PREDICTING SALES VOLUME ON CRIME DATA AND CONSTRUCTION ACTIVITY

JOHN GUINN
METIS, 2020

QUESTION:

IS IT POSSIBLE, WITH LINEAR REGRESSION, TO PREDICT REAL ESTATE SALES VOLUME IN A PARTICULAR AREA IN MANHATTAN USING THE CRIME STATISTICS CONSTRUCTION ACTIVITY IN THAT AREA?

NO

(ATLEAST NOT WITH ANY PRECISION THAT WOULD PROVIDE USEFUL RESULTS,
OR WITH THE DATA AT HAND.)

DATA ENGINEERING

SALES DATA

SALE PRICE|SALE COUNT	LONGITUDE/LATITUDE

SCRAPED DATA

Property sale in Manhattan per year [2016]
~20,000 rows

PERMIT DATA

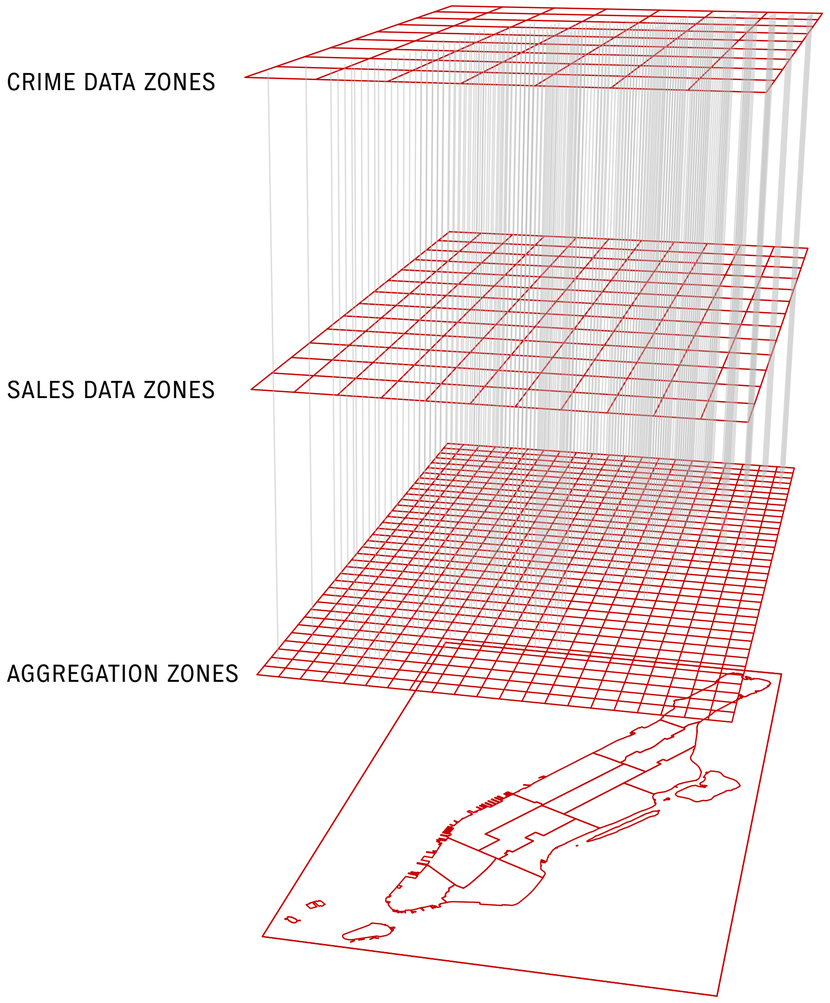
JOB TYPE	LONGITUDE/LATITUDE

Permits pulled for construction in Manhattan [2014-2015]
Categorical, New Building, Alteration, etc.
~150,000 rows

