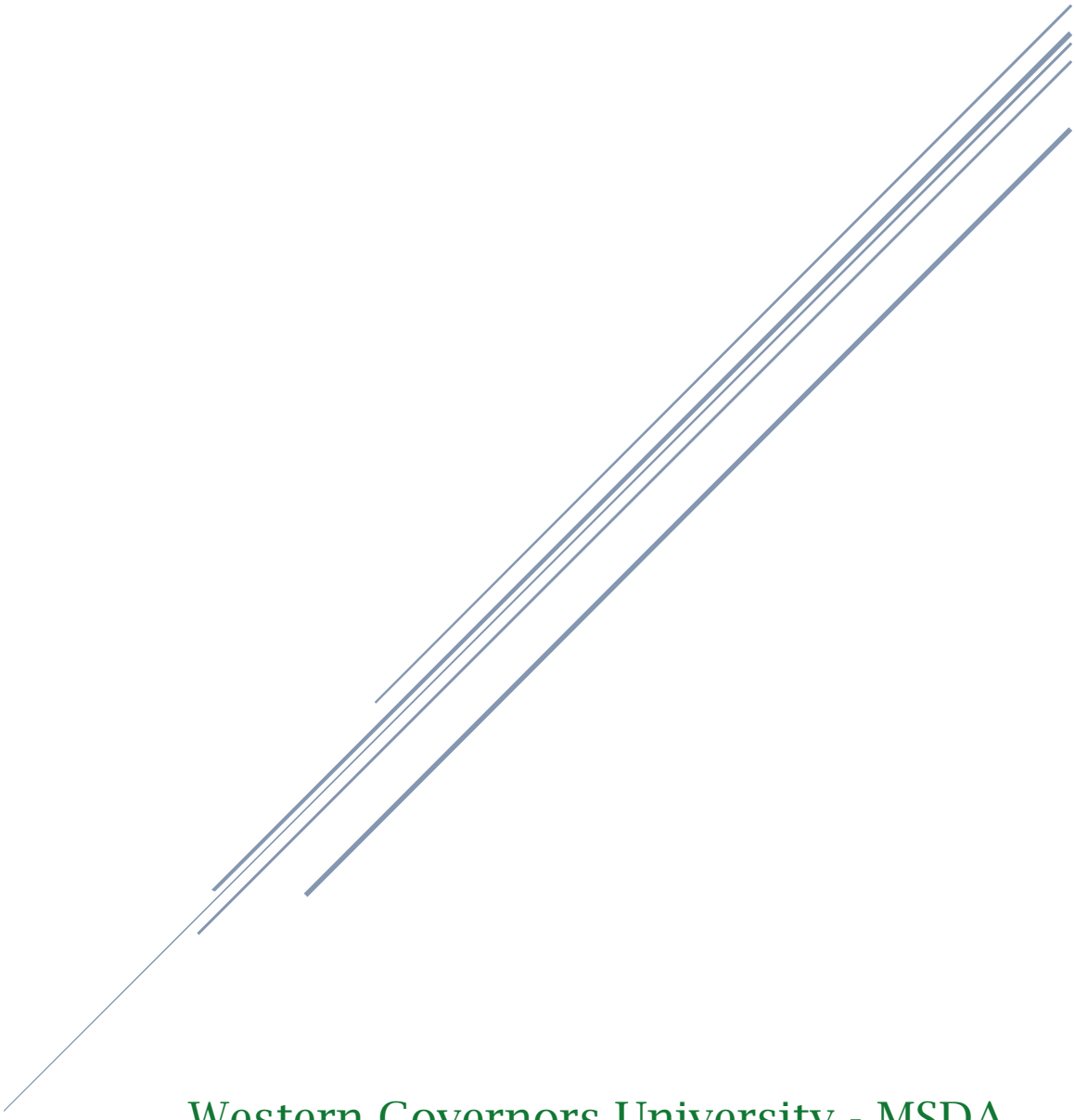
CAPSTONE – EXECUTIVE SUMMARY

Studen: Bryan M. Bravo | ID: 001362039



Western Governors University - MSDA
05/13/24

Executive Summary

Research Question

Can ANOVA with Tukey HSD testing and K-means clustering be used to identify the optimal locations for new E-bike stations in Portland, Oregon?

