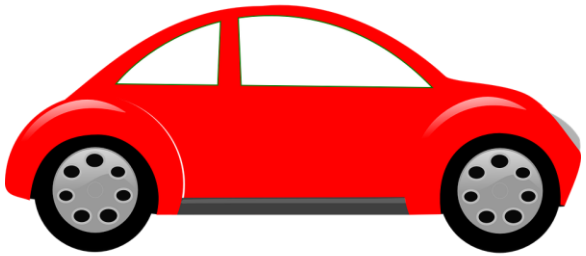
Car Price Statistical Analysis
Technical Presentation

Presenter: Zach Summy
Date: 15 November, 2020



Introduction

To assist them, we sought to answer three questions:

1. Which variables are significant in predicting the price of a car?

2. How do variations in those variables affect price?

3. Can a predictive model be built, and if so, what is it?

The variables in the dataset of cars include Categorical and Numerical variables

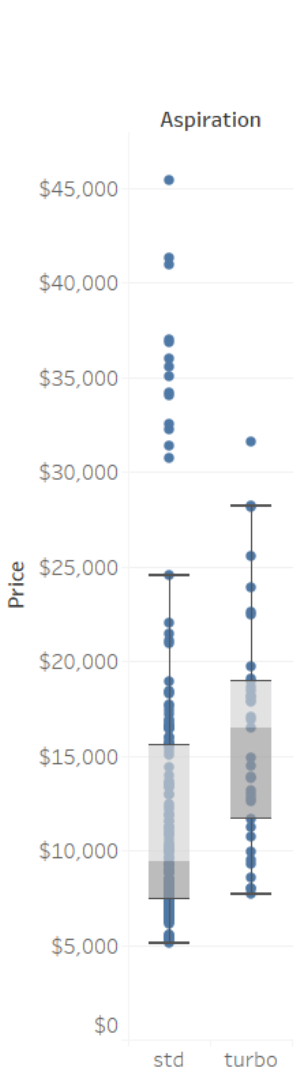
Variable	Define	Numerical or Categorical	Variable	Define	Numerical or Categorical
Aspiration	Air intake (standard or turbo)	Categorical	**Engine size**	Size of the engine, measured in cubic inches	Numerical
Bore ratio	Diameter of the cylinder divided by the piston stroke length	Numerical	**Engine type**	Mainly how the cylinder and engine components are arranged	Categorical
Car body	Car body style	Categorical	**Fuel system**	Components that deliver fuel from the tank to the engine	Categorical
Car length, width, height, curb weight	The dimensions of the car without passengers	Numerical	**Fuel type**	Diesel or gas	Categorical
City mpg	MPG based on frequent starting, stopping, and idling.	Numerical	**Highway mpg**	MPG based on more continuous acceleratio n.	Numerical
Compression ratio	The ratio of the volume of the cylinder and its head space w hen the piston is at the bottom of its stroke to the volume of the head space w hen the piston is at the top of its travel	Numerical	**Horsepower**	A unit of power equal to 550 foot-pounds per second	Numerical
Cylinder number	Number of engine cylinders	Categorical	**Peak rpm**	Typically the engine rpm at which the peak horsepower occurs	Numerical
Door number	Usually 2 or 4	Categorical	**Stroke**	The total distance traveled by the piston	Numerical
Drive wheel	The wheel of the car that transmits force	Categorical	**Wheelbase**	The horizontal distance between the centers of the front and rear wheels	Numerical
Engine location	Front or rear	Categorical	**Price**	The wholesale? price of the car	Numerical

