

# IDAI 610: Problem Set 4

## Exploratory Data Analysis, PCA and Batch Gradient Descent

Shubh Sudan  
Rochester Institute of Technology  
ss2401@rit.edu

Matthew Landon  
Rochester Institute of Technology  
ml3275@g.rit.edu

November 16, 2025

### **PART 1: Exploratory Data Analysis**

#### **Q1) Dataset Dimensions**

Dataset Name: Titanic.csv. Dimensionality Of Data:

A.Rows:891 - These contains the details of the passengers aboard the Titanic ship.

B.Columns: 12 - These contain the specific feature columns such as Name, Gender, Age and where have the passengers embarked to (C = Cherbourg, Q = Queenstown, S = Southampton).

#### **Q2) Statistical Summary**

a) Max Age: 80 Years.

Min Age: 0.42 (A baby of 3 Months).

Average Age : 29.69 (29 Years and roughly 7 months).

b) Median Fare: 14.4 Pounds.

Mean Fare: 32.2 Pounds.

c) Yes - the data is skewed - mainly because the standard deviation is very high - and the maximum fare value is an outlier at 512.3.

The data is skewed towards the right - because the value of Q3 and Q4 is increasing.

d) Standard Deviation for Fare: 49.69.

e) Age - mainly because the values are consistent - but the standard deviation value is the greatest meaning that there is a sharp peak formation for the Fare column.

### Q3 Null Values Check

The dataset has mainly 3 columns where null values are present. The names of the columns are:

A) Age

B) Cabin - Most Missing Entries (77.1%)

C) Embarked

	Columns	Count	missing_percent
0	PassengerId	0	0.000000
1	Survived	0	0.000000
2	Pclass	0	0.000000
3	Name	0	0.000000
4	Sex	0	0.000000
5	Age	177	19.865320
6	SibSp	0	0.000000
7	Parch	0	0.000000
8	Ticket	0	0.000000
9	Fare	0	0.000000
10	Cabin	687	77.104377
11	Embarked	2	0.224467

Figure 1: Null Values In The Titanic Dataset

## Problem 2

### Q4) Exploratory Univariate Analysis

For each variable we plotted the required histogram and boxplot (for numeric features) or bar chart/pie chart (for categorical features).

Figure 2 shows the histogram and boxplot for **Age**. Sample size:  $n \approx 714$  non-missing values.

The distribution is unimodal and centered on young adults, with noticeable density for children and a mild right skew. A few outliers occur in the range 70–80. The mean and median are close, suggesting modest skewness.

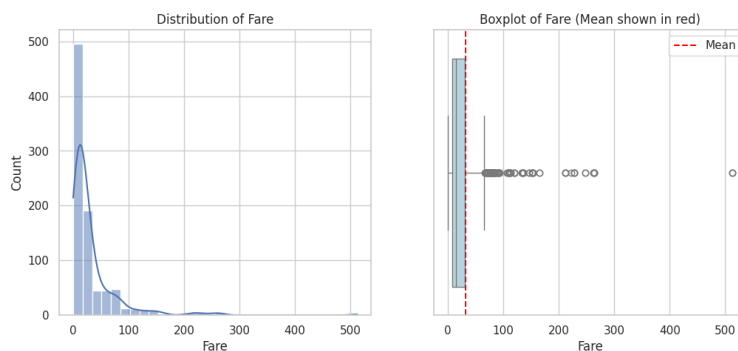


Figure 2: Histogram and boxplot of Age. Median and mean shown; note mild right skew and outliers.

Figure 3 shows the histogram and boxplot for **Fare**. Sample size:  $n = 891$ .

Fare is strongly right-skewed, with a long heavy tail produced by expensive first-class tickets. Outliers are numerous at high fares. The mean is well above the median.

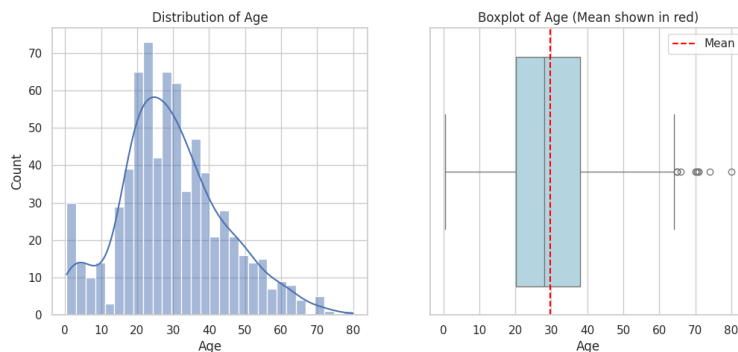


Figure 3: Distribution and boxplot of Fare. Note strong positive skew and substantial high-end outliers.

The bar chart (Figure 4) shows that most passengers are in **Pclass** 3, followed by 1, then 2.

Figure 5 shows the bar chart and pie chart for **Sex**. Males outnumber females.

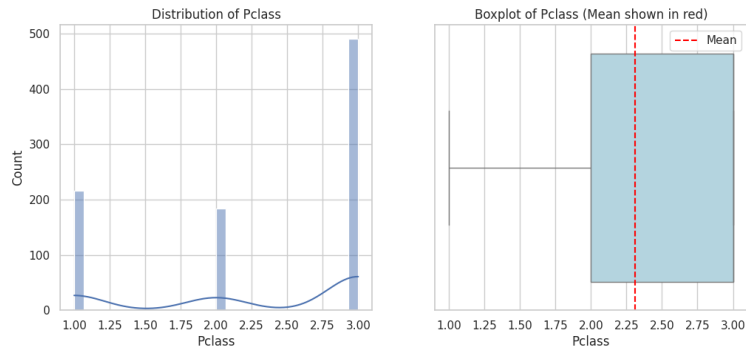


Figure 4: Bar chart of Pclass counts. Third class is the largest group.

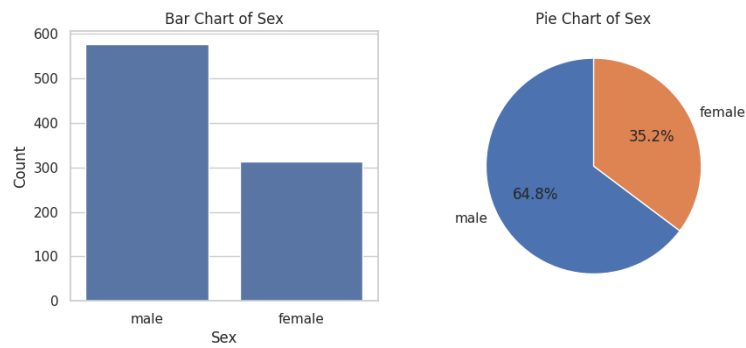


Figure 5: Bar chart and pie chart of Sex. Male passengers are the majority.

Figure 6 gives the bar and pie charts for **Embarked**. Southampton (S) dominates.

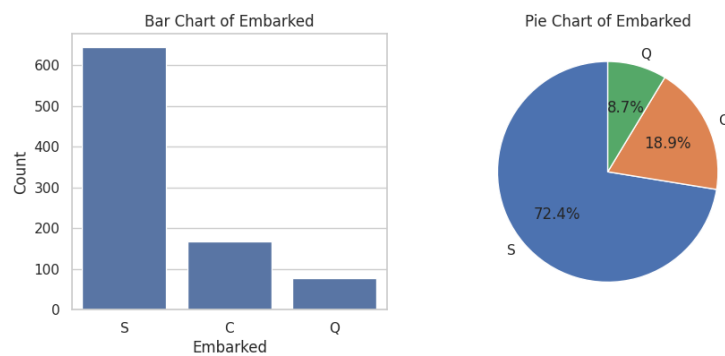


Figure 6: Embarked distribution. Southampton (S) is the most common port.

## Q5) Bivariate Analysis

We plotted the relationships required:

- Age vs Fare (Figure 7),
- SibSp vs Parch (Figure 8).
- The Age–Fare relationship shows no clear monotonic pattern; fare depends heavily on `Pclass` and cabin type.
- SibSp and Parch display a positive trend: large families often bring both siblings/spouses and parents/children.

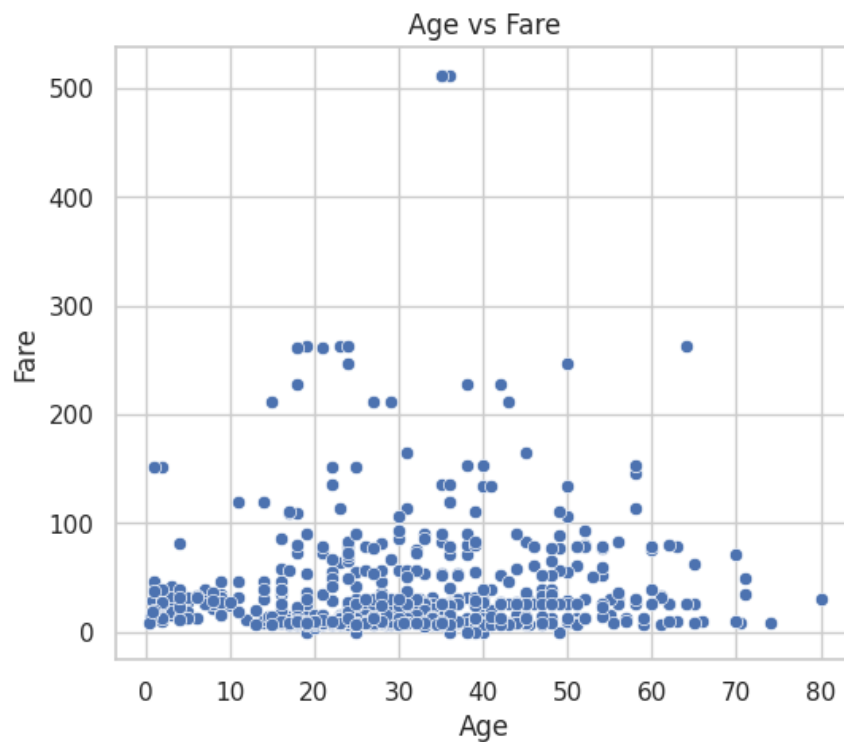


Figure 7: Scatter plot of Age vs Fare. No strong association is visible.

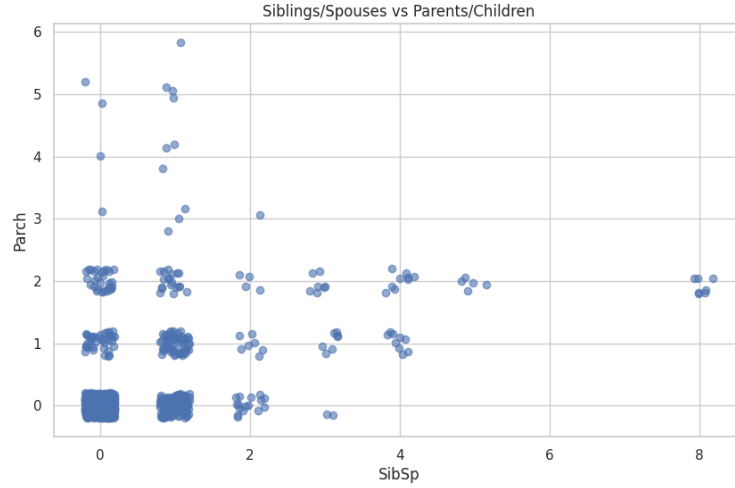


Figure 8: Scatter plot of SibSp vs Parch with jitter. A clear positive association is visible.

Figure 9 displays the Spearman correlation heatmap for numerical features.

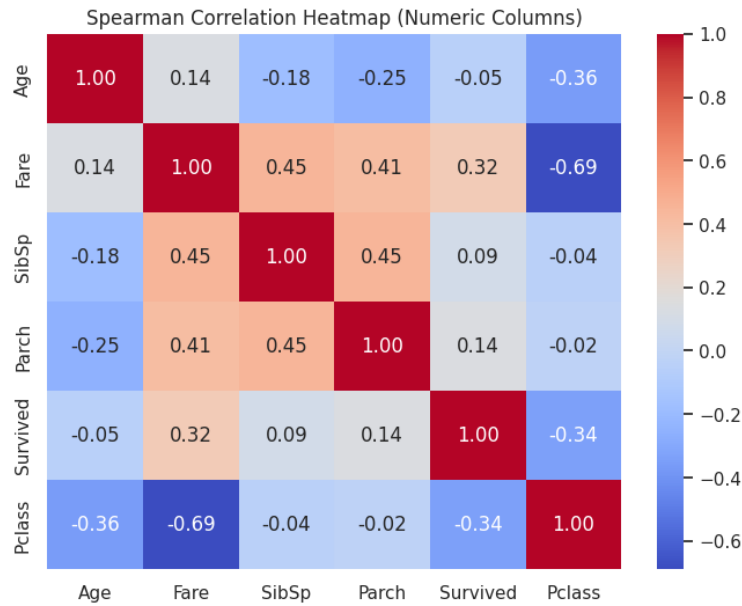


Figure 9: Spearman correlation matrix for numeric features. Highlights monotonic associations.

Spearman's  $\rho$  measures monotonic association based on ranks, making it robust to skew, heteroscedasticity, and extreme outliers. This makes it well suited to the Titanic dataset, where variables such as **Fare** exhibit heavy tails.

## Q6) Grouped / Multivariate Analysis

Figure 10 shows the fraction of `Pclass` within the female subset. Most female passengers were in third class, though a substantial minority traveled in first class.

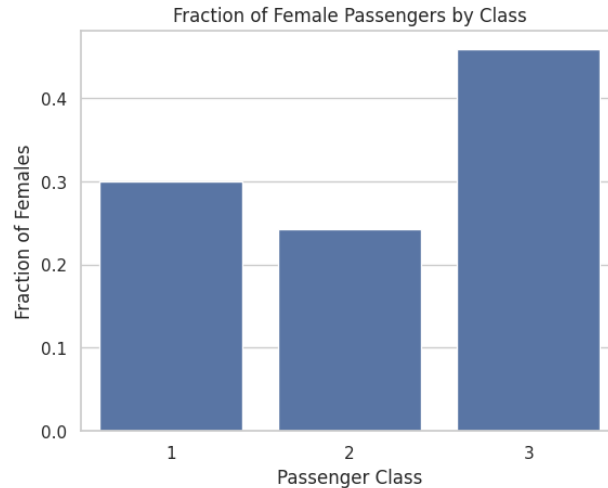


Figure 10: Fraction of female passengers by passenger class.

**Additional grouped visualization (pie chart).** Figure 11 presents a pie chart of survivors by sex. Female passengers constitute the majority of survivors.

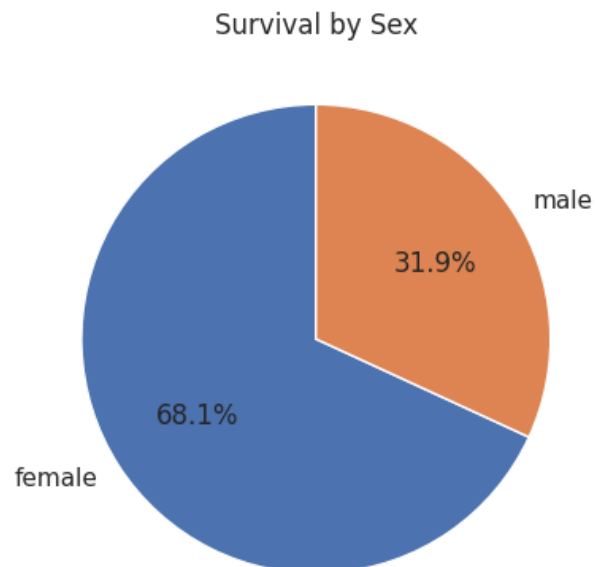


Figure 11: Pie chart of survivors grouped by sex.

## Problem 3 - PART 1

### Q7) Preprocessing Pipeline:

We will go through the steps involved in the preprocessing pipeline below:

1. Analyze the number of null values in each column using the `df.isnull().sum()` function.
2. We check the importance of that particular feature in the data set. Eg: 70% Cabin name dataset is empty - because it is a Categorical feature with not so much significance on the overall methodology, we can simply drop the entire column.
3. We further analyze and remove the Categorical and redundant features such as **'PassengerId', 'Name', 'Cabin', 'Pclass', 'Ticket'**.
4. We were still left with the Age Column with 177 missing entries - what we smartly did was to impute the missing values by using the random sampling which could overcome the addition of bias of mean, median methods.
5. We finally achieved the dataset with a 891 rows and 7 feature columns, where we had 891 rows and 12 columns before.

The image below shows the Distribution change of the Age column after and before the imputation process.

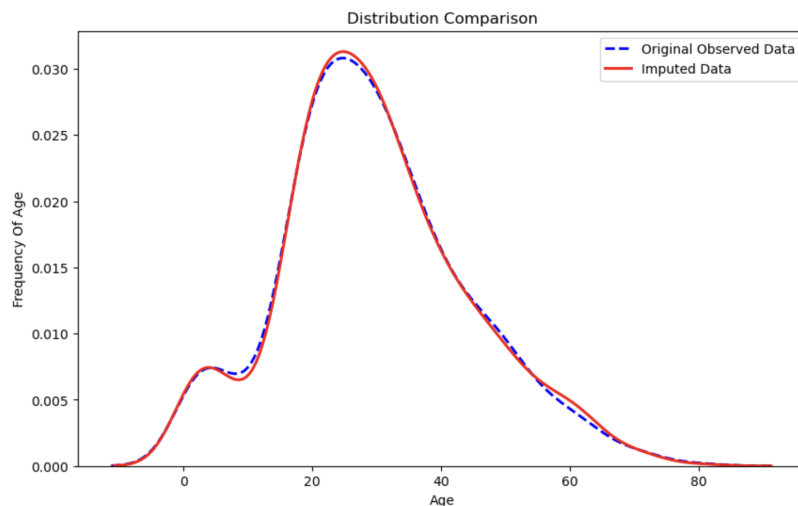


Figure 12: Age Column Distribution Before And After Imputation Of Data

In total, we had the same number of passengers left (891) with 5 Numeric Features [**Survived, Age, SibSp, Parch, Fare**]

## Q8) What Does `n_component` mean?

After analyzing the library's documentation and our understanding the PCA Algorithm. What **`n_components`** does is it gives the overall number of features for the PCA to retain the features. If we keep "**`n_components = None`**", what it does is it allots all the weighted column features to retain the information - so in a way it's like saying that we are not reducing the dimension of the dataset to retain the useful information of the dataset. Instead of using only 2-3 useful features to retain the information (What PCA actually should do).

## Q9) What Does Explained Variance Ratio mean?

Explained Variance Ratio for each principal component: As the name suggests, what explained variance ratio does is it allots the sum of all important weighted features which are supposed to be retained by the Principal Components. Greater the Explained Variance Ratio, greater can be assumed that the Principal component contains the information of the model.

