Have Affordable Housing! Factors of Pricing Trend

CHEN-YUAN HO
Computer Science, Courant Institute of Mathematical Sciences, New York University
New York City, United States
cyh359@nyu.edu

***Abstract*—**

This project is to do analysis on the pricing of house rent and sales in the New York City. The analysis includes two sections, the one is the descriptive statistics, the other is regression model, which will be base of the pricing prediction application. The analysis is implemented on NYU Dumbo HDFS with tool of Spark and Spark SQL.

Keywords—analytics, New York City, Rent, House, Predictive Analytics, Linear Regression Model, Spark, SQL, Machine Learning

I. INTRODUCTION

To begin with, I will collect the following four historical data. The first is dataset of housing cost, including rental and sales price, in New York City. The second is dataset of CPI in New York City. The third is dataset of property tax rate in New York City. The fourth is dataset of employment rate. The fifth is dataset of GDP in New York City. Then, I move to the second step. I will use Spark in Hadoop environment to do ETL and data profiling. Then, I will run statistical tool, such as Chi-Square, Correlation and Coefficient in Spark to know the relationship among factors in housing cost. In the end, I will try to build a linear regression model for housing cost based on factors of CPI, property tax and employment rate.

II. MOTIVATION

Knowing the market price of house sales and rent is important for anyone wants to reside in the New York City, because living expense accounts for almost half of a person's annual spending. Especially, in some leasing office or real estate agent, negotiating price is available, and then making a deal on actual value can benefit buyers or leaseholders. I have experience in bargaining leasing price on my current apartment. I struggled with negotiation with leasing office, since I did not have clear understanding of real estate market. That is the reason that I initiate this project.

III. RELATED WORK

1. Related Work 1
   (1) Title:

   Collective Data Mining: A New Perspective Toward Distributed Data Analysis

   (2) Authors:

   Hillol Kargupta, Byung-Hoon Park, Daryl Hershberger, and Erik Johnson

   (3) Link:

   http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9A9FF7B826DC6F3B0A64F089A9CDA169?doi=10.1.1.34.4490&rep=rep1&type=pdf

   (4) Summary:

   Nowadays, in order to do data modeling and knowledge discovery, we have to have a centralized collection of data from distributed sites, which requires heavy data communication. Especially, when distributed data is heterogeneous, sites with data with different features, classical statistics and machine learning have to move data to a central site. Contrarily, the traditional approaches, like decomposable data modeling would bring in ambiguous and misleading under this distributed environment. This thesis introduces a new algorithm, Distributed Data Mining (DDM). A typical DDM algorithm works by performing local data analysis for generating partial data models and combining the local data models from different data sites in order to develop the global model.

   To search appropriate data model (constructing representation), we need to compute the basis coefficients by sparse representation and approximate evaluation of the significant coefficients, and by minimizing the mean-square error of model. In the paper, authors recommend decision tree model, CDM regression, and polynomial regression to achieve the above goal.

   This paper is beneficial to my project because it meets the properties of the data sources in my project. I collected 5 datasets from totally different sources. The feature and data schema are not identical. Through DDM algorithm, I can identify the significant features from labels and provide more precise regression model for evaluating the housing and rent prices in the New York City.

2. Related Work 2
   (1) Title:

Data mining: an overview from a database perspective

(2) Authors:

Ming-Syan Chen, Jiawei Han, P.S. Yu

(3) Link:

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=553155

(4) Summary:

Classification is the process which finds the common properties among a set of objects in a database and classifies them into different classes. To construct classification model, we need a training set consists the same set of multiple features. The objective of the classification is to first analyze the training data and develop an accurate description or a model for each class using the features available in the data. Such class descriptions are then used to classify future test data in the database or to develop a better description. There are two major classification methods: decision tree method and clustering.

Decision tree method is a supervised learning method. A typical decision tree learning system adopts a top-down irrevocable strategy that searches only part of the search space. It guarantees that a simple, but not necessarily the simplest, tree is found.

Clustering is the process of grouping physical or abstract objects into class of similar objects, which helps construct meaningful partition of a large set of objects on a divide-and-conquer methodology: decomposes a large scale of system into smaller components to simplify design and implementation.

