Capstone Project Report

Health Insurance Cost Prediction

PGPDSBA.O.OCT23.A

Anirudh Sardiwal

Oct. 27, 2024

Table of Contents

1.0 Introduction

1.1 Problem Statement

Healthcare is a crucial industry that directly impacts the well-being of individuals. Managing medical expenses can become challenging, especially when treatments are costly and an individual lacks sufficient insurance coverage. Medical insurance companies aim to reduce financial risk by optimizing insurance costs. Healthier lifestyles—such as proper diet and regular exercise—significantly reduce the likelihood of illness, influencing insurance premiums.

The goal of this project is to develop a model that predicts the optimal insurance cost for an individual, based on their health and lifestyle habits.

1.2 Need of the Project

The need for a project that predicts insurance costs based on health and lifestyle parameters is more relevant than ever due to rising healthcare expenses and the increasing focus on preventive health. Insurance companies are seeking ways to personalize premiums by accurately assessing an individual's risk profile, which can lead to fairer pricing for both insurers and policyholders. With advancements in data analytics, leveraging health habits like diet, exercise, and smoking status allows for a more accurate prediction of future medical expenses, promoting healthier lifestyles while reducing costs. This approach aligns with the global shift toward preventive healthcare, offering individuals incentives to adopt healthier habits and reducing the overall financial burden on healthcare systems.

1.3 Understanding Business Opportunity

The business opportunity in this project lies in creating a more personalized and data-driven insurance pricing model, which can greatly benefit insurance companies, healthcare providers, and customers. By accurately predicting insurance costs based on an individual's health and lifestyle habits, insurers can reduce risk and improve profitability by setting fairer premiums. This can also attract health-conscious customers who seek lower premiums as a reward for maintaining healthy lifestyles.

2.0 Data Report

The data consists of 25000 Rows and 23 Columns. There were 0 Duplicate rows.

2.1 Data Dictionary

S.No.	Feature Name	Description
1	applicant_id	Applicant unique ID
2	years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
3	regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
4	adventure_sports	Customer is involved with adventure sports like climbing, diving etc.
5	Occupation	Occupation of the customer
6	visited_doctor_last_1_year	Number of times customer has visited doctor in last one year
7	cholesterol_level	Cholesterol level of the customers while applying for insurance
8	daily_avg_steps	Average daily steps walked by customers
9	age	Age of the customer
10	heart_decs_history	Any past heart diseases

11	other_major_decs_history	Any past major diseases apart from heart like any operation
12	Gender	Gender of the customer
13	avg_glucose_level	Average glucose level of the customer while applying the insurance
14	bmi	BMI of the customer while applying the insurance
15	smoking_status	Smoking status of the customer
16	Year_last_admitted	When customer have been admitted in the hospital last time
17	Location	Location of the hospital
18	weight	Weight of the customer
19	covered_by_any_other_company	Customer is covered from any other insurance company
20	Alcohol	Alcohol consumption status of the customer
21	exercise	Regular exercise status of the customer
22	weight_change_in_last_one_year	How much variation has been seen in the weight of the customer in last year
23	fat_percentage	Fat percentage of the customer while applying the insurance
24	insurance_cost	Total Insurance cost

2.2 Other Details

	years_of_insurance_with_us	regular_checkup_last_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps
0	3	1	1	Salaried	2.0	125 to 150	4866.0
1	0	0	0	Student	4.0	150 to 175	6411.0
2	1	0	0	Business	4.0	200 to 225	4509.0
3	7	2	0	Business	2.0	175 to 200	6214.0
4	3	1	0	Student	2.0	150 to 175	4938.0

2.1 - Partial Data Head

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	he
count	25000.0	25000.0	25000.0	25000.0	25000.0	25000.0	25000.0	
mean	17500.0	4.0	1.0	0.0	3.0	5216.0	45.0	
std	7217.0	3.0	1.0	0.0	1.0	1053.0	16.0	
min	5000.0	0.0	0.0	0.0	0.0	2034.0	16.0	
25%	11250.0	2.0	0.0	0.0	2.0	4543.0	31.0	
50%	17500.0	4.0	0.0	0.0	3.0	5089.0	45.0	
75%	23749.0	6.0	1.0	0.0	4.0	5730.0	59.0	
max	29999.0	8.0	5.0	1.0	12.0	11255.0	74.0	

2.2 - Partial Descriptive Statistics

```
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)
```

2.3 - Feature data types

Following features were categorized as Numerical:

‘insurance_cost', ‘years_of_insurance_with_us', ‘regular_checkup_last_year', ‘daily_avg_steps', ‘age', ‘visited_doctor_last_1_year', ‘avg_glucose_level', ‘bmi', ‘weight' ,’weight_change_in_last_one_year', ‘fat_percentage'

Following features were categorized as Categorical:

‘adventure_sports', ‘Occupation', ‘cholesterol_level', 'heart_decs_history', ‘other_major_decs_history', ‘Gender', ‘smoking_status', ‘Location_Zone', ‘covered_by_any_other_company', ‘Alcohol', 'exercise'

3.0 Exploratory Data Analysis

3.1 Removal of Unnecessary Variables & Correcting Errors

The following variables were removed:

1. **'applicant_id'** — because it is purely for administrative purposes.
2. **'Year_last_admitted'** — because it had > 30% missing values.
3. **'Location'** — Cities were clubbed into Geographical zones and a feature **'Location_Zone'** was added.

Following errors were corrected:

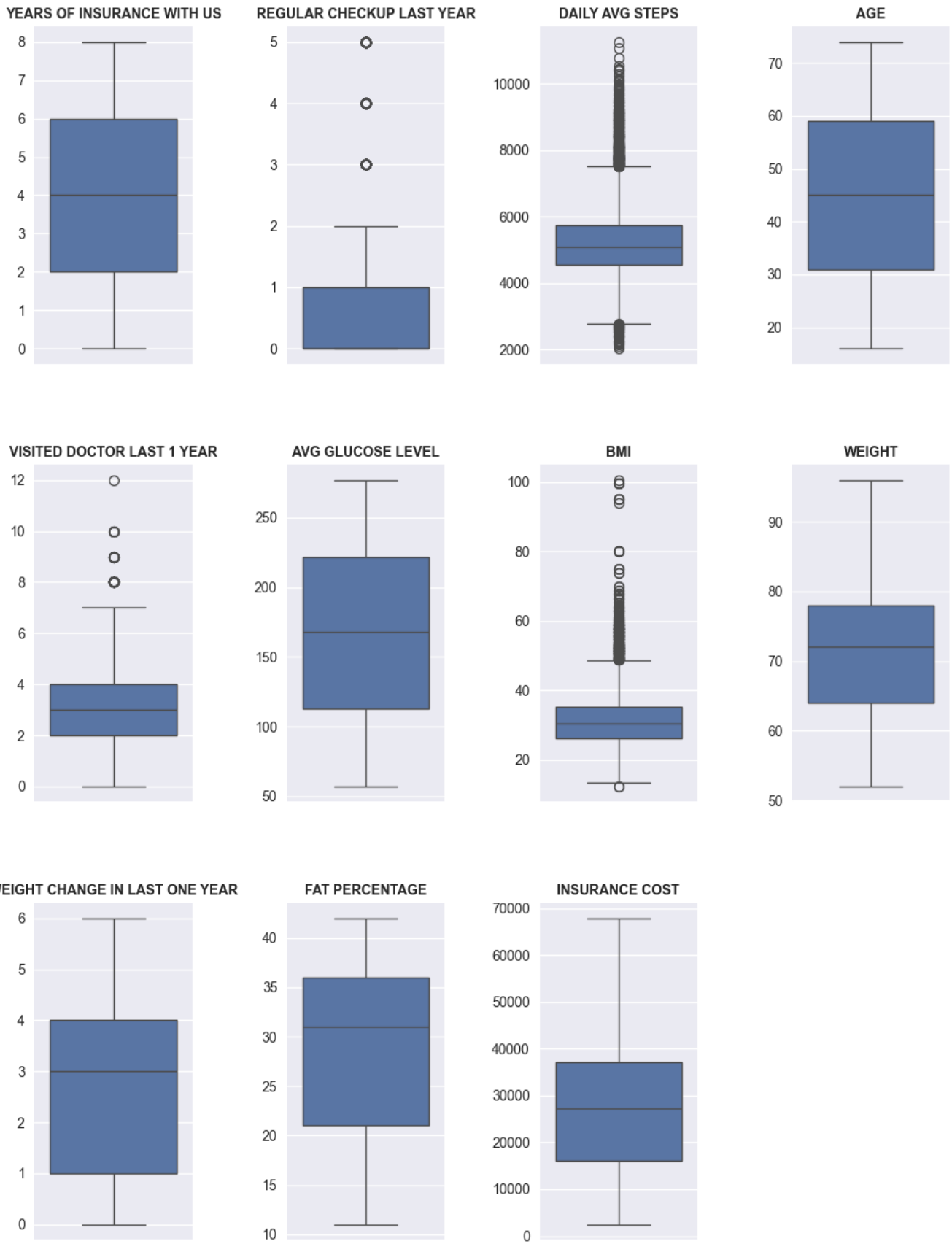
1. 'regular_checkup_lasy_year' was renamed 'regular_checkup_last_year'
2. 'Salried' was renamed 'Salaried'

3.2 Missing Value Treatment

The feature 'bmi' had 990 missing values which is < 5% of values, therefore it was imputed with Median values.

3.3 Univariate Analysis

3.3.1 Numerical Variables — Boxplots


YEARS OF INSURANCE WITH US
REGULAR CHECKUP LAST YEAR
DAILY AVG STEPS
AGE
VISITED DOCTOR LAST 1 YEAR
AVG GLUCOSE LEVEL
BMI
WEIGHT
WEIGHT CHANGE IN LAST ONE YEAR
FAT PERCENTAGE
INSURANCE COST

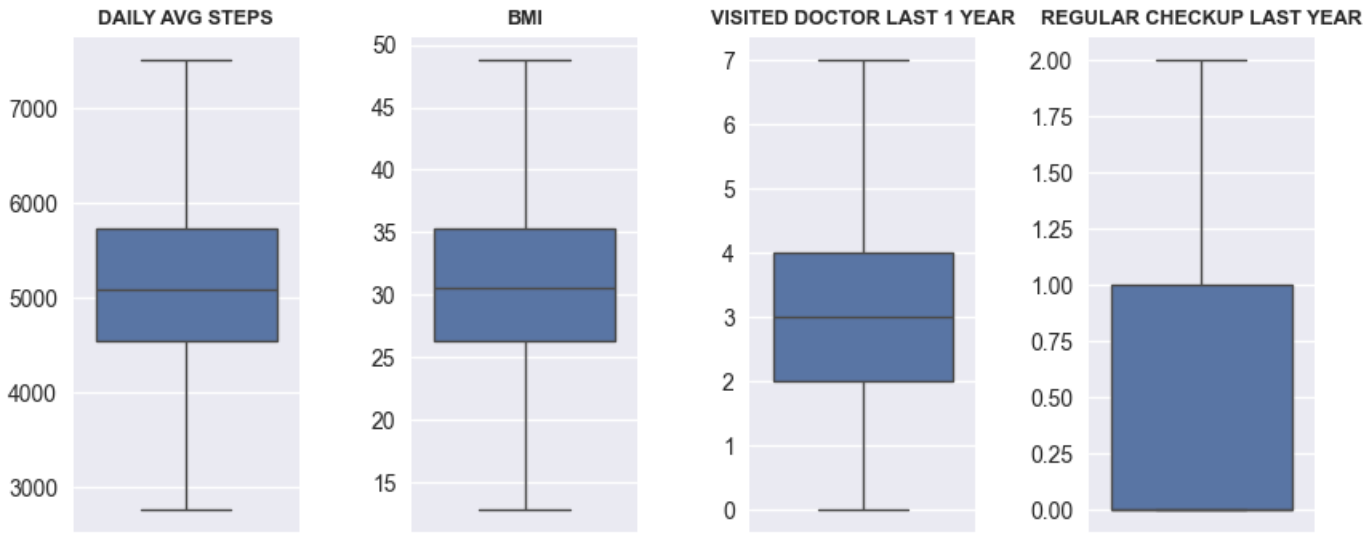
3.1 - Boxplots

3.3.2 Outlier Removal

The features **bmi, daily_avg_steps, visited_doctor_last_1_year, and regular_checkup_last_year** are showing outliers.