Problem Statement

According to the 2023 LMR BIKETOWN One-Pager report, there is an increase in the popularity of BIKETOWN, a bike-sharing service in Portland, Oregon, with 30,000 new riders in 2022. This increase in demand has made it necessary to construct additional E-bike stations, expanding the areas where people can rent and ride bikes. K-means Clustering analysis is a method that can be used to identify potential locations for new E-bike stations by grouping observations based on their geolocation coordinates (George, 2023). Additionally, a combination of ANOVA and Tukey HSD testing can determine the most statistically significant locations where the average distance traveled in miles is the response variable, and each cluster from the cluster analysis represents a sample group for ANOVA testing (Singh, 2024). If ANOVA testing reveals a statistically significant difference between sample groups, then Tukey HSD testing can help identify the distinct group samples that are statistically different from the others (Bobbitt, 2019). The clusters found to be the most significant through Tukey HSD testing would be considered optimal for installing new E-bike stations at or near identified cluster centers.

The Null and Alternative Hypothesis

H_0: The mean distance traveled within the generated clusters does not exhibit statistically significant differences.

$$H_0: \mu_{C\_1} = \cdots = \mu_{C\_k}$$

H_1: The mean distance traveled within the generated clusters exhibits statistically significant differences in at least one cluster group.

$$H_1: \mu_{C\_1} \neq \cdots \neq \mu_{C\_k}$$

The Significance level is set to: $\alpha = 0.05$ (5%). (Bobbitt, 2022)

Summary of the Data-analysis Process

The analysis presented here uses publicly available data from biketownpdx.com, obtained from the system data webpage (BIKETOWN PDX, 2020). The data spans from July 2016 through August 2020. The analysis process involved several critical phases, including data importation, combining similar columns, identifying and removing irrelevant null values, dropping observations outside the City of Portland, excluding trips with distances greater than 25 miles, identifying established E-bike stations and their coordinates, and filtering out observations that are too close to existing E-bike stations.

After cleaning and preprocessing, the dataset includes 7,176 E-bike rental observations and 235 established E-bike stations. The variables used are 'routeid', 'starthub', 'startdate', 'start_latitude', 'start_longitude', and 'distance_miles'. A new variable, 'cluster_labels', was created later through K-means clustering.

Figure 1 displays a map of Portland, where the blue circles with a bike icon represent existing E-bike stations. The heatmap shows the density of E-bike rental observations (Plotting with Folium, n.d.; Valkenburg, n.d.).

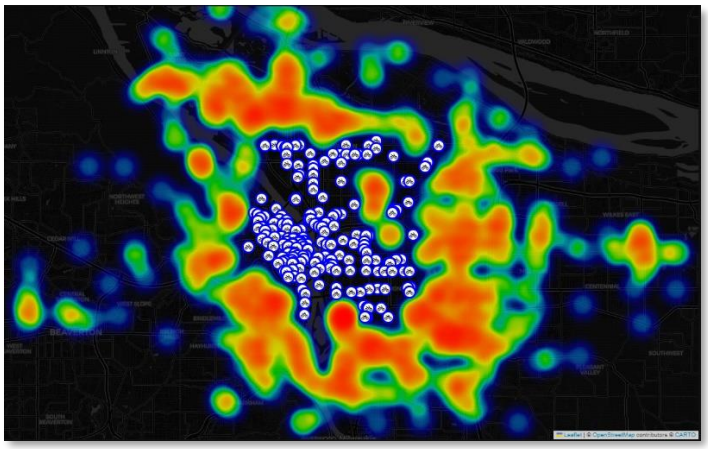


Figure 1 – City of Portland

Figure 2 shows 14 potential locations as cluster centroids. After performing K-means Clustering, the variable 'cluster_label' was created to represent each cluster as a categorical value (George, 2023). An elbow plot was used to identify the optimal K-means clustering solution as 14 clusters (Daityari, n.d.), as shown in Figure 3. The Silhouette Score of **0.609** confirmed a well-separated clustering solution with minimal overlap (Pulkit, 2024).

