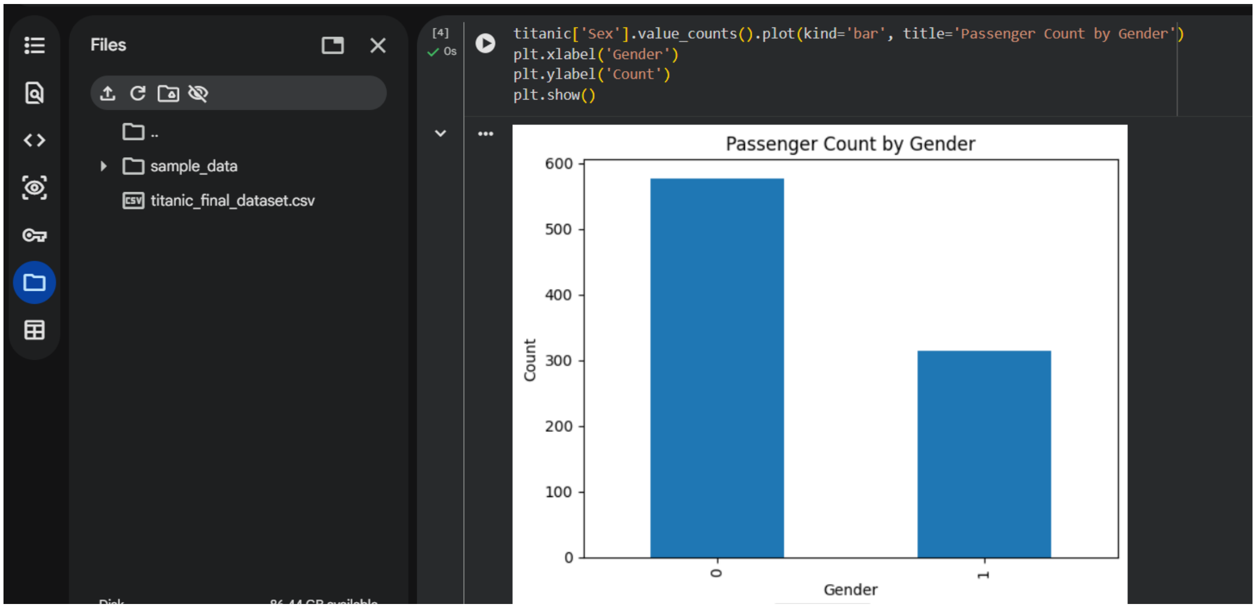
University of Colombo School of Computing

IS 2210 Applied Data Science

Lab Sheet 05

23020212

A.

