Analyzing Titanic Survival Rates Using Big Data and Machine Learning

Contents

Intro

We are going to use the dataset of TitanicData to evaluate the survival rate of a person depending on their details. The data consists of 891 records, each a single passenger. 891 **1. (History, 2009)** "of the 2240 passengers and crew on board", so only 39.7% of the entire titanic population. Each records holds the persons PassengerId, Pclass, Name, Sex, Age, Number of siblings / spouses on board (SibSp), number of parents / children on board (Parch), Ticket number, Fare and Survived. We will look for patterns and correlations between the data, more specifically what do all who survived have in common. I will be using Kaggle platform, along with libraries to create visual diagrams and numerical statistics to analyze the Titanic data.

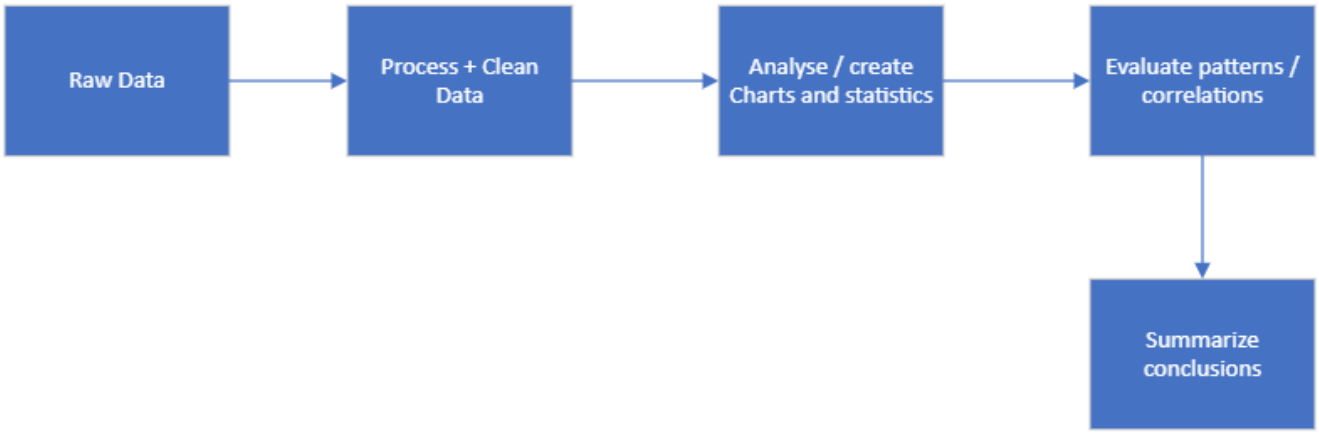
If we are to analyze this data, we must first abstract the relevant data for survival. We can use data sources to determine which data to focus on. **2. (SHIFT, No Data)** Stated " Women had a much higher chance of survival — regardless of what class they were in — then men did". The titanic was famous for initiating the Woman and Children First protocol. With this, we can determine the data of Sex to be relevant.

Then check this data is 'clean'. Replace missing data and outliers. We will use Graphs (such as a boxplot) to see any outliers that could decrease the accuracy of our results.

Once the data is processed, we can now create statistics and charts to locate and visualise correlaction and patterns.

Approach

This will be the summary of my Approach



Data Cleaning

We import the raw data, named Titanic using pandas library

Titanic = pd.read_csv('/kaggle/input/titanicdata/TitanicData (1).csv')

First, the data needs to be cleaned and pre-processed. We need to remove / alter any empty data values. So our results later can be more accurate.

