ESKOM

ELECTRIFICATION BY PROVINCE

PYTHON PROJECT

ABSTRACT

South Africa's electrification (access to electricity) drive aims to bring clean and reliable electricity to unserved communities, fostering social and economic development. However, achieving these goals requires robust data analysis and insightful metrics. This project explores the use of Python functions to calculate electrification metrics. It details how Python functions can analyse factors like population demographics, geographic distribution, and existing infrastructure to assess electrification progress and identify areas most in need. By employing Python's computational power and flexibility, this project demonstrates the potential for data-driven decision-making in achieving South Africa's ambitious electrification goals.

ABOUT THE DATASET

The datasets used in this project are “Electrification by Province” and “Twitter Nov 2019” csv files. The “Electrification by Province” dataset consists of 10 columns:

- Financial Year (1 April – 30 March)
- Limpopo
- Mpumalanga
- North West

- Free State
- Kwazulu Natal
- Eastern Cape
- Western Cape
- Northern Cape
- Gauteng

The “Twitter Nov 2019” dataset consists of 2 columns:

- Tweets
- Date

OBJECTIVES

1. Write a function that calculates metrics and outputs data as a dictionary
2. Write a function that returns a dictionary of the 5-number summary
3. Write a function that takes in a date and output a string with a correct format
4. Write a function that returns a data frame with (1).an added column of extracted hashtags from each tweet, and (2).an added column of the municipality mentioned in each tweet
5. Write a function that returns number of tweets per day
6. Write a function that splits a sentence into a list of individuals words
7. Write a function that returns a data frame with an added column of the tweets without stop-words

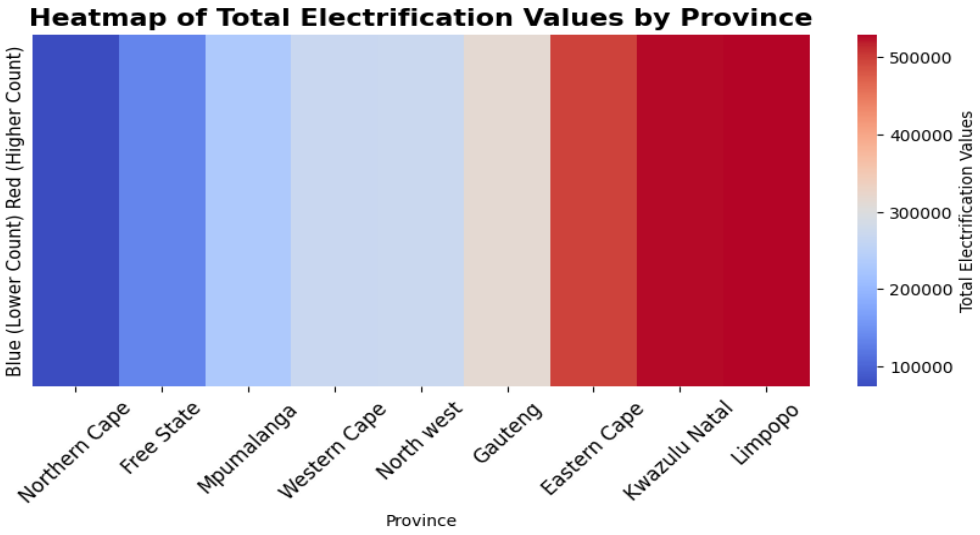
DESCRIPTIVE STATISTICS

Below is a snippet of the data frame that shows the first few rows of the dataset as well as the columns needed for the stats of this project.

