



Medical Insurance Cost Analysis

Tanya Arya, Mia Saavedra,
Jisoo Park, Amolika Kondapalli



Table of contents

01. Introduction

02. Data Description

03. Methodology

04. Results

05. Conclusion



Problem:

What factors influence medical expenses the most?

Factors include: age, sex, BMI, smoking status, number of children, and region.





Hypothesis

Groups with the highest charges:

Older Adults

Smokers

More
Children

Women

High BMI





Introduction to Dataset



Steps

1.

Origin

Dataset "Medical Insurance
Cost Prediction"
from Kaggle.com

2.

Pre-processing

2a

Cleaning

Deleted duplicates and null
values

2b

Transformation

Convert data to
dataframe in Jupyter
Notebook

Introduction to Dataset

Project Goal

Analyze the factors that influence medical expenses such as age, sex, BMI, smoking status, number of children, and region.



Audience

The audience would be stakeholders interested in healthcare analytics, insurance companies, government officials

Purpose

The project aims to gain insights into how different factors influence medical expenses.




Context

Perform statistical tests, such as linear regression, correlation analysis, t-test, and ANOVA




Data Description





Variable	Data Type	Description
Age	Integer	Range 18-64
Sex	Object	Male, Female
Body Mass Index (BMI)	Float	Range 15.96 - 53.13
Children	Integer	Treated as categorical
Smoker	Object	Yes, No
Region	Object	NE, NW, SE, SW
Charges	Float	Range \$1,121 - \$63,770





Methodology



Steps

1.

EDA

Exploratory Data Analysis

2.

Statistical Tests

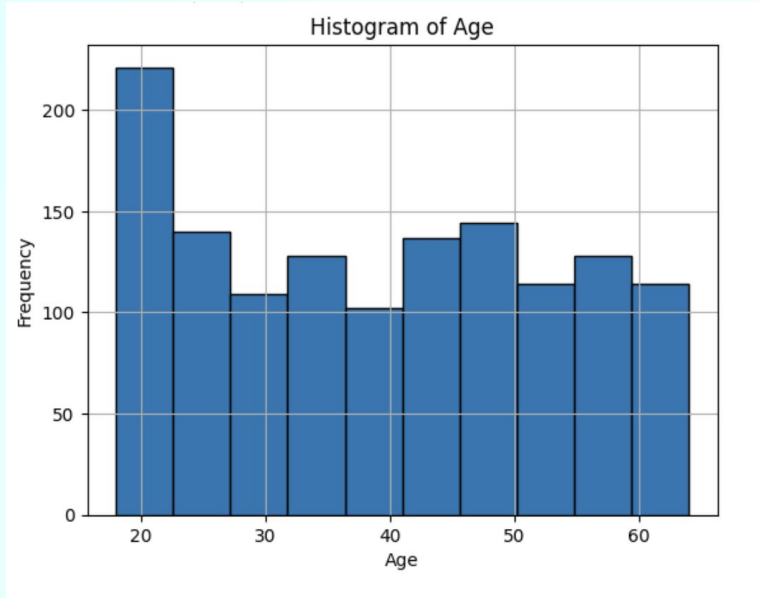
Correlation, T-test, ANOVA,
linear regression

3.

Data Visualization

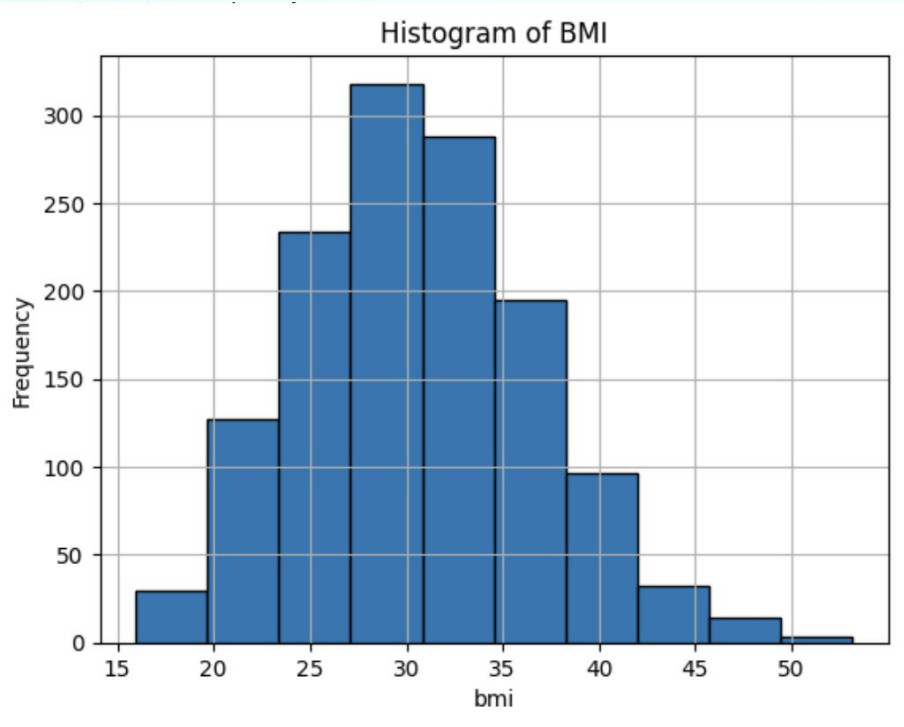
Box plots between variables
and charges

Descriptive Stats- Age



Count	1337.00
Mean	39.22
Standard Deviation	14.04
Min	18.00
Median	39.00
Max	64.00

