Homework 1 - Solution

Eris Azizaj

September 27, 2018

Problem 1. Use summary() on the dataset for a preliminary exploration. Comment on the results. Are the features balanced? Are they in the right range? How many NAs? How do you suggest dealing with possible problems?

Answer: First thing to do is to check the summary statistics and the class of the variables. When we do that we see that some variables like *cylinder, origin, model_year* should be categorical because it takes only a few values. Another thing that captures my attention are the NA's at *horsepower* variable which tells us that the data is not balanced. Since there are only 6 NA's (few compared to the total number of observations we have), I suggest removing these observations before doing any analysis and this will not have a significant impact in our analysis results.

Problem 2. Report the mean value and standard deviation for numerical attributes. Remove NA's as necessary before computing these statistics.

Answer:

Table 1: Mean and Standard deviations

	mpg	displacement	horsepower	weight	acceleration
means	23.446	194.412	104.469	$2,977.584$	15.541
std_dev	7.805	104.644	38.491	849.403	2.759

Problem 3. Plot the correlation matrix for attributes 1-7. Interpret the results.

Answer: Table 2 shows the correlation between the numerical variables. Variables *displacement, horsepower, weight* have a strong and negative correlation with *mpg.* This tells us that cars that tend to have higher *mpg* tend to have lower *displacement, horsepower* and *weight.* It is the opposite for for the variable *acceleration* since it is positively correlated with *mpg* but has a weaker correlation. We can say that cars that tend to have high *mpg* tend to have high *acceleration* but the correlation is not that strong as it used to be for *displacement, horsepower* and *weight.* Same logic applies to other results. Variable *displacement* has a high and positive correlation with *horsepower* and *weight* but a negative correlation with *mpg* and *acceleration.* The sign of each number in the correlation matrix tells us the direction while the number tells the strength of the correlation between the two variables.

Table 2: Correlation Matrix

	mpg	displacement	horsepower	weight	acceleration
mpg	1	-0.805	-0.778	-0.832	0.423
displacement	-0.805	1	0.897	0.933	-0.544
horsepower	-0.778	0.897	1	0.865	-0.689
weight	-0.832	0.933	0.865	1	-0.417
acceleration	0.423	-0.544	-0.689	-0.417	1

Problem 4. Create a scatter plot for attributes 1 and 4 of your dataset and a second scatter plot for attributes 3 and 5. Then, add the number of cylinders to both scatter plots using dots of different colors. Add the 4 scatter plots to your report and interpret your findings.

Answer: This question support our findings from the correlation matrix. As we can see from the Fig 1 upper plots, cars that have high *mpg* tend to have lower *horsepower* and cars that have high displacements tend to be more heavy in weight. By adding the *cylinder* variable we are distinguishing between the categories. As we can see from the plots, cars that have 8 cylinders have low *mpg* but high *displacement, horsepower* and *weight* which is exactly the same as we found from the correlation matrix. Cars of cylinder category of 4 have high *mpg* but low *horsepower, displacement* and *weight.*

