CAPSTONE II: NYC HOUSE PRICE PREDICTION

HUA-LAI, TSENG(JUSTIN)

Methodologies

- **Purpose:**
  - To predict house price in NYC.
- **Data Preprocessing:**
  - Combine and organize NYC house price data from 2013-2019 acquire from NYC.gov
  - Filter, categorize, format data into different types of data (ex: w/ np.nan, w/0) for data analysis and machine learning purposes.
- **Statistics Inference & story telling:**
  - By doing explanatory data analysis, try to find correlation between features and house price in NYC.
  - Explore growth rate by plotting out different features versus house price with time frame.
- **Machine learning:**
  - Part one will be focusing on how to improve the result of NYC house price prediction by using features in every row.
  - Part two, since the result of part one is not ideal, planning to use time series to forecast monthly median house price for different house type in different borough in NYC.
- **Data source:**
  - The data is a combination of 35 data sets download from: https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page

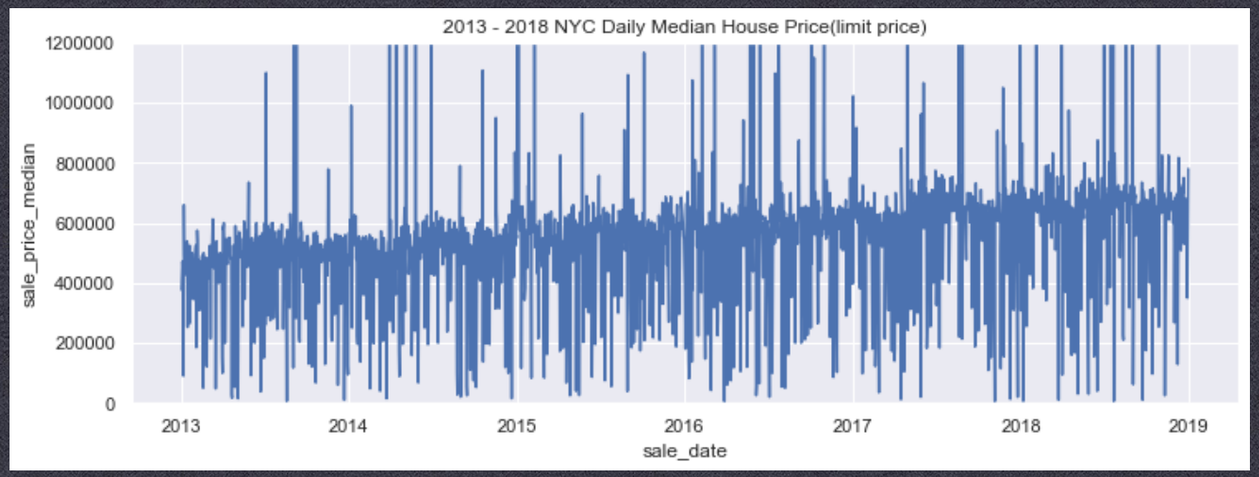
Data Preprocessing

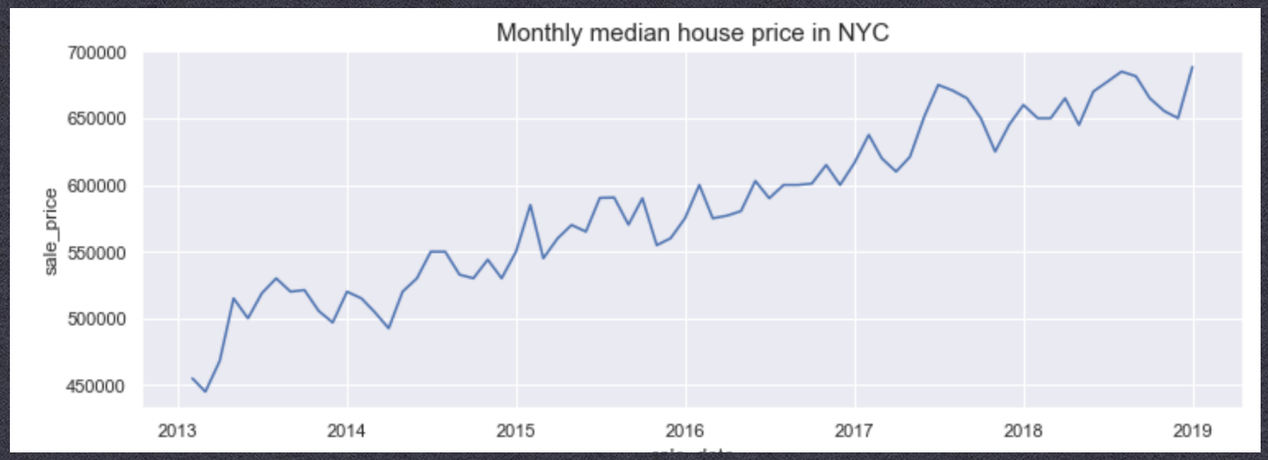
- **Combine and organize:**
  - Transform downloaded data sets from .xls format to .csv
  - Since there are 35 data sets downloaded from the NYC.gov site, I have to first combine and organize them, then separate them into training set (2013-2018) and testing set (2019).
- **Understand, clean & format each columns:**
  - First cleaned the target variable (sale_price) column by removing all the symbols($, /t -, etc), then drop rows with sale_price=0.
  - For independent variables, format and remove all symbols, and replace any abnormal or NaN values with 0.
- **Outliers or non sense values:**
  - For columns with units, residential units and commercial units should add up to total units, dropping the rows which do not add up.
  - To clean abnormally low sale price, have been trying a lot of methods (add price per sqft column, etc), since there also exist abnormal sqft values, dropping the bottom 1% sale price is still the best method to clean this column.
- **Data sets for different purpose:**
  - Try to limit rows I drop for machine learning purposes, I wrote a function to only drop rows with more than half of the values missing.
  - Data sets for different purposes: Data w/0 (Traditional machine learning method, and plotting), Data w/ NaN (Catboosting, LightGBM), Data with no 0 and NaN (Experiment).

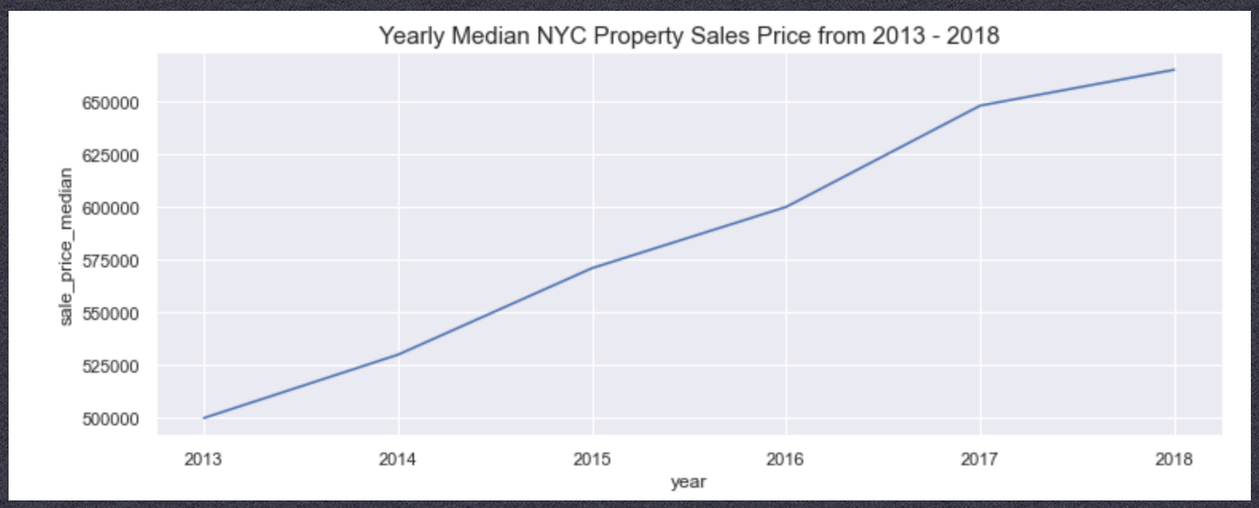
Statistics Inference & Story telling

Note I: Since none of the columns or counts are normally distributed, going to use median instead of mean here for analysis purpose.

Note II: All the analysis are for data between year 2013 - 2018 if not specifically mentioned.

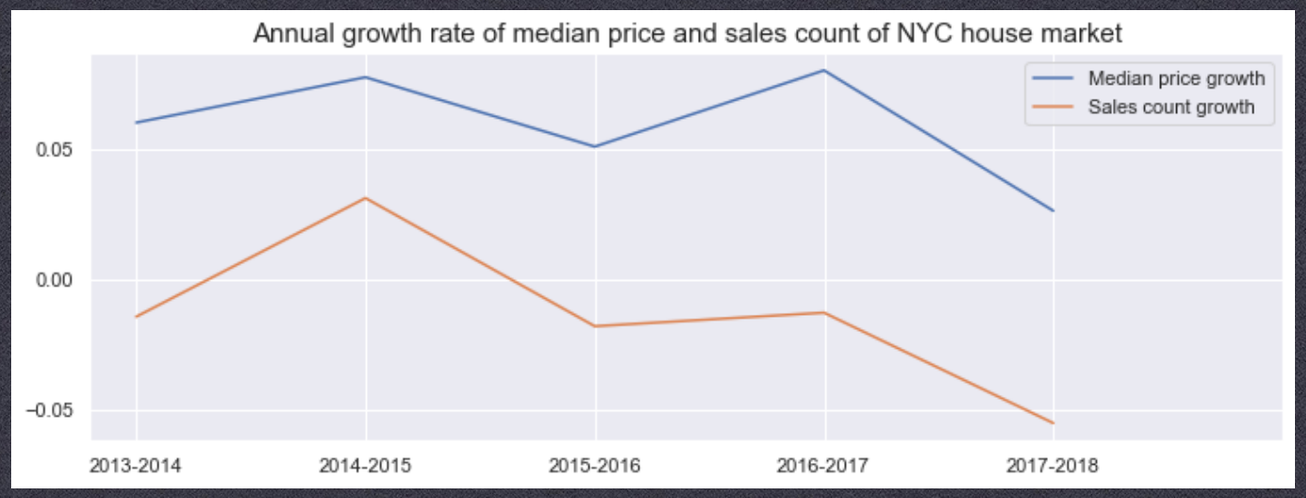


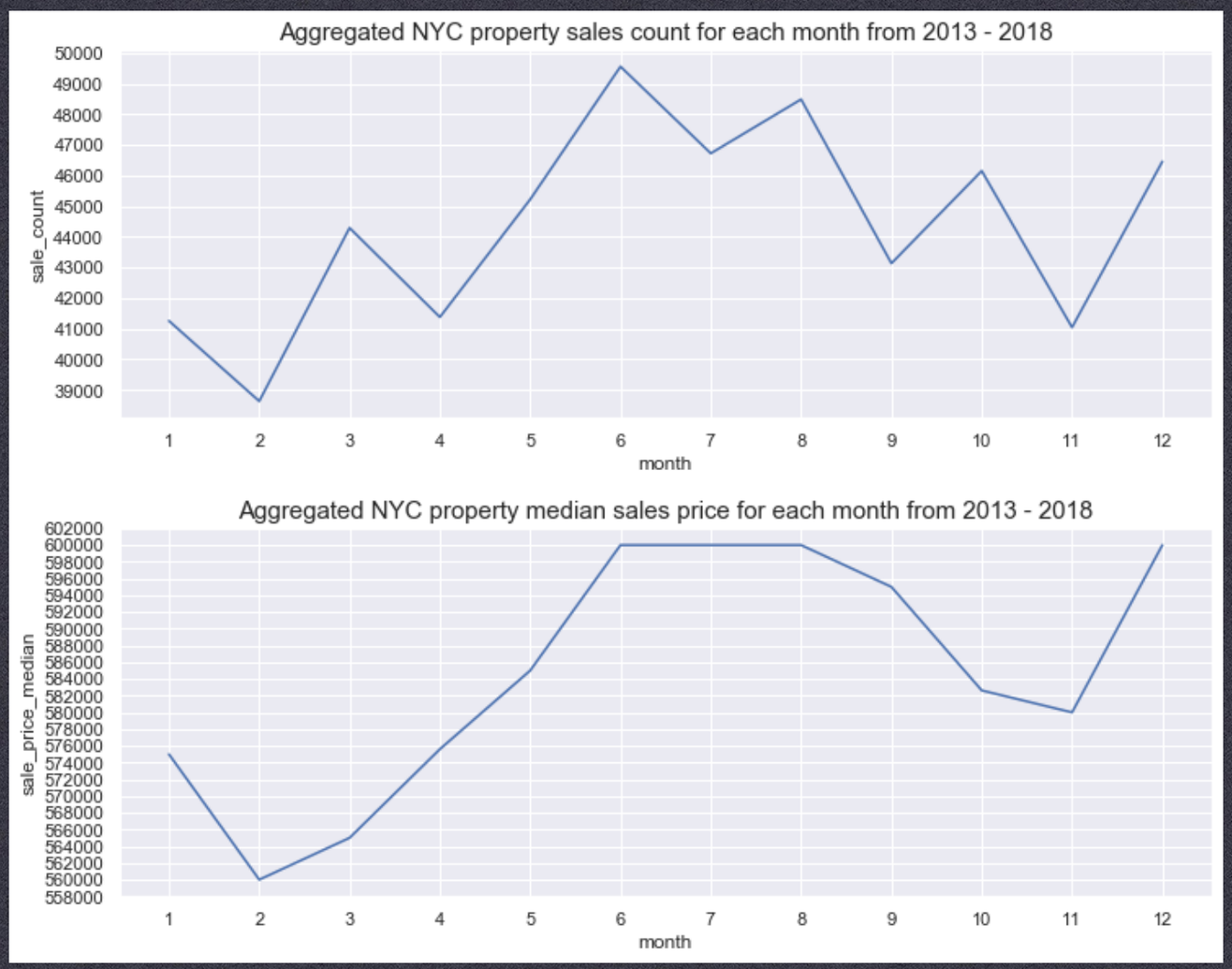




From daily, monthly and annual median house price plots on left hand side, we can see that the median house price in NYC shows an upward trend.

