Medical Charge Prediction

**Done by:**

**Indumathi S**

**Setting the research goal**

1. **Domain Specification**

A health insurance company can only make money if it collects

more than it spends on the medical care of its beneficiaries. On the other hand, even though some conditions are more prevalent for certain segments of the population, medical costs are difficult to predict since most money comes from rare conditions of the patients.

1. **Objective of the Study**

To explore all factors of medical charges

To predict the medical cost of the patient using various factors.

1. **Business Problem**

Medical costs are difficult to predict since most money comes

from rare conditions of the patients. Based on some factors of the policyholders, we need to accurately predict insurance costs based on the policyholder. We will also determine what the most important variable influencing insurance costs is.

1. **Business Questions**

* To observe for whom charges is more: male or female
* For whom charges is high: smokers or non-smokers
* To observe for which region of people BMI higher
* To observe for which region the charges is high
* To observe the BMI of smokers and non-smokers
* To observe the region-wise smokers and non-smokers charges
* To find which variable is highly correlated with charges using correlation matrix.

**Retrieving data**

1. **Dataset Source**

The dataset has been collated from kaggle. The link has been shared.

[***https://www.kaggle.com/mirichoi0218/insurance***](https://www.kaggle.com/mirichoi0218/insurance)

1. **Dataset Specification**

The dataset has 7 variables with 1338 observations. The dataset includes information about the insurance policy holder, their dependents, and their medical charges throughout a year.

**Age** - This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).

**Sex** - This is the policy holder's gender, either male or female.

**BMI** - This is the body mass index (BMI), which provides a sense of how over or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.

**Children** - This is an integer indicating the number of children / dependents covered by the insurance plan.

**Smoker** - This is yes or no depending on whether the insured regularly smokes tobacco.

**Region -** This is the policy holder's place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest.

**Charges​ -** Yearly medical charges billed by the medical insurance provider ($).

1. **Tools used**

* Power BI
* Python

**Data Preparation**

1. **Preprocessing and Transformations**

Since the dataset is cleaned well, we didn’t need more preprocessing work. But In the dataset, we had four categorical variables which are sex (male, female), smoker(yes, no), region (southeast, northeat, southwest, northwest), children(0,1,2,3,4,5). Since the children variable is the number of children covered by policy, that variable is numeric. Therefore, we need to change the datatype as a category for EDA.

1. **Tools and Techniques**

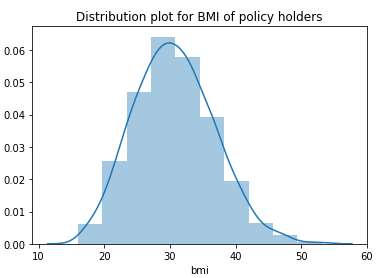
Python has been used for transform the data for our analysis

**Data Exploration**

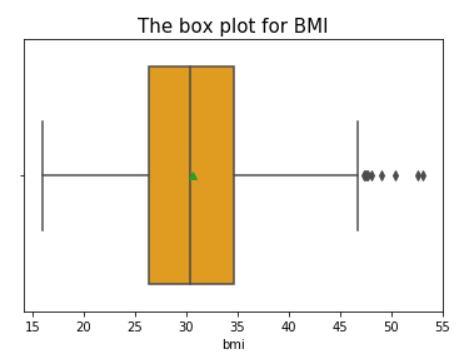
**1) EDA Specifications**

**Univariate Analysis:**

**BMI:**

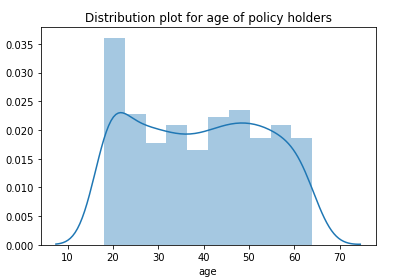
****

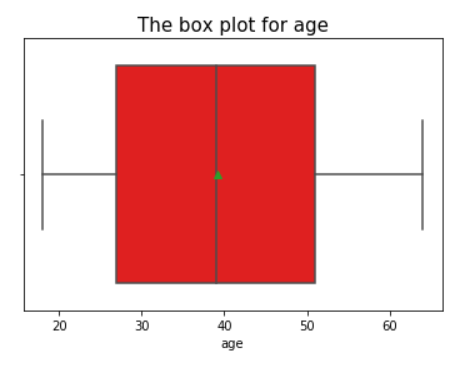
BMI of policyholders is normally distributed.



50% of BMI lies between 26 kg/m^2 to 35 kg/m^2 of the policy holders under consideration and the **maximum BMI of the data is 46 kg/m^2** and there exist few outliers in the data.

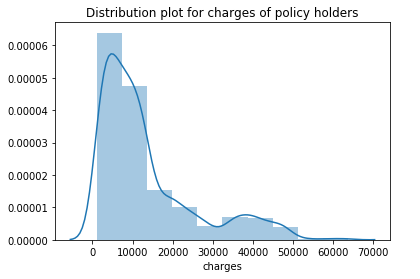
**Age:**

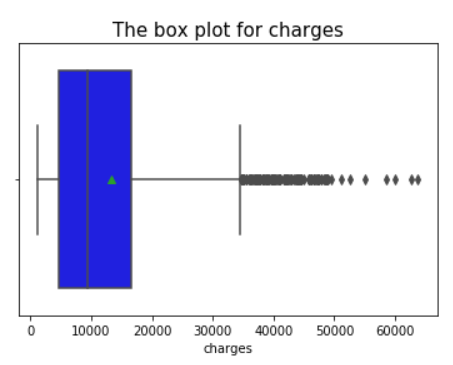
****



50% of the policyholders age between 26 to 52 for consideration.

