A Visual Analysis of Used Car Features
INFO 633 Project C

Katy Matulay
Drexel University

Abstract – The US used car market represents a multi-billion-dollar industry that continues to grow because of supply chain pressures, inventory shortages, and economic uncertainties brought about by the COVID-19 pandemic. The overall goal of this project is to examine a dataset scraped in 2020 from CarGurus inventory website in order to visualize and explore used car features that influence price, and uncover patterns, relationships, and/or trends. Multiple visualization tools such as Tableau and Orange are used to create a broad array of information visualizations to examine trends, relationships, and patterns. Discussions of the influence of various features on price, relationships, trends (or lack thereof), and visualization design decisions are undertaken. The major influencers of price are determined to be *mileage*, *horsepower*, *owner count*, and *year*. A conclusion is reached that no single visualization or feature can fully explain how price is determined, and that a broad multi-dimensional array of features must be considered in determining used car price. However, it was found that a single binary indicator representing if a car is new or used can be used to visually segregate distinct clusters along a clear margin using t-SNE.

Keywords – used cars, automotive industry, CarGurus, pricing algorithms, car price, sales, Tableau, Orange, word clouds, EDA, machine learning.

1. INTRODUCTION

Used car sales represent a significant market in the United States with an estimated 41 million used cars sold annually as of 2020 [1]. From 2019 to 2021, the total value of the used car market in the U.S. grew from $134 billion to $138 billion [2]. Increasing supply chain pressures, rising costs, delays, and inventory challenges brought about by the COVID-19 pandemic are predicted to accelerate demand for used cars. Even before the pandemic, the used car market was estimated at more than twice the size of the new car market [1].

There are multiple avenues for used car sales: trade-in at a franchise dealership, used car dealerships, online auctions, and most commonly through private sales. Private used car sales can occur on platforms like Craigslist or Facebook Marketplace or be facilitated by online car marketplaces such as Autotrader, CarGurus, Carvana, and CarMax. Platforms like Kelly Blue Book, Edmunds, and CarGurus provide used car pricing tools that use unique algorithms to compare pricing from different dealerships and factor in various features and attributes of the car to determine a range of fair retail prices [3]. Understanding the impact of features like year, make, model, mileage, transmission, and condition can help sellers better value their car and buyers be more informed when buying a used car.

1.1. Purpose and Goals

The purpose of this study will be to visually uncover trends, patterns, and relationships between used car listing features and sales price using a dataset scraped from CarGurus inventory in September 2020 [4]. The aim is to create a limited scope buyers' and sellers' visual guide to the U.S. used car market, to facilitate understanding, knowledge acquisition, and decision making via information visualizations.

2. OVERVIEW

2.1. Dataset Description

The main dataset used for this study comes from Kaggle.com and was published in September 2020 using a snapshot of data scraped from the CarGurus.com inventory at that time [4]. An additional dataset containing U.S. zip code and state pairs was obtained to expand upon the dealer zip code column contained in the CarGurus dataset [5]. CarGurus was founded in 2006 as a blog and messaging site, but now focuses almost exclusively on classified listings and connecting buyers with sellers for any of their over 5 million listings for new, certified pre-owned, and used cars [3].

The original dataset contains over 3 million records (9.98 Gb) and contains 66 columns. Each row of the dataset represents 1 unique listing at that point in time, denoted by the unique primary key column *VIN (Vehicle Identification Number)*. Because of the high dimensionality of the dataset, a data dictionary has been included in the Appendix.

The features of the dataset relevant to this study include: *price, mileage, year, make, model, body type, horse power, dealer zip code, longitude, latitude, is new, owner count, description, days on market, listing date, average fuel economy*, and *fuel type*.

2.2. Literature Review

One key application of this dataset is in developing machine learning algorithms to predict price, as accurately determining fair market value in the automotive industry is a crucial business problem to be solved [6]. As such, many manufacturers have already integrated machine learning into their price prediction tools [7]. However, no published studies of this nature using this dataset could be found, so literature using similar datasets was examined.

One study in particular found that predicting a car price is challenging due to the high number of attributes that must be taken into account, amount of data cleaning necessary, and resource requirements [8]. Much work has been done in identifying key features that affect a car's resale value, such as make, body style, mileage, and accident history [9]. Several different pricing tools exist, but each uses a confidential and proprietary algorithm and as a buyer or seller it may not be clear when using these tools which features impact the price the most. Machine learning models such as random forest model and K-Means clustering using 3 clusters with linear regression have produced the most accurate price predictions in a study using a similar US car dataset [10].

2.3. Tools

To perform initial data cleansing, exploratory data analysis, natural language processing, and feature engineering, a Jupyter Notebook using Python with various libraries (pandas, numpy, matplotlib, nltk, seaborn) was used. The software Orange [11] was used to perform unsupervised learning t-SNE and K-means visualizations, cluster analysis, and text mining. Tableau Desktop [12], a free visualization platform with an academic license was used to create all other visualizations except Sankey Diagrams, as this visualization was only available on Tableau Public for a limited time [13]. Lastly, to create text visualizations, the free platform *wordart.com* was used, as it had robust built in text-processing features, unique output shapes, and accepted and processed up to 10k of tokenized words in csv format [14].

3. METHODOLOGY

3.1. Data Cleansing and Feature Engineering

Using python and various libraries in a Jupyter notebook, a random sample of the dataset (900k rows) was first cleaned to strip string characters from numeric features, drop nulls, and perform exploratory data analysis (EDA). Based on findings from EDA, 38 of the original 66 columns were retained and 2 derived features were added to calculate average fuel economy and consolidate owner counts from 12 to 3 ordinal categories (0,1, and 2 to represent 2+ owners). Additional manipulation was done using Python and nltk to generate a tokenized word dataset from the *description* field, which is further described in Section 4.7.1. A correlation matrix heatmap was generated using Seaborn [15] to find informative features and can be found in the Appendix. Lastly, due to size and resource constraints of various software used, a smaller subset of 100k rows (0.3 Gb) was randomly sampled and exported as a *csv* for use in Tableau and Orange.

