1. Census Dataset

Relationachip Chart

Relationship

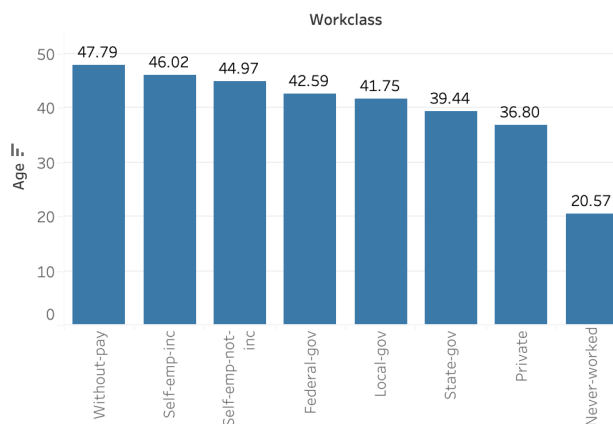| Relationship	Count of adult
Other-relative	981
Wife	1,568
Unmarried	3,446
Own-child	5,068
Not-in-family	8,305
Husband	13,193
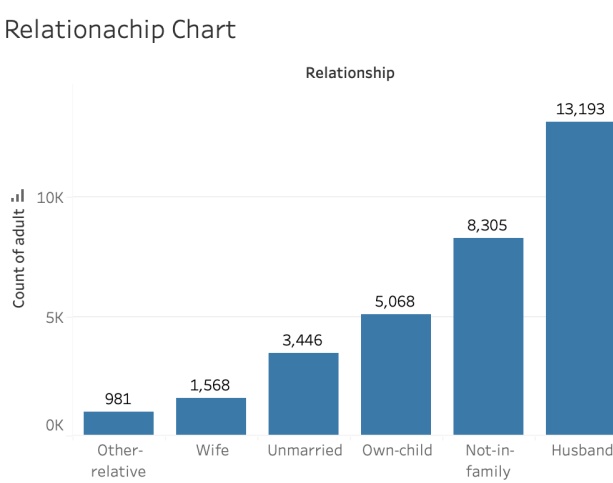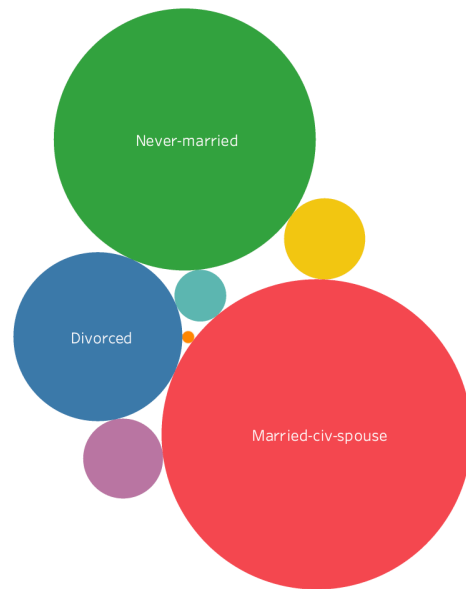
Average Age against Workclass

Workclass

Workclass	Age
Without-pay	47.79
Self-emp-inc	46.02
Self-emp-not-inc	44.97
Federal-gov	42.59
Local-gov	41.75
State-gov	39.44
Private	36.80
Never-worked	20.57

Marital-Status

- Divorced
- Married-AF-sp..
- Married-civ-sp..
- Married-spous..
- Never-married
- Separated
- Widowed

