

Functional PCA for Exploring Bidding Activity Times for Online Auctions

**Valerie Hyde
Eulus Moore
Andie Hodge**

Section 1: Introduction

In the past decade the use of the internet has become a major part of most people's daily lives. The large increase in internet activity has also sparked the growth of new industry, in particular, e-commerce. There are thousands of well-known e-commerce companies. Ebay is one of the most popular and well-known. Ebay is the world's largest online auction house and plays a large part in the e-commerce industry as a whole. Founded in September of 1995, eBay is known as "The World's Online Marketplace®" for the sale of goods and services by a diverse community of individuals and small businesses. Today, the eBay community includes more than a hundred million registered members from around the world. People spend more time on eBay than any other online website, making it the most popular shopping destination on the internet. Visitors buy and sell items in numerous categories. The categories range from collectibles like trading cards, antiques, dolls, and coins to practical items like used cars, clothing, books, CDs, and electronics. Buyers have the option to purchase items in an auction-style format or at a fixed price with the Buy It Now option.

With eBay being such a popular destination for trade on the internet one may wonder how to devise ways to trade in the most efficient manner on eBay? Efficiency in participating in an eBay auction can be described as the best possible method to minimize time spent on an auction. In particular, bidding strategy has been studied in an attempt to reduce that amount of time spent "watching" the auction while still winning the item. There are several scientific techniques that can be used to analyze eBay data to improve the efficiency of auction bidding.

The objective of this paper is to explore bidding activity throughout the auction's duration using principal component analysis. Since the times at which bids are placed during an auction occur at different times for each auction, a technique known as functional data analysis (FDA) is used. A smooth function is fit to the data points: time of bid and bid amount. In this case, a monotone increasing function is fit since proxy bids will always increase. Our analysis looks at the log of the bid; however, log is also a monotone increasing function. When defining a function to fit the data, there is a trade-off between fitting the data too well (overfitting) and producing a "kinky" curve and not fitting the data as well and fitting a smoother curve (underfitting). A tradeoff must be made using the same smoothing parameter for each auction in order to capture the useful information without getting bogged down with noise. Each auction will have a unique function that describes the bidding activity over the duration. Using FDA and principal components, this paper attempts to determine how bidding activity varies though time for

the auctions that are being investigated. One of the features of FDA that is so unique is that since curves are created, derivatives can also be taken (up to one less than the degree of the polynomial). Thus, the velocity and acceleration will also be examined through the course of the auction.

Section 2: Description of the Data

The eBay auctions dealt with in this paper come from two types of wristwatches: Rolex and Cartier. They are both considered 'luxury' watches and command a 'luxury' price. The data set consists of 472 distinct auctions with a total of 6913 bids between them. All auctions were run for seven days. Ebay uses a form of bidding known as proxy bidding. Proxy bidding means the bidder specifies a maximum amount she would like to bid if bidding gets that high and eBay will keep bidding for her until her maximum is obtained. At this point eBay sends an email informing her that she has been outbid. Ebay will not automatically bid the maximum bid if the bid has not been driven up that high; it will bid just enough to beat the current highest bid.

Of the 472 auctions 97 were for Cartier watches (1348 of the bids) and 375 for Rolex watches (5565 of the bids). There was an average of 14.65 bids made per auction. Cartier had an average of 13.90 bids per auction while Rolex had an average of 14.84 bids per auction. Thus, Rolex watches had almost one extra bid per auction more than Cartier watches. For the combined watch sample there were 7.38 unique bidders. Cartier had an average of 6.88 unique bidders per auction while Rolex had 7.51 unique bidders. The bid range for Cartier sales was \$103.50 to \$5,400.00. The mean price was \$936.10 and the median was \$540.00. For Rolex watches, the price range was \$70.00 to \$24,500. The mean price was \$2,300 and the median was \$1,500.

Bidders help determine with how much each watch is sold for. The average experience per bid (which includes the same bidder when necessary) was 49.92. For Cartier watches, the average experience was 33.86, and for Rolexes the average experience was 53.80. The average experience per bidder in an auction (where only unique bidders are examined) was 64.62, with 43.81 for Cartier, and 69.55 for Rolex.

Section 3: Smoothing Splines and the Discretized Approach

For each auction, bids come in at different times during the seven day period. It would be highly unlikely for bids to come in at the exact same time (down to seconds) for any of the auctions, except maybe at the very end of the auction. One way to examine the behavior of different auctions is to determine what the bid position is likely to be at the same time points for all the auctions. To do this, some method of approximation is necessary.

The method used in this analysis is the following. Create 71 time points yielding ten evenly spaced points for each day plus an extra one for the bid position at time zero (which will always be zero). Based on the actual time of the bids for each auction, use linear interpolation to determine what the bid would be at each of the 71 time points. Next, a kernel regression smoother is applied to the data at the knots. There are a total of 14 knots, with the greatest knot density from days 6.75 to 7, the next most between days 6 and 6.75, and the least between days 0 and 6. The highest concentration of knots corresponds to where the most bidding activity is likely to occur. From the kernel smooth, 14 new lightly smoothed bids are obtained. Finally, a 9th degree smoothing spline is fit to the data using a smoothing penalty of 50. From this smoothing spline, one can obtain the smoothed values of the 71 time points that are used for the analysis. This method may also be thought of as the “grid approach.” The “grid approach” is best described as such: put a vertical grid down at the 71 time points and locate the y points by where the grid hits the auction curve.

Section 4: Smoothed Log of Bid Curves

Functional data analysis using the discretized approach is performed on the log of the bids to bring the large bids “in” and make the data more normal. For each of the 472 auctions there are ‘points’ that corresponded to the log of the bids at time t . The 71 points correspond to tenth of a day. The points are then connected with a smoothing spline. A good visual of the new discretized auction data is to plot the smoothed log of the bids versus time on the x-axis. Since smoothed functions are being used, first and second derivatives can also be taken. The original data can be considered the bid position at time t . Similarly, the first derivative corresponds to the velocity of the bid price and the second derivative corresponds to the acceleration of the bid price.

The smoothed log of the bid position, velocity and acceleration for all the auctions can be found in Appendices A, B, and C, respectively. Each graph also includes the mean curve (coded as yellow), median curve (coded as red), and the 95% upper and lower confidence curves from the mean (coded as green and blue, respectively). The mean line is calculated independently at each point. That is, at each time point, the bid position is summed over the auctions and divided by 472. The median and the upper and lower confidence intervals are calculated in the same manner.

The graphs of the smoothed log bids tell us a great deal of information. The graph of bid position shows that while all of the auctions seem to vary in their bid pricing over the life of the auction, most of the variability occurs during the middle of the auction. The graph of bid velocity shows similar information to the bid position graph: there is a lot of variability in the middle of the auction for bid speed. The acceleration graph is perhaps the most interesting in its contrast: there seems to be a lot of variability in auction acceleration during the beginning and end of the auction.

Section 5: Principal Components Overview

Principal components are used for data reduction and interpretation. They are linear combinations of the original variables. Principal components come into play when there are numerous explanatory variables. It allows for the reduction the number of variables used in analysis without losing much of the valuable information they contain.

If there are p original variables (labeled X_1, X_2, \dots, X_p), then there will be p principal components as well. Consider a linear combination of the variables:

$$Y_1 = \mathbf{a}_1' \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$Y_2 = \mathbf{a}_2' \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

.

.

.

$$Y_p = \mathbf{a}_p' \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p$$

Notice that $Var(Y_i) = \mathbf{a}_i' \Sigma \mathbf{a}_i \quad i = 1, 2, \dots, p$
 $Cov(Y_i, Y_k) = \mathbf{a}_i' \Sigma \mathbf{a}_k \quad i = 1, 2, \dots, p$. In order for the “new” variables Y_i to be

principal components, $var(Y_i)$ is maximized subject to $\mathbf{a}_i' \mathbf{a}_i = 1$ and

$$Cov(\mathbf{a}_i' \mathbf{X}, \mathbf{a}_k' \mathbf{X}) = 0 \text{ for } k < i.$$

The most popular method of obtaining principal components is to find the eigenvectors and eigenvalues associated of the correlation matrix although a similar analysis can be performed on the covariance matrix. Without loss of generality, the eigenvector associated with the largest eigenvalue is set to be the first principal component, the eigenvector associated with the second largest eigenvalue is set to be the second principal component, etc.

R uses different methods for eigenvectors, depending on which function is used. The function `prcomp` uses a singular value decomposition of the data (after it is centered and scaled) while the function `princomp` uses spectral decomposition of the correlation or covariance matrix (as is determined by the user). The default setting is the covariance matrix, so a coding adjustment must be made to perform the eigenanalysis on the correlation matrix. The function that is used in this paper is the `princomp` function because S-Plus uses similar methods for calculating the eigenvectors. For this project, some of the researchers were using S-Plus while others were using R requiring consistency across statistical packages.

For this analysis, the correlation matrix is obtained from the predicted values of bid position as interpolated over a seven day period. Therefore, the discretized approach uses the log of bids at the 71 time points to create the correlation matrix.

Section 6: PCA for Bid Position

Principal components analysis is performed on the smoothed curve of the log of the bids. The function represents the “position” at different points of time. The grid approach is used on 71 evenly spaced time points. The resulting 71 eigenvalue-eigenvector pairs found are based on the correlation matrix. This means that the sum of the eigenvalues totals 71 and, therefore, the average of the eigenvalues is one. An eigenvalue is the variance that its corresponding eigenvector explains.

The first eigenvector contains 97.22% of the variation. The second, third, and fourth correspond to 2.45%, 0.28%, 0.05% of the variation in bid position, respectively. Obviously, almost all of the variation can be explained by the first four eigenvectors. Not all of the eigenvectors need to be retained to describe the variation in the data. There are a number of different tests that decide the number of eigen-pairs to be retained. The first is the 80% test: keep the minimum number of eigenvectors to explain 80% of the variation. By this rule, only eigenvector 1 is retained. Another test is to keep eigenvectors where the variance (eigenvalue) is greater than the average of the eigenvalues. The first three eigenvectors are kept under this rule. The final test is to keep those eigenvectors that fall before the elbow on the scree plot. A scree plot graphs the order of the sorted eigenvectors (1, 2, 3, etc.) against the eigenvalue corresponding to it. See Appendix D for the scree plots for bid position and other retention criteria. Using the scree plot test, only the first eigenvector is retained. Overall, it is a judgement call to decide how many principal components to retain. It is probably a good idea to keep the first two; however, as an academic exercise, the first three will be described.

The first three principal component curves are plotted over the duration of the auction in Appendix E. The first principal component (eigenvector) takes on all negative values ranging from -0.11 to -0.21. The largest values (in absolute terms) of the principal component occur during the middle of the auction. The first principal component explains the variation that occurs during the middle of the auction.

The second principal component has negative values for the first half of the auction and positive values for the second half of the auction. After accounting for the variation in the auction bid position during the middle of the auction, the second principal component compares the variation in the beginning of the auction with that in the second half of the auction. When there is much more variation in the bid position in the second half of the auction than the first half, the principal component score will be high.

The third principal component explains almost no variability in the auction. Usually, components that do not explain much variation are hard to interpret. In this case, the eigenvector has a negative parabolic shape. It crosses the x-axis (i.e. is zero) at days two and six. After accounting for the variation in bid position explained by the first two principal components, the third principal component compares the variability in the bid position between days two and six with the variability that occurs in the first two and last one day.

Section 7: PCA for Velocity

The first derivative is taken of the smoothed curve of the log of the bid position. This represents the velocity of the bids being placed. As was the case with bid position, a principal components analysis is performed on the velocity of the log bids.

The first eigenvector accounts for 71.94% of the variation in the velocity of the auctions. The second, third, and fourth components contain 19.29%, 7.61%, and 1.10% of the variation in the velocity, respectively. The first three eigenvectors comprise more than 98% of the variation in the data. By the 80% test, the first two components should be retained. The average eigenvalue test suggests that the first three eigenvectors should be retained. Finally, looking at the scree plot, the first two components should be retained. The eigenvalue retention tests for velocity can be found in Appendix F. Overall, the first two components are sufficient in explaining the variation in auction velocity. As for bid position, the first three components will be examined as an academic exercise.

The first eigenvector takes on all negative values ranging from -0.14 to -0.02 . Each of the first three eigenvectors is plotted against the duration of the auction (Appendix G). The largest values (in absolute terms) of the eigenvector occur during the middle of the auction. This implies that most of the weight of the eigenvector is placed on the middle of the auction. As a result the greatest variation in auction velocity occurs in the middle of the auction. The first component can be described as the variability in the velocity of the bidding during the middle of the auction.

The second component has both negative and positive values ranging from -0.2 to 0.2 . The eigenvector is monotone increasing. The first part of the auction duration contains negative values of the eigenvector and the second part contains positive values. The eigenvector changes sign around day four. This second component compares the bidding velocity in the first half of the auction to that in the second half of the auction.

The third component has both positive and negative values ranging from values of -0.3 to 0.2 . This principal component appears parabolic with the lowest point being negative. The third principal component compares the velocity of days 3.5 to day 6 with the first 3.5 days and the last 1 day.

Section 8: PCA for Acceleration

Taking the derivative of bid velocity (otherwise known as acceleration) is equally interesting. Principal components analysis is performed on the acceleration of the log bids using the same procedures as for the position and velocity.

The first principal component of the acceleration function only has 38.31% of the variation. The second, third and fourth components correspond to 35.57%, 22.40% and 3.72% of the variation in the acceleration of the log bids, respectively. The 80% test suggests that the first three components are kept. The average eigenvalue test results in

keeping the first four components. The scree plot test also says to retain the first four components. As before, the first three components were kept. See Appendix H for details.

The first eigenvector takes on both positive and negative values ranging from -0.15 to 0.15 . This eigenvector (as well as the other two that are examined) is plotted over the duration of the auction in Appendix I. This graph seems to have a sinusoidal shape. The extreme values of the component occur on days 0, 1.5, 4, and 7. This component compares the bidding activity during the times of the negative portions with that of the positive portions.

The second eigenvector takes on all negative values ranging from -0.20 to -0.0 . This eigenvector has one maximum and two minimums. The largest values (in absolute terms) of the eigenvector occur at days 1 and 5. Day 2.5 is the baseline since the eigenvector has zero weight at that time. Using the baseline, the variability in auction acceleration at all other times in the auction can be compared to the baseline.

The third component has both positive and negative values ranging from -0.05 to 0.25 . The eigenvector is zero at days 1 and 4.75. This component is comparing the acceleration at the beginning and end of auction with the acceleration at the middle of the auction.

Section 9: Object Approach

There are several ways in which functions can be created. The discretization approach previously described finds the functions manually. Another way to create a function is with a functional data object. This is a built in function in the R programming language. A smooth function is created for each auction from the discretized data. Each function is called an “observation.” The object approach works as follows. First, a basis function is created using some initial parameters that the programmer selects. One parameter selects the length of the time interval. Another parameter is the nbasis. The nbasis is the number of time point that is used in the basis. The parameter norder is the highest degree of the polynomial to create. Knots tell you how many time points and the location where the polynomial spline may change concavity.

