# Assignment – Health Insurance claim

**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 should the company charge a customer availing an insurance policy.

The insurance company has collected a dataset of 1338 customers-claims. Please refer to the data dictionary below:

**Data Dictionary:**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Age | Age of the customer/claimant who has claimed insurance for medical treatment charges |
| Sex | Gender of the customer/claimant |
| bmi | Health parameter: person's weight in kilograms divided by the square of height in meters |
| Children | No. of children the claimant has |
| Smoker | Whether the claimant smokes or not |
| Region | Region to which the claimant belongs |
| Charges | The exact medical charges for which the claimant has claimed insurance |

**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:

1. Perform the Exploratory Data Analysis on the data.
   1. Identify the categorical and continuous variables

Solution 1.a)

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **Categorical Variables** | **Continuous Variables** |
|  | Sex | Age |
|  | Childern | Bmi |
|  | Smoker | Charges |
|  | Region |  |

* 1. Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis)

Solution 1.b) EDA interpretation:

1.The demography that we are looking at has more no. of young people in the age bracket of 18-30

2. More no of people have lesser insurance charges, whereas lesser no of people have high insurance charges, It would be interesting to know who are these people

3. Age and charges graphs slightly hint that middle aged people and probably older people might be getting billed more under the insurance

4. Is age the only factor or are there more factors contributing to the insurance charges ?

5. BMI is distributed normally

6.Charges and bmi have outliers, these might be some points to study separately, whether higher bmi is causing higher health problems and claims as result

* 1. Make relevant Pivot tables and charts for:
     1. Male/Female ratio and share information on which gender has more smokers

Solution 1.c) Males smokers exceed female smokers by: 38%

Male to female ratio is: 1.021148036

* + 1. Charges vs Age

Solution1.c) ii. There is a positive relation between Insurance charges and Age; Insurance bills increase with Age

* + 1. Charges vs BMI

Solution1.c) iii. Seems like BMI has no majr relationship with insurance bills/charges

* + 1. Charges for Smokers vs Non-smokers

Solution 1.c) iv. Smokers get billed much more as compared to non-smokers

* 1. Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts

Solution 1.d) There are 20% smokers on average basis, in our sample Southeast Region has the largest no of smokers

* 1. Region-wise charges for smokers vs non-smokers

Solution 1.e) It seems like, on an average, the Smokers from Southeast region get billed the most.This can only be vaildated through regression analysis

Ratio of smokers/non-smokers

|  |  |
| --- | --- |
|  | 3.237514 |
|  | 3.528561 |
|  | 4.338155 |
|  | 4.023933 |

* 1. Has charges got something to do with the number of dependents?

Solution 1.f) The claimants with 2 -3 children make more claim on an average.

Among non-smokers the charges claimed increase with increase in no. of children whereas amongst smokers it increases till 3 chikdren then falls down; however it is still more than that of non-smokers

* 1. Do a similar dependants-charges analysis, Region-wise

Solution 1.g) At first look it seems like a person from southeast region and three chidren claims the most insurance money, so is the case with northwest region. Likewise there are a few other red-marked areas

* 1. Do at least one more pivot table and chart of your own choice on the remaining variables.

Solution 1.h) We can extract regionwise charges using filters and copy-paste separately for more graphs and insights

* 1. Give your interpretation for observations made in point (b)

Solution 1.i) EDA interpretation:

1.The demography that we are looking at has more no. of young people in the age bracket of 18-30

2. More no of people have lesser insurance charges, whereas lesser no of people have high insurance charges, It would be interesting to know who are these people

3. Age and charges graphs slightly hint that middle aged people and probably older people might be getting billed more under the insurance

4. Is age the only factor or are there more factors contributing to the insurance charges ?

5. BMI is distributed normally

6.Charges and bmi have outliers, these might be some points to study separately, whether higher bmi is causing higher health problems and claims as result

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

Solution 1.j) 1.Males smokers exceed female smokers by : 38%

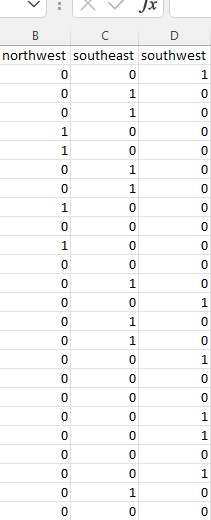
Male to female ratio is : 1.021148036

2. There is a positive relation between Insurance charges and Age; Insurance bills increase with Age

3. Seems like BMI has no majr relationship with insurance bills/charges

4. Smokers get billed much more as compared to non-smokers

1. Edit the data as following, to obtain dummy variables: **(5 marks)**
   1. 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 the “Replace with Match entire cell content” option. Do a replace all to save time.
   2. Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.
   3. 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 the “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest of the column (Southeast). Do a similar operation on the “Southwest” column. Please refer to the below image for your understanding,



Solution 2. We drop Northeast region as we would be taking it as the reference for dummy variables. The data on the left hand side will be used to run regression analysis

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.

Solution 3. Final interpretation:

1) Smoker, age, bmi have a very strong and direct impact on increasing insurance charges claim

2) Also a smoker in southeast and southwest region will claim higher charges

3) If we look at only the region as an idependant variable then people from south east and south west region claim lesser charges w.r.t our reference region northeast

4) Claimants having more no. of children also show higher claims

# Learning Outcome (Result):

* Understand implementation of Exploratory Data Analysis to understand the nature of different data-attributes, through pivot tables and different types of visualizations
* Understand how to use various statistical/analytical tools in MS Excel like Summary statistics, Histogram, correlation table, Pivot tables, Regression analysis (using Data analysis tool pack)