Assignment 3

1. In this analysis, first, we are going to predict the IMDb score of a film based on the running time. This is not very related to engineering, but this will give you an opportunity to try something different.
   IMDb is an online database of information related to movies, television programs, home videos, video games, and online streaming content. IMDb registered users can cast a vote (from 1 to 10) on every released title in the database. Individual votes are then aggregated and summarised as a single IMDb score, visible on the title's main page.
   So, let us assume that you are working for a film production company which produces films on Action, Animation, Biography and Comedy. You are asked to provide some insights for the management using the IMDb scores for the next production. To achieve this, we are going to look at the data in the CSV - movies.csv. The data has been cleaned and processed already.

   a) Read the data into MATLAB.
   See MATLAB code in question h.

   b) Which movie has the highest IMDb score, and what is that score?
   The movie with the highest score is 'Batman: The Killing Joke' with a score of 8.7000.

   See MATLAB code in question h.

   c) Produce a scatter plot of score against run time. Include your plot and a caption. Interpret the plot.

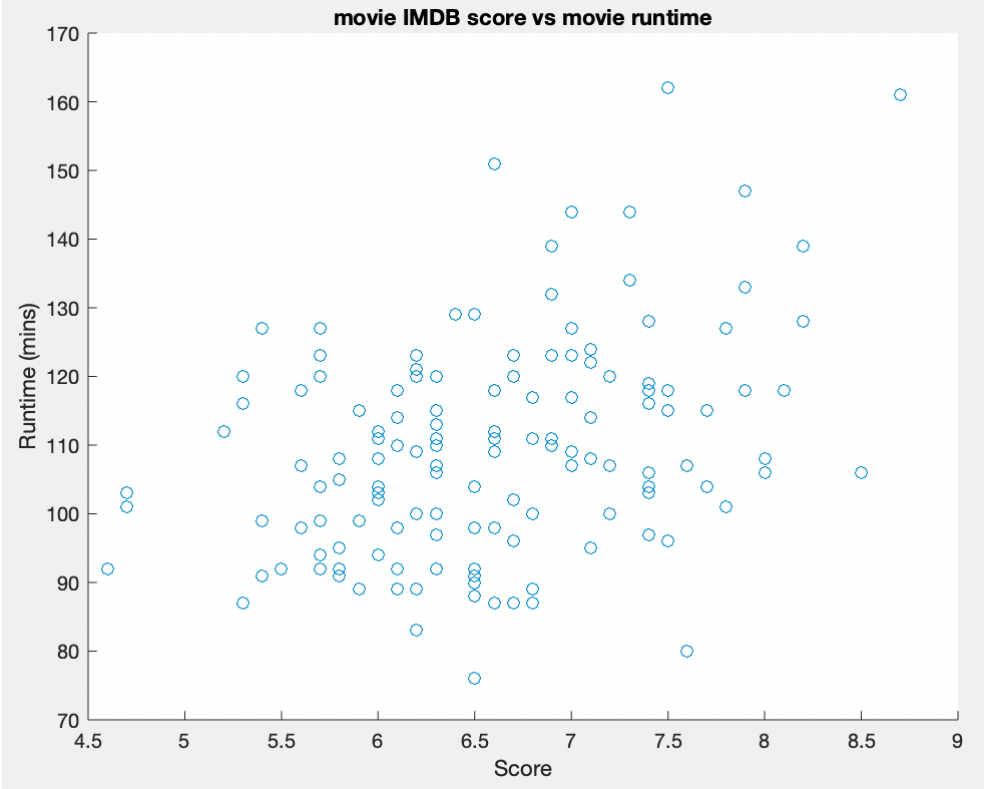

Figure 1: Scatterplot of IMDB score vs runtime

From the scatterplot we can see that majority of the movie ratings lie between 5.5 and 8 and most of the movie runtimes are between 90 to 130 minuets long.

See MATLAB code in question h.

d) Fit a linear model of score against run time. If the run time increases by one minute, what is the expected change in score? Is there a statistically significant linear relationship between score and run time? Justify your conclusion with reference to the P-value. Include the output from your linear regression.

