**Project Name:** Data Visualization

**Group 3:** Seth Phillips, John Allard, Yousaf Khaliq, Clay Skiles

**Dataset:** RMS Titanic (Kaggle)

**Summary:** The Titanic dataset consists of 891 entries with 12 variables, providing information about passengers on the RMS Titanic. The key variables include survival status (38.4% survived), ticket class (majority in 3rd class), sex (64.8% male, 35.2% female), age (mean of 29.7 years, ranging from 0.42 to 80), number of siblings/spouses and parents/children aboard, ticket number, fare (mean of 32.20, with a maximum of 512.33), cabin number, and port of embarkation (majority from Southampton). Notably, the dataset contains missing values in the 'Age' (177 entries), 'Cabin' (687 entries), and 'Embarked' (2 entries) columns. Most passengers traveled alone, and the majority of them embarked from Southampton. The dataset provides a rich source of demographic and socioeconomic information, useful for predictive modeling and survival analysis.

**General Information:**

* Total entries: 891
* Total columns: 12
* Data types:
  + Numerical: PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
  + Categorical: Name, Sex, Ticket, Cabin, Embarked

**Missing Data:**

* Age: 177 missing values
* Cabin: 687 missing values
* Embarked: 2 missing values

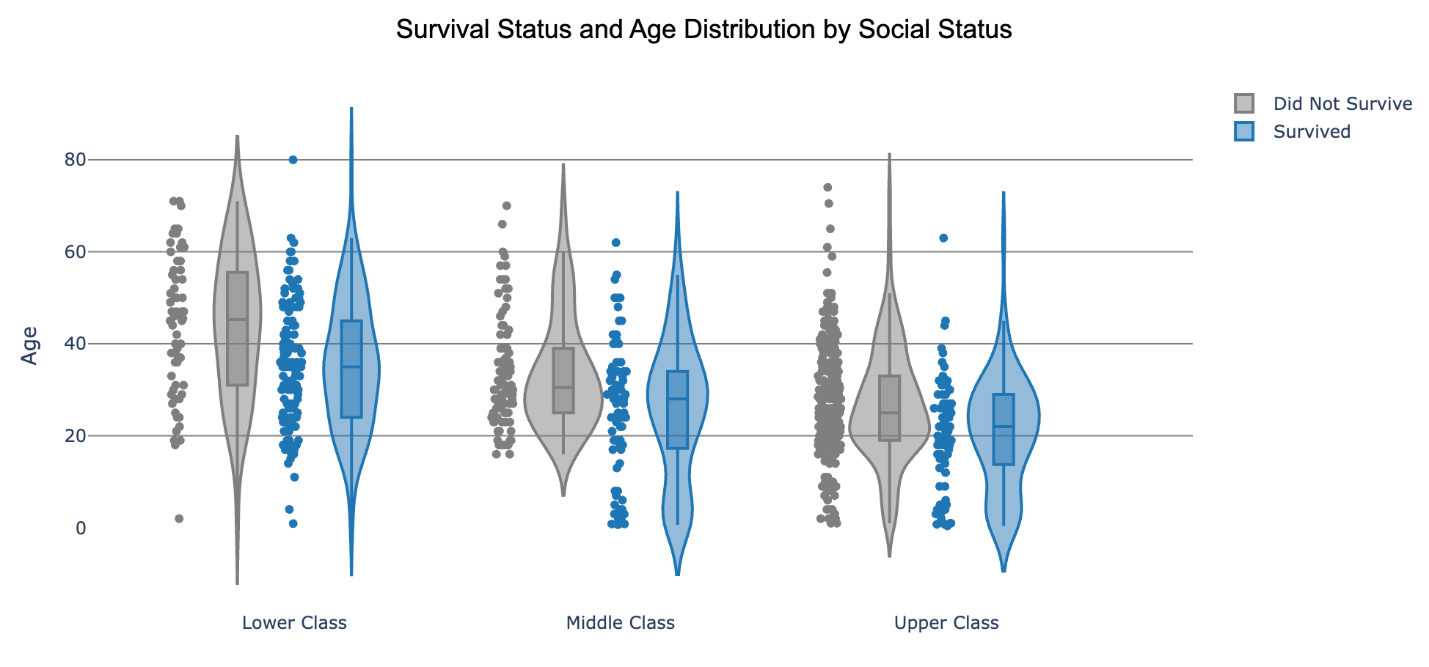
**Descriptive Statistics:**

* **Survived:**
  + Mean: 0.384 (38.4% survival rate)
* **Pclass:**
  + Most passengers were in 3rd class.
* **Age:**
  + Mean: 29.7 years
  + Min: 0.42 (infant)
  + Max: 80
* **SibSp (Siblings/Spouses Aboard):**
  + Most passengers traveled alone.
* **Parch (Parents/Children Aboard):**
  + Most passengers did not have parents/children aboard.
* **Fare:**
  + Mean: 32.20
  + Max: 512.33
* **Embarked:**
  + Most common port: Southampton (S)

