HEALTH INSURANCE CLAIM

OBJECTIVE

- Identifying the primary factors that inflate claim.
- Creating a model that will predict future claim based on the features.
- Interpreting the model and how it affects claim.

ABOUT DATASET AND SOURCE

This is a health insurance dataset which consist of information that are collected from the policy holder. This information contains details that will be used for analysis, prediction and interpretation. The dataset is made up of 13 columns and 15000 rows.

The data was downloaded from Kaggle.com, a subsidiary of google LLC. It is public and isn't bounded by privacy law.

DATA DEFINITION

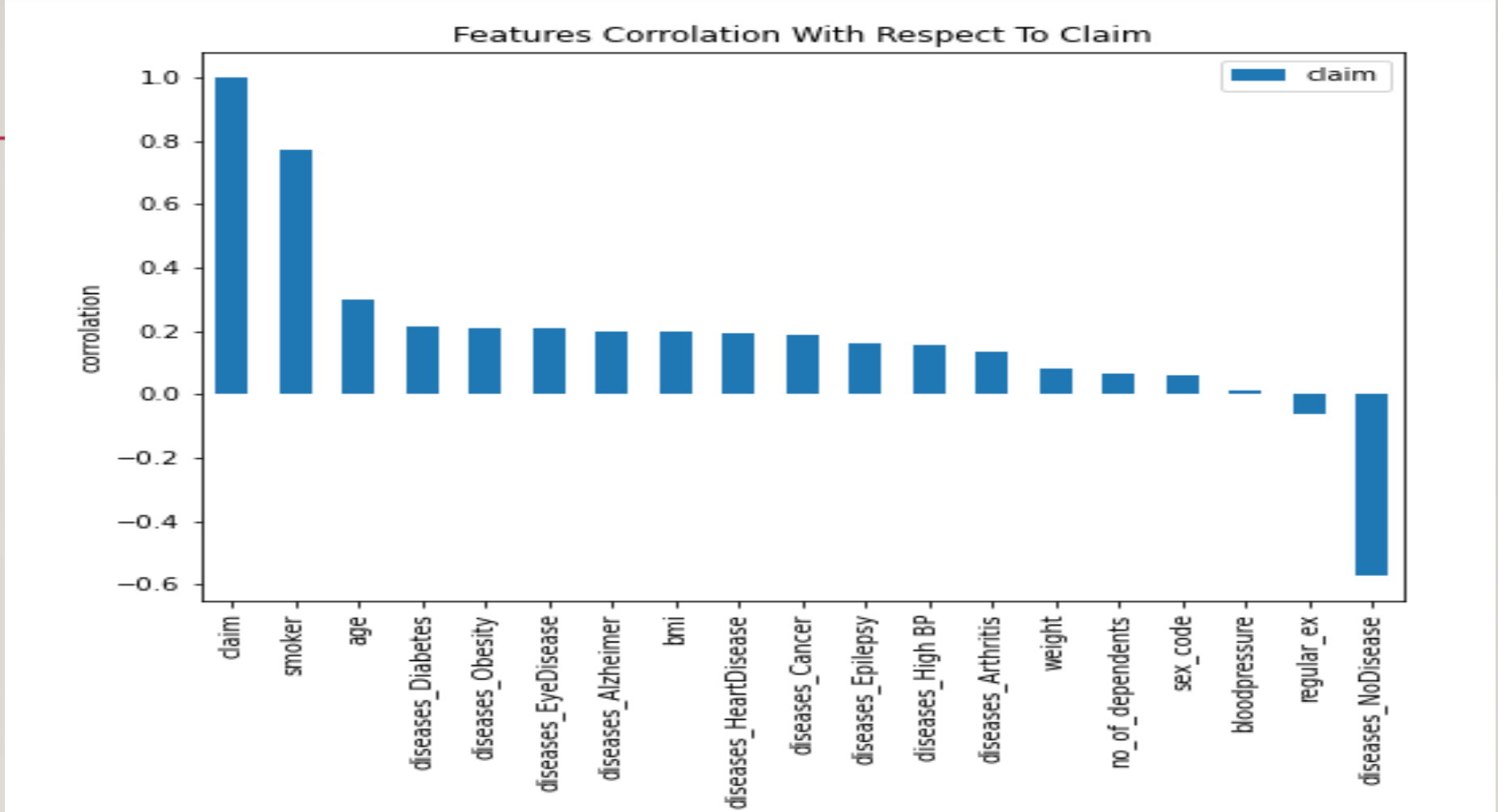
- AGE : Age of the policyholder (Numeric)
- SEX: Gender of policyholder (Categoric)
- WEIGHT: Weight of the policyholder (Numeric)
- BIM: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight (Numeric)
- NOOF DEPENDENT: Number of dependent persons on the policyholder (Numeric)
- SMOKERS: Indicates policyholder is a smoker or a non-smoker (non-smoker=0;smoker=1) (Categoric)
- CLAIMS: The amount claimed by the policyholder (Numeric)
- BLOOD PRESSURE: Blood pressure reading of policyholder (Numeric)
- DIABETISES: Indicates policyholder suffers from diabetes or not (non-diabetic=0; diabetic=1) (Categoric)
- REGULAR_EX: A policyholder regularly exercises or not (no-exercise=0; exercise=1) (Categoric)
- JOB_TITTLE: Job profile of the policyholder (Categoric)
- CITY: The city in which the policyholder resides (Categoric)
- HEREDIRORY DISEASES: A policyholder suffering from a hereditary diseases or not (Categoric)

