**Project 1** – Modeling Beer Advocate Reviews

Andrew Leppla, Huy Hoang Nguyen, Ikenna Nwaogu

**Introduction**

In this project, we will study the Beer Advocate dataset [1]. ..

<https://www.ahoulette.com/2019/03/11/beer-advocate-reviews/> (for some ideas in EDA and introduction)

Beer review data from Beer Advocate were used to address two project objectives:

1. Build predictive regression models using cross validation with metrics to compare multiple models. Provide interpretation of the regression model(s), including: hypothesis testing, interpretation of regression coefficients, and confidence intervals as well as practical vs. statistical significance.
2. Perform a secondary analysis using Time Series and address if the assumption of independent errors is valid for the final regression model.

**Data Description**

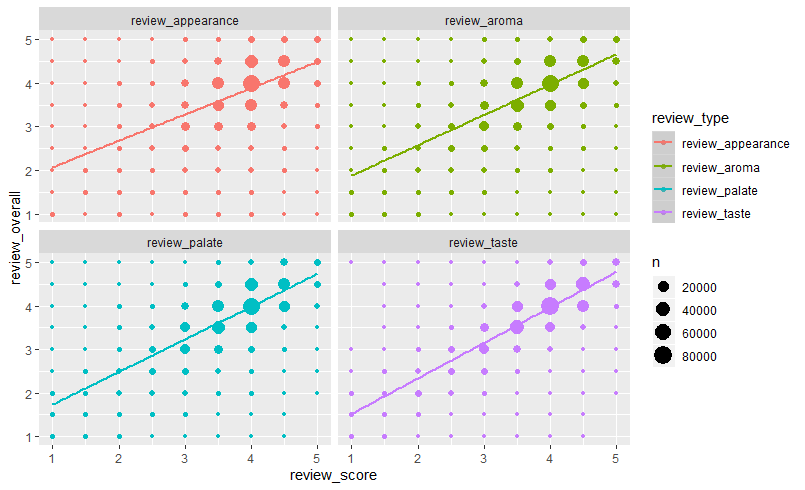
The data set from the online forum BeerAdvocate.com has over 1.5 million individual beer reviews that cover 66,055 unique beers from 5,840 breweries. Most of these reviews focus on craft beers and are not representative of mass-market beers like Budweiser, Miller, Coors, etc. The reviews span about 15 years from 1996 to the end of 2011. In addition to brewery and beer names, reviews include 5 different ratings: overall, taste, appearance, aroma, and palate. This project focused on Overall Rating as the primary response. Additional data are provided on beer style, alcohol by volume (ABV), and review time.

**Exploratory Analysis**

Taking a first look at the data, the beer ratings are discrete with values of 1 to 5 in increments of 0.5 (see Figure 1), 1 being the worst rating and 5 being the best. Ordinal logistic regression or a discrete choice model would be most appropriate here, but they’re not within the scope of this project. With over 1.5 million reviews, the ratings can be averaged across different variables to normalize them per the Central Limit Theorem (Figure 2). These will be called Average Reviews to distinguish them from the raw Individual Reviews.

One. Short phrase on missing value.

Using Overall Rating as the response, the other 4 ratings are correlated with Overall Rating and with each other, indicating a potential issue with multicollinearity (Table 1). This issue is exacerbated when using Average Reviews by beer name (Table 2, Figure 2).



**Figure 1** – Scatterplots of Individual Overall Rating vs. other individual ratings in year 2011. Ratings data are discrete.

**Table 1** – Correlation matrix of individual review ratings data for year 2011

review\_overall review\_aroma review\_appearance review\_palate review\_taste

review\_overall 1.0000000

review\_aroma 0.6872329 1.0000000

review\_appearance 0.5236376 0.4922312 1.0000000

review\_palate 0.7221078 0.5617439 0.5165652 1.0000000

review\_taste 0.8483747 0.6856092 0.4839489 0.6948401 1.0000000

ABV

**Table 2** – Correlation matrix of average review ratings by beer name for year 2011. Average reviews are more correlated.

