# A PROJECT PRESENTATION ON

# INSURANCE PRICE PREDICTION


**Presented by:**

Deepak Kumar Saini
Deepali Mathur
Deeptanu Mandal
Dheeban Jayraj R
Digvijay Singh
Faizan arif
Gaurav Singh
Gurmeet Singh

GROUP NO.- 4

# OUTLINE OF PRESENTATION

- Business and Problem Statement
- Literature Review
- Limitations of dataset
- Data Overview
- Initial EDA
- Data cleaning and pre-processing
- EDA
- Model building, Tuning and Validation
- Business Recommendations

# BUSINESS AND PROBLEM STATEMENT


## BUSINESS 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 some times **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.

# BUSINESS AND PROBLEM STATEMENT


## PROBLEM STATEMENT

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

# LITERATURE REVIEW

We did some research on similar projects and here are the insights of our literature review:

https://github.com/MezbanS/Healthcare-Insurance-Analysis/blob/main/Healthcare_Insurance_Analysis_Mezban.html

- History of transplants, HBA1C test score and family history seems to be key influencing variables which have been provided in the dataset, The company can seek to get the data from the customer to make better predictions

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4366801

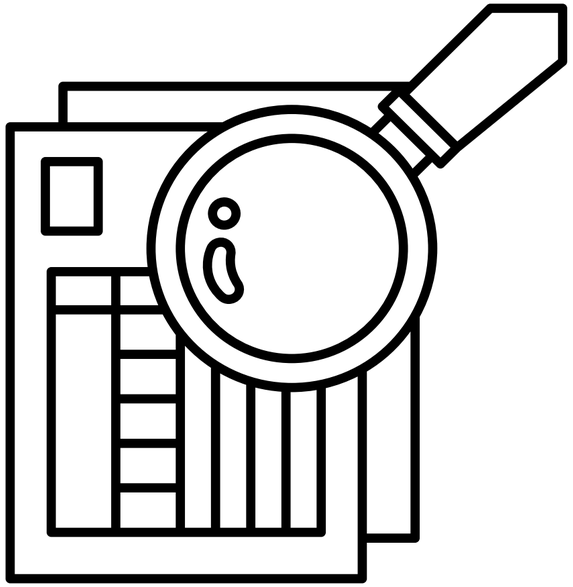
- The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, other factors including geography, marital status, BMI, and family medical history also come into play.

# LIMITATIONS OF DATASET

- Our dataset lacks family history of people which can be considered as a key factor in determining insurance cost.

- Further, the occupation element only has 3 categories namely student, salaried and business. This can be explored further to understand what kind of occupations lead to more hospitalizations. Ex. we can include occupations like doctors, engineers, construction workers, etc.

- Mental health can also be considered as a factor because as per Mental Health Act 2017, it is compulsory to be included in the health insurance coverage.



# DATA OVERVIEW

- Our dataset comprises 25000 rows and 24 columns, with some missing values present. Among the columns, 16 are numerical, and the remaining are object data types.

Upon reviewing the statistical summary of numerical data types, we gained the following insights:

- Nearly 77% of the applicants had a checkup in the previous year, visiting the doctor approximately three times.
- The average glucose level is 167.53, and the BMI ranges from 12.3 to 100.6.
- On average, applicants walk 5216 steps daily. Approximately 8.17% of applicants engage in adventure sports, and the average fat percentage is 28.81%.
- Around 5.46% of applicants have a history of heart disease, and 9.82% have other significant illness. The average weight is 71.61 kg, and the weight change in the last year is 2.52 kg.
- The average insurance cost amounts to 27,147.41 dollars, with a minimum of 2,468 dollars. The last hospital admission year ranges from 1990 to 2018, averaging approximately 2003.89.

# DATA OVERVIEW

The statistical summary of the object datatypes provides several noteworthy insights:

- The majority of participants are students (40.7%) and male (65.7%), primarily residing in Bangalore (7%).
- A significant number of participants have never smoked (55%), with a majority engaging in moderate exercise (58%) and occasional alcohol consumption (55%).
- Cholesterol levels in the range of 150 to 175 are prevalent (35%), with a large number of participants (69%) not covered by any other insurance company.
- Students form the most prominent occupational group (41%), followed by other categories.
- Moderate exercise and rare alcohol consumption are the most common lifestyle habits among participants, indicating a balanced lifestyle. Additionally, "never smoked" is the dominant smoking status among participants.

# INITIAL EXPLORATORY DATA ANALYSIS (EDA)

```
df.duplicated().sum()
```

```
0
```

- There are no duplicate values.

```
# Dropping the "applicant_id" column
df.drop("applicant_id",inplace=True, axis=1)
```

## Checking for missing value

```
df.isna().sum()[df.isna().sum().values>0]
```

```
bmi                    990
Year_last_admitted    11881
dtype: int64
```

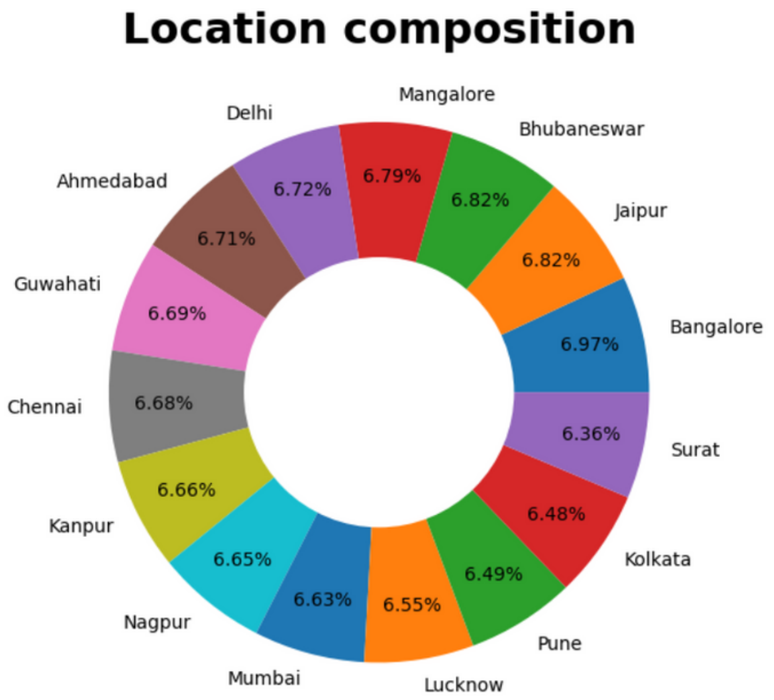
- The `bmi` and `Year_last_admitted` columns have missing values.

```
df['covered_by_any_other_company'].value_counts()
```