3.2. Motivating Questions

The primary motivating question underlying this dataset is to understand what determines the price at which a used car is listed for sale, given a set of internal features that describes the car, and external features such as its geographic location, point of sale, and date of listing. Given the fact that CarGurus computes an Instant Market Value (IMV) based upon features such as make, model, trim, year, mileage, options,

region, and vehicle history—these features as much as possible will be used as starting points to inform the analysis [16]. Thus, the report reads like a buyers' or sellers' visual guide to the used car market—what feature(s) and characteristic(s) of cars command a premium or alternatively negatively influence price, what are the overall trends, and what if any interesting patterns can be found in the data.

3.3. Exploratory Data Analysis

To determine which features to focus on, a Sankey Diagram was chosen to examine the data distribution and flows, as it provided a quick snapshot of the predominant features by number of vehicles (**Figure 1)**. First, it showed that the dataset is nearly split 50/50 between new and used cars, which was surprising given that the dataset is labeled as a 'Used Car Dataset', but based on CarGurus own website description, this aligns with what types of vehicles can be listed for sale on the platform [4]. Because the focus of this analysis is on the used car market, this subset based on the *is new* field was used to filter in Tableau. Breaking down by body type, we can deduce that SUVs represent the most frequent body type and most often are AWD and automatic transmissions.

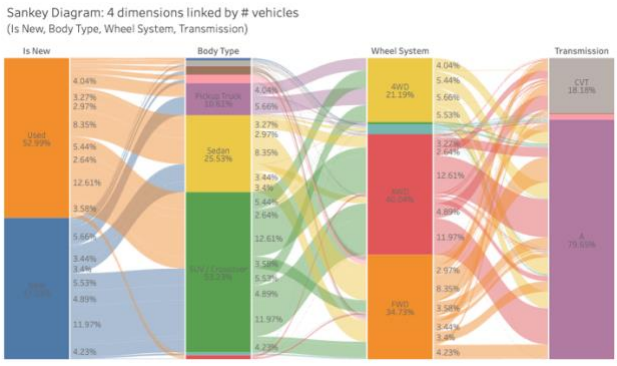

***Figure 1**: Sankey Diagram shows the flows and distributions of # vehicles by new/used car, body type, wheel system, and transmission.*

Given that the dataset is split nearly evenly by used and new cars, it was critical to examine the price ranges. The original dataset spanned a large range of $650--$3.2 million USD; however, during EDA it was found that over 99% of the vehicles are at or below $80k. In order to utilize gestalt principles of similarity and proximity in the visualizations and eliminate outliers that would skew the analysis, this subset of price was retained for ease of visualizing patterns and facilitating understanding for the average car buyer/seller.

The average price for used cars is $19k, while new cars are $32k, which explains the rightward shift in frequency of prices seen in the **Figure 2** below for new cars.

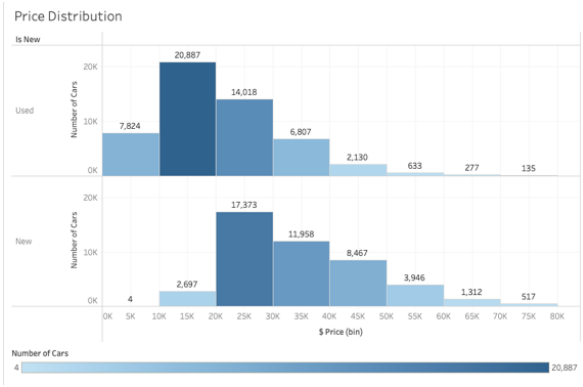

***Figure 2**: Price Distribution histogram shows the price frequencies of used vs new cars (top vs. bottom).*

4. VISUALIZATIONS

4.1. Mileage

Based on the findings from the correlation matrix (Appendix A2), mileage was the highest negatively correlated feature with price. After creating a scatterplot to visualize mileage vs. price (**Figure 3**), it was evident that there was a negative and possibly linear relationship between the two variables. The visualization confirms our mental model of the relationship between mileage and price, as this relationship is very intuitive. The commonly understood concept that a car loses value as soon as you drive it off the lot can be understood to mean that increasing mileage results in decreasing price. Using color to represent year and size to represent number of owners aimed to depict the multi-dimensional relationships between features in a simplified manner. To better visualize the distribution of mileage by number of vehicles and price, a histogram bar chart was also created (**Figure 4**). In this chart, gradient color was used to represent average price to highlight the pattern of decreasing price with increasing mileage.

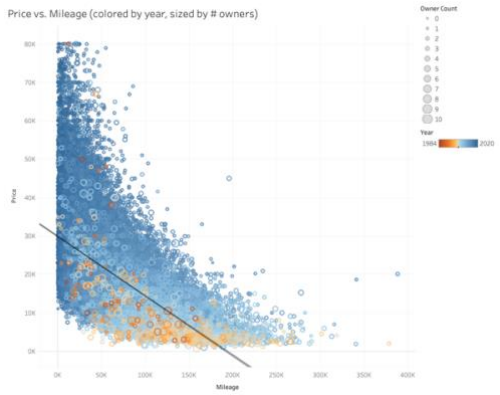

***Figure 3**: Price vs. Mileage scatterplot depicts year by color, # owners by circle size and a linear negative trendline.*

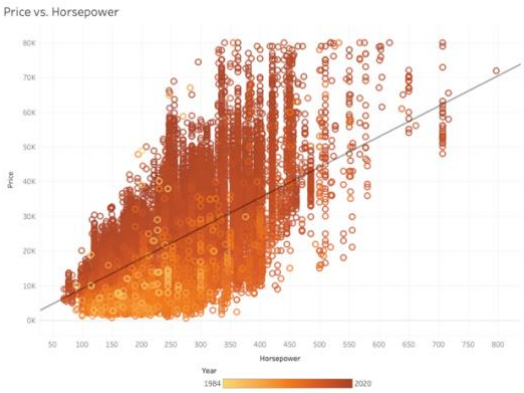

***Figure 5**: Price vs. Horsepower scatterplot depicts year by color and a linear positive trendline.*

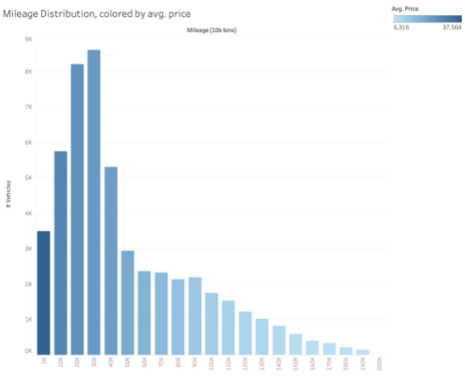

***Figure 4**: Mileage Distribution histogram shows the number of vehicles by 10k mileage bins, colored by average price.*