CRIME DATA

CRIME CLASSIFICATION	LONGITUDE/LATITUDE

Felonies, Misdemeanors, and Violations [2014-2015]
~225,000 rows



EARLY DIAGRAM DESCRIBING
ENGINEERING PROCESS

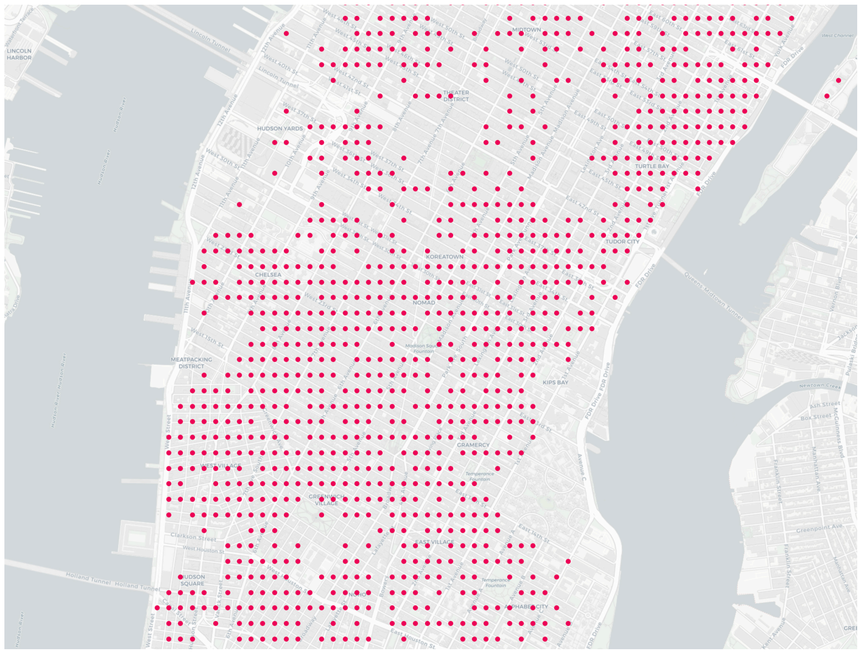
DATA ENGINEERING

Location data must be ‘binned’ in order to aggregate across data sets, despite precieved organization of the raw data as different agencies have different means by which to record their data.

The binning strategy ultimately offered two options.

NYC CRIME LOCATIONS 2014-2015 [AS RECORDED]

APPROXIMATELY 2500 ‘BINS’ -OR- ROWS OF DATA



NYC CRIME LOCATIONS 2014-2015 [STANDARDIZED LOCATION]

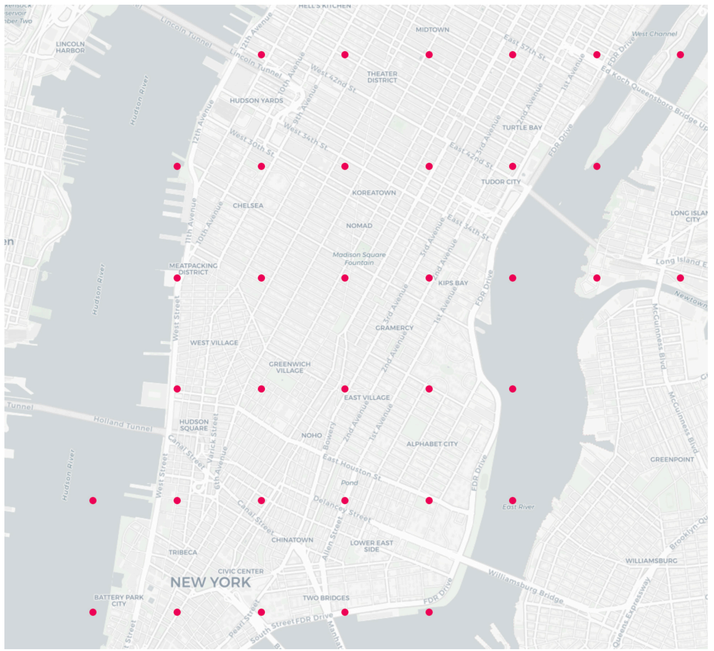
DATA ENGINEERING

Location data must be ‘binned’ in order to aggregate across data sets, despite precieved organization of the raw data as different agencies have different means by which to record their data.

