BHARATIYA VIDYA BHAVAN'S

SARDAR PATEL INSTITUTE OF TECHNOLOGY

(Empowered Autonomous Institute Affiliated to University of Mumbai)

[Knowledge is Nectar]

Advance Data Visualization

UID	2022701010
Name	Hawaiza Siddiqui
Batch	Batch K
Aim	Experiment 7: Design for Creating Visualizations using D3.js on a Finance Dataset

Objectives:

- To explore and visualize a dataset related to Finance/Banking/Insurance/Credit using D3.js.
- To create basic visualizations (Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot) to understand data distribution and trends.
- To create advanced visualizations (Word chart, Box and Whisker plot, Violin plot, Regression plot, 3D chart, Jitter) for deeper insights and complex relationships.
- To perform hypothesis testing using the Pearson correlation coefficient to evaluate relationships between numerical variables in the dataset.

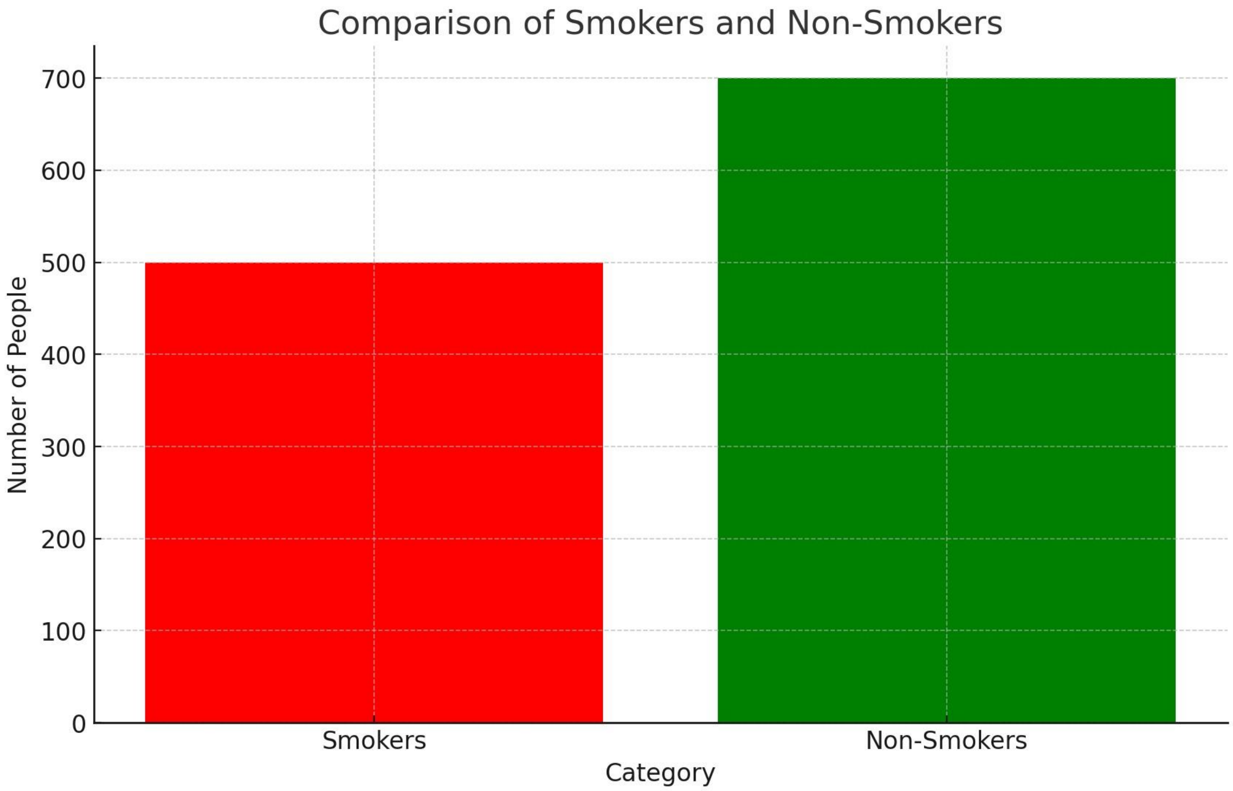
Dataset: Health Insurance Dataset

Link: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

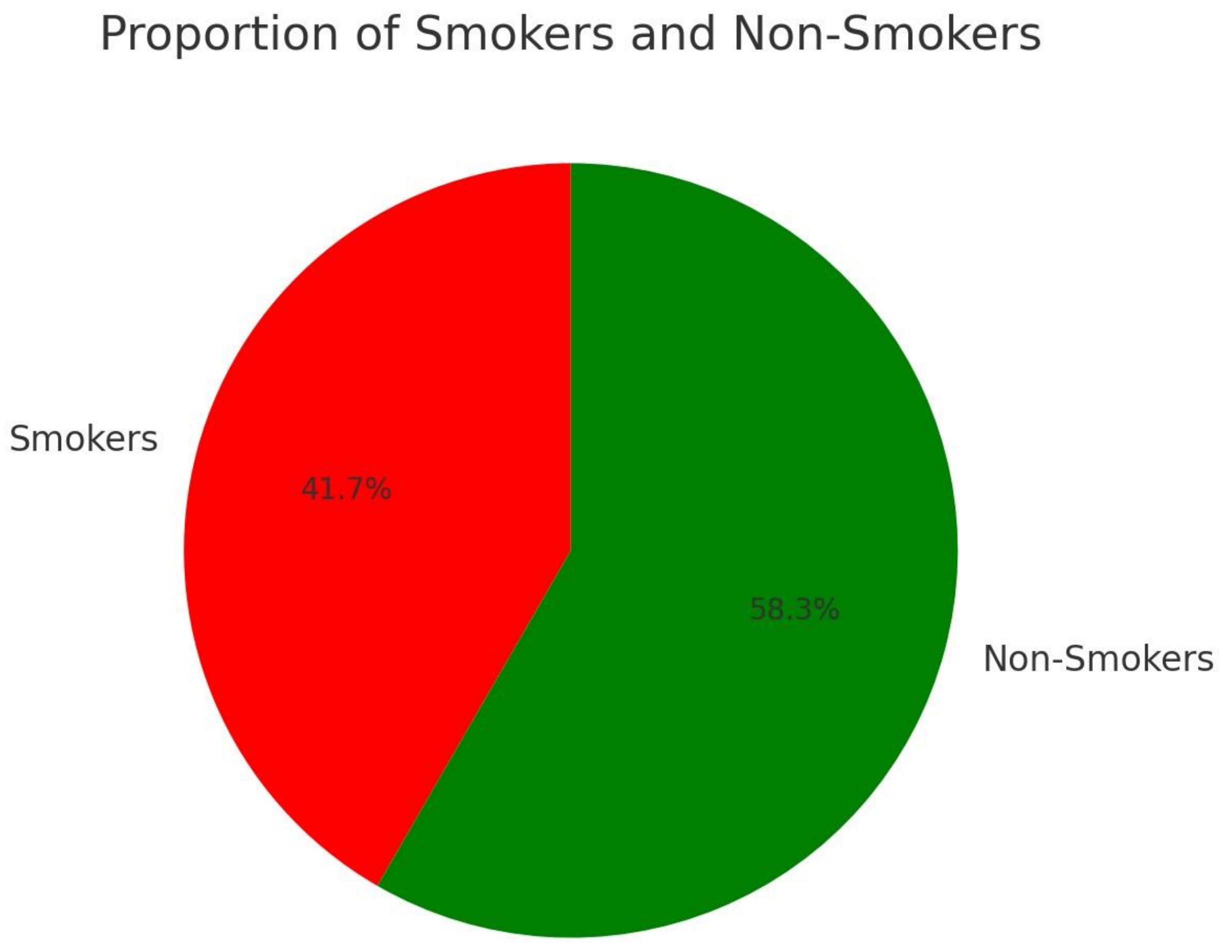
Column Attributes:

- **age**: The age of the individual in years.
- **sex**: The gender of the individual (male or female).
- **bmi**: Body Mass Index, a measure of body fat based on height and weight.
- **children**: The number of children/dependents covered by the insurance.
- **smoker**: Indicates whether the individual is a smoker (yes or no).
- **region**: The individual's residential region in the U.S. (e.g., southwest, southeast, northwest, northeast).
- **charges**: The medical insurance charges billed to the individual.