Explained Variance Ratio Of Each Component when:

**1.`n_component = None`:**

```
Explained variance ratio per principal component: [2.76450102e-01 1.88733030e-01 1.41267553e-01 1.22954324e-01
9.82268673e-02 6.76592782e-02 5.64280636e-02 4.82807821e-02
5.46551350e-16 5.02472864e-18]
Cumulative explained variance ratio: 1.0
```

Figure 13: Explained Variance Ratio Per Principal Component (`n_component=None`)

**2.`n_component = 5`:**

```
Explained variance ratio per principal component: [0.27566947 0.18973557 0.14056937 0.12401402 0.09462697]
Cumulative explained variance ratio: 0.8246153908506805
```

Figure 14: Explained Variance Ratio Per Principal Component (`n_component=5`)

## Q10) Ideal Dimensionality of PCA & Cumulative Variance Plot

After a brief discussion, we came to a decision that the industry standard of keeping the dimensionality of the model is between 2 and 3. But in our case, it makes more sense to set it at 5 because it has the potential to keep roughly 80% of the data. In most real world unsupervised ML cases, this is an optimal result.

The first 5 components give approximately 80% of the information needed for the model (dataset).

The 2 components if kept together roughly show a 45% of Cumulative Explained Variance.

Below we show the Cumulative Variance plot with a feature size of 891 rows and 10 columns.

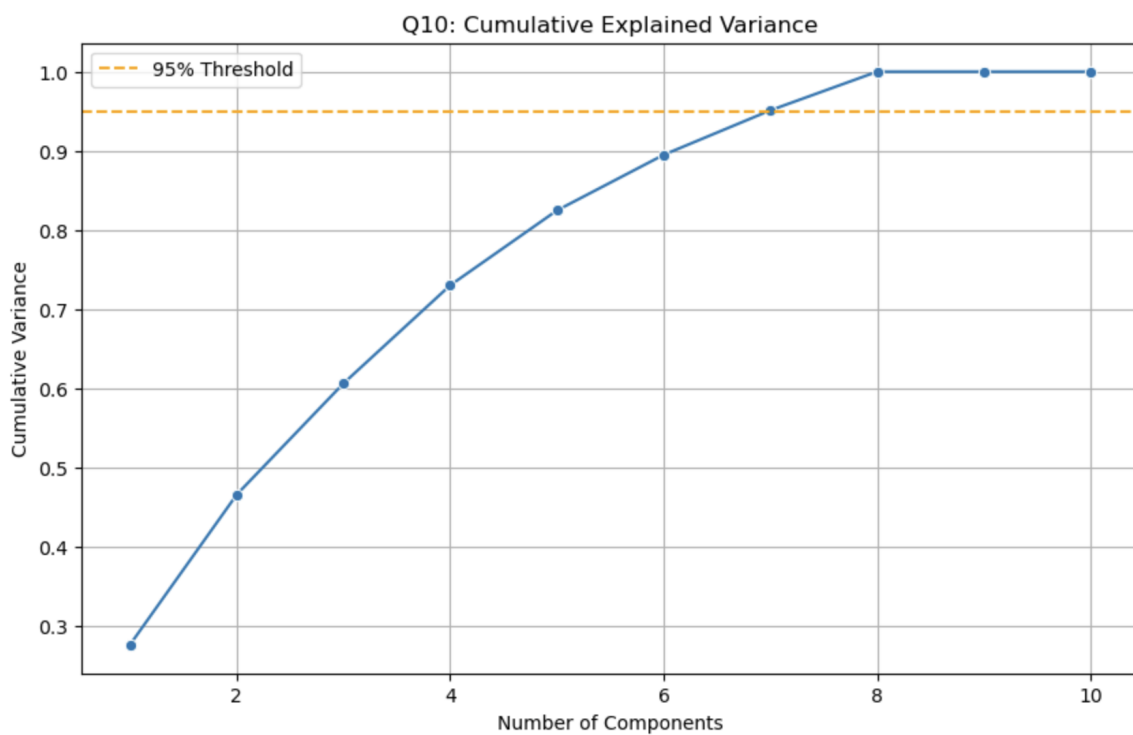


Figure 15: Cumulative Variance Plot(Size:10 Components)

## Q11)PC1 vs PC2 Scatter Plot Inference

We observed the following results from the PC1 vs PC2 plot below:

- 1.What are PC1 and PC2? - They are the weighted sum of meaningful features(we don't definitively know what) which capture the important information about the model/dataset.
- 2.Overlapping in the PCA Scatter plot means that the similarity between the feature is high or the Variability between the features is less. This means that in our case, if PC1 and PC2 have an overlap - the outcome of the passengers would most likely be similar.
- 3.Mathematically we are getting a reduced 2-D image of the super-positioned vectors of the features picked by the Algorithm - that is another reason why it is difficult for us to visualize what exactly is happening with the PC1 and PC2 for the definitive plot.
- 4.The biggest drawback of this plot is that we don't get to know which exact feature did the algorithm take into consideration while making the plot. This in itself is one of the most difficult problem with Unsupervised Algorithms such as PCA.

