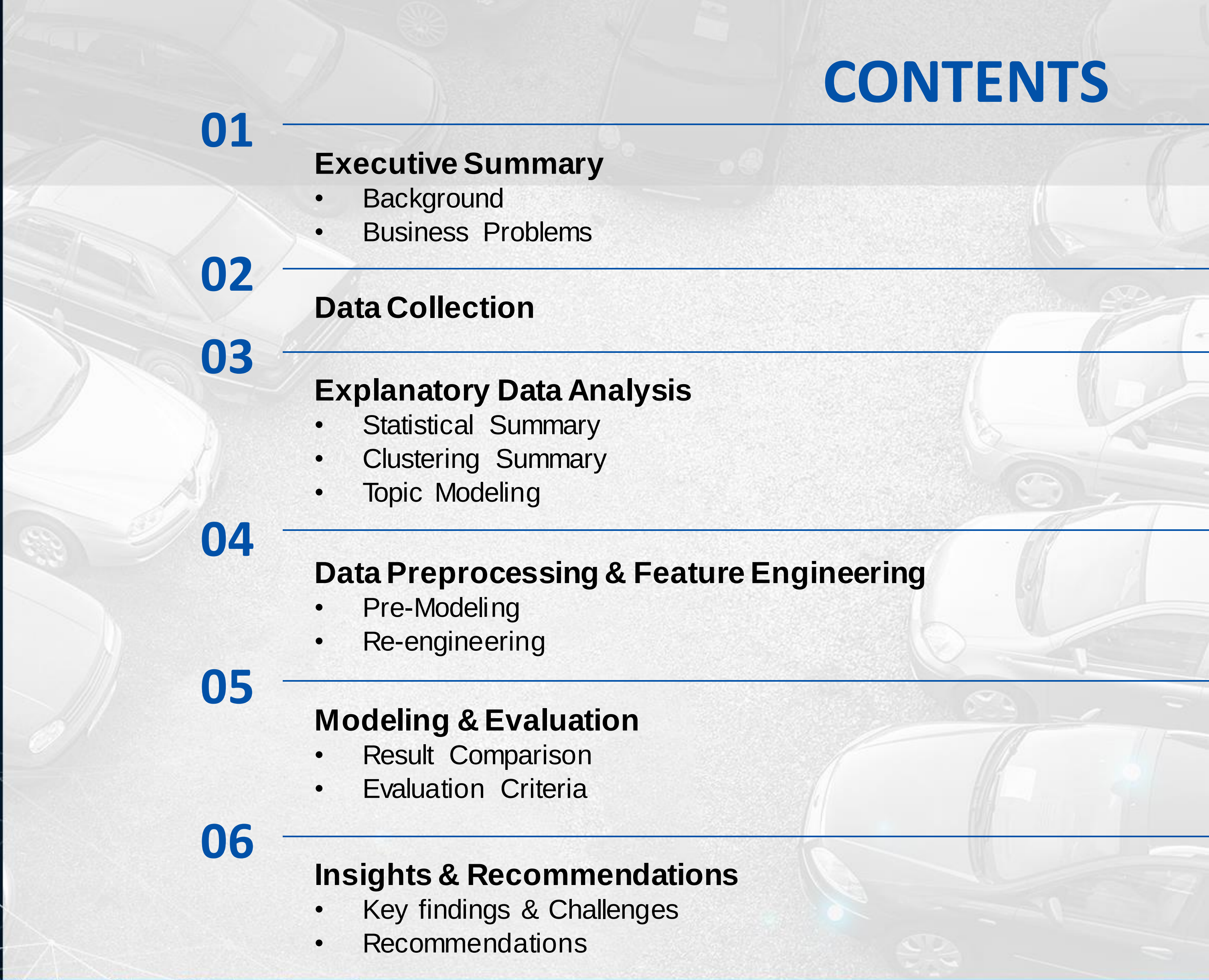
USED CAR MARKET

Team: Ahjeong Yeom, Akhir Syabani, Jessica Addai, Kenji Laurens, Ruofan Yao, Zongyuan Yu



CONTENTS

01

Executive Summary

- Background
- Business Problems

02

Data Collection

03

Explanatory Data Analysis

- Statistical Summary
- Clustering Summary
- Topic Modeling

04

Data Preprocessing & Feature Engineering

- Pre-Modeling
- Re-engineering

05

Modeling & Evaluation

- Result Comparison
- Evaluation Criteria

06

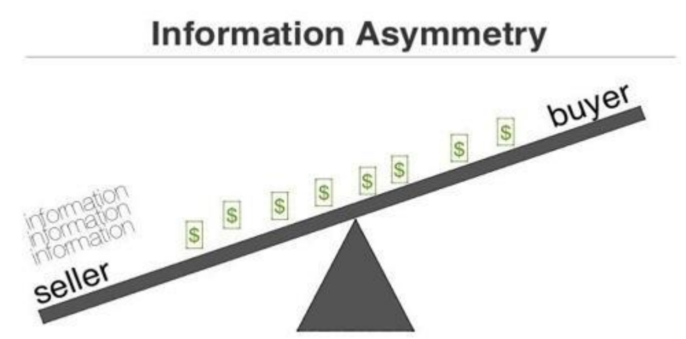
Insights & Recommendations

- Key findings & Challenges
- Recommendations

Executive Summary

Background

- **Asymmetric information**, also known as “information failure,” takes place during a transaction where one party has greater material knowledge or better information than the other party.
- **“The Market of Lemons:** Quality Uncertainty and the Market Mechanism,” by George Akerlof



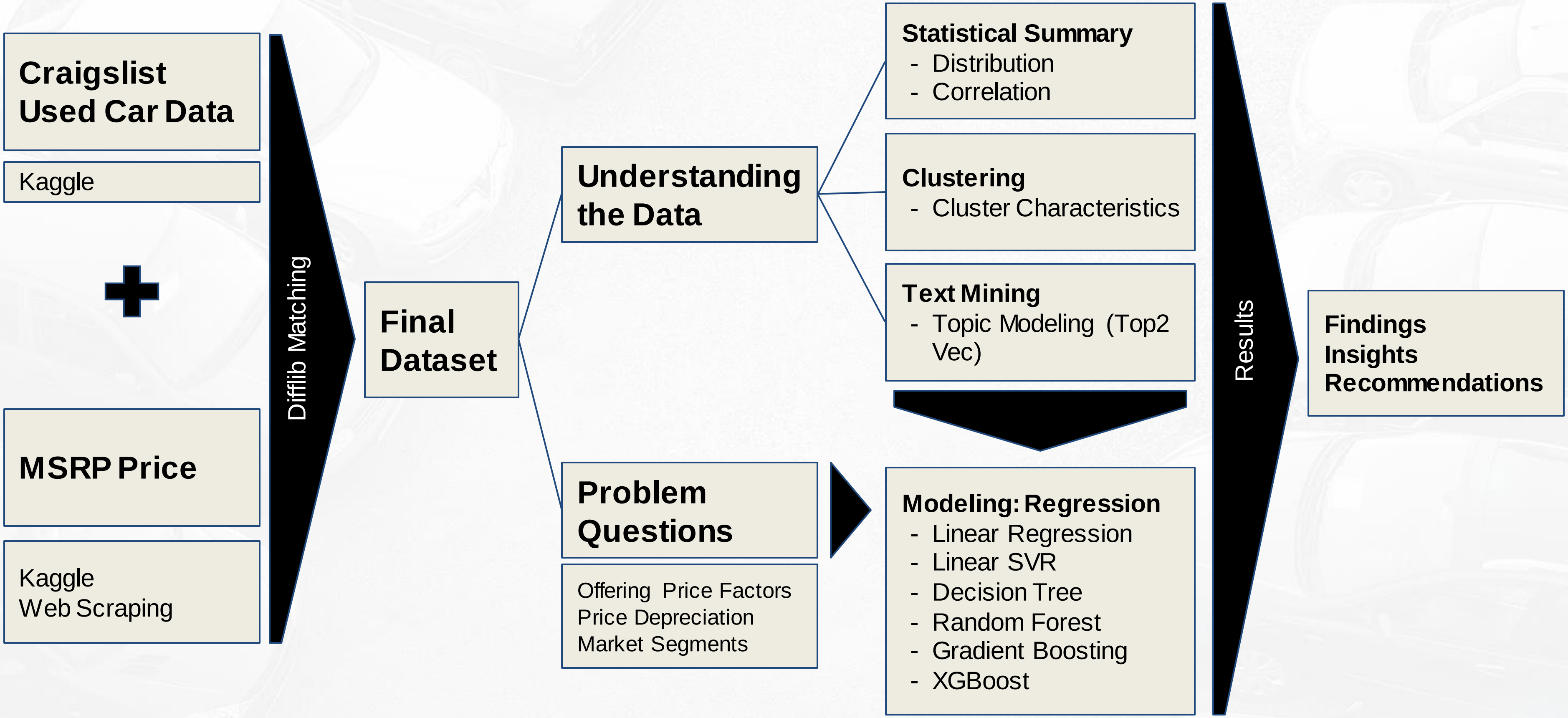
Business Problem

- Consumers face incomplete information on used cars
- Eventual Market Failure

Our Solution

- Help consumers stay informed on what features determine car price
- Help consumers and sellers have baseline for reasonable price points for cars
- Better information leads to better decisions on major purchases like cars

