

**EDA ANALYSIS**

A Minor Project work report Submitted to Verzeo

SUBMITTED BY: **SHRUTI KATHURIA**

EMAIL: [**SHRUTIKATHURIA2001@GMAIL.COM**](mailto:SHRUTIKATHURIA2001@GMAIL.COM)

**Computer engineering**

**Thapar Institute of Engineering and Technology**

**EDA (EXPLORATORY DATA ANALYSIS)**

Firstly, the movie data set that has to go through three different steps:

1)Understanding the data

2)Cleaning of the data

3)Analysis between the variables

UNDERSTANDING AND CLEANING OF DATA

●import pandas as pd

import numpy as np

import seaborn as sns

I started with importing the libraries of python to use them later in the EDA.

●data = pd.read\_csv("C:/Users/SHRUTI KATHURIA/Desktop/Verzeo Project/Minor/tmbd.csv")

I did read the excel sheet and I imported it as a csv file in the Jupiter notebook. And I stored the csv in a variable named

●data.head()

data.tail()

With that part as I wanted to display the number of row and columns at the start and the end, I used two functions.

● data.shape

Used to know the shape of the dataset that is given

●data.describe()

used for calculating some statistical data like percentile, mean and std of the numerical

values of the Series or DataFrame

● data.columns

Used for displaying all columns in the dataset.

● data.nunique()

To check Unique values we used this command.

● data.isnull().sum()

To check null values of the dataset.

● data.mean()

data.fillna(data.mean())

data

I tried replacing all the NaN values with their mean

● data.drop(columns=['vote\_count','cast','overview','imdb\_id','revenue\_adj','homepage','tagline','budget\_adj','keywords'],inplace=True,axis=0)

As I saw there were columns which weren’t important for the analysis soo I dropped those to make it easier for the further process.

● sum(data.duplicated())

I checked for the duplicated columns in the data set that was given.

● data.drop\_duplicates()

I removed the duplicated rows that were present in the data.

● print("Rows With Zero Values In The Budget Column:",data[(data['budget']==0)].shape[0])

print("Rows With Zero Values In The Revenue Column:",data[(data['revenue']==0)].shape[0])

This displayed the number of roes and columns that contain zero.

**ANALYSE THE DATA**

● import seaborn as sns

import matplotlib.pyplot as plt

fig , ax = plt.subplots(figsize=(20,10))

data.groupby('release\_year').count()['id'].plot(kind='bar',fontsize=14,color="#2E86c1")

plt.title('No. of movies released per year', fontsize=19)

plt.xlabel('Years (1967-2015)' , fontsize=17)

plt.ylabel("Number of Movies Released", fontsize=17)

plt.show()

So I plotted a graph between No of movies released and in which years. So as to analyse the data.

In the year 2014 there were maximum number of releases.

● data.hist(color='DarkBlue',figsize= (10,10));

I plotted a histogram and analysed that the graphs for release\_year and the vote\_average was different as compare to others.

● import seaborn as sns

import matplotlib.pyplot as plt

sns.distplot(data['budget'],kde=False,hist=True,bins=11,hist\_kws=dict(edgecolor="k", linewidth=1))

sns.distplot(data['revenue'],kde=False,bins=[0, 5, 10, 15, 20, 25, 30],hist\_kws=dict(edgecolor="g", linewidth=1),color='Green')

Again I plotted graphs for Revenue and budget and they seemed quite similar.

● sns.lmplot('vote\_average', 'revenue', data=data)

● sns.boxplot(x=data['budget'])

To see if there any different points which exists I plotted a boxplot. Hence I saw two differed a lot.

**Questions**

**Q1) Which are the movies with the third lowest and third highest budget?**

data.groupby('release\_year').count()['id'].plot(xticks = np.arange(1960,2016,5))

sns.set(rc={'figure.figsize':(10,5)})

plt.title("Year Vs Number Of Movies",fontsize = 14)

plt.xlabel('Release year',fontsize = 13)

plt.ylabel('Number Of Movies',fontsize = 13)

sns.set\_style("whitegrid")

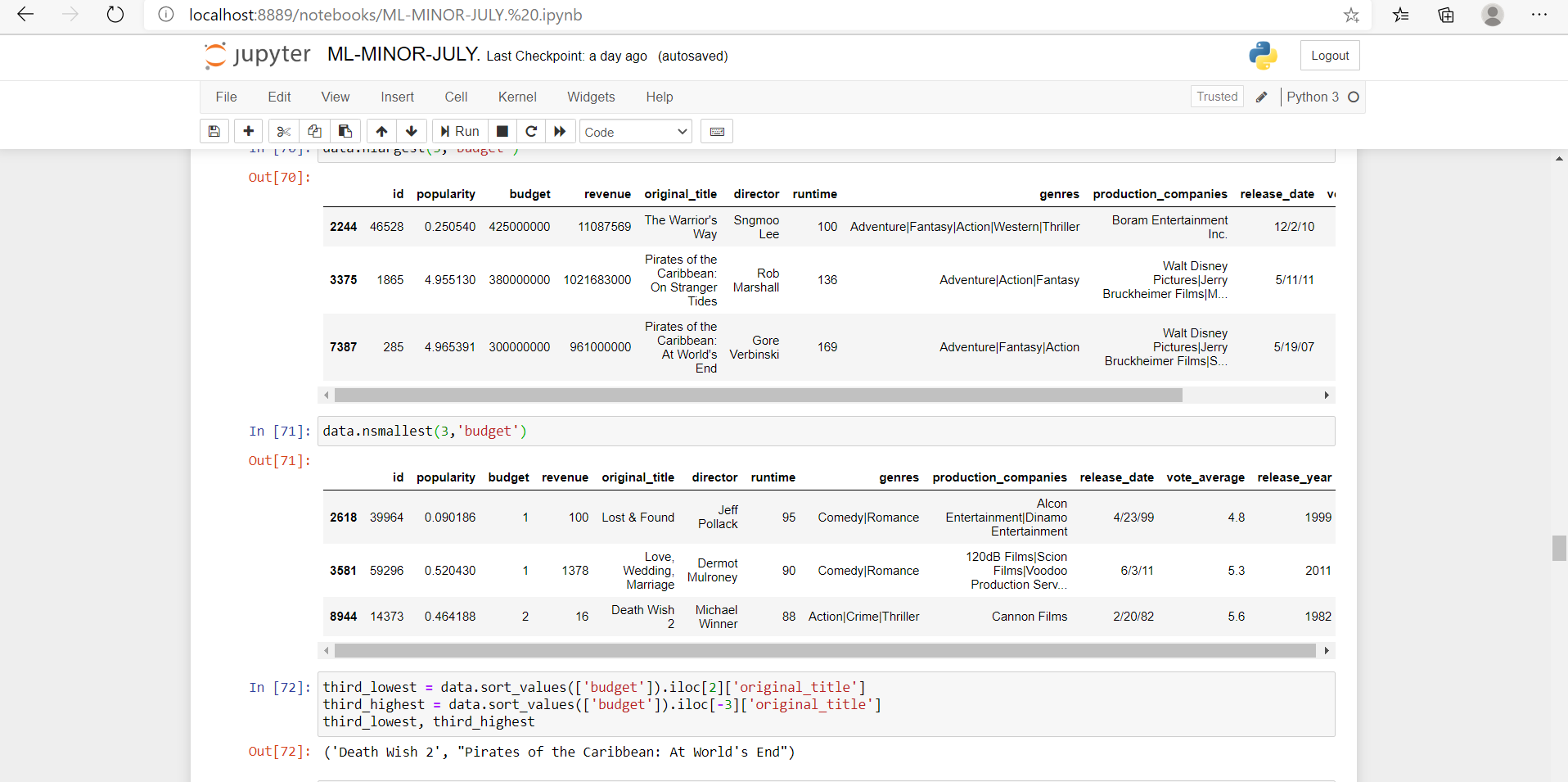
data.nlargest(3,'budget')

data.nsmallest(3,'budget')

third\_lowest = data.sort\_values(['budget']).iloc[2]['original\_title']

third\_highest = data.sort\_values(['budget']).iloc[-3]['original\_title']

third\_lowest, third\_highest



**Q2) What is the average number of words in movie titles between the year 2000-2005?**

import pandas as pd

data = pd.read\_csv("C:/Users/SHRUTI KATHURIA/Desktop/Verzeo Project/Minor/tmbd.csv")

li=data.loc[2000: 2005, "original\_title"]

data.columns

temp = data.where(data['release\_year'] <= 2005)

temp = temp.where(temp['release\_year'] >= 2000)

temp = temp[temp['release\_year'].notna()]

temp.sort\_values('release\_year')

temp['original\_title']

count = 0

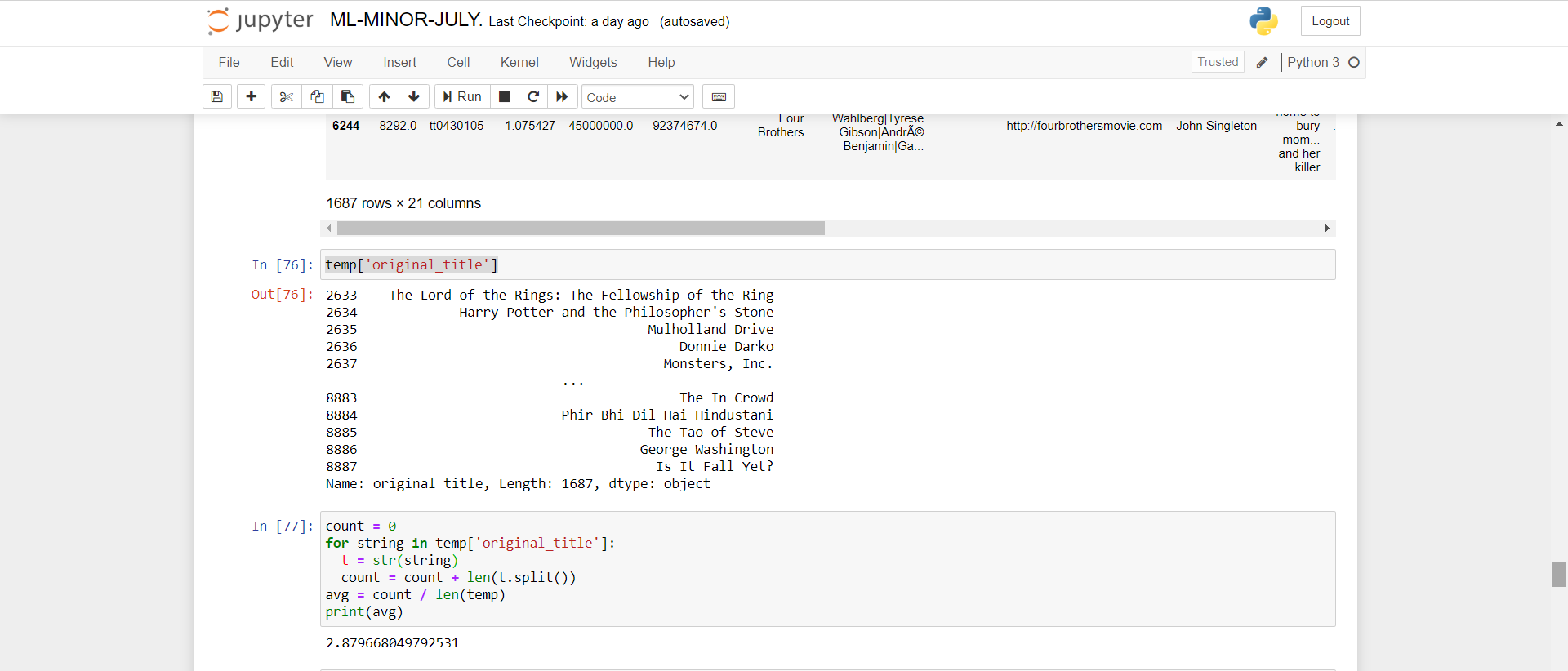
for string in temp['original\_title']:

t = str(string)

count = count + len(t.split())

avg = count / len(temp)

print(avg)



**Q3) What is the most common Genre for Vin Diesel & Emma Watson movies?**

# We need to create two empty Dictionaries

vd={}

em={}

for j in range(int(data.shape[0])):

if type(data.cast[j]) == str:

.

# Movies in which "Vin Diesel" is present

if "Vin Diesel" in data.cast[j] :

if data.genres[j] in vd:

vd[data.genres[j]]+=1

else:

vd[data.genres[j]] = 1

# Movies in which "Emma Watson" is present

if "Emma Watson" in data.cast[j] :

if data.genres[j] in em:

em[data.genres[j]]+=1

else:

em[data.genres[j]] = 1

V={}

# Finding the count of each genre in entire Dataset in which "Vin Diesel" is present

for k,v in vd.items():

# Splitting based on "|" as delimeter

tem= k.split("|")

for ea in tem:

if ea in V:

V[ea]+=1\*int(v)

else:

V[ea]=1\*int(v)

E={}

# Finding the count of each genre in entire Dataset in which "Emma Watson" is present

for k2,v2 in em.items():

# Splitting based on "|" as delimeter

tem2= k2.split("|")

for ea in tem2:

if ea in E:

E[ea]+=1\*int(v2)

else:

E[ea]=1\*int(v2)

In [ ]:

# Finding the Keys with max value of Genre Count

Vmax= max(V, key=V.get)