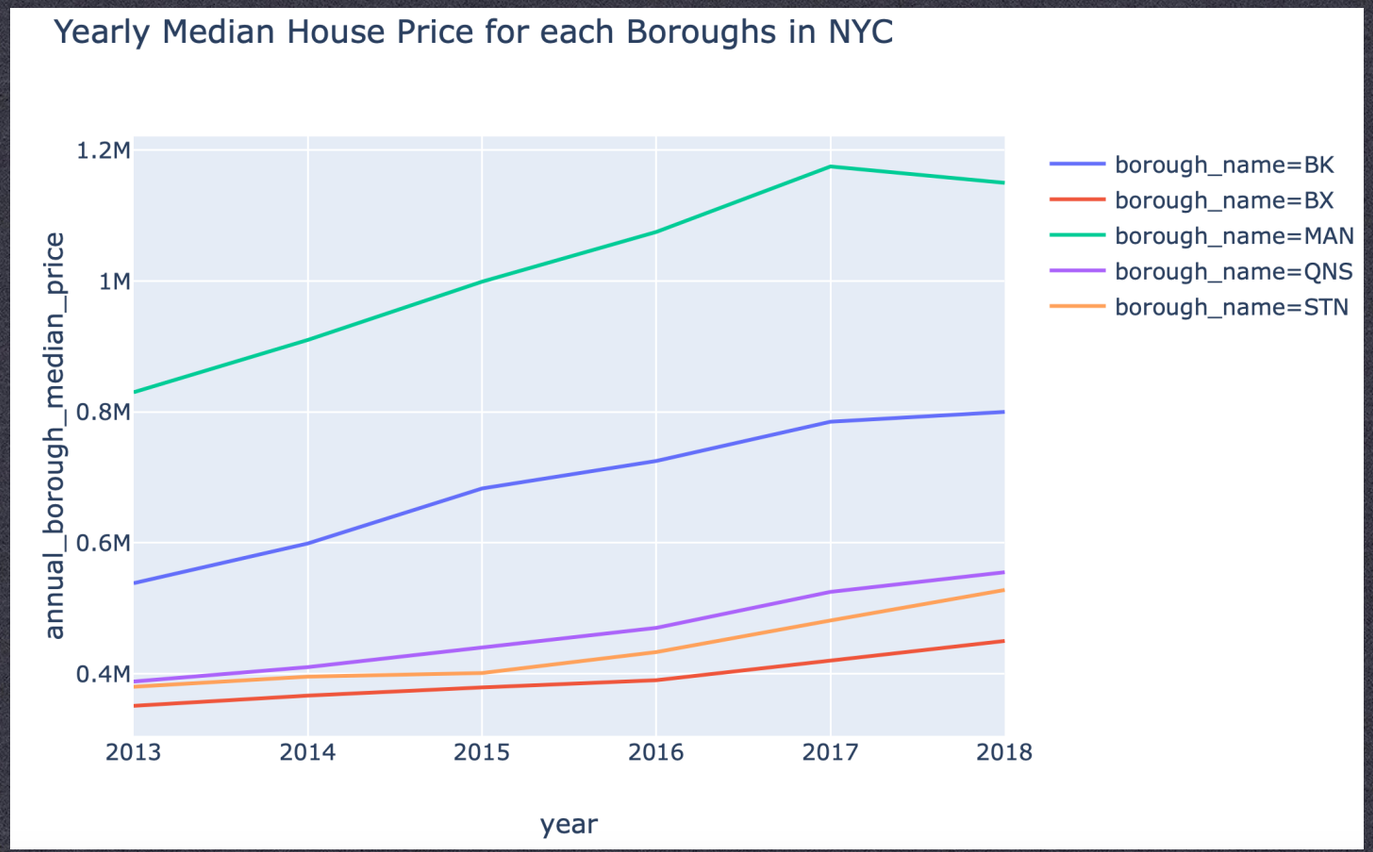
Statistics Inference & Story telling

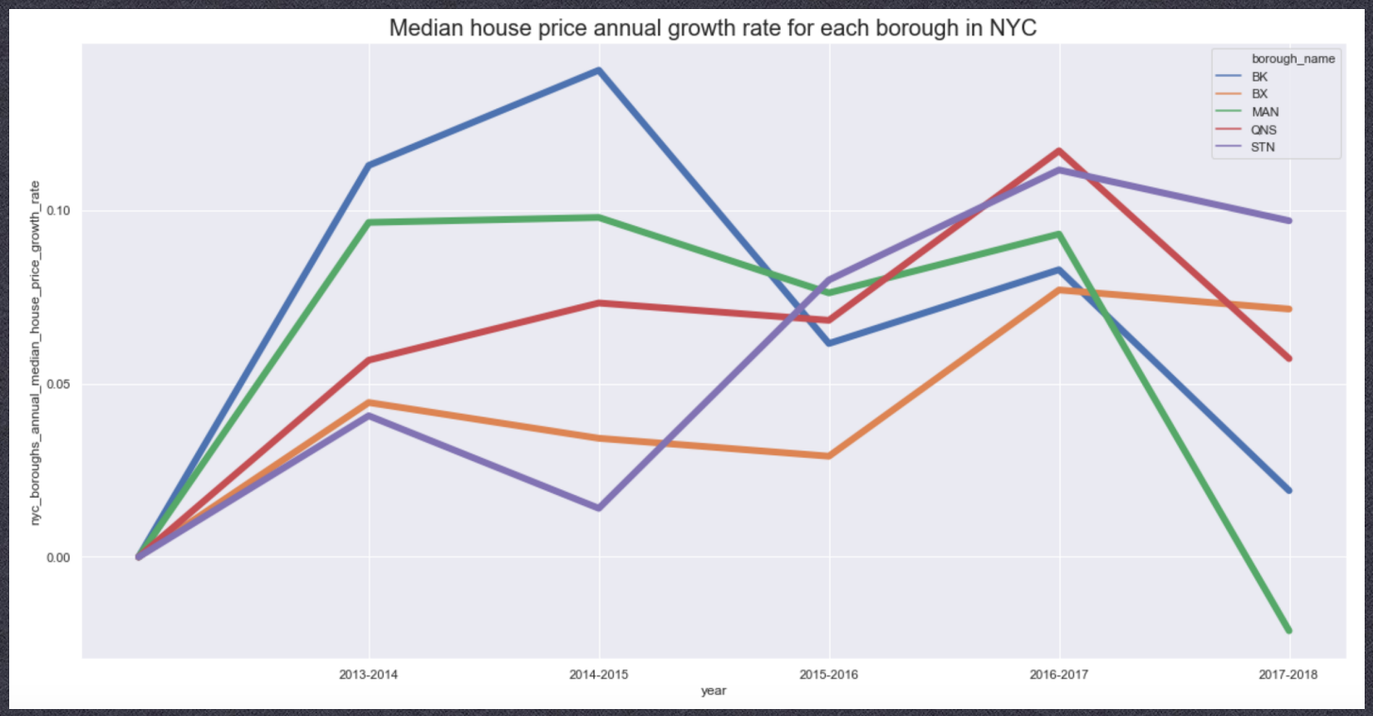




- **First plot: we can see that the annual growth rate of NYC median house price fluctuate with the annual sale count growth rate.**
- **Second plot: after aggregating months from each year, we can see that the trend of NYC's median house price of each month fluctuate with the amount of sales in each month.**

Statistics Inference & Story telling

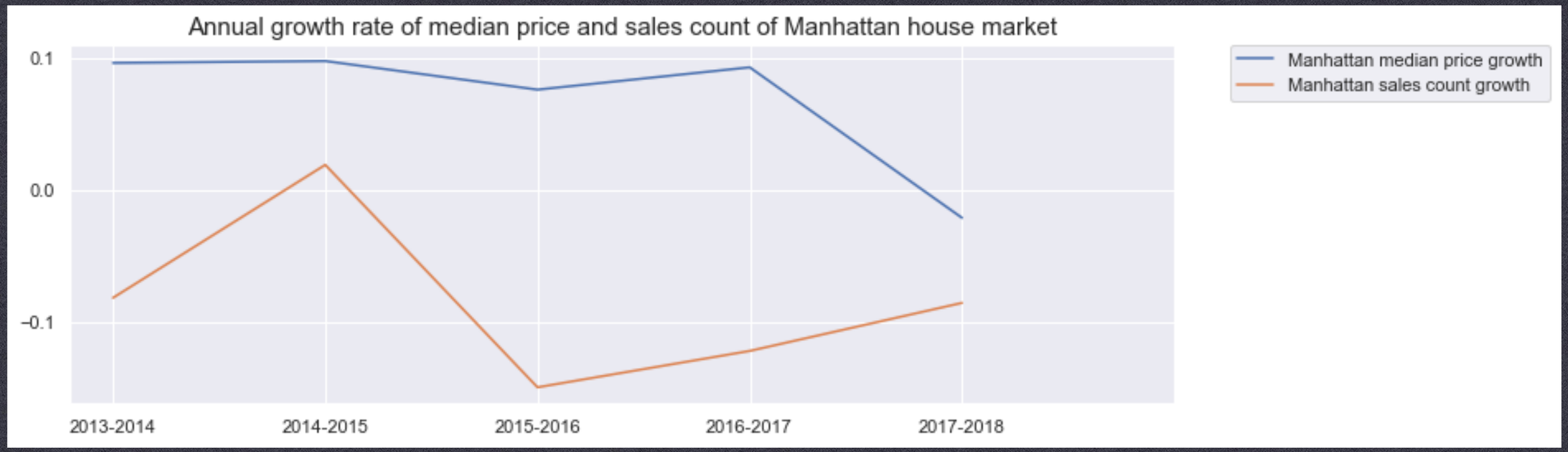




- **First plot: it is not surprising that the median house price in each borough of NYC from high to low is: MAN > BK > QNS > STN > BX, and there only exist two median house price declination in this plot, one happened between 2017-2018 in Manhattan, the other happened between 2014-2015 in Staten Island.**
- **Second plot: interesting facts in this plot is that growth rate of each borough in NYC all increase between 2016-2017 and all decline between 2017-2018.**
- **The chart below shows that Brooklyn have the highest average annual median house price growth rate from 2013 - 2018**

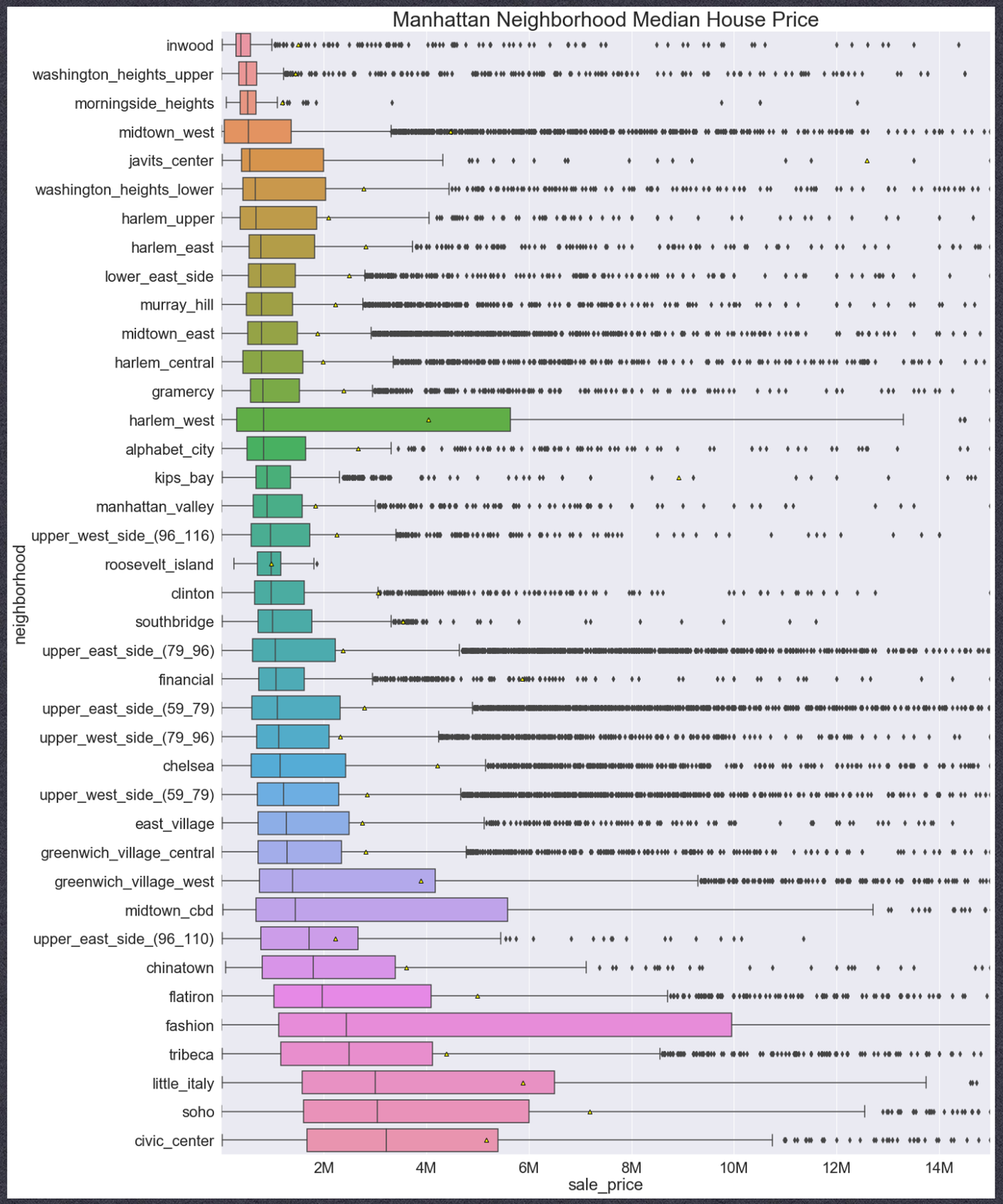
	borough_name	median_growth_rate_avg
0	BK	0.079459
1	MAN	0.072657
2	QNS	0.063015
3	STN	0.049199
4	BX	0.036926

