**DATA ANALYSIS HW 1**

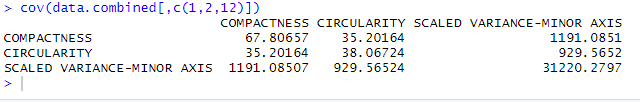
VEHICLE SILHOUETTES

**PART A:**

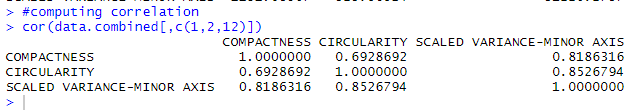
For part A, I will be evaluating the COMPACTNESS, CIRCULARITY, SCALED VARIANCE and the class variable.

1. **COVARIANCE:**

Creating a covariance matrix:



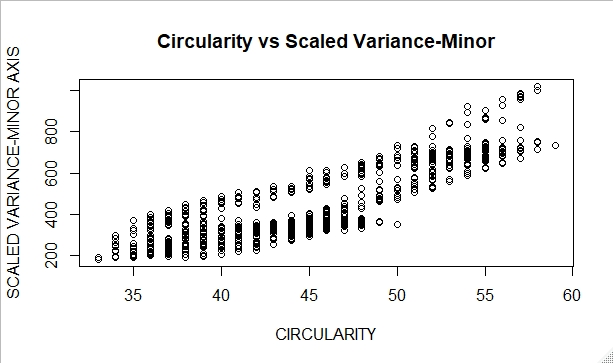
We know that the covariance tells us how changes in one variable affects changes in another. From the table we can see that the SCALED VARIANCE(MINOR) and COMPACTNESS have a high covariance. This shows that changes in one variable will highly trigger a change in the other variable. We can also note that CIRCULATORY AND SCALED VARIANCE(MINOR) also have a high covariance. From this we can see that the SCALED VARIANCE(MINOR) plays a major role in our data. Might be something we delve into later to see why this is.



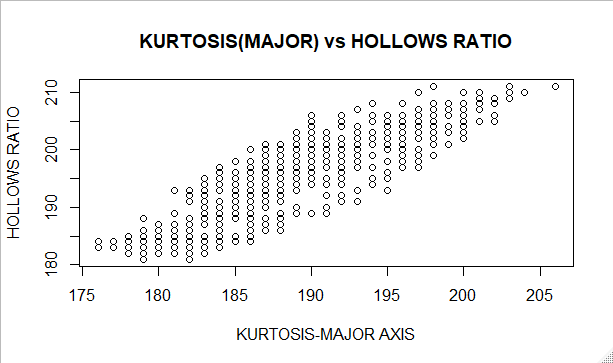
Through the correlation, we can see that the SCALED VARIANCE(MINOR) has a relative high correlation with the other two variables. This tells us that these three parameters have a high relation to one another. More importantly, that the compactness and Circularity have great affects on the Scaled Variance(minor) attribute. Something to take note of, since we know that this plays an impact on our data.

1. **SCATTER PLOT:**

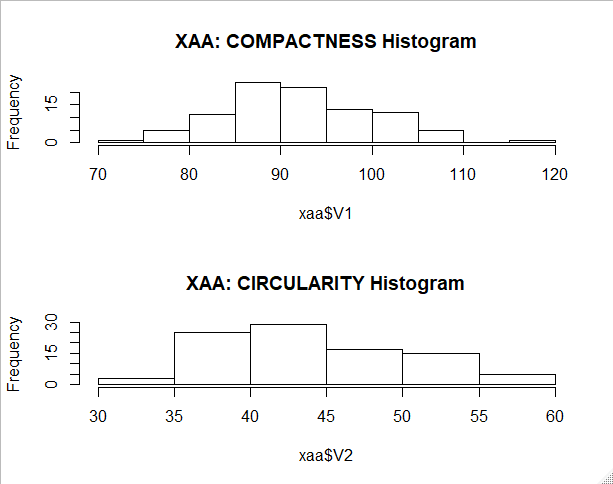
Creating Our Scatter Plot for our last two variables: CIRCULARITY AND SCALED VARIANCE(MINOR):

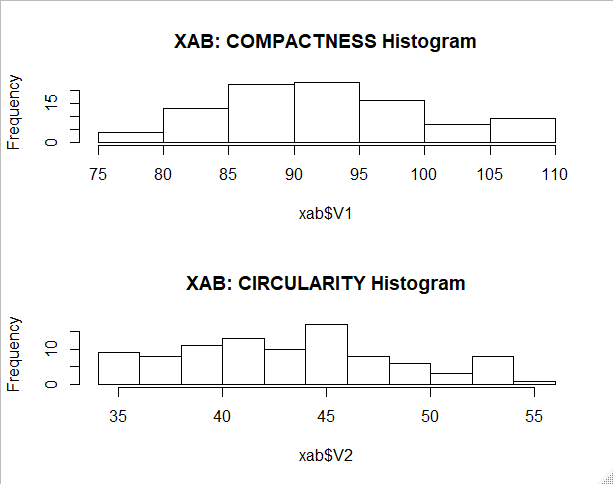


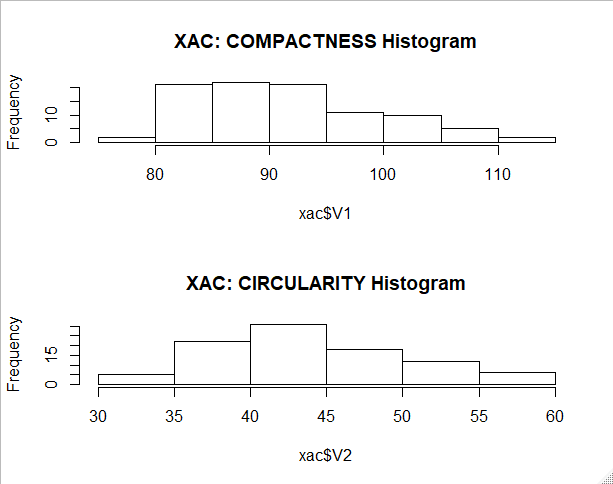
As we can see, there’s a strong correlation between the two attributes. We can see a strong trend between the two sets as many of the data point overlap in multiple areas. We also see an upward trend indicating a strong relationship between the two variables.

