**Interim Report of**

**HealthCare Data Analysis**



**HEALTH CARE INSURANCE - DATA ANALYSIS**

**Introduction:**

A health insurance policy is a legal agreement between an insurer and the insured. The policyholder pays a premium to the insurance company, and the insurer pays for the insured's medical expenses, such as hospitalization, daycare, post, and pre-hospitalization, etc. So, cost prediction is what this study is needed for.

**Problem Statement:**

Health care is a critical market segment. It is directly related to the individual's life. So, we must always be proactive in this arena. Money is important in this arena because treatment can be very expensive at times, and if an individual is not covered by insurance, he or she will be in a very difficult financial situation.

**Objective and Scope:**

The goal of this analysis is to check the quality of the data like the presence of missing/null values, duplicated rows , outliers etc. and do basic EDA (Exploratory Data Analysis)

**As we know that the target variable is insurance cost which is continuous in nature, this is the Regression Problem.** Regression is a supervised learning technique that helps in finding the correlation between variables and enables us to predict the continuous output variable based on one or more predictor variables.

**The steps that you are going to follow to achieve the objective are:**

* Problem Statement
* Data Acquisition (rows, columns) and Data Cleaning
* Visualizations

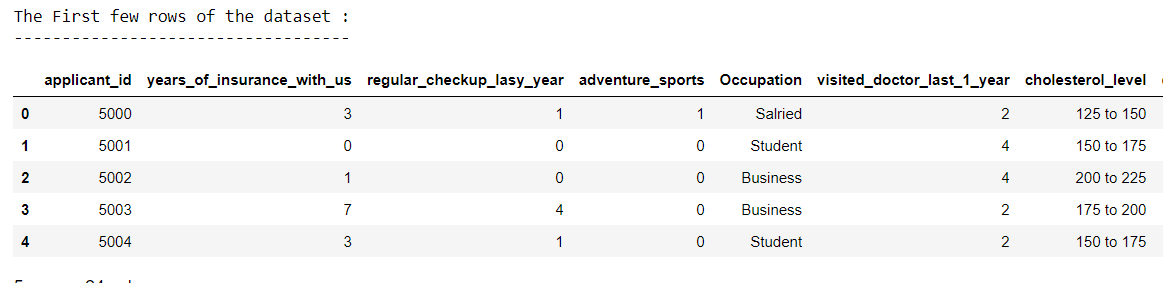
**Sample of the Dataset:**

|  |
| --- |
|  |

Table no 1: Sample of the dataset

**Exploratory Data Analysis**

**Importing the dataset: Head and Tail of the dataset**



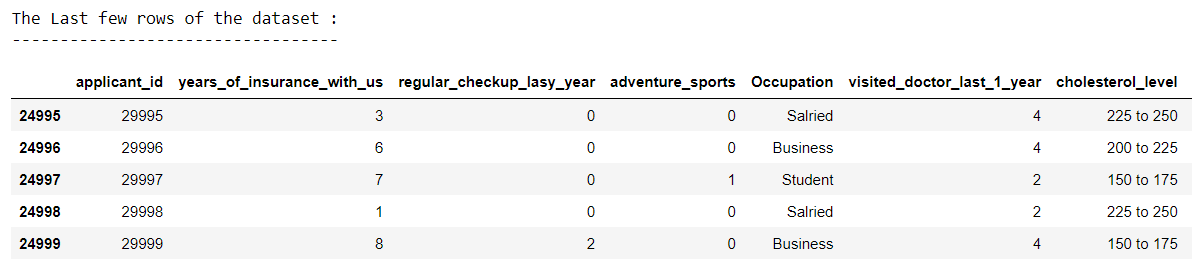


Table no 2: Head and Tail of the dataset

**Inferences from the above table:**

* The first five rows of the dataset which is the head of the dataset and the last five rows of the dataset which is the tail of the dataset. We have 25000 entries in the provided dataset.