Statistics Inference & Story telling – MAN



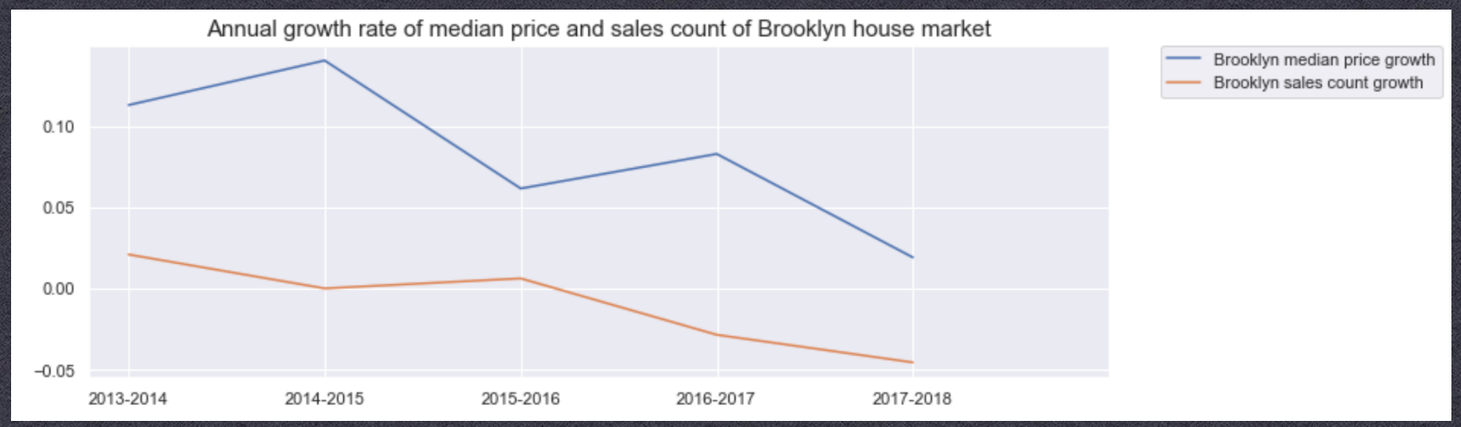
- From the plot above we can see that in Manhattan, trend of annual median sales price growth rate seems to generally follow the trend of growth rate of amount of sales.
- From the box plot next page, we can see that Civic center, SOHO, Little Italy are the neighborhood which have the highest median house price in Manhattan; Inwood, Washington Heights Upper, Morningside Heights are the neighborhoods with the lowest median house price in Manhattan.
- From the chart on next page, we can see that neighborhood Midtown West has an enormous 114% annual average median house price growth rate. SOHO, neighborhood with one of the highest median house price, has one of the lowest average annual growth rate compare to other top median price neighborhood. Harlem Upper is the only neighborhood in Manhattan with a negative average annual median house price growth rate.

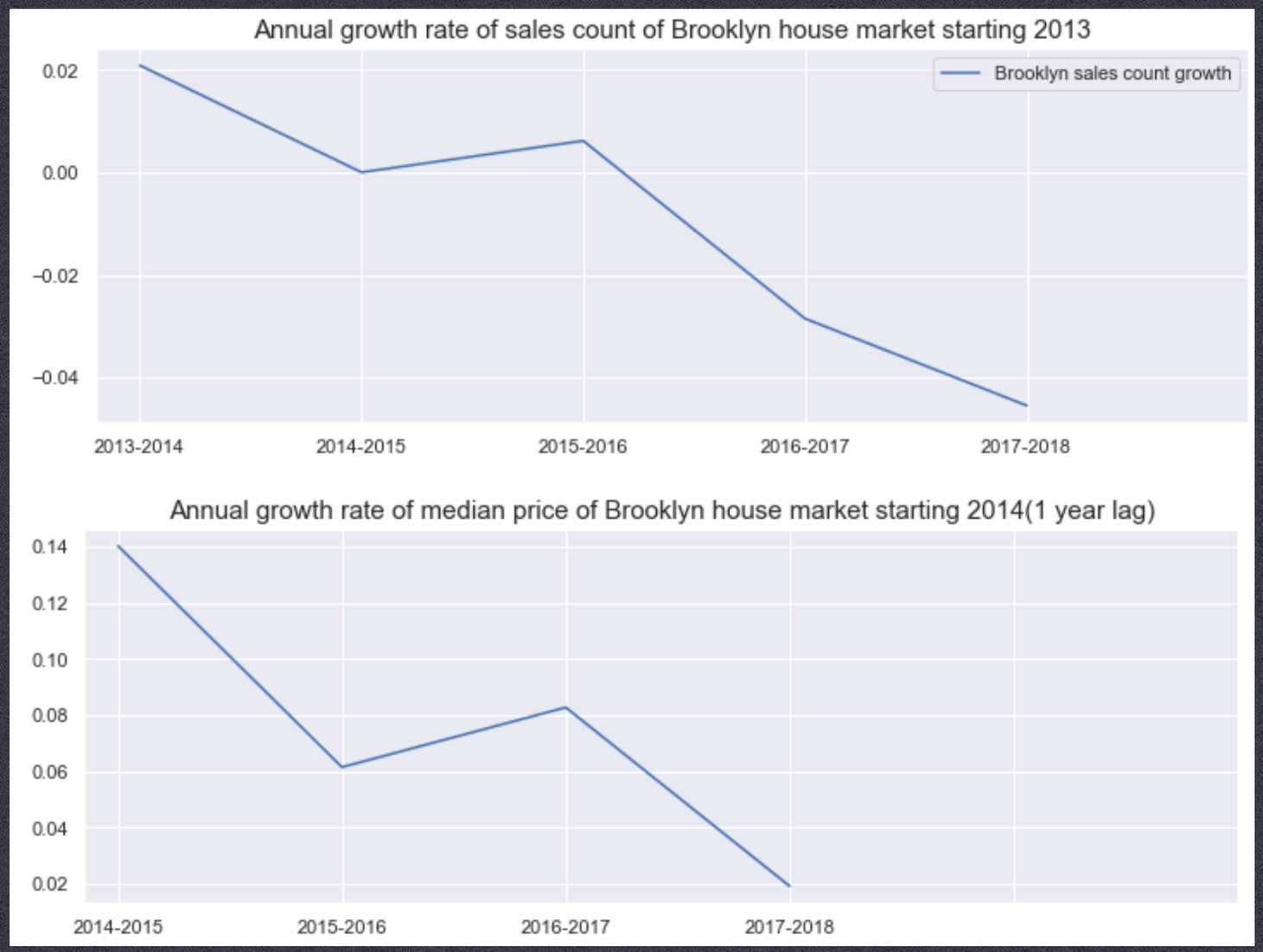
Statistics Inference & Story telling – MAN




```
neighborhood
midtown_west                  1.141422
midtown_cbd                   0.341483
lower_east_side               0.165223
civic_center                  0.149967
little_italy                  0.134494
east_village                  0.124090
greenwich_village_west        0.115136
harlem_west                   0.114528
harlem_central                0.107419
morningside_heights           0.102454
flatiron                      0.090285
greenwich_village_central     0.087445
washington_heights_upper      0.079361
gramercy                      0.076555
fashion                       0.076307
chinatown                     0.071865
upper_east_side_(79_96)       0.069263
upper_west_side_(79_96)       0.067495
clinton                       0.062630
inwood                        0.060263
kips_bay                      0.059040
upper_west_side_(59_79)       0.055660
manhattan_valley              0.054938
tribeca                       0.054903
upper_east_side_(59_79)       0.054540
chelsea                       0.050790
upper_west_side_(96_116)      0.043522
financial                     0.043162
javits_center                 0.037261
harlem_east                   0.034424
midtown_east                  0.030759
murray_hill                   0.028092
soho                          0.027881
upper_east_side_(96_110)      0.020980
washington_heights_lower      0.017806
southbridge                   0.008605
alphabet_city                 0.007069
harlem_upper                 -0.006035
Name: median_growth_rate_avg, dtype: float64
```

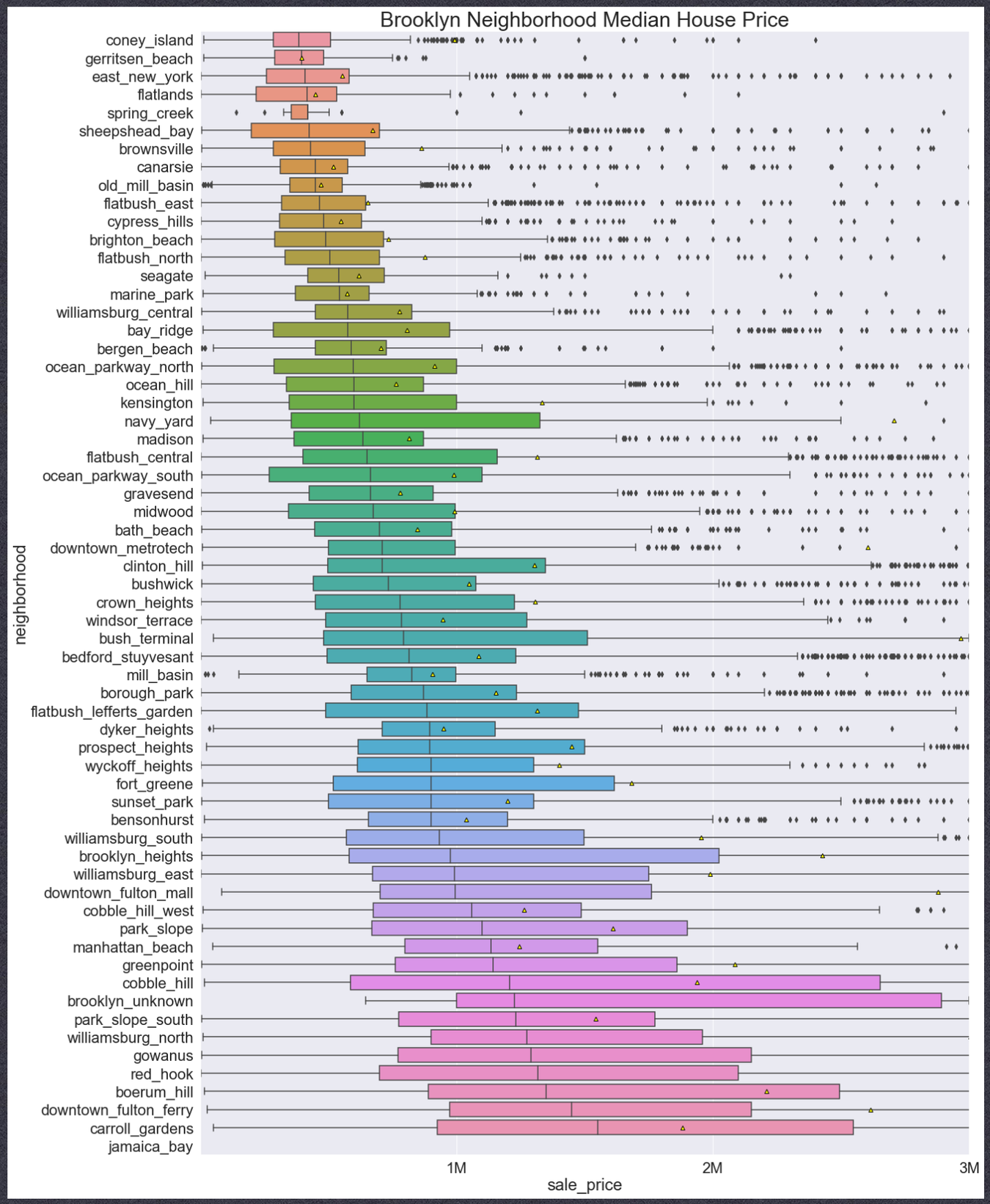
Statistics Inference & Story telling — BK





- **When first plot out the line plot I realized that the annual amount of sales growth rate in Brooklyn seemed to fluctuate with annual median house price growth rate in Brooklyn with one year lagged. From the plot below, we can see that both lines seemed to share the same pattern.**
- **From the box plot next page, Jamaica Bay, Carroll Gardens and Downtown Fulton Ferry are neighborhoods with the highest median house price in Brooklyn. On the other hand, Coney Island, Gerritsen Beach and East New York are neighborhoods with the lowest median house price.**
- **From chart on next page, we can see that Navy Yard is the Brooklyn neighborhood with the highest average annual median house price growth rate, 56.6%. One thing to notice is that there are no neighborhoods in Brooklyn with negative average annual median house price growth rate.**

