Data-Driven Identification of Optimal Urban Tree Planting Sites

K V Shanmukhasai (50596117), TRL Rajeswari (50593975)
Sreehari Rayannagari (50608557) , Tarun Gangadhar Vadaparthi (50592389)
University at Buffalo

***Abstract*—Urban tree plantation will play a major role in this era to compensate environmental challenges like pollution and global warming. At the same time it is equally important to identify the location of planting. This project focuses on identification of optimal place to plan trees in urban locality and primarily focuses within the City of Buffalo with the help of dataset provided by Bureau of Forestry. It includes processing of data analysis, exploratory data analysis in phase 1. This analysis identifies high-impact zones which are under green coverage and areas to improve air quality, noise reduction and overall environmental sustainability.**

***Index Terms*—Machine Learning, EDA, Data Intensive Computing,Data Cleaning**

I. Problem Statement

By verifying existing tree coverage, environmental advantages, and topographical features including stormwater drainage requirements, this project intends to determine the best places to plant new trees in urban areas. The aim is to give urban planners insights based on data collected so that they can select areas like flood control and air quality improvement that would help a large amount of people from tree planting activities

Background and its significance: Rapid urbanization,climate change, and the loss of green space are all factors to the environmental issues that urban regions are currently facing. Excessive stormwater runoff, poor quality of air and temperature rise made worse by the urban heat island effect is one of the most urgent problems. These issues are serious in crowded cities because the terrain is dominated by impermeable surfaces like asphalt and concrete, which traps heat and prevents water from absorbing. To preserve natural balance, green infrastructure,especially trees become increasingly important as cities expand

Why Trees?: An ample amount of ecological advantages are provided by trees, including the reduction of stormwater runoff and the strain on drainage systems, the air quality can be improved by the removal of pollutants and the reduction of urban temperatures through the shade. However, it is not possible to plant trees everywhere because cities constantly have limitations concerning existing areas, budgets, and resources. To prove that tree-planting measures are focused in the most effective locations. Hence, a more practical,evidence-based strategy is now required.

What does our project do? By giving a clear idea on spots where to plant trees that includes environmental aspects like stormwater drainage requirements, present tree coverage, and geographic data. By using this approach, urban planners may find the perfect locations for future tree planting i.e., locations where trees can most effectively reduce flooding, improve air quality and regulate temperature.Cities can step up their tree planting activities and also ensure that their limited resources are put to optimal use by making use of data.

Impact and Value: This project more than any other seeks to prioritize tree planting projects in a way which in turn provides a whole new perspective on urban sustainability and resiliency. What the project can do in assisting cities achieve multiple environmental goals is help in pointing high impact tree planting to where urban planners need to go. For example, strategically placed trees can be utilized as natural flood management systems in flood-prone areas thus reducing excess rainfall runoff and the chances of flooding occurring. It may also decrease the stress on municipal drainage systems, thus cutting down on maintenance costs and preventing flooding.
Likewise, the program can help identify areas in cities with poor air quality where trees would filter the greatest amount of pollutants, thus protecting the general population's health. This program may improve lung health by lowering dangerous airborne particles, particularly among children, the elderly, and individuals with asthma.

This project is valuable because it will help promote the value of urban spaces in the long run apart from its short-term returns. The project fosters the establishment of strong green infrastructure that will be planted in strategic areas that will help address future challenges on the environment.As an added benefit, the concept further encourages the economical utilization of resources such as the expenses and costs that may be incurred during tree planting campaigns ensuring that they remain resource friendly. Urban planners are therefore able to design parts of the city which are more robust, healthier and more ecologically sound.

II. Data Sources

This project is completely developed based on the dataset from the City of Buffalo's Bureau of Forestry. Data is col-

lected from inventory of street trees within the buffalo region. This dataset includes 28 distinct columns and almost 133230 distinct records and suitable to perform EDA techniques.This data[1] enabled to perform complete analysis on tree coverage and it's relation to the quality if life.

List of each Feature and their details:

- **Botanical Name:** The scientific name of the plants.
- **Common Name:** The commonly used name for the tree species.
- **DBH (Diameter at Breast Height):** Helps us to find tree size. The diameter of the tree's trunk measured at about 4.5 feet above the ground.
- **Total Yearly Eco Benefits ($):** Total economic benefits trees provided.
- **Stormwater Benefits ($):** Economic benefits trees helped in reducing stormwater runoff.
- **Stormwater Gallons Saved:** The volume of stormwater the tree helped in absorbing.
- **Greenhouse CO2 Benefits ($):** Tree's ability to absorb carbon dioxide.
- **CO2 Avoided (in lbs.):** CO2 emissions prevented by the tree.
- **CO2 Sequestered (in lbs.):** The amount of CO2 absorbed and stored by the tree.
- **Energy Benefits ($):** The economic benefits from the tree in reducing energy consumption.
- **kWh Saved:** Energy saved with the help of Trees.
- **Therms Saved:** Thermal energy saved with the help of Trees.
- **Air Quality Benefits ($):** The economic value of the tree's role in improving air quality.
- **Pollutants Saved (in lbs.):** The quantity of air pollutants removed by the tree.
- **Property Benefits ($):** Monetary benefits due to the location of the tree.
- **Leaf Surface Area (in sq. ft):** The total surface area of the tree leaves.
- **Address:** The street address where the tree is located.
- **Street:** The name of the street where the tree is located.
- **Side:** Position of the street where the tree is located.
- **Site:** Location of the tree.
- **Council District:** The council district where the tree is located.
- **Park Name:** Name of the park (if any) where the tree is located.
- **Latitude:** The latitude coordinate of the tree's location.
- **Longitude:** The longitude coordinate of the tree's location.
- **Site ID:** A unique address of the specific tree or location.
- **Location:** The geographic coordinates (latitude, longitude) combined as a point location.
- **location:** Geographic coordinates in WKT format.

