Capstone 2 Milestone Report

Problem Statement

This capstone project aims to build a machine-learning tool capable of precisely and accurately classifying cars by their make, model, body style and year from images acquired from the Stanford car dataset using deep learning.

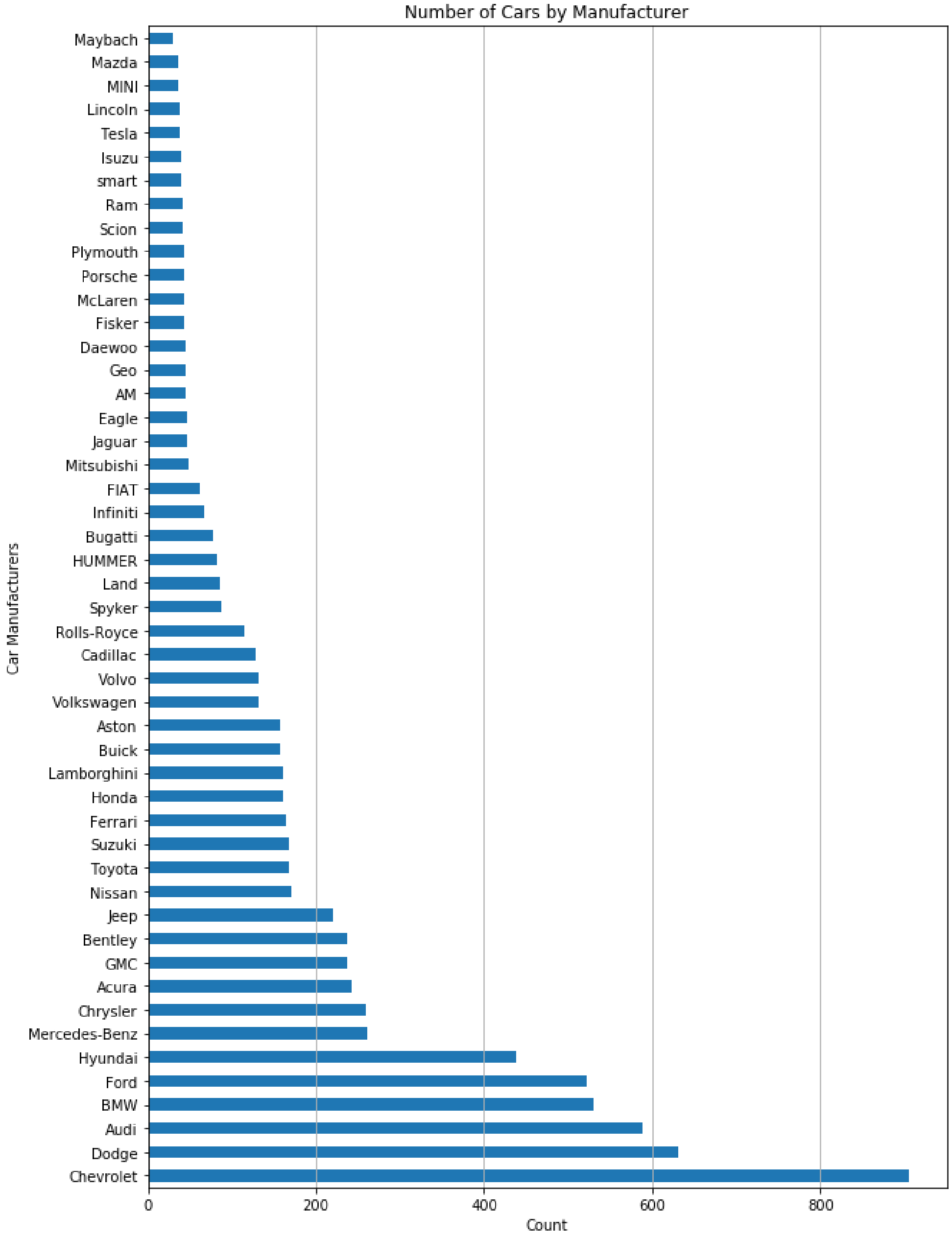
Data Wrangling

The data for the project was collected from a Kaggle dataset based on the Stanford car dataset, which included 196 classes for over 16,000 car images. Each class included the car manufacturer, the model, the body style, and the year the car was released. The data was roughly split 50-50 into training and test folders. All the images were stored in the JPG format within folders providing the class. Alongside the images, three CSV files were included. The first file called names contained the labels for each of the 196 classes within the dataset. The other two files provided insightful information about each image in the training and test data.

