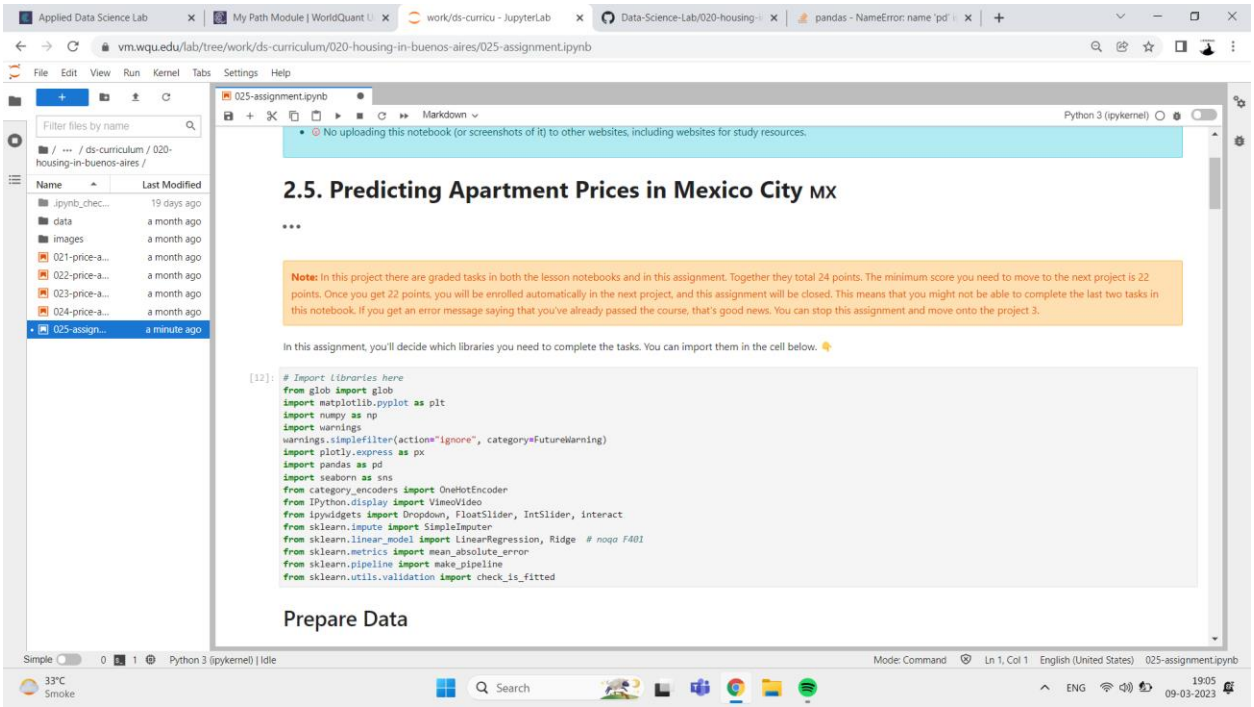
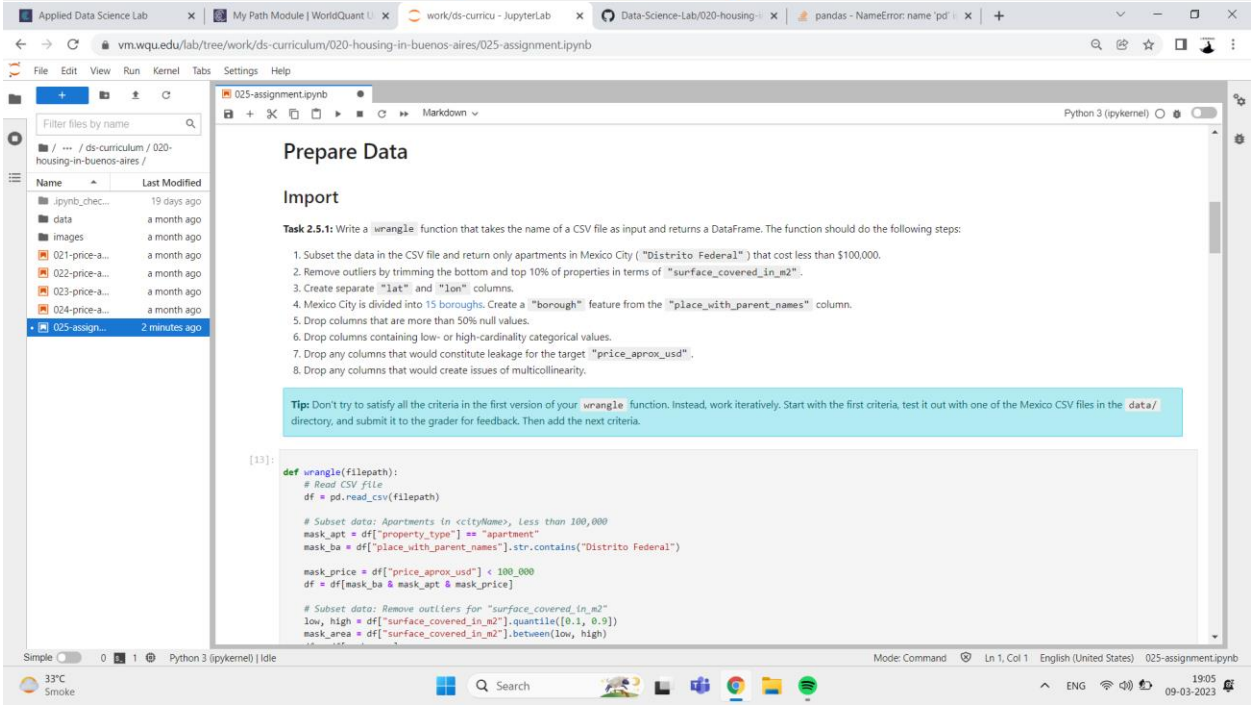

Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/020-housing-
pandas - NameError: name 'pd'
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images a month ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... a minute ago
No uploading this notebook (or screenshots of it) to other websites, including websites for study resources.
2.5. Predicting Apartment Prices in Mexico City MX
...
Note: In this project there are graded tasks in both the lesson notebooks and in this assignment. Together they total 24 points. The minimum score you need to move to the next project is 22 points. Once you get 22 points, you will be enrolled automatically in the next project, and this assignment will be closed. This means that you might not be able to complete the last two tasks in this notebook. If you get an error message saying that you've already passed the course, that's good news. You can stop this assignment and move onto the project 3.
In this assignment, you'll decide which libraries you need to complete the tasks. You can import them in the cell below.
[12]:
# Import libraries here
from glob import glob
import matplotlib.pyplot as plt
import numpy as np
import warnings
warnings.simplefilter(action="ignore", category=FutureWarning)
import plotly.express as px
import pandas as pd
import seaborn as sns
from category_encoders import OneHotEncoder
from IPython.display import VimeoVideo
from ipywidgets import Dropdown, FloatSlider, IntSlider, interact
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LinearRegression, Ridge # noqa F401
from sklearn.metrics import mean_absolute_error
from sklearn.pipeline import make_pipeline
from sklearn.utils.validation import check_is_fitted
Prepare Data
Simple 0 1 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 025-assignment.ipynb
33°C Smoke
Search
ENG 19:05 09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/020-housing-
pandas - NameError: name 'pd'
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images a month ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... 2 minutes ago
Prepare Data
Import
Task 2.5.1: Write a wrangle function that takes the name of a CSV file as input and returns a DataFrame. The function should do the following steps:
1. Subset the data in the CSV file and return only apartments in Mexico City ("Distrito Federal") that cost less than $100,000.
2. Remove outliers by trimming the bottom and top 10% of properties in terms of "surface_covered_in_m2".
3. Create separate "lat" and "lon" columns.
4. Mexico City is divided into 15 boroughs. Create a "borough" feature from the "place_with_parent_names" column.
5. Drop columns that are more than 50% null values.
6. Drop columns containing low- or high-cardinality categorical values.
7. Drop any columns that would constitute leakage for the target "price_aprox_usd".
8. Drop any columns that would create issues of multicollinearity.
Tip: Don't try to satisfy all the criteria in the first version of your wrangle function. Instead, work iteratively. Start with the first criteria, test it out with one of the Mexico CSV files in the data/ directory, and submit it to the grader for feedback. Then add the next criteria.
[13]:
def wrangle(filepath):
# Read CSV file
df = pd.read_csv(filepath)
# Subset data: Apartments in <cityName>, Less than 100,000
mask_apt = df["property_type"] == "apartment"
mask_ba = df["place_with_parent_names"].str.contains("Distrito Federal")
mask_price = df["price_aprox_usd"] < 100_000
df = df[mask_ba & mask_apt & mask_price]
# Subset data: Remove outliers for "surface_covered_in_m2"
low, high = df["surface_covered_in_m2"].quantile([0.1, 0.9])
mask_area = df["surface_covered_in_m2"].between(low, high)
Simple 0 1 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 025-assignment.ipynb
33°C Smoke
Search
ENG 19:05 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing- | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ... / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	a month ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	2 minutes ago

[13]:

```
def wrangle(filepath):
    # Read CSV file
    df = pd.read_csv(filepath)

    # Subset data: Apartments in <cityName>, less than 100,000
    mask_apt = df["property_type"] == "apartment"
    mask_ba = df["place_with_parent_names"].str.contains("Distrito Federal")

    mask_price = df["price_aprox_usd"] < 100_000
    df = df[mask_ba & mask_apt & mask_price]

    # Subset data: Remove outliers for "surface_covered_in_m2"
    low, high = df["surface_covered_in_m2"].quantile([0.1, 0.9])
    mask_area = df["surface_covered_in_m2"].between(low, high)
    df = df[mask_area]

    # split lat-lon column
    df[["lat", "lon"]] = df["lat-lon"].str.split(",", expand=True).astype(float)
    df.drop(columns="lat-lon", inplace=True)

    # Extract newColumnName
    df['borough'] = df["place_with_parent_names"].str.split("|", expand=True)[1]
    df.drop(columns="place_with_parent_names", inplace=True)

    # Drop feature with high null count
    df.drop(columns=["surface_total_in_m2", "price_usd_per_m2", "floor", "rooms", "expenses"], inplace=True)

    # Drop low- and high- categorical variables
    df.drop(columns=["operation", "property_type", "currency", "properati_url"], inplace=True)

    # Drop leaky columns
    df.drop(columns=["price", "price_aprox_local_currency", "price_per_m2"], inplace=True)

    # Drop columns with multi-colinerLity
    #df.drop(columns=["surface_total_in_m2", "rooms"], inplace=True)

    return df
```

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:05 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing- | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

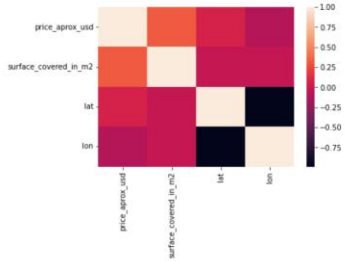
/ ... / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	a month ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	2 minutes ago

[14]:

```
df = wrangle("data/mexico-city-real-estate-1.csv")
corr = df.select_dtypes('number').corr()
sns.heatmap(corr)
```

[14]: <AxesSubplot:>



[15]:

```
wqet_grader.grade(
    "Project 2 Assessment", "Task 2.5.1", wrangle("data/mexico-city-real-estate-1.csv")
)
```

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

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Saving started | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:05 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing-i | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ... / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	a month ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	seconds ago

[15]:
```
wqet_grader.grade(
    "Project 2 Assessment", "Task 2.5.1", wrangle("data/mexico-city-real-estate-1.csv")
)
```

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

Task 2.5.2: Use glob to create the list `files`. It should contain the filenames of all the Mexico City real estate CSVs in the `./data` directory, except for `mexico-city-test-features.csv`.