III. Data Cleaning/Processing

In the data cleaning process, various steps were performed to make sure the dataset is ready for analysis:

Handling Missing Values: Data with null values should be handled since it highly impacts the performance of the model. Since the rows with null values are very few, all the values were dropped. After removing the missing values, the same was verified.

Removing Redundant Values: Duplicate values were identified and removed from the dataset. Removing this is necessary since it may impact model training because if any value is redundant, it can actually give different values altogether. We removed almost 25 redundant rows from the dataset. After removing redundant values, the same had been verified again.

Standardizing Categorical Data: Non-numerical columns were checked for consistency since learning depends on similarity. Here, inconsistent values for "Buffalo" were changed to a proper title format using string formatting, ensuring uniformity in the dataset.

Handling Data Type Issues: Inconsistent data types were addressed, mainly in the "Park Name" column, where '0' values were replaced with "Not a park" to standardize the entries. This is because a tree can be anywhere and not only in a park. So the dataset contains various locations that are not parks. The values were replaced with 0 in order to maintain consistency.

Outlier Removal: From the bar graph mentioned in the exploratory data analysis, we need to remove outliers, which are values that are too high or too low compared to the average set of values. To remove the outliers, the IQR method was used in the *DBH* and *Total Yearly Eco Benefits* columns. The new data points were again written to the data.

IV. Exploratory Data Analysis

EDA is an approach to do analysis on huge dataset which can provide meaningful insights and helps in boosting of the any product outcome. This is crucial step in data analysis process since it allows us to understand the data, find patterns, detect anomalies.

In this work, we have use various types of EDA techniques and will be discussed about each in the following section.

The key components of EDA include:

- **Descriptive Statistics:** Perform calculations such as mean,median,mode,variance, and standard deviation.
- **Data Visualization:** Techniques to plot graphs and find overview of the entire dataset and each will be discussed next.
- **Data Cleaning:** Handling of missing values and outliers. This is important since unstable data result in bad models.
- **Feature Engineering:** Additional of new features to improve model performance.
- **Correlation Analysis:** Finding the relationships with each feature to understand the data better.

In-Depth Analysis of Tree and Eco-Benefit Data Visualizations

Distribution of Tree Diameter (DBH)

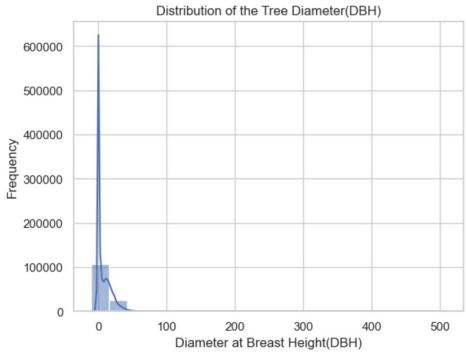


Fig. 1. Distribution of Tree Diameter (DBH)

In the above fig, we used kernel density estimation(KDE) histogram to understand the distribution of tree diameters at breast height. From this graph, it is observed that it's a highly skewed distribution saying that most trees have very small DBH values and only few with bigger values which may be outliers. Most of the trees fall within the 0 to 50 inches size range.

Distribution of Total Yearly Eco Benefits

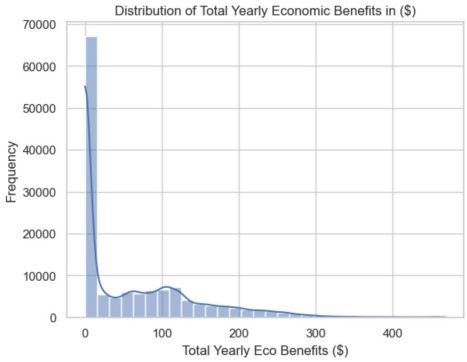


Fig. 2. Distribution of Total Yearly Eco Benefits

In the above fig, we again used kernel density estimation(KDE) histogram to understand the distribution of total years eco benefits. From this graph, like the above EBH graph, this is distribution is also right-skewed. From the graph, most of the trees give almost low eco-benefits but there are few with significantly high benefits.From the above both skewed visuals, there are few trees which have high impact in the contribution.

