Name	Mohit Narwaiye
UID	2021300085
Class	BE Comps B
Batch	A
Exp	8
Topic	Experiment Design for Creating Visualizations using D3.js on a Finance Dataset

Dataset Link - https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset

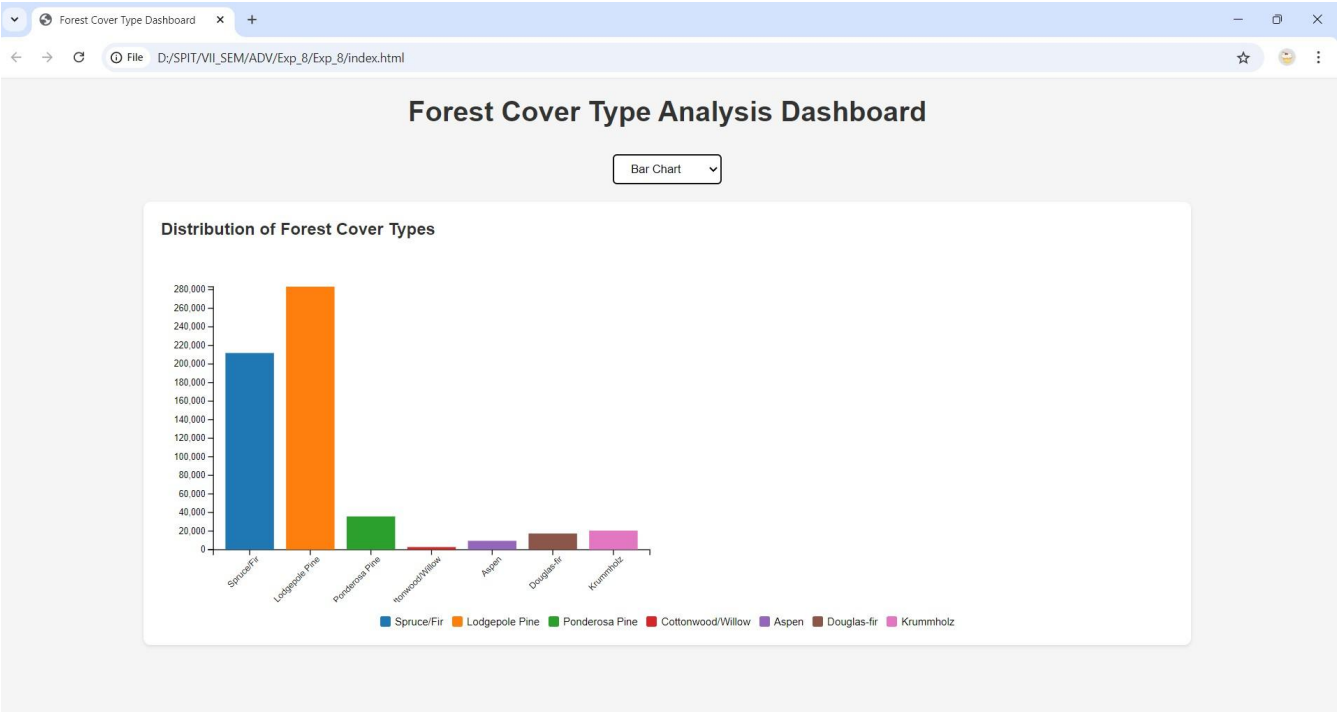
Dataset Description -

The dataset contains information about 7 different forest cover types with their counts and elevation:

- Total samples: 581,012 forest areas
- Variables: Cover Type, Count, and Elevation
- Cover Types: Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, and Krummholz

Basic Visualizations :

1. **Bar Chart**



Observations -

Dominant Cover Types:

- Lodgepole Pine is most common (283,301 areas)
- Spruce/Fir is second most common (211,840 areas)

