**Analysis of revenue sources for wineclubreviews.net**

by: Ken Wallace

***Introduction***

I have chosen to analyze data from an affiliate marketing business, wineclubreviews.net.

I’m looking to gain a solid understanding of the performance of the top performing wine clubs by looking at revenue by club group, bottle price, shipping costs, and over time. Using sales data, collected over more than 4 years, and sales trends, I hope to predict which wine clubs are trending upward, and which, if any, would be beneficial to highlight or otherwise promote on the website.

***Data Sources***

The most important fields and information that the data set provides are:

* + - Click data for over 635,000 click observations of 77 variables, including sales and commissions for each club, a unique session id for each user, and when each sale was converted, with more than 27000 sales over a nearly 4.5 year period
    - Dates for most, if not all, sales, which will allow me to show trends over time as well as within and between specific peak (i.e. holiday) periods, such as the time between Thanksgiving and Christmas, Valentine's’ Day, Mother’s Day, and Father’s Day.
    - There are 832 observations of 66 variables for the clubs, which would allow comparisons of almost any aspect one could consider in terms of impact to revenue. The variable data include club and club group identifiers, club ratings, and detailed information about the clubs such as price per bottle, number of bottles in the club, shipping costs, and club offerings

The data seem complete, so the only thing that may be difficult to assess are evaluations beyond my skill level. I have only performed limited regressions for modelling future predictions, and the error seems higher than I would expect, so this either means that I’m choosing the wrong variables or methods, or the data is not predictive of future results.

In terms of data wrangling, the data I received came in two .csv files, one with click data and one with club specific data, which I merged together in order to pull the necessary information to compare sales data by club.

In addition, the data was pretty clean, though because there are over 635K observations, the main filtering I did was to work mainly with the observations where an actual sale was made (i.e. revenue earned). I also created a couple of new variables, including a revenue column, which is a sum of the totals for each group, and a column called cdate, which is the converted date variable in date format.

***Findings from Exploratory Data Analysis:***

Glimpse of data (using dplyr::glimpse):

Observations: 27,704

Variables: 19

$ clubgroup\_id.x (chr) "31", "70", "51", "70", "53", "31", "31", "27", "42", "49"...

$ club\_id (chr) "1124", "1001", "1116", "1001", "1133", "1124", "1123", "1...

$ sale\_amt (chr) "192.9", "150", "135.8", "150", "93.9", "527.4", "43.95", ...

$ cdate (date) 2011-10-01, 2011-10-01, 2011-10-01, 2011-10-01, 2011-10-0...

$ total (dbl) 38.58, 45.00, 27.16, 45.00, 18.78, 105.48, 8.79, 4.20, 10....

$ price (dbl) 46.00, 21.95, 39.00, 21.95, 130.00, 46.00, 46.00, 37.00, 2...

$ price\_per\_bottle (dbl) 23.00, 16.35, 19.50, 16.35, 10.83, 23.00, 23.00, 23.75, 14...

$ bottles (dbl) 2, 2, 2, 2, 12, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 12, 2,...

$ shipfreq (int) 3, 6, 3, 6, 2, 3, 3, 6, 0, 0, 3, 3, 6, 3, 3, 6, 4, 2, 3, 6...

$ selcolor (dbl) 2, 4, 2, 4, 2, 2, 2, 1, 1, 4, 2, 4, 4, 2, 2, 4, 4, 4, 2, 4...

$ shipprice (dbl) 0.00, 10.75, 0.00, 10.75, 0.00, 0.00, 0.00, 10.50, 0.00, 1...

$ shipdesc (int) 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1...

$ all\_around (chr) "3.45", "3.95", "3.45", "3.95", "3.45", "3.45", "3.45", "4...

$ rating (chr) "4.307692308", "4.923076923", "4.5", "4.923076923", "4.461...

$ selcolor.name (chr) "Red-only option", "Red-only and White-only options", "Red...

$ company\_name (chr) "Cellars Wine Club", "Wine of the Month Club &reg;", "Cell...

$ converted\_at (chr) "10/1/2011 8:17", "10/1/2011 0:00", "10/1/2011 15:13", "10...

$ keywords (chr) "international global ecard holiday-sale last-minute cyber...

$ session\_id (chr) "8ac997b571eccd0860194d30c6dea40f", "86acf2334c5ce590160d1...

Summary of main subset of data (using psych::describe):

