EAST WEST UNIVERSITY

PROJECT REPORT

Course Name: Statistics for Data Science

Course Code: CSE303

Section: 03

Project Name : **Insurance Charges Analysis**

Prepared By

Name	**ID**
Ajmaeen Abid	2023-1-60-175
Aditya Debnath	2022-260-124
Umme Habiba	2022-1-60-362

Submitted to

Dr. Mohammad Manzurul Islam

Assistant Professor

CSE Dept.

Submission Date

1st September,2025

Insurance Charges Analysis Report

Generated analysis and model evaluation for two datasets: Dataset 1 (insurance.csv) and Dataset 2 (insurance_uncleaned_realistic.csv).

Introduction

This report analyzes two insurance datasets to explore data characteristics, perform preprocessing, visualize relationships between features, and build machine learning regression models to predict insurance charges.

Dataset Description

Dataset 1: filename `insurance.csv` — rows: 1338, columns: 7

Dataset 2: filename `insurance_uncleaned_realistic.csv` — rows: 1338, columns: 10

Key features

Dataset 1	Dataset 2
age — numeric (years)	**age** — numeric (years)
sex — categorical (male/female)	**sex** — categorical (male/female)
bmi — numeric (body mass index)	**bmi** — numeric (body mass index)
children — numeric (number of children)	**children** — numeric (number of children)
smoker — categorical (yes/no)	**smoker** — categorical (yes/no)
region — categorical (northeast/northwest/southeast/southwest)	**region** — categorical (northeast/northwest/southeast/southwest)
charges — numeric (insurance charges) — target variable	**blood_pressure** -- numeric
	exercise_level – categorical (low/medium/ high)
	medical_history – categorical (diabetes/ hypertension/none)
	charges — numeric (insurance charges) — target variable

Data Preprocessing

Below are the preprocessing steps applied to both datasets:

- Standardized column names.

- Checked and handled missing and null values: numeric columns were imputed with mode, mean etc. categorical columns were mapped to convert them into numerical values and use them in regression model

- Removed duplicate rows (if present).

- Converted categorical columns to numeric encodings: smoker -> smoker_num (yes=1, no=0); sex -> sex_num (male=0, female=1); region -> region_num (ordinal mapping).

- Ensured charges column is numeric and filled any conversion-caused missing with median.

Preprocessing summary - Dataset 1

Initial rows: 1338 | Duplicates removed: 1

Missing values per column (after imputation):

- age: 0 (original count)

- sex: 0 (original count)

- bmi: 0 (original count)

- children: 0 (original count)

- smoker: 0 (original count)

- region: 0 (original count)

- charges: 0 (original count)

Preprocessing summary - Dataset 2

Initial rows: 1338 | Duplicates removed: 0

Missing values per column (after imputation):

- age: 0 (original count)

- sex: 0 (original count)

- bmi: 50 (original count)

- children: 0 (original count)

- smoker: 30 (original count)

- region: 0 (original count)

- blood_pressure: 353 (original count)

- exercise_level: 540 (original count)

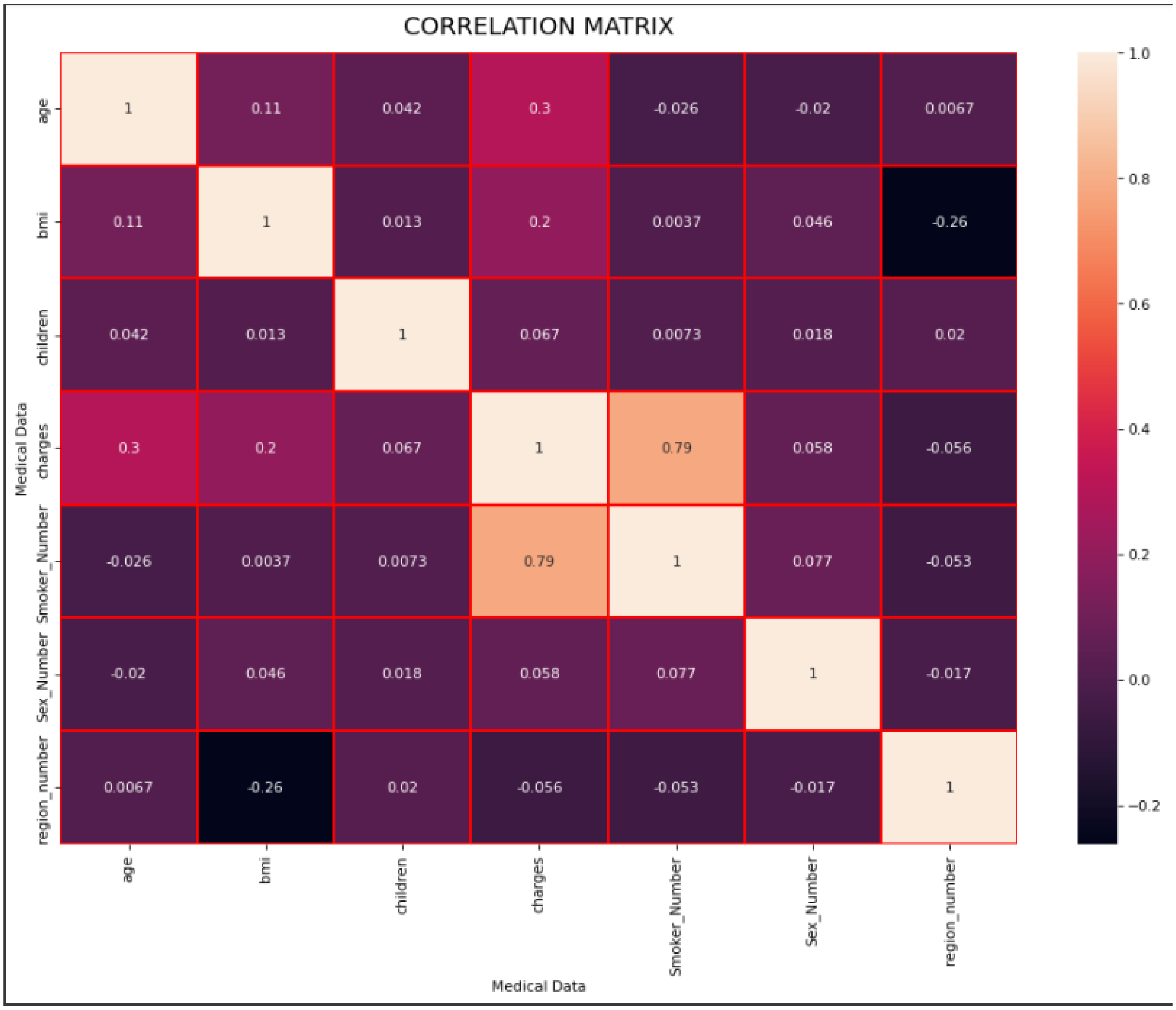
- medical_history: 259 (original count)

- charges: 0 (original count)

Exploratory Data Analysis (EDA)

Correlation matrices and visualizations were generated. Attachments below show key plots.

