<u>Statistics of Data Science Mini Project</u>

Project Title: Analysis of passenger’s survival in titanic disaster

Section: D

Team Members:

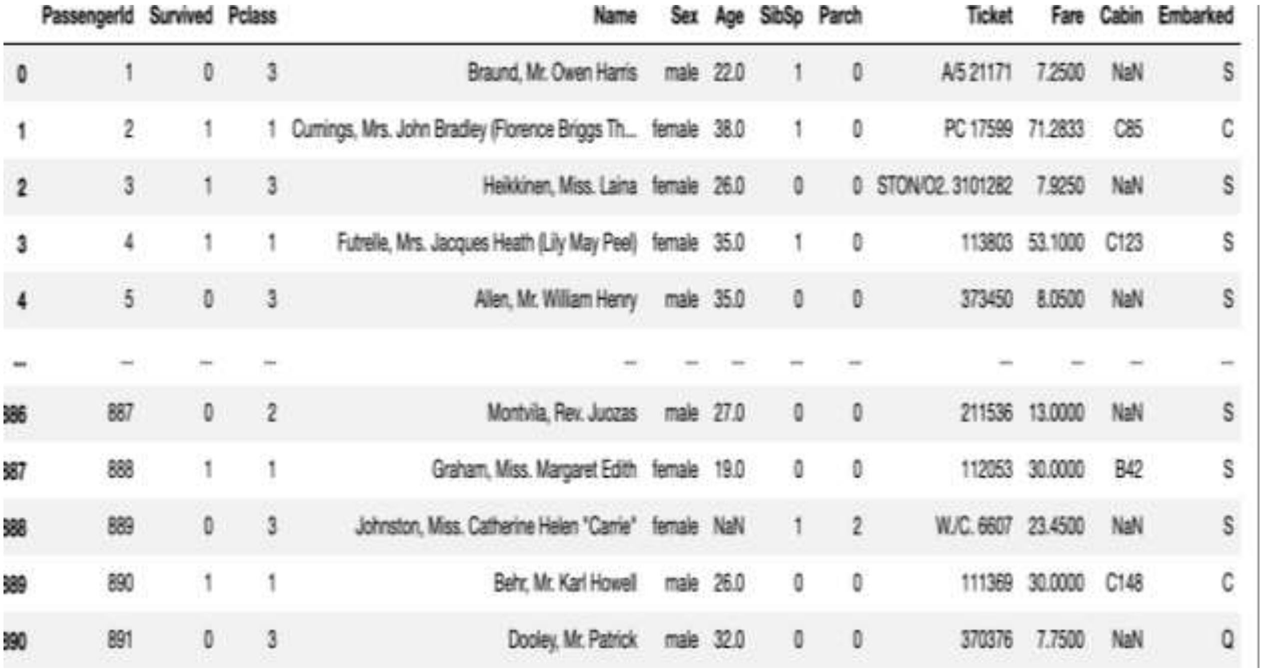
1) ManjunathGowda S - PES1UG19CS264
2) Lithesh Shetty - PES1UG19CS245
3) K S Abhisheka - PES1UG19CS202
4) Manideep P R - PES1UG19CS258

Abstract

Through the process of selecting the data set, cleaning the data, performing exploratory analysis, testing hypotheses along with normalization and correlation. The team was able to generate meaningful insight on the people who travelled on titanic ship at 14thapril 1912.

Introduction

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

Dataset

The Titanic data contains a mix of textual, Boolean, continuous, and categorical variables.
It exhibits interesting characteristics such as missing values, outliers, and text variables ripe for text mining–a rich database that will allow us to demonstrate data transformation

This dataset is taken from GITHUB.

Name: awesomedata

Repository:awesome-public-datasets

link: https://github.com/awesomedata/awesome-public-datasets

The dataset consists of 12 columns and 891 rows.

Variable description.

Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

survival - Survival (0 = No; 1 = Yes)

name - Name
sex - Sex
age - Age
sibsp - Number of Siblings/Spouses Aboard
parch - Number of Parents/Children Aboard
ticket - Ticket Number fare Passenger Fare (British pound)
cabin - Cabin
embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Types of data

Categorical data

Survived

Pclass

Sex

Embarked

Discrete data

Age

Sibsp

Parch

Continuous data

Fare

Qualitative data

Name

Ticket

Cabin

Data Cleaning

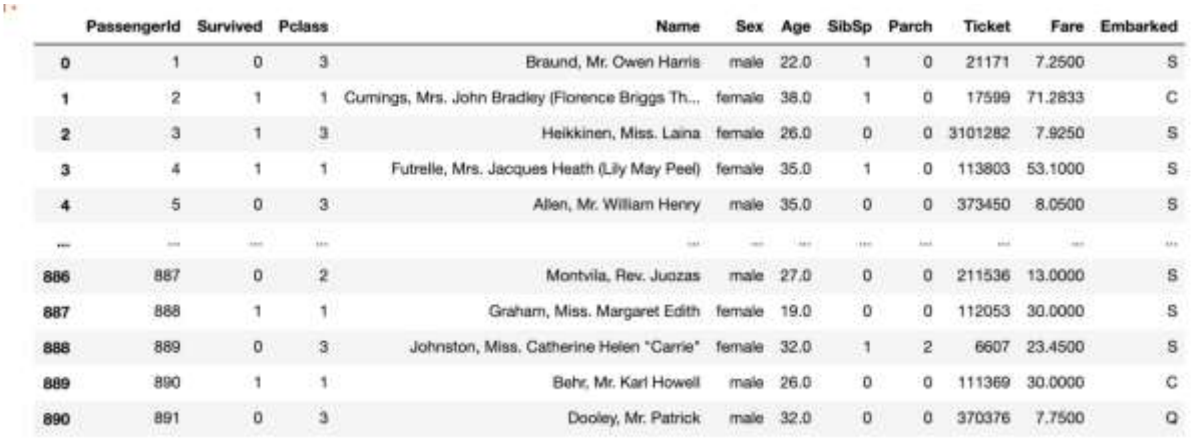
Data cleaning is the procedure of correcting or removing inaccurate and corrupt data. This process is crucial and emphasized because wrong data can drive a dataset to wrong decisions, conclusions, and poor analysis, especially if the huge quantities of big data are into the picture.

There were 177 missing values in Age column , 687 in Cabin column and 2 in Embarked column.

The dataset we obtained contained unwanted and redundant column which would simply over complicate the analysis or provide it with zero value. Hence we dropped the column ‘Cabin’. We then proceeded to check all of the remaining columns for null values and found that one numerical and one categorical variable had null values, them being Age and Embarked respectively.

We handled the missing values in Age column by using the mean age of passenger’s based on title’s.Wecategorised each passenger based on their titles.We made 4 groups based on titles like Mr,Master,Mrs and Miss.Then we found the mean of each title .We then filled the missing values by considering the titles of the passengers appropriately.

We handled the missing values in Embarked column by using the mode of that column.

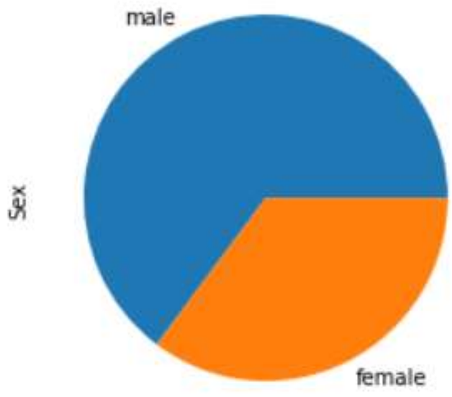


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	32.0	1	2	6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

