greatlearning
Power Ahead

Capstone Presentation

PGPDSBA.O.OCT23.A - Health Insurance Project

Anirudh Sardiwal

Table of Contents

1.0 Business Problem Understanding

1.1 Problem Statement

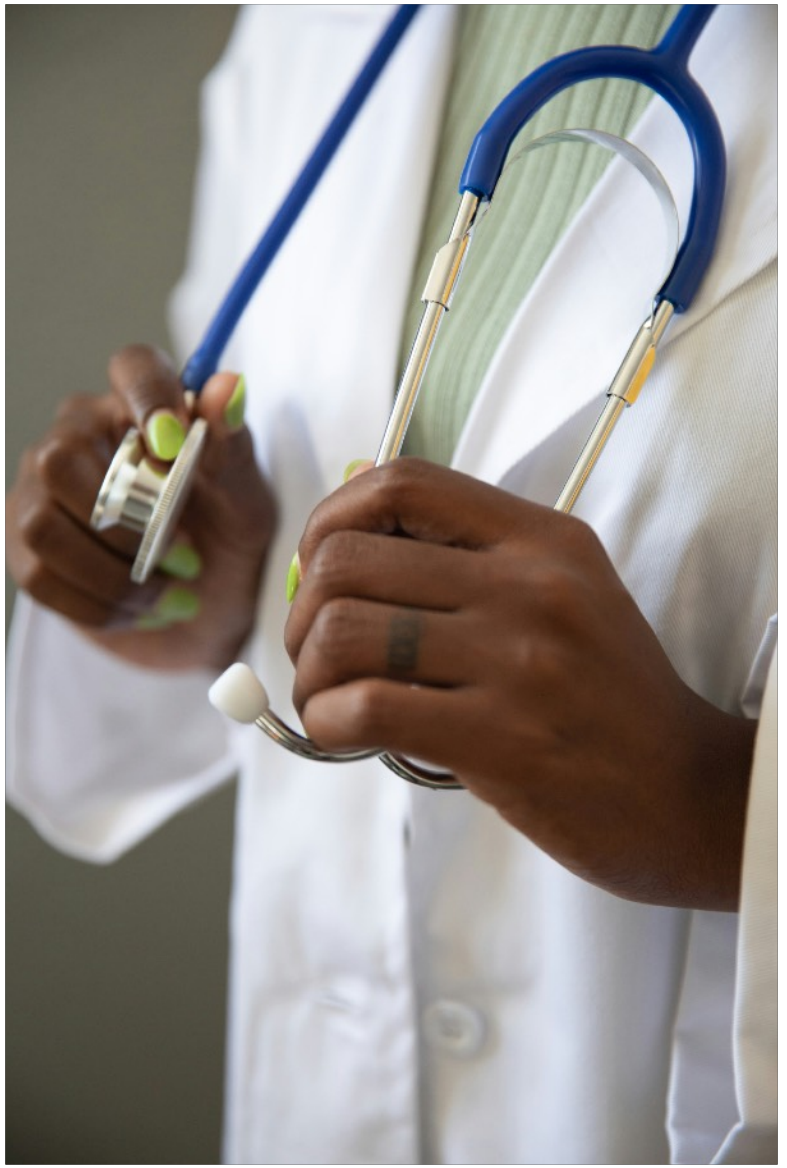
- Managing medical expenses for a layperson can become challenging, therefore health insurance is necessary in modern life.
- Medical insurance companies need to get insurance cost just right to avoid customer churn or revenue loss.

1.2 Objective

To develop a model that predicts the optimal insurance cost for an individual, based on their health and lifestyle habits.

1.3 Need of the Project

The need for the project is because insurance companies need to get the cost just right and need to see how lifestyle habits impact this cost, as majority diseases these days are lifestyle diseases.



1.4 Data

The data has 25,000 rows and 23 columns.

Numerical Features

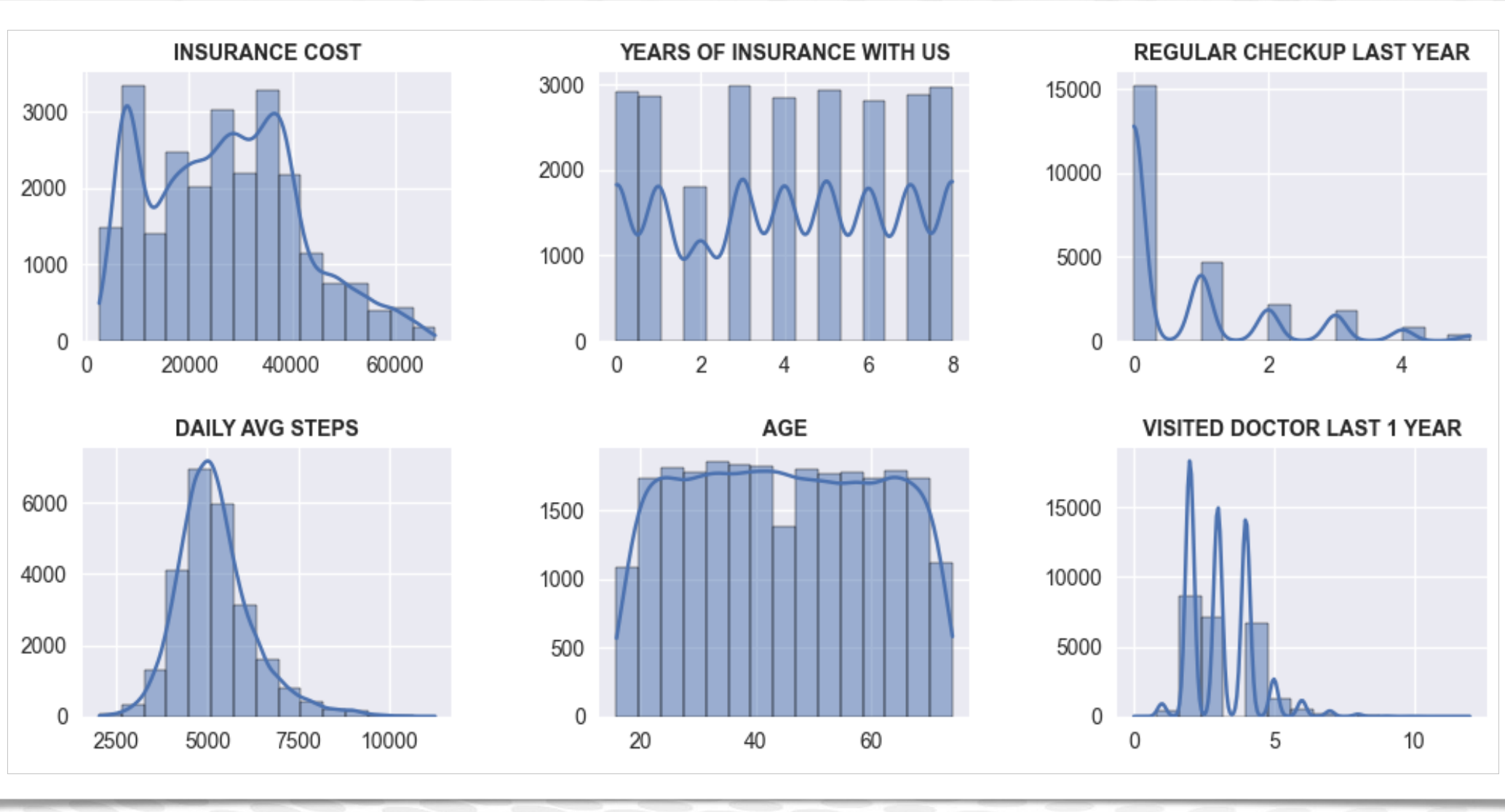
1. insurance_cost
2. years_of_insurance_with_us
3. regular_checkup_last_year
4. daily_avg_steps
5. age
6. visited_doctor_last_1_year
7. avg_glucose_level
8. bmi
9. weight
10. weight_change_in_last_one_year
11. fat_percentage