In my project, it is likely to identifying the distribution of housing and rent pricing among factors by clustering. For example, it is possible that each borough in New York City has unique pattern for pricing factors. Clustering here is a optimal method to characterize it.

3. Related Work 3

(1) Title:

Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors.

(2) Authors:

Frank E. Harrell Jr., Kerry L. Lee, Daniel B. Mark

(3) Link:

https://onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291097-0258%2819960229%2915%3A4%3C361%3A%3AAID-SIM168%3E3.0.CO%3B2-4

(4) Summary:

The authors of this thesis introduce multivariable regression model to do accurate estimation of patient prognosis. Through this method, physicians can do analysis on factor significances in different cases and predict possible outcome of patient prognosis.

To achieve the goal, the authors adopted the following step. First, they did imputation of missing data, pre-specification of interaction and choosing the outcome model. Second, they did data reduction. Third, they validate whether the hypothesized model fits the data. Last, they measure the predictive accuracy.

In my project, the methodology in this thesis might be possible solution to determine the weights of factors for property pricing. It is possible to build up a predictive analysis based on the multivariable regression model which the thesis implements.

4. Related Work 4

(1) Title:

Data Mining Applications in Healthcare

(2) Authors:

Hian Chye Koh, Gerald Tan

(3) Link:

https://www.researchgate.net/publication/7869635_Data_Mining_Applications_in_Healthcare

(4) Summary:

Data mining can be seen as the process of finding previously unknown pattern and trends in databases and using this information to build up predictive models.

The methodology for data mining in business world is following five phases: business understanding, data understanding and preparation, modeling, evaluation, and deployment.

In healthcare, data mining is increasingly prominent. There are following four major data mining applications in healthcare. For example, when new case occurs, data scientists can compare the data of the case with the historical data to predict

the possible outcome of patients, which is beneficial for physicians to practice precise treatment, or for hospital to prevent from fraud or abuse.

In my project, I will adopt the Bayesian classification model and confusion matrix which the case in the paper employs to do prediction analytics of sales and rent pricing. Then, by training the given historical datasets, I will do cross-validation on new sales or rent data from other sources. By doing so, I can enhance the precision of my prediction model.

5. Related Work 5

(1) Title:

Data quality for data science, predictive analytics, and big data in supply chain management: An introduction

to the problem and suggestions for research and applications.

(2) Authors:

Benjamin T. Hazen, Christopher A. Boone, Jeremy D. Ezell, L. Allison, Jones-Farmer

(3) Link:

https://www.sciencedirect.com/science/article/pii/S0925527314001339

(4) Summary:

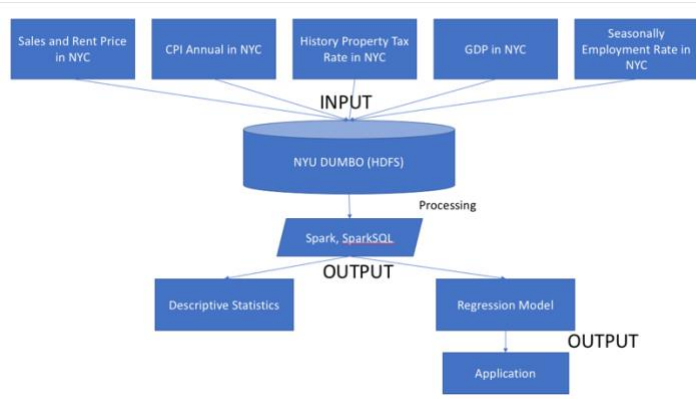
The authors put emphasis on how to maintain high standard of data quality. They suggest that data quality is comprised of two major dimensions: intrinsic and contextual dimensions. While intrinsic dimension is about attributes that are objective and native to the data, contextual dimension is about attributes that are dependent on the context in which the data are observed or used, including relevancy, value-added, quantity, believability, accessibility, and reputation of the data.

