SVKM'S NMIMS

Mukesh Patel School of Technology Management & Engineering

Department of Mechatronics

Subject- Python Programming

Project Report

Mentor- Prof. Ainal Abdul Azeez

Name: Jheel Kamdar
Roll No.: H030
Sap ID: 70041019034
Branch: B.Tech Mechatronics Semester V
Email: jheel.kamdar34@nmims.edu.in

Abstract:

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device.
You can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.
Netflix content varies by region and may change over time. You can watch from a wide variety of award-winning Netflix Originals, TV shows, movies, documentaries, and more.
Each Netflix plan determines the number of devices you can watch Netflix on at the same time and if you prefer to view in Standard Definition (SD), High Definition (HD), or Ultra High Definition (UHD).

Index

Aim:

To analyze your personal Netflix data using Python programming

About the Project:

'How much time have I spent watching a show on Netflix?'
That's a question that has run through my head repeatedly over the past one year. Due to the ongoing pandemic, Netflix was literally a life savior. However, the data Netflix allowed users to download about their activity was extremely limited. Now, though, Netflix allows you to download a veritable treasure-trove of data about your account. With a just Python and pandas programming, we can now get a concrete answer to the question: how much time have I spent watching a show on Netflix?
During the pandemic, I was hooked to a show called 'The Vampire Diaries'. I was so addicted to the show that I used to spend 8 hours a day watching it. So, in this project, I'll be plotting a chart of my viewing habits by day of the week as well as the same data by hour. The two plots will tell you that on which day of the week and during which hours of the day was the show watched the most.

Algorithm:

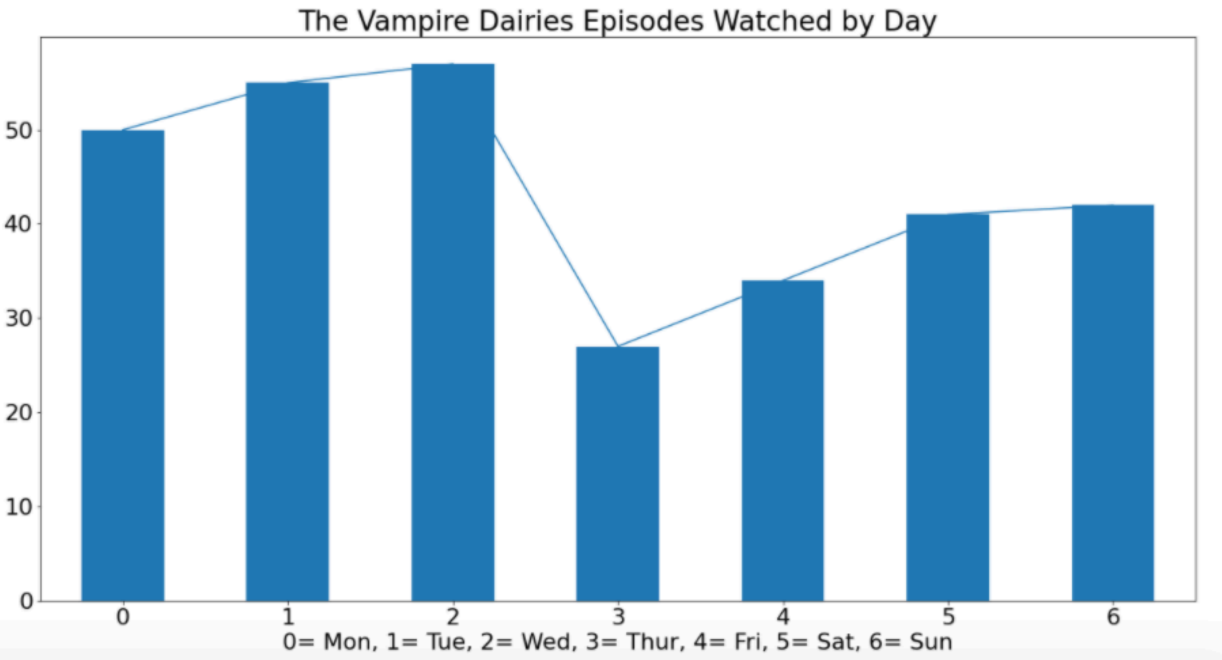
Step 1: Download your Netflix Data (Request a copy of your information from Netflix)
Step 2: Familiarize with the data (Main file needed is ViewingActivity.csv) & load the file into Jupyter notebook.
Step 3: Import the pandas library.
Step 4: Create a pandas framework to make the work easier.
Step 5: Remove unnecessary columns from the .csv file.
Step 6: Convert strings to datetime and timedelta in Pandas.
Step 7: Convert the given data in the IST time zone using datetime.
Step 8: Choose the show you would like to analyse. (Here 'The Vampire Diaries')
Step 9: Filter the string into substring keeping only the rows which contain 'The Vampire Diaries'
Step 10: Filter out short duration (<1 minute) using timedelta.
Step 11: Add up the total duration of the show watched.
Step 12: Import matplotlib
Step 13: Plot a weekly graph showing the maximum number of episodes watched on the weekdays.
Step 14: Plot an hourly graph the maximum number of episodes watched during which hour of the day.

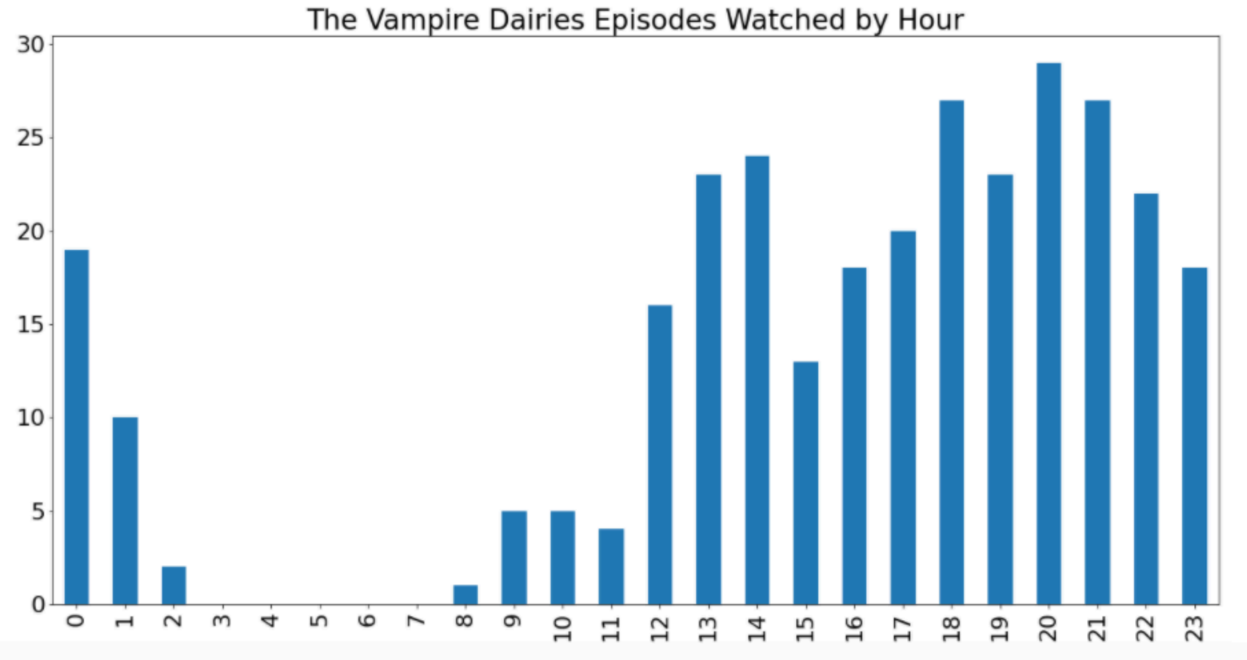
Code:

```
import pandas as pd
data= pd.read_csv('/Users/apple/Downloads/Netflix-report/CONTENT_INTERACTION/
ViewingActivity.csv')
data.shape #checking the total no. of rows and columns in the file
data.head(1) #checking the first row of the file
data = data.drop(['Profile Name', 'Attributes', 'Supplemental Video Type', 'Device Type',
'Bookmark', 'Latest Bookmark', 'Country'], axis=1) #removing the unnecessary columns
data.head(1) #checking if the columns are dropped
data.dtypes #finding the datatypes of the given data

#Convert Start Time to datetime using pandas's pd.to_datetime() (a data and time format pandas can
understand and perform calculations with)
#argument utc=true so that we can confirm that its according to UTC timezone
data['Start Time'] = pd.to_datetime(data['Start Time'], utc=True)
data.dtypes

data = data.set_index('Start Time') # setting our Start Time column as the index using set_index()
data.index = data.index.tz_convert('Asia/Kolkata') #convert from UTC timezone to Indian time
data = data.reset_index() # reset the index so that Start Time becomes a column again
data.head(1) #checking that it worked

#Convert Duration to timedelta (a time duration format pandas can understand and perform
calculations with)
data['Duration'] = pd.to_timedelta(data['Duration'])
data.dtypes #checking that it worked

# create a new dataframe called tvd that that takes from data
# only the rows in which the Title column contains 'The Vampire Diaries'
tvd = data[data['Title'].str.contains('The Vampire Diaries')]
tvd.shape #finding the no. of rows and columns of this particular show
tvd = tvd[(tvd['Duration'] > '0 days 00:01:00')] #removing the autoplayed episodes
tvd.shape #finding the total no. of rows and columns of this particular show
tvd['Duration'].sum() #adding the total no. of hours watched

#creating new columns for "weekday" and "hour"
tvd['weekday'] = tvd['Start Time'].dt.weekday
tvd['hour'] = tvd['Start Time'].dt.hour
tvd.head(1) # check to make sure the columns were added correctly


%matplotlib inline
import matplotlib
# set our categorical and define the order so the days are plotted Monday-Sunday
tvd['weekday'] = pd.Categorical(tvd['weekday'], categories= [0,1,2,3,4,5,6], ordered=True)
# create tvd_by_day and count the rows for each weekday, assigning the result to that variable
tvd_by_day = tvd['weekday'].value_counts()
# sort the index using our categorical, so that Monday (0) is first, Tuesday (1) is second, and so on.
tvd_by_day = tvd_by_day.sort_index()
# making the font size to make it a bit larger and easier to read
```

```
matplotlib.rcParams.update({'font.size': 22})
# plot tvd_by_day as a bar chart with the listed size and title
tvd_by_day.plot(kind='bar', figsize=(20,10), title='The Vampire Dairies Episodes Watched by Day')
tvd_by_day.plot(xlabel= '0= Mon, 1= Tue, 2= Wed, 3= Thur, 4= Fri, 5= Sat, 6= Sun')

# set our categorical and define the order so the hours are plotted 0-23
tvd['hour'] = pd.Categorical(tvd['hour'], categories=
[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19, 20,21,22,23], ordered=True)
# create tvd_by_hour and count the rows for each hour, assigning the result to that variable
tvd_by_hour = tvd['hour'].value_counts()
# sort the index using our categorical, so that midnight (0) is first, 1 a.m. (1) is second, and so on.
tvd_by_hour = tvd_by_hour.sort_index()
# plot tvd_by_hour as a bar chart with the listed size and title
tvd_by_hour.plot(kind='bar', figsize=(20,10), title='The Vampire Dairies Episodes Watched by Hour')
```

Output:





Conclusion:

I have successfully managed to analyse my personal Netflix data using the pandas library in Python programming. It has been a great learning and I look forward to doing more such projects.