Health Insurance Claim

Data analysis

gurvinder Singh

[Year]

**About data set:**

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.

The insurance company has collected a dataset of 1338 customers-claims.

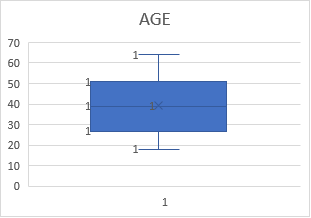
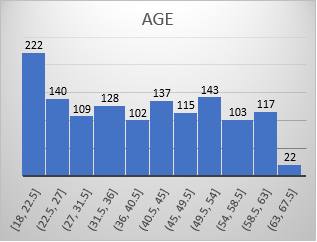
**Problem Statement:**

1. Identify the categorical and continuous variables from dataset.
2. Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis). Make the observations.

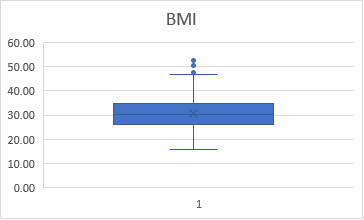
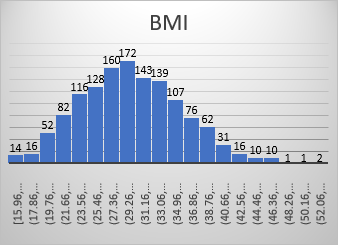
**Solution:**

A separate table is made of different variables according to data classification by using logics. After that histograms are plotted for each continuous variables separately along with box plots. The correlation has been done for the continuous variables with the help of Data Analysis Tool pack.

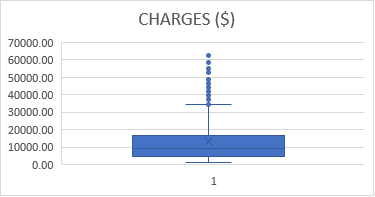
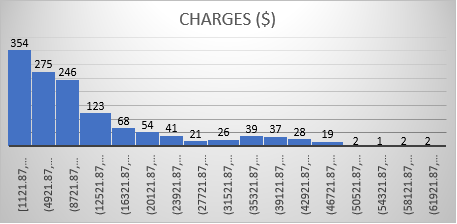
|  |  |
| --- | --- |
| **Variables** | **Data Classification** |
| Age | Continuous |
| Sex | Categorical |
| Bmi | Continuous |
| Children | Discrete |
| Smoker | Categorical |
| Region | Categorical |
| Charges ($) | Continuous |

|  |  |
| --- | --- |
| Observation - **Age** | The majority of the age lies within the range (18,22.5) followed by (49.5,54). The age variable is having a good range between 18 to 67.5. It is an independent variable. There are no outliers . There is huge difference of age at the starting point/range and ending/range point. It is highly skewed towards right side . |

|  |  |
| --- | --- |
| Observation - **Bmi** | In this variable the majority of Bmi lies within the range (28.35,31.45) followed by (25.25,28.35) and 31.45,34.55). It shows normal distribution. There are also outliers in the data. It is also an independent variable. |

|  |  |
| --- | --- |
| Observation - **Charges ($)** | It is a dependent variable. There are so many outliers in the data. The data is highly skewed towards right side. There is huge difference of charges ($) at the starting range/point and ending range/point.The majority of charges ($) lies within range of (1121.87,..). |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *age* | *bmi* | *charges($)* | **Correlation of Continuous variables** |
| age | 1 |  |  |
| bmi | 0.109272 | 1 |  |
| charges($) | 0.299008 | 0.198341 | 1 |

|  |  |
| --- | --- |
| Observation - **Correlation** | The correlation indicates the positive relation between dependent variable and independent variables. The highest correlated pair is of age and charges followed by bmi and charges. |

**Problem Statement:**

Make relevant Pivot tables and charts for:

1) Male/Female ratio and share information on which gender has more smokers

2) Charges vs Age

3) Charges vs BMI

4) Charges for Smokers vs Non-smokers

Make the observations.

**Solution:**

To insert the pivot table, we will select any cell within the data and then range must be selected for pivot table. We can display the table either on same worksheet or even on new worksheet. After do analysis as per the requirements.

