Name	**Harditya Shah**
UID No.	**2021300114**
Batch	**L**

Objectives:

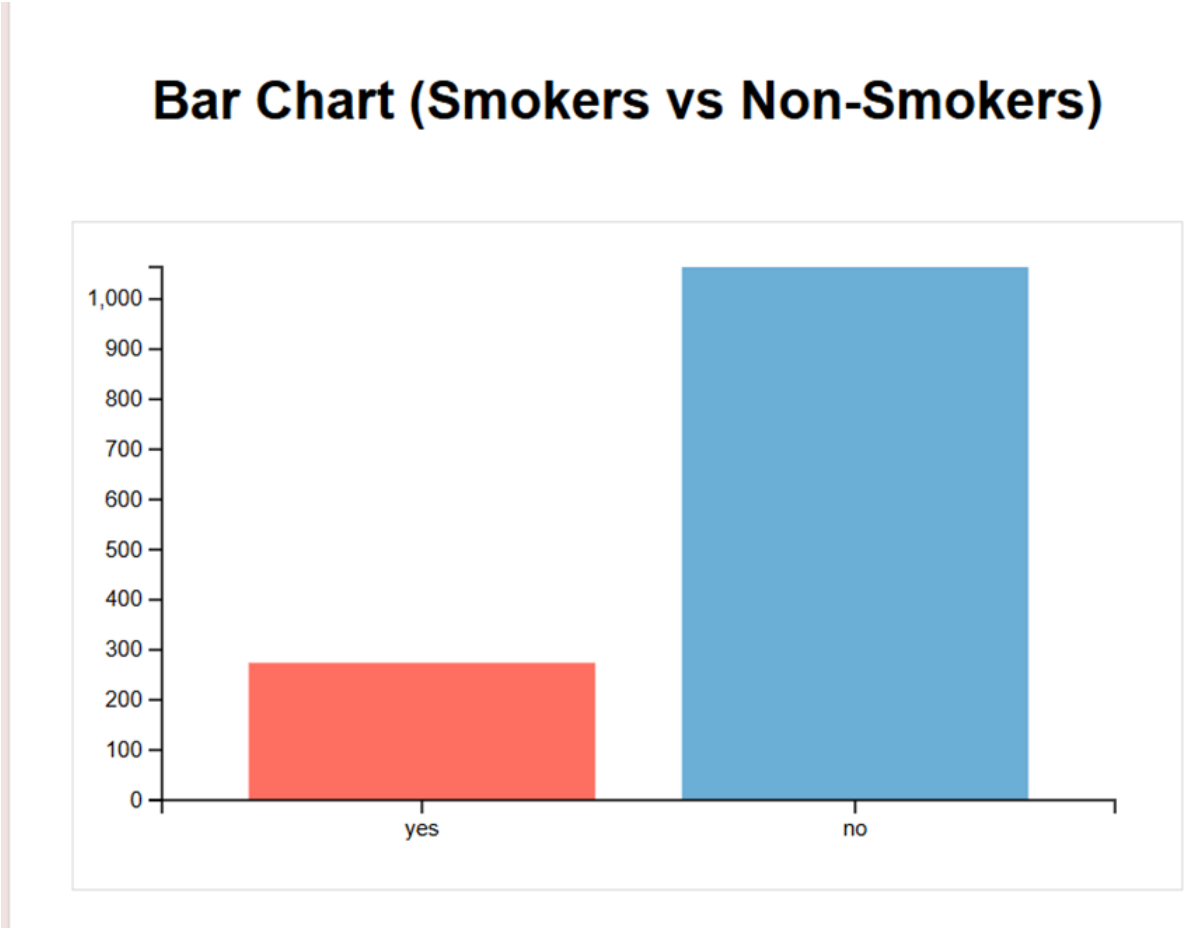
- To explore and visualize a dataset related to Finance/Banking/Insurance/Credit using D3.js.
- To create basic visualizations (Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot) to understand data distribution and trends.
- To create advanced visualizations (Word chart, Box 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: US Health Insurance Dataset

