|  |
| --- |
| PROJECT PROPOSAL - ETL PROJECT  Diversity – Household Income – Unemployment  Data Sets |
| PRAGATHI MADUSHINIE |



**Relationship Between Income Unemployment Diversity**

**Project Proposal for ETL Project**

INSTRUCTIONS

Summary page:

a. Project title: ETL Project Income Unemployment Diversity

b. Name of implementing person(s): Pragathi M Porawakara Arachchige

c. Project location: Data Analytics Boot Camp - WashU

d. Proposed starting date: 07/09/2019

e. Project duration: 1 week

The Project Proposal for ETL Project has three sections:

I. Extracting data from different data sources

II. Transform the data to a presentable format

III. Load them to a sql database

I. EXTRACTING DATA FROM DIFFERENT DATA SOURCES

This section should provide a brief introduction on how to find different data sources and make a relationship in between them. The main reason behind that is, finding a primary key or a composite key which is a combination of two unique keys, is crucial in this step.

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data: data. World and Kaggle.

You can also use APIs or data scraped from the web.

The first step should also describe:

a. Find different data sources

b. Look for different data sets which needs in the project

c. Define the primary key or the composite key

d. Look for more relations or relational data sets which will give a meaning to the project

e. Finalize the data sets

f. Finalize what you would like to implement in this project

g. Time to extract the selected data sets in to python

h. Extract all these data sets to python and check the data

Finally, the extract step should include, the defined data sets, primary keys and the what the final project objective.

II. TRANSFORM DATA IN TO A PRESENTABLE FORMAT

This section should discuss proposed extract methods and data sets, and also focus on transform that data in to a meaningful format. Data clean up and analysis is the main focus in this section. That includes cleaning the data sets, renaming the column names, setting the indexes, extracting just the needed columns & etc.

III. LOAD THE CLEANED DATA SET IN TO A SQL DATABASE

Specifically, the discussion should indicate how will the cleaned transformed data load in to a database. It can be either relational or non-relational. This is the last step of the ETL project.

## Project instructions

### **Finding Data**

### Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data:

### data.world

### Kaggle

### You can also use APIs or data scraped from the web. However, get approval from your instructor first. Again, there is only a week to complete this!

### **Data Cleanup & Analysis**

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

* The sources of data that you will extract from.
* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).
* The type of final production database to load the data into (relational or non-relational).
* The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

### **Project Report**

### At the end of the week, your team will submit a Final Report that describes the following:

### Extract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

### Transform: what data cleaning or transformation was required.

### Load: the final database, tables/collections, and why this was chosen.

### Please upload the report to GitHub and submit a link to Bootcampspot.