Stephen Anderson

Jacquelyn Cleary

Megan Dodson

**Wk. 13 ETL Project Proposal**

We are going to be looking at the US Data Scientist job market and the cost of living in the US. We will be extracting both datasets from Kaggle using Pandas into Jupyter Notebook. We will also use Pandas to transform and clean those datasets. Then we will use Postgres to transport the data and showcase the overlap between them.

Data Scientist job market:

* <https://www.kaggle.com/milan400/glassdoordata-scientist>
* CSV format file

Cost of Living dataset:

* <https://www.kaggle.com/debdutta/cost-of-living-index-by-country>
* CSV format file

US Cities – Primary Key

Based on our initial reviews of the data set, here are some of the cleaning steps we will be pursuing:

* Limiting scope to just US cities
* Ensuring the primary keys (US Cities) look similar to each other (standardizing the format)
* Determining how we will handle cities with no job postings and how we will handle job postings that we don’t cost of living data for
* How to handle if there are duplicated job titles (Data Scientist at company A vs company B, etc.)

Initial Questions we want to answer:

* “Best” Place to live and be a Data Scientist based on salary and cost of living
* Amount of job postings by city
* City size/demographics
* Salary ranges for major cities