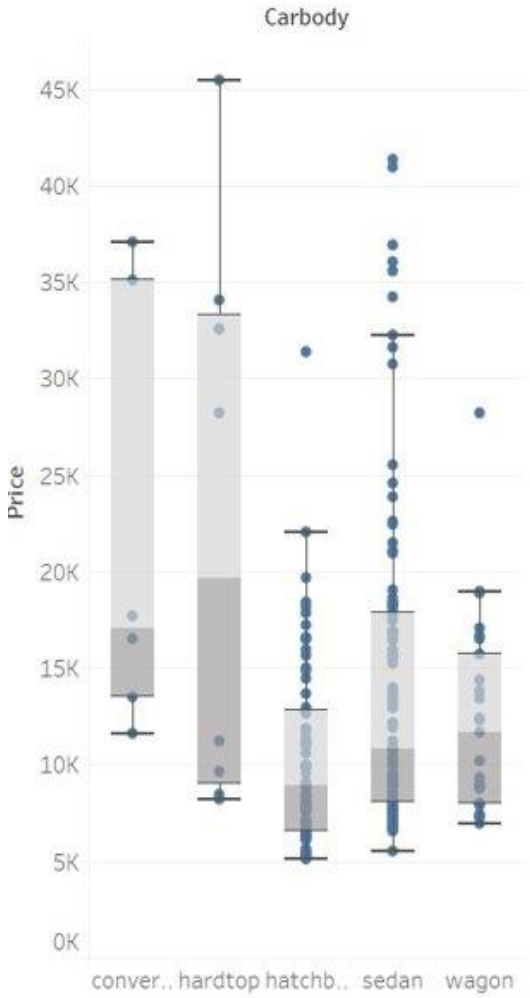
We employed both descriptive and inferential statistics to analyze these variables and relate them to price

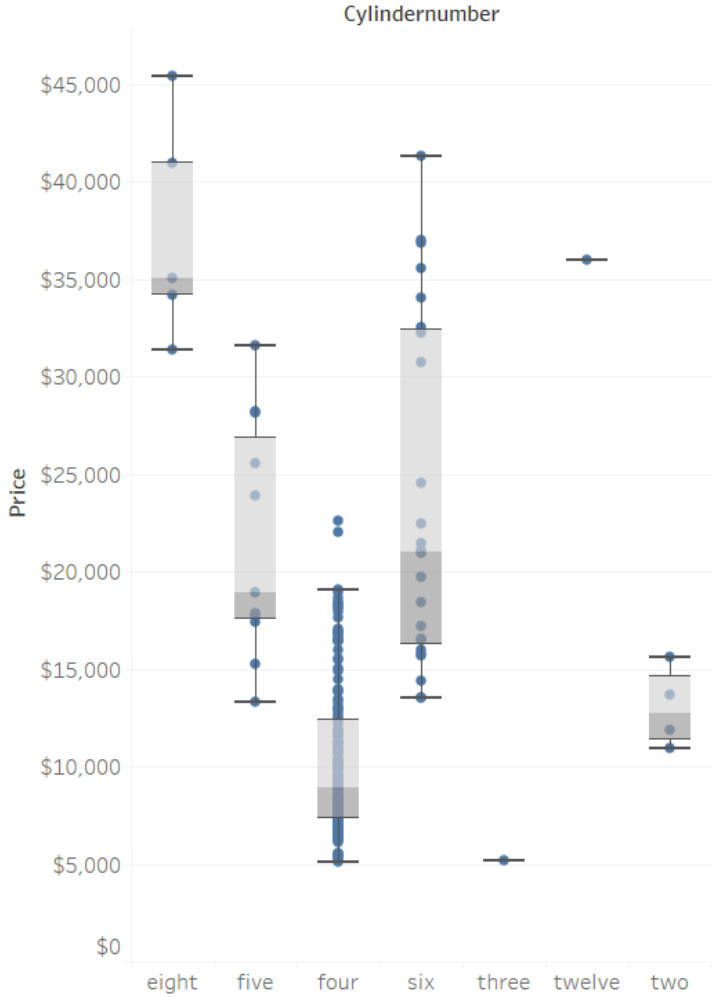
Methodology

- Use Descriptive Statistics (Box Plots) to see which Categorical Variables correlate highly with price
- Use Inferential Statistics (Correlations) for a precise measure of which Numerical variables correlate highly with price
- Examine for Cross-Correlations and pick a set a variables with high correlations and lower cross-correlations to perform a regression
- Compare the R2 of the regression with the R2 of the highest correlating variable to see how much "extra predictive power" we gained through the correlation

