Median Asking Rent in New York

Britton Blaize Olle

Fordham University, New York,
United States

Phone: 970-688-1777
Email: Bolle@fordham.edu

Contents

List of Figures

Abstract.
This is an analytical paper on the monthly median asking rent of New York City from 2010 to 2018. It is written in a question then answer format. For additional reference I have attached a Jupiter Notebook file with all the code and analysis used.

1 Retrieval

1.1 *Where is the data from?*

The data set, "E2 Median Asking Rent", was pulled from street easy.

2 Initial Look Up

2.1 *How big is the data set?*

The data set has 202 different 'Areas' each with 100 months of rent data.

2.2 *Are there any duplicates?*

There are no duplicate rows, meaning that the data set was generally exported correctly.

2.3 *Is there any missing data?*

Yes, there are 8,813 missing observations.

2.4 *Any issues with data typing?*

There are two types of data; integers, that act as the median monthly rent in dollar amounts, and strings, that categorize the locations into: "Area", "Boro", and "AreaType".

The blanks and gaps between months are filled with NA.

2.5 *Handle all issues and explain what you are doing.*

The biggest problem is the missing data, as mentioned before there is 8,813 monthly rent prices missing. The missing months aren't distributed evenly across locations. For example, some areas are missing the first year and then nothing else. Others are missing a single month, while some areas have no rent data all together. This issue is not easily remedied and could undertake an analysis project of its own.

It would make the most sense to find the distribution of the month to month changes in rent. Then, fill the missing rent prices by drawing randomly from the distribution and applying it to the available months, thus simulating the missing data instead of filling the observations with arbitrary prices.

For now though, any months with missing rent values will be assumed to have remained unchanged from the most recent valid past month. If their very first month is not valid I will assume that it is equal to the first month with a valid rent value. Locations with no rent data were removed from the data set entirely. This strategy of filling requires there to be at least one observation to use as a base rent amount.

2.5.1 Justification for filling missing values.

This specific issue doesn't have an easy solution. If I remove the rows with NA values, then I miss out on valuable data that could provide significant insight. Additionally, I bias the population by deleting neighborhoods from the data set.

I face a similar issue if I fill the data. Clearly it wouldn't make sense to fill the missing data with 0, given median rent in that area likely didn't drop to $0 for a single month. This led me to make the assumption that the most reasonable way to fill the data would be to use the previous month's rent. The change month to month is generally minimal. Of course this will bias the data to assume rent is much more constant than it actually is. While this

isn't ideal, I believe for now it is the best option for analyzing the population as a whole.

2.5.2 Additional analysis on NA data

I think it could be useful to separate the NA data from the rest of the data to see if there is anything telling about the locations that go underreported.

For example I did some quick analysis and found that Queens and the Bronx appear to be the most underreported, while Manhattan appears to be the most reported.

3 Wrangling

3.1 *What is the current index on the DataFrame? Do you like it?*

Currently, the default index created by Pandas is being used. While this process does ensure uniqueness in the index, this index, in my mind, is not unique to the structure of the data.

3.2 *Would you create a different index and/ or orientation for this DataFrame? Why or why not?*

Yes, I will create a different orientation. The descriptive variables all specify location. 'AreaType' being the most general, followed by 'Boro', with 'Area' as most specific. With this in mind I will replace the current index with the descriptive variables. This will allow me to organize the data as specifically or as generally as I like.

			2010-01	2010-02	2010-03	2010-04	2010-05	2010-06	2010-07	2010-08
AreaType	Boro	Area								
submarket	Manhattan	All Downtown	3200.0	3200.0	3050.0	3100.0	3100.0	3200.0	3195.0	3200.0
		All Midtown	2895.0	2800.0	2800.0	2850.0	2900.0	2950.0	3000.0	3000.0
		All Upper East Side	2459.5	2450.0	2400.0	2500.0	2550.0	2575.0	2595.0	2500.0
		All Upper Manhattan	1825.0	1810.0	1795.0	1800.0	1823.0	1850.0	1875.0	1850.0
		All Upper West Side	2895.0	2800.0	2750.0	2800.0	2800.0	2795.0	2800.0	2875.0

Fig. 1. New DataFrame Index

4 Analysis

4.1 *Which Area saw the greatest increase in median rent between the start and end times there?*

4.1.1 Greatest rent increase in $.

$$R_{t_i} - R_{t_0}$$

Central Park South has had the greatest rent change in pure magnitude $\$1500.00$.

