

Used Car Value Prediction



[Click for Github docs and code](#)

by Mustafa Aytug KAYA

Problem Statement

Car Value Prediction Is Essential for Individuals and Companies

- Banks
- Insurance
- Car Retailers
- Vehicle Manufacturers

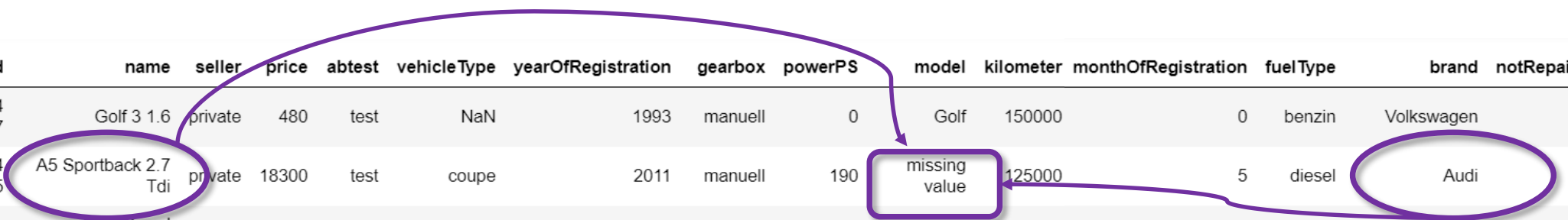


Kelley Blue Book



Dataset

- Arbitrary – Not professionally prepared.
- German Language
- Missing Values
 - Not predictable
 - Filled by special code



The diagram illustrates data flow and missing values using purple annotations. A circle highlights the 'name' column for the second row ('A5 Sportback 2.7 Tdi'). An arrow points from this circle to the 'model' column of the same row, which contains the text 'missing value'. Another circle highlights the 'brand' column for the second row ('Audi'). A second arrow points from the 'missing value' in the 'model' column to this 'Audi' brand entry, indicating that the model information is used to determine the brand.

dateCrawled	name	seller	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage
2016-03-24 11:52:17	Golf 3 1.6	private	480	test	NaN	1993	manuell	0	Golf	150000	0	benzin	Volkswagen	NaN
2016-03-24 10:58:45	A5 Sportback 2.7 Tdi	private	18300	test	coupe	2011	manuell	190	missing value	125000	5	diesel	Audi	ja
2016-03-14 12:52:21	Jeep Grand Cherokee "Overland"	private	9800	test	suv	2004	automatik	163	Grand Cherokee	125000	8	diesel	Jeep	NaN
2016-03-17 16:54:04	GOLF 4 1.4 3TÜRER	private	1500	test	kleinwagen	2001	manuell	75	Golf	150000	6	benzin	Volkswagen	nein
2016-03-31 17:25:20	Skoda Fabia 1.4 TDI PD Classic	private	3600	test	kleinwagen	2008	manuell	69	Fabia	90000	7	diesel	Skoda	nein
2016-04-04 17:36:23	BMW 316i e36 Limousine Bastlerfahrzeug	private	650	test	limousine	1995	manuell	102	3er	150000	10	benzin	Bmw	ja

