Executive Summary

This report analyzes a Kaggle-sourced dataset to predict car prices based on different attributes. It aims to provide insights for the automotive industry. Our regression model predicted car price from input values with a 98.2% accuracy (R-squared). This means our model was a good fit, and hence strong to base assumptions off of and make insights from.

Hypotheses centered on attribute significance and pattern identification within the dataset. The data wrangling process involved importing the dataset, addressing empty values by removing corresponding rows, and eliminating unnecessary columns. In the model creation phase we ran a linear regression, necessitating the generation of dummy variables for categorical attributes. The initial model achieved a 94% R-squared value, refined to 98.2% by eliminating insignificant variables. The resulting model also had a low RMSE of $2613, indicating minimal prediction deviation.

The model is most effective within the $5,000 - $25,000 price range. For instance, the model indicates an average price increase of $122.23 for every additional cubic centimeter in engine size. This is a very strong prediction for cars within the range. Before using the model to make predictions outside of that range, testing the model on unseen data and re-fitting would be good for further validation.

Practical applications include informing car dealerships about feature-specific pricing impacts and aiding customers in making informed decisions based on preferences and budget constraints. For example, our model highlights having less curb weight reduces the price of a car.

Introduction

Our project is to analyze a Kaggle-sourced dataset containing a diverse set of attributes of cars. Our goal is to extract meaningful insights, using the data to form a regression model which can predict car prices given certain characteristics. This should provide valuable perspectives for decision-makers in the automotive industry. This report will address the business problem, data processing, analytical methods, and conclude with actionable insights for organizations operating in this sector.

In the dynamic automotive landscape, understanding market dynamics and consumer preferences is crucial. The dataset's comprehensive view of car attributes- from technical specifications to performance metrics and pricing- is what motivates our project. Our goal is to make an accurate model for car price, enabling informed decisions about product development, marketing strategies, and pricing structures.

The selection of this specific dataset is based on the significance of the automotive industry and its continuous transformation. Analyzing car characteristics allows us to address questions such as consumer preferences, pricing strategy, and market trends. The importance of these insights lies in their potential to offer strategic advantages to automotive manufacturers and stakeholders. By understanding the factors influencing consumer choices and market trends, organizations can tailor their offerings, enhance competitiveness, and make data-driven decisions. Questions to consider are what features are most valued by consumers, and how do these preferences vary across different market segments? How do various car attributes influence the pricing of vehicles? Are there patterns that suggest optimal pricing structures?

The hypotheses we tested revolve around the significant impact of different car attributes on pricing and the variation of preferences based on different car specs. The coefficients of our regressors tell us the ceteris paribus effect of each item on car price. We also looked at the

existence of identifiable patterns in the dataset. In the subsequent sections of this report, we will explore the data wrangling process and talk about our data analytics techniques.

Data Wrangling Process

Knowing we wanted to create a project based on the automotive industry, specifically car prices, we had to find a dataset that detailed components of cars that could potentially influence its prices. Eventually, we were able to find a dataset on Kaggle that had all the data we needed. This dataset included a variety of variables, both numeric and categorical, that can be used to determine car prices (Exhibit 1). To get the data into our Google collab file, we simply had to import drive and direct it to where it is downloaded in our individual google drive's. This allowed access to the data directly in our ipynb file.

However, the dataset we imported simply includes raw data, meaning there are some empty values scattered throughout. Our first step to curating this dataset was to figure out what to do with these empty values. We had two options: fill in the values with the mean/median of the column, or remove them altogether. Because we had plenty of rows in the dataset already, we would still have enough data points after removing rows. Therefore, rather than filling in the values, we removed the rows with null values altogether. By dropping all rows with empty values, we now have a full table of data to continue on with the analysis. Additionally, working with raw data means some of our columns may not be useful for our regression. For example, the column car_ID is useless in our dataset, as it is simply used to number each car in the dataset. The CarName is also useless as each car in the dataset has a different name. Because of this, both of these columns can be dropped from the dataset.

