Statistical Analysis using Excel and Python

Contents

Abstract

This work presents the statistical analysis of the world population dataset using Excel and Python. This report includes statistical information, data visualization, and inference of statistical data and visualization.

The project began with an in-depth analysis in Excel, where statistical metrics like mean, median, mode, quartiles, variance, standard deviation, Quartiles, and lower and upper limits. Further analysis was conducted in Python to validate Excel results. Python's capabilities enabled precise calculations and visualization of data distributions through box plots, standard deviation diagrams, and linear regression plots.

By comparing results from Excel and Python, this report highlights any discrepancies and the unique advantages of each tool for statistical analysis. This project provides valuable insights

into global demographic patterns, offering a comprehensive view of the population and area dynamics across countries and continents.

Introduction

This dataset has historical population data for every country and territory worldwide and various demographic and geographic parameters, including the area size, continent, capital city, population density, growth rate, population rank, and world population share. This report aims to perform a statistical analysis of global population distribution and area characteristics using Excel and Python to explore key metrics and validate findings.

The analysis covers a range of questions to understand demographic trends and population distributions across countries and continents. Specifically, this report investigates:

1. What are the most and least populated countries in 2022?
2. Which country has the largest and smallest land area?
3. How does the population distribution differ among countries?
4. Determining the skewness of the data based on the statistical information.
5. The comparison of statistical analysis using Python and Excel

This report provides a detailed summary of the dataset's core metrics using descriptive statistics such as mean, median, variance, and interquartile range. Additionally, visualizations like box plots and standard deviation diagrams enhance the interpretation of results, while linear regression offers insight into potential relationships within the data.

Solution

Statistical Calculation in Excel

The statistical analysis is done for the 2022 Population and Area. For each of the selected columns, we compute the following:

1. **Min**
2. **Max**
3. **Count**
4. **Range**
5. **Interquartile Range**
6. **Median**
7. **Mean**
8. **Mode**
9. **1st and 3rd Quartiles**
10. **Sample Variance**
11. **Sample Standard Deviation**
12. **Finding Outliers limits – Lower and Upper Limit**

To compute the statistical measures, we use the following formulas in Excel for the 2022 Population and Area.

In the following table, we have the generic Excel formula and what has been used in the Excel Workbook for each measure for the 2022 Population Feature.

STATISTICAL MEASURE	EXCEL FORMULA	FORMULA USED IN EXCEL WORKBOOK
MIN	=MIN(Range)	=MIN(A:A)
MAX	=MAX(Range)	=MAX(A:A)
COUNT	=COUNT(Range)	=COUNT(A:A)
RANGE	=MAX (range) - MIN (range)	=E6-E5 As we have already calculated the Max and Minimum, we subtract them
INTERQUARTILE RANGE	=QUARTILE (range, 3) - QUARTILE (range, 1)	=E13-E12 As we have calculated the Q1 and Q3, we subtract them
MEDIAN	=MEDIAN (range)	=MEDIAN(A:A)
MEAN	=AVERAGE (range)	=AVERAGE(A:A)
MODE	= MODE (range)	=MODE(A:A)
Q1 – 1ST QUARTILE	=QUARTILE (range, 1)	=QUARTILE(A:A, 1)
Q3 – 3RD QUARTILE	=QUARTILE (range, 3)	=QUARTILE(A:A, 3)
VARIANCE	=VAR.S(range)	=VAR.S(A2:A235)
STANDARD DEVIATION	=STDEV.S(range)	=STDEV.S(A:A)
QUARTILE LOWER LIMIT		
QUARTILE UPPER LIMIT		

