Day 2

Question given to us:-

```
Let X denote the number of defective items in a shipment of
7600 items, and assume that the probability of an item being
defective is p = 5%. What is the distribution of X, and what
is
the expectation, variance, and standard deviation for this
distribution?
```

1. The distribution should be binomial, since it is a collection of Bernoullis where each event is either defective or not. The x-axis will be 0->7600, the y-axis is probability that the xth value number of defective items show up.
2. Expected value will be 380 because in the problem itself it was stated that 5% will be defective

Things to look at during my own time.(EDA)

The below is used to clean data up:-

Confusion matrix in data science/probability.
=> confusion matrix is a way of validating a machine learning model, it gives us total number of correct and incorrect predictions by comparing the predicted values vs the actual reality values.

Correlation of data in a correlation matrix.
=> Correlation matrix is a matrix of how correlated certain variables are to each other (look at this chat:- https://chatgpt.com/share/6817f04c-3810-8009-83ff-e262d08c486a)

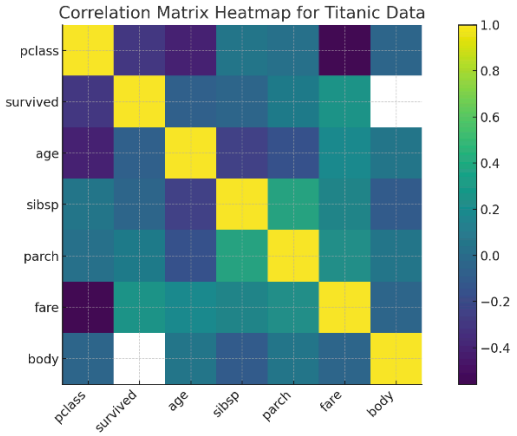
Here is an example of how a dataset's correlation data between each other was converted to a matrix.

Thought for a couple of seconds

Titanic Dataset Correlation Matrix

#		pclass	survived	age
1	pclass	1.0	-0.31246936264967645	-0.40810573614103
2	survived	-0.31246936264967645	1.0	-0.05551252019214
3	age	-0.40810573614103396	-0.05551252019214636	1.0
4	sibsp	0.06083200757490756	-0.02782511923058187	-0.2436994877410
5	parch	0.01832220200978667	0.08265957038609871	-0.1509174344799
6	fare	-0.5586287323271729	0.24426546891481193	0.17873932314646

Correlation Matrix Heatmap for Titanic Data



Correlation between variables is calculated by the Pearson's r method

What is pca

=> PCA is the process of transforming a dataset to be smaller but still hold the same meaning, this helps in processing and classification of data downstream.

Coverage of dataset:- How much the dataset accurately represents the real world.

Try all types of graphs to present your data, and use the one which makes the most intuitive sense.

What is EDA

=>Exploratory Data Analysis (EDA) is the process of getting to know your data—its structure, patterns, anomalies, and relationships—before you jump into modeling or formal statistical inference. A solid EDA helps you:

- **Understand** the variables (data types, ranges, distributions)
- **Detect** data quality issues (missing values, outliers, duplicates)
- **Reveal** initial patterns and correlations
- **Formulate** hypotheses and guide feature engineering

This is how you do EDA-> https://chatgpt.com/share/681808e4-a200-8009-9056-ecd61422aec5

Today's assignments

Assignment1:- Scrape indeed, to scrape roles for any role with the state and city as MA, Boston. The script should scrape the first 5 pages, and store the roles in a csv file.(done look at day2assignment1.py)

Assignment2:- A dataset will be given, do an EDA on it.

