Data-Driven Strategies for Insurance Fraud Prevention and Vehicle Risk Assessment

Abdul Aroworamimo

Agenda

01 **Summary**
Overview of methods and key findings

02 **Data**
Data overview and cleaning

03 **Vehicle Segmentation**
Segmentation technique, evaluation, and insights

04 **Anomaly Detection**
Technique, evaluation, insights, and limitations

05 **Temporal Analysis**
Technique, evaluation, insights, and limitations

06 **Recommendations**
Business Impact for Definity

Executive summary

Approach	Key Findings
K-Prototypes: Clustering Vehicles Based on Price, Mileage, Engine Size, Power, Age, and Vehicle Type	Vehicles were segmented into three distinct clusters identified as **Older City Commuters**, **Modern Power Vehicles,** and **Seasoned Family Travelers**
Isolation Forest: Detecting Anomalous Vehicle Offers	Anomalous vehicle listings were identified, including vehicles with over **1 billion** kilometers in mileage and cars priced as high as **2 million CAD**
Time-Based Analysis: Using **Autocorrelation** Analysis to Discern Actionable Temporal Patterns in Daily Vehicle Offer Volume	A relationship exists between the number of listings on any given day and the number of listings on previous days.

RECOMMENDATIONS

1. Leverage cluster-based insights and trends to **identify frivolous claims** that deviate from cluster norms, enhance underwriting efficiency, and develop targeted marketing strategies

2. Use anomaly detection to address information asymmetry by **flagging claims that significantly deviate from the norm**

3. Use observed temporal patterns **timing of marketing campaigns** especially on online car sales platforms

Problem statement

- Over **1.6 billion CAD** annually in increased premiums due to fraud
- **Information asymmetry** complicates the creation of accurate vehicle risk profiles
- High marketing costs

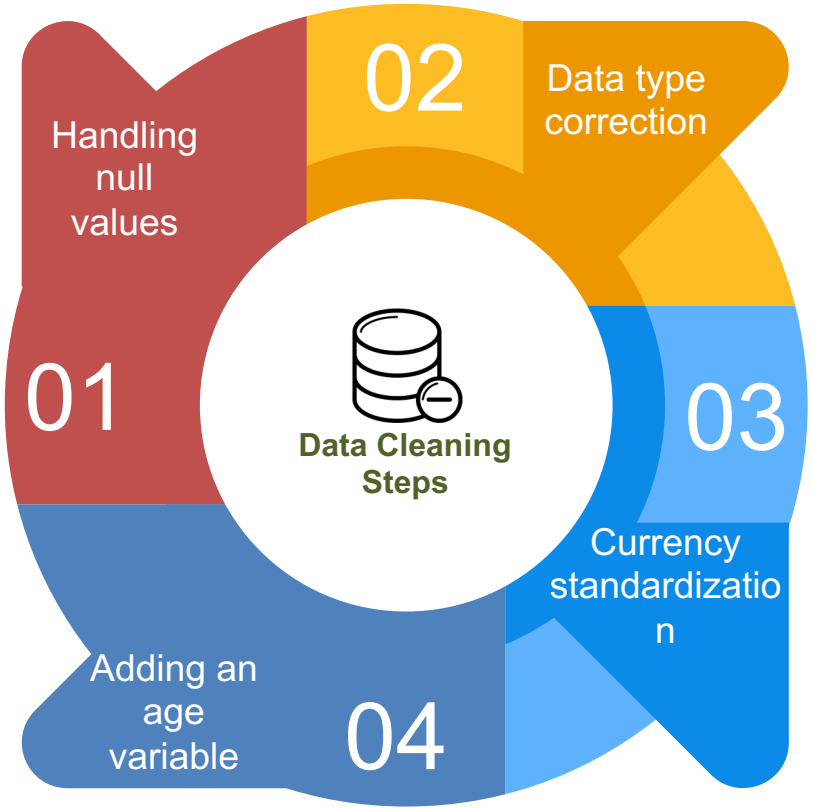
The Dataset

Dataset overview

Data Source	Kaggle
Description	Car advertisements scraped from a Polish online car ads website, including car features such as engine size, price, production year, mileage, and more
Size	The dataset contains **208304** records and **24** features. It includes a mix of numerical, date, textual, and categorical variables relevant to vehicle characteristics

Data cleaning

This step constitutes most of the data cleaning process



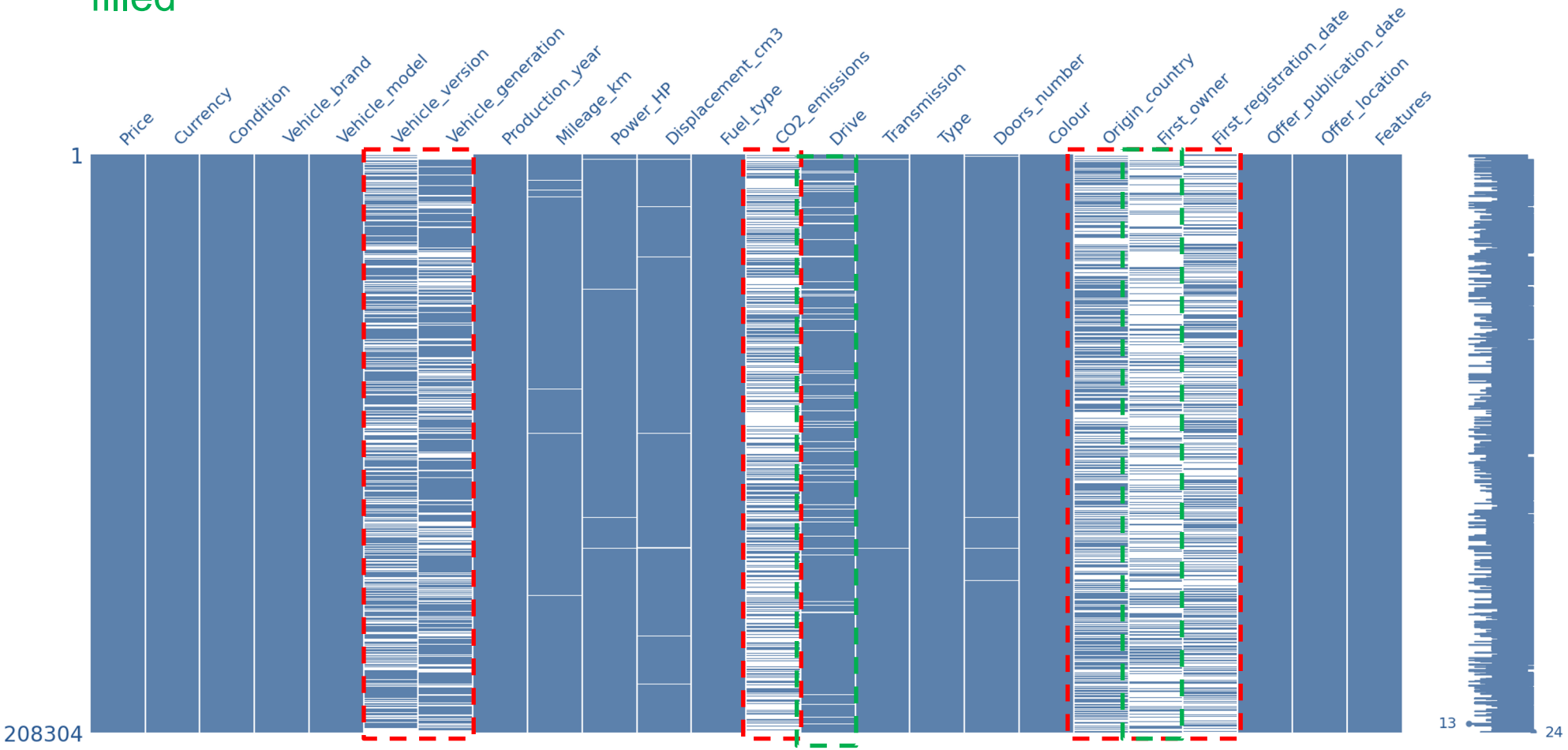
The two time-based variables, production year and offer date, were converted to datetime format

