PREDICTING CAR PRICE

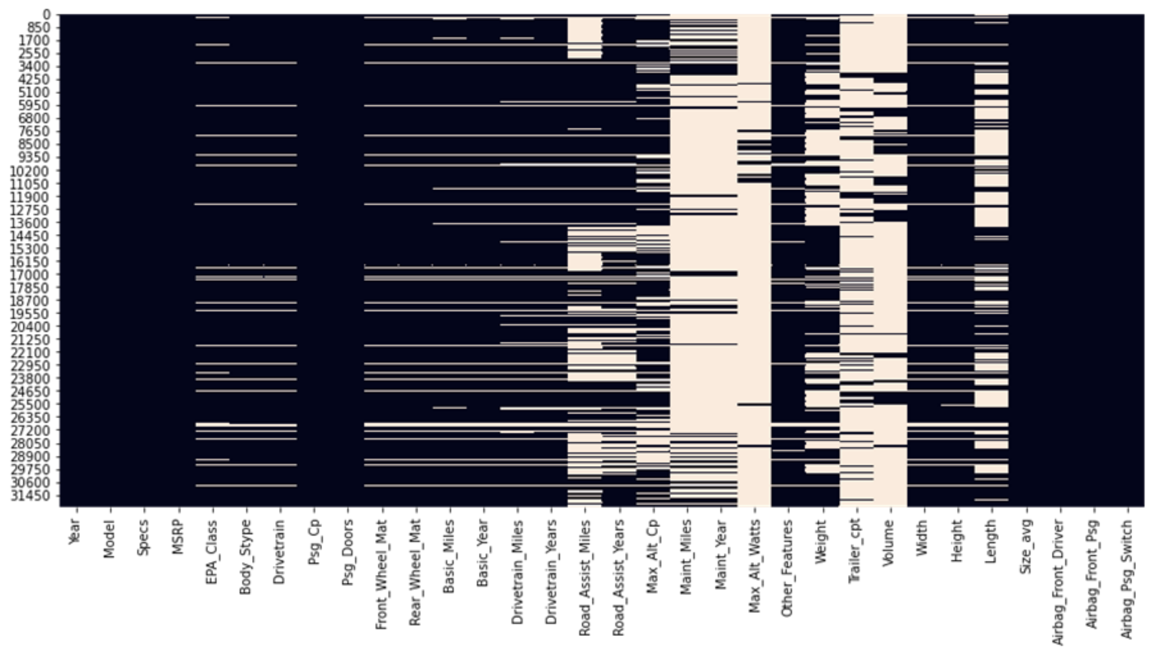
AUTO MACHINE LEARNING

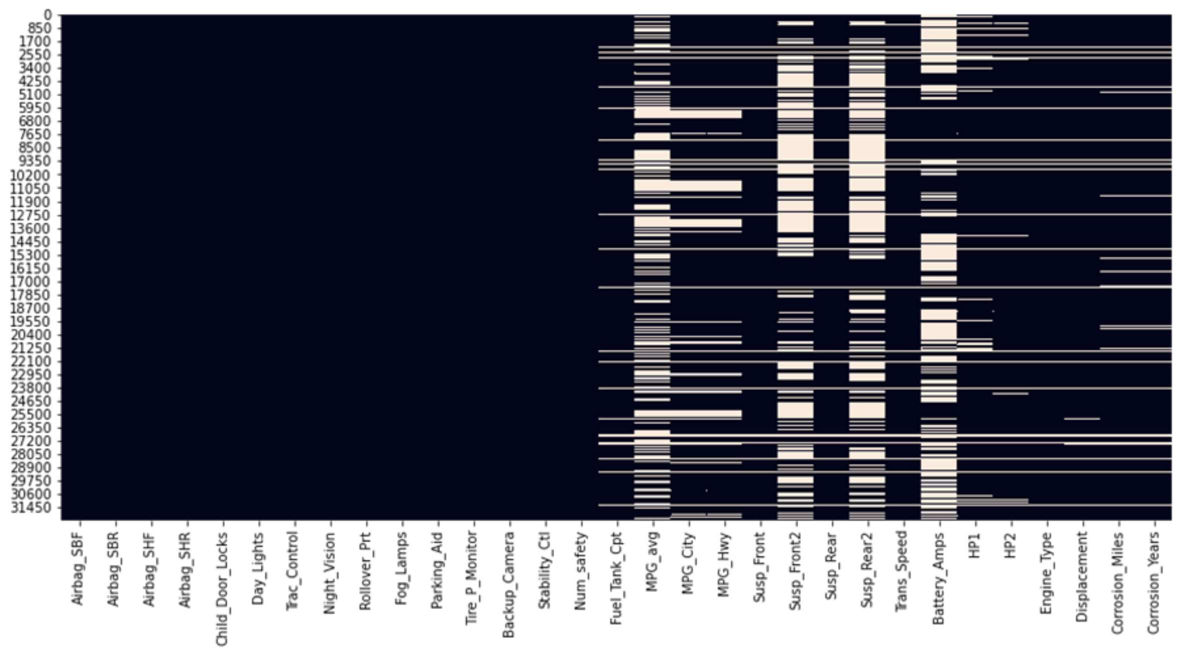
Dongtao Jiang, Data Science Career Track, Springboard
8/2/2020

DATASET

- Data source: thearconnection.com
- Features
  - Dimensions
  - Fuel economy
  - Performance specs
  - Safety features
  - Warranty

