Statistical Data Science: Project
Academic year 2022 – 2023

Thomas Decorte, Tim Verdonck

The following assignment must be made **individually**. Using R and what you learned in the course you need to analyse the given data set and answer the questions below in a written document. The assignment consists of an analysis of a car data set issued by the government. The given data set contains the fuel consumption and emission data alongside various other characteristics of cars, from 2000 to 2013. The rows of the data set will represent cars and the columns their attributes (see table on the next page for variable explanation). Each student draws an individual data set of a random 3000 cars. You use the following code to obtain your subset of the data, where you change 160597 by your birthday (day, month and year):

```
#Load in data
cars_data <- read.table("SDS_Project_22_23.txt",sep="",header=T)

#Generate subset
set.seed(160597)
cars_data$euro_standard <- as.factor(cars_data$euro_standard)
cars_data$transmission_type <- as.factor(cars_data$transmission_type)
cars_data$fuel_type <- as.factor(cars_data$fuel_type)
subset_cars = cars_data[sample(nrow(cars_data),3000,replace=FALSE),]
```

You answer the questions by performing an appropriate analysis with R. The discussion of the results and the necessary figures are reported in a written text that consists of a maximum of 12 pages (12pt font size). Only report results and interpretations, do not repeat theory from the course! Additionally a separate file with the full R script should be provided.

Your report and R script should be send to Tim Verdonck (tim.verdonck@uantwerpen.be) and Thomas Decorte (thomas.decorte@uantwerpen.be) before **January 4 2023, 23.59**. **Good luck!**

The following table gives you the description of the variables:

manufacturer	Car manufacturer or importer.
model	Car model.
description	Further details on the car model.
euro_standard	Euro Standard to which the record applies.
transmission_type	Transmission type. Either Automatic or Manual.
engine_capacity	Engine capacity in cubic centimeters (cc).
fuel_type	Fuel type this car uses, Diesel, Petrol.
urban_metric	Fuel consumption in urban conditions in liters per 100 Kilometers (l/100 Km).
extra_urban metric	Fuel consumption in extra-urban conditions in liters per 100 Kilometers (l/100 Km).
combined_metric	Combined fuel consumption: average of the urban and extra-urban tests, weighted by the distances covered in each part, in liters per 100 Kilometers (l/100 Km).
noise_level	External noise emitted by a car shown in decibels.
co2	CO2 emissions in grammes per kilometer (g/km).
co_emissions	Carbon monoxide emissions in milligrammes per kilometer (mg/km).
nox_emissions	Nitrogen oxides emissions in milligrammes per kilometer (mg/km).

1 Exploratory Data Analysis

Perform an exploratory analysis on your **full** data set (`subset_cars` in the code above). Remove outliers that could have a strong influence on the outcome of your analysis and potentially transform the quantitative variables to a more normal distribution if necessary. State your main findings (max 2 pages).
For the following exercises continue working with the, whether or not transformed, continuous, cleaned (without the outliers) variables and call the dataframe `subset.trans`.

2 PCA

Next you create your own training and validation set. Here you also drop the variables `manufacturer`, `model` and `description` (as these will not be used in further analysis):

```
set.seed(160597)
train_ind <- sample.int(n=nrow(subset.trans),size=1500,replace=F)
subset.trans <- subset.trans[ , -which(names(subset.trans) %in% c("manufacturer"
                                                  ,"model","description" ))]
Xtrain <- subset.trans[train_ind,]
Xval <- subset.trans[-train_ind,]
```

In this code, 160597 should be replaced by your birthday.

1. Perform a PCA analysis on your **training** data. The data matrix should only consist of the continuous variables from the training data set. Please argue why you base the analysis on the correlation or covariance matrix of the data. Explain how you choose the number of components.

2. Make biplots of the your chosen number of scores. Can you recognize the groups determined by one of the categorical variables (`euro_standard`,`transmission_type` or `fuel_type`)?

3. Discuss whether the training data set contains PCA outliers. Compare your analysis with a robust PCA analysis.

4. Continue with the PCA analysis you find most appropriate. Consider now the observations from the **validation** set, and compute their scores and predicted values.

5. Make an outlier map with the observations from both the training and the validation set (use a different symbol or color). Discuss the result.

3 Clustering

Only consider the variables of the **training** data set. Again for this analysis only use the continuous variables. The cluster analysis should be performed using the observations, not the PCA scores.

1. Perform a *partitioning* cluster analysis on the observations. Discuss your choice of the method, and the choice of the number of clusters. Does any of the clustering solutions represent the groups determined by one of the categorical variables (`euro_standard`,`transmission_type` or `fuel_type`)?

2. Perform a *hierarchical* cluster analysis on the observations. Discuss your choice of the method, and the choice of the number of clusters. Does any of the clustering solutions represent the groups determined by one of the categorical variables (`euro_standard`,`transmission_type` or `fuel_type`)?

3. Perform a *hierarchical* cluster analysis on the variables. Discuss your choice of the method.

4. Consider the clustering of the observations that corresponds mostly to the different types of fuel (available in the `fuel_type` variable). Make a heatmap where the observations and variables are sorted according to the selected clusterings. What are your findings?

4 Linear regression

1. Construct a good linear model using the **training** dataset by selecting and/or transforming variables (do not include interaction terms). The response variable is the `co2` variable. Try to build a model that does not suffer from multicollinearity. Comment on the variable selection process and techniques used.

2. Does your final model from step (1) satisfy the assumptions of the normal linear regression model? If you do not succeed in meeting all these conditions explain the shortcomings of your model and try to adapt the model to satisfy these conditions.

3. Perform an ANOVA analysis on the final model you created in step (2). Consider this your final model in further questions. Discuss the results.

4. Compute the R^2 and R_a^2 of the training dataset of your final model. Compute RMSEP of the validation set (`Xval` in the code above). Compare with the RMSE of the training dataset. Discuss your findings.

5. Based on the final model, compute a 99% confidence interval and prediction interval for the average `co2` of a car with `euro_standard=4,transmission_type="Automatic", engine_capacity=1924,fuel_type='Petrol', urban_metric=9.899, extra_urban_metric=6.898,combined_metric=8.436,noise_level=71.00, co_emissions=142.35,nox_emissions=429.00`. Also perform this analysis for a car with the following attributes: `euro_standard=5,transmission_type="Automatic", engine_capacity=9076,fuel_type="Petrol",urban_metric=25.467, extra_urban_metric=19.437,combined_metric=22.443,noise_level=88.00, co_emissions=291.00,nox_emissions=8077.00`. What are your findings? Are the estimates equally trustworthy?

5 Classification

Build a classifier that attempts to identify the `euro_standard` of a car based on the continuous variables included in the dataframe. For these exercises use the above defined **training** dataset to train a classifier and the **validation** dataset to validate the classifier and compare.

1. Train an LDA and QDA classifier of the data. Evaluate both of the classifiers based on the mosaic plots.

2. Apply the k-nearest neighbours classifier. Choose a good value for k and explain your choice.

3. Compare all of the previous classifiers and choose the most suitable one, elaborate on your decision process.