University of Colorado Boulder

Cover Page

Milestone 4

CSCI 5502: Data Mining

SECTION 872

Instructor: Dr. Alfonso G. Bastias

By

Joel Hoversten & Aloken Chaudhari

Boulder, Colorado

28 April 2025

Website:

https://sites.google.com/colorado.edu/5502-chaudhari-hoversten/home

GitHub Repository:

https://github.com/joelhov09/CSCI-5502/blob/main/Milestone%202

Using Machine Learning to analyze Public Transportation effect on Environmental Sustainability

Joel Hoversten & Aloken Chaudhari

April 28 2025

Instructor: Dr. Alfonso G. Bastias

Section 872

Abstract

Public transport and its accessibility have been greatly increasing in the recent years. This issue is something highly specific to regions and cities in the US, where certain cities have higher access and infrastructure in place for public transport than others. In order to examine which cities perform the best in terms of carbon emissions it be beneficial to look into the political influences in that area to see it sustainability is a priority for the region. The popularity of environmental sustainability between rural and urban areas may also indicate similarly to political influence. The initiatives for some of the key decision-makers when it comes to possibly could be heavily influenced by analysis like this. Insights into these factors would allow us to draw conclusions about which cities have the most sustainable models.

1 Introduction

The role of public transport and its accessibility has become increasingly crucial in recent years, particularly against the backdrop of growing concerns over environmental sustainability and climate change. As urban populations expand and greenhouse gas (GHG) emissions continue to rise, the need for efficient, low-emission transportation alternatives has never been more urgent. Public transit systems offer a promising solution, providing high-capacity, lower-emission mobility options compared to private vehicle use.

This analysis utilizes the Intermodal Passenger Connectivity Database (IPCD) and various other datasets to examine the relationship between public transit accessibility and greenhouse gas emissions across various U.S. cities and regions. By integrating transportation infrastructure data with emission statistics, this

study seeks to identify whether enhanced transit accessibility correlates with reduced emissions levels.

Current research has addressed public transportation benefits in isolated cases; however, comprehensive, data-driven studies that empirically link transit accessibility to emission outcomes across multiple cities remain limited. This project aims to fill that gap by offering systematic analysis and comparative insights. Through this research, actionable findings can be generated for policymakers, urban planners, and sustainability advocates, potentially informing decisions about resource allocation, urban development, and climate policy initiatives.

Furthermore, recognizing that access to public transportation and resulting environmental impacts are not uniform across the United States, the study pays particular attention to regional differences. Cities with robust transportation infrastructures will be compared against those with limited access, enabling a nuanced understanding of how transit systems contribute to—or fail to mitigate—urban carbon footprints. In addition to infrastructure and demographic variables, political and policy contexts are considered to account for how local priorities and governance might influence sustainability outcomes.

2 Related Work

A growing body of literature underscores the pivotal role public transportation can play in mitigating greenhouse gas emissions and promoting sustainable urban development. Among the most comprehensive sources, the Federal Transit Administration's 2010 report, Public Transportation's Role in Responding to Climate Change, provides empirical evidence that public transit systems are significantly more environmentally efficient than private automobiles. According to national-level data, modes such as heavy rail transit produce up to 76% less carbon dioxide per passenger mile compared to single-occupancy vehicles, with light rail and bus transit following closely behind.

In addition to direct emissions savings, public transportation fosters compact land use patterns that inherently reduce the need for long car trips. These transit-oriented developments (TODs) encourage walkability, mixed-use zoning, and reduced car ownership—factors that cumulatively shrink the urban carbon footprint. Empirical studies cited by the FTA show that for every mile traveled on public transit, automobile travel may decrease by 1.4 to 9 miles due to shifts in land use and travel behavior. Such findings support the notion that transit investment can have a ripple effect across both direct and indirect sources of emissions.

Further, life-cycle analyses show that even when accounting for the energy and emissions associated with constructing and maintaining transit infrastructure, public transportation retains its environmental advantages. For instance, incorporating emissions from infrastructure into calculations still results in a 55–70% savings in per-passenger emissions compared to car travel, particularly for systems like San Francisco's BART and New York's subway network.