Overall Aroma Visual Palate Taste

Overall 1.0000000

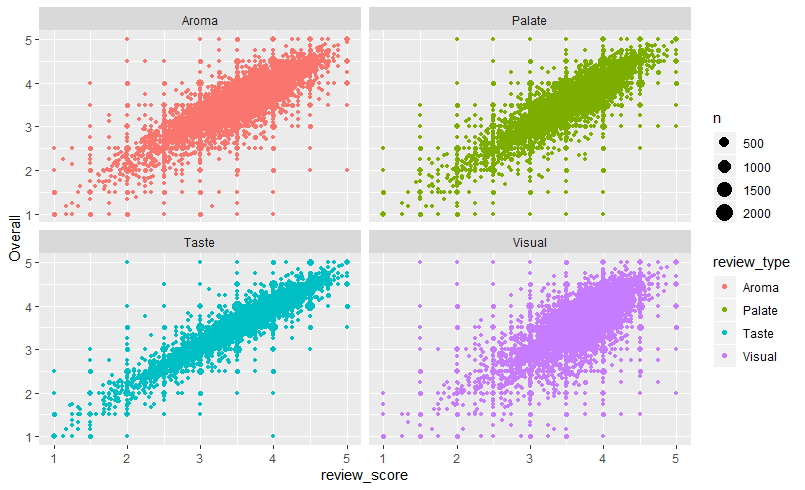
Aroma 0.7800110 1.0000000

Visual 0.6272636 0.5996442 1.0000000

Palate 0.8121640 0.6842859 0.6282888 1.0000000

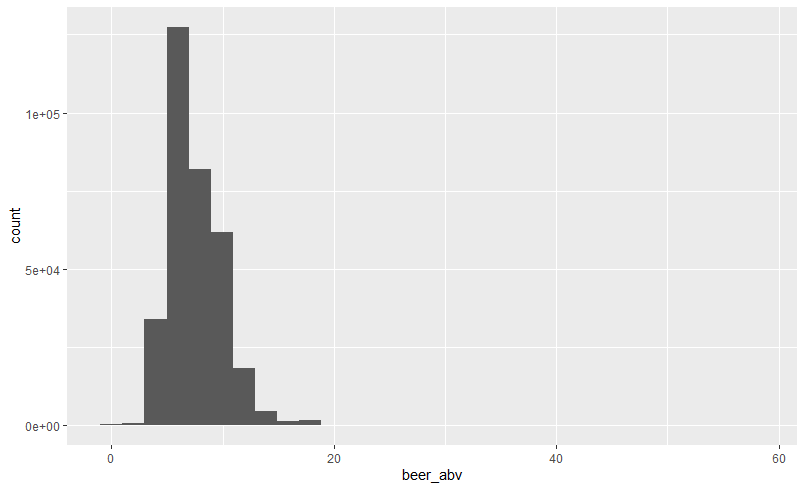
Taste 0.8984971 0.7882234 0.6015011 0.7896848 1.0000000

ABV

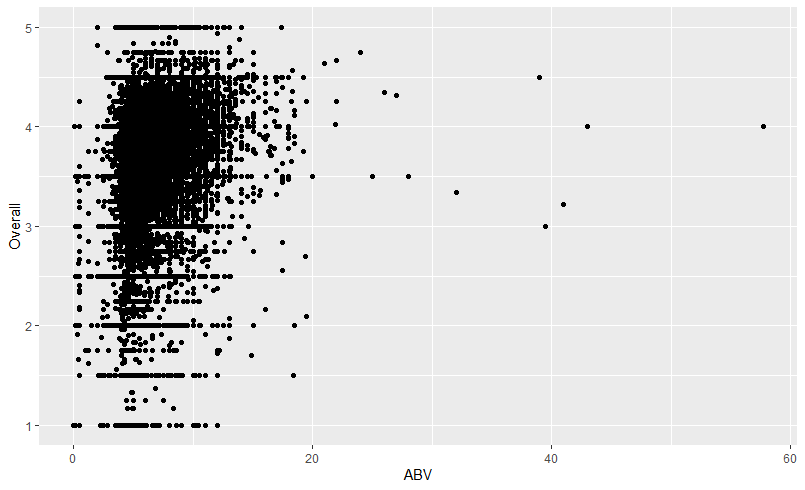


**Figure 2** – Scatterplots of Overall Rating vs. other ratings averaged by beer name in year 2011. Average Ratings data are more normalized and more correlated.