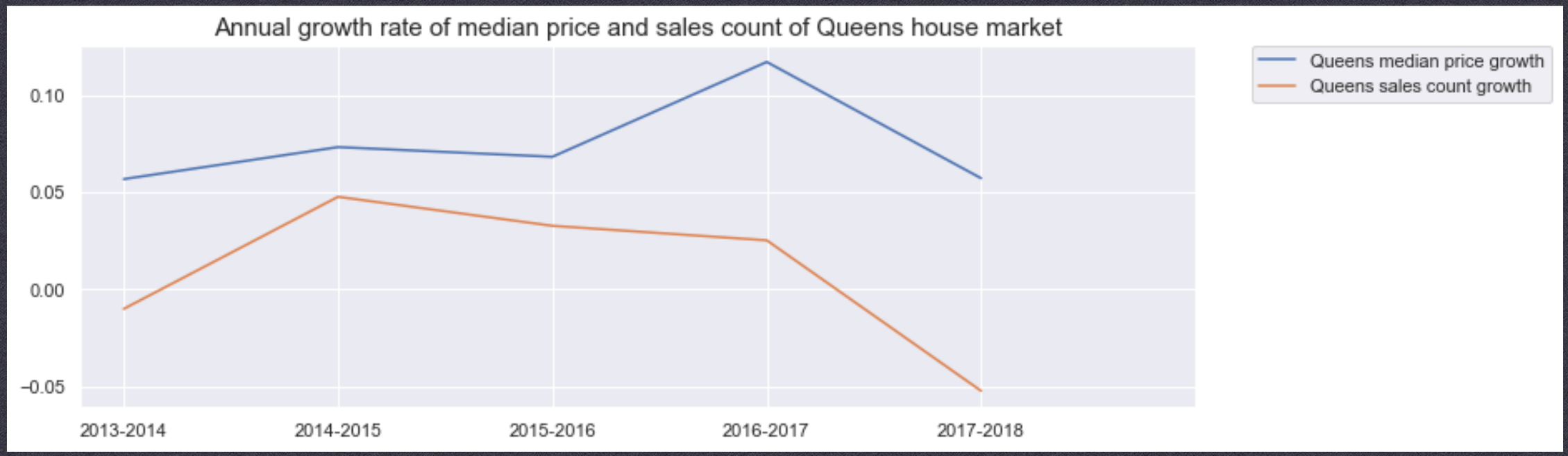
Statistics Inference & Story telling — BK



neighborhood	
navy_yard	0.566543
cobble_hill	0.360498
red_hook	0.234527
kensington	0.208621
ocean_parkway_north	0.174349
ocean_hill	0.165042
ocean_parkway_south	0.161155
bushwick	0.159441
windsor_terrace	0.151552
prospect_heights	0.150978
crown_heights	0.145629
bedford_stuyvesant	0.145396
bush_terminal	0.144528
cobble_hill_west	0.142504
downtown_fulton_ferry	0.136295
manhattan_beach	0.135620
boerum_hill	0.133720
seagate	0.133285
gowanus	0.132051
east_new_york	0.130256
sunset_park	0.129635
flatbush_lefferts_garden	0.125307
clinton_hill	0.124157
flatbush_central	0.118073
greenpoint	0.117725
downtown_fulton_mall	0.114537
wyckoff_heights	0.112740
park_slope_south	0.110515
williamsburg_central	0.106826
brooklyn_heights	0.104942
flatbush_north	0.104335
cypress_hills	0.101866
park_slope	0.099461
bay_ridge	0.097426
fort_greene	0.094468
brownsville	0.092390
williamsburg_east	0.092095
flatbush_east	0.086091
coney_island	0.084073
carroll_gardens	0.079338
marine_park	0.079311
midwood	0.078105
madison	0.075098
canarsie	0.074657
gerritsen_beach	0.073039
dyker_heights	0.071867
borough_park	0.071550
williamsburg_south	0.069602
old_mill_basin	0.069143
bath_beach	0.066956
sheepshead_bay	0.066739
gravesend	0.065956
bensonhurst	0.065274
downtown_metrotech	0.060912
mill_basin	0.048346
brighton_beach	0.046071
bergen_beach	0.041183
williamsburg_north	0.027152
flatlands	0.018984

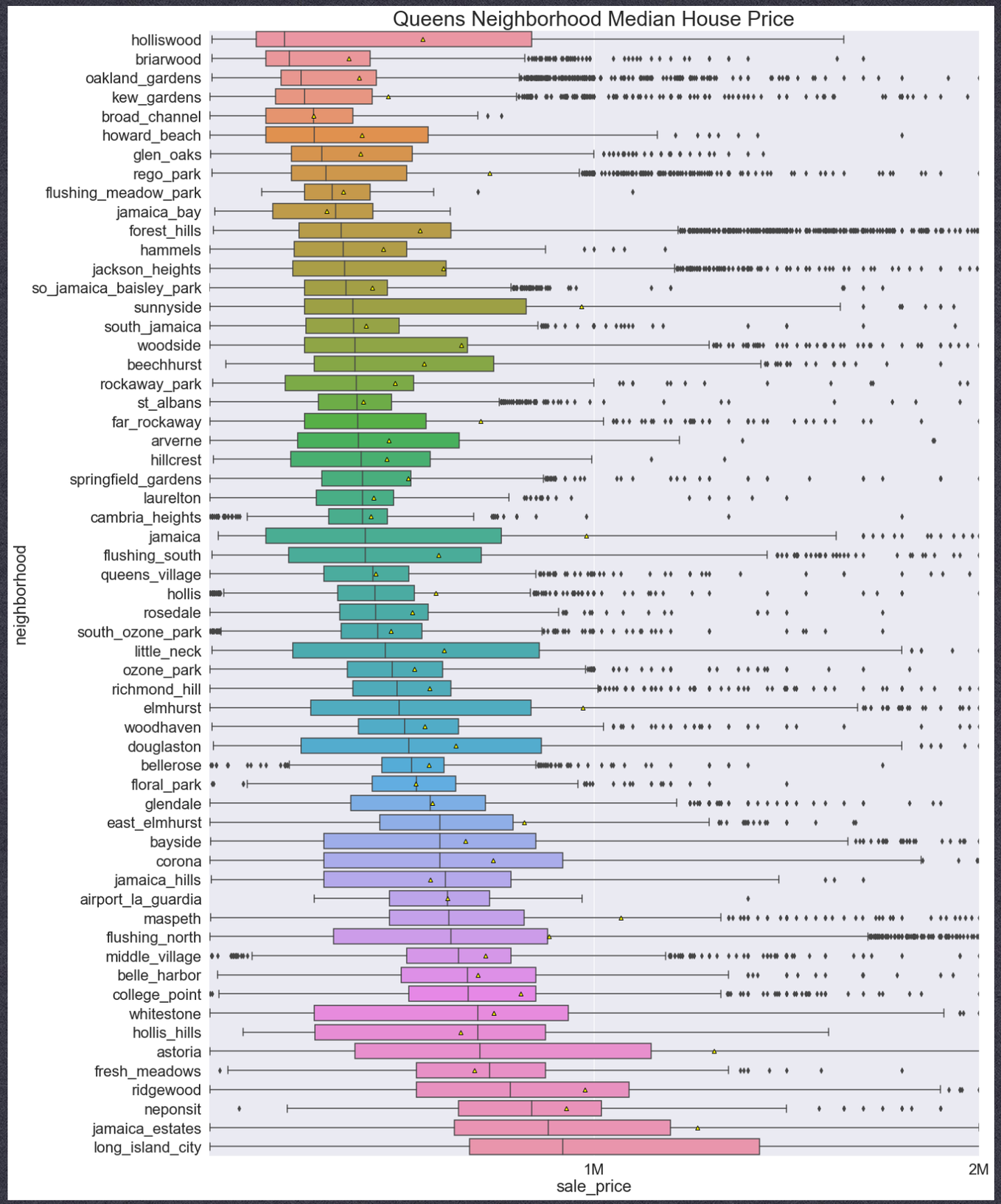
Name: median_growth_rate_avg, dtype: float64

Statistics Inference & Story telling — QNS



- Line plot above showed that in Queens, annual median house price growth rate generally fluctuate with growth rate of annual amount of sales.
- From the box plot next page, Long Island City, Jamaica Estate and Neponsit are the neighborhoods with the highest median house price in Queens. Holliswood, Briarwood and Oakland Gardens are the lowest.
- Though Holliswood has the lowest median house price, but it has the highest average annual median house price growth rate in Queens, 26%. There are three neighborhood with negative average annual median house price growth rate in Queens: Arverne (-4.7%), Little Neck (-3.4%) and Hillcrest (-0.006%).

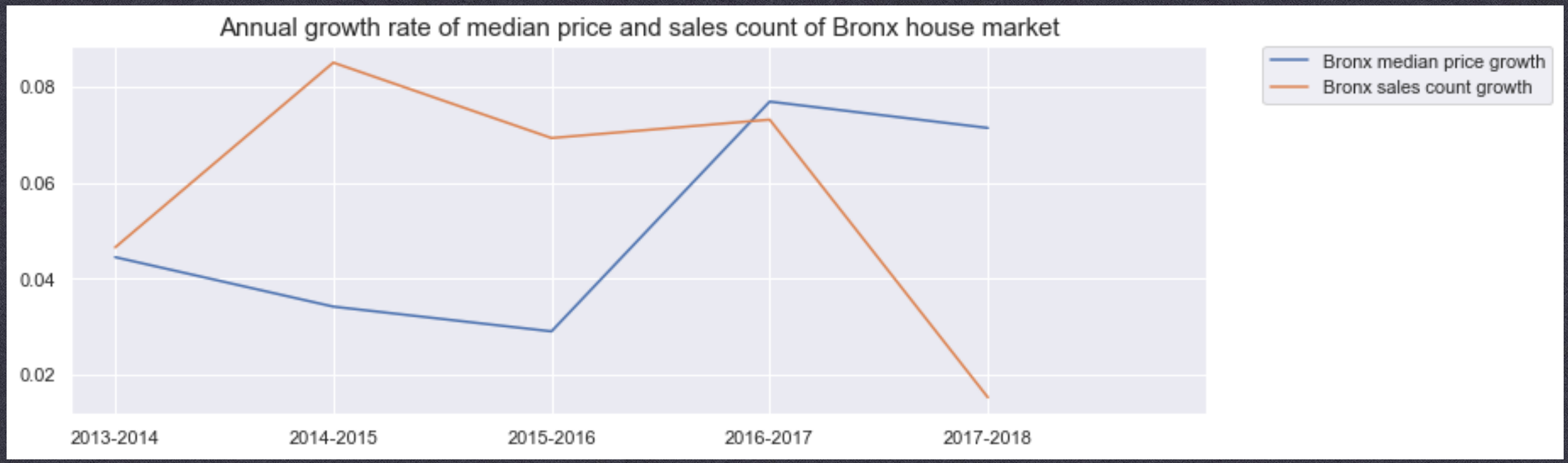
Statistics Inference & Story telling – QNS



neighborhood	
holliswood	0.261308
broad_channel	0.231671
jamaica_bay	0.146689
flushing_meadow_park	0.144200
howard_beach	0.138793
jamaica_hills	0.121599
airport_la_guardia	0.118292
jackson_heights	0.114243
far_rockaway	0.112606
so_jamaica_baisley_park	0.110169
woodside	0.108786
rockaway_park	0.107835
east_elmhurst	0.106015
ridgewood	0.104420
elmhurst	0.103473
woodhaven	0.102652
rego_park	0.101461
belle_harbor	0.096413
briarwood	0.095053
hammels	0.094795
maspeth	0.093729
college_point	0.092495
richmond_hill	0.090039
flushing_north	0.089368
forest_hills	0.088857
south_jamaica	0.087327
glendale	0.086821
hollis	0.085712
sunnyside	0.085383
corona	0.085174
queens_village	0.084255
ozone_park	0.083749
jamaica	0.082846
south_ozone_park	0.082227
springfield_gardens	0.078227
middle_village	0.075762
st_albans	0.075616
astoria	0.075412
whitestone	0.074350
bayside	0.073487
beechhurst	0.073302
hollis_hills	0.072494
jamaica_estates	0.071894
oakland_gardens	0.069699
kew_gardens	0.068975
cambria_heights	0.068553
fresh_meadows	0.067411
rosedale	0.062168
floral_park	0.060615
bellerose	0.060527
laurelton	0.059852
glen_oaks	0.059062
neponsit	0.058815
flushing_south	0.048854
long_island_city	0.034080
douglaston	0.011250
hillcrest	-0.006104
little_neck	-0.034509
arverne	-0.047478

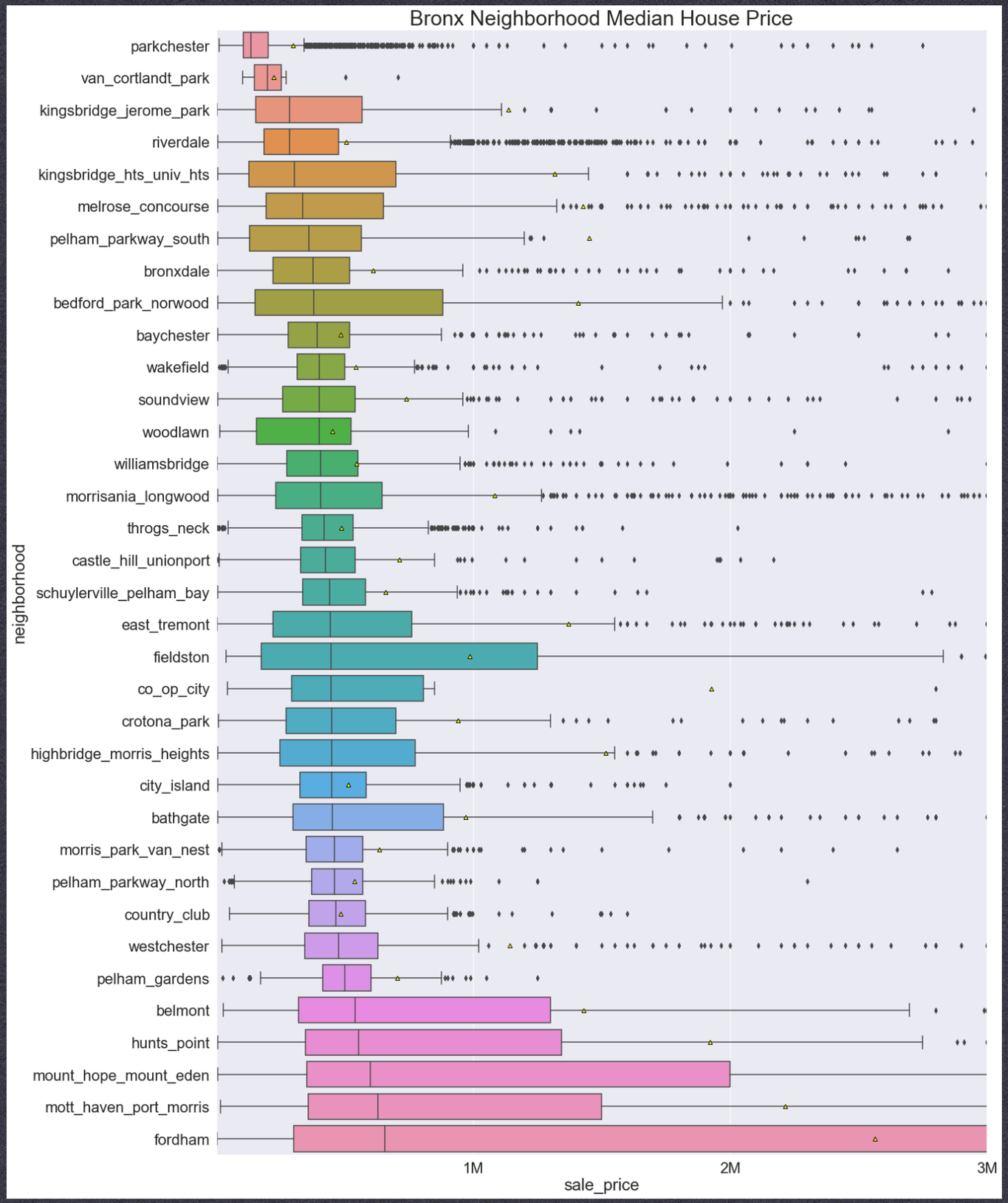
Name: median_growth_rate_avg, dtype: float64