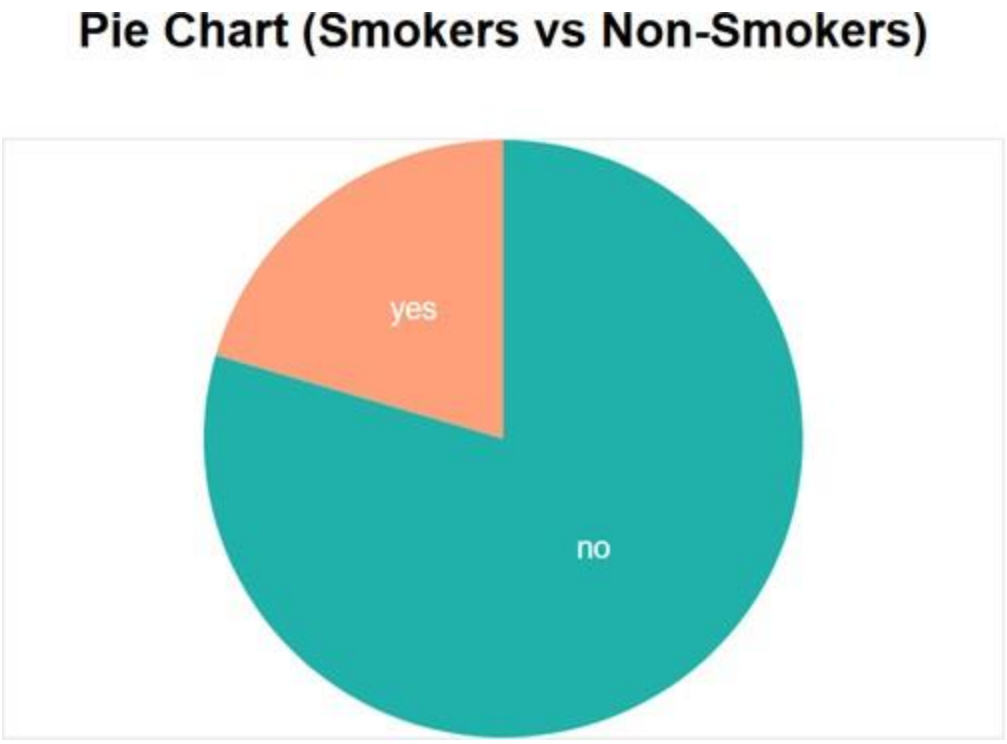
About Dataset:
This dataset is of insured data, where the Insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker and Region. The attributes are a mix of numeric and categorical variables.
Column Attributes:

- 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 individuals residential region in the U.S. (e.g., southwest, southeast, northwest, northeast).
- charges: The medical insurance charges billed to the individual.

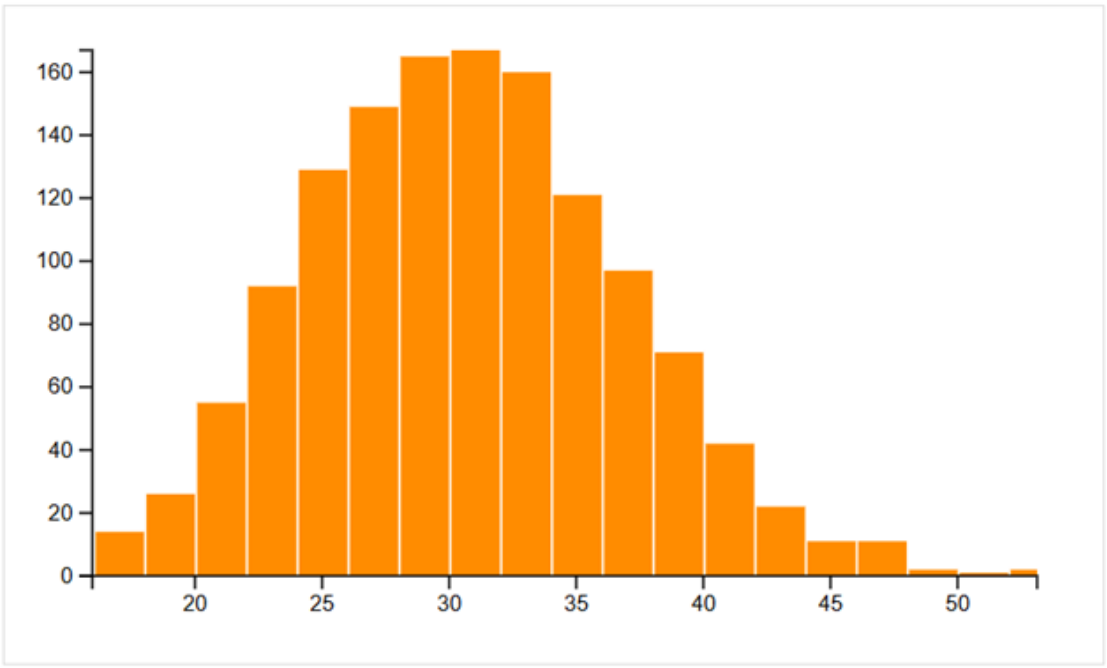
Basic Visualizations:



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

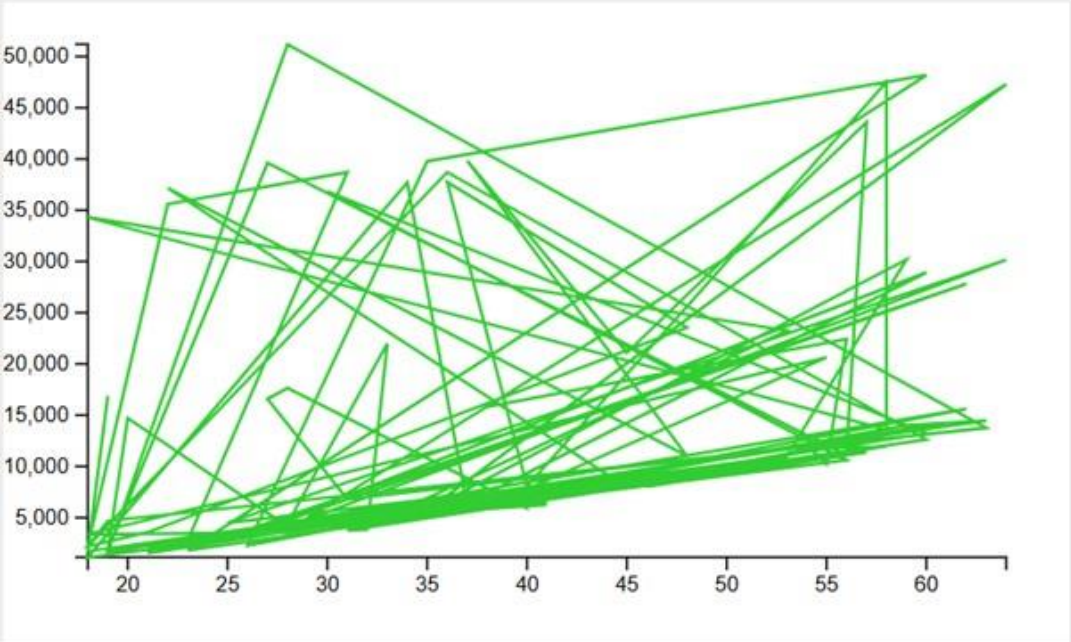


The visual disparity between the “yes” and “no” segments highlights the significant difference in smoking habits within the population.