Relationship Between DBH and Total Yearly Eco Benefits In the below fig,We used a scatter plot to find the correlation between tree diameter at breast height and

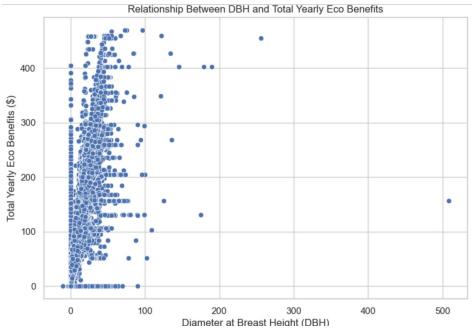


Fig. 3. Relationship Between DBH and Total Yearly Eco Benefits

total yearly eco-benefits. Since it is important to understand relation between ecological benefits and the height of the tree. This plot shows good positive relation between both and even though we have few outliers but it doesn't effect much. This outliers may be due to external factors such as health conditions or environmental conditions. So from the above graphs, we can conclude that most of the trees in the dataset are small and it has less contribution in the ecological benefits.

Pair Plot of Selected Variables

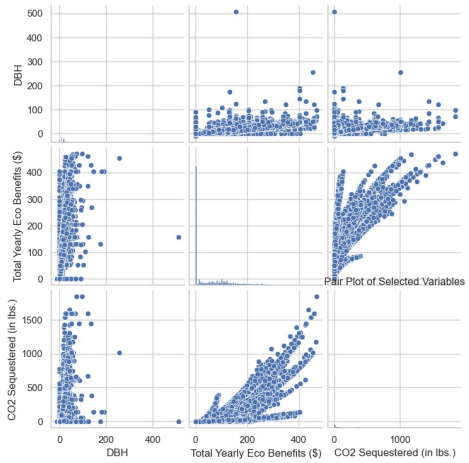


Fig. 4. Relationship Between DBH and Total Yearly Eco Benefits

We chose pair plot visualization because it effectively shows the relationships between multiple variables. This type of plot is great for spotting trends,correlations and outliers all in one view. In this case, it highlighted the right-skewed distribution of tree diameter (DBH), eco benefits, and CO2 sequestration, showing that most trees are smaller and offer fewer benefits.The scatterplots made it clear that larger trees generally provide greater economic and environmental value, as seen in the positive correlations btw DBH and both eco benefits and CO2. It also revealed that trees offer high benefits are also better at capturing CO2, emphasizing their role in environmental conservation. Additionally, the plot helped identify some outliers, such as trees with unusually large diameters, which contribute significantly to ecological benefits.

Relationship Between DBH and Total Yearly Eco Benefits

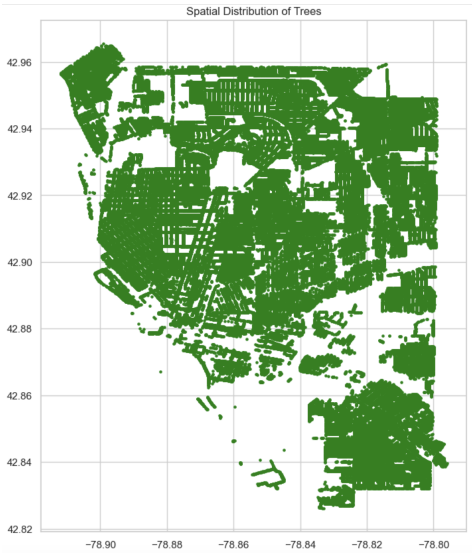


Fig. 5. Relationship Between DBH and Total Yearly Eco Benefits

In this project, it is important to understand the trees distribution with in the region and here GeoDataFrame technique is leveraged to find the high density areas and sparse areas. From the latitude and longitude points this graph provides broad idea of the tree areas. From this graph, next will try to establish the relation between the sparse areas and environmental conditions.This will also help city planners and data team which require about classification of dense and parse areas.

Box Plot for DBH to detect outliers

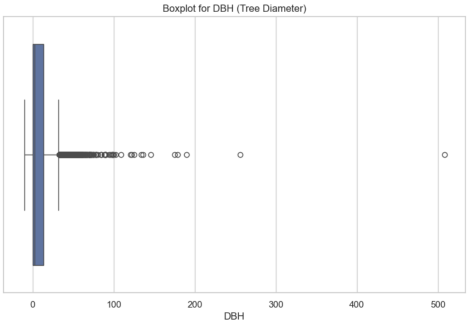


Fig. 6. Box Plot for DBH to detect outliers

For any dataset, the next crucial step is to identify the outliers since it have high impact on the model performance. So in this graph, we did analysis on DBH using box plot. This gave detailed insights of the tree diameter distribution and coming to the graph, box plot provided outliers ranging from 50 to 100 and there are few between 100 to 200 and very few after 200. Most of the diameters lie within 30 showing the consistency of the data.

