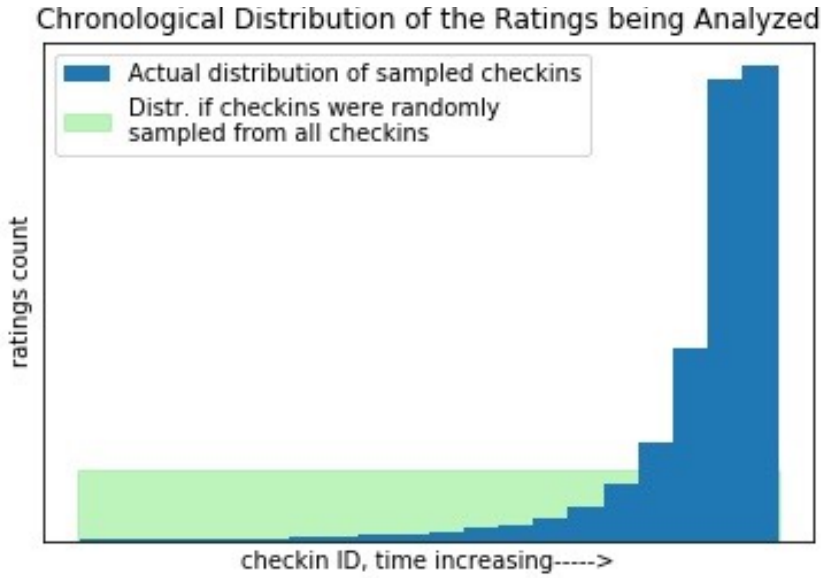
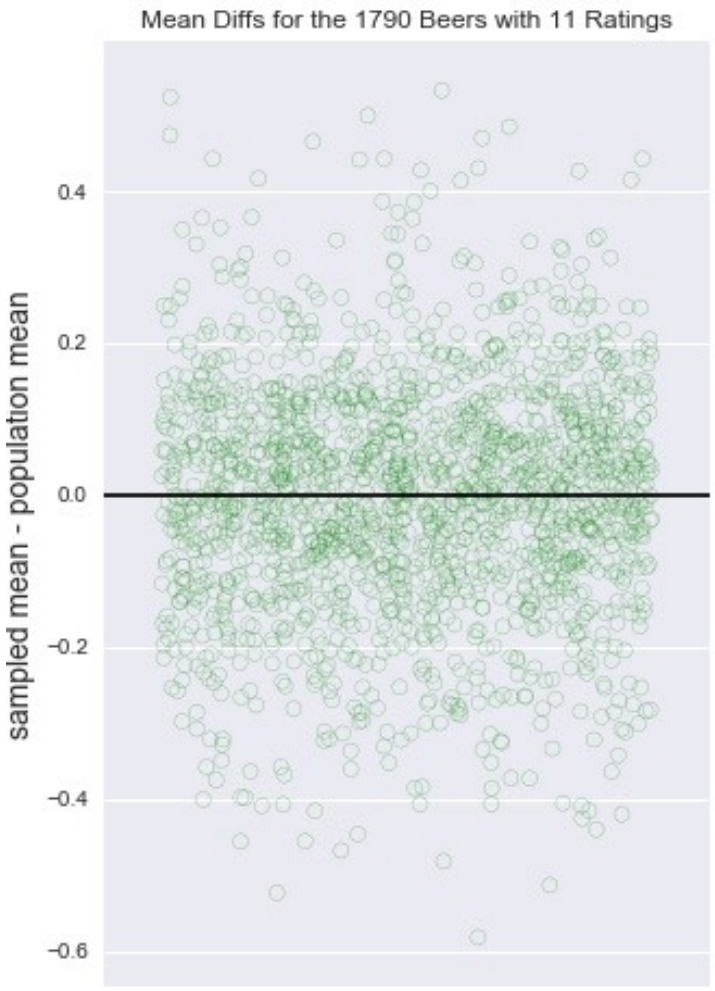
One potentially important question about the ratings analyzed in this project is “How representative of the entire dataset is this subset of ratings?” Starting with the question of the checkin dates, we could either place them in equidistant date range bins and compare them to how many overall ratings occurred in those same date bins, or more simply we could just place the checkin ID’s in bins, use those (chronologically assigned, conveniently) ID’s as a proxy for dates, and just consider the distribution of the histogram area as a relative proportion of all checkins during the timeframe that the ID’s increased in each bin. Here’s how that looks:



Since the data samples are so heavily weighted toward the present, there might be concerns that conclusions drawn analyzing them will too heavily reflect the rater’s most recent tendencies. But that actually turns out to be a benefit to this project, which seeks to predict the most recent ratings of users, in an effort to simulate the intended situation, where a machine recommends a purchase for a user, based on how the user previously rated items. No matter how a user’s tendencies and tastes may change over time, the most accurate conclusions about current preferences will be based on most recent preferences. Furthermore, many beers come and go fairly quickly in today’s market, and brewing techniques are evolving quickly enough that particular beers or styles of beer from which a user is choosing may only have been rated by this user or other users during the previous few months.

