Lab 2 | Data Science

# Data Cleaning the Titanic

Orange Painting of the Sky

## Main Goal & Objectives

The main goal of this lab is is to work with the Titanic Data Set. For background information, the titanic was a passenger ship that famously sank in 1912. The dataset is a classic, canonical data set in the field of data science and machine learning.

The main focus of this lab is to load the data, build an understanding of what is in the dataset, and cleaning the data for later use.

### Section 1 | Reading in the Data

In this section, we work with pandas to read data from a CSV file, and then read that same file manually by parsing the file line by line into a dictionary. We then convert that dictionary into a Pandas data frame.

#### Question 1 | What do the rows and columns of the data frames you have read mean? What kind of information is contained in each?

Upon a closer inspection at the data frame, a row consists of details of each passenger, from the passenger ID to port of which the passenger had embarked. When we look at the columns in the data frame, they are either in the form of categorical or numerical.

#### Question 2 | Use .info() and .describe() methods on the data frame from .read_csv() and on the comes from the manual approach? What information is generated from these two methods, and what differences do you observe between results of the two data frames?

After reading a CSV file using .read_csv(), we can use .info() to find more information about the data frame, such as the name of the column, a count of non-null values, as well as the data type. When using .describe(), we can see statistics relating to each column of the data frame, such as the count, mean, standard deviation, quartiles, and the min and max.

With regards to the manual approach, when we use .info() to find more information about the data frame, the count of non-null values is 891 for all of them, which may suggest that a mistake was made during the process. When it came to using .describe(), we are shown statistics such as count, unique, top and frequency, which is less than what we get when we use Pandas.

**Question 3 | For each variable in the data, describe what the variable measures and any relevant information such as the meanings of individual codes, the units, etc.**

Each variable in the data provides information about each passenger that was aboard the Titanic. With regards to what each variable measure is, and relevant information, they are presented as follows in the bullet points below:

- **PassengerId:** this is the ID number of the passenger that was aboard on the Titanic
- **Survived:** this is whether the passenger survived. 0 means "No" and 1 means "Yes"
- **Pclass:** this is the ticket class, where 1 is "1st", 2 is "2nd" and 3 is "3rd"
- **Name:** this is the name of the passenger that was aboard on the Titanic
- **Sex:** this is the sex of the passenger, with regards to being "male" or "female"
- **Age:** this is the age of the passenger that was aboard on the Titanic
- **SibSp:** this is the number of siblings or spouses that was aboard on the Titanic
- **Parch:** this is the number of parents / children that was aboard on the Titanic
- **Ticket:** this is the ticket number of the passenger that was aboard on the Titanic
- **Fare:** this is the fare for the passenger to board the Titanic, but no currency indicated
- **Cabin:** this is the number of the cabin the passenger used
- **Embarked:** this is the port at which the passenger embarked, where S means "Southampton", C means "Cherbourg" and Q means "Queenstown"

## Section 2 | Representing in the Data

In this section, we count the unique values in the Pclass, SibSp, Parch, Fare and Cabin variables to determine if they should be represented as integers, floats, or categorical data. We then convert them to categorical variables and create a bar chart of the Survived variable using Seaborn's countplot  function.

### Question 1 | Regarding Pclass, SibSp, Parch, Fare, and Cabin variables, why did you choose integer, float, or categorical for each?

With regards to the variables mentioned, the reasoning is as follows below:

- **Pclass:** The best way to represent this feature is categorically, as it is a categorical feature representing the class of ticket using 1, 2 and 3.

- **SibSp:** The best way to represent this is as a integer, as it isn't possible to have floating point values for a discrete feature.

- **Parch:** The best way to represent this is as a integer, as it isn't possible to have floating point values for a discrete feature.

- **Fare:** The best way to represent this is as a floating point value, as currency can have floating point values.

- **Cabin:** The best way to represent this is as a integer, as it is a discrete numerical feature.

### Question 2 | Describe the results of plotting the Survived variable. What did you learn about the data set? How does it compare to the historical survival rate aboard the Titanic?

After plotting the "Survived" variable, it can be seen that there are more people dead than alive, with 342 people alive and 549 people dead. According to the RMS Titanic, the survival rate was 32%. We can find the survival rate found in the dataset by performing the calculation 342 / (342 + 549) * 100 = 38%. As we can see, there is a 6% difference between the survival rate from RMS Titanic and the survival rate from the dataset.

## Question 3 | Identify a variable which would make a good response variable for a supervised machine learning problem – would that variable be suit-able for classification or regression?

A variable that would serve as a good response variable would be the gender of the passengers. This would make a great variable for a supervised machine learning problem.

One justification for that would be is that women and children are given the most priority for safety in a moment of crisis, hence when the titanic was sinking, they were most likely given the rescue boats first. It would be interesting to test this theory by seeing which gender was more likely to have a higher chance of survival during that event.

This variable would be suitable for a classification, as the sex of a passenger can only be male or female. As a result, we will be able to get more insight into Figure 1 presented below, rather than displaying just a simple bar chart showing the numbers of people who survived and those who didn't.

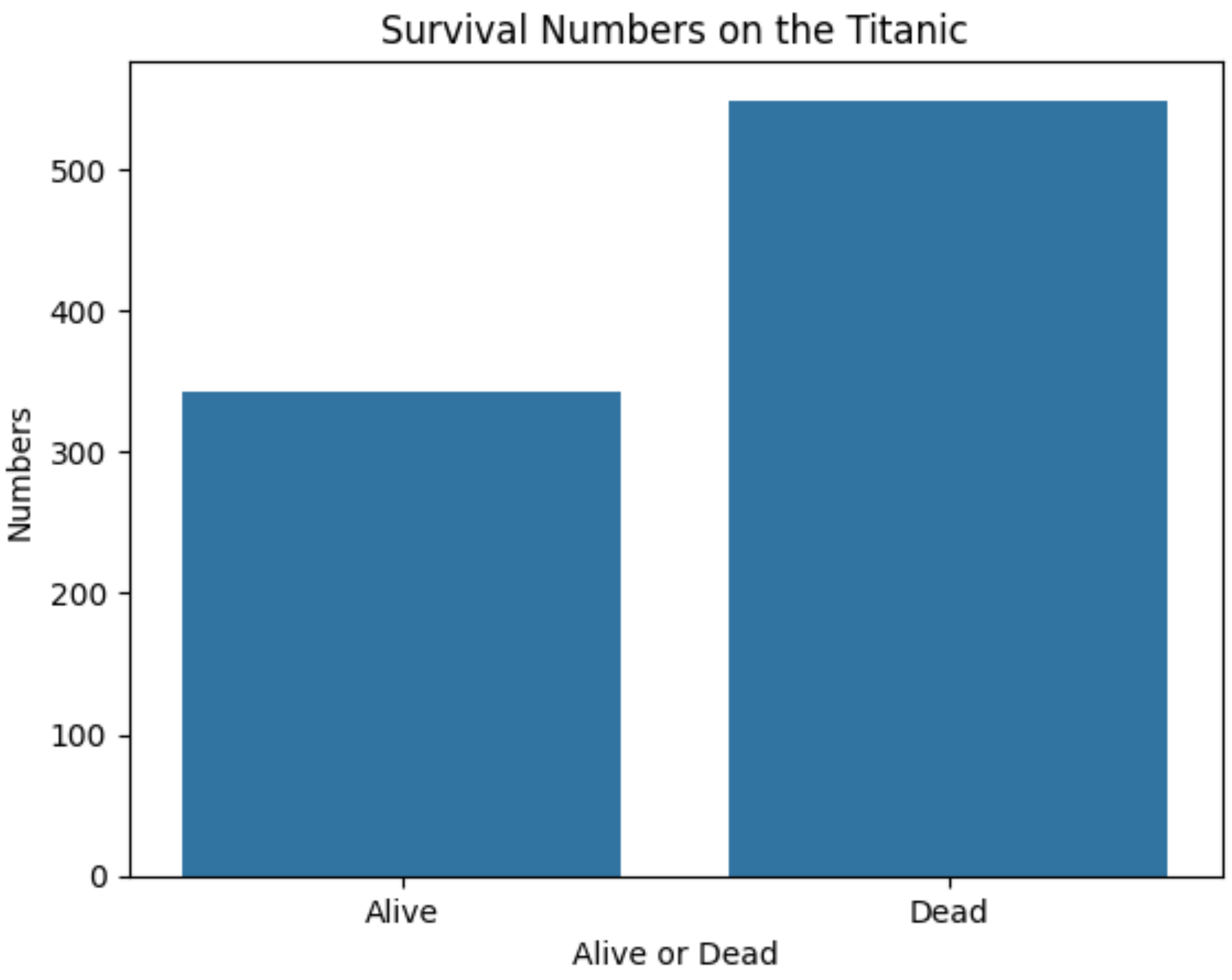

Figure 1

## Section 3 | Missing Values

In this final section, we fill the missing Age values by creating three columns with the mean, median and mode. We also use a supervised learning model (KNN) to predict the missing values and fill them in a new column.

### Question 1 | How many passengers in this data set are missing age information?

By using .isna().sum() on the titanic dataset, we can determine that there are a total of 177 passengers that have missing age information

### Question 2 | Describe what you did in applying the KNN algorithm to fill missing values.

To fill the column *Age_fill_knn* with the missing values, we first started by importing KNeighborsRegressors and not KNeighborsClassifier because we are trying to predict numerical values.