An age variable was created by subtracting the vehicle's production year from the offer year (2021)

The dataset included both euros and PLN, so all values were standardized to CAD

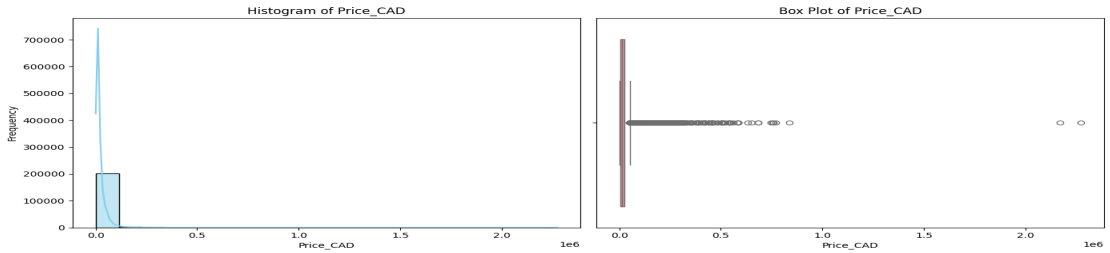
Handling null values

Most variables with substantial number of null values were dropped besides the first owner and Drive columns which were filled

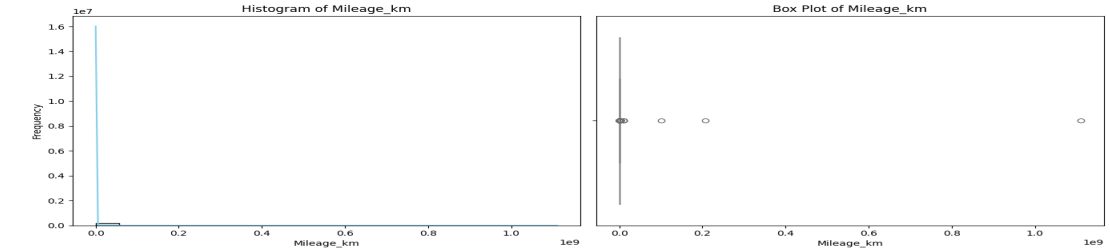


Exploratory Data Analysis

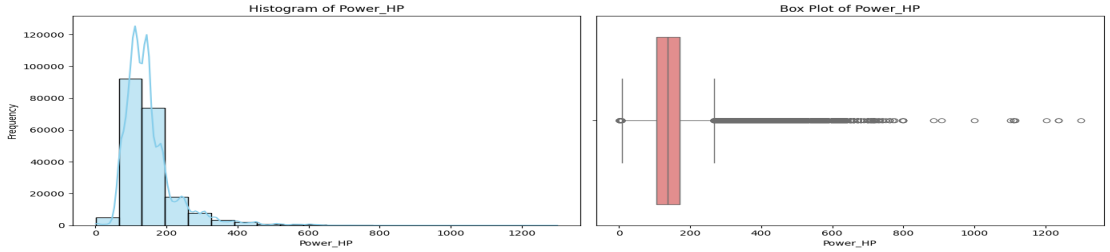
Many outliers observed in the numeric variables