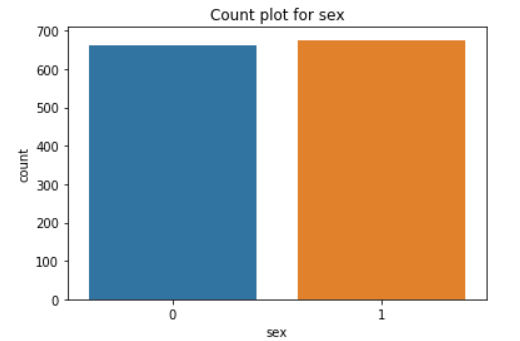
**Charges:**





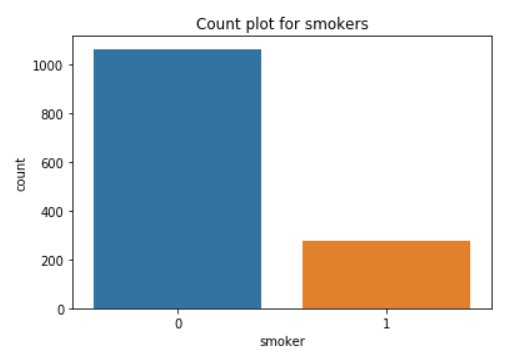
Charges distribution is unimodal and is right skewed. **Average charges billed by insurance companies are approximately $13000 and highest charge is approximately $63000.**There are a lot of outliers at the upper end.

**Sex:**

****

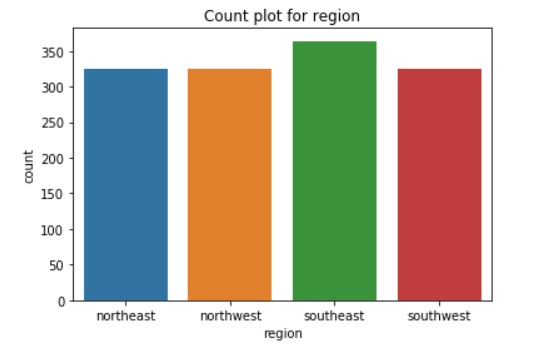
The number of policyholders for both male and female is the same.

**smokers:**

****

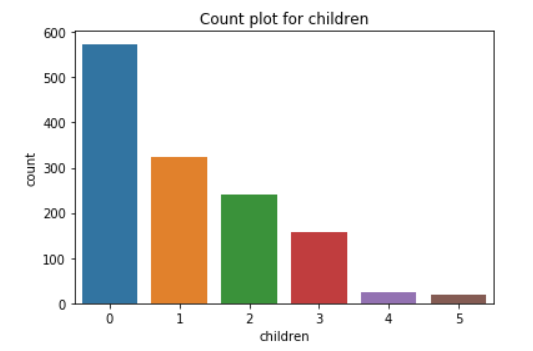
The non-smoker count is 1100 and for smokers 500.

**Region:**

****

The number of policyholders from **southeast is higher** than others

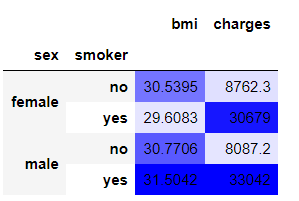
**Children:**

****

**Most of the policyholders have no childrens** that is approximately 580 out of 1338 policyholders have no childrens.

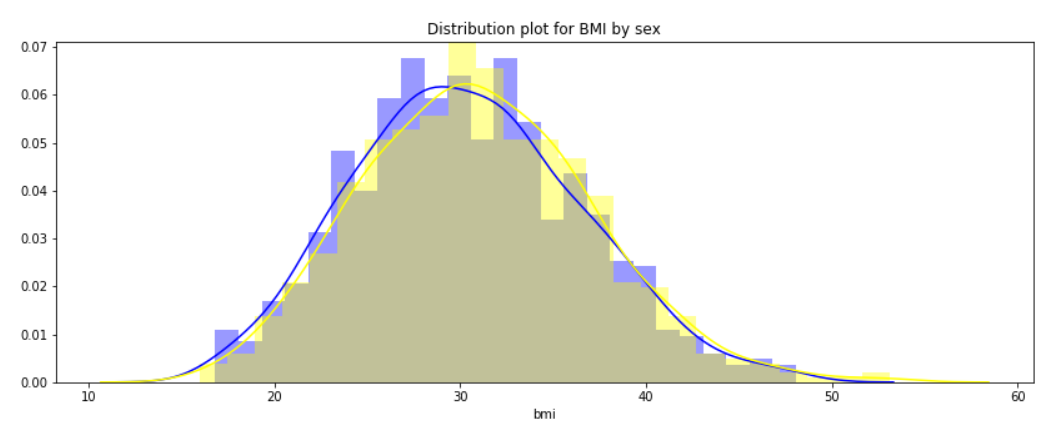
**Bivariate Analysis:**.

**Pivot Table for bmi and charges of policy holders by sex and smokers**

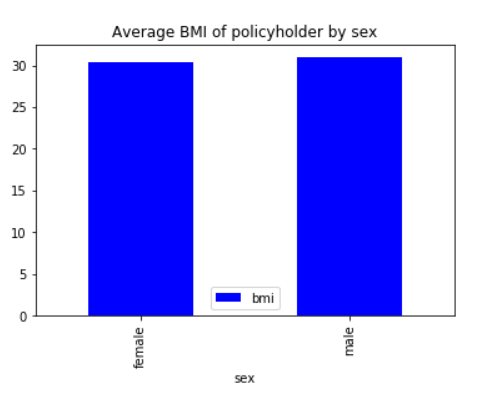
****

The **smokers of both sex have the very high charges** than non smokers. The **female smokers have less bmi** than others.

