August 2022 | Accenture DS



**Project Report on Data Visualization using Power BI for FIFA 2018 data**

**Submitted by**

**Dev Tripathi**

**Contents**

**List of Figures**

[Figure 1: Sample of Insurance Dataset 2](#_Toc109471595)

[Figure 2: Feature Info and null values count of Insurance Dataset 2](#_Toc109471596)

[Figure 3: Statistical Description of numeric data type features 3](#_Toc109471597)

[Figure 4: Univariate Curves for different features in the dataset 4](#_Toc109471598)

[Figure 5: Pairplot for Insurance Dataset 7](#_Toc109471599)

[Figure 6: Correlation plot for different features 8](#_Toc109471600)

[Figure 7: Histogram of charges for different smoking status 9](#_Toc109471601)

[Figure 8: Distribution of BMI for different genders 10](#_Toc109471602)

[Figure 9: Count plot of smokers for different genders 11](#_Toc109471603)

[Figure 10: Distribution of BMI for women with 0,1 and 2 children 12](#_Toc109471604)

**List of Tables**

[Table 1: Data Dictionary 1](#_Toc109471605)

## Problem

### **Insurance Dataset Analysis**

### Information about dataset

Leveraging customer information is paramount for most businesses. In the case of an insurance company, attributes of customers like the ones mentioned below can be crucial in making business decisions. The data at hand contains medical costs of people characterized by certain attributes.

### Data Dictionary for Insurance Dataset

The dataset contains the following features and the information about these features is mentioned below:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| age | age of primary beneficiary |
| sex | insurance contractor gender, female, male |
| bmi | Body mass index, providing an understanding of body, weights that are relatively high or low relative to |
| height | objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9 |
| children | Number of children covered by health insurance / Number of dependents |
| smoker | Smoking |
| region | the beneficiary's residential area in the US, northeast, southeast, southwest, northwest. |
| charges | Individual medical costs billed by health insurance. |

Table : Data Dictionary

### EDA

#### Sample of Dataset

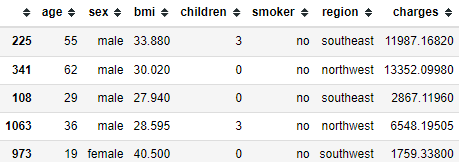


Figure : Sample of Insurance Dataset

#### Variable Information

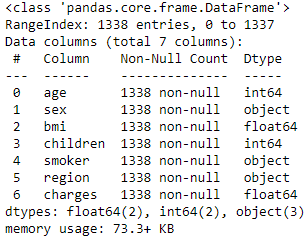
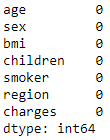
 

Figure : Feature Info and null values count of Insurance Dataset

The dataset contains a total of 1338 entries and 7 features. None of the variables contains null values. There are three columns i.e., ’sex’, ’smoker’, and ‘region’ which are object data types. All other variables are of numeric data type. Though the ‘children’ feature should also be treated as ‘object’ data type because the statistical analysis on this variable like mean or standard deviation will not be of any use, as fractional ‘children’ does not make any sense. Hence, the data type was corrected for further analysis.

#### Statistical Description

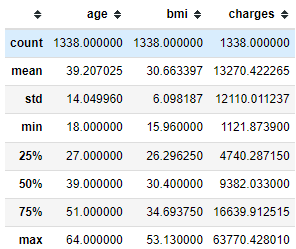


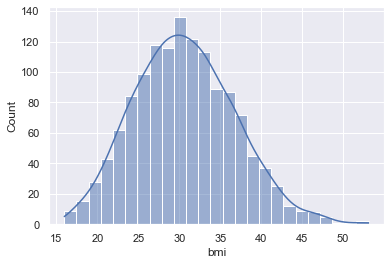
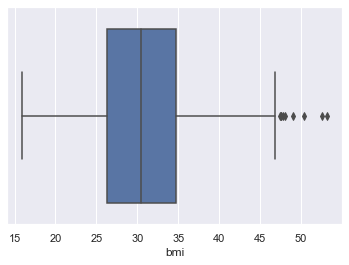
Figure : Statistical Description of numeric data type features

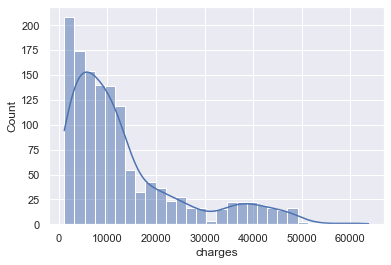
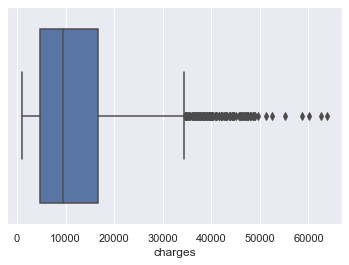
The statistical description contains mean, standard deviation and 5-point summary of the variables. Following inferences can be drawn from this description:

