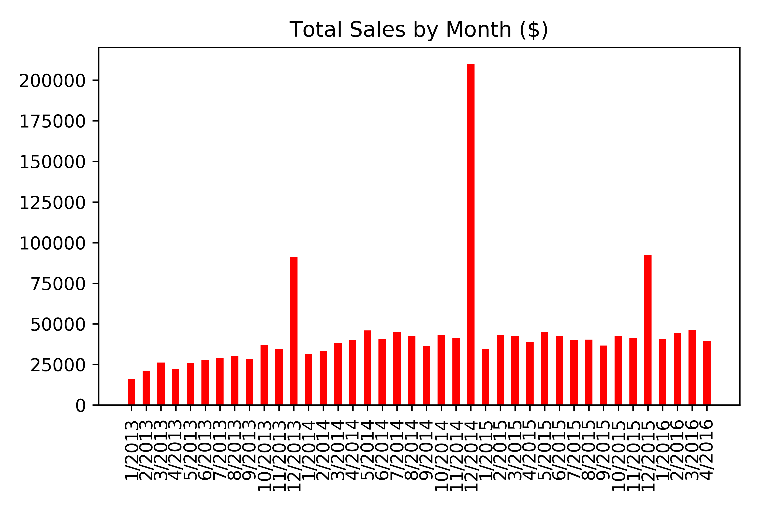
# Data Analysis Challenge

## Executive Summary:

* Sales are increasing, but growth is weak and increasing the customer base will be difficult
* There is significant seasonality in the business – on average December sales are over 300% the other months
* Significant differences in sales and growth across different regions – three regions make up 50% of sales
* Individual customer quality seems to be increasing, but measuring difficulty present challenges here

Methodology:

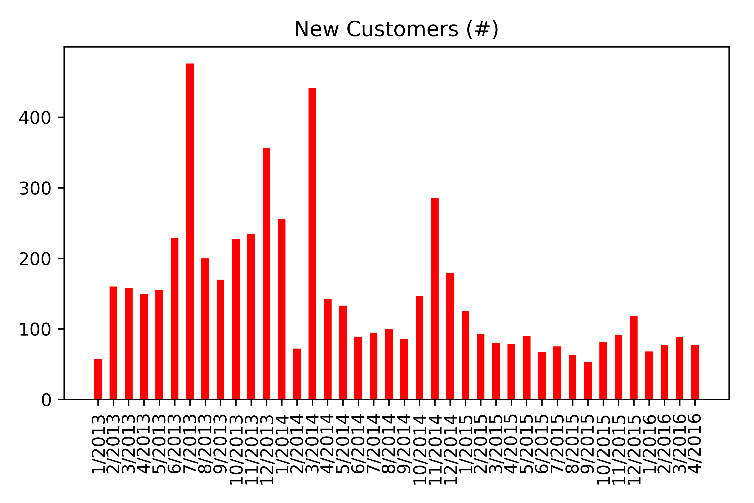
I used python in order to create my report, using libraries that included pandas, numpy, and matplotlib. To clean the data, I removed the rows where data is missing, and also removed the rows where the join date was later than the transaction date, since I assumed that the join date is the date of the first transaction. 6973 transactions were dropped in this way.

## Sales and Growth:

There was a slight upward trend in sales. Using linear regression, on average sales increased by around 670$ per month (though the correlation is very weak as r2=.06). However, the sales for 2015 were lower than those for 2014, which seemed to have occurred due to the great variance in December sales (since sales for December 2014 were abnormally high) rather than structural weakness in the business since sales for other months increased across years fairly consistently. The sales for 2016 were projected using the following method: the sales for the first four months were averaged and then extrapolated for the entire year assuming that every month would have the same sales except December, which would be multiplied accordingly (see seasonality). With that method, the expected sales for 2016 would be $595,096 which is a ten percent increase from 2015, though still less than 2014.