ANALYSIS PROCESS

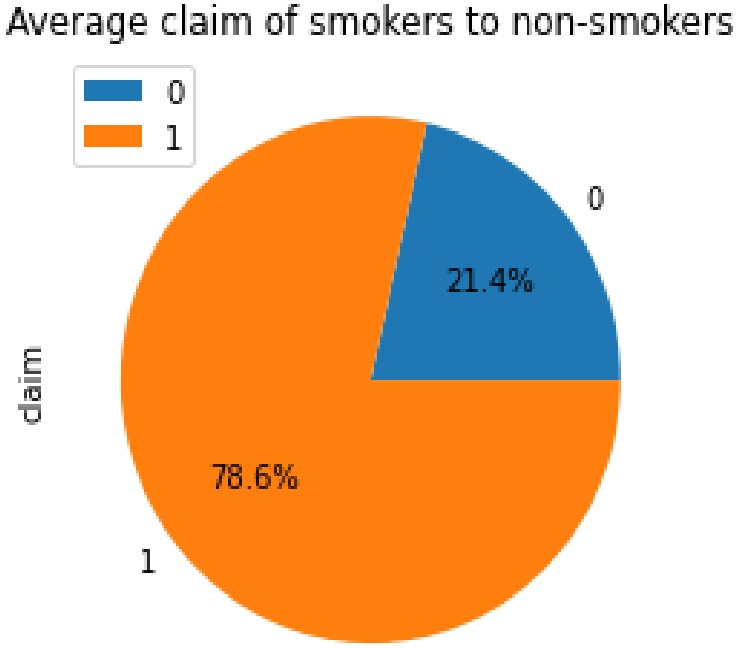
- Data Cleaning
  - Replacing Empty Roles With The Average Value.
  - Removing Duplicates
  - Checking for Outliers
- Exploratory Data Analysis
- Analysis and Insight From Dataset
- Model Development
  - Prediction
  - Interpretation
- Conclusion

