New York Sales document

Robert van der Heijden

3/1/2020

Content

1) Summary section

For this project, we will be creating a sales prediction system using the New York City Sales dataset. This dataset is a record of every building or building unit (apartment, etc.) sold in the New York City property market over a 12-month period. (source: https://www.kaggle.com/new-york-city/nyc-property-sales)

The goal of this project is to train a machine learning algorithm using the inputs in the New York City Sales dataset. The predictions from this machine learning algorithm will be compared to the true sales price in the validation set using RMSE.

2) Analysis Section

The first step is to create an New York City Sales (NYC_Sales) set and validation set.

```
--> Create NYC_Train set, NYC_validation set
```

Data Exploration

What is in the New York Sales data set?

1. The first ten lines of the dataset:

	Borough	Building	ZipCode	Resid.	Commer.
1	1	07 RENTALS - WALKUP APARTMENTS	10009	5	0
2	1	07 RENTALS - WALKUP APARTMENTS	10009	10	0
4	1	07 RENTALS - WALKUP APARTMENTS	10009	8	0
5	1	08 RENTALS - ELEVATOR APARTMENTS	10009	24	0
6	1	08 RENTALS - ELEVATOR APARTMENTS	10009	10	0
7	1	09 COOPS - WALKUP APARTMENTS	10009	0	0
8	1	09 COOPS - WALKUP APARTMENTS	10009	0	0
9	1	09 COOPS - WALKUP APARTMENTS	10009	0	0
10	1	09 COOPS - WALKUP APARTMENTS	10009	0	0
11	1	09 COOPS - WALKUP APARTMENTS	10009	0	0

	Year	TaxClass	Class	Sale Date	Sale Price
1	1900	2	C2	2017-07-19	6625000
2	1913	2	C4	2016-09-23	3936272
4	1920	2	C4	2016-09-23	3192840
5	1920	2	D9	2016-11-07	16232000
6	2009	2	D1	2016-10-17	10350000
7	1920	2	C6	2017-03-10	499000
8	1920	2	C6	2017-06-09	529500
9	1920	2	C6	2017-07-14	423000
10	1925	2	C6	2017-03-16	501000
11	1920	2	C6	2016-09-01	450000

2. The summary of the statistics of the dataset:

Borough	Building	ZipCode	Resid.	Commer.	Year
Min. :1.0	Length:52248	Min. :10001	Min. : 0.0	Min. : 0.00	Min. : 0
1st Qu.:2.0	Class :character	1st Qu.:10302	1st Qu.: 0.0	1st Qu.: 0.00	1st Qu.:1920
Median :3.0	Mode :character	Median :11207	Median : 1.0	Median : 0.00	Median :1940
Mean :2.9	NA	Mean :10824	Mean : 1.7	Mean : 0.17	Mean :1829
3rd Qu.:4.0	NA	3rd Qu.:11357	3rd Qu.: 2.0	3rd Qu.: 0.00	3rd Qu.:1966
Max. :5.0	NA	Max. :11694	Max. :1844.0	Max. :2261.00	Max. :2017

TaxClass	Class	Sale Date	Sale Price
Min. :1.0000	Length:52248	Min. :2016-09-01 00:00:00	Min. : 2
1st Qu.:1.0000	Class :character	1st Qu.:2016-11-30 00:00:00	1st Qu.: 385000
Median :2.0000	Mode :character	Median :2017-03-01 00:00:00	Median : 636355
Mean :1.6274	NA	Mean :2017-03-01 00:02:28	Mean : 1507998
3rd Qu.:2.0000	NA	3rd Qu.:2017-06-01 00:00:00	3rd Qu.: 1085047
Max. :4.0000	NA	Max. :2017-08-31 00:00:00	Max. :2210000000

3. The number of rows is:	52248
4. The number of columns is:	10
5. How many different zipcodes are in the NYC dataset?	182
6. The average saleprice is:	1507998