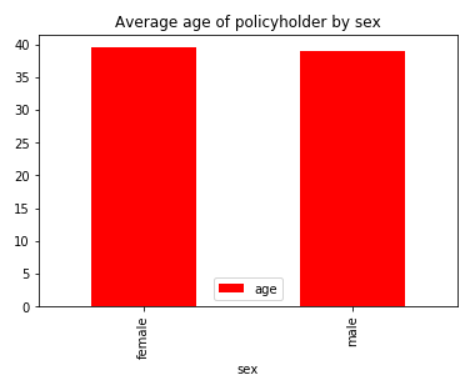
**Sex:**

****

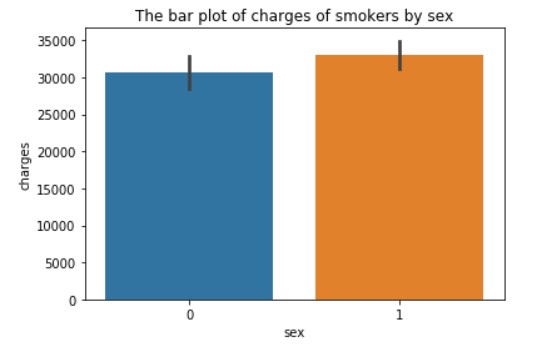
The distribution plot of bmi for both male and female have the normal distribution.

****

The average bmi of both female and male is the same which is 30 kg/m2.

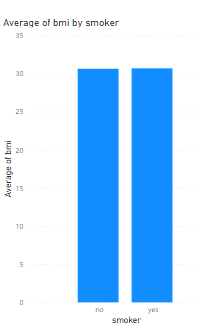


The **average age of male and female is between 37 to 40.**

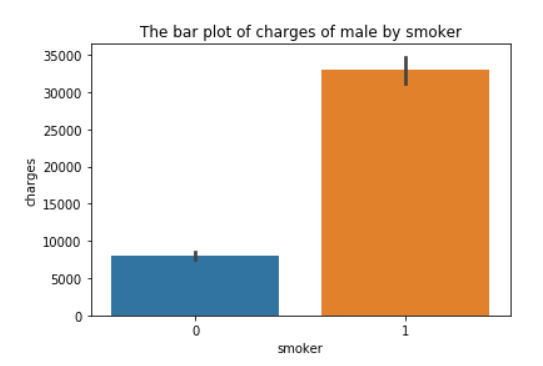
****

The average charges of female smoker is $ 30000 and male smokers is $ approximately $ 33000

**Smoker:**

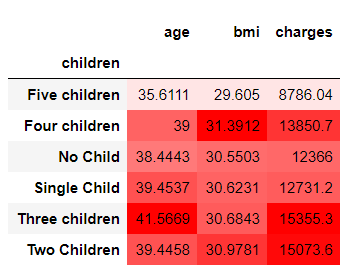


The average BMI of non-smokers and smoker is almost same

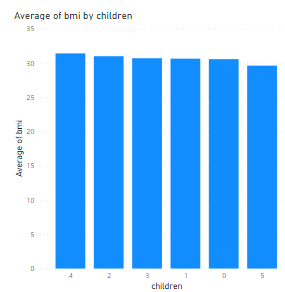
****

The average charges of male non-smoker is appr $ 8000 and male smokers is approximately $ 33000

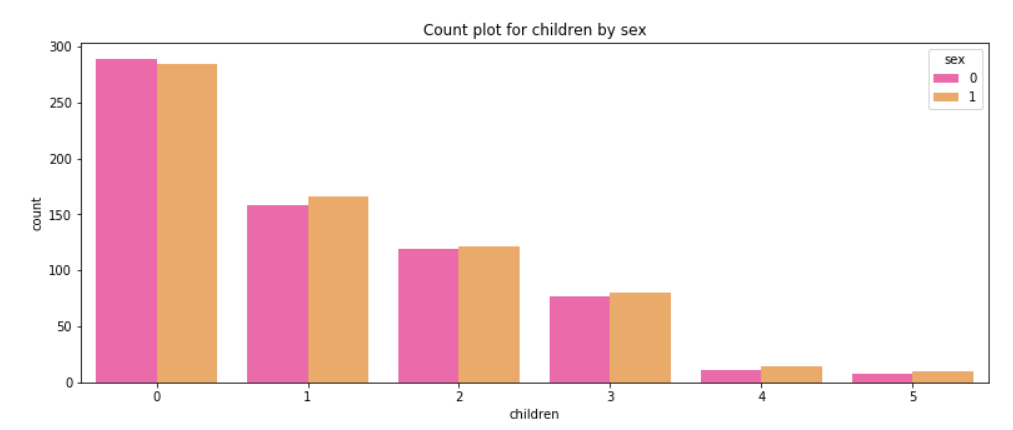
**Children:**

****

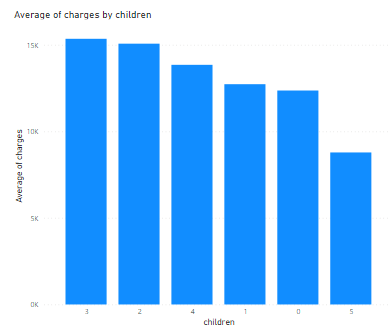
**The age, bmi and charges of policy**

****

The average BMI policyholders with of 4 covered under insurance children’ is little higher which is the maximum average BMI of the other category.



