

DSC 540 Data Preparation (DSC540-T301 225-1)

Bellevue University

Project:

Milestone: 1

Author: Jake Meyer

Date: 04/10/2022

Project Subject Area:

The scope for this project is to explore the top cities to live in the United States along with useful insights for Data Scientist job outlooks within these recommended locations. This project will focus on the following inquiries:

- Identify the top cities to live in the United States.
- Explore basic demographic information about the top cities (population, age, income, sex, ethnicity, count of families, etc....)
- Investigate Data Scientist job opportunities for top cities (potential income, job counts, job opportunities, etc....)
- Explore which cities or states would be highly desired to settle as a Data Scientist.

The project will entail finding data sources, cleaning/formatting the data, loading the data into a database, and a final analysis to review the items listed above.

Data Sources:

1) Flat File: metro_areas.csv and state_ranks.csv

- a. Description of metro_areas.csv: This file will consist of 17 columns and 125 rows. The file contents are centered around the top 125 cities from the US News World Report generated in 2019. Unfortunately, there was no report available for 2022 after hours of searching. An additional file can be generated with an updated report for the top cities in 2021 and will be addressed later in the project. The columns for the existing file are:
 - i. Metro Area (City, State), 2019 Population, Average Salary, Rent Index from Zillow, Average Rainfall, Lowest Monthly Average High Temperature, Highest Monthly Average Temperature, Violent Crime Rate, Property Crime Rate, US News Overall Score (0-10), Nomad Scores Seasonally (December, June, August), and Data Scientist Job Postings.
 - ii. Link to US News World Report article for top cities to live: [Best Places to Live in the U.S. in 2021-2022 | Places Rankings | U.S. News Best Places \(usnews.com\)](#)
 - iii. CSV file is also attached below:



metro_areas.csv

- b. Description of state_ranks.csv: This file consists of 11 columns and 50 rows. US News ranked each state based on Health Care, Education, Infrastructure, Opportunity, Fiscal

Stability, Crime and Corrections, and Natural Environment. The columns for the existing file are:

- i. State Abbreviation, State Name, Overall Rank, Healthcare Rank, Education Rank, Infrastructure Rank, Fiscal Stability Rank, Crime and Correction Rank, and Natural Environment Rank.
- ii. Link to US News World Report article for state rankings: [Overall Best States Rankings | US News Best States](#)
- iii. CSV file is also attached below:



state_ranks.csv

- c. Link for both .csv files: [Best Cities for Data Scientists | Kaggle](#)
- 2) API: Esri Data and Maps
- a. Description: This API provides demographic data for cities with populations of 10,000 people or greater. The data contains 3886 rows and 51 columns. Some of the key columns to consider from this dataset would be the City Name, State, Population, Ethnicity (White, Black, Asian, Hispanic, etc....), Sex (Male or Female Counts), Age (separated by Age groups <5, 10-14, 15-19, etc...), and Number of Families. This dataset also includes longitude and latitude data if mapping is desired for a particular city.
 - b. Link: [USA Major Cities | USA Major Cities | ArcGIS Hub](#)
- 3) Website: Zippia – Average Data Scientist Salary
- a. Description: This website provides several different visualization charts and tables about Data Scientist's salaries with respect to state, city, and industry.
 - b. Link: [Data Scientist Salary \(April 2022\) - Zippia | Average Data Scientist Salaries Hourly And Annual](#)

Relationships:

All these data sources are related with location attributes, specifically by city or state. The metro_areas.csv file contains data for the top cities to live. This file can also be linked to the other sources by state; however, it will require a split for the 'Metro Area' column. The state_ranks.csv will relate to the other sources by state ('State Abbreviation' or 'State Name'). The demographic data captured in the API for USA Major Cities will relate to the other sources by either city ('NAME') or by abbreviated state ('ST'). The Data Scientist job data, from Zippia's website, contains city or state attributes to connect with the other sources. The city and/or state attributes from each dataset will need to be cleaned for consistency.

