02450 Report 1

Table of authors and their contributions

Name	Gisle Joe Garen	Ignacio Ripoll González	Sebastian Svelmøe Timm
Student ID	s242712	s242875	s243935
Introduction	10%	80%	10%
Task 1	10%	80%	10%
Task 2	45%	45%	10%
Task 3	80%	10%	10%
Task 4	10%	10%	80%
Problems	10%	10%	80%

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.

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 studies have explored different approaches to data preprocessing and model building. Here's a summary of the common steps and results found in the literature:

Data Cleaning and Feature Engineering: Many papers focused on addressing missing values, especially for key variables like Age, Cabin, and Embarked. Missing Age values were often imputed using the median or by predicting age based on other features like Pclass and Sex. The Cabin feature, due to its sparsity, was often dropped or transformed into a binary feature indicating whether or not the cabin number was available. Researchers frequently engineered new features, such as creating a "family size" variable by combining SibSp and Parch, or extracting titles (Mr., Mrs., Miss, etc.) from the Name attribute, which provided additional predictive power for survival models.

Classification Techniques: Classification models like logistic regression, decision trees, and random forests were commonly applied to predict the survival outcome (Survived). Studies showed that factors such as Pclass, Sex, and Age were consistently among the most important predictors of survival, with women and first-class passengers having a significantly higher chance of survival.

Dealing with Imbalanced Data: Given that the Titanic dataset has a higher proportion of passengers who perished (about 62%) compared to those who survived (38%), some studies used techniques to address this imbalance, such as oversampling the minority class (survivors) or using undersampling of the majority class (non-survivors). Advanced techniques such as SMOTE (Synthetic Minority Over-sampling Technique) were also applied to improve the performance of classification models in handling this imbalance.

Regression Analysis: Some papers also performed regression analysis, focusing on predicting Fare as a continuous variable or using logistic regression to estimate the probability of survival as a function of the available features. Regression models helped identify how much socio-economic factors (e.g., Pclass, Fare, Age) contributed to survival probability.

Overall, previous studies demonstrated that proper data cleaning, feature engineering, and handling of imbalanced data significantly improved model performance. They consistently showed that social class, gender, and age were the most critical factors in determining survival, providing a clear picture of the socio-economic dynamics of the Titanic disaster.

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.

THIS IS A REFERENCE 1.

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.

Task 3

TBD

[1]:
```
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

# Read the CSV file
titanic_data = pd.read_csv('dataset/train.csv')

# Rename the data to a dataframe for semantics
df = titanic_data

# Let's see what the attribute values are:
print(df.columns)
```

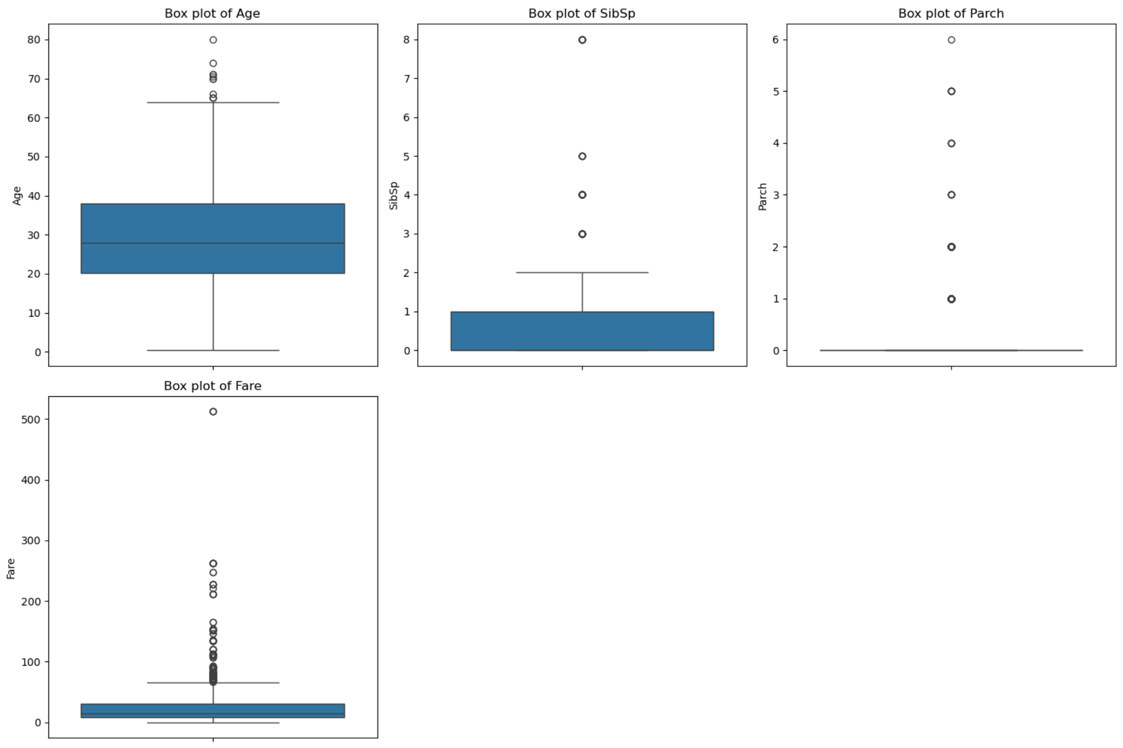
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