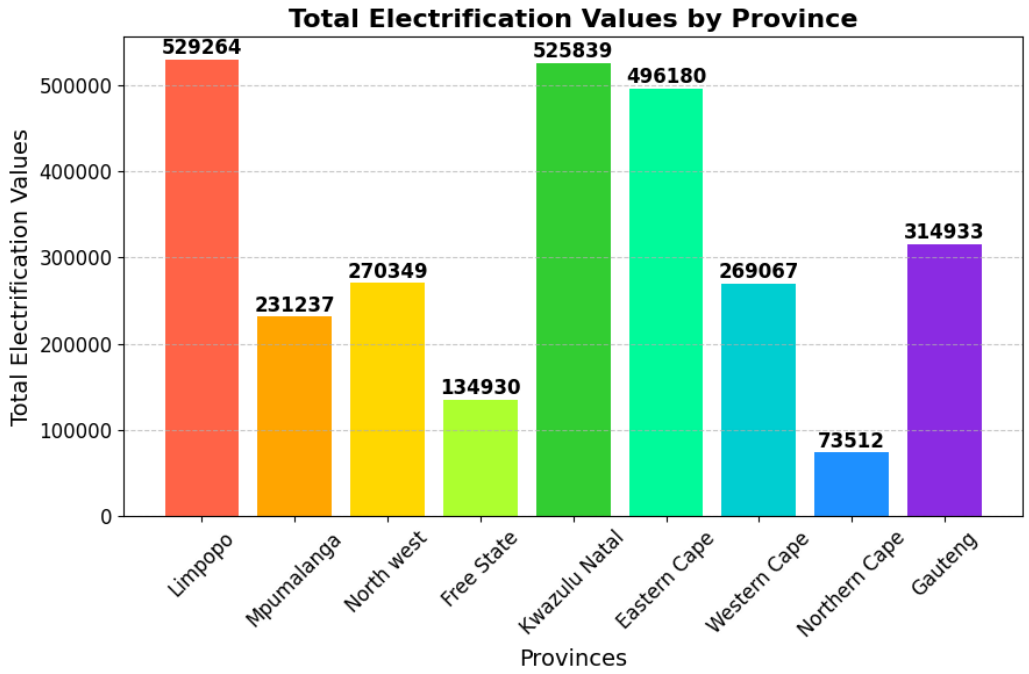
	Financial Year (1 April - 30 March)	Limpopo	Mpumalanga	North west	Free State	Kwazulu Natal	Eastern Cape	Western Cape	Northern Cape	Gauteng
0	2000/1	51860	28365	48429	21293	63413	49008	48429	6168	39660
1	2001/2	68121	26303	38685	20928	64123	45773	38685	10359	36024
2	2002/3	49881	11976	28532	10316	63078	55748	28532	6869	32127
3	2003/4	42034	33515	34027	16135	60282	47414	34027	10976	39488
4	2004/5	54646	16218	21450	5668	37811	42041	21450	6316	18422

This section dives into the analysis of electrification levels across South African provinces. Using Python, we calculate a comprehensive "Total Electrification" metric for each province by summing the relevant values in our dataset (ebp_df). This metric incorporates various factors contributing to electrification progress. To visualize these provincial electrification scores, we create a heatmap. Heatmaps are ideal for representing data with geographical or regional context. Our heatmap utilizes the "cool/warm" colour scheme, where cooler tones represent lower electrification values and warmer tones

represent higher values. This allows for easy identification of provinces with the most and least progress in electrification.



This section complements the heatmap analysis by presenting the "Total Electrification" values for each province in a bar chart format, allowing for easy visualization of provincial differences in electrification progress.

