**Used car data**

Naveen Bagam

Bellevue university – Master of Science in Data Science

Course Name: DSC540-T301 Data Preparation (2227-1)

Instructor: Catie Williams

Due Date: 06/19/2022

**Used car data source**

3 data sources: -

1. Api data: -

Link:- <https://rapidapi.com/apininjas/api/cars-by-api-ninjas/>

This URL Returns the car data by model, it returns other metadata like city\_mpg, year,

Transmisson,fuel\_type etc., we need to pass the following details to the API in code.

This API data has about 13 columns in it.

querystring = {"model":"corolla"}

1. CSV file: -

[dsc540/USA\_cars\_datasets.csv at main · nbagam/dsc540 (github.com)](https://github.com/nbagam/dsc540/blob/main/USA_cars_datasets.csv)

This file contains used car data like price, make, model by state and it has the.

3) Website:-

https://en.wikipedia.org/wiki/List\_of\_U.S.\_state\_and\_territory\_abbreviations

This website data contains states and short codes and other attributes

Project Subject Area:

* Using this data, I am going to analyze all 3 data sets and prepare the data for model to find the avg sale price of all types of cars and attributes which are influencing the price of each used car by model.
* Most of the data is in CSV file which makes relation between the api data and website data.
* I have noticed that data is missing in a few fields of csv file
* Also, as part of all 5 milestones, first we need to clean the data, handle the missing data, remove the invalid value, remove the duplicate data, fix casing and inconsistent data, make the data in more readable format, also need to find if we have any outliners in the data and do the fuzzy matching.
* Once all the steps are completed, data will be ready for the model input.
* We must fix the case of string data and merge the data sets.

**Relationships:-**

The relation between API and CSV file is Make, model and year and the relation between the website and CSV is state.