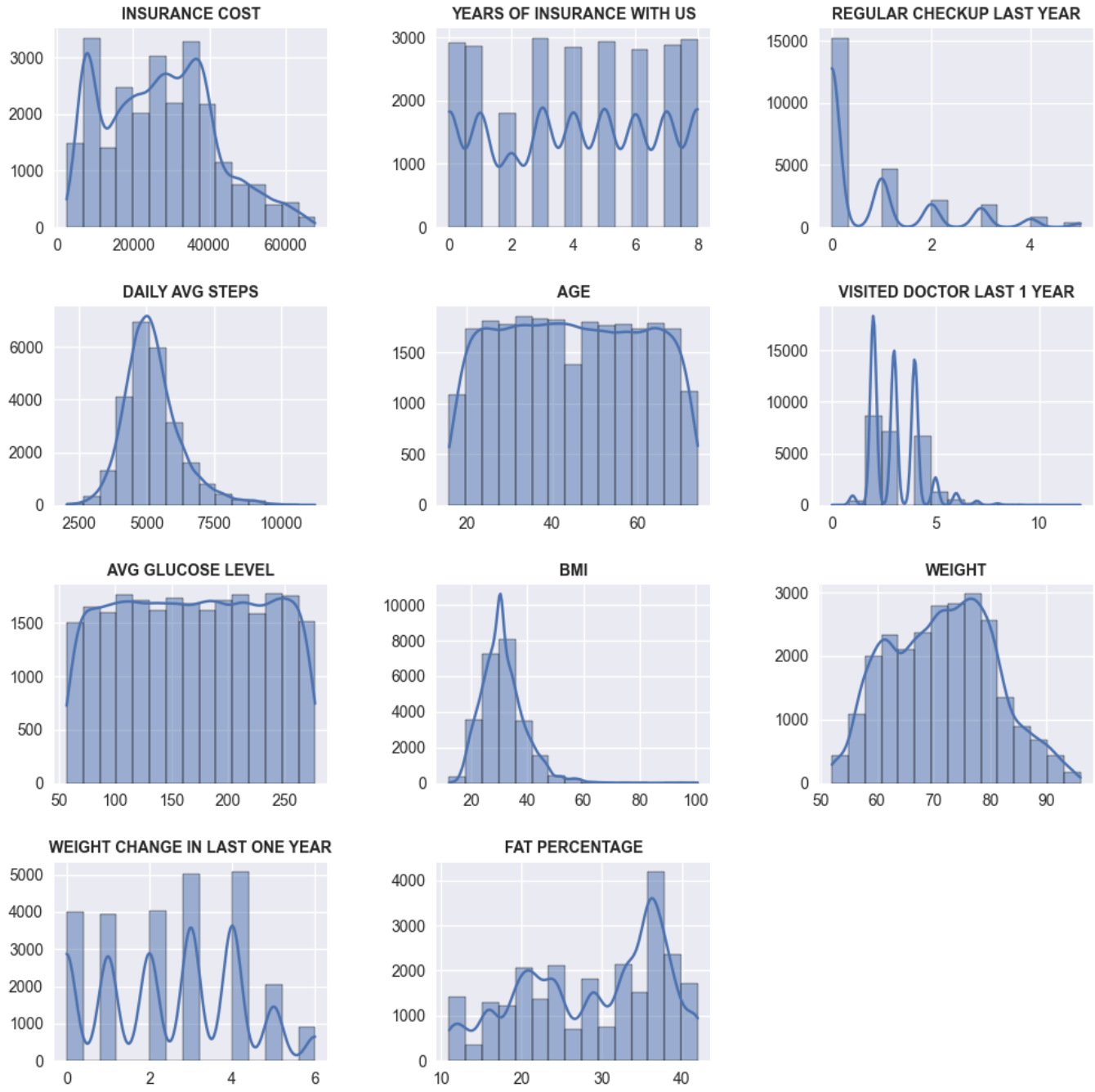
These features were capped at a standard Q1/Q3 ± 1.5 x IQR.

Resulting boxplots:



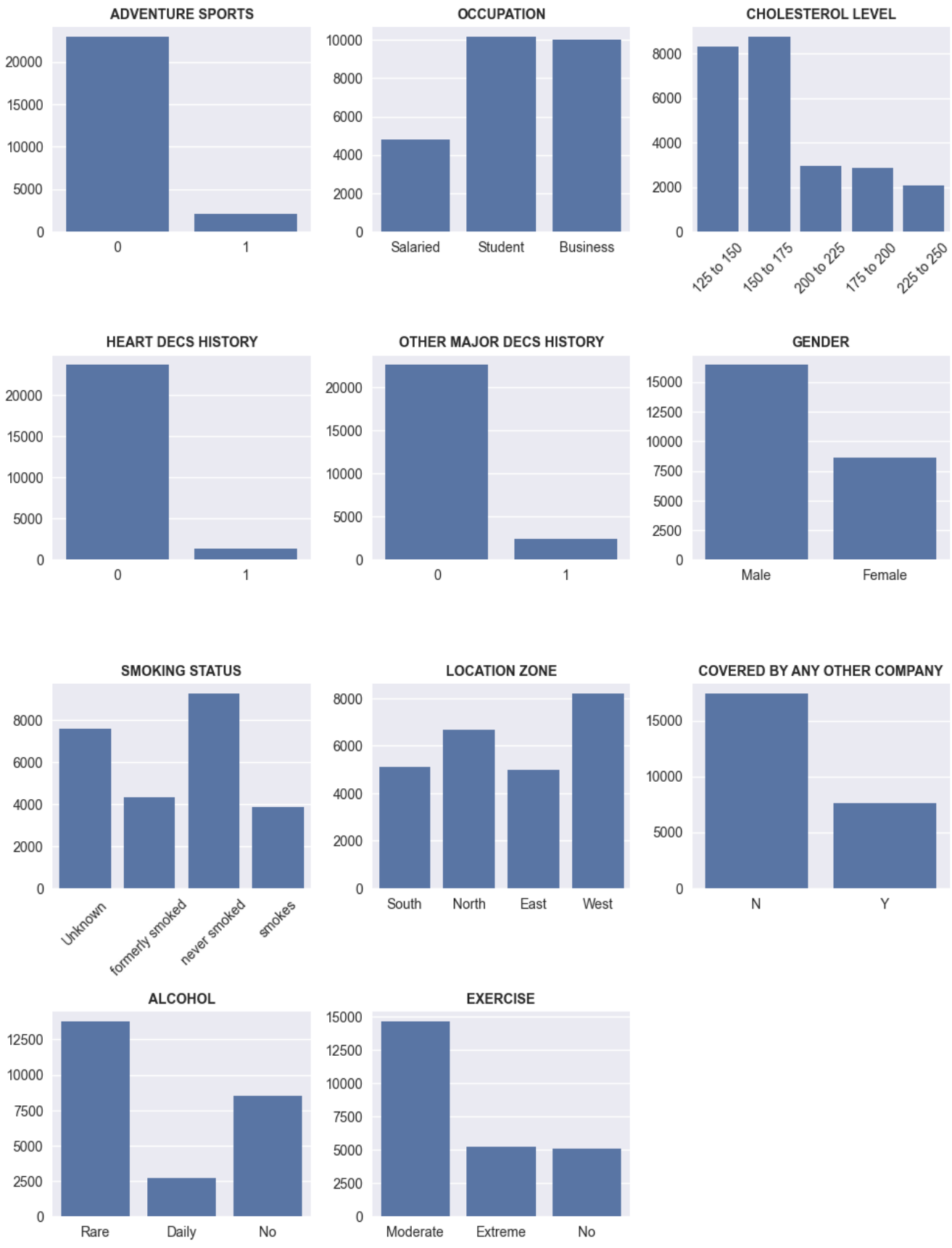
3.2 — Boxplots of features with removed outliers

3.3.3 Numerical Variables — Histograms



3.3 — Histograms after outlier removal

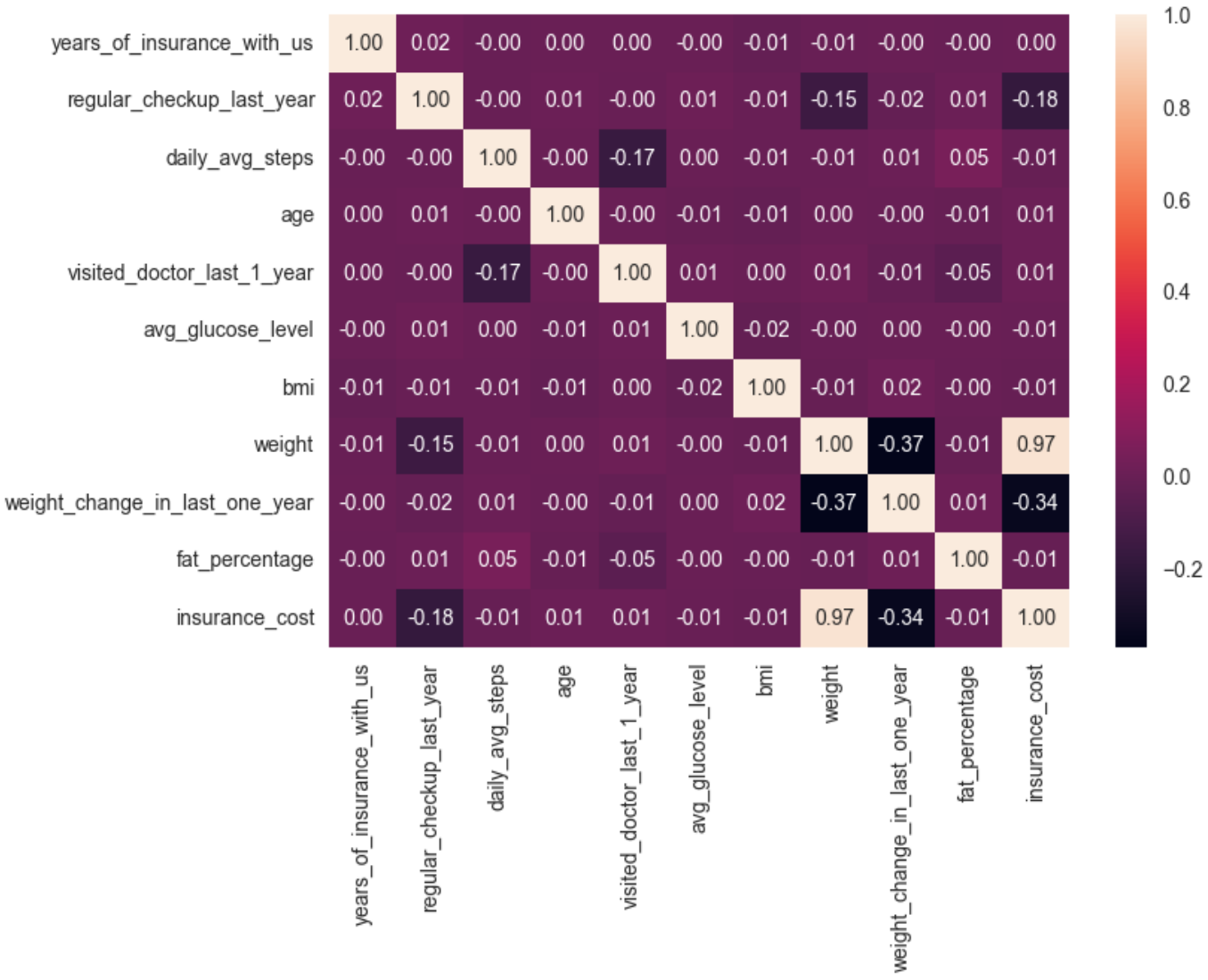
3.3.4 Categorical Variables — Countplots



