FINAL PROJECT

Priya Marla, Susmita Madineni, Muneer Ahmed, Lokesh Tangella

2023-05-27

Contents

1 Introduction

The insurance claim dataset contains insightful information related to insurance claims giving us an in-depth look into demographic patterns of those who are claiming it. The demographic information contains information like the age, gender and other health related paramters such as blood pressure etc. of a patient. Based on these parameters how much insurance amount is claimed by that patient is captured too. This information allows to performed supervised learning on model such as linear regression etc and use this model to predict the insurance claim of new patients based on their demographic patterns as close as possible. These kinds of models can help the insurance agencies/companies to make wiser decisions when considering potential customers for their services. Moreover, this information can inform public policy by allowing for more targeted support and identify the patients who are in most need of the insurance and vulnerable.

2 Data

Let's import the required libraries below:

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

```
## Loading required package: mosaic
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Attaching package: 'mosaic'
## The following object is masked from 'package:Matrix':
##
##     mean
## The following object is masked from 'package:ggplot2':
##
##     stat
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(rapport)
```

```
## Warning: package 'rapport' was built under R version 4.2.3
```

```
library(ggplot2)
library(lattice)
library(stats)
```

2.1 Reading data

The dataset is stored in form of comma seperated values(csv) in a file named insurance_data.csv. This file is imported in the below cell:

```
#reading
initial_insurance <- read.csv("./insurance_data.csv")
```

2.2 Attribute Information

The data set we have used contains the following columns:

- **PatientID:** This is an identifier for the person and contains 1340 records of people.
- **Age:** This is the age of the person in question.The age of the patients ranges from 18 to 60 years.
- **Gender:** This is the gender of the person in question.
- **BMI:** This is the Body Mass Index(weight/height^2) of the person in question.The body mass index (BMI) of the patients ranges from 16 to 53.1.
- **Bloodpressure:** This is the Blood Pressure of the person recorded during the examination and ranges from 80 to 140.
- **Diabetic:** This is an indicator variable, if the person has diabetic or not.

- **Children:** This indicates number of children a person has and ranges from 0 to 5.
- **Smoker:** This is an indicator of if the person smokes or not.
- **Region:** This is the region from which the person is.
- **Claim:** This is the claims made by of the person. The minimum claim is 1122, the 25th percentile is 4720, the median is 9370, the mean is 13253, the 75th percentile is 16604, and the maximum claim is 63770.

Note: The null values in the age column and empty strings in region column indicate that the data is unidentified.

```
#summary
summary(initial_insurance)
```

```
##      index          PatientID           age           gender
##  Min.   :   0.0   Min.   :   1.0   Min.   :18.00   Length:1340
##  1st Qu.: 334.8   1st Qu.: 335.8   1st Qu.:29.00   Class :character
##  Median : 669.5   Median : 670.5   Median :38.00   Mode  :character
##  Mean   : 669.5   Mean   : 670.5   Mean   :38.08
##  3rd Qu.:1004.2   3rd Qu.:1005.2   3rd Qu.:47.00
##  Max.   :1339.0   Max.   :1340.0   Max.   :60.00
##                                    NA's   :5
##       bmi        bloodpressure      diabetic            children
##  Min.   :16.00   Min.   : 80.00   Length:1340        Min.   :0.000
##  1st Qu.:26.27   1st Qu.: 86.00   Class :character   1st Qu.:0.000
##  Median :30.40   Median : 92.00   Mode  :character   Median :1.000
##  Mean   :30.67   Mean   : 94.16                      Mean   :1.093
##  3rd Qu.:34.70   3rd Qu.: 99.00                      3rd Qu.:2.000
##  Max.   :53.10   Max.   :140.00                      Max.   :5.000
##
##     smoker             region              claim
##  Length:1340        Length:1340        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4720
##  Mode  :character   Mode  :character   Median : 9370
##                                        Mean   :13253
##                                        3rd Qu.:16604
##                                        Max.   :63770
##
```

```
dim(initial_insurance)
```

```
## [1] 1340   11
```

2.3 Cleaning Data

We need to clean the data before we perform any analysis,tests on it. We have taken the following steps:

1) We have removed the index, children and patientID column as they don't have any effect on our analysis. The index and patientID are kind of unique identifiers of each patient and has no effect on the insurance claim, while no.of children does not inform about the patients demographic details hence it is removed too.
2) We have removed all the rows which contained null values for any of the columns as this you not give us a complete picture when we analyse the data.
3) Similarly we have removed the empty strings.

```
# removing index and children columns
insurance <- initial_insurance %>% select(-c(PatientID, index, children))

# removing null values rows
insurance <- na.omit(insurance)

#removed empty string
insurance <- insurance[insurance$region != "", ]

#cleaned data statistics
dim(insurance)
```

```
## [1] 1332    8
```

```
summary(insurance)
```

```
##       age           gender               bmi        bloodpressure
##  Min.   :18.00   Length:1332        Min.   :16.00   Min.   : 80.00
##  1st Qu.:29.00   Class :character   1st Qu.:26.20   1st Qu.: 86.00
##  Median :38.00   Mode  :character   Median :30.35   Median : 92.00
##  Mean   :38.09                      Mean   :30.66   Mean   : 94.19
##  3rd Qu.:47.00                      3rd Qu.:34.73   3rd Qu.: 99.00
##  Max.   :60.00                      Max.   :53.10   Max.   :140.00
##    diabetic            smoker             region              claim
##  Length:1332        Length:1332        Length:1332        Min.   : 1122
##  Class :character   Class :character   Class :character   1st Qu.: 4760
##  Mode  :character   Mode  :character   Mode  :character   Median : 9413
##                                                           Mean   :13325
##                                                           3rd Qu.:16781
##                                                           Max.   :63770
```

