**Project on Data Science with Python Programming**

Name of the project:

**Comcast Telecom Consumer Complaints**

**Done by:Mrinmay Saha**

DESCRIPTION

Comcast is an American global telecommunication company. The firm has been providing terrible customer service. They continue to fall short despite repeated promises to improve. Only last month (October 2016) the authority fined them a $2.3 million, after receiving over 1000 consumer complaints.  
The existing database will serve as a repository of public customer complaints filed against Comcast.  
It will help to pin down what is wrong with Comcast's customer service.

**Data Dictionary**

* Ticket #: Ticket number assigned to each complaint
* Customer Complaint: Description of complaint
* Date: Date of complaint
* Time: Time of complaint
* Received Via: Mode of communication of the complaint
* City: Customer city
* State: Customer state
* Zipcode: Customer zip
* Status: Status of complaint
* Filing on behalf of someone

**Analysis Task**

To perform these tasks, you can use any of the different Python libraries such as NumPy, SciPy, Pandas, scikit-learn, matplotlib, and BeautifulSoup.

- Import data into Python environment.  
- Provide the trend chart for the number of complaints at monthly and daily granularity levels.  
- Provide a table with the frequency of complaint types.

* Which complaint types are maximum i.e., around internet, network issues, or across any other domains.

- Create a new categorical variable with value as **Open**and **Closed**. Open & Pending is to be categorized as Open and Closed & Solved is to be categorized as Closed.  
- Provide state wise status of complaints in a stacked bar chart. Use the categorized variable from Q3. Provide insights on:

* Which state has the maximum complaints
* Which state has the highest percentage of unresolved complaints

- Provide the percentage of complaints resolved till date, which were received through the Internet and customer care calls.

The analysis results to be provided with insights wherever applicable.

**Source Code with graphs an insights:**

#importing required libraries

import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

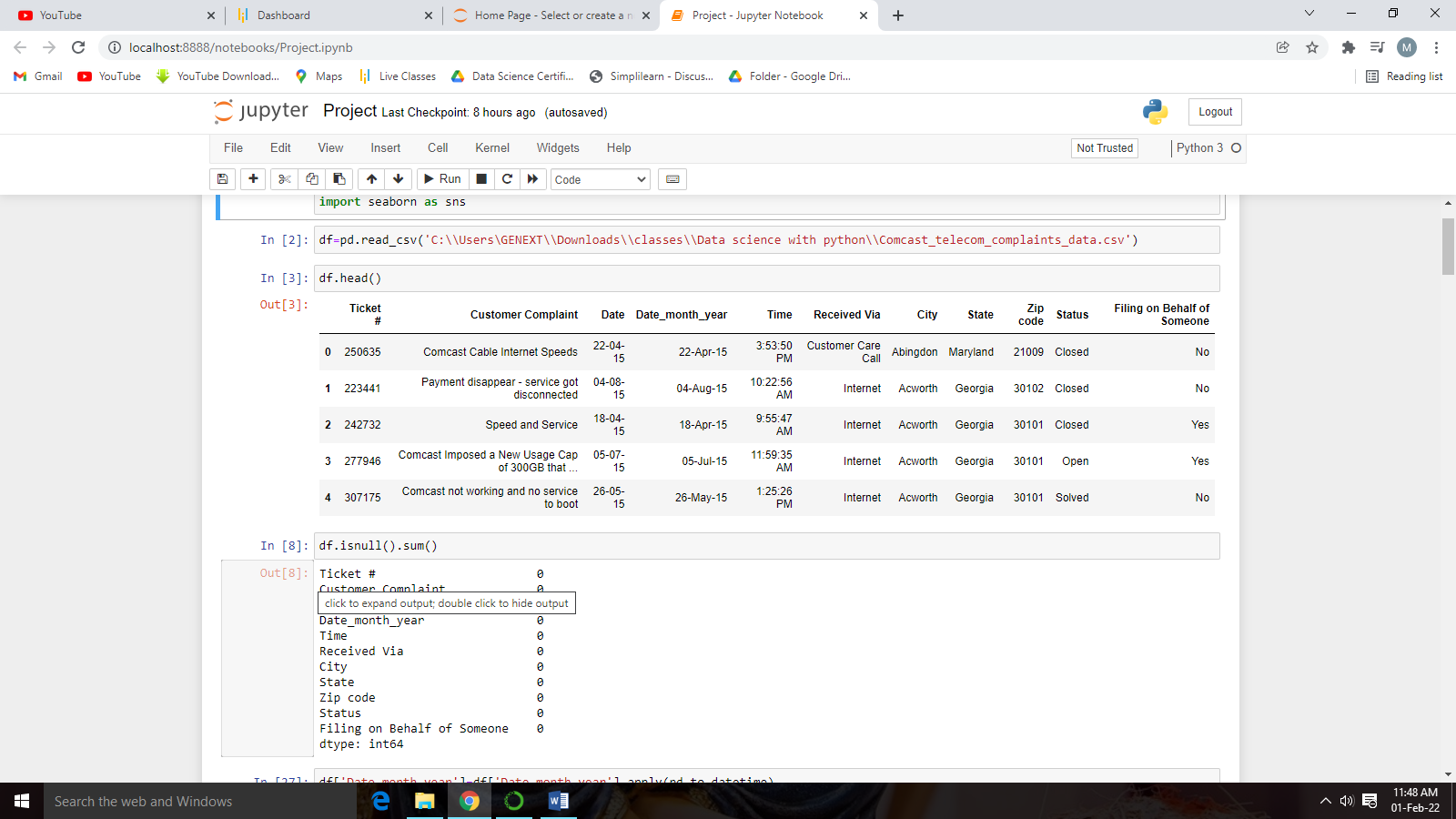
import seaborn as sns

#reading data

df=pd.read\_csv('C:\\Users\GENEXT\\Downloads\\classes\\Data science with python\\Comcast\_telecom\_complaints\_data.csv')