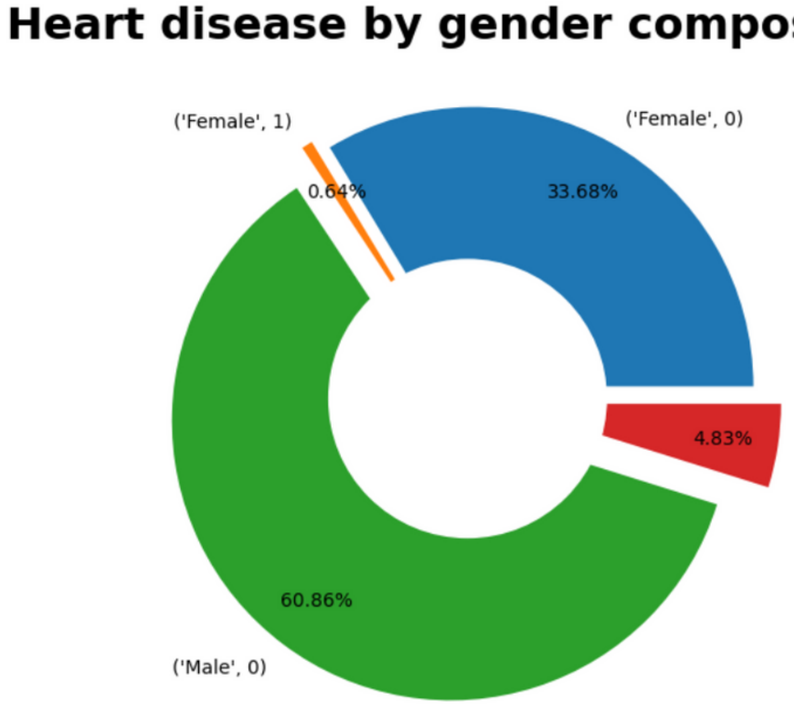
```
covered_by_any_other_company
N    17418
Y     7582
Name: count, dtype: int64
```

- There are 8 catagorical variables.
- There are 15 numerical variables.
- 17418 people are not covered by any other insurance companies
- 7582 people are covered by some other insurance companies

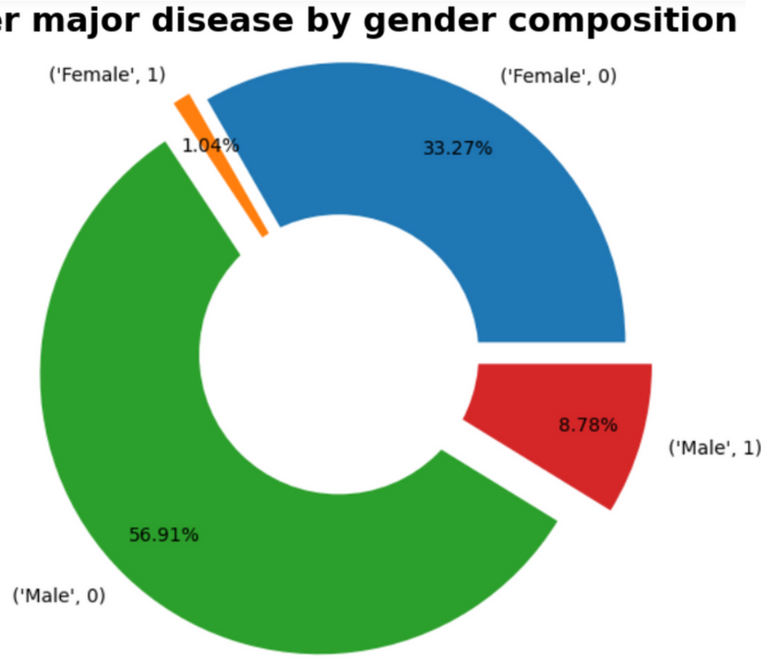
# EDA- PIE PLOTS




- There are total 15 cities in our study. Though majority of the cities have nearly same number of participants, Bangalore has the highest fraction of location composition.

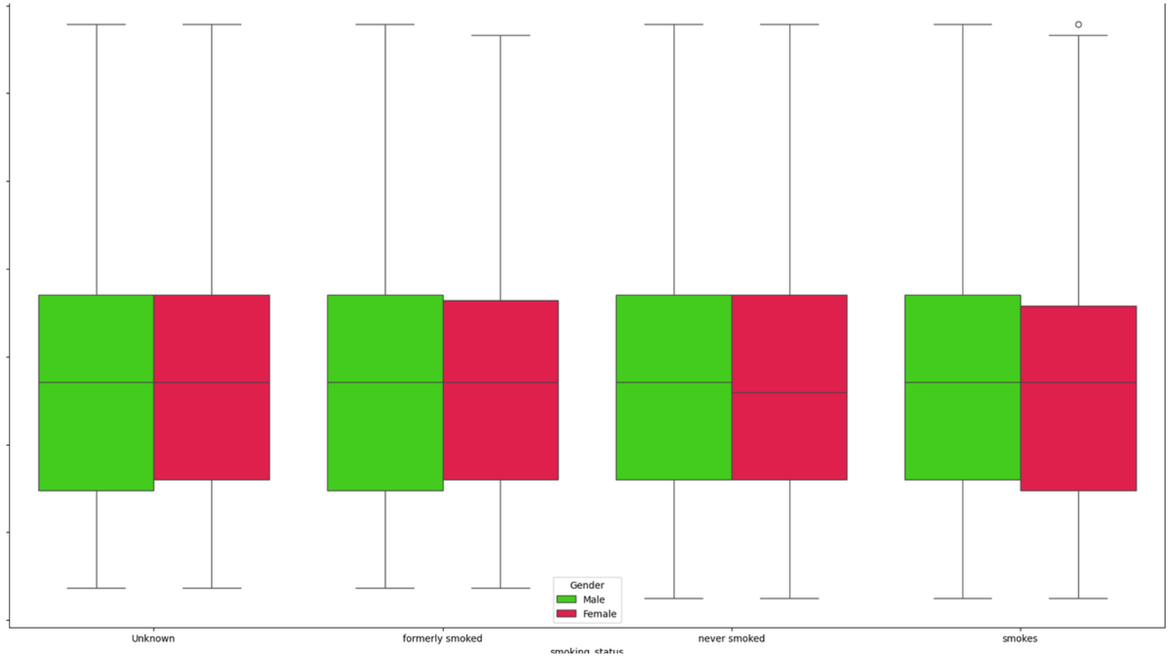



- Out of 65% male participants in study, 4.83% have a history of heart disease. Whereas out of 35% approximate female participants, 0.64% have a history of heart disease.

