EXPERIMENT NO.5

# DATE:

**AIM:** To study and analyse ANOVA for IRIS Data

**SOFTWARE USED:** JUPYTER Notebook

# THEORY:

**Data Visualization**

Data Visualization is the presentation of data in pictorial format. It is extremely important for Data Analysis, primarily because of the fantastic ecosystem of data-centric Python packages. It helps to understand the data, however complex it is, by summarizing and presenting a huge amount of data in a simple and easy-to- understand format, thereby communicating information clearly and effectively.

# Purpose

Data visualization aims to communicate insights from data clearly, efficiently, and effectively. It helps in identifying trends, patterns, correlations, and outliers.

# Types of Visualizations

There are various types of visualizations used in data analysis:

* **Charts:** Bar charts, line charts, pie charts, etc., represent data points using various shapes and lines.
* **Graphs:** Graphs like scatter plots, network graphs, and histograms show relationships between variables.
* **Maps:** Geographic data is often visualized using maps to display spatial patterns and distributions.
* **Infographics:** These combine visual elements with text to convey a story or highlight key points.

# Tools

Several tools are available for creating data visualizations, including:

* **Pandas**: Pandas offer tools for cleaning and processing your data. It is the most popular Python library used for data analysis. In pandas, a data table is called a data-frame.
* **Seaborn:** Seaborn is a visualization library for statistical graphics plotting in Python. Built on top of matplotlib and integrated closely with pandas data structures, Seaborn helps visualize statistical relationships and trends in the dataset.

Types of Plots

Various types of plots can be generated using data visualization tools:

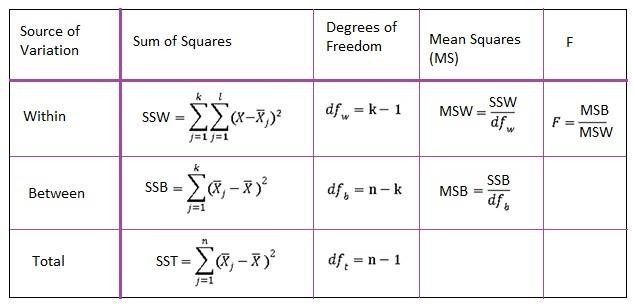
* Line Plot
* Count Plot
* Scatter Plot
* Violin Plot
* Box Plot
* Bar Plot
* Point Plot
* Swarm Plot
* KDE Plot
* Strip Plot

These plots help visualize data relationships, trends, and patterns in the dataset.

**ANOVA** Test is used to compare the means of different groups using various estimate methodologies. ANOVA is an abbreviation for the analysis of variance. The ANOVA analysis is a statistical relevance tool designed to evaluate whether or not the null hypothesis can be rejected while testing hypotheses. It is used to determine whether or not the means of three or more groups are equal. Whenever there are more than two or more independent groups, the ANOVA test is used. The ANOVA test is used to look for heterogeneity within groups as well as variability across groupings.

ANOVA Formula

ANOVA formula is made up of numerous parts. The best way to tackle an ANOVA test problem is to organize the formulae inside an ANOVA table. Below are the ANOVA formulae.

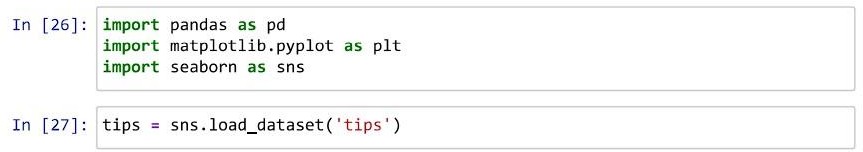


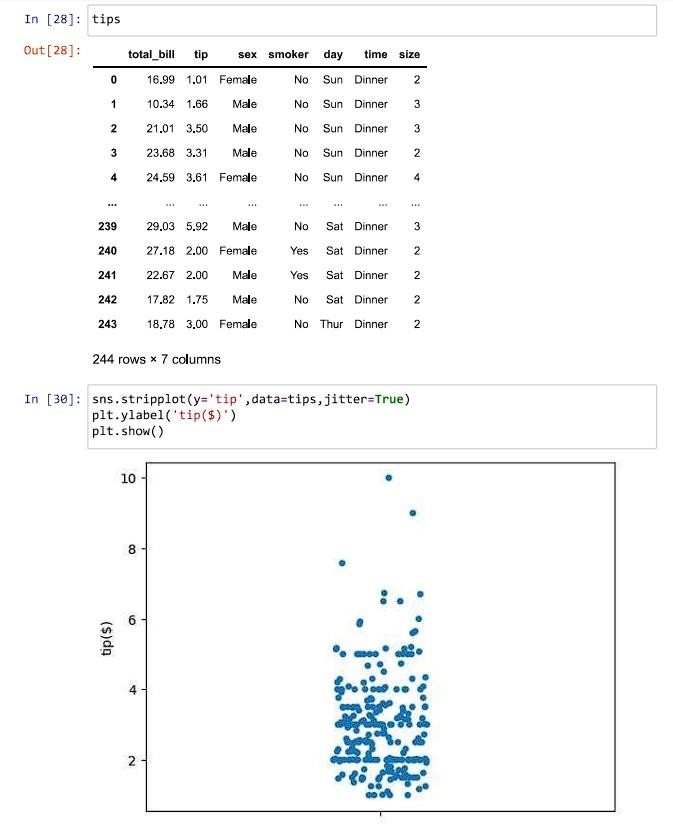
were,

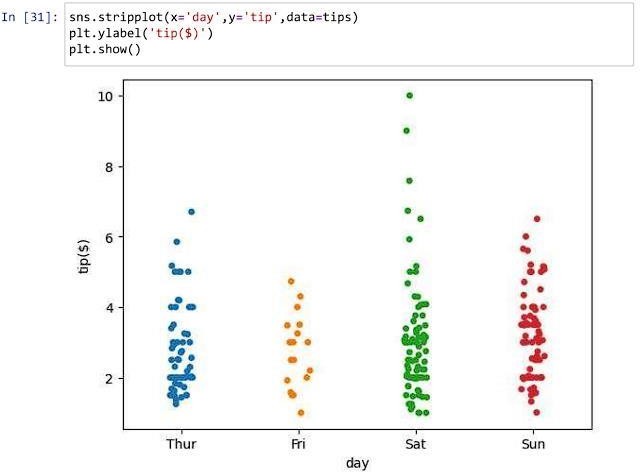
* F = ANOVA Coefficient
* MSB = Mean of the total of squares between groupings
* MSW = Mean total of squares within groupings
* MSE = Mean sum of squares due to error
* SST = total Sum of squares
* p = Total number of populations
* n = The total number of samples in a population
* SSW = Sum of squares within the groups
* SSB = Sum of squares between the groups
* SSE = Sum of squares due to error
* s = Standard deviation of the samples
* N = Total number of observations

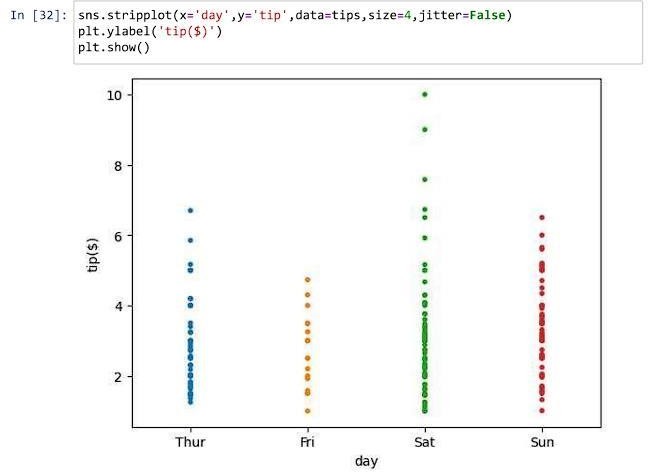
**Iris Dataset** is considered as the Hello World for data science. It contains five columns namely – **Petal Length, Petal Width, Sepal Length, Sepal Width, and Species Type**. Iris is a flowering plant; the researchers have measured various features of the different iris flowers and recorded them digitally.

