ASSIGNMENT 3:

PROJECT:

EXPLORATION AND ANALYSIS OF THE BIKE THEFT IN, OTTAWA.

Yosufi Hasibullah- 041012318

Program: COMPUTER ENGINNERING TECHNOLOGY: COMPUTING SCIENCE

College: ALGONQUIN COLLEGE, OTTAWA

Course: CST8390: BI AND DATA ANALYSIS

Contents

INTRODUCTION

To understand the dynamics and trends associated with bicycle thefts, an analysis was conducted on a dataset comprising 8,134 incidents. This dataset provides valuable insights into various aspects of bike thefts, including the value of stolen bicycles, temporal distribution of thefts, preferred locations for thefts, commonly targeted bicycle makes, and popular bicycle colors among thieves.

BUSINESS UNDERSTANDING

The dataset obtained from the Government of Ottawa contains detailed information on reported bike theft incidents within the city. It includes attributes such as the location of theft, timestamps indicating the time and date of each incident, descriptions of stolen bikes including make, model, and color, circumstances surrounding the theft, police response actions, outcomes of each incident, basic demographics of victims, and contextual factors such as weather conditions and time of day. Analyzing this dataset will yield valuable insights into the patterns, trends, and risk factors associated with bike theft in Ottawa. Leveraging advanced analytics and predictive modeling techniques on this dataset can help stakeholders develop targeted interventions and preventive measures to reduce bike theft occurrences and enhance bike security across the city.

PROJECT PLAN

CLASSIFICATION TASK

Objective: Develop a predictive model to classify incidents of bike thefts based on various attributes (e.g., location, time of day, type of bike) to predict the likelihood of bike theft occurrences.

Methodology:

- **Data Preprocessing:** Clean the dataset by handling missing values, encoding categorical variables, and normalizing numerical features.
- **Feature Selection:** Identify and select relevant features that significantly influence the outcome of a bike theft incident.
- **Model Selection:** Experiment with different classification algorithms Decision Trees, to find the best performing model.

- **Model Training and Evaluation:** Split the data into training and testing sets. Train the model on the training set and evaluate its performance on the testing set using metrics like accuracy,

CLUSTERING TASK

Objective: Group bike theft incidents into clusters based on similarities in their attributes (e.g., location patterns, time patterns) to identify hotspots or patterns of thefts.

Methodology:

- **Data Preprocessing:** Standardize the data to ensure that each feature contributes equally to the distance calculations.
- **Feature Selection:** Choose features relevant to the clustering task, focusing on geographical and temporal aspects.
- **Clustering Algorithm Selection:** Use unsupervised learning algorithms like K-means to identify clusters.
- **Evaluation:** determine the optimal number of clusters and evaluate the clustering performance.

OUTLIER DETECTION TASK

Objective: Identify outliers in the bike theft dataset to detect unusual patterns or anomalies that might indicate sophisticated theft operations or data recording errors.

Methodology:

- **Data Preprocessing:** Clean the dataset and normalize the features to ensure a uniform scale.

- **Outlier Detection Methods:** Apply outlier detection techniques such as Isolation Forest, Local Outlier Factor (LOF to identify anomalies in the dataset.
- **Analysis of Outliers:** Investigate the detected outliers to distinguish between potential theft rings or errors in data collection.

2.0 DATA UNDERSTANDING

2.1 DATA COLLECTION

- The Bike Theft dataset was acquired from the city of Ottawa website containing detailed information on bike theft in Ottawa.

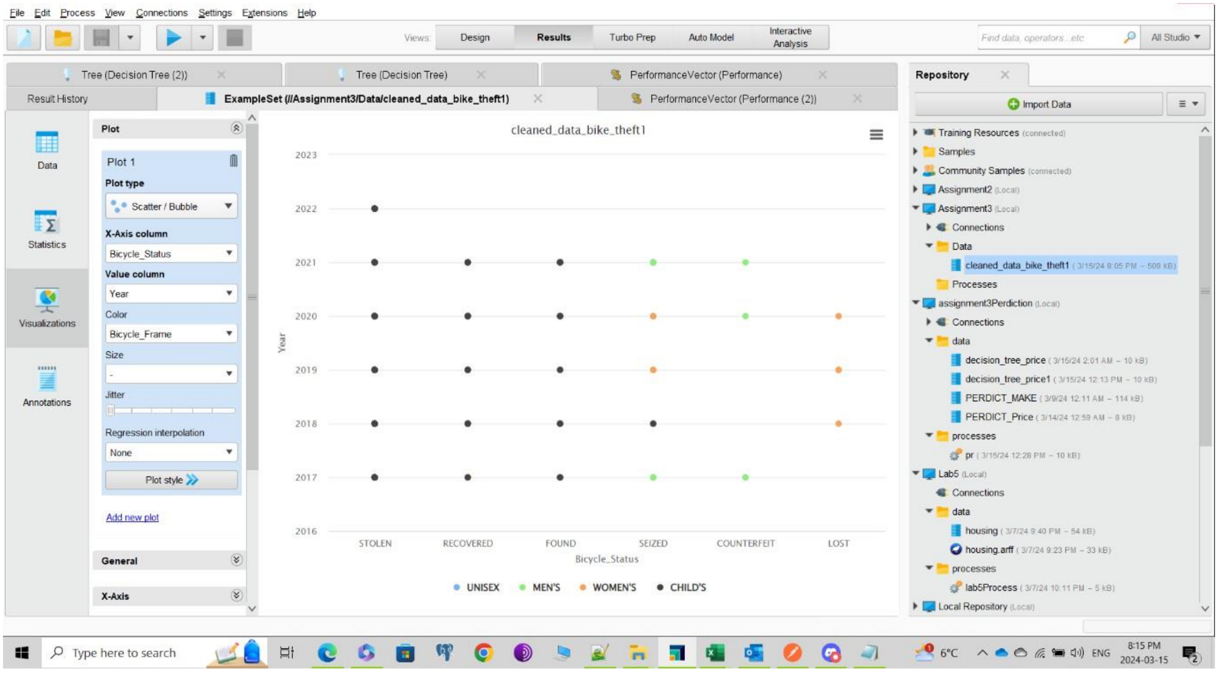
2.2 DATA DESCRIPTION

The dataset comprises records for **8,709 bikes** that were stolen, each described by **28** attributes/columns, including Bicycle Make, Bicycle Model, Bicycle frame, Bicycle Colour, Bicycle Speeds, Ottawa Neighborhood, Census Tract etc.

There was a total of **28** Attributes (Columns) from the dataset comprising of **Numerical attributes**, **Nominal attributes,** and **String attributes**.

We also included 4 new attributes because of the project focus, one of those attributes was the **Month** column which was done based on the occurrence feature. The other attributes created were**, Band popularity** (based on bike make**), Price_affordablity (**categorical version of price for decision tree), **Diff_Rep_Occu** (difference between report day and occurrence day).

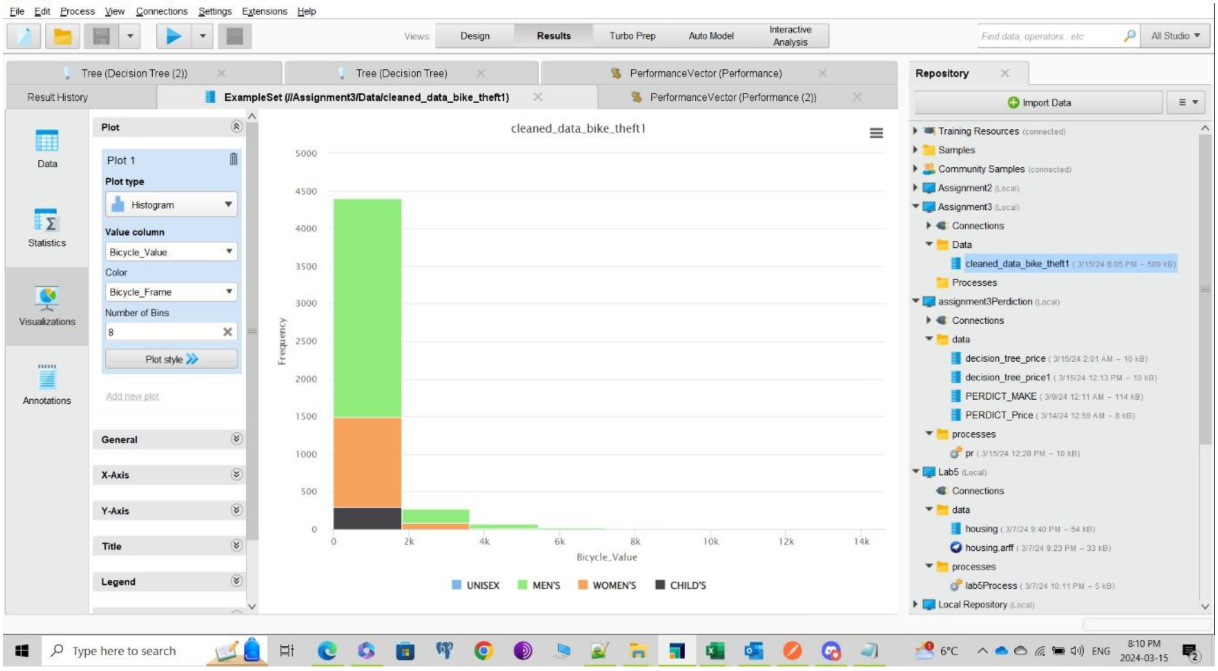
2.3 EXPLORE DATA


ExampleSet (//Assignment3/Data/cleaned_data_bike_theft1)

Plot 1
Plot type: Scatter / Bubble
X-Axis column: Bicycle_Status
Value column: Year
Color: Bicycle_Frame
Size: -
Jitter
Regression interpolation: None

Bikes Status (stolen, recovered, seized, found), Year and Bicycle Frame:

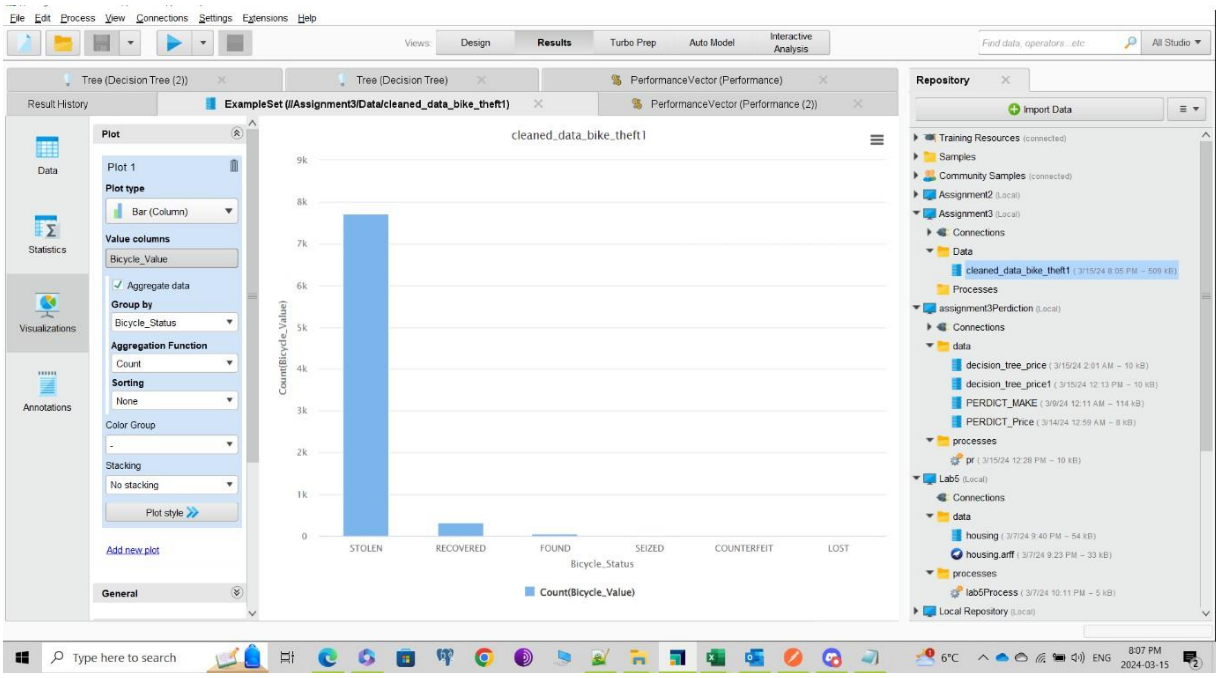
The figure above illustrates the relationships between Bicycle Status, Years, and Bicycle Frame. It is evident that the majority of stolen bikes have child frames, and they are also the most frequently recovered type.


ExampleSet (//Assignment3/Data/cleaned_data_bike_theft1)

Plot 1
Plot type: Histogram
Value column: Bicycle_Value
Color: Bicycle_Frame
Number of Bins: 8

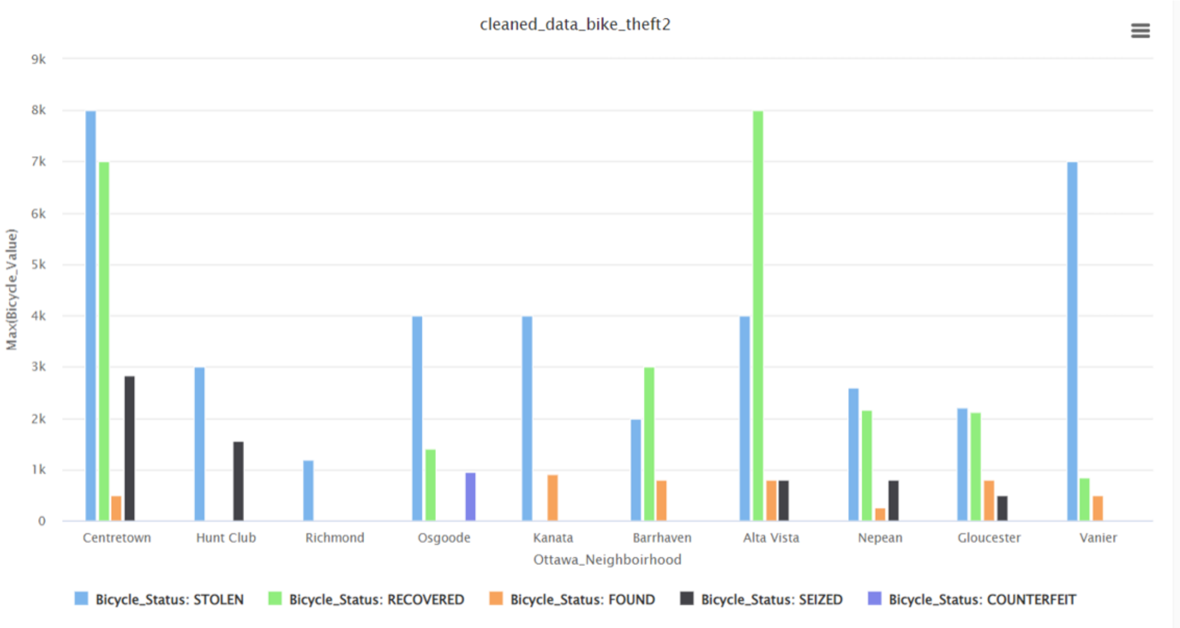
Frequency of bicycle frame and their values:

From the above figure we can see that most of bikes that range from $0 up to $2000 have men’s frame followed by women and children.


Plot 1
Plot type: Bar (Column)
Value columns: Bicycle_Value
☑ Aggregate data
Group by: Bicycle_Status
Aggregation Function: Count
Sorting: None
Color Group: -
Stacking: No stacking

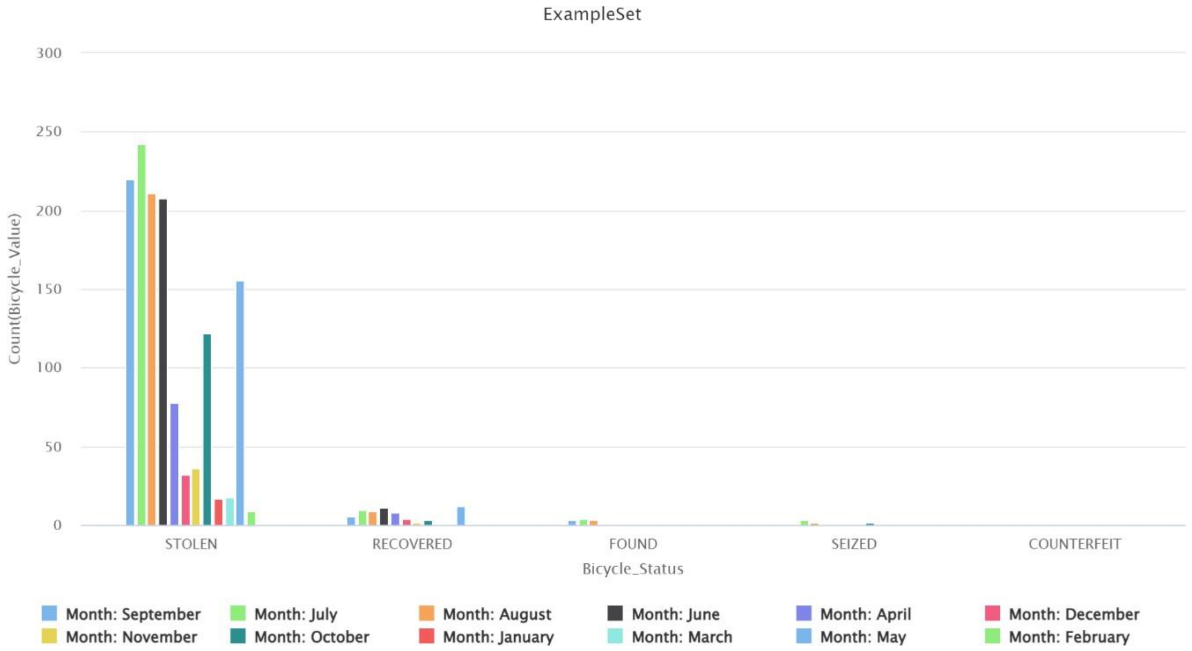
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Tree (Decision Tree (2))
Tree (Decision Tree)
PerformanceVector (Performance)
Result History
ExampleSet (//Assignment3/Data/cleaned_data_bike_theft1)
PerformanceVector (Performance (2))
Data
Statistics
Visualizations
Annotations
Plot
Plot style
Add new plot
General
cleaned_data_bike_theft1
9k
8k
7k
6k
5k
4k
3k
2k
1k
0
Count(Bicycle_Value)
STOLEN
RECOVERED
FOUND
SEIZED
COUNTERFEIT
LOST
Bicycle_Status
Count(Bicycle_Value)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Assignment2 (Local)
Assignment3 (Local)
Connections
Data
cleaned_data_bike_theft1 (3/15/24 8:05 PM – 509 kB)
Processes
assignment3Perdiction (Local)
Connections
data
decision_tree_price (3/15/24 2:01 AM – 10 kB)
decision_tree_price1 (3/15/24 12:13 PM – 10 kB)
PERDICT_MAKE (3/9/24 12:11 AM – 114 kB)
PERDICT_Price (3/14/24 12:59 AM – 8 kB)
processes
pr (3/15/24 12:28 PM – 10 kB)
Lab5 (Local)
Connections
data
housing (3/7/24 9:40 PM – 54 kB)
housing.arff (3/7/24 9:23 PM – 33 kB)
processes
lab5Process (3/7/24 10:11 PM – 5 kB)
Local Repository (Local)
Type here to search
6°C
ENG
8:07 PM
2024-03-15

Figure above Bicycle Status and Their Value



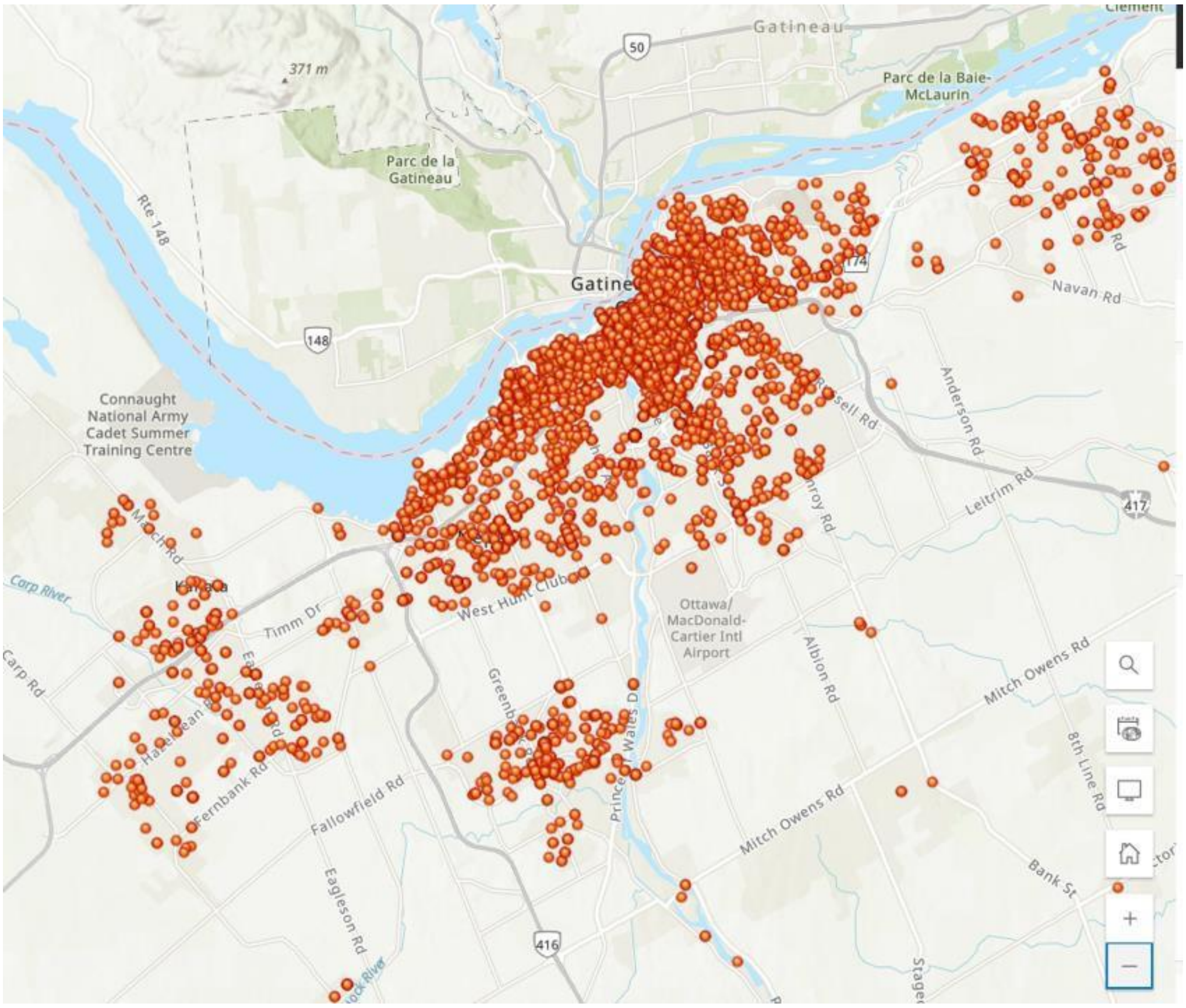
Ottawa Neighborhood, Bicycle Status and Bike Value analysis:

Figure 4 illustrates the relationships between Bicycle Status, Ottawa neighborhoods, and Bicycle Value. We observe that the majority of bike thefts (Stolen Bicycles) occur in the Centretown and Vanier neighborhoods. The stolen bikes range in value from as low as $0 to as high as $8,000. Most bike recoveries, on the other hand, take place in the Alta Vista neighborhood, which encompasses areas such as Glebe, Carleton, Riverside, Billings, and other neighboring areas.