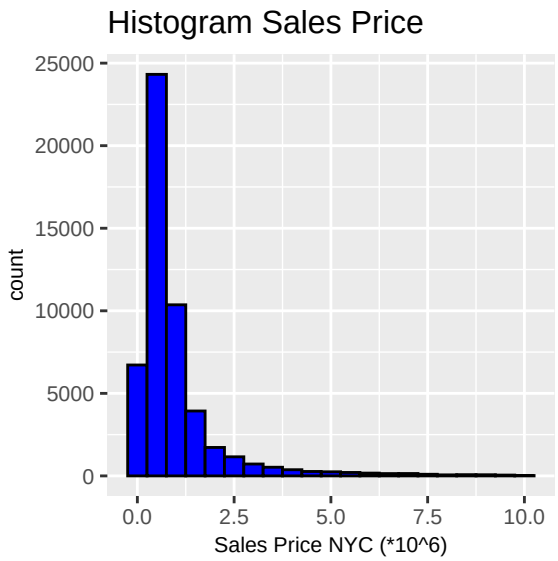
Data Cleaning

- Convert the saledate into Year-Month Format (needed for Data Visualisation)

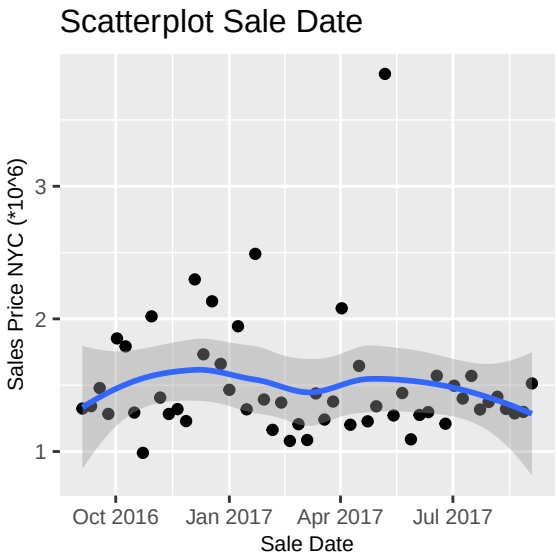
  `--> Year-Month added to train set and validation set`

Data Exploration (After Data Cleaning)

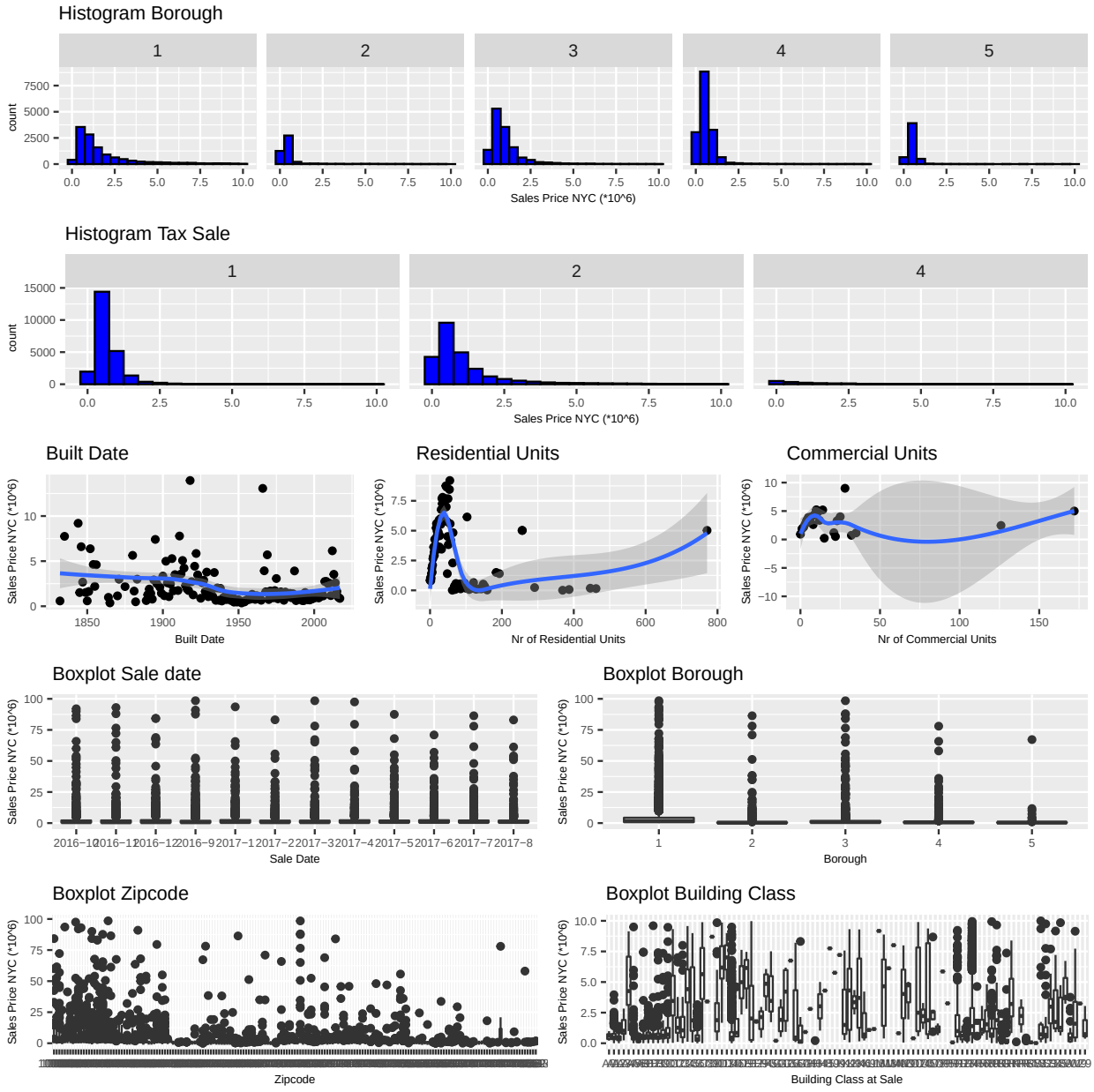
How are the sales prices distributed (histogram)?