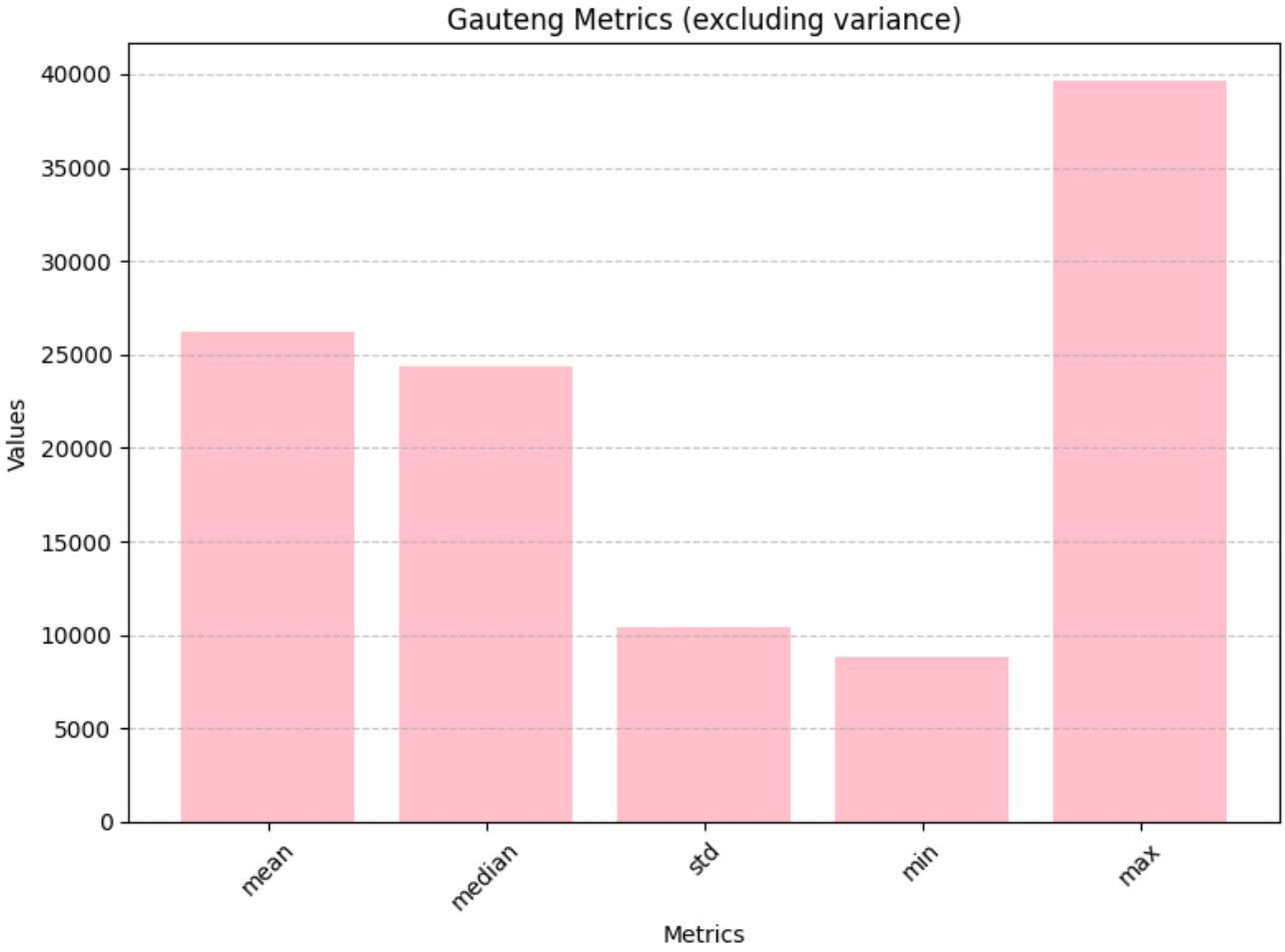


To gain a deeper understanding of electrification distribution within each province, we can calculate various statistical metrics. These metrics provide insights into the average electrification level, the spread of values around that average, and the range of electrification scores across different locations within the province.

