025-assignment.2022-04-18T20-25-03-795Z

April 20, 2022

Assignment: Predicting Apartment Prices in Mexico City

[1]:
```
import warnings

import wqet_grader

warnings.simplefilter(action="ignore", category=FutureWarning)
wqet_grader.init("Project 2 Assessment")
```

```
<IPython.core.display.HTML object>
```

```
<b>Note:</b> In this project there are graded tasks in both the lesson notebooks and in this as
```

In this assignment, you'll decide which libraries you need to complete the tasks. You can import them in the cell below.

[2]:
```
# Import libraries here
from glob import glob

import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import seaborn as sns
from category_encoders import OneHotEncoder
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
```

1 Prepare Data

1.1 Import

Task 2.5.1: (8 points) 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 16 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.

[3]:
```
# Build your `wrangle` function

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

    # Subset data: Apartments in "Capital Federal", less than 400,000
    mask_ba = df["place_with_parent_names"].str.contains("Distrito Federal")
    mask_apt = df["property_type"] == "apartment"
    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)

    # Get place name
    df["borough"] = df["place_with_parent_names"].str.split("|", expand=True)[1]
    df.drop(columns="place_with_parent_names", inplace=True)

    #remove the columns which are maximum null counts
    df.
↪drop(columns=["surface_total_in_m2","price_usd_per_m2","floor","rooms","expenses"],␣
↪inplace=True)

    #remove low and high cardinality categorical variable
    df.drop(columns=["operation","property_type","currency","properati_url"],␣
↪inplace=True)
```

```
    #remove licky columns
    df.drop(columns=[
         'price',
         'price_aprox_local_currency',
         'price_per_m2',

    ],
            inplace=True)

#     #remove columns with multicollinearity
#     df.drop(columns=["rooms", "surface_total_in_m2"], inplace=True)


    return df
```

[4]:
```
# Use this cell to test your wrangle function and explore the data
df = pd.read_csv("data/mexico-city-real-estate-1.csv")
df.info()
#df.head()
#df.isnull().sum() / len(df)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4628 entries, 0 to 4627
Data columns (total 16 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   operation                   4628 non-null   object
 1   property_type               4628 non-null   object
 2   place_with_parent_names     4628 non-null   object
 3   lat-lon                     4144 non-null   object
 4   price                       4538 non-null   float64
 5   currency                    4538 non-null   object
 6   price_aprox_local_currency  4538 non-null   float64
 7   price_aprox_usd             4538 non-null   float64
 8   surface_total_in_m2         1668 non-null   float64
 9   surface_covered_in_m2       4436 non-null   float64
 10  price_usd_per_m2            1150 non-null   float64
 11  price_per_m2                4249 non-null   float64
 12  floor                       291 non-null    float64
 13  rooms                       136 non-null    float64
 14  expenses                    5 non-null      float64
 15  properati_url               4628 non-null   object
dtypes: float64(10), object(6)
memory usage: 578.6+ KB
```

[5]:
```
wqet_grader.grade(
```

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

```
<IPython.core.display.HTML object>
```

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`.

[6]:
```
files = glob("data/mexico-city-real-estate-*.csv")
files
```

[6]:
```
['data/mexico-city-real-estate-4.csv',
 'data/mexico-city-real-estate-3.csv',
 'data/mexico-city-real-estate-1.csv',
 'data/mexico-city-real-estate-2.csv',
 'data/mexico-city-real-estate-5.csv']
```

[9]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.2", files)
```

```
<IPython.core.display.HTML object>
```

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`.

[10]:
```
frames = [wrangle(file) for file in files]
df = pd.concat(frames, ignore_index=True)
print(df.info())
df.head()
#df.isnull().sum() / len(df)
#df.select_dtypes("object").nunique()
#sorted(df.columns)
```

```
<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
```

[10]:

	price_aprox_usd	surface_covered_in_m2	lat	lon	\
0	52686.48	65.0	19.389011	-99.180415	

