For this boxplot, we can see that there are no specific outliers present. All the people who have file the lawsuit seem to fit inside the minimum and maximum value of box plot region.



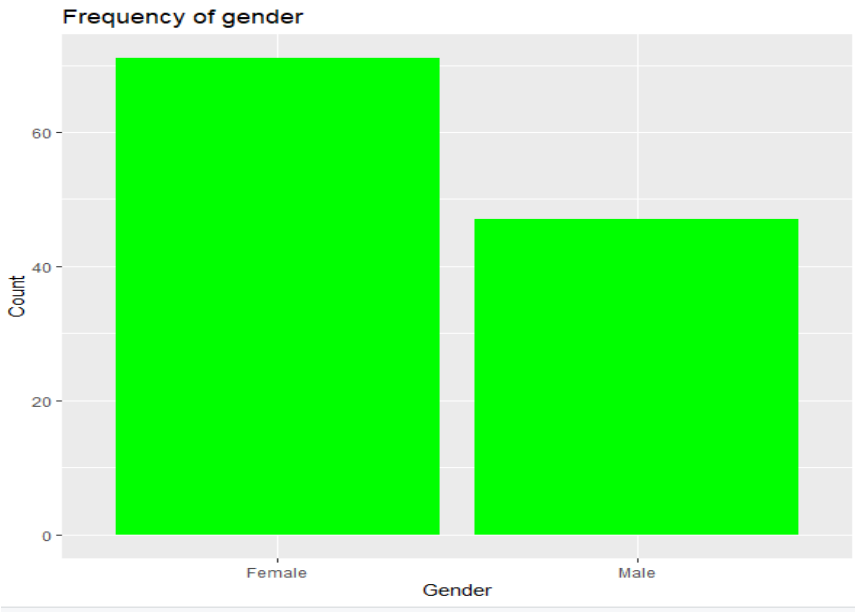
Inferences from univariant charts

Univariate Histogram of variables(numeric)

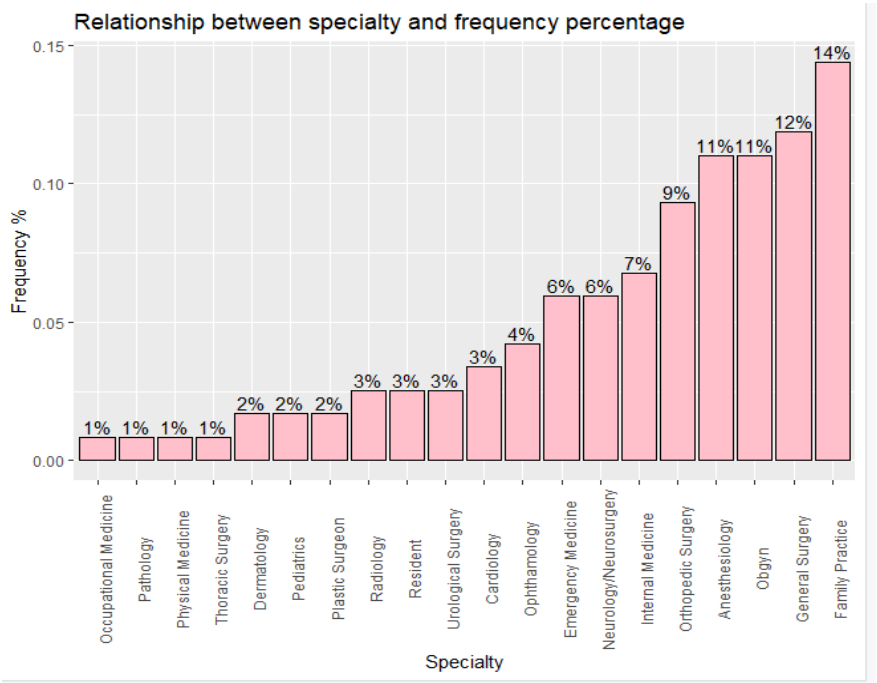


From the above graph, we can conclude the following:

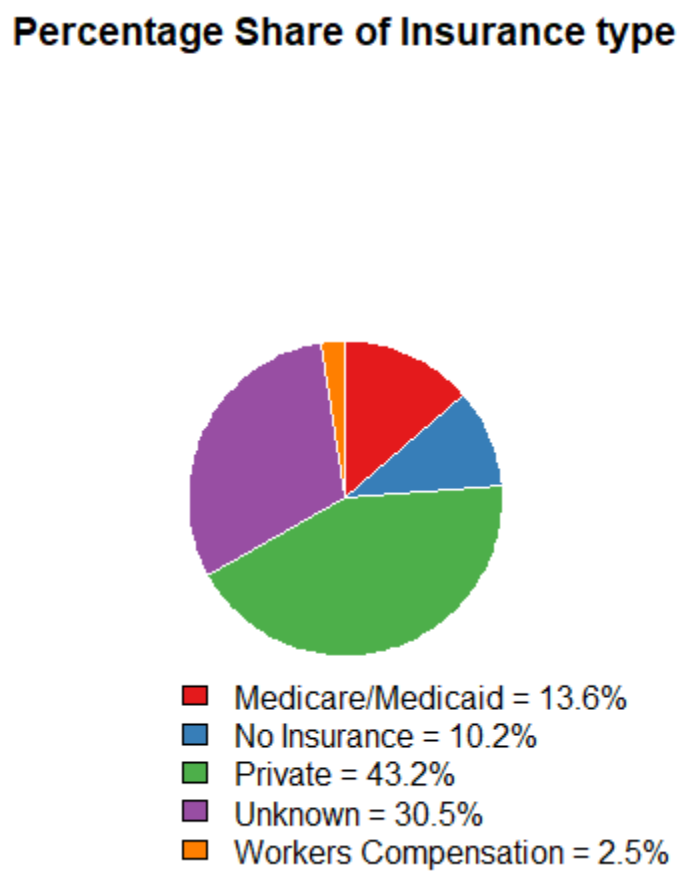
1. Adults who are above 30 and below 40 of age have filed on large frequency against the company.
2. People who are married (2) have filed the most cases while widowed (3) men and women have filed the least number of cases.
3. The company has paid most of its clients within $1000 as compensation.
4. The maximum cases found are of minor temporary damages (45) -severity (3).



