Intro to R Programming: Class Test 1

Class Test Rules & Conditions

- The class test is taken under exam conditions.
- During the test you are not allowed to talk to or otherwise communicate with other students (email, instant messaging, etc.), or access the internet/course material. The only material you can use is the printed copy of R reference manual provided as well as the R help within RStudio.
- **DO NOT** open other web pages - you are only allowed two windows open (R studio and the Moodle Class Test 1 page)
- *Please note we use technological means to check compliance with the above!*
- Students who breech the rules above will be reported to the Clerk of Senate.

Starting the test ...

- You will be already logged in to the computer.
- Please log in to Moodle and navigate to the Introduction to R Moodle page.
- Click on the link to begin the test.
- You should use an R studio (or R) window to trial and test your answers. Once you have written and tested the code for a question, copy the code into the answer field for the question.
- In the answer field for each question, only include the code with answers for that specific question.

During the test ...

- You can move back and forward through questions during the class test period.
- **If you have any issues logging in to Moodle, or cannot locate the link to start the class test, please let one of the tutors know immediately**
- **Once open, do not close the moodle browser window**
- The only external packages you are allowed to use (if you wish, they are not required) are `dplyr`, `ggplot2` and `reshape2`.
- If you have experience any issues with the technology throughout the class test, please speak to a tutor immediately.
- For all parts in the test give the R code which can be used to answer the questions. All questions should be answered programatically and should not be hard coded.
- You should only include the code to answer the question, you do not need to include comments or output from the console window.

MSc Class Test 1: Movies

The data file `julymovies.csv` contains data on movies which were screened in American cinemas on 30^{th} July since 1999. It contains the following columns:

`julymovies.csv`	
title	movie title
distributor	company responsible for distributing movie
gross	gross revenue for movie on given date $1000s
theaters	number of theaters screening the movie on the given date
days	number of days the movie had been in theaters by the given date
date	date the row of data corresponds to
top	Indicates if movie is in top grossing 100 American movies of all time(logical)*

Data originally from the-numbers.com, top grossing movies have taken from figures not adjusted for inflation

1. [**2 marks**] Use R to read in the file `julymovies.csv` correctly and save it as a dataframe called `movies`.
2. [**2 marks**] Define a variable called `lionsgate` which contains the number of movies within the dataframe `movies` which were distributed by Lionsgate.
3. [**2 marks**] Define a vector called `missing` which contains the number of missing values for each variable in the `movies` dataframe.
   In other words, the vector `missing` should have many elements as there are columns in the `movies` dataframe and each element in `missing` should correspond to one of the columns.
4. [**2 marks**] Update the `movies` data frame by removing all rows where the values of `days` and `top` are missing. The updated dataframe should be called `movies`.
5. [**3 marks**] Define a variable called `average.gross` which contains the average revenue per theatre for movies distributed by Walt Disney.
6. [**2 marks**] Define a variable called `log.gross` which contains, for each movie in the dataframe, the log transformed gross revenue. Add this variable to the `movies` dataframe. (The additional column name should be `log.gross` and the resulting dataframe should still be called `movies`).
7. [**2 marks**] Sort the `movies` data frame by number of theaters. The sorted dataframe should be called `movies`.

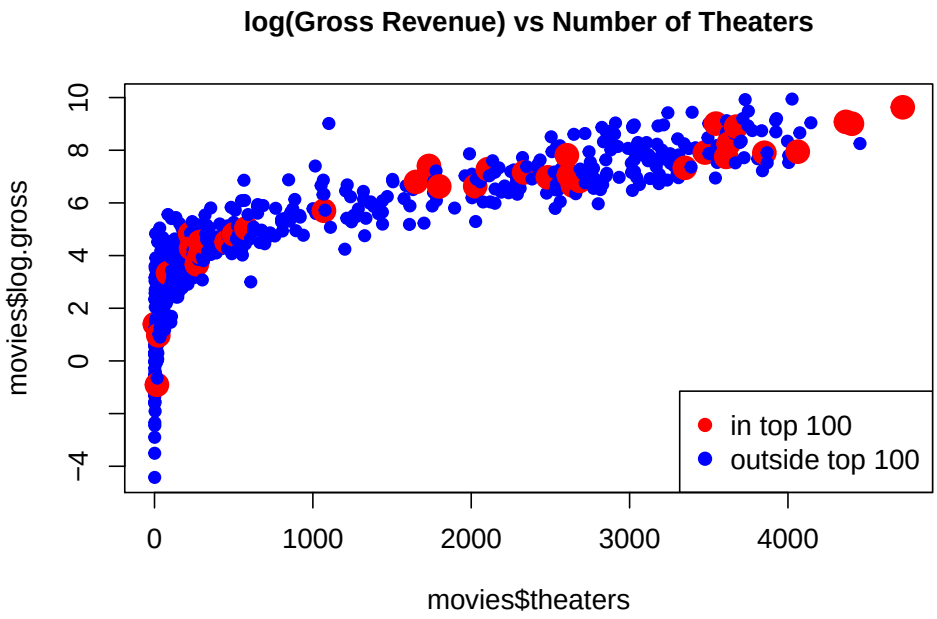
If you have not managed to complete questions 1 to 7 you can use the following code to generate data which can be used to answer questions 8 to 11.

```
theaters <- c(seq(1,1000,length.out=300), seq(1001,6000,length.out=315))
set.seed(123); log.gross <- log(theaters)+rnorm(615)
```

8. [**5 marks**] Produce a plot of `log.gross` by `theaters`.