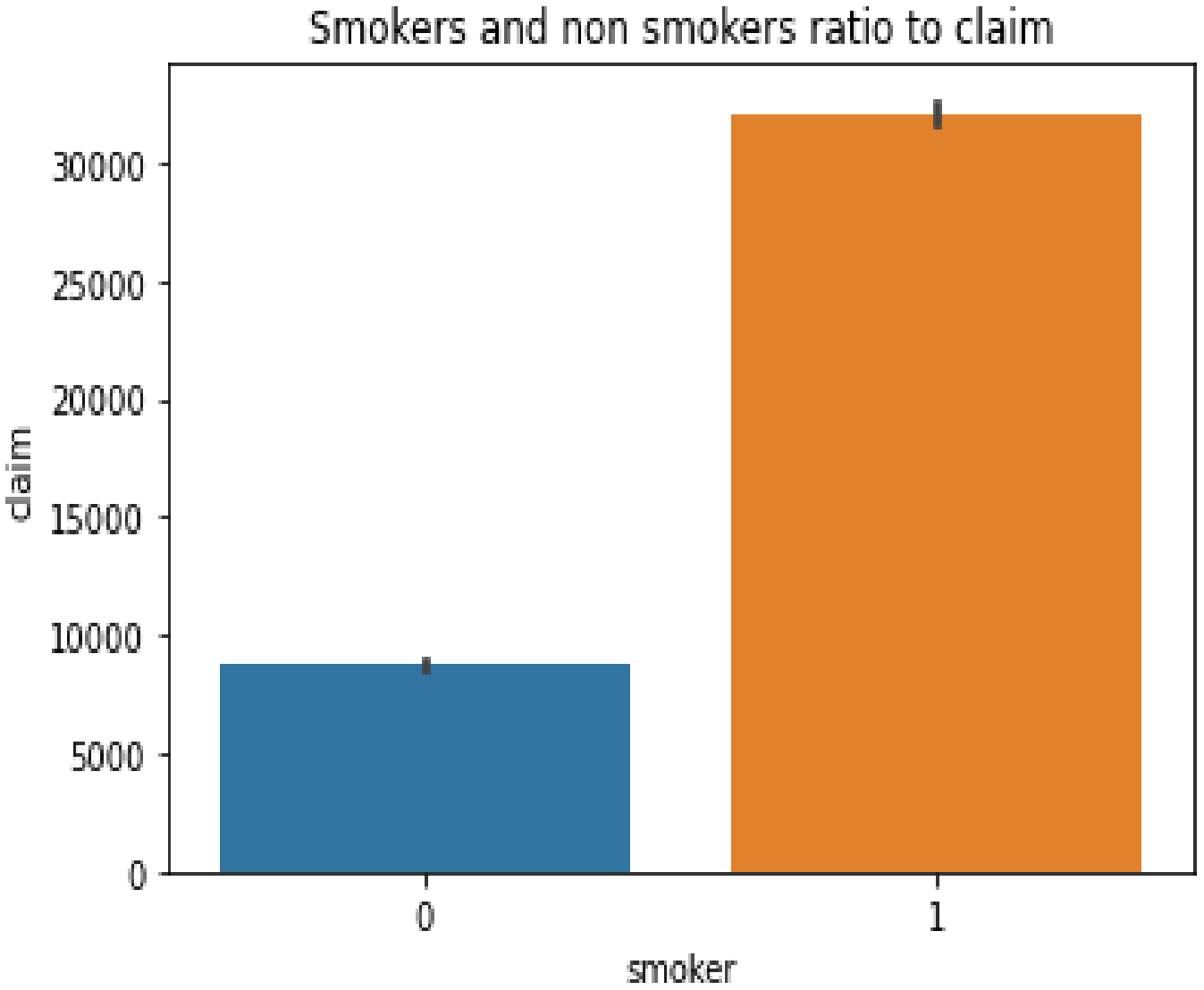
EXPLORATORY DATA ANALYSIS

The table shows the summary of the dataset and how much the features greatly affect claim.