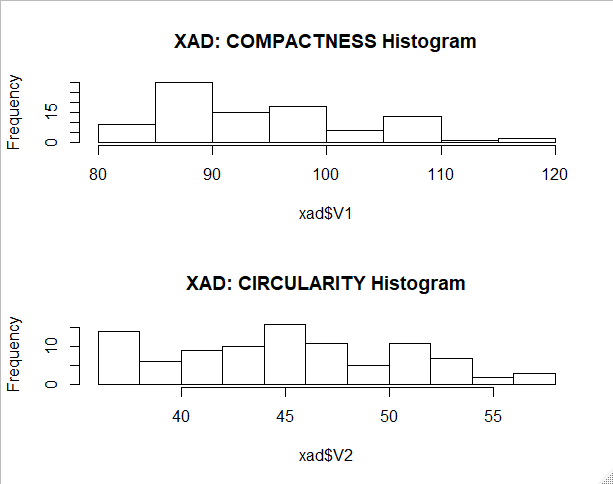


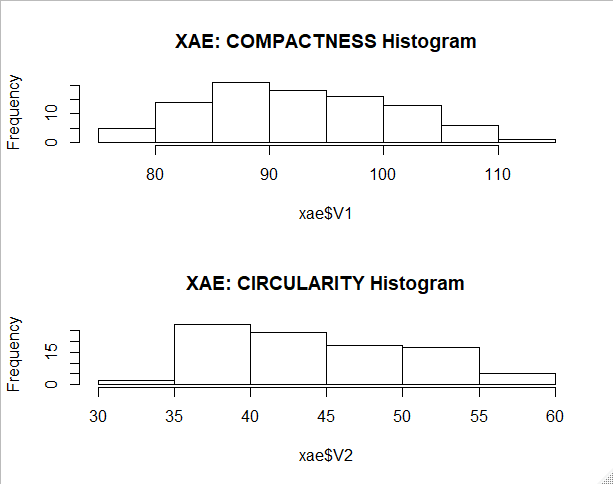
We can see a positive upward trend with our data, telling us that the variables are co-related somehow. The broadness of the data tells us that the correlation may not be as strong as we would like.