(Note- It is very difficult for some the pivot tables to show in Report. It is better to take reference from the excel sheet attached.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of smoker** | **Column Labels** |  |  |
| **Row Labels** | **no** | **yes** | **Grand Total** | **Gender vs Smoker (Yes, No)** | Observation | There are more male smokers than female smokers |
| female | 547 | 115 | 662 |
| male | 517 | 159 | 676 |
| **Grand Total** | **1064** | **274** | **1338** |

**Age vs Average Charges**

|  |  |
| --- | --- |
| Observation - **Age vs Average Charges** | The highest average charges paid to customer is of age 64 and the lowest charges paid to customer is of age 21. The high average charges of $20000 and above are paid to customers of different ages like 60, 61, 63, 64. The data is skewed towards left side of graph. |

(Note- The pivot table can not be inserted in the business report because the content of the pivot table is very large. So, it is better to take reference from the excel sheet. The relation between the **age** and **average charges** is shown in above graph with observation.)

**Bmi vs Charges**

|  |  |
| --- | --- |
| Observation - **Bmi vs Charges** | The Bmi is independent variable. The highest charges paid to customer of bmi 32.30 followed by bmi 38.06. The data is skewed towards left side. |

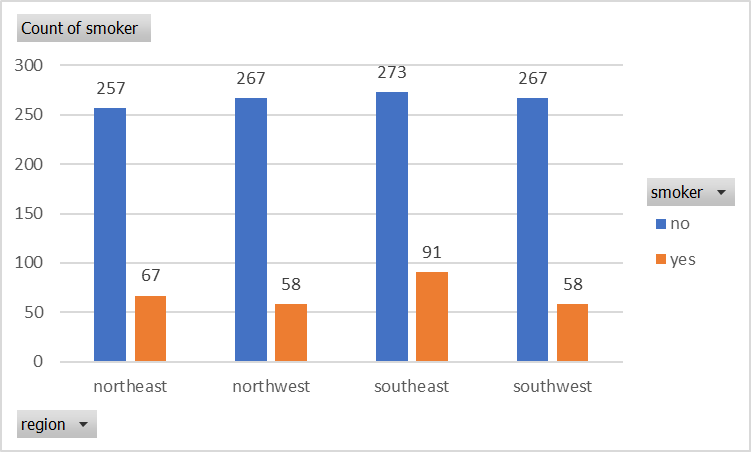
(Bmi- Health parameter: person's weight in kilograms divided by the square of height in meters.)

(Note- The pivot table cannot be inserted in the business report because the content of the pivot table is very large. So, it is better to take reference from the excel sheet. The relation between the **bmi** and **average charges** is shown in above graph with observation.)

**Smokers (Yes,No) vs Average Charges**

|  |  |
| --- | --- |
| **Row Labels** | **Average of charges($)** |
| no | 8434.27 |
| yes | 32050.23 |
| **Grand Total** | **13270.42227** |

|  |  |
| --- | --- |
| Observation - **Smokers (Yes, No) vs Charges** | As per the data, the charges are more paid to smoker customers as compared to non-smoker customers. |



**Problem Statement:**

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

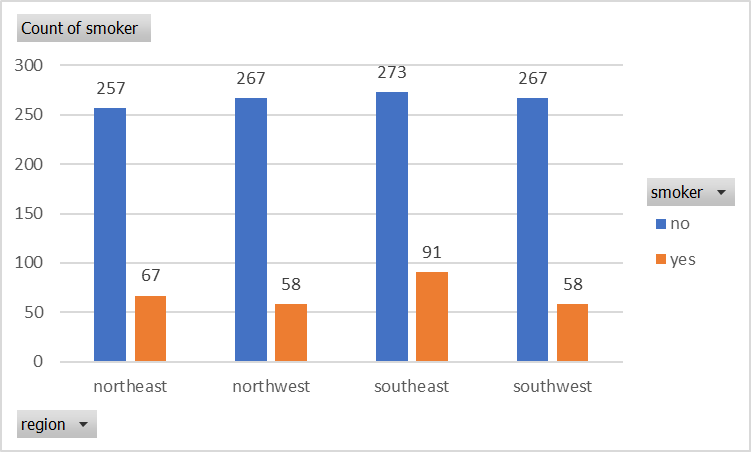
2) Region-wise charges for smoker’s vs non-smokers.

**Solution:**

We will make pivot tables to do analysis.

