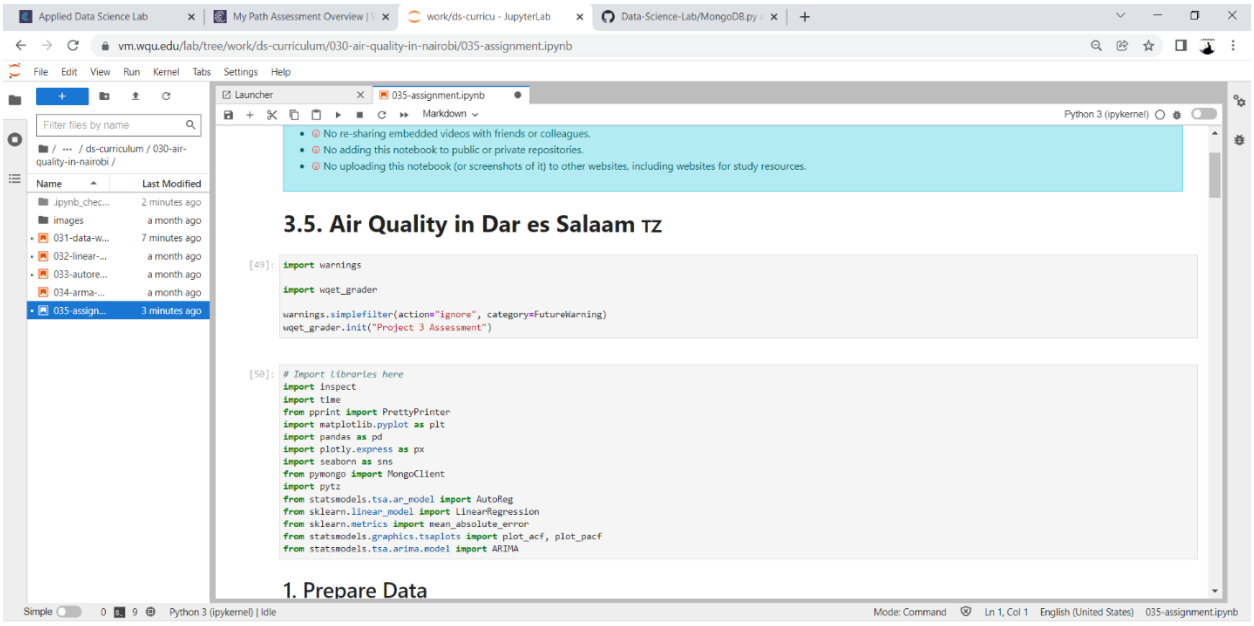
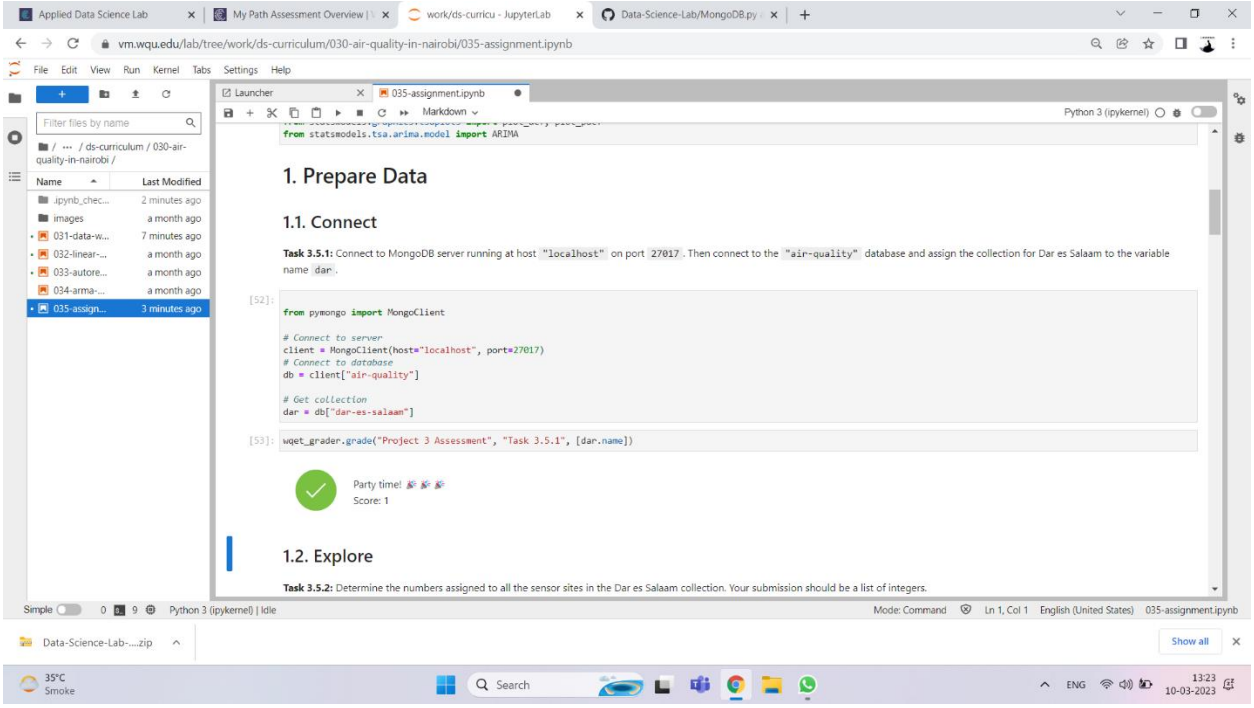
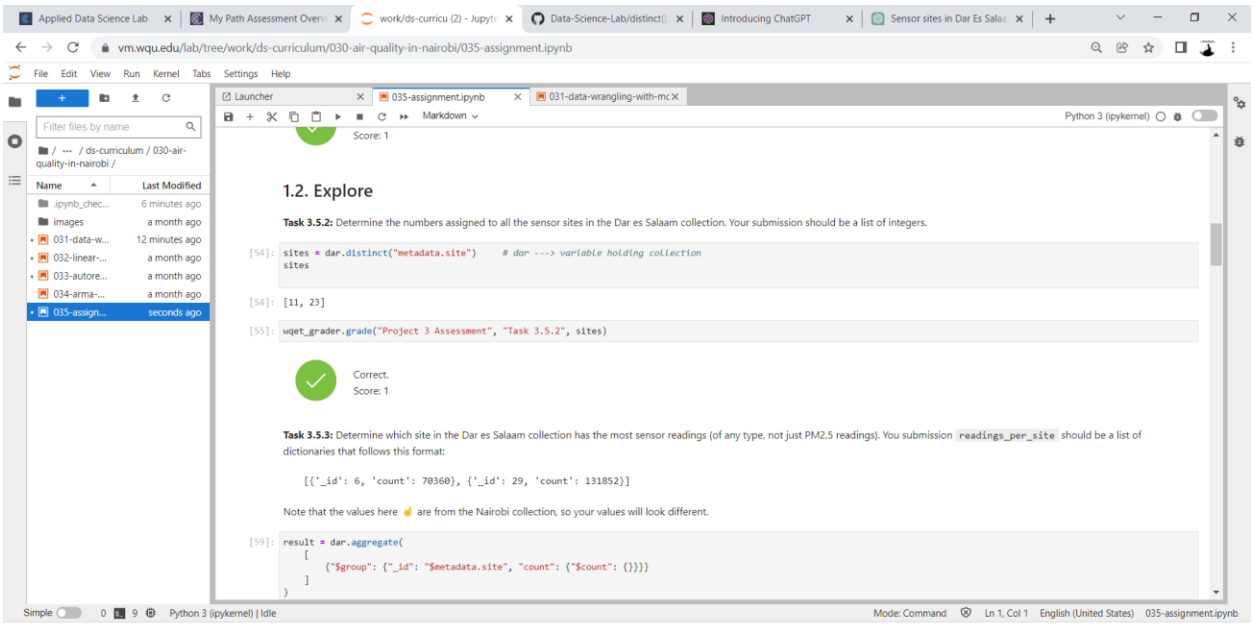
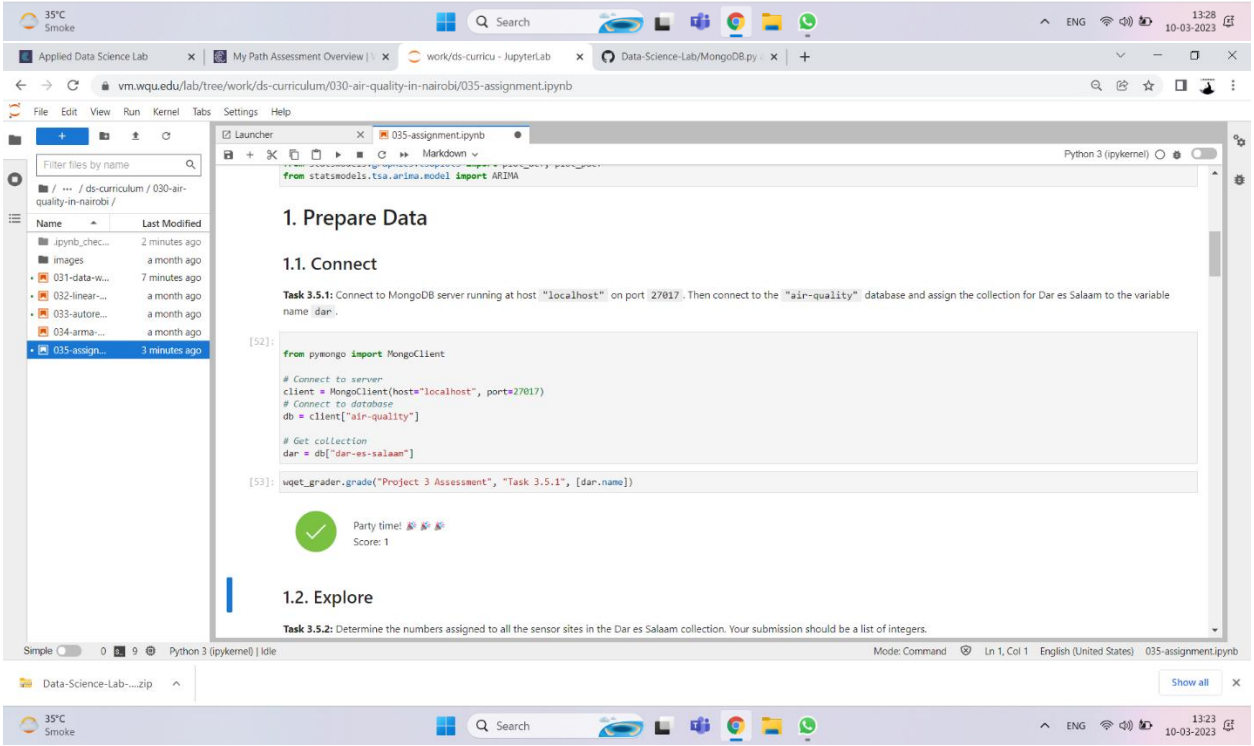
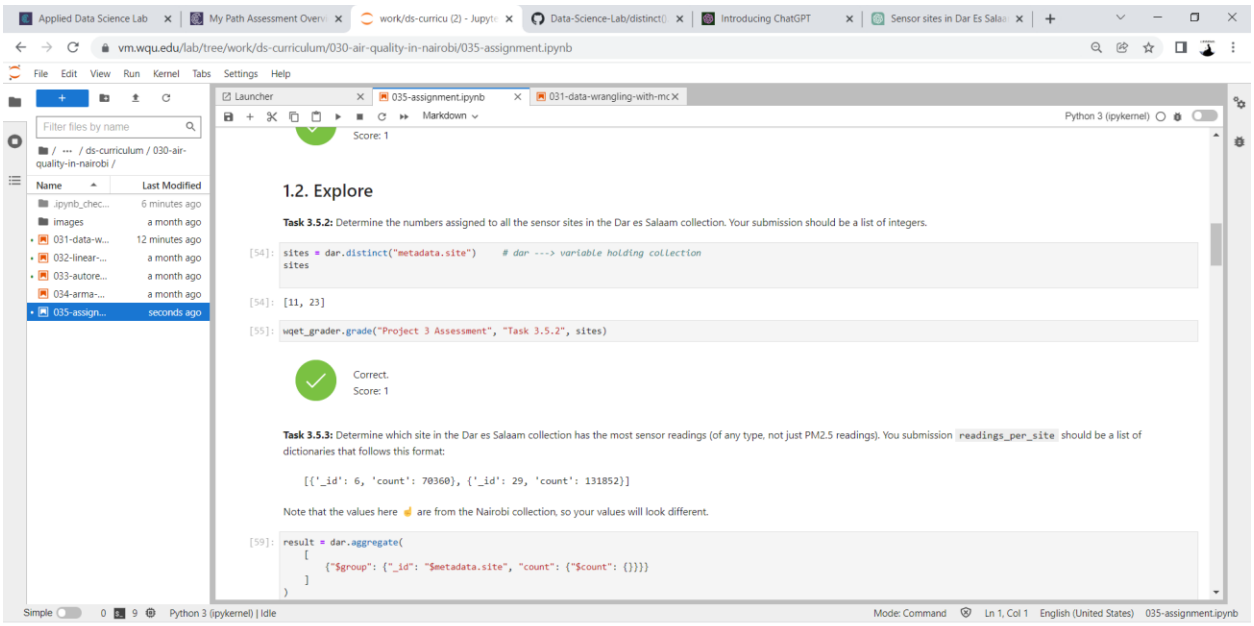
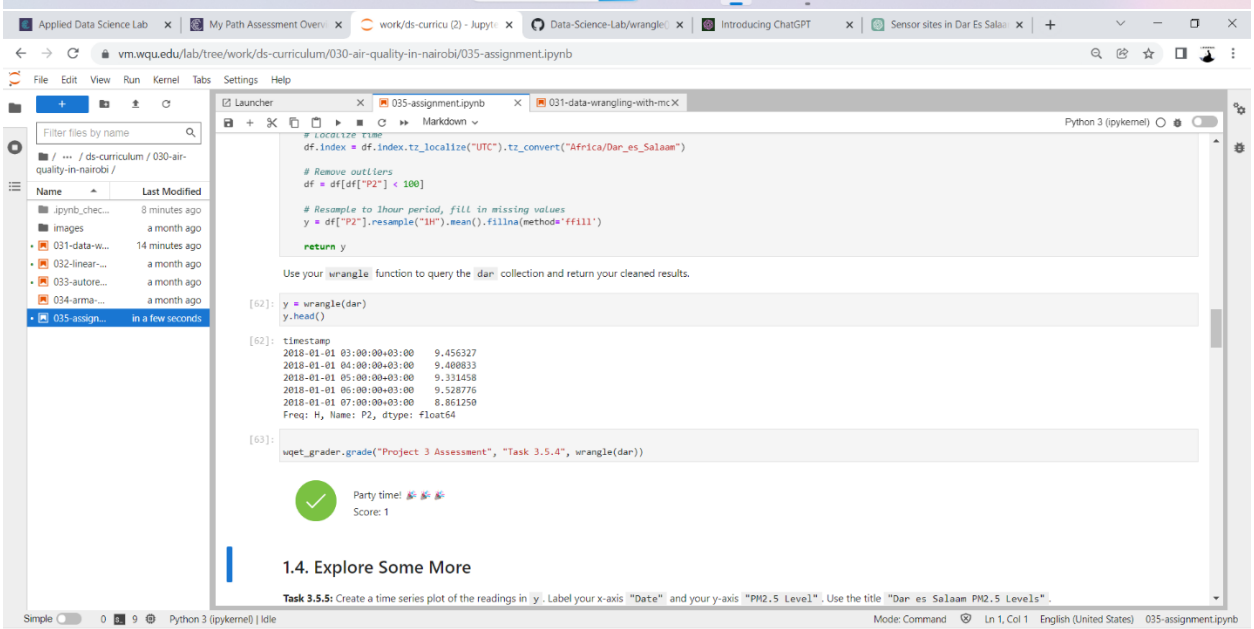

Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu - JupyterLab
Data-Science-Lab/MongoDB.py
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 2 minutes ago
images a month ago
031-data-w... 7 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... 3 minutes ago
No re-sharing embedded videos with friends or colleagues.
No adding this notebook to public or private repositories.
No uploading this notebook (or screenshots of it) to other websites, including websites for study resources.
3.5. Air Quality in Dar es Salaam TZ
[49]:
import warnings
import wqet_grader
warnings.simplefilter(action="ignore", category=FutureWarning)
wqet_grader.init("Project 3 Assessment")
[50]:
# Import libraries here
import inspect
import time
from pprint import PrettyPrinter
import matplotlib.pyplot as plt
import pandas as pd
import plotly.express as px
import seaborn as sns
from pymongo import MongoClient
import pytz
from statsmodels.tsa.ar_model import AutoReg
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
1. Prepare Data
Simple 0 9 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C Smoke
Search
ENG 13:23 10-03-2023

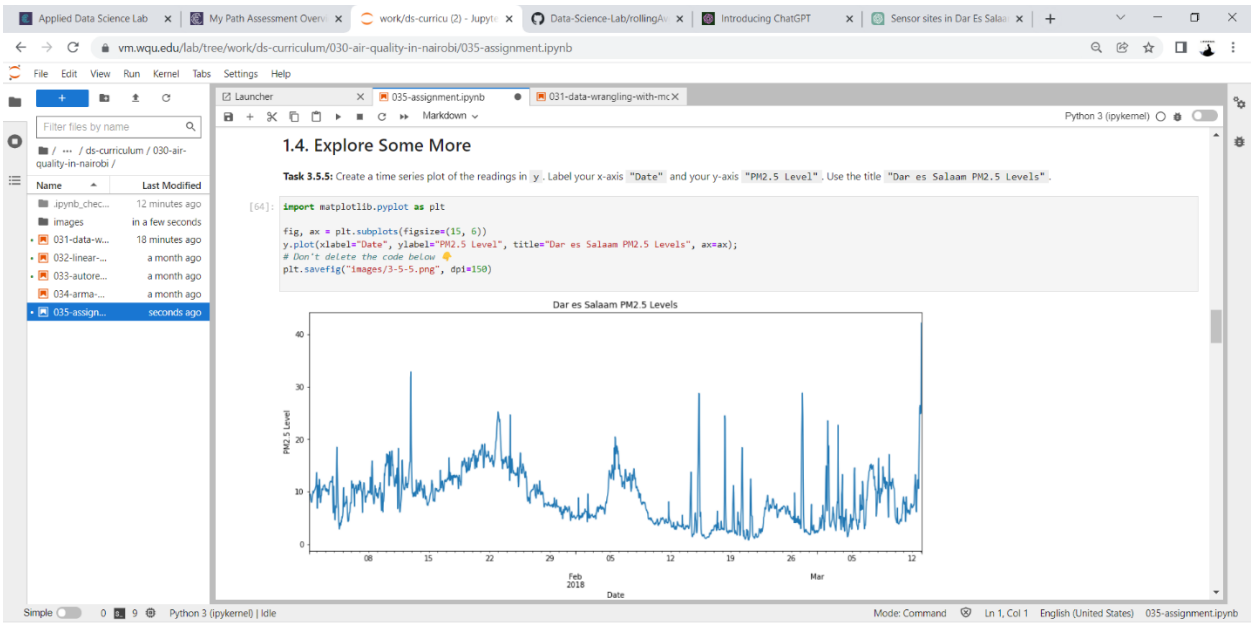

Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu - JupyterLab
Data-Science-Lab/MongoDB.py
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 2 minutes ago
images a month ago
031-data-w... 7 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... 3 minutes ago
from statsmodels.tsa.arima.model import ARIMA
1. Prepare Data
1.1. Connect
Task 3.5.1: Connect to MongoDB server running at host "localhost" on port 27017. Then connect to the "air-quality" database and assign the collection for Dar es Salaam to the variable name dar.
[52]:
from pymongo import MongoClient
# Connect to server
client = MongoClient(host="localhost", port=27017)
# Connect to database
db = client["air-quality"]
# Get collection
dar = db["dar-es-salaam"]
[53]:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.1", [dar.name])
Party time! 🎉🎉🎉
Score: 1
1.2. Explore
Task 3.5.2: Determine the numbers assigned to all the sensor sites in the Dar es Salaam collection. Your submission should be a list of integers.
Simple 0 9 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C Smoke
Search
ENG 13:23 10-03-2023


Applied Data Science Lab
My Path Assessment Overvi
work/ds-curricu (2) - Jupyt
Data-Science-Lab/distinct()
Introducing ChatGPT
Sensor sites in Dar Es Salaa
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 6 minutes ago
images a month ago
031-data-w... 12 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... seconds ago
Score: 1
1.2. Explore
Task 3.5.2: Determine the numbers assigned to all the sensor sites in the Dar es Salaam collection. Your submission should be a list of integers.
[54]:
sites = dar.distinct("metadata.site") # dar ---> variable holding collection
sites
[54]: [11, 23]
[55]:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.2", sites)
Correct.
Score: 1
Task 3.5.3: Determine which site in the Dar es Salaam collection has the most sensor readings (of any type, not just PM2.5 readings). You submission readings_per_site should be a list of dictionaries that follows this format:
[{'_id': 6, 'count': 70360}, {'_id': 29, 'count': 131852}]
Note that the values here ☝ are from the Nairobi collection, so your values will look different.
[59]:
result = dar.aggregate(
[
{"$group": {"_id": "$metadata.site", "count": {"$sum": {}}}}
]
)
Simple 0 9 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C Smoke
Search
ENG 13:28 10-03-2023


Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu - JupyterLab
Data-Science-Lab/MongoDB.py
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 2 minutes ago
images a month ago
031-data-w... 7 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... 3 minutes ago
from statsmodels.tsa.arima.model import ARIMA
1. Prepare Data
1.1. Connect
Task 3.5.1: Connect to MongoDB server running at host "localhost" on port 27017. Then connect to the "air-quality" database and assign the collection for Dar es Salaam to the variable name dar.
[52]:
from pymongo import MongoClient
# Connect to server
client = MongoClient(host="localhost", port=27017)
# Connect to database
db = client["air-quality"]
# Get collection
dar = db["dar-es-salaam"]
[53]:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.1", [dar.name])
Party time! 🎉🎉🎉
Score: 1
1.2. Explore
Task 3.5.2: Determine the numbers assigned to all the sensor sites in the Dar es Salaam collection. Your submission should be a list of integers.
Simple 0 9 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C Smoke
Search
ENG 13:23 10-03-2023


Applied Data Science Lab
My Path Assessment Overv
work/ds-curricu (2) - Jupyte
Data-Science-Lab/distinct()
Introducing ChatGPT
Sensor sites in Dar Es Salaa
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 6 minutes ago
images a month ago
031-data-w... 12 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... seconds ago
Score: 1
1.2. Explore
Task 3.5.2: Determine the numbers assigned to all the sensor sites in the Dar es Salaam collection. Your submission should be a list of integers.
[54]:
sites = dar.distinct("metadata.site") # dar ---> variable holding collection
sites
[54]: [11, 23]
[55]:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.2", sites)
Correct.
Score: 1
Task 3.5.3: Determine which site in the Dar es Salaam collection has the most sensor readings (of any type, not just PM2.5 readings). You submission readings_per_site should be a list of dictionaries that follows this format:
[{'_id': 6, 'count': 70360}, {'_id': 29, 'count': 131852}]
Note that the values here ☝ are from the Nairobi collection, so your values will look different.
[59]:
result = dar.aggregate(
[
{"$group": {"_id": "$metadata.site", "count": {"$count": {}}}}
]
)
Simple 0 9 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C Smoke
Search
ENG 13:28 10-03-2023


Applied Data Science Lab
My Path Assessment Overv
work/ds-curricu (2) - Jupyte
Data-Science-Lab/wrangle()
Introducing ChatGPT
Sensor sites in Dar Es Salaa
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 8 minutes ago
images a month ago
031-data-w... 14 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... in a few seconds
# Localize time
df.index = df.index.tz_localize("UTC").tz_convert("Africa/Dar_es_Salaam")
# Remove outliers
df = df[df["P2"] < 100]
# Resample to 1hour period, fill in missing values
y = df["P2"].resample("1H").mean().fillna(method='ffill')
return y
Use your wrangle function to query the dar collection and return your cleaned results.
[62]:
y = wrangle(dar)
y.head()
[62]:
timestamp
2018-01-01 03:00:00+03:00 9.456327
2018-01-01 04:00:00+03:00 9.400833
2018-01-01 05:00:00+03:00 9.331458
2018-01-01 06:00:00+03:00 9.528776
2018-01-01 07:00:00+03:00 8.861250
Freq: H, Name: P2, dtype: float64
[63]:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.4", wrangle(dar))
Party time! 🎉🎉🎉
Score: 1
1.4. Explore Some More
Task 3.5.5: Create a time series plot of the readings in y. Label your x-axis "Date" and your y-axis "PM2.5 Level". Use the title "Dar es Salaam PM2.5 Levels".
Simple 0 9 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C Smoke
Search
ENG 13:30 10-03-2023


Applied Data Science Lab
My Path Assessment Overv
work/ds-curricu (2) - Jupyte
Data-Science-Lab/rollingAv
Introducing ChatGPT
Sensor sites in Dar Es Salaa
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 12 minutes ago
images in a few seconds
031-data-w... 18 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... seconds ago
1.4. Explore Some More
Task 3.5.5: Create a time series plot of the readings in y. Label your x-axis "Date" and your y-axis "PM2.5 Level". Use the title "Dar es Salaam PM2.5 Levels".
[64]:
import matplotlib.pyplot as plt
fig, ax = plt.subplots(figsize=(15, 6))
y.plot(xlabel="Date", ylabel="PM2.5 Level", title="Dar es Salaam PM2.5 Levels", ax=ax);
# Don't delete the code below 👇
plt.savefig("images/3-5-5.png", dpi=150)
Dar es Salaam PM2.5 Levels
PM2.5 Level
Feb 2018
Mar
Date
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
035-assignment.ipynb
Data-Science-Lab-....zip
Show all


Applied Data Science Lab
My Path Assessment Overv
work/ds-curricu (2) - Jupyte
Data-Science-Lab/rollingAv
Introducing ChatGPT
Sensor sites in Dar Es Salaa
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 13 minutes ago
images seconds ago
031-data-w... 18 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... seconds ago
Task 3.5.6: Plot the rolling average of the readings in y. Use a window size of 168 (the number of hours in a week). Label your x-axis "Date" and your y-axis "PM2.5 Level". Use the title "Dar es Salaam PM2.5 Levels, 7-Day Rolling Average".
[67]:
import matplotlib.pyplot as plt
fig, ax = plt.subplots(figsize=(15, 6))
y.rolling(168).mean().plot(ax=ax, xlabel="Date", ylabel="PM2.5 Level", title="Dar es Salaam PM2.5 Levels, 7-Day Rolling Average");
# --> 168 == num of hours in a week
# Don't delete the code below 👇
plt.savefig("images/3-5-6.png", dpi=150)
Dar es Salaam PM2.5 Levels, 7-Day Rolling Average
PM2.5 Level
Feb 2018
Mar
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
035-assignment.ipynb
Data-Science-Lab-....zip
Show all

Applied Data Science Lab | My Path Assessment Overview | work/ds-curricu (2) - JupyterLab | Data-Science-Lab/PACFplot.py

vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

Launcher | 035-assignment.ipynb | 031-data-wrangling-with-mc

Code | Python 3 (ipykernel)

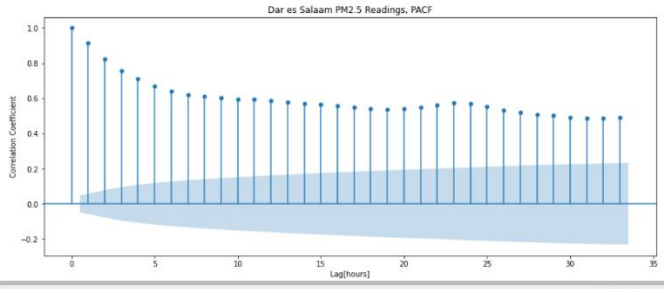
Filter files by name

/ ... / ds-curriculum / 030-air-quality-in-nairobi /

Name	Last Modified
.ipynb_chec...	20 minutes ago
images	seconds ago
031-data-w...	26 minutes ago
032-linear-...	a month ago
033-autore...	a month ago
034-arma-...	a month ago
035-assign...	a minute ago

Task 3.5.7: Create an ACF plot for the data in `y`. Be sure to label the x-axis as `"Lag [hours]"` and the y-axis as `"Correlation Coefficient"`. Use the title `"Dar es Salaam PM2.5 Readings, ACF"`.

[71]:
```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(15, 6))
plot_acf(y, ax=ax)
plt.xlabel("Lag[hours]")
plt.ylabel("Correlation Coefficient")
plt.title("Dar es Salaam PM2.5 Readings, PACF");
# Don't delete the code below 👇
plt.savefig("images/3-5-7.png", dpi=150)
```



Simple | 0 | 9 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 035-assignment.ipynb

Data-Science-Lab-....zip | Show all

35°C Smoke | Search | ENG | 13:41 10-03-2023

Applied Data Science Lab | My Path Assessment Overview | work/ds-curricu (2) - JupyterLab | Data-Science-Lab/PACFplot.py

vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

Launcher | 035-assignment.ipynb | 031-data-wrangling-with-mc

Markdown | Python 3 (ipykernel)

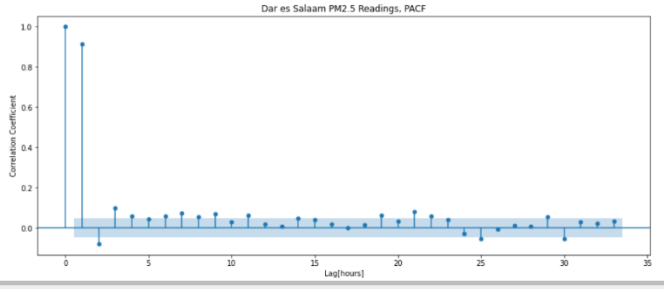
Filter files by name

/ ... / ds-curriculum / 030-air-quality-in-nairobi /

Name	Last Modified
.ipynb_chec...	20 minutes ago
images	seconds ago
031-data-w...	26 minutes ago
032-linear-...	a month ago
033-autore...	a month ago
034-arma-...	a month ago
035-assign...	2 minutes ago

Task 3.5.8: Create an PACF plot for the data in `y`. Be sure to label the x-axis as `"Lag [hours]"` and the y-axis as `"Correlation Coefficient"`. Use the title `"Dar es Salaam PM2.5 Readings, PACF"`.

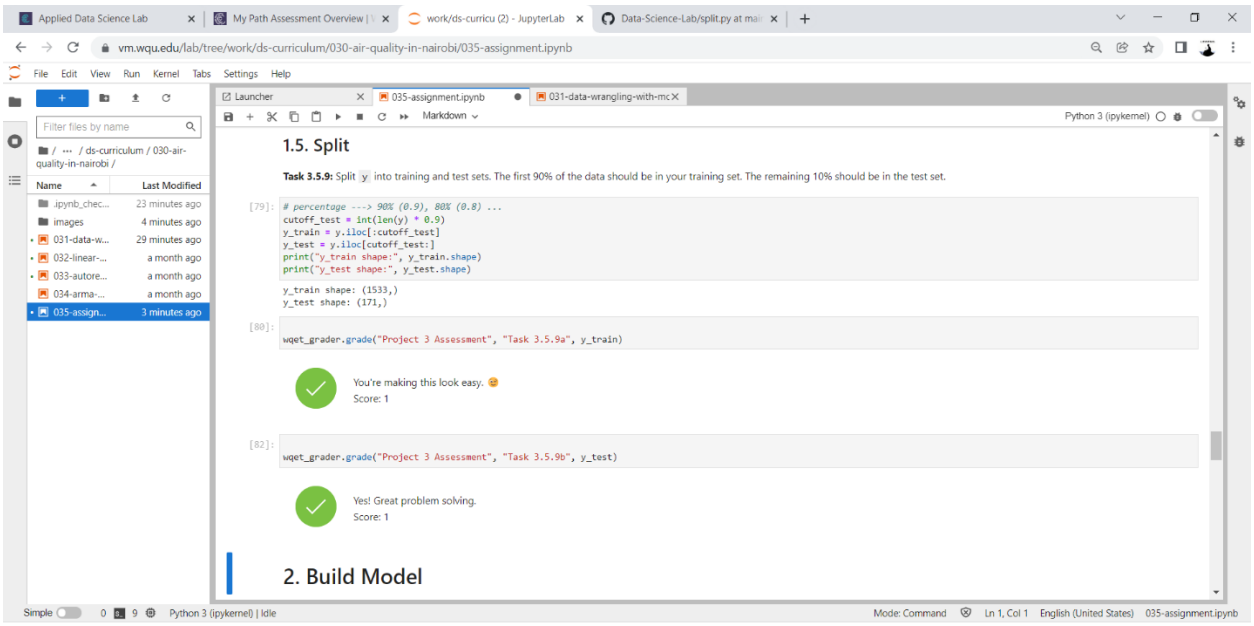
[73]:
```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(15, 6))
plot_pacf(y, ax=ax)  # -----> line showing difference from acf plot <-----
plt.xlabel("Lag[hours]")
plt.ylabel("Correlation Coefficient")
plt.title("Dar es Salaam PM2.5 Readings, PACF");
# Don't delete the code below 👇
plt.savefig("images/3-5-8.png", dpi=150)
```



Simple | 0 | 9 | Python 3 (ipykernel) | Idle | Saving started | Mode: Command | Ln 1, Col 1 | English (United States) | 035-assignment.ipynb

Data-Science-Lab-....zip | Show all

35°C Smoke | Search | ENG | 13:42 10-03-2023



1.5. Split

Task 3.5.9: Split `y` into training and test sets. The first 90% of the data should be in your training set. The remaining 10% should be in the test set.

```
[79]: # percentage ---> 90% (0.9), 80% (0.8) ...
cutoff_test = int(len(y) * 0.9)
y_train = y.iloc[:cutoff_test]
y_test = y.iloc[cutoff_test:]
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

```
y_train shape: (1533,)
y_test shape: (171,)
```

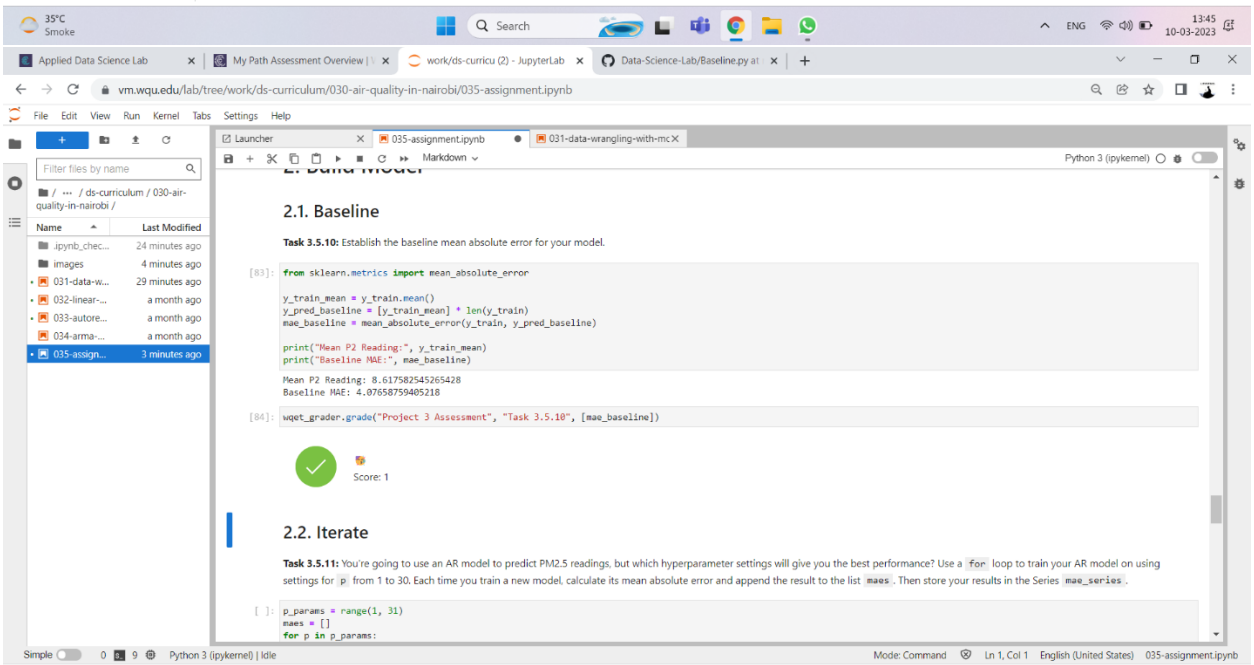
```
[80]: wqet_grader.grade("Project 3 Assessment", "Task 3.5.9a", y_train)
```

You're making this look easy. 😉
Score: 1

```
[82]: wqet_grader.grade("Project 3 Assessment", "Task 3.5.9b", y_test)
```

Yes! Great problem solving.
Score: 1

2. Build Model



2.1. Baseline

Task 3.5.10: Establish the baseline mean absolute error for your model.

```
[83]: from sklearn.metrics import mean_absolute_error

y_train_mean = y_train.mean()
y_pred_baseline = [y_train_mean] * len(y_train)
mae_baseline = mean_absolute_error(y_train, y_pred_baseline)

print("Mean P2 Reading:", y_train_mean)
print("Baseline MAE:", mae_baseline)
```

```
Mean P2 Reading: 8.617582545265428
Baseline MAE: 4.07658759405218
```

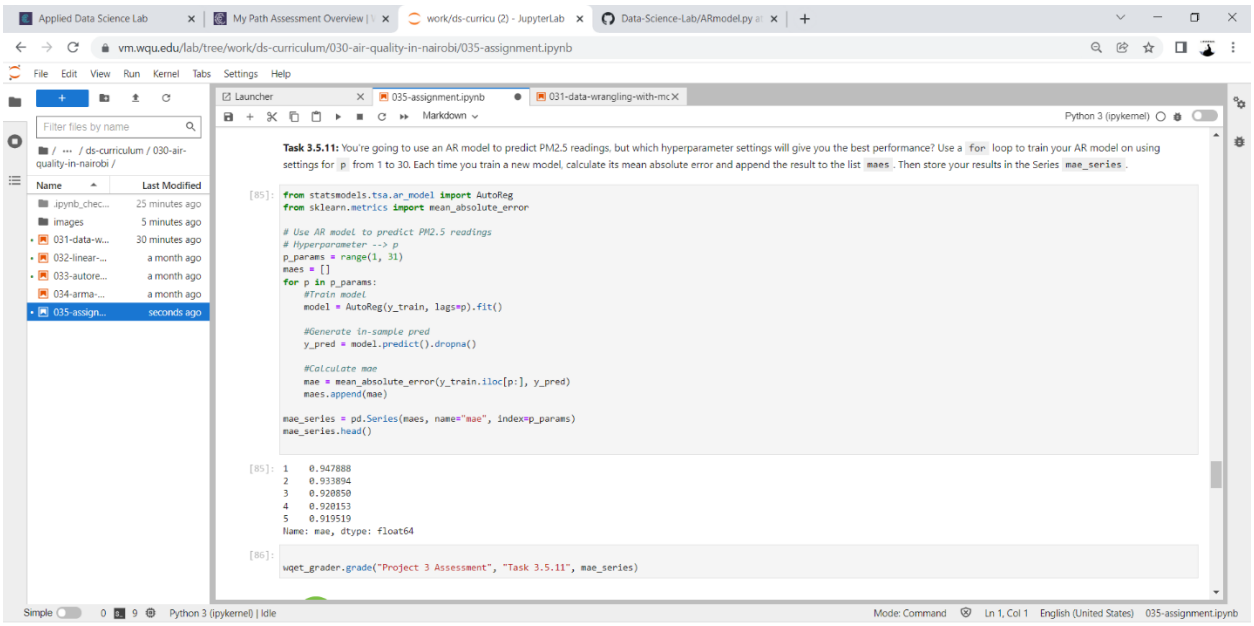
```
[84]: wqet_grader.grade("Project 3 Assessment", "Task 3.5.10", [mae_baseline])
```

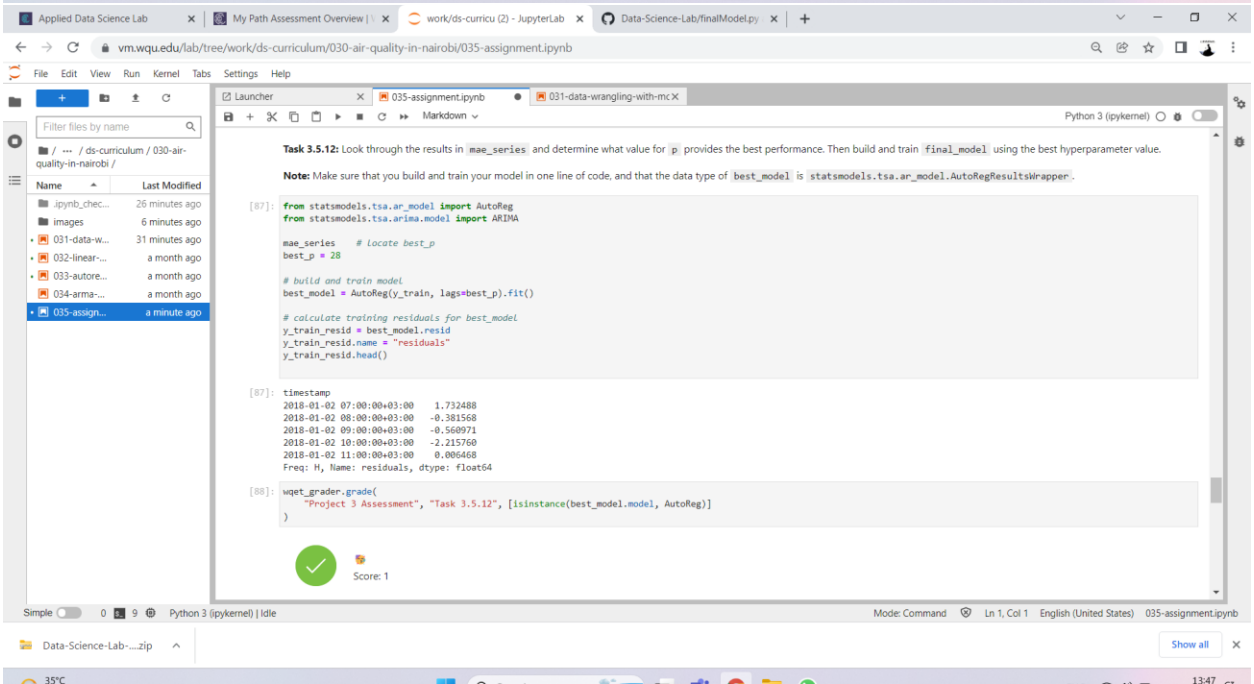
Score: 1

2.2. Iterate

Task 3.5.11: You're going to use an AR model to predict PM2.5 readings, but which hyperparameter settings will give you the best performance? Use a `for` loop to train your AR model on using settings for `p` from 1 to 30. Each time you train a new model, calculate its mean absolute error and append the result to the list `maes`. Then store your results in the Series `mae_series`.

```
[ ]: p_params = range(1, 31)
maes = []
for p in p_params:
```


Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu (2) - JupyterLab
Data-Science-Lab/ARmodel.py
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 25 minutes ago
images 5 minutes ago
031-data-w... 30 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... seconds ago
Task 3.5.11: You're going to use an AR model to predict PM2.5 readings, but which hyperparameter settings will give you the best performance? Use a for loop to train your AR model on using settings for p from 1 to 30. Each time you train a new model, calculate its mean absolute error and append the result to the list maes. Then store your results in the Series mae_series.
[85]:
from statsmodels.tsa.ar_model import AutoReg
from sklearn.metrics import mean_absolute_error
# Use AR model to predict PM2.5 readings
# Hyperparameter --> p
p_params = range(1, 31)
maes = []
for p in p_params:
#Train model
model = AutoReg(y_train, lags=p).fit()
#Generate in-sample pred
y_pred = model.predict().dropna()
#Calculate mae
mae = mean_absolute_error(y_train.iloc[p:], y_pred)
maes.append(mae)
mae_series = pd.Series(maes, name="mae", index=p_params)
mae_series.head()
[85]:
1 0.947888
2 0.933894
3 0.920850
4 0.920153
5 0.919519
Name: mae, dtype: float64
[86]:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.11", mae_series)
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C
Smoke
Search
ENG
13:46
10-03-2023


Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu (2) - JupyterLab
Data-Science-Lab/finalModel.py
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
031-data-wrangling-with-mc
Markdown
Python 3 (ipykernel)
Filter files by name
/ ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... 26 minutes ago
images 6 minutes ago
031-data-w... 31 minutes ago
032-linear-... a month ago
033-autore... a month ago
034-arma-... a month ago
035-assign... a minute ago
Task 3.5.12: Look through the results in mae_series and determine what value for p provides the best performance. Then build and train final_model using the best hyperparameter value.
Note: Make sure that you build and train your model in one line of code, and that the data type of best_model is statsmodels.tsa.ar_model.AutoRegResultsWrapper.
[87]:
from statsmodels.tsa.ar_model import AutoReg
from statsmodels.tsa.arima.model import ARIMA
mae_series # Locate best_p
best_p = 28
# build and train model
best_model = AutoReg(y_train, lags=best_p).fit()
# calculate training residuals for best_model
y_train_resid = best_model.resid
y_train_resid.name = "residuals"
y_train_resid.head()
[87]:
timestamp
2018-01-02 07:00:00+03:00 1.732488
2018-01-02 08:00:00+03:00 -0.381568
2018-01-02 09:00:00+03:00 -0.560971
2018-01-02 10:00:00+03:00 -2.215760
2018-01-02 11:00:00+03:00 0.006468
Freq: H, Name: residuals, dtype: float64
[88]:
wqet_grader.grade(
"Project 3 Assessment", "Task 3.5.12", [isinstance(best_model.model, AutoReg)]
)
Score: 1
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
035-assignment.ipynb
Data-Science-Lab-....zip
Show all
35°C
Smoke
Search
ENG
13:47
10-03-2023

Applied Data Science Lab | My Path Assessment Overview | work/ds-curricu (2) - JupyterLab | Data-Science-Lab/finalModel.py

vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

Launcher | 035-assignment.ipynb | 031-data-wrangling-with-mc

Markdown | Python 3 (ipykernel)

Filter files by name

/ ... / ds-curriculum / 030-air-quality-in-nairobi /

Name	Last Modified
.ipynb_chec...	27 minutes ago
images	7 minutes ago
031-data-w...	33 minutes ago
032-linear-...	a month ago
033-autore...	a month ago
034-arma-...	a month ago
035-assign...	seconds ago

Score: 1

Task 3.5.13: Calculate the training residuals for `best_model` and assign the result to `y_train_resid`. **Note** that your name of your Series should be `"residuals"`.

[91]:
```
y_train_resid = best_model.resid
y_train_resid.name = "residuals"
y_train_resid.head()
```

[91]:
```
timestamp
2018-01-02 07:00:00+03:00    1.732488
2018-01-02 08:00:00+03:00   -0.381568
2018-01-02 09:00:00+03:00   -0.560971
2018-01-02 10:00:00+03:00   -2.215760
2018-01-02 11:00:00+03:00    0.006468
Freq: H, Name: residuals, dtype: float64
```

[92]:
```
wqet_grader.grade("Project 3 Assessment", "Task 3.5.13", y_train_resid.tail(1500))
```

That's the right answer. Keep it up!
Score: 1

Task 3.5.14: Create a histogram of `y_train_resid`. Be sure to label the x-axis as `"Residuals"` and the y-axis as `"Frequency"`. Use the title `"Best Model, Training Residuals"`.

[ ]:
```
# Plot histogram of residuals

# Don't delete the code below 👇
plt.savefig("images/3-5-14.png", dpi=150)
```

Simple | 0 | 9 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 035-assignment.ipynb

Data-Science-Lab-....zip | Show all

35°C Smoke | Search | ENG | 13:49 10-03-2023

Applied Data Science Lab | My Path Assessment Overview | work/ds-curricu (2) - JupyterLab | Data-Science-Lab/030-air-quali

vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

Launcher | 035-assignment.ipynb | 033-autoregressive-models.i

Markdown | Python 3 (ipykernel)

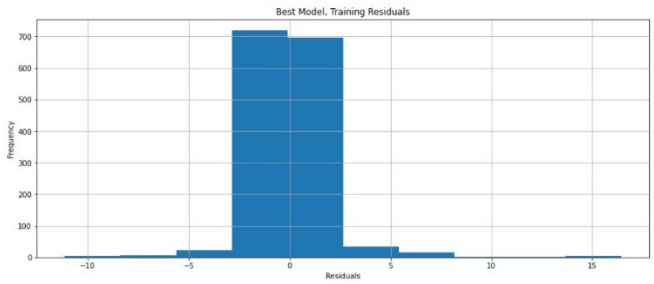
Filter files by name

/ ... / ds-curriculum / 030-air-quality-in-nairobi /

Name	Last Modified
.ipynb_chec...	an hour ago
images	a minute ago
031-data-w...	an hour ago
032-linear-...	a month ago
033-autore...	3 minutes ago
034-arma-...	a month ago
035-assign...	seconds ago

Task 3.5.14: Create a histogram of `y_train_resid`. Be sure to label the x-axis as `"Residuals"` and the y-axis as `"Frequency"`. Use the title `"Best Model, Training Residuals"`.

[106]:
```
fig, ax = plt.subplots(figsize=(15, 6))
# Plot histogram of residuals
y_train_resid.hist()
plt.xlabel("Residuals")
plt.ylabel("Frequency")
plt.title("Best Model, Training Residuals");
# Don't delete the code below 👇
plt.savefig("images/3-5-14.png", dpi=150)
```



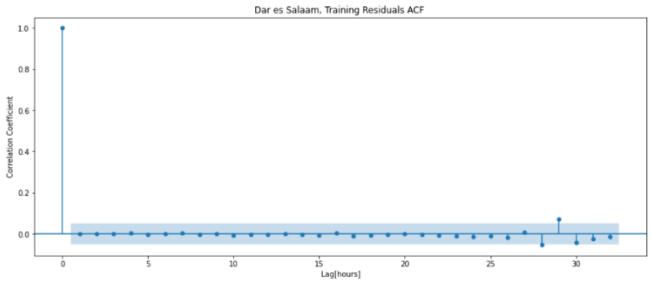
Simple | 0 | 10 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 035-assignment.ipynb

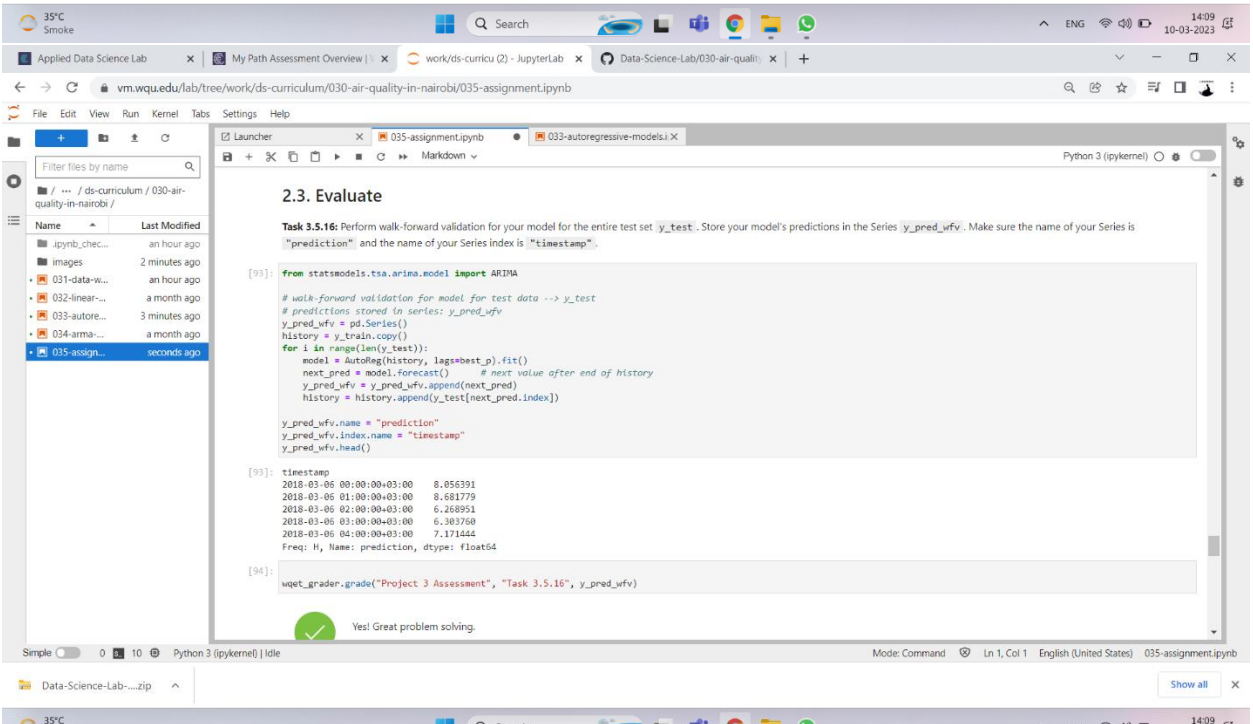
Data-Science-Lab-....zip | Show all

35°C Smoke | Search | ENG | 14:09 10-03-2023

Residuals ACF".

```
[100]: from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
       import matplotlib.pyplot as plt

       fig, ax = plt.subplots(figsize=(15, 6))
       plot_acf(y_train_resid, ax=ax)
       plt.xlabel("Lag[hours]")
       plt.ylabel("Correlation Coefficient")
       plt.title("Dar es Salaam, Training Residuals ACF ");
       # Don't delete the code below 👇
       plt.savefig("images/3-5-15.png", dpi=150)
```





2.3. Evaluate

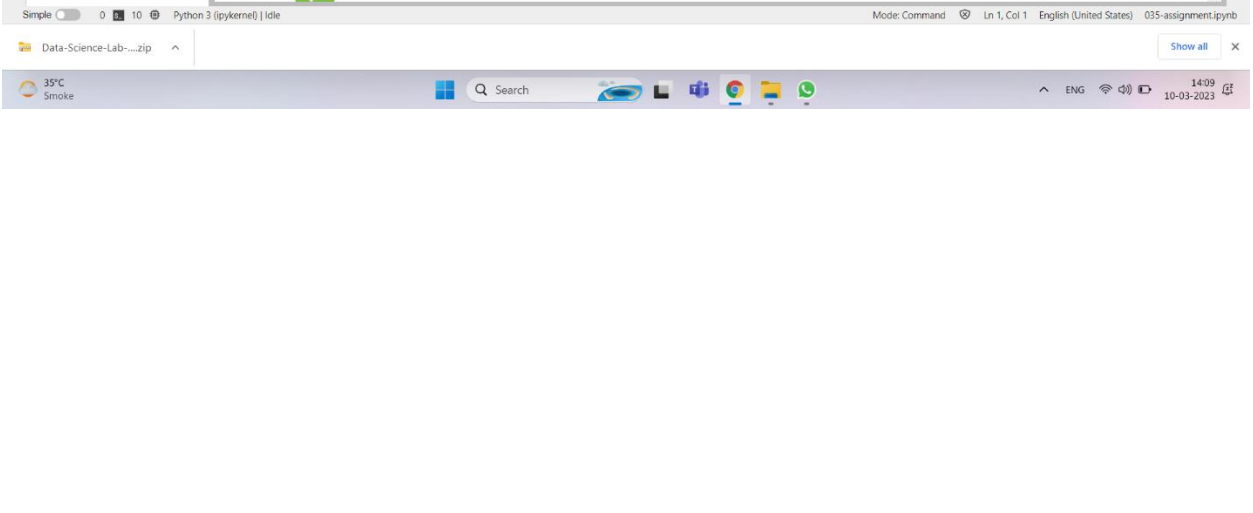
Task 3.5.16: Perform walk-forward validation for your model for the entire test set `y_test`. Store your model's predictions in the Series `y_pred_wfv`. Make sure the name of your Series is `"prediction"` and the name of your Series index is `"timestamp"`.

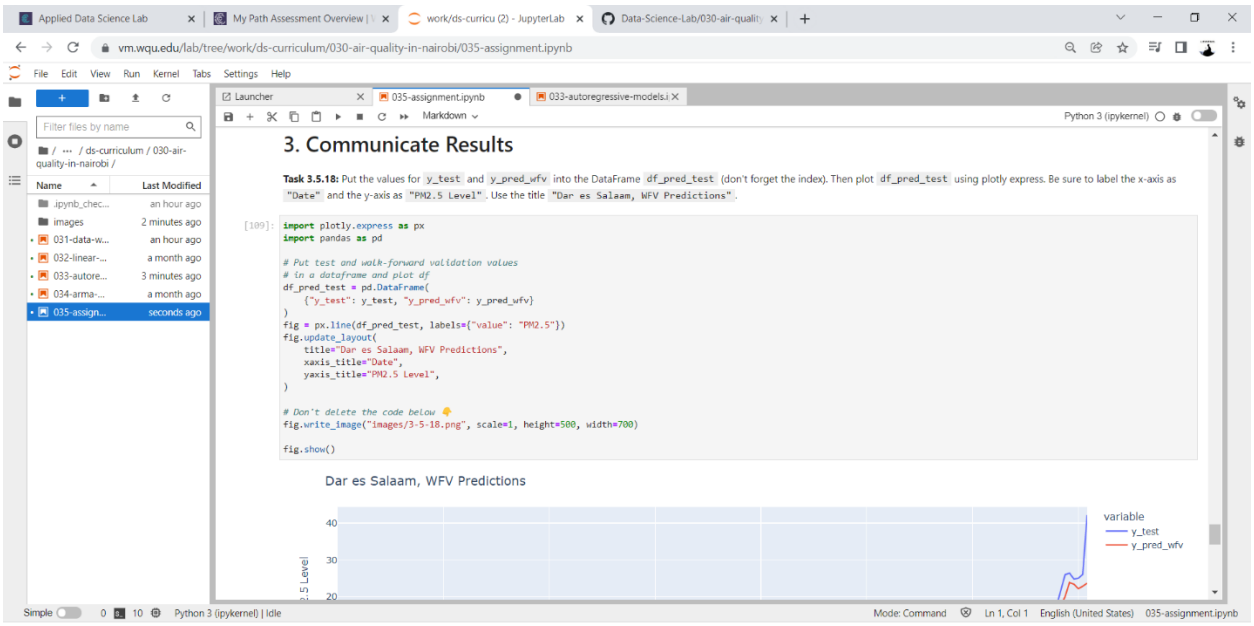
```
[93]: from statsmodels.tsa.arima.model import ARIMA

      # walk-forward validation for model for test data --> y_test
      # predictions stored in series: y_pred_wfv
      y_pred_wfv = pd.Series()
      history = y_train.copy()
      for i in range(len(y_test)):
          model = AutoReg(history, lags=best_p).fit()
          next_pred = model.forecast()      # next value after end of history
          y_pred_wfv = y_pred_wfv.append(next_pred)
          history = history.append(y_test[next_pred.index])

      y_pred_wfv.name = "prediction"
      y_pred_wfv.index.name = "timestamp"
      y_pred_wfv.head()
```

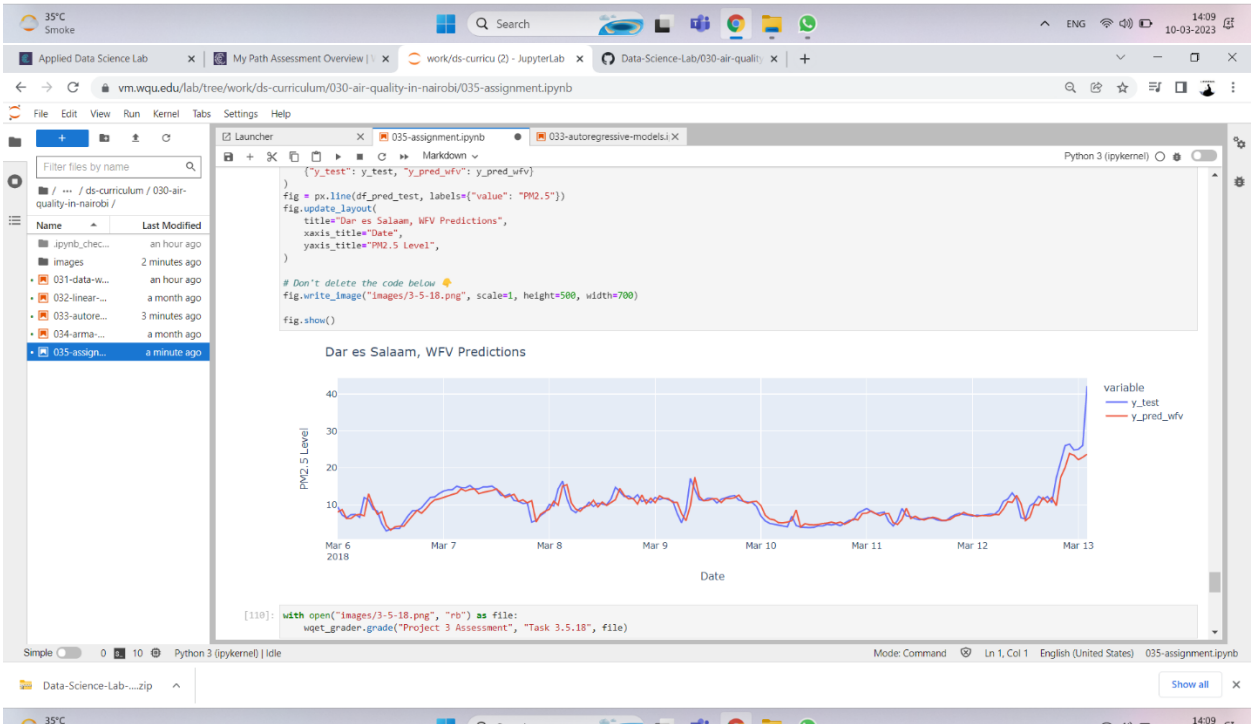
```
[93]: timestamp
      2018-03-06 00:00:00+03:00    8.056391
      2018-03-06 01:00:00+03:00    8.681779
      2018-03-06 02:00:00+03:00    6.268951
      2018-03-06 03:00:00+03:00    6.303760
      2018-03-06 04:00:00+03:00    7.171444
      Freq: H, Name: prediction, dtype: float64
```

```
[94]: wqet_grader.grade("Project 3 Assessment", "Task 3.5.16", y_pred_wfv)
```

Yes! Great problem solving.




Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu (2) - JupyterLab
Data-Science-Lab/030-air-quality
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
033-autoregressive-models.i
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... an hour ago
images 2 minutes ago
031-data-w... an hour ago
032-linear-... a month ago
033-autore... 3 minutes ago
034-arma-... a month ago
035-assign... seconds ago
3. Communicate Results
Task 3.5.18: Put the values for y_test and y_pred_wfv into the DataFrame df_pred_test (don't forget the index). Then plot df_pred_test using plotly express. Be sure to label the x-axis as "Date" and the y-axis as "PM2.5 Level". Use the title "Dar es Salaam, WFV Predictions".
[109]:
import plotly.express as px
import pandas as pd
# Put test and walk-forward validation values
# in a dataframe and plot df
df_pred_test = pd.DataFrame(
{"y_test": y_test, "y_pred_wfv": y_pred_wfv}
)
fig = px.line(df_pred_test, labels={"value": "PM2.5"})
fig.update_layout(
title="Dar es Salaam, WFV Predictions",
xaxis_title="Date",
yaxis_title="PM2.5 Level",
)
# Don't delete the code below 👇
fig.write_image("images/3-5-18.png", scale=1, height=500, width=700)
fig.show()
Dar es Salaam, WFV Predictions
variable
y_test
y_pred_wfv
40
30
20
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
035-assignment.ipynb
Data-Science-Lab-....zip
Show all


Applied Data Science Lab
My Path Assessment Overview
work/ds-curricu (2) - JupyterLab
Data-Science-Lab/030-air-quality
vm.wqu.edu/lab/tree/work/ds-curriculum/030-air-quality-in-nairobi/035-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
035-assignment.ipynb
033-autoregressive-models.i
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 030-air-quality-in-nairobi /
Name
Last Modified
.ipynb_chec... an hour ago
images 2 minutes ago
031-data-w... an hour ago
032-linear-... a month ago
033-autore... 3 minutes ago
034-arma-... a month ago
035-assign... a minute ago
{"y_test": y_test, "y_pred_wfv": y_pred_wfv}
)
fig = px.line(df_pred_test, labels={"value": "PM2.5"})
fig.update_layout(
title="Dar es Salaam, WFV Predictions",
xaxis_title="Date",
yaxis_title="PM2.5 Level",
)
# Don't delete the code below 👇
fig.write_image("images/3-5-18.png", scale=1, height=500, width=700)
fig.show()
Dar es Salaam, WFV Predictions
variable
y_test
y_pred_wfv
PM2.5 Level
40
30
20
10
Mar 6
2018
Mar 7
Mar 8
Mar 9
Mar 10
Mar 11
Mar 12
Mar 13
Date
[110]:
with open("images/3-5-18.png", "rb") as file:
wqet_grader.grade("Project 3 Assessment", "Task 3.5.18", file)
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
035-assignment.ipynb
Data-Science-Lab-....zip
Show all