4.1.2 Greatest percentage rent increase.

$$\frac{R_{t_i} - R_{t_0}}{R_{t_0}}$$

East New York has had the greatest percentage increase in rent, at 91.5%.

4.2 *Which Borough had the highest average median rent?*

The highest median rent of the 5 boroughs is Manhattan at $\$3,216$.

4.3 *If you were to put together an "index" to describe each Borough through summarizing the median prices for each of its Areas, would you use a mean or median, and why?*

Mean is an especially useful measure if you're trying to understand data that is normally distributed. However, the mean value becomes much less useful when there are outliers. The data we are working with now is heavily influenced by outliers, which will bias our understanding of the population. Median on the other hand won't feel this type of bias and therefore median should be used as opposed to mean.

5 Exploration

5.1 *Plot a histogram of the median rents across all Areas for 2018-04*

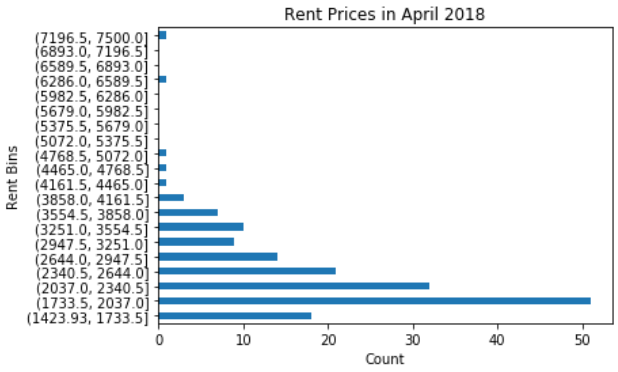

Fig. 2. Histogram: Median Rents Across all Areas

5.1.1 Initial thoughts

At first glance the histogram appears to be log-normally distributed. With this in mind we can do a couple things to test the distribution. First, we will graph a kernel density estimation for the rent price, then we will perform a log transformation on the data and create a QQ-plot to test if the data really is log normal.

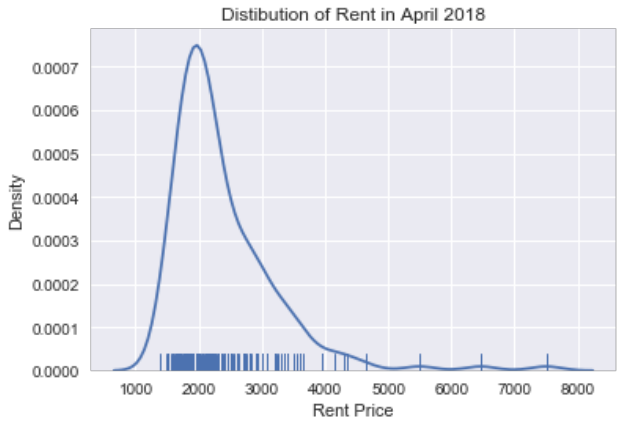

Fig. 3. Distribution: Rent prices

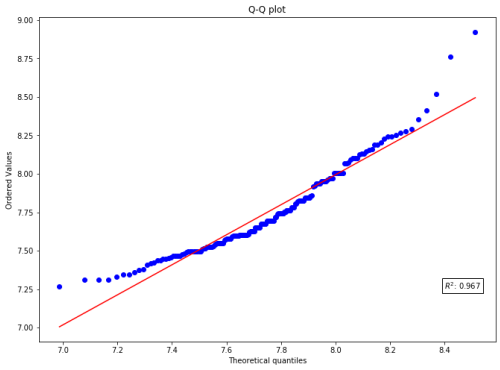

Fig. 4. QQ-Plot: Test for normality

5.1.2 Findings

Using the QQ-Plot we find that while there is a strong fit between the theoretical normal distribution and our actual distribution, the distribution of the log rent price is still skewed to the right. This implies that there is a larger portion of neighborhoods with a high median rent price than would be expected in a normal distribution. While this is telling of the rent prices in New York, I believe the true distribution is likely more symmetrical than our finding would imply. The distribution has been biased by the removal of underreported areas mentioned earlier in the paper. The wealthier, higher rent areas, like in Manhattan, tend to be reported more consistently than those of lower rent locations. Therefore there are more high median rent observations than low median rent and it is skewing the distribution.