[18]:
```
files = glob('data/mexico-city-real-estate-*.csv')
```

[19]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.2", files)
```

Yes! Your hard work is paying off.
Score: 1

Task 2.5.3: Combine your `wrangle` function, a list comprehension, and `pd.concat` to create a DataFrame `df`. It should contain all the properties from the five CSVs in `files`.

[20]:
```
df = pd.concat([wrangle(file) for file in files],ignore_index=True)
print(df.info())
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5473 entries, 0 to 5472
Data columns (total 5 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   price_aprox_usd       5473 non-null   float64
 1   surface_covered_in_m2 5473 non-null   float64
 2   lat                   5149 non-null   float64
```

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:08 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing-i | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ... / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	seconds ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	3 minutes ago

Task 2.5.5: Create a scatter plot that shows apartment price (`"price_aprox_usd"`) as a function of apartment size (`"surface_covered_in_m2"`). Be sure to label your x-axis `"Area [sq meters]"` and y-axis `"Price [USD]"`. Your plot should have the title `"Mexico City: Price vs. Area"`. Use Matplotlib (`plt`).

[26]:
```
# Build scatter plot
plt.scatter(x=df["surface_covered_in_m2"],y=df["price_aprox_usd"])


# Label axes
plt.xlabel("Area [sq meters]")
plt.ylabel("Price [USD]")
# Add title
plt.title("Mexico City: Price vs. Area")

# Don't delete the code below 👇
plt.savefig("images/2-5-5.png", dpi=150)
```

Do you see a relationship between price and area in the data? How is this similar to or different from the Buenos Aires dataset?

[27]:
```
with open("images/2-5-5.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.5", file)
```

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:15 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing-i | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ··· / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	in a few seconds
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	2 minutes ago

[24]:
```
# Build histogram
plt.hist(df["price_aprox_usd"])


# Label axes
plt.xlabel("price [$]")
plt.ylabel("count")
# Add title
plt.title("Distribution of Apartment Prices")

# Don't delete the code below 👇
plt.savefig("images/2-5-4.png", dpi=150)
```

[25]:
```
with open("images/2-5-4.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.4", file)
```

Score: 1

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:12 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing-i | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ··· / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	a month ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	seconds ago

Task 2.5.3: Combine your `wrangle` function, a list comprehension, and `pd.concat` to create a DataFrame `df`. It should contain all the properties from the five CSVs in `files`.

[20]:
```
df = pd.concat([wrangle(file) for file in files],ignore_index=True)
print(df.info())
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5473 entries, 0 to 5472
Data columns (total 5 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   price_aprox_usd      5473 non-null   float64
 1   surface_covered_in_m2  5473 non-null   float64
 2   lat                  5149 non-null   float64
 3   lon                  5149 non-null   float64
 4   borough              5473 non-null   object
dtypes: float64(4), object(1)
memory usage: 213.9+ KB
None
```

[20]:

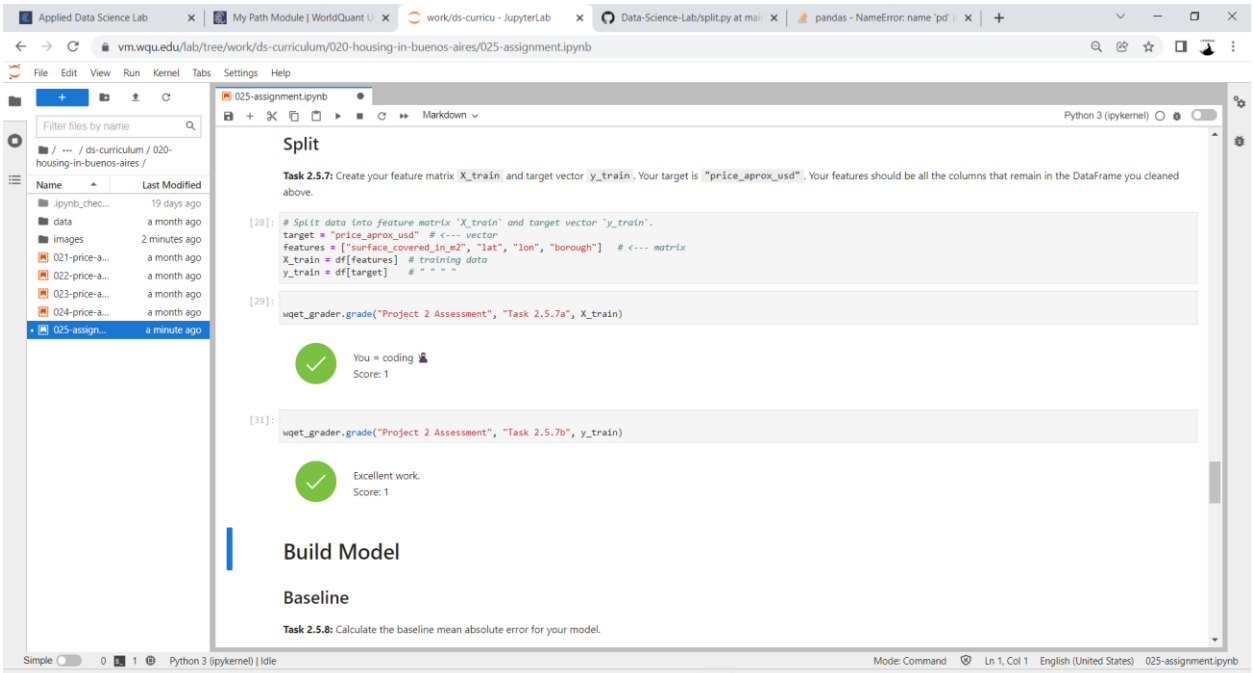
	price_aprox_usd	surface_covered_in_m2	lat	lon	borough
0	29315.91	65.0	19.401350	-99.114726	Iztacalco
1	8693.26	48.0	19.283413	-99.055060	Tláhuac
2	70801.04	80.0	19.484703	-99.211365	Azcapotzalco
3	72250.71	60.0	19.403504	-99.154502	Benito Juárez
4	68492.42	70.0	19.364993	-99.155739	Benito Juárez

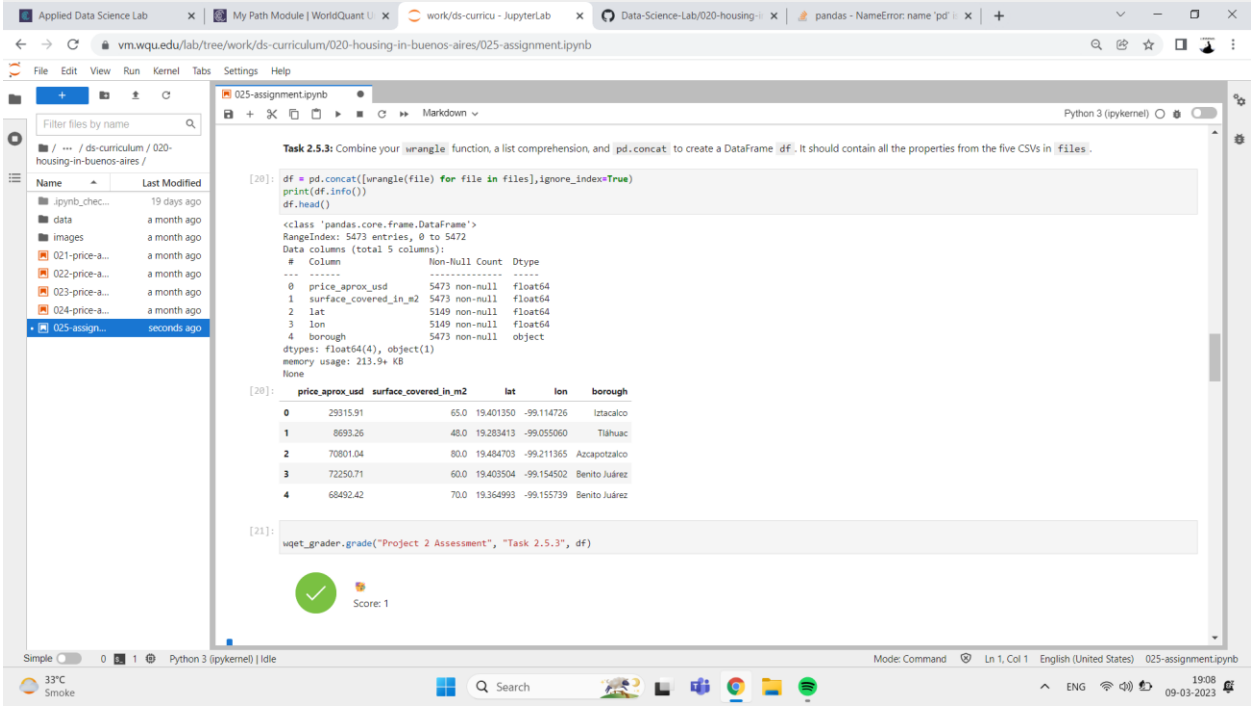
[21]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.3", df)
```