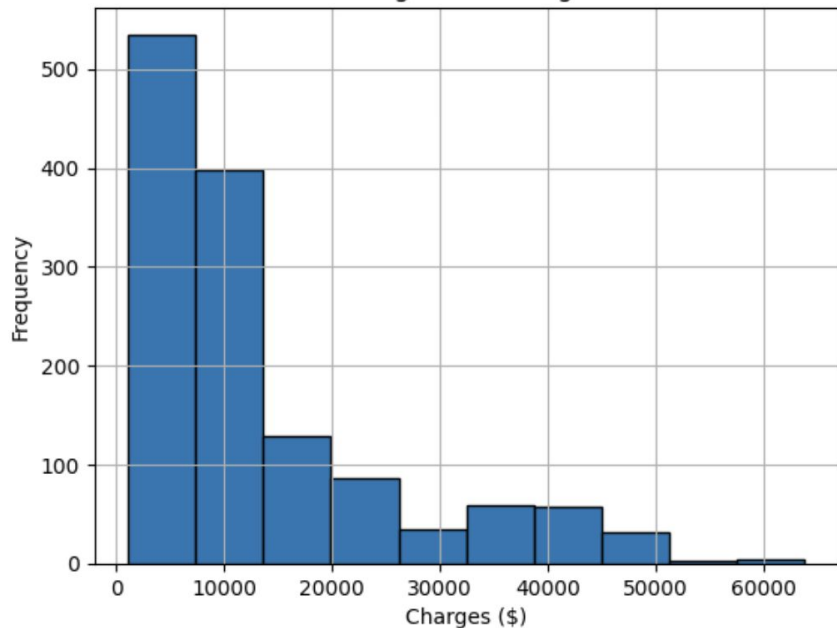
Descriptive Stats- BMI



Count	1337.00
Mean	30.66
Standard Deviation	6.10
Min	15.96
Median	30.40
Max	53.13

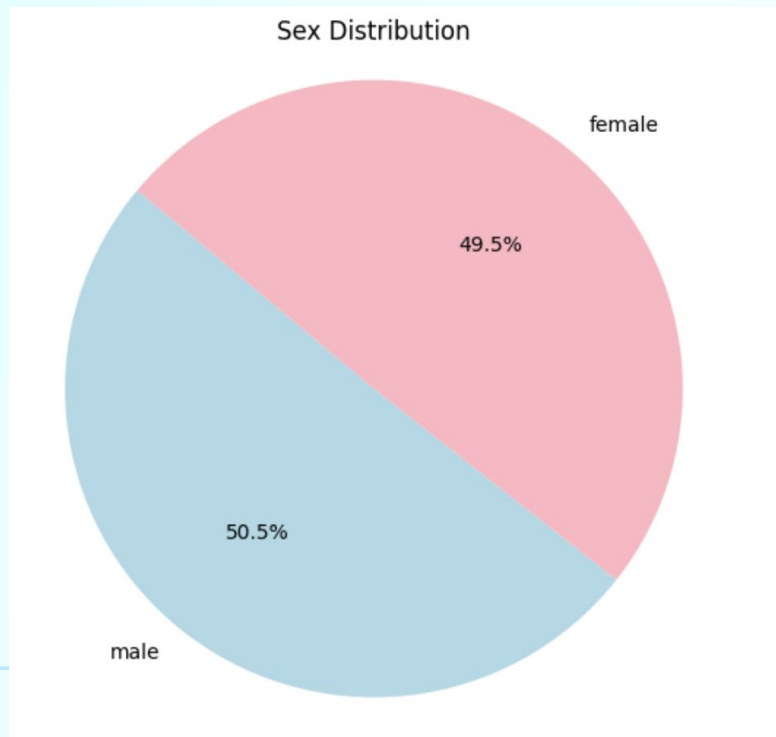
Descriptive Stats- Charges

Histogram of Charges



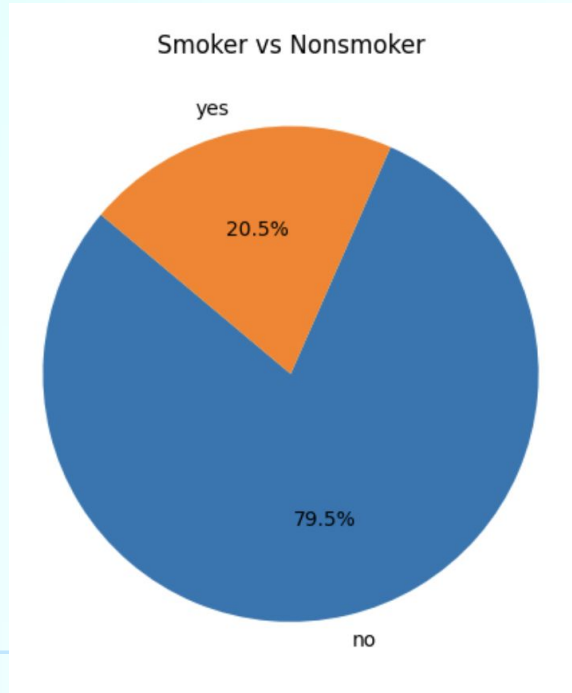
Count	1337.00
Mean	13279.12
Standard Deviation	12110.36
Min	1121.87
Median	9386.16
Max	63770.43