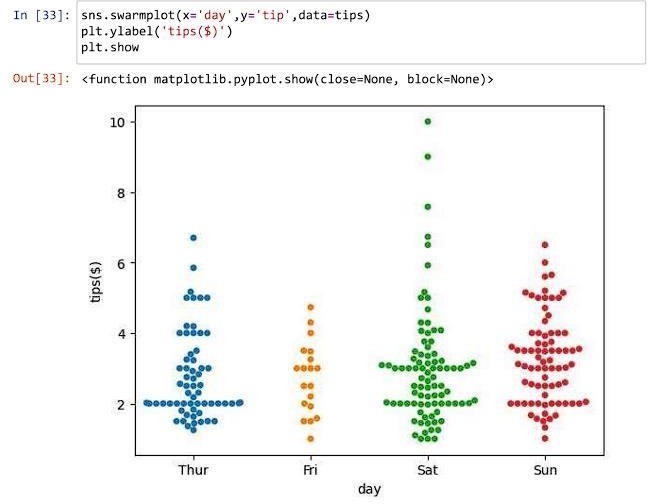
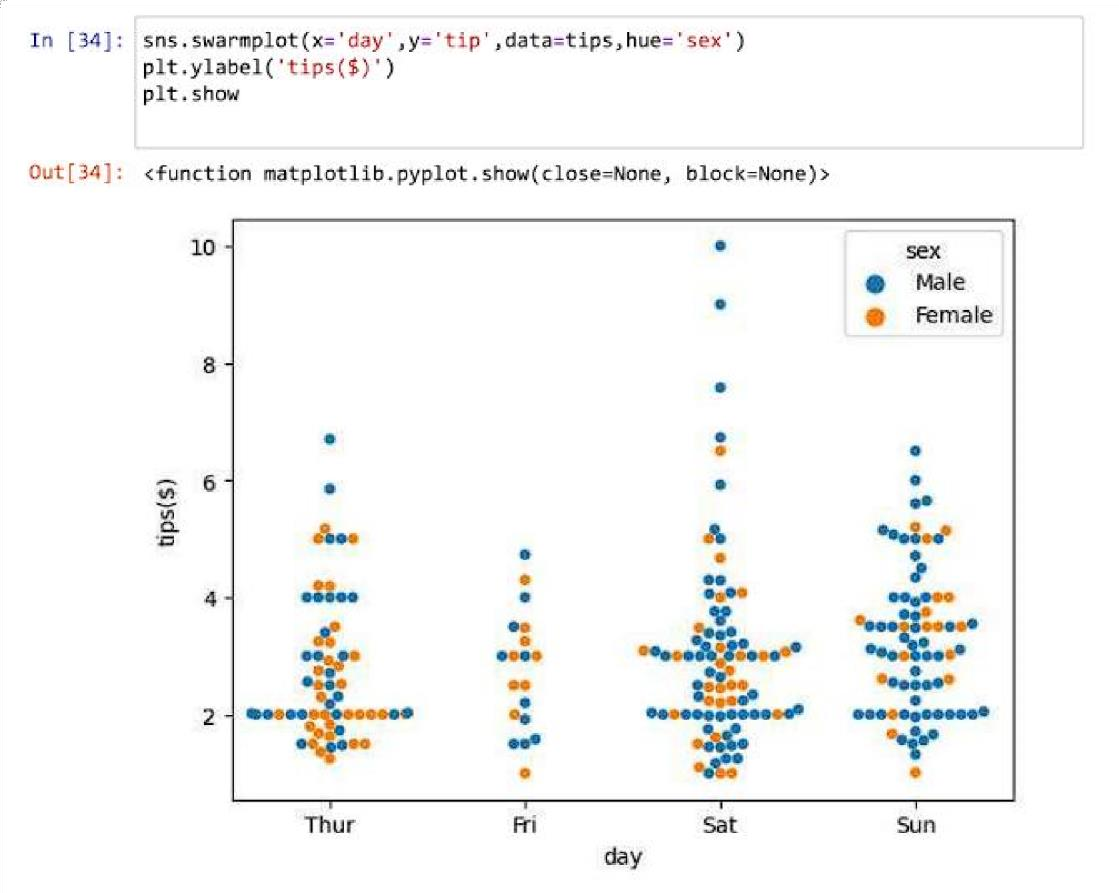
# OUTPUT CODE:

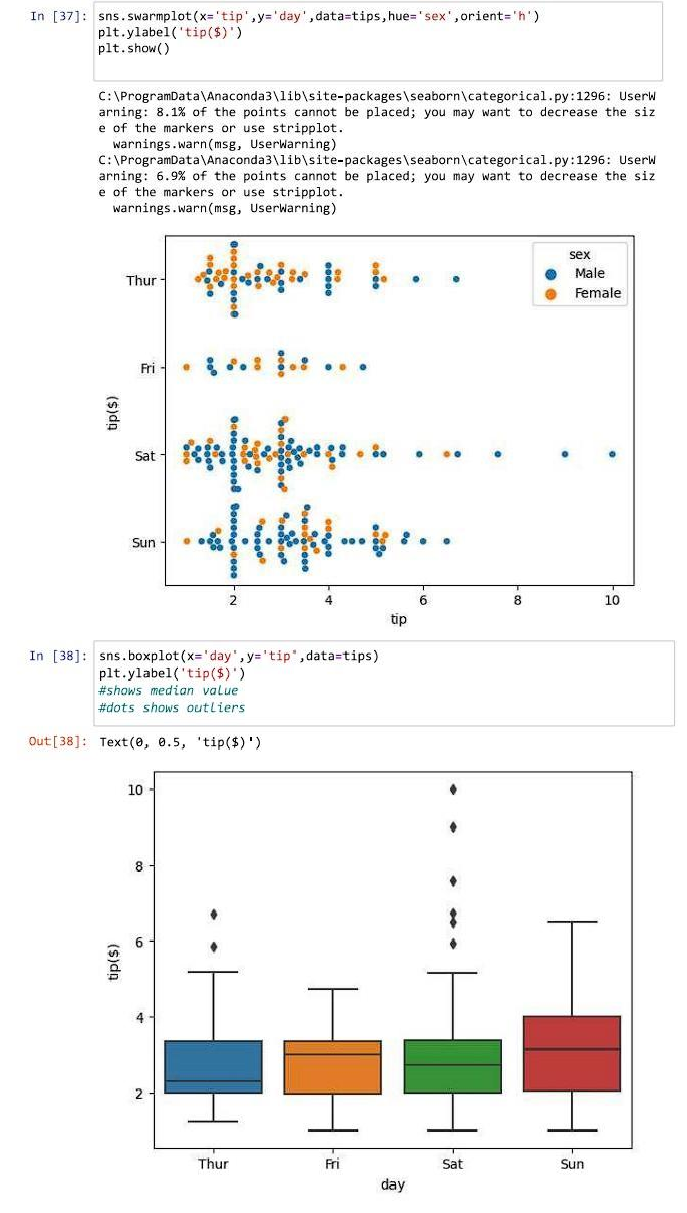


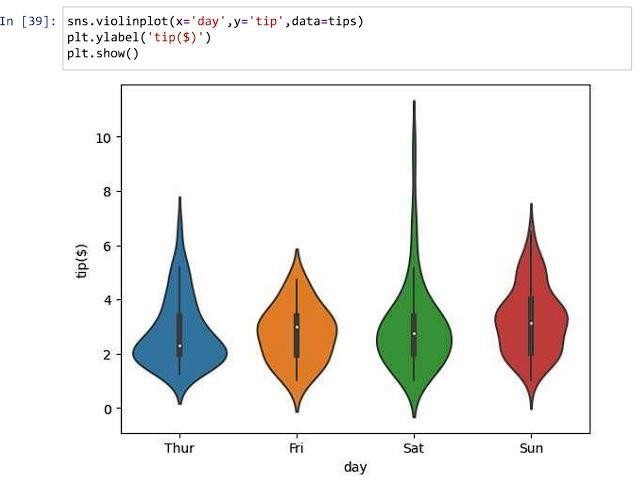
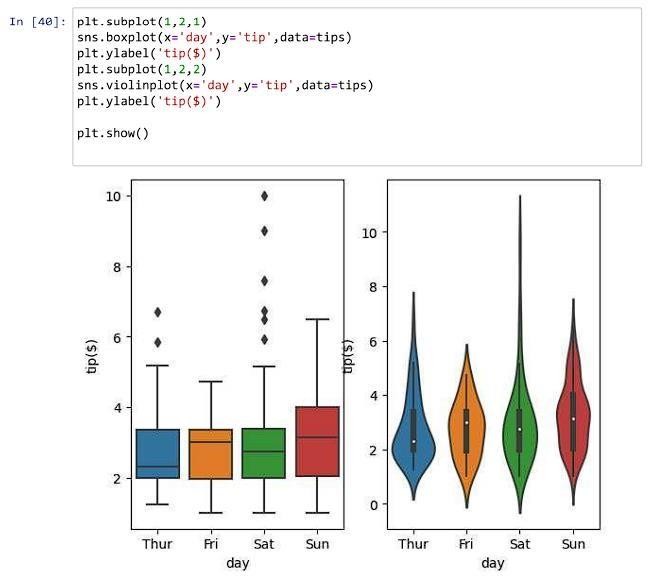


