TUTORIAL 6

Aim:

In this tutorial we will cover:

1. **Revision of exploratory data analysis**
2. **Simple linear regression in R**
3. **Multiple linear regression in R**

Mode of study:

You can work through this tutorial in your own time, or in one of the assigned venues during the tutorial slot. Help is available during the tutorial slot in the venues.

Solutions will be made available in Week 9.

Before you start

If you have not worked through Practical 6 yet, you should do so before starting this tutorial.

FIRST DATASET

In Tutorials 2 and 3, you worked with the `diamonds` dataset. We will use the same dataset this week. If you want to refresh your memory, type `?diamonds` in the R console. You can also explore the dataset by typing `str(diamonds)` and `summary(diamonds)`. Remember that the `diamonds` dataset is part of the `ggplot2` package, so you would have to load the `tidyverse` first.

EDA

1. In Tutorial 3, you constructed various plots for this dataset. You should look at the plots again and take note of your findings based on these plots.

   Now create the following plots as well:

   a. **Boxplot of `price` for each `clarity` category separately.**
   b. **Scatterplot of `carat` and `price`, using different colours for various `clarity` categories.**
   c. **Scatterplot of the *x* and *y* dimensions of the diamonds.**
   d. **Scatterplot of the *x* and *z* dimensions of the diamonds.**
   e. **Scatterplot of the *y* and *z* dimensions of the diamonds.**

2. Your scatterplots in Question 1c. – 1e. should indicate to you that there are some outlying observations, and also some observations which may be incorrect. Based on this, you could decide to remove some or all of these observations.

   Create a new dataset called `cleandiamonds`, which only contains diamonds with non-zero *x*, *y* and *z* dimensions. Also exclude all diamonds with dimensions > 30. Print summary statistics for your new dataset.

For the rest of the practical, only use the `cleandiamonds` dataset.

3. **Calculate the correlation between `price` and each of the numeric attributes in the dataset.**

SIMPLE LINEAR REGRESSION

4. In Question 3 you should have seen that there is a very strong positive linear relationship between `price` and `carat`.
   Fit a simple regression model which predicts the price of a diamond based only on its carat. Store your model output in an object called `model1`. View the summary output of your model and write down the estimated least squares regression line.

5. **Create a scatterplot of `price` and `carat` and include the least squares regression line on this plot.**

6. The plot in Question 5 should indicate that the relationship between `price` and `carat` is perhaps not linear. It is hard to tell though, with some outliers still present. Therefore, **redraw the plot in Question 5 but only include diamonds less than 3 carats.**

7. The plot in Question 6 confirms that the relationship is not linear. One way of dealing with such a relationship is to take logarithmic transformations of the variable(s) instead.
   Create a scatterplot of the logarithm of `price` and the logarithm of `carat` and include the least squares regression line on this plot. (Use the dataset created in Question 6 which excludes diamonds of more than 3 carats.)

SECOND DATASET

The second dataset we will be considering in this tutorial, is the `Carseats` dataset, which is part of the `ISLR` package. This package contains the datasets used in the book *An Introduction to Statistical Learning with Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani. (This book will be used in later Data Science modules.)

`Carseats` is a simulated dataset, containing sales of child car seats at 400 different stores. There are 10 explanatory variables in the dataset.

8. **Install (only if you are working on your own computer) and load the `ISLR` package.**

EDA

9. **Read the help file for the `Carseats` dataset and pay special attention to the variable descriptions.**

   Take note that there are three categorical variables in the dataset: `ShelveLoc`, `Urban` and `US`. (You can verify this by examining the structure of the dataset.)

10. **Construct suitable plots to investigate the relationship between the target variable `Sales` and the numerical variables `CompPrice`, `Income` and `Price`.**

11. **Construct a suitable plot(s) to investigate the relationship between the target variable `Sales` and the categorical variable `ShelveLoc`.**

MULTIPLE LINEAR REGRESSION

12. **Fit a multiple regression model which predicts the sales of carseats based on the other attributes in the dataset. Store your model output in an object called `model2`. View the summary output of your model.**

In your output, you will see that there are now more estimated coefficients than there are attributes in the dataset! This is because R automatically creates dummy variables for the categorical attributes in the dataset.

In the case of a binary categorical variable (i.e. one with two possible levels, such as `Urban` and `US`), there is no need to create dummy variables. However, for `ShelveLoc`, which is a categorical variable with 3 levels (namely "*Bad*", "*Good*" and "*Medium*"), 2 dummy variables will be created. Consider the following example of dummy coding:

Value of original variable	Value of new variable `ShelveLocGood`	Value of new variable `ShelveLocMedium`
Bad	0	0
Good	1	0
Medium	0	1

Note that while a new variable `ShelveLocBad` was not created, we can deduce that a observation has the value *Bad* since it is not *Good* or *Medium*.

You can view the coding that R uses for the dummy variables by calling the `contrasts()` function; i.e. by calling `contrasts(Carseats$ShelveLoc)`.

13. **Use the model constructed in Question 12 to predict the number of carseat units sold at a location with the following attributes:**

Attribute	Value
CompPrice	100
Income	70
Advertising	5
Population	250
Price	150
ShelveLoc	Medium
Age	49
Education	11
Urban	No
US	Yes

(Remember that `Sales` is measured in thousands of units!)

(You can also try to manually replicate this prediction…)