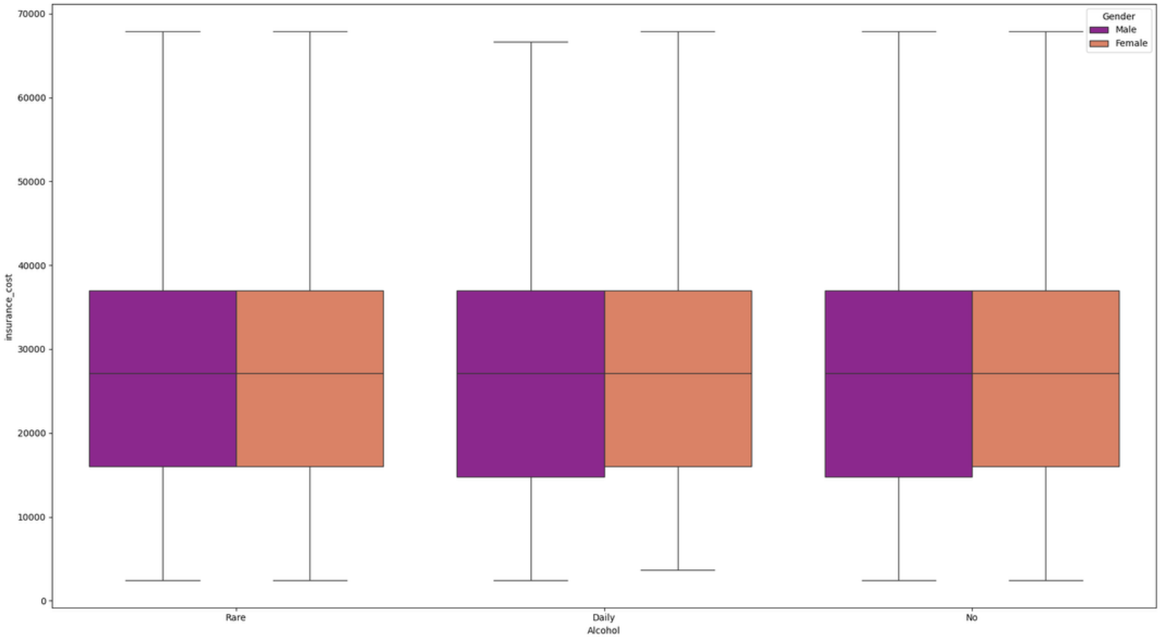



- 8.78% males and 1.04% females have a history of other major diseases. Majority of participants are free from other major diseases.

