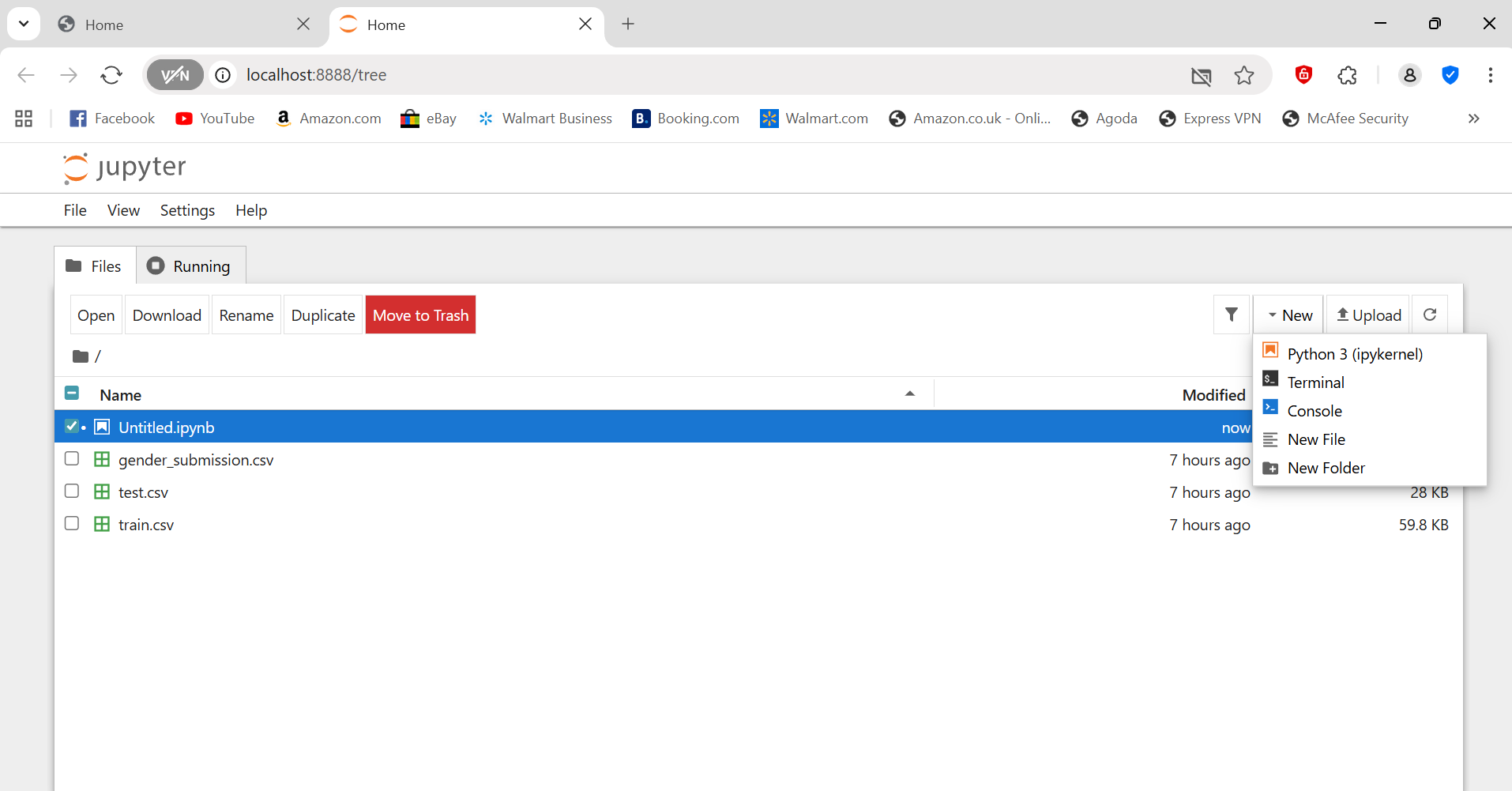
**TASK -5**

cd Documents\Titanic

jupyter notebook

Create a new notebook



-> Click **“New” → Python 3 (ipykernel)** (top-right corner).

-> A new tab opens named **Untitled.ipynb** — this is your notebook.

EDA

# Import libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

df = pd.read\_csv('train.csv')

# Basic exploration

print(df.info())

print(df.describe())

print(df.head())

-----------------------------------------------------------------------------------------------------------

**Data Overview using .info(), .describe(), and .value\_counts()**

import pandas as pd

**Import the visualization libraries**

import matplotlib.pyplot as plt

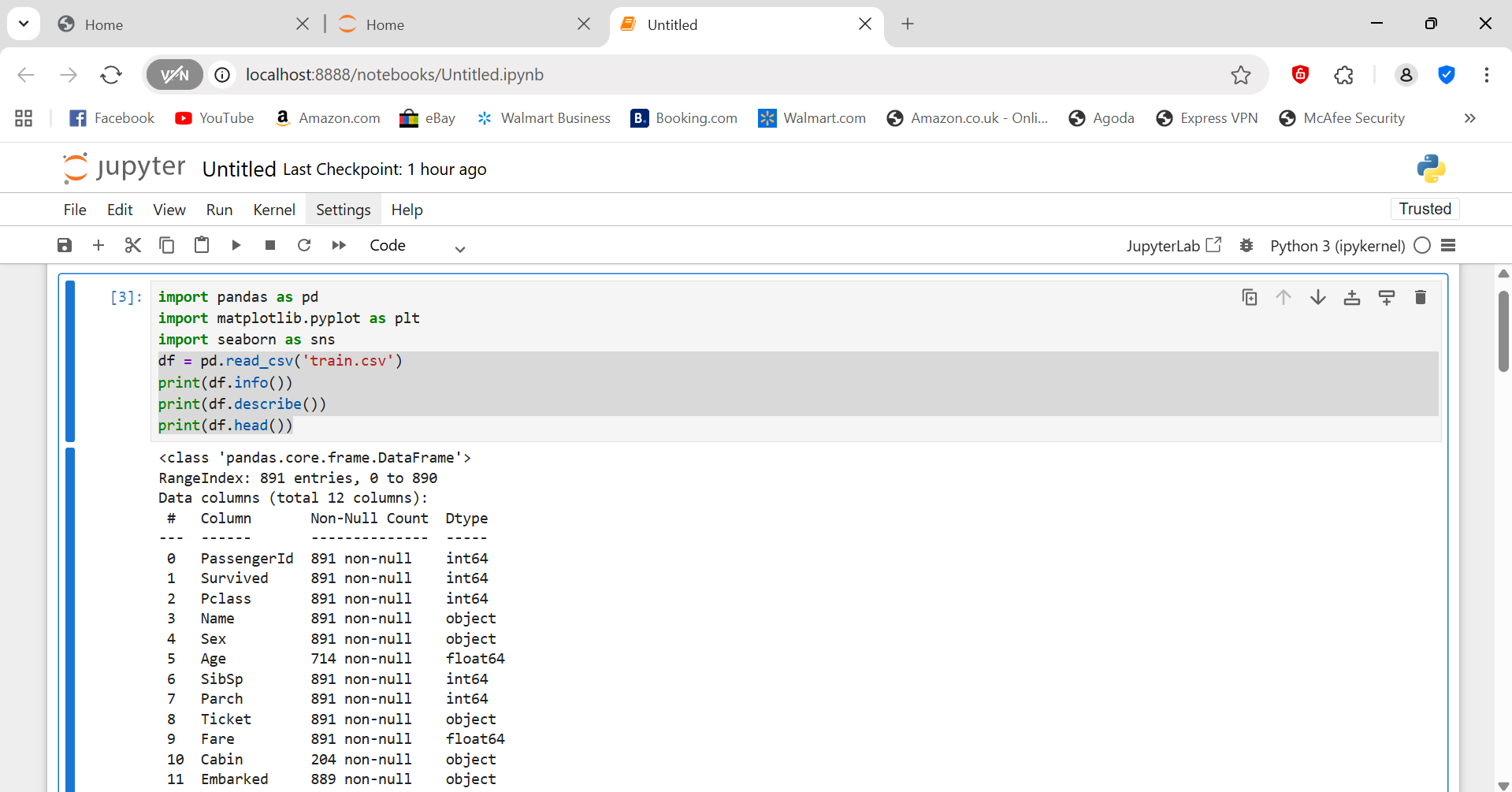
import seaborn as sns

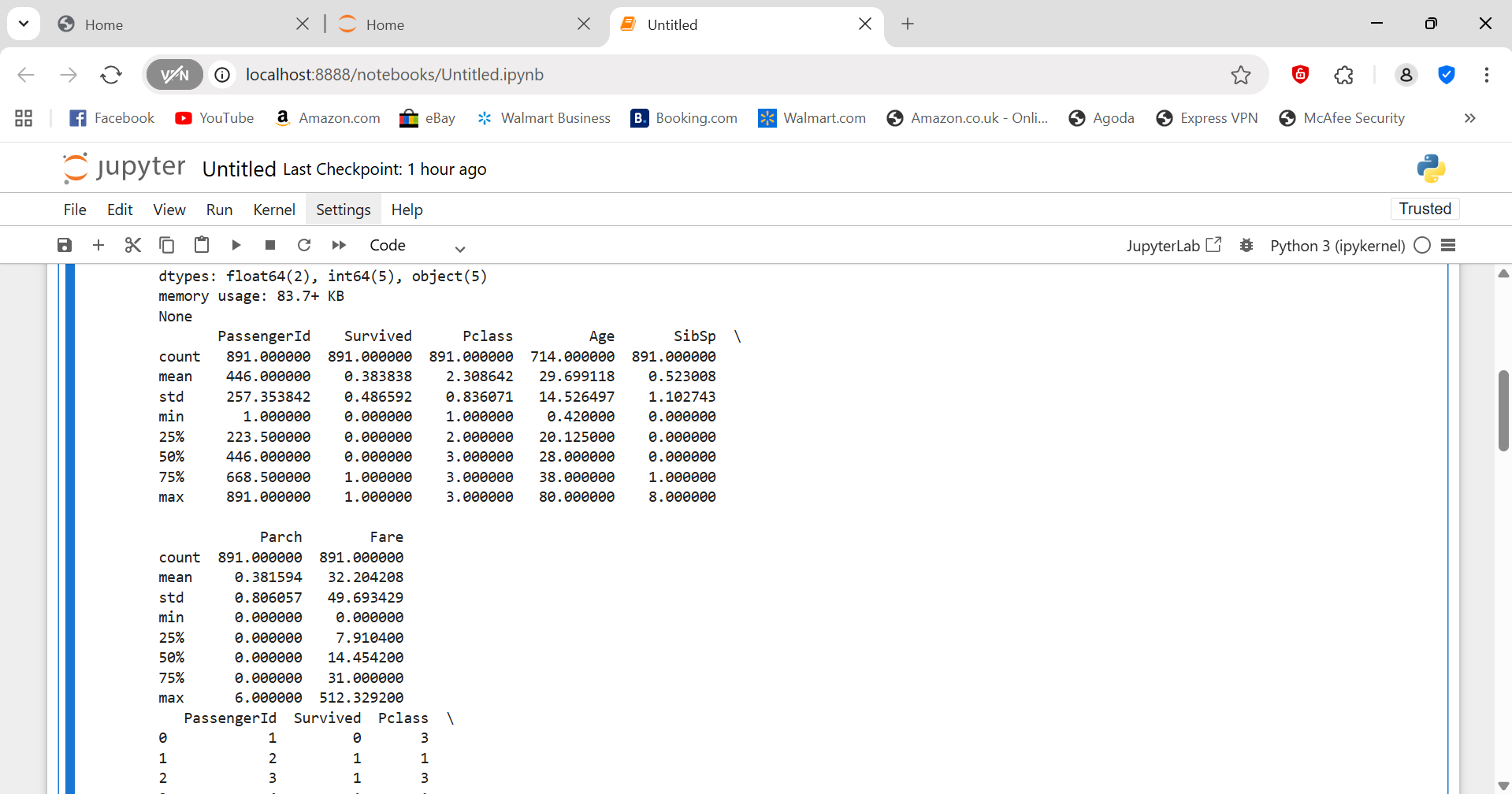
df = pd.read\_csv('train.csv')

