Table of Contents

EDA - Titanic Dataset

Data Modelling

The dataset has been structured using a Star Schema model consisting of one fact table and three-dimension tables. This structure allows for efficient querying and data analysis.

A. Dimension Tables:

1. **Dim Passenger Class:** This table holds information about passenger classes, with columns **'pclass' and 'pclass_type'**. The **'pclass' is used as a primary key** and is populated from raw data with no null values. The 'pclass_type' is a description of the passenger class, created using metadata.

2. **Dim Location:** This table stores information about embarkation ports. It consists of columns **'embarked' and 'location'**, both of which are populated using raw data and metadata with no null values. The **'embarked' column is used as a primary key**.

3. **Dim Passenger X Journey:** This is the most complex dimension table detailing each passenger's journey. It includes data such as 'age', 'body', 'boat', 'cabin', 'home.dest', 'name', 'sex', 'survival_status', 'survived', and 'ticket'. This table contains **null values in several columns**, specifically 'age', 'body', 'boat', 'cabin', and 'home.dest'.

   An additional column **'age_groups'** categorises the ages into 'child' (0-12), teenager (13-18), 'adult' (20-59), and senior (>=60) derived from 'age'. The **'passenger_journey_id', a concatenated field of 'ticket' and 'name', is used as the primary key**.

B. Fact Table

The Fact Table represents factual data about each passenger's journey, such as **'fare', 'parch', 'sibps', and 'Family Member'**. It connects to the dimension tables via the foreign keys **'pclass', 'embarked', and 'passenger_journey_id'**. The table is mostly populated **except for two rows with missing 'embarked' information**. The 'Family Member' column, which is the sum of 'sibps' and 'parch', represents the number of family members for each passenger.

C. Relationships

The Fact Table is related to 'Dim Passenger Class' and 'Dim Location' via a **one-to-many relationship using 'pclass' and 'embarked'**, respectively. 'Dim Passenger X Journey' has a one-to-one relationship with the Fact Table using 'passenger_journey_id'.

D. Derived Columns

Several derived columns have been introduced to enhance the analytical capability of the data model.

1. **Dim Passenger X Journey - 'age(bins)':** categorises ages into bins of 10, aiding in data analysis by age ranges.

2. **Fact Table**

A. **'Traveler Type'** differentiates passengers into 'Solo' and 'Family'.

B. **Percentage distribution columns like '% of Total (Gender)', '% of total Generic', '% Survive', '% Did not Survive', '% of Total' and '% of total (Traveler Type)'** give an understanding of proportional representations.

C. **'Total Passenger'** indicates the total number of passengers on the ship, used for calculating proportions and percentages.

