Case Study On U.S Health Insurance

PROVIDED DATA

A dataset of US health insurance with the attribute insurance charges(Individual medical costs billed by health insurance) a against the following attributes of the insured:

- age – Age of primary beneficiary
- sex – Insurance contractor gender
- bmi – body mass index
- number of children – Number of Children covered by Health insurance
- smoker – Smoker / Non - smoker
- region – The beneficiary's residential area in the US

PROBLEM STATEMENT

As the insured medical charges is determine by factors like age, sex, health of the insurer, predict which factors determine the increase in medical charges based on factors (Age , Sex, No of Children, Smoking habit, Region of residence, BMI as the normal BMI is between 18.5 and 24.9) given in the dataset provided.

DATA PRE-PROCESSING

- Obtaining inference from summary , "sex", "smoker" and "region" are strings while "age", "children", "bmi" and "charges" are numbers.
- None of the columns have any missing values . Its clear that there is no need of any data cleaning.

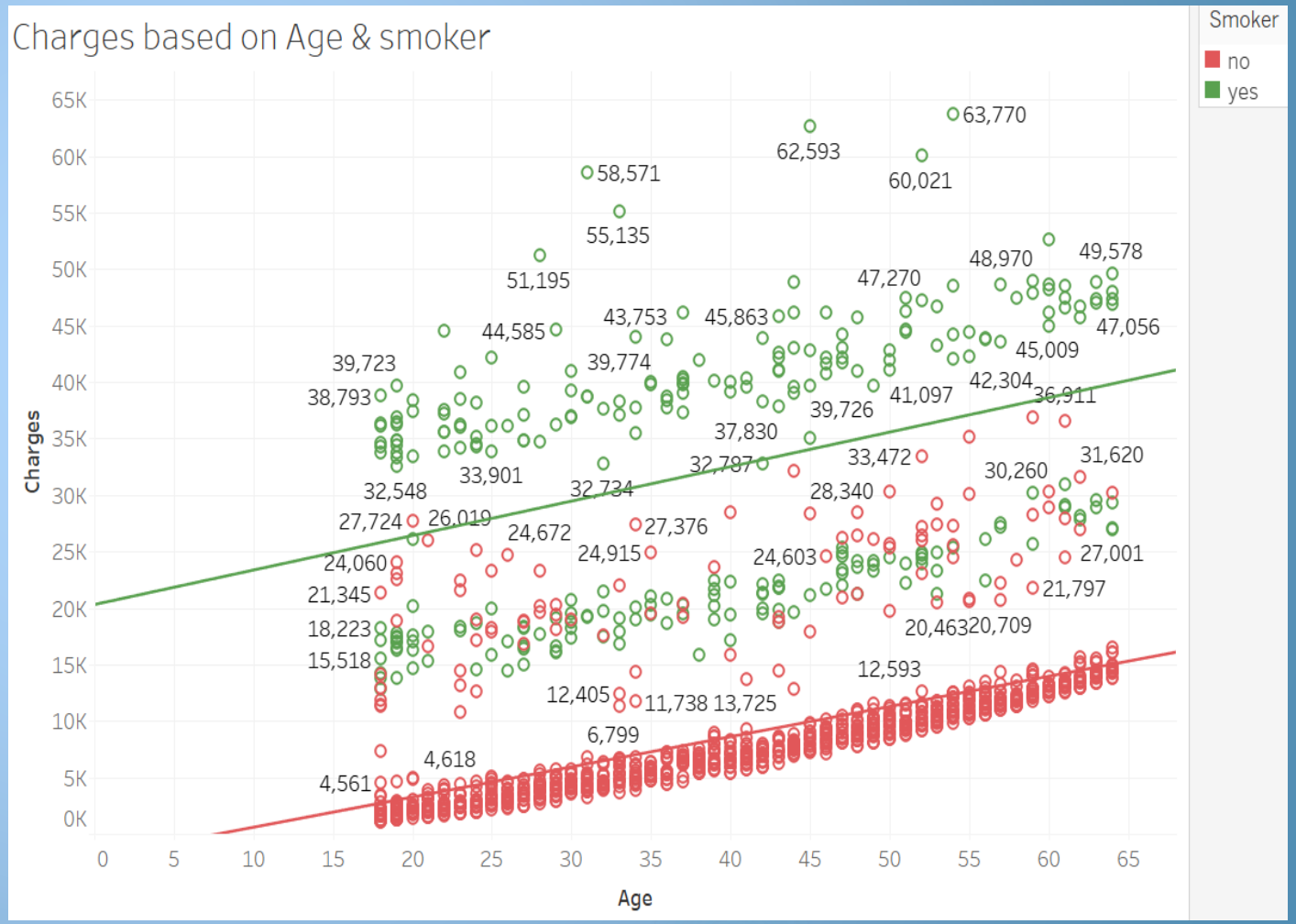
```
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

```
> summary(insurance)
      age            sex                 bmi           children     
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000  
 Mean   :39.21                      Mean   :30.66   Mean   :1.095  
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000  
 Max.   :64.00                      Max.   :53.13   Max.   :5.000  
    smoker             region      charges       
 Length:1338        Length:1338   Min.   : 1122  
 Class :character   Class :character   1st Qu.: 4740  
 Mode  :character   Mode  :character   Median : 9382  
                                       Mean   :13270  
                                       3rd Qu.:16640  
                                       Max.   :63770
```

EXPLORATORY DATA ANALYSIS

Charges based on Age:

- People who are younger are less prone to getting sick and thus company has to pay them less for their medical bills.
- Otherwise every age group in US has equivalent population density.
- So we can see as age increases the medical charges also increases.
- It is also clear that older persons who are smokers receiving high medical bills.