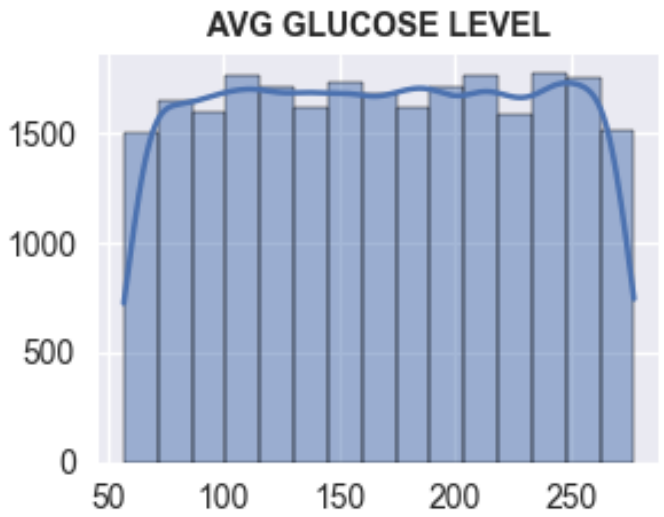
Categorical Features

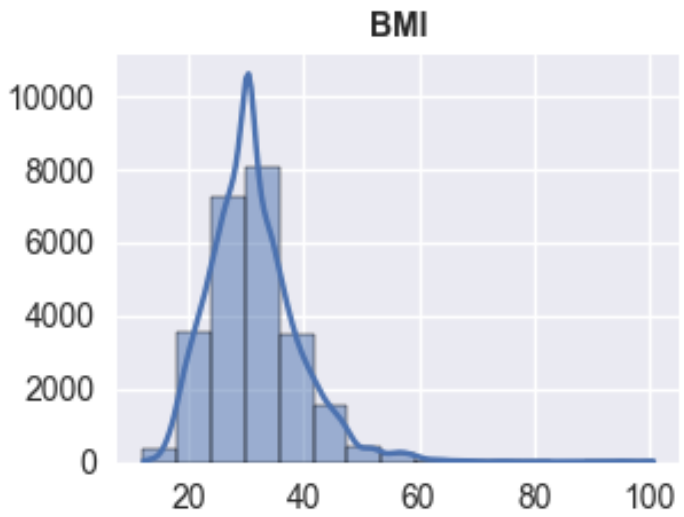
1. adventuresports
2. Occupation
3. cholesterollevel
4. heartdecshistory
5. othermajordecshistory
6. Gender
7. smokingstatus
8. LocationZone
9. coveredbyanyothercompany
10. Alcohol exercise