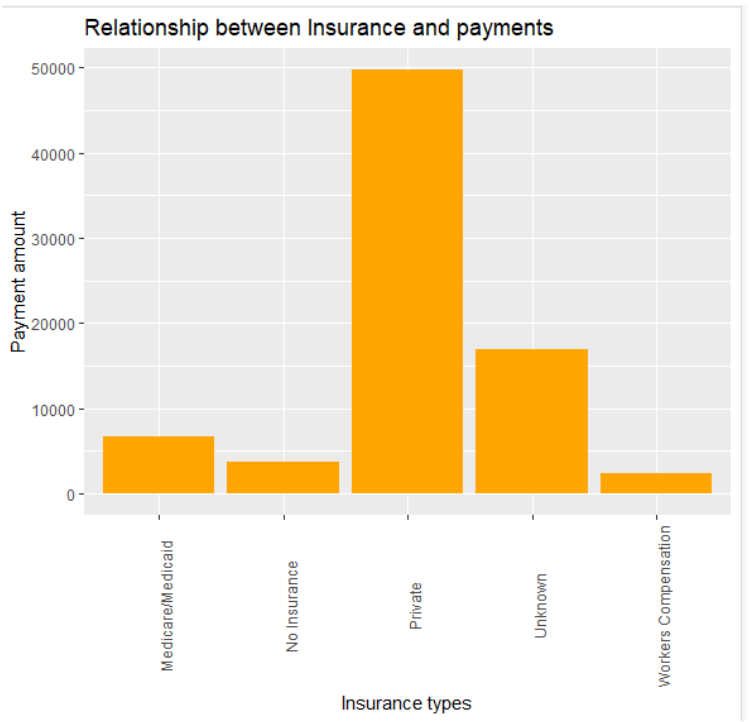
The above graph determines that women have sues the company more than men.



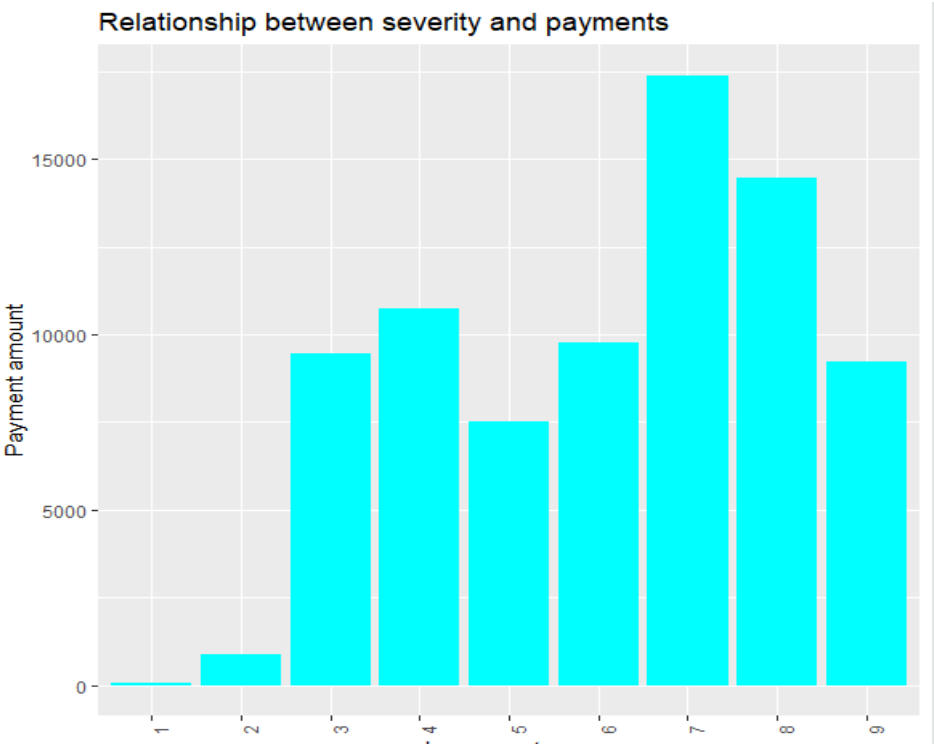
The above graph indicates that doctors practicing on “Family practice” has maximum percentage (14%) of lawsuits filed while “occupational medicine”,” Pathology”, “Physical medicine”, “Thoracic Surgery” have the least number of cases filed against the company.



Inferences from bi-variant charts

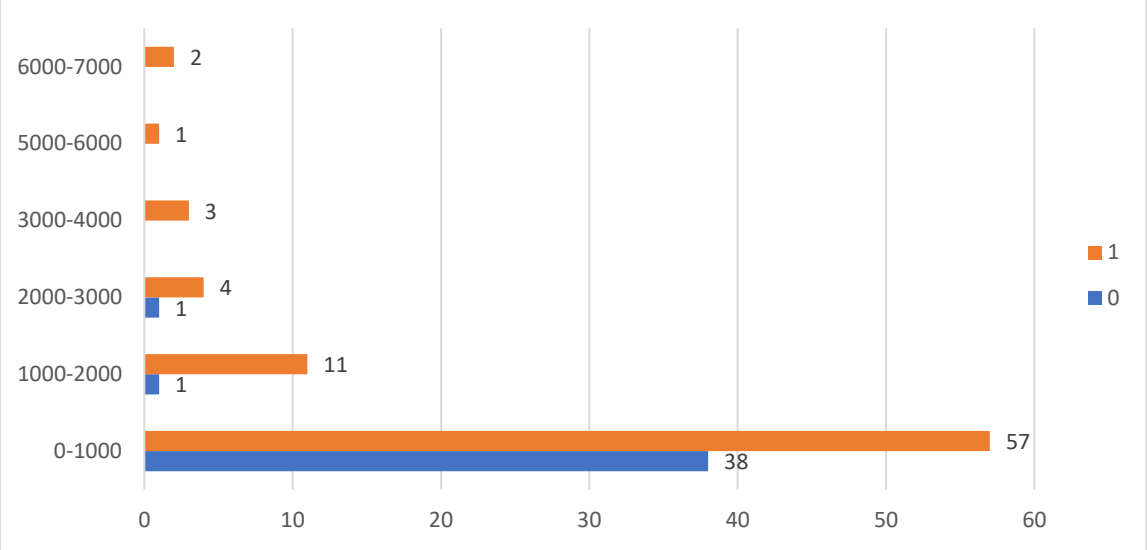


- From the above graph, we can concur that people who have “Private” insurance type have received more payments from the company more than any other category.