**Region vs Smokers (Yes, No)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of smoker** | **Column Labels** |  |  |
| **Row Labels** | **no** | **yes** | **Grand Total** |
| northeast | 257 | 67 | 324 |
| northwest | 267 | 58 | 325 |
| southeast | 273 | 91 | 364 |
| southwest | 267 | 58 | 325 |
| **Grand Total** | **1064** | **274** | **1338** |



(The strength of the smokers is more in the southeast region as compared to other regions. The least smokers are in southwest and northwest region. As we can see 1064 are non-smoker customers and 274 are smoker customers.)

**Region-wise average charges for smoker’s vs non-smokers**

|  |  |  |  |
| --- | --- | --- | --- |
| **Average of charges ($)** | **Column Labels** |  |  |
| **Row Labels** | **no** | **yes** | **Grand Total** |
| northeast | 9165.53 | 29673.54 | 13406.38 |
| northwest | 8556.46 | 30192.00 | 12417.58 |
| southeast | 8032.22 | 34845.00 | 14735.41 |
| southwest | 8019.28 | 32269.06 | 12346.94 |
| **Grand Total** | **8434.27** | **32050.23** | **13270.42** |

(The above table shows the region wise average charges claimed by smoker and non-smoker customers. The highest claim received by non-smoker customers is in northeast region and lowest claim received is in southwest region. The highest claim received by smoker customers is in southeast region and lowest claim received is in northeast region. We can also see that smoker customers claimed more than the non-smoker customers.)

**Problem Statement:**

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

2) Do a similar dependants-charges analysis, Region-wise

3) Do at least one more pivot table and chart of your own choice on the remaining variables

**Solution:**

We will make pivot table to do analysis and give observations.

Children vs Avg Charges

|  |  |
| --- | --- |
| **Row Labels** | **Average of charges ($)** |
| 0 | 12365.98 |
| 1 | 12731.17 |
| 2 | 15073.56 |
| 3 | 15355.32 |
| 4 | 13850.66 |
| 5 | 8786.04 |
| **Grand Total** | **13270.42** |

(The highest charges claimed by customers having 3 children followed by customers having 2 children. The lowest charges claimed by customers having 5 children. According to me, the number of children is not affecting the charges. So, there is no cause-and-effect relationship between children (dependents) and charges.)

Dependents- Charges analysis, Region wise

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Average of charges($)** | **Column Labels** |  |  |  |  |
| **Row Labels** | **northeast** | **northwest** | **southeast** | **southwest** | **Grand Total** |
| 0 | 11626.46 | 11324.37 | 14309.87 | 11938.50 | 12365.98 |
| 1 | 16310.21 | 10230.26 | 13687.04 | 10406.48 | 12731.17 |
| 2 | 13615.15 | 13464.31 | 15728.47 | 17483.49 | 15073.56 |
| 3 | 14409.91 | 17786.16 | 18449.85 | 10402.44 | 15355.32 |
| 4 | 14485.19 | 11347.02 | 14451.02 | 14933.26 | 13850.66 |
| 5 | 6978.97 | 8965.80 | 10115.44 | 8444.16 | 8786.04 |
| **Grand Total** | **13406.38** | **12417.58** | **14735.41** | **12346.94** | **13270.42** |

Average Charges of Males and Females in Region wise

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Average of charges ($)** | **Column Labels** |  |  |  |  |
| **Row Labels** | **northeast** | **northwest** | **southeast** | **southwest** | **Grand Total** |
| female | 12953.20 | 12479.87 | 13499.67 | 11274.41 | 12569.58 |
| male | 13854.01 | 12354.12 | 15879.62 | 13412.88 | 13956.75 |
| **Grand Total** | **13406.38** | **12417.58** | **14735.41** | **12346.94** | **13270.42** |

(In the above pivot table there is not much difference in the insurance claimed by males and females. In the southeast region the insurance is claimed highest as compared to other regions and lowest is claimed by southwest region.)

**Problem Statement:**

Edit the data to obtain dummy cells. Replace all the “Males” with “1” and “Females” with “0”, creating numerical entries for gender. In the same way replace all the “Smokers” with “1” and “Non-smokers” with “0” and the “Regions” like “northwest”, “Southeast”, “Southwest”.

