




Exploratory Data Analysis for Machine Learning

Igor Tomić
February 2022





About Data

- The sinking of the Titanic is one of the most infamous shipwrecks in history.
 - On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.
 - While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.
- 

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton



Data exploration plan

- The analysis is initial step in attempt to build a model to predict survival chances based on sex, having children, spouse, parents or siblings.
 1. Data Overview
 2. Data Cleaning and Feature Engineering: Categorical Data
 3. Data Cleaning and Feature Engineering: Numeric Data
 4. Hypothesis Testing

Data Overview

- The train set has 891 rows and 12 columns.
- There are missing data in only in Age and Cabin columns.

Numeric features: ['Survived', 'SibSp', 'Parch']

Categorical features: ['Sex']

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
SibSp	0
Parch	0
Ticket	0
Fare	0
Embarked	2
Age	177
Cabin	687
dtype:	int64

Categorical Data

1. Data Cleaning:

- Remove features that are not used to discriminate the target: Name, Ticket, Cabin, Embarked
- Also removing PassengerId

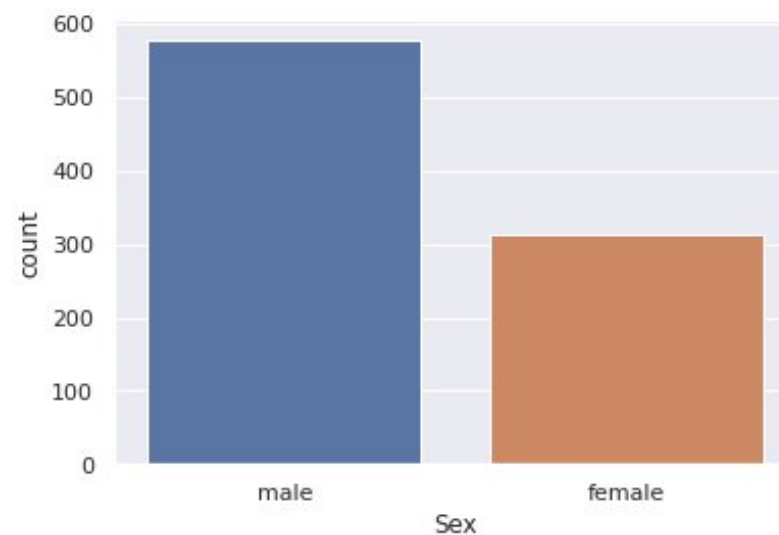
	count	unique	top	freq
Name	891	891	Braund, Mr. Owen Harris	1
Sex	891	2	male	577
Ticket	891	681	347082	7
Cabin	204	147	B96 B98	4
Embarked	889	3	S	644

Categorical Data

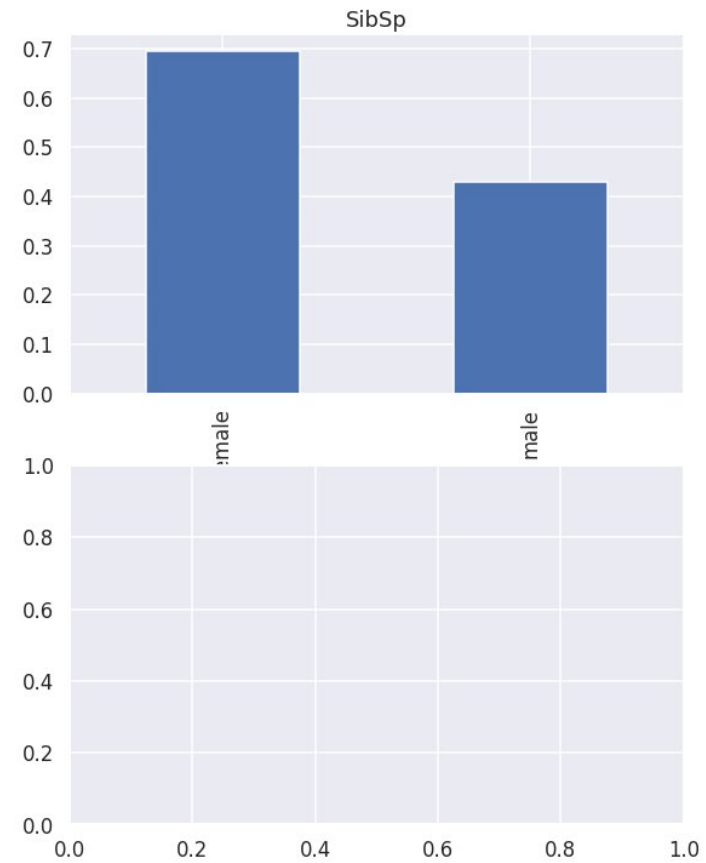
2. Feature Engineering:

- Transforming Sex column to numeric data where male becomes 0 and female becomes 1
- Counting males and females

	Survived	Sex	SibSp	Parch
0	0	0	1	0
1	1	1	1	0
2	1	1	0	0
3	1	1	1	0
4	0	0	0	0



Categorical Data



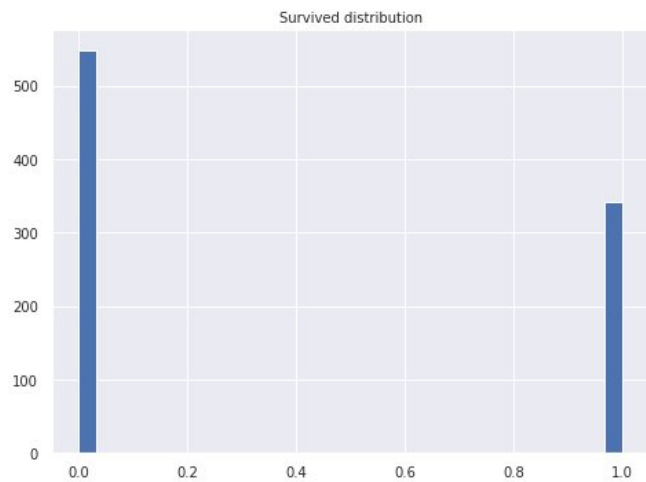
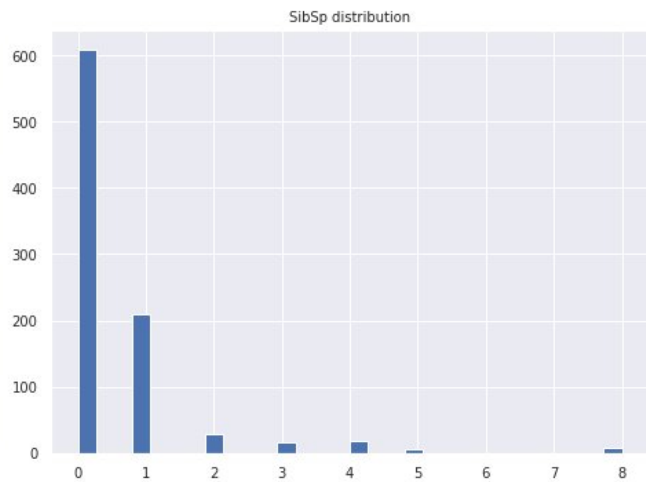
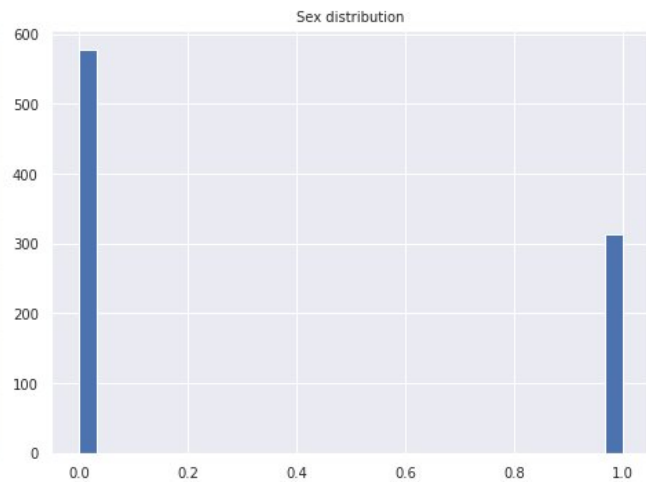
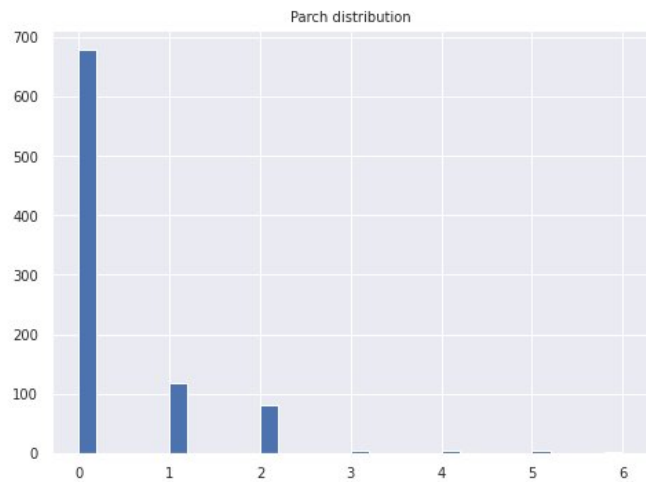
Numeric Data