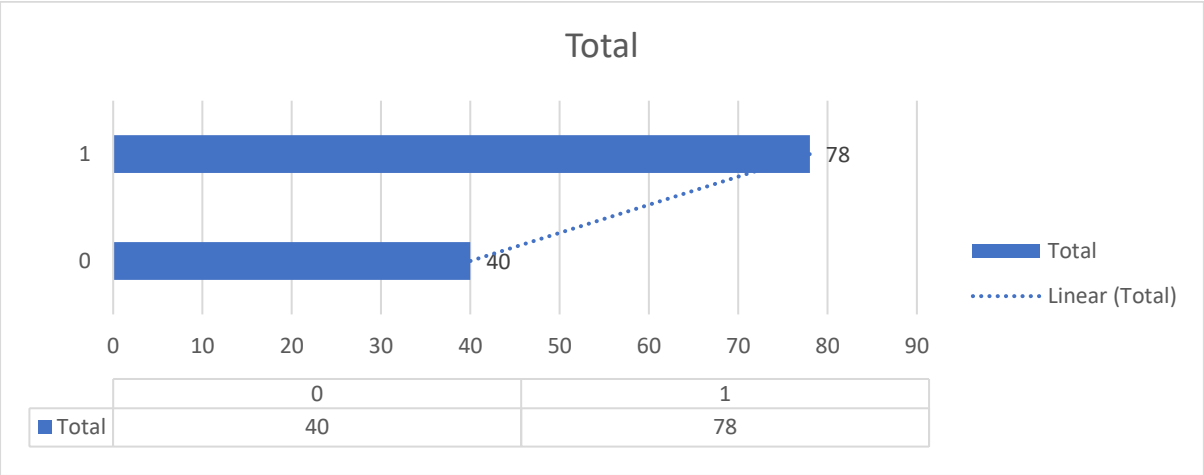


- From the above graph Payments vs Type of Severity, we can conclude that the severity type 7 has the most impact on the company as the company compensated the most with clients sued for major permanent damage(7) whereas the claim for emotional trauma received the least payment from company.

Count of Private Attorney	**Column Labels**		
Row Labels	**0**	**1**	**Grand Total**
0-1000	38	57	95
1000-2000	1	11	12
2000-3000	1	4	5
3000-4000		3	3
5000-6000		1	1
6000-7000		2	2
Grand Total	**40**	**78**	**118**



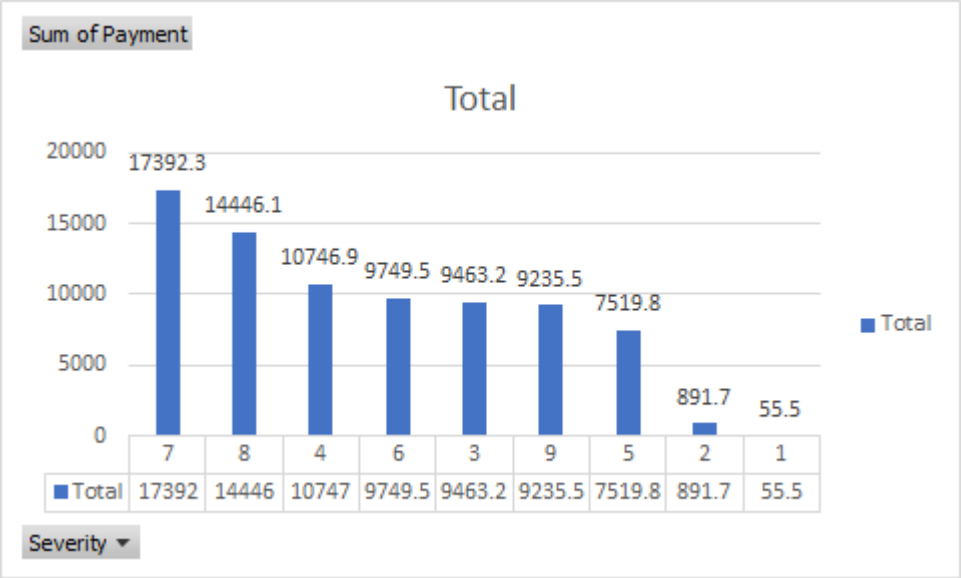
ssRow Labels	Count of Payment
0	40
1	78
Grand Total	**118**

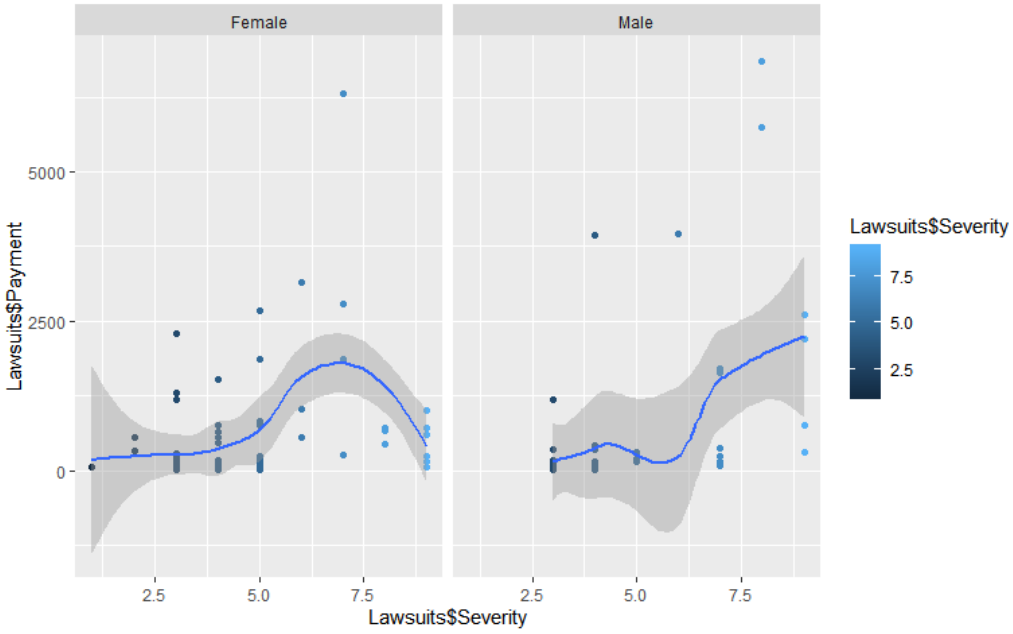


Inferences from multivariate charts

Severity	
Row Labels	**Sum of Payment**
7	**17392.3**
8	**14446.1**
4	**10746.9**
6	**9749.5**
3	**9463.2**