Another concern is that long-term statistics used will be misleading during training that uses short-term statistics. The one long-term statistic that this project will lean on heavily when making its predictions is each beer’s all-time mean global rating for all users who rated it. Rather than training a model to predict the given user’s rating, it works better to start from the historical mean rating for the beer and then target the given user’s preferences/biases. But what if the beer got great ratings 5 years ago and is now considered an average beer? We

don't want that old global average to wrongly inflate our recommendation today. One solution might be to substitute the mean of all ratings for that beer just in our dataset, taking advantage of their relatively current nature. There are 1,366,604 ratings here for 112,970 different beers, so on average we'd have about 12.1 ratings per beer to build a "modern mean," minus the one that we're actually predicting, which we shouldn't include in the training mean. Here's how different our sample means are from their global means at 11 ratings:



In fact, the average for our 112,970 means for all sample sizes is 3.80, same as the average of their global means. The chart shows that while the differences are fairly equally scattered above and below zero, using the mean of the samples instead of the population mean will introduce some prediction bias, which if representative of an underlying trend might improve recommendations but which in fact makes them worse.

Correlation Matrix	Global Rating	Checkin Rating	Checkin ID
ABV	0.5234	0.2516	-0.0014
Checkin ID	0.0707	0.0070	
Checkin Rating	0.4645		

When making recommendations, the highest correlation with Checkin Rating is Global Rating (0.4645), followed by Alcohol Content, or ABV (0.2516).

When predicting global ratings for unrated-yet beers, whether a User needs a guess or a Brewer is making decisions, ABV has the highest correlation (0.5234).

The (Pearson) correlations shown above suggest that global ratings are getting slightly higher as checkin ID's increase (time increases), but at 0.0707, the change isn't large. The individual ratings are even less correlated with time (0.0070), partly due to their much higher variance than the global mean to which they contribute. In our attempts to minimize the root mean squared error (RMSE) of our recommendations, we will be better served using the historic mean, with its lower variance, as our starting point.

An individual rating can be viewed as a combination of the quality of the beer, approximated here by the global mean, and a user bias/preference, which we can attempt to learn. The user bias can in turn be viewed as comprising an overall generosity or stinginess in ratings, plus an individual taste/style preference depending on the specific beer. The first part of that user bias can be predicted fairly well just by comparing how a user rates beers vs. how the rest of the world rates those same beers. Certain users just have a different mental scale for ratings, where they may think of an average beer as being worth 3.0, where others might consider average to be 4.0. But the more variability the user has in those deviations from the global mean, the higher the RMSE for that user's recommendations will be. The mean of the standard deviations for user ratings deviations for users with 5+ checkins is 0.3629, in the bottom row of this table:

Per User (5+ rates)	Standard Deviation
Global Ratings	0.2175
User Ratings	0.4161
User Deviation	0.3629

(These St.Devs are calculated per user, and the mean of all users is shown.)

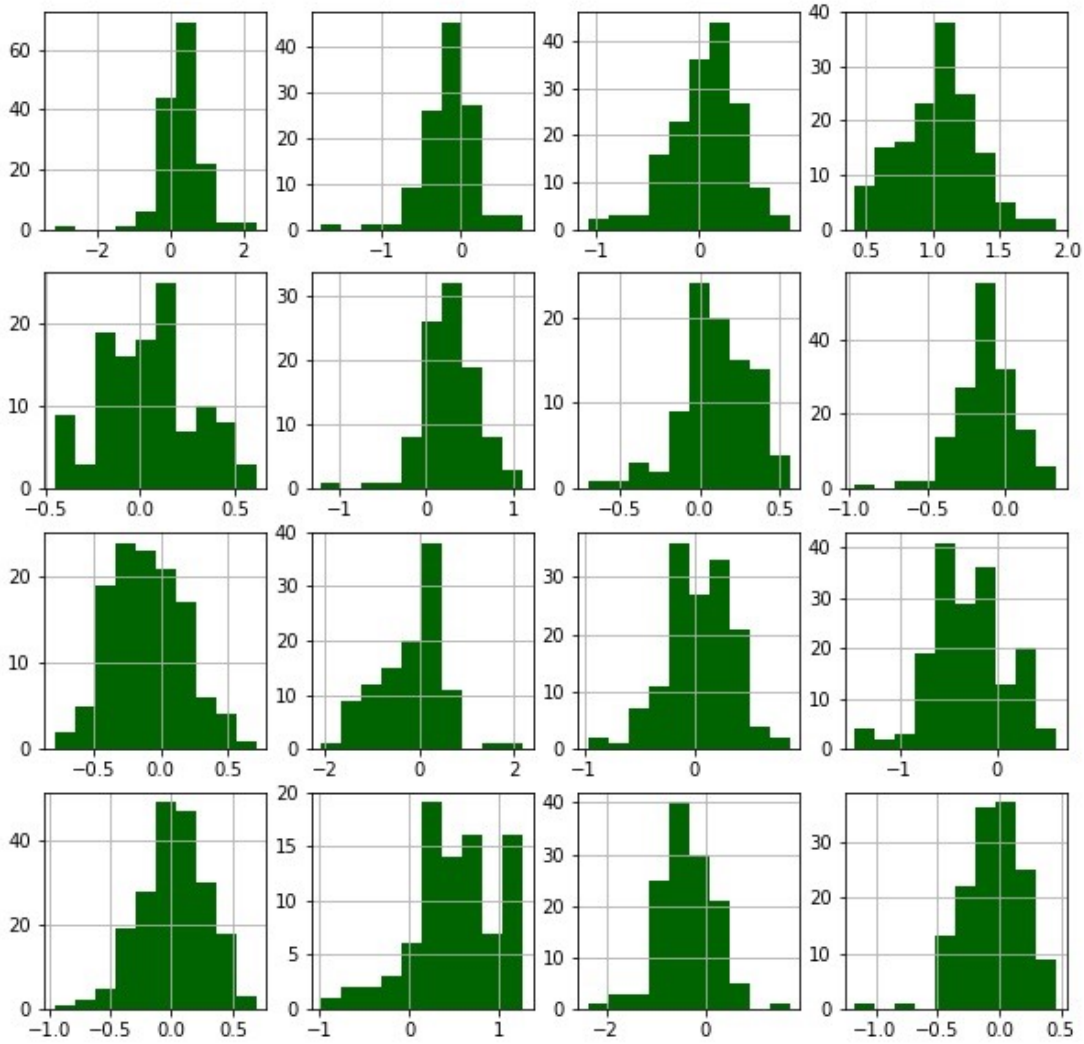
This 0.2175 std for global ratings per user is lower than the overall std for all beers, which is 0.2813. So each user does manage to select and rate beers of a similar quality level.

If we simply used the user's mean rating as our prediction, and the ratings were normally distributed within users, and all users had tried thousands of beers, the RMSE for all predictions would be close to this.

This is the "std of user deviations", showing how much users tend to deviate from the mean rating.

If the last row's stat was normally distributed (it's often close, see plots below, but the sample sizes are sometimes small), and was independent from the top row, its square plus the top row's square would equal the middle row's square.

$\sqrt{(0.3629)^2 + (0.2175)^2} = 0.4230$, which is close the the middle row. A random sample of 16 users with 5+ ratings is shown below, with the x-axes showing their ratings' deviations from the global mean for the beers, and the y-axes showing the rating counts:



Unfortunately, it's not as simple as just starting with the user's mean rating score and tweaking it towards features the user seems likely to prefer or avoid, such as ABV. If you take all of the user's ratings except for the last one and predict their mean, the RMSE for the whole dataset is

not that nice 0.4230 number, but rather 0.4893. This issue is that these predictions don't take into account what differentiates new, unrated beers from previously rated ones. To do so, we'll start with the global mean of the new beer, where available, and make predictions for the global mean where not available (for an unrated or barely-rated beer, for example). Then we'll add a user bias, which is simply how generous the user is in general, and finally we'll try to discern which way from that baseline estimate the user is likely to lean, based both on other users with similar or opposite patterns from our user and on features of the beer that a model can learn to weigh in favor of a higher or lower tweak for this user/beer combination (such as if the user likes or dislikes beers by the same brewer, or the the ABV is low and this user likes or dislikes low-ABV beers).