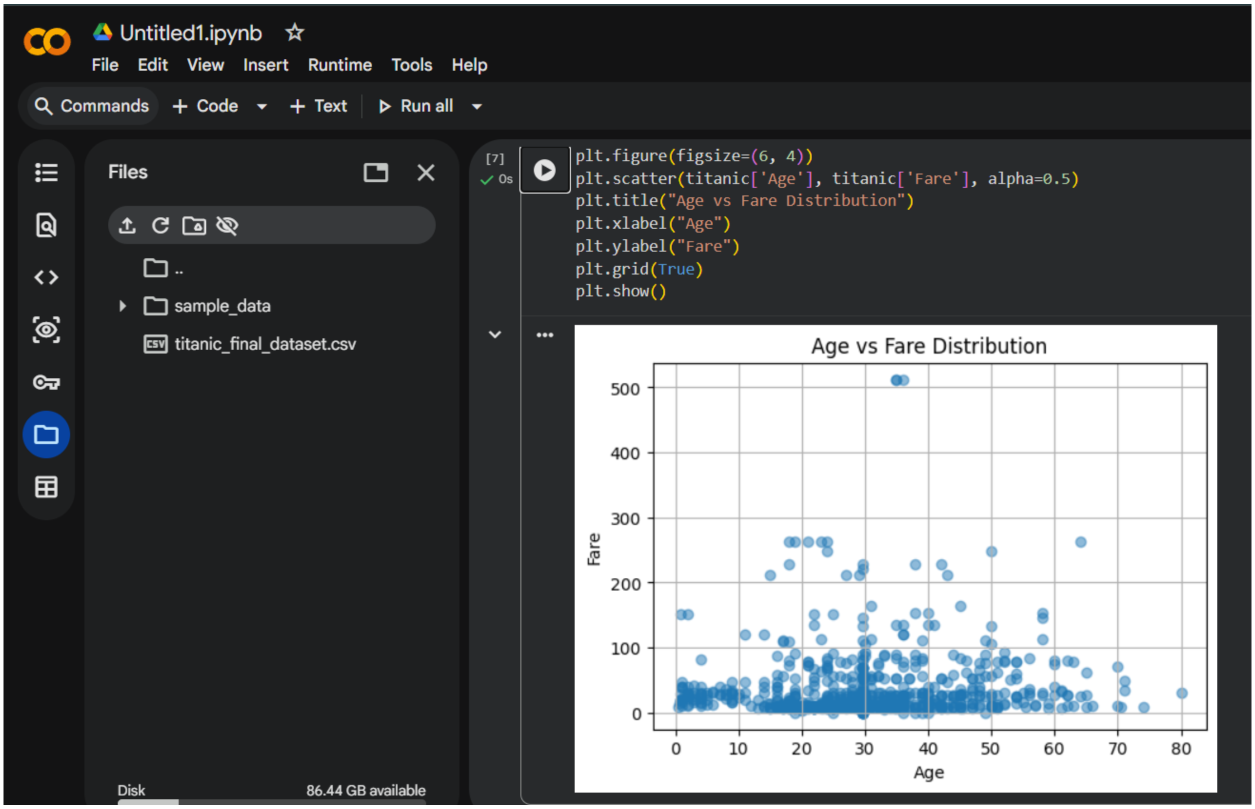


```
titanic['Sex'].value_counts().plot(kind='bar', title='Passenger Count by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

1. A bar chart is suitable because "Gender" is a categorical variable, and we are comparing the frequency of each category. It allows for easy visual comparison of discrete groups.

2. Pandas is sufficient because it allows plotting directly from a Data Frame using simple methods like. plot(), which is ideal for quick, basic visualizations without needing extensive styling code.

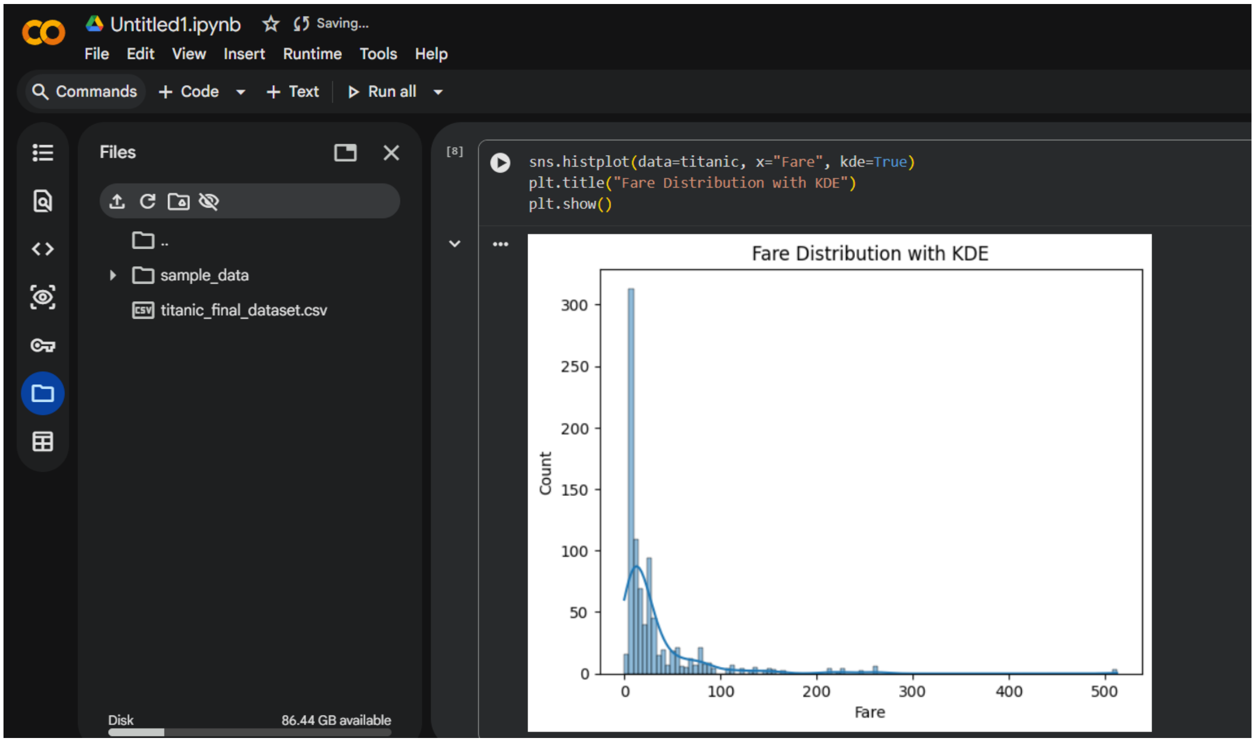
B.