Dataset

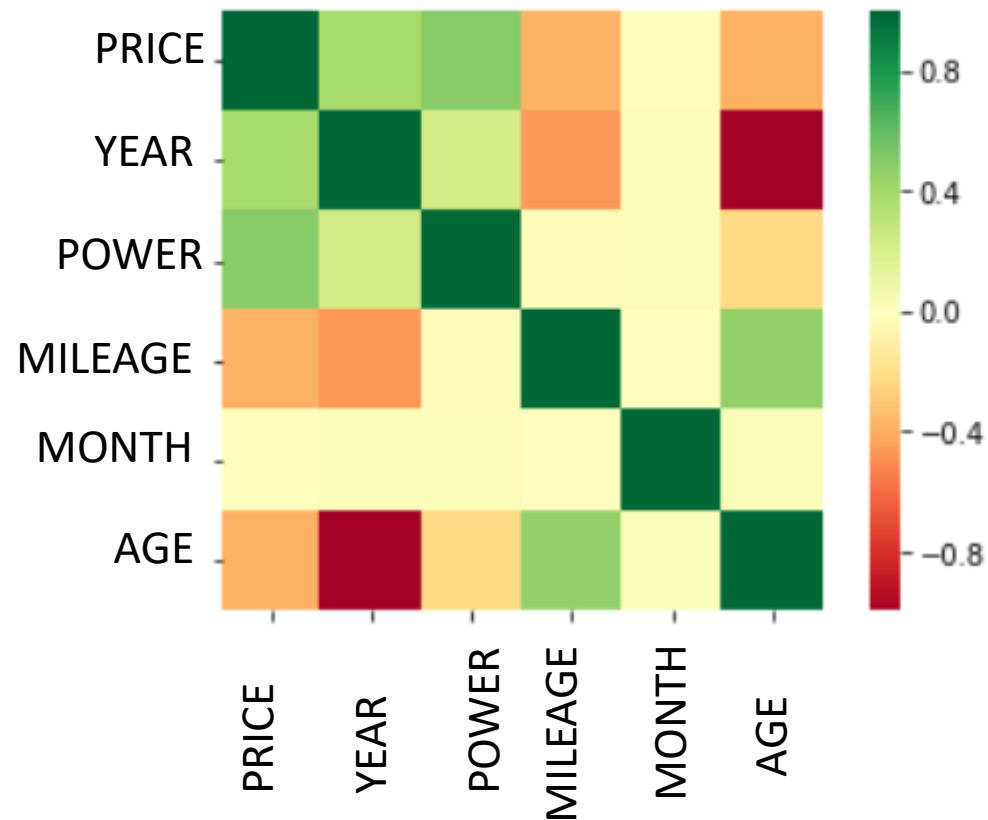
	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration
count	3.715280e+05	371528.000000	371528.000000	371528.000000	371528.000000
mean	1.729514e+04	2004.577997	115.549477	125618.688228	5.734445
std	3.587954e+06	92.866598	192.139578	40112.337051	3.712412
min	0.000000e+00	1000.000000	0.000000	5000.000000	0.000000
25%	1.150000e+03	1999.000000	70.000000	125000.000000	3.000000
50%	2.950000e+03	2003.000000	105.000000	150000.000000	6.000000
75%	7.200000e+03	2008.000000	150.000000	150000.000000	9.000000
max	2.147484e+09	9999.000000	20000.000000	150000.000000	12.000000

38%
Discarded

	dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage
105193	2016-03-30 14:49:32	One too shui fooooo	privat	Angebot	323223	control	coupe	2010	manuell	0	90	5000	4	hybrid	audi	NaN

Exploratory Data Analysis

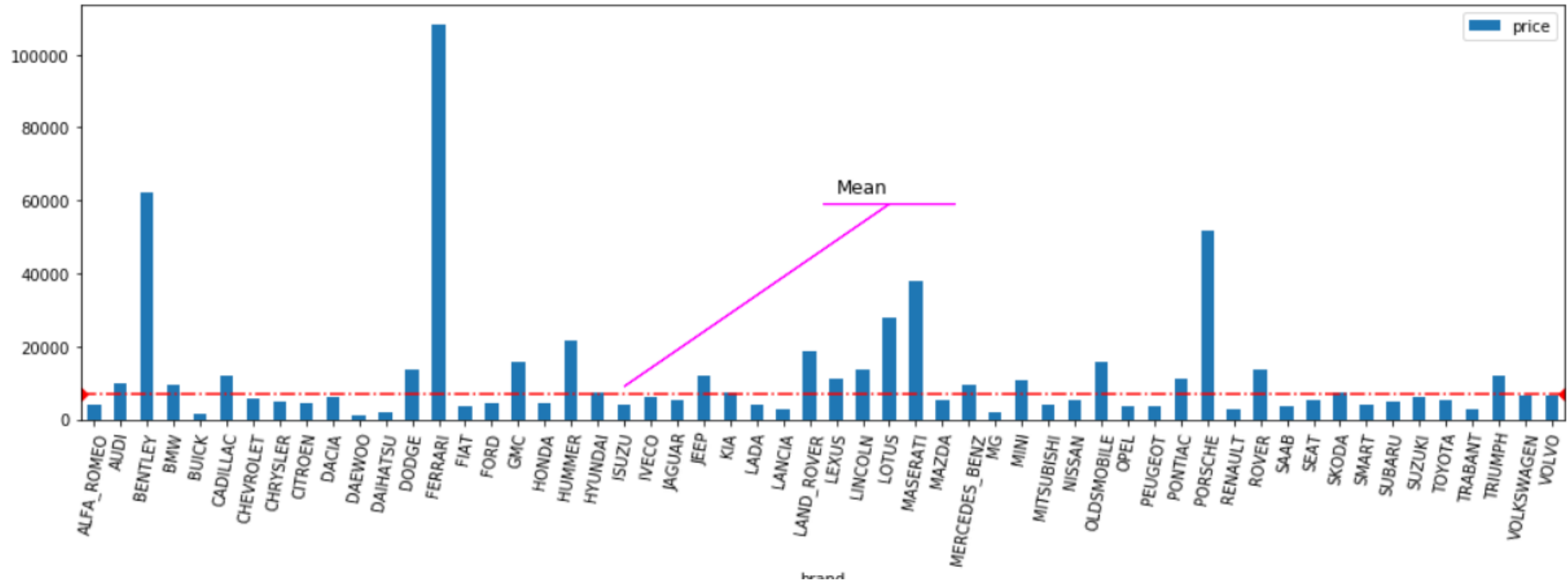
Correlation matrix heatmap reveals major continuous numeric parameters for price:



- Year
- Power
- Mileage
- Age

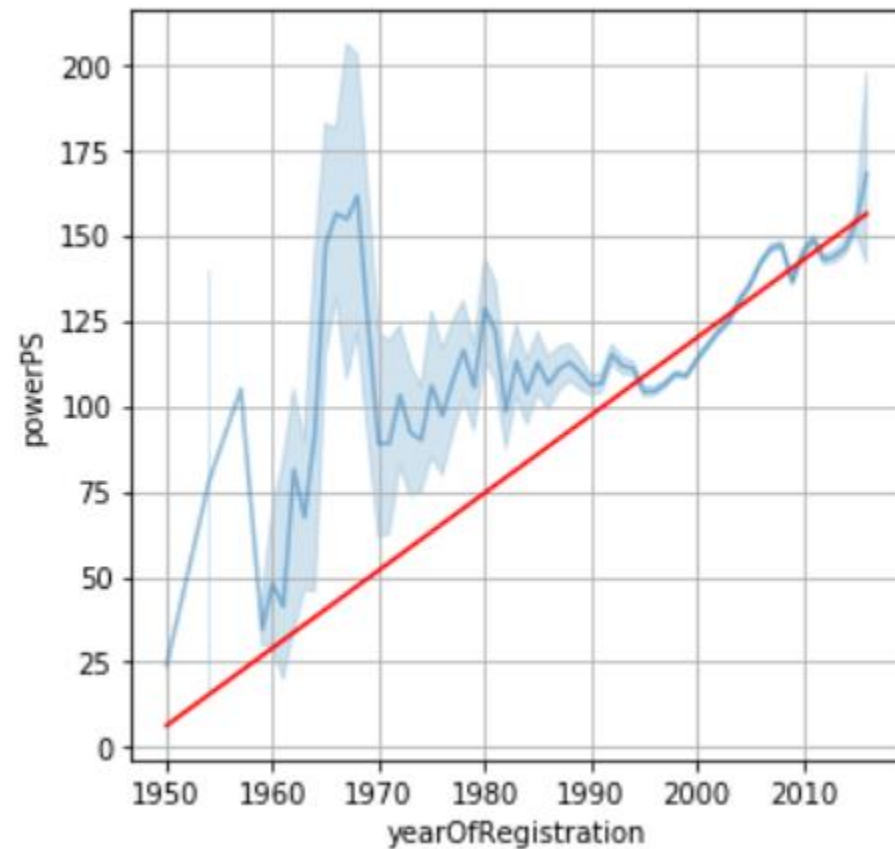
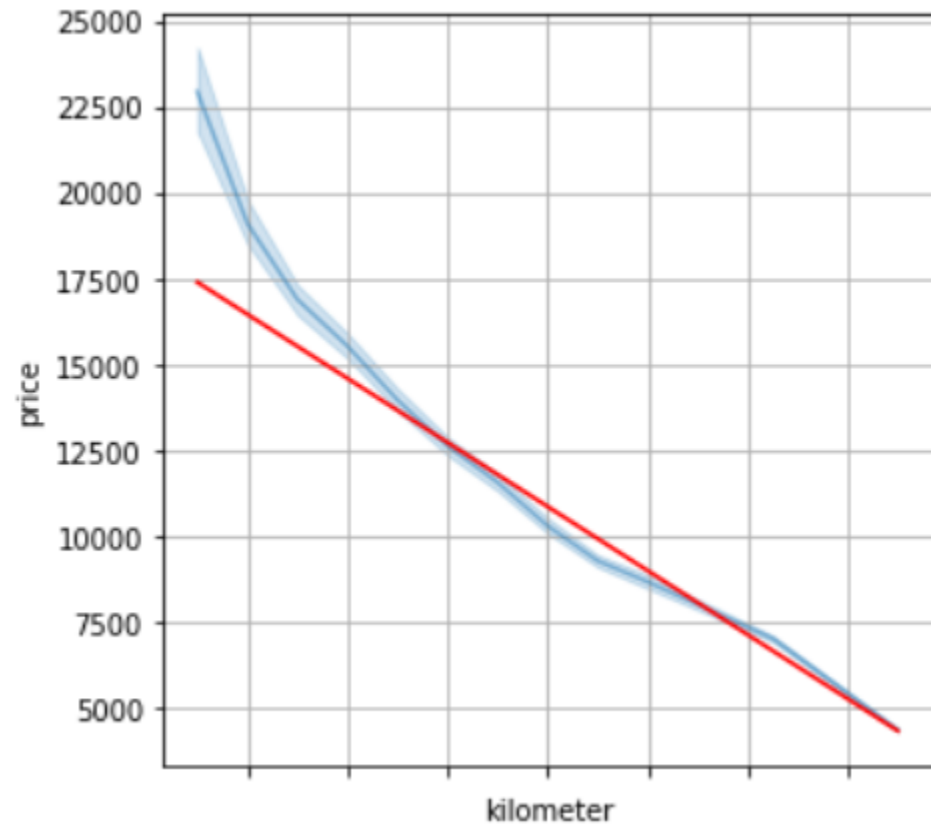
Exploratory Data Analysis