D. **'Survival Rate'** calculates the survival rate of the passengers.

E. **'Family Member Type'** categorises family members into 'Both', 'Siblings/Spouse', and 'Parents/Children'.

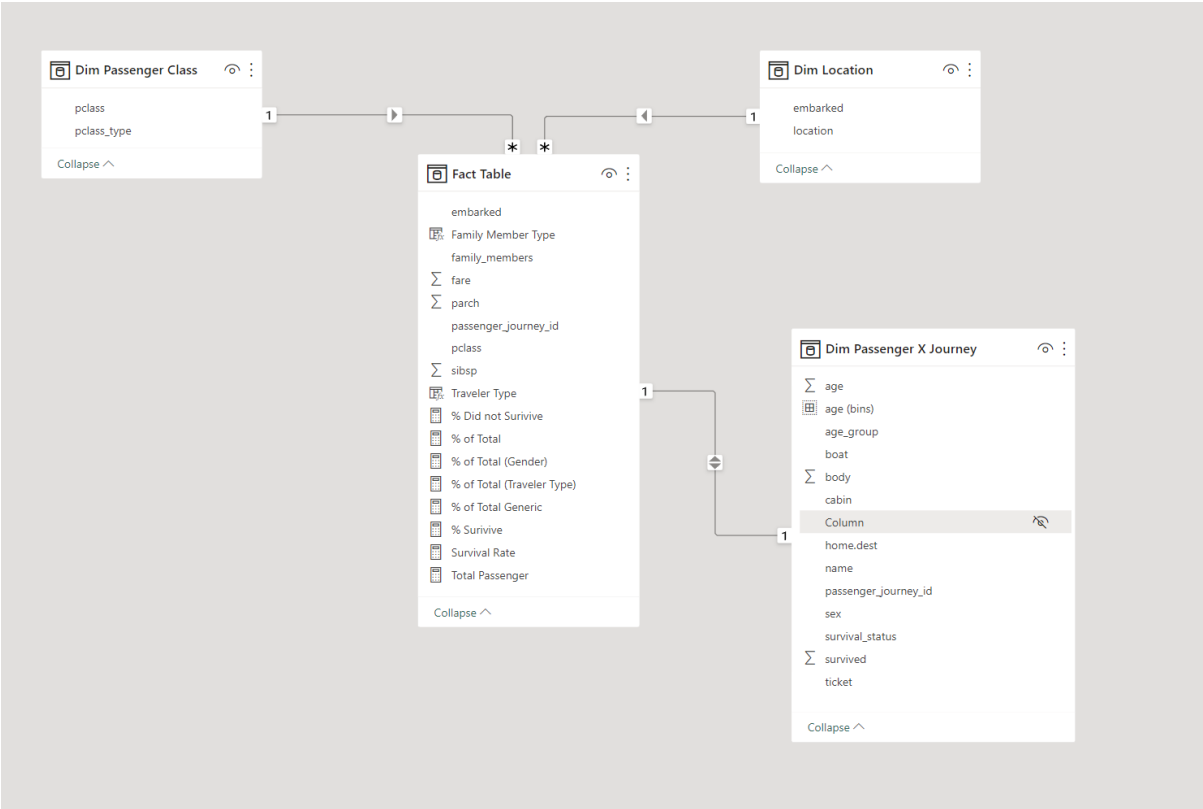

Figure 1: Star Schema

Data Characterisation

The dataset consists of a variety of data types. Identifiers like 'pclass' and status indicators such as 'survived' are whole numbers. Text entries include names, ticket information, descriptions, and categories. Decimal numbers represent continuous data such as 'age' and 'fare'. The range, central tendency, and spread of these continuous data need further exploration.

Missing Information

In **'Dim Passenger X Journey'**, there are 139 missing entries in **'age'**, representing 14% of the data. The **'body' and 'boat'** columns have 90% and 57% missing data, indicating a lack of information on body recovery and lifeboat boarding. These columns context dependent. The **'cabin' and 'home.dest'** fields have 71% and 26% missing data, respectively.

S.No	Dimension Tables	Columns	Description	Data Type	Missing Information
1	Dim Passenger Class	pclass	Passenger Class	Whole Number	No Null
		pclass_type	Passenger Class Description	Text	No Null
2	Dim Location	embarked	Port of Embarkation	Text	No Null
		location	Port of Embarkation Description	Text	No Null
3	Dim Passenger X Journey	passenger_journey_id	concat of ticket & name	Text	No Null
		age	Age of the passenger	Decimal Number	139 rows (14%)
		age(bins)	Grouping Ages (bins=10)	Decimal Number	No Null
		age_groups	Grouping age into category (like adult, child, etc)	Text	139 rows (14%)
		body	Body Number	Whole Number	In general 905 rows (90%) (Context-dependent)
		boat	Lifeboat if Survived	Text	In general 583 rows (57%) (Context-dependent)
		cabin	Cabin	Text	717 rows (71%)
		home.dist	Home/Destination of the passenger	Text	258 rows (26%)
		name	Name of the passenger	Text	No Null
		sex	Gender of the passenger	Text	No Null
		survival_status	Survival Status Description	Text	No Null
		survived	Survival Status	Whole Number	No Null
		ticket	Ticket No. of the passenger	Text	No Null

S.No	Fact Table	Columns	Description	Data Type	Missing Information
1	Fact Table	pclass	Passenger Class	Whole Number	No Null
		embarked	Port of Embarkation	Text	2 rows (<1%)
		passenger_journey_id	concat of ticket & name	Text	No Null
		fare	Ticket Fare	Decimal Number	No Null
		parch	parents or children	Whole Number	No Null
		sibps	siblings or spouse	Whole Number	No Null
		Family Member	sum of sibps and parch	Whole Number	No Null

Primary Key
Foreign Key

Figure 2: Data characterisation & Missing information

Data Visualisations & Insights

The Titanic dataset comprising records of 1,309 passengers. The demographic composition shows a significant gender disparity, with males constituting 64.4% (843 passengers) and females making up 35.6% (466 passengers) of the total populace. An interesting aspect of the journey is the travelling companionship. Most of the passengers, precisely 60.4% (790 passengers), embarked on the voyage solo, perhaps seeking adventure or fortune. In contrast, 39.6% (519 passengers) opted for this transatlantic journey in the company of their relatives - be it parents, children, siblings, or spouses.

