Jalopies, Beaters, and Hoopties

Hey, have I got a deal for YOU!

Our Exploration

Can we predict with a reasonable amount of accuracy the price of used cars based on mileage?

Can we predict with a reasonable amount of accuracy the price of used cars based on City/Highway MPG?

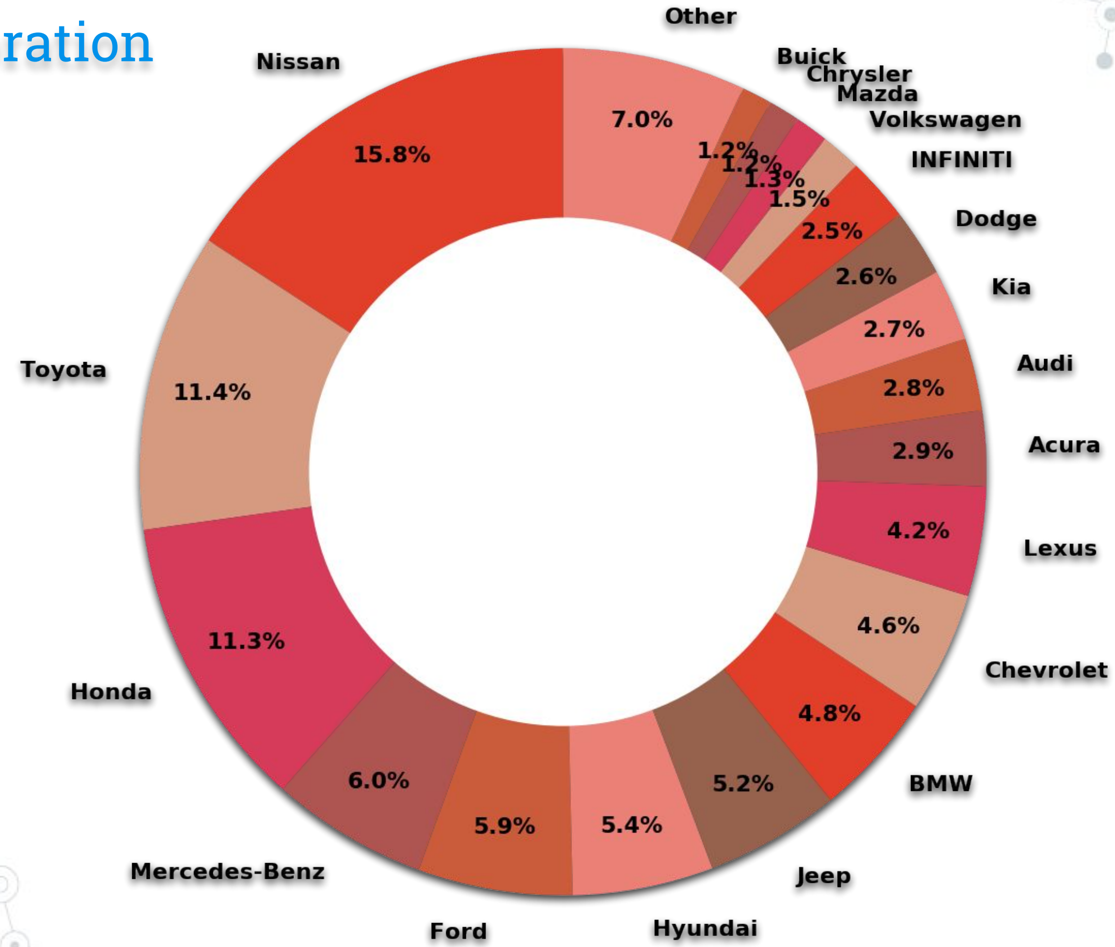
Can we predict with a reasonable amount of accuracy the price of used cars based on any other variables?

