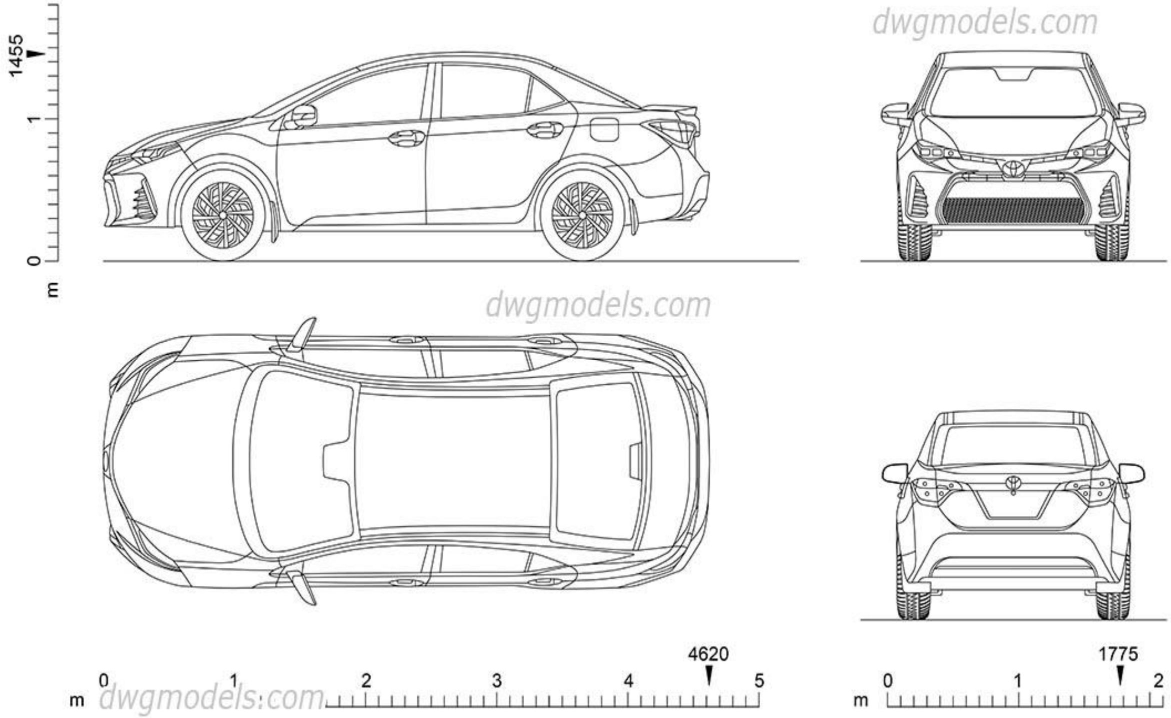
#making resale values transparent



#outline

- ➔ Key take-aways
- ➔ Introduction/background
- ➔ Competitive analysis
- ➔ Approach:
  - Data source/approach/key questions
  - Data exploration/data cleaning/data prep
  - Feature selection
  - Principal component analysis
  - Model parameters selected
  - Results
- ➔ Conclusion/cost benefit analysis

#introduction/background

- ➔ The purchase of a used car is an important life decision that creates some anxiety about whether the price is a true reflection of the value of a car.
- ➔ Our project is to premised on one of the 5 questions data science can answer “How much is the price of a used Toyota Corolla?” framed as “What is the fair price of a used Toyota Corolla?”

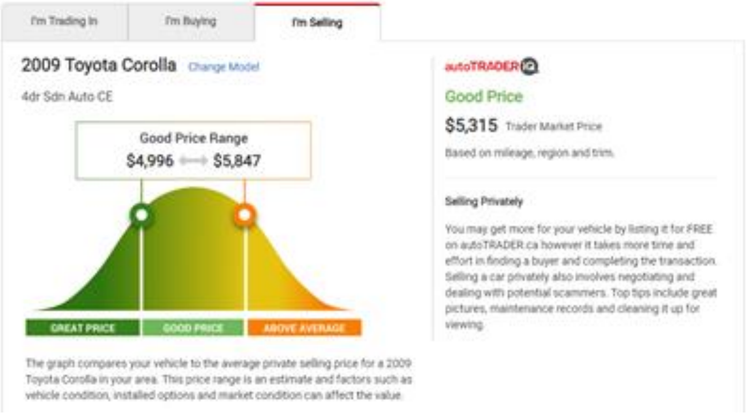
#our value proposition

➔ We offer an unbiased view as the team members of Group 4 have no affiliations with any Toyota Car Dealership for new or used cars

➔ Our methodology is premised on the CRISP-DM approach and

➔ Our model is simple and transparent and includes an evaluation of our results.

#competitive analysis



- ➔ Several companies offer online car valuations for buyers and sellers of used cars
- ➔ But using the same parameters generates different results, and businesses may not be entirely impartial
- ➔ Further, users can’t see the inner workings

#questions and approach

➔ ***Questions:***

- *Which features are most predictive of resale price for used Toyota Corollas?*
- *What is a fair price for a used Toyota Corolla based on these features?*

➔ ***Approach:***

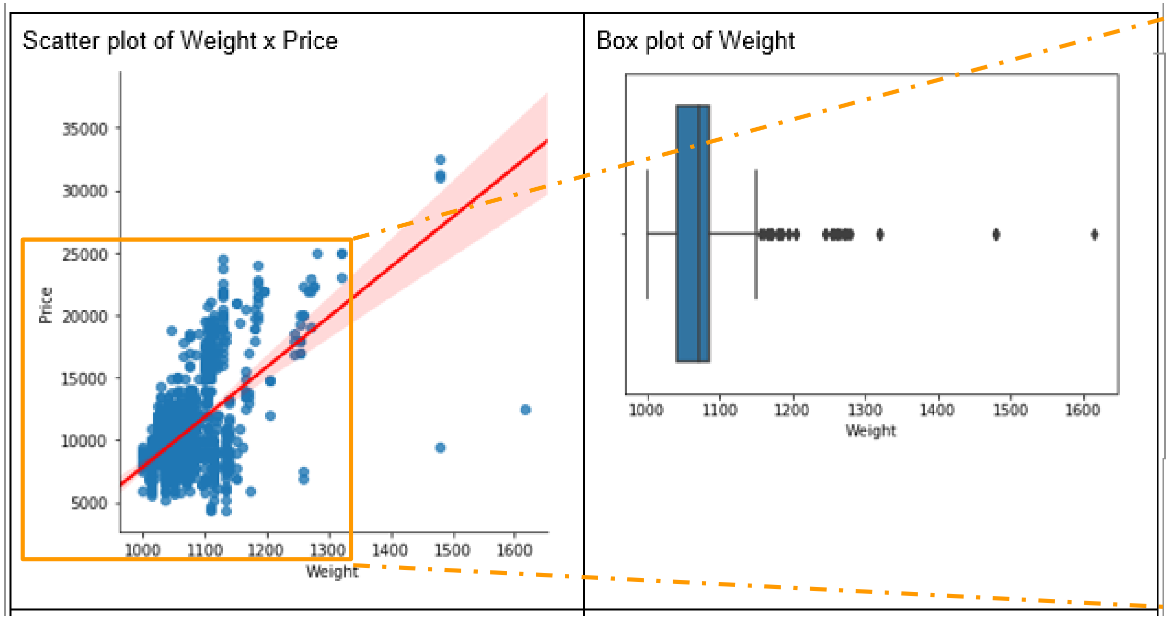
- *Explored data to understand pairwise relationships, identified and removed outliers, and convert qualitative to numeric values.*
- *Ranked correlations of all features*
- *Assessed statistical significance of features with highest correlations,*
- *Conducted Principal Component Analysis to narrow down which components explained maximum variance?*