```
1         48581.99                  66.0  23.634501 -102.552788
2         60589.45                  52.0  19.469681  -99.086136
3         33530.62                  69.0  19.429437  -99.143460
4         23355.39                  54.0  19.408007  -99.069186

              borough
0      Benito Juárez
1          Iztacalco
2  Gustavo A. Madero
3         Cuauhtémoc
4          Iztacalco
```

[11]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.3", df)
```

```
<IPython.core.display.HTML object>
```

1.2 Explore

Task 2.5.4: Create a histogram showing the distribution of apartment prices (`"price_aprox_usd"`) in `df`. Be sure to label the x-axis `"Area [sq meters]"`, the y-axis `"Count"`, and give it the title `"Distribution of Apartment Prices"`.

What does the distribution of price look like? Is the data normal, a little skewed, or very skewed?

[12]:
```
# Plot distribution of price

plt.hist(df["price_aprox_usd"])
plt.xlabel("Area [sq meters]")
plt.ylabel("Count")
plt.title("Distribution of Apartment Prices")

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

[14]:
```
with open("images/2-5-4.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.4", file)
```

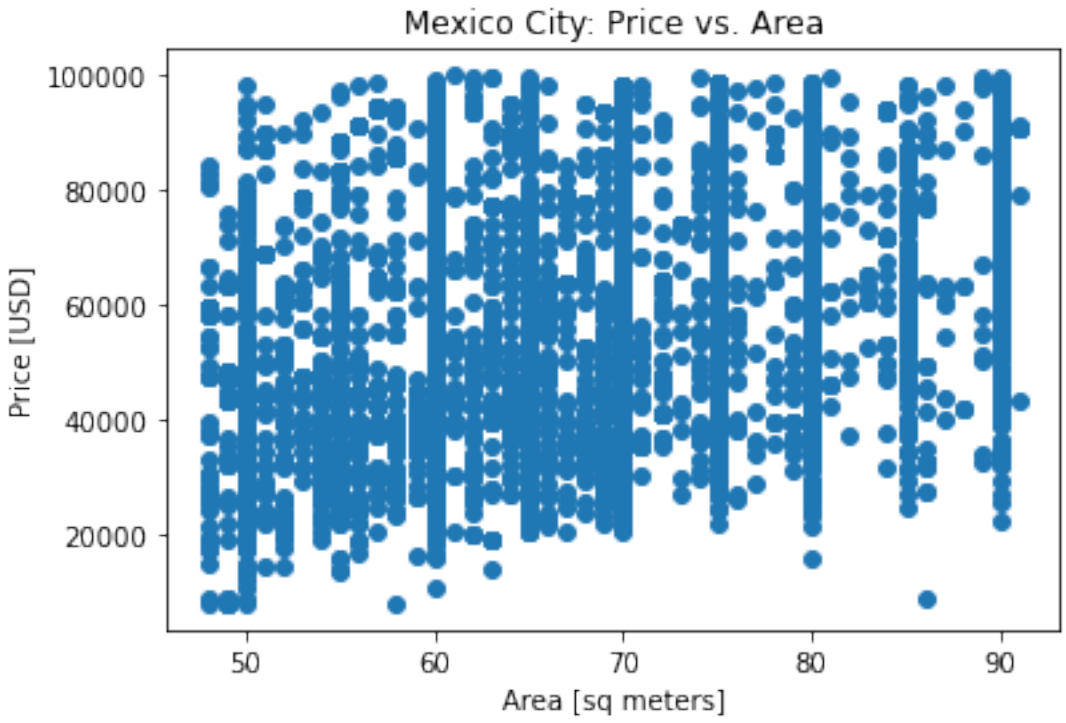
```
<IPython.core.display.HTML object>
```

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 axes `"Price [USD]"` and `"Area [sq meters]"`, respectively. Your plot should have the title `"Mexico City: Price vs. Area"`.

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

[15]:
```
# Plot price vs area
plt.scatter(x=df["surface_covered_in_m2"],y=df["price_aprox_usd"])
plt.xlabel("Area [sq meters]")
plt.ylabel("Price [USD]")
plt.title("Mexico City: Price vs. Area")

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



[16]:
```
with open("images/2-5-5.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.5", file)
```

```
<IPython.core.display.HTML object>
```

Task 2.5.6: (UNGRADED) Create a Mapbox scatter plot that shows the location of the apartments in your dataset and represent their price using color.

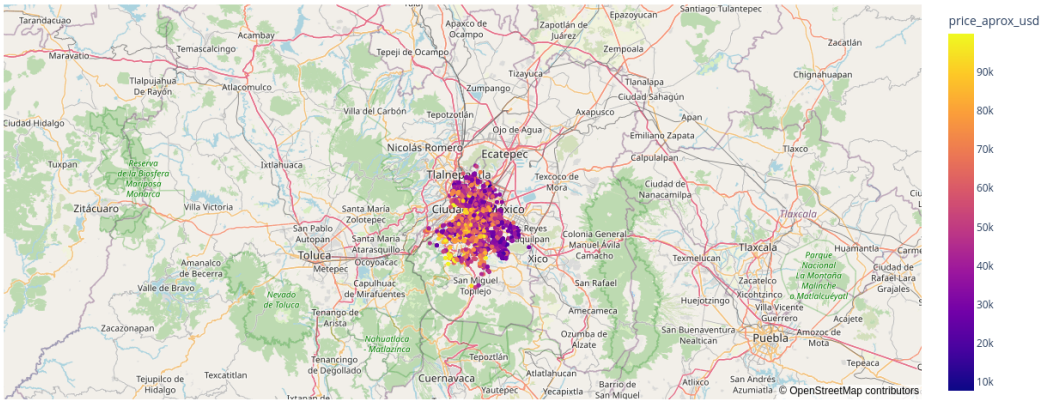
What areas of the city seem to have higher real estate prices?

[17]:
```
# Plot Mapbox location and price

fig = px.scatter_mapbox(
    df,  # Our DataFrame
    lat="lat",
    lon="lon",
    width=600,  # Width of map
    height=600,  # Height of map
    color="price_aprox_usd",
    hover_data=["price_aprox_usd"],  # Display price when hovering mouse over␣
 ↪house
)

fig.update_layout(mapbox_style="open-street-map")
```

```
fig.show()
```



1.3 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.

[18]:
```
# Split data into feature matrix `X_train` and target vector `y_train`.
features = ["surface_covered_in_m2","lat","lon","borough"]
target = "price_aprox_usd"
X_train = df[features]
y_train = df[target]
```

[20]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.7a", X_train)
```

```
<IPython.core.display.HTML object>
```

[ ]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.7b", y_train)
```

2 Build Model

2.1 Baseline

Task 2.5.8: Calculate the baseline mean absolute error for your model.

[21]:
```
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: 54246.5314982643
Baseline MAE: 17239.939475888303
```

[22]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.8", [baseline_mae])
```

```
<IPython.core.display.HTML object>
```

2.2 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.

[23]:
```
# Build Model
model = make_pipeline(
    OneHotEncoder(use_cat_names=True),
    SimpleImputer(),
    Ridge()
)
# Fit model
model.fit(X_train, y_train)
```

[23]:
```
Pipeline(steps=[('onehotencoder',
                 OneHotEncoder(cols=['borough'], use_cat_names=True)),
                ('simpleimputer', SimpleImputer()), ('ridge', Ridge())])
```

[24]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.9", model)
```

```
<IPython.core.display.HTML object>
```

2.3 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.

[25]:
```
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
```

[25]:

	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

[26]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.10", X_test)
```

```
<IPython.core.display.HTML object>
```

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.

[33]:
```
y_test_pred = pd.Series(model.predict(X_test))
y_test_pred.head()
```

[33]:
```
0    53538.366480
1    53171.988369
2    34263.884179
3    53488.425607
4    68738.924884
dtype: float64
```

[34]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.11", y_test_pred)
```

```
<IPython.core.display.HTML object>
```

3 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.

[35]:
```
coefficients = model.named_steps["ridge"].coef_
features = model.named_steps["onehotencoder"].get_feature_names()
feat_imp = pd.Series(coefficients, index=features)
feat_imp
```

[35]:
```
surface_covered_in_m2              291.654156
lat                                478.901375
```

```
lon                                -2492.221814
borough_Benito Juárez              13778.188880
borough_Iztacalco                    405.403127
borough_Gustavo A. Madero          -6637.429757
borough_Cuauhtémoc                  -350.531990
borough_Azcapotzalco                2459.288646
borough_Venustiano Carranza        -5609.918629
borough_Tláhuac                   -14166.869486
borough_Álvaro Obregón              3275.121061
borough_Coyoacán                    3737.561001
borough_Tlalpan                    10319.429804
borough_Miguel Hidalgo              1977.314718
borough_Iztapalapa                -13349.017448
borough_Cuajimalpa de Morelos       9157.269123
borough_Xochimilco                   929.857400
borough_La Magdalena Contreras     -5925.666450
dtype: float64
```

[36]:
```
wqet_grader.grade("Project 2 Assessment", "Task 2.5.13", feat_imp)
```

```
---------------------------------------------------------------------------
Exception                                 Traceback (most recent call last)
Input In [36], in <cell line: 1>()
----> 1 wqet_grader.grade("Project 2 Assessment", "Task 2.5.13", feat_imp)

File /opt/conda/lib/python3.9/site-packages/wqet_grader/__init__.py:180, in␣
 ↪grade(assessment_id, question_id, submission)
    175 def grade(assessment_id, question_id, submission):
    176   submission_object = {
    177     'type': 'simple',
    178     'argument': [submission]
    179   }
--> 180   return␣
 ↪show_score(grade_submission(assessment_id, question_id, submission_object))

File /opt/conda/lib/python3.9/site-packages/wqet_grader/transport.py:143, in␣
 ↪grade_submission(assessment_id, question_id, submission_object)
    141     raise Exception('Grader raised error: {}'.format(error['message']))
    142   else:
--> 143     raise Exception('Could not grade submission: {}'.
 ↪format(error['message']))
    144 result = envelope['data']['result']
    146 # Used only in testing
```

```
Exception: Could not grade submission: Could not verify access to this␣
↪assessment: Received error from WQET submission API: You have already passed␣
↪this course!
```

Task 2.5.13: Create a horizontal bar chart that shows the **10 most influential** coefficients for your model. Be sure to label your x- and y-axis `"Importance [USD]"` and `"Feature"`, respectively, and give your chart the title `"Feature Importances for Apartment Price"`.

[37]:
```
# Create horizontal bar chart

# Don't delete the code below
plt.savefig("images/2-5-14.png", dpi=150)
```

```
<Figure size 432x288 with 0 Axes>
```

[38]:
```
with open("images/2-5-14.png", "rb") as file:
    wqet_grader.grade("Project 2 Assessment", "Task 2.5.14", file)
```

```
---------------------------------------------------------------------------
Exception                                 Traceback (most recent call last)
Input In [38], in <cell line: 1>()
      1 with open("images/2-5-14.png", "rb") as file:
----> 2     wqet_grader.grade("Project 2 Assessment", "Task 2.5.14", file)

File /opt/conda/lib/python3.9/site-packages/wqet_grader/__init__.py:180, in␣
↪grade(assessment_id, question_id, submission)
    175 def grade(assessment_id, question_id, submission):
    176   submission_object = {
    177     'type': 'simple',
    178     'argument': [submission]
    179   }
--> 180   return␣
↪show_score(grade_submission(assessment_id, question_id, submission_object))

File /opt/conda/lib/python3.9/site-packages/wqet_grader/transport.py:143, in␣
↪grade_submission(assessment_id, question_id, submission_object)
    141     raise Exception('Grader raised error: {}'.format(error['message']))
    142   else:
--> 143     raise Exception('Could not grade submission: {}'.
↪format(error['message']))
    144 result = envelope['data']['result']
    146 # Used only in testing

Exception: Could not grade submission: Could not verify access to this␣
↪assessment: Received error from WQET submission API: You have already passed␣
↪this course!
```