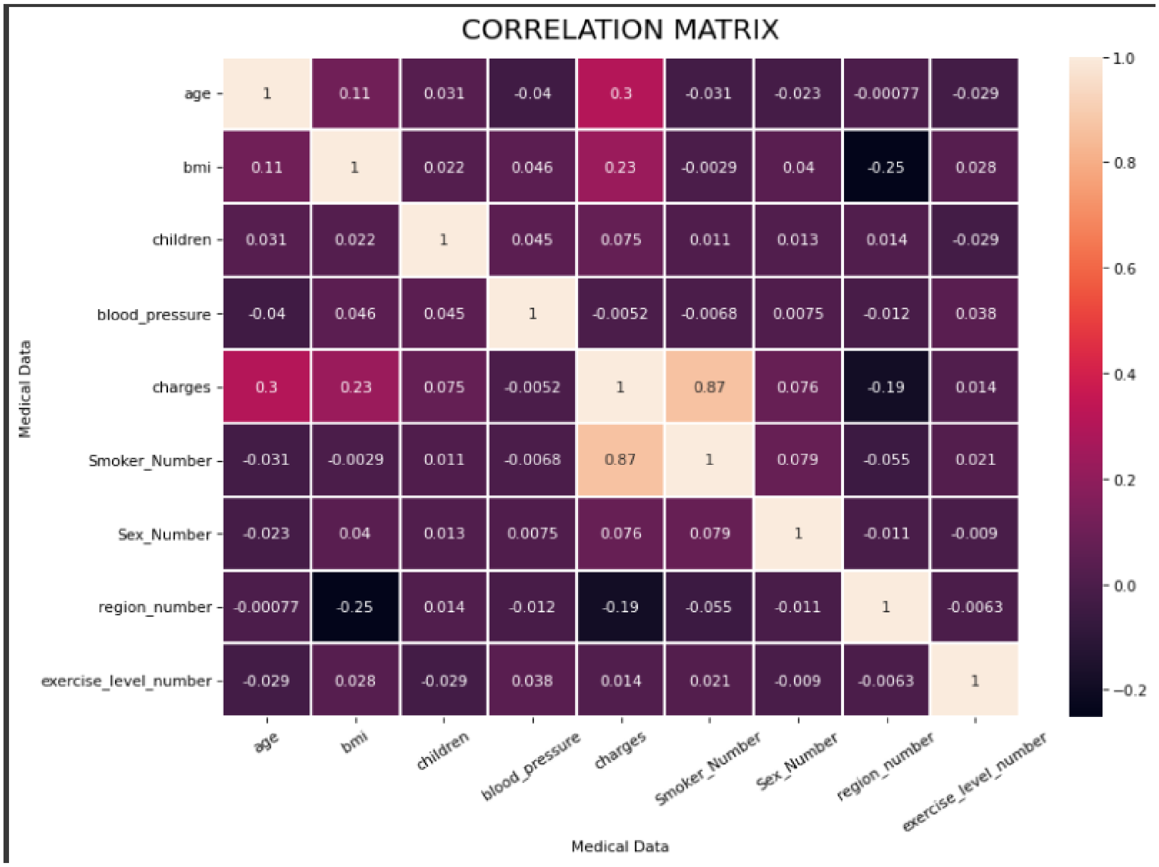
Correlation matrix - Dataset 1



Interpretation

This correlation matrix displays the relationships between different medical data features. The most significant finding is the strong positive correlation (0.79) between Smoker_Number and charges, indicating that smokers tend to have much higher medical costs. Age and BMI also show a positive relationship with charges, although it is not as strong as smoking. Interestingly, there's a negative correlation (-0.26) between bmi and region_number, suggesting some regional differences in BMI. Most other features, like children and sex, have very weak correlations with the final charges.

Correlation matrix - Dataset 2

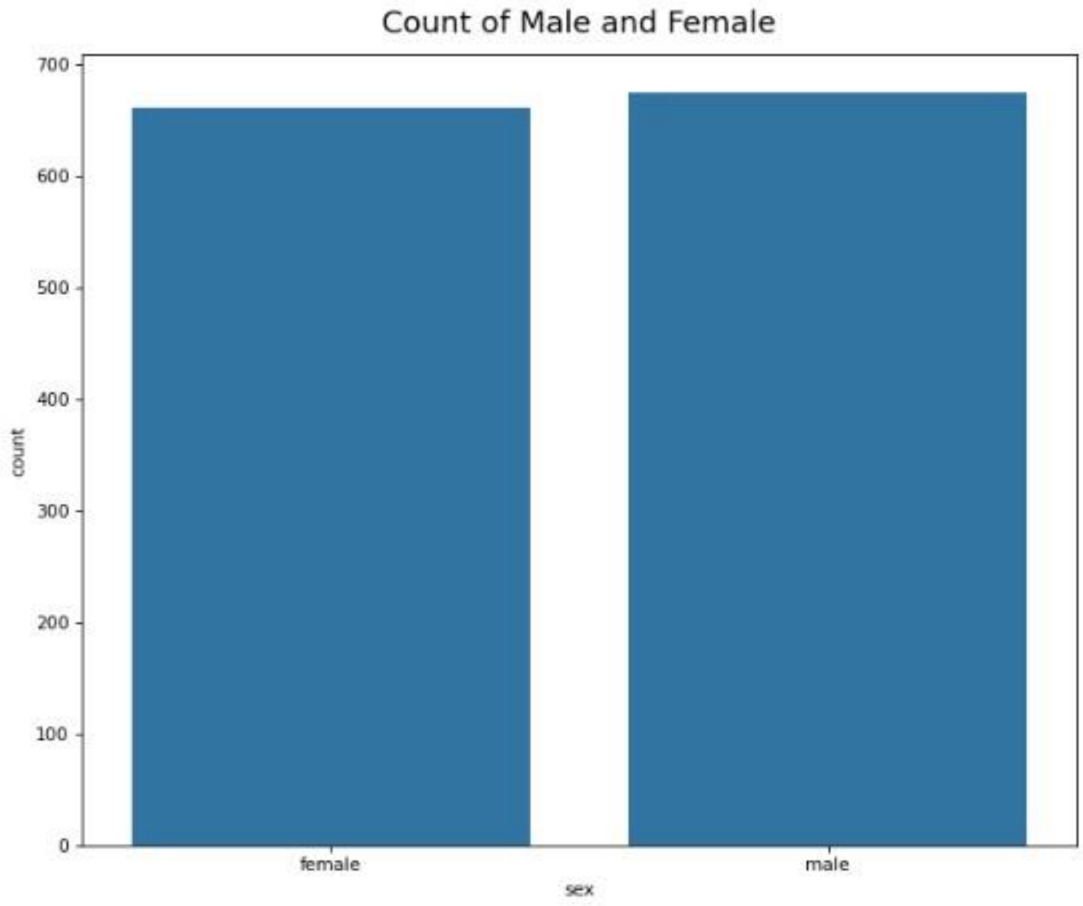


	age	bmi	children	blood_pressure	charges	Smoker_Number	Sex_Number	region_number	exercise_level_number
age	1	0.11	0.031	-0.04	0.3	-0.031	-0.023	-0.00077	-0.029
bmi	0.11	1	0.022	0.046	0.23	-0.0029	0.04	-0.25	0.028
children	0.031	0.022	1	0.045	0.075	0.011	0.013	0.014	-0.029
blood_pressure	-0.04	0.046	0.045	1	-0.0052	-0.0068	0.0075	-0.012	0.038
charges	0.3	0.23	0.075	-0.0052	1	0.87	0.076	-0.19	0.014
Smoker_Number	-0.031	-0.0029	0.011	-0.0068	0.87	1	0.079	-0.055	0.021
Sex_Number	-0.023	0.04	0.013	0.0075	0.076	0.079	1	-0.011	-0.009
region_number	-0.00077	-0.25	0.014	-0.012	-0.19	-0.055	-0.011	1	-0.0063
exercise_level_number	-0.029	0.028	-0.029	0.038	0.014	0.021	-0.009	-0.0063	1

Interpretation

In this matrix, we see that the number of smoker people have strong correlation with charges. Which means the smoking greatly affect the charges. The age and bmi are correlated with charges but they are weak. also region are correlated with bmi and charges negatively.

