

# 02450 Report 1

Gisle Joe Garen (s242712)      Ignacio Ripoll González (s242875)  
Sebastian Svelmøe Timm (s243935)

October 3, 2024

## Table of authors and their contributions

Name Student ID	Gisle Joe Garen s242712	Ignacio Ripoll González s242875	Sebastian Svelmøe Timm s243935
Introduction	15%	70%	15%
Task 1	15%	70%	15%
Task 2	40%	45%	15%
Task 3	70%	15%	15%
Task 4	15%	15%	70%
Problems	30%	30%	40%

## Introduction

This project focuses on the analysis of a dataset using concepts from the initial weeks of lectures on data feature extraction and visualization. In our case, the Titanic dataset is being analyzed to provide a detailed understanding and prepare it for future machine learning tasks, such as classification and regression. The report includes a description of the dataset, a summary of prior analyses, and a discussion of how various attributes might influence survival. It also addresses potential data quality issues, such as missing values and outliers, and presents visualizations, including principal component analysis (PCA), to explore patterns and relationships. This analysis serves as a foundation for evaluating the feasibility of applying machine learning techniques.

All of our calculations are done using Python. We rely on the following packages from the Scientific Python ecosystem.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
from sklearn.preprocessing import StandardScaler, LabelEncoder
from scipy.linalg import svd
import matplotlib.patches as mpatches
```

We will intersperse some of our code in this document to clarify how we have obtained the stated results. To see all our code, look to the file `analysis.py`. That file was originally a Jupyter notebook that we converted to a Python source file using `nbconvert`.

## Task 1: Description of the dataset

### Explain what your data is about. I.e. what is the overall problem of interest?

The Titanic dataset not only provides a valuable resource for statistical analysis but also offers a glimpse into one of the most infamous maritime disasters in history. The RMS Titanic set sail from Southampton, England, on April 10, 1912, bound for New York City. On board were over 2,200 passengers and crew. Tragically, on the night of April 14, 1912, the Titanic struck an iceberg and sank within a few hours. Of the more than 2,200 people on board, over 1,500 lost their lives.

The ship sank due to a combination of design flaws, including insufficient watertight compartments, and the collision with the iceberg, which breached the hull. A critical factor in the high death toll was the lack of lifeboats. The Titanic carried lifeboats for only about half of those aboard, meaning that when the ship began to sink, not everyone had a chance to escape.

The overall problem of interest in studying the Titanic dataset is to identify which demographic, socio-economic, and other relevant characteristics influenced survival rates during the disaster. The dataset includes information on passenger class, gender, age, fare, family size, and more, allowing us to analyze patterns and key factors affecting survival probabilities. It is particularly compelling because the lifeboat shortage necessitated prioritizing passengers, with those in higher social classes, such as first-class passengers, being given precedence during evacuation. By examining these factors, we can gain insights into how class, age, gender, and other variables impacted survival, revealing broader socio-economic dynamics and human behavior in life-and-death situations.

### Provide a reference to where you obtained the data.

The Titanic dataset was obtained from Kaggle, a popular platform for data science competitions and datasets. You can access the data via the following link: [Kaggle Titanic Dataset](#). The data is provided in two separate datasets: one for training and one for testing. The training dataset includes the information used to build predictive models, while the testing dataset is used to evaluate the performance of these models.

### Summarize previous analysis of the data.

In previous analyses of the Titanic dataset, several key data preparation and modelling techniques have been employed.

**Data Cleaning:** This step typically involves handling missing values, which is a common issue in the Titanic dataset, particularly for variables like "Age," "Cabin," and "Embarked." Many studies either impute missing ages using the median or mean, or apply more advanced techniques like regression-based imputation. The "Cabin" attribute is often omitted due to its high percentage of missing data. See [Droste \(2020\)](#) and [Realm \(2023\)](#) for examples of this.

**Classification Techniques:** Many analyses have used classification models to predict survival based on attributes like class, gender, and age. Logistic regression is a commonly used technique for this, along with decision trees, random forests, and support vector machines (SVM). Studies often use accuracy, precision, recall, and F1-score to evaluate model performance. see [Realm \(2023\)](#) for an example of this.

**Dealing with Imbalanced Data:** Since the number of survivors is lower than non-survivors, class imbalance is addressed in some studies. Techniques such as resampling (oversampling the minority class or undersampling the majority class) or using metrics like ROC-AUC, which are

more suited to imbalanced data, have been employed to handle this issue. See [CodeSignal](#) for an example of this.

