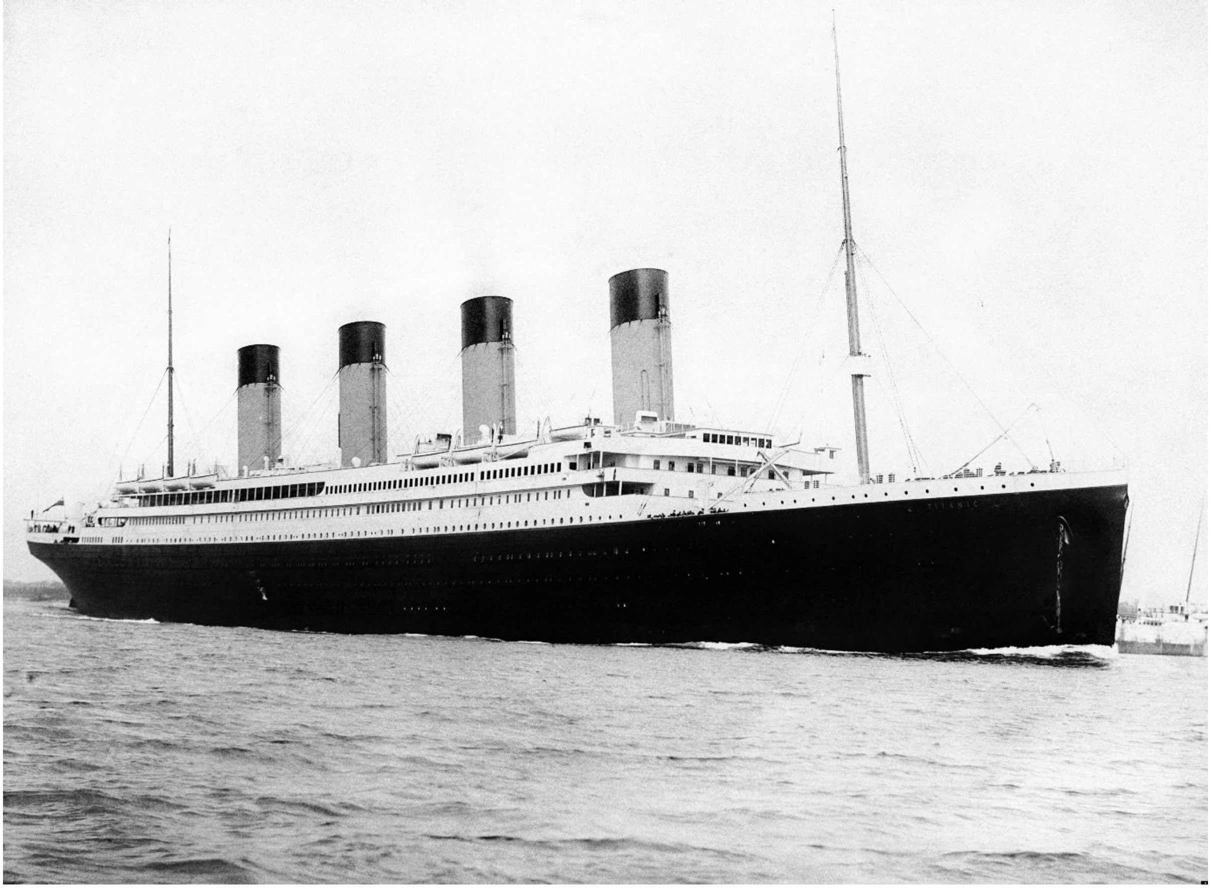
Titanic – Tutorial



Almost everyone knows the story of the Titanic. In April 1912, this magnificent ship left Southampton on its maiden voyage to New York but it never arrived. It hit an iceberg in the Atlantic and sank. There were over 2,000 people on board. Less than half survived.

A century later, this Titanic dataset is a classic case study for rookie data scientist to build a predictive model to determine who is likely to survive or perish (ignoring the fact that this is a matter of historical record). However, we will visualise the data with Power BI and see if we can gain some intuition and who did and did not survive and why. We know from the film that Kate Winslet survived but poor old Leo DiCaprio did not – is that an accurate reflection?

This dataset contains a list of 891 of the passengers on board including variables (columns) such as Name, Age, Sex, and Pclass i.e. whether they travelled 1st, 2nd or 3rd class.

We need to clean the data. Some variables have missing values. The names of the variables, are cryptic and so are the values. We'll start by making the variable names and values easier to understand.

