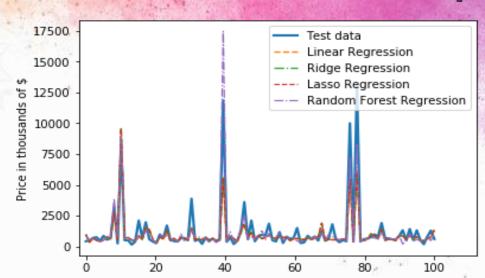
Predicting the price of properties sell in NYC: Using Machine Learning algorithms.

Full code in:

https://github.com/Thaleia18



The prediction task is to determine the price of a property in sale in NYC.

We will create our models using real state data of properties sold in NYC between September 2016 to September 2017.

The machine learning algorithms that I used are:

- Linear regression.
- Ridge regression.
- Lasso regression.
- Random forest regression.

The metrics to evaluate my predictions are:

- R² Coefficient of determination.
- RMSE Root Mean Square Error

THE DATA

This data was extracted from https://www.kaggle.com/new-york-city/nyc-property-sales

Consists on information of properties sold in New York City over a 12-month period from September 2016 to September 2017.

1. BOROUGH 13.	. COMMERCIAL UNIT	S,
----------------	-------------------	----

Attributes. NEIGHBORHOOD 14. TOTAL UNITS

3. BUILDING CLASS CATEGORY, 15. LAND SQUARE FEET

4. TAX CLASS AT PRESENT 16. GROSS SQUARE FEET

5. BLOCK, 17. YEAR BUILT,

6. LOT, 18. TAX CLASS AT TIME OF SALE

7. EASE-MENT 19. BUILDING CLASS AT TIME OF

8. BUILDING CLASS AT PRESENT SALE

9. ADDRESS 20. SALE PRICE,

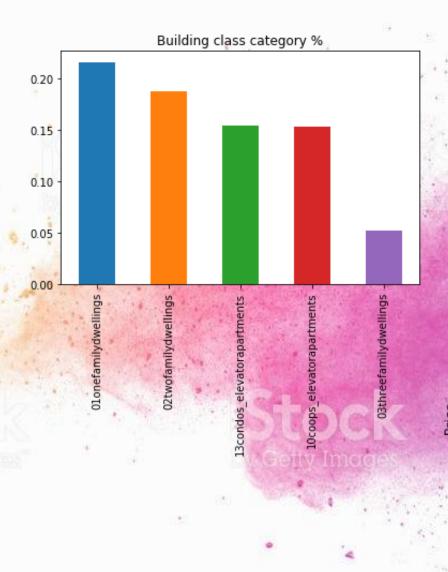
10. APARTMENT NUMBER 21. SALE DATE

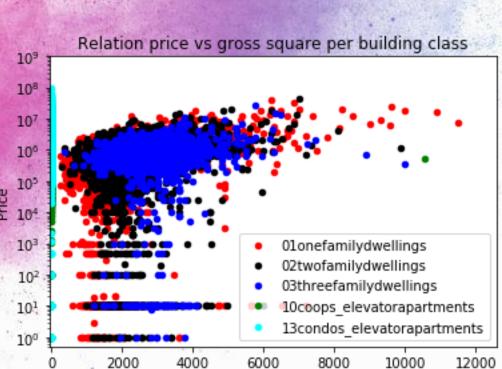
11. ZIP CODE

I cleaned the data: remove duplicate data, clean null data, converted the categorical data in numerical data that I can use in the regression, and normalize the use of strings