Our Data

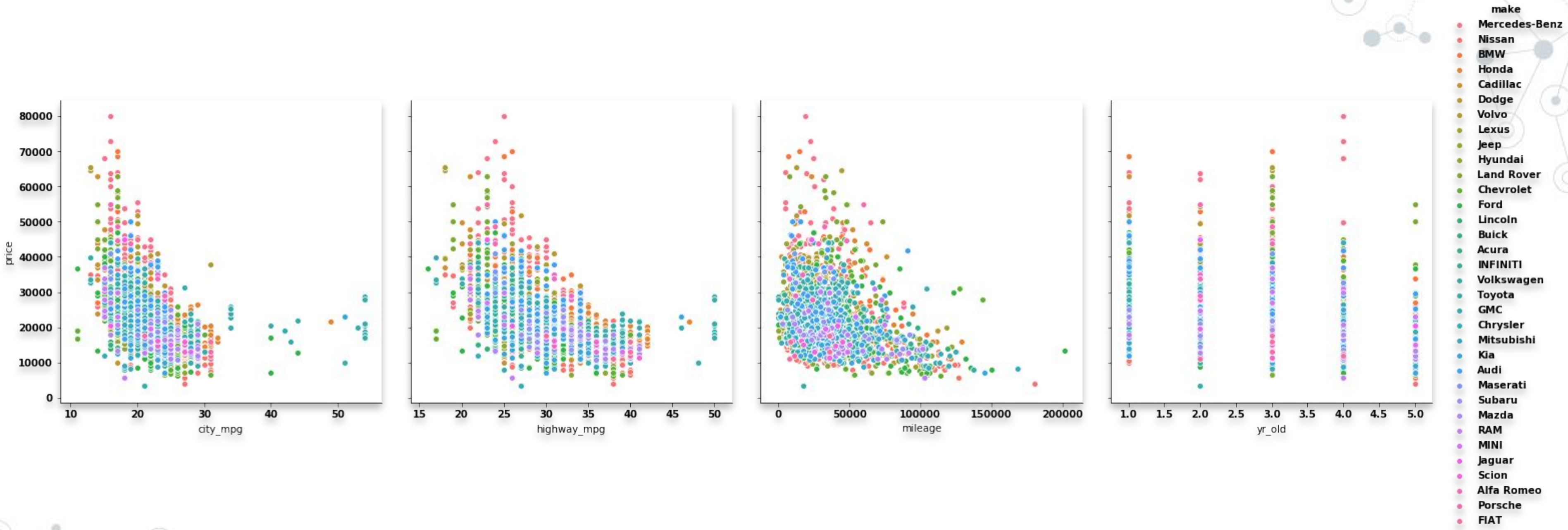
Used Car data

- Scraped from Cars.com
- 2677 entries
- Within 10 miles of zip code 10004
- All Available Makes and Models
- Years from 2014 to 2018
- 13 variables - 4 continuous, 9 categorical

Initial Exploration



More Initial Exploration



First Regression Attempts

Price/City MPG - R^2 = .237, Coeff = -925.69

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.237
Model:	OLS	**Adj. R-squared:**	0.237
Method:	Least Squares	**F-statistic:**	831.3
Date:	Thu, 22 Aug 2019	**Prob (F-statistic):**	2.06e-159
Time:	15:08:17	**Log-Likelihood:**	-27793.
No. Observations:	2677	**AIC:**	5.559e+04
Df Residuals:	2675	**BIC:**	5.560e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.296e+04	734.450	58.493	0.000	4.15e+04	4.44e+04
city_mpg	-925.5909	32.103	-28.832	0.000	-988.540	-862.641

Omnibus:	825.199	**Durbin-Watson:**	1.506
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3020.224
Skew:	1.498	**Prob(JB):**	0.00
Kurtosis:	7.255	**Cond. No.**	111.

Price/Highway MPG - R^2 = .342, Coeff -992.47

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.342
Model:	OLS	**Adj. R-squared:**	0.342
Method:	Least Squares	**F-statistic:**	1392.
Date:	Thu, 22 Aug 2019	**Prob (F-statistic):**	1.26e-245
Time:	15:08:21	**Log-Likelihood:**	-27595.
No. Observations:	2677	**AIC:**	5.519e+04
Df Residuals:	2675	**BIC:**	5.521e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.234e+04	819.084	63.904	0.000	5.07e+04	5.39e+04
highway_mpg	-992.4719	26.603	-37.306	0.000	-1044.637	-940.306

Omnibus:	734.514	**Durbin-Watson:**	1.504
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3068.719
Skew:	1.277	**Prob(JB):**	0.00
Kurtosis:	7.581	**Cond. No.**	180.

