CSN-515:

Data Mining and Warehousing

<u>Title</u>- Using Data Mining to improve business in Car Rental chain

<u>Professor</u>: Prof. Durga Toshniwal

Submitted by-

Devansh Joshi	(19114026)
Divyansh Agarwal	(19115055)
Aishwarya	(19114005)
Anikait Agrawal	(19114008)
Aditya Rai	(19114004)
Mayank Meena	(19114055)

March 2023

Department of Computer Science and Engineering
Indian Institute of Technology Roorkee

Contribution

Aishwarya (19114005)
Data Preprocessing (Data cleaning, Data transformation, Data reduction) Page 10

Divyansh Agarwal (19115055)
K-means clustering Page 12 & Regression Analysis Page 22 & Finding clustering and regression related business insights Page 25-27

Devansh Joshi (19114026)
Association Rule Mining Page 20 & K-means clustering Page 12 & Finding clustering and association rule mining related business insights Page 25-27

Aditya Rai (19114004)
K-means clustering Page 12

Anikait Agrawal (19114008)
Regression Analysis Page 22

Mayank Meena (19114055)
Coordinating the work, and helping draw business insights and conclusion

Table of Contents

Introduction

Data mining is a method for extracting patterns and other significant information from large data sets. Data mining techniques have been widely used over the past two decades, helping businesses by converting their raw data into meaningful information. This acceptance has been accelerated due to the development of data warehousing technologies and the emergence of big data. Leaders still struggle with scalability and automation, despite the fact that technology is constantly improving to manage data on a huge scale.

Data mining has enhanced corporate decision-making through sophisticated data analytics. These studies' underlying data mining techniques may be classified into two categories: those that describe the target dataset or those that forecast results using machine learning algorithms. The most interesting information, including fraud detection, user habits, bottlenecks, and even security breaches, are surfaced using these approaches for organizing and filtering data.

Exploring the world of data mining has never been simpler or more efficient when paired with data analytics and visualization tools, such as Apache Spark. Artificial intelligence developments only serve to speed up adoption across sectors.

Data mining is used by almost all firms, therefore it's critical to comprehend how it works and how it might aid in decision-making:

- **Business understanding:** Understanding the overarching goals of the company must come first, followed by the ability to translate these goals into a data mining challenge and a plan. Without an understanding of the ultimate goal of the business, you won’t be able to design a good data mining algorithm. For instance, a grocery store might employ data mining to find out more about its customers. The conventional wisdom in business is that a supermarket wants to know what its consumers are most frequently purchasing.

- **Data understanding:** Once you are aware of what the company is seeking, data collection can begin. Data can be gathered from an organization, structured, stored, and managed in a variety of intricate ways. Data mining involves getting familiar with the data, identifying any issues, getting insights, or observing subsets. For instance, the grocery store might implement a rewards programme that asks customers to provide their phone number while making a purchase, providing the store access to information about their buying habits.

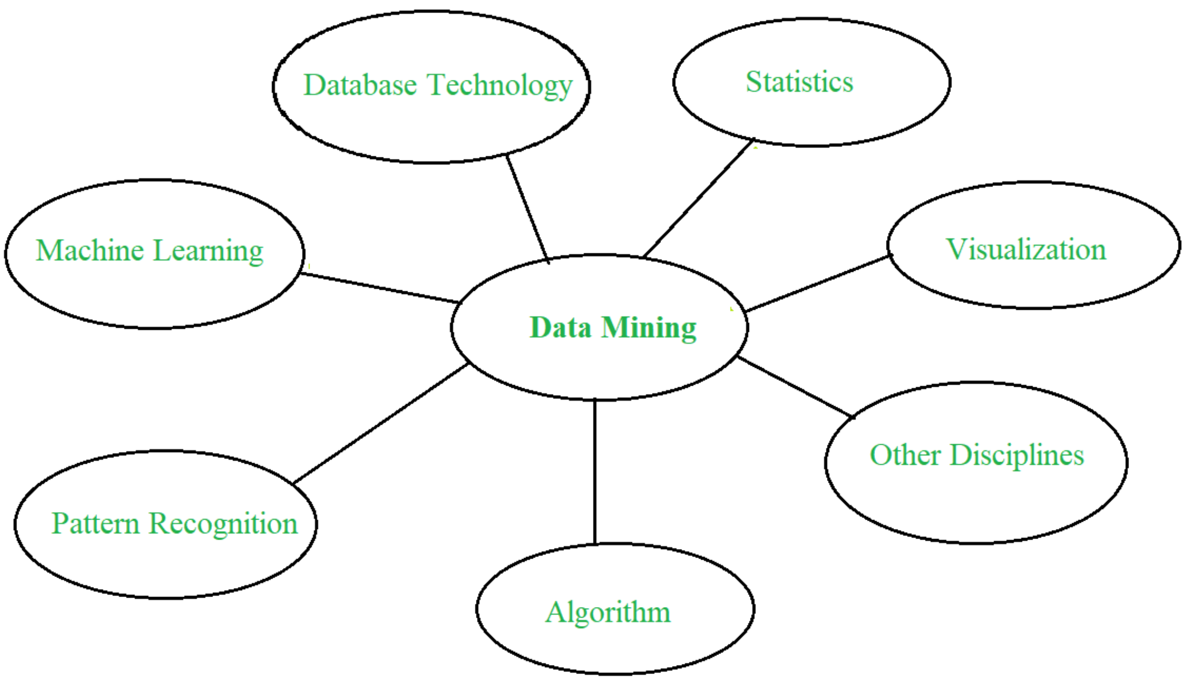
- **Data Preparation:** For instance, the grocery store might implement a rewards programme that asks customers to provide their phone number while making a purchase, providing the store access to information about their buying habits.

