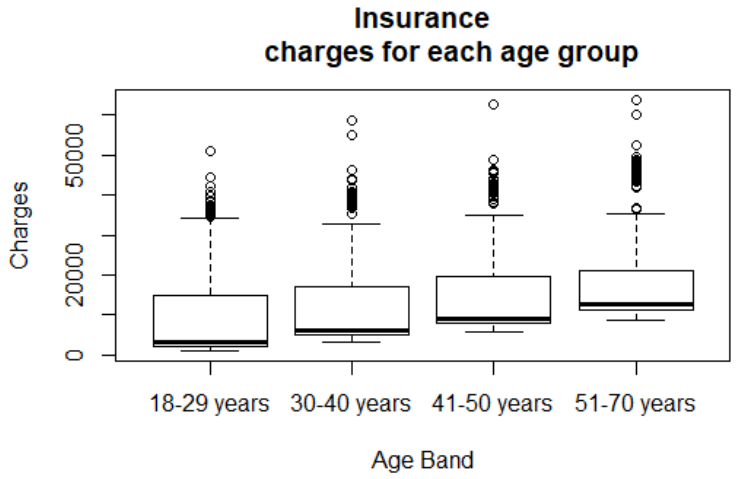
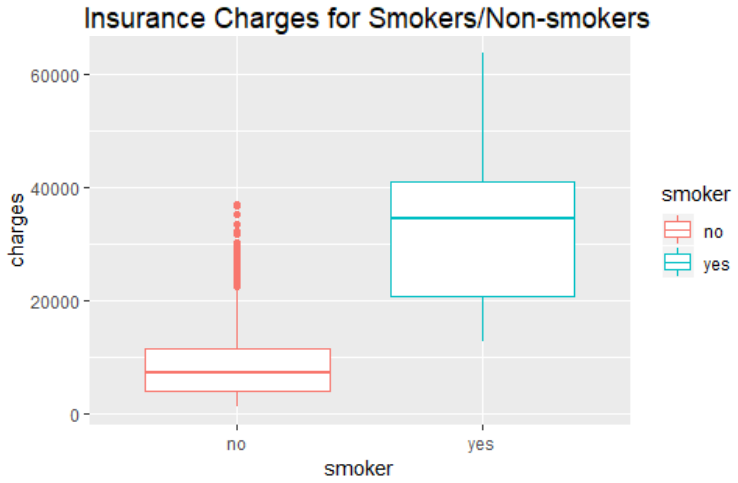
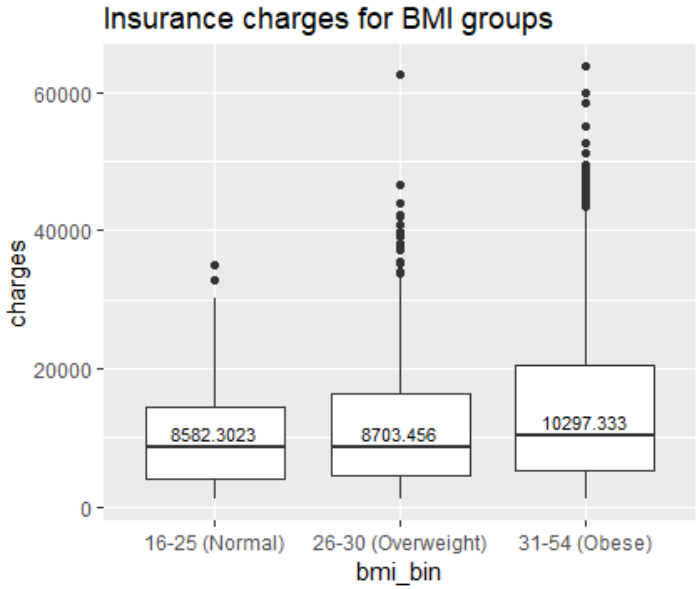
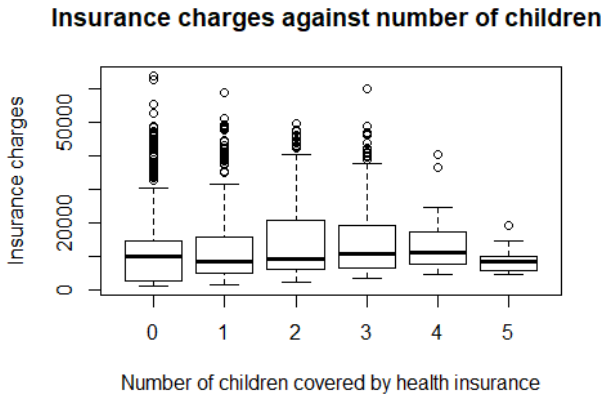
Statistical Analysis & Prediction of Insurance Charges Analysis

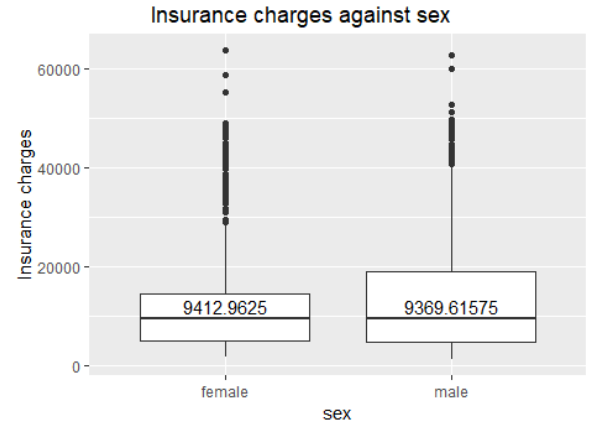
Relationship between insurance charges and factors affecting it





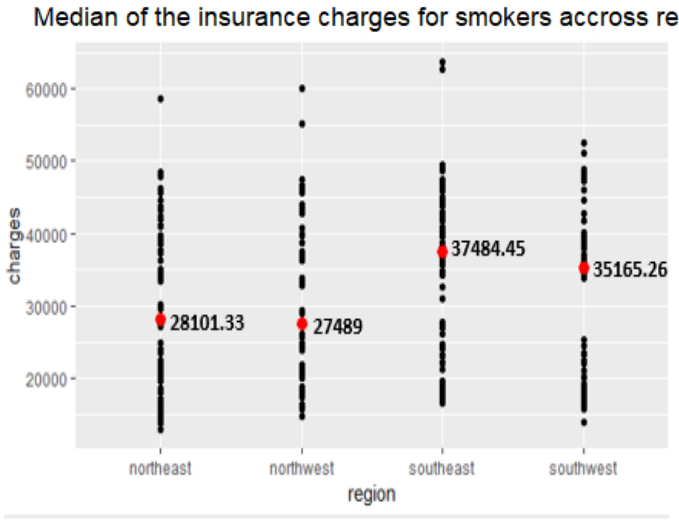


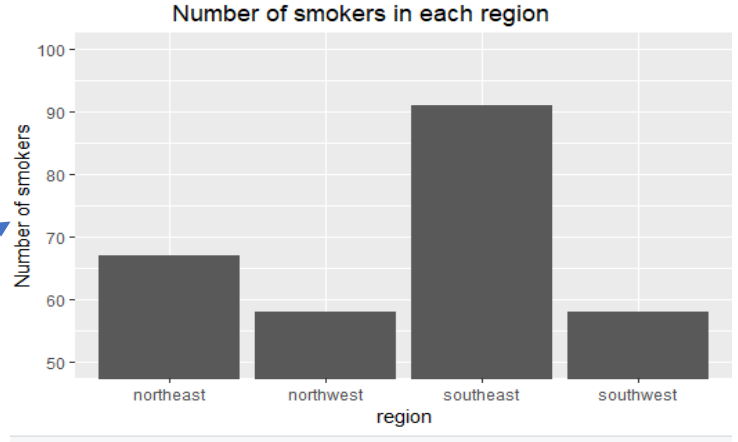




Insurance Charges Analysis for the smokers in southeast regions

we can see that 50 % of the smokers in southeast region has insurance charges above $ 37484.45 which is higher than other regions





Number of smokers in southeast region are high as compared to other regions

BMI for male and female smokers in southeast region is high as compared to other regions

Note : This clearly suggest that insurance charges increase in a specific region, as BMI and number of smokers increases, but overall insurance charges do not increase or decrease w.r.t regions

Advanced Statistical Analysis

For the smokers in southeast region

Identifying 95 % confidence interval for insurance charges and BMI of the smokers in southeast region

- Average insurance charges of the population who smokes in southeast region will be between $32518.21 and $37171.78
- Mean of the Body Mass Index of the population who smokes in southeast region will be between 31.66 and 34.54

For smokers and non smokers (irrespective of sex and region)

Identifying 95 % confidence interval for the difference between average insurance charges for smoker and no- smoker

- Average insurance charge difference for smokers and non-smokers of population will be between $22202.72 & $25029.21

Hypothesis Testing

- Upon doing, hypothesis testing, we found that there is significant difference between average insurance charges for “smokers, non-smokers”
- Also, variance of the insurance charges of “smokers, non-smokers” are different

Identifying variables affecting Insurance charges by using linear regression

- As we have seen earlier, that insurance charges has almost linear relationship with age, BMI and number of children covered by health insurance. Therefore we don't have to transform any of our predictor

