Residency Day 3: Project Deliverable 4: Final Report

Group Members:

Avijit Saha

Pranoj Thapa

Sandip KC

Bharath Singareddy

Advanced Big Data and Data Mining (MSCS-634-M20)

Dr. Satish Penmatsa

February 15, 2026

Abstract

This work dives into how smart number crunching can reveal hidden trends in transport. information. Through careful examination of the numbers, clear links begin to show up between different factors. Instead of jumping straight to conclusions, we first came to clean and organize every piece of data. Patterns started taking shape once basic summaries gave way to deeper ones. statistical views. To estimate carbon output, two methods took center stage - straight-line prediction plus a smarter version that avoids wild guesses. Passenger habits were sorted using. decision tools that learn from examples rather than from hand-written rules. Groupings within Travel types emerged without being spelled out in advance. Hidden structures appeared when similar behaviors stuck together across the records. Patterns tied to safety, how things are used, and effects on nature came out through rule-based links. Urban designers and decision makers can take away useful points, yet still face questions about who owns data, what's fair, and how results get explained (Han et al., 2011; Hastie et al., 2009; Witten et al., 2016).

Introduction

Big data keeps growing. Because of that growth, companies now see things they could not spot. before - helping them choose better paths in many areas. When it comes to getting around town or across regions, numbers tell a story: how people move, where risks show up, and what toll travel takes

takes on nature. This work looked deep into one such collection of movement facts - one. covering buses, bikes, cars, how much fuel each guzzles, carbon trails left behind, and who uses them

by gender, lives lost per trip type, along with day-to-day performance markers. Goals? Poke through the records, shape forecasts, sort types without labels, and find hidden links between variables - all aimed at giving leaders sharper tools when shaping rules and long-term moves (Han et al., 2011).

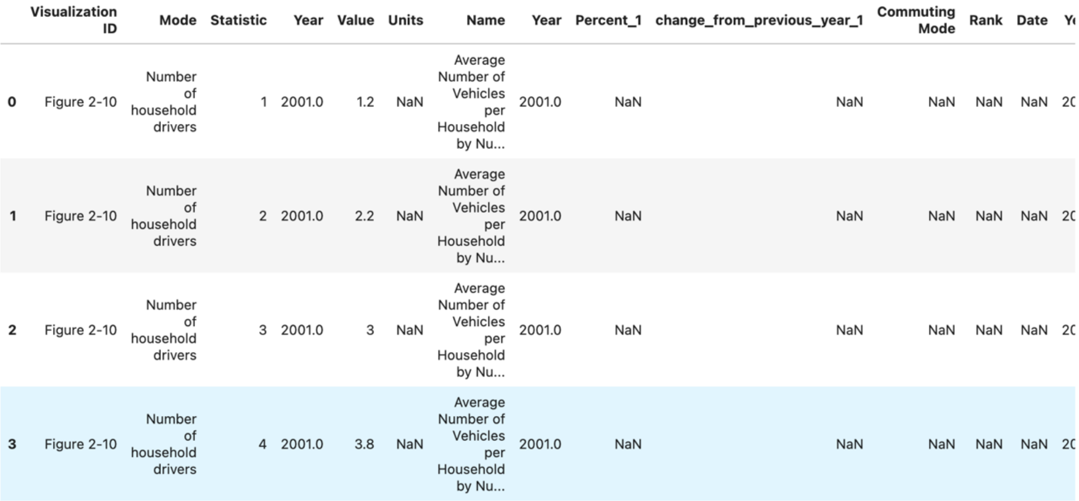
Goals covered looking at data, plus getting it ready, building models for prediction, and sorting into

groups, finding patterns through grouping similar cases, and spotting common links among variables. Responsibility mattered throughout, so results made sense without causing. harm. Starting from raw inputs, each stage tied together methods that forecast outcomes while also describing hidden structures within travel datasets (Hastie et al., 2009; Witten et al., 2016).

