UNIVERSITY OF TEXAS ARLINGTON

Data Analysis & Modelling Techniques

Insurance Cost Prediction

Fall 2022- Project group 2

Sravanisree Mangala
Mary Pranavi Allam
Charan Sai Kondapaneni
Jithin Krishna Kongara

Project Goal

DATASET

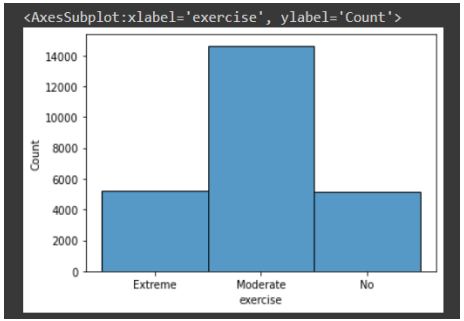
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    25000 non-null  int64
 1   years_of_insurance_with_us      25000 non-null  int64
 2   regular_checkup_lasy_year       25000 non-null  int64
 3   adventure_sports                25000 non-null  int64
 4   Occupation                      25000 non-null  object
 5   visited_doctor_last_1_year      25000 non-null  int64
 6   cholesterol_level               25000 non-null  object
 7   daily_avg_steps                 25000 non-null  int64
 8   age                             25000 non-null  int64
 9   heart_decs_history              25000 non-null  int64
 10  other_major_decs_history        25000 non-null  int64
 11  Gender                          25000 non-null  object
 12  avg_glucose_level               25000 non-null  int64
 13  bmi                             24010 non-null  float64
 14  smoking_status                  25000 non-null  object
 15  Year_last_admitted              13119 non-null  float64
 16  Location                        25000 non-null  object
 17  weight                          25000 non-null  int64
 18  covered_by_any_other_company    25000 non-null  object
 19  Alcohol                         25000 non-null  object
 20  exercise                        25000 non-null  object
 21  weight_change_in_last_one_year  25000 non-null  int64
 22  fat_percentage                  25000 non-null  int64
 23  insurance_cost                  25000 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