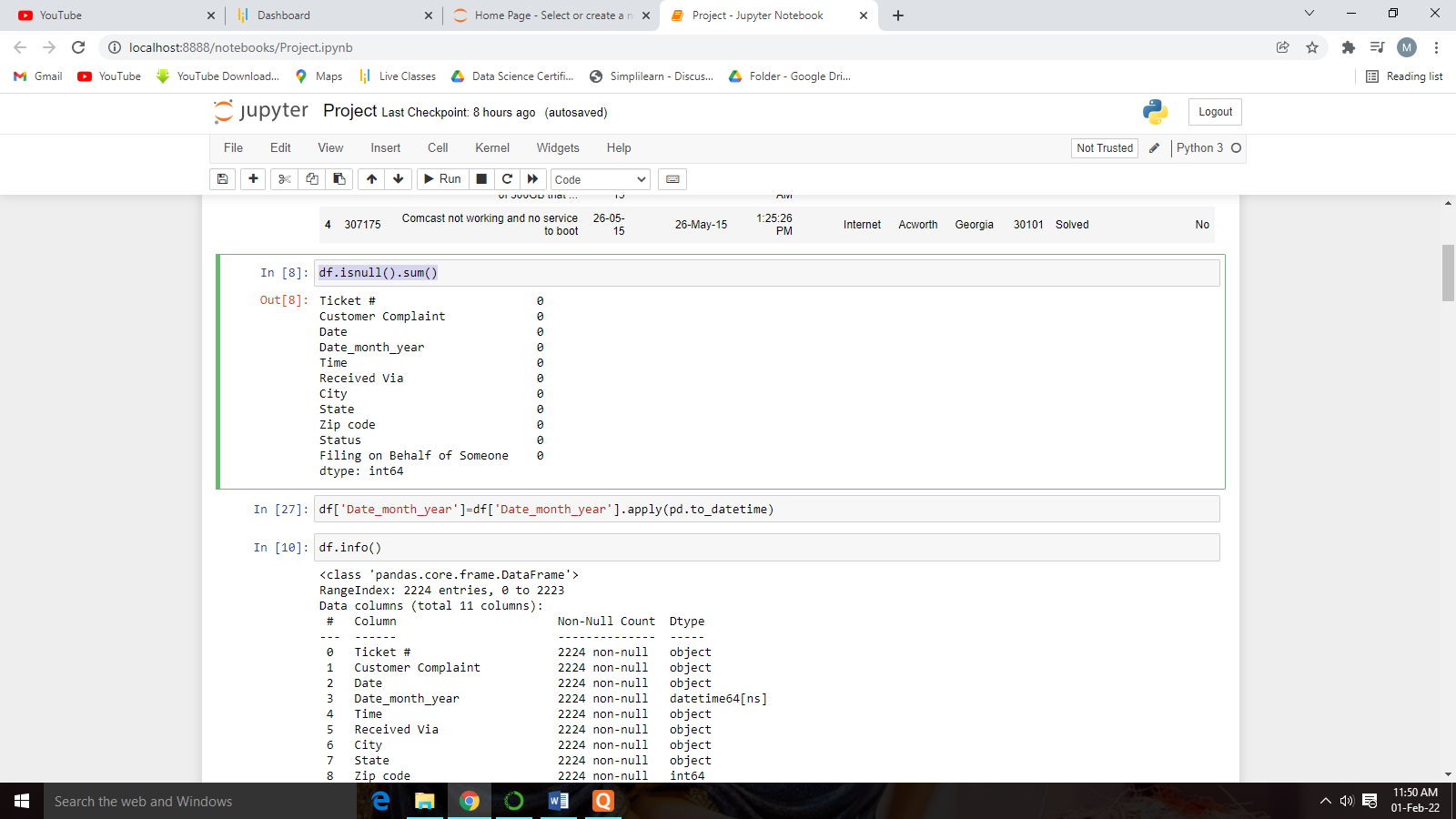
#looking for first 5 rows

df.head()

**Output:**

#Checking for Null values

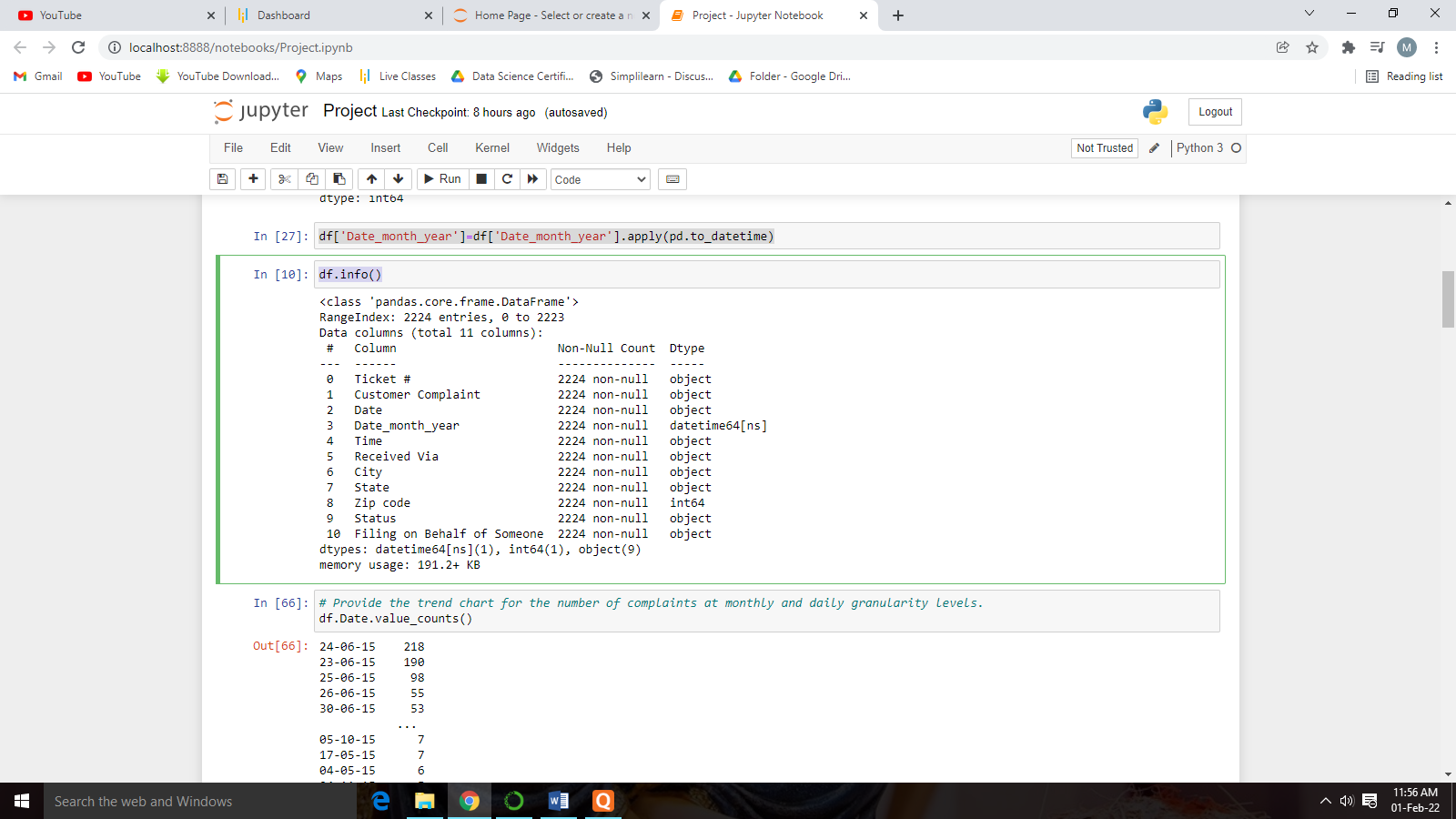
df.isnull().sum()



\*\*No null values found in the Dataset.

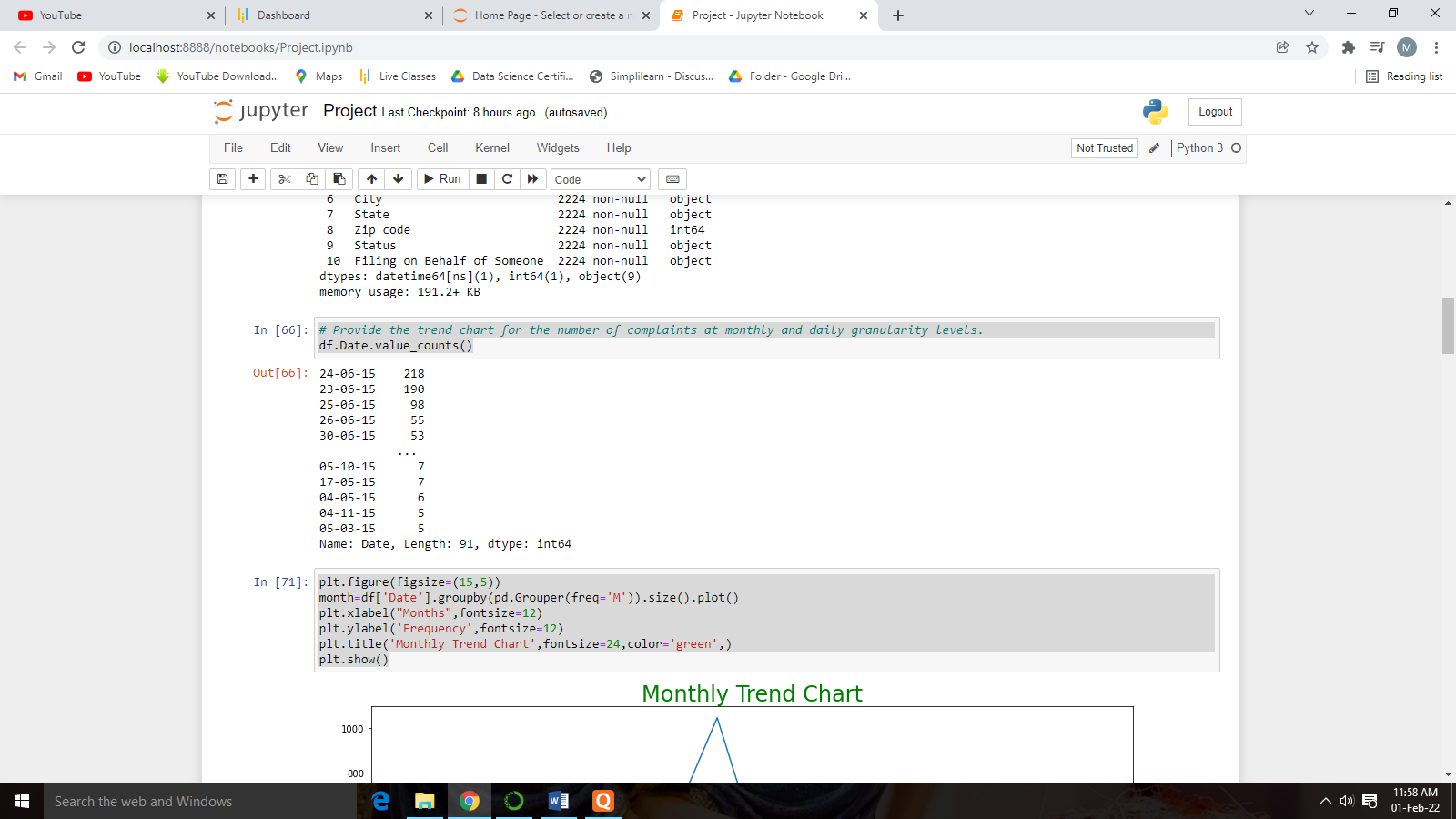
#Converting object type data type to datetime.

df['Date\_month\_year']=df['Date\_month\_year'].apply(pd.to\_datetime)

df.info()

# Provide the trend chart for the number of complaints at monthly and daily granularity levels.

df.Date.value\_counts()

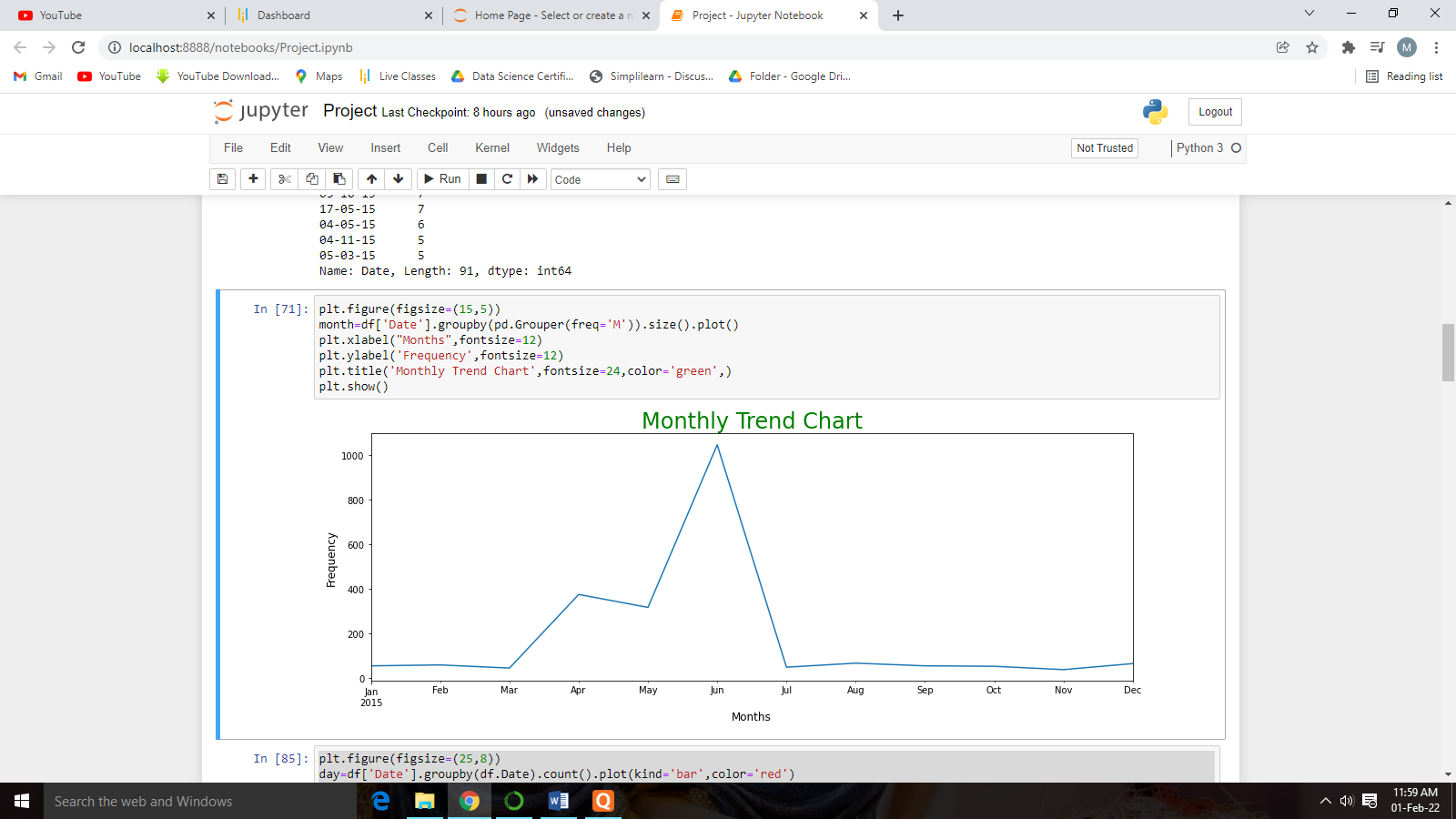


plt.figure(figsize=(15,5))

month=df['Date'].groupby(pd.Grouper(freq='M')).size().plot()

plt.xlabel("Months",fontsize=12)

plt.ylabel('Frequency',fontsize=12)

plt.title('Monthly Trend Chart',fontsize=24,color='green',)

plt.show()

**Insights:** As we can clearly see from the graph that the complaint was increased drastically in the month of June due to some reason.

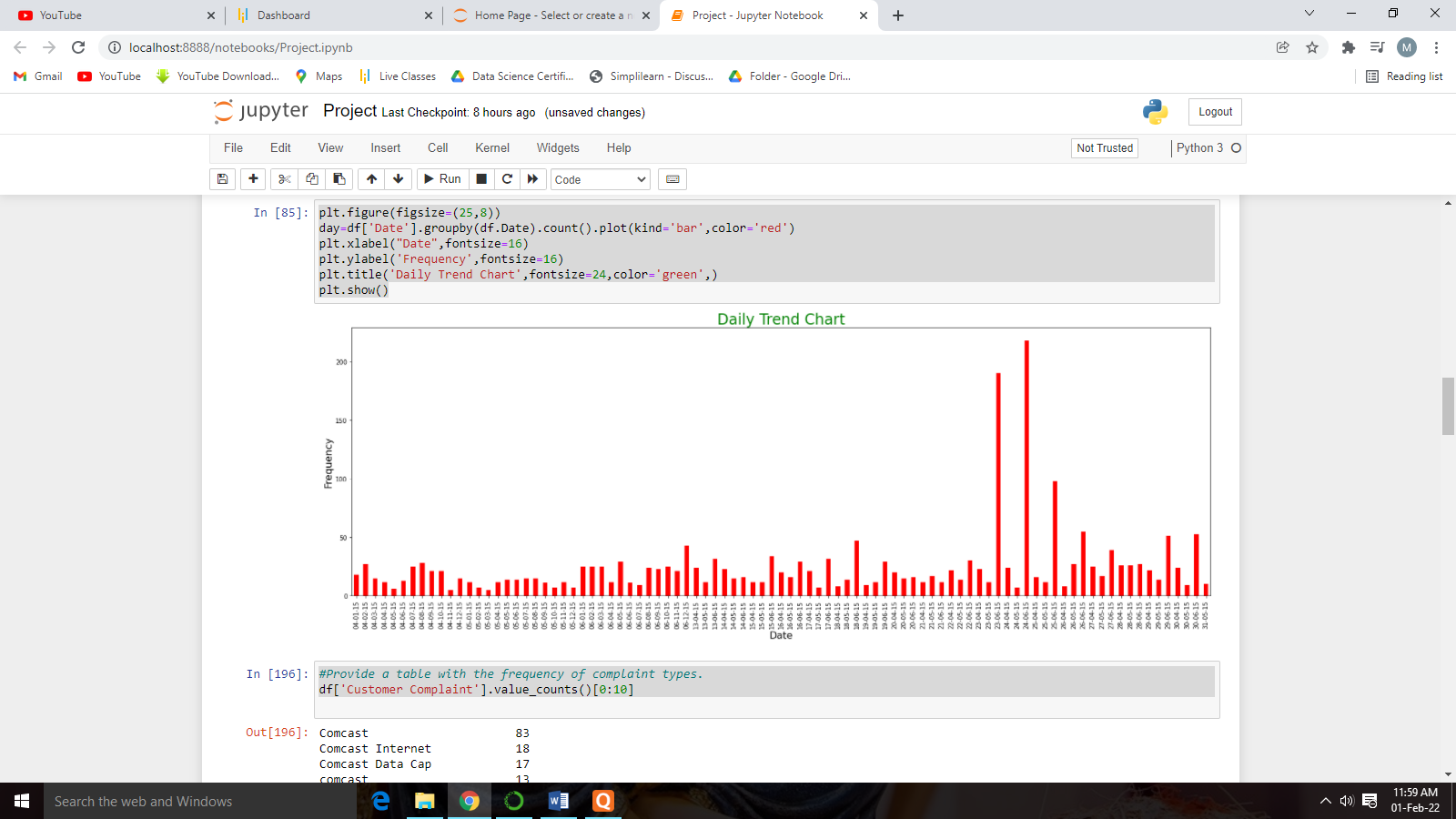
plt.figure(figsize=(25,8))

day=df['Date'].groupby(df.Date).count().plot(kind='bar',color='red')

plt.xlabel("Date",fontsize=16)

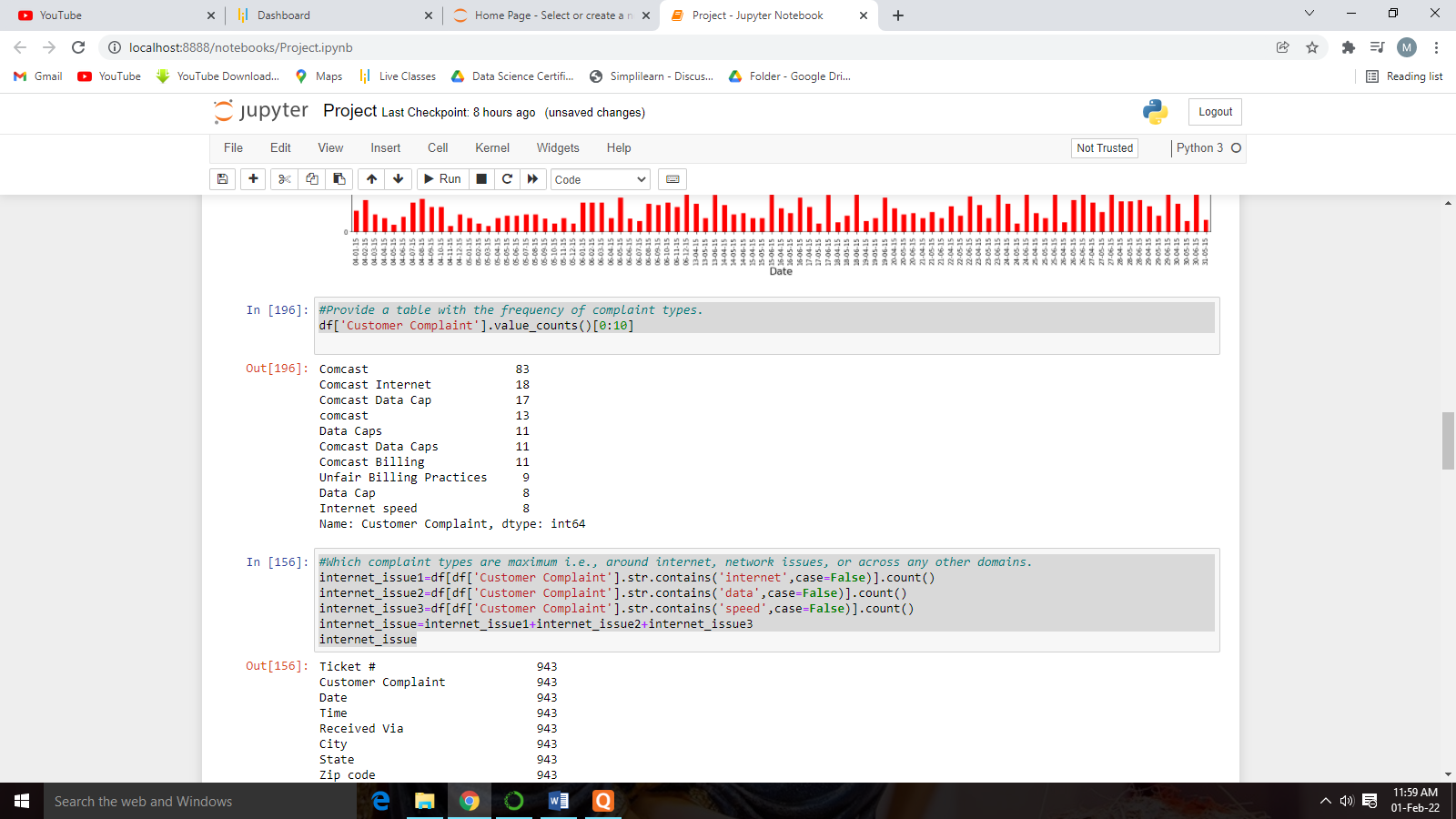
plt.ylabel('Frequency',fontsize=16)

plt.title('Daily Trend Chart',fontsize=24,color='green',)

plt.show()

**Insight:** With the help of the daily report chart it is clearly observed the complaint increased in the second half of the month of June.

#Provide a table with the frequency of complaint types.

df['Customer Complaint'].value\_counts()[0:10]

#Which complaint types are maximum i.e., around internet, network issues, or across any other domains.

internet\_issue1=df[df['Customer Complaint'].str.contains('internet',case=False)].count()

internet\_issue2=df[df['Customer Complaint'].str.contains('data',case=False)].count()

internet\_issue3=df[df['Customer Complaint'].str.contains('speed',case=False)].count()

internet\_issue=internet\_issue1+internet\_issue2+internet\_issue3

internet\_issue

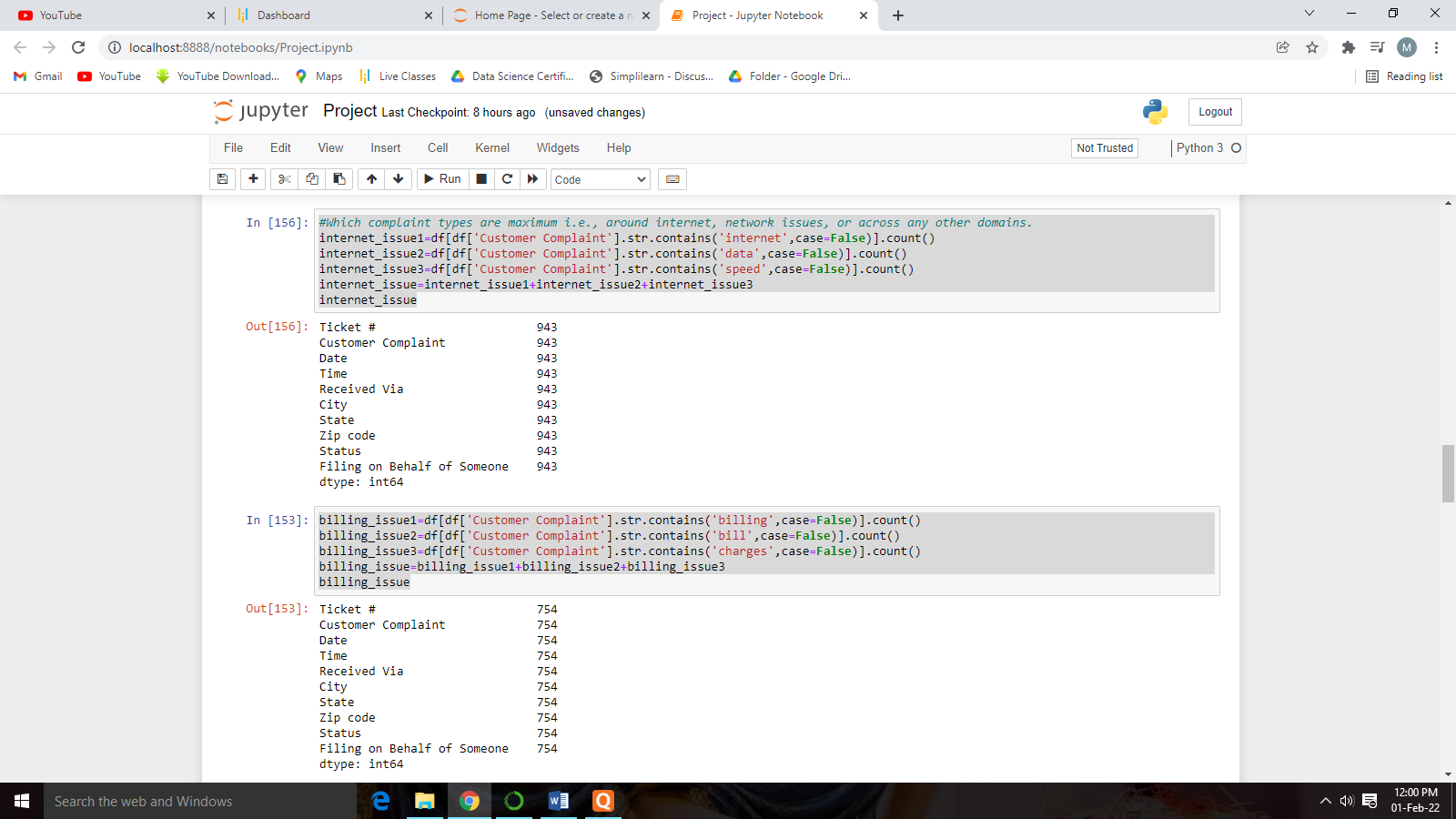
billing\_issue1=df[df['Customer Complaint'].str.contains('billing',case=False)].count()

billing\_issue2=df[df['Customer Complaint'].str.contains('bill',case=False)].count()

billing\_issue3=df[df['Customer Complaint'].str.contains('charges',case=False)].count()

billing\_issue=billing\_issue1+billing\_issue2+billing\_issue3

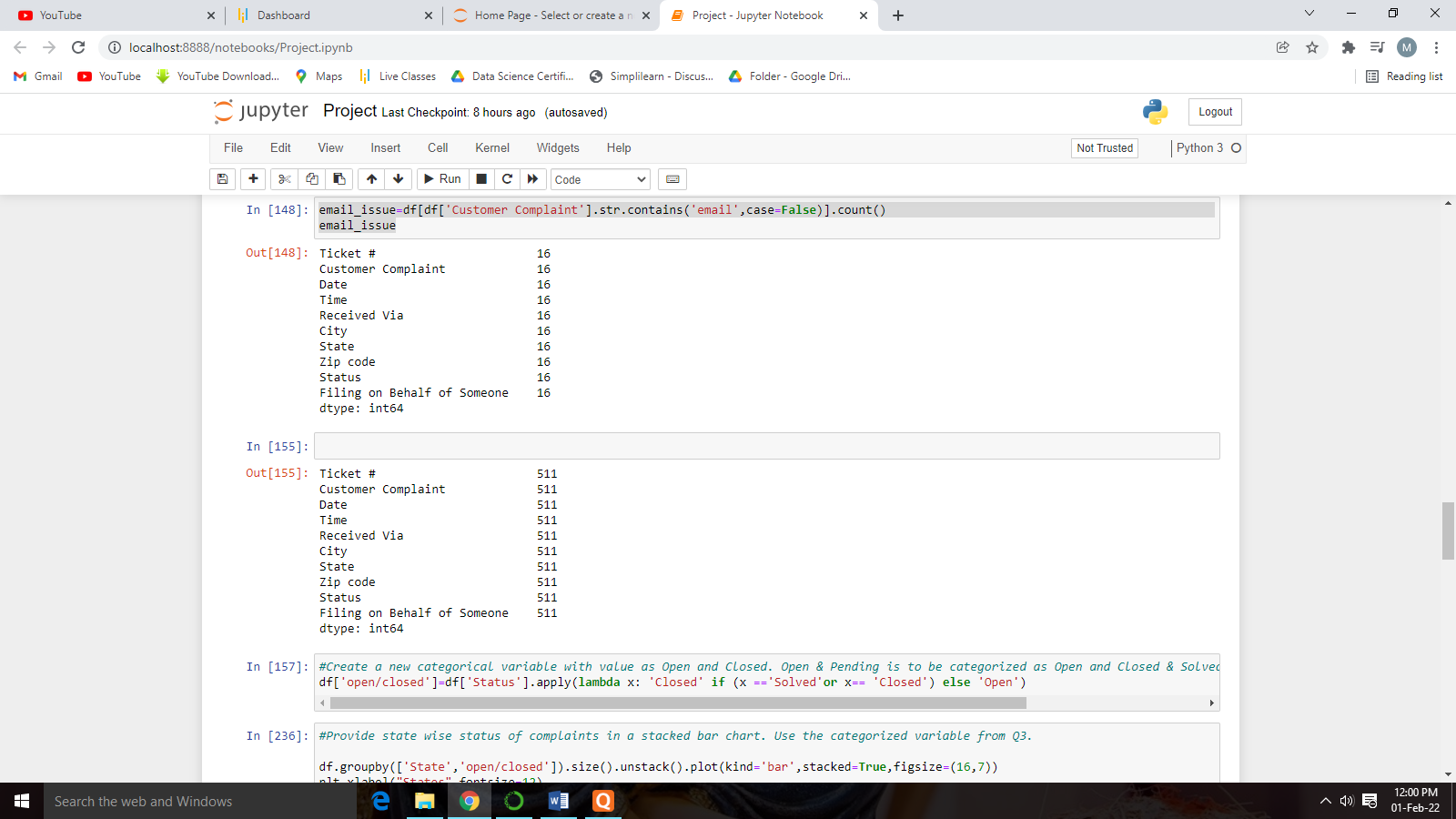
billing\_issue



email\_issue=df[df['Customer Complaint'].str.contains('email',case=False)].count()

email\_issue

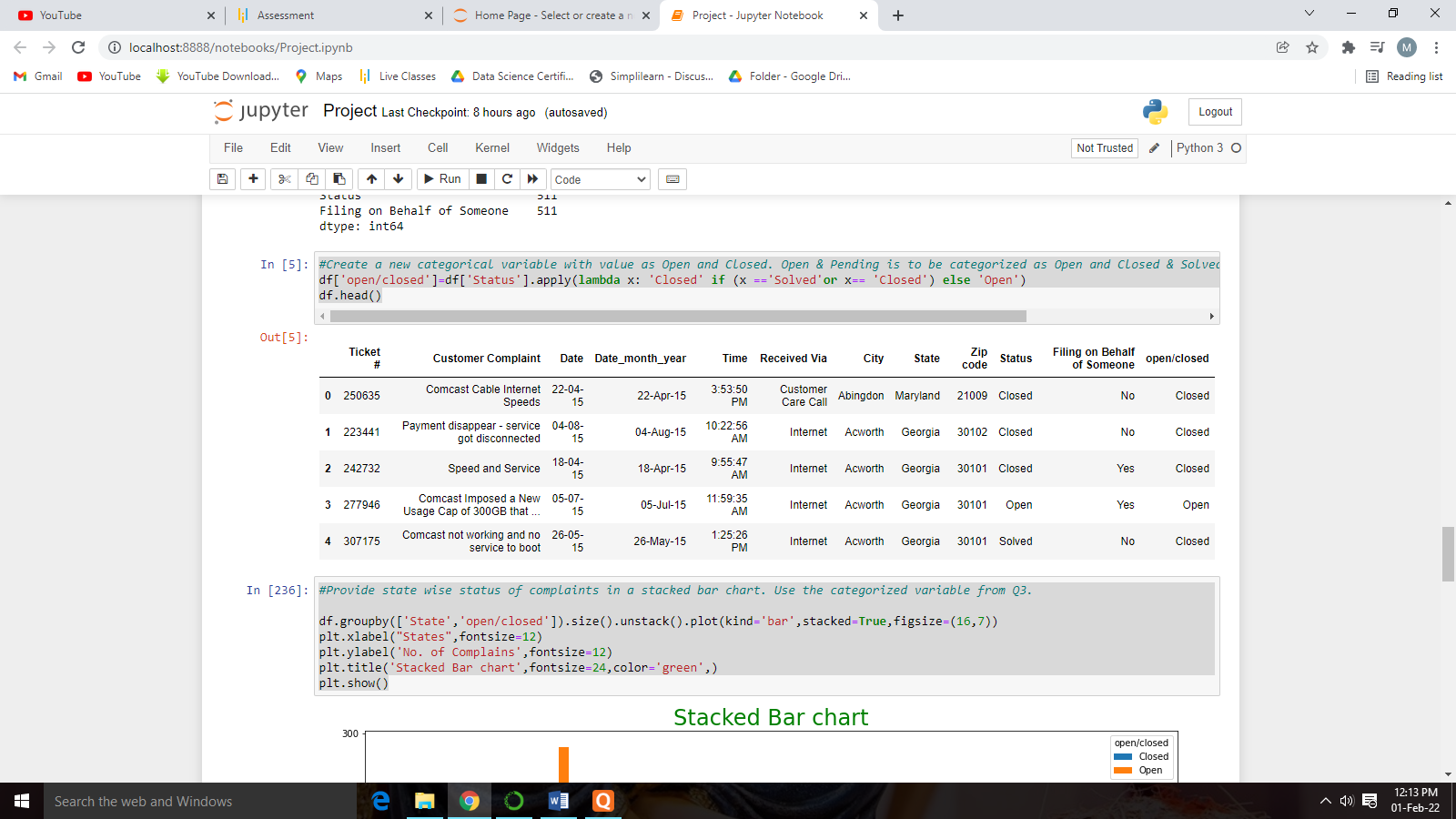
other\_issue=df['Customer Complaint'].count()-internet\_issue-billing\_issue-email\_issue

other\_issue

**Insight:** As we can observe that there are some complaints from different-different categories and we combined them into one, i.e.- others. So most of the complaints are related to Internet issue.

#Create a new categorical variable with value as Open and Closed. Open & Pending is to be categorized as Open and Closed & Solved is to be categorized as Closed.

df['open/closed']=df['Status'].apply(lambda x: 'Closed' if (x =='Solved'or x== 'Closed') else 'Open')

df.head()

#Provide state wise status of complaints in a stacked bar chart. Use the categorized variable from Q3.

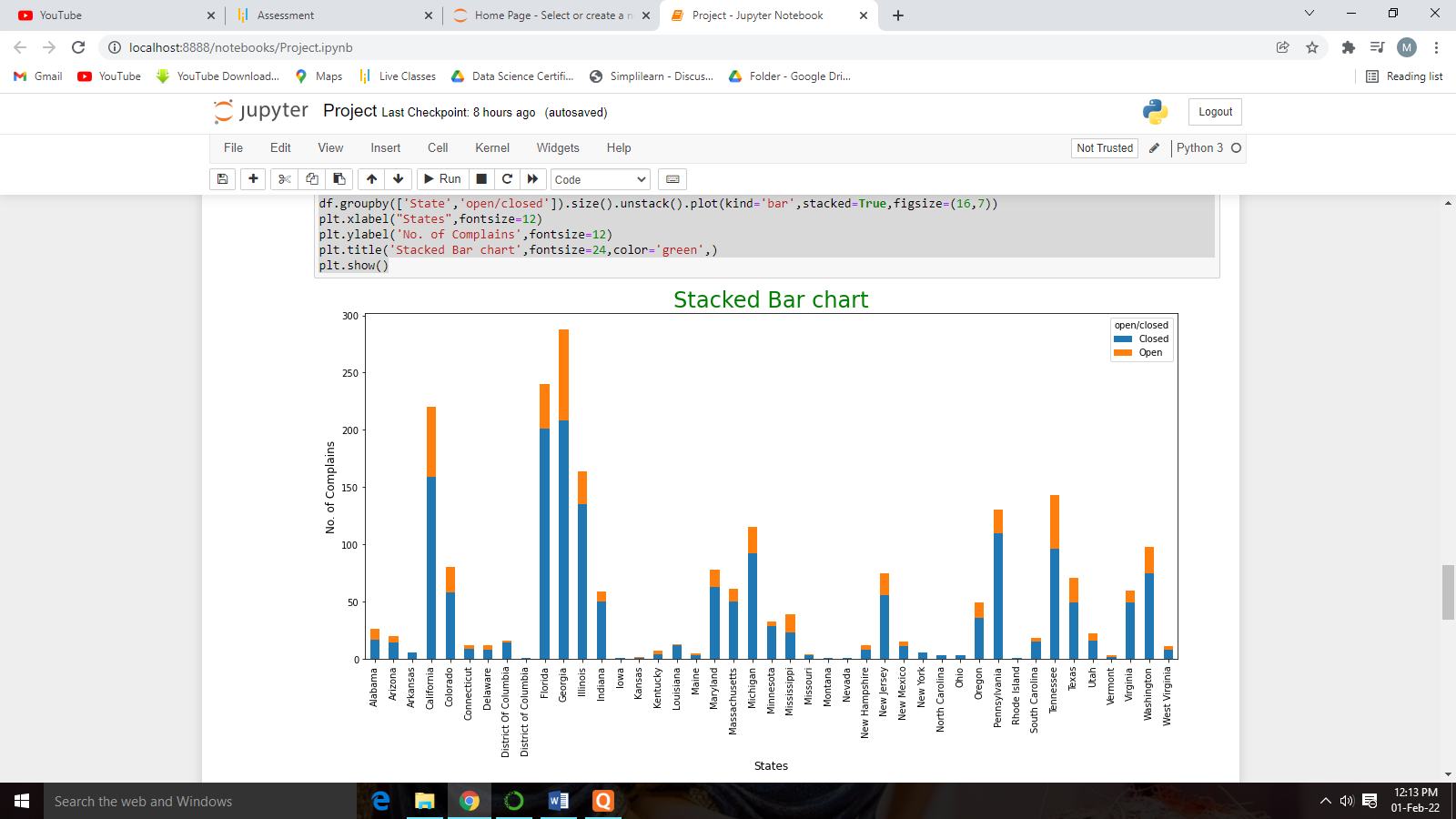
df.groupby(['State','open/closed']).size().unstack().plot(kind='bar',stacked=True,figsize=(16,7))

plt.xlabel("States",fontsize=12)

plt.ylabel('No. of Complains',fontsize=12)

plt.title('Stacked Bar chart',fontsize=24,color='green',)

