Dataset Description:

This dataset, published on Kaggle, is titled “Zillow Rent Index, 2010 – Present”, and it includes the median estimated monthly rental price and rental price per square foot (the “ppsf”) for defined areas in the United States from November 2010 to January 2017. It covers all multifamily, single family, condos, and cooperative homes in the Zillow database, whether or not they are listed for rent. (https://www.kaggle.com/zillow/rent-index) Zillow is a leading online real estate and rental marketplace that endeavors to help consumers buy, sell, rent, finance, and remodel homes, with a database of more than 110 million homes. (https://www.zillow.com/corp/About.htm) Zillow produces these rent estimates, aka “Zestimates”, using a proprietary algorithm that uses public property data, local properties listed for rent, home features and location, and market conditions. (https://www.zillow.com/wikipages/What-is-a-Rent-Zestimate/)

It is interesting to note that the data includes the *median* estimated monthly rental price and ppsf as opposed to the average, or *mean*, estimated monthly rental price and ppsf. The median is generally used when a small number of extreme outliers will skew the mean up or down because such outliers do not affect the median, and the data in this particular set bear this out. Below you see the mean, median, and standard deviation of the median estimated monthly rental prices in January 2017, and one will notice that the mean price for January 2017 ($1,467.29655015) is a full $203.29655015 more than the median price for January 2017 ($1,264.00). Further, the mean ppsf for January 2017 ($0.964355902341) is a full $0.092355902341 more than the median ppsf for January 2017 ($0.872). The boxplot below that represents the median estimated rent price for January 2017 of the entire dataset further emphasizes the number and extremity of the outliers in the data. Even when one focuses the boxplot on the top 100 cities by population, as seen in the next graph, there are still some extreme outliers affecting the mean. So what were those extreme outliers that pull the mean away from, perhaps, the better measure of central tendency, the median?

Question 1: Which Cities Have the Highest and Lowest Rent Price/PPSF?

Below are the top 10 highest and lowest median estimated rent prices for January 2017. The city with the highest rent in January 2017 was Jupiter Island, FL, a city north of West Palm Beach, with a median estimated rent price of $17, 985.00. That rent price is $16,517.70344984 more than the mean for January 2017 and $16,721 more than the median for January 2017. The city with the lowest rent in January 2017 was Beecher, MI, part of metro Flint, MI, with a median estimated rent price of $518.00. This is $949.29655015 less than the mean for January 2017 and $746 less than the median price for January 2017. These numbers help explain why the mean rent of the entire dataset for January 2017 is so much higher than the median rent of the entire dataset for January 2017 because the lowest outliers are less than $1000 away from both central tendency of the data no matter how it is measured, and the highest outliers are more than $16,000 away from the central tendency.

Also below are the top 10 highest and lowest median estimated ppsf for January 2017. The city with the highest ppsf is Fisher Island, FL, part of the Miami-Ft. Lauderdale metro area, with a ppsf of $6.406, which is $5.4416440997659 more than the mean ppsf for January 2017 and $5.534 more than the median ppsf for January 2017. The city with the lowest ppsf is Elkin, NC, a small town in northwest North Carolina, with a ppsf of $0.378, which is $0.586355902341 less than the mean ppsf for January 2017 and $0.494 less than the median ppsf for January 2017. Like the data for rent price, these numbers help explain why the mean ppsf of the entire dataset for January 2017 is so much higher than the median ppsf of the entire dataset for January 2017 because the lowest outliers are less $0.60 away from the central tendency no matter how it is measured, and the highest outliers are more than $5.40 more than the central tendency.

It is interesting to note that the 6 of 10 of the cities with the most expensive rent are in California, and 5 of 10 of the cities with the lowest rent are in Ohio and another 3 of 10 are in Michigan. It seems that the West Coast is the most expensive and the Midwest “Rustbelt” states are the least expensive. It is also interesting to note that the top 10 cities with the highest ppsf are all in California and Florida (CA: 7, FL: 3), and the top 3 cities with the lowest ppsf are all in Mount Airy, a metro area in North Carolina.