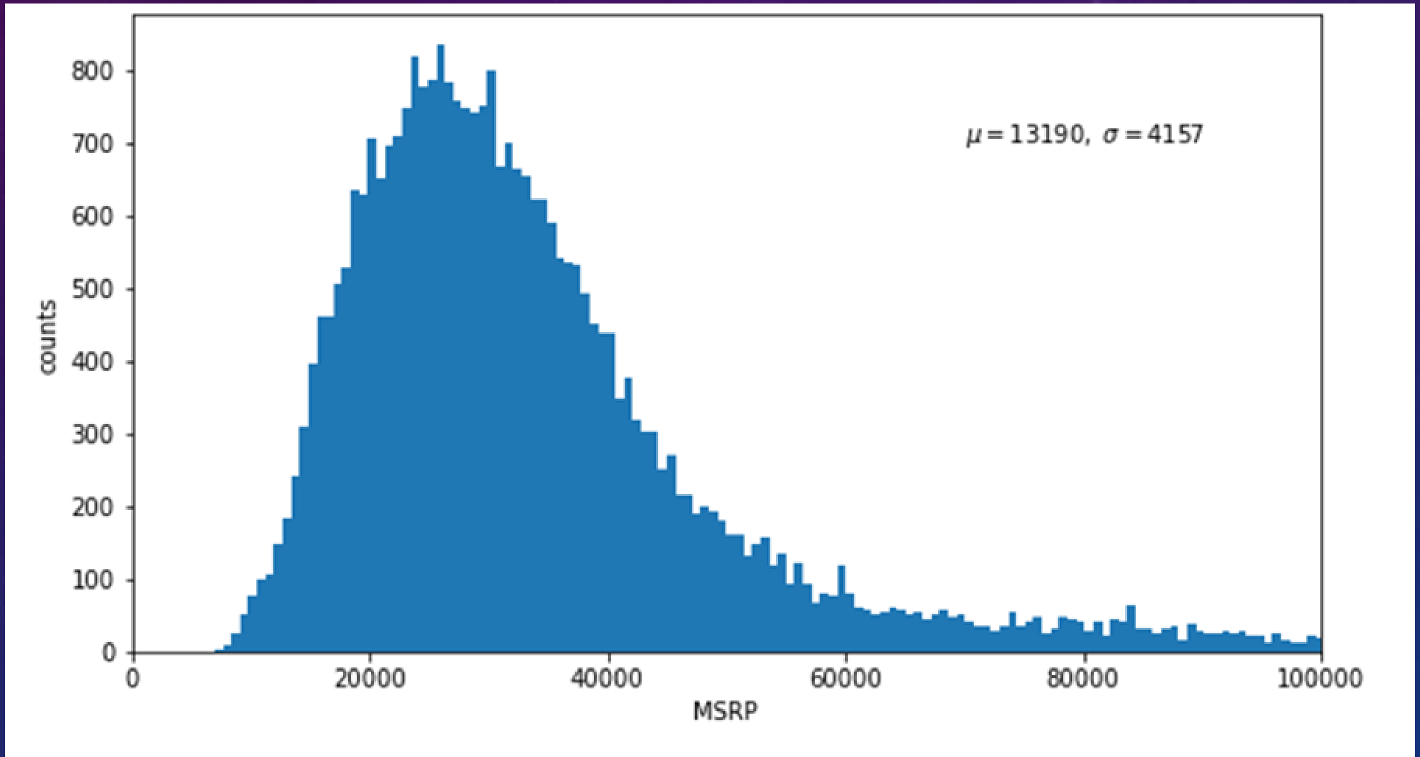
MISSING VALUES





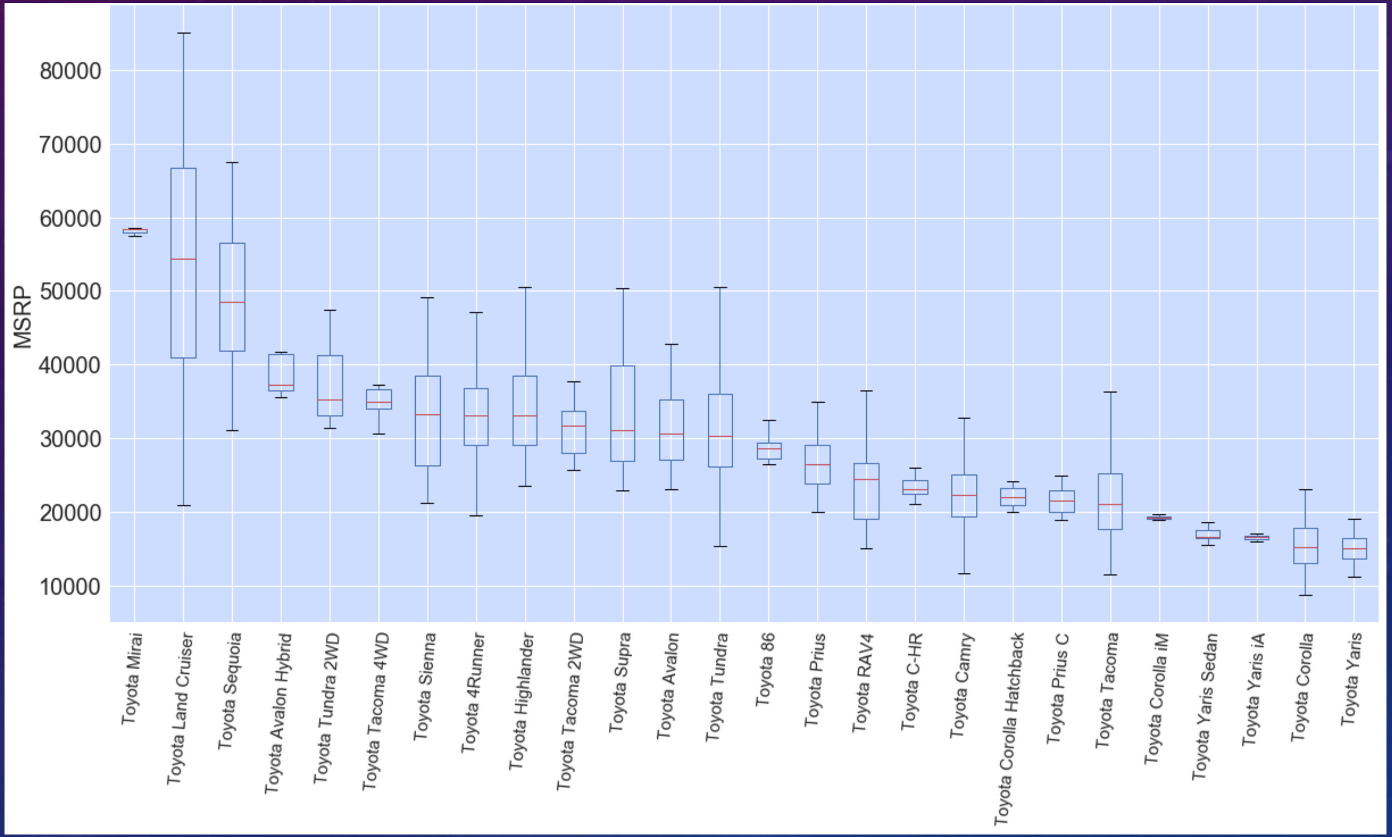
- 15.7% of total dataset are missing.
- Top missing features: Max_Alt_Watts Maint_Miles, Maint_Year, Volume

