Unit 3: Correlation and Regression

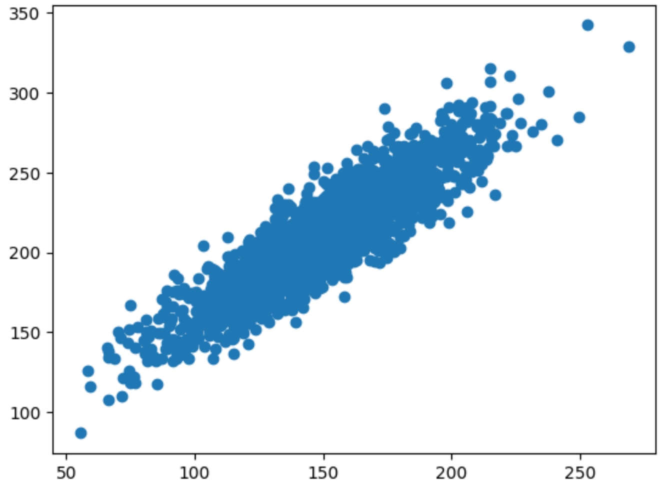
Unit 3 Artefact

Correlation and Regression

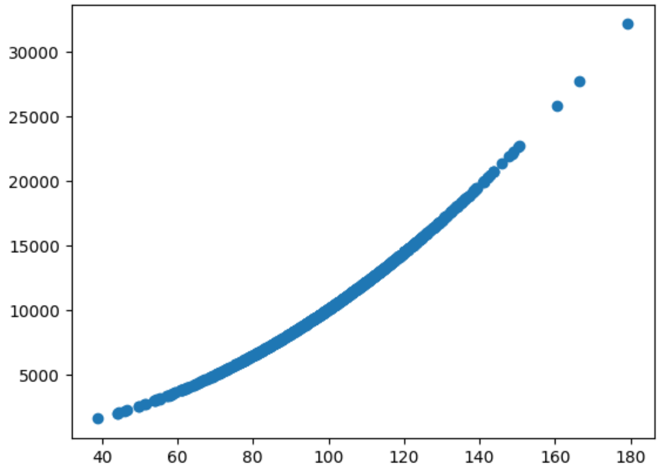
I ran these four programmes via Google Colab and changed certain variables to observe the effect of this change in data points on correlation and regression analyses. I have documented some of the key experimental steps under each programme heading in this document and have included screenshots out output for efficiency of sharing in one place on my e-portfolio.

covariance_pearson_correlation.ipynb

I modified the input data as below:
```
# prepare data
data1 = 30 * randn(2000) + 150
data2 = data1 + (15 * randn(2000) + 60)
```

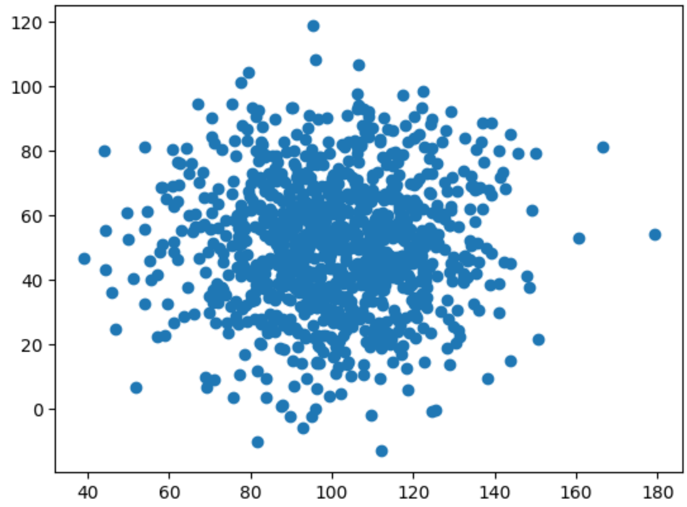


I then introduced a non-linear relationship:
```
# prepare data
data1 = 20 * randn(1000) + 100
data2 = data1 ** 2 + (10 * randn(1000) + 50)  # Non-linear relationship
```



And finally, I looked at the effect of independent data:

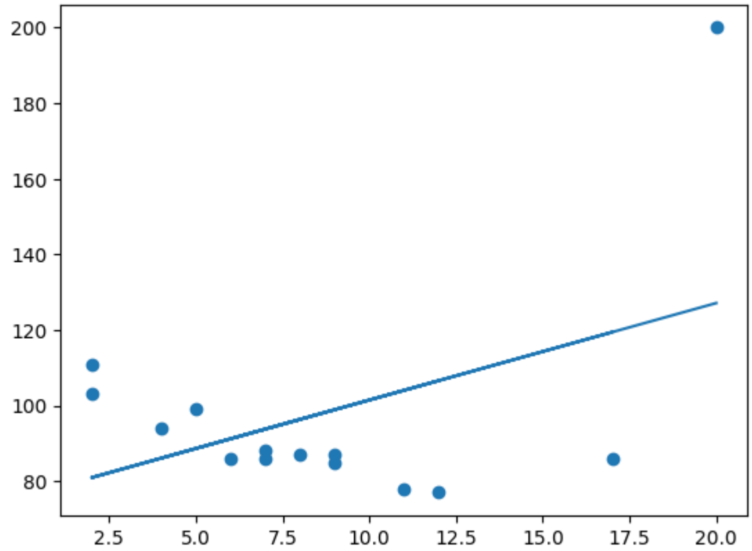
```
# prepare data
data1 = 20 * randn(1000) + 100
data2 = 20 * randn(1000) + 50  # Independent data
```



linear_regression.ipynb

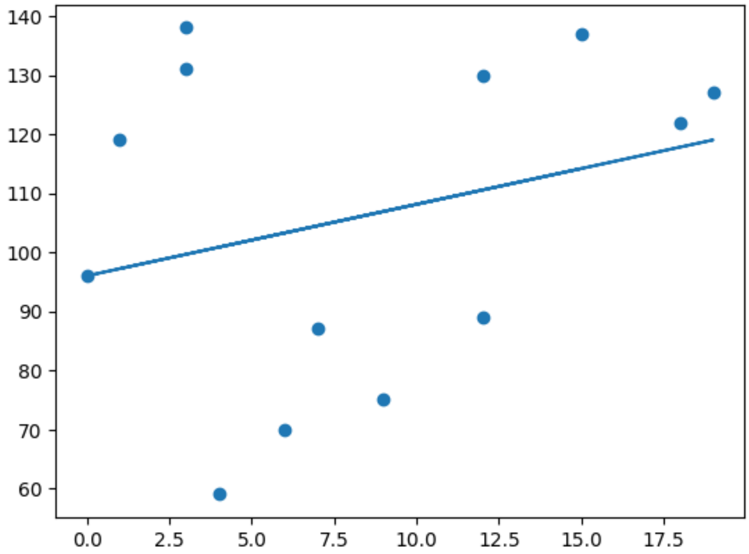
First, I introduced **outliers**:

#Create the arrays that represent the values of the x and y axis
x = [5,7,8,7,2,17,2,9,4,11,12,9,6,**20**]
y = [99,86,87,88,111,86,103,87,94,78,77,85,86,**200**]



Then I used completely random data:

```
# Create the arrays that represent the values of the x and y axis with random data
np.random.seed(0)
x = np.random.randint(0, 20, size=13)
y = np.random.randint(50, 150, size=13)
```



