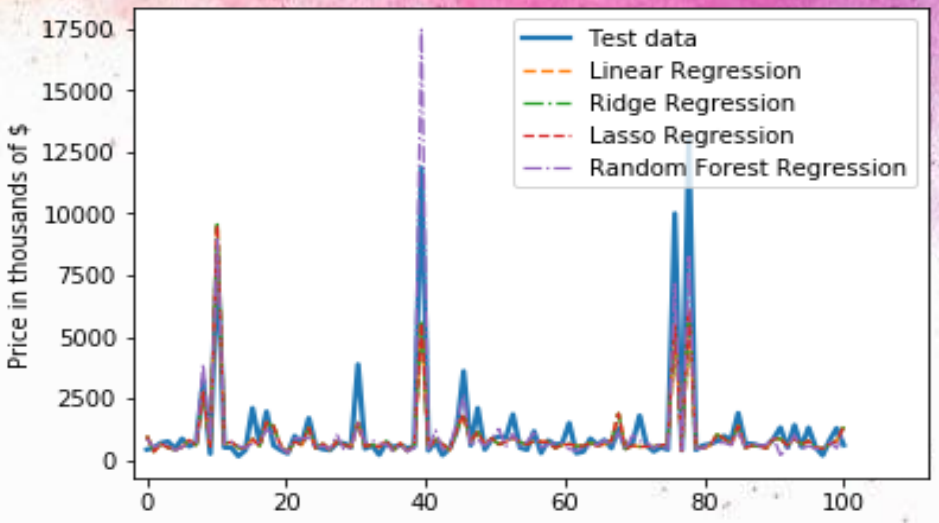
Predicting the price of properties sell in NYC:

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:

- R^2 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.

Attributes:

1. BOROUGH
2. NEIGHBORHOOD
3. BUILDING CLASS CATEGORY,
4. TAX CLASS AT PRESENT
5. BLOCK,
6. LOT,
7. EASE-MENT
8. BUILDING CLASS AT PRESENT
9. ADDRESS
10. APARTMENT NUMBER
11. ZIP CODE
13. COMMERCIAL UNITS,
14. TOTAL UNITS
15. LAND SQUARE FEET
16. GROSS SQUARE FEET
17. YEAR BUILT,
18. TAX CLASS AT TIME OF SALE
19. BUILDING CLASS AT TIME OF SALE
20. SALE PRICE,
21. SALE DATE

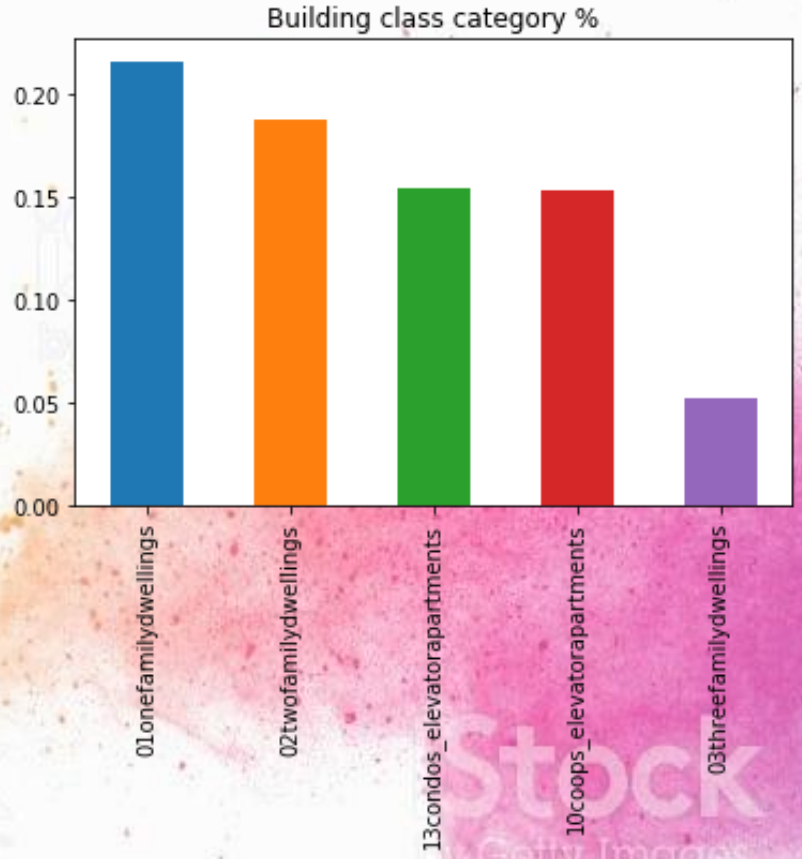
I cleaned the data: remove duplicate data, clean null data, converted the categorical data in numerical data that I can use in the regression, and normalize the use of strings

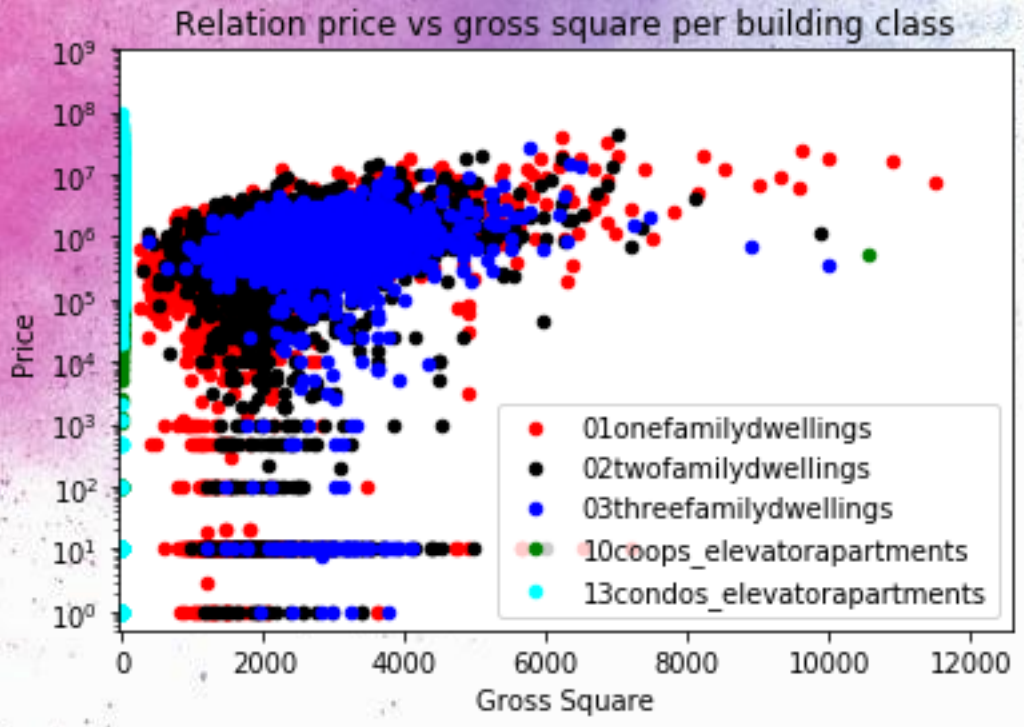
	unnamed: 0	borough	block	lot	zip_code	residential_units	commercial_units	total_units	year_built
count	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.0000
mean	10344.359878	2.998758	4237.218976	376.224015	10731.991614	2.025264	0.193559	2.249184	1789.32297
std	7151.779436	1.289790	3568.263407	658.136814	1290.879147	16.721037	8.713183	18.972584	537.344993
min	4.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4231.000000	2.000000	1322.750000	22.000000	10305.000000	0.000000	0.000000	1.000000	1920.00000
50%	8942.0000								
75%	15987.250								
max	26739.000								

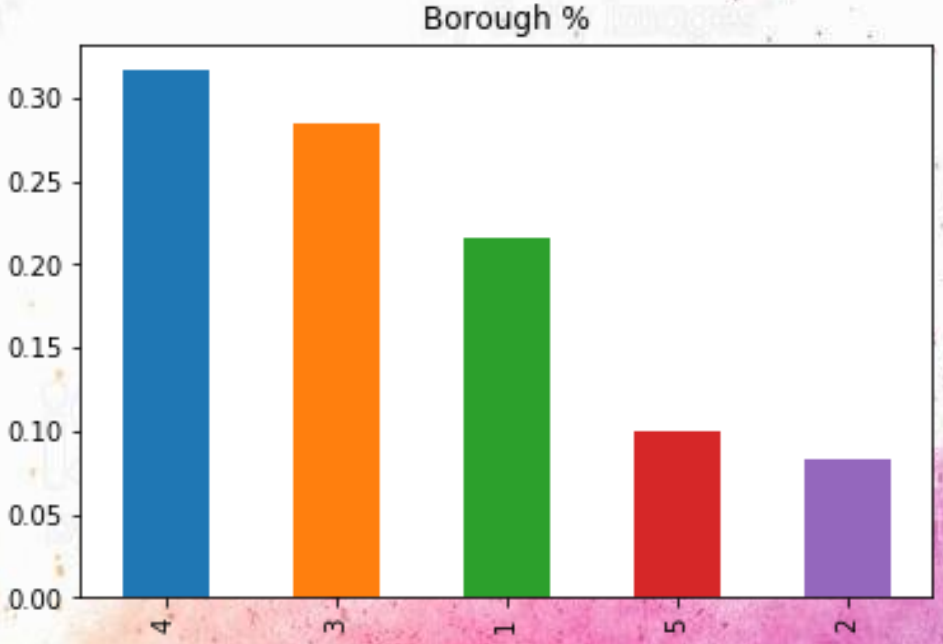
year_built	tax_class_at_time_of_sale	sale_date	price	land_square	gross_square	taxclass1	taxclass2	taxclass4
84548.000000	84548.000000	8.454800e+04	8.454800e+04	8.454800e+04	8.454800e+04	84548.000000	84548.000000	84548.000000
1789.322976	1.657485	2.457811e+06	1.056623e+06	2.717793e+03	2.724445e+03	0.491236	0.434380	0.074336
537.344993	0.819341	1.037445e+02	1.038794e+07	3.490950e+04	2.881080e+04	0.499926	0.495678	0.262319
0.000000	1.000000	2.457632e+06	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000
1920.000000	1.000000	2.457722e+06	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000
1940.000000	2.000000	2.457812e+06	4.150000e+05	1.770000e+03	1.076000e+03	0.000000	0.000000	0.000000
1965.000000								
2017.000000								

