Michael Richardson

Preston Williams

Ian Goodman

Quimera Interactive – App Data Analysis

Hypothesis

1. Our hypothesis is that given a few months’ worth of data which contained App transactions we would be able to predict the next day’s transaction amount and money made based on today’s amount, and the current month.
2. We also saw a potential trend in purchasing after the 15th of every month, it increased. We wanted to see if there was a correlation between pay day and buying trends.
3. As we looked at the data further, we wanted to see if there was a line that would be an approximate fit for the amount of transactions day by day.

Introduction

Quimera Interactive is a company that creates Windows 8 applications, which can be purchased in the Windows Store. The Windows Store is available to anyone who has a Windows 8 machine. Some information about the Windows store overall (population) is available through the developer portal and the website appannie.com. There are currently 2625 transactions in the data, transactions include: the date, app name, transaction type, country/regiona, app proceeds(USD), and the payment date. We will perform statistical analysis to discover predictive trends in sales. We will be looking at sales overall, sales by category, sales by day of week, sales by day of month, sales by month, sales by app, and sales by country.

Theory

In order to perform a prediction on transactions, you must have the current amount of transactions (which we will call xi), and the following days transaction (which we will call xi+1). Then you will take those numbers and plot them on a scatter plot. Using the linear trendline you will be able to get an equation in the form of

y= mx+b (eq. 1)

And you will then be given an R2 value which being the closer to 1 will tell you how good of a fit it is on the line.

We used this equation

(eq. 2)

as we looked at the day to day transaction count. The count looked similar to that of a damped oscillator, so this equation helps you get a good fit curve for the data points.

When you are going to compute ANOVA tables, you will need the following equations:

H0: µ1 = µ2 = µ3

Ha :µ1 != µ2 != µ3



Two degrees of Freedom: N-I (Number of totals – number of groups

I -1 (Number of groups -1)

Additionally, these three conditions must be met in order to have an ANOVA test.

* Random samples taken
* All of the standard deviations are the same
* All of the populations are normally distributed

Results

During our analysis of the app sales information provided by Quimera Interactive, it was noted that the average amount of transactions by day of the month peaked on the seventeenth and eighteenth. We formulated a hypothesis that sales could increase on these days every month due to the nearly universal bi-monthly payday that falls on the 15th and on the 1st of every month. This hypothesis was weakened by the fact that the average amount of transactions for the second and third days of the month does not also increase. We decided to conduct an analysis of variance to determine whether or not the change in the mean amount of transactions by day of month was more than just chance. The results of our initial ANOVA can be seen in figure 1.

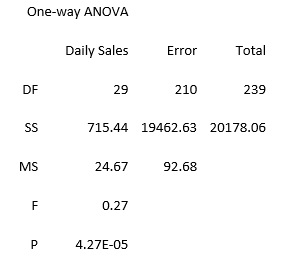


Figure 1 - 8 Month ANOVA

The result leads us to believe that there is a high probability that variance exists in the number of transactions by day of month. To gain a better understanding of this variance we produce a scatter plot as seen in figure 2.

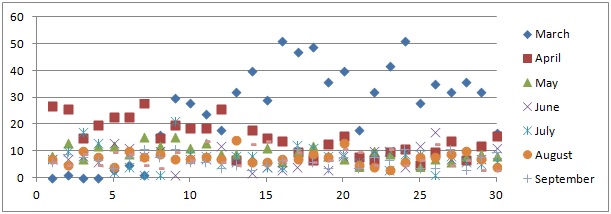


Figure 2 - Scatter plot by day of month

This scatter plot reveals that the months of March and April are producing a large amount of outliers in the data. It appears that the increased mean number of transactions that we observed to be on the 17th and 18th of the month are a result of the outliers in the March data alone. Because of this observation, another analysis of variance is run, this time excluding the months of March and April. The ANOVA should still indicate the existence of variance if the hypothesis is correct.

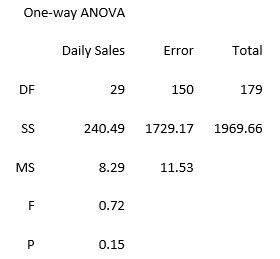


Figure 3 - 6 Month ANOVA

This P-value does not allow us to reject the null hypothesis. We can assume that any variation in the means is a result of random deviations. This result compared with the ANOVA for the full 8 month period indicates that most of the variation in the means is a result of the first two months. See figure 4.

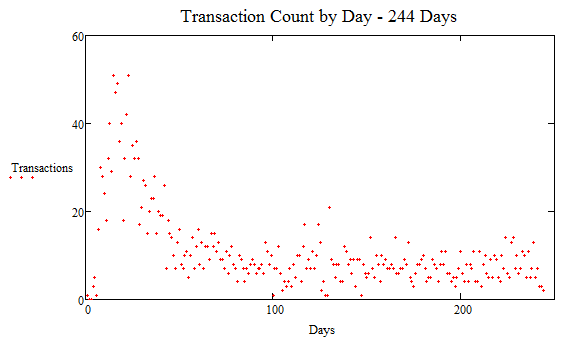


Figure 4

The peak of this scatter plot occurs on the 17th of March, and is the main cause of the variance in the first ANOVA. The scatter plot seems to follow a curve similar to an over damped oscillator. In an attempt to determine the quantity that the amount of transactions is trending towards, a fitted curve is used. Using Mathcad it is first necessary to start with estimate values for our over damped oscillator equation. Here is the equation for an over damped oscillator.

And here are the estimates used to begin the fitting process.

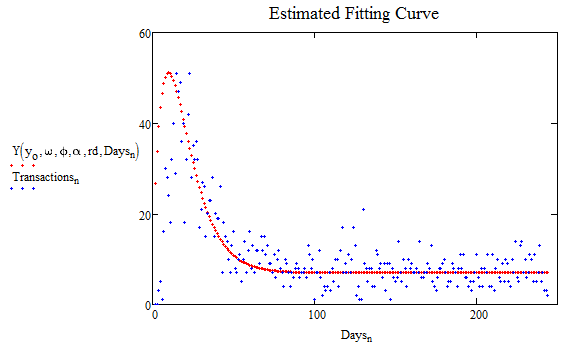
This gives a relatively poor fit, but a good beginning estimate for our purposes as can be seen in figure 5. 

Figure 5

By setting up a proper equation for computing the chi squared, and then telling the computer to find the values for the estimated variables that minimize the chi squared value, we are able to compute the best possible fit that uses the equation for an over damped oscillator.

These values result in the fitting curve seen in figure 6.

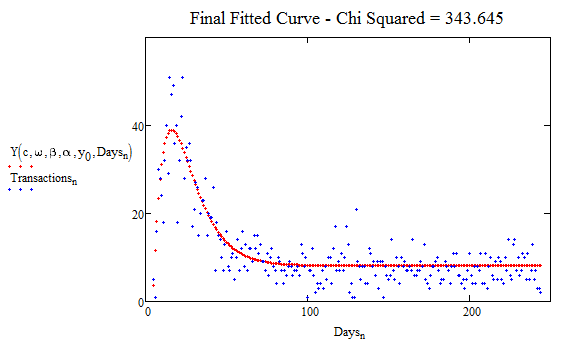


Figure 6

Using the mean standard error in our chi squared equation give us a chi squared value of 343.645. With 243 degrees of freedom this means that our P-Value is 1.265 · 10-5. This means that while this fit is the best possible using the equation for an over damped oscillator, it is still a poor fit. The equation found is still useful in that it indicates that the amount of transactions per day trends to 8.138. While this provides a good current average, future trends are unpredictable. The introduction of any new apps to the market in the future will likely increase that value.

Prediction

We wanted to be able to see if we could predict the transaction amount from month to month and from day to day. Using the equation in the theory (eq. 1) we can see if there is any correlation between the day before and the next day. These are the charts which were created from the data.

As we can see from these charts here there is no way we can predict what the next months’ number of transactions or purchase amount is because the data has not leveled off. It has only been in recent months where the data has approached a medium. It first started off at a very real high. If we were to take the first few months’ data off we would not be able to have a good fit for the data.

Conclusion