HEALTHCARE PROJECT

Submitted by,

Jiya Jacob

PGP DSBA July B Batch

Date: 29/05/2022

|  |  |  |
| --- | --- | --- |
| **Sl.no** | **Questions** | **Page no** |
| **1)**  a. | **Introduction of the business problem**  Defining problem statement | 5 |
| b. | Need of the study/project | 5 |
| c. | Understanding business/social opportunity | 5 |
| **2)**  a. | **Data Report**  Understanding how data was collected in terms of time, frequency and methodology | 5 |
| b. | Visual inspection of data (rows, columns, descriptive details) | 5 |
| c. | Understanding of attributes (variable info, renaming if required) | 7 |
| **3)**  a. | **Exploratory data analysis**   Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) | 9 |
| b. | Bivariate analysis (relationship between different variables, correlations) | 21 |
| c. | Removal of unwanted variables (if applicable) | 32 |
| d. | Missing Value treatment (if applicable) | 33 |
| e. | Outlier treatment (if required) | 34 |
| f. | Variable transformation (if applicable) | 36 |
| g. | Addition of new variables (if required) | 36 |
| **4)**  a. | **Business insights from EDA**  Is the data unbalanced? If so, what can be done? Please explain in the context of the business | 37 |
| b. | Any business insights using clustering (if applicable) | 42 |
| c. | Any other business insights | 42 |

List of figures

|  |  |  |
| --- | --- | --- |
| Sl.no | Figure name | Page no |
| 1. | Univariate analysis of continuous columns using histogram | 10-15 |
| 2. | Univariate analysis of categorical columns using count plot | 16-21 |
| 3. | Bivariate analysis of continuous columns using scatter plot | 22-26 |
| 4. | Bivariate analysis of categorical columns using box plot | 27-32 |

List of tables

|  |  |  |
| --- | --- | --- |
| Sl.no | Table name | Page no |
| 1 | Table 1: Head of the given dataset | 5 |
| 2 | Table 2: Summary of the given data | 7 |
| 3 | Table 3: Information of the given variables | 8 |
| 4 | Table 4: Information about the given variables | 9 |
| 5 | Table 5: Head of the dataset after removal of unwanted variables | 32 |
| 6 | Table 6: Transpose of the head of the dataset after encoding the categorical variables | 37 |
| 7 | Table 7: head of the data set after clustering | 43 |
| 8 | Table 8: description of the newly formed clusters | 44 |

1. **Introduction of the business problem**
2. **Defining problem statement**

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance, then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

1. **Need of the study/project**

The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual. You have to use the health and habit related parameters for the estimated cost of insurance.

1. **Understanding business/social opportunity**

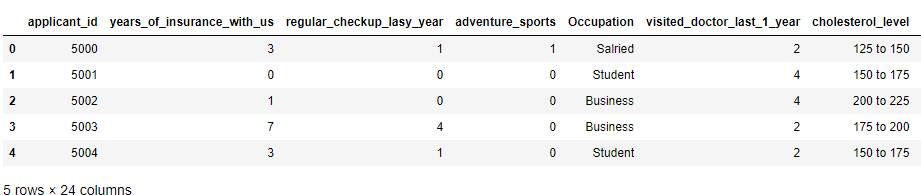
* To reduce risk of our health.
* Create awareness

1. **Data Report**
2. **Understanding how data was collected in terms of time, frequency and methodology**

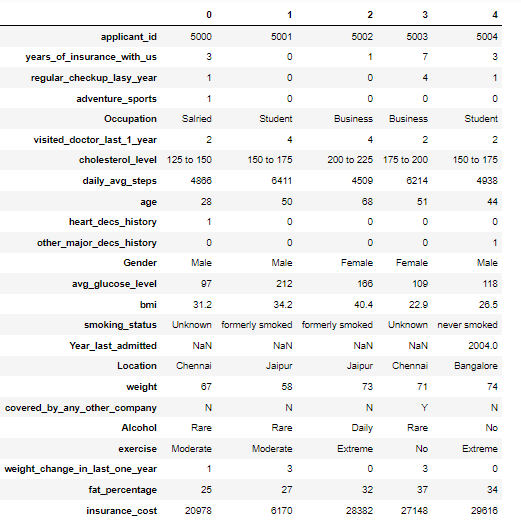
The data was collected with the cooperation of several health insurance companies

1. **Visual inspection of data (rows, columns, descriptive details)**

* **Head of the dataset**



**Table 1: Head of the given dataset**



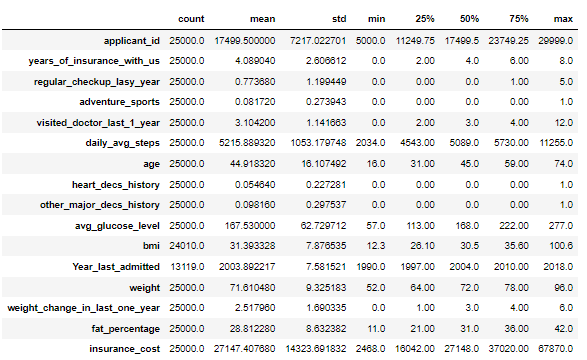
**Head of the given dataset in the transpose view**

* **Shape of the given dataset**



The given data set has 25000 rows and 24 columns. The given data set has no duplicated values.

* **Description of the given dataset**



**Table 2: Summary of the given data.**

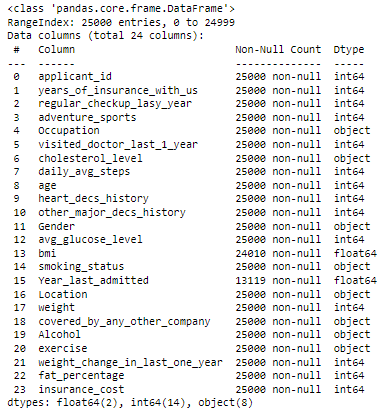
From the summary of the dataset, we can see that mean of the insurance\_cost is the highest, followed by Year\_last\_admitted, while the mean for heart\_decs\_history is least among all. The variables adventure\_sports and other\_major\_decs\_history share almost the similar mean whereas visited\_doctor\_last\_1\_year has slightly lower mean than years\_of\_ insurance with us. Similarly, insurance\_cost has the highest standard deviation whereas heart\_decs\_history has the lowest standard deviation, followed by other\_major\_decs\_history that has slightly higher standard deviation. The median age is 45 and the median insurance\_cost is 271480. The median average\_glucoe\_level is 168 and the median bmi is 30.5. The maximum daily\_average\_steps is 11255, the minimum being 2034 and the median being 5089.

