IT 403
Project – Beer Advocate Analysis

1. Exploratory Data Analysis (EDA)

Beer Advocate is a membership-based reviews website where members rank different beers based on a wide number of categories. The data covers overall ratings of beer by different users, more granular ratings (like aroma, color, etc), beer style, beer name, time of review, ABV (alcohol by volume), and beer brewer. The data set used for this report covers over 260,000 records for the year of 2006. Data was transformed from unstructured to CSV format using Open Refine, as well as a combination of custom shell scripts that were created using various tools like sed to strip out unwanted data and artifacts, as well as convert Unix epoch time to standard years and months. It was then loaded into SPSS for visualization.

The main challenges with working with this dataset in SPSS arose from its size. At over 260,000 records, and checking in at over 100MB, SPSS ran quite slow when ingesting and analyzing the data. Furthermore, when running cross tabulation reports, issues arose with specific limitations (likely due to memory allocation) that SPSS has put in place. This necessitated a dramatic reduction in the size of the dataset (only for the cross tabulation portion of the analysis). Size reduction was done by randomly sampling records using a shell script written in the command line.

The main questions we ask in this report deal with the kinds of beer that people review, and what affects the review scores that they give to the beers that they drink. Since the craft beer movement was created largely as a response to the homogeneity of American beer offerings, it's of interest to see if it actually offers people a dramatic difference from the previous "pilsner beer" paradigm, where American Pilsners like Budweiser, Coors, and Miller were the most popular. The analysis then ends with a review of users on the site, and how they might affect the data in the website's database.

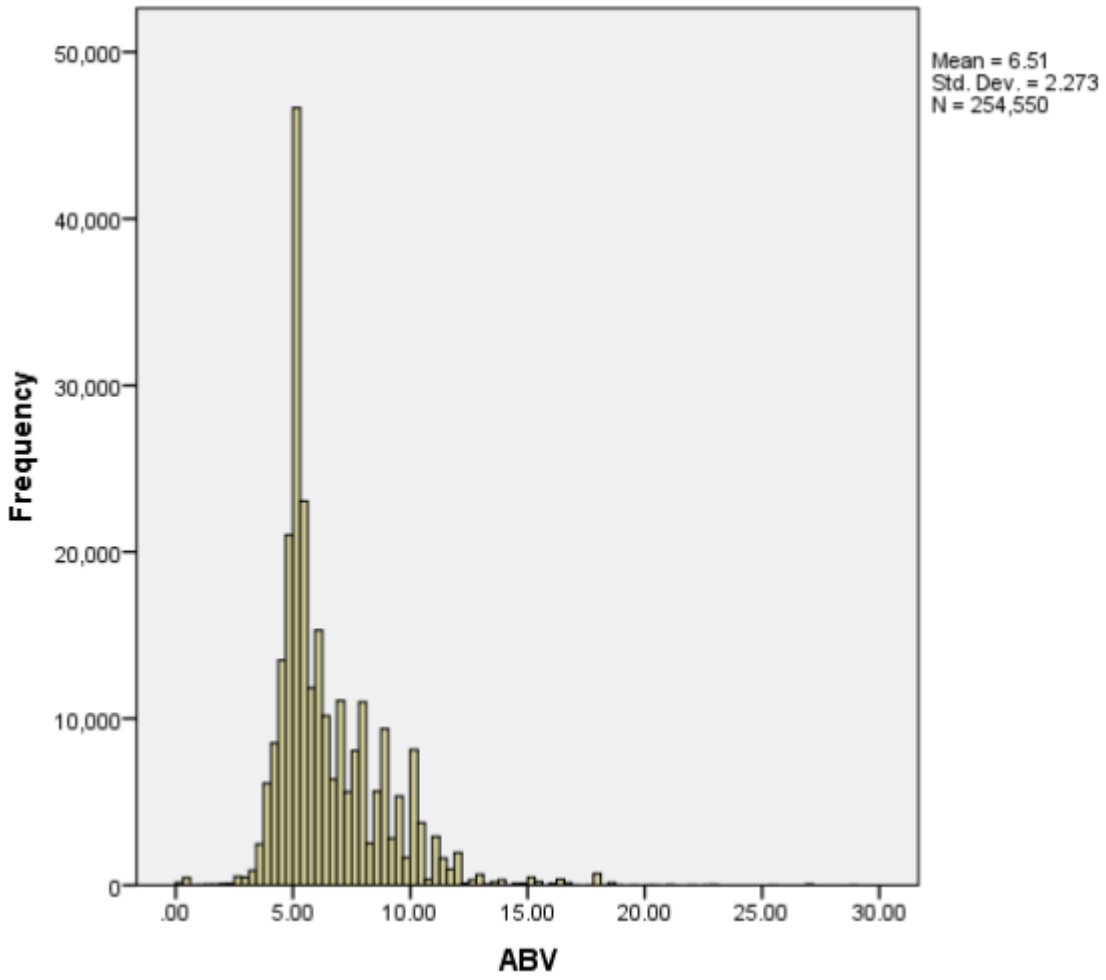
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ReviewAppearance	266508	1	5	3.40	.828
ReviewAroma	266508	1	10	6.24	1.698
ReviewPalate	266508	1	5	3.21	.848
ReviewTaste	266508	1	10	6.34	1.674
ReviewOverall	266508	1	20	13.01	3.443
ABV	254550	.01	29.00	6.5095	2.27291
Month	266508	JANUARY	DECEMBER	JUNE	00:00:03.480
Valid N (listwise)	254550				