After cleaning the data the following have changed:

```
The dataset now contains 1332 records of people.
```

```
The dataset contains information on the claim amounts. The minimum claim is 1122, the 25th percentile is
```

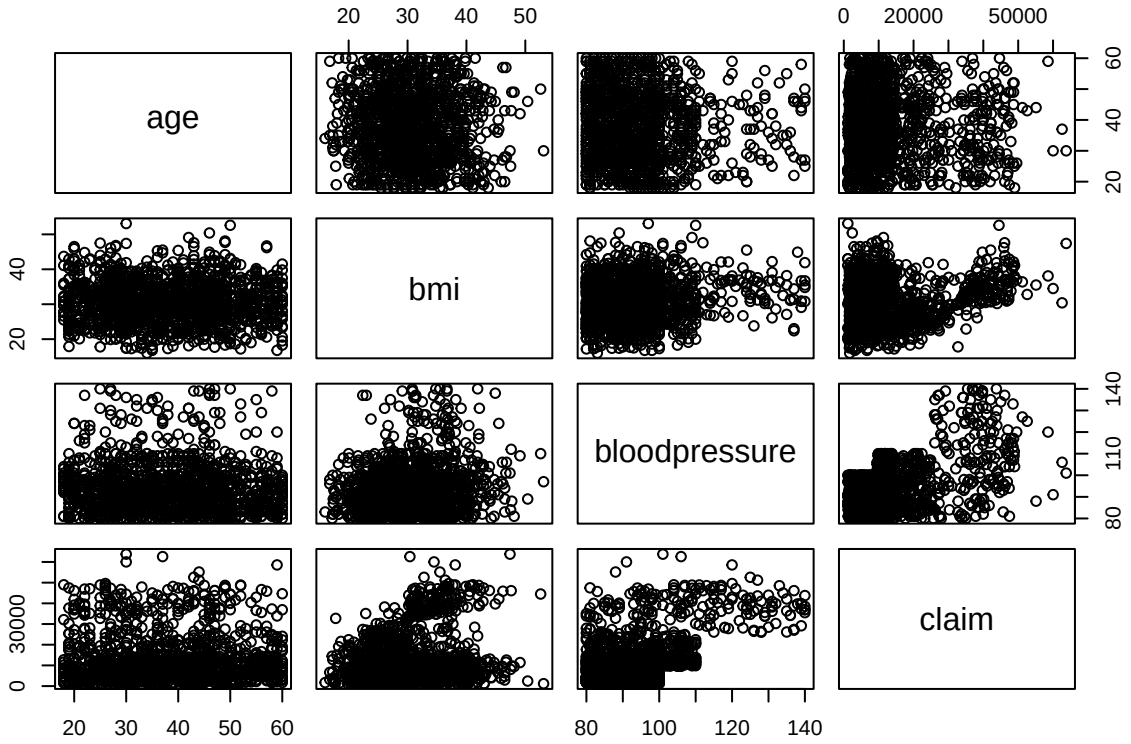
These changes can be attributed to removing the empty values.

3 Analysis

3.1 Linear regression

Paired Scatter Plots:

```
analysis <- insurance %>% select(-c(gender,diabetic,smoker,region))
pairs(analysis)
```



From the above paid scatter plot, we can see that there is a slight relation between blood pressure and the amount of claims made.

```
m4 <- lm(claim ~ bloodpressure , data=insurance)
summary(m4)
```

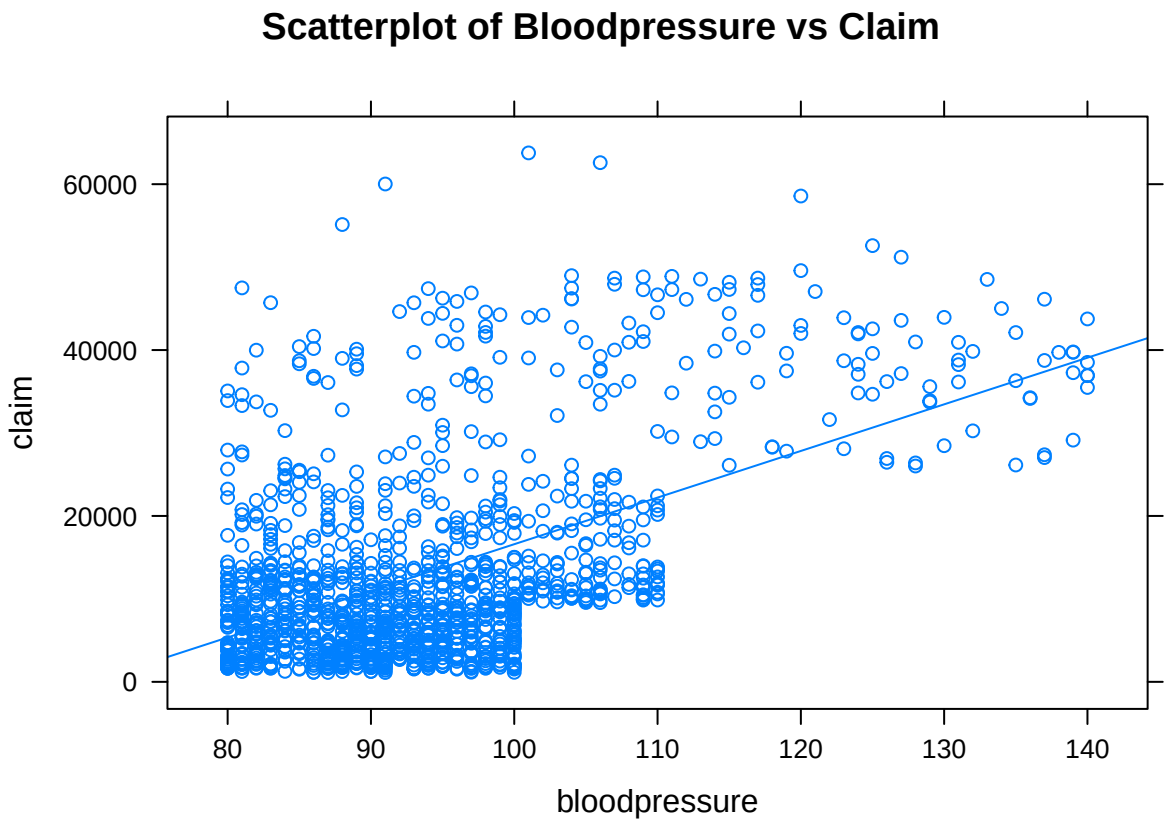
Bloodpressure vs Claim

```
##
## Call:
## lm(formula = claim ~ bloodpressure, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15445  -7025  -2667   3859  48489
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -39619.05    2332.13  -16.99   <2e-16 ***
## bloodpressure    562.11      24.58   22.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10260 on 1330 degrees of freedom
## Multiple R-squared:  0.2822, Adjusted R-squared:  0.2817
## F-statistic:   523 on 1 and 1330 DF,  p-value: < 2.2e-16
```