The binning strategy ultimately offered two options.

APPROXIMATELY 100 ‘BINS’ -OR- ROWS OF DATA

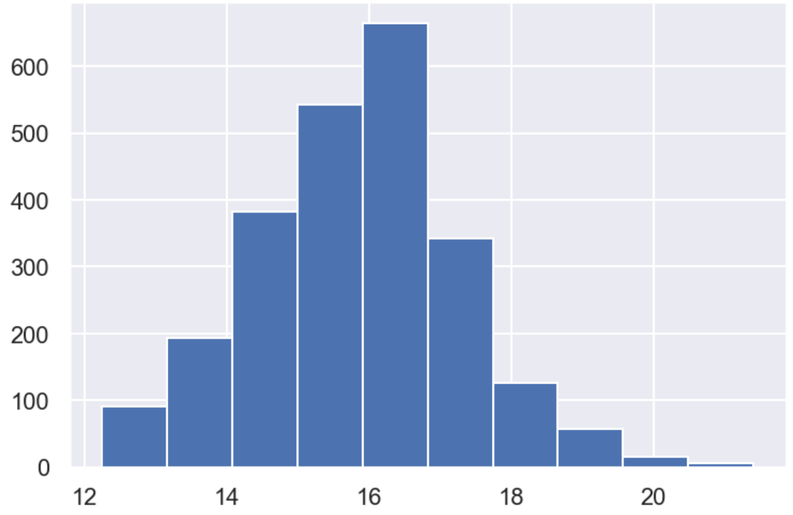
NYC CRIME LOCATIONS 2014-2015 [AS RECORDED]



NYC CRIME LOCATIONS 2014-2015 [STANDARDIZED LOCATION]

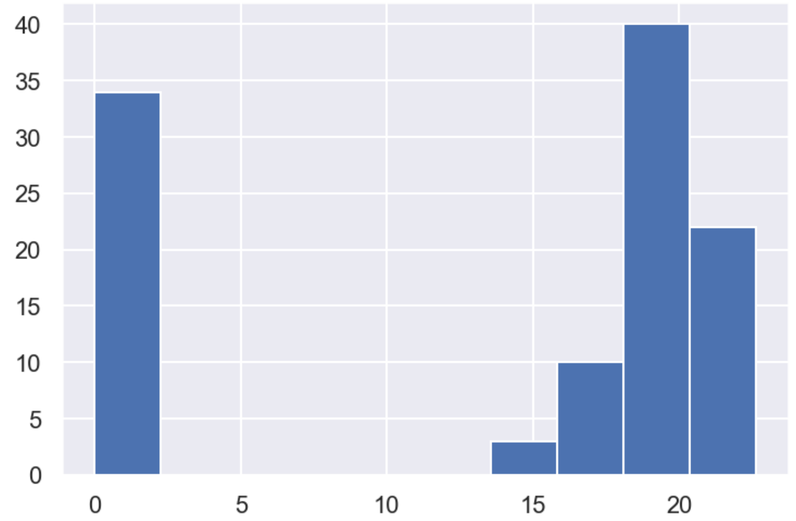
SIMPLE EDA, FEATURE ENGINEERING

~2500 ROWS DISTRIBUTION OF TARGET



AFTER LOG TRANSFORM, AND DROPPING ANY SALE VALUE LESS THAN 100,000

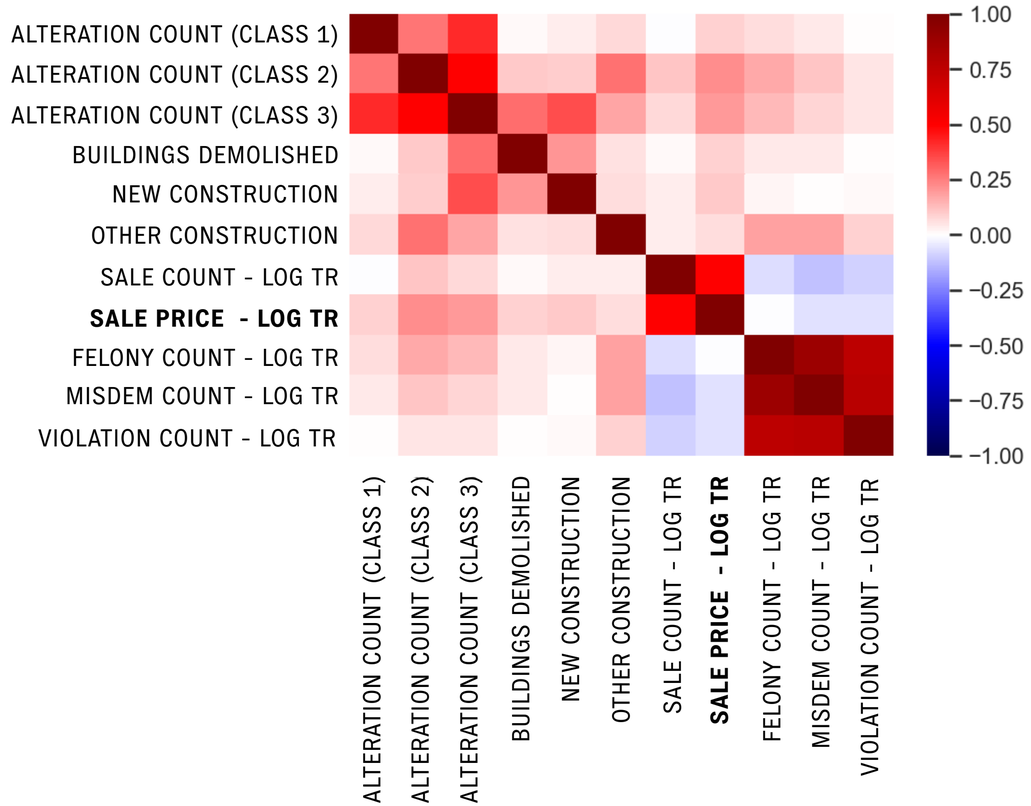
~100 ROWS DISTRIBUTION OF TARGET



AFTER LOG TRANSFORM, UNABLE TO DROP LOW SALE VALUES AND MAINTAIN DECENT SAMPLE SIZE

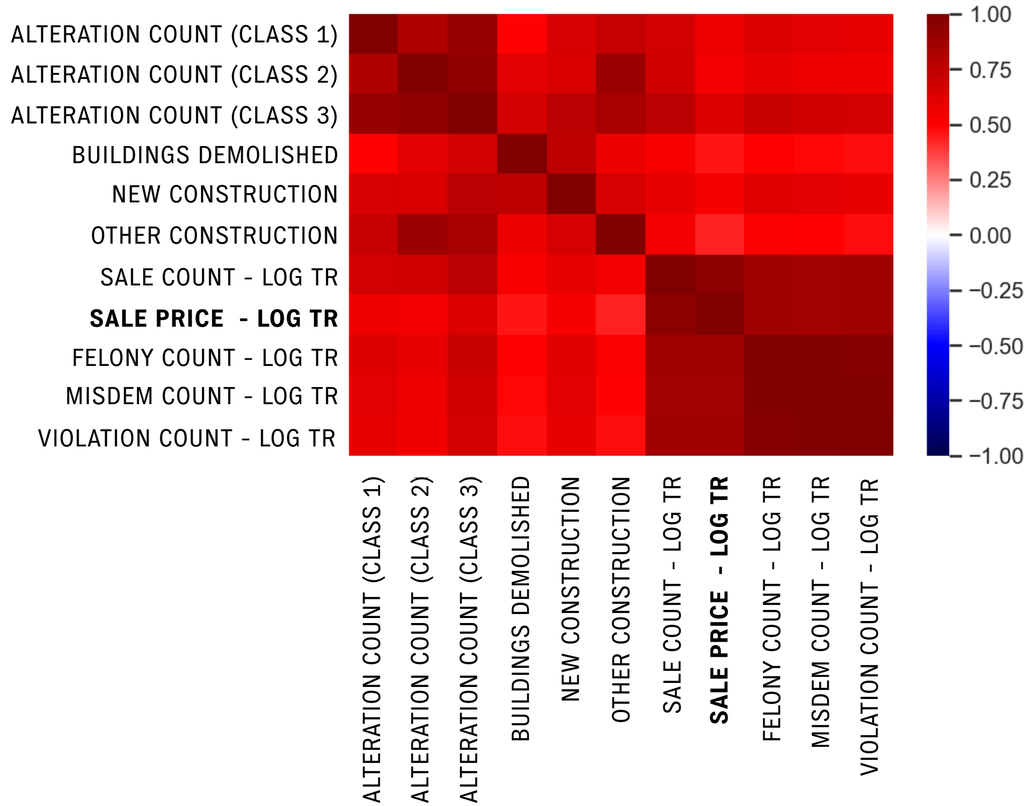
REGRESSION TESTING

~2500 ROWS FORMAT CORRELATION PLOT



AFTER FEATURE ENGINEERING, R^2 : 0.077

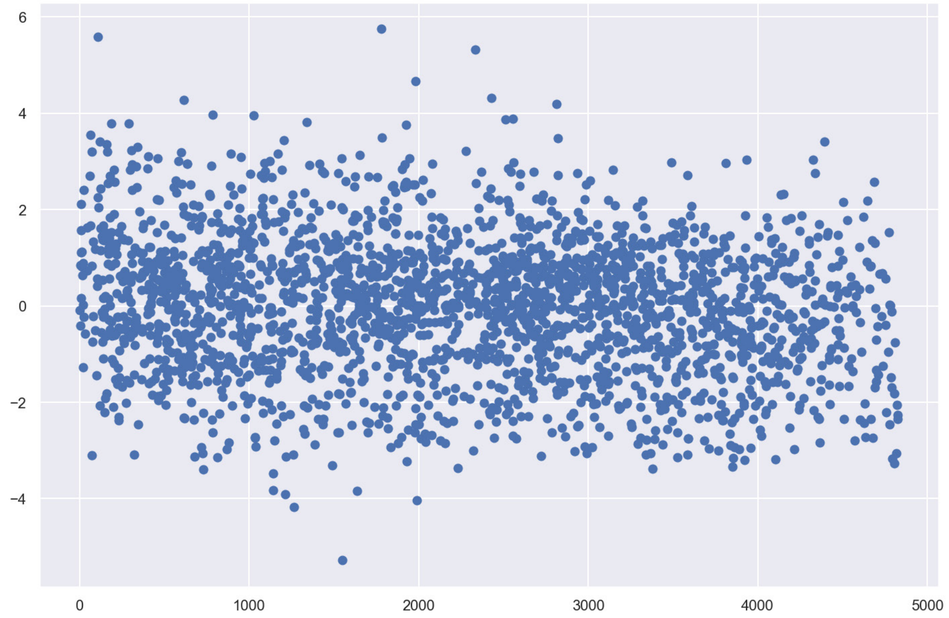
~100 ROWS FORMAT CORRELATION PLOT



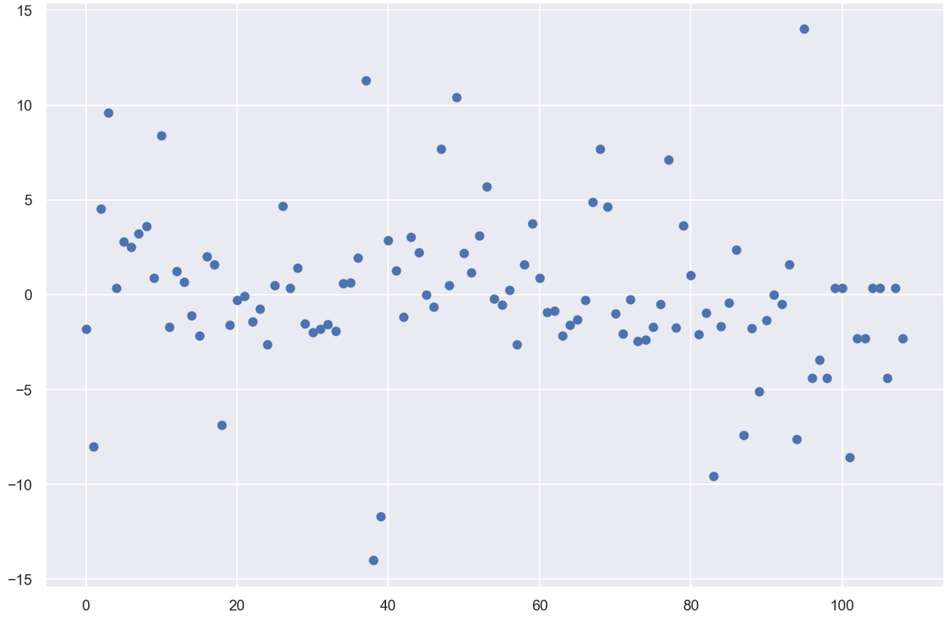
AFTER FEATURE ENGINEERING, R^2 : 0.78

REGRESSION TESTING

~2500 ROWS RESIDUAL PLOT

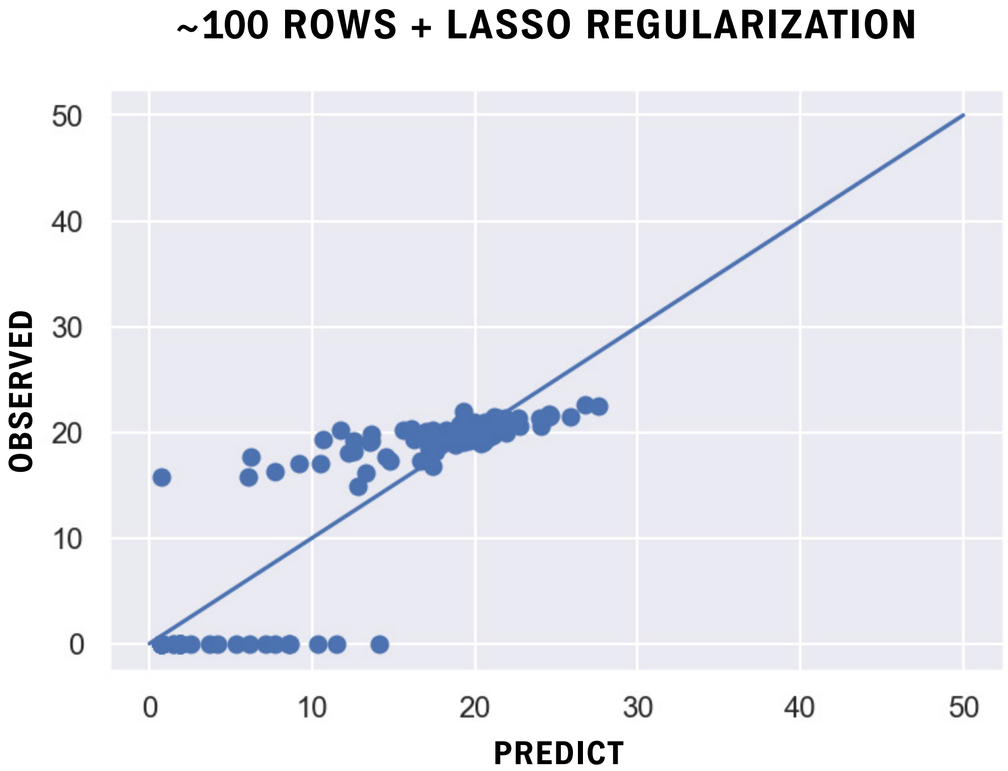


~100 ROWS RESIDUAL PLOT



NOTE Y SCALE CHANGE BETWEEN GRAPHS