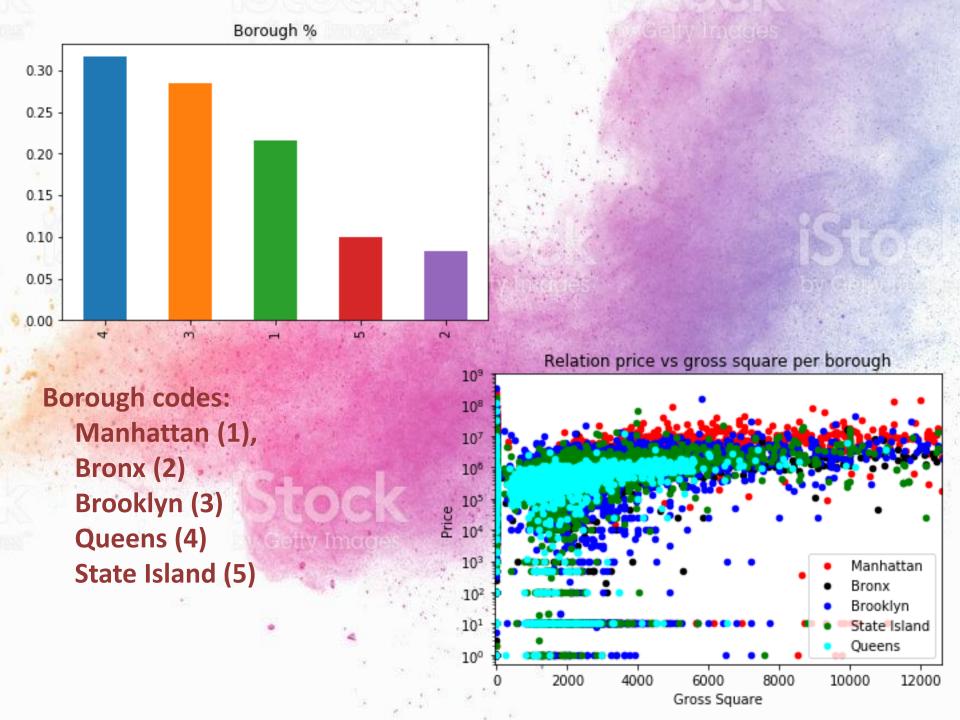
					122	4			1.5	100		CK NIGHT	i = L A	33,000	200		A STATE OF THE PARTY OF THE PAR	
	unnamed:	0 borough	block	lo	ot	zip_co	zip_code ı		al_units	comme	cial_units	total_un	total_units		built			
count	84548.000	000 84548.00	0000 84548.00	00000	4548.000000	84548	3.000000	84548.000000		84548.000000		84548.000000 8		8454	3.0000			
mean	10344.359	10344.359878 2.998758 4237.218976		3976 3	76.224015	10731	.991614	2.025264		0.193559		2.249184 1		1789.	32297		THE P	370
std	7151.7794	36 1.289790	3568.263	3407 6	58.136814	1290.	879147	16.721037		8.713183		18.972584		537.3	44993		100 THE	
min	4.000000	1.000000	1.000000) 1	.000000	0.000000		0.000000		0.000000		0.000000		0.000	000			
25%	4231.0000	00 2.000000	1322.750	0000 2	22.000000		0.000000	0.000000 0.		0.000000		1.000000 192		1920.	.00000			
50%	8942.0000	0 000000	0011 000	2000	0.00000	11000	000000	4 000000		0.00000	·^	4 00000	N	1040	00000	00 (100)00	MANAGEMENT SO	
75%	15987.250	year built	tax class at tin	ne of sale	sale date	pr	ice	land s	quare	aross	square	taxclass1	L	taxclas	ss2	taxclas	ss4	
max	26739.000	84548.000000	84548.000000			+04 8.4	8,454800e+04		800e+04			84548.000000				84548.000000		
		1789.322976	1.657485		2.457811e	+06 1.0	1.056623e+06		793e+03			0.491236 0.43		0.4343	34380 0.074		336	FALLS.
1	537.344993 0.000000 1920.000000 1940.000000		0.819341		1.037445e	1.037445e+02 1.0		038794e+07 3.4909		4 2.881080e+04		0.499926	99926 0.495		6678 0.262		319	
1			1.000000		2.457632e	+06 0.0	0.000000e+00		0.000000e+00 0.		0000e+00 0.0000		0.0000		0.000		000	
100			1.000000	1.000000		+06 0.0	0.000000e+00		0.000000e+00		0.000000e+00		0.000		0.000		000	NULL OF
Selfa!			2.000000		2.457812e	+06 4.	150000e+0	5 1.770	000e+03	1.076	000e+03	0.000000)	0.0000	0.000		000	4
	11	1965.000000		l	0 +	-14	N4		D		Des eldese				0+-+- -		La di dia di di di	.a L
		2017.000000	axclass1	taxclass		class4	Manha		Bronx		Brooklyn		ueens	0000	State_Isl		buildingclass	
100			34548.000000	84548.0		48.00000		.0000000	84548.0		84548.00		4548.00		84548.0		84548.00000	
1).491236	0.43438		74336	0.2165		0.08337		0.284418		.316223		0.09947		0.215676	0.18
	1	NO SERVICE).499926			52319	0.4118					51140 0.465004					0.411293	0.390
		The state of the s	0.000000	0.00000		00000	0.0000		0.00000		0.000000		.000000		0.00000	_	0.000000	0.000
	TV, TV, THEETER		0.000000	0.00000		00000	0.000000		0.00000		0.000000						0.000000	0.000
			0.000000	0.00000			0.0000		0.00000		0.000000		0.000000		0.000000		0.000000	0.000
		3	1.000000	1.00000		0.000000		000	0.00000								0.000000	0.000
			1.000000	1.00000	00 1.0	00000	1.0000	000	1.00000	00	1.000000) 1	.000000)	1.00000	0	1.000000	1.000

