CAPSTONE (LIFE INSURANCE), FINAL REPORT

PG-DSBA

Written by

Priyamvada Singh

Dated: 19-11-2023

(Format: dd-mm-yyyy)

Table of Contents

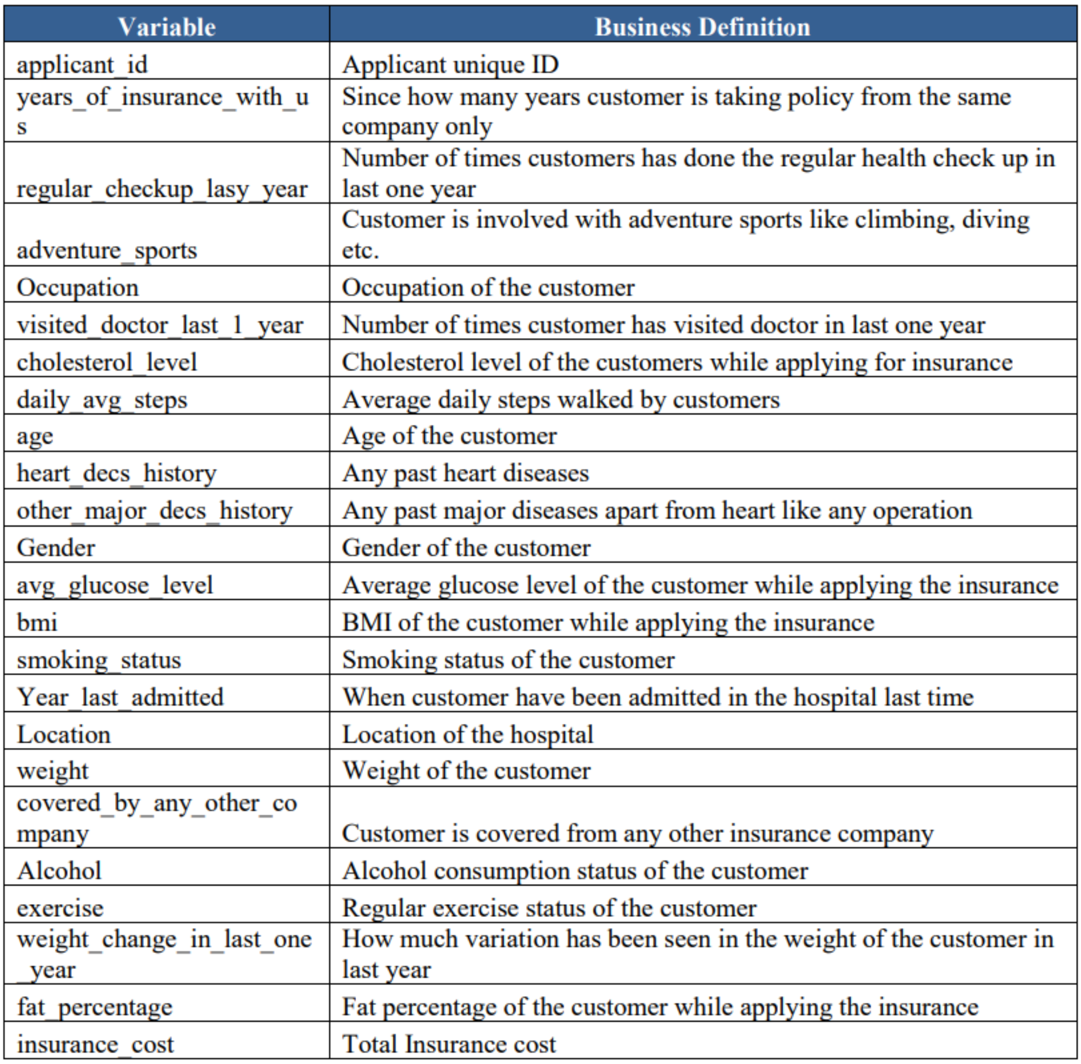
Table of Figures

1. Brief introduction about the problem statement and the need of solving it.

- The life insurance company wants to build a model to optimise their insurance cost for each individual that they acquire. The optimisation of the health insurance cost becomes important because each individual has a different lifestyle and health parameters, which vary their health risks, healthcare cost, and frequency of needing medical attention.

- The project will help build a dynamic linear regression model so that our dependent continuous variable can be optimally predicted as new customers come on board.

- The following parameters will help us further analyse trends, correlations, and segments amongst these records of customers and help build an efficient model by learning how these parameters impact the insurance cost tied to an individual:



Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
adventure_sports	Customer is involved with adventure sports like climbing, diving etc.
Occupation	Occupation of the customer
visited_doctor_last_1_year	Number of times customer has visited doctor in last one year
cholesterol_level	Cholesterol level of the customers while applying for insurance
daily_avg_steps	Average daily steps walked by customers
age	Age of the customer
heart_decs_history	Any past heart diseases
other_major_decs_history	Any past major diseases apart from heart like any operation
Gender	Gender of the customer
avg_glucose_level	Average glucose level of the customer while applying the insurance
bmi	BMI of the customer while applying the insurance
smoking_status	Smoking status of the customer
Year_last_admitted	When customer have been admitted in the hospital last time
Location	Location of the hospital
weight	Weight of the customer
covered_by_any_other_company	Customer is covered from any other insurance company
Alcohol	Alcohol consumption status of the customer
exercise	Regular exercise status of the customer
weight_change_in_last_one_year	How much variation has been seen in the weight of the customer in last year
fat_percentage	Fat percentage of the customer while applying the insurance
insurance_cost	Total Insurance cost

a. Business and Social Opportunity

- Some individuals might need to use their insurance more often than the others. For this reason, the insurance cost also varied so that the company can minimize its own business risks along with providing their customers with suitable health cover.
- This lets the company effectively adjust their costs and profit, in turn providing their customers with the optimum cover that will be useful to their own specific needs in the long-term.