Titanic_Dropnullcolumns = Titanic.dropna(axis = 1). This new Dataset has 9 columns instead of 12. Columns Age, Embarked and Cabin have null values. So we have 3 options, either replace the nulls in the columns, remove the column or if the null percentage is so low it can be ignored. . This depends on the importance of the columns. I think the most relevant columns to use is Age, Sex, Pclass, fare and survived. So for the Age column, I replaced all null values with the mean of the entire column. Creating a dataset for the 'cleaned' data

Titanic_clean = Titanic

Titanic_clean['Age'] = Titanic['Age'].fillna(Titanic['Age'].mean())

For column Cabin and Embarked, this is not a relevant for survival rate. So if null value percentage is too high, bes to just remove from cleaned data. (Code below labelled percent_missing)

percent_missing = Titanic_clean.isnull().sum() * 100 / len(Titanic_clean)

print('percantage of null in entire dataset')

print(percent_missing)

Output:

PassengerId 0.000000
Pclass 0.000000
Name 0.000000
Sex 0.000000
Age 0.000000
SibSp 0.000000
Parch 0.000000
Ticket 0.000000
Fare 0.000000
Cabin 77.104377
Embarked 0.224467
Survived 0.000000

Remove Cabin column

Titanic_clean = Titanic_clean.drop('Cabin', axis = 1)

Embarked is not a priority column, but with a low null percentage, we will keep and alter null value with mean value of column. Column in string value. So find quantity if each value in column, and find mean by converting to int

print(Titanic_clean['Embarked'].value_counts())

Output: S 644
C 168
Q 77

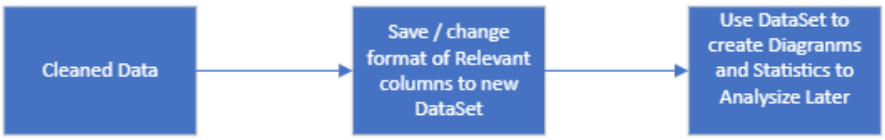
S = 1, C = 2. Q = 3 (1 * 644) + (2 * 168) + (3 * 77) = 1211 / (644 + 168 + 77) = 1.36 (round to 1 so null = S)

Titanic_clean['Embarked'] = Titanic['Embarked'].fillna('S')

Percent_Missing code ran again, embarked now 0%. Titanic_clean dataset now got no nullls, data has been 'cleaned'.

Data Pre-Process

First is to Decide which data to use from the Titanic_clean dataset. **3. (Mental Floss, 2022) "**He reportedly told his first and second officers to "put the women and children in and lower away". Along with **2.** we can deduce column Age and Sex are relevant. **4. (BBC, No Data)** Pclass was the class that person was. **4.** "First class passengers were some of the richest and most important people of the time. "While **4.** "Second class passengers were tourists" And **4. "**Third class were mainly immigrants**"**. Pclass may have influenced who was prioritized onto the lifeboat, so This column will be classed a probably relevant. Fare will come hand in hand with Pclass, Fare = Amount paid for ticket onto titanic.



The last column to be relevant is Parch. **5 (*Titanic—Not Enough Lifeboats, 2022)*** *States about the lifeboats, "*Even if all had been filled to capacity, only half the people would have been saved." This puts into question would a woman with children be prioritized over a woman with no children? Would the age of the child matter? For this reason, parch is a relevant column

Most Relevant Columns: Age, Sex, Survived

Moderately relevant Columns: Pclass, Parch, Fare

Least relevant Columns: PassengerId, Name, SibSp, Ticket

Now to process data to use it. We will first put the 5 relevant columns into a new dataset: Titanic_Relevant

Titanic_Relevant = Titanic_clean[['Age', 'Sex', 'Fare', 'Pclass', 'Survived','Parch']]

Will begin with using them, all in int data type instead of Sex. We will map current data to int version Male = 0, Female = 1

Titanic_Relevant ['Sex'] = Titanic_Relevant ['Sex']. map({'male':0,'female':1})

Titanic_Relevant Dataset prepared to be used for results

Predictions

Confident: Females will have a significantly higher survival rate than Males.

Fairly sure: People with Pclass of 1 will have higher survival rate with Pclass 2 and 3.

Unsure: Will a Female of Pclass 3 have a higher chance of survival than a Male of Pclass 1?

Unsure: Will elderly people have a higher survival rate?

Fairly sure: Higher percentage of Pclass 3 survived than Pclass 2 and 1.

Unsure: what had a bigger impact of survival, Age or Pclass

Confident: Pclass and fare will have a very strong positive correlation.

Unsure: Will people with children have a higher survival rate?

Confident: Children (16 and younger) have an incredibly high survival rate.