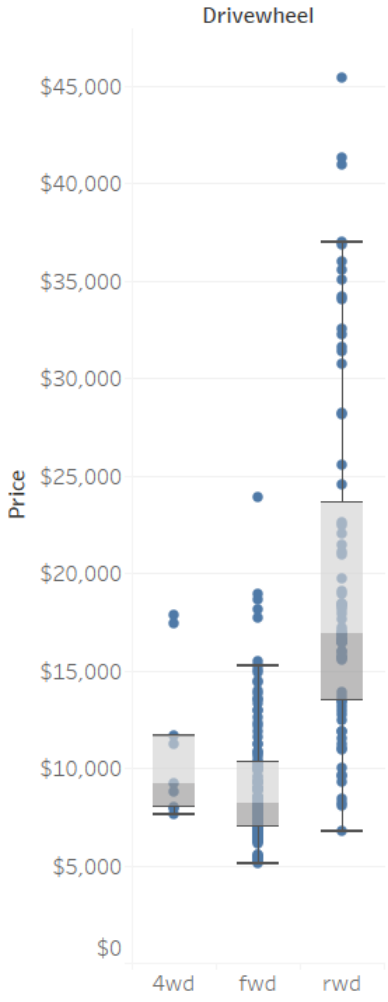
Prices increase greatly with cylinder number, from a median of approx $9,000 for four cylinder to $35,000 for eight cylinders; an ~50% increase exists across Aspiration type, and among the populated car body types very little increase

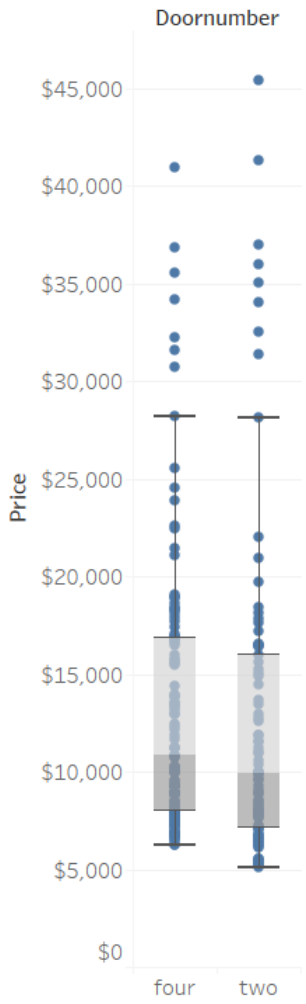




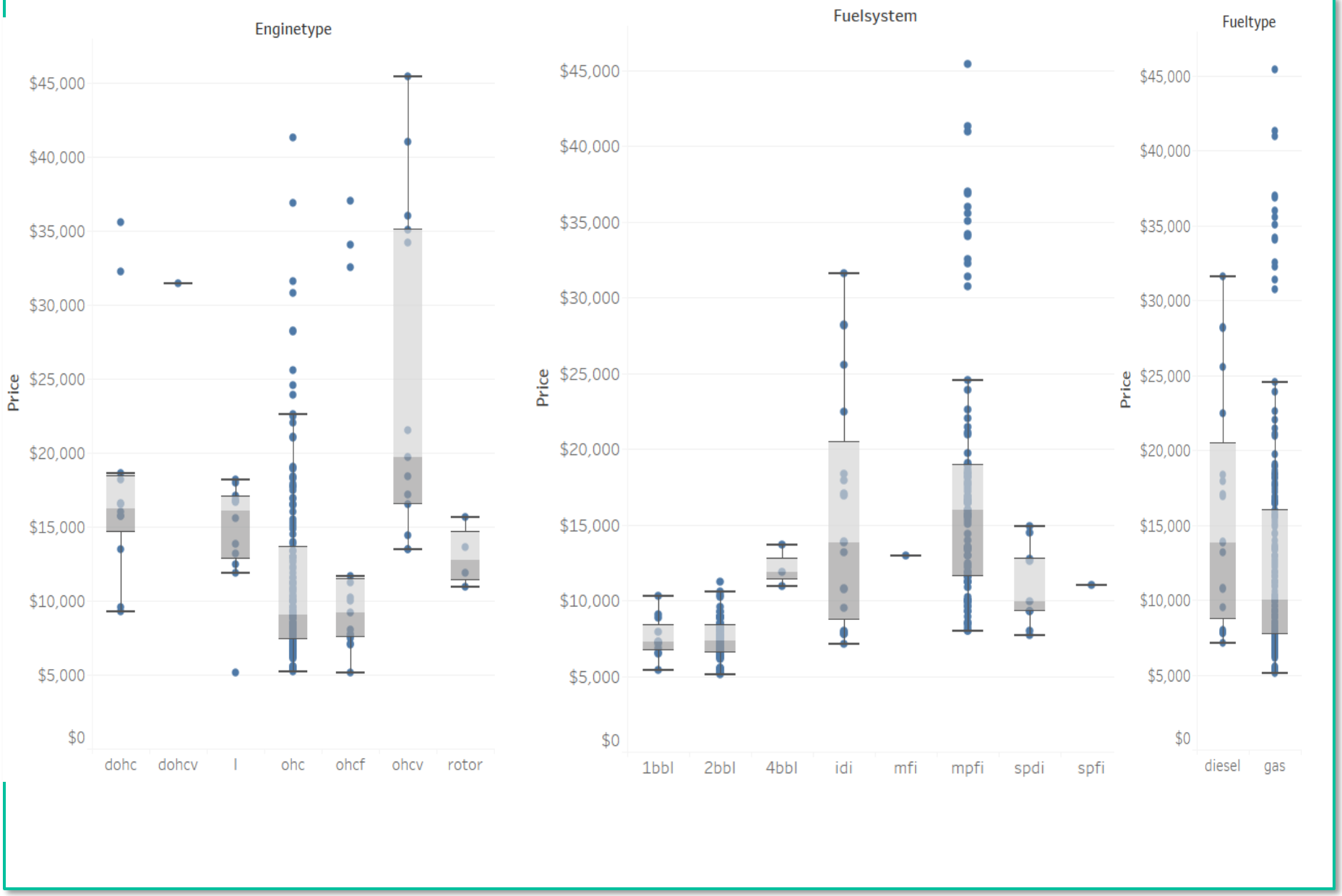


A shift to Rear-wheel-drive increases median prices from approx. $8,000 to $17,000; door number has a small effect





Enginetype and Fuelsystem show a doubling of the median price across the different types while Fueltype shows an ~50% increase