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    ______    __________

    (Intercept)      4.5119      0.45019    10.022     3.724e-18
    x1              0.018827    0.0040693    4.6268    8.3768e-06


Number of observations: 142, Error degrees of freedom: 140
Root Mean Squared Error: 0.758
R-squared: 0.133,  Adjusted R-Squared: 0.126
F-statistic vs. constant model: 21.4, p-value = 8.38e-06
```

Figure 2: MATLAB terminal showing the Linear Regression Model

If the runtime increase by one minute the expected change in score is 0.018827.

The p-value being <= 0.05 indicates that there is a < 5% chance that the results are random, meaning you can reject the null hypothesis. From our results we can see that the p-value is 8.38e-06 which is <= 0.05, therefore there is a statistically significant linear relationship between score and run time.

See MATLAB code in question h.

e) Produce the correct plot to check for constant spread of the residuals. Describe what you see. Do you think there is constant spread?

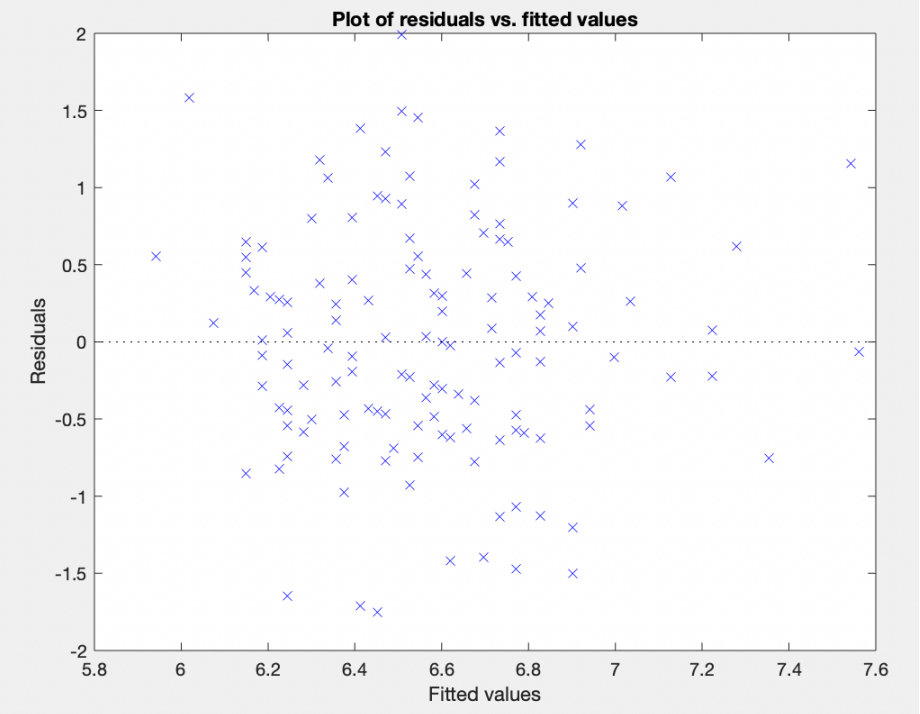

Figure 3: Plot displaying constant spread of the residuals

From the plot we can see that the data appears to be randomly scattered around the zero, therefore it appears as though there is a constant spread.

See MATLAB code in question h.

f) Produce the correct plot to check the normality of the residuals.

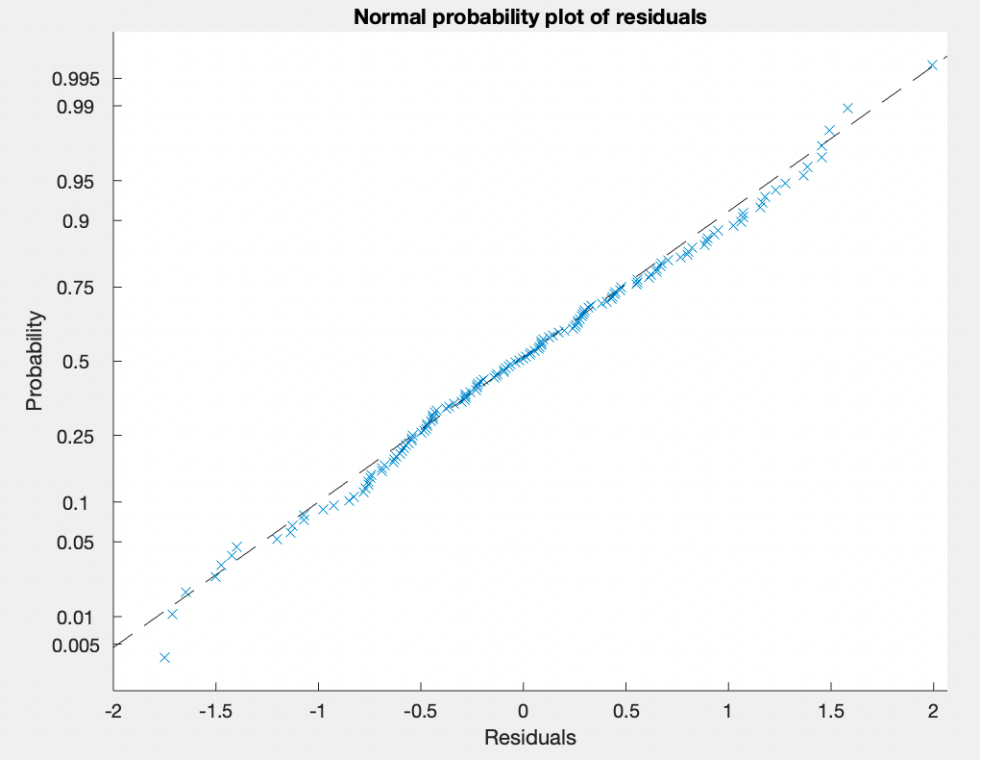

Figure 4: Plot displaying normality of the residuals

g) The boss will fund an advertising campaign for the upcoming movie production if the predicted IMDb score is greater than 6 out of 10. You know that the running time is 120 minutes for the new movie. Use matlab to calculate an appropriate interval to see if this is the case. What is your recommendation?

