MACHINE LEARNING WITH PYTHON

MINOR PROJECT

(ML-MINOR-SEP)

NAME – MANSI SHARMA

THE FOLLOWING IS MY PYTHON DATA CODE SCRIPT –

(I attached the code here itself for easy reference)

```
import pandas as pd
import numpy as np
csv_file = pd.read_csv(r"C:\Users\Mansi Sharma\Downloads\tmdb-movies.csv")
missing_values_count = csv_file.isnull().sum()
total_missing = missing_values_count.sum()
#csv_file.info()

drop_csv_file =
["imdb_id","popularity","homepage","vote_count","vote_average","director","overview","tagline","prod
uction_companies"]
csv_file = csv_file.drop(drop_csv_file , axis = 1)
csv_file = csv_file.drop_duplicates(keep = 'first')
csv_file.dropna(how = "all", inplace = True)
csv_file.drop(csv_file.loc[csv_file['budget']==0].index,inplace= True)

csv_file["cast"].fillna('empty',inplace = True)
csv_file["genres"].fillna('empty', inplace= True)
csv_file["keywords"].fillna('empty', inplace= True)
#csv_file.shape
#csv_file.info()
#1st question
def top_3(col_name,size = 3):
    csv_file_sort = pd.DataFrame(csv_file[col_name].sort_values(ascending= True))[:size]
    csv_file_sort['original_title'] = csv_file['original_title']
    print(csv_file_sort)
def bot_3(col_name,size = 3):
    csv_file_sort2 = pd.DataFrame(csv_file[col_name].sort_values(ascending= False))[:size]
    csv_file_sort2['original_title'] = csv_file['original_title']
    print(csv_file_sort2)
```

```
top_3('budget')
bot_3('budget')

#3rd question
csv_file_split_genre = csv_file.copy()
split_genre = csv_file_split_genre['genres'].str.split('|').apply(pd.Series , 1).stack().reset_index(level = 1, drop =
True)
split_genre.name = 'genre_split'
csv_file_split_genre = csv_file_split_genre.drop(['genres'], axis = 1).join(split_genre)
genres_cast = csv_file_split_genre.groupby(['cast'])['genre_split'].value_counts()
print(genres_cast.groupby(level = 0).nlargest(2).reset_index(level = 0 , drop = True))
#4th question
def find_min_max(col_name):
    min_index = csv_file[col_name].idxmin()
    max_index = csv_file[col_name].idxmax()
    low = pd.DataFrame(csv_file.loc[min_index,:])
    high = pd.DataFrame(csv_file.loc[max_index,:])

    print('Movie which has the highest' +col_name+ ':', csv_file['original_title'][max_index])
    print('Movie which has the lowest' + col_name + ':', csv_file['original_title'][min_index])
    return pd.concat([high,low], axis = 1)
find_min_max('revenue')

#5th question
csv_file_2006 = csv_file[csv_file['release_year'] == 2006]
no_of_rows = csv_file_2006.shape[0]
print(csv_file_2006['runtime'].sum()/no_of_rows)
```

According to the cleaning I performed on my data , I found the following answers to the asked questions -

1)Which are the movies with the third-lowest and third highest budget ?

-> Third lowest – Love , Wedding , Marriage

Third highest – Pirates of the Caribbean : At World's End

3)What is the most common genre for Vin Diesel and Emma Watson movies?

-> For Vin Diesel – Action

Emma Watson - Drama

4)Which are the movies with most and least earned revenue ?

->Most earned revenue – Avatar

Least earned revenue – Wild Card

5)What is the average runtime of movies in the year 2006 ?

-> Average runtime for year 2006 was found to be 105.32 (time units)