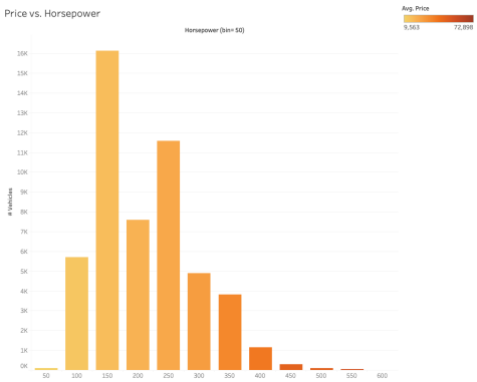

***Figure 6:** Horsepower Distribution histogram shows the number of vehicles by 50hp bins, colored by average price.*

4.2. Horsepower

Horsepower was another highly correlated feature to price, however unlike mileage it is positively correlated. Similar charts were created to visualize the relationship between horsepower and price (**Figure 5**) and the distribution of horsepower by number of vehicles and average price (**Figure 6**). Maintaining consistent graphs but altering the color palette was intended to make it easier to perceive differences in trends between the two sets of variables and create pre-attentiveness. Based on lessons learned from the first set of mileage graphs, circle size was kept constant and only year was depicted by color on **Figure 5**.

4.3. Location

In order to view aggregate state and detailed city visualizations, a second data set of zip code-state lookups was used. Examining the distribution of cars for sale based on location showed that the Northeast and Mid-Atlantic had the largest number of listings (**Figure 7**). Predictably, larger metro areas had the largest number of listings (**Figure 8**), with Pittsburgh surprisingly having the highest average price (**Figure 9**). Location, population, cost of living, and availability of vehicles for sale locally can heavily influence price, thus the need to understand both the number of available used cars for sale and the average price. Using color to represent density of vehicle listings created pre-attentiveness and reinforced viewer mental models. Choosing to use a dual axis in **Figure 9** did leave room for ambiguity and decoding errors, but labeling and consistent color gradients were used to guide viewers and provide contextual clues.

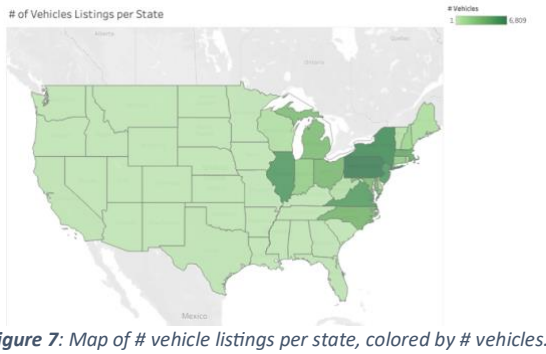

***Figure 7**: Map of # vehicle listings per state, colored by # vehicles.*

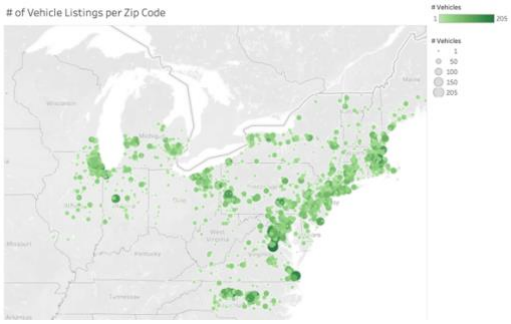

***Figure 8**: Map of # vehicle listings per Zip Code, colored by # vehicles.*

***Figure 9**: Top 10 cities by # vehicle listings (bar) and average price (line), colored by # vehicles.*

4.4. Multi-dimensional: Make, Model, Body Type, Year

The dataset contained over 50 distinct car makes (manufacturer, i.e. BMW), which were filtered down to depict only the top 25 makes (by number of vehicle listings) for the following visualizations. This best represents a high-level abstraction of the car market, as these are the most popular or most common makes for sale in the US at the time.

4.4.1. Make

A heat map of average price by $10k bins was created to depict an abstracted high-level view (**Figure 10**). To further enhance the visualization, it was filtered in descending order to depict the most frequent make to enhance the viewers mental model of the range of prices for these popular car makes. After creating **Figure 10**, it was clear that it attempted to convey too much detail, and the subset would need to be reduced going forward, however with over 50 makes, it was important to show the variance and dispersion of the dataset by make. To create an alternative lower complexity visualization with less detail, a TreeMap inspired by the finviz TreeMap [17] was created of the same top 25 makes to depict average price by make, sorted by highest average price (**Figure 11**). The TreeMap greatly reduced the amount of noise and ambiguity and facilitated faster comparison by utilizing color and size as salient features.