Plan for Project:

US News World Report publishes the best cities for living based on five general categories. The five categories are:

- 1) Job Market Index (unemployment rate and average salary)
- 2) Housing Affordability Index (median annual household income and annual housing cost)
- 3) Quality of Life Index (crime rates, healthcare quality/availability, education quality, well-being, and commuter time)
- 4) Desirability Index (survey of 3600 individuals to determine most desirable locations to live)
- 5) Net Migration (people moving to or from the location).

A score is calculated for each city, through Z-scores and T-scores, which will return a value between 0-10. The cities were then ranked based on these scores and placed into the metro_areas.csv file. The top cities were then sorted based on the US News score to provide a recommendation for top places to live. This was a little background from the article that sparked the interest for this project. This project aims to explore supplementary data around these top cities. In addition, it will be interesting to understand how a Data Scientist may fair if considering one of the recommended locations to live. The methods followed during this project will be documented and executed with data ethics in mind. The intent is to provide additional insights, from a Data Scientist's perspective, on best cities for living based on a study conducted by US News World Report. The steps throughout this process will be documented for transparency. My plan consists of five main phases. Each phase contains a tentative approach for how this project will be completed.

The first phase involves finding data sources of various types. The sources that have been identified were outlined in the sections above. There are two .csv files, an API source, and a website source. The .csv files provide the top city recommendations for living. The API source provides demographic data for each of the cities. The website data provides valuable insight for a Data Scientist's projection for job opportunities and salary in specified city or states. Any additional data sources used in this project will need to be documented later.

The second phase involves consolidating, cleaning, and or transforming the .csv files into a single dataset. The metro_areas.csv file will be the starting point for the data cleaning. The columns will be renamed, 'Metro Area' will be broken out into 'City' and 'State' columns, and any non-essential columns will be removed. The state_ranks.csv file will be merged with metro_areas.csv to consolidate the data into one source. An additional column of 2022 rankings will be added to the dataset. Lastly, any missing values will be identified.

The third phase involves cleaning and formatting the website data. The website data will be filtered for the cities identified from the metro_areas.csv. Non-essential columns will be dropped from the dataset. The column names will be updated accordingly and the string text for the city names may need to be cleaned. The website has a few different tables, so these will need to be merged into one dataset. Lastly, the data will be reviewed for missing values and outliers. This section will be new for me as I have not scraped data from a website previously. The biggest challenge here will be scraping the data from the website.

The fourth phase involves connecting to the API source and cleaning/formatting the data. The first step for this phase will be connecting to the API. I foresee this being one of the biggest challenges for the entire project. Like the previous step(s), the columns will need to be renamed accordingly. Any columns that do not add value will be dropped. The data will be reviewed for missing values and outliers. I foresee this as being one of the more challenging phases since I've not worked much with API data sources. I've connected to an open weather API for one of the DSC 510 course assignments but will need to refresh on this topic.

The fifth, and final, phase for this project consists of merging the cleaned/formatted .csv, website, and API data together. The datasets will be loaded into a database and data visualization charts will be produced to help illustrate the findings. Some example visualization charts will most likely be histograms, scatterplots, and box plots. This is another phase that will be challenging since I have not loaded datasets into a database previously.

References:

Corliss, J. (2020). *Best Cities for Data Scientists*. [Dataset]. [Best Cities for Data Scientists | Kaggle](#)

Esri. (2022). *USA Major Cities*. [Dataset]. [USA Major Cities | ArcGIS Hub](#)

U.S. News. (2022). *Best State Rankings*. [Best Places to Live in the U.S. in 2021-2022 | Places Rankings | U.S. News Best Places \(usnews.com\)](#)

U.S. News. (2022). *150 Best Places to Live in the U.S. in 2021-2022*. [Best Places to Live in the U.S. in 2021-2022 | Places Rankings | U.S. News Best Places \(usnews.com\)](#)
















Zippia. (2022). *Average Data Scientist Salary*. [Data Scientist Salary \(April 2022\) - Zippia | Average Data Scientist Salaries Hourly And Annual](#)

Appendix: Milestone 1 Assignment Criteria

Identify Datasets

The first milestone of this project will be to select the data you want to work with. You will need to select 3 different data sources that have different file types of information – and the data will need to have a relationship between them. If one doesn't exist, you will have to create one. It is likely you will need to manipulate the data to create a relationship. Finding the data, you want to work with for this project, will likely be the hardest part of the project. You must have one of each of the following types of datasets – and you need a minimum of 1000 rows across all datasets. Each dataset should have a minimum of 10 columns/variables.