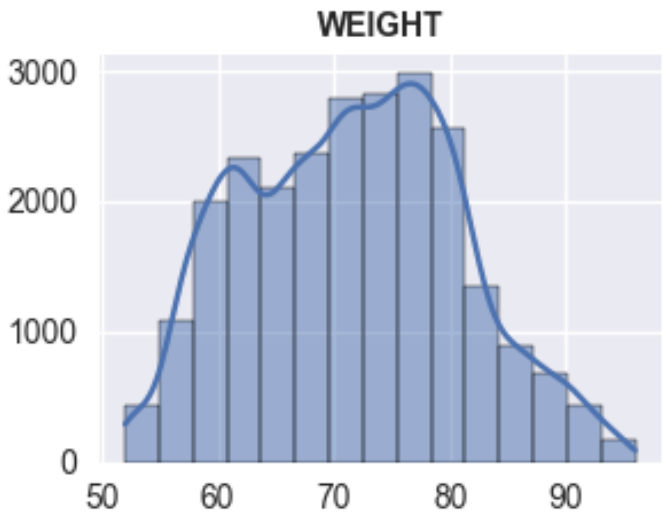
2.0 EDA

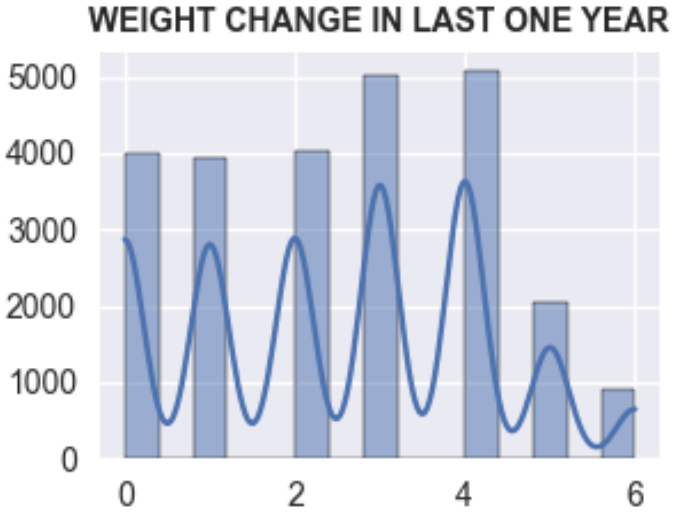
2.1 Numerical Variables

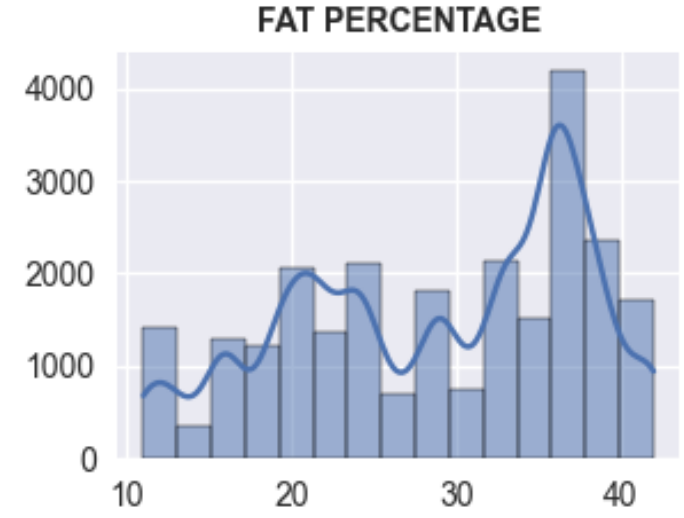



AVG GLUCOSE LEVEL
1500
1000
500
0
50
100
150
200
250


BMI
10000
8000
6000
4000
2000
0
20
40
60
80
100

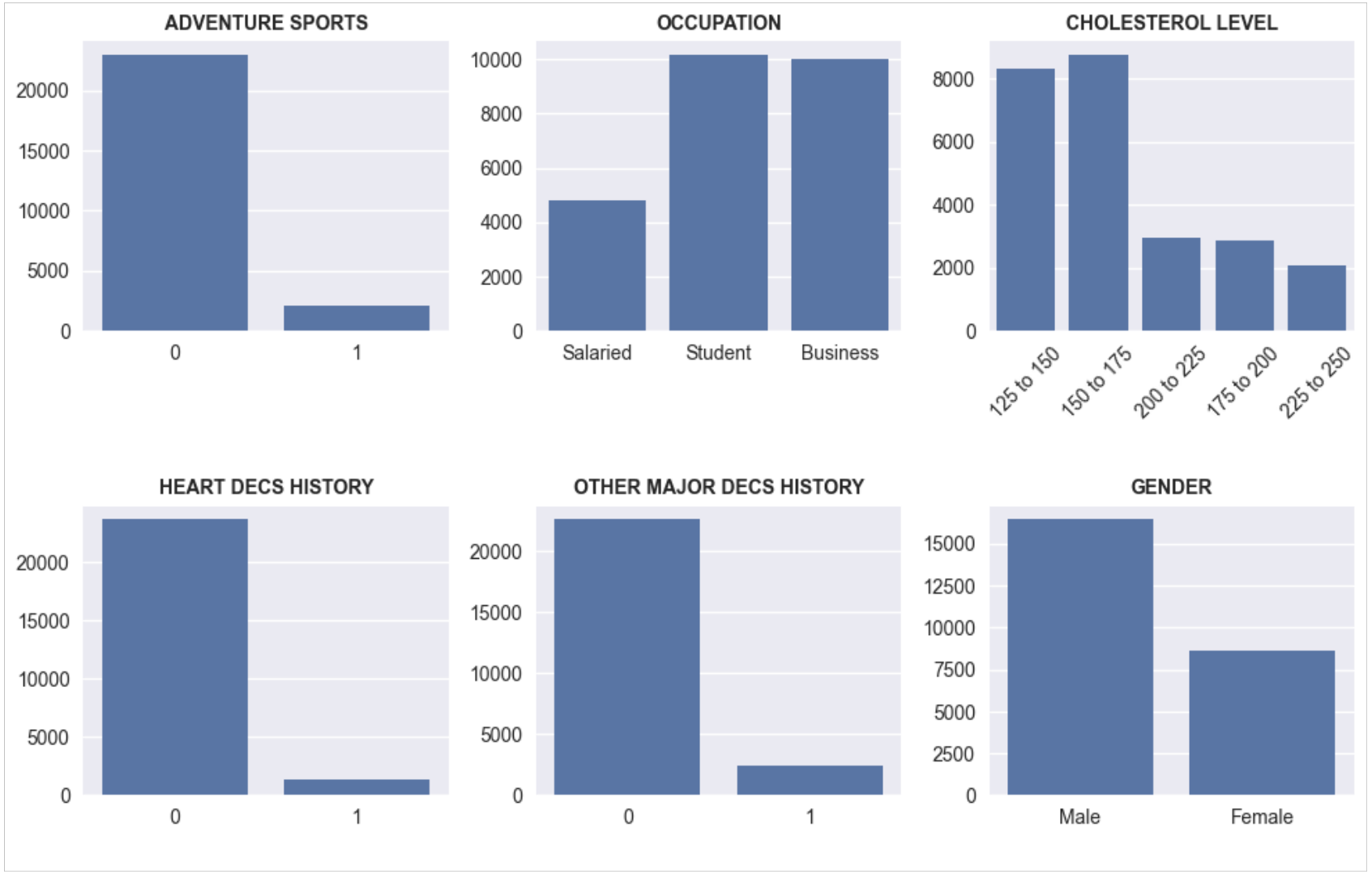

WEIGHT
3000
2000
1000
0
50
60
70
80
90

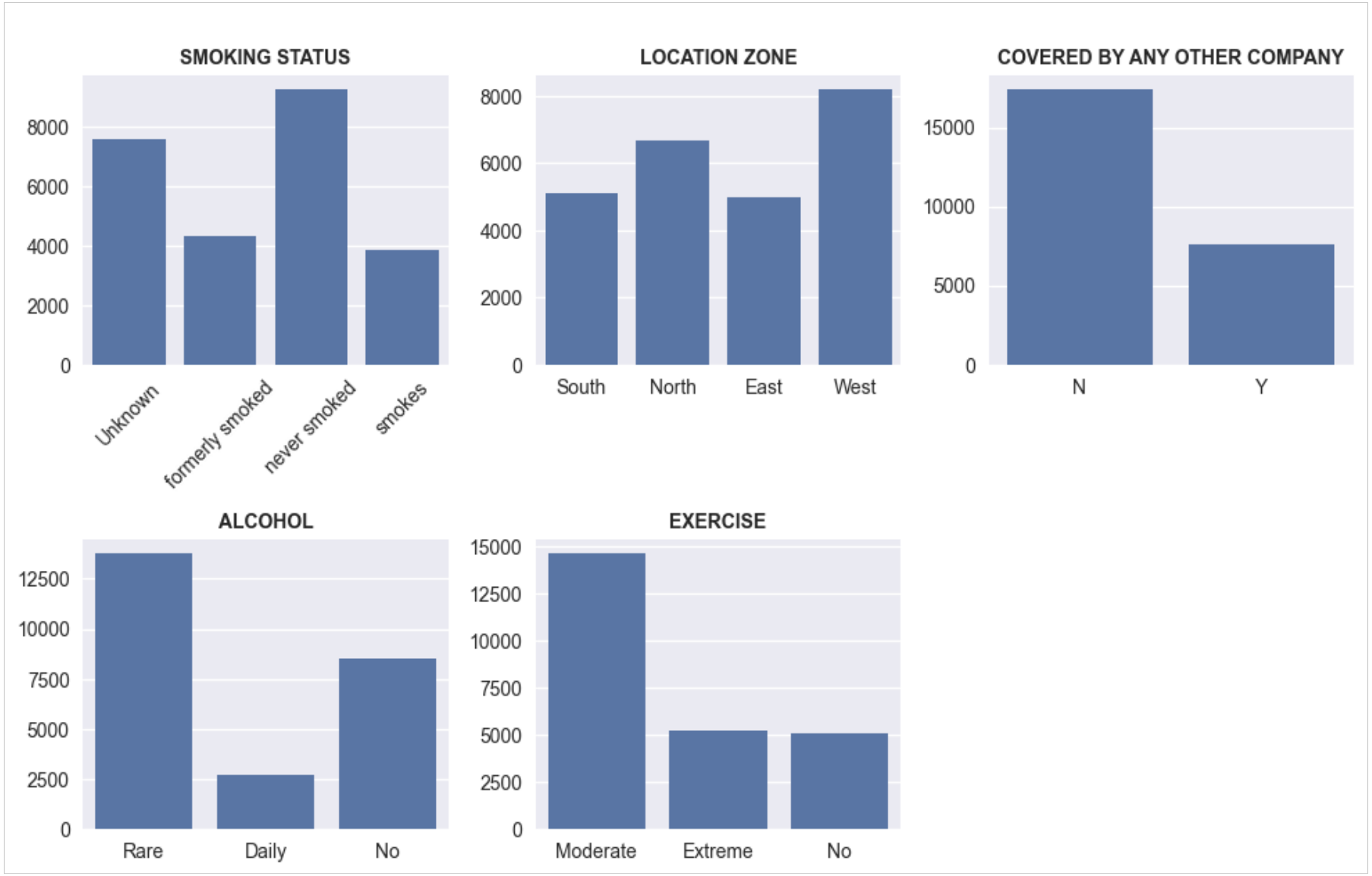

WEIGHT CHANGE IN LAST ONE YEAR
5000
4000
3000
2000
1000
0
0
2
4
6


FAT PERCENTAGE
4000
3000
2000