9	**9235.5**
5	**7519.8**
2	**891.7**
1	**55.5**
Grand Total	**79500.5**

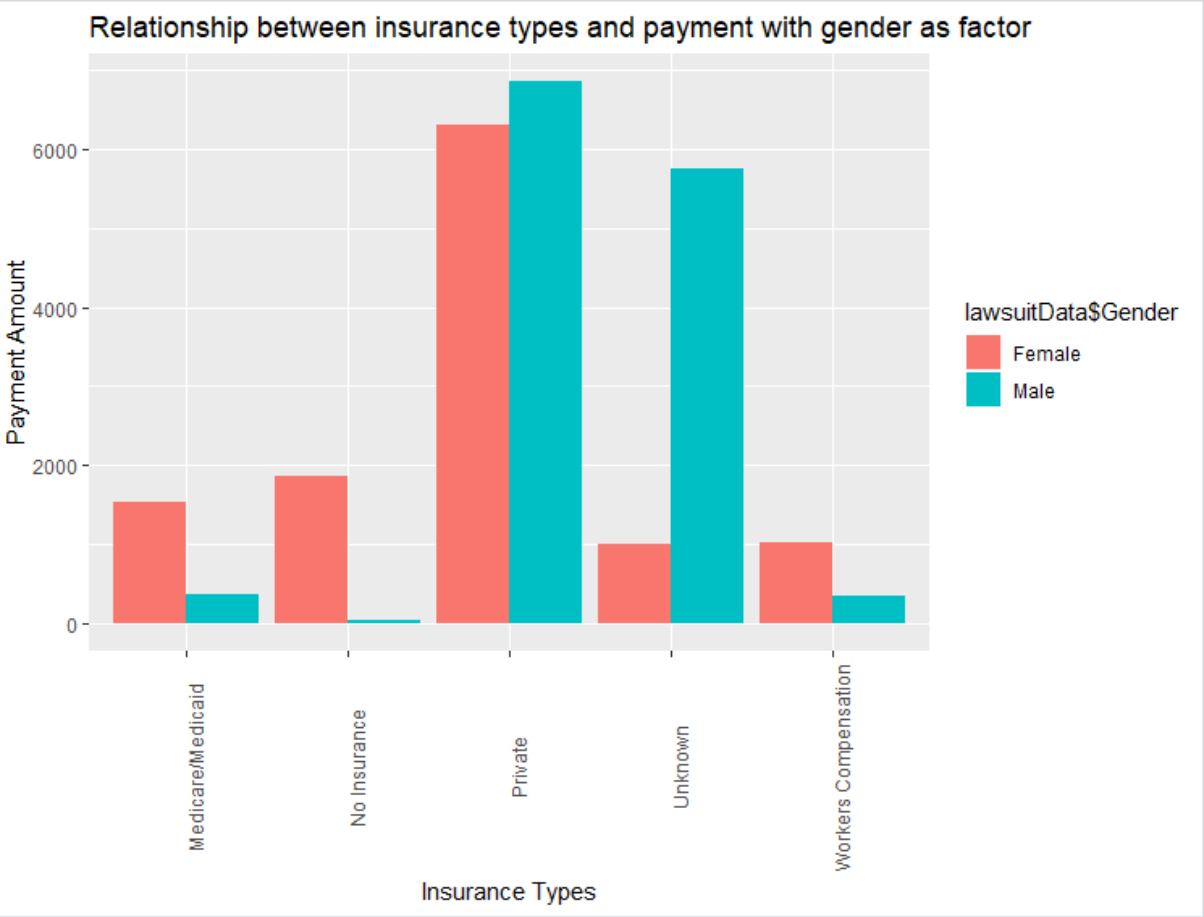




qplot(Lawsuits$Severity, Lawsuits$Payment, data = Lawsuits, color = Lawsuits$Severity ,geom = c("point","smooth"), facets = .~ Lawsuits$Gender)

From the data produced above, we can conclude that there are maximum number of clients sued for minor temporary damage. Although, the company has compensated to 45 people from

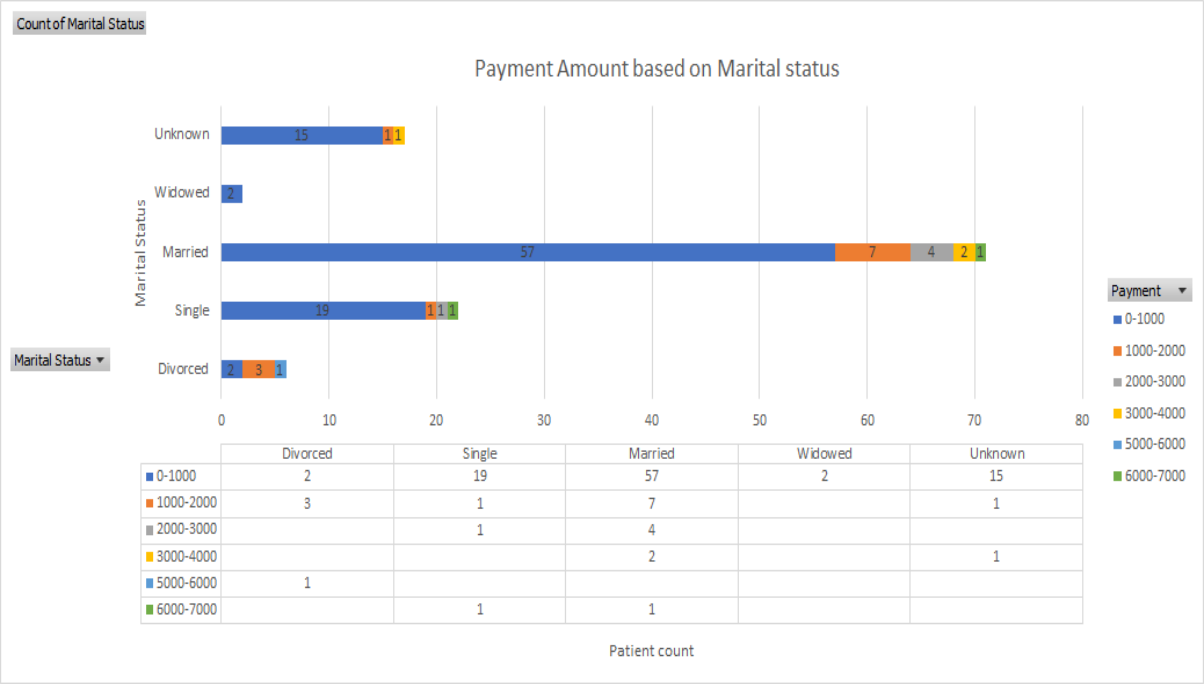
the minor temporary damage, the highest paid severity remained as 7 which is major permanent damage. From the qplot, we can see that married women have filed more complaints than married men.



