**Project 2 Summary Report**

**Data Extraction, Transformation, Load (ETL)**

**Team Members:**

* Russ McGrath
* Sharmila Sainani
* Tony Joy

**Objective:**

To help a Data Science job seeker find a job to suits their preferences.

**Method:**

Collect job postings and combine them with weather and cost of living information for the city each job is located in.

**Datasets:**

**Github Jobs API –** jobs.github.com

**City Weather Summary (web scrape) -** <https://weatherspark.com>

**Cost of Living Data (scrape)** - <https://www.numbeo.com/cost-of-living/region_rankings_current.jsp?region=019>

**Data Collection and Cleaning:**

See jupyter notebook files (in order):

*Github\_jobs.ipynb* – job posting API query

*Weather\_blurb.ipynb* – city weather summary website scrape

*COL\_jobs.ipynb* – cost of living data scrape

**Data combining and loading:**

See jupyter notebook file:

*pandas\_combiner.ipynb* – combining data and uploading to mongoDB

(continued on page 2….)

**“Schema” Summary,:**

**Database: mongoDB**

**Number of collections: 1**

**Date structure:**

[{'creation\_date': <JOB POSTING DATE>,

'title': <JOB TITLE>,

'company': <COMPANY NAME>,

'location': <JOB LOCATION>,

'description': <JOB DESCRIPTION,

'url': <JOB POSTING URL>,

'weather': <CITY CLIMATE SUMMARY>,

'COL Index': <CITY COST OF LIVING INDEX>,

'Rent Index': <CITY RENT COST INDEX>,

'Groceries Index': <CITY GROCERY COST INDEX>,

'Restaurant Price Index': <CITY RESTAURANT COST INDEX>,

'Local Purchasing Power Index': <CITY PURCHASING POWER INDEX>},

...]

**Cost of living indices explanation:**

From numbeo.com:

*These indices are****relative to New York City (NYC)****. Which means that for New York City, each index should be 100(%). If another city has, for example, rent index of 120, it means that on an average in that city rents are 20% more expensive than in New York City. If a city has rent index of 70, that means on average rent in that city is 30% less expensive than in New York City.*

***Cost of Living Index (Excl. Rent)****is a relative indicator of consumer goods prices, including groceries, restaurants, transportation and utilities. Cost of Living Index does not include accommodation expenses such as rent or mortgage. If a city has a Cost of Living Index of 120, it means Numbeo has estimated it is 20% more expensive than New York (excluding rent).*

***Rent Index****is an estimation of prices of renting apartments in the city compared to New York City. If Rent index is 80, Numbeo has estimated that price of rents in that city is on average 20% less than the price in New York.*

***Groceries Index****is an estimation of grocery prices in the city compared to New York City. To calculate this section, Numbeo uses weights of items in the "Markets" section for each city.*

***Restaurants Index****is a comparison of prices of meals and drinks in restaurants and bars compared to NYC.*

***Cost of Living Plus Rent Index****is an estimation of consumer goods prices including rent comparing to New York City.*

***Local Purchasing Power****shows relative purchasing power in buying goods and services in a given city for the average net salary in that city. If domestic purchasing power is 40, this means that the inhabitants of that city with an average salary can afford to buy on an average 60% less goods and services than New York City residents with an average salary.*