***Milestone 1 – Introduction***

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

* 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)
* The relationships between them, or the relationship you will make between them
* What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words
* 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, the ethical implications of your project scenario/topic, and what challenges you might face.

Submit via a PDF to the assignment link.

Example:

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.

Milestone 1 is due Sunday, by Midnight of Week 4. Refer to the rubric for more grading detail.

***Milestone 2 – Cleaning/Formatting Flat File Source***

Perform at least 5 data transformation and/or cleansing steps to your flat file data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone.

* Replace Headers
* Format data into a more readable format
* Identify outliers and bad data
* Find duplicates
* Fix casing or inconsistent values
* Conduct Fuzzy Matching

Make sure you clearly label each transformation (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

Milestone 2 is due Sunday, by Midnight of Week 6. Refer to the rubric for more grading detail.

***Milestone 3 – Cleaning/Formatting Website Data***

Perform at least 5 data transformation and/or cleansing steps to your website data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone.

* Replace Headers
* Format data into a more readable format
* Identify outliers and bad data
* Find duplicates
* Fix casing or inconsistent values
* Conduct Fuzzy Matching

Make sure you clearly label each transformation step (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

Milestone 3 is due Sunday, by Midnight of Week 8. Refer to the rubric for more grading detail.

***Milestone 4 – Connecting to an API/Pulling in the Data and Cleaning/Formatting***

Perform at least 5 data transformation and/or cleansing steps to your API data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if you needed to clean your data. The goal is a clean dataset at the end of the milestone.

* Replace Headers
* Format data into a more readable format
* Identify outliers and bad data
* Find duplicates
* Fix casing or inconsistent values
* Conduct Fuzzy Matching

Make sure you clearly label each transformation step (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

***Milestone 5 – Merging the Data and Storing in a Database/Visualizing Data***

Now that you have cleaned and transformed your 3 datasets, you need to load them into a database. You can choose what kind of database (SQLLite or MySQL, Postgre SQL are all free options). You will want to load each dataset into SQL Lite as an individual table and then you must join the datasets together in Python into 1 dataset.

Once all the data is merged together in your database, create 5 visualizations that demonstrate the data you have cleansed. You should have at least 2 visualizations that have data from more than one source (meaning, if you have 3 tables, you must have visualizations that span across 2 of the tables – you are also welcome to use your consolidated dataset that you created in the previous step, if you do that, you have met this requirement).

For the visualization portion of the project, you are welcome to use a python library like Matplotlib, Seaborn, or an R package ggPlot2, Plotly, or Tableau/PowerBI.

PowerBI is a free tool that could be used – Tableau only has a free web author. If your use Tableau/PowerBI you need to submit a PDF with your assignment vs the Tableau/PowerBI file. /p>

Clearly label each visualization. Submit your code for merging and storing in the database, with your code for the visualizations along with a 250-500-word summary of what you learned and had to do to complete the project. In your write-up, make sure to address the ethical implications of cleansing data and your project topic. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

**Remember** – your GitHub repository can act as a portfolio for potential employers! I would highly suggest using this to submit your work, so you can fill it with good content that demonstrates the projects you are working on!

Rodrigo Ivan Rodriguez DSC-540 Winter

**Milestone 1**

**Three Data Sources & Relationships**

*Instructions:*

3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)

The relationships between them, or the relationship you will make between them

*Answer:*

API

* **Name**: House canary API
* **URL**: <https://api-docs.housecanary.com/#getting-started>
* **Description**: House Canary API has detailed real-estate information at the house level. Data includes details about an individual house as well as market-evaluation and sale price.
* **Relationship**: MSA is the Foreign Key (FK) that ties this dataset to the other two dataset, with a many-to-one relationship. Ex: One ‘home\_address’ from the API can only be matched to many ‘MSAs’ from the Flat File or Web Table.

Flat File

* **Name**: Zillow Research – Sales – Median Sale Price (Smoothed & Seasonally Adjusted, All Homes, Monthly)
* **URL**: <https://www.zillow.com/research/data/>
* **Description**: This dataset is provided by Zillow Research and is comprised of the median price at which homes across various geographies were sold.
* **Relationship:**  RegionName is the FK that ties this dataset to the other two datasets, with a one-to-one relationship. Ex: One RegionName from this dataset can be matched to One MSA from the API or Web Table.

