Titanic dataset analysis

Bogachev Aleksei

September 24, 2022

Abstract

Data analysis and models for the legendary machine learning competition on Kaggle.

Contents

Chapter 1

Introduction

Perhaps, the sinking of the RMS Titanic is the most infamous shipwreck in history. According to the Wikipedia, the RMS Titanic was the largest ocean liner in service at the time. It had advanced safety features, such as watertight compartments and remotely activated watertight doors. The ship was widely considered "unsinkable". However, the Titanic sank in the early morning of 15 April 1912 in the North Atalntic Ocean during her maiden voyage from Southampton to New York City. There were an estimated 2224 people on board when the ship collided with an iceberg [5],[4].

In accordance with existing practice, Titanic's lifeboat system was designed to ferry passengers to nearby rescue vessels, not to hold everyone on board simultaneously; therefore, with the ship sinking rapidly (the ship had sank in 2 hours and 40 minutes) and help still hours away, there was no safe refuge for many of the passengers and crew with only 20 lifeboats. Poor management of the evacuation meant many boats were launched before they were completely full [4].

The shipwreck resulted in the deaths of more than 1500 people, makng it one of the deadliest in history [4].

Without a doubt, there was an element of luck involved in surviving, but, possibly, some groups of people were more likely to survive than others. The Titanic ML competition on Kaggle offers participants to predict which of the passengers survived the shipwreck using passenger data[7].

In this report I'm going to describe my solution of the Titanic ML competition's task. My workflow will be based mostly on the "Machine Learning project checklist" from the book [2]. I really appreciate this book and highly recommend reading it to anyone starting to learn about machine learning.

Dozens of articles dedicated to this competition and hundreds of solutions of this task are available in the Internet. Therefore, I won't cite to all materials seen, but I'll try to give several useful refrences. In exploratory analysis, I relied a lot on the tutorial [9] and borrowed several ideas from it.

Let's get started.

Chapter 2

Task Description

In this section the task is described according to "Machine Learning project checklist" [2].

2.1 Goal

To predict if a passanger survived the sinking of the Titanic or not.

2.2 Crrent Solutions

There are dosens of solutions available on the discussion forum and on the Internet.

2.3 Frame the Problem

- Supervised learning
- Classification
- Binary classification (survived of not)
- Batch learning (no continuous flow of data and the dataset is small)

2.4 Performance metrics

This competition evaluates the **percentage of correctly predicted passengers** (accuracy).

There are also several useful metrics for evaluating the performance of a classification system:

- precision,

- recall,
- F_1 score,
- precision/recall curve,
- ROC curve,
- ROC AUC score.

2.5 Target performance

The leaderboard of this competition contains almost 14000 entries. It's available in the form of the csv-file. An excerpt from the leaderboard is presented in the table 2.1.

Table 2.1: Excerpt from the leaderboard

TeamId	**TeamName**	**SubmissionDate**	**Score**
6987444	no name	2022-08-23 18:16:28	1.0
720238	rosh	2022-06-26 10:58:42	1.0
8814675	nikolai otvetchikov #2	2022-06-26 13:59:39	1.0
8821160	Vibhav Rathkanthiwar	2022-06-26 15:28:12	1.0
6590016	Osman Altuntas	2022-07-24 15:40:15	1.0

The descriptive statistics is shown in the table 2.2.

Table 2.2: Descriptive statistics of scores

Statistics	**Value**
count	13915.000000
mean	0.760751
std	0.075145
min	0.000000
25%	0.765550
50%	0.775110
75%	0.777510
max	1.000000

The median score is about 0.775, but less than 3% of the solutions have a score above 0.8. Thus, **an accuracy score equal to or greater than 0.8 would be a very good result**. Figure 2.1 shows ECDF of the scores in the leaderboard. In this figure, the red lines mark the score 0.8 and the corresponding proportion of solutions.

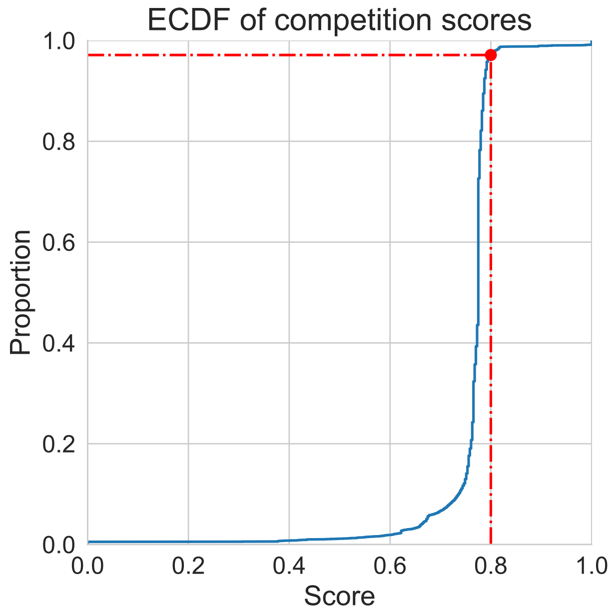

Figure 2.1: Leaderboard Scores ECDF

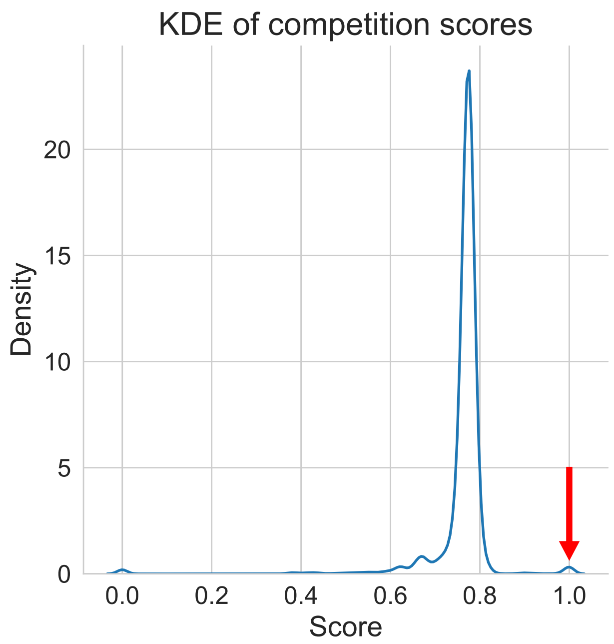

Figure 2.2: Leaderboard Scores KDE

There are several solutions with a score equal to 1.0. These solutions are marked with the red arrow in the figure 2.2. Have authors reached perfection?

I guess, this solutions appears, because there is an exact solution on GitHub. Possibly it is the data extracted from Encyclopedia Titanica[8] or from OpenML [6]. Some authors in their notebooks honestly warn other users about the existence of such a possibility, for example, this one [3].

2.6 Data Dictionary

1. **PassengerId** – Passenger ID.
2. **Survived** – Survival:
   - 0 = No,
   - 1 = Yes.
3. **Pclass** – Ticket class:
   - 1 = 1st,
   - 2 = 2nd,
   - 3 = 3rd.
4. **Name** – Passanger's name, for example, "Braund, Mr. Owen Harris".
5. **Sex** – Gender:
   - male,
   - female.
6. **Age** – Age in years, for example 38.0.
7. **SibSp** – Number of siblings or spouses aboard the Titanic.
8. **Parch** – Number of parents or children aboard the Titanic.
9. **Ticket** – Ticket number, for example, A/5 21171.
10. **Fare** – Passenger fare, for ecample, 71.2833.
11. **Cabin** – Cabin number, for example, C85.
12. **Embarked** – Port of Embarkation:
    - C = Cherbourg,
    - Q = Queenstown,
    - S = Southampton.

2.6.1 Features

PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

2.6.2 Target

Survived

2.6.3 Variable Notes

- **Pclass**: socio-economic status
  - 1st = Upper
  - 2nd = Middle
  - 3rd = Lower
- **Age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- **SibSp** number of sibling/spouses aboard the Titanic
  - sibling = brother, sister, stepbrother, stepsister
  - spouse = husband, wife (mistresses and fiancés were ignored)
- **Parch** number of parents (mother, father)/children (daughter, son, stepdauter, stepson) aboard the Titanic. Some children travelled only with a nanny, therefore **Parch**=0 for them.

2.7 File Paths

- **training set**: ../datasets/train.csv
- **test set**: ../datasets/test.csv
- **example of a submission file**: ../datasets/gender_submission.csv

2.8 Assumptions

Women were more likely to survive than men.

Chapter 3

Preliminary Analysis

3.1 Shape of the dataset

The dataset contains:

- 891 rows,
- 12 columns.

3.2 First rows of the dataset

An excerpt from the dataset is presented in the table 3.1.

Table 3.1: Excerpt from the dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

The **"PassengerId"** feature is the ID of the passanger. It won't help in the analysis and will be dropped. Also, there are several missing values, and some values are categorical, for example, **"Pclass"** and **"Sex"**.

