K-Means Distance-Based Anomaly Detection Compared to Isolation Forest on Philippine Used-Car Listings

Zeus Morley S. Pineda
Torralba St., Brgy. Poblacion, Iligan City, Lanao Del Norte, 9200, Philippines
zeusmorley.pineda@g.msuiit.edu.ph

Abstract—Most studies use K-Means clustering with extras like a hybrid algorithm or algorithms with built-in outlier detection. No one has ever taken vanilla K-Means and used the Euclidean distance of entries to their centroids to flag outliers, then compared its performance against an algorithm like Isolation Forest on Philippine used-car listings.

The listings were scraped from two Philippine platforms (Philkotse and Carmudi), cleaned, and then went through unmodified K-Means. By calculating each car listing's distance from its centroid (K=1 or K=2), the outer 5 percent is flagged as an outlier. Isolation Forest (100 trees, contamination = 0.04) is used on the same dataset for a direct comparison. The goal is to compare the performance of K-means being repurposed for outlier detection against Isolation Forest, and what factors might affect the similarity between the results of both algorithms in terms of outlier detection.

Results show that in some datasets K-means was similar enough to the results of isolation forest, while some datasets made the result K-means be significantly different from the results of Isolation Forest.

Keywords: *Isolation Forest, K-means Clustering, Anomaly, Outlier*

I. INTRODUCTION

The car resale market is a big and fast-growing industry, with both physical stores and online platforms providing buyers and sellers a convenient way to connect and transact. However, a vast amount of car listings usually lead to irregularities in pricing, such as unusually high or low prices for vehicles of similar models and qualities. Identifying these irregularities is necessary for ensuring fair pricing.

Isolation Forest, a machine learning algorithm for detecting outliers, and K-means Clustering, an algorithm that groups the dataset into pre-defined and non-overlapping groups. It is usually used to group the dataset into clusters, but in this paper, it will be used for outlier or anomaly detection. Unlike Isolation Forest which is suited for outlier detection, K-means clustering here is repurposed to perform outlier detection. Both algorithms will detect outliers in car listing prices in terms of price, mileage, and age, and also all the cars on each set of the dataset are of the same car model.

This paper aims to help both car buyers and sellers in the car resale market by providing a method for detecting pricing anomalies and helping them make decisions. It will also show the comparison of the performance between both algorithms, one is for outlier detection, while the other is repurposed to detect outliers. Overall, it shows the potential of machine learning models in solving real-world problems.

II. REVIEW OF RELATED LITERATURE

2.1 Anomaly Detection

Anomaly detection refers to the recognition, identification, or flagging of entries, data, or patterns that deviate significantly from the majority [3].

2.2 K-means Clustering

2.2.1 Definition of K-Means Clustering

K-Means is an algorithm that starts with a chosen number of centers (centroids) and then iteratively assigns each data point to its nearest center, updating those centers until they no longer move significantly. In other words, K-Means splits a dataset into K groups so that points in the same group are as similar as possible in Euclidean space [4].

2.2.2 Application of K-Means Clustering

In Tabianan et. al. [5]. K-means was used on a large retail customer database, grouping customers with similar purchases.

In Park et. al. [6]. K-means was used on maritime waypoints to extract representative routes. By converting latitude/longitude into numerical feature vectors, they show how location‑based clustering can group together large geographic datasets.

2.3 Isolation Forest

2.3.1 Definition of Isolation Forest

Isolation Forest flags outliers by repeatedly picking a random feature and cutting that feature's range in two until each point stands alone in its own leaf node. If a data point is easy to isolate (i.e., it ends up in a leaf after just a few splits), the model gives it a high anomaly score. In practice, you build many such random trees, average each point's 'splitting depth,' and convert that into an anomaly likelihood. By design, anomalies tend to sit far from the main bulk and so get isolated quickly [7]..

2.3.2 Application of Isolation Forest

In Fang et. al. [8]. Shows how Isolation Forest isolates unusual sequences in business‑process logs by converting each event sequence into a feature vector and flagging short path‑length points.

In Geng et. al. [9]. Applies Isolation Forest with path‑weight tweaks to detect outliers in Lidar data, giving insight into tuning contamination and tree parameters.

In Binetti et. al. [10]. Demonstrates Isolation Forest's scalability on large satellite image feature sets and discusses choices for contamination and subsample size.

2.4 K-Means Clustering as an Outlier Detector
Gan [11] proposes a variant called KMOD that builds outlier detection directly into K-Means. During each clustering pass, KMOD penalizes the points that lie farthest from their cluster center, which effectively recognizes them as anomalies.

