# Datathon Semester 2 Preparation Workshop

Monash Data Science Society | 2023

# Today's Agenda

1. Introduction to the MDSS Datathon

2. Overview of Data Cleaning/Wrangling

3. Data Visualization

4. Q&A

# What is a Datathon?

Collaborate, innovate, and demonstrate data science skills to analyse and visualise data in a insightful and creative way.

This event is for both seasoned data experts and beginners

# Data Cleaning/ Wrangling

{mdss}

MDSS
DATATHON
WORKSHOP

# Data Cleaning Basics

Data cleaning refers to the process of identifying and correcting errors, inconsistencies, and inaccuracies in datasets.

1. Handling duplicates
2. Handling missing values
3. Handling outliers
4. Handling inconsistent data

# Data Wrangling

It is the process of transforming and reshaping raw data into a format suitable for analysis.

1. Data acquisition
2. Data cleaning
3. Data integration
4. Data reshaping
5. Data reduction
6. Data validation

Repeat as many times necessary

# Notebook Demonstration

# Data Exploration

{mdss}

MDSS
DATATHON
WORKSHOP

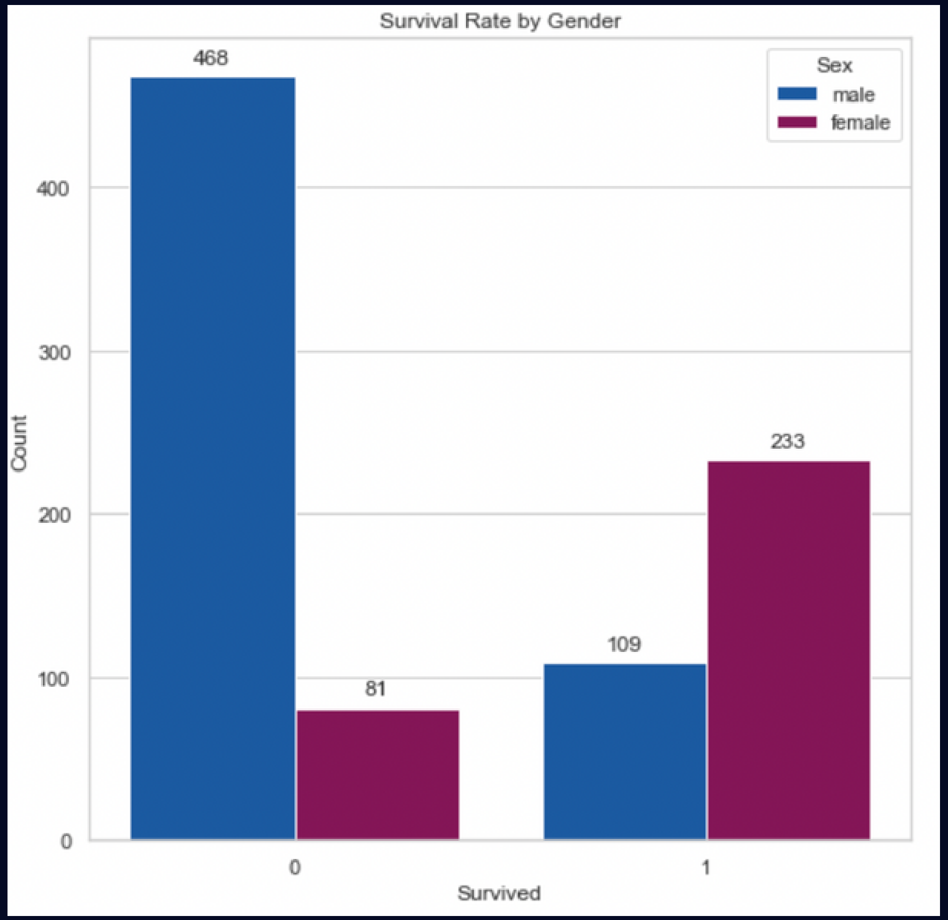
