**Problem Statement (Situation):**

“Finding out the health parameters that affect health insurance claims”

An insurance company in the US is reviewing its insurance claims/charges and is trying to do a cause and effect analysis for future business decisions. It has collected data for its customers who have made claims till recent time. The data-points collected are Age, Gender, BMI, Number of children/dependents, smoking habit, region they belong to, charges/bills claimed under the insurance. This analysis would have a bearing on what premium the company charge a customer availing an insurance policy should. The insurance company has collected a dataset of 1338 customers-claims.

**Objective (Task):**

* **To do a cause and effect analysis on historic-data of insurance claims.**

You have been appointed as the “Analyst” for this project to achieve the objective of the study, your tasks are as under:

Q1. f) Has charges got something to do with the number of dependents?

Answer:-

The observation tells us from the plotted pivot table is as follows,

For Charges Vs Age:

We can say that as the Age range is increasing the charges for medical cost are also increasing simultaneously.

For Charges Vs BMI:

By observing this pivot chart we interpret that here the BMI range affects the Charges of medical cost affects effectively.

Charges Vs Smoker/Non-Smoker:

Here we can say that charges of medical cost for smoker are higher as compared with the non-smokers. That means as the smoking rate increases the charges increases and vice versa for non-smokers.

Region wise Charges for Smoker/Non-smoker:

By observing along with region the charges for smoker and non-smoker does not differ much and remains same with charges of smoker and non-smoker.

Hence, we can definitely say that charges of medical cost billed by insurance, changes with several number of dependent variables depending upon the condition.

Q1. g) Do a similar dependents-charges analysis, Region-wise.

Answer:

Region wise Smoker/Non-Smoker:

Yes definitely we can say that according to the region the rate of smoker are increasing. Southeast region has the highest smoking rate and northeast has the least smoking rate.

Region wise Charges for Smoker/Non-Smoker:

By observing along with region the charges for smoker and non-smoker does not differ much and remains same with charges of smoker and non-smoker.

Q.1.j) Give your interpretation for observations made in point (c)

Answer:

For Male Vs Female:

As we see we can interpret that smoking rate for number of males are higher than compared with females.

For Charges Vs Age:

We can say that as the Age range is increasing the charges for medical cost are also increasing simultaneously.

For Charges Vs BMI:

By observing this pivot chart we interpret that here the BMI range affects the Charges of medical cost affects effectively.

Charges Vs Smoker/Non-Smoker:

Here we can say that charges of medical cost for smoker are higher as compared with the non-smokers. That means as the smoking rate increases the charges increases and vice versa for non-smokers.

Q.3) Descriptive Summary analysis

Answer:

* Std\_error for smoker is almost less than 0.05 which indicates 0 error.
* Age has a variance of 197.
* Kurtosis for all values goes from normal to playtikurtic.
* Skewness here mostly indicates a relatively symmetric distribution as it near to 0 accept for charges.

Q3) Multiple Linear Regression analysis

Answer:

1. The F- value:

The F-value is very high and the P-value of the F- stats is 0 which means that means that we can trust this regression model coefficients for our analysis.

2. The multiple R:

This value is 0.8 which is a high positive correlation between the actual and predicted values.

3. R-square%-

The R- square value of this model is 75% which is not the good proportion of variance that is explained by this model. At least 90% of variance is expected to explain. This gives us an idea that we need to look our independent variable and optimize the model further.

4. Adjusted R square:

The adjusted- R square is almost the same as R- square which indicates that the current is not affected by the number of independent variables.

5. Standard Error:

The STD error of the model seems to be too the high which indicates that we may want to optimize the model further to reduce the STD error. This tells that on average the predicted charges of medical cost billed by insurance could deviate from the actual charges by around 6063 units.

6. P-value:

Sex, regions like southwest, southeast, northwest & northeast have very high p values which gives us an indication that there are significant variables and we should try to improve the model performance by eliminating these variables.

* To note after making changes in the regression model F- value seems to increases that the previous number which is a good indication of explained data.
* And P- value seems to approximately normal with the repaired dataset

Hence, we conclude that by eliminating Sex and regions give us a perfect fit data for our model.