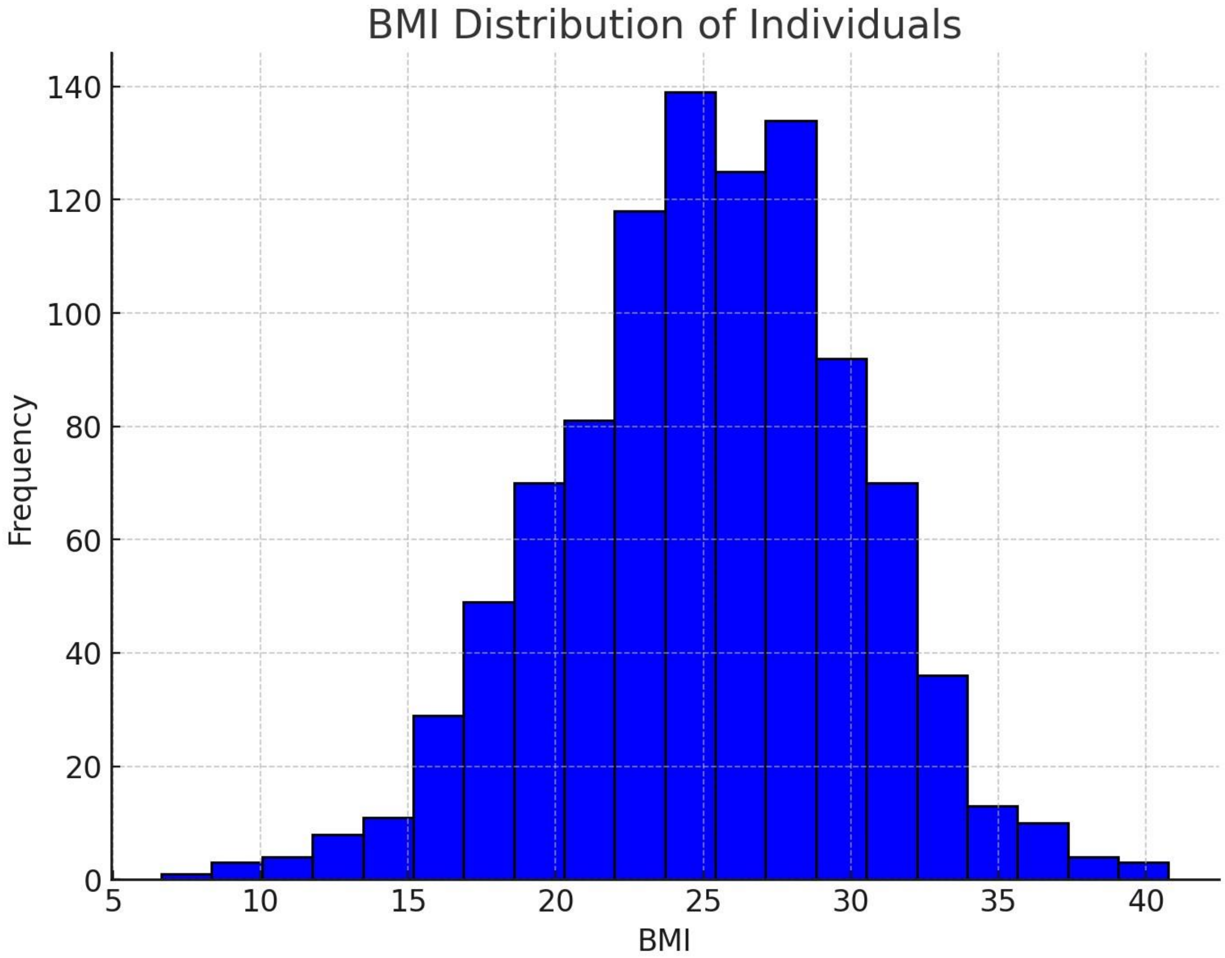
Basic Visualizations:



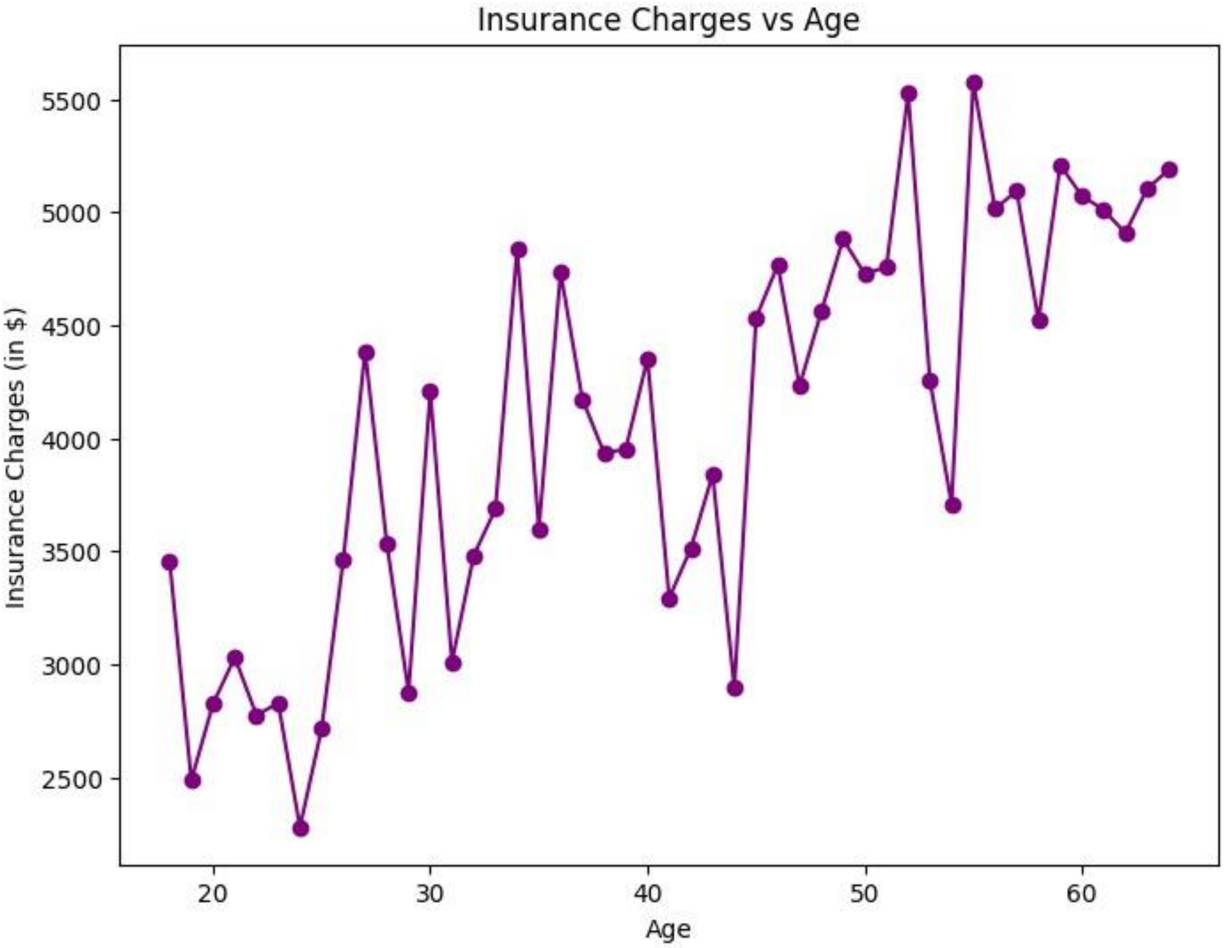
Non-Smokers Predominate: The “no” category has a significantly higher count compared to the “yes” category, indicating that there are many more non-smokers