vars n mean sd median trimmed mad min

clubgroup\_id.x\* 1 27704 80.04 55.61 55.00 73.97 41.51 1.00

club\_id\* 2 27704 1277.07 210.61 1190.00 1263.12 148.26 1001.00

sale\_amt\* 3 27704 155.92 145.03 128.00 133.82 101.81 0.00

cdate\* 4 27628 NaN NA NA NaN NA Inf

total 5 27704 36.33 28.65 31.20 32.70 16.01 0.80

price 6 27704 76.23 58.25 39.95 69.22 14.83 21.95

price\_per\_bottle 7 27704 21.34 9.78 22.50 20.34 5.90 5.55

bottles 8 27704 4.60 4.08 2.00 4.01 0.00 1.00

shipfreq 9 27704 3.15 2.48 3.00 3.15 4.45 0.00

selcolor 10 27704 3.19 1.58 4.00 3.15 0.00 0.00

shipprice 11 27704 9.47 7.39 10.50 9.28 14.07 0.00

shipdesc 12 27704 0.70 0.46 1.00 0.75 0.00 0.00

all\_around\* 13 27704 7.87 25.17 4.72 4.35 0.41 2.55

rating\* 14 27704 4.82 0.58 4.92 4.89 0.68 0.00

selcolor.name\* 15 27704 NaN NA NA NaN NA Inf

company\_name\* 16 27704 NaN NA NA NaN NA Inf

converted\_at\* 17 27704 NaN NA NA NaN NA Inf

keywords\* 18 27704 NaN NA NA NaN NA Inf

session\_id\* 19 27704 NaN NA NA NaN NA Inf

max range skew kurtosis se

clubgroup\_id.x\* 230.00 229.00 0.83 -0.61 0.33

club\_id\* 1887.00 886.00 0.73 -0.57 1.27

sale\_amt\* 4724.12 4724.12 6.54 105.63 0.87

cdate\* -Inf -Inf NA NA NA

total 944.83 944.03 7.47 139.74 0.17

price 1800.00 1778.05 3.44 44.70 0.35

price\_per\_bottle 441.90 436.35 9.43 245.31 0.06

bottles 12.00 11.00 1.20 -0.45 0.02

shipfreq 11.00 11.00 0.16 -1.30 0.01

selcolor 15.00 15.00 0.20 0.74 0.01

shipprice 90.00 90.00 0.12 -0.23 0.04

shipdesc 1.00 1.00 -0.86 -1.27 0.00

all\_around\* 188.00 185.45 6.93 46.15 0.15

rating\* 5.46 5.46 -0.99 1.50 0.00

selcolor.name\* -Inf -Inf NA NA NA

company\_name\* -Inf -Inf NA NA NA

converted\_at\* -Inf -Inf NA NA NA

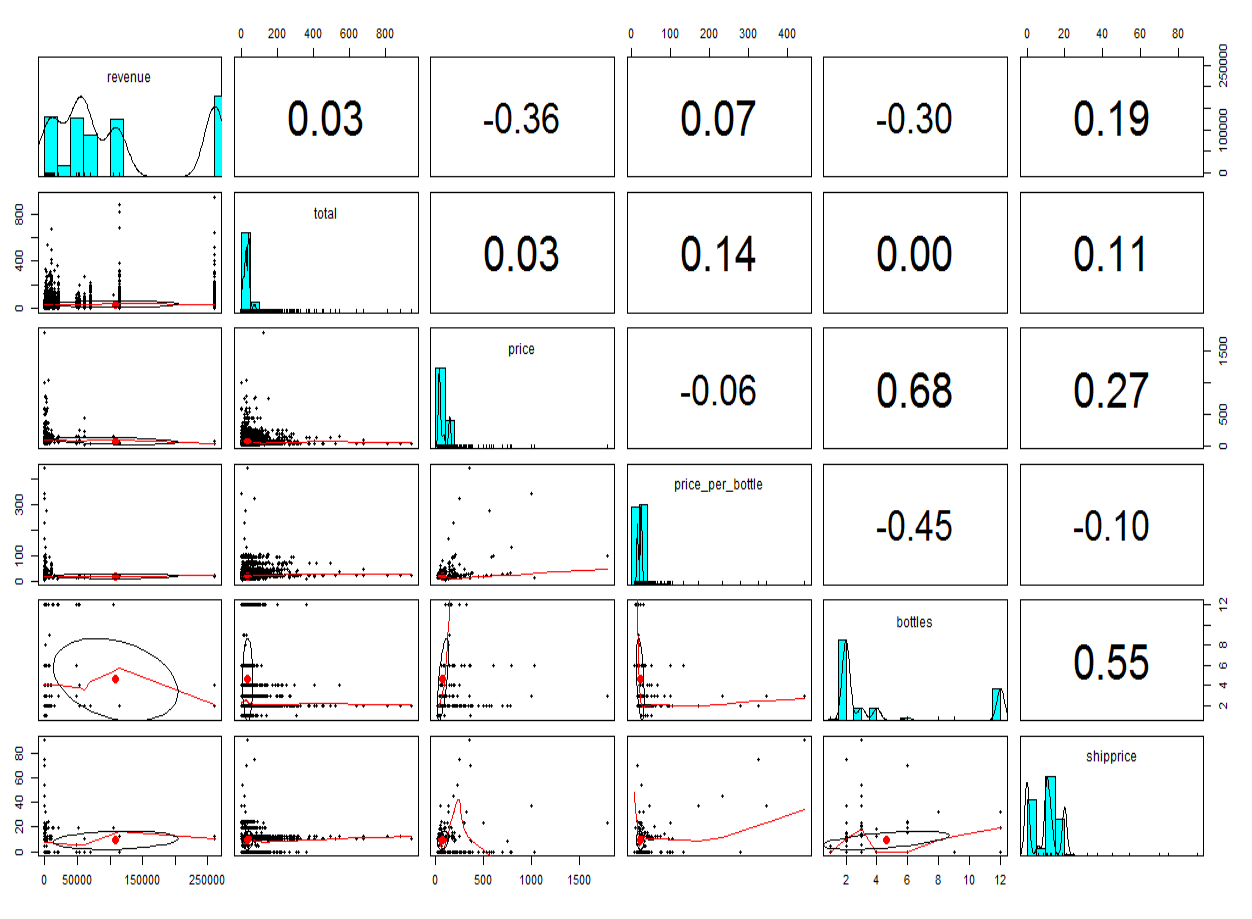
keywords\* -Inf -Inf NA NA NA

session\_id\* -Inf -Inf NA NA NA