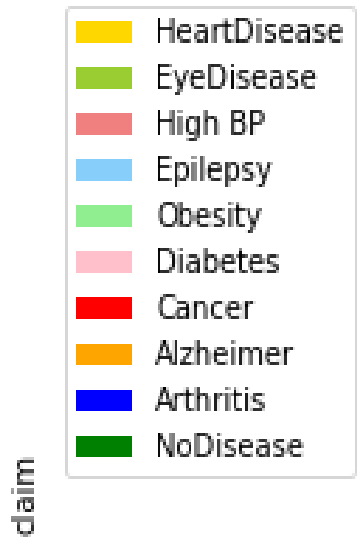


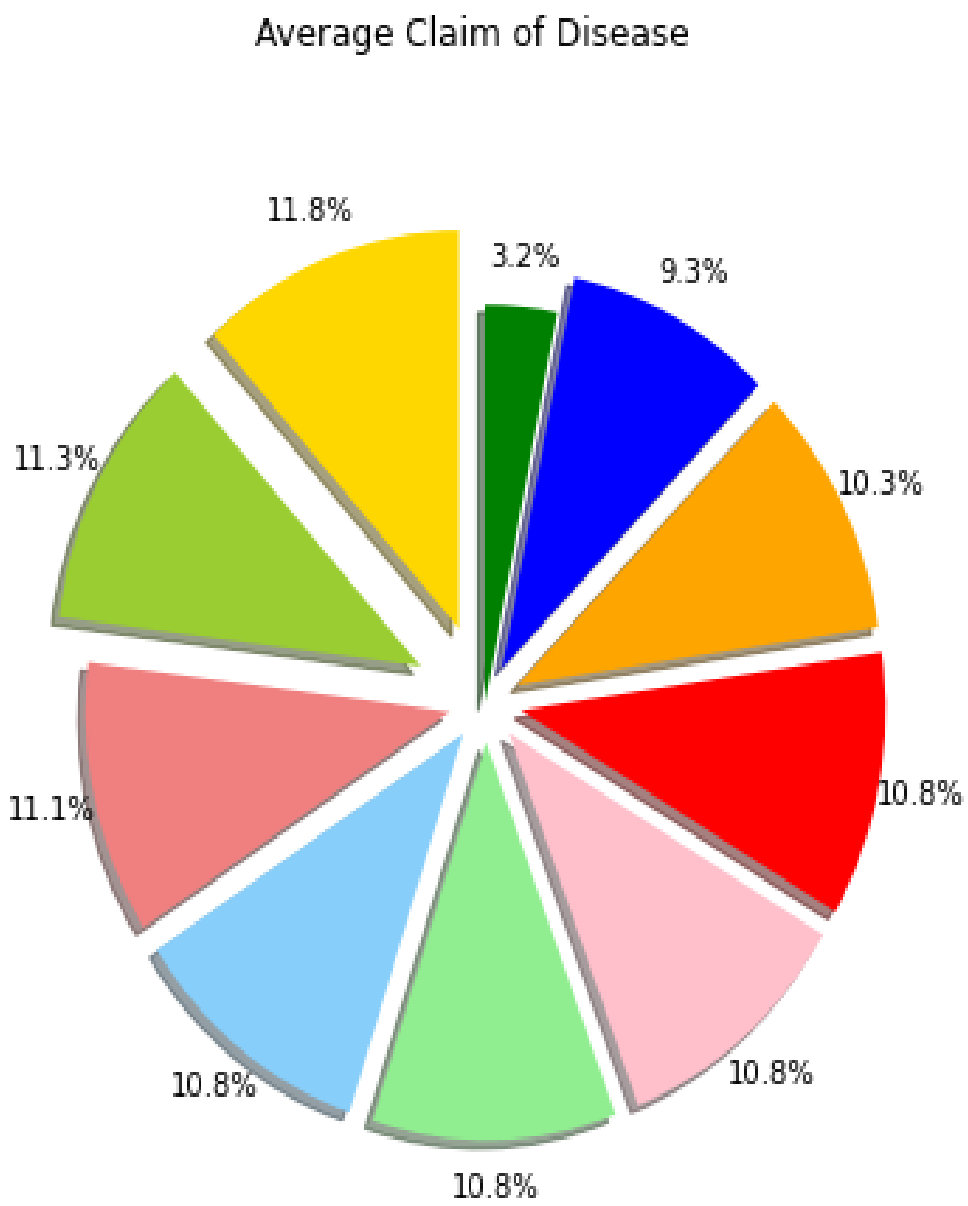
smoker	claim
0	8745.04
1	32101.65





diseases	claim
HeartDisease	43086.02
EyeDisease	41188.11
High BP	40605.91
Epilepsy	39692.06
Obesity	39589.33
Diabetes	39448.45
Cancer	39445.37
Alzheimer	37540.53
Arthritis	33899.27
NoDisease	11557.29