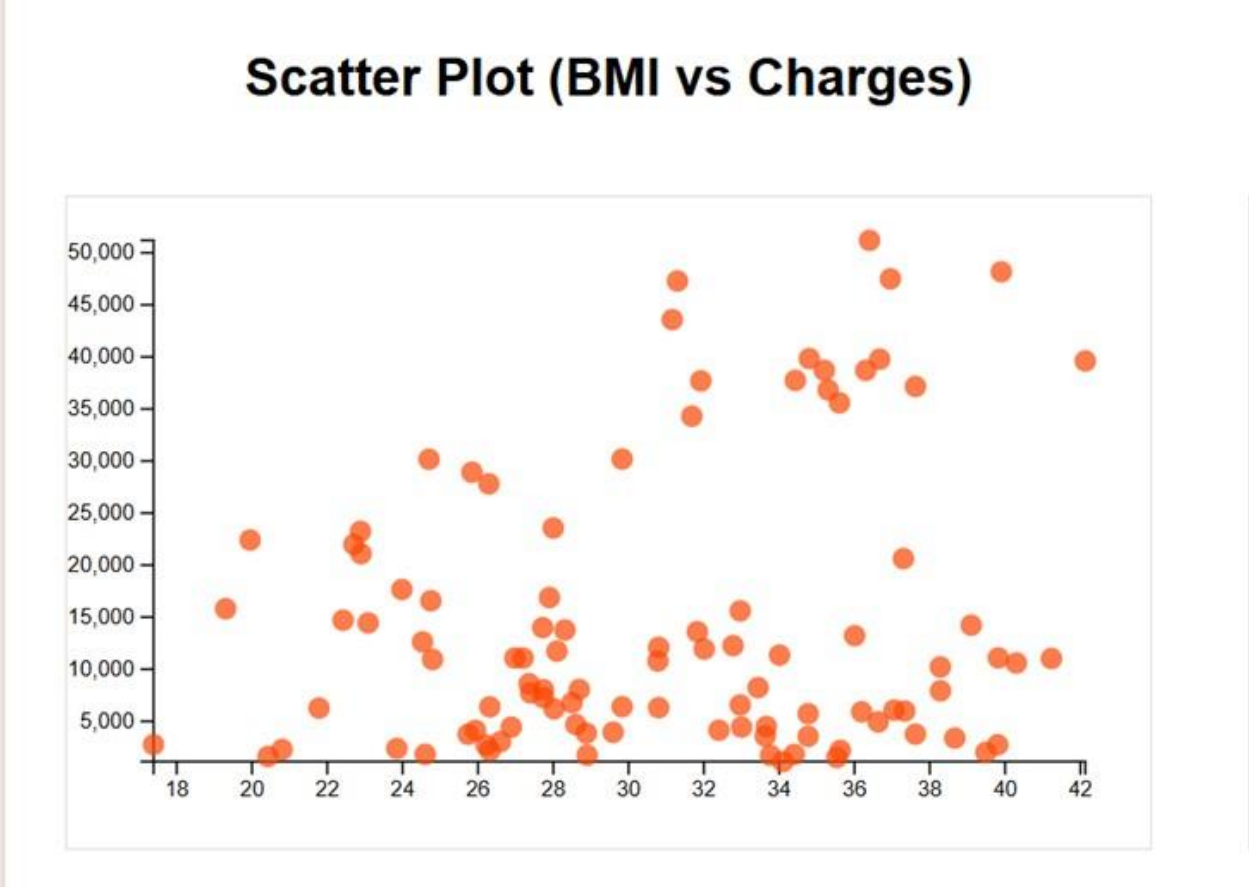


The highest frequency of BMI values is centered around 30, which indicates that most individuals in this dataset have a BMI close to this value.

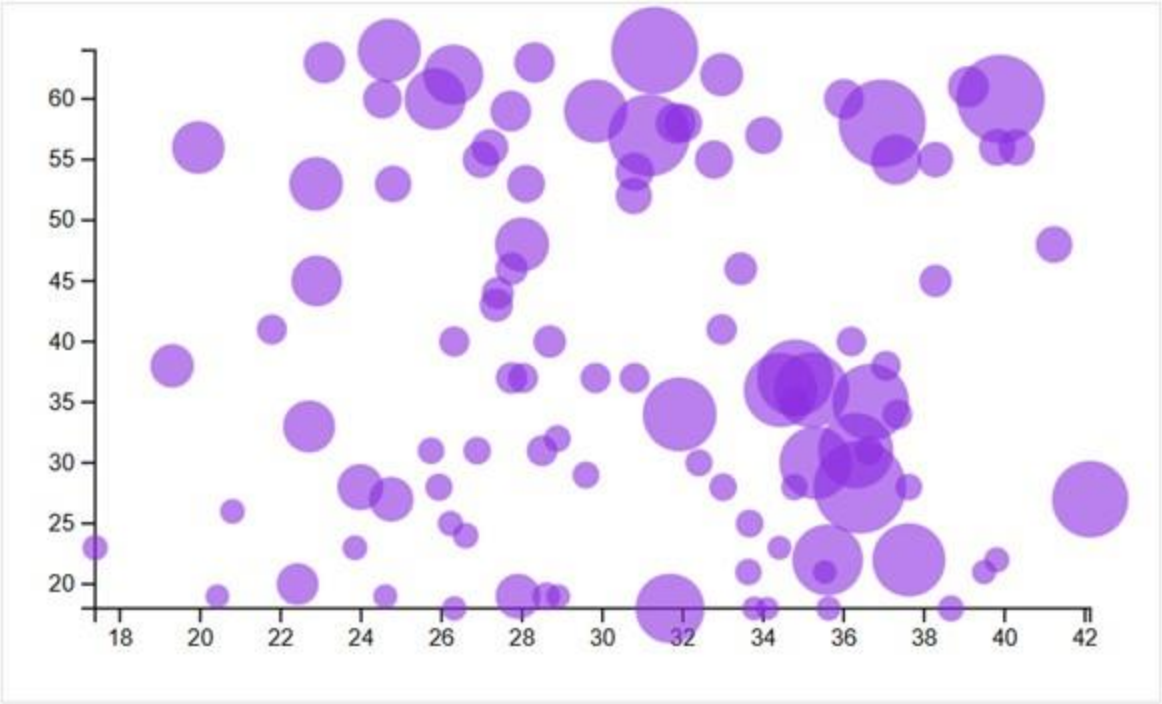
Timeline Chart (Charges vs Age)



