Dear Danielle,

Thank you for trusting us again for your research. In this email you will find some conclusions we can draw from data on cars and iris flowers after working on it with R:

* **Our predictions concerning how far a certain car can travel based on speed**:

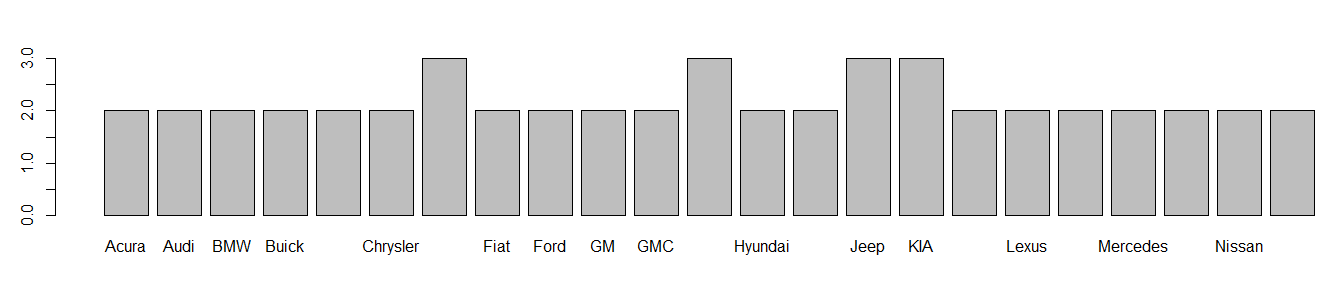
*You will find the R script used in the following link:* <https://github.com/xeniafl/cars>

**Data understanding**:

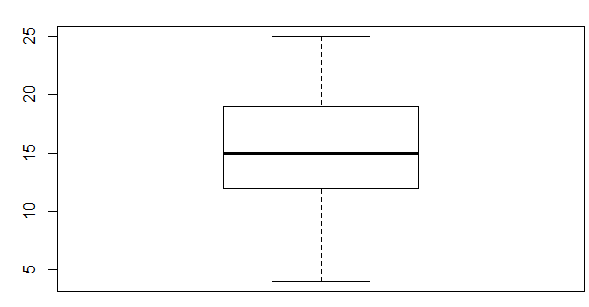
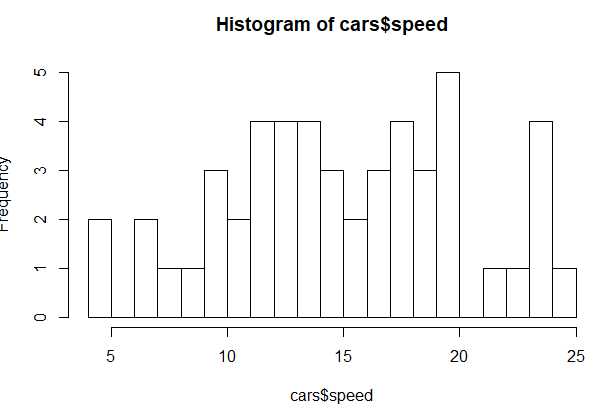
The dataset includes 3 variables: ‘name of car’ (from now on ‘name’), ‘speed of car’ (from now on ‘speed’) and ‘distance of car’ (from now on ‘distance’).

We visualized each variable to see how data is distributed, to see if it can have an impact or not on the predicted variable, distance:

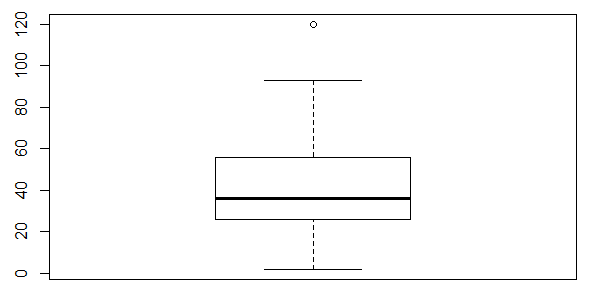
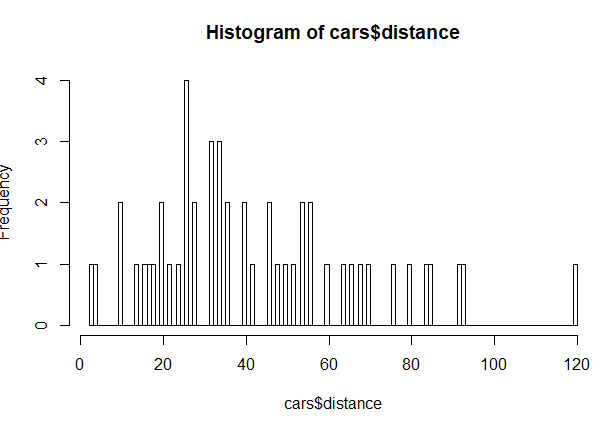
* **Name**. We can see that data is very evenly distributed with only 2 or 3 cars with the same name, so a prediction taking name into account wouldn’t be representative enough:



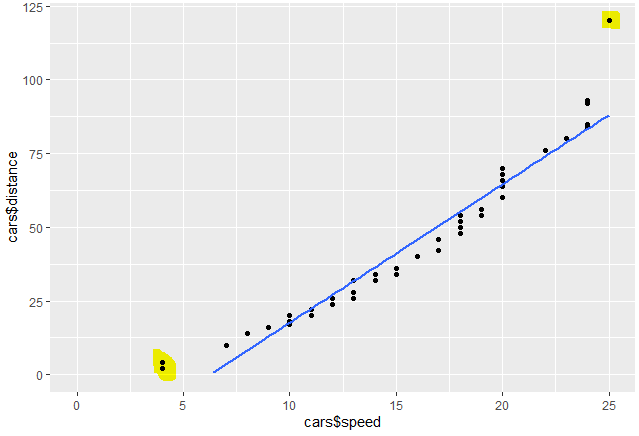
* **Speed**. Instead, we can use speed for our prediction and just as it is, as we can’t really spot any anomaly in terms of frequency and outliers:



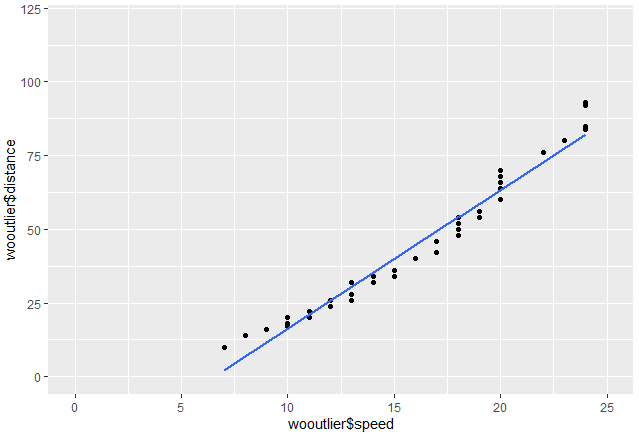
* **Distance**. Finally, we inspected distance’s data distribution and we found an important outlier. In the next steps we’ll see if we want to keep it or not:



We also visualized speed and distance together for us to spot any further interesting information. We plotted the datapoints with its linear regression to help us see better the distribution of our data:



Thanks to this scatter plot we can see that the data we have can be explained by the linear regression with a limited error when speed is between 6 and 24 mph. However, points beyond these speeds fall far away from the line. We plotted the same graph but removing the outliers, but we can see that the straight line doesn’t change much in terms of slope, intercept or error with the remaining speeds:



**Pre-processing**

After getting to know our data, we can decide that for the next steps we will:

* Disregard car name, as it brings no relevant information and use only speed for our predictions
* Outliers don’t seem to change our linear regression much, but we want to check, so for now we’ll keep two datasets: one with outliers and one without them