EXPLORATORY DATA ANALYSIS

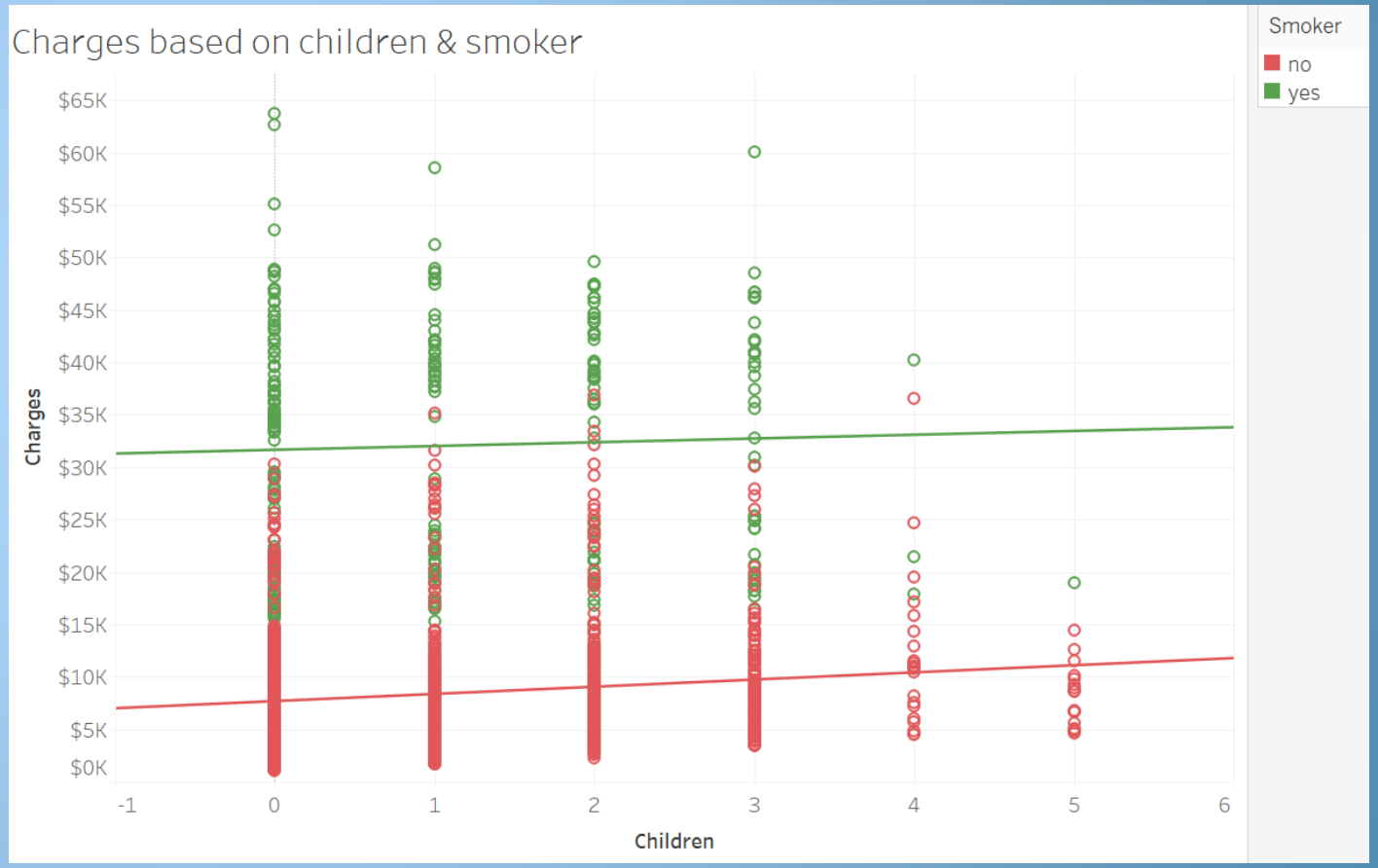
Charges based on Region:

- We can see that southeastern part of U.S is leading in charge but majority of all customers from all parts of US are charged between 0-20k only.
- Also we can observe smokers are insured high amount in all parts of U.S

EXPLORATORY DATA ANALYSIS

Charges based on children:

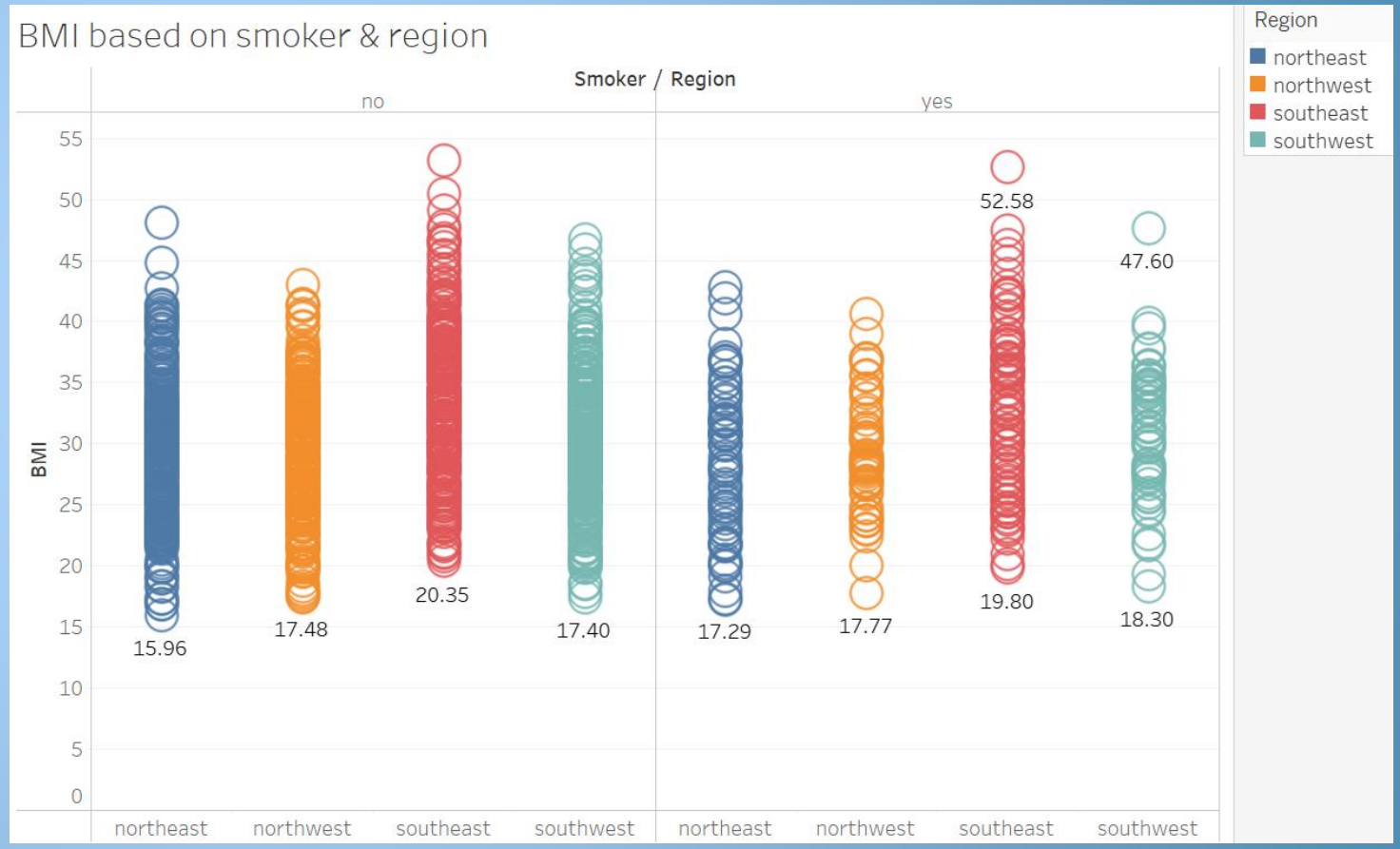
- There is no that strong trend among the variables.