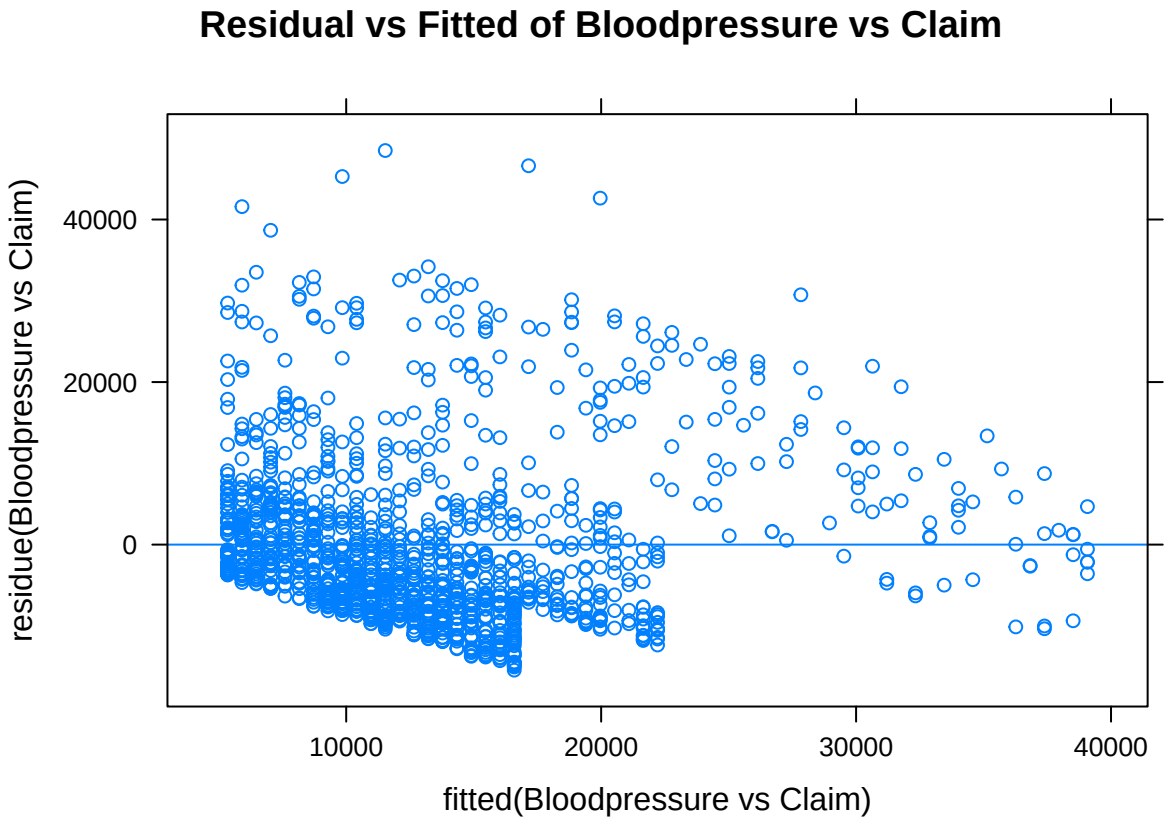
```
cor.test(insurance$bloodpressure, insurance$claim, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 22.869, df = 1330, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4915791 0.5687458
## sample estimates:
##       cor
## 0.5312634
```