1. **Understanding of attributes (variable info, renaming if required)**

|  |  |  |
| --- | --- | --- |
| **SL.NO** | **Variable** | Definition |
| 1. | applicant\_id | Unique identity of the person |
| 2. | years\_of\_insurance\_with\_u s | Count of years that person has insurance deal with the company |
| 3. | regular\_checkup\_lasy\_year | How many times the person has gone for regular check up to the hospital during the last year |
| 4. | adventure\_sports | Whether the person is active in adventure sports, which may increase the chances of his insurance claims |
| 5. | Occupation | Occupation of the person which indicates how much amount he can claim |
| 6. | visited\_doctor\_last\_1\_year | Number of times the person has visited the doctor last year which increases his claims |
| 7. | cholesterol\_level | Indicates the level of cholesterol; high cholesterol indicates high risk and high insurance amount |
| 8. | daily\_avg\_steps | Number of steps the person takes which indicates his fitness |
| 9. | age | Age of the person |
| 10. | heart\_decs\_history | Whether the person has the history of heart disease |
| 11. | other\_major\_decs\_history | Whether the person has the history of any other major diseases |
| 12. | Gender | Gender of the person |
| 13. | avg\_glucose\_level | Average glucose level of the person which indicate the health risk of the person |
| 14. | bmi | Bmi of the person |
| 15. | smoking\_status | Whether the person possess the smoking habit or not |
| 16. | Year\_last\_admitted | Number of times the person was last admitted in the hospital |
| 17. | Location | Place to which the person belongs |
| 18. | weight | Weight of the person |
| 19. | covered\_by\_any\_other\_company | Whether the insurance is covered by any other company |
| 20. | Alcohol | Whether the person comsumes alcohol or not |
| 21. | Exercise | How frequently the person does exercise.This indicates his or her fitness |
| 22. | weight\_change\_in\_last\_one \_year | Change in the weight of the person in last one year. |
| 23. | fat\_percentage | The percentage of fat in the body of the person |
| 24. | insurance\_cost | The total insurance cost considering all the above parameters |

**Table 3: Information of the given variables.**

1. **Information about the variables**



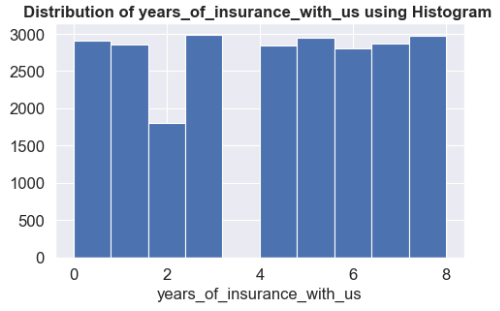
**Table 4: Information about the given variables.**

From the above table, we see that out of total 26 variables, 14 are of integer data types, 8 are of object data types and 2 are of float data types. Out of all these 14 variables, two variables, namely Year\_last\_admitted and bmi has null values present in them.

1. **Exploratory data analysis**
2. **Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)**

* **Continuous columns:**

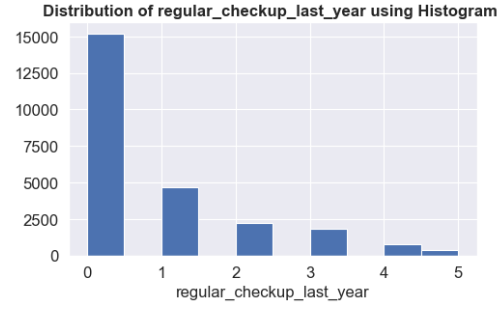
**1.years\_of\_insurance\_with\_us**



**Fig 4: Histogram for years\_of\_insurance\_with\_us.**

From the univariate analysis using histogram, **years\_of\_insurance\_with\_us** (in years) is plotted along the x-axis. We see that the 2990 people have 2 years association with the company, thereby, making them the largest group, followed by 8 years association by 2970 people. Only 1808 people belong to the group of 2 years association, thereby making them the least group.

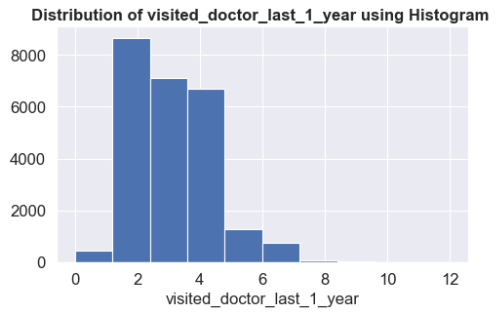
**2.** **regular\_checkup\_last\_year**



**Fig 5: Histogram for regular\_checkup\_last\_year.**

From the univariate analysis using histogram, **regular\_checkup\_last\_year** is plotted along the x-axis. We see that the 15215 people had no check-up last year, thereby, making them the largest group, followed by only one check-up last year by 4644 people. Only 348 people had done 5 check-ups last year, thereby making them the least group. We can consider that people who are doing regular check-ups are at low risk, while people who do not do check ups are at high risk.

**3.** **visited\_doctor\_last\_1\_year**

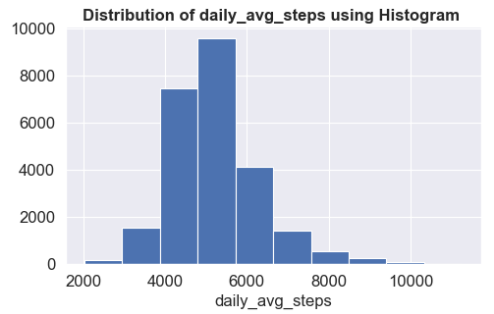


**Fig 6: Histogram for visited\_doctor\_last\_1\_year.**

From the univariate analysis using histogram, **visited\_doctor\_last\_1\_year**

is plotted along the x-axis. We see that the 8669 people have visited doctor 2 times last year, thereby, making them the largest group, followed by 3 times visit by 7094 people. Only 1 person has visited the doctor 12 times and the other person never visited the doctor at all the last year. More visits to the doctor indicate more health risk.

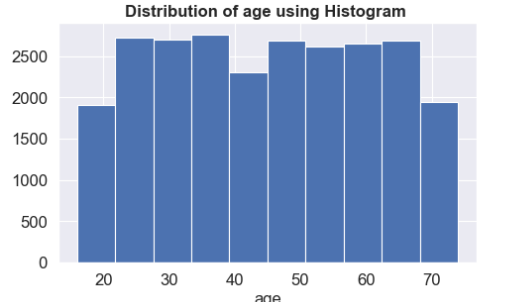
**4. daily\_avg\_steps**



**Fig 7: Histogram for daily\_avg\_steps.**

From the univariate analysis using histogram, **daily\_avg\_steps** is plotted along the x-axis. We see that most people have taken daily average steps between 5000-6000 steps, followed by those who take a daily average of 4000-5000 steps. Least people have taken steps between 9500-1000 daily steps on average.

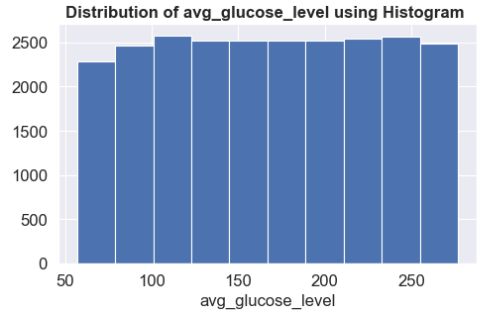
**5. age**



**Fig 8: Histogram for age.**

From the univariate analysis using histogram, **age** is plotted along the x-axis. We see that data of people of age group between 5-75 is present.Almost 59 unique ages are present in the group.

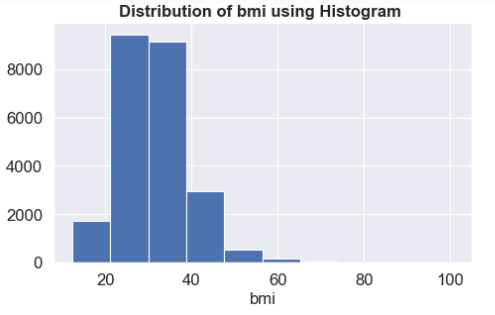
**6. avg\_glucose\_level**



**Fig 9: Histogram for avg\_glucose\_level.**

From the univariate analysis using histogram, **avg\_glucose\_level** is plotted along the x-axis. Average glucose level indicates the sugar level of the person thereby implies whether the person is diabetic or not. Almost 221 unique glucose levels are present in the given dataset.

