USED CARS DATASET..

By : Soaad Aljafr.
Reem Alghamdi.

• Abstract:

- The goal of the project is to predict the prices of cars that will be displayed in the future, The prediction will be made from the data already in the database https://www.kaggle.com/austinreese/craigslist-carstrucks-data By training the machine with the previous data.

Interdiction:

- Craigslist is an American classified advertisements website with sections devoted to jobs, housing, for sale, items wanted, services, community service, gigs, résumés, and discussion forums.
- Craigslist is the world's largest collection of used vehicles for sale, dataset which includes every used vehicle entry within the United States on Craigslist. it contains most all relevant information that provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories.

Design:

- Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

Data:

- The dataset includes 26 columns and more than 400,000 rows, The most important of them are price, color, year , odometer and condition.
- It differs between numeric and category, This information helped an adequate understanding of the database.
- Almost is category so we should convert to numeric.

Num. null

	0
region	0
price	0
year	1205
manufacturer	17646
model	5277
condition	174104
cylinders	177678
fuel	3013
odometer	4400
title_status	8242
transmission	2556
drive	130567
size	306361
type	92858
paint_color	130203
county	426880
state	0
posting_date	68

Object describe

	COUNT	UNIQUE	TOP	FREQ
region	426880	404	columbus	3608
manufacturer	409234	42	ford	70985
model	421603	29667	f-150	8009
condition	252776	6	good	121456
cylinders	249202	8	6 cylinders	94169
fuel	423867	5	gas	356209
title_status	418638	6	clean	405117
transmission	424324	3	automatic	336524
drive	296313	3	4wd	131904
type	334022	13	sedan	87056
paint_color	296677	12	white	79285
state	426880	51	ca	50614
posting_date	426812	381536	2021-04-23T22:13:05-0400	12

Algorithms:

- Cleaned the data for many steps
- First remove the null from year and fill null for mean in numeric columns and mode for categories.
- Then remove duplicate.
- Cheek if there any null or duplicate.

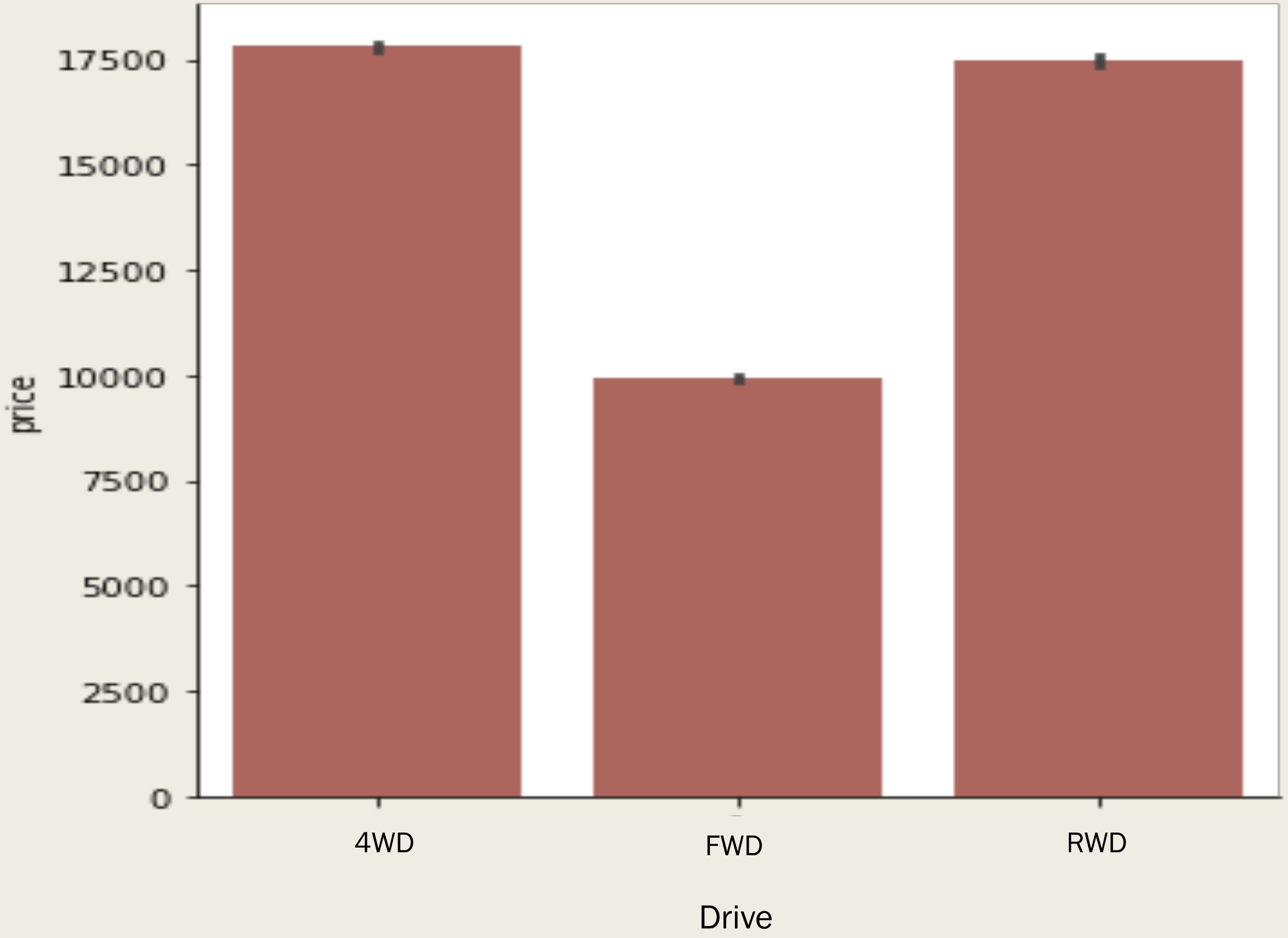
Tools:

- The dataset includes 26 columns and more than 400,000 rows, The most important of them are price, color, year , odometer and condition.
- It differs between numeric and category, This information helped an adequate understanding of the database.
- Almost is category so we should convert to numeric.

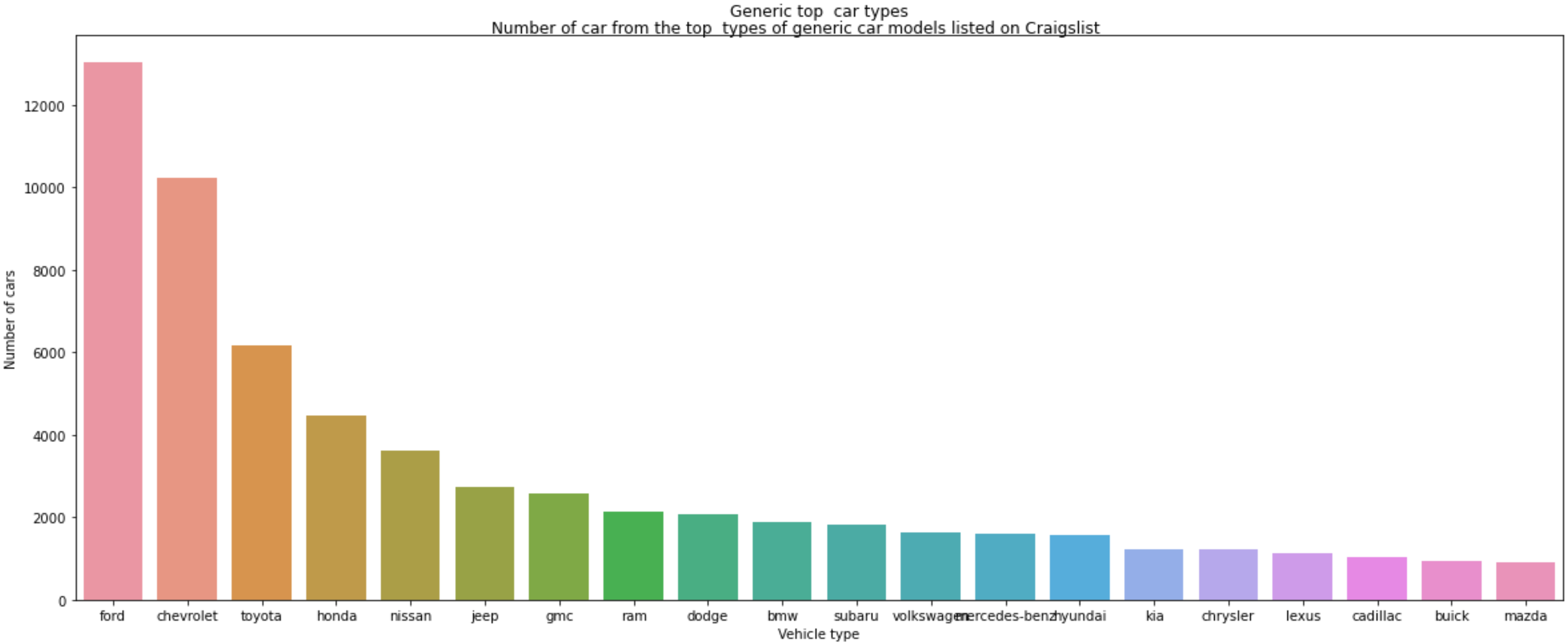
Data Cleaning

- Read The Data from cvs
- Drop some columes
- Delet the Null values
- Fill Null values
- Change the object columes to numaric