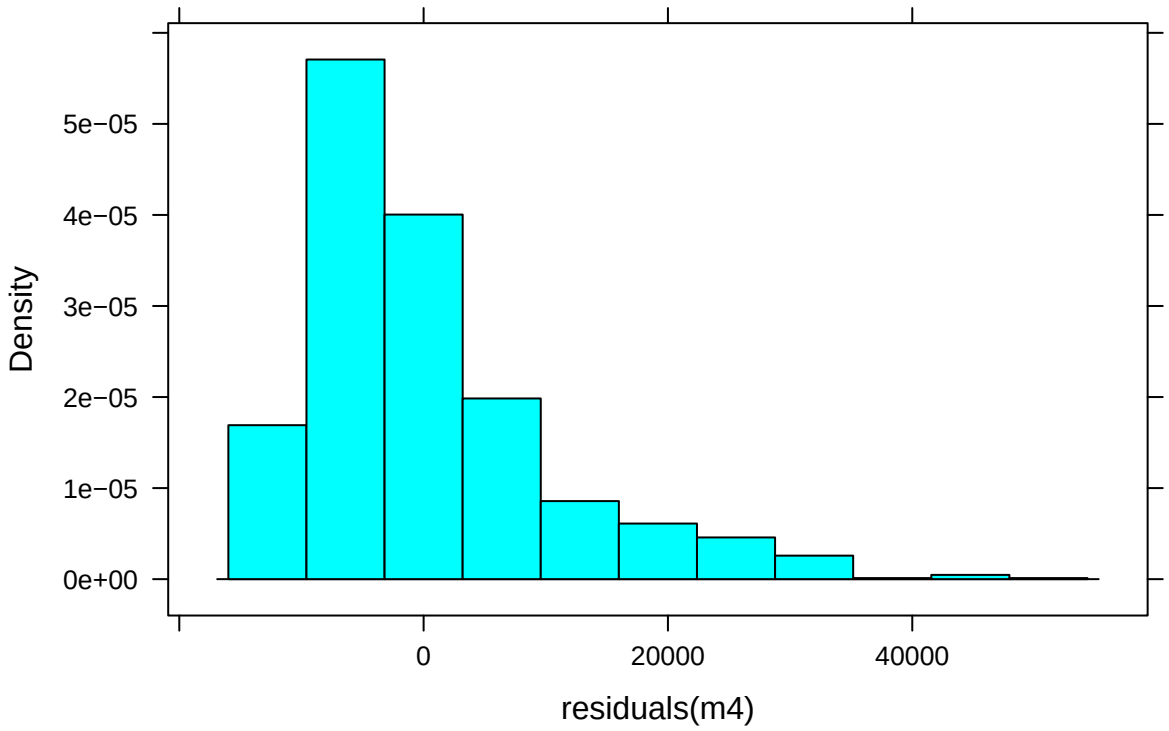
```
xyplot(claim ~ bloodpressure, data=insurance, type=c("p", "r"), main="Scatterplot of Bloodpressure vs C
```




```
xyplot(resid(m4) ~ fitted(m4), data=insurance, type=c("p","r"), main="Residual vs Fitted of Bloodpressu
```




```
histogram(~residuals(m4), main = "Histogram for residuals")
```



Assumptions:

- **Linearity**: The relationship between blood pressure and claim is moderately linear and positively correlated.
- **Normal errors**: From the histogram of the residuals, we can say that it has a normal distribution and there seems to be few outliers
- **Equal Variance**: From the Residual vs fitted plot, we can see that the data points are around zero, hence assuming equal variance
- **Independence**: The data points are independent as each person has his own data.

From the summary R-squared = 0.2822, which is less than 0.3, thus this is weak model. Whereas correlation coefficient is 0.5312 which is greater than 0.5, thus blood pressure vs claim has moderate linear positive correlation i.e if the independent variable blood pressure increases, the dependent variable claim also increases sometimes.

```
m1 <- lm(claim ~ bloodpressure+bmi+age+region+diabetic+smoker+gender , data=insurance)
m2 <- lm(claim ~ diabetic+smoker+bloodpressure , data=insurance)
m3 <- lm(claim ~ smoker , data=insurance)
```

```
summary(m1)
```

All columns vs claim

```
##
## Call:
## lm(formula = claim ~ bloodpressure + bmi + age + region + diabetic +
```

```
##     smoker + gender, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17202.2  -4219.7   -866.8   3180.5  31007.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -21640.06    2029.73 -10.662  < 2e-16 ***
## bloodpressure      226.01      17.76  12.729  < 2e-16 ***
## bmi                354.47      30.91  11.467  < 2e-16 ***
## age                 16.46      17.85   0.922 0.356709
## regionnorthwest  -1911.60     567.42  -3.369 0.000776 ***
## regionsoutheast  -2850.34     547.95  -5.202 2.29e-07 ***
## regionsouthwest  -2128.70     578.94  -3.677 0.000246 ***
## diabeticYes       -300.36     364.88  -0.823 0.410551
## smokerYes        20718.21     499.04  41.516  < 2e-16 ***
## gendermale          15.37     397.27   0.039 0.969138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6629 on 1322 degrees of freedom
## Multiple R-squared:  0.7024, Adjusted R-squared:  0.7004
## F-statistic: 346.7 on 9 and 1322 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = claim ~ diabetic + smoker + bloodpressure, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18717  -4400  -1198   3409  32183
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -15074.92    1712.91  -8.801   <2e-16 ***
## diabeticYes     -416.34     382.42  -1.089    0.276
## smokerYes      20521.43     520.21  39.448   <2e-16 ***
## bloodpressure    258.82      18.38  14.084   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6969 on 1328 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6688
## F-statistic: 896.8 on 3 and 1328 DF,  p-value: < 2.2e-16
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = claim ~ smoker, data = insurance)
##
## Residuals:
```

