REGRESSION TOYOTA CAR PRICE

BUSINESS OBJECTIVE

The business goal is to determine the Toyota Corolla price by using characteristics that describe this model.

DATA SCIENCE TASK

We will perform a linear regression to predict the price of the Toyota Corolla car by using characteristics that describe this model.

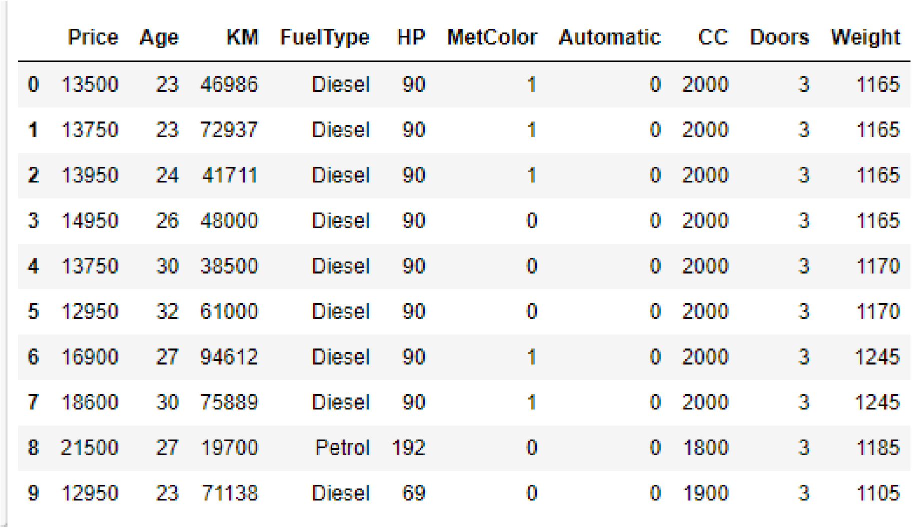
SELECT DATA

We will use 10 characteristics of the Toyota Corolla model to predict its price.

[,1] Age	Age in years
[,2] KM	Accumulated (Kilometers on odometer)
[,3] FuelType	Fuel Type (Petrol, Diesel, CNG)
[,4] HP	Horse Power
[,5] MetColor	Metallic Color? (Yes=1, No=0)
[,6] Automatic	Automatic ((Yes=1, No=0)
[,7] CC	Cylinder Volume in cubic centimeters
[,8] Doors	Number of doors
[,9] Weight	Weight(Kilograms)
[,10] Price	Offer Price (EUROs)

DATA EXPLORATION

An example of 10 rows of data is show below:



	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight
0	13500	23	46986	Diesel	90	1	0	2000	3	1165
1	13750	23	72937	Diesel	90	1	0	2000	3	1165
2	13950	24	41711	Diesel	90	1	0	2000	3	1165
3	14950	26	48000	Diesel	90	0	0	2000	3	1165
4	13750	30	38500	Diesel	90	0	0	2000	3	1170
5	12950	32	61000	Diesel	90	0	0	2000	3	1170
6	16900	27	94612	Diesel	90	1	0	2000	3	1245
7	18600	30	75889	Diesel	90	1	0	2000	3	1245
8	21500	27	19700	Petrol	192	0	0	1800	3	1185
9	12950	23	71138	Diesel	69	0	0	1900	3	1105

After changing several variables to factors the summary statistics for the variables are shown below:

```
> summary(toyota)
     Price            Age              KM            Automatic             CC
          Doors           Diesel
 Min.   : 4350   Min.   : 1.00   Min.   :     1   Min.   :0.00000   Min.   :1
300   Min.   :2.000   Min.   :0.0000
 1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000   1st Qu.:0.00000   1st Qu.:1
400   1st Qu.:3.000   1st Qu.:0.0000
 Median : 9900   Median :61.00   Median : 63390   Median :0.00000   Median :1
600   Median :4.000   Median :0.0000
 Mean   :10731   Mean   :55.95   Mean   : 68533   Mean   :0.05571   Mean   :1
567   Mean   :4.033   Mean   :0.1079
 3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021   3rd Qu.:0.00000   3rd Qu.:1
600   3rd Qu.:5.000   3rd Qu.:0.0000
 Max.   :32500   Max.   :80.00   Max.   :243000   Max.   :1.00000   Max.   :2
000   Max.   :5.000   Max.   :1.0000
      CNG               Age2
 Min.   :0.00000   Min.   :   1
 1st Qu.:0.00000   1st Qu.:1936
 Median :0.00000   Median :3721
 Mean   :0.01184   Mean   :3476
 3rd Qu.:0.00000   3rd Qu.:4900
 Max.   :1.00000   Max.   :6400
```

It is necessary to standardize these results applying a linear method to see the relationship between the variables, it is going to show the actual prediction of the data and we can make a new prediction with the variables. Also, the tree model is going to show the different inputs variables to predict a target value and make decisions.