1.1. Quantitative Variables

The quantitative variables for this dataset are ABV (alcohol by volume), Review Appearance, Review Aroma, Review Palate, Review Taste, Review Overall, Time, Month.

1.2. Histogram



The histogram of this data looks at the beers reviewed by their ABV, and indicates that it the data is right skewed. When analyzing the records of beers reviewed on a beer rating website, the ABV (alcohol by volume), on average, of beers reviewed is 5.75, with three quarters being at 7.7 or below. The skewness of the data, combined with the presence of outliers, indicates that using a mean average is going to result in, to some degree, data that is less reliable.

1.3. Center of Distribution

The median ABV of beers reviewed on the website is 5.75, and is a good measure of center. Due to the number of outliers in the data's ABV, there's too much risk in relying on the mean, which is 6.5.

1.4. Five-number-summary

Statistics

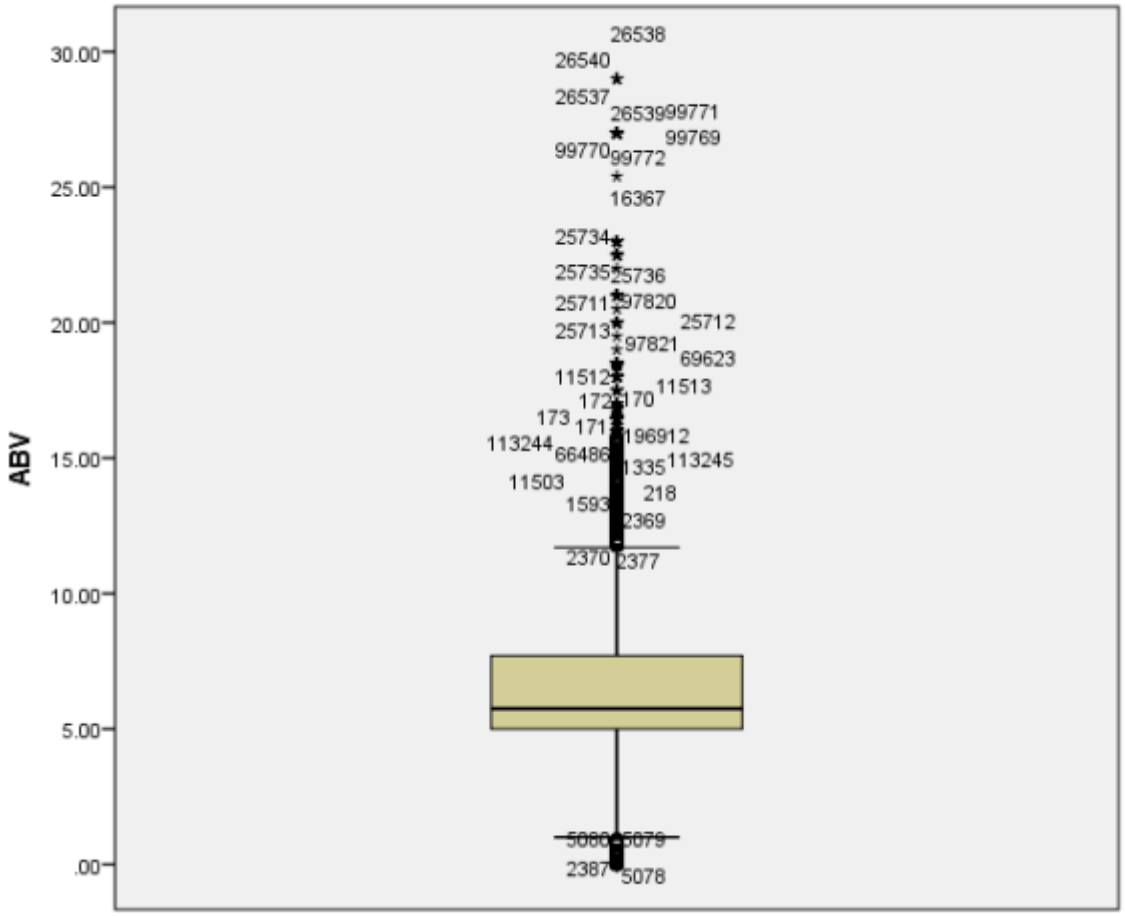
ABV

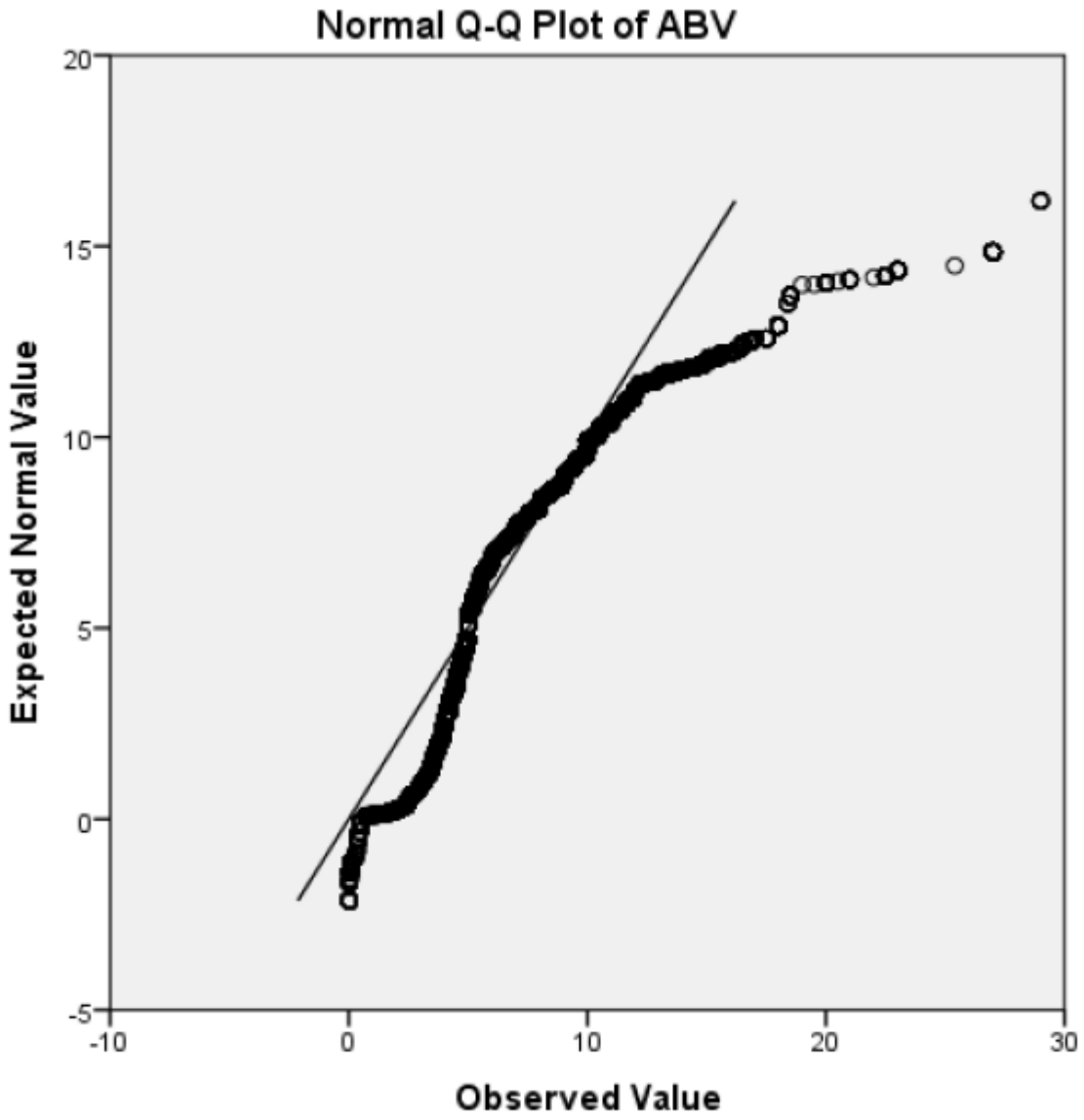
N	Valid	254550
	Missing	11958
Mean		6.5095
Median		5.7500
Std. Deviation		2.27291
Minimum		.01
Maximum		29.00
Percentiles	25	5.0000
	50	5.7500
	75	7.7000

The five-number-summary for ABV is: .01, 5, 5.75, 7.7, 29

The minimum ABV of a beer reviewed is .01, the maximum is 29. Under 25% of beers reviewed have an ABV of 5 or less. 50% of beers reviewed have an ABV of 5.75 or less, and the top 25% of beers reviewed have an ABV of 7.7 or more. The median average ABV of beers reviewed is 5.75.