Dataset

Titanic.xlsx

Have a look at the dataset in Excel. It has two tabs. The passengers tab contains the data. The Comments tab explains the variables – this is useful as the variable names are difficult to understand.

<u>Load Data</u>

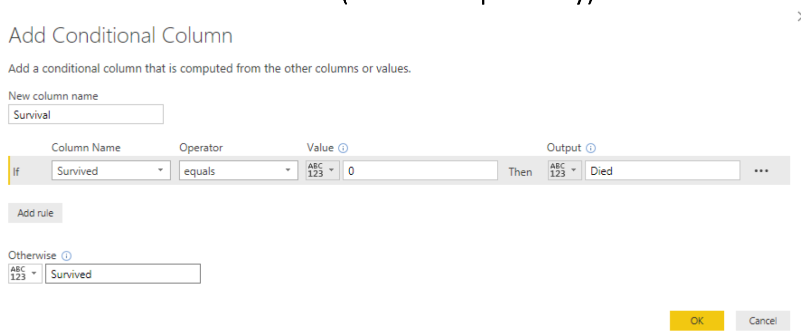
Load the data in the Passengers tab of the Titanic workbook into the Query Editor. (Get Data -> Excel -> Select the Titanic workbook -> Edit)

<u>Transform Data in the Query Editor</u>

Remove the Ticket and Cabin columns *(we don't need them in this exercise).*

Rename the Sex column to Gender.

The Survived column has two values 0 and 1 to indicate whether the passenger died or survived. These values are not intuitive so let's correct that.
Add a conditional column. Name it Survival. It will have two values Died or Survived based on the value of the Survived column (0 and 1 respectively).



Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name: Survival

If Column Name: Survived, Operator: equals, Value: 0, Then Output: Died

Otherwise: Survived

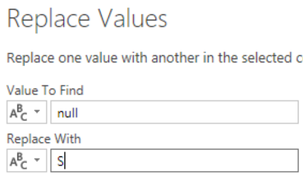
Remove the Survived column – we no longer need it.

The Pclass column has values 1, 2 and 3. Perhaps integer values are not best in this case – is a 2nd class passenger somehow twice as much as a 1st class?
Replace Pclass with a new column PassengerClass with values 1st, 2nd and 3rd. Hint: use Conditional Column again.

Change the datatype of both PassengerClass and Survival from Any to Text.

Add a custom column, FamilySize, with formula = [SibSp]+[Parch]+1. Change the datatype to whole number.
This in data science terms is called by the fancy name of 'feature engineering'. FamilySize may be a useful variable in determining survival chances.

In the Embarked Column, replace any blank values with "S". Use Replace Values like this.



Replace Values

Replace one value with another in the selected columns.

Value To Find: null

Replace With: S

Most people embarked at Southampton– but don't take my word for it - check this for yourself. If we were to build a predictive model, (which we are not going to do, at least in this session) then removing missing values is important since models don't get on well with them.

In the Embarked column, replace S, C and Q values with Southampton, Cherbourg and Queenstown respectively.
The Titanic left Southampton on the English south coast, went across the Channel to pick up some

passengers in Cherbourg, France then across the Irish Sea to pick up more people at Queenstown before setting on its journey across the Atlantic.

Close & Apply the query editor

The Titanic is an unusual dataset in that it does not have any measures (quantities). Typically these are values such as Sales, Amount, Exposure… Every visual needs a measure so let's create one.

Create a new measure. Use the formula

```
PassengerCount = COUNTROWS('Passengers')
```

<u>Explore the data</u>

Create a new page and name it 'Explore

Add the following four bar charts. All have PassengerCount on the Values. On the axis they have

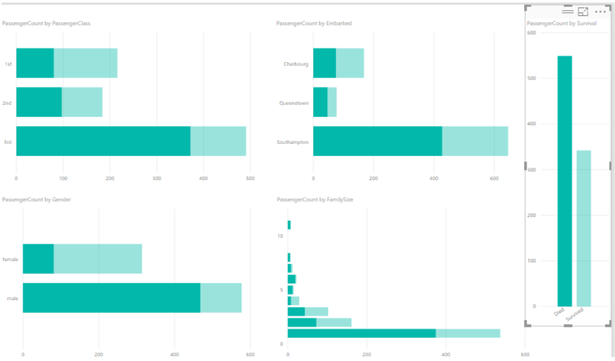
- PassengerClass
- Gender
- Embarked
- FamilySize

Add a stacked column chart of Passenger Count (Values) vs Survival

You may want to change the colours of the Survived and Died values.