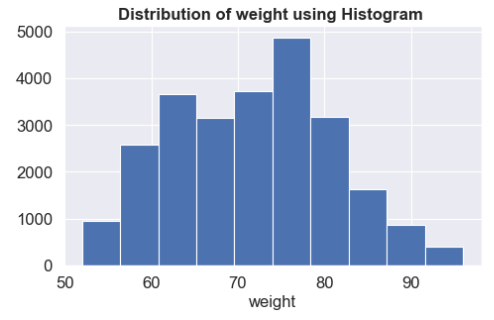
**7. bmi**



**Fig 10: Histogram for bmi.**

From the univariate analysis using histogram, **bmi** is plotted along the x-axis. Almost 465 unique bmis are recorded in the given dataset, wherein almost 189 people possess the bmi of 29.7.BMI stands for body mass index which is calculated by taking our weight (in kg) and height (in m) into consideration. A healthy bmi is in the range of 18.5 to 24.9 .By looking at the dataset , we see that there are many people who has unhealthy bmi and are at high risk.

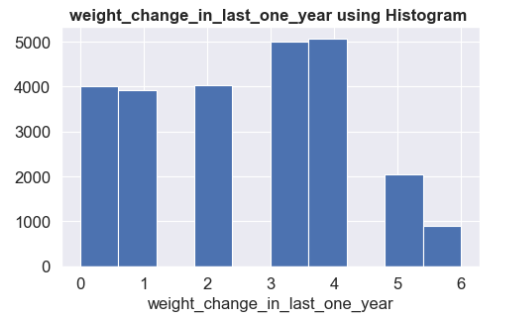
**8. weight**



**Fig 11: Histogram for weight.**

From the univariate analysis using histogram, **weight** is plotted along the x-axis. We see that the 2990 people have 2 years association with the company, thereby, making them the largest group, followed by 8 years association by 2970 people. Only 1808 people belong to the group of 2 years association, thereby making them the least group.

**9. weight\_change\_in\_last\_one\_year**

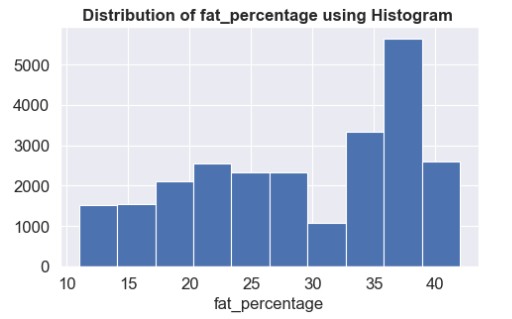


**Fig 12: Histogram for weight\_change\_in\_last\_one\_year.**

From the univariate analysis using histogram, **weight\_change\_in\_last\_one\_year**

is plotted along the x-axis. We see that the 5076 people have seen 4 times variation in their weight in the last one year, thereby, making them the largest group, followed by 3 times weight variation observed in 5006 people. Only 908 people belong to the group who have observed 6 times weight variation in the last year, thereby making them the least group.

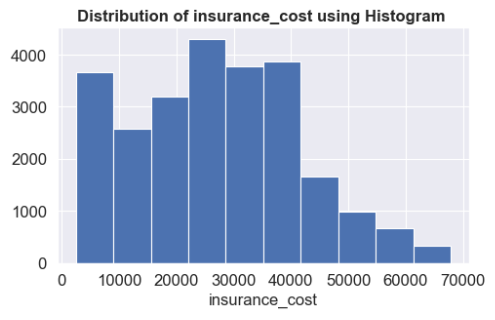
**10. fat\_percentage**



**Fig 13: Histogram for fat\_percentage.**

From the univariate analysis using histogram, **fat\_percentage** is plotted along the x-axis. We see that the 2908 people have seen 36% fat in their body, thereby, making them the largest group, followed by 33% fat noted in 1828 people. Only 65 people belong to the group who have 17% fat in their body, thereby making them the least group.

**11. insurance\_cost**

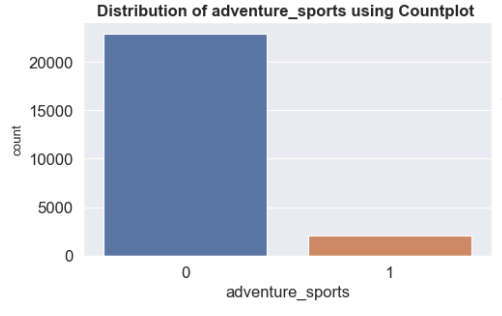


**Fig 14: Histogram for insurance\_cost.**

From the univariate analysis using histogram, **insurance\_cost** is plotted along the x-axis. We see that the 1214 people have their total cost as 7404, thereby, making them the largest group, followed by the insurance cost of 38254 by 977 people. Only 12 people belong to the group who have insured for the cost of 2468, thereby making them the least group.

**Categorical varaibles**

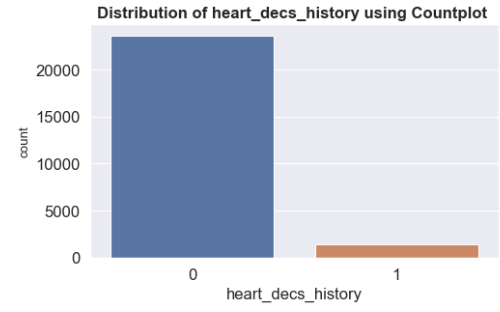
1. **adventure\_sports**



**Fig 15: Count plot for adventure\_sports**

In the univariate analysis using Count plot, whether the person has participated in adventure sports is plotted. From the above count plot, we can see that 22957 have no involvement in such activities (low risk) while 2043 have involvement (high risk).

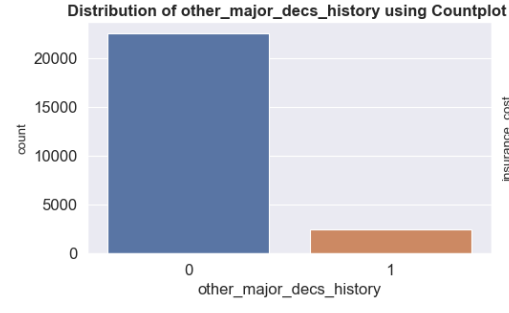
1. **heart\_decs\_history**



**Fig 16: Count plot for heart\_decs\_history**

In the univariate analysis using Count plot, history of heart disease of the people is plotted. From the above count plot, we can see that 23634 people have no such history (low risk) while 1366 people have heart disease history (high risk).

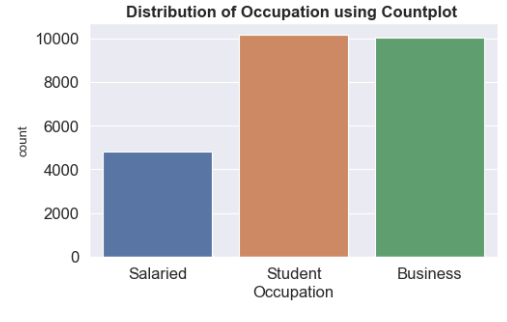
1. **other\_major\_decs\_history**



**Fig 17: Count plot for other\_major\_decs\_history**

In the univariate analysis using Count plot, whether the person has any other major disease is plotted. From the above count plot, we can see that 22546 have no such history (low risk) while 2054 have some other health issues (high risk).