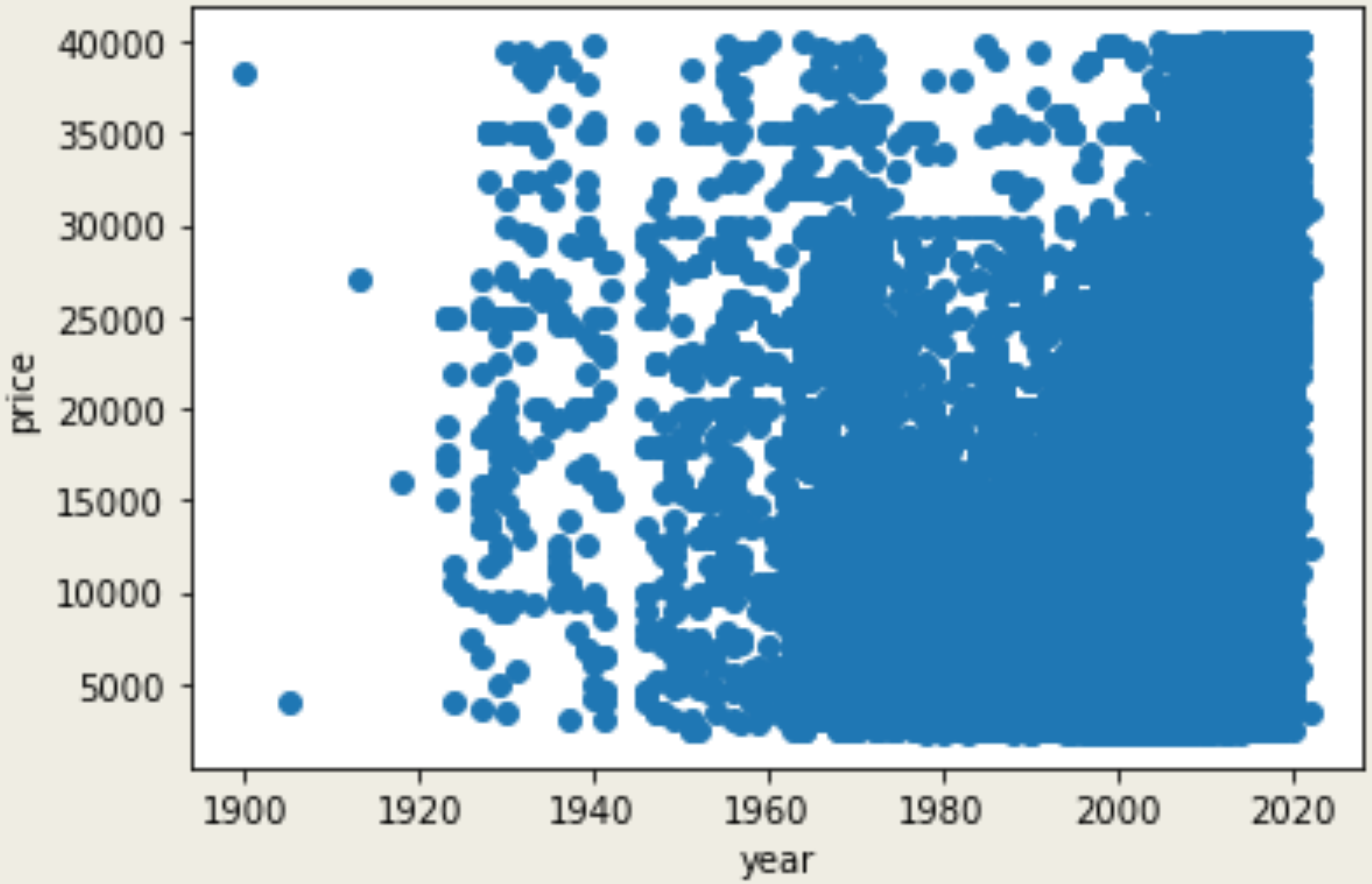
Bar plot between type and price



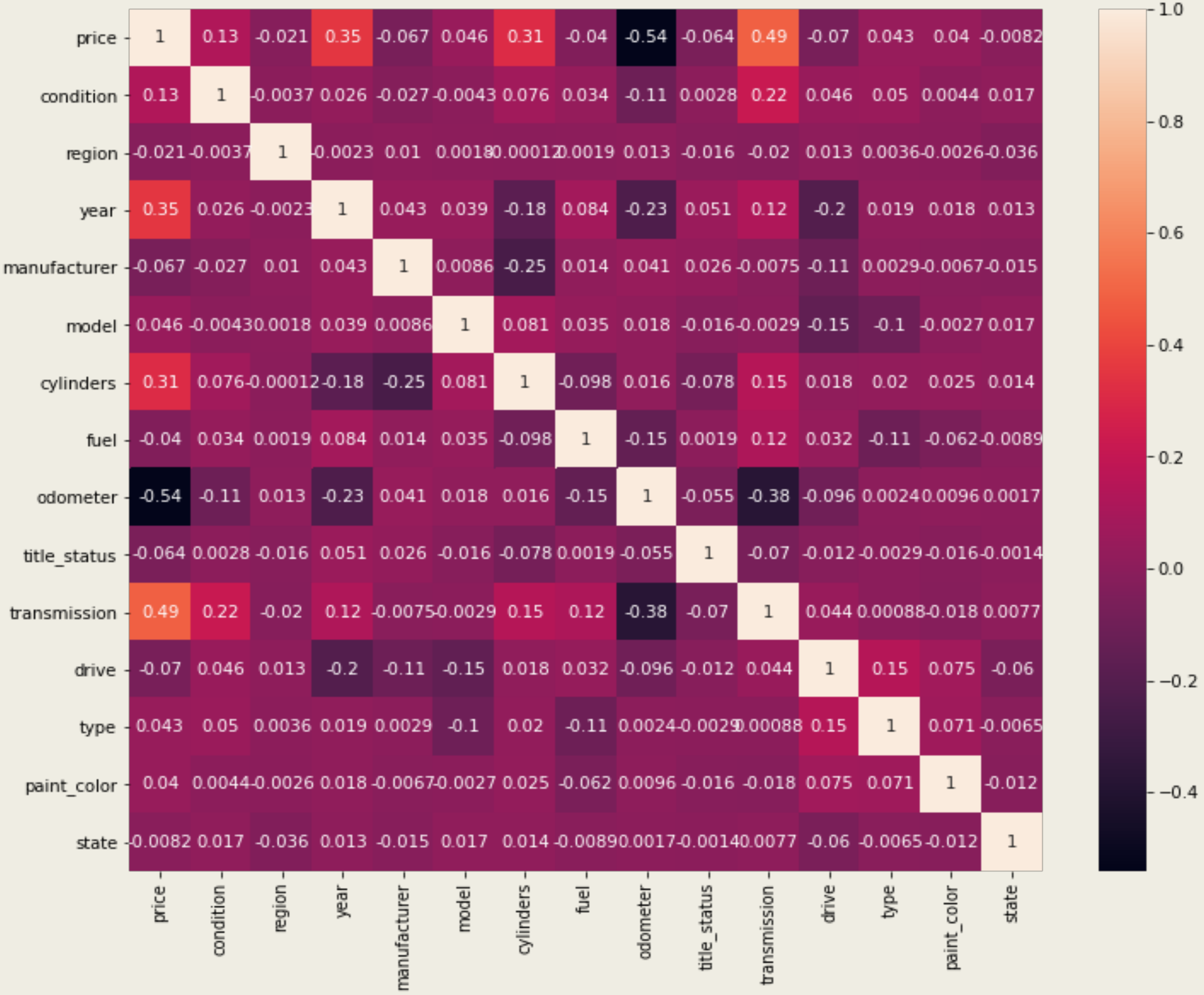
Manufacturer



Scatler Years and Price



Data Correlation



	price	condition	region	year	manufacturer	model	cylinders	fuel	odometer	title_status	transmission	drive	type	paint_color	state
price	1	0.13	-0.021	0.35	-0.067	0.046	0.31	-0.04	-0.54	-0.064	0.49	-0.07	0.043	0.04	-0.0082
condition	0.13	1	-0.0037	0.026	-0.027	-0.0043	0.076	0.034	-0.11	0.0028	0.22	0.046	0.05	0.0044	0.017
region	-0.021	-0.0037	1	-0.0023	0.01	0.0018	-0.00012	0.0019	0.013	-0.016	-0.02	0.013	0.0036	-0.0026	-0.036