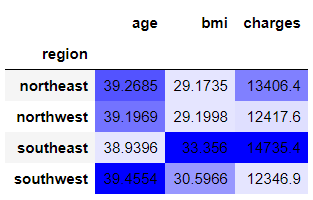
The number of female policyholders with no children under policy is slightly higher than male. But policyholders with some number of children under policy are higher male than female.

****

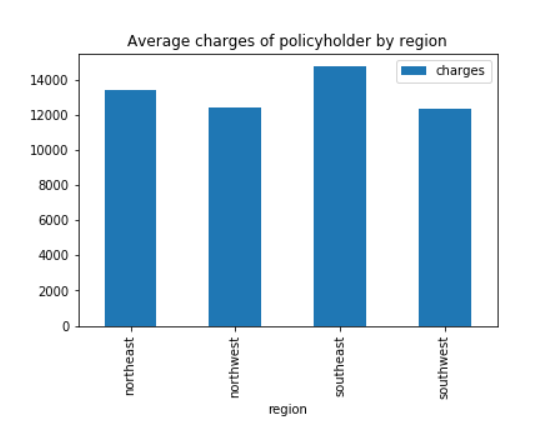
The average charges of policyholders with 3 children covered under policy is higher than others.

**Region:**

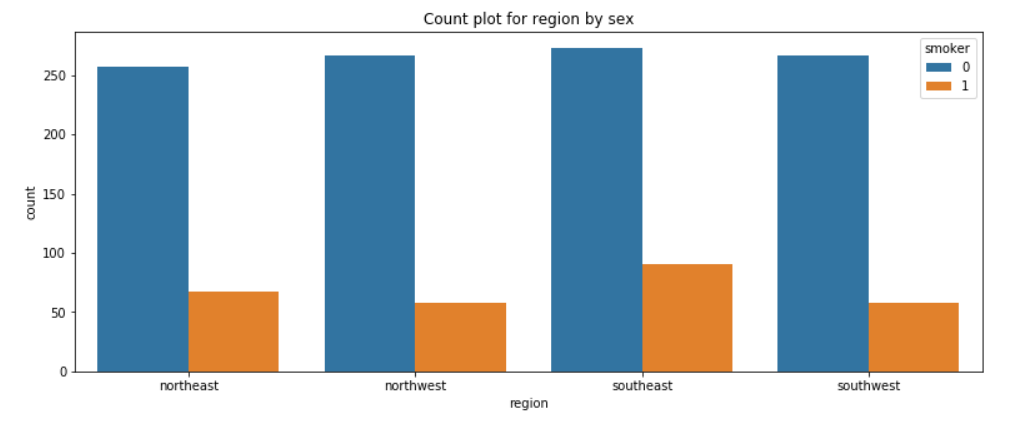
**Pivot Table:**

****

The policyholders from southeast have higher bmi and medical charges than others. The average age of southwest policy holders is high.

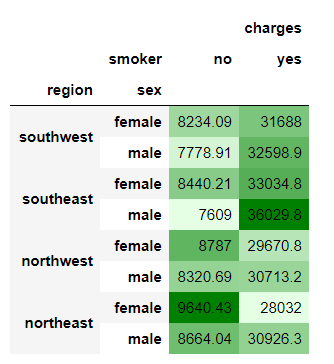
****

The average charges of the policyholders is **maximum for southeast which is above $14000** and the second most is northeast. Rest of the region is almost the same.

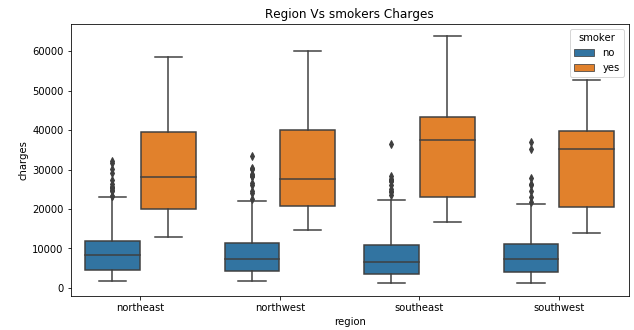
****

The number of smoker policy holders is maximum in the southeast than other regions. The number of non-smoker policy holders is almost the same in all regions.

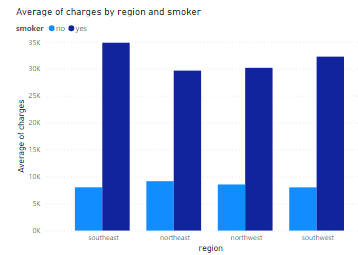
**Multivariate Analysis:**

****

**Box and whisker plot**

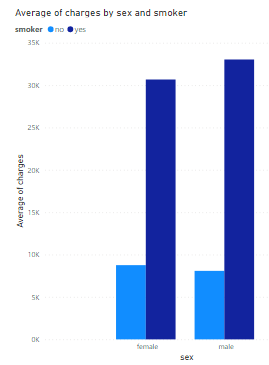


From this plot we can infer that almost in all regions non-smoker minimum, maximum, median and inter quartile range of charges is the same. There exist many outliers in the data of non-smoker. The smokers' charges are different for different regions. 50% of the charges lie between $20000 to $40000 for smokers and lie between $5000 to $10000 for non-smokers. Southeast and Southwest of smokers are negatively skewed Other regions are positively skewed.