```
plt.figure(figsize=(6, 4))
plt.scatter(titanic['Age'], titanic['Fare'], alpha=0.5)
plt.title("Age vs Fare Distribution")
plt.xlabel("Age")
plt.ylabel("Fare")
plt.grid(True)
plt.show()
```

1. There is generally no strong linear correlation between Age and Fare; however, we may observe that higher fares are distributed across all age groups, whereas lower fares are clustered more heavily among younger adults.
2. Matplotlib is often preferred when specific customization is needed "explicitly," such as manually setting titles, labels, and grid lines to build the plot from scratch.

C.



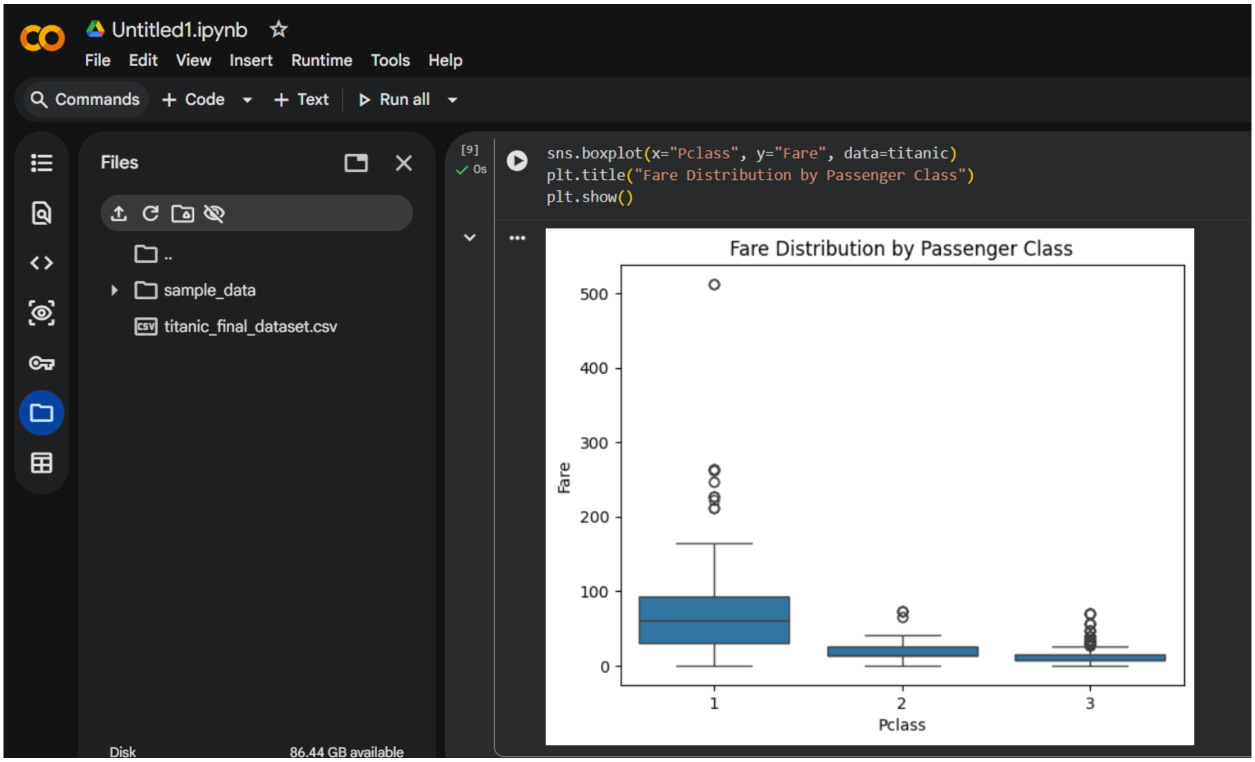
