**Group Members:**

Renzo Bejarano Varela | **Arron Nguyen**| Callum Linnegan | **Carmen Jiaman Li** | **Nicklaus Ng** |**Tealiie**

**Project Title:** Using a machine learning model to forecast a life expectancy for a country.

**Objective:** Can we correctly predict life expectancy? And, if so, which features are most relevant for prediction?

In this project, we examine life expectancy as output for our machine learning app.

* **DATA SOURCES**

-World Bank Data Bank. World Development Indicators.

- https://www.kaggle.com/

**TRANSFORM** - Proposed clean-up and analysis

• What are the transformations you will apply to the data? (e.g. filtering, aggregation, derived columns) Filtering and derived columns • What steps will you take to clean the data and ensure its validity (e.g. messy data, duplicated data, incorrectly formatted data) Deduplication, format revision and cleaning values. • How will you identify potential issues with your data sources? (e.g. exploratory data analysis, simple statistics etc) Exploratory data analysis • How will the data be integrated? (e.g. joins, merges) Merges • How will you apply these transformations (e.g. jupyter notebook, python script) Jupyter Notebook • IMPORTANT → Why did you apply these transformations? How did this enrich your data? We have chosen these transformations as they best suit the data we have selected.

* **DATABASE**

• What type of database (relational, document) will you store the data? Relational database • Why did you choose this database over another database? We chose this database as it will allow us to easily apply analytical functions and derived data • What are your expected tables / documents and relationships between tables / documents in your database? Our tables will be Book data, Review Ratings, Authors

**Potential limitations •** What are the potential limitations of your above proposed steps? ~ The potential limitations we can run into is the dataset having limited data than expected ~ There could be missing fields that we need

• How can you control these potential issues? ~ We will explore different sources of data.