**Structure of the Dataset:**

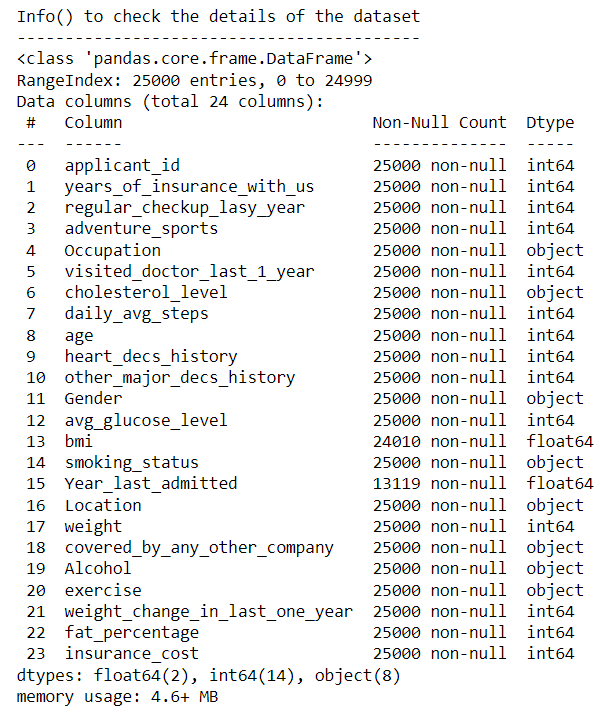


Table no 3: Structure of the dataset

**Inferences from the above table:**

* There are 24 different variables and 25000 records in the data set for each variable.
* The “bmi” and “Year\_last\_admitted” columns have some null values in the dataset. Excluding those null values, we have 24010 and 13119 entries respectively.
* The variables are of different data types which are of float, integer, and object.

**UNIVARIATE ANALYSIS:**

Univariate analysis entails the examination of a single variable. The primary goal is to summarize and identify trends in the data.