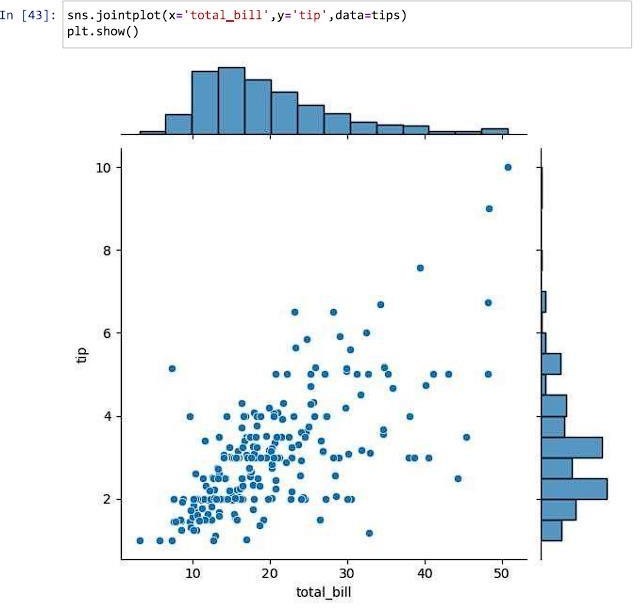


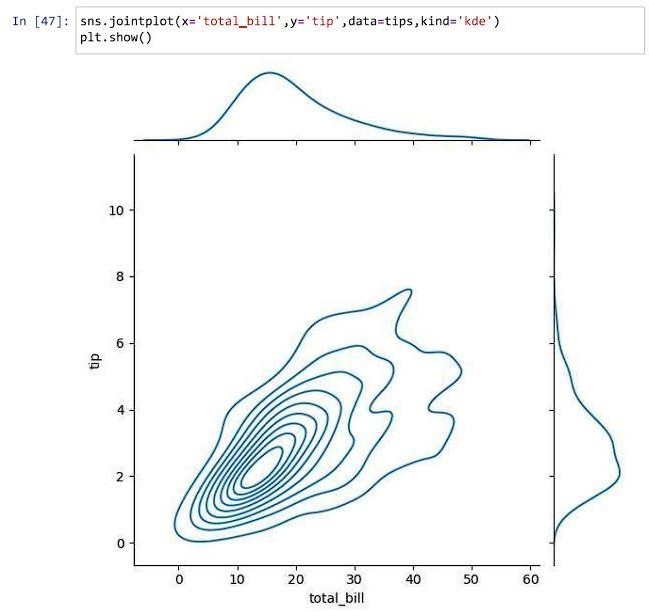


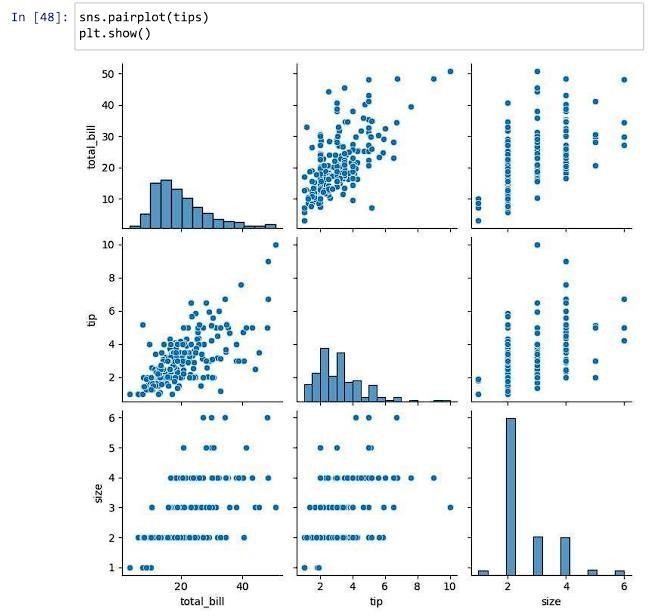


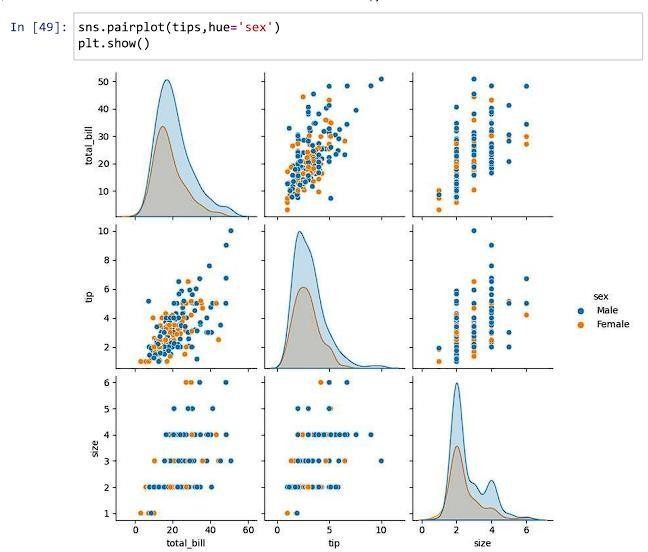




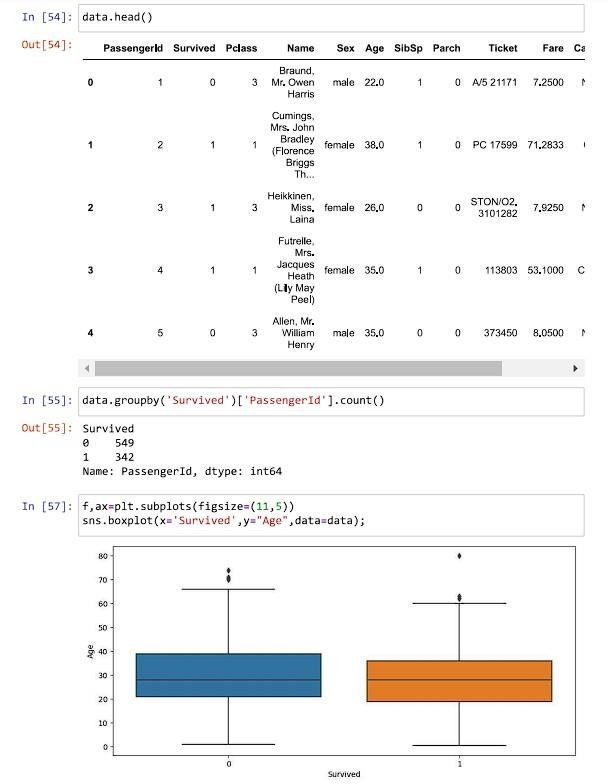
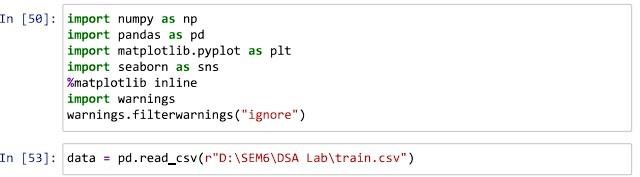