Next, we created two new data frames, one data frame consists of no missing values for age, and another which contains the missing values for age. I wanted to cross-check by adding up the number of elements in both data frames to see if they equal to the number of elements in the original age column.

Afterwards, we created two new data frames from the dataset with no missing values. The first dataset is a *x_train* which are the features, and the other is *y_train* which is the target variable. The last data frame is the *x_predict* which contains all the rows with missing age values.

We then instantiated a KNN model with the neighbors parameter being set to 5, and then we fit the model with the training data *x_train* & *y_train* and then perform the predictions using the *x_predict* dataset.

Lastly, we use boolean masking to fill in the missing values from the *Age_fill_knn* using the predicted values. We achieve this by creating a value that contains the true / false missing values from the *Age_fill_knn* with regards to whether there is a missing value or not. We then fill in missing values from the list of predicted values using *.loc*.

**Question 3 | Use density plots to compare the results of these different methods of filling missing values. For example, import the seaborn library and use kdeplot. How do you interpret the resulting plots? What do the x and y axes mean? What do they tell us about the effects of these different methods of filling in missing values for Age?**

To begin our analysis, we start off by looking at the Kernel Density Estimation (KDE) Plots being listed as Figure 2. The x-axis indicates the age of the passengers on the titanic, and y-axis indicates the probability density.

When it comes to interpreting the plot, the peaks, such as the peak for the mode corresponds to the mode of the data, and so on for the median and mean. For the KNN, the peak may mean that the predictions are clustered around a central value.

When we look at the shape of the peak, we can see that the Mode, Median and Mean peaks are sharp, compared to the KNN peak where it is lower and more spread out. This may mean that the KNN fills out the missing values based on the nearby numbers, thus this means there is more variability, hence a more spread out plot.

