CS5590/490 - Python and Deep Learning Programming
In Class Programming Report - 5
Class ID 24 - Anurag Thantharate
Date Submitted: 02/23/2020

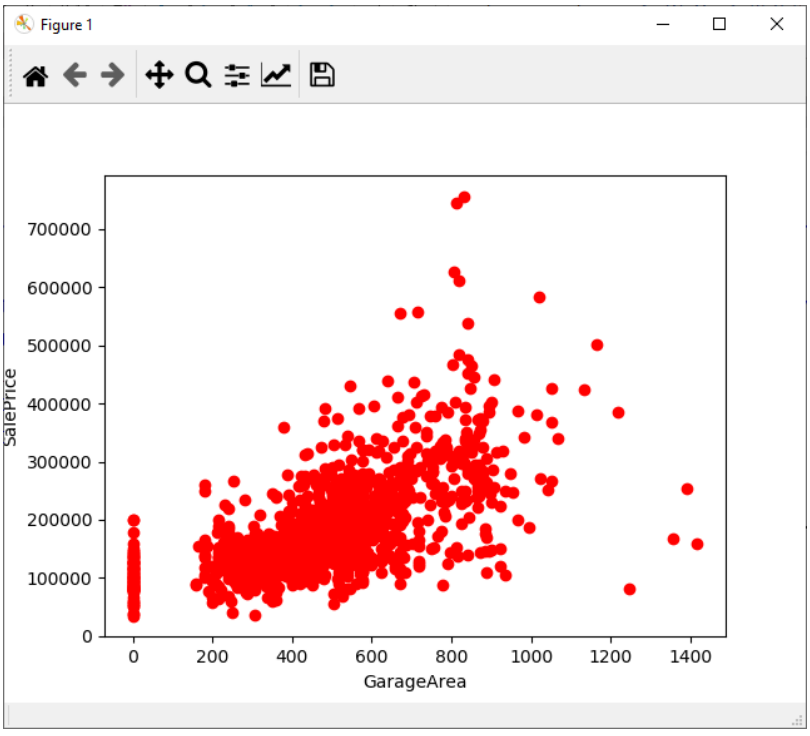
Video Link: https://www.loom.com/share/ddac364d37a6404d9e5affd78d719a77

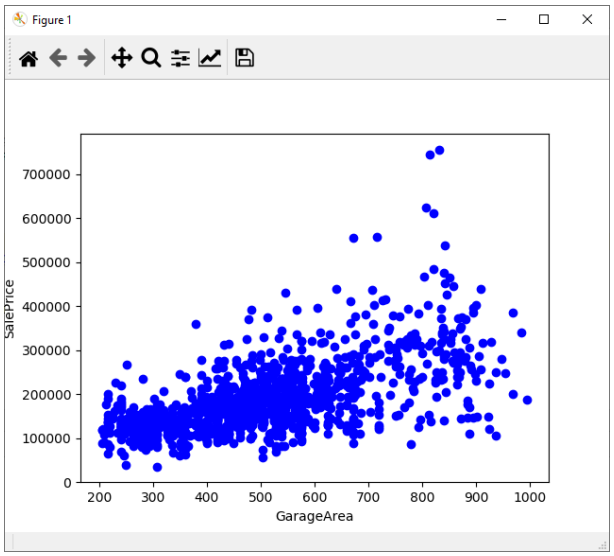
---

1. Delete all the outlier data for the Garage Area field (for the same data set in the use case: House Prices).* for this task you need to plot Gaurage Area field and Sale Price in scatter plot, then check which numbers are anomalies.

```
import pandas as pd
import matplotlib.pyplot as plt

#Read data from dataset
train = pd.read_csv('./train.csv')

#Display the scatter plot of GarageArea and SalePrice
plt.scatter(train.GarageArea, train.SalePrice, color='red')
plt.xlabel('GarageArea')
plt.ylabel('SalePrice')
plt.show()

#Delete the outlier value of GarageArea
outlier_drop = train[(train.GarageArea <1000) & (train.GarageArea >200)]

##Display the scatter plot of GarageArea and SalePrice after filtering
plt.scatter(outlier_drop.GarageArea, outlier_drop.SalePrice, color='blue')
plt.xlabel('GarageArea')
plt.ylabel('SalePrice')
plt.show()

print(train.SalePrice.describe())
```

```
Run:  guagearea_1
C:\ProgramData\Anaconda3\python.exe "C:/Users/1d63
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64

Process finished with exit code 0
```





---

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.
   https://umkc.box.com/s/4hb8p4de88gg1osc10jdex2qscm2l4af
   **You need to delete the null values in the data set
   **You need to find the top 3 most correlated features to the target label(quality)

multipleregression_2.py

```
from pathlib import Path
import pandas as pd
import numpy as np

train = pd.read_csv(Path('./winequality-red.csv'))

#Working with Numeric Features and top features
n_features = train.select_dtypes(include=[np.number])
corr = n_features.corr()
print(corr['quality'].sort_values(ascending=False)[:3], '\n')

#Null values
nulls = pd.DataFrame(train.isnull().sum().sort_values(ascending=False))
nulls.columns = ['Null Count']
nulls.index.name = 'Feature'
print(nulls)

#Handling missing value
data = train.select_dtypes(include=[np.number]).interpolate().dropna()
print(sum(data.isnull().sum() != 0))

#Build a linear model
y = np.log(train.quality)
X = data.drop(['quality'], axis=1)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, test_size=.33)

from sklearn import linear_model
lr = linear_model.LinearRegression()
model = lr.fit(X_train, y_train)
```

Run: multipleregression_2

```
C:\ProgramData\Anaconda3\python.exe "C:/U
quality      1.000000
alcohol      0.476166
sulphates    0.251397
Name: quality, dtype: float64

                      Null Count
Feature
quality                        0
alcohol                        0
sulphates                      0
pH                             0
density                        0
total sulfur dioxide           0
free sulfur dioxide            0
chlorides                      0
residual sugar                 0
citric acid                    0
volatile acidity               0
fixed acidity                  0
0
R^2 is:  0.3433962675485104
RMSE is:  0.013811674843193882

Process finished with exit code 0
```