year	0.35	0.026	-0.0023	1	0.043	0.039	-0.18	0.084	-0.23	0.051	0.12	-0.2	0.019	0.018	0.013
manufacturer	-0.067	-0.027	0.01	0.043	1	0.0086	-0.25	0.014	0.041	0.026	-0.0075	-0.11	0.0029	-0.0067	-0.015
model	0.046	-0.0043	0.0018	0.039	0.0086	1	0.081	0.035	0.018	-0.016	-0.0029	-0.15	-0.1	-0.0027	0.017
cylinders	0.31	0.076	-0.00012	-0.18	-0.25	0.081	1	-0.098	0.016	-0.078	0.15	0.018	0.02	0.025	0.014
fuel	-0.04	0.034	0.0019	0.084	0.014	0.035	-0.098	1	-0.15	0.0019	0.12	0.032	-0.11	-0.062	-0.0089
odometer	-0.54	-0.11	0.013	-0.23	0.041	0.018	0.016	-0.15	1	-0.055	-0.38	-0.096	0.0024	0.0096	0.0017
title_status	-0.064	0.0028	-0.016	0.051	0.026	-0.016	-0.078	0.0019	-0.055	1	-0.07	-0.012	-0.0029	-0.016	-0.0014
transmission	0.49	0.22	-0.02	0.12	-0.0075	-0.0029	0.15	0.12	-0.38	-0.07	1	0.044	0.00088	-0.018	0.0077
drive	-0.07	0.046	0.013	-0.2	-0.11	-0.15	0.018	0.032	-0.096	-0.012	0.044	1	0.15	0.075	-0.06
type	0.043	0.05	0.0036	0.019	0.0029	-0.1	0.02	-0.11	0.0024	-0.0029	0.00088	0.15	1	0.071	-0.0065
paint_color	0.04	0.0044	-0.0026	0.018	-0.0067	-0.0027	0.025	-0.062	0.0096	-0.016	-0.018	0.075	0.071	1	-0.012
state	-0.0082	0.017	-0.036	0.013	-0.015	0.017	0.014	-0.0089	0.0017	-0.0014	0.0077	-0.06	-0.0065	-0.012	1

Model and Result

	Training	Validation	Test
Liner Regression	0.377	0.358	0.339
Polynomial	0.801	0.875	0.845

Thanks..