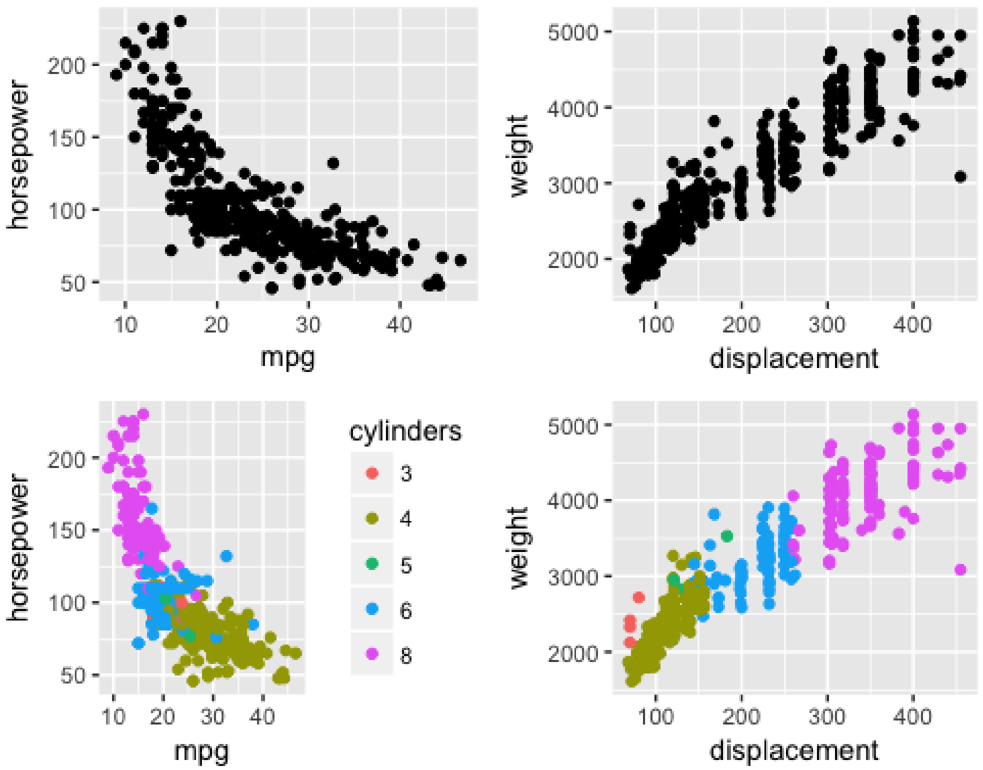

Figure 1: mpg_horsepower AND displacement_weight plots

Problem 5. Create histograms and density plots for attributes 3, 4 and 5. How did you choose the bin size for the different histograms? What do these plots tell us about the variable distributions? Do you notice outliers or skewedness in the data?

Answer: In the beggining I tried different possible values for the bin size before deciding for the final one. I tried to choose the bin size that best describe the variable distribution. Another indicator of deciding the bin size might be the summary statistics. For example the variable *weight* has a much wider range than the other two variables, so the bin size will be larger. Later, I found this website article useful. We take the inverse of the cubic root of the number of observations, multiply it with the 3.49 times the standard deviation of the variable. As we can see from Fig 2 the distribution from the histogram plots is very similar to the distribution of the density plots. We can see that *displacement* and *horsepower* has outliers as the distribution in the left hand side is very small, however we can not say the same thing for *weight*.