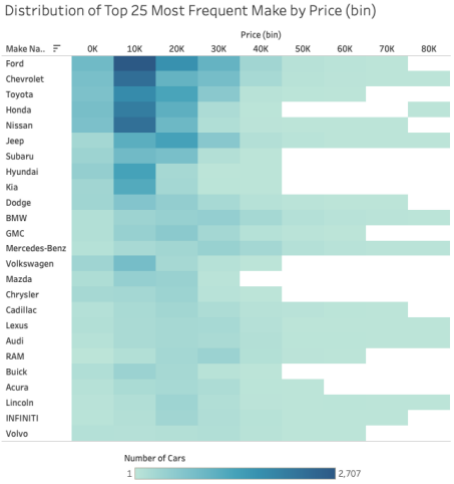

***Figure 10**: Heat Map of top 25 most frequent makes, in descending rank, vs. price bin, colored by # vehicles.*

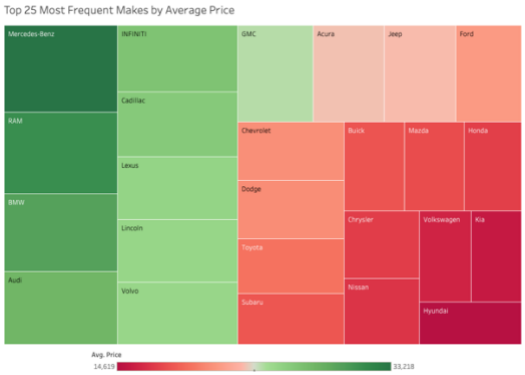

***Figure 11**: TreeMap of top 25 most frequent makes colored and sized by average price.*

4.4.2. Body Type & Model

A Box and Whisker plot of body type vs. price ranges, colored with contrasting colors to depict vehicle year— new (blue) and old (red)— was used to demonstrate the range of prices within a body type (**Figure 12**). Using the y-axis to represent price in this case was much more salient as each category was then easily read and compared from left to right. It is evident from the plot that some body types have a much larger range of acceptable prices, while others like hatchbacks, minivans, and vans, occupy a narrower range.

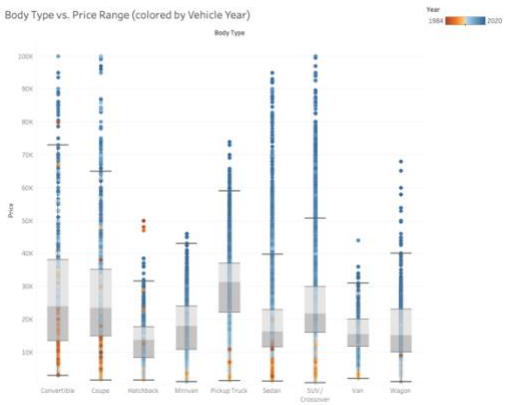

Figure 12: *Box and Whisker plot depicting price ranges by body type, colored by vehicle year to emphasize old (red) and new (blue).*

Further filtering was necessary to visualize car makes and models, as the range of categorical values in these features was too large to depict on a single visualization. Using the top 10 car makes (by number of vehicles) provided the most salient visualization, as it was then possible to further group by body type, making use of the gestalt principle of common region. In **Figure 13** the make name is color-coded, while the model is labeled on the x-axis. Average price and make name are then labeled on the individual bar elements for faster comprehension. Distinct and vibrant coloring of the 7 make categories facilitates comparison across dimensions. It is easy to then spot the frequency of Ford across 2 body styles and relate that back to knowledge gleaned from **Figure 10** which showed Ford as the most frequent make.

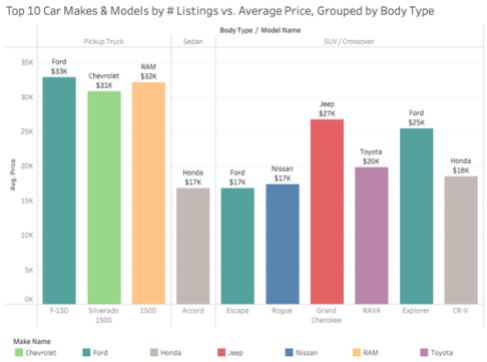

Figure 13: *Top 10 car makes & models (grouped by body type) vs. average price, colored by make name.*

4.4.3. Year

Examining year vs. average mileage and price provided another intuitive comparison, as cars get older, they accrue more milage and price decreases (**Figure 14**). The price reaches a global minimum as the mileage reaches a global maximum around year 2000. Visualizing the number of vehicles by model year and average price in **Figure 15** provides another perspective as the line for average price increases as model year increases and highlights that 2017 models are the most frequent.

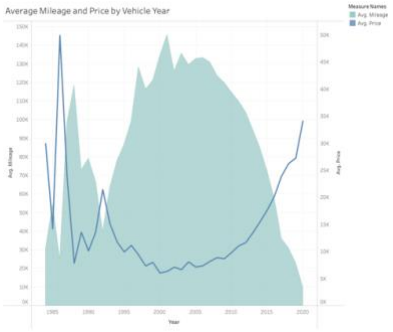

Figure 14: *Average mileage (area) and price (line) vs. vehicle year.*

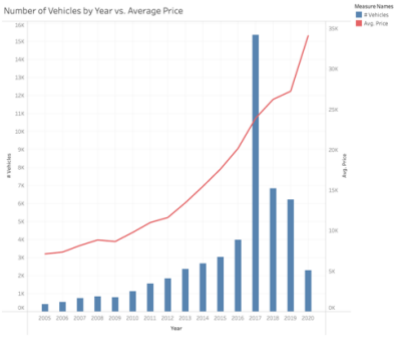

Figure 15: *Number of vehicles by model year (bar) vs. average price (line).*

4.5. Cluster Analysis (t-SNE)

Using Orange, t-SNE and K-Means unsupervised learning algorithms were created in the workflow and scatter plots generated to depict the relationships between mileage, price, make, and owner count on a reduced random sample of 10k cars with price <=$80k. Data was purposely not filtered on *is new* in order to visualize the distinct clusters in the dataset and uncover patterns that may distinguish new/used cars. t-SNE was configured using perplexity = 20, exaggeration = 1, and PCA components =8. K-means was configured using cluster size = 3 and K-Means++ initialization.

4.5.1. Mileage vs. Price

Owner Group & Body Type

Using t-SNE, *mileage* was plotted against *price* with color used to highlight number of owners. The calculated column *owner group* represents the engineered feature derived from *owner count* used to condense values >= 2 to the category 2. Data points were then individually labeled using body type to investigate patterns. As seen in **Figure 16,** 3 fairly distinct clusters based on number of owners can be seen, with 0 owner vehicles typically representing new cars with 0 mileage.