Descriptive Stats- Sex



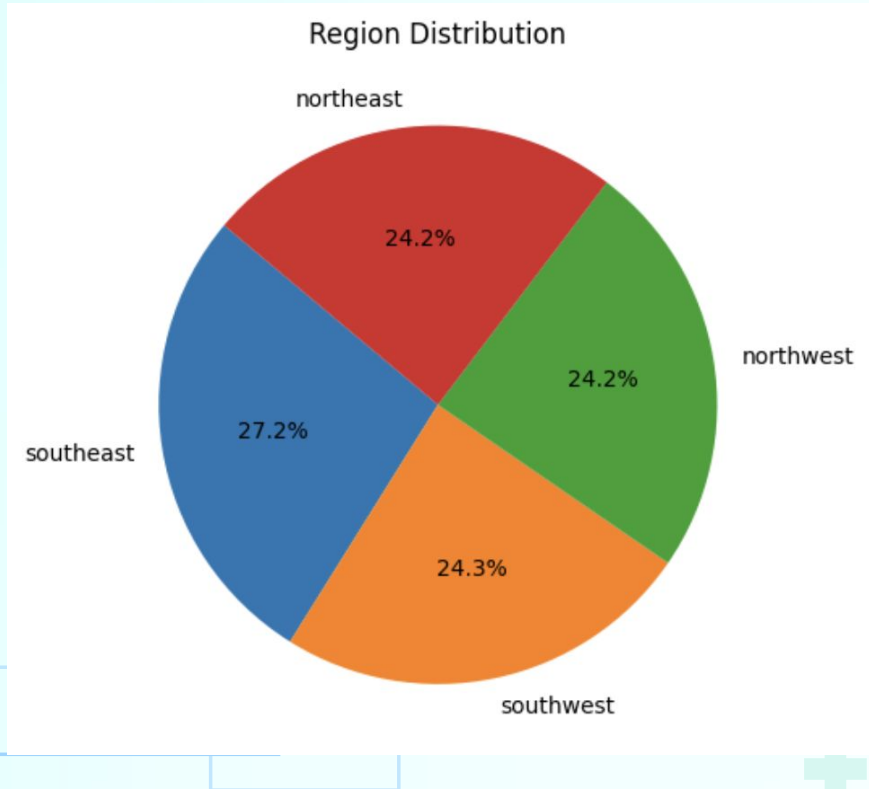
Count	1337.00
-------	---------

Descriptive Stats- Smoker



Count	1337.00
-------	---------

Descriptive Stats- Region



Count	1337.00
-------	---------



Results



Correlations

Testing each variable against the insurance charges to find strongest correlation using the Pearson Coefficient

Age:

$$r = 0.298$$

$$R^2 = 0.888$$

8.88% of the variability in charges is attributed to age

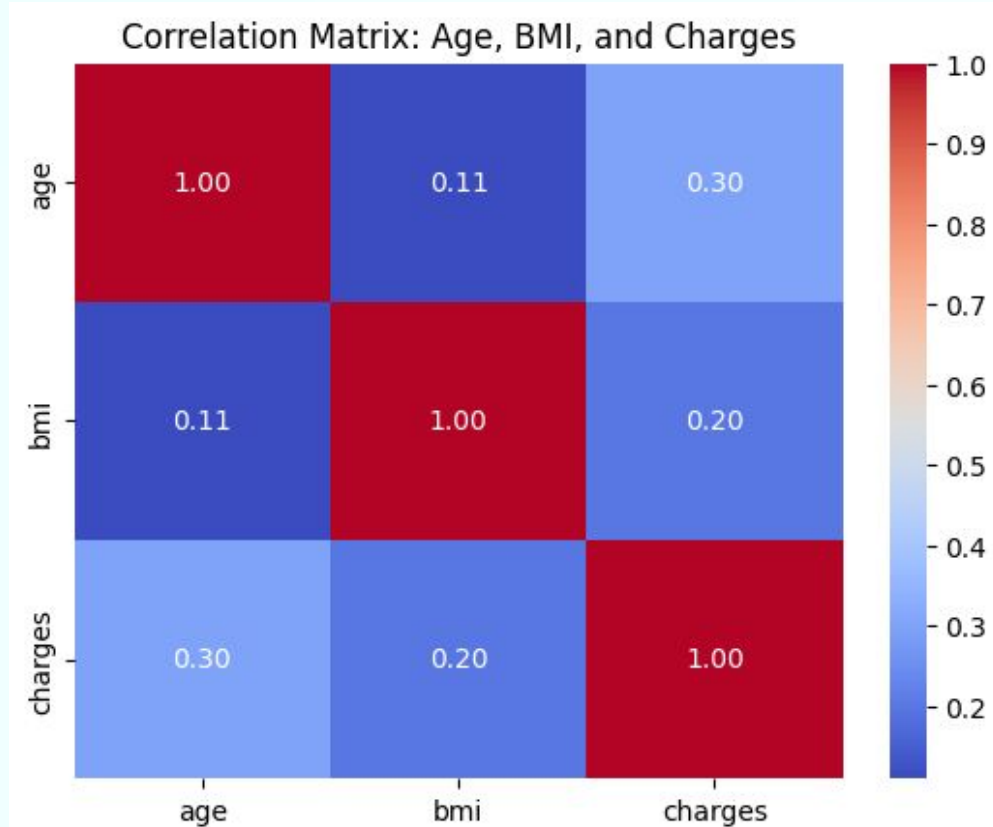
BMI:

$$r = 0.198$$

$$R^2 = 0.0392$$

3.92% of the variability in charges is attributed to BMI

Correlation Matrix





T-Test Hypotheses

Sex:

Null:

The mean charges of the male and female groups are the same.