1000
0
10
20
30
40

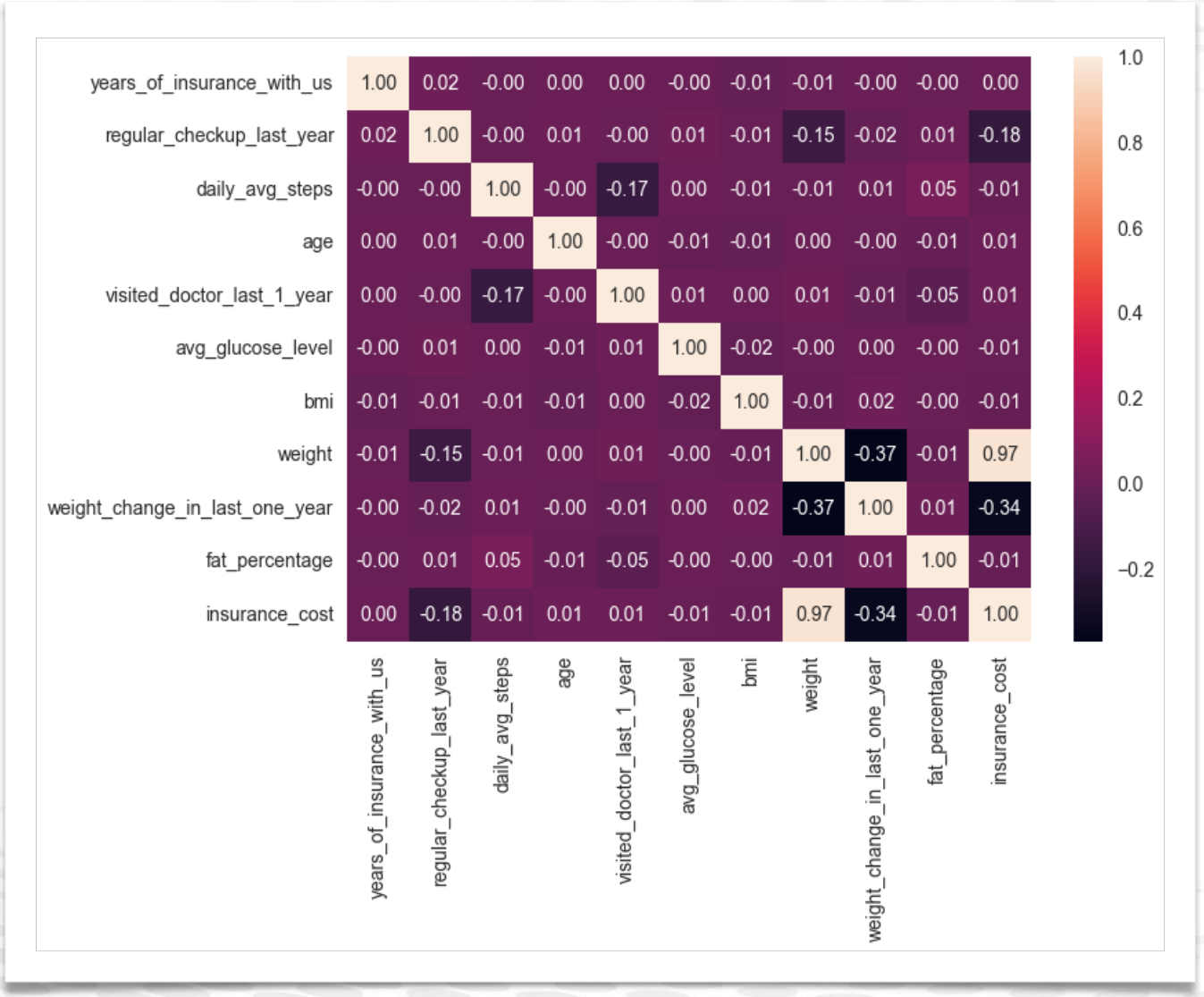
Back

2.2 Categorical Variables

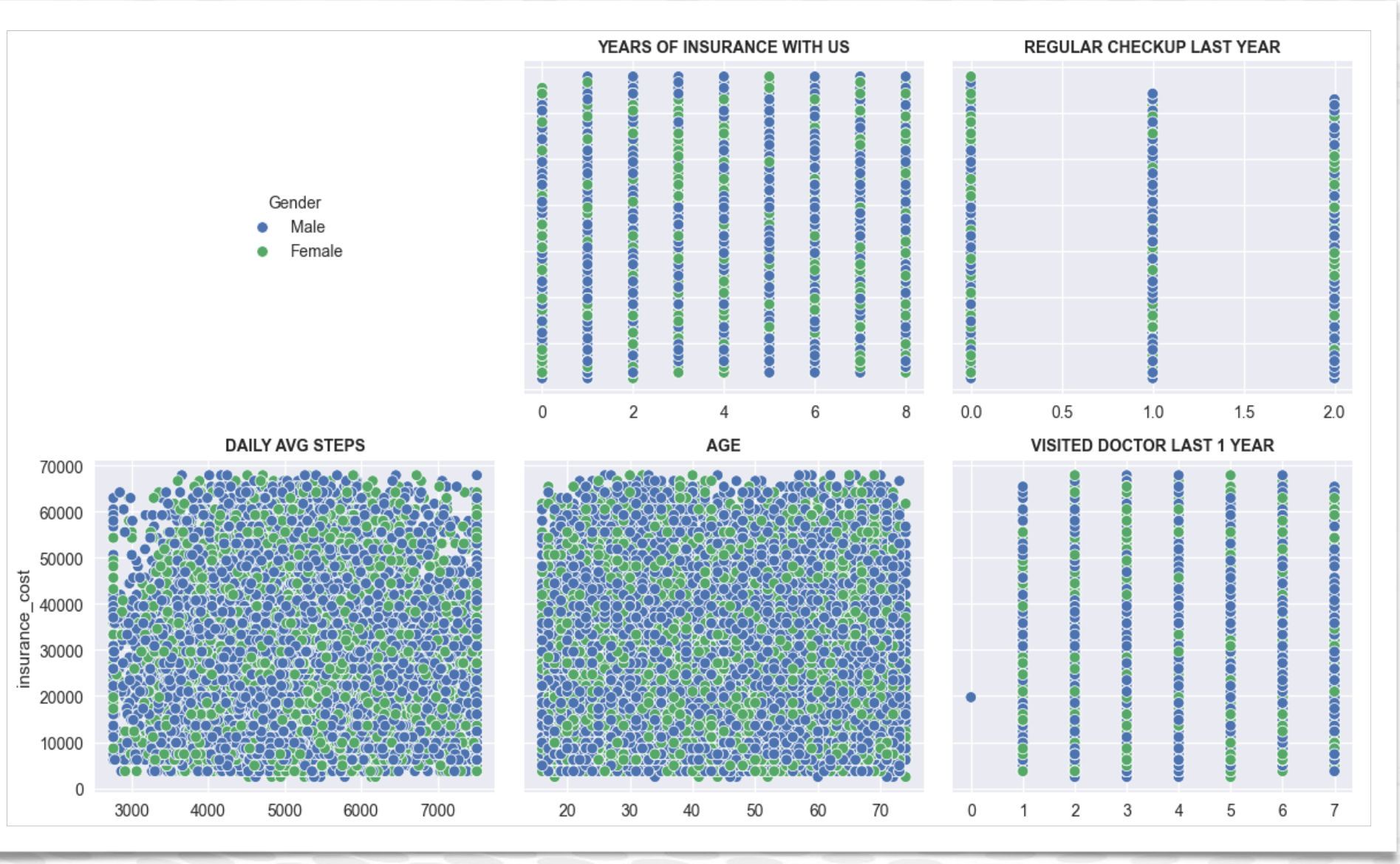


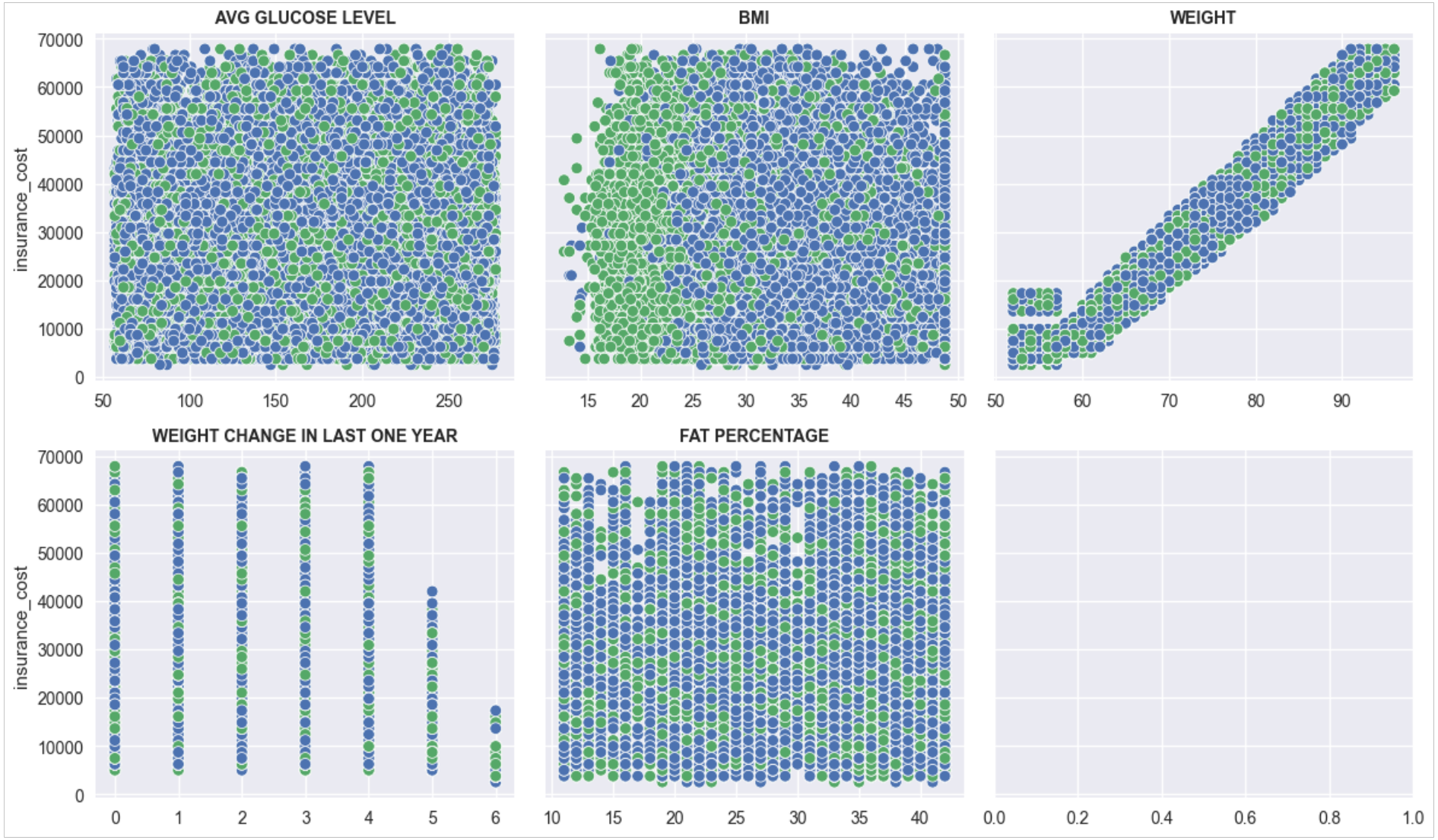

SMOKING STATUS
8000
6000
4000
2000
0
Unknown
formerly smoked
never smoked
smokes
LOCATION ZONE
8000
6000
4000
2000
0
South
North
East
West
COVERED BY ANY OTHER COMPANY
15000
10000
5000
0
N
Y
ALCOHOL
12500
10000
7500
5000
2500
0
Rare
Daily
No
EXERCISE
15000
12500
10000
7500
5000
2500
0
Moderate
Extreme
No