[2]:
```
# First let's explore the initial data where the task asks us to view any␣
 ↪issues with outliers:

# Drop passengerId, survived and Pclass columns because they don't have any␣
 ↪outliers, they are more categorical or unqiue identifier
numerical_columns = df.drop(columns=['PassengerId', 'Survived','Pclass'])

# Get the summary statistics of the numerical columns
numerical_columns = numerical_columns.select_dtypes(include=[np.number])
summary_statistics = numerical_columns.describe()

# Display the summary statistics
print(summary_statistics)
```

```
              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  512.329200
```

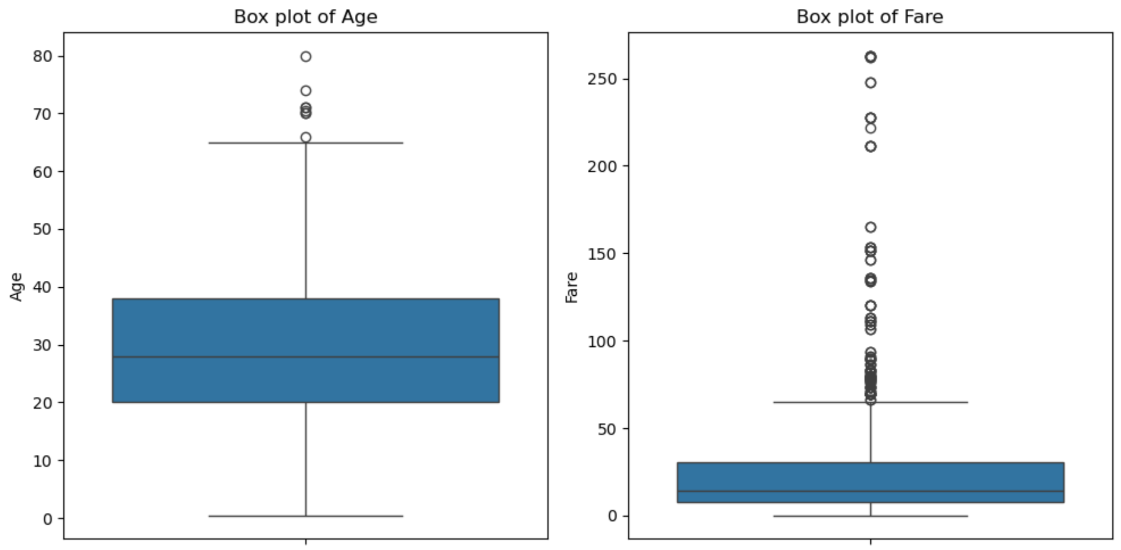
[3]:
```
# Plot box plots for each numerical column to visualize outliers
plt.figure(figsize=(15, 10))
```

```python
for i, column in enumerate(numerical_columns.columns, 1):
    plt.subplot(2, 3, i)
    sns.boxplot(data=df, y=column)
    plt.title(f'Box plot of {column}')
    plt.tight_layout()

plt.show()
```



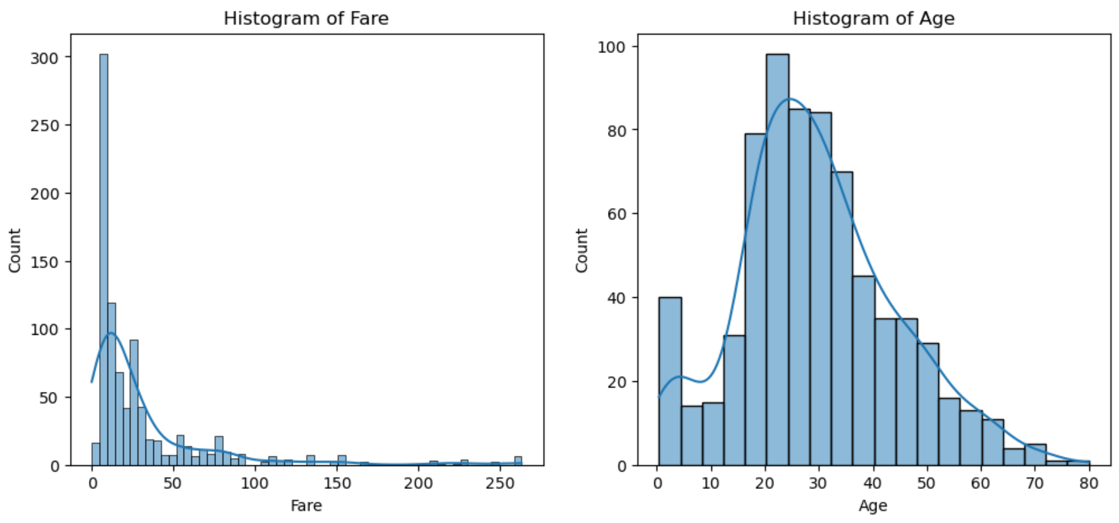
[4]:
```python
# We can see from the boxplots that parch and sibsp have most values clustered␣
↪around 0 and the rest are basically outliers. There's not much
# interest in continuing with these columns as they don't provide much␣
↪information. We can drop these columns.
numerical_columns = df.drop(columns=['Parch','SibSp'])

# 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 and with age the␣
↪ones slightly above 60 are outliers and with the fare
# the ones above around 80 look to be outliers.

# I 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. Only outlier␣
↪I would remove would be the fare price of 512.3292.

# Code below to remove the observation with the fare price of 512.3292 as this␣
↪is a significant outlier almost twice as much as the next highest fare price.
df = df[df.Fare <= 500]

# Select only the Age and Fare columns for plotting
columns_to_plot = ['Age', 'Fare']

# Plot box plots again
plt.figure(figsize=(10, 5))
for i, column in enumerate(columns_to_plot, 1):
    plt.subplot(1, 2, i)  # Adjust subplot grid to 1x2
    sns.boxplot(data=df, y=column)
    plt.title(f'Box plot of {column}')
    plt.tight_layout()

plt.show()
```