REGRESSION TESTING



LASSO ALPHA VAL: 3

R^2 : .76

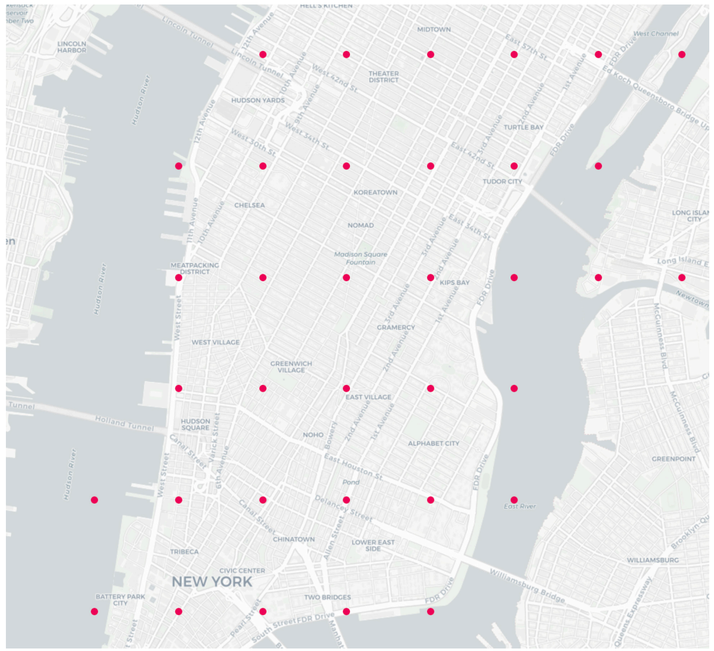
MAE : 3.08

Lasso dropped every feature except Felonies and Misdemeanors. and they are positively correlated.

CONCLUSION

Early regression modeling shows almost no relationship between target variable and features.

Only when data is binned into much less rows, and thereby increasing the detail of each data point does a relationship begin to show, but in so doing the value of the prediction is diminished.



CONCLUSION

Early regression modeling shows almost no relationship between target variable and features.

Only when data is binned into much less rows, and thereby increasing the detail of each data point does a relationship begin to show, but in so doing the value of the prediction is diminished.