Statistics Inference & Story telling — BX



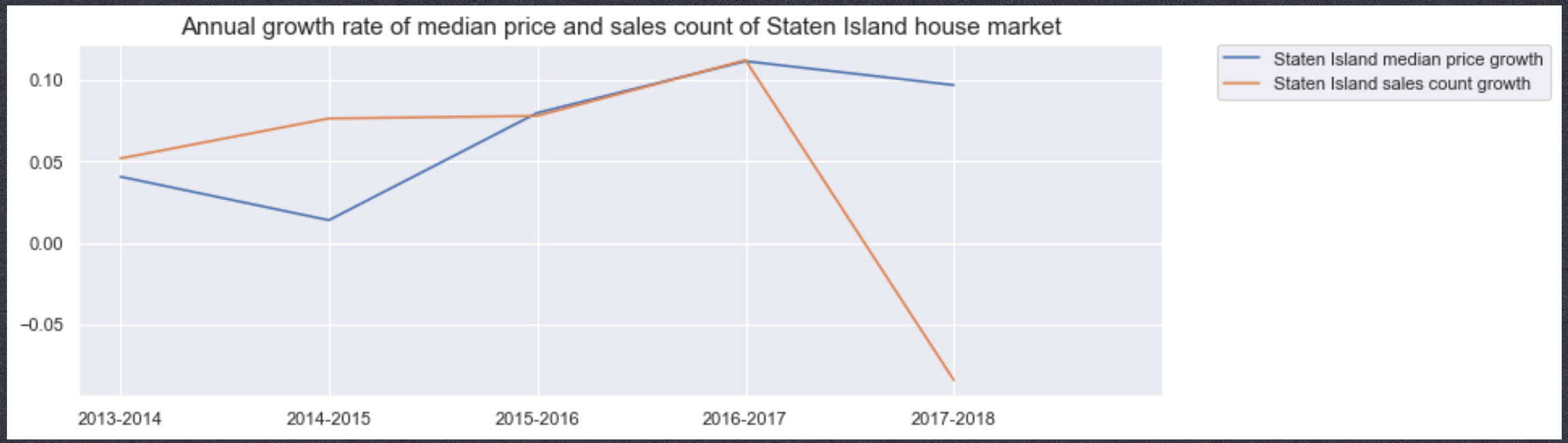
- From the line plot above, we cannot say there exist any correlation between Bronx's average annual median house price growth rate and growth rate of Bronx's annual amount of sales.
- Fordham, Mott Haven Port Morris and Mount Hope Mount Eden are the neighborhoods in Bronx with the highest median house price. Neighborhood Parkchester and Van Cortland Park, on the other hand, have the lowest median price.
- From the growth rate chart on second page, average annual median house price growth rate for neighborhoods in Bronx has the smallest gap between maximum and minimum growth rate. Kingsbridge Hts Univ Hts is the neighborhood with the highest average annual median house price growth rate (22.7%). Fordham, which is has the highest median house price, has the lowest average annual median house price growth rate (1.9%).

Statistics Inference & Story telling — BX



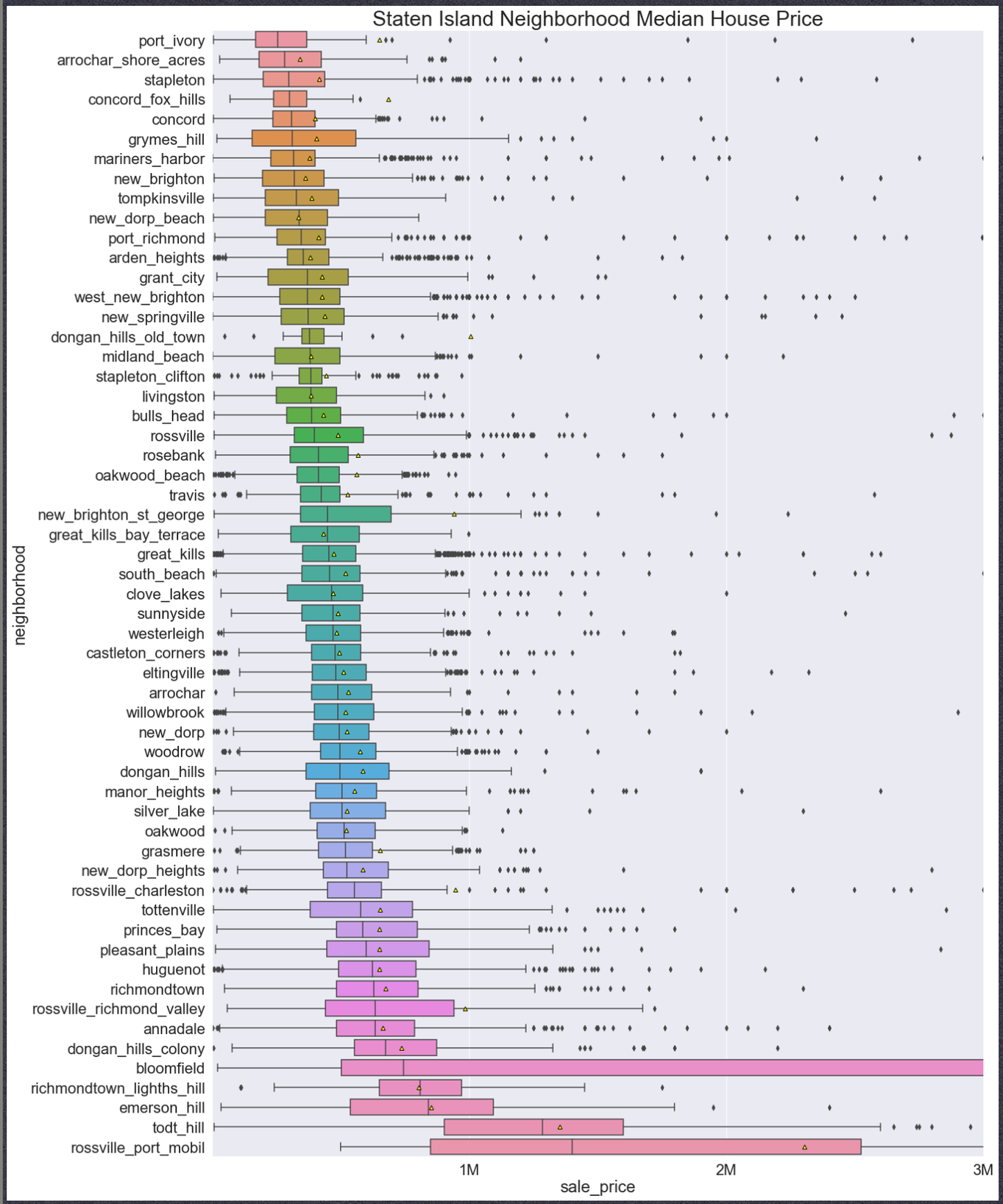

```
neighborhood
kingsbridge_hts_univ_hts      0.227404
co_op_city                    0.216535
fieldston                     0.148937
mount_hope_mount_eden         0.133666
east_tremont                  0.132409
van_cortlandt_park            0.125215
mott_haven_port_morris        0.098850
melrose_concourse             0.092638
kingsbridge_jerome_park       0.089516
castle_hill_unionport         0.083315
pelham_parkway_south          0.077297
bedford_park_norwood          0.075233
parkchester                   0.074220
belmont                       0.072037
morrisania_longwood           0.071800
bronxdale                     0.070167
williamsbridge                0.067054
baychester                    0.066405
morris_park_van_nest          0.066323
bathgate                      0.065689
schuylerville_pelham_bay      0.062049
pelham_gardens                0.061969
wakefield                     0.061507
country_club                  0.060990
pelham_parkway_north          0.059336
city_island                   0.055872
westchester                   0.054190
highbridge_morris_heights     0.048923
throgs_neck                   0.047205
hunts_point                   0.039572
crotona_park                  0.038854
woodlawn                      0.037958
soundview                     0.037809
riverdale                     0.027738
fordham                       0.019096
Name: median_growth_rate_avg, dtype: float64
```

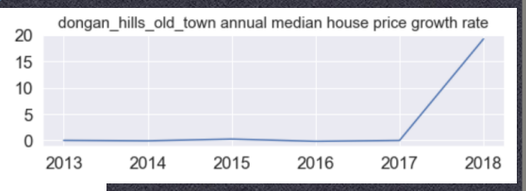
Statistics Inference & Story telling – STN



- We cannot really said that there exist a relationship between the annual growth rate of median house price in Staten Island and the annual growth rate of amount of sales in Staten Island from the line plot above.
- Neighborhoods Rossville Port Mobil and Todt Hill have the highest median house price in Staten Island , the median house price gap between these two neighborhoods to the neighborhood with the third highest median house price is noticeable.
- The chart on next page shows that Dongan Hills Old Town have the highest average annual median house price growth rate, with incredible 385%, after examination, the abnormality is due to that there is only one sale in 2018 with a fairly high price. Neighborhood Pleasant Plains is the only neighborhood in Staten Island with negative growth rate (-0.5%).

Statistics Inference & Story telling – STN

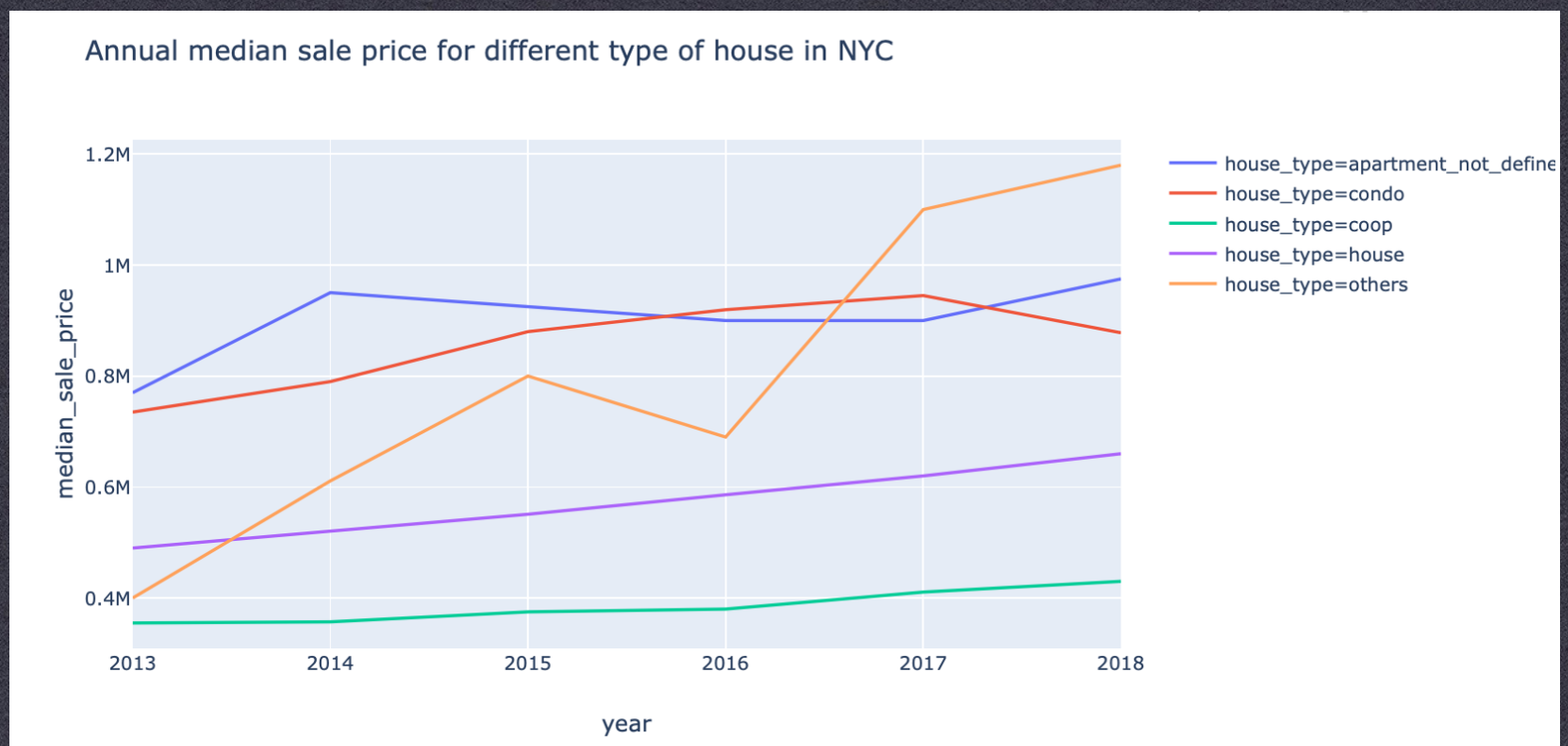




neighborhood	
dongan_hills_old_town	3.850851
bloomfield	1.050590
rossville_port_mobil	0.406050
grymes_hill	0.191610
new_dorp_beach	0.180245
stapleton_clifton	0.170260
new_brighton_st_george	0.155618
rossville_richmond_valley	0.141849
concord_fox_hills	0.140548
arrochar_shore_acres	0.135402
midland_beach	0.112999
great_kills_bay_terrace	0.108210
emerson_hill	0.106586
new_brighton	0.102503
grant_city	0.096443
bulls_head	0.095584
todt_hill	0.094730
dongan_hills	0.093837
mariners_harbor	0.093465
clove_lakes	0.092019
stapleton	0.091559
sunnyside	0.090159
arrochar	0.089875
rosebank	0.089057
oakwood_beach	0.087869
west_new_brighton	0.084507
south_beach	0.084468
richmondtown_lighths_hill	0.083938
oakwood	0.080003
port_richmond	0.078386
castleton_corners	0.078062
new_dorp	0.076001
manor_heights	0.074692
westerleigh	0.071306
willowbrook	0.070092
great_kills	0.068433
eltingville	0.068288
grasmere	0.067310
rossville	0.066896
woodrow	0.066151
silver_lake	0.065878
livingston	0.064158
richmondtown	0.064083
tottenville	0.063356
tompkinsville	0.062234
princes_bay	0.061599
new_springville	0.059969
arden_heights	0.059527
dongan_hills_colony	0.058585
concord	0.056726
travis	0.048745
huguenot	0.048561
rossville_charleston	0.047979
annadale	0.046019
port_ivory	0.038582
new_dorp_heights	0.018150
pleasant_plains	-0.005327