3.4 — Countplots

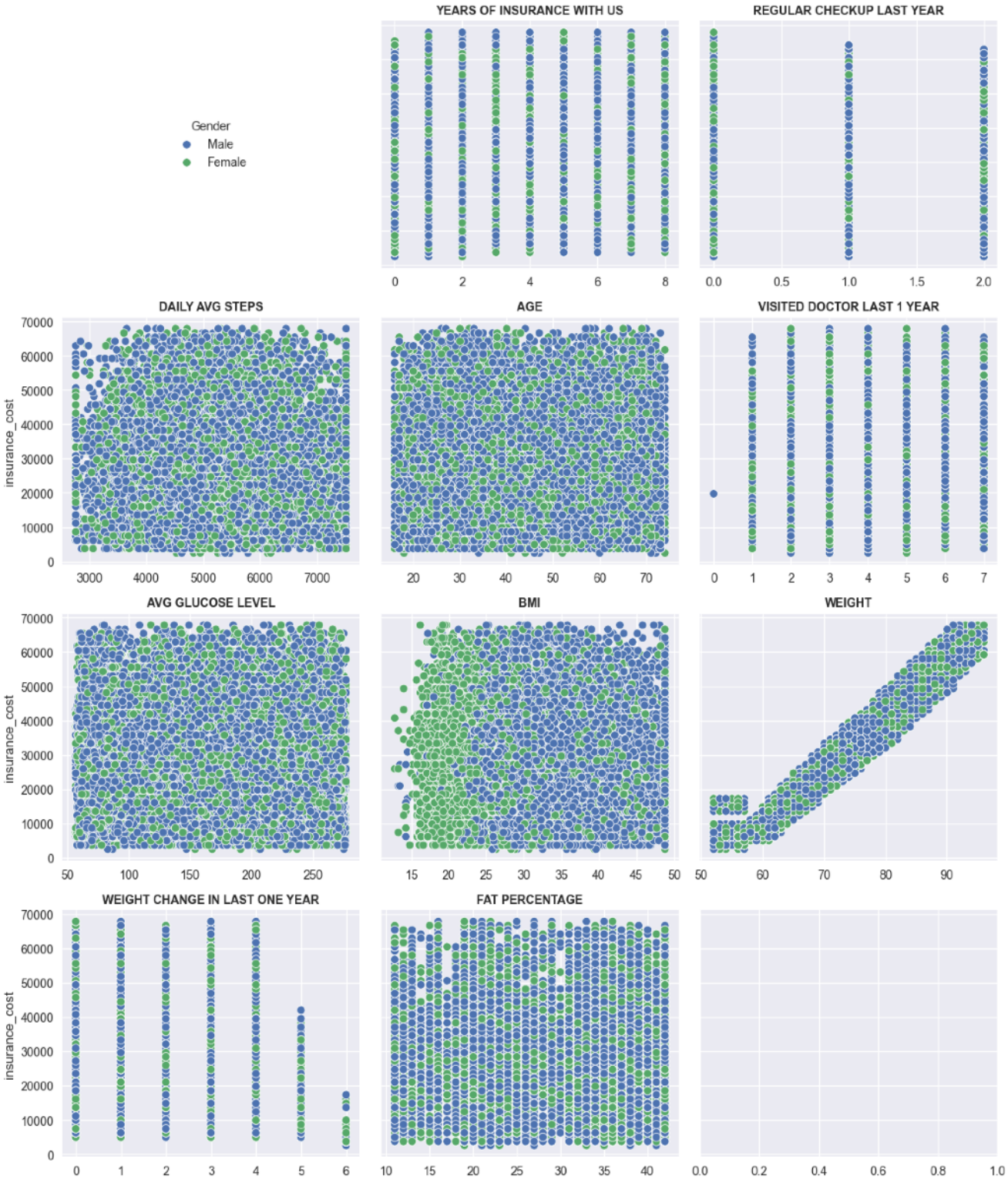
3.4 Bivariate Analysis

3.4.1 Correlation Heatmap



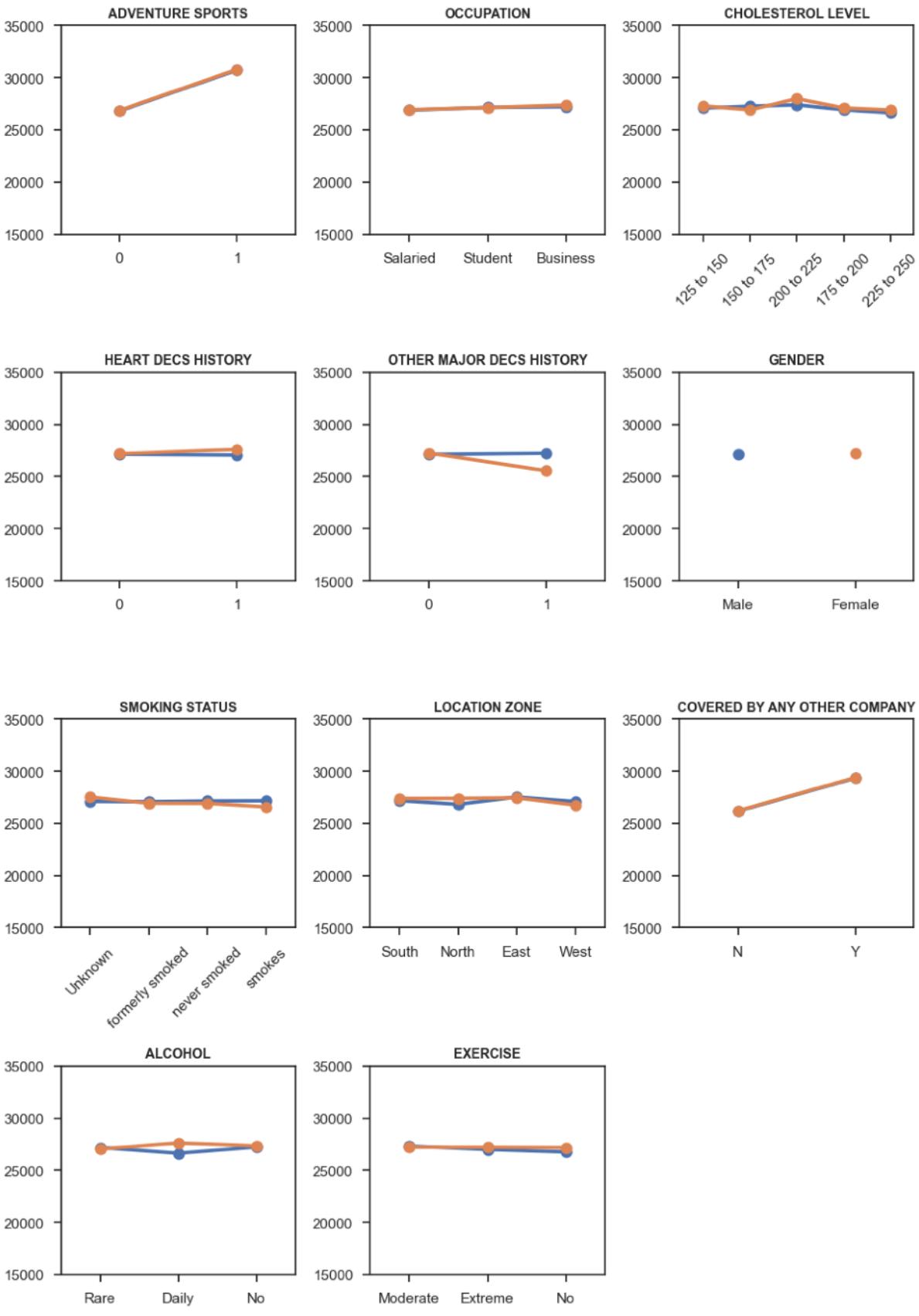
3.5 — Correlation Heatmap

3.4.2 Relation of Target Variable to Numerical Variables



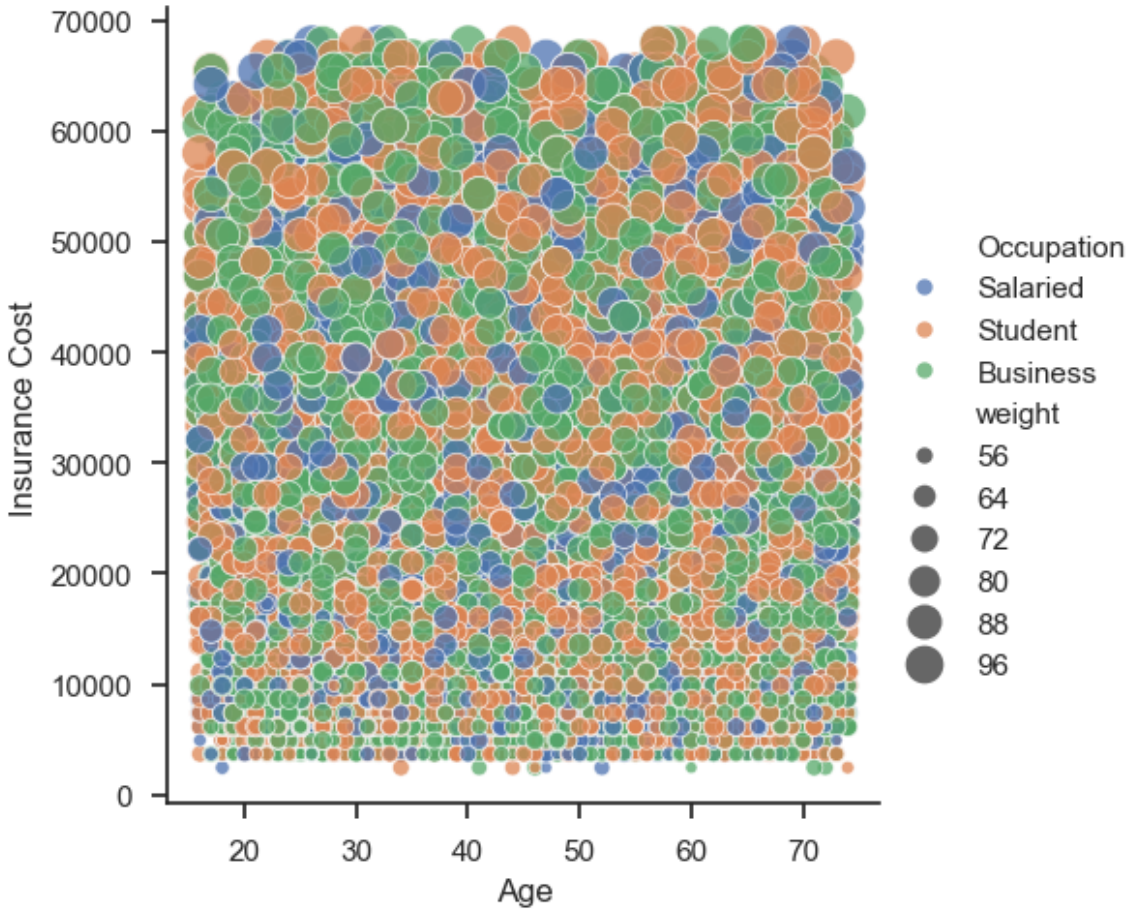
3.6 — Relation of Target Variable to Numerical Variables differentiated by Gender