Boxplot for Total Yearly Eco Benefits to detect outliers

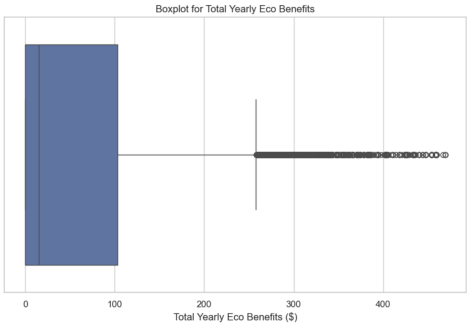


Fig. 7. Boxplot for Total Yearly Eco Benefits

This box plot give insights about the total eco-benefits and their distribution of the data. This step is again for the enhancement of the model. It has many outliers on the upper end showing that some trees have more ecological benefits. This is compared with the median since includes number of trees. The majority of the trees are in 0 to 50 dollars.Now in the data cleaning, will remove the outliers from the data for better model performance. More details are discussed in the Data Cleaning section.

Correlation Matrix

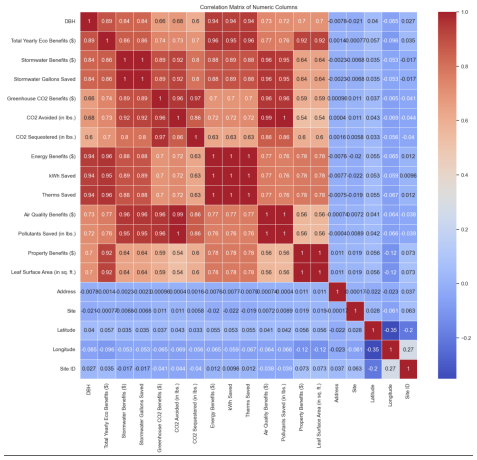


Fig. 8. Correlation Matrix

Heatmap shows the correlation between the numeric variables in the dataset. This heatmap provides various major insights about the dataset. DBH have a great impact with eco-benefits, storm water benefits, greenhouse CO2 benefits, and energy benefits. So from the spatial observation and this observation we can conclude that larger trees contribute more to the environment. Now energy benefits have a great impact with the trees and eco benefits. Latitude and Longitude have no impact and correlation is Zero since it doesn't have any relation with the properties of the tree.

Total Yearly Eco Benefits by Council District

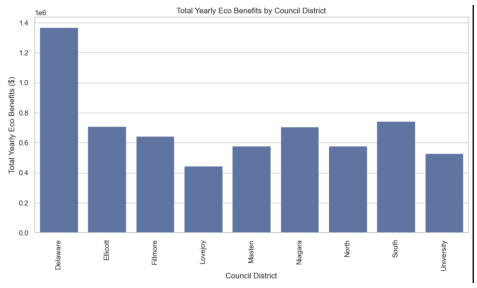


Fig. 9. Total Yearly Eco Benefits by Council District

Bar plot provides insights about the yearly eco-benefits by council district and Delaware District contributes more in eco-benefits compared to other districts and the reason might be due to larger tree cover, more mature trees which we would like to figure out after this step. other districts like Niagara,South and University provide moderate benefits and Lovejoy being at the least.

Storm water Gallons Saved by Council District

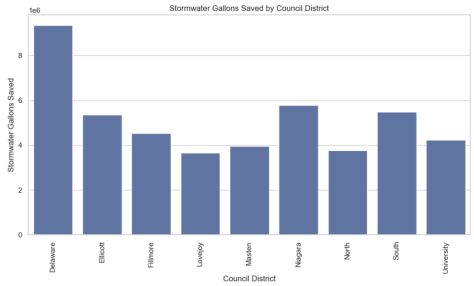


Fig. 10. Storm water Gallons Saved by Council District

Storm water Gallons Saved by Council District This plot highlights storm water savings across council districts. We see Delaware is top in storm water savings, which may be due to presence of large trees that intercept more rainfall. Between districts there isn't large variation compared to other benefits, saying storm water savings could be more uniform across different types of trees.

CO2 Sequestered by Council District

This plot shows CO2 sequestered by trees in different council districts. There is a similar pattern of distribution to the eco-benefits, with Delaware district leaning in CO2 sequestration. This can be because of big, old trees or more tree coverage in this area. Also there is variation between districts which suggests some areas have more potential for C02 sequestration if we increase tree planting programs.