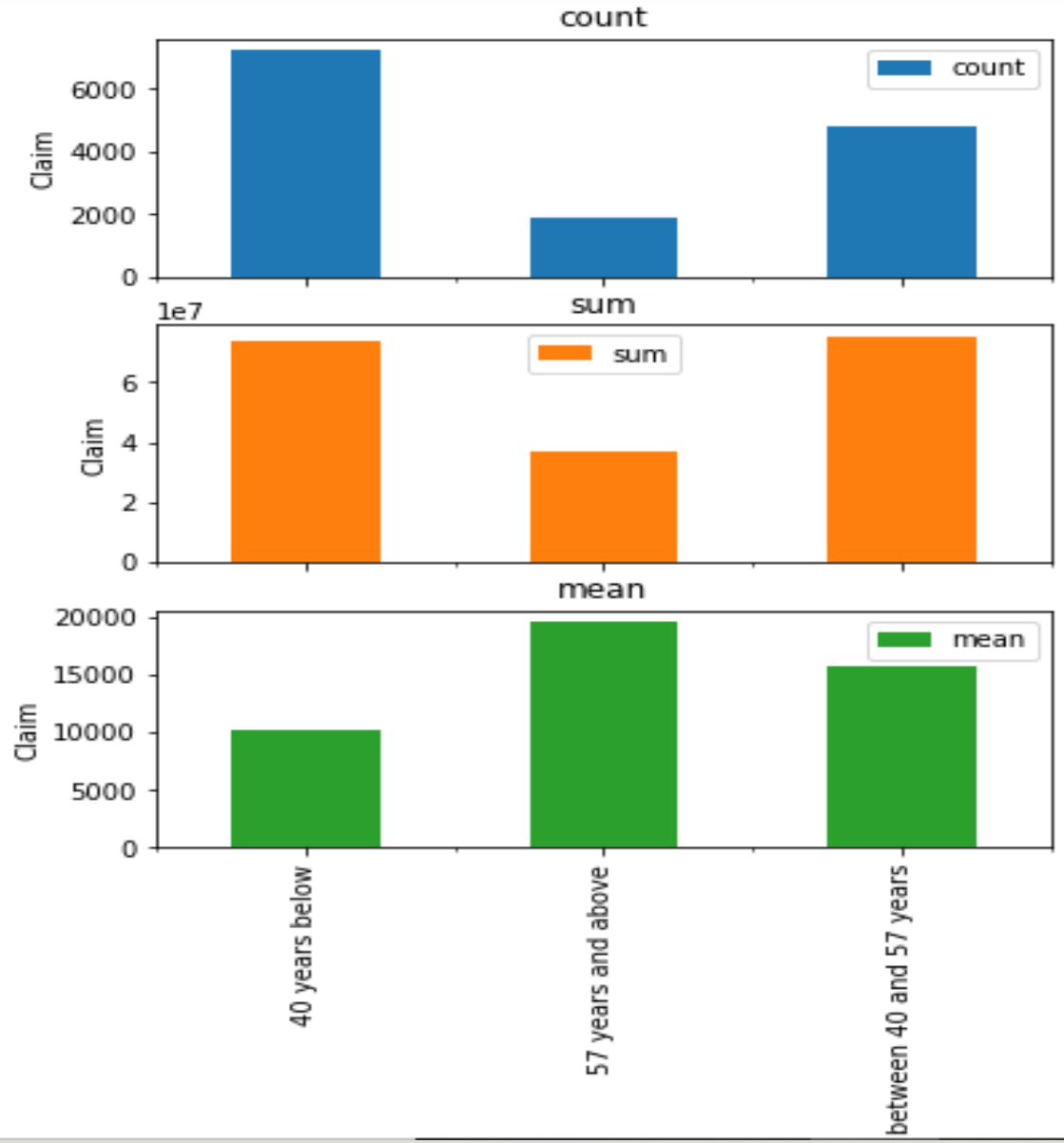




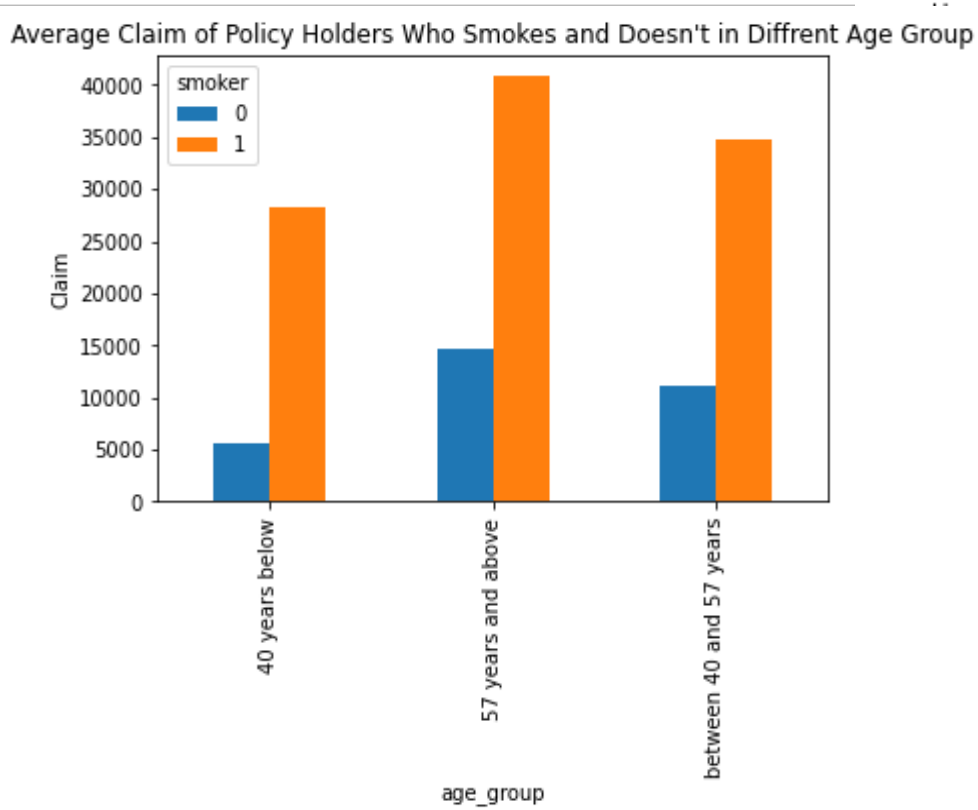
age_group	count	sum	mean
40 years below	7229	74104209.7	10250.96
57 years and above	1893	37127111.4	19612.84
between 40 and 57 years	4782	75524617.2	15793.52