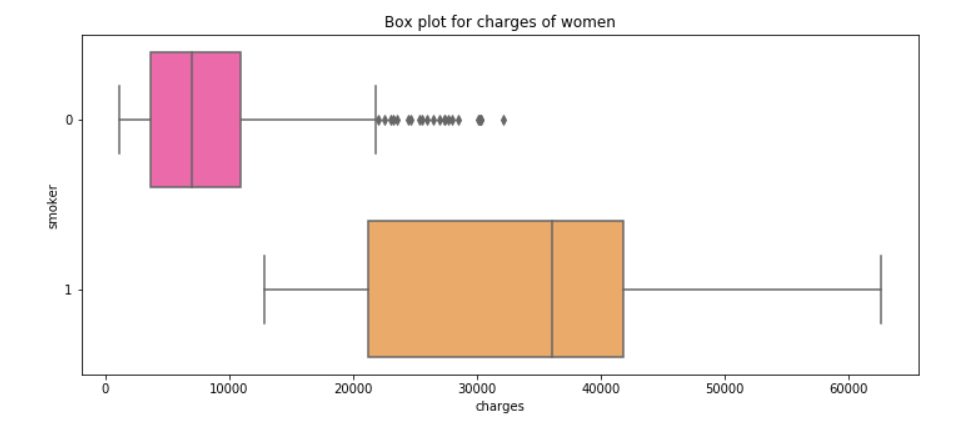


The average charges of smoker’s policy holders of southeast is higher than the other regions and minimum for northeast. Non-smoker’s average charges are almost the same for all regions.

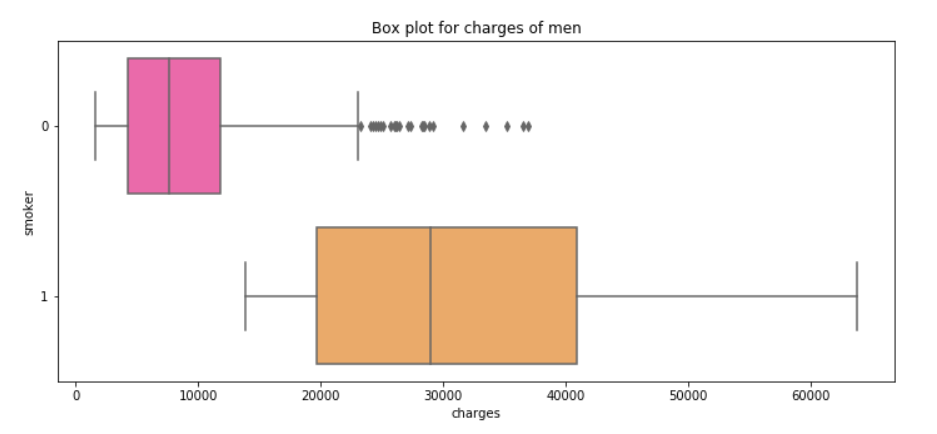
**Average of charges by sex and smoker:**



When we analyse by sex wise smoker category charges we infer that there is difference in charges between smokers and non-smokers for both female and male. Approximately $20000 more charge for smokers than non-smokers in females and male. There may be $1000 rupees less charge for males than females for non-smokers.

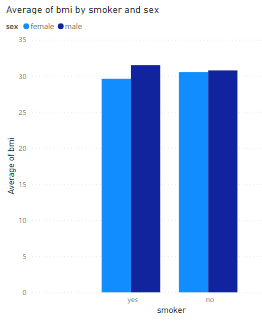


The 50% charges of women non-smoker lie between $5000 to $12000 and charges of women non-smoker are slightly negatively skewed but have more outliers. For women smokers 50% of charges lie between $22000 and $43000 and it is negatively distributed. The maximum charge for women smokers is approximately $63000 and minimum is $13000 which is more than the non-smoker interquartile range.



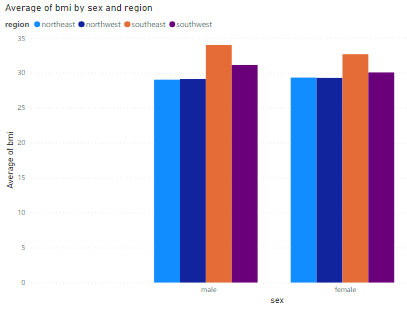
The 50% charges of men non-smoker lie between $5000 to $12000 and charges of men non-smoker are slightly negatively skewed but have more outliers. For men 50% charges lie between $19000 and $42000 for smokers and it is negatively distributed. The maximum charge for men smokers is approximately $6500 and minimum is $14000 which is more than the non-smoker interquartile range.

**Average of bmi by smoker and sex**

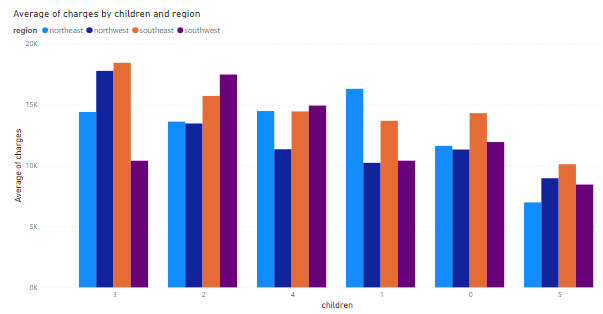


For non-smokers, the average BMI is the same for males and females. For smokers, the average BMI is 3 kg/m^2 less for females than males.