Alternate:

The mean charges of the male and female groups are not the same.

Smoker:

Null:

The mean charges of the smoker and non-smoker groups are the same.

Alternate:

The mean charges of the smoker and non-smoker are not the same.





T-Tests

Sex:

T-Statistic: 2.124

P-Value: 0.0338

Reject the null hypothesis.

There is a significant
difference between the two
groups (male, female)

Smoker:

T-Statistic: 46.645

P-Value: 1.407e-282

Reject the null hypothesis.

There is a significant
difference between the two
groups (non-smoker, smoker)





ANOVA Hypotheses

Region:

Null:

The mean charges for NE, NW, SE, and SW are the same.

Alternate:

The mean charges for NE, NW, SE, and SW are not the same.

Number of Children:

Null:

The mean charges for 0, 1, 2, 3, 4, and 5 children are the same.

Alternate:

The mean charges for 0, 1, 2, 3, 4, and 5 children are not the same.



ANOVA

Region:

F-Statistic: 2.926

P-Value: 0.03276

Reject the null hypothesis.
There is a significant difference
in charges among different
regions.

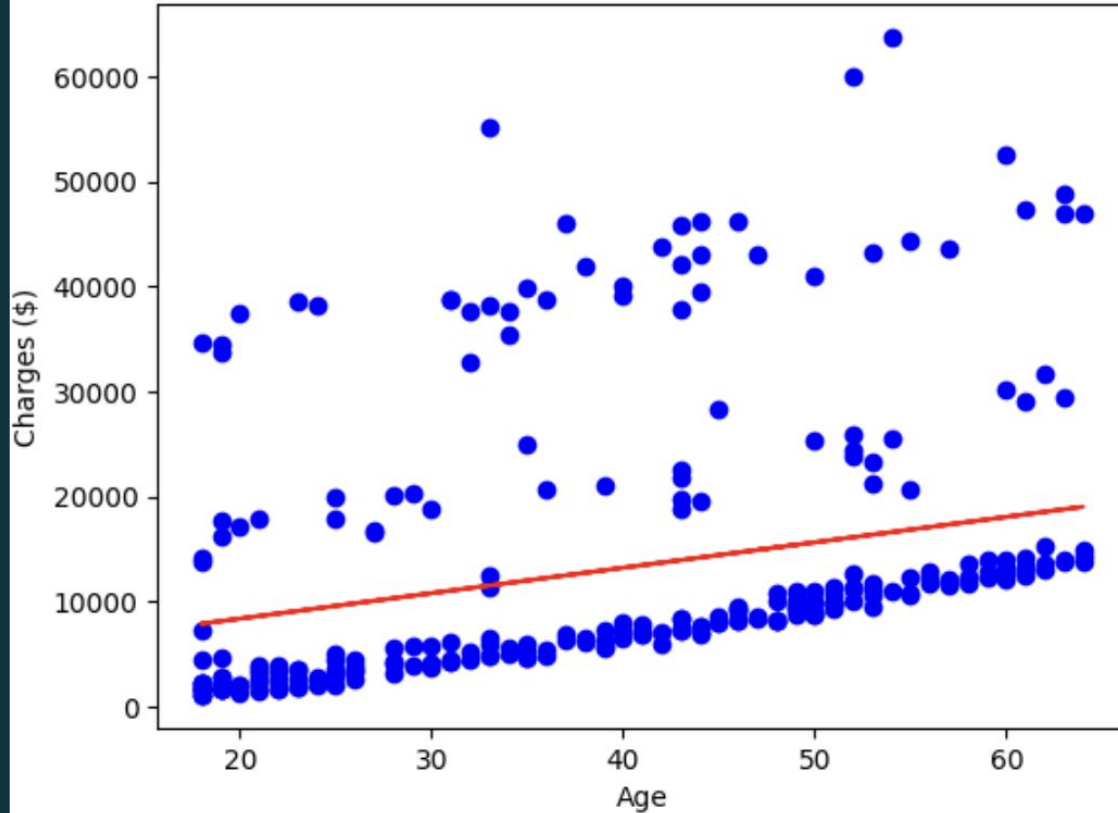
Number of Children:

F-Statistic: 3.268

P-Value: 0.0061

Reject the null hypothesis.
There is a significant difference
in charges based on the
number of children.

