Problem Set 4: Exploratory Data Analysis and Batch Gradient Descent

by Victor Lockwood and Jacob Mongold

November 2025

Part I: Exploratory Data Analysis (EDA)

Problem 1: Data Loading

Q1 - Dataset Dimensions

There are 891 rows and 12 columns. The rows represent individual passengers on the Titanic and the columns are the features/their characteristics.

Q2 - Summary Statistics

The following table summarizes the key statistics for the numerical variables:

Variable	Min	Max	Mean	Median	Std. Dev.
Age	0.42	80.00	29.70	28.00	14.52
Fare	0.00	512.33	32.20	14.45	49.69
SibSp	0.00	8.00	0.52	0.00	1.10
Parch	0.00	6.00	0.38	0.00	0.81

- The max *age* of the passengers is 80, the minimum *age* is 0.42, and the average *age* is 29.70.
- The median *fare* was £14.45.
- The mean *fare* was £32.20, which is greater than the median, making the distribution right-skewed.
- The standard deviation of *fare* was 49.69.
- *Fare* has the largest spread since it has the largest standard deviation and the biggest difference between its minimum and maximum values.

Q3 - Missing Data

The table below summarizes the variables containing missing values:

Variable	Missing Values	Percent Missing (%)
Age	177	19.87
Cabin	687	77.10
Embarked	2	0.22

The *Cabin* variable contains the largest proportion of missing data (over 77%), followed by *Age* (approximately 20%), while only two records are missing *Embarked* values.

Problem 2: Exploratory Visualization and Relationship Analysis

Q4 - Univariate Analysis

Below are the plotted distributions for the classes: *Fare*, *Age*, *Pclass*, *Sex*, and *Embarked*.

Age

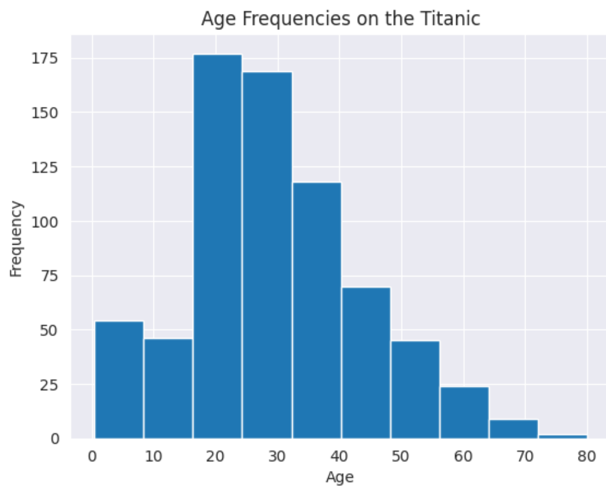


Figure 1:

The histogram above shows a single-peaked (unimodal) distribution, roughly bell-shaped but slightly right skewed. Most passengers are 20 to 40 years old, with the largest group in their 20's.

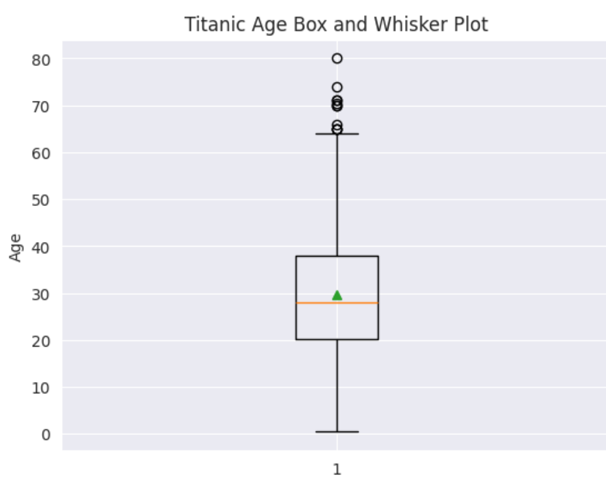


Figure 2:

The box plot reinforces this observation: the median age is ≈ 28, with the mean (green triangle) ≈ 29.7 sitting slightly higher, indicating a minor positive skew. The interquartile range captures the central 50% of the data, spanning roughly 20 to 38 years. The histogram further clarifies the presence of outliers in the 60-80 year range. There are no negative ages or impossible values, so the variable is clean, aside from missing entries handled earlier. Figure 1 and figure 2 have a sample size of 714 non-null values.

Fare

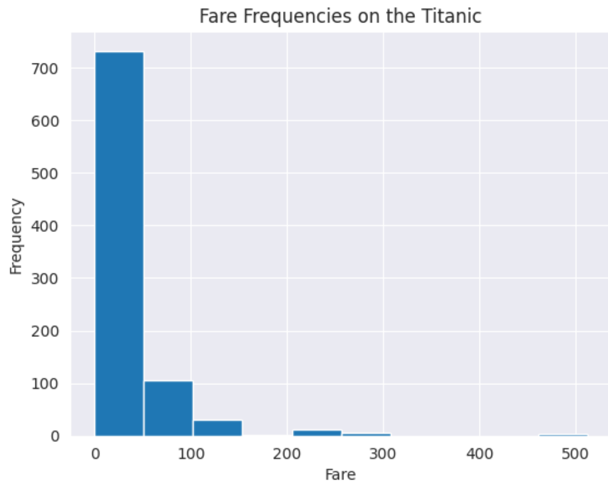


Figure 3:

The histogram for passenger fares is highly right-skewed, confirming that most fares are concentrated in the lower range (primarily under £50). However, the distribution is characterized by a significant long tail, indicating a small number of fares exceeding £500.

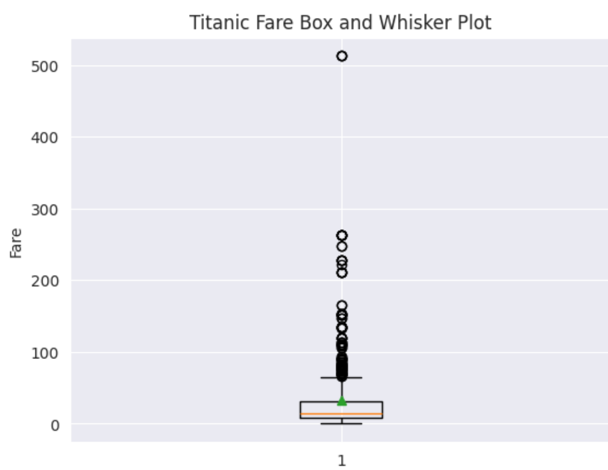


Figure 4:

The box-plot clearly shows many high-value outliers, with the mean (green triangle) far above the median. This confirms that a few luxury cabin fares strongly affect the distribution. Figure 3 and figure 4 have a sample size of 891 non-null values.

Pclass (Passenger Class)

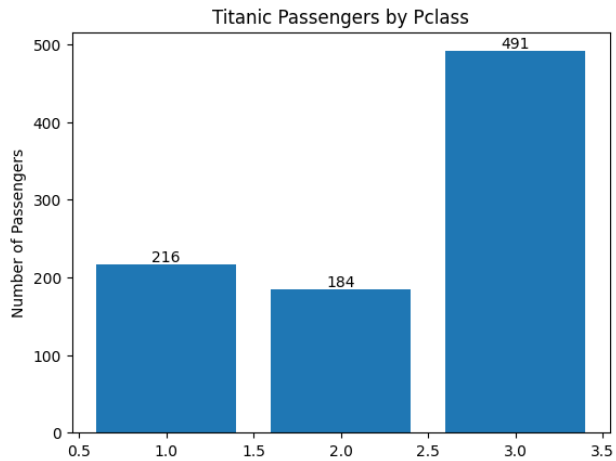


Figure 5:

The bar chart clearly shows a class imbalance in the passenger data, which is discrete and ordinal ($1^{st} < 2^{nd} < 3^{rd}$), with the majority of passengers belonging to the lower classes: 3rd class (491 passengers) dominates, followed by 1st class (216 passengers) and 2nd class (184 passengers), reflecting the higher representation of lower socioeconomic tiers among the total passengers.