In Patel et. al. [12] used Z-score normalization followed by K-Means to isolate anomalies in transaction data. It is not the same as pure centroid-distance, but it shows how distance-based thresholds like the 95th percentile can be used on K-Means.

Statman et. al. [13] introduce a modification that simultaneously clusters and flags a fixed number of top-distance outliers. This provides an example of how researchers have altered K-Means to directly remove or flag anomalies

2.5 Anomaly Detection in the Car Industry

In Guerreiro et. al. [14]. Implemented K-Means and other clustering methods as well to flag irregular car component or part prices. Although focused on parts, its dataset characteristics and methodology for distance-based outlier scoring can be used for pricing anomaly detection in used-car prices.

Fortela et. al. [15] apply K-Means to fuel-economy features (mpg, emissions, age) to detect unusual vehicle behavior. The focus is on fuel and emissions, which shows how distance-based clustering can detect unusual vehicle behavior.

2.6 Summary of Gaps and Questions
K-Means and Isolation Forest each have a lot of applications from clustering customer segments or routes to flagging anomalies in business processes. However, every time K-Means shows up in anomaly work, it's been modified in some way like, KMOD's built-in outlier penalty. Similarly, Isolation Forest has been tuned and extended for high-dimensional data, but it's rarely compared against a truly "vanilla" clustering approach like K-means clustering.

III. THEORETICAL BACKGROUND

3.1 K-Means Clustering
K-Means is an unsupervised algorithm that partitions n data points into K non-overlapping clusters. Each cluster C_j is represented by its centroid μ_j, and the algorithm seeks to minimize the cluster sum of squares:

$$J = \sum_{j=1}^{k} \sum_{x_i \epsilon C_j} ||x_i - \mu_j||^2,$$

The method alternates between two steps:

1) **Assignment**: Assign each point x_i to the nearest centroid:
$$c(i) = arg\,min_{1\le j\le K} ||x_i - \mu_j||^2$$
2) **Update**: Recompute each centroid as the mean of its assigned points:
$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \epsilon C_j} x_i$$

These steps repeat until centroid movement is negligible. Since features such as price (in hundreds of thousands), mileage (in tens of thousands), and age (in years) are on different scales, all numeric features are standardized before clustering.

The value of K is determined by the elbow method: plotting total within-cluster variance versus K and selecting the point beyond which improvements diminish. In practice, the elbow almost always suggests K=1 or K=2 for these datasets.

3.2 Repurposing K-Means for Outlier Detection
Instead of using K-Means to group the listings into clusters, it can supply one or two points against which all data are compared. After centroids stabilize, each listing x_i has a Euclidean distance.

$$d_i = min_{1\le j\le K}||x_i - \mu_j||$$

The 95th-percentile $d_{0.95}$ of d1,…,dn is computed, and any listing with $d_i > d_{0.95}$ is flagged as an anomaly. When K=1, all distances refer to the single global centroid. When K=2, distances are computed to each listing's nearest of two centroids, and the top 5 % cutoff is applied within each cluster.

3.3 Isolation Forest
Isolation Forest isolates anomalies by constructing random binary trees. To build an isolation tree on a subsample of size m:

1) Randomly select a feature (e.g., price, mileage, or age).
2) Pick a split value uniformly between that feature's minimum and maximum.
3) Partition the data according to whether each point's feature value is below or above the split.
4) Recursively apply steps 1–3 until each point occupies its own leaf or a maximum depth is reached.

A data point x in a sparse region will be isolated after fewer splits, producing a short path length h(x). To score anomalies, IF builds T trees on random subsamples and averages each point's path length:

$$E[h(x)] = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

This average is converted into an anomaly score:
$$s(x) = 2 - \frac{E[h(x)]}{c(m)}$$

where c(m) is the expected path length of unsuccessful searches in a binary search tree of size m. Shorter E[h(x)] yields a higher ss(x), indicating a greater likelihood of being an outlier. In this study, T=100 trees and contamination = 0.04 (4 % expected anomalies) are used. Numeric features are scaled (e.g., to [0, 1]) so that no single attribute dominates random splits.

3.4 Feature Representation and Scaling
Only three numeric features are used: price, mileage, and age, so categorical encoding is unnecessary. All three are standardized to ensure comparable magnitudes. This normalization prevents the price's large range from dominating Euclidean distances in K-Means.

IV. METHODOLOGY

This section explains the process of applying the CRISP-DM framework in comparing Isolation Forest vs K-means Clustering in detecting anomalies in car resale prices.

A. Business Understanding

The primary goal of this paper is to compare the performance of Isolation Forest vs K-means clustering in identifying anomalies in the prices of car resale listings of various models, which may be classified as either overpriced or underpriced listings.

Detecting these anomalies will assist both car buyers and sellers to make better decisions and also help the market by having fair prices. Buyers can easily determine which car listings are worth checking on since they are cheaper than usual, and which cars to avoid since they are overpriced. Sellers can also adjust their asking price by determining whether they are asking more or less for what they are offering.

B. Data Understanding

The dataset for this paper was scraped from online car resale listing platforms from the Philippines only, Philkotse [1] and Carmudi [2], then including features like:

- Model (ex: Fortuner, Civic)
- Car Age (Current year - Model year)
- Mileage (ex: 2,000 km)
- Price (P 250,000)

During this phase, the data were explored through Exploratory Data Analysis (EDA) to identify distributions, correlations, and data quality issues (missing values). Scatter plots and 3D visualizations were created to visualize the relationship between Age, Mileage, and Price across different car models.

C. Data Preparation

The raw data underwent the following preprocessing steps:

1) **Feature Engineering**:
The age of the car was computed as:
Current Year - Model Year
2) **Cleaning**: (during scraping)
Commas, symbols and units(“₱”,”km”) were removed from the Price and Mileage columns.

3) **Handling Duplicates**: (during scraping)
Sometimes the same car will be listed multiple times on the same website, maybe to increase its exposure. Therefore during scraping, listings with the same Age, Mileage, and Price are omitted into just one.

4) **Anomaly Handling**:
Both algorithms are tested in their performance of anomaly or outlier detection, therefore any outliers detected by both algorithms will not be removed but rather listed or recorded.

D. Modeling

There are two Unsupervised Learning models used which are the Isolation Forest and K-means clustering.

The Isolation Forest algorithm was configured with 100 estimators and a contamination level of 4%, which assumes that around 4% of the data points are anomalies. The Isolation Forest assigned two outputs for each listing:

1) **Anomaly Score**: Indicates how anomalous the listing is.
2) **Anomaly Label**: Categorizes listings as either normal (1) or anomalous (-1).

The K-means clustering algorithm was usually configured to work with two clusters, and in terms of outlier detection, it is set so that if a data point is outside the 95% or beyond that percentile from their respective centroids or their mean, then it will be marked as an outlier.

E. Evaluation

The outputs of both models were analyzed to determine their differences in terms of how many data points in the same dataset they marked as anomalous, and how many data points are marked by both, and by each of them only.

For the Isolation forest, it also labels whether the anomaly it detected was either overpriced or underpriced, but the focus of this paper is comparing the performance and how similar the results of both algorithms are.

F. Deployment

The results were summarized and visualized through tables and side-by-side graphs to enable interpretation and easier comparison between similar graphs. A table of all the anomalies was also generated to display their details and if they were either detected by both, or one of the algorithms only.

The findings can be used both by the online platforms to recognize potential pricing errors or absurd pricing. For the buyers, it can be a tool for them to narrow down their selection, by avoiding the overpriced listings and looking further into the underpriced listings. For the sellers, they can assess whether they are asking too low or too high based on the condition of the car that they are offering.

V. RESULTS AND DISCUSSION

This section discusses the results of the scatter plots, 3D graphs, and anomaly distribution.

There were 12 different car models in the dataset, mainly from Toyota and Honda. The models are Honda: Civic, City, CR-V, and BR-V, and Toyota; Avanza, Fortuner, Hiace, Hilux, Innova, Rush, Vios, and Wigo.

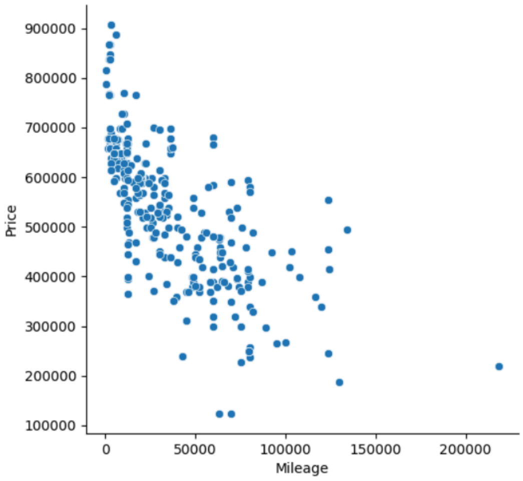


Figure 1. Toyota Vios Mileage vs Price

This scatter plot shows the distribution of the samples for Toyota Vios in terms of Mileage and Price. The price is in pesos and the mileage is in kilometers.