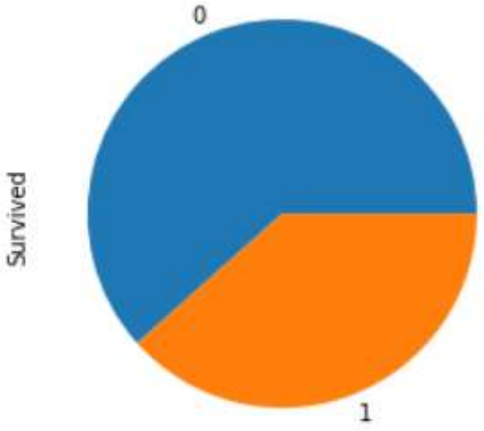
Cleaned Dataset

Exploratory Data Analysis

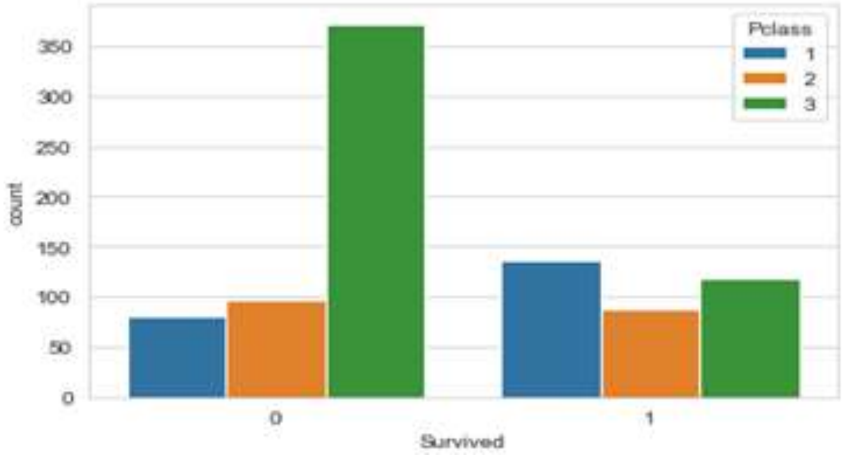
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

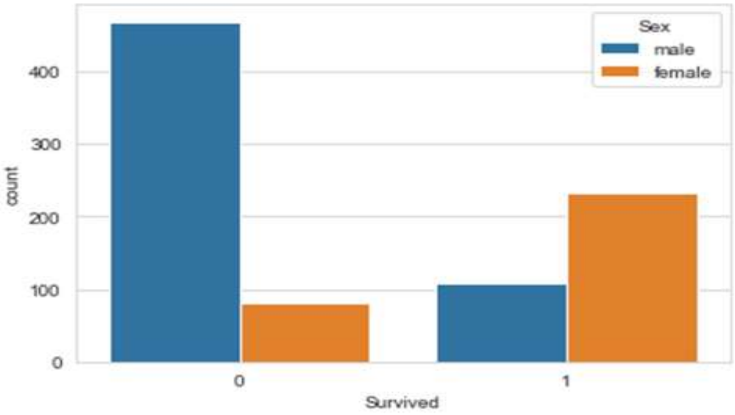


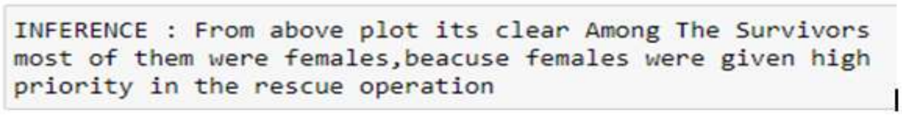
Above graph gives a clear picture of percentage of males and females onboard, 65% were male and 35% were female and the below graph depicts number of passengers survived(62%).



From the above graph its quite clear that passengers with class 1 tickets had higher chances of survival than class 3 because the rescue team gave them higher priority in rescue operation

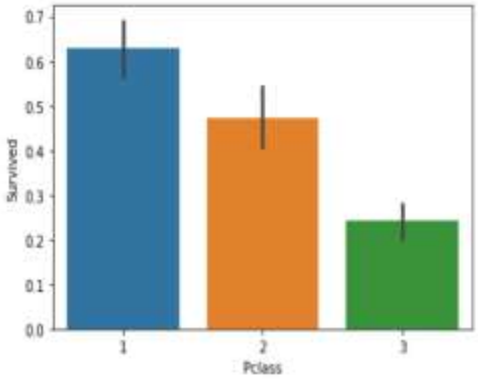


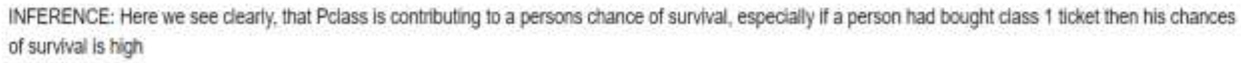




INFERENCE : From above plot its clear Among The Survivors most of them were females,beacuse females were given high priority in the rescue operation