Statistical Analysis in Python

Using the Jupiter Notebook, we calculate the statistical measures previously calculated using Excel formulas. A statistical_analysis function calculates all the statistical measures for a particular column to get the statistical measures.

```
In [10]: def statistical_analysis(column, column_name):
             print(f"\nStatistical Analysis for {column_name}:")

             # Count
             count = column.count()
             print(f"Count: {count}")

             # Minimum and Maximum
             min_val = column.min()
             max_val = column.max()
             print(f"Minimum: {min_val}")
             print(f"Maximum: {max_val}")

             # Range (Max - Min)
             range_val = max_val - min_val
             print(f"Range: {range_val}")

             # Interquartile Range (IQR)
             Q1 = column.quantile(0.25)
             Q3 = column.quantile(0.75)
             IQR = Q3 - Q1
             print(f"Interquartile Range (IQR): {IQR}")

             # Lower and Upper Limits for Outliers
             lower_limit = Q1 - 1.5 * IQR
             upper_limit = Q3 + 1.5 * IQR
             print(f"Lower Limit (Q1 - 1.5*IQR): {lower_limit}")
             print(f"Upper Limit (Q3 + 1.5*IQR): {upper_limit}")

             # Median
             median = column.median()
             print(f"Median: {median}")

             # Mean
             mean = column.mean()
             print(f"Mean: {mean}")

             # Mode
             mode = column.mode()
```

Statistical Analysis function

COLUMN	2022 POPULATION	AREA
COUNT	234	234
MINIMUM	510	1
MAXIMUM	1425887337	17098242
RANGE	1425886827	17098241
INTERQUARTILE RANGE	22056766.25	427775.75
Q1	419738.5	2650.0
Q3	22476504.75	430425.75
LOWER LIMIT	-32665410.875	-639013.625
UPPER LIMIT	55561654.125	1072089.375
MEDIAN	5559944.5	81199.5
MEAN	34074414.70940171	581449.3846153846
MODE	510	21
VARIANCE	1.870505495387683e+16	3104083230282.7534
STANDARD DEVIATION	136766424.80476278	1761840.8640631405

Statistical Measures calculated using Python

Definition

2022 Population

1. **Min**: The minimum population is 510, belonging to Vatican City.
2. **Max**: The maximum population is 1,425,887,337, belonging to China.
3. **Count**: There are 234 countries in the dataset.

4. **Range**: The 2022 Population range is 1,425,886,827, indicating a substantial difference between the least populated country (Vatican City) and the most populated country (China).
5. **Interquartile Range (IQR):** The IQR is approximately 22,056,766.25, meaning that the middle 50% of countries have populations within this range.
6. **Median**: The median population is approximately 5,559,944.5, indicating that half of the countries have populations below this value.
7. **Mean**: The mean population is around 34,074,414.71, which is considerably higher than the median due to the skewed distribution caused by a few highly populated countries like China and India.
8. **Mode**: No mode (N/A) indicates that each country has a unique population size.
9. **Q1 (1st Quartile):** 419,738.5, indicating that 25% of countries have populations below this value.
10. **Q3 (3rd Quartile):** 22,476,504.75, indicating that 75% of countries have populations below this value.
11. **Variance:** The variance is approximately 1.87 x 10^{16}, showing high variability due to a few extremely populous countries.
12. **Standard Deviation**: The standard deviation is approximately 136,766,424.8, reflecting the substantial spread in population sizes.
13. **Lower Limit**: The calculated lower limit is -32,665,410.88 (not meaningful since the population cannot be negative).
14. **Upper Limit**: The calculated upper limit is 55,561,654.13. Countries with populations above this limit (e.g., India, United States) are considered unusually populous, although no countries in the dataset were classified as outliers based on the IQR method.

Area

1. **Min**: The minimum area is 1 km^2, belonging to Vatican City.
2. **Max**: The maximum area is 17,098,242 km^2, belonging to Russia.
3. **Count**: There are 234 countries in the dataset.
4. **Range**: The area range is 17,098,241 km^2, indicating a substantial difference between the smallest country (Vatican City) and the largest country (Russia).
5. **Interquartile Range (IQR):** The IQR is approximately 427,775.75 km^2, meaning that the middle 50% of countries have areas within this range.
6. **Median**: The median area is approximately 81,199.5 km^2, indicating that half of the countries have areas below this value.
7. **Mean**: The mean area is around 581,449.38 km^2, which is considerably higher than the median due to a few exceptionally large countries skewing the distribution.
8. **Mode**: The mode is 21 km^2, indicating that this area value appears most frequently. However, if all values were unique, Excel would return #N/A while Python defaults to showing the first value.