* Median value for `age` is 39 which is approximately equal to 39.2.
* Max and min values for `age` is within
* Mean value for `bmi` is 30.4 which is approximately equal to 30.66
* For `bmi` more than 50 percentile of the values lies within
* `charges` is having mean value as 13270.422 and median as 9382.033 which is far less than the mean value. Hence, the `charges` variable is right skewed

#### Univariate Analysis

Numeric Data type variable:

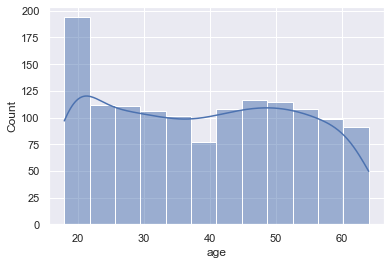
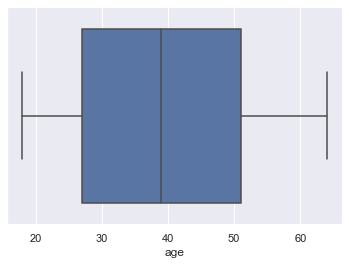
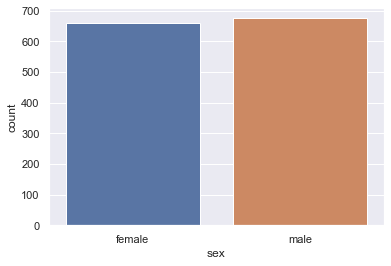
 

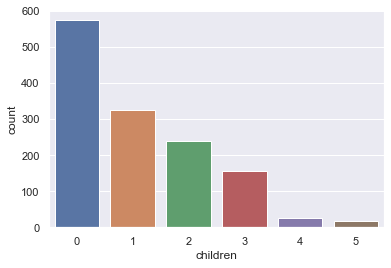
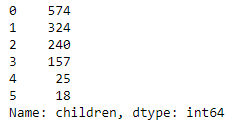
Figure : Univariate Curves for different features in the dataset

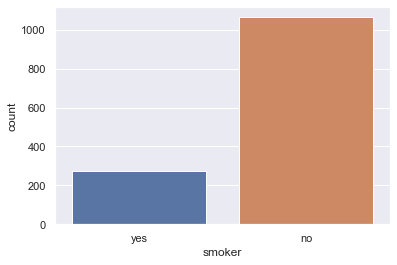
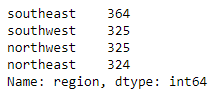
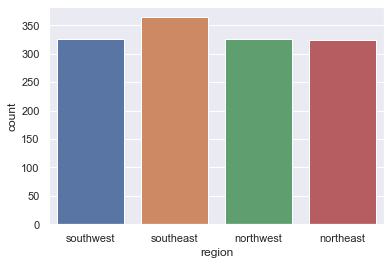
Now, conclusions drawn from this analysis can be briefly explained in the following points:

* The distribution of bmi looks fairly symmetrical and bell shaped
* bmi also contain some outliers.
* The distribution of charges is right skewed distribution denoting the higher medical costs are less frequently billed
* charges contain plentiful outliers
* The distribution of age is uniformly distributed except for the age group around 20 which means higher proportion of people are around 20 y/o.
* age contains no outliers and the boxplot seems to be symmetrical around the median value

Categorical Data Type Variable:





The following inferences can be drawn from the analysis:

* The proportion of ‘male’ and ‘female’ is fairly equal
* The number of ‘children’ shows a deceasing pattern. Number of people with ‘0’ ‘children’ are highest (574) and number of people with ‘5’ ‘children’ are lowest (18)
* Large proportion of people are ‘non-smokers’ (1064)
* The people from all the regions are fairly equal (around 325)

#### Bi-Variate Analysis

Now for bivariate analysis, we have created a pair plot that simply shows a scatterplot for all the pairs of features present in the dataset.