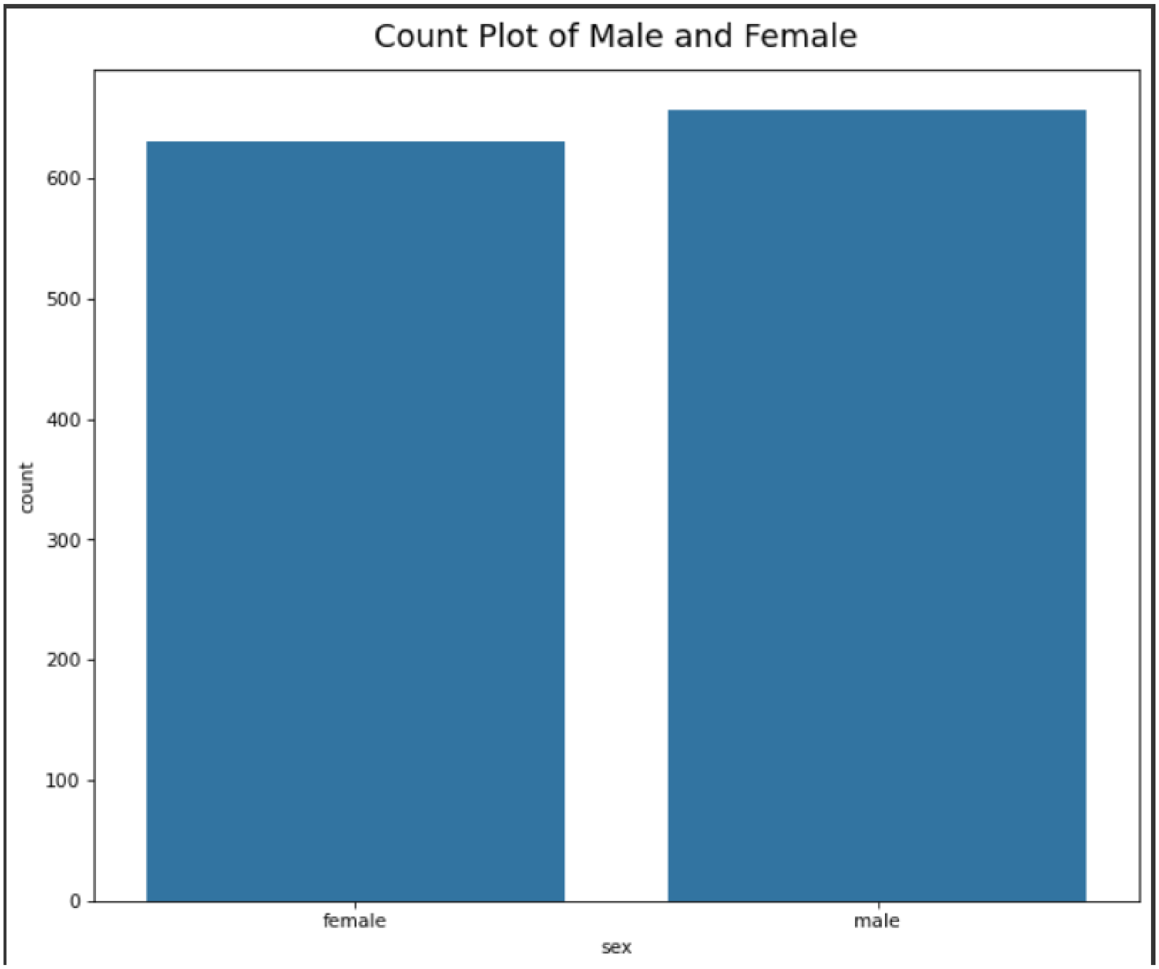
Count Plot of Male vs. Female Dataset 1



Interpretation

This count plot shows a nearly equal distribution of males and females in the dataset.

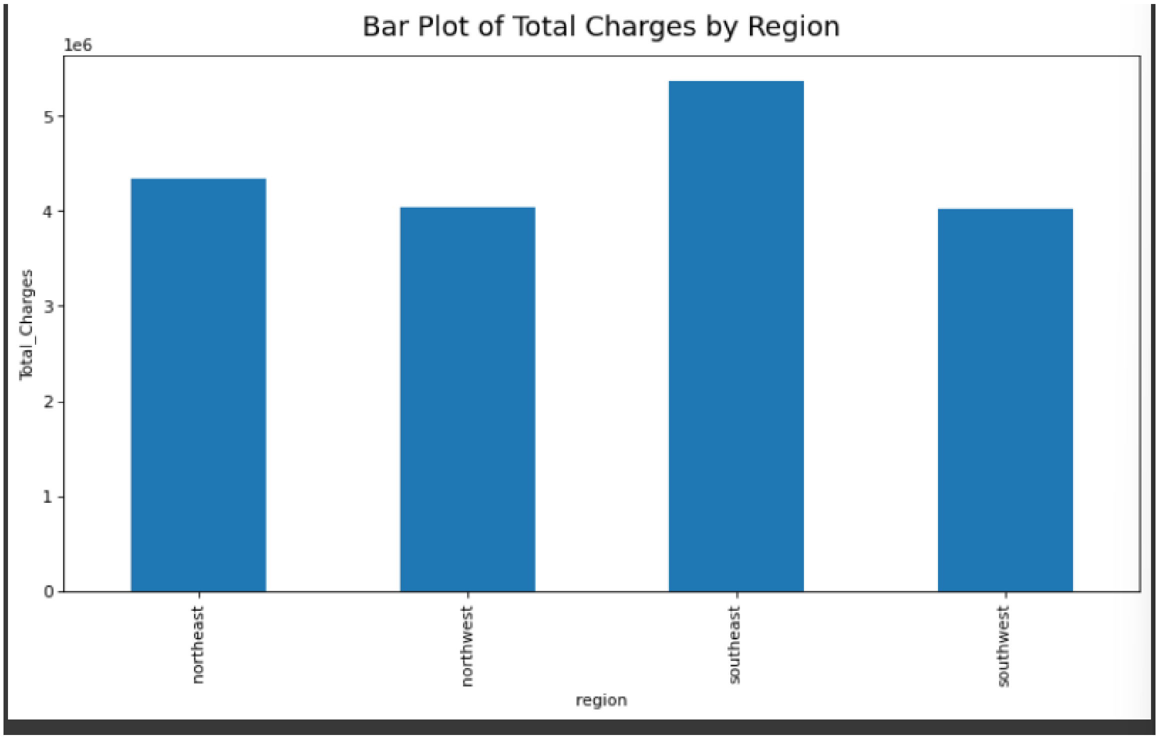
Count Plot of Male vs. Female Dataset 2



Interpretation

In this count plot we can say that male are in great numbers than female. It seems male are around 650-700. on the other hand female are lower than 650. But this difference are not much as for numbers

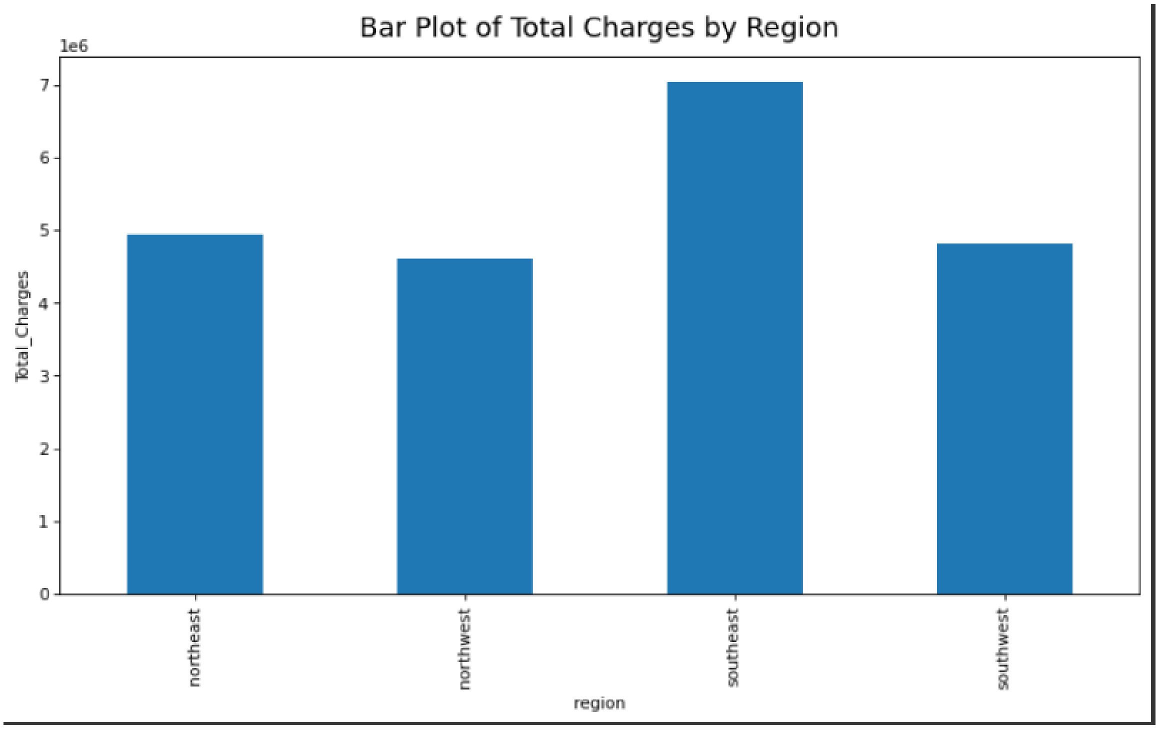
Bar Plot of Total charges by Region- Dataset 1



Interpretation

This bar plot displays the aggregated medical charges across different regions. It highlights that the southeast region incurs the highest total medical costs, exceeding 5 million. The northwest and southwest regions show similar, lower total charges, both around 4 million, suggesting regional disparities in healthcare spending or prevalence of health issues.