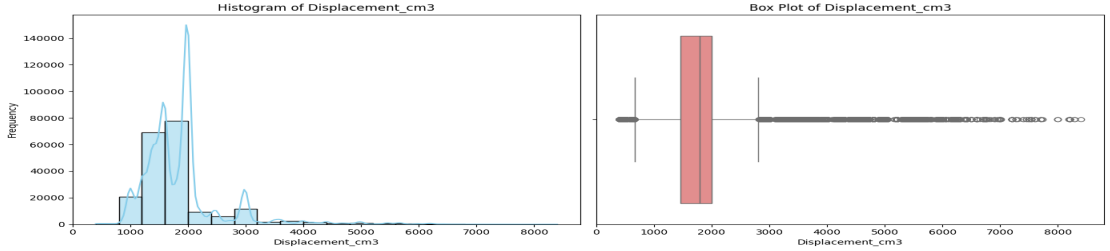
Price



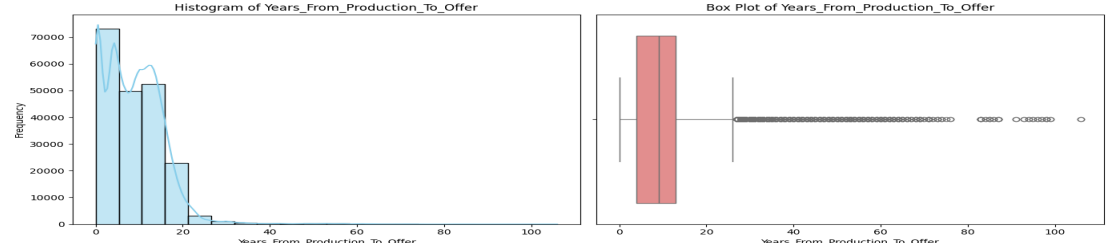
Mileage



Power

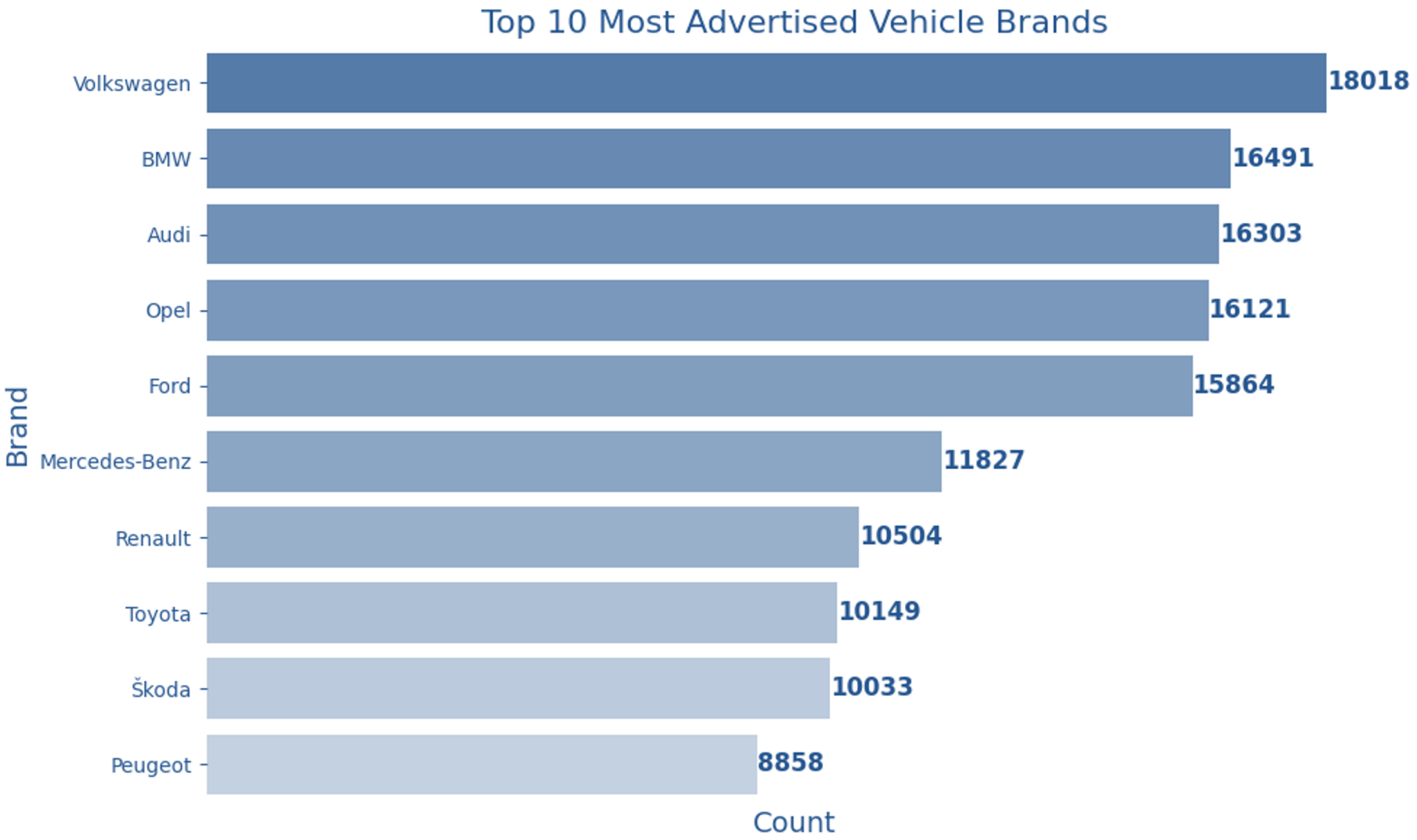


Engine Size

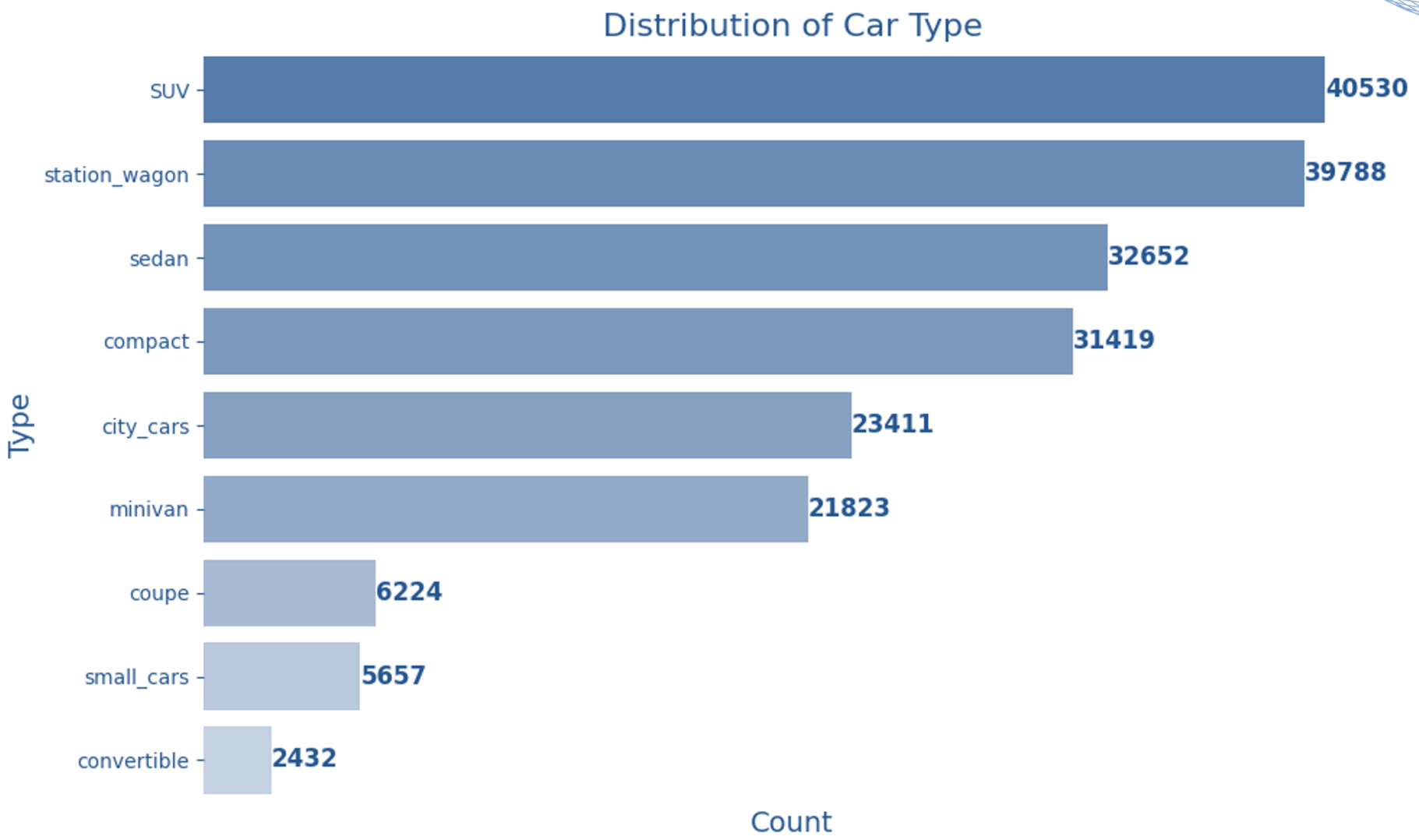


Age

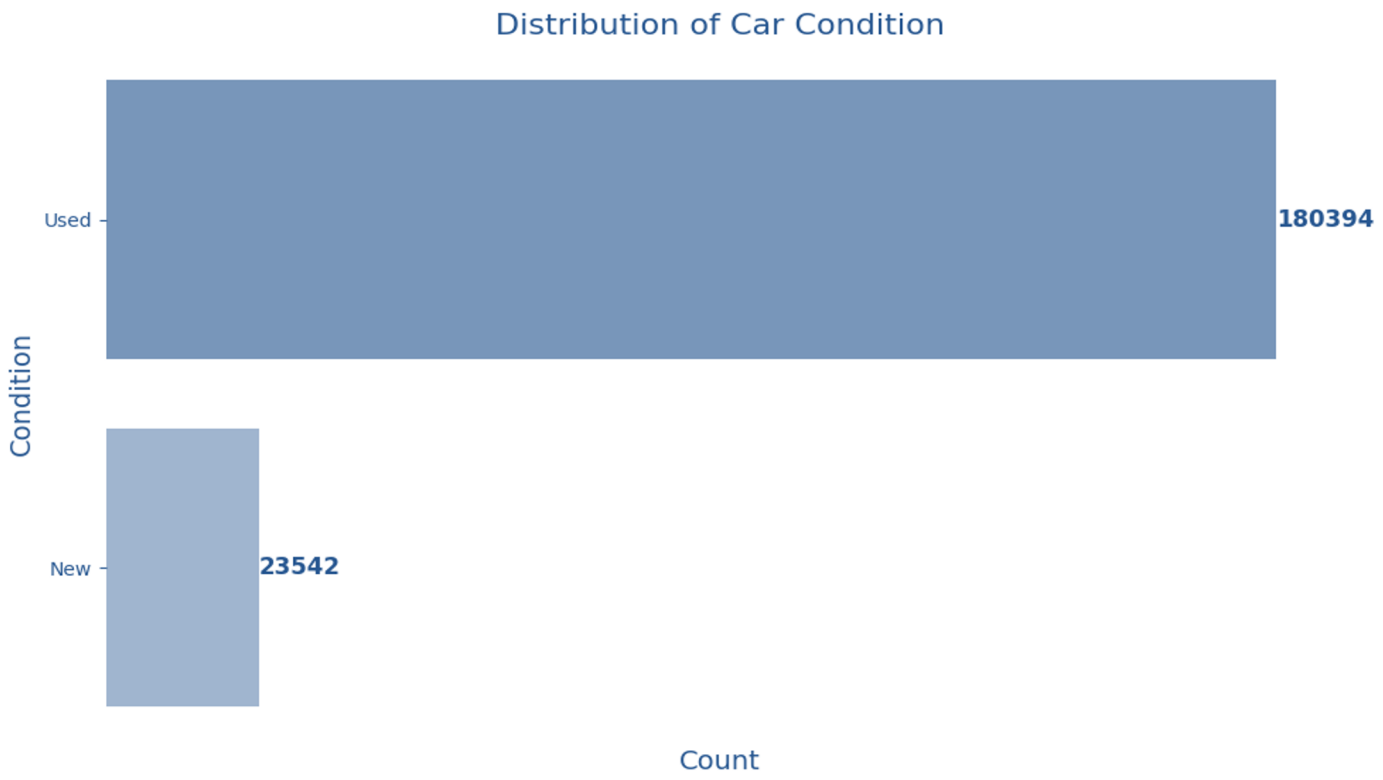
Volkswagen was the most advertised brand



SUVs were the most advertised



Most of the advertised vehicles were used



Vehicle Segmentation

Segmentation Process

1. Feature Selection
2. Algorithm Selection
3. Features Standardization
4. Finding Optimal Number of Clusters
5. Cluster visualization and Evaluation
6. Cluster Analysis
7. Business Impact

Six features were used for clustering