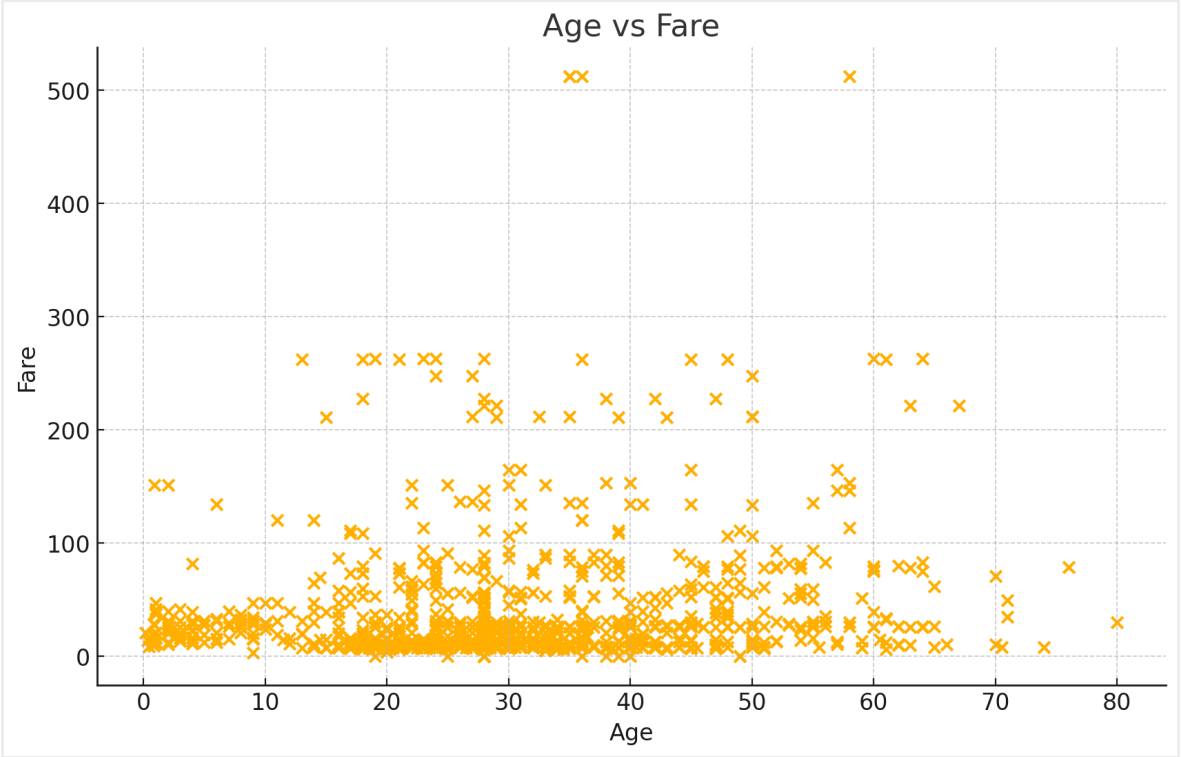
Dataset:-

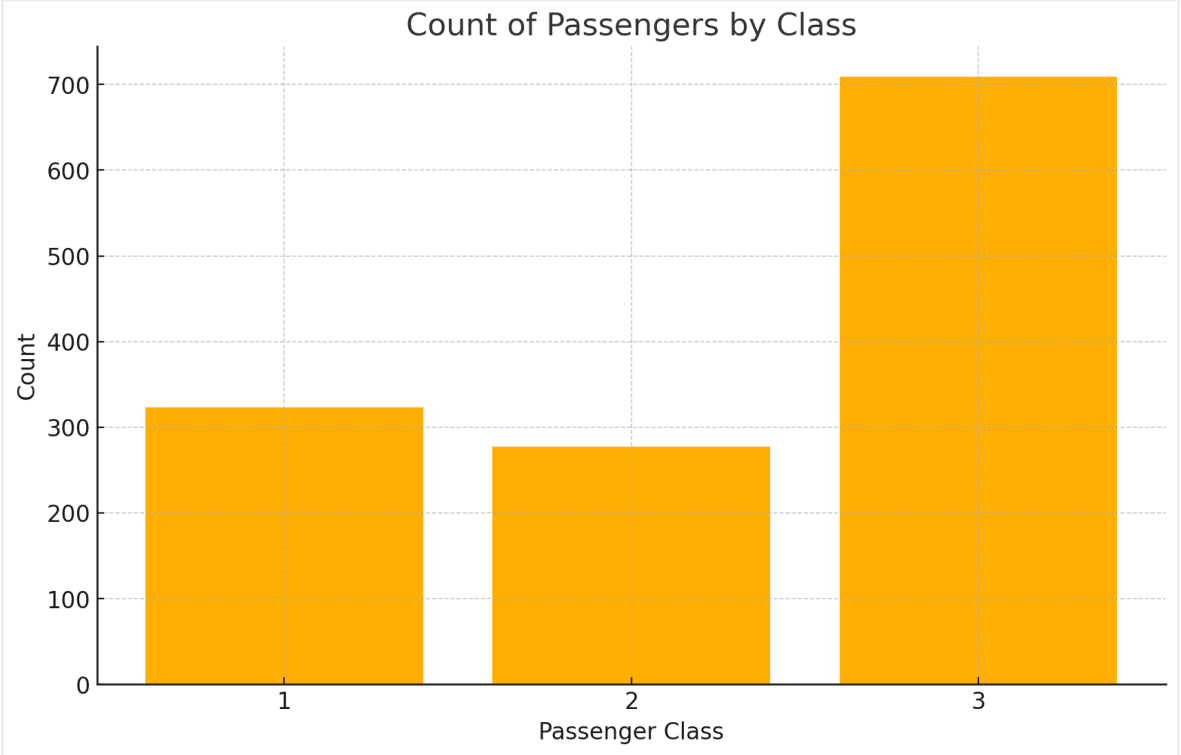
Titanic Data.csv
