DAE Group Project

2180804, 2136369, 1898593, 1712359 and 2228361

April 24, 2023

[1]

Question 1

Single Variable Regression

(a) Model 1a:

With the aim of predicting the maximum elevation of a country using only it's total area, we will first investigate the linear correlation between the two. Evidence of significant linear correlation between the two attributes will enable us to make (most likely) accurate predictions of one using the other.

In Model 1a, we use the respective country's total area as our predictor in measuring a country's maximum elevation.

In terms of the intuition regarding the slope of this model, we expect a positive relationship between the total area and the country's highest point, since the larger the country, the more likely it is to have higher levels of elevation. This intuitively makes sense, however, the magnitude of this relationship will vary depending on factors such as the topographical structure of a country, as well as its location.

To check whether the slope corresponds to our intuition from the scatter plot, we can visually inspect the scatter plot of total area and highest point. If the points appear to form a roughly linear trend with a positive slope, then the regression result is consistent with our intuition.

When considering outliers, we must understand that outliers can be problematic for our regression models as they impose a disproportionate influence on the slope of our regression line. To deal with this, can simply remove them from the model in total or use robust regression techniques that reduce the model's

[1] Group members: Sean Terespolsky(2180804), Robert Schwarz(2136369), Lindelani Delisa Dlamini(1898593), Lindani Dlamini(1712359) and James Thackeray(2228361)

sensitivity to outliers. In our case, we simply use the mean value of a specific attribute in place of the outlier in order to not affect the total size of our dataset and simultaneously lessen the impact that the outlier values may have on our model.

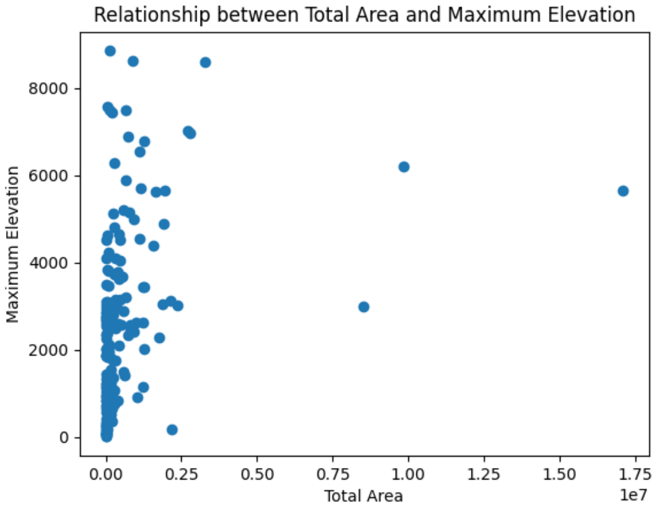
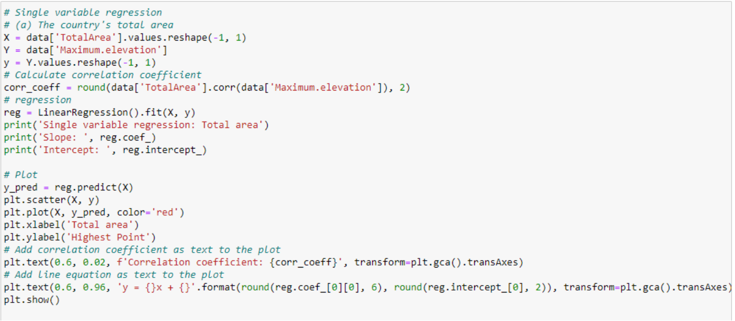