axclass1	taxclass2	taxclass4	Manhattan	Bronx	Brooklyn	Queens	State_Island	buildingclass1	buildin
84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.000000	84548.
.491236	0.434380	0.074336	0.216516	0.083373	0.284418	0.316223	0.099470	0.215676	0.1872
.499926	0.495678	0.262319	0.411872	0.276446	0.451140	0.465004	0.299294	0.411293	0.3900
.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
.000000	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.0000
.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.0000

SOME DATA VISUALIZATIONS.

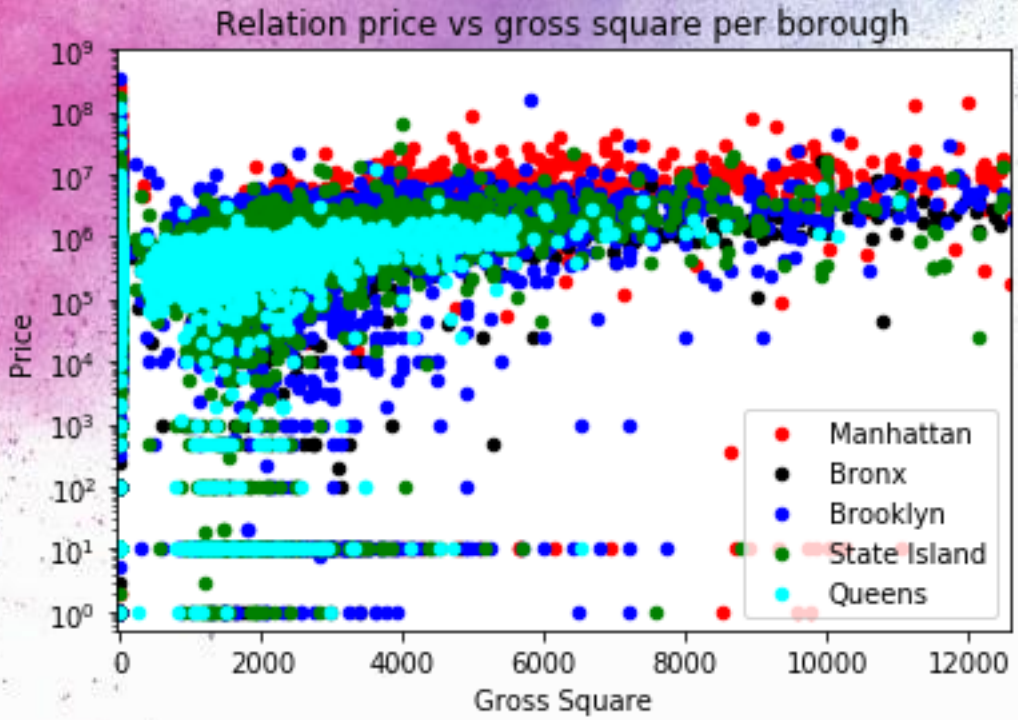




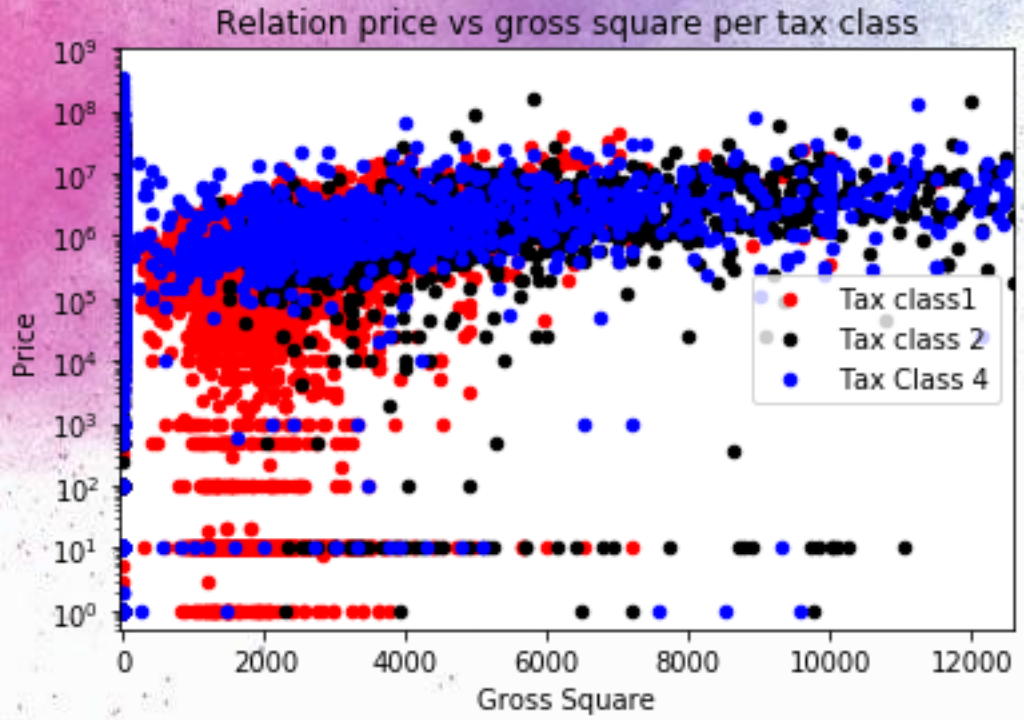


Borough codes:

- **Manhattan (1),**
- **Bronx (2)**
- **Brooklyn (3)**
- **Queens (4)**
- **State Island (5)**

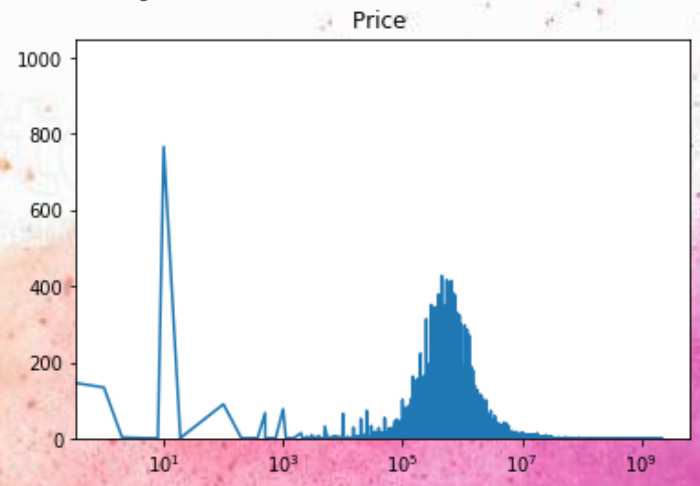


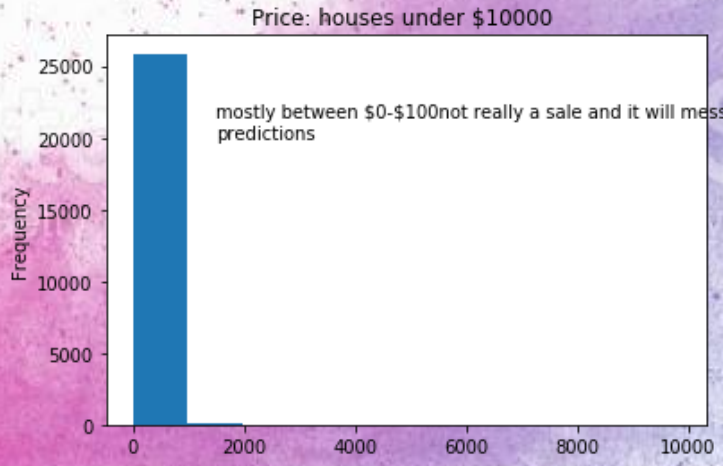

Tax Class at time of sale %
0.5
0.4
0.3
0.2
0.1
0.0
1
2
4
3


Relation price vs gross square per tax class
Price
Gross Square
Tax class1
Tax class 2
Tax Class 4
0
2000
4000
6000
8000
10000
12000