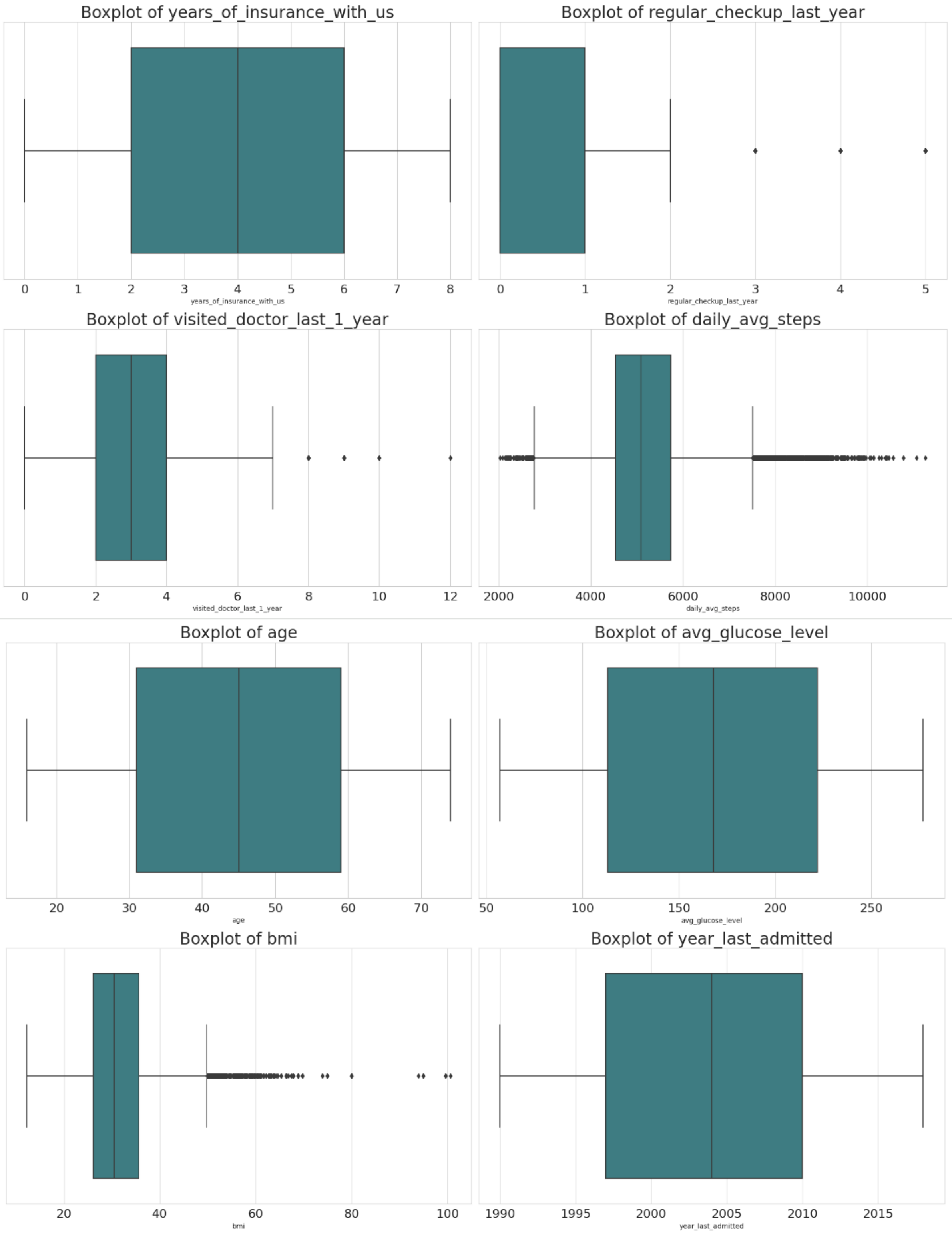
2. EDA and Business Implication

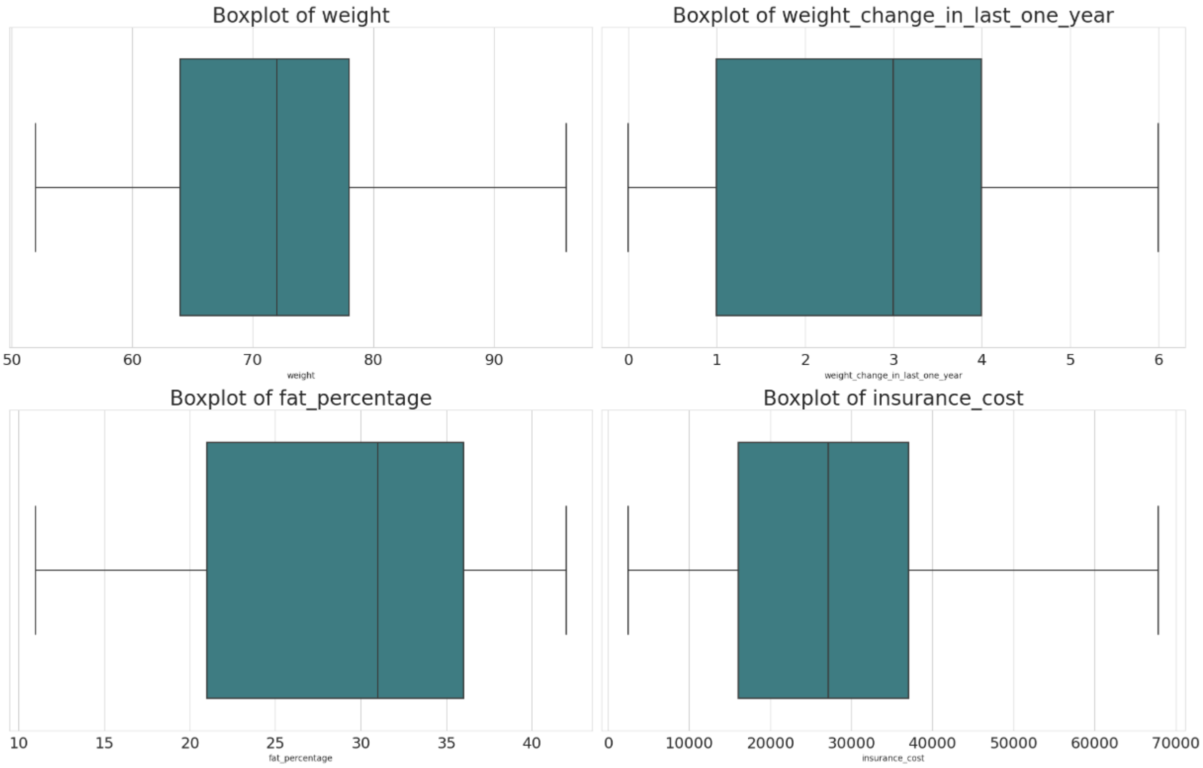
a. Uni-variate / bi-variate / multi-variate analysis to understand relationship between variables. How your analysis is impacting the business?

➢ *Univariate Analysis:*

- Following is the spread for each continuous variable:

Figure 1 - Boxplot for all numeric variables with outliers

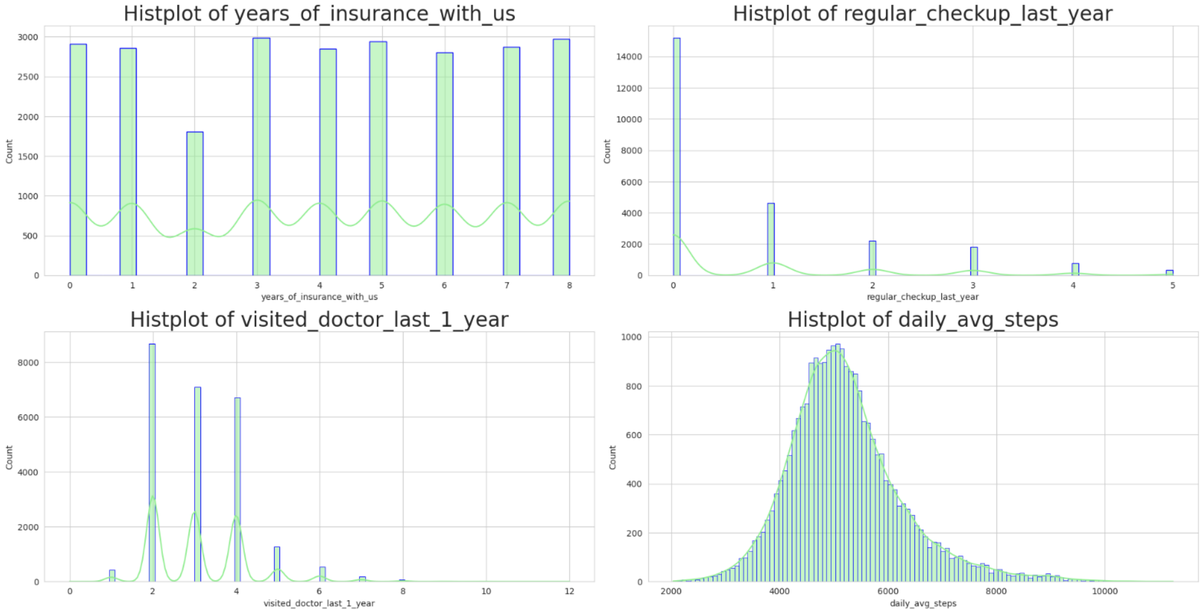


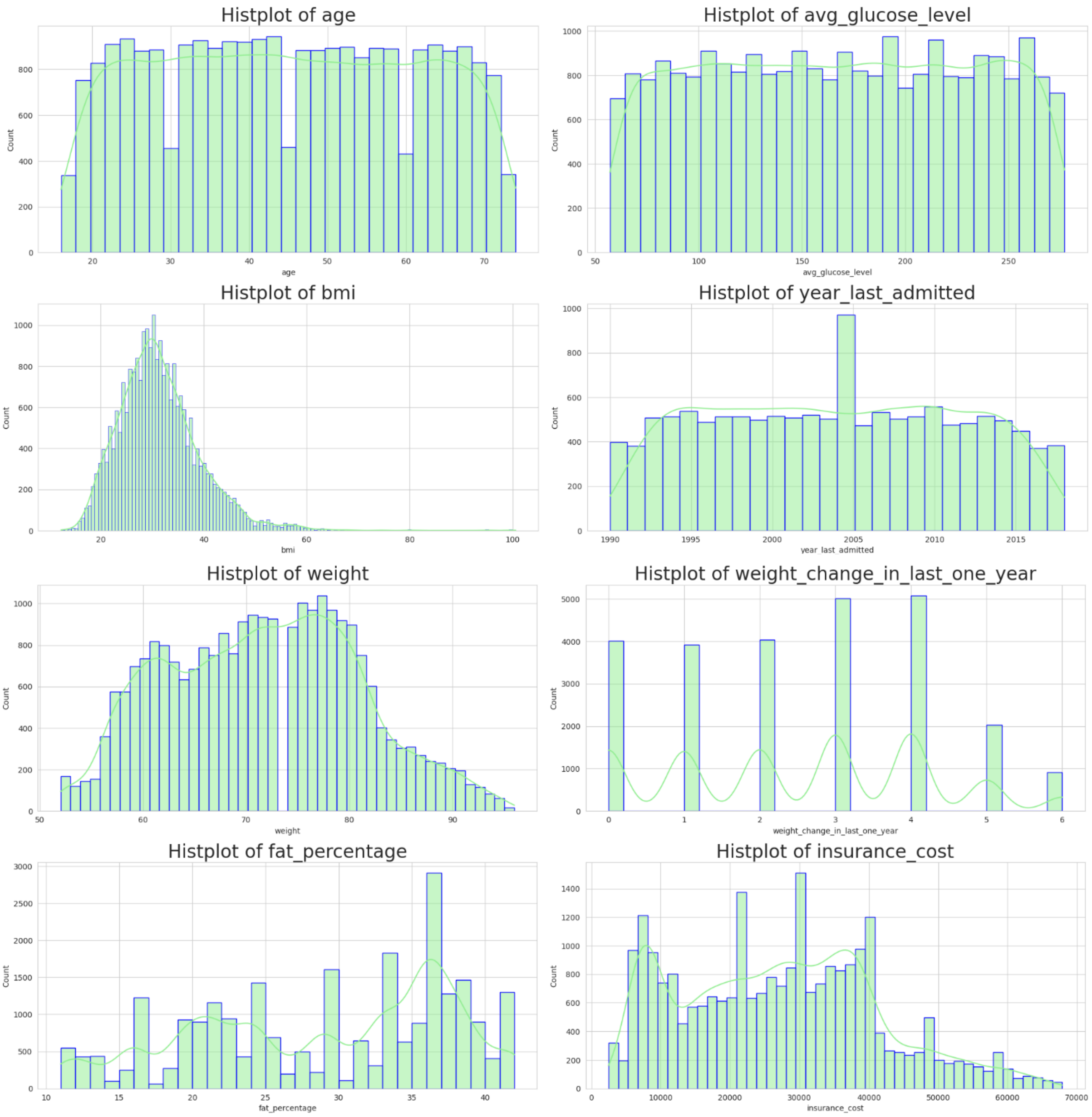


i. Columns: **regular_checkup_last_year**, **visited_doctor_last_1_year**, **daily_avg_steps** and **bmi** have outliers present.