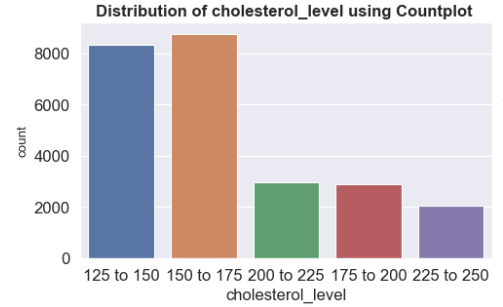
1. **Occupation**



**Fig 18: Count plot for Occupation**

In the univariate analysis using Count plot, occupation of the person is plotted. From the above count plot, we can see that 10169 are students, 10020 are having their own business commitments while 4811 people are employees.

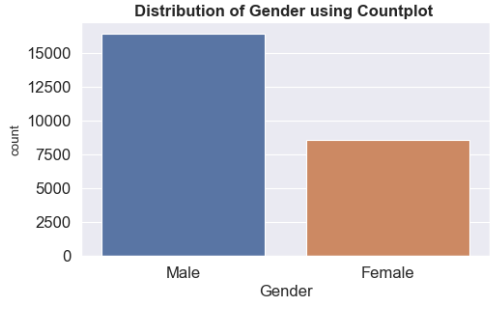
1. **cholesterol\_level**



**Fig 19: Count plot for cholesterol\_level**

In the univariate analysis using Count plot, cholesterol\_level of the person is plotted. From the above count plot, we can see that 8763 have their cholesterol level from 150-175, 8339 have their cholesterol level from 125-150, 2963 have their cholesterol level from 200-225(highly risky group), 2881 have their cholesterol level from 175-200 and 2054 have their cholesterol level from 225-250(the least group).

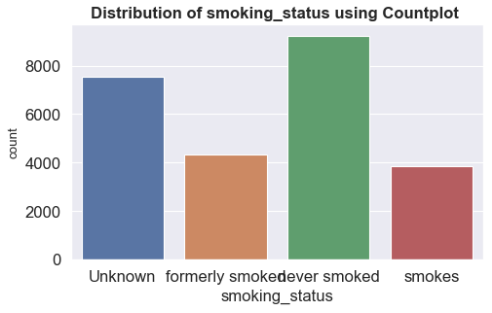
1. **Gender**



**Fig 20: Count plot for Gender**

In the univariate analysis using Count plot, gender of the person is plotted. From the above count plot, we can see that 16422 are male and 8578 are female. Hence, we see that female population is just the half of the male population, indicating an imbalance data.

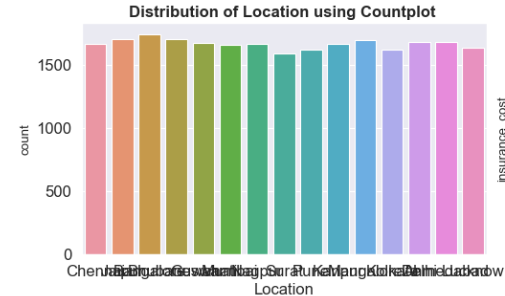
1. **smoking\_status**



**Fig 21: Count plot for smoking\_status**

In the univariate analysis using Count plot, smoking status of the person is plotted. From the above count plot, we can see that 9249 people never smoked at all(majority), while the smoking status of 7555 people are unknown.4329 people were formerly smokers and 3867 are active smokers(minority).

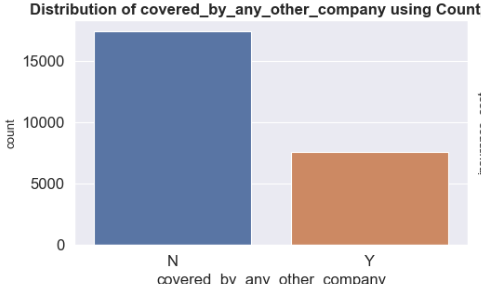
1. **Location**



**Fig 22: Count plot for Location**

In the univariate analysis using Count plot, location of the person is plotted. From the above count plot, we can see that 1742 people reside in Bangalore(largest group) and 1706 people reside in Jaipur. Only 1589 people reside in Surat, thus forming the smallest group.

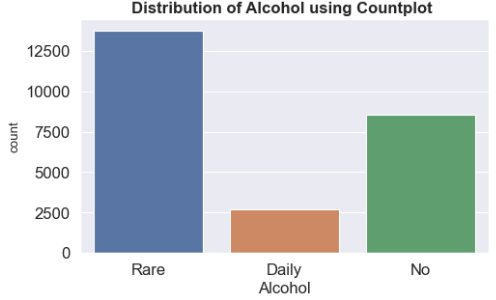
1. **covered\_by\_any\_other\_company**



**Fig 23: Count plot for covered\_by\_any\_other\_company**

In the univariate analysis using Count plot, whether the health insurance of the person is covered by any company is plotted. From the above count plot, we can see that for 17418 people, their insurance is not covered by any company, while for 7582 people their insurance is covered by some other company. Hence, we see that people whose insurance is covered by some other company is just the half of the population whose insurance is not covered by any company , indicating an imbalance data.

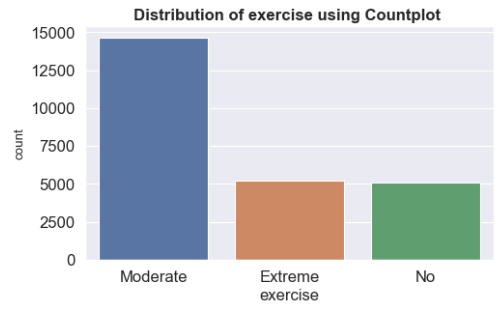
1. **Alcohol**



**Fig 24: Count plot for Alcohol**

In the univariate analysis using Count plot, nature of alcohol intake of the person is plotted. From the above count plot, we can see that 13752 people rarely intake alcohol(majority),while 2707 people are regular drinkers(minority).8541 people have never taken alcohol at all.

1. **exercise**



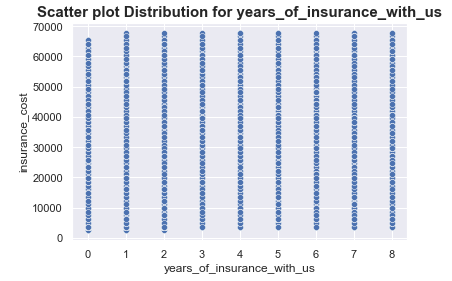
**Fig 25: Count plot for exercise**

In the univariate analysis using Count plot, exercise nature of the person is plotted. From the above count plot, we can see that 14368 people practice moderate exercise, whereas 5114 people have never done exercise at all.5248 people are the ones who does extreme exercise. From the above graph, it is quite evident that this is also an imbalanced column.

**Bivariate analysis (relationship between different variables, correlations)**

**Continuous columns**

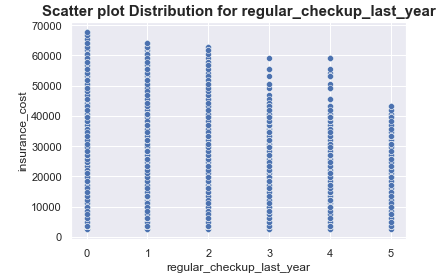
1. **years\_of\_insurance\_with\_us**



**Fig 26: Scatter plot for years\_of\_insurance\_with\_us**

From the above scatter plot, we see that insurance cost has a significant impact on number of years the person has insurance association with the company.