To maintain the data quality, in past, data scientists adopted fishbone diagram, Pareto charts, and histogram is simple but effective tools to clean up data production process. Contrarily, the authors propose Standard Process Control (SPC), which is common in supply chain management (SCM), to deal with this issue. Under the concept of SCM, like manufacturing process of a product, they focus on the data production process, including data collection, storage, retrieval, and processing. For example, in the dataset of jet engine compressors, they examined

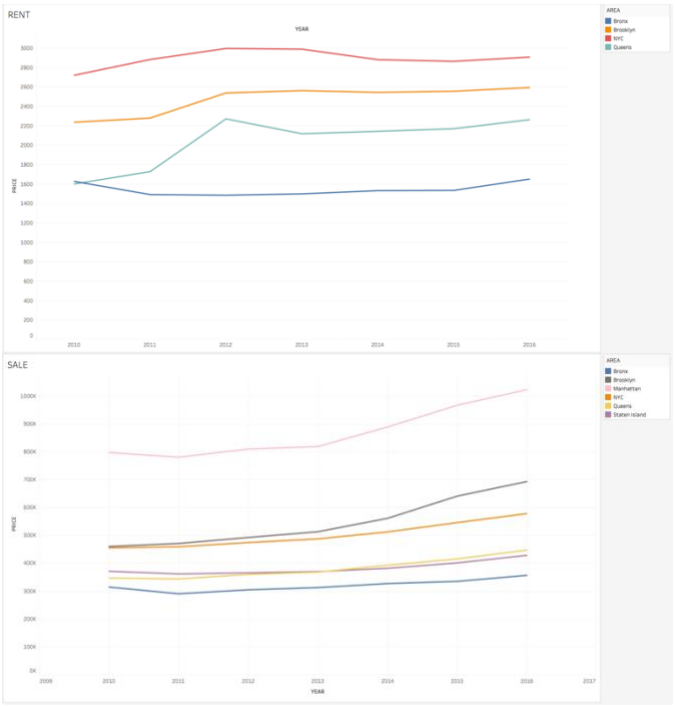
the completeness of data by SPC chart to monitor. Then, we identified 448 records of data were incomplete.

In my project, since I collect five datasets from different website, the robustness and completeness of raw data is important. Adopting the idea in the paper is helpful for me to ensure less garbage-in-garbage-out risks. I will monitor all datasets from collecting, to storage, to retrieval, and analysis.

IV. Application Design



Store collected datasets as input in NYU Dumbo HDFS, including Sales & Rent Price, CPI, Property Tax, GDP, Employment. Then, do ETL and analysis by Spark & SparkSQL. Finally, generate two kind of output files, one is descriptive statistics, the other is regression model of pricing. Finally, based on the regression model, develop application for user to predict the sales or rent price in given area according to input parameters. Besides, the project will do visualization about the price trend of both sales and rent in New York City, as the following two charts:



V. Datasets

There are five datasets in this project.

1. Real Estates Sales and Rent Price

(1) Source

https://streeteasy.com/blog/download-data/

(2) Schema

Rent Schema (Count = 70)

root

|-- Area: string (nullable = true)

|-- Boro: string (nullable = true)

|-- AreaType: string (nullable = true)

|-- 2010-01: float (nullable = true)

|-- 2010-02: float (nullable = true)

|-- 2010-03: float (nullable = true)

|-- 2010-04: float (nullable = true)

|-- 2010-05: float (nullable = true)

|-- 2010-06: float (nullable = true)
|-- 2010-07: float (nullable = true)
|-- 2010-08: float (nullable = true)
|-- 2010-09: float (nullable = true)
|-- 2010-10: float (nullable = true)
|-- 2010-11: float (nullable = true)
|-- 2010-12: float (nullable = true)
|-- 2011-01: float (nullable = true)
|-- 2011-02: float (nullable = true)
|-- 2011-03: float (nullable = true)
|-- 2011-04: float (nullable = true)
|-- 2011-05: float (nullable = true)
|-- 2011-06: float (nullable = true)
|-- 2011-07: float (nullable = true)
|-- 2011-08: float (nullable = true)
|-- 2011-09: float (nullable = true)
|-- 2011-10: float (nullable = true)
|-- 2011-11: float (nullable = true)
|-- 2011-12: float (nullable = true)
|-- 2012-01: float (nullable = true)
|-- 2012-02: float (nullable = true)
|-- 2012-03: float (nullable = true)
|-- 2012-04: float (nullable = true)
|-- 2012-05: float (nullable = true)
|-- 2012-06: float (nullable = true)
|-- 2012-07: float (nullable = true)
|-- 2012-08: float (nullable = true)
|-- 2012-09: float (nullable = true)
|-- 2012-10: float (nullable = true)
|-- 2012-11: float (nullable = true)
|-- 2012-12: float (nullable = true)
|-- 2013-01: float (nullable = true)
|-- 2013-02: float (nullable = true)
|-- 2013-03: float (nullable = true)
|-- 2013-04: float (nullable = true)
|-- 2013-05: float (nullable = true)
|-- 2013-06: float (nullable = true)
|-- 2013-07: float (nullable = true)
|-- 2013-08: float (nullable = true)
|-- 2013-09: float (nullable = true)
|-- 2013-10: float (nullable = true)
|-- 2013-11: float (nullable = true)
|-- 2013-12: float (nullable = true)
|-- 2014-01: float (nullable = true)
|-- 2014-02: float (nullable = true)
|-- 2014-03: float (nullable = true)
|-- 2014-04: float (nullable = true)
|-- 2014-05: float (nullable = true)
|-- 2014-06: float (nullable = true)
|-- 2014-07: float (nullable = true)
|-- 2014-08: float (nullable = true)
|-- 2014-09: float (nullable = true)
|-- 2014-10: float (nullable = true)
|-- 2014-11: float (nullable = true)
|-- 2014-12: float (nullable = true)
|-- 2015-01: float (nullable = true)
|-- 2015-02: float (nullable = true)
|-- 2015-03: float (nullable = true)
|-- 2015-04: float (nullable = true)
|-- 2015-05: float (nullable = true)
|-- 2015-06: float (nullable = true)
|-- 2015-07: float (nullable = true)
|-- 2015-08: float (nullable = true)
|-- 2015-09: float (nullable = true)
|-- 2015-10: float (nullable = true)
|-- 2015-11: float (nullable = true)
|-- 2015-12: float (nullable = true)
|-- 2016-01: float (nullable = true)
|-- 2016-02: float (nullable = true)
|-- 2016-03: float (nullable = true)
|-- 2016-04: float (nullable = true)
|-- 2016-05: float (nullable = true)
|-- 2016-06: float (nullable = true)
|-- 2016-07: float (nullable = true)
|-- 2016-08: float (nullable = true)
|-- 2016-09: float (nullable = true)
|-- 2016-10: float (nullable = true)
|-- 2016-11: float (nullable = true)

|-- 2016-12: float (nullable = true)
|-- 2017-01: float (nullable = true)
|-- 2017-02: float (nullable = true)
|-- 2017-03: float (nullable = true)
|-- 2017-04: float (nullable = true)
|-- 2017-05: float (nullable = true)
|-- 2017-06: float (nullable = true)
|-- 2017-07: float (nullable = true)
|-- 2017-08: float (nullable = true)
|-- 2017-09: float (nullable = true)
|-- 2017-10: float (nullable = true)
|-- 2017-11: float (nullable = true)
|-- 2017-12: float (nullable = true)

Sale Schema (Count = 62)

root
|-- Area: string (nullable = true)
|-- Boro: string (nullable = true)
|-- AreaType: string (nullable = true)
|-- 2010-01: float (nullable = true)
|-- 2010-02: float (nullable = true)
|-- 2010-03: float (nullable = true)
|-- 2010-04: float (nullable = true)
|-- 2010-05: float (nullable = true)
|-- 2010-06: float (nullable = true)
|-- 2010-07: float (nullable = true)
|-- 2010-08: float (nullable = true)
|-- 2010-09: float (nullable = true)
|-- 2010-10: float (nullable = true)
|-- 2010-11: float (nullable = true)
|-- 2010-12: float (nullable = true)
|-- 2011-01: float (nullable = true)
|-- 2011-02: float (nullable = true)
|-- 2011-03: float (nullable = true)
|-- 2011-04: float (nullable = true)
|-- 2011-05: float (nullable = true)
|-- 2011-06: float (nullable = true)
|-- 2011-07: float (nullable = true)
|-- 2011-08: float (nullable = true)
|-- 2011-09: float (nullable = true)
|-- 2011-10: float (nullable = true)
|-- 2011-11: float (nullable = true)
|-- 2011-12: float (nullable = true)
|-- 2012-01: float (nullable = true)
|-- 2012-02: float (nullable = true)
|-- 2012-03: float (nullable = true)
|-- 2012-04: float (nullable = true)
|-- 2012-05: float (nullable = true)
|-- 2012-06: float (nullable = true)
|-- 2012-07: float (nullable = true)
|-- 2012-08: float (nullable = true)
|-- 2012-09: float (nullable = true)
|-- 2012-10: float (nullable = true)
|-- 2012-11: float (nullable = true)
|-- 2012-12: float (nullable = true)
|-- 2013-01: float (nullable = true)
|-- 2013-02: float (nullable = true)
|-- 2013-03: float (nullable = true)
|-- 2013-04: float (nullable = true)
|-- 2013-05: float (nullable = true)
|-- 2013-06: float (nullable = true)
|-- 2013-07: float (nullable = true)
|-- 2013-08: float (nullable = true)
|-- 2013-09: float (nullable = true)
|-- 2013-10: float (nullable = true)
|-- 2013-11: float (nullable = true)
|-- 2013-12: float (nullable = true)
|-- 2014-01: float (nullable = true)
|-- 2014-02: float (nullable = true)
|-- 2014-03: float (nullable = true)
|-- 2014-04: float (nullable = true)
|-- 2014-05: float (nullable = true)
|-- 2014-06: float (nullable = true)
|-- 2014-07: float (nullable = true)
|-- 2014-08: float (nullable = true)
|-- 2014-09: float (nullable = true)
|-- 2014-10: float (nullable = true)
|-- 2014-11: float (nullable = true)
|-- 2014-12: float (nullable = true)

|-- 2015-01: float (nullable = true)

|-- 2015-02: float (nullable = true)

|-- 2015-03: float (nullable = true)

|-- 2015-04: float (nullable = true)

|-- 2015-05: float (nullable = true)

|-- 2015-06: float (nullable = true)

|-- 2015-07: float (nullable = true)

|-- 2015-08: float (nullable = true)

|-- 2015-09: float (nullable = true)

|-- 2015-10: float (nullable = true)

|-- 2015-11: float (nullable = true)

|-- 2015-12: float (nullable = true)

|-- 2016-01: float (nullable = true)

|-- 2016-02: float (nullable = true)

|-- 2016-03: float (nullable = true)

|-- 2016-04: float (nullable = true)

|-- 2016-05: float (nullable = true)

|-- 2016-06: float (nullable = true)

|-- 2016-07: float (nullable = true)

|-- 2016-08: float (nullable = true)

|-- 2016-09: float (nullable = true)

|-- 2016-10: float (nullable = true)

|-- 2016-11: float (nullable = true)

|-- 2016-12: float (nullable = true)

|-- 2017-01: float (nullable = true)

|-- 2017-02: float (nullable = true)

|-- 2017-03: float (nullable = true)

|-- 2017-04: float (nullable = true)

|-- 2017-05: float (nullable = true)

|-- 2017-06: float (nullable = true)

|-- 2017-07: float (nullable = true)

|-- 2017-08: float (nullable = true)

|-- 2017-09: float (nullable = true)

|-- 2017-10: float (nullable = true)

|-- 2017-11: float (nullable = true)

|-- 2017-12: float (nullable = true)

2. Customer Price Index (CPI) Annual Index

   (1) Source:

   http://www.baruch.cuny.edu/nycdata/consumer_prices/cpi.htm

   Parsed by Python script.

   (2) Schema: (Count = 22)

   root

   |-- Year: integer (nullable = true)

   |-- US-CPI: float (nullable = true)

   |-- Percent Change Over Period US: float (nullable = true)

   |-- CPI-NYC: float (nullable = true)

   |-- Percent Change Over Period NYC: float (nullable = true)

3. Property Tax History Data

   (1) Source: (Count = 36)

   http://www1.nyc.gov/site/finance/taxes/property-tax-rates.page

   Parsed by Python script.