Used Car Price Analysis



Craigslist Used Car Data

Kaggle

MSRP Price

Kaggle
Web Scraping

Difflib Matching

Final Dataset

Understanding the Data

Statistical Summary
- Distribution
- Correlation

Clustering
- Cluster Characteristics

Text Mining
- Topic Modeling (Top2 Vec)

Problem Questions

Offering Price Factors
Price Depreciation
Market Segments

Modeling: Regression
- Linear Regression
- Linear SVR
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost

Results

Findings
Insights
Recommendations

Data Collection
Collection

kaggle craigslist iSeeCars

- Our dataset is taken from Kaggle public dataset (1.45 GB)
- Based on resale car listings on Craigslist
- Columns include selling price, car attributes, color, condition, VIN, mileage, make and model, etc.
- Enhanced and imputed some data with data from iSeeCars

- Problems arose with dirty and bad data
- Since each listing data is 100% based on seller entry, sellers tend to input false data, intentionally or otherwise

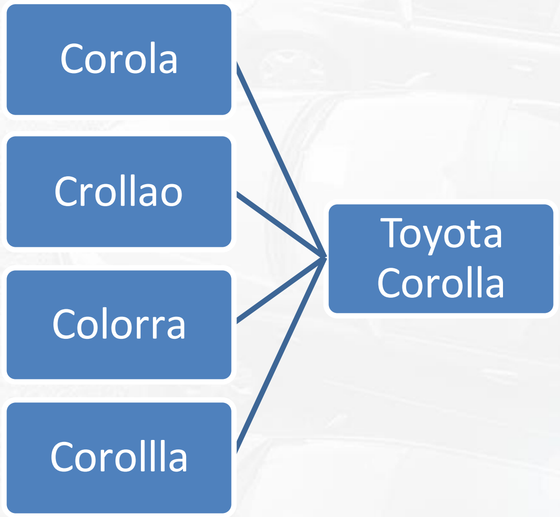
Data Collection

Incorporating External Data

- To impute some of the columns, we looked at the **most common value** for each car model

- **Assumptions:**
  - Cars of the same make and model share common attributes such as cylinders, type and size
  - For customizable attributes (drive type) we are assuming the most common values for each model
- We collected data from iSeeCars through web scraping
- Data scraped include MSRP and car attributes for all makes and models available

- **Problem:**
  - Due to user-input values, car model names may not match exactly with the clean names from iSeeCars
  - To overcome this problem, we used OpenRefine to try and fix some of the entries based on naming clusters
  - We also used **SequenceMatcher** from the **difflib library** to programmatically fix model names based on similarity index and assigned model with the highest name similarity



Data Collection
Available Columns

Data Types

Irrelevant

High Cardinality:
- ID
- URL
- Image_URL
- VIN

Null:
- County (100%)
- Size (>70%)

Text:
- Description

Categorical:
- Region
- Region_URL

Geolocation:
- Longitude
- Latitude

Relevant

Categorical:
- *Manufacturer*
- *Model*
- Condition
- Cylinders
- Fuel
- Title Status
- Transmission
- Drive
- Type
- Paint Color
- State

Continuous:
- Price
- MSRP
- Odometer

Time:
- Year
- *Posting Date*

Data Collection
Feature Selection and Imputation

Dropped columns:

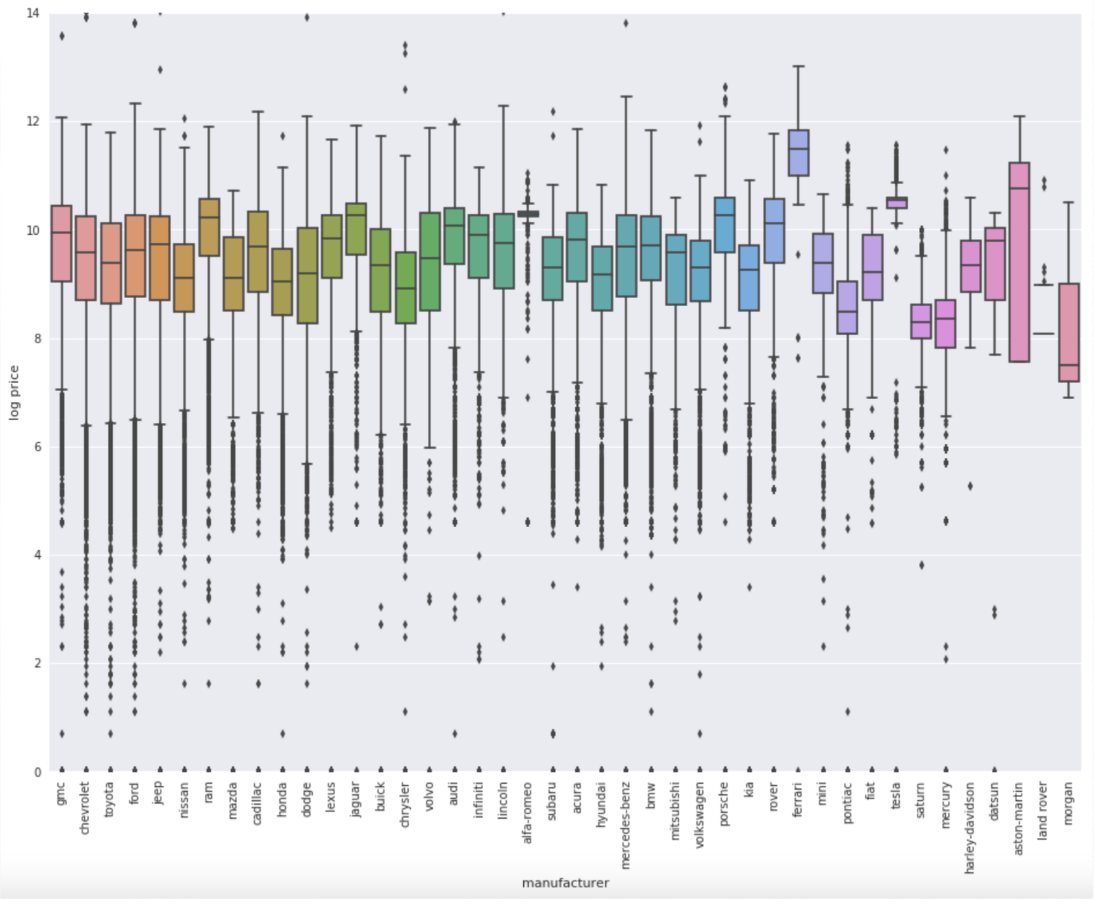
ID	VIN	URL	Latitude	Longitude	Image URL
Region	Region URL	County	Size	Description	

Imputed Columns:

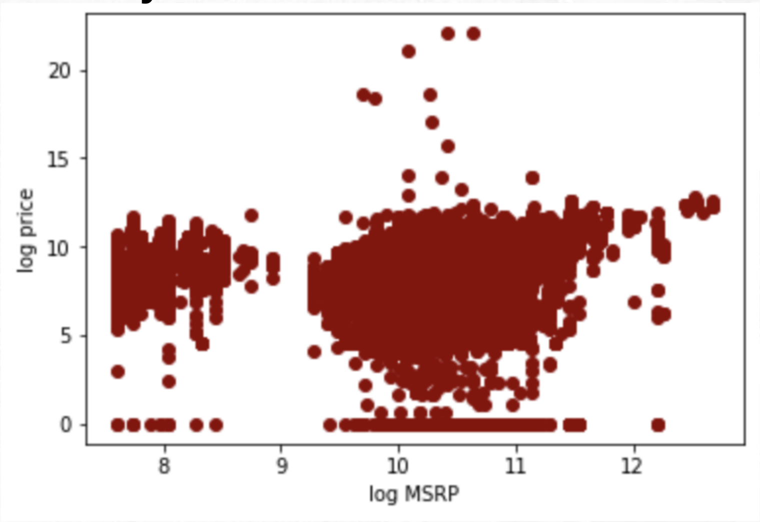
Method	Column / Features
Mode	Title Status Fuel Color
Based on other columns	Condition
Based on external dataset	Manufacturer Model Drive Cylinder Type