- CSV/Excel/PDF or another flat file source.
- Website you want to pull data from--you will want to identify a website that has data stored in a table, similar to the screenshot below.

Rank ↕	Country (or dependent territory) ↕	Population ↕	% of world population ↕	Date ↕	Source
1	 China ^[b]	1,403,496,680	18.0%	12 Jul 2020	National population clock ^[3]
2	 India ^[c]	1,364,603,167	17.5%	12 Jul 2020	National population clock ^[4]
3	 United States ^[d]	329,940,508	4.23%	12 Jul 2020	National population clock ^[5]
4	 Indonesia	269,603,400	3.46%	1 Jul 2020	National annual projection ^[6]
5	 Pakistan ^[e]	220,892,331	2.83%	1 Jul 2020	UN Projection ^[2]
6	 Brazil	211,782,426	2.72%	12 Jul 2020	National population clock ^[7]
7	 Nigeria	206,139,587	2.64%	1 Jul 2020	UN Projection ^[2]
8	 Bangladesh	168,940,146	2.17%	12 Jul 2020	National population clock ^[8]
9	 Russia ^[f]	146,748,590	1.88%	1 Jan 2020	National estimate ^[9]
10	 Mexico	127,792,286	1.64%	1 Jul 2020	National annual projection ^[10]
11	 Japan	125,930,000	1.61%	1 Jun 2020	Monthly provisional estimate ^[11]
12	 Philippines	108,881,966	1.40%	12 Jul 2020	National population clock ^[12]
13	 Egypt	100,608,449	1.29%	12 Jul 2020	National population clock ^[13]
14	 Ethiopia	98,665,000	1.27%	1 Jul 2019	National annual projection ^[14]
15	 Vietnam	96,208,984	1.23%	1 Apr 2019	2019 census result ^[15]

- API you will pull data from.

Some places you can find datasets are listed below:

- [Tableau Community](#)
- [Kaggle Datasets](#)
- [Data.Gov](#)
- [Science.Gov](#)
- [Data.Gov.UK](#)
- [NORC](#)
- [European Social Survey](#)
- [API List](#)
- [PrommableWeb](#)
- [Public APIs](#)
- [OpenWeatherMap](#)

Wikipedia is a good source to find data that is in a table - and the structure of the HTML is usually very similar.

There are no restrictions on what dataset you use, other than you cannot use the specific datasets used in the book(s).

For the first milestone, you need to submit the following:

- Project Subject Area: Describe your project in 1-2 sentences
- Data Sources:
 - Flat File:
 - Description
 - Link or Flat File uploaded
 - API:
 - Description
 - Link
 - Website:
 - Description
 - Link
- Relationships
 - Describe how the data from each source is connected (see example below).
 - If there isn't an obvious relationship, explain how you will make one
- 250 Words describing how you plan to tackle the project, what the data means, ethical implications of your project scenario/topic, and what challenges you might face.

Submit via a PDF to the assignment link.

Example of Relationships:

In case you are confused what is meant by a relationship between the data sources here is an example (this is a very simple example and I would expect your datasets to have more variables)

CSV File: Contains a list of stores by store ID and other metadata about the stores

Website: Contains a list of store locations, by location ID and store ID and the various departments each store has by department ID.

API: Contains the transactions at each store – contains a transaction ID and store ID.

All 3 of these data sources are related by Store ID. The CSV file has a 1 to many relationship with the Website by StoreID and has a one to many relationship with the API data by StoreID as well.