- The highest payment of the claim is done by male as well as female patients with private insurance which is more than $6000.

Payment made versus marital status

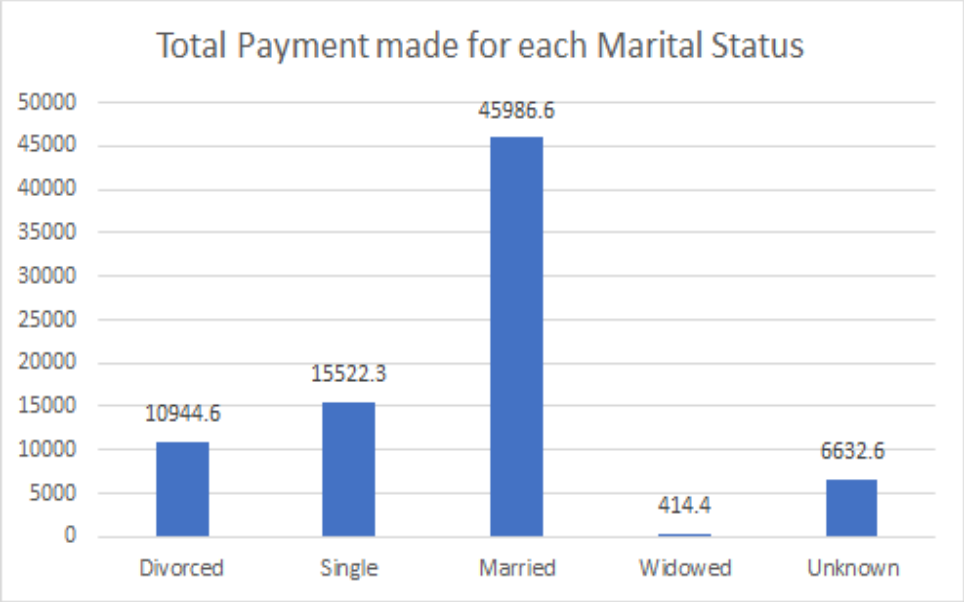
Payment Range	Marital Status						
Count of Marital Status	Column Labels						
Row Labels	0-1000	1000-2000	2000-3000	3000-4000	5000-6000	6000-7000	Grand Total
Divorced	2	3			1		6
Single	19	1	1			1	22
Married	57	7	4	2		1	71
Widowed	2						2
Unknown	15	1		1			17
Grand Total	**95**	**12**	**5**	**3**	**1**	**2**	**118**

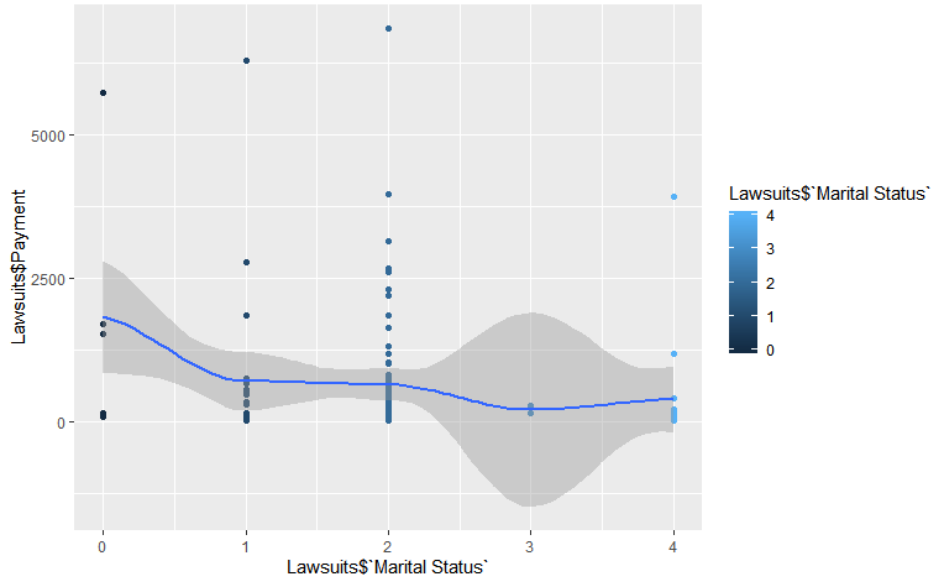


	Divorced	Single	Married	Widowed	Unknown
0-1000	2	19	57	2	15
1000-2000	3	1	7		1
2000-3000		1	4		
3000-4000			2		1
5000-6000	1				
6000-7000		1	1		

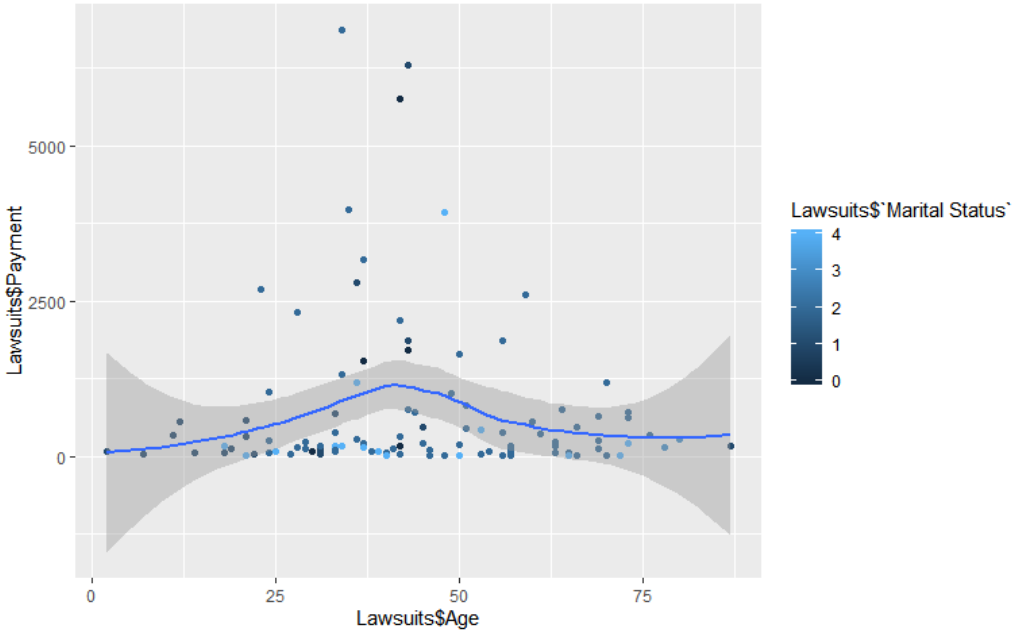
- From the above graph, we can conclude that people who are married filed the greatest number of lawsuits against the company. Although around 71 married people have made the complaint, most of them (57) have been compensated by paying less than $1000. The total expense of the organization for paying married clients was $45986.