[5]:
```
# Do the attributes appear to be normal 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 and from previous␣
↪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
# Plot histogram for Fare
plt.figure(figsize=(12, 5))

# plot.subplot(1, 2, 1) means 1 row, 2 columns, and the first plot
plt.subplot(1, 2, 1)
sns.histplot(df['Fare'].dropna(), kde=True)
plt.title('Histogram of Fare')

# Plot histogram for Age
plt.subplot(1, 2, 2)
sns.histplot(df['Age'].dropna(), kde=True)
plt.title('Histogram of Age')
```

[5]:
```
Text(0.5, 1.0, 'Histogram of Age')
```



[6]:
```
# 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.

# Are variables correlated?
# We can use a correlation matrix to determine if the variables are correlated,␣
  ↪but before that we need to encode categorical
# variables to numerical ones. There are several categorical variables, but sex␣
  ↪and pClass are believed to be the most significant
```

```
# in determining survivability of a passenger.

# Make another copy of the dataframe
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
```

[7]:
```
# Above we can see the correlation matrix of the various attributes together.
# We can see that when it comes to sex and survived, we can see that 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.

# 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.
```

[8]:
```
# First thing we need to do is to clean the data to prepare it for PCA.␣
↪Remember, PCA works on numerical data,
# whilst most of the attributes below are qualitative.
# Drop columns that are not useful for PCA or have too many missing values
df_clean = df.drop(columns=['Name', 'Ticket', 'Cabin', 'PassengerId'])

# Probably better ways to do this, but for now imma just drop the rows with␣
↪missing values
df_clean = df_clean.dropna()

# Convert categorical variables into dummy/indicator variables
# Converts a category attribute like sex and embarked into a binary attribute.␣
↪This creates a column sex_male and drops sex_female as
# it's redundant, either its true or not. Embarked has value S, C or Q so it␣
↪creates embarked q and s, dropping embarked c as it's redundant,
# done by drop_first=true.
df_clean = pd.get_dummies(df_clean, columns=['Sex', 'Embarked'],␣
↪drop_first=True)

# Let's see how the dataset looks like now:
print(df_clean.head())
```

```
   Survived  Pclass   Age  SibSp  Parch     Fare  Sex_1  Embarked_Q  \
0         0       3  22.0      1      0   7.2500   True       False
1         1       1  38.0      1      0  71.2833  False       False
2         1       3  26.0      0      0   7.9250  False       False
3         1       1  35.0      1      0  53.1000  False       False
4         0       3  35.0      0      0   8.0500   True       False

   Embarked_S
```

```
0        True
1       False
2        True
3        True
4        True
```

[9]:
```
# The attributes above are of interest when it comes to the target variable␣
  ↪being survivability and we can see that we have no more
# missing attributes that are not numerical.

# 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
# with Z = (X - myu) / sigma. This forms the basis of a normal distribution.
scaler = StandardScaler()
# fit calculates the mean and standard deviation of each attribute and␣
  ↪transform applies the Z score formula to each attribute.
df_standardized = scaler.fit_transform(df_clean)

# Let's see how it looks like now:
print(df_standardized)
```

```
[[-0.81985963  0.90464078 -0.52506921 …  0.75634786 -0.20277082
   0.52894555]
 [ 1.21972099 -1.49308002  0.57772613 … -1.32214296 -0.20277082
  -1.89055377]
 [ 1.21972099  0.90464078 -0.24937037 … -1.32214296 -0.20277082
   0.52894555]
 …
 [ 1.21972099 -1.49308002 -0.73184333 … -1.32214296 -0.20277082
   0.52894555]
 [ 1.21972099 -1.49308002 -0.24937037 …  0.75634786 -0.20277082
  -1.89055377]
 [-0.81985963  0.90464078  0.16417788 …  0.75634786  4.93167604
  -1.89055377]]
```