Sex

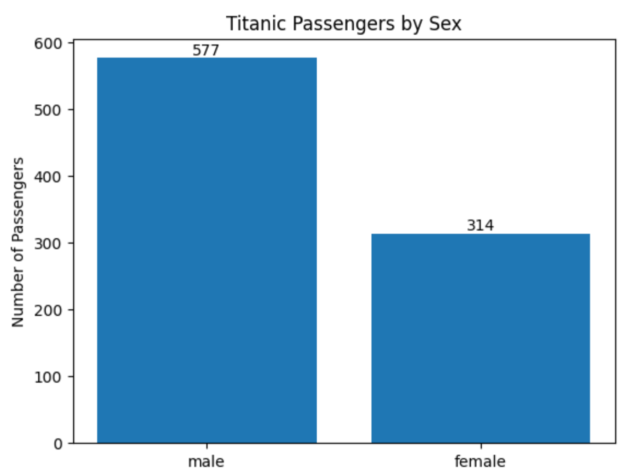


Figure 6:

The distribution by sex is binary categorical, with males dominating the sample (577 passengers, $\approx 65\%$ of the total) over females (314 passengers, $\approx 35\%$ of the total), resulting in a male-to-female ratio of nearly 2:1.

Embarked

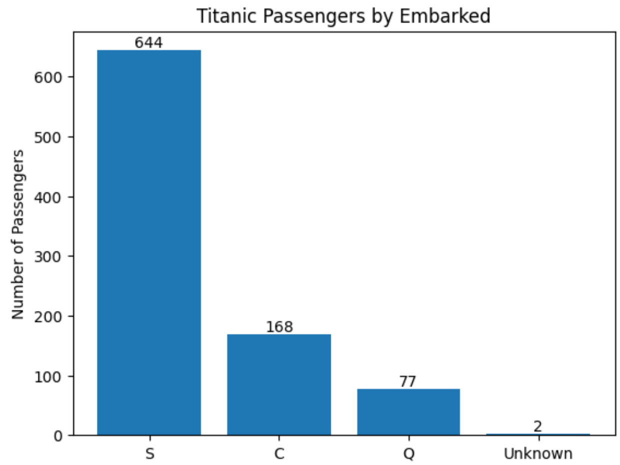


Figure 7:

The distribution of passengers by their nominal categorical embarkation port shows a clear dominance of Southampton (S) with 644 passengers ($\approx 72\%$), followed by Cherbourg (C) with 168 passengers ($\approx 19\%$) and Queenstown (Q) with 77 passengers ($\approx 9\%$), with only two passengers ($< 1\%$) having an unknown port.

Univariate Analysis Summary

The Titanic dataset shows that passenger ages were mostly between 20 and 40 years, forming an approximately normal distribution with a slight right skew and a few elderly outliers. Ticket fares were highly right-skewed, with most passengers paying less than £50 but a small number of first-class travelers paying several hundred pounds, creating many high-value outliers. The class distribution was uneven, with nearly half of passengers in third class and fewer in first and second. Males made up roughly two-thirds of the passengers, and most individuals embarked from Southampton, with smaller groups from Cherbourg and Queenstown. Overall, the data exhibits moderate skewness and strong class and gender imbalances, patterns that are typical of socioeconomic stratification in early 20th-century travel.

Q5 - Bivariate Analysis

Below are scatter plots with the following variable pairs: *Age* vs *Fare*, *SibSp* vs *Parch*, and Spearman's rank correlation (ρ) heatmap.

Age vs. Fare

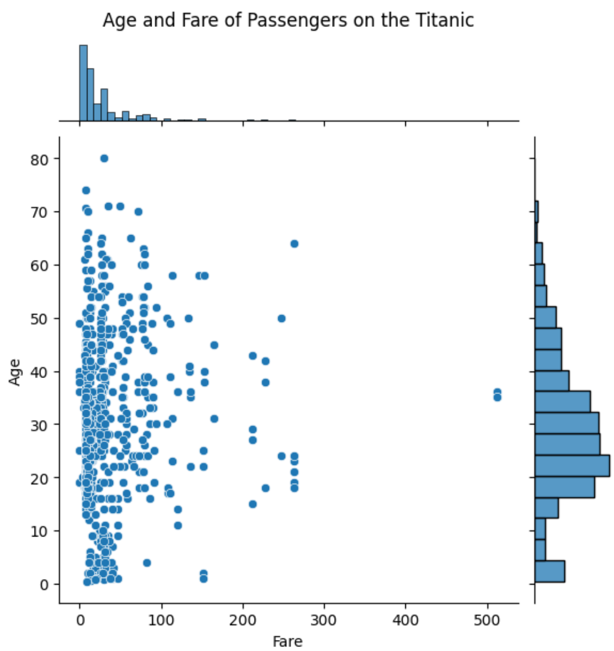


Figure 8:

The scatter plot reveals no clear linear trend between *age* and *fare*, which is consistent with the very weak Spearman correlation coefficient of $\rho \approx 0.09$ (Figure 10), indicating that ticket price was likely determined by class and cabin, not passenger age. The majority of data points form a dense cluster near the lower-left corner, as most passengers across all age groups paid low fares ($<$ £50), though a few outliers; typically older passengers paid very high fares ($>$ £200), likely representing wealthy first-class travelers.

SibSp vs. Parch

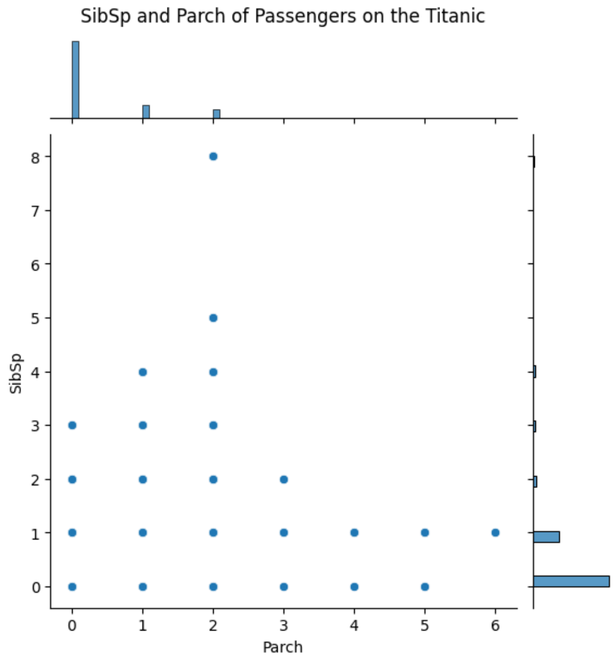


Figure 9:

The plot indicates a positive but moderate relationship between the number of siblings/spouses (*SibSp*) and parents/children (*Parch*), which is confirmed by the Spearman correlation coefficient of $\rho \approx 0.41$ (Figure 10). The distribution is highly discrete, with most passengers having zero for both variables (traveling alone). As Parch increases, SibSp tends to increase slightly, which is expected when families, including both parents and siblings, travel together; however, most points are clustered at integer combinations between 0 and 3.

Spearman Correlation Heatmap

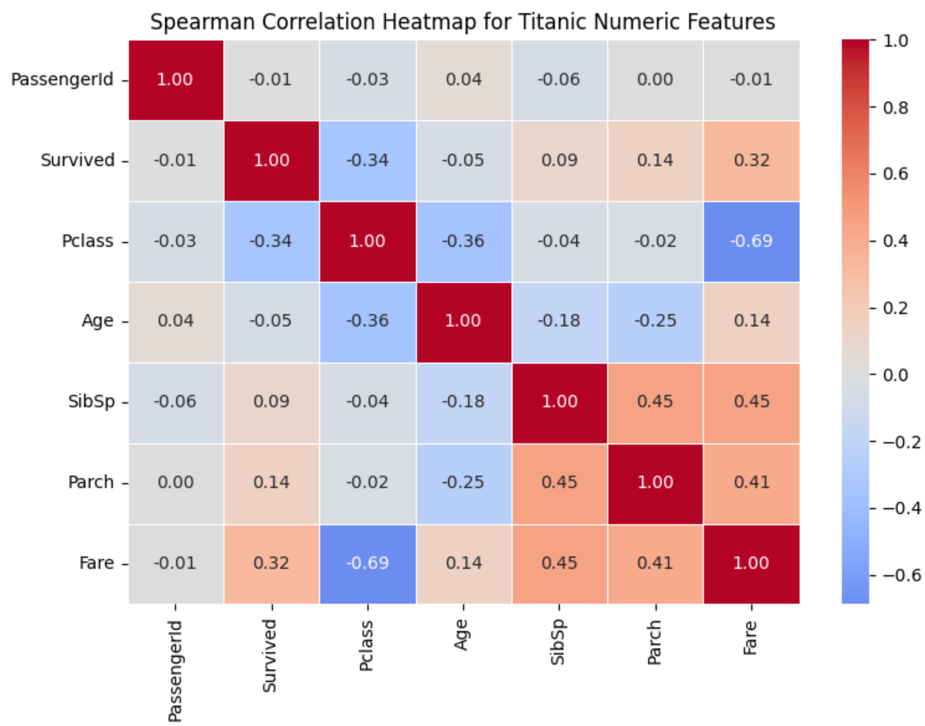


Figure 10:

The Spearman correlation matrix interpretation indicates that few relationships found among the numeric features (*Age*, *Fare*, *SibSp*, and *Parch*) are primarily linked to passenger class and family composition.

The key findings are:

- Class and Fare: A strong negative correlation exists between *Pclass* and *Fare* ($\rho < 0$), meaning that as the class number increases (e.g., from 1st to 3rd), the ticket fare tends to decrease significantly.
- Family Composition: A moderate positive correlation exists between *SibSp* (siblings/spouses) and *Parch* (parents/children) ($\rho \approx 0.41$), suggesting that people traveling with more siblings/spouses also tend to travel with more parents/children.
- Age: *Age* has only a weak correlation with all other numeric variables, indicating that age is largely independent of ticket price, family size, or passenger class.

Q6 - Multivariate/Grouped Analysis

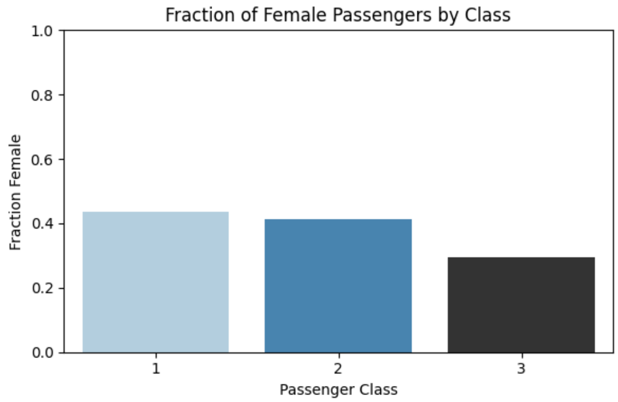


Figure 11:

The analysis shows a strong association between passenger class and sex, where the proportion of female passengers decreases significantly across lower classes: nearly half of 1^{st} ($\approx$ 46%) and 2^{nd} ($\approx$ 44%) class passengers were female, but this dropped to only one-third ($\approx$ 33%) in 3^{rd} class.

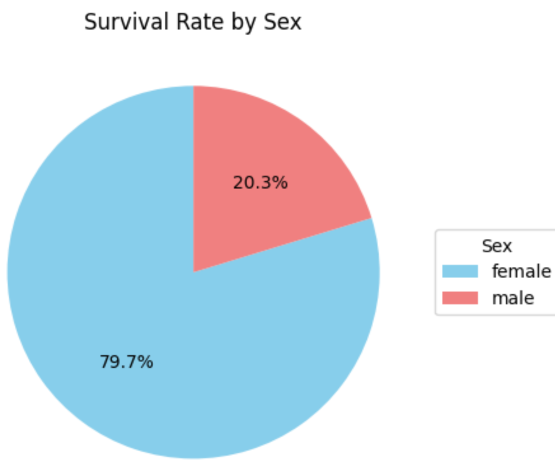


Figure 12:

This disparity in survival rates with $\approx$ 74% of females surviving versus only $\approx$ 19% of males is a striking difference that clearly demonstrates the "women and children first" evacuation policy.

The analysis reveals that both class and gender significantly influenced survival outcomes on the Titanic. The data shows a socioeconomic stratification where higher classes had a larger proportion of female passengers, particularly in 1^{st} class, reflecting wealthier families or women traveling independently. This class/gender dynamic, combined with the striking survival rate difference ($\approx$ 74% of females versus $\approx$ 19% of males), confirms a strong gender bias in survival, directly reflecting the "women and children first" evacuation priority. Passengers who were both higher-class and female were the most likely to survive.

Problem 3: Dimensionality Reduction with PCA and t-SNE

Q7 - Preprocessing Check

All rows now have complete data. Any unnecessary columns were dropped, and any missing data was replaced with the mode of that column, a method suggested int [2]. Remaining categorical data columns were converted to numerical by getting a list of unique values for that column and replacing the string value with the index of it in the list. The full matrix was then standardized with z scores to place all features on the same scale. After these steps, the dataset contained all 891 observations and had no missing entries.

Q8 - Understanding the PCA

None of the components are dropped - in other words, all components are kept and considered in the PCA, and their ratios sum to 1.

All seven principal components were extracted. The table below shows the explained variance ratios.

Component	Variance Ratio
PC1	0.27620418
PC2	0.20296099
PC3	0.15490986
PC4	0.13185055
PC5	0.09723652
PC6	0.07937840
PC7	0.05745951

PC1 captures about 27.6 percent of the variance. PC1 through PC2 capture about 48 percent combined. PC1 through PC3 capture about 63 percent. PC1 through PC4 capture about 77 percent. All components together capture the full variance.

Q9 - Explaining Variance Ratio of Each Principal Component

The cumulative variance plot shows a steady increase as components are added. The curve levels off after the fourth component. This indicates that the first three or four components retain most of the structure in the data. Beyond this point, each additional component contributes a smaller amount of variance. A reduced dimensionality of three or four components provides a good balance between compression and information retained. The explained variance ratios are [0.27620418 0.20296099 0.15490986 0.13185055 0.09723652 0.0793784 0.05745951].

Q10 . Creating a Cumulative Variance Plot

The resource [9] was utilized for assistance with reading the plot.

The first two components explain a little less than 50% of the variance. It looks like at most we would want to remove at most 2 components; nearly 90% of the variance is explained by 6 components. This is not a very dramatic curve, and it seems as though removing columns would come at more of a cost than we really need - of note, we aren't considering very many features to begin with, so dimensionality reduction doesn't seem as vital in this case. There is still overlap between the classes. PCA does not aim to separate labels and only captures the directions of greatest variance. The first two components still show visible structure related to the survival outcome.

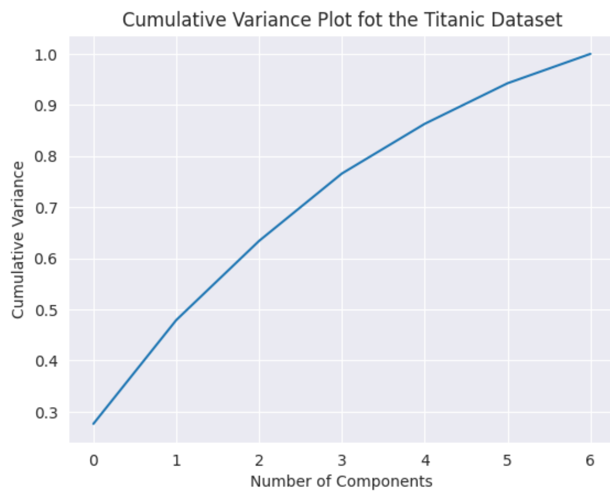


Figure 13: A cumulative variance plot from the explained variance ratios resulting from the PCA.

Q11. Making a PCA Scatterplot

Considering Q10 asked for the variance described by the first two components, I am assuming this is what is expected to be plotted on the scatter plot. Referred to here for assistance.

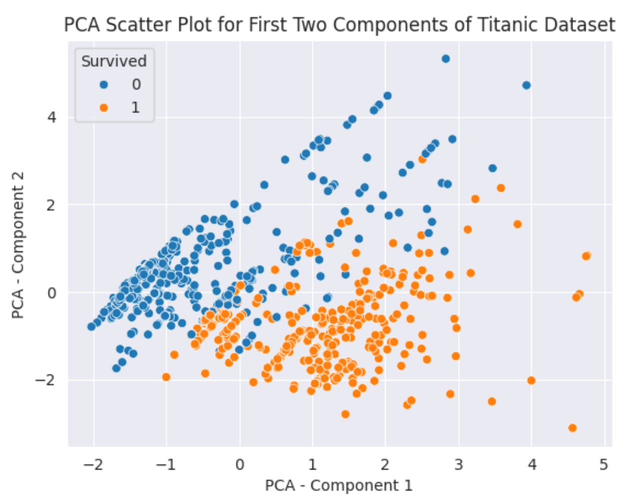


Figure 14: A scatter plot of the PCA results, with the first component as the X-axis and the second component as the Y-axis. They are colored according to whether or not a passenger survived, with orange being yes and blue being no.

There appears to be a loose linear divide between the two groups, with some overlap. This indicates that this dataset would be a good candidate for linear regression or gradient descent, and we can also observe that it confirms the PCA analysis that indicated 65% of variance was explained by the first two components.

Q12. t-SNE Analysis

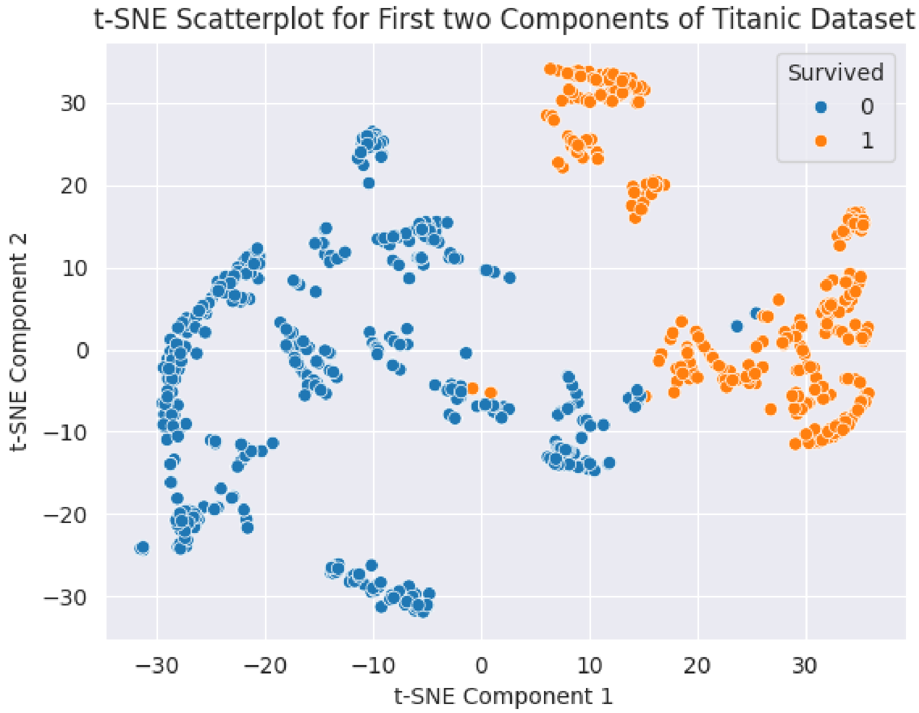


Figure 15: Scatter plot results from the t-SNE analysis, where the first component is the X-axis and the second component is the Y-axis. Once again, survivors are orange and those who did not survive are blue.

Compared to PCA, t-SNE places more emphasis on local similarity. t-SNE often reveals grouping or clustering that PCA cannot show. PCA preserves global structure and variance directions, but it is linear. t-SNE can capture curved or layered structures. In the context of the Titanic dataset, PCA provides a broad view of the data structure. t-SNE gives a more detailed view of how passengers with similar attributes group together. PCA is useful for checking variance structure and the effects of preprocessing. t-SNE is better for visual exploration of potential clusters.[6] [7] [3] [4][5][8]

Part II: Batch Gradient Descent

Bonus: Derivation of Gradient Descent

We consider a simple linear regression model with one input feature x and target y, where the prediction for example i is

$$\hat{y}^{(i)} = wx^{(i)} + b$$

The Mean Squared Error (MSE) cost function for m training examples is

$$J(w,b) = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2.$$

Define the prediction error for each training example as

$$e^{(i)} = \hat{y}^{(i)} - y^{(i)} = wx^{(i)} + b - y^{(i)}.$$

Bonus 1: Partial derivatives

The partial derivative of the cost with respect to w is

$$\frac{\partial J}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} e^{(i)}x^{(i)}$$

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} e^{(i)}$$

Bonus 2: Gradient descent update rules

Gradient descent updates parameters by moving them in the negative direction of the gradient:

$$w := w - \alpha\frac{\partial J}{\partial w}, \qquad b := b - \alpha\frac{\partial J}{\partial b}$$

where α is the learning rate.

Substituting the gradients, the explicit update rules become:

$$w := w - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\left(wx^{(i)} + b - y^{(i)}\right)x^{(i)}\right),$$

$$b := b - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\left(wx^{(i)} + b - y^{(i)}\right)\right).$$

Bonus 3: Why the sum over all training examples?

These expressions sum over all m examples because this is batch gradient descent. The algorithm computes the gradient using the entire dataset at each iteration, producing stable and smooth updates compared with stochastic or mini-batch methods.