3.3 Data types and missing values

Table 3.2 contains types of the data in each column and numbers of non-null values. Table 3.3 contains numbers of missing values in each column.

Table 3.2: Data types and non-null counts

#	Column	Non-Null Count	Dtype
0	PassengerId	891 non-null	int64
1	Survived	891 non-null	int64
2	Pclass	891 non-null	int64
3	Name	891 non-null	object
4	Sex	891 non-null	object
5	Age	714 non-null	float64
6	SibSp	891 non-null	int64
7	Parch	891 non-null	int64
8	Ticket	891 non-null	object
9	Fare	891 non-null	float64
10	Cabin	204 non-null	object
11	Embarked	889 non-null	object

Table 3.3: Number of missing values in each column

#	Column	Number of missing values
0	PassengerId	0
1	Survived	0
2	Pclass	0
3	Name	0
4	Sex	0
5	**Age**	**177**
6	SibSp	0
7	Parch	0
8	Ticket	0
9	Fare	0
10	**Cabin**	**687**
11	**Embarked**	**2**

3.4 Number of unique values

Table 3.4 contains numbers of unique values in each column.

There are high-cardinality features with *object* dtype:

- Name
- Ticket
- Cabin
- PassengerId

Table 3.4: Number of unique values in each column

Column	Number of unique values	Column	Number of unique values
Name	891	Survived	2
Sex	2	Pclass	3
Ticket	681	Age	88
Cabin	147	SibSp	7
Embarked	3	Parch	7
PassengerId	891	Fare	248

This features, possibly, will need special preprocessing. Earlier, I noticed that the **"PassengerId"** feature is the ID of the passanger. It won't help in the analysis and will be dropped.

Features **"Age"** and **"Fare"** are continuous.

3.5 Summary statistics

In this section, summary statistics for numerical and non-numerical attributes are presented in tables 3.5 and 3.6 respectively.

Table 3.5: Summary statistics for numerical attributes

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Table 3.6: Summary statistics for non-numerical attributes

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Braund, Mr. Owen Harris	male	347082	B96 B98	S
freq	1	577	7	4	644

3.6 Proportion of target values

Let's check if one class of target values is more frequent than another.

Figure 3.1 illustrates these numbers. There are 342 (38.38%) survived passengers and 549 (61.62%) drowned passengers in the dataset, so the dataset is a bit skewed. Therefore, the accuracy may not be sufficient to assess the performance of classifiers.

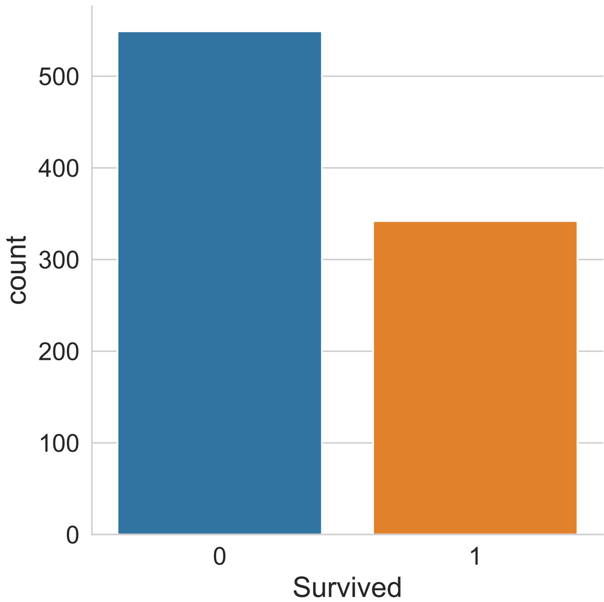

Figure 3.1: Number of survived and drowned passangers in whole dataset

Chapter 4

Sample a Test Set

The test set will be used to evaluate performance of a very final model and forecast the score in the competitions leaderboard. It may seems like it's to early to create a test set, but I'll do it to prevent data snooping.

I'm going to do stratified sampling with scikit-learn's `StraifiedShuffleSplit` to maintain equal ratio of men and women in the train set and the test set. Women seem to have had a better chance of surviving due to the "women and children first" protocol for loading lifeboats.

Next, let's check the proportion of women among all survivors (figure 4.1), and check the proportion of survived women in each **"Pclass"** (figure 4.2).

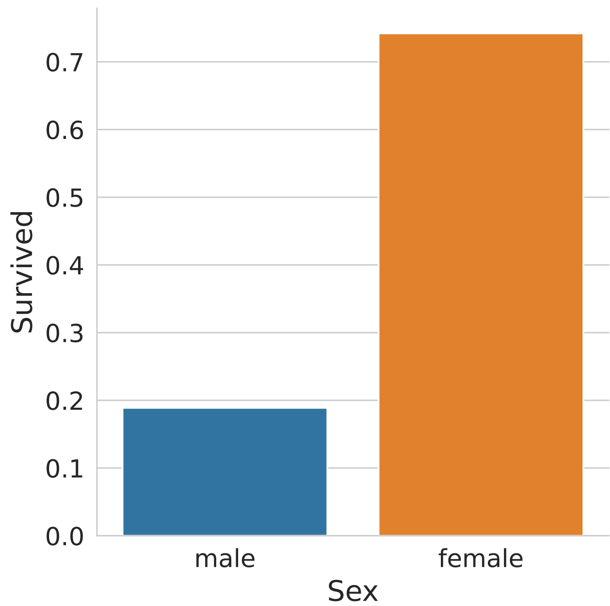

Figure 4.1: Proportions of survived men and women

Figures 4.1 and 4.2 show that in the entire dataset and in each **"Pclass"**, there are more female survivors than males. Thus it is reasonable to do stratification based on the passenger's gender. I will use 80% of the data for training and hold out 20% for testing, refer to the Jupyter Notebook for details.

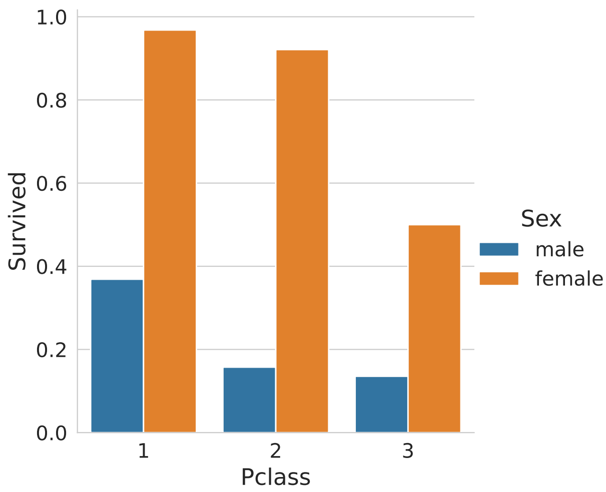

Figure 4.2: Proporotion of survived women and men in each **"Pclass"**

Chapter 5

Exploratory Analysis

In this chapter, I will copy the training set to an exploratory set and work with it, because the exploratory analysis may require some data transformations, however I want to keep the training set untouched.

I will explore each attribute and its characteristics. Here is a list of dataset attributes:

- PassengerId
- Survived
- Pclass (page 16)
- Name (page 18)
- Sex (page 22)
- Age (page 24)
- SibSp (page 26)
- Parch (page 28)
- Ticket (page 30)
- Fare (page 30)
- Cabin (page 34)
- Embarked (page 34)

I won't consider **"PassengerId"** and **"Survived"** attribures. **"PassengerId"** attribute contains 891 passenger's IDs (see section First rows of the dataset and table 3.4), which won't help in building a model. **"Survived"** attribute is a target. A passenger survived if their **"Survived"** attribute is 1, else the passenger drowned (**"Survived"** attribute is 0). Figure 3.1 shows numbers of survived and drowned passengers in whole dataset.

5.1 Pclass

5.1.1 Description

The **Pclass** attribute contains information about socio-economic status of the passenger:

- 1st = Upper
- 2nd = Middle
- 3rd = Lower

The table 5.1 contains excerpt from attribute's column. The table 5.2 contains the attribute characteristics.

Table 5.1: Excerpt from the **Pclass** column

Index	788	347	629	734	106
Value	3	3	3	2	3

Table 5.2: Characteristics of the **Pclass** attribute

Characteristic	**Value**
Type	category
Number of values	712
Number of non-null values	712
Number of unique values	3
Most frequent value	3

The table 5.3 contains the number of occurrences of each value.

Table 5.3: Number of occurrences of each value in the **Pclass** attribute

Value	**Number**
3	402
1	172
2	138

5.1.2 Analysis

Let's estimate the number of passengers of each class in the exploratory set. The figure 5.1 illustrates this estimation.

The figure 5.2 shows the proportions of survived passengers for each **Pclass** in the exploratory set.