Exploratory Data Analysis (EDA)
Raw & Filtered Data

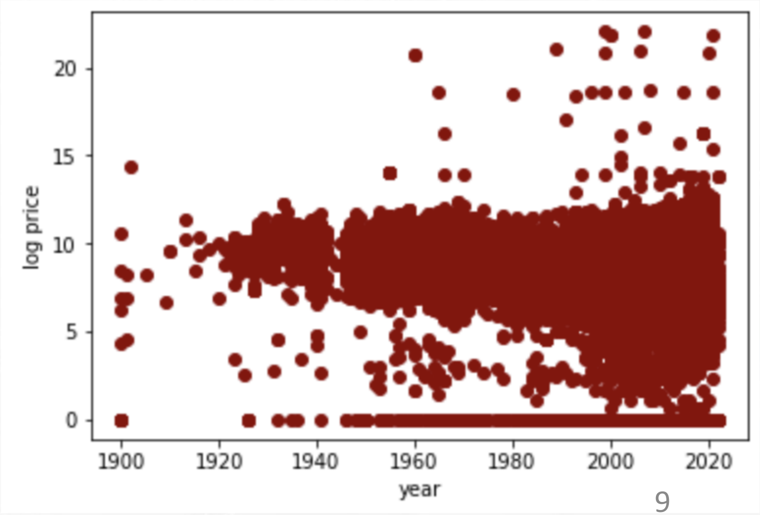
Price by Manufacturer



Price by MSRP



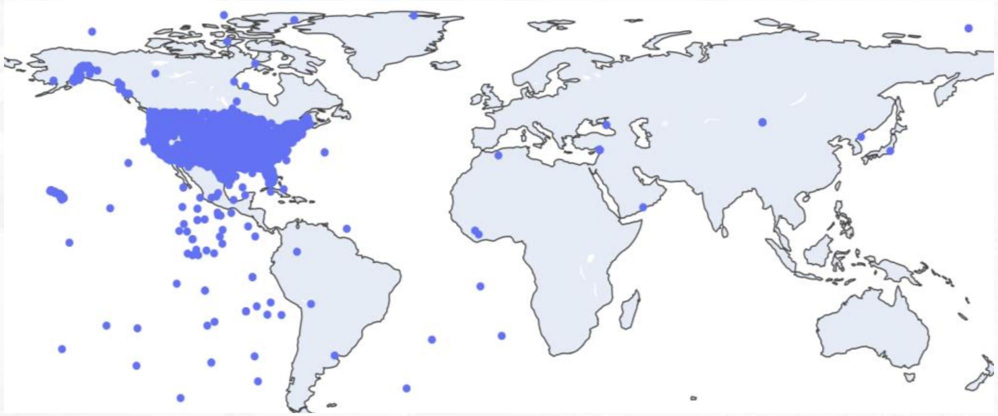
Price Production Year



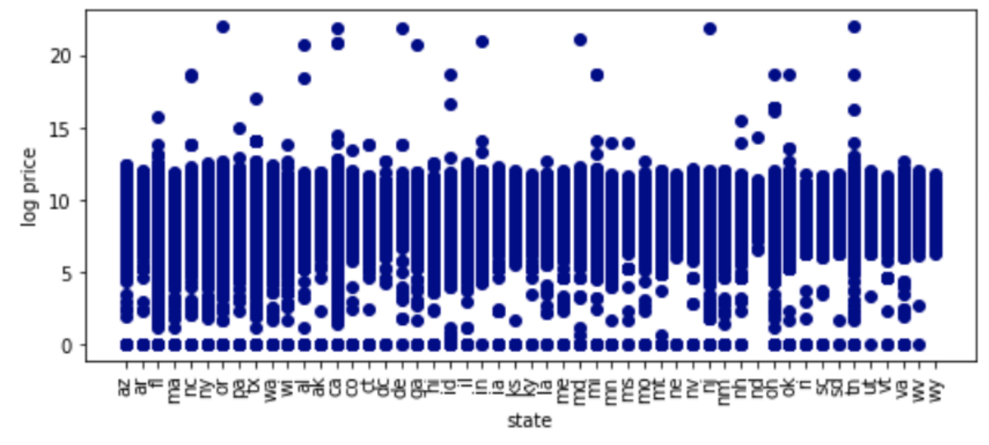
Exploratory Data Analysis (EDA)

Raw Data by Region

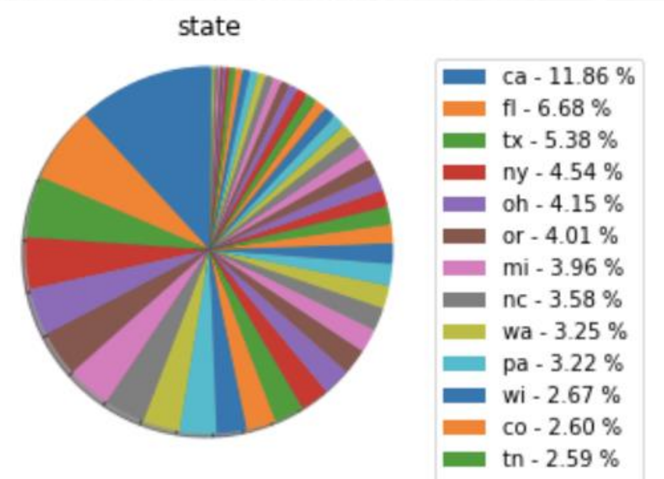
Geo Location: Longitude, Latitude



Price by State



Distribution



Exploratory Data Analysis (EDA)
Highlight: Odometer

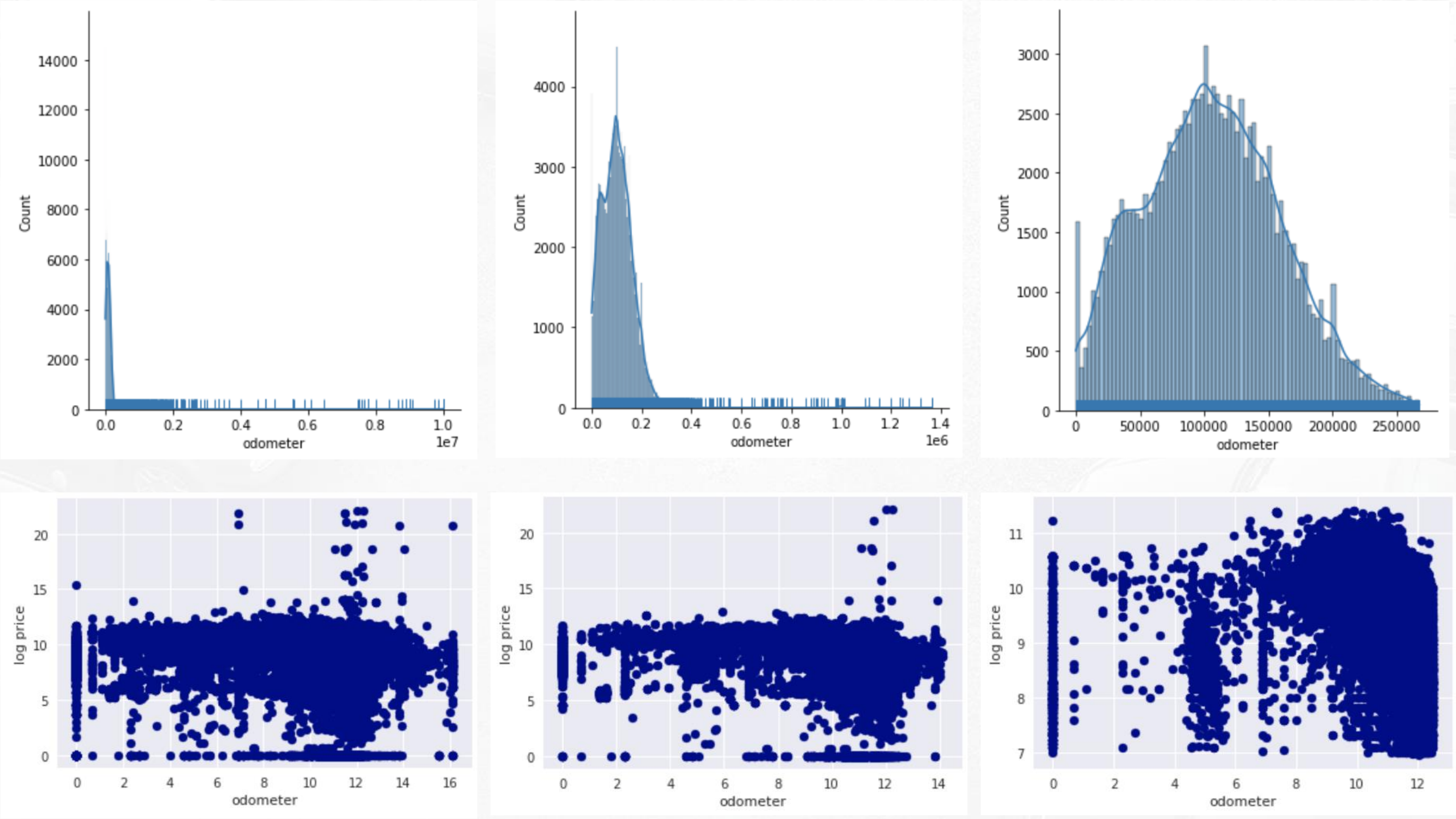
Raw Data

Filtered Data (by MSRP)