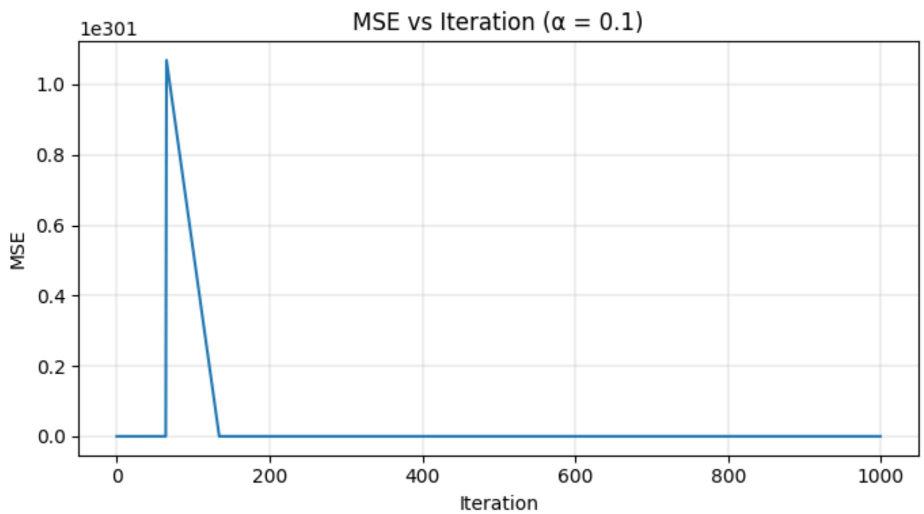
Bonus 4: Convexity and convergence

For linear regression, the cost function $J(w,b)$ is a convex quadratic function. A convex function has a single global minimum, so gradient descent is guaranteed to converge to this minimum for appropriately chosen learning rates.

Problem 4: Batch Gradient Descent on the Titanic Dataset

NOTE: Missing data was once again replaced by using the mode of the appropriate column.

Q13. Analysis of MSE vs Iteration Plot ($\alpha = 0.1$)



Based on the MSE vs. Iteration graph:

Where does the curve taper off?

The MSE curve tapers off (reaches convergence) very quickly around iteration 150-200 in the graph. But when looking at the data the weights and bias explode to NaN and $-\infty$ with the MSE showing as 0 due to the overflow. The graph is misleading because it plots these overflow values as 0, creating the false appearance of convergence.

Is the curve monotonically decreasing, or not?

No, the curve is not monotonically decreasing. The data shows severe divergence:

- **Initial MSE**: 1,751.88 (iteration 0)
- **Iteration 1-4: MSE explodes exponentially**:
  - Iteration 1: MSE = 1.93×10^7
  - Iteration 2: MSE = 7.49×10^{11}
  - Iteration 3: MSE = 2.91×10^{16}
  - Iteration 4: MSE = 1.13×10^{21}
- **By iteration 995+**: Numerical overflow causes $w = \text{NaN}$, $b = -\infty$, and MSE displays as 0.

What went wrong?

The data and graph indicate that there is divergence due to gradient descent instability and not successful convergence. The likely causes are:

1. $\alpha = 0.1$ is large for unscaled features, causing the algorithm to overshoot the minimum repeatedly with increasing magnitude.

2. Without standardization, *Age* (range 0-80) and *Fare* (large variance) create a poor optimization landscape where gradients have very different magnitudes.

Overshooting seems to be the most likely cause (more on overshooting in [1]).

Q14. Final Parameters and Slope Interpretation

After running batch gradient descent for 1000 iterations with a learning rate of $\alpha = 0.1$, the final parameters are:

$$w = \text{NaN}$$
$$b = -\infty$$

Interpretation of the Slope

In a properly converged linear regression model, $\text{Fare} = w \cdot \text{Age} + b$, the weight w would represent Δ, the change in *Fare* for each one-year increase in *Age*. Using a learning rate of $\alpha = .001$, we get a slope of roughly 0.8234, indicating a positive linear correlation. The MSE at that point though is around 1237.56, so a linear correlation likely does not explain the data well, especially since the other weights failed.

Due to the divergence of the gradient descent algorithm for the other rates, we cannot provide a meaningful interpretation of the slope. The parameters exploded beyond representable numerical values, resulting in:

- **Weight (w)**: NaN - indicates numerical overflow
- **Bias (b)**: $-\infty$ - indicates that the bias grew unboundedly negative

Since our implementation failed to converge, we cannot determine this relationship from the results, further indicating that a linear model does not explain the data well.

Q15. Effect of Learning Rate on Convergence

We ran batch gradient descent for $y = wx + b$ with $w_0 = 0$, $b_0 = 0$ for 1000 iterations using four different learning rates $\alpha \in \{0.001, 0.01, 0.1, 0.5\}$ and recorded the MSE at each step.

For the smallest learning rate, $\alpha = 0.001$, the cost curve decreases smoothly and monotonically. The MSE drops steadily but quite slowly and still changes slightly near the end of the 1000 iterations, indicating convergence in the right direction but with very small parameter updates.

For $\alpha = 0.01$, the MSE decreases much faster at the beginning, reaching a region close to the minimum in far fewer iterations. However, the curve shows a sharp spike to extremely large values (on the order of 10^{303}) before collapsing, which is consistent with numerical overflow and the parameter updates stepping too far in parameter space.

For larger learning rates, $\alpha = 0.1$ and especially $\alpha = 0.5$, the behaviour is unstable. The cost initially decreases but then rapidly explodes, and the implementation reports numerical warnings (overflow in the sums and invalid operations), after which the final parameters become `nan` or ∞ and the recorded MSE degenerates to zero due to these non-finite values. This corresponds to gradient descent diverging instead of converging.

The comparison shows that small learning rates (e.g. $\alpha = 0.001$) yield stable but slow convergence, while moderate to large learning rates (especially $\alpha = 0.1$ and $\alpha = 0.5$ in our implementation) cause the updates to overshoot the minimum, leading to exploding costs and numerical instability. The learning rate therefore directly controls the trade-off between speed of convergence and the risk of divergence.

Problem 5: Critical Discussion of LR vs. DNNs

Q16. DNNs vs. Linear Regression

Neural networks can pick up on non-linear patterns in data that gradient descent can't. They can additionally handle data noise better with their internal activation functions, and can be tweaked with many hyperparameters to get better results - with gradient descent, you are limited to a linear function with very little room for change.

The added complexity of DNNs means that they require a LOT more compute resources than a linear model. You also potentially over-engineer the problem where a linear model does effectively explain the data and miss out on a simpler and more elegant solution. Lastly, setting up a good DNN model is a time-consuming art; the hyperparameters serve as both an advantage and disadvantage in this regard, as a lot of tweaking comes into play when obtaining a good model.

Q17. Advantages and Disadvantages of Synthetic Data

Synthetic data has the advantage of being quick to generate if the data has not already been collected, and it is much more likely to be complete. We have seen a large portion of the Titanic dataset was incomplete, and there is very little chance of recovering that data (especially as time goes on).

Data generated by an LLM may have inaccuracies since the LLM was trained on more than just the original dataset's data - biases from other sources may come into play and make the dataset less accurate than the real one. In the case of data missing from the real dataset, while we would get a complete dataset from an LLM, we do not know how it will fill in the blanks. We opted to use a metric based on the real dataset to fill holes, but an LLM may generate something based on entirely different metrics that have nothing to do with the real data.