ABV data are also not normal and have a long right-tailed distribution (Figure 3). It is also non-linear when plotted vs. Overall Reviews with a weaker correlation (Figure 4). Non-linearity will be addressed by filtering out the very rare beers & styles with extreme ABV values.



**Figure 3** – Right-skewed ABV data.



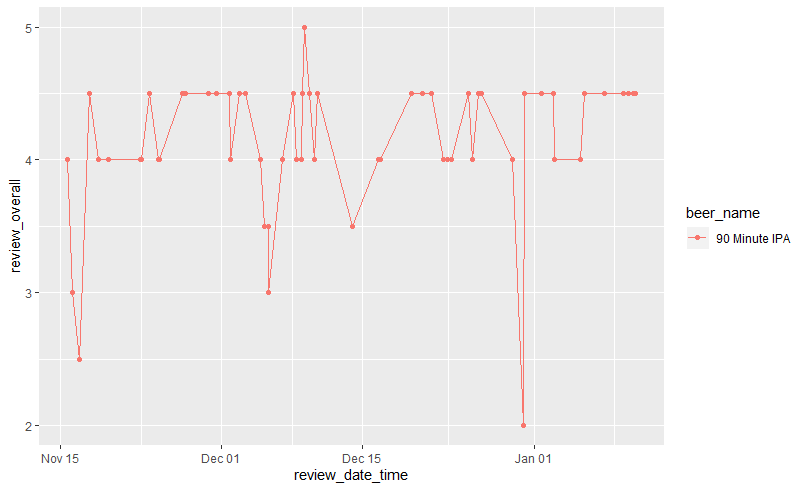
**Figure 4** – Overall Ratings vs. ABV averaged by beer name in year 2011. Non-linear and weakly correlated.

The most reviewed beer styles, breweries, and beers were identified to simplify the data for analysis. The most reviewed beer styles were American IPA, American Double/Imperial IPA, and American Pale Ale. The most reviewed breweries that made these styles were Dogfish Head, Sierra Nevada, and Stone? The most reviewed individual beer was Dogfish Head’s “90 Minute IPA”.

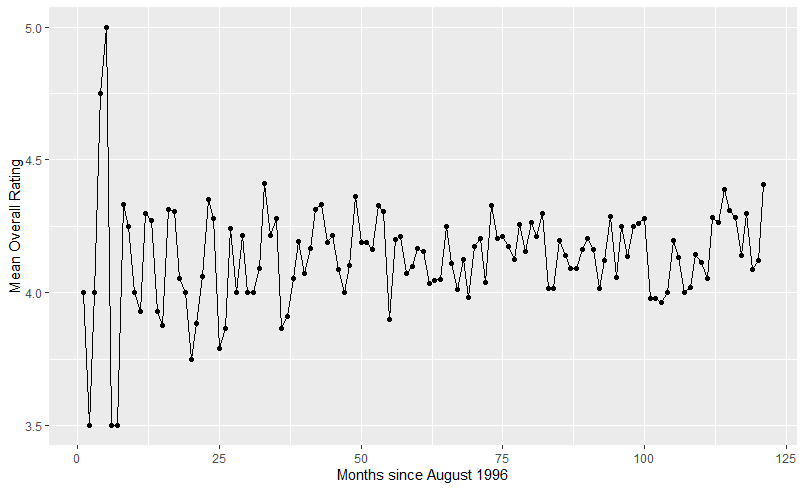
<Insert top 3 tables with counts>

The Overall reviews of Dogfish Head’s “90 Minute IPA” were plotted vs. time to explore the time series data. Per Figure 5, the review data are not spaced equally over time (a key assumption of time series analysis). This is expected given that beers aren’t being reviewed every second of every day in the primarily U.S.-based market.