After the basis object is created, a smoothing basis is used to smooth the raw data. Within the smooth basis object there are several parameters that also have to be set. The first parameter is the data parameter. The second parameter is the degrees of freedom of the basis (or the number of knots). The third parameter is the basis parameter or the basis function parameter that was described above. Another parameter is the LDF. This parameter controls the number of degree of freedom of the derivative function. One the most critical parameters in the smooth basis object is the lambda. This parameter controls the smoothness of the object functions. The lambda parameter has a range anywhere from 0.000000000001 to 87794951 , where 1×10^{-13} is the extreme value that control jagged curves and 8×10^7 is the extreme value the control for the straightest of the

curves. There are several other options that can be set but the default options in R are usually sufficient. When the smooth basis object is fully operational there several results that are created. One result is `fdoject`. This object contains all the coefficient values of the smooth functions or curves over the specified time range along with the order of the spline. There are also the degrees of freedom and the sum of squared error for each function.

One of the disadvantages or limitations of the object approach is that it can only model one variable at a time. If it is desired to create another object, the same basis values can be used but there has to be the creation of a new smoothing basis object.

Section 10: Comparison of Output from Discretized and Object Approach

Another method to aid in the interpretation of principal components is to examine plots of the overall mean function and the function obtained by adding and subtracting a multiple of the principal component that is being examined. The mean function is obtained in the discretized method previously described by taking the mean of all the auctions at each of the 71 time points. The object approach calculates its mean curve in a similar manner but with a function defined in R. Ramsey and Silverman present how to calculate the “suitable multiple” of the principal component. However, in practice, one can try various multiples and obtain neatly identical results. The object approach, which uses a predefined function in R, calculates the suitable multiple.

Plots of the components as perturbations of the mean are created to compare the discretized and object approach. The top picture is the discretized approach and the bottom picture is the object approach. This is done first three principal components for the bid position, velocity, and acceleration curves although a discussion is forthcoming only for the bid position. Appendix J, K, and L contain the graphs for the first three components of bid position.

For the first principal component of the bid position, neither of the perturbations cross the mean curve on any of the graphs. This is because the first principal component has all negative values for each of the 71 time points. Because of this, the perturbation graph is somewhat difficult to interpret. It is difficult to see, but the perturbations are slightly closer to the mean curve at the beginning and end of the auction. This means that the first principal component is comparing the variability in bid position of the middle of the auction.

For the second principal component, the perturbations cross the mean curve at day 3.5. Principal component 2 compares the bid position in the first half with that of the second half of the auction. The same result is found for principal component 3 as was discussed earlier.

Notice how similar the plots for all three graphs are for each of the principal components examined. Judging from the graphs, these methods provide nearly identical results

although the principal components and mean curves vary slightly when the actual numbers are examined.

Section 11: Varimax Rotations of PC Curves

Orthogonal transformation of the principal components can be performed without changing the fit to correlation matrix. Therefore, if interpretation of the original principal components is difficult, a rotation may prove helpful. One popular rotation is the Varimax rotation. It attempts to make the value at the principal component for each time point either zero or large in absolute terms. Using the object approach, there is a predefined function in R to perform the rotation. The Varimax rotation is applied to the principal components of bid position to determine if the components are more easily interpretable. It is important to note that after a rotation, the principal components may now explain different amounts of variation.

When rotated, the first principal component accounts for 30.70% of the variability. As can be seen in Appendix M, this component has a slightly different interpretation than the unrotated set of principal components previously discussed. This component compares the variability in bid position during the middle of the auction with that at the beginning and end. However, the middle of the auction is now days 3.5 through 6.75.

The second principal component contains 61.50 % of the variability. After rotation, the second principal component now contains most of the variability in bid position. This component also compares the variability in the middle of the auction with that at the beginning and the end. Here, the middle of the auction is considered to be days 0.75 to 6.50.

The third rotated principal component contains 7.70% of the variability. This principal component also compares the middle of the auction with the beginning and the end. Days 1 to 5 are the middle of the auction for this principal component.

Section 12: Comparison of PC Curves for Different Products

Are these results wristwatch specific? The results obtained for wristwatches may be interesting in themselves; however, it would also be informative to see if the principal components for wristwatches are the same for other, completely different, products. Auctions for Callaway and Titleist golf balls are used to answer this question. There are two reasons for choosing golf balls. First of all, golf balls and watches are in completely different product categories. Secondly, the market value of a high end wristwatch is much larger than the market value of a box of golf balls. Finally, all of the items have a true market value. There should be no additional private value based on these items.

Judging from Appendix N, the principal component curves for bid position of the four items are very similar. For PC 1, the curves have the same general shape for all items.

However, Callaway golf balls have the least middle of the auction variability and Titleist golf balls have the most. Since the two product categories are NOT grouped together, it does not appear as though PC 1 is product or item specific. The differences are probably due to variability in the auctions chosen.

All the PC 2 curves appear to be very similar as well. They all cross zero between days 3.5 and 4.5. This means that the comparison of first and second parts of the auction have different part lengths. Again, none of the products are grouped together. The results for PC 3 are almost identical for the four items. There do appear to be two groups: Rolex/Titleist and Cartier/Callaway. Obviously, these results are not product specific since the groups contain one watch and one golf ball brand in each group.

The velocity curves are nearly identical for all items (Appendix O). None of the products seem to have any grouping. There is a problem, visually, with these plots. Eigenvectors are not unique. A rotation will provide the same results with the opposite signs for each time point without changing the interpretation of the data. The interpretation would simply be the opposite and also have the opposite sign. For PC 1, Cartier watches would be in line with the other three curves with a reflection over the x-axis.

The acceleration curves are different for the different items (Appendix P). The first PC has Cartier watches slightly “ahead” of the other three items. This may be due to the variability in the auctions, not something special about Cartier watches. With more imagination, it can be seen that the second principal component curves are similar except for differences in location and scale. This is the only set of curves where it appears as though there may be item specific differences. More research is necessary to make this claim. The third principal component is suffering the non-unique problem described above. A reflection over the time axis would remedy this problem. These curves would then be in line.

It appears as though the eigen-analysis that has been discussed thus far is similar, not only across items within a specific product category, but also across items in general. An idea for further research is to look at the principal component curves for many more product categories and items within product categories.

Section 13: PC1*Auction Plots

In order to get a better understanding of the relationship between principal component score one and some of the original variables unique to each auction, graphs are constructed with the auction id on the x-axis and score one on the y-axis. The points are divided into groups by the quartiles of certain static variables to see if a relationship could be established between the score and these variables. The color grouping for quartiles is blue, green, yellow, and red, representing lowest to highest quartile. See Appendix Q for details.

The first variable investigated is the predicted y-value in the middle (day 3.5) of the auction. As can be seen from the graph, the colors are very localized. Those auctions with the highest value (in absolute terms) of the log bid in the middle of the auction have the highest score 1; those with the lowest value of the log bids in the middle of the auction have the lowest score 1. This further illustrates the fact that the first principal component score represents variability in the auction middle. When the bid is high in the middle, there can be a lot of variability in the bidding. If it is low, there is no room for the price to move, so there is not much variability.

It does not seem as though opening bid is related to the first score since a projection of the colored data points onto the y-axis would have all the colors equally mixed. There does appear to be a little separation for the fourth quartile in that it tends to have higher values of score one. When the opening bid is large, there is more variability during the middle of the auction. This is because some people will start bidding close to the market valuation of the item and others will wait for the end to bid.

High bid does appear to be related to the first score. The divisions of the quartiles are not quite as clear as for the middle predicted y values, but there are definite clear breaks in the first score for the different quartiles. There is an inverse relationship between high bid and the first principal component score. The higher the high bid, the lower, in absolute terms, the first score. The high bid is the final winning bid placed in the auction. The higher the bid, the more variability in bid position in the middle of the auction. When an auction ends with a high price, there was more room during the middle of the auction for different bids to have been placed.

Section 14: PC1*PC2 Plots

Plotting the values of the principal component scores against each other can be a valuable tool in determining which variables play an important role in defining the scores. For the bid position data, it was determined that the first principal component should be retained and possibly the second and third. Since the first two principal components account for over 99.50% of the variability in the bid position, only the first two scores are investigated in this manner. Recall that the first principal component corresponds to the variability in auction price in the middle of the auction. The second principal component compares the variability in auction price of the first half to the second half of the auction.

The principal component scores are plotted against each other with the first principal component on the x-axis and the second principal component on the y-axis. Obviously, there is one point on the graph for each auction. One of the ways to distinguish between the auctions is based on how they act in different aspects of the auction. Some static variables of interest are: opening bid, highest bid (sale price), seller experience, number of bids, and number of bidders. It is also interesting to compare the values of the scores for the two types of watches. Each of the numeric variables mentioned above are divided into quartiles. The graphs (in Appendices R through W) color code the auction data points that are in a particular quartile. The first to fourth quartiles are color coded as follows: blue, green, yellow, and red.

For opening bid, there seems to be some separation of the quartiles. The highest values of opening bid are high on score 1 (in absolute terms) and low in score 2. This means that when the opening bid starts off high, there is variability in the middle of the auction, and not much difference between the bid at the beginning and end of the action. Score 1 can be explained since sometimes people will bid close to the market value in the middle to discourage other bidders and others will wait until the end of the auction to bid close to the market value. Score 2 also makes sense since bids starting off high, close to market value, will not vary much from the end price. The other three quartiles follow suit, with each quartile taking on lower values of Score 1 and higher values of score 2; however, the separation is not nearly as remarkable as it was for the largest opening bids.

High bid is the final price paid by the bidder. High bid has the most separation of the quartiles of any plot that will be discussed. All of the separation is along score 1, the variability in the middle of the auction. The auctions with the highest final bid correspond to the highest score 1. This means that auctions with high selling price have the most variability in the middle portion of the auction. Oppositely, auctions with the lowest final selling price have the least variability in the middle of the auction. This is logical since there is not much movement in the selling price if the final auction price is low. Also, since most bidding activity takes place at the beginning and end of the auction, low auctions don't have many bids and certainly not much activity in the middle of the auction.

Seller experience does not illustrate any separation among the quartiles on either of the principal component scores. Seller experience does not affect the variability in bidding activity for auctions.

Not surprisingly, the number of bids and number of *unique* bidders display very similar patterns of the data points. There is no separation of either bids or unique bidders as a projection onto the score 1 axis would verify. Low number of bids (bidders) corresponds to very low values of score 2. This means that when there are not many bids (bidders), there is not much of a difference between bid position at the beginning and end of the auction. This is probably due to a high opening bid in which case there would not be much bid activity in general. (Comment to group: would be interesting if we could plot size of opening bid as size of point, but I'm not sure how we would do this). The highest three quartiles are not separated on score 2. The first two scores are using information only from low number of bids (bidders).

The last score plot looks at the differences between watch types. There is no difference between Rolex and Cartier watches on score 2; however, Rolex watches have larger (in absolute terms) values on score 1. For some reason, Rolex watches have more bid position variability in the middle of the auction.

Section 15: Conclusion

Some of the main objectives accomplished were displaying the auction data so as to highlight its various characteristics and help study the important source, patterns, and variation among the data. These objectives were met using functional data analysis (FDA). The information obtained from FDA helped to identify the patterns within the data, patterns not normally discernable.

To further study the behavior of the auctions and their rates of change, principal components analysis (PCA) was used. PCA showed the different variations within the auction. Using functional data analysis and principal components analysis, the data characteristics began to jump off the page. Intricacies not normally seen were noticed, and this allowed for much more thoroughness in the data scrutiny. With functional PCA, the entire auction duration became important, not just the end of the auction. Now, auctions are understood from beginning to end, and each part of an auction reveals some new, previously hidden information. This paper laid the groundwork for future work to be done in auction analysis, and functional data analysis will be at the helm.

Section 16: Webcrawler

The objective of webcrawler is to use PHP, MySQL and Apache to facilitate systematic real time longitudinal data capture from the web. Code was written to log in to the MySQL database with the user id and password created during MySQL's installation. This code should create a database with n rows and 7 columns. The seven column headings correspond to the information to be obtained from eBay such as: item type, start and end time of the auction, bid amount, etc. The code will display a table with the extracted data columns listed by the auction id.

To get the webcrawling process to work, one first has to install an Apache server, PHP, and MySQL. Apache is a freely available source code implementation of an HTTP (Web) server, PHP is a scripting language that is especially suited for Web development and can be embedded into HTML, and, MySQL is a database software that works in conjunction with PHP and the Apache server.

There have been a couple of issues in trying to implement the code (available in Appendix X). One problem was trying to get the Apache server to work on the network at school. The Apache server did not work on campus because the school blocks servers from being set up on its network. In order for the server to work, a "hole" would have to be put in the firewall, which could subsequently be exploited by predators. Also, there are many different methods for extracting information from websites. Everyone that was talked to in regards to webcrawling used different methods, making it hard to get a consensus. This made it challenging to get the help needed to fix errors in the code.

While the code is by no means complete, its foundation is solid and, with a little tweaking, can be used to gather more rich eBay data for future analysis.

References

Bajari, Patrick and Ali Hortaçsu, "Economic Insights from Internet Auctions," Duke University of Chicago, 2004.

Fox, John, An R and S-Plus Companion to Applied Regression, Sage Publications, 2002.

Johnson, Richard A. and Dean W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New Jersey, 1998.

Ramsey, J.O. and B.W. Silverman, Functional Data Analysis, Springer-Verlag, New York, 1997.