```
sns.histplot(data=titanic, x="Fare", kde=True)
plt.title("Fare Distribution with KDE")
plt.show()
```

1. The distribution is likely right-skewed (positively skewed), meaning most passengers paid low fares, with a long tail extending toward high fares.

2. The KDE shows the overall distribution shape as a smooth curve, helping to visualize the probability density alongside the frequency counts.

3. Seaborn is more suitable because it is built for statistical plotting and can add a KDE curve to a histogram with "minimal code" and "better default styling".

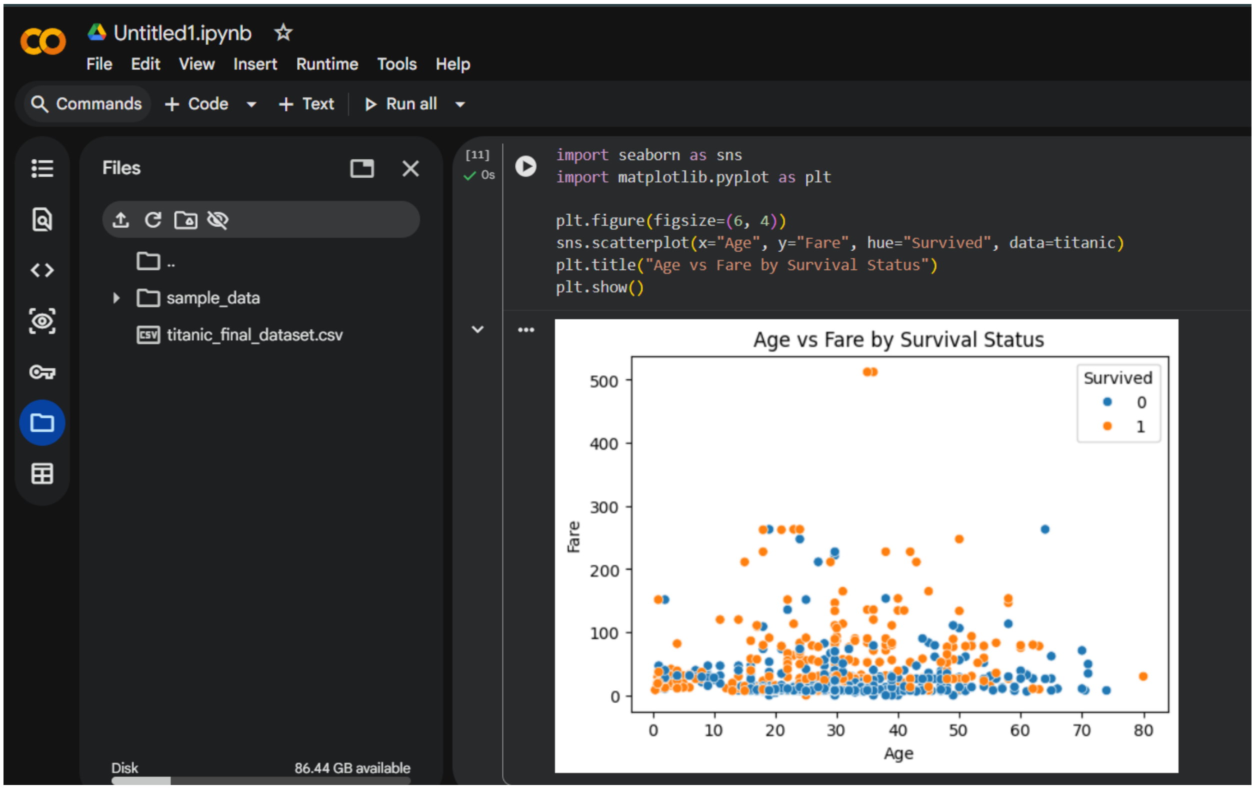
D.