Linear Regression: Age vs Charges



Coefficient: 242.26

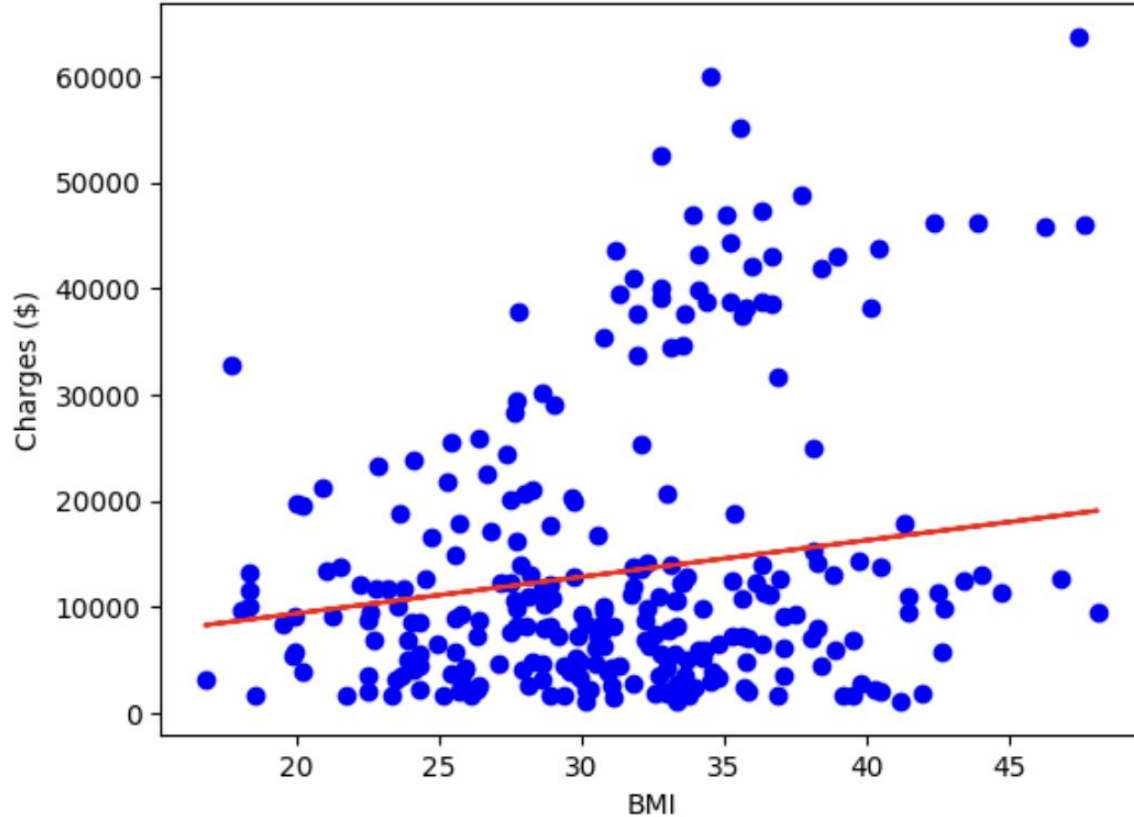
Intercept: 3532.09

$$\text{Charges} = 242.26(\text{Age}) + 3532.09$$

Mean Squared Error:
166,275,348.23

$$R^2 = 0.095$$

