Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA   Batch - 61
EADC Practical 18

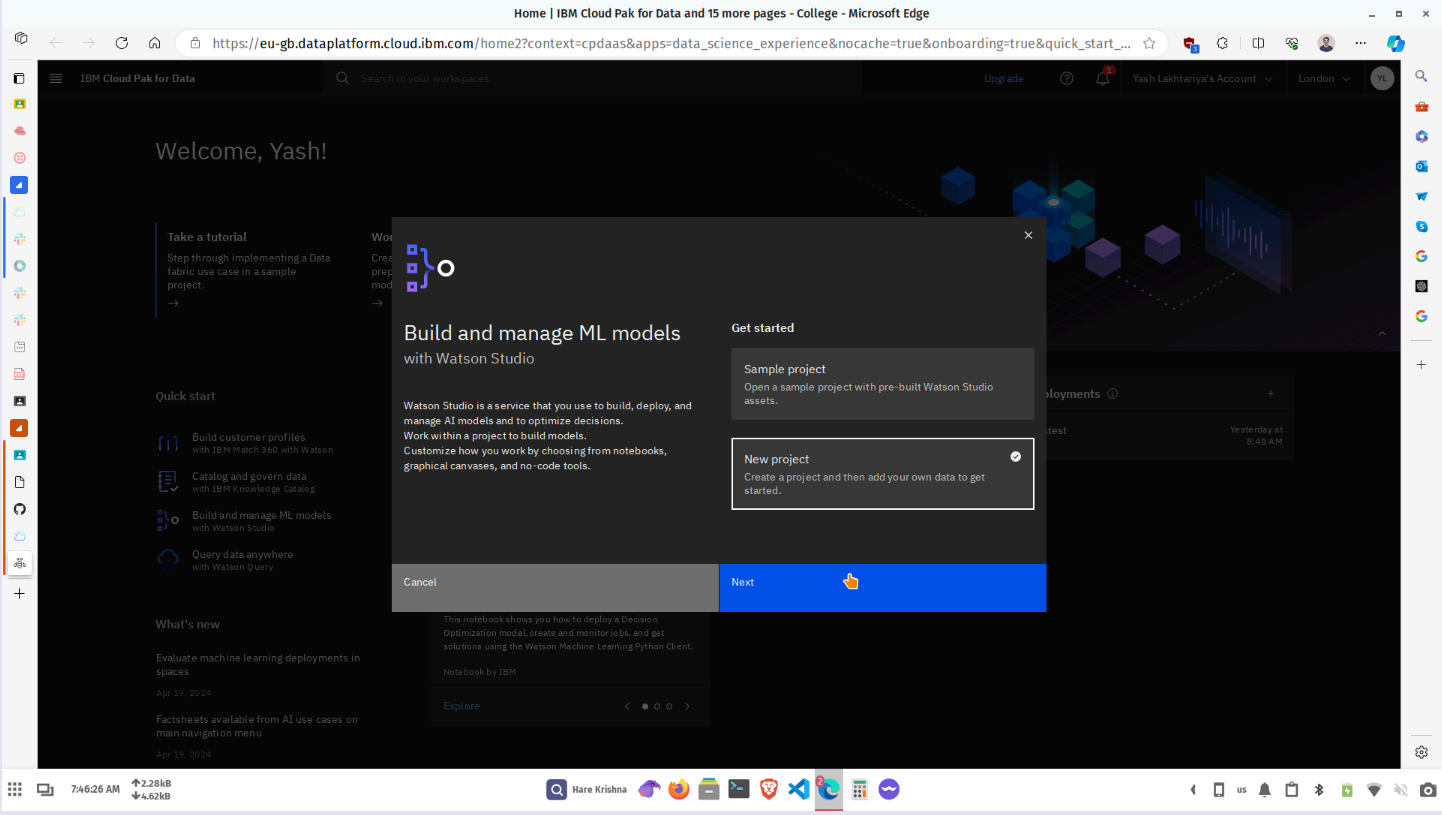
Aim : Assume you are working in a company where you need to extract useful insights from the data collected by organization. Demonstrate how to analyze large datasets with Python data science packages. We'll provide an example use case of analyzing hourly air quality data provided by the EPA.

Perform the following Tasks :

1. **Create a Juypter notebook in Watson Studio**
2. **Extract patterns from datasets using pandas**
3. **Visualize data trends via matplotlib graphs**

Steps and Screenshots :

1. Create new project on Watson Studio IBM instance


Home | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge
https://eu-gb.dataplatform.cloud.ibm.com/home2?context=cpdaas&apps=data_science_experience&nocache=true&onboarding=true&quick_start_...
IBM Cloud Pak for Data
Upgrade
Yash Lakhtariya's Account
London
Welcome, Yash!
Take a tutorial
Quick start
What's new
7:46:26 AM

Build and manage ML models
with Watson Studio

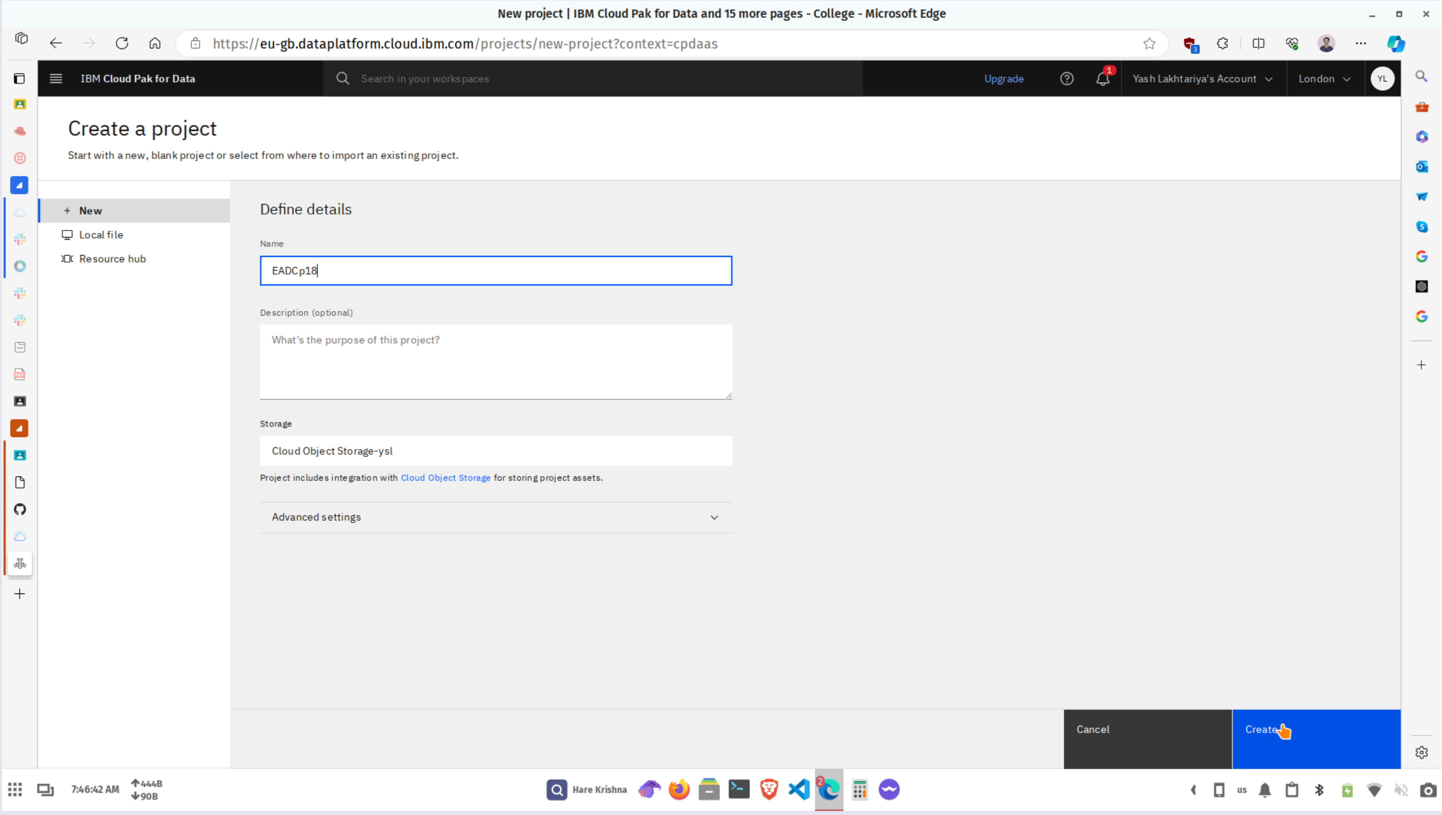
Watson Studio is a service that you use to build, deploy, and manage AI models and to optimize decisions.
Work within a project to build models.
Customize how you work by choosing from notebooks, graphical canvases, and no-code tools.

Get started

Sample project
Open a sample project with pre-built Watson Studio assets.

New project
Create a project and then add your own data to get started.

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA Batch - 61
EADC Practical 18



New project | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge

https://eu-gb.dataplatform.cloud.ibm.com/projects/new-project?context=cpdaas

Create a project

Start with a new, blank project or select from where to import an existing project.

Define details

Name: EADCp18

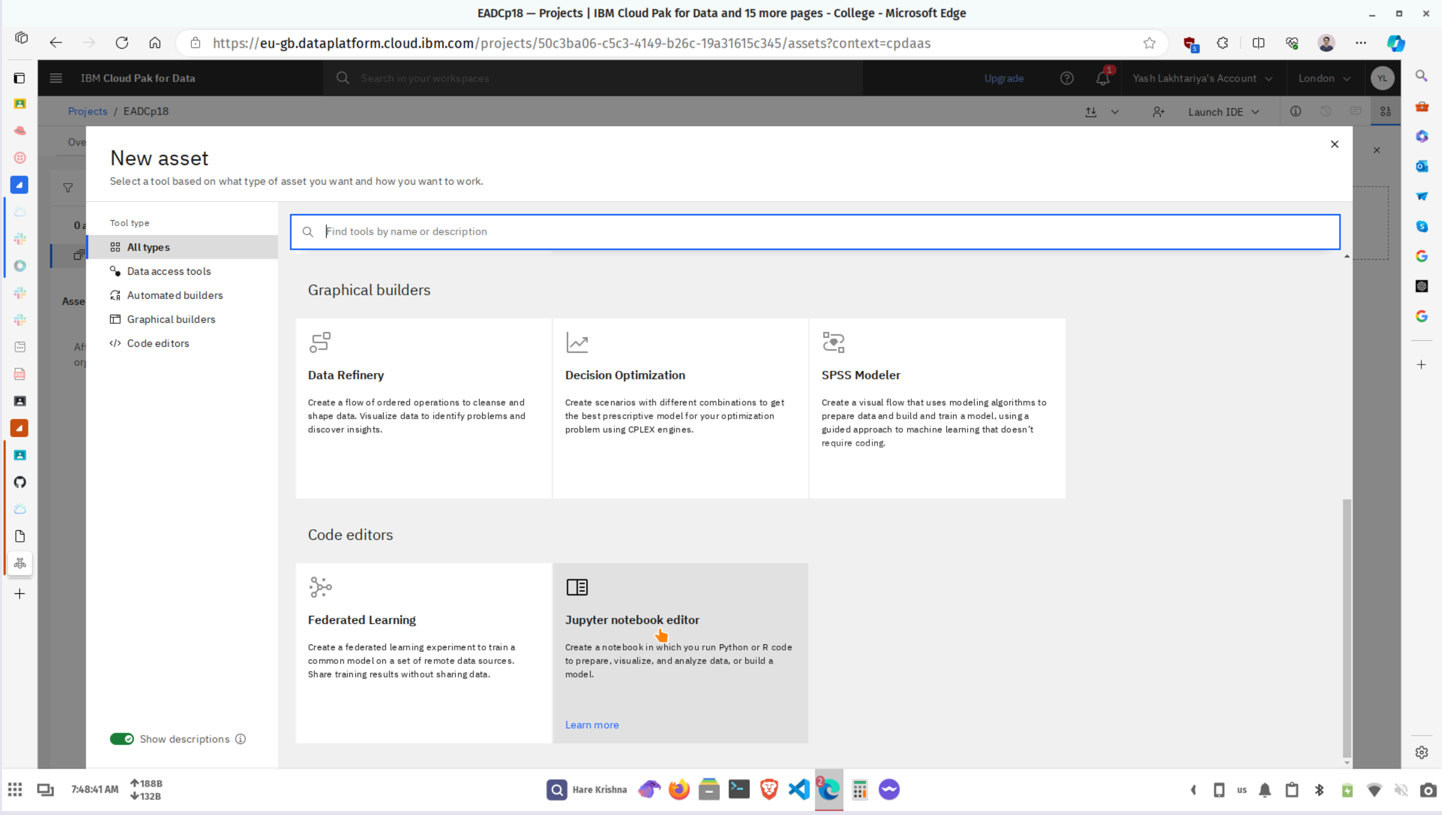
Description (optional): What's the purpose of this project?

Storage: Cloud Object Storage-ysl

Project includes integration with Cloud Object Storage for storing project assets.

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA Batch - 61
EADC Practical 18

2. Add asset of Jupyter notebook



EADCp18 — Projects | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge

https://eu-gb.dataplatform.cloud.ibm.com/projects/50c3ba06-c5c3-4149-b26c-19a31615c345/assets?context=cpdaas

New asset

Select a tool based on what type of asset you want and how you want to work.

Graphical builders

Data Refinery

Create a flow of ordered operations to cleanse and shape data. Visualize data to identify problems and discover insights.

Decision Optimization

Create scenarios with different combinations to get the best prescriptive model for your optimization problem using CPLEX engines.

SPSS Modeler

Create a visual flow that uses modeling algorithms to prepare data and build and train a model, using a guided approach to machine learning that doesn't require coding.

Code editors

Federated Learning

Create a federated learning experiment to train a common model on a set of remote data sources. Share training results without sharing data.

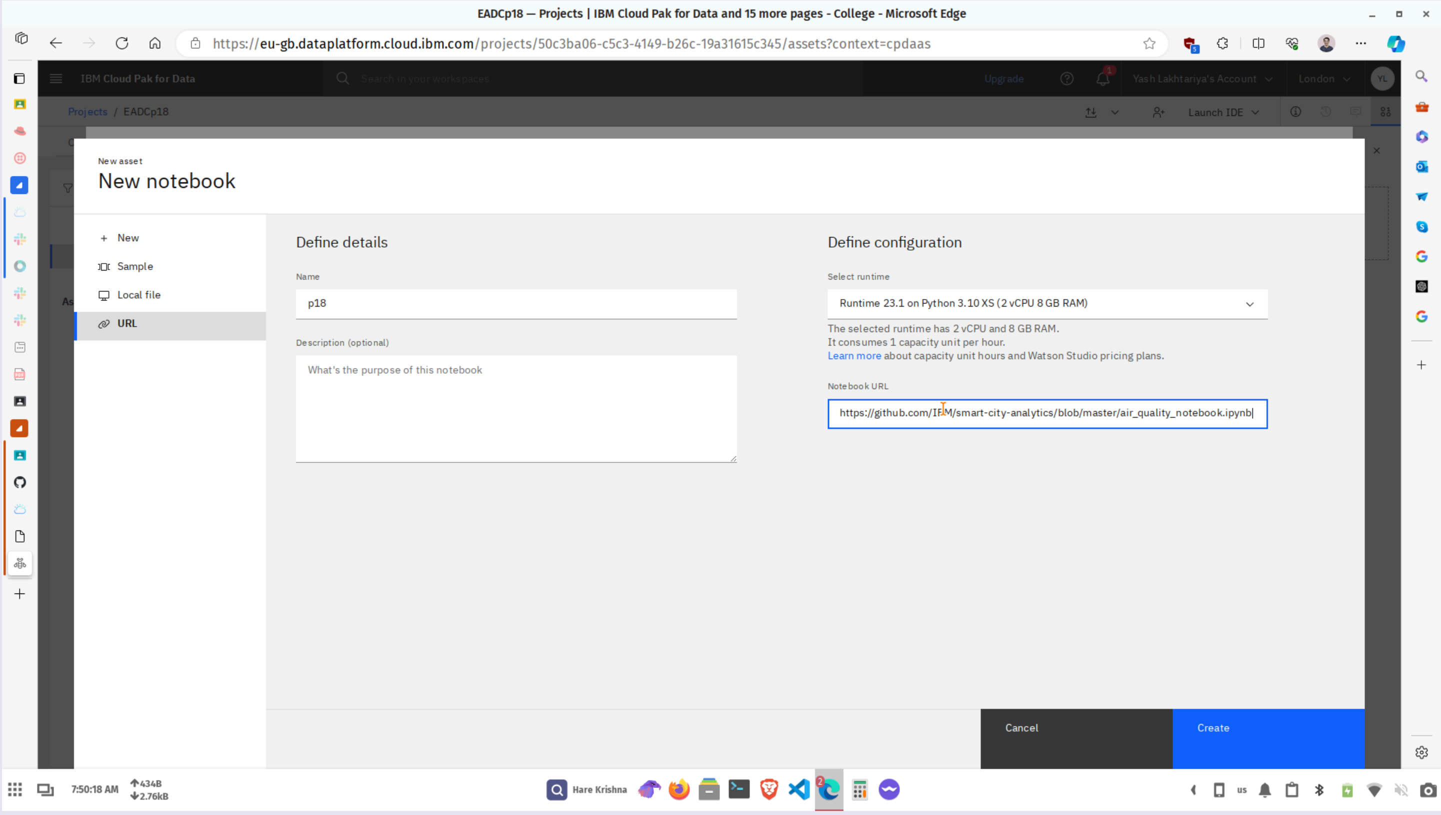
Jupyter notebook editor

Create a notebook in which you run Python or R code to prepare, visualize, and analyze data, or build a model.

Learn more

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA   Batch - 61
EADC Practical 18

3. Add new notebook from URL source and provide URL for ipynb file, here from Github



EADCp18 — Projects | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge

https://eu-gb.dataplatform.cloud.ibm.com/projects/50c3ba06-c5c3-4149-b26c-19a31615c345/assets?context=cpdaas

New asset

New notebook

Define details

Name: p18

Description (optional): What's the purpose of this notebook

Define configuration

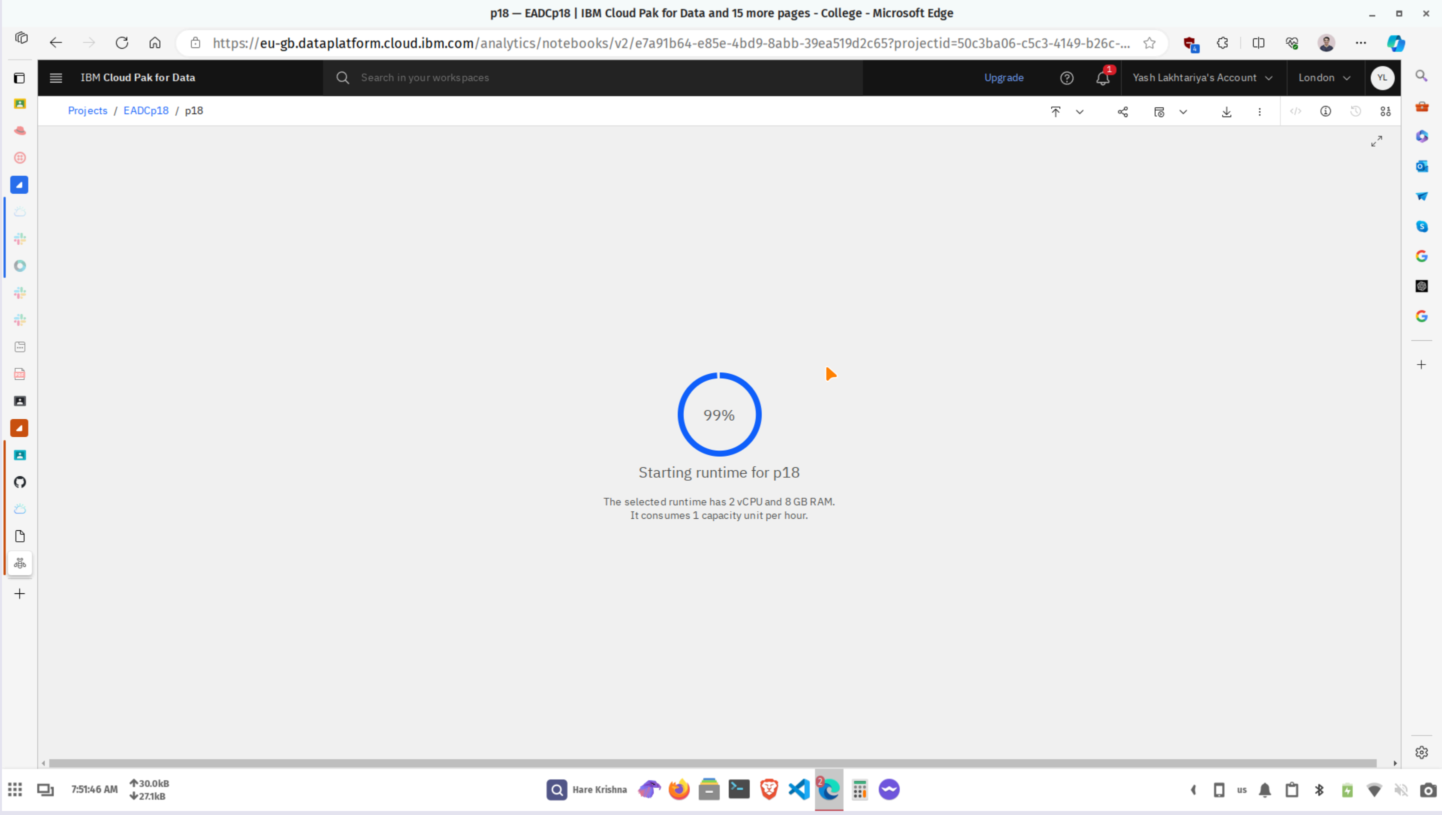
Select runtime: Runtime 23.1 on Python 3.10 XS (2 vCPU 8 GB RAM)

The selected runtime has 2 vCPU and 8 GB RAM.
It consumes 1 capacity unit per hour.
Learn more about capacity unit hours and Watson Studio pricing plans.

Notebook URL: https://github.com/IBM/smart-city-analytics/blob/master/air_quality_notebook.ipynb

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA Batch - 61
EADC Practical 18

4. Wait for creation to be completed


p18 — EADCp18 | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge
https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/e7a91b64-e85e-4bd9-8abb-39ea519d2c65?projectid=50c3ba06-c5c3-4149-b26c-...
IBM Cloud Pak for Data
Search in your workspaces
Upgrade
Yash Lakhtariya's Account
London
YL
Projects / EADCp18 / p18
7:51:46 AM
Hare Krishna

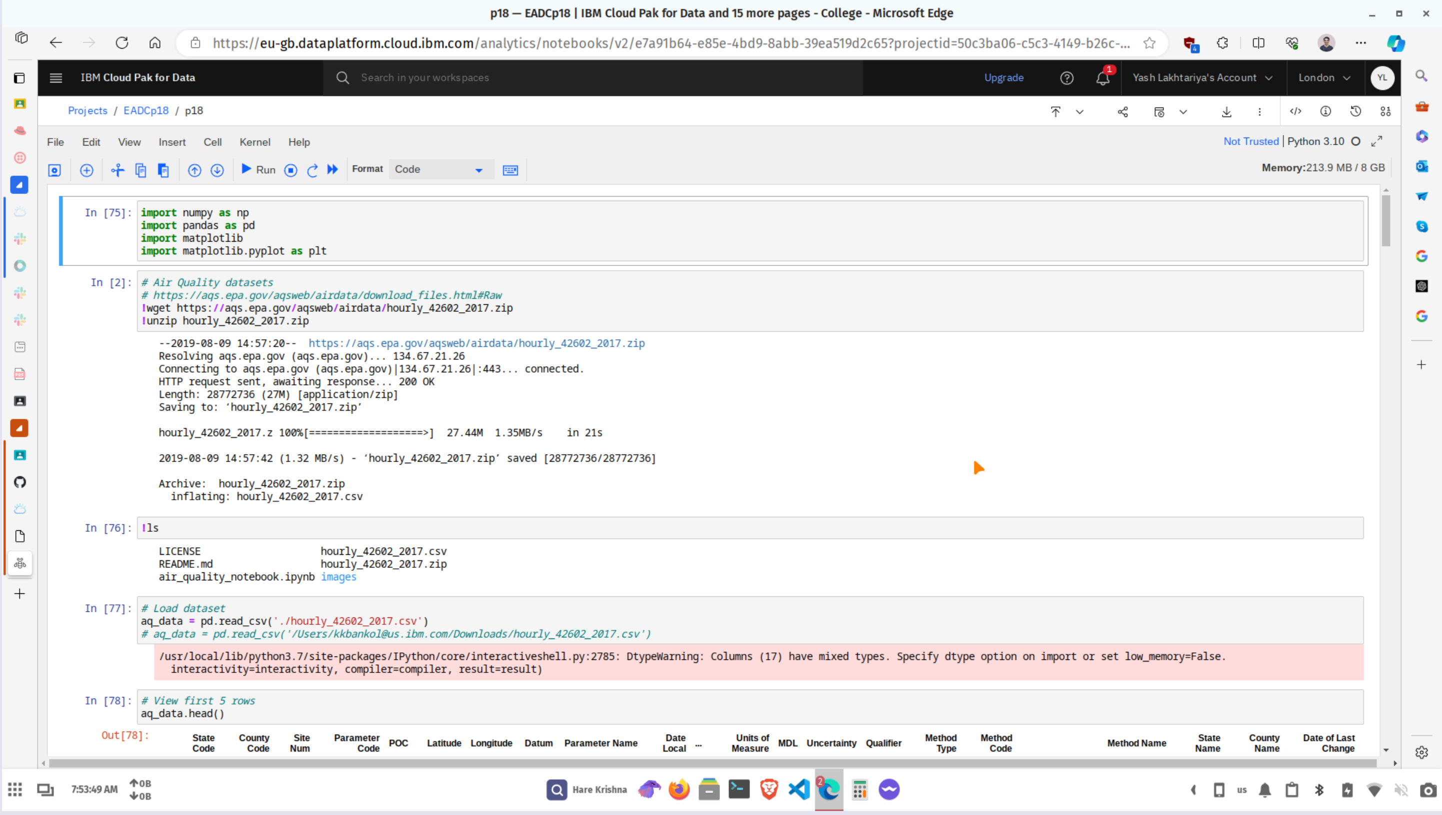
99%

Starting runtime for p18

The selected runtime has 2 vCPU and 8 GB RAM.
It consumes 1 capacity unit per hour.

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA   Batch - 61
EADC Practical 18

5. Now explore the notebook, first clear all outputs



p18 — EADCp18 | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge

https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/e7a91b64-e85e-4bd9-8abb-39ea519d2c65?projectid=50c3ba06-c5c3-4149-b26c-...

Projects / EADCp18 / p18

Not Trusted | Python 3.10

Memory:213.9 MB / 8 GB

```
In [75]: import numpy as np
         import pandas as pd
         import matplotlib
         import matplotlib.pyplot as plt
```

```
In [2]:  # Air Quality datasets
         # https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw
         !wget https://aqs.epa.gov/aqsweb/airdata/hourly_42602_2017.zip
         !unzip hourly_42602_2017.zip
```

```
--2019-08-09 14:57:20--  https://aqs.epa.gov/aqsweb/airdata/hourly_42602_2017.zip
Resolving aqs.epa.gov (aqs.epa.gov)... 134.67.21.26
Connecting to aqs.epa.gov (aqs.epa.gov)|134.67.21.26|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 28772736 (27M) [application/zip]
Saving to: ‘hourly_42602_2017.zip’

hourly_42602_2017.z 100%[===================>]  27.44M  1.35MB/s    in 21s

2019-08-09 14:57:42 (1.32 MB/s) - ‘hourly_42602_2017.zip’ saved [28772736/28772736]

Archive:  hourly_42602_2017.zip
  inflating: hourly_42602_2017.csv
```

```
In [76]: !ls
```

```
LICENSE                    hourly_42602_2017.csv
README.md                  hourly_42602_2017.zip
air_quality_notebook.ipynb images
```

```
In [77]: # Load dataset
         aq_data = pd.read_csv('./hourly_42602_2017.csv')
         # aq_data = pd.read_csv('/Users/kkbankol@us.ibm.com/Downloads/hourly_42602_2017.csv')
```

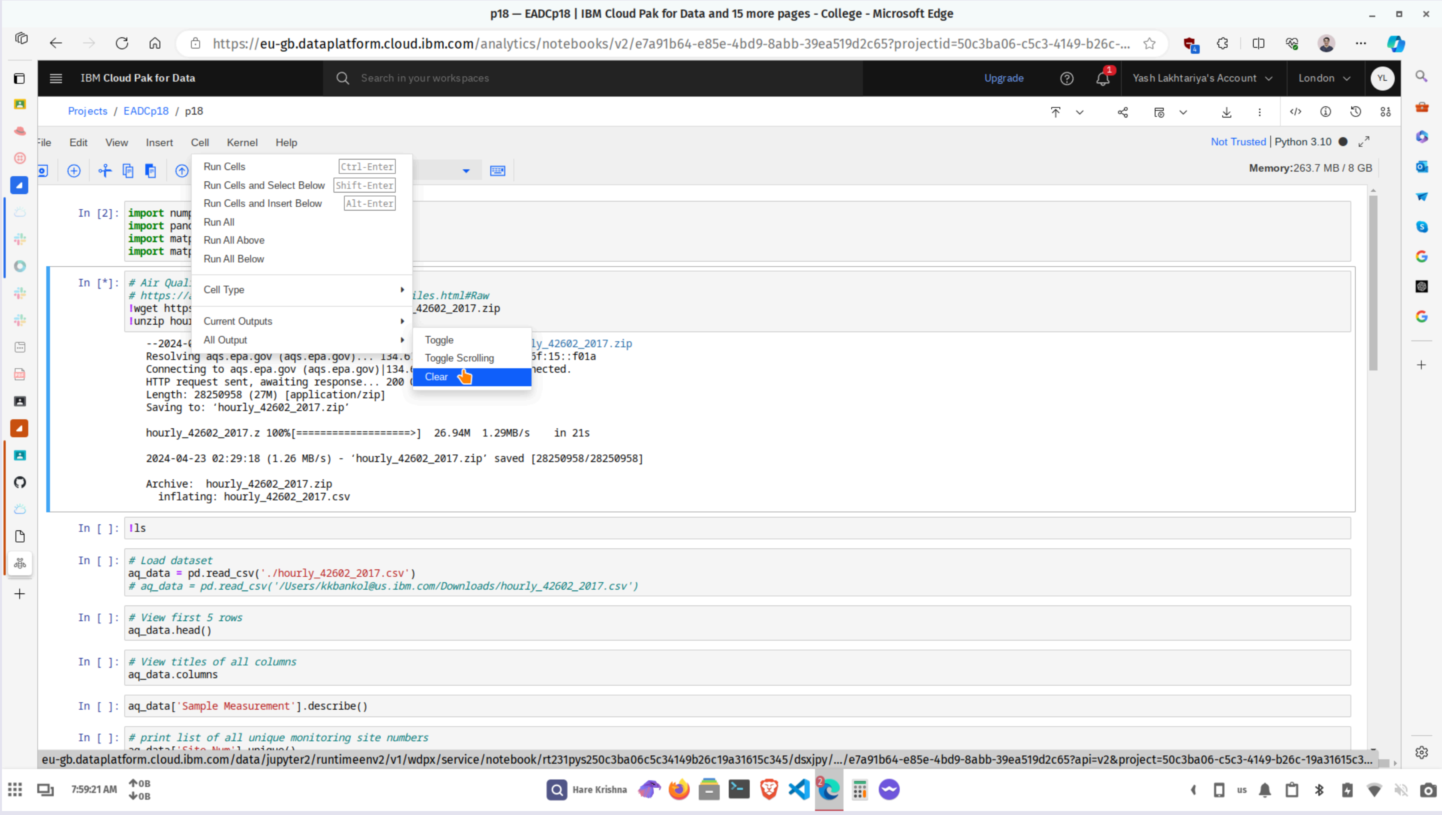
```
/usr/local/lib/python3.7/site-packages/IPython/core/interactiveshell.py:2785: DtypeWarning: Columns (17) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

```
In [78]: # View first 5 rows
         aq_data.head()
```

Out[78]:

State Code	County Code	Site Num	Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Date Local	...	Units of Measure	MDL	Uncertainty	Qualifier	Method Type	Method Code	Method Name	State Name	County Name	Date of Last Change

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA   Batch - 61
EADC Practical 18


p18 — EADCp18 | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge
https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/e7a91b64-e85e-4bd9-8abb-39ea519d2c65?projectid=50c3ba06-c5c3-4149-b26c-...
IBM Cloud Pak for Data
Search in your workspaces
Upgrade
Yash Lakhtariya's Account
London
YL
Projects / EADCp18 / p18
File
Edit
View
Insert
Cell
Kernel
Help
Not Trusted | Python 3.10
Memory:263.7 MB / 8 GB
Run Cells Ctrl-Enter
Run Cells and Select Below Shift-Enter
Run Cells and Insert Below Alt-Enter
Run All
Run All Above
Run All Below
Cell Type
Current Outputs
All Output
Toggle
Toggle Scrolling
Clear
eu-gb.dataplatform.cloud.ibm.com/data/jupyter2/runtimeenv2/v1/wdpx/service/notebook/rt231pys250c3ba06c5c34149b26c19a31615c345/dsxjpy/.../e7a91b64-e85e-4bd9-8abb-39ea519d2c65?api=v2&project=50c3ba06-c5c3-4149-b26c-19a31615c3...
7:59:21 AM
0B
0B
Hare Krishna
us

```
In [2]: import nump
        import panc
        import matp
        import matp
```

```
In [*]: # Air Qual                                  iles.html#Raw
        # https://a
        !wget https                                 _42602_2017.zip
        !unzip hour

          --2024-(                                                ly_42602_2017.zip
          Resolving aqs.epa.gov (aqs.epa.gov)... 134.6            5f:15::f01a
          Connecting to aqs.epa.gov (aqs.epa.gov)|134.(           nected.
          HTTP request sent, awaiting response... 200 (
          Length: 28250958 (27M) [application/zip]
          Saving to: ‘hourly_42602_2017.zip’

          hourly_42602_2017.z 100%[===================>]  26.94M  1.29MB/s    in 21s

          2024-04-23 02:29:18 (1.26 MB/s) - ‘hourly_42602_2017.zip’ saved [28250958/28250958]

          Archive:  hourly_42602_2017.zip
            inflating: hourly_42602_2017.csv
```

```
In [ ]: !ls
```

```
In [ ]: # Load dataset
        aq_data = pd.read_csv('./hourly_42602_2017.csv')
        # aq_data = pd.read_csv('/Users/kkbankol@us.ibm.com/Downloads/hourly_42602_2017.csv')
```

```
In [ ]: # View first 5 rows
        aq_data.head()
```

```
In [ ]: # View titles of all columns
        aq_data.columns
```

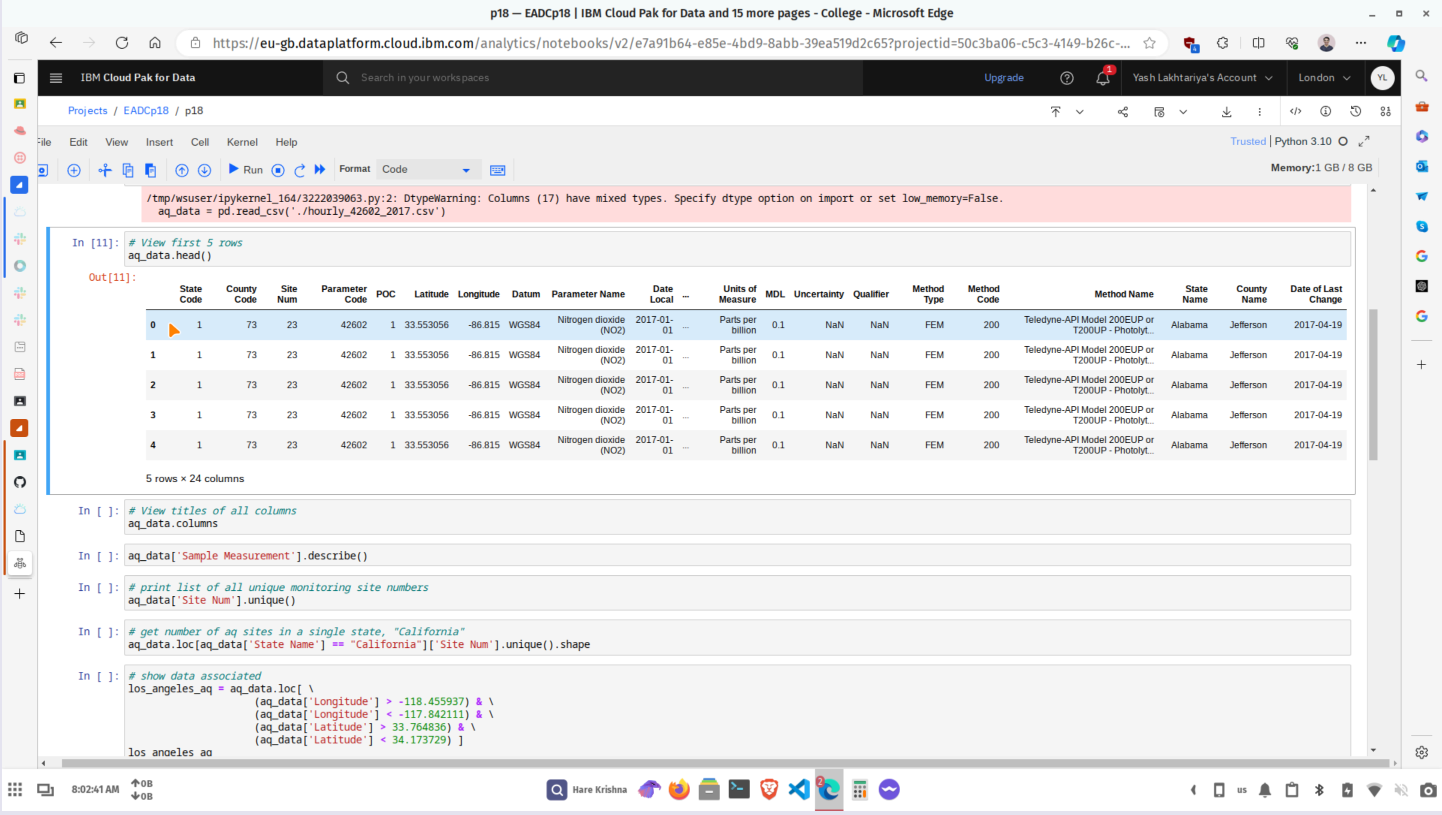
```
In [ ]: aq_data['Sample Measurement'].describe()
```

```
In [ ]: # print list of all unique monitoring site numbers
        aq_data['Site Num'].unique()
```

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA Batch - 61

EADC Practical 18

6. Check data imported from csv file downloaded via wget



p18 — EADCp18 | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge
https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/e7a91b64-e85e-4bd9-8abb-39ea519d2c65?projectid=50c3ba06-c5c3-4149-b26c-...
IBM Cloud Pak for Data
Search in your workspaces
Upgrade
Yash Lakhtariya's Account
London
YL
Projects / EADCp18 / p18
File Edit View Insert Cell Kernel Help
Run
Format
Code
Trusted | Python 3.10
Memory:1 GB / 8 GB

```
/tmp/wsuser/ipykernel_164/3222039063.py:2: DtypeWarning: Columns (17) have mixed types. Specify dtype option on import or set low_memory=False.
  aq_data = pd.read_csv('./hourly_42602_2017.csv')
```

In [11]:
```
# View first 5 rows
aq_data.head()
```

Out[11]:

	State Code	County Code	Site Num	Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Date Local	...	Units of Measure	MDL	Uncertainty	Qualifier	Method Type	Method Code	Method Name	State Name	County Name	Date of Last Change
0	1	73	23	42602	1	33.553056	-86.815	WGS84	Nitrogen dioxide (NO2)	2017-01-01	...	Parts per billion	0.1	NaN	NaN	FEM	200	Teledyne-API Model 200EUP or T200UP - Photolyt...	Alabama	Jefferson	2017-04-19
1	1	73	23	42602	1	33.553056	-86.815	WGS84	Nitrogen dioxide (NO2)	2017-01-01	...	Parts per billion	0.1	NaN	NaN	FEM	200	Teledyne-API Model 200EUP or T200UP - Photolyt...	Alabama	Jefferson	2017-04-19
2	1	73	23	42602	1	33.553056	-86.815	WGS84	Nitrogen dioxide (NO2)	2017-01-01	...	Parts per billion	0.1	NaN	NaN	FEM	200	Teledyne-API Model 200EUP or T200UP - Photolyt...	Alabama	Jefferson	2017-04-19
3	1	73	23	42602	1	33.553056	-86.815	WGS84	Nitrogen dioxide (NO2)	2017-01-01	...	Parts per billion	0.1	NaN	NaN	FEM	200	Teledyne-API Model 200EUP or T200UP - Photolyt...	Alabama	Jefferson	2017-04-19
4	1	73	23	42602	1	33.553056	-86.815	WGS84	Nitrogen dioxide (NO2)	2017-01-01	...	Parts per billion	0.1	NaN	NaN	FEM	200	Teledyne-API Model 200EUP or T200UP - Photolyt...	Alabama	Jefferson	2017-04-19

5 rows × 24 columns

In [ ]:
```
# View titles of all columns
aq_data.columns
```

In [ ]:
```
aq_data['Sample Measurement'].describe()
```

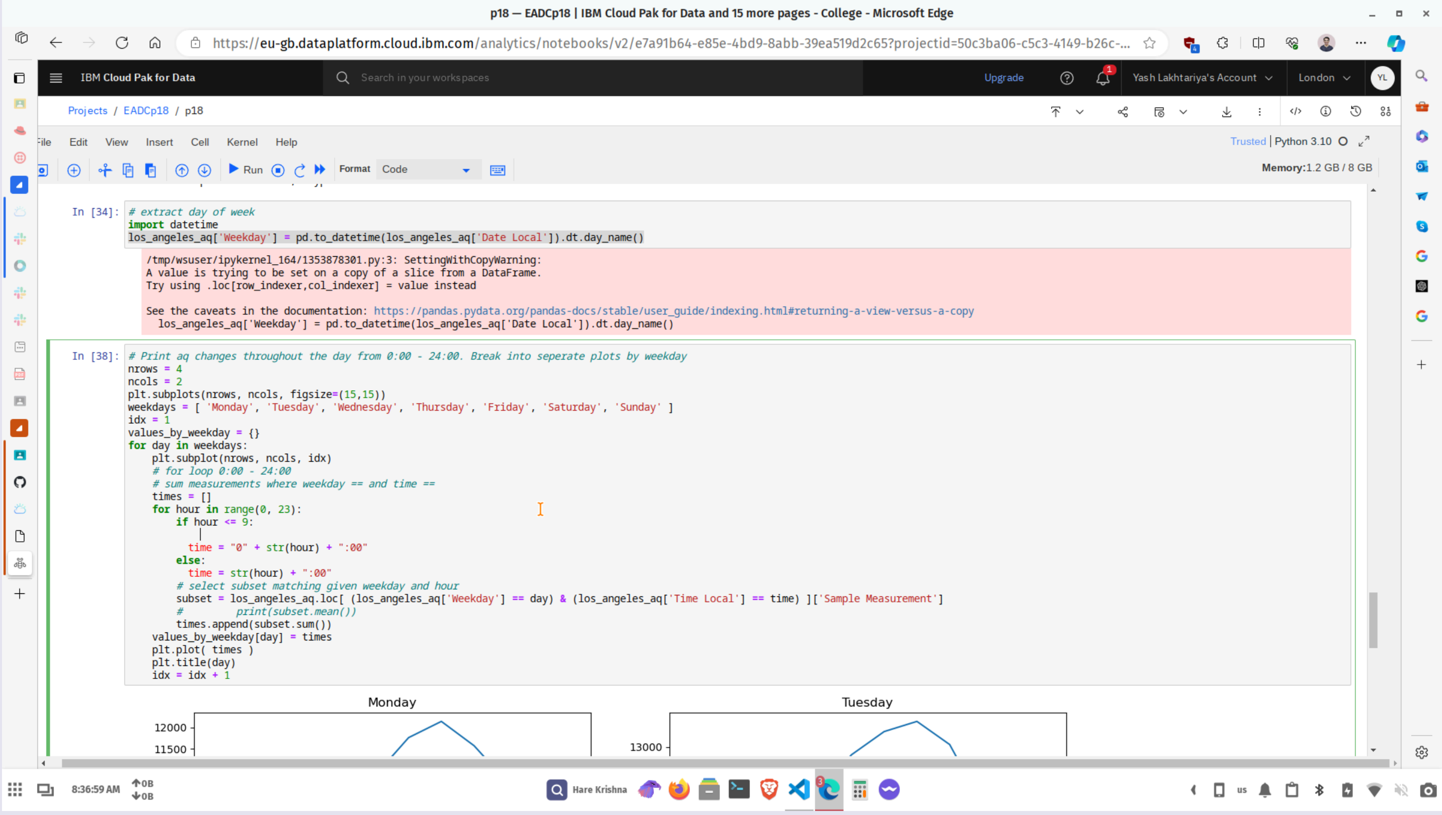
In [ ]:
```
# print list of all unique monitoring site numbers
aq_data['Site Num'].unique()
```

In [ ]:
```
# get number of aq sites in a single state, "California"
aq_data.loc[aq_data['State Name'] == "California"]['Site Num'].unique().shape
```

In [ ]:
```
# show data associated
los_angeles_aq = aq_data.loc[ \
                      (aq_data['Longitude'] > -118.455937) & \
                      (aq_data['Longitude'] < -117.842111) & \
                      (aq_data['Latitude'] > 33.764836) & \
                      (aq_data['Latitude'] < 34.173729) ]
los_angeles_aq
```

Name - Yash Lakhtariya
Enrollment number - 21162101012
Branch - CBA   Batch - 61
EADC Practical 18

7. Try adding columns derived from given columns of data, like Weekday column using Date given and getting day by datetime library, and plotting data day and timewise



p18 — EADCp18 | IBM Cloud Pak for Data and 15 more pages - College - Microsoft Edge

https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/e7a91b64-e85e-4bd9-8abb-39ea519d2c65?projectid=50c3ba06-c5c3-4149-b26c-...

```
In [34]: # extract day of week
         import datetime
         los_angeles_aq['Weekday'] = pd.to_datetime(los_angeles_aq['Date Local']).dt.day_name()
```

```
/tmp/wsuser/ipykernel_164/1353878301.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  los_angeles_aq['Weekday'] = pd.to_datetime(los_angeles_aq['Date Local']).dt.day_name()
```

```
In [38]: # Print aq changes throughout the day from 0:00 - 24:00. Break into seperate plots by weekday
         nrows = 4
         ncols = 2
         plt.subplots(nrows, ncols, figsize=(15,15))
         weekdays = [ 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday' ]
         idx = 1
         values_by_weekday = {}
         for day in weekdays:
             plt.subplot(nrows, ncols, idx)
             # for loop 0:00 - 24:00
             # sum measurements where weekday == and time ==
             times = []
             for hour in range(0, 23):
                 if hour <= 9:

                   time = "0" + str(hour) + ":00"
                 else:
                   time = str(hour) + ":00"
                 # select subset matching given weekday and hour
                 subset = los_angeles_aq.loc[ (los_angeles_aq['Weekday'] == day) & (los_angeles_aq['Time Local'] == time) ]['Sample Measurement']
                 #         print(subset.mean())
                 times.append(subset.sum())
             values_by_weekday[day] = times
             plt.plot( times )
             plt.title(day)
             idx = idx + 1
```