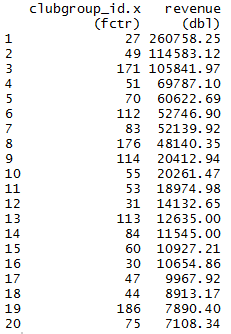
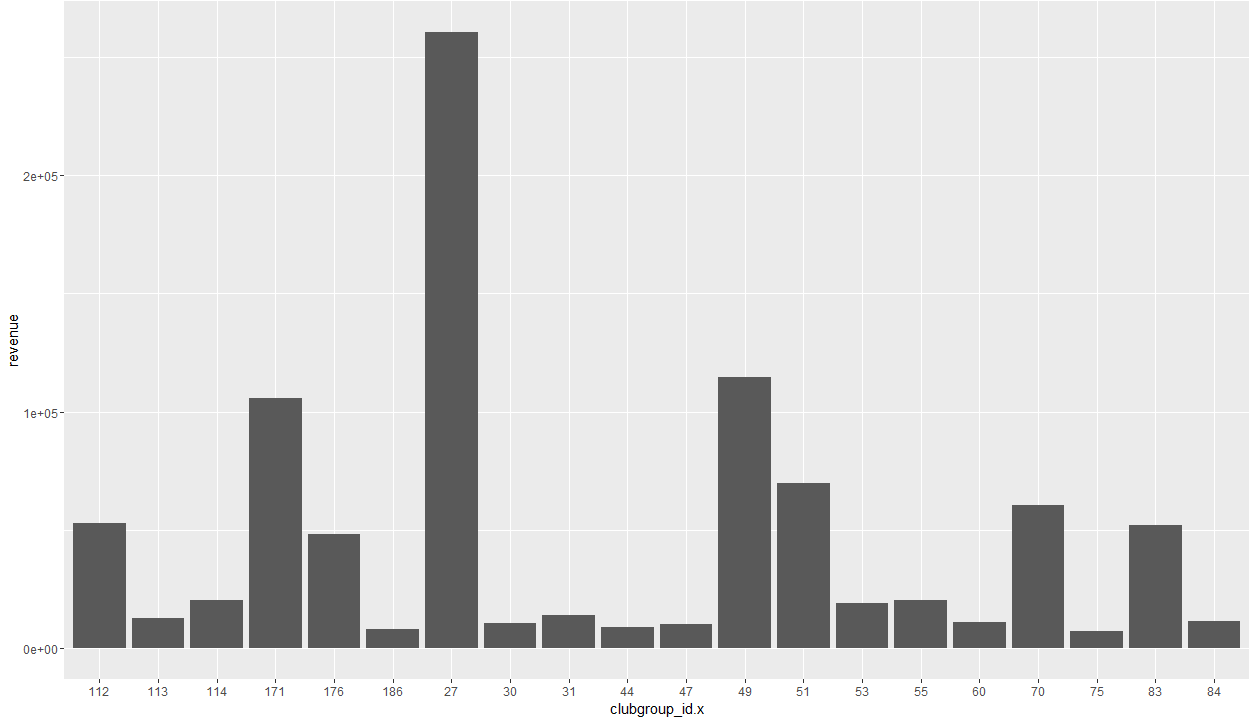
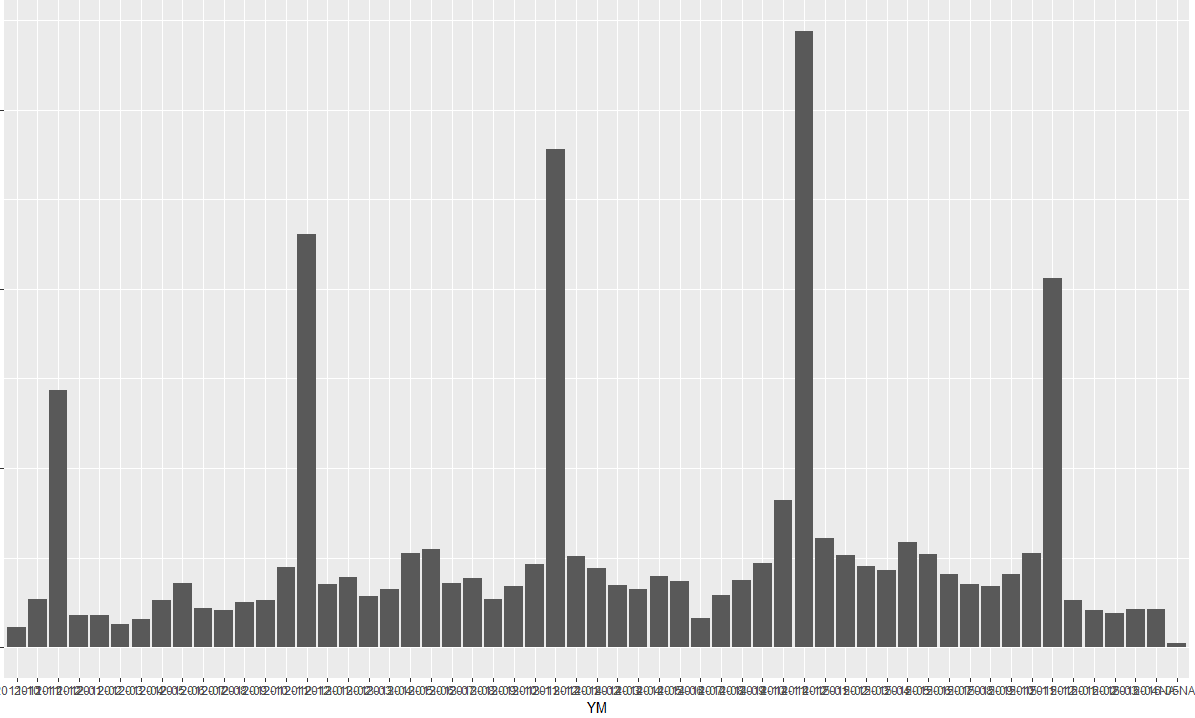
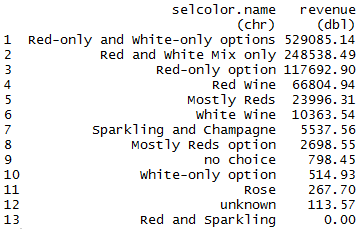
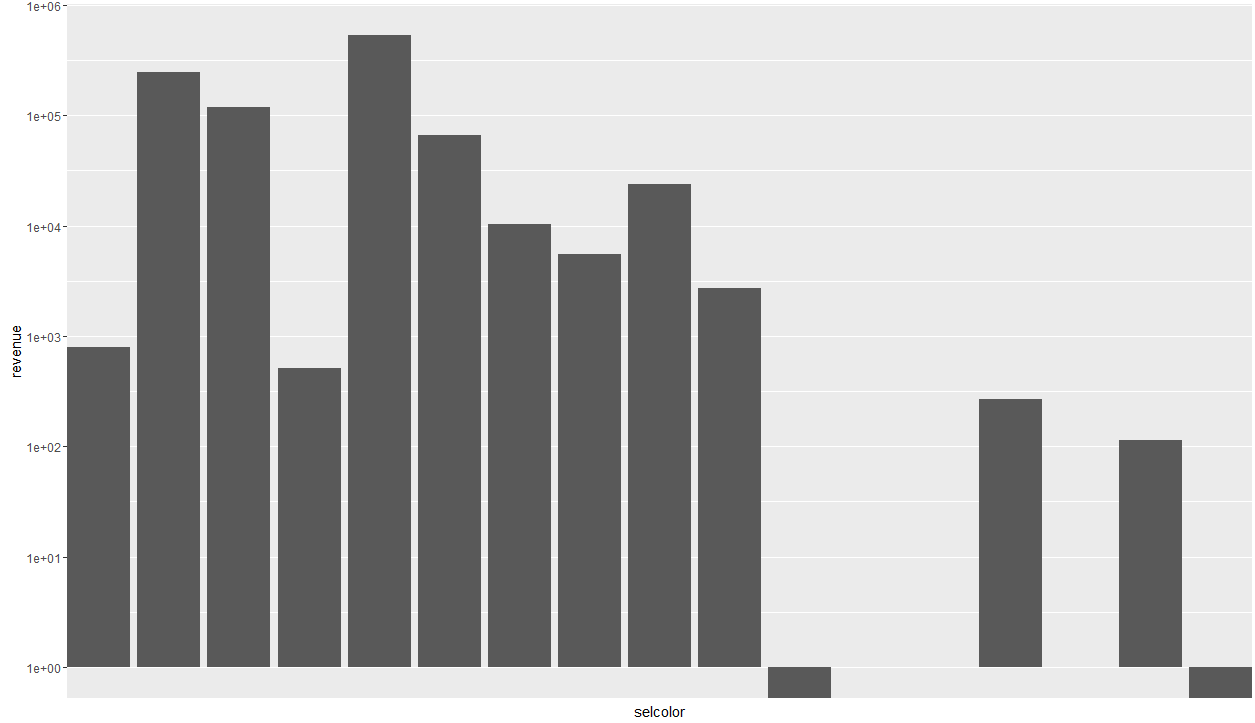
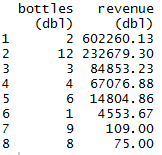
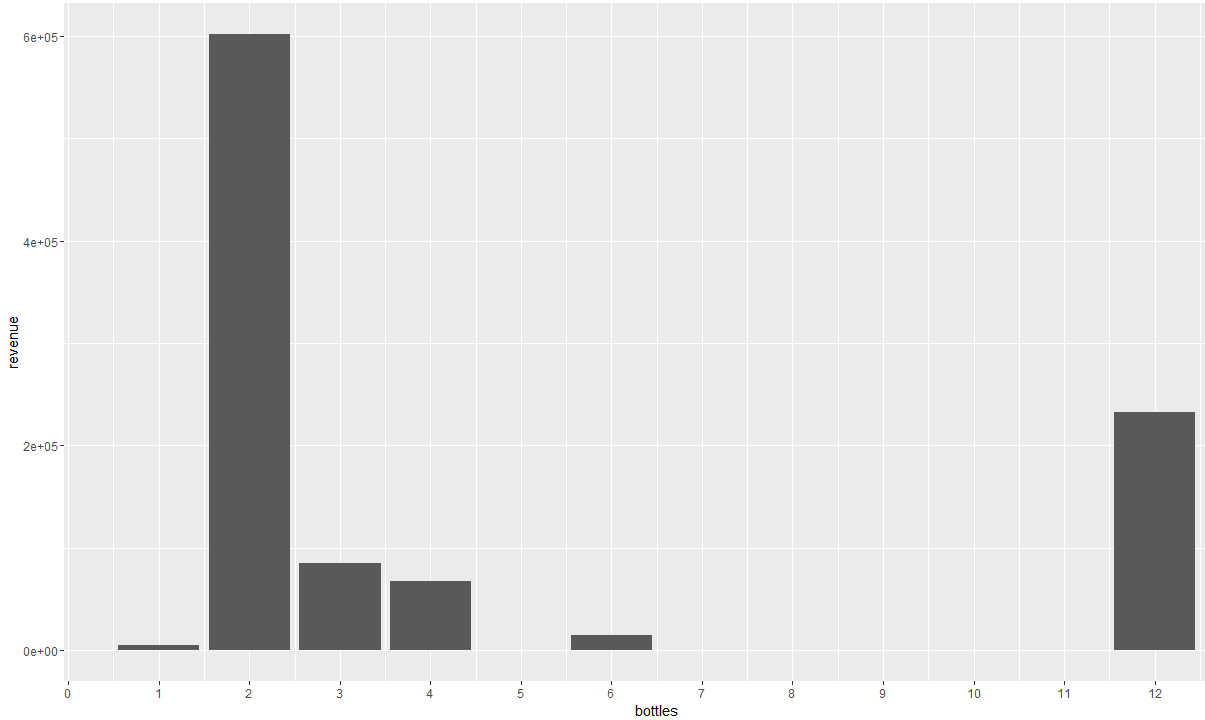
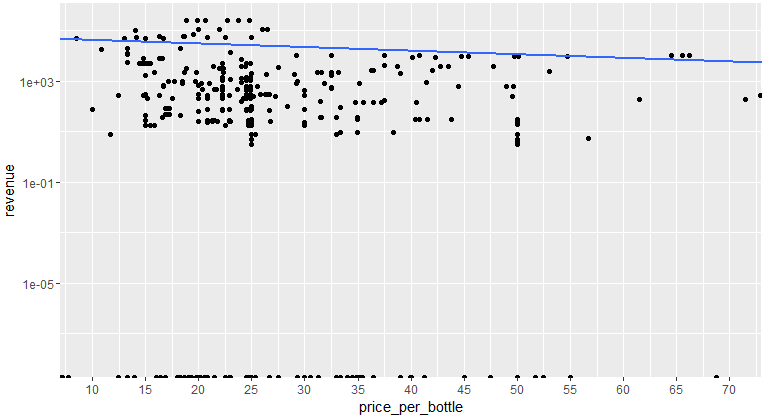
* An introduction to the problem (based on your earlier Capstone submissions).
  + I have chosen to analyze data from an affiliate marketing business, wineclubreviews.net. There are a couple of reasons I chose this company to analyze, mainly because of the volume of data available to me, and because it’s my sister’s company. I know that she has not had the resources to fully analyze or visualize the data, so for me to perform this analysis for her serves to benefit both of us.
  + The problem I’m looking to solve is 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.
* What important fields and information does the data set have?
  + The most important fields and information that the data set provides are:
    - Click data for over 635,000 click observations of 77 variables 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, Valentines’ 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.
* What are its limitations i.e. what are some questions that you cannot answer with this data set?
  + I haven’t yet found any questions that couldn’t be answered with the data that I have, barring time and skill level to perform certain analyses.
  + In addition, as of this milestone report, I have not yet performed any predictive analysis, so it’s unclear whether I will obtain any useful data from efforts to predict future sales.
* What kind of cleaning and wrangling did you need to do?
  + 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).
  + In addition, 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.
* Any preliminary exploration you’ve performed and your initial findings.
  + Top 20 club groups by revenue
    - 
    - 
  + Total revenue by month
    - 
  + 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
* Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?
  + For the most part, the approach is exactly the same. EDA followed by some predictive modeling, though some of the exploratory data analysis has not yielded as useful of information as I had expected. That being said, there are many other paths of exploration I can take, only limited by time and skill. It also remains to be seen if the predictive analysis will lead anywhere.