Results Analysis

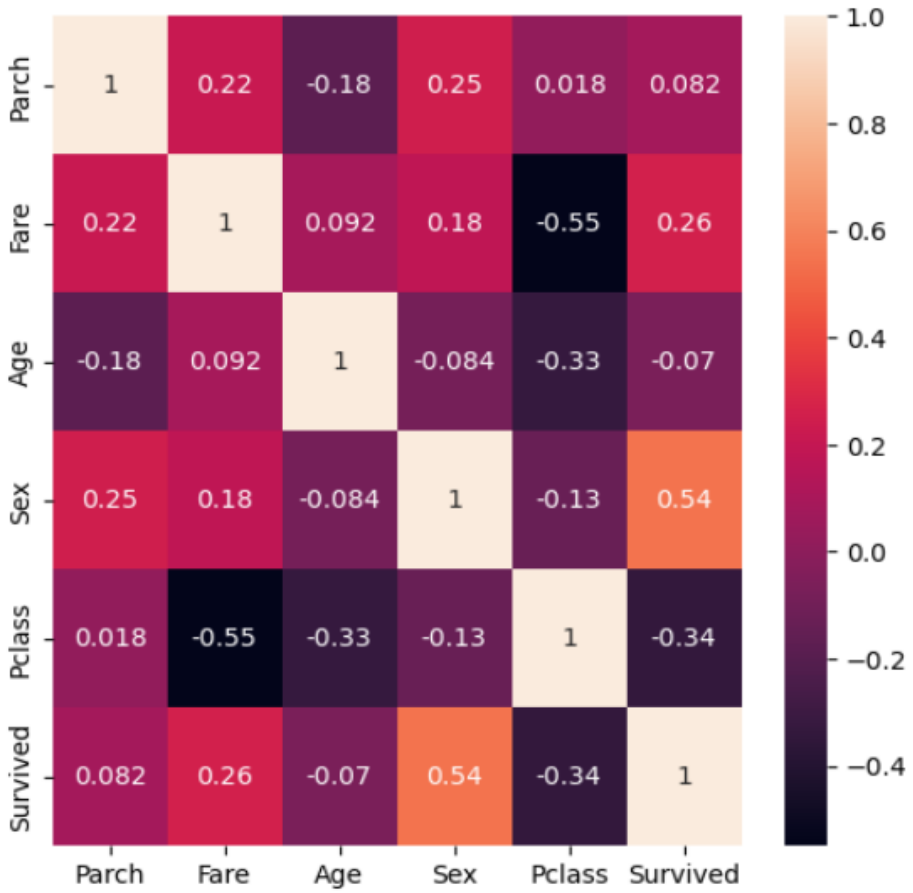
Relevant Diagrams to visualize Data

Using a HeatMap with DataSet Titanic_Relevant

fig = plt.figure(figsize=(6,6))

sns.heatmap(data=Titanic_int_diagrams.corr(), annot = True)

Output: (Labelled HeatMap1.0)



Here are the 1:1 correlation between The columns Fare, Age, Sex, Pclass and Survived. This is a good diagram to show where the strongest and most relevant correlations are. We can use this as a basis to explore more complex patterns that are between more than 2 columns. Strong correlations are Fare – Pclass, Age – Pclass, Sex – Survived, Pclass – Survived,

The two strongest correlations for Survived (the primary column) is with Pclass and Sex. So we will create diagrams with these three more specifically

Here is a bar plot consisting of those three columns.