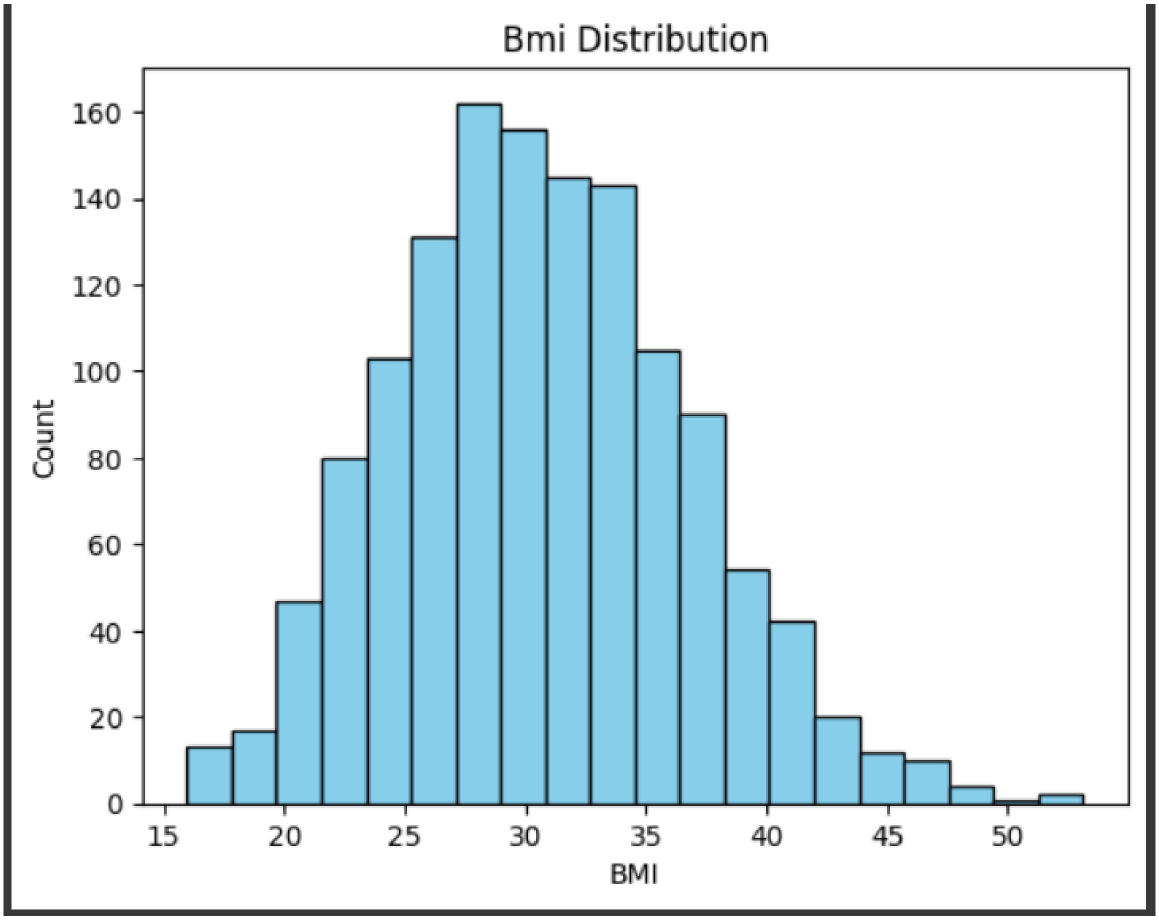
Bar Plot of Total charges by Region- Dataset 2



Interpretation

The bar plot shows, the total charges according to regions. Here the southeast region have highest charges than other regions, which indicates the people in this region spent great amount of money for treatment also they may ignore health rules or they may have higher medical cost. On the other hand, northwest have lowest patient of health issues. And as for northeast and southwest have almost similar amount of charges.

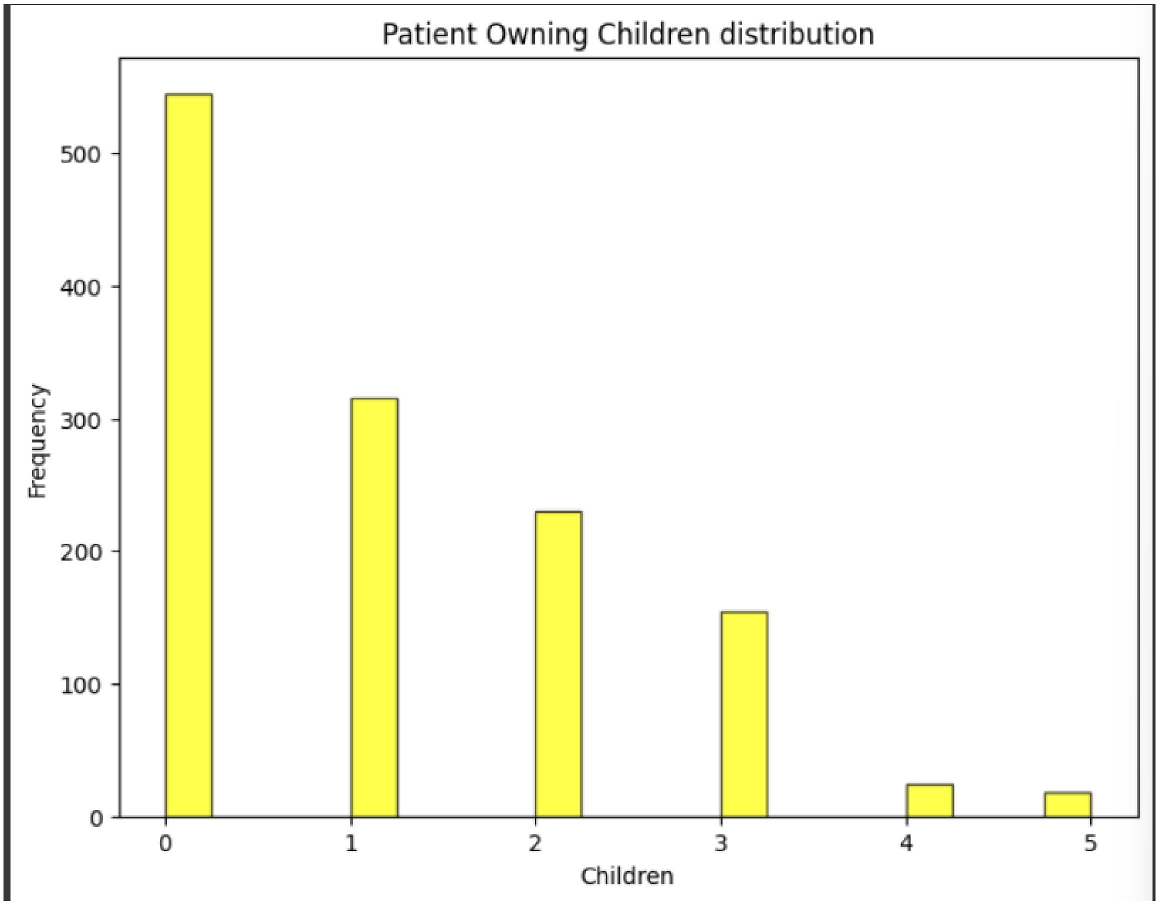
Histogram of bmi distribution - Dataset 1



Interpretation

This histogram shows a roughly normal distribution for BMI, with the most frequent values concentrated around 30.