Figure 3: Topline Stats

1. **Survival Analysis by Gender**

   Out of 843 male passengers, 596 were solo travelers and 292 traveled with family. Only 161 males (19%) survived; 97 were solo and 64 travelled with family. Out of 466 female passengers, 339 (73%) survived, 142 of which were solo travellers and 197 travelled with family members. **This suggests females had a significantly higher survival rate whether they are traveling solo or with family.**

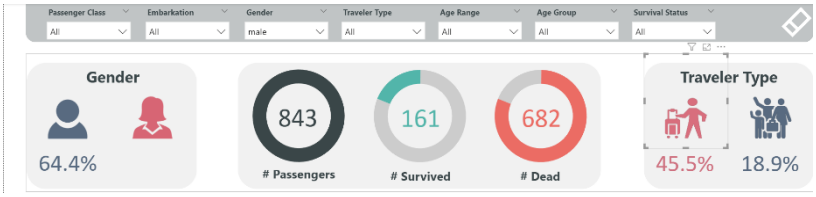

Figure 4: Male Stats

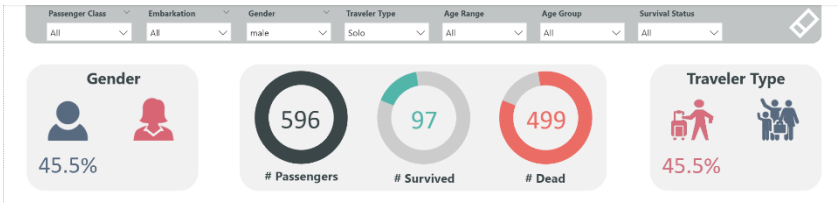

Figure 5: Male & Solo Traveler Stats

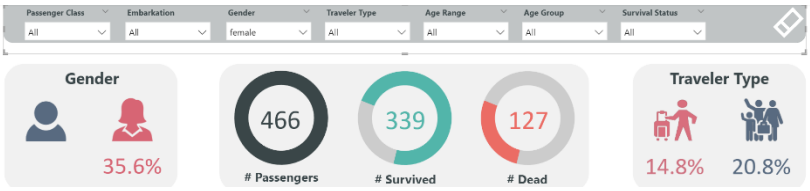

Figure 6: Female Stats

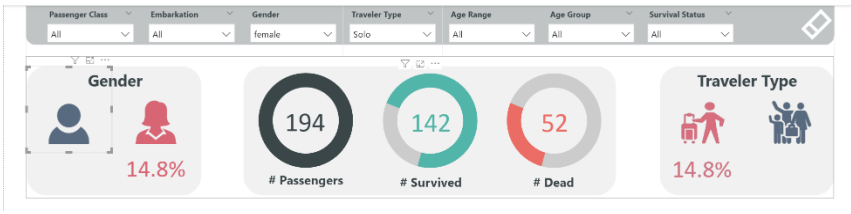

Figure 7: Female & Solo Traveler Stats

2. Survival Analysis by Passenger Class

A. **Class Impact:** Survival odds were highest for 1st class passengers, reflecting the influence of socio-economic status on survival.
B. **Travel Companions:** Survival rates were higher for passengers with families, especially in 1st class, compared to solo travellers, predominantly in 3rd class.
C. **Gender Priority:** Irrespective of their class, females had higher survival rates, indicating a rescue priority for women.