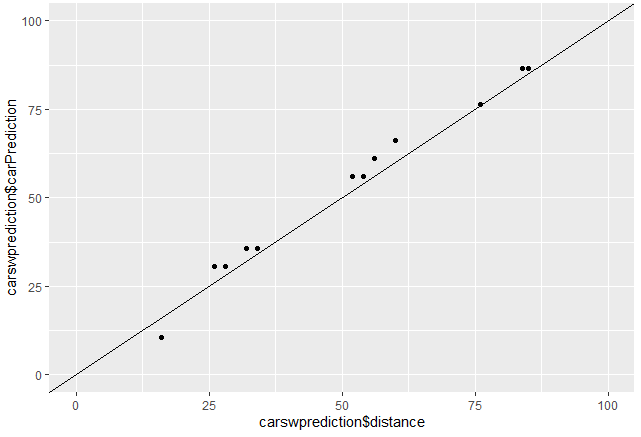
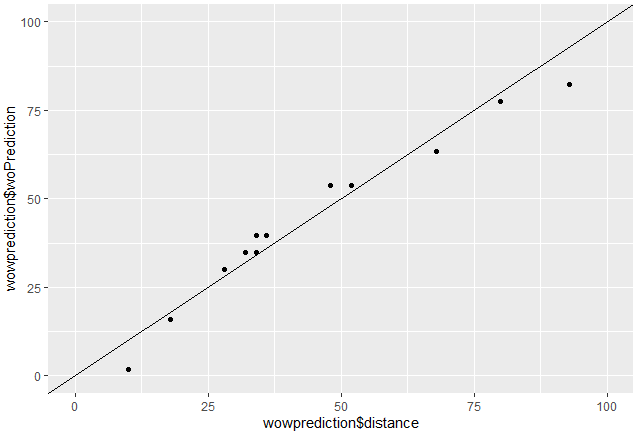
**Modelling & Evaluation**

We will start modelling two linear regressions, one with the data with the outliers and one without them:

|  |  |
| --- | --- |
| Parameters |  |
| Seed | 123 |
| Training | 70% |
| Testing | 30% |

|  |  |  |
| --- | --- | --- |
|  | With outliers | Without outliers |
| Formula |  |  |
| Residuals | Min: -9.0  Max: 28.4 | Min: -7.1  Max: 9.7 |
| Standard error | Intercept: 4.1  Slope: 0.25 | Intercept: 2.5  Slope: 0.15 |
| Multiple R squared | 0.92 | 0.97 |
| MAE | 5.2 | 4.5 |
| RMSE | 20.18 | 16.82 |

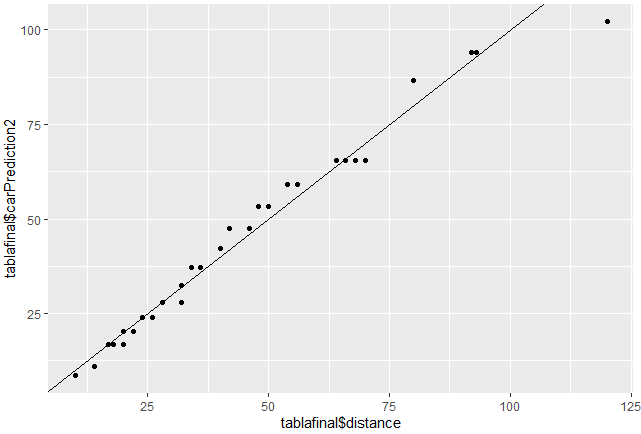
In these two graphs we plotted predicted distance depending on actual distance, and we can see that while the prediction without the outliers has errors both above and below actual distance; the prediction with the outliers will almost always overpredict because it’s trying to fit the outlier:

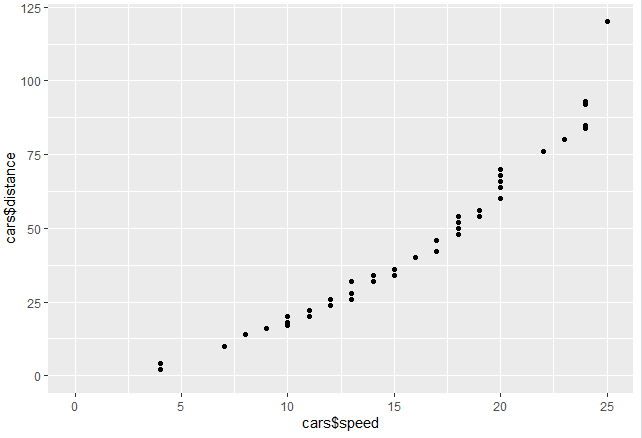
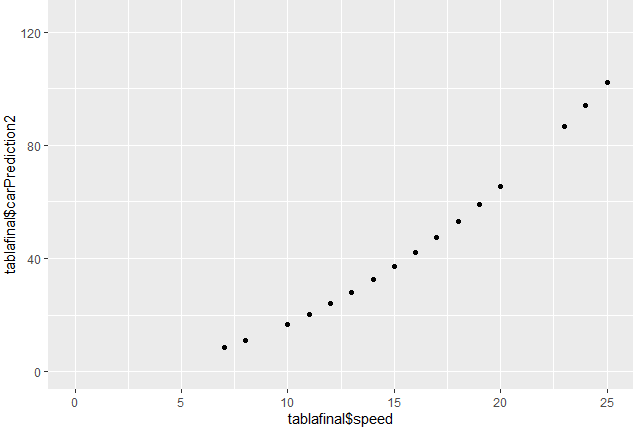
Thus, if we would choose one linear function, we would choose the model which doesn’t take into account the outliers, as it fits better the majority of the rest of the data. However, data doesn’t seem to be linear, this is why we’ll also try on the data set with outliers a quadratic function:

|  |  |  |
| --- | --- | --- |
| With outliers | **Linear function** | **Quadratic function** |
| Formula |  |  |
| Residuals | Min: -9.0  Max: 28.4 | Min: -6.5  Max: 17.9 |
| Standard error | Intercept: 4.1  Slope: 0.25 | Intercept: 1.39  Slope: 0.00 |
| Multiple R squared | 0.92 | 0.97 |
| MAE | 5.2 | 2.7 |
| RMSE | 20.18 | 15.99 |

If we plot actual and predicted distance with the new function we can see that now most of the points are very close to each other, except for the data point with speed = 24 with an error of 17.9 which is strongly penalizing the RMSE.



However, this is the model with the best performance, and therefore we’ll choose this one over the others. If we have a look at the prediction function (left) and the original data (right) we can further note how similar they are.

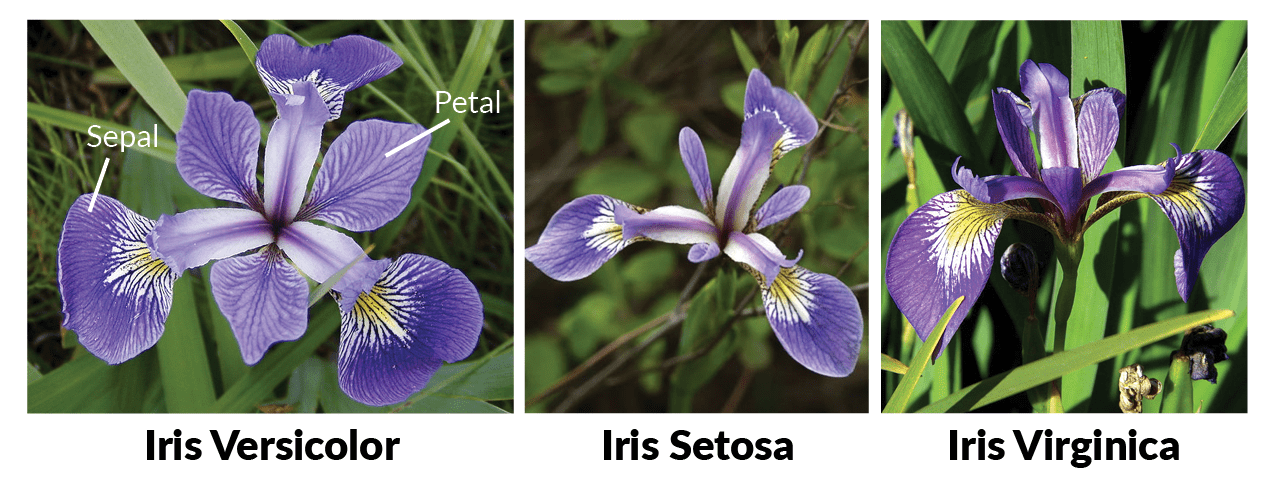


* **Our predictions concerning the petal length through using the petal’s width:**