3.4.3 Relation of Target Variable to Categorical Variables



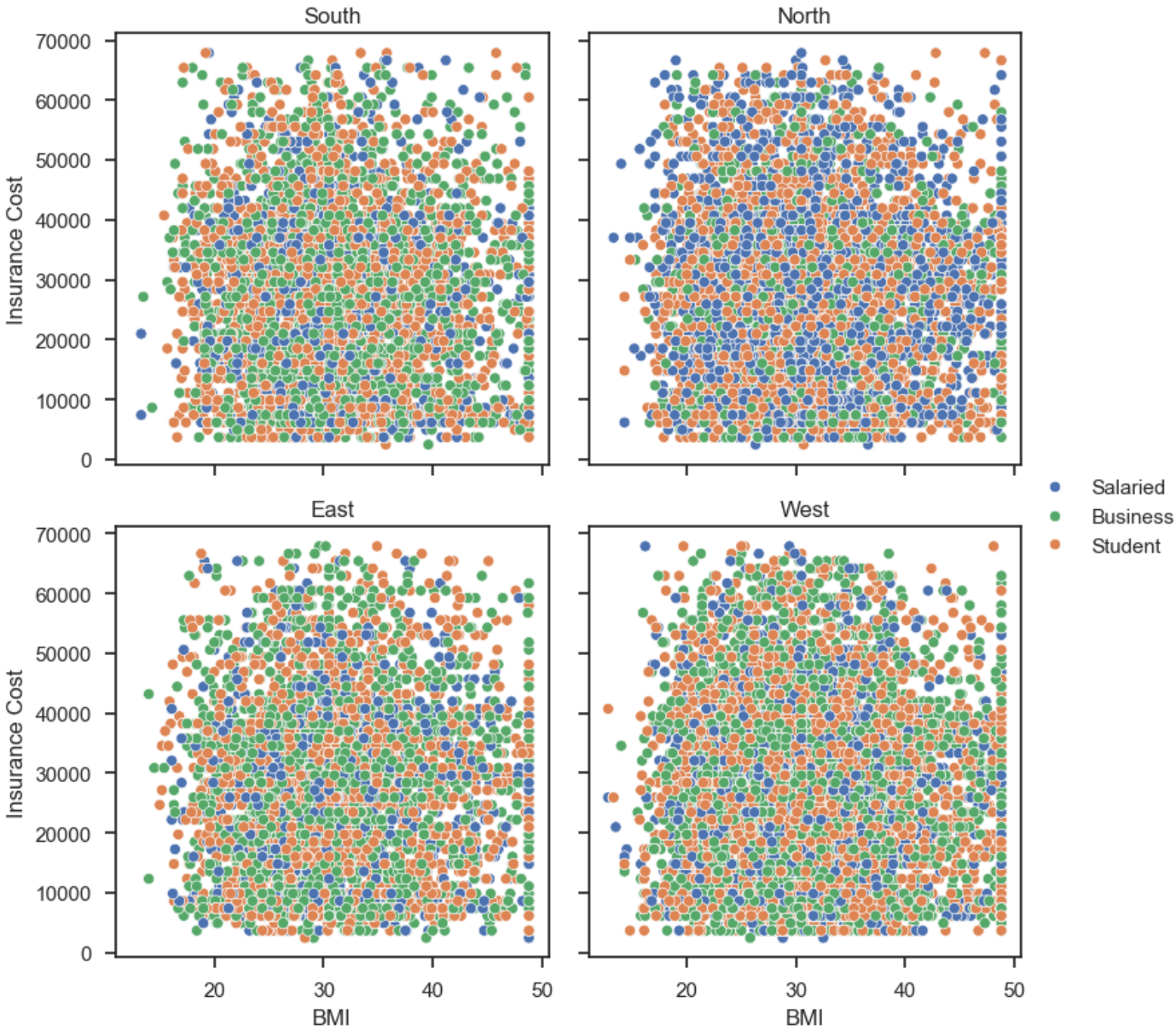
3.7 — Mean Insurance cost by Categorical Variables differentiated by Gender

3.4.4 Insurance Cost relation to Age detailed



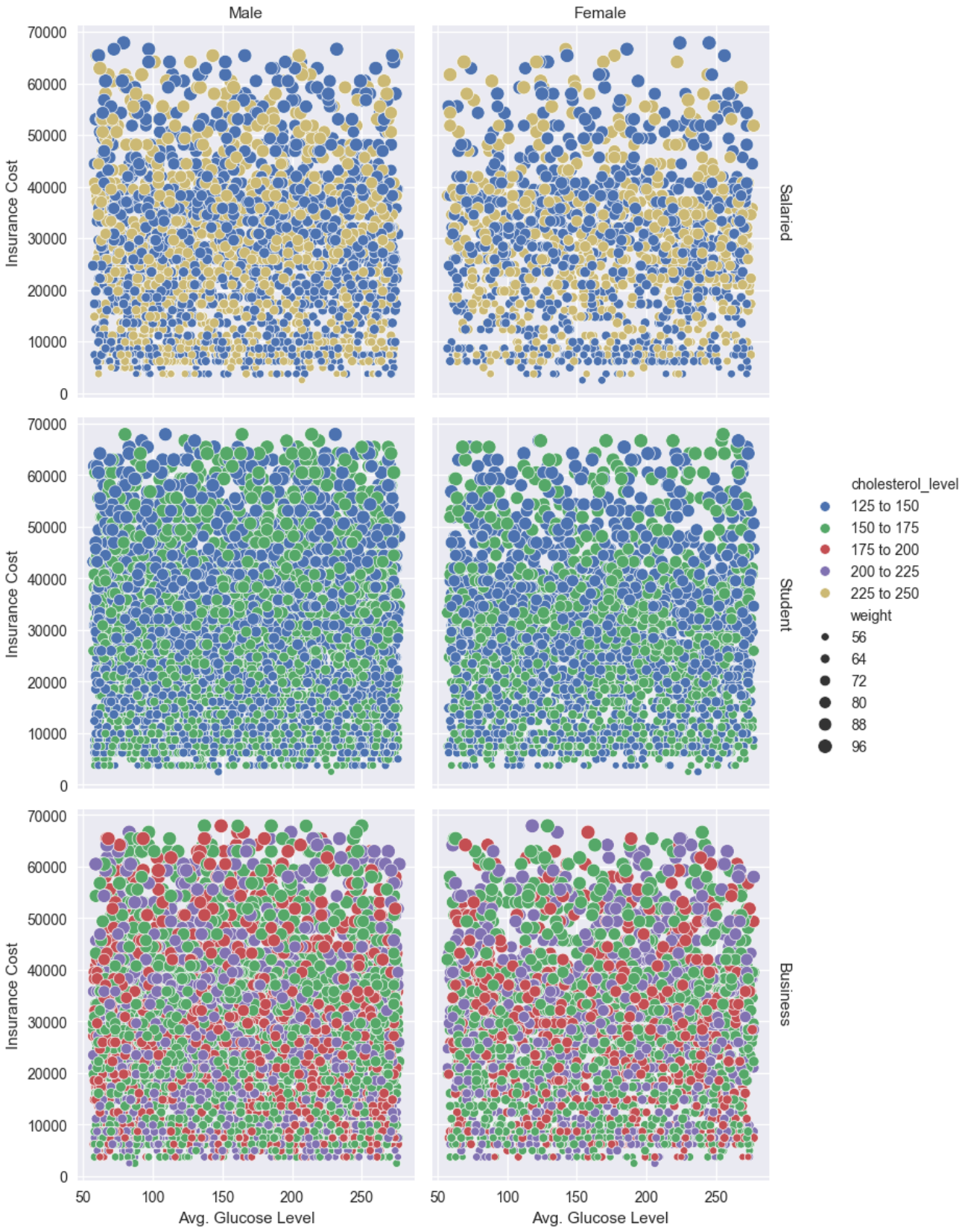
3.8 — Relation of Insurance Cost to Age differentiated by Occupation detailed by Weight

3.4.5 Relation of Insurance Cost to BMI



3.8 — Relation of Insurance Cost to BMI differentiated by Occupation for each Zone

3.4.6 — Relation of Insurance Cost to Glucose Level



3.9 — Relation of Insurance Cost to Glucose Level with a grid of Occupation and Gender detailed by Weight

3.5 Clustering

3.5.1 Encoding & Scaling

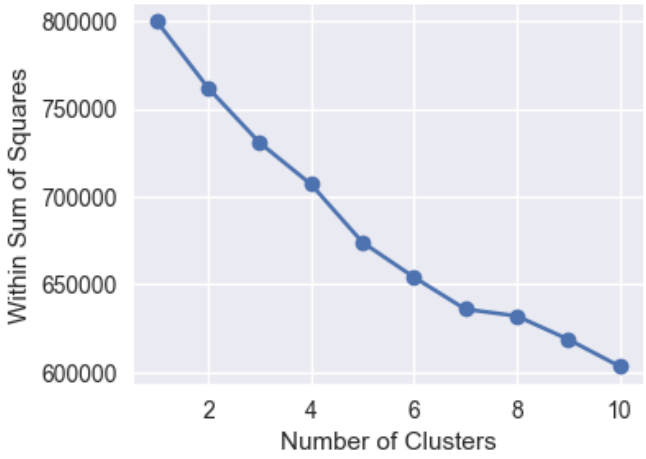
..	smoking_status_never smoked	smoking_status_smokes	Location_Zone_North	Location_Zone_South	Location_Zone_West	covered_by_any_other_comp
...	0	0	0	1	0	
...	0	0	1	0	0	
...	0	0	1	0	0	
...	0	0	0	1	0	
...	1	0	0	1	0	

3.10 — Partial encoded data head

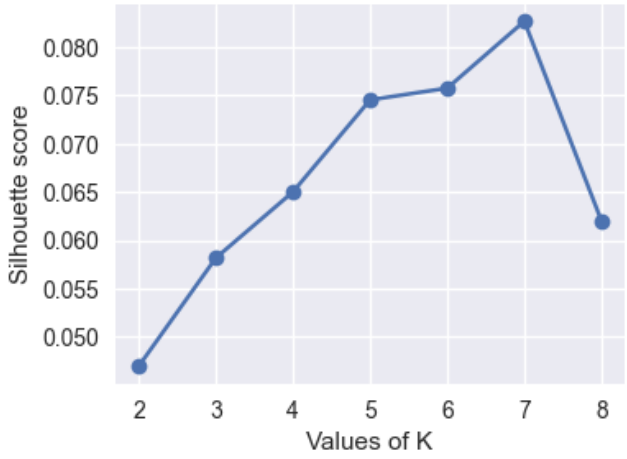
	years_of_insurance_with_us	regular_checkup_last_year	visited_doctor_last_1_year	daily_avg_steps	age	avg_glucose_level	bmi	
0	-0.417807	0.499096	-0.980772	-0.333160	-1.050360	-1.124370	0.002231	-0.4
1	-1.568750	-0.739478	0.803748	1.260326	0.315492	0.708929	0.422682	-1.4
2	-1.185102	-0.739478	0.803748	-0.701364	1.433007	-0.024391	1.291613	0.
3	1.116783	1.737669	-0.980772	1.057144	0.377576	-0.933069	-1.161015	-0.(
4	-0.417807	0.499096	-0.980772	-0.258901	-0.057013	-0.789594	-0.656474	0.2

3.11 — Partial data head — encoded and scaled

3.5.2 Clustering



3.12 — Within sum of squares by cluster



3.13 — Silhouette score by cluster

Based on silhouette score analysis, the optimum number of clusters is 7.

eight	covered_by_any_other_company	Alcohol	exercise	weight_change_in_last_one_year	fat_percentage	insurance_cost	Location_Zone	Clusters
67	N	Rare	Moderate	1	25	20978	South	5
58	N	Rare	Moderate	3	27	6170	North	0
73	N	Daily	Extreme	0	32	28382	North	3
71	Y	Rare	No	3	37	27148	South	1
74	N	No	Extreme	0	34	29616	South	0

3.14 — Partial data head with labels in the column 'Clusters'