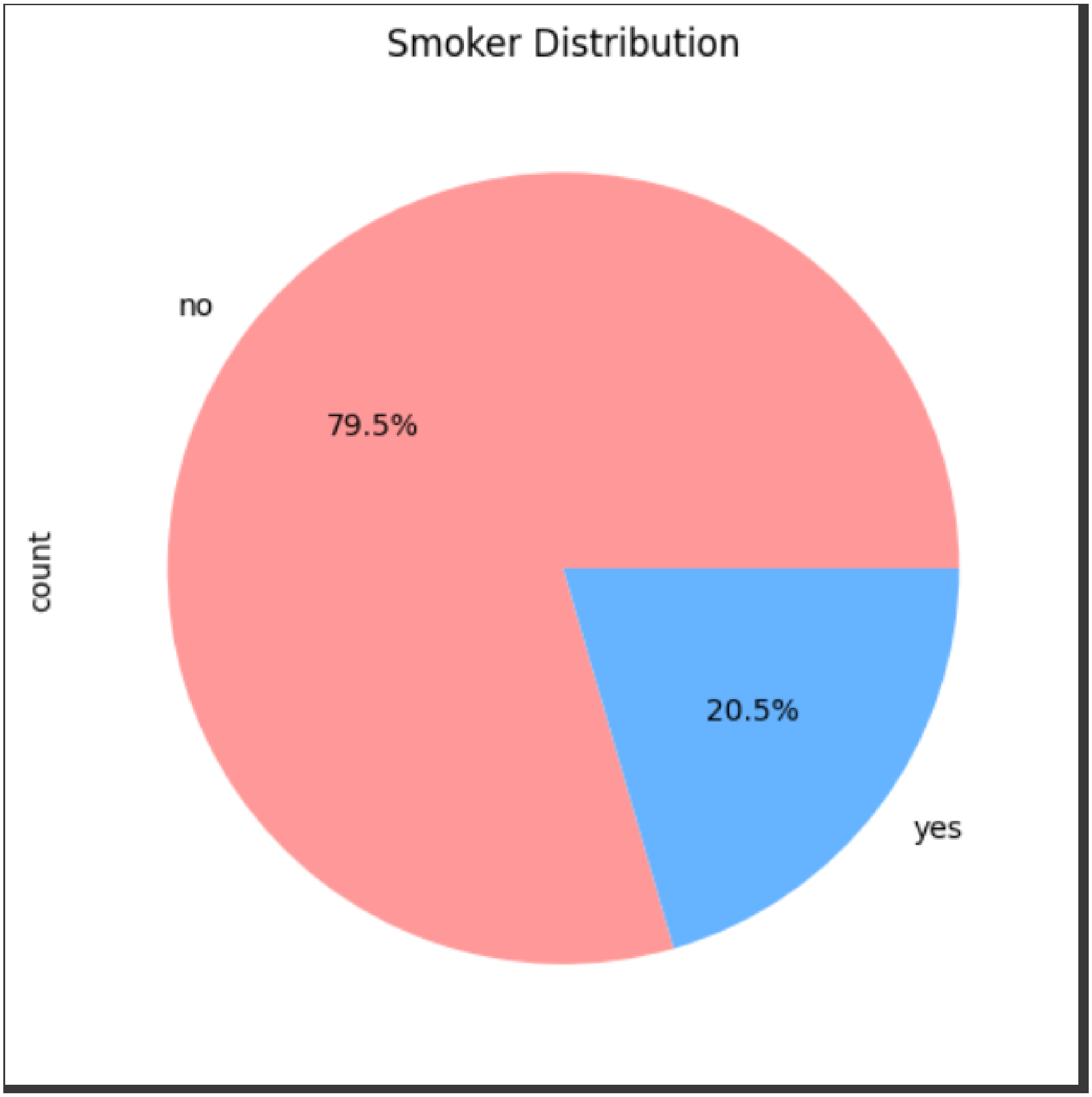
Histogram: Children distribution of the Patients - Dataset 2



Interpretation

The histogram shows the patient who don’t have no children are highest in number and it gradually decreases until 5. as for who have children 4 or 5 are very little numbers

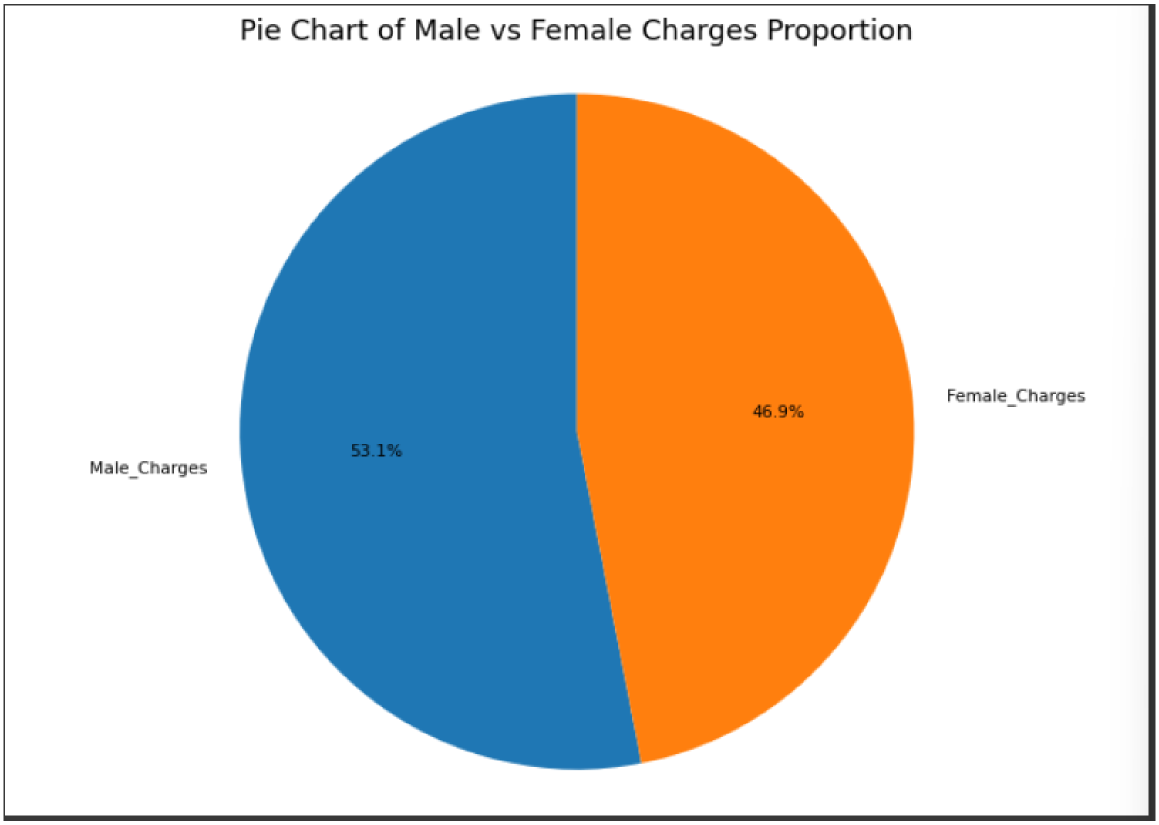
Pie Chart of distribution of smoker’s vs non-smokers - Dataset 1



Interpretation

The pie chart clearly shows that non-smokers are the dominant group, representing nearly 80% of the dataset.