2.3 Correlation

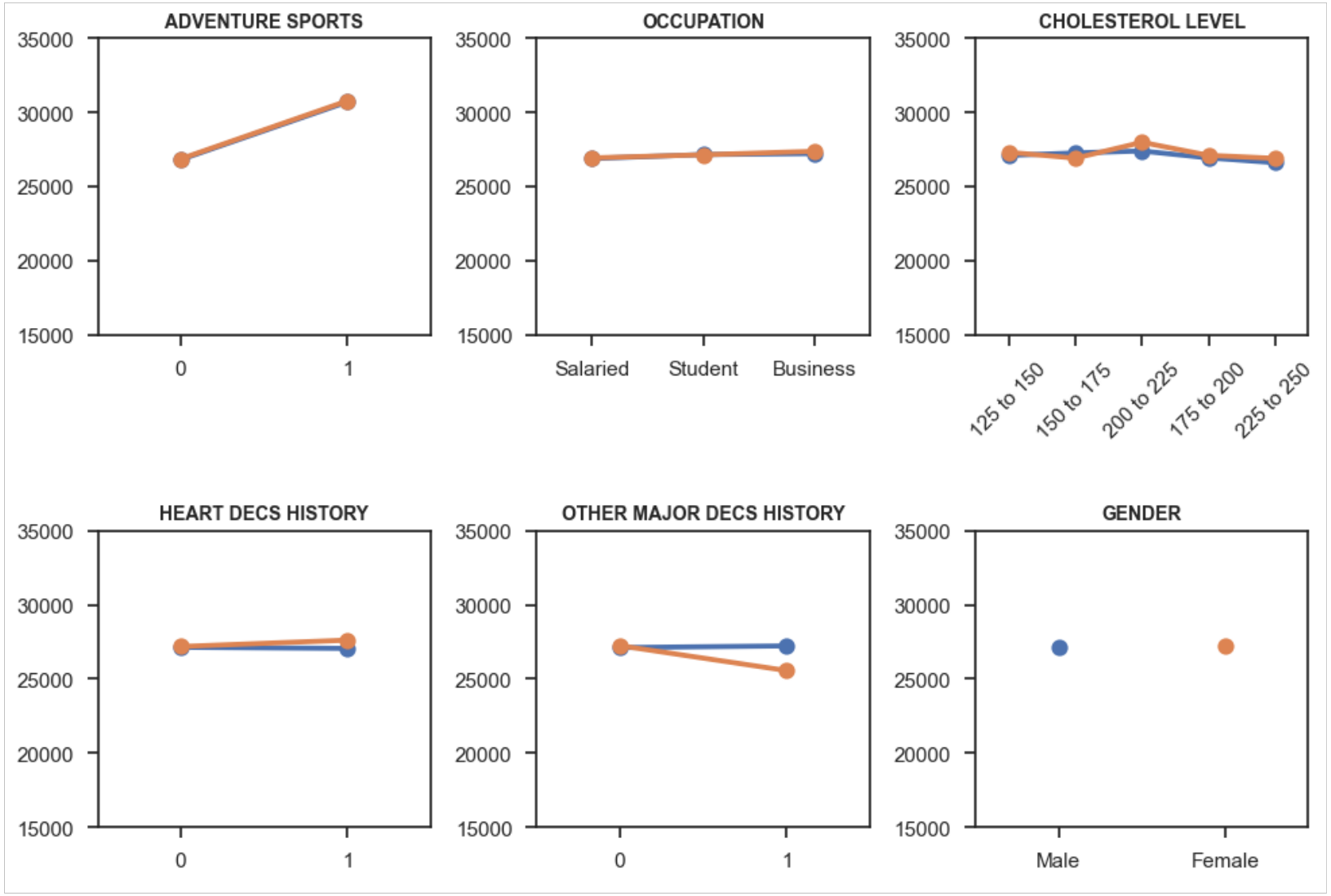


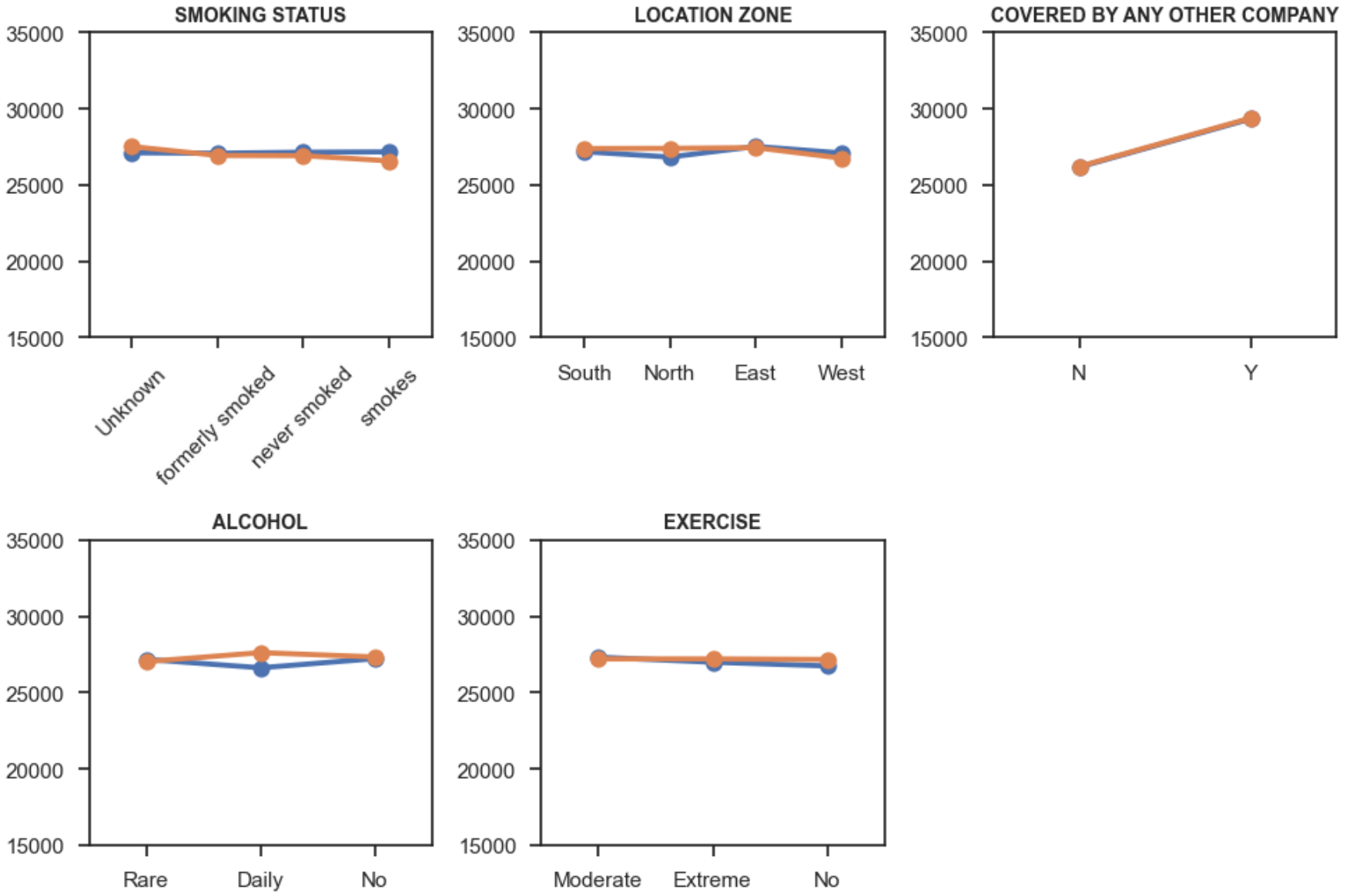
2.4 Target Variable to Numerical Variables




AVG GLUCOSE LEVEL
BMI
WEIGHT
WEIGHT CHANGE IN LAST ONE YEAR
FAT PERCENTAGE
insurance_cost
insurance_cost
70000
60000
50000
40000
30000
20000
10000
0
50 100 150 200 250
15 20 25 30 35 40 45 50
50 60 70 80 90
0 1 2 3 4 5 6
10 15 20 25 30 35 40
0.0 0.2 0.4 0.6 0.8 1.0

2.5 Target Variable to Categorical Variables




SMOKING STATUS
35000
30000
25000
20000
15000
Unknown
formerly smoked
never smoked
smokes
LOCATION ZONE
35000
30000
25000
20000
15000
South
North
East
West
COVERED BY ANY OTHER COMPANY
35000
30000
25000
20000
15000
N
Y
ALCOHOL
35000
30000
25000
20000
15000
Rare
Daily
No
EXERCISE
35000
30000
25000
20000
15000
Moderate
Extreme
No

2.6 Clusters

K-means Clustering recommended 7 clusters:

Cluster 0 — Mostly Students, Low Cholesterol, Moderate Exercisers

Cluster 1 — Medium Cholesterol, Business Owners

Cluster 2 — All non-exercisers

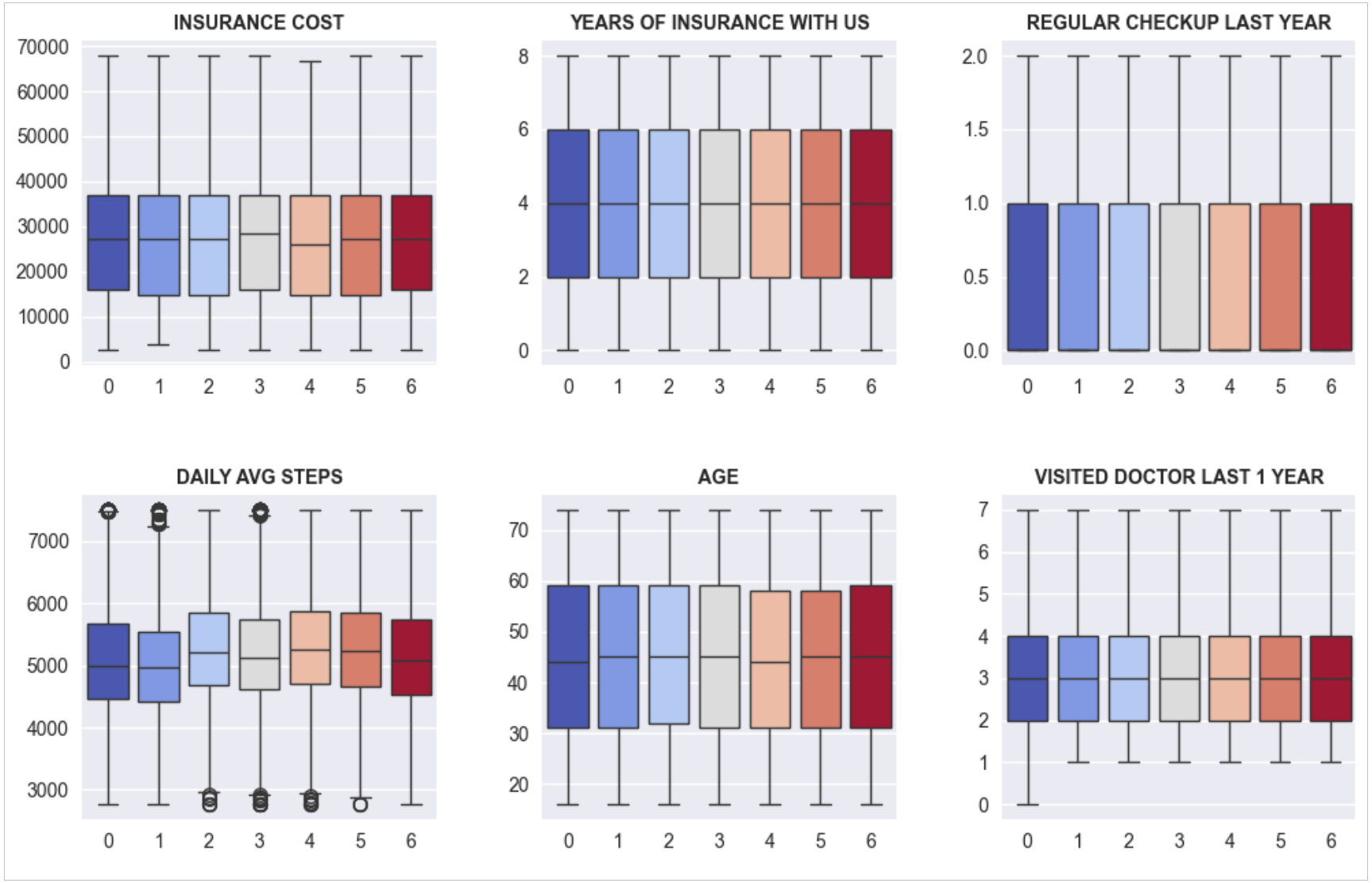
Cluster 3 — High Cholesterol, High Fat Percentage, Business Owners