Price/Mileage - R^2 = .088, Coeff - 0.1150

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.088
Model:	OLS	**Adj. R-squared:**	0.087
Method:	Least Squares	**F-statistic:**	257.5
Date:	Thu, 22 Aug 2019	**Prob (F-statistic):**	2.17e-55
Time:	15:08:26	**Log-Likelihood:**	-28033.
No. Observations:	2677	**AIC:**	5.607e+04
Df Residuals:	2675	**BIC:**	5.608e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.642e+04	308.546	85.625	0.000	2.58e+04	2.7e+04
mileage	-0.1150	0.007	-16.046	0.000	-0.129	-0.101

Omnibus:	833.662	**Durbin-Watson:**	1.588
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2791.534
Skew:	1.552	**Prob(JB):**	0.00
Kurtosis:	6.923	**Cond. No.**	8.04e+04

Let's Smash them All Together!!

Dep. Variable:	price	**R-squared:**	0.709
Model:	OLS	**Adj. R-squared:**	0.705
Method:	Least Squares	**F-statistic:**	173.6
Date:	Thu, 22 Aug 2019	**Prob (F-statistic):**	0.00
Time:	15:08:39	**Log-Likelihood:**	-26504.
No. Observations:	2677	**AIC:**	5.308e+04
Df Residuals:	2639	**BIC:**	5.331e+04
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.708e+04	871.075	65.533	0.000	5.54e+04	5.88e+04
make[T.Alfa Romeo]	5920.8642	2494.842	2.373	0.018	1028.820	1.08e+04
make[T.Audi]	3231.9277	793.885	4.071	0.000	1675.228	4788.627
make[T.BMW]	4545.6909	704.724	6.450	0.000	3163.824	5927.558
make[T.Buick]	-4727.1023	1037.406	-4.557	0.000	-6761.313	-2692.891
make[T.Cadillac]	5626.3107	1106.880	5.083	0.000	3455.869	7796.752
make[T.Chevrolet]	-3113.6917	708.928	-4.392	0.000	-4503.803	-1723.581
make[T.Chrysler]	-5002.6407	1025.450	-4.878	0.000	-7013.409	-2991.873
make[T.Dodge]	-3283.9446	809.400	-4.057	0.000	-4871.067	-1696.823
make[T.FIAT]	-5192.8379	2254.114	-2.304	0.021	-9612.848	-772.828
make[T.Ford]	-4414.1757	682.625	-6.466	0.000	-5752.709	-3075.642
make[T.GMC]	-927.2595	1078.684	-0.860	0.390	-3042.411	1187.892
make[T.Honda]	-2332.8421	632.262	-3.690	0.000	-3572.620	-1093.064
make[T.Hyundai]	-7597.6693	691.250	-10.991	0.000	-8953.117	-6242.222
make[T.INFINITI]	-2278.3524	823.883	-2.765	0.006	-3893.875	-662.830
make[T.Jaguar]	4788.7573	3484.871	1.374	0.170	-2044.598	1.16e+04
make[T.Jeep]	-2911.2742	699.921	-4.159	0.000	-4283.725	-1538.824
make[T.Kia]	-6781.6771	800.813	-8.468	0.000	-8351.961	-5211.393
make[T.Land Rover]	1.606e+04	1131.645	14.191	0.000	1.38e+04	1.83e+04
make[T.Lexus]	4216.8402	729.772	5.778	0.000	2785.856	5647.824
make[T.Lincoln]	-4200.1212	1715.728	-2.448	0.014	-7564.430	-835.813
make[T.MINI]	-3329.9167	2863.865	-1.163	0.245	-8945.564	2285.731
make[T.Maserati]	5850.5880	2247.893	2.603	0.009	1442.776	1.03e+04
make[T.Mazda]	-2823.8576	1011.839	-2.791	0.005	-4807.935	-839.781
make[T.Mercedes-Benz]	8132.6908	681.234	11.938	0.000	6796.885	9468.497
make[T.Mitsubishi]	-1.156e+04	1722.878	-6.709	0.000	-1.49e+04	-8180.700
make[T.Nissan]	-4342.7912	613.190	-7.082	0.000	-5545.174	-3140.409
make[T.Porsche]	1.911e+04	2496.940	7.652	0.000	1.42e+04	2.4e+04
make[T.RAM]	-1750.8237	1467.937	-1.193	0.233	-4629.248	1127.600
make[T.Scion]	-5601.4672	1925.215	-2.910	0.004	-9376.551	-1826.384
make[T.Subaru]	-3941.0746	1097.019	-3.593	0.000	-6092.178	-1789.971
make[T.Toyota]	-1754.9042	637.001	-2.755	0.006	-3003.977	-505.832
make[T.Volkswagen]	-7502.0744	943.669	-7.950	0.000	-9352.480	-5651.669
make[T.Volvo]	5701.0397	1224.362	4.656	0.000	3300.233	8101.847
mileage	-0.0921	0.005	-18.812	0.000	-0.102	-0.083
city_mpg	164.2698	46.641	3.522	0.000	72.812	255.727
highway_mpg	-968.1110	41.309	-23.436	0.000	-1049.113	-887.109
yr_old	-1508.3209	112.209	-13.442	0.000	-1728.348	-1288.294