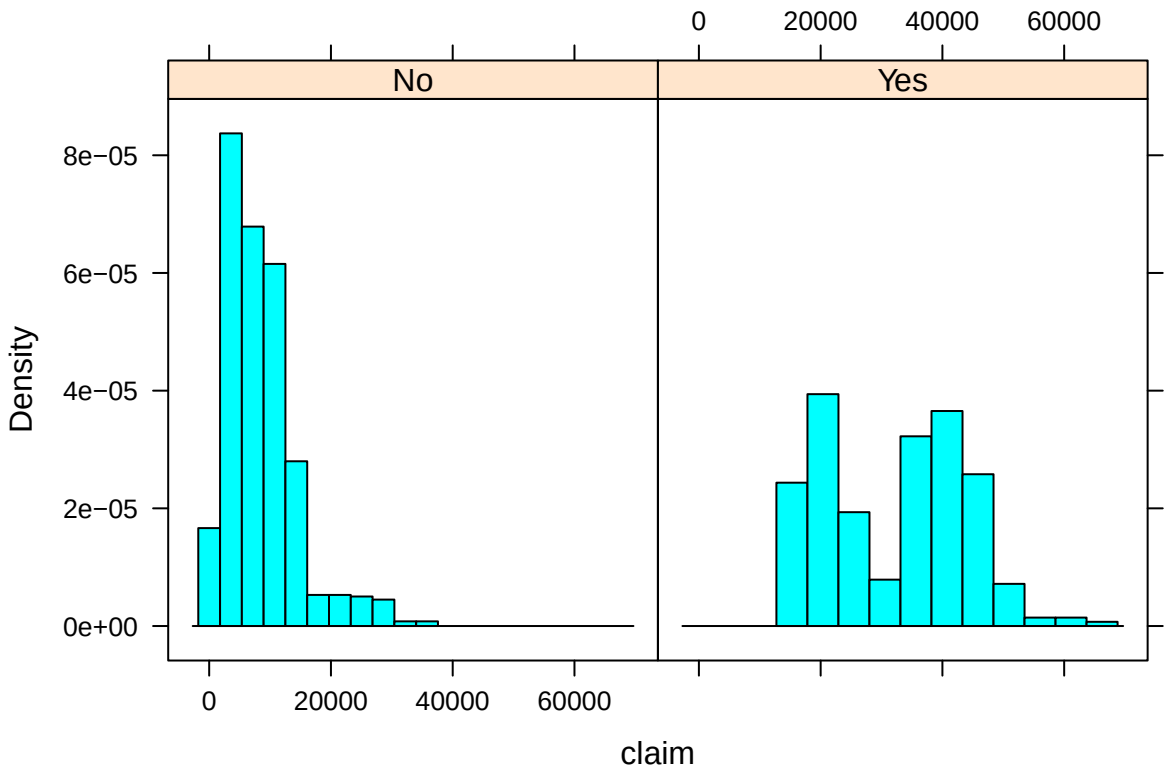
```
##    Min     1Q Median     3Q    Max
## -19221  -5017   -896   3690  31720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8475.9      229.7   36.90   <2e-16 ***
## smokerYes    23574.4      506.4   46.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7471 on 1330 degrees of freedom
## Multiple R-squared:  0.6197, Adjusted R-squared:  0.6194
## F-statistic:  2167 on 1 and 1330 DF,  p-value: < 2.2e-16
```

Based on the above models m1, m2, m3, The column smoker impacts the claim amount majorly than any other demographic details of a given patient. we can see this from the R-squared value for the linear model for smoker vs claim. Another observation supporting this statement is that the R-squared value increases only slightly when the claim is made to be dependent on smoker + some other columns(demographic details). Consider the below plot analzing the insurance claimed by patients who smoke vs who do not smoke.

```
favstats (~claim |smoker, data=insurance)
```

```
##   smoker      min       Q1    median       Q3      max      mean       sd    n
## 1     No  1121.87  4034.32  7364.975 11365.28 36910.61  8475.865  5985.19 1058
## 2    Yes 12829.46 20826.24 34456.350 41019.21 63770.43 32050.232 11541.55  274
##   missing
## 1       0
## 2       0
```

```
histogram (~claim |smoker, data=insurance)
```



Based on the above statistics, the average claim value for a non-smoker is 8475.865, and the average claim value for a smoker is 32050.23. The average claim value for a smoker is four times of a non-smoker. This clearly shows that smokers has a high impact on the claim value.

Histogram for a non-smoker's claims is right-skewed, it means that the data is concentrated towards the left side and has a longer tail towards the right side. This indicates that there are relatively more low claim values and a few high values.

3.2 Tests

Assumptions:

- **Random and Independent**: The data plots do not follow any patterns and hence we can say it is random. Moreover, the data is of individual persons, hence it is independent of other sample.
- **Normally distributed Sample**: From the histogram of residuals vs density above, we can say that the sample is Normal.
- **Equal Variance**: From paired scatter plot, we can see that the data points are around zero, hence assuming equal variance

1. Test the hypothesis that the mean blood pressure of the patient is greater than 120 (Normal Human blood pressure).

Hypothesis:

H0: mean blood pressure is less than or equal to 120

Ha: mean blood pressure is greater than 120

Test Statistics:

```r
t.test(~ bloodpressure, data=insurance, alternative="greater", mu=120)
```

```
##
##  One Sample t-test
##
## data:  bloodpressure
## t = -82.306, df = 1331, p-value = 1
## alternative hypothesis: true mean is greater than 120
## 95 percent confidence interval:
##  93.67301      Inf
## sample estimates:
## mean of x
##  94.18919
```

p-value: At significance level 0.05, p-value is greater than alpha (0.05), which means we can't reject null hypothesis and conclude that the mean blood pressure of the patient is less than or equal to 120.

CI Interval: 120 lies in the confidence interval range, hence it is consistent will the p-value analysis and we can conclude that mean blood pressure is less than or equal to 120

Conclusion: The above analysis suggests that having high blood pressure doesn't indicate high claim rate.

2. Test the hypothesis that the mean bmi is greater than 24.9 (Healthy BMI)

Hypothesis:

H0: mean bmi is less than or equal to 24.9

Ha: mean bmi is greater than 24.9

Test Statistics:

```r
t.test(~ bmi, data=insurance, alternative="greater", mu=24.9)
```

```
##
##  One Sample t-test
##
## data:  bmi
## t = 34.346, df = 1331, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 24.9
## 95 percent confidence interval:
##  30.38237      Inf
## sample estimates:
## mean of x
##  30.65833
```

p-value: At significance level 0.05, p-value is less than alpha (0.05), which means that we can reject the null hypothesis and have enough evidence to conclude that the mean bmi is greater than 24.9.

CI Interval: 24.9 doesn't lie in the confidence interval range, hence we can reject the null hypothesis and have enough evidence to conclude that the mean bmi is greater than 24.9.

Conclusion: The above analysis suggests that having high bmi might indicates high claim rate.