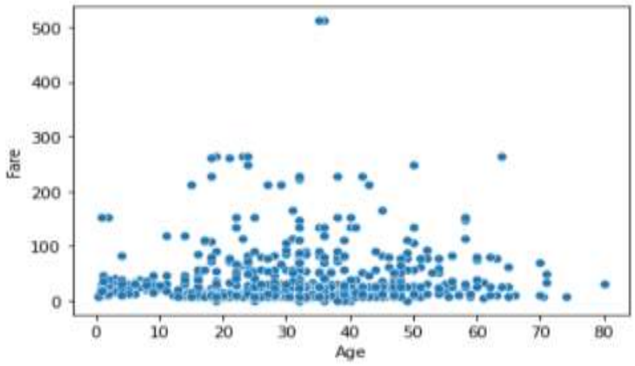
From the above graph its quite clear that among the survivors most of them were females,because females were given higher priority in rescue operation

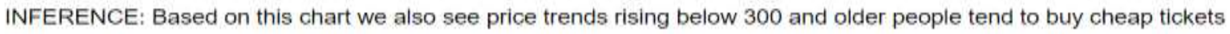




INFERENCE: Here we see clearly, that Pclass is contributing to a persons chance of survival, especially if a person had bought class 1 ticket then his chances of survival is high

From above graph we can infer that passengers with pclass 1 tickets had higher chances of survival

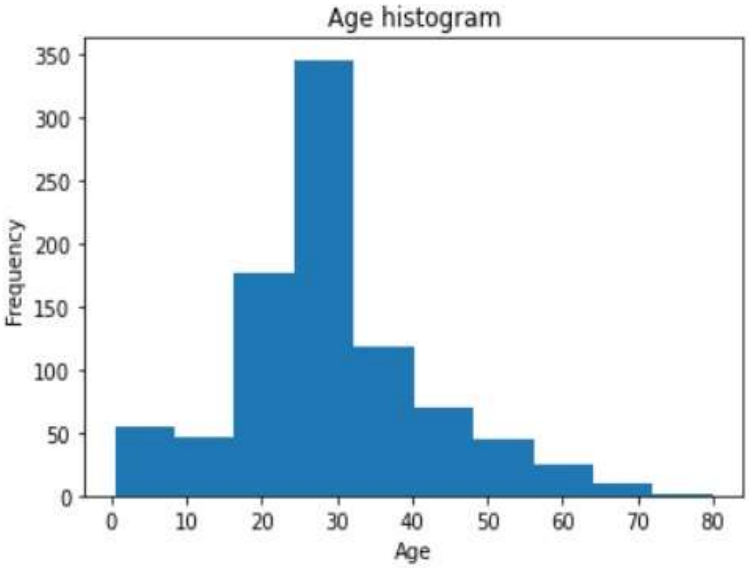




INFERENCE: Based on this chart we also see price trends rising below 300 and older people tend to buy cheap tickets

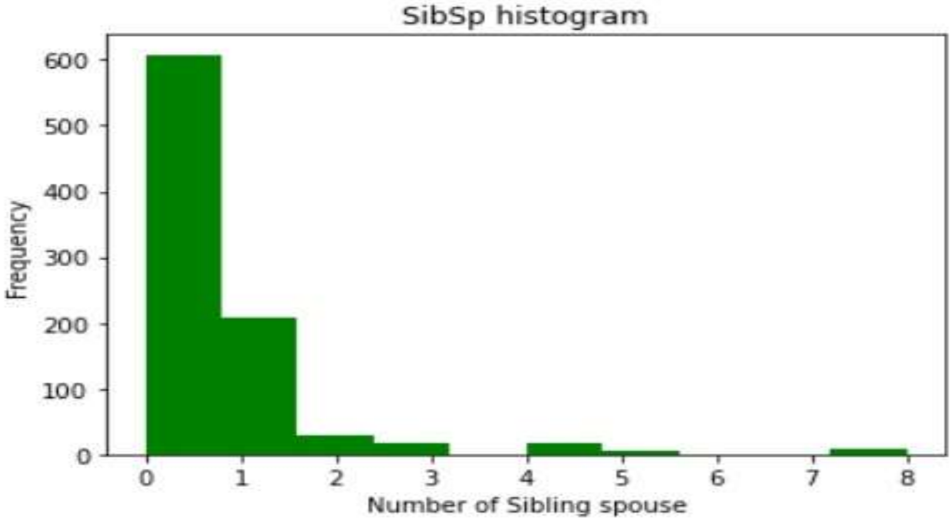
Inference:

From above graph we can infer that older people tend to buy cheaper tickets



Inference:

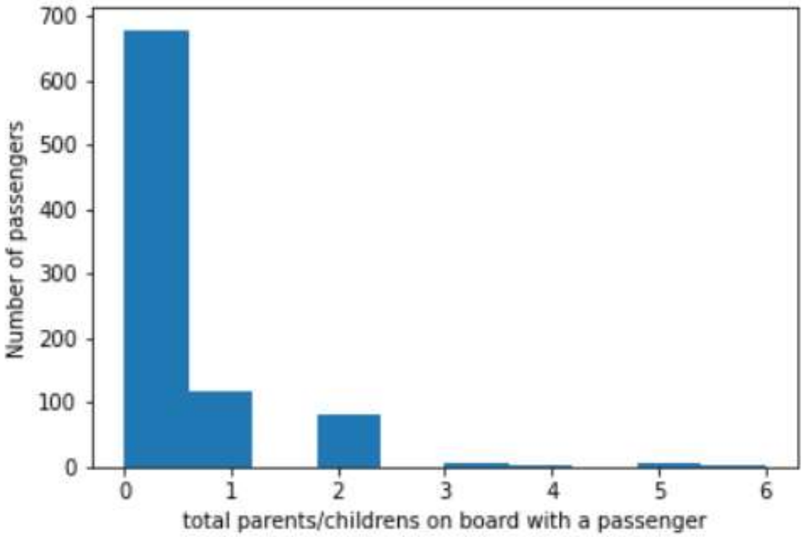
Looking at the histogram,we can infer that the mean lies in between 20 and 40 as the data is normally distributed.



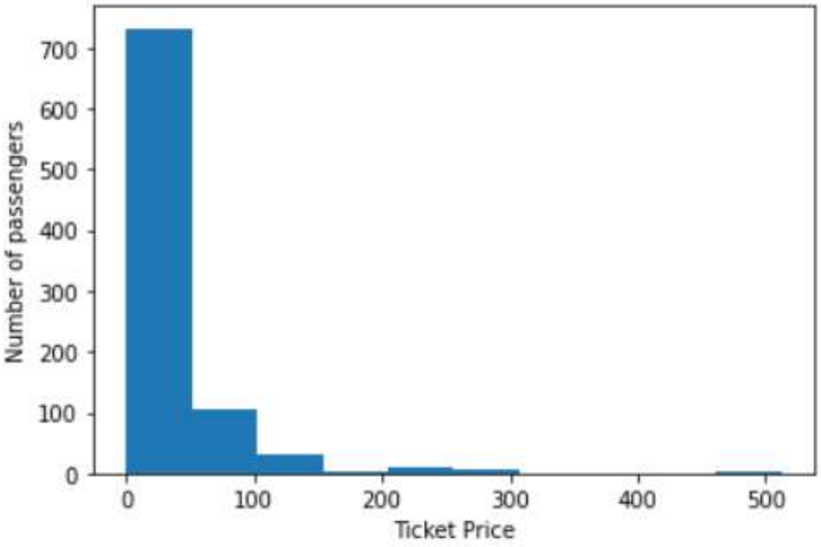
Inference

One can find that the societies in the western countries had a sense of family planning

A histogram is used to summarize discrete or continuous data. The following histogram shows the number of passengersvs total parent/children on board with a passenger.



From the above graph it is clear that around 700 passengers had no chidren .around 100 passengers had 1 child travelling with them. Around 80 to 90 passengers had 2 children travelling with them. Very few passengers had 3 or 4 chidren travelling with them in the ship. Some passengers also had 5 to 6 children with them in the ship.



From the above the graph it is clear that more than 700 passengers had purchased a ticket whose fare was less than 100. Around 100 passengers had purchased a ticket whose fare was about 100. Few passengers (around 25 to 30 passengers) had purchased a ticket whose fare was more than

100 but less than 200. Very few passengers (around 5 to 10 passengers)had purchased a ticket whose fare was more than 200.

Normalization

Normalization is a technique often applied as part of data preparation for machine learning. The point of normalization is to change the observations so that they can be described as a normal distribution.

It changes the values of numeric columns in the dataset to a common scale, without distorting differences in the range of values while also reducing and eliminating data redundancy.Hence we normalized all the numerical columns in the dataset to a mean of zero and variance one.