PRICE DISTRIBUTION



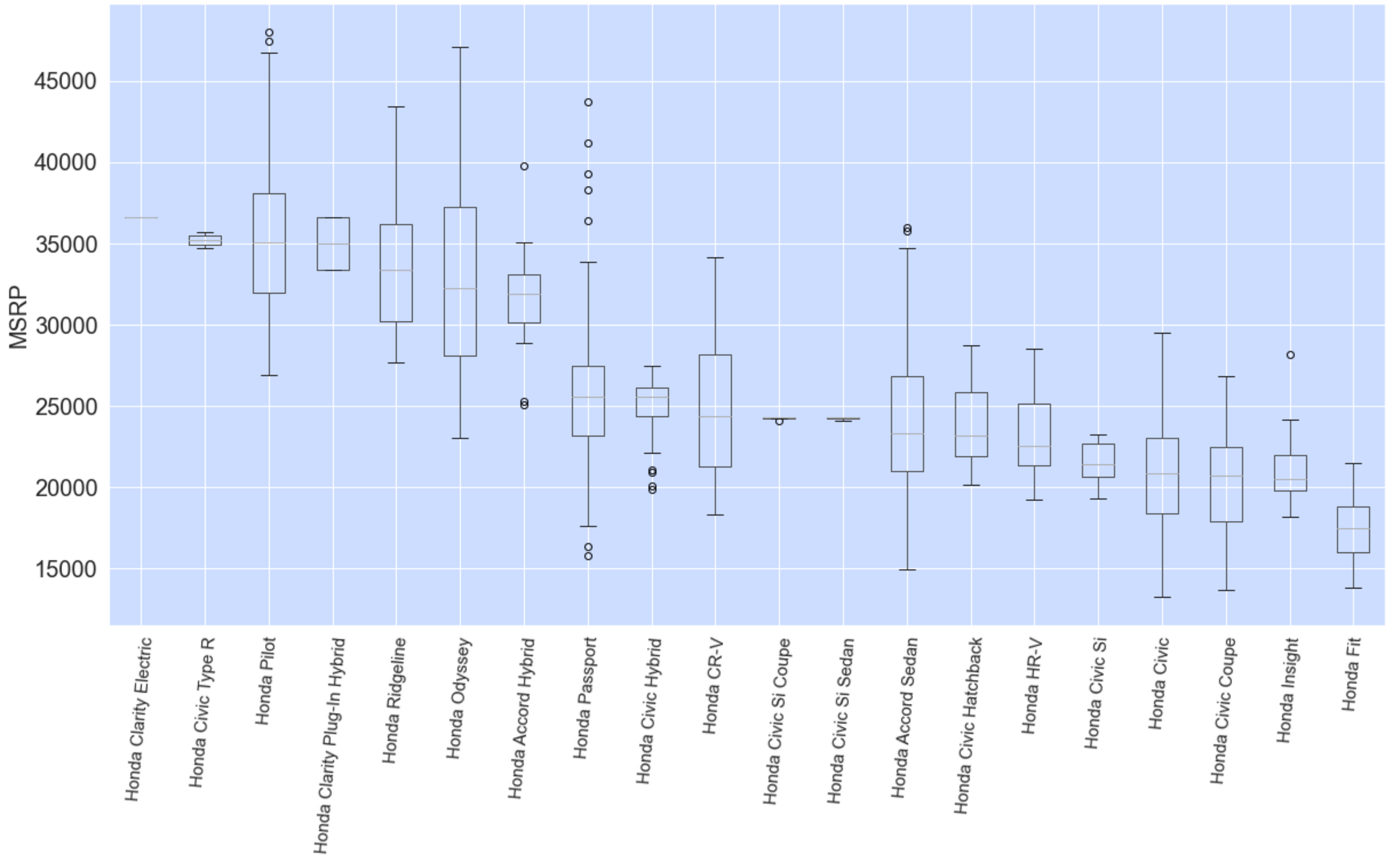
- Average price is $13200.
- Standard deviation is 4160.

TOYOTA PRICES SORTING – BOX PLOT



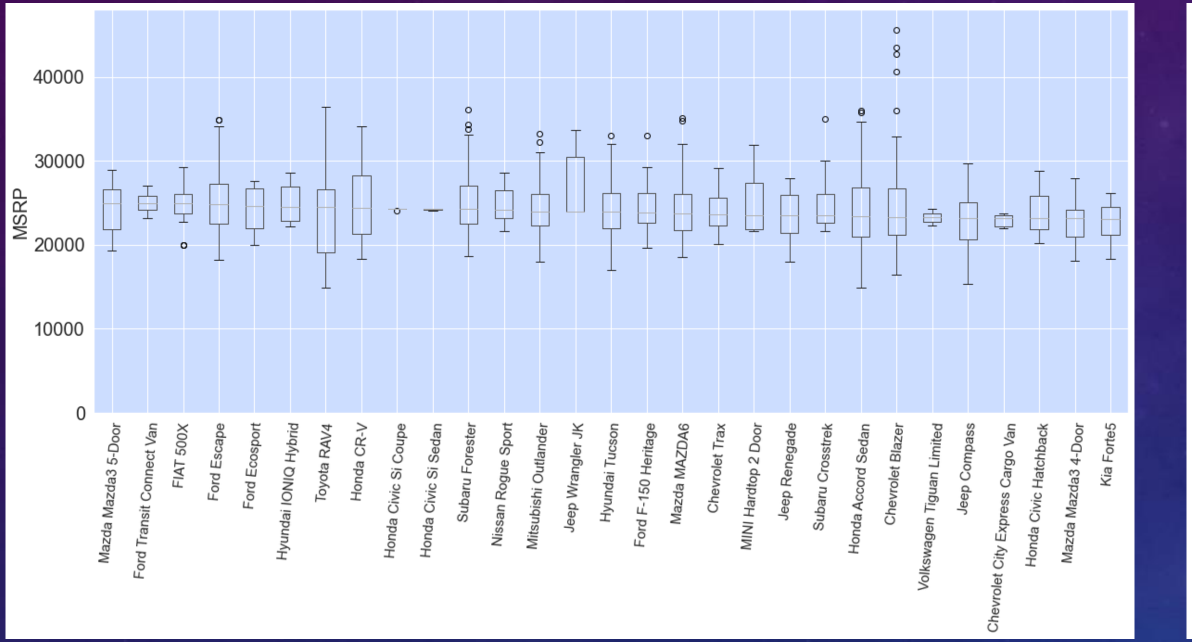
- Luxury model to basic model in sequence.