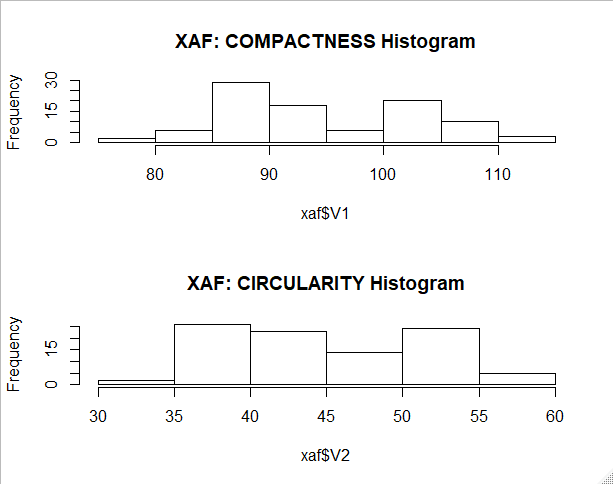
1. **HISTOGRAMS:**

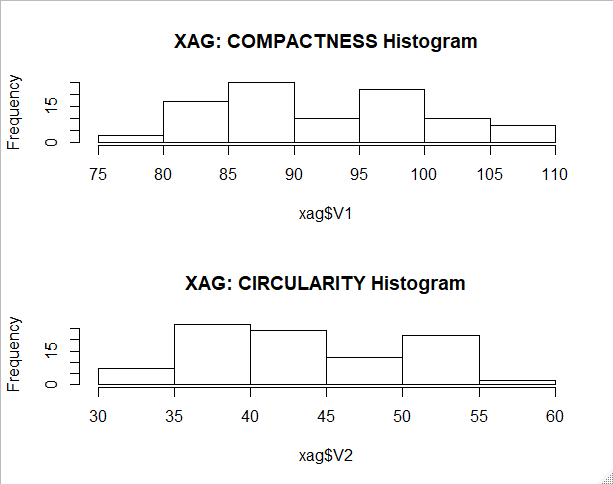


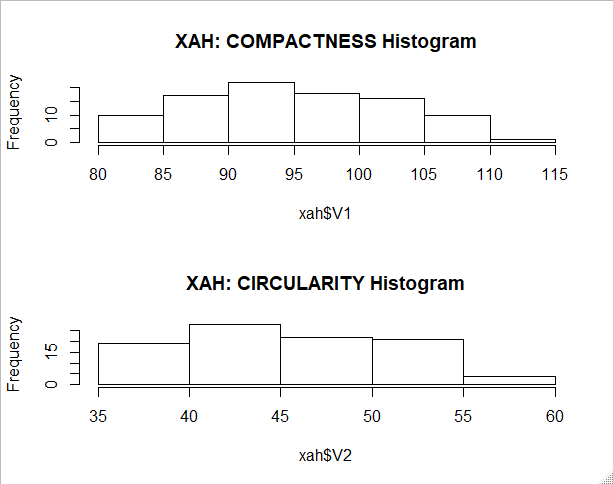


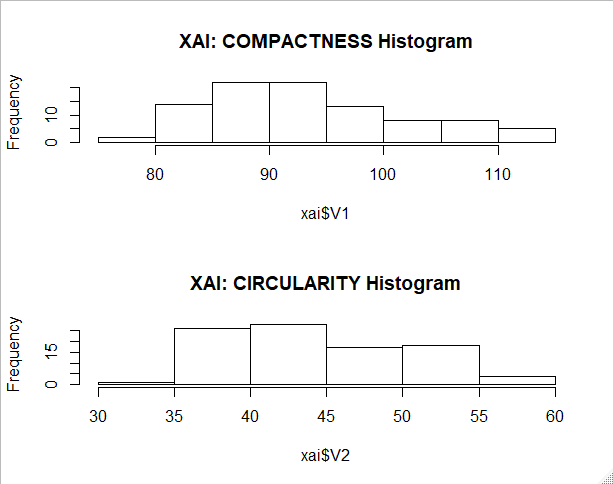


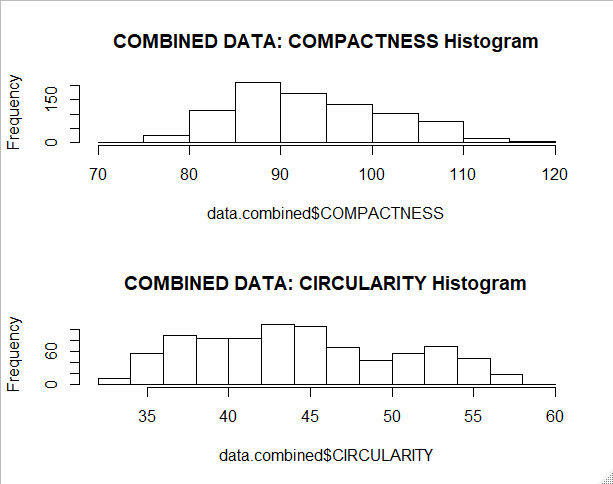




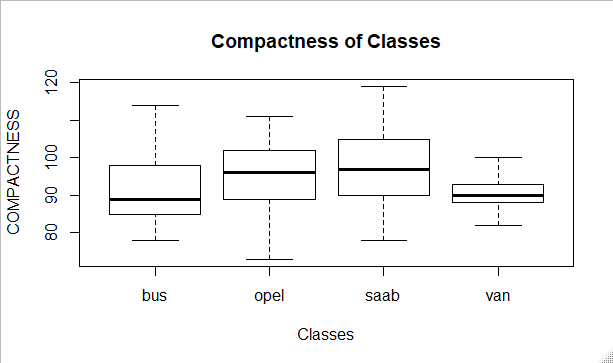
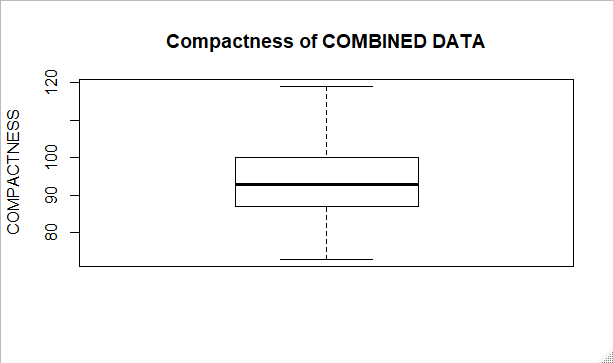
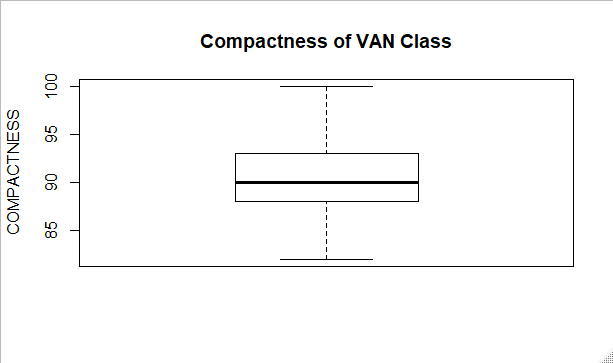
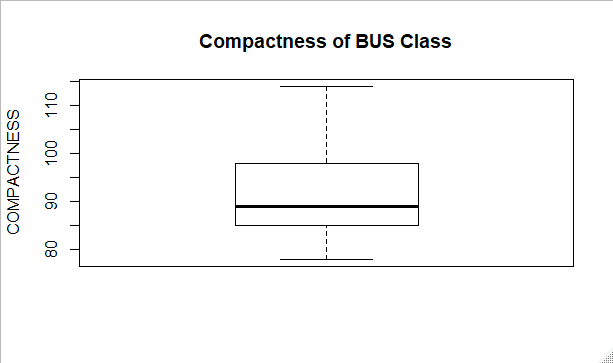
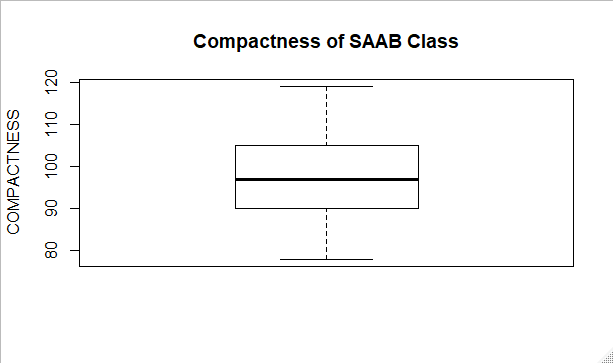
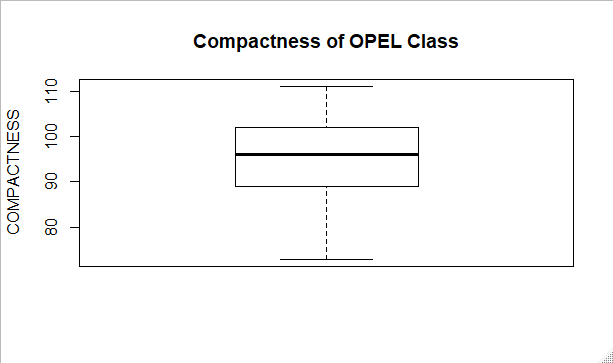