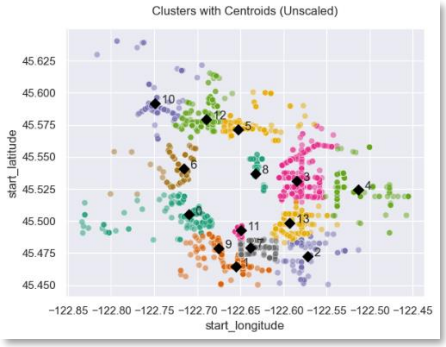


Figure 2 - Clusters with Centroids

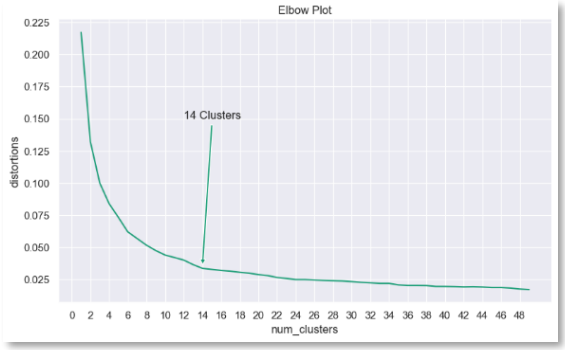


Figure 3 – Elbow Plot

Finally, an ANOVA and Tukey HSD test was conducted to compare each group's average distance traveled in miles, defined by the 'cluster_labels' variable (Singh, 2024). The Hedge's g statistic, provided by Pingouin's 'pairwise_tukey()' method, was utilized in the Tukey HSD test pairwise comparisons to identify the most statistically significant group pairs and optimal clusters.

Summary of Analysis Results

K-Means Clustering Results

K-means clustering was used to group observations into 14 clusters based on their proximity in the City of Portland. To assess the quality of the clustering solution, a silhouette score of **0.609** was computed, indicating well-separated clusters with minimal overlap (Pulkit, 2024). The identified clusters were distributed across different parts of Portland: **Southwest** Portland (clusters 0 and 9), **Southeast** Portland (clusters 1, 2, 7, 11, and 13), **Northwest** Portland (clusters 5, 6, 10, and 12), and **Northeast** Portland (clusters 3, 4, and 8). Each cluster will be treated as a group sample for later ANOVA testing, which will allow for comparing the statistical means of the 'distance_miles' variable between the clusters.

Figure 4 displays the E-bike station locations on a heatmap of the density of observations in Portland and the cluster centroids. The cluster centroids, shown as gray circles, represent proposed locations for new E-bike stations. (Plotting with Folium, n.d.; Valkenburg, n.d.)