Marital Status

Never-married, Separated, Divorced, Married-civ-spouse, Widowed
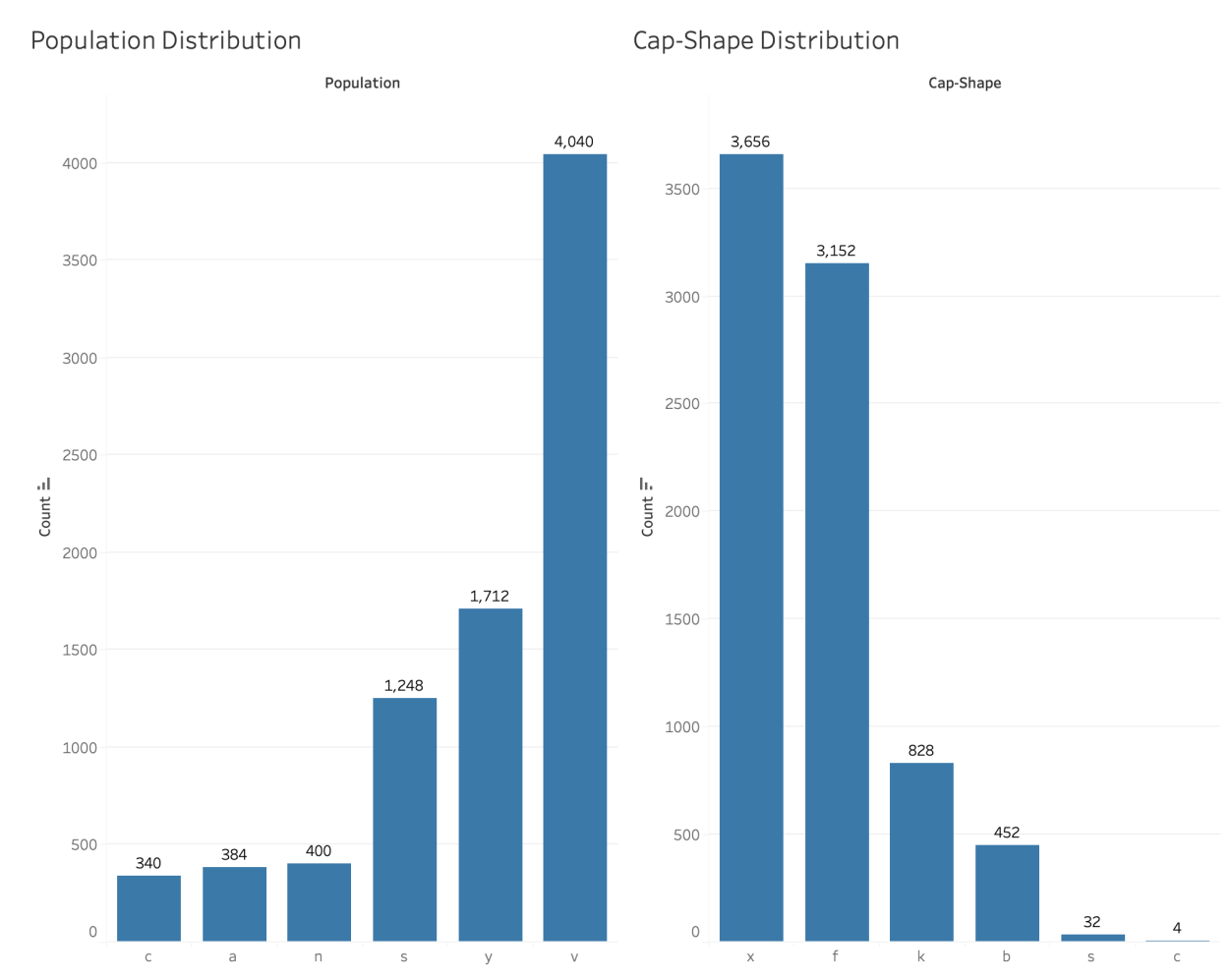
From above visualizations we can see that from Average Age distribution against workclass people with out pay are over the age of 45 and people who never worked are under the age of 25 approximately. And majority of the adults who are working are Husbands.

The census dataset contains demographic information about individuals, including their age, employment status, education level, marital status, occupation, relationships, race, gender, financial gains and losses, hours worked per week, and native country. It is often used for socioeconomic analysis and understanding population characteristics.

2. Mushroom Dataset

Population Distribution

Population

### Cap-Shape Distribution

Cap-Shape

From the above visualization we can see that count of mushroom of several population and Cap-shape with convex are higher in number.

The Mushroom Database is a collection of data that contains details about various types of mushrooms with gills, specifically focusing on 23 species from the Agaricus and Lepiota families. This dataset serves the purpose of categorizing mushrooms into two groups: those that are "edible" (referred to as 'e') and those that are "poisonous" (referred to as 'p').

- Number of Attributes: There are 22 attributes in this dataset, all of which are classified as nominal (categorical) variables.
- Attribute Information: These attributes describe various characteristics of the mushrooms, such as cap shape, cap surface, cap color, presence of bruises, odor, gill attachment, gill spacing, gill size, gill color, stalk shape, stalk root type, stalk surface above and below the ring, stalk color above and below the ring, veil type and color, ring number and type, spore print color, population, and habitat.