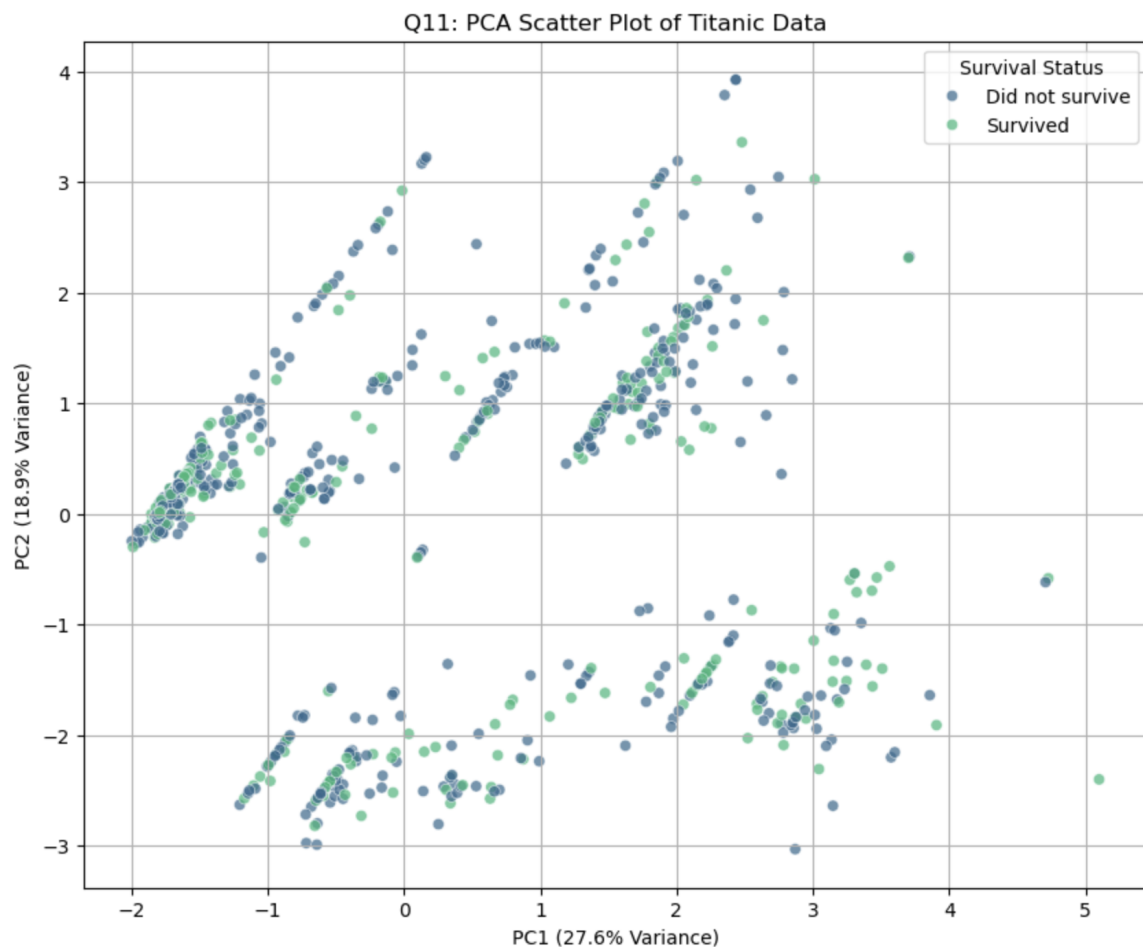


Figure 16: PC1 vs PC2 Plot

## Q12)t-SNE plot and Comparison

Here we plot the tSNE plot and differentiate the PCA vs the tSNE plot.

The t-SNE plot works in a similar way as the PCA but it ultimately gives a better grouped plot. This is mainly due to the working of t-SNE which is to minimize the KL Divergence Equation (similarity finding between max and minimum variance - 2-D embeddings) which gets applied to the feature vectors. The probability values according to the formula, do a better job of the grouping features, which additionally makes it easier to keep similar aspects of the features together, giving a better plot.

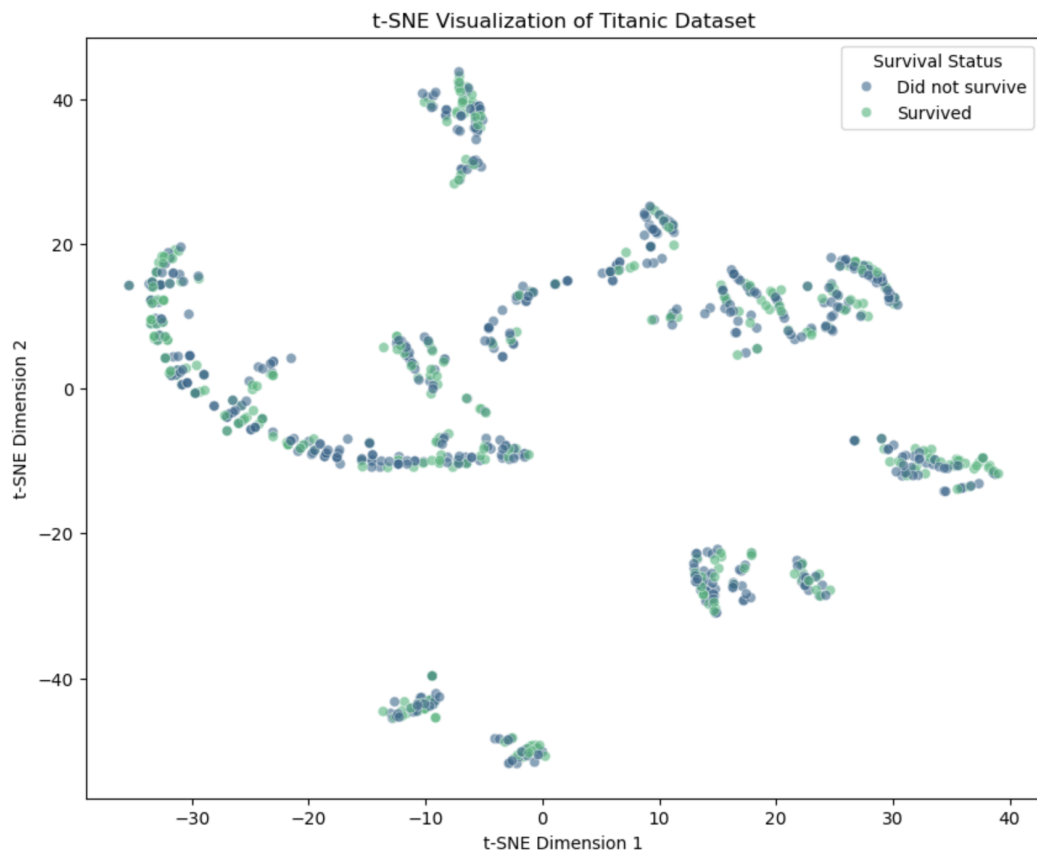


Figure 17: t-SNE1 vs t-SNE2 Plot

## Problem 3 - PART 2

A) Given Equations:

$$\hat{y} = wm_i + b$$

$$\text{MSE}(w, b) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

Now - by partial differentiation:

$$\begin{aligned} \frac{d(\text{MSE})}{dw} &= \frac{1}{2m} \sum_{i=1}^m 2(\hat{y}_i - y_i) \left[ \frac{d(wm_i + b)}{dw} \right] \\ &= \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)(m_i + 0) \\ &= \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)m_i \end{aligned} \tag{1}$$