1.5. Boxplot and Normal Probability Plot


ABV
30.00
25.00
20.00
15.00
10.00
5.00
.00
26538
26540
26537
26539
99771
99769
99770
99772
16367
25734
25735
25736
25711
97820
25713
25712
97821
69623
11512
11513
172
170
173
171
196912
113244
66486
1335
113245
11503
218
1593
2369
2370
2377
5080
5079
2387
5078



The box plot shows a clear indication of the data being right-skewed with outliers on both ends, but the majority on the right and including extreme outliers (indicated by the stars on the graph). The normality test indicates that the data is non-normal, with it being positively skewed.

1.6. Outliers using the 1.5*IQR rule. Interpret and explain in plain English your finding. Include list of possible outliers and what might be the possible reasons why they are outlier(s)

IQR = Q3-Q1

Q3= 7.7, Q1=5.

IQR=2.7*1.5=4.05.

Outliers will be outside of Q1-4.05, and Q3+4.05, or 0.95 and 11.75, respectively.

In other words, the majority of beers should fall within 0.95 and 11.75 ABV. The outliers could exist for several reasons: 1) Error in entry by the website user. Perhaps the person entering the data fat-fingered a number and put in the incorrect ABV. 2) Also quite likely is that there are beers that have a low ABV (consider the low-to-non-alcoholic O'Doul's) or extremely high ABV (Dogfish Head 120 Minute IPA, for instance, has an ABV of 18).

1.7. Findings

The data indicates that beers reviewed on this particular website typically have an ABV of 5.75, and more broadly speaking, generally fall between 0.95 and 11.75 ABV. The highest ABV of a beer reviewed is 29, and the lowest is 0.01, which is quite interesting in terms of beers existing with such extreme ABVs in them. Quite frankly, I'm surprised that the people reviewing the 29 ABV beer were capable of typing up a review after drinking it.

2. Regression Analysis: 43 points

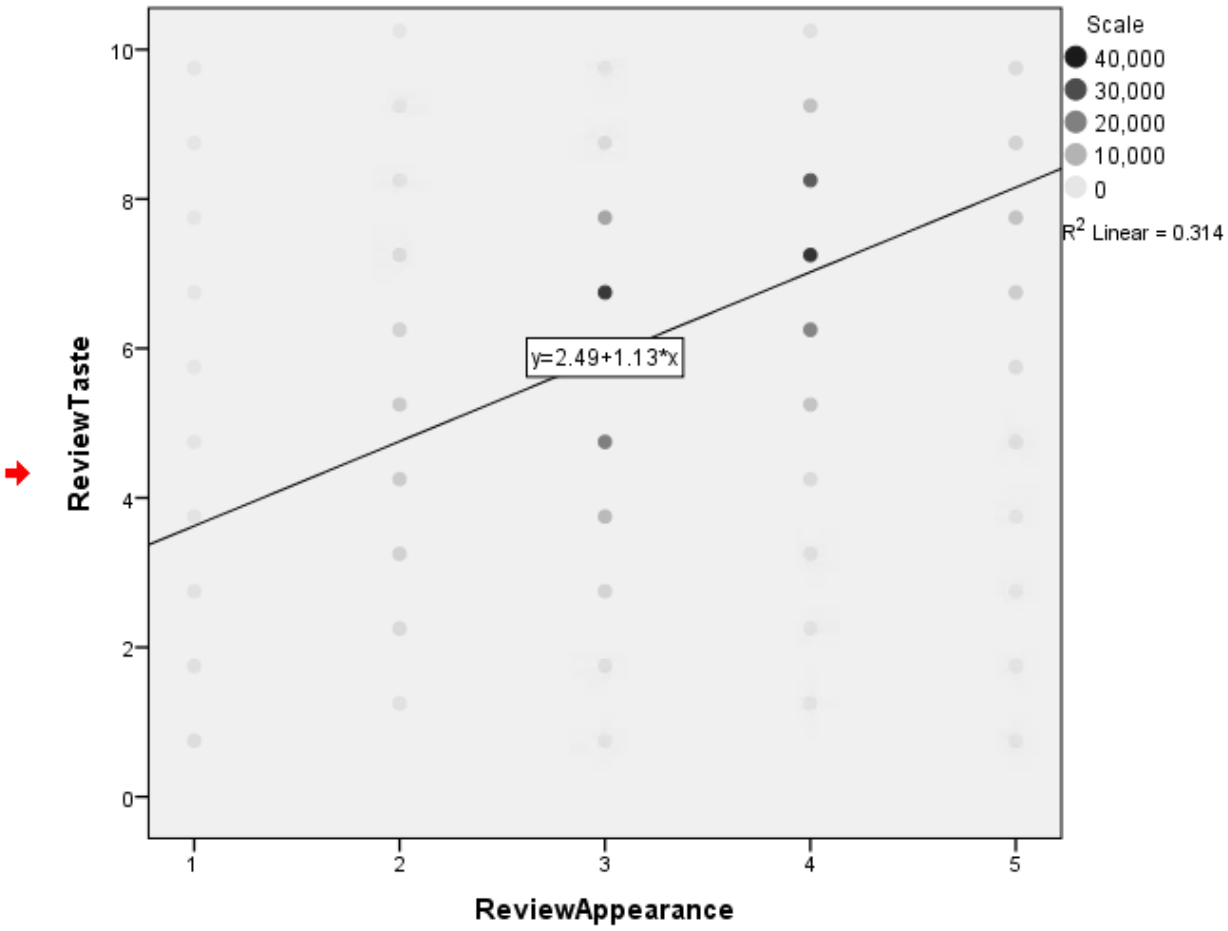
For this regression analysis we begin with a simple question: will the way a beer looks affect the way a person perceives its taste? We know, for instance, that restaurants will serve their food on ornate plates and spend great amounts of time working on its presentation. Similarly, will the way we perceive a beer's look (it's foaminess, clarity, and color) affect how we rate its flavor?

2.1. Response and Predictor Variables

Response: Review Taste
Predictor: Review Appearance

The question being asked is: does the way a beer look influence the way a beer tastes? In order to do this, we take the Review Taste (out of 10 total possible points) and compare it against the Review Appearance (which is out of 5 total possible points).

2.2. Scatter Plot with Regression Line



(Due to the number of data points at over 200,000, I also went ahead and binned the points on the scatterplot to provide a clearer view of distribution and weight)

Review taste and review appearance seem to have a strong positive correlation. As the rating for an appearance in a beer goes up, so does the rating on its taste. Furthermore, by looking at the strength of the binned points, we can determine that there appears to be no significant outliers in the data. This could indicate that the way a beer looks affects how the drinker perceives its flavor.

2.3. Correlations

Correlations

		ReviewTaste	ReviewAppearance
ReviewTaste	Pearson Correlation	1	.560**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	747101.831	206982.843
	Covariance	2.803	.777
	N	266508	266508
ReviewAppearance	Pearson Correlation	.560**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	206982.843	182727.472
	Covariance	.777	.686
	N	266508	266508

**. Correlation is significant at the 0.01 level (2-tailed).

$r^2 = 0.314$

$r = 0.560$

At .560, the correlation coefficient is above .5, and thus indicates that there is a strong positive correlation between the way a beer looks and the way that it tastes. This is promising, as it indicates that there are factors outside of the actual taste of a beer that effect how we experience its flavor.

2.4. Coefficient of Determination

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.560[a]	.314	.314	1.387	.314	121886.614	1	266506	.000

a. Predictors: (Constant), ReviewAppearance

b. Dependent Variable: ReviewTaste

$r^2 = 0.314$, which indicates that 31% of the taste ratings received can be explained by the appearance ratings received. This is likely due to the data being skewed, but is to be expected since we know there are other variables that contribute to how we experience a beer's flavor such as appearance, actual taste, etc.

2.5. Slope and Intercept

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.492	.011		219.374	.000
	ReviewAppearance	1.133	.003	.560	349.123	.000

a. Dependent Variable: ReviewTaste

$y = mx+b$

Slope = 1.133

Intercept (b) = 2.492

1.133 *0 + 2.492 = 2.823

This indicates that, if an appearance rating were to equal 0, a beer's flavor would likely to be rated at 2.823.

2.6. Regression Equation to Predict

ReviewAppearance	ReviewAroma	ReviewPalate	ReviewTaste	ReviewOverall
4	6	3	6	13
4	6	4	7	13
2	4	2	4	8
2	4	2	4	8
3	6	3	6	12
3	5	2	5	9
4	7	4	7	15
3	6	4	7	13
3	7	3	7	15
3	7	3	7	15
4	7	3	7	15
3	7	3	6	12

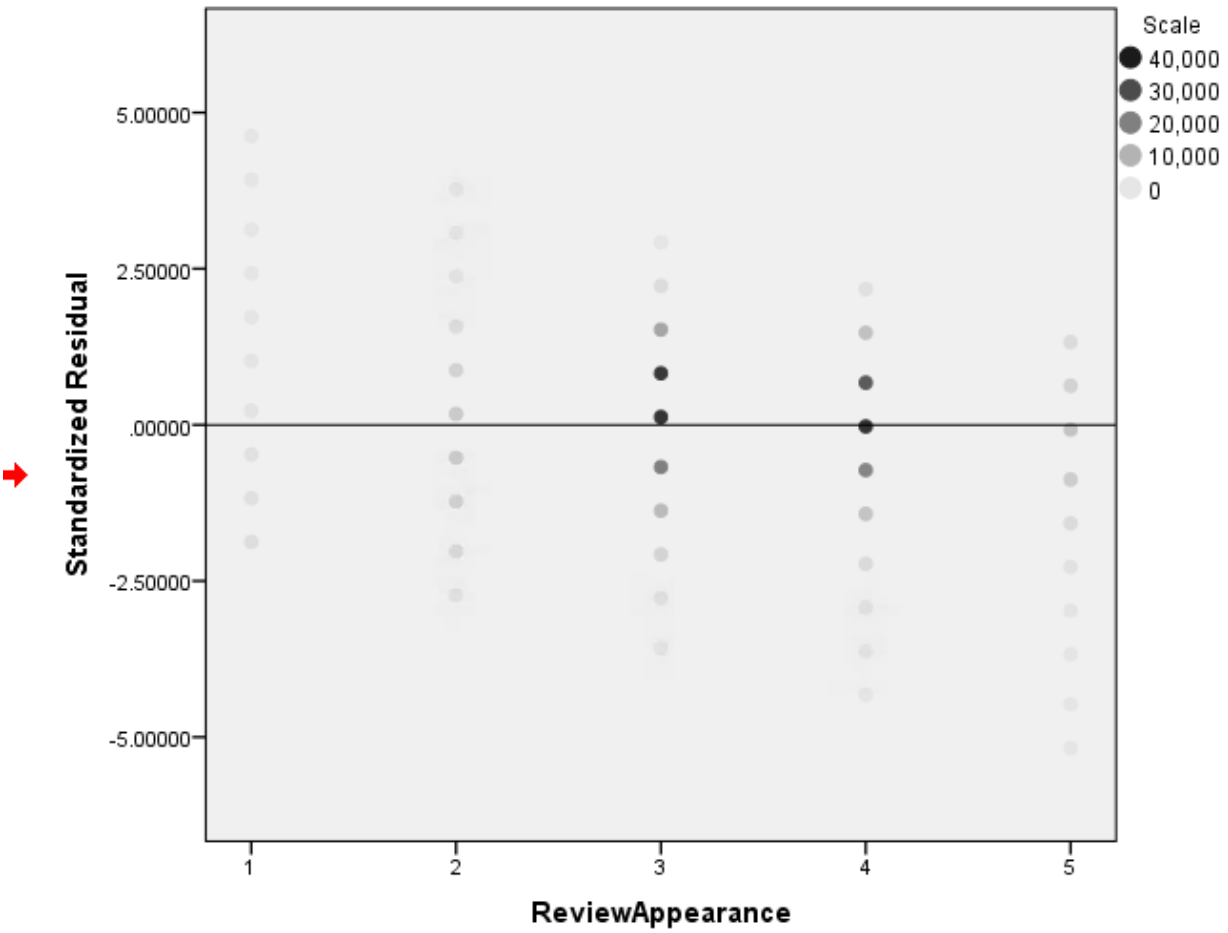
The highlighted record was chosen to test our model. In the record, the review appearance is rated at 2. According to our model, then, taste *should* be rated at 1.133*2+2.492 = 4.758. The actual Y value is 4, which is a difference of 0.758 in ratings, or about 7.5% difference from the actual value to the predicted value. This seems to suggest that the model is fairly accurate for determining how a beer will rank in taste given a rating for its appearance.

2.7. Residual and Residual Plot

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.62	8.16	6.34	.938	266508
Residual	-7.155	6.376	.000	1.387	266508
Std. Predicted Value	-2.900	1.931	.000	1.000	266508
Std. Residual	-5.159	4.597	.000	1.000	266508

a. Dependent Variable: ReviewTaste



The data for the residual plot is binned again in order to show strength of the overlapping data points. Looking at the strength of the points, no strong pattern seems to emerge. Furthermore, (again, looking particularly at the weight of the binned points) the residual error seems to be constant, and within -2 and 2. This seems to indicate that appearance ratings are a good predictor of flavor ratings for beer.

2.8. Conclusion

In conclusion, the appearance rating of a beer (or, the way a person perceives the look of a beer) has an effect on the flavor rating of a beer (or, the way a person perceives the quality of the flavor of a beer). The results also show that beer appearance accounts for 31 percent of the changes we see in flavor ratings, which is an acceptable number considering we know of other factors at play that will influence how we experience tasting a beer. Additional factors that could influence the flavor of a beer (aside from the flavor itself) would be factors like actual taste (of course), as well as the aroma and palate (also known as mouthfeel) of the beer.

3. Contingency Table (Two-way Table): 21 points

For the next part of the analysis, the data had to be reduced in size in order to be able to run a crosstab on it. SPSS has some limitations around record numbers in cross tabulation reports, with the upper end being at 1000 values.