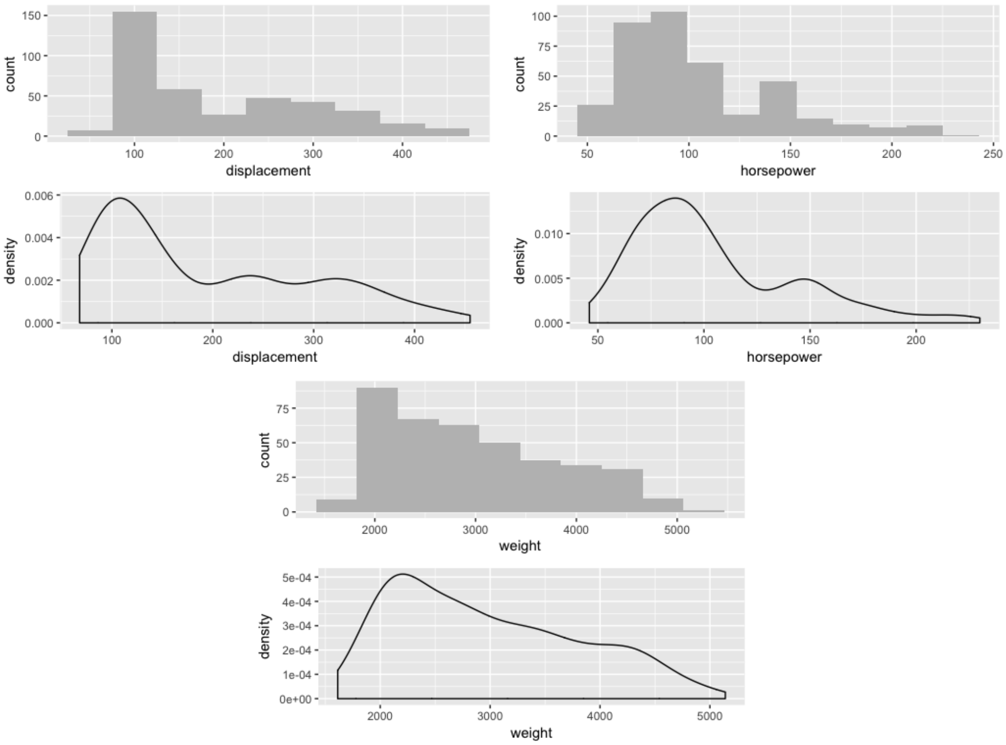

Figure 2: Histograms and Density plots

Problem 6. Create box plots for attributes 4, 5 and 6; one for the whole dataset, one divided by cylinder number, and one divided by origin. Interpret the obtained 9 boxplots.

Answer: Variable **horsepower**'s boxplot seem to have a median of 93.5 with 25th and 75th percentile of value 75 and 126 respectively, 50% of the data lie between these two values. The min and max non-outlier values are around 47 and 200. It has outliers above the value 200. When divided by the number of cylinders, 50% of the data lie on a smalller range. The median is around 35, the 25th and 75th percentiles are 27 and 43 respectively. The min and max non-outliers values are around 18 and 58. It has outliers above the value 90. When divided by the origin the datapoint are more spread as we can see by the box being at larger ranges. **Weight** variable have median around 2804 with 25th and 75th percentiles 2225 and 3615, respectively. It has no outliers. The min and max non-outlier values are around 1600 and 5100. When divided by the number of cylinders seems to have a median of 950 with 25th and 75th percentile 800 and 1125 respectively. It starts having outliers. When divided by the origin the distribution range is much larger than it used to be in the beginning where now 50% of the data lie between 1000 and 3550. **Acceleration** variable have median 15.5 with 25th and 75th percentile of value 13.78 and 17.02, respectively. It has outliers from below and above. When divided by the number of cylinders it has median of 6.5 and 50% of the data lie between values 3 and 8. When divided by the origin median is 12.5 with 25th and 75th percentile of value 7.5 and 16 respectively