Compared to the database mean, the top three brands are clearly seen to be dwarfing others.



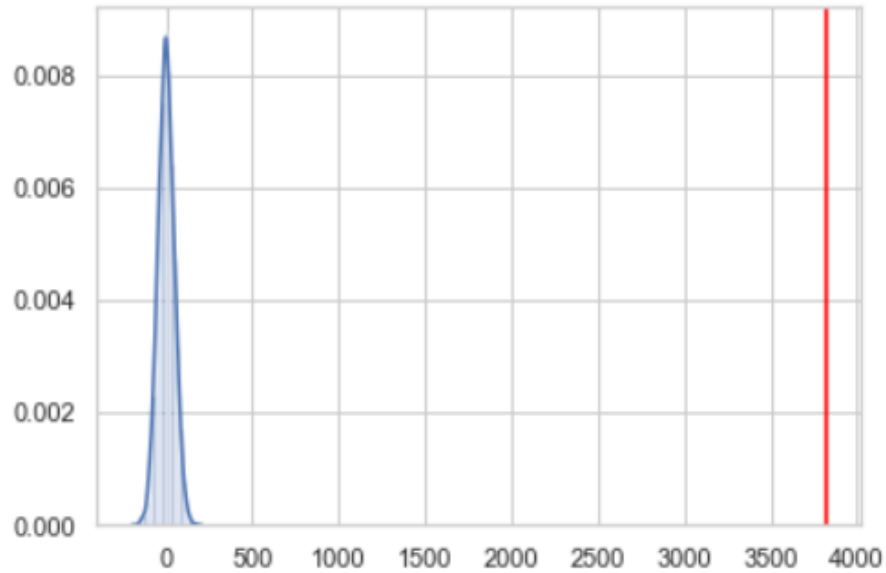
Exploratory Data Analysis

Checking integrity of the data...