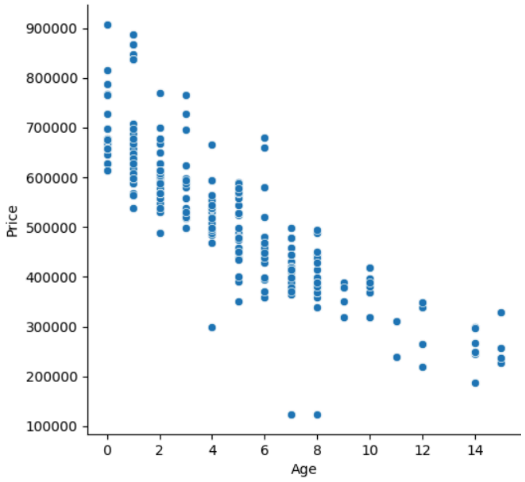


Figure 2. Toyota Vios Age vs Price

This scatter plot shows the distribution of the samples for Toyota Vios in terms of Age and Price. The price is in pesos and age in years.

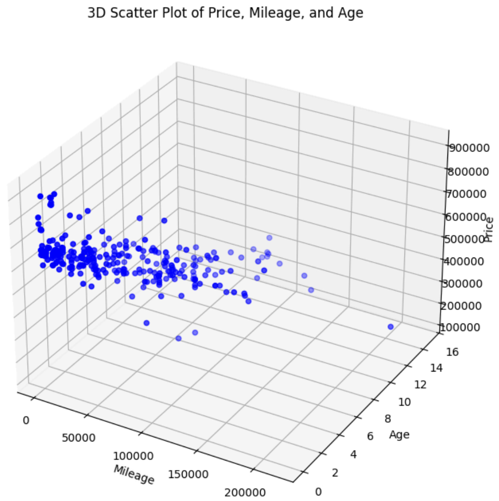


Figure 3. Toyota Vios Age vs Price vs Mileage

This 3D graph shows the distribution of the samples for Toyota Vios in terms of Age, Price, and Mileage.

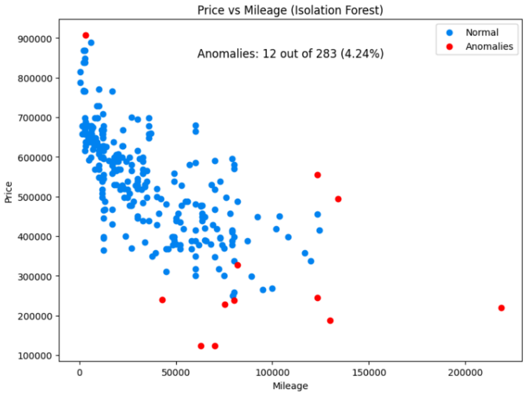


Figure 4. Anomalies in Toyota Vios Price vs Mileage using Isolation Forest

This scatter plot shows the distribution of the samples for Toyota Vios with the anomalies in terms of Mileage and Price using the Isolation Forest algorithm. The price is in pesos and the mileage is in kilometers.

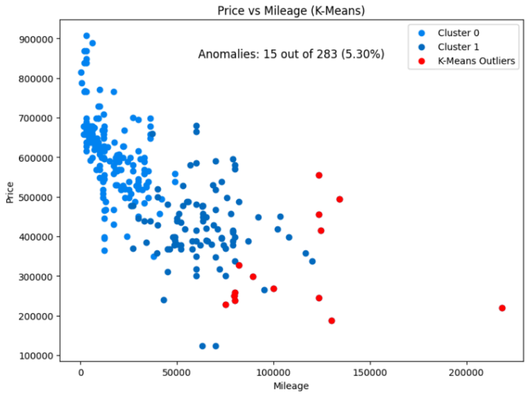


Figure 5. Anomalies in Toyota Vios Price vs Mileage using K-means Clustering

This scatter plot shows the distribution of the samples for Toyota Vios with the anomalies in terms of Mileage and Price using the K-means Clustering algorithm. The price is in pesos and the mileage is in kilometers.

