ETL Project Report:

By: Nishant, John and Anam

Pre-Processing

Data science 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	insert Username = <'username'> and Password = <'password'>	Add with .gitignore
5.	Project Proposal	Add all requirements, Data sources and brief descriptions about project	Project proposal is added to GitHub
6.	Dependencies	import pandas as pd from sqlalchemy import create_engine from sqlalchemy import inspect import config	

Before extraction using python and pandas, create a new database called 'datascience_db' in pgAdmin. In the newly created database, create two tables and inner join them using query tools. This joined table (currently empty) will later hold the data that we're interested in at the end of the ETL process. You'll use python and pandas for ETL process in Jupiter notebook. And at the end of the process, you'll load the DataFrames into the postgreSQL table that we created in the beginning. Note that the names of the columns in postgreSQL ad Pandas should be same to save yourself falling in troubles during the loading process.

Extraction

The main aim of extraction is to fetch data from sources.

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 function “read” to extract data from our csv sources by setting the path to the csv file kept in the “Resources” folder and then we displayed it into the data frames named as data_science_df and job_listing_df.

The following sources for our datasets were used:

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

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

Note: In second data source links follow only listings 2020_2022.csv file.

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.
- For transformation, clean the dataframes by keeping the copy of the columns that you're interested in. Using a copy and not the original data will save us from troubles.
- Remove the unnecessary columns and rename so columns match with table in postgres. As part of data cleaning and keep only three columns (job_ location, salary, and job_title) in data science job opportunities datasets.
- Remove the unnecessary columns and rename so columns match with table in postgres. As part of data cleaning and keep only two columns (job_ location and job_type) in data science job listings data set.
- Remove duplicates from job_location in both data frames so that we are able to join both dataframes in loading phase.
- After cleaning, review the files and transformed into transformed data frames.

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

- Connect to the local database. Here create a config.py file and keep your username and password in it and save the config.py file in .gitignore file to keep your username and password confidential. If it's not confidential, you can put it straight away in the code and you won't have to create config.py or .gitignore file then.
- Inspect method to find the table names (datascience and joblisting).
- Load csv converted DataFrames into database (datascience_db)
- Confirm data has been added by querying the tables in both pandas and postgreSQL
- Join the two tables in pgAdmin or join the two tables in with Pandas and SQLAlchemy.

Joining of datascience and joblistings:

	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:

```
16
17  -- Query to check successful load
18  SELECT * FROM datascience;
19
20  SELECT * FROM joblisting;
21
22
23  -- Join tables on job_title
24  SELECT datascience.job_location, datascience.job_title, datascience.salary,joblisting.job_class
25  FROM datascience
26  INNER JOIN joblisting
27  ON datascience.job_location = joblisting.job_location;
```

Data output | Messages | Notifications

	job_location text	job_title text	salary integer	job_class text
1	Sydney	Data Scie...	125000	Science & Technology
2	Melbourne	Analyst	95917	Information & Communication Technolo...
3	Perth	Entry Lev...	65520	Information & Communication Technolo...
4	Brisbane	Graduate ...	128589	Mining, Resources & Energy
5	Darwin	Data Man...	82988	Science & Technology
6	Adelaide	Software ...	85000	Science & Technology
7	Hobart	Biostatist...	106500	Science & Technology

Total rows: 8 of 8 | Query complete 00:00:00.076

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.