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



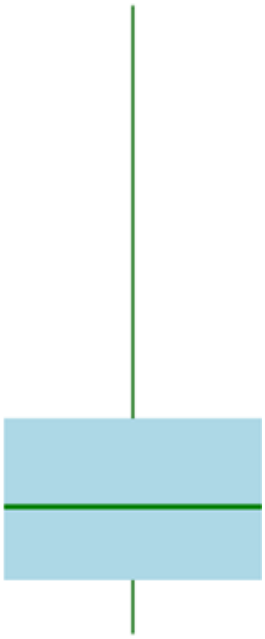
There is a positive correlation between BMI and charges, meaning that as BMI increases, the charges also tend to increase.



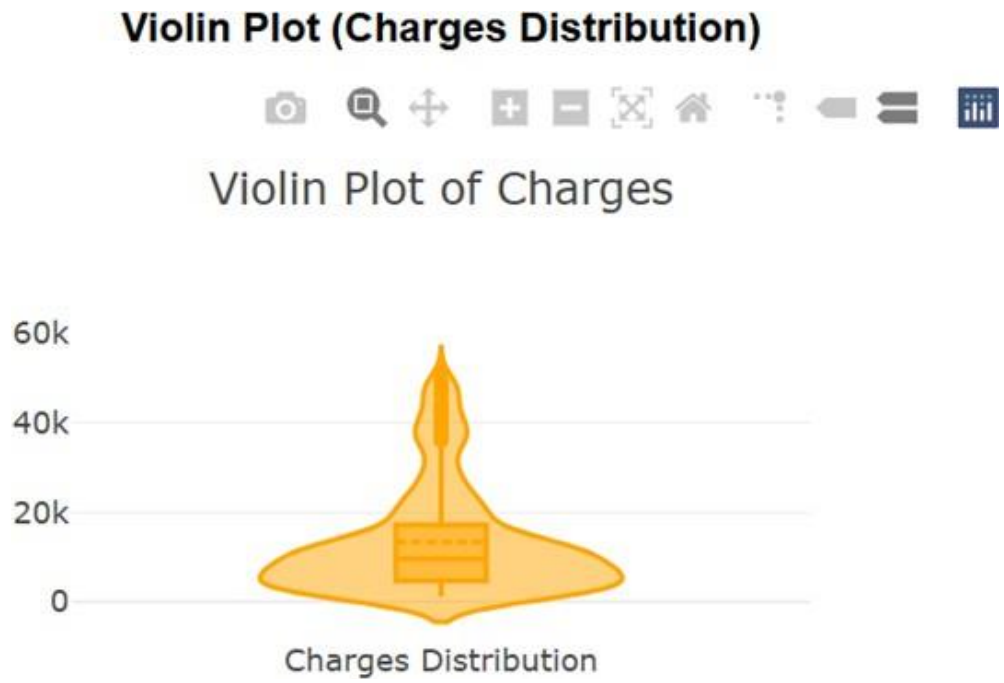
There are clusters of larger bubbles (higher BMI) at various age points, suggesting that higher BMI individuals incur higher charges.

Advanced Visualizations:

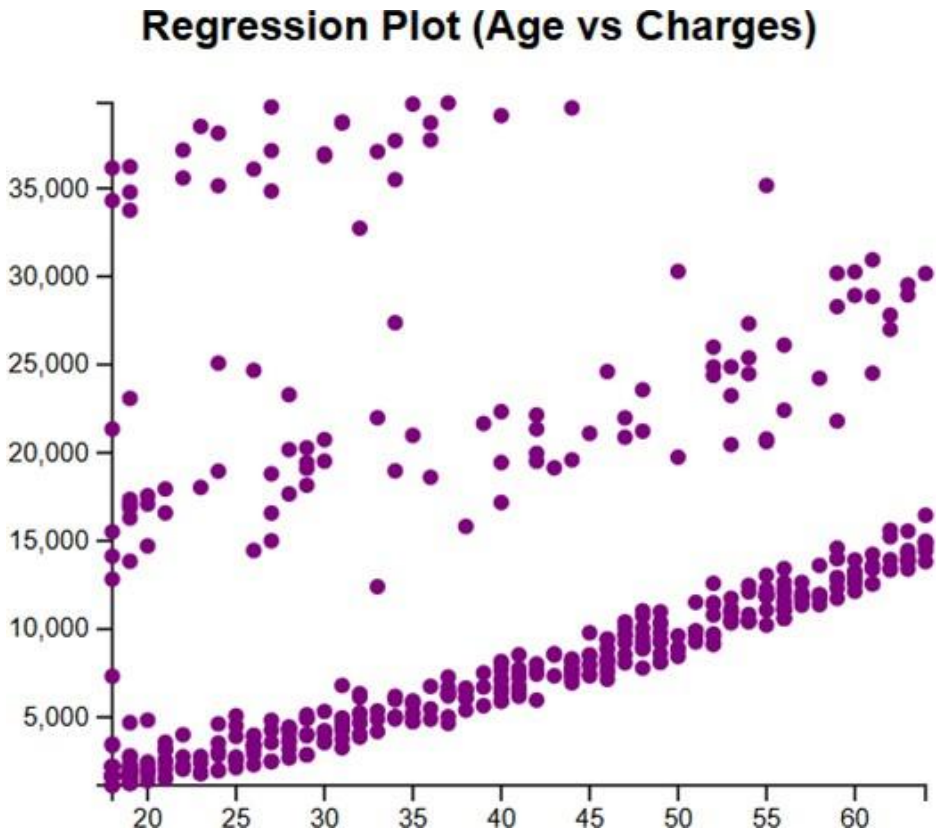
Box Plot (Charges)



The box plot summarizes the distribution of insurance charges. It shows the median, quartiles, and potential outliers in the charges data

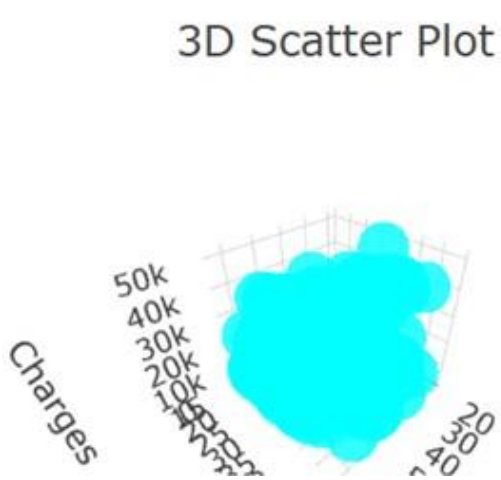


The violin plot provides a detailed view of the distribution of charges, combining the box plot's summary statistics with a density estimation. This helps visualize where the bulk of the charges lie and whether there are multiple modes in the data.



This scatter plot with a regression line illustrates the relationship between age and insurance charges. The slope of the regression line indicates how charges increase with age.

3D Scatter Plot (Age, BMI, Charges)



The 3D scatter plot allows for the visualization of how age, Body Mass Index (BMI), and insurance charges interact. Each point represents an individual in the dataset, with three dimensions of data being plotted.

The jitter plot adds a slight random displacement to points in a scatter plot to prevent overlap, thus visualizing the distribution of charges more clearly. It can reveal clustering of points and density in specific areas of charge values.

