Rystad Energy (Round1)

Name: **Deepak Kumar Mandal**
B.Tech, IIT Guwahati
Email: **dkmiitg@gmail.com**
Contact: **+91-7764933459**
Github Link (for assignment): https://github.com/deepak-mandal/RystadEnergy

I. Data Gathering

Deliverables:
- Script used to download data
- Raw data (one file in csv or xlsx format)
i. The data units s hould be: Thousand barrels per day
ii. Data granularity (period) : Monthly

There are five PADDs or “Petroleum Administration for Defense Districts” in the United States, PADD 1, PADD 2, PADD 3, PADD 4, PADD 5. These are geographical subdivisions or areas.

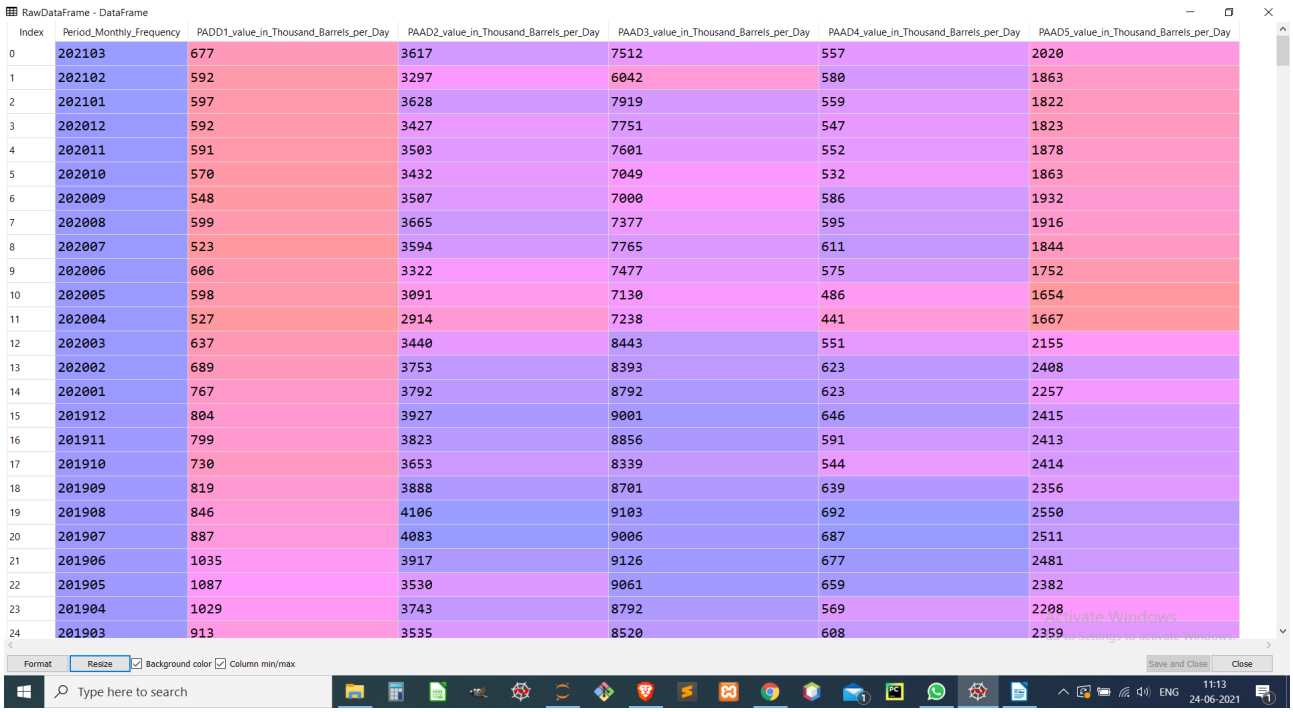
Solution:

Intermediate Generated data Files are:

- index1.html; PADD1.csv
- index2.html; PADD2.csv
- index3.html; PADD3.csv
- index4.html; PADD4.csv
- index5.html; PADD5.csv

Required Raw Data: RawData.csv or RawData.xlsx

Overview:



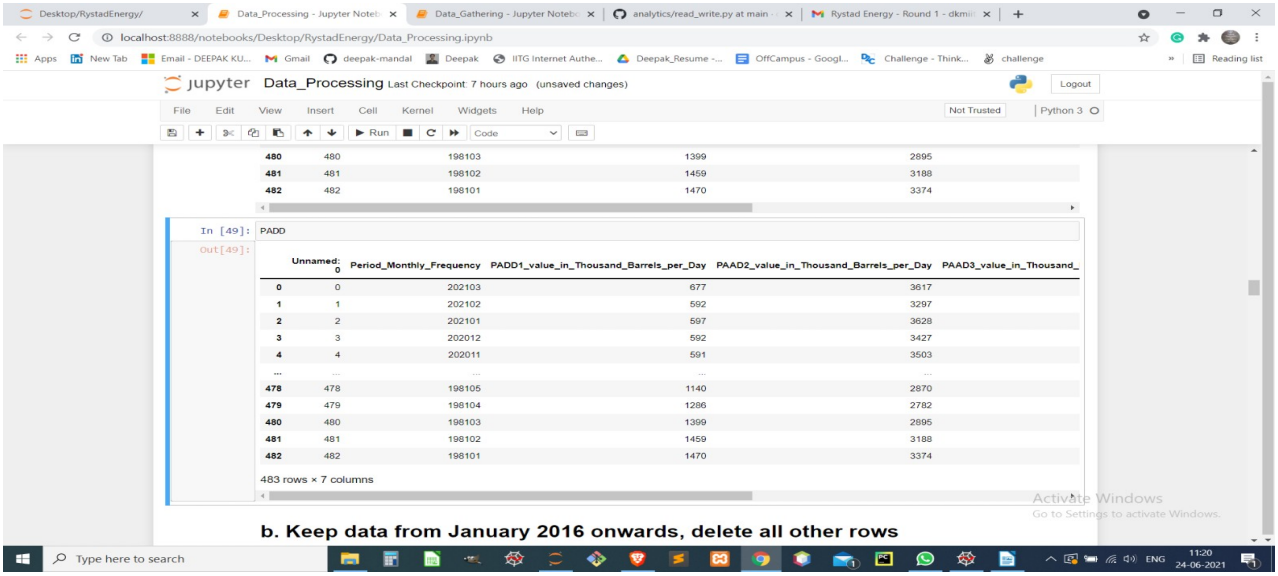
Index	Period_Monthly_Frequency	PADD1_value_in_Thousand_Barrels_per_Day	PAAD2_value_in_Thousand_Barrels_per_Day	PAAD3_value_in_Thousand_Barrels_per_Day	PAAD4_value_in_Thousand_Barrels_per_Day	PAAD5_value_in_Thousand_Barrels_per_Day
0	202103	677	3617	7512	557	2020
1	202102	592	3297	6042	580	1863
2	202101	597	3628	7919	559	1822
3	202012	592	3427	7751	547	1823
4	202011	591	3503	7601	552	1878
5	202010	570	3432	7049	532	1863
6	202009	548	3507	7000	586	1932
7	202008	599	3665	7377	595	1916
8	202007	523	3594	7765	611	1844
9	202006	606	3322	7477	575	1752
10	202005	598	3091	7130	486	1654
11	202004	527	2914	7238	441	1667
12	202003	637	3440	8443	551	2155
13	202002	689	3753	8393	623	2408
14	202001	767	3792	8792	623	2257
15	201912	804	3927	9001	646	2415
16	201911	799	3823	8856	591	2413
17	201910	730	3653	8339	544	2414
18	201909	819	3888	8701	639	2356
19	201908	846	4106	9103	692	2550
20	201907	887	4083	9006	687	2511
21	201906	1035	3917	9126	677	2481
22	201905	1087	3530	9061	659	2382
23	201904	1029	3743	8792	569	2208
24	201903	913	3535	8520	608	2359

II. Data Processing

Deliverable:

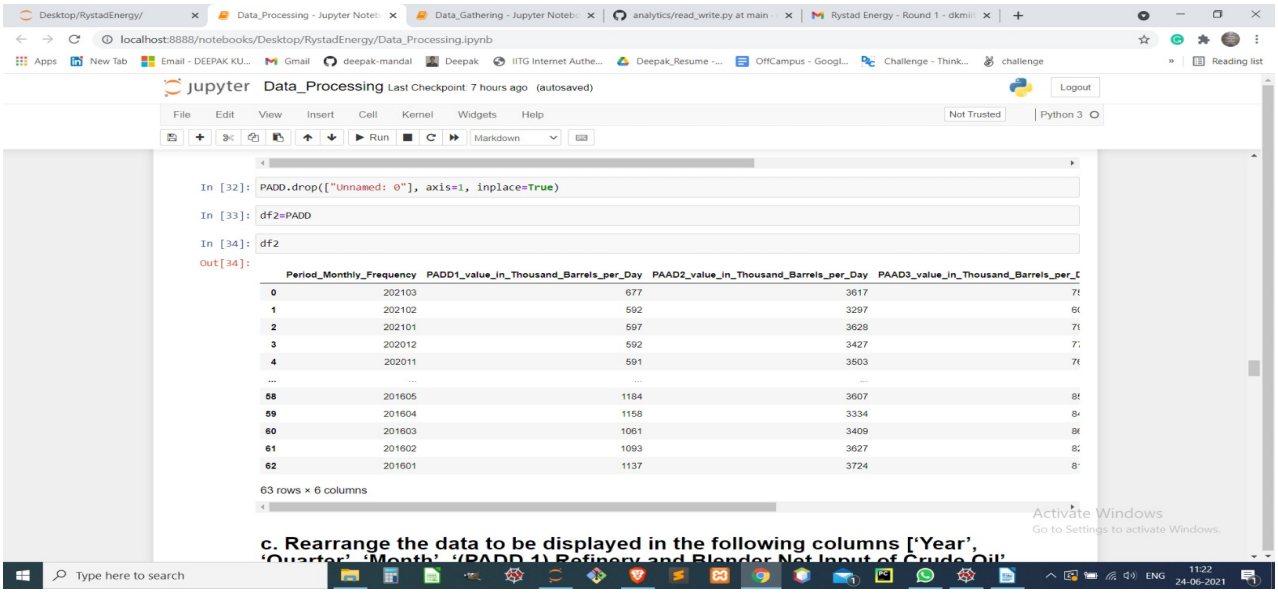
- Script or spreadsheet used for this step
- Three (3) output tables (one file in csv or xlsx format with three different tables: one with data by month, another with data summarized by quarter, another with data summarized by year)

Solution: a. From the data downloaded in the previous step, read the time series for “US Refinery and Blender Net Input of Crude Oil” by PADD



	Unnamed: 0	Period_Monthly_Frequency	PADD1_value_in_Thousand_Barrels_per_Day	PAAD2_value_in_Thousand_Barrels_per_Day	PAAD3_value_in_Thousand_
0	0	202103	677	3617	
1	1	202102	592	3297	
2	2	202101	597	3628	
3	3	202012	592	3427	
4	4	202011	591	3503	
...	...	...	...	...	
478	478	198105	1140	2870	
479	479	198104	1286	2782	
480	480	198103	1399	2895	
481	481	198102	1459	3188	
482	482	198101	1470	3374	

b. Keep data from January 2016 onwards, delete all other rows



```
In [32]: PADD.drop(["Unnamed: 0"], axis=1, inplace=True)

In [33]: df2=PADD

In [34]: df2
```

