

# SCT\_DS\_2 :- Exploratory Data Analysis on the Titanic Dataset: A Deep Dive into Data Insights

The screenshot displays a Jupyter Notebook environment with the following components:

- Top Bar:** Includes the Colab logo, the file name "SCT\_DS\_2.ipynb", and icons for star, cloud, chat, settings, share, Gemini, and a profile icon.
- Menu Bar:** Contains "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help".
- Toolbar:** Features "Commands", "+ Code", "+ Text", "Run all", and status indicators for RAM and Disk.
- Left Panel (Files):** Shows a file explorer with folders like "bin", "boot", "content", "sample\_data", "test", "datalab", "dev", "etc", "home", "kaggle", "lib", "lib32", "lib64", and "libx32". The "test.csv" file is highlighted.
- Code Cells:**
  - Cell [3]: `import pandas as pd`
  - Cell [7]: `pd.read_csv("/content/test.csv")`
  - Cell [8]: `df.head()`
- Data View:** A table showing the first 18 rows of the dataset. The columns are PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.
- Bottom Bar:** Includes "Variables", "Terminal", a star icon, and a status bar showing "10:16" and "Python 3".

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

418 rows x 11 columns

SCT\_DS\_2.ipynb

☆

FileEditViewInsertRuntimeToolsHelp

Q Commands

+ Code+ Text

▶ Run all ▼

RAM

Disk

ShareGemini

p

Files

bin

boot

content

sample\_data

test

test.csv

datalab

dev

etc

home

kaggle

lib

lib32

lib64

libx32

Disk69.03 GB available

[8] df.head()

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

Next steps:

Generate code with df

View recommended plots

New interactive sheet

[9] df.tail()

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

Variables

Terminal

10:16

Python 3



Commands

+ Code

+ Text

Run all



RAM



Disk



Files



- bin
- boot
- content
  - sample\_data
  - test
  - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk 69.03 GB available

[10] df.shape

(418, 11)

df.dtypes

	0
PassengerId	int64
Pclass	int64
Name	object
Sex	object
Age	float64
SibSp	int64
Parch	int64
Ticket	object
Fare	float64
Cabin	object
Embarked	object

dtype: object





Q Commands + Code + Text ▶ Run all ▼



## Files



- bin
- boot
- content
  - sample\_data
  - test
  - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk 69.03 GB available

[12] `df.isnull().sum()`

0

PassengerId 0

Pclass 0

Name 0

Sex 0

Age 86

SibSp 0

Parch 0

Ticket 0

Fare 1

Cabin 327

Embarked 0

dtype: int64

[13] `df['Age'] = df['Age'].fillna(df['Age'].median())`▶ `df.drop(columns=['Cabin'])`

CO

SCT\_DS\_2.ipynb

☆

FileEditViewInsertRuntimeToolsHelp

Q Commands

+ Code

+ Text

▶ Run all

▼

RAM

Disk

✓

▼

▲

Files

bin

boot

content

sample\_data

test

test.csv

datalab

dev

etc

home

kaggle

lib

lib32

lib64

libx32

Disk69.03 GB available

df.drop(columns=['Cabin'])

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	S
...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	27.0	0	0	A.5. 3236	8.0500	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	S
416	1308	3	Ware, Mr. Frederick	male	27.0	0	0	359309	8.0500	S
417	1309	3	Peter, Master. Michael J	male	27.0	1	1	2668	22.3583	C

418 rows × 10 columns

[16] df['Fare'] = df['Fare'].fillna(df['Fare'].median())

[17] df.duplicated().sum()

np.int64(0)

Variables

Terminal

10:16

Python 3

CO

SCT\_DS\_2.ipynb

☆

☁

File Edit View Insert Runtime Tools Help

Q Commands

+ Code + Text

▶ Run all

✓

RAM

Disk

Files

bin

boot

content

sample\_data

test

test.csv

datalab

dev

etc

home

kaggle

lib

lib32

lib64

libx32

Disk69.03 GB available

✓0s

[17] df.duplicated().sum()

np.int64(0)

✓0s

[21] df['PassengerId'] = df['PassengerId'].astype(str)

df['Ticket'] = df['Ticket'].astype(str)