**Regression Analysis:** Regression techniques have been used to predict continuous variables like "Fare" based on attributes like "Pclass," "Age," and "SibSp." This allows for an exploration of how socio-economic factors affected fare prices and the relationships between other numerical features. See [Realm \(2023\)](#) for an example of this.

**Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques.**

In the context of the Titanic dataset, applying classification and regression techniques will help us achieve a deeper understanding of the factors affecting survival and predict outcomes based on these factors.

CLASSIFICATION:

**Objective:** The goal of classification is to predict whether a passenger survived or perished based on their characteristics.

**Application:** We can use classification algorithms, such as logistic regression, decision trees, or support vector machines, to build a model that predicts the survival status (survived or not) of passengers in the test dataset.

**Learning Outcome:** This will help us identify the most significant factors influencing survival, such as class, age, and gender. It will also allow us to evaluate how well these factors can predict survival and how different characteristics are weighted in the prediction process.

REGRESSION:

**Objective:** The goal of regression is to predict a continuous outcome based on the passenger's characteristics. In the Titanic dataset, this can be approached by predicting a variable like the fare paid by the passengers or, more creatively, estimating the probability of survival as a continuous variable.

**Application:** We can use regression algorithms, such as linear regression, to model relationships between predictors (e.g., age, fare, family size) and the continuous outcome.

**Learning Outcome:** This approach can reveal how numerical factors, such as the fare paid or age, correlate with survival probability or other outcomes. It can provide insights into how these continuous variables influence the likelihood of survival and offer a quantitative perspective on their impact.

**Explain which attribute you wish to predict in the regression based on which other attributes? Which class label will you predict based on which other attributes in the classification task?**

In the Titanic dataset, we can approach the classification and regression tasks as follows:

CLASSIFICATION TASK:

**Class Label to Predict:** Survived (a binary outcome where 0 indicates the passenger did not survive and 1 indicates survival).

**Predictor Attributes:** We will predict survival based on attributes such as Pclass, Age, Sex, Fare, SibSp, Parch, and Embarked. These attributes are likely to influence whether a passenger survived the sinking of the Titanic, and analyzing their impact can provide insights into survival patterns and key factors affecting survival rates.

REGRESSION TASK:

**Attribute to Predict:** Fare.

**Predictor Attributes:** We will predict the fare based on attributes such as Pclass (passenger class), Age, SibSp (number of siblings/spouses aboard), Parch (number of parents/children aboard), Sex, and Embarked (port of embarkation). These predictors can help us understand how different factors influence the fare a passenger paid.

By using these attributes in classification and regression, we aim to build models that can predict the survival status and fare of passengers, respectively, based on their characteristics and other relevant information.

**If you need to transform the data in order to carry out these tasks, explain roughly how you plan to do this.**

To effectively carry out the classification and regression tasks, we will need to transform the Titanic dataset to address various quality issues and prepare it for analysis. Here's a rough plan for transforming the data:

- **Handling Missing Values:**

- *Age:* Fill missing values using imputation techniques such as the mean or median age, or use more sophisticated methods like predicting age based on other attributes (e.g., passenger class and title).
- *Embarked:* Fill missing values with the most common port of embarkation or use imputation based on other features.
- *Cabin:* Since the cabin feature is sparse, we might consider dropping it or extracting and using only the information about whether a cabin was known or not.

- **Standardizing and Encoding:**

- *Categorical Variables:* Convert categorical variables like Sex, Embarked, and Pclass into numerical format using techniques such as one-out-of-K encoding or label encoding. For example, Sex can be converted into binary (0 for male, 1 for female), and Embarked can be one-out-of-K encoded into separate columns for each port of embarkation.
- *Passenger Name:* This feature can be dropped or used to extract titles (e.g., Mr., Mrs., Miss) that might provide additional insights.

- **Handling Outliers:**

- *Fare:* Check for outliers in the Fare attribute and decide whether to remove or transform them. For instance, applying a log transformation might help normalize the distribution.

- **Feature Engineering:**

- *Family Size:* Combine SibSp and Parch to create a new feature representing the total family size aboard.

- *Title Extraction:* Extract titles from the Name attribute to create a feature that might be useful for predicting survival.

- **Normalization and Scaling:**

- *Numerical Features:* Scale numerical features like Age and Fare to ensure they are on a comparable scale, which can improve the performance of many machine learning algorithms.

- **Data Splitting:**

- *Train-Test Split:* Ensure proper splitting of the dataset into training and testing sets for validation of the models.

By addressing these data quality issues and performing these transformations, we will prepare the dataset for accurate and meaningful analysis, enabling us to build effective regression and classification models.