Unequal data spacing over time was addressed by averaging the reviews by Month as well as by Week to ensure at least one review for any given time spacing. This also normalized the Overall Rating data. See Figure 6. Unequal spacing could also be addressed by assigning each review an index value rather than a date and time value.



**Figure 5** – Individual Overall Reviews for beer “90 Minute IPA” in Winter 2011. Data are discrete and not equally spaced across time.



**Figure 6** – Average Overall Reviews for beer “90 Minute IPA” by Month. Data are equally spaced across time and more normalized.

**Objective 1**

**I will put sth until tomorrow for introduction.**

Problem Statement - Build predictive regression models using cross validation with metrics to compare multiple models. Provide interpretation of the regression model(s), including: hypothesis testing, interpretation of regression coefficients, and confidence intervals as well as practical vs. statistical significance.

Model Selection

There were multiple methods of model and variable selection used. We used forward, stepwise, backward and a mixture of both forward and backward selection type. We used the AIC method to check importance of variable based on the variables that were selected as well as the interactions between the variables. We first ran selection for the main 5 predictors and plot the importance. Then we used the combinations of top 3 variables as our interaction. We limited it to only combinations of 2 of the 3 variables. The 3 top variables are Taste, Palate and ABV content.

In our analysis we chose to look at the average data of beer name per brewery. We aggregated and counted number of reviews each beer got. We added count column as a predictor to see if it would help with a better model.

Type of Selection: LASSO, RIDGE, ELASTIC NET,

Stepwise, Forward, Backward,

Manual / Intuition,

A mix of all of the above.

Checking Assumptions **Required**

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Compare Competing Models **Optional (Helpful if using 2 model strategy)**

Via: Training and test set split or CV

**Possible Metrics**: (ASE, AIC, BIC, adj R2, etc)

Parameter Interpretation

Interpretation ***Required***

Confidence Intervals ***Required***

Final Conclusions

*From the analyses of Objective 1* ***Required***

*In addition to overall conclusions, feel free to include additional insights or concerns gleaned from the analysis. What needs to be done next or how could we do it better next time?*

**Objective 2**

**Some ideas of introduction maybe pictures I sent to Andrew.**

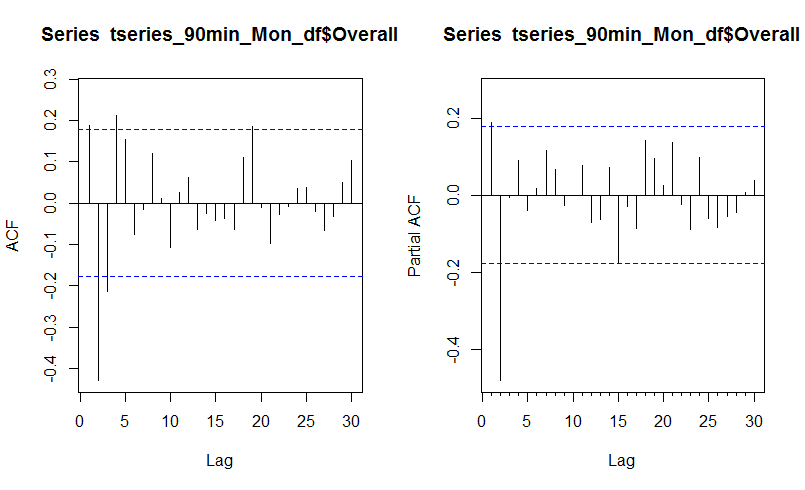
Goal Summary - Perform a secondary analysis using **Time Series** and address if the assumption of independent errors is valid for the final regression model. Determine if there are any meaningful time-based trends in the beer review data.

Main Analysis Content

To reiterate from the Exploratory Analysis section, unequal data spacing over time was addressed by averaging the beer reviews by Month as well as by Week. Unequal spacing could also be addressed by assigning each review an index value rather than a date and time value. Both approaches result in some information loss but enable time series modeling.