DATA ANALYSIS

We applied the following steps in our analysis.

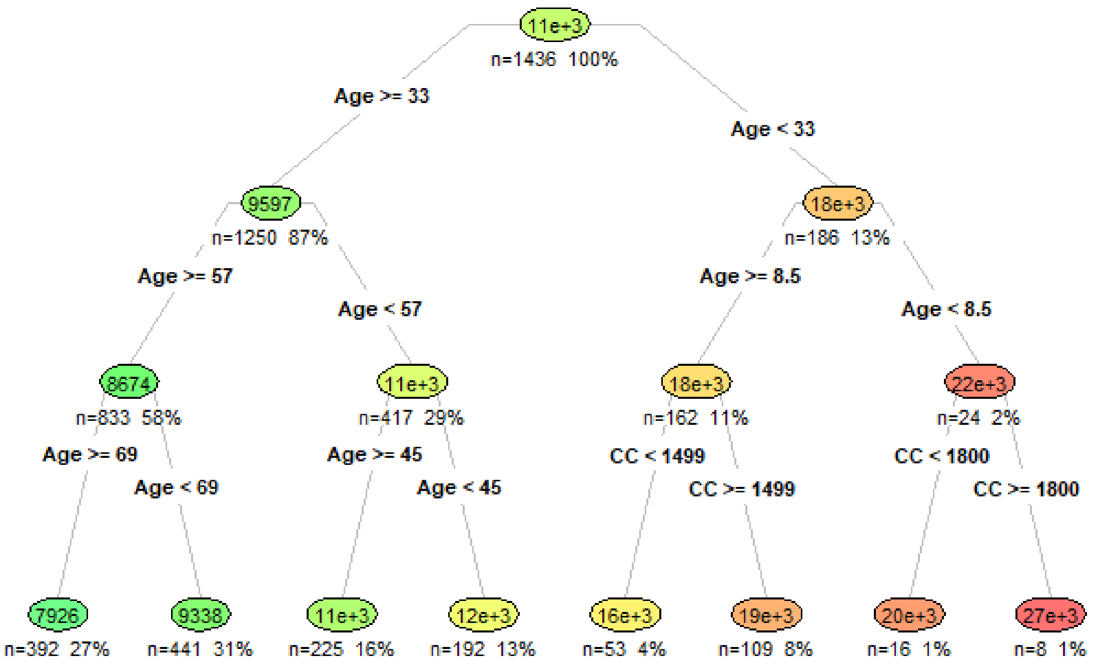
Predicted Price = -1.457e+02 Age + -2.045e-02KM + 7.505e+02 Automatic +3.451e+00CC +1.806e+02Doors + -7.840e+02 Diesel + -4.636e+02 CNG

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.419e+04  5.650e+02  25.121  < 2e-16 ***
Age         -1.457e+02  2.877e+00 -50.633  < 2e-16 ***
KM          -2.045e-02  1.559e-03 -13.114  < 2e-16 ***
Automatic    7.505e+02  1.828e+02   4.106 4.26e-05 ***
CC           3.451e+00  3.500e-01   9.862  < 2e-16 ***
Doors        1.806e+02  4.476e+01   4.034 5.76e-05 ***
Diesel      -7.840e+02  2.241e+02  -3.498 0.000482 ***
CNG         -4.636e+02  3.954e+02  -1.173 0.241122    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1577 on 1428 degrees of freedom
Multiple R-squared:  0.812,	Adjusted R-squared:  0.811 
F-statistic: 880.8 on 7 and 1428 DF,  p-value: < 2.2e-16