```
ypred =

    6.7712


ypi =

    5.2647    8.2778
```

Figure 5: MATLAB terminal showing predicted score and appropriate interval

The predicted score of a movie with a 120-minute runtime is 6.7712 with an appropriate interval between 5.2647 and 8.2778. Considering these results, I would recommend that the boss funds the advertising campaign. I would make this recommendation because the predicted value > 6. Furthermore, the lower bound of the appropriate interval is only 0.7353 less than the desired score of at least 6 whereas the upper bound is 2.2778 greater than 6, therefore within this interval there is a higher likelihood that the movie will score a 6 or above.

h) Include your code for question 1. There are marks for well commented, clear code.

```
% Read in data
movies = readtable('movies.csv');

% Loop through the list of movies to find which has the highest IMDB score
highest = 0;
highest_index = -1;
for i = 1:1:(length(movies.name))
    if movies.score(i) > highest
        highest = movies.score(i);
        highest_index = i;
    end
end

% Display the name and score of the highest scoring movie
disp('The movie with the highest score is');
disp(movies.name(i));
fprintf('with a score of %.2f', highest);

% Scatter plot of score vs runtime
figure; scatter(movies.score , movies.runtime);
xlabel('Score');
ylabel('Runtime (mins)');
title('movie IMDB score vs movie runtime');

% Linear model comparing runtime and score
movies_lm = fitlm(movies.runtime,movies.score)
anova(movies_lm, 'summary')

% plot to check for constant spread of the residuals
figure; plotResiduals(movies_lm,'fitted');

% normality of the residuals
figure; plotResiduals(movies_lm,'probability');