- We can see the charges of the customer having 4 to 5 children are low comparatively than the charges received by the customers having 3 to 0 children.



EXPLORATORY DATA ANALYSIS

BMI based on Smoker and Region:

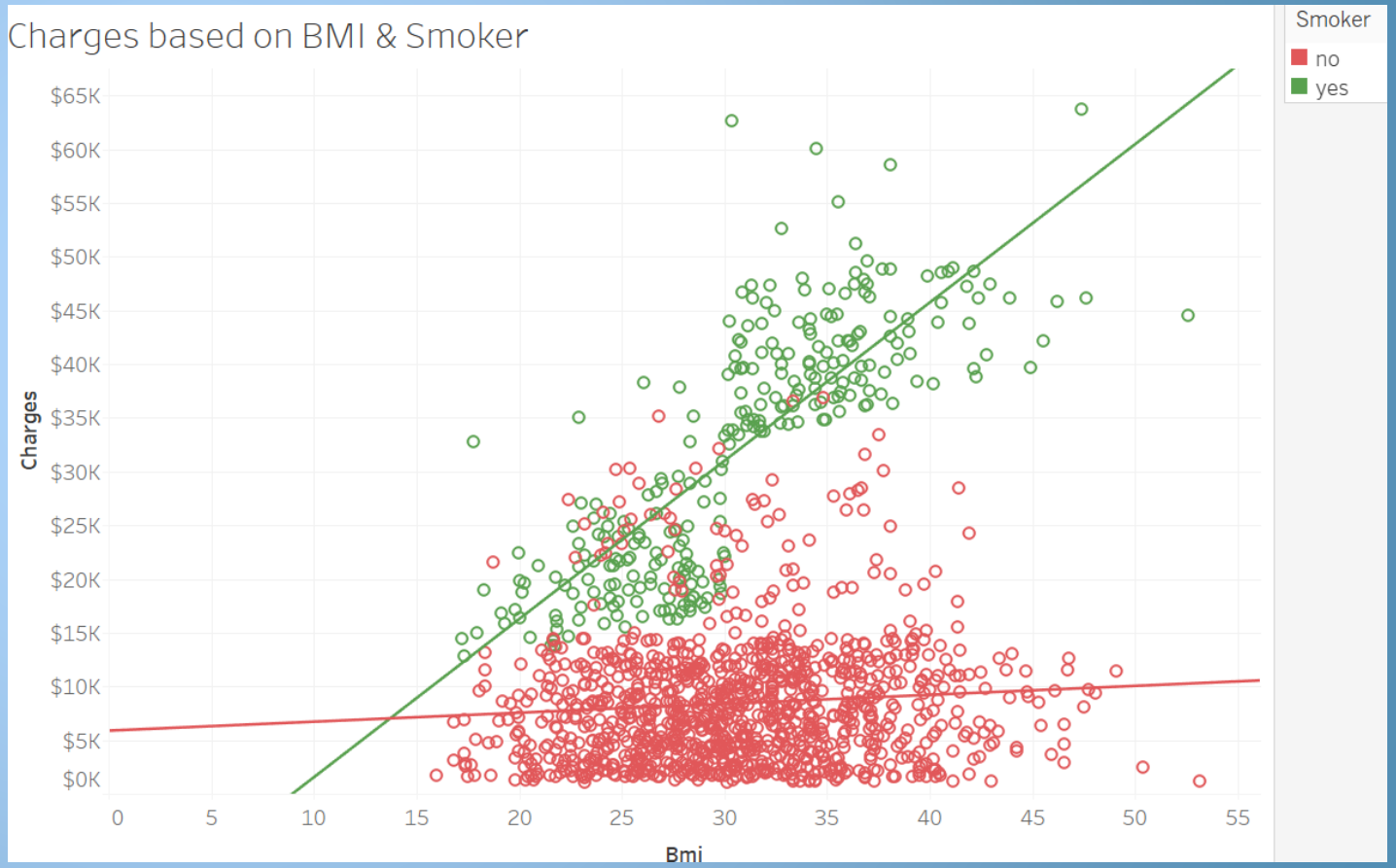
- We can see the BMI of the person who are a smoker is high when compare with others.
