SC1015 Car Project (Overview):

Technical Components:

1. 10% for coming up with your own problem definition based on a dataset
2. 20% for the use of machine learning techniques to solve specific problem

Analytical Components:

1. 20% for exploratory data analysis/visualization to gather relevant insights
2. 20% for the presentation of data-driven insights and the recommendations

Supporting Components:

1. 10% for data preparation and cleaning to suit the problem of your choice
2. 10% for the quality of your final team presentation and overall impressions

Note: 10% for learning something new and doing something beyond this course

**Problem Statement:**

Can the features of a car predict it’s MSRP?

Data Set: <https://www.kaggle.com/CooperUnion/cardataset?select=data.csv>

Exploratory Analysis(Used):

* Linear Regression
* Classification tree
  + Task: To explore in scikit-earn existing model to fit the project.
* Explore train-test split for the data set(such as 80-20 split)
* Data Cleaning(Content)

Visualisation:

* Boxplot
* Correlation
* Data Cleaning (Presentation)
* Bar Plot

16th March: Explore other Scikit-learn model(Different models to be used)

1. do filter\_rule = (2000 == train\_set["MSRP"])

filter\_set = train\_set[filter\_rule] (CLEANING PROCESS) For cleaning, drop unnecessary columns etc

1. then .head
2. remove columns with NaN values
3. display year, make, msrp , change filter to 2000

4.1 have single data between 2000 to 10000 (remove this?) plot data b4 and after filter

1. replot rest of data
2. do random 80:20 (get all data from train test split)
3. which feature affect price most(boxplot – to know IQR median etc, regression etc, buyers’ POV) (do brand first then transmission)
4. auto, manual, auto-manual, direct drive, unknown (combine auto, auto-manual and direct drive? Then compare concat and manual) to determine which car to buy from buyer POV(license etc) using auto and manual(Boolean - tree)
5. a. bivariate plot (make and year of production against MSRP) year will affect MSRP even though from same make until a certain year then make determines MSRP (tree)

b. find out which year this happens, filter out years, then see if make determine MSRP

9. do boxplots for all makes, refer to exercise 3, manually set range of MSRP then show which makes match the range, refer to exercise 5

Include sth outside course code, give up on 10%?

New technique for cleaning data, model, evaluation matrix don’t rly need machine learning but if include then is considered extra etc

Good flow