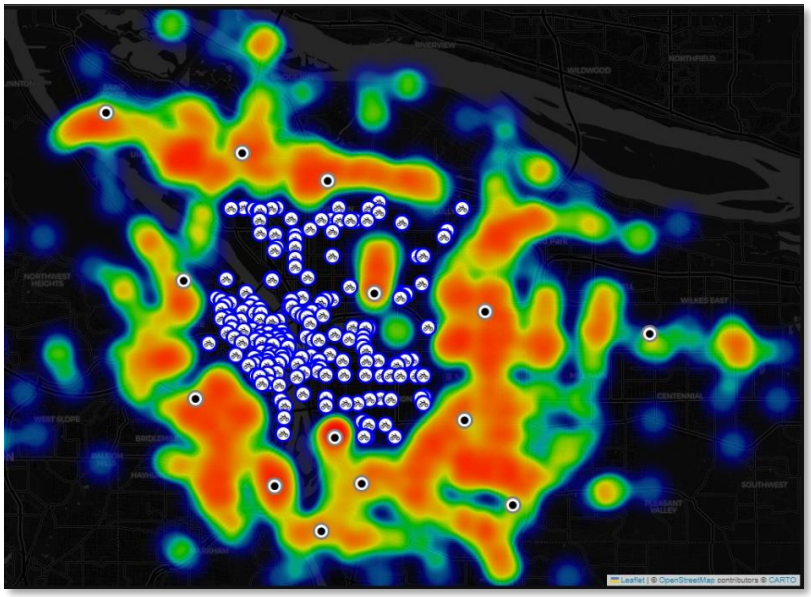


Figure 4 - City of Portland with Proposed Locations of New E-bike Stations

	cluster_labels	latitude	longitude
0	0	45.504974	-122.710020
1	1	45.464736	-122.655386
2	2	45.472621	-122.572445
3	3	45.531345	-122.584313
4	4	45.524640	-122.512927
5	5	45.571096	-122.652901
6	6	45.540744	-122.715197
7	7	45.479177	-122.638062
8	8	45.536890	-122.632437
9	9	45.478484	-122.675625
10	10	45.591755	-122.748786
11	11	45.493019	-122.649501
12	12	45.579395	-122.689865
13	13	45.498507	-122.593380

Figure 5 - Cluster Centroids in Latitude and Longitude Coordinates

ANOVA and Tukey HSD Results

A one-way ANOVA test was performed to compare the distance traveled by different groups of samples generated through clustering. The test revealed a significant difference in at least one group, which led to the **rejection of the null hypothesis** ($F(13) = 61.382364,\ p = 2.069228x10^{-153}$). A boxplot was used to visually confirm the differences between the group samples to plot the distribution of the distance traveled per cluster (Figure 6).

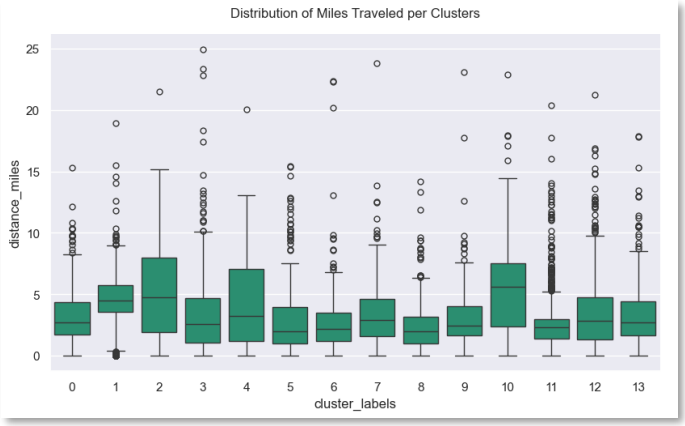


Figure 6 - Boxplot of the Distribution of Miles Traveled between Clusters

The Tukey HSD post-hoc test has made 91 pairwise comparisons between two sample groups and presented the mean difference, standard error, T-values, Tukey adjusted p-values, and effect size using Hedge's g statistic. Based on the top ten highest Hedge's g statistics, all having an adjusted p-value < 0.01 and a g-stat > 0.80, clusters marked as "1", "2", "4", and "10" were found to have the most practical significance compared to other clusters. These clusters were selected from the unique variables in the 'A' column of the pairwise table. Therefore, the most suitable locations to build new E-bike stations, or at least near these locations, are the cluster centers of cluster labels "1", "2", "4", and "10", as shown in Figure 7.

```
#- Filter by the top 10 rows of hedge's g value.
tukey_significant = tukey.sort_values('hedges', ascending=False).head(10)
print(f"\n##-- Clusters with the highest practical significance: {np.sort(tukey_significant['A'].unique())} --## \n")
tukey_significant
```

```
##-- Clusters with the highest practical significance: [ 1  2  4 10] --##
```