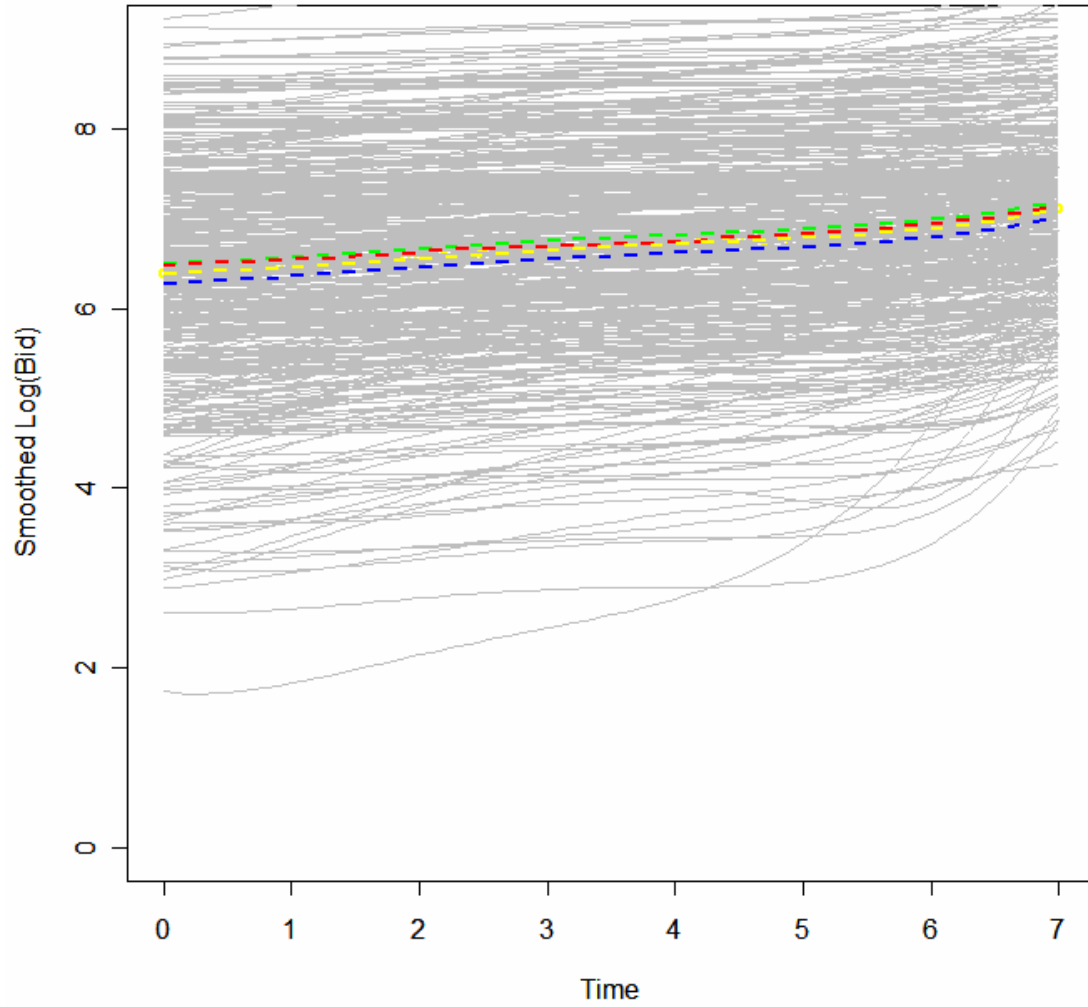
Venables, W. N. and Brian Ripley, Modern Applied in Statistics with S-Plus, Springer Verlag, 1999.

Zantek, Paul, BMGT883 Spring 2004 Class Notes: Principal Components and Factor Analysis Sections

<http://www.sba.uconn.edu/users/rbapna/courses/403eeis/>

Appendix A

Functions, Mean, Median, and Upper and Lower Conf Curve



Yellow: mean curve

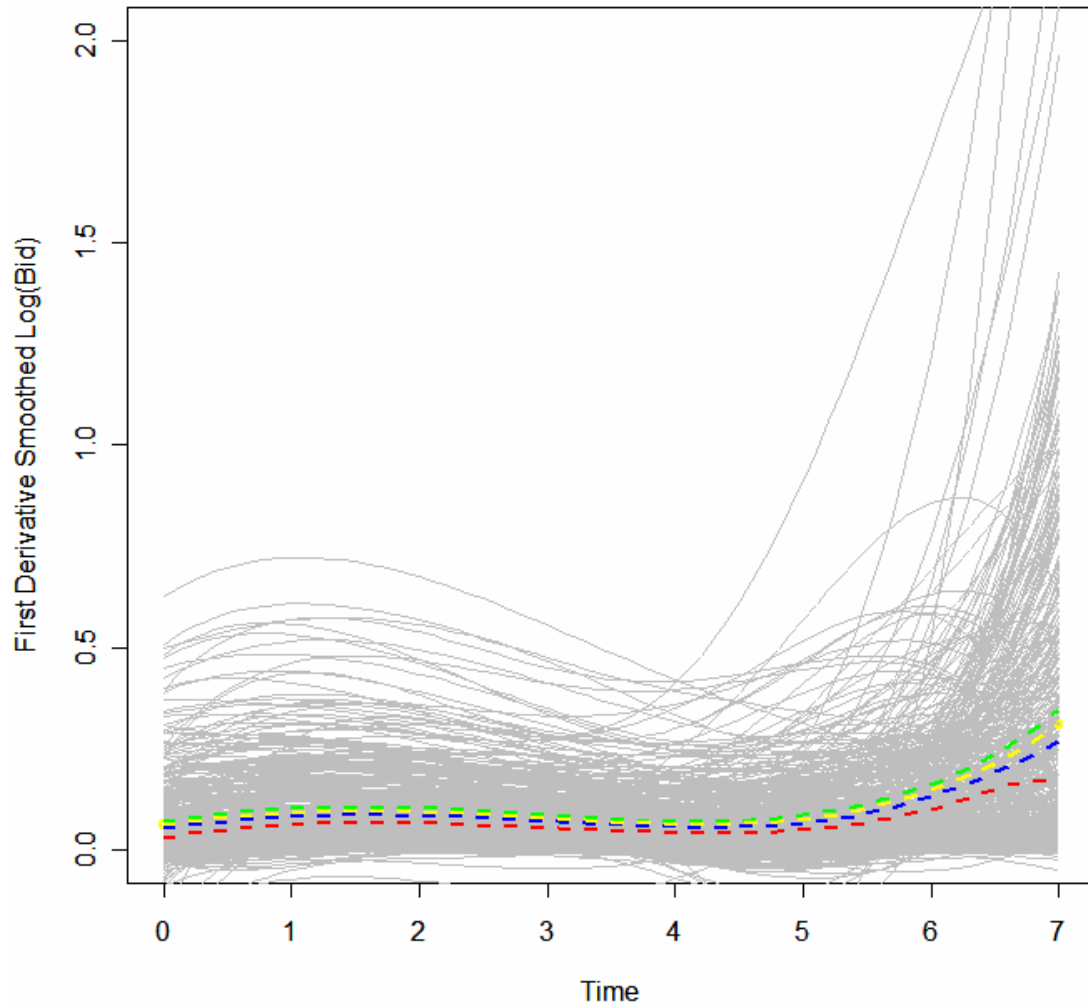
Red: median curve

Green: upper 10% confidence curve

Blue: lower 10% confidence curve

Appendix B

Functions, Mean, Median, and Upper and Lower Conf Curve First Der



Yellow: mean curve

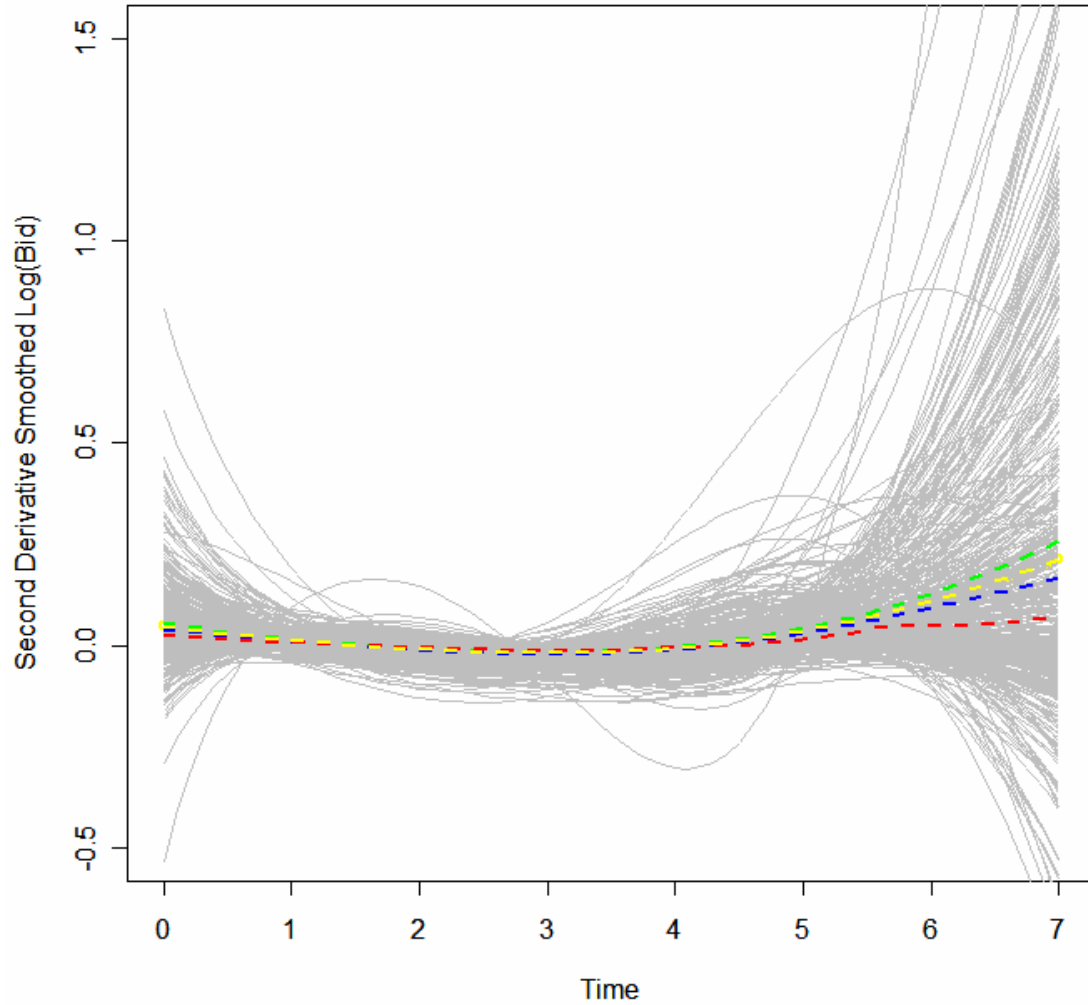
Red: median curve

Green: upper 10% confidence curve

Blue: lower 10% confidence curve

Appendix C

Functions, Mean, Median, and Upper and Lower Conf Curve Second Der



Yellow: mean curve

Red: median curve

Green: upper 10% confidence curve

Blue: lower 10% confidence curve

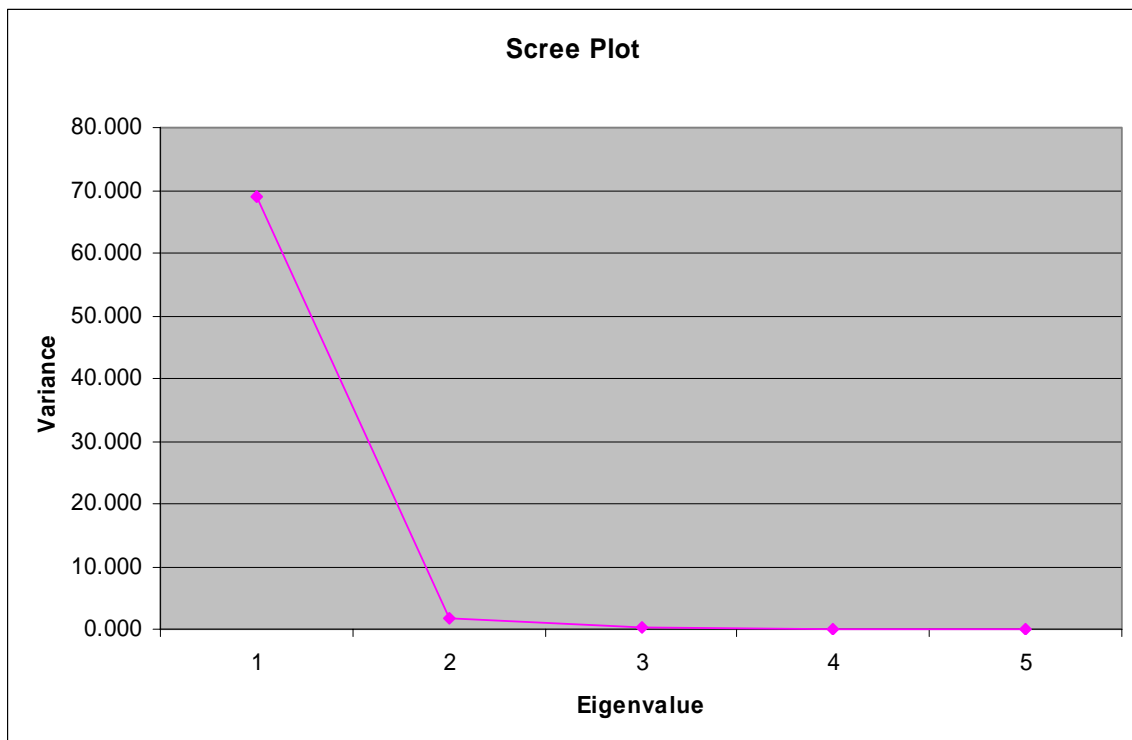
Appendix D

Principal Component Analysis for Watch Data Based on Smoothed Log(Bid)

<u>Eigenvalue</u>	<u>Variance</u>	<u>Percent of Total Variation</u>
1	69.026	97.22%
2	1.739	2.45%
3	0.200	0.28%
4	0.032	0.05%
5	0.003	0.00%
.	.	.
.	.	.
71	0.000	0.00%

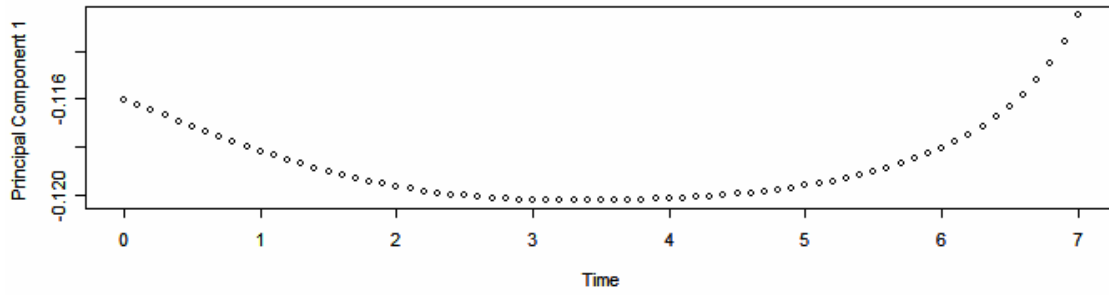
How many principal components should be retained?

1. Keep only those principal components that correspond to 80% of the total variance.
Retain Principal Component 1.
2. Keep only those principal components whose variance is greater than the average eigenvalue.
Retain Principal Components 1, 2, and 3.
3. Keep only those principal components before the "elbow" on a scree plot.
Retain Principal Component 1.

