Data Science Blog Post – Module 4

Evan Okin

As a data scientist, it is important to understand how to perform time series analysis. A time series is a series of data points sequentially listed in time order. For example, this Module 4 project is all about analyzing monthly data (equally spaced data) on median house prices, found on Zillow.

The first step is to load in data, and preview it to make sure that it looks right. For example, we will want to test if we have missing values. The dataset has nearly 15,000 rows, each one corresponding to a different region of the country. In our dataset, we see that we have complete data for 93% of regions, which means that 7% of the data is incomplete. There are many ways we can handle the missing data, but because it represents less than 10% of the data, I decided to drop all regions with missing data points. Another possibility would be to remove all data in which there are certain number of missing months. The benefit to dropping all missing data is that we get to have all of our data on the same time scale, so that when we compare returns later, we are on a more apples-to-apples basis.

We then must decide how we want to define risk versus return. This will aid us in answering the following question – we have $100 million, how should we invest it across 5 regions? Regarding returns, we can calculate the percent change of a time series by using the pct_change() function, a useful trick that enables us to view returns without having to write our own functions. Since we have 22 years of annual data, and monthly data tends to jump around a lot, I decided to find annual returns instead of monthly returns. Another benefit of this is that investors understand annual returns and are more likely to be able to understand in a return is sufficient on an annual basis than a monthly basis, because finance is usually quoted in terms of annual returns.

Next, we must decide how we want to define risk. We want to invest in high returning assets, but we want to do so in a responsible manner. If we observe the 5 regions with the highest annual returns, we can see that 4 are in New York and 1 is in New Jersey. If we were to recommend this list of 5 regions, there would be major concentration risk. This could be problematic for tail risk scenarios such as terrorist attacks in the Tri-State area (it is a morbid discussion, but one that has to be taken into account). If we instead observe the top 5 separate regions, we get New York, Jersey City, Washington, Los Angeles, and Philadelphia. This passes the "sniff test" as we would likely expect these major cities to have higher historical returns. However, if we observe a correlation matrix of returns in these regions against each other, we see a high degree of correlation – over 90% for each region with each other region. This is problematic – if there is a down market, we could get hit doubly bad with this investment. Instead, we should look to add some hedge against this high level of correlation. At this point, it can become more of an art than a science (unless we come up with a specific objective function to define the amount of risk we are willing to take, and optimize on it). I decided to substitute out one region that exhibits more favorable correlation (i.e., lower correlation) and keep the high returns. I removed Jersey City (due to its proximity to New York) and replaced it with Rotonda West, which has a correlation in the 60s-70% instead of high 90s%. The top five regions to invest in are New York (62033), Washington (66125), Los Angeles (96127), Philadelphia (65792) and Rotonda West (72928).

We can then see how our investment would have performed over the 22-year period. It turns out that a $100 million investment, split evenly among the 5 regions, would have grown to $870 million. This corresponds to a healthy compound annual growth rate (CAGR) of 10.33% per year. But, this would hold more weight if we had an understanding of how this compares against other portfolios. By taking a random sample of 5 regions (which generated regions in Florida, Michigan, Texas, North Carolina, and California), a $100 million investment would have grown to just $298 million – a 5% CAGR and a very big difference from our more carefully selected collection of regions.

We can also use simulations to analyze a spectrum of results across different rate environments. By using the historical annual mean and standard deviation figures for each region, we can perform thousands of Monte Carlo simulations with ease. Doing so shows that the median annual CAGR is 11%, with a maximum nearing 32%. Even in a bad year, which I define as the 10^{th} percentile scenario, our investment would earn 3% - even higher than if we had invested directly in US Treasuries / government bonds.

We should also test if our time series are stationary. By plotting the rolling mean and rolling standard deviation of all 5 regions, we can see that the time series have means that are not constant over time (although standard deviations are), which tells us that the time series are not stationary. We can also run a Dickey-Fuller test to confirm that the time series are not stationary. Our null hypothesis (Ho) is that the time series are not stationary, and our alternative hypothesis (ha) is that the time series are stationary. All p-values exceed our alpha threshold of 5%, which implies that the time series are not stationary. In fact, all but Rotanda West, Florida have very high p-values above 80%.

This analysis could be improved going forward in several ways. (1) We dropped null values, but this removed 7% of the dataset that did not have values as early as 1996. We could come up with a different methodology for data removal. (2) While we looked at correlations, we could have done an actual minimization on the correlations to really capture benefits from diversification. (3) We could do more forward-looking research. Our analysis assumes that historical performance is an indication of future success. (4) We could explore further datasets, as this dataset only contains median house prices. Some regions could be risky below the median, which isn't captured here.