Final Exam

The file `nba_logreg.csv` contains the statistics that `1329` players got during their first year in the NBA. The goal of this review is to develop a system that, in the future, will discover those players who will play more than 5 years in the NBA.

Important:

- As some methods may be time consuming, make a logical selection of the possible parameter values.
- Exercises should be solved using only the R packages that have been seen either in master class or in a small group. The use of any other package will have a penalty.

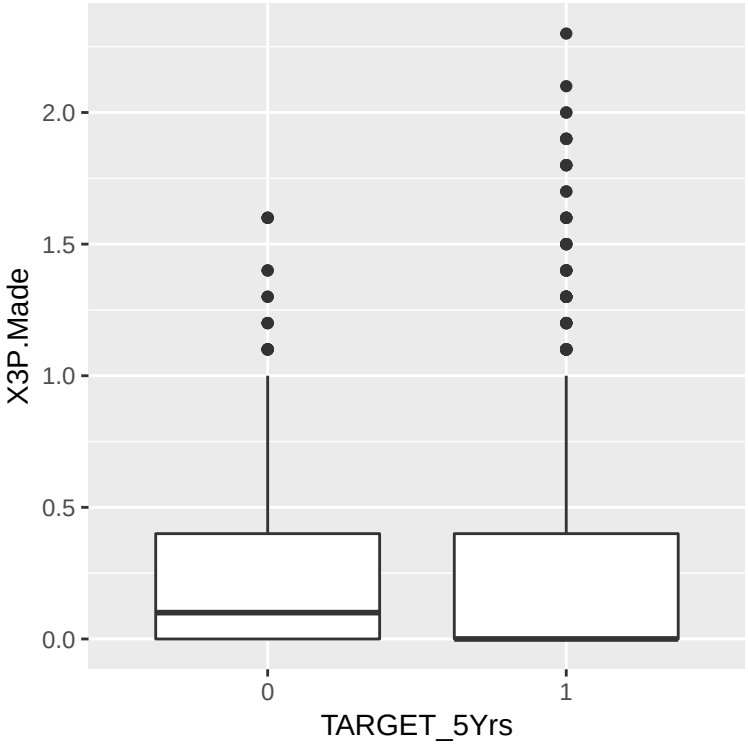
The file contains the following variables:

`Name`: Name of the player	`FTM`: Free Throw Made
`GP`: Games Played	`FTA`: Free Throw Attempts
`MIN`: Minutes Played	`FT%`: Free Throw Percent
`PTS`: Points Per Game	`OREB`: Offensive Rebounds
`FGM`: Field Goals Made	`DREB`: Defensive Rebounds
`FGA`: Field Goal Attempts	`REB`: Rebounds
`FG%`: Field Goal Percent	`AST`: Assists
`3P Made`: 3 Points Made	`STL`: Steals
`3PA`: 3 Points Attempts	`BLK`: Blocks
`3P%`: 3 Point Attempts	`TOV`: Turnovers

