Beer Reviews Data Insights

EVOLENT HEALTH

Hello!

I am Abhishek Jadhav

Email : abhishek.jadhav317@gmail.com

Contact : 9623139647

LinkedIn : https://www.linkedin.com/in/abhishek-jadhav-80b92382/

The Data

1

Overall Data Statistics

Dataset statistics

Number of variables	13
Number of observations	528870
Missing cells	20514
Missing cells (%)	0.3%
Duplicate rows	0
Duplicate rows (%)	0.0%

Variable types

Numeric	9
Categorical	4

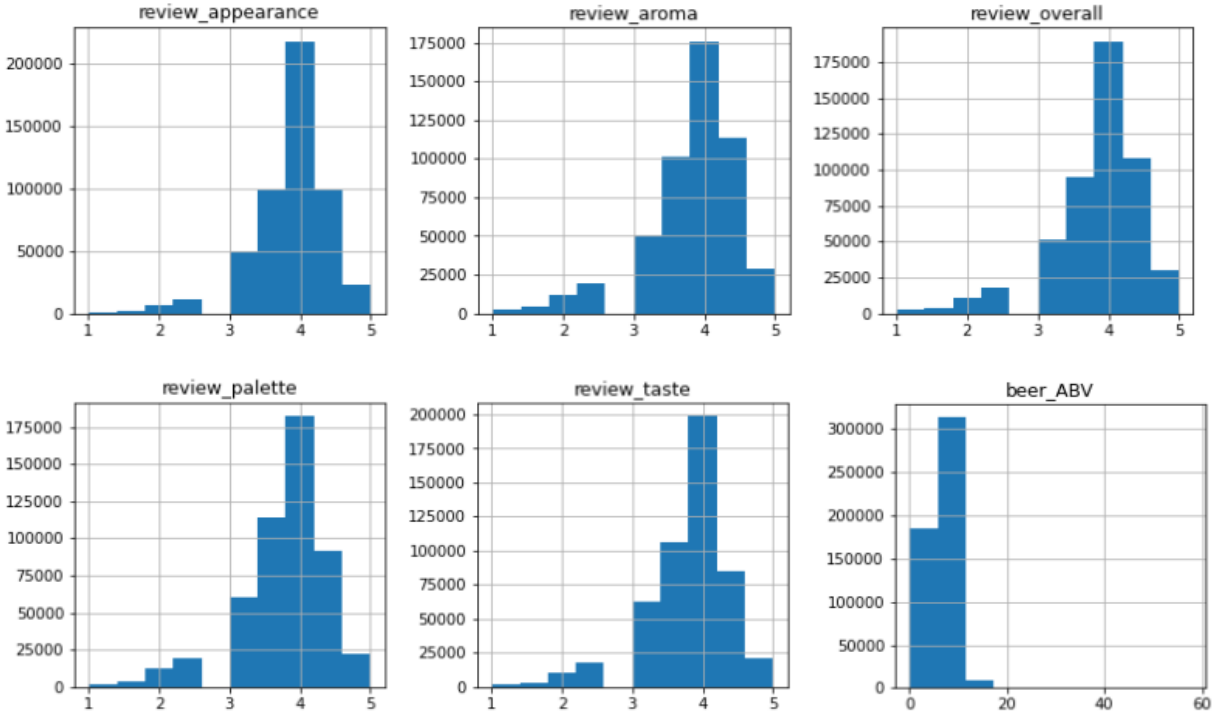
FEATURES

- **Beer_Abv**
  - This column specifies the alcohol percentage in beer.
  - Range is between 0.01% - 57.7% . Mean value is 7%
  - 3.8% of rows have unknown beer_abv %
- **Beer and Brewery Identification**
  - Beer_BeerId : Unique Id of Beer
  - Beer_BrewerId : Unique Id of Brewery
  - Beer_name : Name of the Beer
  - Beer_Styple : Style of the Beer
- **Beer Reviews and Rating information**
  - Review_appearance : Ratings on Appearance of the Beer
  - Review_palette : Ratings on Beer Palette
  - Review_taste : Ratings on Beer's taste
  - Review_aroma: Ratings on Beer's aroma
  - Review_Overall : Overall rating of the Beer