- The following is the distribution for every continuous variable (count on the y-axis and the feature on the x-axis):

Figure 2 - Distribution for all numeric variables with KDE

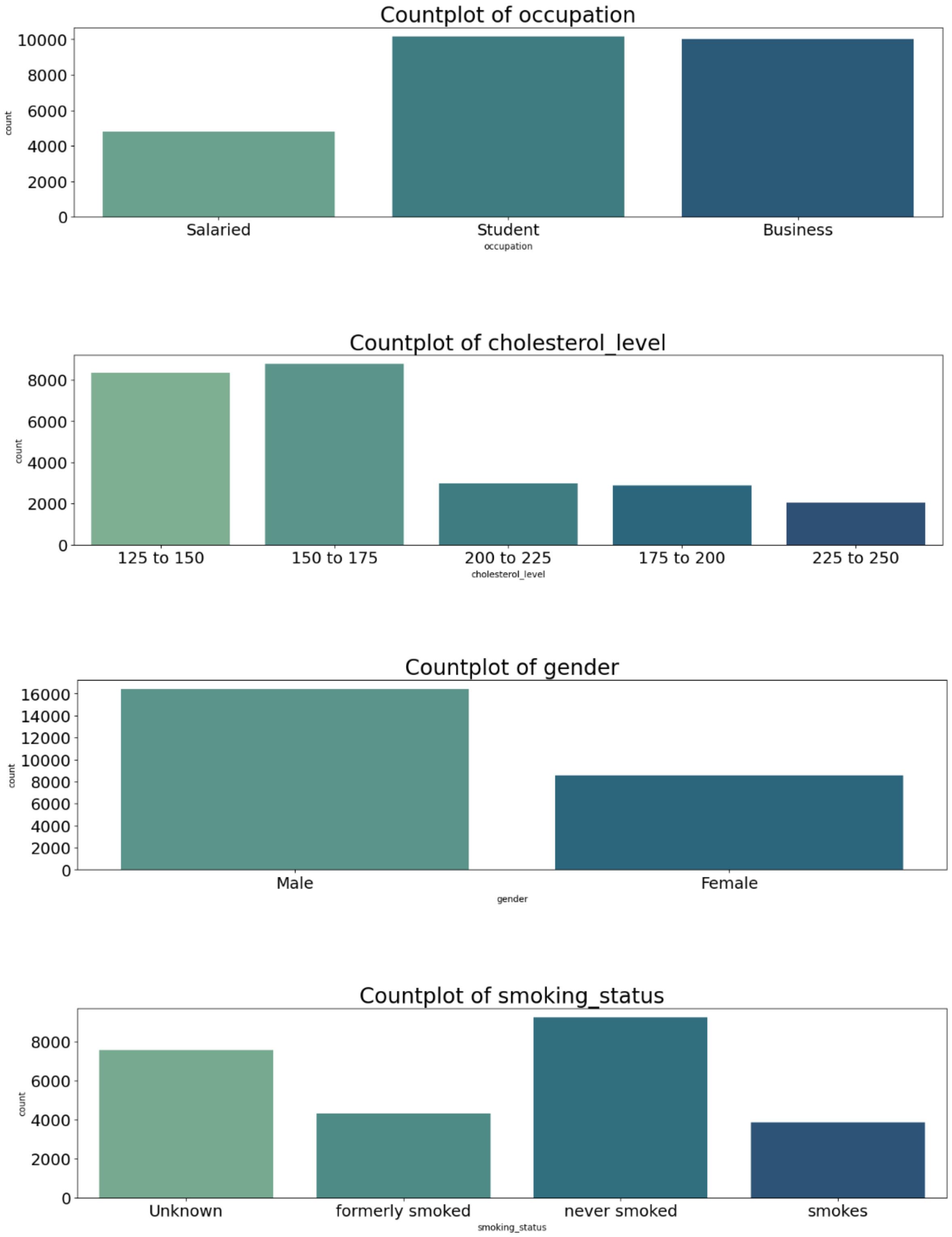


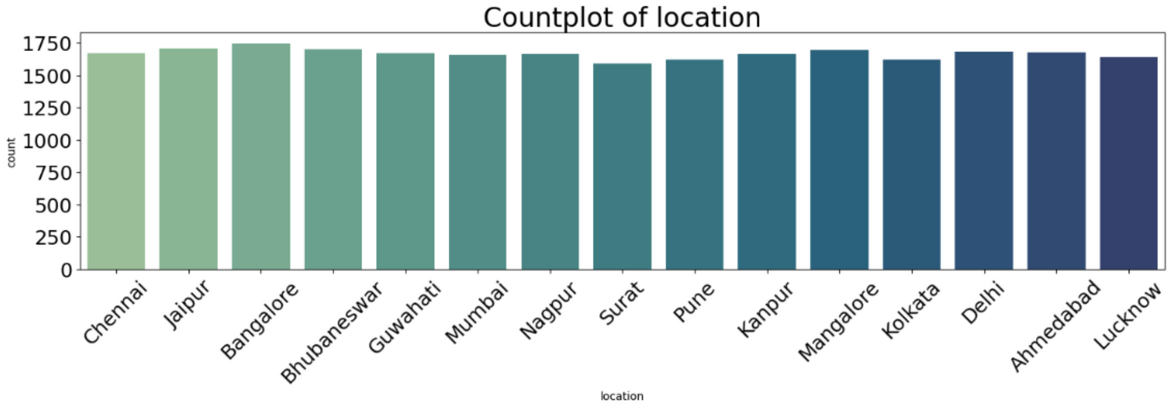


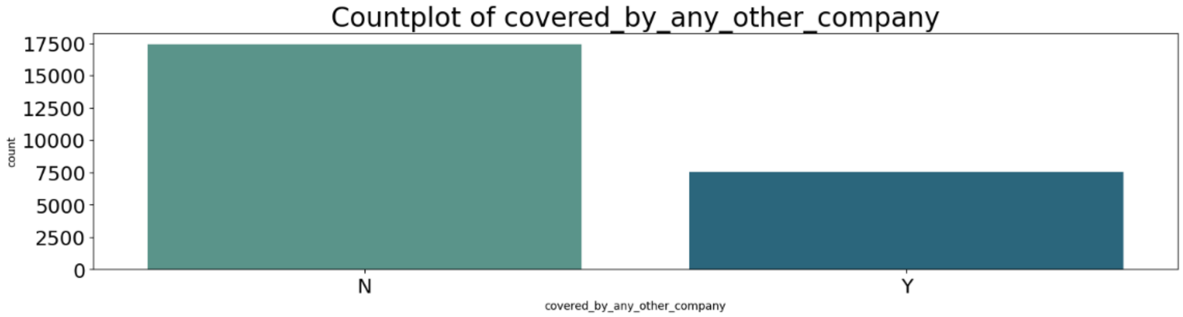
- **daily_avg_steps** and **bmi** are normally distributed with skew on the right tail.

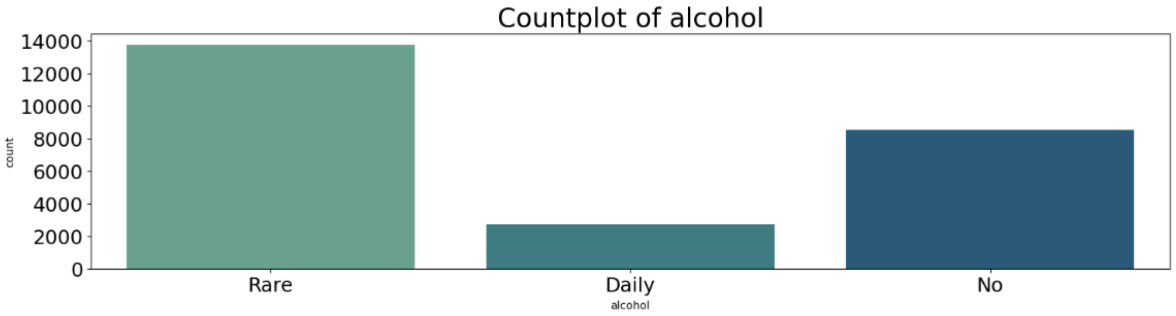
- The following is the count of each categorical variable:

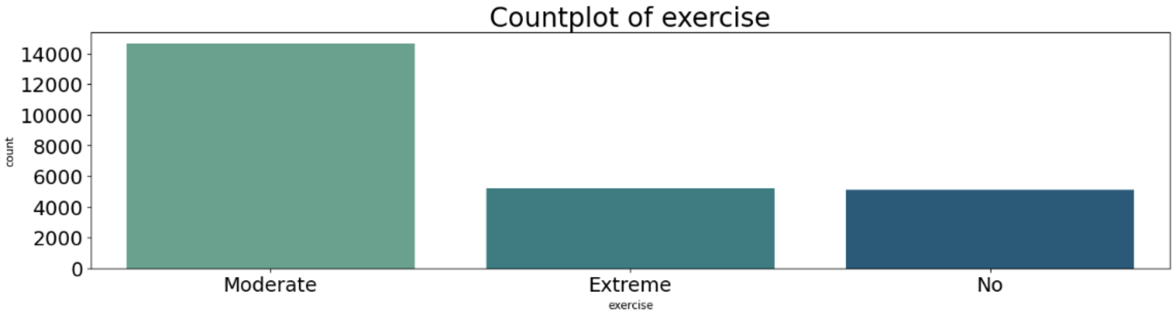
Figure 3 - Countplot for all categorical variables











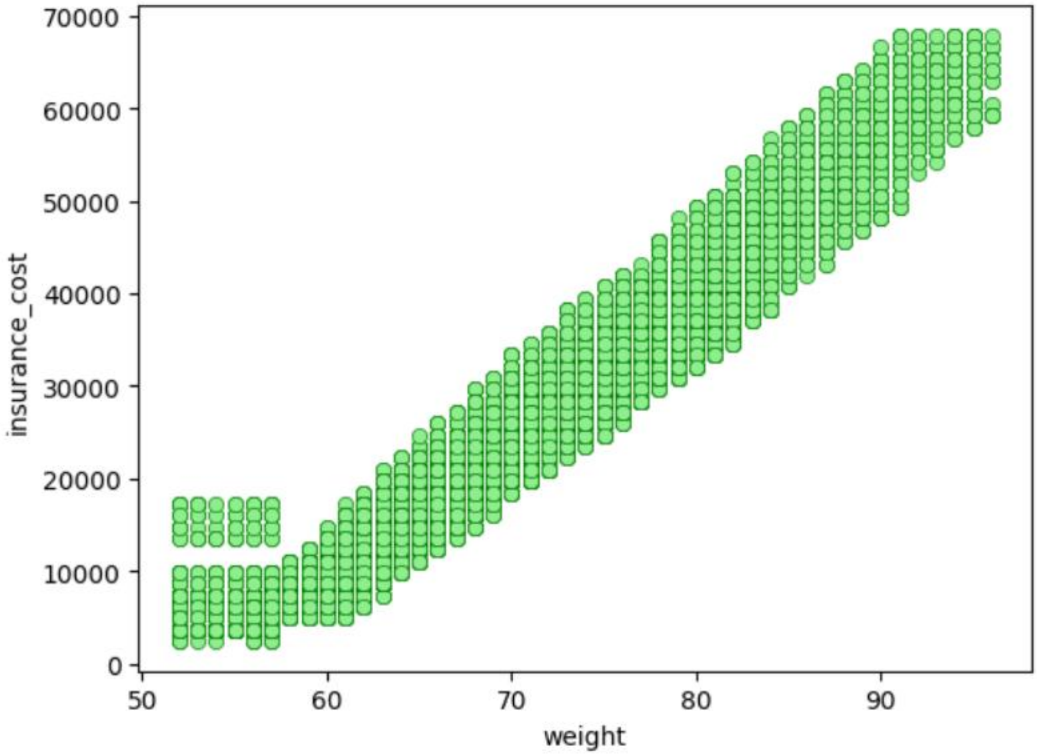
- Following are the most frequent values in each categorical variable:

  a. Occupation → Student (41% of total)
  b. Cholesterol level → 150 to 175 (35% of total)
  c. Gender → Male (66% of total)
  d. Smoking status → Never smoked (37% of total)
  e. Location → Bangalore (7% of total)
  f. Covered by any other company → No (70% of total)
  g. Alcohol intake → Rare (55% of total)
  h. Exercise → Moderate (59% of total)