FUTURE WORK

1.Explore other models
2. Reconsider feature engineering for smaller data set.

3. Introduce a time factor
4. Find an intermediate ‘binning’ row value

APPENDIX

OLS Regression Results

Dep. Variable:	log_sale_price	**R-squared:**	0.077
Model:	OLS	**Adj. R-squared:**	0.074
Method:	Least Squares	**F-statistic:**	25.13
Date:	Fri, 17 Jul 2020	**Prob (F-statistic):**	1.63e-37
Time:	07:38:33	**Log-Likelihood:**	-4231.9
No. Observations:	2419	**AIC:**	8482.
Df Residuals:	2410	**BIC:**	8534.
Df Model:	8		
Covariance Type:	nonrobust		

click to scroll output; double click to hide

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	15.6997	0.065	240.142	0.000	15.572	15.828
log_felony	0.1919	0.057	3.394	0.001	0.081	0.303
log_misd	-0.2080	0.051	-4.111	0.000	-0.307	-0.109
log_viol	-0.0704	0.049	-1.440	0.150	-0.166	0.025
A1	0.0037	0.010	0.387	0.699	-0.015	0.022
A2	0.0067	0.001	7.372	0.000	0.005	0.008
A3	0.0172	0.006	2.835	0.005	0.005	0.029
NB	0.0282	0.011	2.625	0.009	0.007	0.049
DM	0.0526	0.026	2.020	0.044	0.002	0.104

Omnibus:	4.476	**Durbin-Watson:**	1.860
Prob(Omnibus):	0.107	**Jarque-Bera (JB):**	5.061
Skew:	0.014	**Prob(JB):**	0.0796
Kurtosis:	3.222	**Cond. No.**	122.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

Dep. Variable:	log_sale_price	**R-squared:**	0.787
Model:	OLS	**Adj. R-squared:**	0.770
Method:	Least Squares	**F-statistic:**	46.12
Date:	Fri, 17 Jul 2020	**Prob (F-statistic):**	3.22e-30
Time:	07:44:49	**Log-Likelihood:**	-311.68
No. Observations:	109	**AIC:**	641.4
Df Residuals:	100	**BIC:**	665.6
Df Model:	8		
Covariance Type:	nonrobust		

