HW-1 Auto Dataset

Jose Rodriguez, Johnny Nino Ladino, Dayana Sosa; University of Houston

This paper was prepared as partial completion of the course MATH 6350 (Statistical Learning & Data Mining) taught during the Fall 2020 semester at University of Houston

Background

This report will cover the questions presented by Dr. Azencott over the Auto dataset covered in the textbook, Statistical Learning. The cleaned auto dataset sent by Dr. Azencott is made up of 6 columns and 392 rows. The 6 columns are mpg, cylinders, displacement, horsepower, weight and acceleration. For the purpose of this dataset, mpg will be our target variable with the other 5 columns being our features. This dataset has no null values and the feature columns will be referred to each by the first 3 letters of each other name, respectively. This report will aim to answer all 14 questions presented in Dr. Azencott's document in a systematic way while analyzing each question and result to have an in depth understanding of this dataset. The underlying code sustaining this report will be created using python.

Methodology

The following methodology was used for developing the code and answering the questions in the document

1. Write code via python
2. ?
3. ?

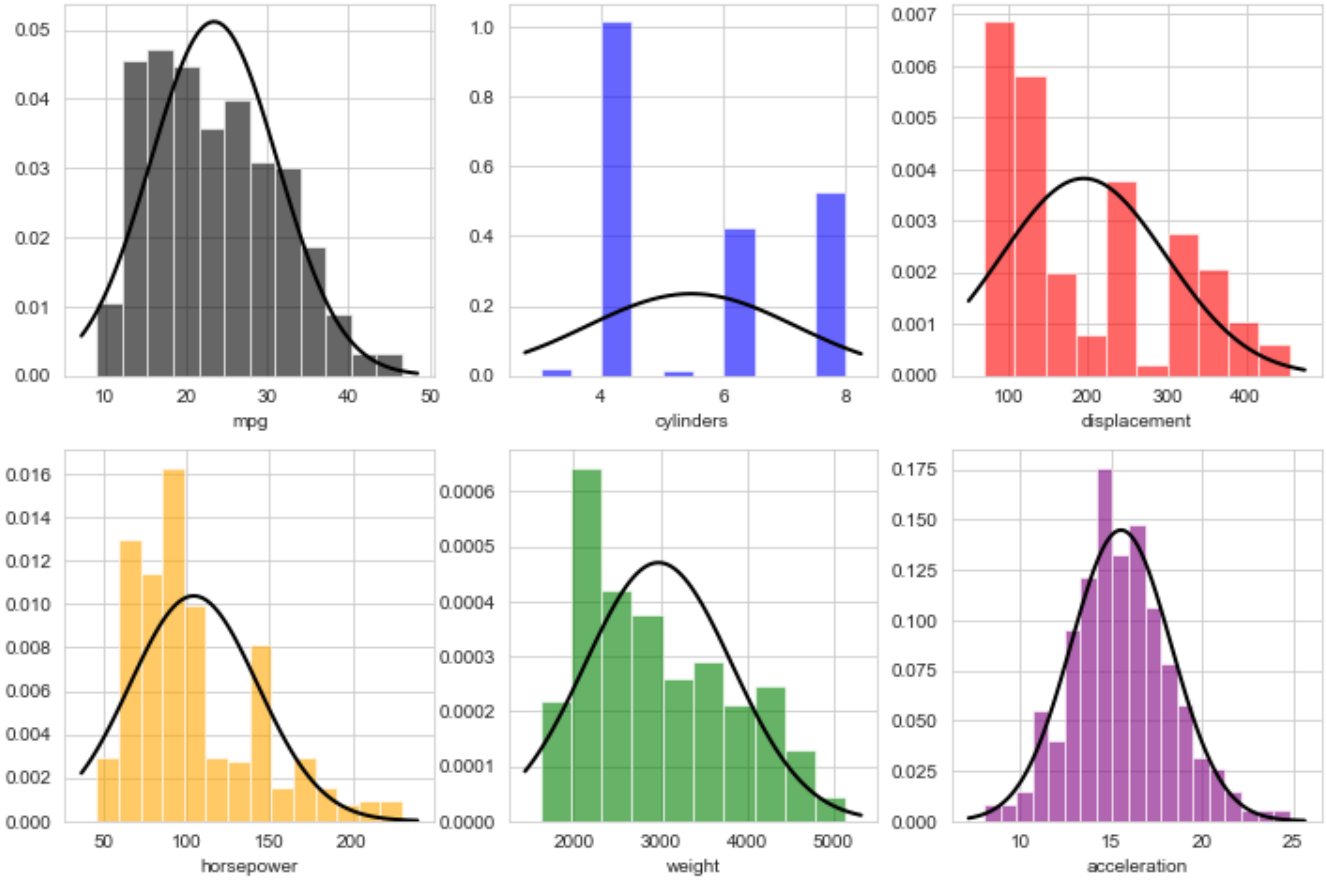
Question 1

The mean and standard deviation for each feature is show in the table below. Weight has the highest mean and standard deviation value while cylinder has the lowest mean and standard deviation. The values are not normalized so the mean has little meaning in terms of trying to compare the values between one another, but the standard deviation can be a better indicator of how spread out the data is. Both values need to be taken with a grain of salt however, since neither the mean nor standard deviation are normalized. If the user has knowledge about the underlying data, these values are a good way to check the data and see if anything seems out of place.

FEATURE	MEAN	STD. DEV.
CYLINDER	5.47	1.71
DISPLACEMENT	194.41	104.64
HORSEPOWER	104.47	38.49
WEIGHT	2977.58	849.40
ACCELERATION	15.54	2.76

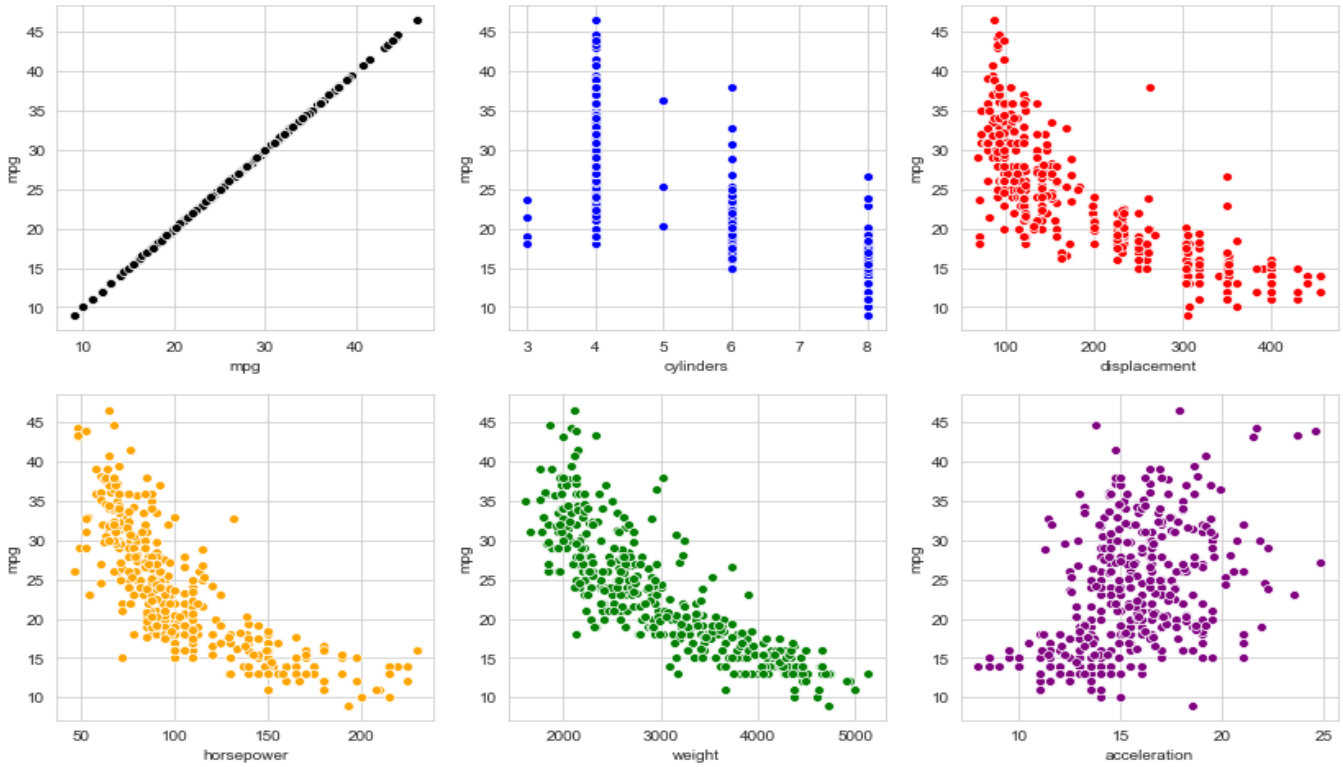
Question 2

The histogram of every feature and mpg is shown below. MPG and acceleration are the series that are the closest to a normal distribution. Weight, horsepower, and displacement's distribution seem to be skewed right. Cylinder's histogram does not have a distribution. This is mostly because most car engines are made with 4, 6, or 8 cylinders.