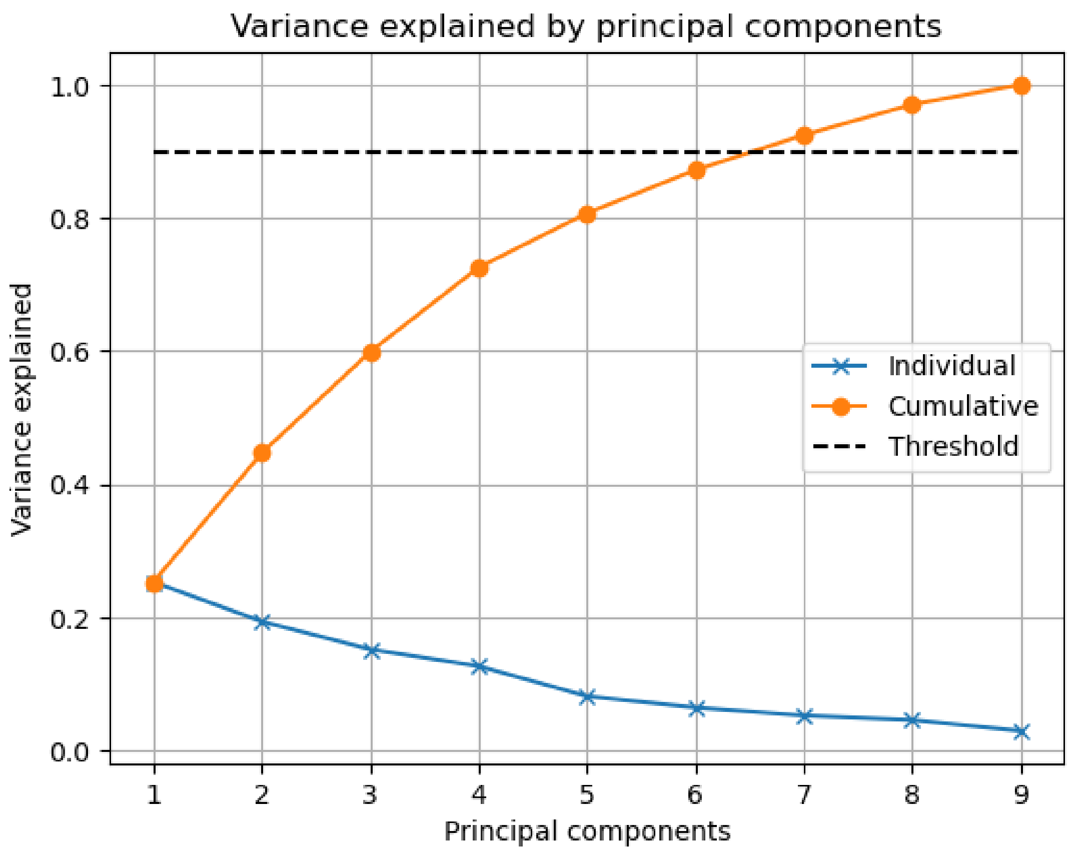
[10]:

```
# All the values above show how many standard deviations away from the mean the␣
↪data is. Notice how there's one value that is 4.94
# standard deviations away from the mean. This is an outlier that we should␣
↪take into account when we do PCA.

# Another thing to note is that by standardizing the data, we don't need to do
# Y = df_standardized - np.mean(df_standardized, axis=0) anymore because␣
↪standardizing already transforms the attribute so that the mean is 0.

# Let's finally perform PCA via Single Value Decomposition svd:
U, S, Vt = svd(df_standardized, full_matrices=False)

# Compute variance explained by principal components
# For example: if S is [3, 2, 1] then S * S is [9, 4, 1] and (S * S).sum() is 9␣
↪+ 4 + 1 = 14.
# rho is [9/14, 4/14, 1/14] = [0.64, 0.29, 0.07] where we get the variance␣
↪explained by each principal component!
rho = (S * S) / (S * S).sum()

# Let's see how much variance is explained by the first 3 principal components:
print("The first principal component explains", rho[0] * 100, "% of the␣
↪variance.")
print("The second principal component explains", rho[1] * 100, "% of the␣
↪variance.")
print("The third principal component explains", rho[2] * 100, "% of the␣
↪variance.")
```

```
The first principal component explains 25.3115822312159 % of the variance.
The second principal component explains 19.3735852803376 % of the variance.
The third principal component explains 15.200345453427424 % of the variance.
```

[11]:
```
# We can see that the first three principle components are ordered from highest␣
↪to lowest variance explained.
# Usually we want to define a threshold, which in our case will be 90% of the␣
↪variance explained.

threshold = 0.9

# Let's try to visualize this better via plots:
plt.figure()
# First plot shows the variance explained by each principal component
plt.plot(range(1, len(rho) + 1), rho, "x-")
# Second will show the cumulative variance explained by the principal components
plt.plot(range(1, len(rho) + 1), np.cumsum(rho), "o-")
# Let's now add the threshold
plt.plot([1, len(rho)], [threshold, threshold], "k--")
# Add a title
```

```
plt.title("Variance explained by principal components")
# Label the axis
plt.xlabel("Principal components")
plt.ylabel("Variance explained")
# Add legend and grid as visual aid
plt.legend(["Individual", "Cumulative", "Threshold"])
plt.grid()
```