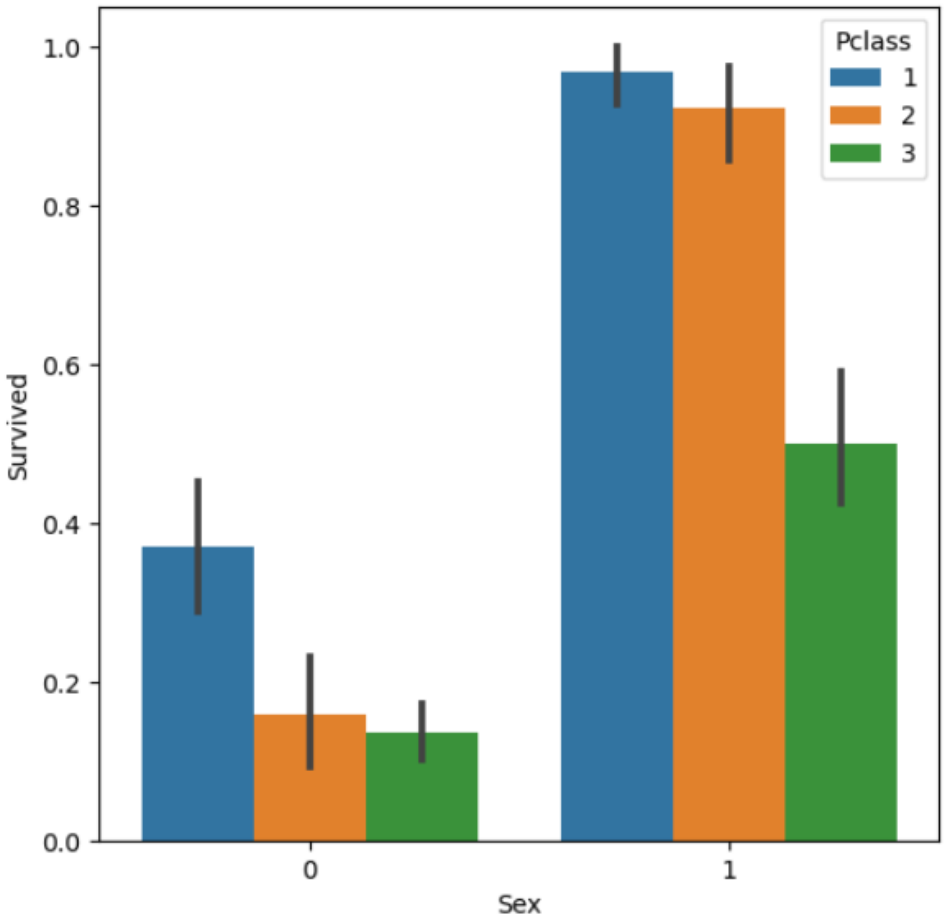
plt.figure(figsize=(6,6))

sns.barplot(x="Sex", y="Survived", hue="Pclass", data=Titanic_Relevant)

plt.show()

Output: (Labelled Bar plot 1.0)

Here we see the difference between the number of men and women who survived grouped by their class. This shows an overwhelming Survivability rate of women (especially in class 1 and 2) compared to all classes of men. In here we see that more 3rd Class Females survived than 1st Class Males, giving evidence that Sex was a stronger factor to survivability



rate that Pclass. Although when it came to men, it seemed that a significant more 1st class Males survived than 2nd and 3rd class Males.

When it comes to Males, we should see the survivability rate between the classes

Titanic_Relevant[['Survived', 'Sex', 'Pclass']].groupby(['Pclass','Sex']).mean()

Output(Labelled GroupBy 1.0)

		Survived
Pclass	**Sex**	
1	**0**	0.368852
	1	0.968085
2	**0**	0.157407
	1	0.921053
3	**0**	0.135447
	1	0.500000

Here we see the percentage of survival for each sex in each class. 36% of 1st class Males survived, with 15% of 2nd class Males and 13% 3rd class males. Seems when it came to 2nd and 3rd class males, it made a merge 2% difference of chances of survival. But if you were 1st class Male, it would more than double your chance up to 36%, still low relative to Females, but significant nonetheless