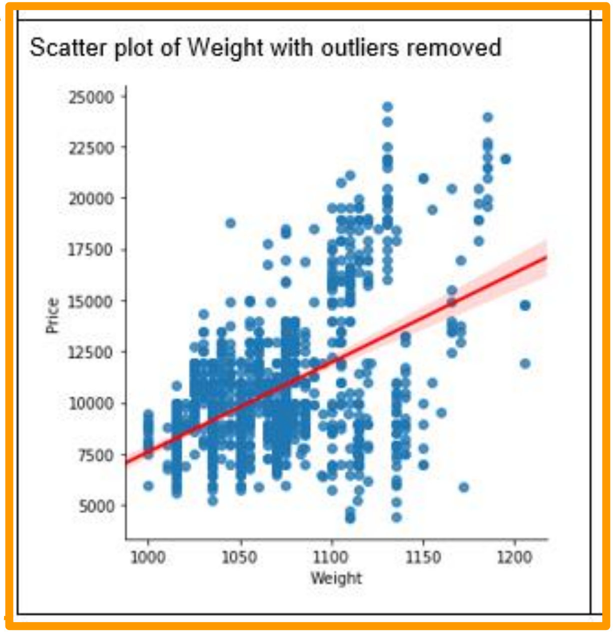
#key take-aways on approach

- *We ranked **42** features (independent variables) in terms of their correlation with price, and found that **4** of these correlations were statistically significant, and together explained 87.3% of variance in price (dependent variable)*

#data exploration and cleaning

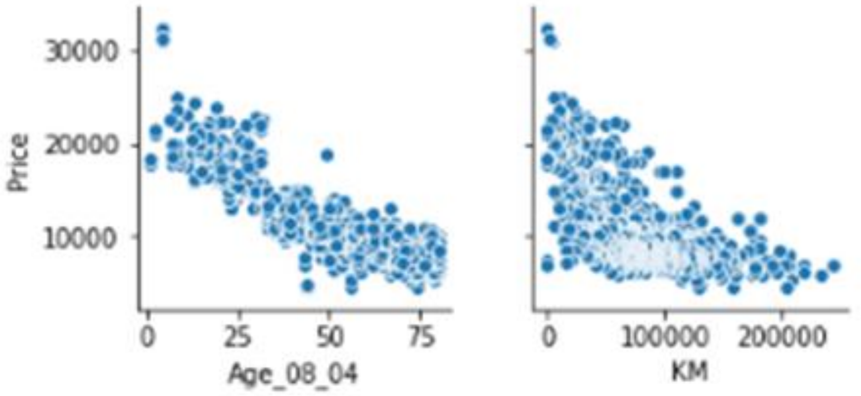
visualizing pairwise relationships → identifying and removing outliers





#data exploration and cleaning

Pair plot analysis gives us some clear relationships to look at.



#feature engineering

- Qualitative Fields were converted to numerical values
- Fuel_Type was broken down into 3 new features one for each of the unique values
- It was observed that Model contained several separate attributes that were broken down into 14 new features

#feature selection

Age_08_04	-0.876377
Boardcomputer	0.603482
Automatic_airco	0.586273
Weight	0.579851
KM	-0.569268
isVVT	0.553608
isLINEA	-0.516154
CD_Player	0.479845
Airco	0.428618
isSOL	0.426239
isD4D	0.37328
Powered_Windows	0.355858
Central_Lock	0.342814
isCOMFORT	0.314709
HP	0.314693
ABS	0.305954
isTERRA	-0.265774
isHATCH	-0.259344
isSPORT	0.248991
Airbag_2	0.248474
Mistlamps	0.223439
Quarterly_Tax	0.219102
isVVTLI	0.207679
Mfr_Guarantee	0.199523
Doors	0.184118
greater.than.85	0.177664
Tow_Bar	-0.171719
Sport_Model	0.165477
cc	0.165085
Guarantee_Period	0.147322
between.69.and.85	-0.129099
isLIFT	-0.110523
Metallic_Rim	0.109572
Met_Color	0.10802
Backseat_Divider	0.105774
Airbag_1	0.093482
isSEDAN	-0.092214
isLUNA	-0.065401
Power_Steering	0.064152
Gears	0.06344
IF_Diesel	0.054734
isWAGON	-0.050678
Radio_cassette	-0.042606
Radio	-0.04131
IF_cng	-0.039434
IF_petrol	-0.039171
BOVAG_Guarantee	0.032916
less.than.69	-0.030947
Automatic	0.026783