Web Table

* **Name**: Standard Metropolitan statistical areas by PCPI, adjusted by regional price parity
* **URL:** <https://en.wikipedia.org/wiki/List_of_U.S._cities_by_adjusted_per_capita_personal_income>
* **Description:** This dataset is obtained from Wikipedia and has the top 50 U.S. MSA’s by adjusted per capital personal income.
* **Relationship:** Metropolitan Areas is the FK that ties to the other two datasets, with a one-to-one relationship. EX Metropolitan Areas from this dataset can be matched to One MSA from the API and Flat File datasets.

**Accomplishing all 5 Milestones**

*Instructions:*

What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words

*Answer:*

The data will help accomplish all 5 milestones functionally and contextually. What I mean by contextually is that after the three datasets are cleaned, prepped, and joined, they can be used to conduct an in-depth analysis on the housing market.

The milestones involve finding the three datasets, cleaning and prepping each different type of data source, joining the datasets, and finally creating visualizations from the data. All datasets have multiple columns and relationships between them. Thus, the datasets can all be cleaned and joined together. Additionally, because the datasets all share relationships with each other, creating visualizations will be efficient and insightful.

The flat file has 68 columns with the majority of columns consisting of month columns. This dataset has the median sale price by MSA by month, with each month having its own column.

The API has a large number of variables. There are variables that have details at a lower granularity than the other two datasets. Data has information about individual house as well as market-evaluation and sale price. The data from this API can be rolled up at the MSA level to match with the other two datasets.

The Web data is a table similar to a flat file but hosted online on Wikipedia. The web table has data at the MSA level and holds information on adjusted per capital personal income.

The three datasets can be joined together by the MSA variable and each data source will be able to satisfy the requirements in each of the 5 milestones.

**Project Subject**

*Instructions:*

Project Subject Area: Describe your project in 1-2 sentences.

*Answer:*

The project focuses on cleaning, prepping, joining, and visualizing data. My project will aim to visualize housing market data at the MSA level by various attributes. The intent is to derive insight on what variables affect the housing market in different MSAs.

**Relationships**

*Instructions:*

* Describe how the data from each source is connected (see example below).
* If there isn't an obvious relationship, explain how you will make ones.

*Answer:*

**API Relationship**: The ‘MSA’ variable is the Foreign Key (FK) that ties this dataset to the other two dataset, with a many-to-one relationship. This variable is an ID that uniquely identifies the MSA.

**Flat File Relationship:** ‘msaID’ is the FK that ties this dataset to the other two datasets, with a one-to-one relationship with the Web Table & one-to-many relationship with the API.

**Web Table Relationship:** ‘metro\_id’ is the FK that ties to the other two datasets, with a one-to-one relationship with the flat file and one-to-many relationship with the API.

While both the Flat File and Web Table are at the MSA level, both data sources need the inclusion of an ‘msa’ variable that matches that of the API. This means that, a 5-digit US census MSA ID would need to be added to Zillow Flat File and Wikipedia Web Table

Here is an ERD diagram that shows these relationships:

A screenshot of a computer

Description automatically generated

**Description**

*Instructions:*

250 Words describing how you plan to tackle the project, what the data means, the ethical implications of your project scenario/topic, and what challenges you might face.

*Answer:*

My project revolves around finding three different data sources that have relationships between them. Then cleaning, prepping, joining, and creating visualizations from the final joined data.

I chose real-estate data from the HouseCanary API at the house level, median sale price data from a Zillow flat file at the MSA level, and personal income data from a Wikipedia Web Table at the MSA level. The focus is to use the variables available in these three data sources to identify any trends or insights in the housing market for some of the top MSAs in the country.

I have reviewed the three data sources and identified that both the Zillow and Wikipedia datasets need to the inclusion of a Census MSA ID field to match with the 'msa' field in the API dataset. After doing this I can begin cleaning and prepping each dataset to be joined.

To avoid ethical concerns I will ensure that any data used for this analysis does not contain any Personal Identifiable Information and is aggregated where possible to protect privacy. Additionally, some potential challenges I could face would be around data integration. I mentioned that I will need to add variables to the flat file and web table that will act as foreign keys. I will be using the census MSA IDs to match these IDs to the MSA primary keys in the respective data sources. I will need to ensure that this step is taken accurately in order to get accurate and consistent results.