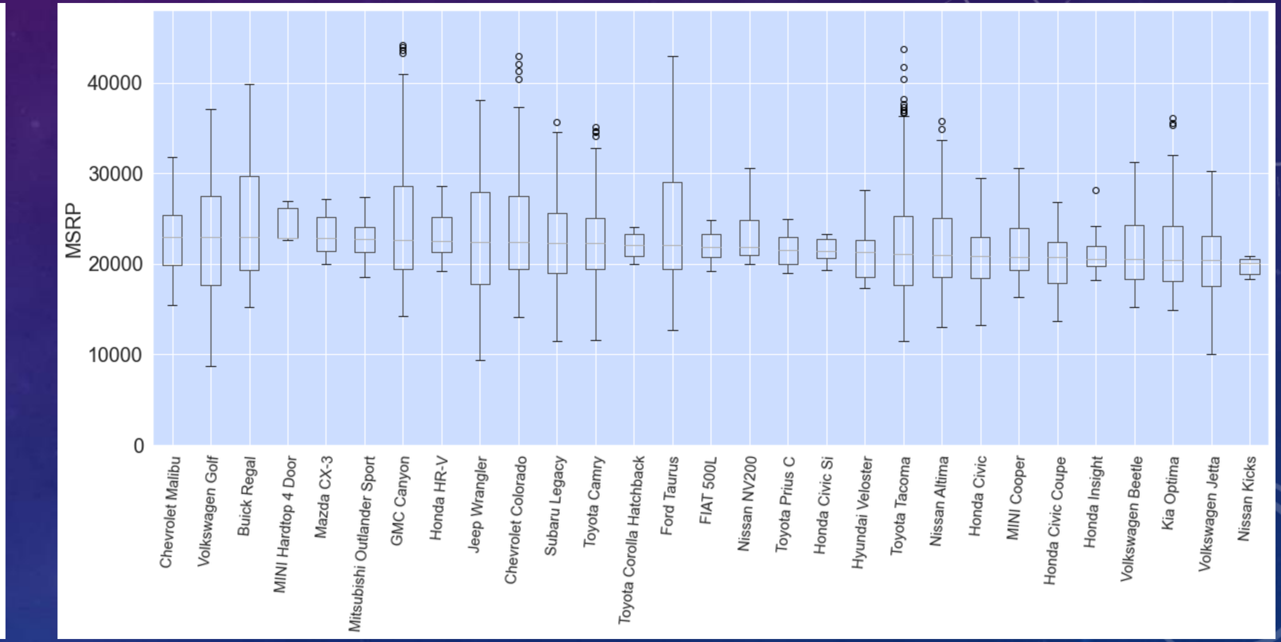
HONDA PRICES SORTING – BOX PLOT



- Luxury model to basic model in sequence.

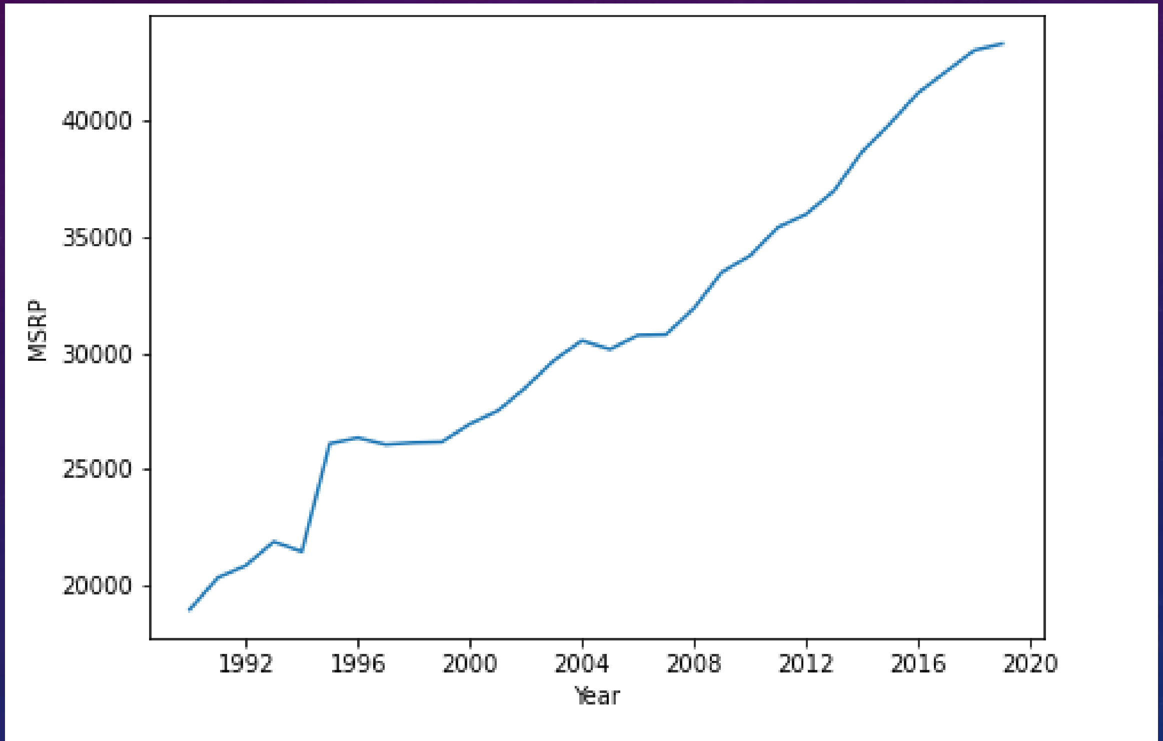
BUYING GUIDE FOR BUDGET-TIGHT CUSTOMERS





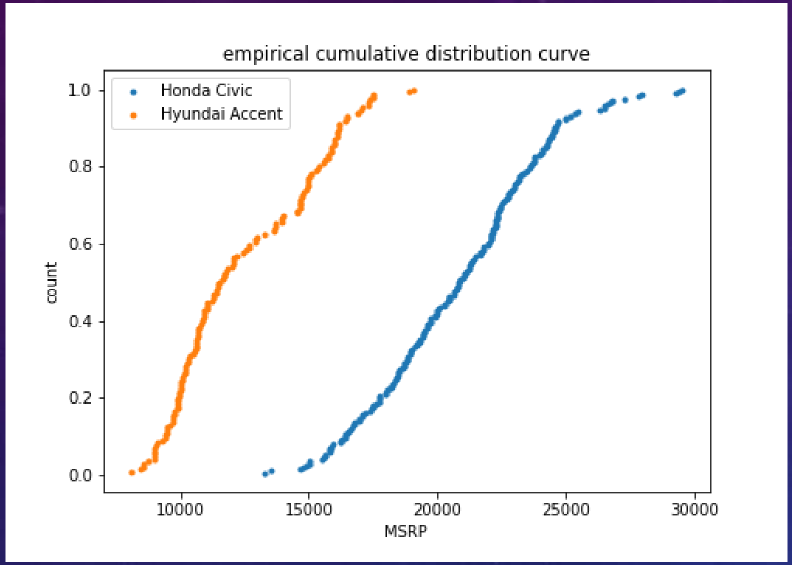
- Selection of low-end models priced about $22,000.