Hypothesis Testing

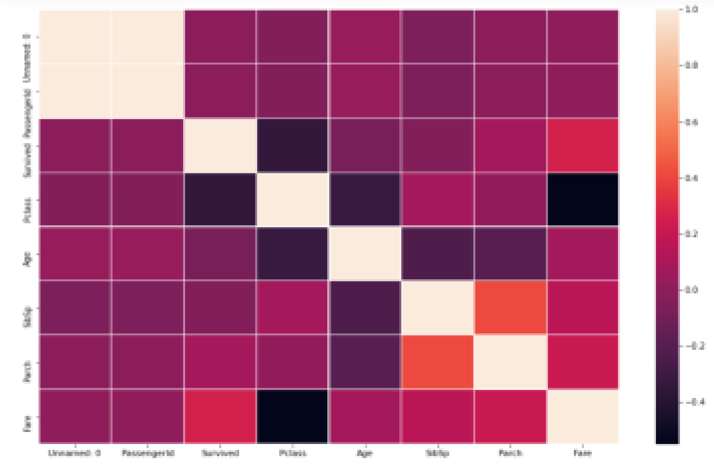
Looking at the dataset, we found that a large number of people was middle aged (25-35) in the sample. So we wanted to research whether the average age of the population in the ship could be 29 or not. We therefore took null hypothesis to imply that mean would be equal to 29 and alternate hypothesis to imply that mean would not be equal to 29. After computing the p-value (0.008) and comparing it with significance level alpha (0.025), it was found that p-value was smaller. Hence this implied that we had to reject the null hypothesis and accept the alternate hypothesis that the average age of people who travelled in the titanic is not 29.

Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).

We used heat map to show the correlation between various columns.

On plotting the heat map we were able to identify a positive correlation between parent children and number of sibling spouse i.e. the values of parent children increases when the number of sibling spouse increases.
We also see that pclass and number of people survived, Age of the people and fare, pclass and fare are negatively correlated.



Results

Hence by performing these various analyses we were able to determine that children and female passengers were given first priority while saving lives when compared male passengers. We could also see negative correlation between many data columns as explained in the correlation part. We also performed a hypothesis whether the average age of the population in the ship could be 29 or not. But from the hypothesis testing we found that the average age of the population in the ship could not be 29.

Individual contribution

Manjunath gowda : data cleaning
Lithesh shetty : visualization of categorical data,standardization and normalization of data
Manideep P R : visualization of numerical data,Correlation
K S Abhisheka : visualization of numerical data,hypothesis testing

