David Kotaev
Watson
CS301

CS301 Project: Individual Contribution 2

See the project repo at https://github.com/flexadecimal/cs301-project - specifically, 'deliverables.ipnyb'.

For the second part of the project, I created a Juypter notebook to perform a stastical summary on our data. The first part of this involved importing them to Pandas dataframes and separating the horsepower and torque trial sets from our original data. I then wrote a couple of utility functions to perform the traditional five-figure stats summary (min, max, quartiles 1 – 3, outliers) on a dataframe that returns rows rather than just single numbers.

Our data set is fundamentally a set of trials, each with their own mods, where each trial is a continuous function of RPM to horsepower/torque. The owner-to-car relationship is one-to-many, and is seen in the table. Each car can have multiple runs. To perform a statistical summary, I chose to use an integral on fixed bounds for each trial. For an RPM vs. horsepower graph, this integral would be the total power produced over the run. For an RPM vs. torque graph, this would be the total force produced over the run. Like horsepower and torque, these two measures are very related but different.

A trial run is encoded as a list of (x,y) points – every run can have a different number of points. To do this statistical summary on the trials, we must use the same bounds, so I wrote a function that returns integration bounds by taking all the of the trials' start and end points and finding the median of these separately. The median start is the beginning bound, the median of the end trial points is the end bound. Then, for the horsepower and torque trial tables, I added a column for each trial's integral power/force integral. The statistical summary was then performed on this column for horsepower and torque sets – see the notebook for details.

Other than writing the notebook code to do this statistical summary, I explained to the group the reason we needed to use an aggregate function like the integral in order to "flatten" the trial runs into single numbers for statistical summary.