*You will find the R script used in the following link:* <https://github.com/xeniafl/iris>

**Task understanding**

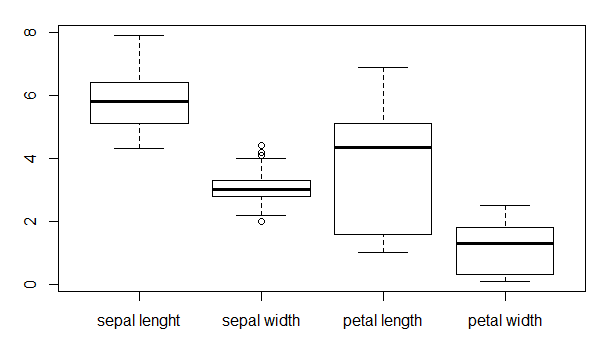
Iris data set has 5 variables, 4 of which are numerical and regard iris size, and the 5th one refers to iris species. First, we did some basic research on iris and its varieties to frame the task:



Thanks to this initial research we can see that species is a variable that could have an impact on our predictions, and that we could need to mix numeric and categoric values to get our results.

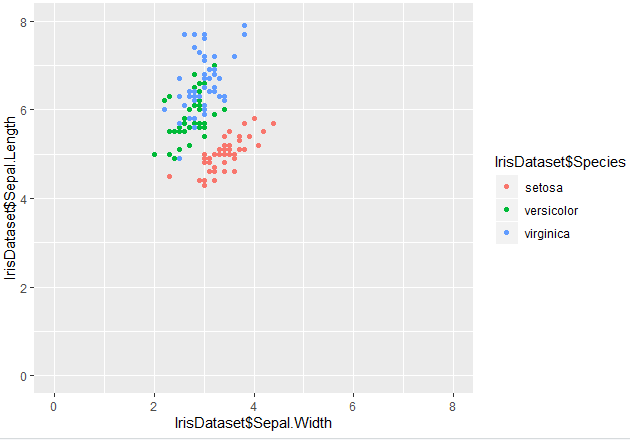
**Data understanding**

After that, we plotted a general boxplot to help us visualize how data for the 4 numerical variables was distributed:

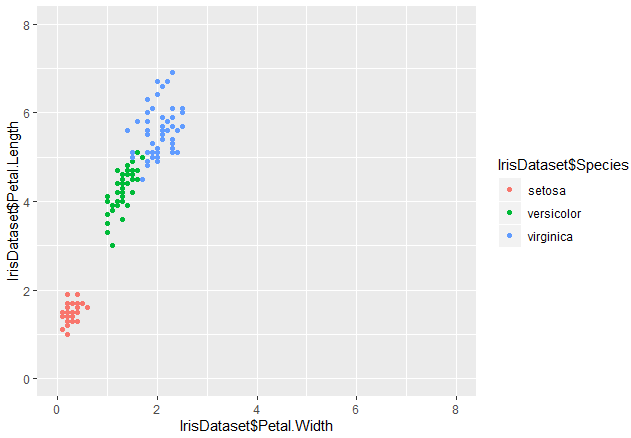


This graph confirms what we can see already see in the pictures above for understanding the task: that iris sepals are bigger than the petals, and also that petal length is very scattered as it varies across species. We can also notice that sepal width has some outliers, but we’ll make sure in the next step if they depend on species or if they are actual outliers.

To see these differences across species we plotted sepal and petal size, using different species as colours:

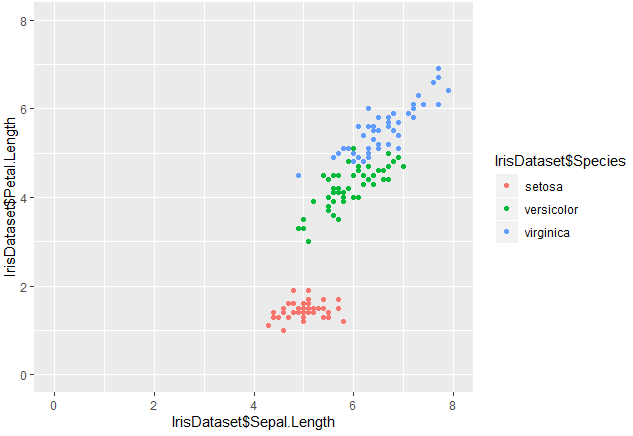
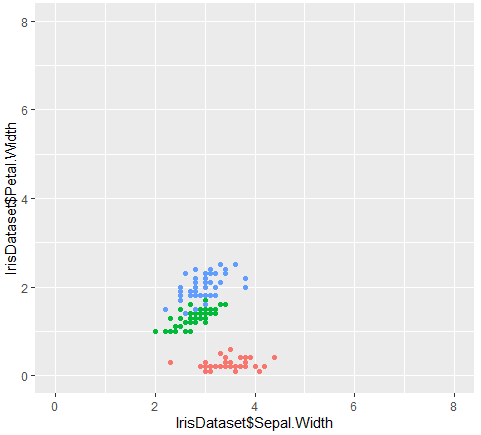


Thanks to this graph we can spot, for example, that setosa iris’ sepal should be within a clear size frame, while versicolor and viriginica are more difficult to tell apart from each other. Also, we can see that the outliers we spotted in the previous graph are not really outliers, and therefore we’ll keep them.



When it comes to petal size instead, we can better distinguish species: with setosa with the smallest petals, versicolor in a distinct second place with greater length and width, and finally viriginica with the biggest petals. This means that petal size and species would be “colinear” if those could be compared, and therefore we could not need species in the end to predict petal length and only petal width. We can also see that, although at pretty much all petal widths we have different petal lengths; the wider the petal, the more the length gets scattered, so we will surely lose precision there.

We also plotted petal and sepal length and petal and sepal width to find further correlations. However, both are very difficult to make predictions on, because on the same sepal length/width we could get pretty much any petal length/width value or even species:



**Pre-processing**

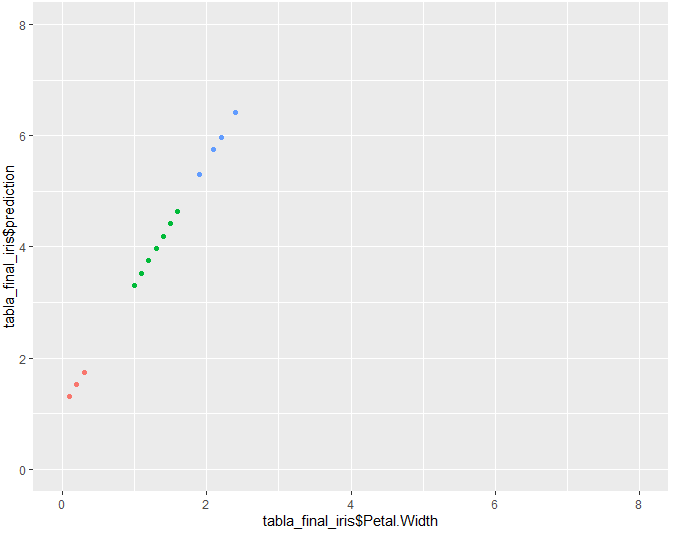
* Outliers: after seeing that the outliers are not such within iris species, we will not remove them
* Turning species into numerical values: we have seen in the previous point that species variable is closely related with petal width, therefore we’ll use this last one and disregard species

**Modelling & Evaluation**

As we have seen along the process, the simplest way to predict petal length is through linear regression in terms of petal width:

|  |  |
| --- | --- |
| Parameters |  |
| Seed | 405 |
| Training | 80% |
| Testing | 20% |

Here are the prediction and the actual data side by side:



For this model we have the following performance:

|  |  |
| --- | --- |
| Residuals | Min: -1.3, Max: 1.4 |
| Standard error | Intercept: 0.08, Slope: 0.05 |
| Multiple R squared | 0.92 |
| MAE | 0.07 |
| RMSE | 0.17 |

As anticipated in the data understanding part, the bigger the iris flower, the higher the error margin when predicting petal length, as we can see in the plot below. As this error was anticipated and overall error is low enough, we will use this model for our predictions.