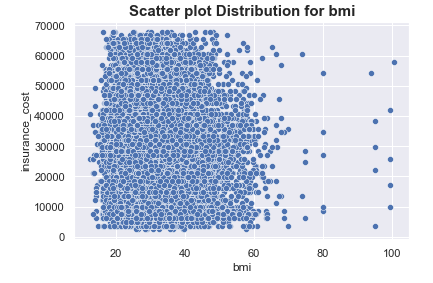
1. **regular\_checkup\_last\_year**



**Fig 27: Scatter plot for regular\_checkup\_last\_year**

From the above scatter plot, we see that insurance cost has a significant impact on number of times the person had regular check-ups last year.

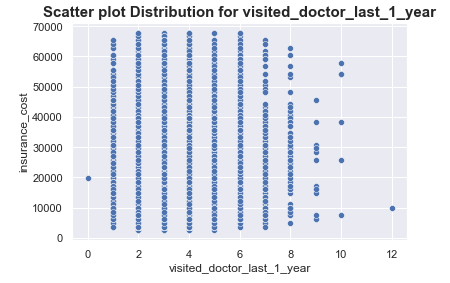
1. **bmi**



**Fig 28: Scatter plot for bmi**

From the above scatter plot, we see that insurance cost has a significant impact on bmi of the person.

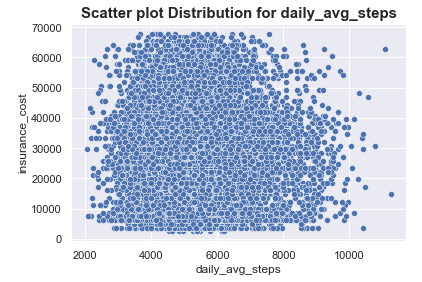
1. **visited\_doctor\_last\_1\_year**



**Fig 29: Scatter plot for visited\_doctor\_last\_1\_year**

From the above scatter plot, we see that insurance cost has a significant impact on number of times the person visited the doctor last year.

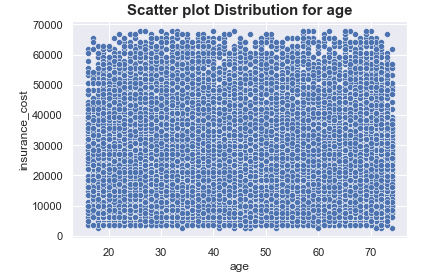
1. **daily\_avg\_steps**



**Fig 30: Scatter plot for daily\_avg\_steps**

From the above scatter plot, we see that insurance cost has a significant impact on daily average steps taken by the person.

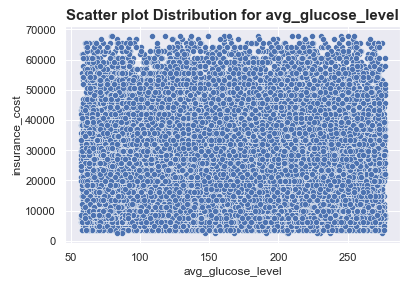
**6. age**



**Fig 31: Scatter plot for age**

From the above scatter plot, we see that insurance cost has a significant impact on age of the person.

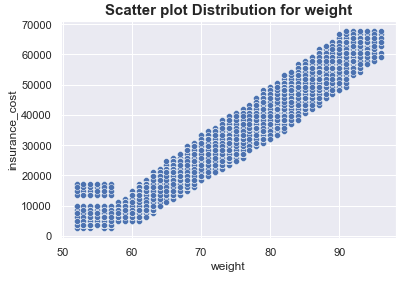
**7. avg\_glucose\_level**



**Fig 32: Scatter plot for avg\_glucose\_level**

From the above scatter plot, we see that insurance cost has a significant impact on average glucose level of the person.

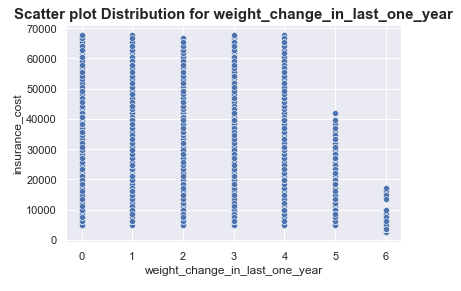
**8. weight**



**Fig 33: Scatter plot for weight**

From the above scatter plot, we see that insurance cost has a significant impact on weight of the person.

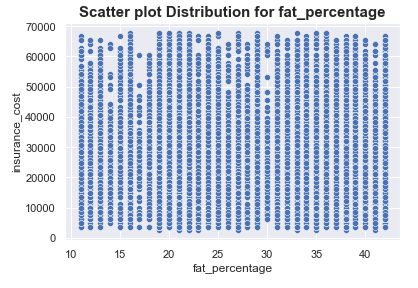
**9. weight\_change\_in\_last\_one\_year**



**Fig 34: Scatter plot for weight\_change\_in\_last\_one\_year**

From the above scatter plot, we see that insurance cost has a significant impact on weight change of the person in last one year.

**10. fat\_percentage**

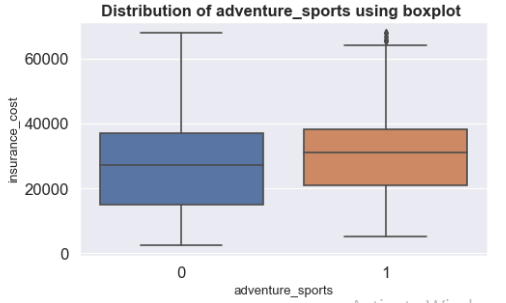


**Fig 35: Scatter plot for fat\_percentage**

From the above scatter plot, we see that insurance cost has a significant impact on fat percentage of the person.

**Categorical columns**

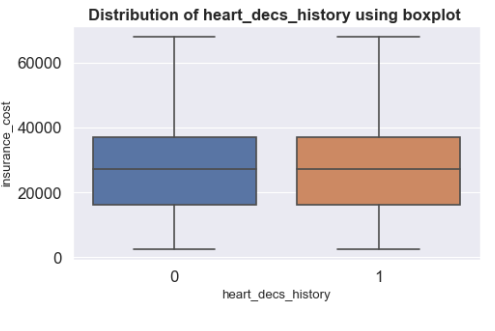
1. **adventure\_sports**



**Fig 36: Box plot for adventure\_sports**

From the above boxplot, we see that insurance cost has an impact on the involvement of the person in adventure sport since the median varies across the two boxes. Median of the insurance cost of the person involved in adventure is slightly higher than the one has no involvement in such activities.

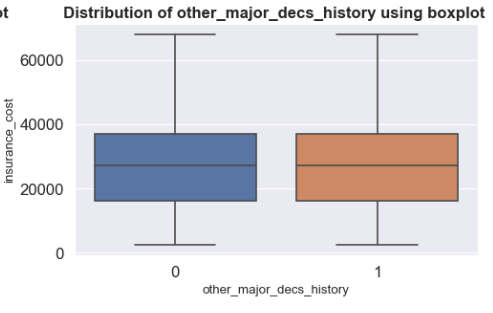
**2. heart\_decs\_history**



**Fig 37: Box plot for heart\_decs\_history**

From the above boxplot, we see that insurance cost has no impact on the heart disease history of the person since the median across the two boxes remains at the same level.

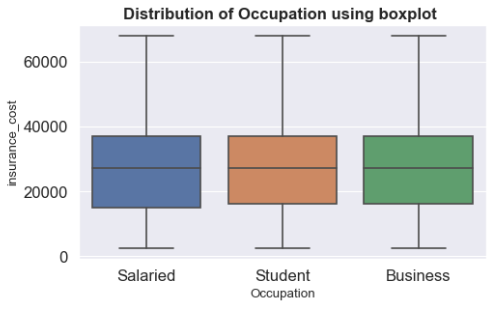
**3. other\_major\_decs\_history**



**Fig 38: Box plot for other\_major\_decs\_history**

From the above boxplot, we see that insurance cost has no impact on the history of other major diseases of the person since the median across the two boxes remains at the same level.