9. **Q1 (1st Quartile):** 2,650 km^2, indicating that 25% of countries have areas below this value.
10. **Q3 (3rd Quartile):** 430,425.75 km^2, indicating that 75% of countries have areas below this value.
11. **Variance**: The variance is approximately 3.10408 x 10^{12} km^2, indicating a high degree of variability due to a few countries with extremely large areas.
12. **Sample Standard Deviation:** The standard deviation is approximately 1,761,840.86 km^2, reflecting a substantial spread in land sizes.
13. **Lower Limit**: The calculated lower limit is -639,013.625 km^2 (not meaningful since area cannot be negative).
14. **Upper Limit:** The calculated upper limit is 1,072,089.375 km^2. Countries with areas above this limit (e.g., Russia, Canada) are considered unusually large, though no countries in the dataset were classified as outliers based on the IQR method.

Comparison between Excel and Python

2022 Population	Excel	Python
Min	510	510
Max	1425887337	1425887337
Count	234	234
Interquartile Range	22056766.25	22056766.25
Mean	34074414.71	34074414.71
Median	5559944.5	5559944.5
Mode	#N/A	510
Q1	419738.5	419738.5
Q3	22476504.75	22476504.75
Quartile Upper Limit	55561654.13	55561654.13
Quartile Lower Limit	-32665410.88	-32665410.9
Variance	1.87051E+16	1.87E+16
Standard Deviation	136766424.8	136766424.8
Range	1425886827	1425886827

Statistical Measure for 2022 Population

The statistical analysis using Excel and Python reveals a range of insights, from the minimum population of 510 (Vatican City) to a maximum population of over 1.4 billion (China), indicating significant variation among countries. The mean population is around 34 million, much higher

than the median of approximately 5.6 million, suggesting a right-skewed distribution. This skewness is driven by a few densely populated countries like China and India, resulting in high variance and a large standard deviation. The interquartile range (IQR) of about 22 million indicates that the middle 50% of countries fall within this range, while the upper limit of 55.6 million marks the threshold for unusually populous countries. However, no outliers were identified using the IQR method, as all values fall within the calculated limits.

A key difference between Excel and Python emerged in calculating the mode. Excel returned #N/A because it does not consider a mode to exist when all values are unique. In contrast, Python returned 510 as the mode, defaulting to the first value in the dataset when there is no repetition.

Area	Excel	Python
Min	1	1
Max	17098242	17098242
Count	234	234
Interquartile Range	427775.75	427775.75
Mean	581449.3846	581449.3846
Median	81199.5	81199.5
Mode	21	21
Q1	2650	2650
Q3	430425.75	430425.75
Quartile Upper Limit	1072089.375	1072089.375
Quartile Lower Limit	-639013.625	-639013.625
Variance	3.10408E+12	3.10408E+12
Standard Deviation	1761840.864	1761840.864
Range	17098241	17908241

Statistical Measures for Area

The statistical analysis of country areas using Excel and Python shows significant variation, ranging from a minimum of 1 km² (Vatican City) to a maximum of 17,098,242 km² (Russia). The mean area is around 581,449 km², much higher than the median of 81,199.5 km², indicating a right-skewed distribution driven by a few large countries. The interquartile range (IQR) is approximately 427,775.75 km², with an upper limit of 1,072,089.375 km² beyond which countries are considered outliers, although no true outliers were found based on the IQR method. The high standard deviation of 1,761,840.864 km² reflects the spread due to a few exceptionally large countries. Excel returned #N/A for mode due to unique values, while Python

returned 21 km², defaulting to the first value when no mode exists. This analysis highlights the uneven global land distribution dominated by a few large countries.