Score: 1

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:08 09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/split.py at mai
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 2 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... a minute ago
Split
Task 2.5.7: Create your feature matrix X_train and target vector y_train. Your target is "price_aprox_usd". Your features should be all the columns that remain in the DataFrame you cleaned above.
[28]:
# Split data into feature matrix `X_train` and target vector `y_train`.
target = "price_aprox_usd" # <--- vector
features = ["surface_covered_in_m2", "lat", "lon", "borough"] # <--- matrix
X_train = df[features] # training data
y_train = df[target] # " " " "
[29]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.7a", X_train)
You = coding
Score: 1
[31]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.7b", y_train)
Excellent work.
Score: 1
Build Model
Baseline
Task 2.5.8: Calculate the baseline mean absolute error for your model.
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
025-assignment.ipynb
33°C
Smoke
Search
ENG
19:17
09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/020-housing-
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images a month ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... seconds ago
Task 2.5.3: Combine your wrangle function, a list comprehension, and pd.concat to create a DataFrame df. It should contain all the properties from the five CSVs in files.
[20]:
df = pd.concat([wrangle(file) for file in files],ignore_index=True)
print(df.info())
df.head()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5473 entries, 0 to 5472
Data columns (total 5 columns):
 # Column Non-Null Count Dtype
--- ------ -------------- -----
 0 price_aprox_usd 5473 non-null float64
 1 surface_covered_in_m2 5473 non-null float64
 2 lat 5149 non-null float64
 3 lon 5149 non-null float64
 4 borough 5473 non-null object
dtypes: float64(4), object(1)
memory usage: 213.9+ KB
None
[20]:
price_aprox_usd surface_covered_in_m2 lat lon borough
0 29315.91 65.0 19.401350 -99.114726 Iztacalco
1 8693.26 48.0 19.283413 -99.055060 Tláhuac
2 70801.04 80.0 19.484703 -99.211365 Azcapotzalco
3 72250.71 60.0 19.403504 -99.154502 Benito Juárez
4 68492.42 70.0 19.364993 -99.155739 Benito Juárez
[21]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.3", df)
Score: 1
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
025-assignment.ipynb
33°C
Smoke
Search
ENG
19:08
09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing-i | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ··· / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	in a few seconds
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	2 minutes ago

[24]:
```
# Build histogram
plt.hist(df["price_aprox_usd"])


# Label axes
plt.xlabel("price [$]")
plt.ylabel("count")
# Add title
plt.title("Distribution of Apartment Prices")

# Don't delete the code below 👇
plt.savefig("images/2-5-4.png", dpi=150)
```

[25]:
```
with open("images/2-5-4.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.4", file)
```

Score: 1

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:12 09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/020-housing-i | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown | Python 3 (ipykernel)

Filter files by name

/ ··· / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	seconds ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	3 minutes ago

Task 2.5.5: Create a scatter plot that shows apartment price (`"price_aprox_usd"`) as a function of apartment size (`"surface_covered_in_m2"`). Be sure to label your x-axis `"Area [sq meters]"` and y-axis `"Price [USD]"`. Your plot should have the title `"Mexico City: Price vs. Area"`. Use Matplotlib (`plt`).

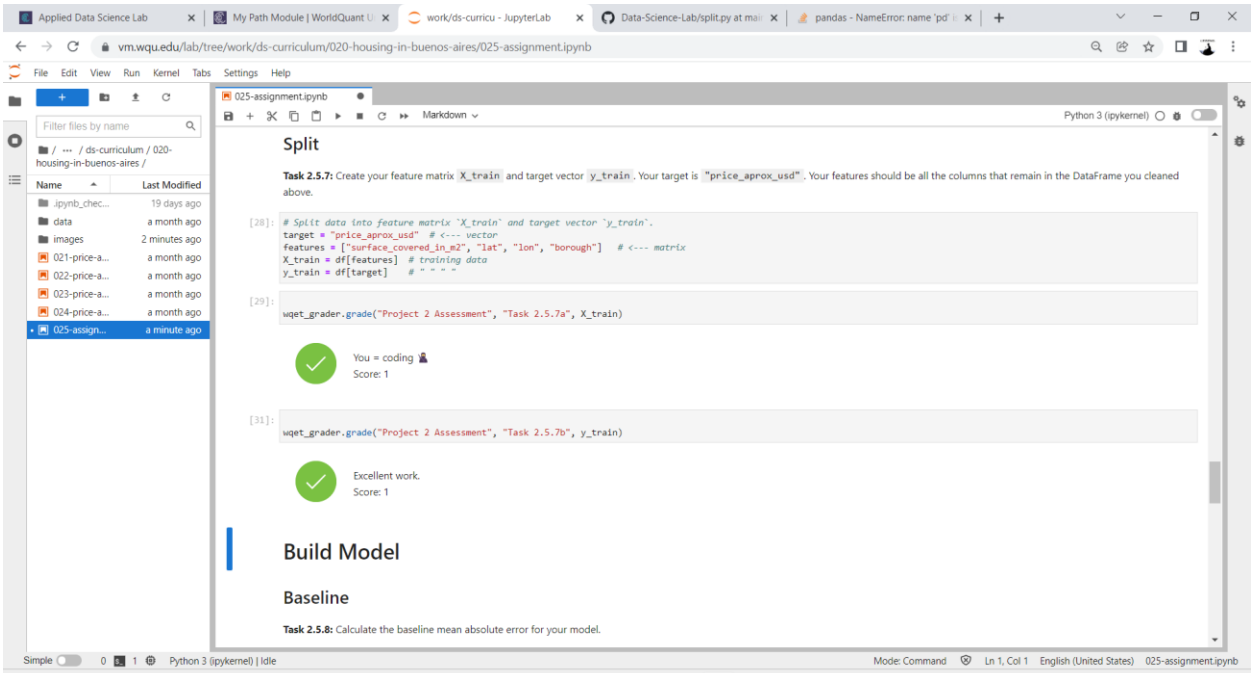
[26]:
```
# Build scatter plot
plt.scatter(x=df["surface_covered_in_m2"],y=df["price_aprox_usd"])


# Label axes
plt.xlabel("Area [sq meters]")
plt.ylabel("Price [USD]")
# Add title
plt.title("Mexico City: Price vs. Area")

# Don't delete the code below 👇
plt.savefig("images/2-5-5.png", dpi=150)
```

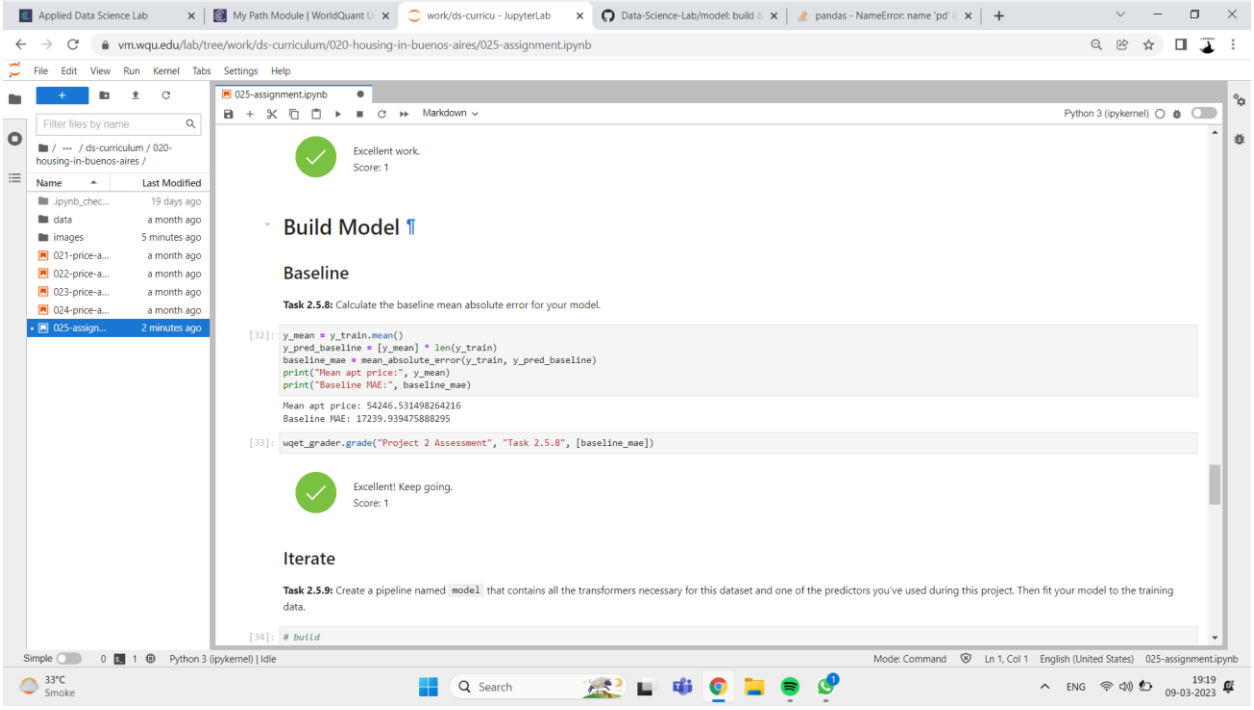
Do you see a relationship between price and area in the data? How is this similar to or different from the Buenos Aires dataset?

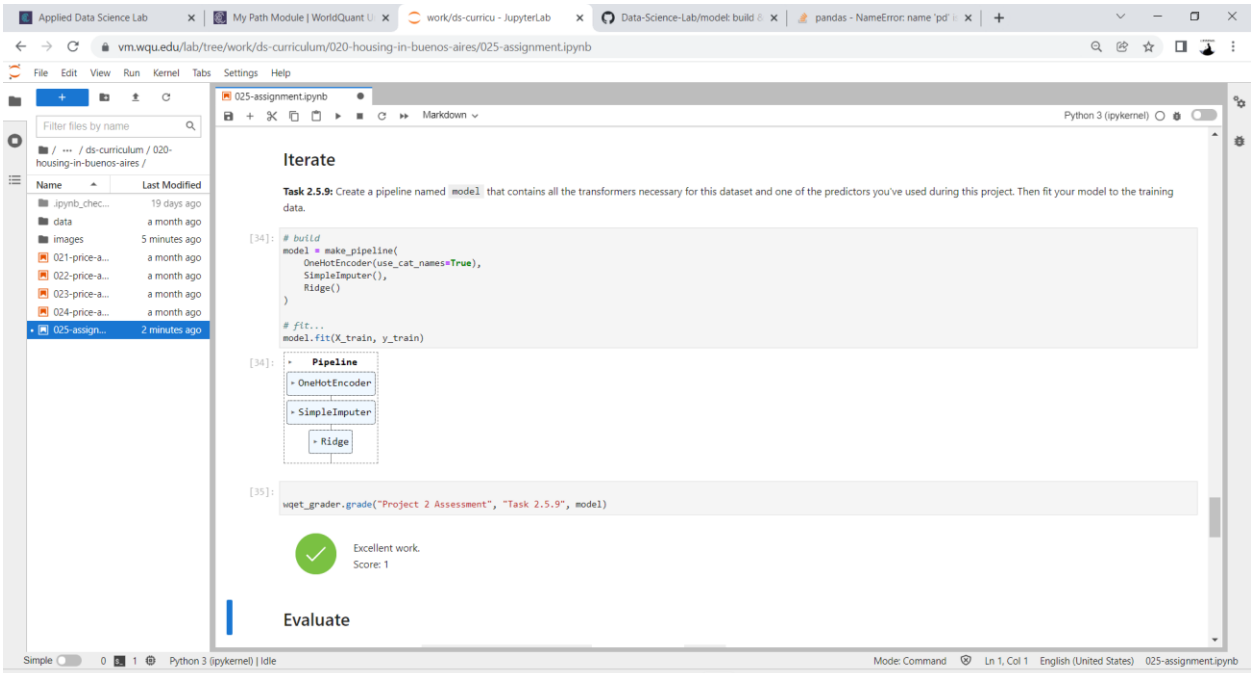
[27]:
```
with open("images/2-5-5.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.5", file)
```

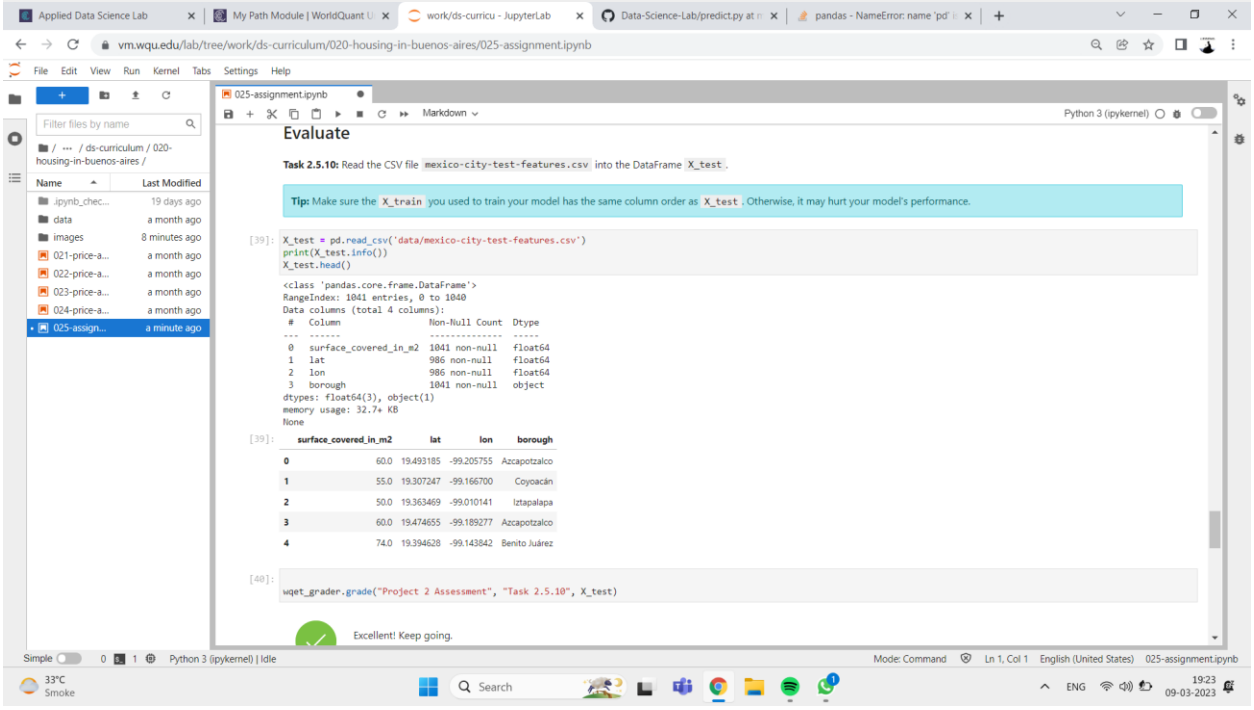
Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

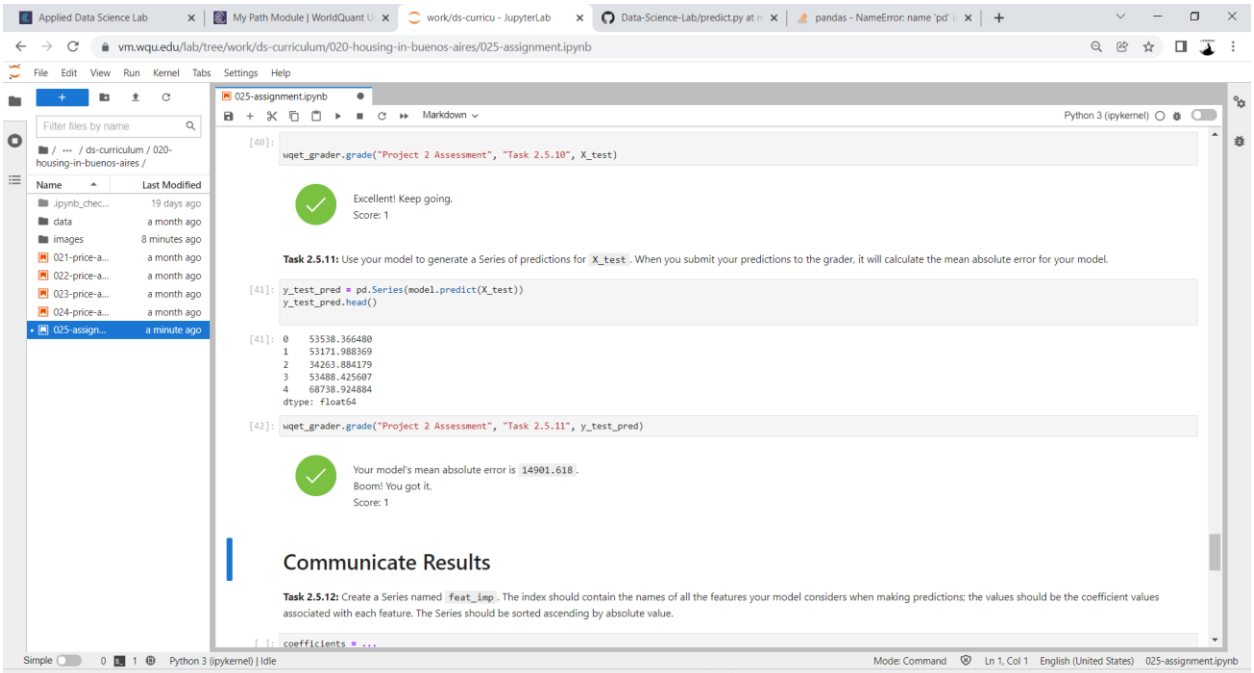
33°C Smoke | Search | ENG | 19:15 09-03-2023

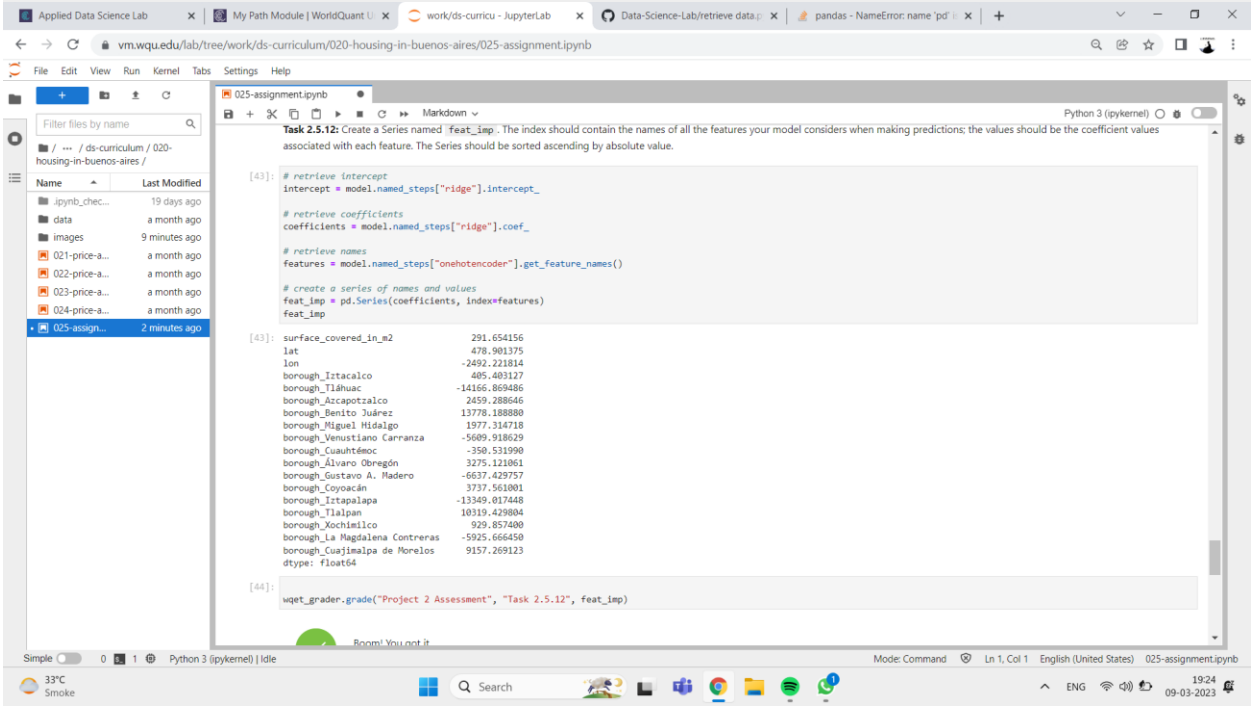

Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/split.py at mai
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 2 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... a minute ago
Split
Task 2.5.7: Create your feature matrix X_train and target vector y_train. Your target is "price_aprox_usd". Your features should be all the columns that remain in the DataFrame you cleaned above.
[28]:
# Split data into feature matrix `X_train` and target vector `y_train`.
target = "price_aprox_usd" # <--- vector
features = ["surface_covered_in_m2", "lat", "lon", "borough"] # <--- matrix
X_train = df[features] # training data
y_train = df[target] # " " " "
[29]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.7a", X_train)
You = coding
Score: 1
[31]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.7b", y_train)
Excellent work.
Score: 1
Build Model
Baseline
Task 2.5.8: Calculate the baseline mean absolute error for your model.
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
025-assignment.ipynb
33°C
Smoke
Search
ENG
19:17
09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/model: build &
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 5 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... 2 minutes ago
Excellent work.
Score: 1
Build Model ¶
Baseline
Task 2.5.8: Calculate the baseline mean absolute error for your model.
[32]:
y_mean = y_train.mean()
y_pred_baseline = [y_mean] * len(y_train)
baseline_mae = mean_absolute_error(y_train, y_pred_baseline)
print("Mean apt price:", y_mean)
print("Baseline MAE:", baseline_mae)
Mean apt price: 54246.531498264216
Baseline MAE: 17239.939475888295
[33]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.8", [baseline_mae])
Excellent! Keep going.
Score: 1
Iterate
Task 2.5.9: Create a pipeline named model that contains all the transformers necessary for this dataset and one of the predictors you've used during this project. Then fit your model to the training data.
[34]:
build
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
025-assignment.ipynb
33°C
Smoke
Search
ENG
19:19
09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/model: build &
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 5 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... 2 minutes ago
Iterate
Task 2.5.9: Create a pipeline named model that contains all the transformers necessary for this dataset and one of the predictors you've used during this project. Then fit your model to the training data.
[34]:
# build
model = make_pipeline(
OneHotEncoder(use_cat_names=True),
SimpleImputer(),
Ridge()
)
# fit...
model.fit(X_train, y_train)
[34]:
Pipeline
OneHotEncoder
SimpleImputer
Ridge
[35]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.9", model)
Excellent work.
Score: 1
Evaluate
Simple 0 1 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 025-assignment.ipynb
33°C
Smoke
Search
ENG
19:20
09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/predict.py at m
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 8 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... a minute ago
Evaluate
Task 2.5.10: Read the CSV file mexico-city-test-features.csv into the DataFrame X_test.
Tip: Make sure the X_train you used to train your model has the same column order as X_test. Otherwise, it may hurt your model's performance.
[39]:
X_test = pd.read_csv('data/mexico-city-test-features.csv')
print(X_test.info())
X_test.head()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1041 entries, 0 to 1040
Data columns (total 4 columns):
Column Non-Null Count Dtype
0 surface_covered_in_m2 1041 non-null float64
1 lat 986 non-null float64
2 lon 986 non-null float64
3 borough 1041 non-null object
dtypes: float64(3), object(1)
memory usage: 32.7+ KB
None
[39]:
surface_covered_in_m2 lat lon borough
0 60.0 19.493185 -99.205755 Azcapotzalco
1 55.0 19.307247 -99.166700 Coyoacán
2 50.0 19.363469 -99.010141 Iztapalapa
3 60.0 19.474655 -99.189277 Azcapotzalco
4 74.0 19.394628 -99.143842 Benito Juárez
[40]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.10", X_test)
Excellent! Keep going.
Simple 0 1 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 English (United States) 025-assignment.ipynb
33°C
Smoke
Search
ENG
19:23
09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/predict.py at m
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 8 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... a minute ago
[40]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.10", X_test)
Excellent! Keep going.
Score: 1
Task 2.5.11: Use your model to generate a Series of predictions for X_test. When you submit your predictions to the grader, it will calculate the mean absolute error for your model.
[41]:
y_test_pred = pd.Series(model.predict(X_test))
y_test_pred.head()
[41]:
0 53538.366480
1 53171.988369
2 34263.884179
3 53488.425607
4 68738.924884
dtype: float64
[42]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.11", y_test_pred)
Your model's mean absolute error is 14901.618.
Boom! You got it.
Score: 1
Communicate Results
Task 2.5.12: Create a Series named feat_imp. The index should contain the names of all the features your model considers when making predictions; the values should be the coefficient values associated with each feature. The Series should be sorted ascending by absolute value.
[ ]:
coefficients = ...
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
025-assignment.ipynb
33°C
Smoke
Search
ENG
19:23
09-03-2023


Applied Data Science Lab
My Path Module | WorldQuant U
work/ds-curricu - JupyterLab
Data-Science-Lab/retrieve data.p
pandas - NameError: name 'pd' i
vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb
File Edit View Run Kernel Tabs Settings Help
025-assignment.ipynb
Markdown
Python 3 (ipykernel)
Filter files by name
/ ... / ds-curriculum / 020-housing-in-buenos-aires /
Name
Last Modified
.ipynb_chec... 19 days ago
data a month ago
images 9 minutes ago
021-price-a... a month ago
022-price-a... a month ago
023-price-a... a month ago
024-price-a... a month ago
025-assign... 2 minutes ago
Task 2.5.12: Create a Series named feat_imp. The index should contain the names of all the features your model considers when making predictions; the values should be the coefficient values associated with each feature. The Series should be sorted ascending by absolute value.
[43]:
# retrieve intercept
intercept = model.named_steps["ridge"].intercept_
# retrieve coefficients
coefficients = model.named_steps["ridge"].coef_
# retrieve names
features = model.named_steps["onehotencoder"].get_feature_names()
# create a series of names and values
feat_imp = pd.Series(coefficients, index=features)
feat_imp
[43]:
surface_covered_in_m2 291.654156
lat 478.901375
lon -2492.221814
borough_Iztacalco 405.403127
borough_Tláhuac -14166.869486
borough_Azcapotzalco 2459.288646
borough_Benito Juárez 13778.188880
borough_Miguel Hidalgo 1977.314718
borough_Venustiano Carranza -5609.918629
borough_Cuauhtémoc -350.531990
borough_Álvaro Obregón 3275.121061
borough_Gustavo A. Madero -6637.429757
borough_Coyoacán 3737.561001
borough_Iztapalapa -13349.017448
borough_Tlalpan 10319.429804
borough_Xochimilco 929.857400
borough_La Magdalena Contreras -5925.666450
borough_Cuajimalpa de Morelos 9157.269123
dtype: float64
[44]:
wqet_grader.grade("Project 2 Assessment", "Task 2.5.12", feat_imp)
Boom! You got it.
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
English (United States)
025-assignment.ipynb
33°C
Smoke
Search
ENG
19:24
09-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/retrieve data.p | pandas - NameError: name 'pd' i

vm.wqu.edu/lab/tree/work/ds-curriculum/020-housing-in-buenos-aires/025-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

025-assignment.ipynb

Markdown

Python 3 (ipykernel)

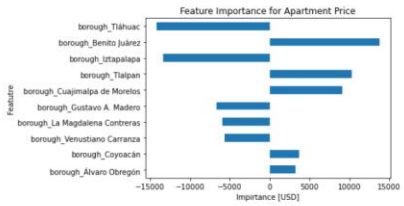
Filter files by name

/ ... / ds-curriculum / 020-housing-in-buenos-aires /

Name	Last Modified
.ipynb_chec...	19 days ago
data	a month ago
images	seconds ago
021-price-a...	a month ago
022-price-a...	a month ago
023-price-a...	a month ago
024-price-a...	a month ago
025-assign...	2 minutes ago

and give your chart the title `"Feature Importances for Apartment Price"`. Use pandas.

[46]:
```
# Build bar chart
feat_imp.sort_values (key=abs). tail(10).plot(kind='barh')

# Label axes
plt.xlabel('Impirtance [USD]')

plt.ylabel('Featutre')


# Add title
plt.title('Feature Importance for Apartment Price')

# Don't delete the code below 👇
plt.savefig("images/2-5-13.png", dpi=150)
```



[47]:
```
with open("images/2-5-13.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.13", file)
```

You = coding

Simple | 0 | 1 | Python 3 (ipykernel) | Idle | Mode: Command | Ln 1, Col 1 | English (United States) | 025-assignment.ipynb

33°C Smoke | Search | ENG | 19:26 09-03-2023