Descriptive Analysis

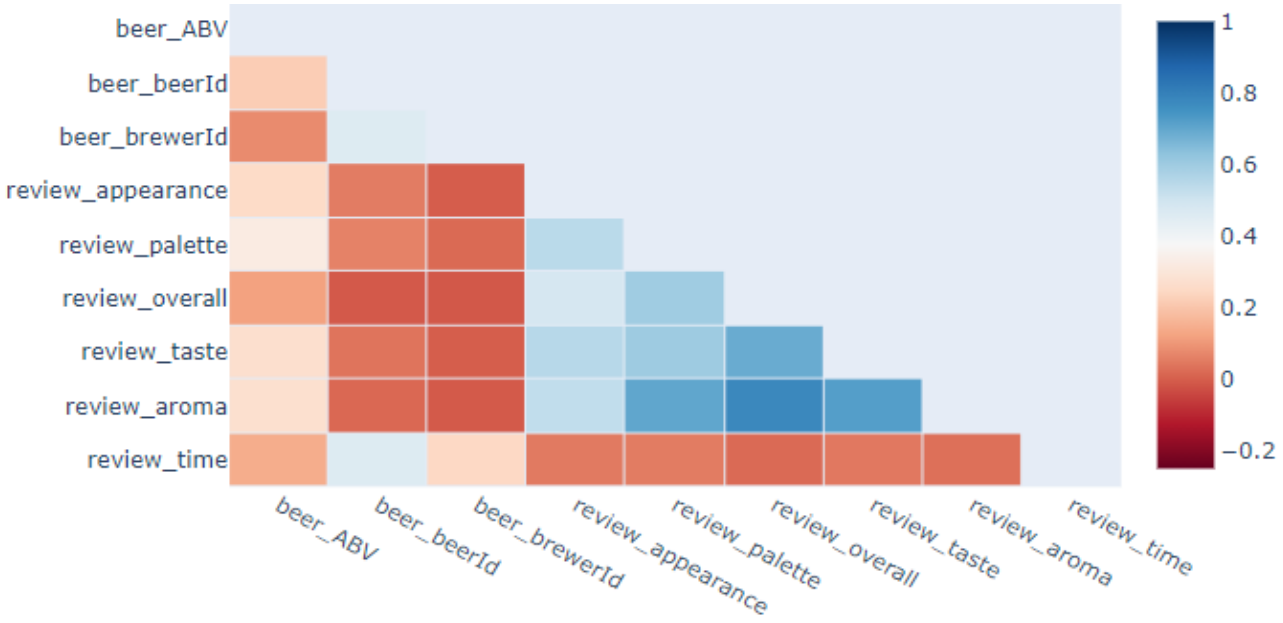
2

Histograms of continuous features



- All the review related features are slightly normally distributed
- Beer_Abv is Right skewed (+ve) distribution

Correlation Matrix



review_aroma have high positive correlation with review_overall(0.78) , review_taste(0.72) , review_palette(0.70)

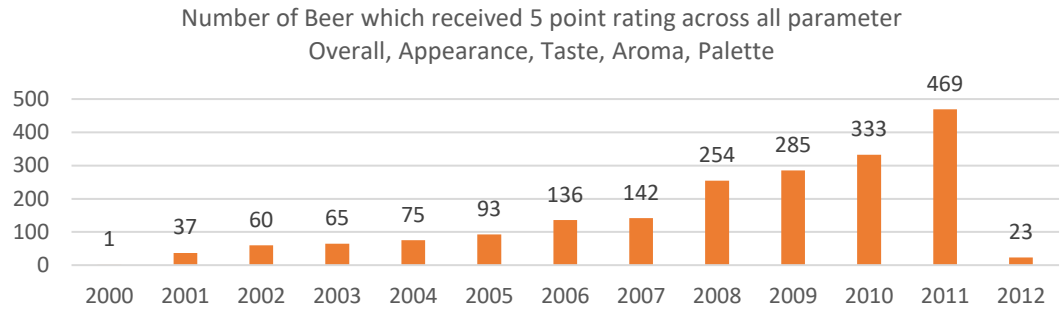
Answers to the Questions

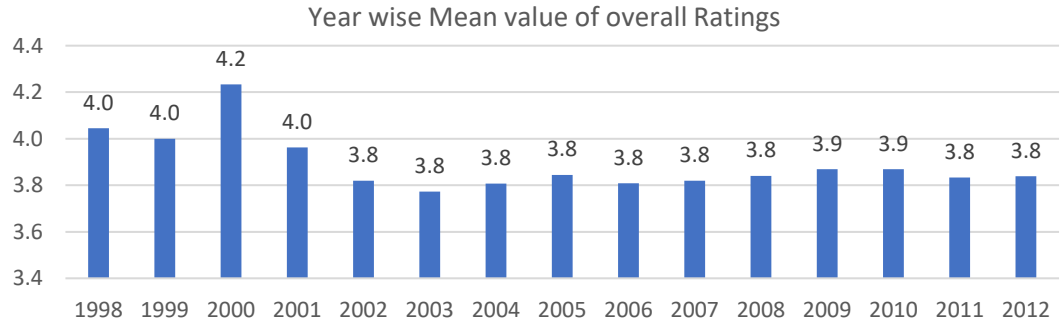
2

Rank top three breweries which produce strongest beers.