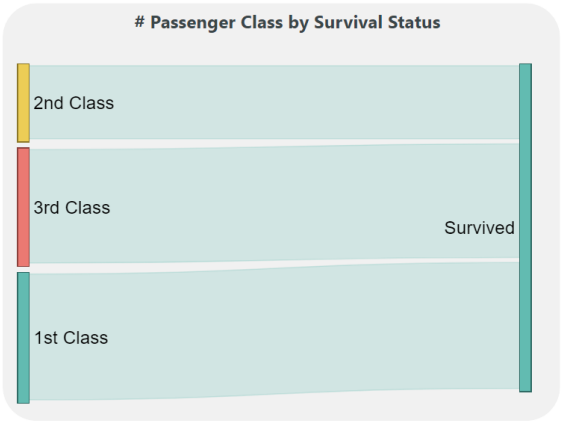

Figure 8: Class Impact On Survival Chances

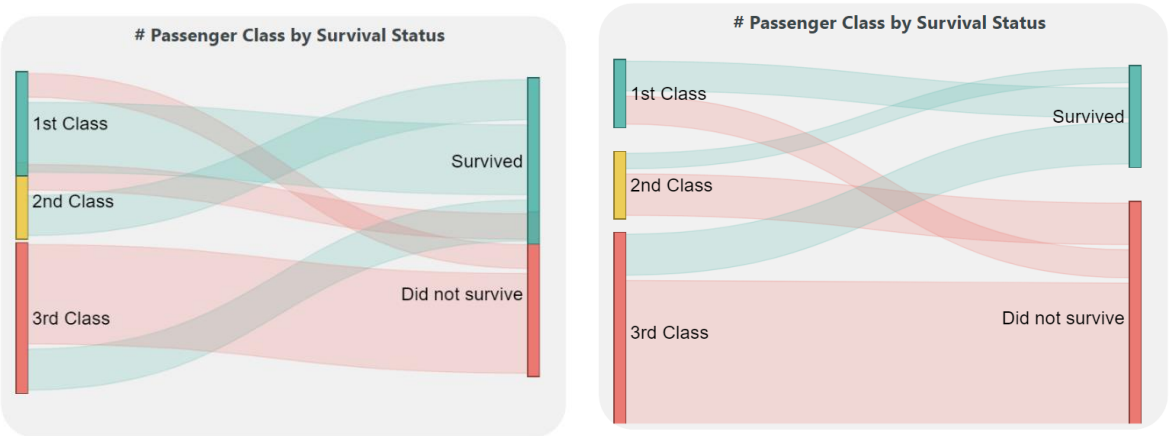

Figure 9: Travel Companions – Frist class higher Survival (with family)(right). Third class least Survival (Solo Traveler)(left)

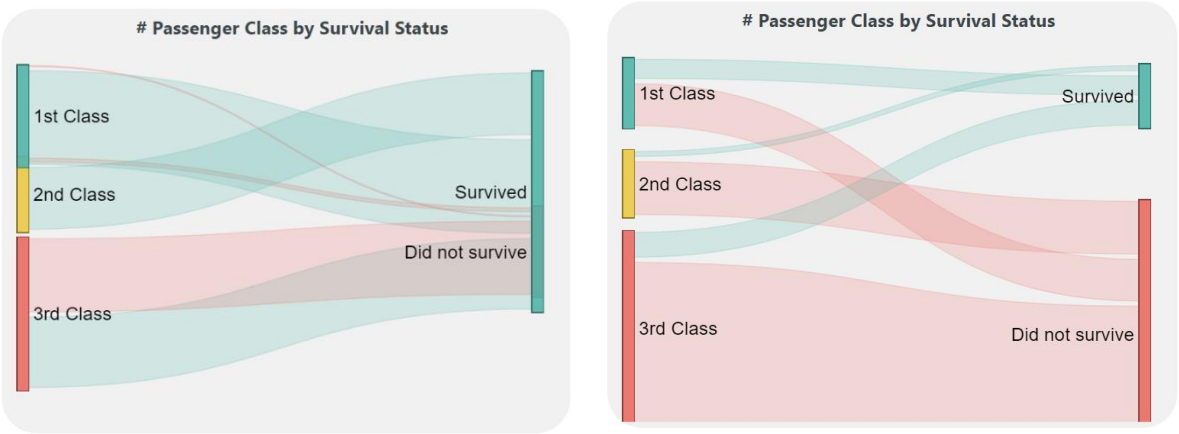

Figure 10: Gender Priority – Females (left) irrespective of class survived the most as compared to Men (right)

3. Survival Analysis by Port

Among 500 survivors, 304 (61%) embarked from Southampton, 150 (30%) from Cherbourg, and 44 (9%) from Queenstown. However, Cherbourg had the highest survival rate (56%) followed by Queenstown (34%) and Southampton (33%), showing that survival rates varied by embarkation location.