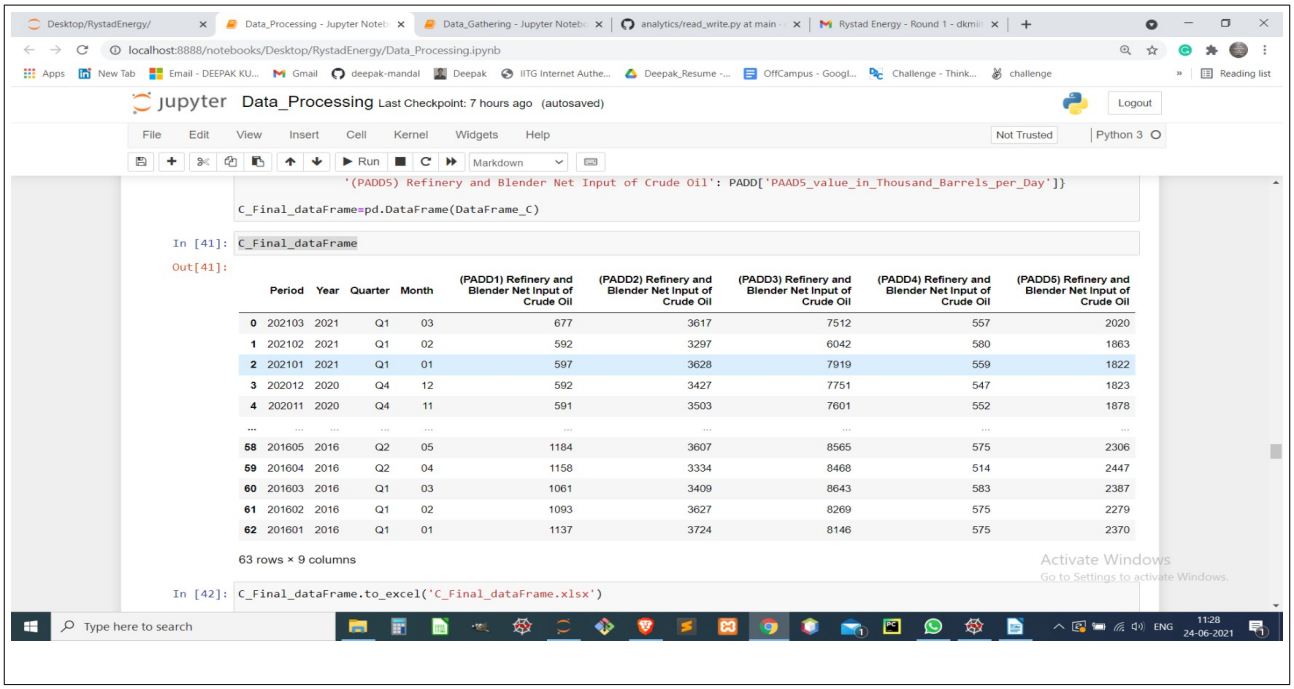
Out[34]:

	Period_Monthly_Frequency	PADD1_value_in_Thousand_Barrels_per_Day	PAAD2_value_in_Thousand_Barrels_per_Day	PAAD3_value_in_Thousand_Barrels_per_D
0	202103	677	3617	75
1	202102	592	3297	6(
2	202101	597	3628	75
3	202012	592	3427	7:
4	202011	591	3503	7(
...	...	...	...	
58	201605	1184	3607	85
59	201604	1158	3334	8-
60	201603	1061	3409	8(
61	201602	1093	3627	8:
62	201601	1137	3724	8'

63 rows × 6 columns

c. Rearrange the data to be displayed in the following columns ['Year', 'Quarter', 'Month', '(PADD 1) Refinery and Blender Net Input of Crude Oil

c. Rearrange the data to be displayed in the following columns ['Year', 'Quarter', 'Month', '(PADD 1) Refinery and Blender Net Input of Crude Oil', '(PADD 2) Refinery and Blender Net Input of Crude Oil', '(PADD 3) Refinery and Blender Net Input of Crude Oil', '(PADD 4) Refinery and Blender Net Input of Crude Oil', '(PADD 5) Refinery and Blender Net Input of Crude Oil']



```
'(PADD5) Refinery and Blender Net Input of Crude Oil': PADD['PAAD5_value_in_Thousand_Barrels_per_Day']}

C_Final_dataFrame=pd.DataFrame(DataFrame_C)

In [41]: C_Final_dataFrame
```

Out[41]:

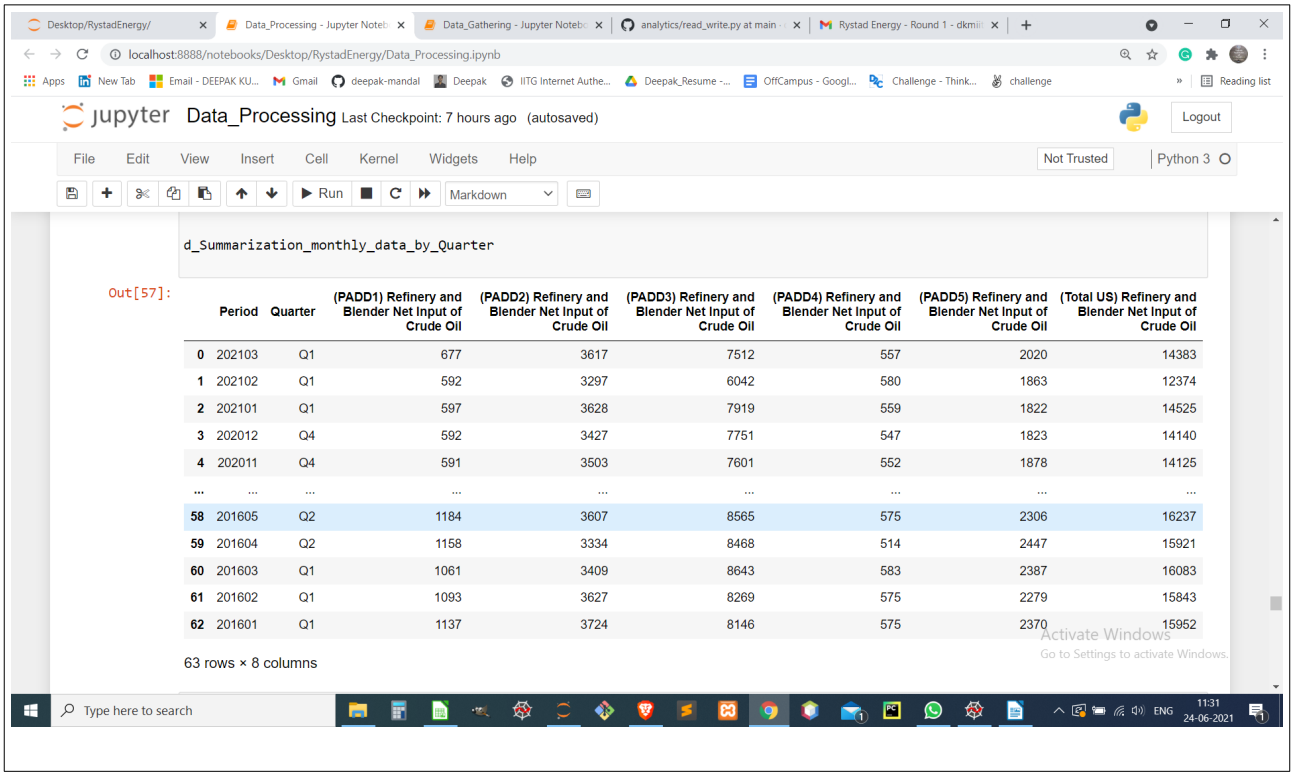
	Period	Year	Quarter	Month	(PADD1) Refinery and Blender Net Input of Crude Oil	(PADD2) Refinery and Blender Net Input of Crude Oil	(PADD3) Refinery and Blender Net Input of Crude Oil	(PADD4) Refinery and Blender Net Input of Crude Oil	(PADD5) Refinery and Blender Net Input of Crude Oil
0	202103	2021	Q1	03	677	3617	7512	557	2020
1	202102	2021	Q1	02	592	3297	6042	580	1863
2	202101	2021	Q1	01	597	3628	7919	559	1822
3	202012	2020	Q4	12	592	3427	7751	547	1823
4	202011	2020	Q4	11	591	3503	7601	552	1878
...	...	...	...	...	...	...	...	...	...
58	201605	2016	Q2	05	1184	3607	8565	575	2306
59	201604	2016	Q2	04	1158	3334	8468	514	2447
60	201603	2016	Q1	03	1061	3409	8643	583	2387
61	201602	2016	Q1	02	1093	3627	8269	575	2279
62	201601	2016	Q1	01	1137	3724	8146	575	2370

63 rows × 9 columns

```
In [42]: C_Final_dataFrame.to_excel('C_Final_dataFrame.xlsx')
```

Main Output Excel Data: “C_Final_dataFrame.xlsx”

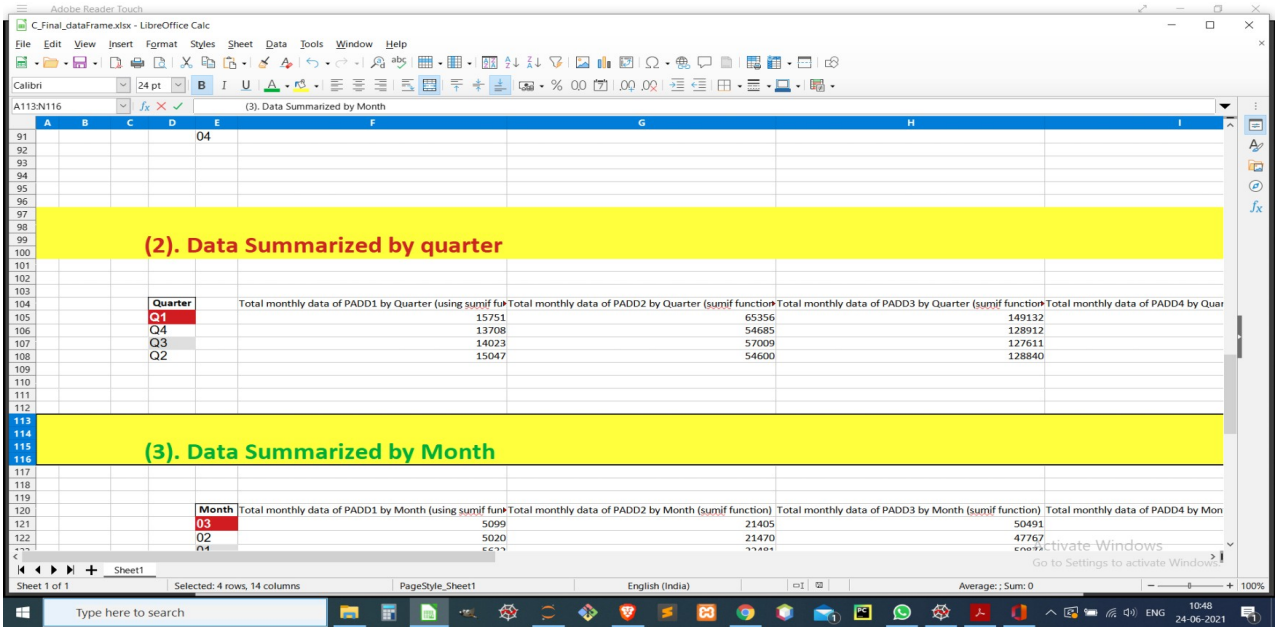
d. For each year/month, sum up the ‘Refinery and Blender Net Input of Crude Oil’ data for the five PADDs and assign the name ‘Total US Refinery Net Input of Crude Oil’ to the time series



	Period	Quarter	(PADD1) Refinery and Blender Net Input of Crude Oil	(PADD2) Refinery and Blender Net Input of Crude Oil	(PADD3) Refinery and Blender Net Input of Crude Oil	(PADD4) Refinery and Blender Net Input of Crude Oil	(PADD5) Refinery and Blender Net Input of Crude Oil	(Total US) Refinery and Blender Net Input of Crude Oil
0	202103	Q1	677	3617	7512	557	2020	14383
1	202102	Q1	592	3297	6042	580	1863	12374
2	202101	Q1	597	3628	7919	559	1822	14525
3	202012	Q4	592	3427	7751	547	1823	14140
4	202011	Q4	591	3503	7601	552	1878	14125
...	...	...	...	...	...	...	...	...
58	201605	Q2	1184	3607	8565	575	2306	16237
59	201604	Q2	1158	3334	8468	514	2447	15921
60	201603	Q1	1061	3409	8643	583	2387	16083
61	201602	Q1	1093	3627	8269	575	2279	15843
62	201601	Q1	1137	3724	8146	575	2370	15952

Now onwards, I have done all calculation in Excel file: (Three (3) output tables (one file in csv or xlsx format with three different tables: one with data by month, another with data summarized by quarter, another with data summarized by year))

e. Create a new output table that summarizes monthly data by ‘ Quarter’ for each of the six time series (‘PADD 1 Refinery and Blender Net Input of Crude Oil’, ‘(PADD 2)…’ ‘(PADD 3)… ’, ‘(PADD 4)… ’, ‘(PADD 5)…’, ‘Total US…’)