- Categorical features such as brand and model were excluded due to their high dimensionality
- Car condition was also not used, as most of the listings were for used cars
- Parsimony was prioritized
- Numeric features were standardized, and outliers were removed

Feature	Description
Price	The Vehicle listed price in CAD
Mileage	Vehicle mileage in Kms
Engine size	Vehicle displacement in cm^3
Power	Engine power in HP
Age	Vehicle age in years
Type	The only categorical feature and it captures the vehicle type, such as SUV, sedan, or city car

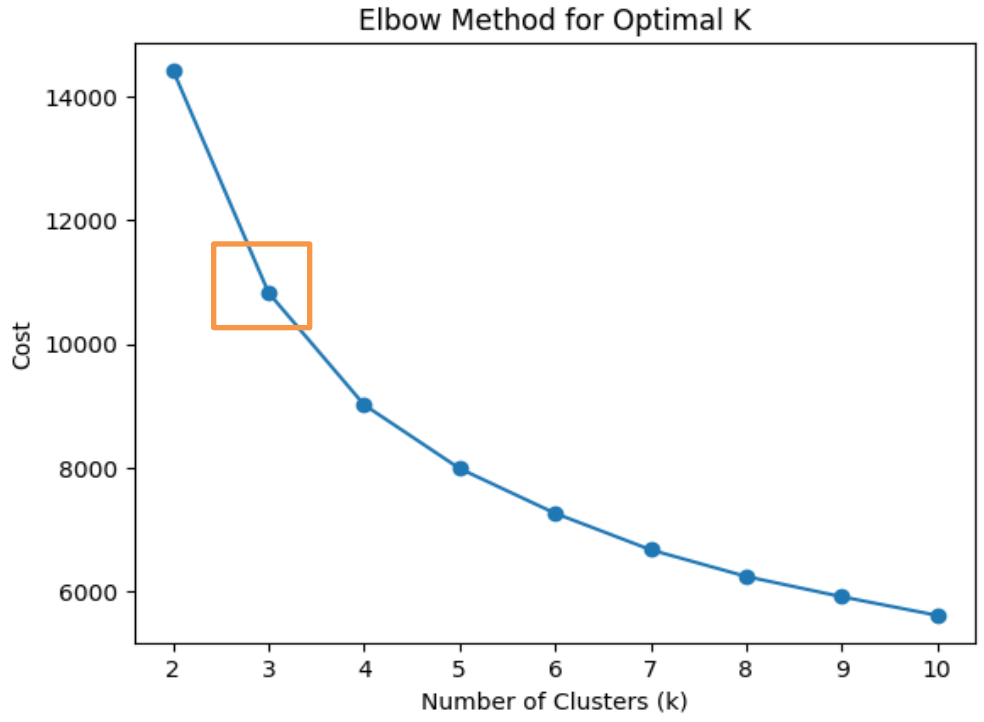
K-Prototypes was the selected clustering technique