# BOX PLOTS



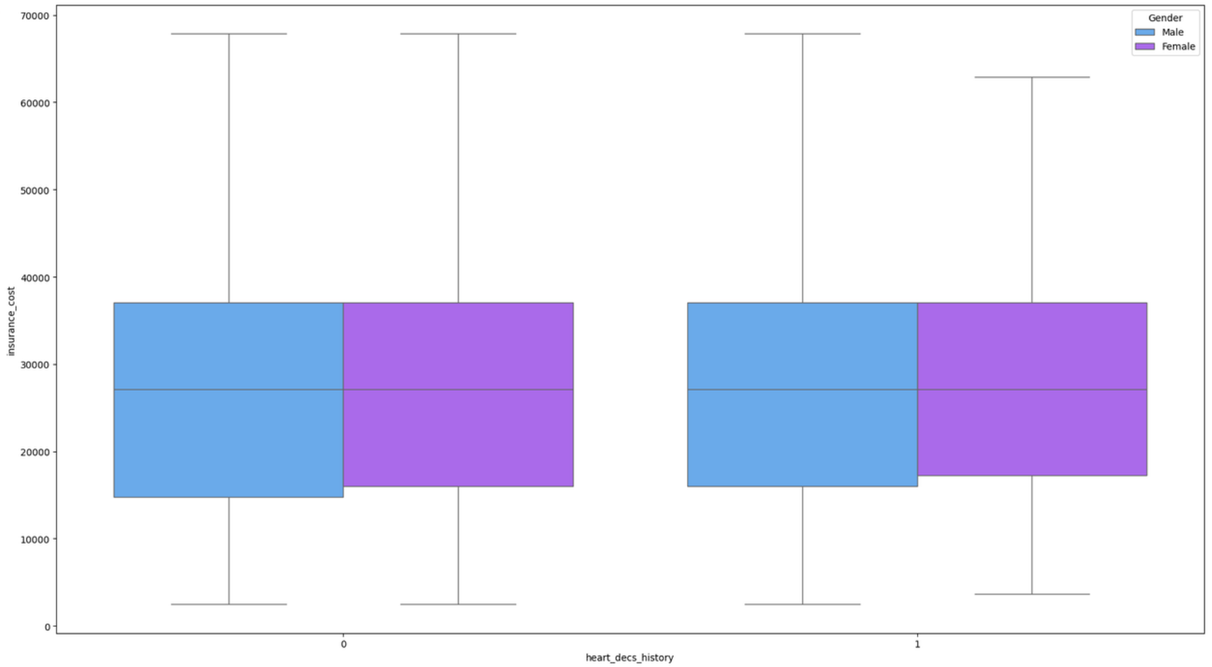

**Insurance Cost and Smoking Status**



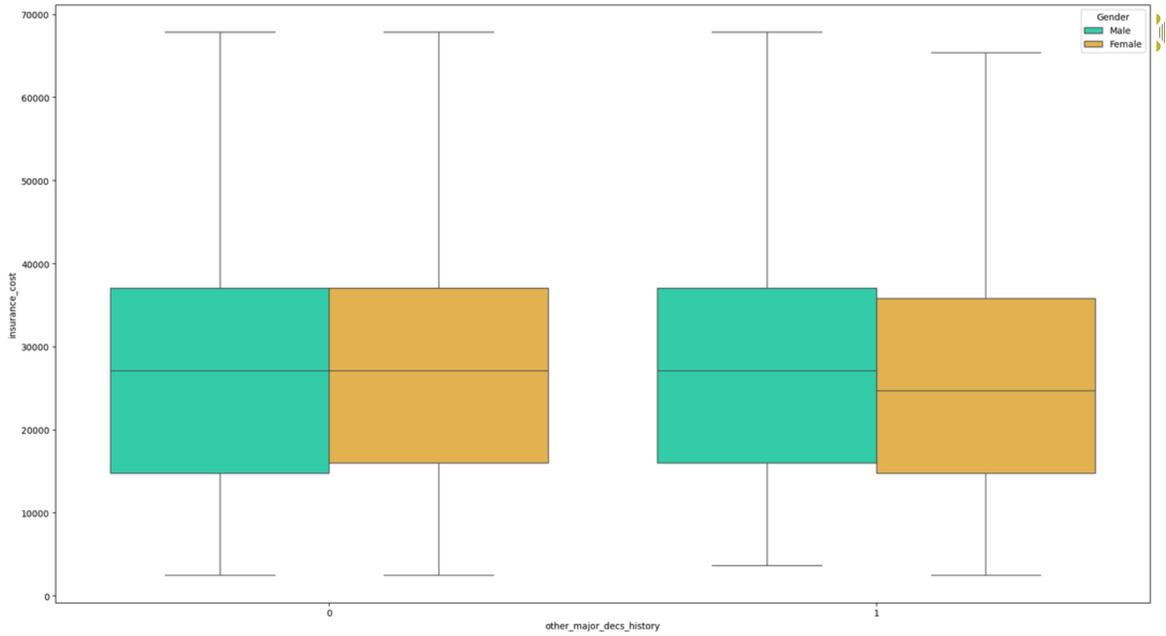

**Insurance Cost and Alcohol Consupmtion**

- Insurance cost remains almost same with respect to smoking status and alcohol consumption by males.

# BOX PLOTS



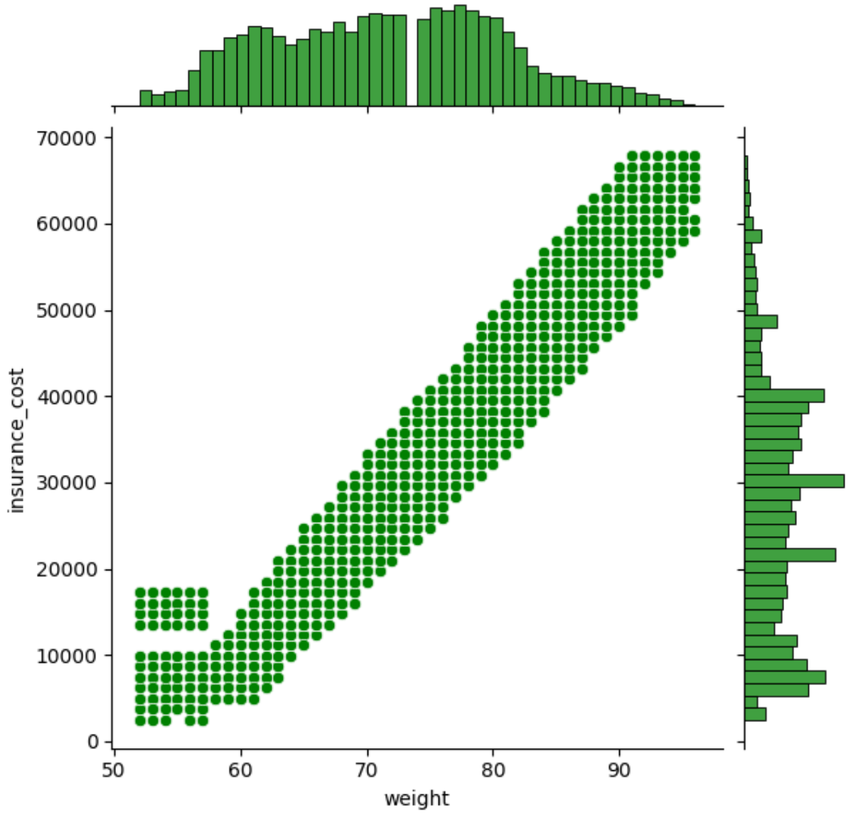

**Insurance Cost and History of Heart Disease**




**Insurance Cost and History of Other Major Disease**

- Insurance cost is also not much affected by history of heart disease. There is a slight decrease in the insurance cost for females with other major disease history. Rest remains the same.

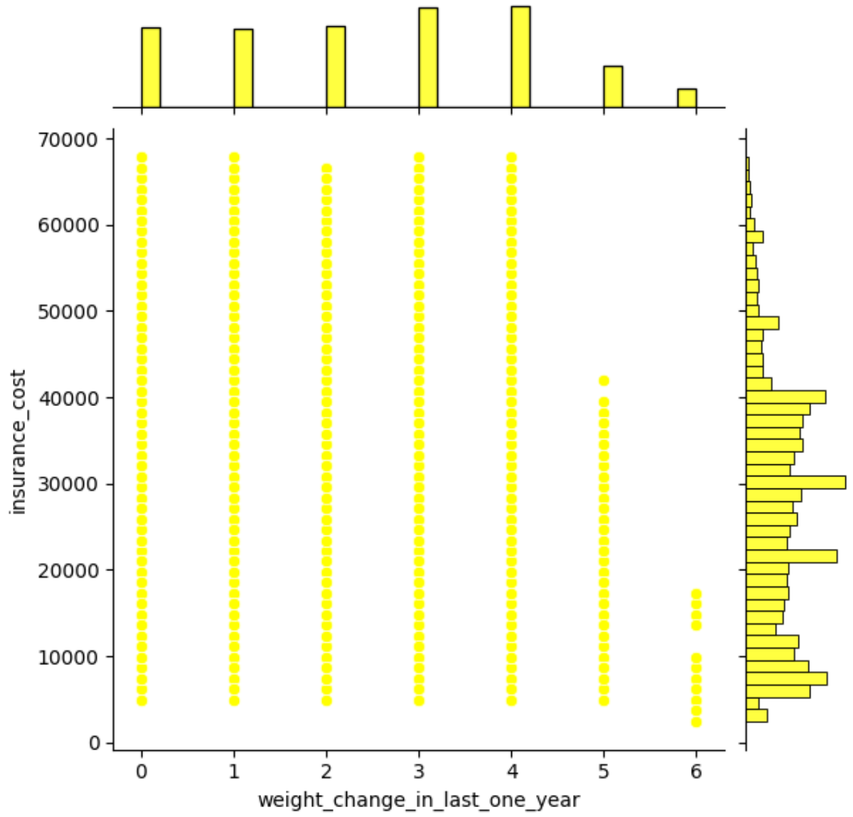
# JOINTPLOTS




## Weight vs insurance cost

We found the insurance cost to increase with weight. This is due to the fact that people with higher weight tend to be more prone to diseases such as heart attack and increased cholestrol level. .

