**ETL PROJECT REPORT**

Team Members: Zach Wilson, Alfred Amachree, Matt Pickett

Date: 4/15/2019

After looking at various data sources we concluded to work on a London dataset that investigates how many people visited London from 2017 – 2019.

We sourced our data from data.london.gov, which is an official website for data from the mayor’s office in London.

EXTRACT

Here’s the link to our data sources:

<https://data.london.gov.uk/dataset/number-international-visitors-london>

<https://ckan.publishing.service.gov.uk/dataset/councillor-travel-and-accommodation-expenses>

TRANSFORM

We proceeded to use Python libraries in jupyter notebook to Transform our data.

Some considerations we investigated and transformed in the data set include:

For the 2018-19.csv and 2017-18-1.csv files

* In our “Trip Type” column we had values (“Train”,” Hotel”, “Flight”) we dropped rows that contained the “Hotel” value since we were not interested in the Hotel information.
* We dropped columns like (“Itinerary ID”, “Total Amount spent”, “VAT”,” Detail”) that were not needed in our Postgres table.
* Renamed columns to remove blank spaces in column names.
* Exported cleaned csv files to Resources folder.
* Merged the two csv in MS Excel.

For the international-visitors-london-raw.csv file

* Dropped rows that were not in the 2017 -2019 date range.
* Dropped columns like (“Market”, “Area”, “Sample”) that did not correspond to the other csv extracted.
* Renamed columns to remove blank spaces in column names to match the other csv.
* Exported cleaned csv files to Resources folder.

LOAD

In Postgres:

* we created the database “Londontrip\_dB”
* we created the tables “International\_travel” and “london\_travel”
* we loaded our exported CSVs into the database