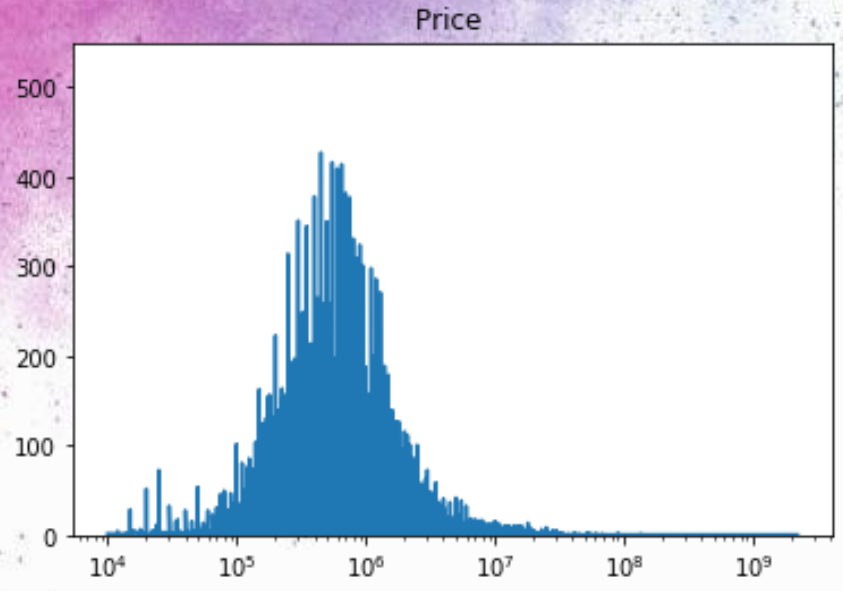
I plotted the frequency of the different sale prices for the properties sold.

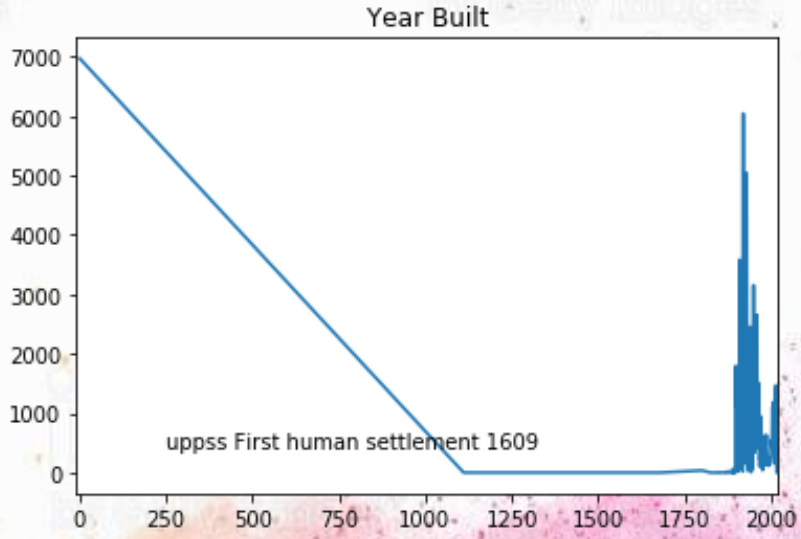
A lot of the houses where sold for low prices, under $10,000. These were not sale, just transfers between family members.