How are the sales prices distributed over the years (boxplot)?



How are the sales prices distributed over the other predictors?



Modelling Approach

- RMSE will be used to evaluate how close the predictions are to the true values in the validation set.

Building the recommendation system

1. We start by building the simplest possible recommendation system. We're going to predict the same sales price for all buildings.

   The average that we predict is: **1,507,998 dollar**.

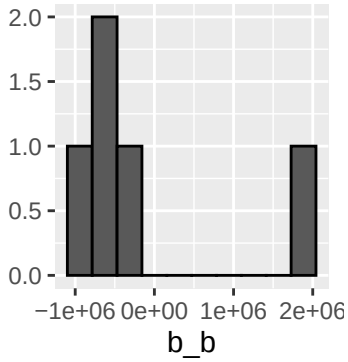
   How well does this model?

   The deviation on average is: **10,107,976 dollar**.

Now because as we go along we will be comparing different approaches, we're going to create a table that's going to store the results that we obtain as we go along.

Method	RMSE
Just the average	10107976

2. In the second step, we are going to take the borough into account.

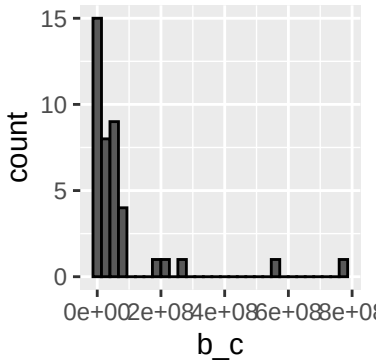


The histogram shows the deviation of the sales price in a borough from the average sales price. You can see that these estimates vary substantially. In some boroughs the sales prices are higher.

Updated RMSE table:

Method	RMSE
Just the average	10107976
BOROUGH Effect Model	10061538

3. In the third step, we are going to take the borough effect and commercial units effect into account.

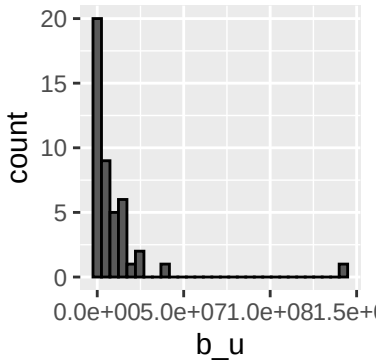


The histogram shows the variability of sales prices of the number of commercial units. There is substantial variability across the number of commercial units.

Updated RMSE table:

Method	RMSE
Just the average	10107976
BOROUGH Effect Model	10061538
BOROUGH & COMMERCIAL Effects Model	9904085

4. In the fourth step, we are going to take the borough, commercial units and building class effect into account.

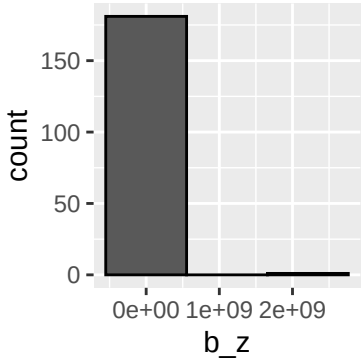


The histogram shows the variability of sales prices of building classes. There is variability across building classes.