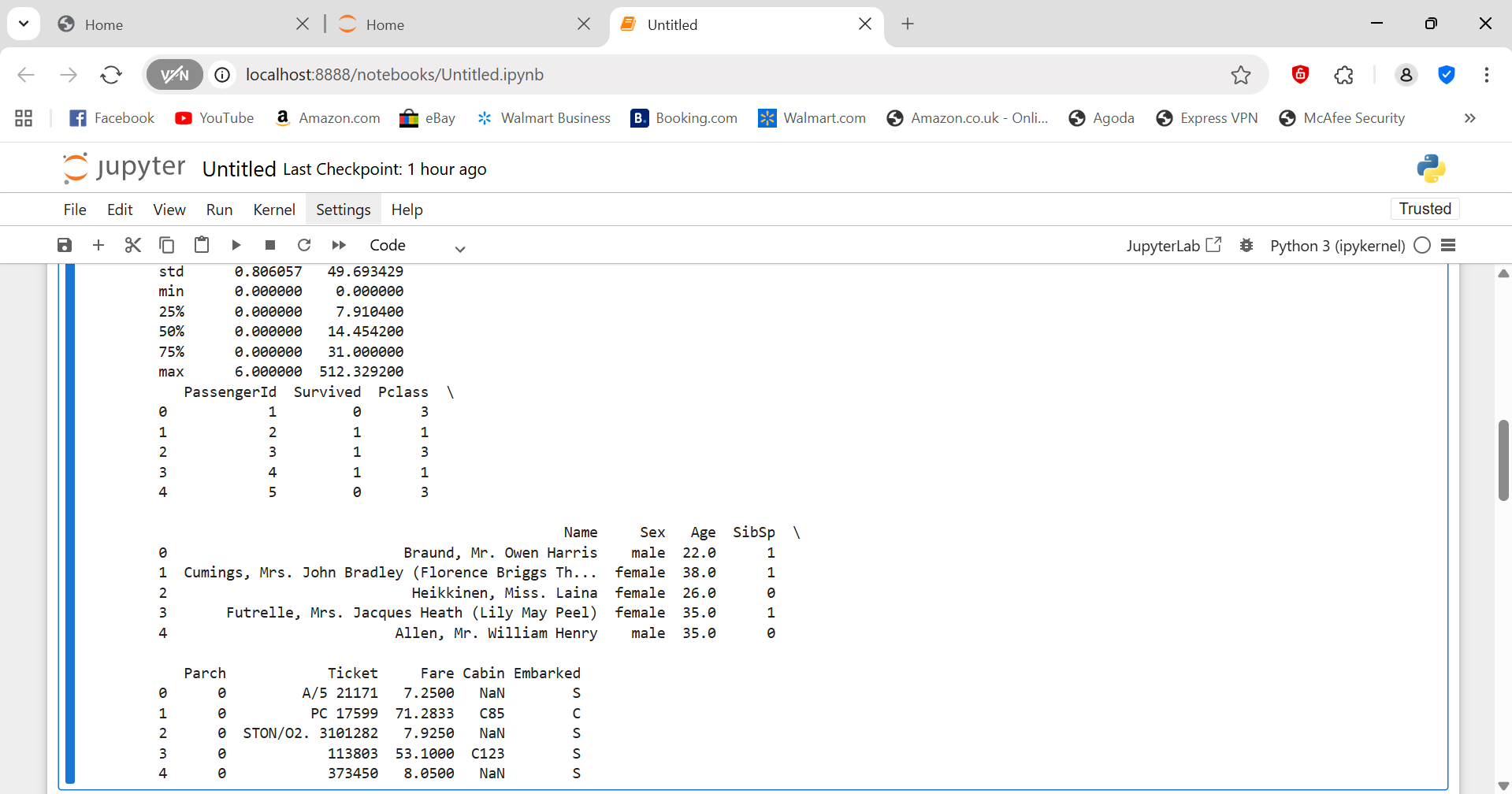
print(df.info())

print(df.describe())

print(df.head())







train = pd.read\_csv("train.csv")

test = pd.read\_csv("test.csv")

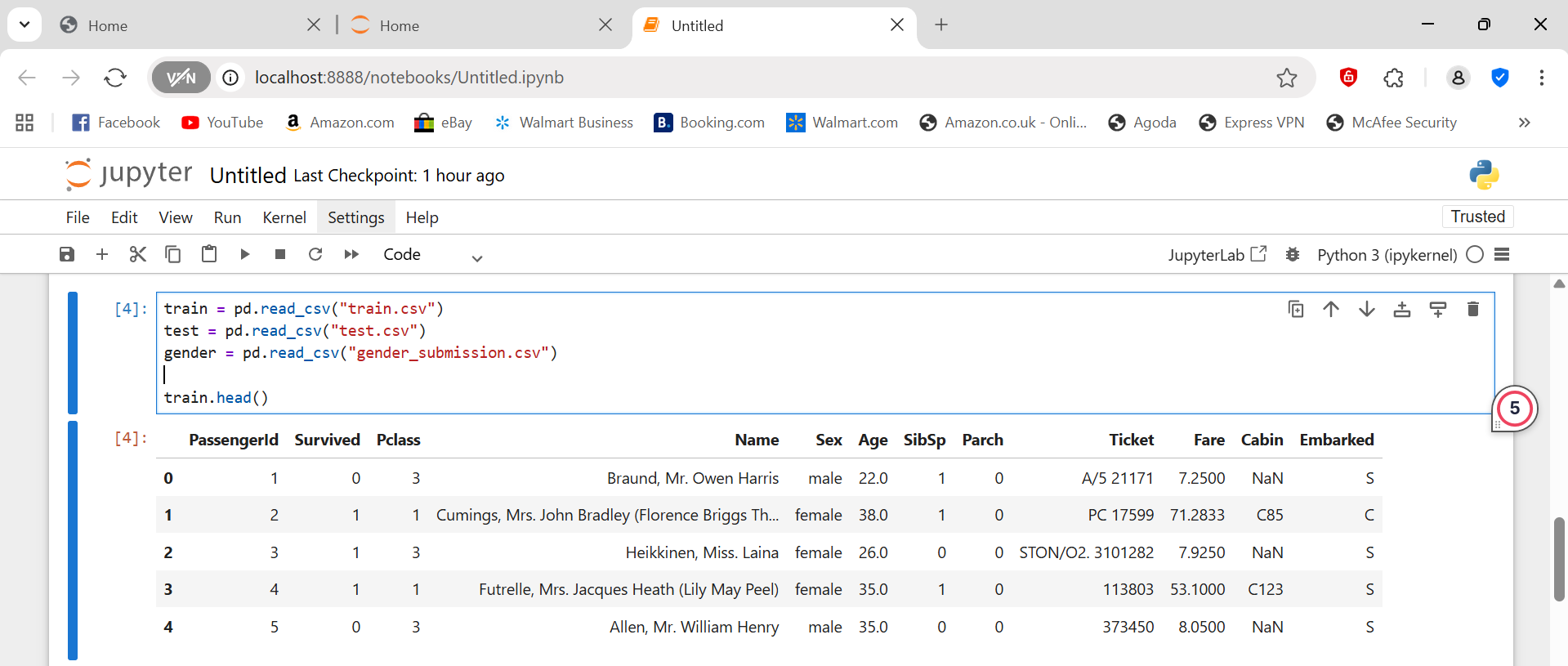
gender = pd.read\_csv("gender\_submission.csv")

train.head()

Observation

- The dataset contains passenger information such as Name, Age, Sex, Ticket, Fare, and Survival status.

- The target column is \*\*'Survived'\*\* (1 = survived, 0 = did not survive).

