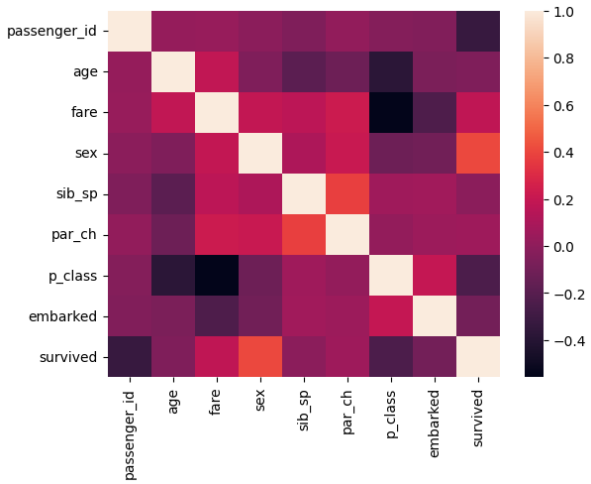
# Exploratory Data Analysis

## Procedure

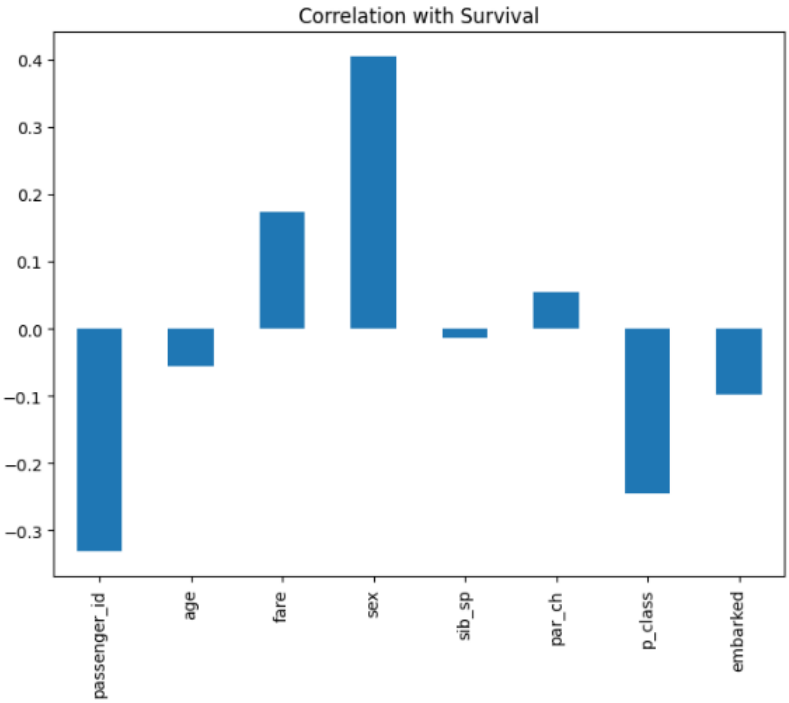
1. Obtained Titanic dataset from Kaggel
2. Performed preprocessing and cleaning of the dataset
    a. Converted columns to appropriate datatypes
    b. Replaced 0 values with NaN, dropped fields where all values were NaN, then changed NaN values back to 0
    c. Standardized the column naming scheme
3. Created multiple plots to explore the relations between the data fields, particularly between each field & 'survived'
    a. Generated heatmap to gain insight into correlation between fields
    b. Plotted correlation of each field with 'survived'
    c. Plotted regression plots for fields having high correlation with 'survived'
    d. Calculated Skewness of each field in the dataframe
4. Encoded the categorical values into seperate fields with binary values for further processing

## Graphs

1. Heatmap of cleaned dataset (brighter colors indicate higher correlation)

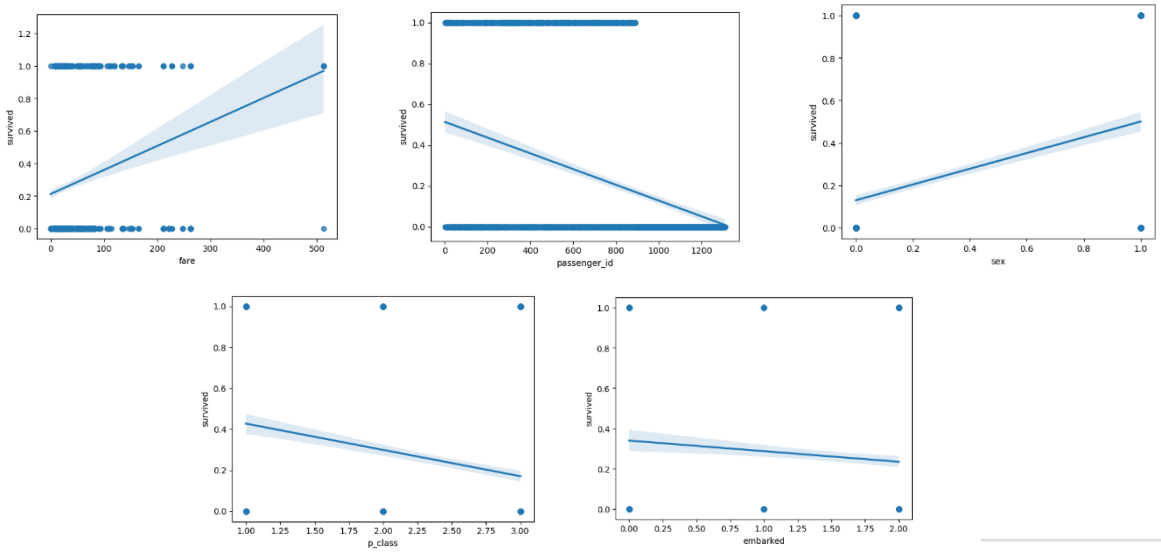


2. Plot of correlation of every field with 'survived' (except itself)




3. Regression plots to identify variation in 'survived' with fields:

'fare', 'passenger_id', 'sex', 'p_class', 'embarked'




4. Calculate skewness of each field in the dataset

| | |
|---|---|
| **passenger_id** | 0.0 |
| **age** | 0.540987 |
| **fare** | 4.36951 |
| **sex** | 0.602189 |
| **sib_sp** | 3.84422 |
| **par_ch** | 3.669078 |
| **p_class** | -0.598647 |
| **embarked** | -1.118807 |
| **survived** | 1.088057 |

## Observations

1. Survival was found to be most correlated with sex, fare, passenger_id, p(assenger)_class & embarked
2. There was positive correlation between fare & survival, with more passengers who had bought costly tickets surviving
3. There was a negative correlation between passenger_id and survival, with less passengers with higher id numbers surviving
4. Women had high survival rate than men
5. Passengers in 1st Class had the highest rate of survival, followed by 2nd Class, and much lower odds for 3rd Class
6. Passengers who embarked early had a higher survival rate
7. Some of the fields, particularly 'fare', 'sib_sp' & 'par_ch' had high skewness resulting from few individuals with outlier values causing the mean and median to diverge