**Categorical Overview:**

* **Sex:**
  + Male: 577 (64.8%)
  + Female: 314 (35.2%)
* **Embarked:**
  + S (Southampton): 644
  + C (Cherbourg): 168
  + Q (Queenstown): 77

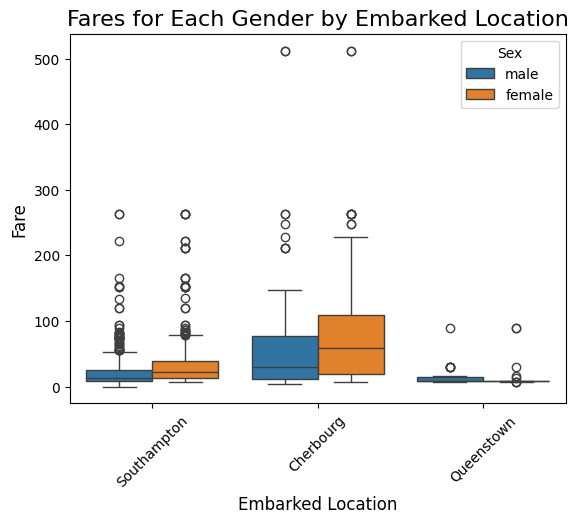
**GOOD VISUALIZATIONS**



Summary:

The violinplot is illustrating the distribution of passenger ages on survival status. This plot is a good example of visualizing a dataset because it effectively communicates the shape and distribution of passenger ages. It visualizes where most data points are concentrated while still showing median values and outliers. The plot is labeled well and maintains clear axis labels and appropriate scales. The plot seems to follow best practices when constructing a visualization by having good color choices, good scaling, and non-misrepresentation in data.

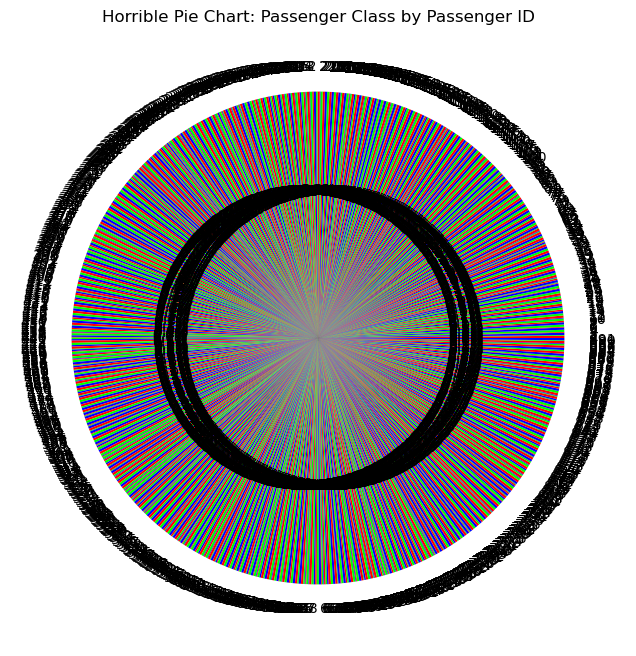
This graph is effective at showing us significant patterns in survivability. We notice that younger people (such as children) survived at a much higher rate, and we can see that social class played a rather significant role in survivability, too. A finding in the distribution for younger people is the noticeable bulge in the graph. This seems to reinforce the well-known survivability pattern of “women and children first.”



Summary: This boxplot demonstrates the differences in social classes amongst those who boarded the Titanic based on Embarkment Location. It is clear that those who boarded from Cherbourg carry the most amount of wealth and financial stability as the median ticket fares and IQR possess greatest values of the three while Queenstown clientele paid for generally the cheapest fares. The plot also shows the differences in fares amongst the genders who boarded the Titanic for each Embarkment Location. It is evident that in Southampton and Cherbourg that women generally paid higher fares than men, whereas it is the opposite in Queenstown.

The visualization itself is effective to notice these observations for various reasons. The boxes are distributed evenly across the Embarkment locations with the hue of sex color-coding the boxes noticeably different between each gender. The axis values for Embarkment location are also rotated 45˚ to discern the axis title and the values, in addition to said values being elaborated with the full title of the location for more accessible reading. The chart also aims takes advantage of the real estate of the variables looking to be analyzed, since the plot investigates the relationship of the fare costs of two types of passengers boarding from three different embarkment locations. This is a lot of information that could be distributed amongst multiple charts, but the above chart condenses these findings into one visualization.