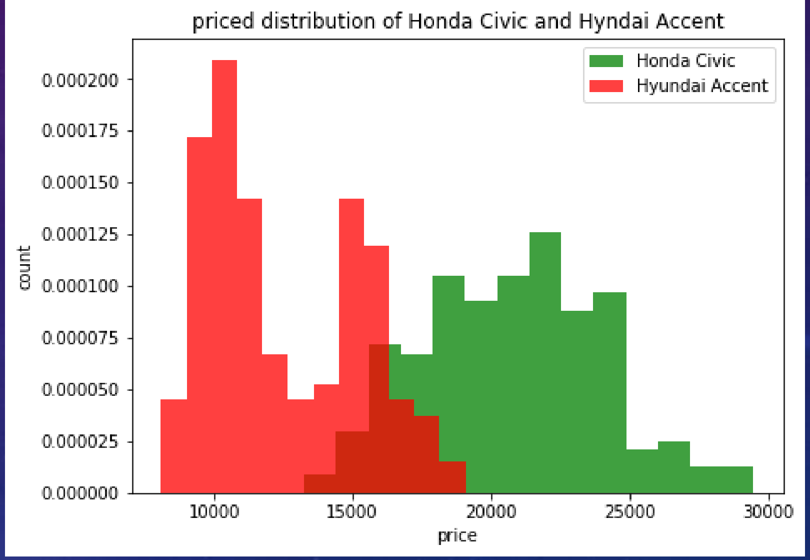
EVOLUTION OF AVERAGE PRICE OVER ALL MODELS



- Car prices have been constantly increasing generally linearly over the last two decades.
- This indicates year is an important feature to predict prices.

PRICE DISTRIBUTION FOR TWO CAR MODELS

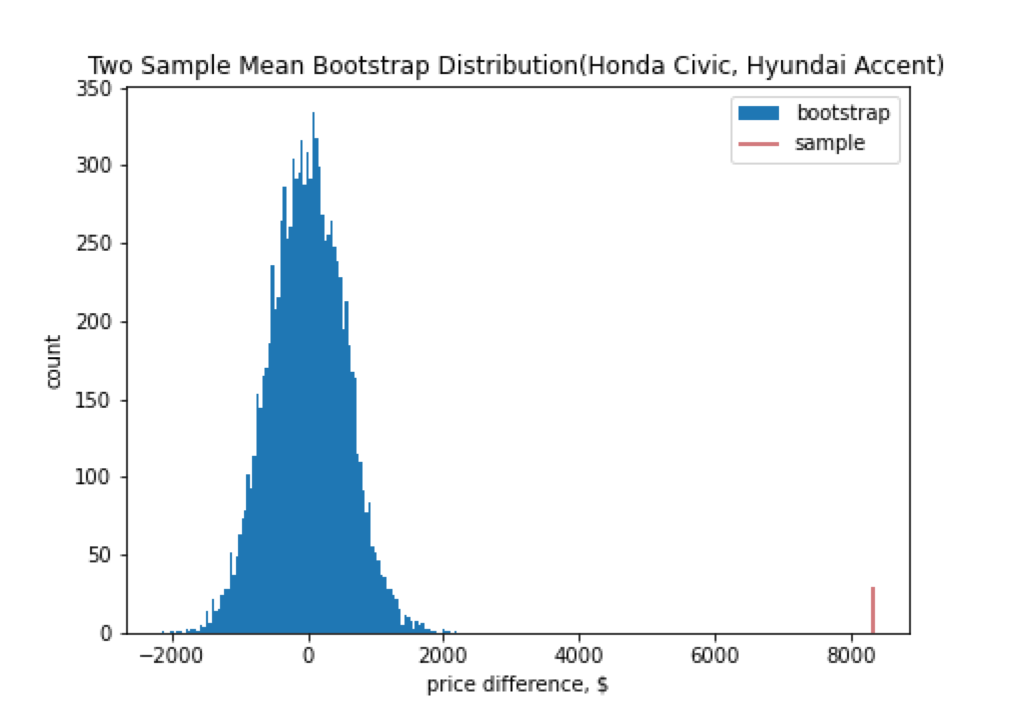




- Difference in prices of the two low-end car models is significant.
- A small fraction of price overlap exists.

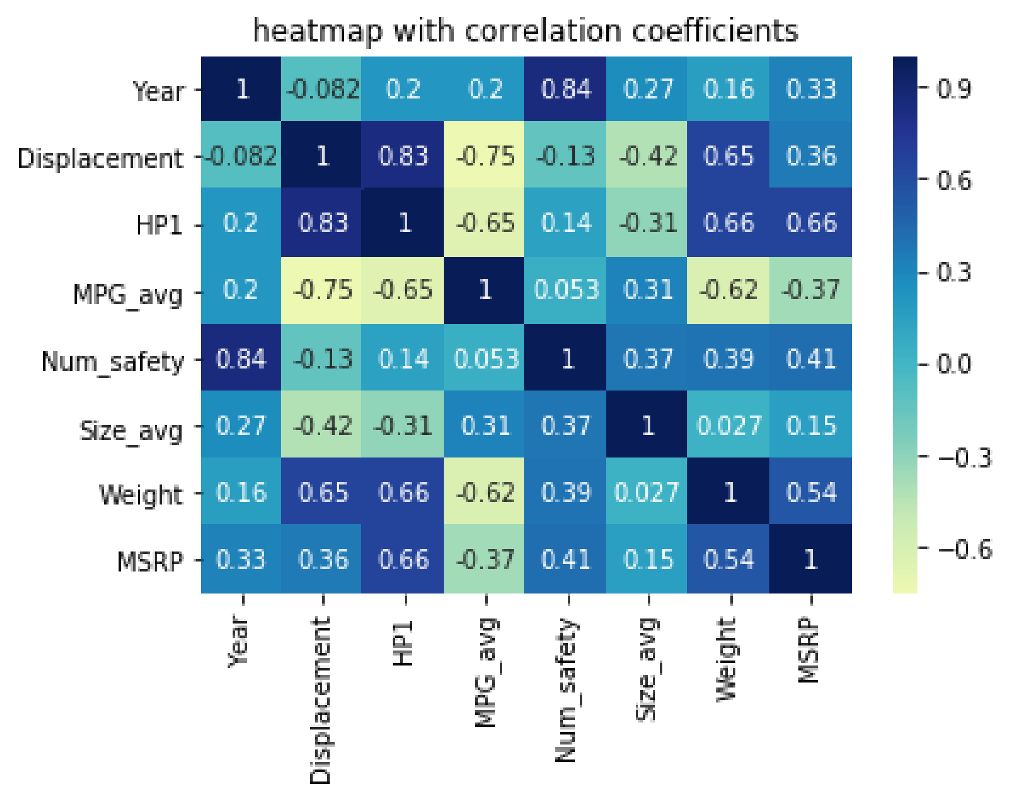
TWO-SAMPLE HYPOTHESIS TEST – BOOTSTRAP APPROACH

- Ho: There is no difference in the mean price between Hyundai Accent and Honda Civic.
- Ha: There is obvious difference in the mean price between Hyundai Accent and Honda Civic.
- alpha: 5%