multiple_linear_regression.ipynb

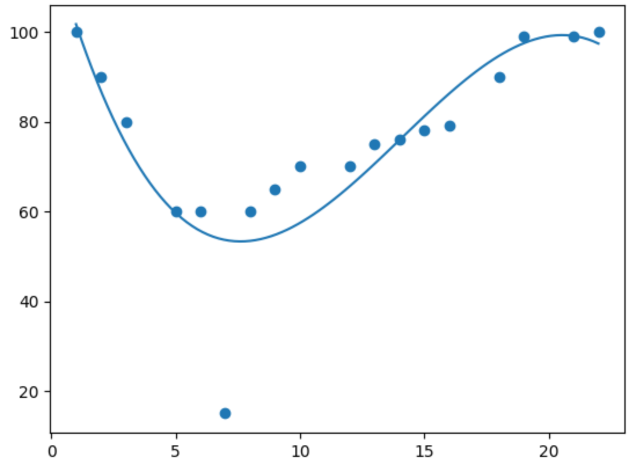
I used the calafornia housing dataset and updated the code to include pandas.DataFrame for prediction in the linear regression model:

```
import pandas as pd
from sklearn import linear_model

# Load the dataset
df = pd.read_csv("/content/sample_data/california_housing_test.csv")

# Prepare the feature matrix X and the target vector y
X = df[['housing_median_age', 'total_rooms']]
y = df['median_income']

# Create and train the linear regression model
regr = linear_model.LinearRegression()
regr.fit(X, y)

# Predict the median income for a given housing_median_age and total_rooms
# Here we use sample values within a reasonable range for these features
housing_median_age = 30
total_rooms = 3000
prediction_data = pd.DataFrame({'housing_median_age': [housing_median_age],
'total_rooms': [total_rooms]})
```

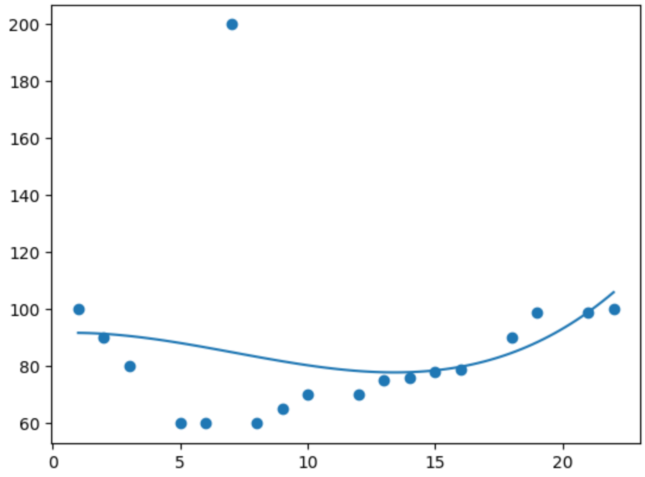
```
predicted_income = regr.predict(prediction_data)

print('Predicted median income:', predicted_income[0])

Predicted median income: 3.861855771345918

print(regr.coef_)
-0.01074958  0.00016731]
```

polynomial_regression.ipynb

I then threw in an **outlier** in the input data:
x = [1,2,3,5,6,7,8,9,10,12,13,14,15,16,18,19,21,22]
y = [100,90,80,60,60,**15**,60,65,70,70,75,76,78,79,90,99,99,100]



And one more **outlier** in the input data:
x = [1,2,3,5,6,7,8,9,10,12,13,14,15,16,18,19,21,22]
y = [100,90,80,60,60,**200**,60,65,70,70,75,76,78,79,90,99,99,100]



The learning outcomes for this e-portfolio activity were:

- Articulate the legal, social, ethical and professional issues faced by machine learning professionals.
- Understand the applicability and challenges associated with different datasets for the use of machine learning algorithms.

Throughout this exercise, I have seen first-hand how manipulation of the data can wildly impact the performance and results from correlation and regression machine learning analyses. This has legal, social and ethical implications as safe and appropriate handling/processing of data and model tuning are critical steps in using machine learning algorithms responsibly. This awareness reflects I am developing a responsible and professional approach to machine learning.

I have gained a greater understanding of the challenges associated with different datasets and their applicability in various machine learning algorithms. By manipulating datasets, observing changes in correlation and regression, and understanding the impact of data points on model performance, I have gained insight into the practical difficulties of working with real-world data. I have learned to identify and address these challenges, thereby enhancing my ability to select appropriate datasets and preprocessing techniques for different machine learning tasks. This understanding is vital for developing robust, accurate, and ethically sound machine learning solutions.