5.2 *Pick 3-5 Areas, for which there is data for the entire provided time period, and form a new table out of these*

Date	All Downtown	All Midtown	All Upper East Side	All Upper Manhattan	All Upper West Side
2010-01-01	3200.0	2895.0	2459.5	1825.0	2895.0
2010-02-01	3200.0	2800.0	2450.0	1810.0	2800.0
2010-03-01	3050.0	2800.0	2400.0	1795.0	2750.0
2010-04-01	3100.0	2850.0	2500.0	1800.0	2800.0
2010-05-01	3100.0	2900.0	2550.0	1823.0	2800.0

Fig. 5. New 5 Area Data Frame

5.3 *Make a time-series plot of the median rents for these Areas*

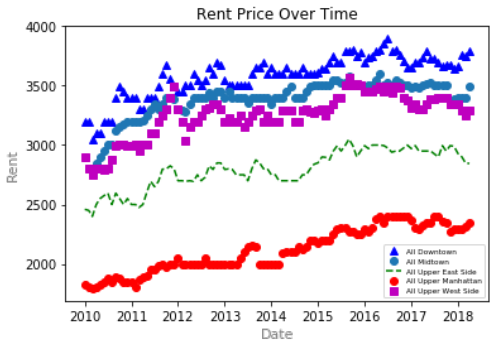

Fig. 6. Scatter Plot: Rent Prices Over Time

This graph shows rent prices tend to increase over time, with some peaks and troughs in the process.

5.4 *Produce a table summary statistics for your selected Areas*

	All Downtown	All Midtown	All Upper East Side	All Upper Manhattan	All Upper West Side
count	100.000000	100.000000	100.000000	100.000000	100.000000
mean	3584.490000	3377.695000	2794.565000	2117.990000	3248.085000
std	178.105887	173.903388	163.347366	186.468591	188.589074
min	3050.000000	2800.000000	2400.000000	1795.000000	2750.000000
25%	3500.000000	3350.000000	2700.000000	1995.000000	3200.000000
50%	3608.250000	3400.000000	2800.000000	2100.000000	3295.000000
75%	3700.000000	3500.000000	2950.000000	2295.000000	3381.250000
max	3895.000000	3595.000000	3050.000000	2400.000000	3572.500000

Fig. 7. Summary Statistics

6 Serialization

6.1 *Write your summary statistics table to an Excel spreadsheet*

The file is called Output

6.2 *Are there any other data formates that could be useful to write data to? Why?*

It can be useful to write data out to other data formats because they can save memory and create additional ease of access. Files such as 'sql' and 'stata', make it much easier for certain languages to access, while 'pickling' the data is absolutely necessary for parallel computing.

7 Modeling

7.1 *What type of model would you use to predict, for one of your selected Areas, its median rent for 2018-05? Why?*

My initial thought is to use an autoregressive model. Below are two such models.

7.2 *What kind of pre-treatment might you preform to help set the stage for modeling? Why, or why might you not do anything?*

Missing Data: Given that the areas I am using have no missing data, nothing needs to be done about this.

Additional: There are a couple things to consider. First, the distribution of the time series. Earlier we looked at the rent distribution across the population at a single point in time. Now, it's important to look at the rent changes throughout time and see if the distribution of the returns tells us anything more about the data. Below is the distribution of the returns for *All Downtown*, *All Upper East Side* and *All Upper West Side*. While the other two areas also share similar distributions, I felt that the graph became to difficult to read when all five locations were included. If you wish to see the individual distributions for *All Midtown* or *All Upper Manhattan* they are both shown in the Jupiter Notebook file.

What we notice is that rent changes actually have a very similar distribution to that of stock reruns. They are both rather symmetric about their means, but with fatter tails than what would be expected of a normal distribution. This suggests that it

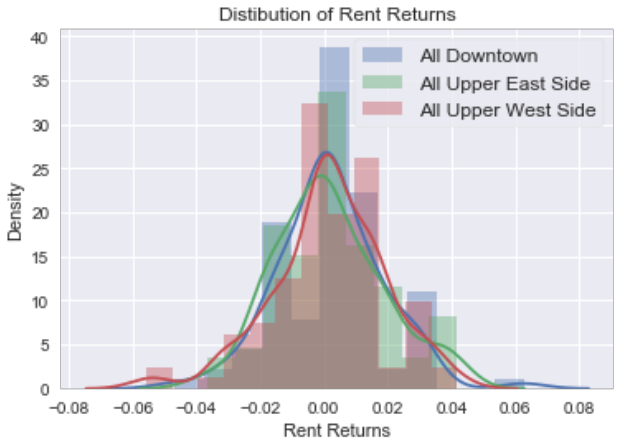

Fig. 8. Summary Statistics

could be a useful to use some techniques more commonly performed for stock analysis, such as using log returns as opposed arithmetic returns.