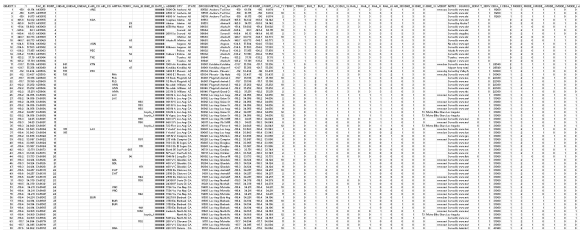
Beyond infrastructure and operational comparisons, policy and planning efforts have also been linked to climate benefits. Programs such as the Transit Investments for Greenhouse Gas and Energy Reduction (TIGGER) initiative demonstrate how federal funding can incentivize local transit agencies to adopt cleaner technologies—such as hybrid buses, solar-powered facilities, and energy-efficient operations. These programs align with broader federal goals of fostering "livable communities" that reduce dependence on fossil fuels, improve accessibility, and minimize emissions.

Despite these advances, most prior work focuses on descriptive or case-specific analysis and lacks a comparative, city-by-city evaluation grounded in spatial and demographic diversity. This gap highlights the need for broader, data-driven research to identify which regions most effectively leverage transit accessibility for environmental benefit. By integrating city-level transit access data with emissions statistics, this study builds on and extends the foundational insights of prior work—shifting the lens from national aggregates to local realities.

3 Methods

3.1 Data Acquisition

One of our sources for data comes from the U.S. Energy Information Administration website. Their site offers a registration for an API that we used to get data on the gas consumption per state. This API provided the state and the corresponding Production, Consumption per Capita and Expenditures per Capita for each of the state. A corresponding rank was assigned for each state in this data and is also included.



The above dataset was merged with a webscraped dataset from the Bureau of Transportation Statistics. The original dataset was at the city level and we merged it at the state level with the EIA dataset to analyze. This dataset gave us the types of transportation offered in each state and how many of each was offered. The API was also used to get CO2 emissions for the state. The API returned the sector of the fuel used by the state as well as the type of fuel used in the set. This can be useful to see which states used what kind of fuel in the future. This data will be extremely helpful to see what kind of transportation

is offered for each state and where those states ranked. The analysis may show a higher ranking for a particular type of transportation offered.

3.2 Data Preprocessing

After the initial data collection phase, the raw dataset required extensive cleaning before analysis. The goal of preprocessing was to prepare the data for integration with the EIA (Energy Information Administration) dataset and ensure it aligned with the research objective of exploring the relationship between transit accessibility and greenhouse gas emissions.

To streamline the dataset, several columns were dropped, including NEAR ID 1, NEAR ID 2, NEAR ID 3, AIR CODE, AIR CODE2, FERRY CODE, AMTRAKCODE, RAIL ID, BIKE ID, WEBSITE, SOURCE, NOTES, CBSA CODE, and CBSA TYPE. These fields were removed for two primary reasons: they contained a high proportion of null or missing values, and they did not contribute meaningfully to the research questions focused on state-level or city-level accessibility and emissions metrics.For columns such as BIKE SYS, ADDRESS, FAC NAME, and METRO AREA, missing values were not dropped but instead replaced with the placeholder "unknown". This decision was driven by the observation that these columns provided region-specific transportation insights—particularly for urban centers—and removing them would have resulted in the loss of potentially valuable categorical indicators. Outliers were identified in some transit-related numerical fields, such as facility counts or system capacity values. Instead of treating these as anomalies, contextual analysis showed that these values were consistent with known population and infrastructure distributions across states. For instance, higher values in California (e.g., for bike share systems or rail stations) reflect its larger and more urbanized environment compared to states like Iowa, which naturally have lower transit availability. These differences were preserved in the dataset to maintain the integrity of regional comparisons.

The STATE column was retained and treated as a categorical variable, containing standardized two-letter state abbreviations used to group and compare observations across regions. Other categorical features of importance included Fuel Type and Sector Name—especially in the merged EIA dataset—which were critical for segmenting energy consumption data by fuel source and end-use context.

The final cleaned dataset consisted predominantly of categorical and numerical data and represented a comprehensive merger of the IPCD transit access data and the EIA emissions data. This unified dataset formed the foundation for downstream analysis using classification, regression, clustering, and frequent pattern mining techniques, with preprocessing ensuring both consistency and analytical relevance.