One such approach is to use a function like dictionary_of_metrics that takes provincial data as input and returns a dictionary containing key metrics. Here's an example of the output for Gauteng province:

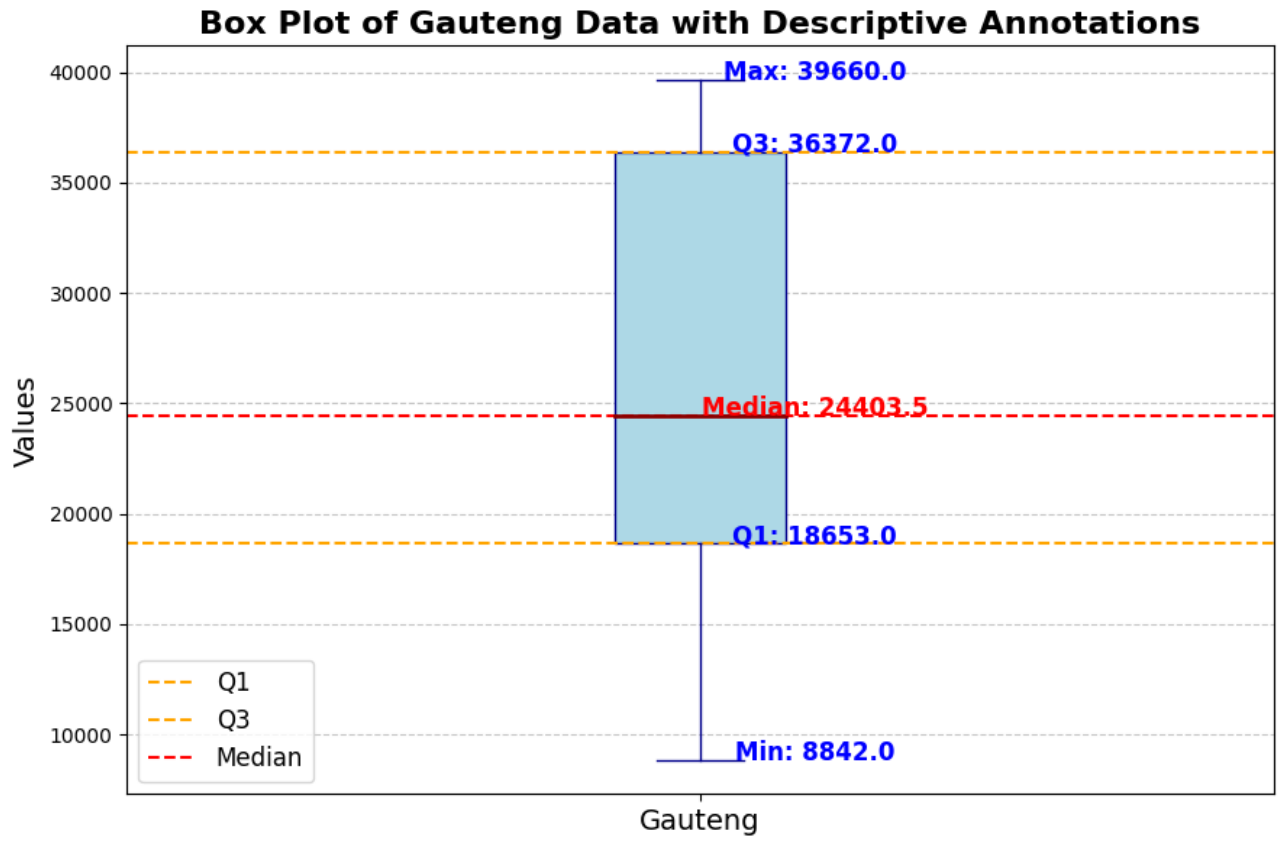
```
dictionary_of_metrics(gauteng) == {'mean': 26244.42,
                                   'median': 24403.5,
                                   'var': 108160153.17,
                                   'std': 10400.01,
                                   'min': 8842.0,
                                   'max': 39660.0}
```

Below is the bar chart representation of the above:



The next section looks at the 5 number summary.