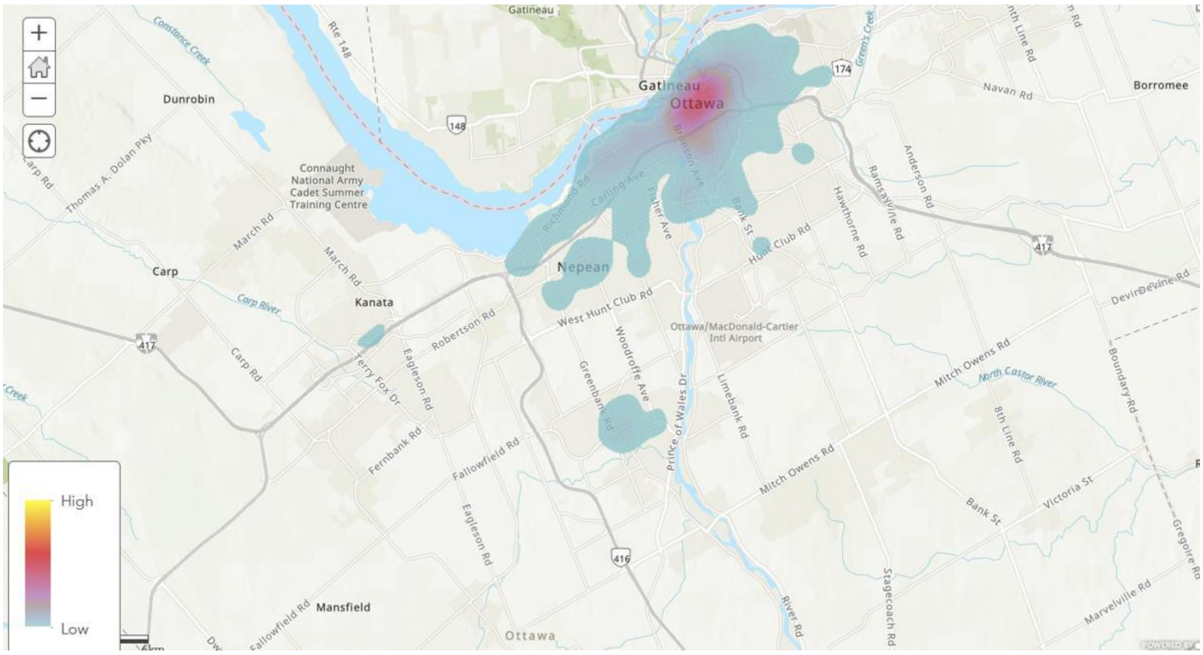


Exploring Bicycle Status, Brand Popularity, and Month:

From the figure above, we observe that bike theft is most prevalent during the months of May, June, July, August, and September. This trend extends from the beginning of summer to early fall, with June experiencing the highest incidence of theft and February the lowest. We conclude that June marks the start of summer break, prompting more people to ride bikes and leave them outdoors, thereby increasing the likelihood of theft. Conversely, February, being the peak of winter, witness fewer instances of theft as bikes are typically kept indoors in buildings or homes. We also conclude that most recovered bikes happen in the months of May, June, and July.



X and Y Coordinates of Bike Theft



Heat Map of Bike Theft

As we can see in the two figures above Centertown is where the most bike theft happened followed by Alta Visa and Vanier.

2.4 Data Quality

The dataset contains 8,709 entries and 28 columns. The dataset includes a mix of float, integer, and object (string) data types. Columns cover a variety of attributes related to bike thefts, such as coordinates (**X**, **Y**), identification numbers (**FID**, **ID**), timing of the report (**Reported_D**, **Reported_T**, **Reported_W**), occurrence details (**Occurrence**, **Occurren_1**, **Occurren_2**), bike specifics (**Bicycle_St**, **Bicycle_Va**, **Bicycle_Ma**, etc.), and location information (**Ottawa_Nei**, **Census_Tra**, etc.).

Certain columns, like **Bicycle_St** (status), **Bicycle_Va** (value), **Bicycle_Ma** (make), **Bicycle_Mo** (model), **Bicycle_Ty** (type), **Bicycle_Fr** (frame), and **Bicycle_Co** (color), have missing values. This could affect analyses that rely on complete data.

Columns like **Reported_T**, **Occurrence**, and **Occurren_1** contain dates but are formatted as object (string) types and include some seemingly incorrect dates (e.g., year 1900 or 1905 in **Reported_T**, which might be placeholders or errors).

Location-related information is detailed, including **coordinates**, **neighborhood names**, and **census tract** data. However, ensuring accuracy and consistency across these columns is essential for geographical analyses.

3.0 DATA PREPARATION

3.1 Select Data

During my selection process, we had to evaluate and see how useful each attribute can be for the Bike theft dataset, and how it could lead to finding usual conclusions about the situation.

We finally decided on, **X, Y**, **Year**, **Occurrence, Occurrence Day, Location_T, Bicycle_Status, Bike_value, Bicycle_Make, Bicycle_Type, Bicycle_Frame, Bicycle_Color, Ottawa_Neighboirhood, Month**

Attributes like **Location Coordinates (X, Y)** are Essential for geographical clustering. Attributes like **Year** was used to observe trends over time.

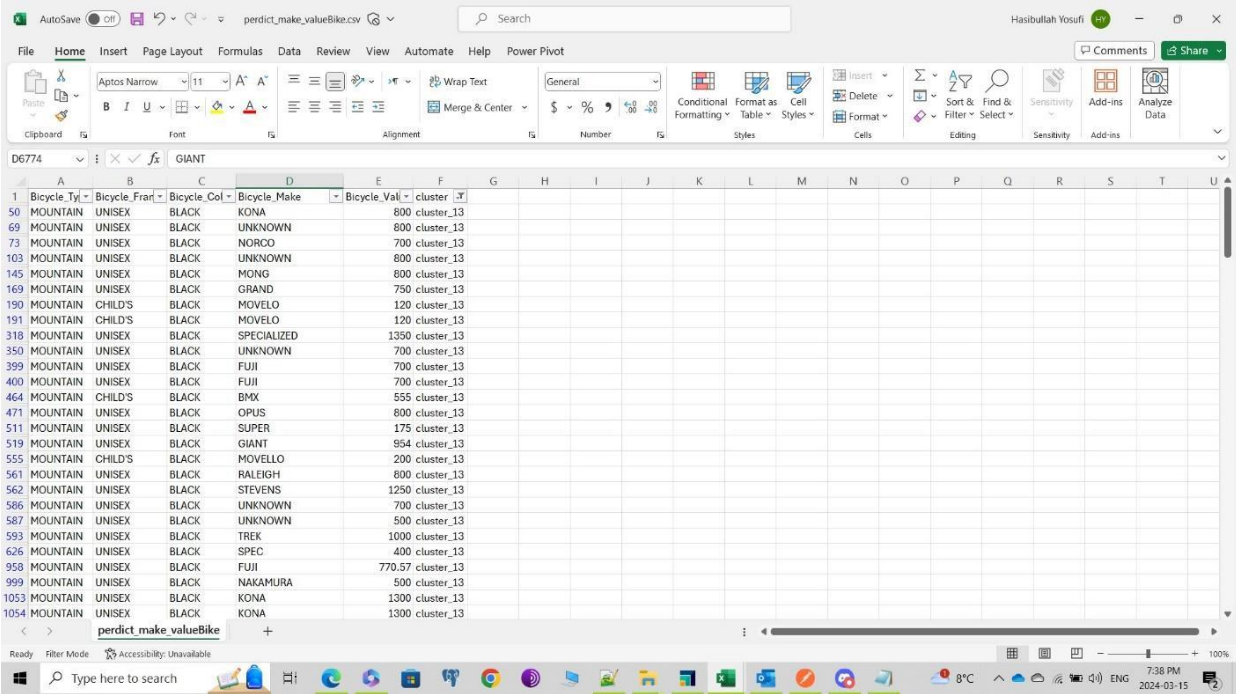
Figure 1 Attributes Selected from the list.

3.1 Clean Data

The feature that contained missing values in the above data set is bike_value which we are going to fill based on decision tree. In addition to that, missing values(N/A)sfor **bicycle_Frame** were predicted based on **Bicycle_Make**, **Bicycle_Type** and **Bicycle_Color.**

In our approach to refining the dataset, particularly in resolving ambiguities related to the 'bicycle_color' attribute, we employed a strategic interpretation technique. For instance, abbreviations such as 'DBL' and 'DGR' were deciphered as 'Dark Blue' and 'Dark Green,' respectively. These interpretations were then thoughtfully categorized under broader color groups, 'Blue' and 'Green,' to ensure consistency and clarity within our dataset.

To address and accurately predict missing values (denoted as N/A) for 'bicycle_make,' we implemented the K-Means clustering algorithm, setting the number of clusters (K) to 40. This sophisticated method allowed us to analyze and group similar values across multiple features—'bicycle_type,' 'bicycle_frame,' 'bicycle_make,' and 'bicycle_value'—based on their inherent characteristics. By clustering these attributes, we were able to identify patterns and relationships within the data, enabling us to predict and fill in the missing values for 'bicycle_make' with a high degree of accuracy. To handle missing values for the Range_Price (Bike_Value) we used decision tree to predict the Range_Price. For the decision tree we considered attributes of Range_Price (label), Brand_Popularity, Bicycle_Frame and, Bicycle_Type. We filled the missing values of numerical Bike_Value based on the predicted intervals of categorical values of Range_Price.



perdict_make_valueBike.csv

D6774 GIANT

	A	B	C	D	E	F
1	Bicycle_Ty	Bicycle_Frar	Bicycle_Col	Bicycle_Make	Bicycle_Val	cluster
50	MOUNTAIN	UNISEX	BLACK	KONA	800	cluster_13
69	MOUNTAIN	UNISEX	BLACK	UNKNOWN	800	cluster_13
73	MOUNTAIN	UNISEX	BLACK	NORCO	700	cluster_13
103	MOUNTAIN	UNISEX	BLACK	UNKNOWN	800	cluster_13
145	MOUNTAIN	UNISEX	BLACK	MONG	800	cluster_13
169	MOUNTAIN	UNISEX	BLACK	GRAND	750	cluster_13
190	MOUNTAIN	CHILD'S	BLACK	MOVELO	120	cluster_13
191	MOUNTAIN	CHILD'S	BLACK	MOVELO	120	cluster_13
318	MOUNTAIN	UNISEX	BLACK	SPECIALIZED	1350	cluster_13
350	MOUNTAIN	UNISEX	BLACK	UNKNOWN	700	cluster_13
399	MOUNTAIN	UNISEX	BLACK	FUJI	700	cluster_13
400	MOUNTAIN	UNISEX	BLACK	FUJI	700	cluster_13
464	MOUNTAIN	CHILD'S	BLACK	BMX	555	cluster_13
471	MOUNTAIN	UNISEX	BLACK	OPUS	800	cluster_13
511	MOUNTAIN	UNISEX	BLACK	SUPER	175	cluster_13
519	MOUNTAIN	UNISEX	BLACK	GIANT	954	cluster_13
555	MOUNTAIN	CHILD'S	BLACK	MOVELLO	200	cluster_13
561	MOUNTAIN	UNISEX	BLACK	RALEIGH	800	cluster_13
562	MOUNTAIN	UNISEX	BLACK	STEVENS	1250	cluster_13
586	MOUNTAIN	UNISEX	BLACK	UNKNOWN	700	cluster_13
587	MOUNTAIN	UNISEX	BLACK	UNKNOWN	500	cluster_13
593	MOUNTAIN	UNISEX	BLACK	TREK	1000	cluster_13
626	MOUNTAIN	UNISEX	BLACK	SPEC	400	cluster_13
958	MOUNTAIN	UNISEX	BLACK	FUJI	770.57	cluster_13
999	MOUNTAIN	UNISEX	BLACK	NAKAMURA	500	cluster_13
1053	MOUNTAIN	UNISEX	BLACK	KONA	1300	cluster_13
1054	MOUNTAIN	UNISEX	BLACK	KONA	1300	cluster_13

perdict_make_valueBike

Ready Filter Mode Accessibility: Unavailable

3.2 CONSTRUCT DATA

Creating new attributes such as "**Month**," "**Brand_Popularity** ", **"Diff_Rep_Occu ",** "**Weekend_or_WeekDay** ", "**Range_Price**" for the Bike dataset is a strategic approach in data preparation that enhances the analysis and predictive modeling.

“**Month**” as an attribute consolidates the information about the month which the theft occurred by converting the occurrence date format to pick only the month makes it easier to predict what month a lot of bikes were stolen, was it higher during the summer or winter?

“**Brand_Popularity**” as an attribute was used to determine what bikes were attractive to buyers, which gives criminals a hint as to what to steal. This attribute was based on bike_make.

“**Range_Price”** was an attribute used to see how much people would spend on bikes, and it was a category for decision tree.

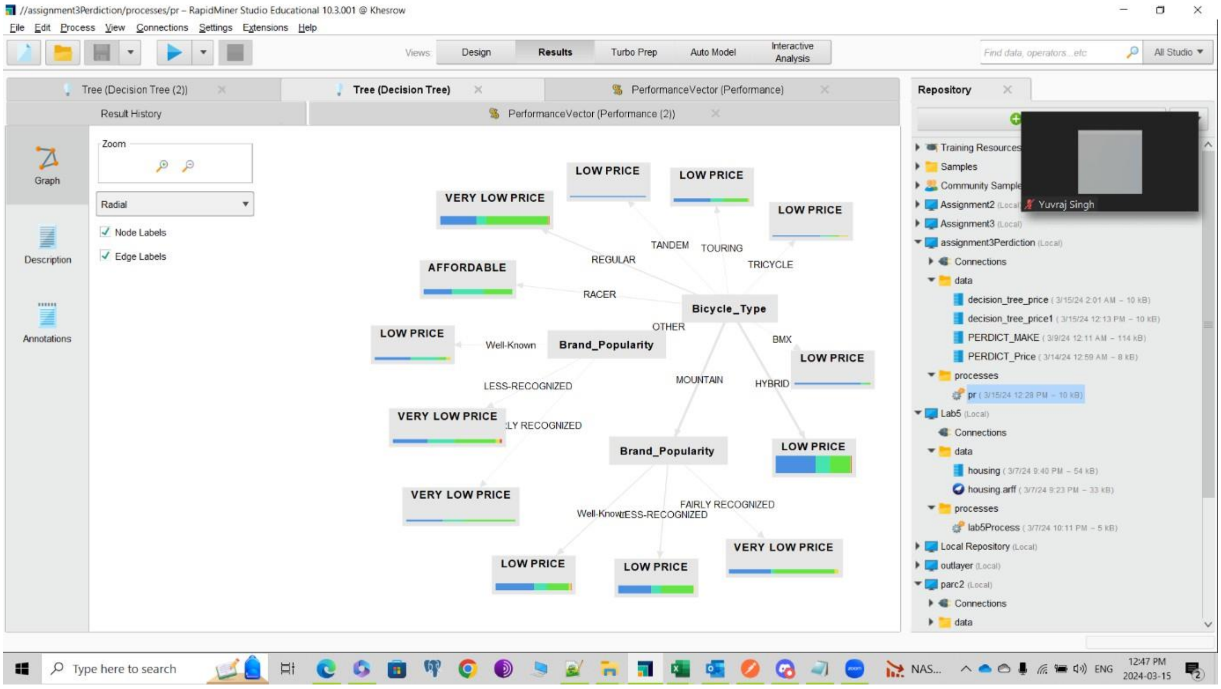


Figure 2 Attributes for Decision Tree

3.3 Integrate Data

- I didn’t need to combine data from multiple sources.

3.4 Format Data

With regards to formatting data, we put the data in numerical data, categorial, date and time, and other attributes. We also converted occurrence into month to extract the month the theft occurred.

Integrating numeric, **categorical**, and **datetime** data to support diverse analyses, including economic impact, temporal trends, and spatial dynamics. Numeric data like the value of stolen bikes, year of theft, and geographical coordinates enable economic, temporal, and spatial studies.

Categorical details such as theft location, bike characteristics, and occurrence specifics allow for profiling high-risk areas and understanding theft patterns. The inclusion of neighborhood data aids in regional analysis, while temporal categorizations and qualitative assessments on brand popularity and price range offer insights into seasonal trends and bike desirability.

The dataset also features datetime information and reporting delay metrics, enhancing its utility for detailed theft pattern analysis and response time evaluation, making it invaluable for addressing bike theft effectively.

4.0 MODELLING

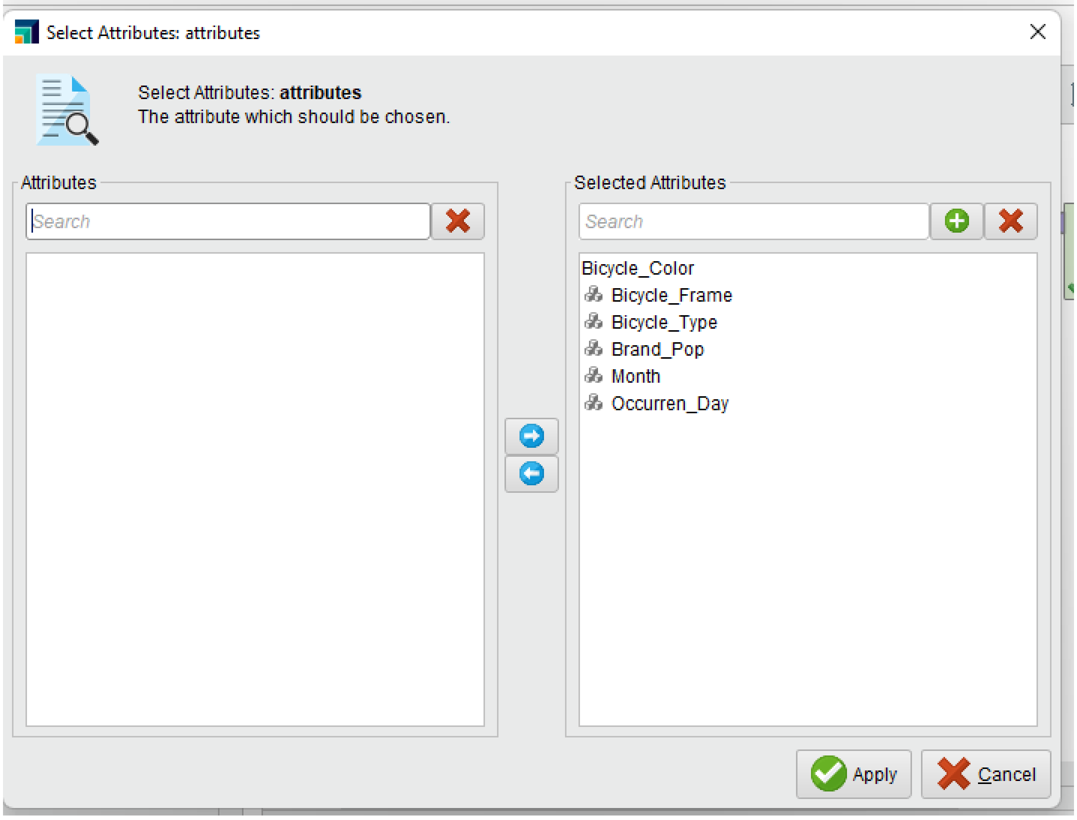
4.1 Select Modelling Techniques

The models used for training are as follows:

- K Means Clustering (Elbow Method)
- KNN Classifier
- Decision Tree, Random Forest
- LOF Outlier Detector

K Means Clustering – conducted by Ken

- In preparing the dataset for k-Means clustering, I ensured all features were numeric. I numerically encoded categorical variables such as "Bicycle_Frame” and 'Month', “Bicycle_Type”,” Month”,”Occuren_Day” allowing for an inclusive analysis that considered all relevant passenger attributes.



Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.

Selected Attributes:
Bicycle_Color
Bicycle_Frame
Bicycle_Type
Brand_Pop
Month
Occurren_Day

Figure 3 Attributes that were converted to for normalization

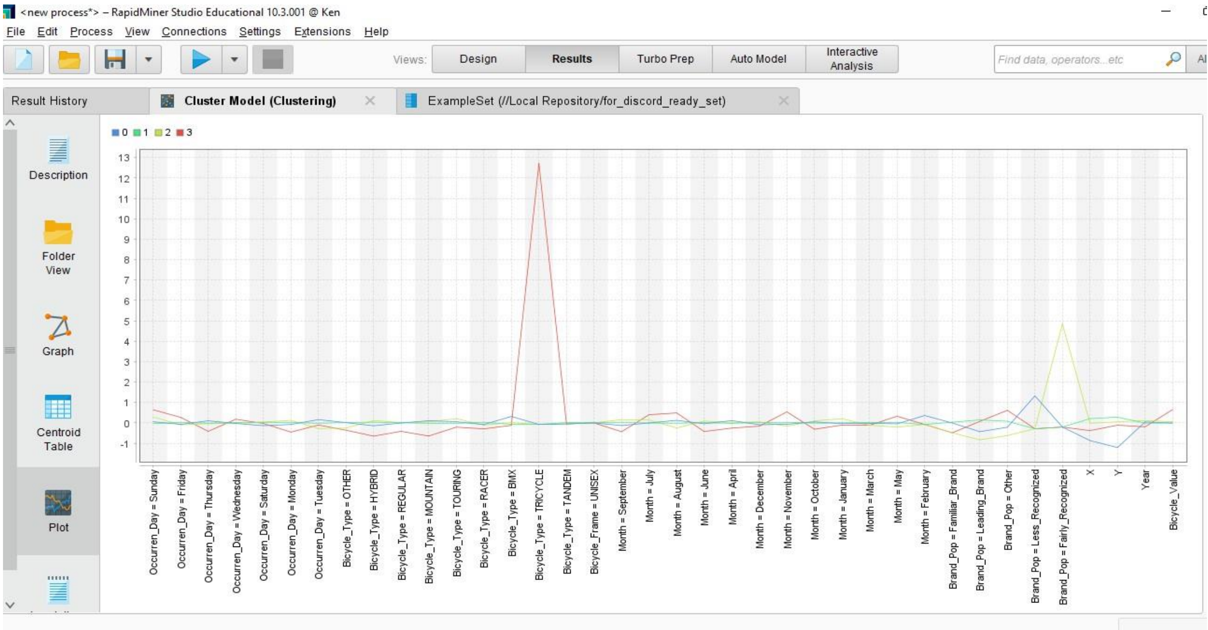
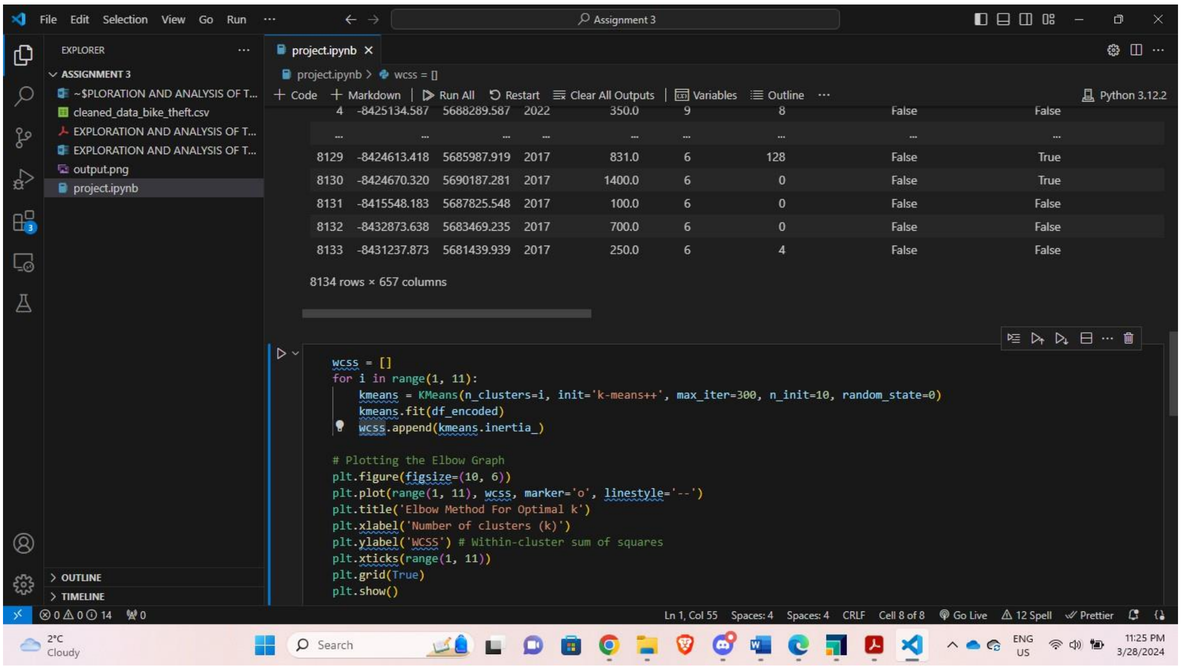


Figure 4 Plot for K Mean cluster.

- I standardized all features to have a mean of 0 and a standard deviation of 1. This essential step guaranteed that variances across features did not skew the clustering process, facilitating an equitable comparison and grouping based on Euclidean distance.
- I used a K value of 4, because when I did the elbow method.
- Performed elbow method with python to generate my graph for K value.


File Edit Selection View Go Run ...
Assignment 3
EXPLORER
ASSIGNMENT 3
~$PLORATION AND ANALYSIS OF T...
cleaned_data_bike_theft.csv
EXPLORATION AND ANALYSIS OF T...
EXPLORATION AND ANALYSIS OF T...
output.png
project.ipynb
OUTLINE
TIMELINE
project.ipynb
project.ipynb > wcss = []
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ...
Python 3.12.2
Ln 1, Col 55 Spaces: 4 Spaces: 4 CRLF Cell 8 of 8 Go Live 12 Spell Prettier
Search
2°C Cloudy
ENG US
11:25 PM 3/28/2024

```
   4  -8425134.587  5688289.587  2022   350.0  9    8  False  False
 ...           ...          ...   ...     ...  ...  ...    ...    ...
8129  -8424613.418  5685987.919  2017   831.0  6  128  False  True
8130  -8424670.320  5690187.281  2017  1400.0  6    0  False  True
8131  -8415548.183  5687825.548  2017   100.0  6    0  False  False
8132  -8432873.638  5683469.235  2017   700.0  6    0  False  False
8133  -8431237.873  5681439.939  2017   250.0  6    4  False  False

8134 rows × 657 columns
```

```python
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
    kmeans.fit(df_encoded)
    wcss.append(kmeans.inertia_)

# Plotting the Elbow Graph
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method For Optimal k')
plt.xlabel('Number of clusters (k)')
plt.ylabel('WCSS') # Within-cluster sum of squares
plt.xticks(range(1, 11))
plt.grid(True)
plt.show()
```

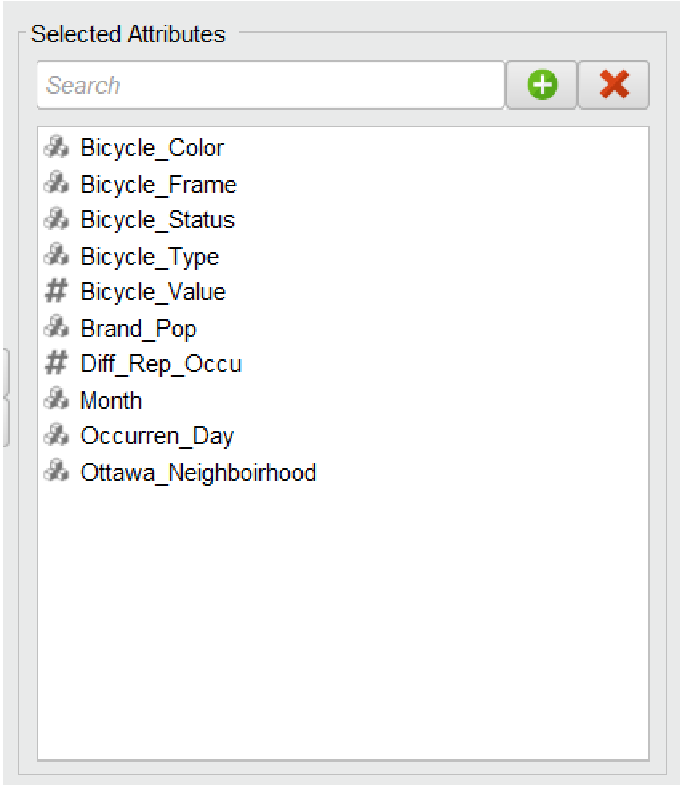
Figure 5 Python Process for my K value using the attributes I selected.

4.1.1 K-NN CLASSIFICATION

Occuanc_, Month, Bicycle_Type

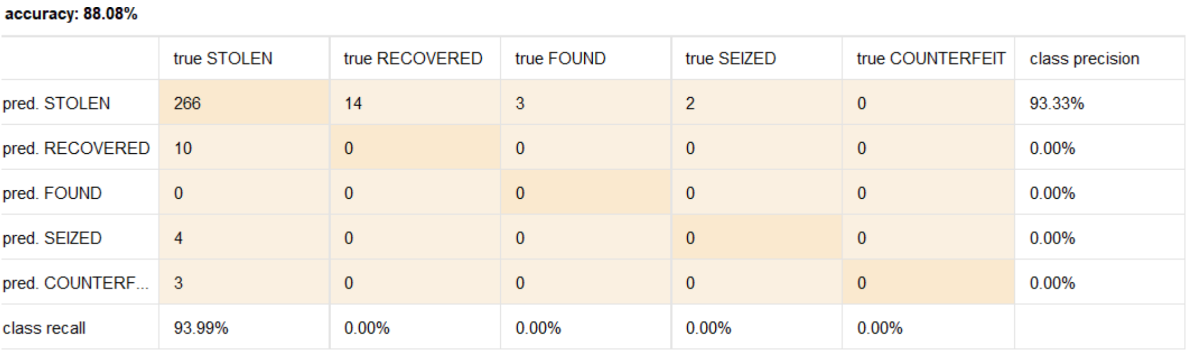
For the K-NN prediction we used Bicycle Status as the class label for prediction. We used attributes Bicycle_Frame, Bicycle_Status, Bicycle_Value, Brand_Pop (Brand Popularity five unique values) , Diff_Rep_Occu (difference between reported date and occurrence date), Month and, Weekend_or_WeekDay. We achieved the highest accuracy up to 93.74% from cross validation with the numbers of fold equals to 10 and 93.71% from the percentage split with the number of neighbors equal to 7.

K-NN Attributes:


Selected Attributes

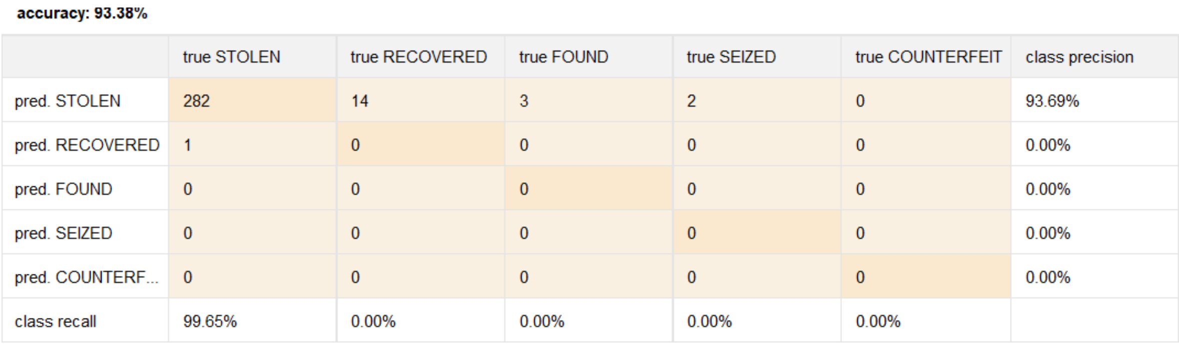
- Bicycle_Color
- Bicycle_Frame
- Bicycle_Status
- Bicycle_Type
- Bicycle_Value
- Brand_Pop
- Diff_Rep_Occu
- Month
- Occurren_Day
- Ottawa_Neighboirhood

K=2


accuracy: 88.08%

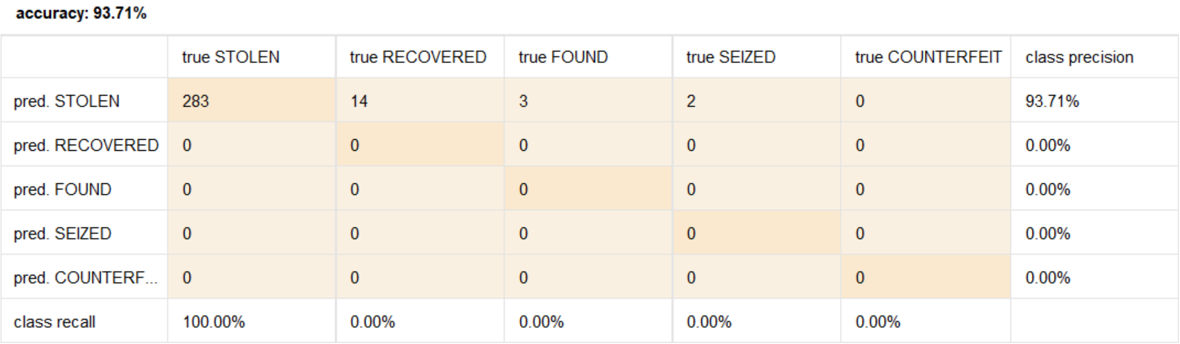
	true STOLEN	true RECOVERED	true FOUND	true SEIZED	true COUNTERFEIT	class precision
pred. STOLEN	266	14	3	2	0	93.33%
pred. RECOVERED	10	0	0	0	0	0.00%
pred. FOUND	0	0	0	0	0	0.00%
pred. SEIZED	4	0	0	0	0	0.00%
pred. COUNTERF...	3	0	0	0	0	0.00%
class recall	93.99%	0.00%	0.00%	0.00%	0.00%	

K=4


accuracy: 93.38%

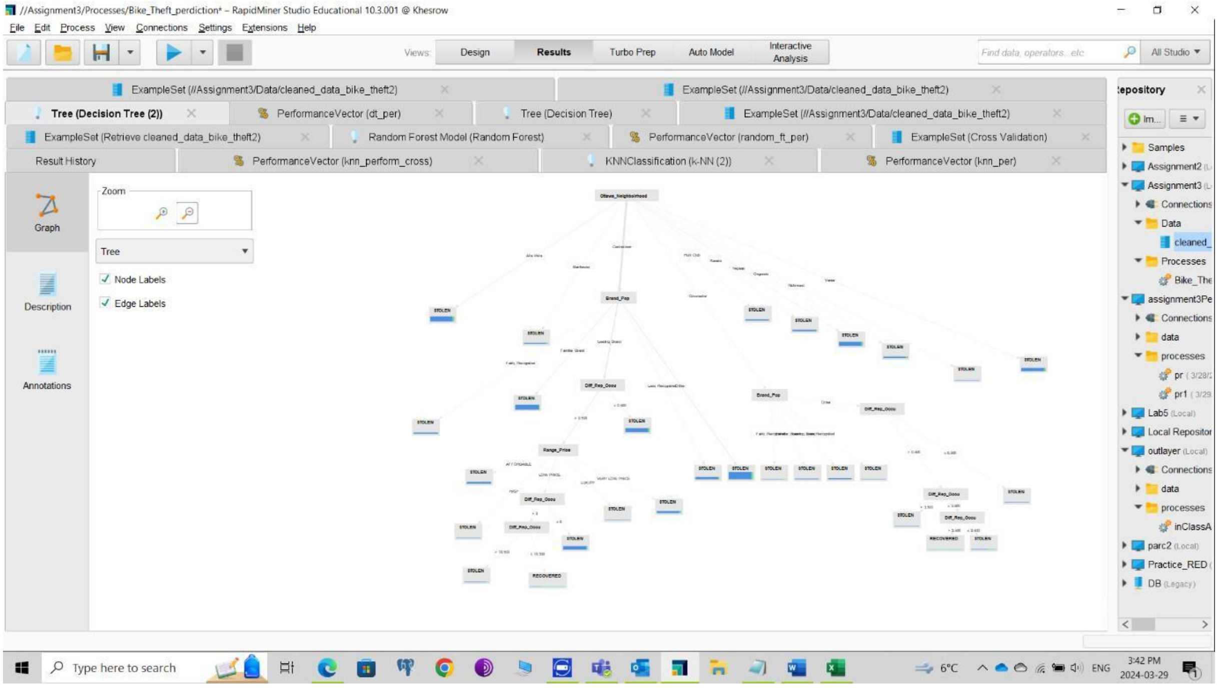
	true STOLEN	true RECOVERED	true FOUND	true SEIZED	true COUNTERFEIT	class precision
pred. STOLEN	282	14	3	2	0	93.69%
pred. RECOVERED	1	0	0	0	0	0.00%
pred. FOUND	0	0	0	0	0	0.00%
pred. SEIZED	0	0	0	0	0	0.00%
pred. COUNTERF...	0	0	0	0	0	0.00%
class recall	99.65%	0.00%	0.00%	0.00%	0.00%	

K=7



accuracy: 93.71%

	true STOLEN	true RECOVERED	true FOUND	true SEIZED	true COUNTERFEIT	class precision
pred. STOLEN	283	14	3	2	0	93.71%
pred. RECOVERED	0	0	0	0	0	0.00%
pred. FOUND	0	0	0	0	0	0.00%
pred. SEIZED	0	0	0	0	0	0.00%
pred. COUNTERF...	0	0	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	0.00%	0.00%	

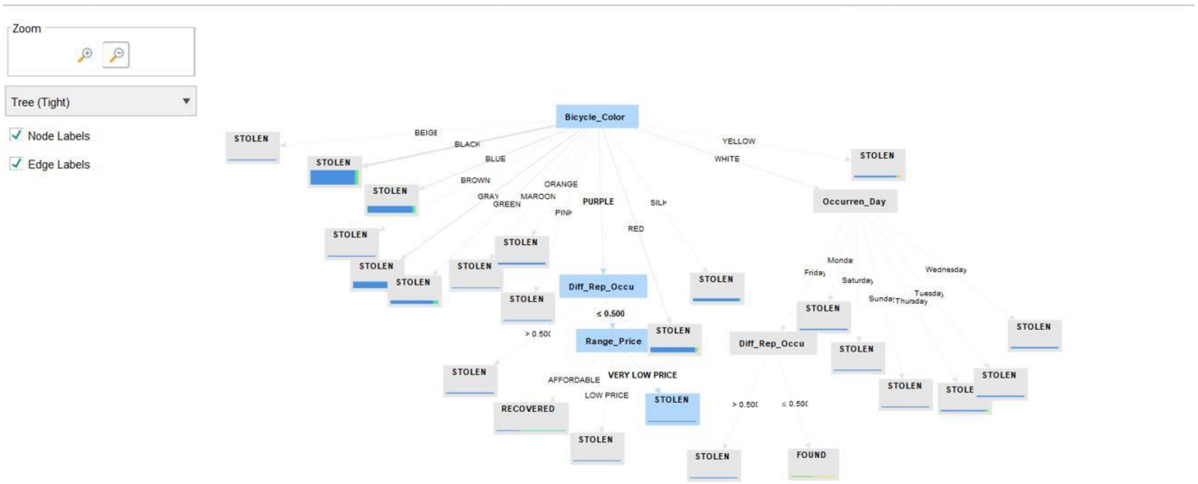


//Assignment3/Processes/Bike_Theft_perdiction* – RapidMiner Studio Educational 10.3.001 @ Khesrow
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
ExampleSet (//Assignment3/Data/cleaned_data_bike_theft2)
ExampleSet (//Assignment3/Data/cleaned_data_bike_theft2)
Tree (Decision Tree (2))
PerformanceVector (dt_per)
Tree (Decision Tree)
ExampleSet (//Assignment3/Data/cleaned_data_bike_theft2)
ExampleSet (Retrieve cleaned_data_bike_theft2)
Random Forest Model (Random Forest)
PerformanceVector (random_ft_per)
ExampleSet (Cross Validation)
Result History
PerformanceVector (knn_perform_cross)
KNNClassification (k-NN (2))
PerformanceVector (knn_per)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
Ottawa_Neighborhood
Brand_Pop
Diff_Rep_Occu
Range_Price
STOLEN
RECOVERED
AFFORDABLE
LOW PRICE
VERY LOW PRICE
LUXURY
HIGH
Repository
Samples
Assignment2
Assignment3
Connections
Data
cleaned_
Processes
Bike_The
assignment3Pe
Connections
data
processes
pr (3/28/
pr1 (3/28
Lab5 (Local)
Local Repositor
outlayer (Local)
Connections
data
processes
inClassA
parc2 (Local)
Practice_RED
DB (Legacy)
Type here to search
6°C
ENG
3:42 PM
2024-03-29

Tree Analysis:

Based on the tree above, we observe that if a bike theft occurs in Centertown and the bike brand is among the leading brands, most expensive bikes tend to get stolen. However, bikes with lower prices, ranging from $400 to $1000, have a higher chance of being recovered.

Root equals Bicycle_Color nodes (Diff_Rep_Occu, Range_Price, Occurance_Day)



Zoom

Tree (Tight)

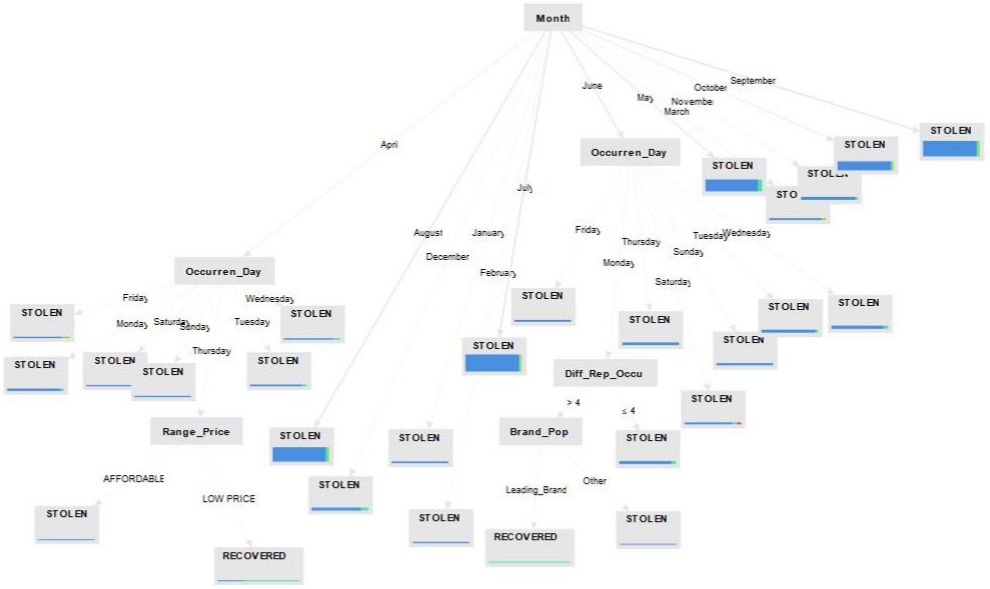
☑ Node Labels

☑ Edge Labels

Tree Analysis:

From the above tree we get the observation that the recovered bikes have purple and white color with affordable price range and the bikes that had white color were found within one day.

Root equals Month nodes (Occurance_Day, Range_Price, Diff_Rep_Occu, Brand_Pop)

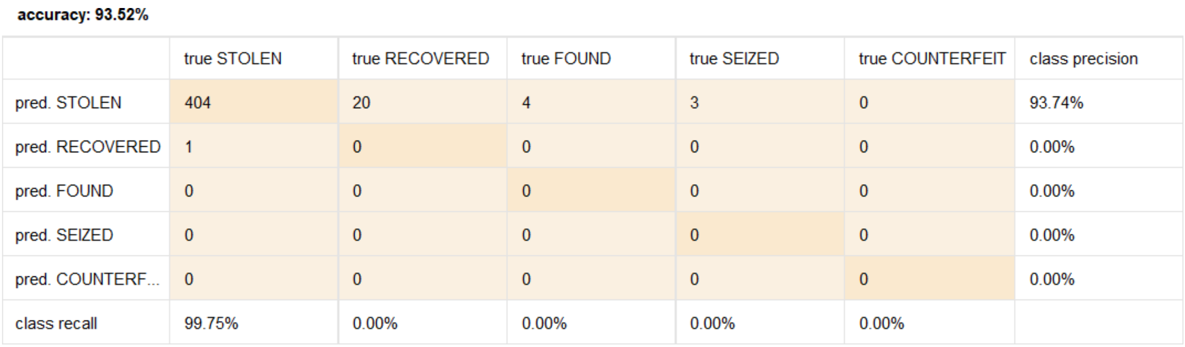


Tree Analysis:

From the tree above, we observe that the majority of recovered bikes were reported during the spring and early summer months, particularly in June and April. These recovered bikes typically had lower prices and belonged to leading brands such as Giant, Specialized, Trek, Cannondale, Scott, Bianchi, Schwinn, GT, and Felt.

4.1.2 RANDOM FOREST

We achieved an accuracy of up to 93.52% using Random Forest with 100 trees and utilizing information gain. The features used for the Random Forest included Bicycle_Color, Bicycle_Frame, Bicycle_Status, Bicycle_Type, Brand_Pop, Diff_Rep_Occu, Month, Occurrence_Day, Ottawa_Neighborhood, and Range_Price.



accuracy: 93.52%

	true STOLEN	true RECOVERED	true FOUND	true SEIZED	true COUNTERFEIT	class precision
pred. STOLEN	404	20	4	3	0	93.74%
pred. RECOVERED	1	0	0	0	0	0.00%
pred. FOUND	0	0	0	0	0	0.00%
pred. SEIZED	0	0	0	0	0	0.00%
pred. COUNTERF...	0	0	0	0	0	0.00%
class recall	99.75%	0.00%	0.00%	0.00%	0.00%	

C4.5, K-NN Random Forest comparison

Algorithm	Accuracy
K-NN K = 7	93.71
C4.5 depth =5 pruned = True	92.3747

K-NN cross Validation	93.74
Random Forest	93.52

4.1.3 CONCLUSION

Our analysis of bike thefts in Ottawa shows a clear pattern. The most at-risk areas are Centretown, Vanier, Alta Vista, and Gloucester. These thefts are most common in the warmer months, from April to September, likely because more people are cycling during this time. Bikes that are easier to steal are also targeted more often. This includes bikes that are purple, white, or silver, cost between $0 and $2000, and are from popular brands like Giant or Specialized.

4.1.4 LOF OUTLIER DETECTION

The use of Local Outlier Factor (LOF) for outlier detection in my analysis of the bike dataset provided a nuanced approach to identifying anomalies within the passenger data. This method was instrumental in uncovering data points that deviated significantly from the norm, which could potentially influence the overall model performance.

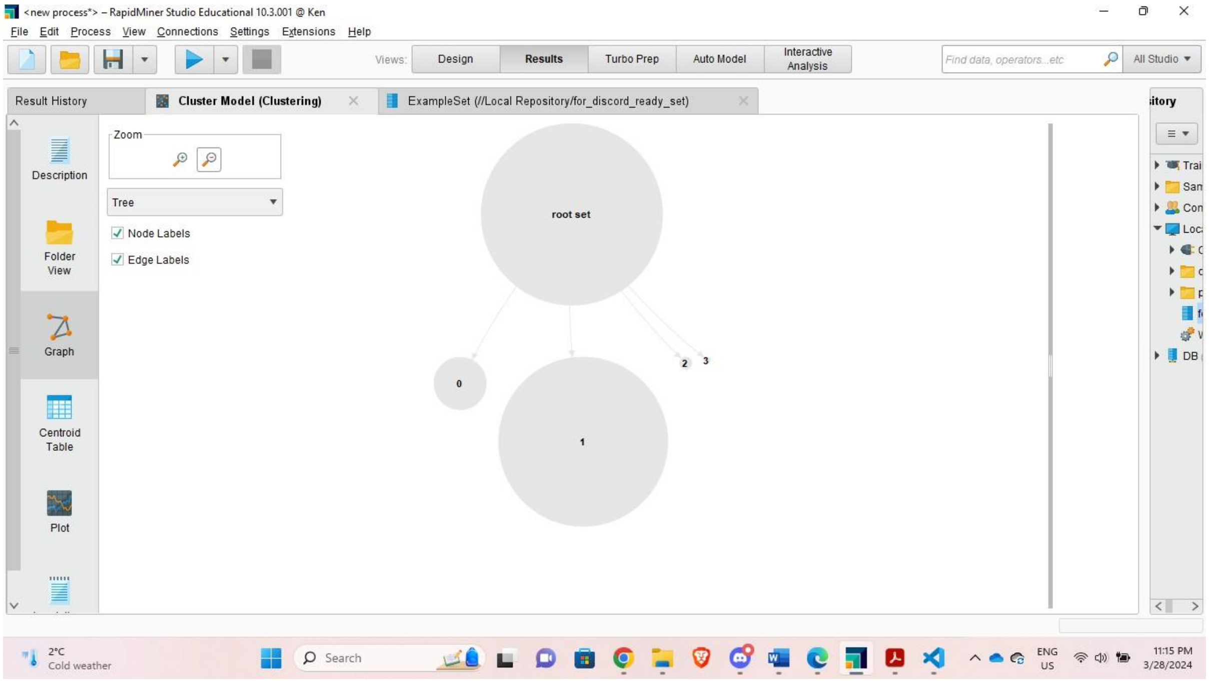
4.1.4 DISTANCE BASED DETECTION

- I also explored the use of distance-based outlier detection techniques. These methods identify outliers by examining the distances between data points, if outliers will have significantly different distances to their neighbors compared to the rest of the data points.

4.2 GENERATE TEST DESIGN

4.2.1 TEST 01 – K MEAN CLUSTER

- All the data was used to generate four clusters.



<new process*> – RapidMiner Studio Educational 10.3.001 @ Ken
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
ExampleSet (//Local Repository/for_discord_ready_set)
Description
Folder View
Graph
Centroid Table
Plot
Zoom
Tree
☑ Node Labels
☑ Edge Labels
root set
0
1
2
3
itory
Trai
San
Con
Loc
DB
2°C
Cold weather
Search
ENG
US
11:15 PM
3/28/2024

Figure 6 cluster for K value 4.

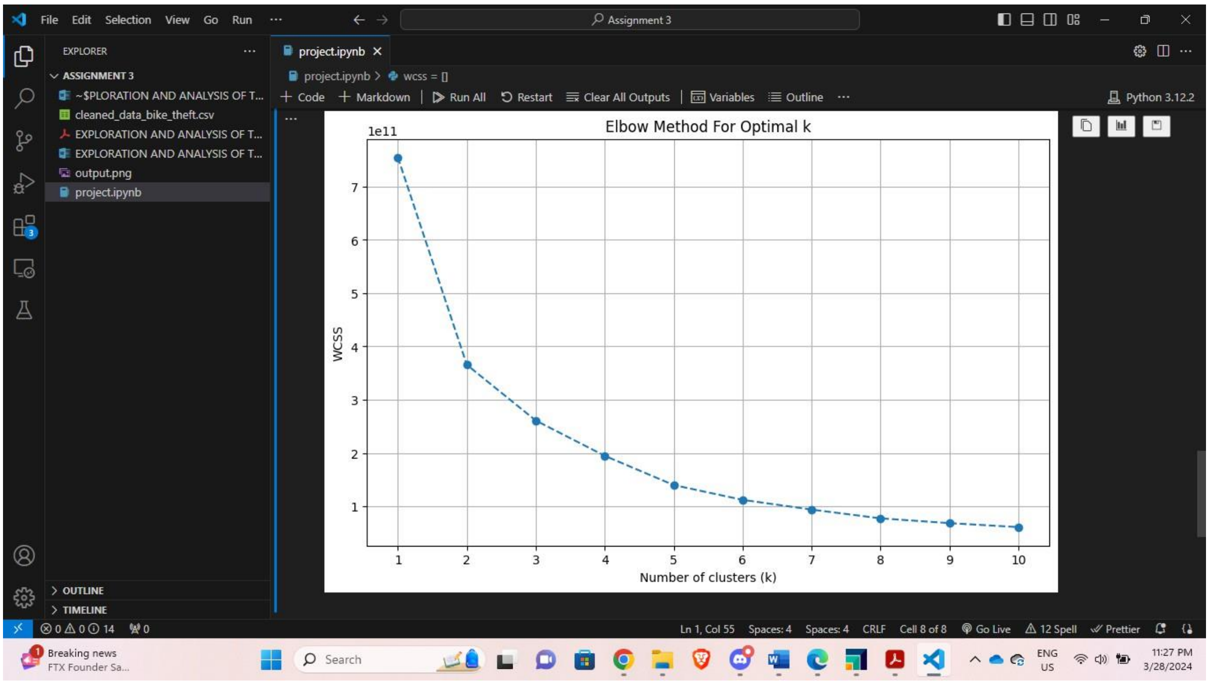


Figure 7 Elbow Plot for clustering.

4.2.2 TEST 02 – DECISION TREE CLASSIFIER

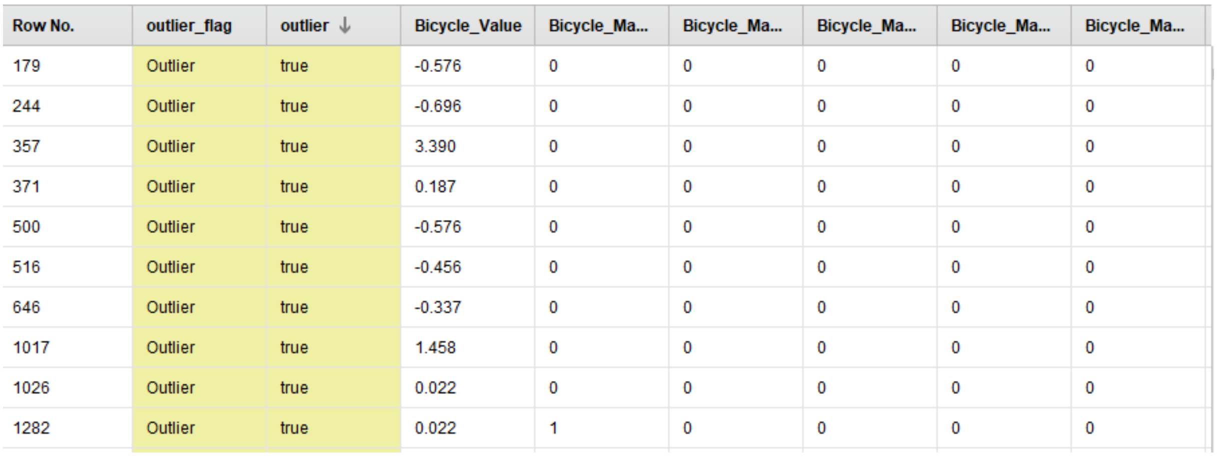
- For the decision tree three main attributes impact the most the tree. They are **Ottawa_Neighbourhood, Bicycle_Color and, Month.** We can get accuracy up to 92.82% using entropy (information gain) if we have **Ottawa_Neighbourhood as root parent and Bicycle_Frame, Bicycle_Status, Brand_Pop, Range_Price, Weekend_Or_WeekDay** as nodes to perdict the bicycle Status.

Figure 8 Decision tree for my results.

4.2.3 TEST 02 – OUTLIER DETECTION

-

  I decided to incorporate Local Outlier Factor (LOF) outlier detection into my analysis of the bike theft dataset to identify and understand anomalies within the bike data.
- I selected the attributes that are relevant to the LOF evaluation. Converted nominal types to numerical and normalized the data.
- I applied the LOF algorithm to the bike theft dataset with a focus on numerical attributes.

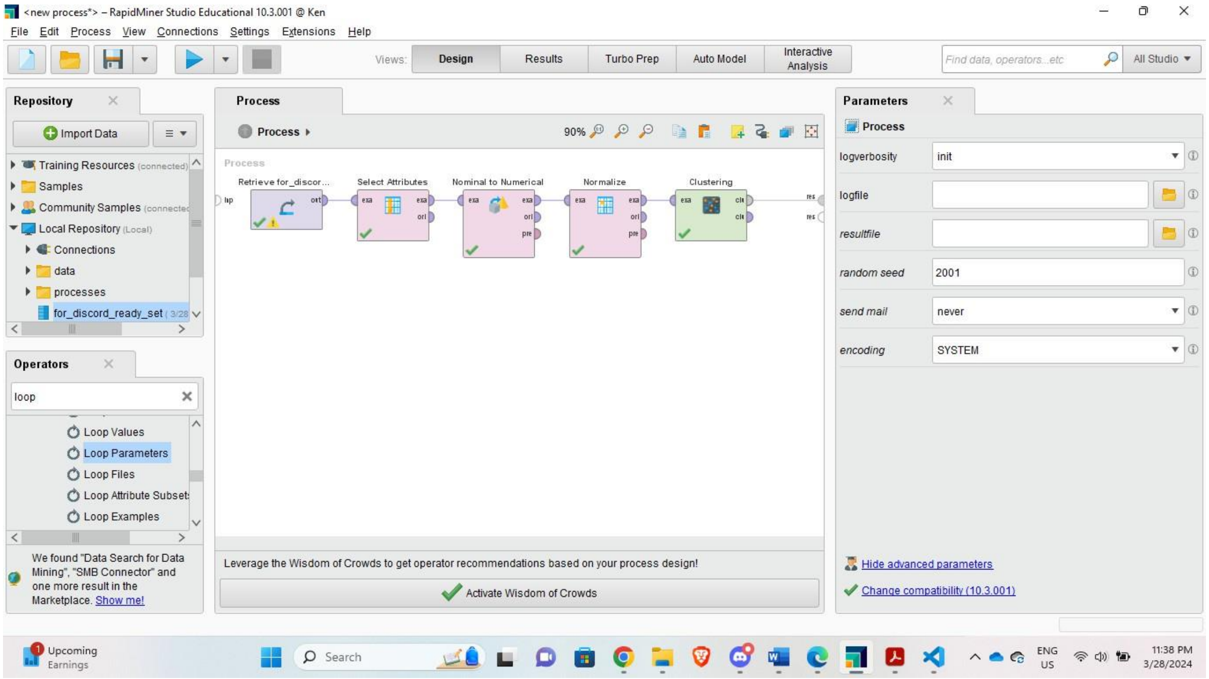


Row No.	outlier_flag	outlier ↓	Bicycle_Value	Bicycle_Ma...	Bicycle_Ma...	Bicycle_Ma...	Bicycle_Ma...	Bicycle_Ma...
179	Outlier	true	-0.576	0	0	0	0	0
244	Outlier	true	-0.696	0	0	0	0	0
357	Outlier	true	3.390	0	0	0	0	0
371	Outlier	true	0.187	0	0	0	0	0
500	Outlier	true	-0.576	0	0	0	0	0
516	Outlier	true	-0.456	0	0	0	0	0
646	Outlier	true	-0.337	0	0	0	0	0
1017	Outlier	true	1.458	0	0	0	0	0
1026	Outlier	true	0.022	0	0	0	0	0
1282	Outlier	true	0.022	1	0	0	0	0

4.3 BUILD MODEL

The Models explained in test design were trained with their respective prepared datset.

We took parts of the models between Decision trees, and the clustering and outlier detection.



<new process*> – RapidMiner Studio Educational 10.3.001 @ Ken
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
for_discord_ready_set (3/28
Operators
loop
Loop Values
Loop Parameters
Loop Files
Loop Attribute Subset
Loop Examples
Process
90%
Retrieve for_discor...
Select Attributes
Nominal to Numerical
Normalize
Clustering
Parameters
Process
Hide advanced parameters
Change compatibility (10.3.001)
Activate Wisdom of Crowds
Upcoming Earnings
Search
ENG US
11:38 PM 3/28/2024

loop

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

We found "Data Search for Data Mining", "SMB Connector" and one more result in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 9 using already cleaned data Ken, performed the K means Clustering task.

4.4 Assess Model

We need to evaluate how well the clusters represent the underlying patterns in the data. Unlike supervised learning models where you have a clear metric (e.g., accuracy, precision, recall) for evaluation, clustering models require different approaches for assessment.

5.0 Evaluation

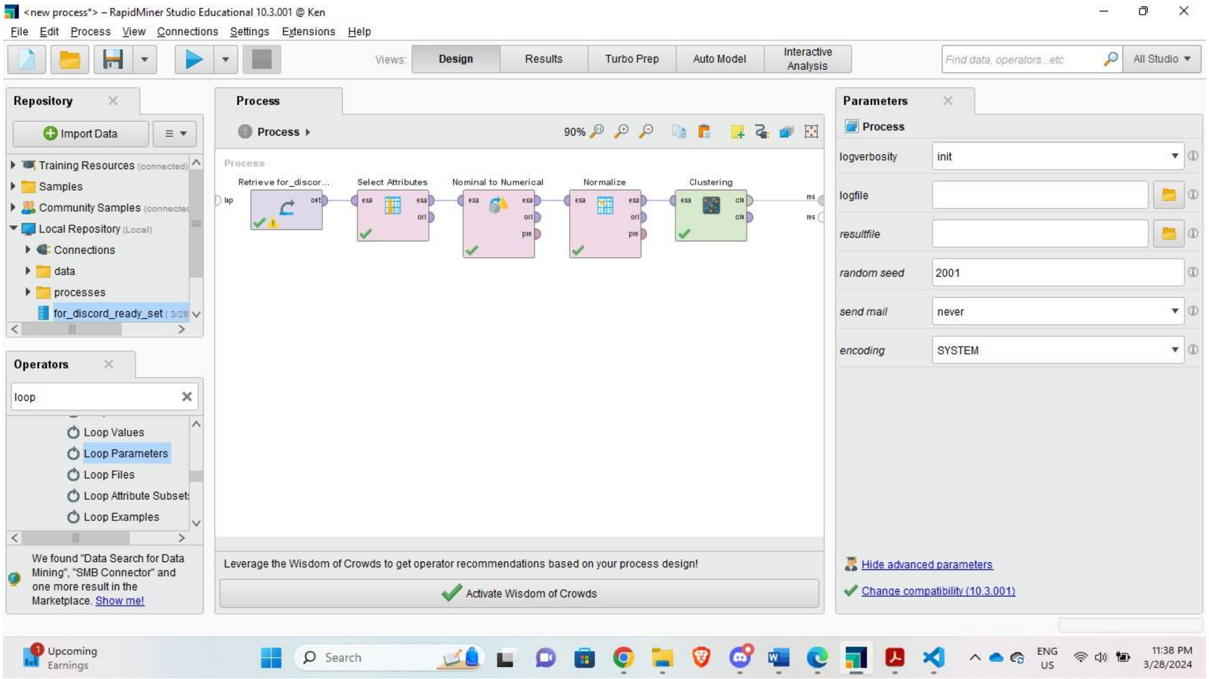
5.1 Evaluate Results

K-Means Clustering: K-means clustering can reveal underlying patterns in bike theft incidents by grouping them based on similarities in various attributes, such as location coordinates (latitude and longitude), time of theft, type of bike stolen, and other available features. This exploratory analysis can identify hotspots, temporal patterns, or common characteristics of thefts, offering insights into how and where bike thefts cluster within the city.

Outlier Detection: The LOF and distance-based outlier detection methods were instrumental in cleaning the dataset and ensuring the robustness of subsequent analyses. However, as standalone models, they did not directly contribute to predicting survival but rather supported the preprocessing phase for other predictive models.

5.2 EVALUATE RESULTS

Based on the evaluation, **the decision tree model emerges as the most aligned with our business objectives**. It not only provides predictions with a reasonable degree of accuracy but also offers insights into the importance of different features for bike theft, meeting the dual objectives of prediction and explanation.



<new process*> – RapidMiner Studio Educational 10.3.001 @ Ken
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
for_discord_ready_set (3/28
Operators
loop
Loop Values
Loop Parameters
Loop Files
Loop Attribute Subset
Loop Examples
Process
Process
90%
Retrieve for_discor...
Select Attributes
Nominal to Numerical
Normalize
Clustering
Activate Wisdom of Crowds
Parameters
Process
Hide advanced parameters
Change compatibility (10.3.001)
Upcoming Earnings
Search
ENG US
11:38 PM 3/28/2024

loop

logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

We found "Data Search for Data Mining", "SMB Connector" and one more result in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 10 Clustering for Bike Thieft

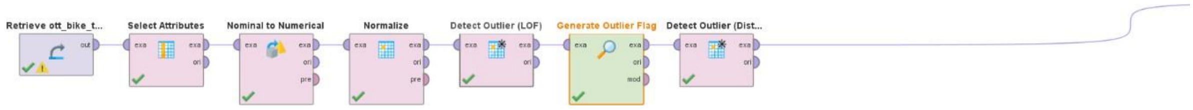


Figure 11 LOF for bike theft

References :

https://open.ottawa.ca/datasets/ottawa::bike-thefts/about

https://rpubs.com/pandeadhi92/bike_theft_toronto