La Vuelta

Laura Cebollero Ruiz, Alexandre Rodríguez Garau

4th January, 2019

In this project we intend to predict the duration of the different stages of the cycling race *La Vuelta* using the information provided in the file `Vuelta0.mtp`.

```
library("readxl")
library("knitr")
library("ggplot2")
```

We have exported the data to an extension that R can read: `.csv`.

```
data<- read.csv("Vuelta01.csv", sep = ';', header = TRUE, dec=",")
```

To predict the length of the stages we will use a set that contains 14 explanatory variables and a response variable called **ForecastedTime**. Let's take a look at the summary of the variables:

```
kable(summary(data[,1:6]))
```

Time	Distance	HeightIncr	AccumIncr	portsE	ports1
Min. :171.6	Min. :111.0	Min. :-940.0	Min. : 190	Min. :0.0000	Min. :0.000
1st Qu.:258.1	1st Qu.:167.2	1st Qu.:-180.0	1st Qu.: 510	1st Qu.:0.0000	1st Qu.:0.000
Median :301.4	Median :196.0	Median : 70.0	Median :1350	Median :0.0000	Median :0.000
Mean :302.6	Mean :193.0	Mean : 227.9	Mean :1477	Mean :0.1524	Mean :0.581
3rd Qu.:345.8	3rd Qu.:219.5	3rd Qu.: 540.0	3rd Qu.:2300	3rd Qu.:0.0000	3rd Qu.:1.000
Max. :439.7	Max. :264.0	Max. :2310.0	Max. :4216	Max. :2.0000	Max. :3.000

```
kable(summary(data[,6:11]))
```

ports1	ports2	ports3	year	week	bef_mount
Min. :0.000	Min. :0.0000	Min. :0.0000	Min. :1.000	Min. :1.000	Min. :0.0000
1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:0.0000
Median :0.000	Median :0.0000	Median :1.0000	Median :4.000	Median :2.000	Median :0.0000
Mean :0.581	Mean :0.4857	Mean :0.8476	Mean :3.619	Mean :2.048	Mean :0.2952
3rd Qu.:1.000	3rd Qu.:1.0000	3rd Qu.:2.0000	3rd Qu.:5.000	3rd Qu.:3.000	3rd Qu.:1.0000
Max. :3.000	Max. :3.0000	Max. :4.0000	Max. :6.000	Max. :3.000	Max. :1.0000

```
kable(summary(data[,12:16]))
```

aft_mount	bef_tt	aft_tt	last	ForecastedTime
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. :180.0
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:263.4
Median :0.0000	Median :0.0000	Median :0.0000	Median :0.00000	Median :307.8
Mean :0.2952	Mean :0.1333	Mean :0.1905	Mean :0.04762	Mean :306.7
3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:348.2
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.00000	Max. :437.5

In the previous tables we can see a short summary of the variables. **ports1**, **ports2** and **ports3** indicate the number of mountain sections in a given stage and their category (1,2 or 3). **year** corresponds to the year in which this data was recorded (1 to 6). **Week** is the week of the race in wich the stage takes place (1 to 3). The variables **bef_mount** and **aft_mount** tell us whether a stage took place before or after a mountain stage, respectively. Similarly, the variables **bef_tt** and **aft_tt** indicate if a stage took place before or after a time trial stage. Finally, the variable **last** tells us if that stage was the last of the whole race.

By the looks of the summary we can't seem to find any outliers or abnormal values. Most of the explanatory variables range from 0 to very low values and are natural numbers. The only continuous variables are **ForecastedTime**, **Distance**, **HeightInc** and **AccumIncr**. This makes sense because the first variable indicates time and time is a continuous variable and the rest indicate distance which can also be continous.

The variable **Distance** seems to be quite balanced with a mean of 193 and min and max values of 111 and 264 respectively. **HeightIncr**, however, is the only variable that presents negative values and has very high variance. Its minimum value is -940 and the maximum value is 2310.

By taking a closer look at the data we detect some abnormal values in the **last** variable: Since this variable indicates if a stage was the last of the whole race, then there should as many stages with this value equal to 1 as years of data have been recorded. Since the recorded stages are from 6 different years then there should be 6 rows, but instead there are only 5, meaning that there is at least 1 missing stage.

```
data[data$last == 1,]
```

```
##         Time Distance HeightIncr AccumIncr portsE ports1 ports2 ports3
## 20  233.4500    157.6          0       300      0      0      0      0
## 38  254.9667    171.2        100       410      0      0      0      0
## 55  266.9167    165.7       -390       780      0      0      0      1
## 90  260.9500    175.0       -410       560      0      0      1      0
## 105 285.0833    169.6       -230       300      0      0      0      1
##     year week bef_mount aft_mount bef_tt aft_tt last ForecastedTime
## 20     6    3         0         0      0      1    1       242.4615
## 38     5    3         0         0      0      1    1       263.3846
## 55     4    3         0         0      0      1    1       254.9231
## 90     2    3         0         1      0      0    1       276.3158
## 105    1    3         0         1      0      0    1       267.7895
```

If we look at the data we can easily see that the missing value belongs to the year 3. However, this will probably not greatly affect our predicting power. There is something to be said about the variables **year**, **week**, **bef_mount**, **aft_mount**, **bef_tt**, **aft_tt** and **last**: all of these variables are categorical and indicate the group or category a row belongs to. For this, we should transform these variables into factors.

```
data$year      <- as.factor(data$year)
data$week      <- as.factor(data$week)
data$bef_mount <- as.factor(data$bef_mount)
data$aft_mount <- as.factor(data$aft_mount)
data$aft_tt    <- as.factor(data$aft_tt)
data$bef_tt    <- as.factor(data$bef_tt)
data$last      <- as.factor(data$last)
```

Now that we have transformed the categorical columns to factors we should take another look at the summary of these variables:

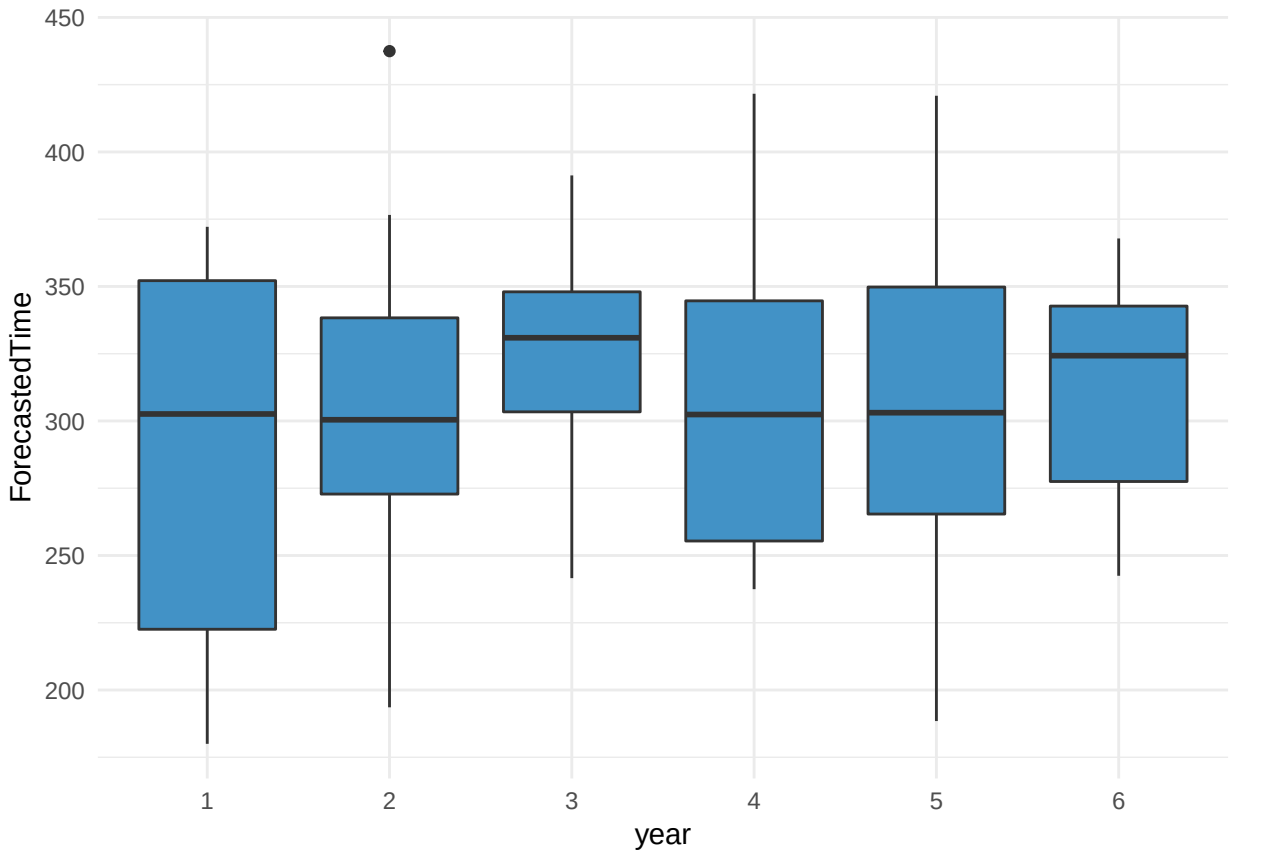
```
kable(summary(data[,9:15]))
```

year	week	bef_mount	aft_mount	bef_tt	aft_tt	last
1:15	1:33	0:74	0:74	0:91	0:85	0:100
2:18	2:34	1:31	1:31	1:14	1:20	1: 5
3:17	3:38	NA	NA	NA	NA	NA
4:17	NA	NA	NA	NA	NA	NA
5:18	NA	NA	NA	NA	NA	NA
6:20	NA	NA	NA	NA	NA	NA

We can se that for the variables **year** and **week** the variables are quite balanced, each group contains a very similar amount of observations. The rest, however, are more unbalanced.

Now, in order to find outliers and to see how the data behaves we will plot some boxplots. First of all we would like to see if the forecasted times are different depending on the year in which the race took place.

```
ggplot(data = data) +
  aes(x = year, y = ForecastedTime) +
  geom_boxplot(fill = "#4292c6") +
  theme_minimal()
```



At first glance we see that the medians are really similar among all years, excepting year 3 and year 6, which are about 25 units higher. The interquartile ranges for every year are different. Even though the Q3 percentile is very similar for all years, the Q1 percentile varies from 225 to 300. It is important to say that there appears

to be an outlier in the second year.

#Fer més boxplots??