CASE STUDY – EDA ON BEER REVIEW

Background:

Here's a fictional dataset of beer reviews with various features/data points. You're required to understand the data set and perform 2 tasks listed below.

Tasks 01: EDA

• Use the Beer dataset and conduct an end-to-end EDA to derive insights using Python on a Jupyter notebook

• Make use of libraries like numpy, pandas, matplotlib, and seaborn to conduct your analysis (feel free to use additional libraries based on your comfort level)

• Please find below the expected EDA process:

1. Data Cleaning (Handling Missing Values, Outliers, etc.)
2. Data Exploration – create hypotheses leading to useful insights and compute the data and visualizations required to support them (e.g., Which beer is the best?)
3. What are the weirdest beers and why?
4. Recommend 3 beers, based on all possible data horizons, to your buddy based on this dataset? And why?
5. What are the most important factors for estimating the overall quality of a beer? 6. Which brewery makes the best beers, according to reviewers?

Task 02: KPI generation:

Let's assume this data set is used by a brewery as way to keep a track of their competition and is refreshed on a weekly basis. The owner of the brewery wants to track certain KPIs that help him know what the competition is upto. Please suggest 2 KPIs from this data set that should be added to a weekly dashboard for owner of the brewery. You just need to state

• What is the KPI and how is it calculated

• What is the value provided by using this KPI

SOLUTION

CONTENT

Understanding the data:

This dataset is the list of beers produced along with its brewery's name, beer style and 5 ratings. The dataset also includes the name of the reviewer, time taken to review and the abv (alcohol by volume) for each beer. The dataset contains the following 13 columns with 1586614 rows

- brewery_id
- brewery_name
- review_time
- review_overall
- review_aroma
- review_appearance
- review_profilename
- beer_style
- review_palate
- review_taste
- beer_name
- beer_abv
- beer_beerid

Out of these, following are the columns representing ratings:

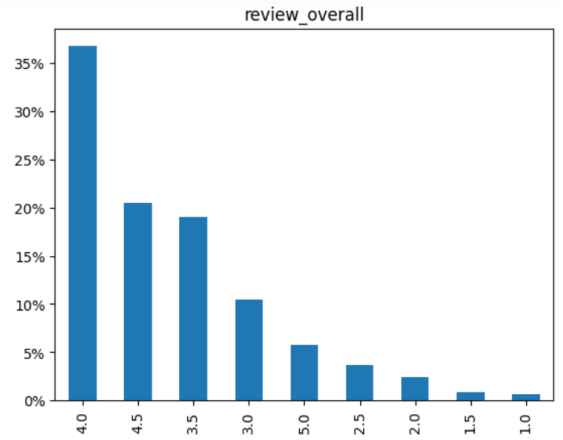
- review_overall: The overall characteristics/quality and the personal experience of the beer
- review_aroma: The smell of the beer
- review_appearance: The color, clarity, head and visual carbonation of this beer.
- review_palate: The body of the beer, carbonation and astringency.
- review_taste: The flavors in this beer, thinking about the palate, bitterness and finish.

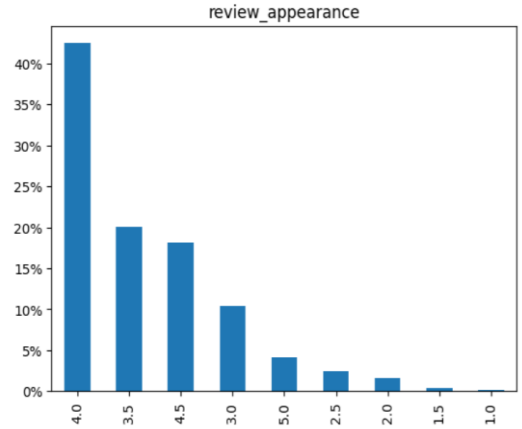
The dataset was then checked for missing values which were later filled with the data of appropriate central tendency.