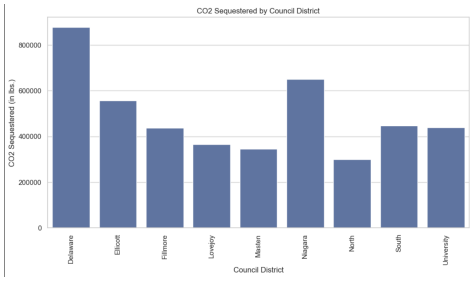


Fig. 11. CO2 Sequestered by Council District

V. Phase - II

Google drive Integration:

The project integrates with Google Drive to access the dataset and other required files. Mount the drive, navigate through its directories to the appropriate folder containing the dataset, run commands to check the current directory and list its contents to ensure the files are available. The dataset, `Tree_Inventory_chngd.csv`, is located in the specified directory and loaded through PySpark. A `SparkSession` has been created with a custom application name, *UrbanTreePlanting*, for dealing with large-scale data. The schema of the dataset is printed for structural verification in terms of column names and data types. Columns in the dataset are listed for a quick overview on the available variables.

Data Exploration

Descriptive statistics are generated for selected columns including important features like *DBH*, *Latitude*, *Longitude*, and Total Yearly Eco Benefits ($). This gives some insights into mean, standard deviation, minimum, and maximum values. Missing or null values for important columns are filtered in order to know the extent of data missing, and the columns that need imputation. Statistical summary has been made for some selected columns, which ends up with means, standard deviations, minimums, and maximums. The missing or null values of critical columns can be filtered in order to know the amount of data the column is missing, as well as which would need imputation.

Data Cleaning

Null and invalid values in numerical columns should be substituted with their mean values. This applies to the dataset, which is now ready for analysis because it is not biased. Re-computations of descriptive statistics were subsequently conducted to ascertain the changes made and validate that no null entries exist in the dataset.

Feature Vectorization and Standardization

Relevant columns can now be merged into a single feature vector using PySpark's `VectorAssembler`. This step is required because all input features should be collected into one-column vector before further processing; thus, this feature vector is put through PySpark's `StandardScaler` for

standardization. Such standardization ensures that features will be uniformly scaled, and hence, perform effectively as part of the model.

VI. K-Means Clustering Analysis (Algorithm 1)

Objective

K-Means clustering is designed to generate clusters from the datasets depending on feature similarity; thus, it mainly serves to discover patterns or groupings in the data.

Implementation

Cluster the data using the scaled dataset so that ranges across the features would be the same. The Elbow Method is performed in order to find the optimal number of clusters (k). The sum of squared distances of the points to their cluster centroid is calculated for $k = 2 : 9$. The elbow point is the point where the cost decreases slowly, and then it can be taken as the optimal k. The value of k is **5**. The clustering model creates prediction for allocation of each data point into clusters.

Output

Cluster assignments (prediction) are appended to the running dataset. Example columns: `DBH`, `Latitude`, and `Longitude`, plus the last column as `prediction`.

Evaluation

Silhouette Score Determine the quality of a clustering scheme with the silhouette coefficient, which indicates how similar data points are within the same cluster compared to those in different clusters. High silhouette values indicate clearer clusters. 0.670 is the value achieved with the current model. *Davies-Bouldin Index* Computes the average of intracluster distance ratios, in comparison with certain random intercluster distances, where lower values would indicate superior scoring on this measure. 0.95 is the value achieved for the K-means Algorithm. *Calinski-Harabasz Index* Measures the ratio of cluster dispersion to that of intra-cluster dispersion. Higher values indicate better-definition clusters. 105868 is the value achieved for the K-means Algorithm.

Visualization

This visual representation will help derive the value by showing the elbow point on the cost versus the number of clusters. *Geographical Visualization:* Displayed on the Folium map, where clusters are shown in different colors to visualize values spatially.

VII. Regression Analysis (Algorithm 2)

Objective

Using linear regression to estimate Total Annual Eco Benefits ($) based on the parameters such as, but not limited to, DBH, Stormwater Benefits ($) and Greenhouse CO2 Benefits ($).

Implementation

Features are assembled into a single vector and then aggregated into a vector. Standardization was applied to all of the features. The dataset was then divided into train (80%) and test (20%). The linear regression model is trained on the training portion to establish the relationship between features and the dependent variable (Total Yearly Eco Benefits ($)).

Output

The model generates predictions for the test set.

Evaluation

R^2 *Score:* Actually measures the proportion of variance in the dependent variable explained by the model. A higher value of R-square indicates a better fit. Mean Absolute Error (MAE): it estimates the average absolute difference between predicted and actual values. Lower the MAE, the better the performance of the model.We have achieved 0.73 R^2 score for the model and MAE is 23.1829 shows is the not the best model for proceed further.

VIII. Decision Tree Classification (Algorithm 3)

Objective

The decision tree classifier is employed for predicting categories of Total Yearly Eco Benefits in dollar amounts ($) as denoted by the following levels: `High`, `Medium`, and `Low`.

Implementation

It creates a categorical label as follows: High: ≥ 100 High, Medium: 50–100 Medium, Low: ¡50 Low. Vectorization and standardization of features are done. Dataset splitting is done for training and testing sets. A decision tree model has been trained on the training set.

Output

It provides predictions for the test set that can be summed up as classifying records in either of the following categories: `High`, `Medium`, or `Low`.