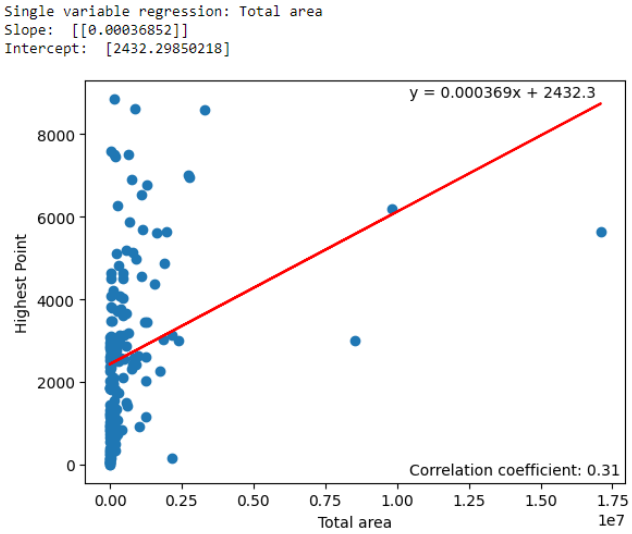
Figure 1: Scatter Plot showing relationship between total area and maximum elevation



```
# Single variable regression
# (a) The country's total area
X = data['TotalArea'].values.reshape(-1, 1)
Y = data['Maximum.elevation']
y = Y.values.reshape(-1, 1)
# Calculate correlation coefficient
corr_coeff = round(data['TotalArea'].corr(data['Maximum.elevation']), 2)
# regression
reg = LinearRegression().fit(X, y)
print('Single variable regression: Total area')
print('Slope: ', reg.coef_)
print('Intercept: ', reg.intercept_)

# Plot
y_pred = reg.predict(X)
plt.scatter(X, y)
plt.plot(X, y_pred, color='red')
plt.xlabel('Total area')
plt.ylabel('Highest Point')
# Add correlation coefficient as text to the plot
plt.text(0.6, 0.02, f'Correlation coefficient: {corr_coeff}', transform=plt.gca().transAxes)
# Add line equation as text to the plot
plt.text(0.6, 0.96, 'y = {}x + {}'.format(round(reg.coef_[0][0], 6), round(reg.intercept_[0], 2)), transform=plt.gca().transAxes)
plt.show()
```

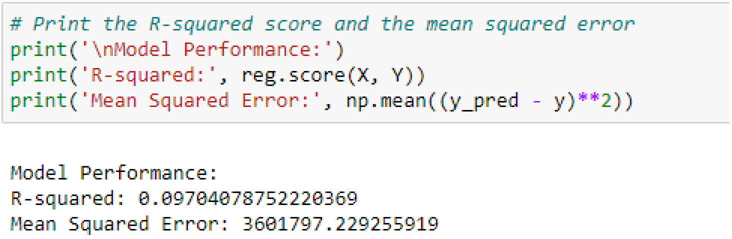
Figure 2: Notebook Code of Model 1a

The low R-Squared score of approximately 0.097 indicates that some of the other attributes are more likely to have an influence on the maximum elevation. A high Mean Squared Error depicts the fact that this model may not be a good fit for the data and thus, the predictions made may not be as accurate as we would like.



```
Single variable regression: Total area
Slope:  [[0.00036852]]
Intercept:  [2432.29850218]
```

Figure 3: Regression Line of Model 1a



```
# Print the R-squared score and the mean squared error
print('\nModel Performance:')
print('R-squared:', reg.score(X, Y))
print('Mean Squared Error:', np.mean((y_pred - y)**2))
```

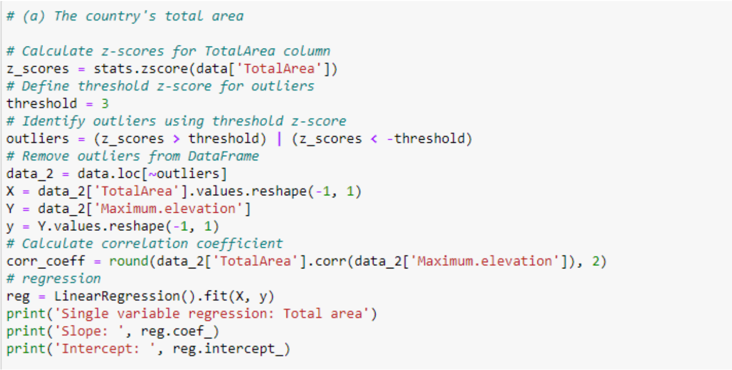
```
Model Performance:
R-squared: 0.09704078752220369
Mean Squared Error: 3601797.229255919
```

Figure 4: Performance metrics of Model 1a

Model 1b: (Outliers Removed)

From our scatterplot in Figure 1, we can clearly notice some extreme outliers that are present in the data. We first identify these outliers and remove them by using their z-scores from the mean. We used a threshold z-score of 3 as our cut-off point for outliers.