✓4s

▶

import seaborn as sns

import matplotlib.pyplot as plt

sns.histplot(df['Age'], kde=True)

plt.show()

Count

120

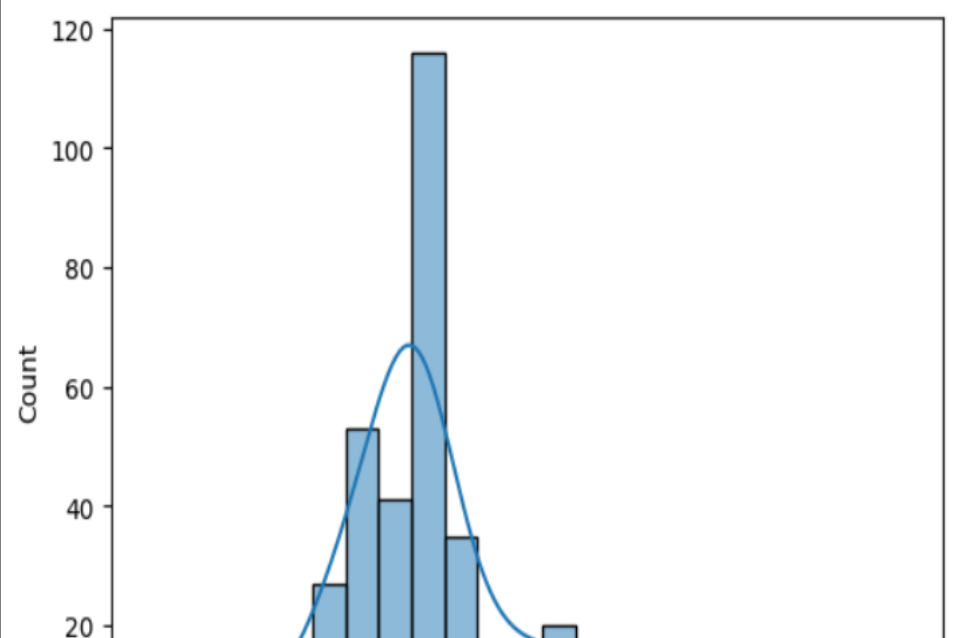
100

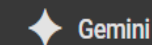
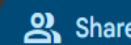
80