**Solution:**

We will create dummy columns for Gender (Dummy Male, Dummy Female). In the same way we will create dummy columns for Smoker (Dummy Smokers, Dummy Non-Smokers) and also for the Regions (Dummy Northwest, Dummy Southeast, Dummy Southwest). Then the formula **“=IF(B11="male",1,0)”** is used to fill the number entries in column (Dummy Male.). Likewise, the same If formula is used to create Dummy columns for Females, Smokers, Non-Smokers, Northwest, Southeast, Southwest. But the content of the formula should be changed according to requirements.

The some the examples of result will be:

|  |  |
| --- | --- |
| Dummy (Male) | Dummy (Female) |
| 0 | 1 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 0 | 1 |

|  |  |
| --- | --- |
| Dummy (Smoker) | Dummy (Non-Smoker) |
| 1 | 0 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |

|  |  |  |
| --- | --- | --- |
| Dummy (Northwest) | Dummy (Southeast) | Dummy (Southwest) |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |

**Problem Statement:**

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.

**Solution:**

A correlation matrix is simply a table which displays the correlation coefficients for different variables. A correlation has done with the help of Data Analysis Tool pack and in it we must select range of data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *age* |  | *Dummy (Male)* |  | *Dummy (Female)* |  |
|  |  |  |  |  |  |
| Mean | 39.20702541 | Mean | 0.505231689 | Mean | 0.494768311 |
| Standard Error | 0.384102419 | Standard Error | 0.013673526 | Standard Error | 0.013673526 |
| Median | 39 | Median | 1 | Median | 0 |
| Mode | 18 | Mode | 1 | Mode | 0 |
| Standard Deviation | 14.04996038 | Standard Deviation | 0.500159569 | Standard Deviation | 0.500159569 |
| Sample Variance | 197.4013867 | Sample Variance | 0.250159595 | Sample Variance | 0.250159595 |
| Kurtosis | -1.245087653 | Kurtosis | -2.002556636 | Kurtosis | -2.002556636 |
| Skewness | 0.055672516 | Skewness | -0.020951397 | Skewness | 0.020951397 |
| Range | 46 | Range | 1 | Range | 1 |
| Minimum | 18 | Minimum | 0 | Minimum | 0 |
| Maximum | 64 | Maximum | 1 | Maximum | 1 |
| Sum | 52459 | Sum | 676 | Sum | 662 |
| Count | 1338 | Count | 1338 | Count | 1338 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *bmi* |  | *children* |  | *Dummy (Smoker)* |  |
|  |  |  |  |  |  |
| Mean | 30.66339686 | Mean | 1.094917788 | Mean | 0.204783259 |
| Standard Error | 0.166714232 | Standard Error | 0.032956155 | Standard Error | 0.01103632 |
| Median | 30.4 | Median | 1 | Median | 0 |
| Mode | 32.3 | Mode | 0 | Mode | 0 |
| Standard Deviation | 6.098186912 | Standard Deviation | 1.20549274 | Standard Deviation | 0.403694038 |
| Sample Variance | 37.18788361 | Sample Variance | 1.453212746 | Sample Variance | 0.162968876 |
| Kurtosis | -0.050731531 | Kurtosis | 0.202454147 | Kurtosis | 0.145755539 |
| Skewness | 0.284047111 | Skewness | 0.93838044 | Skewness | 1.46476616 |
| Range | 37.17 | Range | 5 | Range | 1 |
| Minimum | 15.96 | Minimum | 0 | Minimum | 0 |
| Maximum | 53.13 | Maximum | 5 | Maximum | 1 |
| Sum | 41027.625 | Sum | 1465 | Sum | 274 |
| Count | 1338 | Count | 1338 | Count | 1338 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Dummy (Non-Smoker)* |  | *Dummy (Northwest)* |  | *Dummy (Southeast)* |  |
|  |  |  |  |  |  |
| Mean | 0.795216741 | Mean | 0.242899851 | Mean | 0.272047833 |
| Standard Error | 0.01103632 | Standard Error | 0.011728017 | Standard Error | 0.012170498 |
| Median | 1 | Median | 0 | Median | 0 |
| Mode | 1 | Mode | 0 | Mode | 0 |
| Standard Deviation | 0.403694038 | Standard Deviation | 0.428995407 | Standard Deviation | 0.445180784 |
| Sample Variance | 0.162968876 | Sample Variance | 0.18403706 | Sample Variance | 0.19818593 |
| Kurtosis | 0.145755539 | Kurtosis | -0.559856699 | Kurtosis | -0.949522817 |
| Skewness | -1.46476616 | Skewness | 1.200409261 | Skewness | 1.025621147 |
| Range | 1 | Range | 1 | Range | 1 |
| Minimum | 0 | Minimum | 0 | Minimum | 0 |
| Maximum | 1 | Maximum | 1 | Maximum | 1 |
| Sum | 1064 | Sum | 325 | Sum | 364 |
| Count | 1338 | Count | 1338 | Count | 1338 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Dummy (Southwest)* |  | *charges($)* |  |
|  |  |  |  |
| Mean | 0.242899851 | Mean | 13270.42227 |
| Standard Error | 0.011728017 | Standard Error | 331.0674543 |
| Median | 0 | Median | 9382.033 |
| Mode | 0 | Mode | 1639.5631 |
| Standard Deviation | 0.428995407 | Standard Deviation | 12110.01124 |
| Sample Variance | 0.18403706 | Sample Variance | 146652372.2 |
| Kurtosis | -0.559856699 | Kurtosis | 1.606298653 |
| Skewness | 1.200409261 | Skewness | 1.515879658 |
| Range | 1 | Range | 62648.55411 |
| Minimum | 0 | Minimum | 1121.8739 |
| Maximum | 1 | Maximum | 63770.42801 |
| Sum | 325 | Sum | 17755824.99 |
| Count | 1338 | Count | 1338 |

