DATASCI W200 Project 2 Proposal

Names of team members: Kelianne Heinz, Ivy Chan
Name of your team's GitHub repository: Project2_Chan_Heinz

Introduction
The RMS Titanic, a luxury ocean liner, was considered unsinkable due to a series of compartment doors that could be closed to maintain buoyancy. However, in its maiden voyage on April 15, 1912, it sank after colliding with an iceberg in the North Atlantic, killing 1502 out of 2224 passengers and crew. This infamous tragedy inspired many stories, films, and musicals and numerous myths arise from it. In this project, we will perform exploratory data analysis (EDA) on the raw dataset to illustrate demographic data distribution and the survival rate by various parameters such as age, passenger class, and sex.

Primary Dataset
The primary dataset we intend to analyze is Titanic passenger data. It is available at https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv

This dataset contains the following fields:

Column Name (col = 14)	Description	dtype	null values (rows = 1309)
pclass	Passenger class (1, 2, 3)	int64	0
survived	Boolean; 1=survive, 0=died	int64	0
name	Name	object	0
sex	Sex (Male, Female)	object	0
age	Age (0.17 to 80)	float64	263
sibsp	Num. of siblings/ spouses on board	int64	0
parch	Num. of parents/ children on board	int64	0
ticket	Ticket number	object	0
fare	Fare (0 to 512.3292)	float64	1
cabin	Cabin number	object	1014
embarked	Location embarked; Southampton (S), Cherbourg (C), Queenstown (Q)	object	2
boat	Lifeboat number / ID	object	823
body	Body Identification Number	float64	1188
home.dest	Home Destination	object	564

With this exploratory analysis, we intend to answer the following questions in our final report:

- **Who was on board the Titanic when it sank?** To answer this question, we will investigate the demographic information available on the Titanic passengers, including the age distribution, gender distribution, and class level of passengers. We will also look at what passengers were travelling with relatives, and what passengers were travelling alone.

- **Where were these passengers coming from, and where were they going?** We will look at the passengers' embarking location and home destination data to see where people were travelling on the Titanic.
- **What demographics were most likely to survive the sinking of the Titanic?** We intend to explore how passenger demographic factors such as gender, passenger class, and party size (travelling with relatives vs. travelling alone) related to the likelihood of a passenger surviving.
- **Of passengers who did not survive, is there any pattern to whose body was most likely to be identified?** We will explore the relationship between a passenger's demographic data and whether their body was identified or not.

In order to carry out this EDA, we will look at the pclass, survived, sex, age, embarked, body, and home.dest columns of this data. We will also derive new fields based on existing fields. For example, we can derive the passengers who have relatives travelling with them by combining sibsp and parch. We will also handle missing values in some columns.

In the final report, we will create charts which illustrate the data distribution and relationship of various fields such as pclass, age, gender, and survived and answer the questions listed above. These will include:

- Figure for survival by passenger sex
- Bar chart for passenger class vs survival
- Line chart for passenger sex vs survival
- Figure for single traveller vs family traveller

Initial plots, figures, or tables: