|  |  |  |
| --- | --- | --- |
|  | **PES University, Bengaluru**  (Established under Karnataka Act 16 of 2013) | UE17MC523 |
| **OCTOBER 2018 : IN SEMESTER ASSESSMENT (ISA) MCA III SEMESTER**  UE17MC523 **INTRODUCTION TO DATA SCIENCE** | | |
| Time: 1½ Hours Answer All Questions Max Marks: 40  Part- A – Quiz – 20 Marks  Part B – Practical – 20 Marks | | |

**PART B – 20 Marks**

**Dataset: movies**

**Research Question: Which parameter highly fluctuate the profit of the movie? Is there a relationship between specific genre and year?**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Question** | **Answer**  **write the R code in answer column** | **Marks** |
| 1. | **Read & Examine Data**  Inspect the dataset and write how many observations and variables are there in the given dataset | movies = read.csv('movies.csv')  summary(movies)  **Observations: 2961**  **Variables:** **11** | 2 |
| 2. | Add a column profit to the movie dataset which has the value derived by applying the formula  profit=gross-budget | movies['profit'] = movies$gross - movies$budget | 2 |
| 3. | **Analysis**  Check out how profit fluctuates relative to each movies rating. Do the initial check with the plotting a graph. And a linear line of best fit. Justify your results with values | plot(movies$profit, movies$rating)  cor(movies$profit, movies$rating)  abline(lm(movies$profit~movies$rating))  **Less positive correlation of 0.295** | 3 |
| 4. | Check out which other variables may influence the profit. Represent it in a graph which has the correlation values | plot(movies$profit, movies$cast\_facebook\_likes)  cor(movies$profit, movies$cast\_facebook\_likes)  **Corr: No correlation**  plot(movies$profit, movies$budget)  cor(movies$profit, movies$budget)  **Corr: No correlation**  plot(movies$profit, movies$votes)  cor(movies$profit, movies$votes)  **Corr: High positive correlation 0.5**  plot(movies$profit, movies$reviews)  cor(movies$profit, movies$reviews)  **Corr: Moderate positive correlation 0.38**  plot(movies$profit, movies$duration)  cor(movies$profit, movies$duration)  **Corr: Very less positive correlation 0.12** | 3 |
| 5. | Subset the data of comedy and crime movies of year 2004 and 2005. | movies\_subset = movies[((movies$genre == 'Comedy' | movies$genre == 'Crime') & (movies$year == 2004 | movies$year == 2005)),] | 3 |
| 6. | Is there a relationship exists between genre and year. | table(movies$genre, movies$year)  prop.table(table(movies$genre, movies$year))  table(movies\_subset$genre, movies\_subset$year)  prop.table(table(movies\_subset$genre, movies\_subset$year)) | 2 |
| 7. | **Conclusion** | | 5 |
|  | We examined the data of **movies** dataset. There were about **2961** observations and **11** variables. The following variables that are examined for linear relationship  a. duration  b. Facebook likes  c. votes  d. reviews  e. rating  f. profit  g. budget  It is found that the movie profit greatly fluctuates based on **votes** and it is found be **0.5** correlation (value).  Analysing the comedy and crime genre of year 2004 and 2005, it seems that the proportion of **Comedy** movies are more in the year **2004**. It is also seen that there is **moderate positive** relationship between genre and year. | | |