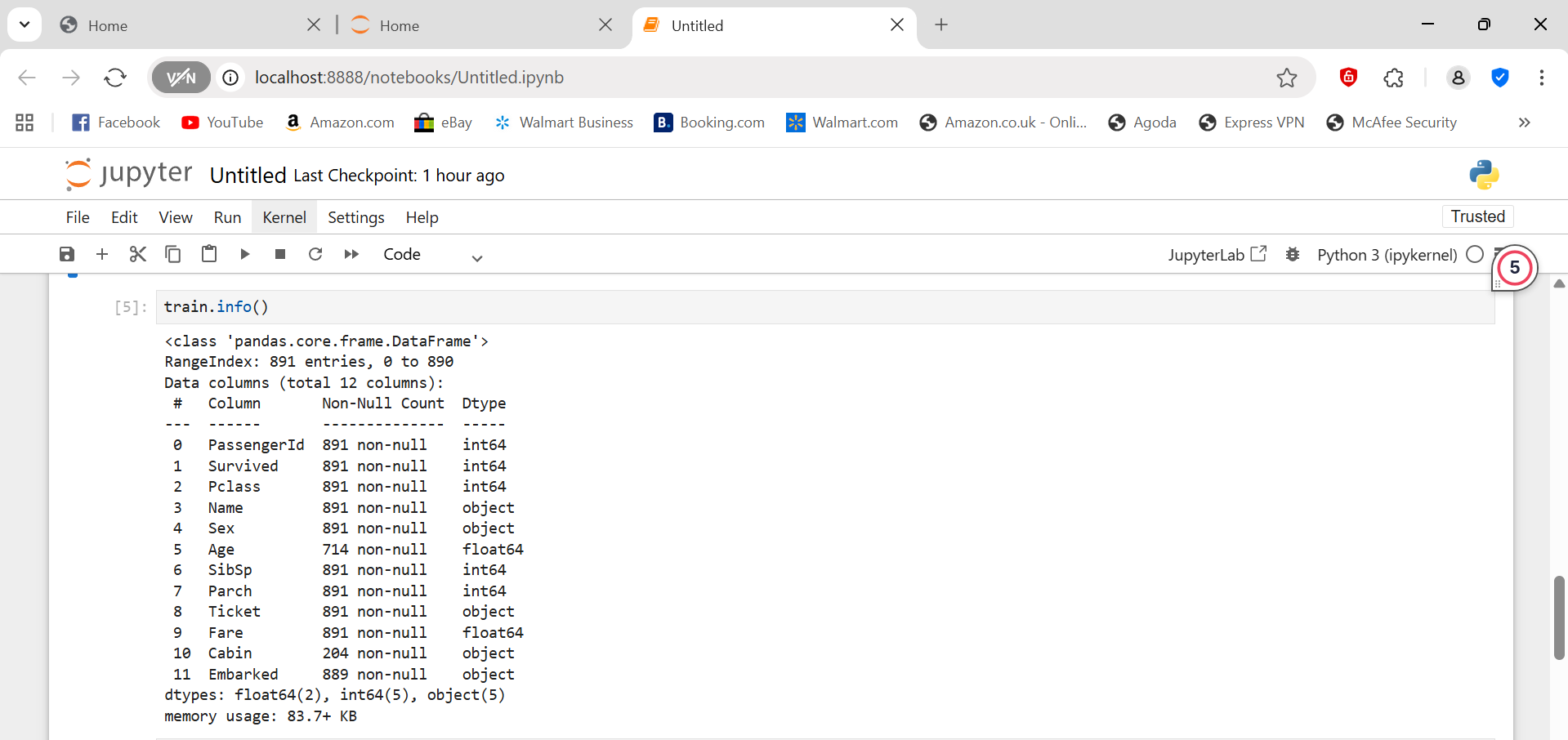


train.info()

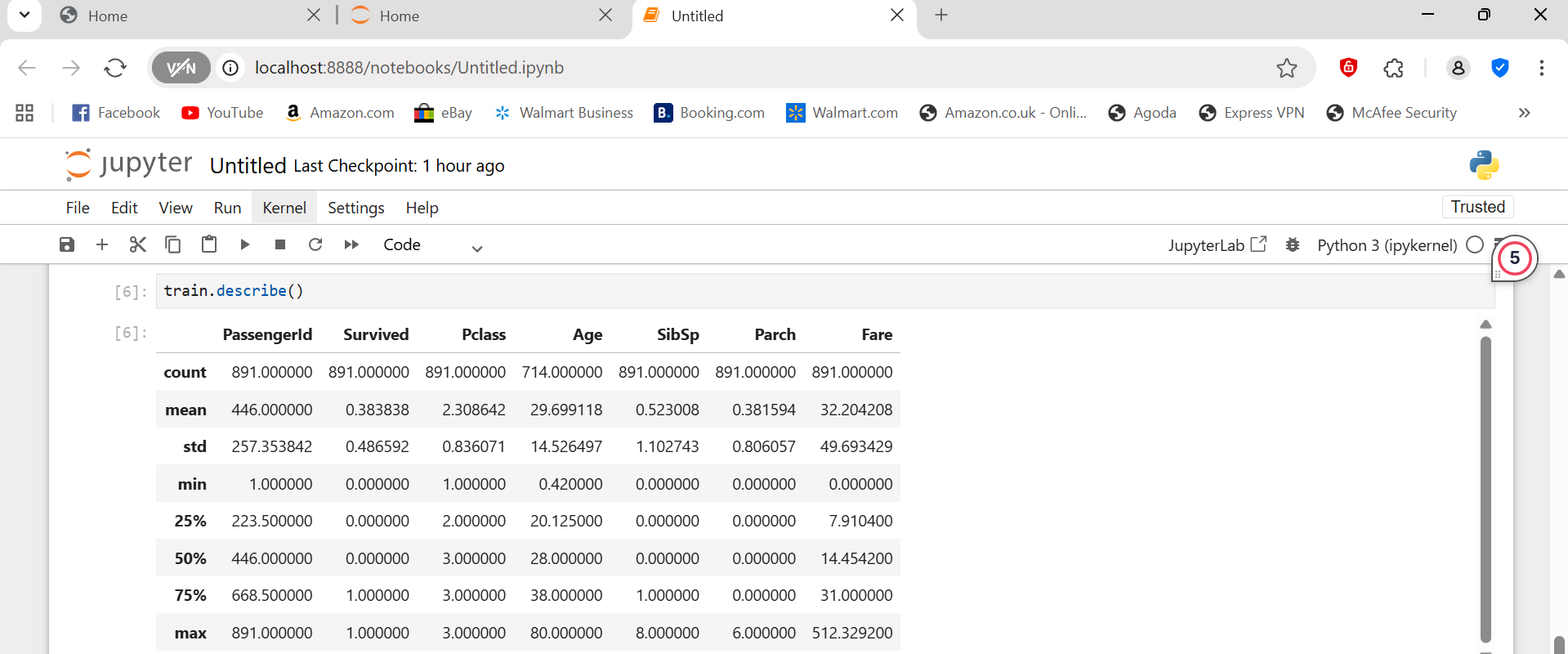
The dataset has \*\*891 rows and 12 columns\*\*.

- Columns like \*\*Age\*\*, \*\*Cabin\*\*, and \*\*Embarked\*\* have missing values.

- Most columns are of type `int64` or `object`.



train.describe()



The average passenger age is approximately \*\*29.7 years\*\*.

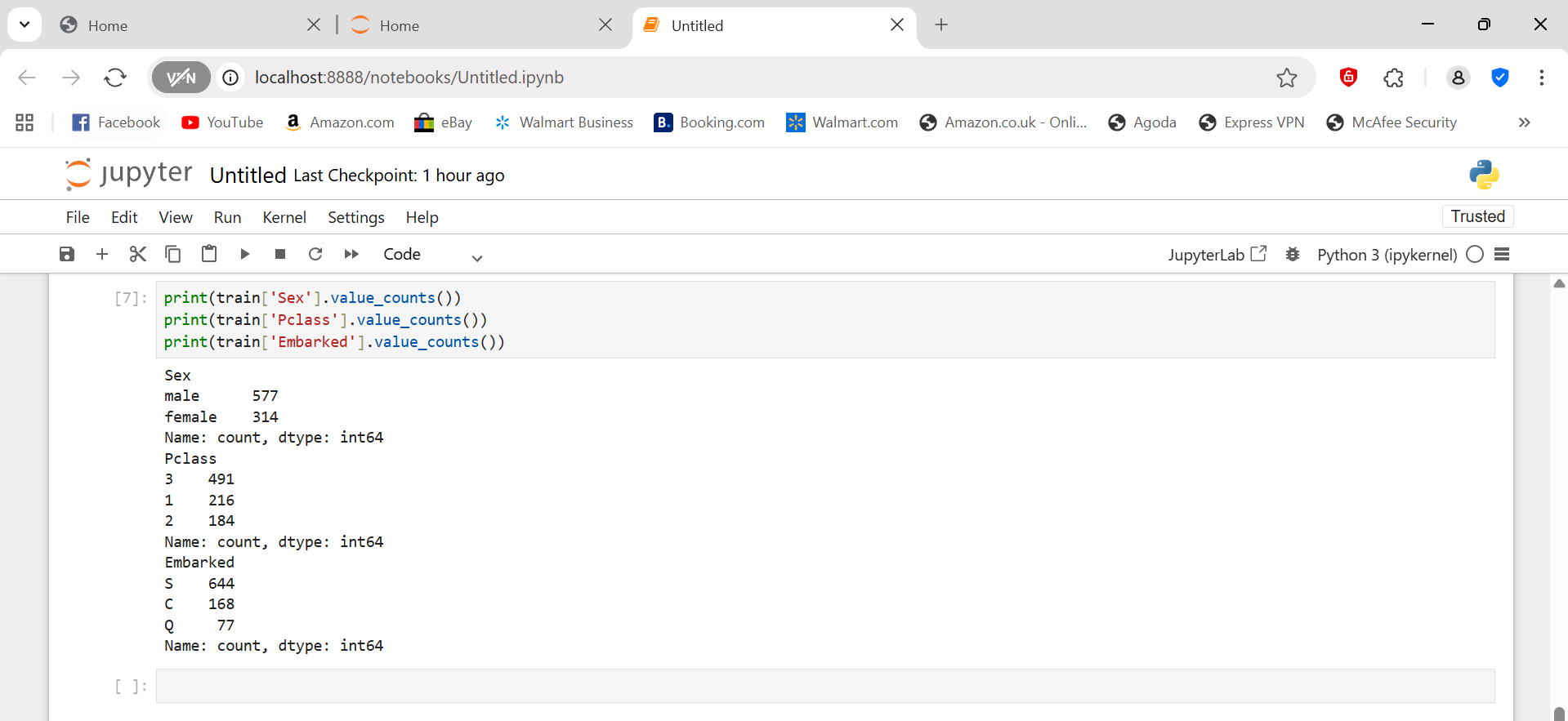
- The average fare is \*\*$32.20\*\*, with a maximum of \*\*$512.33\*\*.

- Around \*\*38% of passengers survived\*\* (mean of 'Survived' = 0.38).

print(train['Sex'].value\_counts())

print(train['Pclass'].value\_counts())

print(train['Embarked'].value\_counts())



- There are \*\*577 males\*\* and \*\*314 females\*\*.