Updated RMSE table:

Method	RMSE
Just the average	10107976
BOROUGH Effect Model	10061538
BOROUGH & COMMERCIAL Effects Model	9904085
BOROUGH & COMMERCIAL & BUILDING CLASS Effects Model	9679346

5. In the fifth step, we are going to take the borough, commercial units, building class and Zip code effect into account.



The histogram shows the variability of sales prices per zip code. There is variability across zip codes.

Updated RMSE table:

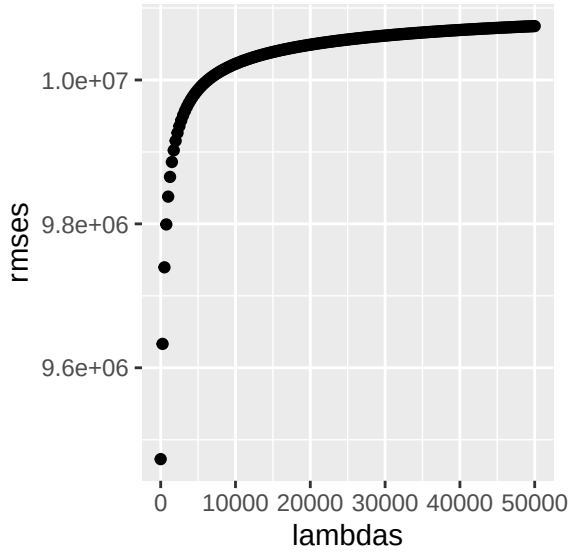
Method	RMSE
Just the average	10107976
BOROUGH Effect Model	10061538
BOROUGH & COMMERCIAL Effects Model	9904085
BOROUGH & COMMERCIAL & BUILDING CLASS Effects Model	9679346
BOROUGH & COMMERCIAL & BUILDING CLASS & ZIPCODE Effects Model	9673145

6. in the fifth step, we are going to regularize the borough, commercial units, building class and Zip code

effect.

Large errors can increase our residual mean squared error, so we would rather be conservative when we're not sure. Regularization permits us to penalize large estimates that come from small sample sizes.

First we have to pick the optimal tuning parameter lambda.




```
The optimal lambda is:  0
```

Updated RMSE table:

Method	RMSE
Just the average	10107976
BOROUGH Effect Model	10061538
BOROUGH & COMMERCIAL Effects Model	9904085
BOROUGH & COMMERCIAL & BUILDING CLASS Effects Model	9679346
BOROUGH & COMMERCIAL & BUILDING CLASS & ZIPCODE Effects Model	9673145
Regularized BOROUGH & COMMERCIAL & BUILDING CLASS & ZIPCODE Effects Model	9473041

3) Results Section

RMSE overview

The RMSE values for the used models are shown below:

Method	RMSE
Just the average	10107976
BOROUGH Effect Model	10061538
BOROUGH & COMMERCIAL Effects Model	9904085
BOROUGH & COMMERCIAL & BUILDING CLASS Effects Model	9679346
BOROUGH & COMMERCIAL & BUILDING CLASS & ZIPCODE Effects Model	9673145
Regularized BOROUGH & COMMERCIAL & BUILDING CLASS & ZIPCODE Effects Model	9473041

The RMSE table shows an improvement of the model over the different assumptions. The simplest model 'Just the average' calculates a RMSE of 10,107,976, which means, on average, we miss the sales price by

10,107,976 dollar. Incorporating 'Borough', 'Commercial Units', 'Building Class' and 'Zipcode' effects in our model gives an improvement of 0.46%, 2.02%, 4.24% and 4.3% respectively.

A deeper insight into the data revealed some data points have large effect on errors. So a regularization model was used to penalize these kind of data points. The final RMSE is 9473041 with an improvement of , 6.28% with respect to the baseline model.

Other sources of variation can be added to the model to further improve the predictability of the model.

4) Conclusion Section

For this project, we created a sales prediction system using the New York City Sales dataset.

The goal of this project is to train a machine learning algorithm using the inputs in the New York City Sales dataset. The predictions from this machine learning algorithm were compared to the true sales prices in the validation set using RMSE.

We started with a simple model, using the 'mean sales price' only, and added 'Borough', 'Commercial Units', 'Building Class' and 'Zipcode' effects. Our final model included regularization, that improved the final RMSE to **9,473,041**. This was an improvement of **6.28%** with respect to the baseline model.

Other sources of variation can be added to the model to further improve the predictability of the model.