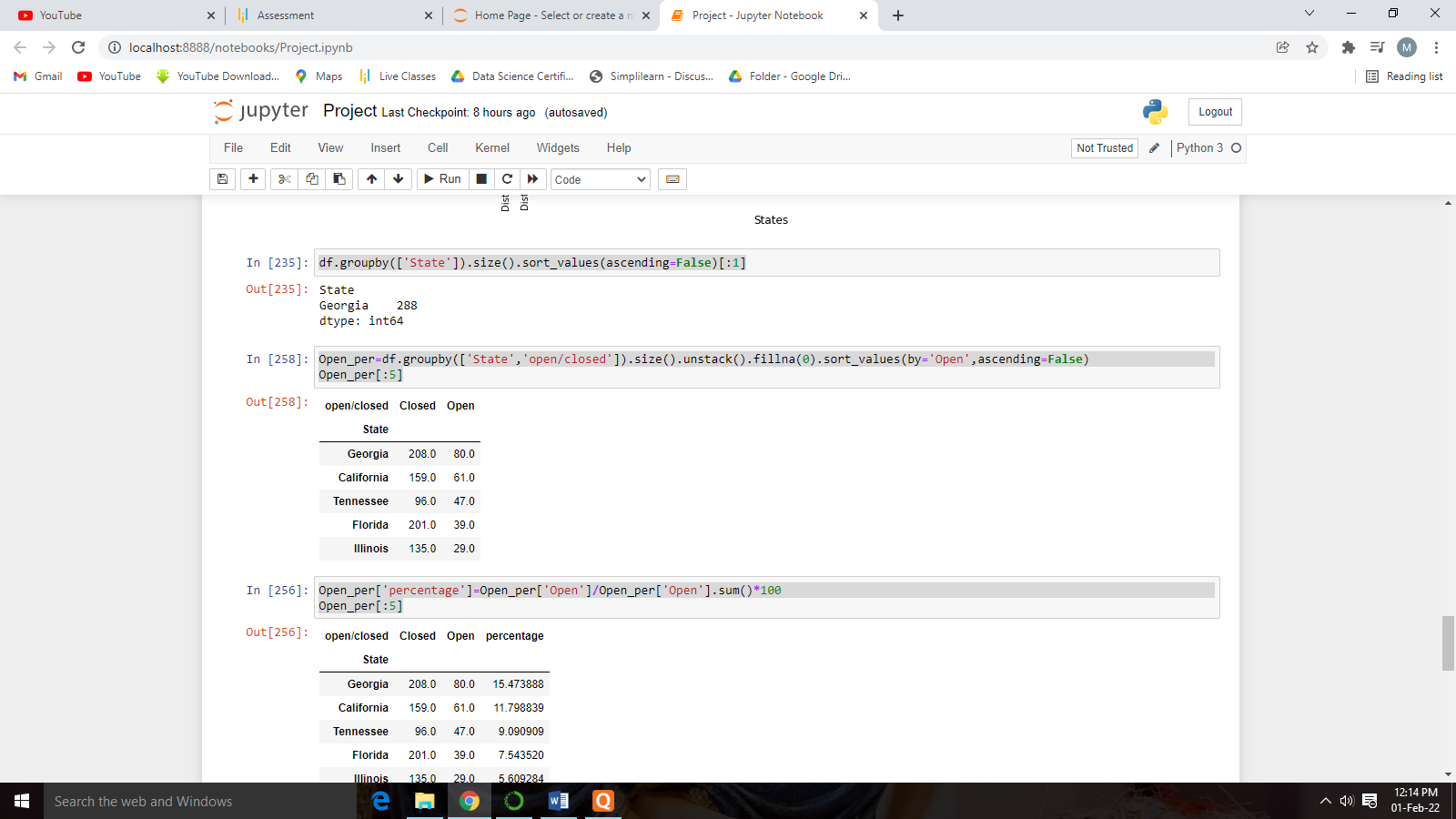
plt.show()



**Insight:** Now it`s clearly shown that the highest number of complaints recorded from the state Georgia and the second highest number of complaints recorded from the state Florida.

#Which state has the maximum complaints

df.groupby(['State']).size().sort\_values(ascending=False)[:1]



**Insight:** As seen from the table the maximum number of open cases is from the state of Georgia.

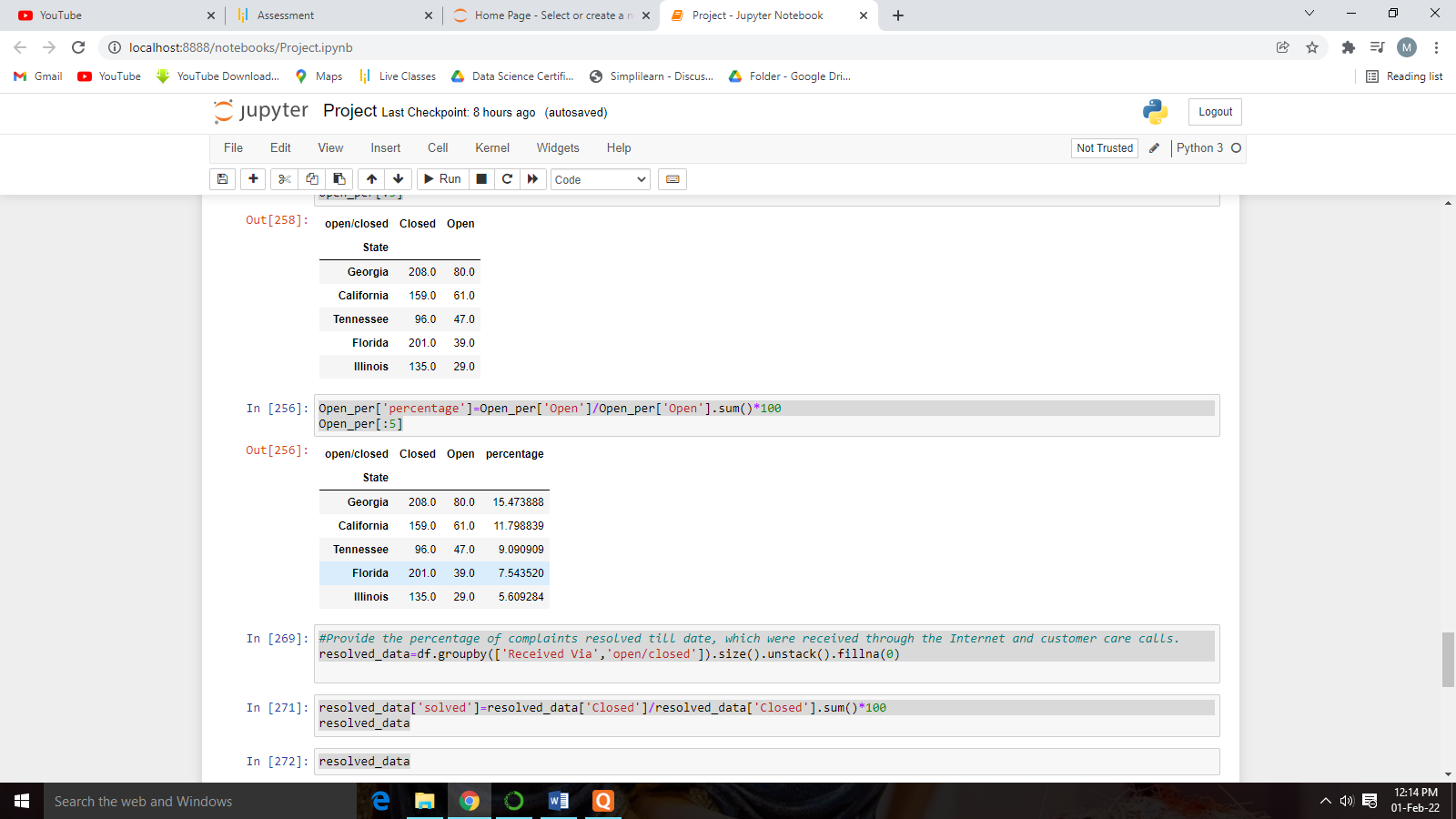
#Which state has the highest percentage of unresolved complaints

Open\_per=df.groupby(['State','open/closed']).size().unstack().fillna(0).sort\_values(by='Open',ascending=False)

Open\_per[:5]

Open\_per['percentage']=Open\_per['Open']/Open\_per['Open'].sum()\*100

Open\_per[:5]



Insight:As from the above table we can clearly see that the highest percentage of unresolved complaints are from Georgia followed by California and Tennessee.