- Most passengers belong to \*\*Pclass 3\*\*.

- The most common embarkation point is \*\*'S' (Southampton)\*\*.

**b.Use sns.pairplot(), sns.heatmap() for visualization**

import pandas as pd

# Load the training dataset

train = pd.read\_csv("train.csv")

# View first few rows to confirm

train.head()

import seaborn as sns

import matplotlib.pyplot as plt

**# Set visual style**

sns.set(style="whitegrid")  
  
**Pairplot — Explore Relationships Between Variables**

The **pairplot** function lets you see how multiple numerical variables relate to each other.

sns.pairplot(train[['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']],

hue='Survived',

palette='coolwarm',

diag\_kind='kde')

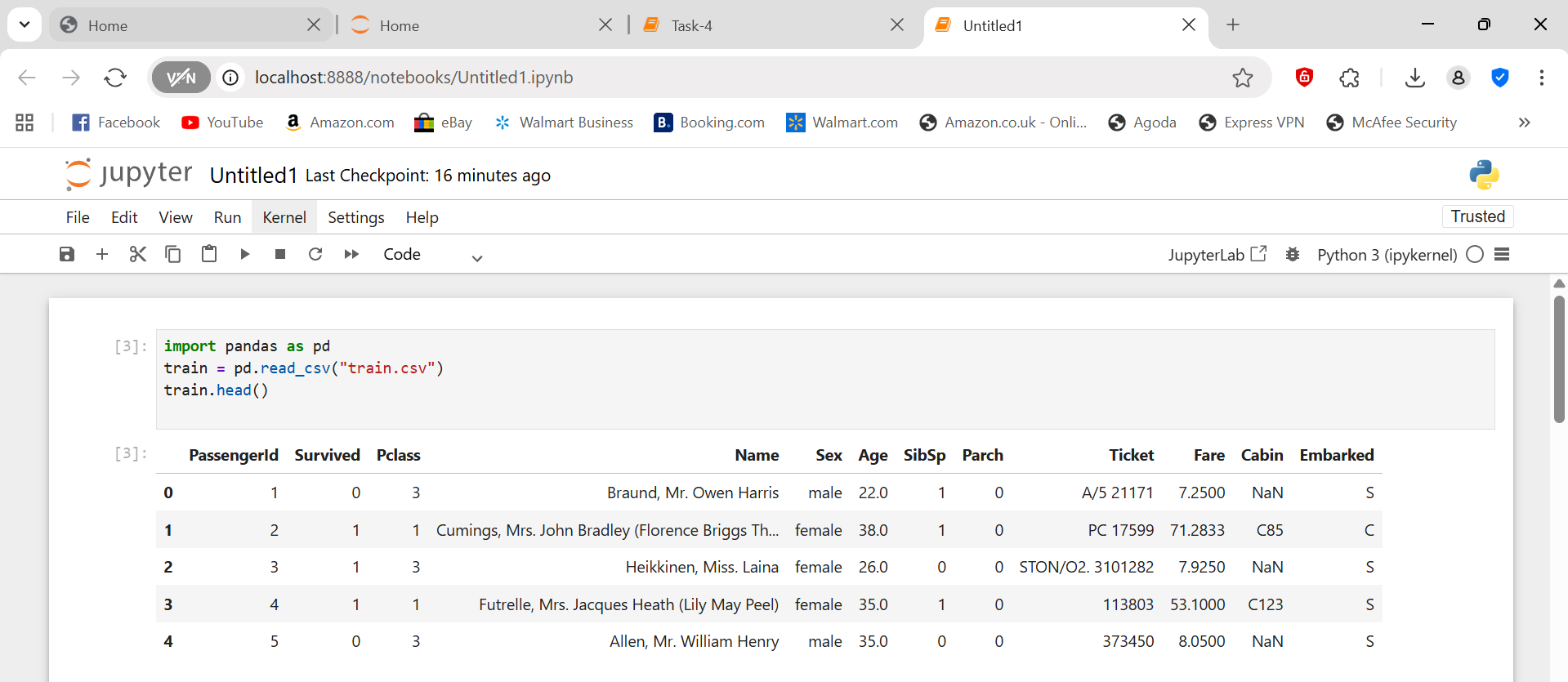
plt.show()

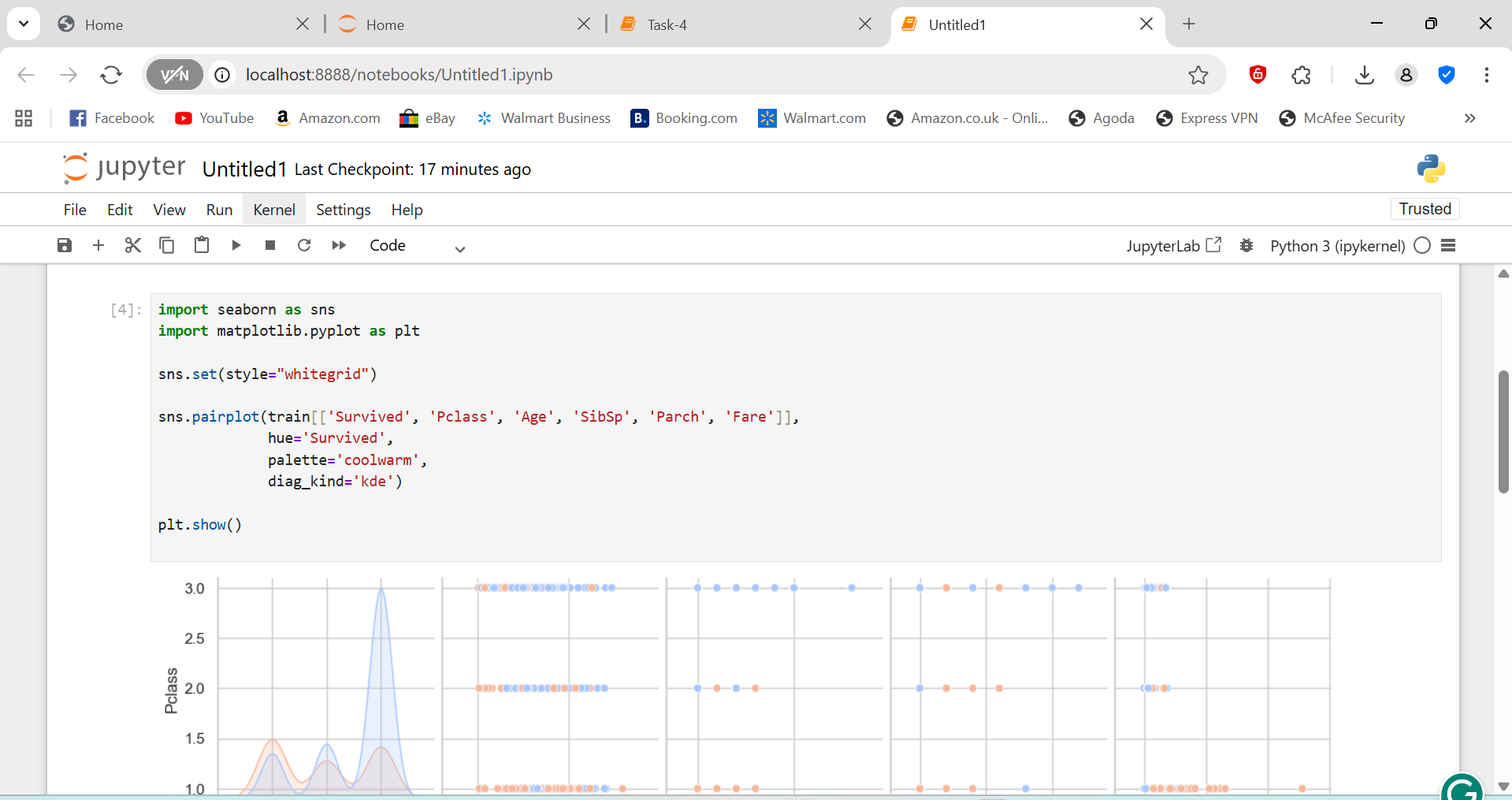
Observation:

- Passengers who paid higher fares had a greater chance of survival.

- Younger passengers and those with fewer siblings/spouses (SibSp) had a slightly better survival rate.

- Class (Pclass) and Fare show a clear relationship: higher class = higher fare = better survival.









**Correlation Heatmap**

After your pairplot, add this new code cell:

# Compute correlation matrix

corr = train.corr(numeric\_only=True)

# Set up the matplotlib figure

plt.figure(figsize=(10, 6))