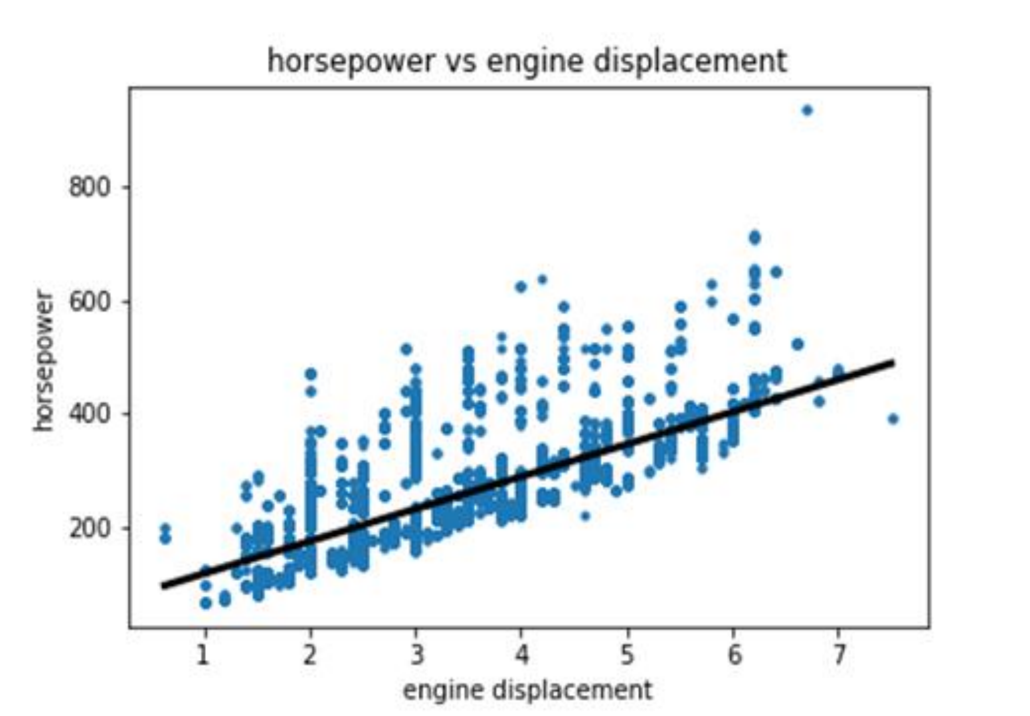


- Resutls:
  - p-value: 0
  - Null hypothesis rejected.
  - Approve the sharp difference in mean.

CORRELATION INVESTIGATIONS

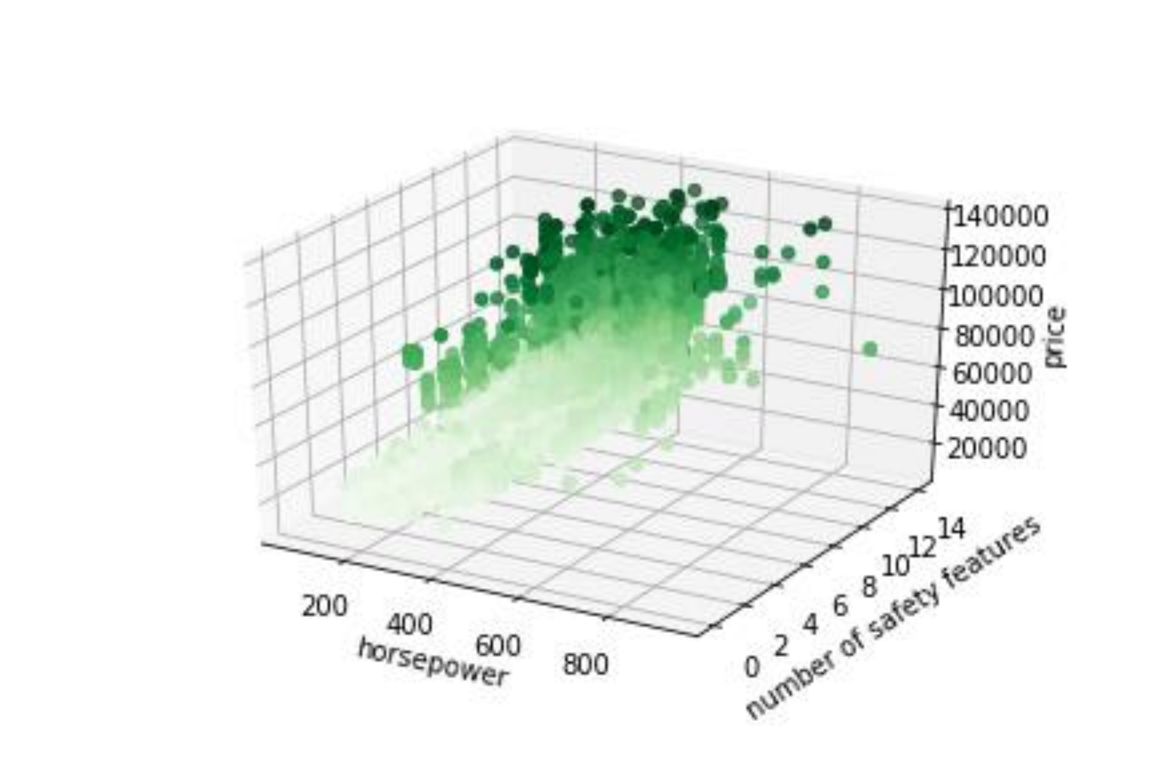


	Year	Displacement	HP1	MPG_avg	Num_safety	Size_avg	Weight	MSRP
Year	1	-0.082	0.2	0.2	0.84	0.27	0.16	0.33
Displacement	-0.082	1	0.83	-0.75	-0.13	-0.42	0.65	0.36
HP1	0.2	0.83	1	-0.65	0.14	-0.31	0.66	0.66
MPG_avg	0.2	-0.75	-0.65	1	0.053	0.31	-0.62	-0.37
Num_safety	0.84	-0.13	0.14	0.053	1	0.37	0.39	0.41
Size_avg	0.27	-0.42	-0.31	0.31	0.37	1	0.027	0.15
Weight	0.16	0.65	0.66	-0.62	0.39	0.027	1	0.54
MSRP	0.33	0.36	0.66	-0.37	0.41	0.15	0.54	1



- Strong correlation between horsepower and engine displacement agrees well with physics and engineering principles.

3-D VISUALIZATION



- 3-D plot is utilized to visualize the effect of two features combined on prices.

MACHINE LEARNING MODELS

- **Linear Regression**

  One of the most well-known and well understood algorithms in statistics and machine learning.

- **Ridge Regression**

  Useful to mitigate the problem of multicollinearity in linear regression.

- **Lasso Regression**

  Ideal for producing simpler models.

- **Decision Tree**

  Easy to interpret.

- **Random Forest**

  Merges multiple decision trees to get a more accurate and stable prediction.