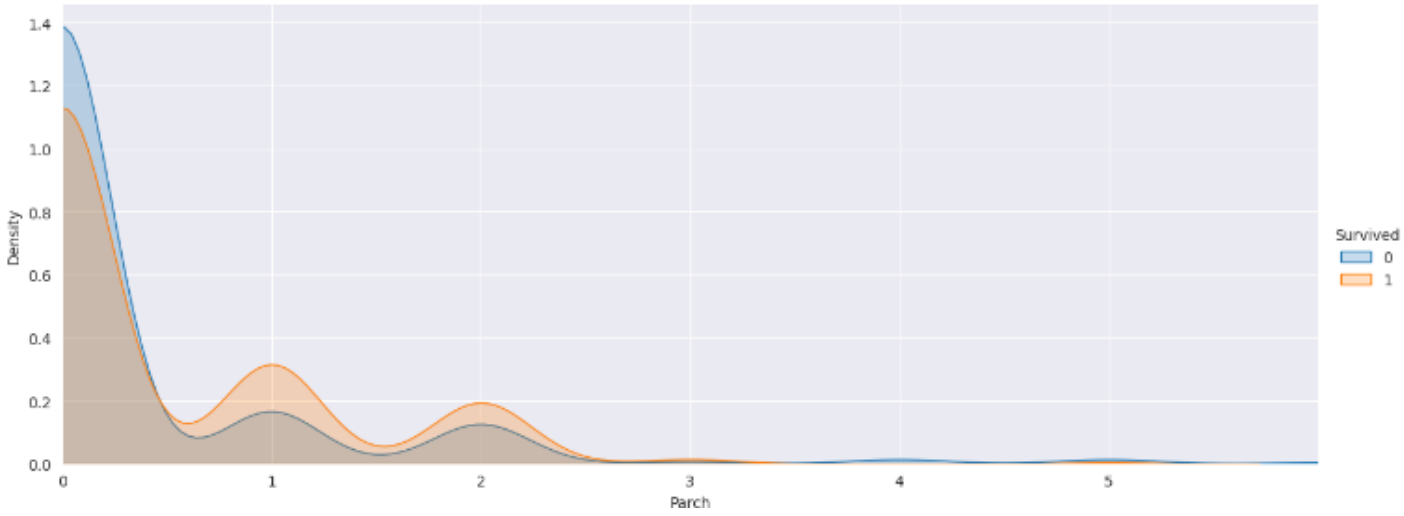
When the Sexes are split into Pclass, it seems the Pclass comes into a bigger role for survivability and creates 2 groups. For Females, class 1 and 2 had an incredibly high survivability rate, while 3rd class women had a relatively low chance of only 50%. But when it came to Males, the 'high survival' group is only class 1, while group 2 is class 2 and 3.

Now we will see the effect of a persons Parch (Number of parents / children aboard) to their survivability rate.

I used a swarmplot to see the Parch number of who survived, to who didn't. This is with all records (Male and Female)

```
with sns.axes_style("darkgrid"):
__g = sns.FacetGrid(Titanic_Relevant, hue='Survived', height=5, aspect=2.5)
__g.map(sns.kdeplot, 'Parch', shade=True)
__g.add_legend()
__g.set(xticks=np.arange(0, Titanic_Relevant['Parch'].max()), xlim=(0, Titanic_Relevant['Parch'].max()))
```

Output (Labelled kdeplot 1.0)