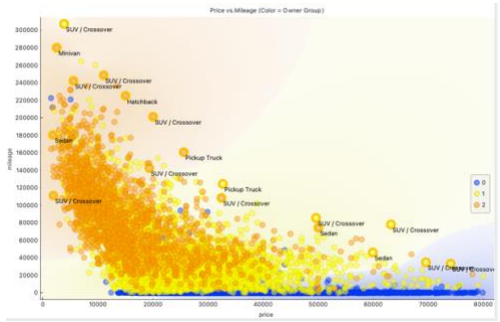

Figure 16: *Price vs. Mileage scatter plot, colored by # owners, labeled by body type.*

4.5.2. t-SNE-x vs. t-SNE-y

Owner Group & Body Type

t-SNE-x was then plotted against t-SNE-y in **Figure 17** using the same *owner group* field to color the data points. A clear delineation can be seen between 0 owner and 1, less so between 1 and 2—comparable to **Figure 16** cluster separations. However now in this figure, vehicles with 1 owner are more distinct than **Figure 16**.

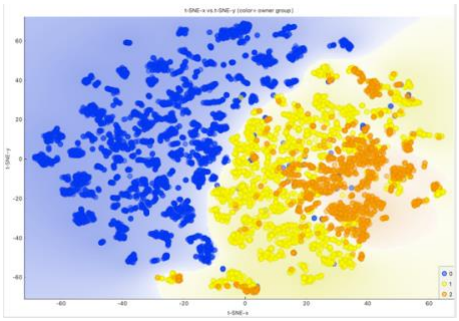

Figure 17: *t-SNE-x vs. t-SNE-y scatter plot, colored by # owners.*

The same t-SNE-x vs. t-SNE-y plot was then colored by *body type* to determine if distinct clusters could be found in this feature (**Figure 18**). Pickup trucks (yellow) formed the most distinctive and cohesive cluster, while SUV (pink) and Sedan (light blue) showed overlap and are not wholly distinct.

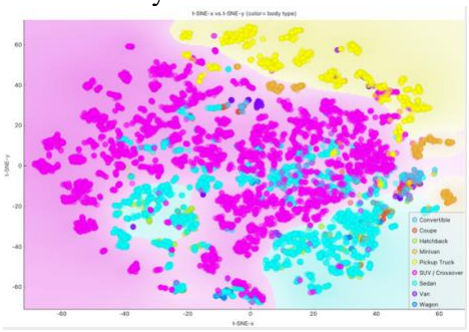

Figure 18: *t-SNE-x vs. t-SNE-y scatter plot, colored by body type.*

New vs. Used

Unlike in the Tableau visualizations, the Orange visualizations were not pre-filtered based on the feature column *is new* to show only used cars. The same t-SNE plot was used, but the color was set to use the *is new* column. Grouping the data points by color using this column showed the clearest cluster delineation and widest margin of separation of all the t-SNE plots, reinforcing the decision to filter on this feature and mental model of new vs. used cars (**Figure 19**). There were some noticeable outliers, but they were largely determined to be from bad data and data uncertainty.

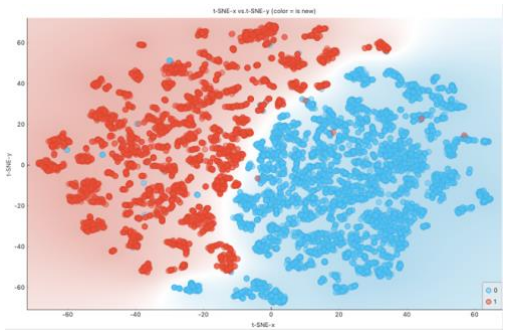

Figure 19: *t-SNE-x vs. t-SNE-y scatter plot, colored by new/used.*

4.6. Cluster Analysis (K-Means)

Mileage, Price, Make

A scatter plot of price vs. mileage was visualized after applying K-means using 3 clusters (**Figure 20)**. Compared to **Figure 16** using t-SNE and colored by *owner count*, the K-Means clusters appear to be at least partially derived using the owner count/group features, with cluster 3 most closely relating to new cars with 0 owners and mileage. Since owner group and body type were already examined, make was chosen as the label for this plot, but was not very informative aside from finding a cluster of Mercedes-Benz predictably between $60k and $80k.

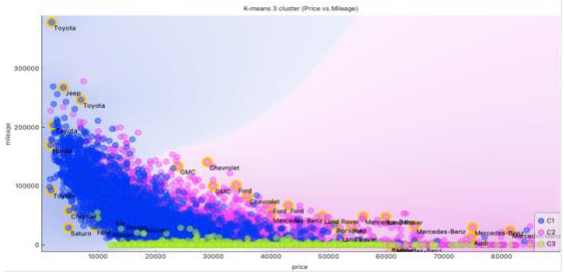

Figure 20: *K-means scatter plot of price vs. mileage, colored by cluster, labeled by make.*

Horsepower, Price, Make

A scatter plot of price vs. horsepower is depicted as **Figure 21**, colored by cluster, and labeled by make. Using the clusters to find patterns between horsepower, make, and price highlighted 3 distinct groups: (C1-blue) – above average price and below average horsepower (avg 238 hp); (C2-red) – wide range of prices and above average horsepower; (C3-green) – below average price and at or below average horsepower. Overall, the clusters were not all as similar within, showing greater dispersion and less central tendency in the red cluster 2. Make yielded no discernable patterns or noteworthy observations.

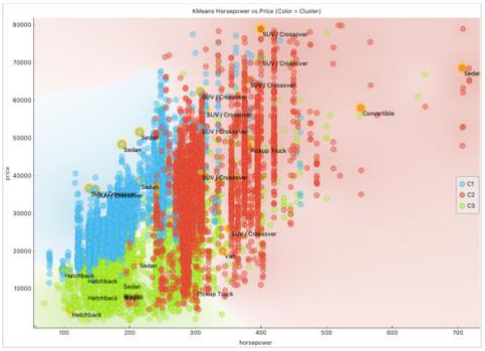

Figure 21: *K-means scatter plot of price vs. horsepower, colored by cluster, labeled by make.*

4.7. Text Analysis

The free text field *description* was evaluated using two approaches to determine word frequency and investigate if this field could highlight any additional impactful feature descriptions not encoded independently as part of the original dataset.