STATISTICAL MEASURE	PYTHON	EXCEL	COMPARISON
MIN	Accurate	Accurate	Both
MAX	Accurate	Accurate	Both
COUNT	Accurate	Accurate	Both
INTERQUARTILE RANGE	Accurate	Accurate	Both
MEAN	Accurate	Accurate	Both
MEDIAN	Accurate	Accurate	Both
MODE	Accurate	Inacurate (# NA for some values)	Excel is more accurate; Python returns the first value if the mode is not available.
Q1	Accurate	Accurate	Both
Q3	Accurate	Accurate	Both
QUARTILE UPPER LIMIT	Accurate	Accurate	Both
QUARTILE LOWER LIMIT	Accurate	Accurate	Both
VARIANCE	Accurate	Slightly Rounded	Python
STANDARD DEVIATION	Accurate	Accurate	Both
RANGE	Accurate	Slightly Rounded	Python

Python provides more accurate calculations for most measures but returns the first value if the mode is unavailable.

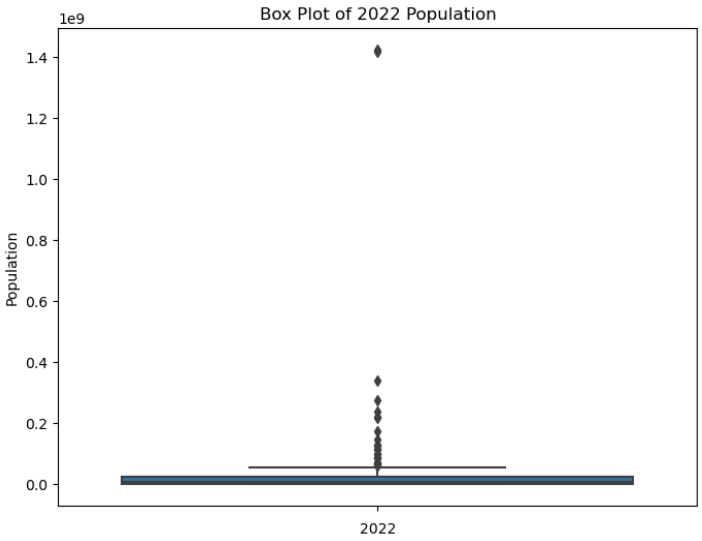
Data Visualization

In this section, we visualize the outliers and the relationship between the columns. We plot the following visualizations:

1. Box Plot
2. Standard Deviation
3. Linear Regression

Box Plot

The Box Plot visualizes the Outliers for a selected column, we visualize the outliers for Area and 2022 Population

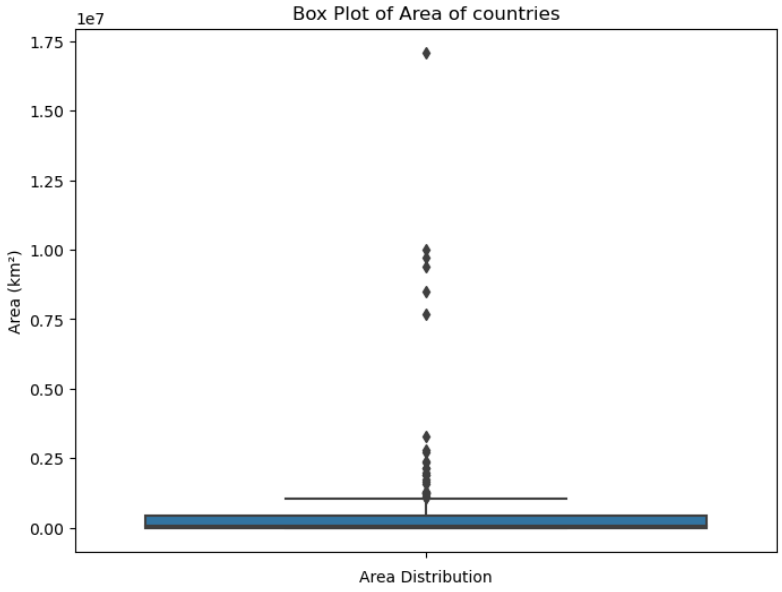


2022 Population Box Plot

This figure shows that most countries have a low population, as seen by compact distribution. However, a few countries have extremely high populations, which are displayed as outliers, making the overall distribution **highly right-skewed.**