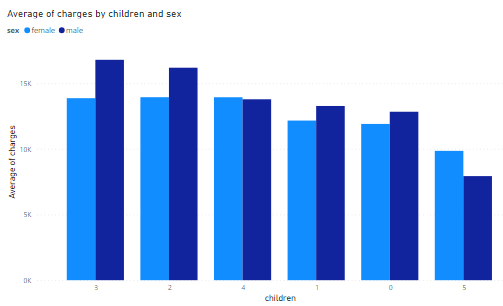
**Average of bmi by sex and region:**



The average BMI of males and females from the southeast is higher than other regions. Northeast and northwest for male and females is approximately the same.



The policyholder with three children covered by insurance from southeast and northwest has higher charges than others. The policyholder with two children covered by insurance from southwest is higher than others with 2 children. The policyholder with one child covered by insurance from the northeast is higher than others with 2 children. The policyholders with no children covered by insurance is less for all regions than others.



The average charges of male policyholders with no children, 1,2,3 children under policy is higher than others



BMI and charges of smokers are positively correlated but have few outliers.

**2) Tools and Techniques**

* Python
* Power BI

1. **Findings**

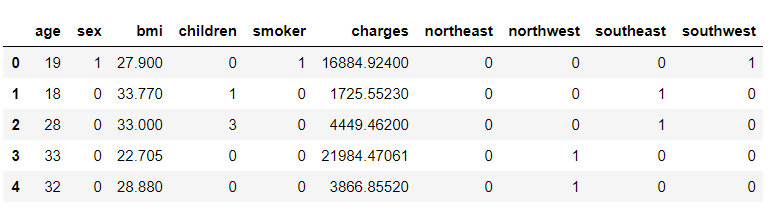
* The policyholders from southeast have high BMI and more smokers of the policyholders are from southeast. The southeast policyholder’s charges are higher than other region policyholders. Number of policyholders from the southeast is also high. Policyholders from the north from the north region have less BMI and charges.
* Policyholders of male and female are equal. The number of female smokers count is slightly higher than male. For smokers and non-smokers the average BMI is almost the same. For non-smokers the average BMI is the same for males and females. For smokers, the average BMI is 3 kg/m^2 less for females than males. The average BMI of both sex is same which is 30 kg/m^2.
* The average charge of smokers is very higher than non-smokers. Male smoker charge is higher than female. The 50% charges of women non-smoker lie between $5000 to $1200.For women smokers 50% of charges lie between $22000 and $43000.The 50% charges of men non-smoker lie between $5000 to $12000.
* The policyholder with three children covered by insurance from southeast and northwest has higher charges than others. The policyholder with two children covered by insurance from southwest is higher than others with 2 children. The policyholder with one child covered by insurance from the northeast is higher than others with 2 children. The policyholders with no children covered by insurance is less for all regions than others.
* Most of the policyholders have no childrens that is approximately 580 out of 1338 policyholders have no childrens.The number of female policyholders with no children under policy is slightly higher than male. But policyholders with other numbers of children under policy are higher male than female. The average charges of male policyholders except with 4 children under policy is higher than other children and females. Policyholders with no children are more benificiary.

**Data Modeling**

1. **Presentation and automation**

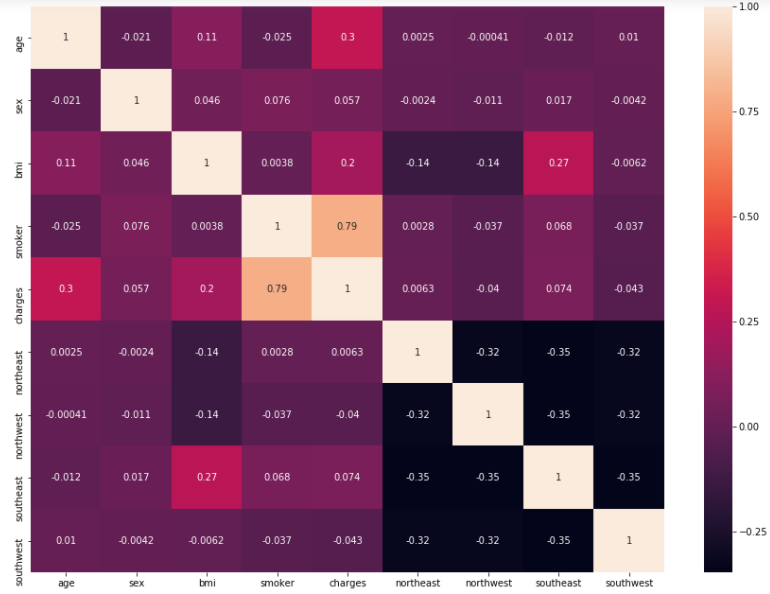
lifestyle and physical factors of the person impact the health condition. So, the expenditure of medical charges of the policyholder is dependent on many factors of the policy holder which include smoking, age, BMI etc. To predict medical charges based on other factors, we need to fit the linear model and check the model adequacy.

**Relabeled dataframe:**



Before modelling, the dummy variables has been created for region and relabelling for sex and smoker has been coded for modelling.