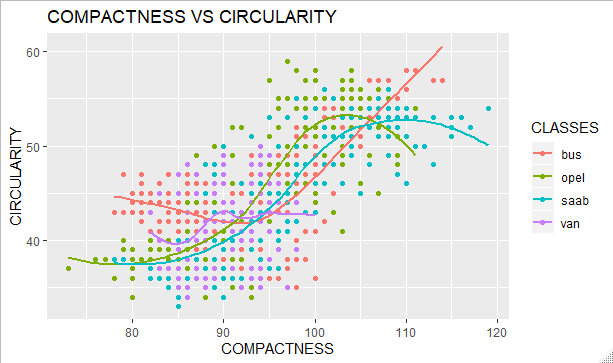


1. **BOX PLOTS:**

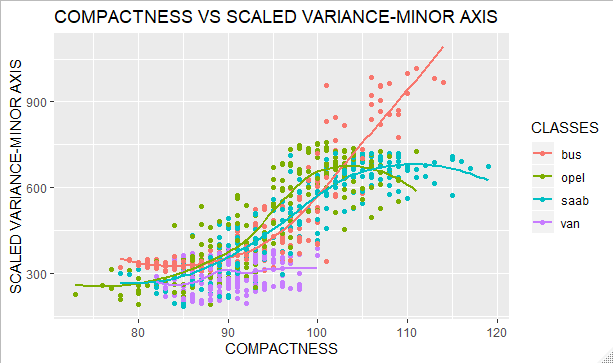


From the Box plots, we can see that the Van had the lowest range. Also the Bus’s median line is not in the ideal middle of the interquartile range. It’s interesting how the combined box plot is similar to that of the van’s; the combined box also has attributes of the other three classes such as it encompasses both the range of the union of classes and the whole dataset of all classes. The median of the combined data set is more approximate to the average of all classes, resting ideally in the middle of the inter-quartile range.

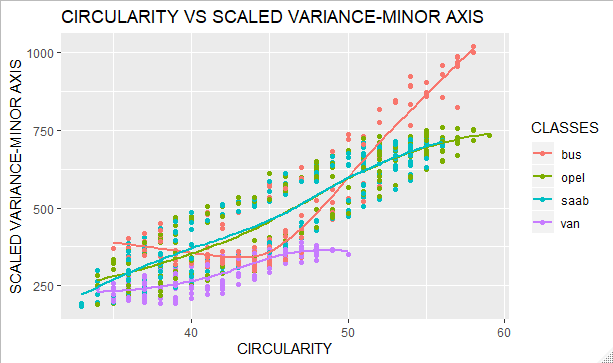
1. **SUPERVISED SCATTER PLOTS:**



From this data we can see that the lower the compactness and circularity, the more prone we are to have a vehicle in the Van class. Whereas the other classes follow a natural regression curve, meaning that we will not be able to distinguish between the classes solely based on the compactness and circularity attributes.



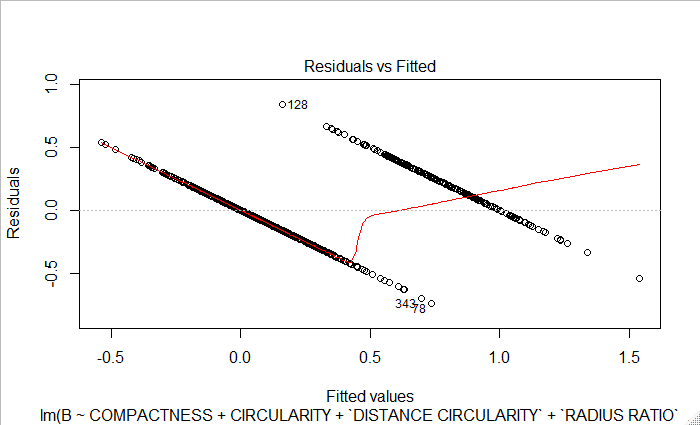
From the data, we can see that although everything is compact, we still have a definite area to find where our Van classes are clustered. We can also determine a better approximation for where our Bus classes are clustered, as we can see them either high in both attributes are very low in both. The opel and saab classes vary evenly throughout the data, so it’s hard to determine if we can be able to predict their classes solely based on this data.