   Your plot should look similar to the one at the top of page 3. On your plot..

- Points representing movies which were in the top 100 grossing American movies of all time should be coloured red, while those which were not in the top 100 should be coloured blue.
- You should set `pch=19` (this changes the plotting character to an solid circle)
- Points representing the movies which were in the top 100 grossing American movies of all time should be twice the size of those representing movies not in the top 100.
- The title of the plot "log(Gross Revenue) vs Number of Theaters".
- There should be a legend the same as the one shown on the plot on page 3.



9. **[3 marks]** If we want to model the relationship between the number of theaters a movie is shown in (`theaters`) and the log transformed gross revenue a film makes (`log.gross`) we can use fractional polynomial regression.

   For a covariate vector $\mathbf{x} = (x_1, ..., x_n)$ the design matrix for fractional polynomial regression of degree 1 takes the form;

$$X = \begin{bmatrix} \frac{1}{x_1} & \frac{1}{\sqrt{x_1}} & 1 & \sqrt{x_1} & x_1 \\ \frac{1}{x_2} & \frac{1}{\sqrt{x_2}} & 1 & \sqrt{x_2} & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{x_n} & \frac{1}{\sqrt{x_n}} & 1 & \sqrt{x_n} & x_n \end{bmatrix}$$

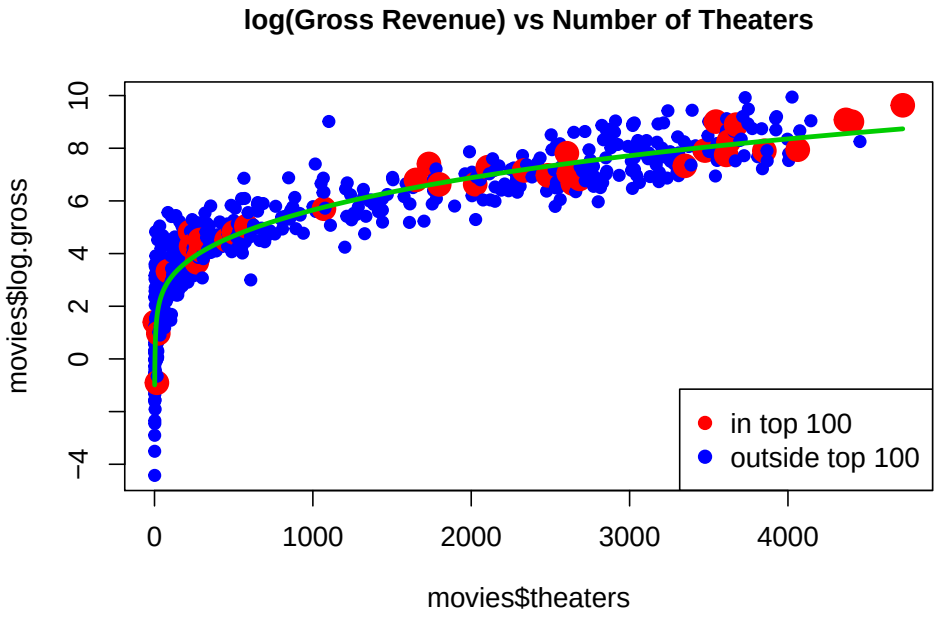
   Define a matrix `X` which is the design matrix required for fitting a fractional polynomial regression model where log transformed gross revenue is the response, $\mathbf{y}$, and number of theaters is the covariate, $\mathbf{x}$.

10. **[3 marks]** Using the design matrix, X, from Question 9 define a vector `y.hat` which contains, $\hat{\mathbf{y}} = (\hat{y}_1, ...\hat{y}_n)$, the fitted values for the fractional polynomial regression between the log transformed revenue (response) and number of theaters (covariate).

    The fitted values can be computed using;

$$\hat{y} = X(X^T X)^{-1} X^T y$$

11. **[2 marks]** Add the fitted fractional polynomial regression line to the plot produced in Question 8. You should add a thick green line to represent the fitted model. Your plot should look like the one at the top of page 4.



12. [**4 marks**] R^2 can be used as a measure of how well a regression line fits a set of data. For a response variable denoted $\mathbf{y} = y_1, ...y_n$, R^2 can be calculated as

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$$

where $\hat{\mathbf{y}} = (\hat{y}_1, ..\hat{y}_n)$are the fitted values and $\bar{y} = \frac{\sum_i^n y_i}{n}$ is the mean of y.

Define a variable `Rsq` which contains the R^2 value for the fitted fractional polynomial regression line computed in Question 10.

13. [**4 marks**] Susan wants to see a movie which starts at exactly 8pm. She leaves her house to go to the cinema randomly at any time between 7.00pm and 7.30pm and her journey there can take anywhere between 30 and 45 minutes depending on traffic. Assume that the length of her journey is also random.

Use a simulation based on 1000 possible journeys to define a vector called `late` which contains the probability Susan arrives after the movie starts at 8pm.

Please enter and run the line of code below before carrying out your simulation.

```
set.seed(123)
```

Hint: You can use the function `runif(n,a,b)` to generate `n` draws from a random uniform distribution with lower limit `a` and upper limit `b`.

14. [**4 marks**] For this question you should answer based on your simulation from Question 13.

i) Define a variable called `waiting` which contains the average number of minutes that Susan has to wait before the film begins.

ii) Define a variable called `unseen` which contains average number of minutes of the movie Susan will miss if she is not there for the start of the movie.