**Correlation Matrix after relabeled:**

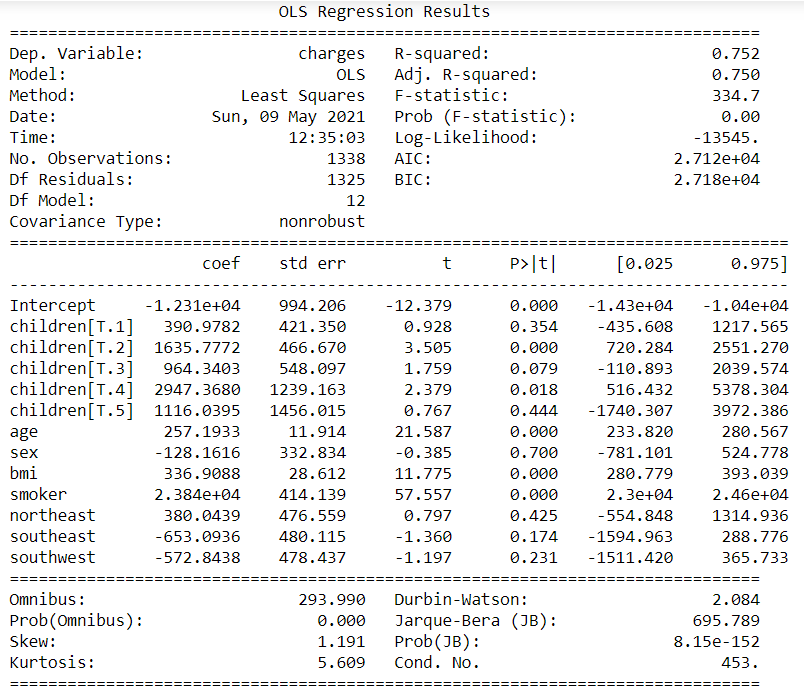


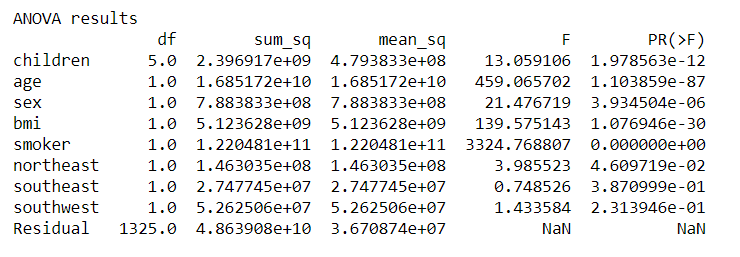
The coefficients of all variables except regions (northeast, northwest, southeast, southwest) are positive which tells that there exists a positive linear relationship between these independent variables & dependent variables (charges). The coefficient of region is negative which tells that there exists a negative linear relationship.

**Multivariate Linear Regression Model:**

To predict the medical cost of the policy holder we will use a linear regression model because the output variable(Charges) is continuous and more than one independent variable will use multiple linear regression.

The multivariate linear regression model has been fitted and the summary of the model has been found.





**Interpretation of model:**

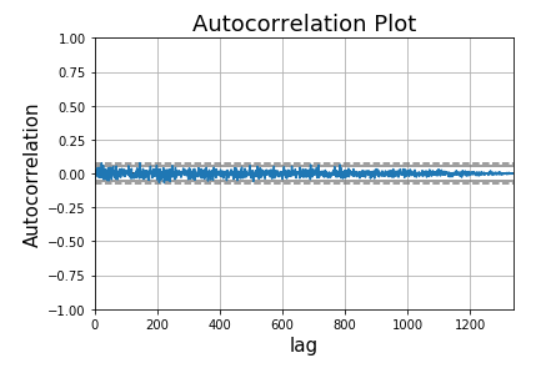
* *The final fitted model is charges = -12307.218 + 390.978\*children[T.1] + 1635.7\*children[T.2] +964.34\*children[T.3]+2947.36\*children[T.4]+ 1116.039\*children[T.5] + 257.193\*age + -128.16\*sex + 336.90\*bmi + 23836.4\*smoker + 380.04\*northeast + ( -653.09)\*southeast + (-572.84)\*southwest.*
* The sign of regression coefficient tells about the linear relationship. The coefficients of all variables except region are positive which tells that there exist the positive linear relationship between these independent variables & dependent variables(charges). The coefficient of region is negative which tells that there exists a negative linear relationship.

#### **Goodness of fit using R square:**

* Adjusted R-square won’t change with respect to the addition of independent variables so we have obtained the adjusted R-square is 0.75 which tells that **75% of charges is predicted by the independent variables(age, bmi, sex, children, smoker, region). Therefore the model is a good fit.**

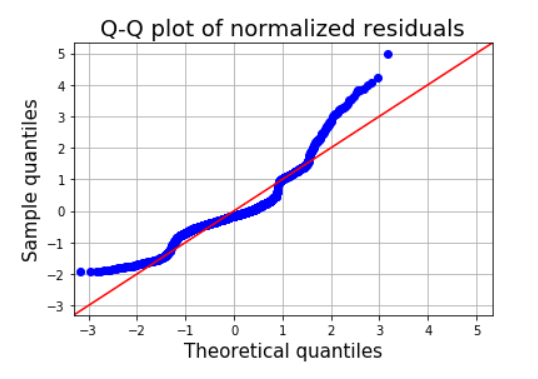
**To Check Model adequacy:**

#### **Autocorrelation test:**



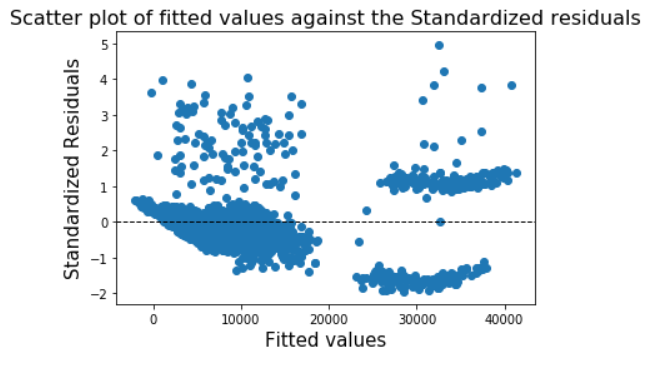
* There are no correlations outside the threshold line that is, correlation of all lags are negligible. Therefore the residuals are uncorrelated which satisfies the assumption on uncorrelated residuals.
* The Durbin-Watson test statistic tests the null hypothesis that the residuals are uncorrelated against the alternative that the residuals are autocorrelated. The Durbin-Watson statistic ranges in value from 0 to 4. A value near 2 indicates non-autocorrelation, a value toward 0 indicates positive autocorrelation, a value toward 4 indicates negative autocorrelation. Here, the Durbin-Watson test statistic value is 1.282 which is near to 2. Therefore, the residuals are uncorrelated i.e) the assumption of residuals is satisfied.