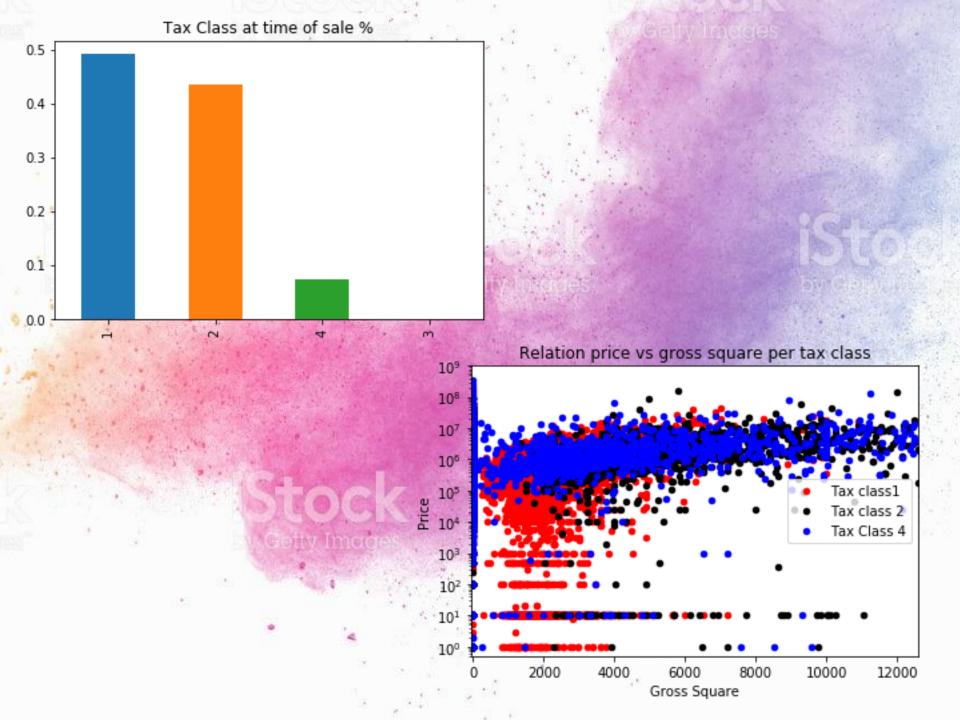
SOME DATA VISUALIZATIONS.





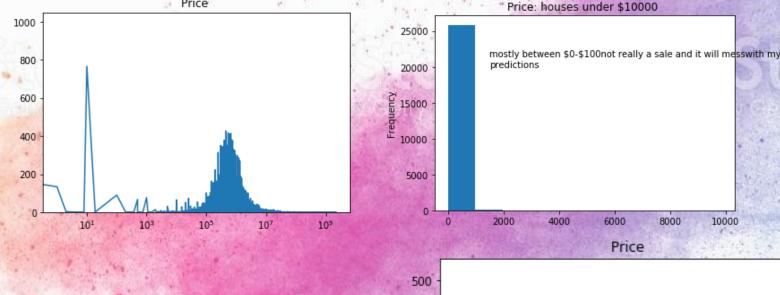
Gross Square



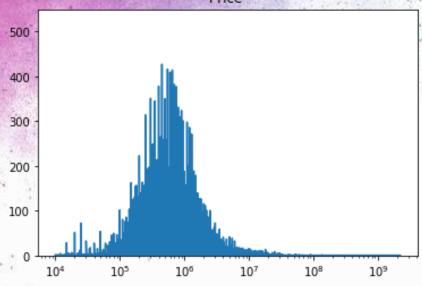


I plotted the frequency of the different sale prices for the properties sold.