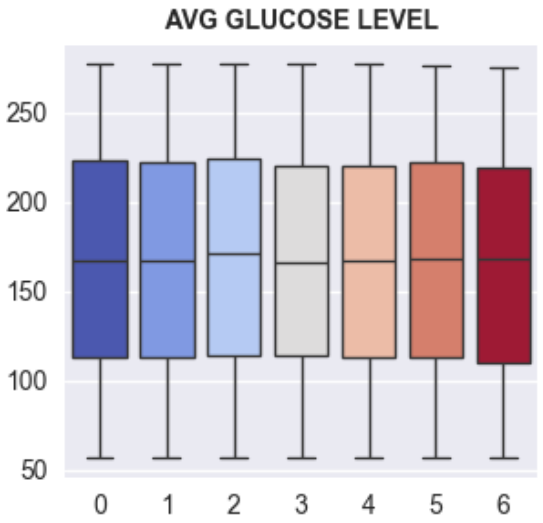
Cluster 4 — High Cholesterol, Salaried

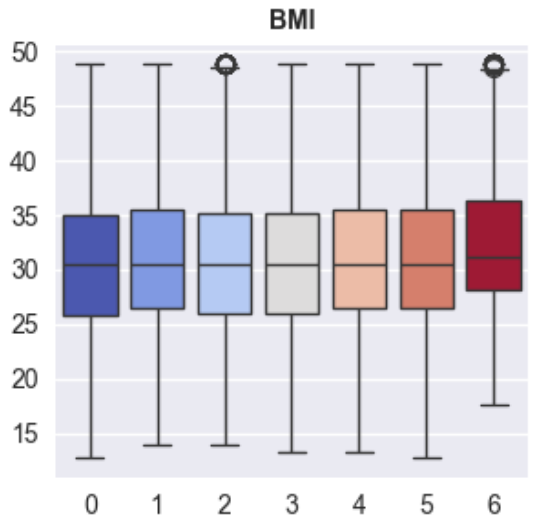
Cluster 5 — Low Cholesterol, Low Fat Percentage, Salaried

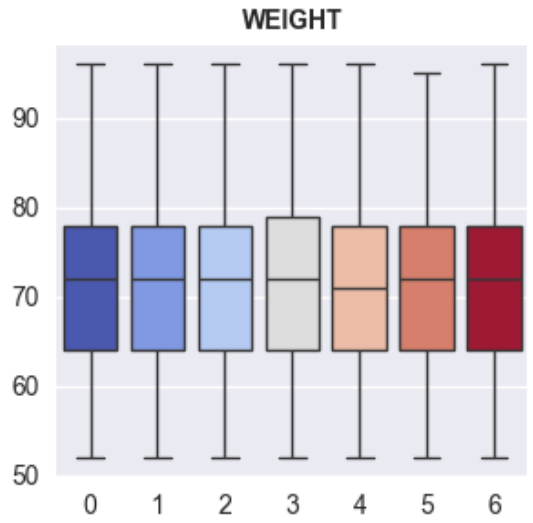
Cluster 6 — All Smokers

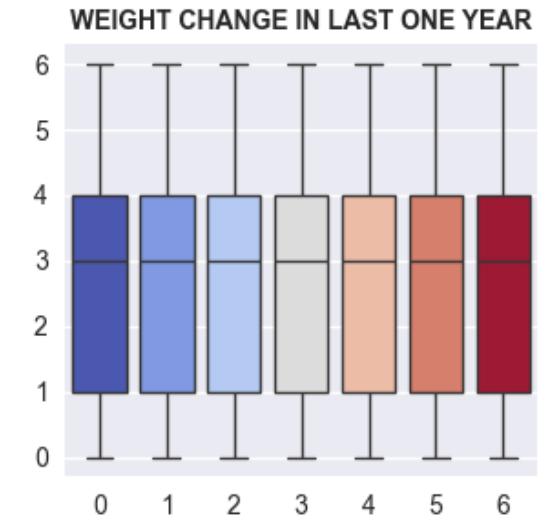
2.7 Clustered Data — Numerical Variables