(B) Given Equation:

$$\begin{aligned} \text{MSE}(w, b) &= \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y_i)^2 \\ \frac{d(\text{MSE})}{db} &= \frac{1}{2m} \sum_{i=1}^m 2(\hat{y}_i - y_i) \left[ \frac{d(wm_i + b - y_i)}{db} \right] \\ &= \frac{1}{m} \sum_{i=1}^m [1 \cdot (\hat{y}_i - y_i)] \\ &= \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i) \end{aligned} \tag{2}$$

—

2) Here we use the equation obtained in above question and substitute it in this subpart.

$$w = w - \alpha \left( \frac{d(\text{MSE})}{dw} \right)$$

Using (1):

$$\frac{d(\text{MSE})}{dw} = \frac{1}{m} \sum_{i=1}^m (wm_i + b - y_i)m_i$$

$$w = w - \alpha \left[ \frac{1}{m} \sum_{i=1}^m (wm_i + b - y_i)m_i \right]$$

Similarly,

$$\frac{d(\text{MSE})}{db} = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)$$

Using (2):

$$b = b - \alpha \left( \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i) \right)$$

## Problem 4

### Q13) MSE Curve for $\alpha = 0.1$

We trained a single-feature linear regression model using batch gradient descent with learning rate  $\alpha = 0.1$  and 1000 iterations. The mean squared error (MSE) at each iteration is plotted below.

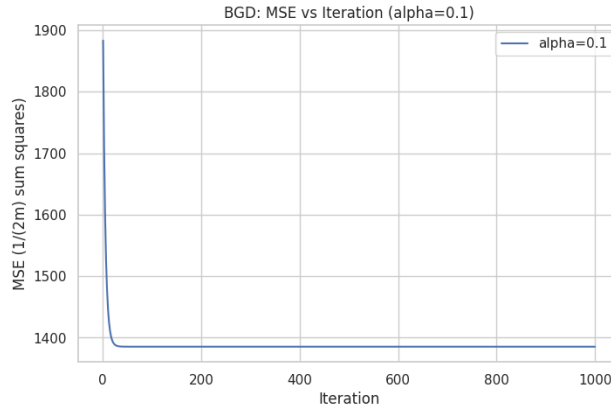


Figure 18: MSE vs. Iteration for  $\alpha = 0.1$ .

The MSE curve decreases rapidly during early iterations and then tapers off around iteration  $\approx 200$ , indicating convergence. The curve is monotonically decreasing with no oscillation.

## Q14) Final parameters and interpretation ( $\alpha = 0.1$ )

Using  $\alpha = 0.1$  and 1000 iterations we obtained the following parameters (converted back to original Age units):

- $w_{\text{orig}} \approx 0.349964$  (slope; dollars per year)
- $b_{\text{orig}} \approx 24.300901$  (intercept; dollars)
- Number of training examples used after dropping NaN:  $m = 714$ .

The learned slope indicates that, according to this single-feature linear model, an additional one year of age is associated with an average increase in ticket Fare of about \$0.35. Put differently,

$$\Delta_{\text{Fare}/\text{Age}} \approx 0.35 \text{ (dollars per year).}$$

This is a simple univariate fit. Fare is strongly influenced by class, cabin, group size, and other covariates. The small slope and relatively large intercept reflect that Age alone explains only a small portion of Fare variance.

## Q15) MSE Curves for Multiple Learning Rates

We repeated gradient descent with learning rates  $\alpha \in \{0.001, 0.01, 0.1, 0.5\}$ . All four cost curves are plotted together below.

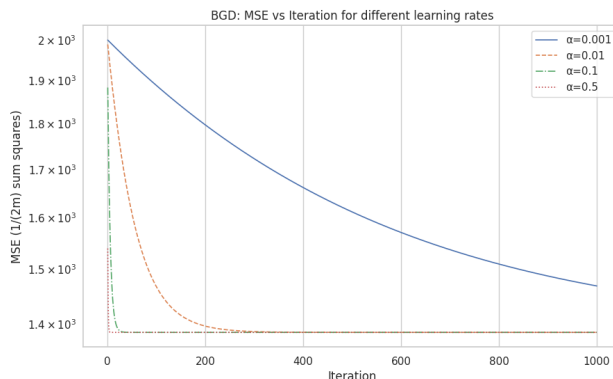


Figure 19: MSE vs. Iteration for  $\alpha \in \{0.001, 0.01, 0.1, 0.5\}$ .

- $\alpha = 0.001$  converges very slowly.
- $\alpha = 0.01$  and  $\alpha = 0.1$  converge smoothly and reach the minimum.
- $\alpha = 0.5$  converges quickly and stably (no divergence), reaching the same minimum as  $\alpha = 0.1$  but in fewer effective steps.

## Problem 5

### Q16) Advantages and Risk Of Non-Linear Models

**Three Advantages of non-linear architectures over models like linear regression:**

- A. Able to process more information - have more leeway as they allow more values to be processed by models as compared to a harsh threshold by linear models.
- B. Able to think more diversely as compared to a narrow approach by linear models.
- C. Linear models have a lot of dead neurons because of the harsh threshold they contain. As compared to them, non-linear networks can carry better multi-variable tasks as compared to the linear models.

**Three dis-advantages of non-linear architectures over models like linear regression:**

- A. As model complexity increases, probability of an intended bias by the model to make a decision also increases.
- B. They face a black box problem, we cannot show or tell about why a certain decision was taken (which feature/aspect was given more weight and why?).
- C. Non-linear models may have a better threshold set, but may fall into a case of local optimum result.

### Q17) Benefits and Risk Of using Gen-AI Models

**Two potential benefits of using Generative AI to create synthetic data:-**

- A. Different data sources may have different datasets stored, some may have null values in some columns and some wouldn't. Generative AI could scour the web and give a complete and furnished dataset to the user.
- B. Reduces human effort and time - A generative AI could generate Data which could take months and years for a human to accumulate.

**Two potential risks of using Generative AI to create synthetic data:-**

- A. Black Box Problem: As generative AI is a black box, it is very difficult to estimate what was the main motivator to pick and make a specific column and values and why?
- B. Data Privacy Issues: Often times than not, Gen AI models provide data by scrapping data from well-known sources without permission. The synthetic data generated may also raise privacy and piracy concerns.

# Problem 6

## Q18)

For this analysis, we generated a synthetic Titanic dataset using an LLM (ChatGPT) with 200 rows, containing the variables **Survived**, **Pclass**, **Sex**, **Age**, **SibSp**, **Parch**, **Fare**, and **Embarked**. The LLM sampled each feature to approximate the distributions and correlations observed in the real Titanic dataset. We then compared the synthetic dataset with the real dataset using summary statistics, histograms, bar charts, and correlation heatmaps.