60

40

20





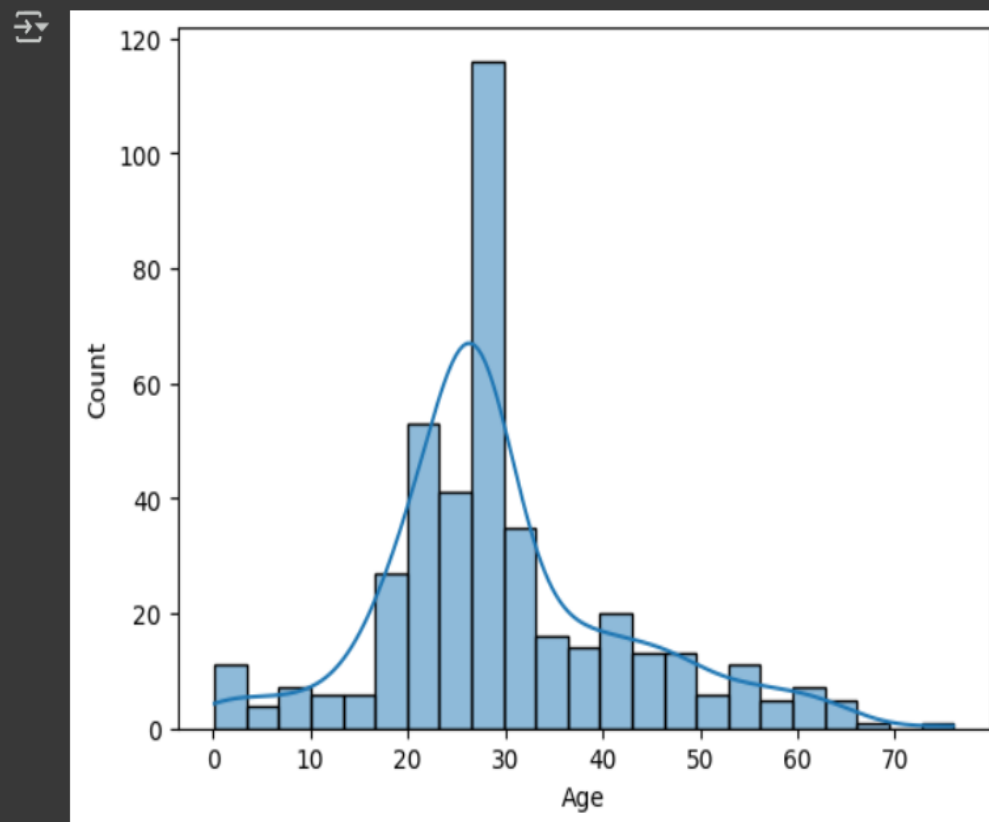
## Files



- bin
- boot
- content
  - sample\_data
  - test
    - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32





Disk 69.03 GB available

```
[22] import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(df['Age'], kde=True)
plt.show()
```



```
sns.boxplot(x='Sex', y='Age', data=df)
```

Files

bin

boot

content

sample\_data

test

test.csv

datalab

dev

etc

home

kaggle

lib

lib32

lib64

libx32

Disk  69.03 GB available

Age

0s  sns.boxplot(x='Sex', y='Fare', data=df)  
plt.show()







A box plot showing the distribution of 'Fare' (Y-axis, 0 to 500) for 'male' and 'female' (X-axis). The plot shows that females generally have higher fares than males. There are several outliers for both groups, with a notable outlier for females at a fare of approximately 510.

Sex	Min	Q1	Median	Q3	Max	Outliers
male	0	10	15	25	55	~15
female	10	10	20	55	110	~10

✓ [241] numerical df = df.select dtypes (include=['number'])



Files

bin

boot

content

sample\_data

test

test.csv

datalab

dev

etc

home

kaggle

lib

lib32

lib64

libx32

Disk 69.03 GB available

Sex

```
numerical_df = df.select_dtypes(include=['number'])
sns.heatmap(numerical_df.corr(), annot=True, cmap='coolwarm')
plt.show()
```

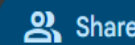
Heatmap showing the correlation matrix for variables Pclass, Age, SibSp, Parch, and Fare. The color scale ranges from -0.4 (blue) to 1.0 (red).

	Pclass	Age	SibSp	Parch	Fare
Pclass	1	-0.47	0.0011	0.019	-0.58
Age	-0.47	1	-0.071	-0.044	0.34
SibSp	0.0011	-0.071	1	0.31	0.17
Parch	0.019	-0.044	0.31	1	0.23
Fare	-0.58	0.34	0.17	0.23	1

Variables Terminal

✓ [In]: df.groupby('Pclass')['Fare'].mean()

✓ 10:16 Python 3



Q Commands + Code + Text ▶ Run all

✓ RAM  
Disk

## Files



- bin
- boot
- content
  - sample\_data
  - test
  - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk 69.03 GB available

```
[25] df.groupby('Pclass')['Fare'].mean()  
df.groupby('Sex')['Fare'].mean()  
df.groupby('Embarked')['Fare'].mean()
```



Fare

Embarked

C 66.259765

Q 10.957700

S 28.179413

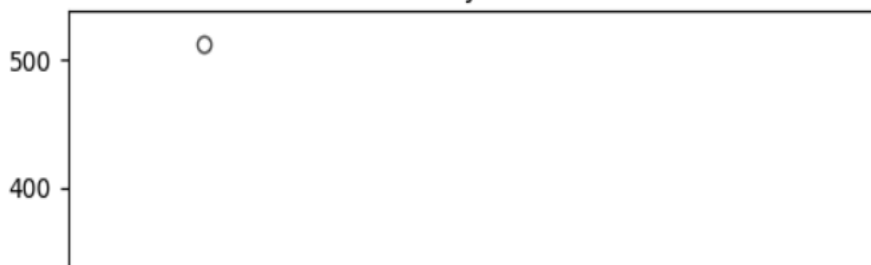
dtype: float64



```
import seaborn as sns  
import matplotlib.pyplot as plt  
sns.boxplot(x='Pclass', y='Fare', data=df)  
plt.title("Fare by Pclass")  
plt.show()
```