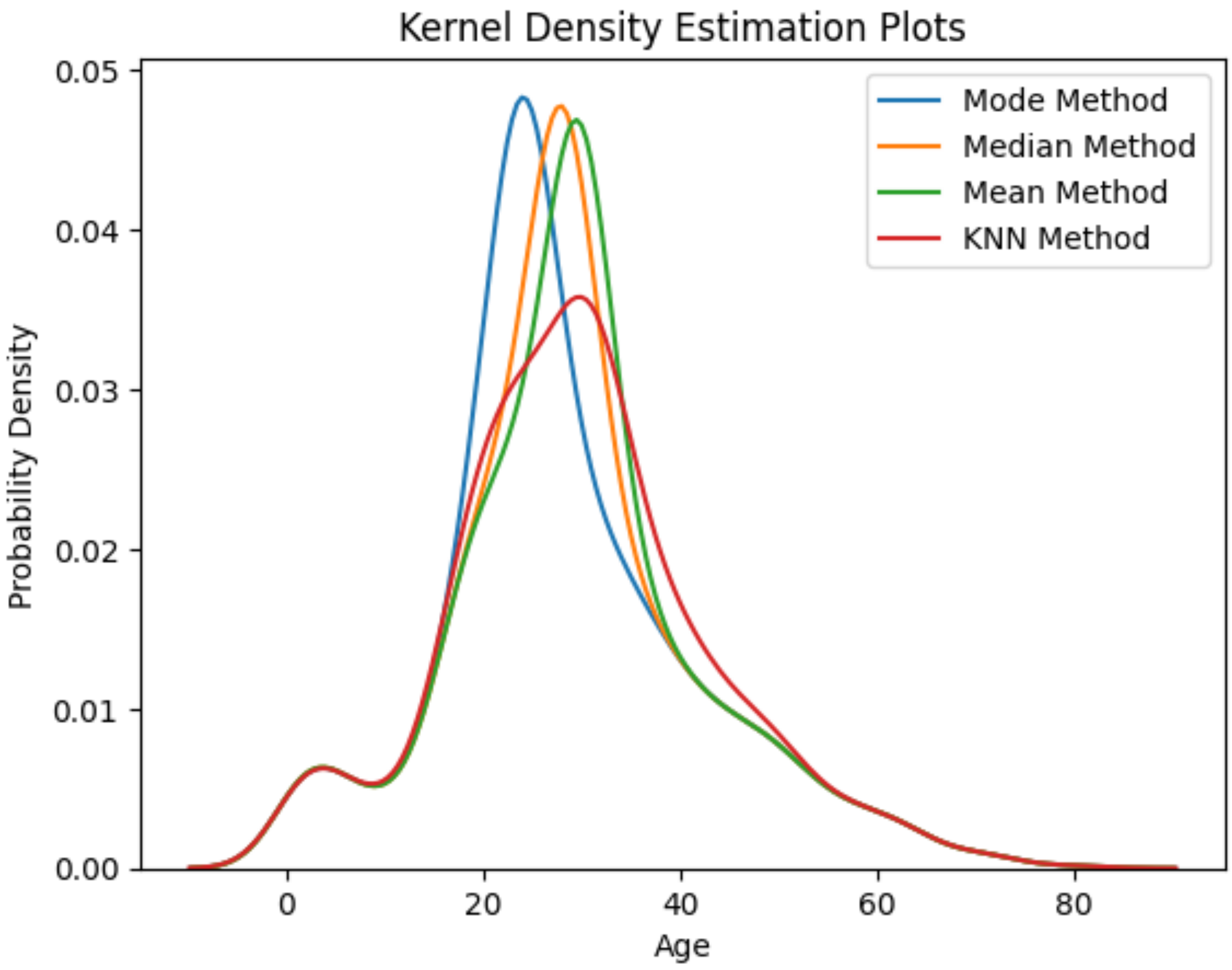

Figure 2

## Question 4 | Consider the Cabin variable. Why would this variable not work well with the methods we applied to fill in missing values for Age? Identify at least one reason that has to do with interpretation rather than with code.

With regards to the Cabin variable, it is as first categorical. This means that we can't calculate the mean or median. While we can find the mode by finding the most repeated Cabin room, this may mean that we can find the most commonly used cabin room, however there's not much we can do with this information.

Moreover, with this categorical feature, we know that it is nominal and not ordinal. This is because we can't apply order to cabin rooms.

If we were to attempt to use KNN for filling in missing values, we would have to start by turning the categorical data into numerical form, and this often involves using a technique using one-hot encoding.

## Sources

“How Many People Survived the Titanic - RMS Titanic Inc.” Expedition Discover Titanic, 2024, expedition.discovertitanic.com/how-many-people-survived-the-titanic/.

Minkyung's blog. “Missing Data Imputation Using Sklearn.” Minkyung’s Blog, 21 Nov. 2020, mkang32.github.io/python/2020/11/21/Missing-data-imputation-using-sklearn.html?utm_source=chatgpt.com. Accessed 4 Feb. 2025.