	A	B	mean(A)	mean(B)	diff	se	T	p-tukey	hedges
30	2	8	5.208980	2.261304	2.947676	0.269927	10.920273	0.0	1.478103
85	10	11	5.730140	2.497285	3.232855	0.186683	17.317387	0.0	1.357537
33	2	11	5.208980	2.497285	2.711694	0.264955	10.234562	0.0	1.236036
19	1	8	4.480476	2.261304	2.219172	0.124254	17.860018	0.0	1.126438
49	4	8	4.461890	2.261304	2.200586	0.215906	10.192355	0.0	1.023365
28	2	6	5.208980	2.623106	2.585873	0.279327	9.257516	0.0	0.979388
22	1	11	4.480476	2.497285	1.983190	0.113046	17.543248	0.0	0.940124
27	2	5	5.208980	2.745062	2.463918	0.276564	8.909043	0.0	0.885087
52	4	11	4.461890	2.497285	1.964605	0.209656	9.370614	0.0	0.859739
17	1	6	4.480476	2.623106	1.857369	0.143534	12.940294	0.0	0.813920

Figure 7 - Top 10 Pairwise Comparisons based on Hedge's g-stat and Code

The map in Figure 8 shows the density of locations where E-bike rentals began in the cleaned dataset. The blue circles on the map represent E-bike stations that are already established. The red circles with star icons indicate the optimal locations to build new E-bike stations. Additionally, gray circles represent other proposed E-bike stations. (Plotting with Folium, n.d.; Valkenburg, n.d.)

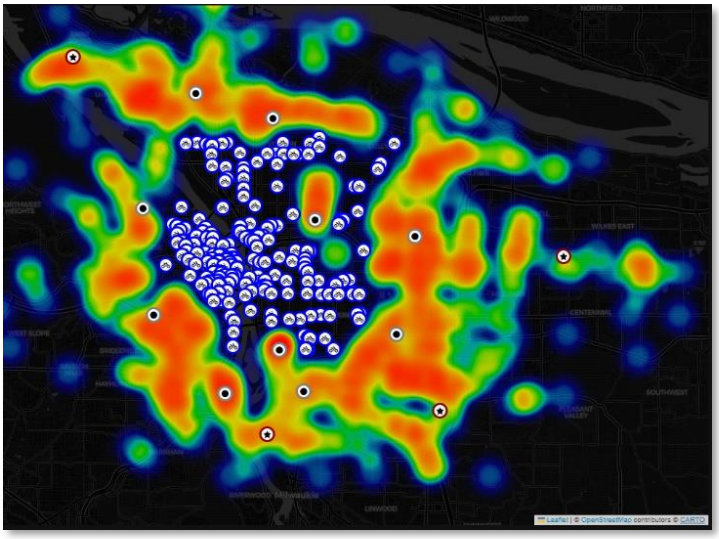


Figure 8 - Final Map with the Most Optimal E-bike Station Proposal Locations

Limitations of the Analysis

The data available for public access through BIKETOWN's dataset is limited to a specific timeframe (BIKETOWN PDX, 2020). This timeframe includes data from July 19th, 2016, to August 31st, 2020. Therefore, any data outside this timeframe is not included in the dataset and may be outdated.

Another limitation when considering cluster centroids as suggested locations for new E-bike stations is that K-means clustering can be sensitive to outliers (Nagar, 2020). The outliers found in the 'start_latitude' and 'start longitude' variables can affect the position of the cluster centroids, making them less reliable. As shown in Figure 9, the centroid for cluster "4" is not in the most densely populated areas of the map due to this limitation.

