1 MAT013 - Practice Sheet

1.0.1 Attempt to do the following in SAS and/or R.

1. Create a new data set by finding the age and height of everybody in your class.

   1. Obtain the mean, minimum, maximum and standard deviation of the age and height.
   2. Create a new frame that gives the above statistics.
   3. Output this to csv.

2. Download the file trees91.csv.

   1. How many variables and observations are there in this data set?
   2. Obtain some summary statistics for all the variables in the data set.
   3. Are there any non numeric variables? If so, obtain frequency tables and a pie chart on all the non numeric variables.
   4. Is there a correlation between any of the variables?

3. Download the file migration.csv. Output a table of summary statistics for the migration data where the variable mass is less than 24 (you might need to do look up a few things).

4. Download the data sets weight.xls and height.xls:

   1. Create a SAS data set containing the bmi of the observations.
   2. Output the data set to csv.

5. For the concatenated data set (of JJJ and MMM):

   1. For individuals over the age of 25, calculate the yearly average savings (for each year after their 25th birthday)
   2. Output a frequency table showing the mean yearly average by sex.
   3. Obtain the mean yearly average savings by sex and age groups:
      - Group A [0,18]
      - Group B [19,65]
      - Group C [66,]
   4. Find the mean, max and min for the variable "random number" and output this to a SAS data set.

6. Create (seperate) data sets containing the following:

   1. The first 200 odd numbers;
   2. The square root of the first 1000 integers;

3. The square root of the first 10000 integers, selecting only those that are integers
4. The first 20 prime numbers (this is slightly harder)

7. Download the file coordinates.csv.

1. Obtain coefficients for a line fitted to the coordinates (x, y) of this data set.
2. For each of the observations create a new variable giving the following:
   1. The product xy;
   2. The ration x/y;
   3. The exponent x^y;
   4. The sum $x + y$;
   5. The difference $x - y$;
   6. The absolute value of x;
   7. The square root of x;
   8. The log to the base 10 $\log_{10}(x)$;
   9. The natural log $\ln(x)$.
3. Obtain the following for the above data set (including the newly created variables):
   1. minimum
   2. maximum
   3. total
   4. mean
   5. median

8. The probability that an $M/M/k$ queue is empty is given by the following formulae:
$$\pi_0 = \frac{1}{\sum_{i=0}^{k-1} \frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^k}{k!(1-\lambda/(k\mu))}}$$

Obtain π_0 for $\lambda \in \{.1, .5, .7, .9\}$, $k = 3$ and $\mu = 1$.

9. Download the data set marks.csv

1. Create a data set that shows the percentage increase or decrease of marks for each student from one month to the next.
2. Obtain a table showing the mean of these percentage changes categorized by gender.

10. Download the data set numbers.csv and create the following variables:

1. $y1 = x1 \times x2 \times \ldots x36$
2. $y2 = (1 + z1/1) \times (1 + z2/2) \times \ldots (1 + z36/36)$

3. $y3 = x1 - x2 + x3 - \cdots - x36$

11. Create code (in SAS: a macro, in R: a function) that gives

    1. The first k even numbers.
    2. The first k odd numbers.
    3. The square root of the first k integers.
    4. The square root of the first k integers, selecting only those that are integers.

12. Write code that creates a data set with the mean of a given variable for a given data set. Experiment on the JJJ and MMM datasets.

13. Write code that merges two named data sets on a given variable. Remember to ensure that the data sets are sorted as required. Use this to merge the datasets height.xls and weight.xls on the variable "obs" (seen in the previous lab sheet).

14. Download the data set demographics.csv. Write code that creates another data set only containing a given postcode.

15. The mean number of people in an $M/M/k$ queue, where λ denotes the inter arrival rate, μ denotes the service rate and k denotes the system capacity is given by:
$$L_q = \frac{(\lambda/\mu)^{k+1} \pi_0}{kk! \, (1 - \lambda/k\mu)^2}$$
where π_0 is given on the week 2 exercise sheet. Write code that gives L_q for given λ, μ and k. Make sure that your code only runs for $\frac{\lambda}{k\mu} < 1$.

16. Write code that calculates the monthly balance on a loan, for a given, initial balance, repayment value, interest rate and number of months.

    1. Adjust the code so that it displays a plot of balance against time.
    2. Adjust the code so that the plot has an appropriate title.

17. Modify the previous code so that it can output the required monthly repayment value to repay the loan within a given amount of time.

18. Solve the following optimisation problem (we have not see how to do this in R):

    Maximise: $f(x_1, x_2, x_3) = x_1 + x_2 + 5x_3 + x_4^2$ subject to:

$$3x_1 + 2x_2 - x_3 + x_4 \leq 1$$
$$-2x_1 - 3x_2 + 2x_3 - x_4 \leq 1$$
$$x_2 - x_4 \leq 8$$

19. Using SQL merge the data sets weight.xls and height.xls and calculate the bmi. Obtain a new data set only with clinically obese individuals.

20. For the data set demographics.csv. Using SQL create 4 data sets that:

    1. contains all columns and all rows.
    2. contains the columns patientnum, surverynum, heightcm, weightkg, postcode and rows where postcode is not missing.
    3. contains mean values for heightcm and weightkg by postcode and ordered by postcode.
    4. contains mean values for bmi by postcode and ordered by postcode.

21. Download the data sets queue_servers.csv, queue_arrival_rates.csv and queue_service_rates.csv:

    1. Using the mathematical formulae seen in previous weeks obtain a new data set with the Number of the Queue, the service rate, the arrival rate and the mean queue length (note that the formulae only holds for $\lambda < c\mu$).
    2. Is there a correlation between the service rate and the queue length? Between the arrival rate and the queue length?
    3. Obtain a regression model for the mean queue length.

22. Attempt to redo all graphs (from the course) with the ggplot pckage (this is only relevant to R).