**Intro to R – Lessons Learned**

**Dataset: CARS**

**Description of the Dataset.**

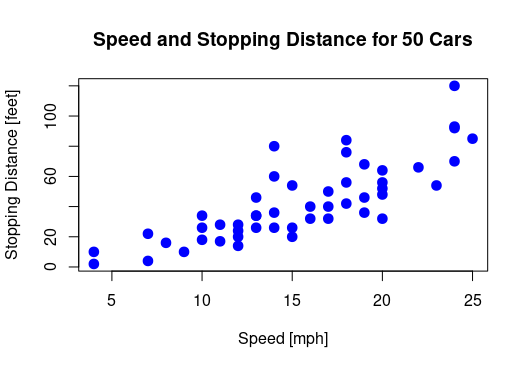
* Consists on 3 columns (‘Name of cars’, ‘Speed” and “Distance”), and 50 rows of records.
* No Missing values.
* ***‘Name of cars’,*** Nominal Categorical Variable.
  + Holds the different brands of cars.
* ***‘Speed’,*** Numerical variable
  + As I didn’t have any extra information about the dataset, I assumed that this variables represent the speed of cars, and it’s units are Mph ( a max velocity of 25kmh is quite low, but 25 mph is more acceptable, around 40kmh).
  + This graph summarizes the numerical values of this variable:

|  |  |
| --- | --- |
| **Speed**  Min. : 4.0  1st Qu.:12.0  Median :15.0  Mean :15.4  3rd Qu.:19.0  Max. :25.0  The most frequent speed of those cars are between 12 to 19 mph. And the average speed is 15.4 mph. |  |

* ***‘Distance’,*** Numerical Variable.
  + I assumed that this variable represents the stopping distance, and that the measuring unit is feet. (also those values have more sense in feet than meters because a stopping distance of 120m is quite exaggerated, but in feet it’s a safer distance of almost 37 meters)
  + This graph summarizes the numerical values of this variable.

|  |  |
| --- | --- |
| **Distance**  Min. : 2.00  1st Qu.: 26.00  Median : 36.00  Mean : 42.98  3rd Qu.: 56.00  Max. :120.00  The typical stopping distance of those cars are between 26 to 56 feet. And the average is 43 feet.  There is one car that has an exaggerated stopping distance of 120 feet. |  |

Useful information comes from plotting both variables together. There is a clear positive correlation  between them. This correlation confirms the assumption about the units and description of the variables: as faster a car goes, more distance will be needed to stop it.

****

**Prediction**

The dataset was divided in train/test set with a rate of 70/30. And then a linear regression model was applied to achieve the goal of the analysis: ***predict the car stopping distance, based on tits speed.***

* Independent Variable: Speed of the cars (speed)
* Dependent Variable: Stopping distances (distance)

And this data fits into the linear regression as: **Distance = slope\*speed + y\_intercept**

With the results, the equation of the model changes to **Distance = 3.582\*speed – 13.082**

**Conclusion**

A detailed output could be found in the output of the script. To summarize, I will focus on the result of the R2, as this statistic provides a measure of how well the model fits the  data. Those values lie between 0 and 1 (where 0 represents a regression that does not explain the relation of the variables at all, and 1 corresponds to a full and perfect correlation). In our case we got a **Multiple R-squared of 0.603.** Our rough 60% of the changes in the response variable (Stopping Distance) could be explained by the predictor variable (Speed). And it also makes sense, there are a lot of metrics that could be chosen to define the stopping distance: type of brakes, type of tires, wet/dry surface, driver’s attention level, etc. But in this case I discovered that, based on the info provided, just using a single independent variable, with a simple linear model, I was able to explain quite well the changes in the dependent variable.

**Dataset: IRIS**

**Description of the Dataset:**

* The data is about dimensions of some specific characteristics of 3 different species of flowers.
* It consists of 6 columns: row\_index, sepal length, sepal width, petal length, petal width, species. And 150 rows. There are 3 species of flowers (setosa, versicolor, and virginica) and there are 50 rows of information for each flower.
* There are no missing values.

**Finding/Solving Errors**:

This step was quite straight forward. It didn’t took me much time to solve it. Most of the errors were spelling, calling unknown variables, repeated code. I copy-pasted the original code into RStudio, and step by step I solved all the error and warnings.

**Analysis Goal**

After a fast overview I noticed that this simple dataset allows a lot of interesting approaches. But, to keep as close as possible to the analysis goal, I defined and just performed a simple study on the variables of interest. The analysis goal is: ***‘predict a petal's length using the petal’s width’.***

* Independent Variable: petal’s width
* Dependent Variable: petal's length

**Variables**

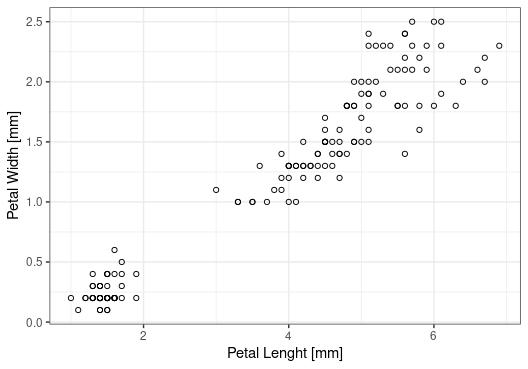
* **Petal’s Length (Petal.Length)** 
  + Floating numerical variables, expressed in mm.

|  |  |  |
| --- | --- | --- |
| Min. : 1.000  1st Qu.: 1.600  Median : 4.350  Mean : 3.758  3rd Qu.: 5.100  Max. : 6.900  The typical length of petals of those 3 species of flowers lies between 1.6 and 5.1mm. With an average of 3.76mm. |  |  |

* **Petal’s Width (Petal.Width)** 
  + Floating numerical variables, expressed in mm.

|  |  |  |
| --- | --- | --- |
| Min. : 0.100  1st Qu.: 0.300  Median : 1.300  Mean : 1.199  3rd Qu.: 1.800  Max. : 2.500    The typical width of petals of those 3 species of flowers lies between 0.3 and 1.8mm. With an average of 1.2mm. |  |  |

When you plot a variable versus another, you realize that there is a very positive correlation between them.



**Prediction**

In this case the Train/Test ratio was 80/20 of the data and the linear regression model was used.

The linear equation used to fit the data was **P.Width = slope \* P.Lenght + y\_intercept**

With the results, the equation of the model changed to **P.Width = 0.393 \* P.Lenght – 0.28**

**Conclusion**

The R2 of this experiment is a beautiful 0.9513. Which means that the length variation explains almost perfectly the width variation. Another desirable aspect is the residual symmetry between Min/Max, 1Q/3Q and a median as close as possible to zero.

Min 1Q Median 3Q Max

-0.36964 -0.10766 0.00591 0.08338 0.47607

I don’t have enough knowledge about biology to explain why this happens, and to make assumptions about this fact. But those virtually perfect characteristics are something that every data analyst looks for when fitting data into a model.

**Lessons Learned**

* Learning a new (programming) language allows you to discover new layers of the same things. Food is ‘quite’ the same all around the world. The different cultures are the ones that combine them in so many great and diverse ways. Using Python, R or a spreadsheet to analyze the same dataset will give “almost” the same output, but each one will let you see and learn new things.
* Something interesting I noticed about R is that it’s a more “academic” language than Python. At least the resources and documentation had this style underlying. I’m concluding this based on just the sites I visited to solve those simple problems. Perhaps my assumption is biased, but I have been reading tutorials and documentation about python since some months ago, and I found some differences.
* Installing R and RStudio in Ubuntu 20.04 was straightforward. (MySQL Workbench was a nightmare).
* The dataset came in a .csv format, and to load it into Rstudio the library ‘readr’ was used. I learned how to install a library and how to instantiate it.
* Something great of RStudio is the possibility of coding in the console at the same time you write your script, something like debugging in real time.
* I loved coding plots with ggplot2. Adding layers to an existing plot is a very intuitive and natural thing.
* I share some links used to solve this task:
  + <https://statquest.org/video-index/> I consult this you tube channel almost everyday. Hard topics explained in a simple and gentile way. Richard Feynman would be proud of him.
  + <https://www.youtube.com/watch?v=u1cc1r_Y7M0> this video is specific about Linear Regression in R.
  + <https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R> additional information of interpreting the results of the linear regression.
  + <https://feliperego.github.io/blog/2015/03/11/Simple-Linear-Regression-Example-in-R> just an example of using R to apply a linear regression.
  + <https://ggplot2.tidyverse.org/>
  + <https://ademos.people.uic.edu/Chapter10.html> a site with good and clear info about ggplot and R