   (2) Schema:

   root

   |-- YEAR: integer (nullable = true)

   |-- CLASS 1: float (nullable = true)

   |-- CLASS 2: float (nullable = true)

   |-- CLASS 3: float (nullable = true)

   |-- CLASS 4: float (nullable = true)

4. GDP of the New York City from 2001 to 2016

   (1) Source: (Count = 16)

   https://www.statista.com/statistics/183815/gdp-of-the-new-york-metro-area/

   (2) Schema:

   root

   |-- Year: integer (nullable = true)

   |-- GDP_Billion: float (nullable = true)

5. NYU Seasonally Adjusted Total Employment from 1990 to 2017

   (1) Source: (Count = 506)

   https://www.labor.ny.gov/stats/nyc/

   (2) Schema:

   root

   |-- Mo-Yr: string (nullable = true)

   |-- Labor Force in K: float (nullable = true)

   |-- Employment in K: float (nullable = true)

   |-- Emp/Pop percent: float (nullable = true)

|-- Unemployed in K: float (nullable = true)

|-- Unemp Rate percent: float (nullable = true)

|-- LFPART percent: float (nullable = true)

VI. Remediation

We could gain insight from linear regression model to know factors affect the pricing of house sales and rents. If the real price of sales or rent is relatively higher than the result prediction model, we seek for other opportunity which the price will be lower to make a deal, and vice versa. This remediation is automated by linear regression model.

VII. Experiments

1. Hypothesis:

   According to the thesis and economic concept I have learned, I assume that CPI, GDP, employment rate, and property tax rate put influence on the price of real estate sales and rental in New York City. Besides, there is positive relationship to pricing in the first three factors, while there is negative relationship between property tax rate and pricing. Then, I collected all of the corresponding datasets for data mining. In order to check whether my assumption is right or wrong, as the linear regression model setting mentioned in the thesis of "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors."[3], the project adopts linear regression model to fit the coefficients of each factor. Furthermore, according to the absolute value of the coefficient for each factor, I can determine the most significant factor to affect the price of sales and rent in New York City.

2. Dataset

   After collecting all of the dataset, illustrated in the previous section, I did extract-transform-load first to clean messy datasets.

   We set up the analysis based on annual data. However, not all of datasets log entries on annual basis. The sales and rental price dataset and the employment dataset record on monthly basis. To make all of the data infrastructure consistent, I merge rows of months with same years into corresponding rows of years.

3. Tools

   I stored all of raw data in NYU Dumbo HDFS. To do data ETL and analysis, I utilized Spark, Spark SQL, and Spark ML package. Furthermore, for data visualization, I employ tableau to complete the job.

4. Limitation

   (1) The datasets of rental and sales price is limited to median price of each borough in New York City, which made me merely do rough analysis on housing market.

   (2) There is not detail data about price for each type of house, like apartment, studio, mansion.

   (3) I cannot collect all of factors which may lay influence on pricing. The only thing that I can do is to collect factors according to result of relative research previously.

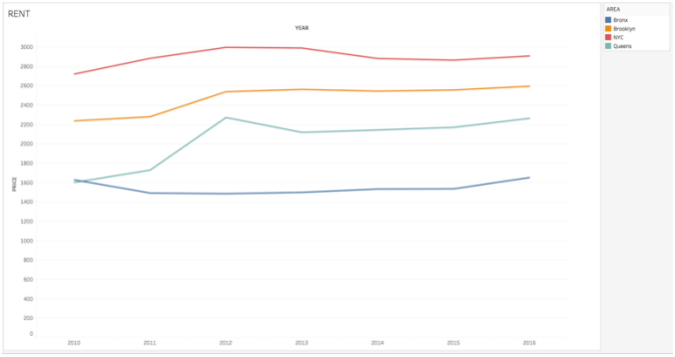
VIII. Result

1. Rent Price Trend

The following table is median rent price in each borough of New York City from 2010 to 2016. The standard deviation of rental pricing in New York City is around 5738.48007.

Area / Year	Year						
Area	2010	2011	2012	2013	2014	2015	2016
Bronx	1628.125	1491.875	1485.041667	1498.958333	1532.666667	1534.958333	1651.708333
Brooklyn	2236.875	2280	2538.541667	2562.458333	2544.5	2556.25	2594.166667
Manhattan	2877.333333	3100	3194.583333	3218.083333	3215.833333	3314.333333	3335.583333
NYC	2720.583333	2882.916667	2996.166667	2989.083333	2881.166667	2864.083333	2907.5
Queens	1601.041667	1728.541667	2271.75	2118.75	2143.458333	2170.416667	2263.416667

As we see in the following plots, the rent price had grown not significantly from 2010 to 2016 significantly other than Queens.