Question 3 & 4

Below are 6 scatterplots with 5 of them being a feature vs the target variable. By visually interpreting these plots, displacement, horsepower and weight all seem to have a very clear relationship with MPG. All three have a negative relationship with MPG that seem to be linear but could also be interpreted as a non-linear relationship. Acceleration seems to have a positive relationship that could also possibly be linear. Cylinder could be interpreted as a categorical variable and seems to have a negative relationship and could possibly fit a logistic regression. Displacement, horsepower and weight probably have the strongest capacity to predict MPG due to the clear trends.



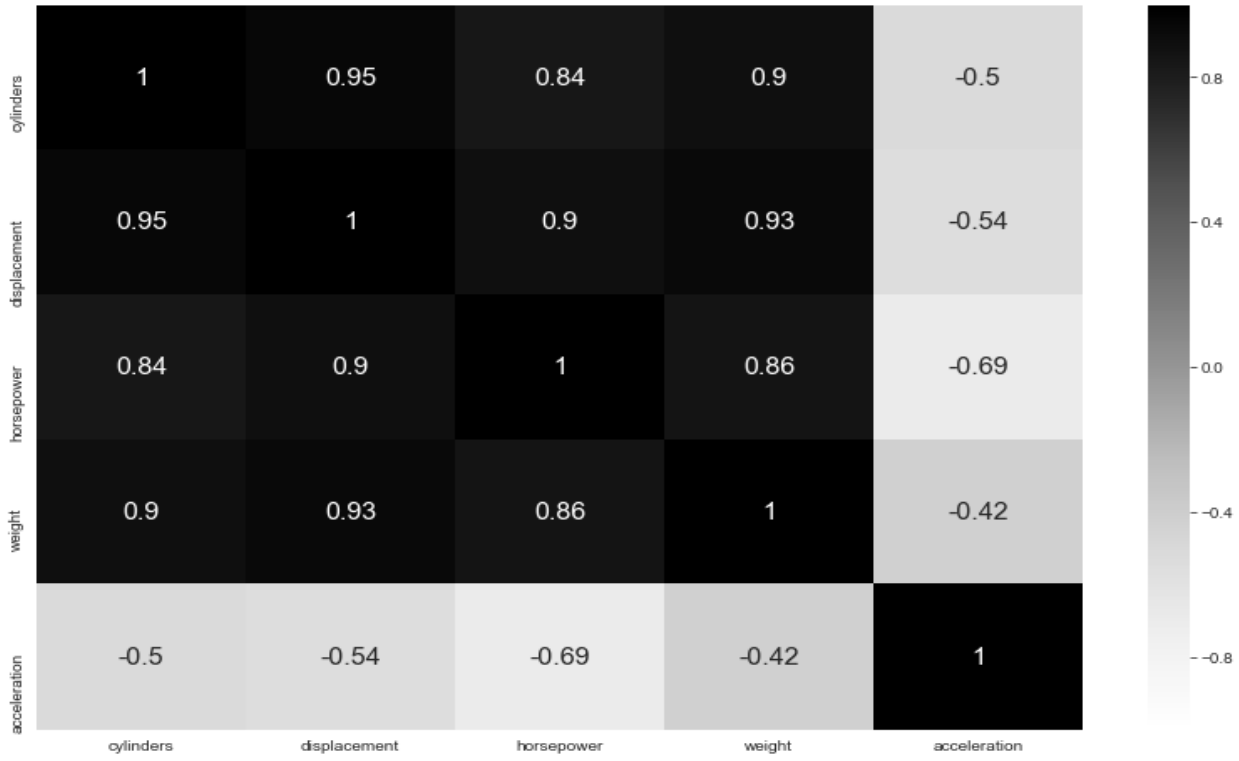
Question 5 & 6

Below are the 5 correlations with each feature and mpg and the correlation matrix of each of the features. According to the 5 correlations below, weight and MPG have the highest correlation with displacement coming in second. Cylinders and horsepower are tied in third place and acceleration has the lowest correlation. All but acceleration have a negative correlation which coincides with our findings in the scatterplot. It is interesting to see cylinder having such a high correlation value when that does not seem to be the case in the scatterplot. It is also interesting that all but acceleration seem to have relatively high correlation value and they are all above >0.7. The same conclusion as before still stands except now weight seems to have the highest predictor capability of the three previously hypothesized.

