Student name: ______________________ Matriculation Number: _____________

CMM020 Data Visualisation and Analysis SAMPLE Practical Examination

Do the following:

1. **Complete the time-limited quiz which is available on CampusMoodle.**
2. **Carry out each of the tasks detailed below.**

Note: some of the tasks below have been covered in the lab. For the real practical examination, the tasks will not have been covered in the labs, but will be similar.

Set Up

I. Create a folder called *CMM020* in the *H:* drive.
II. Open RStudio and create a new RMarkdown file. Delete unnecessary files and save the file on the folder that you created above with name CMM020PE.rmd
III. Load library tidyverse

This assessment consists of three tasks as detailed below. Carry out each of the tasks detailed below using R code where coding is required. Also ensure you include comments/discussions as requested below in your RMarkdown file.

Before you start work on the tasks, download all the files specified below.

Downloads

Download the following files from CampusMoodle (labs section or assessed lab section):

1. **provenOilReservesZerosT.csv:** it contains data regarding proven oil reserves for European countries.
2. **provenOilReserveWEurope.csv:** contains data regarding proven oil reserves for western European countries.
3. **channels.csv:** contains perceived sensation vs stimuli for various channels.
4. **OilQuality.**csv: contains price and API gravity for oil.
5. **UKData:** contains various data related to hydrocarbon production.

Data Visualisation Task (RStudio)

Load the first 3 datasets (provenOilReservesZerosT.csv, provenOilReserveWEurope.csv and channels.csv).

You are required to produce the visualisations below using R. Visualisations must have appropriate titles and labels:

1. A scatterplot of yearly proven oil vs Denmark. Use "darkblue" squares for the line in your points and "blue" for the fill. If you can, make the size of the squares be a factor of the value they represent. Otherwise, make all the squares of size 4. Discuss the quality of the visualisation and what the visualisation suggests. Comment on whether similar conclusions would have been made if the data available was only for the period 1990-2005. Use the ProvenOilReservesZeosT dataset.

2. A line plot containing size of stimuli vs. size of sensation for electricity, length and area.
   Make any adjustments which you feel are necessary to fully appreciate the difference in size of sensation between these 3 stimuli. Use colour appropriately. Use the channels dataset.

3. A line plot of year versus Million Tons of Barrels by country in Western Europe. If you can, make each line a different colour. Discuss the quality of the visualisation that you have produced and interpret the results obtained. Use the provenOilReserveWEurope dataset.

4. Proven oil reserves in Western Europe by country using facets. If you can and make the colour dependent on the country. Discuss the visualisation obtained. Use the provenOilReserveWEurope dataset.

Data Analysis Task (RStudio) – linear regression

Read file **OilQuality.csv** into RStudio.

The file contains paired observations of two variables.

- Price – the price of oil from a particular source. [Dollars per Barrel]
- API – the API gravity of oil from that source. [No units]

5. Calculate the correlation and covariance between the Price and the API gravity.

6. Produce a scatterplot of Price against API gravity, including a straight line fit to the data. Comment on your plot.
7. Calculate a linear regression of Price (response) on API gravity (predictor) and display the summary of the results.
   State the equation of the regression line and the coefficient of determination.
   Two new batches of oil have API gravities of 35.5 and 45.1. Predict the selling price of each. Comment on the validity of these predictions.

Data Analysis Task (RStudio) – Data Manipulation and Statistical Inference

Read file ***UKdata.csv*** into RSudio **-** a file containing the following monthly data for the UK for 2014.

- ***a.*** ***Month*** : specifies a month in 2014 in mm/yyyy format.
- b. ***OPUKShare:*** Oil production UK Share – measured in “m3”.
- c. ***AGPUKS:*** Associated gas production UK share - measured in “Ksm3”
- d. ***IWVUK:*** Injected Water Volume UK - measured in “Ksm3”.
- e. ***OilField:*** the oilfield the data is for.
- ***f.*** ***Operator:*** the operator for the oilfield the data is for.

8. Add a new column to UKData called bigProducerOil which states whether that month the operator was producing more oil (OPUKShare) than the average OPUKShare for the year. How many instances of the operator producing more than average are there?

9. Run an experiment to analyse the true mean for AGPUKS. You want to check whether it is greater than each of the following 3 values: 12000, 10000 and 8000. Justify your answers at 95% confidence level.

10. Analyse whether there is a statistically significant difference in Oil Production based on the operator. Use a 95% confidence level.

Submission

You are required to use the drop box to submit:

- The Rmd file containing the code, the results and your explanations

- An html file which you have obtained by knitting your Rmd file.

You are also required to complete the time-limited Quiz on CampusMoodle.

This specification sheet (i.e. this document) must be submitted to the invigilator at the end of the exam. Your name and matriculation number must appear on the top of the first page.