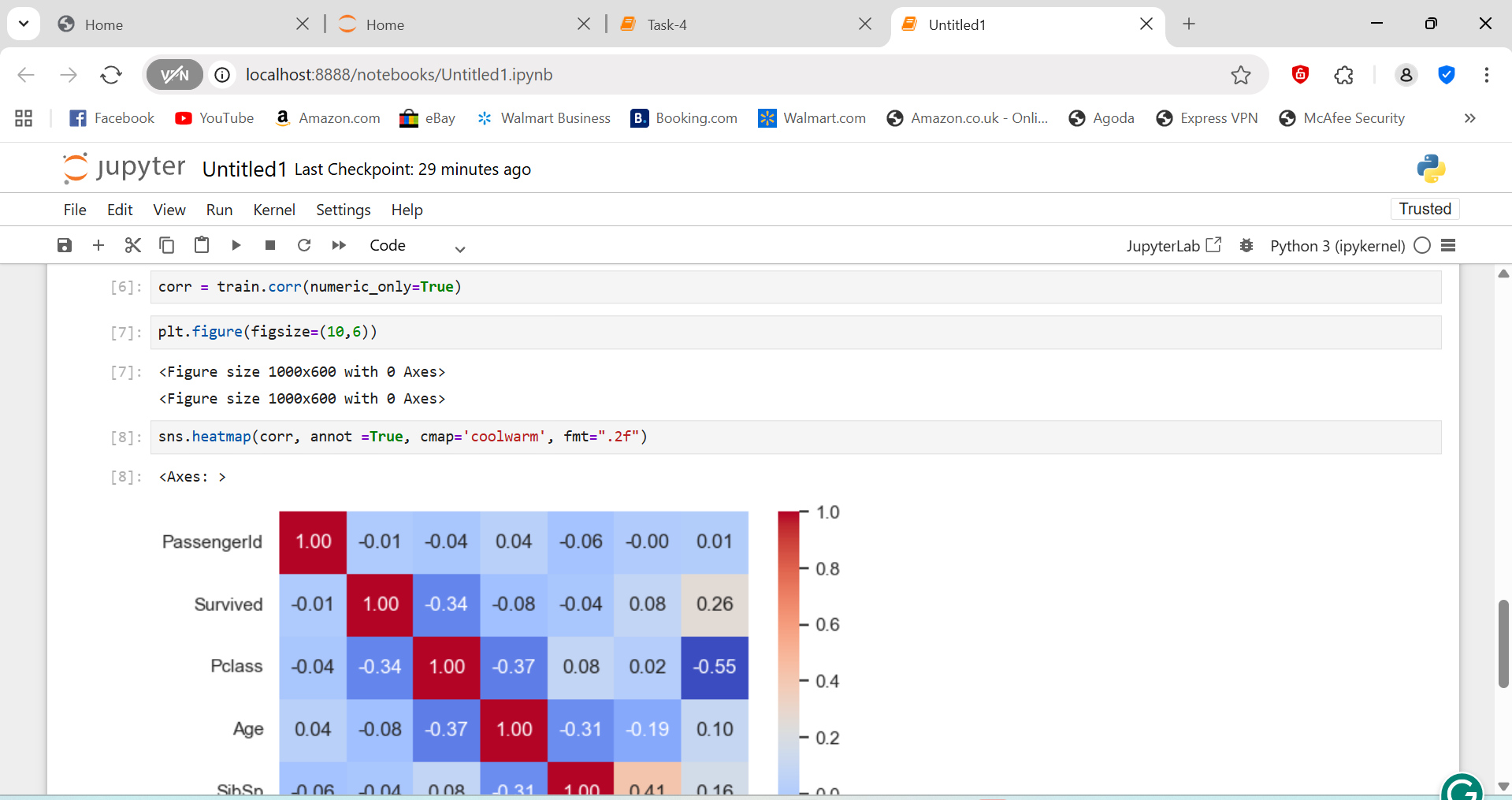
# Create a heatmap

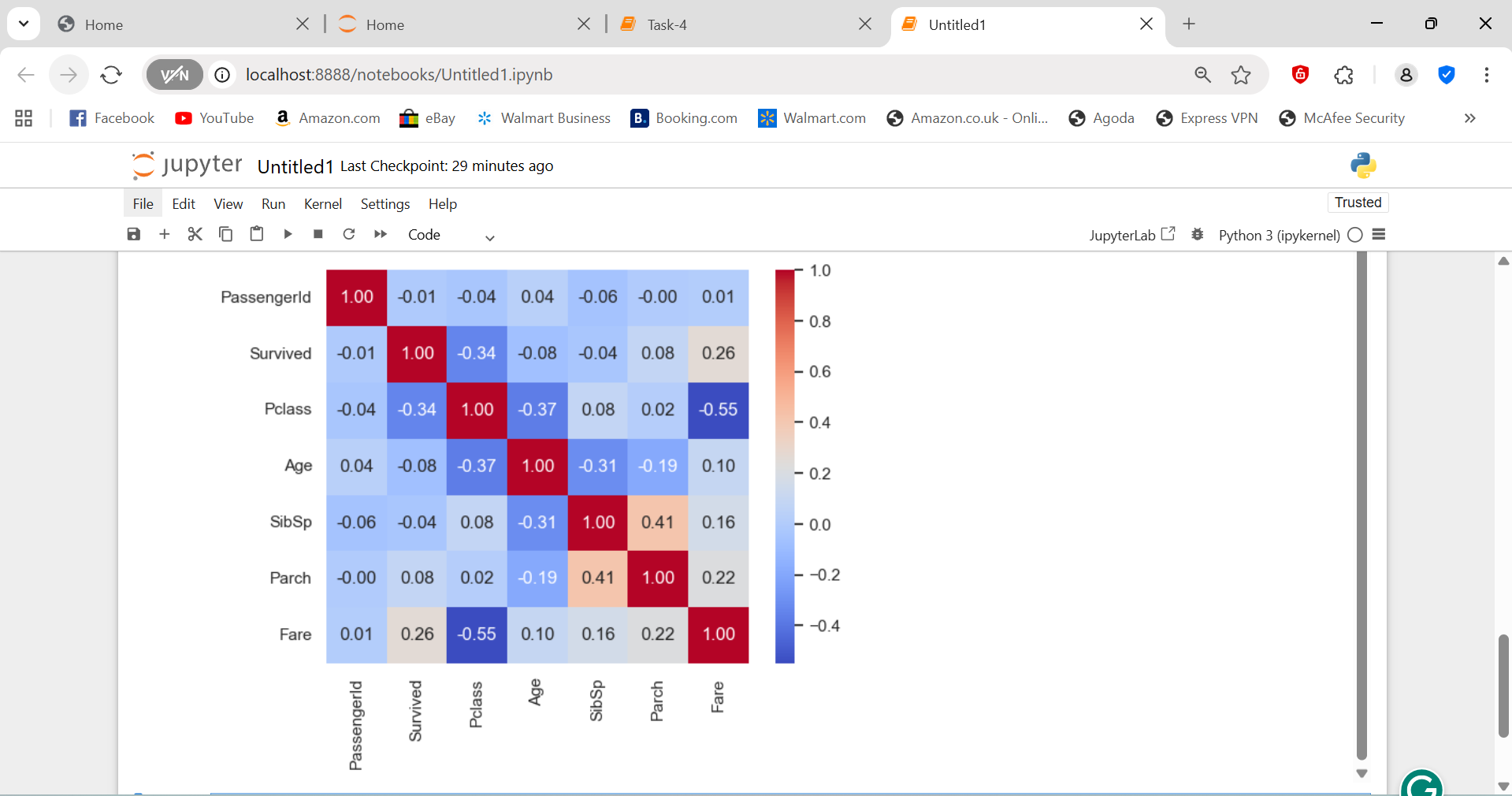
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")

# Add title

plt.title("Correlation Heatmap of Titanic Dataset", fontsize=14)

plt.show()

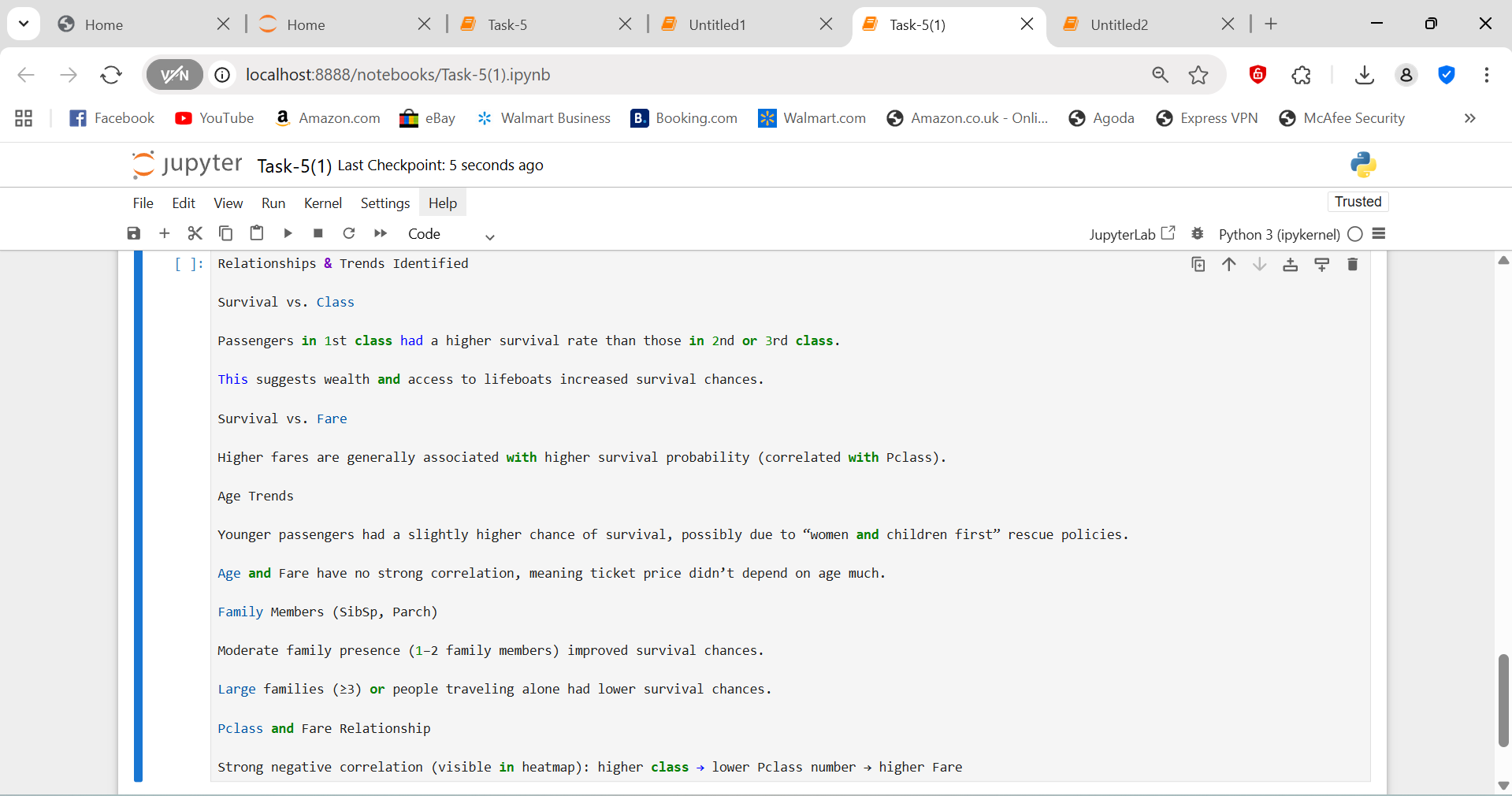




c.Identify relationships and trends

**Relationships & Trends Identified**

1. **Survival vs. Class**
   * Passengers in **1st class** had a **higher survival rate** than those in 2nd or 3rd class.
   * This suggests wealth and access to lifeboats increased survival chances.
2. **Survival vs. Fare**
   * Higher fares are generally associated with **higher survival probability** (correlated with Pclass).
3. **Age Trends**
   * Younger passengers had a slightly higher chance of survival, possibly due to “women and children first” rescue policies.
   * Age and Fare have **no strong correlation**, meaning ticket price didn’t depend on age much.
4. **Family Members (SibSp, Parch)**
   * Moderate family presence (1–2 family members) improved survival chances.
   * Large families (≥3) or people traveling alone had lower survival chances.
5. **Pclass and Fare Relationship**
   * Strong **negative correlation** (visible in heatmap): higher class → lower Pclass number → higher Fare.



**Plot Histograms, Boxplots, and Scatterplots**

**Step 1 — Histogram: Distribution of Age and Fare**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

train = pd.read\_csv("train.csv")

train.head()

# Histogram for Fare

plt.figure(figsize=(8,5))

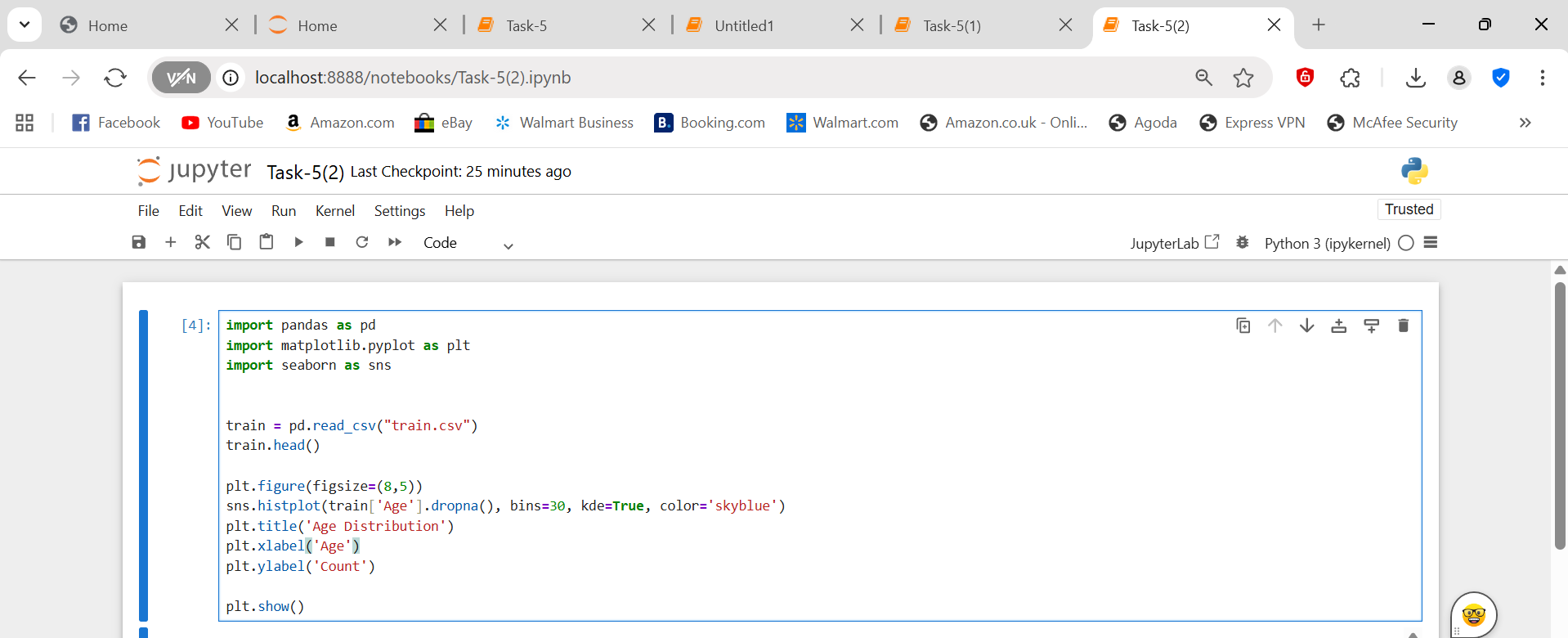
sns.histplot(train['Fare'], bins=30, kde=True, color='orange')

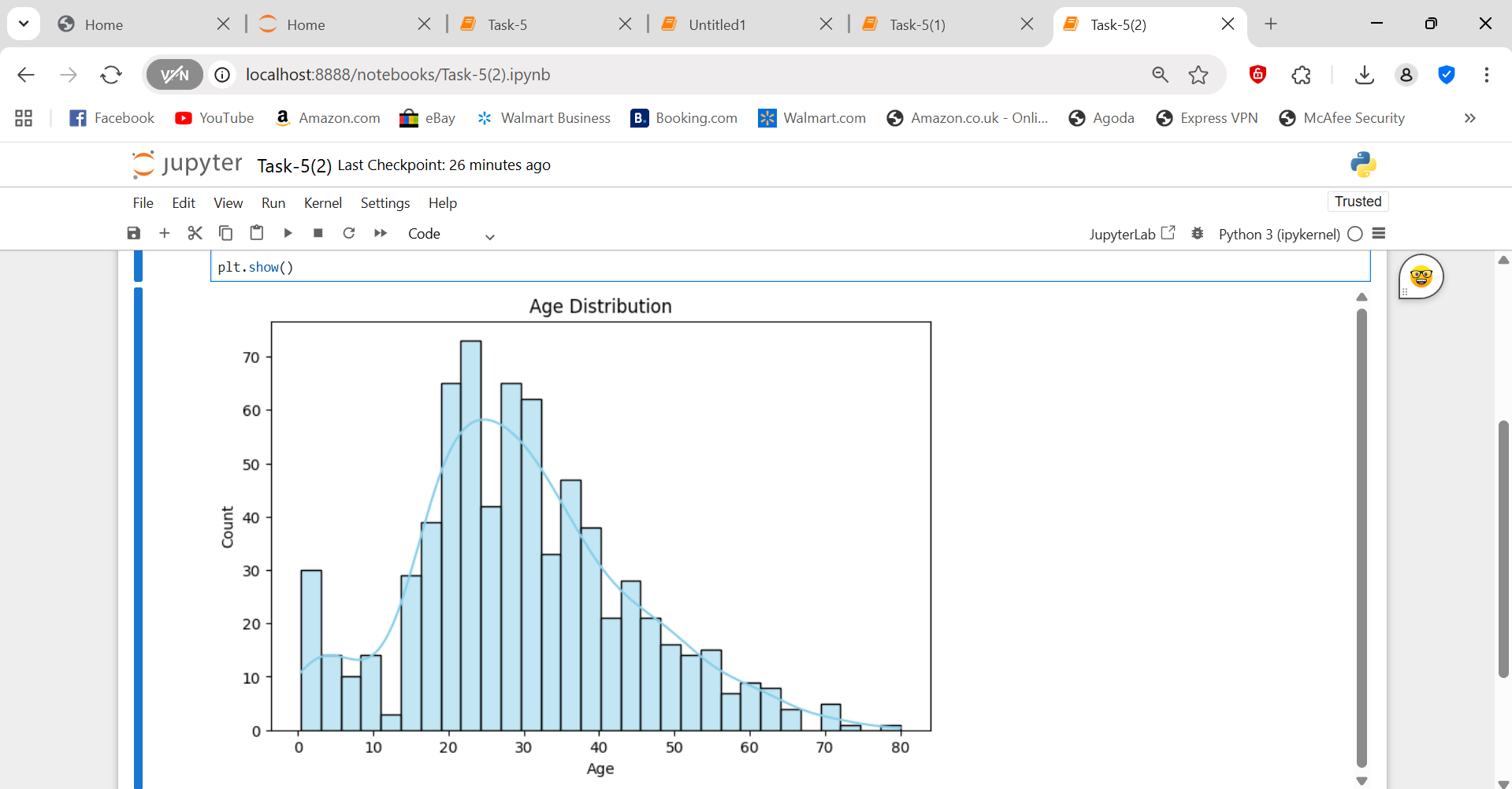
plt.title('Fare Distribution')

plt.xlabel('Fare')

plt.ylabel('Count')

plt.show()





Observation

* Most passengers are between **20–40 years old**.
* Most fares are under **₹100** (or the local equivalent) — showing a majority of economy-class passengers.

**Step 2 — Boxplot: Survival vs. Age**

plt.figure(figsize=(8,5))

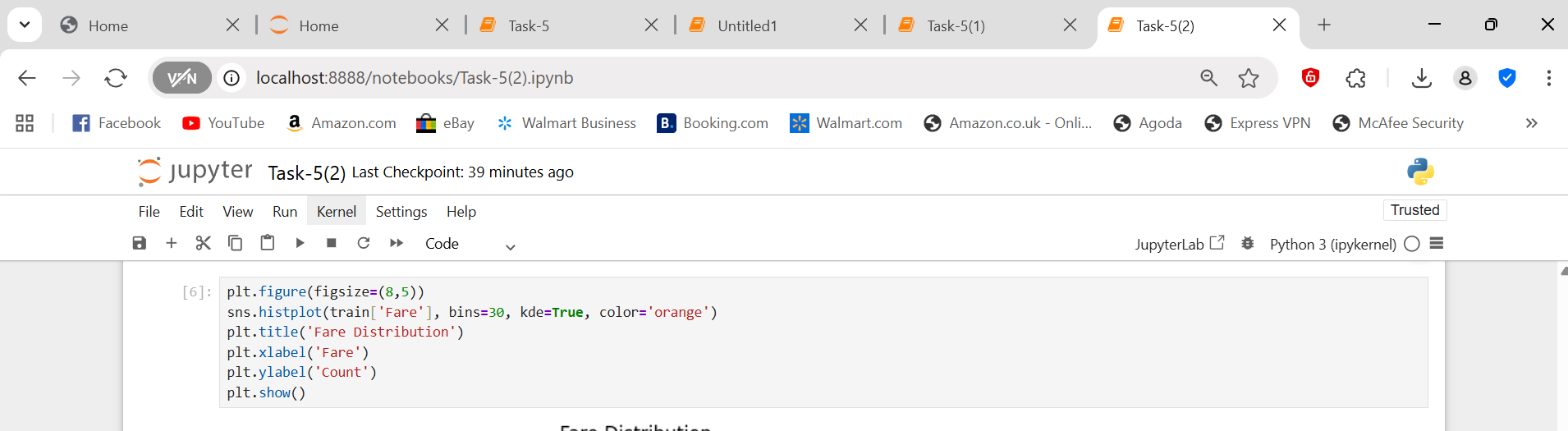
sns.boxplot(x='Survived', y='Age', data=train, palette='coolwarm')