	Divorced	Single	Married	Widowed	Unknown
Sum of Payment	**10944.6**	**15522.3**	**45986.6**	**414.4**	**6632.6**





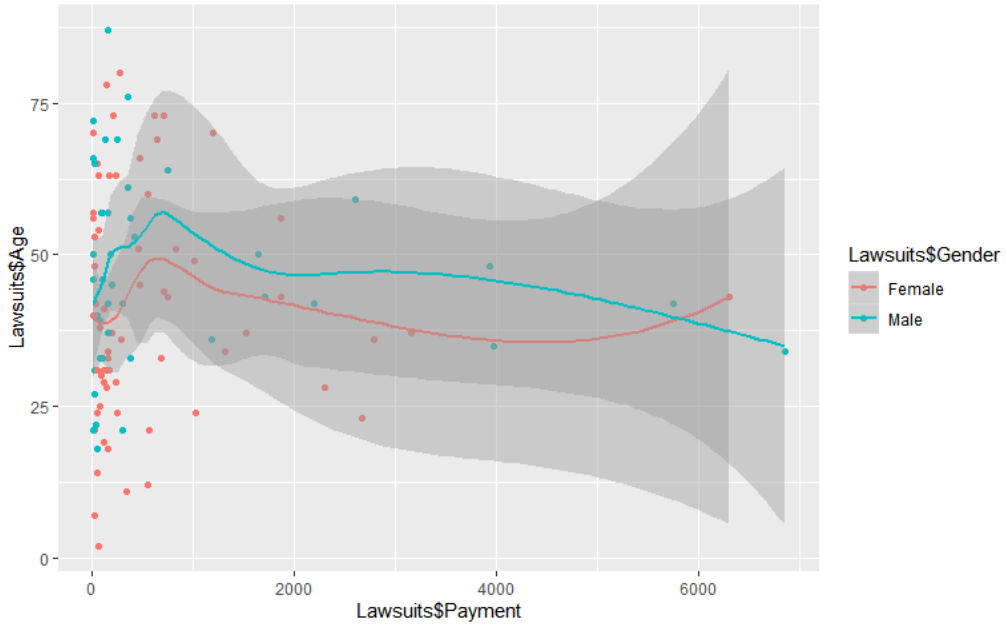
qplot(Lawsuits$`Marital Status` ,Lawsuits$Payment, color = Lawsuits$`Marital Status`,geom = c("point","smooth"))



qplot(Lawsuits$Age, Lawsuits$Payment ,geom = c("point","smooth"), color = Lawsuits$`Marital Status`)

Payment made versus Age with gender as a factor

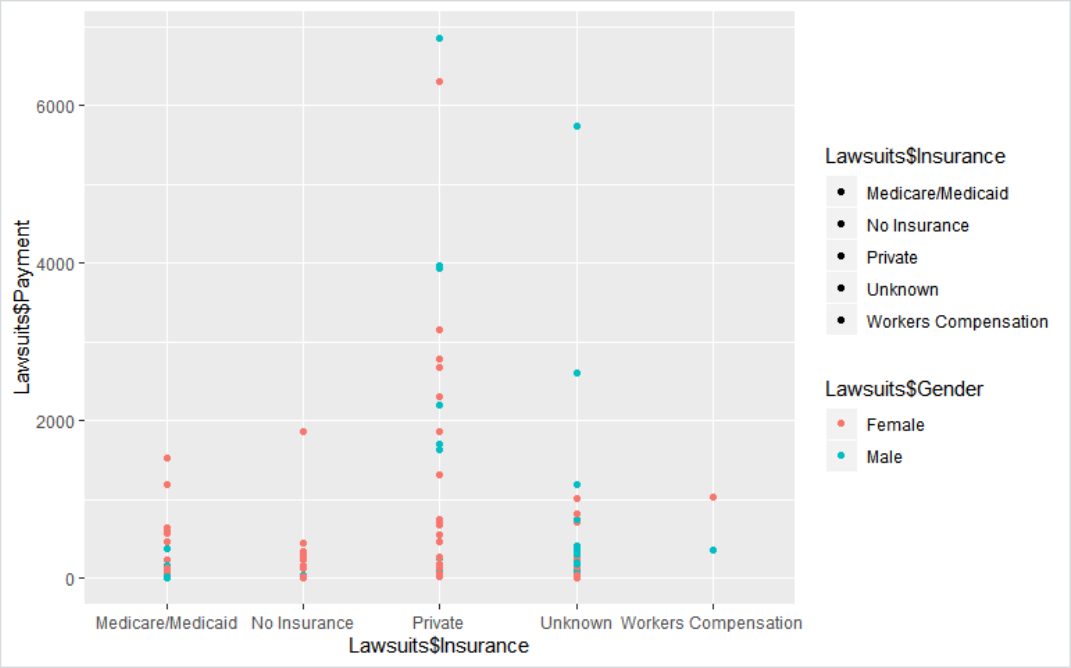
qplot(Lawsuits$Payment,Lawsuits$Age, color = Lawsuits$Gender,geom = c("point","smooth"))



- With the above qplot function, we can calculate that people within 30-40 are the highest in number to file a lawsuit. And analysing closely, we can identify that most of cases who got compensated higher than $2000 are male who age between 25-50.

