Step 1 : Load the data sets

Hint : make sure your path is proper. Inorder to reduce OS dependencies on the filepath use Path

Explore https://pypi.org/project/path.py/ (https://pypi.org/project/path.py/)

Step 2 : Explore the data

- Print the data frame
- Check the data types in the data frame
- Look at the head, and do a describe on the dataset

Task 1 : Provide a trend chart for the number of complaints at monthly and daily granularity levels

Hints - Daily Granualrity

- For this we need to plot date on x-axis and no of complaints on y-axis
- To plot explore plot method available with a dataframe
- Value_counts of data frame gives the frequency of occurence of a record
- Choose the date time coloum from the datat frame and find the number of occurences to identify the max number of complaints.

Hint - Monthly Granualrity

- In order to get monthly granualarity you need to parse the date time object
- Use DatetimeIndex available with pandas https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.DatetimeIndex.html (https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.DatetimeIndex.html)
- Apply DatetimeIndex on the colum having Date information
- After you get the indexes of month and year from the data frame combine year and month colums after converting them as strings.
- so you should get a something like 2015_4 for each row, then find the frequency by using value_counts()
- this approach will help even if there are complaints form multiple years, if you find this difficult just get value counts by month as our data is only pertaining to a single year here.

Task 2 : Find out which complaints are maximum

Hints

- Need to use NLP library for finding this out.
- Use CountVectorizer to tokenize the text.
- Need to play with the parameter ngram_range , as we need to extract meaningful word sequences instead of just frequent words
- N-gram is simply a sequence of N words.

Task 4 Create a new catergorical variable

- 4 status to be merged in to two open (Open & Pending) and closed (Solved & Closed).
- create a new coloum in the original dataframe.
- Use a loop to iterate through the existing status colum and add entries to the new colum.
- Use list comprehension instread of a for loop.

Task 5 : Which state has max complaints.

- Use groupby available with the data frame

Task 6 : which state has highest percentage of un resolved complaints

- Use groupby with two coloums

Task 7 : % of complaints resolved till date

- Groupby status we added, and use size() to see the count of each.
- Findi total complaints and do the %