	STATE	Productio	Productio	Consumpt	Consumpt	Expenditu	Expenditu	Total_Faci	FAC_TYPE	MODE_BU	MODE_AIR	MODE_RAI	MODE_FEI	MODE_BIK	BIKE_SHARE
0	AK	1.4	13	987	1	13051	1	302	5	30	233	26	43	0	906
1	WY	6.1	4	853	4	11221	2	1	1	0	1	0	0	0	0
2	ND	4.2	8	861	3	10507	3	1	1	0	0	1	0	0	3
3	LA	4.9	6	925	2	9781	4	7	5	7	2	5	2	0	21
4	IA	0.8	19	445	7	6927	5	5	3	4	2	3	0	0	15
5	TX	25.5	1	459	6	6748	6	18	6	17	1	17	0	6	21
6	MT	0.7	20	352	15	6378	10	2	2	2	0	2	0	0	6
7	OK	4.3	7	380	11	6132	11	2	2	2	2	0	0	0	6
8	WV	5.9	5	471	5	6129	12	4	2	2	0	4	0	0	6
9	AL	1.1	14	375	12	6110	13	18	3	12	6	3	0	0	54
10	KY	0.9	15	371	14	6062	14	2	2	2	0	2	0	0	6
11	ME	0.1	43	241	34	6008	15	7	4	6	3	2	2	0	21
12	MS	0.3	35	374	13	5985	16	4	2	4	0	4	0	0	12
13	KS	0.6	24	341	17	5873	17	1	1	0	0	1	0	0	3
14	IN	0.8	18	383	10	5800	18	11	4	11	5	9	0	0	15
15	AR	0.7	22	346	16	5709	19	44	4	29	8	19	0	0	132
16	NM	6.8	3	325	18	5562	20	7	4	7	2	5	0	0	21
17	VT	0	47	193	46	5532	21	4	3	4	2	2	0	0	6
18	NH	0.2	42	213	38	5465	22	4	3	4	2	2	0	0	6
19	MN	0.5	31	308	19	5244	23	16	6	15	6	14	0	1	18
20	WI	0.3	36	300	21	5187	24	13	6	12	4	12	0	3	6
21	TN	0.5	30	298	22	5158	25	1	1	1	1	0	0	0	0
22	CA	1.6	12	176	49	5123	26	787	9	668	31	655	29	180	943
23	NV	0.1	45	222	36	5083	27	4	3	4	2	2	0	0	6
24	CT	0.2	41	196	43	5022	29	19	5	19	2	17	4	0	57
25	MO	0.2	39	281	26	4951	31	10	4	10	4	9	0	0	30
26	VA	0.8	17	280	27	4889	32	38	6	38	3	38	0	11	41
27	CO	3.6	9	251	33	4857	33	3	2	3	0	3	0	0	6
28	PA	10.1	2	288	25	4848	34	71	7	69	7	69	0	27	57
29	MA	0.1	46	188	47	4842	35	75	9	70	4	68	7	41	84
30	OH	3	10	298	23	4829	36	12	5	10	2	12	0	0	36
31	IL	2.2	11	292	24	4825	37	68	7	61	7	66	7	26	66
32	DE	0	50	270	29	4725	38	5	3	5	0	5	0	0	0
33	GA	0.6	23	260	31	4668	39	6	4	6	3	6	0	3	3
34	OR	0.7	38	262	41	4594	40	24	5	21	6	20	0	7	47

3.3 Data Visualization

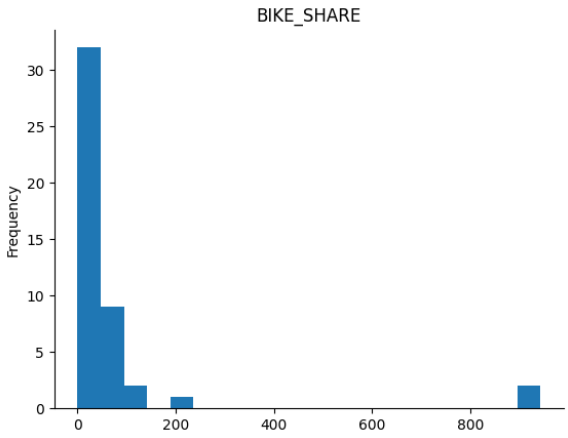


Figure 1: Shows the frequency of Bike Share programs. Most states seem to offer under 200 bike share options with 2 states offering over 800. Those two states have high urban areas and are considered biking cities .