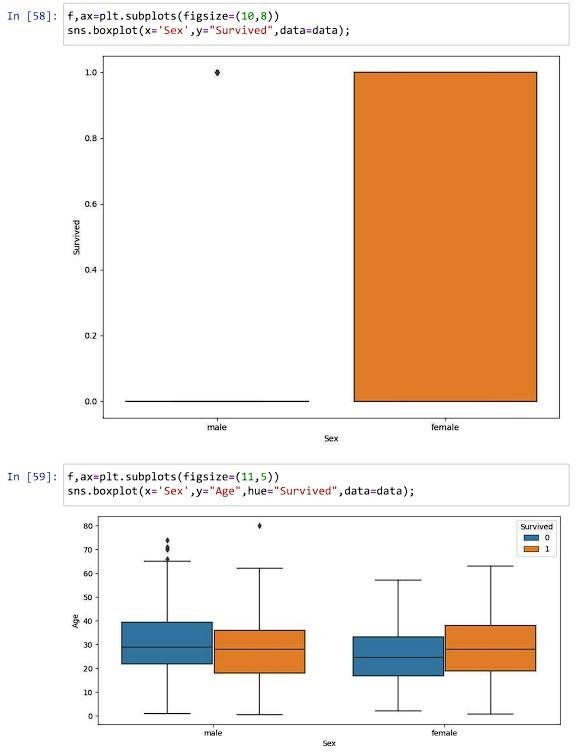


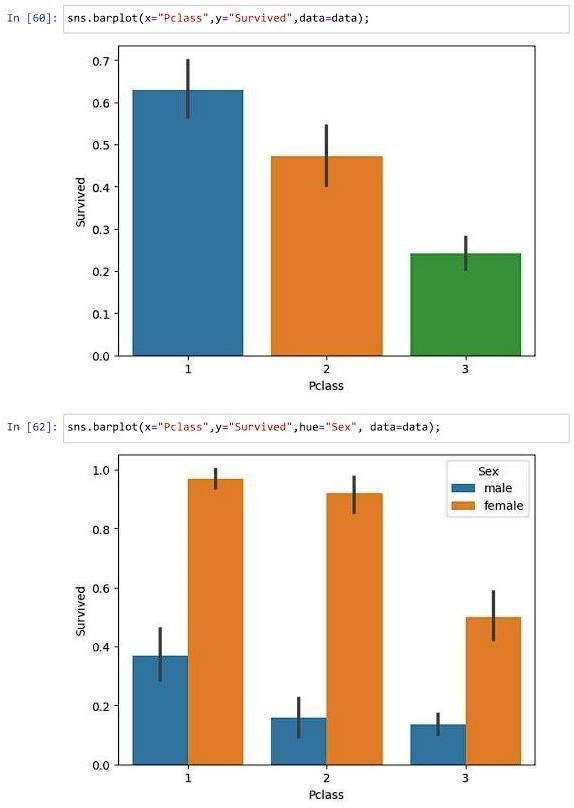


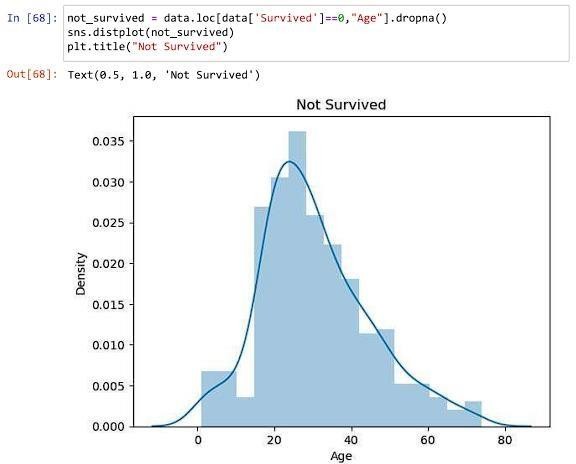
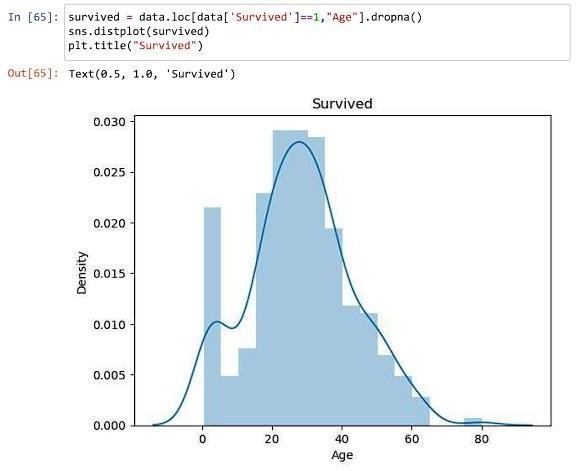
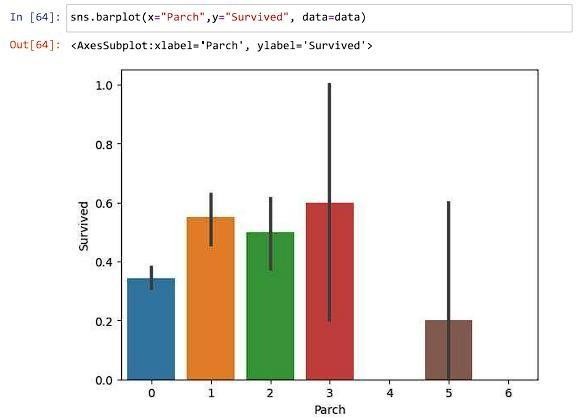
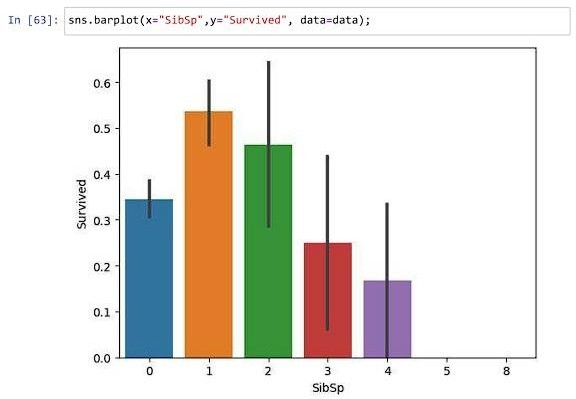