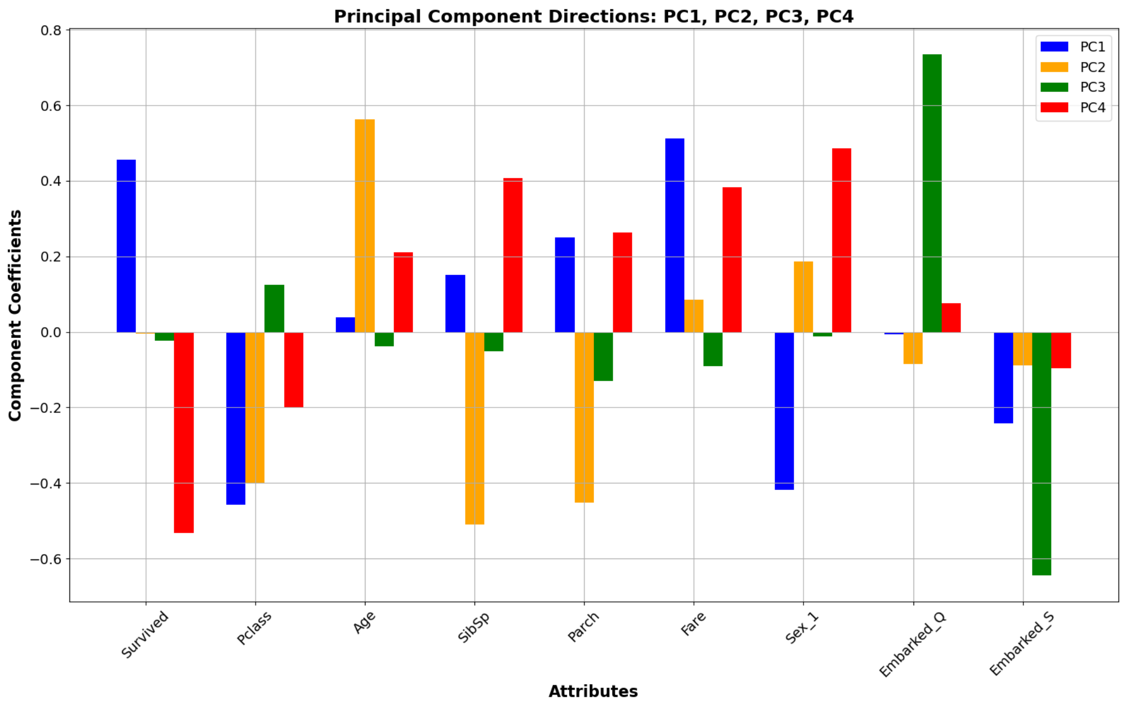


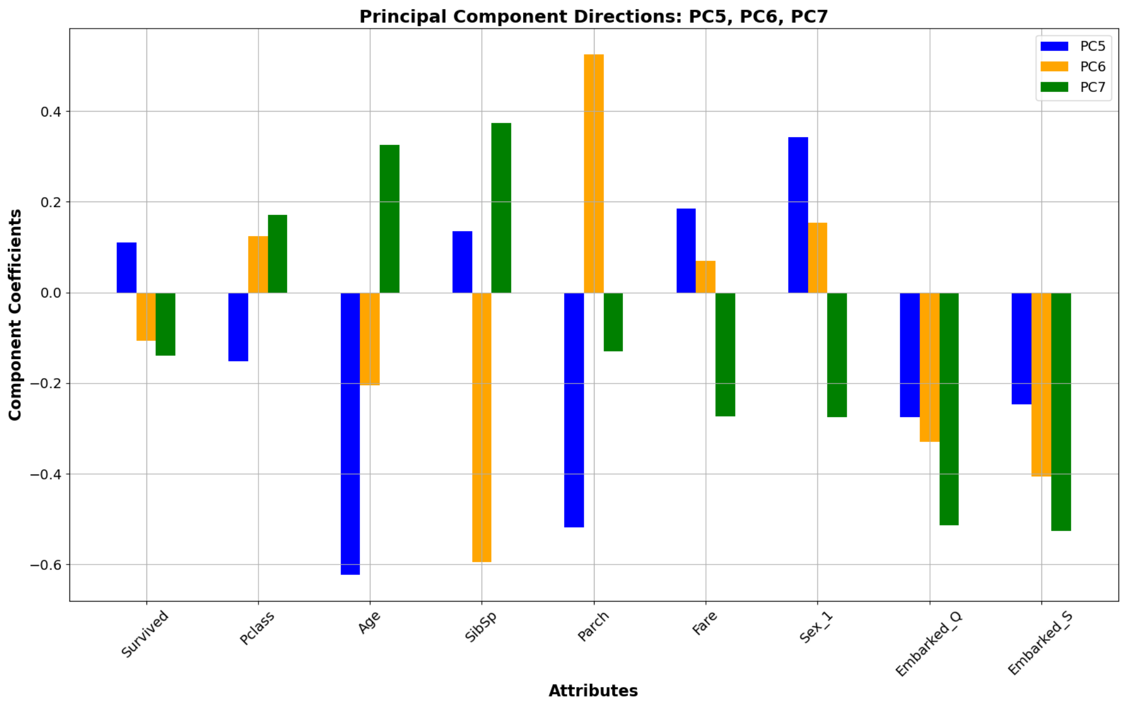
[12]:
```
# How can we interpret this result? It's obvious from the graph above that we␣
↪need at least 7 principal components to explain at least 90% of the variance,
# which is 2 less dimensions than the cleaned dataset and 6 less dimensions␣
↪than the original dataset. This is a reduction in dimensionality 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!
```

```python
# Describe the principle directions of the considered PCA components (either␣
↪find a way to plot them or interpret them
# in terms of the features)

# Remeber earlier, we performed the svd U, S, Vt = svd(df_standardized,␣
↪full_matrices=False)
# Now we extract the Vt, which is a matrix that contains the vectors defining␣
↪the principle directions.
principal_directions = Vt.T

# Begin by printing how to principle directions look like
print("Principal directions 1 - 7:")

# Convert the principal directions to a DataFrame for better formatting
principal_directions_df = pd.DataFrame(principal_directions)

# Transpose the DataFrame to show each principal direction as a column
principal_directions_df = principal_directions_df.T

# Print the DataFrame
print(principal_directions_df.iloc[:, :7])

# Set the components to plot in each subplot
pcs_top = [0, 1, 2, 3]  # First four principal components
pcs_bottom = [4, 5, 6]  # Fifth, sixth, and seventh principal components

# Start by configuring the plot
legendStrs_top = ["PC" + str(e + 1) for e in pcs_top]
legendStrs_bottom = ["PC" + str(e + 1) for e in pcs_bottom]
colors_top = ["b", "orange", "g", "r"]
colors_bottom = ["b", "orange", "g"]
bar_width = 0.175
# Below simply creates an array from 0 to the number of principal directions
r = np.arange(principal_directions.shape[0])

# Create the figure with two subplots
plt.figure(figsize=(16, 20))

# First Subplot: PC1, PC2, PC3, PC4
plt.subplot(2, 1, 1)
for i, pc in enumerate(pcs_top):
    # r + i * bar_width calculates the positions of the bars on the x_axis and␣
↪we plot on the y axis the principal directions of each vector
    plt.bar(r + i * bar_width, principal_directions[:, pc], width=bar_width,␣
↪color=colors_top[i], label=legendStrs_top[i])

# Config the plots more with labels and titles
```

```
plt.xticks(r + bar_width, df_clean.columns, rotation=45, fontsize=14)
plt.yticks(fontsize=14)
plt.xlabel("Attributes", fontsize=16, fontweight='bold')
plt.ylabel("Component Coefficients", fontsize=16, fontweight='bold')
