ETL Project Report:

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
1.	1000+ rows of data needed	Data well observed and relevant to each source	Two csv files
2.	Unique identifier	Job Location is the unique identifier for two of the data sets.	Job Location – Primary key
3.	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.
4.	Config.py	Add postgres username and password	Add with. gitignore
5.	Project Proposal	Add all requirements, Data sources and brief descriptions about project	Project proposal is added to GitHub

Extraction

The main aim of extraction is to fetch data from sources. Then clean that data at professional level so anybody can easily understand.

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

Here we used pandas read to extract data from our csv sources. We used dropna method to remove unnecessary data. We keep only necessary columns in these datasets which we already mentioned at the start of extraction.

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

To transform the public data and use it in our study we performed the following:

- The main aim of this transformation is how we clean data set as professional level so we can join that data set with relevant columns.
- Used Pandas functions such as inspect to find table names in Jupyter Notebook and config to add Pgadmin's password and username.
- 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

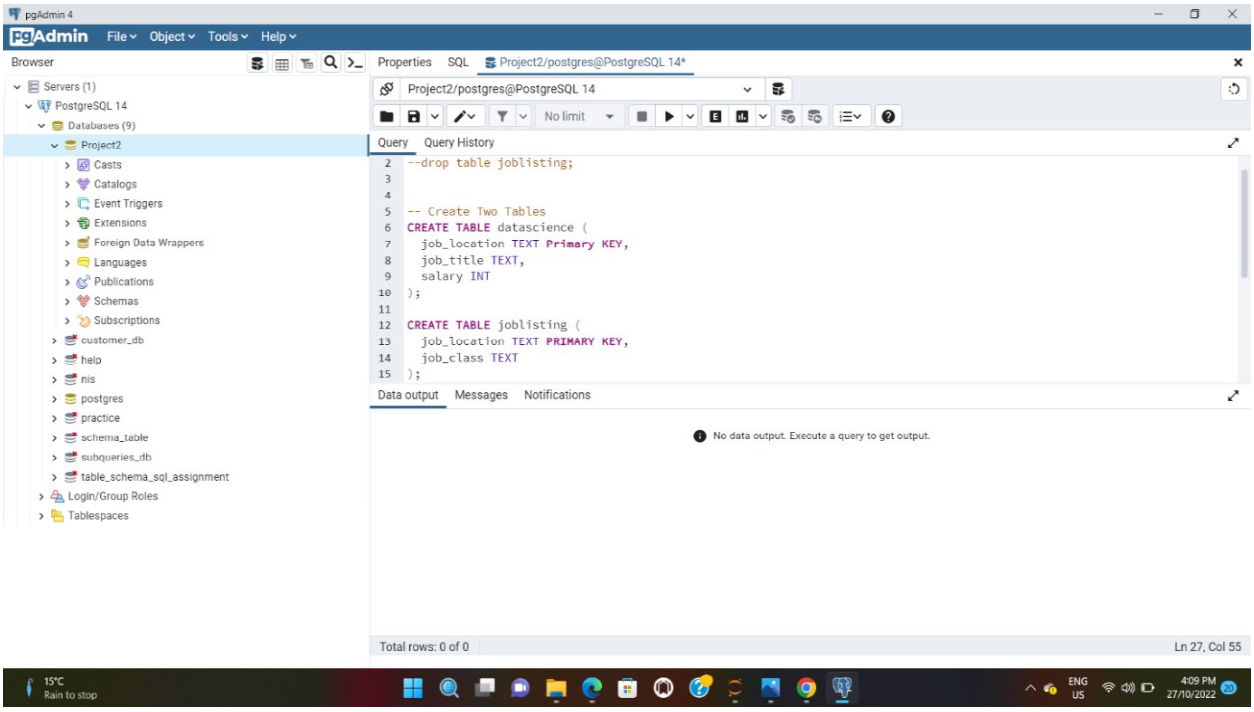
Load

The basic aim of Loading is how we join two datasets by using Job Location as primary key. We successfully linked two data sources which are relevant to each other. Basic steps which we used are first 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.

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

Postgres Database:



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: Is it worth to be a Data scientist in Australia?

Findings: Yes, Data science is very popular job profile in Australia as it offers higher salary. Here in this analysis, we found data scientist making more than 90000$ per annum. Moreover, we can also be able to see that job opportunities are very huge as per the data available for each job location.

Questions: Which Job class are popular in Australia as Data scientist?

Findings: Here in Job listings, we can easily be able to conclude that Information & technology is very popular job class in not only Australia but also in major states of Australia. Moreover, data analyst from this similar job class makes higher annum salary.

Questions: Which job titles getting highest annum salary?

Findings: Based on the analysis, Software engineer makes highest salary with more than 200000$ and they belong to the science and technology job class.