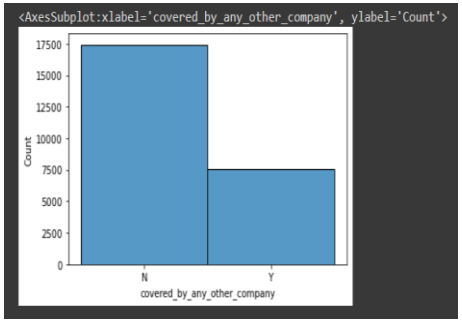
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   years_of_insurance_with_us      25000 non-null  int64
 1   regular_checkup_lasy_year       25000 non-null  int64
 2   adventure_sports                25000 non-null  int64
 3   Occupation                      25000 non-null  int64
 4   visited_doctor_last_1_year      25000 non-null  int64
 5   cholesterol_level               25000 non-null  category
 6   daily_avg_steps                 25000 non-null  int64
 7   age                             25000 non-null  int64
 8   heart_decs_history              25000 non-null  int64
 9   other_major_decs_history        25000 non-null  int64
 10  Gender                          25000 non-null  int64
 11  avg_glucose_level               25000 non-null  int64
 12  bmi                             24010 non-null  float64
 13  smoking_status                  25000 non-null  int64
 14  Year_last_admitted              13119 non-null  float64
 15  weight                          25000 non-null  int64
 16  covered_by_any_other_company    25000 non-null  int64
 17  Alcohol                         25000 non-null  int64
 18  exercise                        25000 non-null  int64
 19  weight_change_in_last_one_year  25000 non-null  int64
 20  fat_percentage                  25000 non-null  int64
 21  insurance_cost                  25000 non-null  int64
dtypes: category(1), float64(2), int64(19)
memory usage: 4.0 MB
```

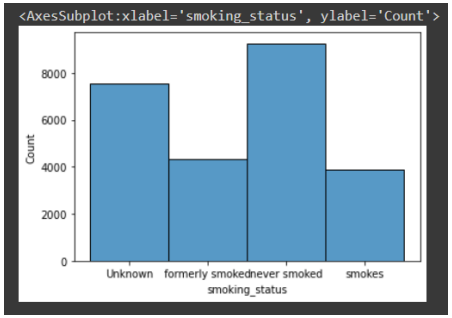
A. The dataset contains 4188 rows and 24 columns.

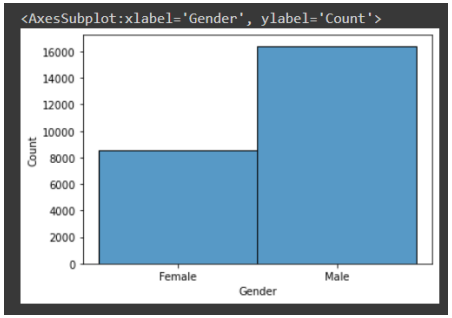
B. After removing outliers

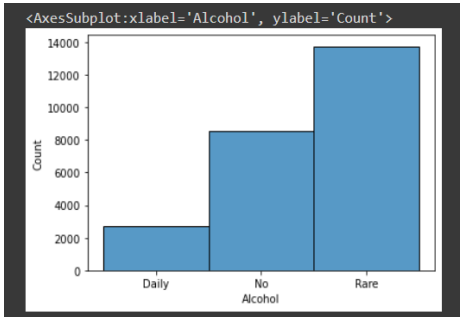
DATA VISUALIZATION

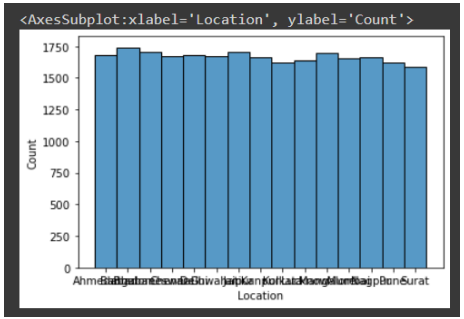


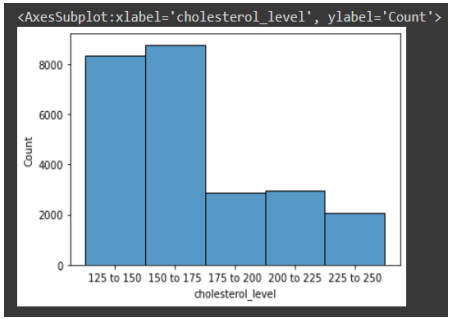


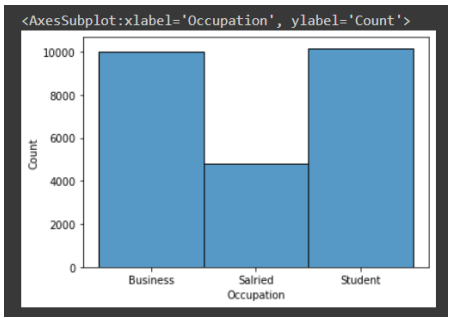




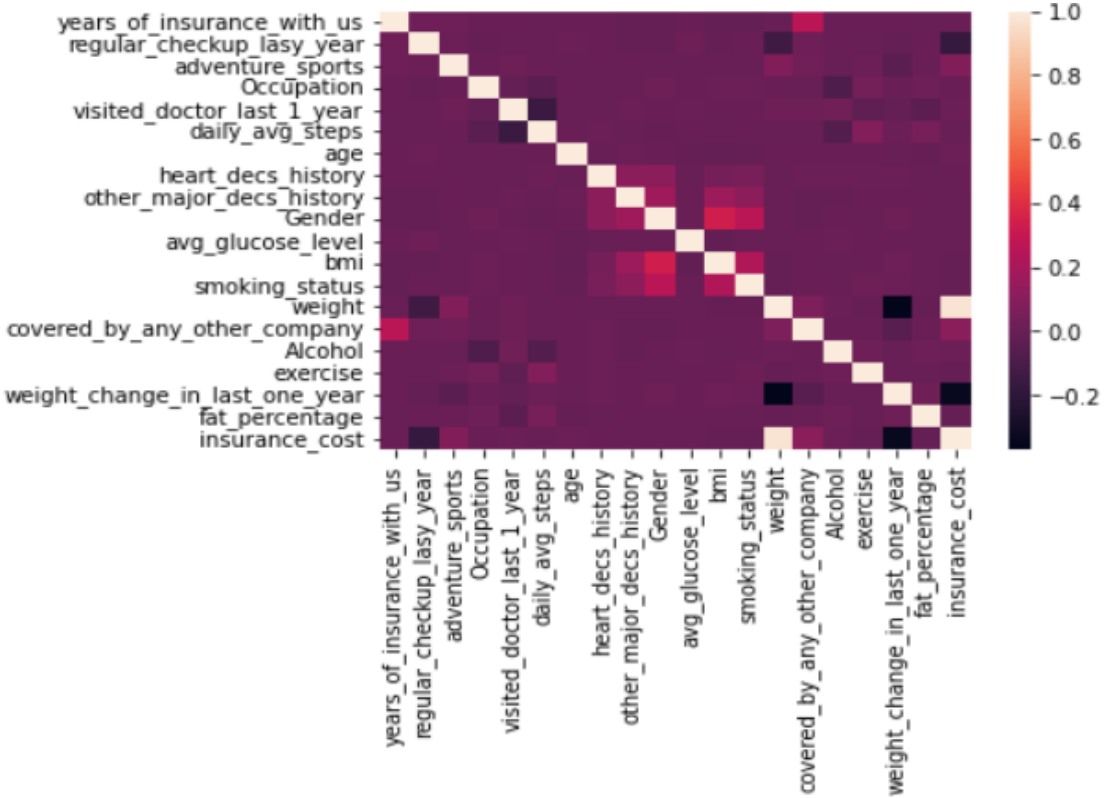




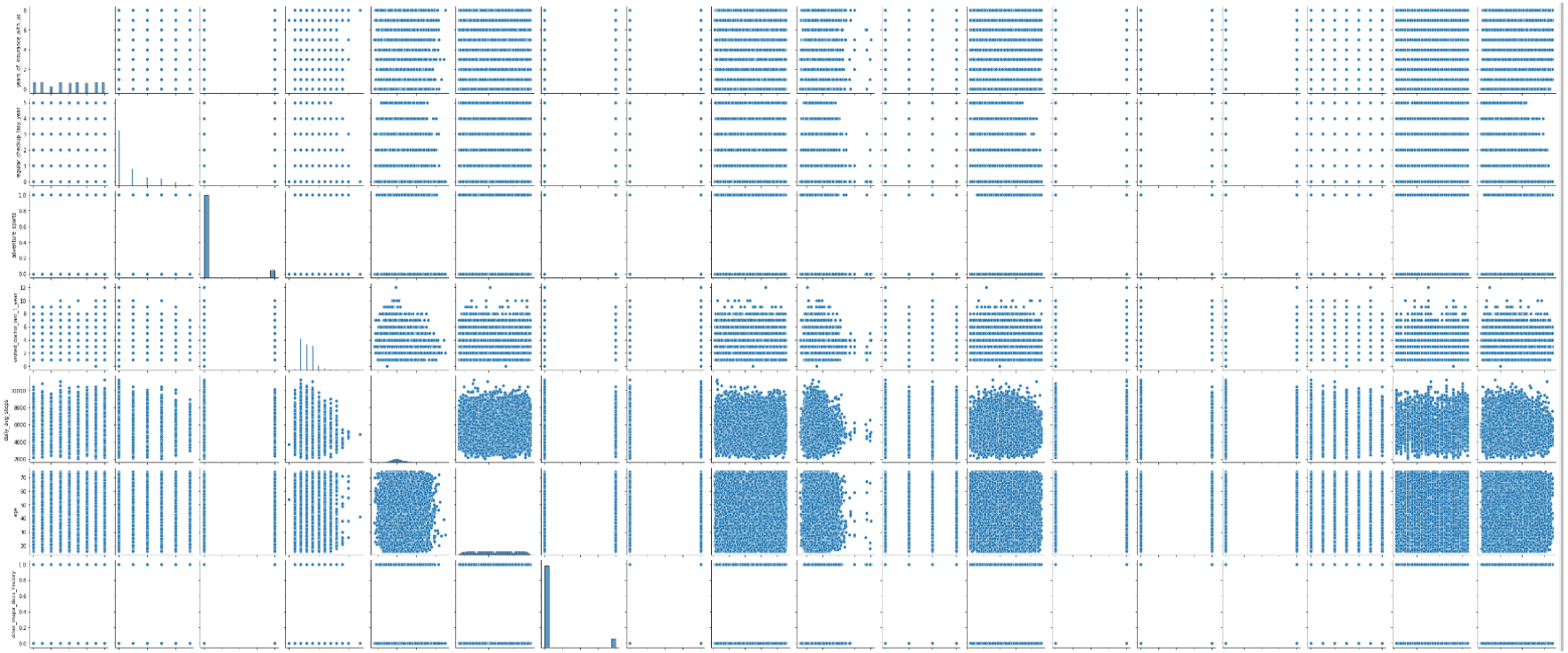




CORRELATION MATRIX



PAIRPLOT



UNIVERSITY OF TEXAS ARLINGTON

THANK YOU