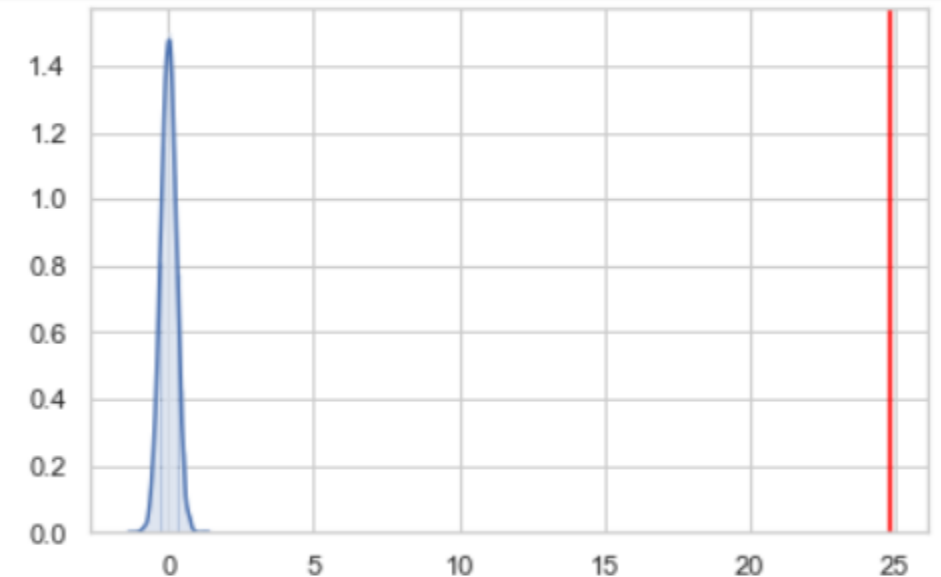
Inferential Statistics

Diesel is much more expensive and powerful. Is this coincidence?



PRICE	T Statistic	P Value	Result
DIESEL vs GAS	83.050	0.00000000	Reject H ₀
MANUAL VS AUTO	120.509	0.00000000	Reject H ₀
SALVAGED VS CLEAN	-64.047	0.00000000	Reject H ₀

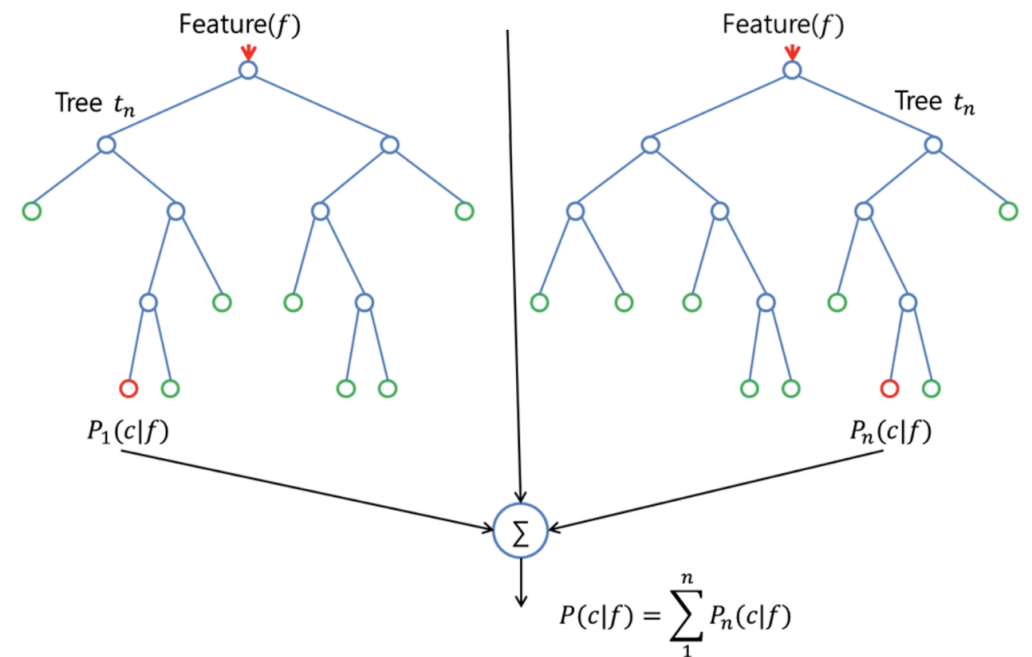
POWER PS	T Statistic	P Value	Result
DIESEL vs GAS	95.365	0.00000000	Reject H ₀
MANUAL VS AUTO	249.735	0.00000000	Reject H ₀
VW vs SKODA*	-0.385	0.70049115	Accept H ₀