In order to further subset the data, I’ve taken the bulk of the numeric variables and run a correlation (stats::cor) on those. This can be shown in the following two ways:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | revenue | total | price | price\_per\_bottle | bottles | shipprice |
| revenue | 1.00 | 0.03 | -0.36 | 0.07 | -0.30 | 0.19 |
| total | 0.03 | 1.00 | 0.03 | 0.14 | 0.00 | 0.11 |
| price | -0.36 | 0.03 | 1.00 | -0.06 | 0.68 | 0.27 |
| price\_per\_ bottle | 0.07 | 0.14 | -0.06 | 1.00 | -0.45 | -0.10 |
| bottles | -0.30 | 0.00 | 0.68 | -0.45 | 1.00 | 0.55 |
| shipprice | 0.19 | 0.11 | 0.27 | -0.10 | 0.55 | 1.00 |

And:

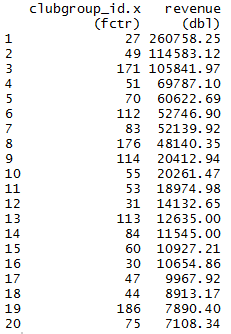


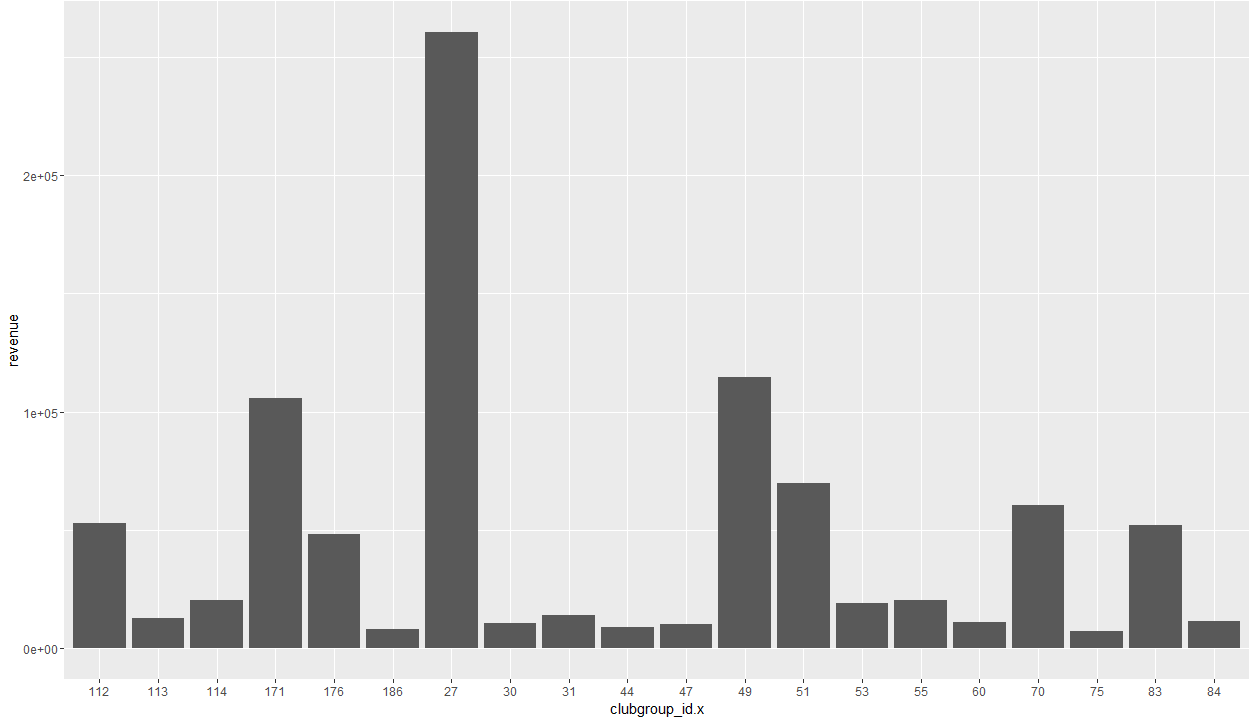
Some of the strongest correlations are in line with expectations, in that it is expected that the more bottles in a shipment, the more that shipment would cost (correlation 0.55), and the more bottles in the shipment the more the shipment would cost (correlation 0.63). The next highest correlation, which happens to be negative, is the relationship between number of bottles and price per bottle. I wouldn’t necessarily expect the correlation for this to be so strongly negative, as shipments could contain any number of bottles of any price, but it appears that the cheaper bottles do sell in larger amounts (and the more expensive bottles are more likely to sell in smaller shipments).

In addition, there is a moderately negative correlation (-0.36) between price and revenue, which indicates that the lower priced clubs yield a greater portion of the revenue for the company. This follows the summary shown below of revenue broken down by the number of bottles in a shipment, where 2 bottle shipments outsell the next category by over 2:1. That being said, the second highest revenue category is 12 bottle shipments. It could be interesting to see the revenue breakdown amongst 2 bottle and 12 bottle shipments.