- Strength of the beer depends on the alcohol percentage present in the beer.
- We can use average value of alcohol percentage of beer per breweries to find strong beer producing breweries
- But the beer_abv column which gives us the beer's alcohol percentage has high positive skewness. And as average is highly sensitive to data skewness median value to represent strength of the beer would be better option.
- Top 3 Breweries which produce strongest beers are
    1. brewer_id = 736
    2. brewer_id = 5562
    3. brewer_id = 6513 and 36

Which year did beers enjoy the highest ratings?

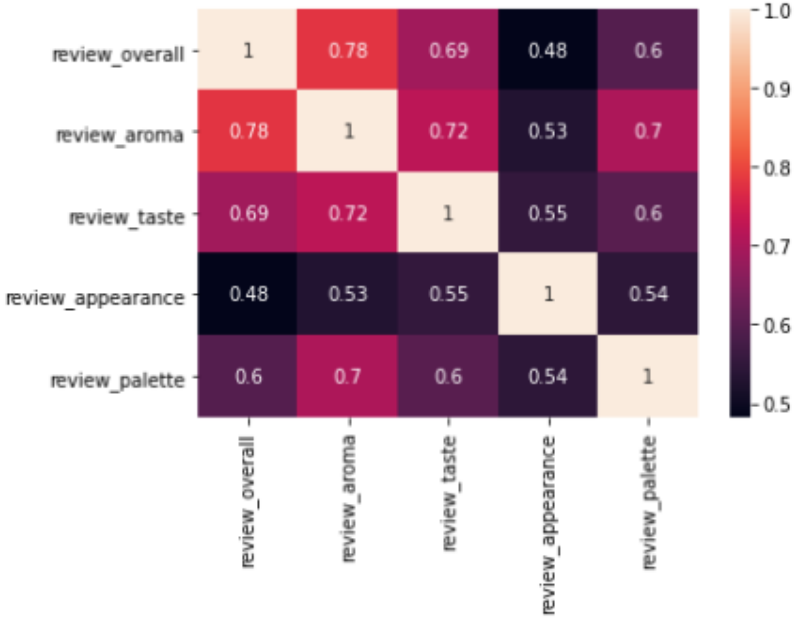




Year 2000 has highest mean values of overall rating.
Year 2011 has highest number(469) of beers having 5 rating in all the parameters

Based on the user's ratings which factors are important among taste, aroma, appearance, and palette?

- Overall Ratings has high correlation with Aroma, Taste and Palette
- We can find out which factors are playing important role in increasing overall rating by finding feature importance using Random Forest Regressor Algorithm by keeping overall ratings as target variable.
- Feature importance after executing Random Forest Regressor
  - review_aroma : 0.92
  - review_taste : 0.056
  - review_appearance : 0.0098
  - review_palette : 0.0088
- Aroma is certainly the largest driving factor for improving overall ratings



If you were to recommend 3 beers to your friends based on this data which ones, will you recommend?

- If I have to recommend a beer to someone who is new to the taste of beer, I will first find out which beers are popular in the industry.
- Number of reviews received can be good parameter to find out the popularity of the beer.
- After finding the popular beers I would pick those beer which are having highest overall ratings.
- Mass pleasing choice can be good choice to do a cold start.
- 3 Beers I would suggest to my friends are :
  - Founders Breakfast Stout
  - Trappistes Rochefort 10
  - La Fin Du Monde

Which Beer style seems to be the favorite based on reviews written by users?

- I used spacy library to clean up the Stop words which are not important for text data analysis.
- I used contractions library to fix contractions words like didn't to did not.
- I used nltk package to find out sentiment behind each written review.
- SentimentIntensityAnalyzer will give polarity of each review if the polarity is between -1 to +1 . Negative polarity states that reviews are not good whereas positive polarity states that reviews are good.
- Eisbock beer style has highest positive polarity score. So, it is definitely the most favourite beer style based on written reviews

How does written review compare to overall review score for the beer styles?

- I used percentile bucketing technique to club the beers into 10 Quantile buckets. Bucket number 10 has higher values and Bucket 1 has lower values.
- Percentile Bucketing is done on Polarity Score of written reviews and Overall Ratings.
- Based on the data I observed that beer styles belonging to lower bucket have lower overall ratings. And beer style belonging to higher bucket i.e. Bucket 10 have higher overall ratings.
- So, we can infer that if the written reviews are good than there are high chances that the overall rating will also be high.

Thanks!