I loaded the train CSV file into a pandas dataframe examining the contents of the files given. The dataframe contained filenames of all the images within the train folder, the class number, which corresponded with the labels in the names CSV file mentioned above, and the positions of the boundary box for the vehicle in each image. While the column labels for the data files were listed on Kaggle the files did not have column headers. Once I updated the column names within the dataframe, I used the labels from the names file to add the make, model, and year for each row of the dataset. Along with that, I added two new columns taking the coordinates of the boundary box to give the height and width of the boundary boxes of each image. Specifically, to aid in my exploratory data analysis, in which I wanted to compare the cars within the dataset, I added columns for make, model, body style and year. The information for all these new columns was already established within the labels for each class, I used the builtin str method in pandas to split and isolate each piece of the label for their respective new columns. Once all the steps required to clean the data and make it serviceable for future action was revealed, I created a new python file called clean_data_files.py which would perform the necessary steps on the given CSV files and create a new clean CSV file. Once this script was run for both the train and test data files, I was ready to perform my exploratory data analysis.

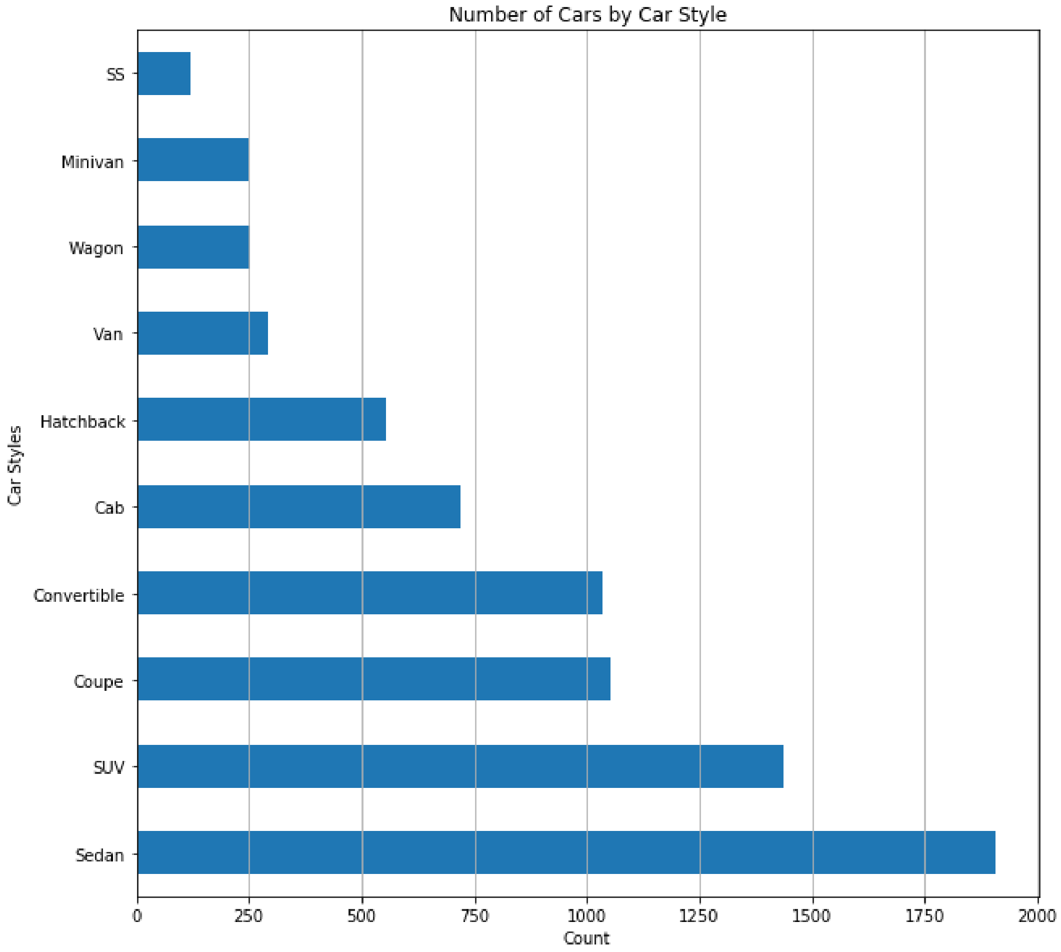
Exploratory Data Analysis