4.7.1. Python + *wordart.com* word cloud

A Jupyter notebook using python and libraries nltk and pandas was used to pre-process the description text, remove stop words, numbers, and non-text characters, and export to csv. Using a base dataset of 300k rows, over 47.5 million tokens were identified, however only 10k tokens were exported due to limitations in the visualization tools used. This csv was then used as the input for the *wordart.com* free word cloud generation tool [14]. The tool provided customization in terms of text-processing, shape, font, layout, and style. This tool was chosen primarily because it offered a pre-made car shape, which reinforced existing mental models and emphasized the topic of the dataset as seen below.

Figure 22: *Wordart.com generated word cloud in the shape of a car; font size represents frequency.*

4.7.2. Orange data mining word cloud

Orange software was used to pre-process the same dataset, randomly down sampled to 10,000 rows, equating to over 1.7 million tokens. It was filtered on the top 100 most frequent tokens and removed stop words, numbers, and non-text characters. The aim was to compare the output from Orange, given the larger number of tokenized words, to the output from the simple *wordart.com* tool and see if the overall landscape of frequency of words changed. As seen below, the word cloud generated from Orange contains many of the same words seen in the first word cloud. Orange lacked many of the customization features that *wordart.com* had, but the general dispersion and frequency of words was similar. Overall, both word clouds were largely comprised of common car component words, with "front" and "power" appearing most frequently in both visualizations.

Figure 23: *Orange generated word cloud, font size represents frequency.*

Another interesting feature that Orange provided was the annotated Corpus Map. Visualizing the salient features of price vs. mileage, it was possible to see the most common tokens in a cluster of data points. The corpus map shows a distinct cluster of 5 words in the high mileage low price range, with one potentially interesting word found- "intermittent".

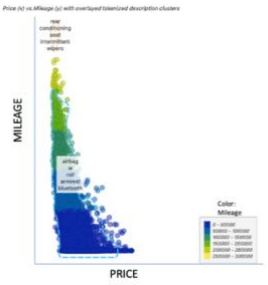

Figure 24: *Orange generated scatterplot of price vs. mileage, with corpus map highlighting noteable tokens in highlighted clusters.*

5. DISCUSSION

5.1. Results – Mileage

Figure 3 showed the intuitive negative relationship between mileage and price, while **Figure 4** depicted the central tendency of used cars to have less than 50k miles. Examining the additional features on **Figure 3** also showed a possible linear relationship between mileage and price, with price approaching zero as mileage approached 200k. The use of color to represent year emphasized the trend of older cars having lower prices, but the use of circle size to represent owner count was not as effective because it was not as easy to observe 'just noticeable differences' in such a densely crowded scatter plot.

5.2. Results- Horsepower

As expected from the correlation matrix, horsepower very clearly had a positive relationship with price. Cars with higher horsepower command a premium, though there is some variance in price seen in **Figure 5** that could likely be explained by other features of the car. Per **Figure 6**, horsepower follows a normal distribution, with most cars falling between 100-350hp, and the average being 238hp.

5.3. Results- Location

Location was not a straightforward dimension to compare to price, as it was highly dimensional and external factors like cost of living may influence the trends seen in price variation by location. However, the maps in **Figures 7&8** did identify high-density regions for used car listings in the Northeast and Mid-Atlantic. For a buyer looking to find a used car, these would be the regions to focus on searching as statistically there are more vehicles available for sale. Interestingly, across the top 10 cities as shown in **Figure 9**, there exists some average price variation that could be due to the mix of vehicles for sale, mileage, road conditions, cost of living, or other external factors not able to be visualized in a 2D graph.

5.4. Results – Multi-dimensional: Make, Model, Body Type, Year

Visualizing make and model did not yield definitive trends, aside from highlighting the dispersion of price within each make as shown in the heatmap **Figure 10**, and reinforcing existing mental models of luxury car brands commanding premium prices as shown in the TreeMap **Figure 11.** Examining by average price was only one aspect of understanding, as there was much price variance within a make; manufacturers produce multiple models at different price points to attract buyers with varying levels of purchasing power. Abstracting to a higher level and visualizing price by body type provided more memorable and actionable insight, as there was more central tendency and similarity within body types along the price dimension shown in the box and whisker plot (**Figure 12)**. It was interesting to see the narrower price ranges for vans, minivans, and hatchbacks compared to other body styles. **Figure 13** was perhaps one of the more busy and complex figures as it aimed to show multiple dimensions on a bar graph, but effectively used grouping, contrasting colors, and layering to focus viewers' attention on salient features. It enabled viewers to ascertain what models within body types are the most common and their average prices.

Lastly, examining year reinforced pre-existing mental models, as it showed the positive relationship between year and price. **Figure 15** showed an interesting trend of 2017 year cars being the most frequent year available for sale, while **Figure 14** pointed out the trend of lowest average price and peak average mileage occurring for cars approximately 20 years old.

5.5. Results – Cluster Analysis

Cluster analysis yielded perhaps the most insightful findings and informed many of the Tableau multi-dimensional visualizations. In particular, the observed relationship between number of owners and price seemed to follow very closely to the 3 clusters determined by K-Means in **Figures 16 & 20**, implying that owner count played a large role in clustering similar data points. Intuitively this makes sense, as vehicles with 0 owners (i.e. new vehicles) were shown to be the most dissimilar from all others, occupying a distinct and clearly separated region in space in **Figure 19**.

Visualizing multiple dimensions using unsupervised learning techniques was successful as it utilized gestalt principles of similarity, proximity, and continuity to enhance understanding and reveal underlying patterns and structures that were not immediately apparent from the raw data. Grouping similar data points together in clusters emphasized by color enabled greater visual understanding by highlighting the whole picture instead of individual parts.

5.6. Results – Text Analysis

The *wordart.com* word cloud (**Figure 22)** was more effective in creating memorability and pre-attentiveness due to the use of the car shape. The whole was more than the sum of the parts in this instance, as the individual words making up the word cloud did not convey much information on their own nor provide any insightful discoveries, but when grouped together to form a shape reinforced the mental model of a car's features. Words such as "front" and "wheel" appeared most frequently, perhaps because FWD vehicles were the most common types of vehicles listed, and not necessarily because they were the most important features. Additional text processing such as TF-IDF may yield greater insight on features than a simple word cloud can.