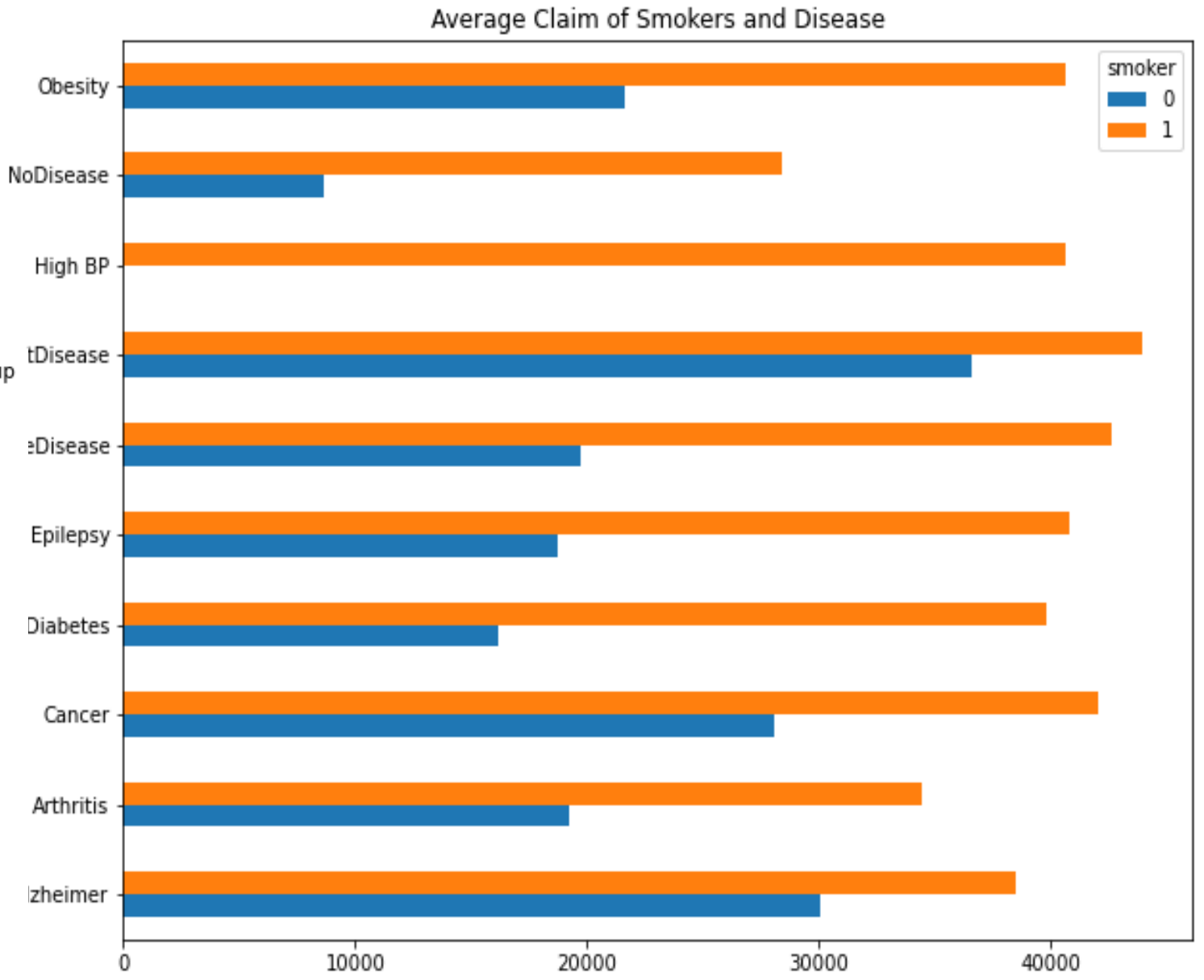
OLS Regression Results

Dep. Variable:	claim	R-squared:	0.084
Model:	OLS	Adj. R-squared:	0.084
Method:	Least Squares	F-statistic:	424.7
Date:	Wed, 05 Oct 2022	Prob (F-statistic):	5.02e-264
Time:	09:10:15	Log-Likelihood:	-1.4988e+05
No. Observations:	13904	AIC:	2.998e+05
Df Residuals:	13900	BIC:	2.998e+05
Df Model:	3		
Covariance Type:	nonrobust		



diseases	age_group	smoker	claim
Arthritis	57 years and above	1	48665.10
Cancer	57 years and above	1	48375.74
Epilepsy	57 years and above	1	48030.80
Obesity	57 years and above	1	47394.35
Diabetes	57 years and above	1	46760.99





ANALYSIS AND INSIGHT FROM DATASET

Why does policy holders with no disease have the lowest average claim and the highest total claim?

Policy holders with no disease have the highest number whose claims are being paid to, this tends to increase the total claim amount, but individual claims are not much compared to the individual claim paid by those with diseases. The average claim of policy holders with no disease is the lowest because the claim paid individually is not much. The total claim will increase because the claim is paid to many people and the average claim reduces because the amount paid individually is not material, meaning its not much. The analysis shown in slide number 3 describes the average claim of diseases of policy holders. It shows that policy holders with heart disease have the highest average claim because they are prone to having more health issues followed by eye disease policy holders, while policy holders with no disease have the lowest average claim because they have the least chance of having health issues.

Does the claim of the policy holder defer by age?

It has been statistically proven with a probability value of 5.02e-264 which is less than the significant level of 0.05 this proves that there is a significant difference in the claim of policy holders among different age group. In the table and graph above, it is clear that the age group of 57 years and above has the highest claim followed by the years between 40 and 57. policy holders above 57 years get higher claims as a result of old age.

Why does policy holders above the age of 57 years have the highest average claim and the lowest total claim?