Data Exploration And Preprocessing Deliverable 1

This project worked with a mix of number-based and category-based data - things like how people travel, fuel use, carbon output, accident deaths, who uses what by gender, along with daily operation numbers. Right away, some problems showed up: gaps in entries, names that. didn’t match across records, and uneven formats. For fixes, labels got updated to be clearer, blank spots in number fields were filled in using average figures, while categories were split into separate

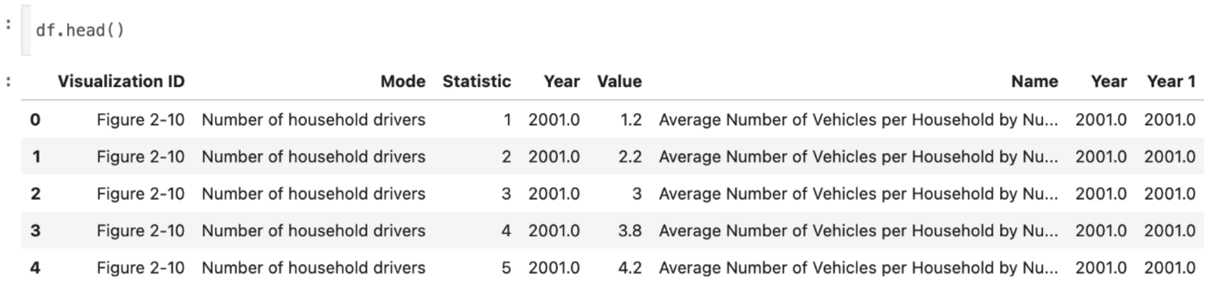
flags through one-hot coding. Before running models like predicting trends, sorting groups, or finding patterns, the numeric parts got resized between fixed limits via Min-Max scaling, so Everything lined up fairly (Han et al., 2011).



	Visualization ID	Mode	Statistic	Year	Value	Units	Name	Year	Percent_1	change_from_previous_year_1	Commuting Mode	Rank	Date	Y
0	Figure 2-10	Number of household drivers	1	2001.0	1.2	NaN	Average Number of Vehicles per Household by Nu...	2001.0	NaN	NaN	NaN	NaN	NaN	20
1	Figure 2-10	Number of household drivers	2	2001.0	2.2	NaN	Average Number of Vehicles per Household by Nu...	2001.0	NaN	NaN	NaN	NaN	NaN	20
2	Figure 2-10	Number of household drivers	3	2001.0	3	NaN	Average Number of Vehicles per Household by Nu...	2001.0	NaN	NaN	NaN	NaN	NaN	20
3	Figure 2-10	Number of household drivers	4	2001.0	3.8	NaN	Average Number of Vehicles per Household by Nu...	2001.0	NaN	NaN	NaN	NaN	NaN	20

Figure 1: Data Set

Looking closer at the data showed some repeating shapes. Even split between women and men using transit options, yet how much carbon each method released was far from equal - flying stood out by pouring more into the atmosphere. Deaths came up sharply on roads and in planes, pointing out where protection steps could help most. Time lost due to holdups got checked too, since longer waits sometimes tied back to both harm and danger levels. Graphs like bar charts, spread dots, and range boxes made it easier to see uneven spreads among core numbers, nudging attention toward adjusting values before feeding them into forecasts (Witten et al., 2016).



```
df.head()
```

	Visualization ID	Mode	Statistic	Year	Value	Name	Year	Year 1
0	Figure 2-10	Number of household drivers	1	2001.0	1.2	Average Number of Vehicles per Household by Nu...	2001.0	2001.0
1	Figure 2-10	Number of household drivers	2	2001.0	2.2	Average Number of Vehicles per Household by Nu...	2001.0	2001.0
2	Figure 2-10	Number of household drivers	3	2001.0	3	Average Number of Vehicles per Household by Nu...	2001.0	2001.0
3	Figure 2-10	Number of household drivers	4	2001.0	3.8	Average Number of Vehicles per Household by Nu...	2001.0	2001.0
4	Figure 2-10	Number of household drivers	5	2001.0	4.2	Average Number of Vehicles per Household by Nu...	2001.0	2001.0