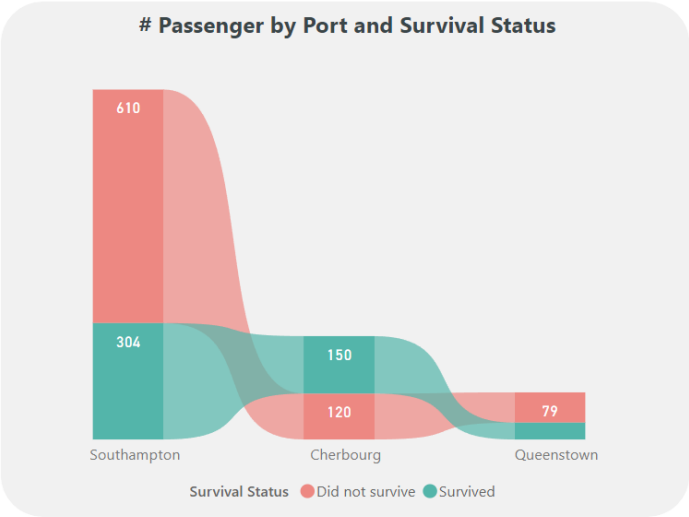

Figure 11: Survival Chances by Embarkation.

4. **Survival Analysis by Age Group**
   Children (0-12 years) had the highest survival rate (54.7%), while seniors (>60 years) had the lowest (30%), indicating that age was a significant factor in survival.

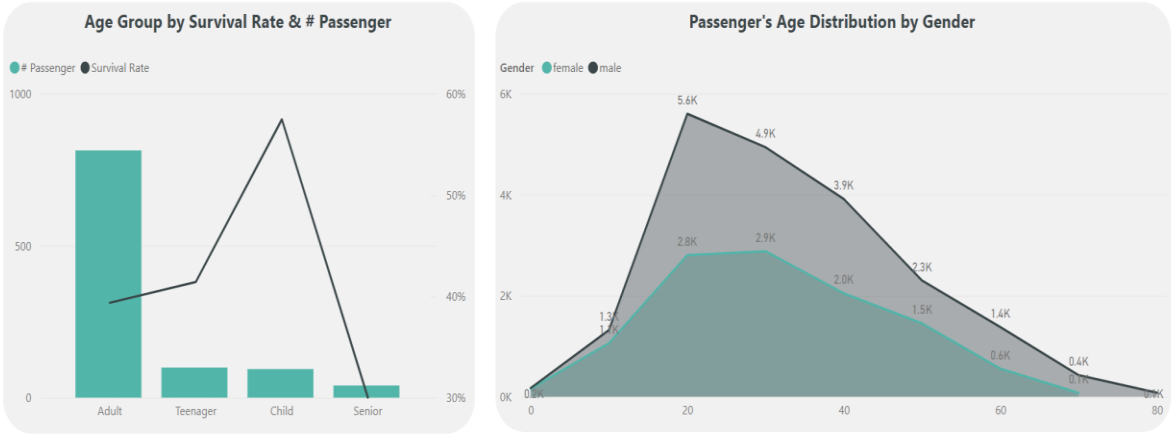

Figure 12: Children had the highest survival rate (left). Passenger's age distribution (right).

5. **Survival Analysis by Class-Port**
   Most passengers who embarked from Southampton belonged to the 3rd class, with a significant proportion of adults. This group had the lowest survival rate, demonstrating a combined effect of class and embarkation location on survival chances.

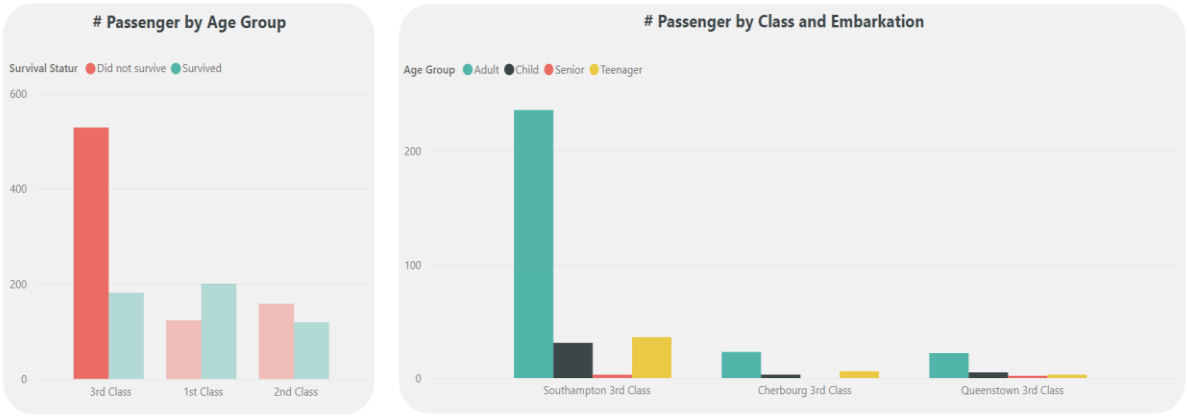

Figure 13: The probability of survival for adults hailing from Southampton and belonging to the 3rd class was notably low.