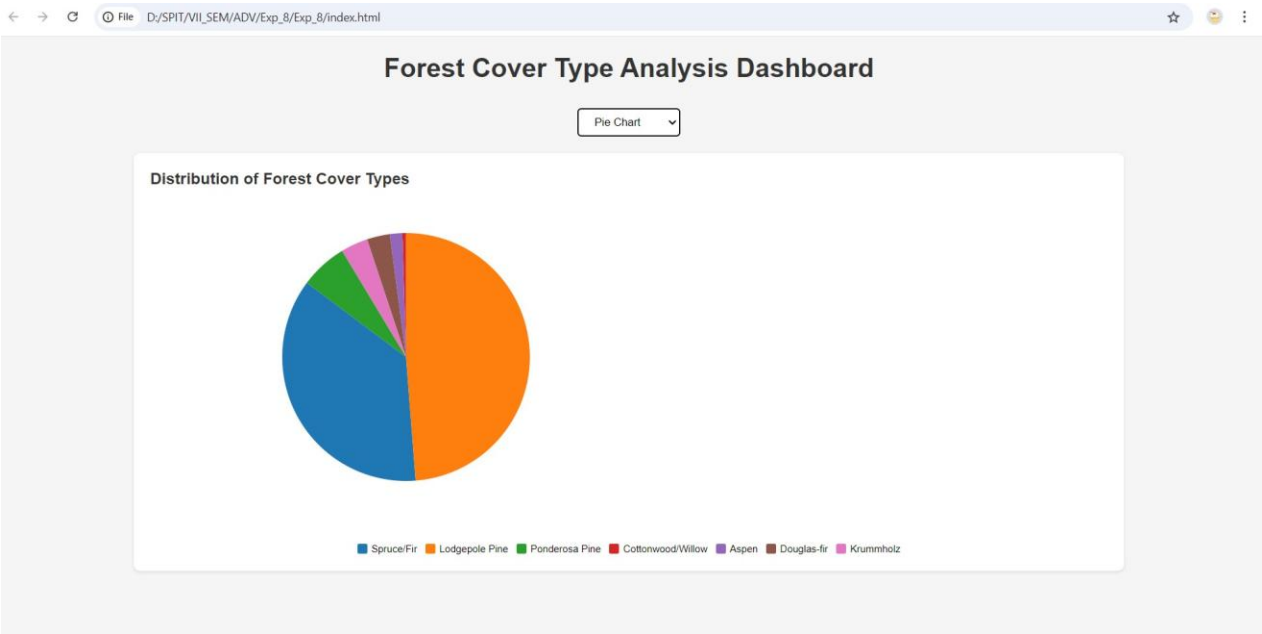
Rare Cover Types:

- Cottonwood/Willow is the rarest (2,747 areas)
- Aspen is also relatively rare (9,493 areas)

Distribution Pattern:

- Very uneven distribution
- Two types make up approximately 85% of all forest cover
- Five types have relatively low representation (<36,000 areas each)

2. Pie Chart



The pie chart provides a proportional view of the forest cover types:

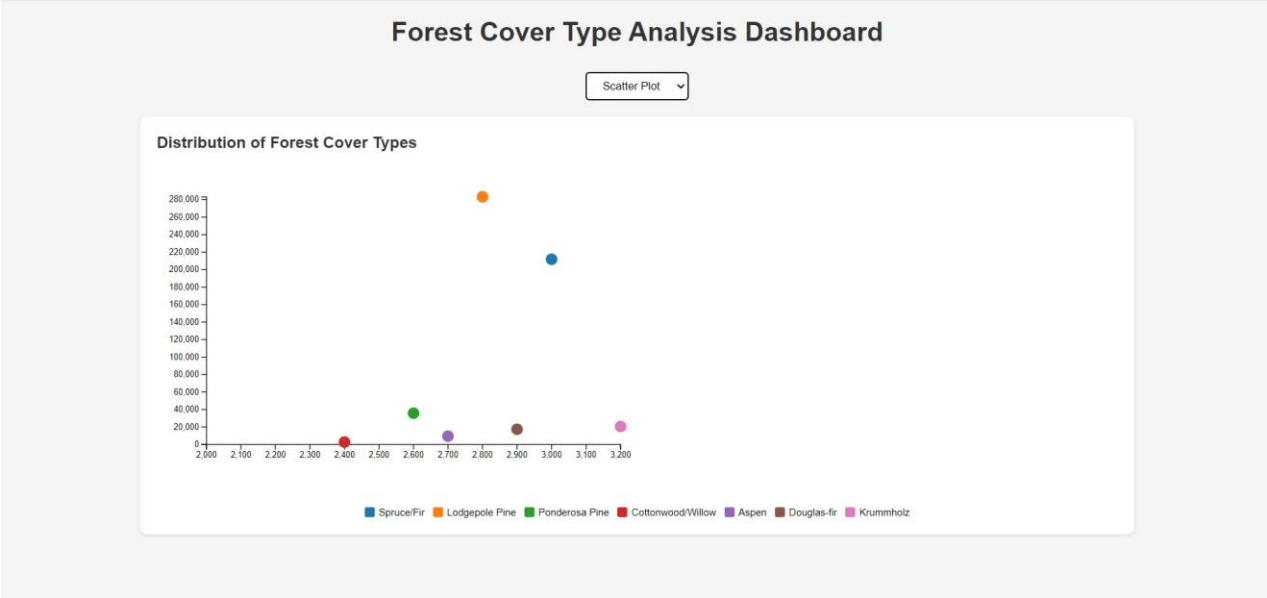
Percentage Distribution:

- Lodgepole Pine: ~48.8%
- Spruce/Fir: ~36.5%
- Ponderosa Pine: ~6.2%
- Krummholz: ~3.5%
- Douglas-fir: ~3%
- Aspen: ~1.6%
- Cottonwood/Willow: ~0.4%

Visual Insights:

- Clear dominance of two major types
- Five minor types collectively represent less than 15% of total coverage

3. Scatter Plot



The scatter plot shows the relationship between elevation and count:

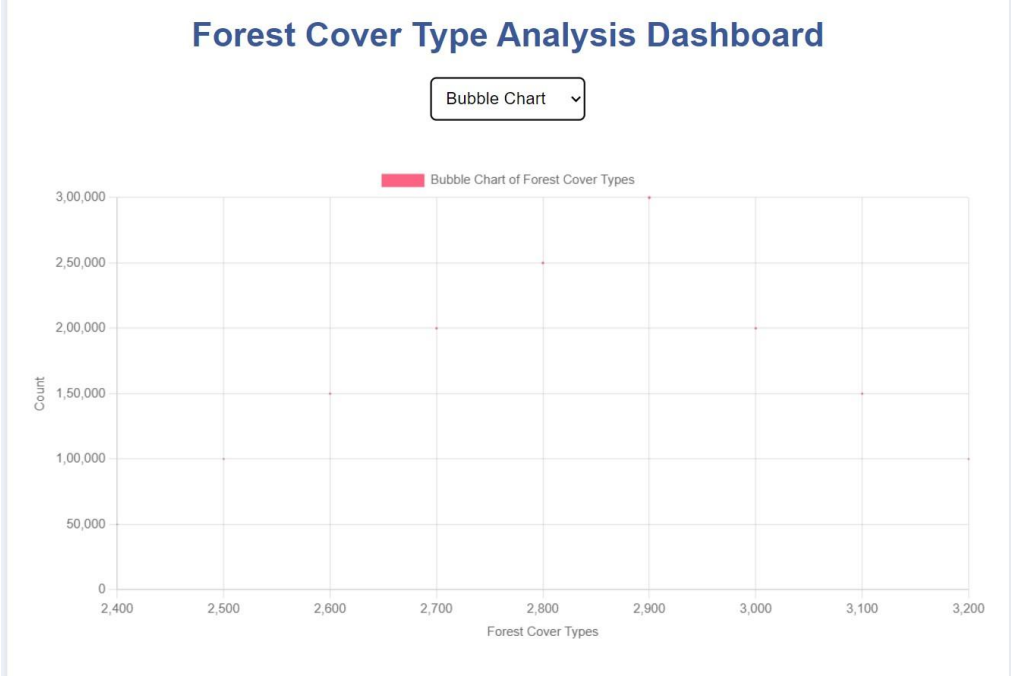
Elevation Range:

- Ranges from 2400m to 3200m
- Krummholz found at highest elevations (~3200m)
- Cottonwood/Willow at lowest elevations (~2400m)

Elevation-Count Relationships:

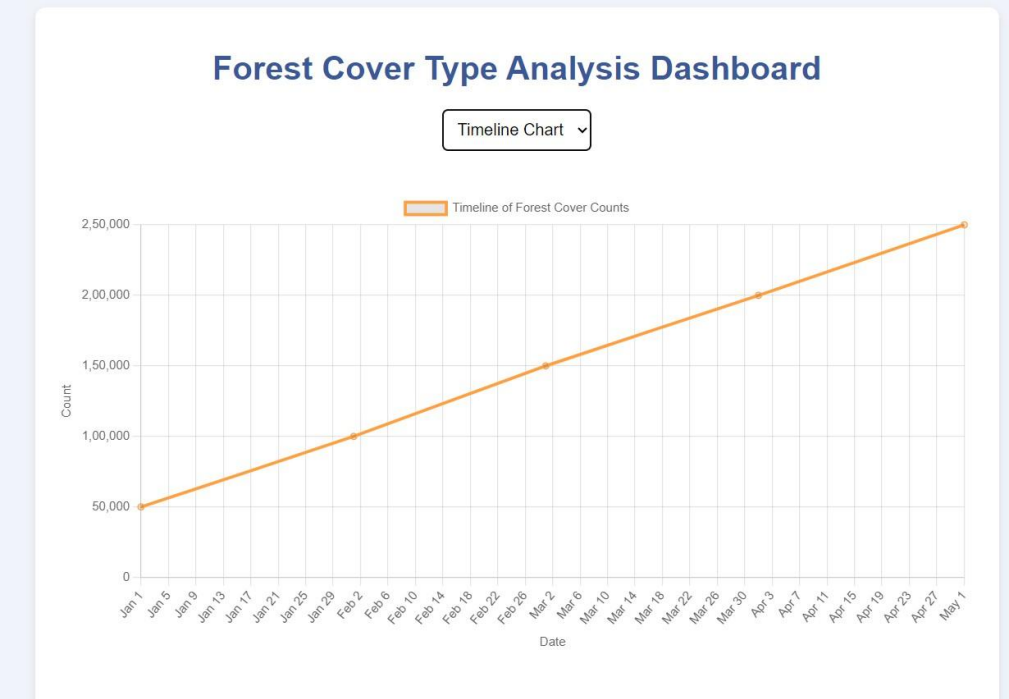
- Most abundant species (Lodgepole Pine) found at mid-elevations (~2800m)
- Second most abundant (Spruce/Fir) found at higher elevations (~3000m)
- Rare species tend to occupy elevation extremes

