**Milestone 1**

Satish Agrawal

Bellevue University

DSC 540

For this project I chose used cars datasets from variety of sources including a flat file like CSV, a webpage in tabular format and an API which can be accessed via restful GET endpoints. Below are the three data sources with description of data in there:

1. **A CSV file**: This CSV file is downloaded from Kaggle [1]. It contains Car’s data, scraped from AUCTION EXPORT.com. This dataset included Information about 28 brands of clean and used vehicles for sale in US. There are twelve features assembled for each car in this dataset. This data set includes basic car attributes like brand, model, year, color, the sale price for the car, vin (vehicle identification number) and mileage. It also includes other information like, state and city of the car location, title status, the lot number and condition.
2. **A web page**: The second source of cars data is from true car website where it has all the cars listed for sale [2]. This web site provides a variety of filters to pinpoint to small number of listed cars to limit the result set. This site provides multiple car attributes like interior color, exterior color, make model, year, mileage and many other attributes that define the car and its condition.
3. **An API**: The third source of data is a marketplace for API which provides information related to cars [3]. This API gateway site has many different APIs that provide data like inventory of cars, VIN history, Dealer info API, and other market statistics related data APIs. For this project I am concentrating on the statistics API which provides popular cars by city state pair. I have subscribed to the API for a trial to review the data and see any possible connections with other two data sources. This API provides the max and min price of the popular cars, with mean, variance and standard deviation. It also mileage statistics like min, max, median and variance. Other basic info like make and model are also provided.

**Relationship**

First, all these three data sources are about used cars. The one common relationship between the three is the make, modal and year and other common car attributes. It will be easy to identify the commons to do “JOIN” on for rest of the project milestones. Even though all these data sources are for used cars, they represent different type of information for analysis. The CSV file form Kaggle is a collection of cars listed for auction. The listing from the true car website is list of cars listed for sale (not auction) from the dealerships. And finally, the API from marketplace accumulates the history of used cars sell in passed in city state and provides the top 50 popular cars based on the sales. It will be interesting to join them into a RDMS (relational database management system) model.

**Possible Steps**

Looking ahead to other project milestones, it looks like we will need to work on these datasets independently first and clean and format the data to get ready for merging data from all the sources in one relations database like My SQL or SQLLite. While working with CSV, we have columns predefined and data is not changing once I have it downloaded. We will have to look for missing data and outliers and plan for handling them. Finding duplicates could be done by using the unique key which is an index in this case.

The second source is a website listing of cars, which is a dynamic data. Data keeps changing as new cars become available for sale and previously listed cars get sold. It is also possible that the website owner changes the look and feel of the site and starts providing information in different format. Individual records addition and removal may not be a big concern to deal with, if the format doesn’t change. It will be interesting to work with the website data using a web crawler. There will definitely be many munging steps with this data format as some of the data points has no header. Data provided in key value pair (for example JSON) is easier to work with as key explains what the data means. Here in this case we have features listed independently and we may have to make connections to create a list of features.

The third source (API) has data already in JSON format and can be handled and processed in Python. There are built in functions in Python to work with APIs using restful verbs. I have installed Postman (a restful client, open source software) to play with the many APIs provided by marketplace [3]. There will still be conversion of data fields to join with data from other sources.

References:

[1] <https://www.kaggle.com/doaaalsenani/usa-cers-dataset>

[2] <https://www.truecar.com/used-cars-for-sale/listings/>

[3] <https://apidocs.marketcheck.com/?version=latest#5be79d1e-2756-4cae-845f-8dc28fbf8a3c>