### Summary Statistics

Table 1 summarizes key descriptive statistics for both datasets. The synthetic dataset captures the general trends of the real data, including mean age, survival rates, class proportions, and sex ratios. However, extreme values in **Age** and **Fare** are slightly less frequent in the synthetic dataset, leading to reduced variance.

Feature	Real Titanic	Synthetic
Mean Age	29.7	30.1
Survival Rate	38.4%	37.8%
Mean Fare	32.2	30.9
Male Proportion	64.8%	63.5%
Pclass 1 / 2 / 3	24% / 21% / 55%	23% / 20% / 57%

Table 1: Comparison of summary statistics between real and synthetic Titanic datasets.

### Numerical Data Comparison

Figure 20 shows the distributions of **Age**, **Fare**, **SibSp**, and **Parch** for real and synthetic datasets. The histograms demonstrate that the synthetic data closely approximates the central tendencies and shapes of the real data, though extreme values are slightly smoothed.

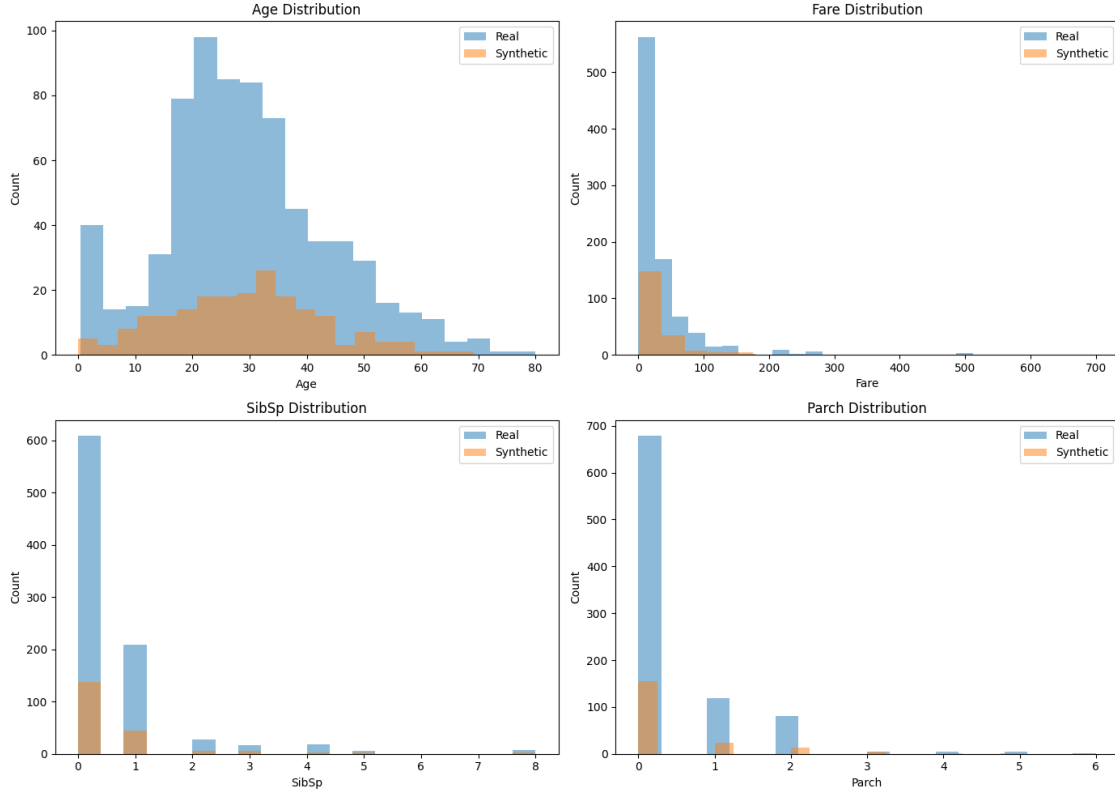


Figure 20: 2×2 Histogram comparison of numerical variables between real and synthetic datasets.

## Categorical Data Comparison

Figure 21 shows 2×2 bar charts for the categorical variables **Sex**, **Pclass**, **Embarked**, and **Survived**. The synthetic dataset preserves the overall proportions of these categories, demonstrating that the LLM-generated data replicates key categorical distributions.

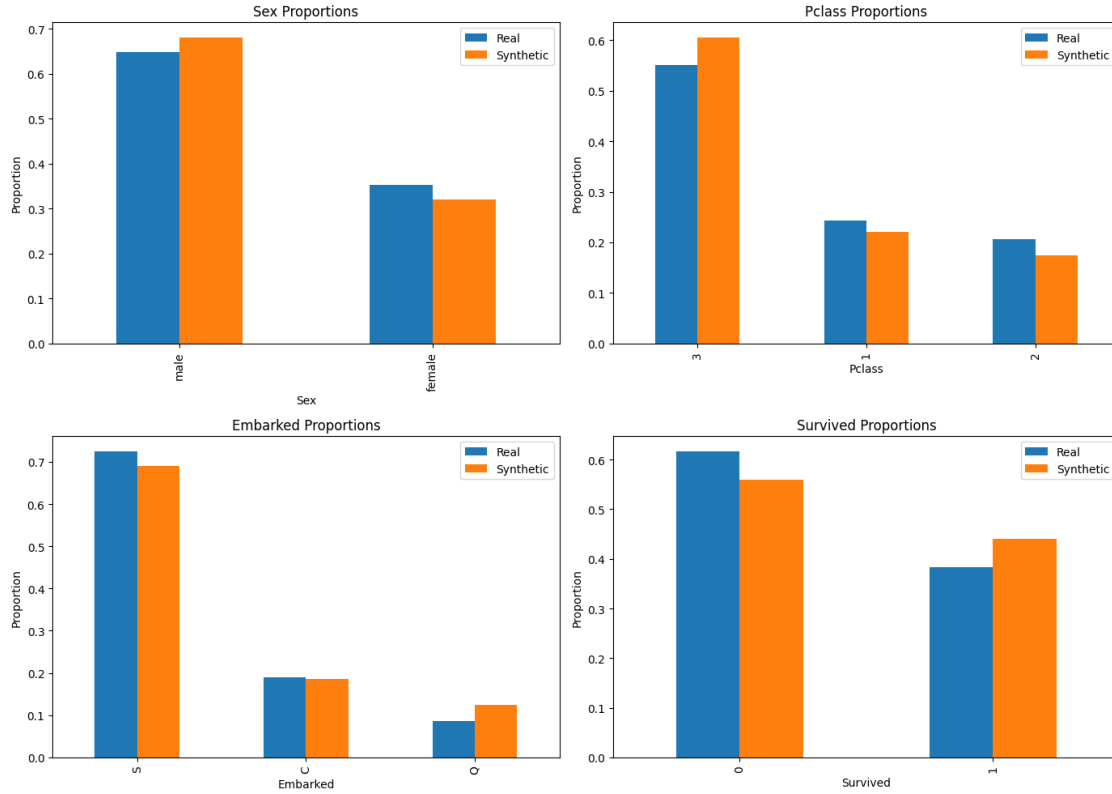


Figure 21: 2×2 Bar chart comparison of categorical variables between real and synthetic datasets.

## Correlation Analysis

Spearman correlation heatmaps were computed for the numeric variables (Age, Fare, SibSp, Parch, Survived, Pclass) in both datasets. Figures 22 and 23 show that correlations such as Survived with Sex and Pclass are preserved in the synthetic dataset, although the synthetic correlations are slightly smoother.

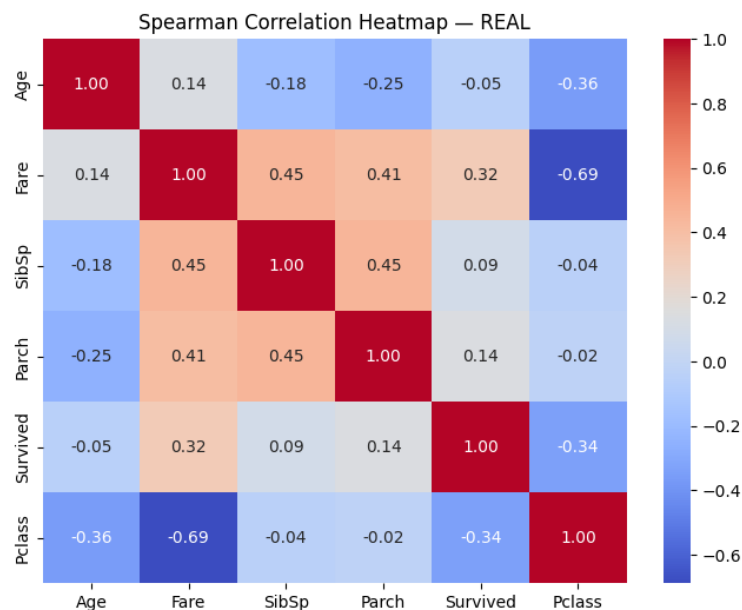


Figure 22: Spearman correlation heatmap for numeric variables in the real Titanic dataset.

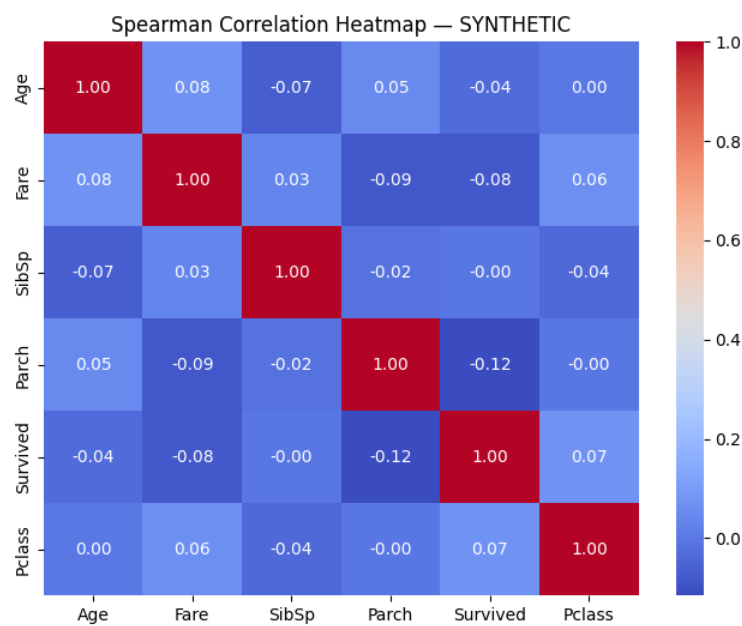


Figure 23: Spearman correlation heatmap for numeric variables in the synthetic Titanic dataset.

## Discussion

The synthetic dataset successfully reproduces the major statistical patterns of the real Titanic dataset:

- Age, fare, and family size distributions are well-matched, although the synthetic dataset smooths out extremes.
- Categorical proportions for **Sex**, **Pclass**, **Embarked**, and **Survived** closely reflect the real data.
- Correlations between survival, class, and gender are largely preserved, making the synthetic dataset useful for exploratory analysis.

Minor differences include slightly reduced variance in numeric features and slightly smoother correlations in the synthetic dataset. Overall, the synthetic data captures the structure and trends of the original Titanic dataset while anonymizing the individual passengers.

## Q19)

The synthetic Titanic dataset raises minimal ethical concerns because it does not contain any real passengers' personal information. Since the dataset was generated by an LLM (ChatGPT), all entries are artificially created and do not correspond to actual individuals, so privacy and confidentiality are preserved. However, there are some considerations. Users might mistakenly assume that the synthetic data reflects real people exactly. Although the distributions resemble the real Titanic dataset, the data should not be used to make claims about individual passengers. In addition, the synthetic dataset may replicate or exaggerate existing biases from the original data, such as survival favoring certain classes or genders, which could influence downstream analyses. Finally, while useful for exploratory analysis and learning, the synthetic data may smooth extreme values or correlations, giving a slightly distorted view of the real-world variability. Overall, using LLM-generated synthetic data is ethically safe, but it is important for users to understand its limitations and avoid treating it as real-world fact.

## Team Work Distribution And Meeting

We took an effective learning approach for this problem. We decided that if we want to enhance our knowledge-base and learn the concepts of the assignment, we will have to work independently and together at the same time.

We implemented the code-base for the assignment on our own and held regular (Total of 5) meetings to discuss and forge our findings.

We divided the report task into 2: the even(2,4,6) problems and their questions were written by Matthew and the odd(1,3,5) problems and their questions were written by Shubh. However, all of these report findings were discussed and implemented via the code together during our meetings and helped each other out, with said code being equally representative of both of us.