The data points above the upper whisker represent outliers. These are countries with populations significantly higher than the majority. The largest outlier (approximately 1.4 billion) likely represents a country like China or India, with populations far exceeding those of other nations. Smaller outliers, near 0.2 to 0.4 billion, represent countries with moderately high populations, such as the United States, Indonesia, and Brazil.

Now, we plot the boxplot for Area.



Box Plot for Area Column

This box plot of **Area (km²)** for countries provides a clear visual of the distribution and highlights the presence of extreme outliers in land area among countries. The plot shows that most countries have a low area, with some having a very large area.

The highest outlier, at around **1.75 x 10^7** km² (17,500,000 km²), likely represents Russia, the largest country by area. Other large outliers may include countries like Canada, China, the United States, and Brazil, which also have significant land masses.

Standard Deviation Plot

Standard Deviation Plot

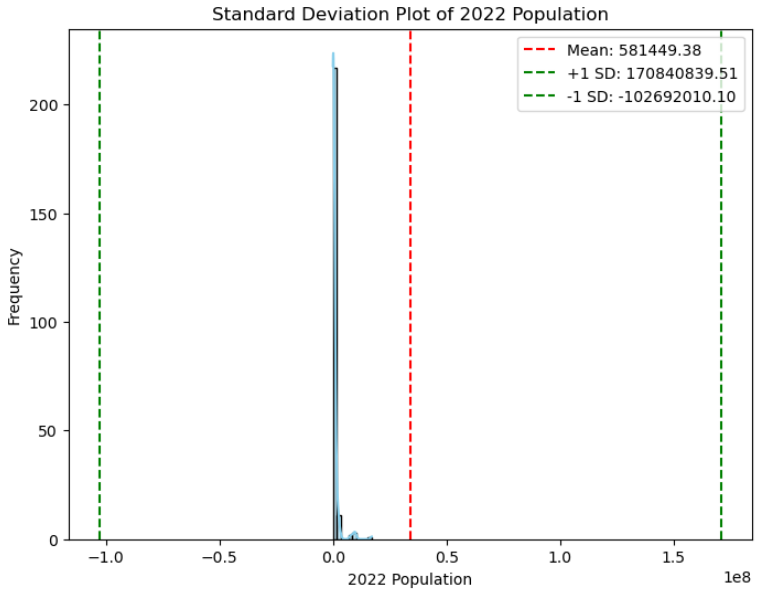
In [17]:
```
mean_area = df['Area (km²)'].mean()
std_dev_area = df['Area (km²)'].std()
```