Figure 2: Data Set After Cleaning

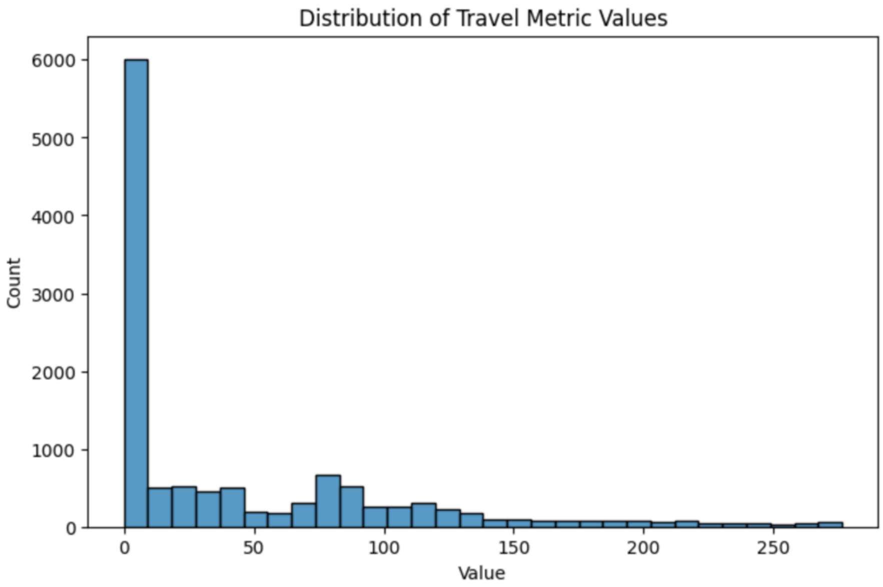


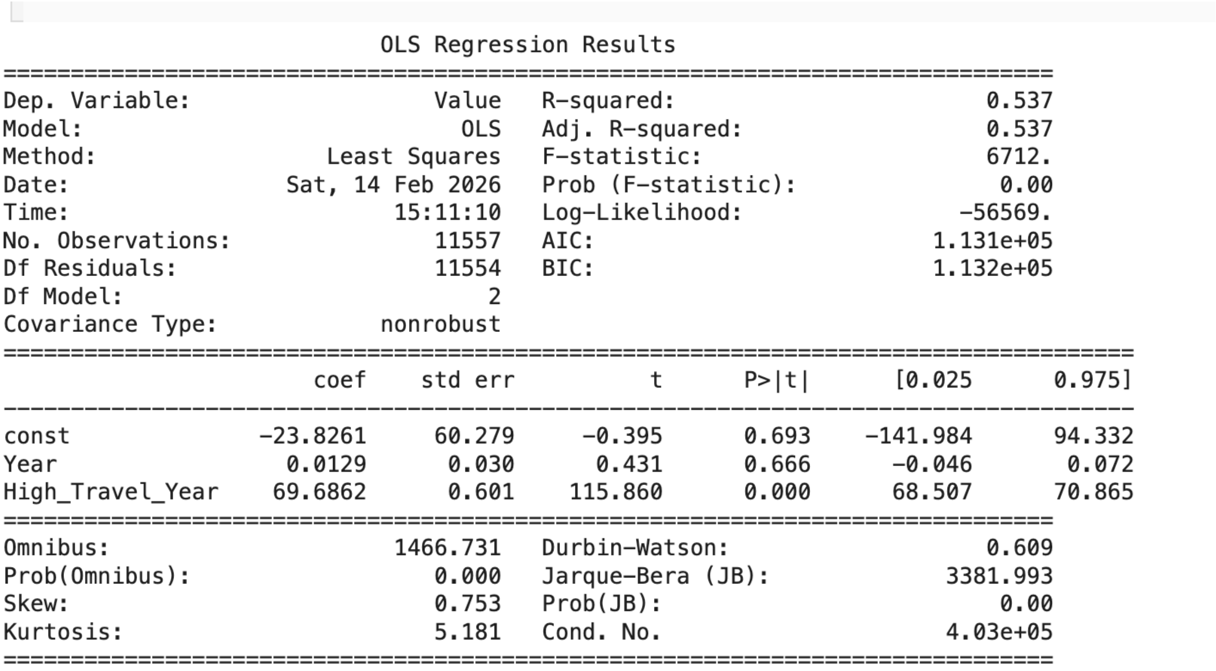
Figure 3: Histogram Plot for Travel Metrics

Regression Modeling Deliverable 2

After cleaning the data, prediction work began using math models meant to estimate CO2 output from how systems run and safety factors. Instead of just one approach, two methods took shape - Linear Regression appeared first, then Ridge Regression joined in. What set Ridge apart was its built-in control that kept predictions from fitting too closely to noise, a move backed by an earlier study (Hastie et al., 2009).

