Data Incubator Capstone Project Proposal

Michelle (Qin) Peng

Dataset Overview

- **Data source:**

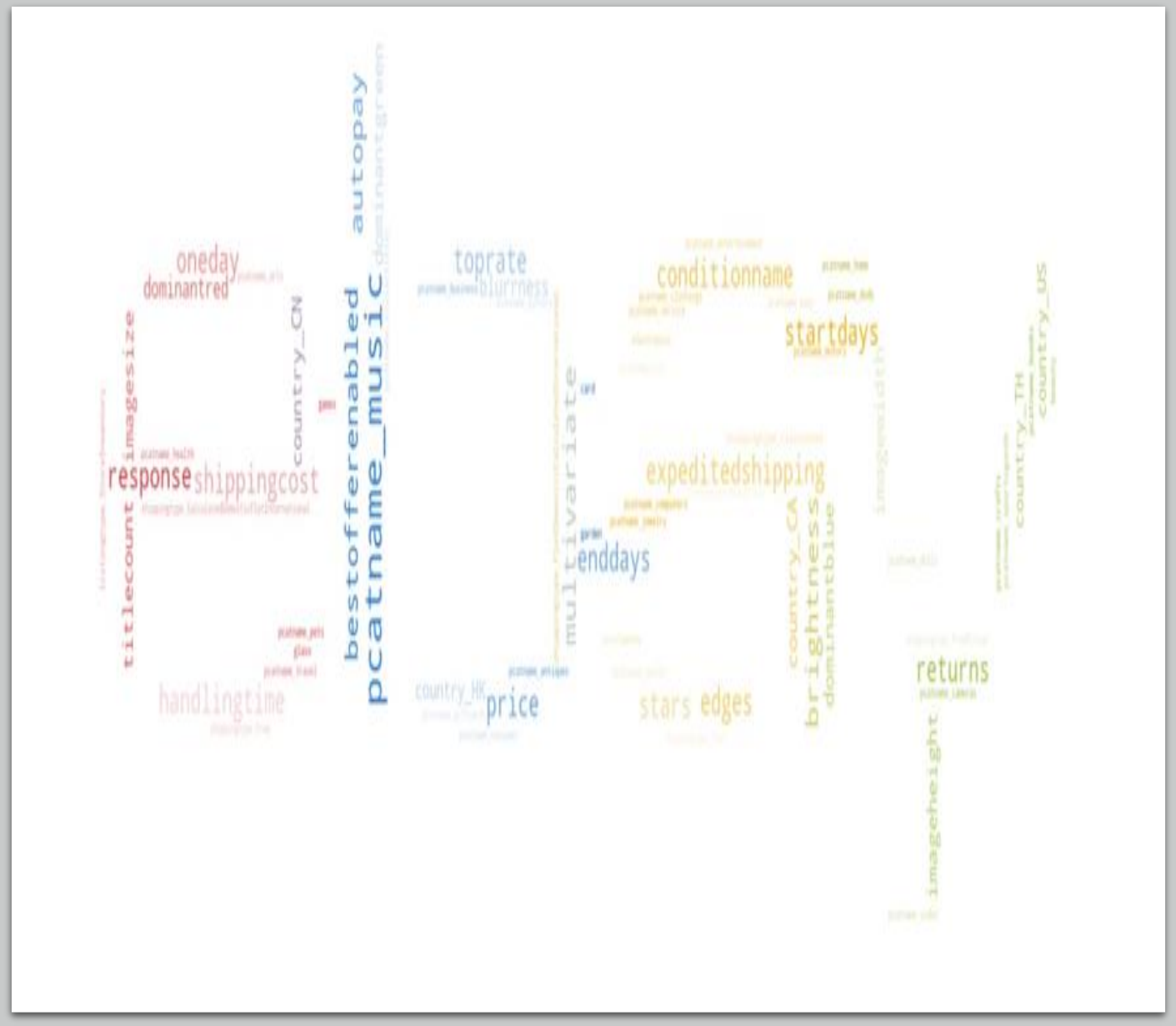
Ebay API and Ebay WebScrapping

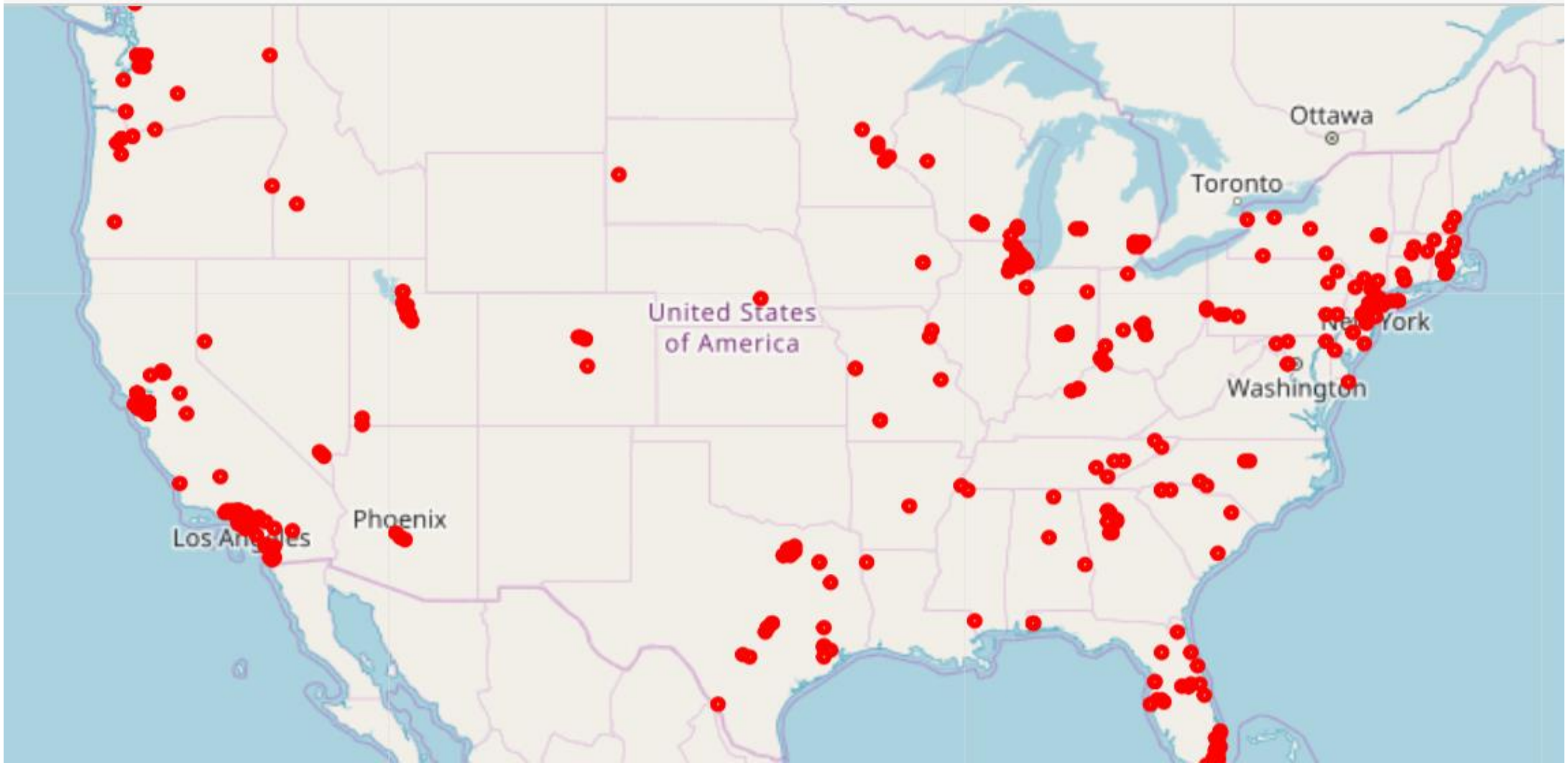
- **Number of Observations:** 708
- **Response variable:**

Sales Percentage (How much is sold)

- **Features:**

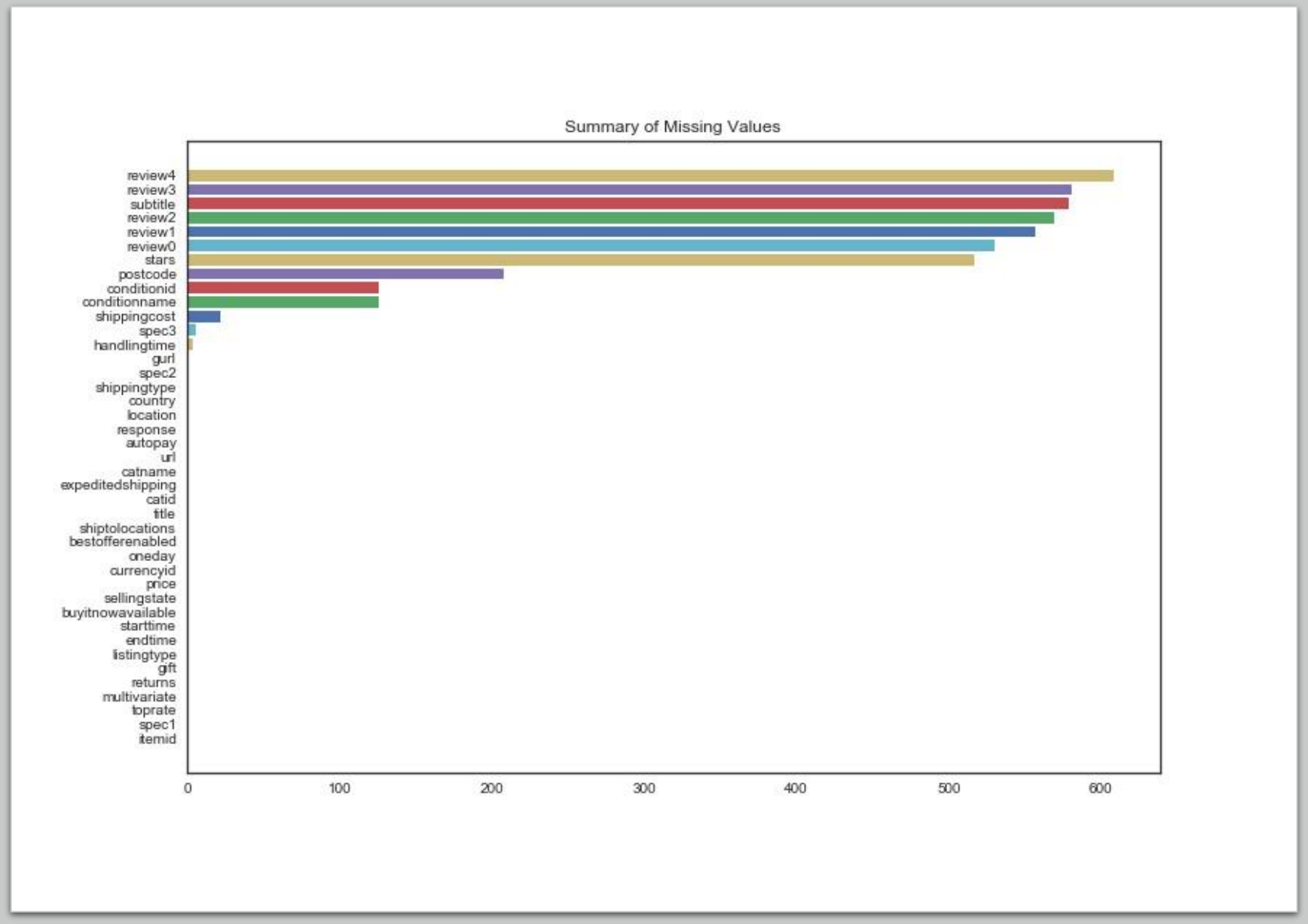
31 (Categorical Variables (6), Numerical Variables (25))

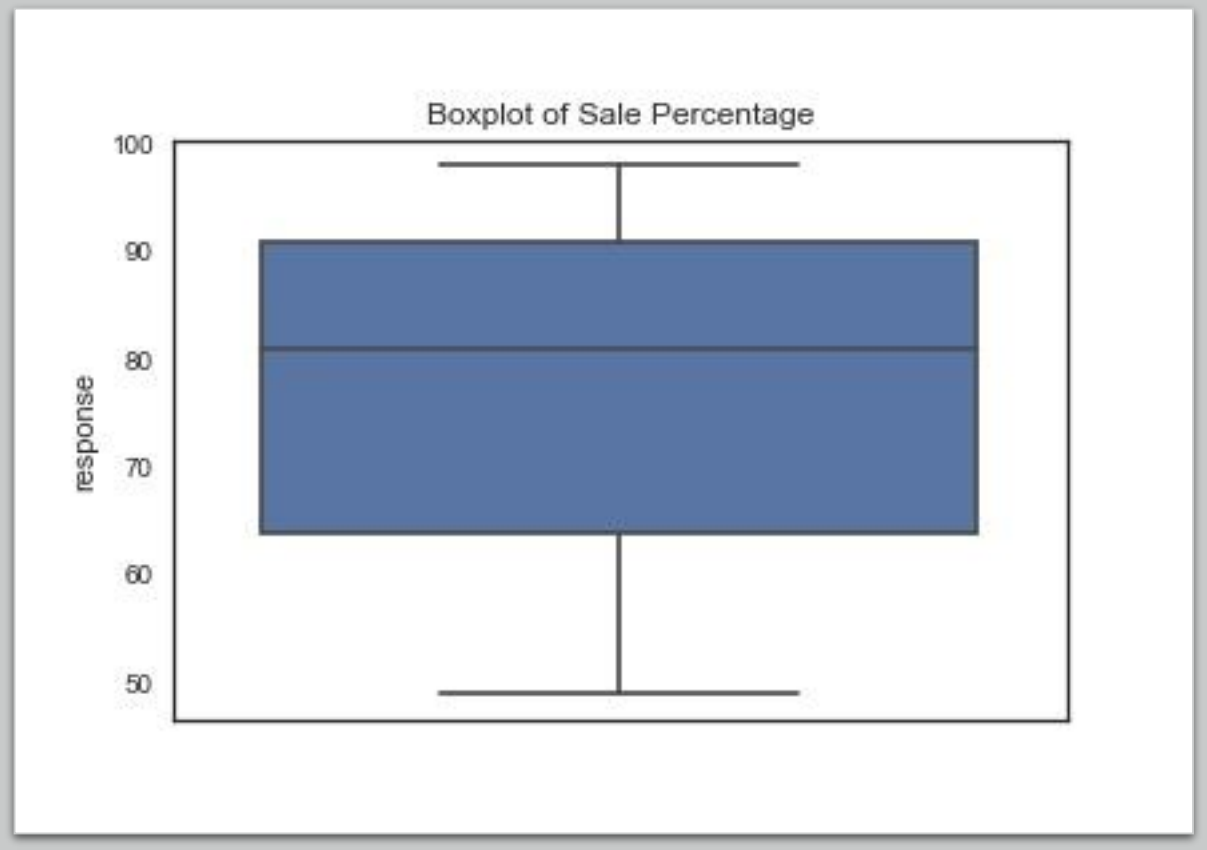


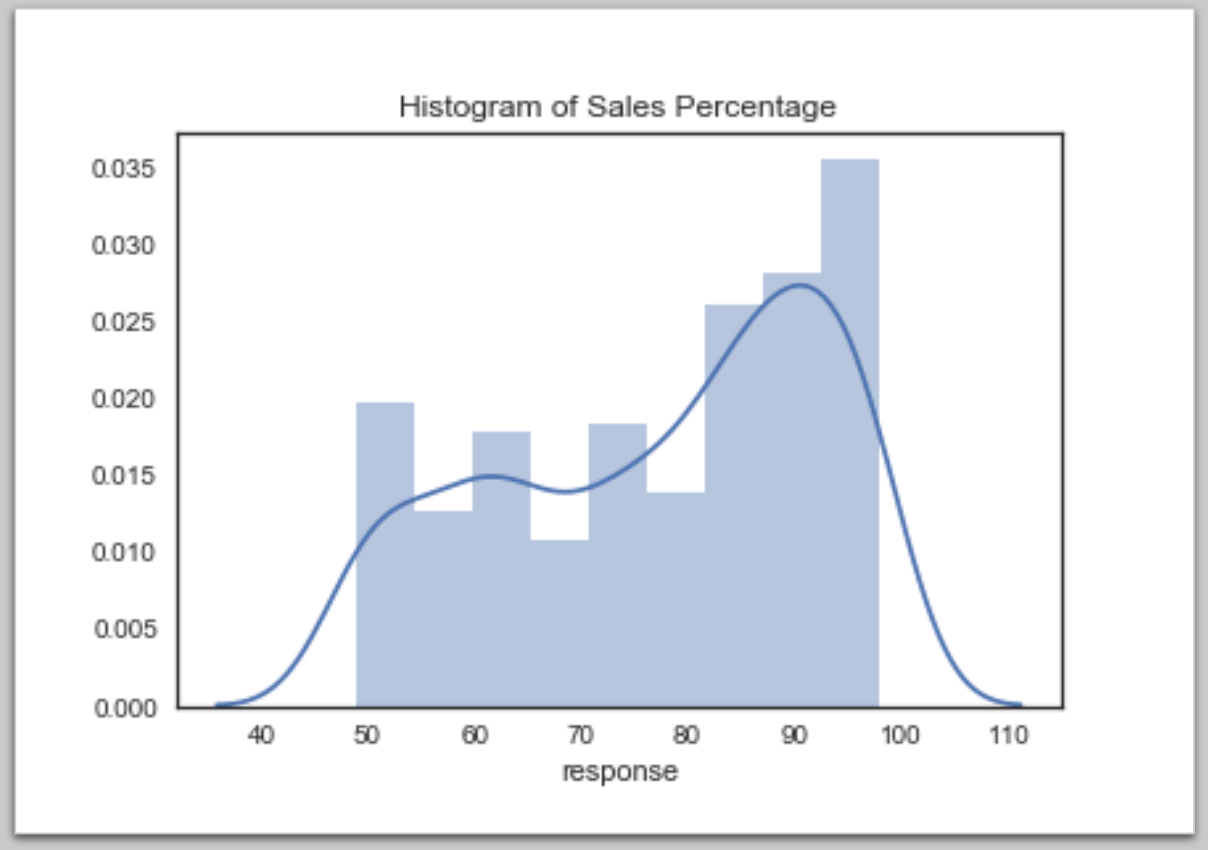

Ottawa
Toronto
United States
of America
Washington
Phoenix

Exploratory Data Analysis

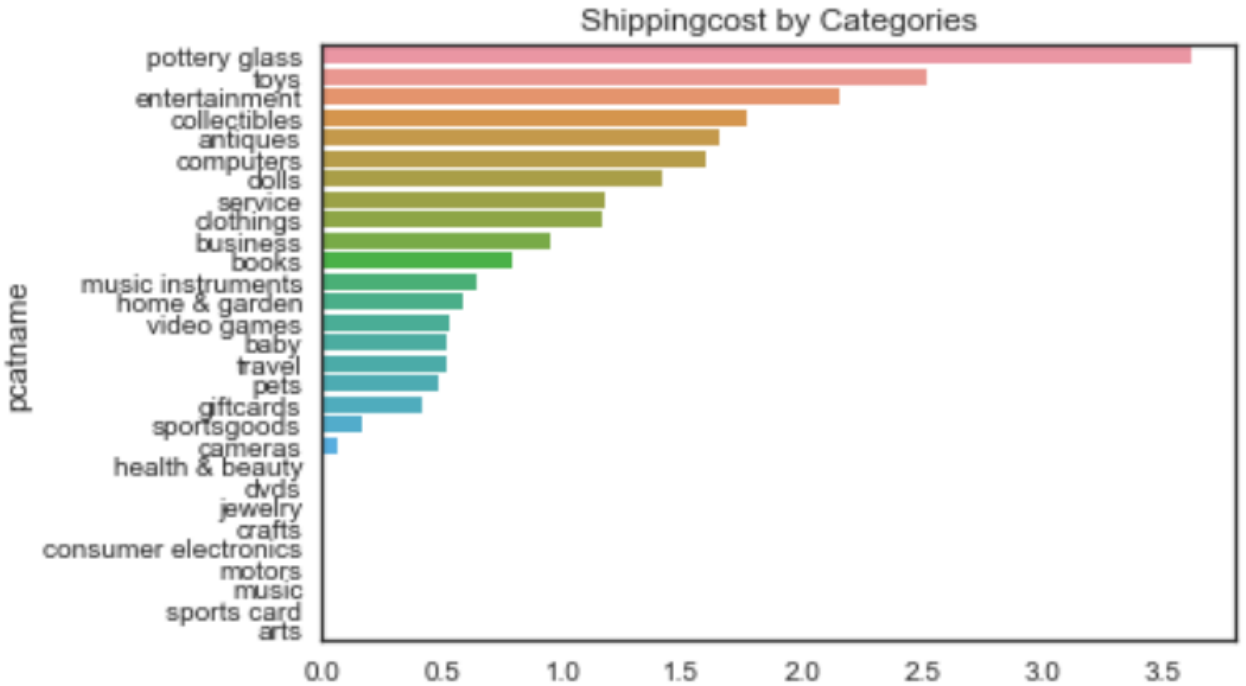
Missing Values

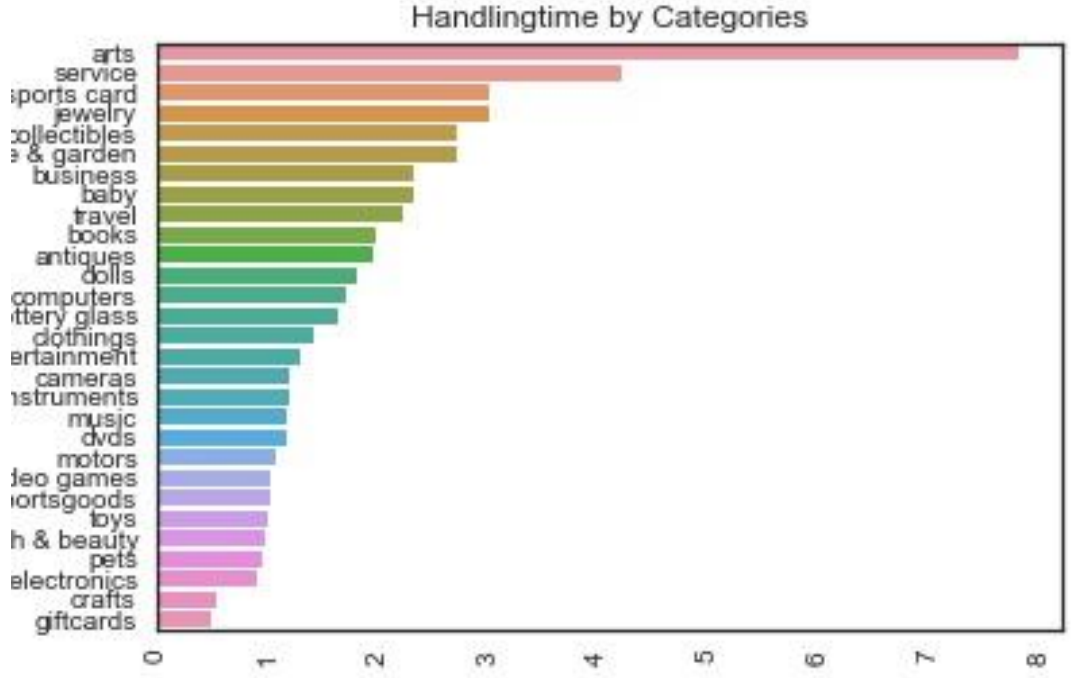


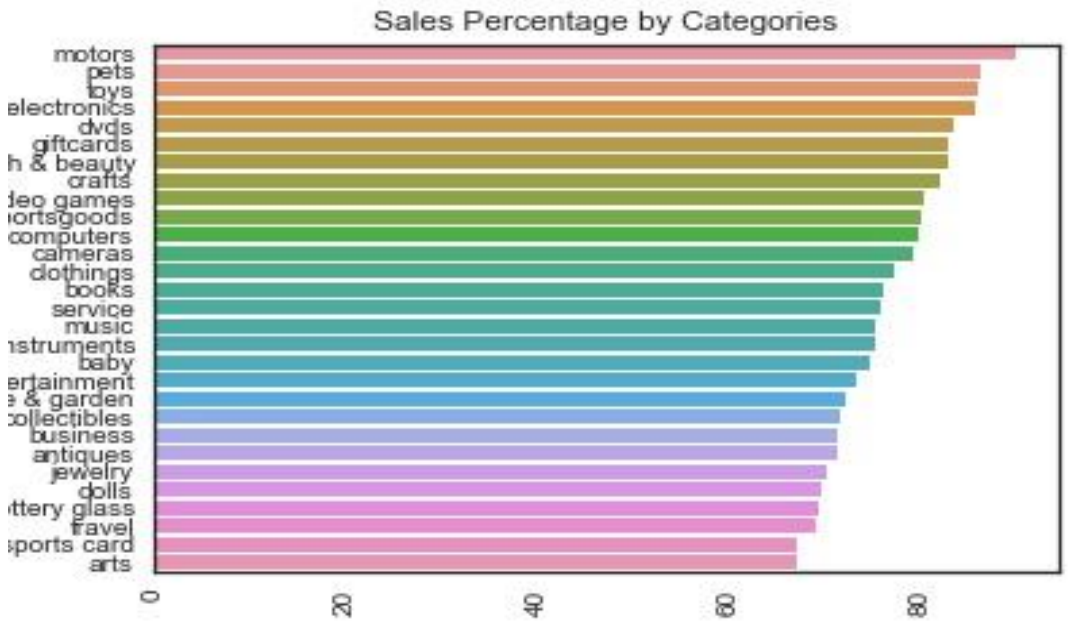


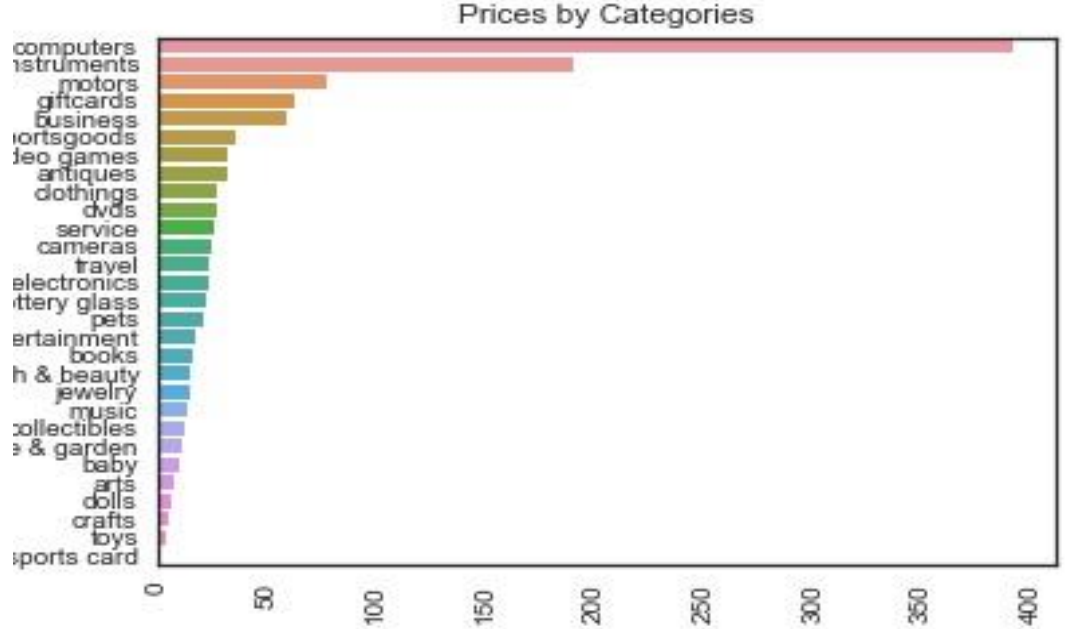


Response Variable: Sales Percentage


Shippingcost by Categories
pcatname
pottery glass
toys
entertainment
collectibles
antiques
computers
dolls
service
clothings
business
books
music instruments
home & garden
video games
baby
travel
pets
giftcards
sportsgoods
cameras
health & beauty
dvds
jewelry
crafts
consumer electronics
motors
music
sports card
arts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Handlingtime by Categories
arts
service
sports card
jewelry
collectibles
& garden
business
baby
travel
books
antiques
dolls
computers
ttery glass
clothings
ertainment
cameras
nstruments
music
dvds
motors
deo games
ortsgoods
toys
h & beauty
pets
electronics
crafts
giftcards
0
1
2
3
4
5
6
7
8

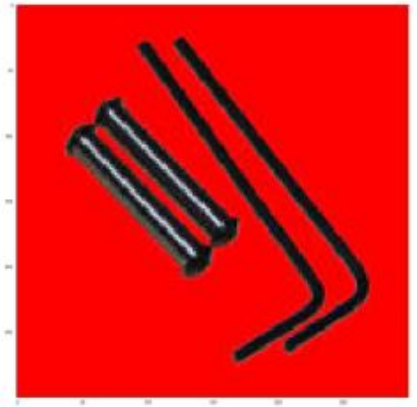

Sales Percentage by Categories
motors
pets
toys
electronics
dvds
giftcards
h & beauty
crafts
deo games
ortsgoods
computers
cameras
clothings
books
service
music
nstruments
baby
ertainment
& garden
collectibles
business
antiques
jewelry
dolls
ttery glass
travel
sports card
arts
0
20
40
60
80


Prices by Categories
computers
nstruments
motors
giftcards
business
ortsgoods
deo games
antiques
clothings
dvds
service
cameras
travel
electronics
ttery glass
pets
ertainment
books
h & beauty
jewelry
music
collectibles
& garden
baby
arts
dolls
crafts
toys
sports card
0
50
100
150
200
250
300
350
400

Feature Engineering

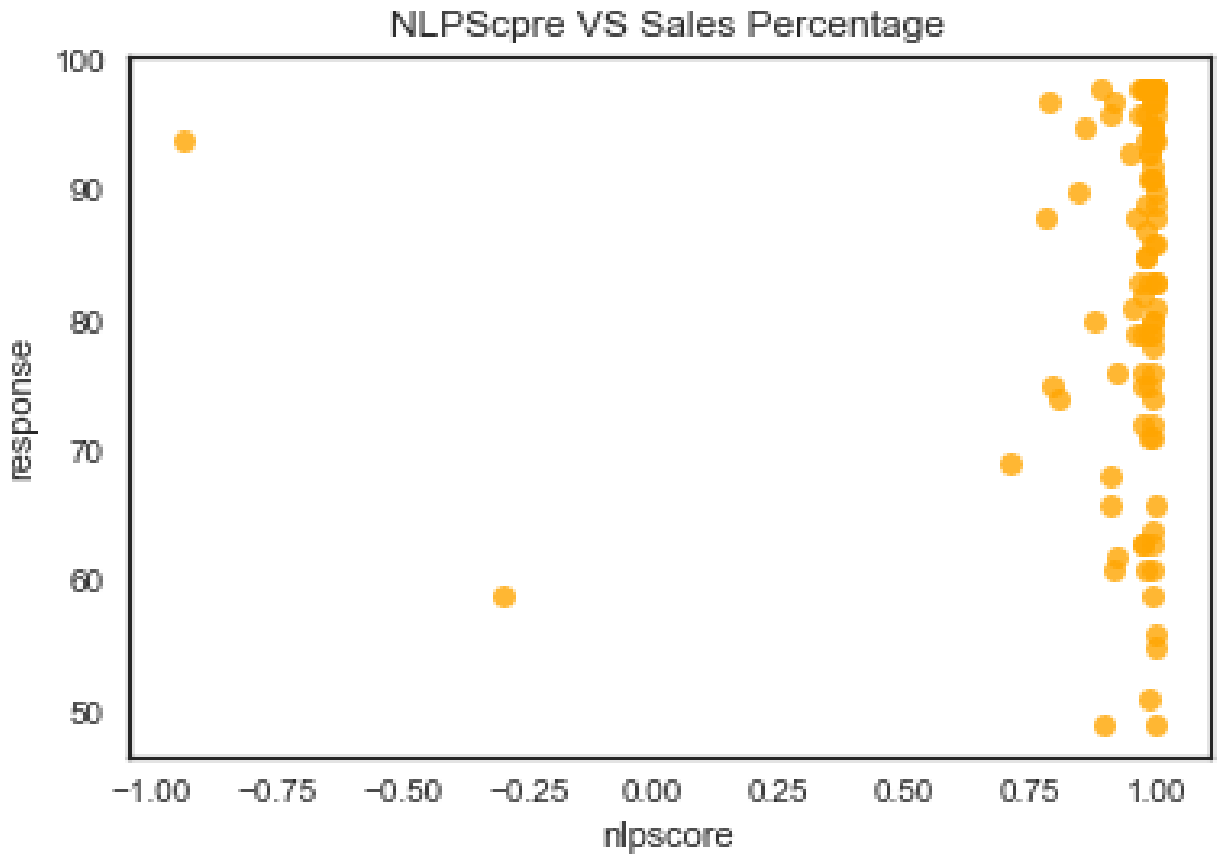
Image Variables:

Each picture is broken into Brightness, Blurness, Dominant Colors, Size


NLP Variables:

A sentiment score is calculated for each review using the VaderSentiment

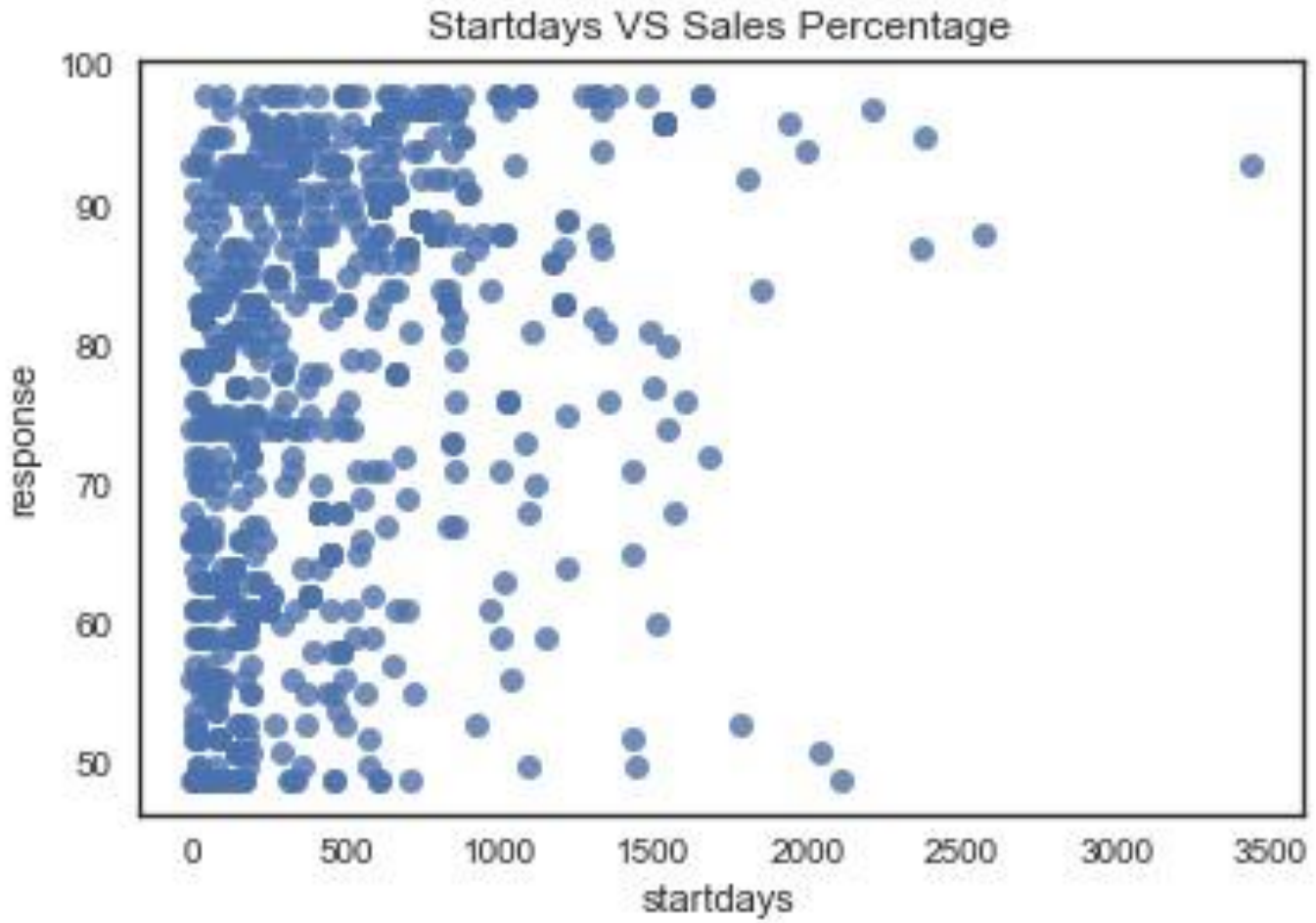


“I\'m a longtime Shania fan but "Now" is a dud in my books.
It really saddens me to say that.
It\'s very different from what she has previously released
and I just cannot get into it.
If your expecting the sounds you\'ve come accustomed to with..”

{'compound': 0.7781, 'neg': 0.08, 'neu': 0.788, 'pos': 0.132}

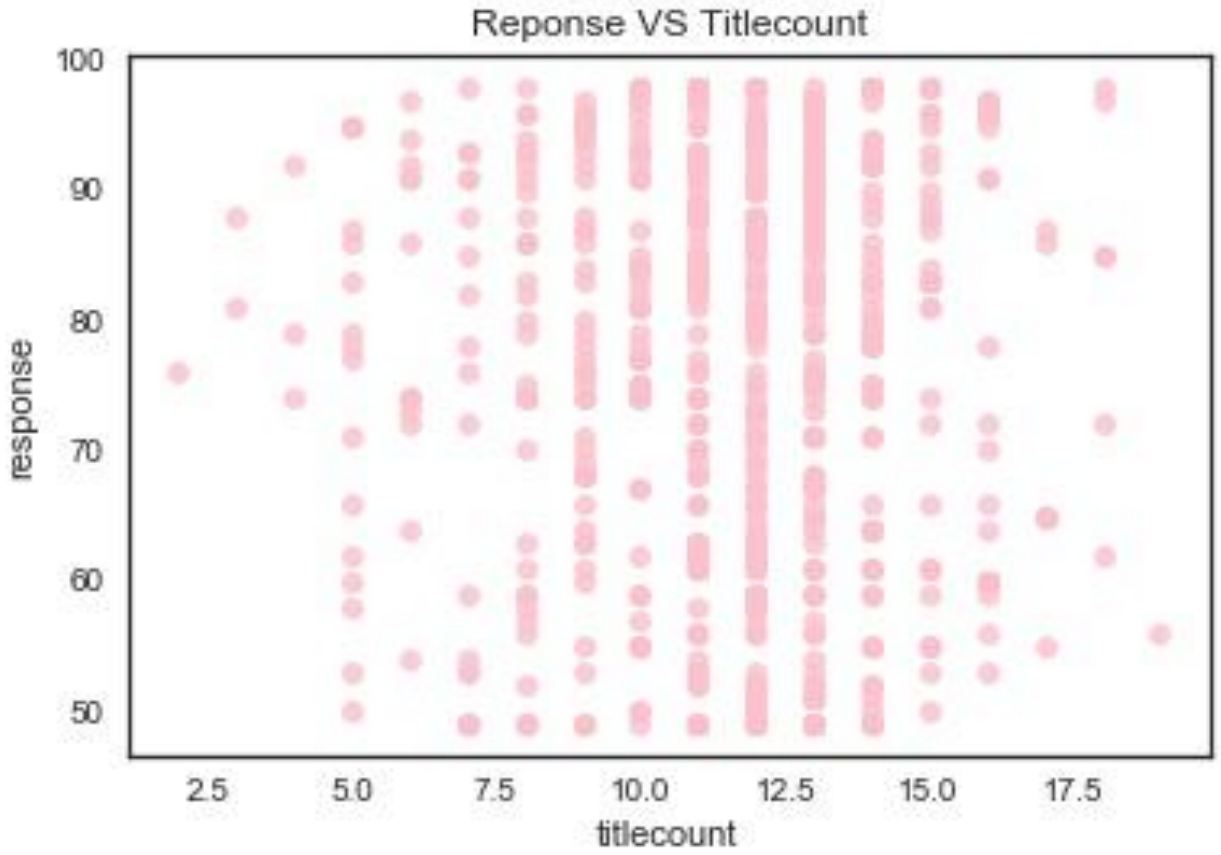
Date Variables

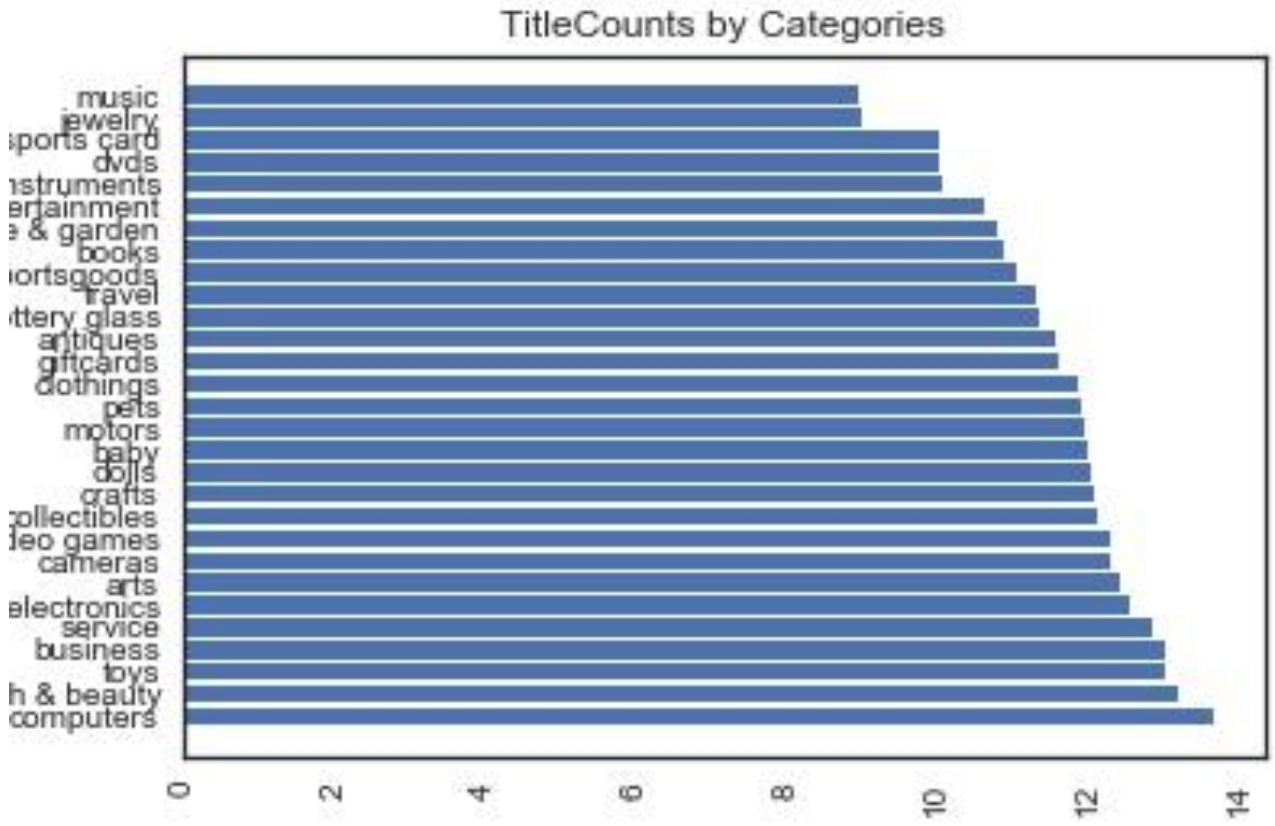
Turns starttime, endtime into startdays, enddays which means how long have the product has been posted(startdays) and how many days left(enddays

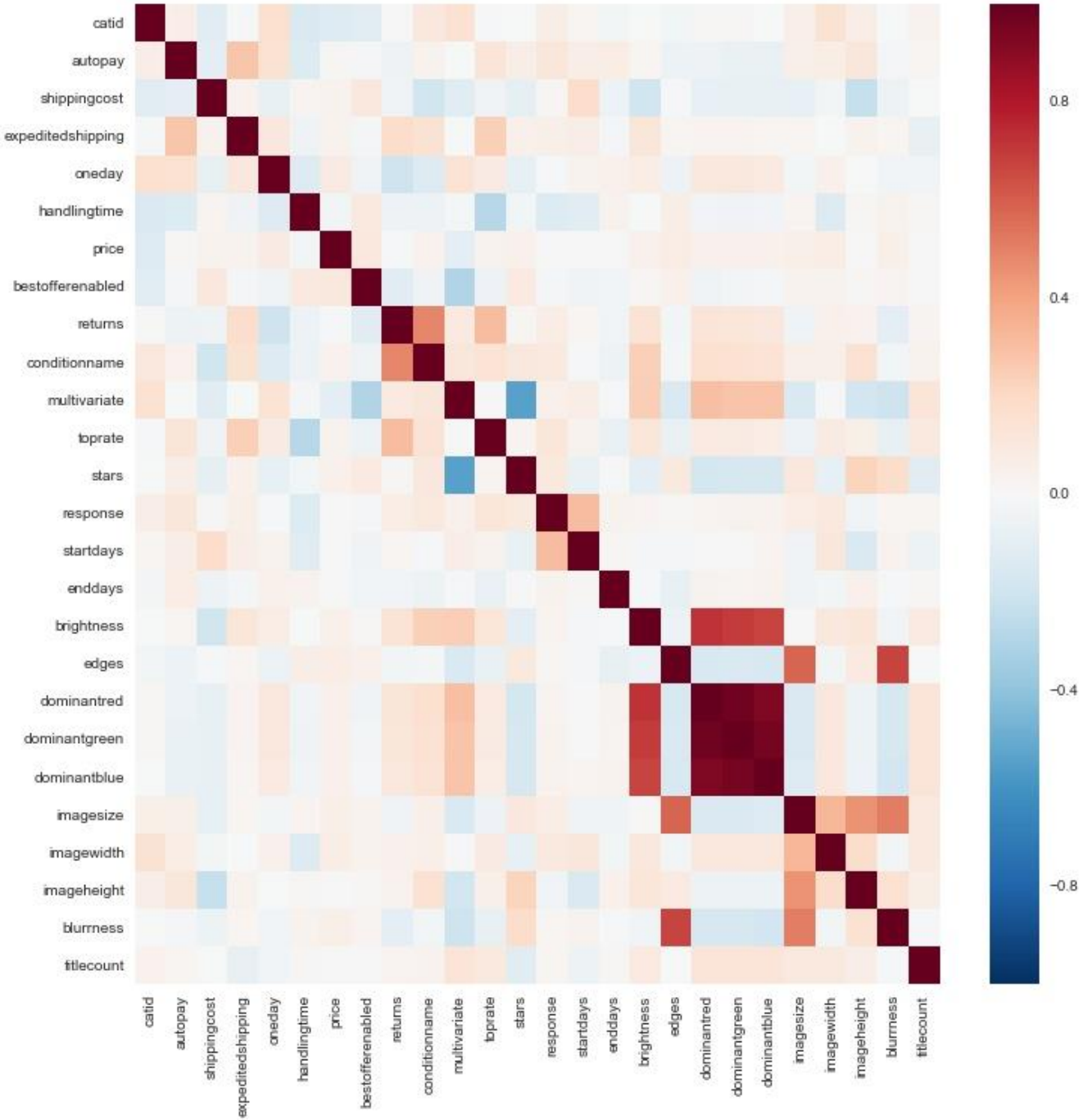


Title Variables

Construct the counts of words in the title.







Strong Correlation pairs:

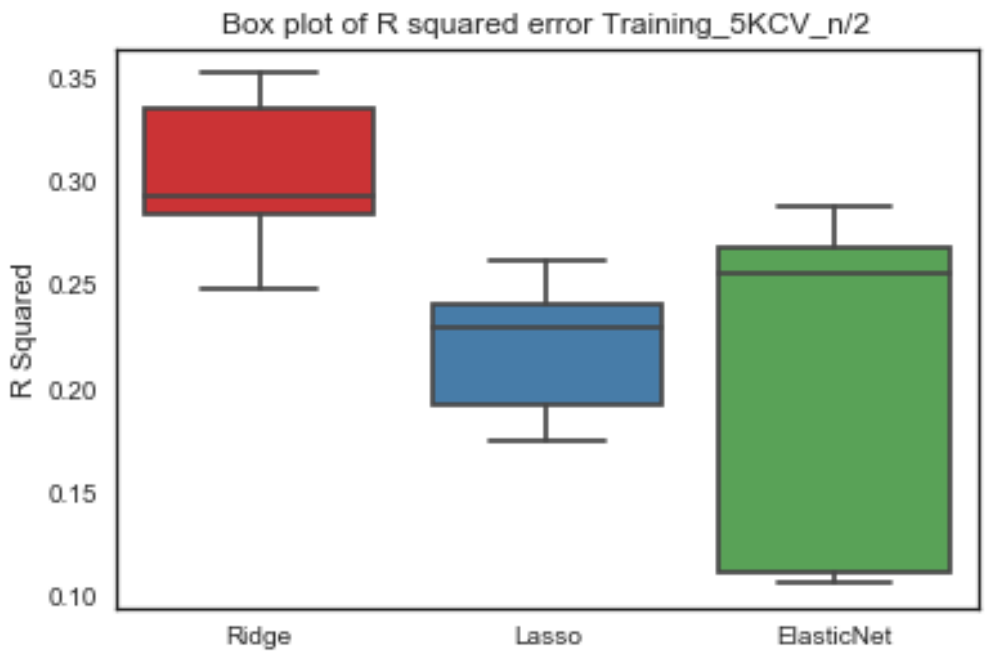
1. Image colors
2. Image size VS Image widths & heights
3. Returns VS ConditionName
4. Number of Stars VS Multivariate Listing
5. Sales Percentage VS
6. Startdays