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



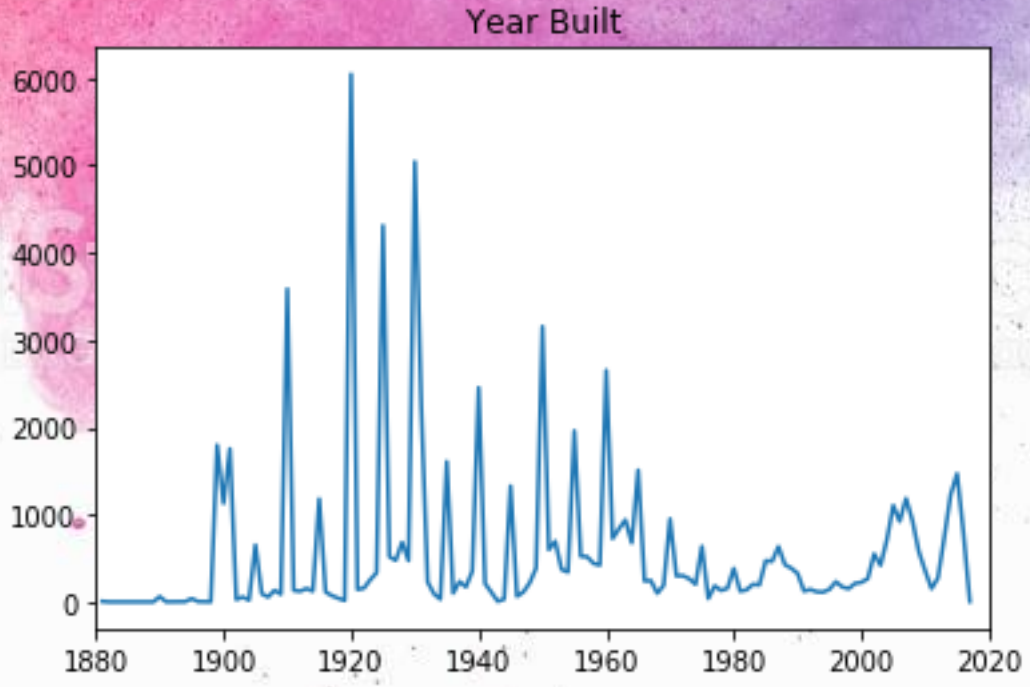


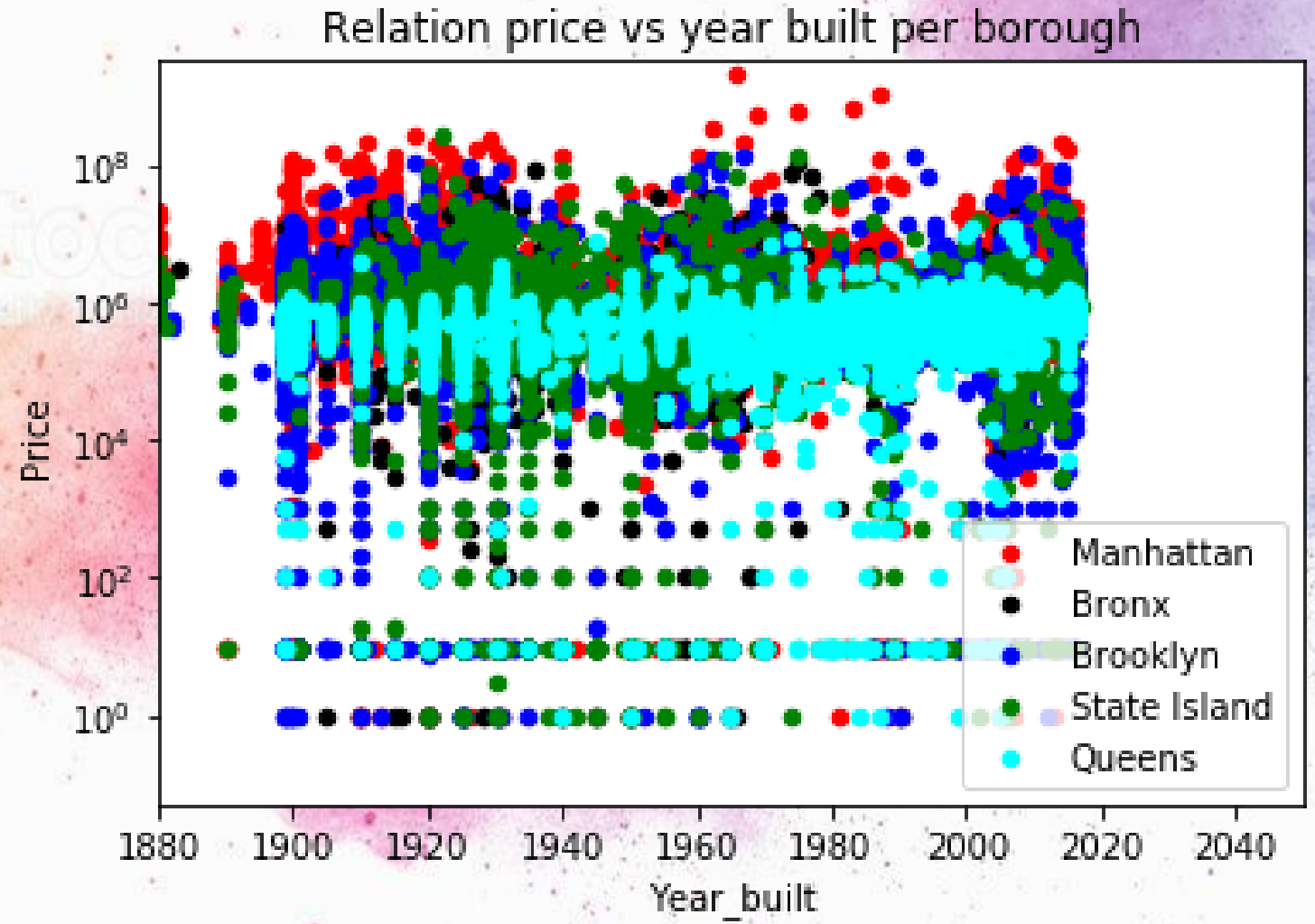
I have data about the year built for each property sold.