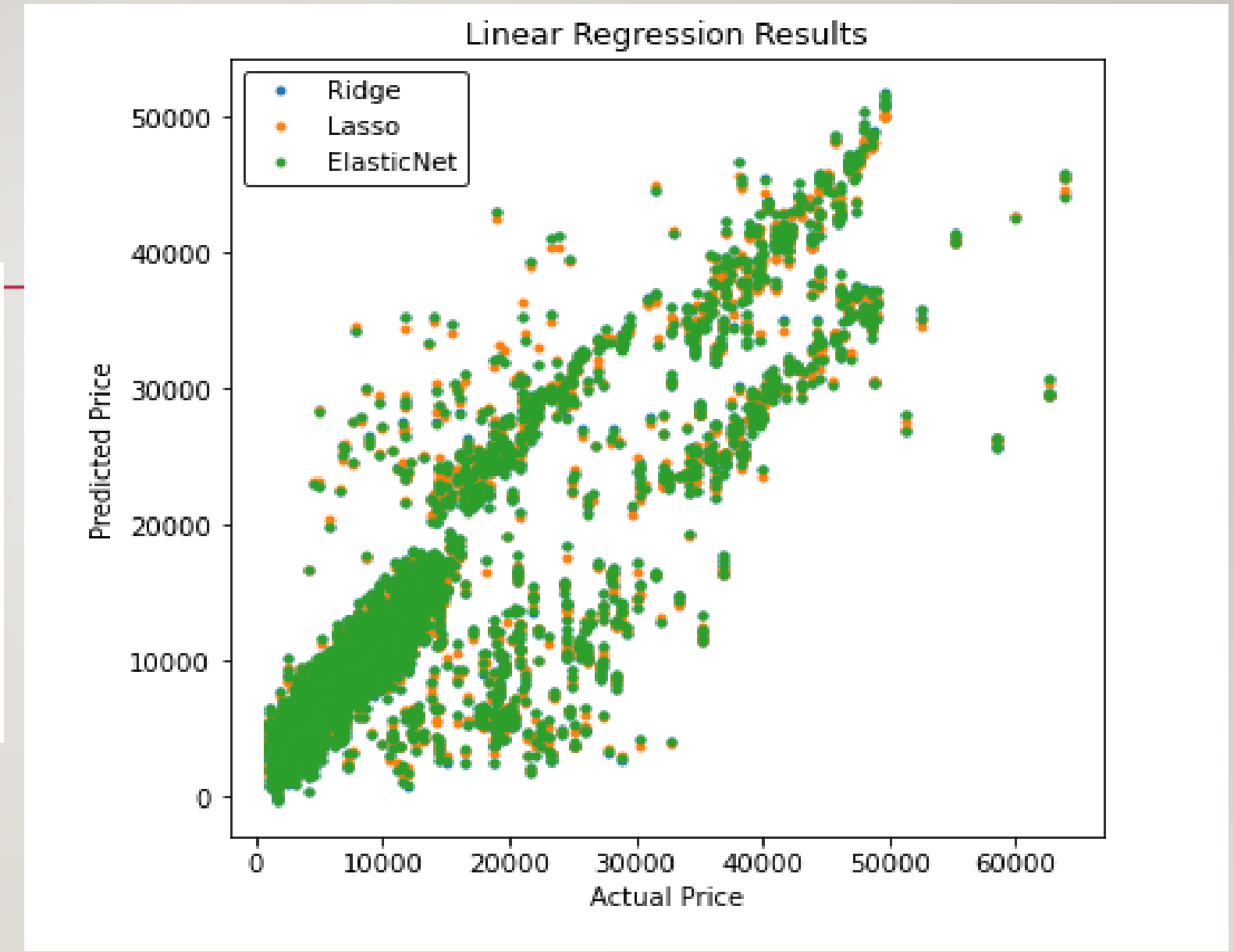
From the data and visualization in slide number 9, its is clear that policy holders above the age of 57 years have the highest claim and this is as a result of the materiality of their claim, meaning that the claim paid are of huge amount. Their claim is material because at this age, it is expected of them to have a weak immune system which makes them prone to illness, and according to the analysis, policy holders at this age have the highest rate of smokers and it is proven statistically that smokers have the highest positive impact on claims. The total claim of policy holders at this age is verry low and this is because this age group have the lowest number of policy holders who claims are paid to.

Why does smokers have a very large impact on the increase in claim?

Naturally, smoking is expected to have a negative effect on the health of a policy holder, according to my analysis, majority of those smoking are the age group above 57 years and they happen to suffer from diseases. From the analysis above, smokers are 78.6% more than does who doesn't, this implies that policy holders who smokes tends to have higher claims compared to those how doesn't. This implies that policy holders who smokes are prone to have health issues more compared to those who doesn't, this will result to an increase in the premium they pay which will directly increase their claims.