Text Document · 107 KB

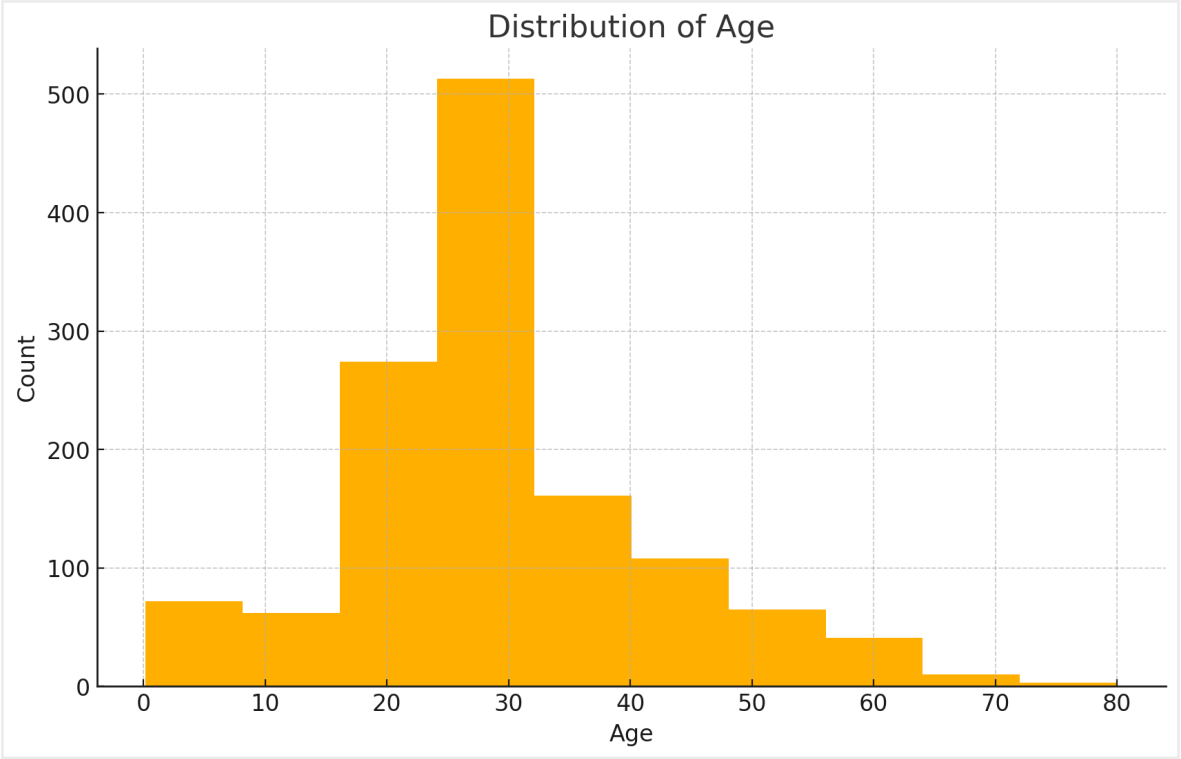
1. Filled dataset's missing values with median of the same column:-