At this point, we can start to create our model. Because our expected outcome is continuous and not categorical, we knew to use a linear regression model to predict the car

prices. We had to create dummy variables for all the categorical variables so they can be interpreted properly. From there, we created our model, which had an R-squared value of 94%. However, multiple x variables had p-values above 0.05, so we refined the model even further by removing the insignificant variables. In the end, we were able to create a final model that had an R-squared value of 98.2% (Exhibit 2). Our next section will dive deeper into the analysis of our final model, data visualizations, and actionable insights.

Data Analysis/Discussion

Overall, we were satisfied at the accuracy of our model from the car characteristics we were given. As mentioned previously, the R-squared value was 98.2%, meaning that 98.2% of the variation in car prices can be explained by all the variables we used. Having such a high R-squared value means that the model fits our data very well. It is important to note that overfitting could have potentially occurred here, so a next step to ensuring the accuracy of the model would be to test unseen data on the model to see its performance. Regardless, the model we made is very accurate in determining the prices of cars based on specific car characteristics. Diving deeper into our analysis, our RMSE (root mean squared error) value was only $2613. This means that on average, our prediction was only off the actual price by $2613.

Looking at Exhibit 3, we can see that our predictions are very close to the actual result most of the time. However, this graph indicates that most of our actual data takes place between the $5000 - $25000 price range, so it is best to use this model when the expected price is between those numbers. After this range, there are minimal data points which makes it difficult for our model to accurately predict these prices.

Knowledge of this information can be very important to car dealerships. For a car dealership, knowing how each individual car feature impacts price is very important to properly

pricing a car for the public. For example, engine size can play a factor in pricing a car. For every additional cubic centimeter increase in engine size, the price of the car increases by $122.23 on average. This can be important information for pricing a new car. For example, say Audi has a car that is priced at $25000 with an engine size of 100. They make a model exactly the same, but instead upgrade it so it has an engine size of 120. The value of the car obviously increased, but by how much? Our model says that the new car's price should be $27444. The individual impact of engine size on car price can be found in Exhibit 4. This same analysis can be used for any variables, including fuel type, peak RPM, car width, and many more. It can be a way car dealerships price their new cars.

In general, this is also good information for a customer. From a customer's perspective, this is valuable because they can see where they can save money. Knowing what features drive up the price of a car can help a customer determine what they are looking for. Customers typically have a price range for their car, so they can see which car characteristics and features fit into that price range. A car with less curb weight, for example, will help reduce prices of the car. As long as the customer doesn't have a preference over this, and it doesn't come at the expense of other features, a slightly smaller curb weight can be very appealing for a cost-saving customer. Exhibit 5 shows the specifics on weight vs price.

In conclusion, our project demonstrates how the power of data can be used to create highly-accurate regression models to predict outcomes. In our case, we were able to use a variety of car attributes to determine the price it should have. This model not only provides crucial insights to the automotive industry for pricing strategies, but it also helps consumers understand the value certain characteristics hold. In a world converting to automation and analytics, these

correlations serve as a way for dealerships and customers to make decisions on pricing strategies and purchasing cars.

Exhibit 1: Dataset Variables and Types

DATA DICTONARY		
1	Car_ID	Unique id of each observation (Interger)
2	Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
3	carCompany	Name of car company (Categorical)
4	fueltype	Car fuel type i.e gas or diesel (Categorical)
5	aspiration	Aspiration used in a car (Categorical)
6	doornumber	Number of doors in a car (Categorical)
7	carbody	body of car (Categorical)
8	drivewheel	type of drive wheel (Categorical)
9	enginelocation	Location of car engine (Categorical)
10	wheelbase	Weelbase of car (Numeric)
11	carlength	Length of car (Numeric)
12	carwidth	Width of car (Numeric)
13	carheight	height of car (Numeric)
14	curbweight	The weight of a car without occupants or baggage. (Numeric)
15	enginetype	Type of engine. (Categorical)
16	cylindernumber	cylinder placed in the car (Categorical)
17	enginesize	Size of car (Numeric)
18	fuelsystem	Fuel system of car (Categorical)
19	boreratio	Boreratio of car (Numeric)
20	stroke	Stroke or volume inside the engine (Numeric)
21	compressionratio	compression ratio of car (Numeric)
22	horsepower	Horsepower (Numeric)
23	peakrpm	car peak rpm (Numeric)
24	citympg	Mileage in city (Numeric)
25	highwaympg	Mileage on highway (Numeric)
26	price(Dependent variable)	Price of car (Numeric)