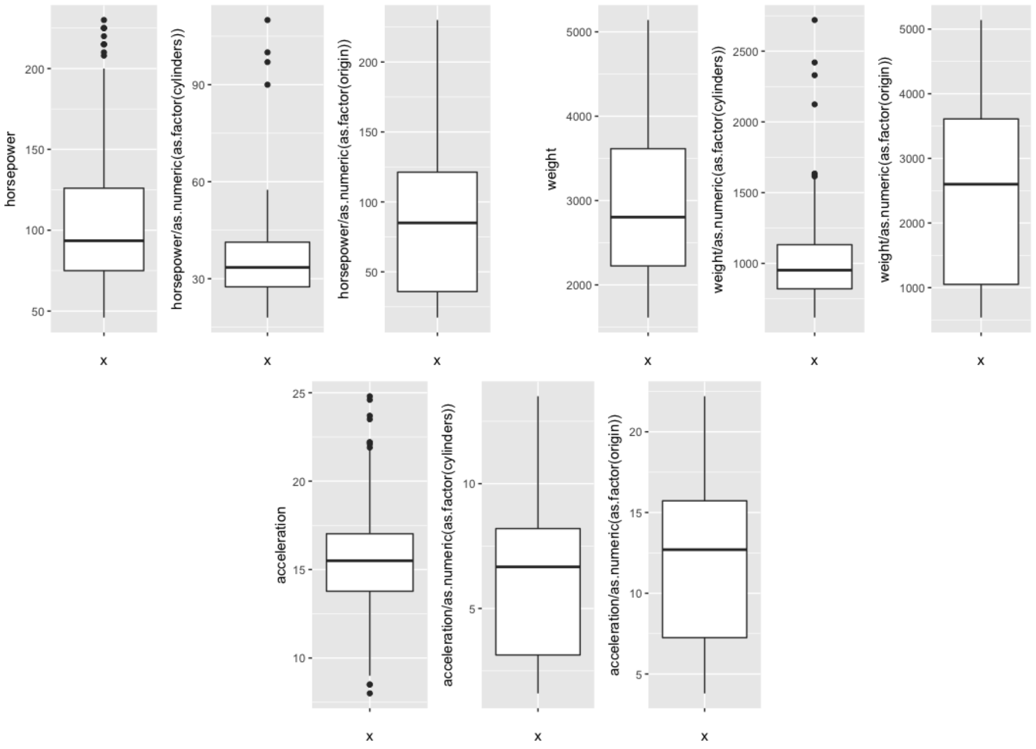

Figure 3: Boxplots for horsepower, weight and acceleration

Problem 7. Next, we will try to predict the mileage per gallon of a car using linear regression. Complete the following steps:

a. Split the dataset in training (75%) and testing set (25%). You can use any method, but explain how you ensured that the final distribution is correct.

   Answer: I am splitting into the main dataset into test and training datasets using a random group mark. We need to use set.seed() function in order to generate the same numbers each time and call the new variable *gp*. Then using the subset() function and the condition on *gp* we create the test and training data.

b. Train a simple linear regression model that predicts mpg as a function of horsepower.

   Answer: Done in the code file

c. Report the R^2 and the $RMSE$ (for training set) of the linear model and the coefficient in the obtained regression function. What can we tell about the relationship between mpg and horsepower? Is it significant?

Answer: $R^2 = 0.614$ and the $RMSE = 4.856$. As we can see from Table 3 the relation between the *mpg* and *horsepower* is negative which is not surprising since we have found the same relation in the previous exercises. The coefficient is significant at 99.9% level.

Table 3: Regression results

	Estimate	Std. Error	t value	Pr($\|t\|$)
(Intercept)	39.858807	0.812233	49.07	<2e-16 ***
horsepower	-0.157608	0.007298	-21.60	<2e-16 ***

*p<0.1; **p<0.05; ***p<0.01

d. Predict mpg for the testing set data. Report the resulting R^2 and $RMSE$. How does it compare to the training set?

Answer: The results are slightly different from the train data. $R^2 = 0.58$ and the $RMSE = 5.0048$. The R^2 in test data is lower compared to the train data while $RMSE$ is larger.

e. Repeat steps b-d using all available variables. Explain your findings. Compare this model to the simple linear regression.

Answer: First I am regressing *mpg* on all other variables except *car_name* from the train dataset . I found $R^2 = 0.879$ and the $RMSE = 2.714$ which are much better than the previous model. Remember that we want R^2 to be as much as possible close to 1 and $RMSE$ as small as possible. So this model does better prediction on mpg than the previous ones. The regression results are at Table 4. We see that coefficients for variables *displacement* and *acceleration* are not significant but *weight* and *horsepower* are significant at 99.9% level. For the categorical variables the model generates dummies for each category and takes as base level one of them. For example category *cylinder3* is the base. When I apply this model to the test dataset and do the *mpg* prediction the results for R^2 and $RMSE$ are 0.846 and 3.029 respectively. We conclude that this model better performs on both train and test datasets compared to the previous models.