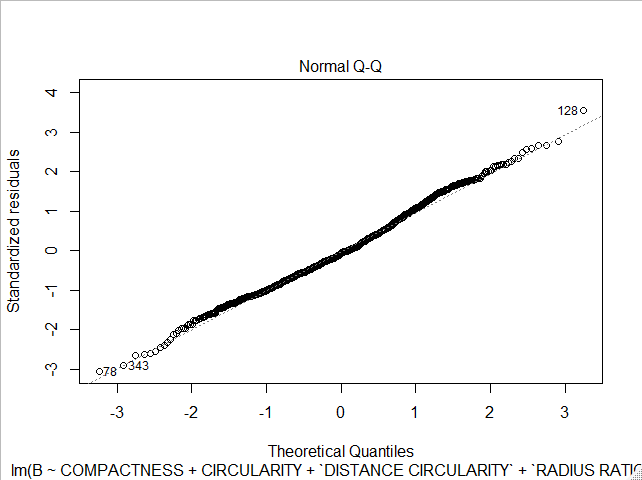


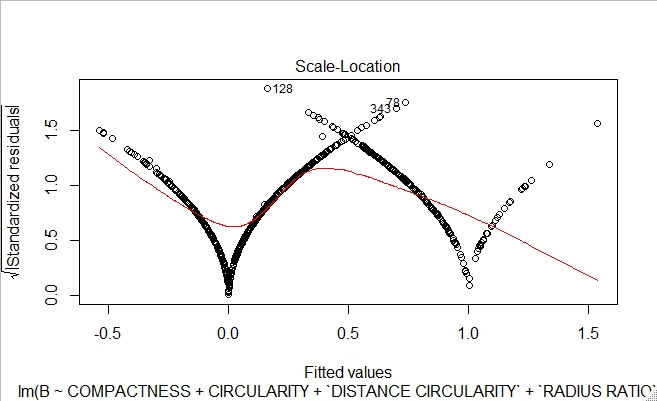
This data clearly shows that there is a big correlation between the opel and the saab classes. It also shows that majority of our data points don’t fall into these category; meaning we will not be able to predict the opel and saab classes based on strictly circularity and scaled variance. In contrast, we can to some extent be able to model a way to predict Vans and Busses based on the Circularity and Scaled Variance.

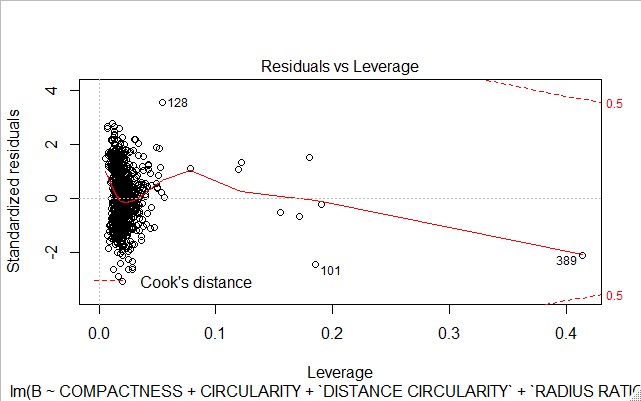
1. **LINEAR MODEL:**

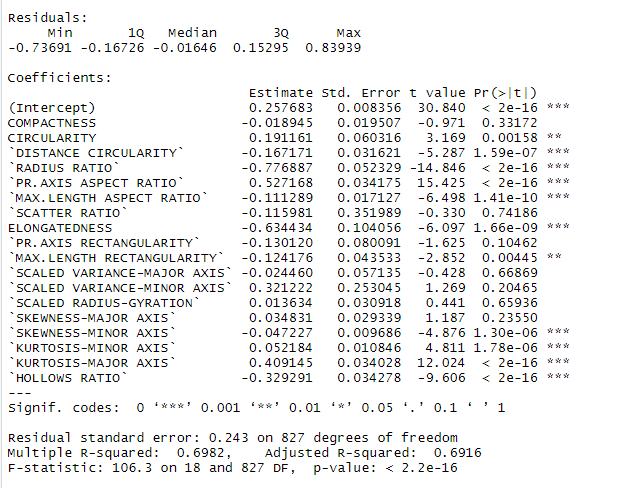
Z-SCORE MEASURES how many standard deviations from the mean.