Name: median_growth_rate_avg, dtype: float64

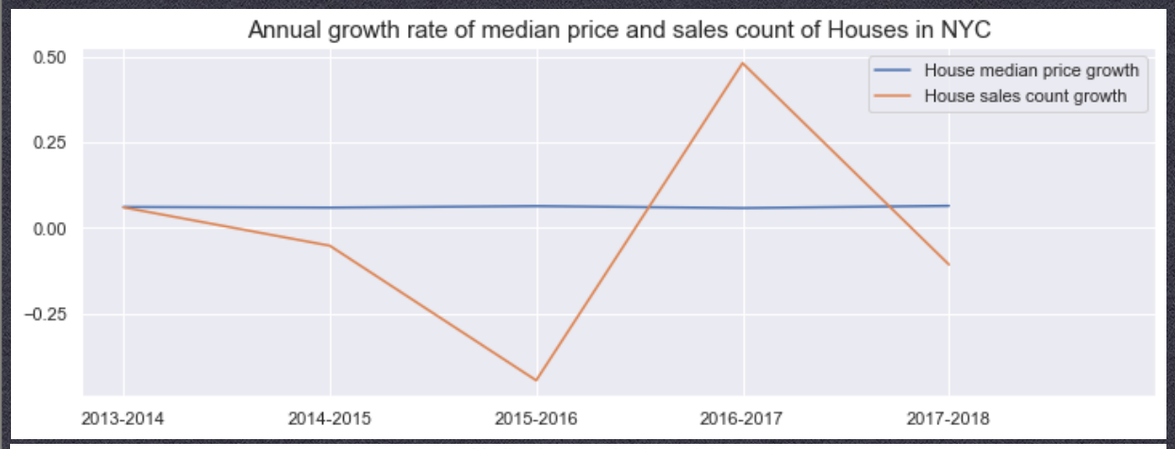
Statistics Inference & Story telling

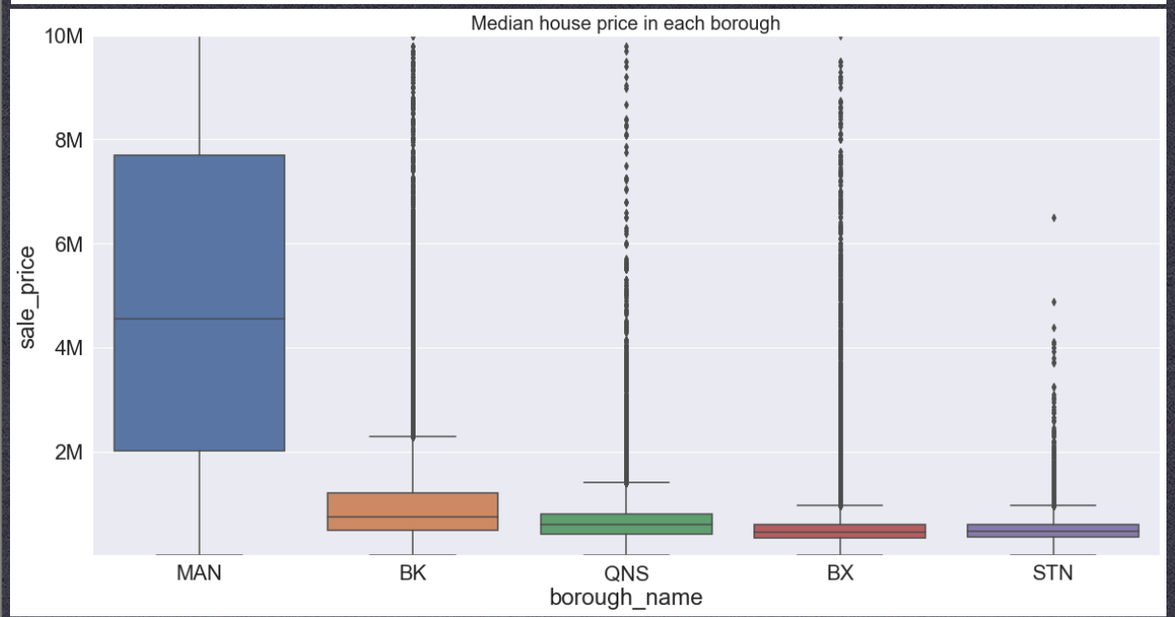


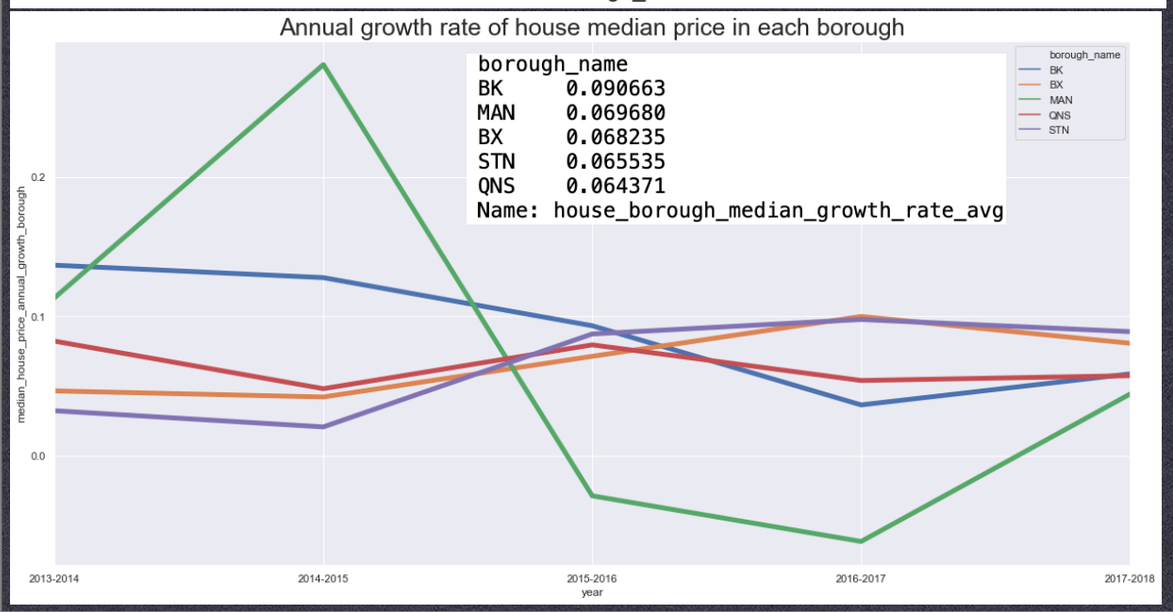
- Though from the violin plot, coop seems to have higher price than house, the statistics test results showed, in general, price of condo > price of house > price of coop.
- From the line plot, we can see that price of condo and house seems to increase in a steady rate, however, price of condo shows a decline between year 2017 - 2018.
- From the charts below, we can see that median price of house have a higher average annual growth rate compare to coop and condo.

```
house_type
others                    0.273252
house                     0.061380
apartment_not_define      0.052766
coop                      0.039433
condo                     0.038093
Name: annual_diff_type_mid_price_growth_rate_avg
```

Statistics Inference & Story telling — House

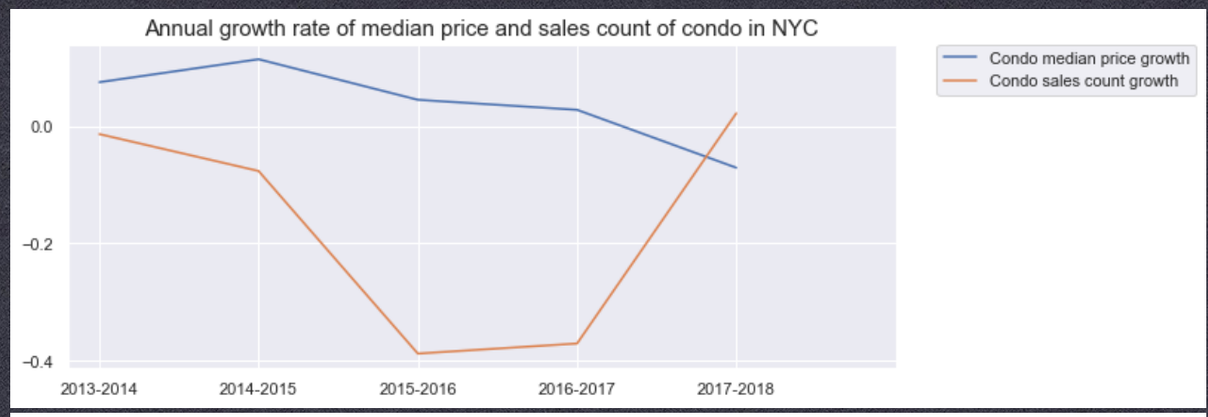


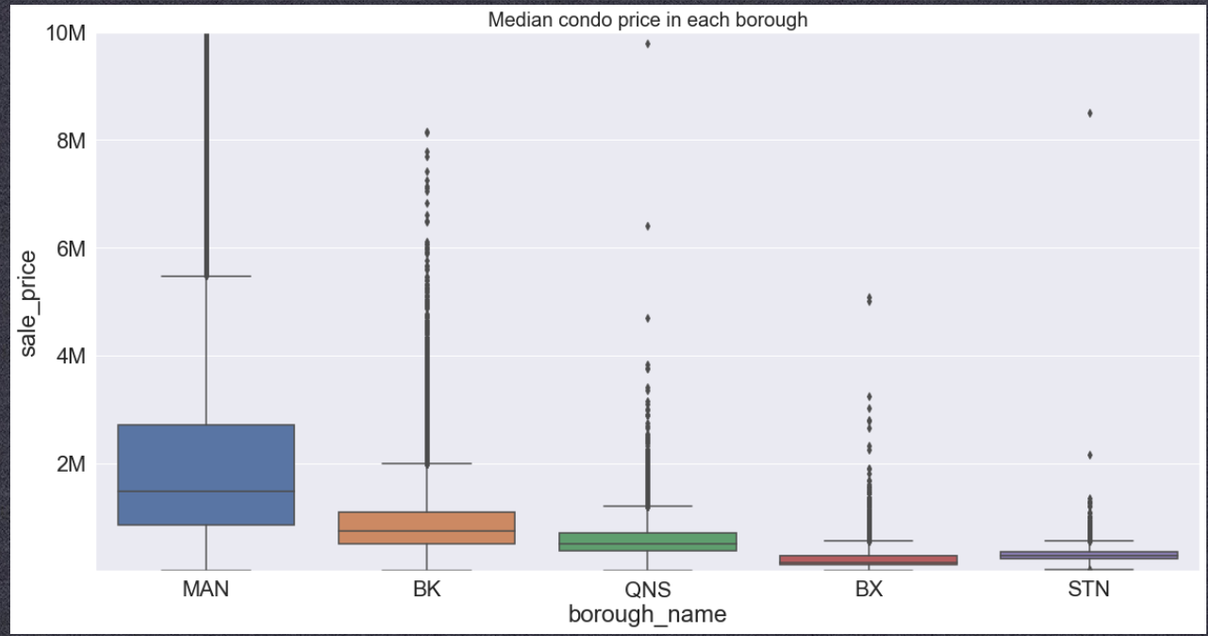


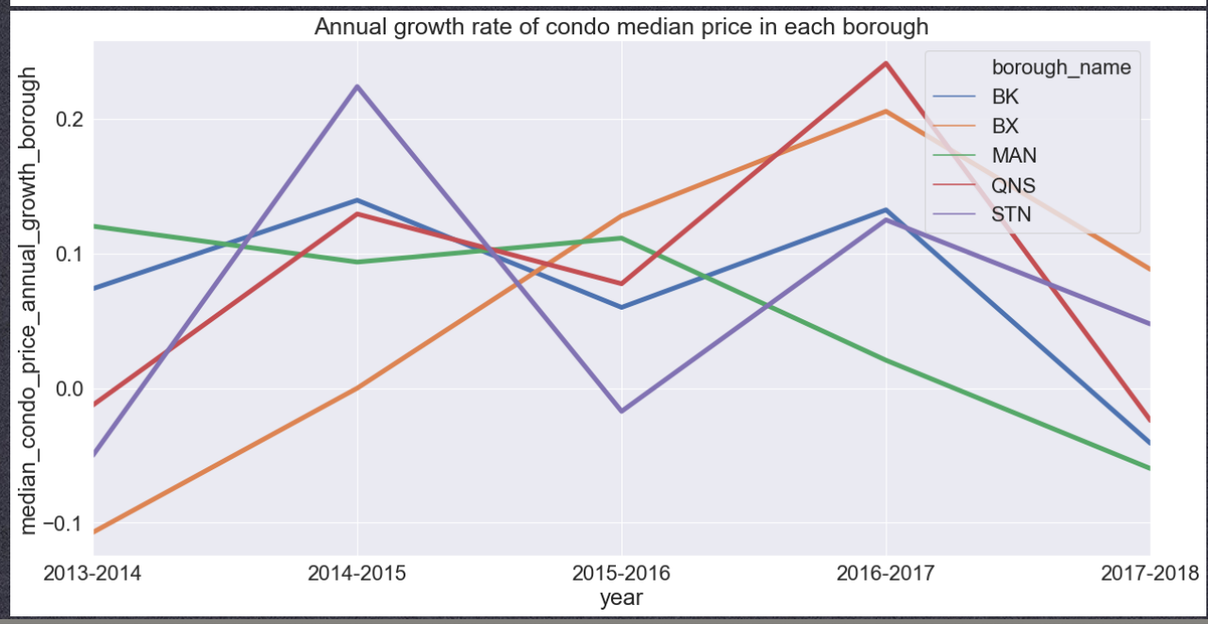


- From the line plot, we can see that the annual growth rate of median house price is comparably calm compare to the annual growth rate of amount of sales of house.
- The box plot showed that houses in Manhattan, with a median of 4.5M USD, are extremely expensive compare to house price in other boroughs.
- The third plot shows that the annual growth rate of house in Manhattan fluctuate the most compare to other boroughs. From the chart, houses in Brooklyn have the highest average annual median house price growth rate, 9%.

