BHARATIYA VIDYA BHAVAN'S

SARDAR PATEL INSTITUTE OF TECHNOLOGY

Advance Data Visualization

UID	2021600051
Name	Dhruvil Patel
Batch	Batch L
Aim	Experiment Design for Creating Visualizations using D3.js on a Finance Dataset

Overview:

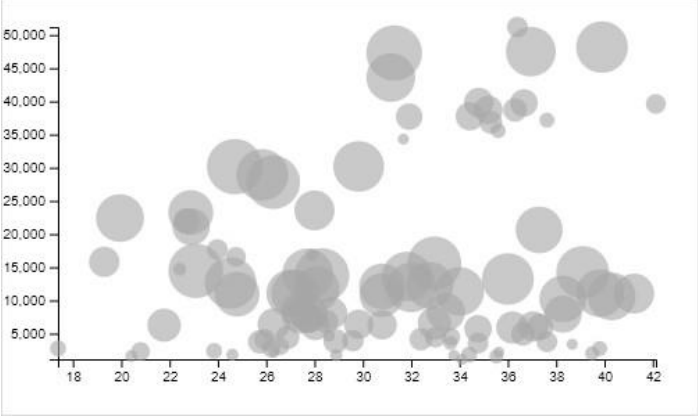
This dataset contains information on 1,338 individuals and their health insurance charges. The data includes various attributes such as age, gender, body mass index (BMI), number of dependents, smoking status, and region. These attributes are a combination of numerical and categorical variables.

Column Attributes:

- **Age:** The age of the insured individual in years.
- **Sex:** The gender of the insured individual (male or female).
- **BMI:** The individual's body mass index, a measure of body fat.
- **Children:** The number of children or dependents covered by the insurance.
- **Smoker:** Indicates whether the individual is a smoker or non-smoker.
- **Region:** The geographic region in the United States where the individual resides.
- **Charges:** The total cost of the medical insurance coverage.

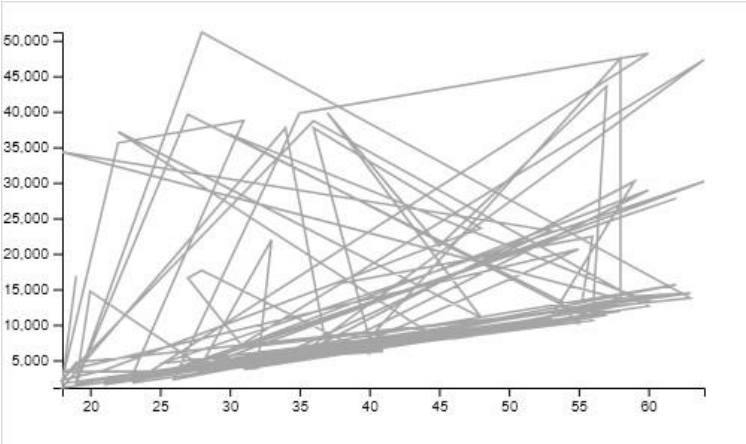
Basic Plots:

Bubble plot (bmi,age and charges)



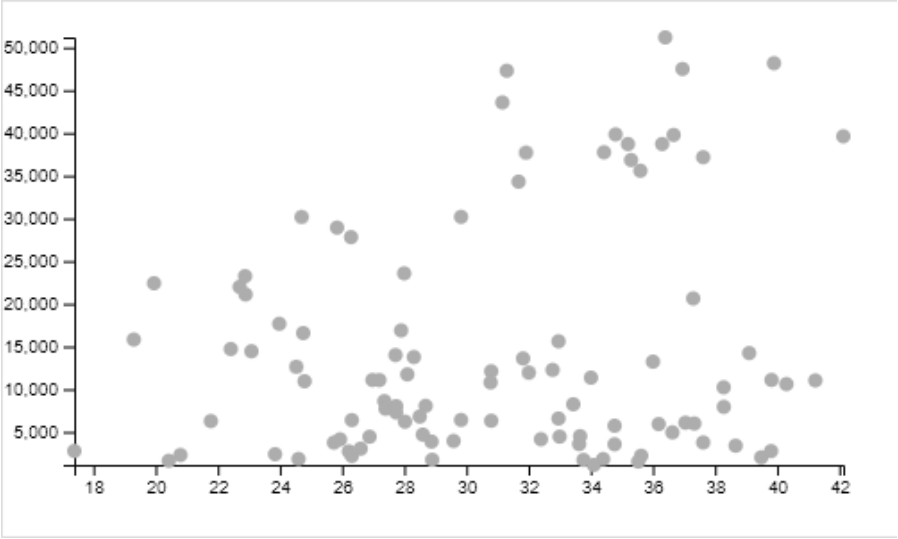
"The visualization reveals distinct clusters of larger bubbles (indicating higher BMI) at different age points, suggesting a correlation between higher BMI and increased insurance charges."