**BUS:** 





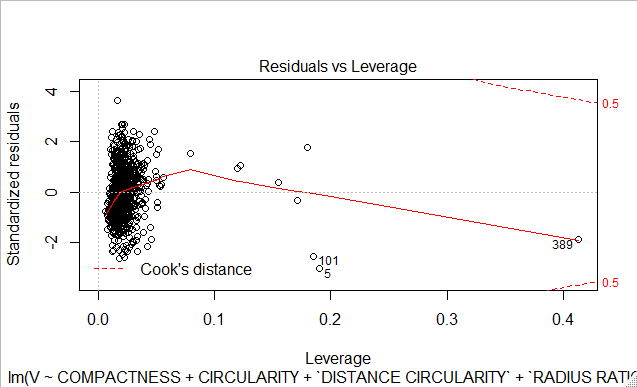


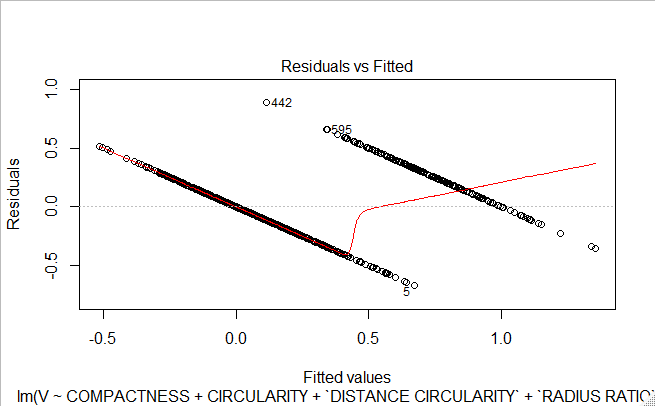


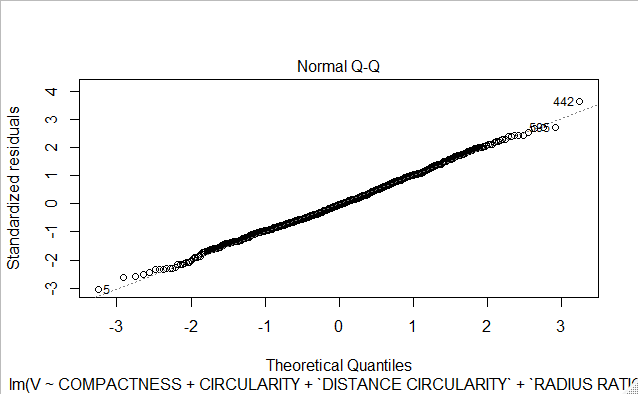
BUS-

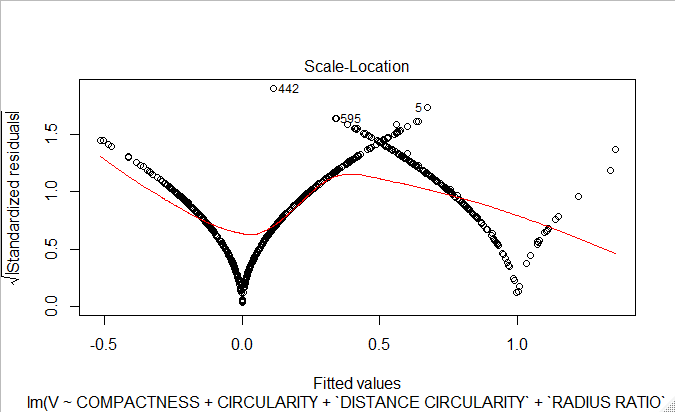
R2 tells us that about 69.82% of the model can help to determine whether a vehicle is a Bus. The coefficient of each attributes helps us understand the degree to which they affect the model. A positive value means a greater dependency on the attribute, whereas a negative has the opposite effect. can be explained through the other 18 variables.

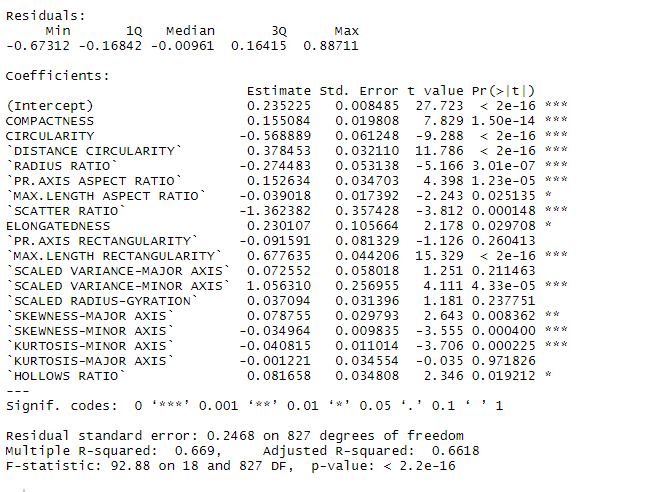
**VAN:**







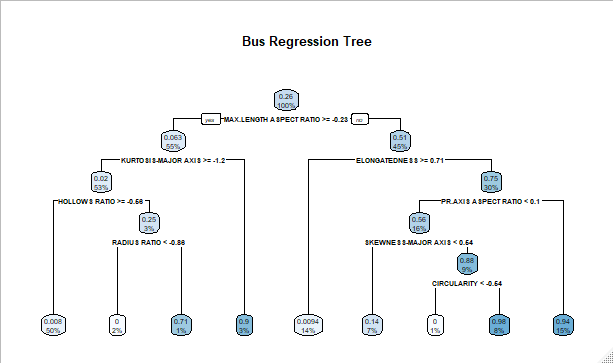


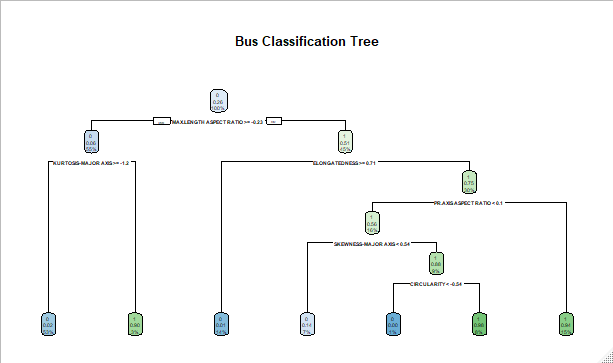


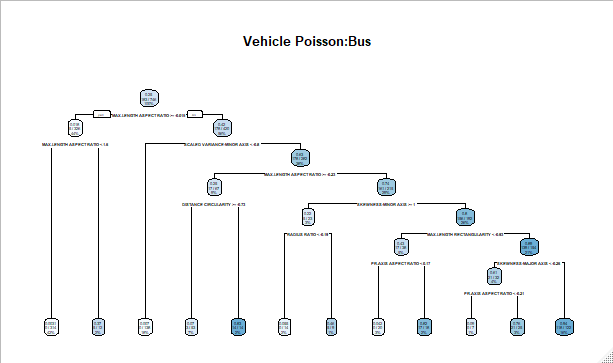
We can see that the scaled Minor Variance plays a very important role in determining whether a vehicle is a Van or not. From the Normal Q-Q graph, we can see that the data fits almost one hundred percent of the fitted regression line. From our R2 valued, we can see that the 18 attributes will help us correctly create a model that predicts if a vehicle is a Van 66.9% of the time.