Payment versus Insurance with gender as factor

qplot(Lawsuits$Insurance, Lawsuits$Payment, color = Lawsuits$Gender, fill = Lawsuits$Insurance)

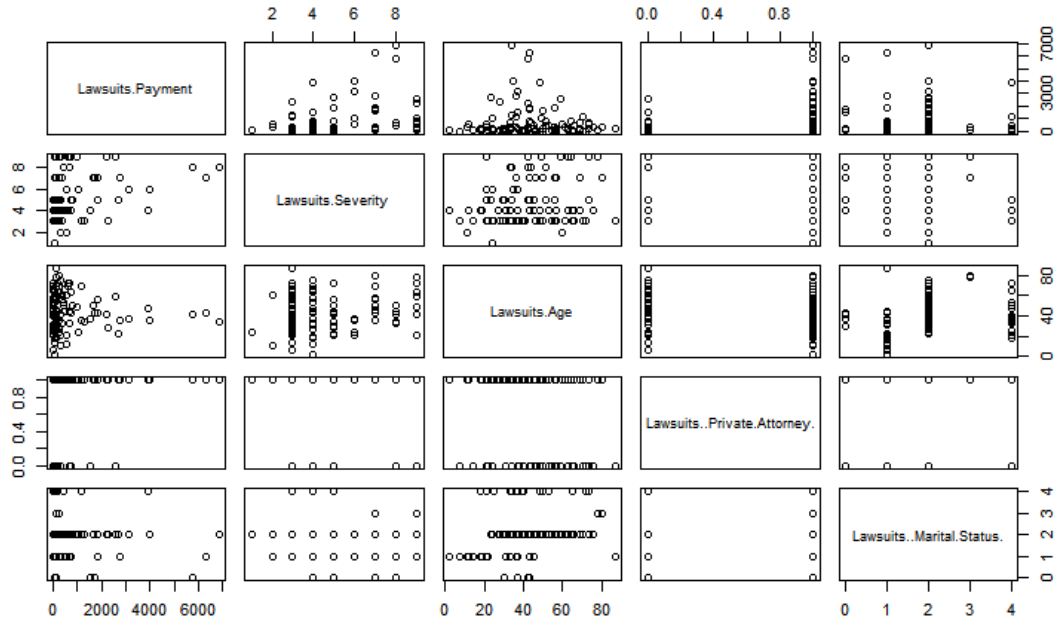


- From the above graph, Private insurance holders have helped the customers file more lawsuits against company than any other insurance type. Here, we could see that both male and female count are equally distributed.

Scatter plot Observation for correlation

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

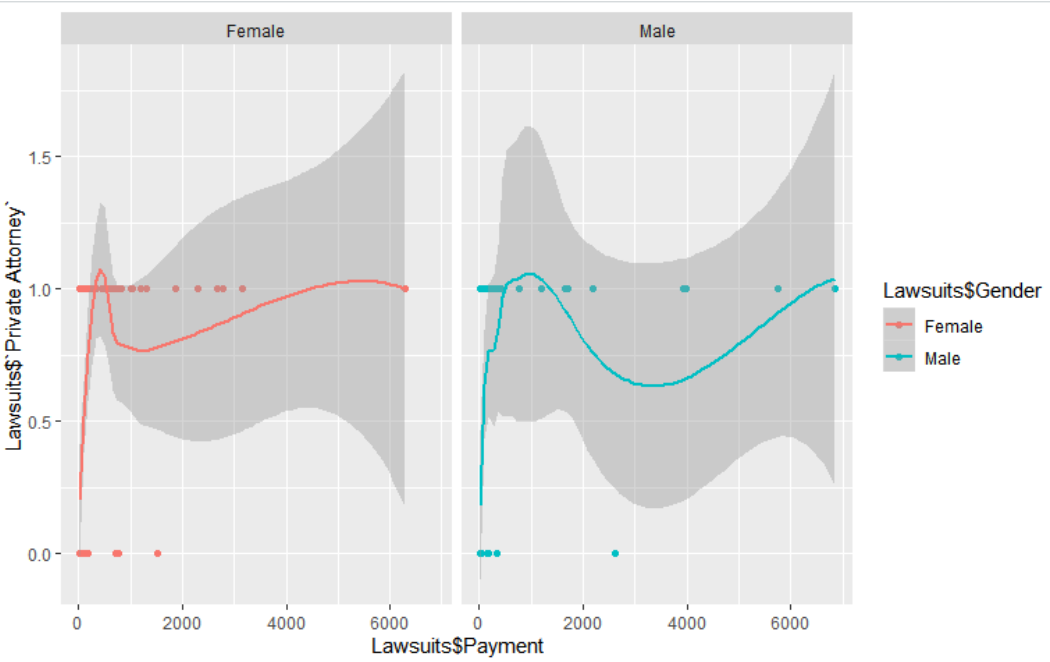
pairs(numData)



- With this scatter plot, we cannot see any strong correlation between any attributes. We can see a very weak correlation with Age vs Payment but that too has many outliers. Taking logarithmic values can give us better understanding of attributes.

Payment versus Private attorney based on gender.

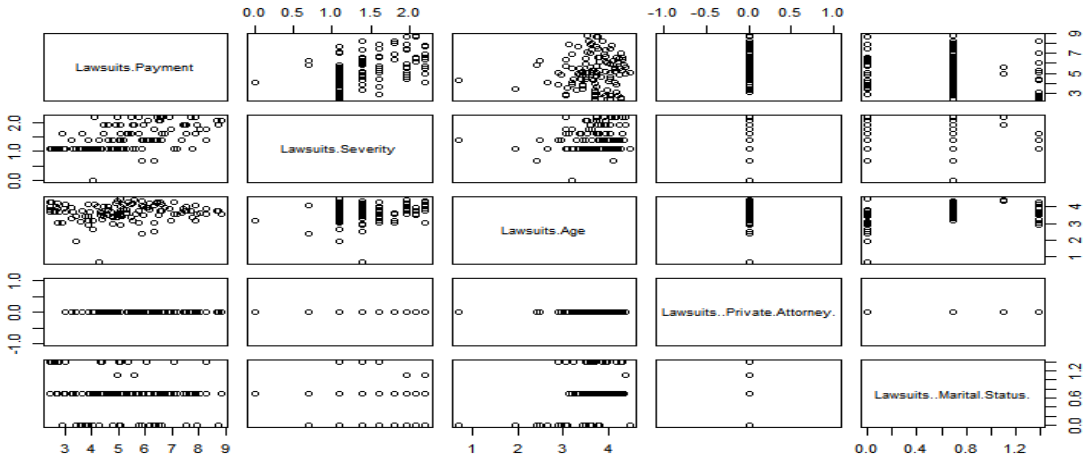
qplot(Lawsuits$Payment,Lawsuits$`Private Attorney`, data = Lawsuits, color = Lawsuits$Gender ,geom = c("point","smooth"), facets = .~ Lawsuits$Gender)



From the graphs above, we can see that people who hired private attorneys are more than people who have not hired. And, because of the outlier present in the graph representing male count, we can see the curve dipping towards value"0" around $2000 and rising again.