## Task 2: Data Attribute Description and Issues

### Attribute Types and Measurement Scales

This section\* provides a detailed explanation of the attributes in the Titanic dataset, categorizing each as discrete or continuous and identifying their corresponding measurement scales: Nominal, Ordinal, Interval, or Ratio. Understanding the nature of these variables is essential for selecting the appropriate analysis techniques in subsequent stages. For a full overview of the attributes, see table 2 in the appendix.

Each of these attributes serves a distinct role in analysis, with discrete variables often used for classification and categorical comparisons, and continuous variables applied in regression or to explore relationships like fares and age.

### Data Quality Issues

The Titanic dataset contains several data quality issues that must be addressed before analysis. The primary problems identified are as follows:

#### MISSING VALUES

- **Age:** A significant number of passengers have missing values for the Age attribute, which could impact the analysis of age as a factor in survival.
- **Cabin:** The Cabin attribute has a large proportion of missing values, making it difficult to use for meaningful analysis. In many cases, this variable is either dropped or transformed into a binary feature (whether a cabin number was recorded or not).
- **Embarked:** A few missing values are present in the Embarked attribute, which indicates the port of boarding.

#### CORRUPTED OR INCONSISTENT DATA

- **Ticket:** The Ticket attribute consists of alphanumeric codes with no consistent format, leading to difficulty in interpreting this feature. It may need to be treated as a nominal attribute or disregarded altogether.

- **Fare:** There are potential outliers in the Fare attribute, as some values seem unusually high or low, which could skew the analysis if not handled properly.

## DATA STANDARDIZATION

- **Name:** The Name attribute contains titles (e.g., Mr., Mrs., Miss) embedded within full names. Extracting the title could be useful for analysis, but without standardization, this information is not immediately usable.

These data issues highlight the need for preprocessing steps, such as handling missing values, addressing outliers, and standardizing certain attributes, to ensure the dataset is ready for analysis.

## Summary statistics

We have calculated summary statistics for all the numerical variables. The statistics are seen in the table below. There was an outlier of roughly 512 in the Fare variable that we have chosen to remove. We will talk more about this when we discuss our principal component analysis.

	Age	SibSp	Parch	Fare
count	714.000000	891.000000	891.000000	891.000000
mean	29.699118	0.523008	0.381594	32.204208
std	14.526497	1.102743	0.806057	49.693429
min	0.420000	0.000000	0.000000	0.000000
25%	20.125000	0.000000	0.000000	7.910400
50%	28.000000	0.000000	0.000000	14.454200
75%	38.000000	1.000000	0.000000	31.000000
max	80.000000	8.000000	6.000000	263.329200

Table 1: Summary statistics for the numerical variables in the dataset.

The summary makes the presence of missing values clear, as we can see the count for the age variable is lower than the other variables. We can also spot the presence of some potential outliers in the SibSp, Parch, and Fare variables, as the respective third quartiles are significantly lower than the respective maxima.

## Task 3

### Issues with outliers

We plot the numerical variables as box plots in figure 1. We can see from the boxplots that Parch and SibSp have most of their values clustered around 0 and the rest appear to be outliers. There is not much interest in continuing with these columns as they do not provide much information. We can drop these columns for now.

The Age and Fare attributes on the other hand are much more interesting. We can see that there are some outliers in the data. Values above the highest whiskers represent the outliers. For Age the ones slightly above 60 are outliers, and for Fare, the ones above around 80 look to be outliers.

We believe that the outliers are still fundamental when it comes to analysis of survivability of a passenger, and Age and Fare prices are believed to be important factors, so we want to keep these. The only outlier we would remove would be the fare price of 512.3292. This results in the box plots of figure 2.

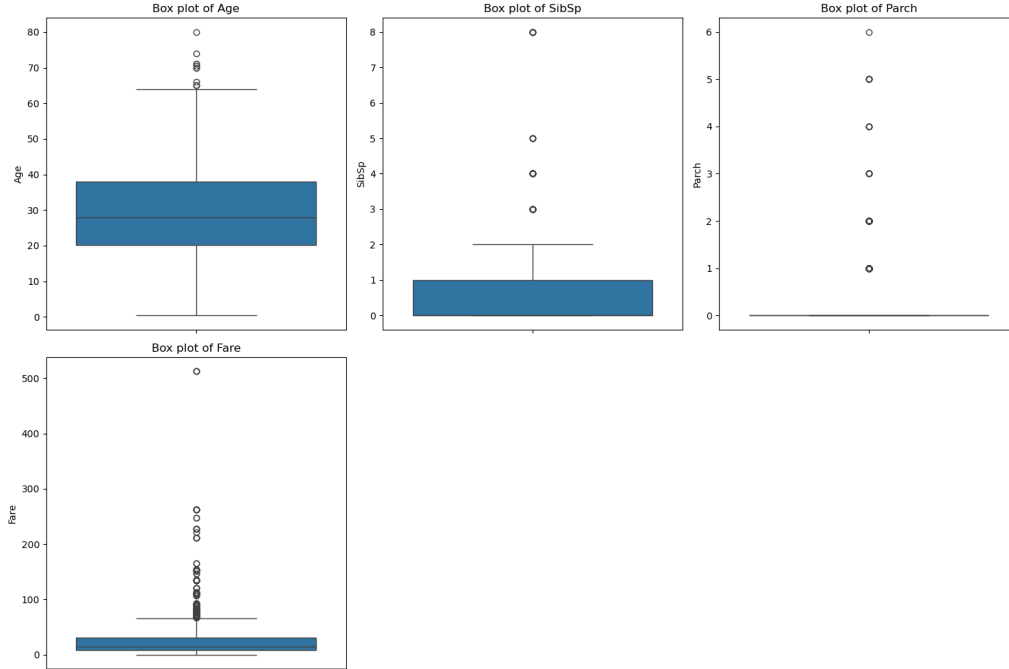


Figure 1: Box plots of the variables Age, SibSp, Parch, and Fare.

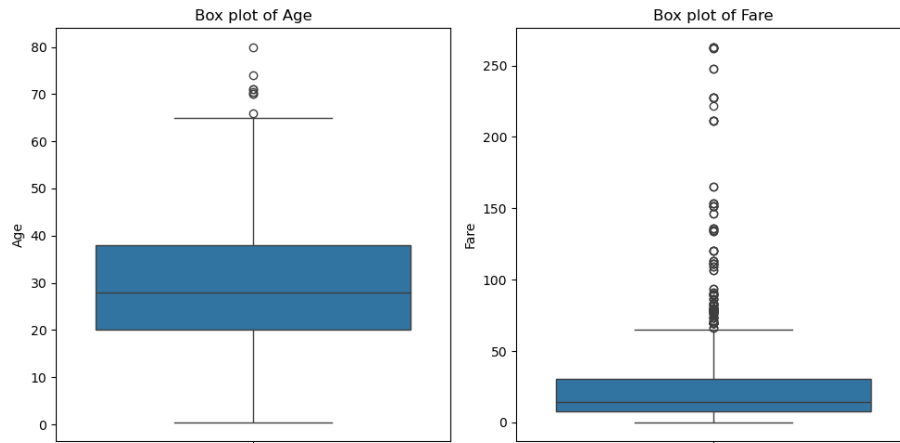


Figure 2: Box plots of Age and Fare with Fare outlier removed.

### Do the attributes appear to be normally distributed?

In order to determine that, we first need to observe our dataset and check if the observations in each attribute are continuous. This is because the normal distribution is continuous. From our analysis in question 2, we see that age and fare are two continuous attributes. One simple way to acknowledge if the data is normally distributed is to plot a histogram as is done on figure 3.

We can see that the result above shows that the data is not very normally distributed. Age is slightly bell shaped, but it skews slightly to the right whilst the fare distribution is heavily skewed to the right. This is due to the numerous outliers present in the attributes. If we want to normally distribute them more, we would have to log transform to reduce skewness.

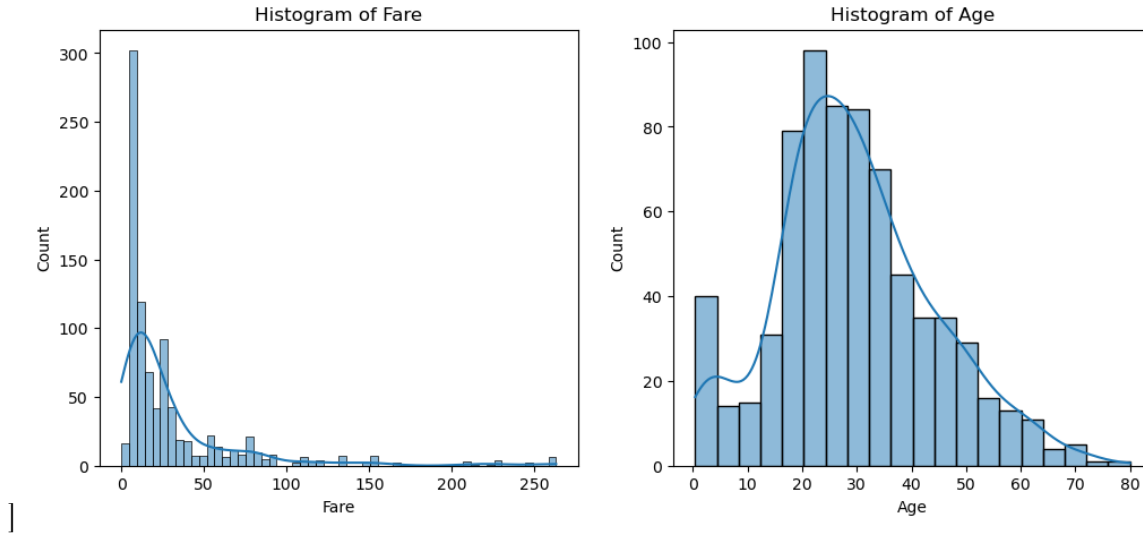


Figure 3: Histogram of Age and Fare

### Are the variables correlated?

We can use a correlation matrix to determine if the variables are correlated, but before that, we need to encode categorical variables as numerical ones. There are several categorical variables, but we believe sex and pClass to be the most significant in determining survivability of a passenger. We compute the correlation matrix as follows:

```
# Make another copy of the dataframe where df contains the entire dataset
cor_df = df
# Code below to encode sex whilst pClass is already numerical
label_encoder = LabelEncoder()
cor_df.loc[:, 'Sex'] = label_encoder.fit_transform(df['Sex'])
# Drop columns that are not of significance
cor_df = df.drop(columns=['PassengerId', 'Name', 'Ticket', 'Cabin', 'Embarked'])
# Compute the correlation matrix
correlation_matrix = cor_df.corr()
# Display the correlation matrix
print(correlation_matrix)
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
Survived	1.000000	-0.334068	-0.545899	-0.079472	-0.033395	0.082157	0.261742
Pclass	-0.334068	1.000000	0.132881	-0.368625	0.080937	0.018212	-0.604960
Sex	-0.545899	0.132881	1.000000	0.093296	-0.114799	-0.247003	-0.222361
Age	-0.079472	-0.368625	0.093296	1.000000	-0.307639	-0.189194	0.100396
SibSp	-0.033395	0.080937	-0.114799	-0.307639	1.000000	0.415141	0.211816
Parch	0.082157	0.018212	-0.247003	-0.189194	0.415141	1.000000	0.263910
Fare	0.261742	-0.604960	-0.222361	0.100396	0.211816	0.263910	1.000000