**Summary of the Dataset:**

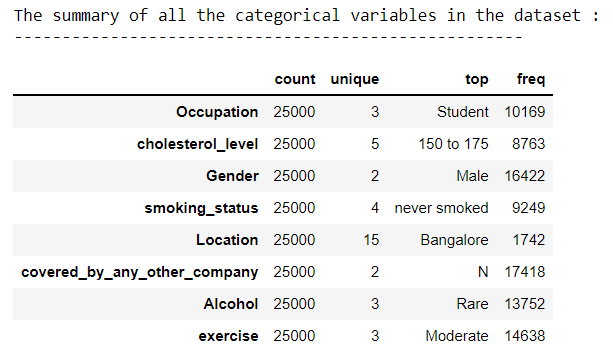


Table no 4: Summary of all categorical variables of the dataset

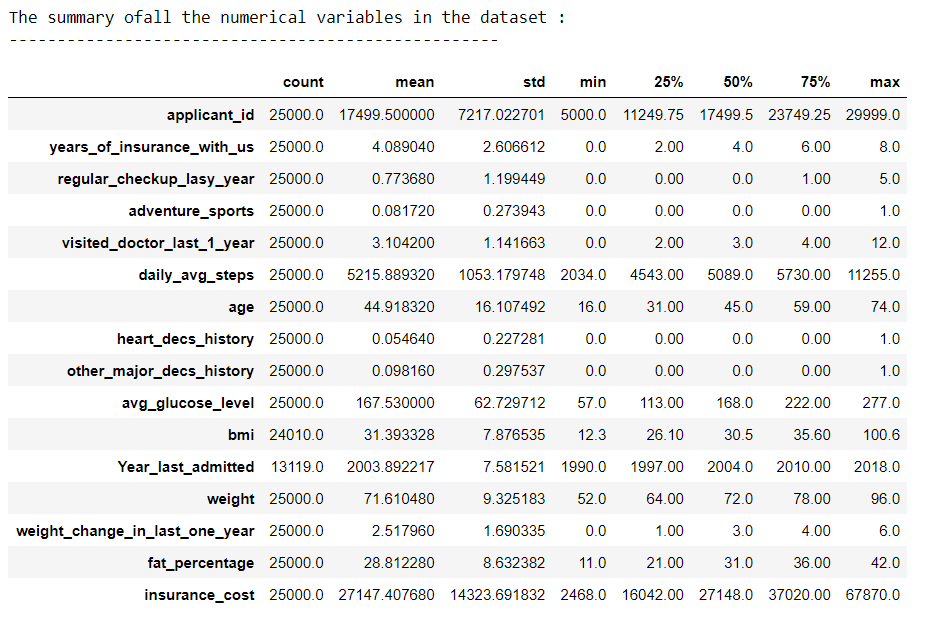


Table no 5: Summary of all numerical variables in the dataset

**Inferences from the above table:**

1. The Data consists of 25000 records with 24 variables (16 numerical and 8 categorical).
2. “Daily\_avg\_steps” has the highest standard deviation followed by “Year\_last\_admitted”.
3. The mean and median are nearly equal for all the variables in the dataset.

**Dimension of the dataset:**

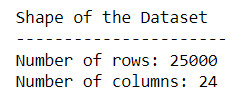


Table no 6: Shape of the dataset

**Duplicate rows check:**

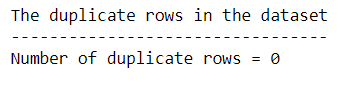
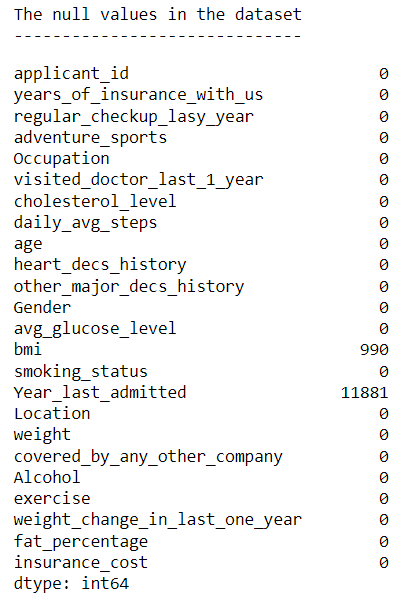


Table no 7: Duplicate values

**Observations**:

There are no duplicate rows found in the Dataset.

**Checking and imputing missing values in the dataset:**



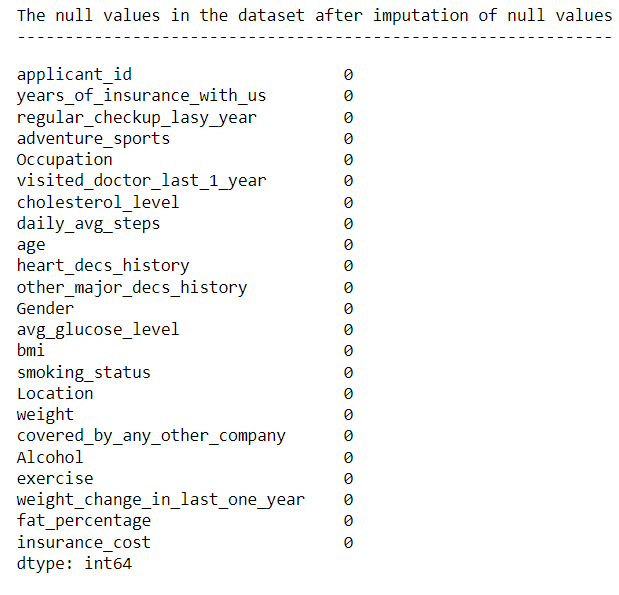


Table no 8: Tables showing the null values before and after imputation

**Inferences from the above table:**

1. There are missing values in the variable “bmi” and “Year\_last\_admitted”.
2. We impute the null values present in “bmi” with median value which is 30.5.
3. As the missing value in “Year\_last\_admitted” is almost equal to 50%, we drop the variable.
4. After imputation of missing values of bmi, the data is clear of null values.