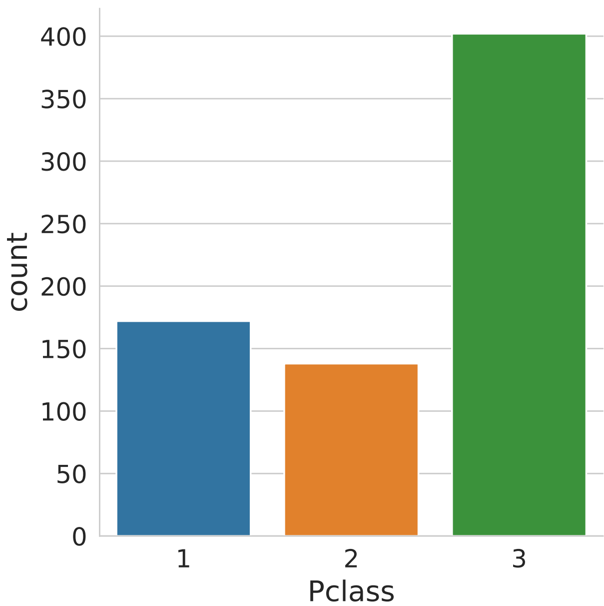

Figure 5.1: Number of passengers in each **Pclass**

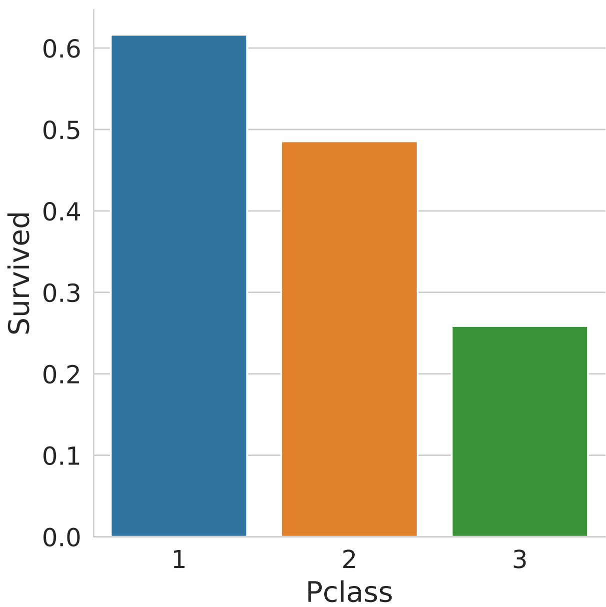

Figure 5.2: Proportion of survived passengers in each **Pclass**

It looks like there were more passengers of the lower socio-economic class (**Pclass**=3), but they had less chance of surviving. **Pclass** is a class of the ticket, so it contains information about the location of the passenger's cabin. It is known that the cabins of passengers with low socio-economic status were located on lower decks, that is, further from the lifeboats, this explains why there are fewer survivors among them [5].

Finally, let's check the proportion of women among the survivors of each **Pclass**. Figure 5.3 show these proportions.

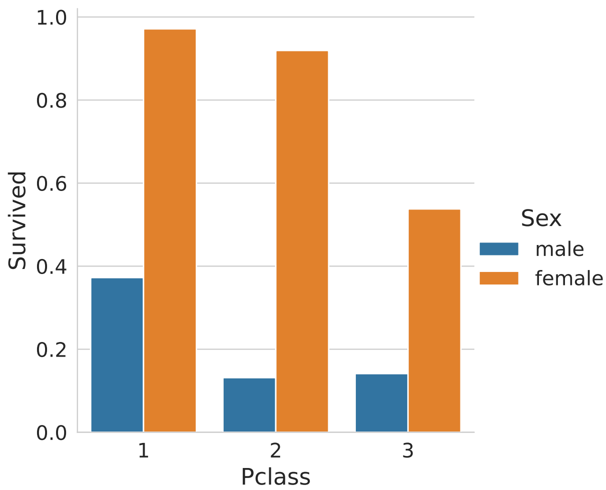

Figure 5.3: Proportion of survived passengers of eache gender in each **Pclass**

More women than men survived in each **Pclass**.

5.2 Name

5.2.1 Description

The **Name** attribute contains names of passengers, and it contains 712 unique values in the exploratory dataset, so each value occurs only once. The table 5.4 shows the exerpt from this column. The table 5.5 contains the attribute characteristics.

Table 5.4: Excerpt from the **Name** column

Index	**Name**
788	Dean, Master. Bertram Vere
347	Davison, Mrs. Thomas Henry (Mary E Finck)
629	O'Connell, Mr. Patrick D
734	Troupiansky, Mr. Moses Aaron
106	Salkjelsvik, Miss. Anna Kristine

5.2.2 Analysis

There seems to be a useful pattern. Each name contains a title, such as "Mrs." or "Master.". The title may contain information about gender, socio-economic status, age, etc.

Let's create new feature **Title** containing the title from each name and count how many times each title occurs in the exploratory set. The results is shown in the table 5.6 and figure 5.4.

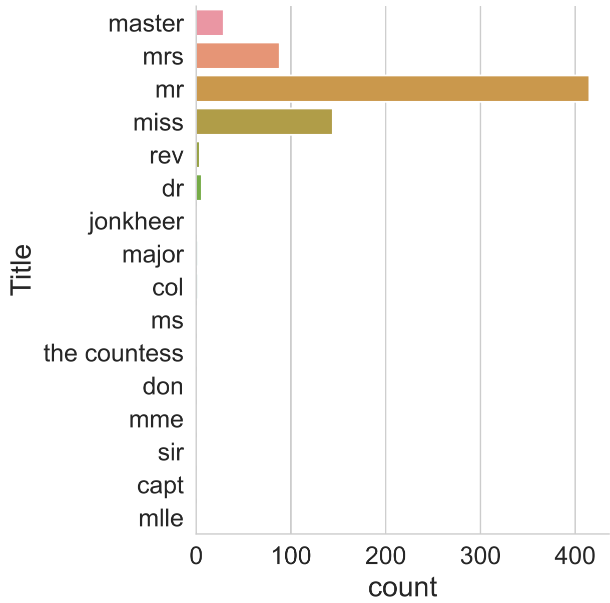

Figure 5.4: Number of occurrences of each title in the **Name** attribute

Several titles occurs extremely rarely, the figure 5.5 shows what **Pclass** these titles belong to.

People with titles *'rev'*, *'lady'*, *'dr'*, *'jonkheer'*, *'major'*, *'col'*, *'ms'*, *'the countess'*, *'don'*, *'mme'*, *'sir'*, *'capt'*, *'mlle'* belong to the first or second **Pclass**, perhaps they are members of the same social class. Let's assign them a new *'aristocratic'* category. The figure 5.6 shows what **Pclass** the titles with new *'aristocratic'* category belong to.

At last, let's count the proportion of survived passengers for each value in the new **Title** feature and the survival rate for men and women with the title *'aristocratic'*. Figures 5.7 and 5.8 show respective results.

Accordingly to the figure 5.7 people with titles *'mrs'*, *'miss'* and *'master'* were most likely to survive. *'Master'* is an English honorific for boys and young men [1]. This pattern is consistent with "women and children first" protocol. As shown in the figure 5.8, for the women and men with **Title** *'aristocratic'* the survival rate is almost the same as survival rates for men and women in the entire dataset (see 4.1).

Table 5.5: Characteristics of the **Name** attribute

Characteristic	**Value**
Type	object
Number of values	712
Number of non-null values	712
Number of unique values	712

Table 5.6: Number of occurrences of each title in the **Name** attribute

Title	**Number**	**Title**	**Number**
mr	415	jonkheer	1
miss	144	ms	1
mrs	88	the countess	1
master	29	don	1
dr	6	mme	1
rev	4	sir	1
major	2	capt	1
col	2	mlle	1