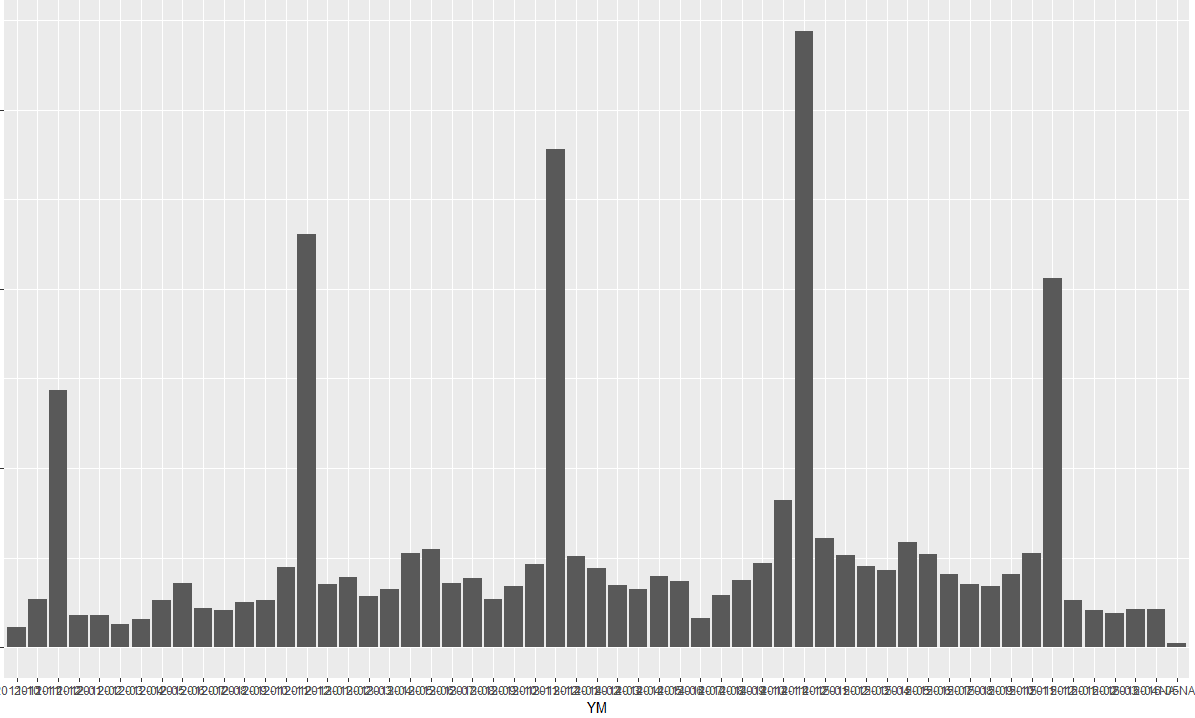
Some other interesting results, below, that can be pulled from the data:

* + Top 20 club groups by revenue



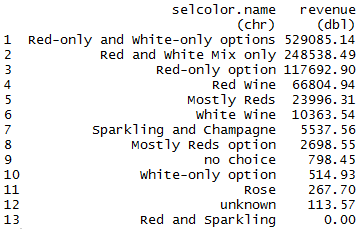


* + Total revenue by month



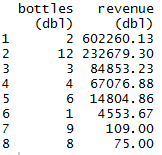
The spikes occur every December, with smaller spikes occurring in May (Mother’s Day), June (Father’s Day), and February (Valentine’s Day).

* + Revenue by club wine color choices

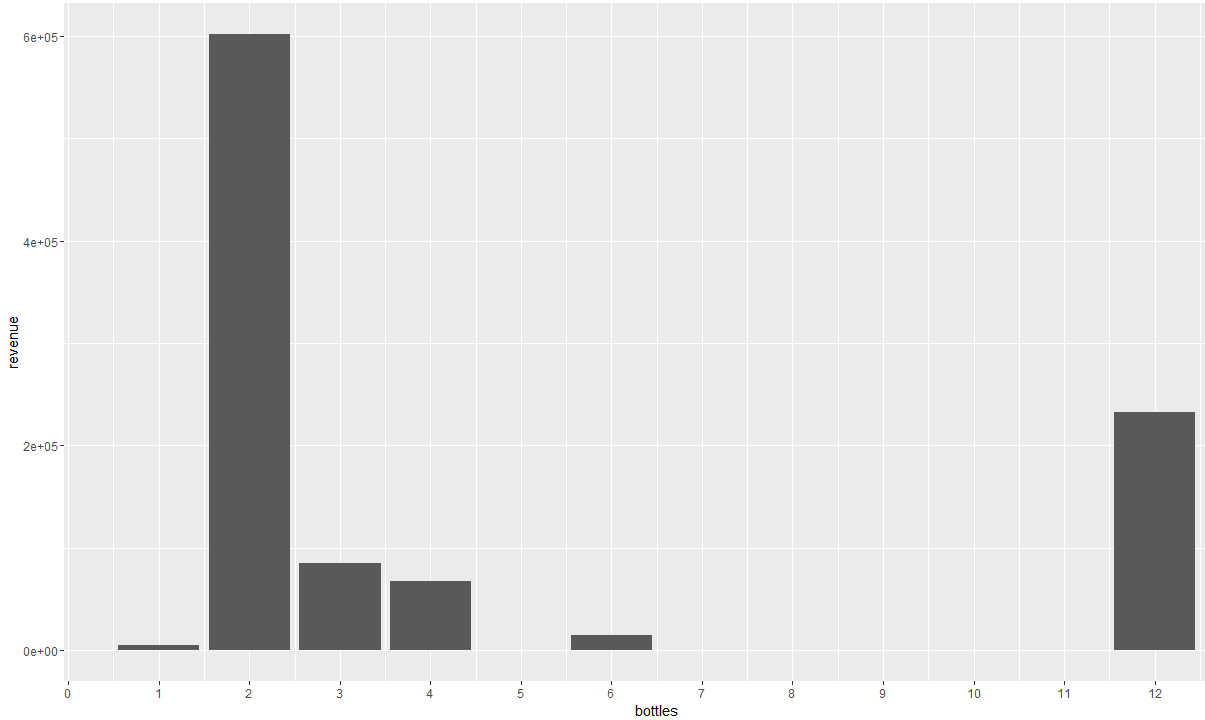


Customers strongly prefer clubs that offer choices including red wine in their shipments over any other choices by over 48 to 1.

* + Revenue by # of bottles in each shipment

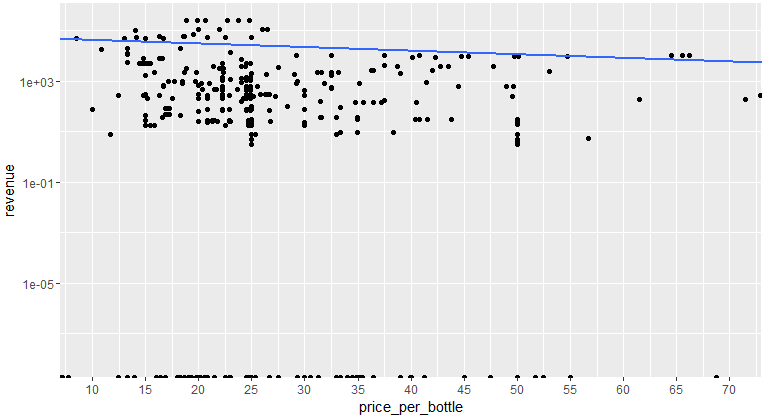


2 bottle shipments account for more than 2x the next highest shipment quantity, and 2 and 12 bottle shipments combined account for over 80% of revenue over the past 4.5 years.



The interesting thing here is that 2 bottle and 12 bottle shipments are the highest sellers, or at least generate the highest revenue for WCR.

* + Scatterplot of revenue by average price per bottle



The interesting thing about this is the cluster around $25/bottle

***Predictions***

Linear Model

In creating a linear model for this, I chose the following variables:

* Price
* Price per bottle
* Bottles (i.e. number of bottles in a shipment)
* Shipprice (the price for each shipment)

Call:

lm(formula = revenue ~ price + price\_per\_bottle + bottles + shipprice,

data = data.train)

Coefficients:

(Intercept) price price\_per\_bottle bottles shipprice

154119.3 -185.3 -1564.2 -14241.3 7205.5

> summary(lm.1)

Call:

lm(formula = revenue ~ price + price\_per\_bottle + bottles + shipprice,

data = data.train)

Residuals:

Min 1Q Median 3Q Max

-237624 -59519 -18119 103471 148436

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 154119.3 6941.8 22.202 < 2e-16 \*\*\*

price -185.3 48.3 -3.837 0.000128 \*\*\*

price\_per\_bottle -1564.2 267.5 -5.848 5.84e-09 \*\*\*

bottles -14241.3 905.9 -15.721 < 2e-16 \*\*\*

shipprice 7205.5 314.7 22.899 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78870 on 1917 degrees of freedom

Multiple R-squared: 0.3218, Adjusted R-squared: 0.3203

F-statistic: 227.4 on 4 and 1917 DF, p-value: < 2.2e-16

>

> # Use the model on the training set:

> lm1.train <- predict(lm.1,data=data.train)

>

> # Show the square root of the mean squared error (MSE): known actual - predicted

> sqrt(mean((lm1.train-data.train$revenue)^2) )

[1] 78765.32

>

> # Similarly, use the model on the testing set and solve for the square root of mean squared error:

> lm1.test <- predict(lm.1,data.test)