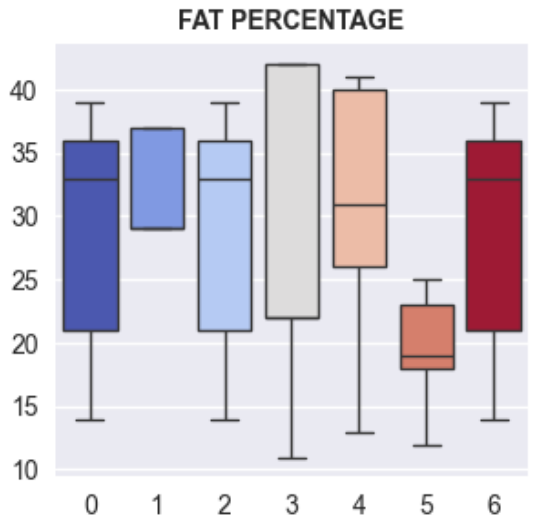






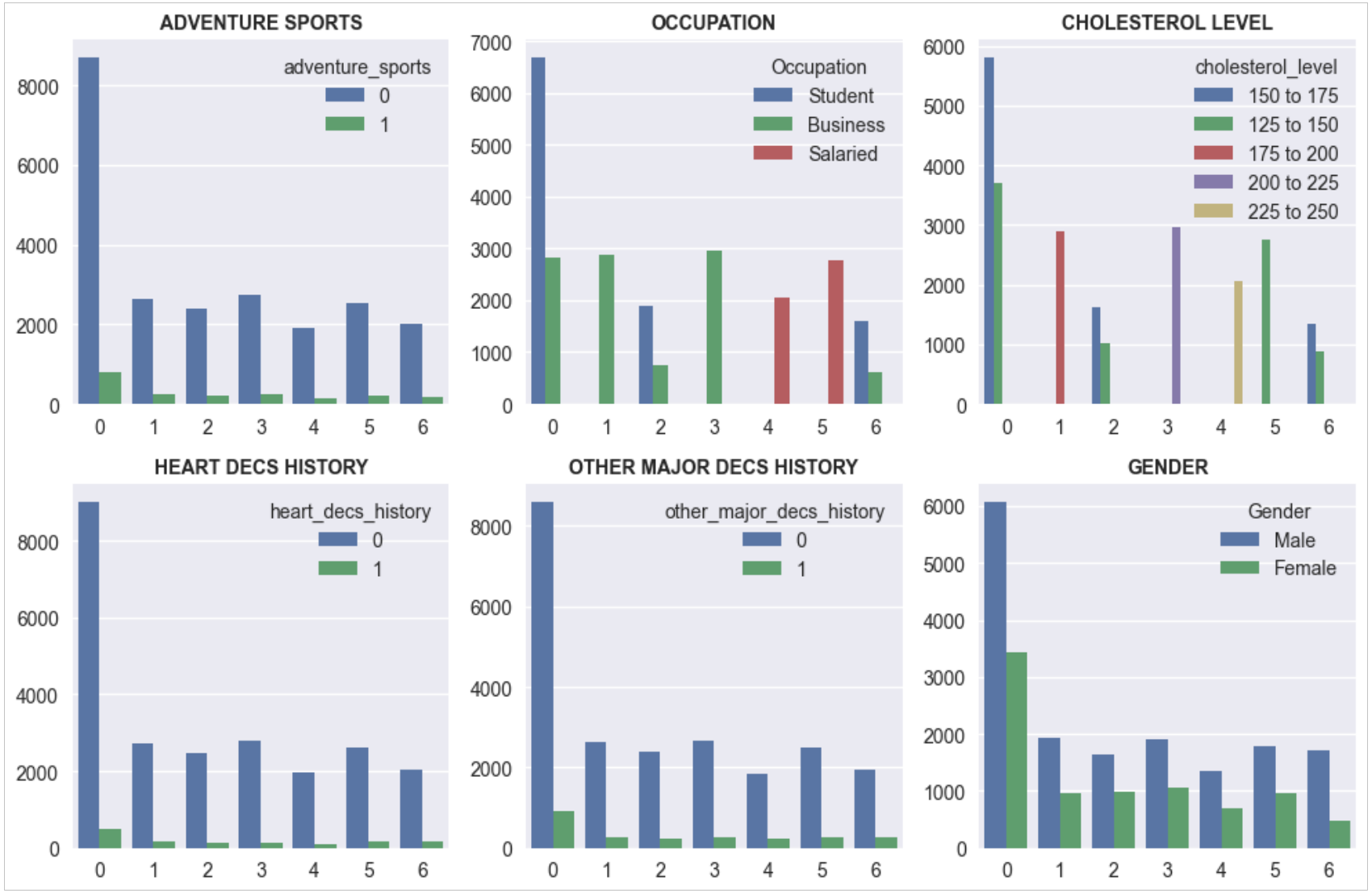




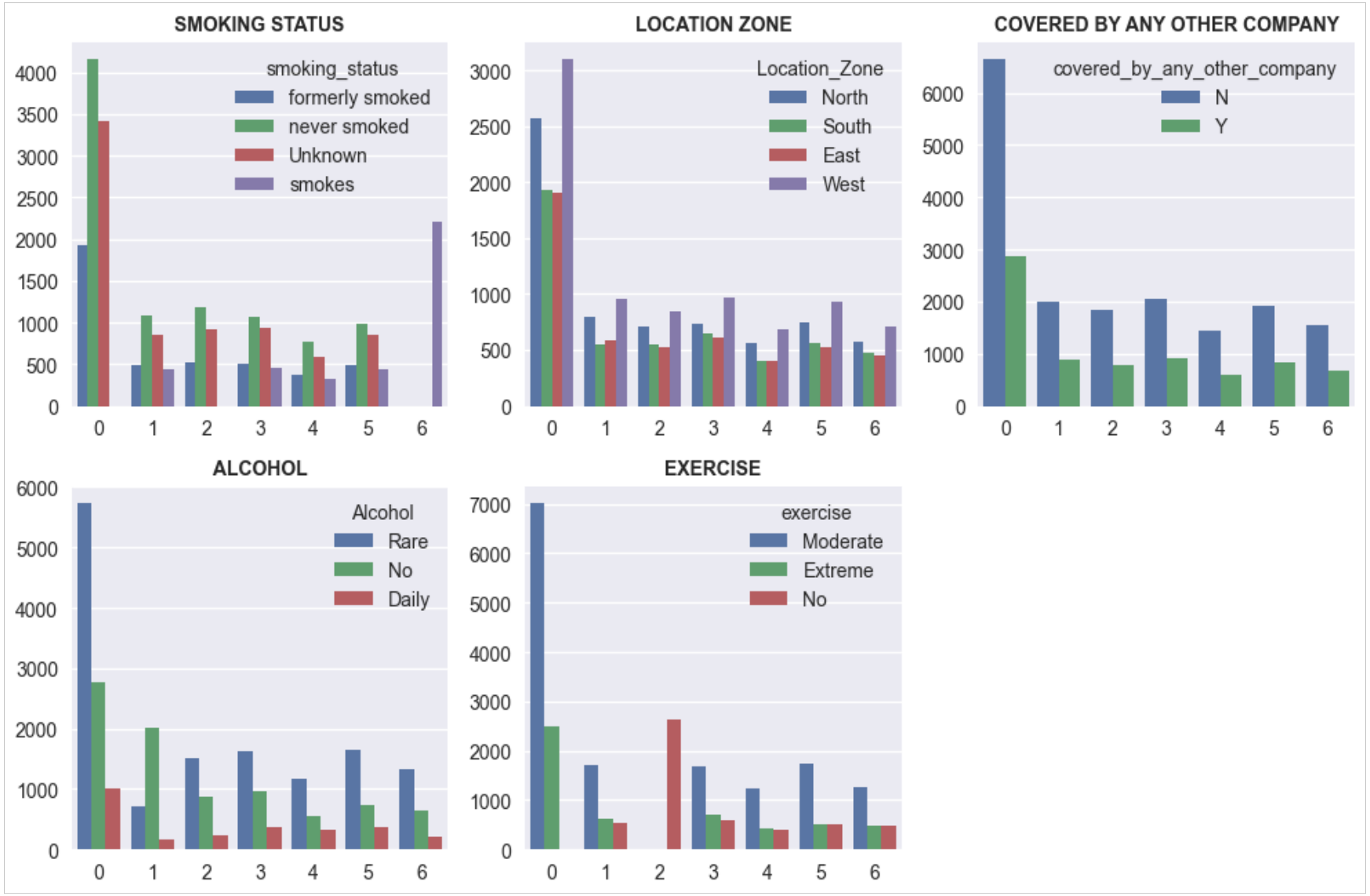


- Cluster 3: Highest Fat Percentage
- Cluster 5: Lowest Fat Percentage

2.8 Clustered Data — Categorical Variables



- Cluster 0: Mostly Students, Low Cholesterol
- Cluster 3 : Only Businesspeople, High Cholesterol
- Cluster 1: Medium Cholesterol Businesspeople
- Cluster 4: Salaried High Cholesterol
- Cluster 5: All Salaried Low Cholesterol



- Cluster 0: Mostly Moderate Exercisers
- Cluster 2: Non-Exercisers only
- Cluster 6: All Smokers

3.0 Modelling Approach

3.1 Models

Following Models were built and tested:

Parametric Models

1. Linear Regression
2. Polynomial Regression
3. Ridge
4. Lasso

Non-Parametric Models

1. Random Forest
2. XG Boost
3. AdaBoost
4. Support Vector Regression

3.2 Comparison

Model	Train				Test				Bias
	MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
Linear Regression	0.15	3374	0.945	0.945	0.15	3339	0.945	0.945	Slight bias
Polynomial Reg. deg. 2	0.14	3166	0.951	0.951	0.14	3237	0.949	0.948	Unbiased
Polynomial Reg. deg. 3	0.11	2683	0.965	0.965	0.16	3795	0.929	0.929	Slight bias
Ridge	0.15	3363	0.945	0.945	0.16	3369	0.944	0.944	Slight bias
Lasso	0.15	3363	0.945	0.945	0.16	3369	0.944	0.944	Slight bias
Random Forest	0.05	1173	0.993	0.993	0.12	3105	0.953	0.953	Unbiased
XG Boost	0.09	2250	0.975	0.975	0.13	3150	0.95	0.95	Unbiased
XG Boost Tuned	0.12	2997	0.956	0.956	0.12	3001	0.956	0.956	Slight Bias
AdaBoost	0.15	3172	0.951	0.951	0.15	3147	0.951	0.951	Biased
AdaBoost Tuned	0.15	3168	0.951	0.951	0.15	3181	0.950	0.950	Slight Bias
Support Vector Reg.	0.80	14333	0.00	0.00	0.80	14271	0.00	0.00	Biased

3.3 Model Selection

Random Forest was selected as compared to XGB Tuned because:

1. **Much Lower MAPE** — Only 5%
2. **Much better Train RMSE** — 1173, with almost equal Test RMSE
3. **Much better Train R^2** — 0.993, with almost equal Test R^2
4. **Unbiased** — Perfectly random residuals

3.4 Important Features

Feature importances from the Random Forest model:

S.No.	Feature	Importance
1	**Weight**	**95.17%**
2	Daily Avg. Steps	0.59%
3	Avg. Glucose Level	0.57%
4	BMI	0.56%
5	Age	0.51%

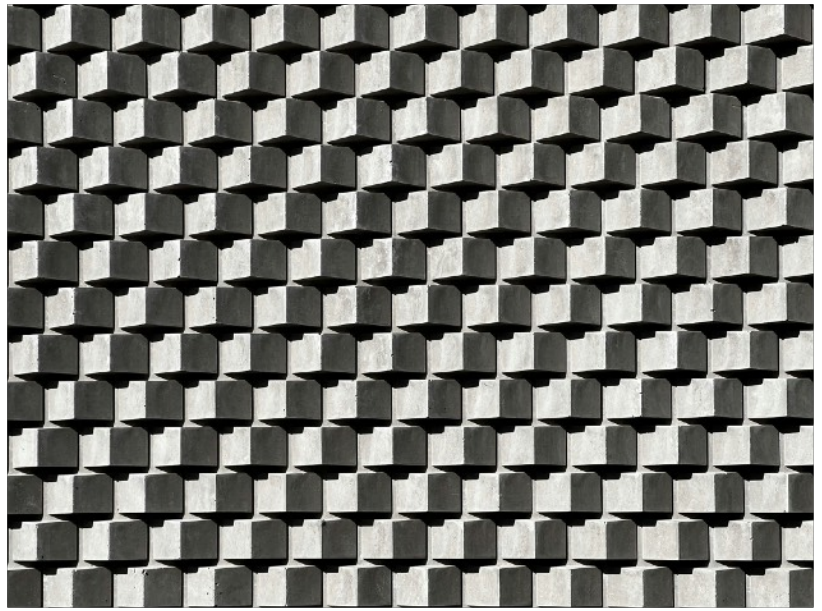
4.0 Key Insights

1. Weight Dominance

Weight is the primary factor determining insurance cost, with a 97% correlation.

2. Homogeneity

The dataset is highly homogenous, meaning there is little variation across most features except for weight.

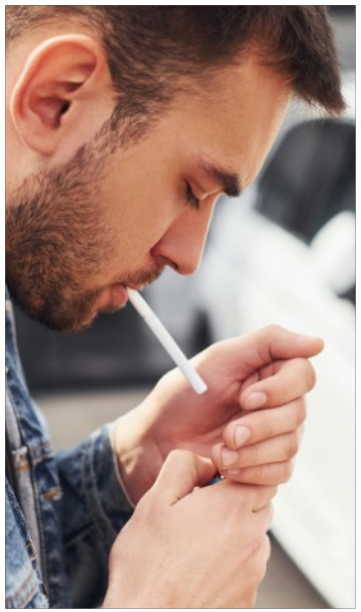


3. Clustering

Some clusters are able to be identified which can help separate specific customer groups.

4. Low Impact of Lifestyle Factors

Lifestyle factors do not show sufficient variation to have an impact on the insurance cost.



5. Cholesterol and Glucose Levels

While high in the given data, they don’t seem to significantly impact insurance cost predictions.

5.0 Recommendations

1. Cannot Follow the Model

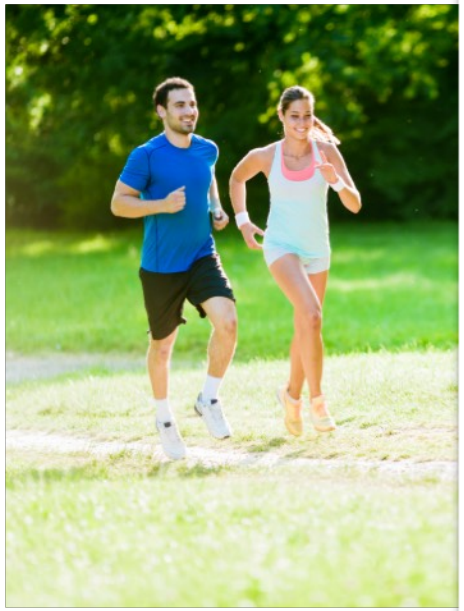
Since Weight is the overwhelming factor determining insurance cost, the model cannot be used because it would give inaccurate results.

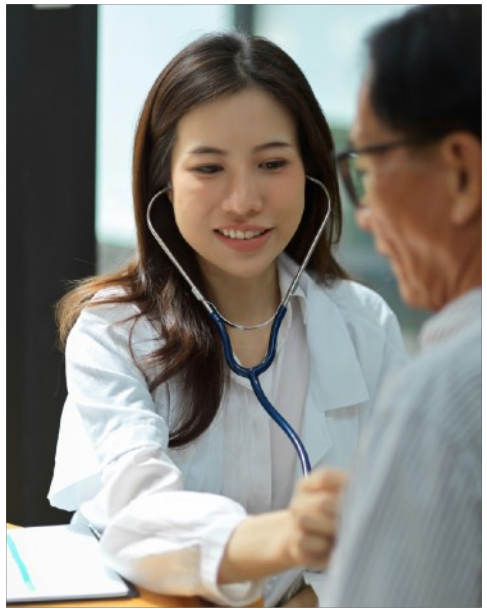
2. Keep Collecting Data

More detailed information on various factors needs to be collected to induce variation in predictive models.

3. Incentivise Healthy Behaviour

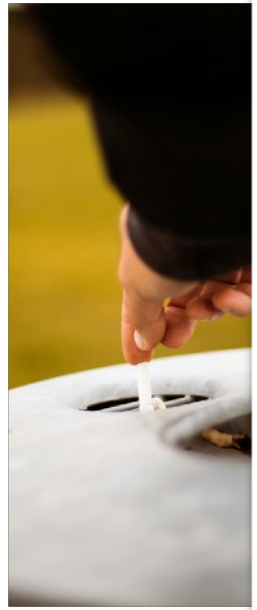
Introduce incentives for regular health checkups and moderate exercise to promote healthier lifestyles.





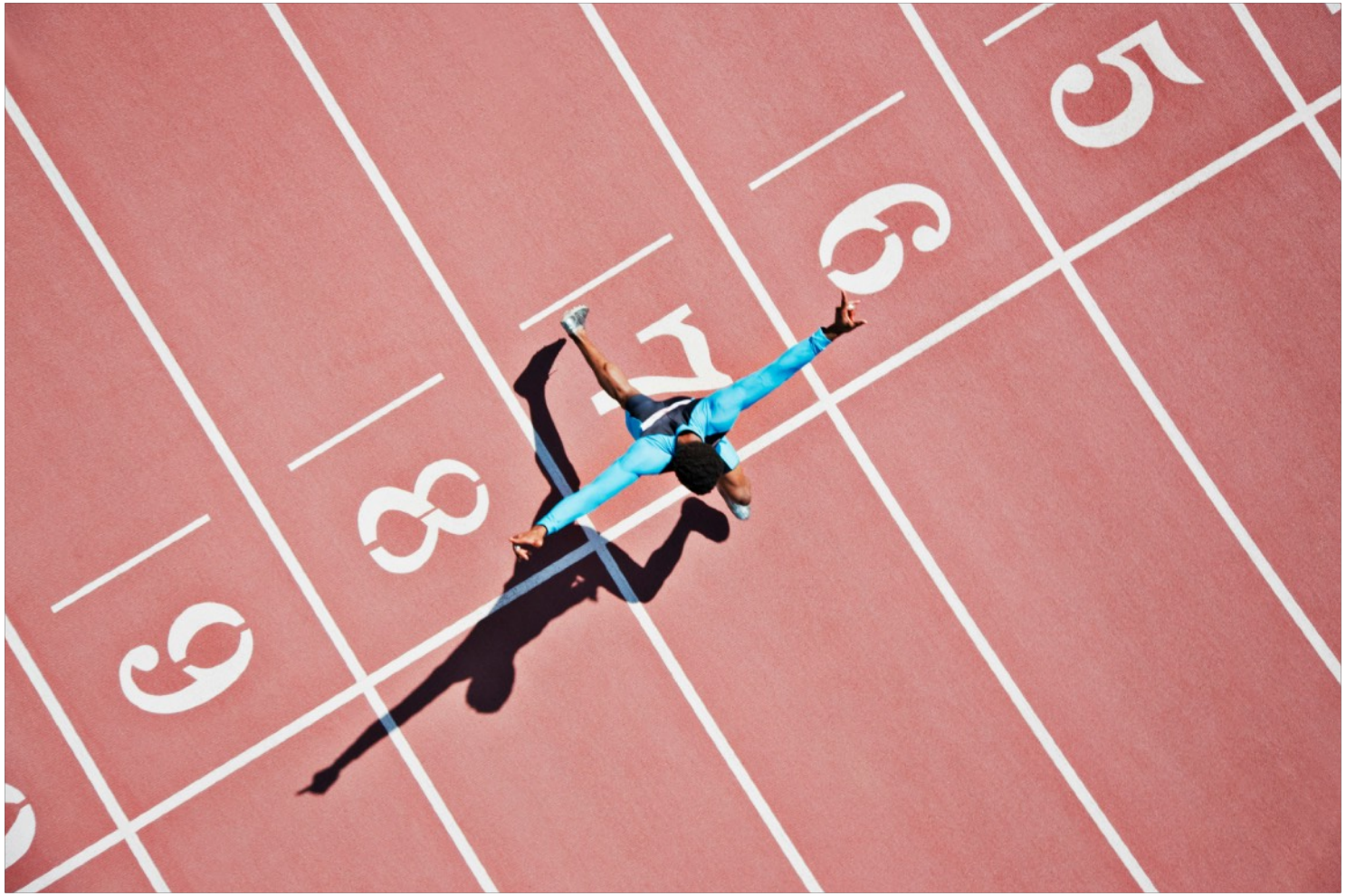
4. Targeted Wellness Programs

The clusters identified can guide the creation of wellness programs aimed at specific groups, like Smokers and High Cholesterol.



5. Refine Marketing Strategy

The clusters provide a basis for segmenting customers and offering tailored products.



Thank you!