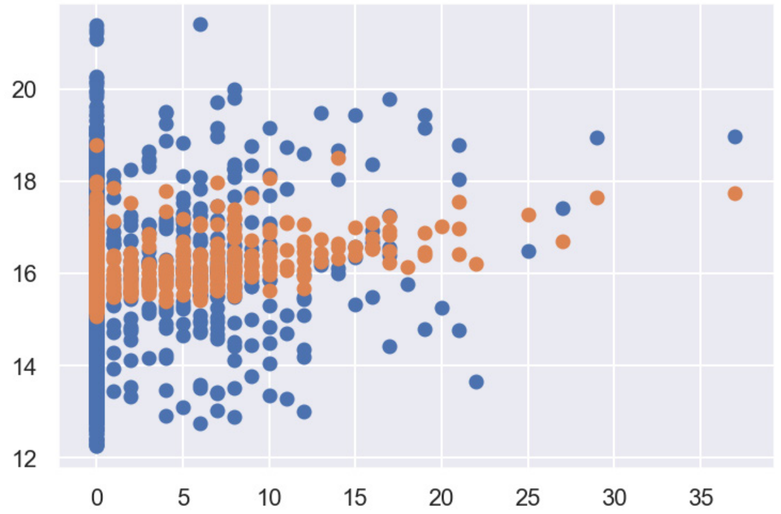
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.7456	1.061	1.646	0.103	-0.359	3.850
log_felony	3.8553	1.131	3.408	0.001	1.611	6.100
log_misd	-3.0927	1.209	-2.559	0.012	-5.490	-0.695
log_viol	2.5281	0.843	2.999	0.003	0.855	4.201
A1	0.0024	0.015	0.166	0.869	-0.027	0.032
A2	0.0006	0.001	0.628	0.532	-0.001	0.002
A3	-0.0041	0.011	-0.393	0.695	-0.025	0.017
NB	-0.0054	0.018	-0.295	0.769	-0.042	0.031
DM	0.0268	0.045	0.594	0.554	-0.063	0.116

Omnibus:	9.490	**Durbin-Watson:**	1.892
Prob(Omnibus):	0.009	**Jarque-Bera (JB):**	22.151
Skew:	0.065	**Prob(JB):**	1.55e-05
Kurtosis:	5.205	**Cond. No.**	7.92e+03

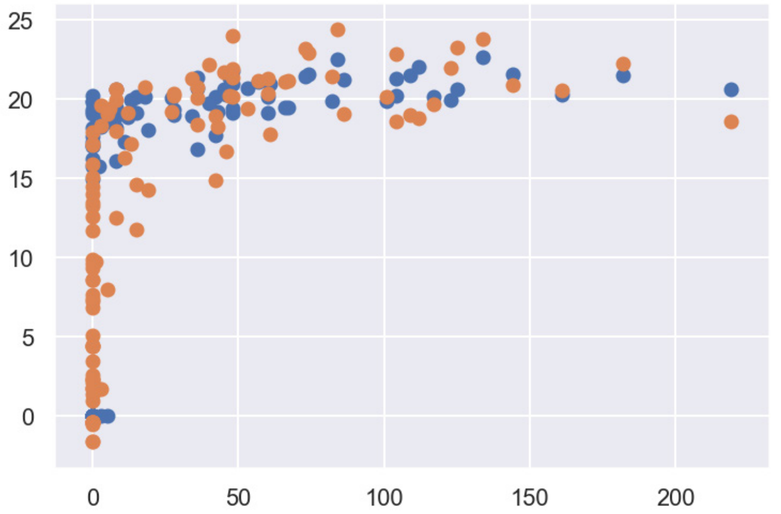
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.92e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

REGRESSION TESTING

~2500 ROWS SIMPLE REGRESSION PLOT AGAINST ONE FEATURE



~100 ROWS SIMPLE REGRESSION PLOT AGAINST ONE FEATURE



ORANGE: PREDICT