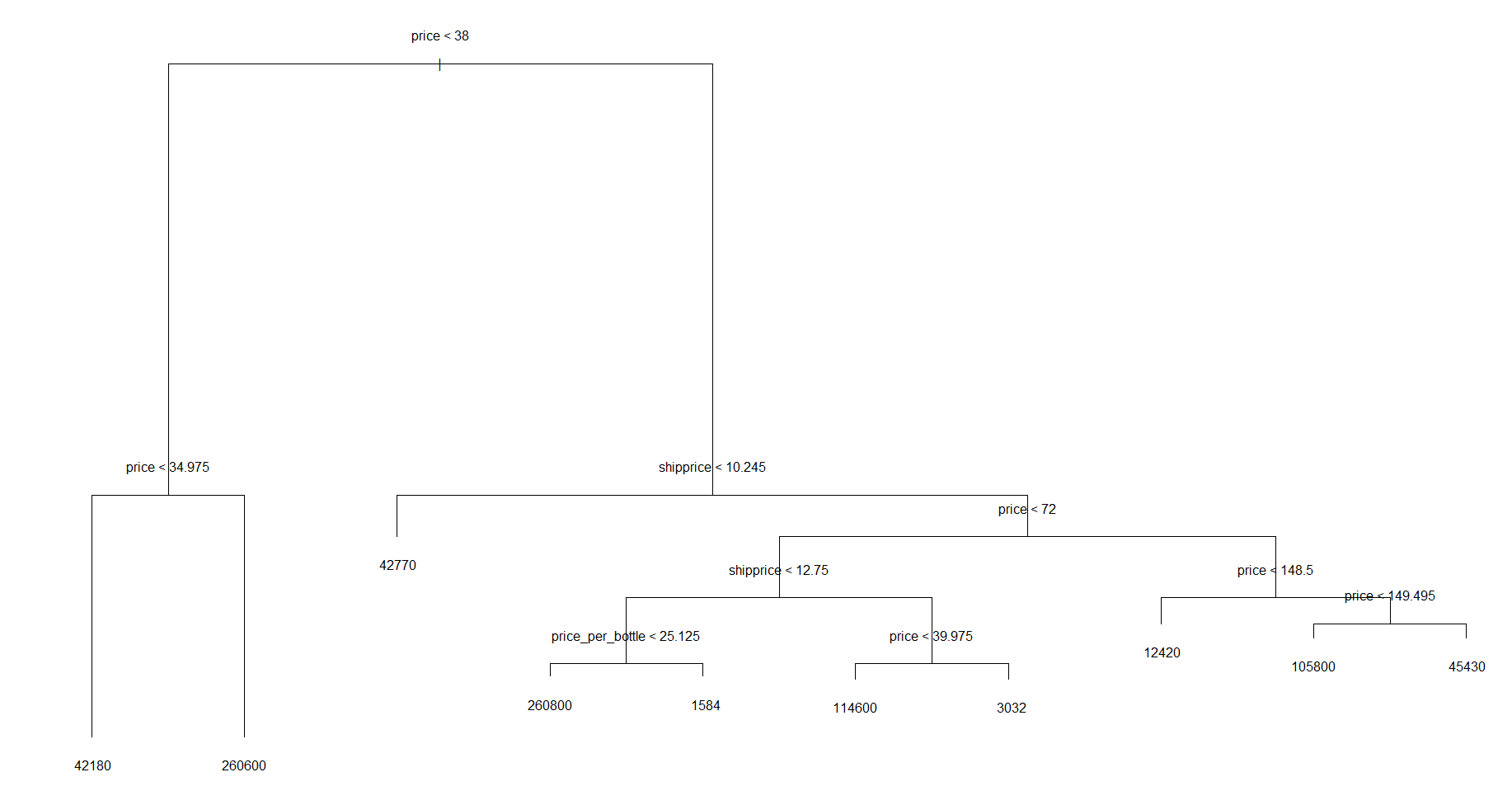
> sqrt(mean((lm1.test-data.test$revenue)^2))

[1] 79028.98

All of the variables in this test appear to be significant, though there may be some multicollinearity happening. The model seems to be error prone, with an RMS value of approximately $79000 for both the training set and the test set.

Trees

Running the tree() function from library(tree) for revenue against all other variables, the result is the following:



What this is telling me is that a large bulk of the revenue come from clubs priced between $35-72, and that shipping costs are not a barrier to purchasing clubs. From this it can be inferred that club selection and value would be more highly regarded by actual customers than shipping cost. I hypothesized that clubs that charge a shipping cost would be less desirable than ones including free shipping, but this just doesn’t appear to be the case based on the tree. Another indicator of that is the relatively low correlations between shipprice and total and revenue.

Random Forest

Running a Random Forest from library(randomForest) yields the following results:

Call:

randomForest(formula = revenue ~ ., data = wine.train.all, importance = TRUE, ntree = 500)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 204124463

% Var explained: 97.77

> importance(rf.1) # Shows the effect on MSE for each specific variable

%IncMSE IncNodePurity

total 19.60801 8.072099e+11

price 37.03576 5.589265e+12

price\_per\_bottle 33.73773 4.124589e+12

bottles 30.97961 1.517100e+12

shipprice 39.96378 4.084574e+12

> varImpPlot(rf.1)

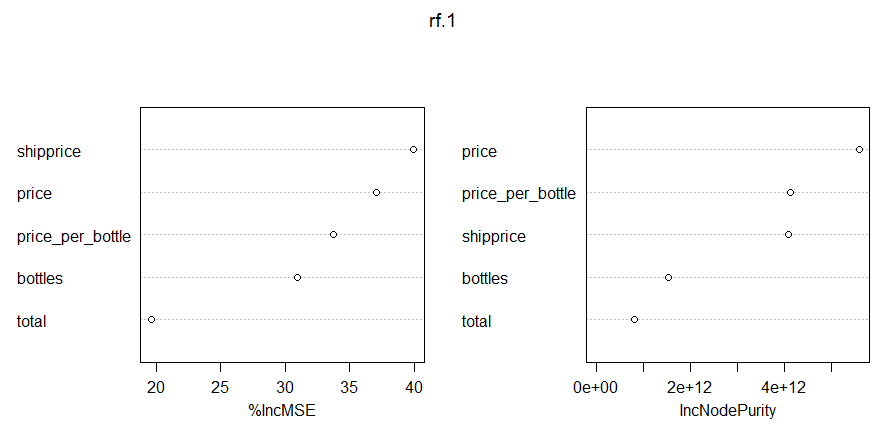
>

> rf.test <- predict(rf.1,wine.test.all)

> sqrt(mean((rf.test-data.test$revenue)^2))