Above we can see the correlation matrix of the various attributes together. We can see that when it comes to sex and survived, there is a moderate negative correlation of -0.54. This suggests that passengers being male (encoded as 1) are less likely to survive. There is also a negative correlation



of class and survival, which suggests that passengers in higher classes are more likely to survive. Finally the fare has a relatively moderate positive correlation with survival, which suggests that passengers who paid more for their fare are more likely to survive.

When looking away from target variables, we can also see some other correlations passenger class and age, where there's a moderate correlation indicating that higher class has a moderate relationship with increased age. We also see a moderate to strong correlation between the class and the fare indicating that the higher the passenger class of the ticket, the higher the fare price. Other notable and fairly obvious correlations are the instances of siblings and the instances of parents, which have a moderate positive correlation.

### **Does the primary machine learning modeling aim appear to be feasible based on your visualizations?**

This dataset is primarily a classification problem, where we are trying to predict whether a passenger survived or not. Here the target variable is survived which has a binary outcome, making it suitable for classification problems.

Based on the correlation matrix results, we can see that there are some attributes that are correlated with the target variable and most are explained in the first paragraph in this cell. Through visualization, we can see that the age and fares are relatively normally distributed but skewed, especially fares, which could affect the model's performance, which might necessitate a transformation such as log transformation to reduce skewness.

Classification is the most feasible given the correlation matrix and the visualizations we have done so far, which could include models such as logistic regression and decision trees.

Linear regression could be used to explore the relationship between age and fare, although given the skewness and the weak positive correlation between the two, it might require some additional data manipulation to improve the model's performance.

### **The principal component analysis**

First thing we need to do is to clean the data to prepare it for PCA. Recall that, PCA works on numerical data. We therefore drop the attributes Name, Ticket, Cabin, and PassengerId as these are qualitative. We further convert the categorical variables Sex and Embarked into indicator variables. The remaining attributes are of interest when it comes to the target variable survivability. Following our conversions, no non-numerical data remains.

One last thing before we apply PCA, we need to standardize the data. This is because PCA is sensitive to the scale of the data. One reason is that if we compare age to fare, age can vary from 0 to 100, whilst fare can vary from 0 to 1000. This means that the variance in the data is dominated by fare. We need to standardize the data so that the variance in the data is not dominated by one attribute. We can do this by subtracting the mean and dividing by the standard deviation of each attribute:

```
# Standardize the data using the StandardScaler import.  
# Mathematically this is fairly simple, as it involves Z-scoring the data.  
scaler = StandardScaler()  
# fit calculates the mean and standard deviation of each attribute  
# and transform applies the Z score formula to each attribute.  
# df_clean is the cleaned dataset.  
df_standardized = scaler.fit_transform(df_clean)
```

All the standardized values show how many standard deviations away from the mean the data is. There is one value that is 4.94 standard deviations away from the mean. Finally, we perform PCA using the singular value decomposition:

```
U, S, Vt = svd(df_standardized, full_matrices=False)
# Compute variance explained by principal components:
rho = (S * S) / (S * S).sum()
```

We define a threshold of 90% of the variance explained to see how many principal components are needed to explain most of the variance. We plot the variance explained to see how many principal components are needed to surpass this threshold. The plot is seen in figure 4.

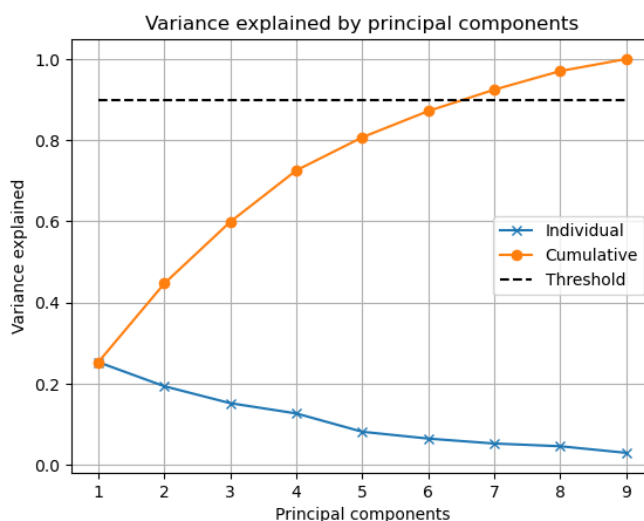


Figure 4: Plot of variance explained by against number of principal components with threshold of 90%.

How can we interpret this result? It is obvious from the graph above that we need at least 7 principal components to explain at least 90% of the variance, which is 2 less dimensions than the cleaned dataset and 6 less dimensions than the original dataset. This is a reduction in dimensionality whilst still being able to explain most of the spread of the original data.

Unfortunately we can only visualize up to 3D data, which in this dataset only amounts to 60% of the variance explained. This is the tradeoff of PCA, we lose some information but we gain interpretability and computational efficiency.

We will now attempt to interpret the 7 principal components needed to explain 90% of the variance. This is seen in figure 5.

How can we interpret this? Each principle direction is an eigenvector orthonormal to each other where the direction yields the maximized variance of the projected datapoints. The first principle component always yields the highest proportion of the variance. This then tapers as we move onto the next few principle components.

The bars represent the contributions of each feature to the principle component, where the sign either positive or negative and magnitude of these coefficients indicate the direction and strength of each feature's contribution. A positive bar means that the feature contributes positively to the component and vice versa. Magnitude shows the strength of each contribution.