The correlation matrix below helps analyze which features may have collinearity and may not be needed for predicting mpg. It is interesting to note that some of the features have a very high correlation value between one another. Displacement and cylinders have a value 0.95 which is very high with displacement and weight have a value of 0.93. This indicates displacement could possibly be removed from a predictor variable since it has such a high collinearity with other features.

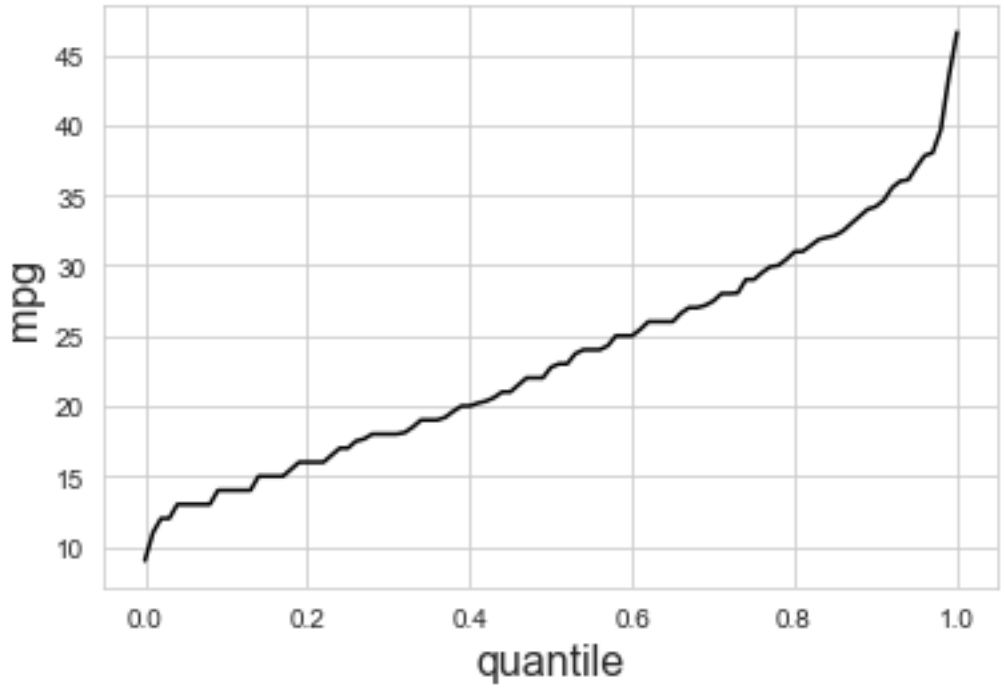
CORRELATION VARIABLES	CORRELATION VALUE
COR(CYL,MPG)	-0.78
COR(DIS,MPG)	-0.81
COR(HOR,MPG)	-0.78
COR(WEI,MPG)	-0.83
COR(ACC,MPG)	0.42

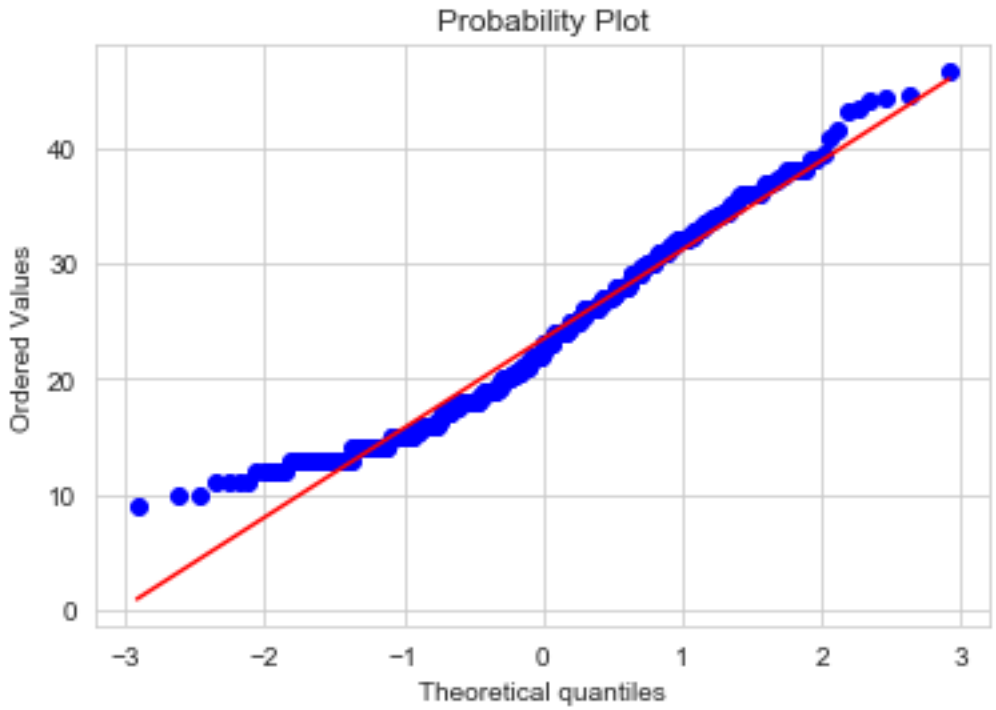
	cyl	*dis*	*hor*	*wei*	*acc*
cyl	**1.00**	**0.95**	**0.84**	**0.90**	**-0.50**
dis	**0.95**	**1.00**	**0.90**	**0.93**	**-0.54**
hor	**0.84**	**0.90**	**1.00**	**0.86**	**-0.69**
wei	**0.90**	**0.93**	**0.86**	**1.00**	**-0.42**
acc	**-0.50**	**-0.54**	**-0.69**	**-0.42**	**1.00**