1. Data Cleaning:

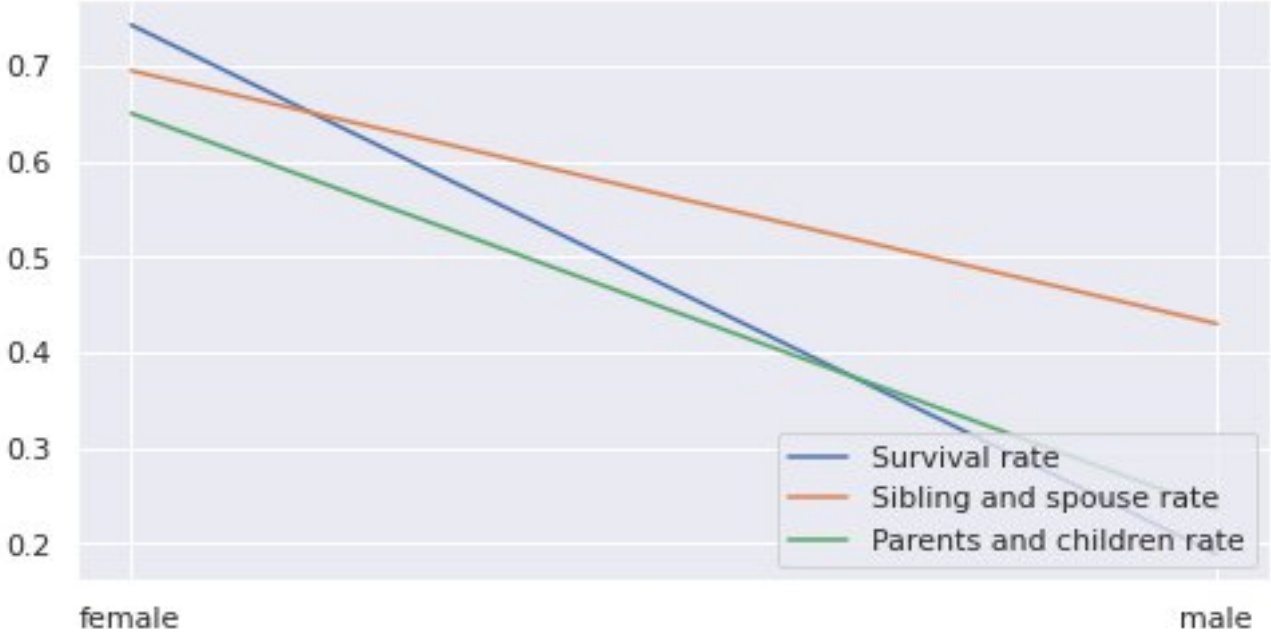
- Removed Pclass and Fare
- Also removed Age because it is not relevant to target and column has null values

	Survived	Sex	SibSp	Parch
count	891.000000	891.000000	891.000000	891.000000
mean	0.383838	0.352413	0.523008	0.381594
std	0.486592	0.477990	1.102743	0.806057
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	1.000000	1.000000	1.000000	0.000000
max	1.000000	1.000000	8.000000	6.000000

Numeric Data Distribution



Relationship Between Data



	Survived	Sex	SibSp	Parch
Survived	1.000000	0.543351	-0.035322	0.081629
Sex	0.543351	1.000000	0.114631	0.245489
SibSp	-0.035322	0.114631	1.000000	0.414838
Parch	0.081629	0.245489	0.414838	1.000000




Hypothesis Testing

Null Hypothesis

- Male have lower survival rate and having children, sibling, spouse or parents with you on ship significantly decreases survival rate.

Alternative Hypothesis

- There is no relationship between male and female survival rate and having children, sibling, spouse or parents with them does not influence survival rate.
- 

Hypothesis Testing

- Since in both calculations $p > 0.05$ we can accept the null hypothesis and we can tell that males with children, sibling, spouse or parents are less likely to survive.

```
ss.kruskal(male_dead['SibSp'], female_dead['SibSp'])  
ss.kruskal(male_survived['SibSp'], female_survived['SibSp'])
```

✓ 0.3s

```
KruskalResult(statistic=2.4242226489487697, pvalue=0.11947249452560628)
```

```
ss.kruskal(male_dead['Parch'], female_dead['Parch'])  
ss.kruskal(male_survived['Parch'], female_survived['Parch'])
```

✓ 0.4s

```
KruskalResult(statistic=2.4435813598496208, pvalue=0.11800652886291178)
```



Quality of Data Set

- Quality of data set is average but considering age of the data and that number of passengers was small and limited it was good enough.



Next Steps

- As this data set show inclination towards linear regression maybe this kind of model will be a good choice for this data set.
- Jupyter notebook can be found at:
<https://github.com/igortomic99/IBM-Exploratory-Data-Analysis-for-Machine-Learning-Project>