3.5.3 Cluster Descriptions

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

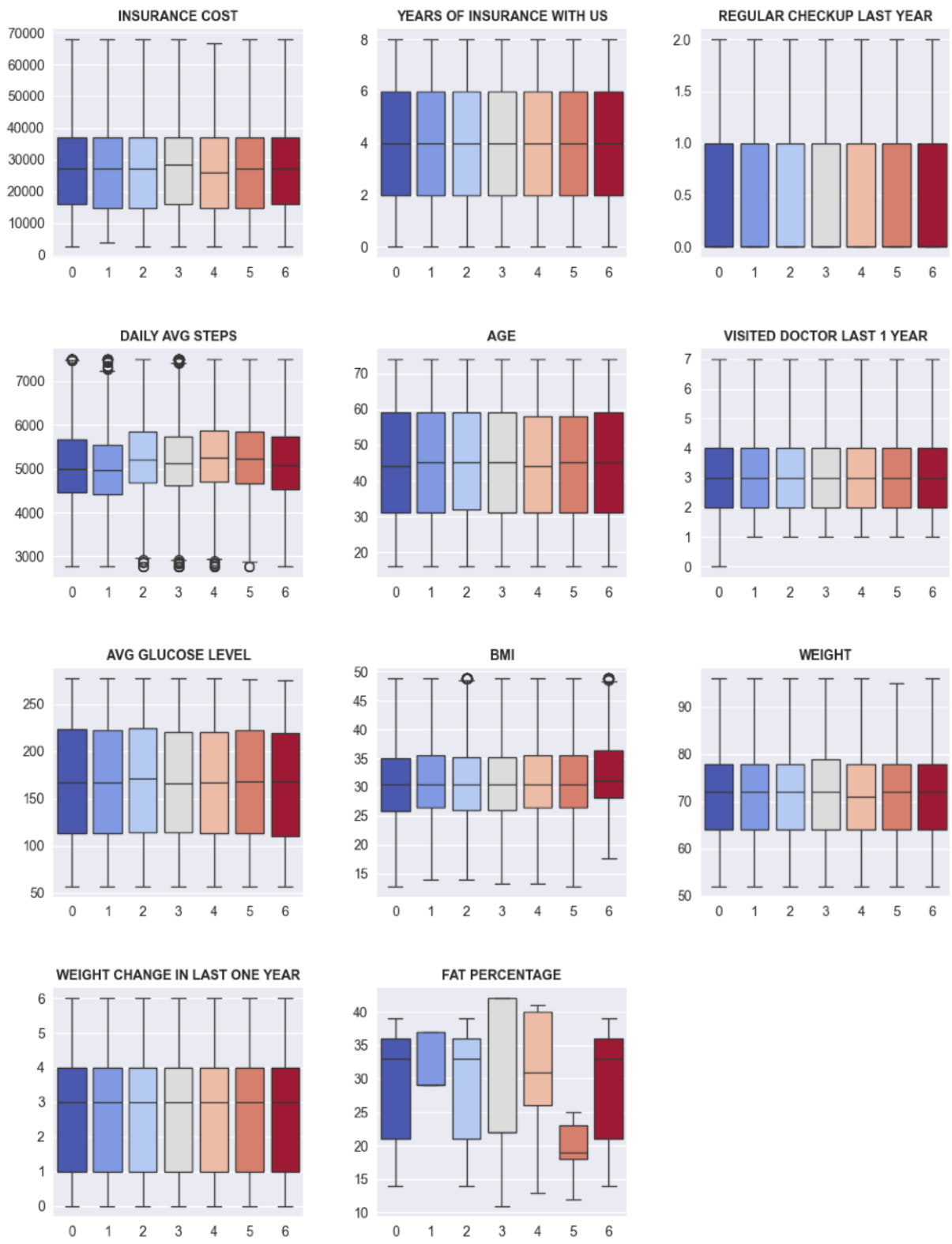
Cluster 4 — High Cholesterol, Salaried

Cluster 5 — Low Cholesterol, Low Fat Percentage, Salaried

Cluster 6 — All Smokers

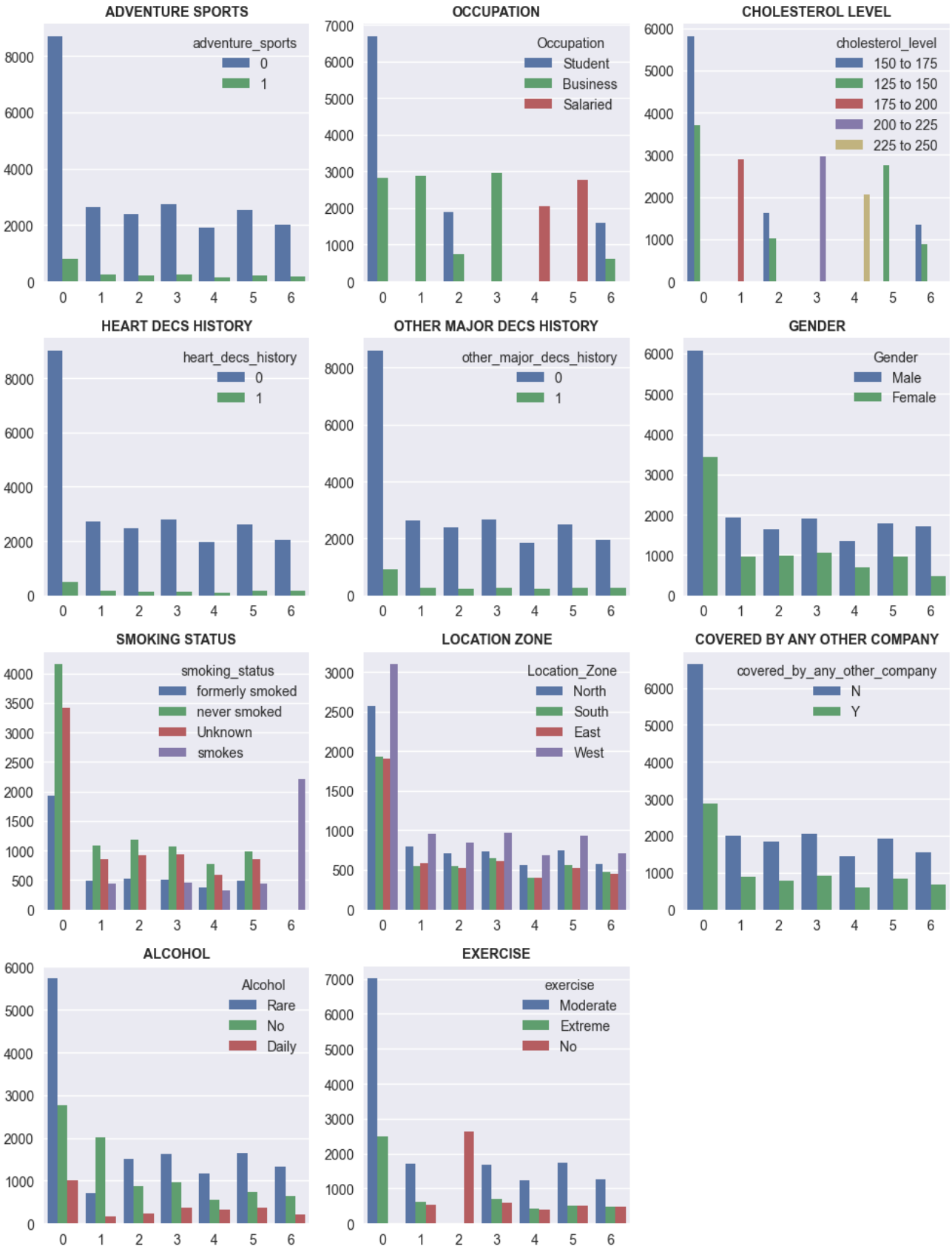
3.6 EDA on Clustered Data

3.6.1 Numerical Variables



3.15 — Boxplots of Numerical variables by cluster

3.6.2 Categorical Variables — Countplots



3.16 — Countplots of Categorical Variables by Cluster

3.6.3 Categorical Variables — Heatmaps

3.17 — Heatmaps of counts of variables by cluster

4.0 Insights from EDA

4.1 On Entire Data

4.1.1 Numerical Variables

- Mean insurance cost is Rs. 27,147 and 16% people have insurance cost > 40,000 and 16% people have insurance cost < 10,000.
- People are equally distributed across number of Years with the company.
- Most people don't have Regular Checkups.
- Daily Avg. Steps follows a perfectly normal distribution.
- Age groups are well distributed, i.e. there are somewhat equal number of people from all age groups.
- Average random Glucose Level should be no more than 125, while there are 17,000+ people above 125, that's 68%+.
- BMI should ideally be < 25 while there are 20,000+ people above 25, which is a staggering 80%+.
- Weight is mostly between 60 and 80 kgs which doesn't tell us much.
- If the upper limit of healthy fat percentage is taken as 25%, then there are 15,000+ people above that which is 75%+ of the population.

Conclusion

- Mean insurance cost is average with plently of people on the lower side.
- People are generally unhealthy when it comes to diabetic and heart health markers like Glucose Level, BMI, and Fat Percentage + they do not have Regular Checkups.
- So we are dealing here with a population at risk.

4.1.2 Categorical Variables

- Very few people indulge in Adventure Sports.
- Salaried people are the fewest, half of other Occupations.
- Cholesterol is mostly on the lower side.
- Very few people have a history of Heart Disease or Other Major Diseases.
- Males are twice more than Females.
- Smokers are ~15% of the sample.
- Twice more people are Covered by Other Company than not.
- Daily Alcohol consumers are ~10% of the sample.
- Most people do Moderate exercise and 20% people don't exercise at all.

Conclusions

- There are positive points here — people playing it safe by not indulging in Adventure Sports, Cholesterol is on the lower side, no history of Heart or other Diseases, Smokers are less, Alcohol consumers are less, Moderate exercisers are more.
- This is will all equate to less risk for our insurance company.

4.1.3 Relation of Target Variable to All Other Variables

- The target variable is mostly not correlated to other numerical variables except Weight. Weight is highly correlated (97%) to the Insurance Cost.
- Women have lesser BMI than Men.
- There is no change in the Mean of Insurance Cost for Males and Females for categorical variables except:
    - Those practicing Adventure Sports have a 20% higher Insurance Cost on average.
    - Those Covered by Other Company have 15% higher cost.

- Age — Insurance Cost is not related to Age at all, there are no patterns emerging when its differentiated by Occupation.
- BMI — Insurance Cost is not related to BMI either. We only see that Salaried people are more in the North.
- Avg. Glucose — Salaried people have the most Cholesterol and Students have the least.

Conclusions

- In the current model Insurance Cost is directly increasing as Weight of the individual increases.
- Other factors influencing the cost are Adventure Sports and being Covered by Other Company.
- It's not taking into account negative factors like Glucose Level, BMI, and Fat Percentage and positive ones like Cholesterol is on the lower side, no history of Heart or other Diseases, Smokers are less, Alcohol consumers are less, Moderate exercisers are more.

4.2 EDA on Clustered Data

- Cluster 0 has most number of values at 9500+ or 38%+.
- The only major difference between the Clusters in Numerical Variables is the Fat Percentage — Cluster 3 has the highest and Cluster 5 has the lowest median fat percentage.
- Occupation — Cluster 0 has maximum number of students, 3 has Business only, and Cluster 5 is Salaried only.
- Cholesterol — 0 has the lowest, 3 and 4 have the highest.
- Heart Disease, Other Diseases, and Gender are all in similar ratios in all clusters.
- Smoking — All Smokers are in Cluster 6.
- Zones — All Zones have a similar distribution across clusters.

- Exercise — Non-Exercisers are in Cluster 2, Moderate Exercisers in Cluster 0.

Conclusions

- Cluster 0 (Largest Group, 38%) – Students, Moderate Exercisers, Low Cholesterol: This cluster presents a young and relatively healthy population, with many being students and moderate exercisers.
- Cluster 3 (Business, High Fat Percentage & Cholesterol): This cluster consists solely of business professionals with the highest median fat percentage and cholesterol levels, indicating potential health risks.
- Cluster 5 (Salaried, Lowest Fat Percentage): This salaried group has the lowest fat percentage, suggesting a relatively healthier lifestyle.
- Cluster 6 (Smokers): All smokers are grouped into this cluster, signifying a high-risk group in terms of health outcomes.

5.0 Modeling — Parametric Models

5.1 Linear Regression

OLS Regression Results

Dep. Variable:	insurance_cost	R-squared:	0.945
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	1.100e+04
Date:	Thu, 03 Oct 2024	Prob (F-statistic):	0.00
Time:	13:27:56	Log-Likelihood:	-1.9085e+05
No. Observations:	20000	AIC:	3.818e+05
Df Residuals:	19968	BIC:	3.820e+05
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.966e+04	354.396	-224.781	0.000	-8.04e+04	-7.9e+04
years_of_insurance_with_us	-10.8282	9.505	-1.139	0.255	-29.459	7.803
regular_checkup_last_year	-692.1947	30.002	-23.072	0.000	-751.000	-633.389
adventure_sports	133.8469	87.369	1.532	0.126	-37.403	305.097
visited_doctor_last_1_year	-34.3922	21.796	-1.578	0.115	-77.113	8.329
daily_avg_steps	-0.0424	0.025	-1.667	0.096	-0.092	0.007
age	3.2608	1.482	2.201	0.028	0.356	6.165
heart_decs_history	153.2690	106.652	1.437	0.151	-55.777	362.315
other_major_decs_history	22.1984	82.981	0.268	0.789	-140.451	184.848
avg_glucose_level	0.2346	0.381	0.616	0.538	-0.512	0.981
bmi	-0.5339	3.668	-0.146	0.884	-7.724	6.656
weight	1488.2576	2.807	530.252	0.000	1482.756	1493.759
weight_change_in_last_one_year	157.5977	15.269	10.321	0.000	127.669	187.526
fat_percentage	-0.8330	3.089	-0.270	0.787	-6.888	5.222
Occupation_Salaried	-11.0116	129.699	-0.085	0.932	-265.233	243.210
Occupation_Student	20.3935	82.600	0.247	0.805	-141.509	182.296
cholesterol_level_150_to_175	-71.5231	76.787	-0.931	0.352	-222.031	78.985
cholesterol_level_175_to_200	-13.7180	121.506	-0.113	0.910	-251.879	224.443
cholesterol_level_200_to_225	4.6667	121.183	0.039	0.969	-232.862	242.195
cholesterol_level_225_to_250	164.5084	114.948	1.431	0.152	-60.800	389.817
Gender_Male	38.6808	56.291	0.687	0.492	-71.654	149.016
smoking_status_formerly_smoked	-42.7818	76.346	-0.560	0.575	-192.427	106.863
smoking_status_never_smoked	-1.8311	61.805	-0.030	0.976	-122.973	119.311
smoking_status_smokes	-54.9208	78.067	-0.704	0.482	-207.938	98.096
covered_by_any_other_company_Y	1190.6396	53.986	22.055	0.000	1084.823	1296.456
Alcohol_No	-1.8877	84.785	-0.022	0.982	-168.073	164.298
Alcohol_Rare	30.9390	79.949	0.387	0.699	-125.768	187.646
exercise_Moderate	48.8713	60.958	0.802	0.423	-70.612	168.355
exercise_No	47.1070	74.638	0.631	0.528	-99.189	193.403
Location_Zone_North	105.4790	70.648	1.493	0.135	-32.998	243.956
Location_Zone_South	32.2895	75.287	0.429	0.668	-115.279	179.858
Location_Zone_West	4.0743	67.705	0.060	0.952	-128.634	136.782