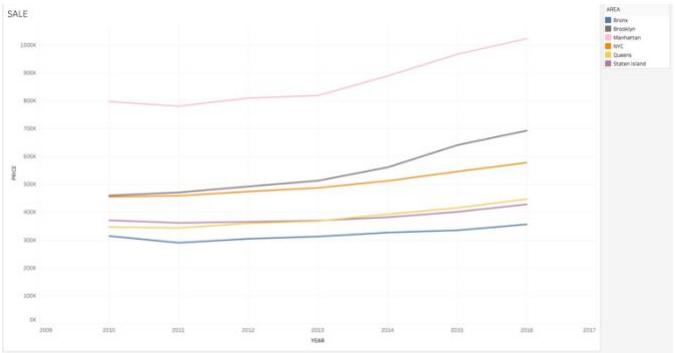


2. Sale Price Trend

The following table is median sales price in each borough of New York City from 2010 to 2016. The standard deviation of sales pricing in New York City is around 42839.75701

Area / Price	Year						
Area	2010	2011	2012	2013	2014	2015	2016
Bronx	313957.9167	289825	304049.1667	312008.7083	326508.3333	334354.1667	355907.1667
Brooklyn	459521.0833	470236	491813.4583	512437.5	560589.4583	640198.7917	692450
Manhattan	796458.8333	779778.0833	809120.3333	817944.1667	888234.0833	966077.9167	1022211.25
NYC	454565.5	458553.125	473598.7083	486798.25	511844.0833	545202.7083	577583.3333
Queens	346505.4167	343005.25	359406.9167	367825	392277.625	414924.3333	446665.9583

As we see in the following plots, the sales price trend had been increasing from 2010 to 2016.



3. Linear Regression Prediction Model

To form up prediction model, I run linear regression model by means of importing Spark ML packages. The following is the configuration of prediction model.

Max Iteration	10
Regression Parameter	0.3
Elastic Net Parameter	0.8

The following is the prediction models for house sales and rental price.

$$Prediction_{Sales\ Price} = -208458.94468347 + 1091.4270482783807 \times \text{CPI} + 165.7274091063199 \times \text{GDP} + 10399.662279225112 \times \text{EmploymentRate} - 57269.21486119139 \times \text{PropertyTaxRate}$$

$$Prediction_{Rental\ Price} = 239.77645898174342 + 10.589805354979298 \times CPI + 0.031920256159639726 \times GDP - 32.26363655775778 \times EmploymentRate + 213.30694059291323 \times PropertyTaxRate$$

IX. Conclusion

1. The price growth of Sales market is much more significant than that of Rental Market from 2010 to 2016 in New York City.
2. Property tax rate weighs the most in pricing of both sales and rent. While the former is negative relationship, the latter is positive. This phenomenon may indicate that if the property tax rate is higher, New Yorkers tends not to buy owned house and rent it instead, and vice versa.

X. Future work

In the future, I will do following two works. The one is to keep track of the precision of the prediction model through input the parameters of CPI, GDP, employment rate, and tax rate in the following years. If the mean squared error is more than 5% between real price and predicted price, I will modify the model. The other thing is to collect other possible factors which affects the pricing and do try-and-error to build up new model.

XI. Acknowledgment

Suzanne McIntosh, Courant Institute of Mathematical Sciences, New York University.

(1) New York City Government.
(2) Department of Labor, New York State Government.

XII. References

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. Linear Regression, Wikipedia, 2018.
3. Classification and Regression, Apache Spark 2.2.0, 2018.
4. Collective Data Mining: A New Perspective Toward Distributed Data Analysis, Hillol Kargupta, Byung-Hoon Park, Daryl Hershberger, and Erik Johnson, 2012.
5. Data mining: an overview from a database perspective, Ming-Syan Chen, Jiawei Han, P.S. Yu, 1996.
6. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors., Frank E. Harrell Jr., Kerry L. Lee, Daniel B. Mark, 1996.
7. Data Mining Applications in Healthcare, Hian Chye Koh, Gerald Tan, 2005.
8. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications., Benjamin T. Hazen, Christopher A. Boone, Jeremy D. Ezell, L. Allison, Jones-Farmer, 2014.