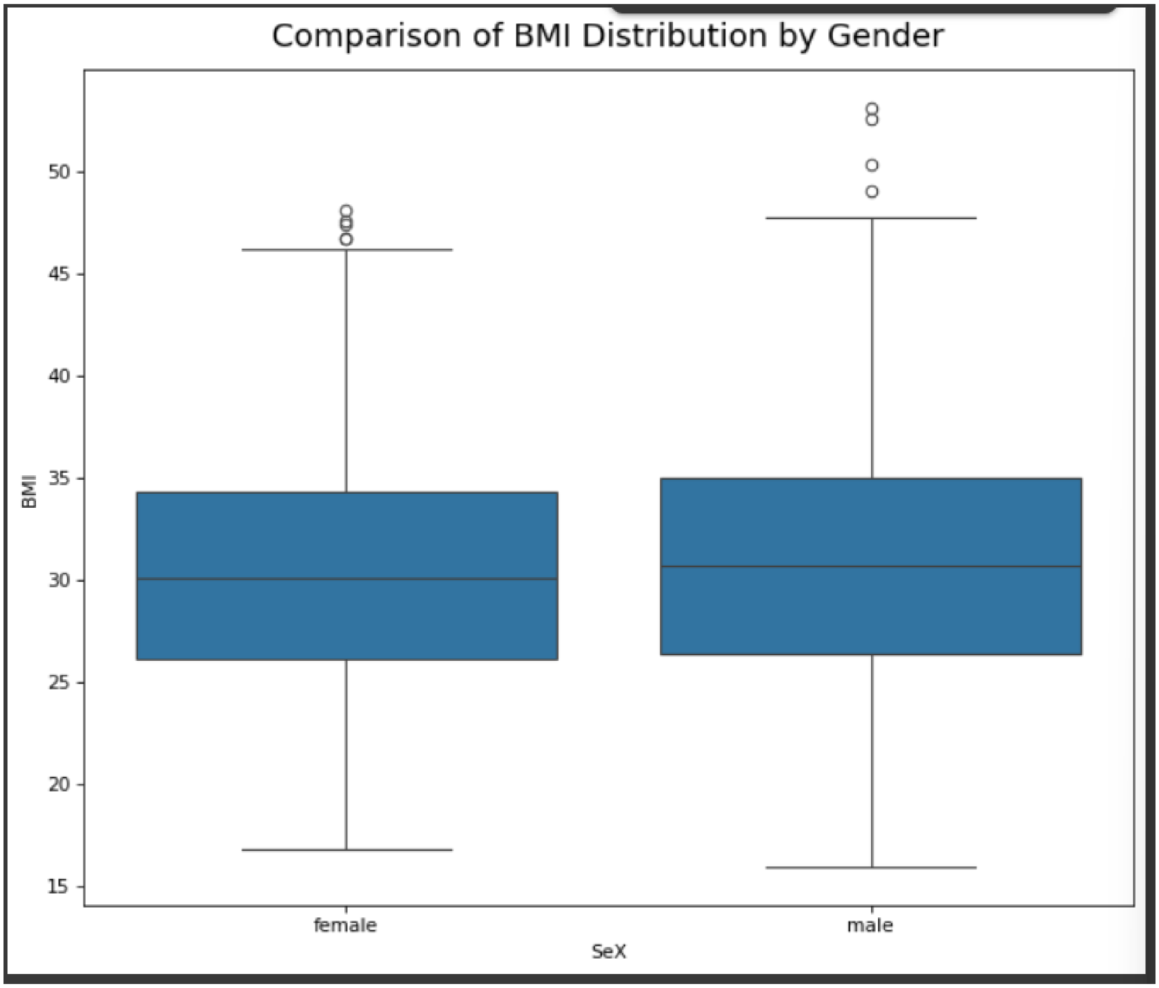
pie Chart of Male vs. Female Charges Proportion- Dataset 2



Interpretation

The pie chat shows, male spend more for medical cost than female. The total charge for this dataset the male charges overcome more than 50% and female less than 50%.

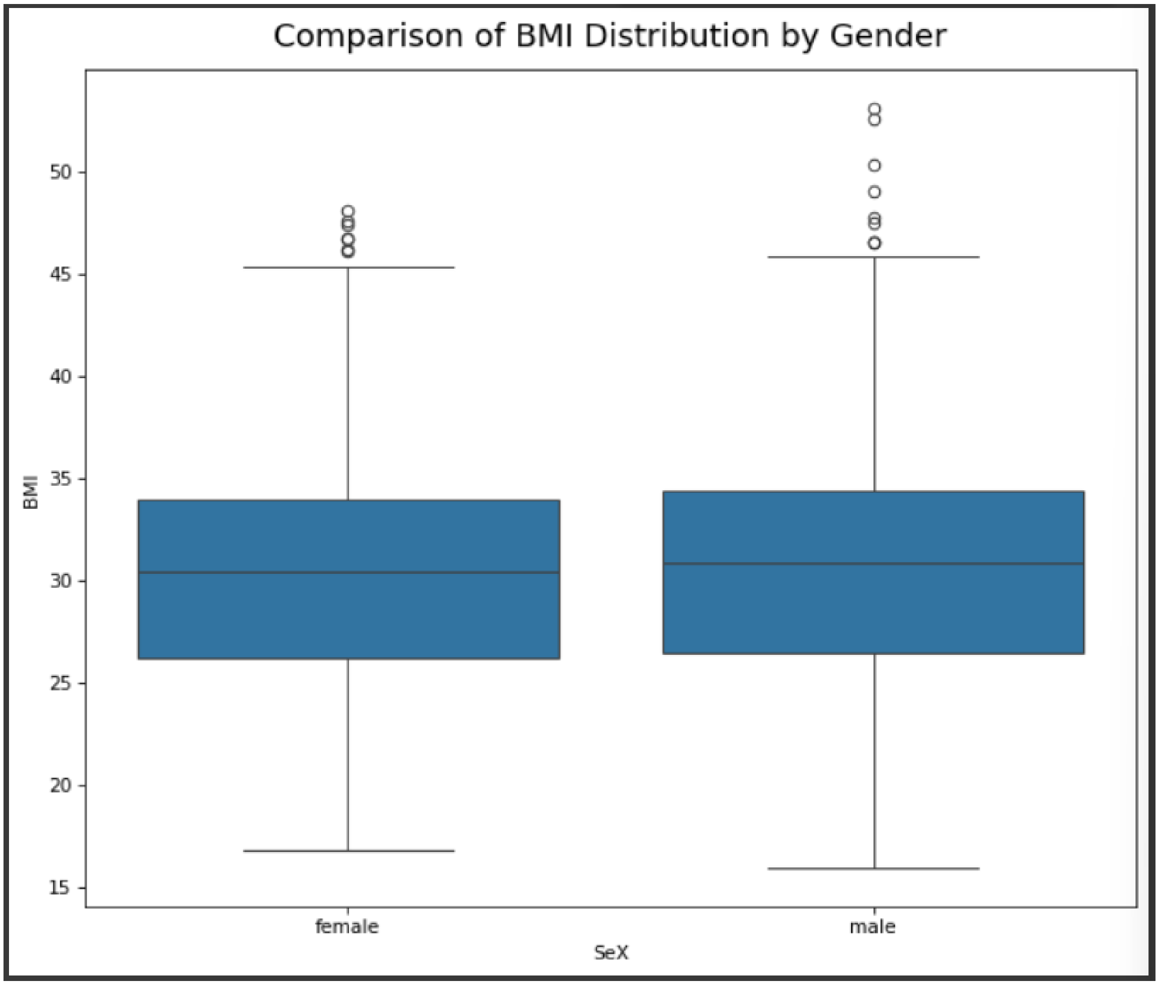
Age comparison of BMI- Dataset 1



Interpretation

This boxplot compares the BMI distributions for males and females, showing very similar median values and interquartile ranges, which suggests no significant difference in typical BMI between the two genders. However, the outliers for males extend to slightly higher BMI values compared to females.