Having created new clean CSV files for both the training and test data, I loaded the clean train datafile into a pandas dataframe aiming to gain further insight into the data through visualization. Since I had isolated the make, model, body style, and release year of each vehicle in the dataset I wanted to aggregate the values to understand how diverse the data was. The graph below illustrates the different car manufacturers present in the dataset alongside the number of cars for each make featured in the training folder.



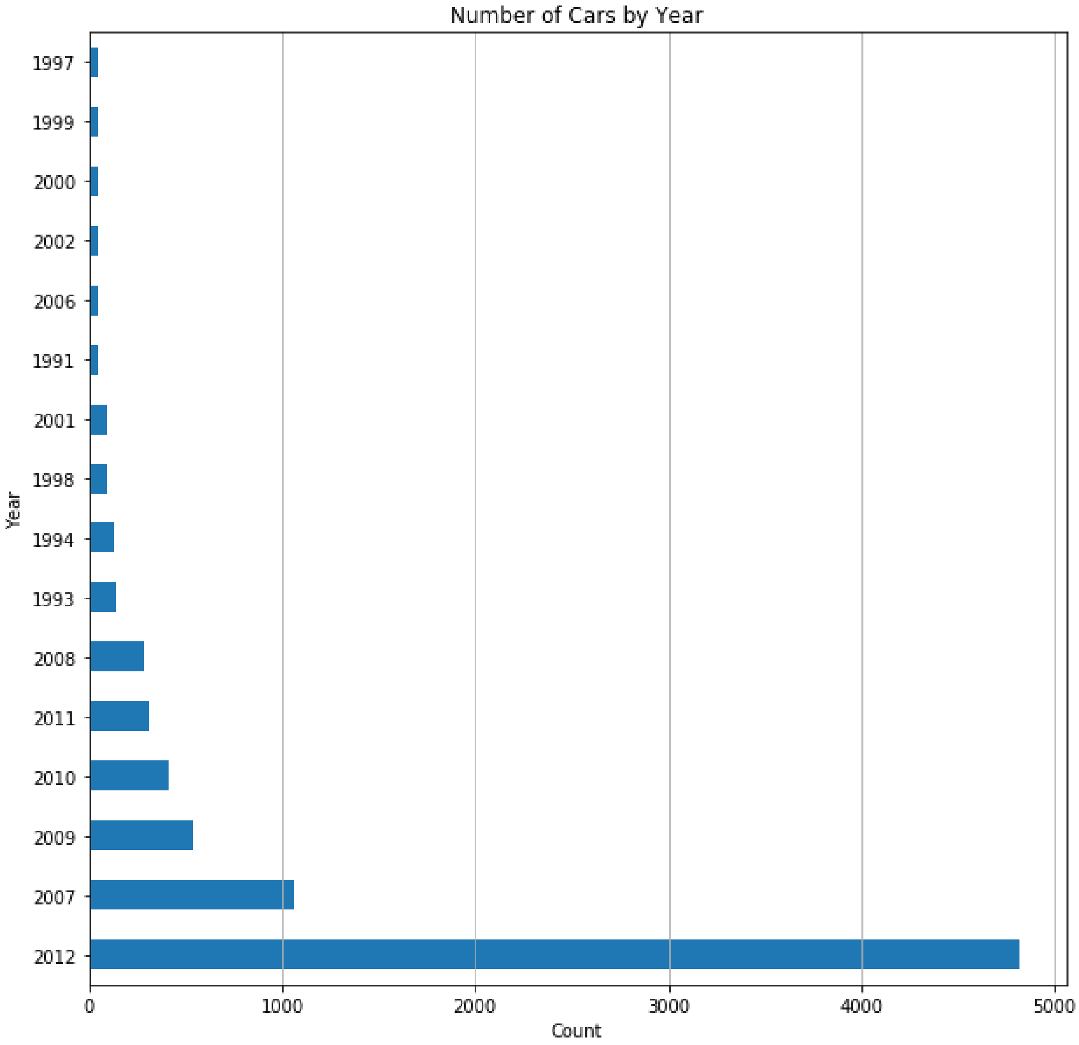
As can be observed from the graph above the dataset is diverse and all major car manufacturers are included. Chevrolet has by far the most vehicles in the dataset with nearly 300 more images than second-placed Dodge, with over 900 images roughly every ninth image in the dataset is a car made by Chevrolet. Twelve car manufacturers have over 200 images validating that a majority of vehicles that can be found on the streets and parking lots are available in the training dataset for our classification model. Many of the manufacturers that have limited photos make far fewer models and body styles than that of the top car manufacturers and so while the images may be fewer there is sufficient data available to classify such vehicles. For example, Tesla in 2012 which is the latest year of vehicles this dataset included made only one car the Model S available as a sedan, whereas Chevrolet manufactured 14 different car models alone not including the different body styles available for each model. The one car manufacturer that makes a variety of models but is largely underrepresented in this dataset is Mazda who has the second-fewest images in the Stanford dataset.