Linear Regression: BMI vs Charges



Coefficient: 345.17

Intercept: 2488.57

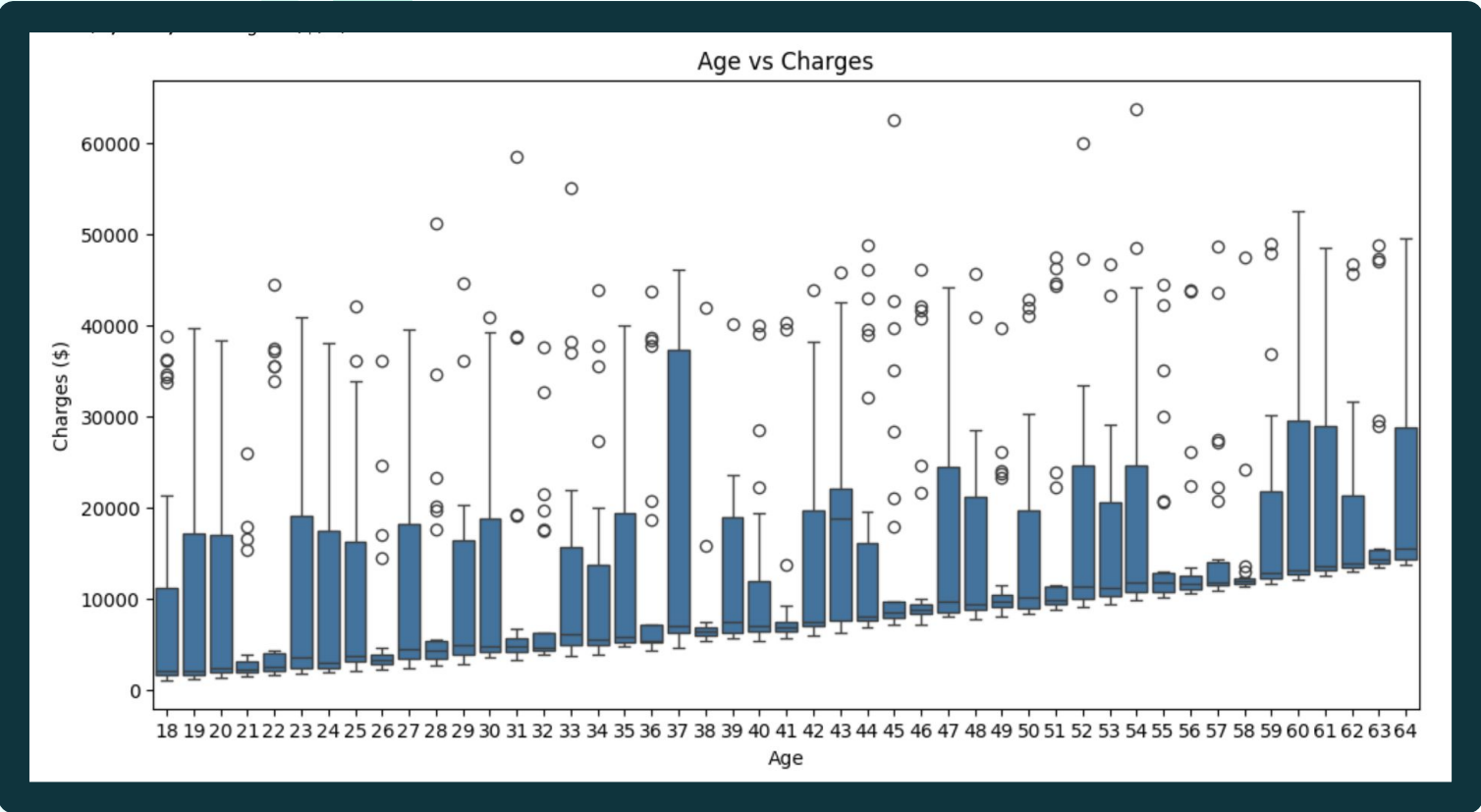
Charges = 345.17(BMI) +
2488.57

Mean Squared Error:
174,251,720.52

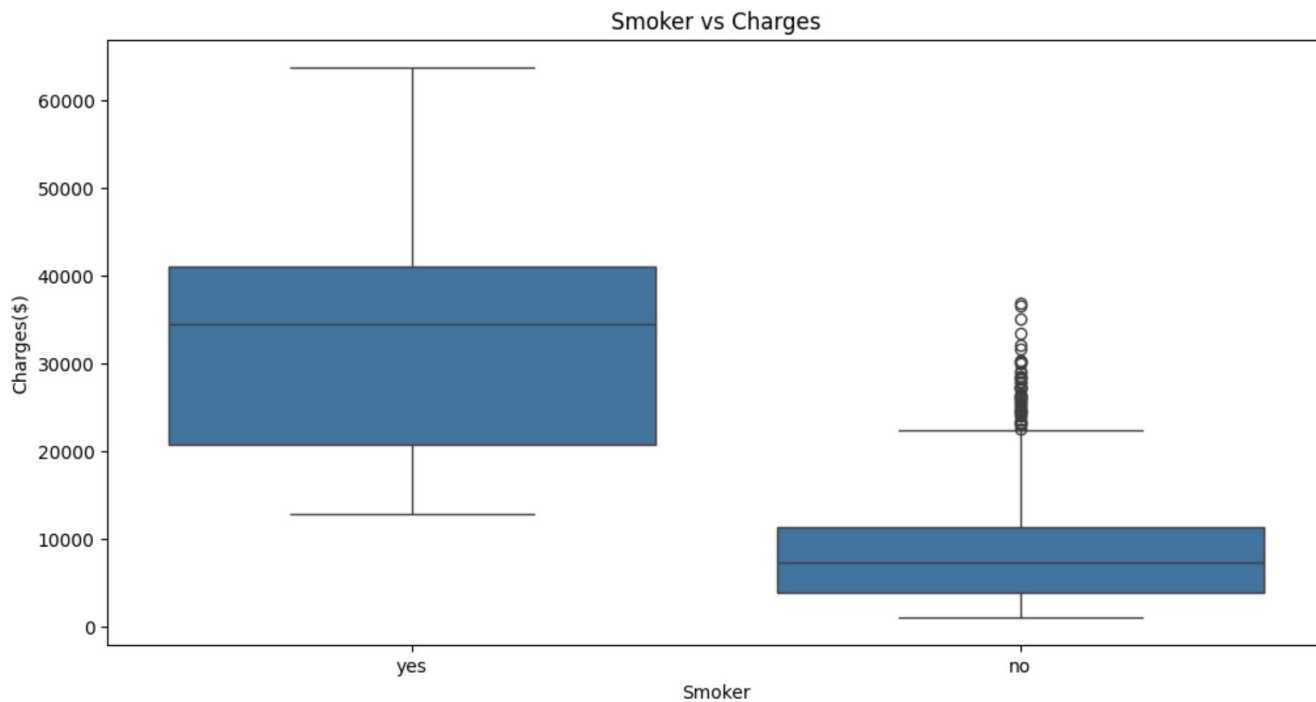