Problem 6: Creating Synthetic Data

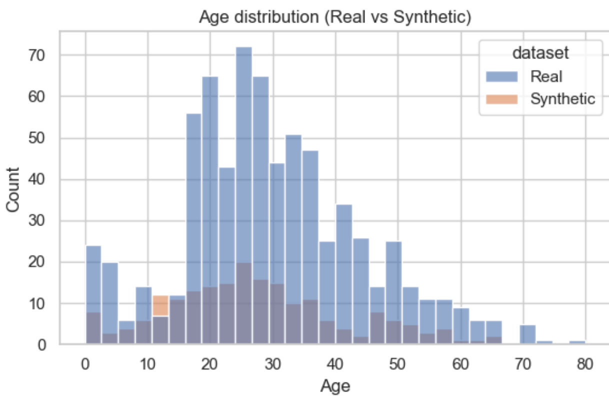
Q18. Comparison and Visualization of Real and Synthetic Titanic Data

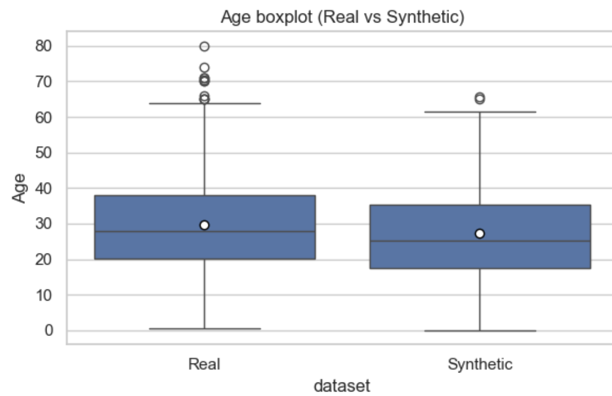
We generated the synthetic dataset by using ChatGPT to produce code that ran correctly in the notebook. This produced a synthetic dataset with 200 passengers.

The table below compares the summary statistics of the real and synthetic data:

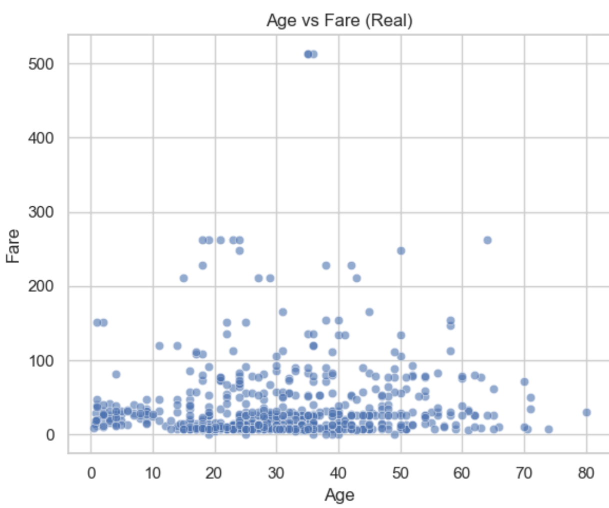
	Age	**Fare**	**SibSp**	**Parch**
Real Data (n=891)				
Mean	29.70	32.20	0.523	0.382
Std. Dev.	14.53	49.69	1.103	0.806
Min	0.42	0.00	0	0
25th pct.	20.13	7.91	0	0
Median	28.00	14.45	0	0
75th pct.	38.00	31.00	1	0
Max	80.00	512.33	8	6
Synthetic Data (n=200)				
Mean	27.24	34.38	0.595	0.360
Std. Dev.	14.47	45.96	1.268	0.757
Min	0.00	0.00	0	0
25th pct.	17.54	7.98	0	0
Median	25.31	17.46	0	0
75th pct.	35.25	37.92	1	0
Max	65.65	267.37	8	6

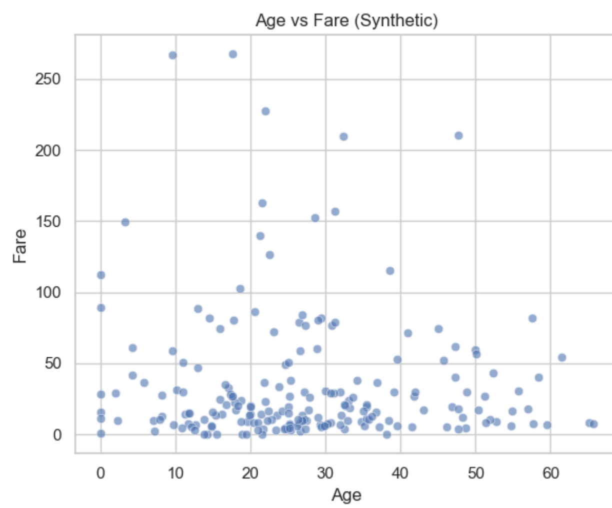
The statistics show that the synthetic data resembles the real dataset in central tendency and spread. The most noticeable differences occur in the upper extremes: the synthetic dataset does not reproduce the highest fares or the oldest passengers. This reflects the simpler sampling strategy, which does not amplify rare outliers. To assess similarity more fully, we repeated the exploratory visualizations from Problem 2.



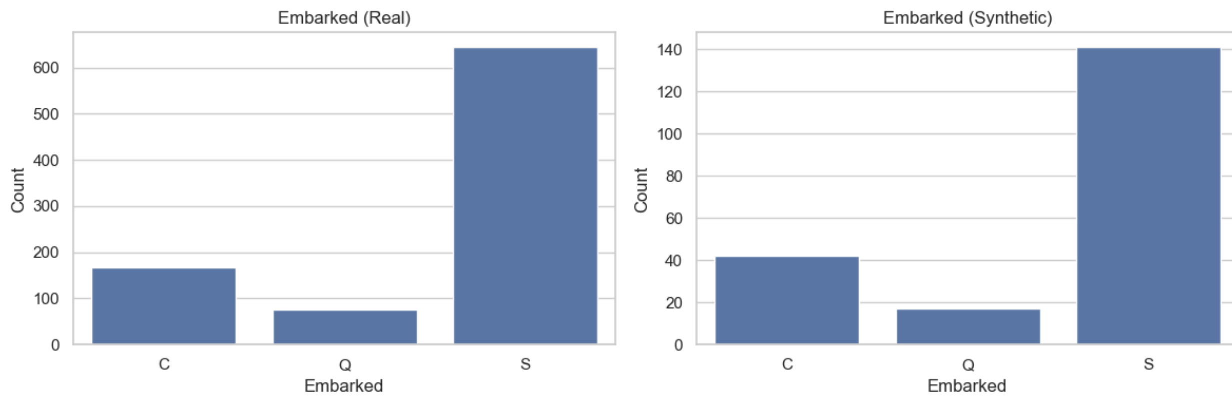


Age (Histogram and Box-plot): The synthetic age distribution follows the same general pattern as the real data, with a peak in the mid-20s and a long right tail. The synthetic dataset is slightly younger on average and lacks the older passengers seen in the real data. The box-plots show similar medians and quartiles, with fewer extreme upper outliers in the synthetic version.