Fare by Pclass





Q Commands

+ Code + Text

▶ Run all ▼



RAM

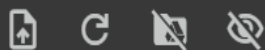
Disk



Files



0s



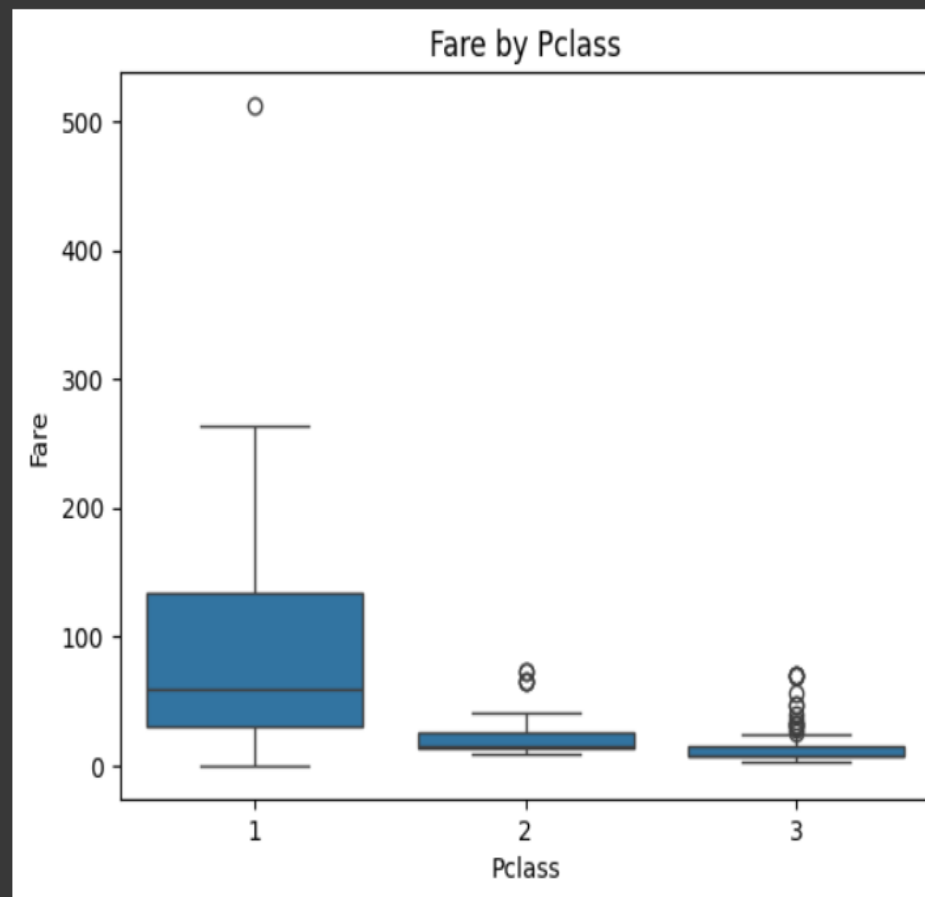
- bin
- boot
- content
  - sample\_data
  - test
    - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk

69.03 GB available



```
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title("Fare by Pclass")
plt.show()
```





## Files

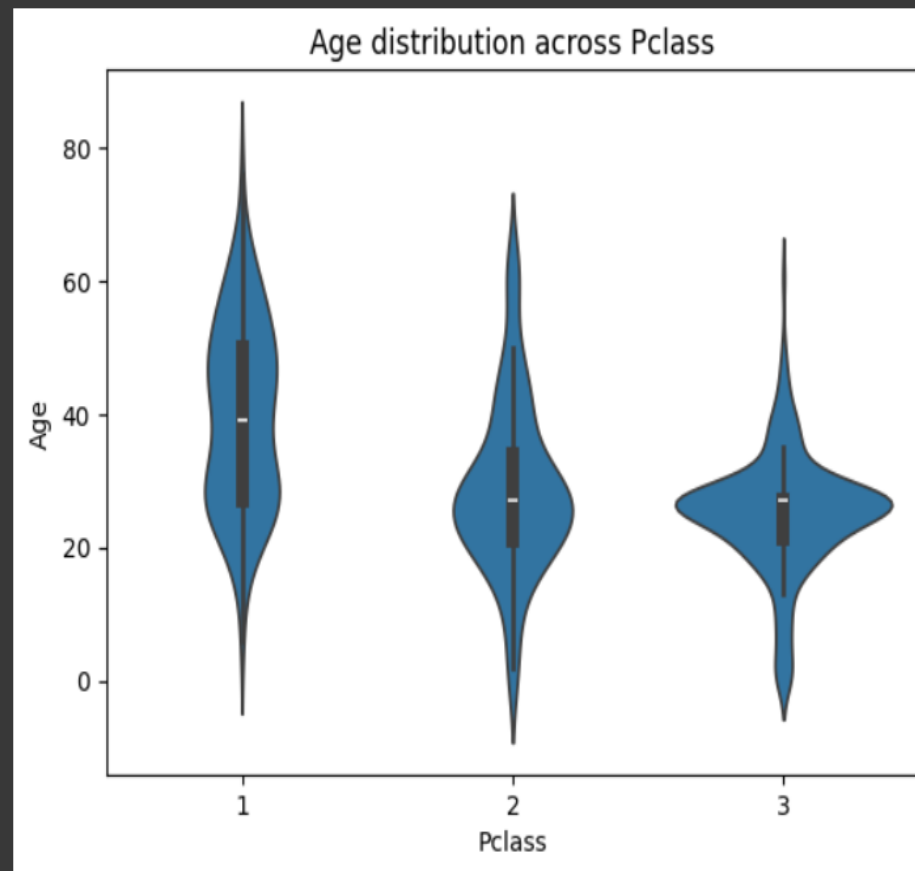


- bin
- boot
- content
  - sample\_data
  - test
    - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk  69.03 GB available

0s

```
sns.violinplot(x='Pclass', y='Age', data=df)
plt.title("Age distribution across Pclass")
plt.show()
```