- We can also infer that southeast part of U.S have higher no of customers who receive high amount of BMI in both cases



EXPLORATORY DATA ANALYSIS

Charges based on BMI:

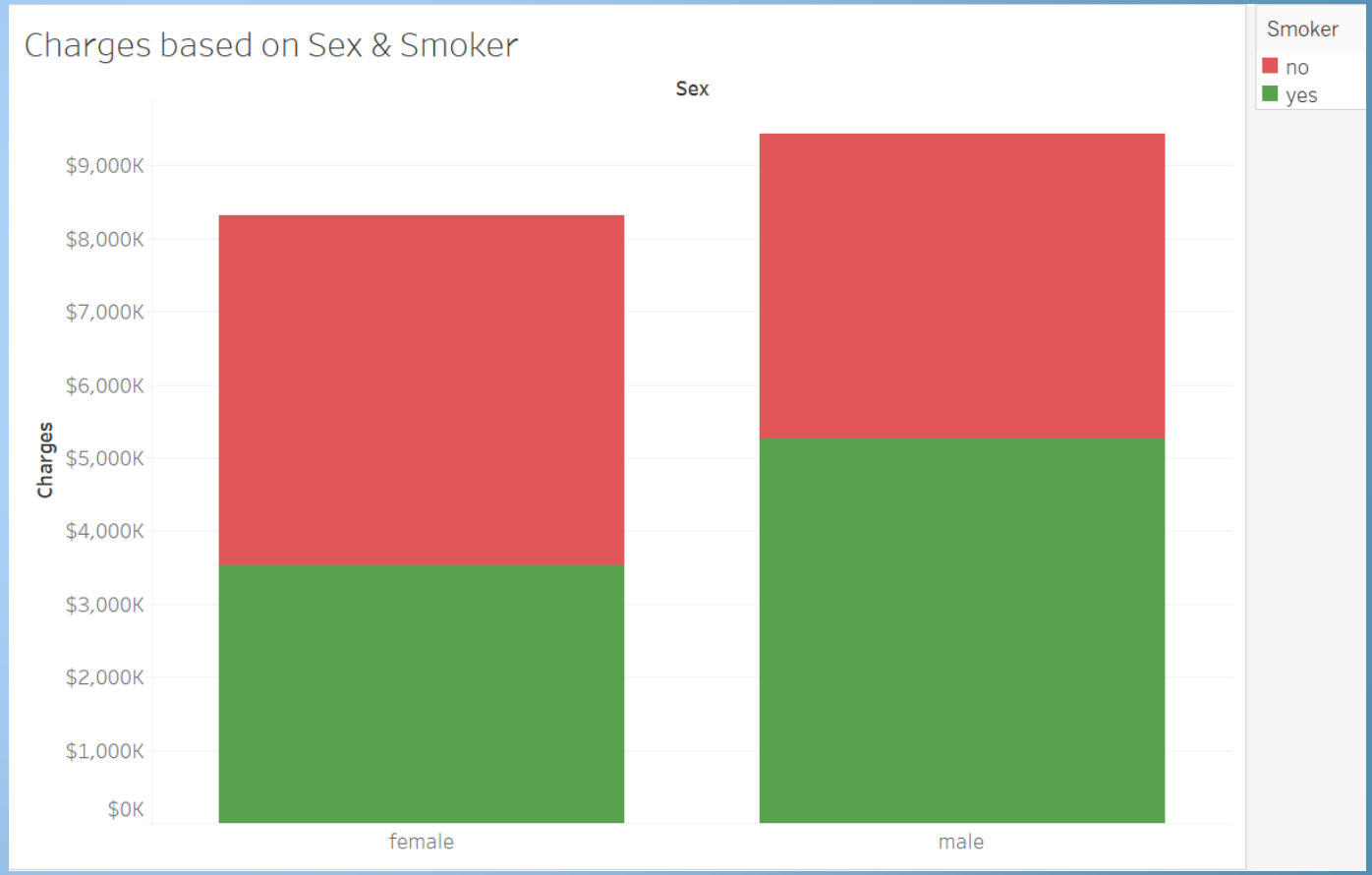
- It clearly shown that the customer who have high BMI received high medical charges.
- And person with high BMI who is a smoker received even higher medical charges. It is clearly shown using trend lines.



EXPLORATORY DATA ANALYSIS

Charges based on Sex:

Its clear as a whole male received high medical charges and especially the male who is a smoker received more than others

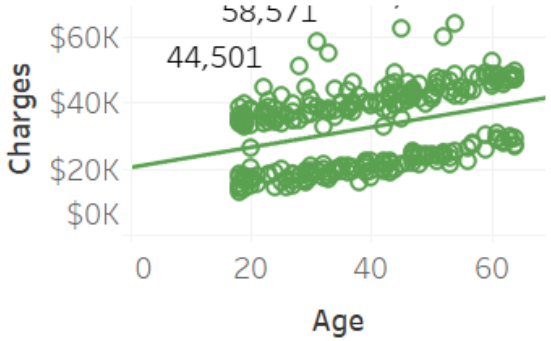


Inference for which factor infuence the charges

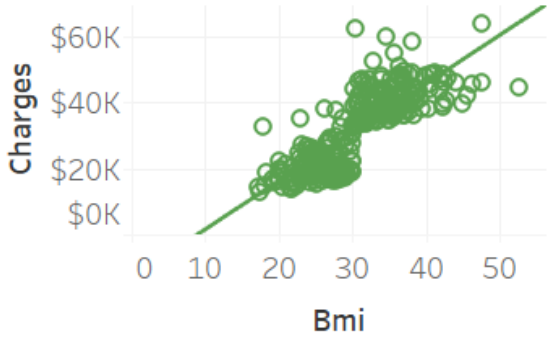
It is clear that the BMI of the person is higher than the normal level when he / she is a smoker.Thus obviously the person will be bad in health condition. Hence it is clearly shown from the charts that the medical charges of those persons are high.