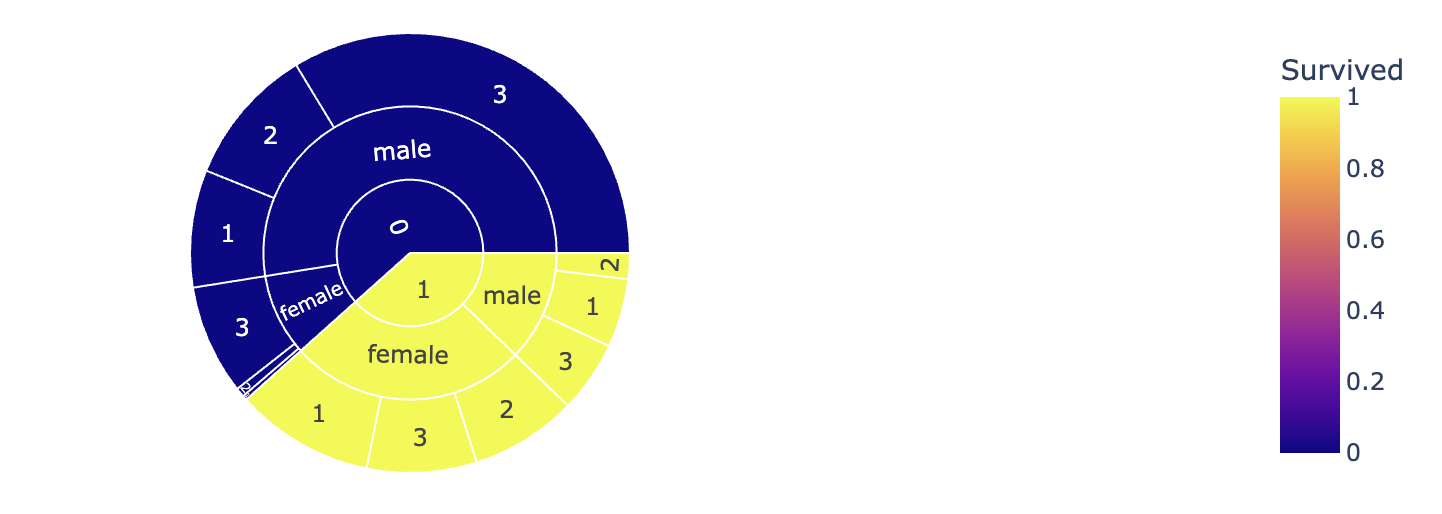
**BAD VISUALIZATIONS**



**Summary**:

This pie chart is a bad visualization due to its ineffectiveness to communicate key data findings in the dataset. This is due to clear overcrowding and excessive fragmentation. The pie chart attempts to display data for each individual passenger leader to unreadability, visual noise, and indistinguishable slices. This is a very overwhelming visualization for any viewers.

To make the chart better, some options could be to group the individuals by survival status, gender, or some other category to attempt to reduce visual clutter. Generally, pie charts effectively compare only a few categories (ideally no more than six). For comparisons involving more groups, a bar chart would likely be a more suitable choice in this scenario.

Another option if one wishes to investigate the relationship between passenger class by passenger ID is to consider another visualization type. With Passenger ID being an unimportant identification variable to discern unique entries of the dataset, it could be possible to take counts of those in each passenger class based on passenger ID with a bar chart. That way it is clear to see the differences in each passenger class of the numbers of individuals who boarded the Titanic without over-cluttering the visualization with an overwhelming count of unique ID variables.

**Summary:**

This is a sunburst chart with an impractical color scheme and unintuitive labels for ‘Survived’ and ‘Pclass’. The plot automatically generates a legend for the initial ‘Survived’ column, despite the fact that the only 2 options are ‘Survived’ and ‘Did not survive’. Also, ‘Pclass’ isn’t specified anywhere on the plot, so the final ring of the chart only show the values 1,2, and 3, with no other context.

The sunburst chart is a good example of a bad visualization due to its poor labeling, unnecessary complexity, and unclear structure. Sunburst charts are useful when you have clear relationships defined for each category, but in this case the sunburst chart fails to make survival rate by passenger class interpretable.

The chart ineffectively uses color coding and labeling. One major issue is the legend, which is redundant since there is only two possible outcomes. The legend adds noise to the visual and is distracting. Another key issue is that there isn’t any clear labeling for passenger class. The ring displays values (1, 2, and 3) without any sort of key to identify what these values represent.

Visually, the chart is confusing and ineffective. While it attempts to represent gender, passenger class, and survivability, it fails to communicate this information clearly.