Omnibus:	574.108	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	662.502
Skew:	0.383	Prob(JB):	1.38e-144
Kurtosis:	3.457	Cond. No.	8.08e+04

5.1 — Summary of initial model

OLS Regression Results

Dep. Variable:	insurance_cost	R-squared:	0.945
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	6.822e+04
Date:	Thu, 03 Oct 2024	Prob (F-statistic):	0.00
Time:	13:29:35	Log-Likelihood:	-1.9086e+05
No. Observations:	20000	AIC:	3.817e+05
Df Residuals:	19994	BIC:	3.818e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.993e+04	232.417	-343.899	0.000	-8.04e+04	-7.95e+04
regular_checkup_last_year	-691.6901	29.964	-23.084	0.000	-750.422	-632.958
age	3.2449	1.481	2.191	0.028	0.342	6.148
weight	1488.4637	2.799	531.877	0.000	1482.978	1493.949
weight_change_in_last_one_year	157.6103	15.256	10.331	0.000	127.708	187.513
covered_by_any_other_company_Y	1172.8538	52.032	22.541	0.000	1070.866	1274.841

Omnibus:	565.279	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	649.896
Skew:	0.380	Prob(JB):	7.53e-142
Kurtosis:	3.448	Cond. No.	833.

5.2 — Model summary after removing features with high p-value

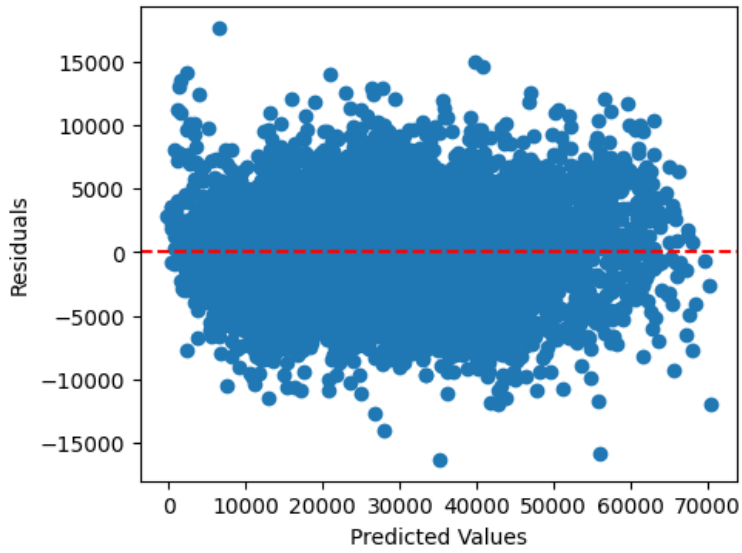
Evaluation — LR

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.15	3374	0.945	0.945	0.15	3339	0.945	0.945	Slight bias

5.2 Polynomial Regression

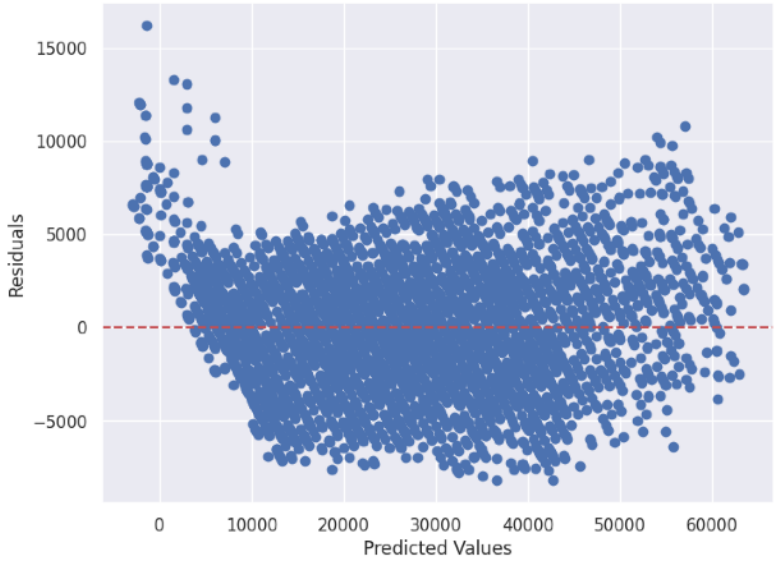
5.2.1 Degree 2

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.14	3166	0.951	0.951	0.14	3237	0.949	0.948	Unbiased

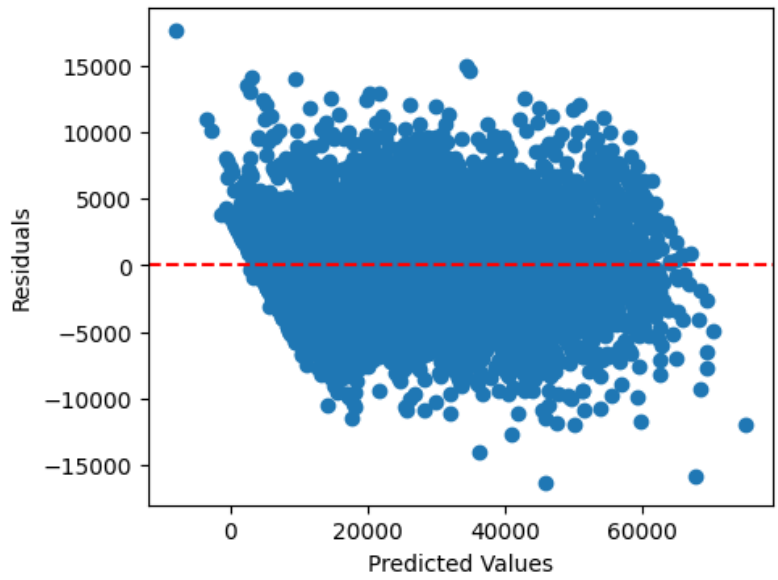


5.4 — Residual Plot PR deg. 2

5.2.2 Degree 3



5.3 — Residual Plot - LR



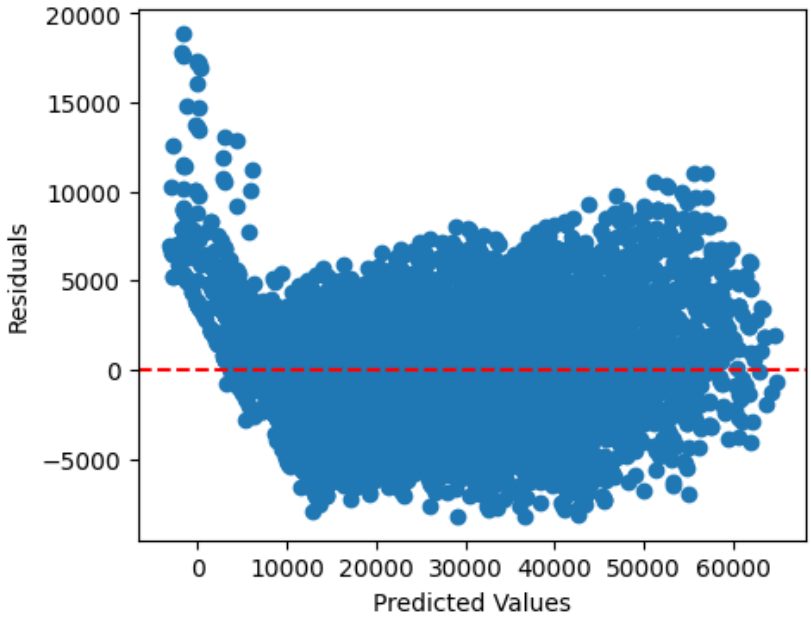
5.5 — Residual Plot deg. 3

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.11	2683	0.965	0.965	0.16	3795	0.929	0.929	Slight bias

5.3 Ridge

Train		Test		

MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	Bias
0.15	3363	0.945	0.945	0.16	3369	0.944	0.944	Slight bias

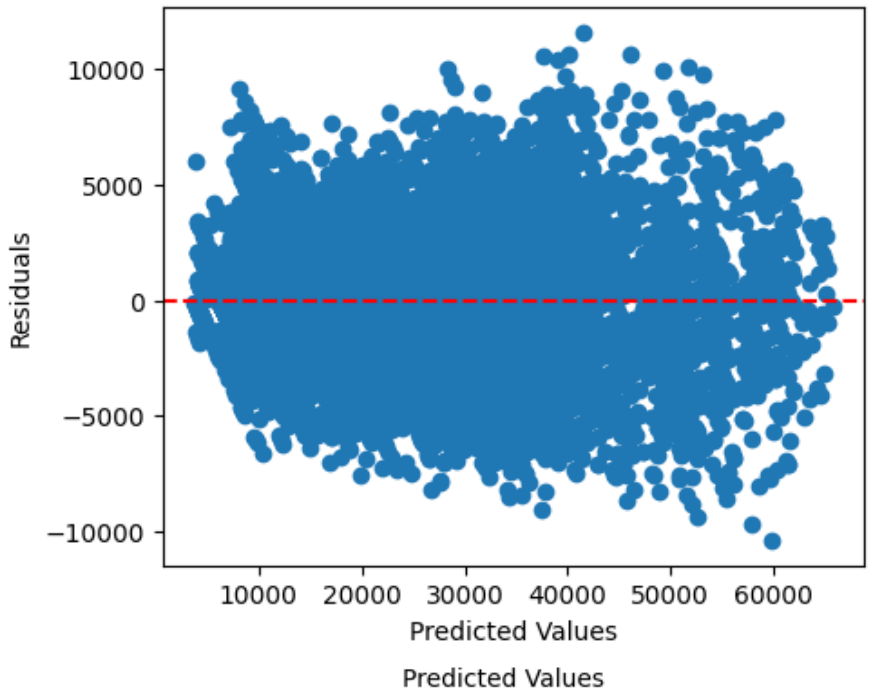


5.6 — Residual Plot - Ridge

5.4 Lasso

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.15	3363	0.945	0.945	0.16	3369	0.944	0.944	Slight bias

6.0



Predicted Values

6.1 — Residual Plot - RF

5.7 — Residual Plot - Lasso

Modeling — Non-Parametric Models

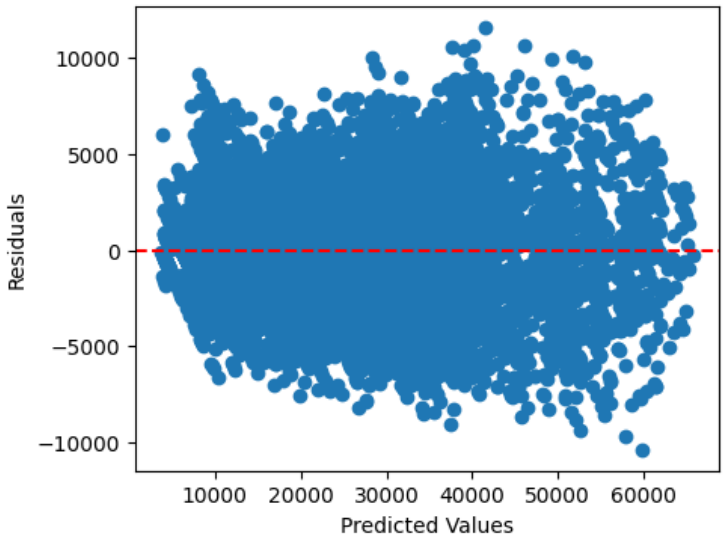
6.1 Random Forest Regressor

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.05	1173	0.993	0.993	0.12	3105	0.953	0.953	Unbiased