Therefore, it became necessary to limit the data being used in this analysis by choosing a random sample. This was done through a script written in the command line.

3.1. Qualitative Variables

In total, the dataset has the following qualitative variables: Beer Name, Beer Style, Profile Name, Beer ID, Brewer, Season, Strength. For this study, we'll be looking at beer strength (mild is <4.5, strong is >4.5 and <9, and very strong is >9 ABV).

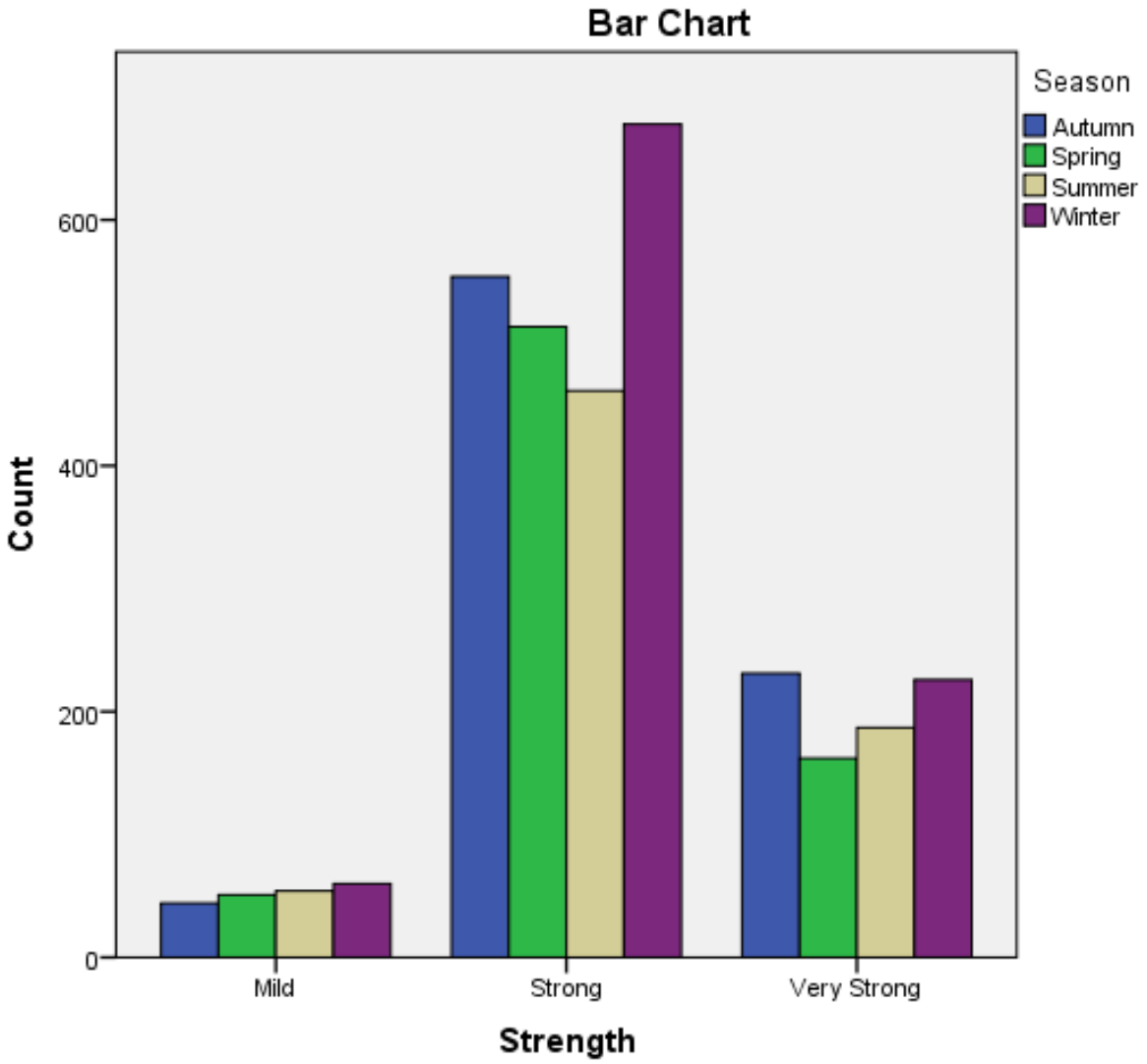
3.2. Contingency Table

Strength * Season Crosstabulation

Count

		Season				
		Autumn	Spring	Summer	Winter	Total
Strength	Mild	44	51	54	60	209
	Strong	554	513	461	678	2206
	Very Strong	231	162	187	226	806
Total		829	726	702	964	3221

3.3. Clustered Bar Graph



Strong beers are clearly the most popular ones to drink, followed by very strong, and then finally mild. Mild beers are most popular in the Winter, followed by Summer, Spring, then Autumn. Strong beers are most popular in the Winter, followed by Autumn, Spring, then Summer. And finally, very strong beers are most popular in Autumn, followed by Winter, Summer, then Spring.

