ETL Project

By: Nishant, John and Anam

**Pre-Processing**

Data analyst is very demanding role. As part of project requirements our team has been hired to analyze data scientist data to address whether data scientists are very demanding or not with location.

The following table illustrates the observations and actions taken by the group to ensure a clean data set.

| **Pre-process Step** | **Data Need** | **Observation** | **Action** |
| --- | --- | --- | --- |
|  | 1000+ rows of data needed | Data well observed and relevant to each source | None taken. |
|  | Unique identifier | Job Location is the unique identifier for two of the data sets. | None taken. |
|  | Verified formats of data and availability of all rows | Job Location, Job title, Salary, and Job class data | Perform data cleaning and keep only necessary columns. |
|  | Findings | *.* |  |
|  | Findings |  |  |
|  | Findings |  |  |

**Extraction**

We used 2 different datasets from the public platform Kaggle which led us to the Data science job opportunities and Data science job listings. The data in the two files included the following information:

* Data science job opportunities
* Data science job listings

The fields of interest include the following:

* Job location
* Job title
* Salary
* Job class

The following sources for our datasets used:

<https://www.kaggle.com/datasets/nadzmiagthomas/australia-data-science-jobs>

<https://www.kaggle.com/datasets/nomilk/data-science-job-listings-australia-20192020>

**Transformation**

In order to transform the public data and use it in our study we performed the following:

* Used Pandas functions in Jupyter Notebook to load all two CSV files.
* Reviewed the files and transformed into data frames.
* We used Pg admin to create table.
* Removed the unnecessary columns as part of data cleaning and keep only three columns (Job location, Salary, and Job title) in data science job opportunities datasets.
* Removed the unnecessary columns as part of data cleaning and keep only two columns (Job location and Job type) in data science job listings data set.
* Identified duplicates by doing an inner merge on the Job Location column across all two data sets.

**Data science:**

job\_location job\_title salary

0 Melbourne Analyst 95917

1 Mulgrave Clinical Research Associate 96555

8 Australia Software Engineer 212000

40 Dandenong Quality Manager 90000

44 Reservoir Food Technologist 75000

**Job listings**

job\_location job\_class

0 Sydney Science & Technology

1 ACT Information & Communication Technology

8 Melbourne Information & Communication Technology

15 Perth Information & Communication Technology

24 Brisbane Mining, Resources & Energy

**Joining data science and job listings:**

job\_location job\_title salary job\_class

0 Sydney Data Scientist 125000 Science & Technology

1 Melbourne Analyst 95917 Information & Communication Technology

2 Perth Entry Level

Media Coordinator 65520 Information & Communication Technology

3 Brisbane Graduate

Data Scientist 128589 Mining, Resources & Energy

4 Darwin Data Manager 82988 Science & Technology

5 Adelaide Software Engineer 85000 Science & Technology

6 Hobart Biostatistician 106500 Science & Technology

7 Gold Coast Data Engineer 82171 Information & Communication Technology

**Load**

After we pulled in the CSV files and loaded them into the data frames, we did an initial connection to the Postgres database using PG admin to store our original clean data sets. We used the quick database website to create the initial table schema that got loaded into the Postgres database that generated the first set of tables. After running the queries and created the new tables with only the relevant information we reconnected to the database and generated additional tables for the data frames.

**Postgres Database:**

**Joining of data science and job listings:**

job\_location job\_title salary job\_class

0 Sydney Data Scientist 125000 Science & Technology

1 Melbourne Analyst 95917 Information & Communication Technology

2 Perth Entry Level

Media Coordinator 65520 Information & Communication Technology

3 Brisbane Graduate

Data Scientist 128589 Mining, Resources & Energy

4 Darwin Data Manager 82988 Science & Technology

5 Adelaide Software Engineer 85000 Science & Technology

6 Hobart Biostatistician 106500 Science & Technology

7 Gold Coast Data Engineer 82171 Information & Communication Technology

**Summary**

There were some limitations to our findings due to the data available and limited time. However, we were able to address some relevant questions in our initial project proposal below:

The data science job is very demanding role across Australia as it offers higher salary and a huge number of opportunities,

Questions:

Findings:

Questions:

Findings: