

# ICP5

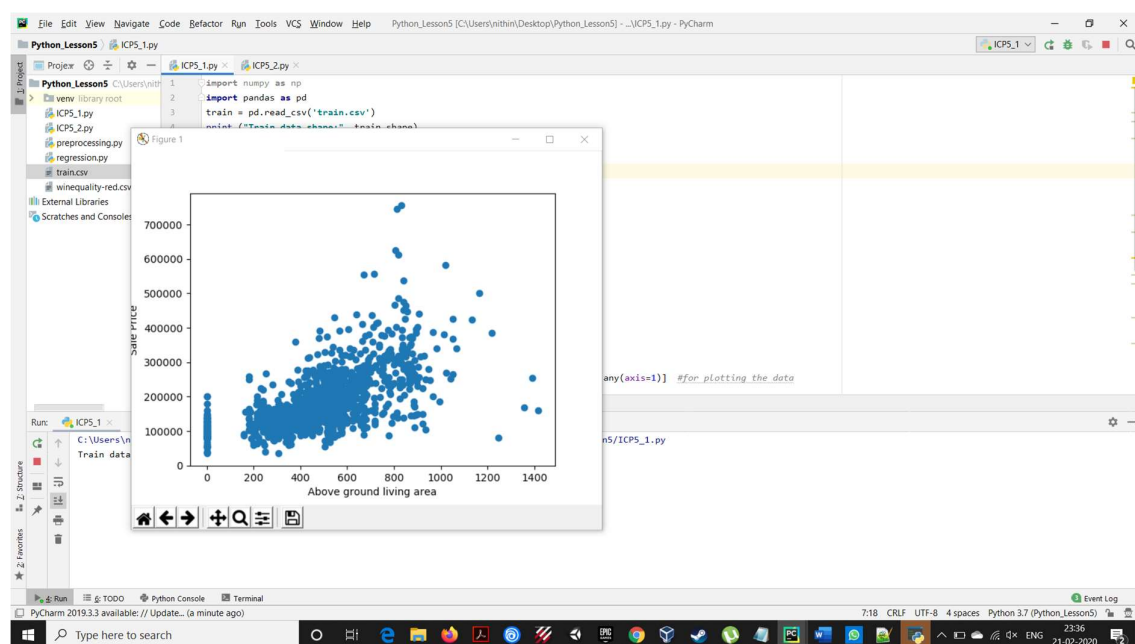
**1. Delete all the outlier data for the GarageArea field (for the same data set in the use case: House Prices).\*** for this task you need to plot GaurageArea field and SalePrice in scatter plot, then check which numbers are anomalies.

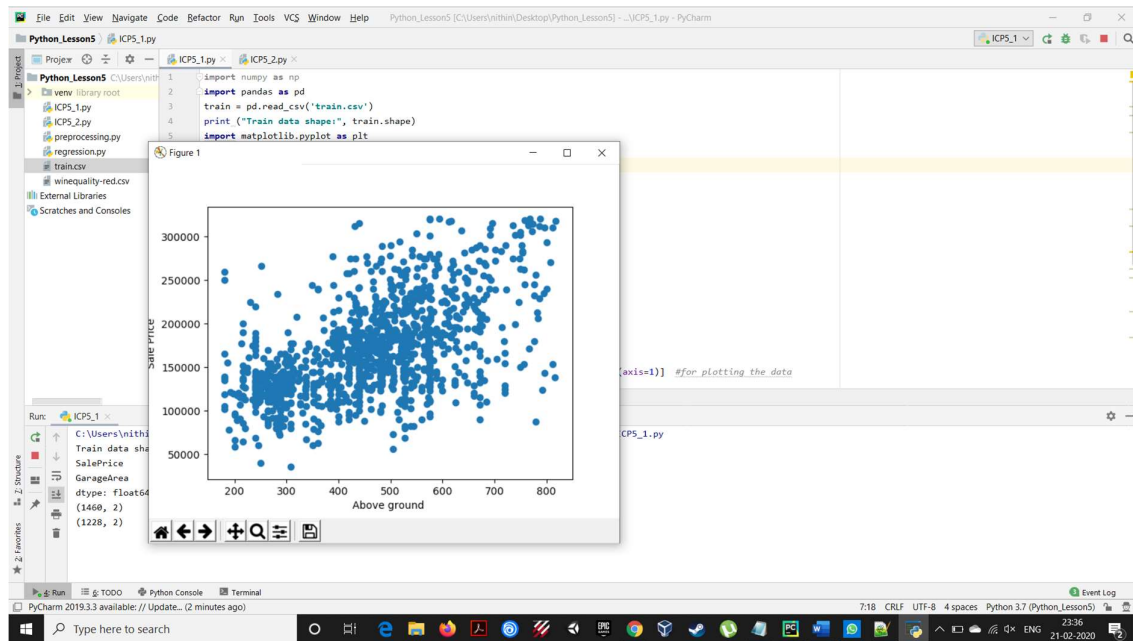
```
Python_Lesson5 [C:\Users\nithin\Desktop\Python_Lesson5] - ICP5_1.py - PyCharm

Python_Lesson5 C:\Users\nithin\Desktop\Python_Lesson5
venv library root
ICP5_1.py
ICP5_2.py
preprocessing.py
regression.py
train.csv
winequality-red.csv
External Libraries
Scratches and Consoles

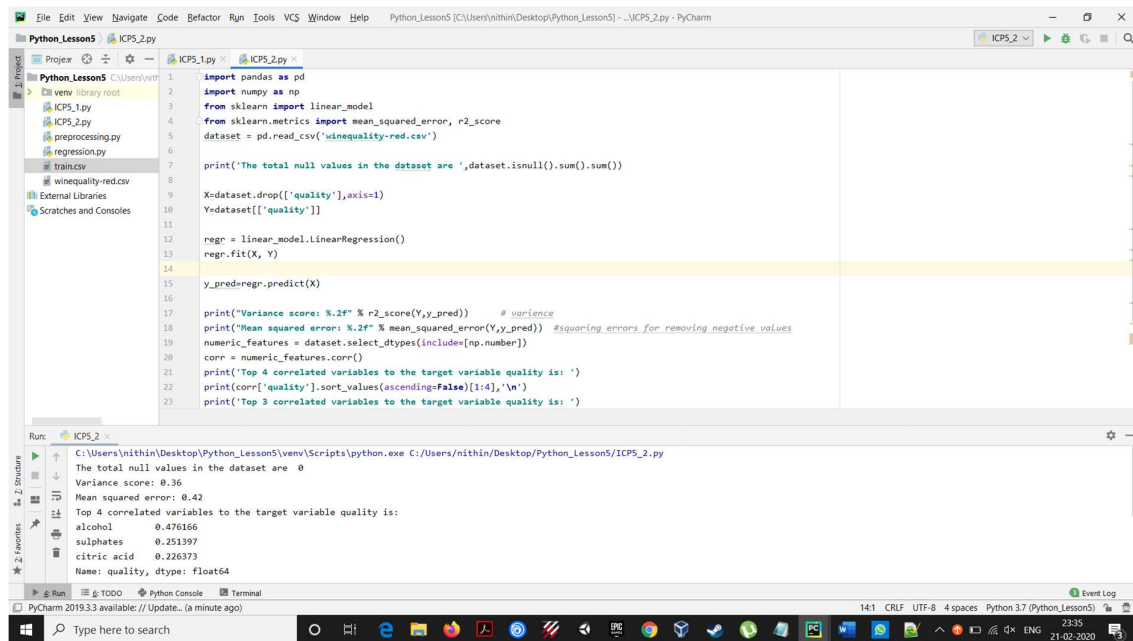
ICP5_1.py 1: import numpy as np
2: import pandas as pd
3: train = pd.read_csv('train.csv')
4: print("Train data shape:", train.shape)
5: import matplotlib.pyplot as plt
6:
7: var = 'GarageArea'
8: data = pd.concat([train['SalePrice'], train[var]], axis=1)
9: plt.scatter(x=train['GarageArea'], y=train['SalePrice'])
10: plt.ylabel('Sale Price')
11: plt.xlabel('Above ground living area')
12: plt.show()
13:
14:
15:
16: Q1 = data.quantile(0.35) # removing the outliers
17: Q3 = data.quantile(0.75)
18: IQR = Q3 - Q1 #training the data
19: print(IQR)
20:
21: data_df_out = data[~((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR))).any(axis=1)] #for plotting the data
22:
23: print(data.shape) #before removing outliers
24: print(data_df_out.shape) #after removing outliers
25:
26:
27:
28: plt.scatter(x=data_df_out['GarageArea'], y=data_df_out['SalePrice'])
29: plt.ylabel('Sale Price')
30: plt.xlabel('Above ground ')
31: plt.show()

Run ICP5_2
C:\Users\nithin\Desktop\Python_Lesson5\venv\Scripts\python.exe C:\Users\nithin\Desktop\Python_Lesson5\ICP5_2.py
PyCharm 2019.3.3 available // Update... (a minute ago)
1:112 CRLF UTF-8 4 spaces Python 3.7 (Python_Lesson5)
2336 21-02-2020
```





**2. Create Multiple Regression for the “wine quality” dataset. In this data set “quality” is the target label. Evaluate the model using RMSE and R2 score.**



By

**DUKKIPATI SRI SAI NITHIN CHOWDARY**

**Class id: 4**

**Lab id: 6**