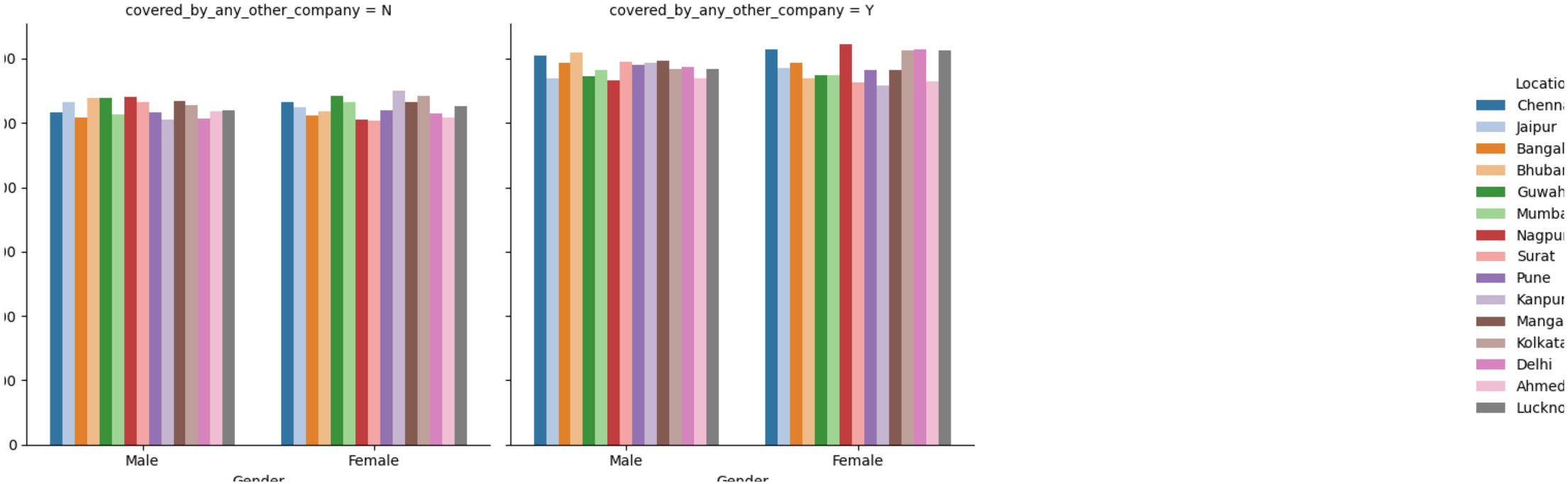
# JOINTPLOTS




## Weight change in last year vs insurance cost

We found the insurance cost to decrease with weight change in last year.This is due to the fact that older people in elderly age tend to be more prone to weight loss.So this particluar set of people are elderly people who dont’t have much life expectancy left decrasing their insurance cost .

# CATPLOT




We can see insurance cost of our company is more than that of other company so the insurance holders can be convinced to have insurance with us.

# PAIRPLOT


1. Number of times visited last year vs insurance cost seems to decrease with increase in checkup frequency.

2. More weight means more insurance cost

3. Weight change in last year and imsurance cost seems to be inversely related.

Points 2 and 3 can be attributed to elderly people/diseased people who visit doctors regulary and prone to weight loss.

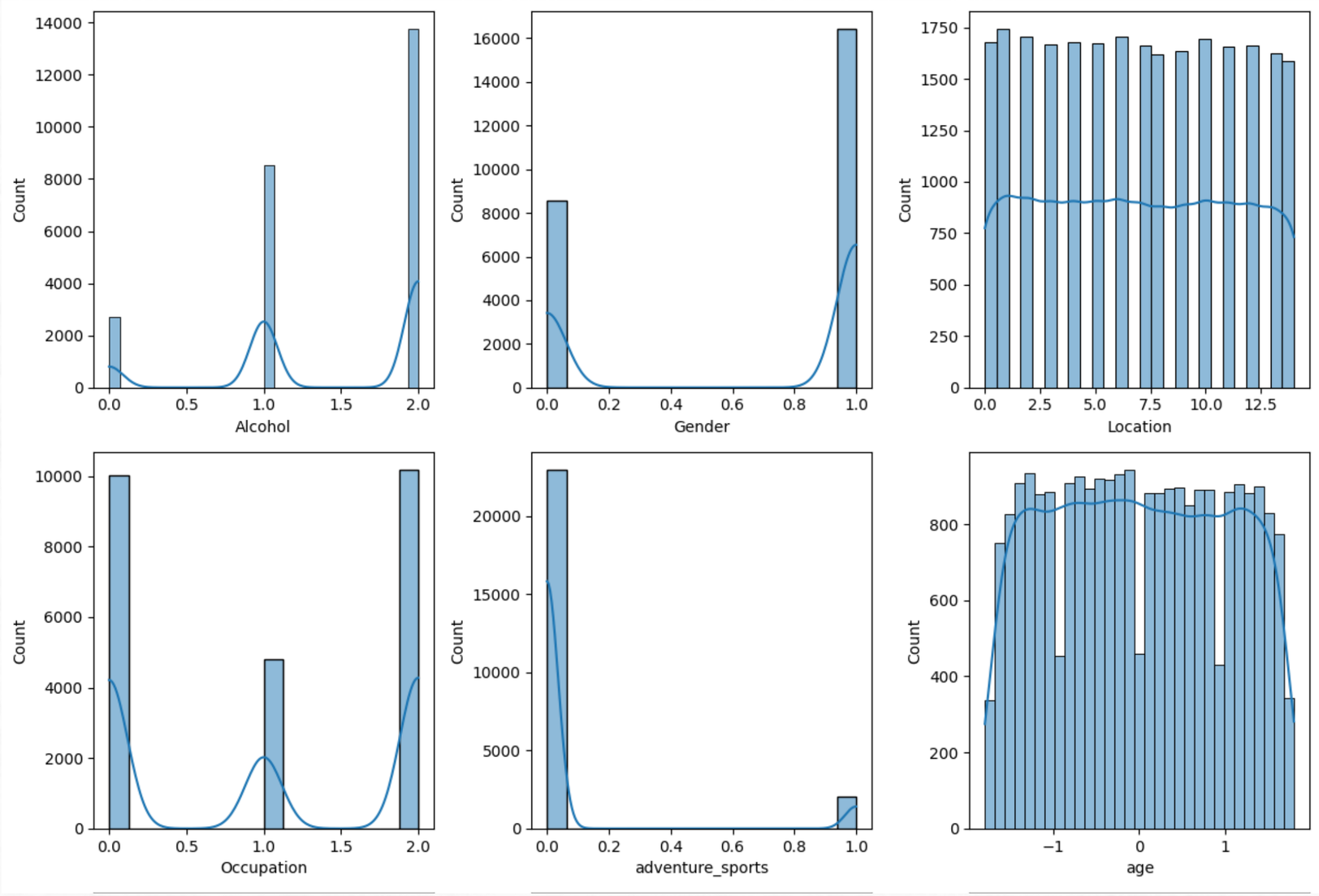
# DATA CLEANING AND PREPROCESSING

- The column “*smoking_status*” has a redundant attribute “***Unknown***” which is included in the “***never smoked***” attribute.
- Cross tabulation of certain variables revealed the following inferences:

1. A higher proportion of females have a 'never smoked' status compared to males while a higher proportion of males have a 'formerly smoked' status compared to females .
2. A higher proportion of males prefer '***Rare***' alcohol consumption compared to females , whereas less females opt for '***No***' alcohol consumption compared to males .
3. A higher prevalence of '***Student***' occupation in both genders , with males slightly outnumbering females . '***Business'*** occupation follows, accounting for 40% of the sample, with more males than females .
4. A significant majority of the sample has no history of other major decisions, with males having a higher proportion than females

- A new column named “*Year_since_last_admitted*” is created in place of “*Year_last_admitted*” by subtracting the latter from 2023 which would make more sense .
- The “*daily_avg_steps*” column is converted into a categorical attribute by binning the values into ‘L’ , ‘M’ and ‘H’ divisions.

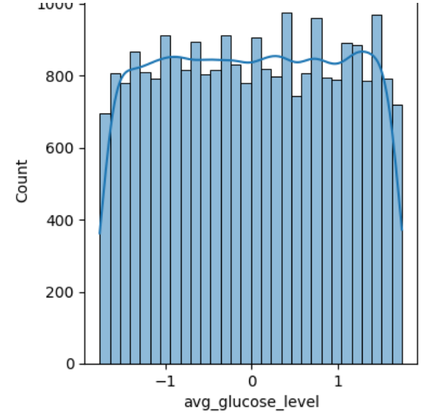
# EXPLORATORY DATA ANALYSIS (EDA)

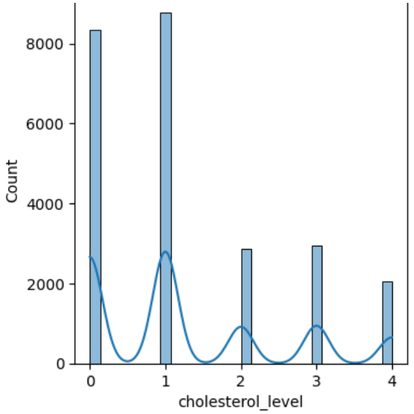


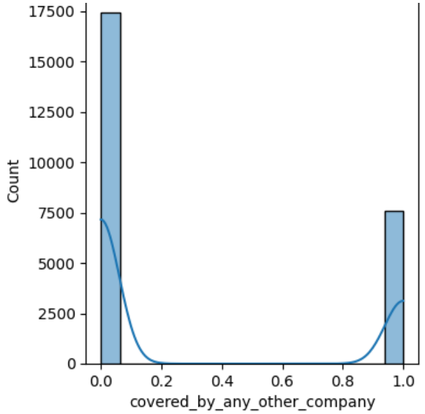

EDA after preprocessing

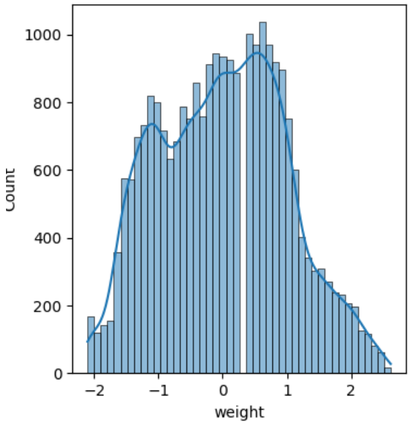
- All the correlation and insights that have been made for the variables are much clearer after preprocessing.

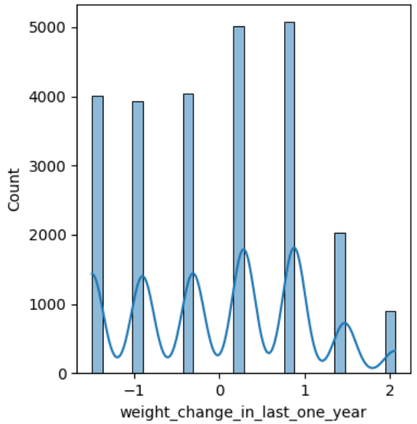
# EXPLORATORY DATA ANALYSIS (EDA)

