The Classification Dataset

The dataset I'm using for this classification task is: The Titanic Disaster (https://www.kaggle.com/competitions/titanic/data). The goal of this dataset is to predict if a person is going to survive the titanic or not based on these variables: PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked.
In this notebook I try to conclude with which variables have a strong correlation with the classifier Survived and if they should be used in a model

The Final Dataset

Data cleaning

Null-values

Some columns contained a lot of null-values, especially Cabin and Age. I chose to remove the column Cabin due to it having over 600 out of 891 null values.
For Age I created three functions:

- one that uses the median age as the value for all null-values.
- another that uses the mean age
- the last one removes every row which has at least one column containing a null value

For Embarked

- For the two missing Embarked values I chose the most common Embarked location which was "S"

Normalizing

I chose to crate bins containing age intervals and then one-hot encoded these intervals.
I could've also normalized the column Fare, by since Fare and Pclass is basically the same thing I ended up with dropping this column.

I also created two function; min-max-normalizing and z-score-normalizing just to have the opportunity to check the possible effects of this on a model

Encoding

I changed the datatypes of Pclass, AgeGroup and Embarked to category, so I could then one-hot encode them.

Feature selection

I also chose to remove some columns. The reason for this is that I know that these columns will not affect the chances of survival. The columns I removed was: PassengerId, Name, Cabin and Ticket.

- Removing PassengerId due to it holding no meaningful data for visualizing or ML
- Removing name for the same reason
- Ticket consists of the ticket number which probably have no correlation with survival
- Cabin has way to many null-values to give us any meaningful info and therefor I removed it
- I also ended up with removing Fare in The Final Dataset, due to it adding little extra value due to the Pclass column.

EDA

Feature engineering

I chose to create a new variable FamilyMembersCount. This was created by adding SibSp and Parch together. This variable gives us and idea of how many people traveled together. A family size of one indicates that the person traveled alone.

Feature Correlation Analysis

I see a great correlation between the ground truth and: Sex and Pclass, but a little unclear how Age correlates with Survived. The heatmap shows close to no correlation between Age and Survived, but independent analysis shows a great correlation.