We interpret each of the principal components to capture the following patterns:

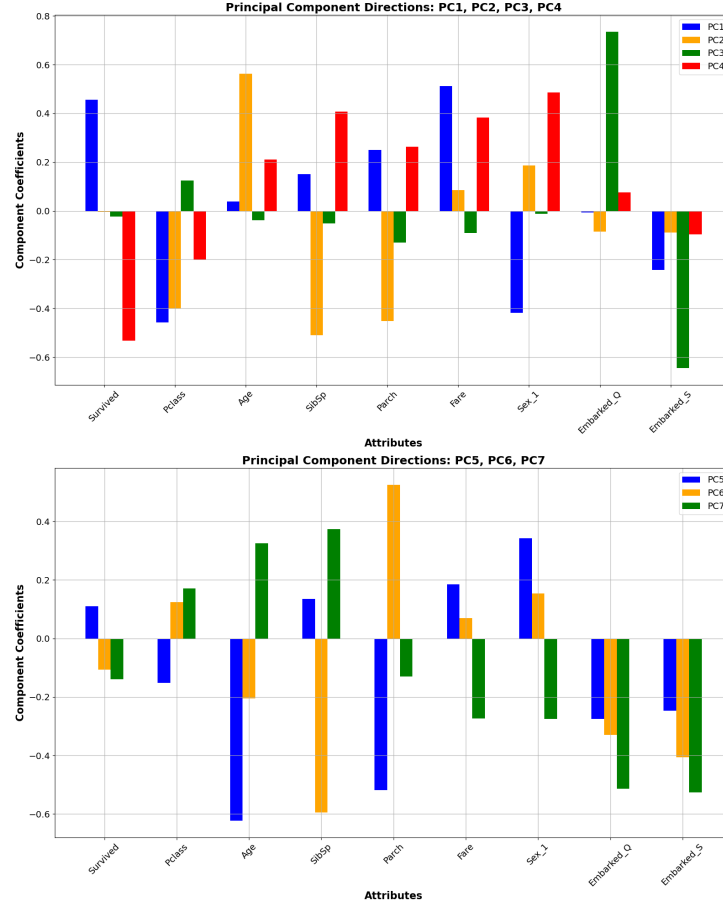


Figure 5: The principal components' values on each attribute.

- Principal component 1: Attributes like “Survived”, “Sex\_1”, “Fare” and “pClass” have relatively high positive coefficients, suggesting they are key factors driving the variance captured by principal component 1. These attributes seem to explain variances concerning socioeconomic status, with survivability being driven by higher class, higher fare prices and not being male.
- Principal component 2: The main attribute with the highest magnitude seems to be the positive age. It looks like higher age is associated with higher pClass, not travelling with children, siblings as well as being male and slightly increased fare prices.
- Principal component 3: Seems to capture where a person is embarked and it seems like the underlying pattern is that if a person embarked from Q, they certainly didn't from S.
- Principal component 4: Seems to revolve around not surviving the journey and the other bar charts are relatively moderate indicating that if a person was a middleclass male, there seems to be a downward forcing survivability rate.
- Principal component 5: Here the underlying pattern seems to be that if a person is young, there seems to be an inverse relationship with number of parents children aboard.
- Principal component 6: Looks like this principle component captures the relationship mostly between instances of parents / children aboard and number of siblings aboard, which makes sense as parents with children onboard would often have several children with them.

- Principal component 7: Difficult to interpret, maybe a pattern of how a person embarked?

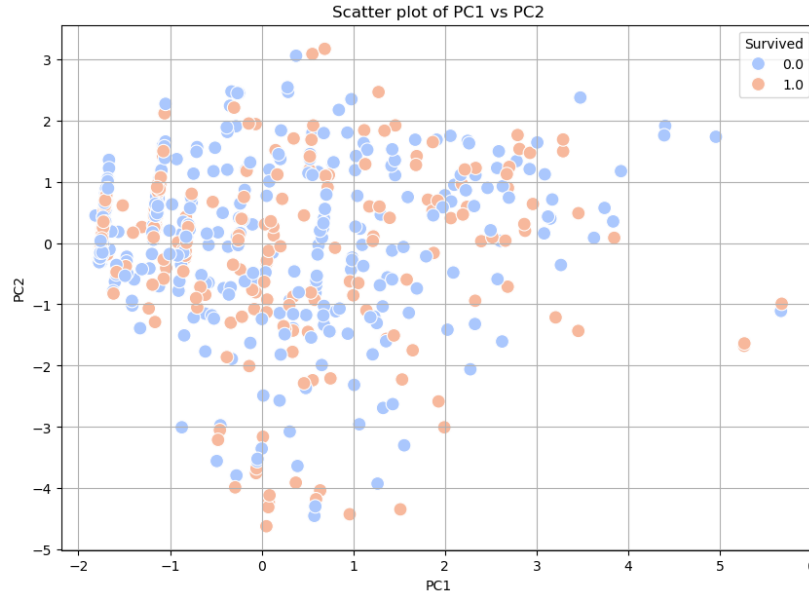


Figure 6: The data projected onto the first two principal components.

Now we describe the data projected onto the considered principal components. The 2-dimensional graph seen in figure 6 do not separate the two binary outcomes into clear clusters, suggesting that the first two principle components do not capture the variance too well. This makes sense given that the first two principle components only explains a little under half of the variance whilst the optimal solution is to get them to explain at least 90%. This proved to be only doable with 7 principle components. Maybe projecting to 3 dimensions would ease this.