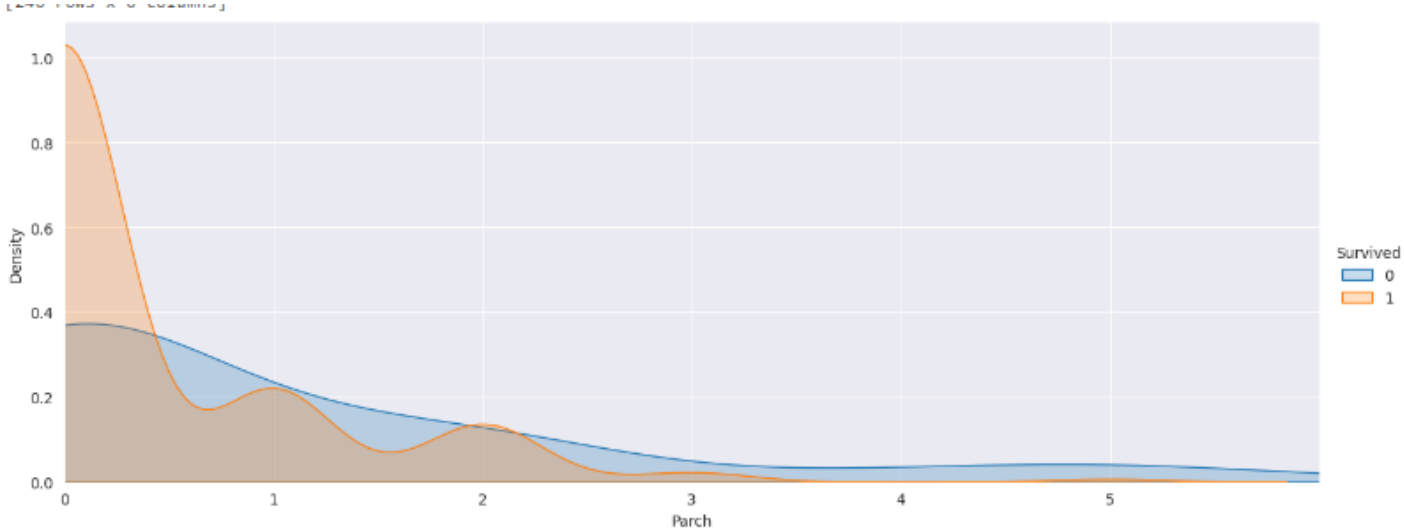


Here we see the density of those who survived / did not survive compared to their Parch number. Seems the most common parch number was 0, followed by 1 then 2. Seems parch number influenced who survived, we can see more people survived when they had 1 or 2 children, while the majority with a parch number of 0 did not survive.

However, Due to the Women and children first protocol issued, we should remove all male records. We know women and children were prioritized over men. But we want to know if people with a higher parch were prioritized. More specifically, women. Adult females (older than 18), so we will filter out the male records and children records into a new dataset.

```
Titanic_Temp = Titanic_Relevant[Titanic_Relevant.Sex == 1][Titanic_Relevant.Age > 18]

print(Titanic_Temp)

with sns.axes_style("darkgrid"):

__g = sns.FacetGrid(Titanic_Temp, hue='Survived', height=5, aspect=2.5)

__g.map(sns.kdeplot, 'Parch', shade=True)

__g.add_legend()

__g.set(xticks=np.arange(0, Titanic_Temp['Parch'].max()), xlim=(0, Titanic_Temp['Parch'].max()))
```

Output(Labelled kdeplot 2.0)

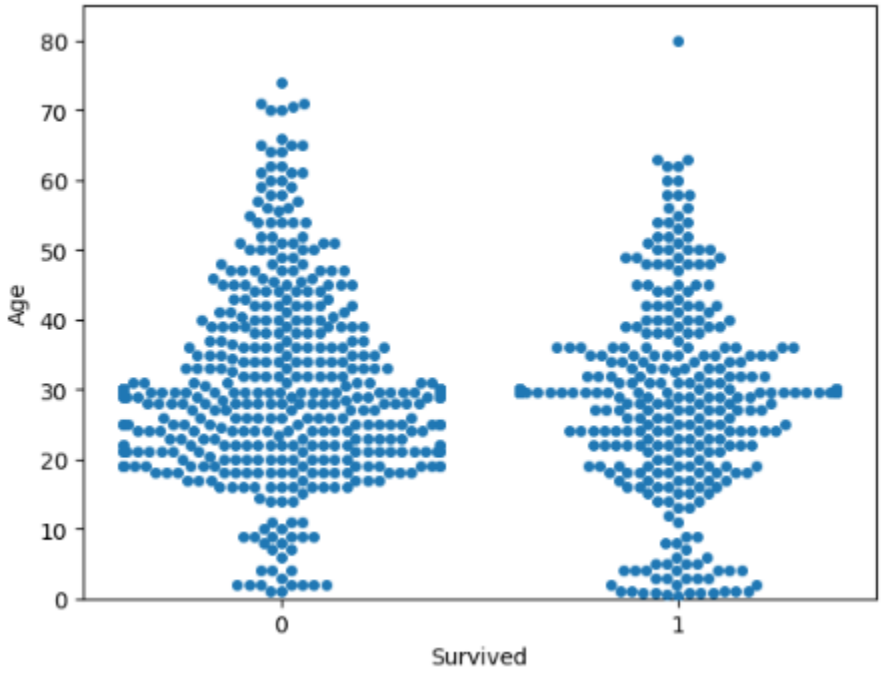


This is females over 18 who survived grouped by their Parch number. Seems the same number of women who had 1 or 2 parch number died to those who survived. Furthermore, more women with 3,4 or 5 parch number died than survived. And the majority who had 0 parch survived. This is clear evidence that if a women had a larger parch number, their chances of survival did not increase. Meaning women who had children were NOT priorities to women with children. Parch is below Sex on effect of survival rate.