Evaluation

`Accuracy:` This imprints the indication of a true prediction. `Precision:` It is an indicator that measures the number of total true positives with regard to total predicted positives. `Recall:` It measures how many true positives there are compared to all positive cases actual. `F1-Score:` A composite method of measurement regarding precision and recall that reconciles their performances in one value.All the values are around 0.91 which shows the model is good. However, for the given problem, we need to find the similarity rather than taking decision.

IX. Aggregated Eco Benefits (MapReduce - Algorithm 4)

Objective

Incorporate MapReduce paradigm to finally aggregate the total yearly eco benefits with respect to each council district.

Implementation

Before processing, null entries in the Council District column are suppressed. The working method is a PySpark RDD (Resilient Distributed Dataset): Such as `Map Phase:` transforms each record into key-value pairs, where the key is the council district and the value is Total Yearly Eco Benefits ($). `Reduce Phase:`, thus, aggregates values (sums) for each district. Final results are collected and shared.

Output

Presents the aggregate total eco-benefits for each council district which gives an insight into how much these districts contribute at the district level.

X. Geographical Analysis (Algorithm 5)

Objective

Geographical analysis and visualization of the clusters would be mainly through the clustering results of K-Means.Since the data is not labeled and there is no true and false data in the dataset. So K-Means clustering algorithm best fits with the requirement and the cluster data is further used to build the recommendation of plantations.

Implementation

The cluster assignments are mapped geographically in Folium using unique colors on the map to represent each cluster. Calculated and displayed are the cluster centroids, which signify the center location of each group.

Evaluation

Centroid Analysis: Average values of some features such as *DBH*, Stormwater Benefits ($), etc., for each cluster are calculated, leading to an understanding of the characteristics of each cluster. `Map Quality:` The clusters on the map can be visually inspected for an evaluation of the spatial distribution of clusters and their effectiveness.

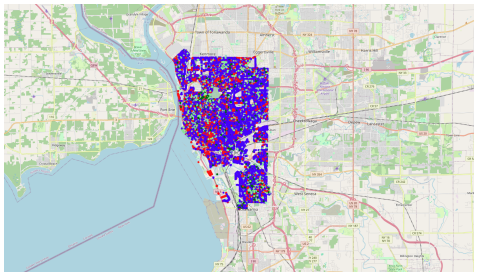


Fig. 12. Cluster Map of the given dataset

XI. Algorithmic Observations

Algorithm diversity ensures that clustering, regression, classification, and MapReduce are used for all-round analysis. *Evaluation metrics:* All the algorithms consider into the analysis as they work according to the relevant performance metrics so that the results become meaningful and actionable. *Visualization*: Using tools such as Folium maps and plots involves Transformational visualization when it comes to understanding how things pattern spatially and cluster-wise.

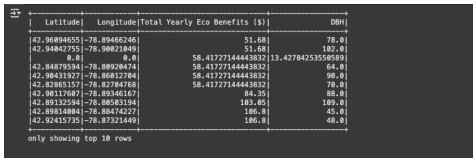
XII. Recommendation of Priority Areas

Filtering Priority Clusters

Output: Priority clusters (Cluster 1 and Cluster 4) correspond to the areas relatively lower in ecological benefits compared to the other clusters.

Details: The dataset is restricted to the priority clusters. The areas within the clusters would then be ranked according to their *Total Yearly Eco Benefits ($)* in ascending order to highlight the least eco-beneficial locations for targeted tree planting.

Analysis: Filtering ensures that the efforts go to areas which have the highest need for ecological improvement. Ranking clearly establishes prioritization direction so that decision-makers can focus resources appropriately.



Latitude	Longitude	Total Yearly Eco Benefits ($)	DBH
42.96094655	-78.89466246	51.68	78.0
42.94042755	-78.90021049	51.68	102.0
0.0	0.0	58.417271444443832	13.427842533508589
42.84879594	-78.88928474	58.417271444443832	64.0
42.90431927	-78.86012704	58.417271444443832	90.0
42.82865157	-78.82704768	58.417271444443832	70.0
42.90117607	-78.89346167	84.35	88.0
42.89132594	-78.88583194	103.05	109.0
42.89814004	-78.88474227	106.8	45.0
42.92415735	-78.87321449	106.8	48.0

only showing top 10 rows

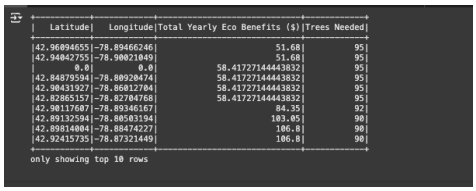
Fig. 13. Top 10 priority locations

Calculating Trees Needed

Output: The calculation is done to establish how many additional trees are needed in each priority cluster area to reach the target level of eco-benefit amounting to $1,000.

Details: Assuming an average annual benefit of $10tree, the difference between the targeted eco-benefit and the present benefit is divided by the average annual benefit per tree. This gives us the number of trees required. Removed from the final recommendation are areas which require no additional trees.