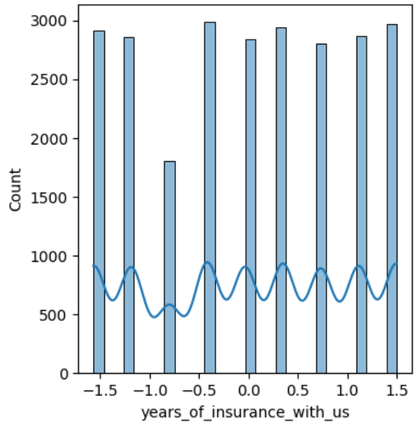


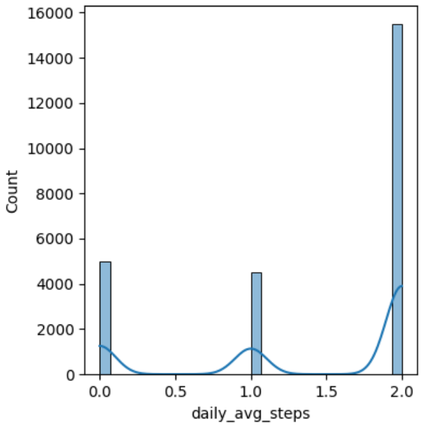



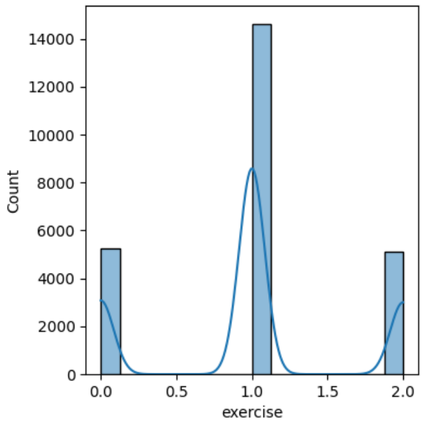



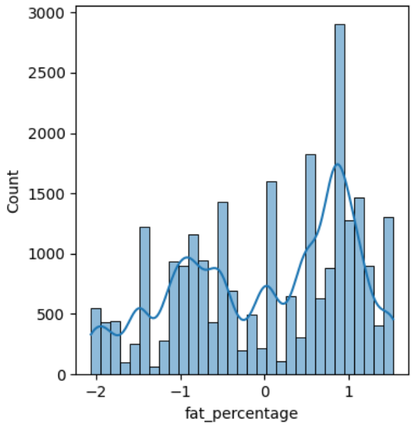



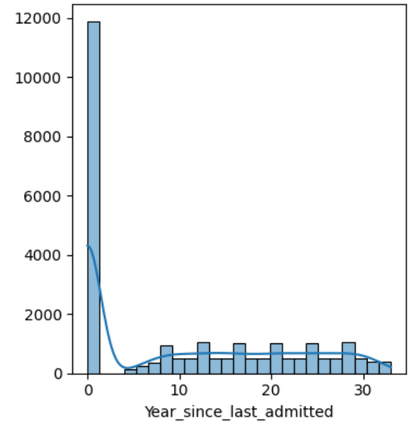



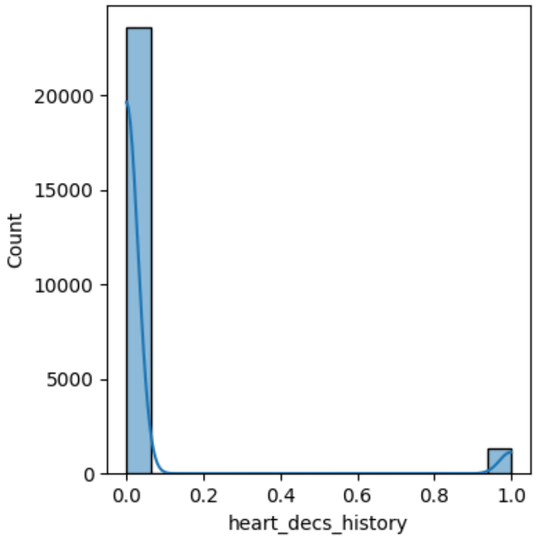



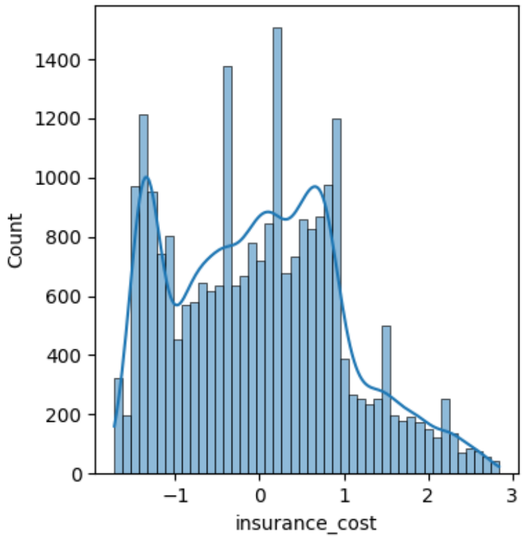



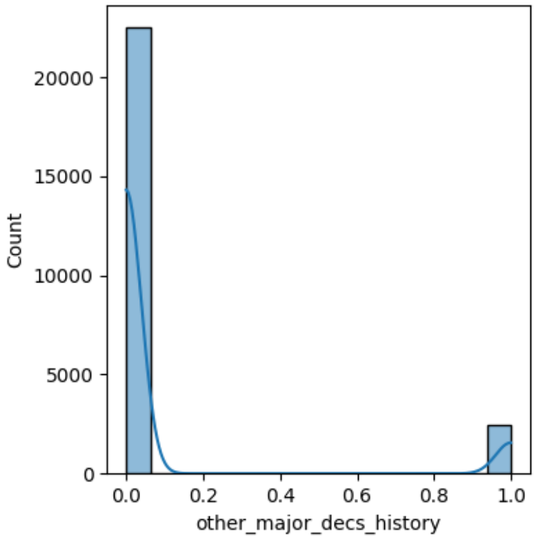



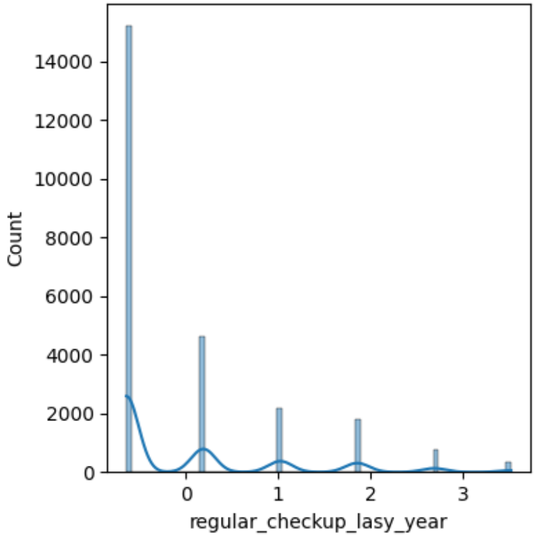

We have higher number of people in the higher fat percentage zone and higher number of people who don’t have any other insurance, the company should capitalise on this demand.

