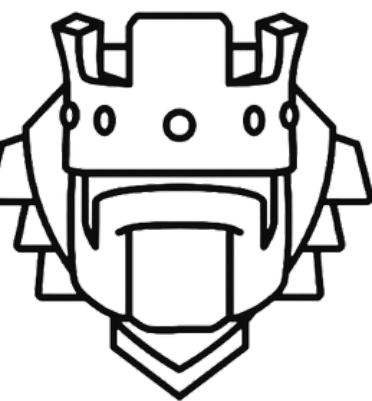


DECONSTRUCTING *DISASTROUS* DATA



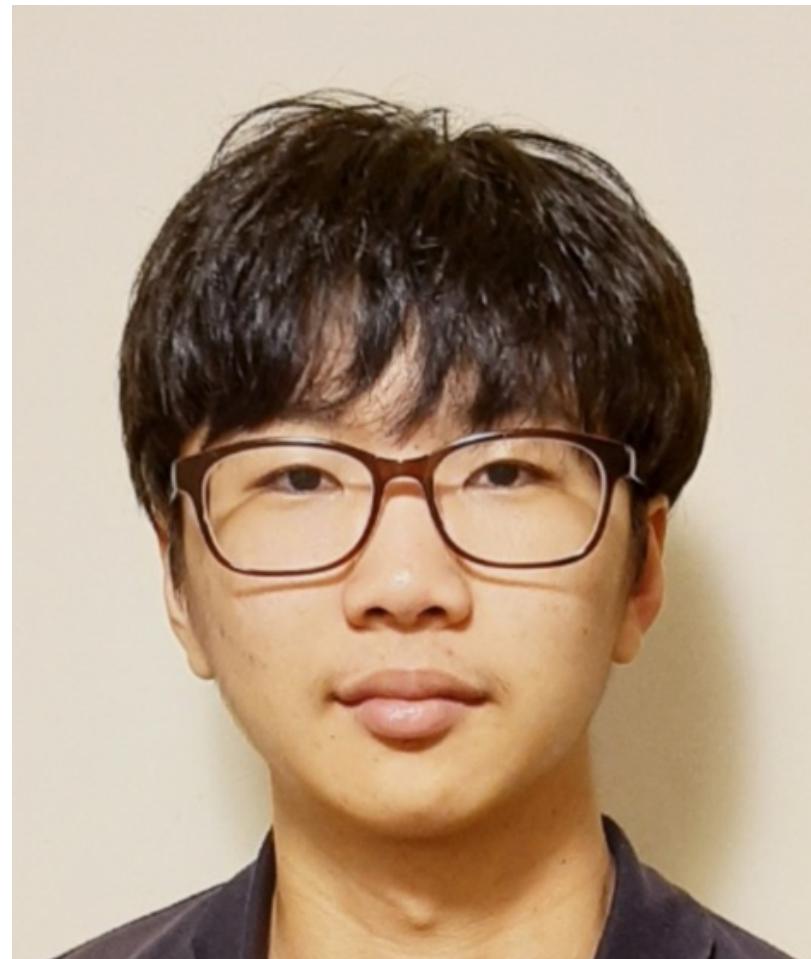


Titans Tech

THE TITANS TECH TEAM



ALEX KIM



STEVE JANG



KLAUS LEUNG

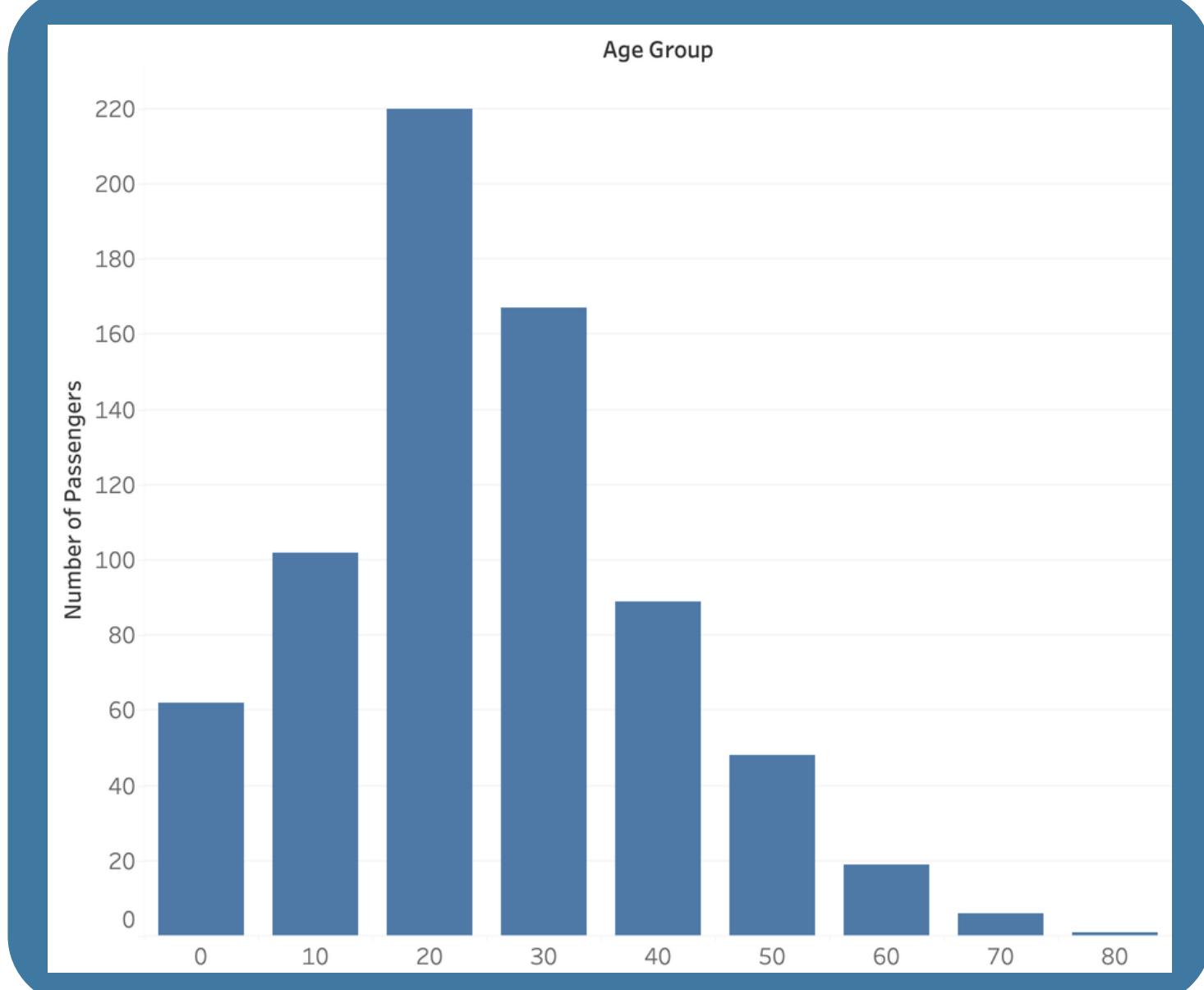
DATA CLEANING

Null Values present in the fields: **Age**, **Cabin** and **Embarked**



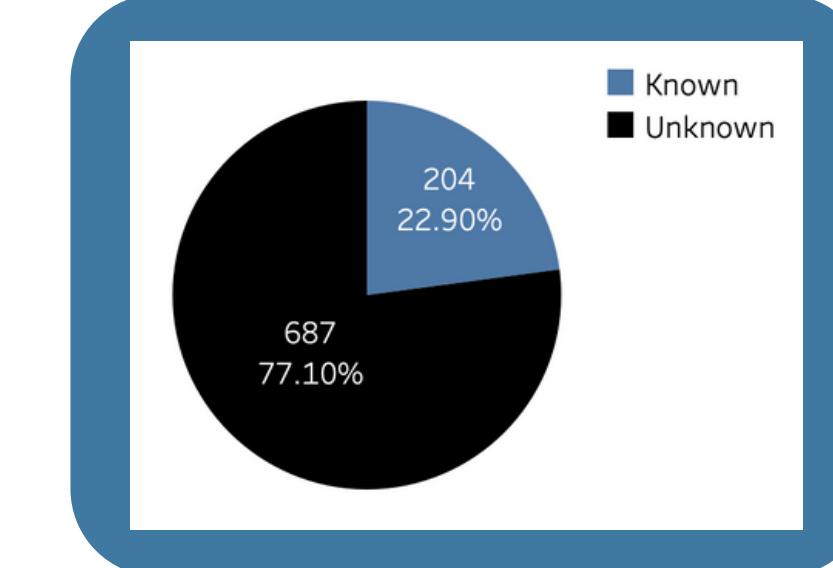
Age

- Backfill the **177** nulls with the **mean** of the ages, as it follows the **Normal Distribution**
 - Normality test used to confirm the appropriateness of fit
 - Backfilling with mean is preferred over median or mode, if normally distributed



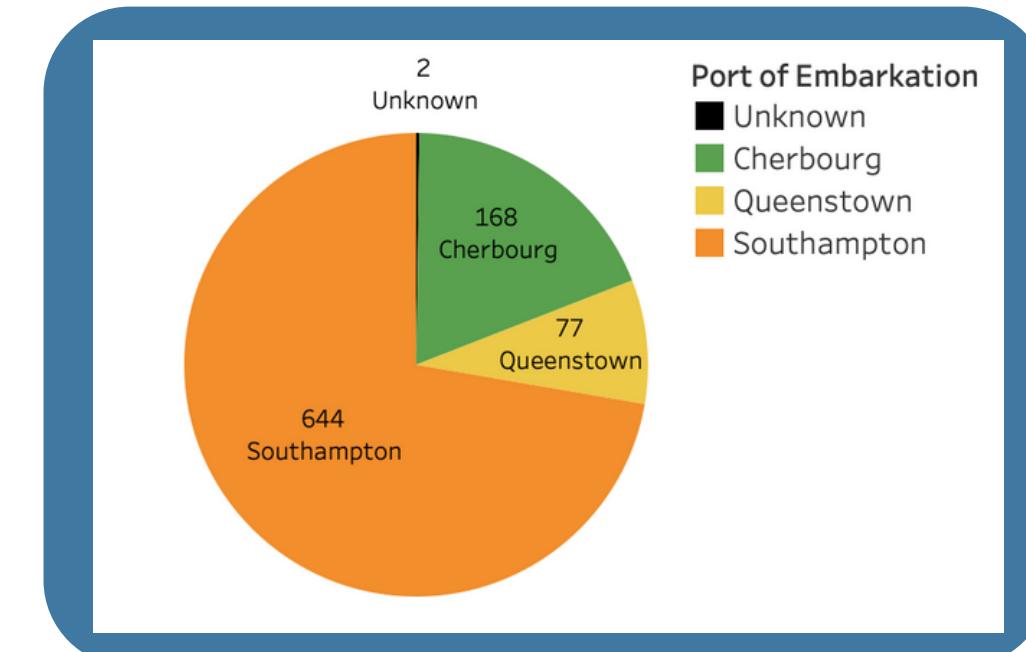
Cabins

- 687 of 891** values are **null**. Replace nulls with a new category, "**Unknown**"
 - Majority of the field is null
 - Cannot backfill with statistical measures as it distorts data



Embarked

- 2** **null** values - Backfill nulls with the highest occurring embarkation port, due to the negligible amount



POPULATION OVERVIEW

AGE/SEX DISTRIBUTION

Significantly more males present in the ship, with over 25% of passengers being males in their 20s

Sex	0	10	20	30	40	50	60	70	80
Female	30	45	125	60	32	18	4		
Male	32	57	272	107	57	30	15	6	1