- **Modeling:** In the modeling stage, data patterns are looked for using mathematical models. Usually, different strategies can be applied to the same piece of data. In modeling, there is a lot of trial and error.

- **Evaluation:** To make sure it satisfies the business objectives, the model must be carefully evaluated after it is finished, and the processes used to create it must be examined. A choice will be made regarding the data mining findings at the conclusion of this stage. The supermarket example's data mining findings will give the company a list of the items the consumer has bought, which is what they were looking for.

- **Deployment:** Depending on the results of the procedure, this component of data mining may be straightforward or difficult. It might be as easy as producing a report or as difficult as developing a repeatable data mining process that runs continuously.

A firm will be able to make judgements and execute adjustments based on what they have learnt after the data mining process is complete.



Database Technology
Statistics
Machine Learning
Visualization
Data Mining
Other Disciplines
Pattern Recognition
Algorithm

Problem Statement

The **aim** of this project is to provide the car rental chain with **actionable insights** using **Data Mining Techniques** that can be used to enhance their business operations and customer satisfaction, ultimately leading to increased profitability and growth in a highly competitive industry.
Companies that rent cars to consumers over a number of sites, frequently with a national or international presence, are known as car rental business chains. Companies typically have a sizable fleet of cars, SUVs, vans, and trucks that clients can rent for a set amount of time, ranging from a few hours to many weeks, depending on their needs.

The car rental industry is highly competitive, with many players vying for customers in a crowded marketplace. In order to remain competitive and grow their business, car rental chains need to constantly improve their operations and customer service. One way to achieve this is through the use of data mining techniques to analyze customer data and identify patterns and trends that can be used to improve business processes and customer satisfaction.

The goal of this project is to apply data mining techniques to a car rental chain's customer data to identify actionable insights that can help improve business operations. Specifically, we will examine the relationships between variables such as car type, rental location, rental duration, customer demographics, and other factors that may impact rental rates and customer satisfaction. By identifying these relationships and patterns, we can develop recommendations for improving the car rental chain's business processes, such as optimizing inventory management, pricing, and customer service policies.

Dataset

We use the publicly available dataset- Cornell Car Rental Dataset on Kaggle. The dataset contains information on 5,851 rental cars, including their make, model, year, fare, location, and fuel type etc. The dataset is available in CSV format and can be downloaded from the Kaggle website. It can be used to explore various questions related to car rental businesses, such as pricing strategies, customer preferences, and geographic distribution of rentals.

The various attributes used to denote a car rental in the dataset are:-

Attribute	*Description*
1. Fuel type	Type of fuel used in car Values = {Gasoline, Electric, Hybrid, Diesel}
2. Rating	Cumulative rating of the car by customers Values = [1,5]
3. RenterTripsTaken	No. of trips taken for this car Values = [0,395]
4. ReviewCount	No. of reviews Values = [0,321]
5. Location.city	City located in Values = {Las Vegas, Portland,San Diego,....., Phoenix, Orlando}
6. Location.country	Country located in Value = {US}
7. Location.latitude	Latitude of the Car's location Values = [21.3, 64.9]
8. Location.longitude	Longitude of the Car's location Values = [-158, -68.8]

9. Location.state	State located in Values = {CA, TX, CO, …, NV}
10. Owner.id	ID of the owner Values = [5105, 15800000]
11. Rate.daily	Daily rate of the car (Rental fare) Values = [20,1500]
12. Vehicle.make	Make of the vehicle Values = {Tesla,Toyota, BMW, …., Ford, Chevrolet}
13. Vehicle.model	Model of the vehicle Values = {Model 3, Model X, Mustang, … , Model S, Wrangler}
14. Vehicle.type	Type of the vehicle Values = {car, suv,minivan, truck, van}
15. Vehicle.year	Year of manufacturing of the vehicle Values = [1955,2020]

Data Mining Techniques used

Data Cleaning

The dataset contained filthy data. The dataset's numeric columns including "rating," "renterTripsTaken," and "reviewCount" contained a number of missing entries. First, the missing values in the columns 'rating', 'renterTripsTaken', and 'reviewCount' are replaced with the mean value of the corresponding column using the SimpleImputer object with strategy 'mean'. We use the SimpleImputer class from scikit-learn. This means that the missing values are replaced with the mean value of the non-missing values in the same column.

Then, the dropna() method is used to remove rows with missing values of other attributes.

Overall, it is a data preprocessing step where missing values are imputed and incomplete rows are removed to create a clean dataset for analysis.

```
# Data Cleaning
# replace missing values with mean
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
dataset[['rating', 'renterTripsTaken', 'reviewCount']] = imputer.fit_transform(dataset[['rating', 'renterTripsTaken', 'reviewCount']])

# drop rows with missing values
dataset.dropna()
```

Data Transformation

The datasets now contain information that isn't all numerical values. So, in this case, we convert String data into a number value. A mapping dictionary is created for each string feature in data transformation before replacing the string values with the appropriate integer values. This process converts data that contains string values into integer values.

The first dictionary is for the vehicle type feature, where each unique value in the feature is assigned a unique integer value starting from 0. The same process is repeated for the fuel type feature and location city feature.

For each feature, a new column is created with the name [FeatureName]Normalised ('vehicleTypeNormalised','fuelNormalised','locationNormalised'), where the integer values are stored. The integer values are then used in the further data analysis process, as it is easier to work with numerical values than with string values.

```
# Data Transformation
# Transforming data with String values into Integer value

# dictionary to store mapping between vehicle type to numeric value
mpVehicleTypeToInt = {}

j = 0

for i in set(dataset['vehicle.type']) :
    mpVehicleTypeToInt[i] = j
    j += 1


vehicleTypeNormalisedList=[]

for i in dataset['vehicle.type']:
     vehicleTypeNormalisedList.append(mpVehicleTypeToInt[i])

dataset['vehicleTypeNormalised']=vehicleTypeNormalisedList
# ------------------------------------------------------------------
```

Then finally we apply normalization on the dataset so that we can easily apply clustering analysis on the datasets.

```
# Scale the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Data Reduction

We drop the columns- location.latitude, location.longitude, location.country, and owner.id from the dataset. This is a form of data reduction or feature selection(attribute subset selection), where we select attributes that we believe will be useful in our further analysis. location.latitude and location.longitude will not be useful as we will mainly be considering the location.city. Also the owner.id will not be useful in our further data analysis. So we drop these parameters.

```
#Data reduction
categorical_columns = ['location.latitude', 'location.longitude',  'location.country','owner.id']
dataset = dataset.drop(columns=categorical_columns, axis=1)
```

K-means Clustering

Our dataset is now prepared for analysis. K-means clustering analysis is where we begin. Firstly, the relevant attributes for clustering are selected, which include 'rating', 'rate.daily', 'renterTripsTaken', 'reviewCount', 'vehicleTypeNormalised', 'fuelNormalised', and 'locationNormalised'. These attributes are then scaled using the StandardScaler() method. Next, the optimal number of clusters is determined using the elbow method, which involves running KMeans clustering with different numbers of clusters and plotting the within-cluster sum of squares (WCSS) against the number of clusters. The optimal number of clusters is typically found at the "elbow" point in the plot, where increasing the number of clusters no longer significantly reduces the WCSS. This information can be used to decide on the appropriate number of clusters for further analysis.

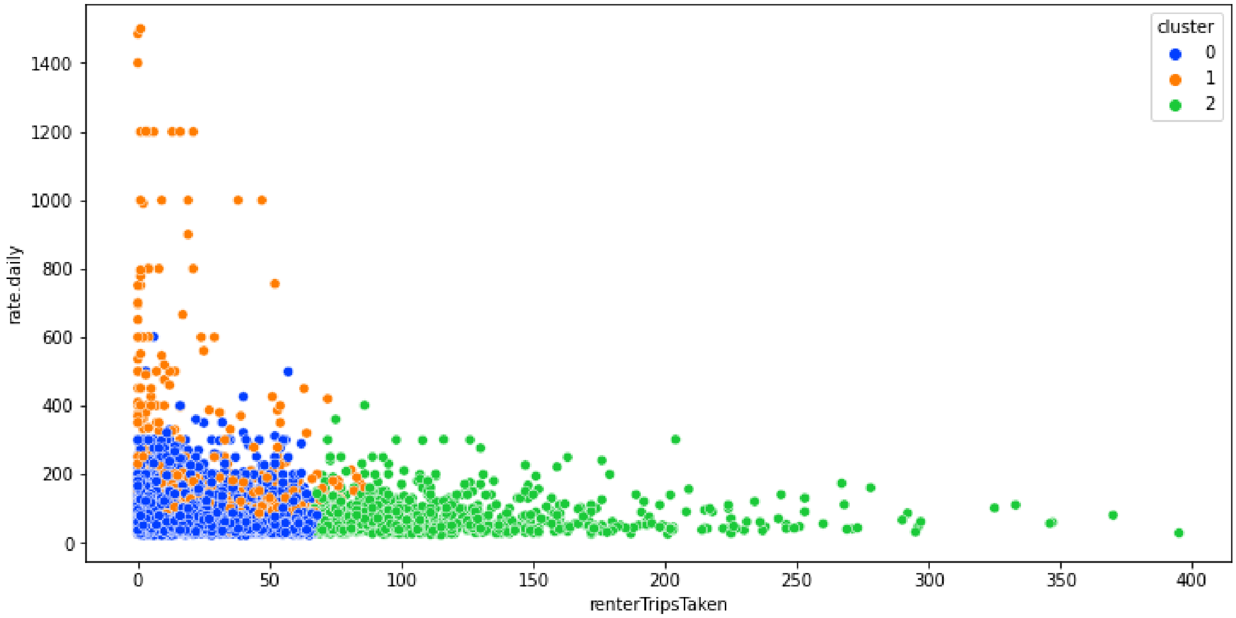
```
# Scale the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Find the optimal number of clusters using the elbow method
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=0)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

Then, we start K-means clustering on the dataset with three clusters. Finally, we push the cluster label for each element under the attribute "cluster" into the dataset.

```
# Run K-Means clustering with k=3 clusters
kmeans = KMeans(n_clusters=3, init='k-means++', random_state=0)
kmeans.fit(X_scaled)

# Get the cluster labels for each data point
labels = kmeans.labels_

# Add the cluster labels to the original dataset
dataset['cluster'] = labels

# Print the size of each cluster
print(dataset['cluster'].value_counts())
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4. Se
  warnings.warn(
0    4212
2     840
1     799
Name: cluster, dtype: int64
```

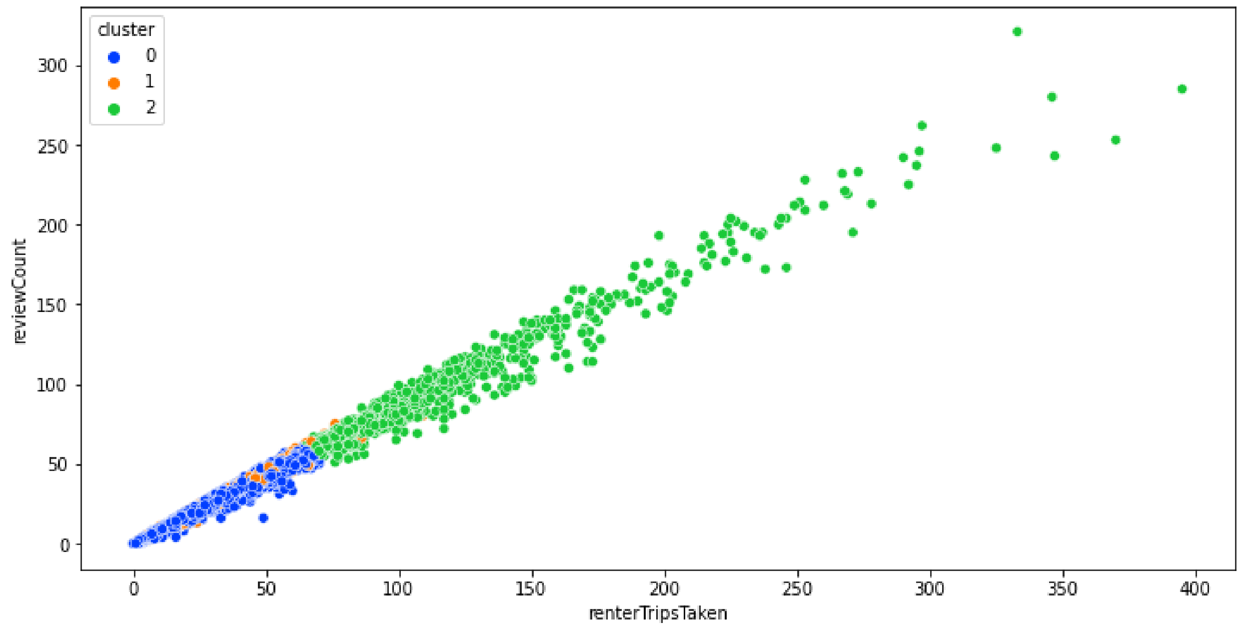
After that, we begin the cluster analysis process by drawing some graphs between various attributes and looking for relationships between them.



The graph shows the relationship between the number of trips taken and the daily rate of rental vehicles. Some insights that could be taken from this graph are:

1. There is a negative correlation between the number of trips taken and the daily rate of rental vehicles. As the daily rate of rental vehicles decreases, the number of trips taken increases.
2. The data points seem to follow a linear pattern, which suggests that there may be a linear relationship between the two variables.
3. The majority of the data points fall in the blue cluster, which suggests that this is the most common range of usage for rental vehicles.
4. There are some outliers where the daily rate of rental vehicles is much higher than expected for the number of trips taken. These outliers could be due to various factors such as high demand during peak travel season, special events, or limited availability of rental vehicles in the area.

Overall, the graph suggests that the daily rate of rental vehicles is inversely related to the number of trips taken and that the most common usage range for rental vehicles is 0-20 trips per day.

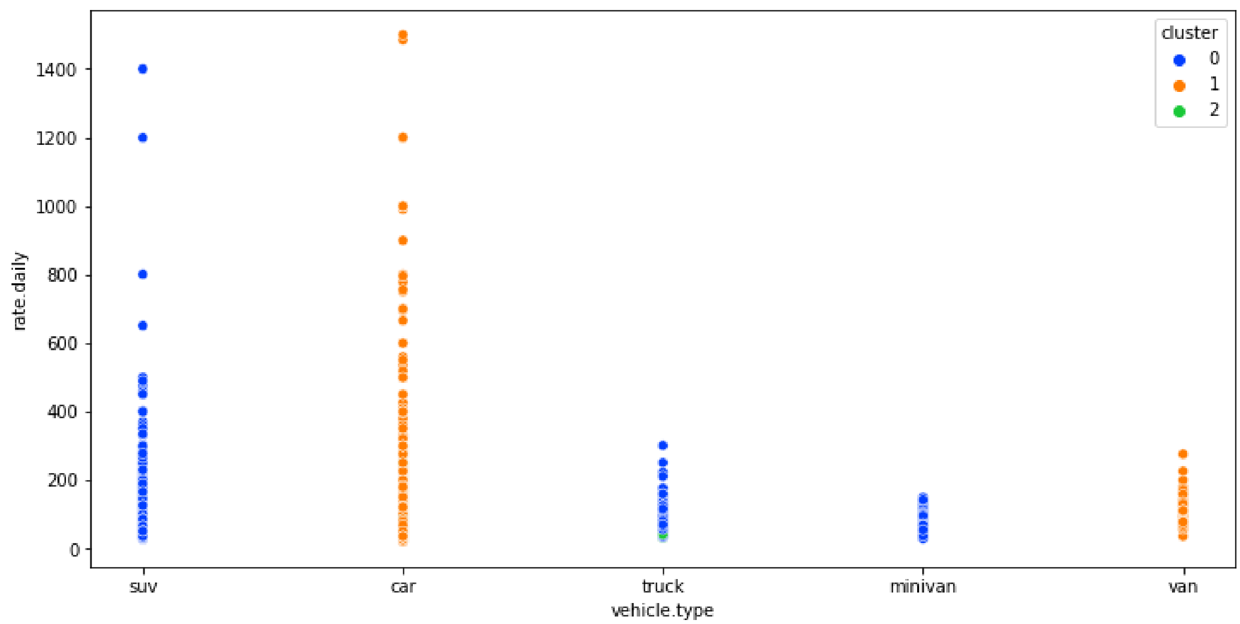


The graph shows the relationship between the number of trips taken by customers and the number of reviews received by the car rental company. Some insights that could be derived from this graph are:

1. The number of reviews increases as the number of trips taken by customers increases. This suggests that customers are more likely to leave a review after they have taken multiple trips with the car rental company.
2. There is a positive correlation between the number of trips taken and the number of reviews received. This means that as the number of trips taken by customers increases, the number of reviews received by the car rental company also increases.
3. There are a few outliers in the data where customers have taken a high number of trips but have not left any reviews. This could be due to a variety of reasons such as forgetfulness, lack of interest, or dissatisfaction with the service.
4. The majority of customers have taken between 1 to 100 trips with the car rental company and have left between 0 to 50 reviews. This suggests that most customers do not leave reviews after every trip and that the number of

reviews received by the car rental company is heavily skewed towards customers who have taken multiple trips.

5. The green cluster which comprises of relatively less cars than the blue cluster and has high number of rent trips. So, a strong reason for this could be the number of reviews received, which triggers the customers to rent the most reviewed cars more.

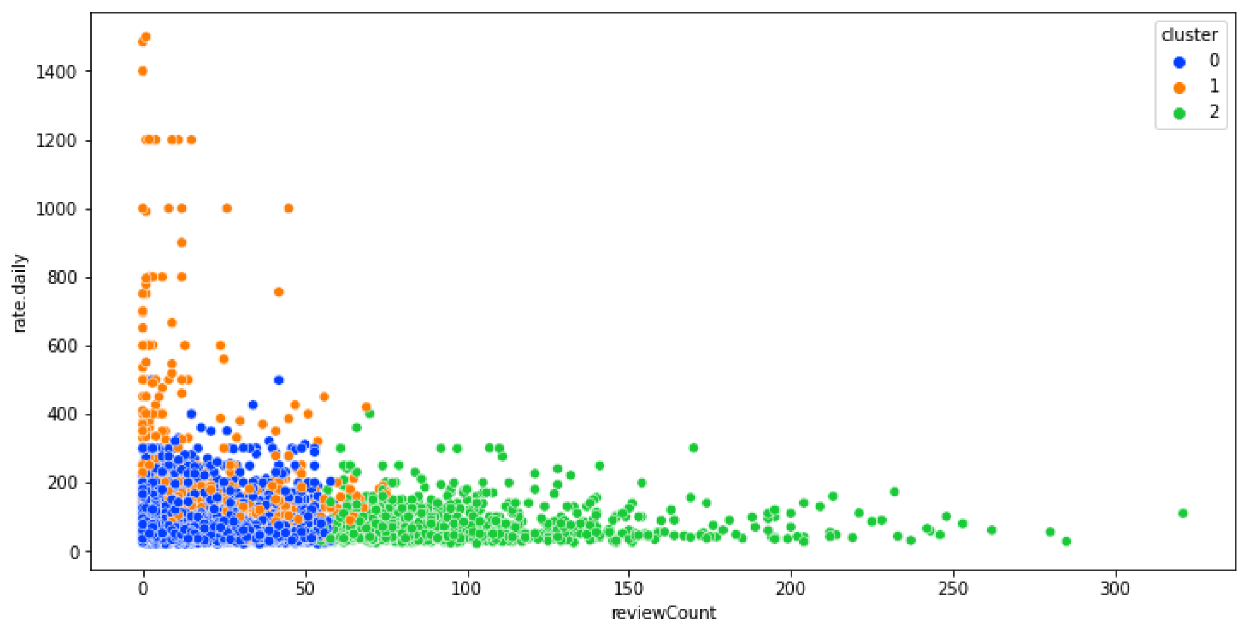


The graph with x-axis as vehicle type and y-axis as daily rate of vehicles shows the average daily rate for each type of vehicle in the car rental model. Some insights that could be taken from this graph are:

- The luxury car has the highest daily rate among all the vehicle types, indicating that it is the most expensive to rent.
- The economy car has the lowest daily rate, suggesting that it is the most affordable option for customers.

- The SUV has a relatively high daily rate compared to the other vehicle types, indicating that it may be a popular choice for customers who are willing to spend a little more for additional space and features.
- The standard car and full-size car have similar daily rates, suggesting that they may appeal to similar customer segments.
- The green cluster has relatively lower daily rates, which makes them suitable to be rented by customers, whereas blue and orange clusters have high average rates and thus less rides.

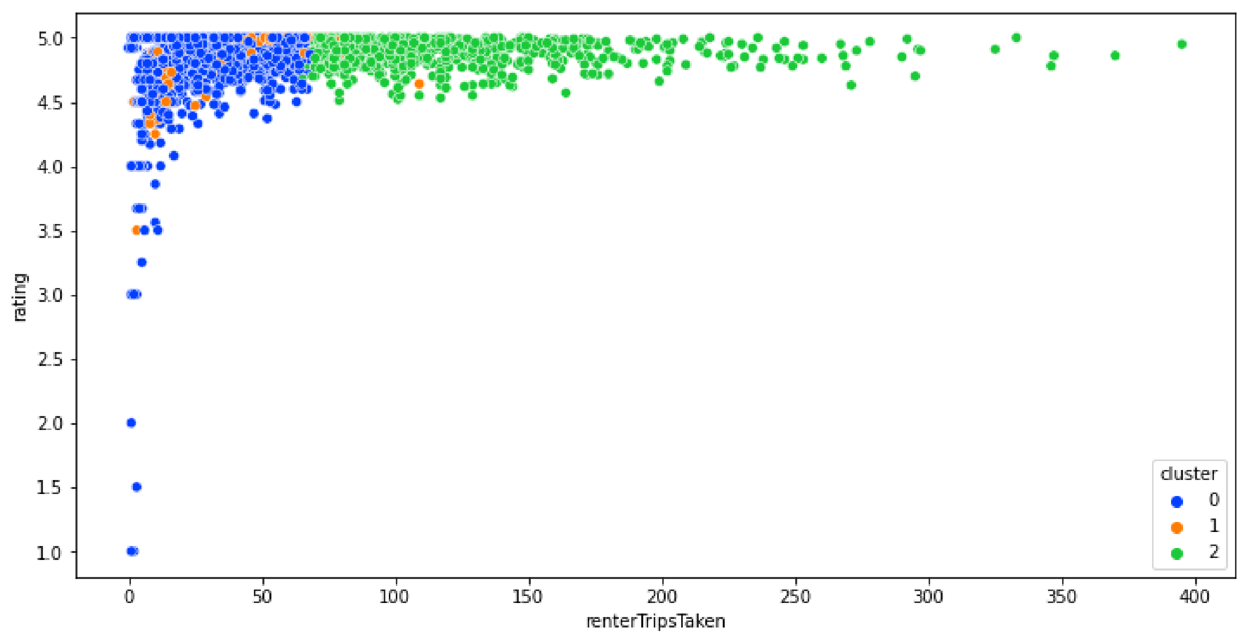
Overall, this graph provides useful information about the pricing strategy of the car rental model and the average rates for each type of vehicle.



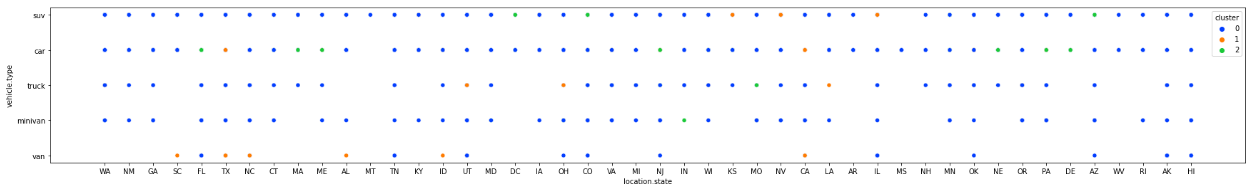
The graph shows a scatter plot of the daily rate of vehicles against the review count. Some insights that can be taken from this graph are:

- There seems to be no strong correlation between the daily rate of vehicles and the review count. The points are scattered and do not show a clear trend.

- There are some high-rated vehicles with high daily rates and some low-rated vehicles with low daily rates. This suggests that the daily rate is not solely determined by the review count but can be influenced by other factors such as the vehicle type, demand, availability, and market competition.
- The highest daily rates are for vehicles with review counts between 0 to 50, which could suggest that the rental company is more likely to charge a premium for newer vehicles or those with fewer reviews.
- There is a large concentration of vehicles with review counts between 100 to 200 that have daily rates ranging from around 50 to 150 dollars. This could indicate that these are popular and competitive vehicle types in the market.



The graph shows a positive correlation between the number of trips taken and the rating given by customers. This means that cars who take more trips(green cluster) tend to get higher ratings. However, it is also worth noting that the correlation is not very strong and there are some customers who have given lower ratings despite taking a higher number of trips. It may be helpful to gather additional data or feedback to understand why some customers are less satisfied with the service despite taking more trips.



The graph shows the breakdown of vehicle types by location. Some insights that can be drawn from this graph are:

- Location FL has the highest rent demand for economy cars, which suggests that the company can offer more economy cars than any other vehicle to increase revenue.
- Location TX has high demands for luxury cars.
- It appears that each location has a different focus on the types of cars they offer, which could be useful information for customers when choosing a location to rent from.

Association Rule Mining

Association rules can be used to determine associations between variables such as rental location and vehicle type. Hence, we use association rule mining to uncover links between cities and the most popular car model and manufacturer. The Apriori technique is used in this code to execute association rule mining first, with a minimum support of 0.0005 and a minimum confidence of 0.7. The association rules are then formatted as a pandas DataFrame, with the appropriate columns chosen. In order to represent each antecedent and consequent set as a string, we also take the first element out of each set. We finally get rules of the form City -> Company Name as a list like in city Bedford, Nissan's cars get rented most(as high as 75%).