[1] 13679.46

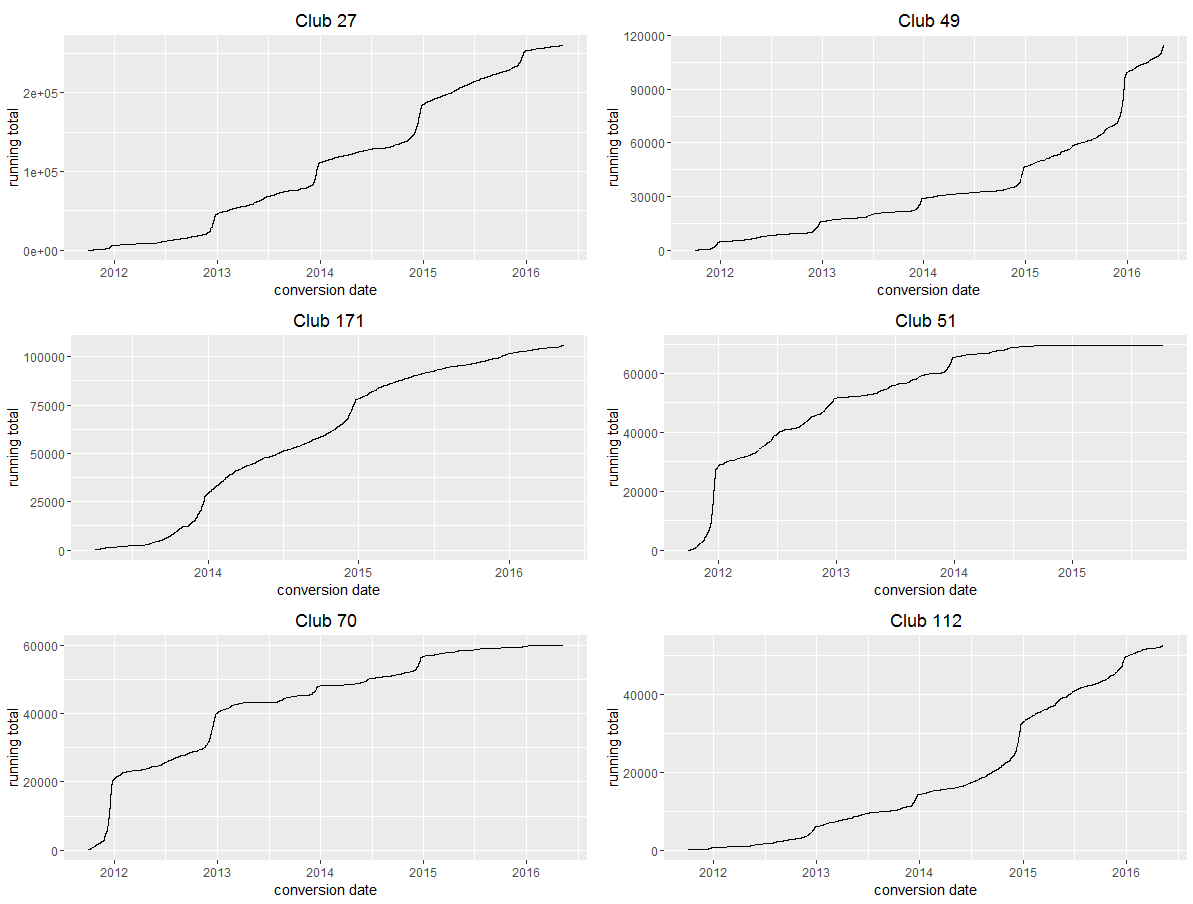
It appears as though price, price\_per\_bottle, and shipprice are the most significant variables of the ones tested. The RMS error is under $14000, which is an improvement in the error compared with the linear model by over 80% ((79000-14000)/79000).