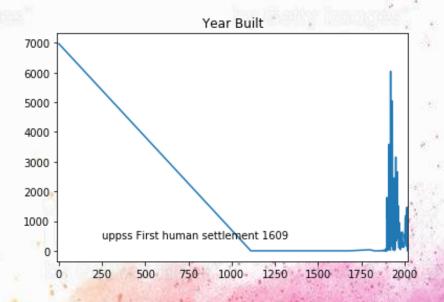
A lot of the houses where sold for low prices, under \$10,000. These were not sale, just transfers between family members.



I used data with prices over \$10,000. I consider noise that will mess my models all the data related with transfers.



10000

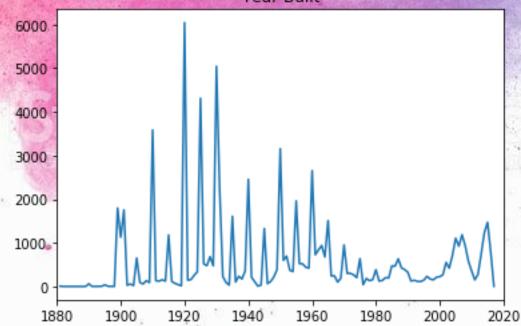


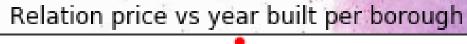
I have data about the year built for each property sold.

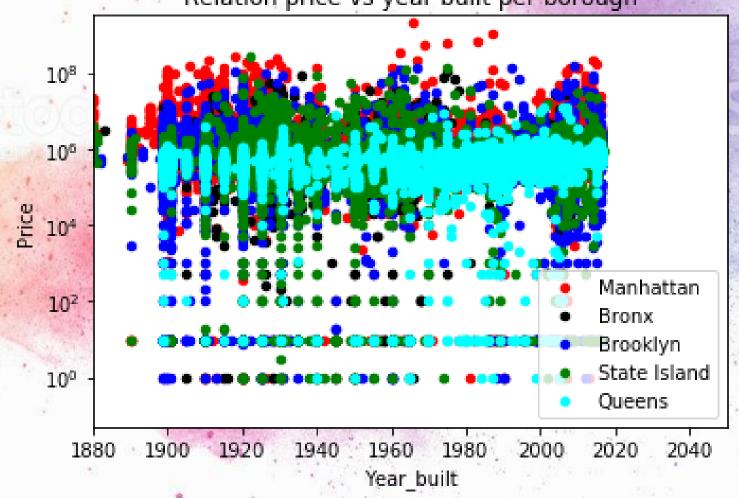
I plotted the frequency of houses built every year.

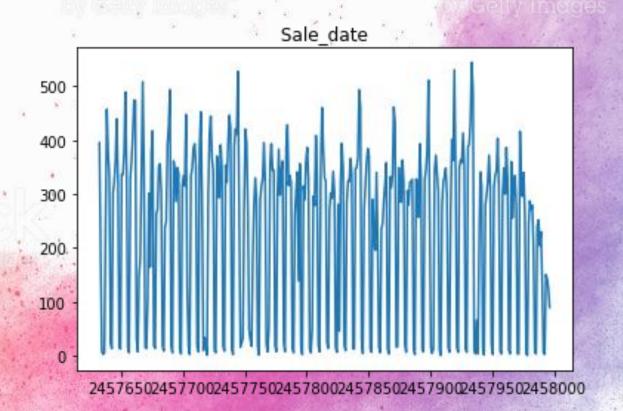
We have some data that says that the properties were built in year 0, this seems wrong.

I removed all the data with houses built before 1880.