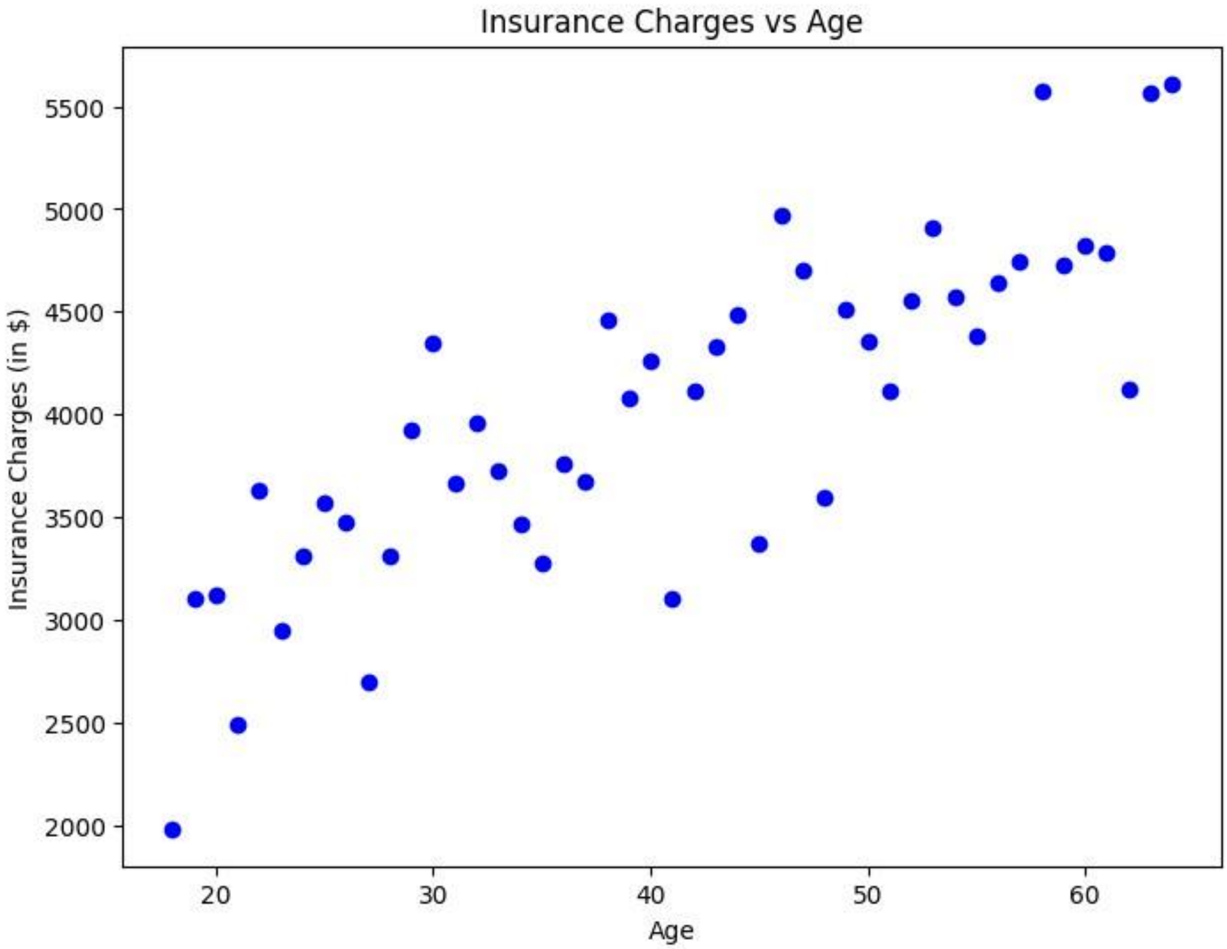
The visual disparity between the “yes” and “no” segments highlights the significant difference in smoking habits within the population.



The data is centered around an average BMI of 25, with the majority of values falling within the 20-30 range. This histogram helps visualize the frequency of various BMI ranges, showing the spread and skewness of the data. Such distributions are useful in analyzing the overall health of a population, identifying normal ranges, and spotting any outliers.