An improve model:
```

Insurance Charges



Better Model:

Predicted Price = -3.295e+02Age + -1.803e-02 KM + 6.499e+02 Automatic +3.431e+00 +1.806e+02Doors + -9.622e+02Diesel + -2.717e+02 CNG + 1.873e+00Age^2

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.803e+04  5.413e+02  33.312  < 2e-16 ***
Age         -3.295e+02  9.873e+00 -33.379  < 2e-16 ***
KM          -1.803e-02  1.395e-03 -12.929  < 2e-16 ***
Automatic    6.499e+02  1.629e+02   3.989 6.98e-05 ***
CC           3.431e+00  3.118e-01  11.003  < 2e-16 ***
Doors        1.384e+02  3.994e+01   3.465 0.000546 ***
Diesel      -9.622e+02  1.999e+02  -4.814 1.64e-06 ***
CNG         -2.717e+02  3.524e+02  -0.771 0.440863    
Age2         1.873e+00  9.712e-02  19.287  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1405 on 1427 degrees of freedom
Multiple R-squared:  0.8508,	Adjusted R-squared:   0.85 
F-statistic:  1017 on 8 and 1427 DF,  p-value: < 2.2e-16
```

APPLY ANALYSIS

There is missing data in the variable FuelType which was not included in the analysis.

DEPLOY MODEL

Give coupons to the 1436 observations it was possible to create a prediction in the price of the Toyota Corolla model using correlation, linear regression and a tree model to see the actual and future data to make decisions.

ASSESS RESULTS

We will evaluate the response rate for the targeted customers compared to the response rate for the random sample to see if there was a greater response rate for the price of the Toyota Corolla model.

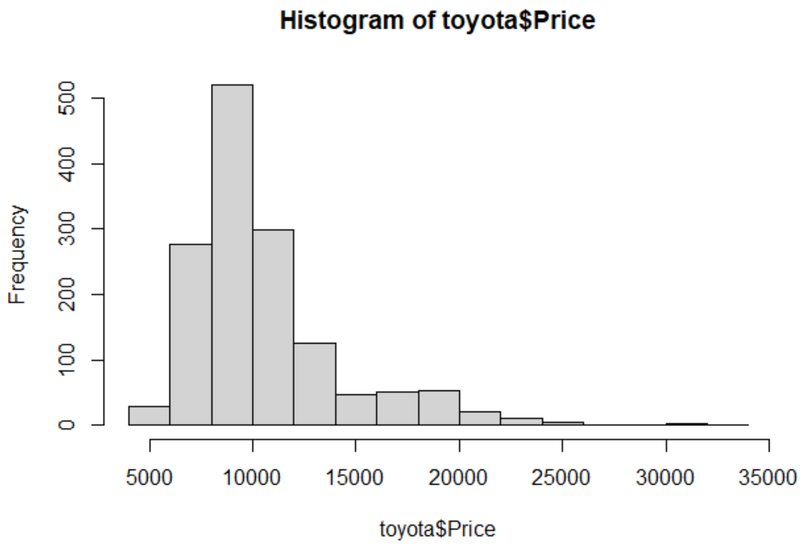
STRENGTHS OF XYZ ANALYSIS

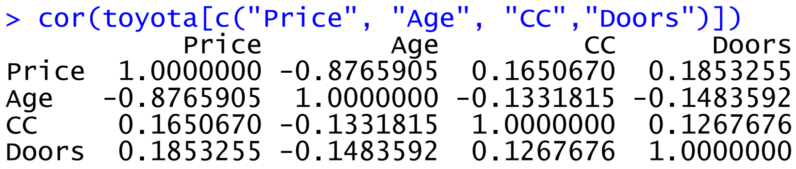
Correlation, linear regression and tree math other analysis has many strengths that allow to calculate the level of change in one variable when other one change.