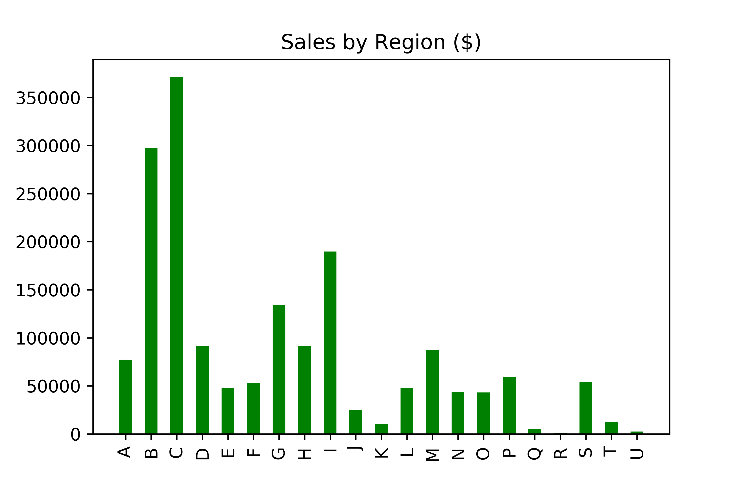
## Total Sales by Year:

|  |  |  |  |
| --- | --- | --- | --- |
| 2013 | 2014 | 2015 | 2016 (projected) |
| $388,799 | $647,343 | $538,975 | $595,096 |

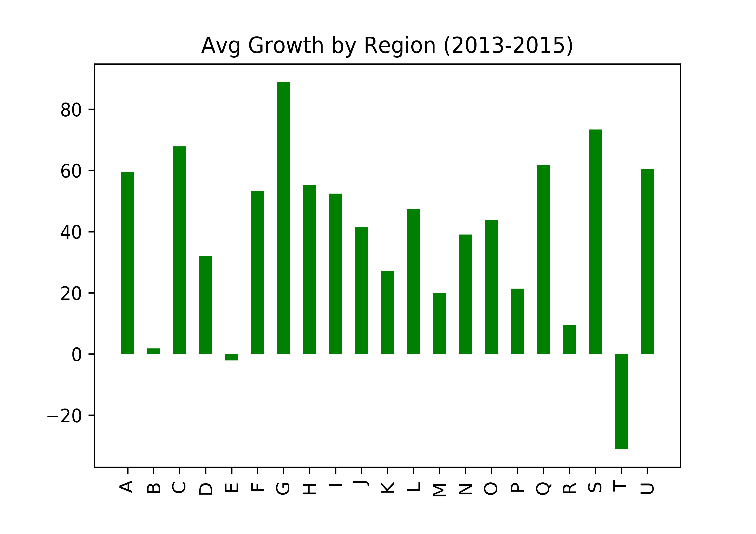


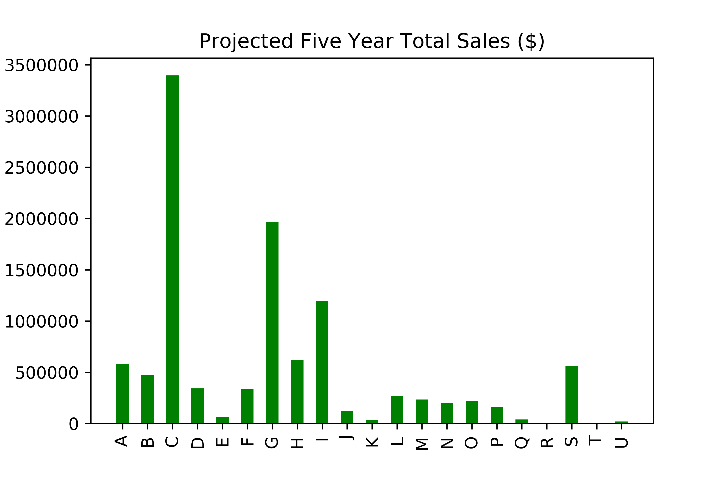
Future customer growth may prove difficult. As demonstrated in the graph, the number of new customers (as defined by a customer who made their first transaction in a given month) decreased significantly over the time period: on average, 4.25 less new customers joined every month, as calculated through linear regression. This indicates that more and more of the business is coming from recurring customers. Without acquisition of new customers, future growth will prove to be difficult, especially if customer quality doesn’t increase.

## Seasonality:

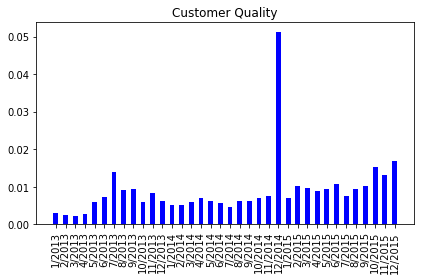
There is significant seasonality in the business, most notably concerning the month of December. In fact, on average, for December, average sales were around 300% the average sales for all months. This was adjusted for in the extrapolation of sales for 2016 above. Other than December, the sales for the rest of months were fairly constant, though there was a slight decrease in January – in January sales were around 70% of average as compared to around 85% for all of the other months excluding December. A possible explanation for this January slump could be that customers already bought the good/service in December. Since December makes up so much of the business’s total sales and there is great variation in December sales (σ = $55,581), the business should consider investing greatly in advertising, inventory, etc. in December to maximize its upside then: a 10% increase in sales for December would be much more impactful for the same increase in other months.

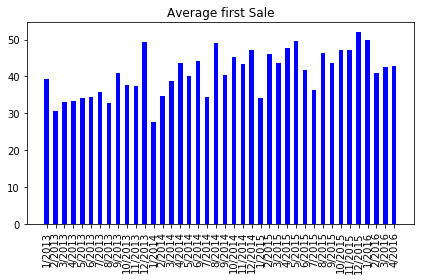
## Regional Variation:

The sales for each region and the growth for each region were calculated). Much of the sales came from just a few regions: out of the 1.75 million in total sales over the time period, around half came from regions C, B, and I while regions such as R,U,Q,K, and T had only a very small portion of sales (<$25,000 apiece). Growth also varied significantly across regions, though for most regions, there was an increase in sales over time. In order to consider both sales and growth in determining the quality of regions, I determined the total expected sales for each region over the next five years by finding the average yearly sales for each region and adjusting for its growth to sum up an extrapolation of sales over the next five years. Of course, since regional growth is very unpredictable (for example, a growth of over 80% per year for five years seems unrealistic) and the extrapolation is occurring on a data set of just three years, this will give at best a very rough estimation. With that extrapolation, the regions that will bring in the most revenue over the future should be regions C, G, and I.

Nevertheless, from a business perspective, it should make sense to invest more resources into the regions for which sales are both high and increasing. The relative weighing of importance placed on current sales and growth should depend on the tolerance for risk as well as the time span of the investment. For example, a high variance in growth in a region imply that investments there may be more volatile and therefore require a higher risk tolerance. Also, in determining which regions to invest in, for investments of longer time spans, it would make sense to place more relative important on growth over current revenues and vice versa for shorter length investments.

## Customer Quality:

In order to measure customer quality over time, I attempted to measure the amount each customer would spend in a given period of time, so I defined customer quality as the average sales a customer who joined in a given month brought in per month. This approach makes it so customers who have joined at different times are able to be compared since customers who joined earlier would have more time to make purchases. However, one issue with this approach is that this approach assumes that customer purchases are evenly distributed over time, when in fact, the first purchase actually will make a considerable difference in the average sales per month for a given customer, and this effect would be greater for customers who joined closer to the end of the measuring period. More sophisticated techniques would need to be used to correct for this effectively, but in order to minimize this effect, I did not consider customers who joined in 2016, which is closest to the current day. With this approach, I found that customer quality was increasing slightly: on average, a customer who joined a month later would spend $.02 more dollars per month than a customer who joined the month earlier, though the correlation was fairly weak (r2=.12)

Another metric I used to measure customer quality is the average value of the first sale made to a customer who joined in a given month. Increasing first sales would indicate that new customers are willing to spend more money. Using this method, there is also a clear upward trend over the months. On average, the value of the first sale to a customer who joined a month later was $.33 higher than a customer who joined the month earlier, and there was a stronger correlation here(r2=.41). Every year, the average first sale for a customer would increase by around 8%. This gives another sign that customer quality is increasing over time, since customers are willing to spend more money.

While the quality of each particular customer may be increasing, another important metric is the quality of the customer base as a whole. In order to calculate that, I created a retention metric that attempted to measure the amount of time each customer stayed as a customer on average, assuming that if the customer’s latest purchase was over a year before the current day, the customer had left the business. Then, for customers who have departed, I calculated the average time between their first and last purchases. I found, that on average, a customer would stay on the business for 812 days, or 2.2 years, though this isn’t perfect; with more complete data stretching further back, the net customer change per month could be more accurately calculated. The average customer retention time will have implications on the future growth of the business: since customer growth was the greatest between 2013-2014, some of those customers leaving between 2015-2016 and beyond have negatively impacted sales and will offset future growth. Therefore, attempting to increase the customer base will be vital for future growth.