% appropriate interval of IMDB score for movie 120mins long
[ypred,ypi] = predict(movies_lm,120, 'Prediction','observation');
```

2. Now, the boss claims that, on average, movies will receive the same IMDb score irrespective of the genre. To check this claim, you have decided to run an ANOVA test.

a) Perform an ANOVA test to check the boss's claim. Include your ANOVA table and justify your conclusion with reference to the P-value.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	13.3508	3	4.45027	7.73	8.21491e-05
Error	79.4228	138	0.57553		
Total	92.7736	141			

Figure 6: ANOVA Table

From this data we can see that the p-value is 8.21491e-15 which is < 0.05. this suggest that there is a statistically significant linear relationship between score and genre. From these results, I disagree with the bosses claim and instead conclude that the genre does impact the IMDB score.

b) Run a multiple comparison test, include the means plots for each genre, and compare the mean scores. According to the plots, which genre can be recommended for the upcoming movie?

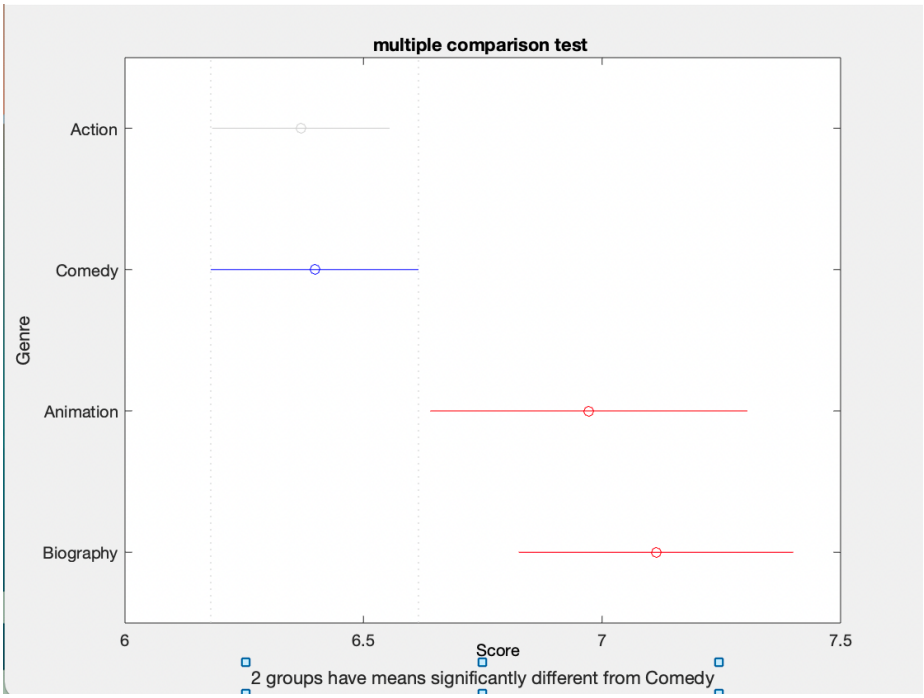

Figure 7: multiple comparison test comedy genre

This figure shows us that the action genre mean is not significantly different to comedy mean. However, the animation and biography genres has significantly different means compared to the comedy genre.

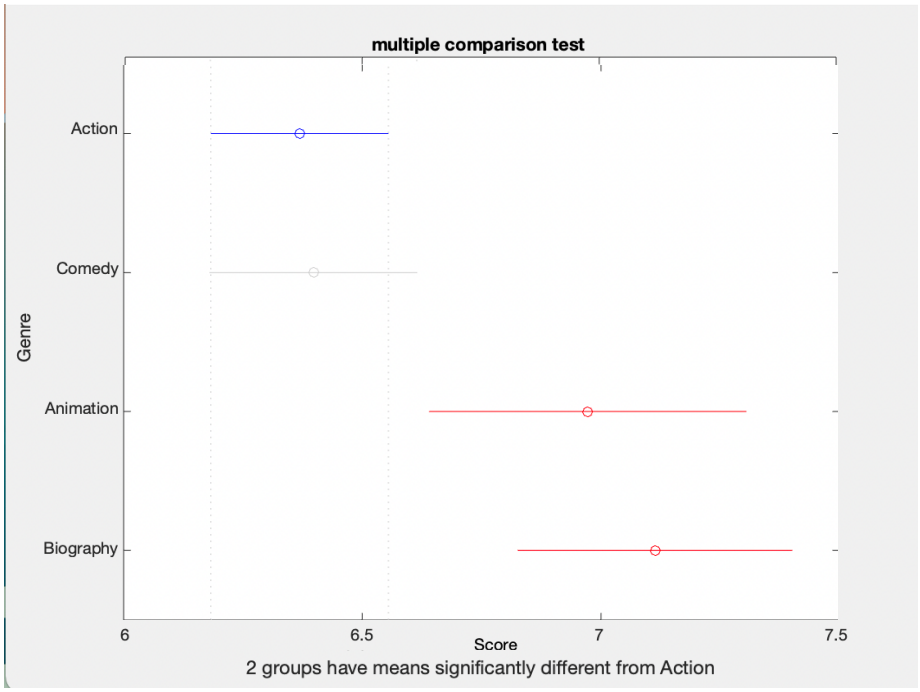