Timeline chart (charges vs age)



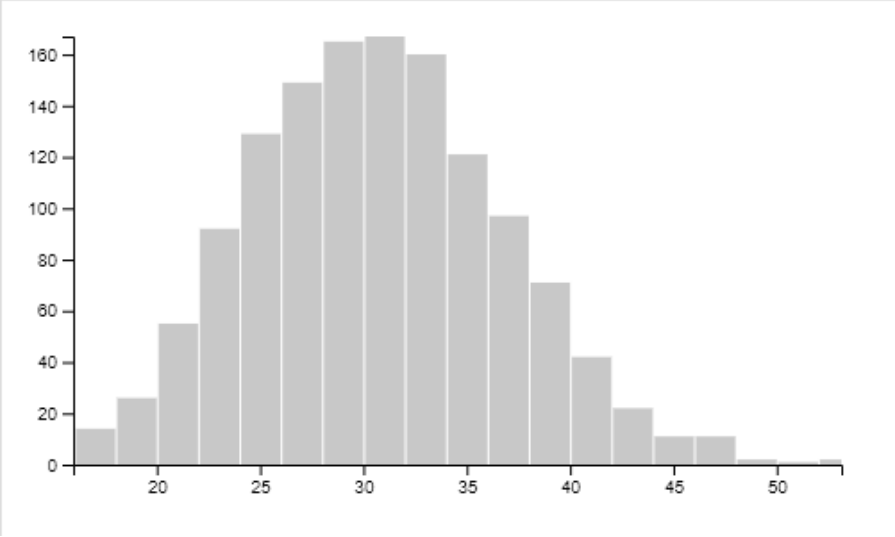
"The overall trend suggests a positive correlation between age and insurance charges, although there are some variations."

Scatter Plot (BMI vs charges)



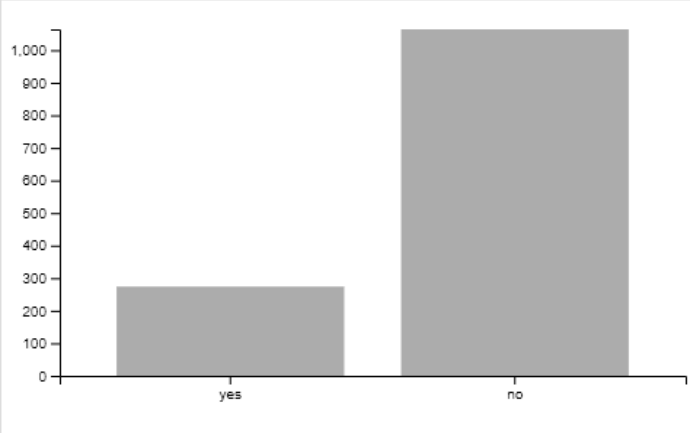
The analysis reveals a clear positive correlation between BMI and insurance charges, indicating that individuals with higher BMIs tend to have higher insurance costs

Histogram (BMI distribution)



The BMI distribution in the dataset exhibits a peak around 30, suggesting that a significant portion of the individuals have a BMI close to this value

Barchart (smokers vs non-smokers)

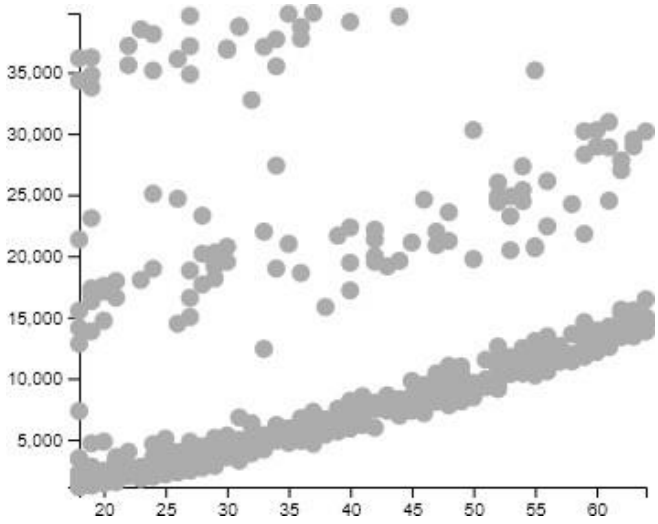


The dataset reveals a significant prevalence of non-smokers, with the "no" category outnumbering the "yes" category by a considerable margin

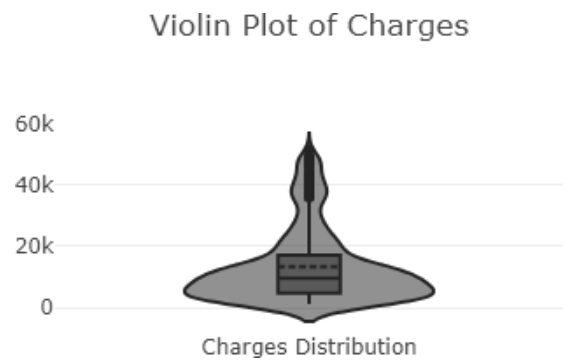
Pie chart(smokers vs non - smokers)

Advanced Visualizations:

regression(age vs charges)

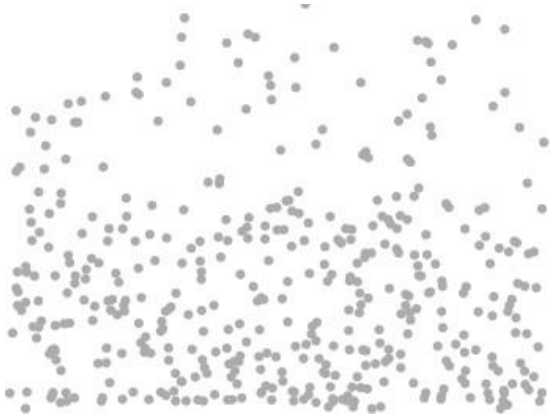


Violin Plot(charges distribution)



Box Plot(charges)

Jitter Plot(charges)



CODE:

```
from scipy.stats import pearsonr
import pandas as pd
# Load dataset
file_path = '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.")
```

OUTPUT:

```
--- 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.
```

Conclusion: Our analysis of the health insurance dataset using D3.js demonstrates its versatility in creating interactive and dynamic visualizations. By leveraging D3.js's capabilities, we were able to effectively explore data patterns and trends through regression plots and custom box plots. These visualizations provided valuable insights into the relationships between various attributes in the dataset, enhancing our understanding of the factors influencing health insurance charges.