- *Bivariate and Multivariate Analysis:*

  - The scatterplot for **weight** and **insurance_cost**:

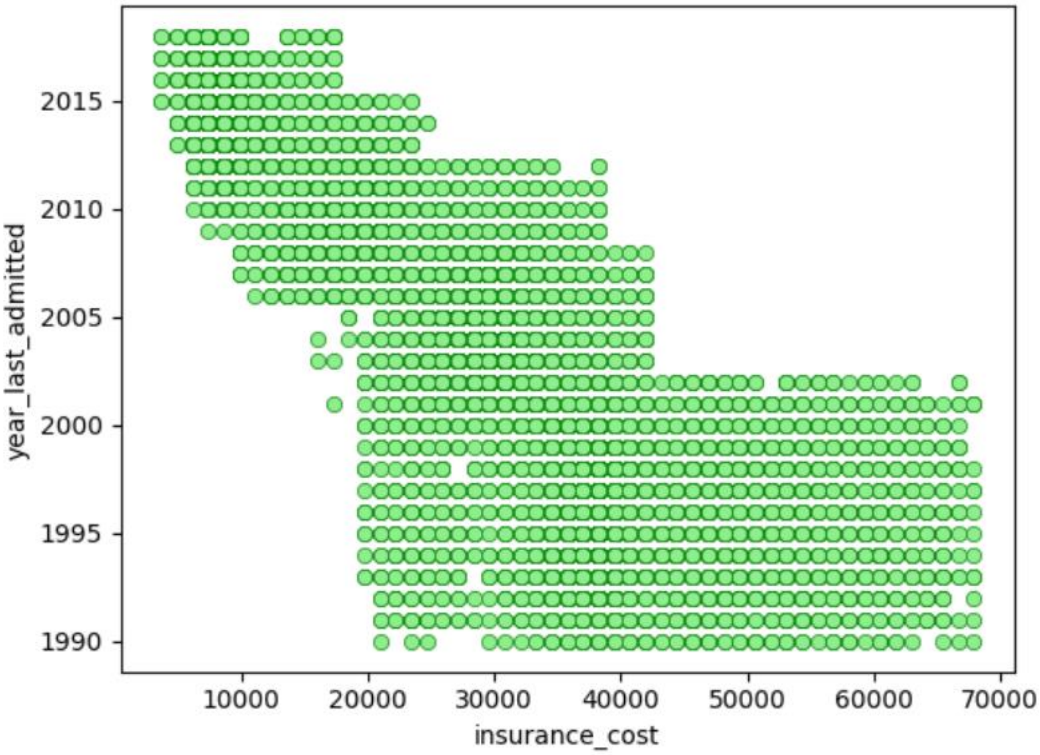
Figure 4 - Scatterplot of weight vs. insurance_cost



  - The insurance cost is positively correlated with the weight of the individual.

  - The scatterplot for **insurance_cost** and **year_last_admitted**:

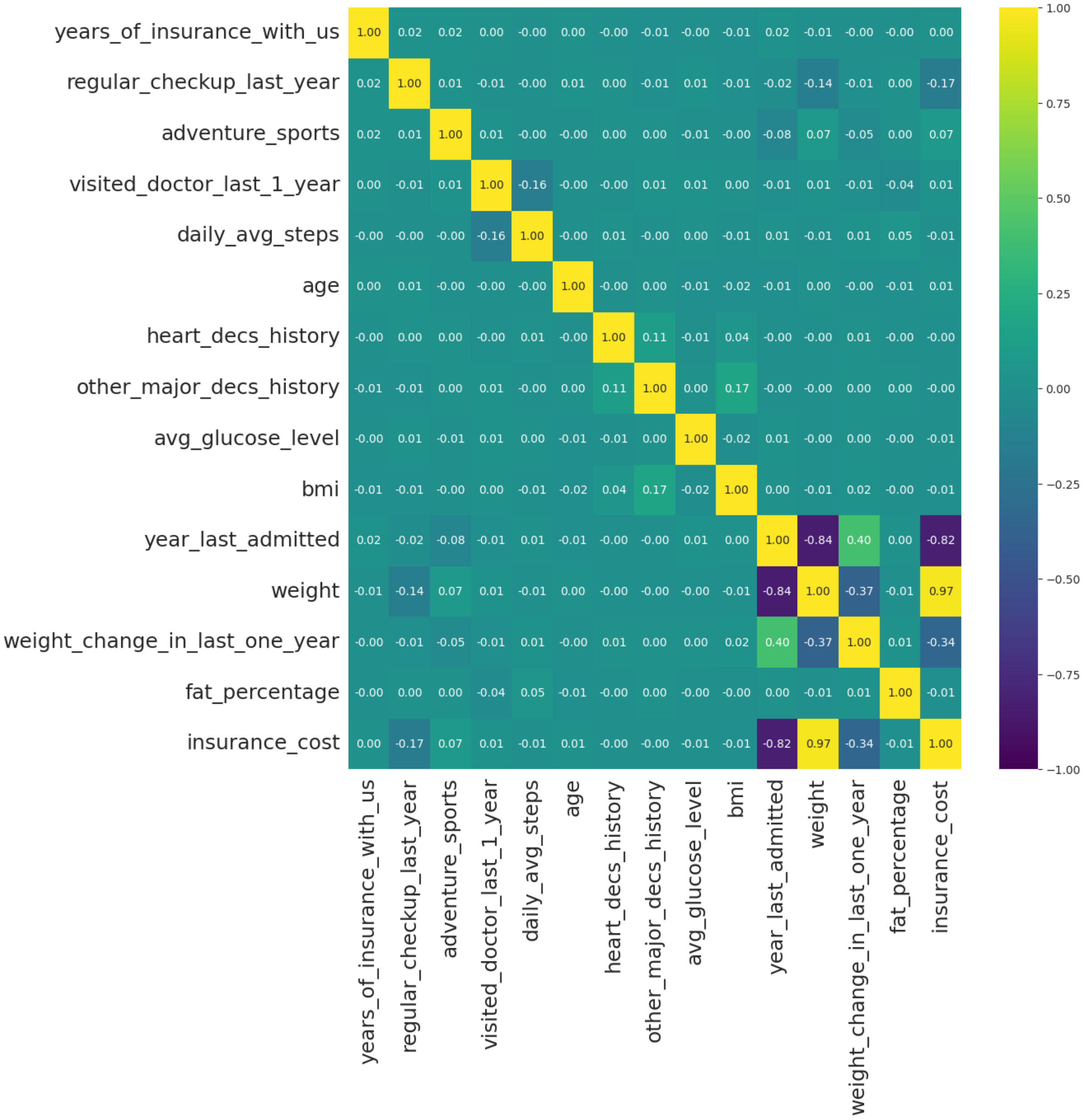
Figure 5 - year_last_admitted vs. insurance_cost



- The insurance cost is negatively correlated with the year the individual was last admitted. This could be because if the individual has been admitted recently, the chances that they will need medical attention again soon decreases. If the individual has not received any medical attention in a long time, the chances of the insurance cost being greater is more.

- The following is a correlation heatmap gives the quantitative relationship amongst all the variables:

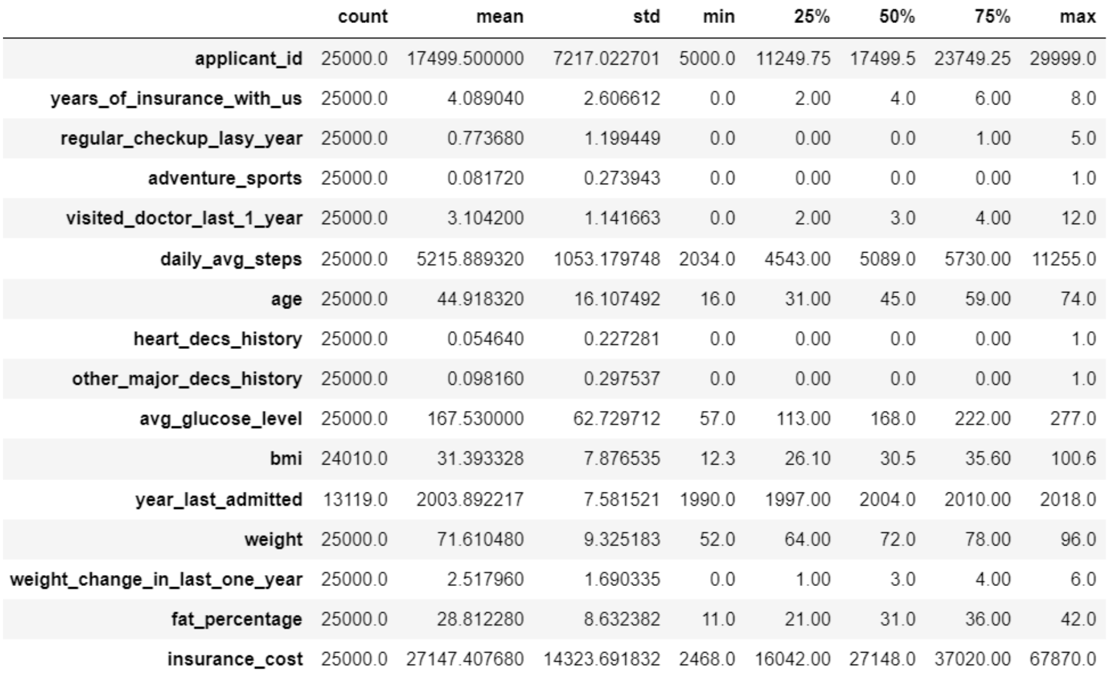
Figure 6 - Correlation heatmap for all numeric variables