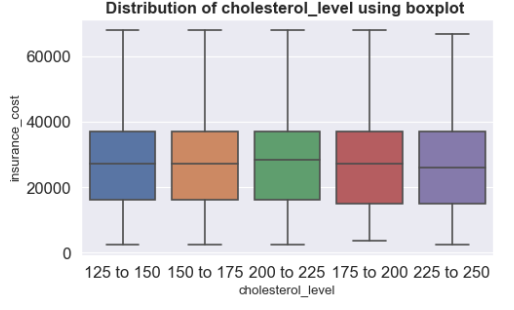
**4. Occupation**



**Fig 39: Box plot for Occupation**

From the above boxplot, we see that insurance cost has no impact on the occupation of the person since the median across all the three boxes remains at the same level.

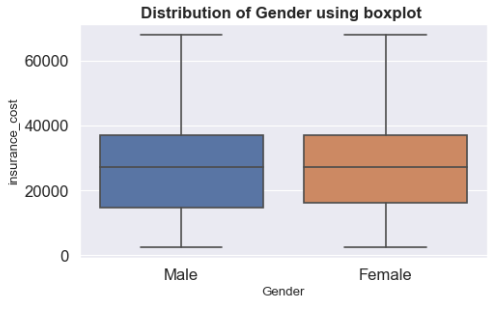
**5. cholesterol\_level**



**Fig 40: Box plot for cholesterol\_level**

From the above boxplot, we see that insurance cost has a very small impact on the cholesterol level of the person. The median of the insurance cost for those whose cholesterol level is between 200-225 is slightly higher than others, thus making them a risky group, whereas for other groups the insurance cost has no impact.

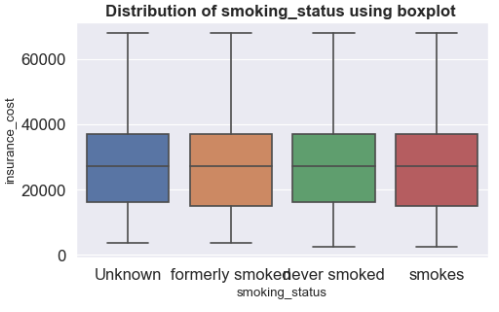
1. **Gender**



**Fig 41: Box plot for Gender**

From the above boxplot, we see that insurance cost has no impact on the gender of the person since the median across all the two boxes remains at the same level.

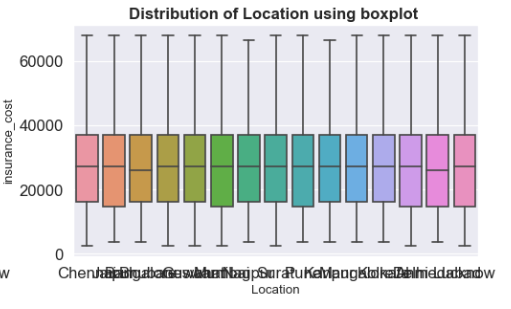
**7. smoking\_status**



**Fig 42: Box plot for smoking\_status**

From the above boxplot, we see that insurance cost has no impact on the smoking status of the person since the median across all the four boxes remains at the same level.

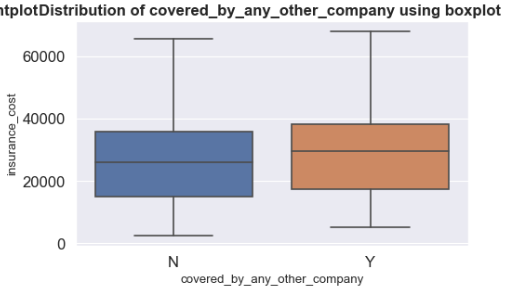
**8. Location**



**Fig 43: Box plot for Location**

From the above boxplot, we see that insurance cost has a very small impact on the location of the person. The median of the insurance cost for some locations is slightly higher than others.

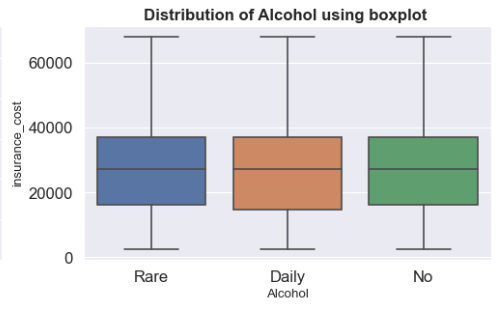
**9. covered\_by\_any\_other\_company**



**Fig 44: Box plot for covered\_by\_any\_other\_company**

From the above boxplot, we see that insurance cost has a significant impact on the insurance cost covered by other company of the person. The median of those people whose insurance cost is covered by other groups is slightly greater than the remaining group. This indicates that in order to attract such customers, the company has to give a higher insurance money.

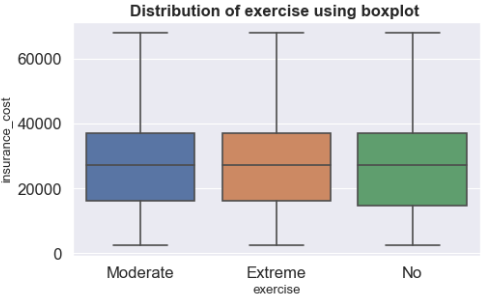
**10. Alcohol**



**Fig 45: Box plot for Alcohol**

From the above boxplot, we see that insurance cost has no impact on the alcoholic nature of the person since the median across all the three boxes remains at the same level.

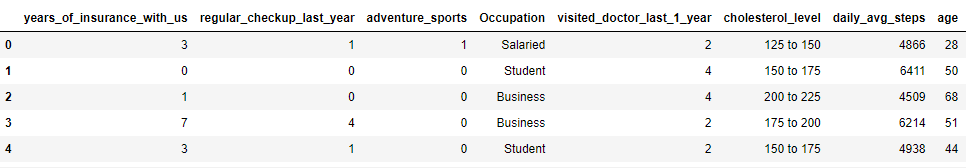
**11. exercise**



**Fig 46: Box plot for exercise**

From the above boxplot, we see that insurance cost has no impact on the exercise of the person since the median across all the three boxes remains at the same level.

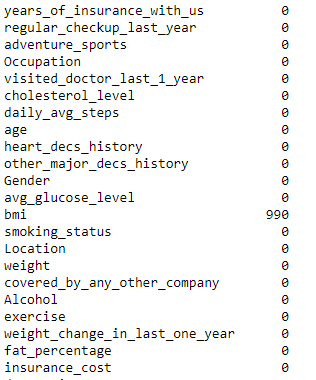
1. **Removal of unwanted variables (if applicable)**



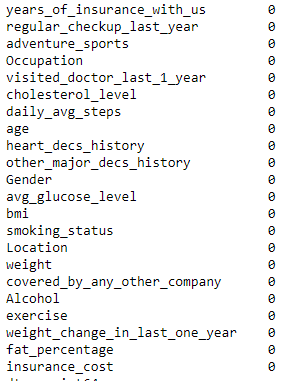
**Table 6: Head of the dataset after removal of unwanted variables**

Here two variables are removed, namely,Year\_last\_admitted, applicant\_id. Year\_last\_admitted is removed due to null values, which cannot be imputed. If it is imputed, then we are corrupting the data and applicant\_id is deleted since it has no relevance to the insurance cost we are calculating.