6.2 XG Boost Regressor

6.2.1 Normal Model

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.09	2250	0.975	0.975	0.13	3150	0.95	0.95	Unbiased

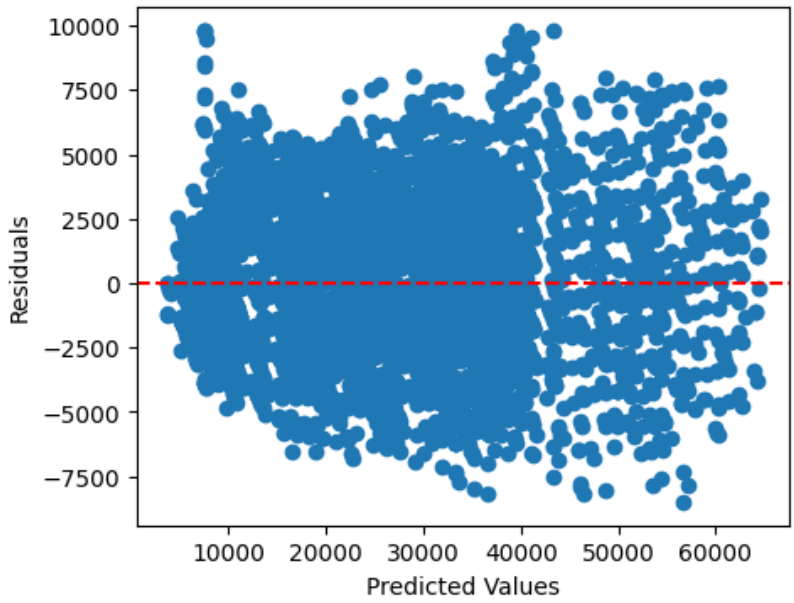


6.2 — Residual Plot - XG Normal

6.2.2 XGB Model Tuning

Best parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.12	2997	0.956	0.956	0.12	3001	0.956	0.956	Slight Bias



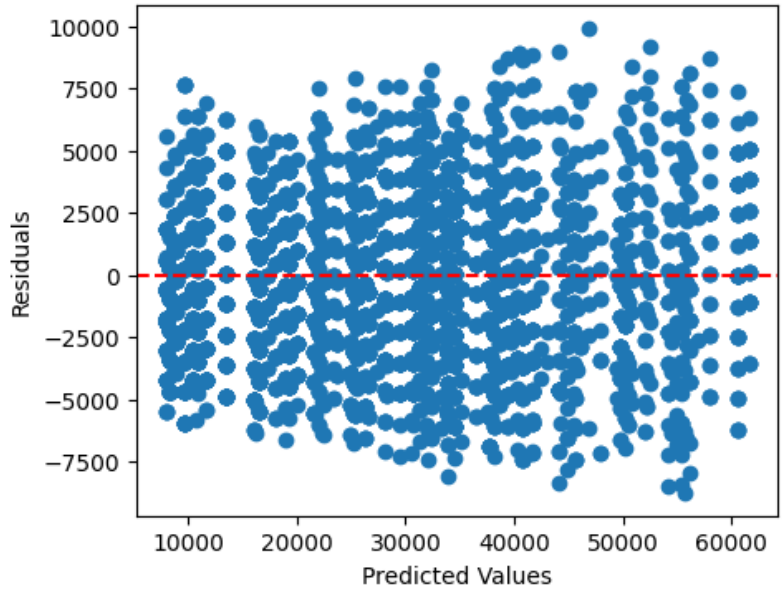
6.3 — Residual Plot - XGB Tuned

6.3 AdaBoost Regressor

6.3.1 Normal Model

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.15	3172	0.951	0.951	0.15	3147	0.951	0.951	Biased

6.3.2

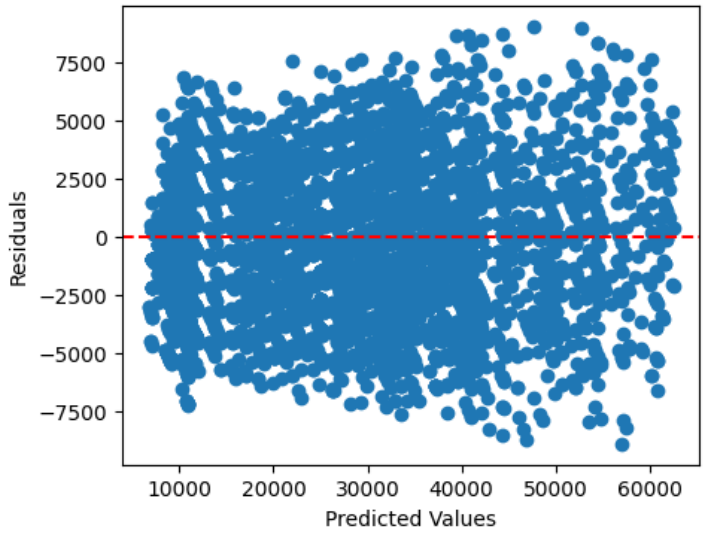


2.4 — Residual Plot - AdaBoost

AdaBoost Model Tuned

Best Params: {'estimator__max_depth': 5, 'n_estimators': 50}

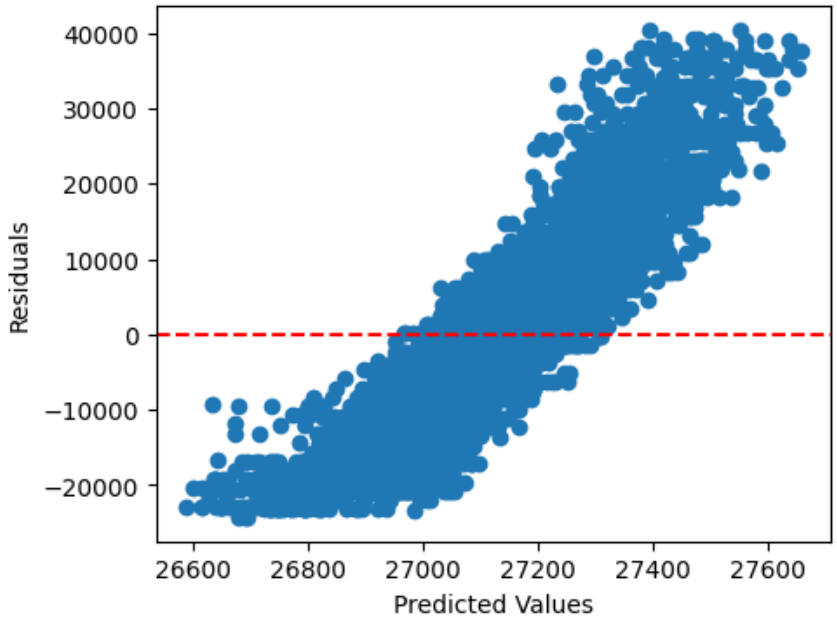
Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.15	3168	0.951	0.951	0.15	3181	0.950	0.950	Slight Bias



2.5 — Residual Plot - ADB Tuned

6.4 Support Vector Regression

Train				Test				Bias
MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
0.80	14333	0.00	0.00	0.80	14271	0.00	0.00	Biased



2.6 — Residual Plot - SVR

7.0 Model Selection

7.1 Comparison

Model	Train				Test				Bias
	MAPE	RMSE	R^2	Adj. R^2	MAPE	RMSE	R^2	Adj. R^2	
Parametric									
LR	0.15	3374	0.945	0.945	0.15	3339	0.945	0.945	Slight bias
PR deg. 2	0.14	3166	0.951	0.951	0.14	3237	0.949	0.948	Unbiased
PR 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
Non-Parametric									
RF	0.05	1173	0.993	0.993	0.12	3105	0.953	0.953	Unbiased
XGB	0.09	2250	0.975	0.975	0.13	3150	0.95	0.95	Unbiased
XGB Tuned	0.12	2997	0.956	0.956	0.12	3001	0.956	0.956	Slight Bias
ADB	0.15	3172	0.951	0.951	0.15	3147	0.951	0.951	Biased
ADB Tuned	0.15	3168	0.951	0.951	0.15	3181	0.950	0.950	Slight Bias
SVR	0.80	14333	0.00	0.00	0.80	14271	0.00	0.00	Biased

7.2 Selection

The selected model is **2.2.1 Random Forest**. Here is the comparion with the 2nd best model 2.2.2 XG Boost Tuned:

RMSE

RF: 1173 (Train), 3105 (Test) → Much lower Train RMSE.

XGB Tuned: 2997 (Train), 3001 (Test) → XGB Tuned has a slightly better test RMSE than RF, but RF still performs better on train data.

R^2

RF: 0.993 (Train), 0.953 (Test) → Higher R^2, indicating better fit.

XGB Tuned: 0.956 (Train and Test) → Slightly lower R^2 on both train and test.

Bias

RF: Unbiased.

XGB Tuned: Slight bias.

Conclusion

Random Forest (RF) model is selected because it has the lowest RMSE (1173 on training and 3105 on testing), the highest R^2 (0.993 on training and 0.953 on testing), and is unbiased. This indicates that it has strong predictive accuracy and generalizes well to new data.

7.3 Important Features

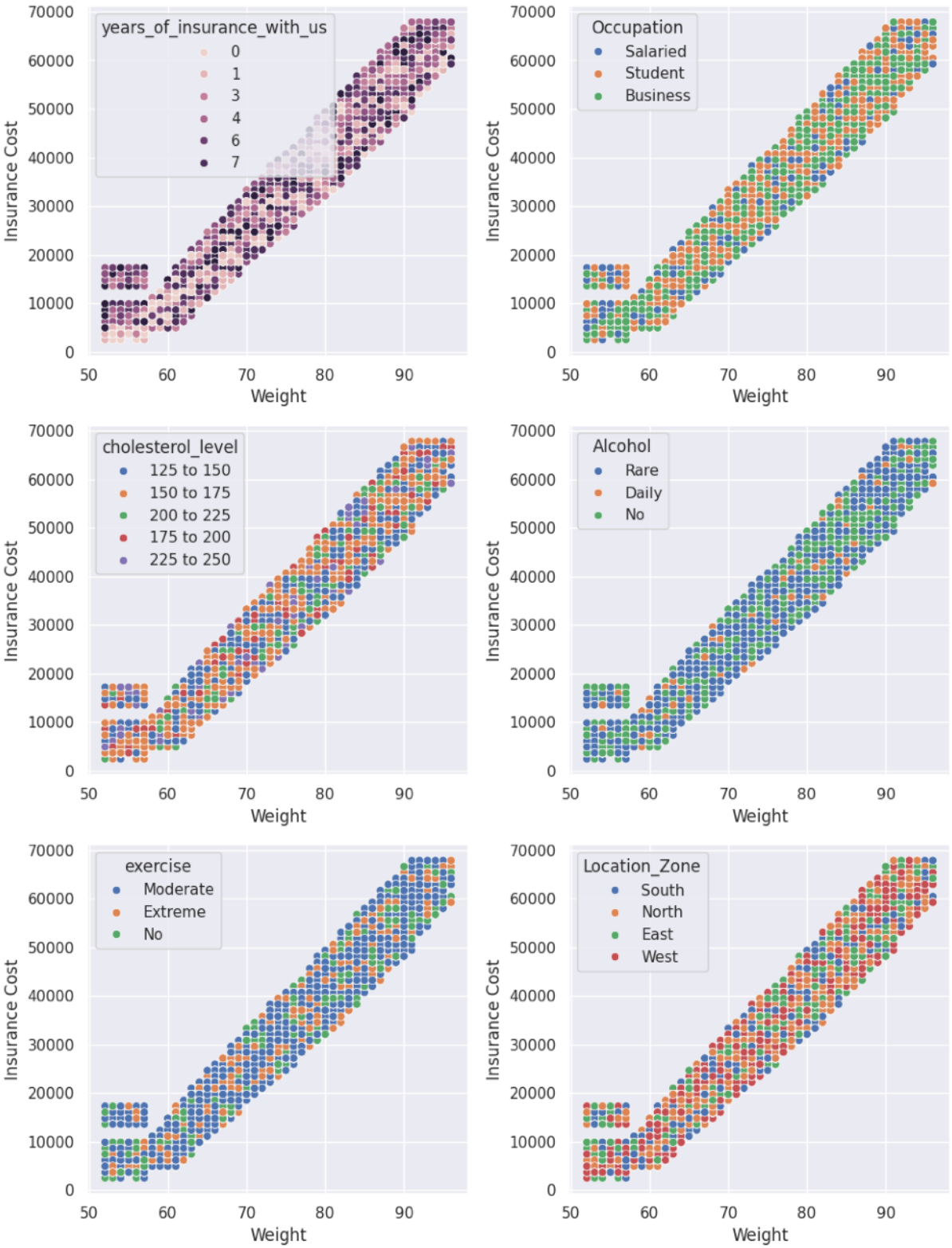
	Feature	Importance
10	weight	0.951700
4	daily_avg_steps	0.005892
8	avg_glucose_level	0.005730
9	bmi	0.005575
5	age	0.005060

7.1 — Important Features - RF Model

There is essentially just one important feature — ‘weight’ (Weight of the Individual) with 95%+ importance. The very next feature has an importance of only 0.5%.

