MACHINE LEARNING MINOR PROJECT(SEPTEMBER-2020)

ML-MINOR-SEP.

1. What is the average number of words in movie titles between the year 2000-2005?

Solution:

Step1: Import pandas library and read the dataset and store it in a data frame.

Step2: Check whether there are any null values using isnull() function and then drop all the null values using dropna().

Step3: Filter the complete data frame depending the release_year such that release_year must be between 2000 – 2005(release_year >2000 and release_year<2005) and store it in another data frame.

Step4: Add a column (no_of_words) to the created data frame and apply the lambda function on the column name(original_title)

LAMBDA FUNCTION:

Df2['no_of_words'] = df2.original_title.apply(lambda x: len(x.split()))

Step5: Apply sum function on the df2 and store it in the variable number_of_words .

Step6: Divide the number_of_words by length of df2 and store it in average_no_of_words.

Step7: Print the average_no_of_words.

2. What is the average runtime of movies in the year 2006?

Solution:
Step1: Import pandas library and read the dataset and store it in a data frame.

Step2: Firstly, filter by the release_year and store it in df_2006.

Step3: Check for the null values and drop them using dropna() function.

Step4: Filter the df_2006 by runtime and store in df_2006_runtime.

Step5: Apply describe() method on df_2006_runtime to know whether there are any outliers or not.

Step6: Since there is a huge difference between the 50%,75% and max value we can conclude that there are outliers.

Step7: To remove these outliers I used Inter-Quartile method(IQR).

Step8: To apply IQR we have to sort the dataframe.

Step9: Import numpy and use the percentile function(q1,q3 = np.percentile(sorted_df,[25,75]).

Step10: Apply formulas IQR = q3-q1, lower_bound=q1-(1.5*IQR), upper_bound = q3+(1.5*IQR).

Step11: We can conclude that the values below lower_bound and values above upper_bound are outliers. Hence we select the data frame which has no outliers.

Step12: Store sum in a variable by applying the function sum() of data frame.

Step13: Store the number of rows by applying the shape method on data frame.

Step14: Print the average by dividing the sum and number of rows.

3. Which are the movies with most and least earned revenue?

Solution:
Step1: Import pandas library and read the dataset and store it in a data frame.

Step2: Firstly, filter by the revenue and store it in df_revenue.

Step3: Check for null values in df_revenue and drop them using dropna() function.

Step4: Apply describe() method and check if outliers are present or not.

Step5: Since there is a huge difference between the 50%,75% and max value we can conclude that there are outliers.

Step6: To remove these outliers present I used Inter-Quartile method(IQR).

Step7: To apply IQR we have to sort the data frame.

Step8: Import numpy and use the percentile function(q1,q3 = np.percentile(sorted_df,[25,75]).

Step9: Apply formulas IQR = q3-q1, lower_bound=q1-(1.5*IQR), upper_bound = q3+(1.5*IQR).

Step10: We can conclude that the values below lower_bound and values above upper_bound are outliers. Hence we select the data frame which has no outliers.

Step11: Applying df[0:1] gives the lowest budget and df[4273:4274] gives the highest budget.

Step12: Now using the row indexes obtained above pass those indexes to df_original_title then we get the names of movies with third highest and third lowest budget.

4. Which are the movies with the third-lowest and third-highest budget?

Solution:

Step1: Import pandas library and read the dataset and store it in a data frame.

Step2: Filter the data frame based on the column original_title and store it in df_original_title and also filter by budget and store it in df_budget.

Step3: Check for null values using isnull() function and remove null values using dropna() function.

Step4: Apply describe() function to check if the outliers are present or not.

Step5: Since there is a huge difference between the 50%,75% and max value we can conclude that there are outliers.

Step6: To remove these outliers present I used Inter-Quartile method(IQR).

Step7: To apply IQR we have to sort the data frame.

Step8: Import numpy and use the percentile function(q1,q3 = np.percentile(sorted_df,[25,75]).

Step9: Apply formulas IQR = q3-q1, lower_bound=q1-(1.5*IQR), upper_bound = q3+(1.5*IQR).

Step10: We can conclude that the values below lower_bound and values above upper_bound are outliers. Hence we select the data frame which has no outliers.

Step11: Applying df[2:3] gives the third lowest budget and df[4740:4741] gives the third highest budget.

Step12: Now using the row indexes obtained above pass those indexes to df_original_title then we get the names of movies with third highest and third lowest budget.

By:
Annamaneni Ajay
Gmail: ajayannamaneni20@gmail.com