- From the above correlation heatmap, it is clear that there are two major correlations as suspected through the scatterplots.
- The correlation between weight and insurance cost is 0.97 while the correlation between insurance cost and year last admitted is -0.82.
- Naturally, there is also a high correlation of -0.84 between weight and year last admitted.
- Apart from this, the dataset does not appear to have strong correlations or much multicollinearity.

b. Visual and non-visual understanding of the data

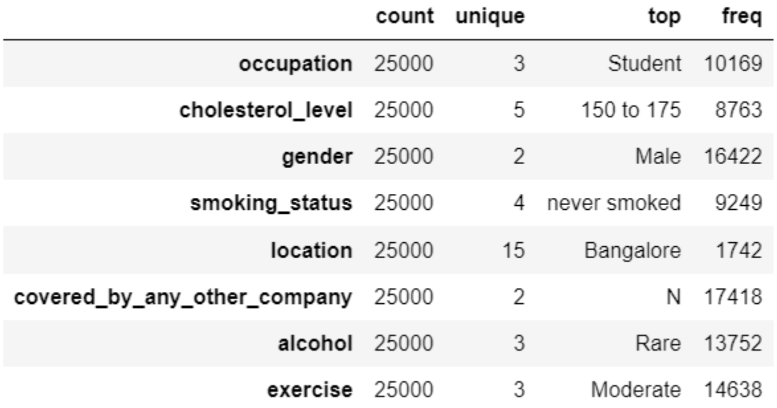
- The following is the 5-point summary of the numeric data fields:



	count	mean	std	min	25%	50%	75%	max
applicant_id	25000.0	17499.500000	7217.022701	5000.0	11249.75	17499.5	23749.25	29999.0
years_of_insurance_with_us	25000.0	4.089040	2.606612	0.0	2.00	4.0	6.00	8.0
regular_checkup_lasy_year	25000.0	0.773680	1.199449	0.0	0.00	0.0	1.00	5.0
adventure_sports	25000.0	0.081720	0.273943	0.0	0.00	0.0	0.00	1.0
visited_doctor_last_1_year	25000.0	3.104200	1.141663	0.0	2.00	3.0	4.00	12.0
daily_avg_steps	25000.0	5215.889320	1053.179748	2034.0	4543.00	5089.0	5730.00	11255.0
age	25000.0	44.918320	16.107492	16.0	31.00	45.0	59.00	74.0
heart_decs_history	25000.0	0.054640	0.227281	0.0	0.00	0.0	0.00	1.0
other_major_decs_history	25000.0	0.098160	0.297537	0.0	0.00	0.0	0.00	1.0
avg_glucose_level	25000.0	167.530000	62.729712	57.0	113.00	168.0	222.00	277.0
bmi	24010.0	31.393328	7.876535	12.3	26.10	30.5	35.60	100.6
year_last_admitted	13119.0	2003.892217	7.581521	1990.0	1997.00	2004.0	2010.00	2018.0
weight	25000.0	71.610480	9.325183	52.0	64.00	72.0	78.00	96.0
weight_change_in_last_one_year	25000.0	2.517960	1.690335	0.0	1.00	3.0	4.00	6.0
fat_percentage	25000.0	28.812280	8.632382	11.0	21.00	31.0	36.00	42.0
insurance_cost	25000.0	27147.407680	14323.691832	2468.0	16042.00	27148.0	37020.00	67870.0

  - In the above table, the minimum and maximum ranges of all the variables look correct except the higher range of BMI, which is 100.6 while the maximum weight is 96. It is possible that BMI has some bad values.

- The following is the description of the categorical data fields:



	count	unique	top	freq
occupation	25000	3	Student	10169
cholesterol_level	25000	5	150 to 175	8763
gender	25000	2	Male	16422
smoking_status	25000	4	never smoked	9249
location	25000	15	Bangalore	1742
covered_by_any_other_company	25000	2	N	17418
alcohol	25000	3	Rare	13752
exercise	25000	3	Moderate	14638

- In the above table, all the top values appear to be in favour of the insurance company. For instance, the top value under 'alcohol' is 'Rare', 'exercise' is 'Moderate', and 'smoking_status' is 'never smoked'. Even the cholesterol level is under 200, which is a sign of a healthy individual.

- Dataset info:

  - The dataset has two float64, fourteen int64, and eight object datatype variables.
  - It has 24 columns and 25000 rows with two columns having null values.
  - The column **bmi** has 990 null values while **year_last_admitted** has 11881 null values.

- Renaming:

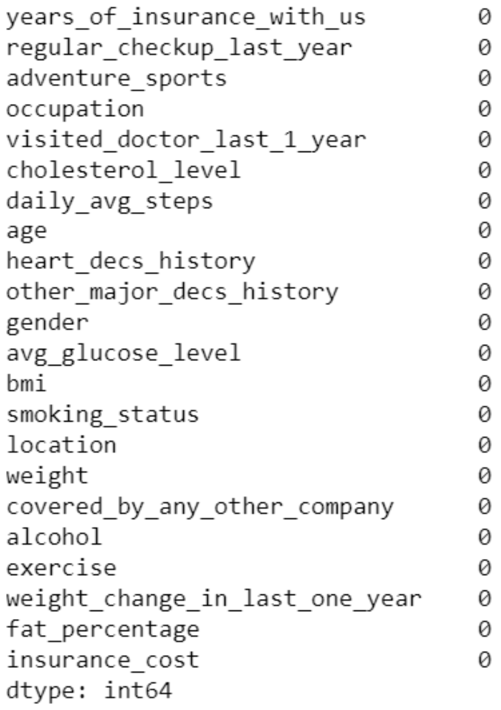
- Since the names of the columns were inconsistent in terms of letter casing, the column names have been converted into lower case for convenience.
- Under the categorical variable called ‘occupation’, the spelling of the value ‘Salried’ has been corrected as ‘Salaried’.
- The column ‘regular_checkup_lasy_year’ has been corrected and renamed ‘regular_checkup_last_year’.

3. Data Cleaning and Pre-processing

a. Approach used for identifying and treating missing values and outlier treatment (and why)

- Missing values treatment:

  - The columns named **bmi** (990) and **year_last_admitted** (11881) have null values.
  - At this point, **year_last_admitted** is dropped from the dataset because it has 47.52% null values. Imputation of such a large number of rows for a feature might negatively impact the analysis.
  - The **bmi** variable is imputed by its median to treat the null values.
  - The following is a snapshot after treating the null values of **bmi** and dropping **year_last_admitted**:

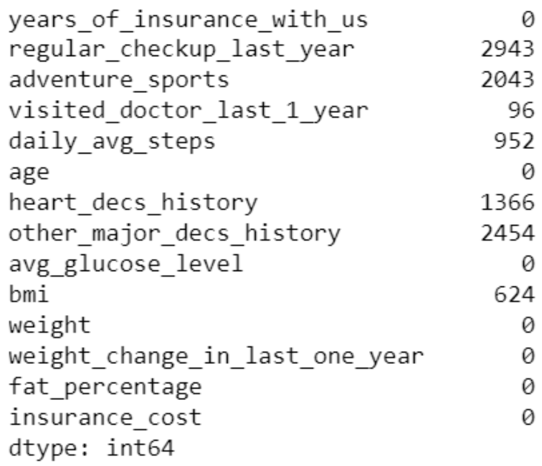


```
years_of_insurance_with_us       0
regular_checkup_last_year        0
adventure_sports                 0
occupation                       0
visited_doctor_last_1_year       0
cholesterol_level                0
daily_avg_steps                  0
age                              0
heart_decs_history               0
other_major_decs_history         0
gender                           0
avg_glucose_level                0
bmi                              0
smoking_status                   0
location                         0
weight                           0
covered_by_any_other_company     0
alcohol                          0
exercise                         0
weight_change_in_last_one_year   0
fat_percentage                   0
insurance_cost                   0
dtype: int64
```

- Outlier treatment:

  - The following shows the number of outliers present in each feature:

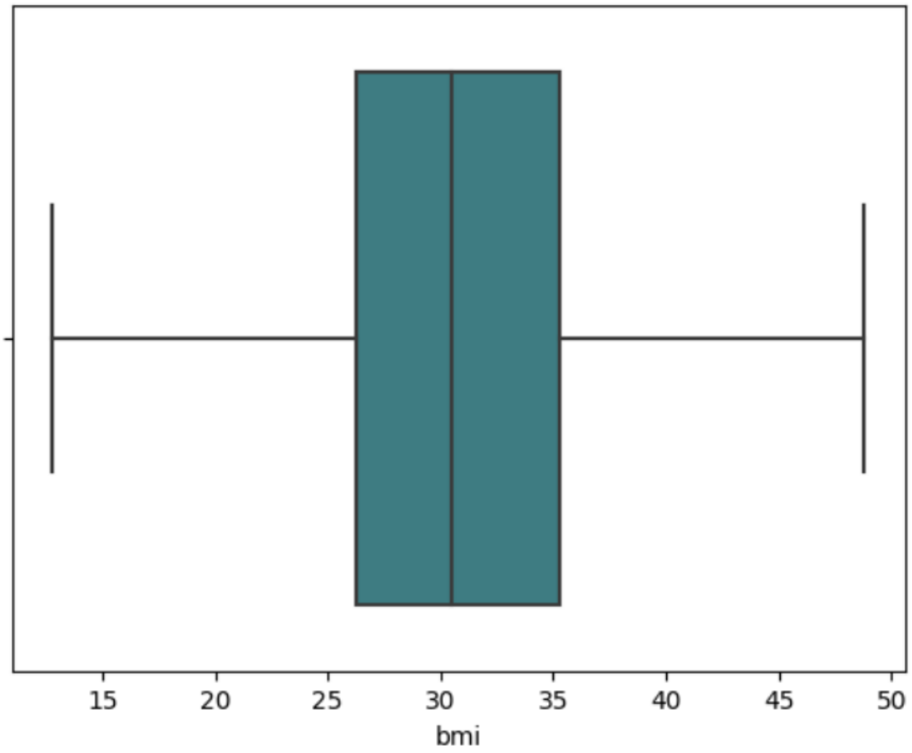
Figure 7 - Outlier count table



```
years_of_insurance_with_us           0
regular_checkup_last_year         2943
adventure_sports                  2043
visited_doctor_last_1_year          96
daily_avg_steps                    952
age                                  0
heart_decs_history                1366
other_major_decs_history          2454
avg_glucose_level                    0
bmi                                624
weight                               0
weight_change_in_last_one_year       0
fat_percentage                       0
insurance_cost                       0
dtype: int64
```

- After analysing the outlier values for all columns, **bmi** appears to have bad values in the form of outliers. It has abnormally high values for many records. At this stage, we will only treat the outliers for **bmi** by capping with the IQR method while keeping the all the other columns untreated.
- Although, the BMI of individuals could actually be higher than 50, the company is advised to ask for the height of the individuals so this column can be calculated accurately from our end.
- The following graph shows that the outliers in the **bmi** feature have been treated:

Figure 8 - Outlier treatment for bmi



- For the **bmi**, the upper range is 48.80 and the lower range is 12.80.

b. Need for variable transformation (if any)

- Since this is a regression problem, the categorical variables were converted into numeric.

- The variables **covered_by_any_other_company**, **cholesterol_level**, **alcohol**, and **exercise** were given numeric ranges manually with the following values:

  - **covered_by_any_other_company**
    N = 0
    Y = 1

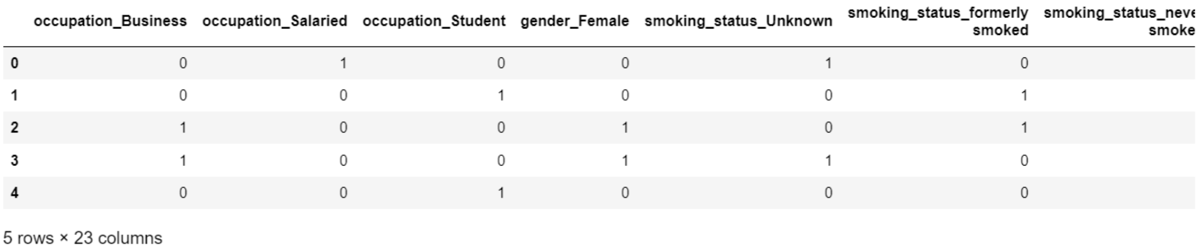
  - **cholesterol_level**
    125 to 150 = 1
    150 to 175 = 2
    175 to 200 = 3
    200 to 225 = 4
    225 to 250 = 5

  - **alcohol**
    No = 0
    Rare = 1
    Daily = 2

  - **exercise**
    No = 0
    Moderate = 1
    Extreme = 2

Note: These variables were converted into int64 datatype after being coded.

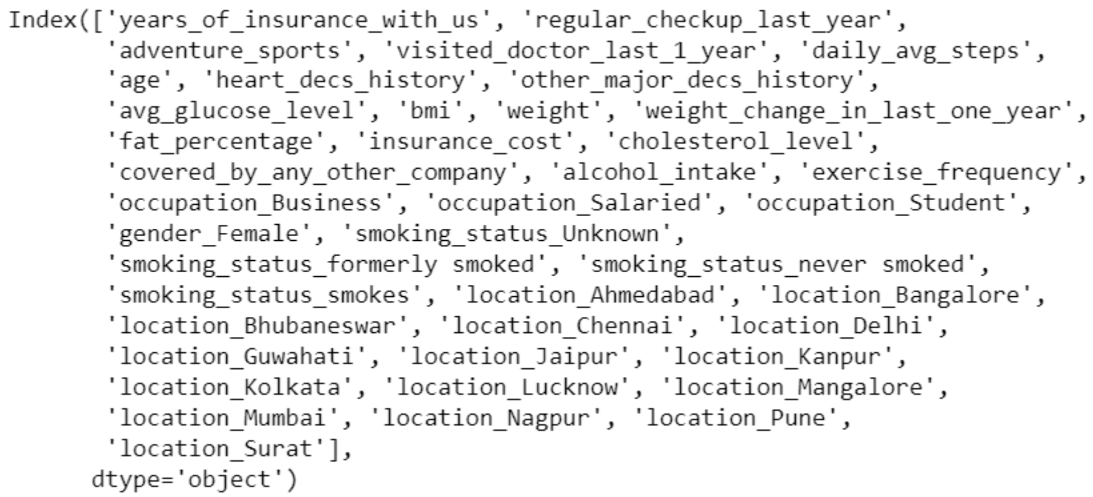
- The other categorical variables, namely, **occupation**, **gender**, **smoking_status**, **location** were converted into numeric variables by one hot encoding.
- The following is a sample of the one hot encoded dataset:



	occupation_Business	occupation_Salaried	occupation_Student	gender_Female	smoking_status_Unknown	smoking_status_formerly smoked	smoking_status_neve smoke
0	0	1	0	0	1	0	
1	0	0	1	0	0	1	
2	1	0	0	1	0	1	
3	1	0	0	1	1	0	
4	0	0	1	0	0	0	

5 rows × 23 columns

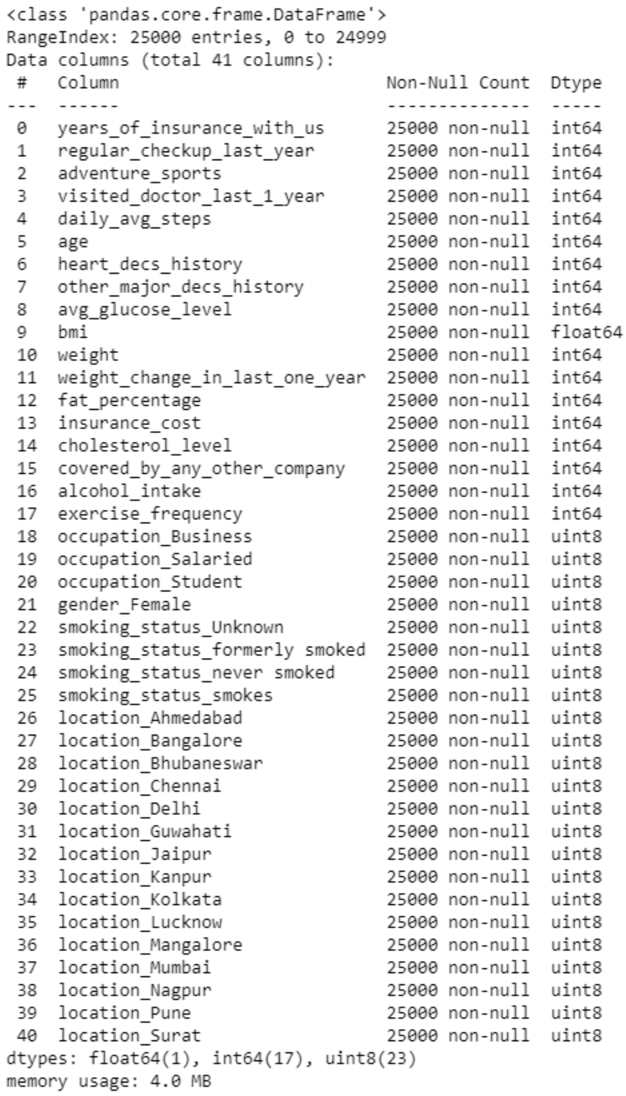
- The new dataset was constructed by concatenating the newly converted numeric variables and the original numeric variables.
- The following are the columns inside the new dataset, which is fully numeric:



```
Index(['years_of_insurance_with_us', 'regular_checkup_last_year',
       'adventure_sports', 'visited_doctor_last_1_year', 'daily_avg_steps',
       'age', 'heart_decs_history', 'other_major_decs_history',
       'avg_glucose_level', 'bmi', 'weight', 'weight_change_in_last_one_year',
       'fat_percentage', 'insurance_cost', 'cholesterol_level',
       'covered_by_any_other_company', 'alcohol_intake', 'exercise_frequency',
       'occupation_Business', 'occupation_Salaried', 'occupation_Student',
       'gender_Female', 'smoking_status_Unknown',
       'smoking_status_formerly smoked', 'smoking_status_never smoked',
       'smoking_status_smokes', 'location_Ahmedabad', 'location_Bangalore',
       'location_Bhubaneswar', 'location_Chennai', 'location_Delhi',
       'location_Guwahati', 'location_Jaipur', 'location_Kanpur',
       'location_Kolkata', 'location_Lucknow', 'location_Mangalore',
       'location_Mumbai', 'location_Nagpur', 'location_Pune',
       'location_Surat'],
      dtype='object')
```

- In this new dataset, the columns **alcohol** and **exercise** have been renamed to **alcohol_intake** and **exercise_frequency** for more clarity.