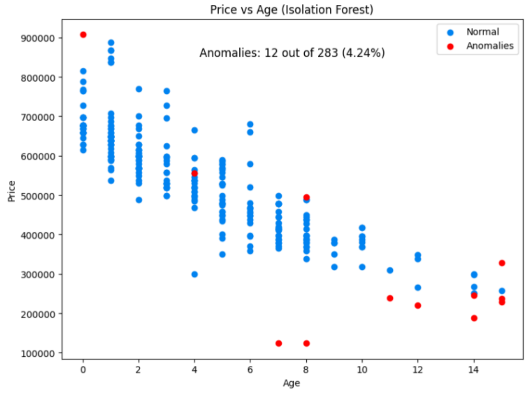


Figure 6. Anomalies in Toyota Vios Price vs Age using Isolation Forest

This scatter plot shows the distribution of the samples for Toyota Vios with the anomalies in terms of Age and Price using the Isolation Forest algorithm. The price is in pesos and age in years.

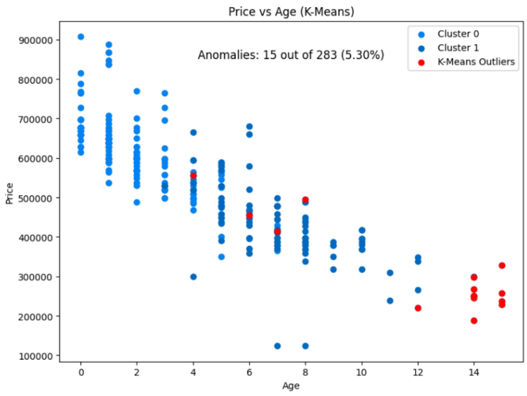


Figure 7. Anomalies in Toyota Vios Price vs Age using K-means clustering

This scatter plot shows the distribution of the samples for Toyota Vios with the anomalies in terms of Age and Price using the K-means clustering algorithm. The price is in pesos and age in years.

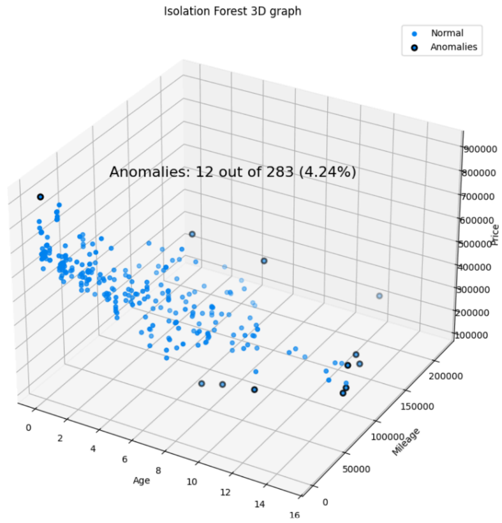


Figure 8. Anomalies Toyota Vios in Age vs Price vs Mileage using Isolation Forest

This 3D graph shows the distribution of the samples for Toyota Vios in terms of Age, Price, and Mileage using Isolation Forest. It also shows that it detected that 4.24% of the samples of Toyota Vios as anomalous.

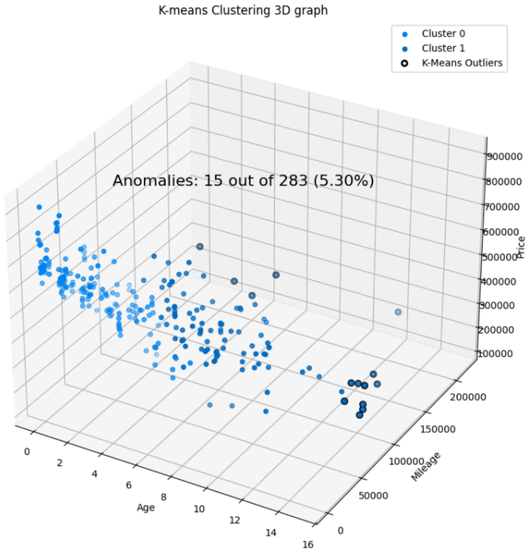
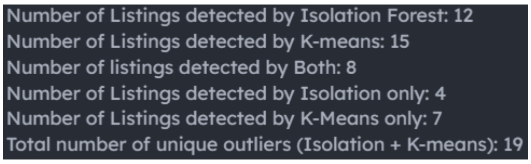


Figure 9. Anomalies Toyota Vios in Age vs Price vs Mileage using K-means Clustering