The multi linear regression model is built with the dependent variable and all the independent variable which are significant predictor of dependent variable (Charges $). The range must be selected for dependent variable (y) and all the independent variables (x1,x2,x3,\_ \_ \_ \_ xn). We can also plot residuals in summary output also if needed.

Summary Output

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.866552 |
| R Square | 0.750913 |
| Adjusted R Square | 0.747909 |
| Standard Error | 6062.102 |
| Observations | 1338 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 10 | 1.47E+11 | 1.47E+10 | 500.8107 | 0 |
| Residual | 1329 | 4.88E+10 | 36749084 |  |  |
| Total | 1339 | 1.96E+11 |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -12069.9 | 999.6492 | -12.0741 | 6.34E-32 | -14030.9 | -10108.8 | -14030.9 | -10108.8 |
| age | 256.8564 | 11.89885 | 21.58666 | 7.78E-89 | 233.5138 | 280.1989 | 233.5138 | 280.1989 |
| Dummy (Male) | 0 | 0 | 65535 | #NUM! | 0 | 0 | 0 | 0 |
| Dummy (Female) | 131.3144 | 332.9454 | 0.394402 | #NUM! | -521.842 | 784.4703 | -521.842 | 784.4703 |
| bmi | 339.1935 | 28.59947 | 11.86013 | 6.5E-31 | 283.0884 | 395.2985 | 283.0884 | 395.2985 |
| children | 475.5005 | 137.8041 | 3.450555 | 0.000577 | 205.1633 | 745.8378 | 205.1633 | 745.8378 |
| Dummy (Smoker) | 23848.53 | 413.1534 | 57.7232 | 0 | 23038.03 | 24659.04 | 23038.03 | 24659.04 |
| Dummy(Non-Smoker) | 0 | 0 | 65535 | #NUM! | 0 | 0 | 0 | 0 |
| Dummy (Northwest) | -352.964 | 476.2758 | -0.74109 | #NUM! | -1287.3 | 581.3704 | -1287.3 | 581.3704 |
| Dummy (Southeast) | -1035.02 | 478.6922 | -2.16219 | 0.030782 | -1974.1 | -95.9473 | -1974.1 | -95.9473 |
| Dummy (Southwest) | -960.051 | 477.933 | -2.00876 | 0.044765 | -1897.64 | -22.4656 | -1897.64 | -22.4656 |

|  |  |
| --- | --- |
| Infer | We can predict that our 75% of the data is accurate by checking R Square value. The independent variables like age, bmi, children, smoker, southeast region, southwest region are the significant predictor of dependent variable. The positive coefficients indicate as independent variables increases the dependent variable also increases and vice versa. The negative coefficient indicates as independent variables increases the dependent variable decreases and vice versa. |

|  |
| --- |
| The variables that decide the insurance charges are: |
| Age |
| Bmi |
| Children |
| Smoker |
| Southeast Region |
| Southwest Region |

**Thank You**