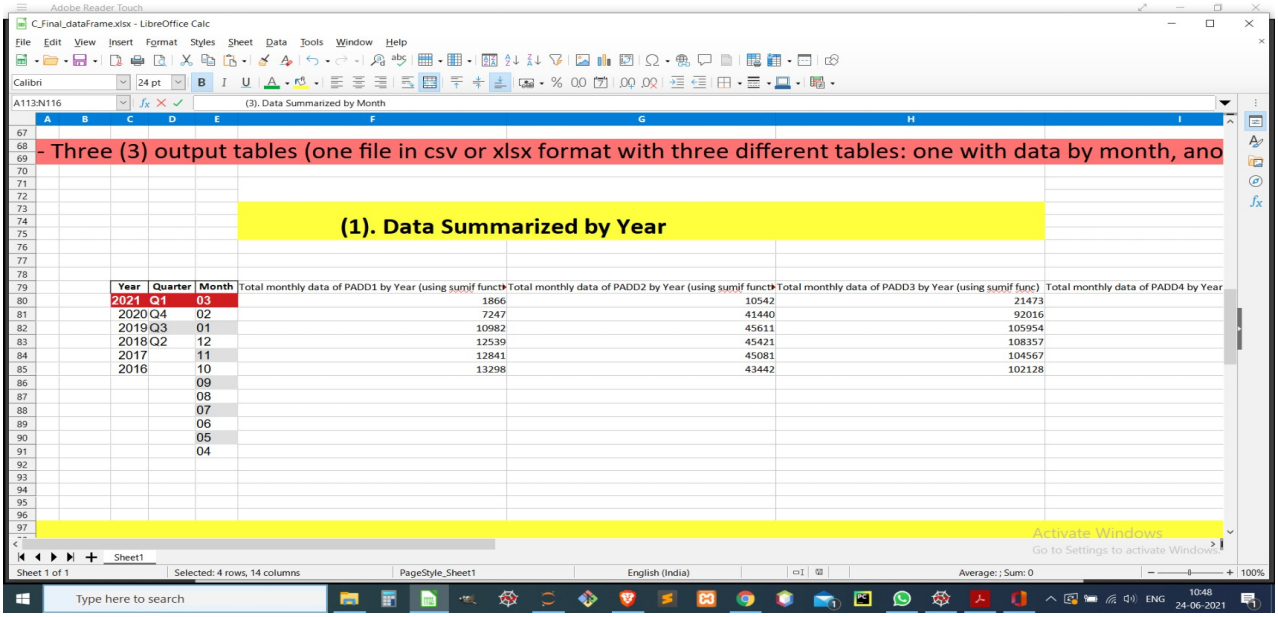
(2). Data Summarized by quarter

Quarter	Total monthly data of PADD1 by Quarter (using sumif fu	Total monthly data of PADD2 by Quarter (sumif function	Total monthly data of PADD3 by Quarter (sumif function	Total monthly data of PADD4 by Quar
Q1	15751	65356	149132	
Q4	13708	54685	128912	
Q3	14023	57009	127611	
Q2	15047	54600	128840	

(3). Data Summarized by Month

Month	Total monthly data of PADD1 by Month (using sumif fun	Total monthly data of PADD2 by Month (sumif function)	Total monthly data of PADD3 by Month (sumif function)	Total monthly data of PADD4 by Mon
03	5099	21405	50491	
02	5020	21470	47767	

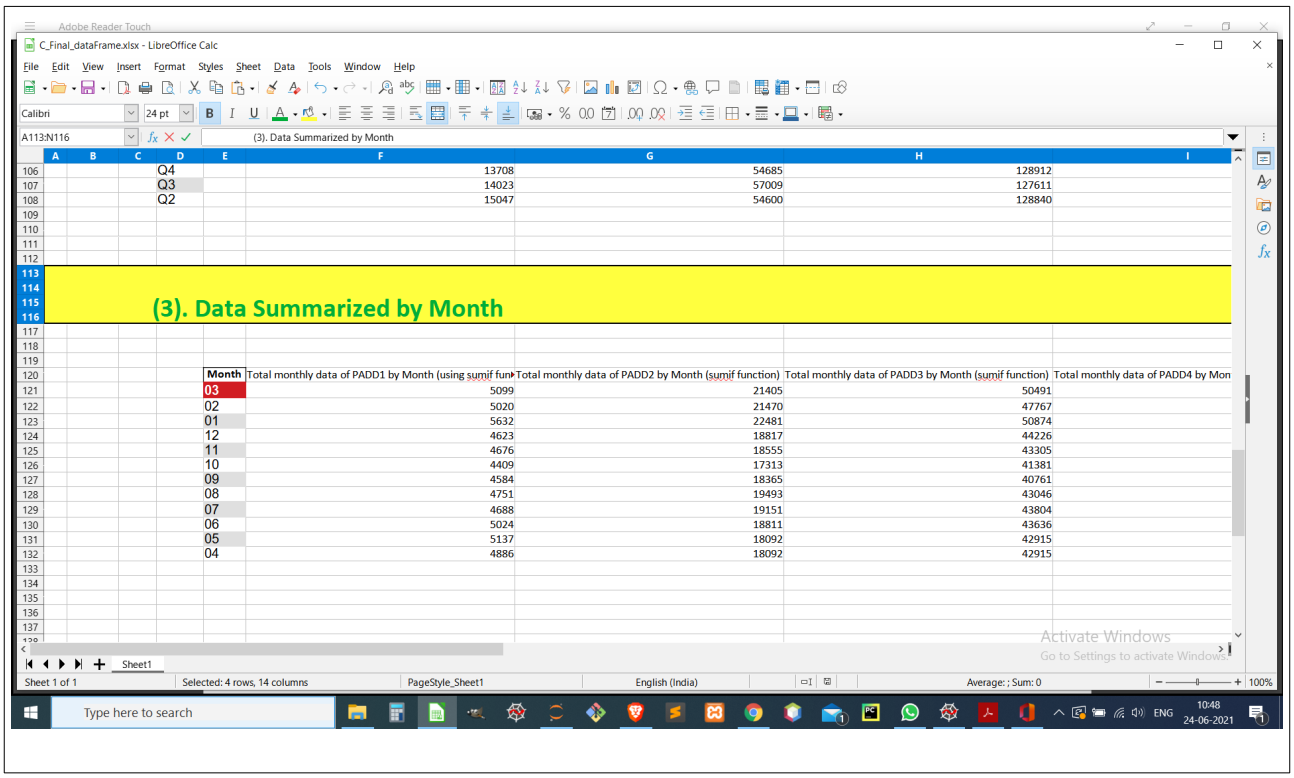
f. Create a new output table that summarizes monthly data by ‘ Year’ f or each of the six (6) time series (‘PADD 1 Refinery and Blender Net Input of Crude Oil’, ‘(PADD 2)…’ ‘(PADD 3)… ’, ‘(PADD 4)… ’, ‘(PADD 5)…’, ‘Total US…’)



- Three (3) output tables (one file in csv or xlsx format with three different tables: one with data by month, ano

(1). Data Summarized by Year

Year	Quarter	Month	Total monthly data of PADD1 by Year (using sumif funct	Total monthly data of PADD2 by Year (using sumif funct	Total monthly data of PADD3 by Year (using sumif func)	Total monthly data of PADD4 by Year
2021	Q1	03	1866	10542	21473	
2020	Q4	02	7247	41440	92016	
2019	Q3	01	10982	45611	105954	
2018	Q2	12	12539	45421	108357	
2017		11	12841	45081	104567	
2016		10	13298	43442	102128	
		09				
		08				
		07				
		06				
		05				
		04				



106			Q4		13708	54685	128912	
107			Q3		14023	57009	127611	
108			Q2		15047	54600	128840	
109								
110								
111								
112								
113–116			(3). Data Summarized by Month					
117								
118								
119								

Month	Total monthly data of PADD1 by Month (using sumif fun	Total monthly data of PADD2 by Month (sumif function)	Total monthly data of PADD3 by Month (sumif function)	Total monthly data of PADD4 by Mon
03	5099	21405	50491	
02	5020	21470	47767	
01	5632	22481	50874	
12	4623	18817	44226	
11	4676	18555	43305	
10	4409	17313	41381	
09	4584	18365	40761	
08	4751	19493	43046	
07	4688	19151	43804	
06	5024	18811	43636	
05	5137	18092	42915	
04	4886	18092	42915	