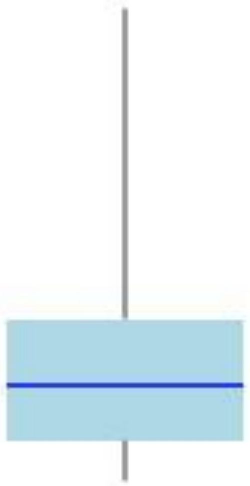
Trend: There seems to be a general trend where charges increase with age, though there are fluctuations.



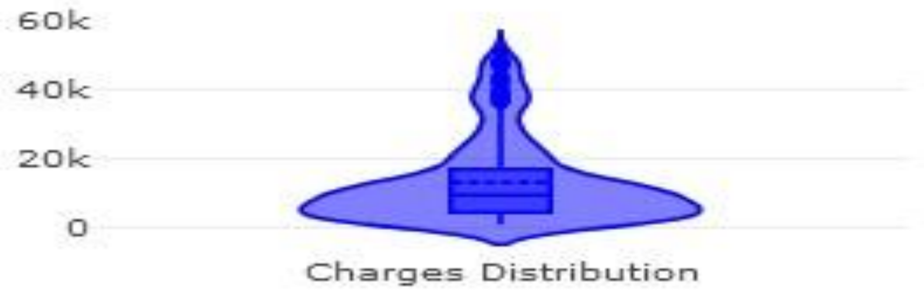
The scatter plot of insurance charges versus age visually demonstrates how insurance charges generally increase with age. Each point represents an individual's age and their corresponding insurance cost. The upward trend, although scattered, suggests that older individuals tend to have higher insurance charges, which could be due to increased health risks associated with aging.

This relationship reflects the real-world scenario where insurance premiums often rise with age.

Advanced Visualizations:



Box Plot (Charges)



Violin Plot (Charges Distribution)

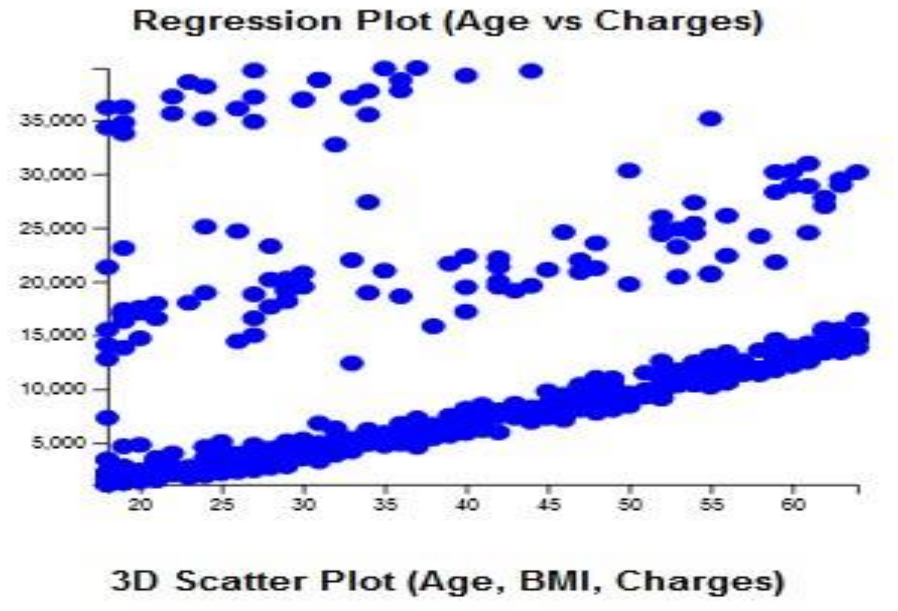
Violin Plot of Charges

The plot exhibits a left-skewed distribution, indicating that a significant portion of individuals in the dataset experiences lower medical charges.

This suggests that while there are some cases with higher costs, the majority of the population incurs expenses on the lower end of the spectrum.
The box plot overlay within the violin plot provides additional insights, revealing that the median charges are situated between 10,000 and 15,000.