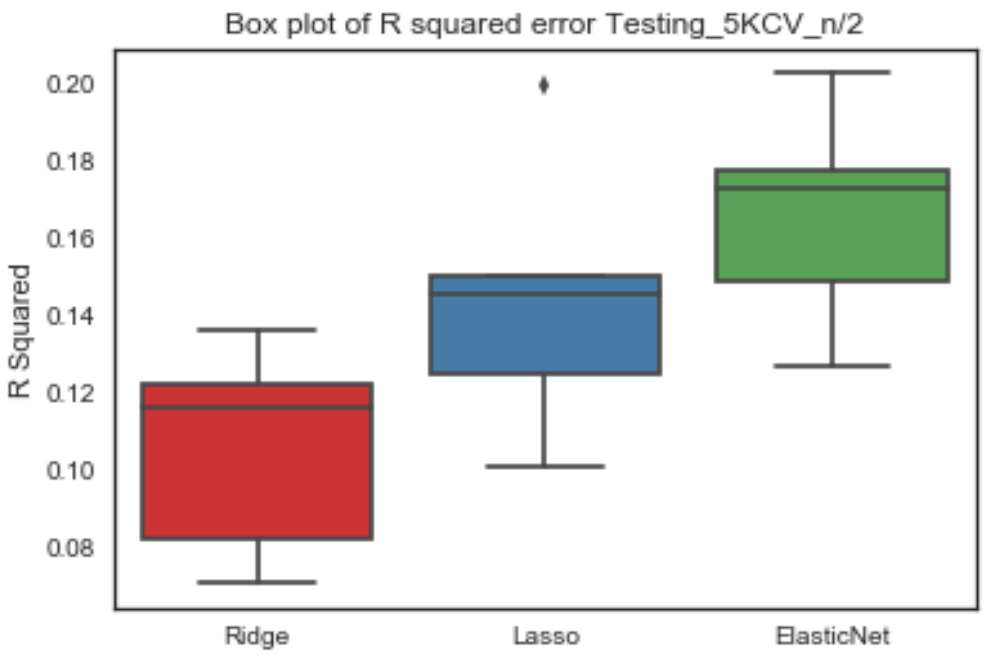
Basic Model Fitting

OLS Regression Results

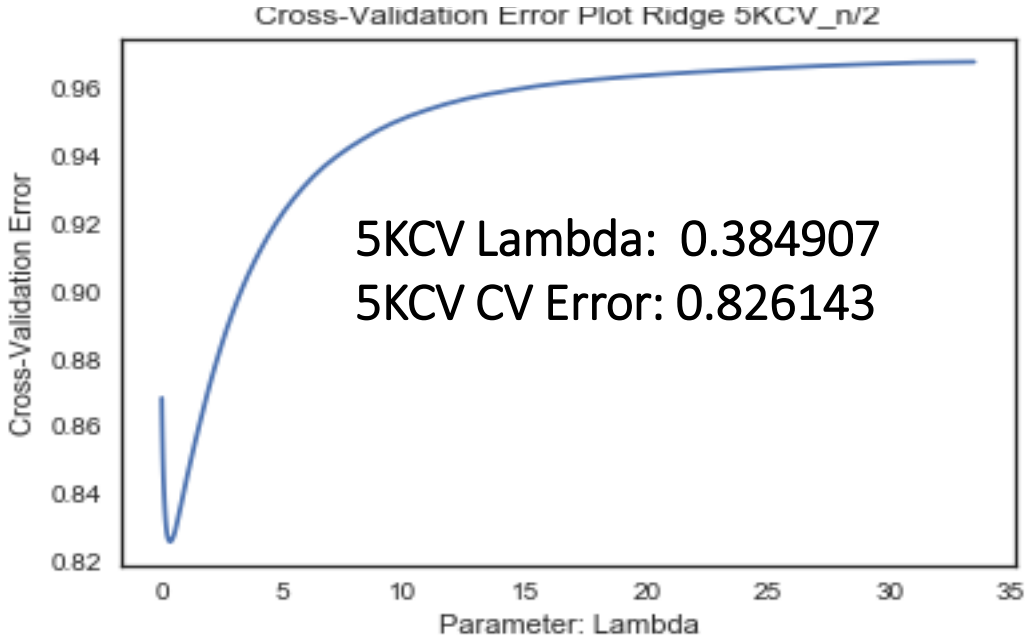
Dep. Variable:	y	R-squared:	0.296
Model:	OLS	Adj. R-squared:	0.227
Method:	Least Squares	F-statistic:	4.276
Date:	Tue, 15 May 2018	Prob (F-statistic):	2.25e-21
Time:	01:57:18	Log-Likelihood:	-856.22
No. Observations:	692	AIC:	1836.
Df Residuals:	630	BIC:	2118.
Df Model:	62		
Covariance Type:	nonrobust		

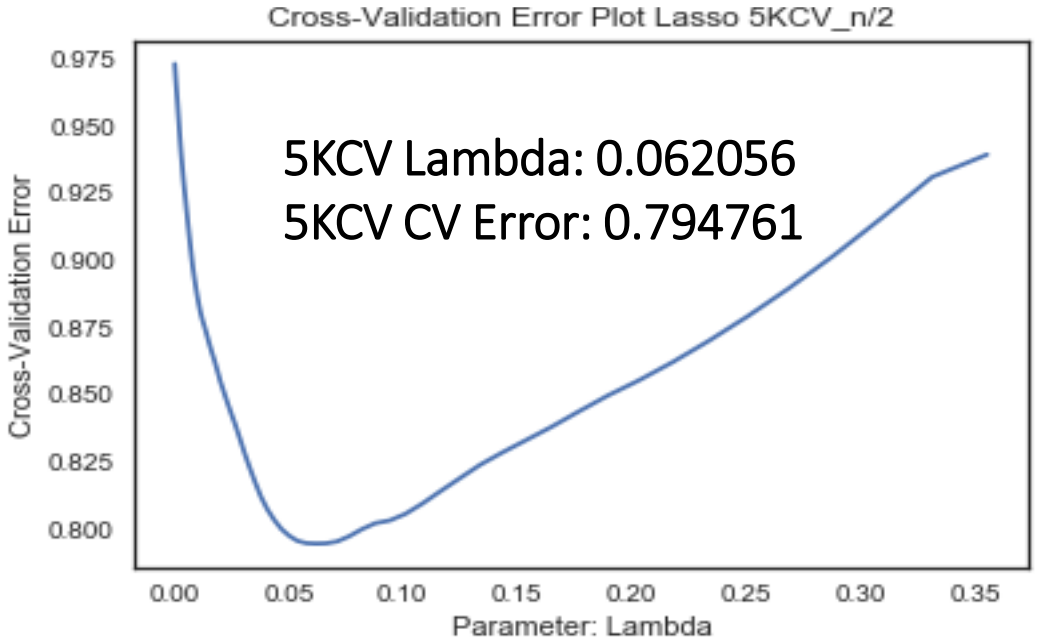
	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.0026	0.042	0.062	0.951	-0.079	0.084
x2	0.0086	0.062	0.138	0.890	-0.114	0.131
x3	0.0023	0.040	0.058	0.954	-0.077	0.081
x4	-0.0408	0.041	-1.001	0.317	-0.121	0.039
x5	-0.0175	0.040	-0.440	0.660	-0.096	0.061
x6	-0.0534	0.041	-1.315	0.189	-0.133	0.026
x7	0.0155	0.038	0.408	0.684	-0.059	0.090
x8	0.1052	0.054	1.944	0.052	-0.001	0.211

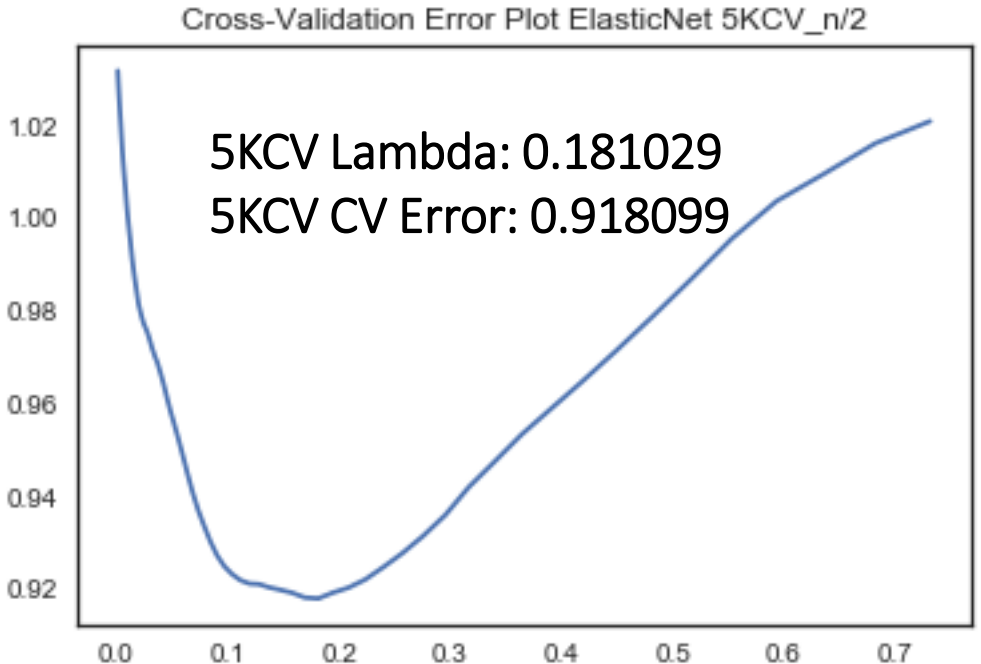

Box plot of R squared error Training_5KCV_n/2
R Squared
0.35
0.30
0.25
0.20
0.15
0.10
Ridge
Lasso
ElasticNet

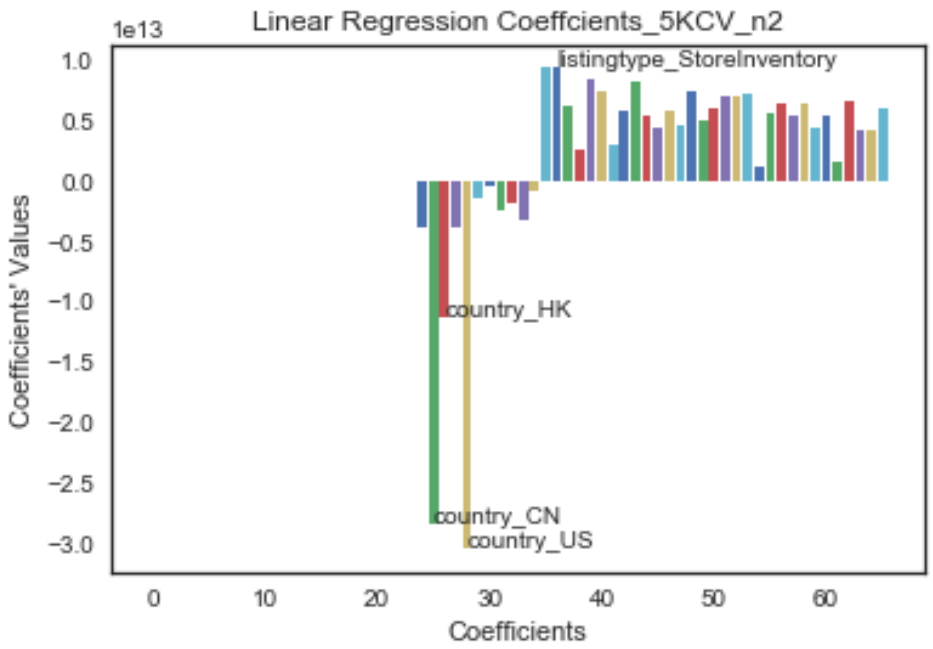

Box plot of R squared error Testing_5KCV_n/2
R Squared
0.20
0.18
0.16
0.14
0.12
0.10
0.08
Ridge
Lasso
ElasticNet