Clean/Final Dataset

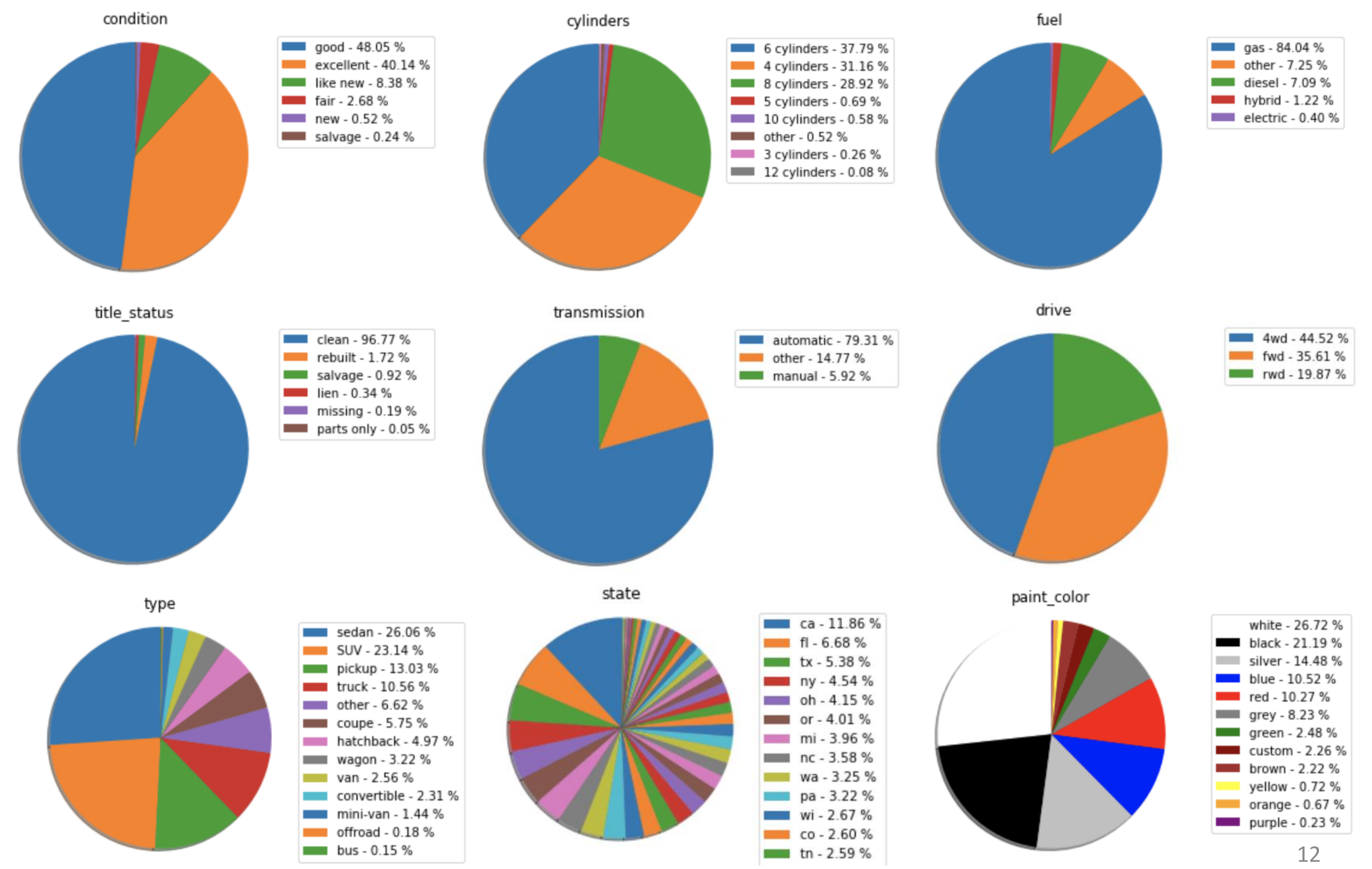
Distribution

Against Log Price



Exploratory Data Analysis (EDA)

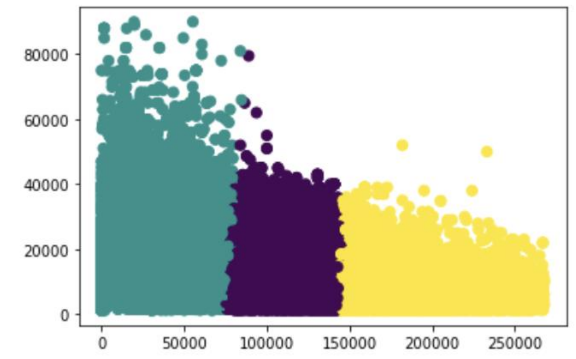
Raw Data: Other Categorical Variables



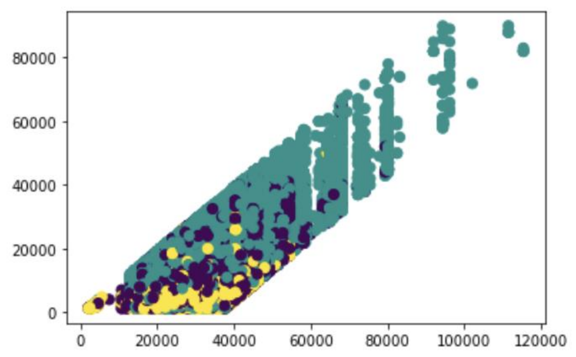
Exploratory Data Analysis (EDA)

Clustering Summary (Clean Dataset, K-Means, 3 Clusters)

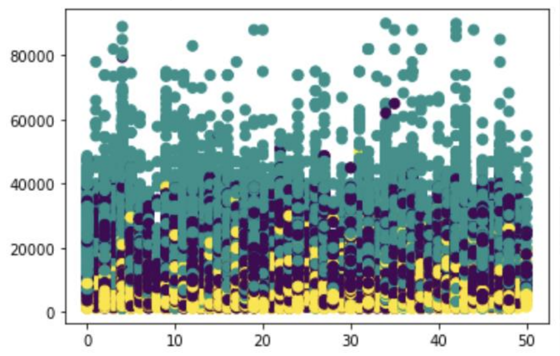
X = Odometer, y = Price



X = MSRP, y = Price



X = State, y = Price

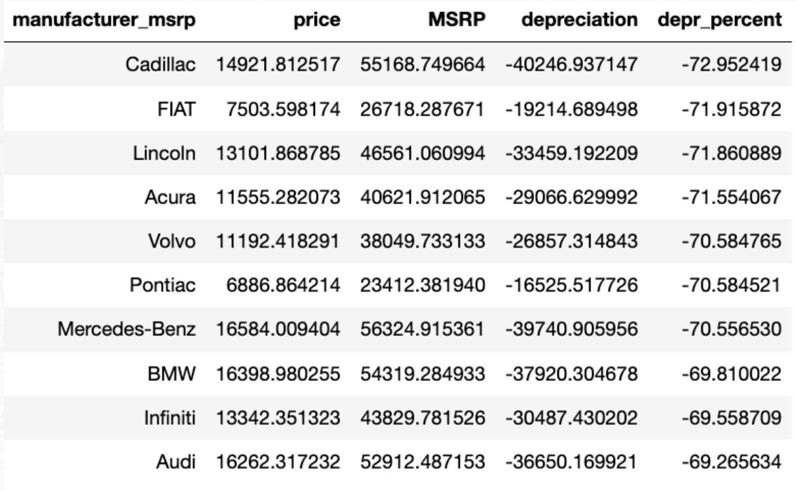


Key Characteristics

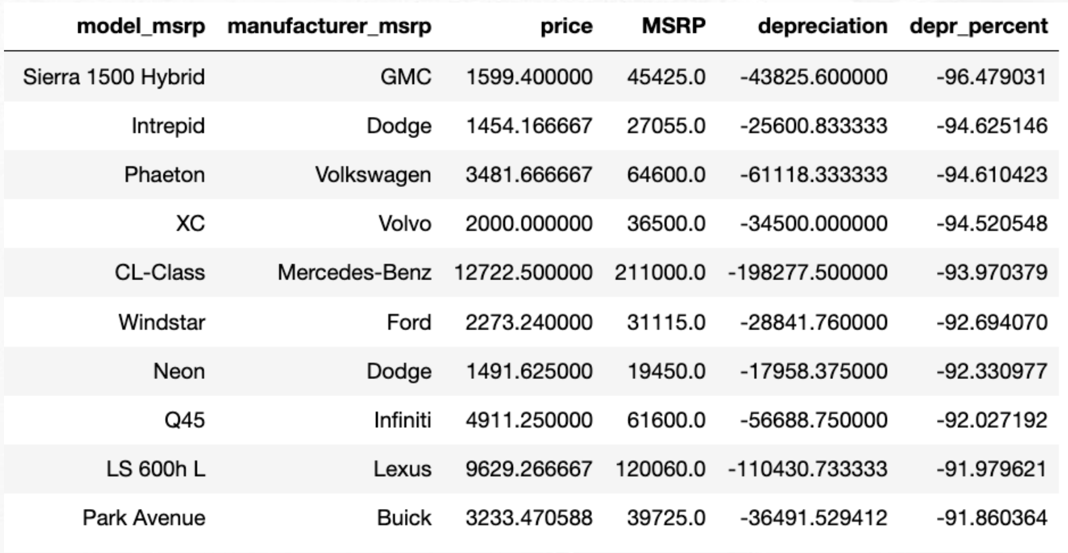
Cluster	Price	Odometer	MSRP	Car Age	Vintage	Color
Cluster0 "Lower"	Min: 1k Max: 52k Mean: 7k	Min: 143k Max: 268k Mean: 177k	Max: 65k Mean: 30k	Mean: 14	0.04%	46% neutral
Cluster1 "Mid"	Min: 1k Max: 80k Mean: 12k	Min: 75k Max: 125k Mean: 110k	Max: 96k Mean: 31k	Mean: 10	0.30%	51% neutral
Cluster2 "Upper"	Min: 1k Max: 90k Mean: 21k	Min: 0 Max: 85k Mean: 45k	Max: 102k Mean: 33k	Mean: 6	0.82%	53% neutral