Another potential Data value that could alter survival chances is age. First we should see the overall Range of ages that survived / didn't survive. We will use a swarmplot. Due to the number of data points. We will go back to using Titanic_Relevant dataset

```
plt.axis([1, 3, 0, 85])

sns.swarmplot(x = 'Survived', y = 'Age', data = Titanic_Relevant)
```

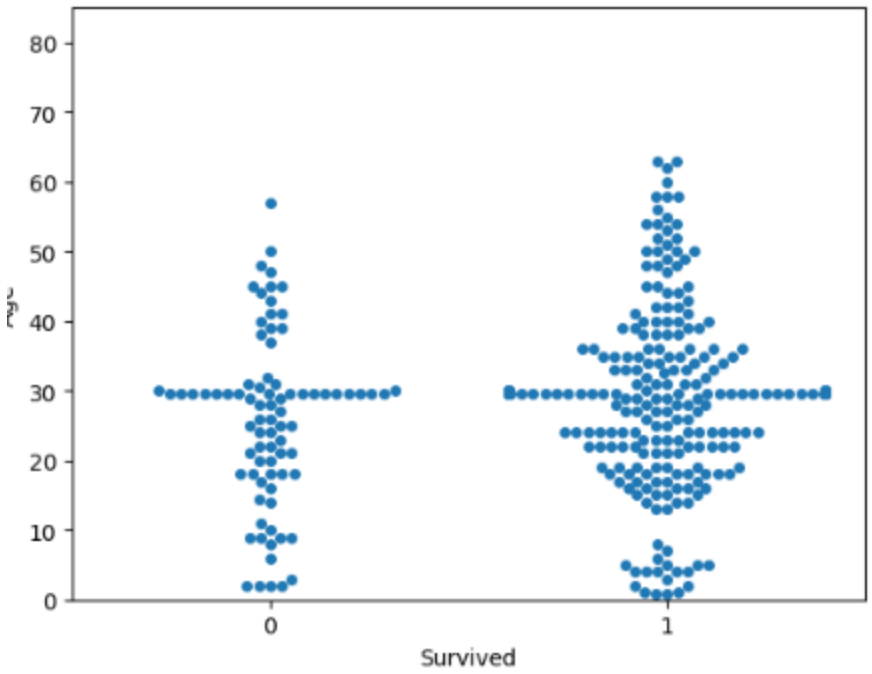
Output (swarmplot 1.0)



Here we see the ages of all people, grouped into who survived and who did not. Does not seem to be a strict pattern. A vague difference is the below 10s, seems a higher number under 10 survived than did not. So young children were prioritized to an extent. Seems an incredible high rate of people between 18 – 30 did not survive, but this may be because that age was the most common among the population. To see clearer patterns, the data should be split up into Sex categories.

```
plt.axis([1, 3, 0, 85])

sns.swarmplot(x = 'Survived', y = 'Age',
data =
Titanic_Relevant[Titanic_Relevant.Sex ==
1])
```