Influence of smoking habit on Medical charges

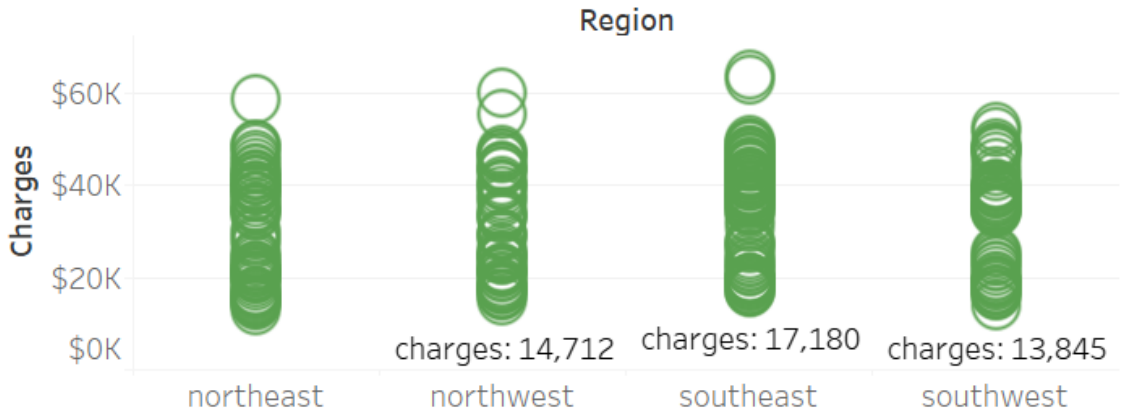
Charges based on Age & smoker



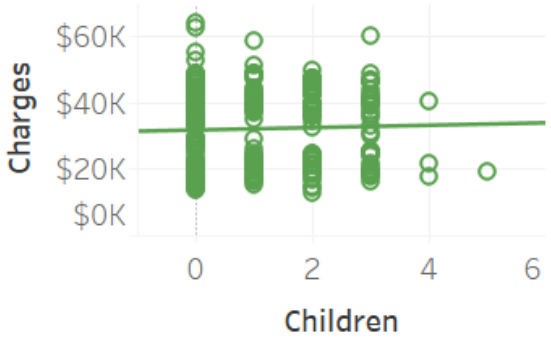
Charges based on BMI & Smoker



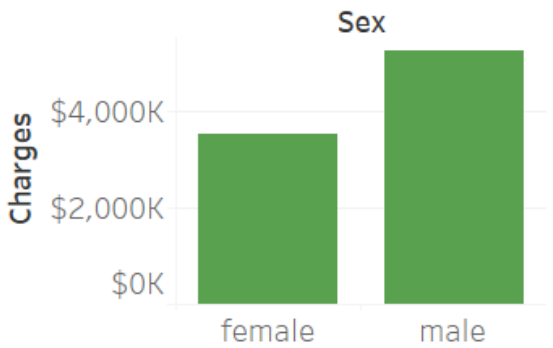
Charges based on region



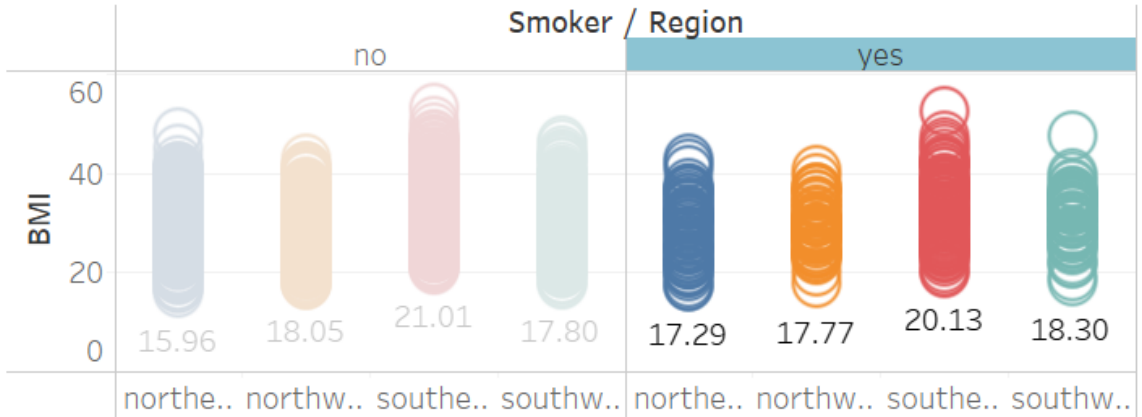
Charges based on children & smoker



Charges based on Sex & Smoker



BMI based on smoker & region

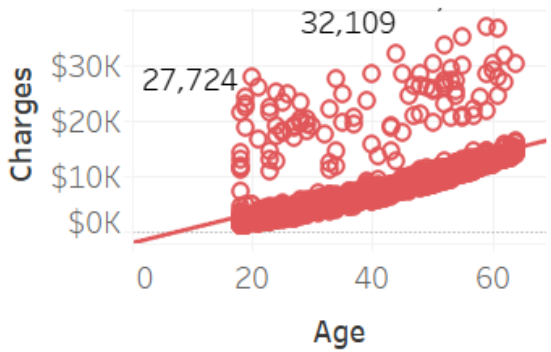


Inference for which factor infuence the charges

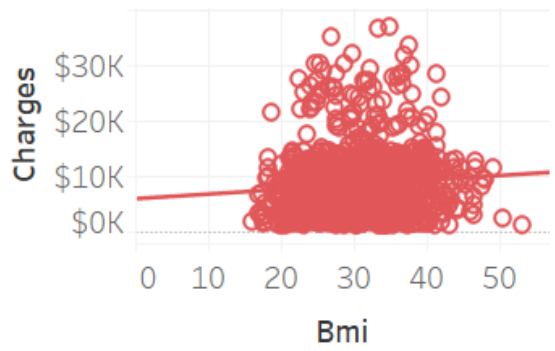
we can see that the BMI of some person who are not a smoker is also comparatively high. It may be due to some health issues like obesity or the person may be met with an accident.