6. CONCLUSION

Examining 10 features out of over 60 available in this dataset provided a great deal of insight into key relationships, patterns, and trends in prices of used cars. Though no one feature alone is solely responsible for determining the price of a used car, through the visualizations presented it is possible to understand how certain features positively or negatively influence price and what trends are currently observed in the used car market. Key takeaways include the observed negative relationships between price, mileage, and number of owners. Positive relationships were identified between price, year, and horsepower. Clusters by number of owners, horsepower, and body type indicate that these

features may have similarity within the group and predictable relationships with price.

Overall, in examining the many features of used car listings, the whole is other than the sum, as many features must be simultaneously examined and incorporated in order to accurately predict the price, but not all are equal. Pricing algorithms for used cars must factor in both internal and external features and consider other dealership listings, local market features, cost of living, and supply and demand. Based on the visualizations presented, a buyer or seller can confidently begin to understand what features impact price and what trends appeared in the US used car market in 2020.

6.1. Future Work

Expanding upon the existing text analysis to uncover less frequent but recurring terms using TF-IDF could yield more insight into how terminology differs across price and other features. Removing a custom list of words common to all cars such as: wheel, tire, brake, seat belts, and mirrors, before creating a word cloud may offer more actionable insights but would require more domain knowledge.

Exploring the time-series analysis of this dataset was intentionally left out, as the listing date ranges were quite narrow, however it could yield some insight into how prices have changed over time, especially given the pandemic.

A focused analysis by make or body type could yield more actionable insights for targeted potential buyers, as the diversity within the make category was too high to fully visualize or glean meaningful insight from. Many buyers have a preconceived notion of what type or make of car they are looking for, so a detail and zoom feature that harnesses this existing bias could be more informative than the generalized overview that was presented herein.

6.2. Limitations

Using Tableau to analyze the location of the listings had many shortcomings, from missing city names to lack of built-in abilities to extrapolate a state from a zip code. Having to use Tableau Public to create the Sankey Diagram was not nearly as seamless as Tableau Desktop at handling large datasets, and didn't allow for saving workbooks as Tableau Desktop did.

In comparison to Tableau, Orange provided a much more robust framework for building an exploratory workflow and visualizations, but it was far more limited in visualization customizations and polished formatting. One unfortunate downside to using Orange is the lack of ability to customize some display options beyond adding a title, such as the x and y axis labels and legend that were manually drawn in **Figure 24**. The ability to create more advanced data science pipelines in Orange is a useful feature, but in order to create more robust visualizations it may be best to export the resulting datasets to Tableau which has more polished visualizations and customization options.

7. APPENDIX

Data Dictionary

Column Name	Datatype	Description
vin	String	Vehicle Identification Number is a unique encoded string for every vehicle
back_legroom	String	Legroom in the rear seat.
bed	String	Category of bed size(open cargo area) in pickup truck. Null usually means the vehicle isn't a pickup truck
bed_height	String	Height of bed in inches
bed_length	String	Length of bed in inches
body_type	String	Body Type of the vehicle. Like Convertible, Hatchback, Sedan, etc.
cabin	String	Category of cabin size(open cargo area) in pickup truck. Eg: Crew Cab, Extended Cab, etc.
city	String	City where the car is listed. Eg: Houston, San Antonio, etc.
city_fuel_economy	Float	Fuel economy in city traffic in km per litre
combine_fuel_economy	Float	Combined fuel economy is a weighted average of City and Highway fuel economy in km per litre
daysonmarket	Integer	Days since the vehicle was first listed on the website.
dealer_zip	String	Zipcode of the dealer
description	String	Vehicle description on the vehicle's listing page
engine_cylinders	String	The engine configuration. Eg: I4, V6, etc.
engine_displacement	Float	Measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers.
engine_type	String	The engine configuration. Eg: I4, V6, etc.
exterior_color	String	Exterior color of the vehicle, usually a fancy one same as the brochure.
fleet	Boolean	Whether the vehicle was previously part of a fleet.
frame_damaged	Boolean	Whether the vehicle has a damaged frame.
franchise_dealer	Boolean	Whether the dealer is a franchise dealer.
franchise_make	String	The company that owns the franchise.
front_legroom	String	The legroom in inches for the passenger seat
fueltankvolume	String	Fuel tank's filling capacity in gallons
fuel_type	String	Dominant type of fuel ingested by the vehicle.
has_accidents	Boolean	Whether the vin has any accidents registered.
height	String	Height of the vehicle in inches
highway_fuel_economy	Float	Fuel economy in highway traffic in km per litre
horsepower	Float	Horsepower is the power produced by an engine.
interior_color	String	Interior color of the vehicle, usually a fancy one same as the brochure.
isCab	Boolean	Whether the vehicle was previously taxi/cab.
is_certified	Boolean	Whether the vehicle is certified. Certified cars are covered through warranty period
is_cpo	Boolean	Pre-owned cars certified by the dealer. Certified vehicles come with a manufacturer warranty for free repairs for a certain time period.
is_new	Boolean	If True means the vehicle was launched less than 2 years ago.
is_oemcpo	Boolean	Pre-owned cars certified by the manufacturer
latitude	Float	Latitude from the geolocation of the dealership.
length	String	Length of the vehicle in inches
listed_date	String	The date the vehicle was listed on the website. Does not make daysonmarket obsolete. The prices is dayson_market days after the listed date.
listing_color	String	Dominant color group from the exterior color.
listing_id	Integer	Listing id from the website, unique
longitude	Float	Longitude from the geolocation of the dealership.
main_picture_url	String	URL for listing
major_options	String	Name of options package, i.e. Adaptive Cruise Control
make_name	String	Vehicle make name, i.e. Jeep
maximum_seating	String	# seats
mileage	Float	# miles on the odometer
model_name	String	Vehicle model name, i.e. Wrangler
owner_count	Float	# previous owners
power	String	Full horsepower description
price	Float	Price for the vehicle
salvage	Boolean	If the vehicle has a salvage title then True
savings_amount	Integer	Amount saved
seller_rating	Float	Rating on a scale of 1-5
sp_id	Float	dealership ID
sp_name	String	dealership name
theft_title	Boolean	If the vehicle has a theft title then True
torque	String	Torque rating of vehicle
transmission	String	Type of Transmission- A (automatic)
transmission_display	String	Type of Transmission detailed display- 9-speed Automatic Overdrive
trimId	String	ID code of trim
trim_name	String	Description of trim- SE FWD
vehicle_damage_category	Float	Null, no data
wheel_system	String	Type of wheel drive system, i.e. FWD
wheel_system_display	String	Detailed description of wheel drive, i.e Front-Wheel Drive
wheelbase	String	Size of wheelbase inches
width	String	Width of vehicle inches
year	Integer	Year of vehicle manufacture for sale
age*	Integer	Derived Col: Bucketed age based on year column (Ordinal from oldest to newest) 1 = Antique (<1997), 2= 1997-2010, 3= 2010-2015, 4=>2015
avg_fuel_economy*	Float	Derived Col: Average of highway + city fuel economy