I transform the date to Julian date. I plot the frequencies to see if in some point of the year more houses were sold.

Gelly Imoges

I didn't find any pattern I decided no to use the sale date in my regression.

		_		75: 340	1.54	ATTER TO SELECT	The Landson			
	land_square	gross_square	year_built	taxclass1	taxclass2	taxclass4	Manhattan	Queens		
count	2.787200e+04	2.787200e+04	27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	27872.00000	96	PROPERTY
mean	4.263921e+03	4.350186e+03	1940.795709	0.871053	0.080547	0.048400	0.031896	0.380848		
std	3.847117e+04	3.385798e+04	30.290694	0.335147	0.272143	0.214613	0.175726	0.485604		
min	2.000000e+02	1.200000e+02	1881.000000	0.000000	0.000000	0.000000	0.000000	0.000000		
25%	2.000000e+03	Queens	Brooklyn	Bronx	State_Island	buildingclass1	buildingclass2	! buildingclass	3 buildingclass10	r€
50%	2.500000e+03	27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	27872.00000	27872.000000	2
75%	4.000000e+03	0.380848	0.292516	0.120336	0.174404	0.446254	0.343427	0.080654	0.000897	3.
max	4.228300e+06	0.485604	0.454926	0.325360	0.379464	0.497112	0.474861	0.272309	0.029936	2
4		0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
-		0.000000	0.000000	0.000	huilding alaast	n nandantial	unita aanan	o cocco	total unita	nrice
+ 4		0.000000	0.000000	0.000 lass3	buildingclass1			_	total_units	price
-		1.000000	1.000000 1.000000 0.000		27872.000000	27872.000	0000 27872	2.000000	27872.000000	2.787200e+04
		1.000000	1.000000		0.000897	3.013634	0.334	493	3.346656	1.718361e+06
	*	4			0.029936	20.124790	14.37	79 <mark>4</mark> 2	24.866141	1.742905e+07
					0.000000	0.000000	0.000	000	0.000000	1.007000e+05
					0.000000	1.000000	0.000	000	1.000000	4.490000e+05
			۰	7	0.000000	2.000000	0.000	000	2.000000	6.408500e+05
				-	0.000000	2.000000	0.000	000	2.000000	9.750000e+05
					1.000000	1844.0000	000 2261.	000000	2261.000000	2.210000e+09

The correlation of these features with the sale price:

																		750		
							F	earso	n Corr	elatio	n of Fe	eature	s	9	26	4			100	
									100	3							-			
land_square -	1	0.44	0.082	-0.25	0.078	0.28	0.023	0.085	-0.21	0.0068	0.13	-0.01	-0.1	-0.1	0.14		0.047	0.24		- 1
gross_square -	0.44	1	-0.031	-0.63	0.49	0.36	0.39	-0.19	0.12	0.095	-0.16	-0.46	-0.026	0.12	0.17		0.069	0.34	0.69	
year_built -	0.082	-0.031	1	0.095	-0.13	0.017	-0.14	-0.026	-0.25	-0.02	0.41	0.13	-0.053	-0.027	0.0086	-0.003	0.00083	-0.0019	-0.12	
taxclass1 -	-0.25	-0.63	0.095	1	-0.77	-0.59	-0.35	0.13	-0.092	-0.034	0.14	0.35	0.28	0.11	-0.078	-0.19	-0.056	-0.18	-0.52	
taxclass2 -	0.078	0.49	-0.13	-0.77	1	-0.067	0.32	-0.13	0.095	0.031	-0.12	-0.27	-0.21	-0.088	0.1	0.24	0.004	0.2	0.38	
taxclass4 -	0.28	0.36	0.017	-0.59	-0.067	1	0.15	-0.043	0.024	0.014	-0.056	-0.2	-0.16	-0.067	-0.0068	-0.019	0.082	0.032	0.34	
Manhattan -	0.023	0.39	-0.14	-0.35	0.32	0.15	1	-0.14	-0.12	-0.067	-0.083	-0.13	-0.096	-0.027	0.022	0.15	0.025	0.14		
Queens -	0.085	-0.19	-0.026	0.13	-0.13	-0.043	-0.14	1	-0.5	-0.29	-0.36	0.14	-0.022	-0.063	0.0086	-0.031	0.0051	-0.022	-0.11	
Brooklyn -	-0.21	0.12	-0.25	-0.092	0.095	0.024	-0.12	-0.5	1	-0.24	-0.3	-0.22	0.1	0.11	-0.0087	-0.0097	-0.0064	-0.012	0.2	
Bronx -	0.0068	0.095	-0.02	-0.034	0.031	0.014	-0.067	-0.29	-0.24	1	-0.17	-0.11	0.039	0.092	-3.1e-05	0.024	-0.0035	0.017	-0.12	
State Island -	0.13	-0.16	0.41	0.14	-0.12	-0.056	-0.083	-0.36	-0.3	-0.17	1	0.24	-0.083	-0.12	-0.011	-0.038	-0.0076	-0.035	-0.22	
buildingclass1 -	-0.01	-0.46	0.13	0.35	-0.27	-0.2	-0.13	0.14	-0.22	-0.11	0.24	1	-0.65	-0.27	-0.027	-0.09	-0.02	-0.084	-0.31	
buildingclass2 -	-0.1	-0.026	-0.053	0.28	-0.21	-0.16	-0.096	-0.022	0.1	0.039	-0.083	-0.65	1	-0.21	-0.022	-0.036	-0.015	-0.038	-0.069	
buildingclass3 -	-0.1	0.12	-0.027	0.11	-0.088	-0.067	-0.027	-0.063	0.11	0.092	-0.12	-0.27	-0.21	1	-0.0089	-0.00021	-0.0069	-0.0041	0.045	
buildingclass10 -	0.14	0.17	0.0086	-0.078	0.1	-0.0068	0.022	0.0086	-0.0087	-3.1e-05	-0.011	-0.027	-0.022	-0.0089	1	0.34	0.0021	0.27	-0.025	
residential units -	0.26	0.37	-0.003	-0.19	0.24	-0.019	0.15	-0.031	-0.0097	0.024	-0.038	-0.09	-0.036	-0.00021	0.34	1	0.011	0.82	0.19	
	0.047	0.069	0.00083		0.004	0.082	0.025	0.0051	-0.0064	-0.0035	-0.0076	-0.02	-0.015	-0.0069	0.0021	0.011	1	0.59	0.043	
ommercial_units -																				
total_units -		0.34	-0.0019	-0.18	0.2	0.032	0.14	-0.022	-0.012	0.017	-0.035	-0.084	-0.038	-0.0041	0.27	0.82	0.59	1	0.18	
price -	0.29	0.69	-0.12	-0.52	0.38	0.34	0.48	-0.11	0.2	-0.12	-0.22	-0.31	-0.069	0.045	-0.025	0.19	0.043	0.18	1	
	land_square -	gross_square -	year built -	taxclass1 -	taxclass2 -	taxclass4 -	Manhattan -	Oneens	Brooklyn -	Bronx -	State Island -	buildingclass1 -	buildingclass2 -	buildingclass3 -	buildingclass10 -	residential units -	ommercial_units -	total units	price	

Ì											
	price	1.000000									
	gross_square	0.689004									
	Manhattan	0.475596									
Š	taxclass2	0.376039									
Š	taxclass4	0.335969									
	land_square	0.292000									
į	Brooklyn	0.200388									
0	residential_units	0.192468									
a de	total_units	0.180459									
ì	buildingclass3	0.045443									
	commercial_units	0.042755									
	buildingclass10	-0.025421									
į	buildingclass2	-0.068948									
	Queens	-0.109829									
	Bronx	-0.120081									
	year_built	-0.123245									
	State_Island	-0.216972									
١	buildingclass1	-0.309559									
	taxclass1	-0.520487									
	Name: price, dtype:	float64									

MODEL RESULTS

Linear Regression.

r2 0.6111805534821673

rmse 0.5340183237978753

Ridge regression.

r2 0.6111805532023231

rmse 0.5340183239900492

Lasso Regression.

r2 0.6113140862445937

rmse 0.5339266166236538

iSto

Getty links ones

Random forest regression.

r2 0.6233097547347948

rmse 0.5256229959990711

The best results are for the