Q Commands + Code + Text ▶ Run all ▼

✓ RAM  Disk 

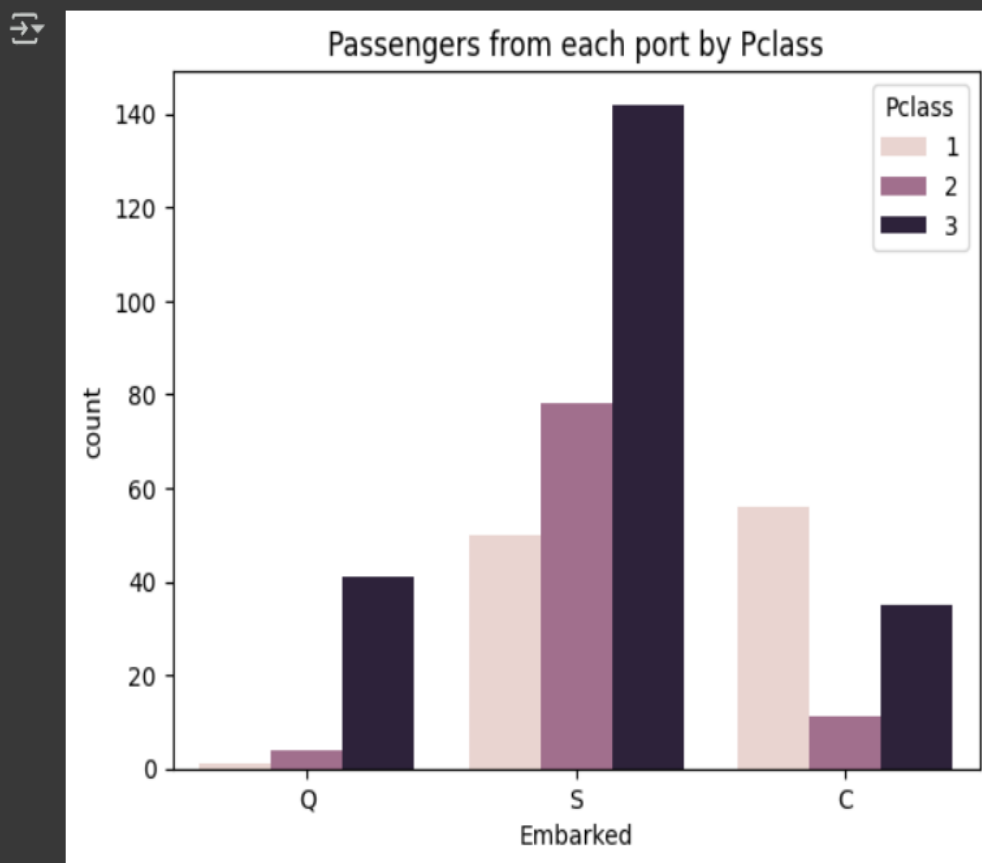
## Files



- bin
- boot
- content
  - sample\_data
  - test
  - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk 69.03 GB available

```
sns.countplot(x='Embarked', hue='Pclass', data=df)
plt.title("Passengers from each port by Pclass")
plt.show()
```



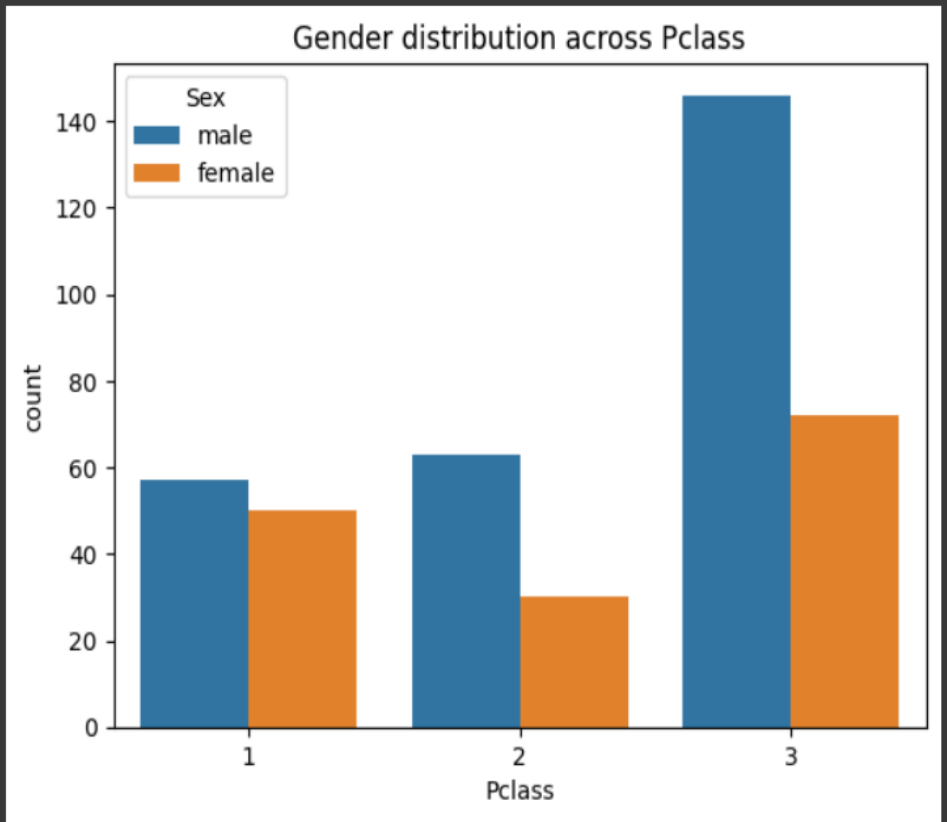
```
[20]: sns.countplot(x='Pclass', hue='Sex', data=df)
```

Files

- bin
- boot
- content
  - sample\_data
  - test
  - test.csv
- datalab
- dev
- etc
- home
- kaggle
- lib
- lib32
- lib64
- libx32

Disk 69.03 GB available

```
sns.countplot(x='Pclass', hue='Sex', data=df)
plt.title("Gender distribution across Pclass")
plt.show()
```



```
pclasses = df['Pclass'].unique()
```