Influence of smoking habit on Medical charges

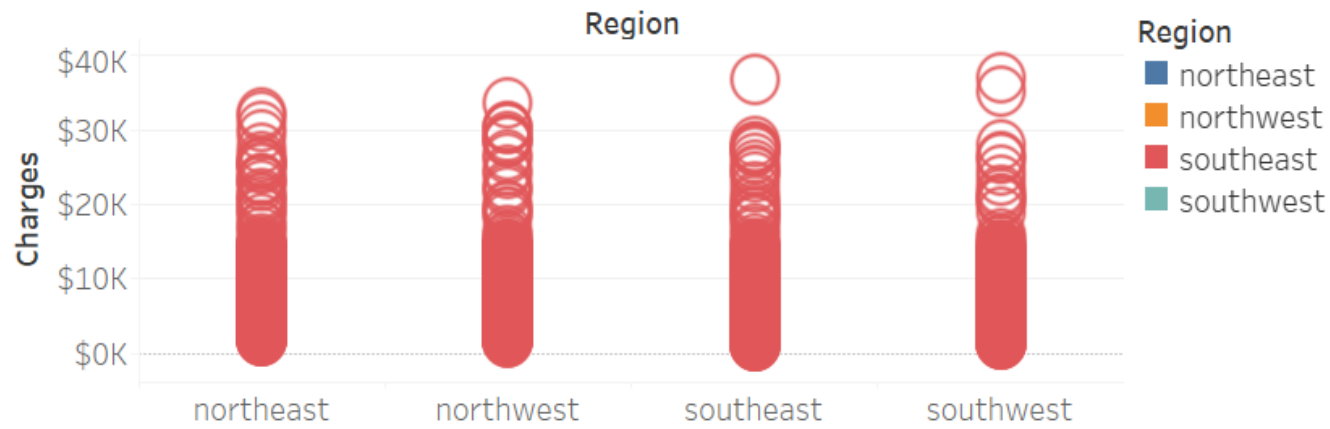
Charges based on Age & smoker



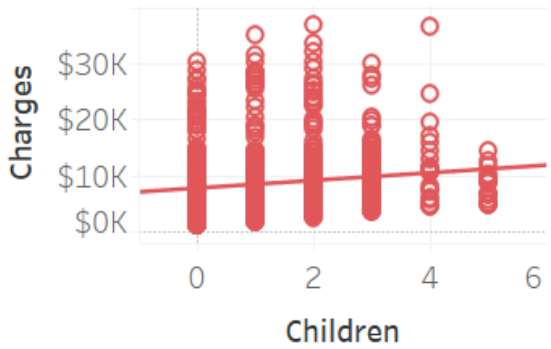
Charges based on BMI & Smoker



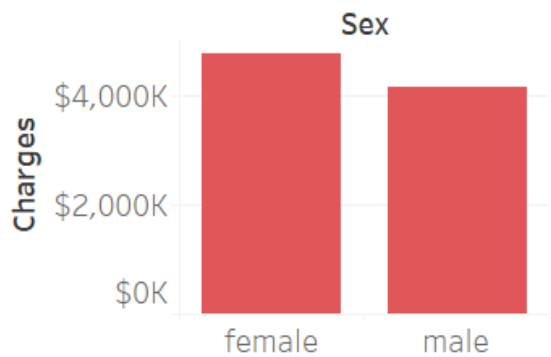
Charges based on region



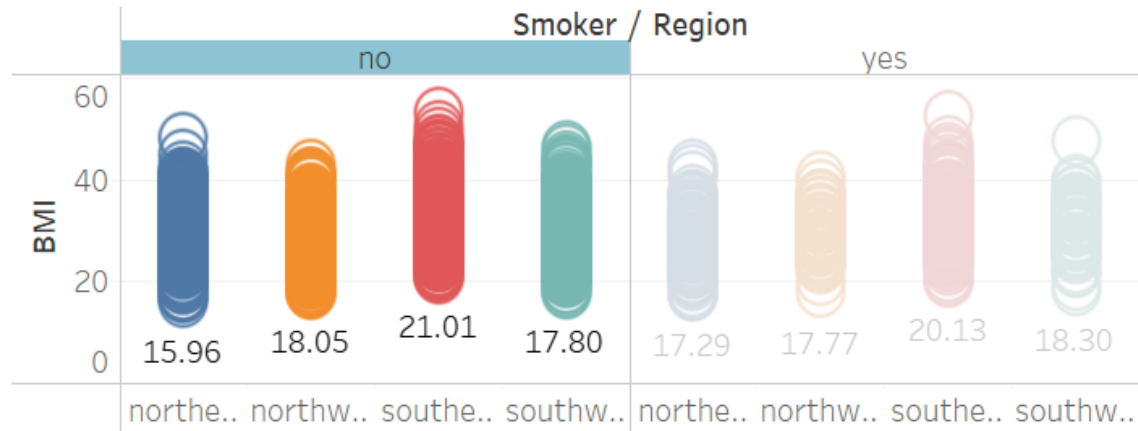
Charges based on children & smoker



Charges based on Sex & Smoker



BMI based on smoker & region

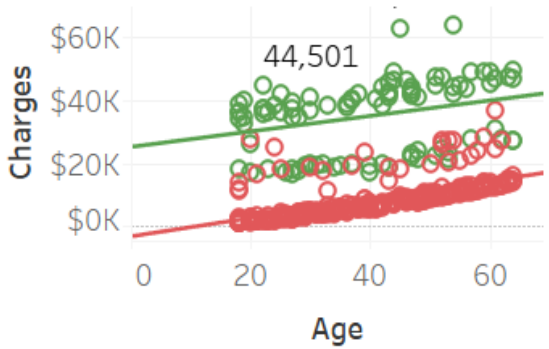


Inference for which factor infuence the charges

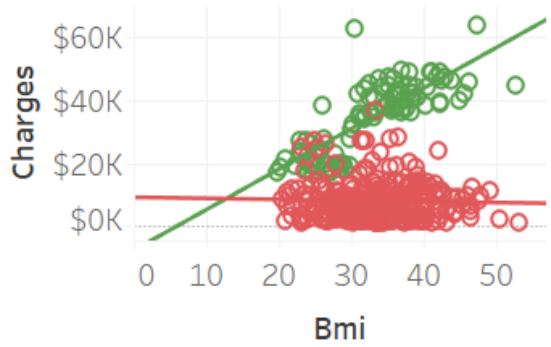
Also it is observed that over all the charges are high in the southeast region.

