Data Analysis (CS40003)

Assignment #2

29th October, 2017

Submitted by:
Manish Agrawal
15IM10032

Project Assignment 2

Topic 1

Reference: MOVIE data with 5043 observations

a) Calculate population mean from all the movies up to 2015 on imdb_score:

Since there were NaN type values, those were removed and thus the mean imdb_score was found to be **5.971**

b) Collect a sample of all the movies in the year 2016

The movies whose title_year was equal to 2016 was separately stored in the dataframe and whose count is **105**

c) Test the hypothesis that "popularity of films (as imdb score) increases".

To test the hypothesis consider following:

i. Population standard deviation is known.

Using Z-test

ii. Population standard deviation is unknown

Using T-test

Topic 2

Reference: NUTRITION data with 80 observations

a) Decide whether rating is correlated with sugar content in the product

For this the cor() package was used. Relevant codes are written in the R script. The correlation among the rating and sugars was found to be around **-0.7557551**

b) If correlation exist then what type of correlation (i.e. positive, negative, linear, nonlinear). Calculate r2 to support your answer. For non-linearity test you should try with up to 3 degree models.

Yes the correlation exist among them. Since the value is found to be negative we can conclude that when rating increases then sugar content decreases and vice versa

To support our answer, we also found three linear model between rating and powers of sugar content (ie, linear,quadratic and linear). The three models gave a R- square value of **0.5654 , 0.5916** and **0.5867**

The linear equation between the variables are : **Rating = 42.182 - 92.541* sugars**

Topic 3

Reference: SALARY data with 1,48,654 observations

Database contains salary information of different employees in different organisations. It is required to test whether Overtime Pay, Other Pay and benefits altogether increases with Basic Pay for the year 2014.

To show that there exists any relation between the two, the correlation between BasePay and Overtime Pay, Other Pay and benefits was calculated. The value suggested

higher degree of correlation. Thus, linear model between each one of them was made such that there are three separate linear equations. Further a linear model involving all three features was made which is as :

BasePay = 32900 + 0.29* OvertimePay + 0.863* OtherPay + 1.517* Benefits

This suggests that there is a relation between these three such that all three increases if BasePay is increased.

Topic 4

Reference: SNACKS data with 100 observations

a) Find the Spearman correlation matrix of all the ordinal attributes

The Spearman correlation matrix was taken out using the package :
cor(,method = "spearman").
The table is as:

	Liking scores	**Saltiness**	**Sweetness**	**Acidity**	**Crunchiness**
Liking scores	1				
Saltiness	0.27220415	1			
Sweetness	0.02989934	-0.12160131	1		
Acidity	-0.03451333	-0.07715498	-0.22275703	1	
Crunchiness	0.46590771	0.11290395	-0.08096516	0.15770467	1

b) Determine the coefficient of determination (Spearman).

	Liking scores	**Saltiness**	**Sweetness**	**Acidity**	**Crunchiness**
Liking scores	100				
Saltiness	7.409	100			
Sweetness	0.089	1.47	100		
Acidity	0.119	0.595	4.96	100	

Crunchiness	21.7	1.27	0.655	2.48	100

c) Interpret the result from the two tables.

From the two tables the second table can be more valuable to draw out information since the square provides more vivid picture of deviation from the mean. It is appropriate if we have to find how scatter the value is and getting the distance between the data from mean value. The former table gives the direction in which the data is located by taking mean value as reference.

Topic 5

Reference: GAMES data with 16,719 observations

Draw the relevance contingency table to test the hypothesis “action video game is highly rated among teens”. T=teens.(rating column in game data)

The given dataset has a column named Rating which has the category of age groups. Our target are teen which is categorized as “T” in the rating. Thus a separate dataframe having data whose Rating is “T”. Then the contingency table of Genre was calculated which showed the maximum number of counts in “Action” genre about **681**. This clearly shows that there is high dependency of action games on teens.

Topic 6

Reference: STOCK data for the year 2016-2017

For the given data from stock exchange predict the stock value in the month 1/10/2017.

Stocks are to be predicted on the basis of time series model given. The given dataset is of one year. Of which there is holidays for 2 days every week.