CODE PART

```
#DATACLEANING
import pandas as pd
importnumpy as np
importmatplotlib.pyplot as plt
importseaborn as sns

df=pd.read_csv("/Users/manjusgowda/Downloads/titanic.csv")
df
df.head
df['Age'].isnull().sum()
df[‘Pclass'].hist()
x=list(df.groupby('Pclass')['Age'].std())
print(x)
importnumpy as np
mr=list()
mrs=list()
master=list()
miss=list()
x=df['Name']
y=df['Age']
z=list(zip(x,y))
for i in z:
try:
int(i[1])
if 'Mr.' in i[0].split():
mr.append(i[1])


elif 'Master.' in i[0].split():
master.append(i[1])

elif 'Miss.' in i[0].split():
miss.append(i[1])
elif 'Mrs.' in i[0].split():

mrs.append(i[1])
else:
pass
except:
pass
print(len(mr)+len(master)+len(miss)+len(mrs))
print(np.var(mr))
print(np.var(master))
print(np.var(mrs))
print(np.var(miss))

def check(inp):
age=inp[1]

name=inp[0]
print(name)
if(pd.isnull(age)):

if 'Mr.' in i[0].split():
return 32
elif 'Master.' in i[0].split():
return 4.5
elif 'Miss.' in i[0].split():
return 21
elif 'Mrs.' in i[0].split():
return 35
else:pass
else:
return age

dtt=df
dtt['Age']=dtt[['Name','Age']].apply(check,axis=1)
df['Age'].std()
df.isnull().sum()
df.head(20)
y=df[['Pclass','Cabin']]
df['Embarked'].fillna(df['Embarked'].mode()[0],inplace=True)
df.isnull().sum()
df.drop('Cabin',axis=1,inplace=True)
df.isnull().sum().sum()
df.isnull().sum()
def change(y):
if(len(y.split())>1):
returny.split()[1]
else:

return y
df['Ticket']=df['Ticket'].apply(
change)
df['Ticket']
df['Ticket'].replace('LINE',np.nan,inplace=True)
df['Ticket'].fillna(0,inplace=True)
df.isnull().sum().sum()
df.to_csv("/Users/manjusgowda/Desktop/titanicdup.csv")
df.describe()




#DATA VISUALIZATION
df=pd.DataFrame(dataset)
df.Sex.value_counts().plot(kind="pie",autopct='%1.0f%%')
plt.show()

df.Survived.value_counts().plot(kind="pie",autopct='%1.0f%%')
plt.show()

df.Survived.value_counts().plot(kind="bar")
plt.xlabel("Dead(0)   |    Survived(1)")
plt.ylabel("Number of Peoples")
plt.show()

df.Embarked.value_counts().plot(kind="bar")
plt.xlabel("Port Of Embarkation\n\nC = Cherbourg; Q = Queenstown; S = Southampton")
plt.ylabel("Number of Peoples")
plt.show()

df.Pclass.value_counts().plot(kind="bar")
plt.xlabel("Passenger Ticket Class (1st,2nd & 3rd)")
plt.ylabel("Number of Peoples")
plt.show()
plt.show()

sns.set_style("whitegrid")
sns.barplot(x='Pclass', y='Survived', data=dataset)

#scaterplot of fare versus Age
sns.scatterplot(x='Age',y="Fare",data=dataset)

#How gender affected a passengers survival
sns.countplot(x="Survived" ,hue="Sex",data=dataset)
```

```
#the below plot explains how number of passengers survived depended on their ticket class
sns.set_style("whitegrid")
sns.countplot(x="Survived",hue="Pclass",data=dataset)


#Histograms for numerical data

plt.hist(df.Age)
plt.xlabel('Age')
plt.ylabel("Frequency")
plt.title("Age histogram")
plt.hist(df.SibSp,color='green')
plt.xlabel('Number of Sibling spouse')
plt.ylabel('Frequency')
plt.title('SibSp histogram')


plt.hist(df.Parch)
plt.xlabel('total parents/childrens on board with a passenger')
plt.ylabel("Number of passengers")

plt.hist(df.Fare)
plt.xlabel('Ticket Price')
plt.ylabel("Number of passengers")

#Normalization
numerical_cols=['Age','Parch','SibSp','Fare']
# Normalizecontinious data
df_norm = df.copy()
for cols in numerical_cols:
df_norm[cols] = (df_norm[cols] - df_norm[cols].min())/(df_norm[cols].max() - df_norm[cols].min())
df_norm.describe()


# Standardize numerical columns
df_std = df.copy()
for cols in numerical_cols:
df_std[cols] = (df_std[cols] - df_std[cols].mean())/(df_std[cols].std())
df_std.describe()

#Normal plots
```

```
sns.distplot(df_norm['Age'],kde=True)
plt.title("Normal plot of the Age")
plt.figure()

sns.distplot(df_norm['Fare'],kde=True)
plt.title("Normal plot of the Fare")
plt.figure()
```

```
#Hypothesis testing

#H0=The mean age of the people in the ship is equal to 29
H0="The mean age of the people in the ship is equal to 29"
#H1:The mean age of the people in the ship is not equal to 29
H1="The mean age of the people in the ship is not equal to 29 "
population_mean_from_hypothesis1=29
number_of_values=len(df)
Sample_data1=df['Age']
sample_mean1=np.mean(Sample_data1)
sample_sd1=np.std(Sample_data1)
print(sample_sd1)
test="two_tailed_test"
z_score1=(sample_mean1-population_mean_from_hypothesis1)/(sample_sd1/np.sqrt(number_of_values))
print(z_score1)
p_value1=scipy.stats.norm.sf(z_score1)
print(p_value1)
#Checking the test type and assigning the value for alpha
if(test=="two_tailed_test"):
number_of_tails=2
alpha=0.025
elif(test=="one_tailed_test"):
number_of_tails=1
alpha=0.05
#Comparingpvalue with alpha
```

```
if p_value1>alpha:
print("Null hypothesis accepted.\nAccepted hypothesis is:",H0)
else:
print("Null hypothesis rejected and Alternate hypothesis is accepted.\nAccepted hypothesis is:",H1)
```

```
#Corelation
df.corr()
plt.figure(figsize = (15,10))
sns.heatmap(df.corr(),linewidths = 0.1,)
```