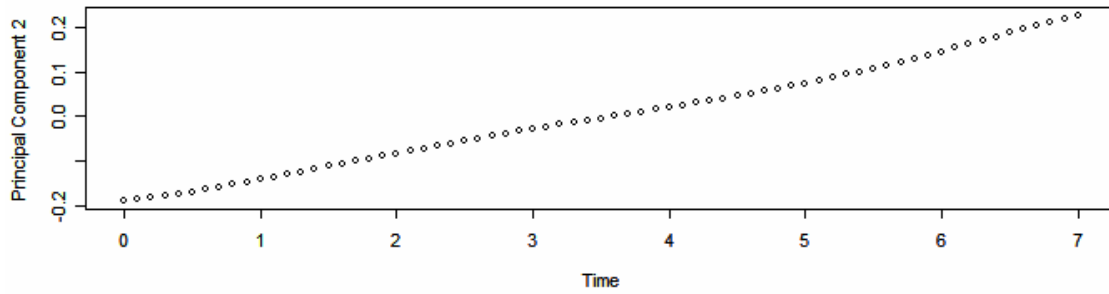


Appendix E

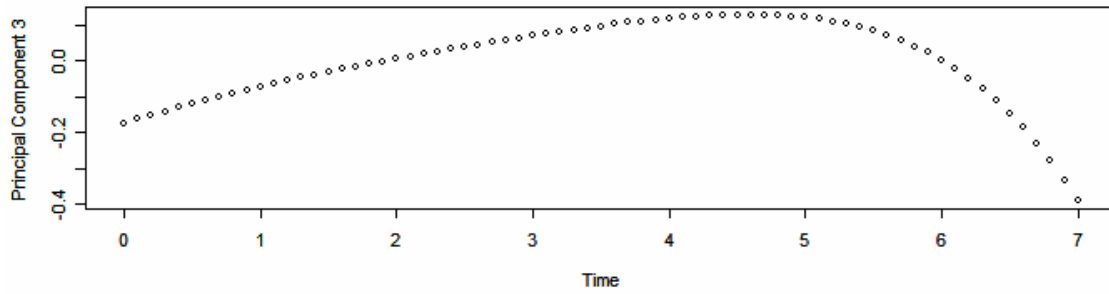
Plot of Time Versus PC1 (97.22%)



Plot of Time Versus PC2 (2.45%)



Plot of Time Versus PC3 (0.28%)



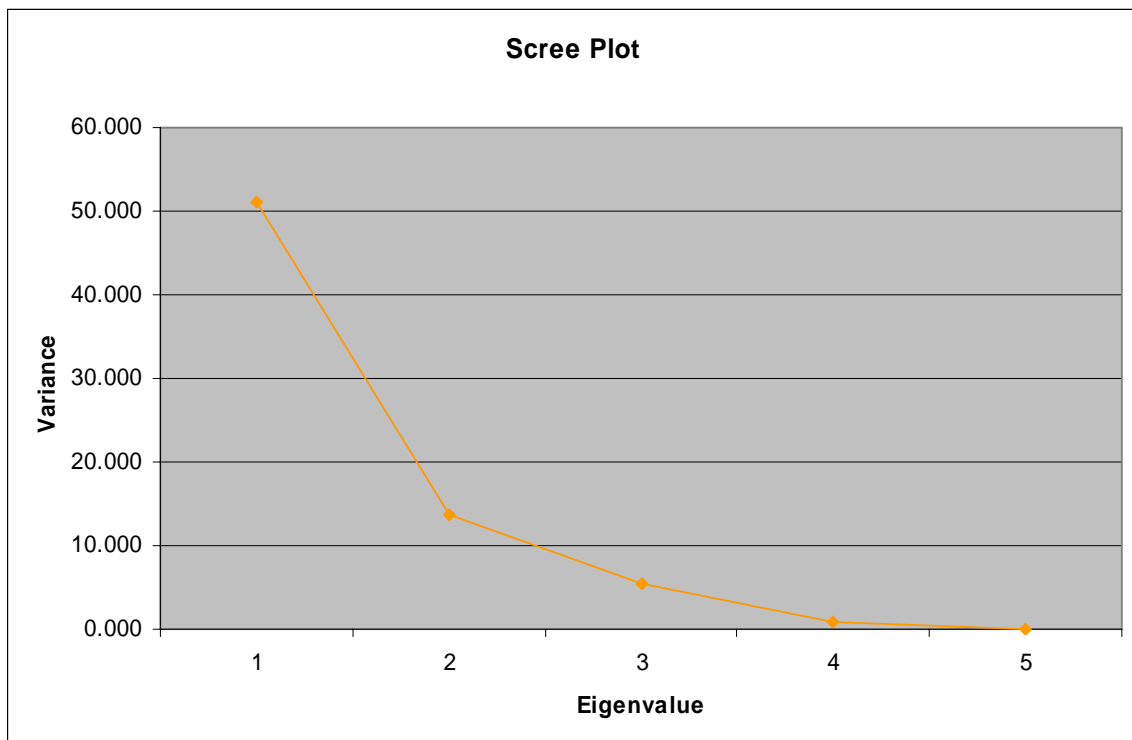
Appendix F

Principal Component Analysis for Watch Data Based on First Derivative of Smoothed Log(Bid)

<u>Eigenvalue</u>	<u>Variance</u>	<u>Percent of Total Variation</u>
1	51.080	71.94%
2	13.695	19.29%
3	5.400	7.61%
4	0.782	1.10%
5	0.043	0.06%
.	.	.
.	.	.
71	0.000	0.00%

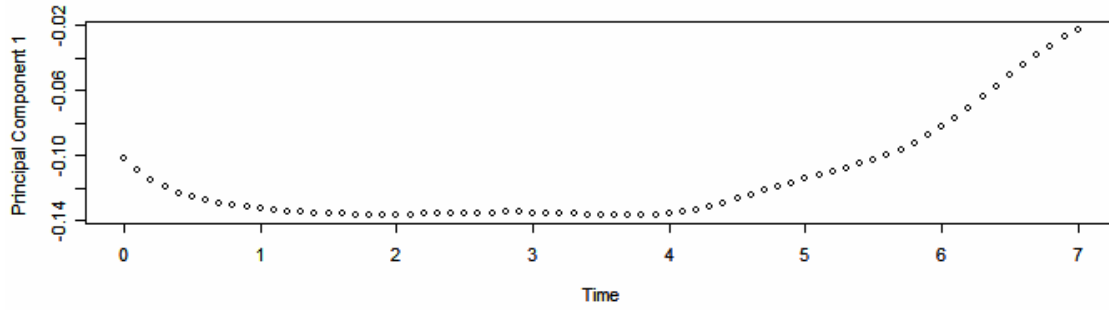
How many principal components should be retained?

1. Keep only those principal components that correspond to 80% of the total variance.
Retain Principal Components 1 and 2.
2. Keep only those principal components whose variance is greater than the average eigenvalue.
Retain Principal Components 1, 2, and 3.
3. Keep only those principal components before the "elbow" on a scree plot.
Retain Principal Components 1 and 2.

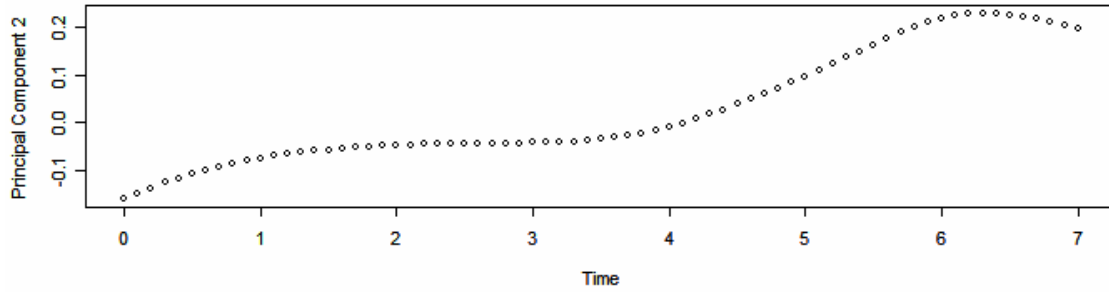


Appendix G

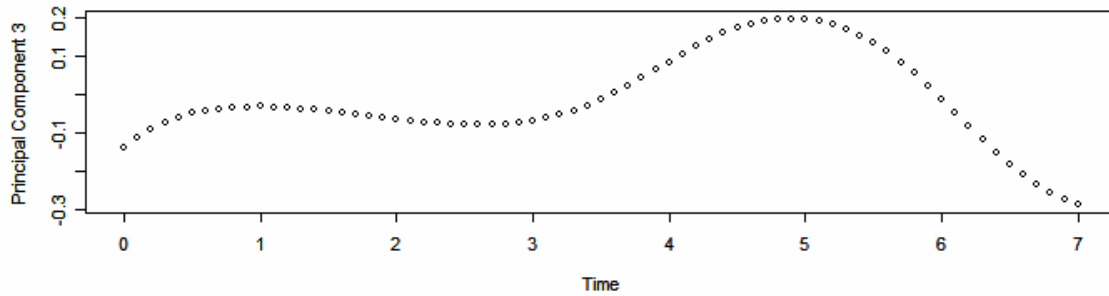
Plot of Time Versus First Derivative PC1 (71.94%)



Plot of Time Versus First Derivative PC2 (19.29%)



Plot of Time Versus First Derivative PC3 (7.61%)



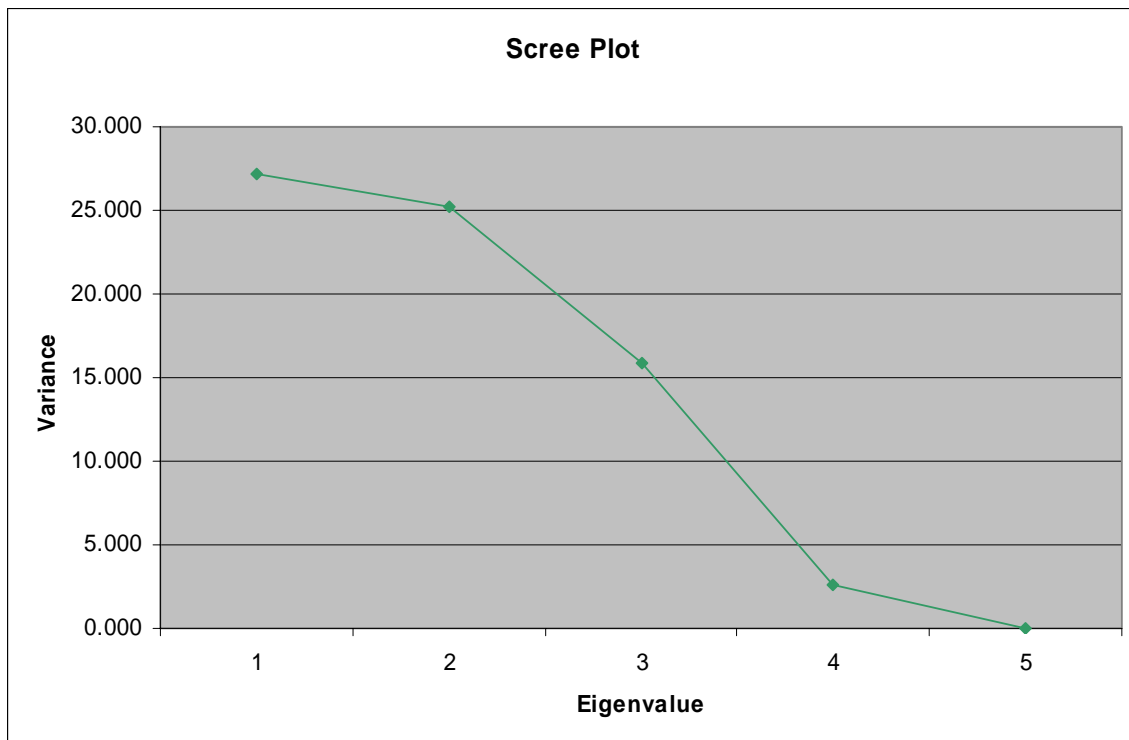
Appendix H

Principal Component Analysis on Watch Data Based on Second Derivative of Smoothed Log(Bid)

<u>Eigenvalue</u>	<u>Variance</u>	<u>Percent of Total Variation</u>
1	27.198	38.31%
2	25.251	35.57%
3	15.901	22.40%
4	2.641	3.72%
5	0.008	0.01%
.	.	.
.	.	.
71	0.000	0.00%

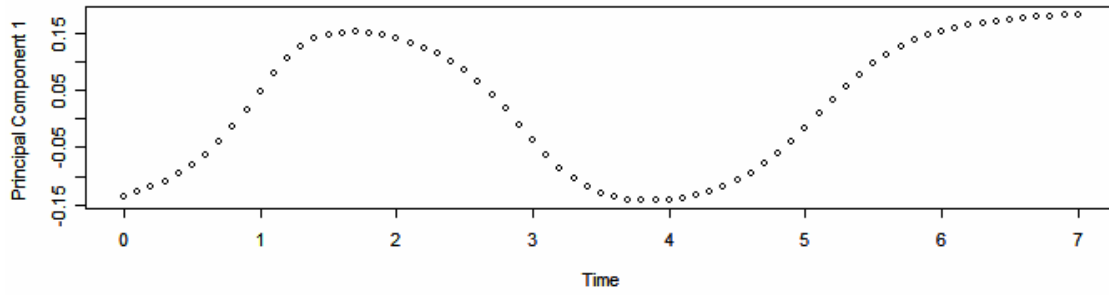
How many principal components should be retained?

1. Keep only those principal components that correspond to 80% of the total variance.
Retain Principal Components 1, 2, and 3.
2. Keep only those principal components whose variance is greater than the average eigenvalue.
Retain Principal Components 1, 2, 3 and 4.
3. Keep only those principal components before the "elbow" on a scree plot.
Retain Principal Component 1, 2, and 3.