Emax= max(E, key=E.get)

print(V)

print("The most common Genre for Vin Diesel :",Vmax)

print(E)

print("The most common Genre for Emma Watson :",Emax,"and Adventure")

**Q4)  Which are the movies with most and least earned revenue?**

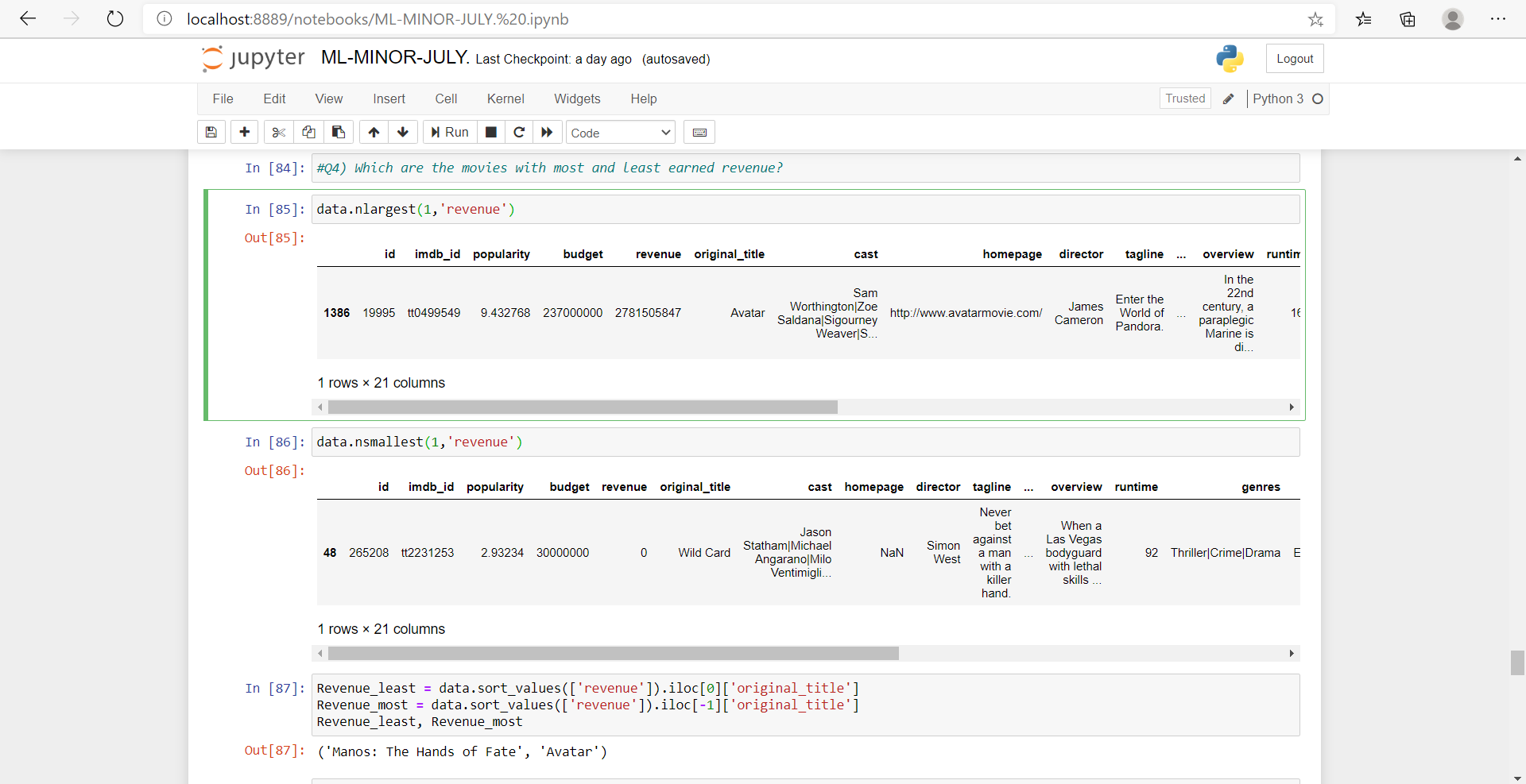
data.nlargest(1,'revenue')

data.nsmallest(1,'revenue')

Revenue\_least = data.sort\_values(['revenue']).iloc[0]['original\_title']

Revenue\_most = data.sort\_values(['revenue']).iloc[-1]['original\_title']

Revenue\_least, Revenue\_most



**Q5) What is the average runtime of movies in the year 2006?**

data[data.release\_year==2006].runtime.mean()

data.groupby('release\_year').mean()['runtime'].plot(xticks = np.arange(1960,2016,5))