1. **DECISION TREE:**

**DECISION TREE: BUSSES(B)**



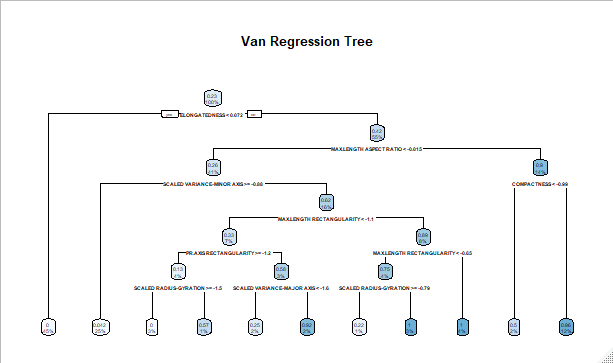


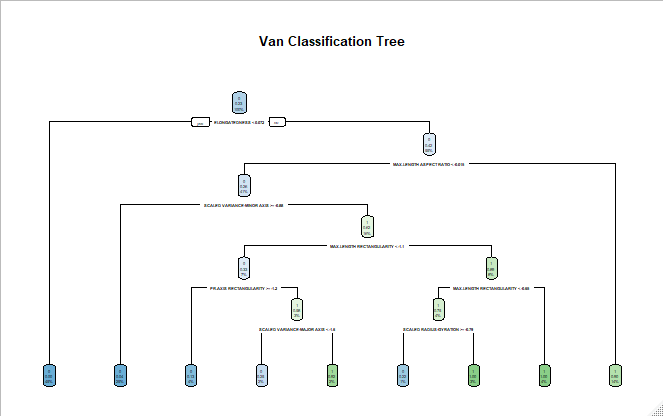


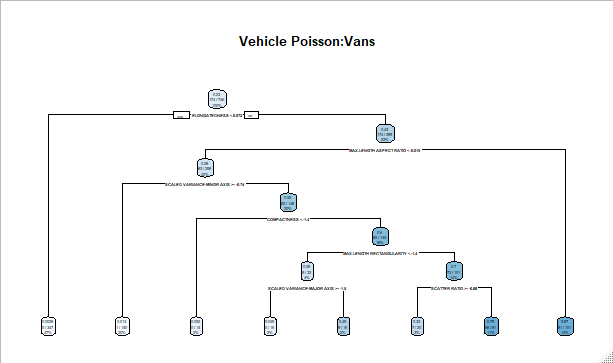
I created 3 different decision tree: classification, regression, and poisson trees. I chose these three types, mainly because these are normally the basic type of way we analyze data. The Poisson tree tells us about how likely it is for each type of event to occur. The classification helps us finds ways to classify an item. The Regression tree tells us more so how each variable affects the type of class a vehicle is placed into.

Though they all help make decision, I like the classification tree the best as it gives us the least amount of attribute needed to determine if a vehicle is a Bus or not. Furthermore, the attributes make sense as you go from a parent node to a child nodes.

**DECISION TREE: VANS(V)**

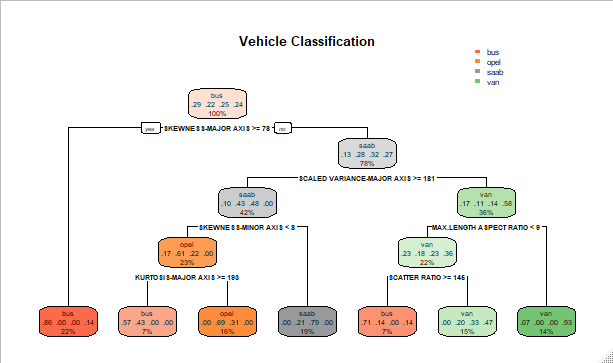






I created 3 different decision tree: classification, regression, and poisson trees. I chose these three types, mainly because these are normally the basic type of way we analyze data. The Poisson tree tells us about how likely it is for each type of event to occur. The classification helps us finds ways to classify an item. The Regression tree tells us more so how each variable affects the type of class a vehicle is placed into.

Interestingly enough, the Poisson tree is the one that give the fastest way to classify a vehicle as a Van.



I created this one just to see how the classifications work in general.

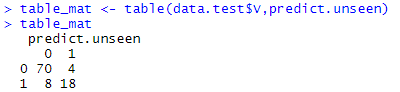
1. **CONCLUSION:**

We’ve identified about 6 major attributes that greatly determines the class of a vehicle: The Skewness (Major and Minor), Scaled Variance(Major), Max Length Aspect Ratio, Scatter Ratio, and the Kurtosis(Major). We can see that the most important attributes will be the Skewness, as having both the major and minor Skews can greatly help determine the classification of a vehicle. The tree above helps display a better overall understanding of which attributes play the greatest roles when determining the classification of a vehicle. Furthermore, from previous scatter plots we can understand why a vehicle is prone to be classify as a Bus or Van rather than a saab or opel. The attributes themselves are more geared towards predicting Vans and Buses as we can see in the Normal Q-Q linear models we conjured earlier.

1. **OBSERVATIONS:**

Even though the original data was evenly distributed within the 4 classes at about 240 vehicle per class. We can see that the actual attributes point towards measuring whether a vehicle is a bus or a van. This is troubling, since if we need to find a model to fit opel, or saab, we would need either more data or new type of data in order to configure the correct prediction model.

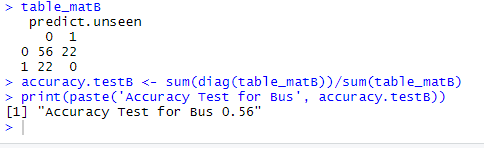
**Testing Model:** Below I have created a test and training set and see how my predictive model fairs.





From my model, we can see that it correctly classifies 70 vehicles as a van and misclassifies 4 as not. It also misclassifies 8 vehicles as vans, that weren’t vans. So about 88% of the time, my model for classification for busses will be right.