$R^2 = 0.0517$



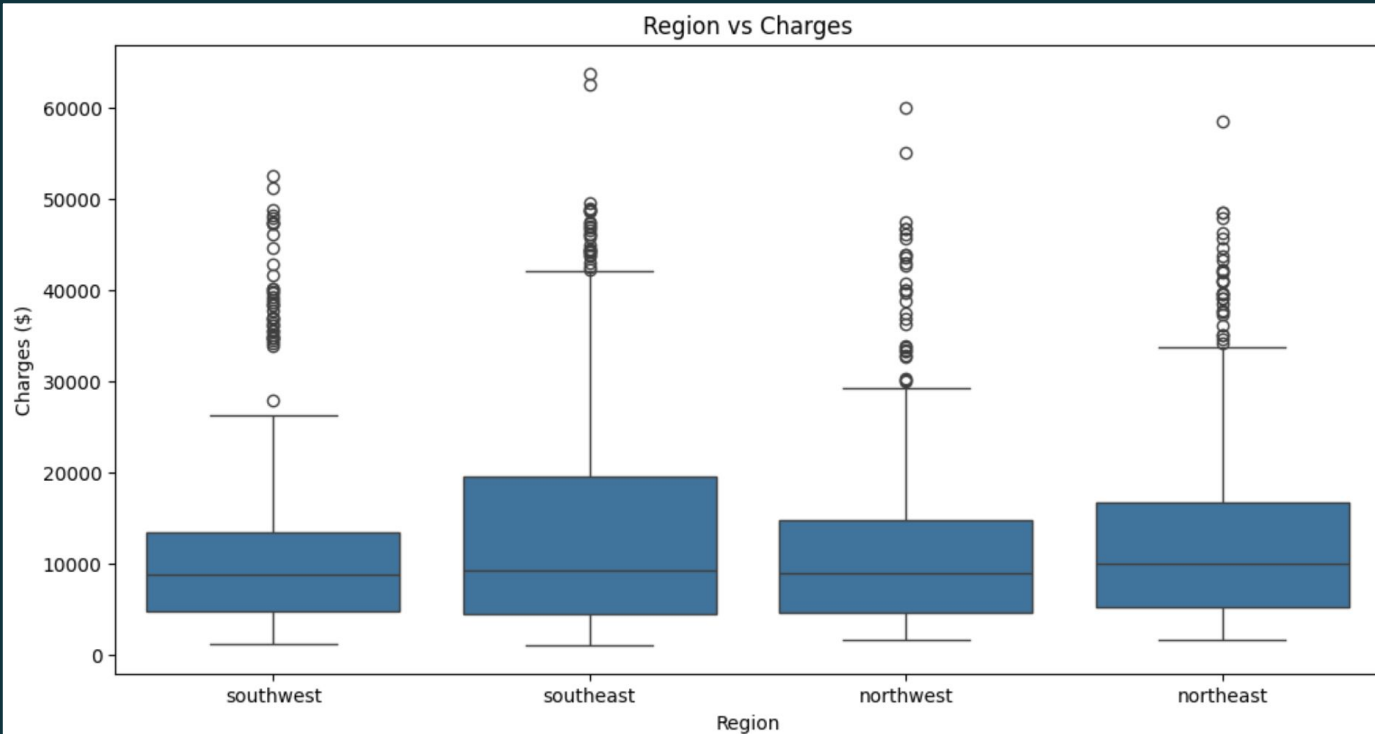
Slight positive trend between Age and Charges



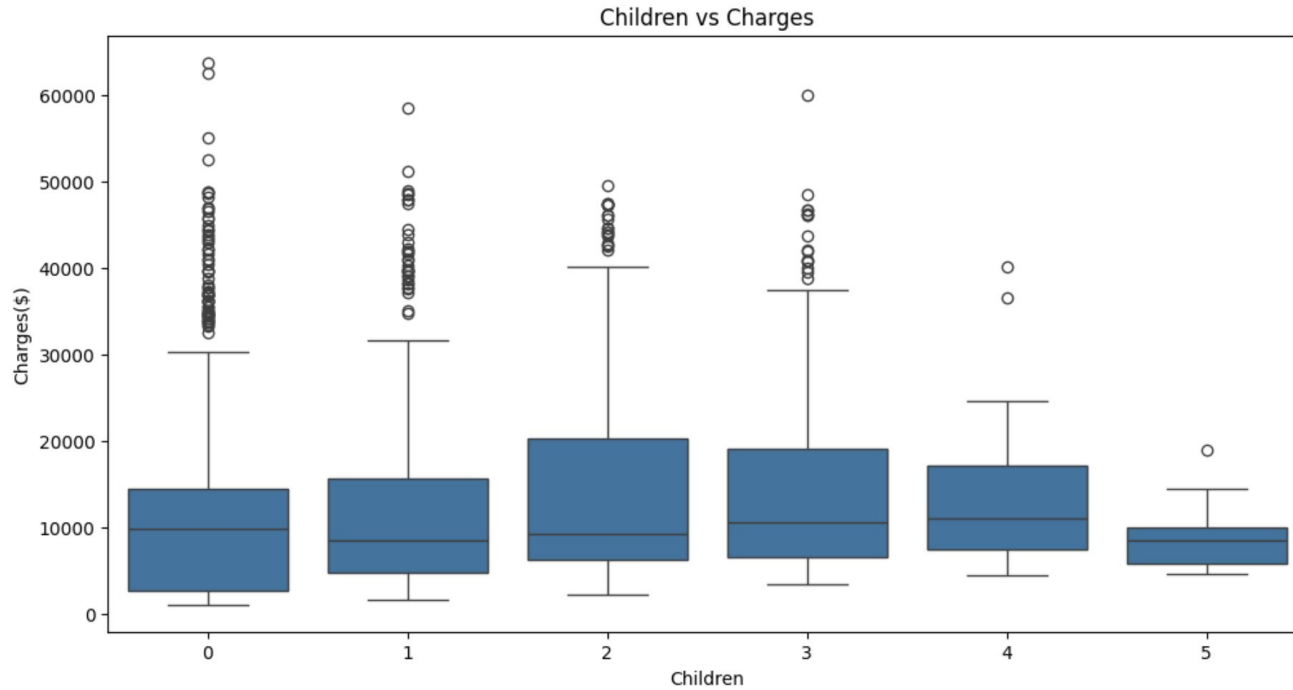
Smokers had higher medical insurance charges overall compared to non-smokers