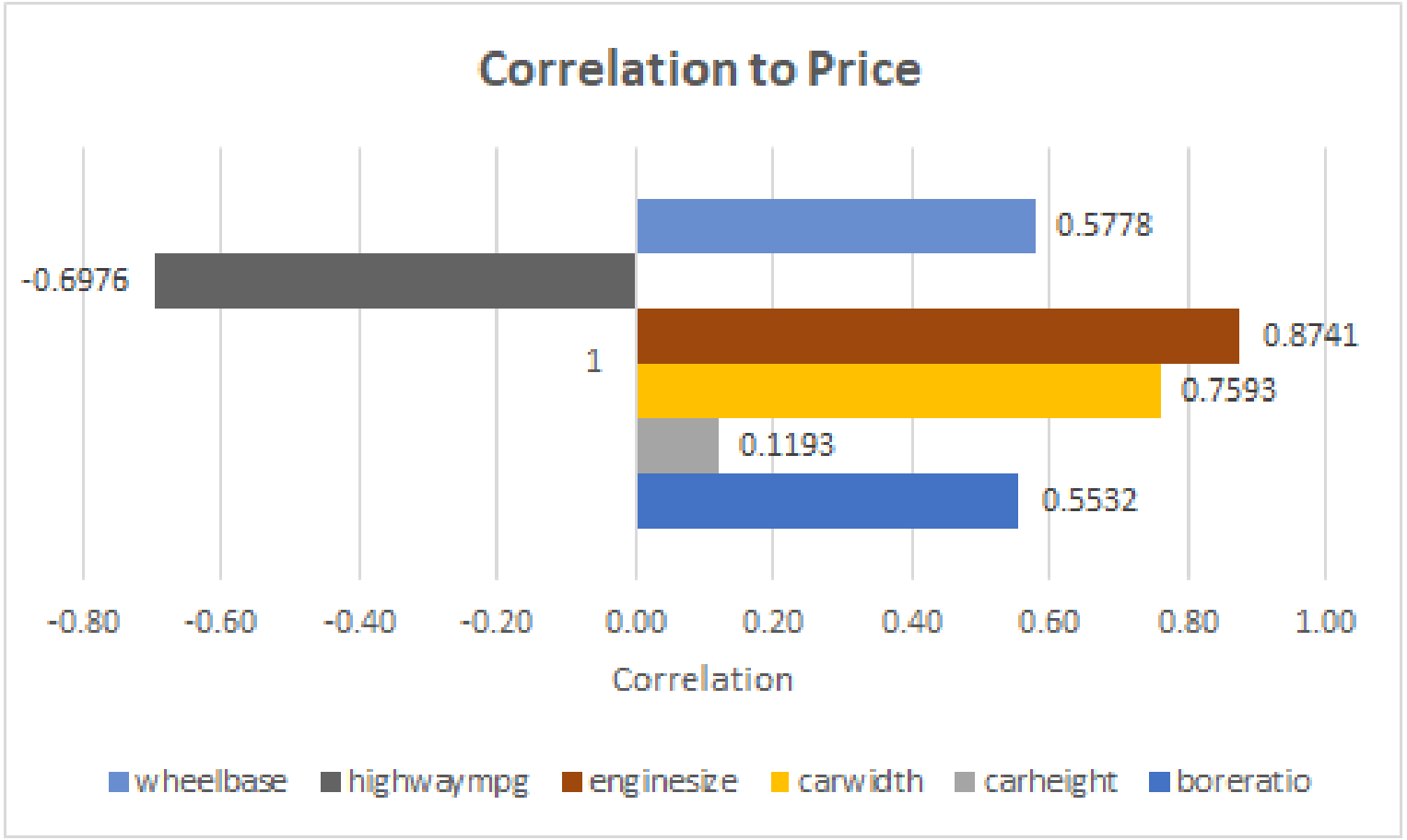


We removed the cross correlations that were higher than 0.8 with other variables (car length, weight, horsepower, city mpg) and the less than 0.1 correlation with price (stroke, compression ratio, peak rpm)

Correlation Matrix

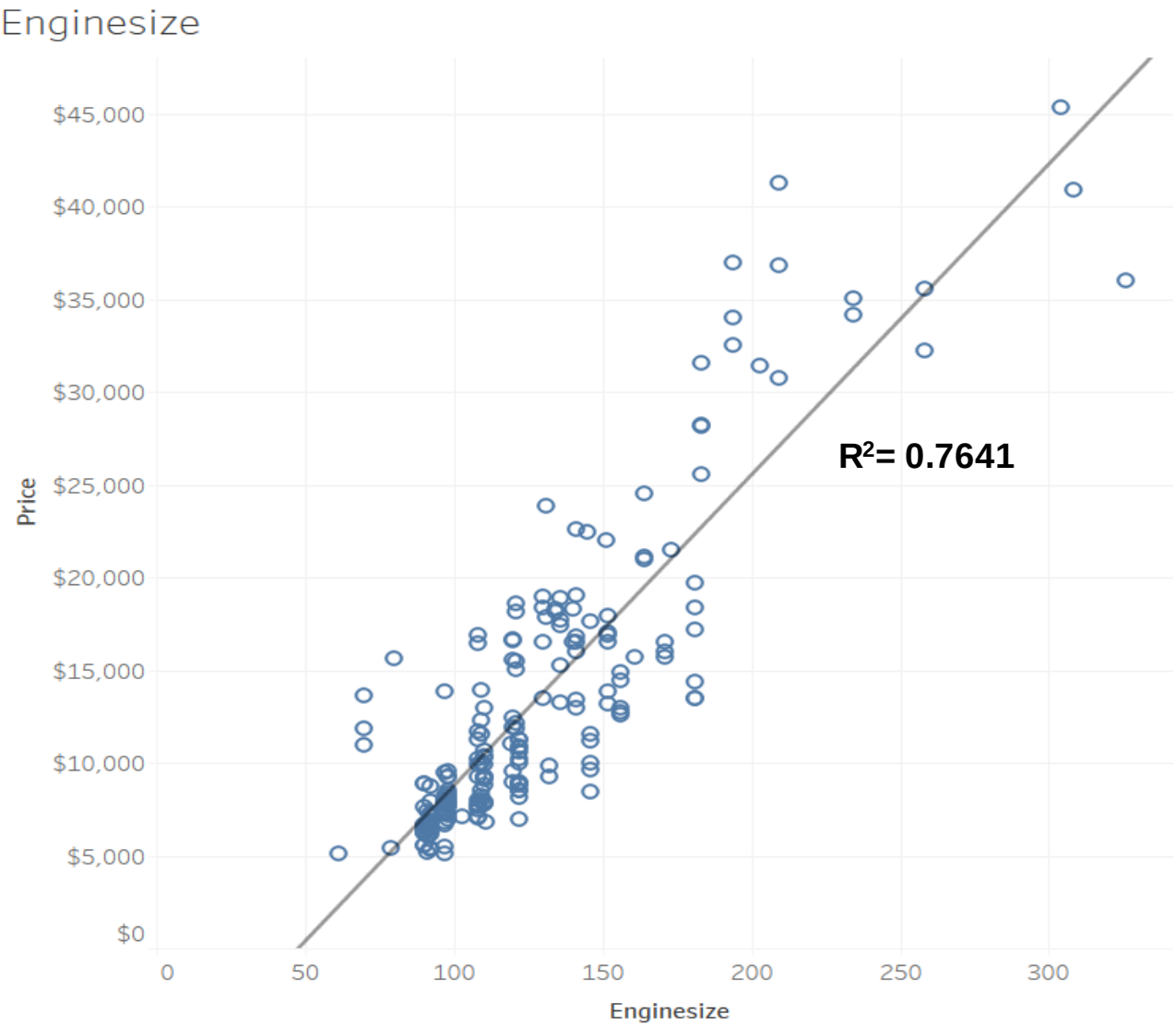
	boreratio	~~carlength~~	carheight	carwidth	~~curbweight~~	~~citympg~~	~~compressio~~	enginesize	highwaym	~~horsepowe~~	~~peakrpm~~	~~stroke~~	wheelbase
boreratio	1.0000												
~~carlength~~	0.6065	1.0000											
carheight	0.1711	0.4910	1.0000										
carwidth	0.5591	~~0.8411~~	0.2792	1.0000									
~~curbweight~~	0.6485	~~0.8777~~	0.2956	0.8670	1.0000								
~~citympg~~	-0.5845	-0.6709	-0.0486	-0.6427	-0.7574	1.0000							