**Inter Quartile Range of the variables in the dataset:**

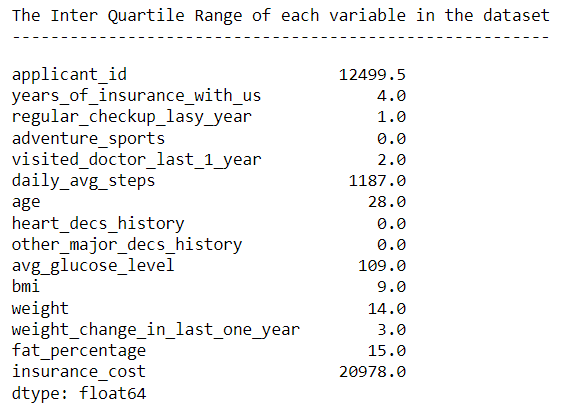


Table no 9: IQR of all the variables in the dataset

**Catplot of all the categorical variables in the dataset:**

|  |
| --- |
|  |

Fig no 1: Catplot of all the categorical variables in the dataset

**Inferences from the above figure:**

* In all the cities, count of customers taking up the health insurance is almost similar/equal which means people are much aware of safety and have secured themselves.
* From the cat plot of cholesterol level, people with cholesterol level 150-175 have taken up insurance mostly followed by 125-150.
* Customers who consume alcohol rarely, customers who have never smoked and work out moderately are seemed to take up insurance at the first place.
* Most of the males have taken up health care insurance as compared to females.
* Students are most likely to take up the insurance followed by people who are involved into business.

**Skewness of all the variables in the dataset:**

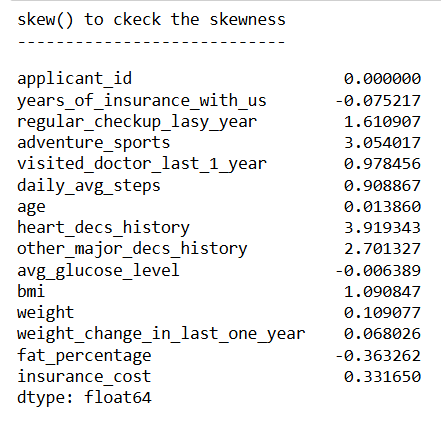


Table no 10: Skewness of all the variables in the dataset

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point, let us check it with the help of distplot. The value of heart\_decs\_history (3.91) is comparatively higher than other variables.

**Distplot to check distribution and density of variables in the dataset:**

|  |
| --- |
|  |
|  |

Fig no 2: Distplot of all the variables in the dataset

**Inferences from the above figure:**

* Daily\_avg\_steps, age and avg\_glucose\_level variables are normally distributed.
* All the other variables in the figure shows the right tailed distribution.

**Histogram of all the variables in the dataset:**

|  |
| --- |
|  |

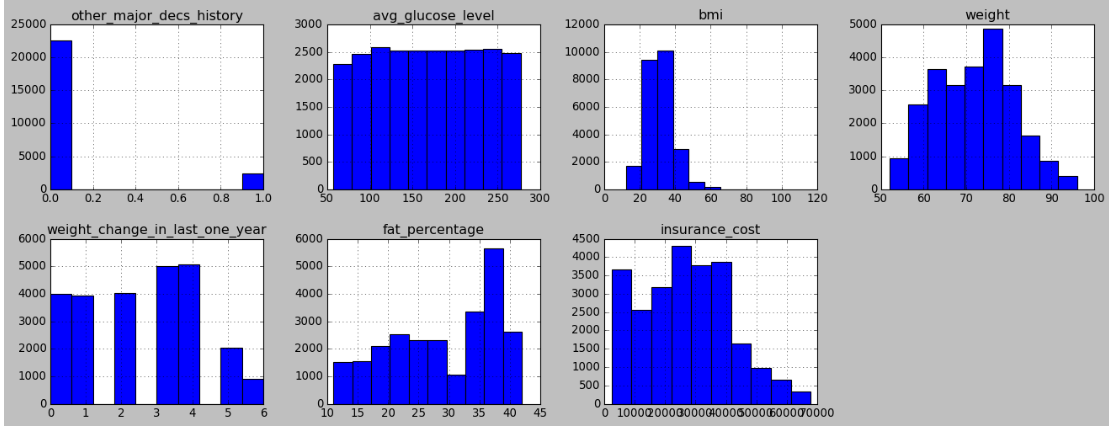


Fig no 3: Histogram of all the variables in the dataset

Fig no 3: Histogram of all the variables in the dataset

**Boxplot of all the variables in the dataset to check the outliers:**

|  |
| --- |
|  |

Fig no 4: Boxplot of all the variables in the dataset

**While reviewing the boxplots, we identified a few outliers in the variables "bmi" and "daily\_avg\_steps". Treating them will increase efficiency and improve analytical quality.**

|  |
| --- |
|  |

Fig no 5: Boxplot of bmi and daily\_avg\_steps after outlier treatment

**We can observe that there are no outliers in bmi and daily\_avg\_steps after the outlier treatment is done. We have kept outliers in few of the variables as they are considered to be valid outliers, which does not affect our model building.**