Preface to Questions 2 & 3

Going forward, I will focus on the top 10 cities by population rank according to the dataset. As of 2014, these cities had a combined population of 25,191,657. (http://www.infoplease.com/ipa/a0763098.html) That is just under 10% of the total population in 2014, which was 318.9 million. (https://www.google.com/publicdata/explore?ds=kf7tgg1uo9ude\_&met\_y=population&idim=country:US&hl=en&dl=en) Below one can find a histogram that shows the distribution of the median estimated rent price for the top 10 cities by population in January 2017 and graphs that show the median estimated rent price for the top 10 cities by population from November 2010 to January 2017 and the median estimated ppsf for the same cities and dates. One can make some interesting notes and hypotheses just from looking at these graphs side-by-side, and questions come from these observations. For instance, the histogram shows that there is a gap between the top three cities with the highest rent and the bottom seven cities. This gap is also apparent when one looks at the graphs showing rent price and ppsf from November 2010 to January 2017. How similar are those cities at the top of the distribution, those with the highest rent? How similar are those cities that are at the bottom of the distribution, those with lower rent? Another observation is that San Diego has the second highest median estimated rent price and ppsf even though it is number 9 the list of 10 by population rank, and it seems to follow a similar pattern of price growth as Los Angeles. How similarly do those cities’ rent price behave?

Question 2: How similar are the rent distributions for the top 3 cities by rent price? For the bottom 7 cities by rent price?

As stated above, after looking at the histogram of rent price distributions for January 2017 and the graph showing rent price from November 2010 to January 2017, I developed a hypothesis that the distributions for cities in the top three cities with the highest rent would be similar, and the distributions for the cities with the lowest rent would be similar. To judge whether my hypothesis was accurate, I calculated the p-value comparing the distributions of two cities from the top 3 and two cities from the bottom 7. Out of the top 3, I chose Los Angeles and San Diego because I am comparing them further later in the report. Out of the bottom group, I chose to compare Las Vegas and San Antonio as well as Phoenix and Philadelphia. I chose to compare more than 2 cities out of the bottom group because it is a larger group. Los Angeles and San Diego have p-value of 0.01241778458855608, Las Vegas and San Antonio have a p-value of 0.12156427439122321, and Phoenix and Philadelphia have a p-value of 0.16389000004422588. The closer to 1 the p-value is, the more similar the distributions being compared. It appears that Philadelphia and Phoenix have the most similar distributions. Los Angeles and San Diego have the least similar distributions.

Question 3: How similarly do LA and San Diego track each other’s rent price?

As shown in the above, the distribution of rent price for Los Angeles and San Diego is not very similar. However, I developed a hypothesis that the rent price of San Diego and Los Angeles would behave in a similar manner because they are both markets in Southern California. As seen in the graphs below, when placed side-by-side, it seems, at first, that the hypothesis is accurate. Though the rent price of San Diego is less than that of Los Angeles, they both appear to follow a similar trajectory of flat growth until around November 2012 (month 25) when steady, largely uninterrupted growth begins. However, if one looks at the difference in rent price of November 2012 and January 2017 for both San Diego and Los Angeles, one will find that a difference does begin to emerge between these two cities. There is a $435 difference in rent price between those dates in San Diego and a $570 difference between those dates in Los Angeles. When do the cities’ rent price growth begin to diverge? By looking at the data and the graphs, it seems that the big divergence begins to occur after April 2014. The graph that shows the difference in Los Angeles and San Diego rent prices tells the story. From November 2010 until April 2014, the difference fluctuates but trends down until it hit $130 in April, its lowest point since December 2010 when it was $125. After April 2014, there were three big increases in the differrence that make up the larger increase over the course of nearly three years. The first increase took place between April 2014 and February 2015 when the price difference increased from $130 to $190 (these points are marked with a red and green dot, respectively). There was a brief pause in that upward trend from February 2015 to June 2015 when the difference decreased to $183, but the difference began to increase sharply again from June 2015 to January 2016 when the difference increased to $236 (these points are marked with a red and green square, respectively). The second sharp increase was followed by an even briefer pause in the upward trend from January 2016 to March 2016 when the difference decreased to $222. The final sharp increase in difference began in March 2016 and continues until January 2017 when our data ends. As of January 2017, the difference in the median estimated rent price of Los Angeles and San Diego was $284 (this point is marked with a green triangle).

An interesting observation is that the increase of the difference in rent prices seems to slow or pause in the winter months until Spring or Summer. The data does not suggest any reason for this, and I have not found any outside sources to shed light on the subject, but it would be interesting to learn if there was a reason behind it.

Proposed Further Research

Can we predict the trajectory of rent prices in Los Angeles and San Diego to determine whether the differences in the markets will continue to widen?