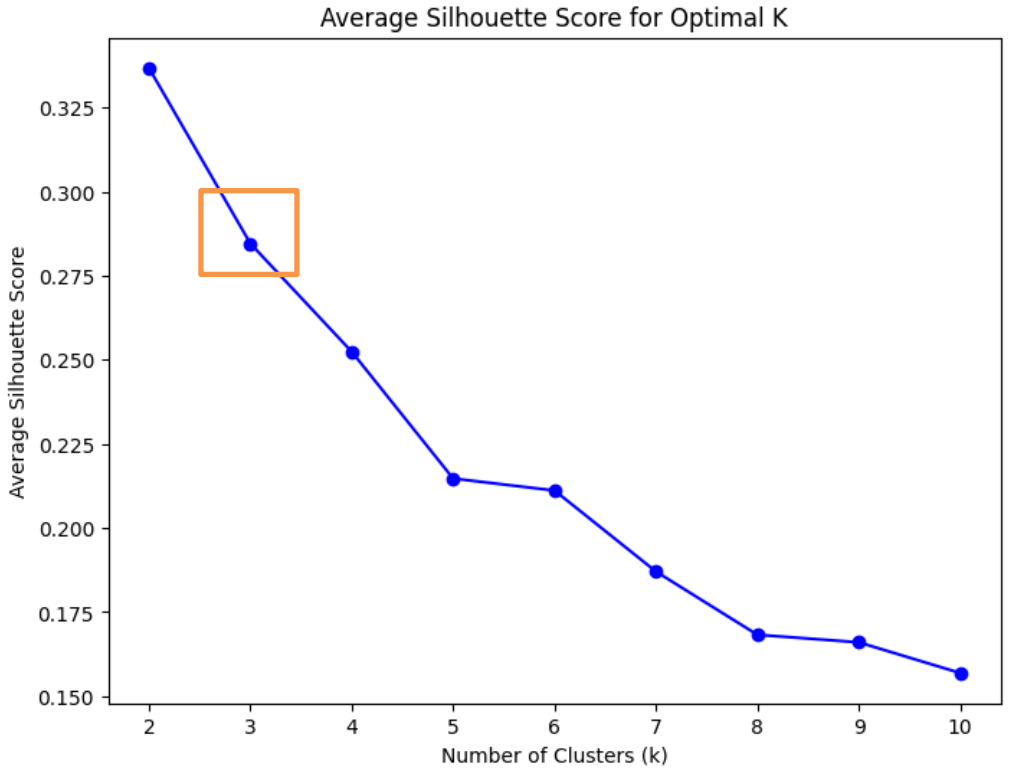
- Can handle both categorical and numeric features
- Easy to interpret even with mixed data types

How does it work?	Groups records into **K clusters** by minimizing within-cluster variation to enhance cohesion within each cluster
Hyperparameters	Number of Clusters (K): 3 Initialization Method: Huang
Dissimilarity Measure	Uses **Euclidean distance** for numeric features and **matching dissimilarity** for categorical features

Three was determined to be the optimal number of clusters (K)

- **Three** clusters achieved a good balance between cohesion and separation





Evaluation of clustering performance

Evaluating the clusters involved a combination of intuition and quantitative metrics:

- **Cost:** Measures the within-cluster variation; lower values indicate better cohesion. However, cost does not account for cluster separation and decreases monotonically as the number of clusters increases.

- **Silhouette Scores:** Evaluate both cluster separation and cohesion, providing a more balanced measure of clustering quality. Scores range from -1 to 1, where **1** indicates well-defined clusters, **0** suggests overlapping clusters, and **-1** reflects poor clustering

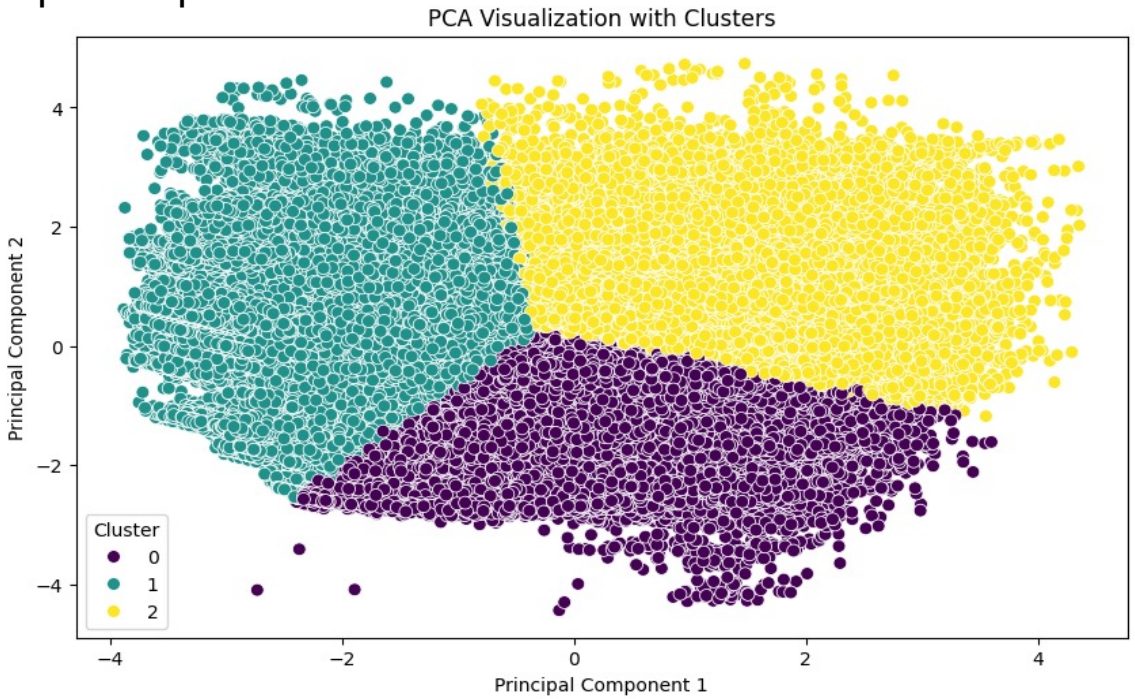
- **Visualization:** Enables the identification of overlaps and patterns

Average Silhouette Score

0.28

Cost

43,783



Cluster analysis

- **Older City Commuters (Cluster 0)**: compact, older cars, predominantly city cars, with low engine power, high mileage, and affordability
- **Modern Power Vehicles (Cluster 1)**: newer, higher-powered vehicles, mainly SUVs, with lower mileage and higher prices
- **Seasoned Family Travelers (Cluster 2)**: older, high-mileage vehicles, mostly station wagons, with larger engine sizes

	Price (CAD)	Age (Years)	Mileage	Engine Size (cm^3)	Power (HP)	Type
Older City Commuters	7,173	10	150,000	1,402	95	city car
Modern Power Vehicles	27,000	3	51,000	1,640	145	SUV
Seasoned Family Travelers	8,097	12	223,000	1,981	146	station wagon

How can insurers leverage the results?