Exploratory Data Analysis (EDA)
Price Depreciation

Avg MSRP	**$ 31,328**
Avg Price	**$ 13,670**
Depreciation	**$ -17,658**
Depreciation %	**-56%**



manufacturer_msrp	price	MSRP	depreciation	depr_percent
Cadillac	14921.812517	55168.749664	-40246.937147	-72.952419
FIAT	7503.598174	26718.287671	-19214.689498	-71.915872
Lincoln	13101.868785	46561.060994	-33459.192209	-71.860889
Acura	11555.282073	40621.912065	-29066.629992	-71.554067
Volvo	11192.418291	38049.733133	-26857.314843	-70.584765
Pontiac	6886.864214	23412.381940	-16525.517726	-70.584521
Mercedes-Benz	16584.009404	56324.915361	-39740.905956	-70.556530
BMW	16398.980255	54319.284933	-37920.304678	-69.810022
Infiniti	13342.351323	43829.781526	-30487.430202	-69.558709
Audi	16262.317232	52912.487153	-36650.169921	-69.265634



model_msrp	manufacturer_msrp	price	MSRP	depreciation	depr_percent
Sierra 1500 Hybrid	GMC	1599.400000	45425.0	-43825.600000	-96.479031
Intrepid	Dodge	1454.166667	27055.0	-25600.833333	-94.625146
Phaeton	Volkswagen	3481.666667	64600.0	-61118.333333	-94.610423
XC	Volvo	2000.000000	36500.0	-34500.000000	-94.520548
CL-Class	Mercedes-Benz	12722.500000	211000.0	-198277.500000	-93.970379
Windstar	Ford	2273.240000	31115.0	-28841.760000	-92.694070
Neon	Dodge	1491.625000	19450.0	-17958.375000	-92.330977
Q45	Infiniti	4911.250000	61600.0	-56688.750000	-92.027192
LS 600h L	Lexus	9629.266667	120060.0	-110430.733333	-91.979621
Park Avenue	Buick	3233.470588	39725.0	-36491.529412	-91.860364

Exemplifies the commonly held notion that Asian car values depreciate less

Exploratory Data Analysis (EDA)

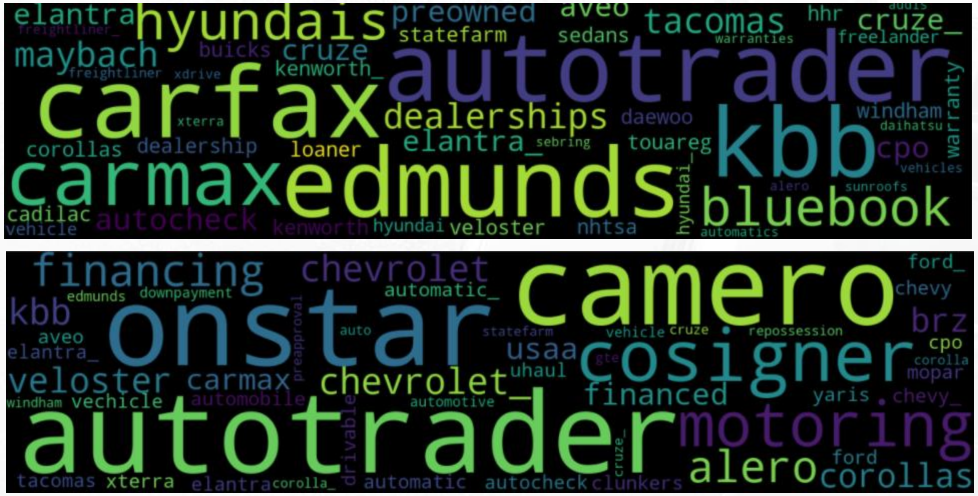
Topic Modeling With Top2Vec

- For the "Description" column of our data, we use **Top2Vec** algorithm to see whether we can extract useful information
- Top2Vec -- An algorithm that can perform topic modeling on text. It returns the number of the topics it finds from the text, and the keywords from each topic. It can also generate word cloud to help us visualize the keywords in certain topic.

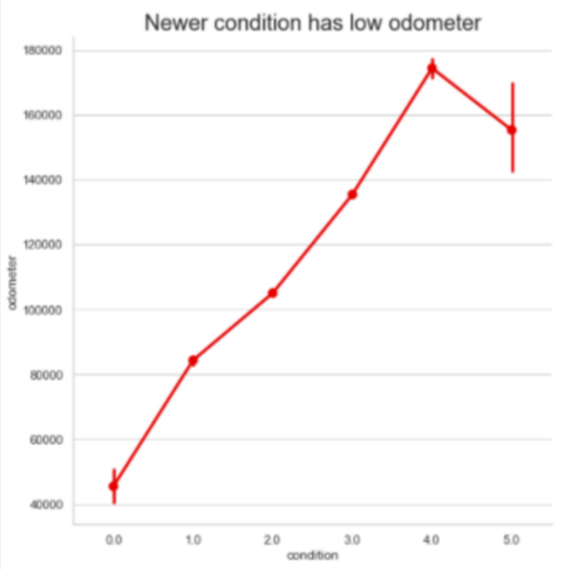
From our result, we get more than 1000 topics back, and the keywords in each topic are not similar with each other. Thus, ***the <u>"Description" column is not informational and should not be included in our model</u>***

However, one insight we have is that ***a lot of used car company advertise their car on craigslist*** since a lot of the keywords revolve the names of some used car companies, like Carfax, Carmax, Autotrader, etc.

Fig 1. Two examples of the generated keyword clouds



Data Preprocessing & Feature Engineering
Pre-Modeling Continued – Imputation Continued


Newer condition has low odometer

Initial Outlier handling & filling in null values

- **Odometer:**
  - Highly right skewed
  - Outliers are removed: Upper threshold +7* stdev
- **Condition:** fill nulls based on odometer
  - Quantile <25% = Excellent
  - Quantile 25%-50% = Good
  - Quantile >50% = Fair
  - NA = Fair
- **Transmission:**
  - Automatic, manual, other -> not possible to assume car transmission type based on other features
  - drop NA rows (~0.6%)
- **Categorical variables:** we filled in NA with mode
  - For cylinders, drive, type, fuel
    - Fill NA with the most common types based on matched model and manufacturer
    - Fill the rest of NA with mode

Data Preprocessing & Feature Engineering

Pre-Modeling - Feature Creation

Feature Variables

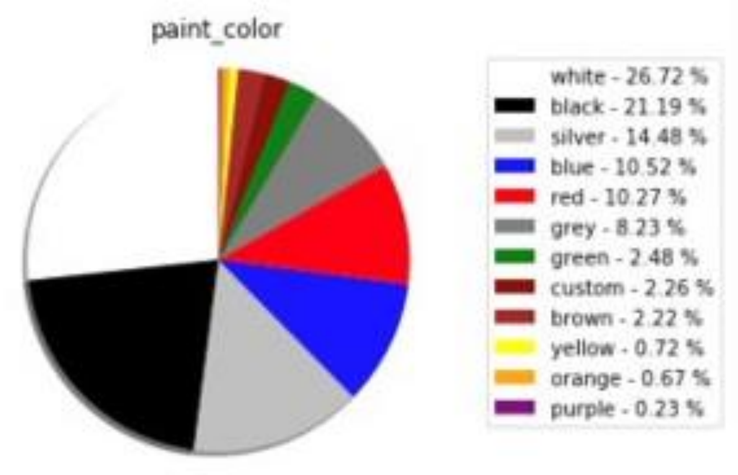
Continuous:
- price : *target variable*
- MSRP
- odometer

Time:
- year
- posting date

Categorical:
- condition
- cylinders
- fuel
- title_status
- transmission
- drive
- type
- state