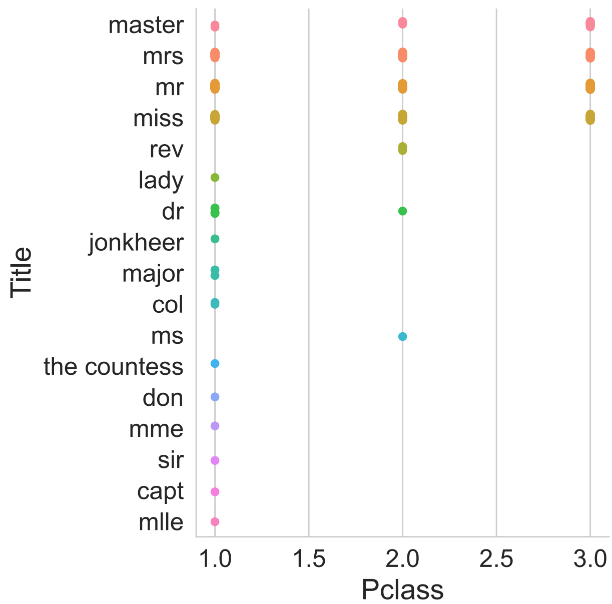

Figure 5.5: **Title**s belonging to **Pclass**es

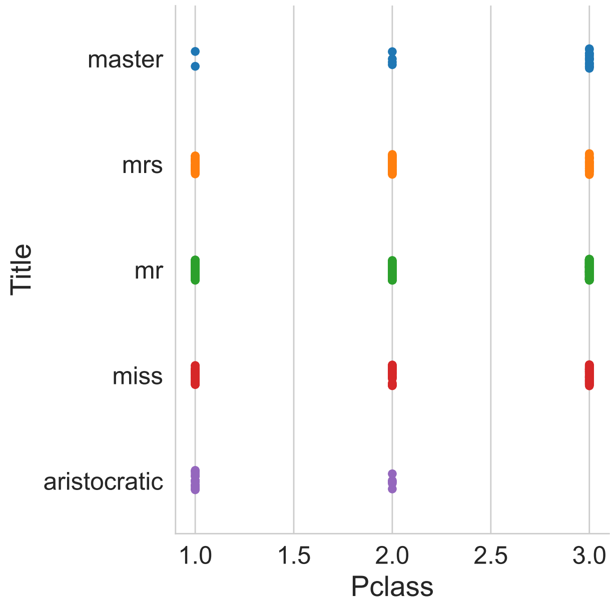

Figure 5.6: **Title**s belonging to **Pclass**es

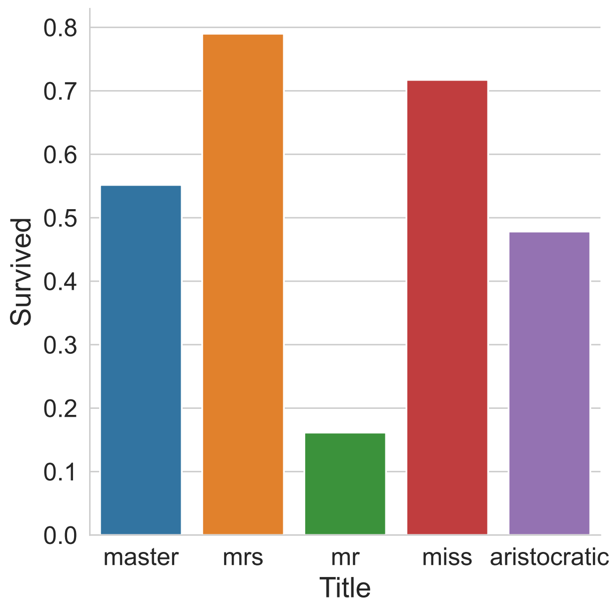

Figure 5.7: proportion of survived passengers for each value in the new **Title** feature

5.3 Sex

5.3.1 Description

The **Sex** attribute contains information about the gender of the passenger. The table 5.7 contains excerpt from this column. The table 5.8 contains the attribute characteristics. The table 5.9 contains the number of occurrences of each value.

Table 5.7: Excerpt from the **Sex** column

Index	788	347	629	734	106
Sex	male	female	male	male	female

Table 5.8: Characteristics of the **Sex** attribute

Characteristic	**Value**
Type	category
Number of values	712
Number of non-null values	712
Number of unique values	2
Most frequent value	male

Table 5.9: Number of occurrences of each value of the **Sex** attribute

Value	**Number**
male	461
female	251

5.3.2 Analysis

We have already studied this attribute in chapter 4 "Sample a Test Set". Figures 4.1 and 4.2 have shown that in the entire dataset and in each **"Pclass"** there are more female survivors than males. So, it's very important attribute, and I will consider other attributes separetely for each gender.

The figure 5.9 shows the number of passengers of each gender on board the Titanic.

The figure 5.10 shows the proportion of survived men and women in the exploratory set.

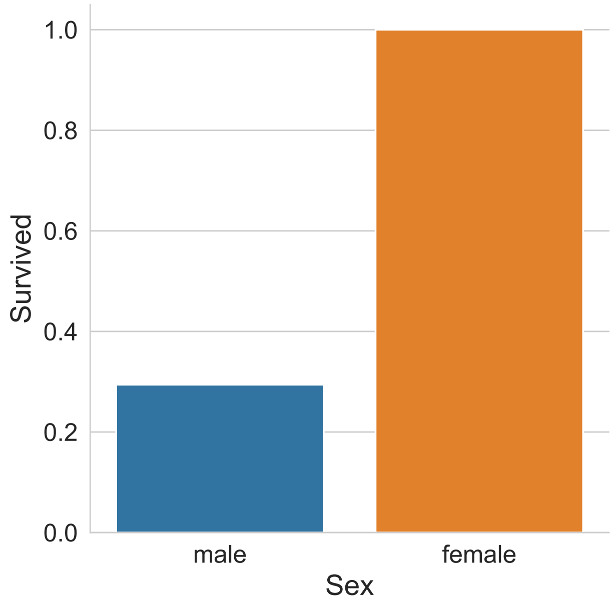

Figure 5.8: Survival rate for men and women with the title *'aristocratic'*

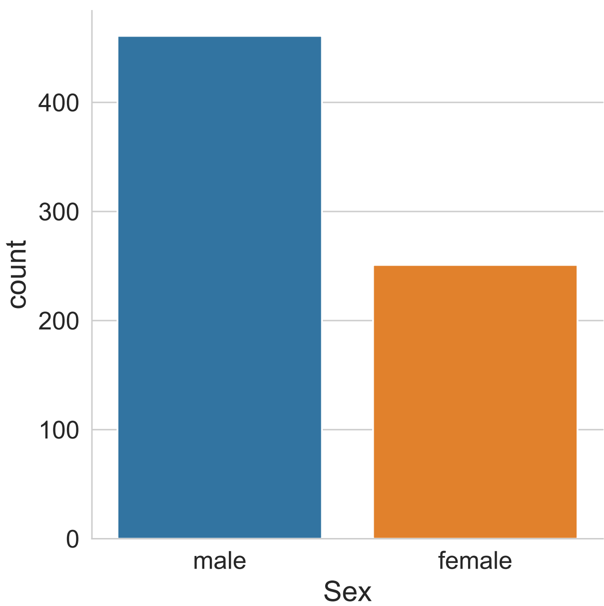

Figure 5.9: Number of men and women on board the Titanic

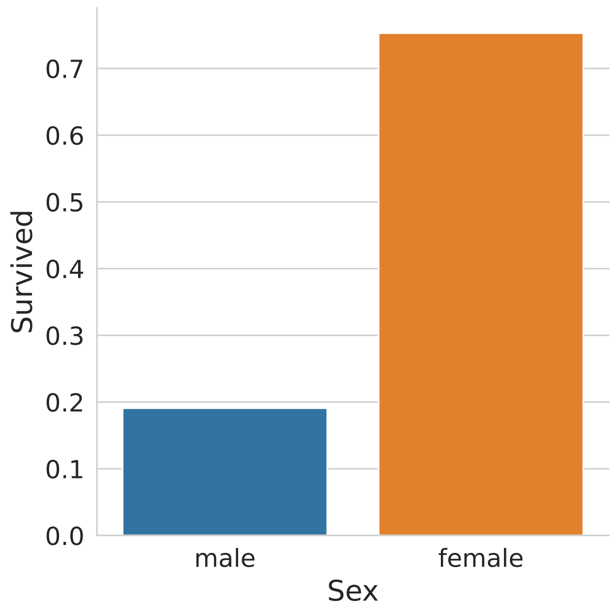

Figure 5.10: Proportion of survived men and women

5.4 Age

5.4.1 Description

The **Age** attribute contains age of a passenger. The table 5.10 contains excerpt from attribute's column. The table 5.11 contains the attribute characteristics.

Table 5.10: Excerpt from the **Age** column

Index	788	347	629	734	106
Attribute	1.0	NaN	NaN	23.0	21.0

Table 5.11: Characteristics of the **Age** attribute

Characteristic	**Value**
Type	numerical (float64), continuous
Number of values	712
Number of non-null values	578
mean	29.781436
std	14.628503
min	0.420000
25%	21.000000
50%	28.000000
75%	38.000000
max	80.000000

5.4.2 Analysis

At first, let's study the distribution of the **Age** attribute. The figures 5.11 and 5.12 show the histogram with KDE and ECDF respectively.

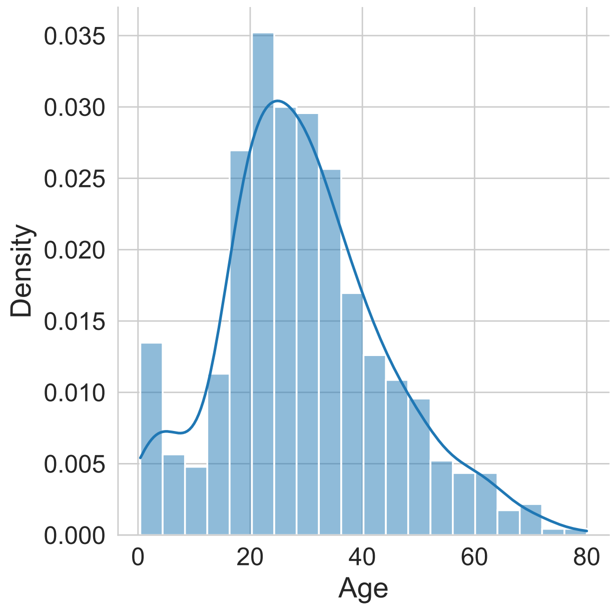

Figure 5.11: Histogram of the **Age** attribute

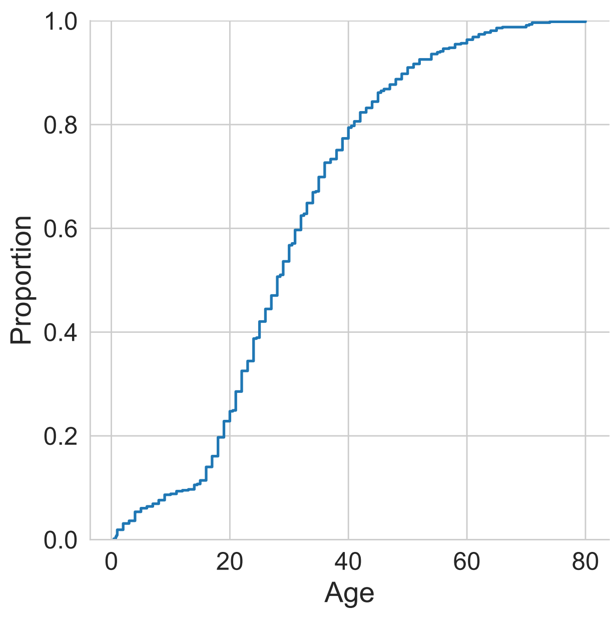

Figure 5.12: ECDF of the **Age** attribute

Obviously, the distribution is skewed and tail-heavy, so we may need to do some transformations, such as standardization. Also, it's continuous feature, so we may need to discretize it.

Next, I'm going to test the assumption that the surviving passengers are younger. The figure 5.13 illustrates distribution of survived and drowned passengers by age. The obvious pattern is not visible here.

Let's check the distribution of passengers by age in each **Pclass**. The corresponding boxplots are shown in the figure 5.14. People from higher socio-economic classes appear to be older.

At last, let's look on the distributions of representatives of each **Title** by age. The corresponding boxplots are shown in the figure 5.15.

5.5 SibSp

5.5.1 Description

The **SibSp** attribute contains the number of siblings (brother, sister, stepbrother, stepsister) or spouses (husband, wife) aboard the Titanic. The table 5.12 contains excerpt from attribute's column. The table 5.13 contains the attribute characteristics.

There is not so many unique values, so let's count number of occurrences of each value. The results is shown in the table 5.14.

Table 5.12: Excerpt from the **SibSp** column

Index	788	347	629	734	106
Value	1	1	0	0	0

Table 5.13: Characteristics of the **SibSp** attribute

Characteristic	**Value**
Type	numerical (int64), discrete
Number of values	712
Number of non-null values	712
mean	0.546348
std	1.110283
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	8.000000

Table 5.14: Number of occurrences of each value of the **SibSp** attribute

Value	0	1	2	4	3	8	5
Number	477	171	25	16	13	5	5

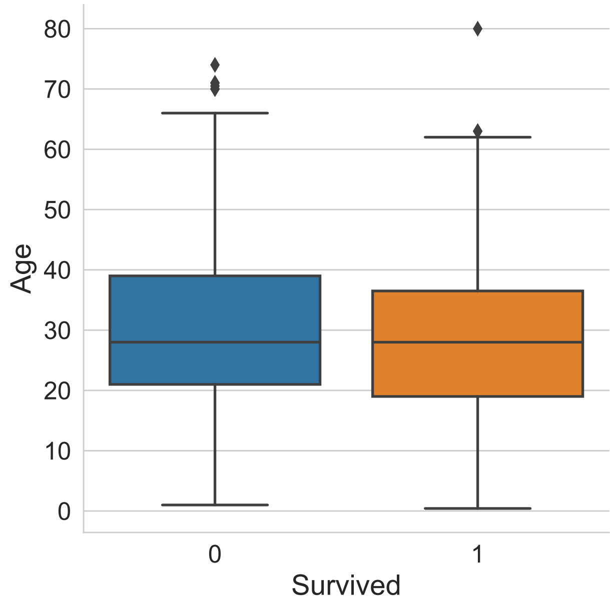

Figure 5.13: Distributions of survived and drowned passengers by age

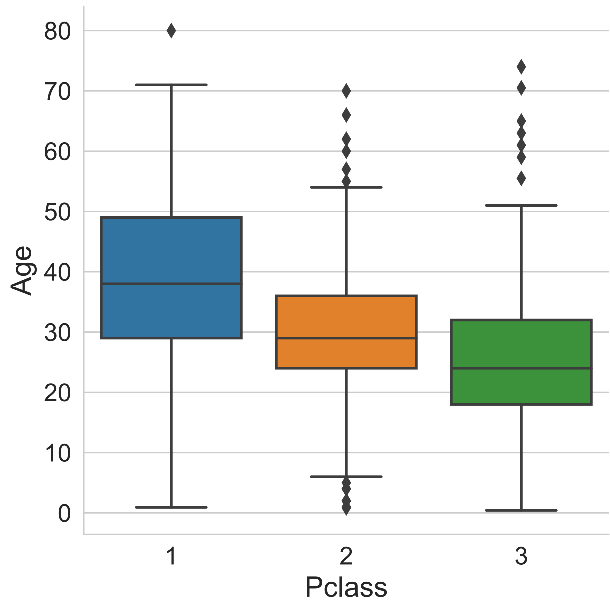

Figure 5.14: Distributions of passengers by age in each **Pclass**

5.5.2 Analysis

Let's first examine the distribution of the attribute's values. The distribution is shown in the figure 5.16. This distribution is skewed and tail-heave. In truth, it's more like an exponential distribution.

The figure 5.17 shows the proportion of survived passengers for each **SibSp** value. It seems that passengers with one sibling or spose had more chances to survive.

5.6 Parch

5.6.1 Description

The **Parch** attribute contains number of parents or children aboard the Titanic. The table 5.15 shows an excerpt from this column. The table 5.16 contains the attribute characteristics. The table 5.17 contains the number of occurences of each value.

Table 5.15: Excerpt from the **Parch** column

Index	788	347	629	734	106
Value	2	0	0	0	0

Table 5.16: Characteristics of the **Parch** attribute

Characteristic	**Value**
Type	numerical (int64), discrete
Number of values	712
Number of non-null values	712
mean	0.373596
std	0.803144
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	6.000000