7.3 *Go ahead and develop a model to predict that next median rent value. How are you training the model, and how are you evaluating its performance?*

7.3.1 AR Model

An autoregressive model is when a value from a time series is regressed on by previous values from that same time series.

$$\beta_0 + \beta_1 * y_{t-1} + \epsilon_t$$

However, it is common to use multiple lag variables instead of a classic AR(1) model.

To start, I separated my 100 months of rent data into two different sections. The first 75 months of data I made my 'training data' and the last 25 months my 'test data'. Meaning I used the first 75 months to fit the AR model and then the last 25 variables to test and see how accurate my predictive model is. I found that the optimal number of lag variables to use is 11, anymore and it starts to lose significance. After training and testing the data, we find that the model predicts with an average error of $45.

I also did a log returns AR model. However, I found the model less telling than the one above. If you are interested, the results are in the Jupiter Notebook file.

7.3.2 Moving Average Model

The moving average model I decided to create myself. First, I started by finding the moving average of rent over a 3 month period, six month period and twelve month period. After, I created my own function to fit the data using all three moving averages as well as the percentage change from the previous month. Then I fed my function the first 75 months worth of variables I had created and optimized for my coefficients by minimizing the error of my fit. Finally, I used my fitted function to test the the accuracy over the remaining 25 months and found an average error of $43.

7.4 Predict values for 2018-05. Now that you have that value, predict values for the following June and July as well.

$$May = 3782.769678$$
$$June = 3779.338886$$
$$July = 3749.836177$$

7.5 Additional Thoughts

There are other linear, non-linear and machine learning techniques to try. It would make sense to go and find some other additional variables (e.g. increases in population, rent availability etc...) Then an individual could see if one of these other variables, or combination of variables has some effect on rent prices using multiple linear regression, neural nets, random forests etc.

7.6 Visualize your predicted data. How does it look?

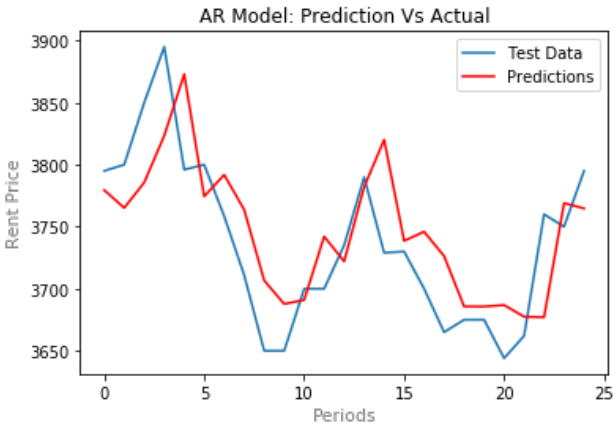

Fig. 9. AR Model (11 lag variables)

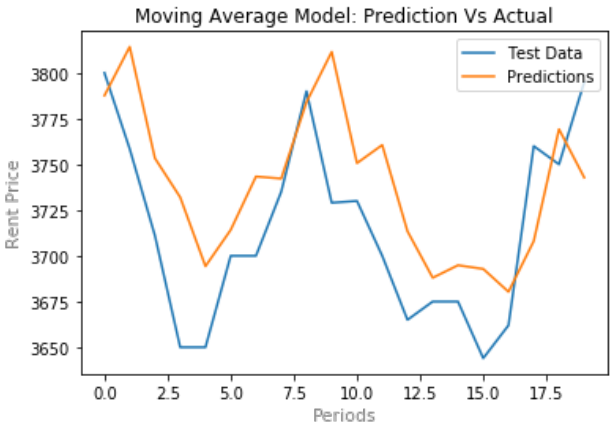

Fig. 10. Moving Average Model