The 3-dimensional graph is seen in figure 7. We get a slightly clearer view here although the first 3 components seem to only capture around 60% of the variance. Based on what we can observe here, we do see some semblance of two clusters being formed. Based on this static 3D plot, it looks like we can see a cluster for individuals not surviving being formed by low values of principal component 1, relatively high values of principal component 2, and non-positive values of principal component 3.

With negative principal component 1 values, we previously interpreted this as an individual having relatively low socio-economic status, which makes sense given that they were often farther down the ship. We also see that the cluster along principal component 2 seems to be based on being slightly older being a cluster of fatalities there too, which seems to make sense given that younger people, especially children would survive more.

Lastly, principal component 3 is much harder to interpret as it seems to show the relationship based on the embarking of a passenger. The effect of this is unclear.

## Task 4

In the preceding three task, we have explored the Titanic dataset to assess its quality and suitability for classification and regression tasks. The key interest in the dataset is the possibility of finding patterns in the demographic factors of the passengers that could explain their survival rate. This is a noteworthy problem, as the Titanic carried insufficient lifeboats to save all of its passengers. It is

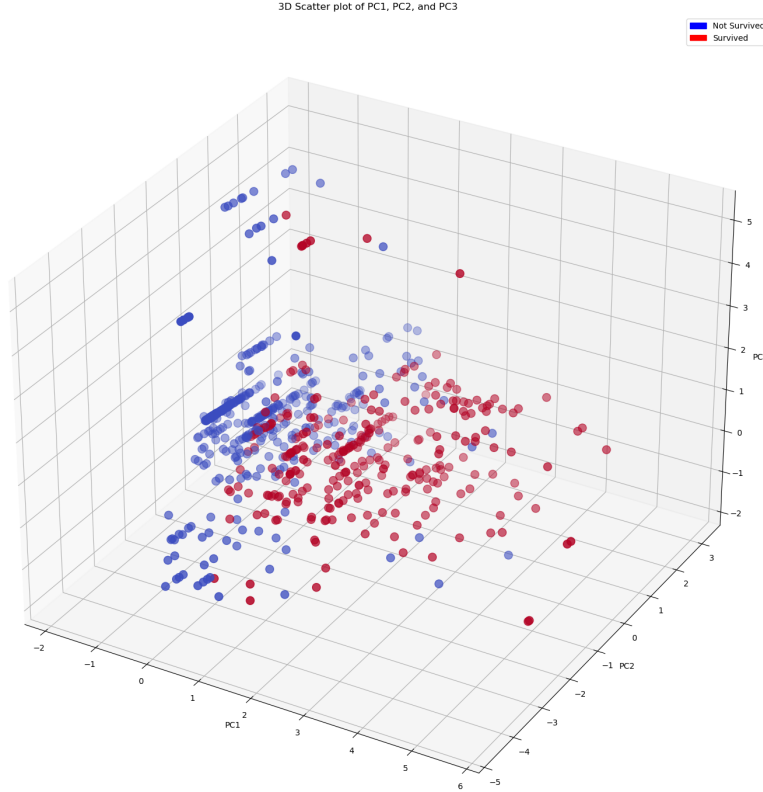


Figure 7: The data projected onto the first three principal components.

therefore insightful to determine whether the crisis adversely affected certain demographic groups or if survival was arbitrary.

This is by no means a novel analysis. Many previous analyses exist that also deal with the problem of detecting any pattern between the other variables and survivorship. The chief challenge that all prior analyses have faced is the quality of the data. Some variables have missing data. Also, survivorship is not a balanced variable. This results in some data wrangling challenges. However, since most of the variables are typical demographic variables, it should be possible to treat them all using standard techniques, e.g. imputation for missing values, and still obtain decent results.

Our principal component analysis reveals that a significant number of principal components is needed to explain most of the variance in the dataset. This, again, convolutes any potential classification or regression analysis, as many of the datasets dimensions will need to be accounted for. And even with the dimensions accounted for, any clear pattern may not be guaranteed.

However, none of these challenges foreclose statistical analysis. Most datasets are somewhat messy and need cleaning. The Titanic dataset is no different. Our principal component analysis demonstrates that it is possible to make sense of the dimensions captured in the dataset. Prior analyses also demonstrate the feasibility of working with the dataset. We therefore judge that we will be able to complete our classification and regression analyses of the data.

## Exam Problems

### Question 1

The only correct statement is C. The justification is:

- The variable  $x_1$  represents the time of day in blocks of 30 minutes that partition a day into a finite number of intervals. The variable is therefore discrete. Since the intervals can be ordered, the variable is ordinal.
- The attribute  $x_6$  is a ratio variable. The number of broken traffic lights is clearly a numeric variable. The canonical zero of the variable is zero broken traffic lights.
- The attribute  $x_7$  is a ratio variable for much the same reasons as for  $x_6$ .
- The congestion level is ordinal as it is a discrete variable that can be ordered.

The statements A, B, and D are all incorrect as they mistake  $x_1$  for something other than an ordinal variable.

### Question 2

A. This statement is correct since  $|26 - 19| = 7$  and that is the maximal difference among the respective coordinates of the two vectors.