You can see the distribution of these four variables and the cross highlighting gives some idea of the interplay between them as in the snapshot below.

You can also filter on several bars across charts now – for example have a look at the chances of survival of women in first class.



<u>Examine how gender and passenger class affect survival chances</u>

Create a new page. Name it 'Survival Analysis.

Add a matrix and show Passenger Count (Values) by Survival (Columns), Gender and Passenger Class (Rows).

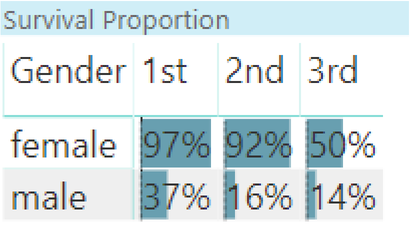
Calculate a new measure as below

Survival Ratio =

```
DIVIDE(
    CALCULATE([Passenger Count], Passengers[Survival] = "Survived"),
    [Passenger Count]
)
```

This gives the proportion of the passengers in each cell that survived.

Plot this against gender and passenger class



Survival Proportion

Gender	1st	2nd	3rd
female	97%	92%	50%
male	37%	16%	14%

<u>The curious case of the women in 3rd class</u>

Look at Survival of women in 3rd class against the fare they paid.

It is helpful here to bin the Fare field values into bins on width £10. You can choose to do this in two ways

a) In the Query Editor create a conditional column with categories 0-10, 10-20, 20-30 and 30-above.
b) In Power BI Desktop, select the Fare field and select New Group. Since this is a numeric filed, you will get a dialog select the bin size.

Which option is better?

Use a 100% stacked bar chart to see the proportions more easily. What does the data show that is surprising given the insights so far? Any possible theories why this might be the case?

<u>Analyse the distribution of the fare paid by passenger class and gender</u>

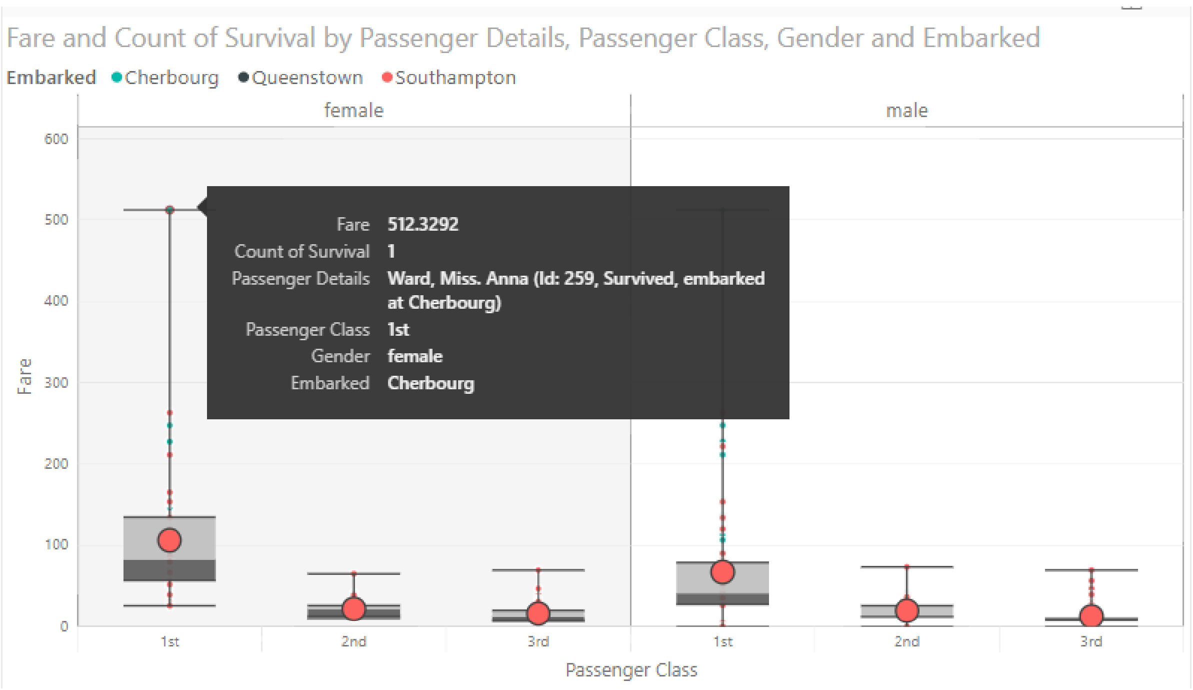
Import a custom visual, the box and whisker chart by MAQ software

Use this custom visual so see the distribution of fares by factors such as passenger class, gender and embarkation.