Table A1: Data Dictionary

Additional Figures

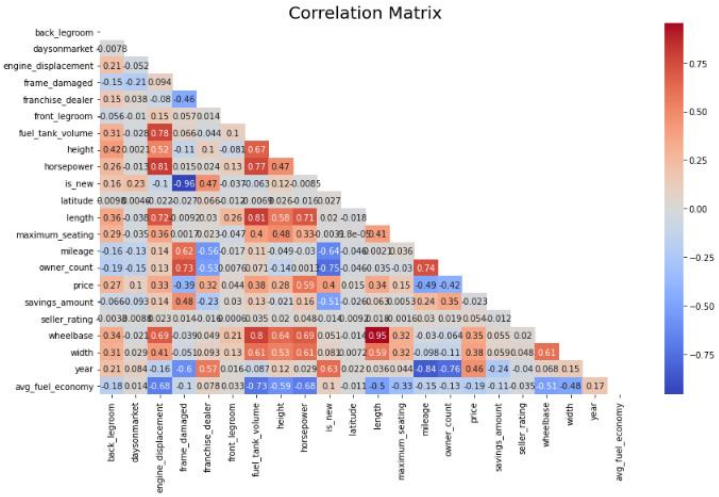

Figure A2: Correlation Heat Matrix. *The correlation heat matrix was built using Seaborn in Python and shows the most negatively correlated features to be mileage and owner count, while the most positively correlated features are horsepower and is_new.*

8. REFERENCES

[1] Rosenbaum, E. (2020, October 15). *The used car boom is one of the hottest, and trickiest, coronavirus markets for consumers*. CNBC. Retrieved May 28, 2023, from https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html

[2] Carlier, M. (2022, September 2). *U.S. used car dealer market size*. Statista. Retrieved May 28, 2023, from https://www.statista.com/statistics/1328700/us-used-car-dealer-market-size/

[3] *CarGurus - Compare KBB vs CarGurus*. (n.d.). Kelley Blue Book. Retrieved May 25, 2023, from https://www.kbb.com/cargurus/

[4] *US Used cars dataset*. (2020, September 16). Kaggle. Retrieved May 28, 2023, from https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

[5] *US Zip Codes Points - United States of America*. (2021, April 27). OpenDataSoft. Retrieved May 31, 2023, from https://data.opendatasoft.com/explore/dataset/georef-united-states-of-america-zc-point%40public/information/

[6] Amik, F.R., Akash L., Ahnaf I., and Sifat M.. 2021. "Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh" *Information* 12, no. 12: 514. Retrieved May 25, 2023, from https://doi.org/10.3390/info12120514

[7] *Automatic Determination of Used Car Prices*. (n.d.). Mercedes-Benz Group. Retrieved June 2, 2023, from https://group.mercedes-benz.com/careers/about-us/artificial-intelligence/for-nerds/pricing.html

[8] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019, February). Car Price Prediction using Machine Learning Techniques. *TEM Journal*, *8*(1), 113-118. 10.18421 Retrieved May 25, 2023 from https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf

[9] Martucci, B. (n.d.). *12 Factors That Affect Your Car's Resale Value*. Money Crashers. Retrieved May 23, 2023, from https://www.moneycrashers.com/factors-affect-used-cars-resale-value/

[10] Kumbar, K., Gadre, P., & Nayak, V. (2019, Decemner 21). *CS 229 Project Report: Predicting Used Car Prices*. CS229. Retrieved June 2, 2023, from https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf

[11] Orange Data Mining - Data Mining. Retrieved May 30, 2023, from https://orangedatamining.com/

[12] *The Tableau Platform | World's #1 analytics*. (n.d.). Tableau. Retrieved June 3, 2023, from https://www.tableau.com/products/our-platform

[13] Bartlett, S., & VanDerAa, K. (2023, April 21). Exploring Sankey and Radial Charts with the New Chart Types Pilot on Tableau Public. Tableau. Retrieved June 3, 2023, from https://www.tableau.com/blog/exploring-sankey-and-radial-charts-new-chart-types-pilot-tableau-public

[14] *WordArt.com*. (n.d.). Word Art. Retrieved June 1, 2023, from https://wordart.com

[15] *seaborn.heatmap — seaborn 0.12.2 documentation*. (n.d.). Seaborn. Retrieved June 1, 2023, from https://seaborn.pydata.org/generated/seaborn.heatmap.html

[16] *CarGurus Instant Market Value*. (2018). CarGurus Instant Market Value. Retrieved May 31, 2023, from https://assets.ctfassets.net/0czyc7nlfvzo/4f2pymo70GTJ6EqnoMB7GO/d50c19b3b16a83f71e4b7e35075f46c3/CarGurus-IMV-one-pager.pdf

[17] *S&P 500 Map*. (n.d.). FINVIZ.com. Retrieved June 1, 2023, from https://finviz.com/map.ashx