**MULTIVARIATE ANALYSIS**

**Correlation matrix of all the variables:**

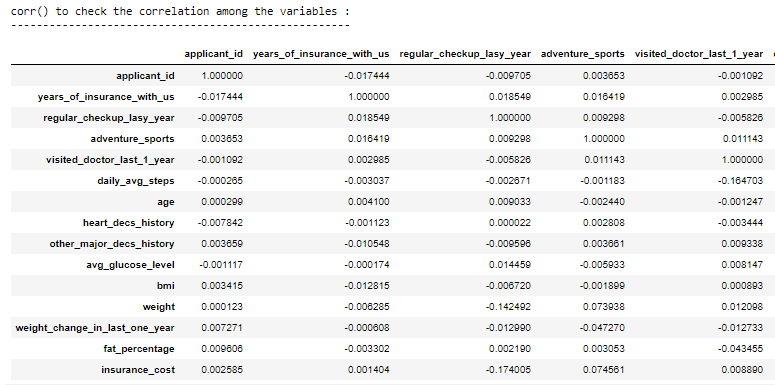


Table no 11: correlation matrix of all the variables in the dataset

**Inferences from the above correlation matrix and boxplot:**

* Weight variable and weight change in last one year variable shows more correlation with insurance cost as compared to other variables.

**Below Heatmap exhibits multicollinearity issue as significant number of high correlation variables pairs /features. When the statistical significance of independent variables is undermined Multicollinearity is observed.**

**Heat Map to check collinearity of Original Data:**

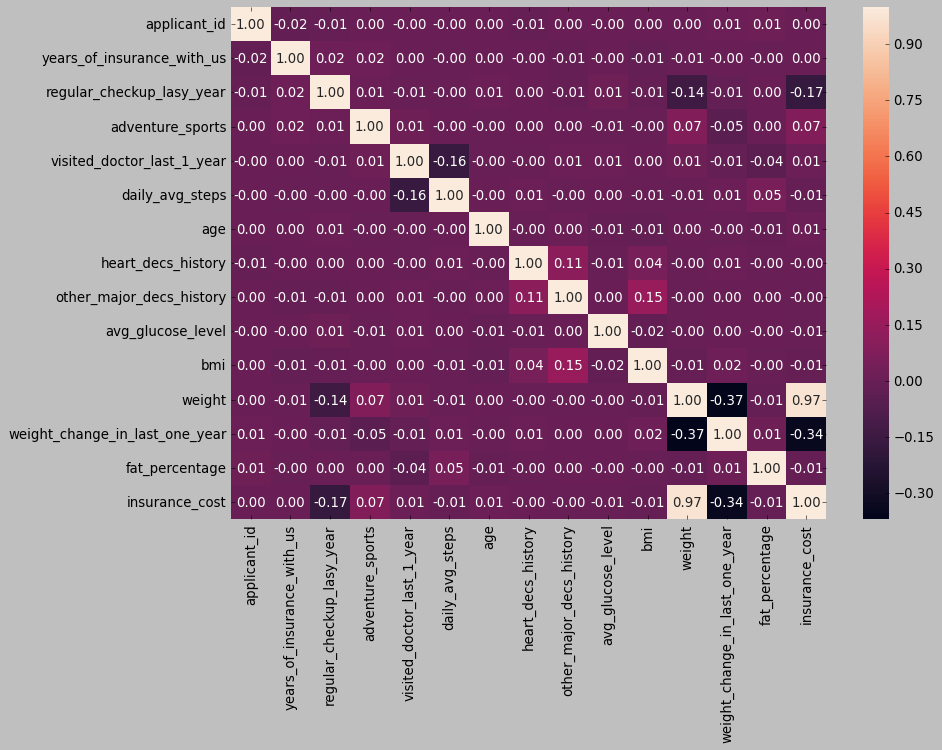
****

Fig no 6: Heat map of all the variables in the dataset

**Inferences from the above figure:**

* Here in this dataset, there is no much correlation among the variables.
* The weight variables show the highest correlation with the target variable which is 0.97, which means the weight variable contributes more in prediction of insurance cost followed by weight change in last one year which is 0.34.