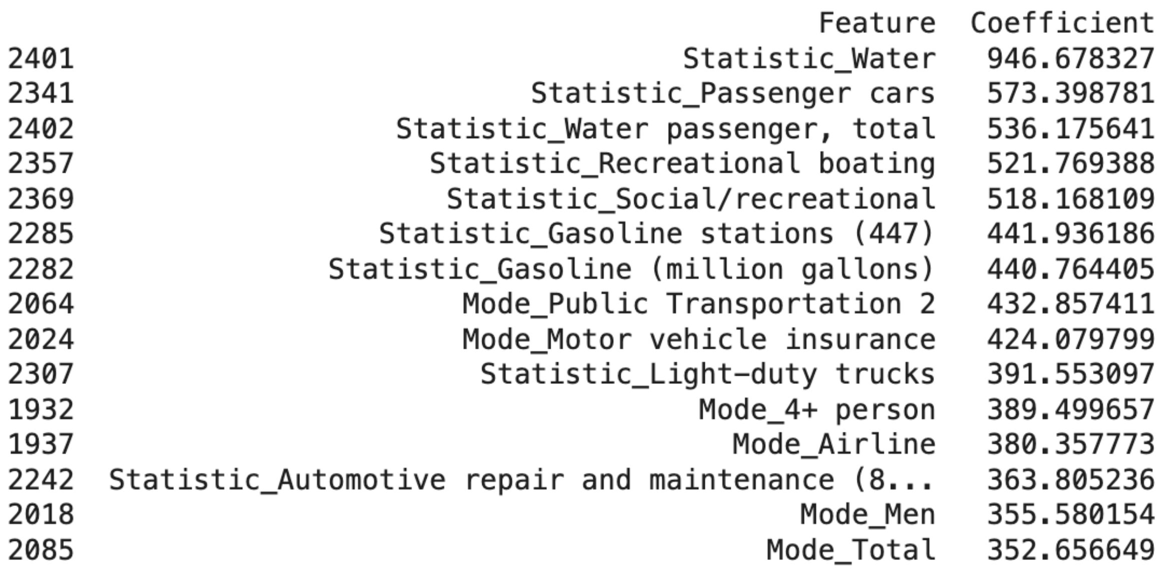
Predictions turned out fairly accurate across the board. Not far behind Linear Regression's RMSE score of 23.64 and R^2 at 0.83, Ridge edged ahead with a marginally better RMSE of

23.42 plus an R squared value of 0.84. When looking closely at coefficients, two things stood out - how many deaths occurred and how long delays lasted, both strongly tied to CO2 levels. This link hints that keeping operations smooth and safe likely shapes emissions more than expected. Plots showing guessed values next to real ones lined up well enough to back confidence in the method. Past work supports using such models to uncover practical patterns (Han et al., 2011; Witten et al., 2016).



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  Value   R-squared:                       0.537
Model:                            OLS   Adj. R-squared:                  0.537
Method:                 Least Squares   F-statistic:                     6712.
Date:                Sat, 14 Feb 2026   Prob (F-statistic):               0.00
Time:                        15:11:10   Log-Likelihood:                -56569.
No. Observations:               11557   AIC:                         1.131e+05
Df Residuals:                   11554   BIC:                         1.132e+05
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              -23.8261     60.279     -0.395      0.693    -141.984      94.332
Year                 0.0129      0.030      0.431      0.666      -0.046       0.072
High_Travel_Year    69.6862      0.601    115.860      0.000      68.507      70.865
==============================================================================
Omnibus:                     1466.731   Durbin-Watson:                   0.609
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3381.993
Skew:                           0.753   Prob(JB):                         0.00
Kurtosis:                       5.181   Cond. No.                     4.03e+05
==============================================================================
```

Figure 4: Linear and Ridge Regression Results