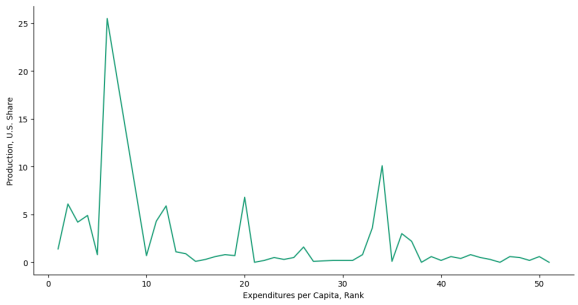


Figure 2: Shows the Expenditures vs the Production by rank and share. The 3 peaks show that those ranks may be outliers when it comes to production, but looking at the graph there may be some indicators in rank that show production.

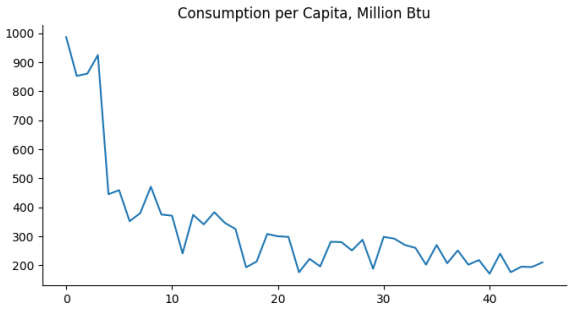


Figure 3: A graph of the frequency of Consumption per Capita

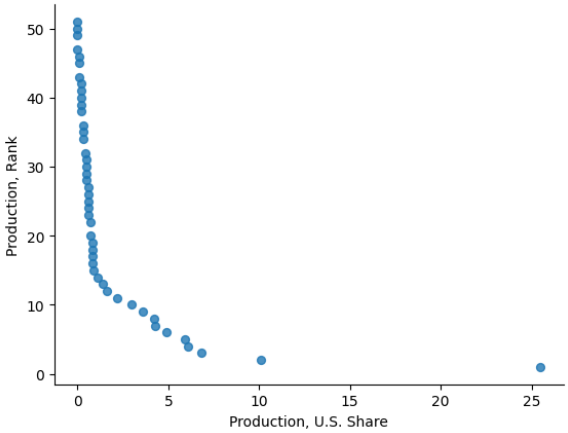


Figure 4: This just showcases that the rank of production is lower when the

produciton oif that stat is higher which makes sense that there is a negative correlation and proves a robust ranking system.

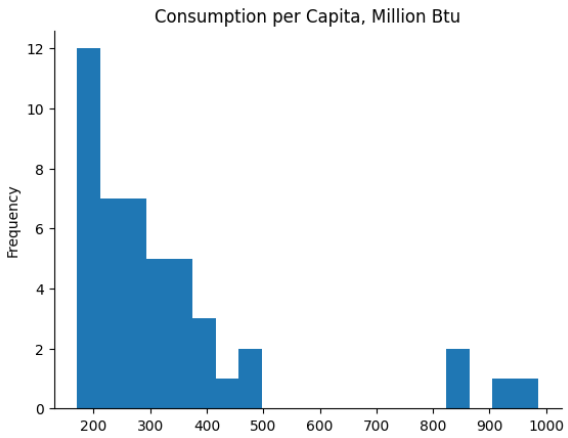


Figure 5: A graph of the frequency of Consumption per Capita. The downward histogram shows that consumption is mainly lower for most states with a few outliers to address.

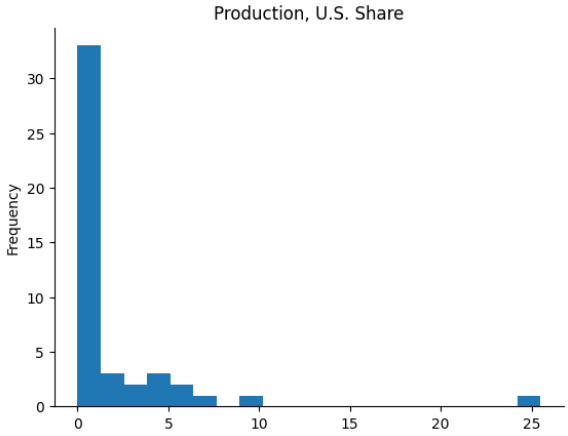


Figure 6: A graph of the frequency of Production. The downward histogram shows that production is mainly lower for most states with a few outliers to address.