plt.title("Principal Component Directions: PC1, PC2, PC3, PC4", fontsize=18,␣
↪fontweight='bold')
plt.legend(fontsize=14)
plt.grid(True)

# Second Subplot: PC5, PC6, PC7
plt.subplot(2, 1, 2)
for i, pc in enumerate(pcs_bottom):
    plt.bar(r + i * bar_width, principal_directions[:, pc], width=bar_width,␣
↪color=colors_bottom[i], label=legendStrs_bottom[i])

plt.xticks(r + bar_width, df_clean.columns, rotation=45, fontsize=14)
plt.yticks(fontsize=14)
plt.xlabel("Attributes", fontsize=16, fontweight='bold')
plt.ylabel("Component Coefficients", fontsize=16, fontweight='bold')
plt.title("Principal Component Directions: PC5, PC6, PC7", fontsize=18,␣
↪fontweight='bold')
plt.legend(fontsize=14)
plt.grid(True)

plt.tight_layout()
plt.show()
```

```
Principal directions 1 - 7:
          0         1         2         3         4         5         6
0  0.455455 -0.457589  0.038203  0.150868  0.250352  0.512659 -0.418343
1 -0.005131 -0.400613  0.563250 -0.510466 -0.452356  0.084931  0.186335
2 -0.023341  0.124021 -0.038500 -0.051926 -0.130117 -0.091428 -0.012132
3 -0.532181 -0.200543  0.210577  0.407874  0.262400  0.382339  0.485859
4  0.110926 -0.152440 -0.623144  0.135432 -0.518809  0.185240  0.343188
5 -0.106948  0.124865 -0.204303 -0.594533  0.525387  0.070379  0.154082
6 -0.138925  0.170370  0.325338  0.373156 -0.130272 -0.273310 -0.274635
7  0.611287 -0.094247  0.181148  0.191157  0.257248 -0.405730  0.557452
8 -0.303412 -0.705539 -0.258375 -0.002969  0.136282 -0.544833 -0.162377
```





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 captures. 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 strenght of each contribution.

Remember that the PCs seem to capture the variances of these patterns the most in each principle component.

PC1: Attributes like "Survived", "Sex_1", "Fare" and "pClass" have relatively high positive coefficients, suggesting they are key factors driving the variance captured by PC1. These attributes seem to explain variances concerning socioeconomic status, with survivability being driven by higher class, higher fare prices and not being male.

PC2: 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.

PC3: Seems to capture where a person is embarked and it seems like the underlying pattern is that if a person embarked from Q, they certaintly didn't from S.

PC4: 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.

PC5: 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.

PC6: 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.