New:
- car_age
- is_vintage
- is_color_neutral

Created Features:
- **Color:** 12 colors into binary column of "is_neutral"
  - is_neutral (1)**:** Black, White, Silver, Grey
  - Is_neutral (0): Colorful



- **car_age :** year of posting_date subtracted by year when car came out
- **is_vintage:** car_age >50
  - To account for vintage cars' higher price due to rarity and originality

Then, change data type into numeric format
  - Label Encoder applied to categorical variables

Data Preprocessing & Feature Engineering
Post-Modeling - Re-engineering

Then after performing low on running models, we re-engineered some of our features

Contextual Anomaly Detection

1. Odometer:

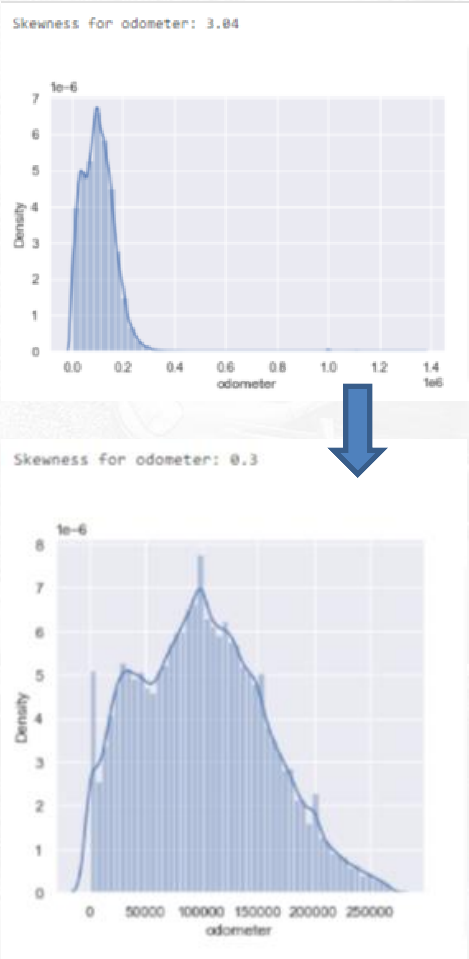
- Standard car odometer should have max 300,000 miles. Trimming data with 75th quantile + 3*IQR, ~268,000, as a cut reduced skewness from 3.04 to 0.3
- Only new car should have 0 odometer so trimmed otherwise

2. Price:

- Dropped cars below $1000 and greater than $200,000 that are in the extreme ranges not fit for our analysis purpose
- Max car price was three billion dollars

3. MSRP:

- MSRP is the manufacturer's suggested retail price (list price)
- Dropped MSRP < car price as MSRP should be higher than used car selling price



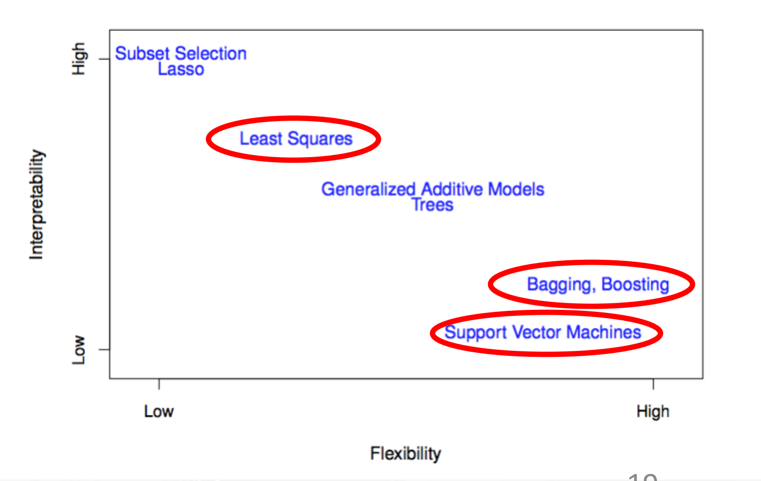
Modeling & Evaluation
Models for Consideration

- Our goal is to predict the sale price of a used car, which is a ***supervised regression*** problem. We pick our models base on two considerations, *flexibility (accuracy)* and *interpretability*.

- We value *model accuracy over interpretability* because:
  - The industry we are in doesn't require we provide explanation for the decision we make.
  - Features of our used car dataset are easy to understand, thus making it easy for us to debug the model even without high model interpretability.

- Models to consider are:
  - *Linear Regression*
  - *Support Vector Regression with linear kernel*
  - *Decision Trees Ensemble method*
    - *Bagging Trees (Random Forest)*
    - *Boosting Trees*

Fig 1. Flexibility vs. interpretability tradeoffs for models



Modeling & Evaluation
Our Approach

Our Guesses for Models:

- *Linear Regression* is not flexible enough to capture all the variance of the model
- *SVR* would be very slow to train. (SVR training time scale badly with large number of training sample)
- *Ensemble Trees* would be the best method as it is flexible and has decent interpretability

Our Approach:

- Train and tune all the models and compare the models' accuracy
- Select the model with the best metric scores

Our metrics:

- R squared: the proportion of the variance explained by the model
- Root Mean Squared Error: $\sqrt{\frac{\sum(x_i - \tilde{x}_i)^2}{N}}$
- Mean Absolute Proportional Error: $\frac{1}{N}\sqrt{\frac{\sum|(x_i - \tilde{x}_i)|}{x_i}}$

x_i $- ith\ observed\ value$

$\tilde{x}_i$ $- ith\ predicted\ value$

$N - Total\ Number\ of\ obsevation$

$i \in [1, N]$

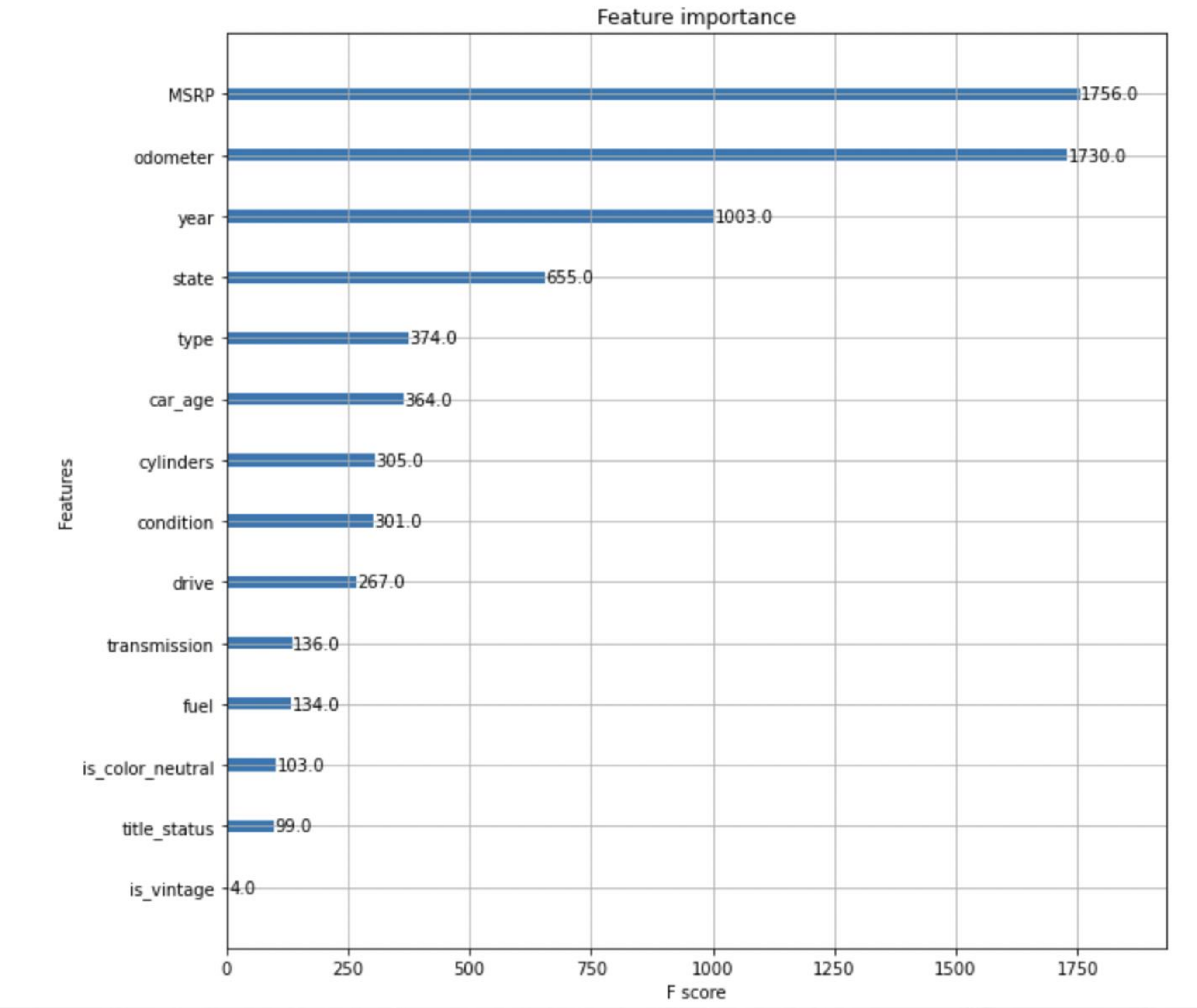
Modeling & Evaluation
Results

Results:

- Just like our expectation, Ensemble Tree methods, specifically ***XGBoost*** has the best overall performance.
- This is not very surprising since ensemble method is known to:
  - Have higher predictive accuracy, compared to the individual models.
  - Be very useful when there is both linear and non-linear type of data in the dataset
- *Linear Regression* and *Linear SVR*, like expected, didn't perform well. From the R squared score, we see that both models could not capture all the variance of our data.
- XGBoost performs better than Random Forest.