This can be done by making the following modification to the original code:



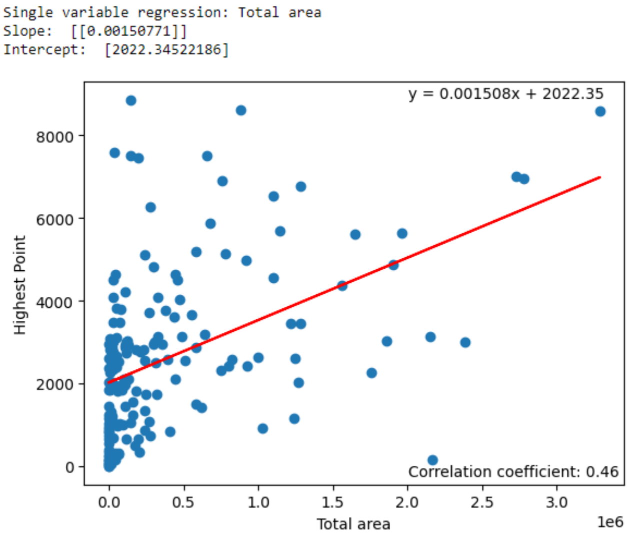
```
# (a) The country's total area

# Calculate z-scores for TotalArea column
z_scores = stats.zscore(data['TotalArea'])
# Define threshold z-score for outliers
threshold = 3
# Identify outliers using threshold z-score
outliers = (z_scores > threshold) | (z_scores < -threshold)
# Remove outliers from DataFrame
data_2 = data.loc[~outliers]
X = data_2['TotalArea'].values.reshape(-1, 1)
Y = data_2['Maximum.elevation']
y = Y.values.reshape(-1, 1)
# Calculate correlation coefficient
corr_coeff = round(data_2['TotalArea'].corr(data_2['Maximum.elevation']), 2)
# regression
reg = LinearRegression().fit(X, y)
print('Single variable regression: Total area')
print('Slope: ', reg.coef_)
print('Intercept: ', reg.intercept_)
```

Figure 5: Code to identify and remove outliers

Thereafter, we are left with a new scatter plot and regression line of the data given by Figure 6:

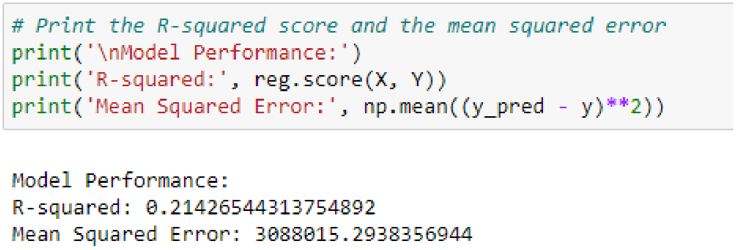
Hence, in Figure 6 we can see that the presence of the outliers had an impact on the slope and intercept of the regression line as well as the correlation coefficient. Removing them has reduced their influence on the data and thus the model overall.

In terms of the performance of Model 1b, we can see that we get a better r-squared score than Model 1A after we've dealt with the outliers - as given by Figure 7. This allows for the assumption that the Total Area may be one of the factors that have an influence on the highest point of a country. The two are relatively linearly correlated. The MSE is still a bit high though and this suggests that even though there is some linear correlation between the total area and the maximum point of elevation, the model may still not be accurate in making predictions.



```
Single variable regression: Total area
Slope:  [[0.00150771]]
Intercept:  [2022.34522186]
```

Figure 6: Scatter Plot with regression line showing relationship between total area and maximum elevation after dealing with outliers.



```
# Print the R-squared score and the mean squared error
print('\nModel Performance:')
print('R-squared:', reg.score(X, Y))
print('Mean Squared Error:', np.mean((y_pred - y)**2))
```

```
Model Performance:
R-squared: 0.21426544313754892
Mean Squared Error: 3088015.2938356944
```

Figure 7: Relevant performance metrics of Model 1b

(b) Model 2a:

In this context, we want to try and predict the highest point of elevation using the lowest point. If we are able to find a strong linear correlation between the two, then our model will be able to make predictions that would most likely be more accurate.