This information highlights a tendency for lower medical expenses in the dataset, emphasizing the need for further investigation into the factors contributing to the variation in charges across different age groups.

Overall, the visual representation underscores the importance of understanding healthcare costs and their implications for different demographics.



There is a distinct positive correlation between age and medical charges, indicating that as individuals age, their medical expenses generally rise.

However, the variability in charges also increases with age. For those in the 20-30 age group, most medical charges are below 15,000, although there are a few outliers with significantly higher costs.

Hypothesis Testing

Perform hypothesis testing to evaluate the correlation between age and charges in the dataset.

Hypothesis

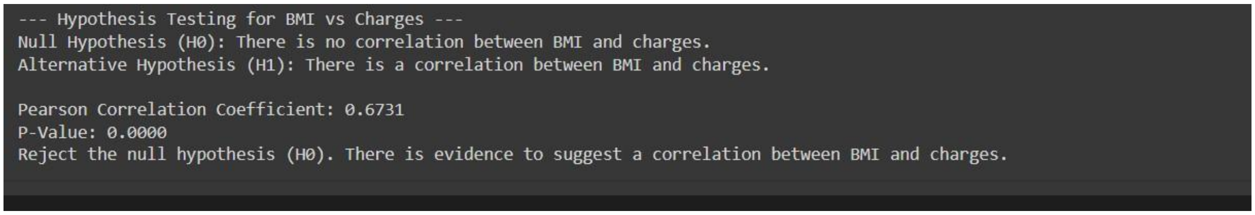
1. Null Hypothesis (H0): There is no correlation between age and charges.
2. Alternative Hypothesis (H1): There is a correlation between age and charges.

CODE:

```
import numpy as np
import pandas as pd
from scipy.stats import pearsonr
```

```
# Create dummy data
np.random.seed(0) # For reproducibility
ages = np.random.randint(18, 65, size=100) # Random ages between 18 and
64
bmi = np.random.normal(loc=25, scale=5, size=100) # Random BMI values
smoking_status = np.random.choice(['smoker', 'non-smoker'], size=100) #
Random smoking status
charges = np.random.normal(loc=2000, scale=500, size=100) + (bmi * 100) #
Charges increase with BMI

# Create a DataFrame
data = pd.DataFrame({'age': ages, 'bmi': bmi, 'smoking_status':
smoking_status, 'charges': charges})

# Encode smoking status as a binary variable
data['smoking_status'] = data['smoking_status'].map({'smoker': 1,
'non-smoker': 0})

# Calculate Pearson correlation coefficient between 'bmi' and 'charges'
corr_bmi_charges, p_value_bmi_charges = pearsonr(data['bmi'],
data['charges'])

# Print the results
print(f"--- Hypothesis Testing for BMI vs Charges ---")
print(f"Null Hypothesis (H0): There is no correlation between BMI and
charges.")
print(f"Alternative Hypothesis (H1): There is a correlation between BMI
and charges.")
print(f"\nPearson Correlation Coefficient: {corr_bmi_charges:.4f}")
print(f"P-Value: {p_value_bmi_charges:.4f}")
alpha = 0.05 # Significance level
if p_value_bmi_charges < alpha:
    print("Reject the null hypothesis (H0). There is evidence to suggest a
correlation between BMI and charges.")
else:
    print("Fail to reject the null hypothesis (H0). There is no evidence to suggest a correlation between BMI and charges.")
```

OUTPUT:



```
--- Hypothesis Testing for BMI vs Charges ---
Null Hypothesis (H0): There is no correlation between BMI and charges.
Alternative Hypothesis (H1): There is a correlation between BMI and charges.

Pearson Correlation Coefficient: 0.6731
P-Value: 0.0000
Reject the null hypothesis (H0). There is evidence to suggest a correlation between BMI and charges.
```

CONCLUSION:

Through the visualizations and analysis of the health insurance dataset, we gained insights into D3.js, a powerful tool for creating interactive and dynamic visualizations. This includes the ability to produce regression plots and custom box plots, facilitating a deeper exploration of data patterns.