Question 7

Below is the quantile curve plot for MPG. 17.5mpg is approximately the 30^{th} quantile for this dataset. 27.5mpg is approximately the 70^{th} quantile for this dataset.





This is the qqplot. Follows normal.

Question 8, 9 & 10

Above are the 5 pairs of histograms, one pair for each feature displayed side by side. Low MPG and High MPG are labeled at the top of each histogram to distinguish the two. Every pair has a shared x- axis to see if there is any bias in the data that may indicate good capacity to discriminate between high mpg and low mpg. According to the plots, Horsepower, Weight and Displacement may have good capacity to discriminate since the high mpg all lean towards the left-hand side. It is hard to tell with Cylinders but acceleration seems to have evenly distributed data. The more evenly distributed the data, the better and less change of data discrimination. Low MPG's histogram is skewed left while High MPG's histogram is skewed right. The same could be said for displacement.

Question 11

For Low MPG

	mpg	*cyl*	*dis*	*hor*	*wei*	*acc*
mean	*15.14539*	*7.407692*	*315.3077*	*145.6231*	*3937.331*	*13.77846*
std	*2.202575*	*1.00923*	*71.11404*	*35.84101*	*557.1857*	*2.649294*

For High MPG

	mpg	*cyl*	*dis*	*hor*	*wei*	*acc*
mean	*32.50606*	*4.083333*	*106.4015*	*74.39394*	*2226.091*	*16.56061*
std	*4.392095*	*0.391546*	*26.42225*	*13.88434*	*345.8779*	*2.519989*

We can see that the mean for High MPG is lower than the Low MPG's mean. The same can be said for cylinders, displacement, horsepower, weight. The difference is between acceleration, where the mean for acceleration for our high MPG is higher than the mean acceleration for Low MPG. The spread for cylinders and acceleration are about the same for both categories.

Question 12

Sample Standard Deviation and Discriminating Power

	s(F)	*disc(F)*
cylinders	*0.09458*	*35.14852*
displacement	*6.628266*	*31.51747*
horsepower	*3.358199*	*21.21052*
weight	*57.29843*	*29.86539*
acceleration	*0.319459*	*8.708918*

Ranking the discriminating power for each feature, cylinder has the highest discriminating power. Displacement and weight follow being the next ones with most discriminating power.

Question 13

Thr(F):

	thr(F)
Cyl	5.01256
Dis	162.9935
Hor	94.28258
Wei	2881.504
Acc	15.20433

ScoreF(n).head()

	cyl	dis	hor	wei	Acc	fullSCORE(n)
0	-1	-1	-1	-1	-1	-5
1	-1	-1	-1	-1	-1	-5
2	-1	-1	-1	-1	-1	-5
3	-1	-1	-1	-1	-1	-5
4	-1	-1	-1	-1	-1	-5

Here we have a threshold that we will use to assign a score for each case n's feature. We will then sum the numbers for each case to give the case a score. After we will use the score to determine what is the classification, in this case, High or Low MPG.

Question 14

When $a = 1$:
Accuracy Score = 65.56%

When $a = 2$:
Accuracy Score = 71.68%

When $a = 3$:
Accuracy Score = 71.68%

We see that when a = 1, our classifier is only choosing between High MPG and Low MPG. This will not always be the case, as there are cases where MPG's are neither high nor low, but instead Neutral. This makes sense, because when a is changed, we can then allow a third classification to be made, which is neutral. In our case, we let our classifier determine that if the full score does not fit neither high or low mpg cases, to classify it as neutral.

Conclusions

To conclude we were able to create a classifier that can classify our target variable with up to 71% accuracy.