Python Project 4 - Polynomial Regression

Are we getting taller? There is no noticeable height change in humans from when record keeping started to the early 1800's but since that time the average height has been increasing. Today, on average, we are taller than our predecessors but these gains vary considerably from country to country. Researchers found that today Dutch men have the highest average heights at 182.5 centimeters (about 6 feet) and Latvian women at 170 cm (about 5 feet 6.5 inches) with Dutch women in second place. A century ago the United States ranked third in adult male height but it currently ranks 37th so the gain here is not as large as in other countries.

We want to use linear regression to model the growth in the average height of men or women (your choice) in the United States and the Netherlands.

The data for this lab was taken from the website ourworldindata.org which got the data from NCD RisC (a network of health scientists around the world).

The data is contained in the file human_heights.txt which consists of 5 columns of numerical data in the following order: year, Dutch men average height, Dutch women average height, US men average height, US women average height. The heights are given in centimeters.

1. Choose to compare either Dutch men to US men or Dutch women to US women
2. Import libraries
3. Enter the recorded values for comparison that I have provided
4. Read in data from file human_heights.txt; you can either put your data in 1D arrays or in a matrix. Plot your data for Dutch and US on same plot; add labels to axis, plot title and legend. To add a legend use the label option on plt.plot and the command plt.legend().
5. Use linear regression to fit the data for either men or women; plot both Dutch and US results on the same plot; add labels to axis, plot title and legend.
6. Calculate the variance
7. Use regression to fit the data for either men or women using a parabola; plot both Dutch and US results on the same plot; add labels to axis, plot title and legend.
8. Calculate the variance
9. Decide which fit (a line or a parabola) is better; justify your answer quantitatively. Then use whichever is better to predict the heights in 1955 and 1995; calculate the percent error in these predictions and the exact values I provided. The percent error should be positive and found by taking the absolute value of the difference in the predicted and actual heights divided by the actual height times 100 (to get percent). Are the predictions at different years comparable?

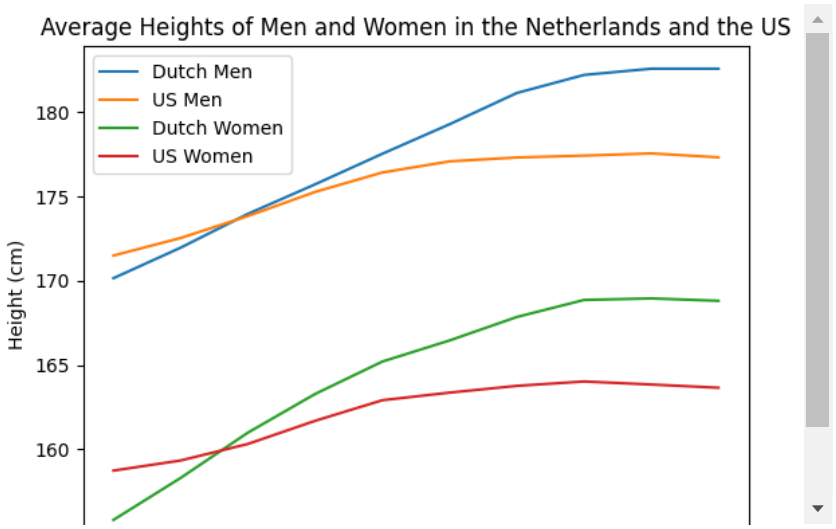
Indicate here whether you are comparing Dutch men to US men or Dutch women to US women

```
# Import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('human_heights.txt', sep='\t')
```

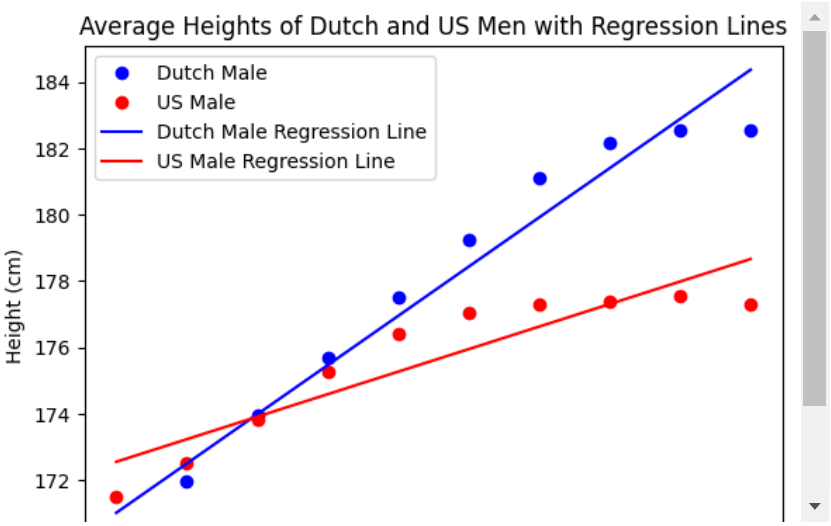
```
# Input exact values for comparison
#
dutch_man_1925 = 174.83; us_man_1925 = 174.53
dutch_man_1955 = 180.23; us_man_1955 = 177.22
dutch_man_1995 = 182.54; us_man_1995 = 177.16
#
dutch_woman_1925 = 162.2; us_woman_1925 = 160.97
dutch_woman_1955 = 167.11; us_woman_1955 = 163.54
dutch_woman_1995 = 168.73; us_woman_1995 = 163.56
```

```
# Read in the data and plot; add axes labels, plot title and legend

# Note: to add legend use label option in plt.plot (e.g., label='Dutch Men') and then use
# the command plt.legend ()

data = np.loadtxt('human_heights.txt', skiprows=1)

print(data.shape)

years = data[:, 0]
dutch_men = data[:, 1]
dutch_women = data[:, 2]
us_men = data[:, 3]
us_women = data[:, 4]

plt.plot(years, dutch_men, label='Dutch Men')
plt.plot(years, us_men, label='US Men')
plt.plot(years, dutch_women, label='Dutch Women')
plt.plot(years, us_women, label='US Women')
plt.xlabel('Year')
plt.ylabel('Height (cm)')
plt.title('Average Heights of Men and Women in the Netherlands and the US')
plt.legend()

plt.show()
```

```
(10, 5)
```




```
# Linear regression fit for both Dutch and U.S.; plot and print out the line
from sklearn.linear_model import LinearRegression

x = np.array(years).reshape(-1, 1)
y_dutch = np.array(dutch_men).reshape(-1, 1)
y_us = np.array(us_men).reshape(-1, 1)

model_dutch = LinearRegression().fit(x, y_dutch)
model_us = LinearRegression().fit(x, y_us)

plt.plot(years, y_dutch, 'bo', label='Dutch Male')
plt.plot(years, y_us, 'ro', label='US Male')
plt.plot(years, model_dutch.predict(x), 'b-', label='Dutch Male Regression Line')
plt.plot(years, model_us.predict(x), 'r-', label='US Male Regression Line')

plt.xlabel('Year')
plt.ylabel('Height (cm)')
plt.title('Average Heights of Dutch and US Men with Regression Lines')
plt.legend()

plt.show()
```