In [21]:
```
import matplotlib.pyplot as plt
# Plot histogram with KDE (Kernel Density Estimate)
plt.figure(figsize=(8, 6))
sns.histplot(df['Area (km²)'], kde=True, color='skyblue', bins=10)

# Add mean and standard deviation lines
plt.axvline(mean_area, color='red', linestyle='--', label=f'Mean: {mean_area:.2f}')
plt.axvline(mean_area + std_dev_area, color='green', linestyle='--', label=f'+1 SD: {mean_area + std_dev_area:.2f}')
plt.axvline(mean_area - std_dev_area, color='green', linestyle='--', label=f'-1 SD: {mean_area - std_dev_area:.2f}')

# Add title and labels
plt.title('Standard Deviation Plot of Area')
plt.xlabel('Area (km²)')
plt.ylabel('Frequency')
plt.legend()

# Show the plot
plt.show()
```

```
C:\Users\daivi\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will
be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

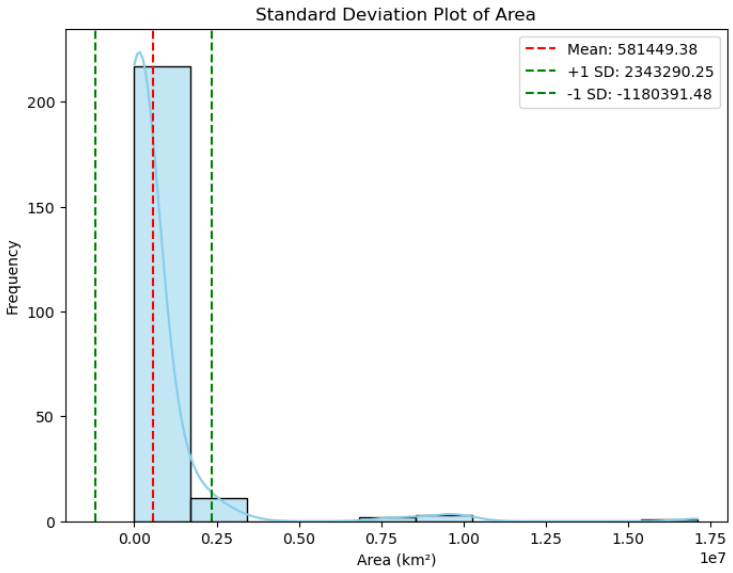
Standard Deviation Plot of Area

Python Code for Standard Deviation Plot



Standard Deviation Plot for 2022 Population

This standard deviation plot of the 2022 Population data shows an extremely right-skewed distribution, with most countries having small populations clustered near zero and a few countries with very high populations causing a significant skew. The mean population is around 58 million, but due to the skewed data, it's not centrally located within the main cluster, as countries like China and India raise the average. The ±1 standard deviation lines, widely spaced at approximately 170 million and -102 million, reflect high variance in population sizes across countries. The negative -1 SD line underscores the limitation of using mean and standard deviation to describe such skewed data, as it is not meaningful for population values.



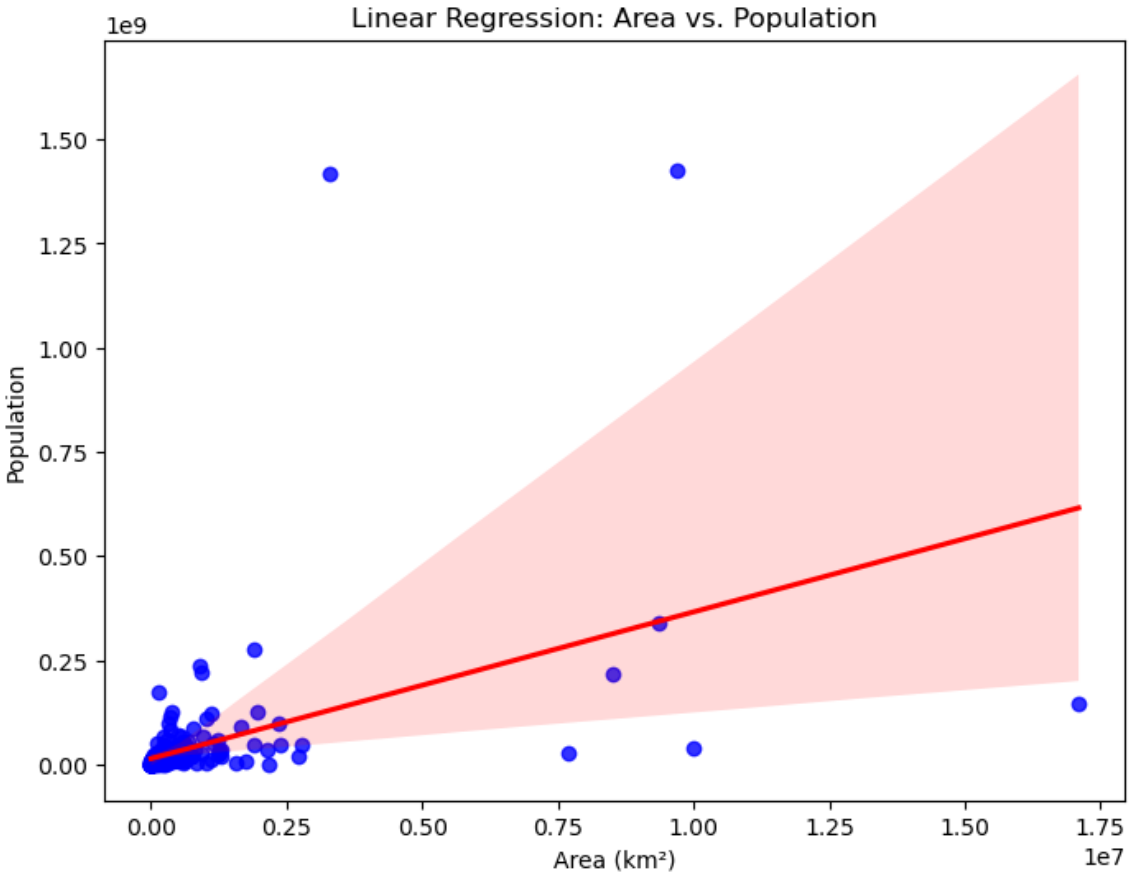
Standard Deviation Plot for Area

This standard deviation plot of **Area (km²)** reveals a highly **right-skewed distribution,** where most countries have relatively small land areas clustered near zero, and a few countries with exceptionally large areas create an extended tail on the right. The mean area is approximately 581,449 km².

However, like the population data, it is not representative of the typical country size due to a small number of extremely large values (e.g., Russia, Canada). The ±1 standard deviation lines are widely spaced, with +1 SD reaching around 2.34 million km² and -1 SD extending to a negative area (around -1.18 million km²), which is not meaningful in this context.

This large spread and a negative SD highlight the limitations of using mean and standard deviation for such skewed data, indicating that measures like median and interquartile range would better represent the central tendency and variability of country areas.