Count of Age



Age Group	Count
1	272

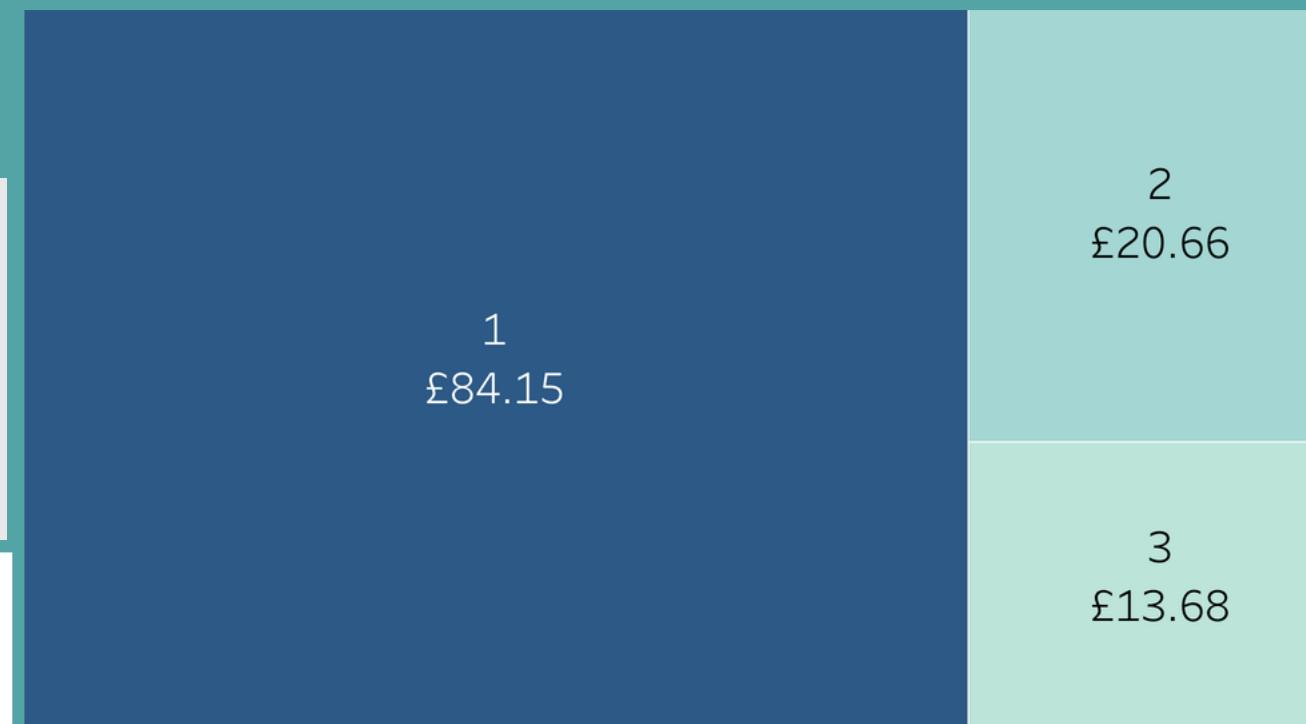
AVERAGE FARE COST BY CLASS

Today's Cost (USD)

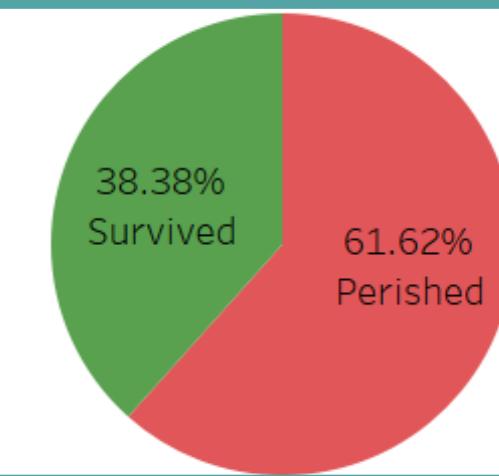
Class 1 - \$420.75

Class 2 - \$103.3

Class 3 - \$68.4



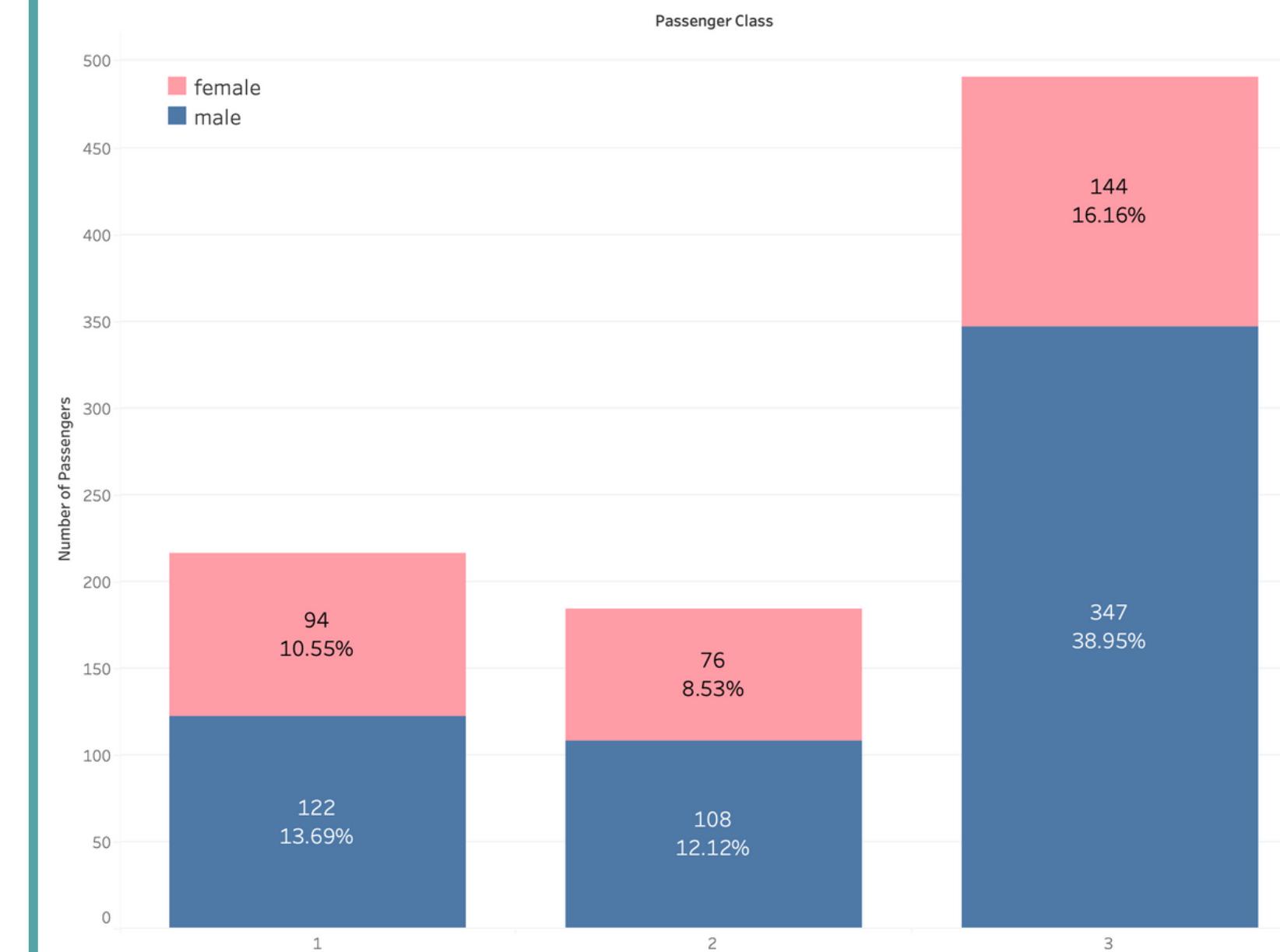
OVERALL SURVIVAL RATE



SEX/CLASS DISTRIBUTION

Most passengers travelled in the 3rd class, which was the cheapest

Population by Class and Gender



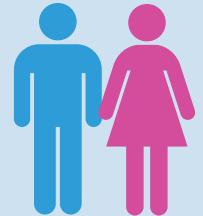
KEY QUESTIONS

WHAT ARE THE MOST IMPORTANT FACTORS AFFECTING SURVIVAL?

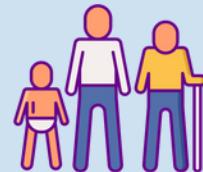
1 Does a higher **passenger class** lead to more safety (lower mortality rates)?

1
2
3

2 Did a specific **sex** have a higher survival rate?



3 How does the survival rate vary between the **ages** (young and old)?



4 Did having **family members** lead to a higher survival rate?



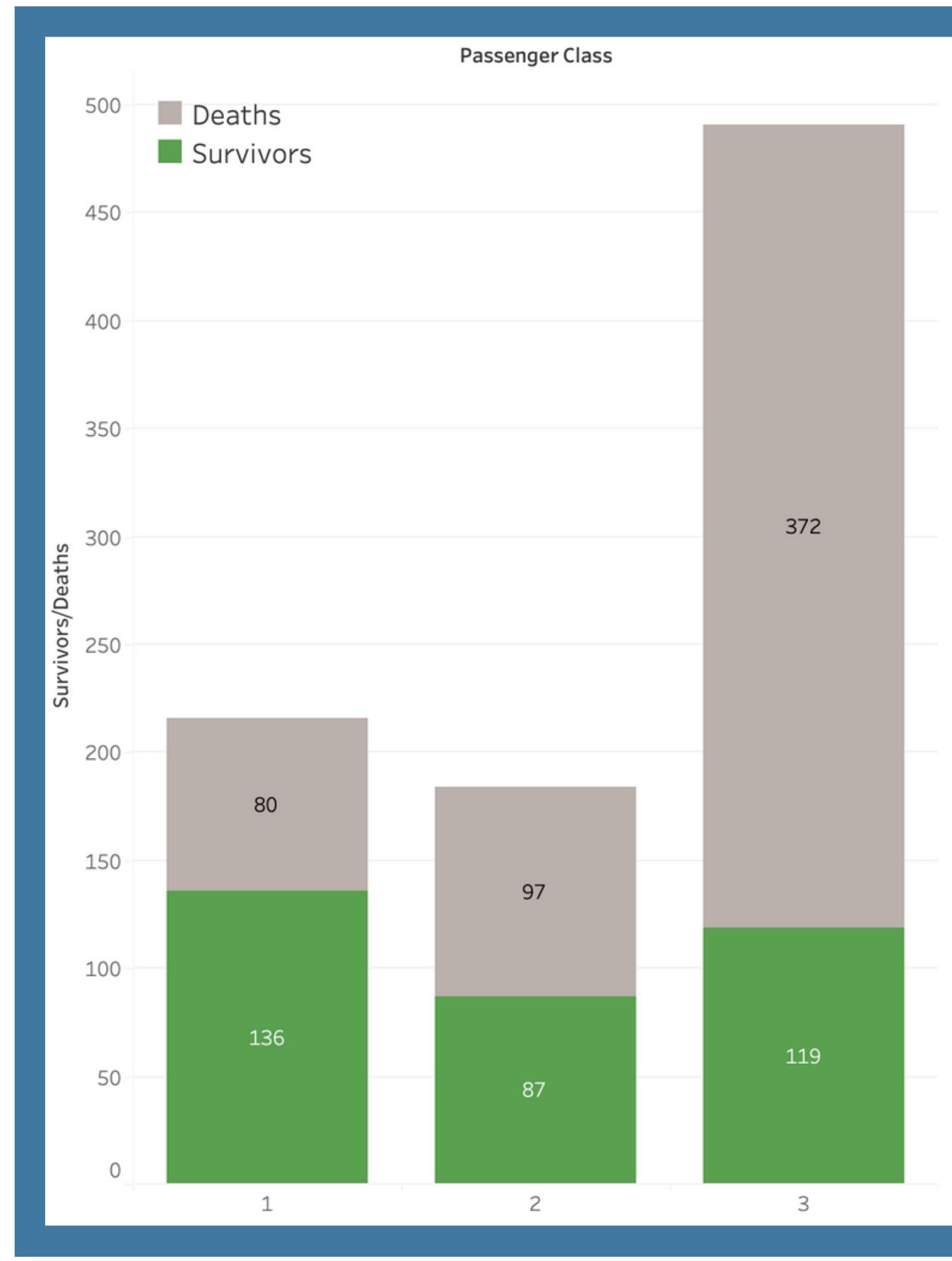


PASSENGER CLASS



PASSENGER CLASS

Insights



Passenger Class

Sex

Age

Family

Higher classes linked to a greater survival rate



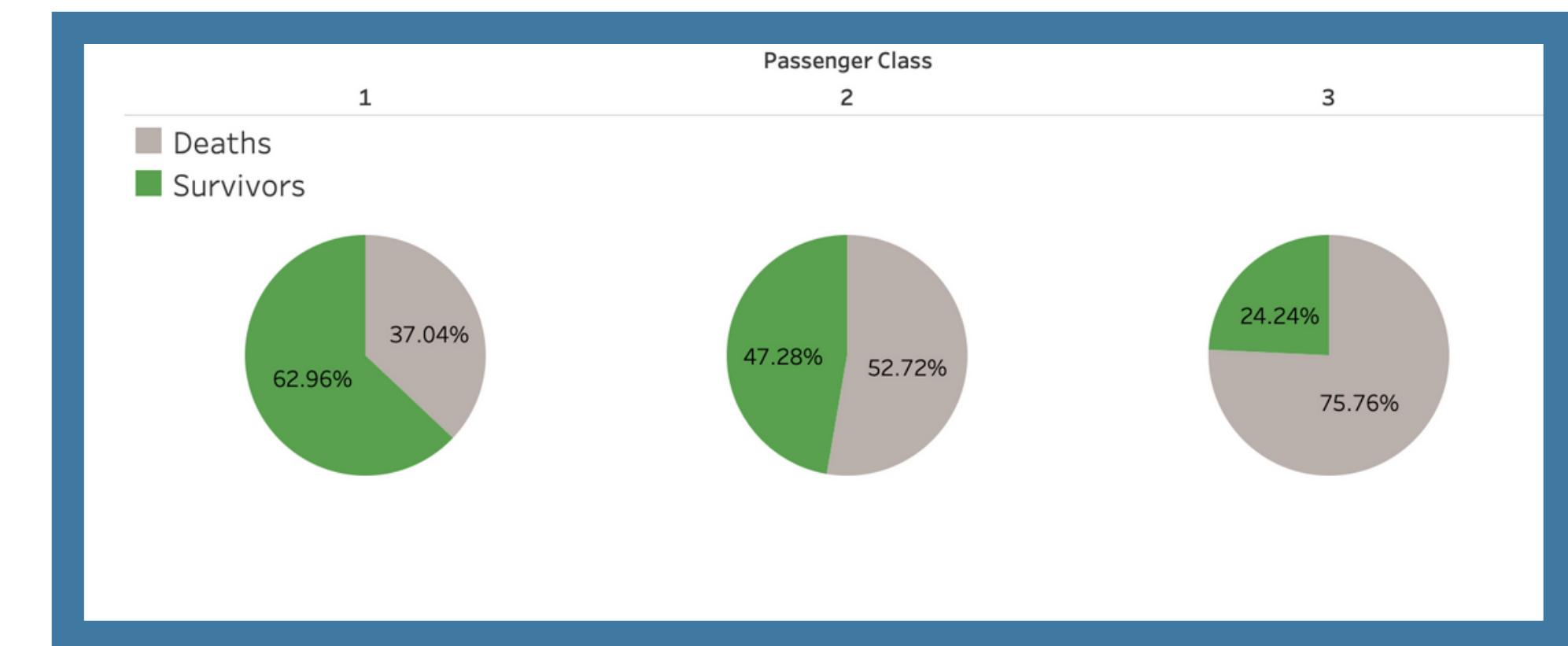
1st class passengers **15% more likely** to survive than 2nd class
2nd class passengers **2 times more** likely to survive than 3rd class



Disparity in survival chances occurred although there were more 3rd class passengers than **1st and 2nd combined**



This disparity + population difference meant that more than **two thirds** of deaths were from the **3rd class**



PASSENGER CLASS: FARE PRICES

 **Greater survival rate with higher fares paid**

Higher fare prices is directly related with higher passenger class

Rich passengers survived more than commoners

This is despite the total number of 1st class passengers being only 44% of the number of 3rd class passengers.



Disparity due to the disadvantageous cabin positions of the 3rd class on the Titanic

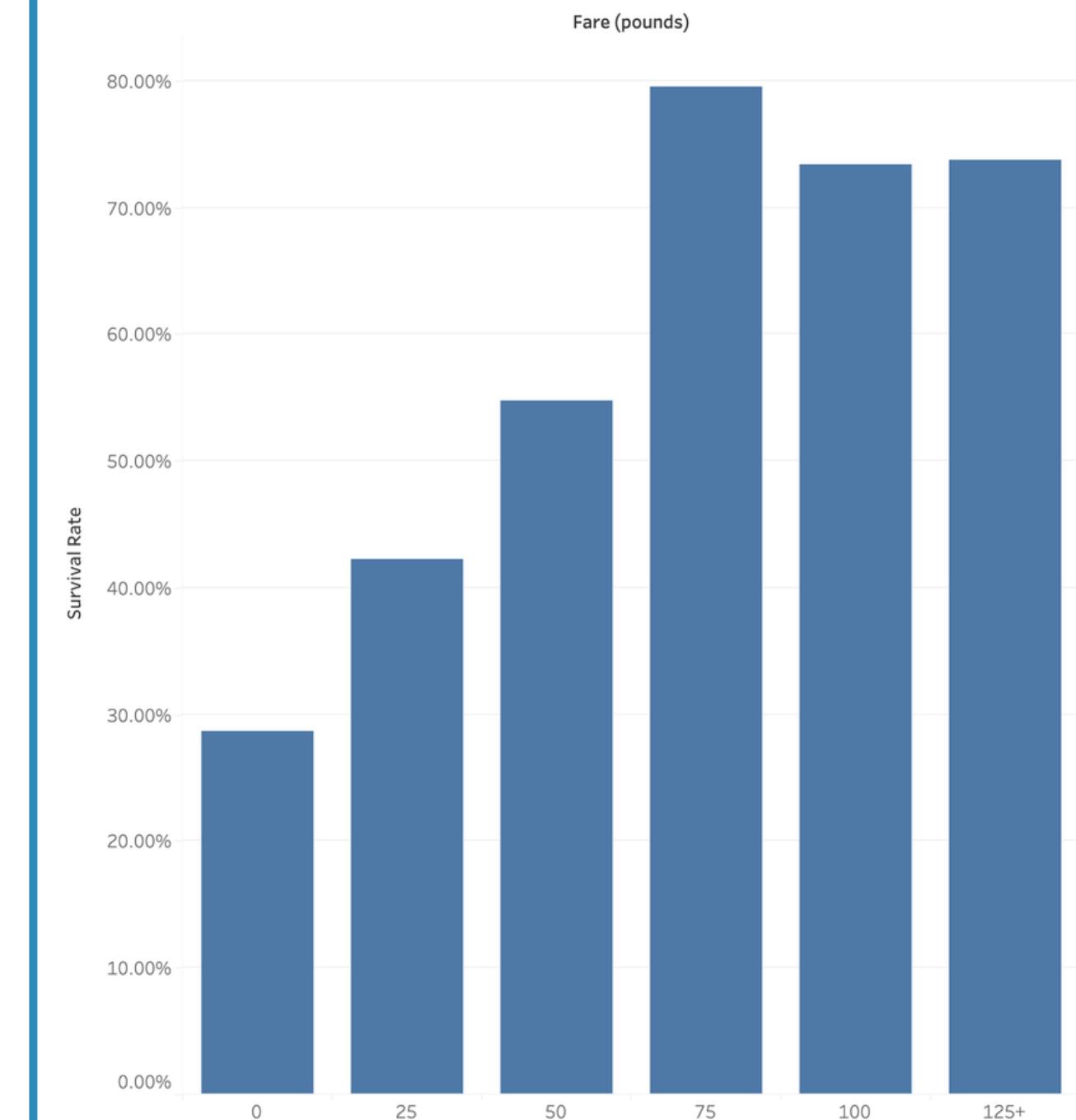
*66% of those who spent £50 or less died
24% of those who spent £75 or more died*

.....
An increased 42% chance of survival!

Survivors/Deaths by Fare

	0	25	50	75	100	125+
Survived	160	73	35	35	11	28
Died	397	100	29	9	4	10

Survival Rate by Fare



PASSENGER CLASSES: CABINS



The Illustrations show that the higher passenger class cabins were more spacious, meaning that passenger density was lower, allowing for **faster evacuation**



The **staircases** would have gotten heavily clogged during the scramble, especially for the **third class**, where the majority of people belonged

The Titanic had a total of 17 stories

FIRST CLASS



SECOND CLASS



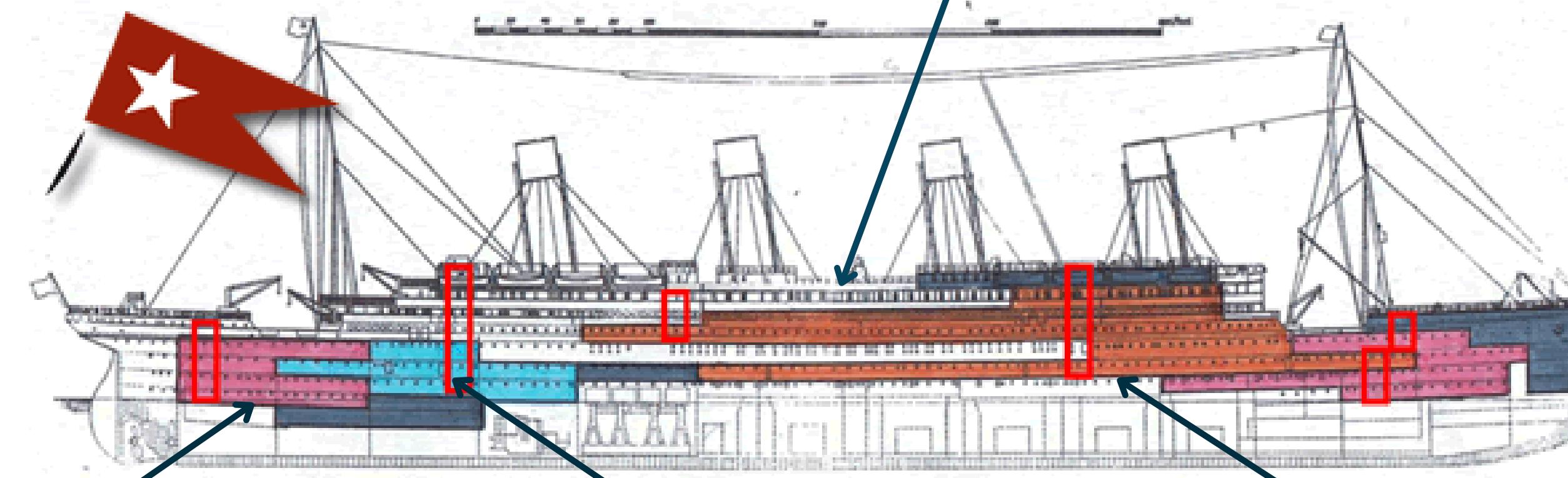
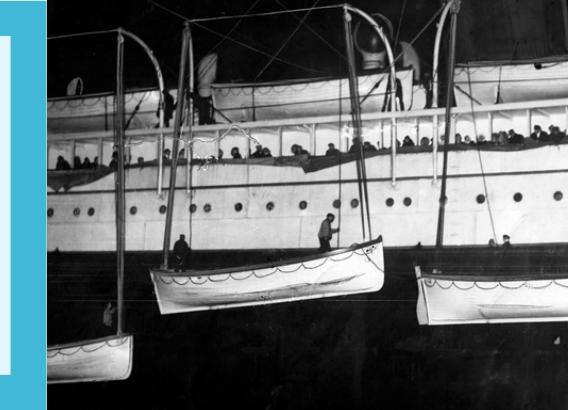
THIRD CLASS



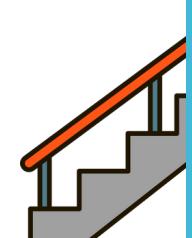
PASSENGER CLASSES: CABINS

The lifeboats were designed to hold only **65 people** each

The lifeboats were stored at the **top & middle** of the Ship, nearest to 1st class



The third class cabins were on each **end** of the Ship, being the **most distant** from the exits and lifeboats



The second class cabins had a **direct** but long staircase leading to the lifeboats



First class passengers were situated **closest** to the 20 lifeboats' storage

Passenger Class

Sex

Age

Family

PASSENGER CLASSES: 3RD CLASS ETHNICITIES



Last names of passengers were analysed with a data warehouse of names to predict possible **ethnic backgrounds**

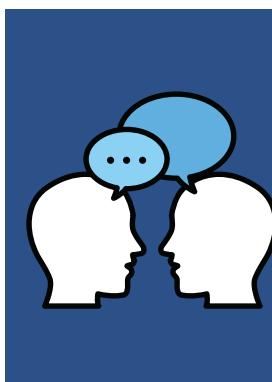
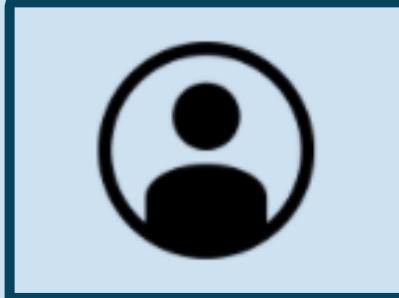