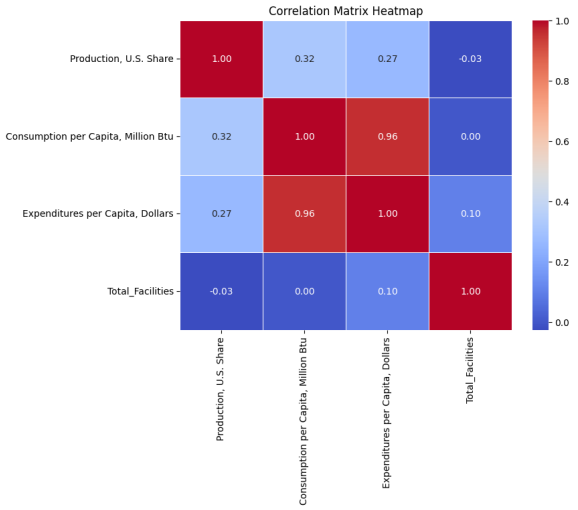


Figure 7: Takes the four categories from the merged dataset (no ranks) and graphs their correlation matrix. With low vfalues for correlation there is little evidence for multicollinearity

4 Evaluation

4.1 Frequent Pattern Mining: APRIORI

FP growth is a bit better than Apriori when it comes to handling sparse data. The tree FP-structure is something that I believed would be able to find patterns more easily with the data we have. The model is assuming the data is transactional, so data structure is important. It also assumes values were converted to binary values based on thresholds especially for Millions of BTUs consumed. The model also operates under the assumption that the presence or absence of one item in a transaction is independent of the presence or absence of other items.

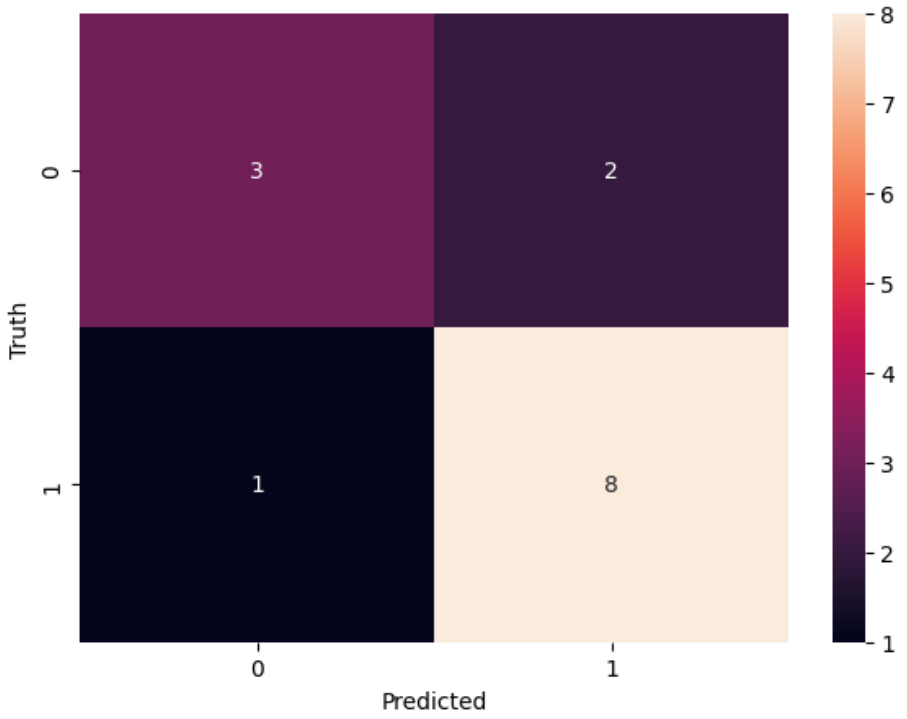
```
Frequent Itemsets:
    support                                    itemsets
0       1.0                                (Petroleum)
1       1.0                              (Natural Gas)
2       1.0                                    (Coal)
3       1.0                                (All Fuels)
4       1.0                  (Petroleum, Natural Gas)
5       1.0                         (Petroleum, Coal)
6       1.0                    (All Fuels, Petroleum)
7       1.0                       (Coal, Natural Gas)
8       1.0                  (All Fuels, Natural Gas)
9       1.0                         (All Fuels, Coal)
10      1.0            (Petroleum, Coal, Natural Gas)
11      1.0       (All Fuels, Petroleum, Natural Gas)
12      1.0              (All Fuels, Petroleum, Coal)
13      1.0            (All Fuels, Coal, Natural Gas)
14      1.0  (All Fuels, Petroleum, Coal, Natural Gas)
```