Average monthly review data for Dogfish Head 90 Minute IPA look fairly stationary, but there is a possible upward trend for the last 20-24 months (Figure 6). Also, the first seven months are highly variable due to a low sample size from the early days of the review platform with few users. Both of these issues will be investigated through this analysis.

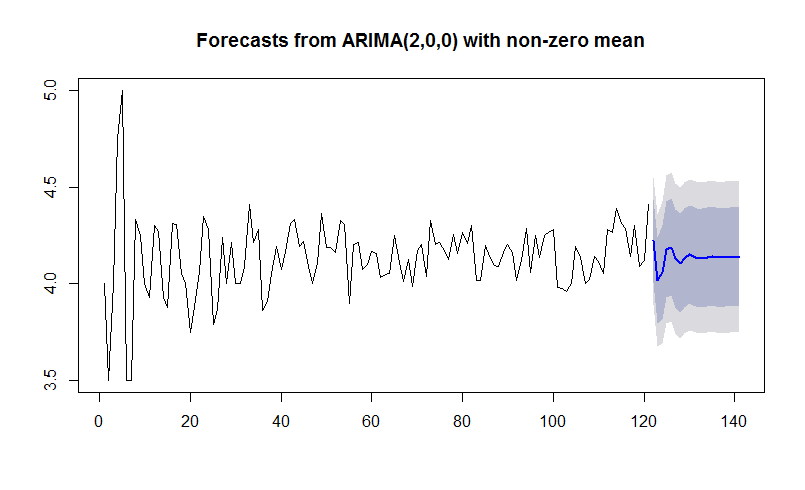
First, the ACF and PACF plots were examined to assess any potential time series models:



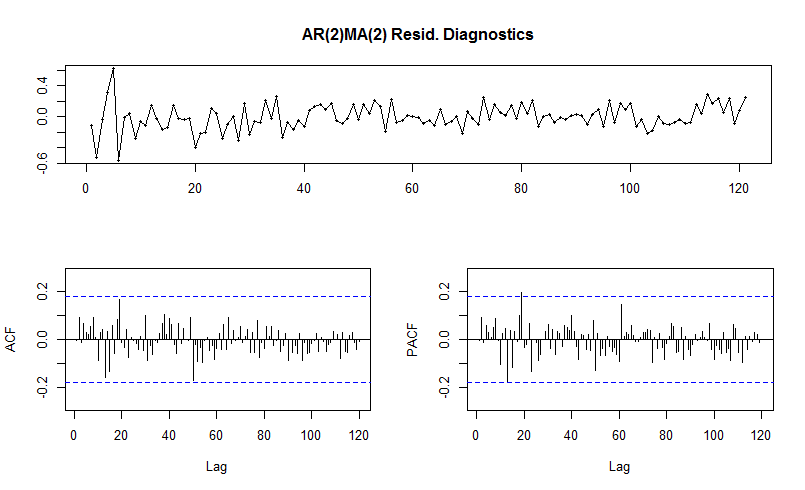
**Figure X1** – ACF and PACF plots for average monthly reviews of Dogfish Head’s 90 Minute IPA.

The plots indicate a possible ARMA model with both the ACF and PACF plots slowly decaying over time. There may be a significant AR(4) model per the ACF plot and a MA(2) per the PACF plot. There may also be some higher order terms or non-stationary behavior for lag 19 (ACF).

The auto.arima function in R was used for initial time series model screening. This game a simple, low bias AR(2) model. This model wasn’t the best fit and missed significant residual behavior at lag 19 per the PACF residual plot. Further, the residuals don’t decay until after lag 50-60 indicating some high order terms or non-stationary behavior.