In Model 2a, we use the respective country's lowest point of elevation as our predictor in measuring a country's maximum elevation.

Intuitively, the slope of the regression line in this model represent the change in the response variable (highest elevation) per unit increase in the predictor variable (lowest elevation). A positive slope here indicates a positive relationship between the lowest and highest points, while a negative slope would indicate a negative relationship in this regard. The intuitive sense made by the slope in this model indicates a positive relationship. This means that in general, the highest and lowest points of a country are often located in the same region.

By analyzing the scatter plot for this dataset, we can examine the strength of the relationship between the 2 aforementioned variables. If the scatter plot shows a clear relationship (either positive or negative) between the lowest and highest points, and the slope of the regression line is consistent with this, then we can ascertain that our intuition coincides with the implications of the slope.

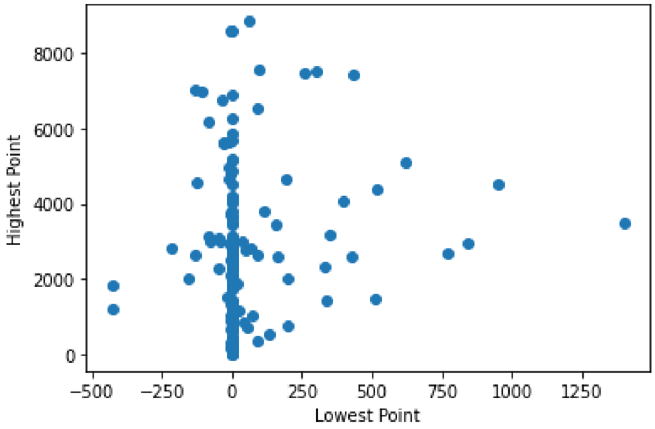
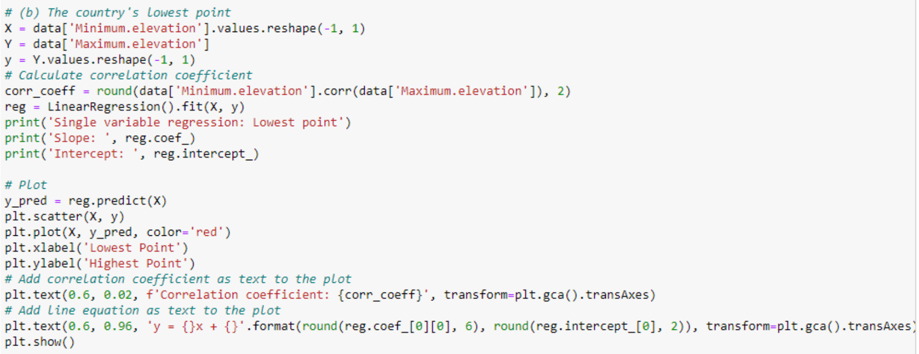
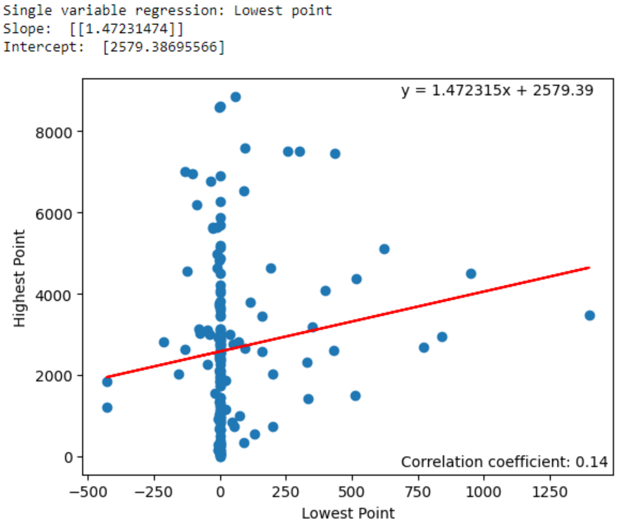


Figure 8: Scatter plot showing the relationship between minimum elevation and maximum elevation.



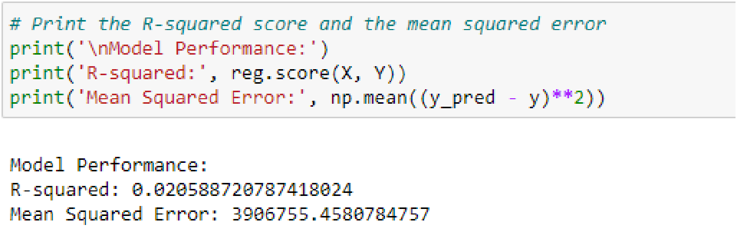
```
# (b) The country's lowest point
X = data['Minimum.elevation'].values.reshape(-1, 1)
Y = data['Maximum.elevation']
y = Y.values.reshape(-1, 1)
# Calculate correlation coefficient
corr_coeff = round(data['Minimum.elevation'].corr(data['Maximum.elevation']), 2)
reg = LinearRegression().fit(X, y)
print('Single variable regression: Lowest point')
print('Slope: ', reg.coef_)
print('Intercept: ', reg.intercept_)

# Plot
y_pred = reg.predict(X)
plt.scatter(X, y)
plt.plot(X, y_pred, color='red')
plt.xlabel('Lowest Point')
plt.ylabel('Highest Point')
# Add correlation coefficient as text to the plot
plt.text(0.6, 0.02, f'Correlation coefficient: {corr_coeff}', transform=plt.gca().transAxes)
# Add line equation as text to the plot
plt.text(0.6, 0.96, 'y = {}x + {}'.format(round(reg.coef_[0][0], 6), round(reg.intercept_[0], 2)), transform=plt.gca().transAxes)
plt.show()
```

Figure 9: Notebook code of the implementation of Model 2a.



```
Single variable regression: Lowest point
Slope:  [[1.47231474]]
Intercept:  [2579.38695566]
```

Figure 10: Regression Line of Model 2a.



```
# Print the R-squared score and the mean squared error
print('\nModel Performance:')
print('R-squared:', reg.score(X, Y))
print('Mean Squared Error:', np.mean((y_pred - y)**2))
```

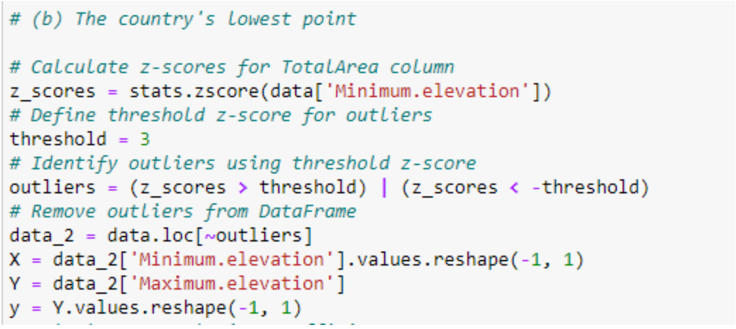
```
Model Performance:
R-squared: 0.020588720787418024
Mean Squared Error: 3906755.4580784757
```

Figure 11: Performance metrics of Model 2a.

Model 2b: (Outliers Removed)

Once again, it is clear that there are some outliers present in the current dataset, and dealing with these appropriately allows us to lessen their impact on the model as a whole. We first identify these outliers and remove them by using their z-scores from the mean.

This can be done by making the following modification to the original code:



```
# (b) The country's lowest point

# Calculate z-scores for TotalArea column
z_scores = stats.zscore(data['Minimum.elevation'])
# Define threshold z-score for outliers
threshold = 3
# Identify outliers using threshold z-score
outliers = (z_scores > threshold) | (z_scores < -threshold)
# Remove outliers from DataFrame
data_2 = data.loc[~outliers]
X = data_2['Minimum.elevation'].values.reshape(-1, 1)
Y = data_2['Maximum.elevation']
y = Y.values.reshape(-1, 1)
```

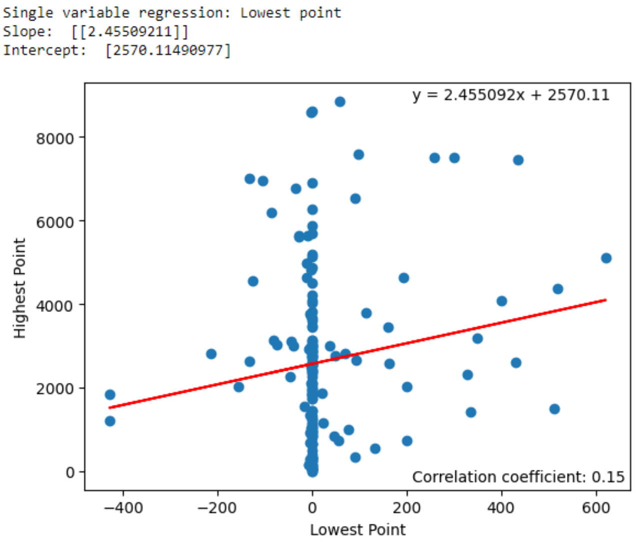
Figure 12: Code modification to remove outliers from Model 2a.

Thereafter, we are left with a new regression line that models the relationship between a country's lowest point of elevation and the country's maximum point of elevation.

Thus, we can see that the presence of the outliers had little impact on the slope and intercept of the regression line as well as the correlation coefficient. The slope of the regression line increased from 1.47231474 to 2.45509211, indicating that for each unit increase in Lowest point (x), the Highest point is expected to increase on average by 2.4550921.

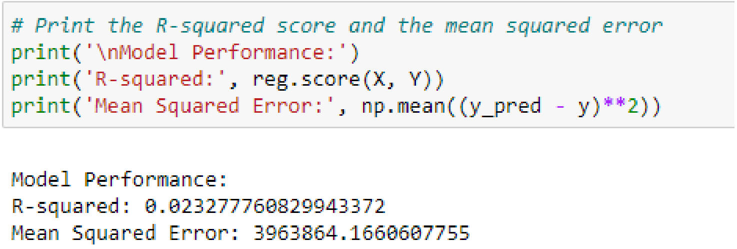
The correlation coefficient also increased from 0.14 to 0.15, indicating that there is almost no linear relationship between the two variables. There isn't much increase in a country's highest point with increase in lowest point. It makes some intuitive sense because there isn't much of a relationship between a country's highest and lowest points in the real world.

The intercept also changed from 2579.38695566 to 2570.11490977. Since the lowest point will continue to decrease, we expect a country with lowest point less than -400 to have a highest point of less than 2570.11490977.



```
Single variable regression: Lowest point
Slope:  [[2.45509211]]
Intercept:  [2570.11490977]
```

Figure 13: Scatter plot and Regression Line of Model 2b showing the relationship between minimum and maximum points of elevation, after dealing with outliers.



```
# Print the R-squared score and the mean squared error
print('\nModel Performance:')
print('R-squared:', reg.score(X, Y))
print('Mean Squared Error:', np.mean((y_pred - y)**2))
```

```
Model Performance:
R-squared: 0.023277760829943372
Mean Squared Error: 3963864.1660607755
```

Figure 14: Relevant performance metrics of Model 2b.

Question 2

Multiple Variable Regression

Now, to compare the correlation between the different predictor variables, we create a correlation matrix to visualize the relevant inter dependencies between the variables.

In Figure 15, we can see that the correlations between the different area metrics of a country are all closely related with one another. Intuitively this makes sense as we would expect the total ares to be a loose sum of the land and water areas respectively. In a similar way, we can understand the relationship that these 3 areas have on one another.

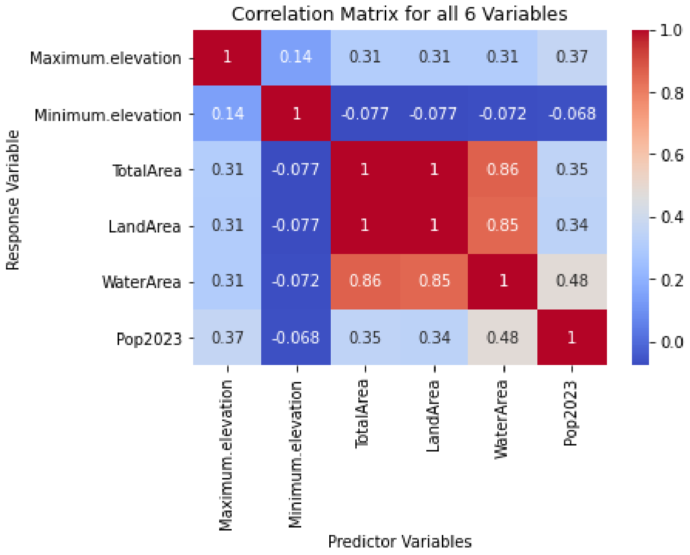


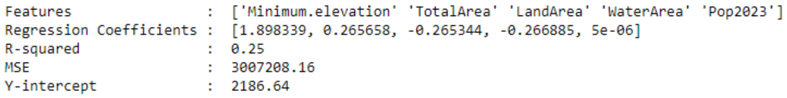
Figure 15: Correlation Matrix for all the given variables.

To determine the significance of each predictor variable in the model, we would look at the p-values associated with their coefficients. A low p-value (< 0.05) indicates that the predictor variable is statistically significant in predicting the response variable.

Additionally, we would want to check for multicollinearity between the predictor variables. Multicollinearity occurs when two or more predictor variables are highly correlated with each other, which can cause issues with the interpretation of the coefficient estimates. One way to check for multicollinearity is to calculate the correlation matrix between the predictor variables and check for high correlations (above 0.7 or 0.8).

In our analysis, we include the following multiple variable models using different combinations of the predictor variables:

- **Model 1:** All 5 available features (minimum elevation, total area, land area, water area and population)



```
Features                : ['Minimum.elevation' 'TotalArea' 'LandArea' 'WaterArea' 'Pop2023']
Regression Coefficients : [1.898339, 0.265658, -0.265344, -0.266885, 5e-06]
R-squared               : 0.25
MSE                     : 3007208.16
Y-intercept             : 2186.64
```

Figure 16: All 5 available features used as predictor variables.

  In the results of this model, as given by Figure 16, we can make some insights into the model itself. For instance, the regression coefficients given show how much the response variable (highest elevation) changes with regard to a unit change in each predictor variable/feature. Clearly, the minimum elevation has the largest regression coefficient, while total area, land area and water area all have similar magnitudes of regression coefficients. This implies that changes to these predictor variables will have different impacts on the overall model.

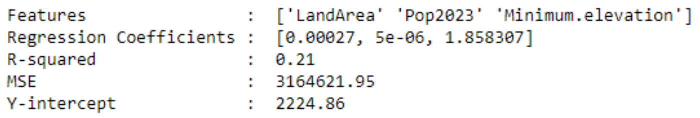
  By analyzing the correlations between all the variables given by Figure 15, we can examine that the correlation between maximum elevation (our response variable) and our various predictor variables. By analysing the various correlations coefficients, we can clearly observe that the largest correlations exist between the response variable and population (0.37), followed by water area, land area and total area (all with 0.31) and finally with minimum elevation (0.14).

  Hence, when deciding which variables to remove in our subsequent models, it is important to understand the overall impact incurred by removing certain features from the regression model.

- **Model 2:** Here, we choose to leave out two different features from our new regression model. We have done so by deciding that since all the area metrics have the same correlation with respect to the response variable, there is little to no benefit of using all 3 simultaneously. Thus, our new revised model will only consider minimum elevation, land area and population as our 3 predictor variables only.

Intuitively, this combination makes sense as the minimum elevation is expected to have a relationship with the maximum elevation since countries with a higher minimum will in all likelihood have a higher maximum elevation. Similarly, the intuition behind using land area as a predictor is based on the fact that this measure excludes the total area of water in a terrain which allows us to use the fact that larger countries (in terms of land) tend to have more varied topography which is usually associated with a higher maximum elevation. Finally, in terms of the use of population as a predictor variable, we first need to understand that population indirectly implies settlement of the people across the country as well as the resources available to that population. This can be seen by the fact that countries with higher population density are more likely to be located at lower elevations, while areas with lower population densities are more likely to be located at higher elevations. Similarly, countries with larger populations may have more resources available to support infrastructure development, which in turn may facilitate access to higher elevations. Additionally, countries with larger populations may have more resources available to support scientific research and exploration, which may lead to the discovery of new mountain ranges or peaks.

The revised model can be seen in the following:



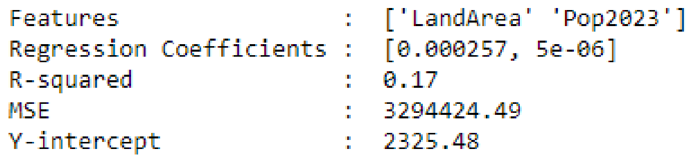
```
Features                : ['LandArea' 'Pop2023' 'Minimum.elevation']
Regression Coefficients : [0.00027, 5e-06, 1.858307]
R-squared               : 0.21
MSE                     : 3164621.95
Y-intercept             : 2224.86
```

Figure 17: Results of using land area, populations and minimum elevation used as predictor variables

- **Model 3:** In this model, we further remove another predictor variable from the model in Figure 17. Specifically, we chose to remove minimum elevation as, according to Figure 15, it has the lowest correlation with our particular response variable (maximum elevation). Another reason for the removal of minimum elevation is that minimum and maximum elevation of a country are not necessarily directly related, since these measures do not capture the terrain topography in its entirety. This is due to the fact that a country with a low minimum elevation could still have high mountains, and a country with a high minimum elevation could still have relatively flat terrain.

In a similar way, it's possible that the range of minimum elevation values is not very large for the countries in the dataset, which could make it difficult to detect any significant relationship with maximum elevation.

Hence our revised model produces the following results:



```
Features                 :  ['LandArea' 'Pop2023']
Regression Coefficients :  [0.000257, 5e-06]
R-squared                :  0.17
MSE                      :  3294424.49
Y-intercept              :  2325.48
```

Figure 18: Results of using land area and population used as predictor variables.

Question 3

If accuracy of the prediction is our only concern when deciding which of the above models would be most effective, then it is clear that that the multiple variable regression model that considers all 5 features as the relevant predictor variables to estimate the response variable (maximum elevation) would be most suitable. This is directly a result of both its highest R^2 value which is 0.25 and its lowest MSE value which is given by 3007208.16.

Although the model given by Figure 16 provides the most accuracy, it is important to consider the computational complexity associated with using every possible feature as a predictor variable as it is obviously higher to do so than to use a smaller number of the given features.

In our above analyses, we considered the dataset in its totality. However, when we investigated the idea of splitting the dataset into testing and training subsets and built our model in this manner, we saw very little benefit in general of the training-testing approach - especially when concerned with overall accuracy.

Finally, it is clear from our various models that, in terms of linear regression on this dataset, the most accurate model would be one that encapsulates as much information as possible with regards to the given features, as seen by Figure 16. However, when deciding which model is most suitable, we should perhaps consider the features that have the most significant positive impact on our final model and build a linear regression model around these.

In our rigorous testing, we ascertained that the most viable combination of features (4 or less) will include the land area, population and minimum elevation as these factors most intuitively contribute to the accuracy of a model in estimating the maximum elevation of a given country.