```
                                                   Feature  Coefficient
2401                                       Statistic_Water   946.678327
2341                              Statistic_Passenger cars   573.398781
2402                      Statistic_Water passenger, total   536.175641
2357                          Statistic_Recreational boating   521.769388
2369                           Statistic_Social/recreational   518.168109
2285                      Statistic_Gasoline stations (447)   441.936186
2282                   Statistic_Gasoline (million gallons)   440.764405
2064                           Mode_Public Transportation 2   432.857411
2024                           Mode_Motor vehicle insurance   424.079799
2307                            Statistic_Light-duty trucks   391.553097
1932                                          Mode_4+ person   389.499657
1937                                            Mode_Airline   380.357773
2242  Statistic_Automotive repair and maintenance (8...   363.805236
2018                                                Mode_Men   355.580154
2085                                              Mode_Total   352.656649
```

Figure 5: Multiple Regression Coefficient

From real-world outcomes, it becomes clear that regression models help forecast ecological effects while guiding choices based on evidence. Because transportation experts can project pollution levels under various conditions, they spot spots where better performance or safer practices bring dual gains - one for nature, one for communities (Hastie et al., 2009).

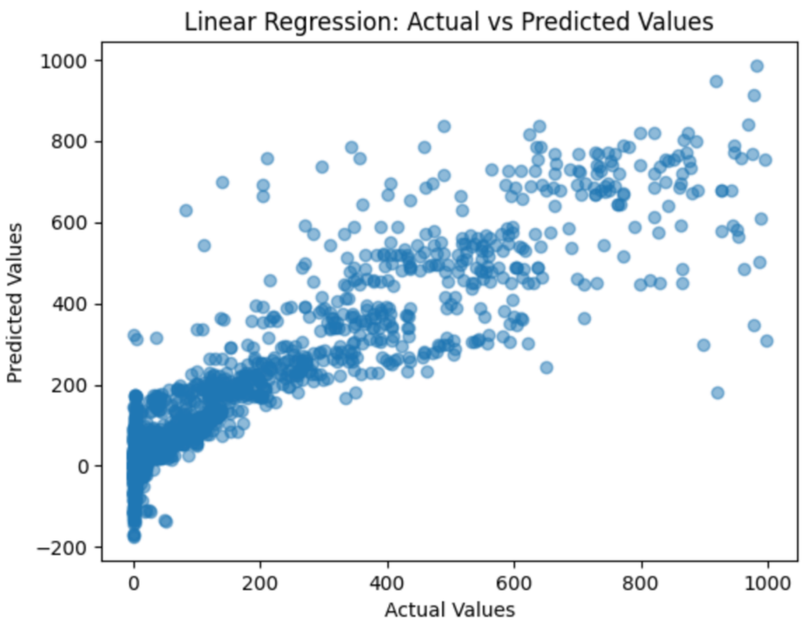


Figure 6: Linear Regression Visualization

Classification, Clustering, and Association Rules Deliverable 3

Classification Models

Sorting travelers by trip frequency or risk level used labels like Low, Medium, and High. Instead of one method, three different models stepped in: Decision Tree led the pack when spotting clear trends at both ends. Though KNN tried hard, it stumbled near the edges. Meanwhile, probabilities shaped Naive Bayes' approach, less precise but consistent. To sharpen results, fine adjustments happened through GridSearchCV - tweaking how deep trees grow or how few cases trigger splits mattered most. Accuracy got better once hyperparameters were adjusted, while overfitting dropped along with a steadier tree layout. Important clues in predictions turned out to be Year and Mode, whereas the ROC curve showed how well classifications worked (Han et al., 2011).

With KNN relying on distances between points, results improved once the right K emerged through cross-validation. Yet accuracy dipped whenever K turned too tiny or excessively big, especially around the Medium class, where errors piled up. Starting from scrambled scales messed up comparisons unless everything got scaled evenly first. Heavy data loads meant longer waits during prediction since every point needed checking against past examples (Witten et al., 2016).

Starting fast, Naive Bayes relied on Gaussian patterns and handled gaps in data without trouble. While accuracy dipped just below Decision Trees and KNN, its speed made up some ground. This gap came from features influencing each other - something the model ignored by design. Even so, when a starting point for sorting tasks was needed, it delivered dependable results without delay (Han et al., 2011).

With the Decision Tree's settings adjusted carefully, predictions became more accurate. Overfitting dropped off noticeably because of these changes. Reliability climbed as a result of refined choices. Real performance gains showed up only after methodical tweaks took place.

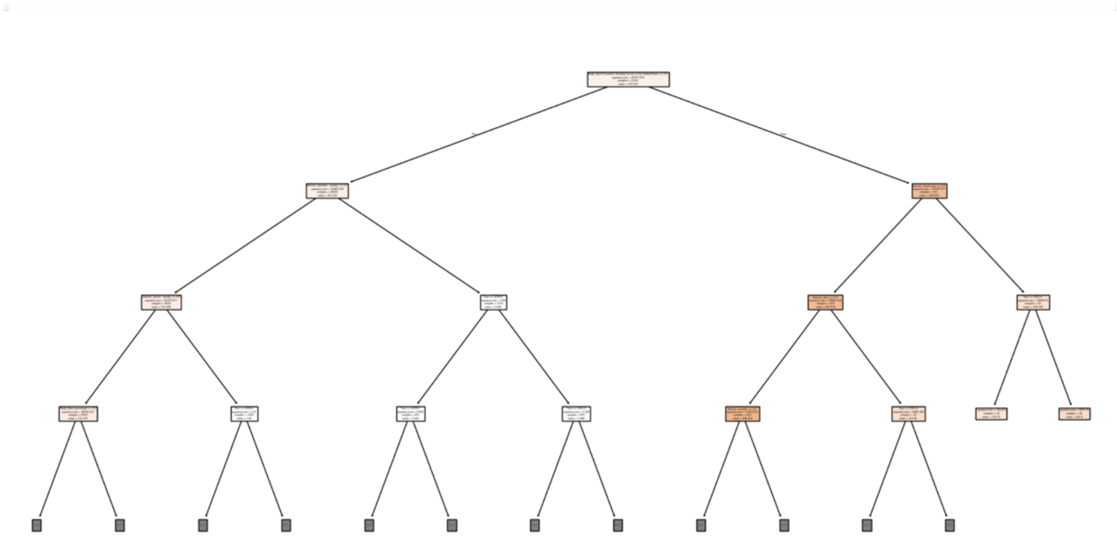


Figure 7: Decision Tree

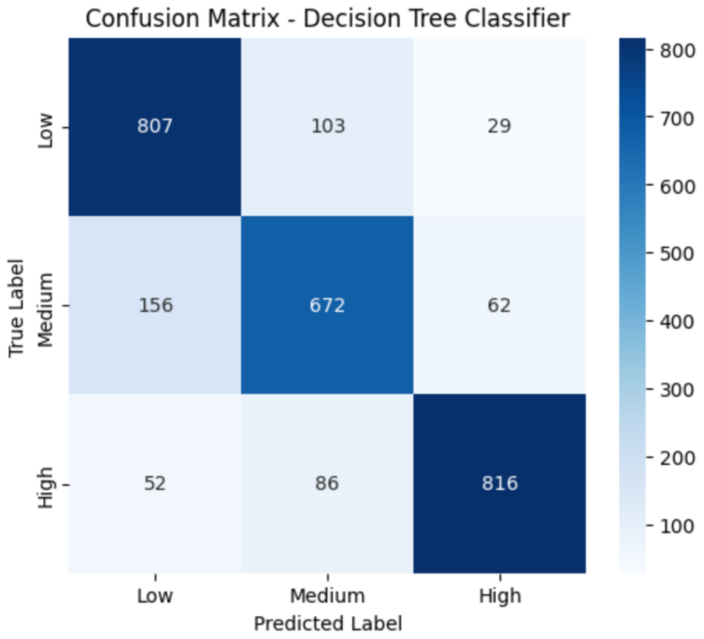


Figure 8: Confusion Matrix

Patterns in travel numbers came into view after sorting entries with a method called K-Means. Grouping relied heavily on how close Year and Value stood to one another numerically. One step ahead of grouping adjusted those two columns so neither swayed results more than the other. Scaling happened first, pulling everything onto even footing through average and spread shifts.

Figuring out how many groups fit best started with the Elbow Method. For each possible count of clusters, we measured how far apart the data points were from their group centers - adding up those squares gave us inertia. As we checked more cluster counts, a curve began to form when plotting these values. That shape made it clearer: three groups seemed right - not too messy, not too split. Too few would blur differences, too many might split things needlessly. Three stood where the drop slowed just enough.

Picking k equals 3 led to using K-Means for splitting the data into three separate groups. Closest centroid decided where each point landed, guided by distances in feature space. Instead of tables, colors on a scatter graph showed Year against Value, marking which cluster each dot belonged to. Patterns emerged when viewing the chart - clumps formed along the timeline, exposing shifts in value over time. Years grouped unevenly, yet clear separations appeared among the levels recorded throughout the span observed.