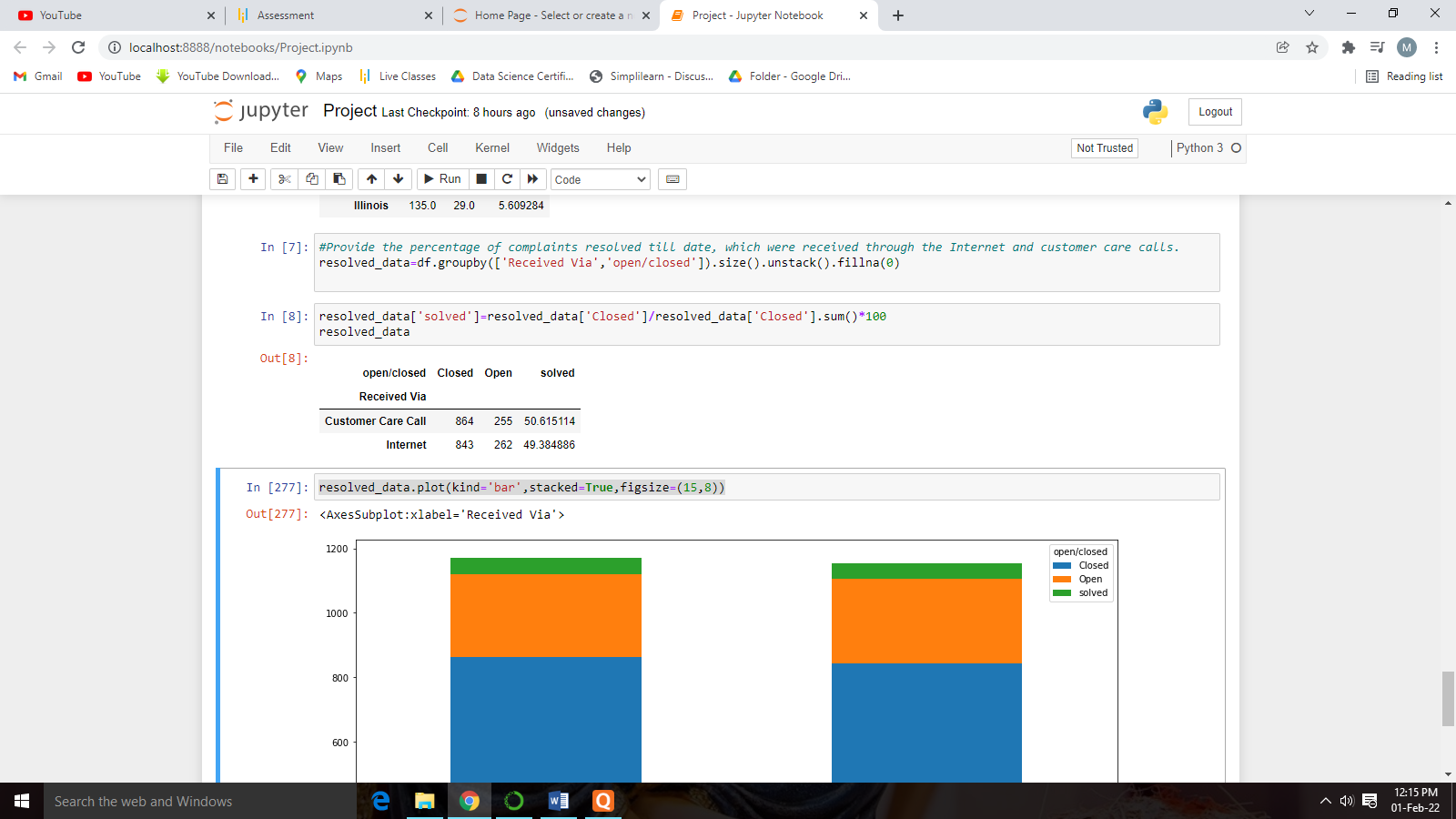
#Provide the percentage of complaints resolved till date, which were received through the Internet and customer care calls.

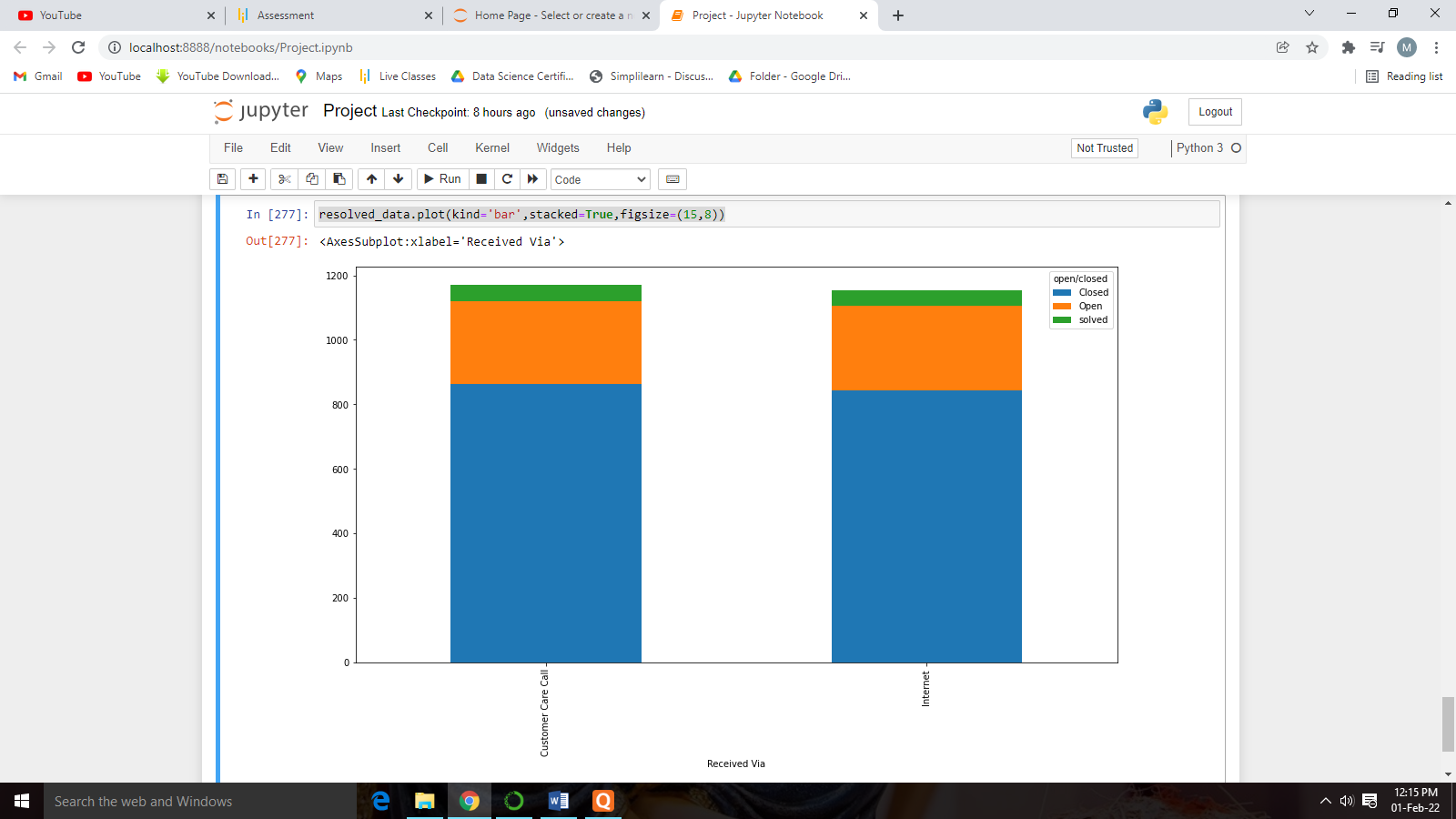
resolved\_data=df.groupby(['Received Via','open/closed']).size().unstack().fillna(0)

resolved\_data['solved']=resolved\_data['Closed']/resolved\_data['Closed'].sum()\*100

resolved\_data

resolved\_data.plot(kind='bar',stacked=True,figsize=(15,8))





From the Category Resolved the complaint received through Internet and Customer care call are near about same of 50%-50.

Conclusion:

As per the above analysis we observe that in the 2nd half of the June month Comcast received high amount of complaints in which most of the complaints are related to internet service issue and the highest amount of complaints are received from the state Georgia. The highest unresolved complaints are related from the state Georgia and the total amount of resolved complaints are 50% in which 50% are received the internet and from the customer care calls.