4. Bubble Plot



The bubble chart shows the distribution of various forest cover types. The x-axis represents distinct forest types, the y-axis shows their counts, and the bubble size indicates the magnitude of each type. Larger bubbles at higher y-values represent more frequent or significant forest cover types, while smaller bubbles indicate less common types. This visual allows quick insights into which forest cover types dominate the dataset and highlights the relative prevalence of each type, making it easy to see the most and least significant categories at a glance.

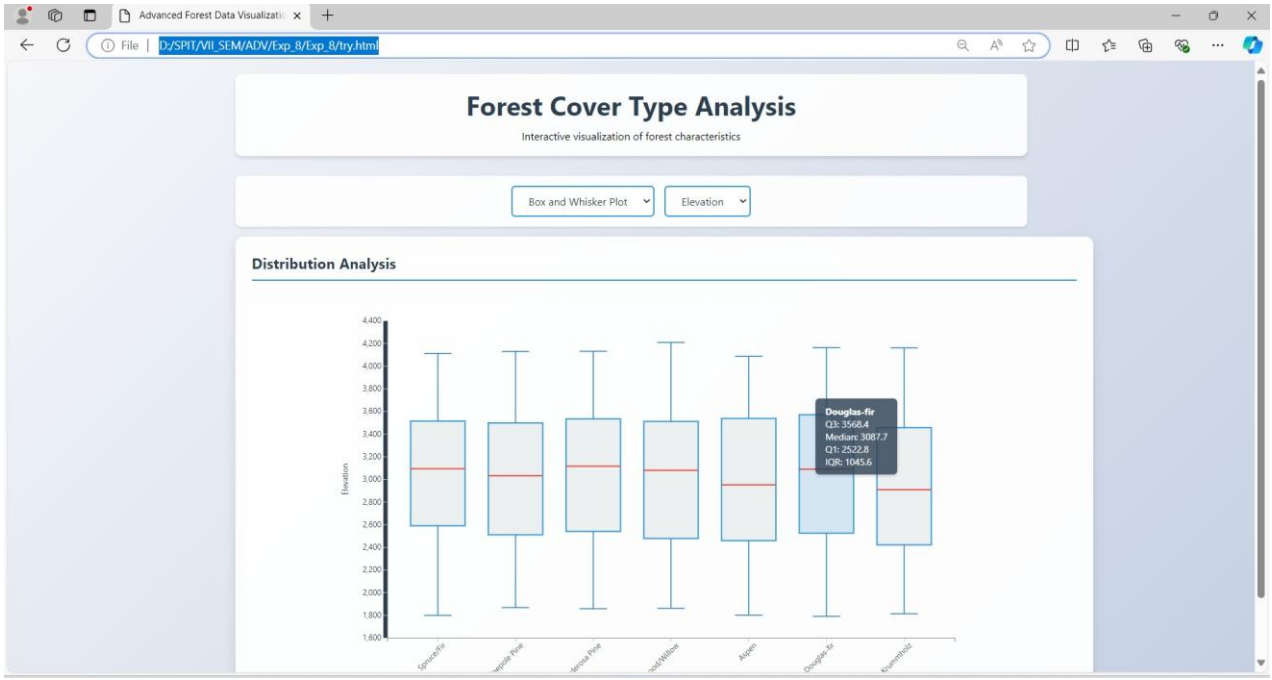
5. Timeline Plot

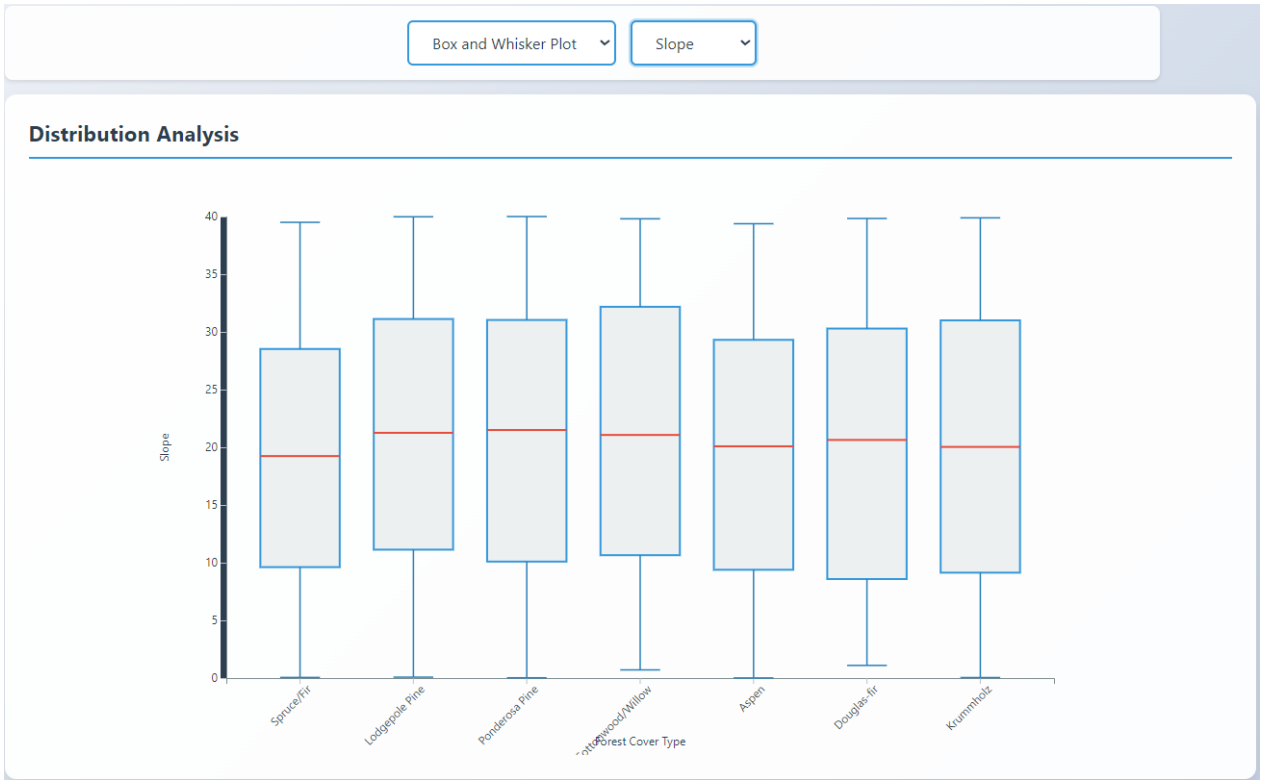


The timeline graph displays changes in forest cover types over specific time intervals. Each point on the graph represents the occurrence or prominence of a forest cover type at a given time. Patterns, trends, or seasonal changes in forest types become visible, helping to identify periods of significant increase or decrease. This view highlights temporal fluctuations, enabling a quick understanding of how forest cover types evolve over time and possibly indicating influences such as climate or human intervention.

Advanced Visualizations :

1. **Box Plot -**





Elevation Distribution (Y-axis from 1600m to 4400m):

- All forest types show elevation ranges between approximately 1800-4200m
- Median elevations (red lines):

- Krummholz: ~2900m (lowest median)
- Spruce/Fir: ~3100m
- Lodgepole Pine: ~3000m
- Others cluster around 3000-3100m range