Model	R Squared	Train RMSE	Validation/ Test RMSE	Test MAPE
Linear Regression	0.68	5222.94	5272.66	0.49
Linear SVR	0.65	5613.04	5626.03	0.43
Random Forest	0.89	2971.65	3215.39	0.25
Gradient Boosting Machine	0.89	3111.86	3179.89	0.25
XGBoost	0.92	2400.81	2674.25	0.208

Feature Importance



Insights

Key Findings:

- MSRP, odometer, and production year are proven to be top 3 strongest determinants of used car prices.
  - *Expected from initial EDA as we observed correlations*
- States determine price range.
- Higher price variance as years go by.
- Some cars are not being sold as advertised (ex. Vintage cars may be lemons).

Challenges/Areas of Improvement:

- Employ highly advanced NLP on textual data (description) excluding Ads, supplement the data with public reviews on each car, and apply topical modeling into our features.
- Perform deeper research on car models with missing values and perform more thorough anomaly detection.
- We could integrate image detection algorithms to see whether car is described as it is and additionally use them as features for modelling (CNN Image Classification)

Recommendations

Proposed Business Application To Problems of Information Asymmetry:

- Craigslist should require sellers to fill in clearly defined forms for used cars so that 'information asymmetry' can be mitigated. (Now, it is not mandatory. 'Condition' criteria is also not clear, while this can be an important indicator.)

- Craigslist or other platforms can present predictions (using the predictive model) of used cars so that buyers can get a sense of what is reasonable and have a base point for comparison.

- Craigslist can also add exception criteria or specific section for vintage cars.

- For reputation and quality assurance purposes, used car companies can use the predictions to target and filter out sellers prone to selling lemons prior to posting for sale.

Eventually, all these adjustments can be expected to improve the quality of used car listings in Craigslist, which in turn, can improve transaction success rate.

References

Data Sources:

- **Used Car Dataset:** https://www.kaggle.com/austinreese/craigslist-carstrucks-data
- **MSRP Dataset:** https://www.kaggle.com/CooperUnion/cardataset
- **iSeeCar:** https://www.iseecars.com/

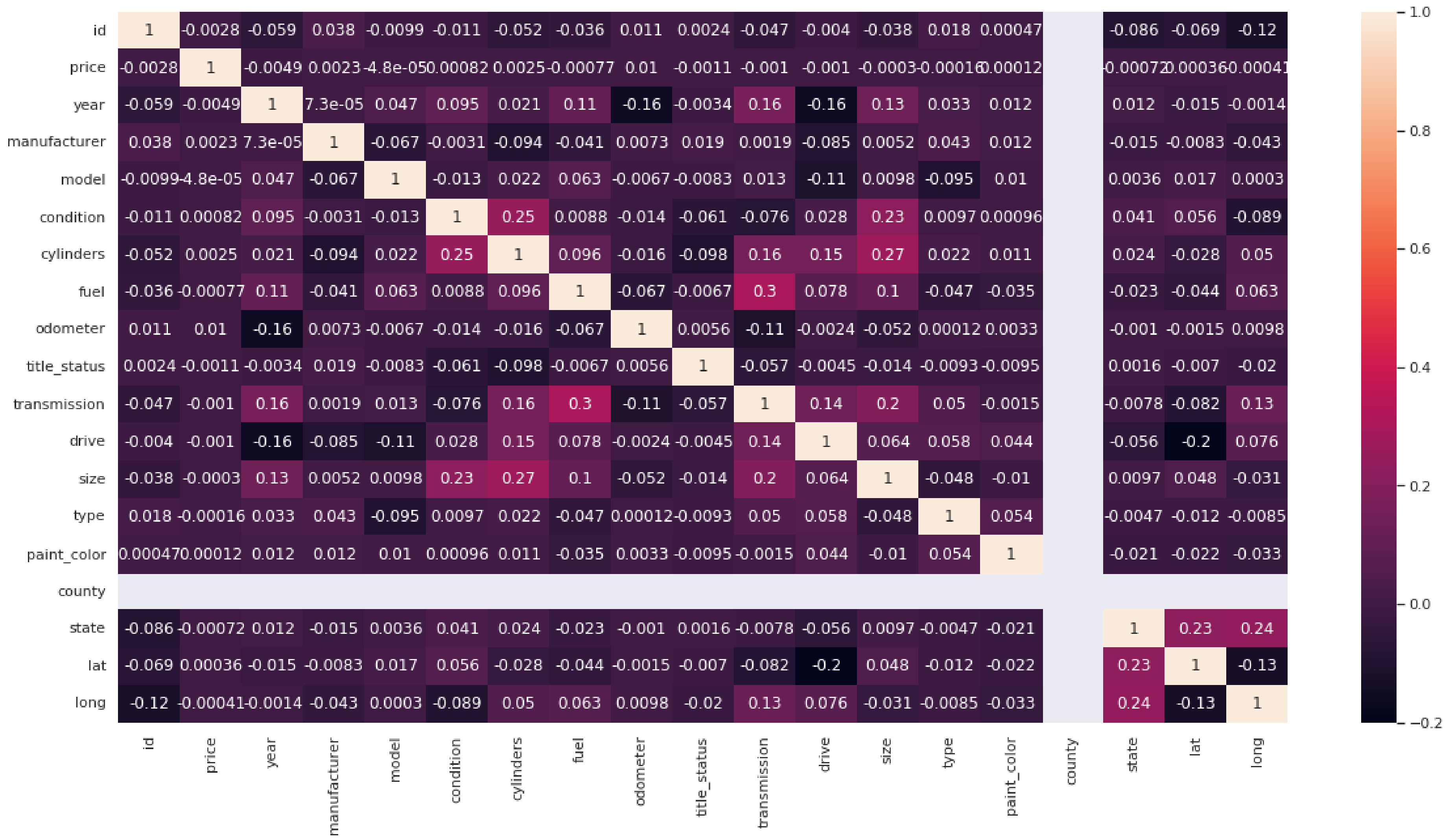
Tools:

- **Top2Vec:** https://github.com/ddangelov/Top2Vec
- **OpenRefine:** https://openrefine.org/
- **DiffLib:** https://docs.python.org/3/library/difflib.html
- **Scrapy:** https://scrapy.org/