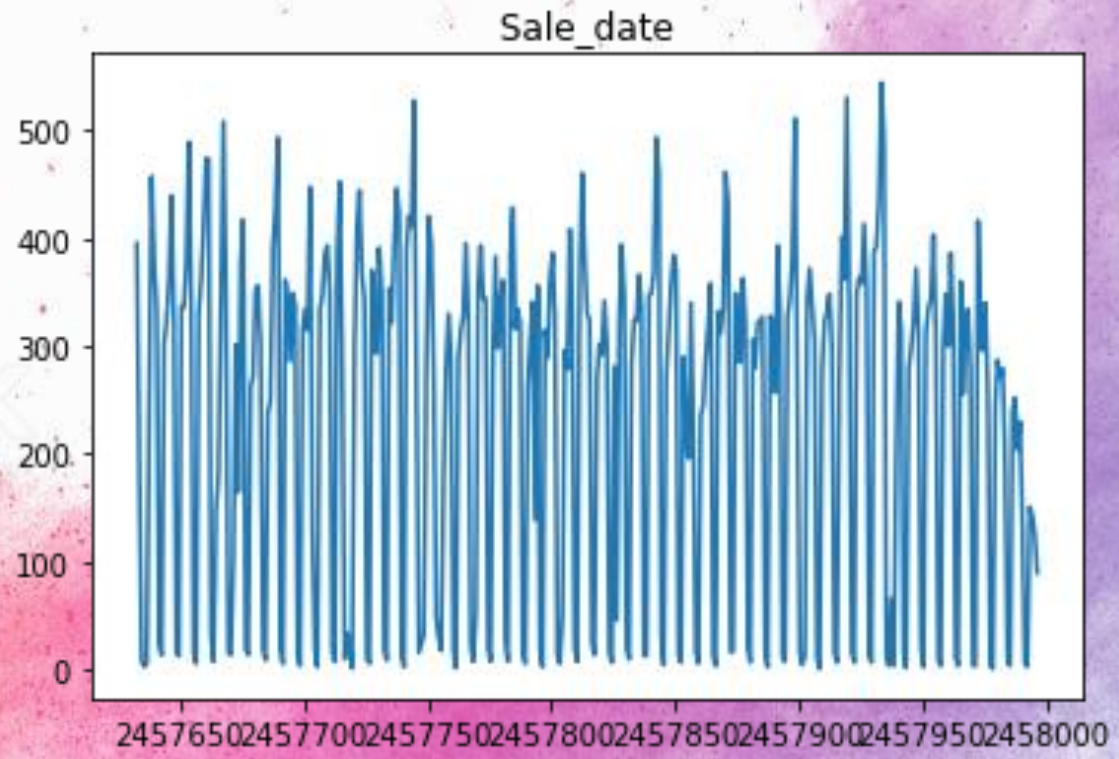
I plotted the frequency of houses built every year.

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.




Relation price vs year built per borough
Price
10^8
10^6
10^4
10^2
10^0
1880
1900
1920
1940
1960
1980
2000
2020
2040
Year_built
Manhattan
Bronx
Brooklyn
State Island
Queens



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

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

So finally, the features that I decided to use in my models:

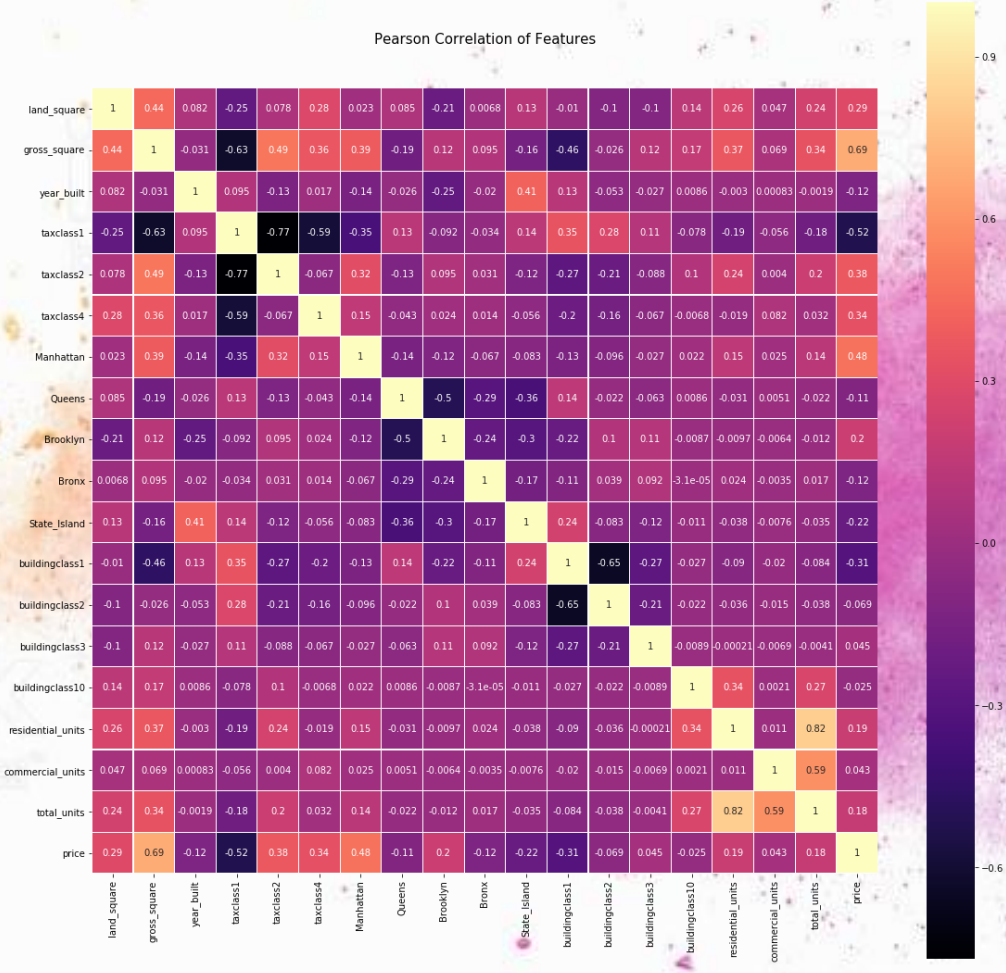
'land_square' 'gross_square', 'year_built', 'taxclass1', 'taxclass2', 'taxclass4', 'Manhattan','Queens','Brooklyn','Bronx','State_Island', 'buildingclass1','buildingclass2','buildingclass3','buildingclass10', 'residential_units', 'commercial_units', 'total_units'.

	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	27872.0000
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							
50%	2.500000e+03							
75%	4.000000e+03							
max	4.228300e+06							

Queens	Brooklyn	Bronx	State_Island	buildingclass1	buildingclass2	buildingclass3	buildingclass10	re
27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	27872.000000	2
0.380848	0.292516	0.120336	0.174404	0.446254	0.343427	0.080654	0.000897	3.
0.485604	0.454926	0.325360	0.379464	0.497112	0.474861	0.272309	0.029936	2(
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
0.000000	0.000000	0.000						
0.000000	0.000000	0.000						
1.000000	1.000000	0.000						
1.000000	1.000000	1.000						

lass3	buildingclass10	residential_units	commercial_units	total_units	price
0000	27872.000000	27872.000000	27872.000000	27872.000000	2.787200e+04
	0.000897	3.013634	0.334493	3.346656	1.718361e+06
	0.029936	20.124790	14.377942	24.866141	1.742905e+07
	0.000000	0.000000	0.000000	0.000000	1.007000e+05
	0.000000	1.000000	0.000000	1.000000	4.490000e+05
	0.000000	2.000000	0.000000	2.000000	6.408500e+05
	0.000000	2.000000	0.000000	2.000000	9.750000e+05
	1.000000	1844.000000	2261.000000	2261.000000	2.210000e+09

The correlation of these features with the sale price:



price	1.000000
gross_square	0.689004
Manhattan	0.475596
taxclass2	0.376039
taxclass4	0.335969
land_square	0.292000
Brooklyn	0.200388
residential_units	0.192468
total_units	0.180459
buildingclass3	0.045443
commercial_units	0.042755
buildingclass10	-0.025421
buildingclass2	-0.068948
Queens	-0.109829
Bronx	-0.120081
year_built	-0.123245
State_Island	-0.216972
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
```

Random forest regression.

```
r2 0.6233097547347948
rmse 0.5256229959990711
```

The best results are for the random forest regression.