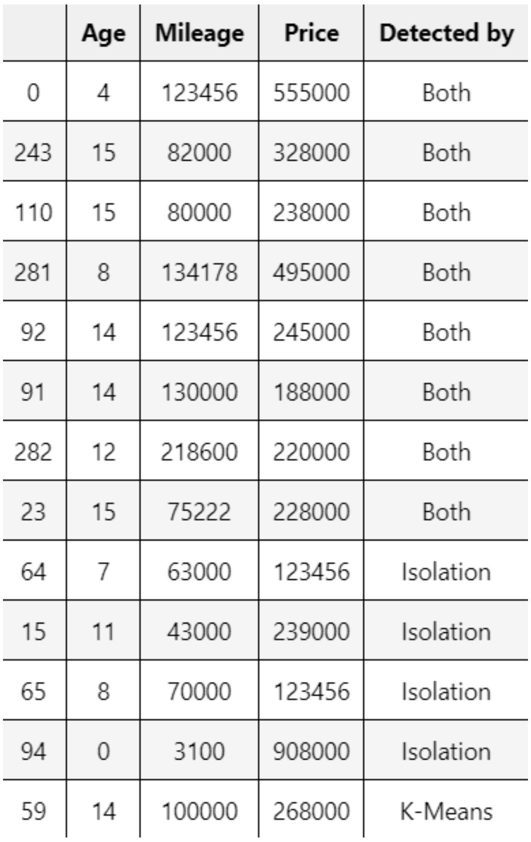
This 3D graph shows the distribution of the samples for Toyota Vios in terms of Age, Price, and Mileage using K-means Clustering. It also detected that 5.30% of the samples of Toyota Vios as anomalous.



```
Number of Listings detected by Isolation Forest: 12
Number of Listings detected by K-means: 15
Number of listings detected by Both: 8
Number of Listings detected by Isolation only: 4
Number of Listings detected by K-Means only: 7
Total number of unique outliers (Isolation + K-means): 19
```

Figure 10. Summary of the detected Anomalies for Toyota Vios

This shows the summary of the anomalous listings detected, and whether each anomalous listing was detected by both or only one of the algorithms.



	Age	Mileage	Price	Detected by
0	4	123456	555000	Both
243	15	82000	328000	Both
110	15	80000	238000	Both
281	8	134178	495000	Both
92	14	123456	245000	Both
91	14	130000	188000	Both
282	12	218600	220000	Both
23	15	75222	228000	Both
64	7	63000	123456	Isolation
15	11	43000	239000	Isolation
65	8	70000	123456	Isolation
94	0	3100	908000	Isolation
59	14	100000	268000	K-Means

Table 1. List of Anomalous listings

This table shows some of the anomalous listings, with their age, mileage, price, and whether they were detected by both algorithms or one of them only.

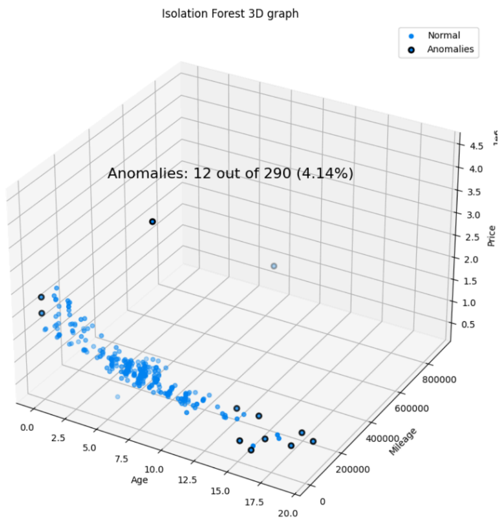


Figure 11. Anomalies Toyota Fortuner in Age vs Price vs Mileage using Isolation Forest

This 3D graph shows the distribution of the samples for Toyota Fortuner in terms of Age, Price, and Mileage using Isolation Forest. It also shows that 4.14% of the samples of Toyota Fortuner are detected as anomalous.

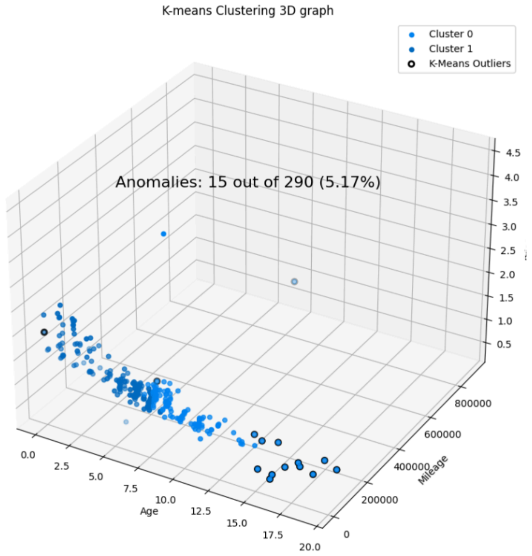
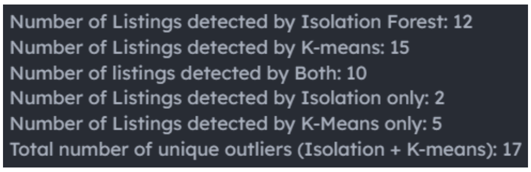


Figure 12. Anomalies Toyota Fortuner in Age vs Price vs Mileage using K-means Clustering

This 3D graph shows the distribution of the samples for Toyota Fortuner in terms of Age, Price, and Mileage using K-means Clustering. It also shows that 5.17% of the samples of Toyota Fortuner are detected as anomalous.



```
Number of Listings detected by Isolation Forest: 12
Number of Listings detected by K-means: 15
Number of listings detected by Both: 10
Number of Listings detected by Isolation only: 2
Number of Listings detected by K-Means only: 5
Total number of unique outliers (Isolation + K-means): 17
```

Figure 11. Summary of the detected Anomalies for Toyota Fortuner

This shows the summary of the anomalous listings detected, and whether each anomalous listing was detected by both or only one of the algorithms.

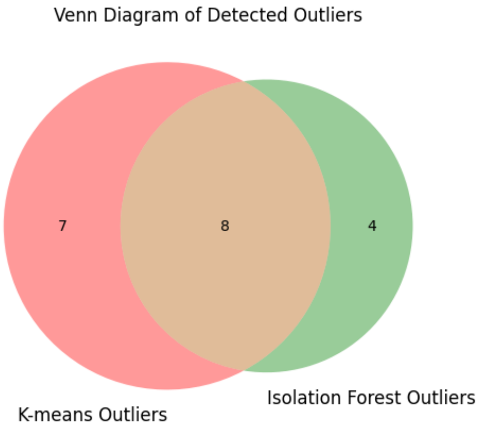


Figure 12. Venn Diagram for the detected outliers for Toyota Vios

This Venn diagram shows how many outliers are detected by both algorithms and how many are detected by one of them only.

Applying McNemar's test, which evaluates whether the results of both differ significantly in the Toyota Fortuner dataset with (N=290):

$$x^2 = \frac{(b-c)^2}{(b+c)} \text{ with b=5 (K-Means only), c=2 (IF only)}$$

$$x^2 = \frac{(5-2)^2}{5+2} = \frac{9}{7} \approx 1.29, \text{ degrees of freedom = 1}$$

$$p \approx 0.26$$

Therefore, since p>0.05, there is no significant difference between K-Means and Isolation Forest in the number of Fortuner listings flagged.

The results show that in some datasets, like for the Toyota Fortuner, where the data points are close together, K-means clustering performed better or is more similar to the Isolation Forest because out of the 12 total anomalous listings detected by Isolation Forest, it also detected 10 of them and it detected 5 more listings which Isolation Forest did not detect as anomalous. Unlike the dataset for the Toyota Vios, where the data points were more spread out, K-means clustering performed worse because out of the 12 listings detected by Isolation Forest, it only detected 8 of them and it also detected 7 more listings which Isolation Forest did not detect as anomalous.

VI. CONCLUSION

The results show the similarity of the modified K-means clustering in terms of anomaly detection and by using Isolation Forest as the benchmark or comparison since it is designed for anomaly or outlier detection, unlike K-means clustering which is for grouping the dataset into clusters but in this paper was modified so that it will detect outliers by marking data points in each cluster that are very far from the centroid or mean.

The results of both algorithms show which listings are cheap for what it offers or which ones are expensive for what it is offering. This study only uses age, mileage, and price as it is only the most common factors that are available on car listings online. The anomalous listings detected that are on the lower area of the graph can be classified as underpriced, while the ones that are on the upper area of the graph can be classified as overpriced.

The similarity of the results of K-means clustering from Isolation Forest can vary depending on the shape of the dataset, it is observed that it is more similar to Isolation Forest when the data points are closer together, and it is less similar when the data points are more spread out.

Other factors that might significantly affect the price but are not always accurately mentioned are condition, whether it has been through an accident, and whether there were parts that were replaced or upgraded.

Overall, anomaly detection in car resale prices is still helpful for both the buyers and sellers, as the buyers can narrow down their selection into the overpriced or underpriced ones so they can ask the sellers themselves for more information that might justify its higher or lower than the average price. The sellers can also use this to gauge their standing amongst their competitors, whether the value they offer is good or bad.