Table 4: Regression results 2

	Estimate	Std. Error	t value	Pr($\|t\|$)
(Intercept)	31.8743252	2.7158575	11.736	<2e-16 ***
horsepower	-0.0442468	0.0145759	-3.036	0.002633 **
cylinders4	5.8787745	1.7722569	3.317	0.001033 **
cylinders5	5.3020945	3.5522134	1.493	0.136696
cylinders6	3.4399947	2.0072451	1.714	0.087707 .
cylinders8	5.1419189	2.3146057	2.222	0.027140 *
displacement	0.0136954	0.0079843	1.715	0.087430 .
weight	-0.0050466	0.0006982	-7.228	4.95e-12 ***
acceleration	-0.0241382	0.1022164	-0.236	0.813495
model_year71	1.1209701	0.9036006	1.241	0.215838
model_year72	-0.0631662	0.8900106	-0.071	0.943472
model_year73	-0.3863413	0.8041067	-0.480	0.631286
model_year74	1.1888119	0.9538691	1.246	0.213725
model_year75	0.8486691	0.9437677	0.899	0.369323
model_year76	2.0329213	0.9152216	2.221	0.027159 *
model_year77	2.7512340	0.9715318	2.832	0.004974 **
model_year78	4.0356277	0.9093815	4.438	1.32e-05 ***
model_year79	5.2639319	0.9301078	5.659	3.85e-08 ***
model_year80	8.6346558	1.0379060	8.319	4.31e-15 ***
model_year81	6.9214457	1.0119831	6.839	5.24e-11 ***
model_year82	8.1760897	0.9370712	8.725	2.71e-16 ***
origin2	2.1601973	0.6009657	3.595	0.000386 ***
origin3	2.6285952	0.5713238	4.601	6.46e-06 ***

*p<0.1; **p<0.05; ***p<0.01

f. Use backward selection to eliminate non significant predictors from your model, until only significant coefficients remain (p<0.05). How does this model compare to the previous two?

Answer: For this question I have already eliminated the variable *car_name*. Using the backward selection, now I will eliminate *accelerate* and *displacement*. The results are those of Table 5. I also removed model year, origin and cylinder and at the end i got only intercept, horsepower and weight significant. However I do prefer to keep model with model year, origin and cylinder even though some of the categories are not significant. This model has slightly smaller $R^2 = 0.878$ and slightly larger $RMSE = 2.73$ for training data and $R^2 = 0.846$ and $RMSE = 3.033$ for test dataset.

Table 5: Regression results 3

	Estimate	Std. Error	t value	Pr($\|t\|$)
(Intercept)	31.0136347	2.2485693	13.793	< 2e-16 ***
horsepower	-0.0341710	0.0112160	-3.047	0.002540 **
cylinders4	6.4663527	1.7206812	3.758	0.000209 ***
cylinders5	6.5371872	3.4697025	1.884	0.060613 .
cylinders6	4.8461970	1.8185168	2.665	0.008158 **
cylinders8	7.3598775	1.9255220	3.822	0.000164 ***
weight	-0.0046700	0.0005724	-8.158	1.24e-14 ***
model_year71	1.1279575	0.9011527	1.252	0.211753
model_year72	-0.2145045	0.8867841	-0.242	0.809046
model_year73	-0.4055364	0.8048435	-0.504	0.614758
model_year74	1.0599431	0.9479953	1.118	0.264508
model_year75	0.7743902	0.9399332	0.824	0.410726
model_year76	1.8986725	0.9113917	2.083	0.038154 *
model_year77	2.6233127	0.9683586	2.709	0.007173 **
model_year78	3.8562806	0.9029536	4.271	2.69e-05 ***
model_year79	5.1233678	0.9236559	5.547	6.85e-08 ***
model_year80	8.5933930	1.0387609	8.273	5.76e-15 ***
model_year81	6.7943446	1.0038980	6.768	7.91e-11 ***
model_year82	8.1093557	0.9350812	8.672	3.80e-16 ***
origin2	1.8420682	0.5740080	3.209	0.001490 **
origin3	2.3690553	0.5528817	4.285	2.53e-05 ***

*p<0.1; **p<0.05; ***p<0.01

Problem 8. Write a conclusion (at most 18 sentences) summarizing the most important findings of the assignment. What did we learn about the dataset?

Answer: We learned how to check for missing variables or outliers and how to deal with them. Plotting the correlation matrix for numerical variables help us see the relationships between two variables and how strong is this relation. Another way to see the relationship is using scaterplots which helps us visualize the relation of two variables. Boxplots helps us better realize the outliers and see in which ranges are most of the data. Using a random group mark we split the data into train and test sets. In order to measure the performance of the models we use measures as R^2 and $RMSE$ where the former is the coefficient of determination and can be thought of as what fraction of the y variation is explained by the model while the latter is the most common goodness of fit.