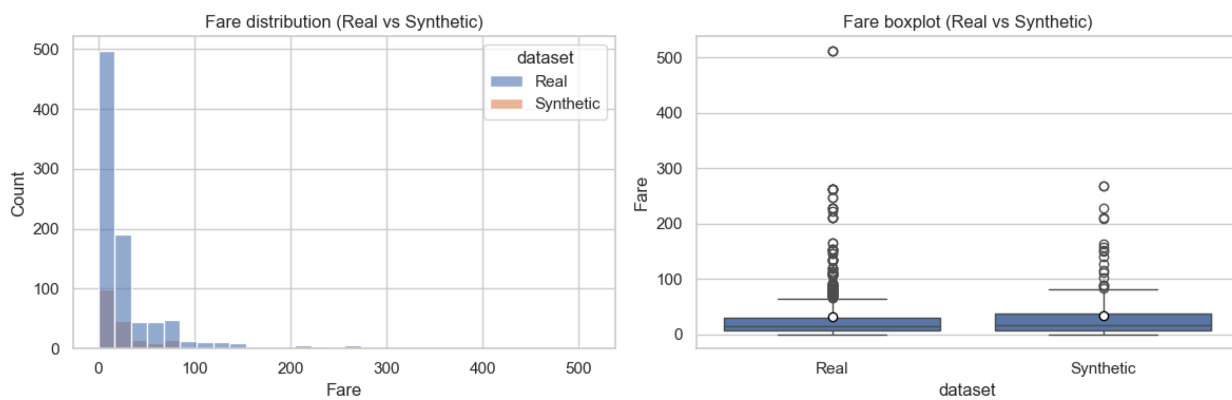




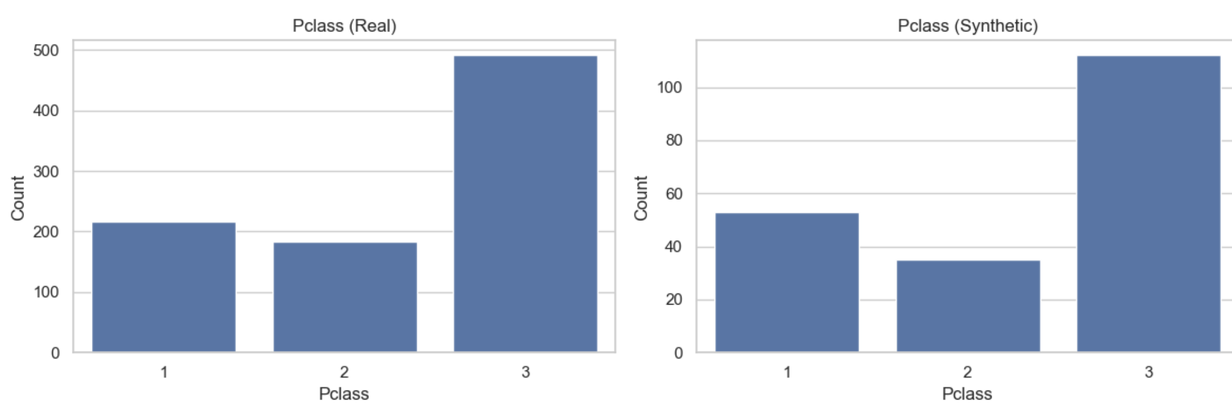
Age vs. Fare: Both datasets show a weak relationship between Age and Fare, with fares concentrated near zero. The real dataset contains several very high fares, creating tall vertical outliers in the scatterplot. These do not appear in the synthetic version, which smooths the tail and produces a more compact cluster.



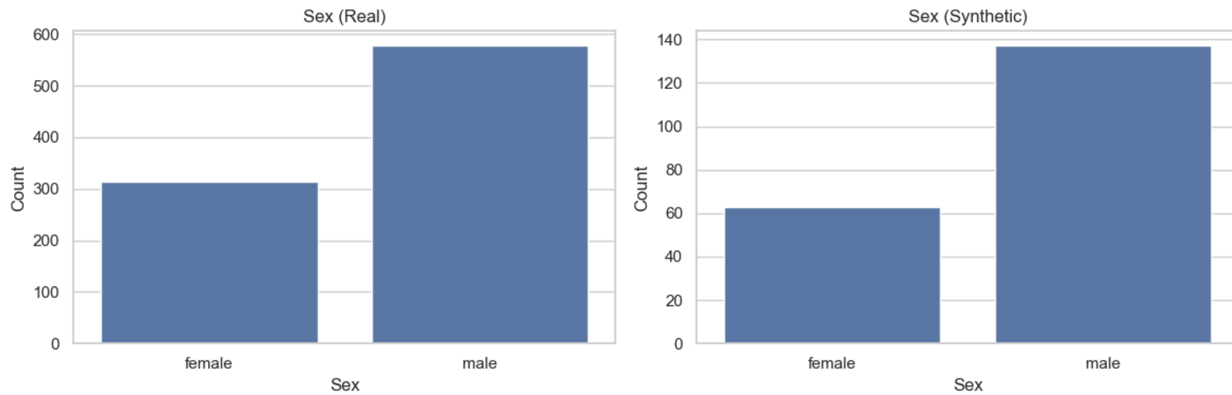
Embarked: The Embarked distribution is very similar between the two datasets. Most passengers depart from Southampton (S), followed by Cherbourg (C) and Queenstown (Q). The synthetic dataset captures these proportions closely, with only minor shifts in the size of the Q group.



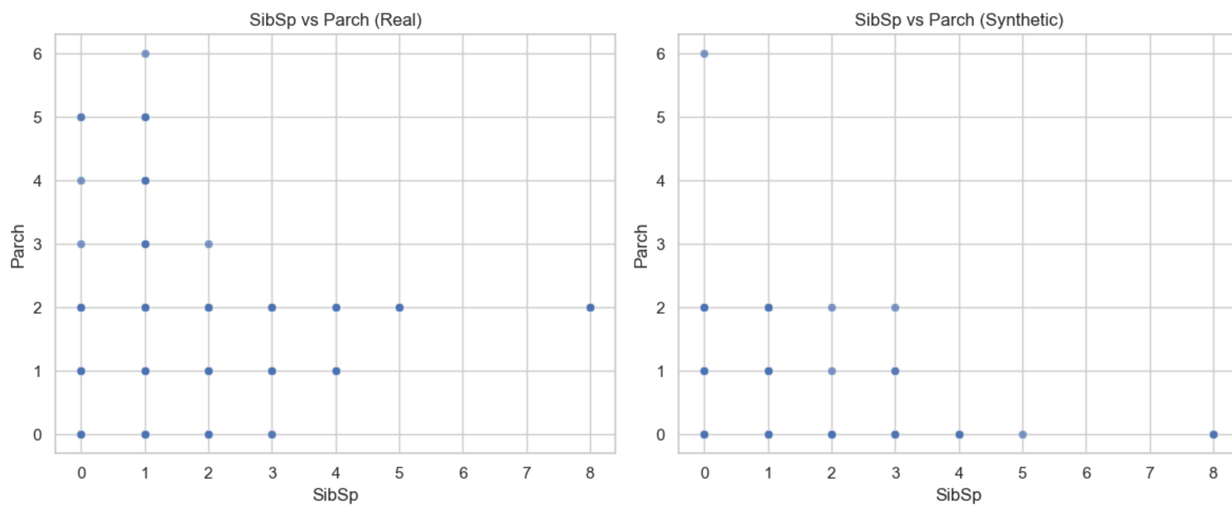
Fare (Histogram and Boxplot): The synthetic fare distribution matches the real dataset's heavy right skew but does not reproduce the extreme high-fare outliers (300–500). As a result, the synthetic boxplot shows a slightly smaller range. Medians and quartiles remain well aligned.