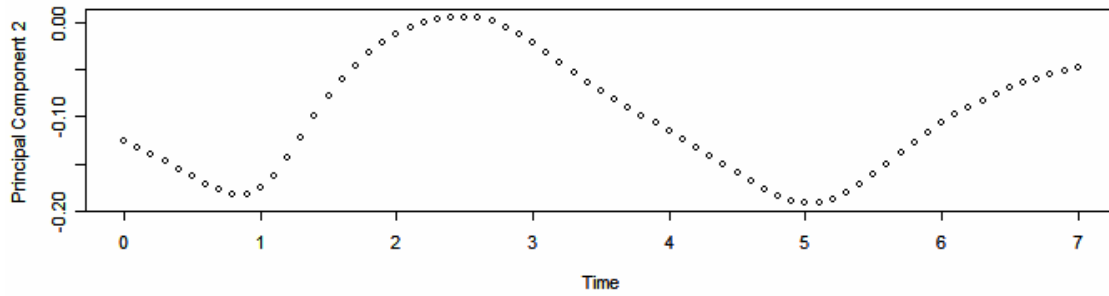


Appendix I

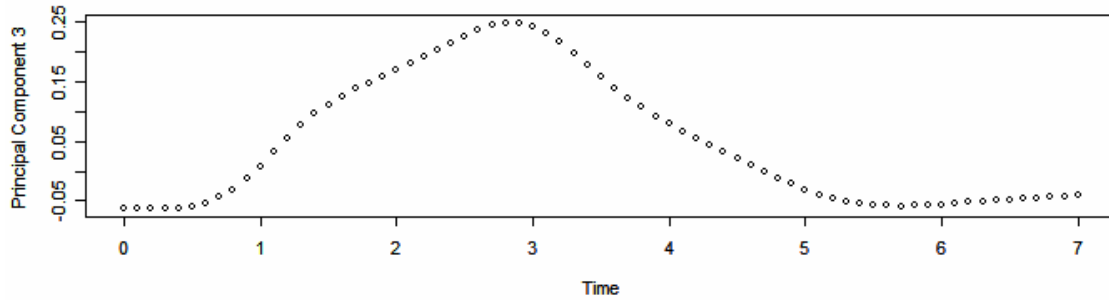
Plot of Time Versus Second Derivative PC1 (38.31%)



Plot of Time Versus Second Derivative PC2 (35.57%)

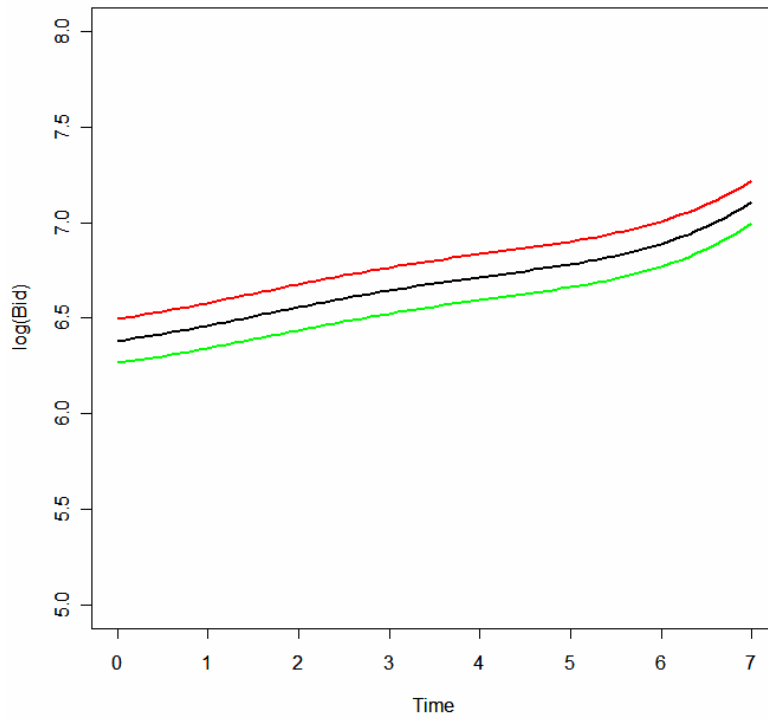


Plot of Time Versus Second Derivative PC3 (22.40%)

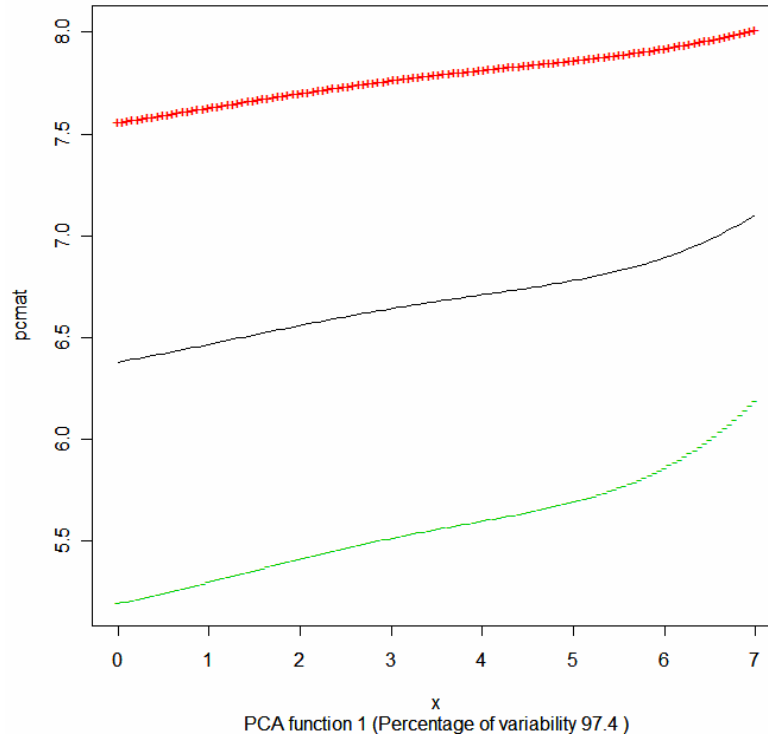


Appendix J

Plot of Time Versus Log(Bid)

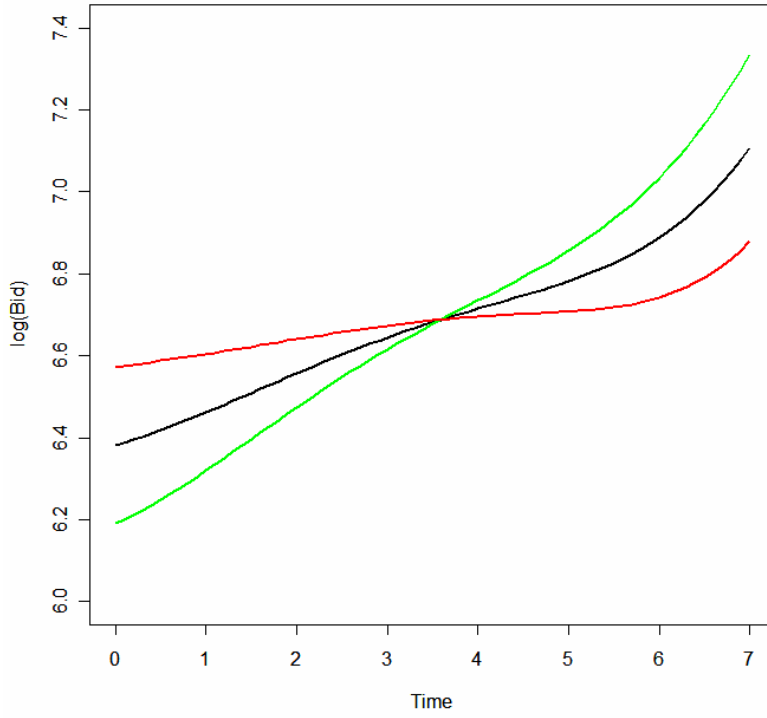


Click to advance to next plot

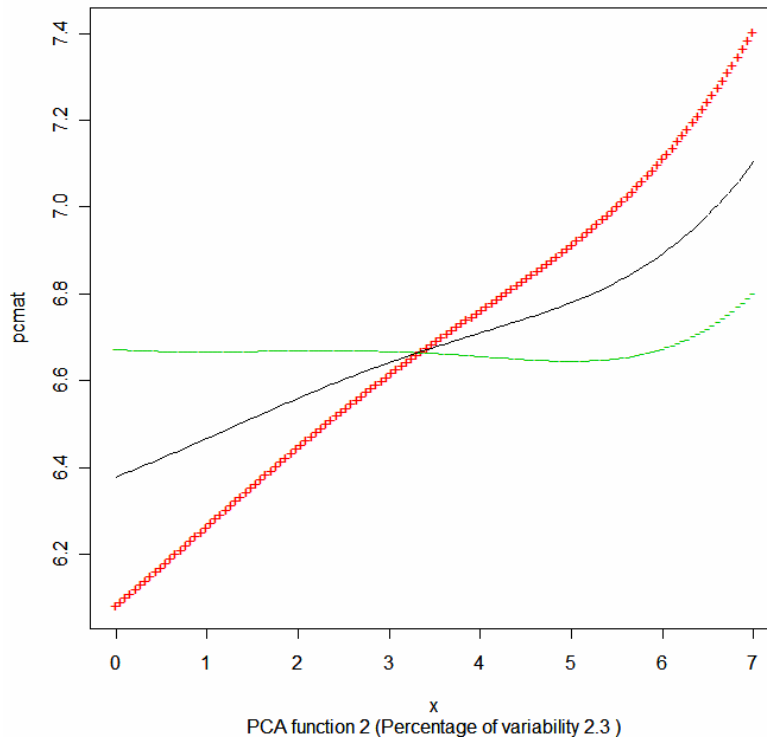


Appendix K

Plot of Time Versus Log(Bid)

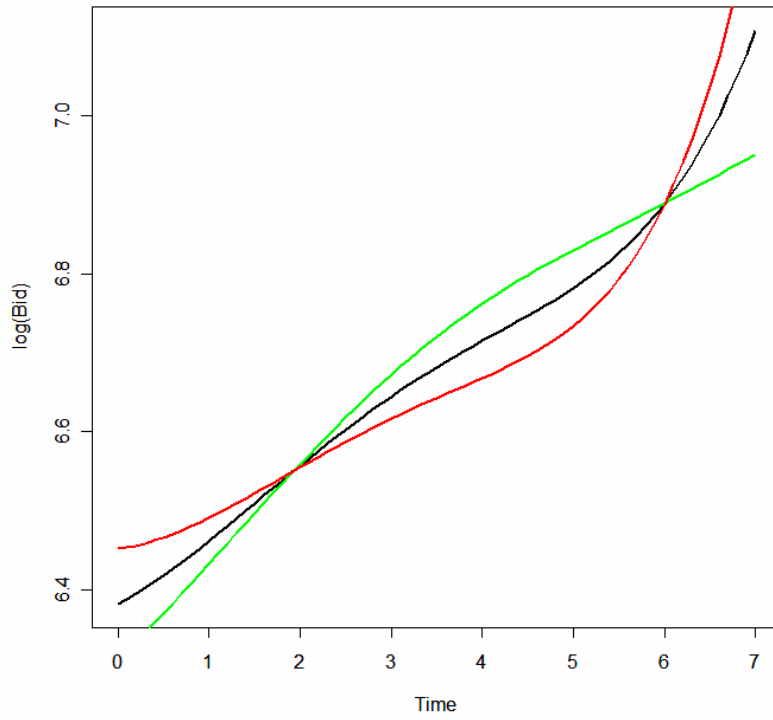


Click to advance to next plot

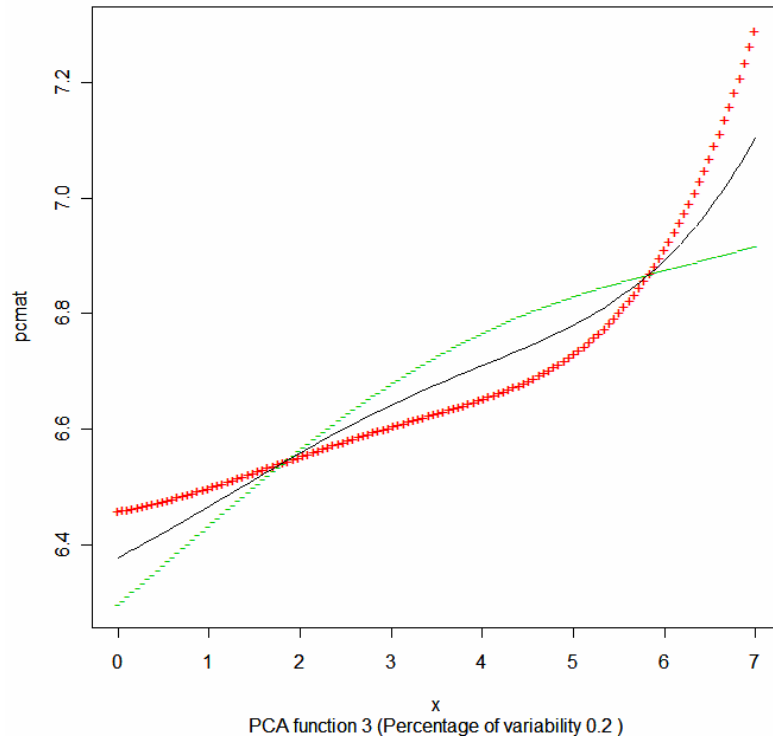


Appendix L

Plot of Time Versus Log(Bid)

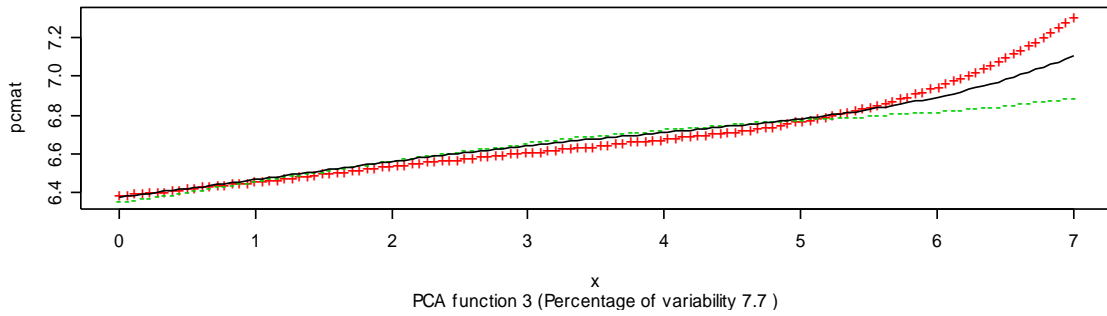
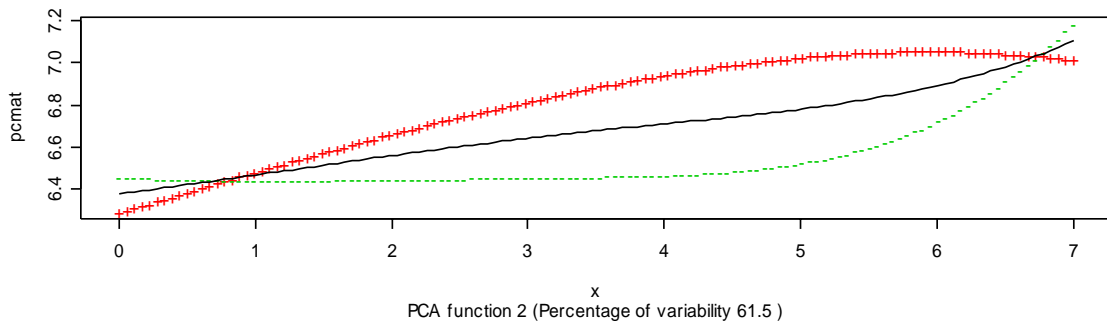
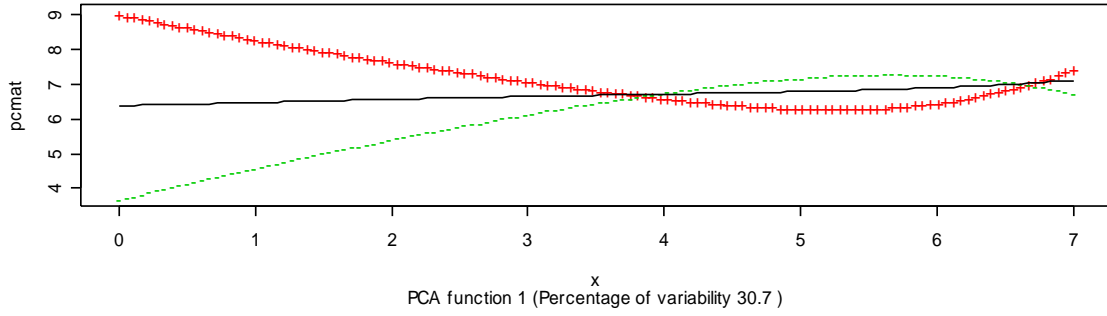