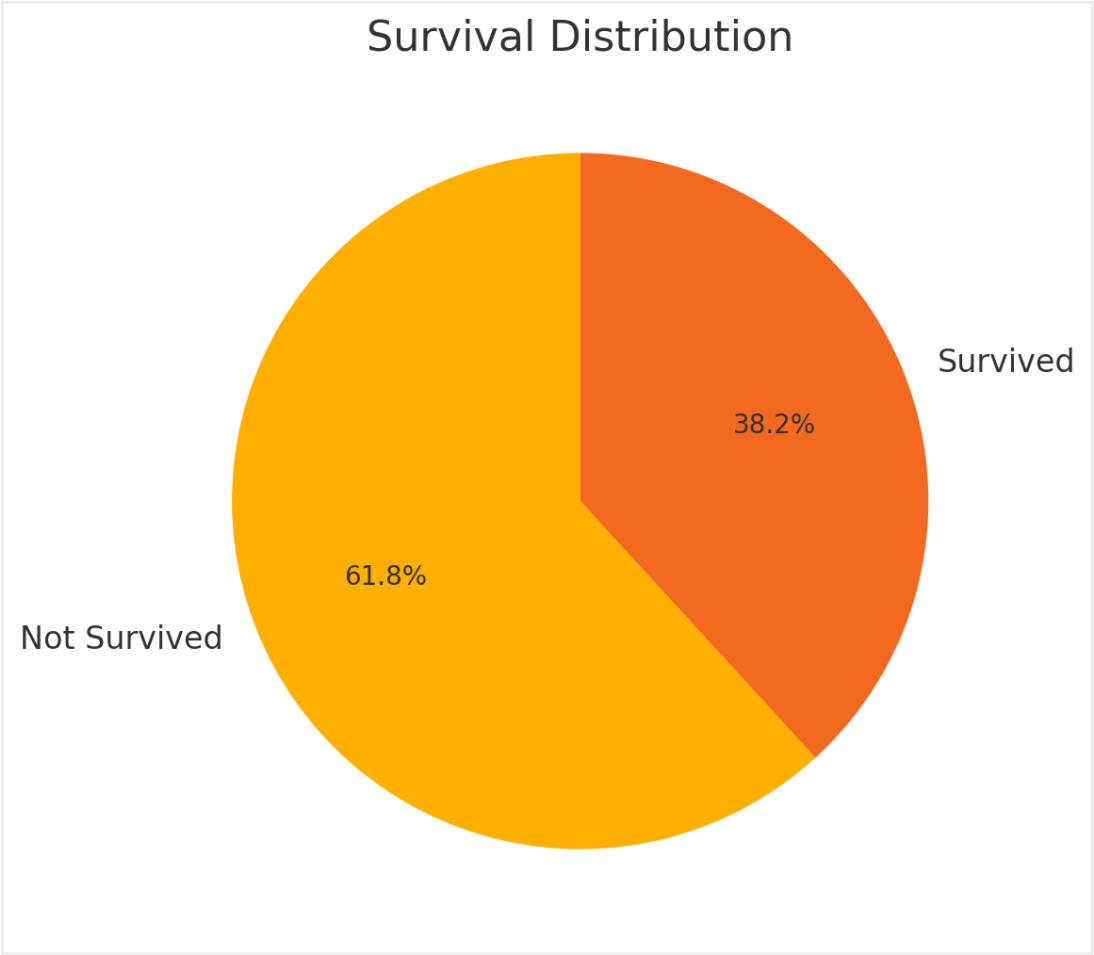
Titanic_Data_filled.csv
Text Document · 135 KB

