UCI Bootcamp Project 2  
etl project proposal

September 2021

# Overview

## Project Background and Description

|  |  |
| --- | --- |
|  | Good quality data is important to organizations and businesses, as data is an asset that organizations cannot overlook to generate revenue, project growth and make sound business decisions. Organizations that have put data available to them to good use have grown in leaps and bounds. The Extract, Transform and Load (ETL) process is used by organizations to guarantee data quality. The ETL process helps organizations ingest data from multiple sources, formats. Clean, shape, filter and output this data in a format that can be used by the organization.  For this project, our team will extract, transform, and load dataset into a relational database. We have identified the Tokyo 2021 – Olympic data files downloaded from Kaggle.  **Team Members are:**  Jasper Chou  Galyna Malynska  Michael Liu  Banke Ogunjumo |

## Project Scope

|  |  |
| --- | --- |
|  | The scope of this project will include researching and identifying datasets that can be used to complete the Extract, Transform and Load – ETL process. Our team will extract data from different sources, transform these data by cleaning, joining, filtering, aggregating, deduplicating etc. Also, the data will be loaded into a relational database.  **EXTRACT**  In the extract phase of this project, we will research datasets of interest, inspect, and collect dataset to use.  **TRANSFORM**  In this process, business rules will be created based on business requirements and need for the dataset. Then, the data collected will be shaped based on business rules.  **LOAD**  This phase involves creating an Entity Relationship Diagram (ERD) for tables created in the transform stage and storing these tables into a relational database. |

## Deliverables

|  |  |
| --- | --- |
|  | **ETL Project deliverables**  Project Proposal  Data Sources – Two Data Sources  Complete ETL and create documentation   * Data Extracting from <https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo> * Data Business Rules * Data Transformation in Python * Data ERD * Data Storage to PostgreSQL an SQL Relational Database   Data Mapping documentation – Source data, Target data and Business rules |

## Specific Exclusions from Scope

|  |  |
| --- | --- |
|  | Considering project scope and time constraint, we are rolling out project at once with all team members assigned specific stages of the project to complete. |

## High-Level Timeline/Schedule

|  |  |
| --- | --- |
|  | **Day 1**   * Discuss and outline project details * Research and identify data sources and datasets * Create Project proposal * Project proposal approved * Research and identify data sources and datasets * Assign tasks   **Day 2 – Day 5**   * Check-ins * Communicate milestones and bottlenecks * Resolve bottlenecks * Project proposal approved * Research and identify data sources and datasets   Day 6   * Meet to discuss milestones * Make changes and corrections * Finalize project for submission   **Day 7**   * Submit project |

# Approval and Authority to Proceed

We approve the project as described above, and authorize the team to proceed.

|  |  |  |
| --- | --- | --- |
| Name | Title | Date |
| Anthony Taylor |  | 09/02/2021 |
| Mark Flynn |  | 09/02/2021 |
|  |  |  |