Statistics Inference & Story telling — Condo



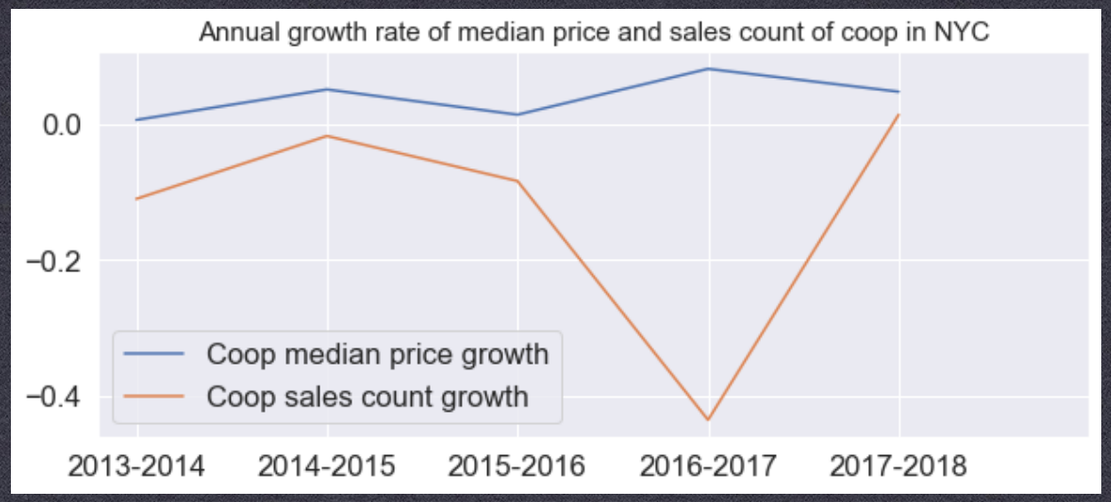


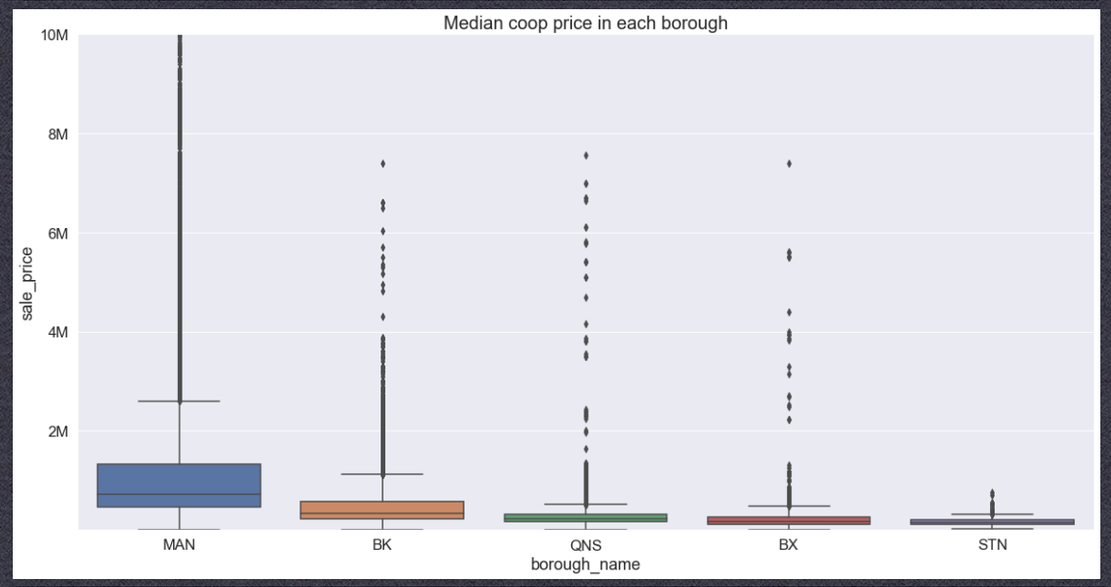


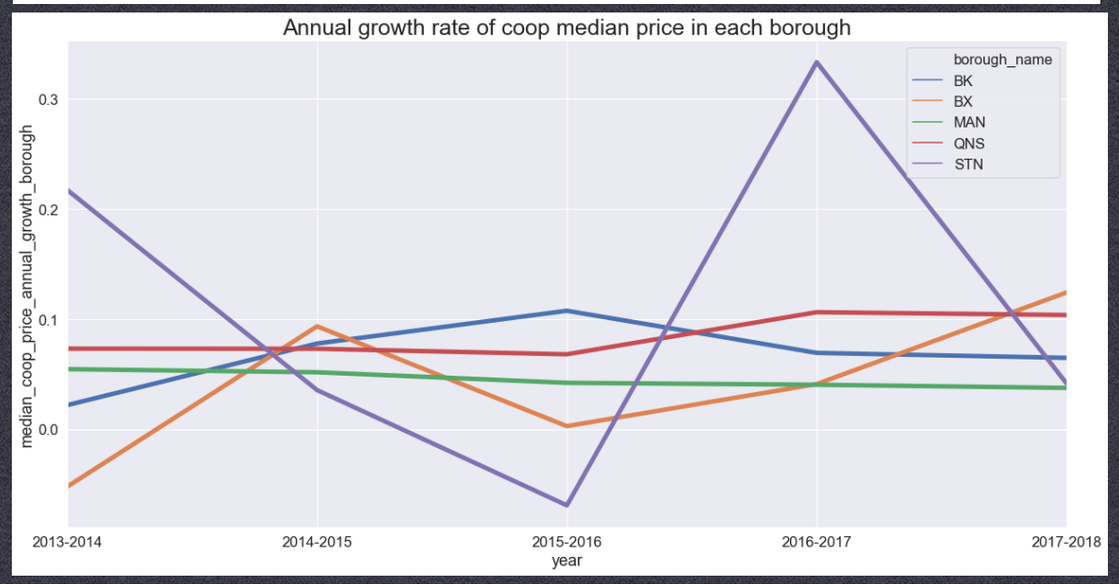
- **Same as house in NYC, the growth rate of amount of sales of condo in NYC seems to fluctuate more compare to annual growth rate of median price of condo.**
- **With a median price around 1.5M USD, condos in Manhattan are the most expensive compare to other boroughs, but the median price gap between Manhattan's condo and other boroughs' condo has been narrow down compare to the median price gap between houses in Manhattan and other boroughs.**
- **For annual median condo price growth rate, Manhattan has a downward trend, other boroughs have increase between year 2014 and 2015, 2016 and 2017. For Condo, Queens has the highest average annual median price growth rate, at 8%.**

```
borough_name
QNS     0.082336
BK      0.073004
STN     0.065896
BX      0.062953
MAN     0.057293
Name: condo_borough_median_growth_rate_avg
```

Statistics Inference & Story telling — Coop



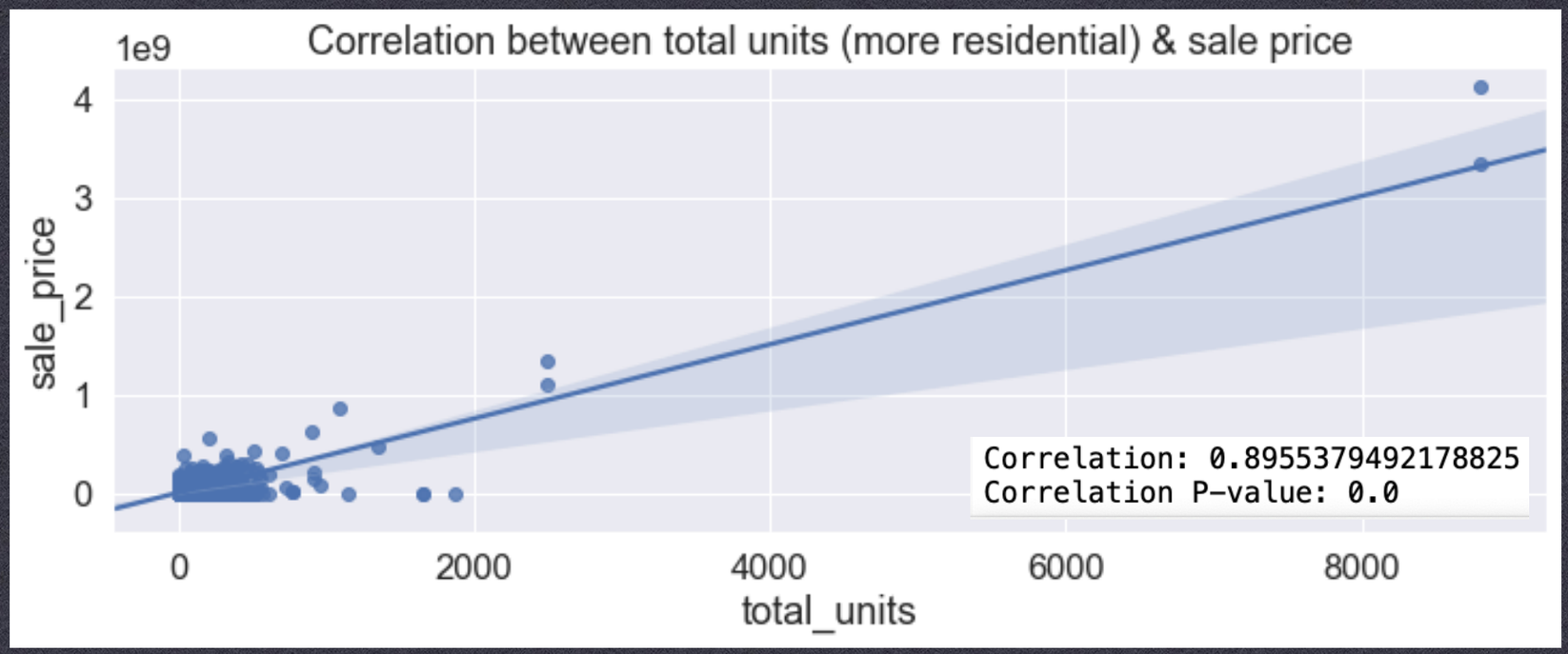


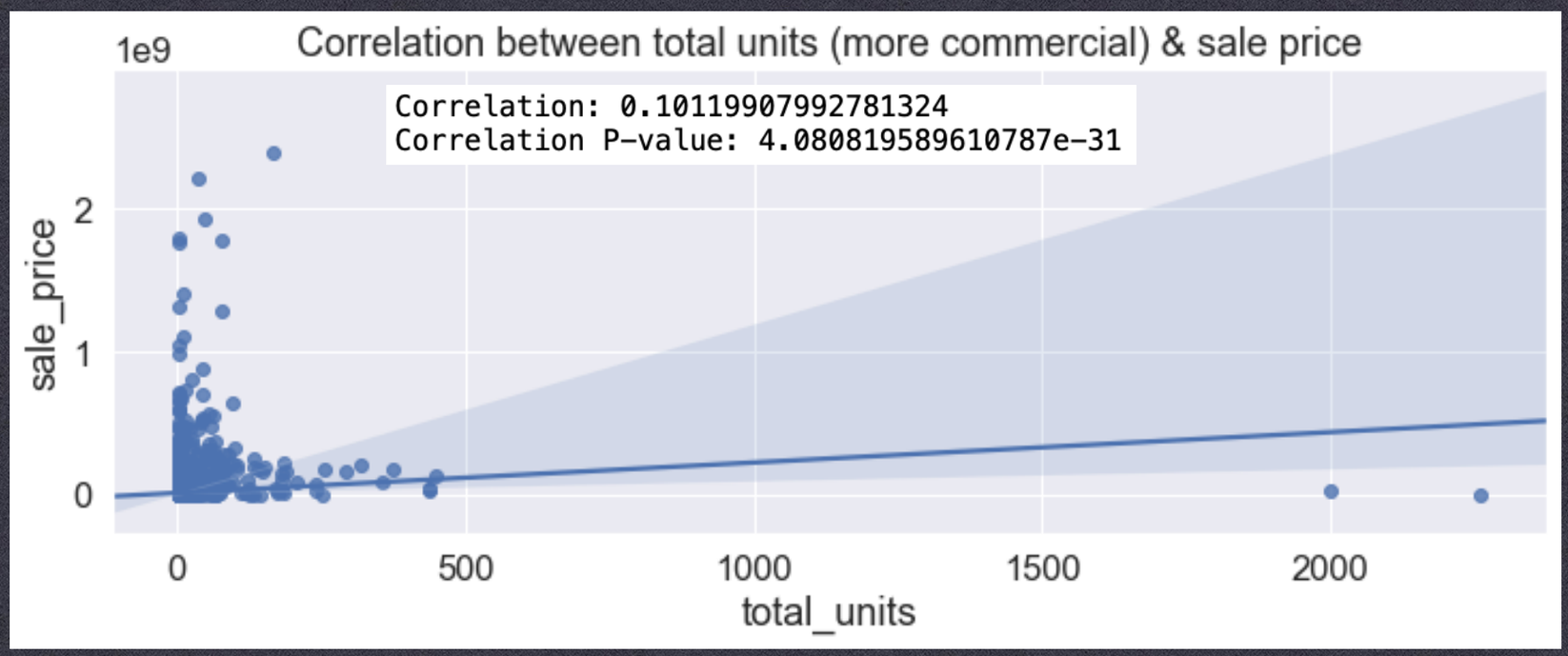


- **Same as house and condo in NYC, the growth rate of amount of sales of coop in NYC seems to fluctuate more compare to annual growth rate of median price of coop.**
- **With a median price around 800,000 USD, coops in Manhattan are the most expensive compare to other boroughs. However, their do not exist big gaps between median price of each boroughs' coops.**
- **For annual median coop price growth rate, Staten Island seemed to fluctuate the most compare to other boroughs. Staten Island, with an average 11%, is also the borough with the highest annual growth rate for the coop market.**

```
borough_name
STN     0.111828
QNS     0.084976
BK      0.068410
MAN     0.045399
BX      0.041996
Name: coop_borough_median_growth_rate_avg
```