2. Representing the data in different ways using different graphs


Age vs Fare
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80


Count of Passengers by Class
Count
Passenger Class
0
100
200
300
400
500
600
700
1
2
3



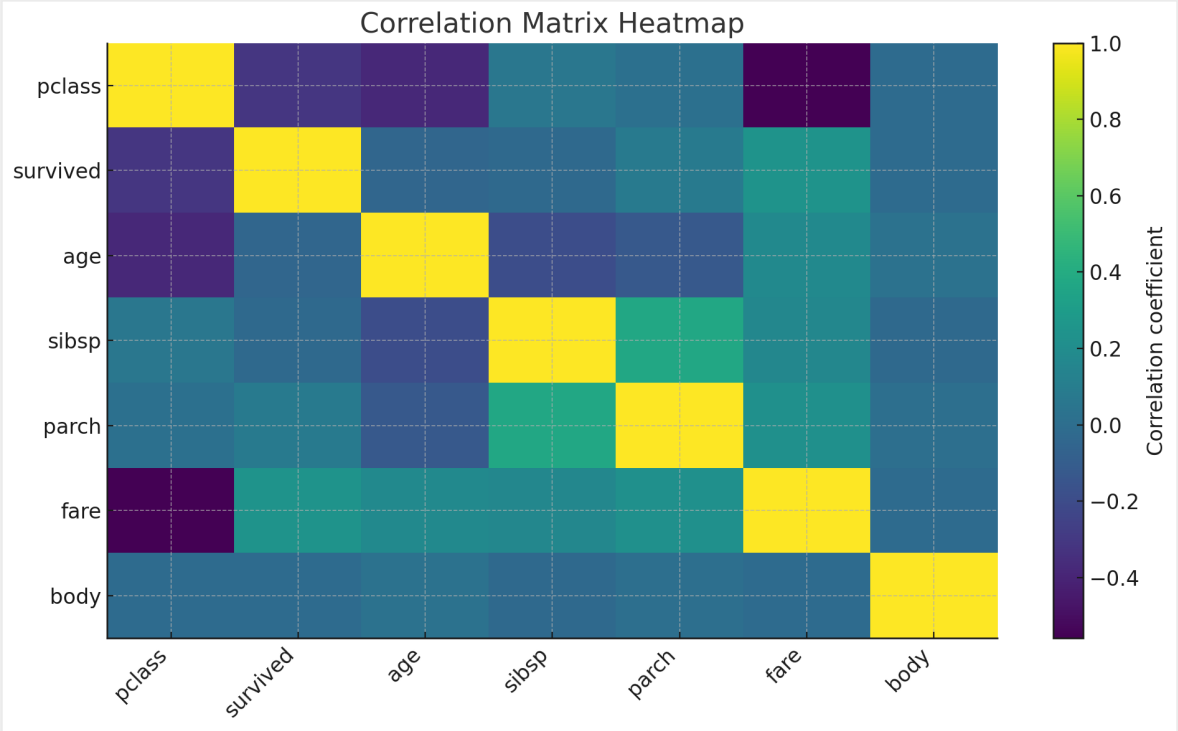


3. Correlation matrix heatmap analysis

Correlation csv:-

Titanic_Dataset_Correlation_Matrix.csv
Text Document · 981 bytes

Correlation heat map:-



From this correlation heat map we can tell the following:-

1. The higher the fare the higher the chance of survival, because the 2 datapoints are mildly correlated.
2. The higher the age the higher the fare, as seen by the mild correlation between the 2 datapoints
3. The higher the sibps the higher the parch as seen by the 0.37 correlation between the 2
4. The higher the parch the higher the fare