Analysis: This process closes the gap between current ecological states and target conditions by quantifying intervention needs in each area and providing actionable measures to urban planners.



Latitude	Longitude	Total Yearly Eco Benefits ($)	Trees Needed
42.96094655	-78.89466246	51.68	95
42.94042755	-78.90021049	51.68	95
0.0	0.0	58.417271444443832	95
42.84879594	-78.88928474	58.417271444443832	95
42.90431927	-78.86012704	58.417271444443832	95
42.82865157	-78.82704768	58.417271444443832	95
42.90117607	-78.89346167	84.35	92
42.89132594	-78.88583194	103.05	90
42.89814004	-78.88474227	106.8	90
42.92415735	-78.87321449	106.8	90

only showing top 10 rows

Fig. 14. Trees needed at each location

XIII. Geospatial Visualization

Mapping Recommended Areas

Output: A geospatial map (`tree_planting_map.html`) is constructed utilizing Folium for visualization of recommended areas.

Details: Indicate the following locations by marker: Latitude, longitude, current eco-benefits, and how many more trees are needed. Each marker has a pop-up for details. It centers

around the average location of all recommendation sites to provide better context.

Analysis: An interactive map that enables everyone to make a decision based on spatial realities. At a glance, urban planners can see the clusters of priority areas and can come up with area-based strategies.

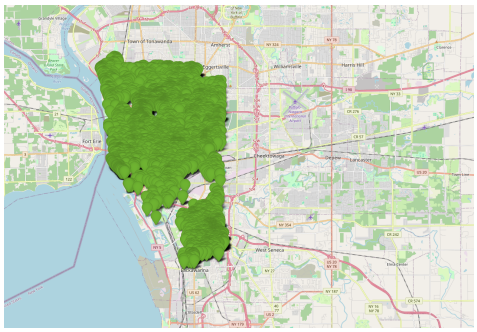


Fig. 15. Geospatial Representation

Heatmap of Clusters

Output: This heatmap (`cluster_heatmap.html`) depicts the concentration of data points in clusters.

Details: It illuminates the patterns in space such as high density zones for even better prioritization.

Analysis: The heatmap pinpoints other concentration areas that poorly score on eco-benefits-adding another dimension to discern the spatialities of ecological conditions.

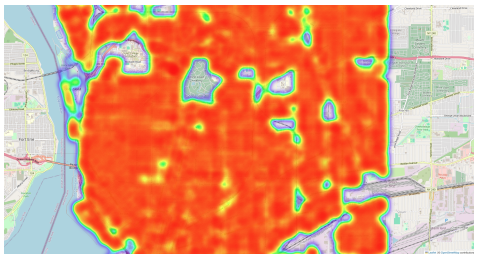


Fig. 16. Heatmap Representation

Aggregate Benefits by Cluster

*Output:*The bar chart shows Total Yearly Eco Benefits ($) average for each cluster.

Details: The given chart elaborates on the differences in welfare among clusters in terms of average values, focusing more on those priority clusters that scored lower on average. The difference between clusters is given here quantitatively; it again provides one of the reasons for the priority given to Cluster 1 and Cluster 4.

Analysis: This visualization adds numerical values towards the differences between clusters to support the prioritization of the two: Cluster 1 and Cluster 4.

XIV. Location-Specific Recommendations

Adding Location Names

Output: Geocoding is a process involving association between latitude and longitude with location names that are easy to read. The best 10 areas are indicated for their location, eco-benefits, and, tree needed.

Details: All null points such as zeros or empty coordinates are filtered out of the final recommendations. Valid places are prioritized based on variables like the number of extra trees required.

	Location
0	188, Gorton Street, Buffalo, Erie County, New ...
3	78, Masten Avenue, Fruit Belt, Buffalo, Erie C...
4	Cazenovia Parkway, Buffalo, Erie County, New Y...
5	South Park 9 Hole Golf Course, Park Drive, Buf...
1	56, Dunston Avenue, Buffalo, Erie County, New ...
6	618, Columbus Parkway, Buffalo, Erie County, N...
12	Cazenovia Parkway, Buffalo, Erie County, New Y...
15	Cazenovia Parkway, Buffalo, Erie County, New Y...
14	Scajaquada Expressway, Buffalo, Erie County, N...

Fig. 17. Top 10 priority locations

Mapping Top Locations

Output: An interactive map topically outlining the top 10 tree planting locations contains further informative pop-ups at each of the indicated places. It is viewable at top_10_tree_planting_map.html.

Details: The mean location of the top 10 areas is what the map emphasizes. Each marker informs about its location name, economic benefits, and tree requirements

Analysis: The mean position of the top ten areas is placed in the map. Each point supplies information on the location name along with its eco-benefits and tree demand while geocoding increases interpretability in recommendations, thus delivering an actionable recommendation applicable at ground level. This final map is able to zoom in on a person's key areas, hence allowing for more targeted action plans.