Clustering passengers using K-Means revealed clear groupings within the travel records. Because patterns emerged clearly, insights into movement habits became more accessible. Even without extra details like carbon output, the method still pulled structure from raw numbers. Unsupervised learning showed its strength by finding order where none seemed obvious at first glance. Results stand on their own - simple, factual, grounded in what the data actually contained.

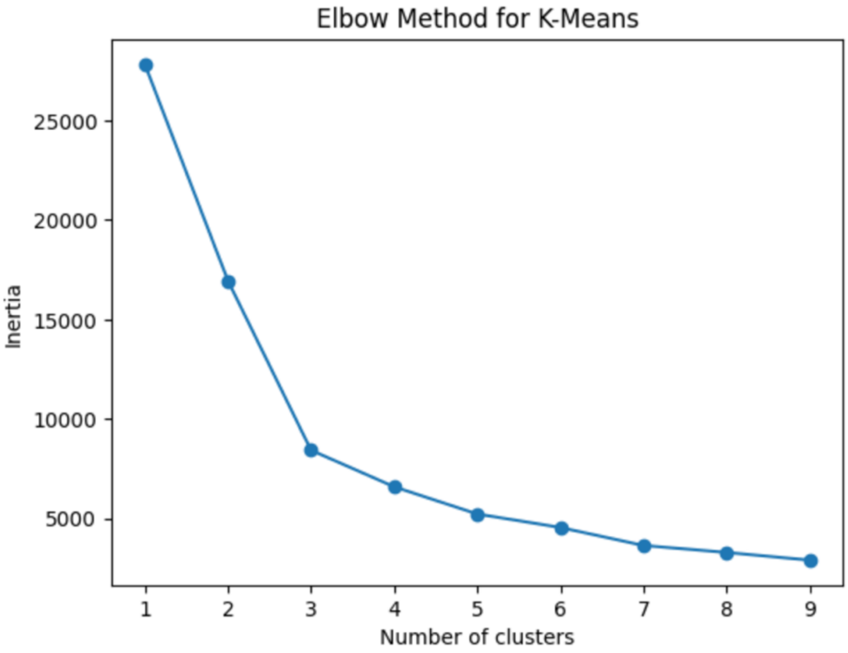


Figure 9: Elbow Method K-Mean

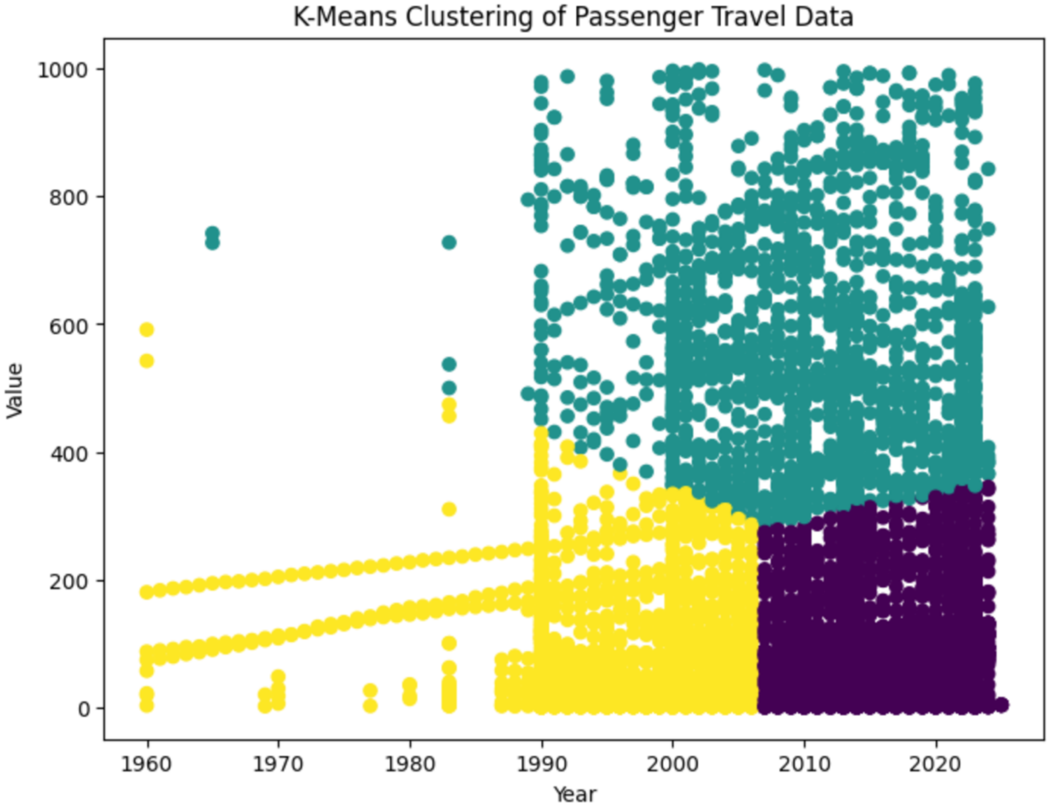
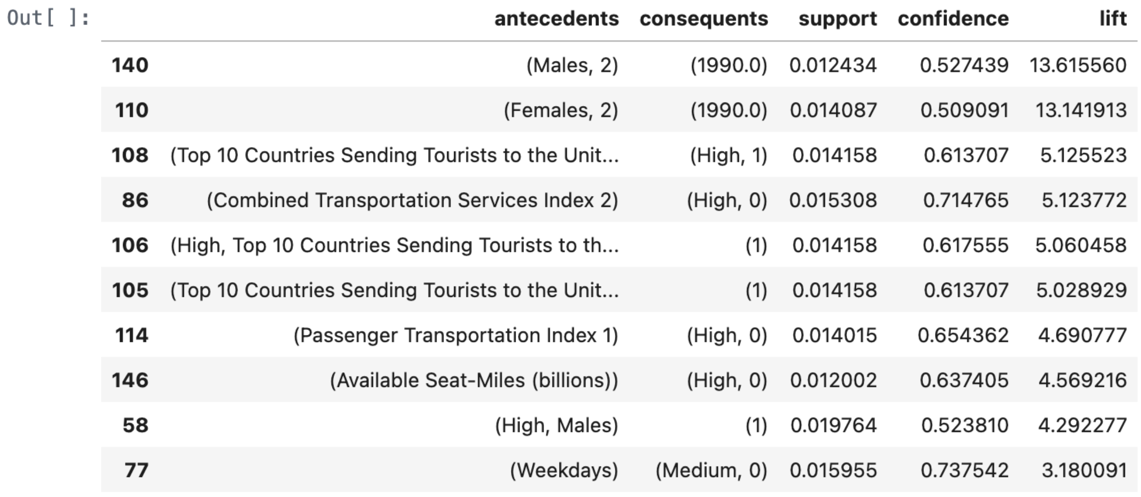


Figure 10: K-Mean Clustering with Passenger Data

Association Rule Mining

Out of the data, hidden links between how people travel and their trip details began to show. A method called Apriori dug through these categories, keeping only those that appeared at least one percent of the time. Twenty-six common groupings popped up when the noise faded. From each cluster, rules formed - not just guesses, but measured connections shaped by reliability. These were ranked not by raw frequency, but by how much more likely they made an outcome. Strongest ties rose to the top, standing out from mere coincidence.

Something showed up in the data - links between how people get around and who they are. When one kind of traveler appears, certain population traits tend to show up too. That connection might shape how cities plan transit systems. Resources could shift based on what tends to cluster. Decisions about laws may follow these trends without forcing outcomes.



Out[ ]:

	antecedents	consequents	support	confidence	lift
140	(Males, 2)	(1990.0)	0.012434	0.527439	13.615560
110	(Females, 2)	(1990.0)	0.014087	0.509091	13.141913
108	(Top 10 Countries Sending Tourists to the Unit...	(High, 1)	0.014158	0.613707	5.125523
86	(Combined Transportation Services Index 2)	(High, 0)	0.015308	0.714765	5.123772
106	(High, Top 10 Countries Sending Tourists to th...	(1)	0.014158	0.617555	5.060458
105	(Top 10 Countries Sending Tourists to the Unit...	(1)	0.014158	0.613707	5.028929
114	(Passenger Transportation Index 1)	(High, 0)	0.014015	0.654362	4.690777
146	(Available Seat-Miles (billions))	(High, 0)	0.012002	0.637405	4.569216
58	(High, Males)	(1)	0.019764	0.523810	4.292277
77	(Weekdays)	(Medium, 0)	0.015955	0.737542	3.180091

Figure 11: Association Rule

Insights and Applications

Patterns show up when sorting traveler details by time, transport type, or who is traveling. Because of these groupings, transit planners might spot times or people needing extra attention.