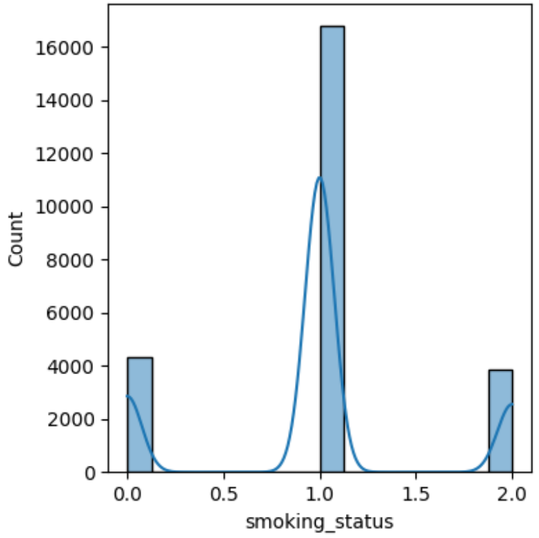
# EXPLORATORY DATA ANALYSIS (EDA)

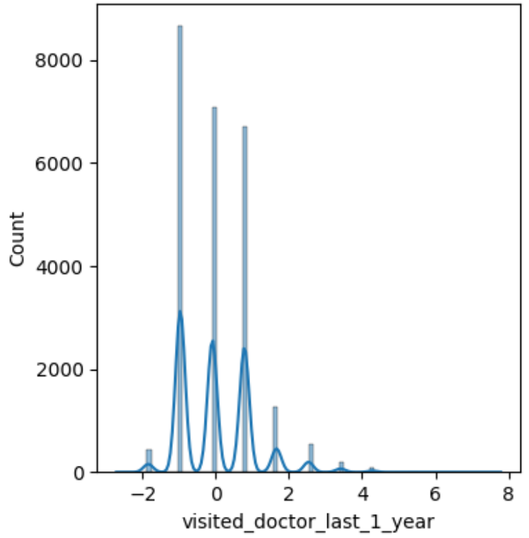



- Most customers don’t visit doctors frequently and don’t take regular check ups which might be harmful as underlying health conditions can go undetected , therefore the company must suggest their to visit doctors frequently and do checkups to reduce risks.

# MODEL BUILDING, TUNING AND VALIDATION

## Model Building

Since The problem is a regression type problem , We have decided to test the following algorithms to see which one gave us the best results and insights

- Linear Regression
- ANN Regression
- Random Forest Regressor
- Decision tree regressor
- Gradient Boost
- Ada Boost

## Model Tuning

We Tried to tune 3 different models on our data:

- Grid search for Decision Tree
- Grid search on Random Forest
- Grid search on ANN

# MODEL BUILDING, TUNING AND VALIDATION

## Model Validation

After Tuning the hypermeters, These are the best scores that we got from each model:

| Model | RMSE | R-squared |
|---|---|---|
| ANN Regressor | 0.26 | 0.93 |
| Decision Tree | 0.301 | 0.908 |
| Random Forest | 0.21 | 0.953 |

# BUSINESS INSIGHTS AND RECOMMENDATIONS

**Significant Variables:**

- years_of_insurance_with_us: This variable has a negative coefficient, indicating that as the number of years with the insurance company increases, the insurance cost decreases. It is statistically significant (p-value < 0.05), suggesting that it plays a significant role in determining insurance costs.

**Business Insight:** Rewarding loyal customers with lower premiums can be a strategy.

- regular_checkup_last_year: It has a negative coefficient, suggesting that individuals who had regular checkups last year tend to have lower insurance costs. It is statistically significant.

**Business Insight:** Encouraging regular checkups can lead to cost savings for both customers and insurance companies.

# BUSINESS INSIGHTS AND RECOMMENDATIONS

weight: This variable has a highly positive coefficient, indicating that weight has a **significant impact on insurance costs.**

**Business Insight:** Encouraging weight management and a healthy lifestyle can be a strategy to reduce claims and insurance costs.

covered_by_any_other_company: This variable has a positive coefficient and is highly statistically significant. It suggests that individuals covered by another insurance company have higher insurance costs.

**Business Insight:** Understanding the nature of other insurance coverage and aligning policies can be explored to reduce costs.

# BUSINESS INSIGHTS AND RECOMMENDATIONS

**Insignificant Variables:**

adventure_sports: This variable has a p-value > 0.05, indicating it's not statistically significant in predicting insurance costs. Business Insight: It may not be a strong predictor of insurance costs.

Occupation, daily_avg_steps, heart_decs_history, other_major_decs_history, Gender, smoking_status, Location, Alcohol, exercise, fat_percentage: These variables have coefficients and p-values suggesting that they are not statistically significant in predicting insurance costs. Business Insight: These variables may not play a crucial role in determining insurance costs.

# FUTURE STRATEGIES FOR HEALTH INSURANCE PRICE PREDICTION:

Based on the analysis of significant variables and insights, the following strategies can be considered for health insurance price prediction:

- **Customer Loyalty Programs**: Continue to reward loyal customers who have been insured with the company for multiple years with lower premiums.

- **Promote Preventive Care**: Encourage policyholders to have regular checkups and lead a healthy lifestyle. This can potentially reduce insurance costs in the long run.

- **Weight Management**: Implement programs or incentives to help policyholders manage their weight and maintain a healthy BMI, which can lead to lower insurance costs.

- **Analyze Competing Insurance**: Investigate the impact of coverage by other insurance companies and consider adjusting pricing strategies accordingly.

# FUTURE STRATEGIES FOR HEALTH INSURANCE PRICE PREDICTION:

**Review and Optimize Additional Variables**: While some variables may not be significant in the current model, it's important to periodically review and analyze their impact as the dataset evolves. New variables or data sources may become relevant in the future.

**Customized Pricing**: Consider offering customized insurance packages based on individual health profiles, as some variables may be significant for certain segments of customers.

**Data Quality**: Ensure the quality and completeness of data. Address missing or inconsistent data to improve the accuracy of predictions.

It's essential to remember that the business and pricing strategies should be developed in collaboration with domain experts and data scientists to balance the needs of the company and the welfare of policyholders. Additionally, regularly updating the model and strategies based on new data and market trends is critical for long-term success in the health insurance domain.

WE WANT TO SAY

# THANK YOU

FOR YOUR ATTENTION