Majority of the 1st and 2nd class passengers were **native Europeans**

1st class: upper class (businessmen, politicians)

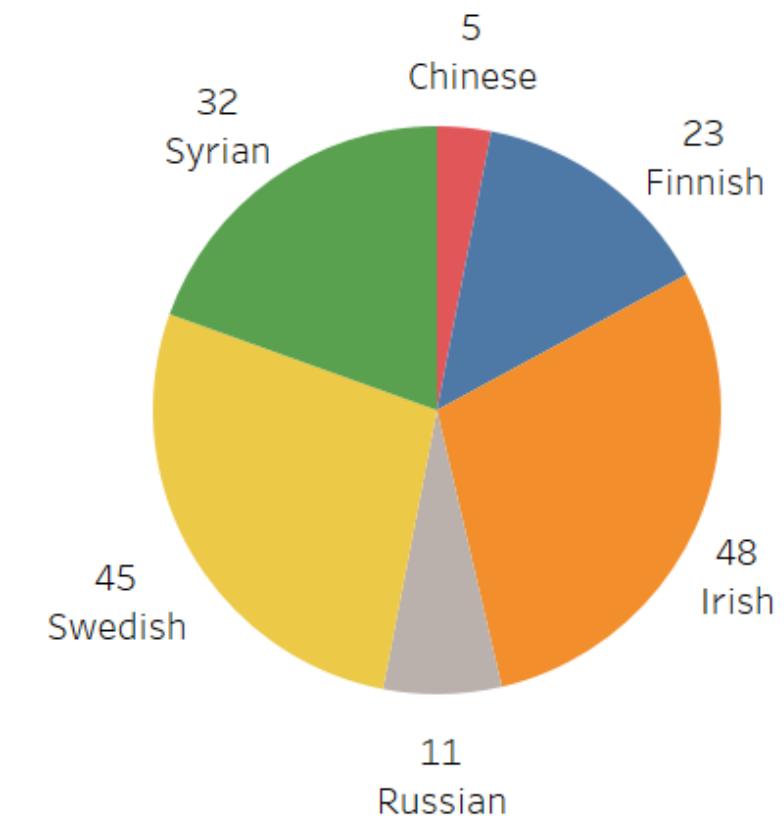
2nd class: tourists, authors

3rd class: immigrants moving to US / Canada

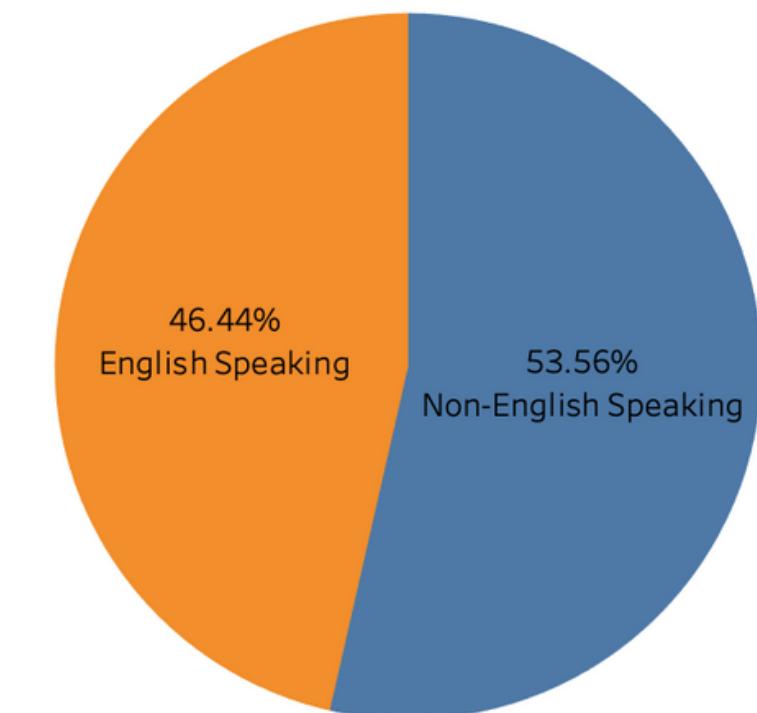


Highly likely that third-class emigrants died due to **poor English skills**, hindering **communication** and lowering their likelihood of survival

Immigrant Ethnicities



3rd Class Passengers by Language





SEX



SEX: ANALYSIS

- Males had a **19%** survival rate
- Females had a **74%** survival rate



Huge disparity caused by:

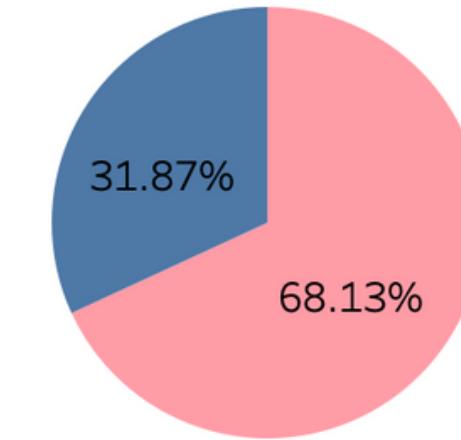
- "Women and children first" mentality
- Men most fit to help others

Interesting Facts:

- **97%** of crew members were male
- The number of males who died **exceeds** total females onboard

Sex Distribution of Survivors

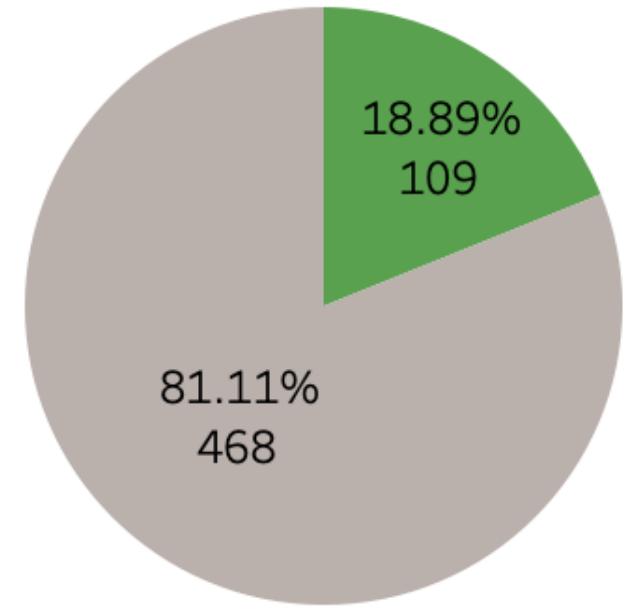
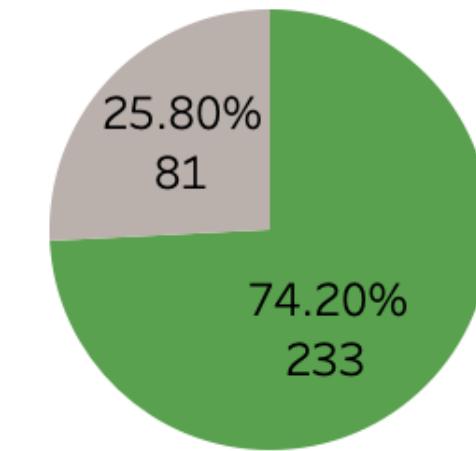
Sex
■ female
■ male



female

male

■ Survivors
■ Deaths



Passenger Class

Sex

Age

Family

SEX: HISTORICAL CONTEXT

Women in society



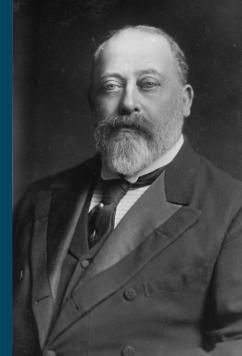
Increased societal interest towards women's rights

Women had increased employment opportunities

Although excluded from leadership roles, women gained a wider access to knowledge sources

These contextual factors likely influenced the prioritising of women's survival in the Titanic evacuation

Edwardian Era



Chivalry common, with a "**Ladies first**" European etiquette

Deemed correct behaviour for men to give women **precedence over themselves**

Large disparity between rich and poor, with great change (political + social)

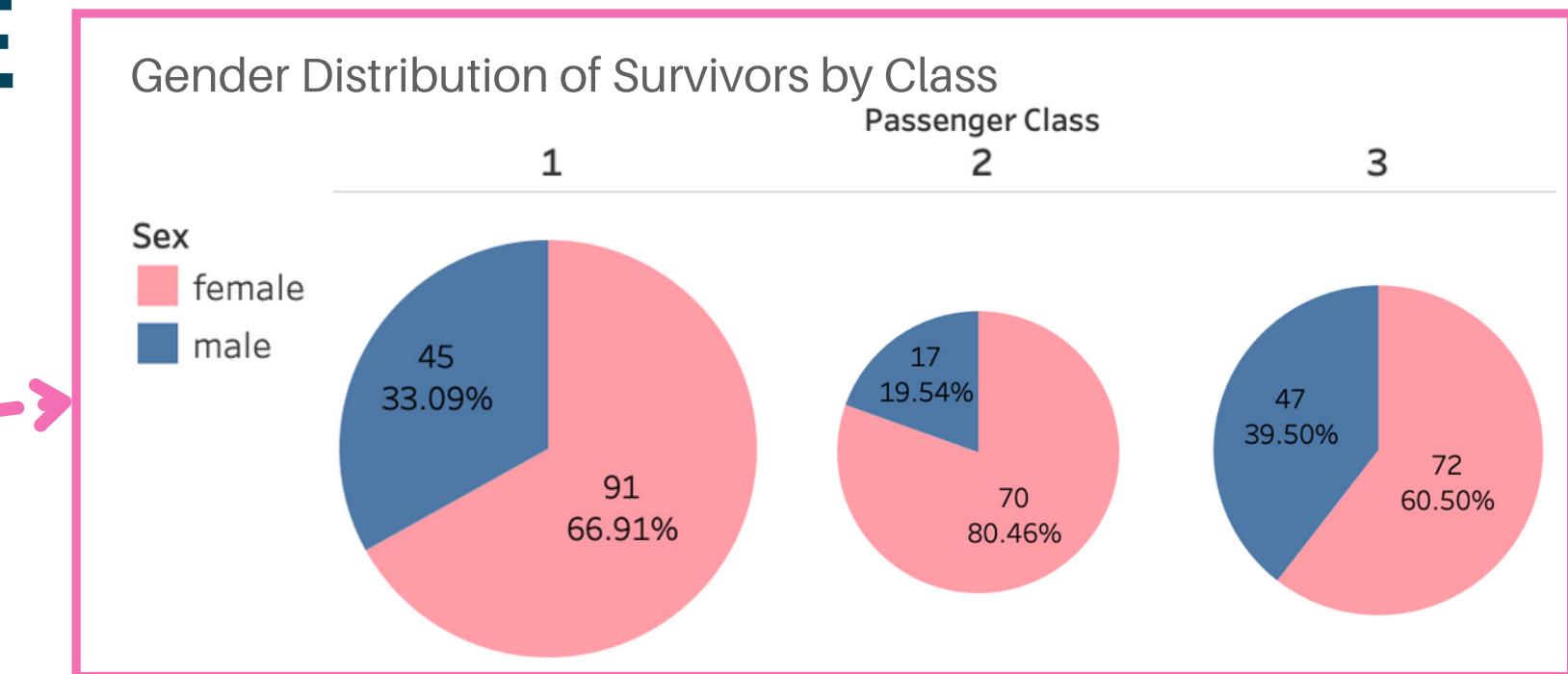
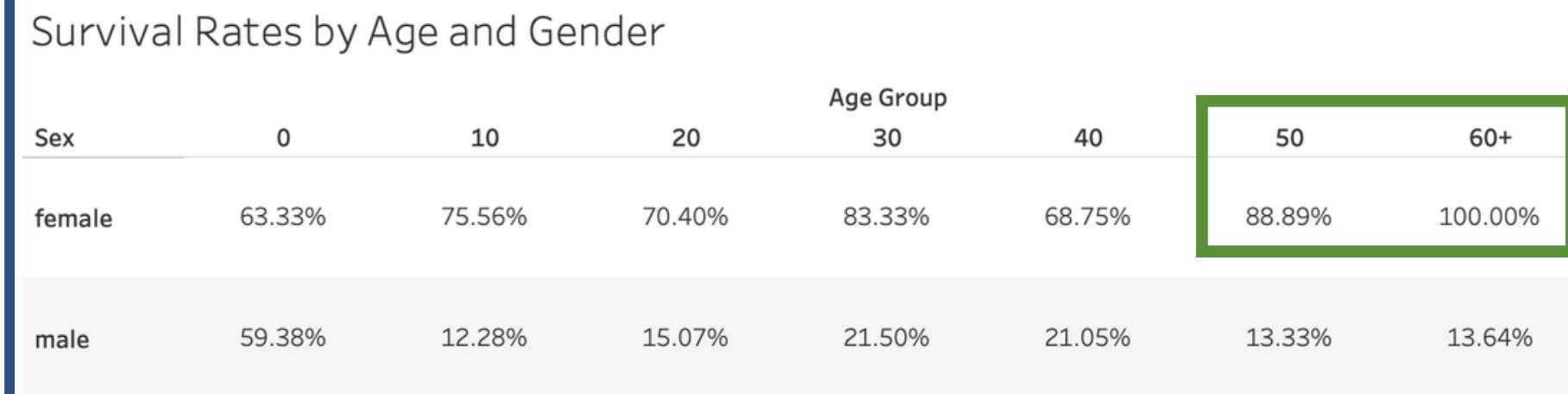