```
sns.boxplot(x="Pclass", y="Fare", data=titanic)
plt.title("Fare Distribution by Passenger Class")
plt.show()
```

1. Passenger Class 1 shows the highest fare variability.
2. Yes, outliers are clearly present
   Representation: These represent passengers who paid significantly higher fares than the typical range for their specific class.
3. A box plot is superior here because it visualizes the distribution spread, median, and outliers all at once.

E.

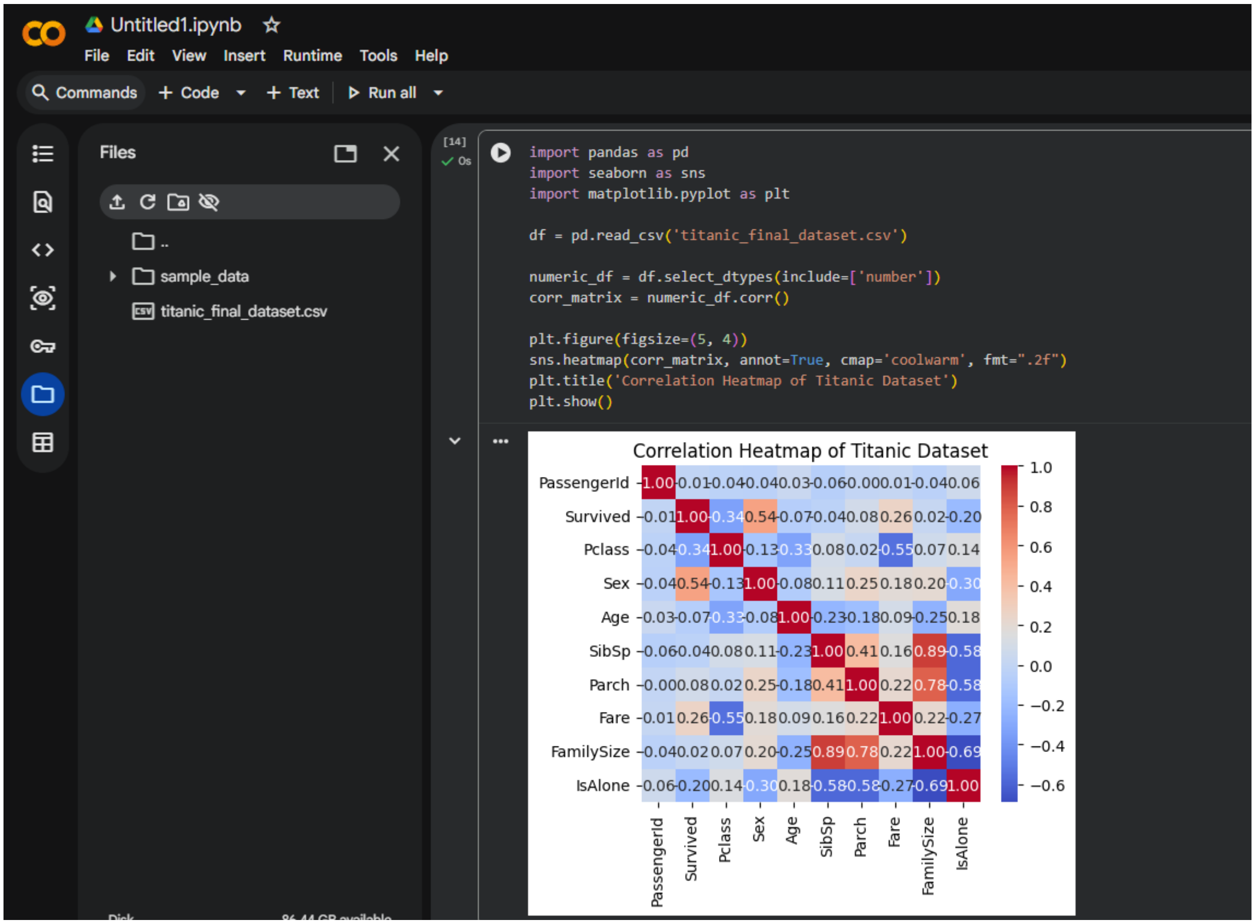


