|  |  |
| --- | --- |
| **Name;** | **Odhiambo Nancy Anyango** |
| **School;** | **Grand Valley State University** |
| **Department;** | **Computing and Informatics** |
| **Major;** | **Data Science and Analytics.** |
| **Course Title;** | **Data engineering, CIS 660** |

***Project title*: *Assessing comorbidity and demographic influence on covid-19 patients outcome from Mexican government covid-19 dataset from kaggle through automated ETL, data analysis and visualization.***

## **1.ProjectOverview This project mainly aims at creating a complete and automated ETL workflow, data analysis and visualization for the covid-19 data set from the Mexican government from kaggle. The extraction of the dara set is directly automated from kaggle site using the kagglehub api key, transformation executed in Kestra platform using python scripts to check various data characteristics and fit them to meet the required features for the project and finally the loading to PostgreSQL. This data will then be used in analysis and visualization in VS Code and Looker studio to come up with insights and recommendations. The project intends to compass these various data engineering and data science application softwares to derive insights to improve health care and public health sector management of the pandemic in the future and any other related outbreaks by implementing the same strategies.**

***Key Questions Addressed:***

1. **Which comorbidities have the strongest association with COVID-19 mortality?**
2. **How does age interact with pre-existing conditions to affect outcomes like death, in-tubing and referral to icu?**
3. **Can we reliably predict mortality risk at hospitals?**
4. **What is the correlation matrix and co-occurrence of comorbidities and health status of patients in the data set?**

## **2. Data Source Selection**

* **Dataset Choice**

**In my project I will use the covid\_19 data set from the Mexican government and directly used from kaggle using the kagglehub. Data set link <https://www.kaggle.com/datasets/meirnizri/covid19-dataset>**

* **Selection criteria and justification**

The data set is directly related to health sector, specifically public health and I intend to focus on how the comorbidities, demographic information and the health outcome such as intubed/ icu in this data set are directly connected to a patients outcome i.e. The status feature which is whether a patient is dead or Alive.

The data set has a total of 21 unique features and 1,048,576 unique patients however I intend to subset it to only have patients that have their classification status as 1-3 (1 - mild case of covid-19, 2 - Moderate case of covid-19, 3 - Severe case of covid-19 and 4 & above for those whose tests were either inconclusive or not carriers of covid-19. I decided to subset this data to mainly focus on those patients that are exclusively affected by covid-19 to b**e able to visualize their patterns and health status. Some of the variables that will be of less significance in the project will as well be dropped in the course such as the medical unit and usmr.**

**The data set has considerable complexities ranging from missing values encoded as 97, 98 and 99 (97 was an outlier value in the categorical features and i made an assumption that it could be signifying the unknown entries from patient feedback which is mainly the encoding used in health systems), the features are somehow encoded in inconsistent format and some encoded in binaries of 1 & 2 which makes it difficult to work with it since most of those variables are discrete. The age feature has some noticeable outliers which either means that some elderly people were also in the data base. The intubed and icu features have noticeable amount of missing values my inference at the moment is that they could have filled the 97, 98 and 99 for those patients that were not hospitalized which made it that ambiguous.**

**This data set was last updated 2 years ago and at the moment does not have any update frequency, I mainly chose it to analyze and visualize impact and derive future recommendations to support the health department and public health in arresting covid-19 situations in Mexico and any other country by enhancing preparedness and probabilistic modeling which I might not apply within the scope of this project.**

## **3. Technology & Tool Requirements**

* **ETL Tools & Platforms**

For the ETL I intend to use five softwares i.e. VS Code, Kestra, Postgres, pg4admin and Docker. VS Code will mainly be used at the initial stages of initializing the kestra and then directly orchestrate my data into kestra platform for automation and managing the ETL pipeline, Postgres and pg4admin will be used for creating a database and running the SQL queries and finally **Docker** for containerization and ensuring consistency across kestra and postgres environments.

* **Database & Storage Options**

**At the moment I settling at using PostgreSQL for the database and storage and querying since. I picked the option for its stability and comfort in use since I have few issues with operating the GCP. This an option that I might reconsider in the process in a case that the GCP issues are resolved and I am finally able to swiftly and able to use the BigQuery platform.**

* **Visualization & Monitoring Tools**Use open-source tools to track pipeline health and visualize business insights
* **Monitoring**

For monitoring since kestra **is capable of monitoring workflows and providing insights into the execution of data pipelines which allows tracking of the status of tasks, detect failures, and ensure that pipelines are running as expected I will stick to using it.**

* **Data Visualization / BI**

My features interactions, comorbidity analysis and outcome analysis will be done through Looker studio which will be directly getting the data from the PostgreSQL. In a case that the GCP issues are resolved I intend to use the Looker studio with data directly from BigQuery platform.

## **4. Implementation Approach**

* **Data Extraction**

The data set is in CSV format and will be directly orchestrated into Kestra using Kagglehub through API key. This will allow for the automation of the workflow process.