SEX: PASSENGER CLASS & AGE



Further Insights

- For every passenger class, at least **60%** of survivors were women
- No women aged 60+ died, likely due to others (men) prioritising their evacuation
- Female survivors aged 50+ are displayed on the right:
 - 14 of 20** from First Class
 - 17 of 20** were native English speakers
- Reinforces that **First Class** Passengers, and **native English speakers** were more likely survive



Background, Age and Class of 50+ Female Survivors

Name	Passenger Class	Age	English Background
Andrews, Miss. Kornelia Theodosia	1	63	Yes
Appleton, Mrs. Edward Dale (Charlotte Lamson)	1	53	Yes
Baxter, Mrs. James (Helene DeLaudeniere Chaput)	1	50	Yes
Bonnell, Miss. Elizabeth	1	58	Yes
Eustis, Miss. Elizabeth Mussey	1	54	Yes
Graham, Mrs. William Thompson (Edith Junkins)	1	58	Yes
Hays, Mrs. Charles Melville (Clara Jennings Gregg)	1	52	Yes
Hogeboom, Mrs. John C (Anna Andrews)	1	51	Yes
Lurette, Miss. Elise	1	58	No
Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	1	56	Yes
Rothschild, Mrs. Martin (Elizabeth L. Barrett)	1	54	No
Stephenson, Mrs. Walter Bertram (Martha Eustis)	1	52	Yes
Stone, Mrs. George Nelson (Martha Evelyn)	1	62	Yes
Warren, Mrs. Frank Manley (Anna Sophia Atkinson)	1	60	Yes
Hewlett, Mrs. (Mary D Kingcome)	2	55	Yes
Hocking, Mrs. Elizabeth (Eliza Needs)	2	54	Yes
Parrish, Mrs. (Lutie Davis)	2	50	Yes
Ridsdale, Miss. Lucy	2	50	Yes
Toomey, Miss. Ellen	2	50	Yes
Turkula, Mrs. (Hedwig)	3	63	No