DECLARATION OF GENERATIVE AI SOFTWARE TOOLS IN THE WRITING PROCESS

During the preparation of this work, the author, Zeus Morley S. Pineda used ChatGPT to assist with the format, the outline, and guide questions of some parts. Grammarly Firefox extension for the readability and accuracy of the document. The author reviewed and edited all content derived from these services and takes full responsibility for the content of this publication.

ACKNOWLEDGEMENT

I acknowledge myself as the author of this paper for the effort that I have exerted. I also acknowledge the software namely: ChatGPT and Grammarly, that I used in guiding me throughout this whole process of creating not only this paper but the project. I also acknowledge the websites that I used to gather my dataset.

References

[1] "Used Cars for sale in the Philippines." Philkotse. [Online]. Available: https://philkotse.com/used-cars-for-sale (accessed Dec. 14, 2024).

[2] "Second Hand Cars for Sale in Philippines." Carmudi. [Online]. Available: https://www.carmudi.com.ph/used-cars/ (accessed Dec. 14, 2024).

[3] R. Al-amri, R. K. Murugesan, M. Man, A. F. Abdulateef, M. A. Al-Sharafi, and A. A. Alkahtani, "A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data," Applied Sciences, vol. 11, no. 12, p. 5320, Jan. 2021, doi: https://doi.org/10.3390/app11125320.

[4] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: https://doi.org/10.3390/electronics9081295.

[5] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, Available: https://www.mdpi.com/2071-1050/14/12/7243

[6] J. Park and M. Choi, "A K-Means Clustering Algorithm to Determine Representative Operational Profiles of a Ship Using AIS Data," *Journal of Marine Science and Engineering*, vol. 10, no. 9, p. 1245, Sep. 2022, doi: https://doi.org/10.3390/jmse10091245.

[7] W. Chua *et al.*, "Web Traffic Anomaly Detection Using Isolation Forest," *Informatics*, vol. 11, no. 4, pp. 83–83, Nov. 2024, doi: https://doi.org/10.3390/informatics11040083.

[8] N. Fang, X. Fang, and K. Lu, "Anomalous Behavior Detection Based on the Isolation Forest Model with Multiple Perspective Business Processes," *Electronics*, vol. 11, no. 21, pp. 3640–3640, Nov. 2022, doi: https://doi.org/10.3390/electronics11213640.

[9] G. Geng, P. Wang, L. Sun, and H. Wen, "Enhanced Isolation Forest-Based Algorithm for Unsupervised Anomaly Detection in Lidar SLAM Localization," *World Electric Vehicle Journal*, vol. 16, no. 4, p. 209, Apr. 2025, doi: https://doi.org/10.3390/wevj16040209.

[10] M. S. Binetti, V. F. Uricchio, and C. Massarelli, "Isolation Forest for Environmental Monitoring: A Data-Driven Approach to Land Management," *Environments*, vol. 12, no. 4, pp. 116–116, Apr. 2025, doi: https://doi.org/10.3390/environments12040116.

[11] G. Gan, "A k-Means Algorithm with Automatic Outlier Detection," *Electronics*, vol. 14, no. 9, p. 1723, Apr. 2025, doi: https://doi.org/10.3390/electronics14091723.

[12] J. Patel, J. Reiner, B. Stilwell, A. Wahbeh, and R. Seetan, "Leveraging K-Means Clustering and Z-Score for Anomaly Detection in Bitcoin Transactions," *Informatics*, vol. 12, no. 2, p. 43, Apr. 2025, doi: https://doi.org/10.3390/informatics12020043.

[13] A. Statman, L. Rozenberg, and D. Feldman, "k-Means: Outliers-Resistant Clustering+++," *Algorithms*, vol. 13, no. 12, p. 311, Nov. 2020, doi: https://doi.org/10.3390/a13120311.

[14] M. Guerreiro *et al.*, "Anomaly Detection in Automotive Industry Using Clustering Methods—A Case Study," *Applied sciences*, vol. 11, no. 21, pp. 9868–9868, Oct. 2021, doi: https://doi.org/10.3390/app11219868.

[15] D. L. B. Fortela *et al.*, "Unsupervised Machine Learning to Detect Impending Anomalies in Testing of Fuel Economy and Emissions of Light-Duty Vehicles," *Clean Technologies*, vol. 5, no. 1, pp. 418–435, Mar. 2023, doi: https://doi.org/10.3390/cleantechnol5010021.