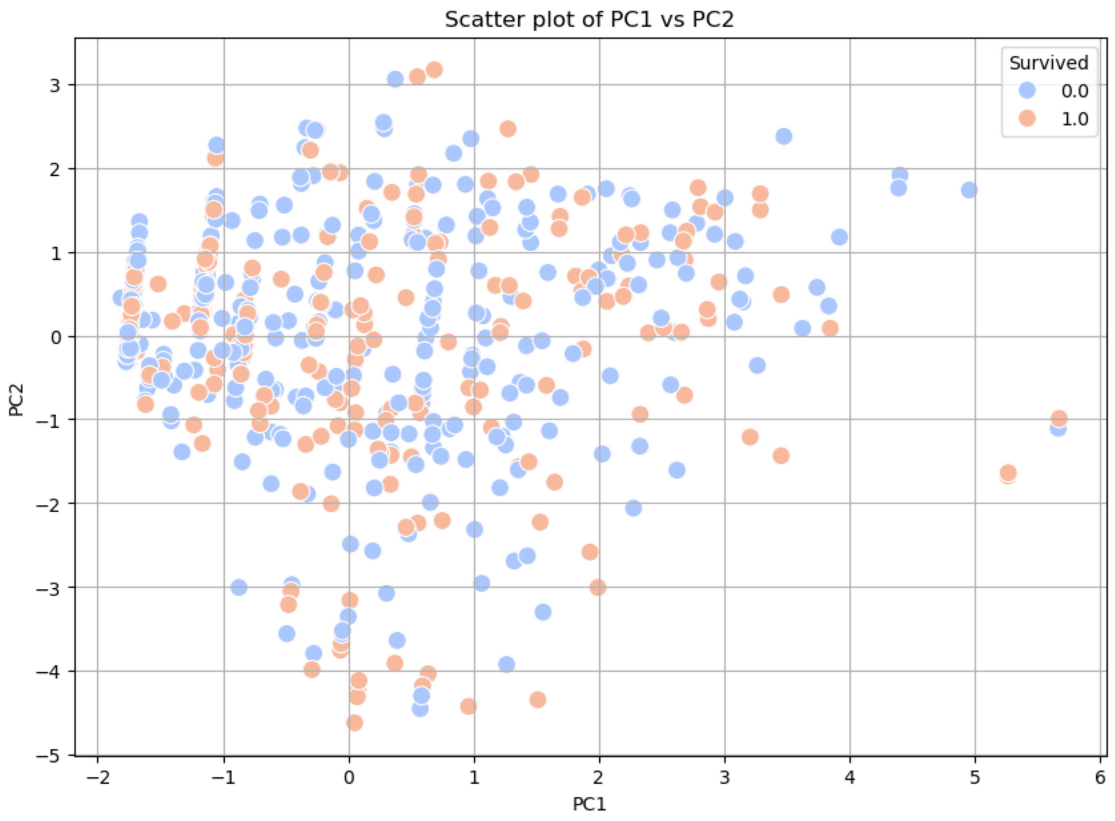
PC7: Difficult to interpret, maybe a pattern of how a person embarked?

[13]:
```
# Now we need to describe the data projected onto the considered principle␣
 ↪component

# Project the data onto the principal components We could also do␣
 ↪projected_data = df_standardized @ principal_directions instead of scores =␣
 ↪U * S
# The scores (projected data) are already given by U * S
scores = U * S  # This scales U by the singular values, giving the principal␣
 ↪component scores

# Convert the projected data into a DataFrame for easier interpretation
projected_df = pd.DataFrame(scores, columns=[f'PC{i+1}' for i in range(scores.
 ↪shape[1])])

# Display the first few rows of the projected data
print(projected_df.head())

# Let's visualize the projection onto the first two principal components
```

```python
plt.figure(figsize=(10, 8))

# Scatter plot of the first two principal components
plt.figure(figsize=(10, 7))
sns.scatterplot(x=projected_df['PC1'], y=projected_df['PC2'],␣
 ↪hue=df_clean['Survived'], palette='coolwarm', s=100)
plt.title('Scatter plot of PC1 vs PC2')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.grid(True)
plt.show()
```

```
        PC1       PC2       PC3       PC4       PC5       PC6       PC7  \
0 -1.601119 -0.629406 -0.254474  0.299565  0.506450 -0.340026  0.136214
1  2.686560  0.897990  0.805582 -0.281035  0.573398 -0.315784  1.238534
2  0.053959 -0.322481 -0.233166 -2.169556 -0.294913 -0.295310  0.108448
3  1.876640  0.528845 -0.706440 -0.717566  0.024512 -1.285645  0.014605
4 -1.719307  0.424923 -0.234925  0.057462 -0.193853  0.116547  0.021988

        PC8       PC9
0 -0.092667 -0.140593
1  0.010376  0.203326
2 -0.166234 -0.498547
3 -0.049888  0.455963
4 -0.143157 -0.379075
```