B. The metric is

$$d_3(x_{14}, x_{18}) = \sqrt[3]{|26 - 19|^3 + |2 - 0|^3} \approx 7.05.$$

Thus, the statement is incorrect.

C. The metric is

$$d_1(x_{14}, x_{18}) = |26 - 19| + |2 - 0| = 9.$$

Thus, the statement is incorrect.

D. The metric is

$$d_4(x_{14}, x_{18}) = \sqrt[4]{|26 - 19|^4 + |2 - 0|^4} \approx 7.01.$$

Thus, the statement is incorrect.

### Question 3

A. The explained variance is

$$\frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.87.$$

The statement is therefore correct.

B. The explained variance is

$$\frac{11.48^2 + 10.03^2 + 9.45^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.48.$$

Thus, the statement is incorrect.

C. The explained variance is

$$\frac{13.9^2 + 12.47^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.52.$$

Thus, the statement is incorrect.

D. The explained variance is

$$\frac{13.9^2 + 12.47^2 + 11.48^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.72.$$

Thus, the statement is incorrect.

#### Question 4

- A. Such an observation will typically have a negative value as the high values of the third and fourth coordinates will make the negative coordinates of the principal component dominate. The statement is false.
- B. For much the same reasons as in A, such an observation will typically have a negative value. This statement is false.
- C. For such an observation, the positive value of the second coordinate of the principal component will dominate the sum in the dot product. The value will therefore typically be positive. The statement is false.
- D. The only negative coordinate of the principal component is the first one. As the observation has a low value in this position, and high values everywhere else, the dot product will typically be positive. The statement is correct.

#### Question 5

We are given the following data:

$$n = 20000, M_{11} = 2, M_{01} = 5, M_{10} = 6.$$

The Jaccard similarity of the documents is

$$\frac{M_{11}}{M_{11} + M_{01} + M_{10}} \approx 0.1538.$$

The correct answer is A.

#### Question 6

The probability  $p(\hat{x}_2 = 0 \mid y = 2)$  can be found by marginalizing on  $\hat{x}_7$ . This results in the probability

$$p(\hat{x}_2 = 0 \mid y = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0 \mid y = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1 \mid y = 2) = 0.81 + 0.03 = 0.84.$$

The correct answer is B.

## Appendix A. Description of dataset

Attribute	Discrete/ Continuous	Type	Explanation
Survived	Discrete	Nominal	This is a binary attribute (0 = did not survive, 1 = survived) with no inherent order between the categories. It simply represents a classification into two distinct groups.
Pclass (Passenger Class)	Discrete	Ordinal	Pclass indicates the socio-economic class of the passenger (1 = first class, 2 = second class, 3 = third class). Although it is a discrete variable, it is ordinal because there is a clear hierarchy (first class is higher in rank than second or third).
Name	Discrete	Nominal	The Name attribute is a string (text) variable that uniquely identifies passengers. Since there is no inherent order or numerical relationship between names, it is nominal.
Sex	Discrete	Nominal	The Sex attribute represents the gender of the passenger (male/female). This is a nominal variable as the categories are distinct and there is no order.
Age	Continuous	Ratio	Age is a continuous variable that represents the passenger's age. Since age has a meaningful zero point (birth) and the differences between values are consistent, it is measured on a ratio scale.
SibSp (Number of Siblings/Spouses Aboard)	Discrete	Ratio	SibSp is a count of how many siblings or spouses a passenger had aboard. It is discrete because it represents a count, and ratio because it has a meaningful zero and equal intervals.
Parch (Number of Parents/Children Aboard)	Discrete	Ratio	Parch is a count of how many parents or children a passenger had aboard. Like SibSp, it is a discrete variable on a ratio scale with a meaningful zero.
Ticket	Discrete	Nominal	The Ticket attribute is an identifier for the passenger's ticket. It is nominal because it consists of text or numerical codes that do not have any intrinsic order.
Fare	Continuous	Ratio	Fare represents the amount of money the passenger paid for the ticket. It is a continuous variable with a meaningful zero (no fare), and the differences between values are meaningful, making it a ratio variable.
Cabin	Discrete	Nominal	The Cabin attribute is a string that represents the cabin number assigned to the passenger. It is nominal because the cabin numbers are simply labels with no inherent numerical or ordered relationship.
Embarked (Port of Embarkation)	Discrete	Nominal	This attribute indicates the port where the passenger boarded the Titanic (C = Cherbourg, Q = Queenstown, S = Southampton). Since there is no natural ordering between these ports, it is a nominal variable.

Table 2: Summary of Titanic Dataset Attributes by Type and Measurement Scale



## References

- CodeSignal. Data preprocessing: The titanic dataset exploration. URL <https://learn.codesignal.com/preview/lessons/739/data-preprocessing-the-titanic-dataset-exploration>.
- Benedikt Droste. Machine learning with the titanic dataset, April 2020. URL [https://people.sc.fsu.edu/~jburkardt/classes/ml\\_2023/datasets/titanic.pdf](https://people.sc.fsu.edu/~jburkardt/classes/ml_2023/datasets/titanic.pdf).
- Hackers Realm. Titanic dataset analysis using python (kaggle) | classification | machine learning project tutorial, June 2023. URL <https://www.hackersrealm.net/post/titanic-dataset-analysis-using-python>.