- The following is the info of the new dataset:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 41 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   years_of_insurance_with_us      25000 non-null  int64  
 1   regular_checkup_last_year       25000 non-null  int64  
 2   adventure_sports                25000 non-null  int64  
 3   visited_doctor_last_1_year      25000 non-null  int64  
 4   daily_avg_steps                 25000 non-null  int64  
 5   age                             25000 non-null  int64  
 6   heart_decs_history              25000 non-null  int64  
 7   other_major_decs_history        25000 non-null  int64  
 8   avg_glucose_level               25000 non-null  int64  
 9   bmi                             25000 non-null  float64
 10  weight                          25000 non-null  int64  
 11  weight_change_in_last_one_year  25000 non-null  int64  
 12  fat_percentage                  25000 non-null  int64  
 13  insurance_cost                  25000 non-null  int64  
 14  cholesterol_level               25000 non-null  int64  
 15  covered_by_any_other_company    25000 non-null  int64  
 16  alcohol_intake                  25000 non-null  int64  
 17  exercise_frequency              25000 non-null  int64  
 18  occupation_Business             25000 non-null  uint8  
 19  occupation_Salaried             25000 non-null  uint8  
 20  occupation_Student              25000 non-null  uint8  
 21  gender_Female                   25000 non-null  uint8  
 22  smoking_status_Unknown          25000 non-null  uint8  
 23  smoking_status_formerly smoked  25000 non-null  uint8  
 24  smoking_status_never smoked     25000 non-null  uint8  
 25  smoking_status_smokes           25000 non-null  uint8  
 26  location_Ahmedabad              25000 non-null  uint8  
 27  location_Bangalore              25000 non-null  uint8  
 28  location_Bhubaneswar            25000 non-null  uint8  
 29  location_Chennai                25000 non-null  uint8  
 30  location_Delhi                  25000 non-null  uint8  
 31  location_Guwahati               25000 non-null  uint8  
 32  location_Jaipur                 25000 non-null  uint8  
 33  location_Kanpur                 25000 non-null  uint8  
 34  location_Kolkata                25000 non-null  uint8  
 35  location_Lucknow                25000 non-null  uint8  
 36  location_Mangalore              25000 non-null  uint8  
 37  location_Mumbai                 25000 non-null  uint8  
 38  location_Nagpur                 25000 non-null  uint8  
 39  location_Pune                   25000 non-null  uint8  
 40  location_Surat                  25000 non-null  uint8  
dtypes: float64(1), int64(17), uint8(23)
memory usage: 4.0 MB
```

- As seen above, all the variables are now numeric.

c. Variables removed or added and why (if any):

- Column named **year_last_admitted** was dropped because it had 47.52% null values. It is not advised to impute such a high number of null values for any feature.
- The **applicant_id** column has also been dropped as it holds no value with respect to the exploratory data analysis and model building.

➢ *Business Recommendations:*

i. Under the **smoking_status** category, the option Unknown can be removed (if it were a customer survey form that was used for data collection).
ii. The company can add another feature named **height** along with the weight (already existing). This will help us to calculate the **bmi** field accurately, hence improving data integrity.
iii. Individuals can be segmented based on the location to run specific campaigns depending on the feature specifications of the city.
iv. There are more records of males than females. The company may also look into the causes for a higher male percentage in the records.
    - Is it due to a fundamental difference between the health patterns of females and males, or
    - is it that the company has not run a proper acquisition campaign that specifically appeals to females?
    - This investigation can be a potential opportunity to acquire more customers.
v. On comparing the values of all the categorical variables against the insurance cost, no strong trend was found after performing the EDA.
vi. The **year_last_admitted** field can be made mandatory to avoid null values. This column had to dropped as it had around 48% null values. This feature has a strong correlation with the insurance cost (target) variable.
vii. Make as many features mandatory to fill as possible to avoid null values.

4. Model building

- Linear models:

    i. Linear Regression
    ii. Ridge
    iii. Lasso
    iv. Elastic Net

- Ensemble models:

    i. Gradient Boosting Regressor
    ii. Bagging Regressor

a. Why the above models were chosen:

- The business wants to predict the ***insurance cost*** of an individual, which is a **continuous variable**. To build a predictive model where the target variable is continuous, **Linear Regression** was used.
- Along with Linear Regression, we also build three other models that **regularize the linear regression algorithm** to generate optimized output. The three additional models were **Ridge, Lasso,** and **Elastic Net**. This is so that we can choose the best performing model for final use.

- Additionally, two more regressors from Ensemble techniques were used to see if they predict the insurance cost with higher accuracy and consistency. The models built were **Gradient Boosting Regressor** and **Bagging Regressor**.

b. Efforts to improve model performance:

- The models were re-built after changing a few key characteristics of the dataset.

  - The train and test dataset split was changed from 70:30 to 20:80.
  - Earlier, we had dropped ‘year_last_admitted’ due to too many missing values. Now, we imputed those values and built the model. The imputation method used was KNN Imputer. The values were float but were converted to int to achieve discrete name for the years.
  - The outliers for ‘bmi’ were left untreated.
  - Grid Search CV was used on all the best performing base models for hyperparameter tuning.

- The **final models chosen** were **Elastic Net** and **Gradient Boosting Regressor**. The Grid Search CV optimization was performed using the following parameters:

  - **Gradient Boosting**

Loss = 'huber’ (default: 'squared error’)
n_estimators = 100 (default: 100)

  - **Elastic Net**

Alpha = 0.01 (default: 1.0)
L1 ratio = 0.3 (default: 0.5)

5. Model validation

a. How was the model validated? Just accuracy, or anything else too?

- The model was validated using R-squared (R2), root mean squared error (RMSE) and mean absolute percentage error (MAPE). Below is the model comparison table with the metrics for each final model:

Model Category	Model Name	R2 Train	R2 Test	RMSE Train	RMSE Test	MAPE Train	MAPE Test
Linear	Linear Regression	0.945	0.944	3368	3374	0.15	0.16
	Ridge	0.945	0.944	3368	3374	0.389	0.394
	Lasso	0.945	0.944	3369	3373	0.39	0.394