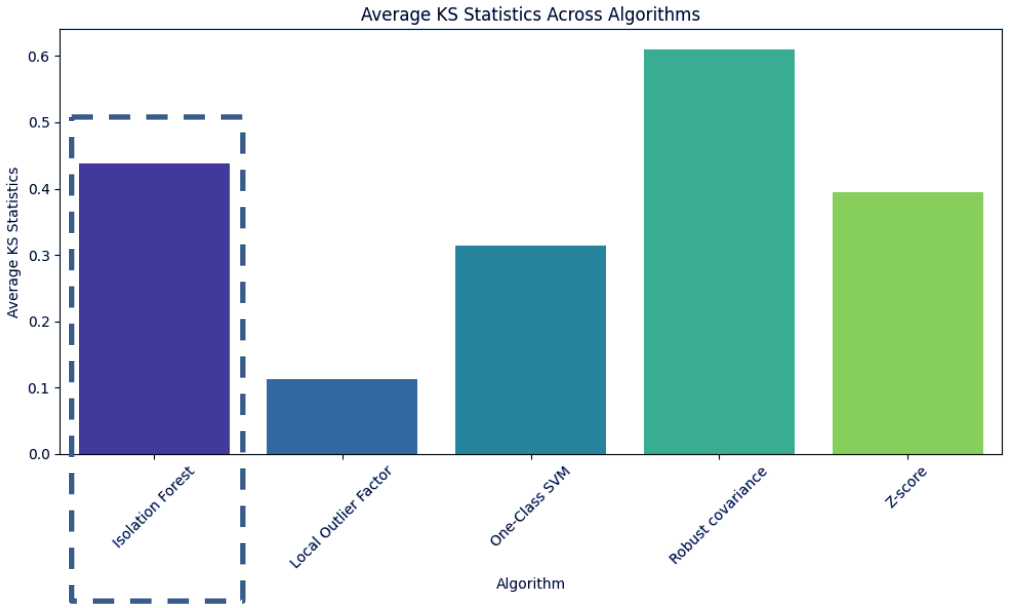
Definity can leverage insights from the identified clusters to optimize operations and customer engagement across three strategic areas:

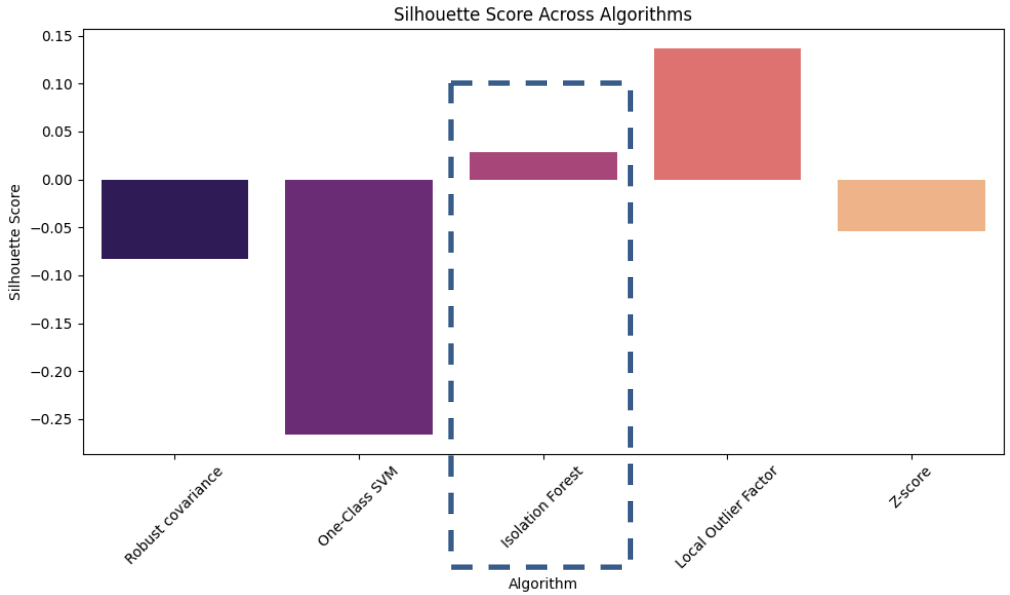
- Fraud prevention
- Underwriting and risk assessment
- Tailored coverage

Enhancing Vehicle Risk Assessment Through Anomaly Detection

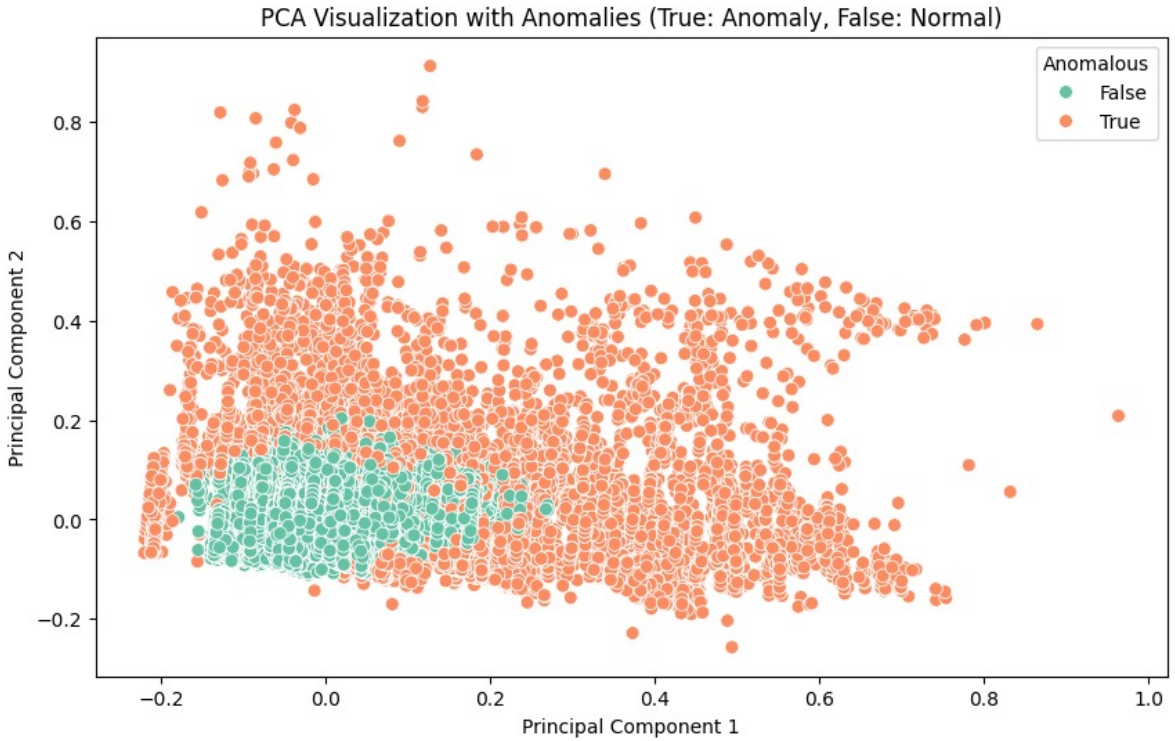
Isolation Forest was the selected anomaly detection technique