Exhibit 2: OLS Regression Results (Between Car Characteristics and Prices)

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.982
Model:	OLS	Adj. R-squared (uncentered):	0.980
Method:	Least Squares	F-statistic:	515.8
Date:	Tue, 12 Dec 2023	Prob (F-statistic):	1.05e-150
Time:	20:18:02	Log-Likelihood:	-1854.7
No. Observations:	205	AIC:	3749.
Df Residuals:	185	BIC:	3816.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
carwidth	401.2611	136.008	2.950	0.004	132.935	669.587
curbweight	3.0824	0.977	3.156	0.002	1.156	5.009
enginesize	122.2333	11.872	10.296	0.000	98.811	145.656
stroke	-5201.9510	641.131	-8.114	0.000	-6466.819	-3937.083
compressionratio	-936.6516	314.488	-2.978	0.003	-1557.095	-316.208
peakrpm	2.3313	0.435	5.355	0.000	1.472	3.190
fueltype_gas	-1.441e+04	4423.695	-3.258	0.001	-2.31e+04	-5685.443
aspiration_turbo	1590.6351	609.295	2.611	0.010	388.574	2792.696
carbody_hardtop	-3166.9477	1208.053	-2.622	0.009	-5550.278	-783.617
carbody_hatchback	-3441.6573	1030.205	-3.341	0.001	-5474.118	-1409.197
carbody_sedan	-2558.8153	1010.191	-2.533	0.012	-4551.791	-565.839
carbody_wagon	-3774.4285	1068.414	-3.533	0.001	-5882.270	-1666.587
enginelocation_rear	8361.5094	1756.333	4.761	0.000	4896.492	1.18e+04
enginetype_ohc	3611.0685	504.131	7.163	0.000	2616.484	4605.653
enginetype_ohcv	-5023.3163	901.024	-5.575	0.000	-6800.919	-3245.713
cylindernumber_five	-8403.8106	1188.837	-7.069	0.000	-1.07e+04	-6058.390
cylindernumber_four	-9727.1400	879.643	-11.058	0.000	-1.15e+04	-7991.719
cylindernumber_six	-5415.8525	913.021	-5.932	0.000	-7217.124	-3614.581
cylindernumber_twelve	-8885.3220	2977.671	-2.984	0.003	-1.48e+04	-3010.765
fuelsystem_spdi	-3011.1908	926.481	-3.250	0.001	-4839.017	-1183.365

Omnibus:	27.779	Durbin-Watson:	1.516
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56.346
Skew:	0.657	Prob(JB):	5.82e-13
Kurtosis:	5.207	Cond. No.	1.81e+05

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.81e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Exhibit 3: Predicted vs Actual Results