Omnibus:	951.951	**Durbin-Watson:**	1.810
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9093.539
Skew:	1.405	**Prob(JB):**	0.00
Kurtosis:	11.581	**Cond. No.**	1.72e+06

Fun Facts:

R^2 = 0.709
P-Values < .05
Control Brand: Acura

Notable Coefficients:

- Used Mercedes will get you $8133 more.
- Used BMW will get you $4546 more.
- GMCs hold their value well
- Hyundai's will get you $7597 less.

What is the price of luxury?

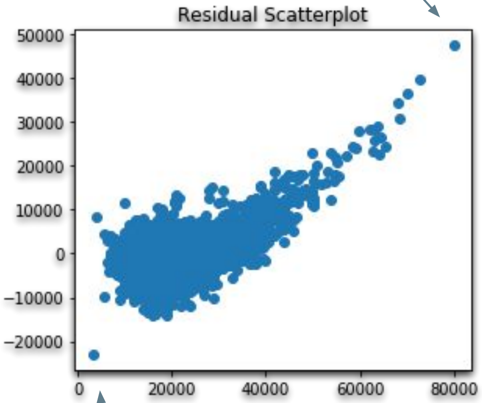
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.625
Model:	OLS	**Adj. R-squared:**	0.624
Method:	Least Squares	**F-statistic:**	889.0
Date:	Thu, 22 Aug 2019	**Prob (F-statistic):**	0.00
Time:	16:04:44	**Log-Likelihood:**	-26844.
No. Observations:	2677	**AIC:**	5.370e+04
Df Residuals:	2671	**BIC:**	5.374e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5.405e+04	737.279	73.315	0.000	5.26e+04	5.55e+04
mileage	-0.1004	0.005	-18.684	0.000	-0.111	-0.090
city_mpg	276.5644	49.124	5.630	0.000	180.239	372.890
highway_mpg	-1084.1649	43.341	-25.015	0.000	-1169.151	-999.179
yr_old	-1301.6877	123.100	-10.574	0.000	-1543.069	-1060.307
luxury	8208.0310	252.329	32.529	0.000	7713.251	8702.811

Omnibus:	918.893	**Durbin-Watson:**	1.765
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6570.900
Skew:	1.437	**Prob(JB):**	0.00
Kurtosis:	10.117	**Cond. No.**	3.01e+05

Outlier:
2015 Certified Mercedes Benz AMG,
18954 miles, $79,901



Outlier:
2017 Toyota Highlander,
17,297 miles, $3,298.

Fun Facts:

$R^2 = 0.625$

P-Values = 0

Notable Numbers:

Luxury coefficient:

$8208.00

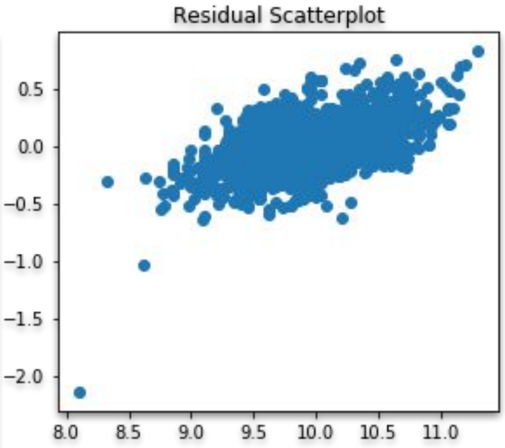
Unless you are selling a Lincoln

Log Log Log

OLS Regression Results

Dep. Variable:	price_log	**R-squared:**	0.752
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	211.0
Date:	Thu, 22 Aug 2019	**Prob (F-statistic):**	0.00
Time:	16:29:16	**Log-Likelihood:**	671.29
No. Observations:	2677	**AIC:**	-1265.
Df Residuals:	2638	**BIC:**	-1035.
Df Model:	38		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.5796	0.034	339.368	0.000	11.513	11.646
make[T.Alfa Romeo]	0.2076	0.097	2.132	0.033	0.017	0.399
make[T.Audi]	0.1037	0.031	3.346	0.001	0.043	0.164
make[T.BMW]	0.1743	0.028	6.338	0.000	0.120	0.228
make[T.Buick]	-0.2231	0.040	-5.509	0.000	-0.302	-0.144
make[T.Cadillac]	0.1342	0.043	3.107	0.002	0.050	0.219
make[T.Chevrolet]	-0.1928	0.028	-6.967	0.000	-0.247	-0.139
make[T.Chrysler]	-0.2599	0.040	-6.495	0.000	-0.338	-0.181
make[T.Dodge]	-0.1560	0.032	-4.940	0.000	-0.218	-0.094
make[T.FIAT]	-0.3294	0.088	-3.744	0.000	-0.502	-0.157
make[T.Ford]	-0.2427	0.027	-9.110	0.000	-0.295	-0.190
make[T.GMC]	-0.0886	0.042	-2.104	0.035	-0.171	-0.006
make[T.Honda]	-0.1156	0.025	-4.686	0.000	-0.164	-0.067



Fun Facts:

R^2 = 0.752 (highest yet!)

P-Values = 0.00 - 0.544

Notable Numbers:

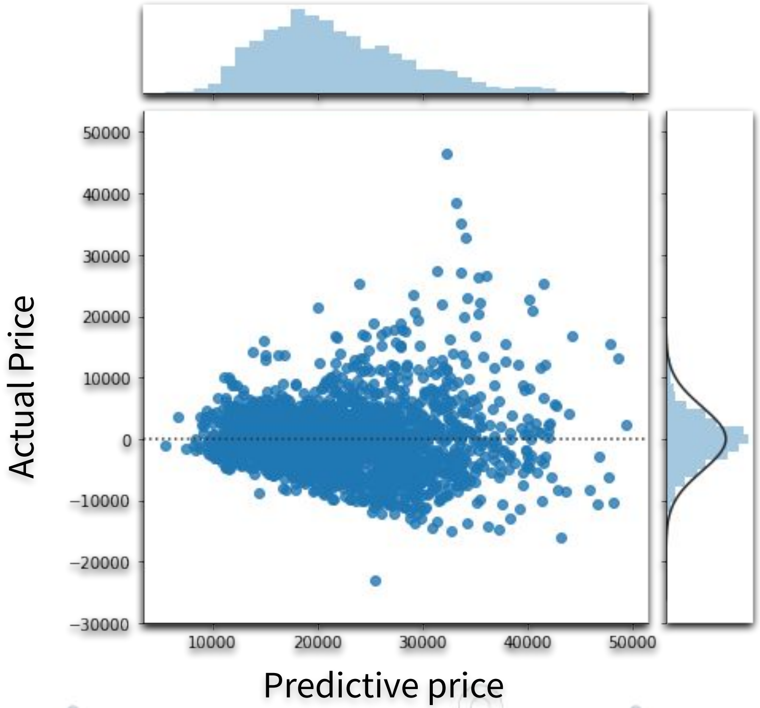
Jaguars!