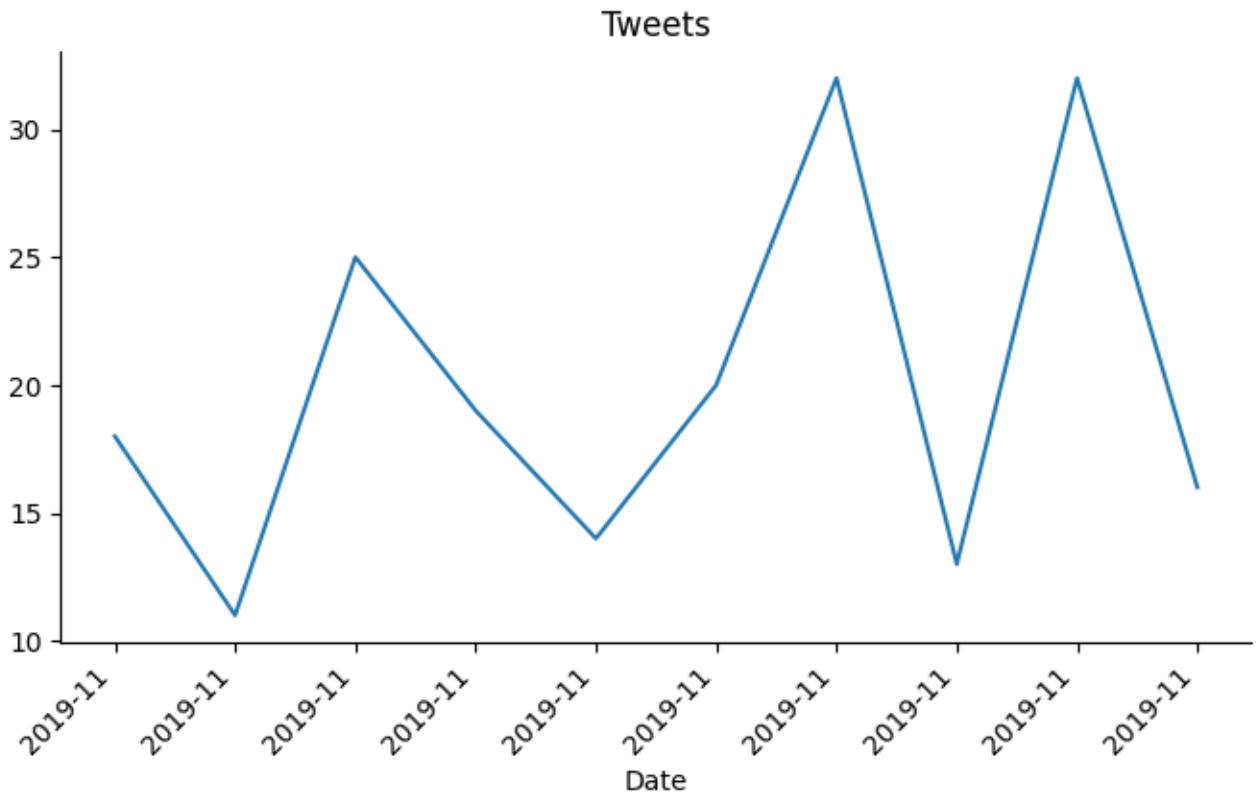
- **Minimum Value:** The smallest value in the data set.
- **First Quartile (Q1):** The value that separates the bottom 25% of the data from the top 75%.
- **Median:** The middle value when the data is arranged from lowest to highest. It represents the "centre" of your data.
- **Third Quartile (Q3):** The value that separates the top 25% of the data from the bottom 75%.
- **Maximum Value:** The largest value in the data set.