Let us now examine the different body styles of vehicles present in the dataset, due to many styles being unique to certain vehicles and having very few images only the ten most popular body styles were included in the graph below.



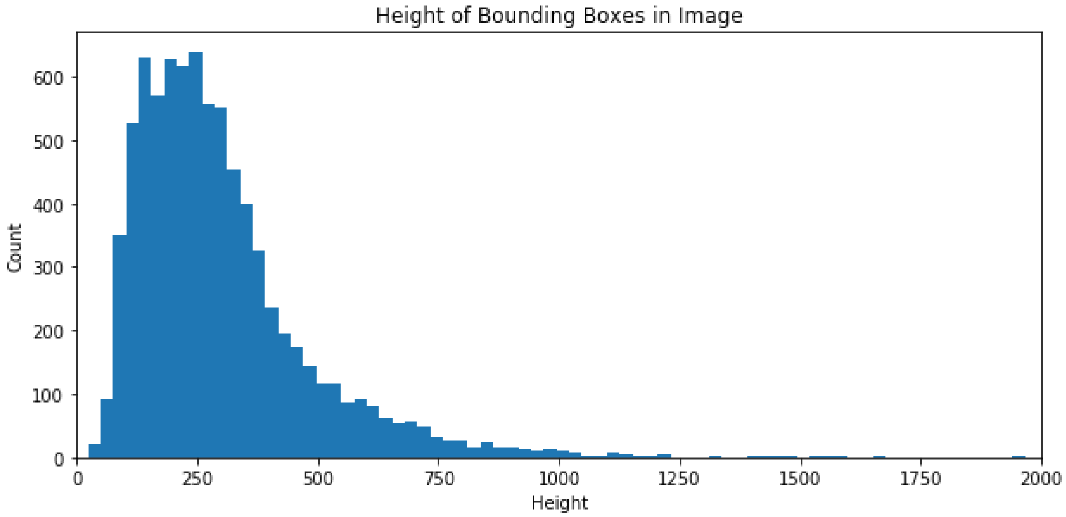
The graph above displays the body styles of the cars present in the Stanford car dataset. Mirroring the popularity of vehicles on the road, sedans and SUVs are the body classes that appear the most in the images provided. Sedans and SUVs account for over 3300 of the 8000 images present in the Stanford car dataset. Unlike the earlier graph, all the major and popular body styles included have a large number of images whereas earlier a leading car manufacturer such as Mazda has substantially fewer images compared to the competing automotive manufacturers of similar size.