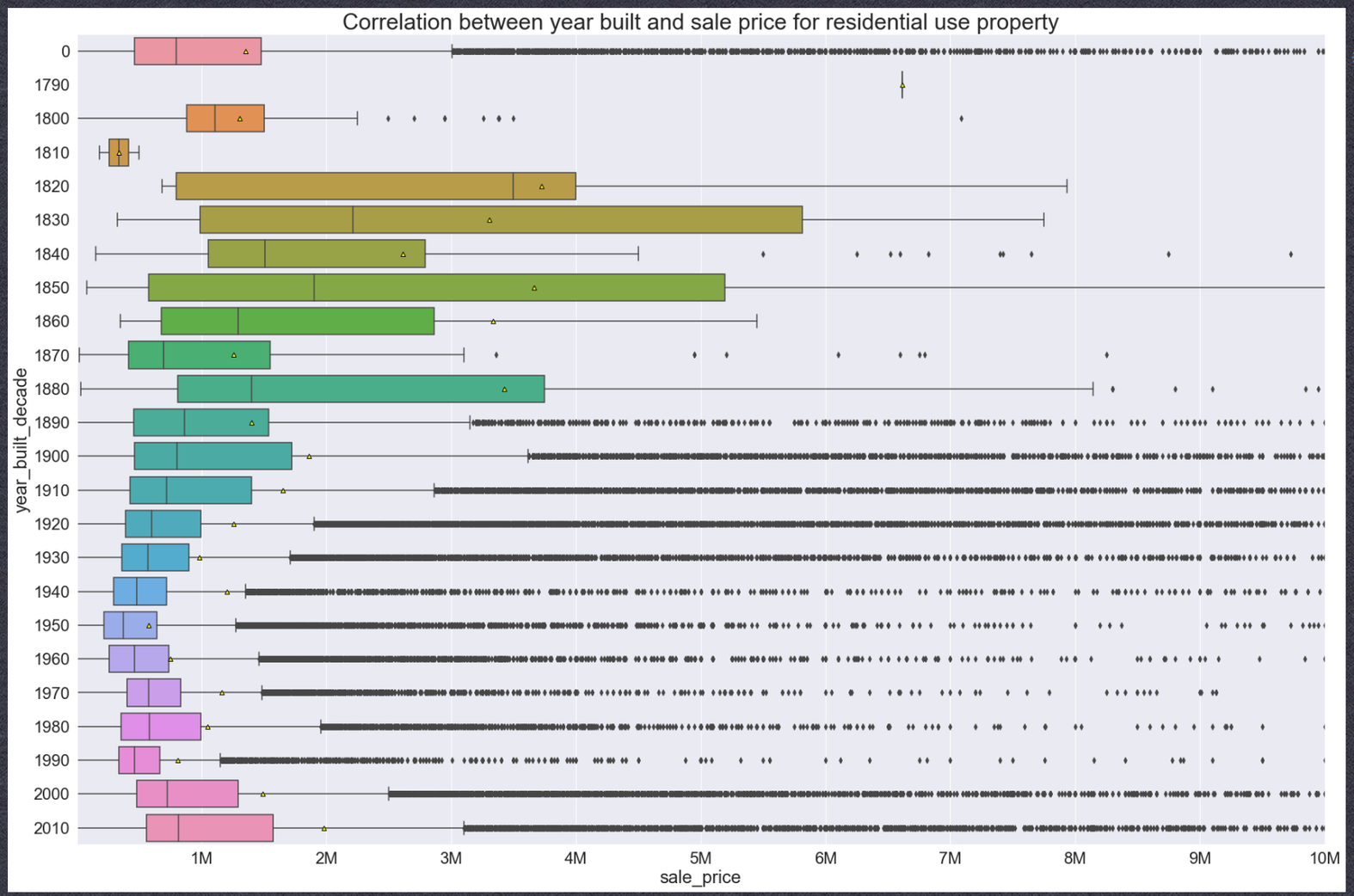
Statistics Inference & Story telling





- **I have separate total units into 2 categories, one with residential units > commercial units, vice versa. From the 2 plots on the left hand side we can see that the building with more residential units seems to be strongly correlated with the sale price compare to the buildings with more commercial units.**

Statistics Inference & Story telling

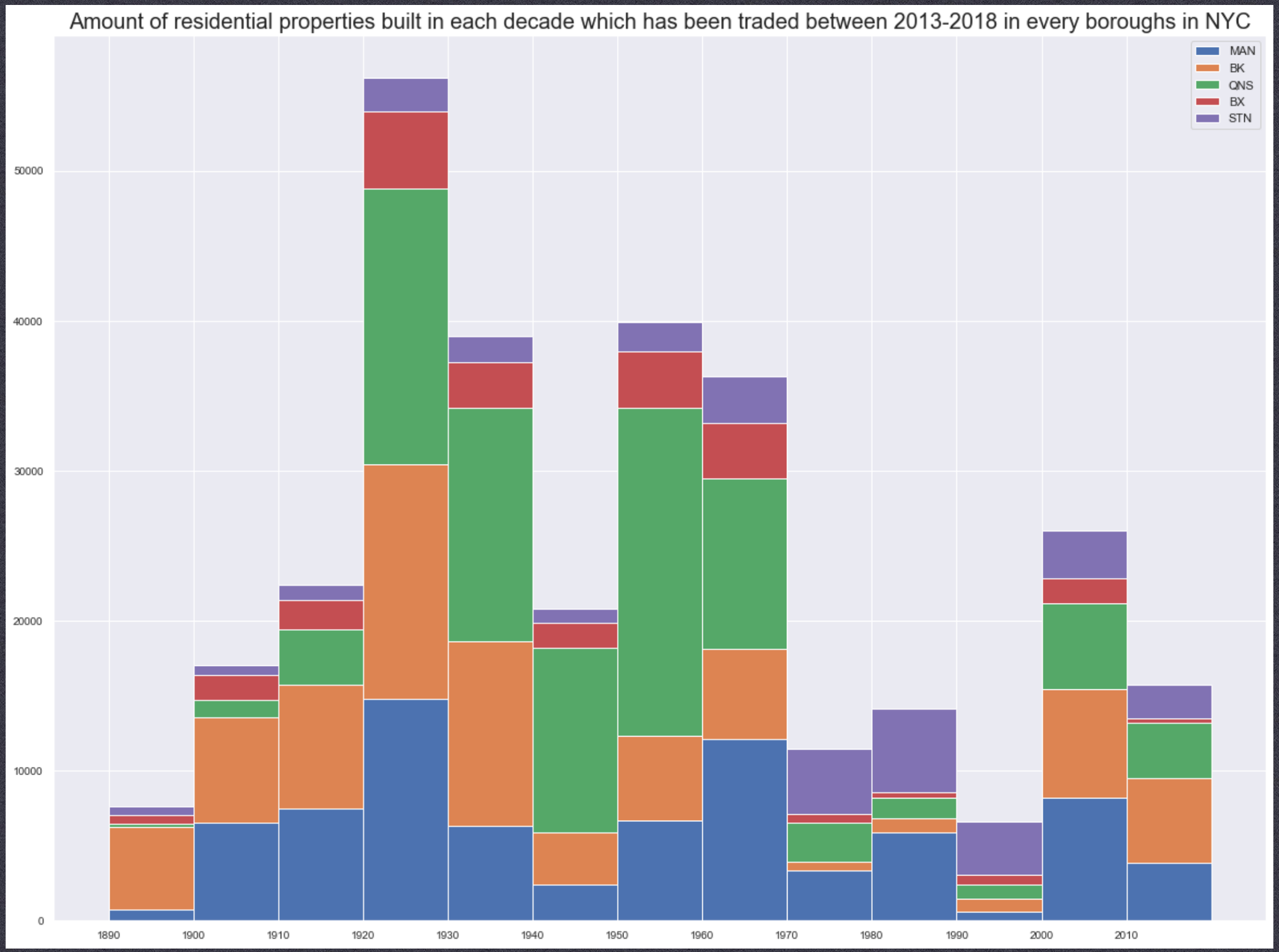


- From the plot on the left, we can see that newly built properties do not necessary results in higher price. Here, we are not taking properties built before 1890 in to account, since the sample size is fairly small compare to properties built after 1890, the high median price of buildings built before 1890 are due to lack of data. Anyway, properties built in the 1890s, 1990s, 2010s seems to have the highest median price in the market.

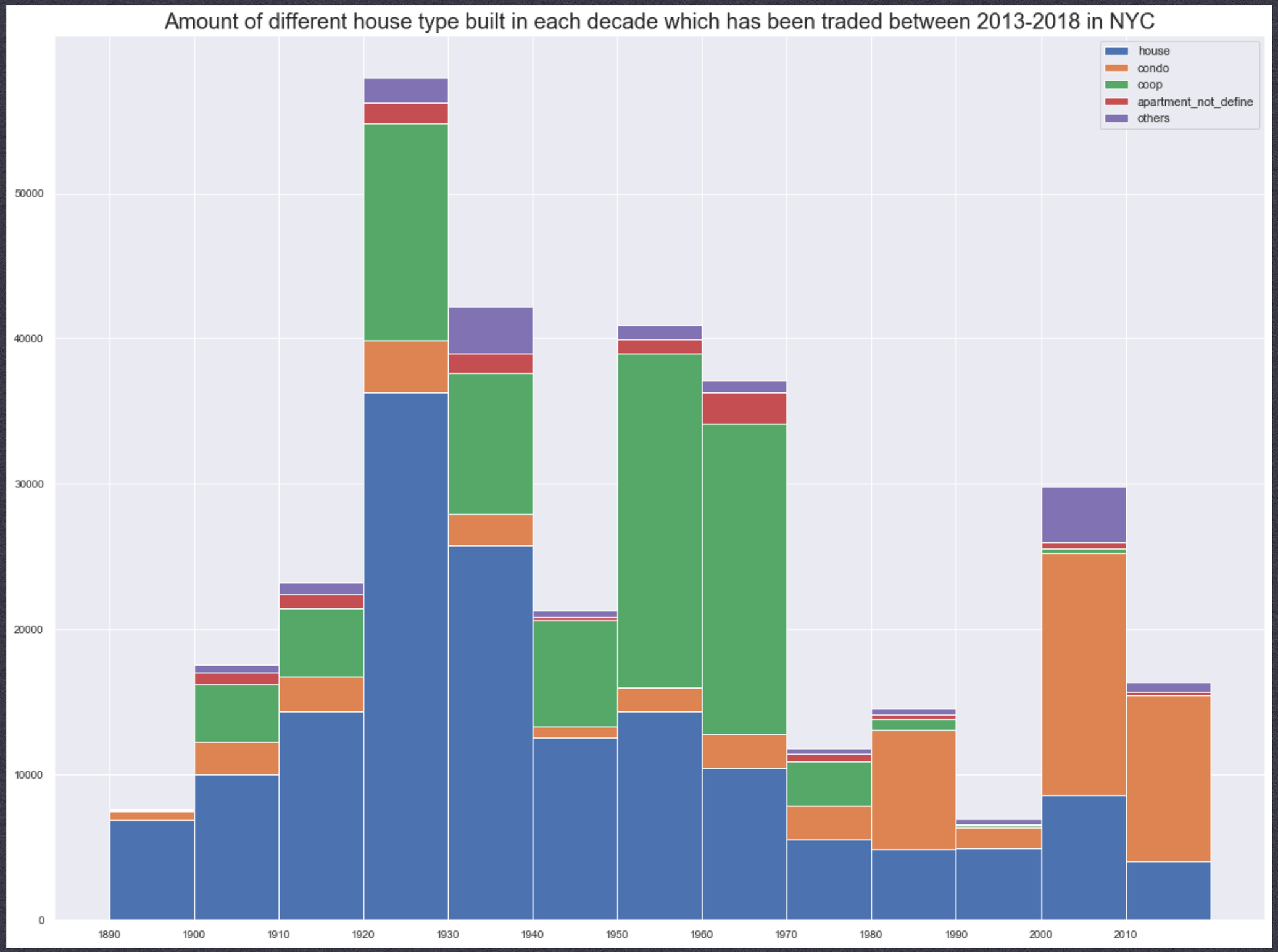
- From the plot on next page, we can see that houses built in the 1920s are the hottest in the market for each borough except for staten island, on the other hand, houses built in the 2000s is the least popular. There seems to have cycles of popularity of the house built in different decades, due to the forward moving time (a phenomenon worth researching).

- From plot on page 3, Coops, which are built before year 1970s are very popular compare to condos built before 1970s in 2013-2018 house market. On the other hand, condos built after 1980s seems to out popular coop for the house market.

Statistics Inference & Story telling



Statistics Inference & Story telling



Machine Learning

Regressors	Best R^2 Score
Linear Regression	0.2392
Random Forest Regressor	0.2626
Gradient Boosting Regressor	0.0856
Cat Boosting Regressor	0.1696
LightGBM Regressor	0.1852
XGBoost Regressor	0.0362

- After applying six kinds of regressors to various forms of datasets, by using different tuning methods and features selection, results are listed below. From the data frame we can see that Random Forest Regressor help achieved the highest R^2 score among all the regressors fitted. Since these results are not predicted by applying regressors with a given time line, I am going to use time series in machine learning part two, to see if there can be an improvement on results.

Future Work

- **Add variables like FOREX, and net immigrants of NYC for data analysis, and see if it helps improve the score for prediction.**
- **Apply machine learning algorithms to the data with no missing value, and see if there can be an improvement in score.**
- **Use time series for machine learning part 2, and change the target to prediction of monthly median house price for different house types in each borough.**

Acknowledgement

- **Mentors(Sorted_value): Dipanjan (DJ) Sarkar, Harsh Singh, Kenneth Gil-Pasquel**
- **Springboard**
- **NYC.gov**