- We created dummy variable for sex, smoker and region

- After using exhaustive method we finalized our top 3 models as below,
    - Charges~ age+bmi+children+smoker
    - Charges~age+bmi+smoker
    - Charges~age+ smoker

p	(Intercept)	age	sex_dummy	bmi	children	smoke_dummy	region_dummy	SSRes	R2	AdjR2	MSE	Cp
2	1	0	0	0	0	1	0	7E+10	0.62	0.6195	6E+07	696.4
2	1	1	0	0	0	0	0	2E+11	0.089	0.0887	1E+08	3528
2	1	0	0	1	0	0	0	2E+11	0.039	0.0386	1E+08	3796
3	1	1	0	0	0	1	0	5E+10	0.721	0.721	4E+07	155.6
3	1	0	0	1	0	1	0	7E+10	0.658	0.6574	5E+07	494.5
3	1	0	0	0	1	1	0	7E+10	0.624	0.623	6E+07	677.9
4	1	1	0	1	0	1	0	5E+10	0.747	0.7469	4E+07	18.41
4	1	1	0	0	1	1	0	5E+10	0.724	0.7231	4E+07	145.1
4	1	1	0	0	0	1	1	5E+10	0.721	0.7208	4E+07	157.5
5	1	1	0	1	1	1	0	5E+10	0.75	0.7489	4E+07	8.568
5	1	1	0	1	0	1	1	5E+10	0.748	0.7477	4E+07	15.24
5	1	1	1	1	0	1	0	5E+10	0.747	0.7467	4E+07	20.3
6	1	1	0	1	1	1	1	5E+10	0.751	0.7498	4E+07	5.155
6	1	1	1	1	1	1	0	5E+10	0.75	0.7488	4E+07	10.42
6	1	1	1	1	0	1	1	5E+10	0.748	0.7475	4E+07	17.13
7	1	1	1	1	1	1	1	5E+10	0.751	0.7496	4E+07	7

- By using Backward variable selection method, we get our final model as
    - Charges~age+bmi+children+smoker

- Also, there is no correlation between age, BMI and children. Therefore there is no problem of multicollinearity

- There are few leverage points available in the data but there is no influential point present in the data

Output

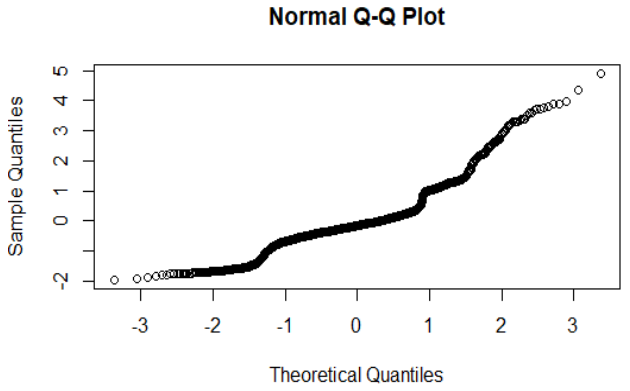
Model Summary

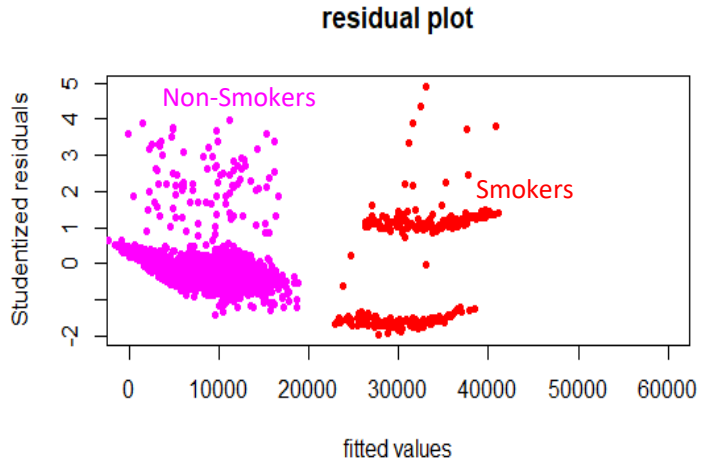
```
Call:
lm(formula = charges ~ age + bmi + children + smoke_dummy, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11897.9  -2920.8   -986.6   1392.2  29509.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smoke_dummy  23811.40     411.22  57.904  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```





- 74.97 % of variation in insurance charges is explained through the regression on age, BMI, children and smoker
- Average insurance charges for smoker is $23811.4 higher than non-smoker at same age, bmi and # of children
- Errors are normally distributed
- Residuals have constant variance for smokers and non smokers