#### **Normality test:**



* The **normal QQ plot** is used to check the normality of residuals. Most of the points are not lying on the straight line which indicates the **residuals are not normally distributed.**
* The **Jarque-Bera test statistic** tests the null hypothesis for the test is that the residuals is normally distributed against the alternate hypothesis is that the residuals does not come from a normal distribution. From the summary of the model, we found that the p value of JB statistics is 2.21e-05 which is less than the significance level 0.05. So the null hypothesis is rejected. Therefore, the residuals are not normally distributed i.e) **the residuals are not normally distributed. Therefore, the normality assumption on residuals is violated.**

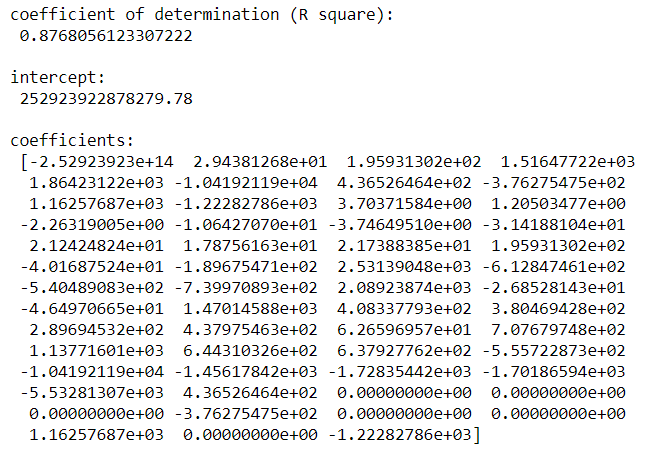
**Linearity test:**



We can clearly see that the **points are following a curve pattern. Therefore the variance of the residuals is not constant and the curve plot indicates polynomial or non-linear model will be suitable for this dataset**

**Polynomial Model:**

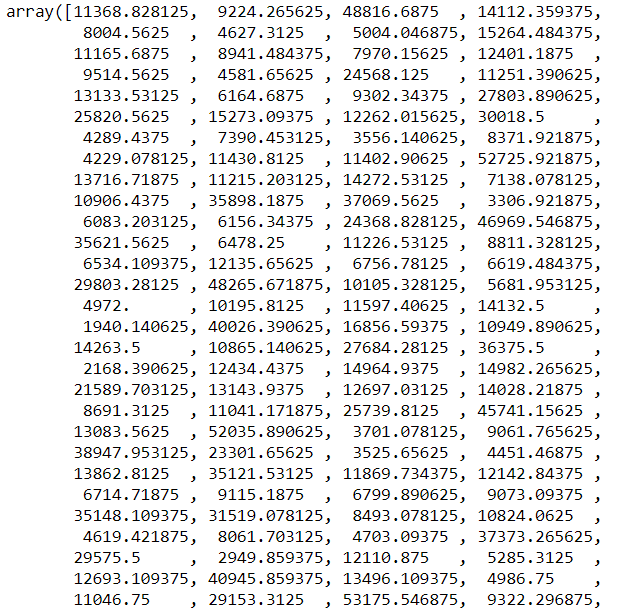
Since multivariate linear model does not satisfy the model assumptions, the polynomial model has been fitted and the summary of polynomial models are,

****

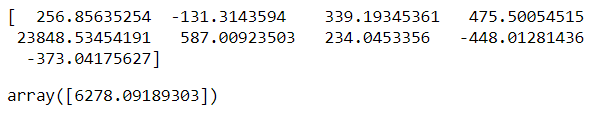
We have obtained the multiple R-square value of polynomial regression model is 0.8768 which is higher than the R square of the multiple linear model and it tells that 88% of medical charges of the patient is predicted by the independent variables. Therefore, the model is a good fit.

**Prediction:**

The predicted values of the model are

****

**Prediction for unknown values**

****

$ 6278.09 is predicted charge for the male non-smoker policyholder from southwest with age 30, bmi 30

**Conclusion:**

The count of male smokers is higher than female smokers. The average age of policyholders is 40 and average BMI is 30 kg/m^2. The average charge of smokers is higher than non-smokers.Since the person who is having a high BMI and high smoking habit causes health issues, that person needs high medical insurance. Here, **the average BMI of policyholders from southeast is high and also more smokers from southeast regions so their average charges of southeast policyholders is also high.** Most of the policyholders have no childrens that is approximately 580 out of 1338 policyholders. So the **average charge of policyholders with** **no children is higher than the other.**

The R-square value of the polynomial regression model is 0.8768 which is higher than the R-square value of the multiple linear regression model and it tells that 88% of the total variation in the medical charges is explained by the independent variables. Therefore, the model is a good fit.

The predicted values have been obtained for the model. And $ 6278.09 is predicted charge for the male non-smoker policyholder from southwest with age 30, bmi 30