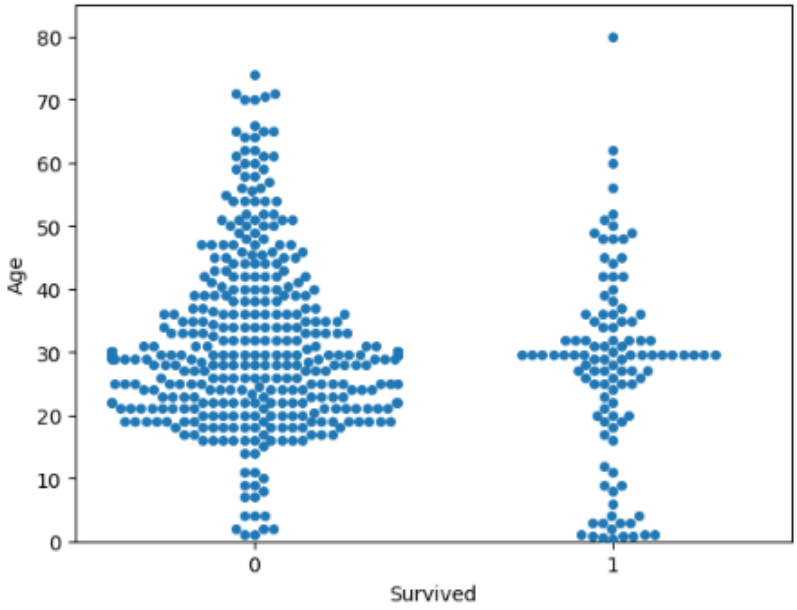
Output (swarmplot 2.0)



For women, seems a decent resemblance of Swarmplot 1.0. The most populated area at around 30 years of age, a decent number of older people above 40 did not survive. However, unlike Swarmplot 1.0. It seems majority of all ages did survive. But that may be due to the records being females.

```
plt.axis([1, 3, 0, 85])

sns.swarmplot(x = 'Survived', y = 'Age',
data =
Titanic_Relevant[Titanic_Relevant.Sex ==
0])
```

Output (swarmplot 3.0)



For all male records, like we saw on Swarmplot 1 and 2, a pattern of a high density of children (age under 10) surviving. Other than that, the diagram follows the pattern of Swarmplot 1.0. Still, we see a high number (and even the majority) of older males (above 50) not surviving.

However, swarmplot 1.0,2.0 and 3.0 does show that they did not prioritize Female children over Male children. That is one group that their survival chances were not ruled by Sex category.

Conclusion

Overall, we can decisively claim that the strongest single data value that could accurately determine if you survived or not was your sex category. However, the 2nd most influential category was Pclass, but this was still heavily influence by your sex, Example, 1st class women had a 96% survival rate, while men only had 36%. Though low, still a major improvement from 2nd class males (15%). Seems Parch had a weak influence, when around half of women with children did not survive.

Age also had a weak influence, only relevant with children (around under 10) but not when it came to more older people (around above 40). However, when it came to children, seams their Sex became irrelevant.

Swarm graphs were good for this due to the number of records (891). Confusion matrix was a good starting point and showed the strongest correlations of 2 Data columns.

Some issues found were formatting the data. Example, I thought about swapping the Pclass system so 3rd was 1st class and vice versa. So, if females = 1, we would see a positive correlation if there were 3rd class (1st class). But this seemed to complicate things unnecessarily.

Suggestions for future work: Use point plots, investigate embarked column, see if patterns of survivability.

References

1. History (2009) Titanic. Available At: https://www.history.com/topics/early-20th-century-us/titanic [Accessed 21 April 2023]
2. SHIFT (No Data) Data-Driven Survival; Titanic Edition. Available At: https://www.shiftcomm.com/insights/never-let-go-titanic-survival-101/#:~:text=Gender%20%26%20Class&text=Turns%20out%2C%20the%20blockbuster%20hit,women%20on%20board%2C%20339%20survived. [Accessed 22 April 2023]
3. Mental Floss (2022) Women and Children First: How a Rarely-Used Maritime Practice Caused Confusion On Board the 'Titanic'. Available At: https://www.mentalfloss.com/posts/women-and-children-first-origins-titanic#:~:text=In%20fact%2C%20in%20the%20decades,maritime%20disasters%20is%20a%20myth. [Accessed 22 April 2023]
4. BBC Bitesize (No Data) What was life like on Board Titanic?. Available at: bbc.co.uk/bitesize/topics/z8mpfg8/articles/zkg9dxs [Accessed 22 April 2023]
5. NonFinction Minute (2022) Titanic – Not Enough LifeBoats. Available At: https://www.nonfictionminute.org/the-nonfiction-minute/titanic-not-enough-lifeboats [Accesses 24 April 2023]