Click to advance to next plot



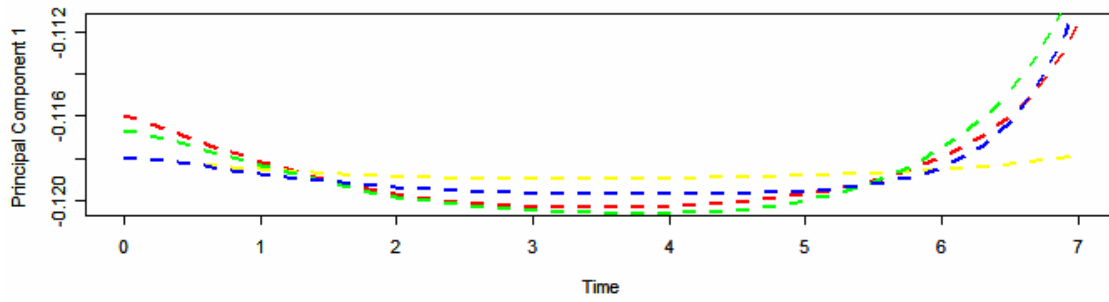
Appendix M

Click to advance to next plot

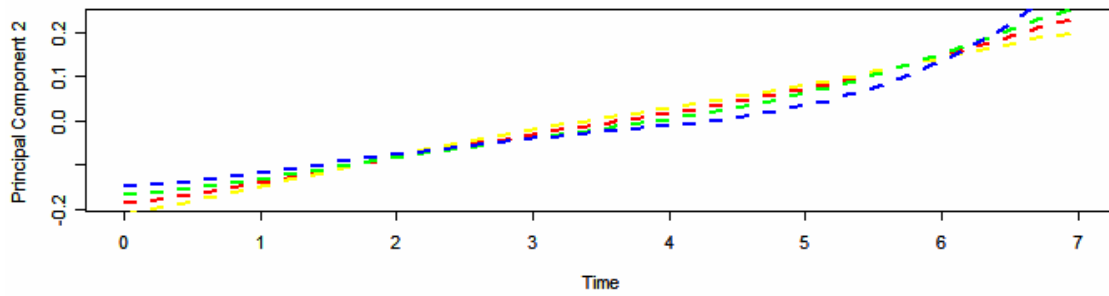


Appendix N

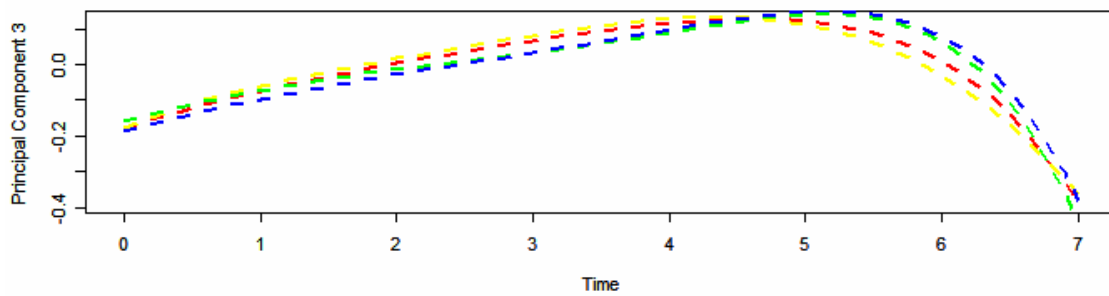
Plot of Time Versus Position Product PC1



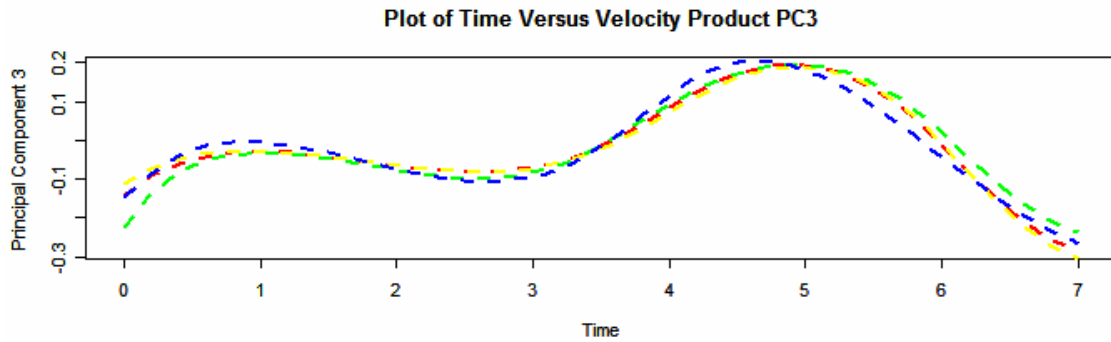
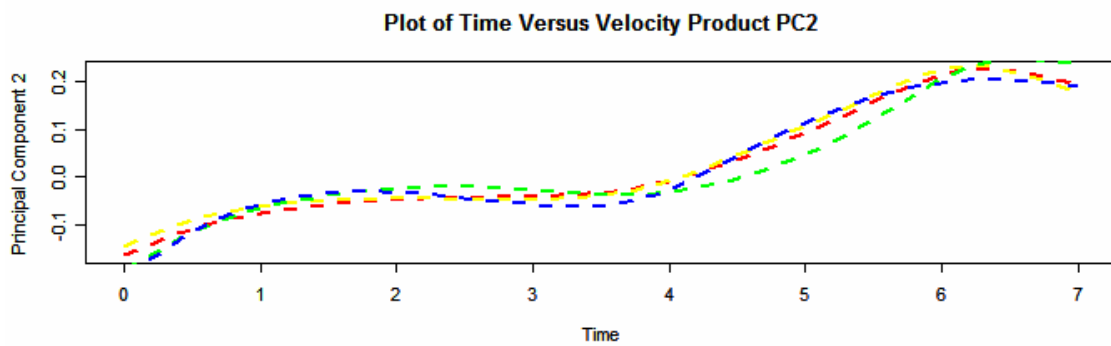
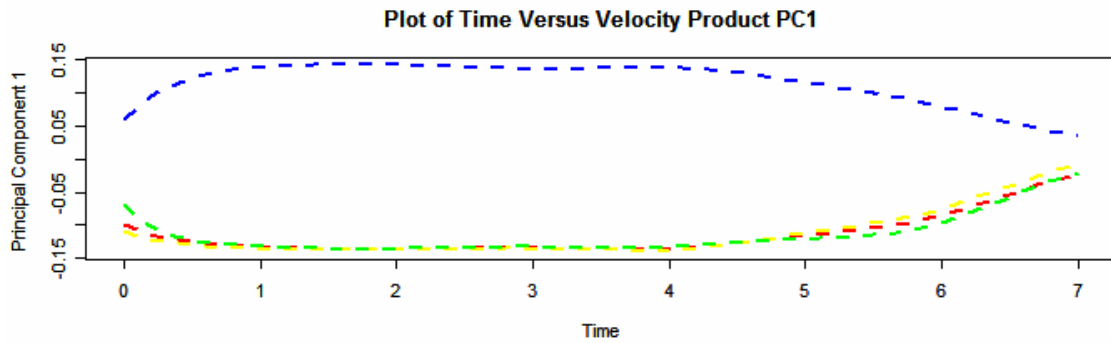
Plot of Time Versus Position Product PC2



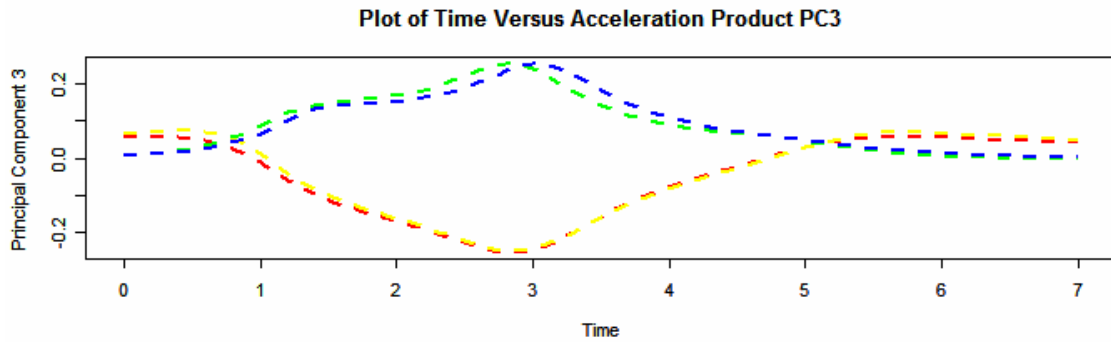
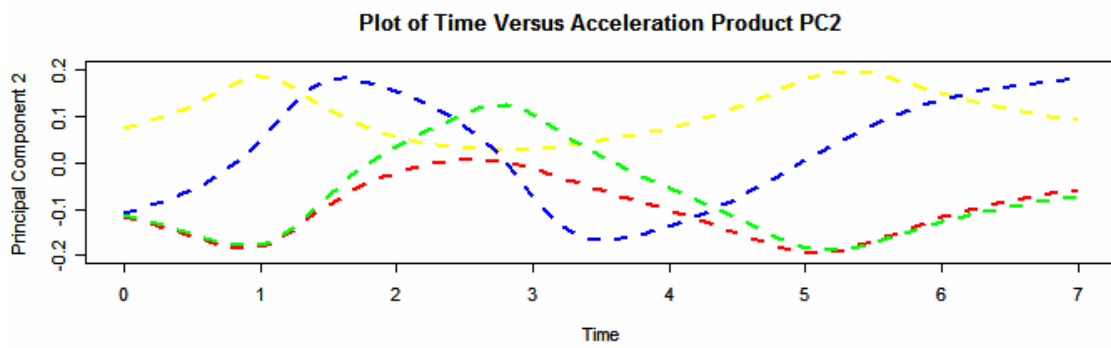
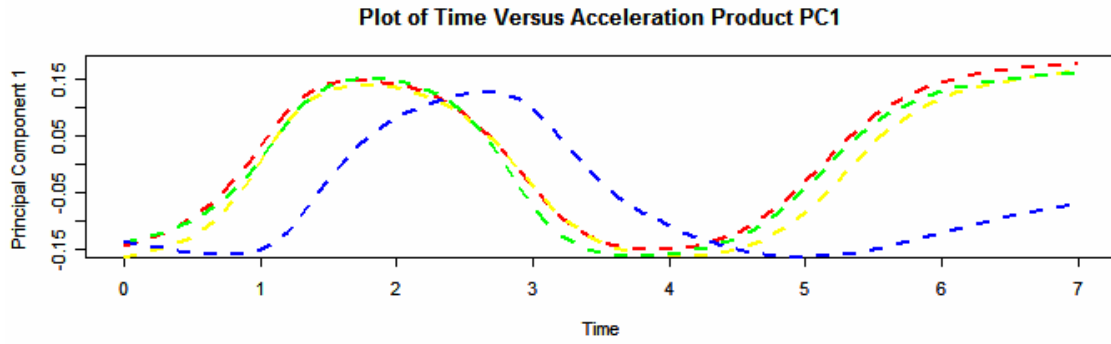
Plot of Time Versus Position Product PC3