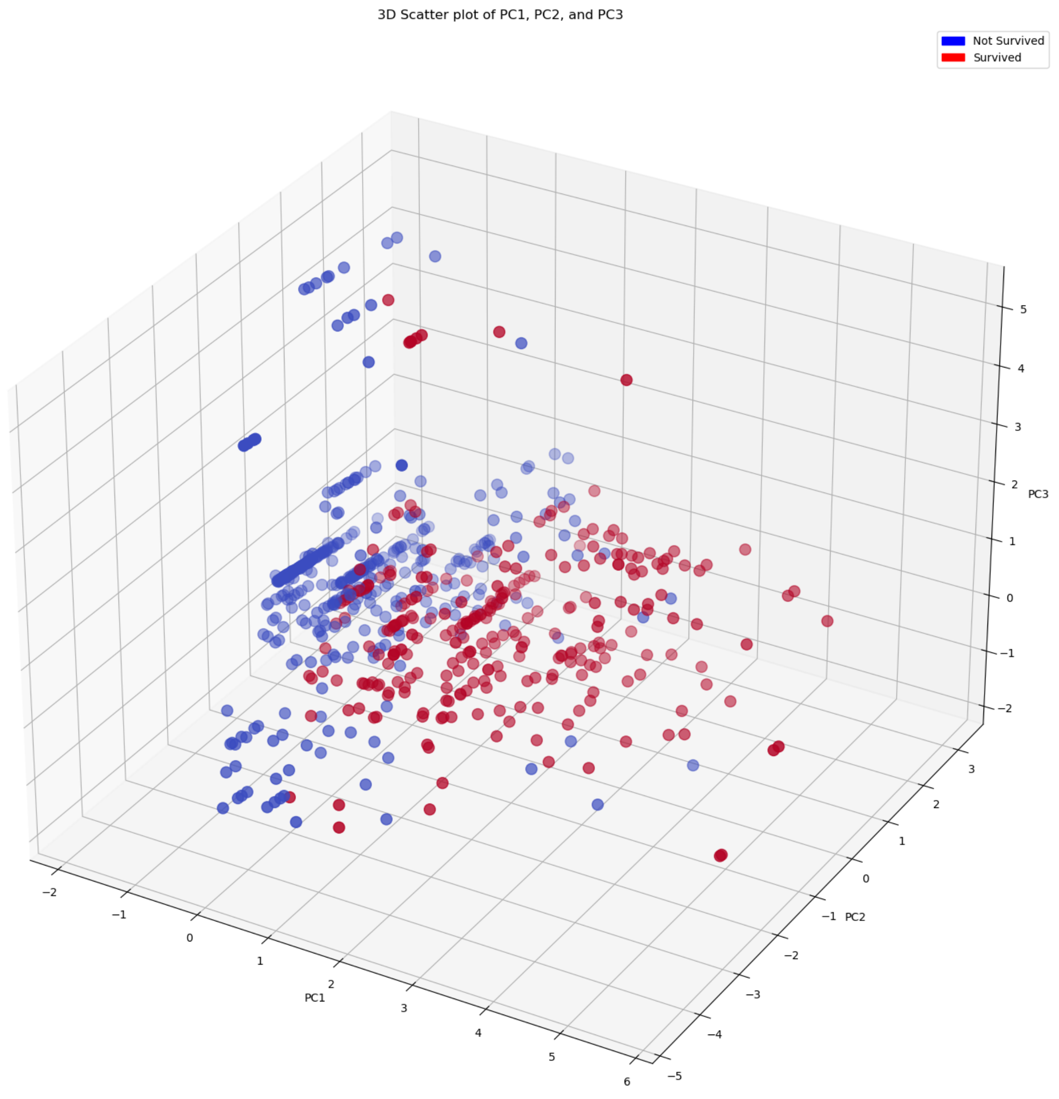
```
<Figure size 1000x800 with 0 Axes>
```



The 2 dimensional graph doesn't separate the two binary outcomes into clear clusters, this suggests that the first two principle components doesn't capture the variance too well, which makes sense given that the first two principle components only explain a little under half of the variance whilst the optimal solution is to get the principle components to explain about 90% ideally. This proved to be only doable with 7 principle components. Maybe by projecting to 3 dimensions would ease up on this.

[14]:
```python
# Create a new figure with interactive 3D projection
fig = plt.figure(figsize=(18, 18))
ax = fig.add_subplot(111, projection='3d')

# 3D Scatter plot of the first three principal components
ax.scatter(projected_df['PC1'], projected_df['PC2'], projected_df['PC3'],␣
 ↪c=df_clean['Survived'], cmap='coolwarm', s=100)

# Set axis labels
ax.set_xlabel('PC1')
ax.set_ylabel('PC2')
ax.set_zlabel('PC3')

# Set the title of the plot
```

```
plt.title('3D Scatter plot of PC1, PC2, and PC3')

# Create custom legend
survived_patch = mpatches.Patch(color='red', label='Survived')
not_survived_patch = mpatches.Patch(color='blue', label='Not Survived')

# Add legend to the plot (top right corner)
plt.legend(handles=[not_survived_patch, survived_patch], loc='upper right')

# Show the plot
plt.show()
```



Remember:

PC1: Attributes like “Survived”, “Sex_1”, “Fare” and “pClass” have relatively high positive coefficients, suggesting they are key factors driving the variance captured by PC1. These attributes seem to explain variances concerning socioeconomic status, with survivability being driven by higher class, higher fare prices and not being male.

PC2: 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.

PC3: Seems to capture where a person is embarked and it seems like the underlying pattern is that if a person embarked from Q, they certaintly didn’t from S.

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 is that we do see some semblance of two clusters being formed. I had some configuration issues, which made it very challenging to create a 3d plot that was movable so that we could see the 3d plot from different angles. But based on this static 3d plot, it looks like we can see a cluster being formed the lower the PC1 with relatively high values of PC2 with PC3 from its negative values to around 0 being the cluster for an individual not surviving the titanic.

With negative PC1 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 PC2 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, PC3 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, but without the ability to rotate the 3d plot, it is difficult to come to a reasonable conclusion of the higher the scale of PC3 the greater the relationship with embarking at Queenstown.

Task 4

TBD

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 cannonical 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

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

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

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

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 1: Summary of Titanic Dataset Attributes by Type and Measurement Scale