**Pair plot to see the relationship of all variables among each other:**

|  |
| --- |
|  |

Fig no 7: Pair plot of the variables in the Dataset

**Insights from Exploratory Data Analysis**

1. Dataset has 25000 rows, and 24 columns dataset, there are 2 float variables, 14 int variables, and 8 object variables.
2. We see no significant difference between the mean and median, thus we infer that the data is normally distributed for the majority of the variables except daily average steps and BMI, which are right-scaled.
3. Years of insurance range from 0 to 8, and 2912 consumers had no prior insurance history with us, while the remaining people in the data have varying years of history, demonstrating their loyalty.
4. Regular check-ups last year demonstrated individuals' health consciousness on a scale of 0 to 5, with 15215 people who have not had a health check-up having the highest value. The remaining 9785 people had a health checkup last year and are aware of their health condition.
5. We have 2043 customers who participate in various adventure sports, which shows they have a high risk of injury and they are more likely to purchase health insurance.
6. Individuals' who have visited doctors in the previous year ranged from 0 to 12. There is a possibility of a direct association between the number of visits and the suffering from any health issue.
7. The acceptable level of cholesterol suggested by WHO is less than 200 mgs, yet we have 2054 customers with cholesterol levels ranging from 225 mgs to 250mgs, which means the person might have a high chance of suffering from any health issue.
8. We have both males and females in our data and owing to biological differences, they may have different health issues which leads to fluctuations in Insurance costs.
9. The WHO recommended glucose level is 135 mgs - 140 mgs, yet our research shows readings ranging from 222mgs -277mgs. This demonstrates the significant likelihood of diabetes and other health problems.
10. "Smoking and drinking are harmful to one's health." We have individuals who smoke and drink alcohol daily, thus we can infer that they may have poor health, resulting in high insurance costs.
11. Weight fluctuation during the past year ranges from 0 to 6 kg. More weight fluctuation can be an indication of both good and bad health, so it is hard to ascertain the customer's health.
12. High body fat can raise the risk of diseases including diabetes and cardiovascular disease, as well as lethargy and inactivity.

**More fat percentage** → more weight → more BMI → **High chances of bad health**

1. Only 5.4% of people have a history of heart disease, and they are more likely to incur higher insurance expenses.
2. From the multivariate analysis, we can conclude that Weight has a positive correlation with insurance cost, which is 0.97. Followed by, Weight change in the last one which is 0.34.
3. The WHO recommended BMI is below 40, but even values ranging between 60-100 indicate severe obesity. Thus, it requires an outlier treatment.

**Recommendations for the business:**

1. Many industrial zones exist in Tier-1 cities, creating a variety of pollution. This has an impact on the health of those living in Tier-1 cities. As a result, the organization should concentrate on Tier-1 customers in predicting the insurance cost.
2. The organization should pay special attention to individuals in earning roles because they make the decisions and are more likely to purchase insurance.
3. The insurance organization should organize numerous awareness campaigns to help individuals understand the benefits of having insurance, which could lead to an increase in customer numbers.
4. Additional concentration should be given on tax paying individuals because they have the benefit of tax redemption.
5. Following up with existing customers and delivering beneficial updates will increase their interest in paying their premium on time.
6. The company should focus on customer experience, especially during the claims process. This will build trust and a favorable influence on the organization on behalf of the consumer, resulting in referrals, which will eventually boost the volume of customers.
7. We have a few null values in the data, which may influence the cost prediction's efficiency. It is recommended that the company acquire correct personal and medical information from all its customers, which will aid in better data analysis and interpretation of insurance costs. For data collection constraints, we recommend that the organization make the filing of a health check-up report a mandatory requirement for determining the cost of insurance.