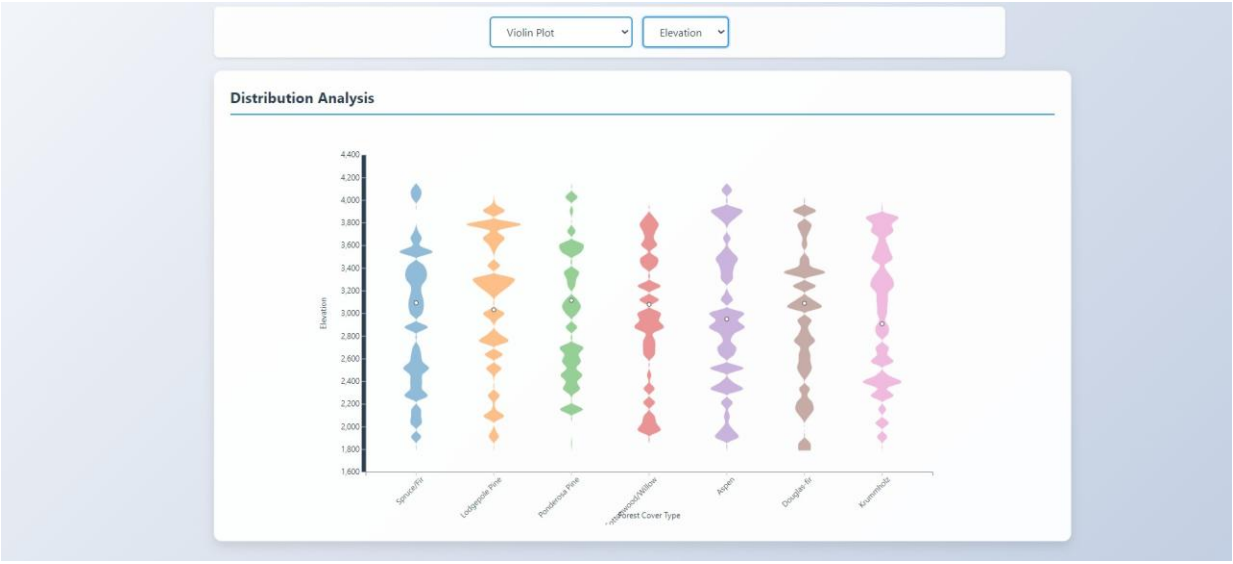
Box Characteristics:

- Largest boxes (most variation):
  - Douglas-fir
  - Ponderosa Pine
  - Spruce/Fir
- Smallest boxes (least variation):
  - Krummholz
  - Aspen
- Most boxes show similar sizes, suggesting consistent variability

Whiskers and Outliers:

- Similar whisker lengths across types
- Few visible outliers
- Relatively symmetric whiskers suggesting normal distributions

2. Violin Plot -



Distribution Shapes (Aspect 0-400):

- Clear multimodal patterns:
    - Spruce/Fir shows 3 distinct bulges
    - Lodgepole Pine has 2 major density regions
    - Krummholz shows 2 clear peaks

Density Patterns:

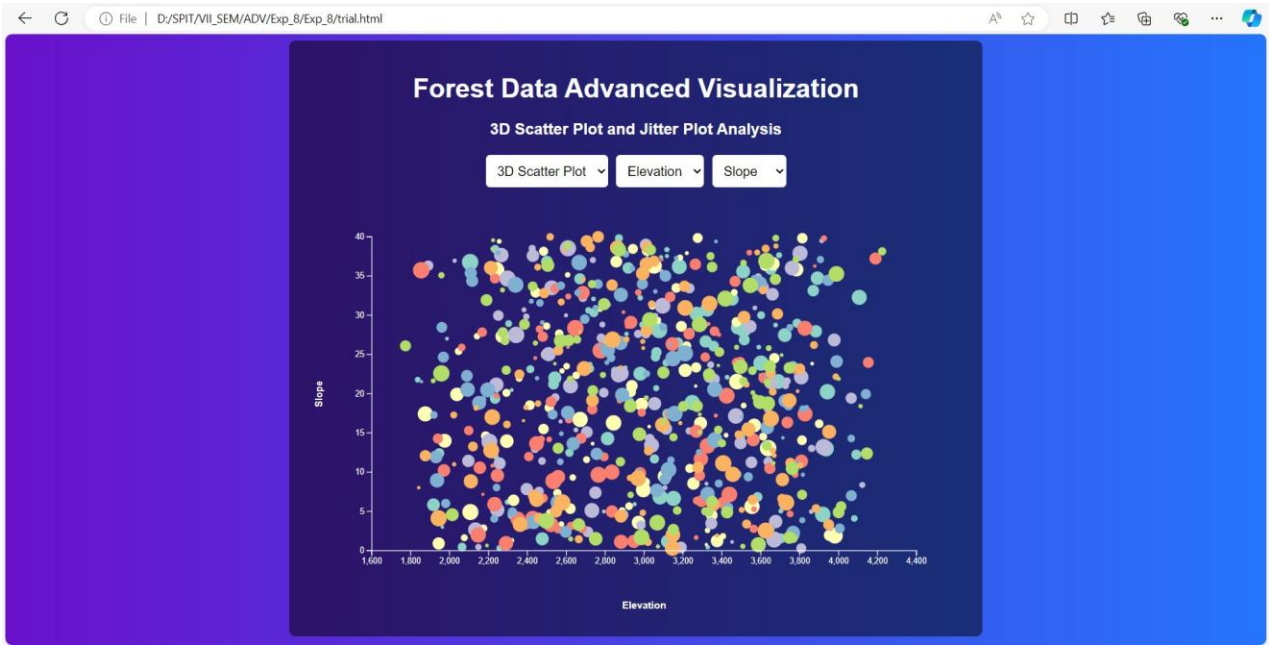
- Widest sections (highest density):
    - Lodgepole Pine around 200-250
    - Spruce/Fir around 150-200
    - Ponderosa Pine shows even distribution