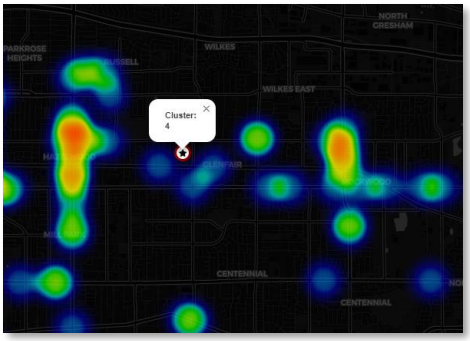


Figure 9 - Example of Cluster Centroid not in Densely Populated Areas

Proposed Actions

Recommended Course of Action

The demand for bike-sharing services from BIKETOWN PDX has been growing, and building more E-bike stations is necessary to meet this demand (2023 LMR BIKETOWN One-Pager, 2023). To achieve this, E-bike stations must be constructed at or near the proposed locations, especially the proposed locations linked to the most optimal clusters "1", "2", "4", and "10."

Most Optimal Locations

Cluster Labels	Latitude	Longitude	Nearest Intersection
1	45.464736	-122.655386	SE Tacoma St. & SE 11th Ave.
2	45.472621	-122.572445	SE Cooper St. & SE 89th Ave.
4	45.524640	-122.512927	NE Everett St. & NE 146th Ave.
10	45.591755	-122.748786	N Hudson St. & N Richmond Ave.

Other Proposed Locations

Cluster Labels	Latitude	Longitude	Nearest Intersection
0	45.504974	-122.710020	SW English Ct. & SW English Ln.
3	45.531345	-122.584313	NE Multnomah St. & NE 77th Ave.
5	45.571096	-122.652901	NE Dean St. & NE Bellevue Ave.
6	45.540744	-122.715197	NW Industrial St. & NW 31st Ave.
7	45.479177	-122.638062	SE Botsford Dr. & SE 28th Ave.
8	45.536890	-122.632437	NE Hancock St. & NE 32nd Ave.
9	45.478484	-122.675625	S Carolina St. & S Corbett Ave.
11	45.493019	-122.649501	SE Boise St. & SE 16th Ave.
12	45.579395	-122.689865	N Terry St. & N Omaha Ave.
13	45.498507	-122.593380	SE Franklin St. & SE 68th Ave.

Recommended Directions for Future Study of the Dataset

One way to handle the sensitivity of outliers for K-means Clustering is to transform the 'start_longitude' and 'start_latitude' variables using a log or square-root function. This transformation may help decrease the number of outliers in the dataset, and it could be worth investigating whether it changes the clustering solution.

Additionally, using the 'startdate' variable, it is possible to identify trends and clusters trending upward in customer usage of E-bike sharing services. Conducting a time-series analysis can help predict future demand for E-bike sharing in Portland (Bordogna, 2016). Based on the insights gained from this analysis, expanding existing E-bike stations or setting up new ones in key locations may be necessary.

Benefits of the Study

This study utilized clustering analysis through K-means Clustering to gain insights into the demand for bike-sharing services in the City of Portland. The results from the clustering analysis were then used to conduct further statistical testing using ANOVA and Tukey HSD tests. Based on the observed demand patterns, these tests aimed to determine the areas within the city where bike-sharing service expansion would be most beneficial.

In 2023, George S., Seles J.K.S., Brindha D., Jabaseeli T.J., and Vemulapalli L. conducted an analysis that introduced using K-means Clustering on geolocation data to extract more patterns and insights. Using this idea, a new variable called 'cluster_labels' with 14 cluster groups was created for this analysis. Using the new variable gives a better understanding of the dataset and provides additional insights into the locations where customer demand is found.

