Kaggle - Titanic: Machine Learning from Disaster

Section 1: Introduction

The purpose of this data science project is to predict who survived the sinking of the titanic. The data we will be using contains the following features to use for prediction:

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
passengerId	ID assigned to passenger	

Section 2: Exploratory Analysis

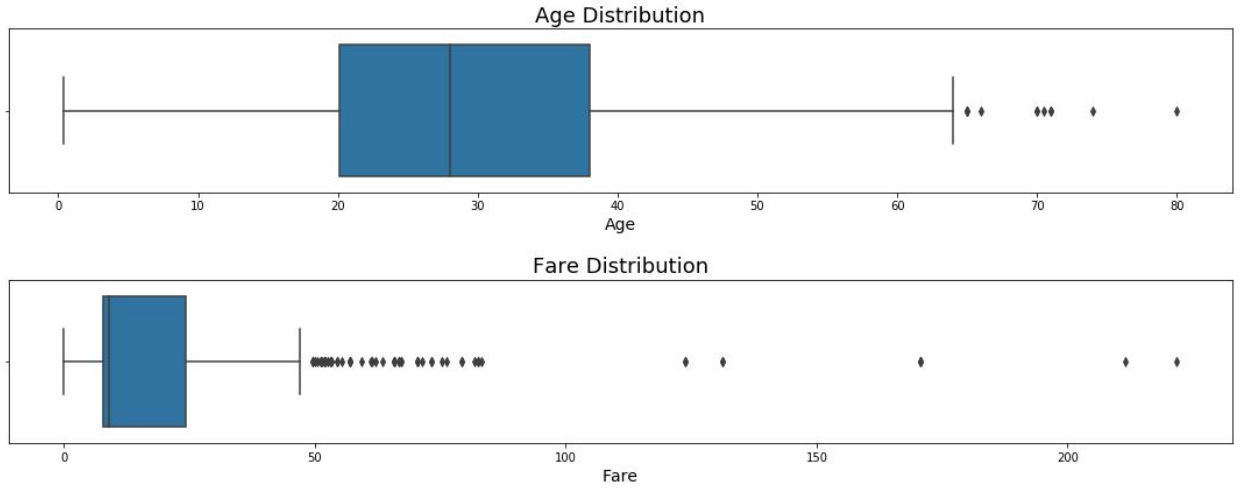
I will start by partitioning the variables into the categories of numerical, discrete, and categorical

Variable	Definition
survival	categorical
pclass	categorical

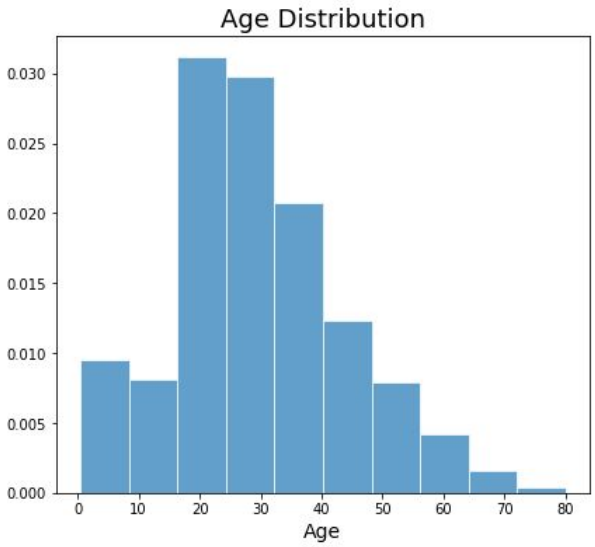
sex	categorical
Age	numerical
sibsp	discrete
parch	discrete
ticket	categorical
fare	numerical
cabin	categorical
embarked	categorical
passengerId	categorical

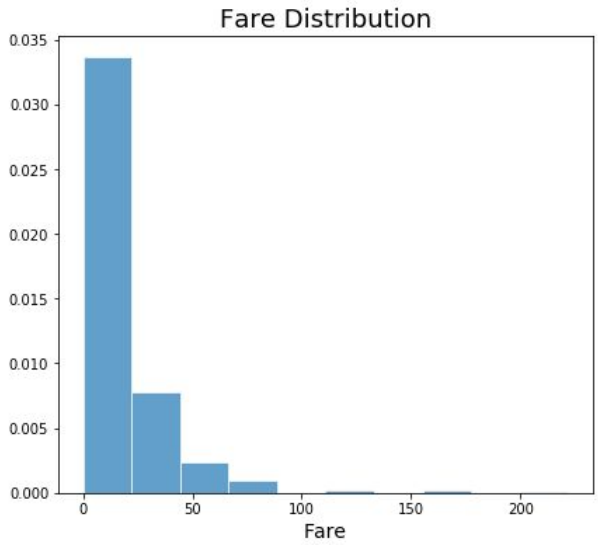
Section 2.1: Outlier Detection

For our numerical variables we need tto be concerned with outliers depending on which machine learning algorithm we use. To get a sense of the extent of outliers, let's start by examining the box and whisker plots of the continuous variables.