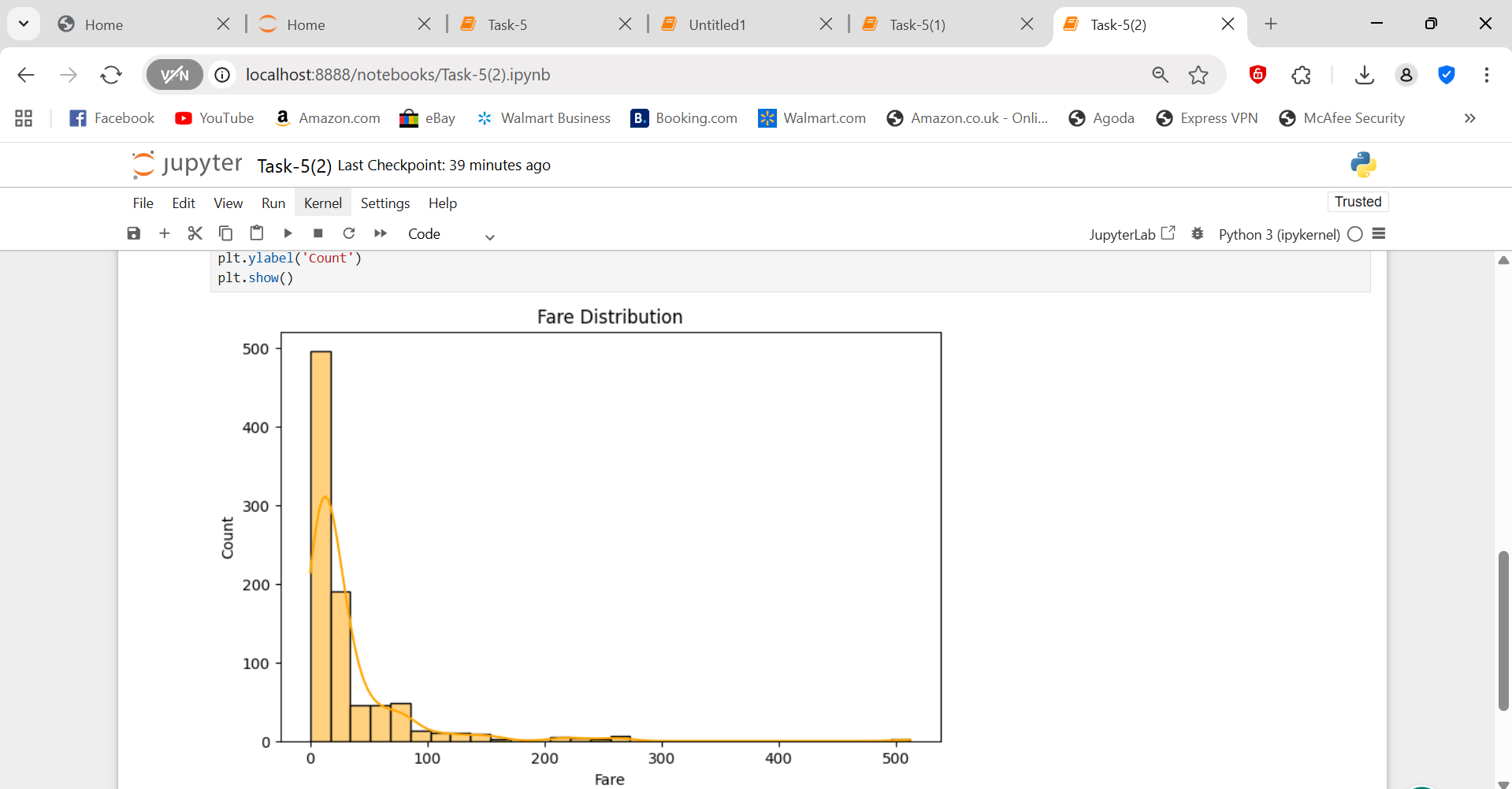
plt.title('Survival vs Age')

plt.xlabel('Survived (0 = No, 1 = Yes)')

plt.ylabel('Age')

plt.show()





**Observation:**

* Median age of survivors is slightly **lower** than non-survivors.
* Younger passengers tended to survive more.

**Step 3 — Boxplot: Survival vs. Fare**

plt.figure(figsize=(8,5))

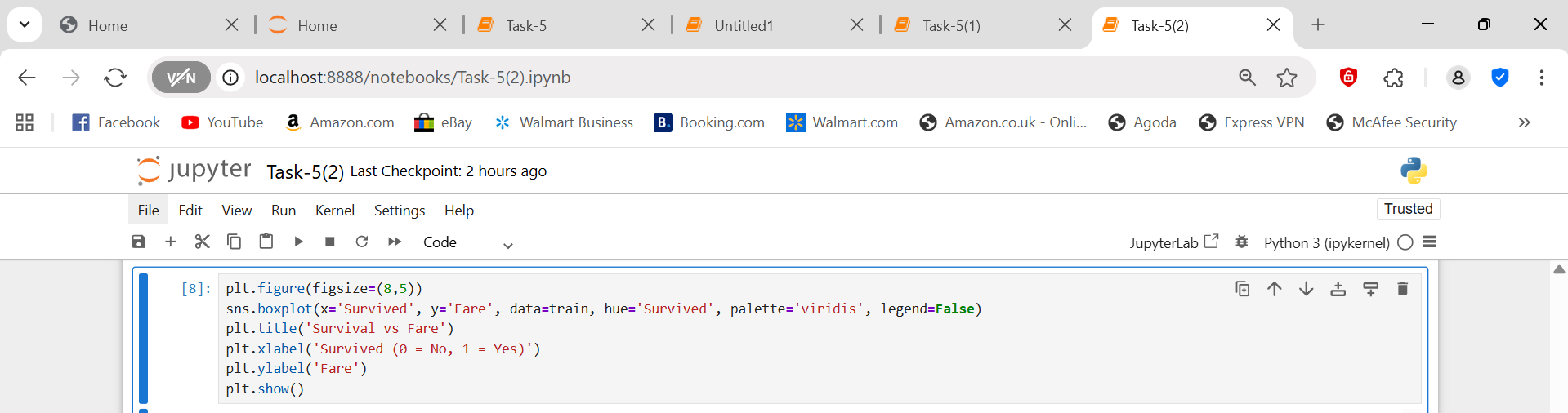
sns.boxplot(x='Survived', y='Fare', data=train, palette='viridis')

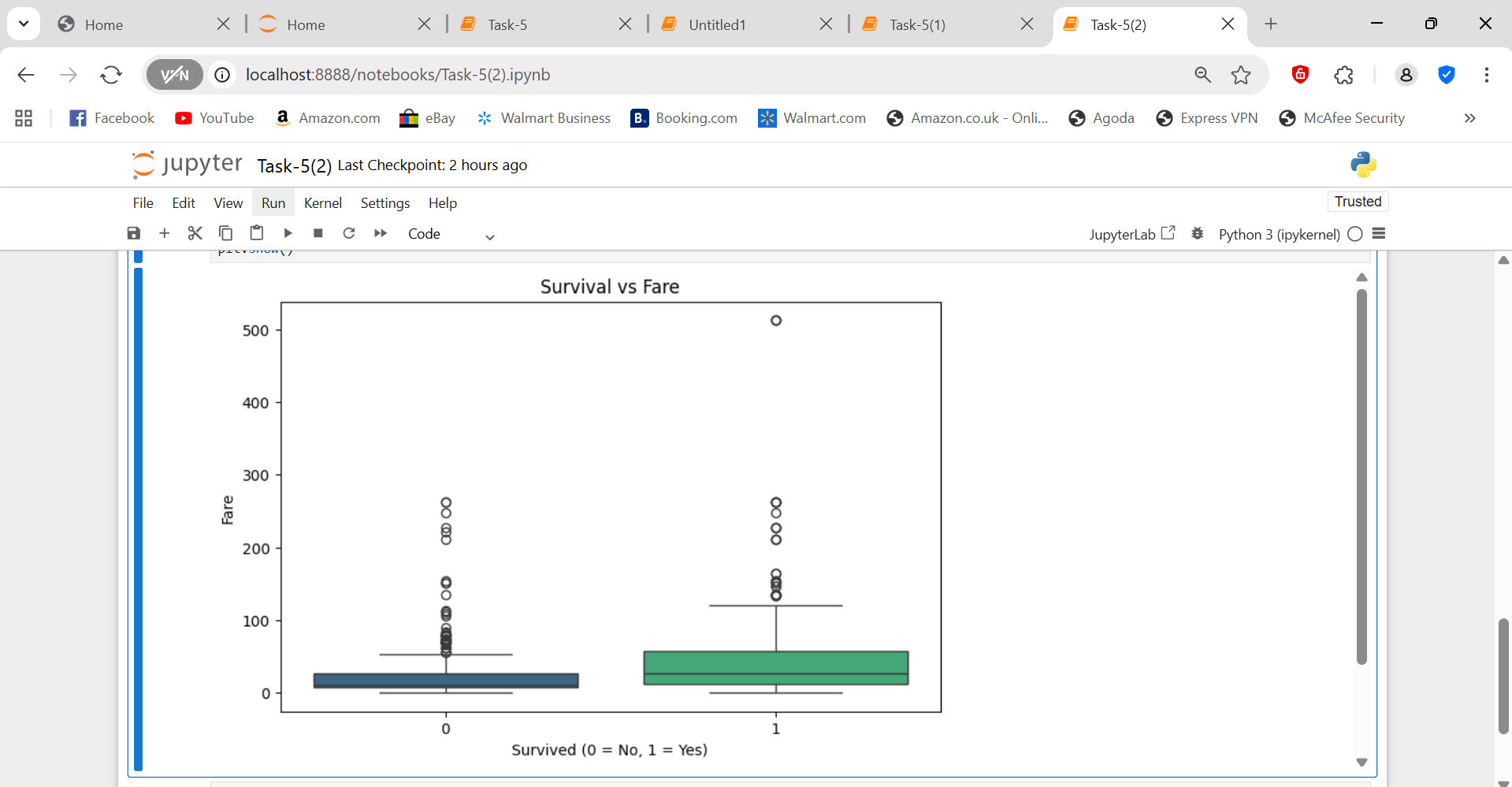
plt.title('Survival vs Fare')

plt.xlabel('Survived (0 = No, 1 = Yes)')

plt.ylabel('Fare')

plt.show()





**Observation:**

* Survivors generally paid **higher fares** → corresponds to **higher-class cabins**.