As a company providing bike-sharing services in Portland, finding ways to optimize decision-making is critical to using company resources efficiently. Hypothesis testing, such as ANOVA and Tukey HSD testing, helps to achieve accurate insight and enables executives to make wise decisions about business initiatives. ANOVA testing helps compare clusters from cluster analysis by treating each cluster as a sample group, allowing executives to make informed decisions. This analysis identified 4 clusters (1, 2, 4, and 10) as the most optimal using Hedge's g-stat from the pairwise table created through Tukey HSD comparisons.

The analysis identified potential locations for new E-bike stations through K-means clustering and optimal locations using ANOVA and Tukey HSD testing. The study's limitations encourage continued exploration of methods, tools, and techniques to meet service demand. Additionally, the methods from this analysis can guide future efforts in meeting business demand and promoting efficiency in the decisions made by executives in BIKETOWN.

Presentation of Findings

Link to Panopto Video:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b4578b5c-90a1-41e3-bcd9-b16f01646441

"Please see Bryan_Bravo_Capstone_Presentation.pptx"

Sources

2023 LMR BIKETOWN One-Pager. (2023). *biketownpdx.com*. Retrieved April 15, 2024, from https://drive.google.com/file/d/1MXgVuQxIMg2cJV66pX1EJqmbG-4VWcEj/view

BIKETOWN PDX. (2020, September 14). Trip Data For Download. *BIKETOWN PDX*. Retrieved April 15, 2024, from https://biketownpdx.com/system-data

Bobbitt, Z. (2019, April 14). A Guide to Using Post Hoc Tests with ANOVA. *Statology*. Retrieved April 16, 2024, from https://www.statology.org/anova-post-hoc-tests/

Bobbitt, Z. (2022, June 16). Understanding the Null Hypothesis for ANOVA Models. *Statology*. Retrieved April 20, 2024, from https://www.statology.org/null-hypothesis-for-anova/

Bordogna, G., Kliment, T., Fiderio, L., Brivio, P., Crema, A., Stroppiana, D., Boschetti, M., Sterlacchini, S. (2016, May 21). A Spatial Data Infrastructure Integrating Multisource Heterogeneous Geospatial Data and Time Series: A Study Case in Agriculture. *ISPRS Int J. Geo-Inf, 5(5)*, 73. https://doi.org/10.3390/ijgi5050073

Daityari, S. (Narrator). *Cluster Analysis in Python* [Online video]. DataCamp. Retrieved April 17, 2024, from https://app.datacamp.com/learn/courses/cluster-analysis-in-python

George, S., Seles, J. K. S., Brindha, D., Jabaseeli, T. J., Vemulapalli, L. (2023, December 11). Geopositional Data Analysis Using Clustering Techniques to Assist Occupants in a Specific City. *Engineering Proceedings, 59(1)*, 8. Retrieved April 16, 2024, from https://doi.org/10.3390/engproc2023059008

Nagar, A. (2020, January 26). K-means Clustering - Everything you need to know. *Medium.com*. Retrieved April 20, 2024, from https://medium.com/analytics-vidhya/k-means-clustering-everythingyou-need-to-know-175dd01766d5#f6a0

Pingouin.pairwise_tukey. (n.d.). Pingouin. Retrieved April 18, 2024, from https://pingouin-stats.org/build/html/generated/pingouin.pairwise_tukey.html

Plotting with Folium. (n.d.). *GeoPandas.org*. Retrieved April 19, 2024, from https://geopandas.org/en/stable/gallery/plotting_with_folium.html

Pulkit, S. (2024, February 20). The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications. *Analytics Vidhya*. Retrieved April 20, 2024, from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-meansclustering/

Singh, G. (2024, February 7). ANOVA: Complete guide to Statistical Analysis & Applications (Updated 2024). *Analytics Vidya*. Retrieved April 15, 2024, from https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

Valkenburg, M. V. (Narrator). *Visualizing Geospatial Data in Python* [Online video]. DataCamp. Retrieved April 15, 2024, from https://app.datacamp.com/learn/courses/visualizing-geospatial-data-in-python