Passenger Class

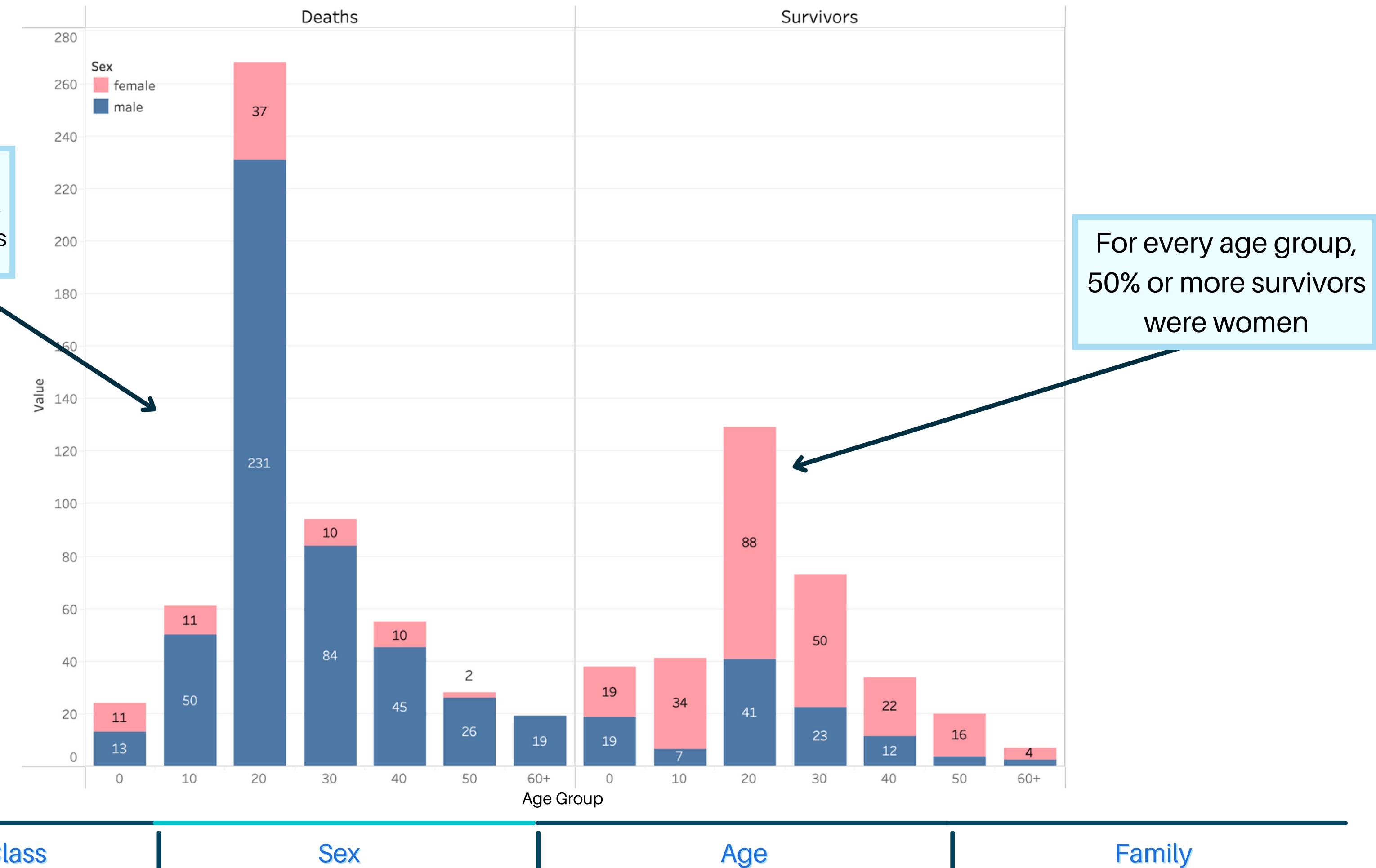
Sex

Age

Family

SEX: AGE DISTRIBUTION BY OUTCOME

A majority of people who perished were young men, especially those in their 20s



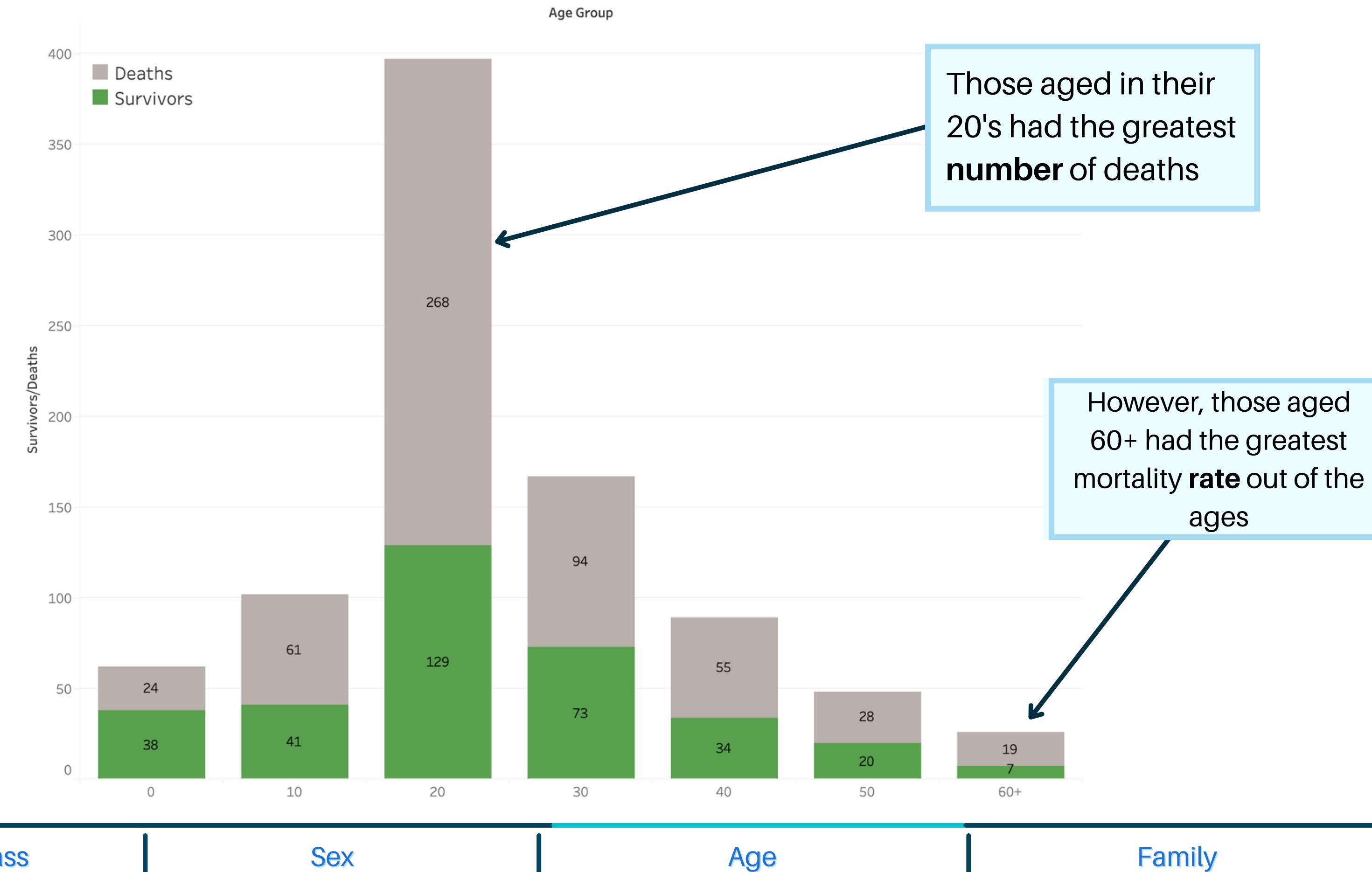


AGE



AGE: SURVIVORS/DEATHS

Survivors/Deaths by Age



AGE: ANALYSIS



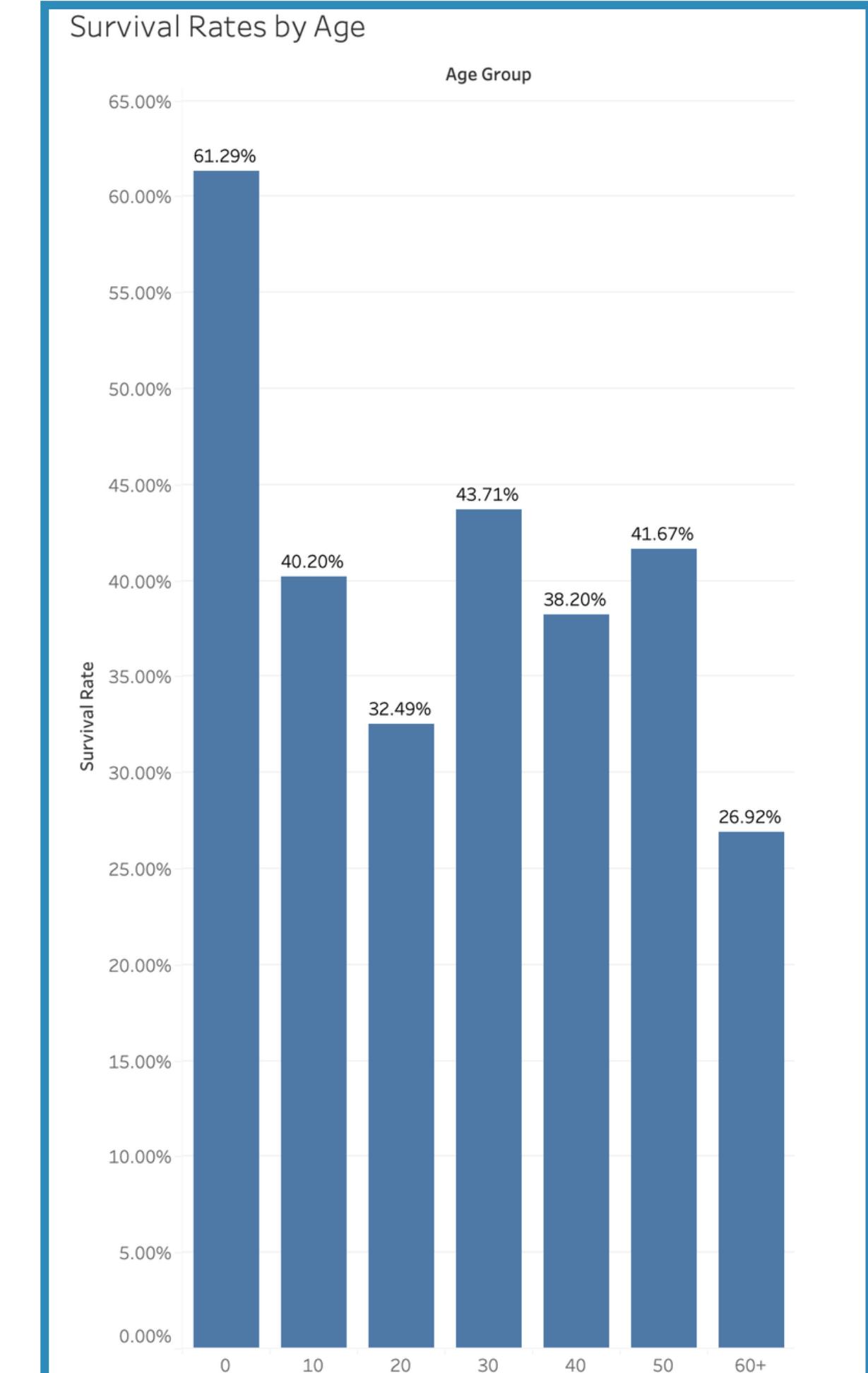
Elderly passengers, especially **ages 60+** had the lowest survival rate - most liable to perish from **hypothermia** in **-2°C waters**



Children aged **0-9** years old had the highest survival rate, likely due to their safety being prioritised by their **parents**



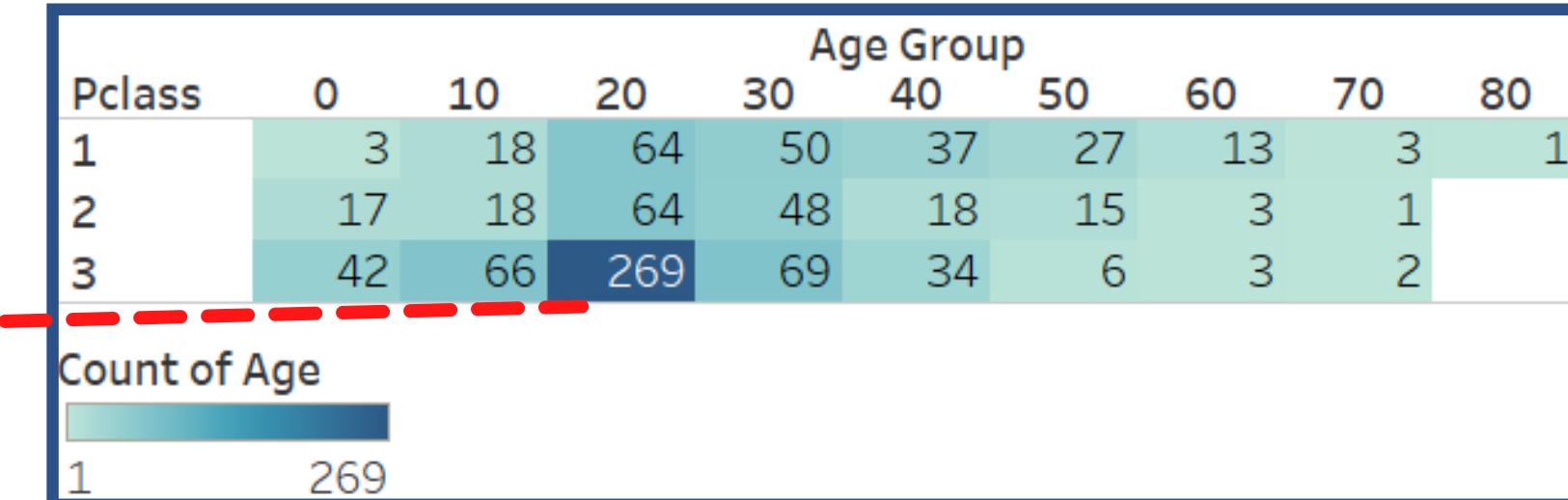
Older children and young men in their **20s** had a surprisingly low survival rate despite their relative fitness - likely due to **confidence** and helping weaker age groups



AGE: PASSENGER CLASS

Those aged in their **20's** died the most, as a majority of them were in 3rd class

- Likely due to inability to afford high class tickets compared to those older



Men aged 20+ had highest mortality rate

- Were confident with their ability to help others
- Combination of the 2 factors with worst survival rate: 3rd Class & Male

Interestingly , the survival rate of **1st class men** was very similar to that of **3rd class children** (both about 35%), despite children being prioritised over men in general

- This affirms that **Passenger Class** was the most important factor for survival

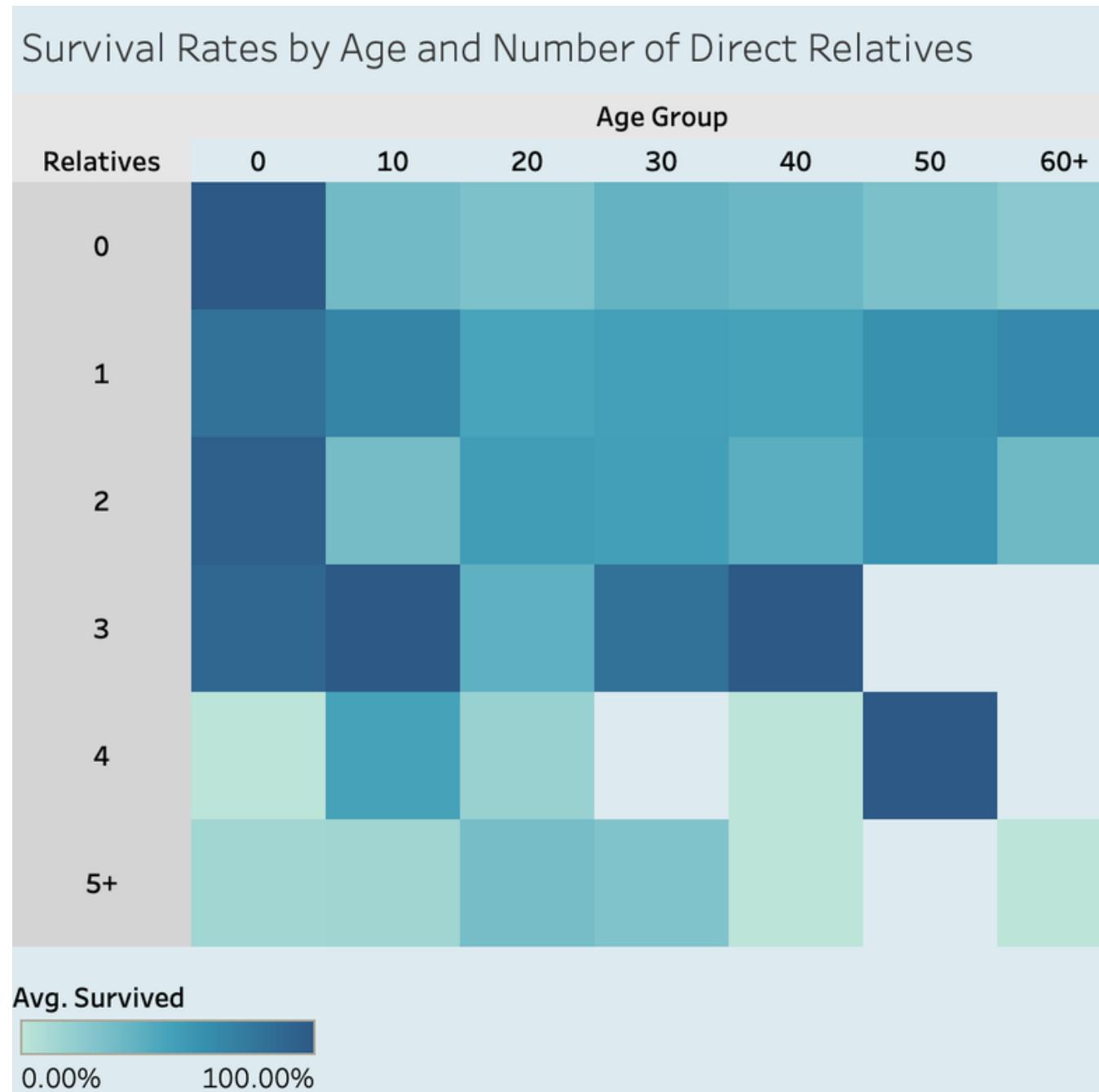


Richard L. Beckwith | 1st Class Stockbroker



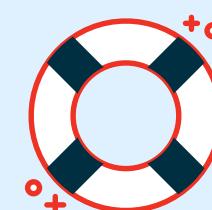
Clarence G.H. Asplund | 3rd Class Child of 2 boarded parents

AGE: FAMILY MEMBERS



Heatmap Analysis

Higher age linked to lower survival rates



As age group increases, more family members increased survival:

- Children with fewer siblings had more attention from parents
- Older passengers with more supporting relatives were safer



Those in their teens with 3 siblings also stand out with a **higher** survival rate, likely due to them being very active with family responsibilities



As expected, those in their 60+ years had very low survival rates, but had a higher chance of surviving with 1 relative