Table 5.17: Number of occurrences of each value of the **Parch** attribute

Value	0	1	2	5	4	3	6
Number	543	99	58	4	4	3	1

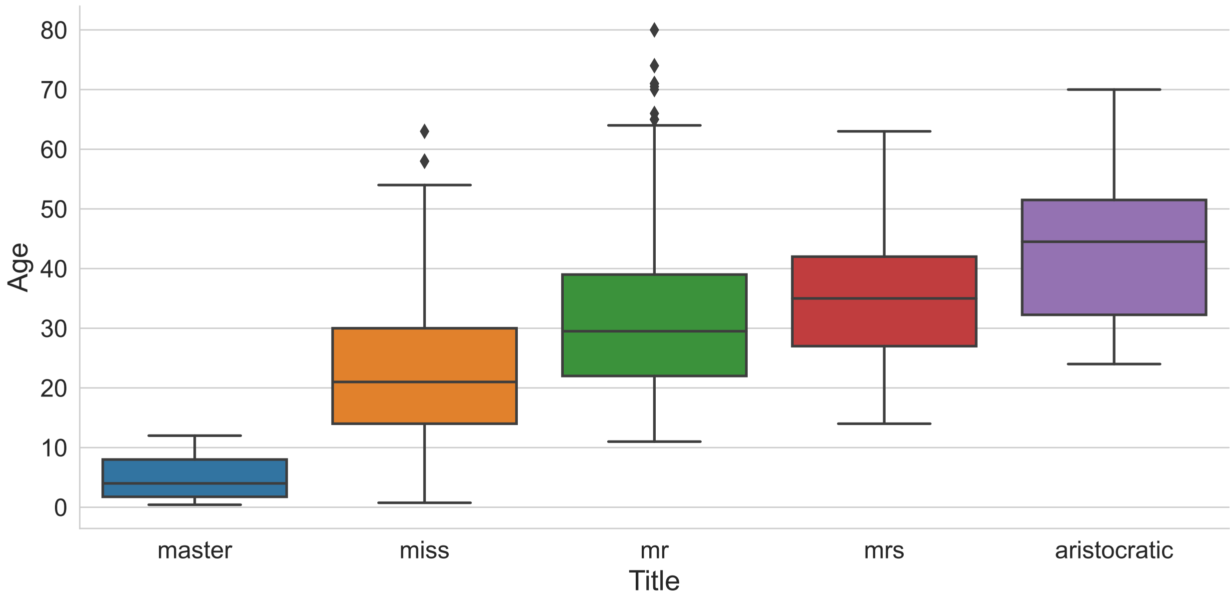

Figure 5.15: Distributions of representatives of each title by age

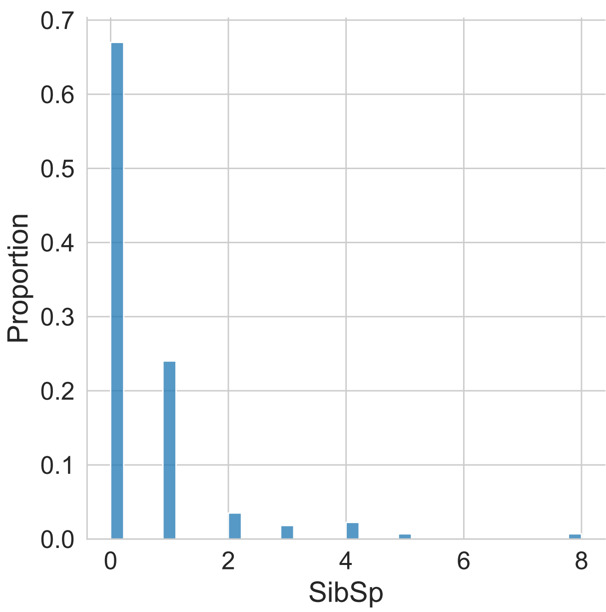

Figure 5.16: Distributions of the **SibSp** attribute's values

5.6.2 Analysis

Let's study the distribution of the attribute's values and the proportion of survived passengers for each **Parch** value. This plots are shown in the figures 5.18 and 5.19.

Again, the distribution is more like an exponential one, and passengers with parents or children seem to be more likely to be saved. Perhaps it would be more efficient to combine the **SibSp** and **Parch** attributes into one.

5.7 Ticket

The **Ticket** attribute contains the ticket number, for example, A/5 21171. There are 566 unique values in this attribute, thus it is most likely useless for analysis. The tables 5.18 and 5.19 show an excerpt from this column and the attribute characteristics respectively.

Table 5.18: Excerpt from the **Ticket** column

Index	788	347	629	734	106
Value	C.A. 2315	386525	334912	233639	343120

Table 5.19: Characteristics of the **Ticket** attribute

Characteristic	**Value**
Type	object
Number of values	712
Number of non-null values	712
Number of unique values	566
Most frequent value	1601

5.8 Fare

5.8.1 Description

This attribure contains the passenger fare. The tables 5.20 and 5.19 show an excerpt from this column and the attribute characteristics respectively.

Table 5.20: Excerpt from the **Fare** column

Index	788	347	629	734	106
Value	20.5750	16.1000	7.7333	13.0000	7.6500

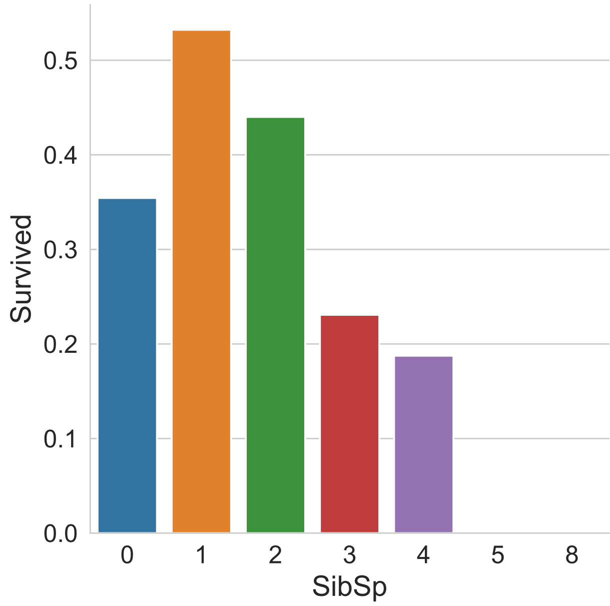

Figure 5.17: Proportion of survived passengers for each **SibSp** value

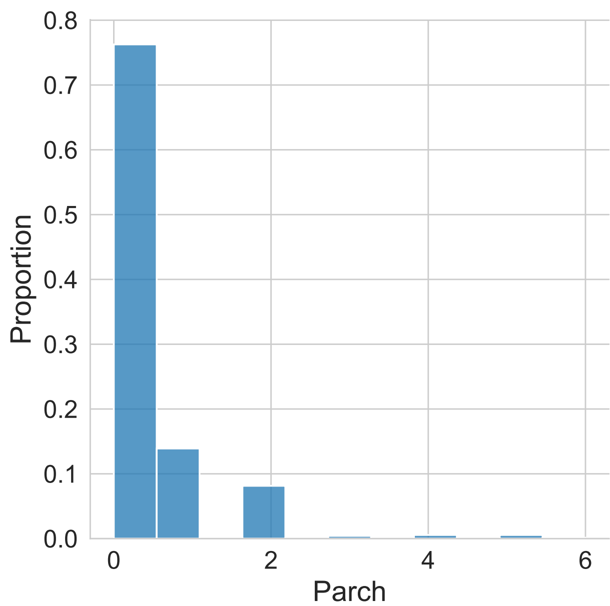

Figure 5.18: Distributions of the **Parch** attribute's values

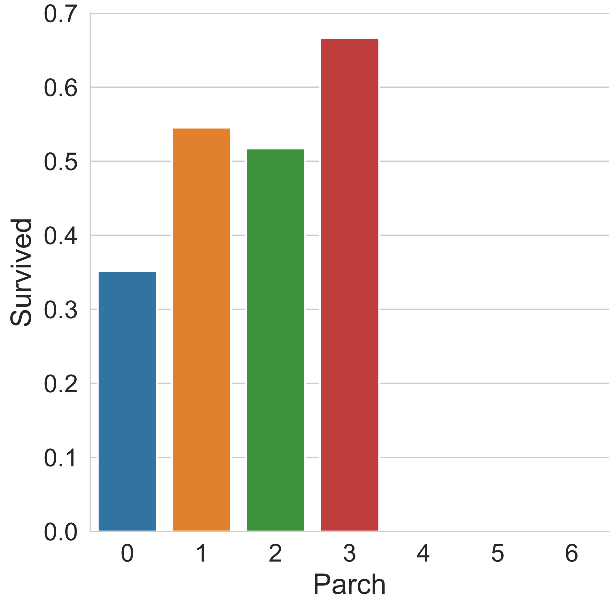

Figure 5.19: Proportion of survived passengers for each **Parch** value

Table 5.21: Characteristics of the **Fare** attribute

Characteristic	**Value**
Type	numerical (float64), continuous
Number of values	712
Number of non-null values	712
mean	31.282893
std	44.377233
min	0.000000
25%	7.895800
50%	14.456250
75%	31.275000
max	512.329200

5.8.2 Analysis

The distribution of values is skewed and tail-heavy, perhaps closer to lognormal distribution. The figure 5.20 shows the histogram.

Figure 5.21 shows boxplots for each **Pclass** value. The distributions of **Fare** in each **Pclass** are quite different, so it's resonable to concider them separately. The figure 5.22 show the boxplots of fare paid by survived and drowned passengers separately for each **Pclass**. There is no obvious pattern, but perhaps in first class, the owners of more expensive tickets had a better chance of surviving.