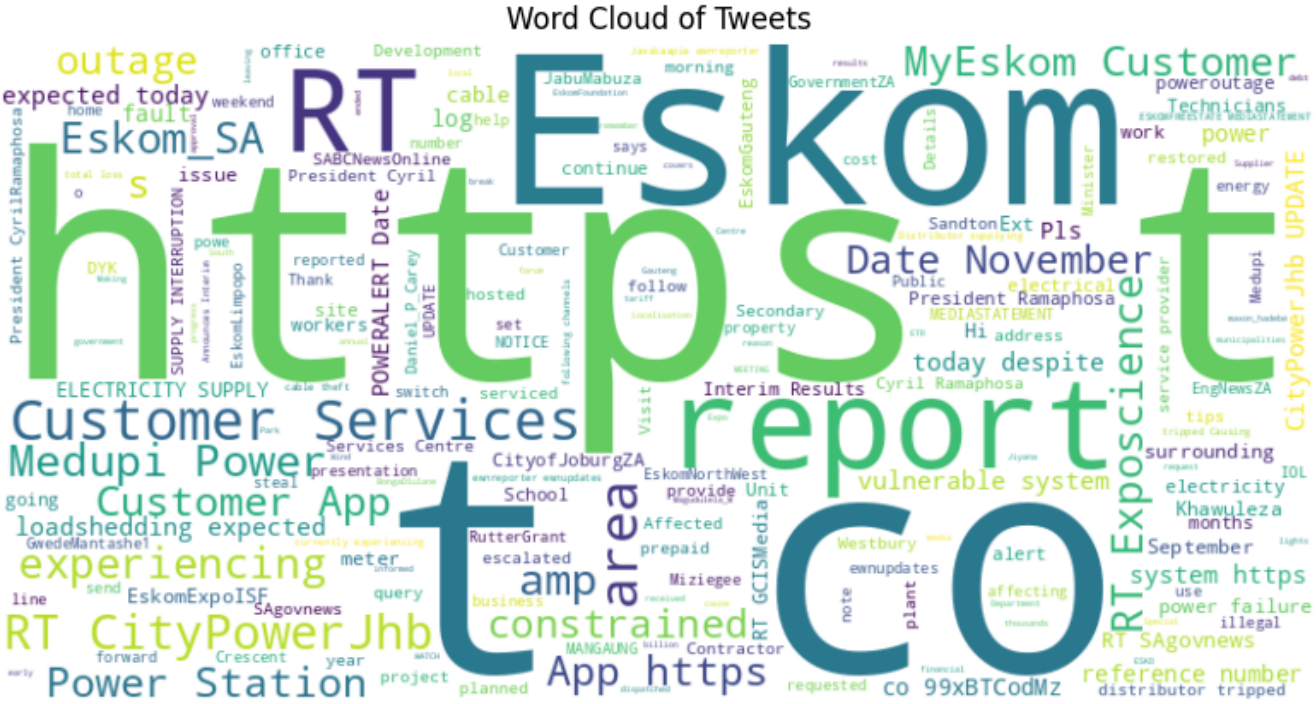
By looking at these five values, we understand that:

- **Spread:** How much the data varies from the centre (median). A large difference between the minimum and maximum or large gaps between quartiles indicate a wider spread.
- **Symmetry:** Whether the data is skewed towards higher or lower values. A median closer to one quartile than the other suggests a potential skew.
- **Outliers:** Potential outliers might exist if there are significant gaps between the minimum/maximum values and the rest of the data.

The line graph below represents the number of tweets sent a day.



Below is a word cloud of tweets.



CONCLUSION

This project explored the use of Python functions to calculate electrification metrics in South Africa. By leveraging Python's data analysis capabilities, we demonstrated how to calculate and analyse electrification metrics across South African provinces. We visualized these metrics using heatmaps and bar charts to gain a comprehensive understanding of the electrification landscape across the nation.

The findings from this analysis can inform data-driven decision-making for achieving equitable access to electricity in South Africa. By identifying areas with lower electrification rates, policymakers and stakeholders can prioritize resource allocation and target interventions to bridge the gap.

Furthermore, the Python code presented in this paper can be adapted and expanded to incorporate additional data sources and perform more complex analyses. This can contribute to ongoing efforts to develop a comprehensive strategy for achieving universal electrification in South Africa.

IN THE FUTURE

Future work in this domain could involve:

- Integrating geospatial data to analyse electrification patterns at a more granular level.
- Incorporating socioeconomic factors to understand the link between electrification and development.
- Developing machine learning models to predict future electricity demand and optimize resource allocation.