Passenger Class

Sex

Age

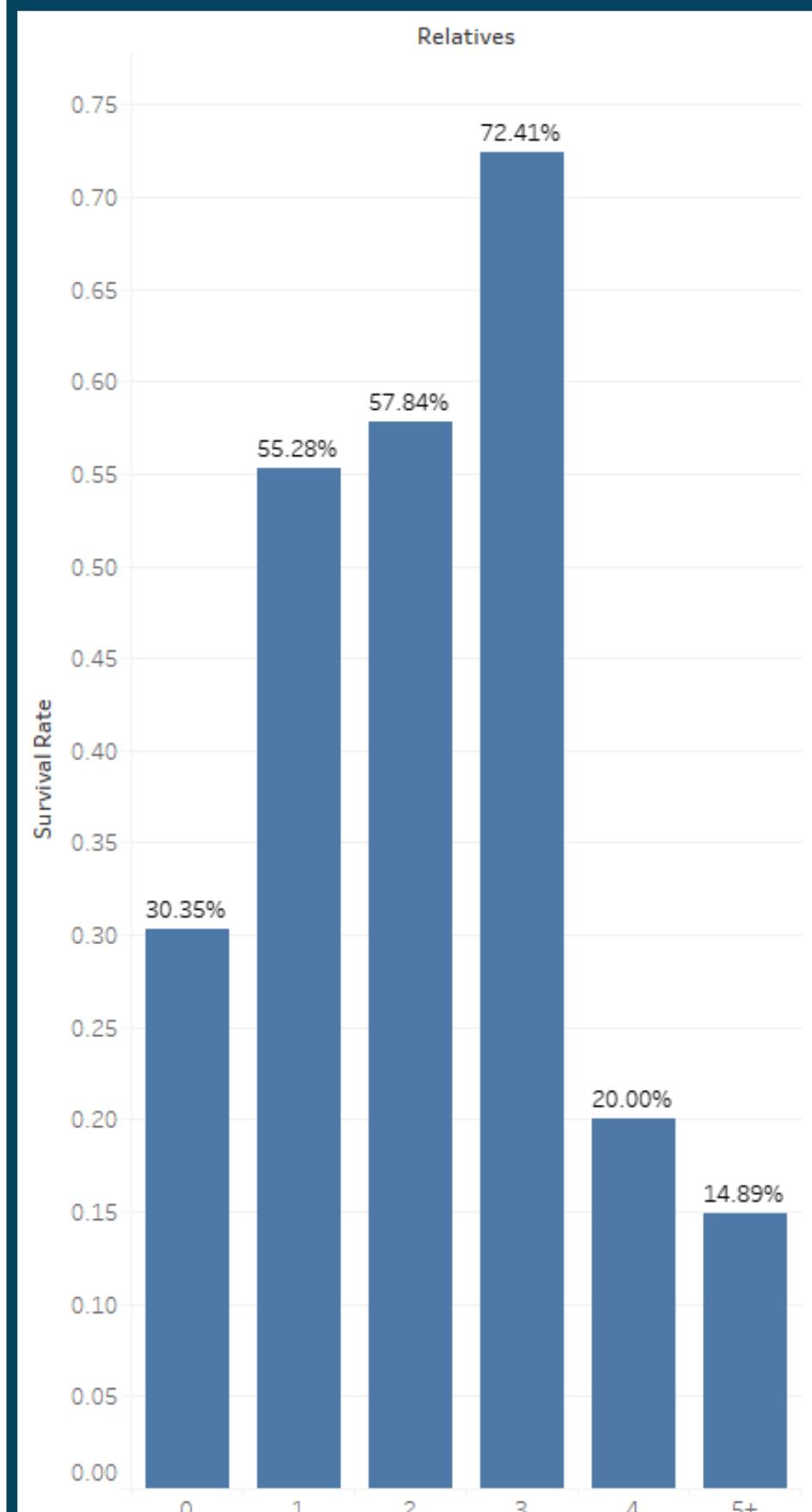
Family



FAMILY



FAMILY: ANALYSIS



Those with **3 direct relatives** had the highest survival rate of **72.4%**

Most individuals had **0 relatives** on board, yet their survival rate was only **30%**

May be linked to **heroism**, as those with 0 relatives may have had the **most intent** to save others

Individuals with 4+ direct relatives were mostly from the 3rd class, likely explaining their low survival rate



Passenger Class	Relatives					
	0	1	2	3	4	5+
1	109	70	24	7	2	4
2	104	34	31	13	1	1
3	324	57	47	9	12	42

Passenger Class

Sex

Age

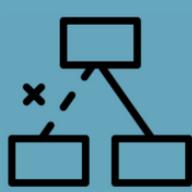
Family

FAMILY: PASSENGER CLASS AND SEX

Survival rates by Siblings/Spouses & Gender

Survival Rate by Siblings/Spouses & Class

Siblings/Spouses	Passenger Class		
	1	2	3
0	54.20%	38.18%	21.30%
1	73.85%	47.73%	26.87%
2+	75.00%	42.86%	13.64%



No matter the number of siblings and spouses, being in a higher passenger class led to a **higher** survival rate



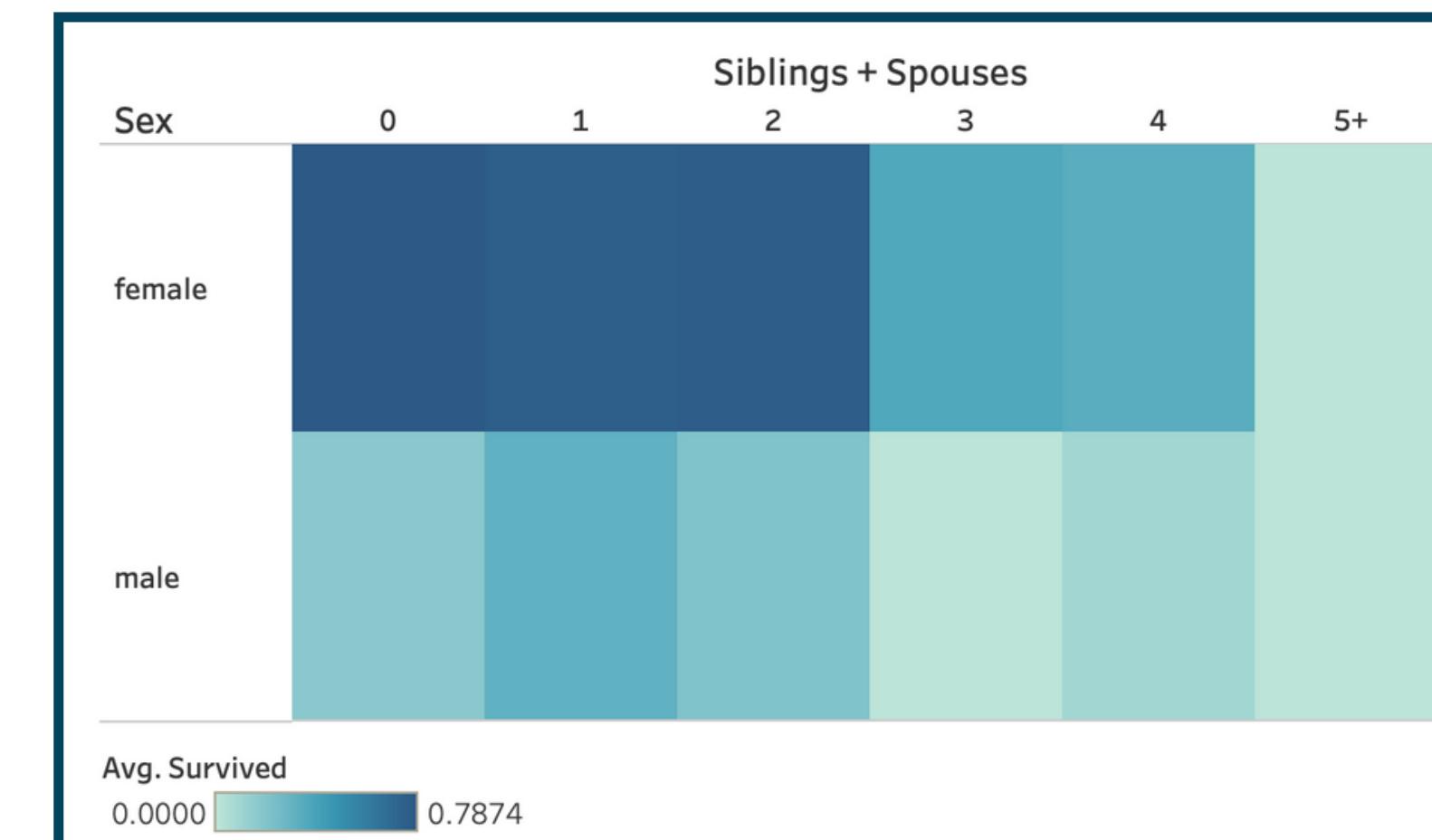
Having **more** siblings and spouses on board was **advantageous** in the first class, but this is not clear for lower passenger classes



For especially women, having **more** siblings or spouses led to **lower** chances of survival



This **opposite trend** hints that there is **no general relationship** between the number of siblings/spouses & survival rate



Passenger Class

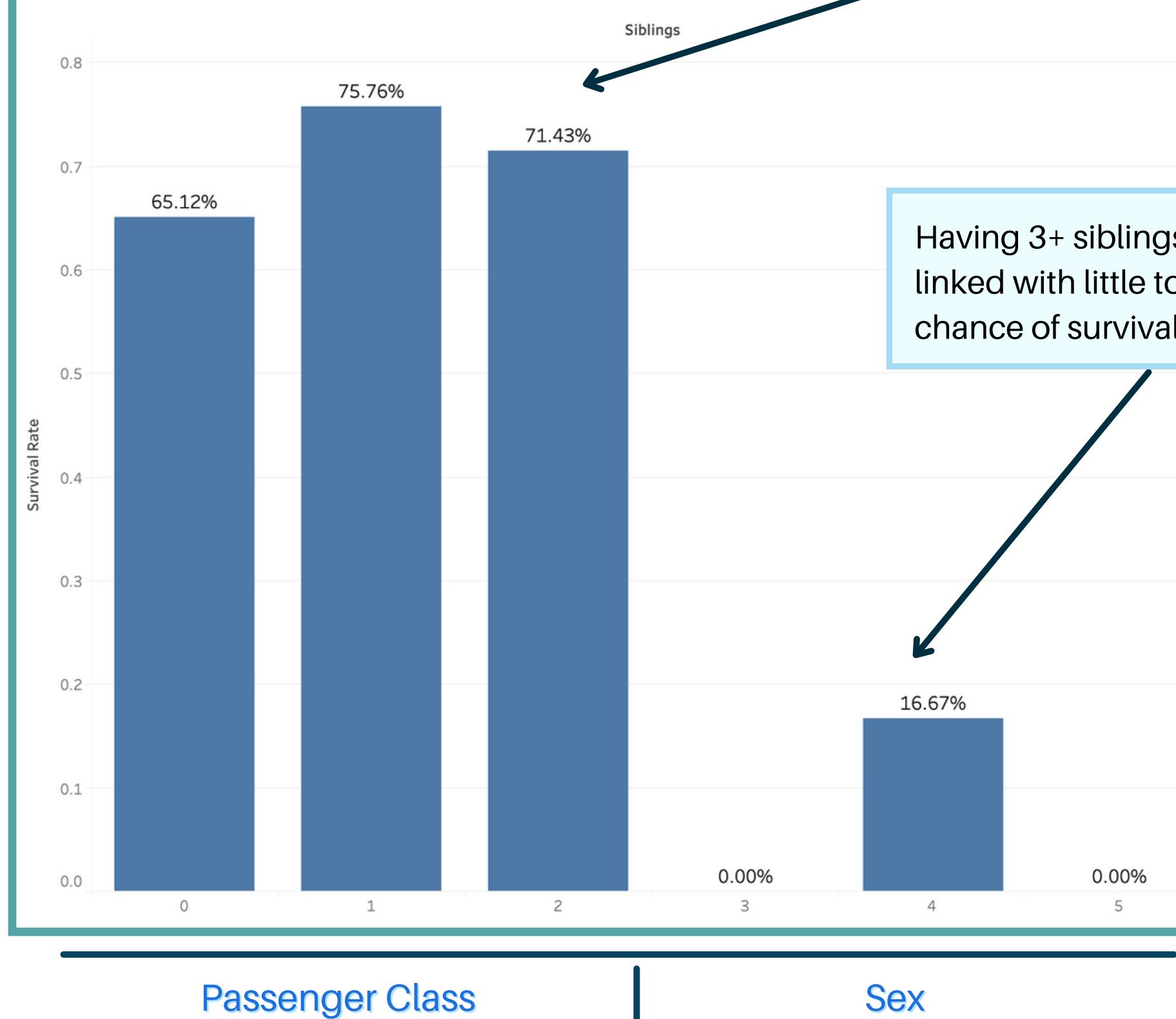
Sex

Age

Family

FAMILY: CHILDREN WITH SIBLINGS

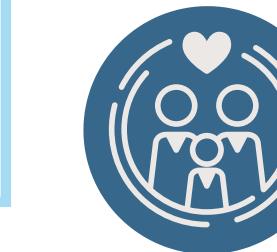
CHILD SURVIVAL RATE BY NUMBER OF SIBLINGS



Children with 0-2 siblings had a high survival rate

Siblings

Having 3+ siblings linked with little to no chance of survival



- Children with **0 to 2** siblings had relatively high survival rates of **over 50%**
- Likely due to them helping each other



- Children with **3 or 5** siblings had a **100% mortality** rate, and those with **4** siblings with **84%**
- Likely due to slower evacuating as more siblings mean each have less attention from parents



Passenger Class

Sex

Age

Family



CONCLUSION



CONCLUSION

WHAT IS THE SINGLE MOST IMPORTANT FACTOR AFFECTING SURVIVAL?