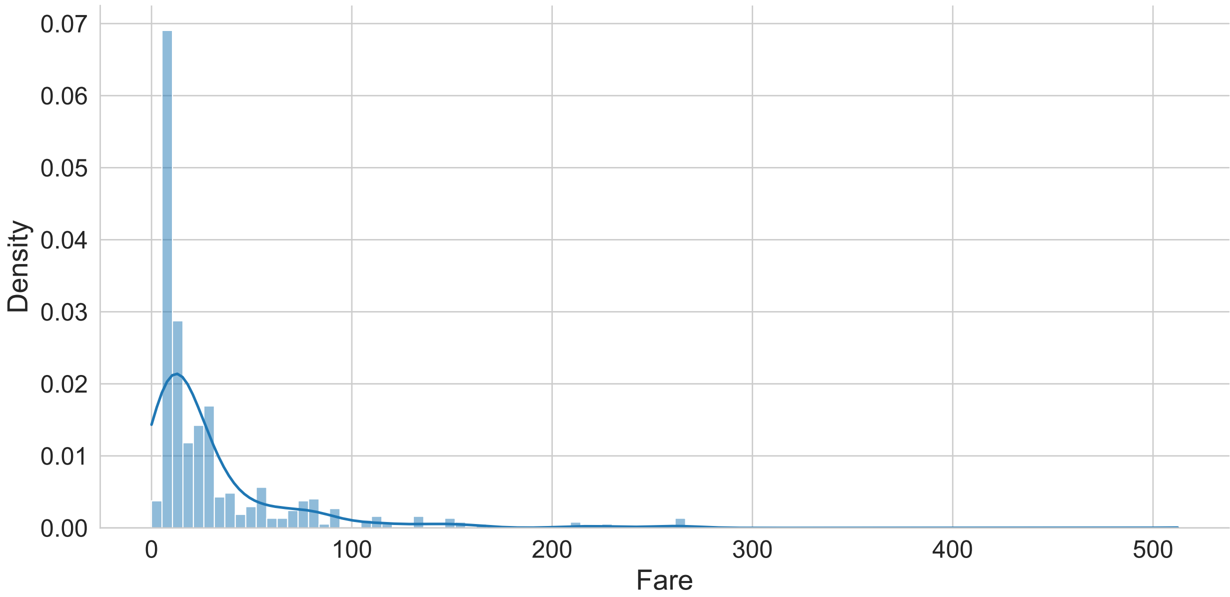

Figure 5.20: Histogram of the **Fare** attribute

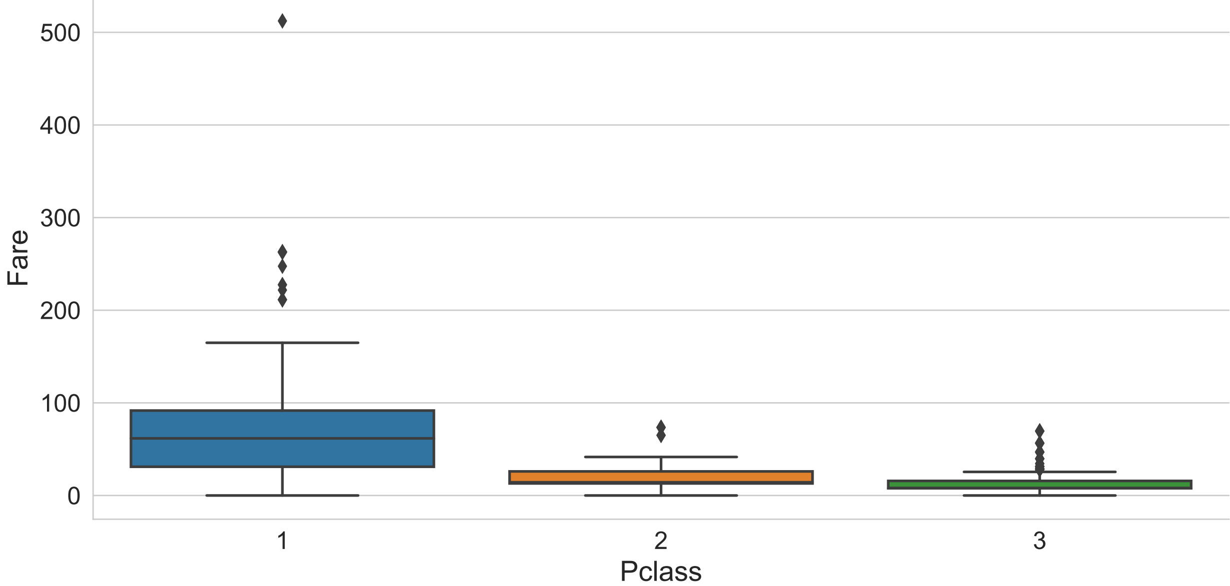

Figure 5.21: Distributions of the passenger fare by **Pclass**

5.9 Cabin

The **Cabin** attribute contains the cabin number, for example, C85. This attribute has 128 unique values out of 164 non-null values, thus it is most likely useless for analysis. However, we can use the presence of a non-nul values as a feature. The tables 5.22 and 5.23 show an excerpt from this column and the attribute characteristics respectively.

Table 5.22: Excerpt from the **Ticket** column

Index	788	347	629	734	106
Value	NaN	NaN	NaN	NaN	NaN

Table 5.23: Characteristics of the **Cabin** attribute

Characteristic	**Value**
Type	object
Number of values	712
Number of non-null values	164
Number of unique values	128
Most frequent value	C23 C25 C27

5.10 Embarked

5.10.1 Description

The **Embarked** attribute contains information about port of embarcation:

- C = Cherbourg,
- Q = Queenstown,
- S = Southampton.

The table 5.24 contains excerpt from attribute's column. The table 5.25 contains the attribute characteristics.

Table 5.24: Excerpt from the **Embarked** column

Index	788	347	629	734	106
Value	S	S	Q	S	S

The table 5.26 contains the number of occurrences of each value.

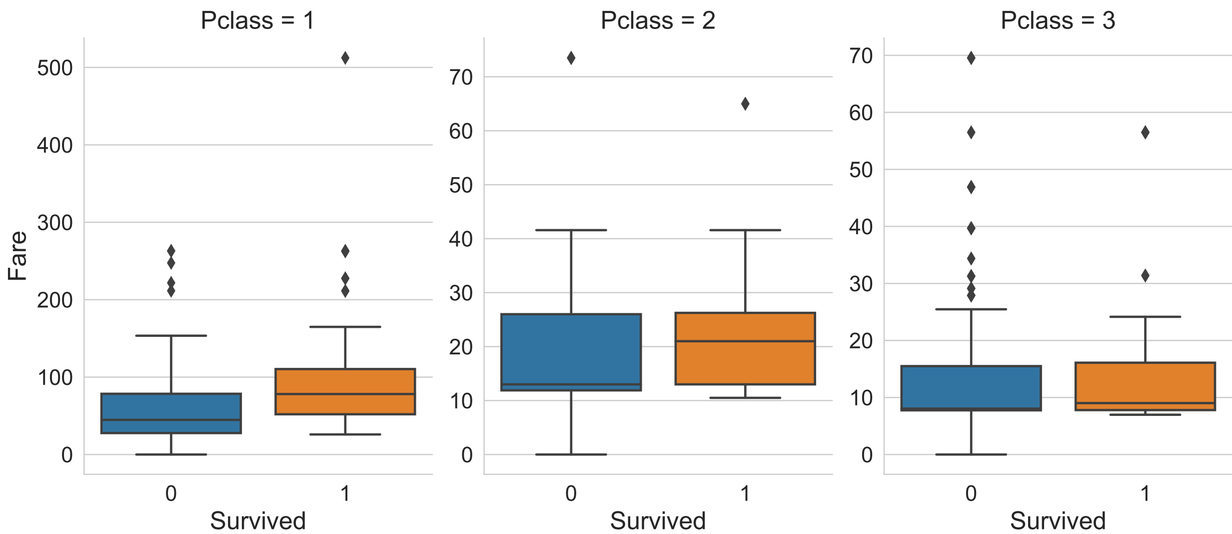

Figure 5.22: Distributions of fares for survivors and drowned passengers grouped by **Pclass**

Table 5.25: Characteristics of the **Embarked** attribute

Characteristic	**Value**
Type	category
Number of values	712
Number of non-null values	710
Number of unique values	3
Most frequent value	S

Table 5.26: Number of occurrences of each value of the **Embarked** attribute

Value	**Number**
S	515
C	132
Q	63

5.10.2 Analysis

As it shown on the figure 5.23, most of the passengers boarded the ship at Southampton.

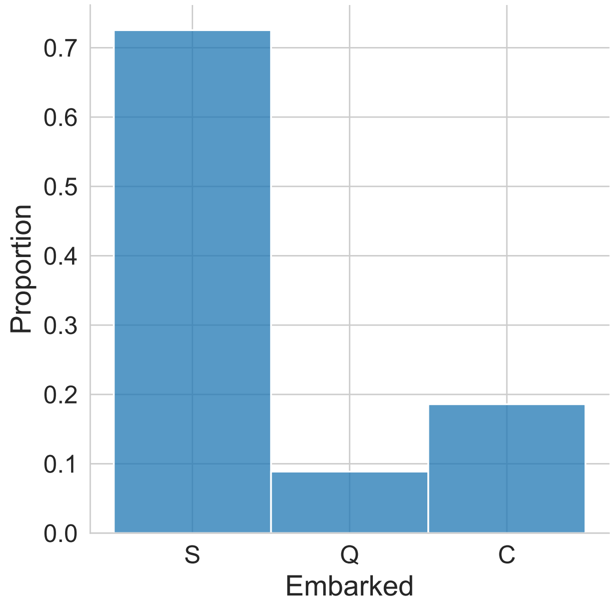

Figure 5.23: Proportion of passengers by the port of embarkation

Let's check if there is a pattern between survival rate and port of embarkation. As shown on the figure 5.24 most of the survivors among passengers boarded the ship in Cherbourg. We remember that people of hier cocio-economic status had better chance to survive, thus let's heck the proportion of survivors by the port of embarkation separetely for each **Pclass**. As you can see, the most of the passenges from the first **Pclass** boarded the ship in Cherbourg. This may explain the observed pattern.

5.11 Correlation among attributes

Finally, let's examine the correlation among the numerical attributes. The figure 5.26 shows the diagonal correlation matrix. Some attribute pairs are highly correlated, thus possibly we will make new decorrelated features from them.

