# Instructions

Data exploration and visualization with R

The dataset ToyotaCorolla.csv contains data on used cars on sale during the late summer of 2004 in the Netherlands.

Please use RStudio to finish the following (refer to DMBA Book Chapter 3 and class lecture for codes). Write your answers to the questions using comment “#” near your codes.

1. Open the R script “Toyota.R” from Assignment 1.
2. Change the working directory use *Session -> Set Working Directory -> To Source File Location* from the Menu Bar.
3. Rerun the read.csv command to import the dataset
4. Use a comment line (“#”) to answer the following question: What prediction question can we answer with this dataset? (3 points)
5. Implement R codes to answer questions from here on. How many different categories does “Color” have? What are they? (3 points)
6. Explore variable “Color” using frequency count and a bar chart. (5 points)
7. Explore variable “Price” using statistics and distribution plots (including all we have learned in class.) (8 points)
8. Is the distribution of “Price” skewed? What might we do to correct the skewness? (3 points)
9. View the data using View() or clicking on the dataset in the Environment tab. Use the arrow beside “Price” column name to sort the data. Check the first three records and the histogram together. They seem to be a lot higher than other cars. Check their manufacture year (which year was the data collected?) and mileage. What would be your conclusion? Are they outliers by error? (3 points)
10. Explore the relationship between “Price” and “Age\_08\_04” with statistics and a graph. Explain your findings. (5 points)
11. Explore the relationship between “Price” and “Mfr\_guarantee” with a graph. Explain your findings. (5 points)

Next, we use ggplot2 to develop more advanced graphs. You should have ggplot2 package installed from class.

1. Import ggplot2 package using library().
2. From Step 10, we explore the relationship between “Price” and “Age\_08\_04” with a graph. Now use ggplot2 to add variable “Fuel\_Type” to this graph. Map “Fuel\_Type” to color. What is the Fuel\_Type for most cars? What Fuel\_Type do the most expensive cars have? (5 points)
3. Following Step 13 (copying the code), use Facet\_wrap() to create multiple panels of the graph based on “Fuel\_Type”, putting the panels in different rows. Is the relationship trend between Price and Age\_08\_04 different across different Fuel\_Types? (5 points)
4. Following Step 14 above (copying the code), add variable “HP”, horsepower, to this graph. Mapping HP to size. Expand the plot window for better observation by clicking the button on the top right corner. Check the cars with CNG fuel type. What can you say about their horsepower? (5 points)