THANK YOU

Appendix

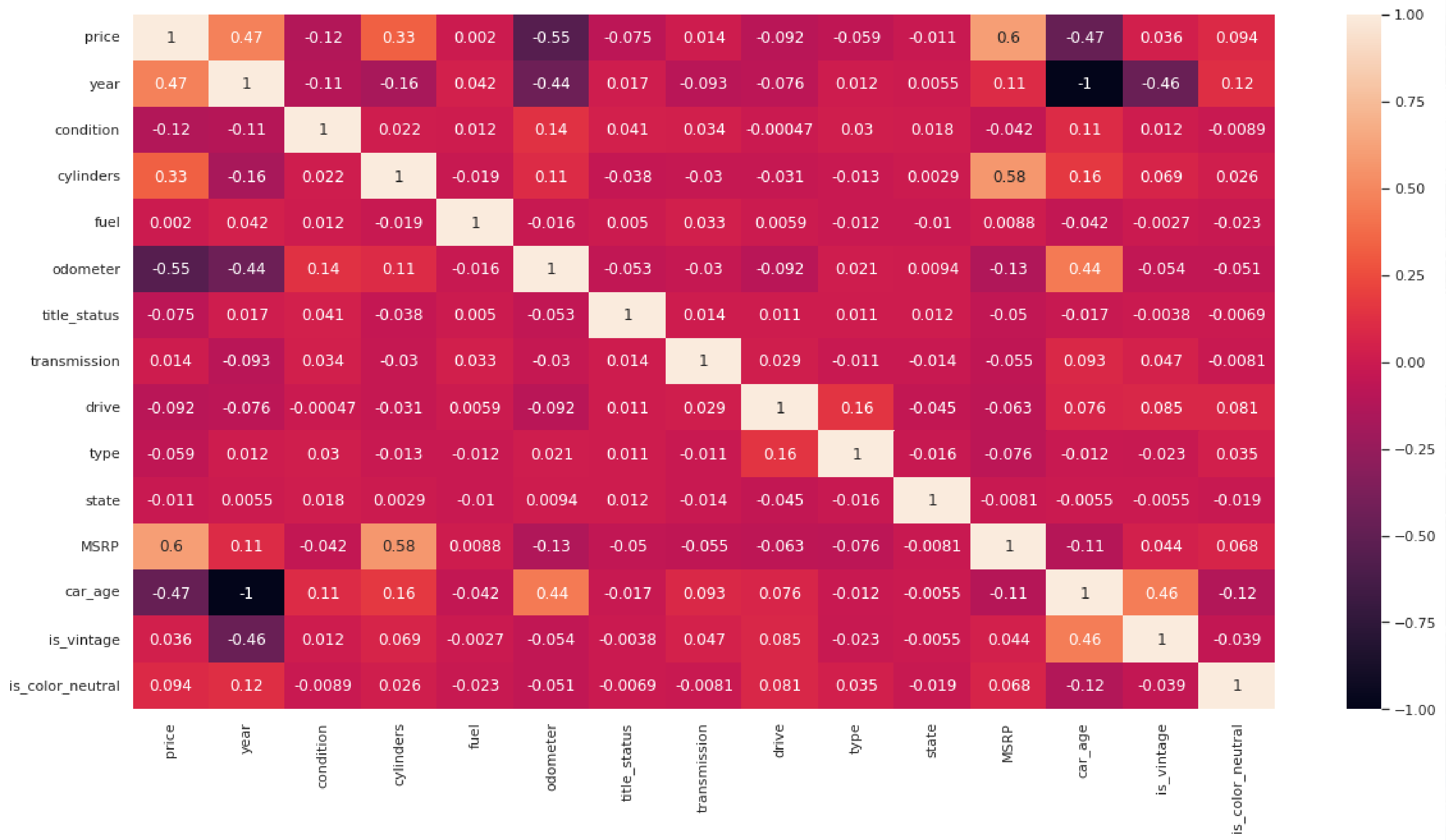
Correlation Table of the Raw Data



	id	price	year	manufacturer	model	condition	cylinders	fuel	odometer	title_status	transmission	drive	size	type	paint_color	county	state	lat	long
id	1	-0.0028	-0.059	0.038	-0.0099	-0.011	-0.052	-0.036	0.011	0.0024	-0.047	-0.004	-0.038	0.018	0.00047		-0.086	-0.069	-0.12
price	-0.0028	1	-0.0049	0.0023	-4.8e-05	0.00082	0.0025	-0.00077	0.01	-0.0011	-0.001	-0.001	-0.0003	-0.00016	0.00012		-0.00072	0.00036	-0.00041
year	-0.059	-0.0049	1	7.3e-05	0.047	0.095	0.021	0.11	-0.16	-0.0034	0.16	-0.16	0.13	0.033	0.012		0.012	-0.015	-0.0014
manufacturer	0.038	0.0023	7.3e-05	1	-0.067	-0.0031	-0.094	-0.041	0.0073	0.019	0.0019	-0.085	0.0052	0.043	0.012		-0.015	-0.0083	-0.043
model	-0.0099	-4.8e-05	0.047	-0.067	1	-0.013	0.022	0.063	-0.0067	-0.0083	0.013	-0.11	0.0098	-0.095	0.01		0.0036	0.017	0.0003
condition	-0.011	0.00082	0.095	-0.0031	-0.013	1	0.25	0.0088	-0.014	-0.061	-0.076	0.028	0.23	0.0097	0.00096		0.041	0.056	-0.089
cylinders	-0.052	0.0025	0.021	-0.094	0.022	0.25	1	0.096	-0.016	-0.098	0.16	0.15	0.27	0.022	0.011		0.024	-0.028	0.05
fuel	-0.036	-0.00077	0.11	-0.041	0.063	0.0088	0.096	1	-0.067	-0.0067	0.3	0.078	0.1	-0.047	-0.035		-0.023	-0.044	0.063
odometer	0.011	0.01	-0.16	0.0073	-0.0067	-0.014	-0.016	-0.067	1	0.0056	-0.11	-0.0024	-0.052	0.00012	0.0033		-0.001	-0.0015	0.0098
title_status	0.0024	-0.0011	-0.0034	0.019	-0.0083	-0.061	-0.098	-0.0067	0.0056	1	-0.057	-0.0045	-0.014	-0.0093	-0.0095		0.0016	-0.007	-0.02
transmission	-0.047	-0.001	0.16	0.0019	0.013	-0.076	0.16	0.3	-0.11	-0.057	1	0.14	0.2	0.05	-0.0015		-0.0078	-0.082	0.13
drive	-0.004	-0.001	-0.16	-0.085	-0.11	0.028	0.15	0.078	-0.0024	-0.0045	0.14	1	0.064	0.058	0.044		-0.056	-0.2	0.076
size	-0.038	-0.0003	0.13	0.0052	0.0098	0.23	0.27	0.1	-0.052	-0.014	0.2	0.064	1	-0.048	-0.01		0.0097	0.048	-0.031
type	0.018	-0.00016	0.033	0.043	-0.095	0.0097	0.022	-0.047	0.00012	-0.0093	0.05	0.058	-0.048	1	0.054		-0.0047	-0.012	-0.0085
paint_color	0.00047	0.00012	0.012	0.012	0.01	0.00096	0.011	-0.035	0.0033	-0.0095	-0.0015	0.044	-0.01	0.054	1		-0.021	-0.022	-0.033
county																			
state	-0.086	-0.00072	0.012	-0.015	0.0036	0.041	0.024	-0.023	-0.001	0.0016	-0.0078	-0.056	0.0097	-0.0047	-0.021		1	0.23	0.24
lat	-0.069	0.00036	-0.015	-0.0083	0.017	0.056	-0.028	-0.044	-0.0015	-0.007	-0.082	-0.2	0.048	-0.012	-0.022		0.23	1	-0.13
long	-0.12	-0.00041	-0.0014	-0.043	0.0003	-0.089	0.05	0.063	0.0098	-0.02	0.13	0.076	-0.031	-0.0085	-0.033		0.24	-0.13	1

Appendix

Correlation Table of the Clean Dataset



	price	year	condition	cylinders	fuel	odometer	title_status	transmission	drive	type	state	MSRP	car_age	is_vintage	is_color_neutral
price	1	0.47	-0.12	0.33	0.002	-0.55	-0.075	0.014	-0.092	-0.059	-0.011	0.6	-0.47	0.036	0.094
year	0.47	1	-0.11	-0.16	0.042	-0.44	0.017	-0.093	-0.076	0.012	0.0055	0.11	-1	-0.46	0.12
condition	-0.12	-0.11	1	0.022	0.012	0.14	0.041	0.034	-0.00047	0.03	0.018	-0.042	0.11	0.012	-0.0089
cylinders	0.33	-0.16	0.022	1	-0.019	0.11	-0.038	-0.03	-0.031	-0.013	0.0029	0.58	0.16	0.069	0.026
fuel	0.002	0.042	0.012	-0.019	1	-0.016	0.005	0.033	0.0059	-0.012	-0.01	0.0088	-0.042	-0.0027	-0.023
odometer	-0.55	-0.44	0.14	0.11	-0.016	1	-0.053	-0.03	-0.092	0.021	0.0094	-0.13	0.44	-0.054	-0.051
title_status	-0.075	0.017	0.041	-0.038	0.005	-0.053	1	0.014	0.011	0.011	0.012	-0.05	-0.017	-0.0038	-0.0069
transmission	0.014	-0.093	0.034	-0.03	0.033	-0.03	0.014	1	0.029	-0.011	-0.014	-0.055	0.093	0.047	-0.0081
drive	-0.092	-0.076	-0.00047	-0.031	0.0059	-0.092	0.011	0.029	1	0.16	-0.045	-0.063	0.076	0.085	0.081
type	-0.059	0.012	0.03	-0.013	-0.012	0.021	0.011	-0.011	0.16	1	-0.016	-0.076	-0.012	-0.023	0.035
state	-0.011	0.0055	0.018	0.0029	-0.01	0.0094	0.012	-0.014	-0.045	-0.016	1	-0.0081	-0.0055	-0.0055	-0.019
MSRP	0.6	0.11	-0.042	0.58	0.0088	-0.13	-0.05	-0.055	-0.063	-0.076	-0.0081	1	-0.11	0.044	0.068
car_age	-0.47	-1	0.11	0.16	-0.042	0.44	-0.017	0.093	0.076	-0.012	-0.0055	-0.11	1	0.46	-0.12
is_vintage	0.036	-0.46	0.012	0.069	-0.0027	-0.054	-0.0038	0.047	0.085	-0.023	-0.0055	0.044	0.46	1	-0.039
is_color_neutral	0.094	0.12	-0.0089	0.026	-0.023	-0.051	-0.0069	-0.0081	0.081	0.035	-0.019	0.068	-0.12	-0.039	1