This gives us a glimpse into the beer's people drink during any season, and begins to paint a broader picture that indicates preference for beer based on the time of year. It also poses some new questions: what leads people to drink stronger beers during colder seasons? Why do people drink less beer during warmer seasons?

3.4. Two Way Table without Percentages

Strength * Season Crosstabulation

Count

		Season				
		Autumn	Spring	Summer	Winter	Total
Strength	Mild	44	51	54	60	209
	Strong	554	513	461	678	2206
	Very Strong	231	162	187	226	806
Total		829	726	702	964	3221

About two thirds of beers drunk are of the strong (between 4.5 and 9 ABV) category. The next most popular type of beer to drink is the very strong variety (9 ABV and above), followed by the mild category, which is below 4.5 ABV. Neither Spring nor Summer have a more popular strength of beer compared to Winter and Autumn, though mild beers do not appear to be popular during the Autumn. More beer is drunk in the colder Winter and Autumn seasons than in the warmer Spring and Summer seasons.

The people who drink beer and rate them on this website clearly prefer drinking strong beer, which is also associated with the craft beer movement. We may begin to infer from this that most people who rate beer on the website are craft beer drinkers, then.

3.5. Two Way Table with Percentages

Strength * Season Crosstabulation

			Season				
			Autumn	Spring	Summer	Winter	Total
Strength	Mild	Count	44	51	54	60	209
		% within Strength	21.1%	24.4%	25.8%	28.7%	100.0%
		% within Season	5.3%	7.0%	7.7%	6.2%	6.5%
		% of Total	1.4%	1.6%	1.7%	1.9%	6.5%
	Strong	Count	554	513	461	678	2206
		% within Strength	25.1%	23.3%	20.9%	30.7%	100.0%
		% within Season	66.8%	70.7%	65.7%	70.3%	68.5%
		% of Total	17.2%	15.9%	14.3%	21.0%	68.5%
	Very Strong	Count	231	162	187	226	806
		% within Strength	28.7%	20.1%	23.2%	28.0%	100.0%
		% within Season	27.9%	22.3%	26.6%	23.4%	25.0%
		% of Total	7.2%	5.0%	5.8%	7.0%	25.0%
Total		Count	829	726	702	964	3221
		% within Strength	25.7%	22.5%	21.8%	29.9%	100.0%
		% within Season	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	25.7%	22.5%	21.8%	29.9%	100.0%

During the Winter and Autumn seasons when it's cold, a higher proportion of stronger beers are drunk (at both strong and very strong) than in the Spring and Summer seasons. This is likely due to the fact that during the colder months, darker beers are more popular, which have a higher ABV, and thus are stronger beers.

Looking at seasons individually, during Autumn very strong beers are preferred. During Spring, mild beers are preferred. During Summer, mild beers are also preferred. And finally, during Winter, strong beers are most preferred. This aligns with our hypothesis earlier that during warmer seasons, people drink less strong beer, and during colder seasons, people drink stronger beer. This aligns well with the "lawn mowing beer" idea, that is to say, a good lawn mowing beer (to be drunk when it's hot outside) is typically a lighter one.

3.6. Findings and Conclusions

Perhaps not surprisingly, during the colder seasons people drink more beer, and drink stronger beer. During the warmer seasons, people tend to drink less beer. This is likely due to stronger

beer styles being sold more commonly during the colder seasons, as well as stronger alcohol generally being preferred when it gets colder outside.

Furthermore, strong beers between 4.5 and 9 ABV are the most popular beers to drink, while very strong beers are the next most favored, and mild beers being the least favored. This indicates that people who use the website do not generally drink traditional American Pilsners, which fall at or under 4.5 ABV.

Given that, it's safe to assume a few things about this website and their data:

1) The people who rate beers on the website have a preference for craft beer, which has a much higher ABV then traditional beers.

2) As a result of that, the beers that people drink are of a higher strength (ABV).

3) When rating the flavor of these beers, as seen in part 2 of this analysis, factors such as the beer's appearance are just as important as its flavor, indicating a more discerning beer taster concerned with multiple facets of their beer's composition.

Beer Advocate users do indeed seem to represent a different market segment than the traditional beer drinker. They drink beers that are of a higher alcoholic content than traditional beers, are influenced by attention to flavor, but also things like how a beer looks. And finally, they change the kinds of beers they drink seasonally, indicating shifting preferences and an eye toward diversity in the beer that they drink.