**ETL Project Proposal**

**Team Number:** 6

**Team Name:** Ryans boys

**Team Members:**

1. Ash Drayton
2. Alex Lawson
3. Amil Walgamage
4. Rishit Sheth

**Project overview**

Covid-19 has been completely taking control of our day to day lives in 2020, hence we decided to analyze the impact of healthcare expenditure of countries on the number of deaths because of Covid-19 in those countries considered.

**Our question of interest**

Does healthcare expenditure impact covid related death count?

**Outcome of your ETL project**

We expect that countries that spend a higher percentage of their GDP on health resources will have lower fatality rates in response to the rising covid-19 cases.

**Expectation**

Our project will provide those interested with a concise data set comparing COVID related deaths to governments health expenditure, to show how well countries might have been prepared for the epidemic as well as how each country prioritizes the health of the population.

**EXTRACT** - **Data sources**

1. Healthcare expenditure of OECD countries in .csv file (limited countries)

<https://stats.oecd.org/Index.aspx?ThemeTreeId=9>

1. Covid deaths and confirmed cases worldwide (data limitation 22/01/2020 - 12/08/2020)

<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/data?select=time_series_covid_19_deaths.csv>

<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/data?select=time_series_covid_19_confirmed.csv>

1. Web Scrapped population data from

<https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)>

**TRANSFORM** - Proposed clean up and analysis

* Covid data set needs to aggregate the death count for each day until the cutoff date derived columns
* Data is up to date, but each column has accumulated values.
* Remove duplicates, check dtypes, sense check with column counts and visualize with scatter plots.
* How will you identify potential issues with your data sources? (e.g. exploratory data analysis, simple statistics etc)
* We used .describe() and visualized with py.scatter to identify potential issues with data sources.
* Merge was used to integrate datasets and left join on ‘country name’ was used to integrate tables.
* The transformation was done using Jupyter Notebook and ERD was done on <https://app.quickdatabasediagrams.com/>
* Transformation helped to convert the data into more meaningful format to start our storytelling.

**LOAD** - Data storage

* We stored our data using a Relational database - PostgreSQL.
* PostgreSQL is simple and clean to use as a db tool.
* Three tables in ERD:

1. Country/% of gdp on health
2. Country/Covid Deaths/Confirmed cases
3. Country/Region

**Potential limitations**

* Two potential limitations of our above proposed steps are:

1. Naming conventions in countries; for example, some countries have multiple names (East Timor and Timor Leste). This will affect the join process and querying the data.
2. We are limited by the total information available, as not all countries have a resource with updated and correct health expenditure.

* To overcome these potential issues, we are using the countries that we have available to analyze, thus we will mention during the conclusion that only some countries had sufficient information for us to use.