Figure 8: multiple comparison test action genre

This figure shows us that the comedy genre mean is not significantly different to action mean. However, the animation and biography genres has significantly different means compared to the action genre.

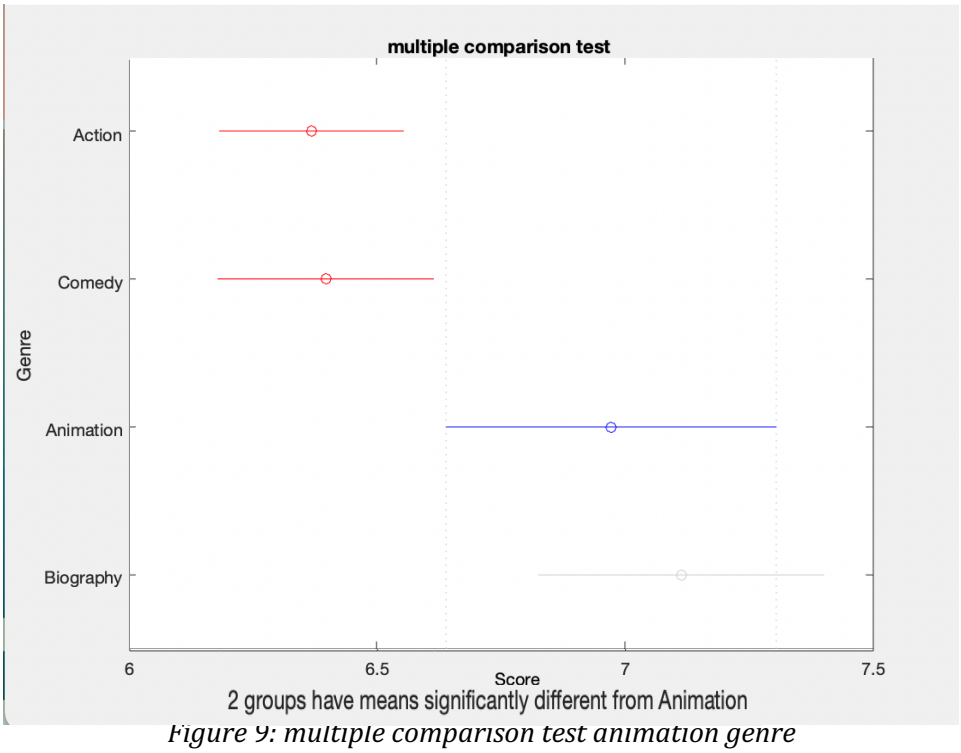

Figure 9: multiple comparison test animation genre

This figure shows us that the animation genre mean is not significantly different to biography mean. However, the comedy and action genres has significantly different means compared to the animation genre.

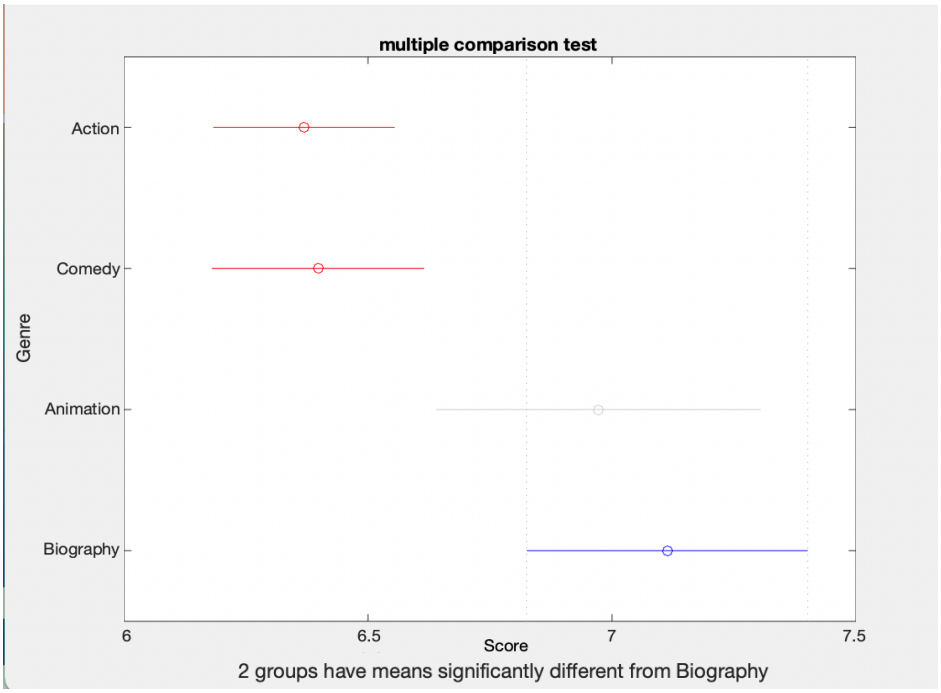

Figure 10: multiple comparison test biography genre

This figure shows us that the biography genre mean is not significantly different to biography mean. However, the comedy and action genres has significantly different means compared to the biography genre.

From looking at the plot I would recommend that the next movie produced is a biography. This is because this genre has the highest lower bound score and the highest upper bound score. Furthermore, biography had the highest mean score. Therefore, the film is more likely to score highly if it is a biography.

c) Include your code for question 2. There are marks for well commented, clear code.

```
% ANOVA test
[p, tbl, stats] = anova1(movies.score, movies.genre);

% multiple comparison test of the score vs genre
multcompare(stats);
```