MODEL PREDICTION

	R2
Linear	0.765026
Ridge	0.764957
Lasso	0.765902
ElasticNet	0.764924

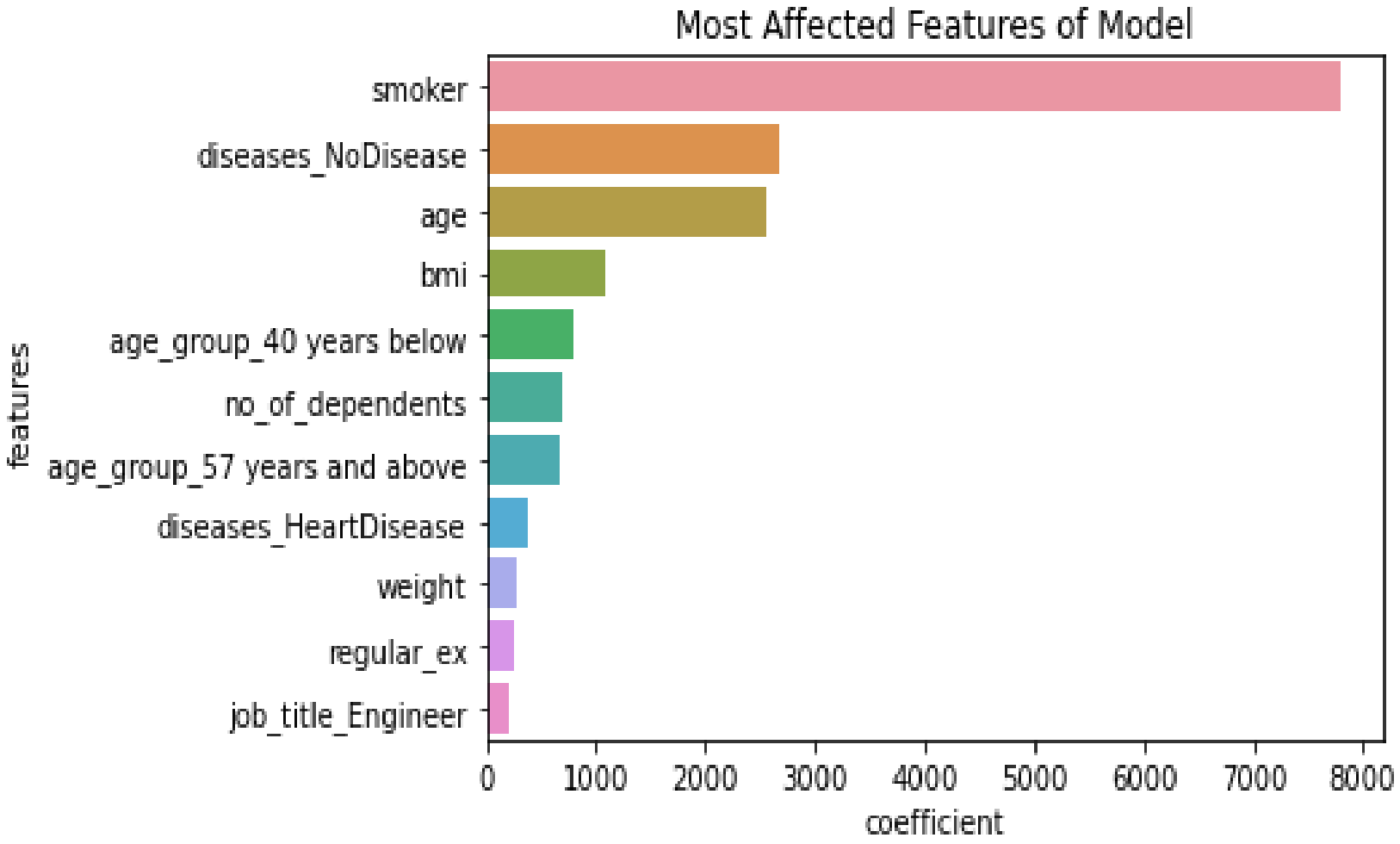


MODEL DESCRIPTION

According to the table and diagram above, there are four different models describing the dataset, each of the models are similar and all working perfectly fine but out of the four, the model built with lasso regression has the highest score for efficiency. The model is built on a linear regression and the use of lasso regression to reduce the model error to its minimum.

MODEL INTERPRETATION

	features	coefficient
4	smoker	7803.772915
17	diseases_NoDisease	2689.595415
0	age	2558.963518
2	bmi	1097.610902
145	age_group_40 years below	782.483952
3	no_of_dependents	684.281324
146	age_group_57 years and above	659.675323
15	diseases_HeartDisease	377.564026
1	weight	269.444701
6	regular_ex	257.845295
126	job_title_Engineer	200.851801



INTERPRETATION PREDICTION

The table and graph in slide number 15 are the summary of my model interpretation. It explains the features that has great impact on the model. From the graph above, it shows how each features affects the model accordingly. Smokers have a very high effect on the model, followed by policy holders with no diseases, age, bmi etc. According to the model, its is clear that smokers has a very high impact in predicting the claims of policy holders. The model would have performed better if there were more data to aid its prediction.

CONCLUSION

Analysis and visualizations drawn from this project work is executed in jupyter notebook with python language and I answered few questions using the root cause analysis approach specifically for understanding the factors that rise claim.

The model is based on supervised machine learning (linear regression).