~~compressionratio~~	0.0052	0.1584	0.2612	0.1811	0.1514	0.3247	1.0000						
enginesize	0.5838	0.6834	0.0671	0.7354	~~0.8506~~	-0.6537	0.0290	1.0000					
highwaympg	-0.5870	-0.7047	-0.1074	-0.6772	-0.7975	~~0.9713~~	0.2652	-0.6775	1.0000				
~~horsepower~~	0.5737	0.5526	-0.1088	0.6407	0.7507	~~-0.8015~~	-0.2043	~~0.8098~~	-0.7705	1.0000			
~~peakrpm~~	-0.2550	-0.2872	-0.3204	-0.2200	-0.2662	-0.1135	-0.4357	-0.2447	-0.0543	0.1311	1.0000		
~~stroke~~	-0.0559	0.1295	-0.0553	0.1829	0.1688	-0.0421	0.1861	0.2031	-0.0439	0.0809	-0.0680	1.0000	
wheelbase	0.4887	0.8746	0.5894	0.7951	0.7764	-0.4704	0.2498	0.5693	-0.5441	0.3533	-0.3605	0.1610	1.0000
price	0.5532	~~0.6829~~	0.1193	0.7593	~~0.8353~~	~~-0.6858~~	~~0.0680~~	0.8741	-0.6976	~~0.8081~~	~~-0.0853~~	~~0.0794~~	0.5778