Influence of smoking habit on Medical charges

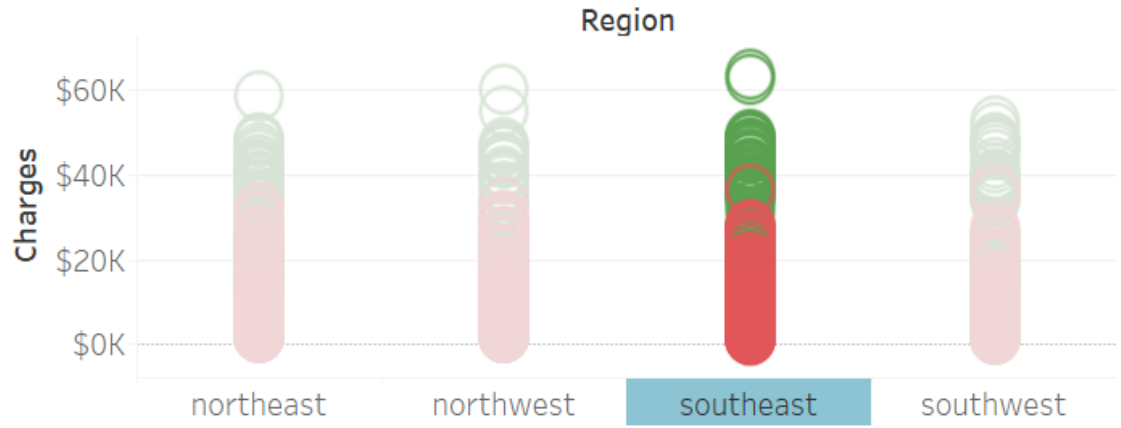
Charges based on Age & smoker



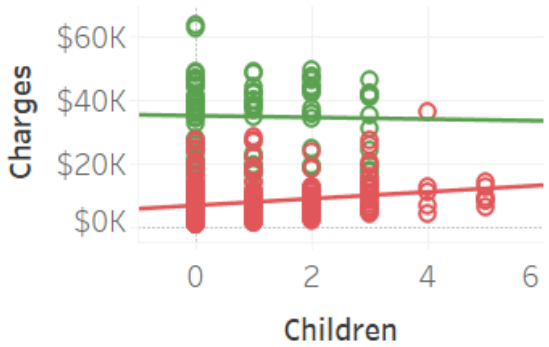
Charges based on BMI & Smoker



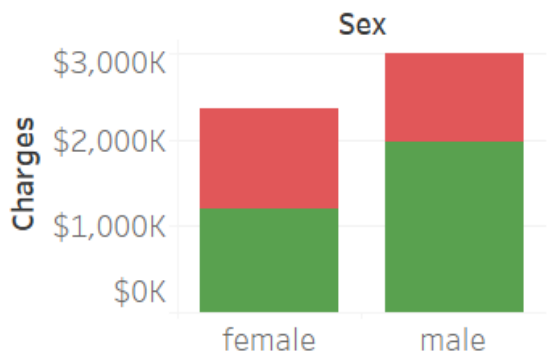
Charges based on region



Charges based on children & smoker



Charges based on Sex & Smoker



BMI based on smoker & region

EXPLORATORY DATA ANALYSIS

- Using above charts ,story with multiple charts and with inference in those story in different perspective , we can easily infer the data set clearly and can say how the charge is determined using other factors.
- It is clear that sex doesn't play a significant role in predicting the charges as the distribution is some how uniform.

LINEAR REGRESSION MODEL

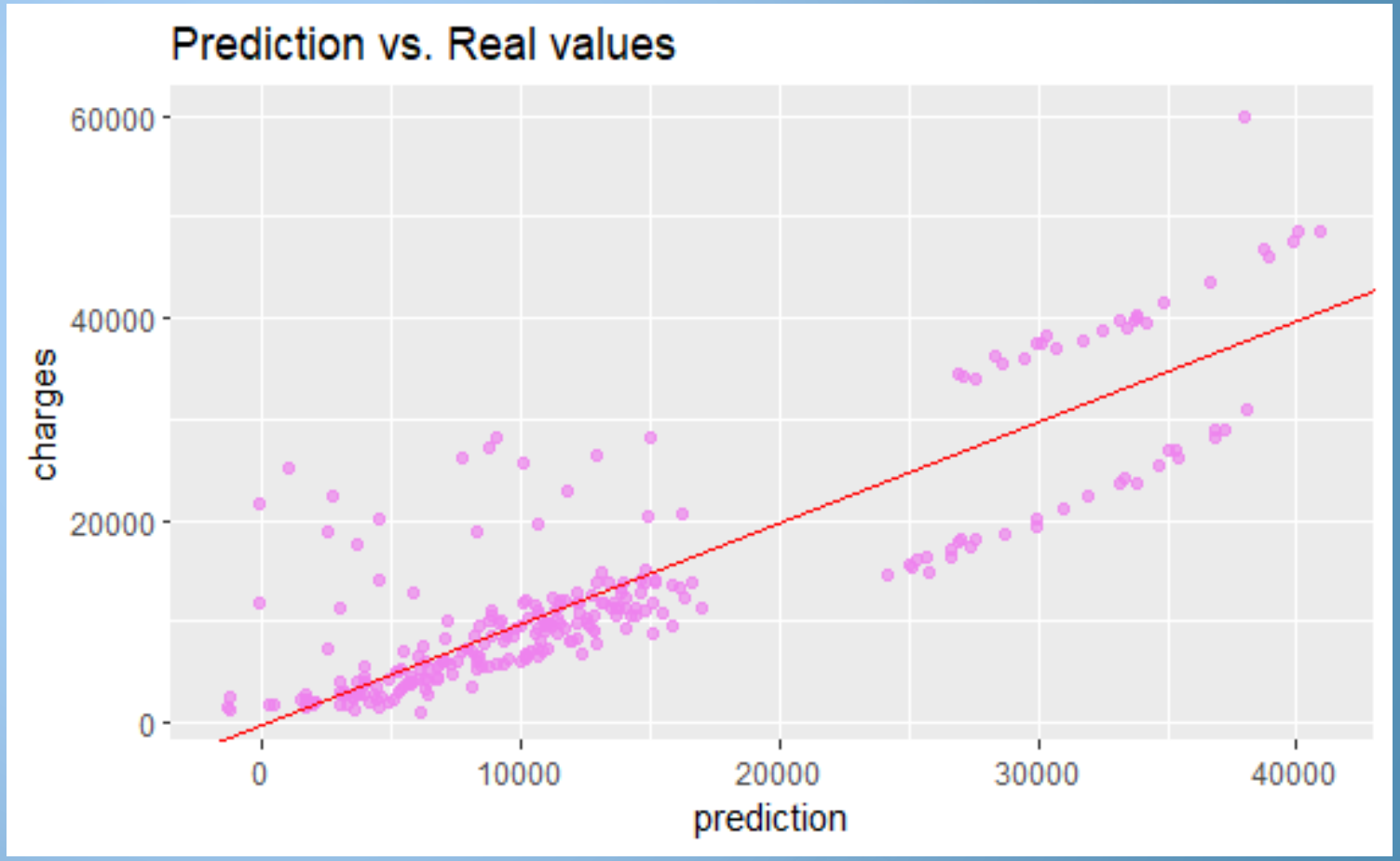
Using regression model, we constructed 5 models. Out of 5 models, we found the model 2 works significantly good by using the high adjusted R- squared value than the other models where,

Multiple R-squared: 0.7509
Adjusted R-squared: 0.7496

```
mod_1<-lm(charges ~ age + sex + bmi + children + smoker + region, data=insurance)