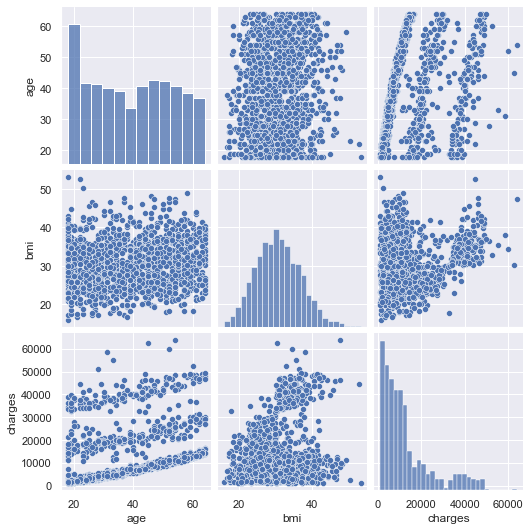


Figure : Pairplot for Insurance Dataset

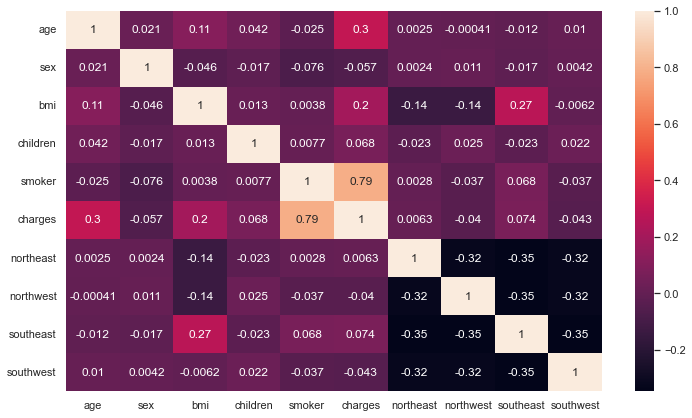


Figure : Correlation plot for different features

For the correlation plot encoding of categorical features is necessary. After encoding the categories in numeric form (please refer to the notebook for the labels/steps involved in encoding), the correlation plot was created.

Now, one can observe that the ‘smoker’ and ‘charges’ features have a high correlation between them. All other features and levels show insignificant correlation among them.

#### Hypothesis Testing

1. **Whether charges of people who smoke differ significantly from the people who don't?**

**Null Hypothesis :** charges for the smokers and non-smokers are equal  
**Alternate Hypothesis :** charges for smokers and non-smokers are not equal

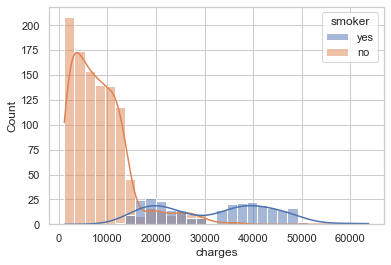


Figure : Histogram of charges for different smoking status

Since, the standard deviation of population is unknown, we can perform t-test to test the hypothesis. The t-statistic and p-value obtained is given below:



Since, the obtained p-value is less than the significance level of **5%** hence, we **can reject** the null hypothesis **()** and can say that the charges for smokers and non-smokers are **not equal.**

1. **Whether bmi of males differ significantly from that of females?**

**Null Hypothesis:** bmi of males is **equal** to bmi of females  
**Alternate Hypothesis:** bmi of males is **different** than that of females

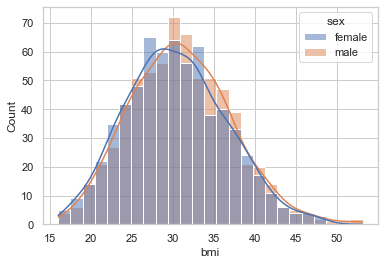
****

Figure : Distribution of BMI for different genders

Since, the standard deviation of population is unknown, we can perform t-test to test the hypothesis. The t-statistic and p-value obtained is given below:



Since, p-value is greater than the significance level of **5%** hence, we **cannot reject** the null hypothesis  and can say that the bmi for male and female does not differ significantly.

1. **Whether proportion of smokers is significantly different in different genders?**

**Null Hypothesis:** proportion of smokers across different genders **is same**  
**Alternate Hypothesis:** proportion of smokers across different genders **is different**

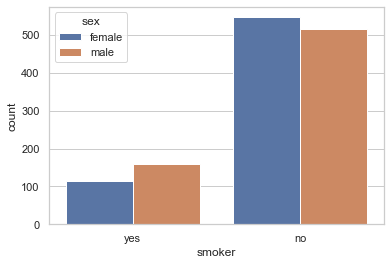


Figure : Count plot of smokers for different genders

Now, for testing proportions, a z-test was performed. The z-statistic and p-value obtained are given below:



Since the p\_value is less than significance level of **5%**, hence we **can reject** the null hypothesis  and can say that the proportion of smokers across different genders **is different**.

1. **Whether the distribution of bmi across women with no children, one child and two children, the same?**

**Null Hypothesis:** bmi across women with 0, 1 and 2 children **are same**  
**Alternate Hypothesis:** bmi across women with 0, 1 and 2 children **are not same**

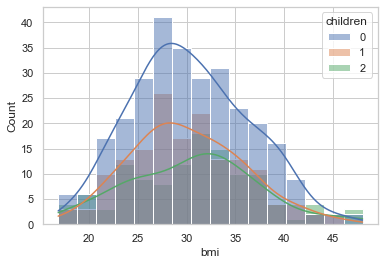
****

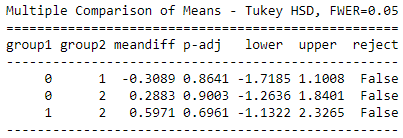
Figure : Distribution of BMI for women with 0,1 and 2 children

Since, the we have to test the means of more than two levels of a factor, we can perform one way ANOVA for testing the hypothesis. The ANOVA table for the same is given below:



Since p-value is greater than the significance level of **5%** the null hypothesis **cannot be rejected**. Hence, we can say the distribution of bmi is same across women with no children, one child and two children.

For testing the same we can also perform pairwise Tukey test. The results of Tukey test are following:

****

Since, the p-value for all the pairs is more than **significance level (0.05),** we can say that there is no statistically significant difference between the means.