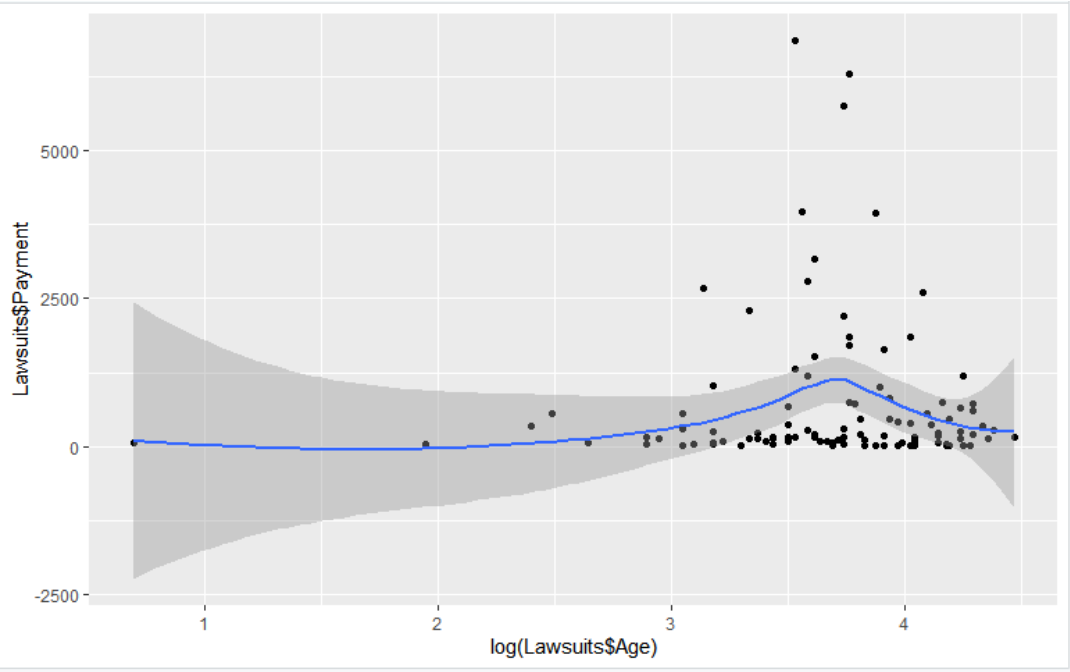
Logarithmic Transformed Graphs

pairs(log(numData))



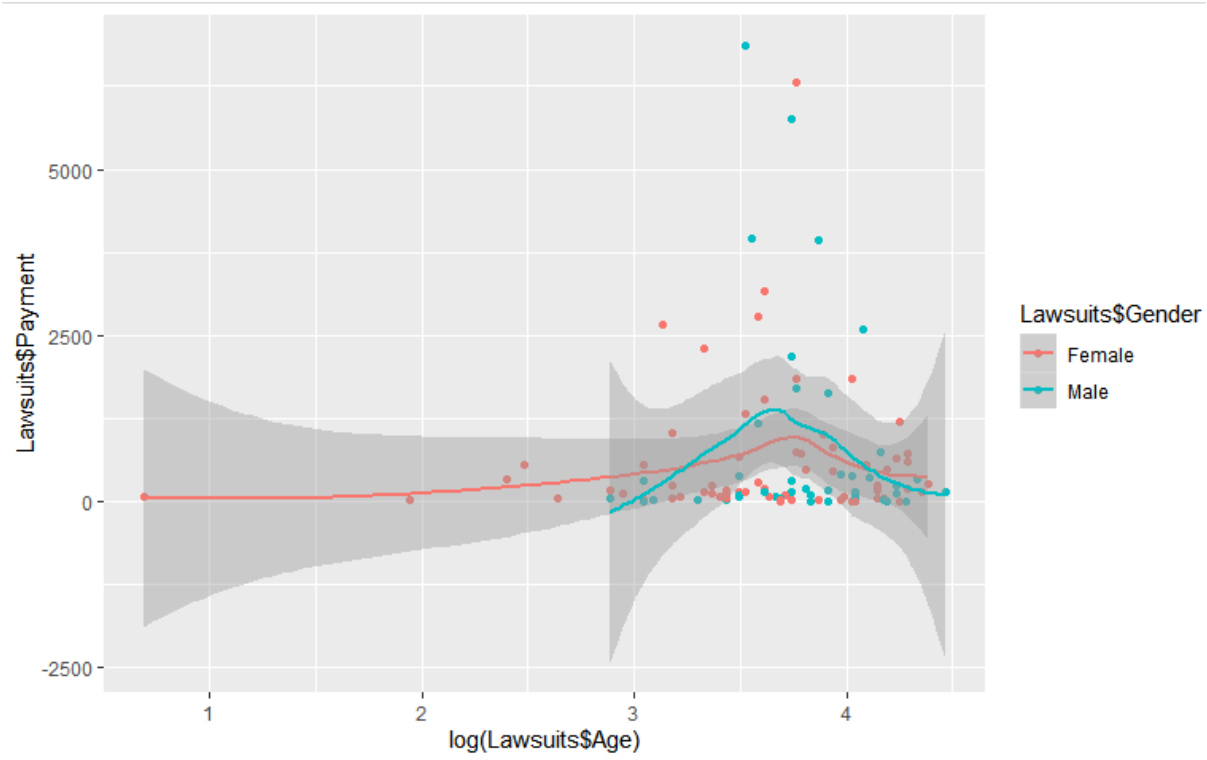
- From the logarithmic values, we can see many correlations between attributes by creating a scatter plot. Following correlations are found from the graph:
- Strong relationship between private attorney and every other attribute in the dataset.
- Correlation between Age vs payment.

qplot(log(Lawsuits$Age),Lawsuits$Payment, data = Lawsuits,geom = c("point","smooth"))



- These graphs plot the relationship between Lawsuit's payment attribute with Age and also with gender. We can see that there are few outliers on payments that has to be considered as they are legit values and can influence the modelling on mean, mode values. Also , these outliers has to be noted carefully as they show that company has paid more amount of money to some of the categories.

qplot(log(Lawsuits$Age), Lawsuits$Payment, data = Lawsuits, geom = c("point","smooth"), color = Lawsuits$Gender)



cor(log(Lawsuits$Age), Lawsuits$Payment)

```
Console ~/
> cor(log(Lawsuits$Age), Lawsuits$Payment)
[1] 0.03751455
>
```

With this correlation function, we can see that between payment and age, there is a weak correlation link of 0.037.