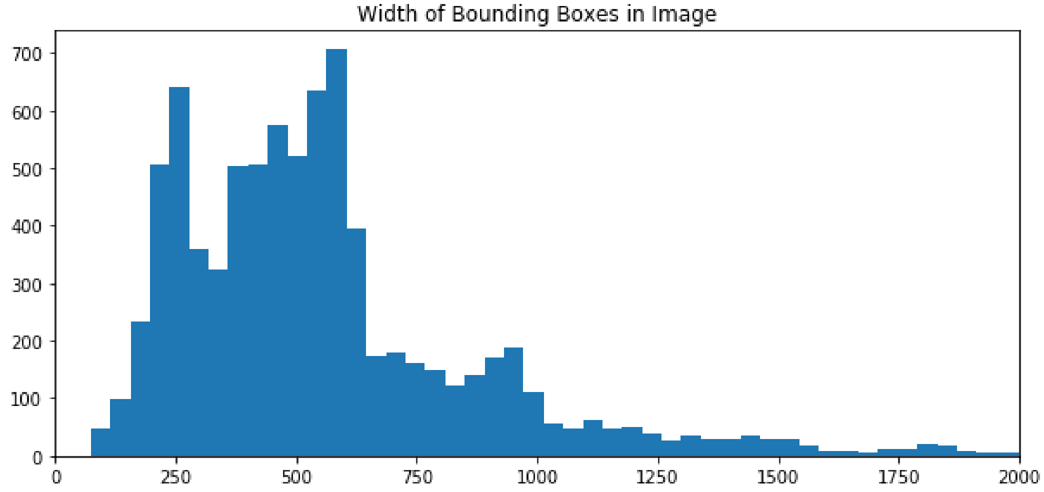
The below graph explores the number of images of the vehicles by their respective manufacturing year.



The recency bias on display in the dataset is acceptable for our model, primarily due to older vehicles being far less likely to be present on the road than newer models. 2012 remains the most popular and most recent manufacturing year for the dataset with nearly 5,000 of the 8,000 images being from cars released in 2012. In contrast, 2007 is the only other year in which cars manufactured have over 1000 images present in the dataset. That is also in line with automotive manufacturing trends as a new model for a car manufacturer arrives every four to five years while vehicles get small incremental updates annually in the years in between. The one shorting coming of the model is that it contains no images for vehicles prior to 1991, while vehicles made earlier can rarely be seen on the road often in shows and parking lots owners have more vintage vehicles that gain popularity after being decades old. The inclusion of images of vehicles decades old today could have proven valuable for predicting vehicles that most would not be able to easily identify themselves, an ideal use case for our model.

Finally, the last graph below illustrates the deviation in the boundary boxes for each vehicle. The two histograms contrast the height and width of the boundary boxes for cars inside the images.





As observed above the height boundary compared to the width is more centralized in size. The width is more varied allowing for the boundary boxes to be wider and more diverse. This can be explained by the dimensions of a vehicle which are almost always wider than are tall. The images plotted in the jupyter notebook displayed similar characteristics with the images being almost twice as wide compared to the height.

The visualizations yielded from the exploratory data analysis have garnered critical insights into our data. We know that the data is diverse amongst most leading automotive manufacturers, Mazda being the one exception. The most popular body styles for vehicles are captured in larger numbers in the dataset, and while the recency bias in manufacturing year is not in complete detriment to us, considering the use cases of our predictive model, the performance on classifying older vehicles through images will be something to pay attention to when modelling. The differences in the boundary boxes dimensions validate, images processed will, almost always, contain a larger number of pixels in an image’s width compared to its height.