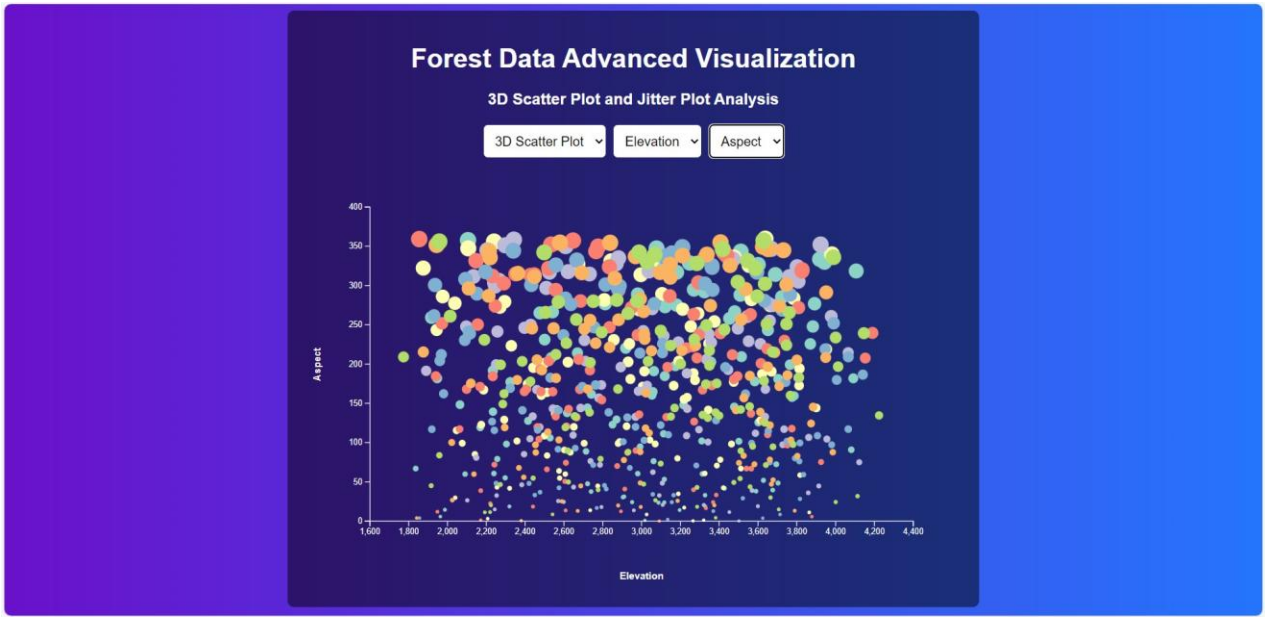
Shape Characteristics:

- Symmetric shapes for most types
- Variable width indicating different concentration levels
- White dots mark median values