You may want to create a calculated column named Passenger Details with a formula similar to this.

```
Passenger Details =
  Passengers[Name]
  & " (Id: " & Passengers[Passenger Id]
  & ", " & Passengers[Survival]
  & ", embarked at " & Passengers[Embarked] & ")"
```

If you add this to the Axis of the box whisker chart (rather than Passenger Id), this makes the details on the tooltip when looking at outliers more useful.



Fare 512.3292
Count of Survival 1
Passenger Details Ward, Miss. Anna (Id: 259, Survived, embarked at Cherbourg)
Passenger Class 1st
Gender female
Embarked Cherbourg

Fare and Count of Survival by Passenger Details, Passenger Class, Gender and Embarked
Embarked ● Cherbourg ● Queenstown ● Southampton
female
male
600
500
400
300
200
100
0
Fare
1st
2nd
3rd
1st
2nd
3rd
Passenger Class

Analyse the passenger age distribution

The default summarisation of the Age field is set to “Sum”. Is this the best choice? If not change it.

Does the dataset record the age of all passengers? Create a calculated column with a formula like this.

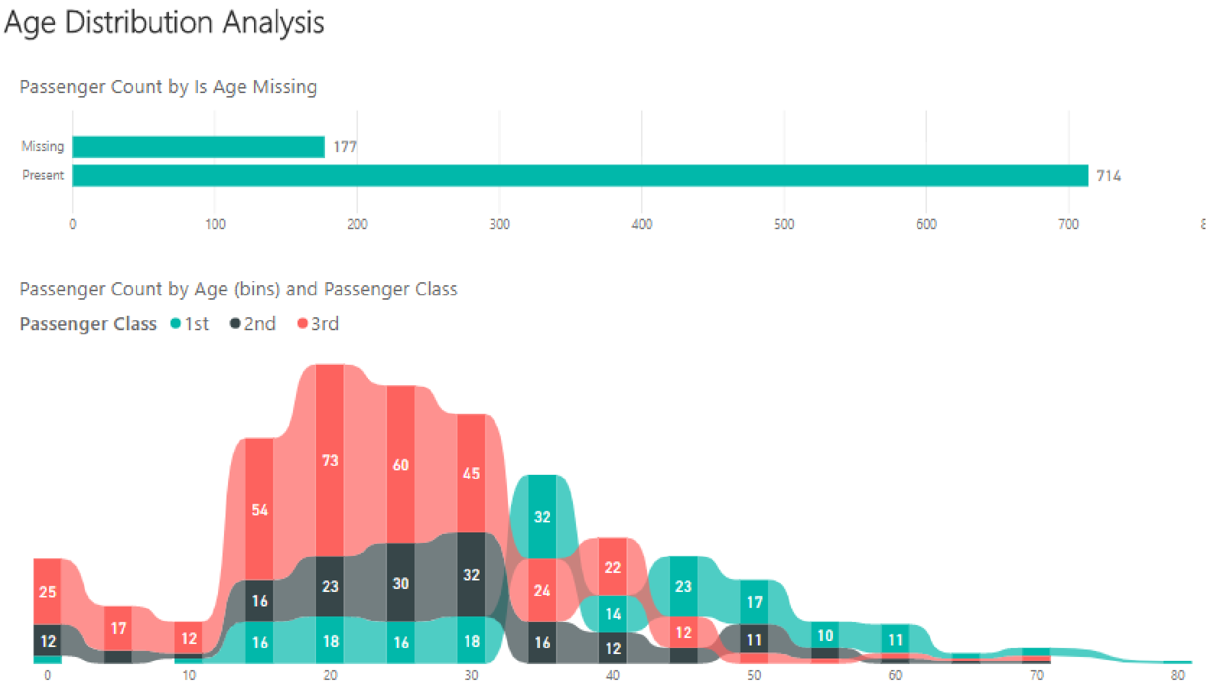
```
Is Age Missing = if(ISBLANK(Passengers[Age]), "Missing", "Present")
```

Bin the passenger ages into groups of size 5 years. To start this, select the Age field and select New Group from the context menu.

Use a bar chart so analyse the age distribution of the passengers. Place the ‘Age (bins)’ field in the Axis well and the ‘Passenger Count’ field in the Values well, and the Passenge Class filed in the Legend well.

Morph the bar chart into a ribbon chart. Is that helpful?

Your page may look something like this



Age Distribution Analysis

Passenger Count by Is Age Missing

Passenger Count by Age (bins) and Passenger Class