MACHINE LEARNING

Models	r2_score	Negative mean_absolute_error
Random Forest	0.986	-1443
Linear regression	0.968	-1002483
Lasso Regression	0.967	-3262
Ridge Regression	0.964	-3525
Decision Tree	0.908	-5470

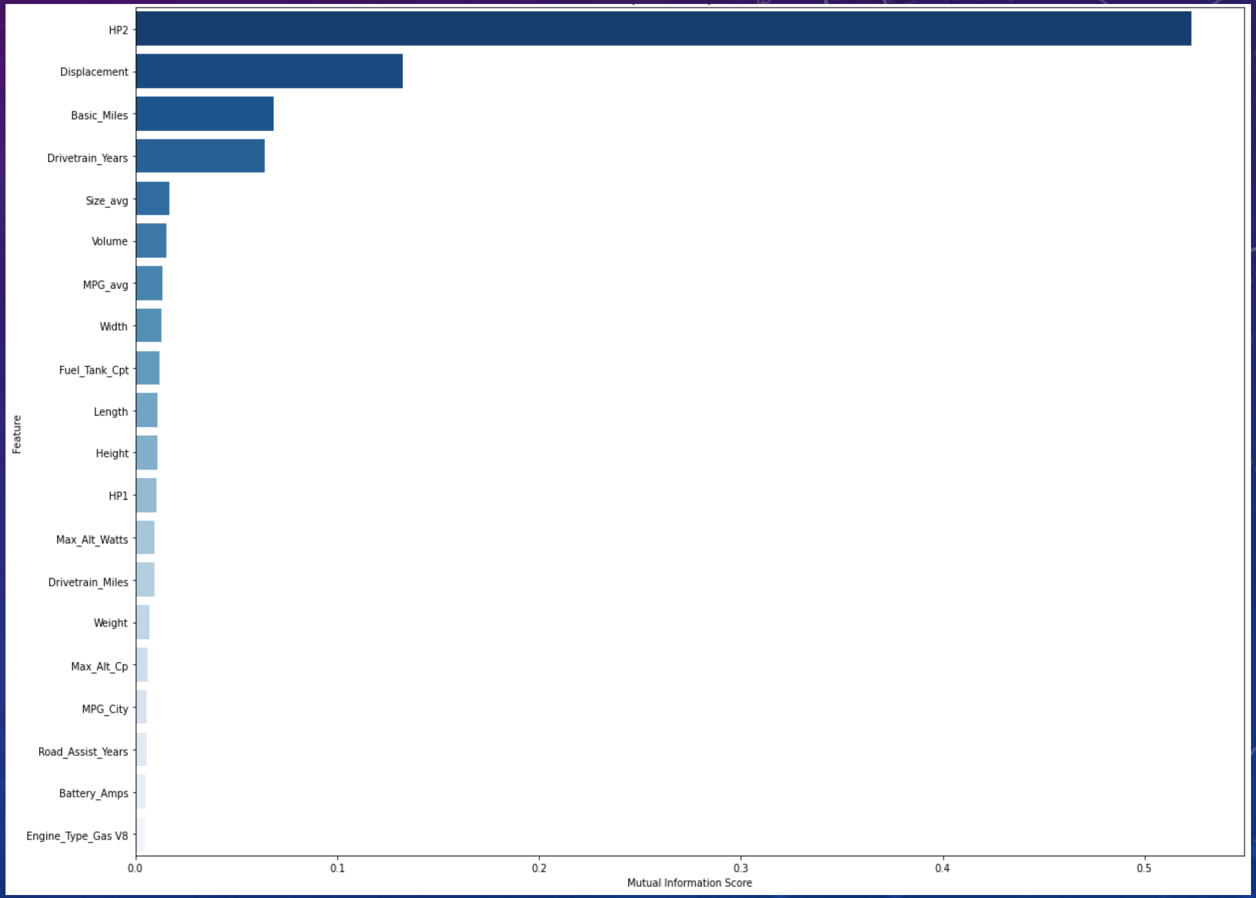
- Most of the models used give excellent goodness of fit that their r2 scores are mostly above 0.96. This indicate these models are of good choice.
- Random Forest gives the best performance with r2 being very close to 0.99. Decision Tree has the poorest performance with a r2_score of 0.908.

DIFFERENT METRICS

- Negative mean absolute error compared with r2.
- Decision Trees still tops the perforce.
- Scoring for Linear Regression is, however, exceedingly low.
- Performance of algorithms varies with metrics.

FEATURE IMPORTANCE

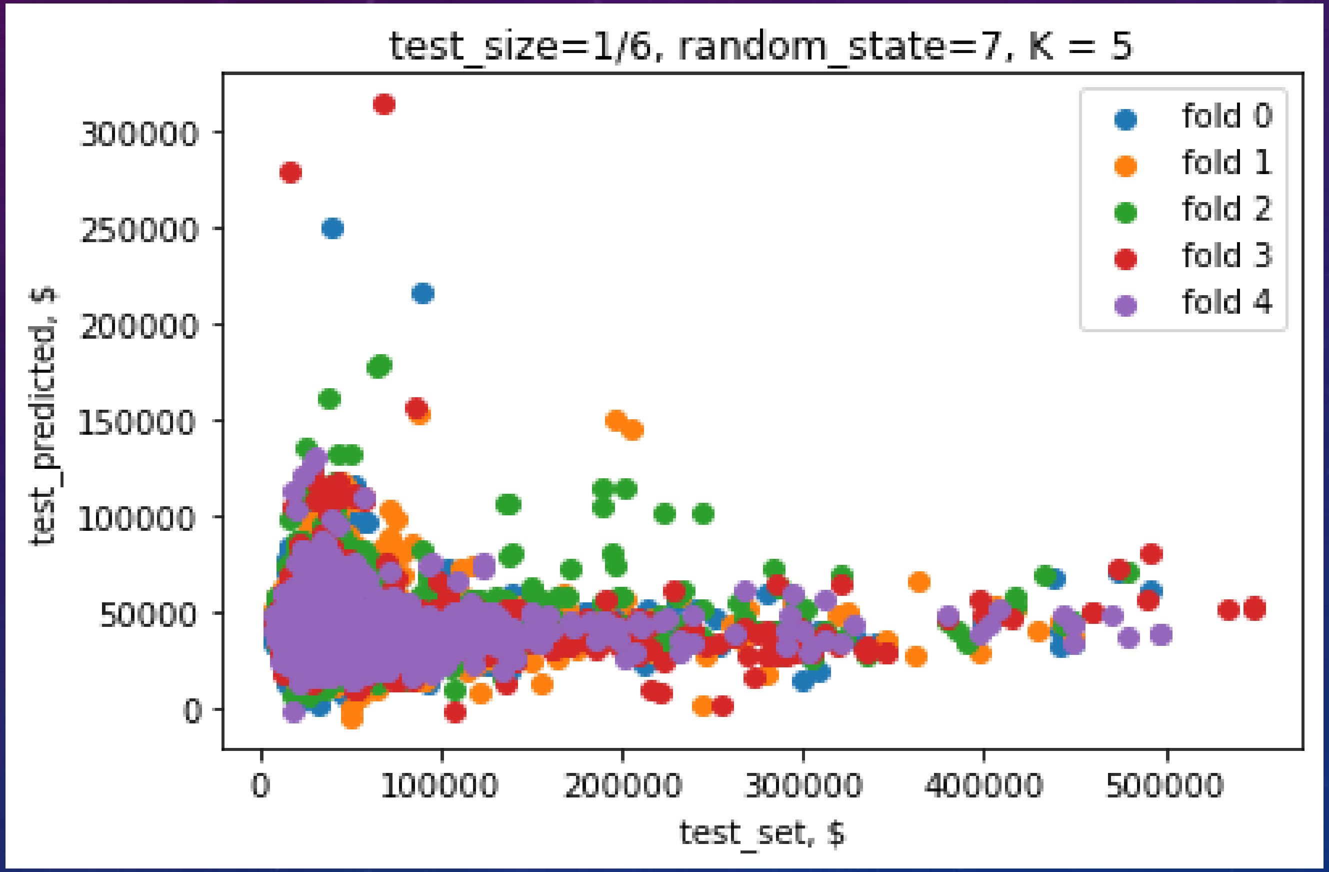
Rank	Feature	Importance
0	HP2	0.523067
1	Displacement	0.132579
2	Basic_Miles	0.068280
3	Drivetrain_Years	0.064147
4	Size_avg	0.016712
5	Volume	0.015565
6	MPG_avg	0.013185
7	Width	0.013044
8	Fuel_Tank_Cpt	0.011811
9	Length	0.011152
10	Height	0.011051