Prediction Testing for Buses



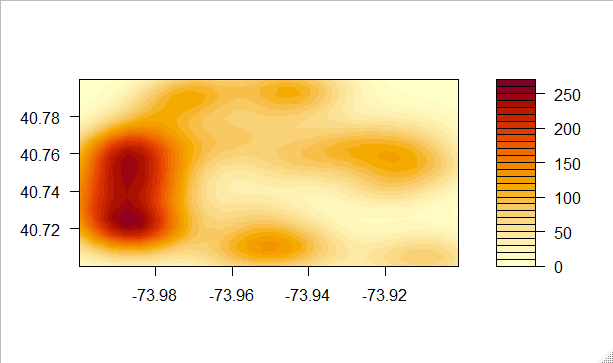
For Buses, my accuracy was only 56%. Which isn’t bad either but it would not be something I will use to predict a real life event.

1. **HEATMAP:**

#heatmap for Harassment .5

k <- with(Harassment0.5, MASS:::kde2d(Harassment0.5$Longitude, Harassment0.5$Latitude))

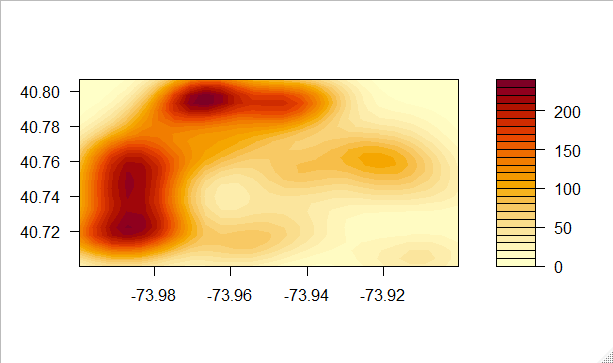
filled.contour(k)



#heatmap for Harassment 12.17

k <- with(Harassment12.17, MASS:::kde2d(Harassment12.17$Longitude, Harassment12.17$Latitude))

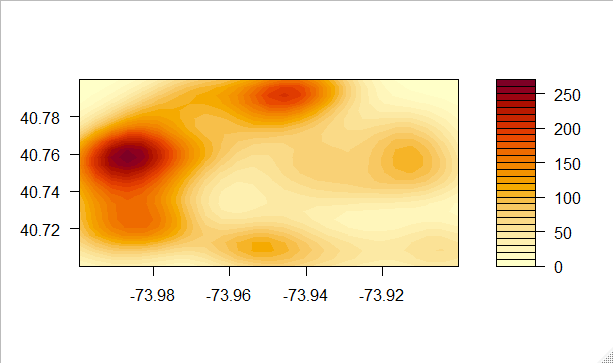
filled.contour(k)



#heatmap for Harassment 6.11

k <- with(Harassment6.11, MASS:::kde2d(Harassment6.11$Longitude, Harassment6.11$Latitude))

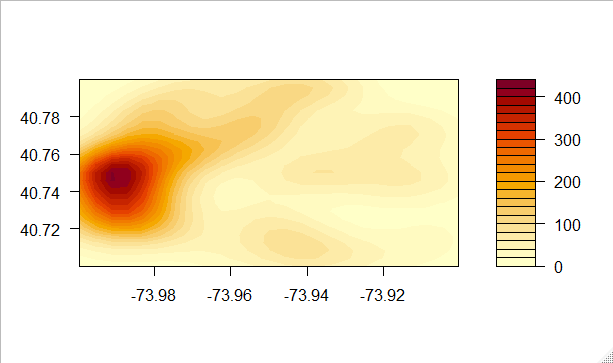
filled.contour(k)



#heatmap for PetitLarcency

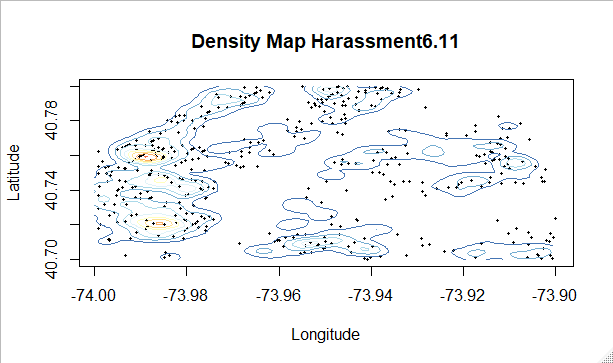
k <- with(PetitLarcency6.11, MASS:::kde2d(PetitLarcency6.11$Longitude, PetitLarcency6.11$Latitude))

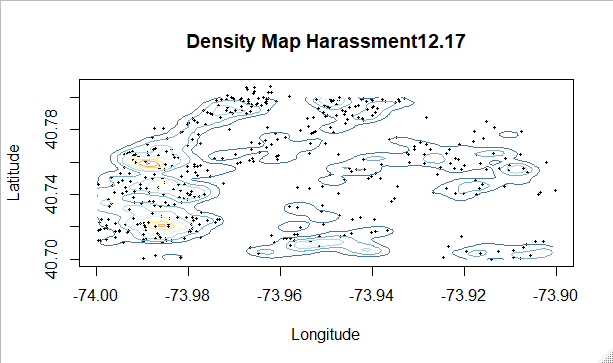
filled.contour(k)



For my heatmap, I chose a bandwidth of 100, to keep it see-able.

1. **DENSITY CONTOUR MAP:**





The two density maps look almost identical! They may have a few points in different areas, but the contours are very similar if not identical. I would reason that these are share some common features in terms of the type of crime or some attribute pertaining to crime.

1. Based on the Larcency heat map and the Harassment6.11 density map, we can see a correlation between the two maps. Area of high crimes are highlighted in both maps. We can say that Larcency is highly correlated to the time slots 6-11.