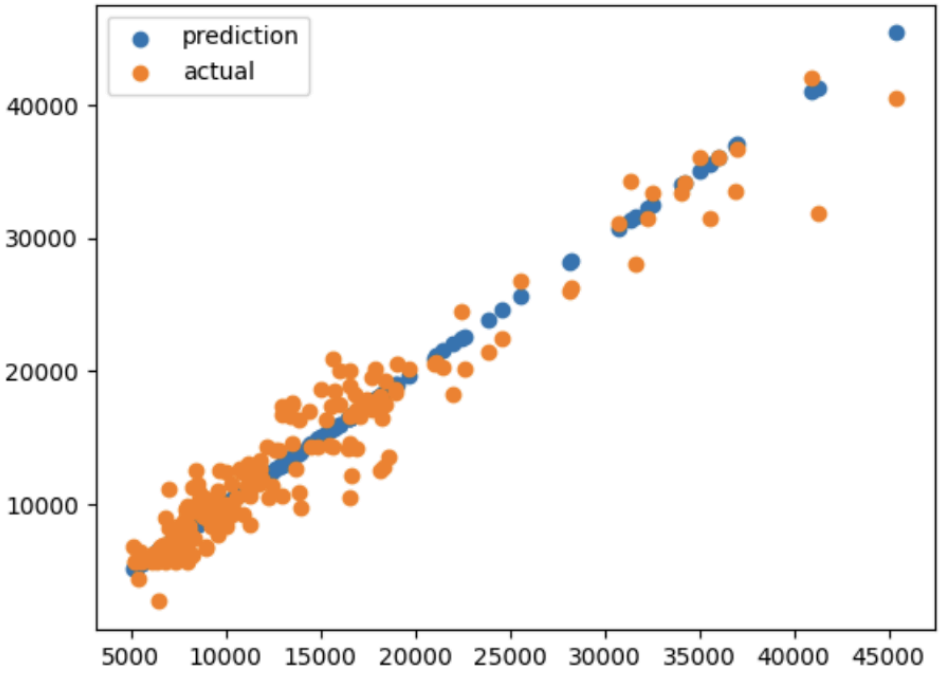

Exhibit 4: Engine Size vs Price

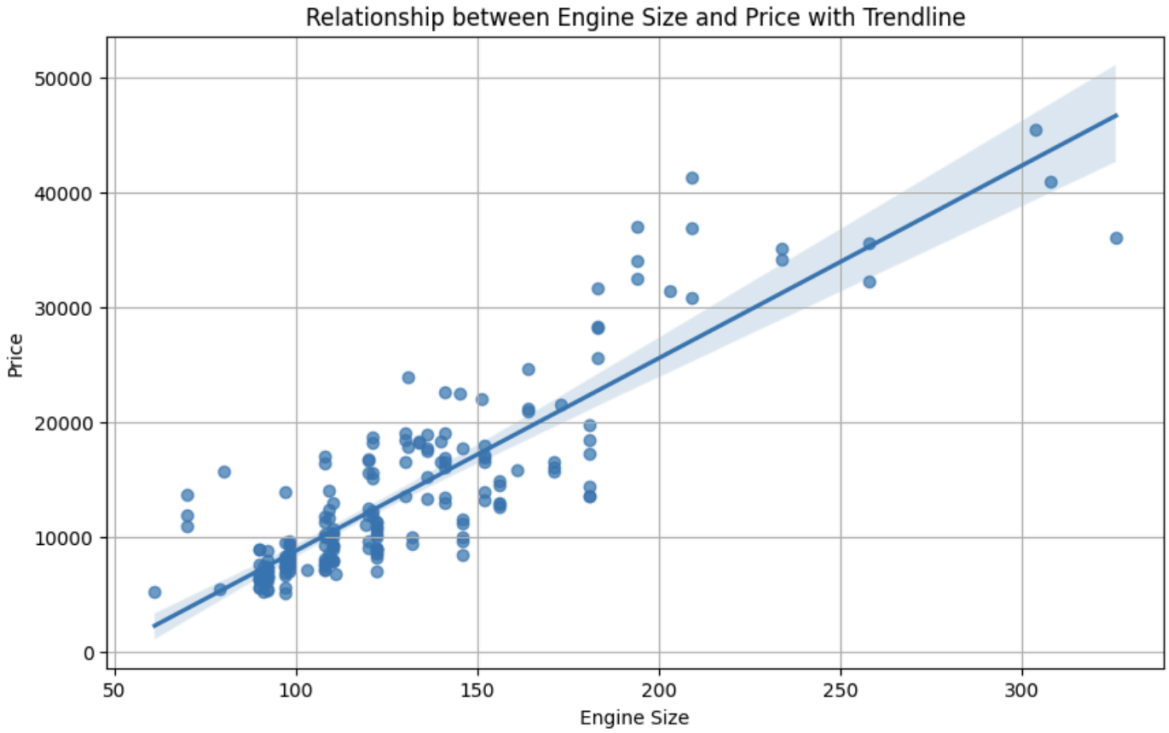

Exhibit 5: Curb Weight vs Price