Appendix O

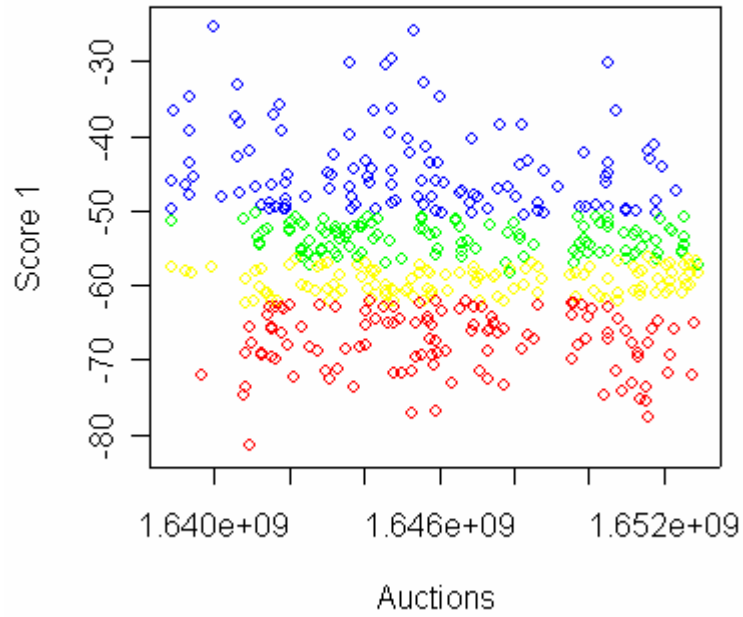


Appendix P

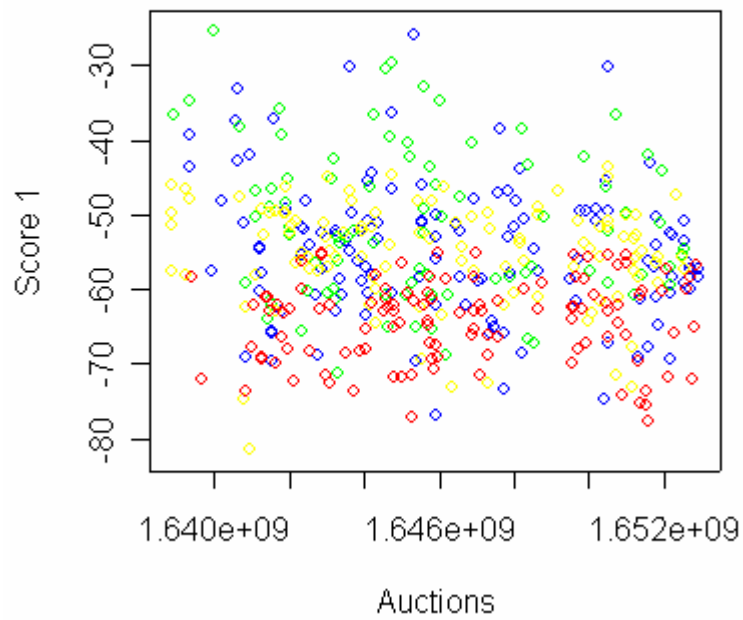


Appendix Q

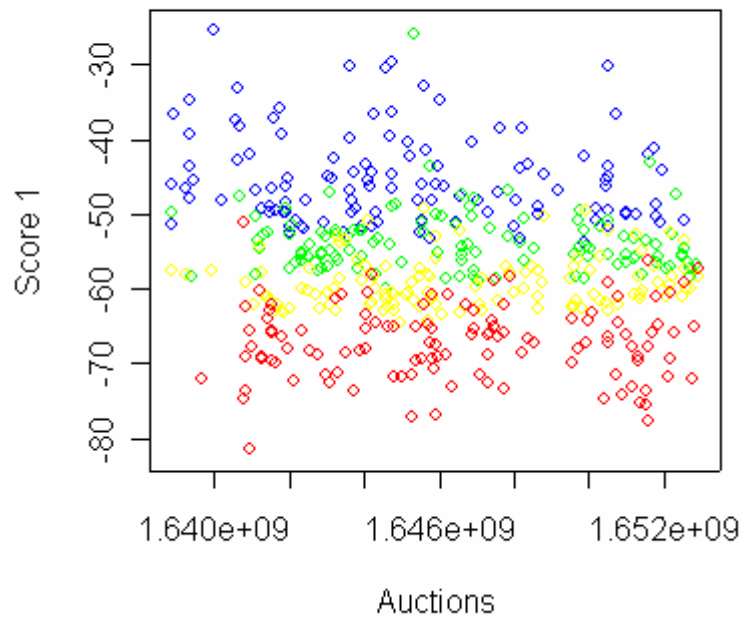
Divided by Ypred Middle



Divided by opening bid



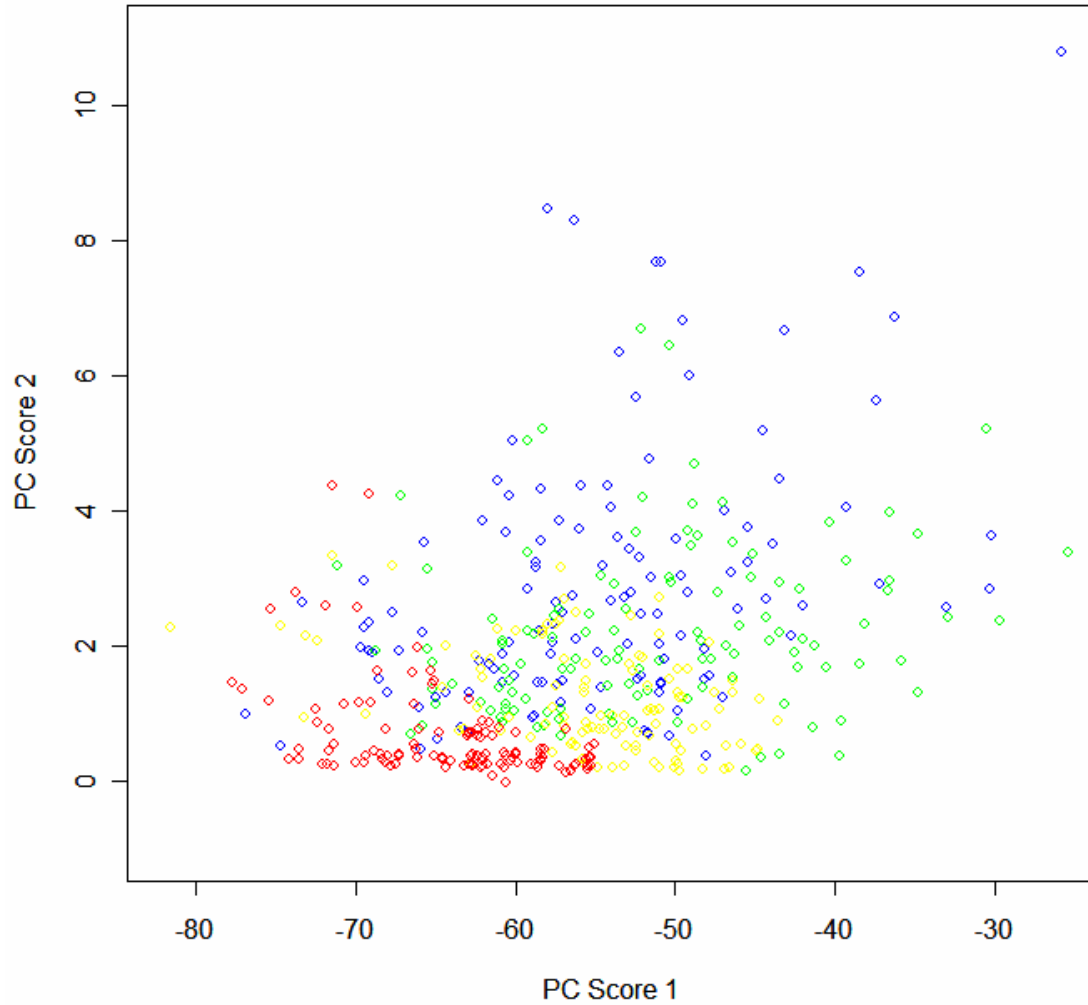
Divided by high bid



- Blue: 1st quartile
- Green: 2nd quartile
- Yellow: 3rd quartile
- Red: 4th quartile

Appendix R

Colors Correspond to Quartiles of Opening Bid



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.01	1.00	100.00	509.70	600.00	6500.00

Blue: 1st Quartile

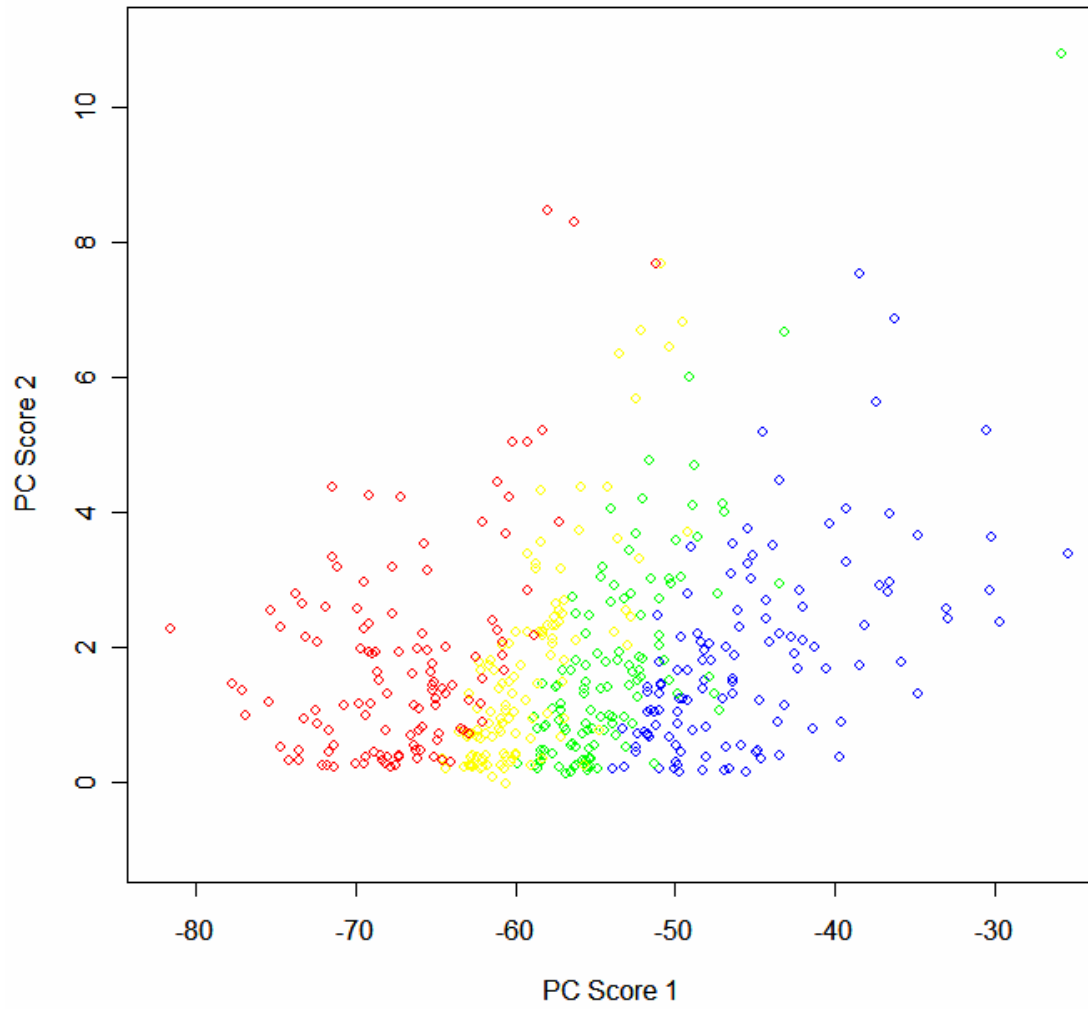
Green: 2nd Quartile

Yellow: 3rd Quartile

Red: 4th Quartile

Appendix S

Colors Correspond to Quartiles of High Bid



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
70	650	1300	2019	2206	24500

Blue: 1st Quartile

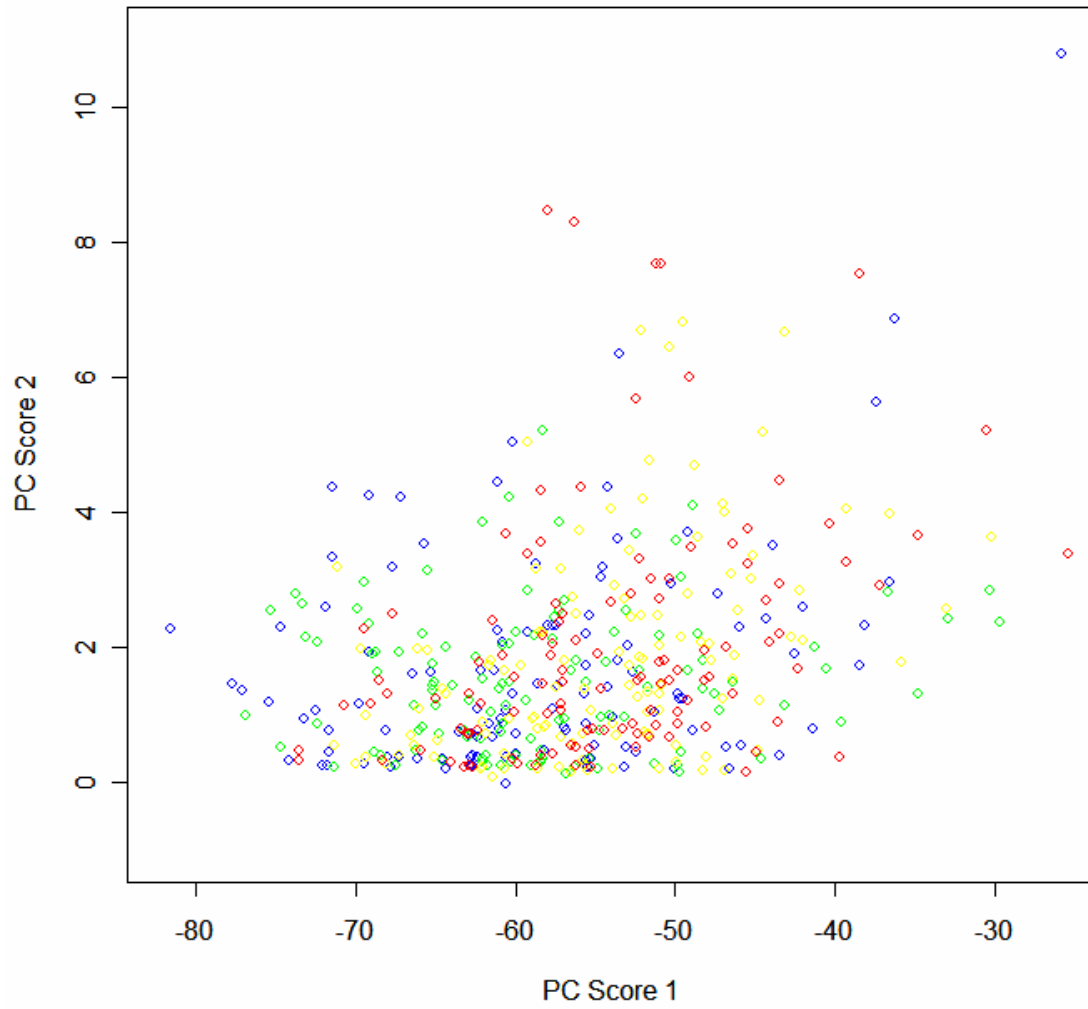
Green: 2nd Quartile

Yellow: 3rd Quartile

Red: 4th Quartile

Appendix T

Colors Correspond to Quartiles of Seller Experience



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.00	27.75	107.00	572.00	358.00	9055.00

Blue: 1st Quartile

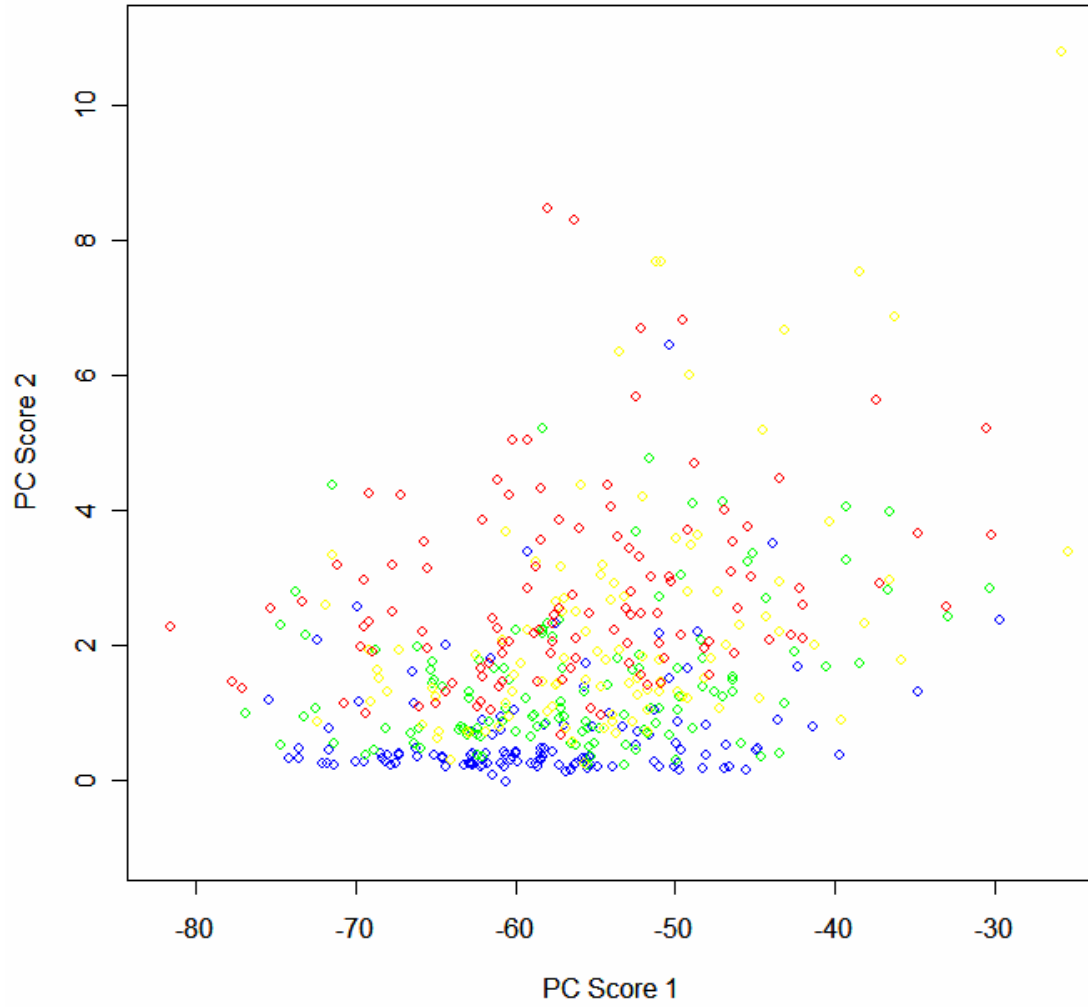
Green: 2nd Quartile

Yellow: 3rd Quartile

Red: 4th Quartile

Appendix U

Colors Correspond to Quartiles of Number of Bids



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	7.00	13.00	14.65	20.00	57.00