1. **Missing Value treatment (if applicable)**



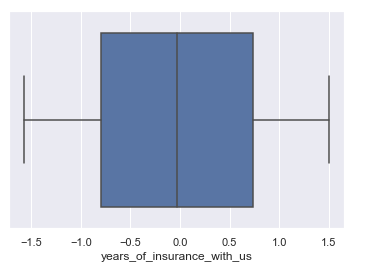
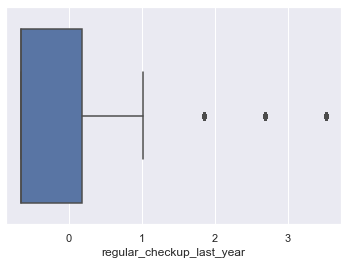
Here we see that the variable ‘bmi’ has 990 missing values which has to be imputed. We are using knn imputer to treat the missing values. KNN imputer works on the idea of k-nearest neighbours where the missing value is determined with the n number of its nearest neighbours. Before imputing the missing value, KNN imputer scales the data using standard scaler technique.Standard scaler removes the mean and scales each variable to unit variance. Hence the mean of each column becomes 0 and standard deviation becomes 1.

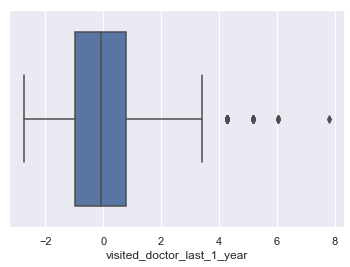
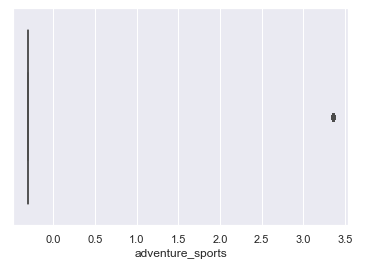


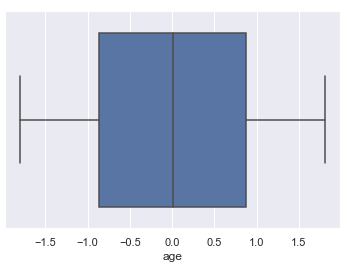
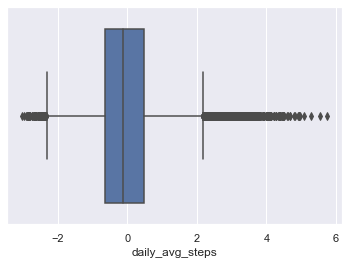
After the missing value treatment, we see that the dataset is free of missing values.

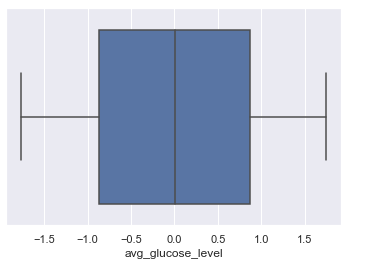
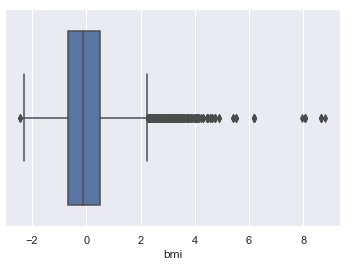
1. **Outlier treatment (if required)**

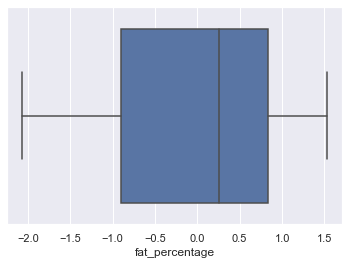
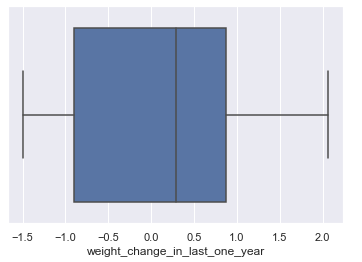
There were outliers present in the numerical columns of the given data set and it was treated. The numerical columns were specially taken out to perform the outlier treatment. First, we checked for the presence of outliers, then treated likewise.

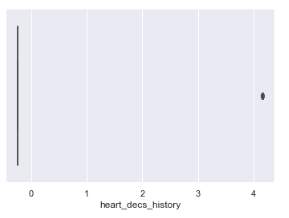
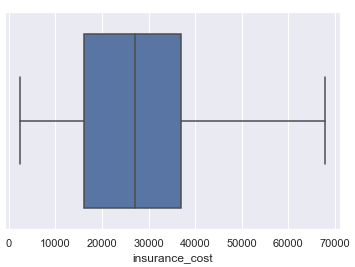
 

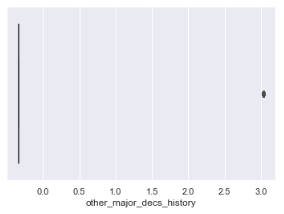






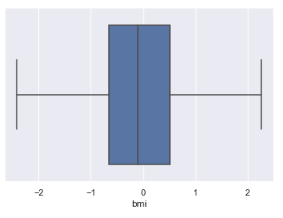
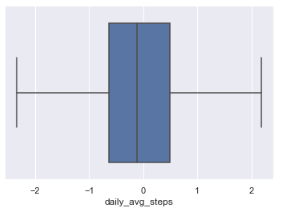






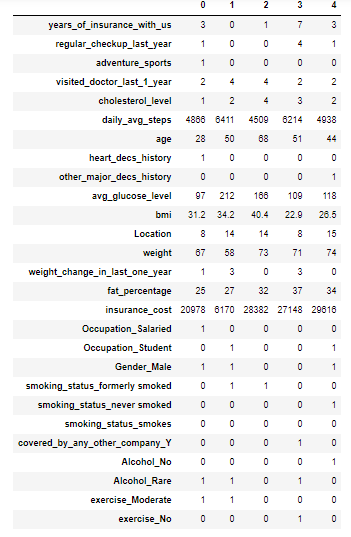
**Fig 44: Checking for outliers in the variables**

From the above figures, it is clear that outliers are present only in some variables such as bmi and daily average steps, hence only those variables need to be treated.



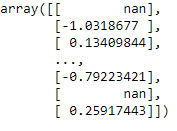
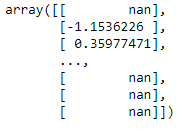
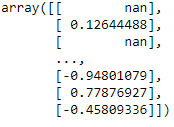
**Fig 46: after the outlier treatment in the selected variables**

1. **Variable transformation (if applicable)**



**Table 7: Transpose of the head of the dataset after encoding the categorical variables.**

Here variables are encoded using label encoding and one-hot encoding.

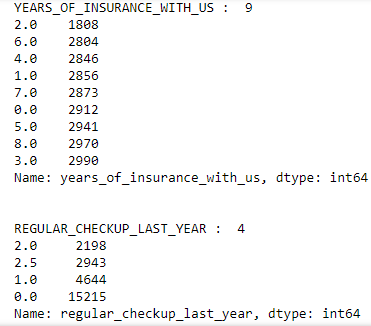
1. (b) (c)

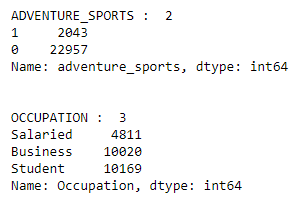
The above three arrays show the logarithmic transformation of the variables namely, age, bmi and daily average steps respectively. This kind of transformation is usually done for transforming highly skewed data into a more normalized data. If this transformation is not done, then when modelling variables with non-linear relationships, the chances of producing errors may also be skewed negatively. Logarithmic transformation is the most basic and simplest form of transformation done in linear regression.

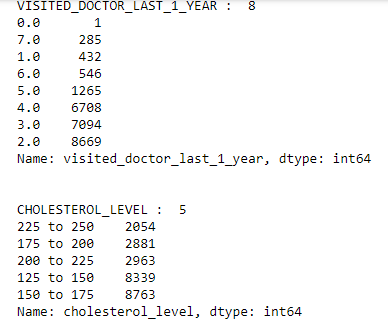
1. **Addition of new variables (if required)**

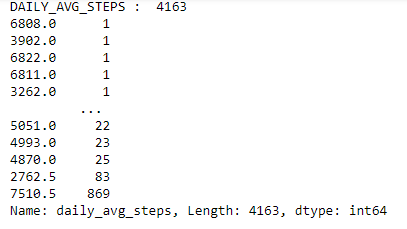
No new variables were added into the dataset. Only those variables who have undergone the one-hot encoding has been added to the dataset, thereby increasing the dimensionality.

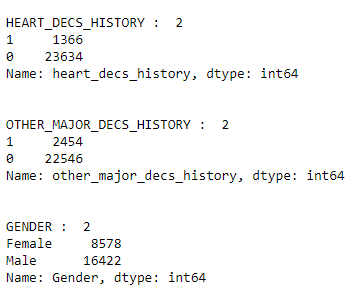
1. **Business insights from EDA**
2. **Is the data unbalanced? If so, what can be done? Please explain in the context of the business**

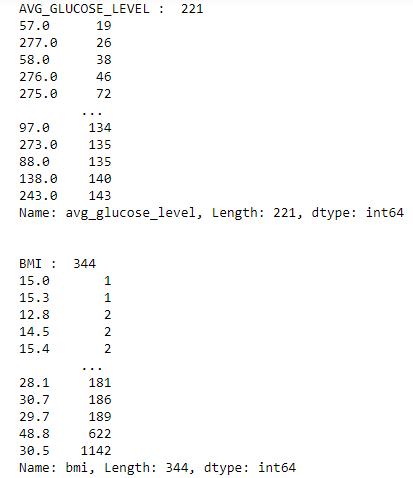


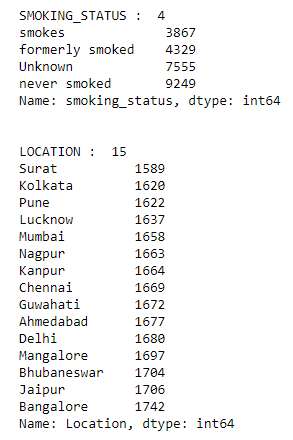


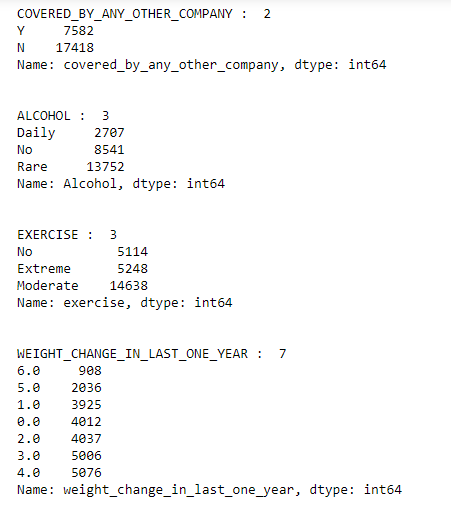








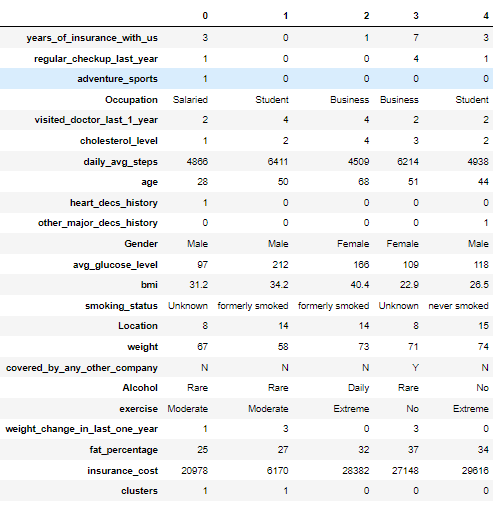




From the above list of unique value count of all the variables, we see that some variables are highly unbalanced, while others are moderately unbalanced.

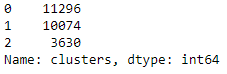
* In the column named regular\_checkup\_last\_year, half of the total population are those who have never gone for check-up last year.
* In the column cholesterol\_level, half of the population are who cholestrol level is very high comparred to those whose cholestrol level is at the normal or low level.
* In the column daily\_avg\_steps, people who have taken higher number of steps are unevenly distributed comparred to people who have taken lower steps.
* In the column named gender, females are only half the population of male, whih indicates the unbalance in the data.
* In the column exercise,people who exercise moderate are more than twice of those who exercise extremely and never have exercised.
* In the column bmi, a high proportion of the population have the bmi 30.5 which indicated unbalanced data.
* In the column covered\_by\_any\_other\_company, those people whose insurance is covered by any other company is less than half of those who insurance is not covered by any company at all.
* In the column heart\_decs\_history, majority of the population does not have any heart diesase history, while a small portion has a history of such aliment.
* In the column other\_major\_decs\_history,majority of the population does not have any such disease history, while a small portion has such a serious medical history.
* In other columns like bmi, age, location the data seems to be more evenly distributed.

1. **Any business insights using clustering (if applicable)**

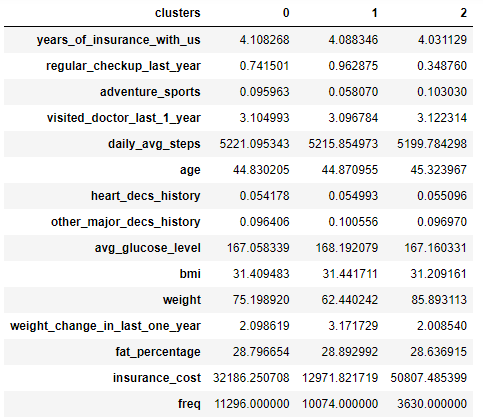


**Table 8: head of the data set after clustering**

From the above data set we see that 3 clusters are formed in the data given. We have used k- means clustering technique for clustering. In k- means, we choose the number of clusters by noticing the difference between the k-means inertia of consecutive clusters and by using elbow method.



Here we see that, highest population is in the first cluster, while the lowest population is in the third cluster.



**Table 9: description of the newly formed clusters**

**From the newly formed clusters, we see that:**

* All the three clusters have an average of 4 years of association with the insurance company.
* First cluster (cluster 0) is considered to the healthy cluster since they have highest daily average steps,
* Second cluster (cluster 1) has the lowest insurance cost, so they are considered to be healthy group of people
* Third cluster (cluster 2) has the highest insurance, so they can be considered as target audience.
* Cluster 1 can be given some bonus as an appreciation for keeping themselves fit.
* Cluster 2 can be assured of certain life time monetary funds and policies thereby gaining their trust and pulling in more such people of risky health condition.

1. **Any other business insights**

* The location has an impact on the insurance cost
* Those who insurance cost is covered by other companies need to be offered higher amount to be brought to this company
* The people who are into adventure sport, who are regular smokers and drinkers, who have visited the doctor many times last year can be treated as target customers.
* The people who exercise regularly can b given as bonus in insurance as an appreciation for their fiteness.