This is my first project where data cleaning was required for the data analysis

to produce a useful result.

There were no duplicate rows and no columns with values that were different by

a space or spelling error so that's good, however, certain columns were removed

to make analysis on the data useful.

Here is a list of all the changes I made to the columns/variables and reasons for those changes.

* Name: This column was removed because it had too many levels and it didn’t seem useful in any way.
* Cabin: This column was removed originally because almost all of it’s values were empty, after performing some analysis by running classification trees, linear models and logistic regression I will try to find some way to accurately predict the empty values in columns.
* Ticket: This column was removed for the same reason as Name
* PassengerId: This column was removed because I ran a logistic regression with it and it wasn’t important or helpful (should have seen that coming).
* Fare: This column was changed to be a numeric/quantitative column because there didn’t seem to be any levels and fare is money which is usually continuous.
* Pclass: This column was changed (when running the regression models) to a factor, this column only had 3 values which were 1, 2, 3, the data dictionary shows the meaning of those values.
* Survived: This column was changed (when running the regression models) to a factor for the same reason as Pclass.