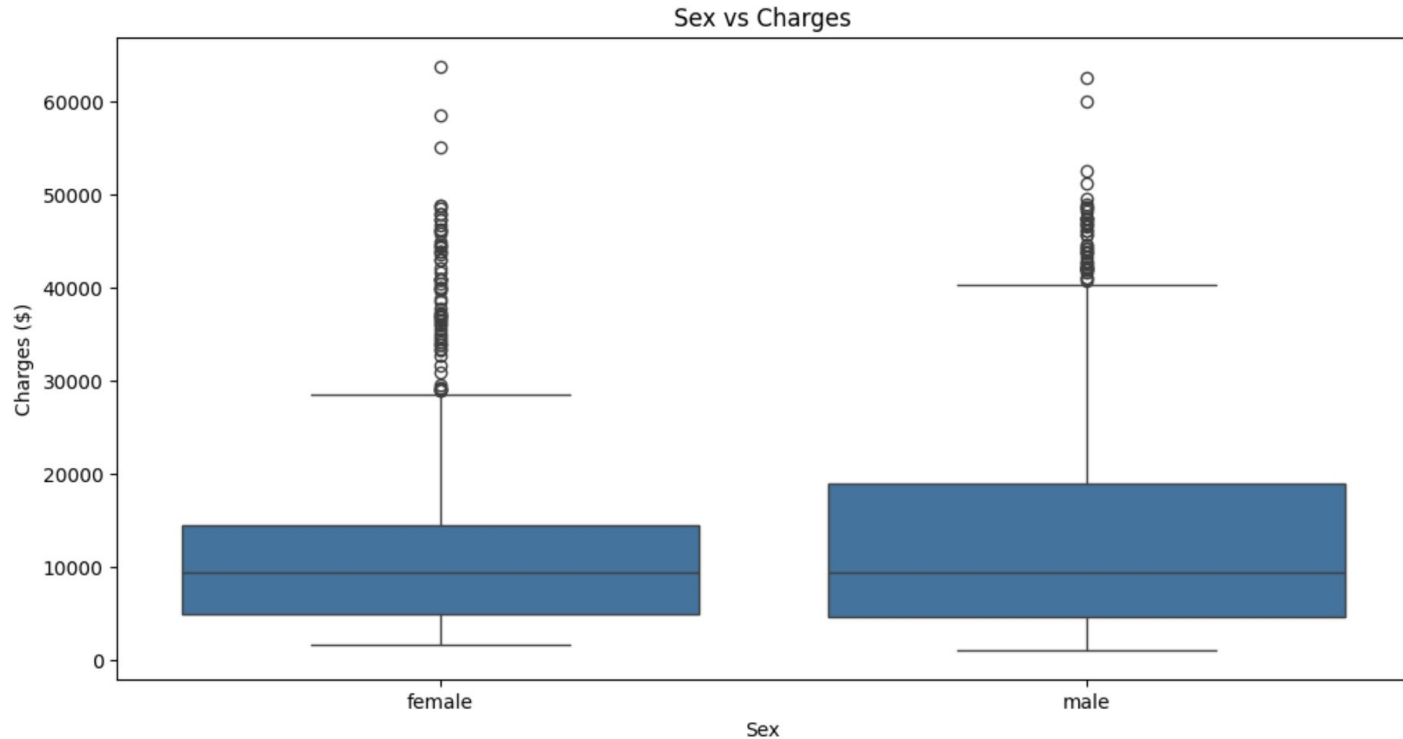
Slight differences between region groups, SE generally had higher charges



Small differences between # Children and Charges, Similar median values



Similar median values, but upper 50% of males had higher charges than females





Conclusion





Factor Results



Age

Moderately Weak Positive Correlation

BMI

Weak Positive Significant Correlation

Smoker

Significant difference in charges
between smoker and non-smoker

Children


Significant difference in charges
based on number of children

Region

Significant difference in charges
among regions

Sex

Significant difference between male
and female



Implications



Healthcare Providers

Tailor services to better meet the needs of different patient demographics.



Insurance Companies

Develop more accurate pricing models and risk assessments



Policy Makers

Make better decisions to address disparities in healthcare access



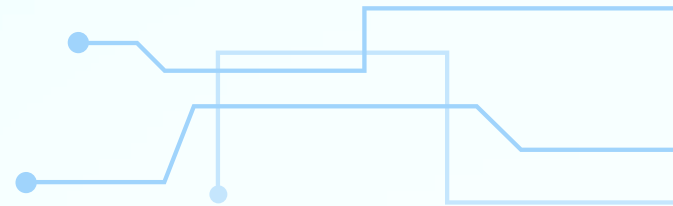
Research

Contribute to researchers of healthcare economics and public health



Public

Transparency about medical costs can help the public make smarter lifestyle and insurance choices



Further Research

Analyze further variables related to insurance charges:

- Disabilities
- Health Conditions
- Diet



Thank you!