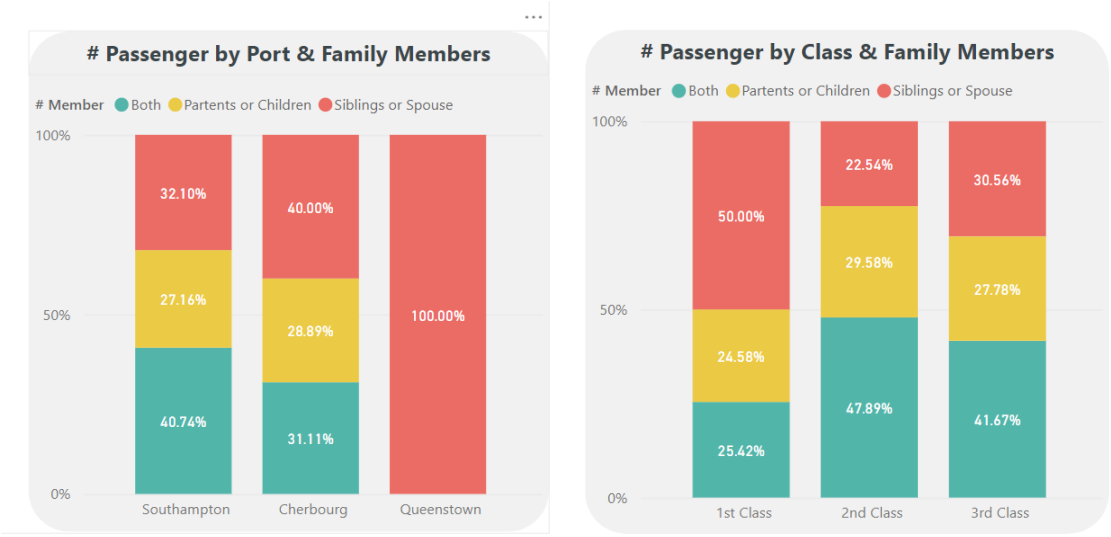

Figure 14: Passengers with parents and children had a higher survival rate when grouped by embarkation(right). In contrast, spouses or siblings had higher survival rates when grouped by class (left). Note: % on the bar reflects % of total passengers, not actual survival rates.

Conclusion

Analysis suggests that females, 1st class passengers, those who embarked from Cherbourg, and children had the highest survival rates on the Titanic. The findings underscore the impact of gender, passenger class, embarkation location, and age on survival chances during the tragedy.

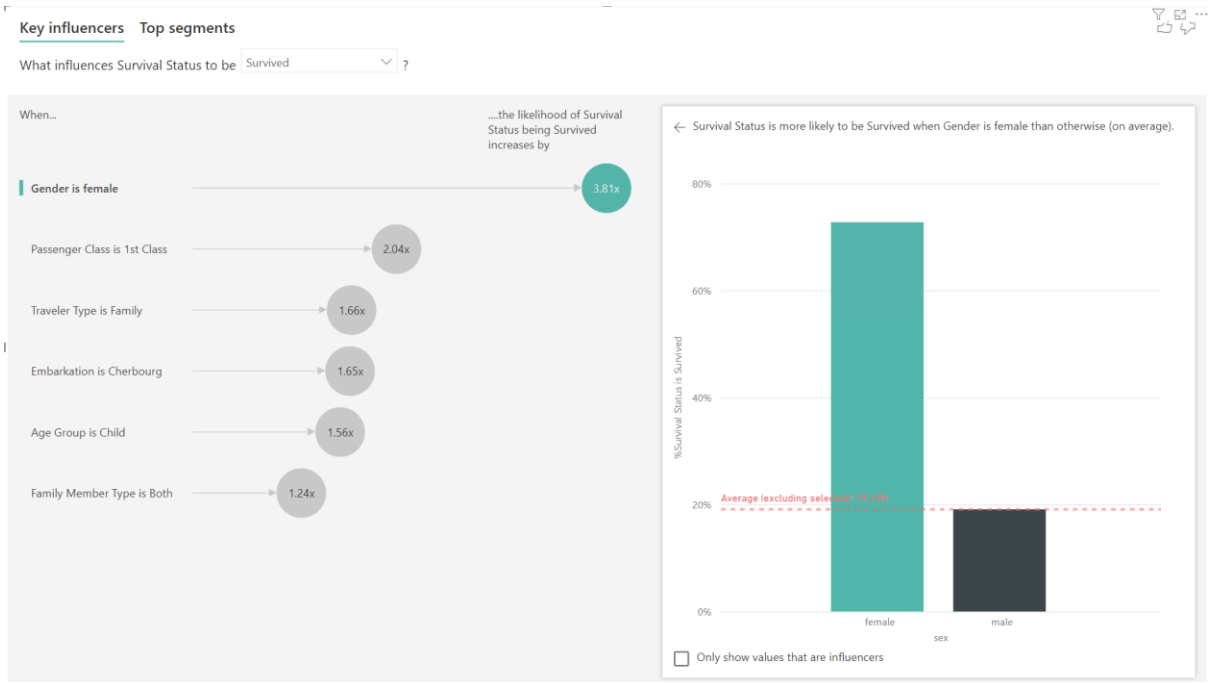

Figure 15: Key Influencers

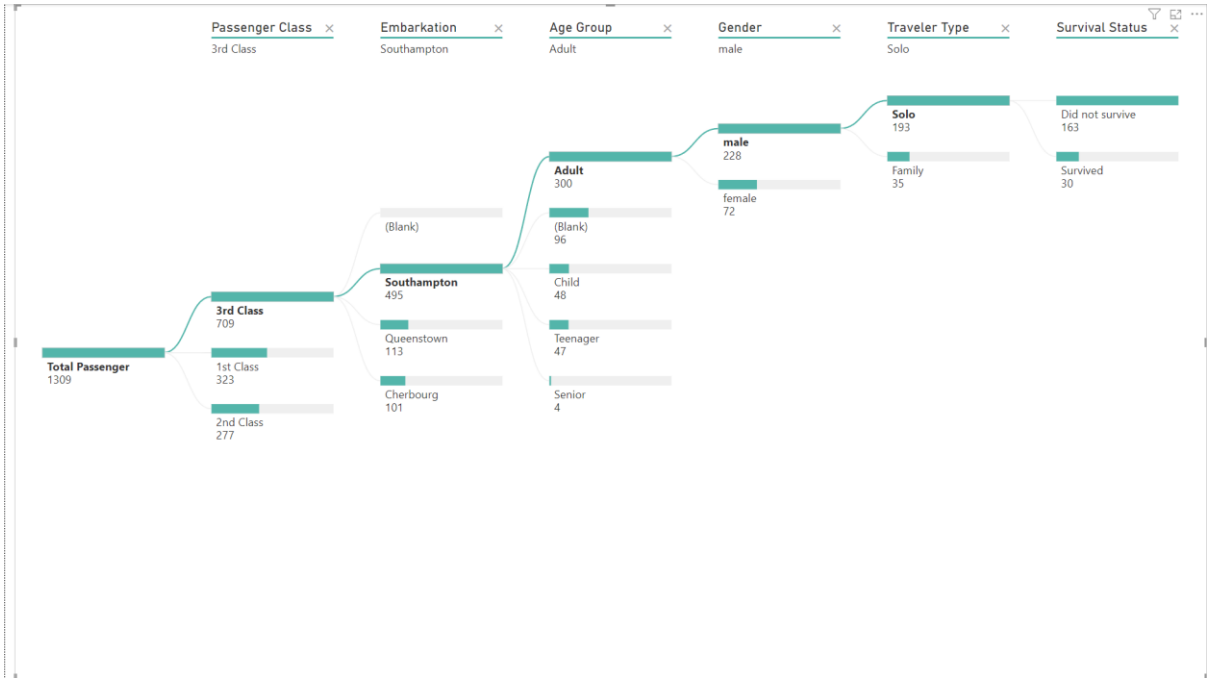

Figure 16: Decomposition Tree