FEATURE IMPORTANCE

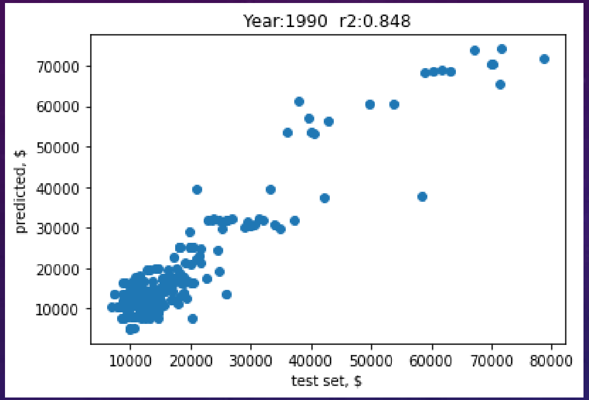
- The HP2 (torque spec) has the highest importance, which is at least 4 times higher than the rest.
- The 2nd highest is Displacement.
- This result agrees well with engineering know-how.
- Engine is the most dominating part for the major performance of car for example the car’s lifetime, speed, driving smoothness, horsepower, fuel efficiency, etc.

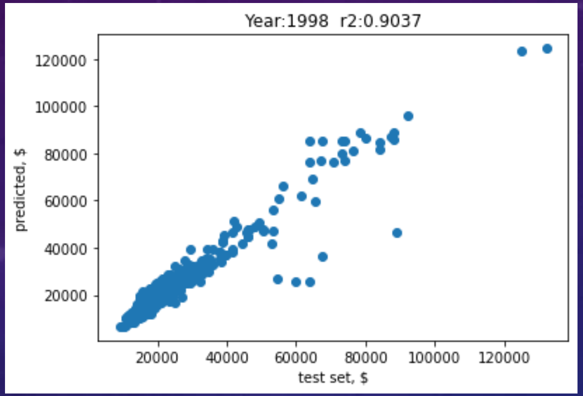
LINEAR REGRESSION ANOMALY

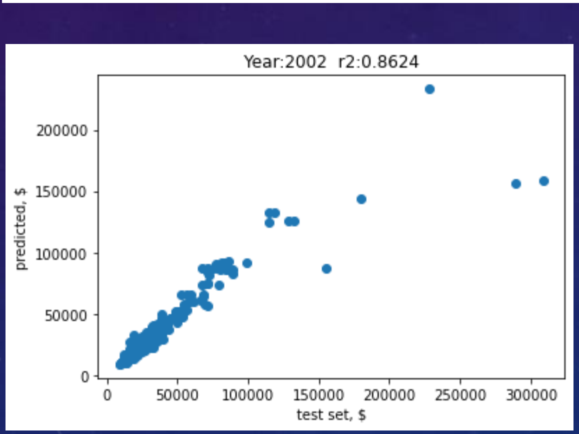


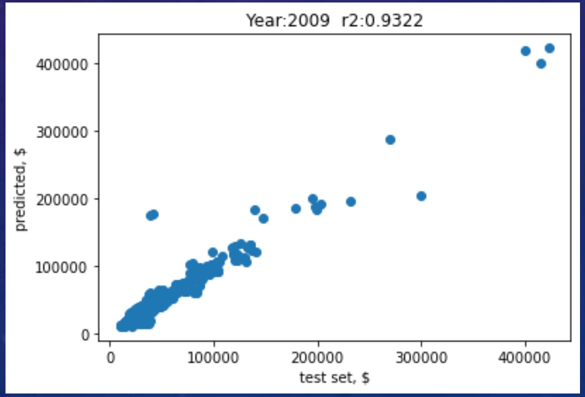
- Random split cause certain cars to be in the test set but not in the train set.

SET ASIDE ONE YEAR AS TEST SET









- Year was used as an integer

CONCLUDING REMARKS

- Hypothesis testing quantified significant difference in the mean price between two low-end popular car models: Hyundai Accent and Honda Civic.
- Exploratory data analysis was conducted to visualize missing values over all dataset, provide buying guide for low-income customers by extracting all lowly-priced car models and sorting in order.
- The pair plot and heatmap indicate a positive correlation between horsepower and engine displacement, which agrees well with physics and engineering principles.
- Machine learning algorithm Missforest was successfully used to imputate missing values, which is one of the reasons we end up with very high predicting accuracy.
- Five models were experimented including linear regression, ridge regression, lasso regression, decision trees and random forest. Except lower performance of decision trees, all the other models deliver very good r2 scores higher than 96%. The best model is random forest that scored close to 99%.
- Feature importance analysis revealed the torque spec (HP1) and displacement are the most important factor determining the car prices. This result indicates the power of random forest because these two features are related to the heart of car: engine.

FUTURE WORK

- Acquire more domain knowledge for the purpose of feature engineering
    - remove unnecessary features
    - create new features
- Tune models hyperparameters to marginally improve performance
- Obtain data for newer car models to test the model