**Step 4 — Scatterplot: Fare vs. Age (colored by Survival)**

plt.figure(figsize=(8,6))

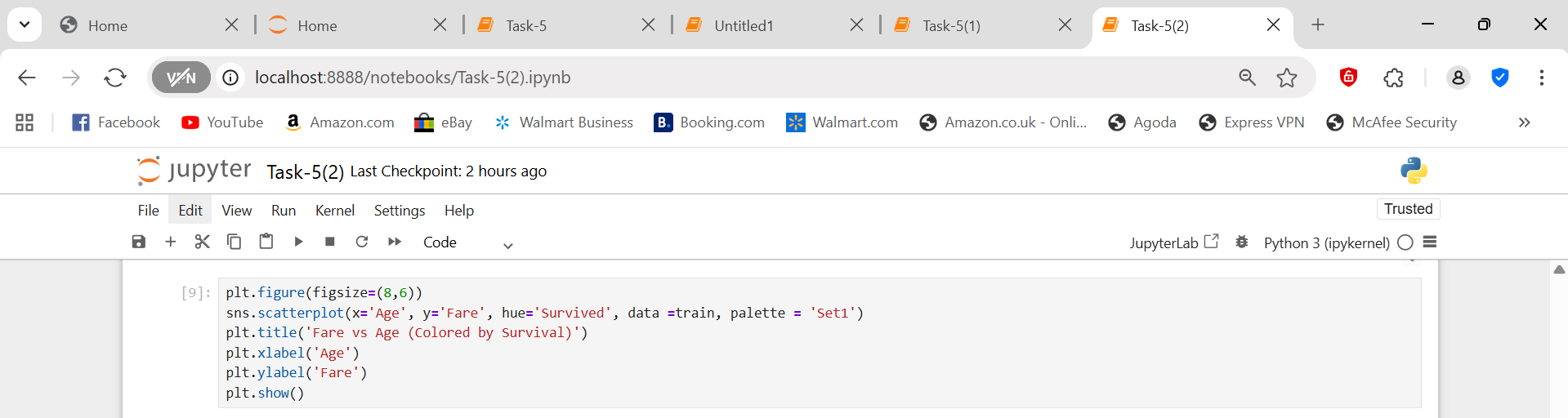
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=train, palette='Set1')

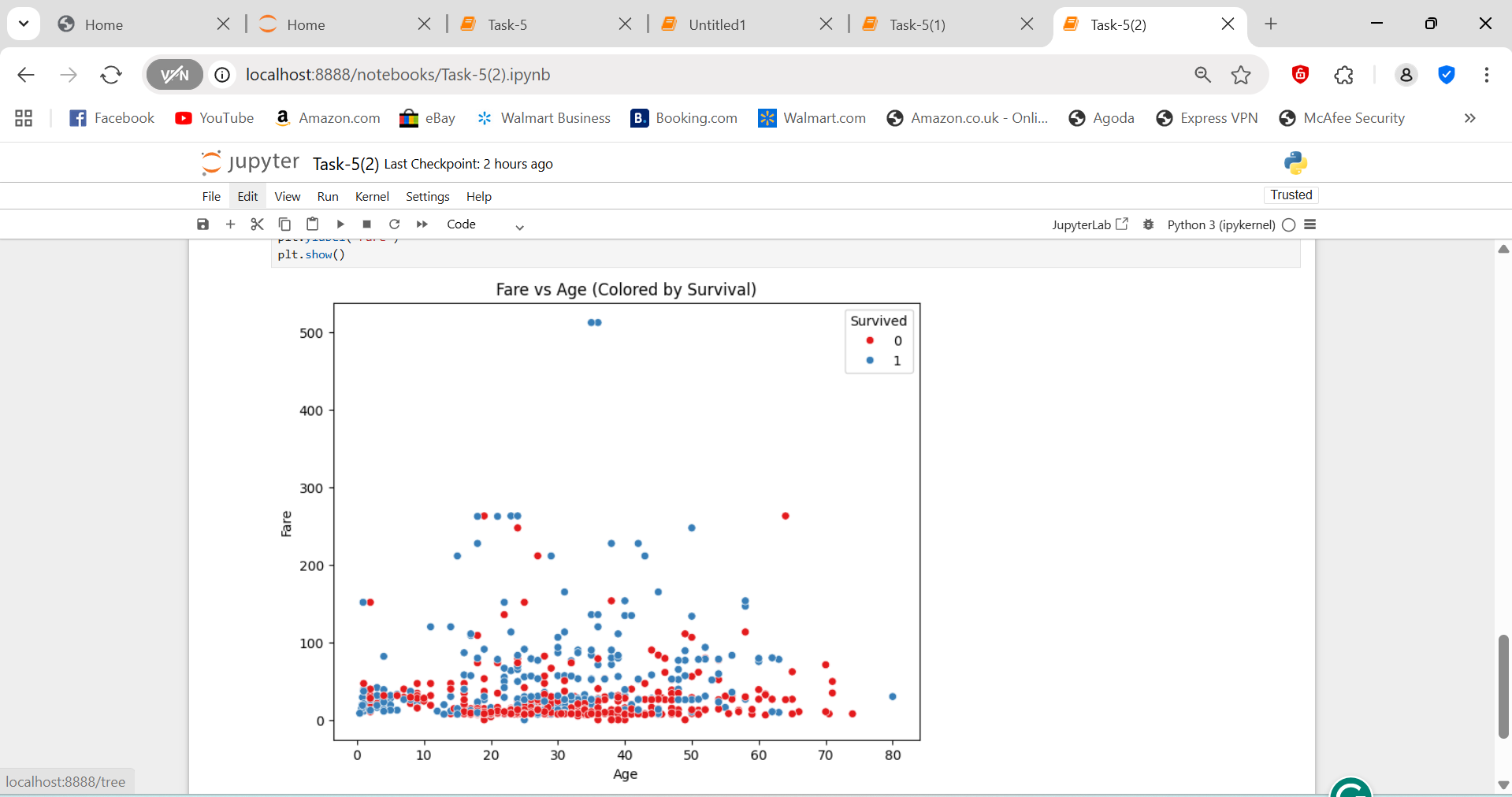
plt.title('Fare vs Age (Colored by Survival)')

plt.xlabel('Age')

plt.ylabel('Fare')

plt.show()





**Observation:**

* Younger and high-fare passengers (top-right corner) show more red dots (Survived = 1).
* Indicates correlation between **wealth + youth → survival**.

**e.Write observations for each visual**

**Observations for Each Visualization**

1. **Survival vs Gender**

* **Observation:** Females had a much higher survival rate compared to males.
* **Insight:** This suggests that during evacuation, women were given priority (“women and children first”).
* **Evidence:** Most female passengers are represented in the “Survived” category.

1. **Survival vs Passenger Class (Pclass)**

* **Observation:** Passengers in **1st class** had the highest survival rate, followed by 2nd class.
* **Insight:** Socio-economic status played a key role — wealthier passengers likely had better access to lifeboats.
* **Evidence:** The majority of 3rd-class passengers did not survive.

1. **Age Distribution by Survival**

* **Observation:** Younger passengers (especially children) had a higher survival probability than middle-aged or older passengers.
* **Insight:** Families and younger individuals were prioritized during rescue operations.
* **Evidence:** The density curve for survivors peaks in the lower age range.

1. **Survival vs Embarkation Port**

* **Observation:** Passengers who boarded at **Cherbourg (C)** had a higher survival rate than those from Southampton (S) or Queenstown (Q).
* **Insight:** This may correlate with passenger class distribution — more first-class passengers embarked at Cherbourg.
* **Evidence:** Countplot shows a higher proportion of survivors for port ‘C’.

1. **Survival vs Fare (Boxplot)**

* **Observation:** Passengers who paid higher fares generally had better survival outcomes.
* **Insight:** Higher fares correlate with first-class cabins, which had more lifeboat access.
* **Evidence:** The median fare for survivors is visibly higher.

**f.Provide summary of findings**

**Summary of Findings – Titanic Dataset Analysis**

1. **Overview**

The analysis aimed to explore patterns in survival among passengers on the Titanic using various demographic and socio-economic variables such as gender, class, age, and fare.  
The dataset consisted of records from the **Titanic passenger manifest**, with details on survival status, passenger class, gender, age, family members aboard, and ticket fares.

1. **Key Insights**

| **Factor** | **Observation** | **Impact on Survival** |
| --- | --- | --- |
| **Gender** | Females had a significantly higher survival rate than males. | Strong Positive |
| **Passenger Class (Pclass)** | 1st class passengers survived the most; 3rd class the least. | Strong Positive |
| **Fare** | Higher fare correlated with higher survival. | Moderate Positive |
| **Age** | Younger passengers (especially children) had slightly better survival. | Mild Positive |
| **Embarkation Port** | Passengers from Cherbourg (C) survived more often. | Moderate Positive |
| **Family Aboard (SibSp, Parch)** | Having 1–2 relatives increased survival chance slightly; too many reduced it. | Mixed Impact |

1. **Relationships Observed**

* **Gender and Class** were the two most influential predictors of survival.
* **Fare** served as a proxy for socio-economic status and access to lifeboats.
* **Heatmap correlations** showed moderate positive relationships between *Fare* and *Survived*, and negative correlations between *Pclass* and *Survived*.
* **Pairplots** visually confirmed these relationships — survivors cluster around lower class numbers (1st class) and higher fares.

1. **Visual Summary**

* **Pairplot:** Highlighted clear separation of survivors by fare, age, and class.
* **Heatmap:** Demonstrated feature correlations; “Pclass” negatively correlates with survival.
* **Histplot:** Revealed more non-survivors across most age groups but a visible young survivor cluster.
* **Boxplot:** Displayed higher fares for survivors, confirming socio-economic influence.

1. **Conclusion**

* **Survival on the Titanic was not random.**
* Passengers who were **female, younger, and from higher classes** had a much higher chance of survival.
* Socio-economic privilege and evacuation protocols (favoring women and children) directly influenced outcomes.
* These findings align historically with documented events of the Titanic tragedy.