Of the variables remaining, Engine Size, Car width, and Highway MPG have high correlations with price; Wheelbase and Bore ratio moderately correlate*



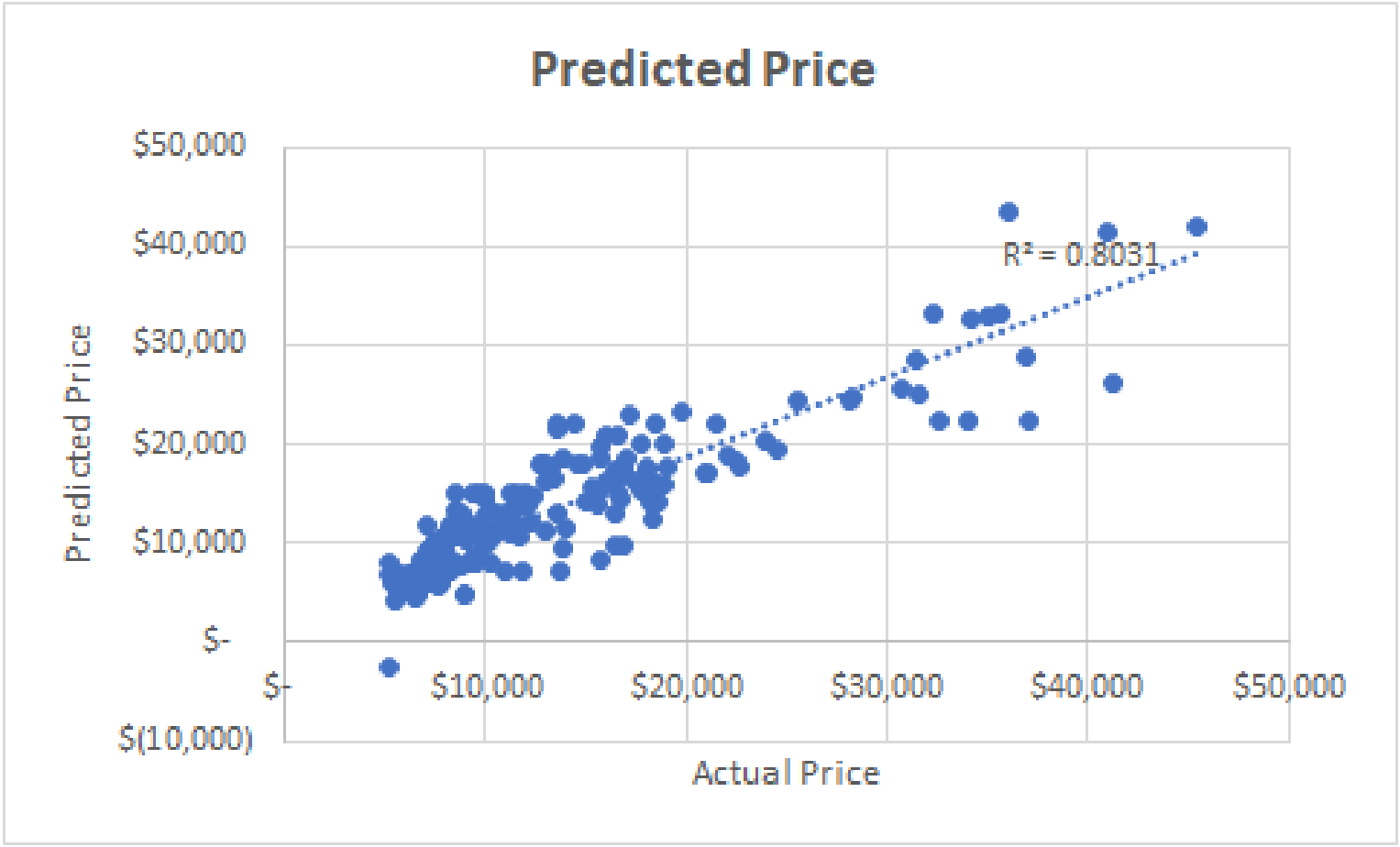
*Scale for correlation: Low < 0.4; 0.4 <= Moderate < 0.7; High >= 0.7

Engine Size by itself returns an R2 of 0.76 with price.



Adding in Bore ratio, car height, car width, engine size, highway mpg, and wheelbase increases the R2 incrementally to 0.8031.

Regression: Predicted Price vs. Actual Price



Adding the categorical variables to the pricing model in a future phase could provide some incremental improvements

Descriptive Statistics Summary

Variable	Impact on Price	Evaluation for Pricing Model Improvement
Cylinder number	High	Important, but will be covered by Engine Size in Numerical Variables
Aspiration Type	~50% increase in Median	Add as dummy variables in next phase
Car body type	Low	Not significant
Drive wheel	50% increase in Median from fwd to rwd	Add as dummy variables in next phase
Door number	Low	Not significant
Engine type	50% increase in Median	Add as dummy variables in next phase
Fuel system	50% increase in Median	Add as dummy variables in next phase
Fuel type	50% increase in Median	Add as dummy variables in next phase

Recommendation: We can plan our product line in categories primarily based on engine size (i.e. economy to luxury), and use the remaining variables to fine tune the price of each category.

Conclusions

- **Among the categorical variables, Price shows the most variation with Aspiration Type, Drive Wheel, Engine Type, Fuel System, and Fuel Type**
- **Among the numerical variables, after eliminating high cross-correlations, Engine Size, Car width, and Highway MPG have high correlations with price; Wheelbase and Bore ratio moderately correlate**
- **Engine Size by itself returns an R^2 of 0.76 with price**
- **Adding in Bore ratio, car height, car width, engine size, highway mpg, and wheelbase increases the R^2 incrementally to 0.8031**