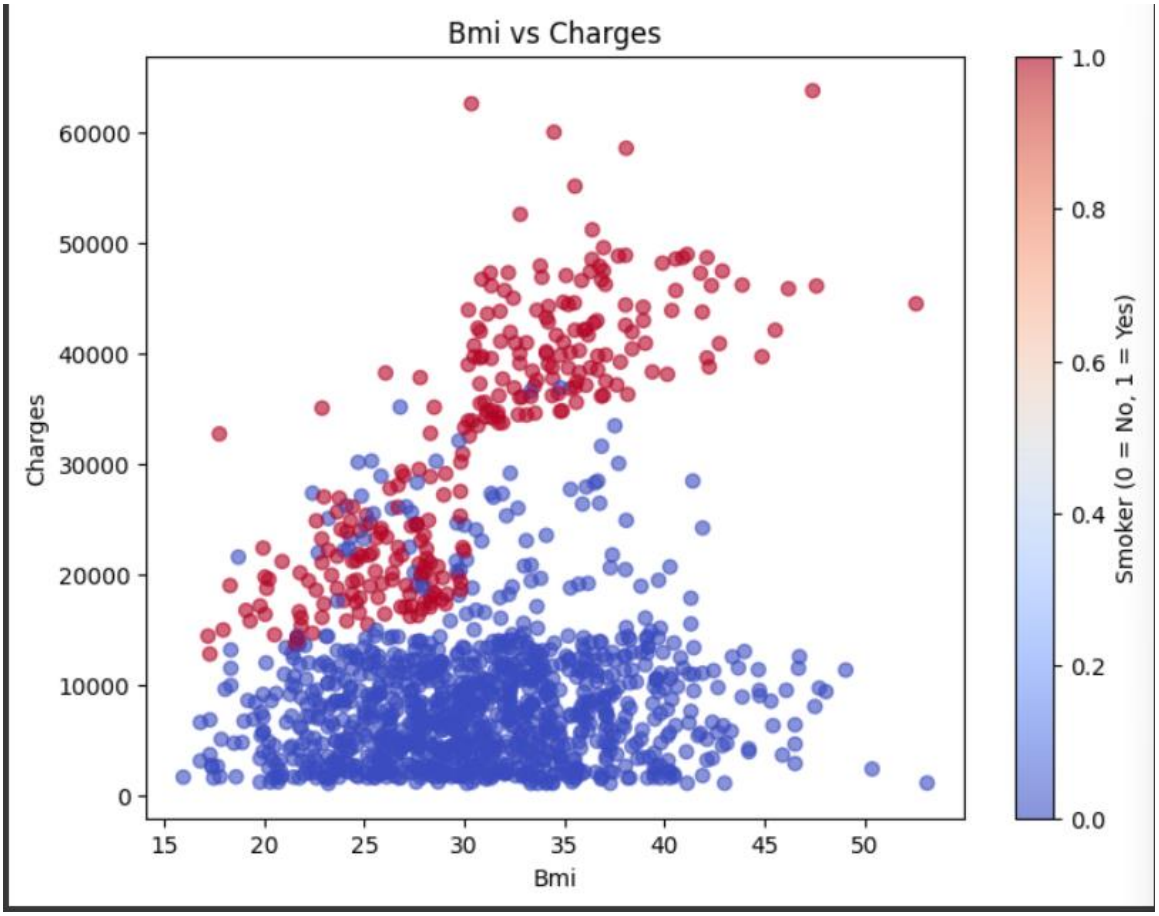
Age comparison of BMI- Dataset 2



Interpretation

This shows the bmi value for both male and female start after 26 and it continues until 34 for these people. the bmi outliers for male are greater in numbers and higher than female. The bmi outliers for male exceed 50 and as for female it remains under 50

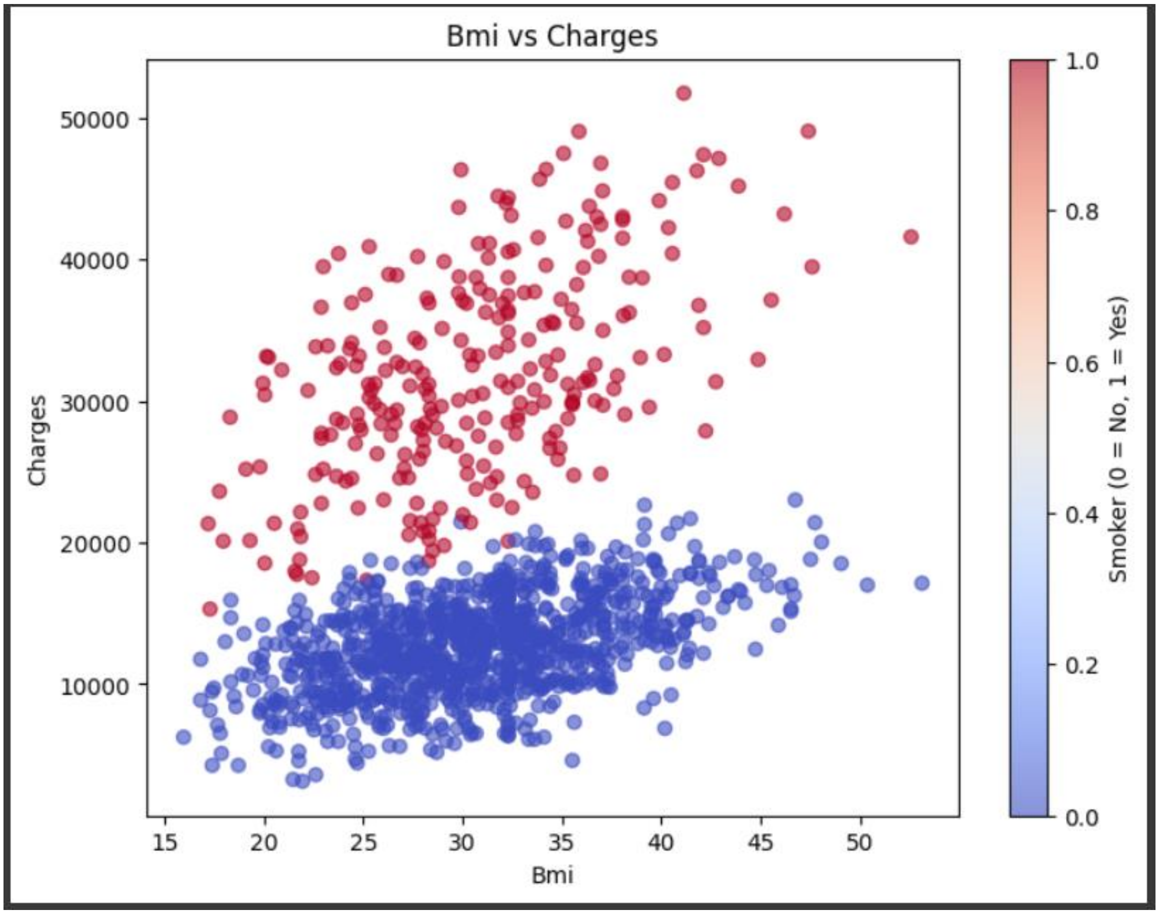
Scatter plot Bmi vs Charges- Dataset 1



Interpretation

This scatter plot reveals a clear distinction in medical charges based on smoking status. Non-smokers (blue dots) generally have lower costs, forming a dense cluster at the bottom of the graph. In contrast, smokers (red dots) face significantly higher charges, with a noticeable positive trend where costs increase with BMI.

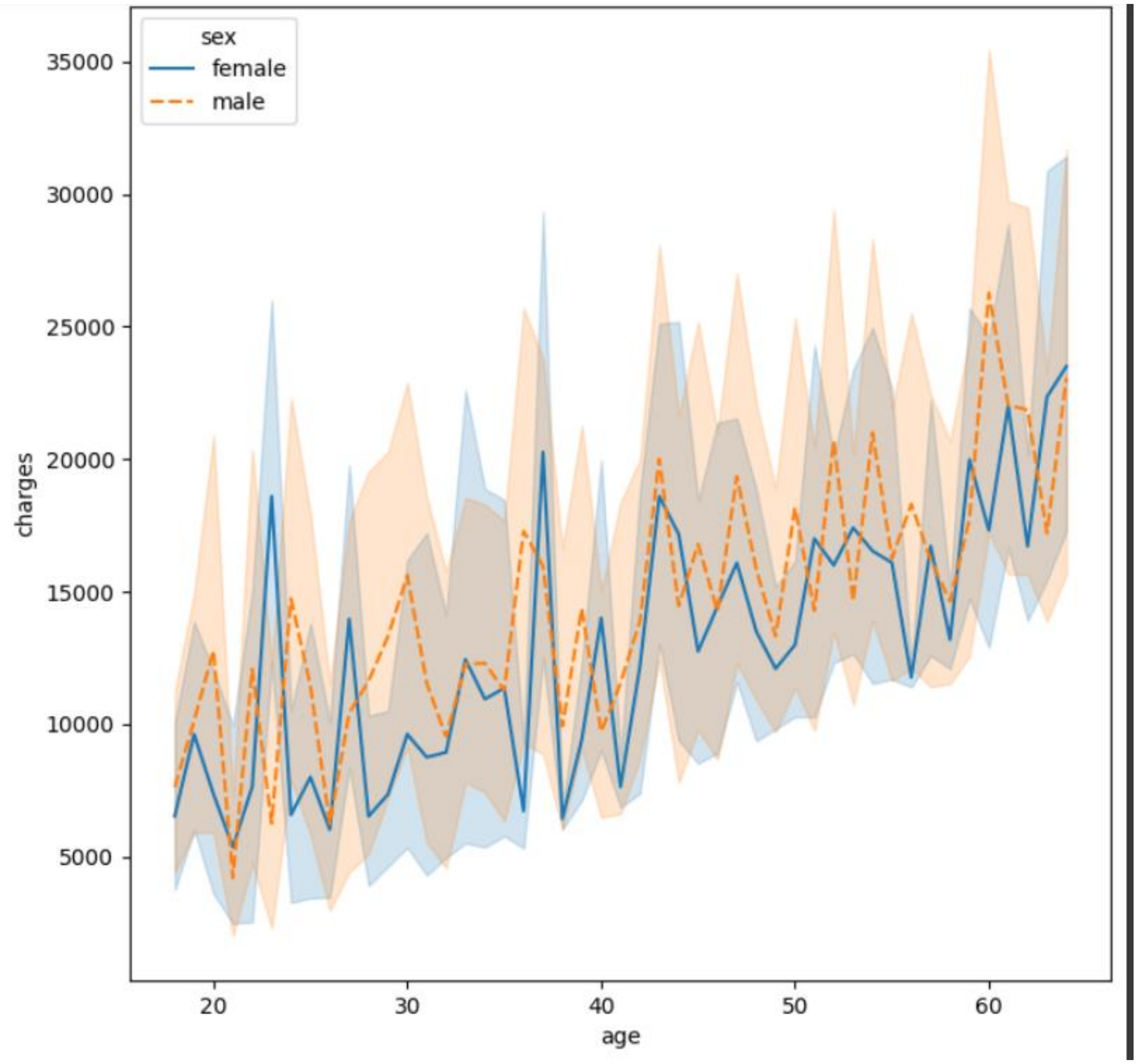
Scatter plot Bmi vs Charges- Dataset 2



Interpretation

The scatter plot shows, the great relation of bmi and charges alongside with smoking. The plot shows the smoker have higher medical charges. As bmi increases the charges for smoker increase also gradually. The charges for non-smoker are consized. On the other hand, the smoker charges are wide spreaded.