Blue: 1st Quartile

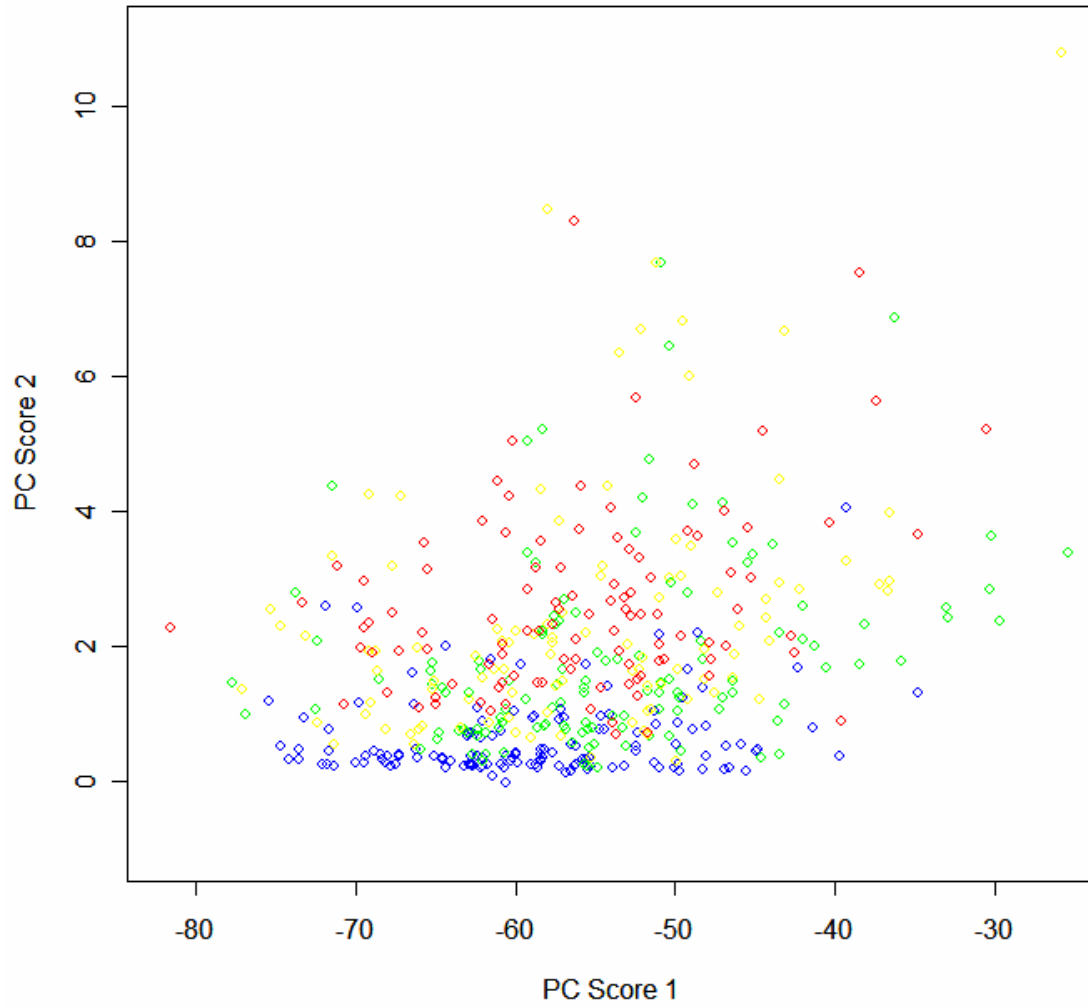
Green: 2nd Quartile

Yellow: 3rd Quartile

Red: 4th Quartile

Appendix V

Colors Correspond to Quartiles of Number of Unique Bidders



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	4.000	7.000	7.377	10.000	21.000

Blue: 1st Quartile

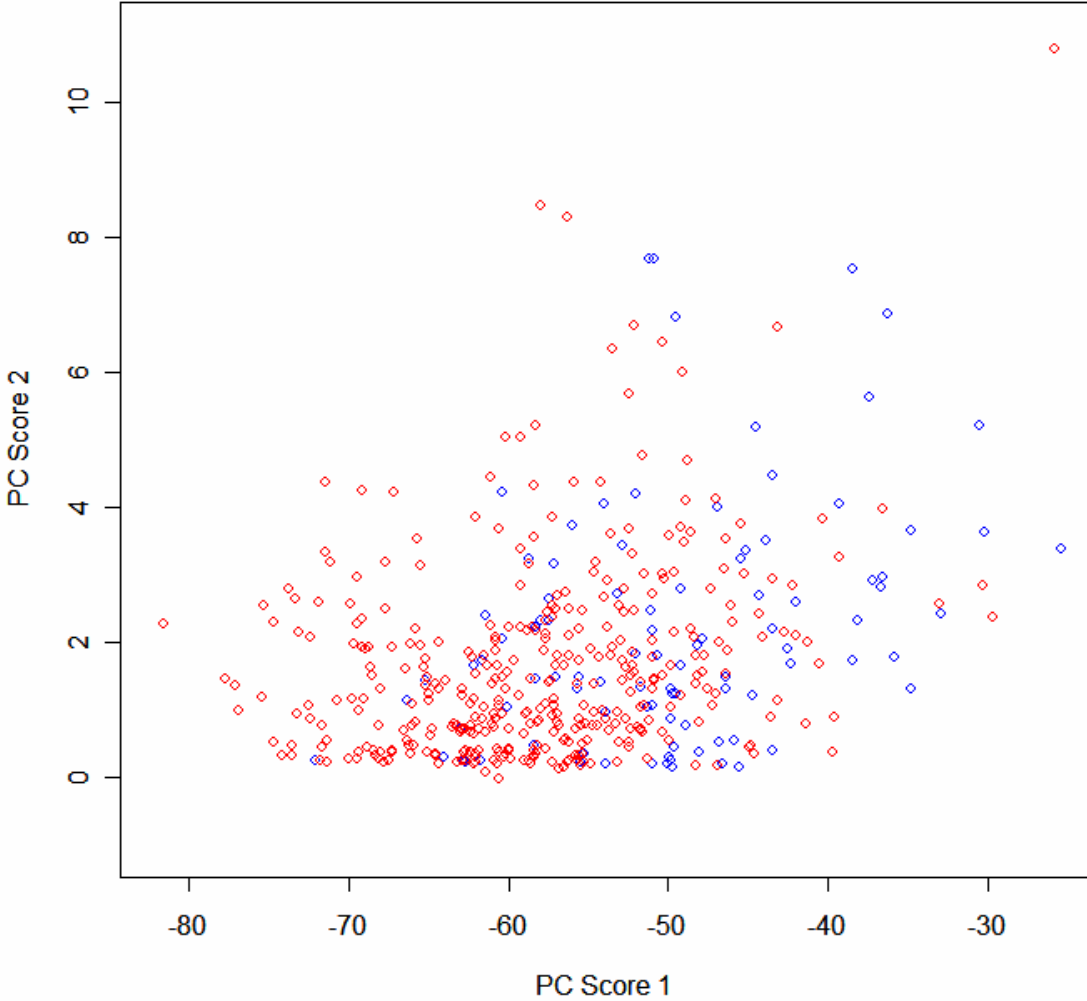
Green: 2nd Quartile

Yellow: 3rd Quartile

Red: 4th Quartile

Appendix W

Colors Correspond to Watch Type, Blue=Cartier, Red=Rolex



Appendix X

```
<?php
echo "Hello";
//Connect to "assignments" database using username 'root' and password
'root'
$conn = mysql_connect('localhost', 'root', 'root');
mysql_select_db('assignments', $conn);
//Retrieve all rows from "contacts" table
$sql1 = "SELECT * FROM contacts;";
$result = mysql_query($sql1, $conn);
//Show total number of rows extracted by the SQL query
$numRows = mysql_num_rows($result);
echo " Number of rows is $numRows <p>";
//Display table header
$displayString1 = "<table>
<th colspan='10' rowspan='0' bgcolor='#FFFFFF'>Names and
Email</th></table>";
echo $displayString1;
while ($namesArray = mysql_fetch_array($result)) {
    $Item_Number=$namesArray['Item_Number'];
    $Start_time = $namesArray['Start time'];
    $Ends = $namesArray['Lastname'];
    $History = $namesArray['Bid History'];
    $UserID=$namesArray['UserId'];
    $Bid_Amount=$namesArray['Bid Amount'];
    $Date_of_Bid=$namesArray['Date of Bid'];
    $displayString = "
    <table>
    <tr align='center' bgcolor='#CCFFFF' style='border-style:
solid'>
    <td colspan='20' rowspan='0'> $Item_Number</td>
    <td colspan='20' rowspan='0'> $Start_time</td>
    <td colspan='20' rowspan='0'> $Ends</td>
    <td colspan='20' rowspan='0'> $History</td>
    <td colspan='20' rowspan='0'> $UserID</td>
    <td colspan='20' rowspan='0'> $Bid_Amount</td>
    <td colspan='20' rowspan='0'> $Date_of_Bid</td>

    </tr>
    </table>";
    echo $displayString;
}
?>
```

This some extra code that I have gotten a chance to test yet either.

```
<?php

//function to print out an array
function displayArray($currentArray) {
    $i=0;
    foreach ($currentArray as $k) {
```

```

        echo "element number: ". $i. " " . $k;
        echo "<br>". "\n";
    }
}

function getEbayFrontPageFile ($urlName, $fileName) {

    $fd = fopen($urlName , 'r');
    $counter =0;
    //open output file handle
    $fileOutputHandle = fopen($fileName, 'a');

    while (!feof($fd)) {
        $buffer = fgets($fd, 8096);
        if (preg_match('/findSortArray =/',
$buffer)) {
            $buffer = fgets($fd, 58096);
            $temp = preg_split("/item="/, $buffer);
            displayArray($temp);
            foreach ($temp as $k) {
                if ($counter == 0) {
                    $buffer1 = substr($k,
0,10). "\n";
                    $counter ++;
                }
                else {
                    $buffer1 = substr($k,
0,10). "\n";
                    echo $buffer1;
                    echo "<br>". "\n";

                    fwrite($fileOutputHandle, $buffer1);
                    $counter ++;
                }
            }
        }
    }

    fclose($fd);
    fclose($fileOutputHandle);

} // end of function getEbayFrontPageFile

function getGivenDateeBayAuctionIDs() {
//open file that contains list of all auctions

$ctr = 0;
$localfile = "091704eBayAuctionIDsGucci.txt";
set_time_limit(600000);
//getEbayFrontPageFile($url, $localfile);
while ($ctr <= 529) {

```

```

        $url = "http://search.ebay.com/gucci-  
handbag\_W0QQfromZR8QQsorecordstoskipZ"  
        . trim($ctr) ;  
        $ctr= $ctr+50;  
        //echo $url;  
        //echo "<br>". "\n";  
        //if ($ctr == 200) break;  
        //$fileFrontName = $urlFront . $fileNamebuffer;  
        getEbayFrontPageFile($url, $localfile);  
    }  
}  
set_time_limit(30);  
getGivenDateeBayAuctionIDs();  
  
?>

```

```

function getWebSiteFrontPageFile ($urlName, $fileName, $directoryName)
{

/*echo $urlName;
echo "<br>". "\n";
echo $fileName;
echo "<br>". "\n";*/

$fd = fopen($urlName , 'r') or die();
$ctr = 0;

//open output file handle where URL's content will be stored
$outFrontFileName = trim($directoryName) . trim($fileName);
echo "writing $outFrontFileName";
echo "<br>". "\n";

$fileOutputHandle = fopen($outFrontFileName, "w");

while (!feof($fd)) {
    $buffer = fgets($fd, 4096);
    fwrite($fileOutputHandle, $buffer);
    $ctr++;
}

fclose($fd);
fclose($fileOutputHandle);

} // end of function getWebSiteFrontPageFile

```

Such a function can be called using a command such as

```

getWebSiteFrontPageFile(someURL, someFileName, someLocalPath );

        Lets try above for this auction
        Reading input from text files to grab a set of files
function processAllFiles() {
    //open file that contains list of all auctions
    $auctionListFile = fopen("problemRemainingAucs.txt", "r");

    //$buffer = fgets($auctionListFile, 1024); //skip first line
    $urlFront =
http://offer.ebay.com/ws3/eBayISAPI.dll?ViewBids&item=;
    $ctr = 0;
    set_time_limit(600000);
    while (!feof($auctionListFile)) {
        //$fileNamebuffer = fgets($auctionListFile, 4096);
        $fileNamebuffer = fgets($auctionListFile, 4096);
        $ctr++;
        /*echo $fileNamebuffer;
        echo "<br>". "\n";*/
        //if ($ctr == 2) break;
        $fileFrontName = $urlFront . $fileNamebuffer;

        $directoryName = "/some_path/";

        getWebSiteFrontPageFile($fileFrontName, $fileNamebuffer,
        $directoryName );
    }
    set_time_limit(30);
    fclose($auctionListFile);
}

```

Try this with the following text file that contain auction ids for Gucci handbags.

(091004eBayAuctionIDsGUCCI.txt)

Processing all files in a folder

(think about when you might need to do this)

```

//takes a directory name and process all file in directory
//checks whether a certain file has .html extension
function processAllFilesInDirectory($directoryName) {
    $i = 0;
    $directoryHandle = opendir($directoryName) or die($php_errormsg);
    while (false !== ($fileHandle = readdir($directoryHandle))) {
        $path_parts = pathinfo($directoryName. "/"$fileHandle);
        /*echo $path_parts["dirname"];
        echo "<br>". "\n";*/

if ($i > 6810) {
    if ($path_parts["extension"] == "html") {
        echo $path_parts["basename"];
        echo "<br>". "\n";
        echo $path_parts["extension"];
        echo "<br>". "\n";*/
    }
}

```

```
$urlFront =
"http://offer.ebay.com/ws3/eBayISAPI.dll?ViewBids&item=";

$fileNamebuffer = $path_parts["basename"];
/*echo $fileNamebuffer;
echo "<br>". "\n";*/
//if ($i == 4) break;
$fileFrontName = $urlFront . $fileNamebuffer;
$temp = $directoryName . "/some_path/";
getWebSiteFrontPageFile($fileFrontName, $fileNamebuffer, $temp);
}
}
$i++;

} //close while loop to process all files
} // close function processAllFilesInDirectory1
```