Top 7 features with strongest individual correlation to car price:

1. age
2. board computer
3. automatic aircon
4. weight
5. KMs
6. isVVT
7. isLINEA

#feature selection

Pearson Correlation Analysis

Full Correlation Matrix

	Age_08_04	KM	Weight	Automatic_airco	Boardcomputer	isLINEA	isVVT
Age_08_04	1	0.504953	-0.46902	-0.4244	-0.697177	0.635095	-0.67845
KM	0.504953	1	-0.02681	-0.25618	-0.354703	0.300964	-0.43118
Weight	-0.469018	-0.02681	1	0.427774	0.275686	-0.22195	0.065032
Automatic_airco	-0.424404	-0.25618	0.427774	1	0.27591	-0.19749	0.207213
Boardcomputer	-0.697177	-0.3547	0.275686	0.27591	1	-0.68982	0.741605
isVVT	-0.678453	-0.43118	0.065032	0.207213	0.741605	-0.67539	1
isLINEA	0.635095	0.300964	-0.22195	-0.19749	-0.689817	1	-0.67539

#principal component analysis

first attempt

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1133.9822	1.03E+03	1.104	0.26994	
Age_08_04	-112.6529	3.59E+00	-31.409	< 2.2e-16	***
KM	-0.0211	1.11E-03	-19.041	< 2.2e-16	***
Weight	15.9166	8.59E-01	18.533	< 2.2e-16	***
Automatic_airco	2963.8422	1.73E+02	17.163	< 2.2e-16	***
Boardcomputer	-233.4155	1.30E+02	-1.8	0.07212	.
isVVT	352.1773	1.29E+02	2.737	0.00628	**
isLINEA	114.3505	1.03E+02	1.115	0.2652	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Min	1Q	Median	3Q	Max
-8056.3	-721.1	13.7	711.4	6200.6

second attempt

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1249.2376	1.03E+03	1.218	0.22331	
Age_08_04	-109.5095	3.35E+00	-32.649	< 2.2e-16	***
KM	-0.02135	1.11E-03	-19.322	< 2.2e-16	***
Weight	15.70796	8.56E-01	18.355	< 2.2e-16	***
Automatic_airco	2978.8541	1.73E+02	17.264	< 2.2e-16	***
isVVT	177.84254	1.07E+02	1.657	0.09773	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Min	1Q	Median	3Q	Max
-8065.77	-711.45	3.44	728.8	6452.64

#model parameters selected

1 **Report for Linear Model Linear_Regression_6**

2 *Basic Summary*

Call:

3 lm(formula = Price ~ Age_08_04 + KM + Weight + Automatic_airco, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-8057.7	-711.6	-3.2	710.6	6489.8

6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2047.4308	9.06E+02	2.26	0.02394	*
Age_08_04	-113.1616	2.53E+00	-44.73	< 2.2e-16	***
KM	-0.02142	1.11E-03	-19.39	< 2.2e-16	***
Weight	15.21296	8.02E-01	18.96	< 2.2e-16	***
Automatic_airco	2974.1462	1.73E+02	17.23	< 2.2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1288.3 on 1430 degrees of freedom