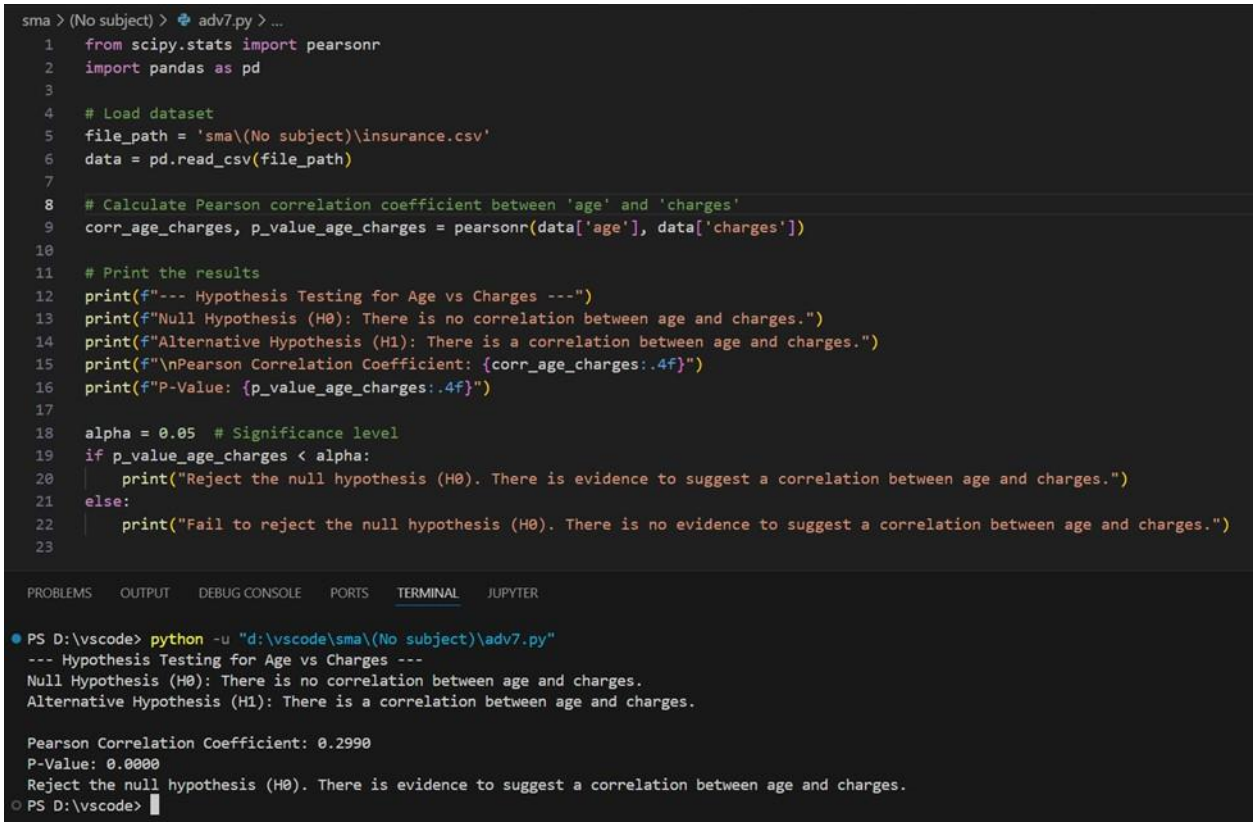
Hypothesis Testing

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 AND OUTPUT:



```
sma > (No subject) > adv7.py > ...
```

```python
from scipy.stats import pearsonr
import pandas as pd

# Load dataset
file_path = 'sma\(No subject)\insurance.csv'
data = pd.read_csv(file_path)

# Calculate Pearson correlation coefficient between 'age' and 'charges'
corr_age_charges, p_value_age_charges = pearsonr(data['age'], data['charges'])

# Print the results
print(f"--- Hypothesis Testing for Age vs Charges ---")
print(f"Null Hypothesis (H0): There is no correlation between age and charges.")
print(f"Alternative Hypothesis (H1): There is a correlation between age and charges.")
print(f"\nPearson Correlation Coefficient: {corr_age_charges:.4f}")
print(f"P-Value: {p_value_age_charges:.4f}")

alpha = 0.05  # Significance level
if p_value_age_charges < alpha:
    print("Reject the null hypothesis (H0). There is evidence to suggest a correlation between age and charges.")
else:
    print("Fail to reject the null hypothesis (H0). There is no evidence to suggest a correlation between age and charges.")

```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  PORTS  TERMINAL  JUPYTER

PS D:\vscode> python -u "d:\vscode\sma\(No subject)\adv7.py"
--- Hypothesis Testing for Age vs Charges ---
Null Hypothesis (H0): There is no correlation between age and charges.
Alternative Hypothesis (H1): There is a correlation between age and charges.

Pearson Correlation Coefficient: 0.2990
P-Value: 0.0000
Reject the null hypothesis (H0). There is evidence to suggest a correlation between age and charges.
PS D:\vscode>
```

Conclusion: Using the health insurance dataset, visualizations created with D3.js revealed significant insights. These interactive and dynamic visualizations, such as regression plots and custom box plots, facilitated deeper exploration of data patterns and trends. These tools are invaluable for understanding complex relationships within the dataset and enhancing data analysis capabilities.