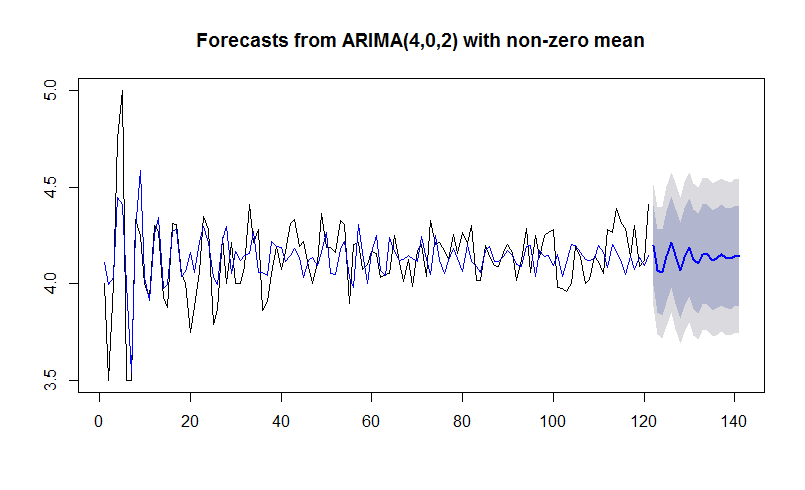


**Figure X2** – AR(2) time series fit and forecast for 90 Minute IPA Overall Rating by Month.

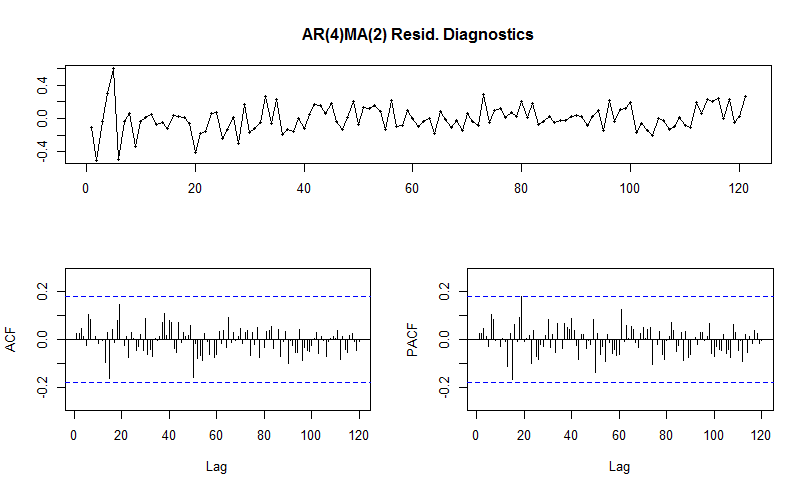


**Figure X3** – AR(2) time series residuals for “90 Minute IPA” Overall Rating by Month. Significant lag 19 residuals with decay after lags 50-60.

To try to select a higher order model, auto.arima was run without stepwise selection and with a max order >10. This produced an AR(4)MA(2) model that was consistent with the initial ACF & PACF plots. The lag 19 residual is just below the significance threshold, but the residuals still don’t decay until after lags 50-60:



**Figure X4** – AR(4)MA(2) time series fit and forecast for 90 Minute IPA Overall Rating by Month.



**Figure X5** – AR(4)MA(2) residuals for “90 Minute IPA” Overall Rating by Month. Significant lag 19 residuals with decay after lags 50-60.

Conclusion/Discussion ***Required***

*The conclusion should reprise the questions and conclusions of objective 2.*

References:

[1] Beer Advocate dataset [https://www.beeradvocate.com](https://www.beeradvocate.com/)

[2] G. James, D. Witten, T. Hastie, R. Tibshirani*, An Introduction to Statistical Learning with Application in R*, Springer, 2017.

[3] R. H. Shumway, D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, 4th Edition, Springer, 2017.

[4] Stats business – I will put the name when at home

[5] another book forget name

ANDREW LEPPLA – Southern Methodist University – *Email:* [aleppla@smu.edu](mailto:aleppla@smu.edu)

HUY HOANG NGUYEN - Southern Methodist University - *Email address:* hoangnguyen@smu.edu

IKENNA NWAOGU - Southern Methodist University - *Email address:* inwaogu@smu.edu

**Appendix**

*Well commented SAS/R Code* ***Required***

*Graphics and summary tables (Can be placed in the appendix or in the written report itself.)*