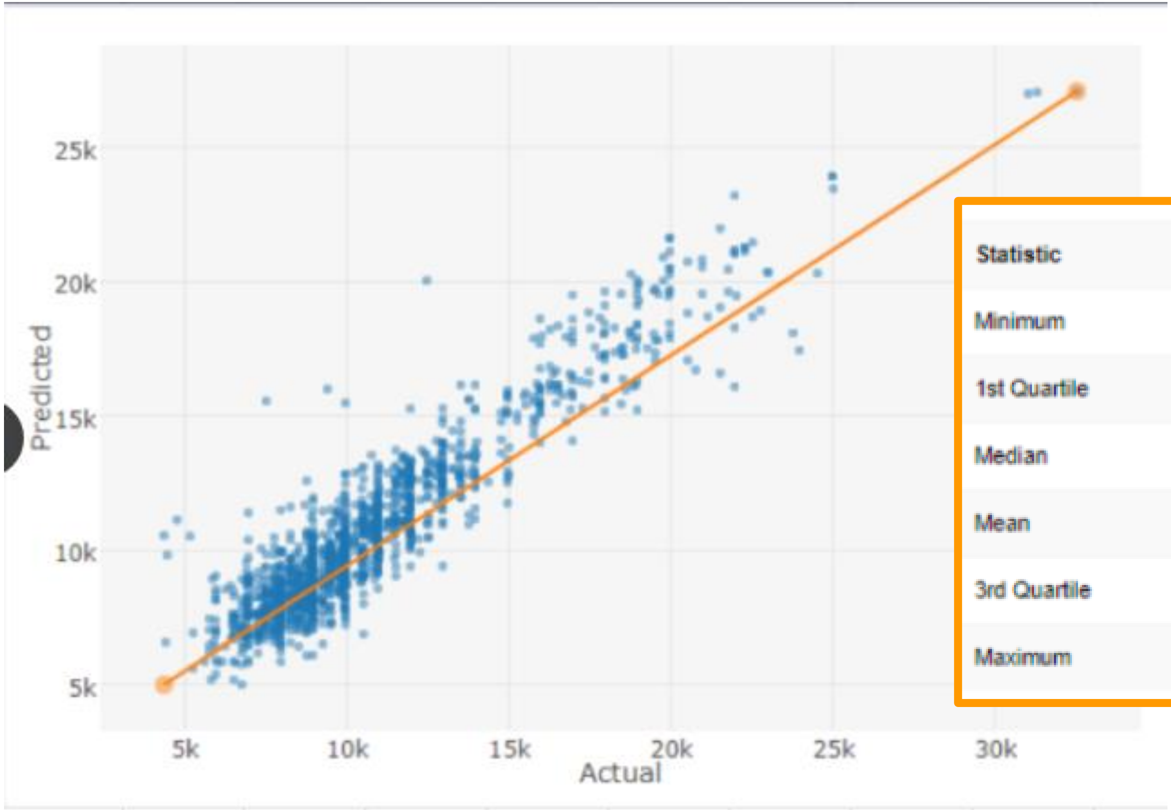
8 Multiple R-squared: 0.8738, Adjusted R-Squared: 0.8735

F-statistic: 2476 on 4 and 1430 degrees of freedom (DF), p-value < 2.2e-16

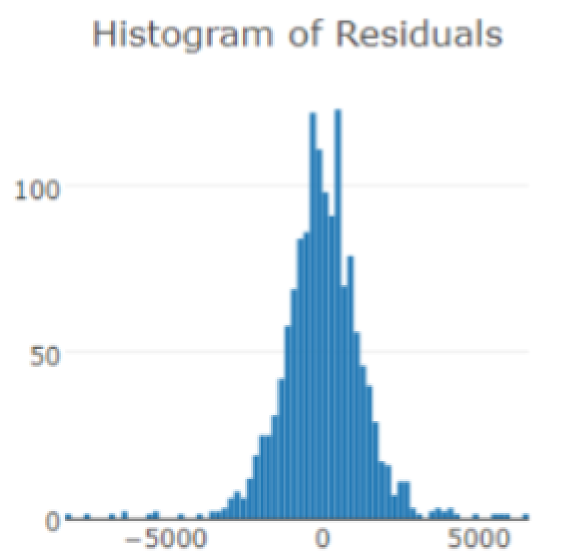
9 *Type II ANOVA Analysis*

Predictive equation:
2047.4308
-113.16166xAge_08_04
-0.02142xKM
+15.21296xWeight
+2974.14627xAutomatic_airco

#model results



Statistic	Value
Minimum	-8057.7
1st Quartile	-711.6
Median	-3.2
Mean	0.0
3rd Quartile	[illegible]
Maximum	6489.8



#conclusion

- *Not surprisingly, age and mileage were most predictive of resale value -- this is consistent with competitor models*
- *More surprisingly, weight was the third-ranked feature in terms of predictive power, suggesting that ‘size matters’*
- *This may be useful information for prospective buyers to find value in smaller vehicles*
- *Air conditioning, also not included in competitor models, was next most predictive*
- *Communicating model results with respect to fair pricing could help them avoid purchasing lemons!*