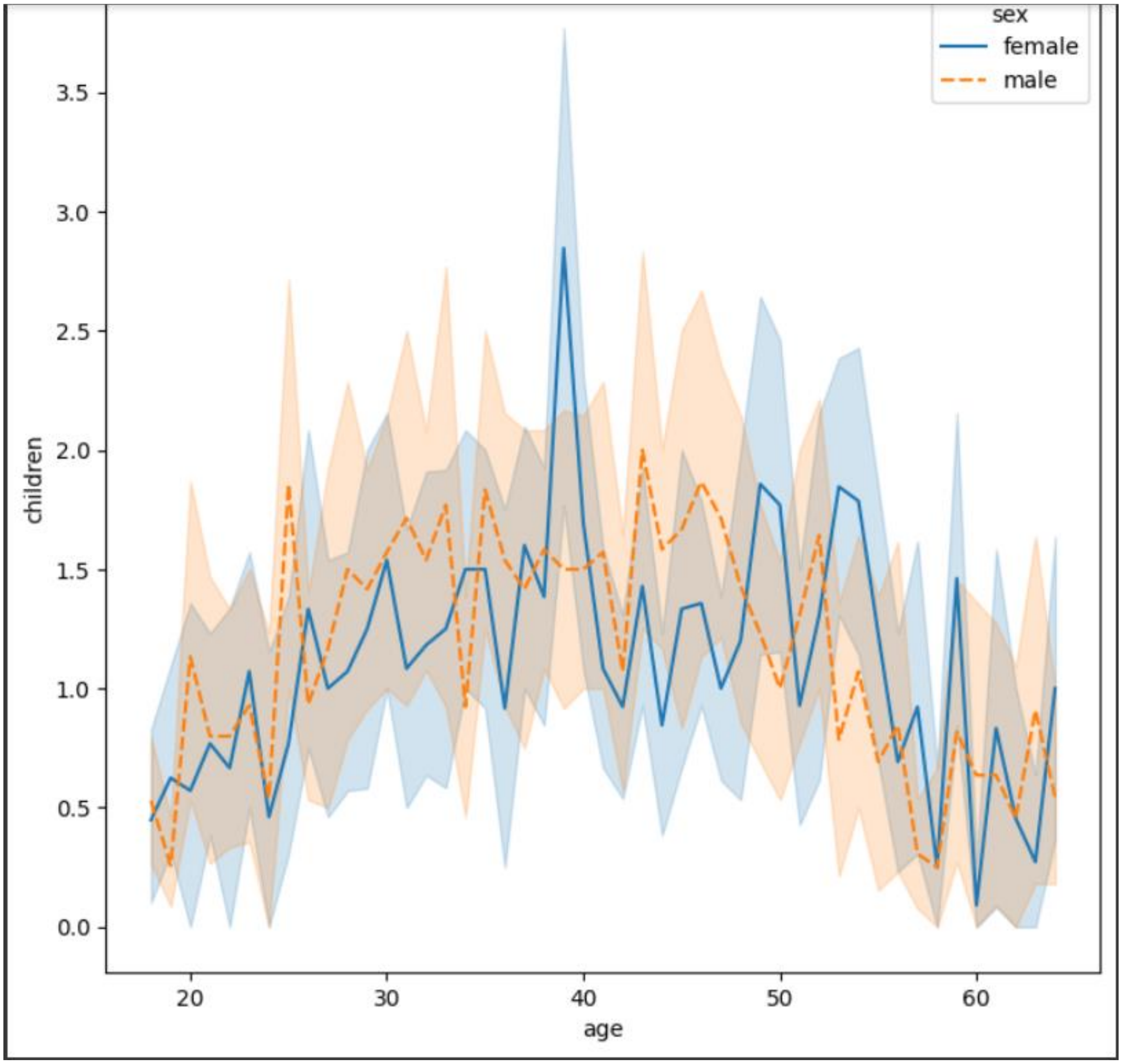
line plot of AGE vs charges according to gender- Dataset 1



Interpretation

This line plot illustrates a clear positive trend where medical charges generally increase with age for both males and females. Despite the overall upward slope, the significant fluctuations and overlapping shaded areas suggest high variability and no consistent difference in charges between the sexes.

line plot of AGE vs children according to gender- Dataset 2



Interpretation

The line plot shows the children distribution of the patients according with age. It shows lower age patient have small number of children but around 25 years age the children of male rise suddenly and decreases then after. Around 40 age the number of children for female increases highly which is the highest number for male also. after 40 years female having children moves upside down but for male the numbers usually decreases. Again the numbers come back at same at the last stage of the age years. Also we can say that the data for this plot are normally skewed.

Machine Learning Models

We chose the regression model cause our datasets contains continuous values of medical cost. As we will build model for predicting medical cost , regression model is the appropriate choice for predicting and work with continuous values.

Two regression models were trained to predict `charges` for each dataset: Linear Regression and Decision Tree. Models were evaluated using MAE, MSE, RMSE, and R² on a 80/20 train/test split.

Dataset 1 - Model Performance

Metric	Linear Regression	Decision Tree
Mean Target (Medical Cost)	14272.01	14272.0075
MAE	4181.35	2591.6682
MAE as % of Mean Target	29.29%	18.16%
MSE	35604894.07	18722739.9352
RMSE	5966.99	4326.9782
RMSE as % of Mean Target	41.81%	30.3179%
R^2 Score	0.8062	0.8981

Dataset 2 - Model Performance

Metric	Linear Regression	Decision Tree
Mean Target (Medical Cost)	15836.71	15836.71
MAE	2143.18	2066.35
MAE as % of Mean Target	13.53%	13.05%
MSE	7157336.70	7066064.17
RMSE	2675.32	2658.21
RMSE as % of Mean Target	16.89%	16.79%
R^2 Score	0.9006	0.9019

Model Interpretation

For Dataset1:

Both models perform strongly, with R^2 values around 0.80, meaning they explain about 80% of the variance in medical costs. The Decision Tree achieves slightly lower MAE (2591.6682 vs. 4181.35) and RMSE (4326.9782 vs. 5966.99), and marginally higher R^2 (0.8981 vs. 0.8062) compared to Linear Regression. This indicates that while both models are reliable, the Decision Tree provides a slightly better predictive performance.

For Dataset2:

Both models perform strongly, with R^2 values around 0.90, meaning they explain about 90% of the variance in medical costs. The Decision Tree achieves slightly lower MAE (2066.35 vs. 2143.18) and RMSE (2658.21 vs. 2675.32), and marginally higher R^2 (0.9019 vs. 0.9006) compared to Linear Regression. This indicates that while both models are reliable, the Decision Tree provides a slightly better predictive performance.

Conclusion

Summary of findings

The comparative analysis across both datasets shows that Linear Regression and Decision Tree models demonstrate strong predictive capability for medical cost prediction, with high R^2 values (≈0.80 for Dataset1 and ≈0.90 for Dataset2). However, in both cases, the Decision Tree consistently outperforms Linear Regression by achieving lower MAE and RMSE, along with slightly higher R^2 values. This suggests that while both models are reliable, the Decision Tree offers superior accuracy and robustness in capturing the underlying patterns of the data, making it the preferred choice for this prediction task.

.