3. Test the hypothesis that the mean age is greater than 35

Hypothesis:

H0: mean age is less than or equal to 35

Ha: mean age is greater than 35

Test Statistics:

```
t.test(~ age, data=insurance, alternative="greater", mu=35)
```

```
##
##  One Sample t-test
##
## data:  age
## t = 10.136, df = 1331, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 35
## 95 percent confidence interval:
##  37.58515      Inf
## sample estimates:
## mean of x
##  38.08634
```

p-value: At significance level 0.05, p-value is less than alpha (0.05), which means that we can reject the null hypothesis and have enough evidence to conclude that the mean age is greater than 35.

CI Interval: 35 doesn't lie in the confidence interval range, hence we can reject the null hypothesis and have enough evidence to conclude that the mean bmi is greater than 35.

Conclusion: The above analysis suggests that having age greater than 35 might indicates high claim rate.

4. Test the hypothesis that difference between the means of claims based on smoker is not equal to zero

Hypothesis:

H0: difference in means is equal to zero

Ha: difference in means is not equal to zero

Test Statistics:

```
t.test(claim ~ smoker, data=insurance) # Unpooled
```

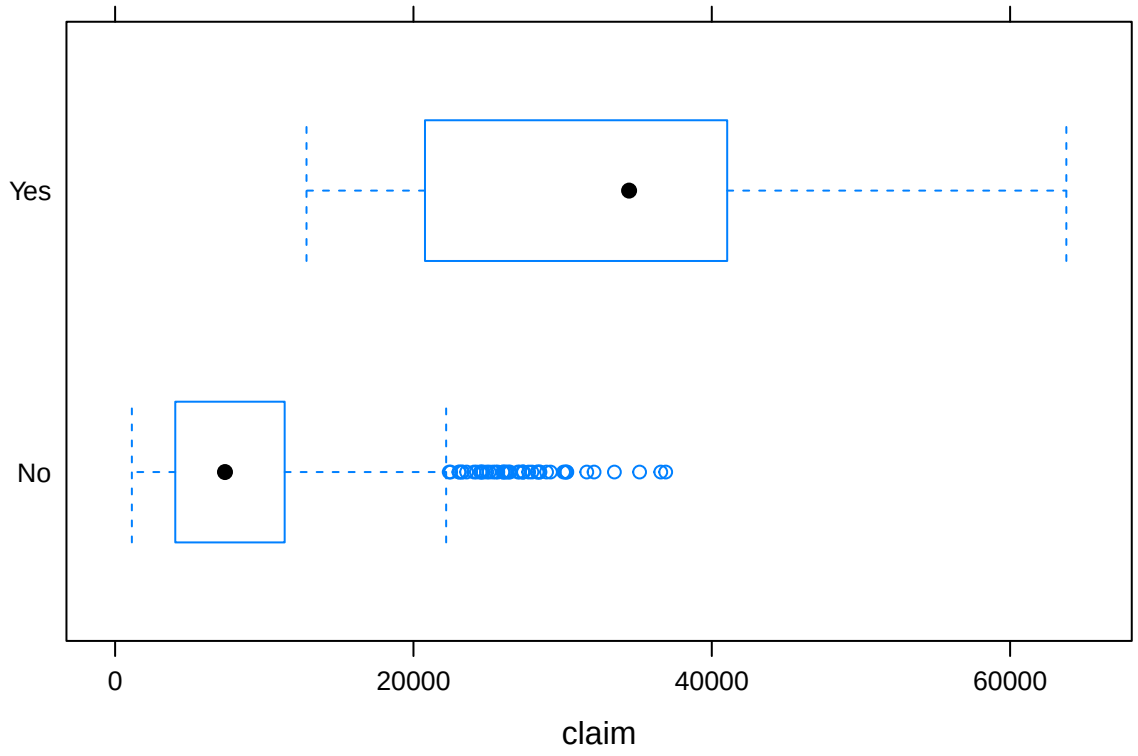
```
##
##  Welch Two Sample t-test
##
## data:  claim by smoker
## t = -32.691, df = 311.96, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -24993.25 -22155.49
## sample estimates:
##  mean in group No mean in group Yes
##          8475.865         32050.232
```

```
t.test(claim ~ smoker, var.equal=TRUE, data=insurance)   # Pooled
```

```
##
##  Two Sample t-test
##
## data:  claim by smoker
## t = -46.552, df = 1330, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -24567.81 -22580.93
## sample estimates:
##  mean in group No mean in group Yes
```

```
##          8475.865         32050.232
```

```
bwplot(smoker ~ claim, data=insurance)
```



p-value: At significance level 0.05, p-value is less than alpha (0.05), which means that we can reject the null hypothesis and have enough evidence to conclude that the difference in means is not equal to zero.

CI Interval: 35 doesn't lie in the confidence interval range, hence we can reject the null hypothesis and have enough evidence to conclude that the difference in means is not equal to zero.

Conclusion: The above analysis suggests that there is difference in mean claims of smokers and non-smokers i.e people who smoke tend to claim more than non-smokers.

5. Test the hypothesis that difference between the means of claims based on gender is not equal to zero

Hypothesis:

H0: difference in means is equal to zero

Ha: difference in means is not equal to zero

Test Statistics:

```
t.test(claim ~ gender, data=insurance) # Unpooled
```

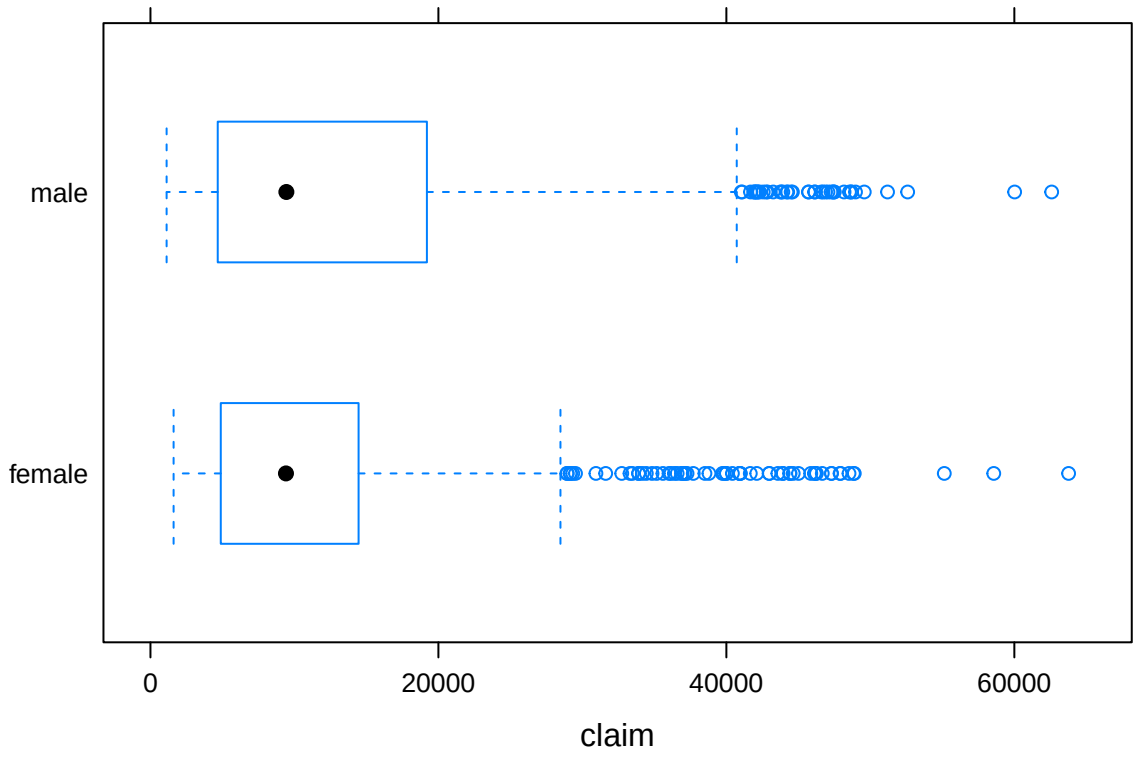
```
##
##  Welch Two Sample t-test
##
## data:  claim by gender
## t = -2.2694, df = 1304.4, p-value = 0.02341
## alternative hypothesis: true difference in means between group female and group male is not equal to
```

```
## 95 percent confidence interval:
##  -2800.9741  -203.6502
## sample estimates:
## mean in group female   mean in group male
##             12569.58             14071.89
```

```
t.test(claim ~ gender, var.equal=TRUE, data=insurance)   # Pooled
```

```
##
##  Two Sample t-test
##
## data:  claim by gender
## t = -2.2673, df = 1330, p-value = 0.02353
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -2802.1382  -202.4861
## sample estimates:
## mean in group female   mean in group male
##             12569.58             14071.89
```

```
bwplot(gender ~ claim, data=insurance)
```



p-value: At significance level 0.05, p-value is less than alpha (0.05), which means that we can reject the null hypothesis and have enough evidence to conclude that the difference in means is not equal to zero.

CI Interval: 35 doesn't lie in the confidence interval range, hence we can reject the null hypothesis and have enough evidence to conclude that the difference in means is not equal to zero.

Conclusion: The above analysis suggests that there is difference in mean claims based on gender is different

i.e one of the gender has more claim than other

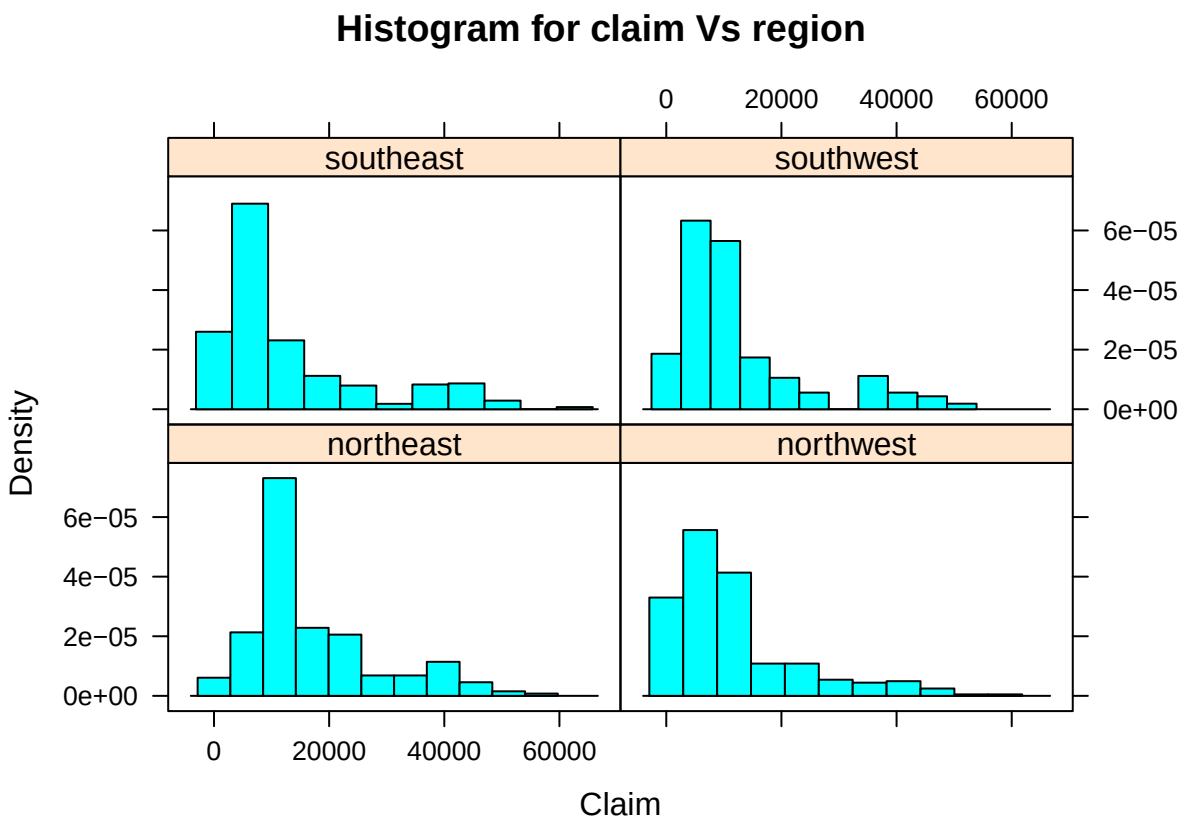
3.3 analysis of variables

```
# Claim values for region

favstats (~claim |region, data=insurance)
```

```
##      region     min       Q1    median       Q3      max     mean       sd   n
## 1 northeast 1694.80 9521.135 13129.600 21715.64 58571.07 16889.04 11578.10 231
## 2 northwest 1146.80 4134.080  8310.840 14256.19 60021.40 11794.22 11036.72 345
## 3 southeast 1121.87 4430.440  7403.290 17074.79 63770.43 13085.50 13179.73 442
## 4 southwest 1261.44 5029.378  9123.185 13839.08 52590.83 12723.13 11578.52 314
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
```

```
histogram (~claim |region, data=insurance, xlab = "Claim",
     ylab = "Density",
     main = "Histogram for claim Vs region")
```