	Elastic Net	0.945	0.944	3368	3374	0.152	0.155
Ensemble	Gradient Boosting	0.956	0.955	3013	3026	0.121	0.121
	Bagging	0.991	0.946	1355	3320	0.052	0.131

b. Detailed recommendations for the management/client based on the analysis done.

i. The equation for linear regression is:

insurance cost = -79954.55083061793 + (-448.07 * regular_checkup_last_year) + 197.80 * (adventure_sports) + (-56.37 * (visited_doctor_last_1_year)) + 3.76 * (age) + 286.24 * (heart_decs_history) + 1488.98 * (weight) + 163.83 * (weight_change_in_last_one_year) + 1243.08 * (covered_by_any_other_company)

ii. The equation for ridge regression:

Insurance cost = -79954.45 + (-447.86 * regular_checkup_last_year) + (196.38 * adventure_sports) + (-56.32 * visited_doctor_last_1_year) + (3.77 * age) + (283.19 * heart_decs_history) + (1488.99 * weight) + (163.73 * weight_change_in_last_one_year) + (1239.70 * covered_by_any_other_company)

iii. The equation for lasso regression:

Insurance cost = 79944.54 + (-440.12 * regular_checkup_last_year) + (65.09 * adventure_sports) + (-48.23 * visited_doctor_last_1_year) + (3.70 * age) + (95.78 * heart_decs_history) + (1489.11 * weight) + (159.02 * weight_change_in_last_one_year) + (1195.58 * covered_by_any_other_company)

iv. The equation for elastic net:

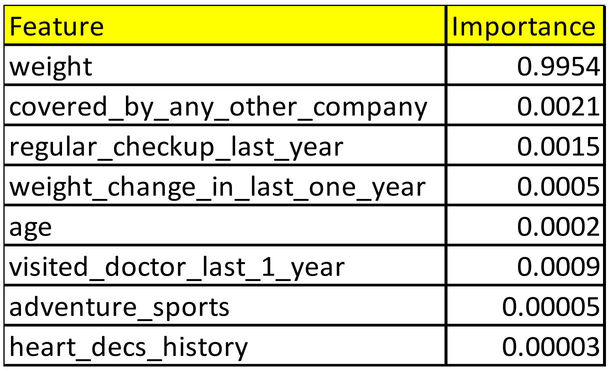
Insurance cost = -79941.93 + (-445.51 * regular_checkup_last_year) + (181.69 * adventure_sports) + (-55.83 * visited_doctor_last_1_year) + (3.77 * age) + (252.86 * heart_decs_history) + (1489 * weight) + (162.66 * weight_change_in_last_one_year) + (1203 * covered_by_any_other_company)

➢ *Inferences:*

- According to the models, regular checkup last year is negatively correlated with the insurance cost. The lower the number of checkups, the higher the insurance cost for the individual.
- The feature ‘visited_doctor_last_1_year’ is also negatively correlated to the insurance cost. The higher the number, the lower the insurance cost for the individual.
- All the other variables are positively correlated with the target variable.
- Weight seems to have the highest impact on the insurance cost of an individual. The correlation is positive.
- Another significant learning is that as age and the number of adventure sports increase, the insurance cost goes up.
- All the equations have similar weights for the variables. However, the regularization models have lowered the weights when compared to the original linear regression model. Elastic Net has lowered the weights the most as it uses a combination of ridge and lasso to regularize the model and implement penalty on the features.

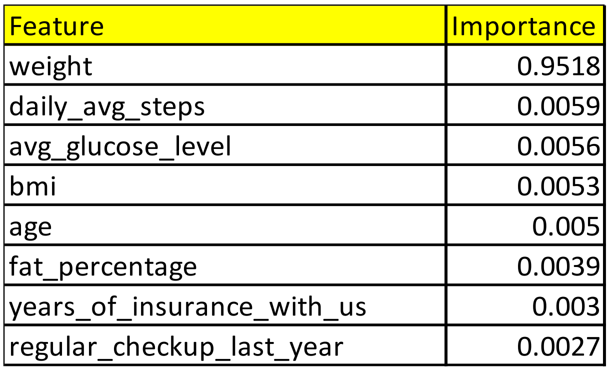
➢ *Feature importances:*

- Gradient boosting regressor



Feature	Importance
weight	0.9954
covered_by_any_other_company	0.0021
regular_checkup_last_year	0.0015
weight_change_in_last_one_year	0.0005
age	0.0002
visited_doctor_last_1_year	0.0009
adventure_sports	0.00005
heart_decs_history	0.00003

- Bagging regressor



Feature	Importance
weight	0.9518
daily_avg_steps	0.0059
avg_glucose_level	0.0056
bmi	0.0053
age	0.005
fat_percentage	0.0039
years_of_insurance_with_us	0.003
regular_checkup_last_year	0.0027

Feedback taken and tested:

- Bagging regressor was rebuilt with more variables that were initially dropped by the linear regression model through higher than 0.05 p-values. This was to check if other variables can be utilized and result in high feature importance alongside 'weight'.

- The result of this exercise has been useful as the Bagging regressor has used few different features like **daily_avg_steps**, **avg_glucose_level**, **bmi**, **fat_percentage** and **years_of_insurance_with_us**. We can rely on this analysis due to the fact that the model validation metrics for Bagging regressor were almost as good as the finalized models.

6. Final interpretation / recommendation

a. Detailed recommendations for the management/client based on the analysis done.

- The best performing model is Elastic Net since it has a great performance not only on the train set but on the test set as well. Recapping its performance:
  - R-squared train = 0.945; R-squared test = 0.944
  - MAPE train = 0.152; MAPE test = 0.155
  - RMSE train = 3368; RMSE test = 3374

- The R-squared value is high on the test set as well. Additionally, the mean absolute percentage error is also not inflated on the test set and is comparable with that of the training set. Therefore, we can assume that the model is stable and will not run into the problem of overfitting.

- **Insurance cost = -79941.93 + (-445.51 * regular_checkup_last_year) + (181.69 * adventure_sports) + (-55.83 * visited_doctor_last_1_year) + (3.77 * age) + (252.86 * heart_decs_history) + (1489 * weight) + (162.66 * weight_change_in_last_one_year) + (1203 * covered_by_any_other_company)**

  - The above equation produced by the Elastic Net model is helpful in understanding the relationship between the independent variables with the dependent variable.

  - Business insights:

    i. If an individual had a checkup last year, then her insurance cost will reduce by 446 units. Individuals that go for regular checkups are less likely to have high insurance costs, given that all the other variables remain constant.

    ii. If an individual is involved in adventure sports, their insurance cost will go up by 182 units, given that all the other variables remain constant. Health risks in individuals who attempt adventure sports increase, leading to higher insurance costs.

    iii. If an individual visited a doctor last year, it can bring down the insurance cost by 56 units, given that all the other variables remain constant. The more regular the checkups, the less chance of bigger health risks since big health problems could have an early detection.

    iv. As an individual gets older, the insurance cost will increase 3.77 units, given that all the other variables remain constant.

    v. If an individual has history of a heart decease, the insurance cost is expected to increase by 253 units, given that all the other variables remain constant.

    vi. A unit increase in weight will increase the insurance cost by 1489 units, given that all the other variables remain constant.

    vii. If a person has experienced weight change in the last year, it could result in an increase of 163 units in the insurance cost, given that all the other variables remain constant.

    viii. If a person is covered by any other company, the insurance cost can have an increase of 1203 units, given that all the other variables remain constant.

Feedback taken and implemented:

- *Visual interpretation for key business insights:*

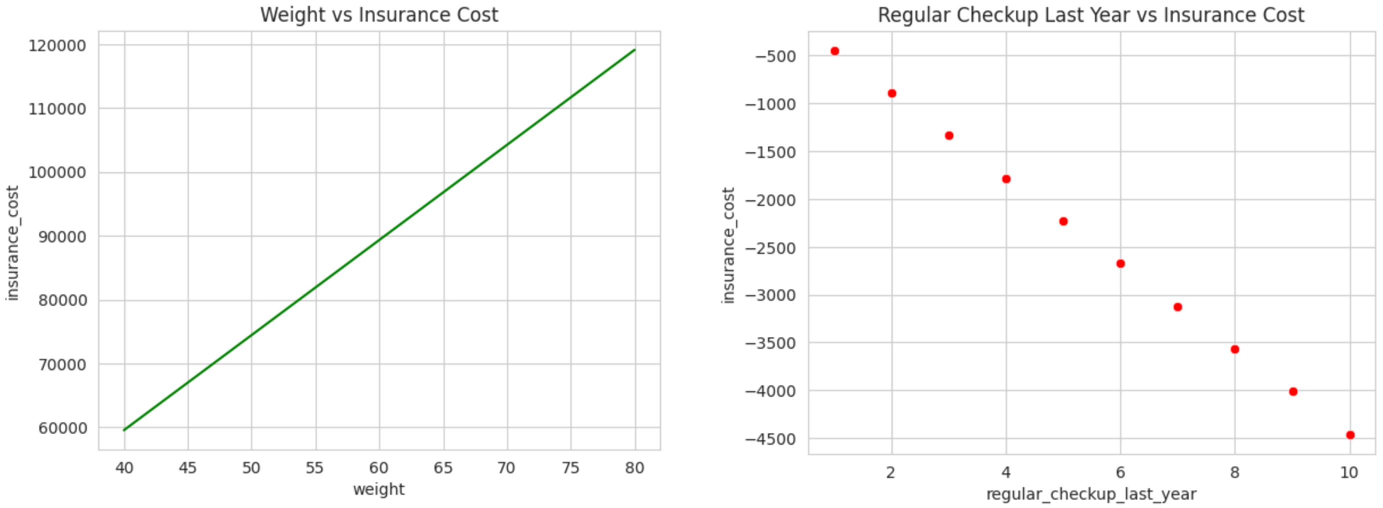
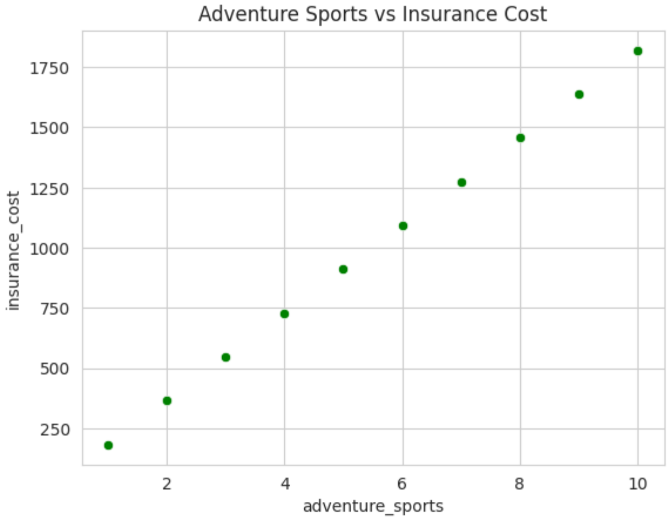
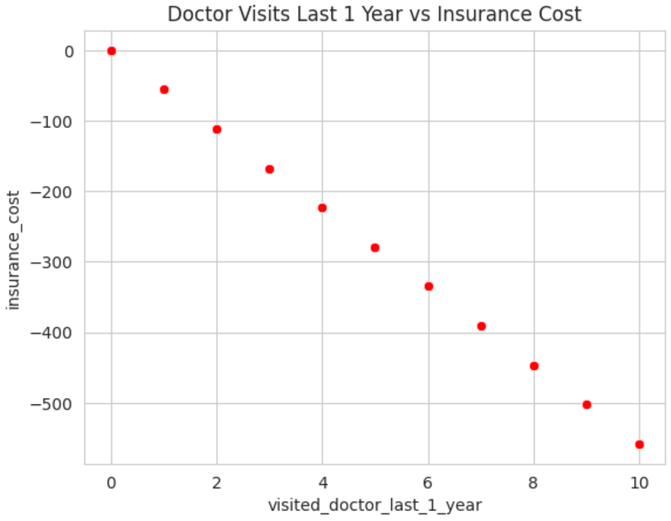
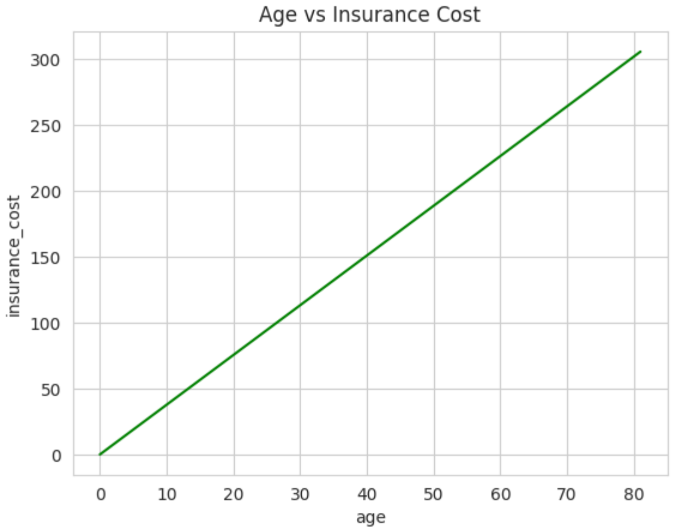
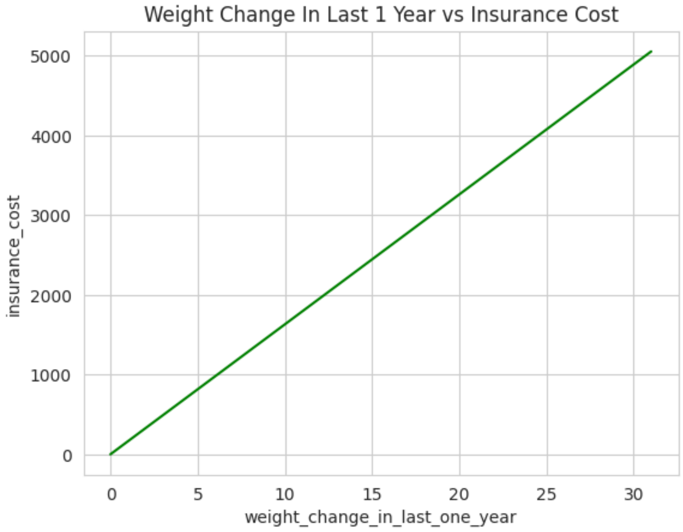


Figure 9 - Visual Interpretation Graphs









END