This concludes the exploratory analysis. In the next section, we will clean the existing data and create new features.

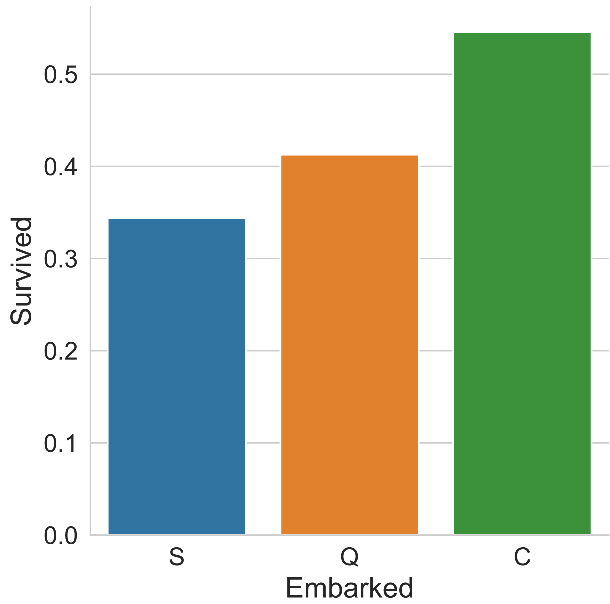

Figure 5.24: The proportion of survived passengers by the port of embarkation

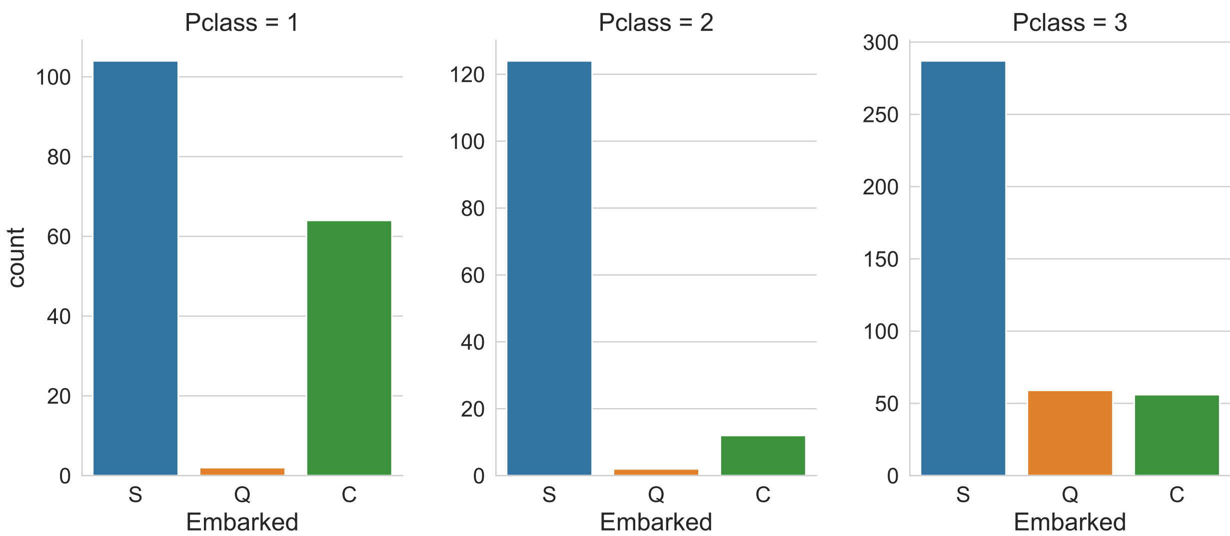

Figure 5.25: The proportion of survived passengers by the port of embarkation separately for each **Pclass**

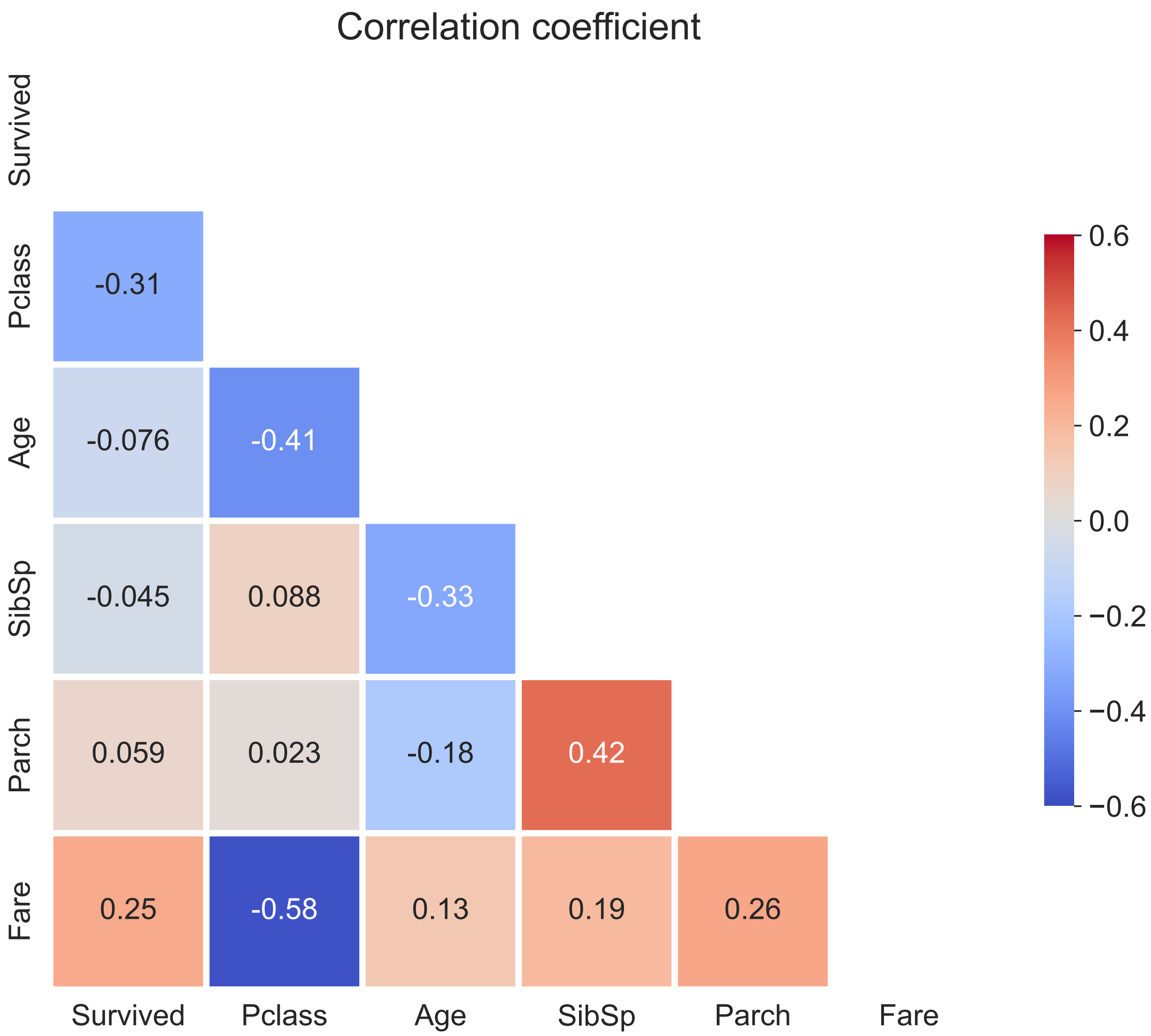

Figure 5.26: Correlation matrix

Chapter 6

Data preparation

Some preparations.

References

[1] *English honorifics*. URL: https://en.wikipedia.org/wiki/English_honorifics (visited on 09/10/2022).

[2] Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. O'Reilly Media, Inc., 2019. ISBN: 1492032646.

[3] *How to be a Top LB Explained for Beginners*. URL: https://www.kaggle.com/code/suzukifelipe/how-to-be-a-top-lb-explained-for-beginners/notebook?scriptVersionId=99817039 (visited on 08/24/2022).

[4] *Sinking of the Titanic*. URL: https://en.wikipedia.org/wiki/Sinking_of_the_Titanic (visited on 08/24/2022).

[5] *Titanic*. URL: https://en.wikipedia.org/wiki/Titanic (visited on 08/24/2022).

[6] *Titanic*. URL: https://www.openml.org/search?type=data&sort=runs&id=40945&status=active (visited on 08/24/2022).

[7] *Titanic - Machine Learning from Disaster*. URL: https://www.kaggle.com/c/titanic (visited on 08/24/2022).

[8] *Titanic Survivors - Names of all passengers and crew that survived. Complete list of Titanic survivors*. URL: https://www.encyclopedia-titanica.org/titanic-survivors/ (visited on 08/24/2022).

[9] *Титаник на Kaggle: вы не дочитаете этот пост до конца*. URL: https://habr.com/ru/company/mlclass/blog/270973/ (visited on 09/10/2022).