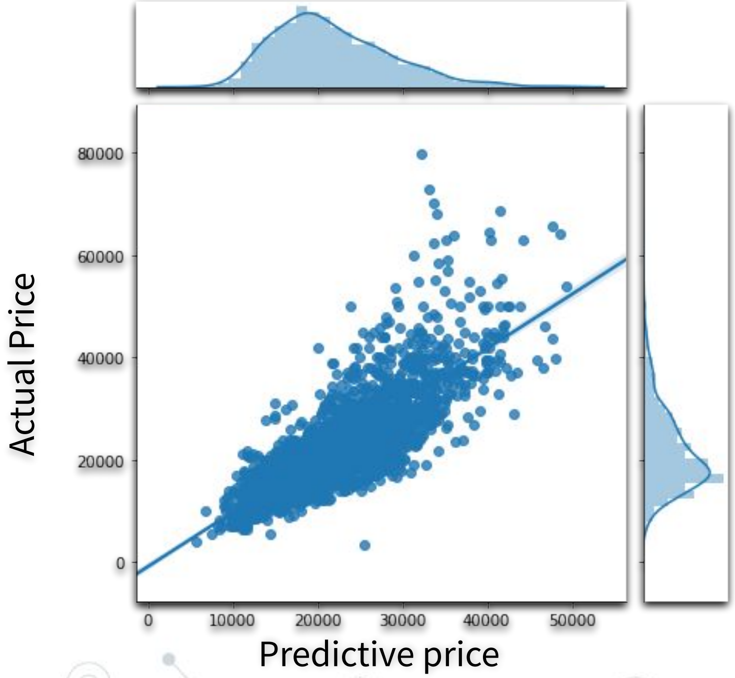
Simplified Prediction Model

Used Car Price = 11.3940 + (-5.499e-6 * (mileage)) +
(0.0044 * (low_mileage)) +
(0.0133 * (city_mpg)) +
(-0.0489 * (highway_mpg)) +
(-0.0569*(yr_old)) +
(0.3272 * (luxury))

Value above is the Natural Log of Price

Testing our Predictive Model





Our Initial Predictive Formula

```
In [21]: ins = input("""Enter the parameters in the order shown below separated by commas
                        1. Max mileage(integer and less than 80000mi),
                        2. required city_mpg(integer and between 11 - 54 only),
                        3. required highway_mpg(integer and between 16-50 only),
                        4. year (integer and between 2018-2014 only),
                        5. luxury brand? (if yes enter 1 otherwise 0)""")

         Enter the parameters in the order shown below separated by commas
                        1. Max mileage(integer and less than 80000mi),
                        2. required city_mpg(integer and between 11 - 54 only),
                        3. required highway_mpg(integer and between 16-50 only),
                        4. year (integer and between 2018-2014 only),
                        5. luxury brand? (if yes enter 1 otherwise 0)50000, 15, 20, 2018, 0

In [22]: x = ins.split(",")
         x.insert(1, 'replace me!')
         x[0] = int(x[0])
         x[1] = 1 if x[0] <= 7500 else 0
         x[2] = int(x[2])
         x[3] = int(x[3])
         x[4] = 2019 - int(x[4])
         x[5] = int(x[5])

In [23]: model1(x)

Out[23]: 'Expected resale value is $29324.0'
```

Questions left to answer

What effect does "Certified" have on luxury vehicles?

What effect does engine type have on all vehicles?

How can we control for MPG, which seems to be less valuable in newer cars?

Why are our outliers… outliers? Can we control for that?

Can we further refine our model to take into account car models and sizes?

Will our model hold true outside of the Tri-state area?