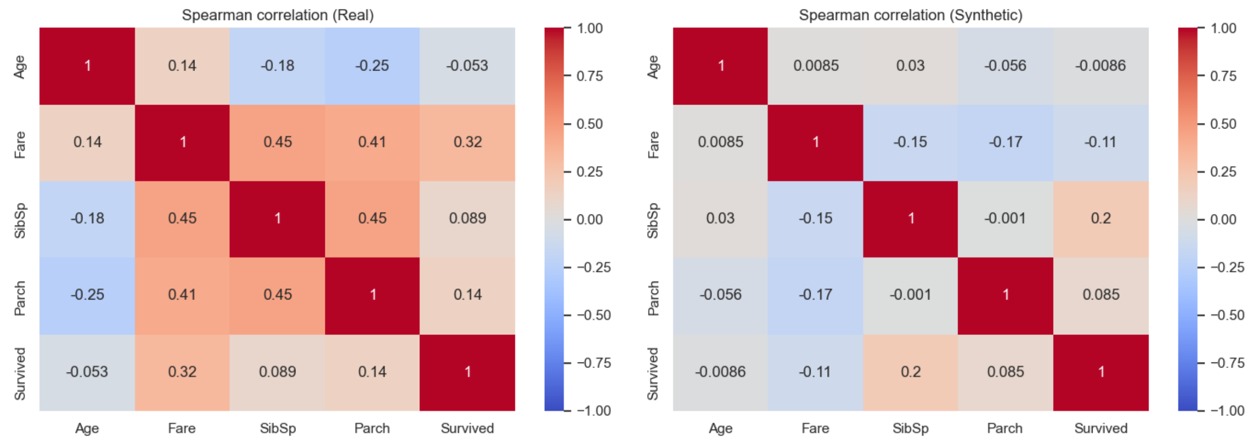
Pclass: The synthetic class distribution maintains the same rank order as the real dataset: most passengers are in third class, then first, then second. The proportions shift slightly but preserve the overall structure of the original data.



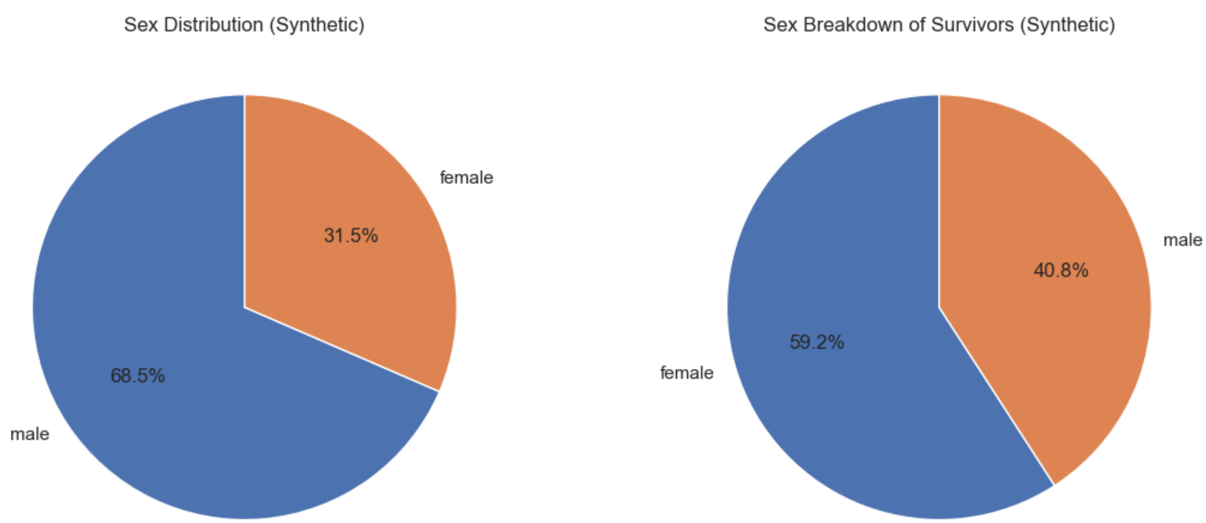
Sex: The synthetic male–female ratio is consistent with the real dataset. There are slightly fewer females in the synthetic version, but the imbalance remains in the same direction, with more males overall.



SibSp vs. Parch: The structure of family groups is captured well in the synthetic dataset. Both datasets show that most passengers travel alone or with one family member. The synthetic version has fewer large family groups, but the grid-like pattern of possible (SibSp, Parch) combinations is preserved.



Spearman Correlation: The real dataset shows moderate positive correlations among Fare, SibSp, and Parch, with weak negative correlations between Age and the family variables. The synthetic correlations follow the same direction but with lower magnitudes. This is expected given the simpler, independent sampling strategy used for data generation.



Sex Breakdown of Survivors: The synthetic survivor distribution shows the expected pattern: females represent a larger share of survivors than males. The exact percentages differ slightly from the real dataset, but the general trend remains intact.

Overall, the synthetic dataset captures the main patterns, distributions, and relationships present in the real Titanic data. The main deviations occur in the tails of the distributions and in the strength of correlations, both of which reflect the limited complexity of the generative method used.

Q19. Ethical Considerations of the Synthetic Data

The synthetic dataset does not contain identifiable passenger information, but it still raises a few ethical concerns.

Potential Risks:

- Even though the generator samples distributions rather than names or identifiers, the structure of the Titanic dataset includes sensitive attributes such as gender, age, class, and survival. Poorly designed generators may create synthetic records that resemble actual passengers or reproduce rare combinations that existed in the real dataset. This risks misrepresentation of real individuals.
- **False confidence in data realism:** Synthetic datasets can appear statistically similar to the real data while missing important edge cases or structural features. If someone later uses the synthetic data for modeling or decision-making, the results may be biased or unreliable.

Potential Benefits:

- Because the synthetic data is generated from distributions rather than real passenger records, it removes direct links to specific individuals and lowers re-identification risk compared to releasing the raw dataset. This results in reduced exposure of sensitive attributes.
- Synthetic data allows testing code, pipelines, and visualizations without manipulating or distributing the original dataset, which may be subject to usage constraints and allows us to experiment more safely.

The synthetic data used here reduces direct privacy concerns but still carries the risk of misinterpretation and overgeneralization. It should be treated as an approximation of the real dataset, not as a substitute for the original data when accurate modeling or historical analysis is required.

Workload

- Problem 1 was primarily handled by Jacob, with Victor looking it over and giving the okay.
- Problem 2 was split relatively evenly, with Victor handling Univariate Analysis and half of Bivariate Analysis, and Jacob handling the other half of Bivariate Analysis and then Multivariate Analysis.
- Problem 3 was primarily handled by Victor, with Jacob okaying the code and handling much of the conceptual analysis.
- The bonus problem was handled primarily by Jacob, with Victor independently deriving the gradients as well and we compared answers. Jacob's was the most correct, and later code was tweaked accordingly.
- Problem 4 was primarily handled by Victor for the algorithm code, but both of us worked through what was happening in the code when results were unexpected and was adjusted accordingly. Jacob handled visualizations and much of the interpretation.
- Problem 5 was primarily handled by Victor.
- Problem 6 was entirely handled by Jacob.
- The report is as in-depth and pretty as it is because of Jacob's hard work and dedication to make sure all conceptual and math questions were accurately responded to. We both worked on the report itself overall.

References

[1] Gradient Descent Optimization: A Simple Guide for Beginners — Udacity — udacity.com. https://www.udacity.com/blog/2025/02/gradient-descent-optimization-a-simple-guide-for-beginners.html. [Accessed 17-11-2025].

[2] How to Deal with Missing Data — mastersindatascience.org. https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/. [Accessed 17-11-2025].

[3] Visually Explained. Principal component analysis (pca) — easy and practical explanation. YouTube, 2021. Accessed: 2025-11-16.

[4] Visually Explained. Latent space visualization: Pca, t-sne, umap — deep learning animated. YouTube, 2024. Accessed: 2025-11-16.

[5] Visually Explained. Visualizing complex data: Pca vs t-sne techniques. YouTube, 2024. Accessed: 2025-11-16.

[6] Stuart Russell and Peter Norvig. *Artificial Intelligence*. Pearson, 4 edition, 2021.

[7] Jonathon Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014.

[8] Zach Star. Breaking the curse of dimensionality. YouTube, 2024. Accessed: 2025-11-16.

[9] Jake VanderPlas. In Depth: Principal Component Analysis — Python Data Science Handbook — jakevdp.github.io. https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html. [Accessed 17-11-2025].