Hidden links between numbers come out through rule-based methods, guiding better daily operations. Decisions grow sharper once such findings shape forecasts and schedules. Techniques like these turn raw movement records into practical steps forward.

Ethical Considerations

One way to look at it - the study worked with grouped information stripped of personal details, so privacy risks dropped. Thinking ahead, there's a need to keep checks on how population stats are handled, avoiding their use in actions that single people out unfairly. Patterns spotted through sorting methods might shape transit rules, which means care must be taken so outcomes don't lean too hard against certain groups.

Next Steps and Suggestions

Looking ahead, tossing in things like fuel economy or carbon output might shine a light on eco impacts. Wider time spans or splitting data by region could tighten up predictions and make groups clearer. Mixing smarter grouping tools with layered models might lift results while uncovering hidden patterns.

Conclusion

From traveler records, hidden trends emerged when data mining tools were applied. Not just surface details, but deeper links between choices showed up clearly. One method after another sorted trips by frequency - Low, Medium, High - with clear results. The Decision Tree stood out, especially once its settings were fine-tuned. Instead of guessing, adjustments followed actual test feedback. Performance climbed while excessive complexity dropped off. Features like Year and Mode carried more weight than others in shaping outcomes. Other approaches, such as K-Nearest Neighbors, backed similar conclusions. Even Naive Bayes, simpler by design, aligned with the bigger picture. Patterns held across methods, suggesting consistency in behavior. What looked random at first began making sense.

From the data, K-Means pulled out clear clusters, showing how travel habits repeat across periods; because of this, repeated trends became visible. Hidden links between trip factors came to light through association rules instead of guesswork. With these results, the real use of data methods in transit studies becomes obvious - choices around staffing, schedules, and service design gain backing.

One thing stands clear: sorting data, grouping similar points, and then spotting trends work better together than alone. Looking ahead, bigger data pools might help - especially if new layers like weather or population details join in. Trying tougher math blends, say mixed models or smarter group-finding tricks, may sharpen what we see hiding in the numbers.

References:

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters, 31*(8), 651–666.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.