The threshold for binary conversions of the Millions of BTUs consumption per capita was originally set to the value proposed for sustainability, but a lot the states do not meet this threshold, so it had to be tweaked. Figuring out how to organize the data so that this would be a functional method was one of the biggest challenges faced. The analysis revealed that states with above-median public transportation access were frequently associated with below-median energy consumption per capita (Support = 0.7826, Confidence = 0.7826, Lift = 1.00), suggesting a potential inverse relationship between transit access and emissions. One major challenge was transforming state-level numeric features into meaningful transactional items without introducing noise.

4.2 Classification: SVM

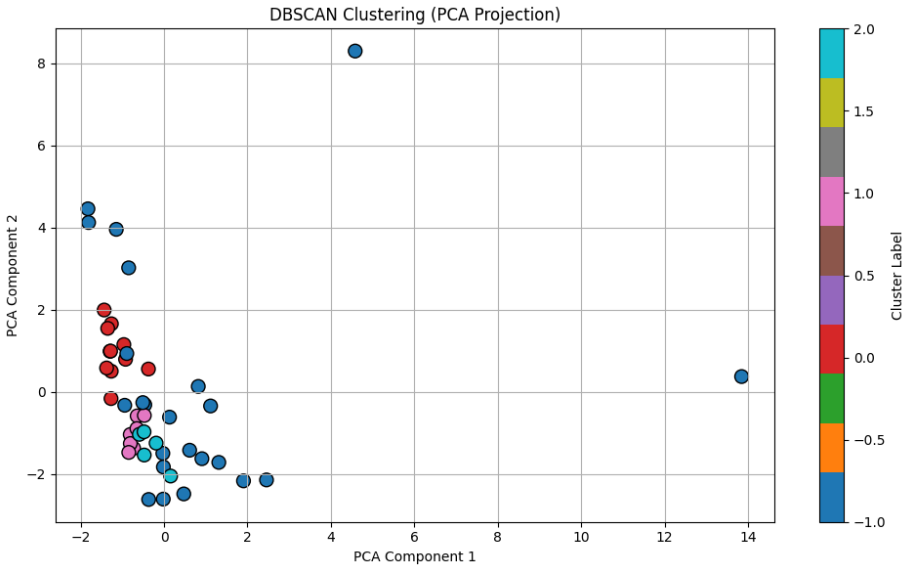
For the classification model of the data, the Support Vector Machine approach was used. This is useful because our data has a high dimension with lots of numerical data for only 50 states. With a small sample size of high dimensions SVM is preferred to find a hyperplane that separates that classify the data. Our SVM model separates the Red vs Blue states in our data. This supports the SVM assumption that the data is linearly separable. We scale our features first and then apply the SVM model using a linear kernel. We also tried using an "rbf" and "poly" kernel, but the linear kernel returned the best results. It is also important to note that our data included some redundant and unnecessary columns that we had to drop like the State abbreviation and the State name

before we were able to conduct an SVM analysis. The model returned an Accuracy of 0.79, a Precision 0.8 , a Recall of 0.889, an F1-score of 0.842, and a ROC-AUC of 0.822.



4.3 Clustering: DBSCAN

We used the DBSCAN method for clustering the dataset. This is good because DBSCAN does not assume the number of clusters and does not assume the clusters form a pattern. DBSCAN assumes that the density of the clusters are similar to each other. DBSCAN works especially well for our data since some of the data varies drastically form others. California and Alaska are standout "noise" states since they have very different access to transportation and largely different populations for those states. DBSCAN also works well with scaled data which we scale before applying the method. We selected DBSCAN for its ability to discover non-spherical clusters and handle outliers in our small dataset of 50 states. After scaling the data, we used a k-distance plot to select an epsilon of 0.5 and min_samples of 3. Our model returns a silhouette score of 0.431 and a Davies-Bouldin Index of 0.818, this indicates moderate cluster seperation. The eps and min_samples for our DBSCAN is 1.2 and 4



4.4 Regression: Logistic Regression

The choice between Logistic and Linear Regression came down to the response variable. Since our response variable was binary in classifying if the state is red or blue we chose to go with logistic regression. The features we have were numerical, so it was easy to implement a logistic regression after dropping some identifier columns like state name. Similar to SVM we only had the 50 states as a sample, so it fits a logistic regression. Logistic regression assumes a linear relationship between the response variable and the independent variables in log-odds form. It is also nice to have an easy equation as an outcome. The only preprocessing required was dropping the unneeded columns and then the model returned a RMSE of 0.598, an MSE of 0.357 and an R-Squared of -0.556. Key predictors included energy consumption per capita and presence of bike systems, with the former having a strong negative coefficient, indicating higher consumption is associated with red states.

