Examining Titantic Data

hol_5_1_l

Description

A common data science beginner's challenge is to examine Titantic data to look for patterns in survivability. You are considering pursuing a career as a data scientist so you have decided to look at the Titantic data to get a cursory look at the career.

Below you should review the data and try to answer the questions:

1. What part did age play?
2. What part did gender play?
3. Did the passenger class make a difference?

Titanic Data: Factors Affecting Survivability

This data was collected from a web search. It is available from many different organizations. The data provides specific data about passengers on the Titanic and whether they survived the disaster or not.

The various data available is defined as:

- PassengerId - Indexed starting at 1
- Survived - Survival (0 = No; 1 = Yes)
- Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- Name - Name
- Sex - Sex
- Age - Age
- SibSp - Number of Siblings/Spouses Aboard
- Parch - Number of Parents/Children Aboard
- Ticket - Ticket Number
- Fare - Passenger Fare
- Cabin - Cabin
- Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

The questions we are asking:

1. What part did age play?
2. What part did gender play?
3. Did the passenger class make a difference?

Load the CSV Data Into a Dataframe

```
import matplotlib.pyplot as plt
import pandas as pd

%matplotlib inline

titanic_df = pd.read_csv('titanic.csv')

titanic_df.head()
```

Examine The Affect of Age on Survivability

- Under 12
- 13 - 24
- 25 - 49
- 50 - 74
- 75 and Older

```
#### Under 12
passengers_under_12 = titanic_df[titanic_df.Age < 12]
passengers_under_12_survived =
passengers_under_12[passengers_under_12.Survived == 1]
passengers_under_12_percent_survived =
passengers_under_12_survived.Age.count() / passengers_under_12.Age.count()

# Under 13 - 24
passengers_13_to_24 = titanic_df[(titanic_df.Age >= 13) & (titanic_df.Age <
25)]
passengers_13_to_24_survived =
passengers_13_to_24[passengers_13_to_24.Survived == 1]
passengers_13_to_24_percent_survived =
passengers_13_to_24_survived.Age.count() / passengers_13_to_24.Age.count()

# 25 to 49
passengers_25_to_49 = titanic_df[(titanic_df.Age >= 25) & (titanic_df.Age <
50)]
passengers_25_to_49_survived =
passengers_25_to_49[passengers_25_to_49.Survived == 1]
passengers_25_to_49_percent_survived =
passengers_25_to_49_survived.Age.count() / passengers_25_to_49.Age.count()

# 50 to 74
passengers_50_to_74 = titanic_df[(titanic_df.Age >= 50) & (titanic_df.Age <
74)]
passengers_50_to_74_survived =
passengers_50_to_74[passengers_50_to_74.Survived == 1]
passengers_50_to_74_percent_survived =
```

```
passengers_50_to_74_survived.Age.count() / passengers_50_to_74.Age.count()

# 75 and over
passengers_75_over = titanic_df[titanic_df.Age > 74]
passengers_75_over_survived = passengers_75_over[passengers_75_over.Survived
== 1]
passengers_75_over_percent_survived = passengers_75_over_survived.Age.count()
/ passengers_75_over.Age.count()

print(f'Under 12:\t{passengers_under_12.Age.count()} -
{passengers_under_12_percent_survived}')
print(f'13 - 24:\t{passengers_13_to_24.Age.count()} -
{passengers_13_to_24_percent_survived}')
print(f'25 - 49:\t{passengers_25_to_49.Age.count()} -
{passengers_25_to_49_percent_survived}')
print(f'50 - 74:\t{passengers_50_to_74.Age.count()} -
{passengers_50_to_74_percent_survived}')
print(f'75 & Over:\t{passengers_75_over.Age.count()} -
{passengers_75_over_percent_survived}')
```

```
# Show data as a bar chart
groups = ('Under 12', '13 - 24', '25 - 49', '50 - 74', '75 & Over')
percentages = [0.57, 0.37, 0.41, 0.36, 1]
plt.bar(groups, percentages, align='center', alpha=0.5)
plt.ylabel("Percent Survived")
plt.title("Titanic Survivablity by Age Group")
```

The suggests that children under 13 may have been given some perferential treatment for lifeboats. However, it is not clear if survivability is only those that died in the event. It may be that some of the children may have been more susceptible to environment factors, such as, temperature, and died in the lifeboat.

Since there was only one passenger in the 75 & Over group, the survivability of that group is not useful and should not be considered.

Examine the Affect of Gender on Survivability

```
#### Male
passengers_male = titanic_df[titanic_df.Sex == "male"]
passengers_male_survived = passengers_male[passengers_male.Survived == 1]
passengers_male_percent_survived = passengers_male_survived.Sex.count() /
passengers_male.Sex.count()

#### Female
passengers_female = titanic_df[titanic_df.Sex == "female"]
```

```
passengers_female_survived = passengers_female[passengers_female.Survived ==
1]
passengers_female_percent_survived = passengers_female_survived.Sex.count() /
passengers_female.Sex.count()

print(f'Male:\t{passengers_male.Sex.count()} -
{passengers_male_percent_survived}')
print(f'Female:\t{passengers_female.Sex.count()} -
{passengers_female_percent_survived}')
```

```
# Show data as a bar chart
groups = ('Male', 'Female')
percentages = [0.18, 0.74]
plt.bar(groups, percentages, align='center', alpha=0.5)
plt.ylabel("Percent Survived")
plt.title("Titanic Survivablity by Gender")
```

It is obvious female passengers were given preference over male passengers for lifeboats. It would be interesting to break down the male survivors by age group. Hypothesis: Younger males survived at a great rate.

Examine the Affect of Passenger Class on Survivability

```
#### Passenger Class 1
passengers_class_1 = titanic_df[titanic_df.Pclass == 1]
passengers_class_1_survived = passengers_class_1[passengers_class_1.Survived
== 1]
passengers_class_1_percent_survived =
passengers_class_1_survived.Pclass.count() /
passengers_class_1.Pclass.count()

#### Passenger Class 2
passengers_class_2 = titanic_df[titanic_df.Pclass == 2]
passengers_class_2_survived = passengers_class_2[passengers_class_2.Survived
== 1]
passengers_class_2_percent_survived =
passengers_class_2_survived.Pclass.count() /
passengers_class_2.Pclass.count()

#### Passenger Class 3
passengers_class_3 = titanic_df[titanic_df.Pclass == 3]
passengers_class_3_survived = passengers_class_3[passengers_class_3.Survived
== 1]
passengers_class_3_percent_survived =
passengers_class_3_survived.Pclass.count() /
passengers_class_3.Pclass.count()
```

```
print(f'Class 1:\t{passengers_class_1.Pclass.count()} -
{passengers_class_1_percent_survived}')
print(f'Class 2:\t{passengers_class_2.Pclass.count()} -
{passengers_class_2_percent_survived}')
print(f'Class 3:\t{passengers_class_3.Pclass.count()} -
{passengers_class_3_percent_survived}')
```

```
# Show data as a bar chart
groups = ('Class 1', 'Class 2', 'Class 3')
percentages = [0.63, 0.47, 0.24]
plt.bar(groups, percentages, align='center', alpha=0.5)
plt.ylabel("Percent Survived")
plt.title("Titanic Survivablity by Passenger Class")
```

It is clear that Class 1 passengers were more likely to be saved, whether they were closer to the lifeboats or a genuine perference cannot be deteremined. One again looking at this data by age and gender would be interesting for furtehr study.

This is not an exhaustive review of the data available, but simple review based on three attributes treating them as independent. Much more data could be analyzed for deeper more specific ideas of how the surviving passengers were selected.