Linear Regression Plot



The linear regression plot between **Area** and **Population** shows a weak positive correlation, suggesting that larger areas are slightly associated with higher populations, but the relationship is not strong. The data points are widely scattered, with many large-area countries having low populations and some small-area countries with high populations. Outliers, like densely populated countries (e.g., China, India) and vast but sparsely populated regions (e.g., Russia), distort the trend, indicating that area alone cannot predict population size. This highlights that population is influenced by multiple factors beyond just land area.

Summary

Using Excel and Python, the statistical summary of the World Population Data's 2022 Population and Area Column yielded several insights into the global demographics and geographic distributions.

The analysis highlighted the concentration of population and land area among a few countries, with most countries having relatively modest sizes and populations. Python proved to be more effective for comprehensive statistical analysis, and for creating visualizations, while Excel was more straightforward for quick calculations. The experience

of using both tools provided a deeper understanding of global demographic patterns and the strengths and limitations of each tool in statistical analysis.

Reflection

This project gave me valuable experience performing statistical analysis using Python and Excel. I learned to compute essential statistical measures in each tool, from mean and median to interquartile range and standard deviation. While Python provided higher accuracy and flexibility, I appreciated the ease of using Excel's built-in formulas, which make basic calculations accessible and quick.

Working with Python allowed me to explore various visualizations, including box plots, standard deviation plots, and linear regression plots, and deepened my understanding of interpreting these visuals.

I learned that box plots are useful for identifying outliers and understanding data spread. In contrast, the standard deviation plot showed me how data points are distributed around the mean, especially in skewed datasets.

Additionally, creating and analyzing the linear regression plot helped me grasp the relationship (or lack thereof) between area and population, reinforcing how visualizations can uncover trends and patterns that may not be immediately visible in raw data. Overall, this project has equipped me with a solid statistical analysis and data visualization foundation, enhancing my ability to interpret complex demographic data effectively.