```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(6, 4))
sns.scatterplot(x="Age", y="Fare", hue="Survived", data=titanic)
plt.title("Age vs Fare by Survival Status")
plt.show()
```

1. Yes, there is a noticeable clustering pattern.
2. Color encoding adds a third dimension to the visualization.

F.



```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('titanic_final_dataset.csv')

numeric_df = df.select_dtypes(include=['number'])
corr_matrix = numeric_df.corr()

plt.figure(figsize=(5, 4))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap of Titanic Dataset')
plt.show()
```

Correlation Heatmap of Titanic Dataset

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	FamilySize	IsAlone
PassengerId	1.00	-0.01	-0.04	-0.04	0.03	-0.06	-0.00	0.01	-0.04	0.06
Survived	-0.01	1.00	-0.34	0.54	-0.07	-0.04	0.08	0.26	0.02	-0.20
Pclass	-0.04	-0.34	1.00	-0.13	-0.33	0.08	0.02	-0.55	0.07	0.14
Sex	-0.04	0.54	-0.13	1.00	-0.08	0.11	0.25	0.18	0.20	-0.30
Age	0.03	-0.07	-0.33	-0.08	1.00	-0.23	-0.18	0.09	-0.25	0.18
SibSp	-0.06	-0.04	0.08	0.11	-0.23	1.00	0.41	0.16	0.89	-0.58
Parch	-0.00	0.08	0.02	0.25	-0.18	0.41	1.00	0.22	0.78	-0.58
Fare	0.01	0.26	-0.55	0.18	0.09	0.16	0.22	1.00	0.22	-0.27
FamilySize	-0.04	0.02	0.07	0.20	-0.25	0.89	0.78	0.22	1.00	-0.69
IsAlone	0.06	-0.20	0.14	-0.30	0.18	-0.58	-0.58	-0.27	-0.69	1.00

1. Strong Correlation: SibSp (Siblings/Spouses) and FamilySize have a very strong positive correlation (0.89). This makes sense because FamilySize is calculated using SibSp. Another strong negative correlation is between Pclass and Fare (-0.55), meaning as the class number goes up (1st -> 3rd), the fare goes down.

   Weak Correlation: PassengerId and Parch (Parents/Children) have an extremely weak correlation (~0.00). This confirms that the randomly assigned ID has no relationship with family size.
2. Yes, there is a moderate positive correlation
3. Correlation only shows that two variables move together, not that one causes the other.

G.

1. Pandas is best for quick exploration. It allows you to create plots directly from a Data Frame with a single line of code, making it the fastest way to get a basic visual of your data without importing extra libraries or writing complex syntax.

2. Matplotlib provides maximum customization. It is the foundational library that Pandas and Seaborn are built on. It gives you low-level control over every single element of the plot, allowing you to build any visualization from scratch.

3. Seaborn is most suitable for statistical visualizations. It is specifically designed to handle statistical aggregation (like calculating means or error bars automatically), visualizing distributions (like KDE or violin plots), and plotting complex relationships (like heatmaps or pair plots) with simple, high-level commands and better default aesthetics.

H. Correlation Heatmap

1. The cell intersecting Pclass and Fare is dark blue, indicating a negative value of -0.55.

2. This strong negative correlation confirms the pricing structure of the Titanic: "1st Class" corresponds to the lowest number but the highest price, while "3rd Class" corresponds to the highest number but the lowest price. The mathematical direction perfectly aligns with the real-world hierarchy.