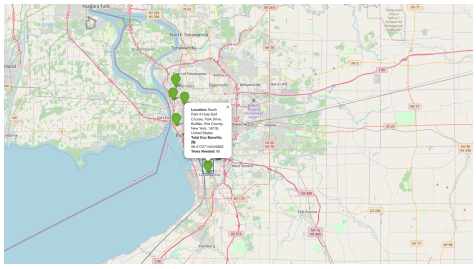


Fig. 18. Final top 10 priority locations

XV. Overall Observations and Strengths

Actionable Outputs: The code duly discerns and quantifies the interventions as it requires to improve the ecology. *Geo-Spatial Integration:* Using geo-spatial tools such as Folium and heatmaps, it renders practical visualization of how eco-benefits are distributed in space. *Data-Driven decision making:* It bases its calculations and visualizations on extensive data processing in terms of reliability and exactitude.

XVI. Limitations

The analysis assumes an average per tree ecological benefit of $10, which can vary by species or by environmental conditions. Geocoding relies on external libraries, which may not resolve all coordinates correctly.

XVII. Conclusion

The analysis offered and the visualizations employed comprehensively meet the objective of urban prioritization for tree planting based on ecological gains. An invaluable mix of clustering, ranking, geospatial visualization, and quantitative measures has been used in the code to develop actionable recommendations for enhancing urban green coverage. This structured approach emphatically secures resource allocation aligned with the sustainable urban health goals. A summary of the grist yet to come will be given as follows:

Clustering Analysis: K-Means Clustering Analysis: Low ecological benefits with priority areas were identified by using K-Means clustering. The reliability of the clustering results was warranted through the evaluation metrics of Silhouette Score and Davies-Bouldin Index. The spatial visualization of clusters from actionable insights on how they are located in relation to different ecological conditions.

Regression and Prediction: The relationship between tree-associated attributes and the total aggregate annual eco-benefits was well captured by linear regression. The predictive performance of the model was validated with several evaluation metrics like R^2 and MAE. This is an analysis that conveys a measurable understanding of how varying factors influence ecological benefits.

Classification for Decision-Making: The areas which were classified into High, Medium, and Low eco-benefit zones through decision tree classification enabled interventions specific to the areas of concern. The measurement of the precision, recall, and F1 score testified to the strength of the classification model.

MapReduce for Aggregated Insights: The MapReduce implementation aggregates eco-benefits. This highlighted the regional disparity of ecological condition. This macro perspective gives a balance to the micro-level analysis of clustering and ranking. The ecobenefits were mapped to council districts under MapReduce, pinpointing thereby the discrepancies between regions, on conditions that are ecological. That gave a macro view, which complements that of a micro-level analysis at clustering and ranking.

Prioritized Recommendations: Final ranking of the areas in terms of eco-benefits and number of trees needed was the evident guidance toward ecological improvement. Geospatial maps, interactive heat maps, and ranked recommendations offered intuitive yet accessible top-down perspective insight into data-driven findings.

Visualization and Spatial Analysis: The heatmaps, charts and maps in Folium-theirs were very instrumental in displaying the findings. Additionally humans could sort the readable location names into the very practical component for implementation in the real world.

XVIII. Remarks and Future Scope

This project showcases a data-driven approach to dealing with serious urban ecological challenges. The application of advanced analytical and visualization techniques offers a scalable and replicable approach for urban planners worldwide. With strong algorithms, accurate calibration, and insightful visualization, the user is ensured that the solutions offered are not only impactful but also actionable. This project will be instrumental in future cities where the population continues to grow and will face increasing environmental pressures in ensuring that they are sustainable, greener cities.This project can be extended further with recommendation of right tree species based on the climatic conditions,plant oxygen levels and other parameters.

Following the data preparation and EDA phases, the refined data is now ready to use the knowledge acquired to define certain areas to be targeted for tree planting. The results received from these preliminary phases will be the basis for moving forward as decisions will be data and analyses driven.

XIX. References

1) *Tree Inventory Data*, https://data.buffalony.gov/Quality-of-Life/Tree-Inventory/n4ni-uuec/about_data
2) Rhyne, Theresa Marie and MacEachren, Alan, "Visualizing geospatial data," ACM SIGGRAPH 2004 Course Notes, Los Angeles, CA, 2004, pp. 31–es, doi: 10.1145/1103900.1103931.
3) Stančin, I. and Jović, A., "An overview and comparison of free Python libraries for data mining and big data analysis," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 977-982, doi: 10.23919/MIPRO.2019.8757088.
4) Zaharia, M., et al., "Spark: Cluster Computing with Working Sets." *HotCloud*, 2010.
5) Kraska, T., et al., "MLbase: A Distributed Machine-learning System," *CIDR*, 2013.
6) Dean, J., & Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM*, 51(1), 2008.
7) Folium Documentation, "Folium: Python Data, Leaflet.js Maps," available at: https://python-visualization.github.io/folium/.
8) Hunter, J. D., "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, 9(3), 90–95, 2007.