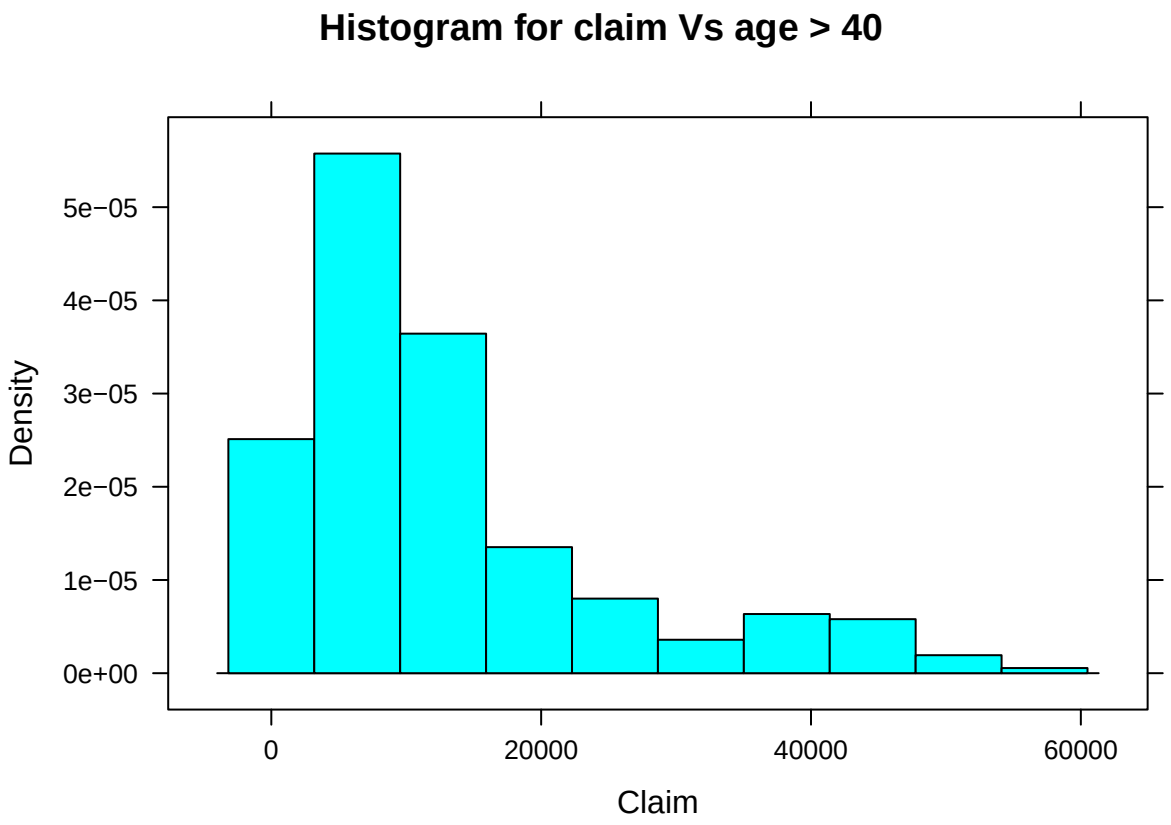



```
# Claim values - for age greater than 40

favstats(~(insurance%>% filter(age > 40))$claim)
```

```
##      min      Q1 median       Q3      max     mean       sd   n missing
##  1261.86 4661.29 9249.5 16085.13 58571.07 13007.49 11938.85 569       0
```

```
histogram(~(insurance %>% filter(age > 40))$claim, xlab = "Claim",
     ylab = "Density",
     main = "Histogram for claim Vs age > 40")
```



Based on the above histograms, the variables age and region very slightly effects claim. Similarly other columns also show very little impact on the claim amount. On the other hand the smoker columns shows a moderate positive linear relationship with claim.

4 Conclusion

Based on the above analysis, we can conclude that if a patient is a smoker then his claim amount is expected to be higher than a non smoker with same demographic details. Although other details like blood pressure, bmi etc show a little impact on the insurance, it's overshadowed by the effect shown by smoking. Hence the patients with smoking habit are more vulnerable and the insurance companies can expect more number of and higher claim amount from these patients.