- Missing Attribute Values: In attribute #11, which pertains to "stalk-root," there are 2,480 instances where data is missing and represented by the symbol "?".
- Class Distribution: The dataset is divided into two main classes:
    1. "Edible" mushrooms (abbreviated as 'e') constitute 51.8% of the dataset, accounting for 4,208 instances.
    2. "Poisonous" mushrooms (abbreviated as 'p') constitute 48.2% of the dataset, with a total of 3,916 instances.
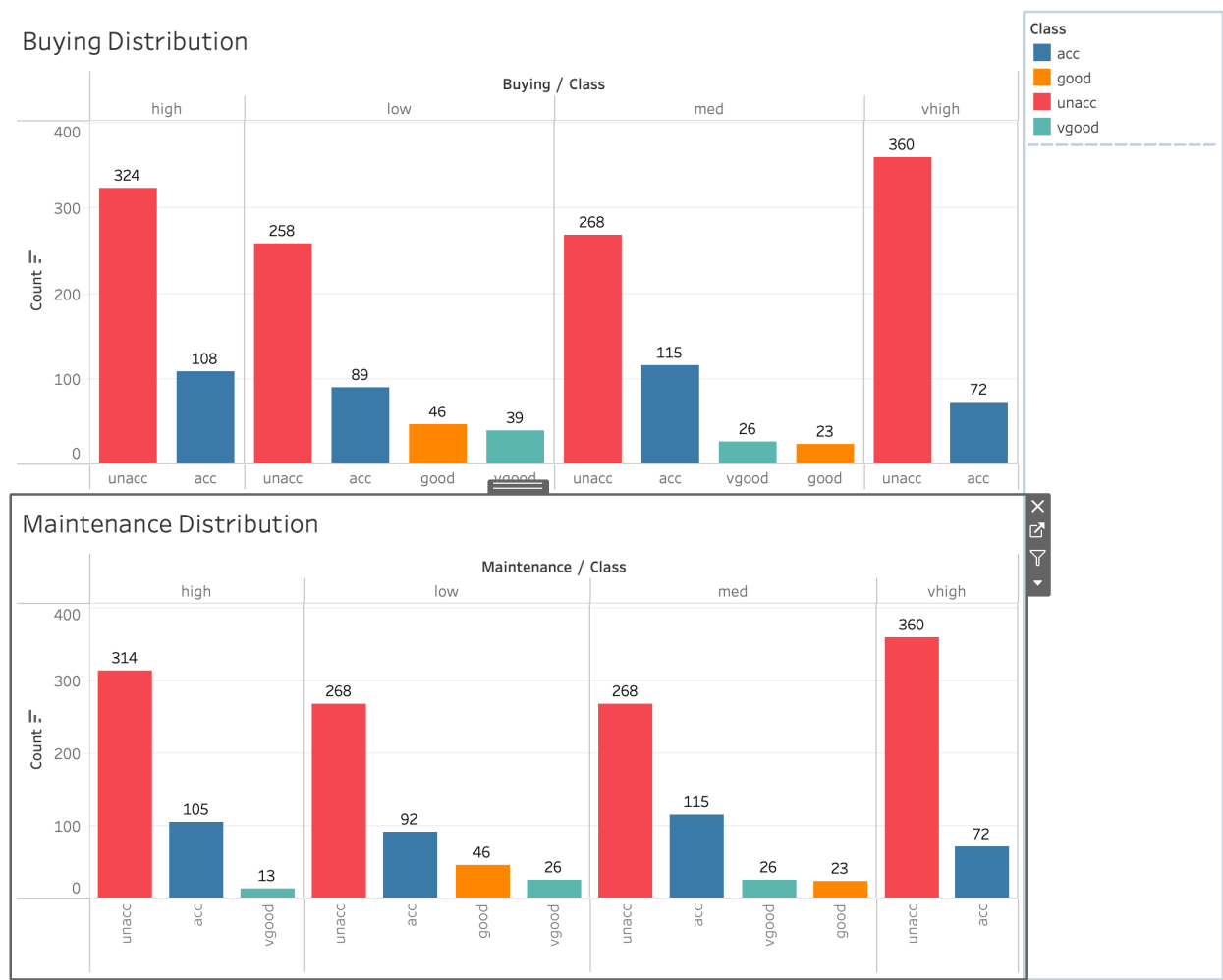
3. Iris Dataset
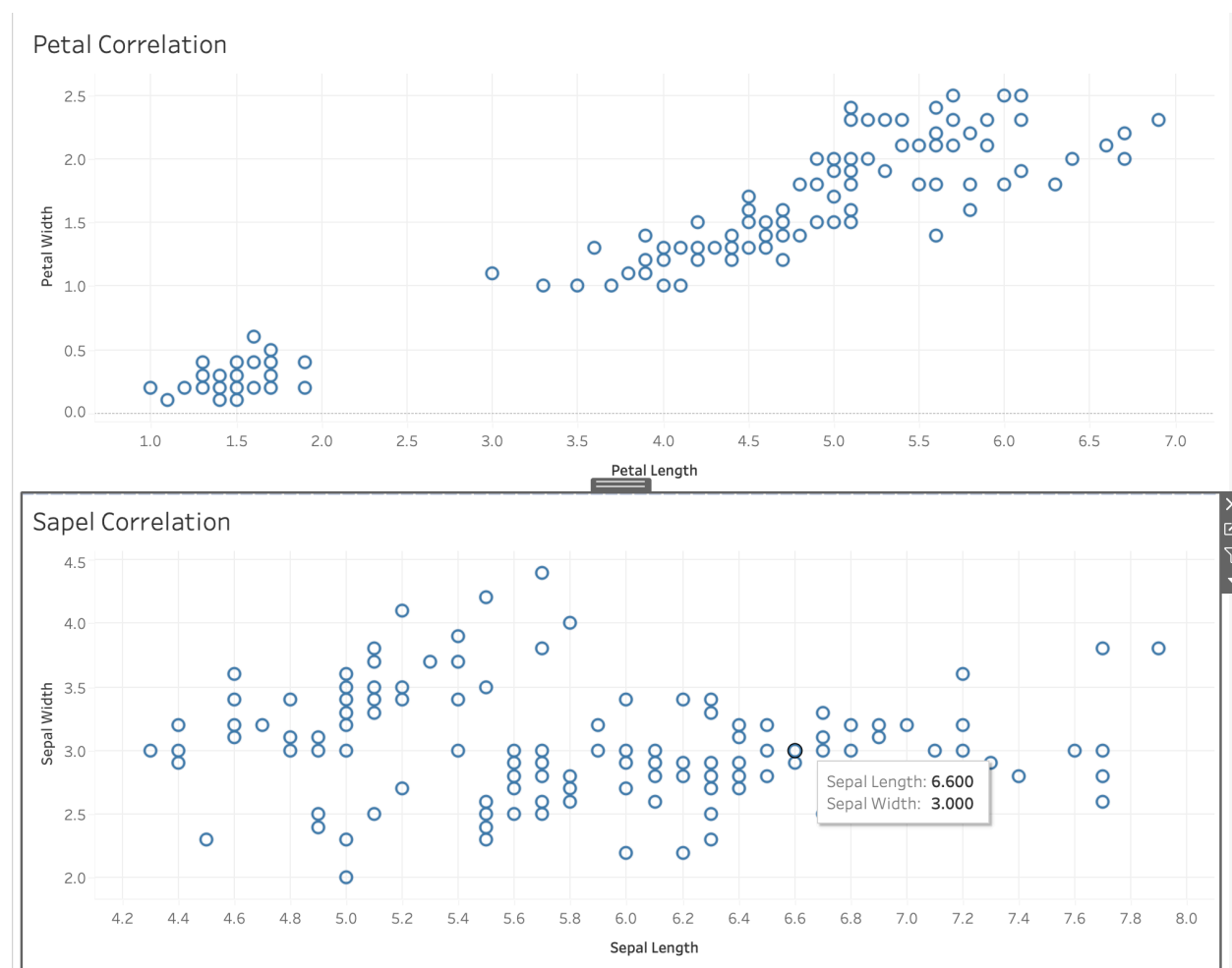
From the scatter plots we can see that there is a high correlation between Petal Width and Petal Length.

The dataset comprises 150 instances, each representing an iris plant categorized into one of three classes: Iris Setosa, Iris Versicolour, and Iris Virginica. It includes four numerical predictive features measured in centimeters, which are sepal length, sepal width, petal length, and petal

width. This dataset is widely employed for classification tasks and is renowned for its simplicity and effectiveness in evaluating classification algorithms. It has no missing data, and the summary statistics offer valuable insights into the distribution and relationships among the attributes within each class. This dataset's historical importance is underscored by its frequent utilization in numerous publications over the years.

4. Car+Evaluation Dataset

The above visualization shows the distribution of Buying and Maintenance features for various classes of car.

This dataset has been created to assess the desirability of automobiles based on six factors related to their cost, upkeep, technical specifications, and safety features. These factors encompass qualitative variables like purchase price, maintenance expenses, the number of doors, passenger capacity, trunk size, and an estimated safety rating.

The dataset is complete, with no missing values, making it suitable for thorough analysis. When examining the distribution of classes within the dataset, it's evident that a significant portion of

the cars are classified as "unacc," indicating that the majority of the vehicles in the dataset are considered unacceptable. Nonetheless, there are also cars classified as "acc" (acceptable), "good," and "v-good."