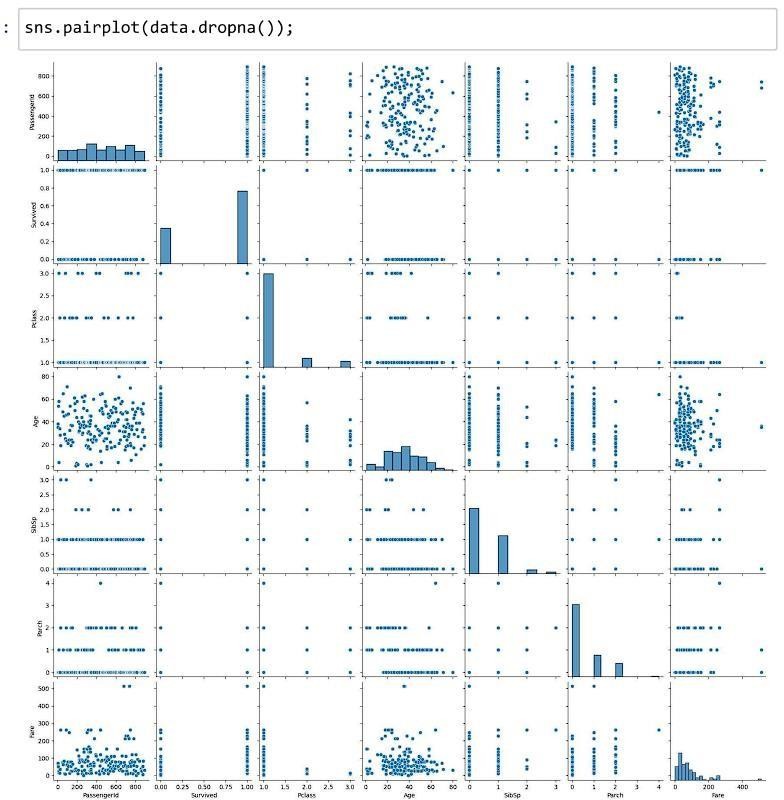
Train Dataset

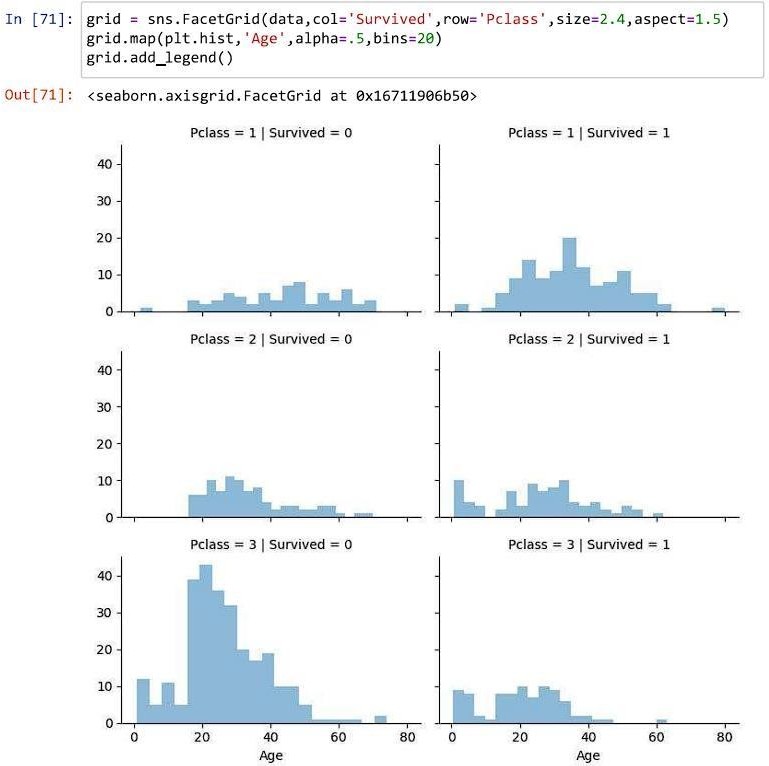


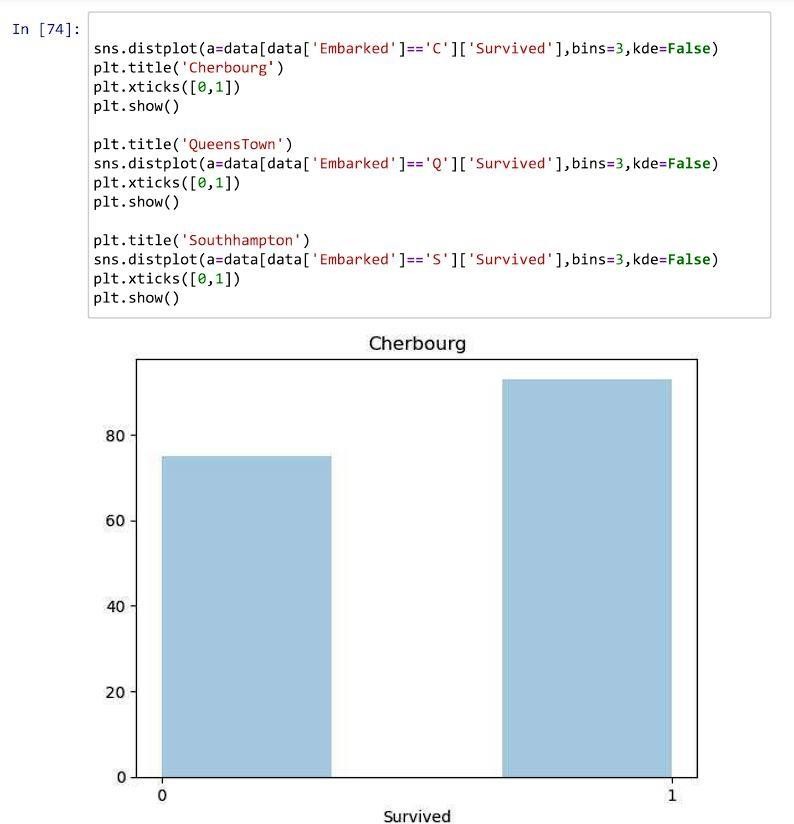


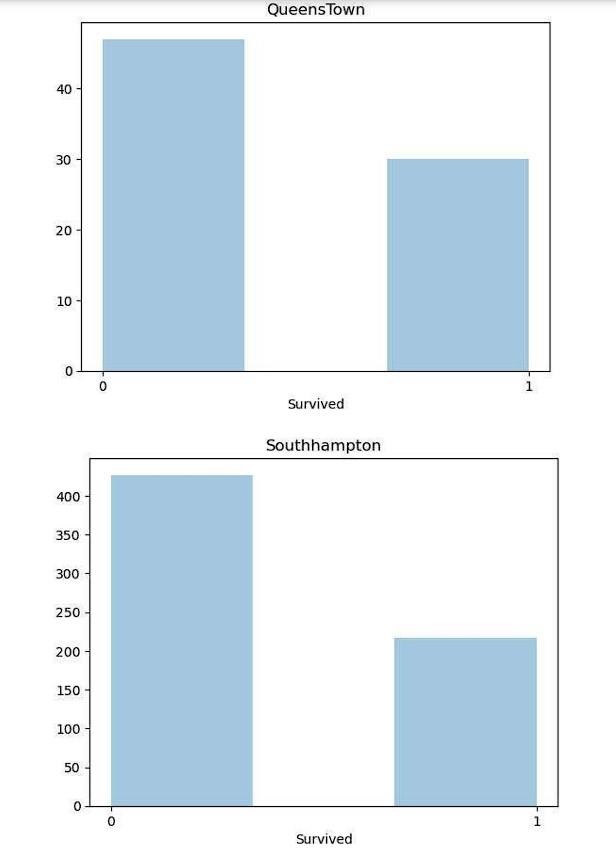












**CONCLUSION:** In conclusion, the data visualization with Seaborn provided valuable insights into the Iris dataset, showcasing various relationships and patterns. The subsequent ANOVA analysis aimed to determine if there were significant differences in the means of different features across the three Iris species. The specific outcomes of the ANOVA test, which are missing in the provided information, would be essential to draw a final conclusion regarding the statistical significance of these differences