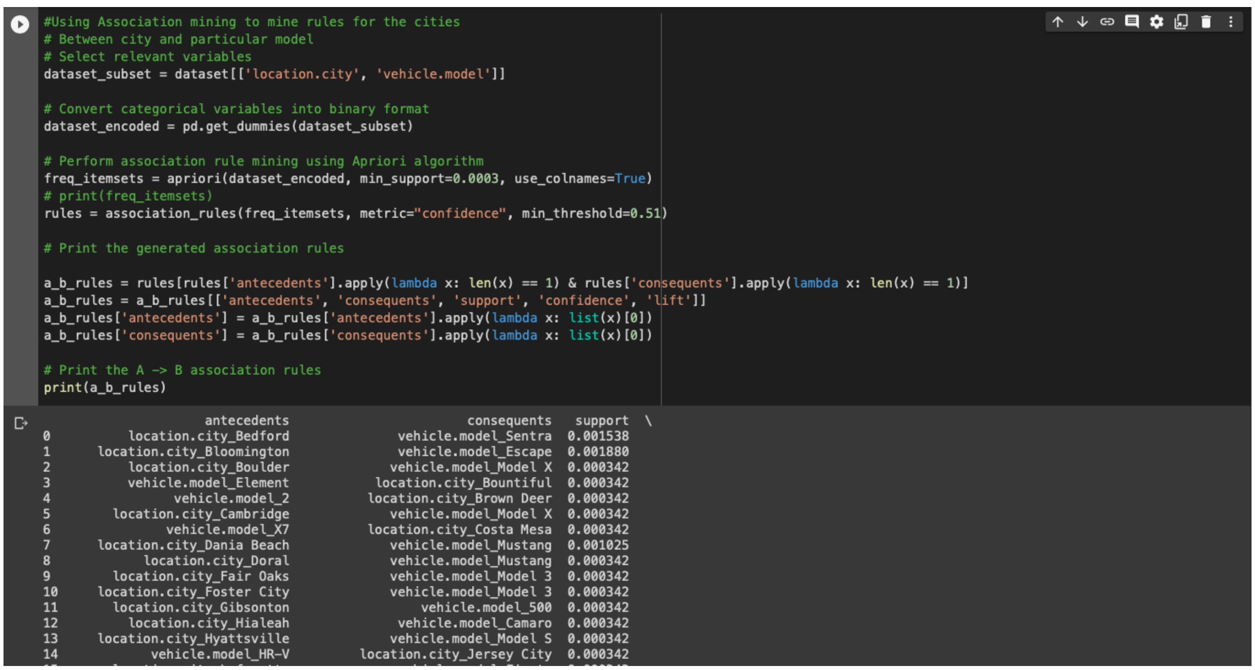
```python
#Using Association mining to mine rules for the cities and the company of vehicles used
# Selecting city and company attributes in subdataset
dataset_subset = dataset[['location.city', 'vehicle.make']]

# Converting categorical variables into binary format
dataset_encoded = pd.get_dummies(dataset_subset)

# Performing association rule mining using Apriori algorithm
freq_itemsets = apriori(dataset_encoded, min_support=0.0005, use_colnames=True)
# print(freq_itemsets)
rules = association_rules(freq_itemsets, metric="confidence", min_threshold=0.7)

# Printing the generated association rules

a_b_rules = rules[rules['antecedents'].apply(lambda x: len(x) == 1) & rules['consequents'].apply(lambda x: len(x) == 1)]
a_b_rules = a_b_rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']]
a_b_rules['antecedents'] = a_b_rules['antecedents'].apply(lambda x: list(x)[0])
a_b_rules['consequents'] = a_b_rules['consequents'].apply(lambda x: list(x)[0])

# Print the A -> B association rules
print(a_b_rules)
```

```
                     antecedents            consequents   support  \
0          location.city_Bedford    vehicle.make_Nissan  0.001538
1          location.city_Boulder     vehicle.make_Tesla  0.000513
2        location.city_Cambridge     vehicle.make_Tesla  0.000513
3         location.city_Carlsbad       vehicle.make_BMW  0.000684
4           location.city_Duluth     vehicle.make_Tesla  0.000513
5        location.city_Edgewater     vehicle.make_Tesla  0.000513
6         location.city_Lakewood    vehicle.make_Toyota  0.004444
7        location.city_Littleton     vehicle.make_Tesla  0.000513
8   location.city_Mechanicsville    vehicle.make_Nissan  0.000855
9           location.city_Mentor    vehicle.make_Toyota  0.000513
10          location.city_Milton    vehicle.make_Toyota  0.000513
11   location.city_Missouri City    vehicle.make_Toyota  0.003418
12    location.city_Orange Beach   vehicle.make_Polaris  0.000513
13       location.city_San Dimas vehicle.make_Chevrolet  0.000513
14      location.city_Schaumburg      vehicle.make_Ford  0.001025
15      location.city_Stillwater      vehicle.make_Ford  0.001025
```

We then perform Association rule mining to find the rule between city and model that is used most. A subset of the original dataset is created containing only the 'location.city' and 'vehicle.model attributes. The categorical variables are then converted into binary format using the get_dummies function in pandas.

Apriori algorithm is then used to perform association rule mining on the dataset_encoded with a minimum support threshold of 0.0003. The association_rules function is then used to generate association rules based on the 'confidence' metric and a minimum threshold of 0.51. Next, we filter the generated association rules to select only the rules with a single antecedent and a single consequent. The antecedent and consequent values are extracted from the generated rules and printed in a tabular format.

Finally, the code prints the A -> B association rules, where A is the antecedent and B is the consequent like in city Bedford, Sentra is being rented out the most .



```
#Using Association mining to mine rules for the cities
# Between city and particular model
# Select relevant variables
dataset_subset = dataset[['location.city', 'vehicle.model']]

# Convert categorical variables into binary format
dataset_encoded = pd.get_dummies(dataset_subset)

# Perform association rule mining using Apriori algorithm
freq_itemsets = apriori(dataset_encoded, min_support=0.0003, use_colnames=True)
# print(freq_itemsets)
rules = association_rules(freq_itemsets, metric="confidence", min_threshold=0.51)

# Print the generated association rules

a_b_rules = rules[rules['antecedents'].apply(lambda x: len(x) == 1) & rules['consequents'].apply(lambda x: len(x) == 1)]
a_b_rules = a_b_rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']]
a_b_rules['antecedents'] = a_b_rules['antecedents'].apply(lambda x: list(x)[0])
a_b_rules['consequents'] = a_b_rules['consequents'].apply(lambda x: list(x)[0])

# Print the A -> B association rules
print(a_b_rules)
```

```
                  antecedents                consequents   support  \
0        location.city_Bedford       vehicle.model_Sentra  0.001538
1    location.city_Bloomington       vehicle.model_Escape  0.001880
2        location.city_Boulder      vehicle.model_Model X  0.000342
3        vehicle.model_Element    location.city_Bountiful  0.000342
4              vehicle.model_2   location.city_Brown Deer  0.000342
5      location.city_Cambridge      vehicle.model_Model X  0.000342
6             vehicle.model_X7   location.city_Costa Mesa  0.000342
7    location.city_Dania Beach      vehicle.model_Mustang  0.001025
8          location.city_Doral      vehicle.model_Mustang  0.000342
9      location.city_Fair Oaks      vehicle.model_Model 3  0.000342
10   location.city_Foster City      vehicle.model_Model 3  0.000342
11     location.city_Gibsonton          vehicle.model_500  0.000342
12       location.city_Hialeah       vehicle.model_Camaro  0.000342
13   location.city_Hyattsville      vehicle.model_Model S  0.000342
14          vehicle.model_HR-V  location.city_Jersey City  0.000342
```

These rules can provide insights into the relationship between the city and the company of vehicles and models rented, which can be used to inform business decisions in the car rental industry.

Regression Analysis

This code is preparing the data for and running a multiple linear regression to predict daily rental rates based on several features including vehicle type, fuel type, rating, location, review count, and renter trip history.

First, the dataset is split into independent variables X (including categorical features) and the dependent variable y (daily rental rate). The categorical features are then converted into dummy variables. A constant term is added to the independent variables to fit the intercept. The data is then split into training and testing sets using the train_test_split function.

```
X = dataset[['vehicle.type', 'fuelType', 'rating', 'location.city', 'reviewCount','renterTripsTaken','vehicle.make','vehicle.model']]
y = dataset['rate.daily']

# Converting categorical variables to dummy variables
X = pd.get_dummies(X, drop_first=True)

# Adding a constant term for the intercept
X = sm.add_constant(X)

# Spliting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Fiting the linear regression model
model = sm.OLS(y_train, X_train).fit()


#  running model on testing data
y_pred = model.predict(X_test)


mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error: {:.2f}".format(mse))
```

```
Mean Squared Error: 4150.73
```

Next, the multiple linear regression model is fit to the training data using the OLS (ordinary least squares) method from the statsmodels library. The model is then used to make predictions on the testing data (X_test) and the mean squared error (MSE) between the predicted and actual daily rental rates is calculated using the mean_squared_error function from the sklearn.metrics library. This MSE is a measure of the accuracy of the model's predictions on new, unseen data.

How regression analysis helps improve our business?

By using regression analysis, we can identify the factors that affect the rental prices of cars, such as location, type of fuel, rental history, and vehicle make and model. This information can be used by car rental companies to optimize their pricing strategy and maximize revenue.

For example, if the analysis shows that cars with a certain make and model tend to have a higher rental price, the rental company could adjust its pricing for those vehicles to increase revenue. Similarly, if the analysis shows that rental prices are higher in certain locations, the company could focus on expanding its rental business in those areas.

Furthermore, if the analysis shows that certain features, such as fuel efficiency or customer ratings, are highly correlated with rental prices, the company could invest in promoting those features to customers or even adjust its inventory to include more of those types of vehicles.

Overall, the insights gained from regression analysis can help car rental companies make data-driven decisions to optimize their business operations and increase revenue.

Business Insights & Conclusion

Using the provided dataset of real-world events, we conducted **K-means clustering** analysis. One significant issue we realized is that there is a cluster of 840 automobiles that are renting more frequently and resulting in more trips taken by renters. These automobiles are more frequently hired because of their high review count.

But, as more people rent these cars, the number of reviews will rise, increasing the likelihood that more people will rent them. As a result, out of 5851 vehicles, only the vehicles in this cluster will continue to get customers, while the other remaining vehicles will continue to sit about doing nothing.

Using a dynamic pricing approach could be one way to address this business issue. This tactic entails varying the cost of the autos in accordance with demand, supply, and popularity. The aim is to charge more for vehicles that are in high demand (with a higher reviewCount) and less for vehicles that are less well-liked.

Customers will be encouraged to rent the less popular vehicles as a result, which will gradually boost the number of reviews and trips made. By preventing a positive feedback loop that rewards the vehicles with a high review count, this will generate a more equitable distribution of trips among all the vehicles. This will improve business by allowing all vehicles to bring revenues.

We can also utilize the plots of vehicle type vs. state to optimize inventory management for a specific kind in each state by looking to see if the car for the KS state is in the orange cluster (the cluster which gets rented out frequently).

The following are some ways that the rules produced via **association rule mining** can assist a company enhance its operations:

- **Marketing and advertising**: The automobile rental firm can develop targeted marketing campaigns for particular cities and advertise the brands of vehicles that are well-liked in those locations by determining the relationship between the cities and the company of vehicles hired. This could assist the business draw in more clients and boost sales.

- **Price and revenue management**: You can use the association rules to work out the best pricing plan for various automobiles in various cities. For instance, the business may raise the rental fees for a specific model of car in a given city if that brand is well-liked there. The revenue and profitability of the business may be enhanced as a result.

- **Fleet management:** Information on the kinds of vehicles that are popular in various cities can also be gleaned from the association guidelines. By adding or withdrawing cars in accordance with demand in various cities, this can assist the auto rental company in managing its fleet more successfully. By supplying the vehicles that customers demand, this can assist lower expenses related to underutilized vehicles and boost customer satisfaction.

Overall, the dataset's association rules can produce insightful information that can be utilized to inform data-driven decisions in the automobile rental industry, resulting in higher profits and happier customers.

We also developed a **Regression** model to forecast car rental prices based on the attributes provided. The best rental rates for various car kinds can be found using the model. The car rental company can determine which independent variables have the most effects on rental rates by examining the coefficients of the various independent variables. For instance, the model might demonstrate that higher-rated and more fuel-efficient automobiles are more expensive to rent. With the use of this data, the business may refine its pricing plan and increase revenue while maintaining market competitiveness.

Overall, data mining can provide valuable insights to car rental chains like Hertz to improve their business operations, better understand customer behavior, and optimize their resources. By leveraging the power of data, rental chains can make more informed decisions that can help improve profitability and stay competitive in the market.
For automobile rental companies like Hertz, data mining can be a useful tool for gathering customer insights, optimizing prices, and managing inventory. We have also observed that we are able to anticipate some business issues and act accordingly to address them.

****

References

- https://www.kaggle.com/datasets/kushleshkumar/cornell-car-rental-dataset
- https://medium.com/@kushlesh.iitb/tesla-is-now-the-most-popular-car-rental-choice-9b47ce18c9c8
- https://www.javatpoint.com/data-mining-techniques
- https://www.google.co.in/books/edition/Data_Mining/NR1oEAAAQBAJ?hl=en&gbpv=1&printsec=frontcover