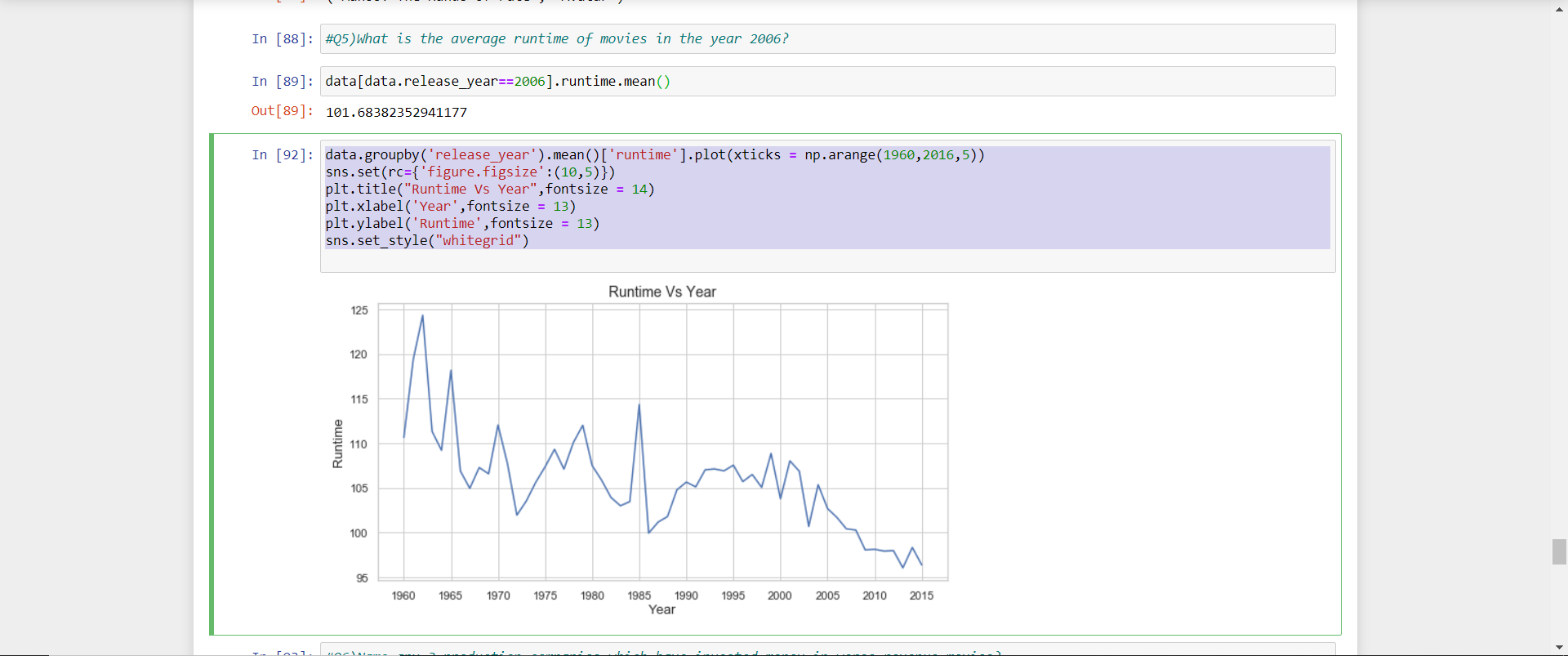
sns.set(rc={'figure.figsize':(10,5)})

plt.title("Runtime Vs Year",fontsize = 14)

plt.xlabel('Year',fontsize = 13)

plt.ylabel('Runtime',fontsize = 13)

sns.set\_style("whitegrid")



**Q6) Name any 3 production companies which have invested money in worse revenue movies?**

data.nsmallest(3,'revenue')

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

data = pd.read\_csv("C:/Users/SHRUTI KATHURIA/Desktop/Verzeo Project/Minor/tmbd.csv")

data.info()

df=data.drop(columns=['id', 'popularity', 'budget', 'original\_title','cast','director', 'keywords', 'runtime', 'genres', 'release\_date','vote\_count', 'vote\_average', 'release\_year'])

df.isnull().sum()

import warnings

warnings.filterwarnings("ignore")

df.dropna(subset=['production\_companies'],inplace=True)

df=df[(df['revenue']>0)]

df.nunique()

# We are calculating the mean revenue of each production company. As there are like x different production companies, we should get x rows.

df\_=df.groupby('production\_companies').mean()

df\_=df\_.sort\_values('revenue')

df\_.head(10)