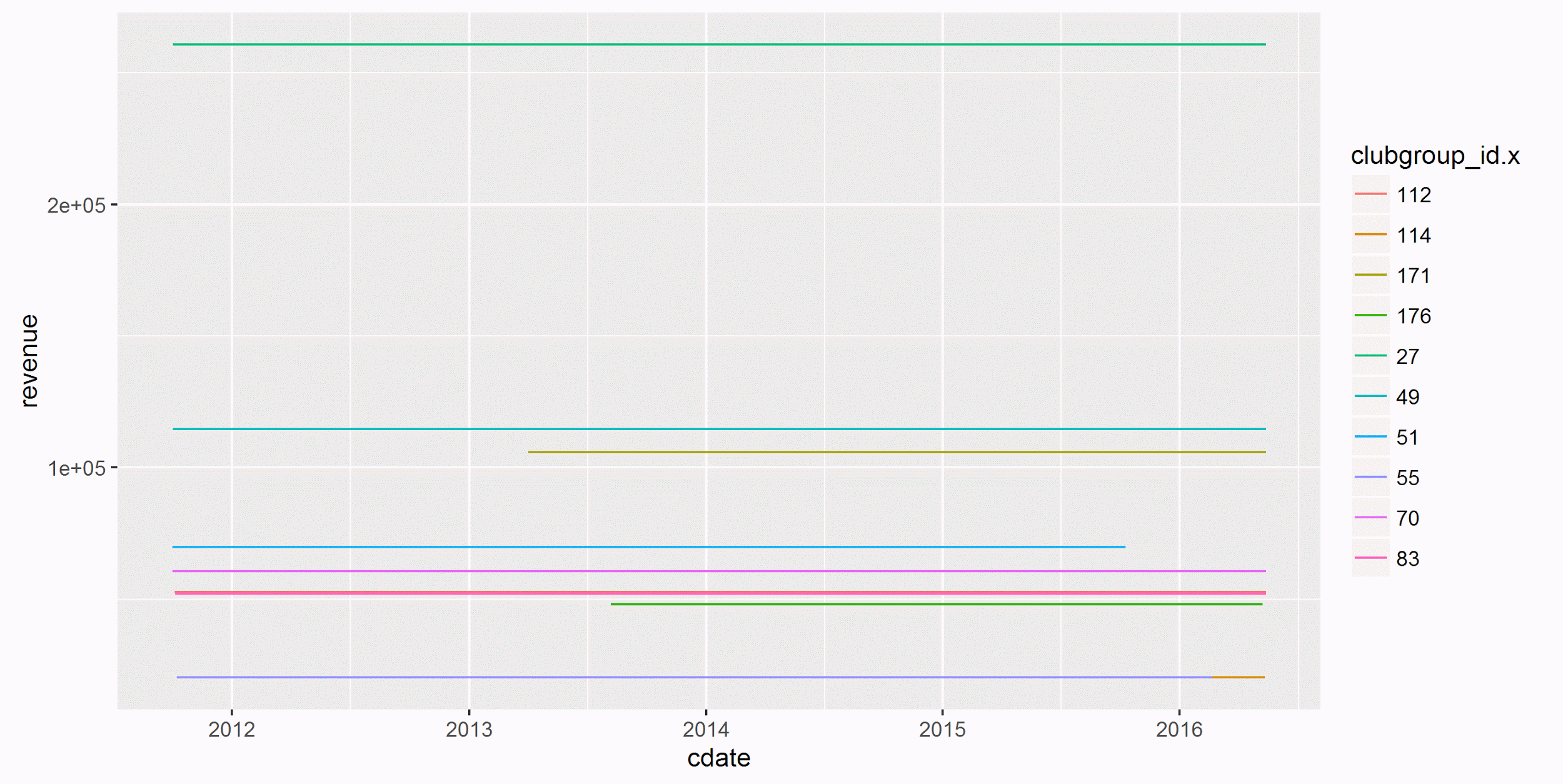
***Future considerations***

Future analysis and modeling that I believe would be useful and that I plan to perform are:

1. Trend plots showing top revenue generating clubs with revenue over time
   1. Plot with cumulative sales totals vs. time for all clubs, with each club using a different line on the plot (i.e. looking to combine the following plots into one).



* 1. This is an example of one showing the final revenue number, indicating which time period that club had sales:



* 1. Plot for individual clubs with a separate line for each year

1. Analysis of clubs over time, particular focused on comparing sales during “holiday” periods compared with “non-holiday” periods. “Holiday” periods would be defined as those in the 2 weeks preceding Valentine’s Day, Mother’s Day, Father’s Day, and the period from US Thanksgiving through New Years. The goal for this analysis is to look at what, if anything, sells better or worse during these periods (e.g. clubs with fewer or more bottles, higher or lower prices), and to look at any historical information -- such as whether being featured, for example -- for any effects on sales during these periods.
2. Additional predictive modelling to see if it’s possible to predict which clubs are trending higher and may sell better during the next holiday periods.

***Conclusion***

Though there are more ways to slice and dice this data, depending on the story to be told, the analysis in this report does clearly indicate the following things:

1. There were more than 27000 sales over a nearly 4.5 year period, accounting for a total revenue of over $1MM (commissions, not sales).
2. The top n revenue generating clubs account for the following percentages of total revenue:
   1. n=5 -> over 60%
   2. n=10 -> over 80%
   3. n=20 -> over 90%
3. Clubs that charge for shipping are weakly correlated to revenue (cor=.11), which was counter to the hypothesis that clubs with free shipping sell better than clubs that charge for shipping. Another way to say this is that the clubs that charge for shipping have revenue numbers exceeding clubs that don’t.
4. Even within the top 6 revenue generating clubs, 2 of these earned the majority of their revenue early and have tapered off in recent months/years. For example, the club 51 has had only $203 in revenue in the past 2 years (and none after 10 Oct 2015), while club 112, which has lower total revenue, has earned $35451 in the same time period. This is probably already obvious to the client, but promoting club 112 would seem to provide much greater benefit than club 51.

Based on the above report, I would make the following recommendations to the client:

1. Try to understand why clubs 49 and 112 have had significant upticks in revenue since the beginning of 2015, and see if there is any way to influence the same factors for other clubs.
2. Focus marketing efforts on clubs that offer red wine in their offerings, as they outsell white only or even champagne or rosé clubs by 48:1.
3. Focus efforts on clubs that offer smaller shipment sizes, as these outsell even the next most common shipment size (albeit those are shipments with 12 bottles) by 2:1. Or more likely, recommend to the wine clubs themselves that 2 bottle and 12 bottle shipments are the most purchased, outselling all others by 5:1.
4. Focus efforts on clubs offering a price per bottle of around $25, as there is a distinct grouping there.
5. Don’t worry too much about shipping costs impacting sales, though shipping costs should stay around $10-15 for clubs in the $38-72 range. Above that point the data doesn’t indicate what the preferred shipping price limit might be.