For both Age and Fare, outliers are present and will need to be dealt with. It will also be helpful to get an understanding of the distributions for both features are so let's take a look at their histograms.





Fare is heavily skewed and thus I will use the inner-quartile range to detect outliers. Age is gaussian enough that I will use the standard deviation to detect the outliers. Later I will be doing some feature engineering on this data so I will not be addressing the outliers for now.

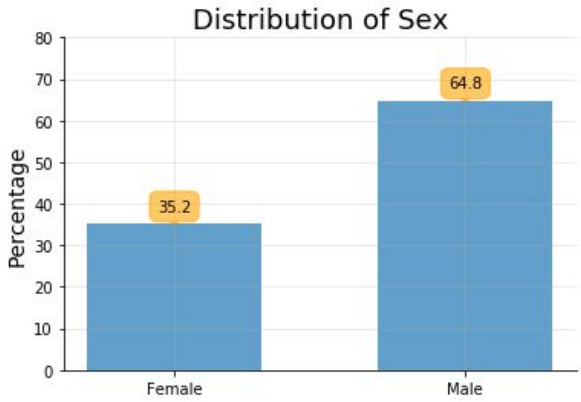
Section 2.2: Missing Data

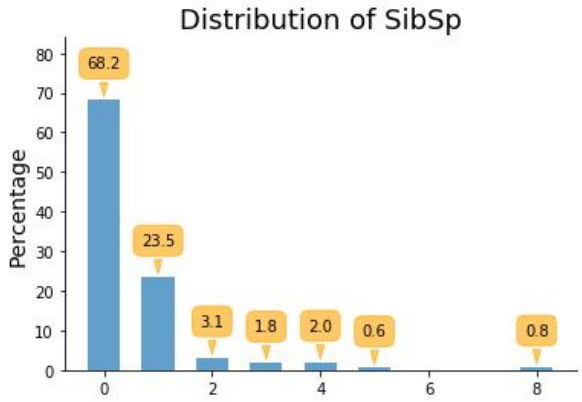
The following table summarizes the variables in which data is missing.

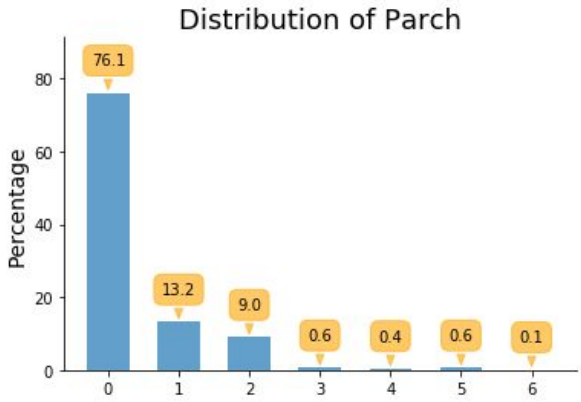
Variable	**Percent of Missing Data**
Age	19.8%
Cabin	77.1%
Embarked	0.2%

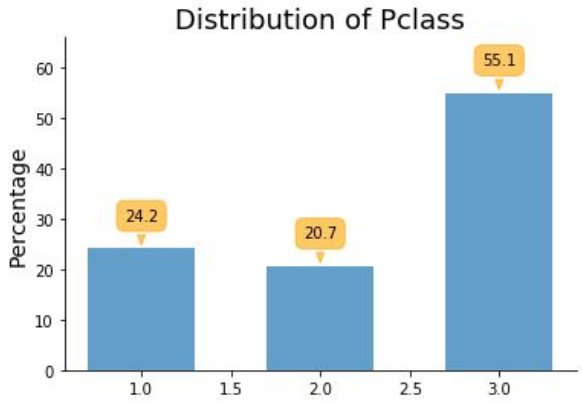
Section 2.3: Rare Categories

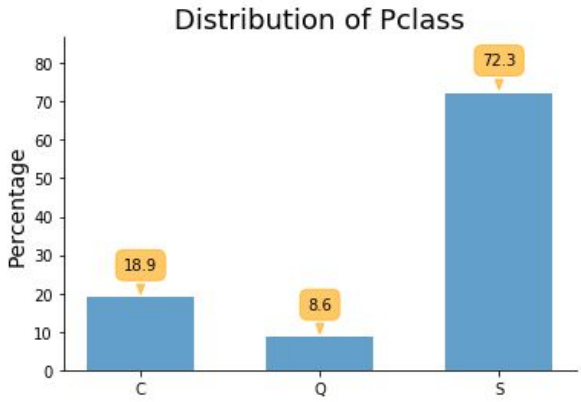
The following charts will be useful for detecting which features contain rare categories.











In both SibSp and Parch there are categories that appear in less than one percent of the data. With only 891 observations, I would classify these values as rare, but I'm going to do some feature engineering with them so I'm going to keep them as they are for now.

Section 3: Feature Engineering

In this section, I will explain the changes I've implemented to the data to improve the performance as well as explain how new features were made.

Section 3.1: Dealing with Missing Values

Age, Cabin, and Embarked all contained missing data, which I'm now going to address.

Section 3.1.1: Age

Age contained missing data for nearly 20% of the observations. I could address this by choosing to impute the missing values with the average of the miss