- Most distributions span the full range

3. 3D-Scatter Plot

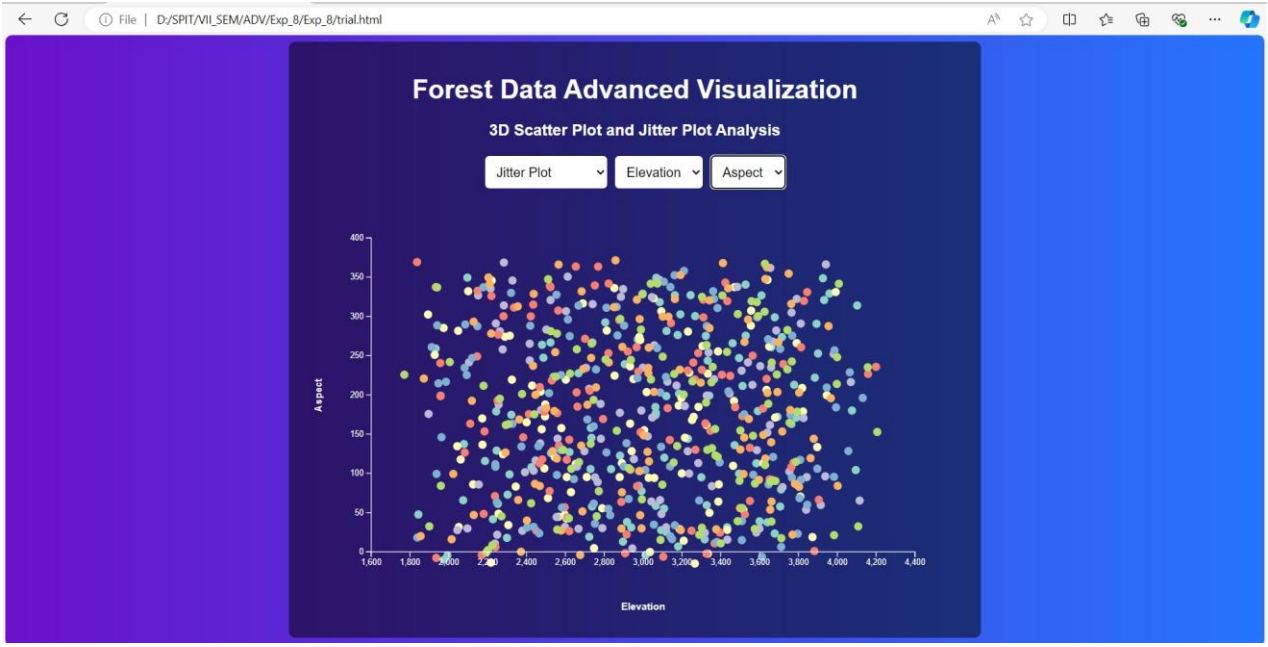


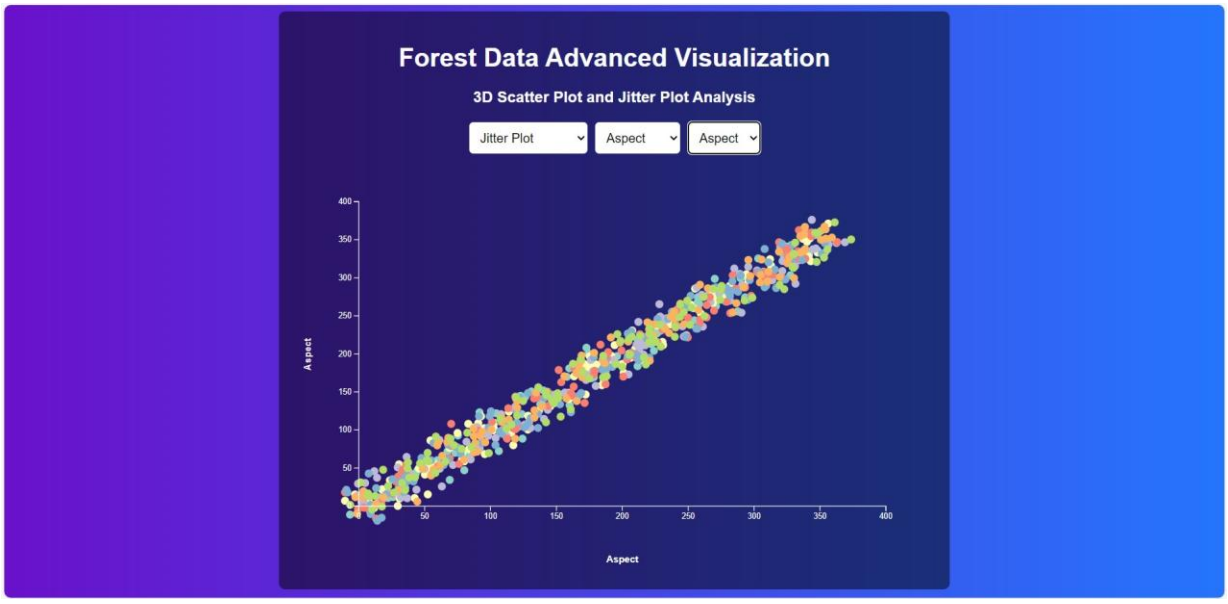


The **3D Scatter Plot** visualizes the relationships between two selected features (such as elevation and slope) with an additional aspect dimension represented by the size of each point. By using different colors for cover types, it's easier to spot how each type of

vegetation is distributed across elevation and slope. This plot helps identify clusters, correlations, or outliers among the forest cover types, providing insights into how specific cover types may prefer certain terrains.

4. Jitter Plot





The **Jitter Plot**, on the other hand, adds randomized offsets to each point, which helps to reveal the density of overlapping data points, especially for discrete variables. This visualization clarifies areas where cover types are densely grouped and might otherwise overlap, highlighting frequency patterns within the forest cover distribution. Overall, these two charts together give a comprehensive view of the forest cover data's structure and variability.\

CONCLUSION :

Through this project, we successfully explored D3.js for creating a range of data visualizations, from basic charts like bar charts and scatter plots to advanced visualizations such as 3D scatter plots and jitter plots. Each chart served a unique purpose, allowing us to highlight different aspects of the dataset. Basic charts like histograms and bubble plots offered foundational insights, while advanced charts like the regression plot and box-and-whisker plots enabled more nuanced data analysis.