APPENDIX

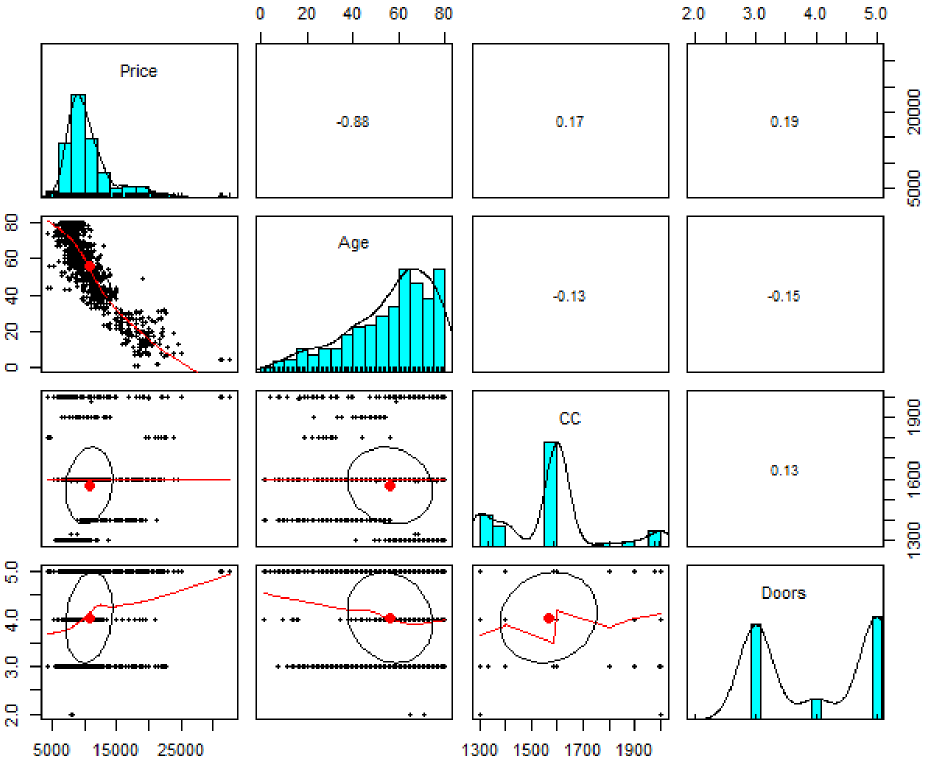
A.

```
> summary(toyota)
     Price            Age              KM           Automatic      
 Min.   : 4350   Min.   : 1.00   Min.   :     1   Min.   :0.00000  
 1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000   1st Qu.:0.00000  
 Median : 9900   Median :61.00   Median : 63390   Median :0.00000  
 Mean   :10731   Mean   :55.95   Mean   : 68533   Mean   :0.05571  
 3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021   3rd Qu.:0.00000  
 Max.   :32500   Max.   :80.00   Max.   :243000   Max.   :1.00000  
       CC           Doors           Diesel            CNG         
 Min.   :1300   Min.   :2.000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:1400   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :1600   Median :4.000   Median :0.0000   Median :0.00000  
 Mean   :1567   Mean   :4.033   Mean   :0.1079   Mean   :0.01184  
 3rd Qu.:1600   3rd Qu.:5.000   3rd Qu.:0.0000   3rd Qu.:0.00000  
 Max.   :2000   Max.   :5.000   Max.   :1.0000   Max.   :1.00000  
```





```
> cor(toyota[c("Price", "Age", "CC","Doors")])
           Price        Age         CC      Doors
Price  1.0000000 -0.8765905  0.1650670  0.1853255
Age   -0.8765905  1.0000000 -0.1331815 -0.1483592
CC     0.1650670 -0.1331815  1.0000000  0.1267676
Doors  0.1853255 -0.1483592  0.1267676  1.0000000
```



Code

```
library(dplyr)
library(rpart)
library(caret)

# Load data
toyota <- read.csv("ToyotaCorolla.csv")
summary(toyota)
# Convert fuel type to dummy variables
toyota$Diesel <- ifelse(toyota$FuelType == "Diesel",1,0)
toyota$CNG <- ifelse(toyota$FuelType == "CNG",1,0)
# Remove unnecessary columns
toyota <- toyota %>% select(-c("HP", "Weight", "FuelType","MetColor"))
#Step 2: Train the model on the data, we are using all the data
ins_model = lm(Price ~ ., data=toyota)

# see the estimated beta coefficients
ins_model



## Step 1: Exploring and preparing the data ----
#start with original insurance data, not edited data from another analysis

# examine the data
str(toyota)

# the distribution of quality ratings
```

```
hist(toyota$Price)

# summary statistics of the wine data
summary(toyota)

## Step 2: Training a model on the data ----
# regression tree using rpart package
ins_tree = rpart(Price ~., data=toyota)
# get basic information about the tree
ins_tree
# get more detailed information about the tree
summary(ins_tree)
```