`TARGET_5Yrs` (dependent variable): 1 if career >= 5 years, 0 otherwise

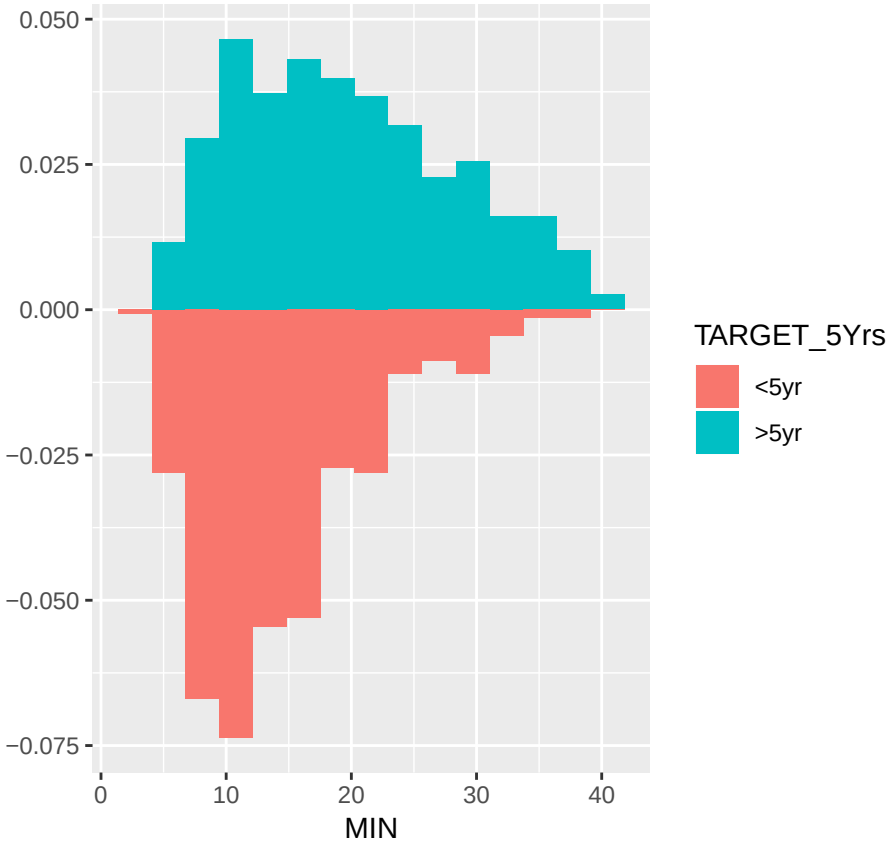
Question 1

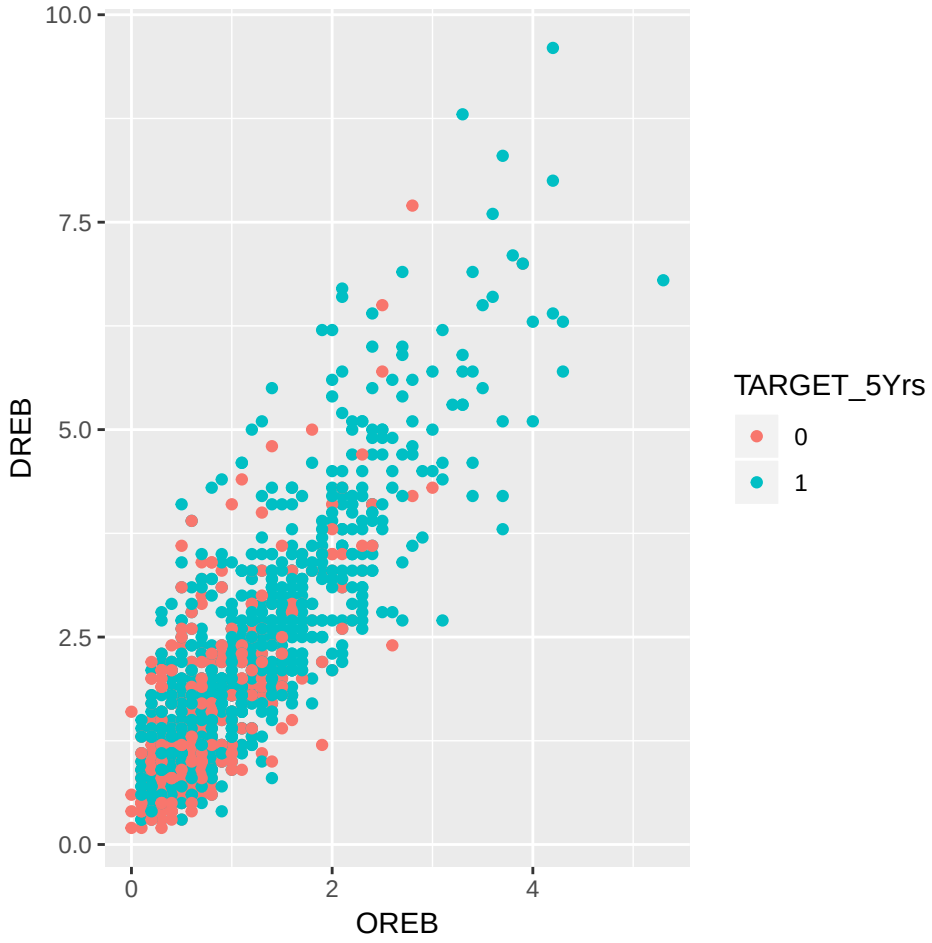
Write a code in R, using the `ggplot` package that produces the following displays. (1.5 points)



Histograms

Minutes played





Question 2

Describe the Principal Component Analysis technique, providing as much detail as you can and emphasizing one of its main applications (i.e., an example in which it is often used). (1 point)

Question 3

a. Build a training data set, named `data_tr`, that contains the first `1000` observations and a test set, `data_tst`, that contains the remaining `329` observations. Hint: Remember that variable `TARGET_5Yrs` must be a factor. (0 points)

b. Implement a decision tree, using `data_tr`, that maximizes the correct classification rate (`ccr`) of the set `data_tst`. Choose the parameters that maximize this `ccr`. (1 point)

c. Visualize the decision tree that you have built. (0.5 points)

d. Implement a random forest, using `data_tr`, that maximizes the `ccr` of the set `data_tst`. Choose the parameters that maximize this `ccr`. (1 point)

e. Implement a k-nearest neighbors, using `data_tr`, that maximizes the `ccr` of the set `data_tst`. Choose the parameters that maximize this `ccr`. (1 point)

f. Implement a support vector machine, using `data_tr`, that maximizes the `ccr` of the set `data_tst`. Choose the parameters that maximize this `ccr`. (1 point)

Question 4

a. Explain in detail the k-means algorithm. That is, what problems does it tackle, how does it work, what parameters have to be supplied in order to obtain an optimal solution. (1 point)

b. Write a script in R to perform color segmentation using k-means on the image `elektra.jpg` and visualize the result, as we did in Lab.4 with the Altamira picture. Explain why you have selected that number of centers. (1 point)

Question 5

Replace the following **for** loop to execute it using parallel programming to speed it up. Note: you cannot remove the `stupid_function` from the loop. (1 point)

```
if (!require(tictoc))
  install.packages("tictoc")
library(tictoc)

stupid_function=function(){
 Sys.sleep(0.5)
}

tic()
a=0
for (i in 1:21){
  a[i]=sqrt(i)
  stupid_function()
}
toc()
```