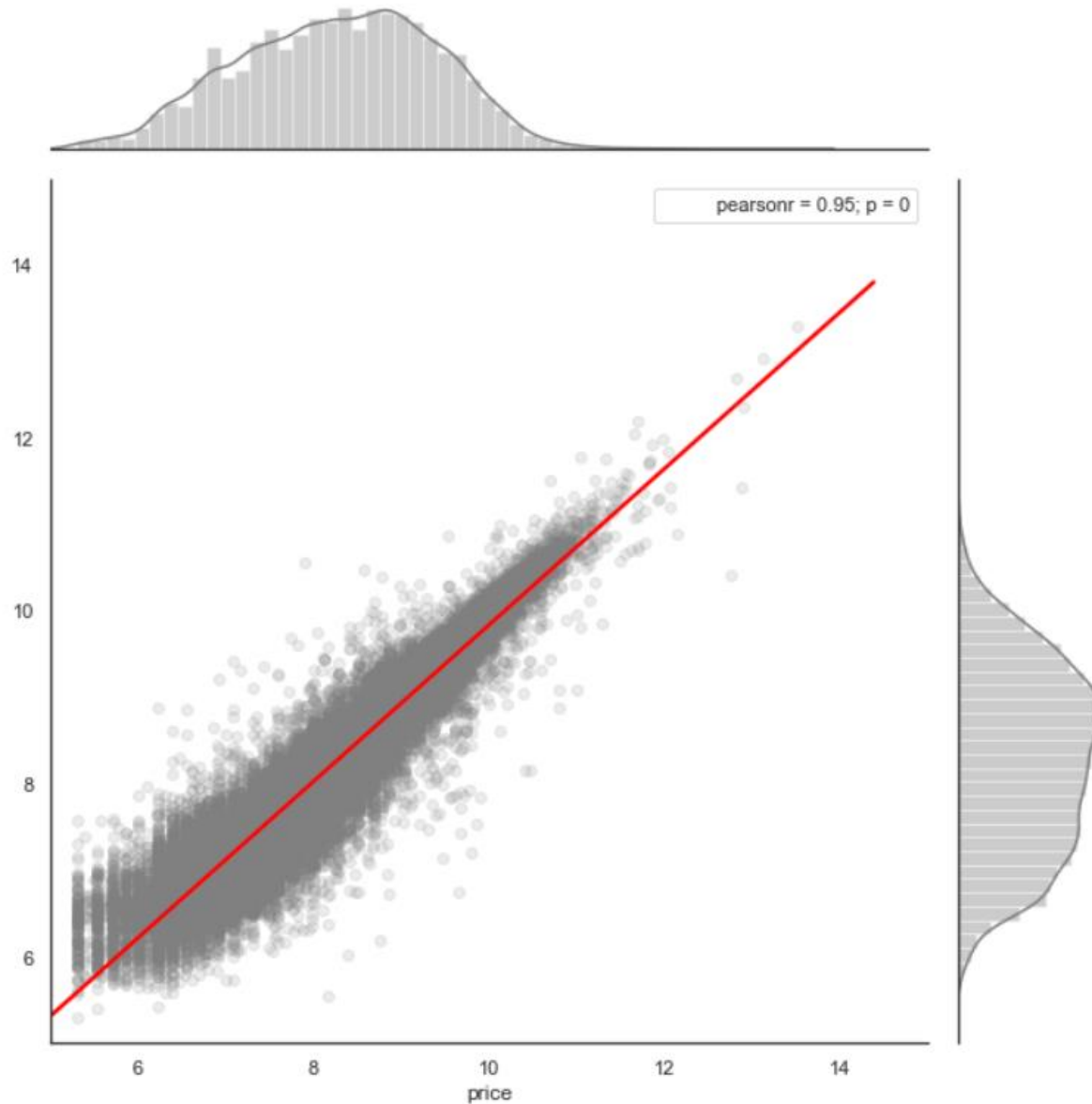
Machine Learning



Non-linear Decision Tree based regression models were used.



Machine Learning



Correlation between predicted prices and actual prices was 0.95.

Prediction Error



Error = 1.40\$