# Exploration

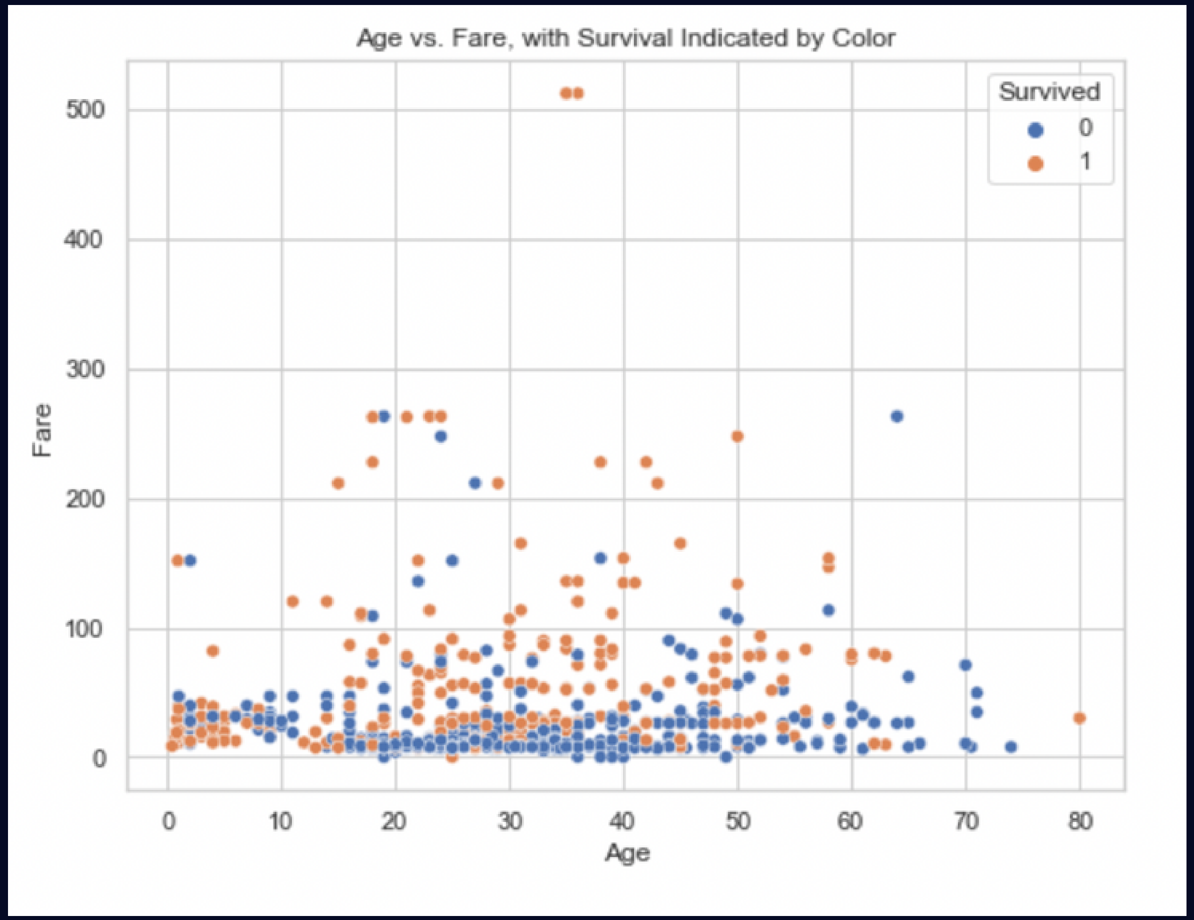
## Common techniques:

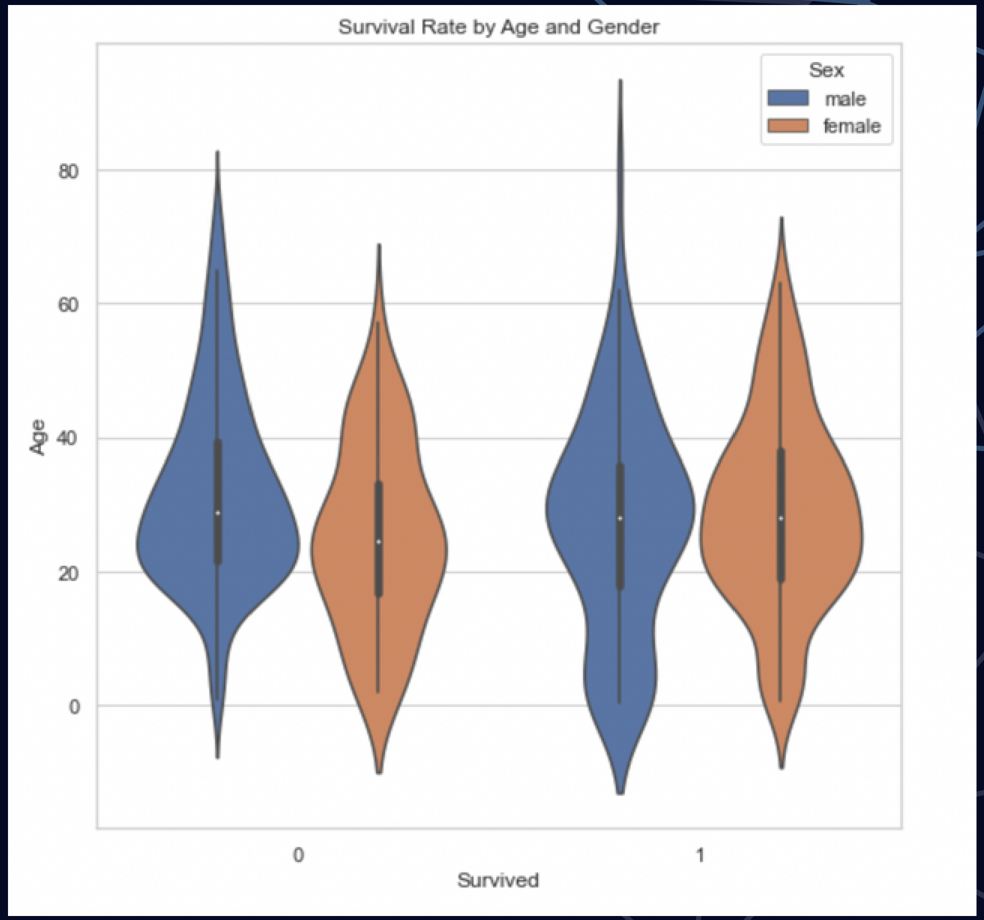
- Summary statistics - basic statistics, such as mean, median, and standard deviation.
- Data visualization - creating graphs and charts, such as histograms, scatterplots, and box plots, to visually explore the data and identify patterns and relationships.
- Data transformation - converting or manipulating the data, such as scaling or normalizing variables, to better understand the relationships between variables.
- Dimensionality reduction - involves reducing the number of variables in the data, such as using principal component analysis (PCA), to simplify the analysis and identify the most important variables.

"the process of examining and understanding the data before performing any formal analysis."

# Example Graphs







These visualisations are created using the famous Titanic ML competition dataset - which we will explore shortly!

# Notebook Demonstration

MDSS
DATATHON
WORKSHOP

# Data Set Used

https://www.kaggle.com/c/titanic

# Q&A