* **Data Transformation**

**From the above analysis of data issues from the data set or the subset that I intend to work with I have the intention to check out for all the missing values, calculate their ratios from the total number of patients and if it is less and cannot impact insights I will remove them from the analysis, for the outliers too i will check their occurrence and how significant they are and remove. Looking at the binary features, to make it easier to visualize, I will change the 1 to Yes and 2 to No for the respective features and during the correlation analysis I will encode them to 1 for Yes & 0 for No the standard binary encoding. For the status I will change all that have the date of death to Dead and those with the alternative i.e. 9999-99-99 to Alive for ease in interpretation.**

**The summary statistics for the age feature will be calculated to give clear picture of the mean age, median age, SD, minimum age and maximum age of the group of patients to derive insights on distributions of per comorbidity and the comparison of the outcome for the status depending on the age differences.**

The libraries to be used for data transformation will be pandas and numpy with techniques like IQR for calculating the the quartile ranges and removal of outliers.

* **Data Loading**

**The data will be loaded into PostgreSQL by using Kestra’s SQL task. This SQL task will involve creating a table named covid\_data and inserting the features(variables) in the PostgreSQL. After this process the rows in the data set will be inserted in an iterated manner to add into the columns. Finally the PostgreSQL will be queried using the SELECT \* FROM covid\_data LIMIT 10; to assess whether all columns have been added and the relevant entries made in the rows. It will also be important to count the total number of values in a column to assess whether all rows have been inserted.**

* **Schema Design:**

The data consists of 21 features with only 1 numeric feature which will be subsetted to only have the most relevant features for the project. The rest of the features are of type binary i.e. Yes/No and multinomial which makes them all discrete. There is only 1 large file of type CSV and the data is anonymized with only the number of patient per row and no name or ID used in the data set as a key feature. For joins since the intubed and icu data might show some correlations , I will join the two columns to understand the number of patients who are both intubed and in icu and their probable status (outcome)

* **Data Analysis & Visualization**

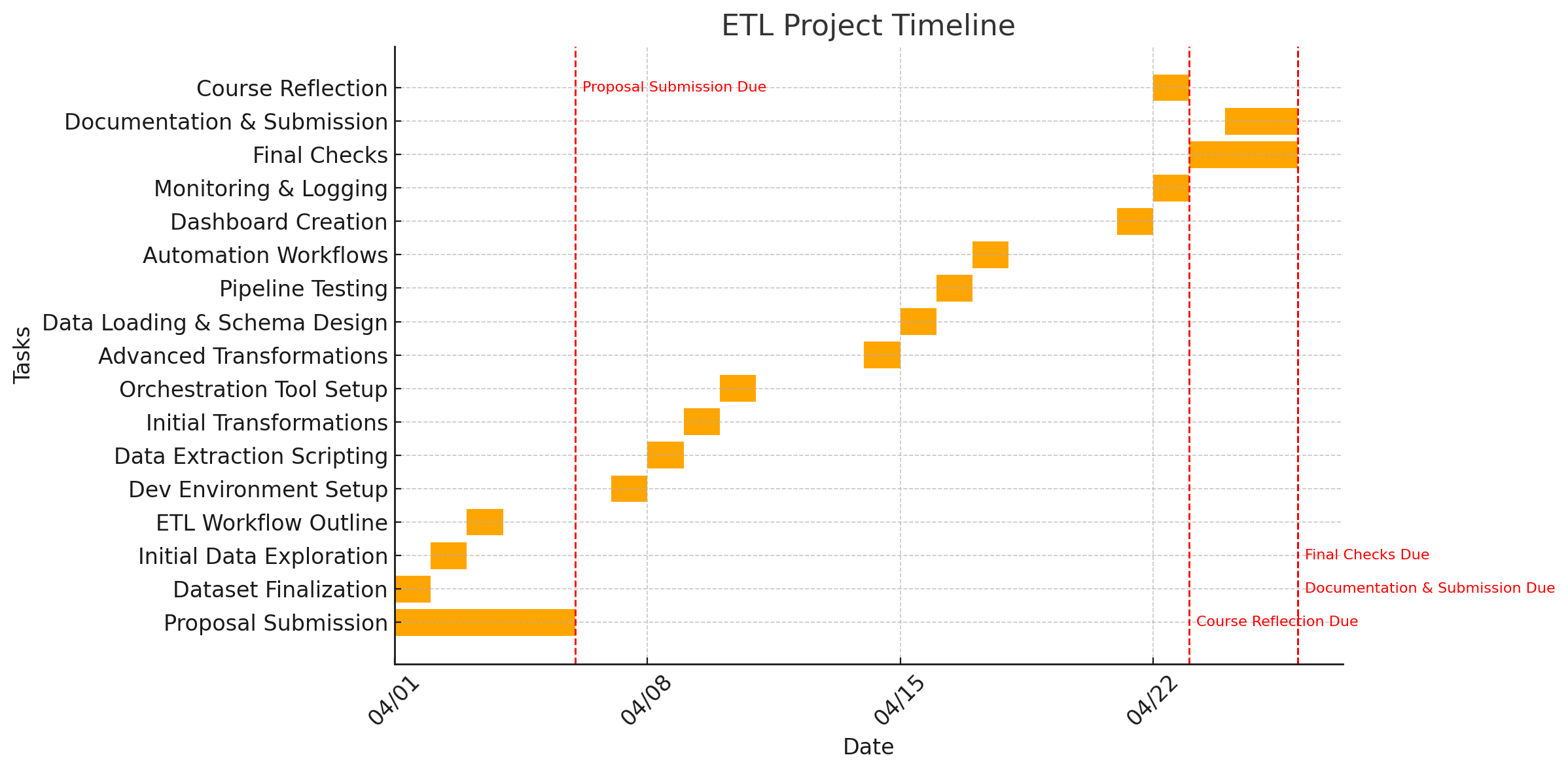
After the data is loaded into Postgres an instance of PostgreSQL will be created in the cloud to allow the ease in data retrieval to make the visualizations and the insightful dashboard in the Looker studio platform. The project will also use VS Code specifically python, (Jupyter notebook) to perform exploratory data analysis aside from the queries written in the PostgreSQL to come up with a good EDA analysis for the features and load these to Looker studio to compliment the visualizations done in the platform like the correlation matrix and comorbidity analysis and cooccurrence in patients.