Passenger Class

- 1
- 2
- 3

Passenger Class: First class survived more, due to cabins' proximity to lifeboats



Sex: Females survived more, due to the values of the time



Age: Younger people survived more, due to their fitness



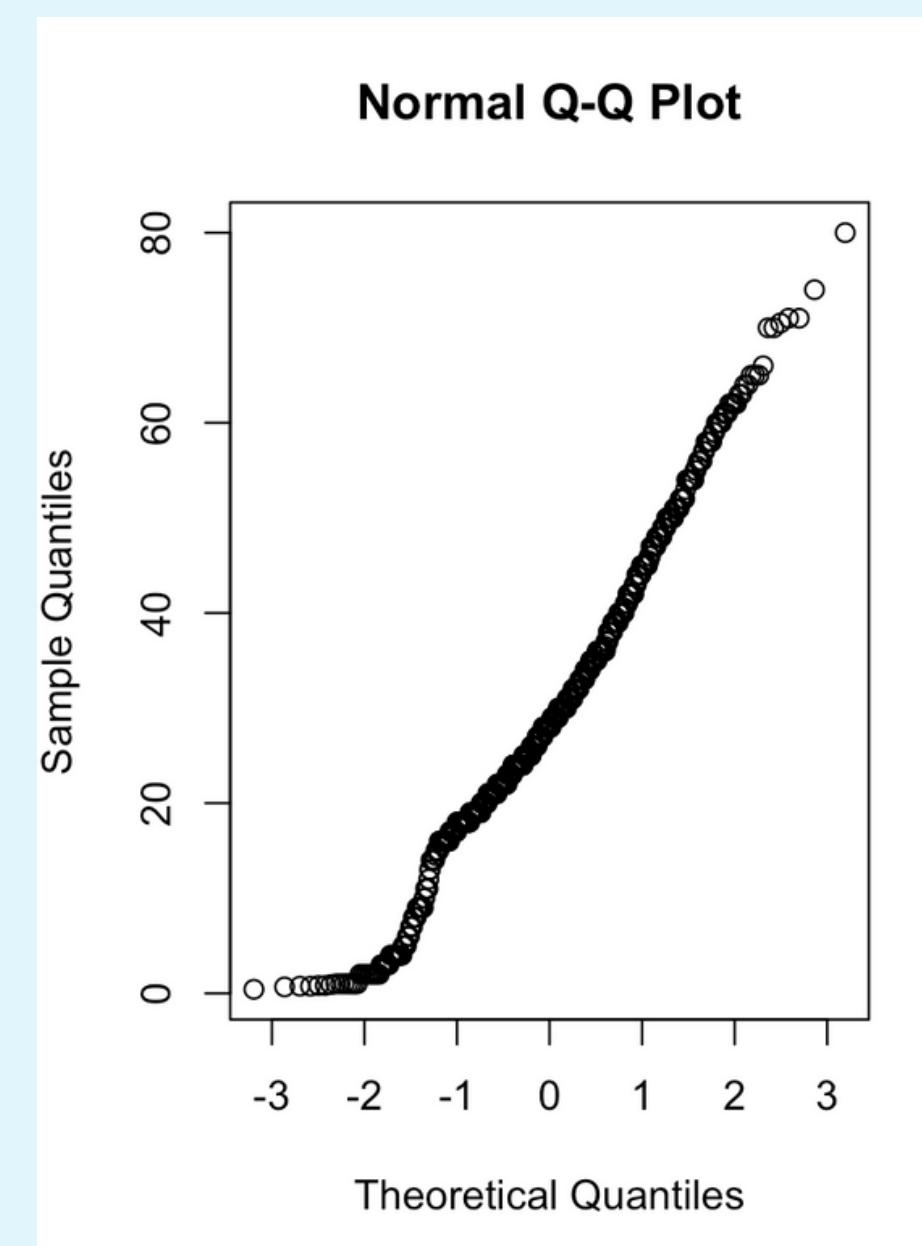
Family: Younger people with less relatives survived more, opposite for old



APPENDIX



RATIONALE FOR BACKFILLING NULL VALUES



- Dropping of column is only appropriate if it doesn't affect analysis
- Removing data leads to loss of information, which returns biased data
- Null values are replaced with:
 - **Mean** if dataset is normally distributed
 - **Mode** if there are frequent values & number of nulls is small
 - **Median** if data comprises of significant outliers

```
stats.normaltest(df['Age'])
```

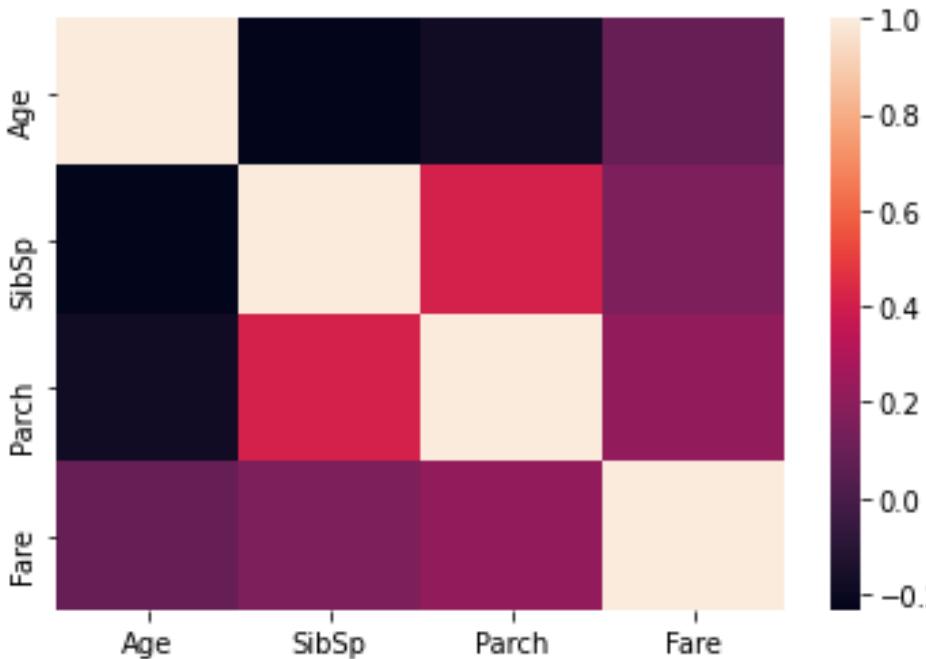
```
NormaltestResult(statistic=43.85896585051567, pvalue=0.299327)
```

Methodology to decide which variables to analyse

1. Chose the most intuitive fields that could influence survival
2. Confirmed hypothesised relationships by computing correlation and testing for independence (Chi-square Test)
3. Concluded that Passenger Class, Sex, Age & No. of Relatives were the most influential

Output from Python code for checking correlation between variables

→ The p-value of the Chi-square test comparing Pclass with Survived is: 4.549251711298793e-23
The p-value of the Chi-square test comparing Sex with Survived is: 1.1973570627755645e-58
The p-value of the Chi-square test comparing Embarked with Survived is: 2.3008626481449577e-06



Methodology to answer the 4 main questions

- For each variable of concern (Class, Sex, Age, Siblings/Spouse), subquestions were constructed
- These subquestions involved comparing **Survival Rate** with the chosen variable with other 3 variables
 - e.g. For Sex, Class and Age died the most?

Technologies Used



- Used Python (pandas) to check rough relationships of both categorical & numerical values
 - Also used to export filtered CSV files (e.g. only children data)
- Used R for statistical measures of numerical rows
- Used Shell to derive Passenger Ethnicities from last names
- Visualised with Tableau after data cleaning

Methodology for Mapping Surnames to Backgrounds

1. Used the website www.familysearch.org, which maps surnames to countries based on a **large worldwide database**
2. Compiled a list of surnames appearing in our dataset
3. Wrote a shell script using linux tools including *curl* to automatically map every surname to candidate countries
4. Filtered out all surnames which had the top candidate country as an English speaking country

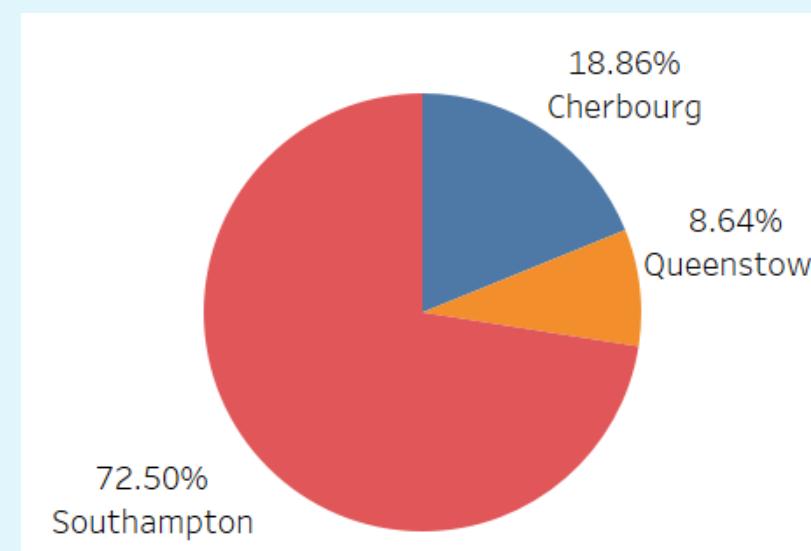
Shell Script Output

Abbing,Netherlands,Germany,United States
Abbott,United States,England,Canada
Adahl,Sweden,United States,Finland
Adams,United States,England,Canada
Ahlin,Sweden,United States,Samoa
Aks,United States,England,Austria

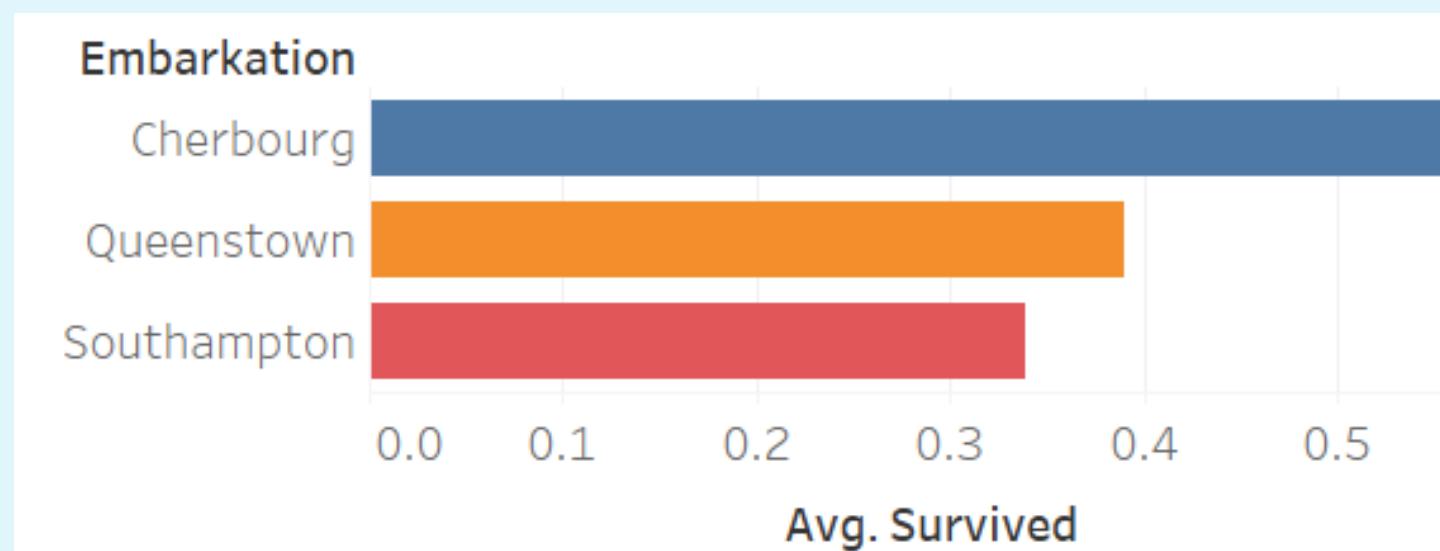
Additional Information: Embarked Field

- Data shows that someone from "Cherbourg" had a higher chance of surviving than those from "Queenstown" and "Southampton"
- Most passengers boarded from Southampton (73%)
- In a real life context, where a passenger came from would have had **no impact on survival rate**, as any region could purchase higher class tickets

Passengers by Embarkation



Survival Rate by Embarkation



Survival Count by Embarkation

Embarkation	Count
Cherbourg	168
Queenstown	77
Southampton	646