**Case Study for presentation**

An insurance company in US is reviewing its insurance claim/charges

and trying to do a cause and effect analysis for future business decisions. It has collected data for its

customers’ age, gender, bmi, number of children/dependents, smoking habit, region they belong to,

charges/bills claimed under the insurance.

Your tasks are as under:

(Ensure that you are interpreting the operations at all steps as you shall be presenting your findings

as an assignment) – Two files to be prepared and submitted by each group (Excel + PPT).

1. Perform the basic Exploratory Data Analysis on the sample data.

a. Identify the categorical and continuous variables

b. Make Histograms and box plots for continuous variables, do a correlation analysis.

c. Make relevant Pivot tables and charts for :

i. Male/Female ratio and which gender has more smokers

ii. Charges vs Age

iii. Charges vs BMI

iv. Charges for Smokers vs Non-smoker

v. Region-wise Smokers vs non-smokers analysis with one or more pivot table

and charts

vi. Region-wise charges for smokers vs non-smokers

vii. Has charges got something to do with no. of dependents ?

viii. Do a similar dependents-charges analysis, Region-wise

ix. Do at least one more pivot table and chart of your own choice, if needed

d. Give your understanding from the patterns observed in point (b)

e. Give your interpretation for observations made in point ( c )

2. Edit the data as following, to obtain dummy variables:

a. Sex : Replace all the “Males” with “1” and “Females” with “0”, creating numerical

entries for gender this way will help you do analysis further. You can use Replace

with “Match entire cell content option. Do a replace all to save time.

b. Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.

c. Region: We always create one less category column for the dummy data w.r.t the

categories available for that original variable. So for Region, we will create three

dummy columns, assuming “Northeast” as zero and omit the column for it. Now

create three columns for “northwest”, “Southeast”, “Southwest”. Whichever row

has “northwest” region as an entry will take “1” as an entry otherwise “0” in

“northwest” column. Similarly in “Southeast” column, which ever row had

“southeast” as an entry will take “1” as the new entry and “0” for rest of the rows.

Do a similar operation on “Southwest” column.

1. Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.