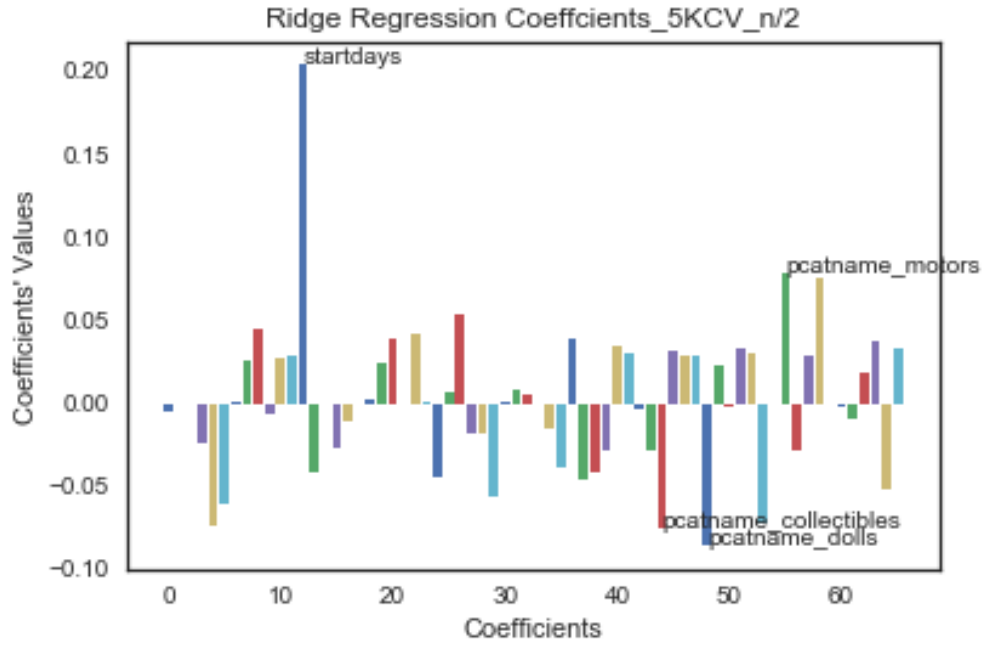
0.384907

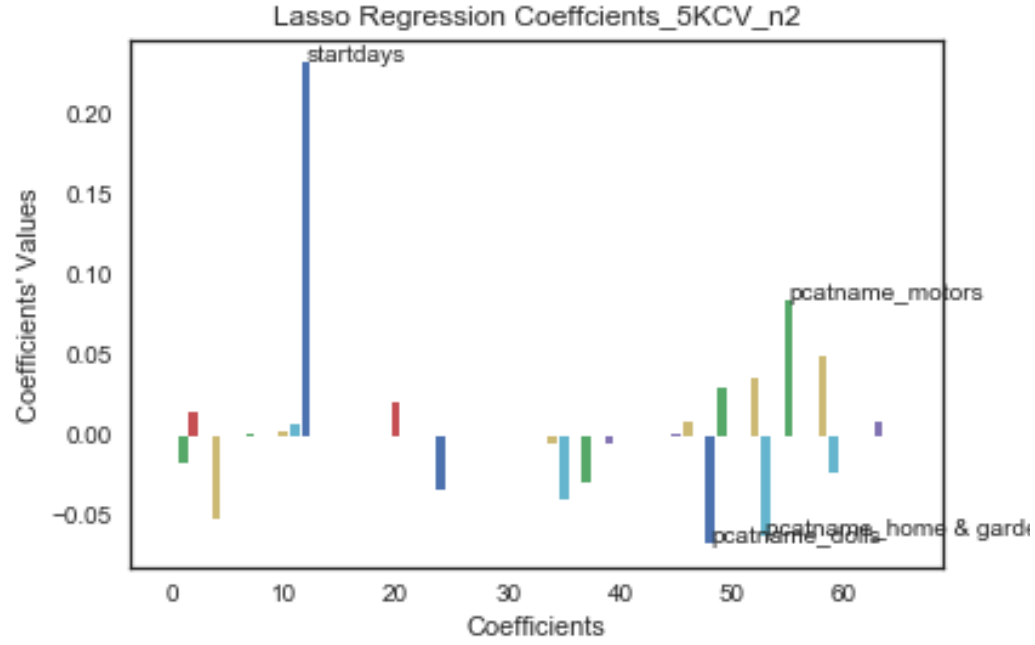

Cross-Validation Error Plot Ridge 5KCV_n/2
Cross-Validation Error
Parameter: Lambda
5KCV Lambda: 0.384907
5KCV CV Error: 0.826143

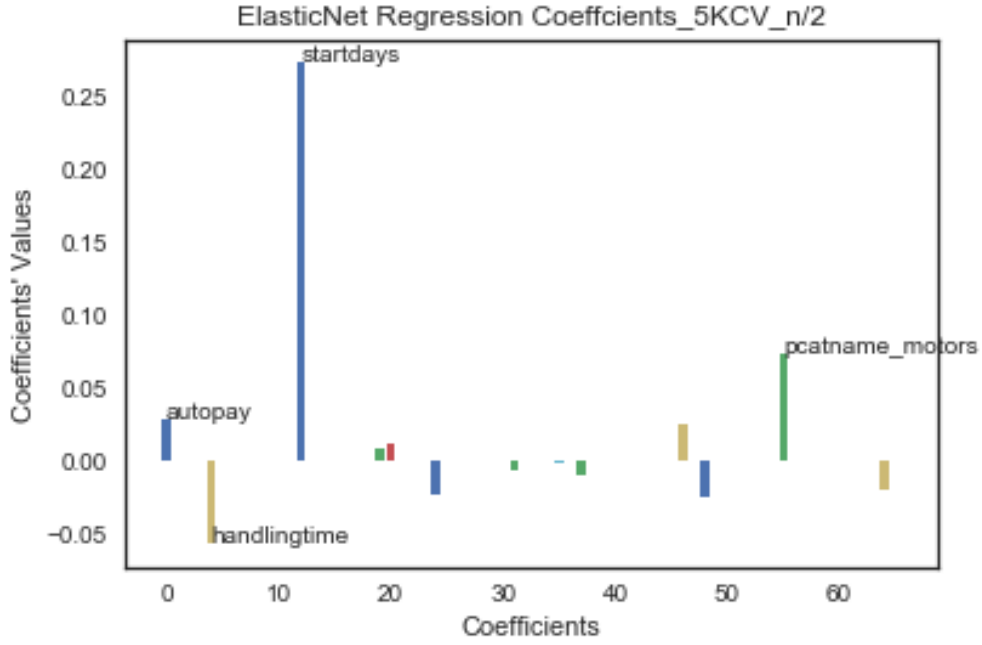

Cross-Validation Error Plot Lasso 5KCV_n/2
Cross-Validation Error
Parameter: Lambda
5KCV Lambda: 0.062056
5KCV CV Error: 0.794761


Cross-Validation Error Plot ElasticNet 5KCV_n/2
5KCV Lambda: 0.181029
5KCV CV Error: 0.918099


Linear Regression Coeffcients_5KCV_n2
1e13
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
Coefficients' Values
listingtype_StoreInventory
country_HK
country_CN
country_US
0
10
20
30
40
50
60
Coefficients


Ridge Regression Coeffcients_5KCV_n/2
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
Coefficients' Values
startdays
pcatname_motors
pcatname_collectibles
pcatname_dolls
0
10
20
30
40
50
60
Coefficients


Lasso Regression Coeffcients_5KCV_n2
0.20
0.15
0.10
0.05
0.00
−0.05
Coefficients' Values
startdays
pcatname_motors
pcatname_dolls
pcatname_home & garde
0
10
20
30
40
50
60
Coefficients


ElasticNet Regression Coeffcients_5KCV_n/2
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
Coefficients' Values
startdays
autopay
handlingtime
pcatname_motors
0
10
20
30
40
50
60
Coefficients

Thanks for watching!