Monash University

Data Analytics Bootcamp

ETL - PROJECT PROPOSAL  
Task List

\_\_\_

# Team Members Rubric Preferences

* Hayley W (HW) **(PP)**
* Tuflikha P (TP) **(??)**
* Hatim M (HM) **(??)**
* John N (JN) **(PR)**

# Rubric: Project 2

## Project Proposal (PP)

* Proposal cites minimum 2 sources
* Includes production type DB
* Description of findings

## Project Report (PR)

* **Extract:** Original data source and formatting
* **Transform:** Data cleaning/transformation required
* **Load:** Explains the final DB, tables/collections and end goal

## GitHub (GH)

* Uploaded all to GitHub
* Clean repo of redundant/unused files/folders
* The read me is written in to a professional level

# Project Summary

The project is designed to create a database for the HR department at Youtube by extracting, transforming, and loading the world happiness index report from the last thirteen years from 2008 to 2020.

The HR team wanted a stable database for the reporting dashboard to be able to pull data, demonstrate the changes of index through time and as well to track the results of it.

## 

## Global Population Happiness Levels

4 Data sources that have been chosen are related to:

* Global Population levels 2020
* historic World Happiness Index scores from 2005 (2 Sources)
* Wikipedia iso country codes

# Group Project Dynamics

* 1 dedicated person for GitHub Repo
* 1 dedicated person for Project Report
* 1 dedicated person for PG Admin Schemata
* Project Vision and Data Sources collectively agreed to **early**
* Once Structure is agreed upon team can flesh out data cleansing and divide mini tasks to each member

# Data Sources

## world-happiness-report.csv

## world-happiness-report-2021.csv

## population\_by\_country\_2020.csv

**Link to source data:**

https://www.kaggle.com/joshuaswords/awesome-eda-2021-happiness-population/data

# ETL Project Steps

## Brain Storming Stage

* ~~Group to agree on project focus~~
* ~~2+ data source agreed upon~~
* ~~Group Github repo established~~
* ~~Understand data collectively~~
* ~~Establish required data cleaning/transformation~~
* ~~Write prototype schemata/table nomenclature~~

## DB Set Up

* Set up and load data in to DB
* Extract ERD for reference in Project Report
* Short Summary description of table structure to be written

## Project Report

* Summary written of data sources and background
* Short Summary description of table structure to be written
* Cleaning/transformation methods chosen/why
* Final tables/collections schema described in DB (reference ERD)
* Obstacles overcome/future analysis for Project for Database

**NB:** Above potential project is to use Happiness data and contrast and compare youtube views/trends during 2020-2021 pandemic period

<https://www.kaggle.com/rsrishav/youtube-trending-video-dataset?select=BR_youtube_trending_data.csv>

## Github

* All files used uploaded:
  + ~~Raw data sources~~
  + ~~Cleaned up data source~~
  + ~~Jupyter Notebook (if used for cleaning)~~
  + ~~SQL Script that sets up DB structure~~
  + ~~ERD Diagram of Database~~
  + ~~Project Proposal~~
  + ~~Readme.md~~
  + Presentation (potentially)
* GitHub Repository cleaned up
* Remove unused/redundant data/folders
* Uniform/Professional naming conventions/structure
* Readme.md file updated with report/proposal

# Data Source Structure

## population\_by\_country\_2020.csv

* Country (or dependency)
  + This column contains different country's name (235 countries)
* Population (2020)
  + This columns contains the population of different countries
* Yearly Change
  + This columns contains the population change by yearly
* Net Change
  + This columns contains the net change of the population
* Density (P/Km²)
  + The column contains the density of the population
* Land Area (Km²)
  + This column contain the land area in terms of kilometer square
* Migrants (net)
  + This column represents the migrants of the countries
* Fert. Rate
  + This column represents the fertility or the growth rate of individual countries
* Med. Age
  + This column represents the median age (Middle Age or the average age) lifespan of the country
* Urban Pop %
  + This column represents the urban population

## world-happiness-report-2021.csv

* Country name
  + Name of the country
* Regional indicator
  + Region the country belongs to.
* Ladder score
* Standard error of ladder score
* upperwhisker
* lowerwhisker
* Logged GDP per capita
* Social support
* Healthy life expectancy
* Freedom to make life choices
  + The extent to which Freedom contributed to the calculation of the Happiness Score
* Generosity
* Perceptions of corruption
* Ladder score in Dystopia
* Explained by: Log GDP per capita
* Explained by: Social support
* Explained by: Healthy life expectancy
* Explained by: Freedom to make life choices
* Explained by: Generosity
* Explained by: Perceptions of corruption
* Dystopia + residual

## world-happiness-report.csv

* Country name
  + Name of the country
* Year
  + Region the country belongs to.
* Life Ladder
* Log GDP per capita
* upperwhisker
* Healthy life expectancy at birth
* Freedom to make life choices
  + The extent to which Freedom contributed to the calculation of the Happiness Score
* Generosity
* Perceptions of corruption
* Positive affect

## wikipedia-iso-country-codes.csv

* English short name lower case
  + Name of the country
* Alpha-2 code
* Alpha-3 code
* Numeric code
* ISO 3166-2
  + ISO Code reference for Country