8.0 Visual Analysis with Important Features

8.1 With Target Variable

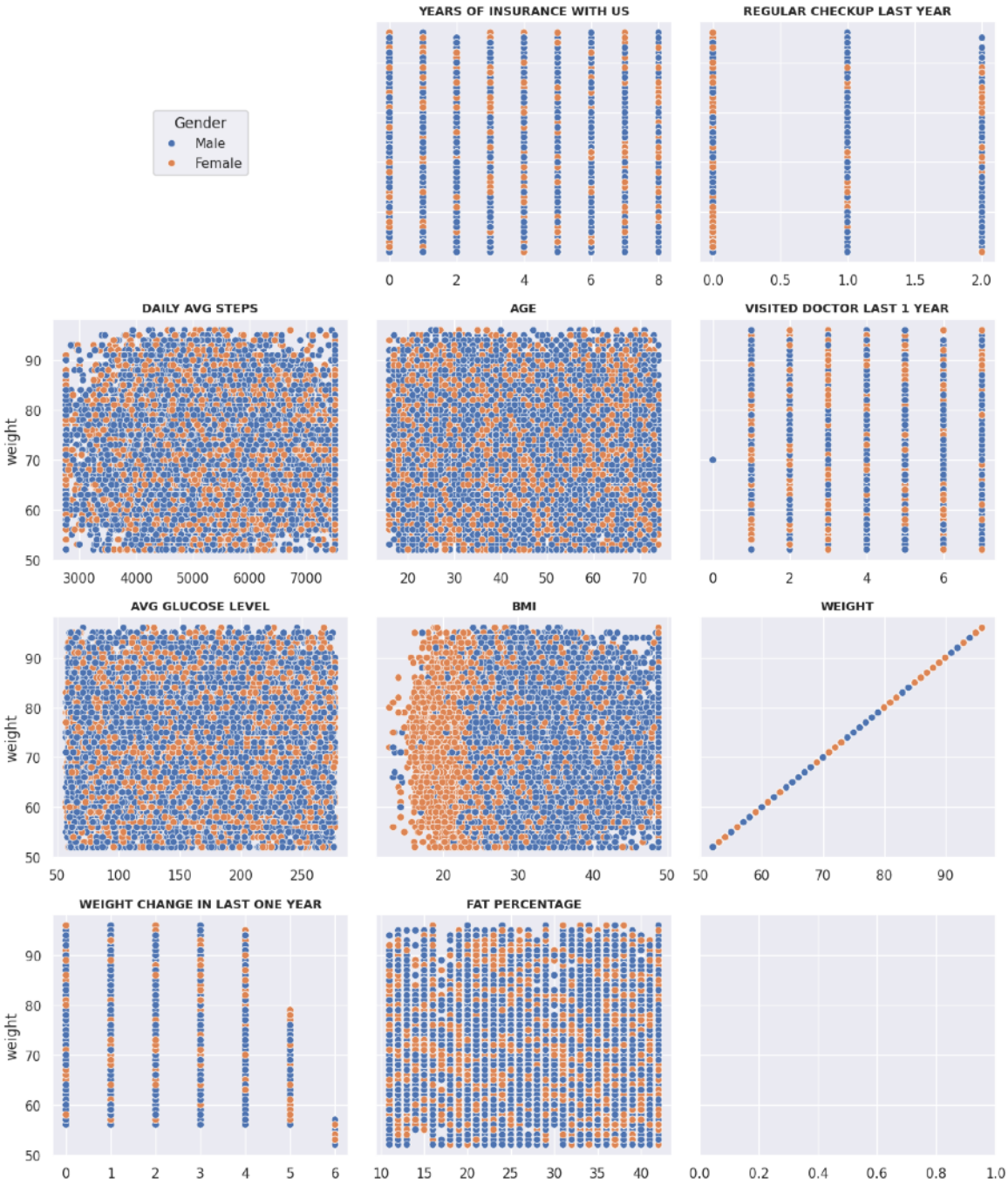


8.1 — Weight vs Insurance Cost differentiated by selected categorical features

Analysis

- Weight is highly (97%) correlated to Insurance Cost
- None of the other important categorical variables differentiate the scatter plot in any meaningful way.

8.2 With Numerical Variables

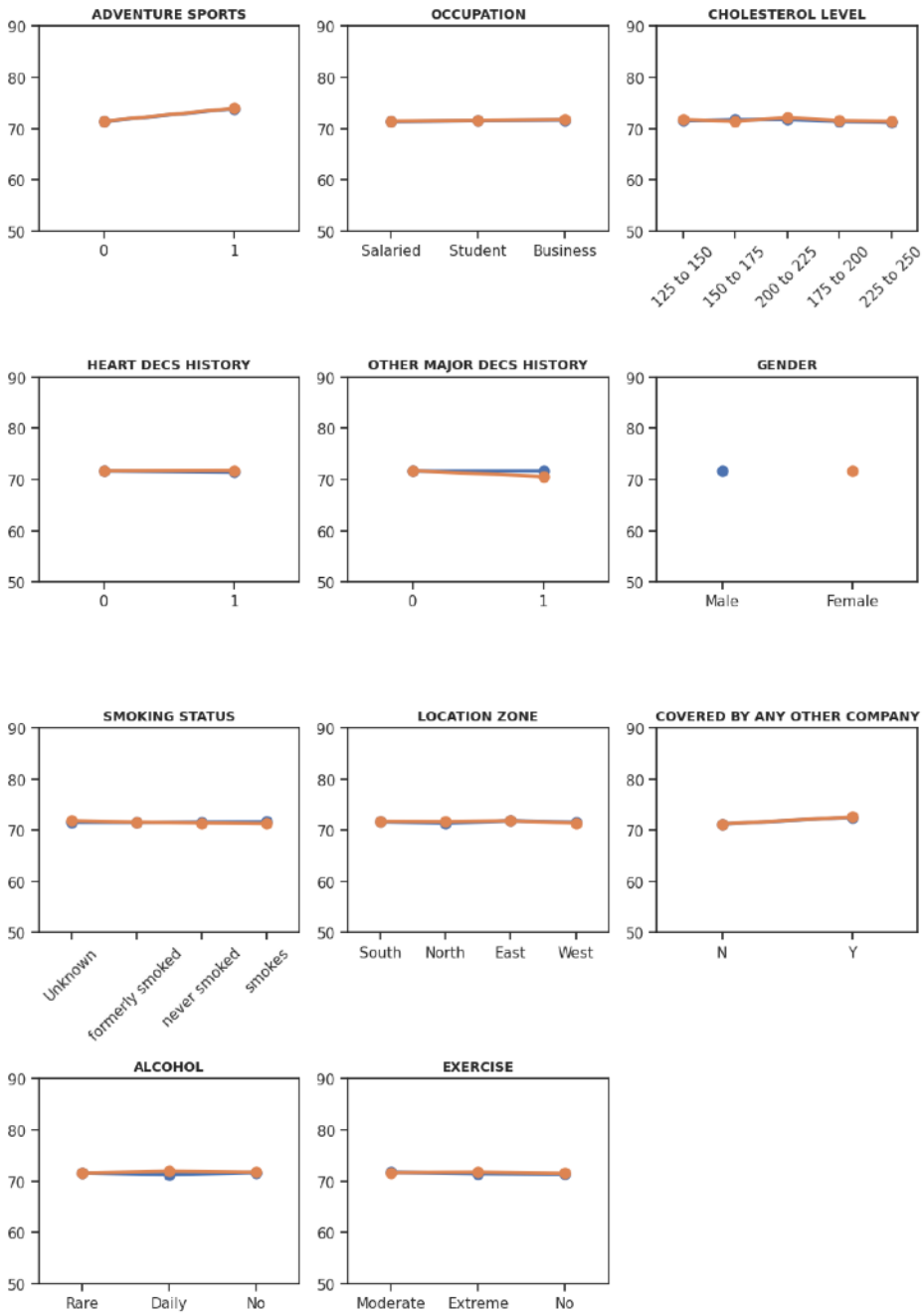


8.2 — 'weight' in relation to Numerical features

Analysis

- There is no meaningful pattern between ‘weight’ and other numerical features. The data is highly homogenous.
- Gender differentiation only shows us that Females have lesser BMI.

8.3 With Categorical Variables



8.3 — Mean of ‘weight’ for Categorical variables differentiated by Gender

Analysis

- There is very little variation between the mean of ‘weight’ for Categorical variables for both Genders.
- Those participating in Adventure Sports have slightly higher weight for both men and women.

9.0 Final Insights & Recommendations

9.1 Key Insights

1. **Weight Dominates Insurance Cost Predictions:**
   - Weight is by far the most influential factor in predicting insurance costs (97% correlation), suggesting that heavier individuals are associated with higher costs.
   - No other feature, including lifestyle or categorical variables, significantly impacts cost predictions.
2. **Low Variability in Other Features:**
   - The minimal differentiation in categorical variables (e.g., Gender, Adventure Sports) means that insurance cost predictions are largely unaffected by demographic or behavioral factors outside of weight.
   - Adventure sports participants have slightly higher weights, but the impact is minimal.
3. **Homogeneity in the Data:**
   - The data shows little variability, especially in key health and lifestyle indicators (e.g., cholesterol, exercise). This suggests that the insured population may be relatively similar in terms of health risks, making weight the primary differentiator.
4. **Gender Differences:**
   - Females generally have lower BMI compared to males, but this difference is not substantial enough to drive cost variations significantly.

9.2 Recommendations

1. **Model is not Recommended to be Used**
    - The fact that the model is giving such overwhelming importance to the 'weight' variable (96.6%) raises some concerns regarding its reliability and generalizability for insurance cost predictions.
    - While the model's performance metrics are excellent, this over-reliance on a single feature can be problematic.
    - Relying too heavily on 'weight' could result in biased predictions where individuals with similar weights but very different health profiles are treated the same.
2. **Weight Management Programs**
    - Since weight is the main driver of insurance costs, introducing or enhancing weight management programs could help lower long-term costs for both the company and the insured.
    - Incentivizing healthier lifestyles through premium discounts for weight loss or maintaining a healthy BMI could be beneficial.
3. **Targeted Wellness Programs**
    - Smokers (Cluster 6): These customers might represent a higher health risk. Offering smoking cessation programs or higher premiums for smokers could mitigate potential risks.
    - Non-exercisers (Cluster 2): Encouraging this group to adopt exercise habits by offering discounts for those participating in fitness programs or tracking steps could reduce long-term costs.
    - Cholesterol and Fat Management (Clusters 3, 4, 5): Tailoring diet and cholesterol-lowering programs to individuals with higher cholesterol levels or fat percentages may help manage risks in these groups.

4. **Personalized Premium Adjustments**
    - Adventure Sports Participation: Since adventure sports enthusiasts tend to have slightly higher weights, their premiums could be adjusted to reflect the potential for higher risk, though this increase should be moderate due to the marginal difference in weight observed.
    - Gender-Based Insights: Given that females generally have a lower BMI, gender-specific incentives for maintaining healthy weight and BMI could be introduced, potentially resulting in lower premiums for women who meet specific health targets.
5. **Cluster-Based Pricing and Offers**
    - Cluster 0 (Students): Younger, healthier individuals (low cholesterol, moderate exercise) may warrant lower premiums or introductory offers, encouraging them to stay long-term with the company.
    - Business Owners (Clusters 1 and 3): These customers might be more risk-prone due to high cholesterol and fat percentage. Premiums for this group could be slightly higher, with incentives for health improvement initiatives.
6. **Expand Data Collection**
    - The homogeneity and low feature importance of other factors suggest that additional data might be necessary to capture a fuller picture of health risks. Consider incorporating more granular health data (e.g., metabolic health indicators, stress levels, or genetic predispositions) to refine your cost models further.