## **5. Project Timeline**

This timeline is structured to guide your project to completion within the project window. Each week focuses on key ETL pipeline stages, from planning and setup to deployment, visualization, and documentation. Progress should be iterative, with refinement and testing occurring alongside development to ensure quality delivery by the end of Week 4.

| **Week** | **Milestone** | **Description** |
| --- | --- | --- |
| **Week 1** | **Proposal Submission & Planning** | - Finalize the dataset selection, ensuring it meets volume, complexity, and relevance criteria.  - Conduct initial data exploration (sample subset) to validate the feasibility of transformations and storage strategy.  - Outline ETL workflow structure (diagram or notes) to inform Week 2 development.  - Submit the project proposal document, clearly outlining the data source, technology stack, tools, and implementation strategy.  - Set up the development tools requred for the whole process of extraction and transformation. |
| **Week 2** | **Environment Setup & Initial Pipeline** | - Begin scripting for data extraction from API  - Implement and test initial transformation logic  - Set up containers (Kestra) for reproducibility using Docker.  - Set up an orchestration tool Kestra to schedule data extraction jobs from Kaggle |
| **Week 3** | **Complete ETL Integration** | - Expand transformation logic with enrichment and advanced cleaning/aggregation tasks.  - Finalize data loading into the target storage system (PostgreSQL or BigQuery).  - Review of the dataset using basic queries.  - Begin testing for pipeline performance and data accuracy. |
| **Week 4** | **Visualization, Monitoring, & Final Documentation** | - Conduct EDA using python.  - Create interactive dashboards GCP’s Looker Studio  - Conduct final data quality checks and optimize transformation scripts.  - Document all implementation steps, architectural decisions, and visual outputs.  - Finalize project for submission/presentation, including README, code, dashboard links, and diagrams. |

**Gantt Chart: ETL Project Timeline**



## **6. Expected Challenges**

**Identify potential challenges you anticipate in the project and propose strategies to address them.**

Anticipating potential obstacles during ETL pipeline development is key to delivering a successful and maintainable project. Below are some common challenges that may arise throughout the project lifecycle, along with proposed strategies to proactively mitigate them:

| **Challenge** | **Description** | **Proposed Strategy** |
| --- | --- | --- |
| **Data Quality Issues** | The covid\_data dataset contains missing values, inconsistent formats, and outliers. | Implement comprehensive data validation and cleaning routines using pandas and IQR for the outliers. |
| **API Instability or Rate Limits** | Extraction of data from kaggle and use of kaggle hub API. The is an error in this process on the library importation and pip installation process | Research more on troubleshooting of the extraction phase and ensure that the stability is attained. |
| **Tooling Configuration & Integration** | Kestra set up and the whole ETL automations is quite a hurdle. Kestra keeps crashing in the process. | Allocate buffer time in Week 2 for environment setup. Use official documentation, class lectures & walkthrough videos and community forums for troubleshooting, and containerize environments to ensure consistency. |
| **Performance & Scalability** | Large datasets may slow down processing or exceed local system limits. | Use chunking, streaming, or batch processing strategies. For large-scale data if needed incorporate use of either databricks or BigQuery to optimize queries and transformations throughout. |
| **Dashboard Usability & Clarity** | Creating visualizations that clearly convey insights without overwhelming users can be difficult. | Define clear KPIs and design dashboards with usability in mind. Use consistent formatting, filters, and interactive elements to make the visuals accessible to technical and non-technical audiences. |

## **7. References**

<https://www.kaggle.com/datasets/meirnizri/covid19-dataset>