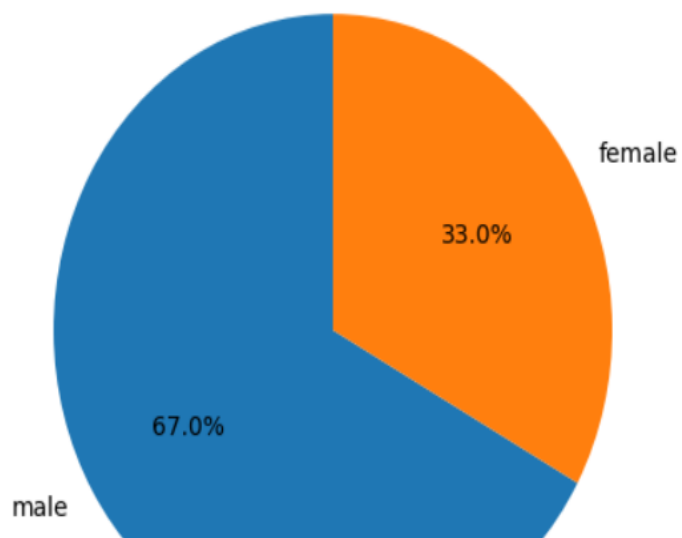


```
pclasses = df['Pclass'].unique()
```

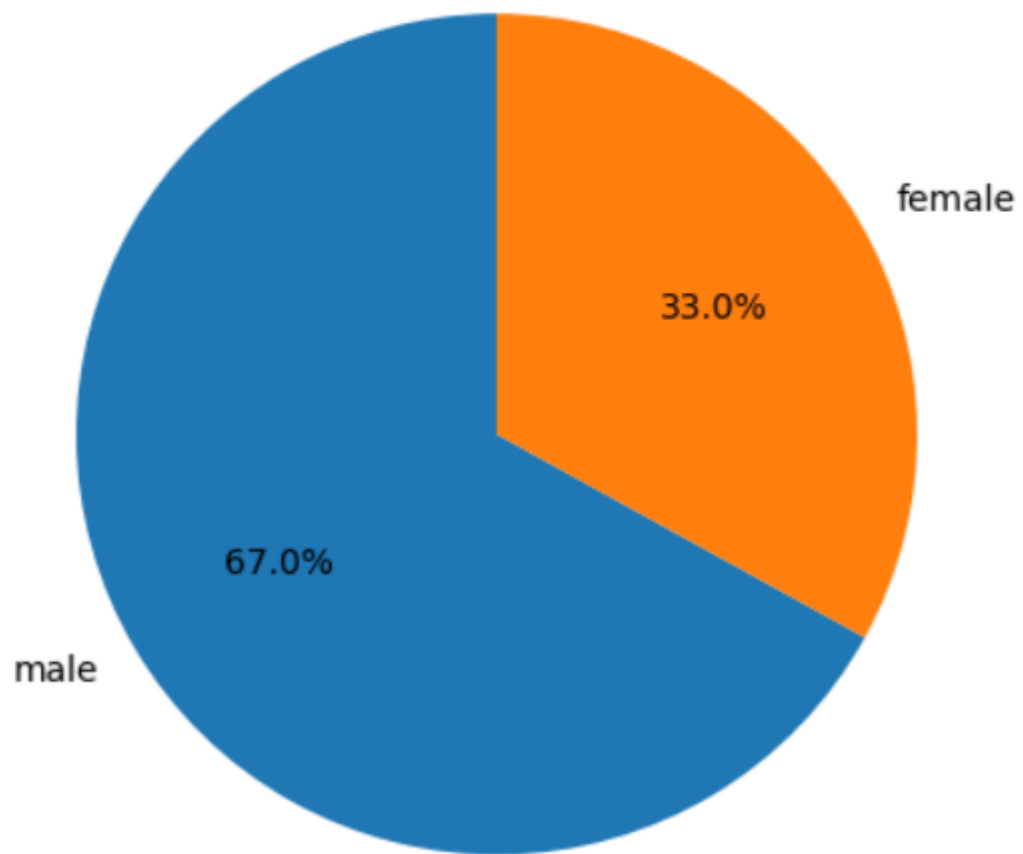
```
for pclass in pclasses:  
    subset = df[df['Pclass'] == pclass]  
    gender_counts = subset['Sex'].value_counts()  
  
    plt.figure()  
    plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%', startangle=90)  
    plt.title(f"Gender Distribution in Pclass {pclass}")  
    plt.axis('equal')  
    plt.show()
```



Gender Distribution in Pclass 3



Gender Distribution in Pclass 3



Gender Distribution in Pclass 2