- Five techniques were tested, including **Robust Covariance**, **One-Class SVM**, **Z-Score**, and **Local Outlier Factor**
- **Isolation Forest** was deemed the best based on silhouette scores and results from the **Kolmogorov-Smirnov Test**
- **Contamination** was set to 10%

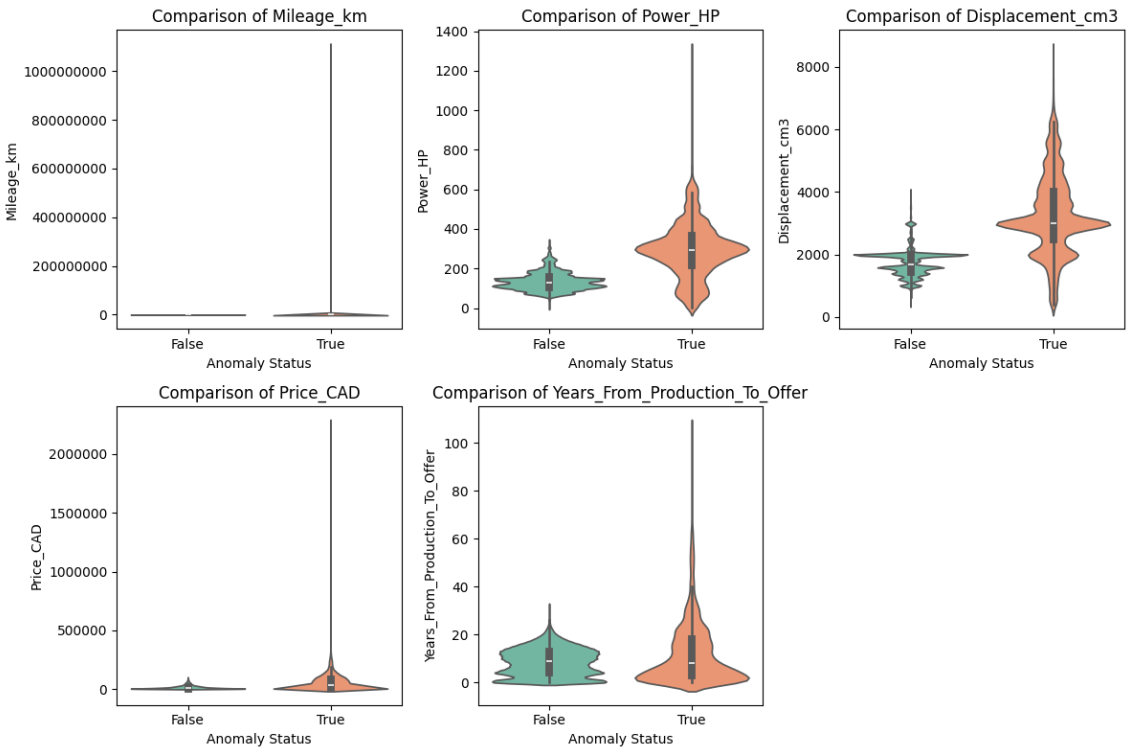




Comparison of anomalies and inliers distributions



Legend
- Anomalies
- Inliers



Comparison of anomalous and normal vehicle offers

Inliers

Feature	Mean	Std Dev	Median	Min	Max
Mileage (km)	140,098.25	93,314.31	149,035.00	1.00	2,930,000.00
Power (HP)	135.54	44.92	130.00	1.00	340.00
Engine Size (Displacement_cm3)	1,730.11	426.24	1,685.00	400.00	4,015.00
Price (CAD)	16,105.11	15,277.50	10,689.00	190.00	100,426.00
Age (Years)	8.60	5.66	9.00	0.00	32.00

Anomalies

Feature	Mean	Std Dev	Median	Min	Max
Mileage (km)	204,455.95	7,945,569.33	91,000.00	1.00	1,111,111,111.00
Power (HP)	291.99	135.22	292.00	1.00	1,300.00
Engine Size (Displacement_cm3)	3,254.33	1,257.29	2,993.00	400.00	8,400.00
Price (CAD)	55,512.87	65,629.16	32,165.00	211.00	2,273,975.00
Age (Years)	12.29	13.33	8.00	0.00	106.00

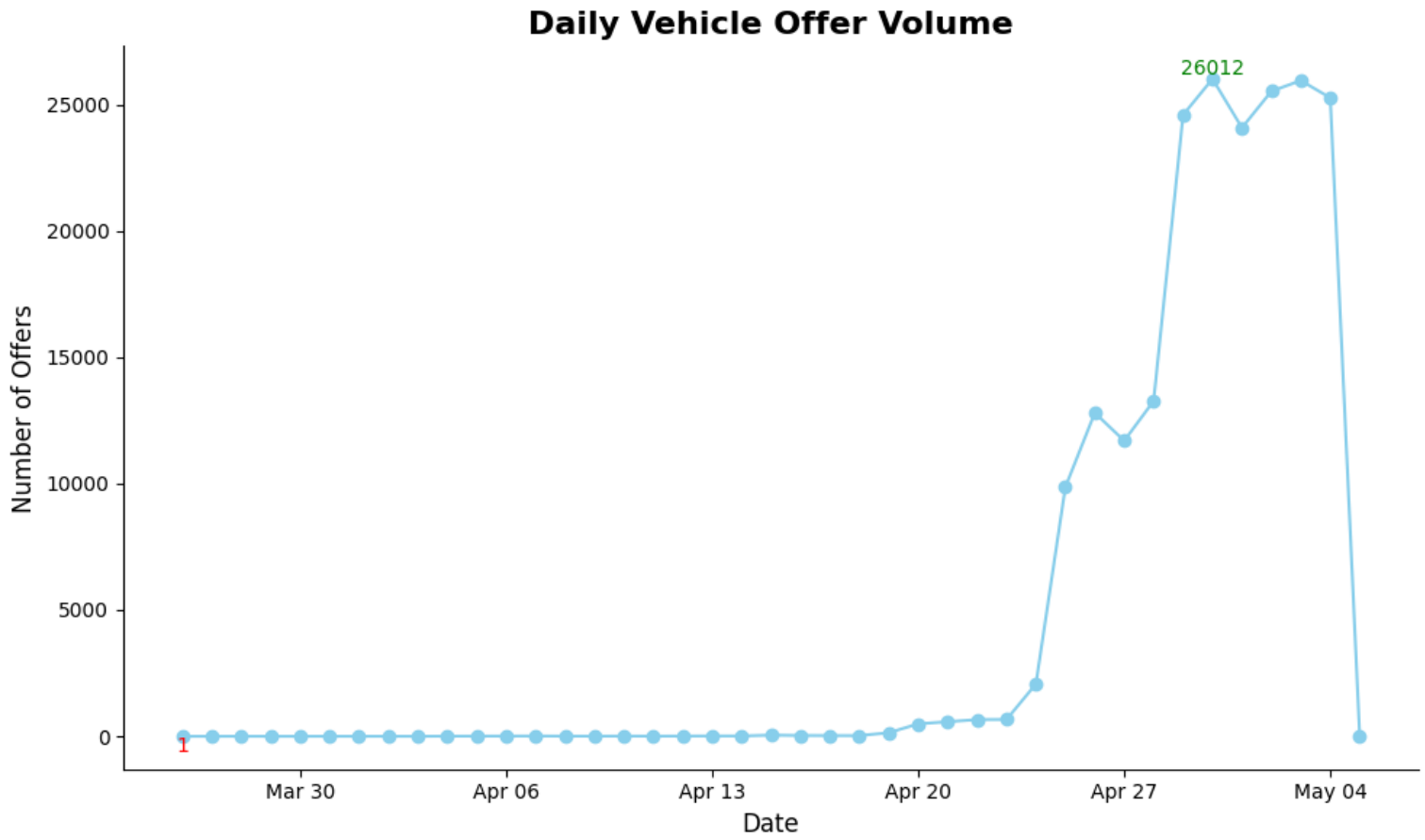
How can insurers leverage the results?

Definity can leverage the anomaly detection model to :

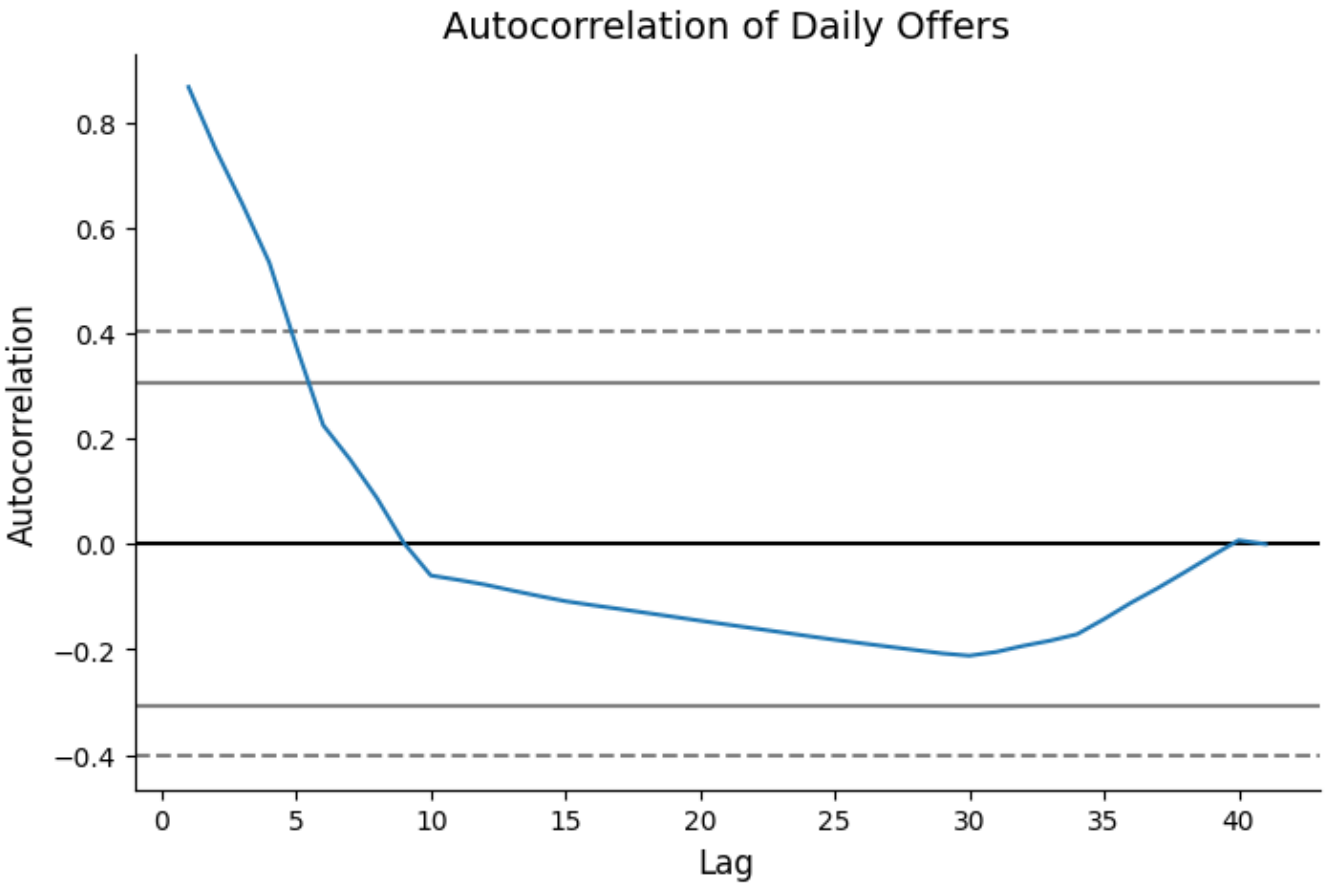
- Streamline underwriting and risk assessment
- Prevent Overinsurance and underinurance

Leveraging Temporal Patterns in Vehicle Offers to Streamline Marketing Strategies

Abrupt spike in offer volume



Correlation between offer volumes and previous days' offer volumes



How can insurers leverage the observed temporal pattern?

Definity can leverage the observed temporal pattern to:

- Strategically time marketing campaigns on online car sales platforms, ensuring optimal engagement and conversion rates

Next steps

- Gather data spanning a longer time period to enable a more comprehensive and robust time-based analysis
- Validate the selection of the number of clusters by consulting with domain experts to ensure alignment with industry knowledge
- Experiment with alternative clustering algorithms and incorporate different feature sets to explore diverse perspectives and improve results
- Confirm the chosen percentage of anomalies with domain experts to ensure the threshold reflects practical relevance and accuracy.

Thank you for listening!