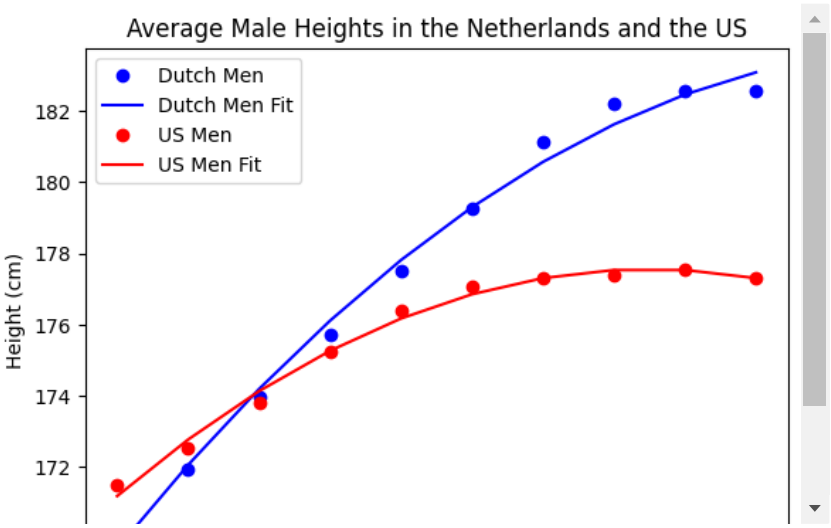



```
# Calculate the variance for each fit; use the function that we wrote in a previous notebook

#
# Input: the x and y arrays for the data points, coefficients of line found using LR
#
# Output: variance
#
def linreg_var(x, y, coefficients):
    y_pred = coefficients[0] + coefficients[1] * x
    residuals = y - y_pred
    squared_residuals = residuals ** 2
    variance = np.sum(squared_residuals) / (len(x) - 2)
    return variance

us_men_coeffs = np.polyfit(years, us_men, 1)
dutch_men_coeffs = np.polyfit(years, dutch_men, 1)

var_us_men = linreg_var(years, us_men, us_men_coeffs)
var_dutch_men = linreg_var(years, dutch_men, dutch_men_coeffs)
print(var_us_men)
print(var_dutch_men)
```

```
8883150425.100662
58515564152.00359
```

```
# Quadratic regression fit for Dutch and U.S.; plot and print out the parabolas
# Calculate the quadratic regression fit for Dutch
from scipy.optimize import curve_fit

def parabola(x, a, b, c):
    return a * x**2 + b * x + c
```

```
popt, pcov = curve_fit(parabola, years, dutch_men)

plt.plot(years, dutch_men, 'bo', label='Dutch Men')
plt.plot(years, parabola(years, *popt), 'b-', label='Dutch Men Fit')

popt, pcov = curve_fit(parabola, years, us_men)

plt.plot(years, us_men, 'ro', label='US Men')
plt.plot(years, parabola(years, *popt), 'r-', label='US Men Fit')
plt.xlabel('Year')
plt.ylabel('Height (cm)')
plt.title('Average Male Heights in the Netherlands and the US')
plt.legend()
plt.show()
```




```
# Calculate variance for the quadratic fits
def polyreg_var(x, y, coefficients):
    y_pred = np.polyval(coefficients, x)
    residuals = y - y_pred
    squared_residuals = residuals ** 2
    variance = np.sum(squared_residuals) / (len(x) - len(coefficients))
    return variance

us_men_coeffs = np.polyfit(years, us_men, 1)
dutch_men_coeffs = np.polyfit(years, dutch_men, 1)


var_us_men = polyreg_var(years, us_men, us_men_coeffs)
var_dutch_men = polyreg_var(years, dutch_men, dutch_men_coeffs)
print(var_us_men)
print(var_dutch_men)
```

```
0.8917043939393932
0.9458953030302724
```

Conclusion

Decide which fit is best; justify

```
# Use best fit to predict average heights in 1955 and 1995 for both Dutch and U.S.; compute percent error;
# round values to 2 decimal places
#
coefficients, _ = np.polyfit(years, dutch_men, 2, cov=True)
a, b, c = coefficients

predicted_1955 = a*(1955**2) + b*1955 + c
predicted_1995 = a*(1995**2) + b*1995 + c

actual_1955 = 182.9
actual_1995 = 182.8
percent_error_1955 = abs(predicted_1955 - actual_1955) / actual_1955 * 100
percent_error_1995 = abs(predicted_1995 - actual_1995) / actual_1995 * 100

predicted_1955 = round(predicted_1955, 2)
predicted_1995 = round(predicted_1995, 2)
percent_error_1955 = round(percent_error_1955, 2)
percent_error_1995 = round(percent_error_1995, 2)

print("Dutch men:")
print(f"Predicted height in 1955: {predicted_1955}")
print(f"Percent error in 1955: {percent_error_1955}%")
print(f"Predicted height in 1995: {predicted_1995}")
print(f"Percent error in 1995: {percent_error_1995}%")
```

```
Dutch men:
Predicted height in 1955: 179.96
Percent error in 1955: 1.61%
Predicted height in 1995: 183.31
Percent error in 1995: 0.28%
```

```
print("The quadratic fit is more accurate than the linear fit as the model follows a polynomial. Using the pc
```

```
The quadratic fit is more accurate than the linear fit as the model follows a polynomial. Using the polynomial regression we ca
```