summary(mod_1)
mod_2<-lm(charges ~ age + bmi + children + smoker + region, data=insurance)
summary(mod_2)
mod_3<-lm(charges ~ age + bmi + children + smoker ,data=insurance)
summary(mod_3)
mod_4<-lm(charges ~ bmi + children + smoker ,data=insurance)
summary(mod_4)
mod_5<-lm(charges ~ bmi + age + smoker  ,data=insurance)
summary(mod_5)
```

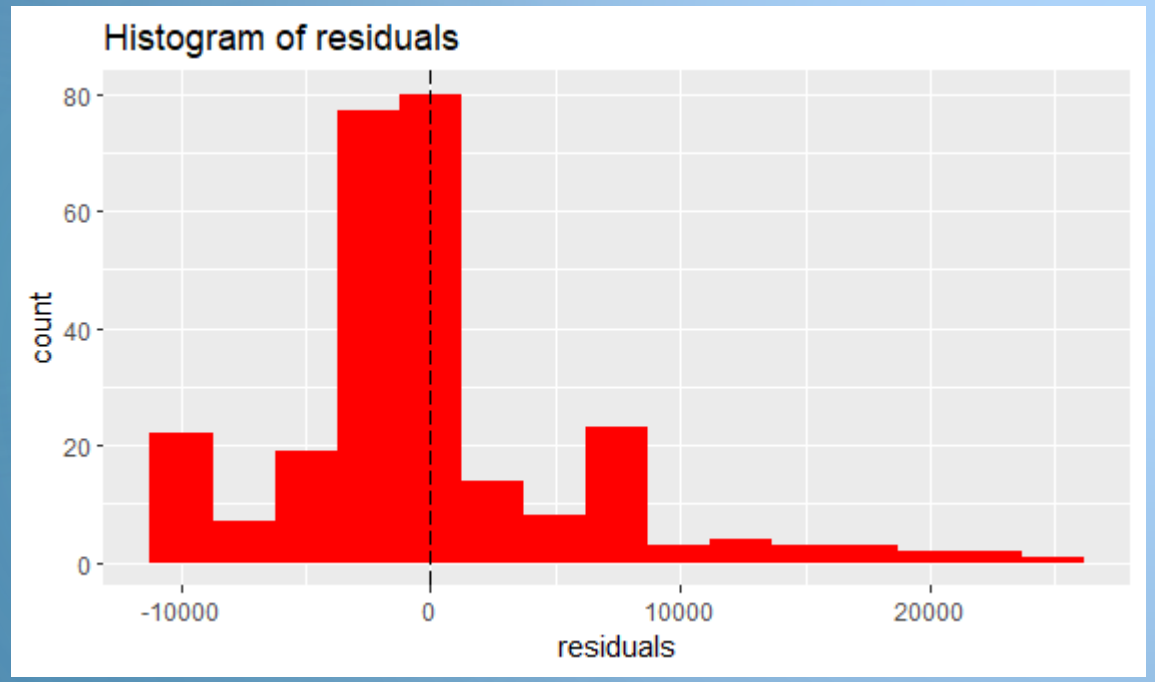
MODEL PERFORMANCE

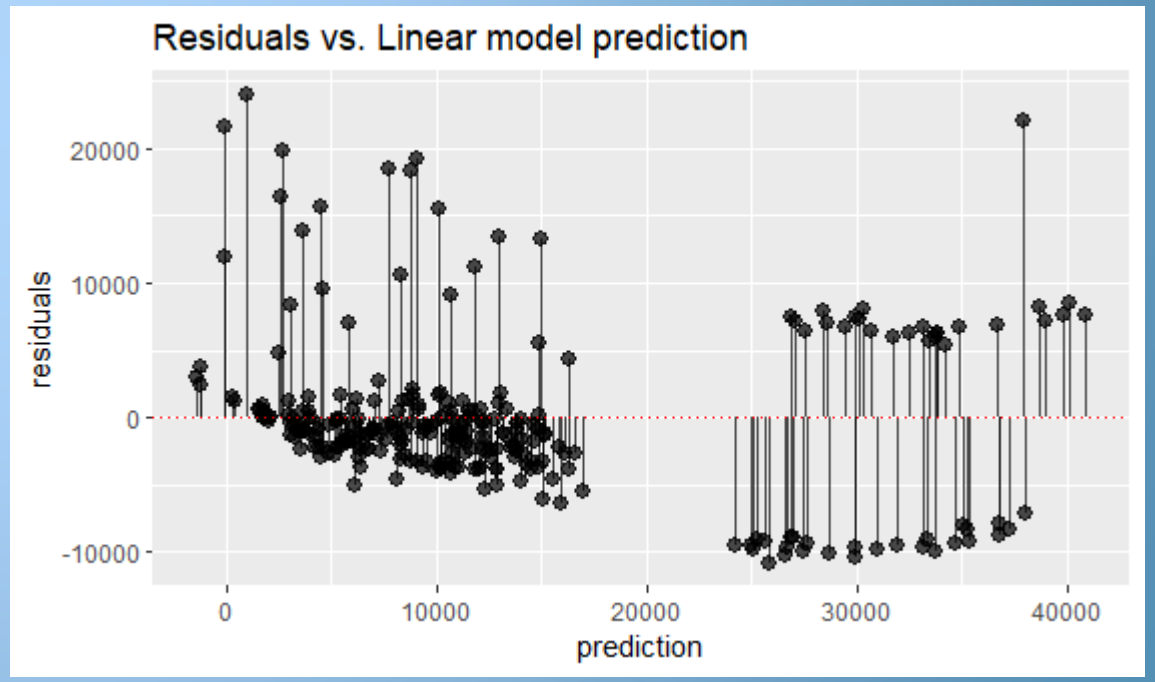
As the trend line fit to the graph of predicted values of tested model VS the real values of the actual data. Hence the model performs well.



MODEL PERFORMANCE

- Residuals are normally distributed
- They are independent.





CONCLUSION

By above observations from the charts and the model which we got using linear regression clearly predict that the factors age, children , BMI, Smoking habit and Region plays a vital role in the determination of the medical charges to be insured to the customer. As sex the doesn't show that much connection with charges determination. In the first place Smoking habit and BMI contribute huge part then the others.

THANK YOU