5 Conclusion

5.1 Results & Insights

Overall, the SVM model outclassed the other models with a higher Accuracy, F-1 Score and ROC-AUC. The SVM model did a better job at handling any class imbalance and gave a solid hyperplane that separated the data. Logistic Regression still had strong numbers but in general does a better job about independent variable explanation. If you wanted to see which variables affect

red vs blue states then Logistic regression would be easier to explain. SVM is very effective in high dimensional datasets like the one we use and used all the features to develop the separating hyperplane. The resulting confusion matrix misclassified 3 out of 14 and properly classified 11 out of 14 in our dataset. It correctly predicted 3 out of the 5 blue states and 8 out of the 9 red states. Clustering with DBSCAN revealed distinct groupings of states based on transportation-emission profiles, identifying states like California and Alaska as structural outliers. Frequent pattern mining uncovered strong associations between low energy consumption and presence of multiple transit options. These findings collectively support the hypothesis that transit accessibility is linked to environmental sustainability and offer a foundation for targeted policy recommendations.

5.2 Ethical Reflection

It is important to note that modeling based off the two political parties can oversimplify the complex political landscape. The ecological fallacy is that a group-level pattern of a particular political party cannot be assumed to be applied on a smaller scale or to an individual. This classification of "red" and "blue" states obscures the the differences between the states. Some states have larger divides in socioeconomic status, rural and urban areas as well as demographic differences.

Since our dataset is only the 50 states it is limited to some of the risks of a small dataset. Variance, reliability and overfitting can lead to misleading predictions. Our Dataset includes all 50 states, but does not have the goal to use state as a group to predict urban and rural differences or even demographic differences. Other countries aiming to use this research shoudl be aware of the more urban and rural states to accurately interpret the results.

5.3 Future Work

While the current study has provided meaningful insights into the relationship between public transportation accessibility and greenhouse gas emissions, several avenues exist to deepen and expand the analysis in future work.

5.4 Model Enhancements

One limitation of the current approach is its reliance on relatively simple models, such as SVMs and Logistic Regression. These models offer interpretability but may fall short in capturing complex, nonlinear relationships within the data. Future studies could explore ensemble learning techniques such as Random Forests or XGBoost, which are better suited for identifying intricate feature interactions and typically provide improved classification performance.

For clustering tasks, K-Means++ may offer a complementary perspective to DBSCAN by enabling more interpretable, hard-cluster assignments while addressing initial centroid selection bias. Additionally, deep learning approaches,

such as artificial neural networks (ANNs), may provide enhanced generalization capabilities—particularly in handling the high dimensionality present in state-level datasets.

5.5 Dataset Additions

The analysis could be substantially enhanced by incorporating additional datasets that capture political, urban, and socioeconomic dimensions more precisely. For example, integrating state-level voting records or environmental policy adoption scores would allow for a more nuanced understanding of political alignment as it relates to sustainability practices.

Furthermore, more granular urban data, such as city-specific infrastructure metrics from OpenStreetMap or U.S. Census microdata, could provide finer resolution on transportation access and usage. Adding socioeconomic indicators—such as median household income, educational attainment, and healthcare accessibility—could contextualize the observed patterns and control for potential confounding variables.

Finally, refining the emissions data to distinguish CO_2 emissions by transportation mode (e.g., personal vehicles vs. public transit) would enable a more targeted evaluation of public transportation's environmental impact, helping to isolate the benefits of specific systems and policies.

6 References

- https://www.eia.gov/opendata/pdf/EIA-APIv2-HandsOn-Webinar-11-Jan-23.pdf
- https://www.eia.gov
- https://www.eia.gov/state/
- https://data-usdot.opendata.arcgis.com/datasets/usdot::intermodal-passenger-connectivity-database-ipcd/explore?location=34.522898