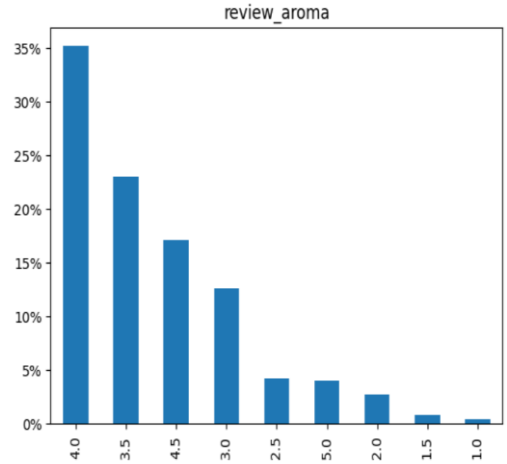
EDA:

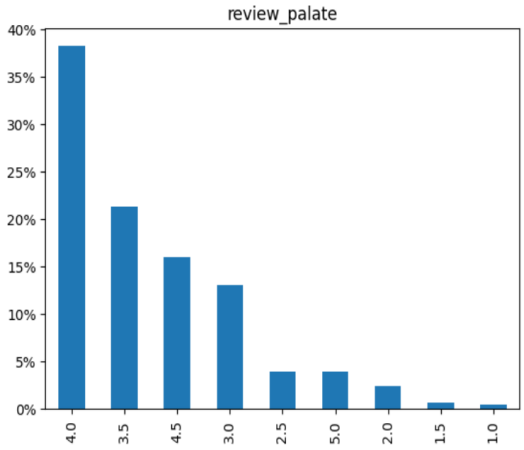
On performing univariate analysis on the ratings variables, we find that the ratings are 1 – 5 in the interval of 0.5. Key insights about the distribution of ratings:

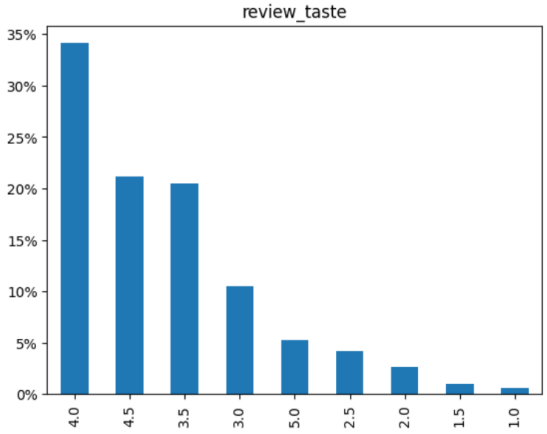
- review_taste, review_palate, review_aroma, review_appearance: 85% of the rating is between 3 - 4.5
- review_appearance: 90% of the rating is between 3 - 4.5











Hypothesis:

On researching about beers, I found there has been few statements mentioned related to range of abv, average abv of beers in the world, relationship of beer alcohol content to people's choice. Thus, I hypotheses the following statements:

1. Average abv of the beers in the world is 5%. This sample dataset represents population
2. The overall characteristics and your personal experience of the beer is strongly influenced by aroma and taste
3. Stronger beers (above 10% abv) are rated 5 for overall characteristics
4. More the review time, better is the rating i.e beers rated 4, 5 for overall characteristics has taken more time for review
5. Atleast 80% of the beers have abv between 4% - 7%

Hypothesis 01: Average abv of the beers in the world is 5%. This sample dataset represents population

On performing 2 tailed test of significance, taking level of significance as 95%, I confirmed that the data represents the population.

Hypothesis 02: The overall characteristics and your personal experience of the beer is strongly influenced by aroma and taste

Taking overall as target variable and aroma, appearance, taste, palate as independent variable, I fit a linear regression model. The model has r-square value of 0.65 indicating a good fitted model. The coefficients of independent variables were:

	Coefficients	Confidence Interval		P-Value
	0	0	1	1
review_aroma	0.069482	0.068066	0.070899	0.0
review_appearance	0.096297	0.095060	0.097534	0.0
review_taste	0.556976	0.555437	0.558515	0.0
review_palate	0.284283	0.282800	0.285765	0.0

All variables have p-values=0, indicating that they are statistically significant predictors of overall
The value of r2> 0.6, we can consider this model to be a good fit for the data.
Looking at the coefficient of the parameters of the model, palate & taste has strong relationship with overall quality. Thus, hypothesis 02 is partially correct - Taste influences the overall but aroma doesn't

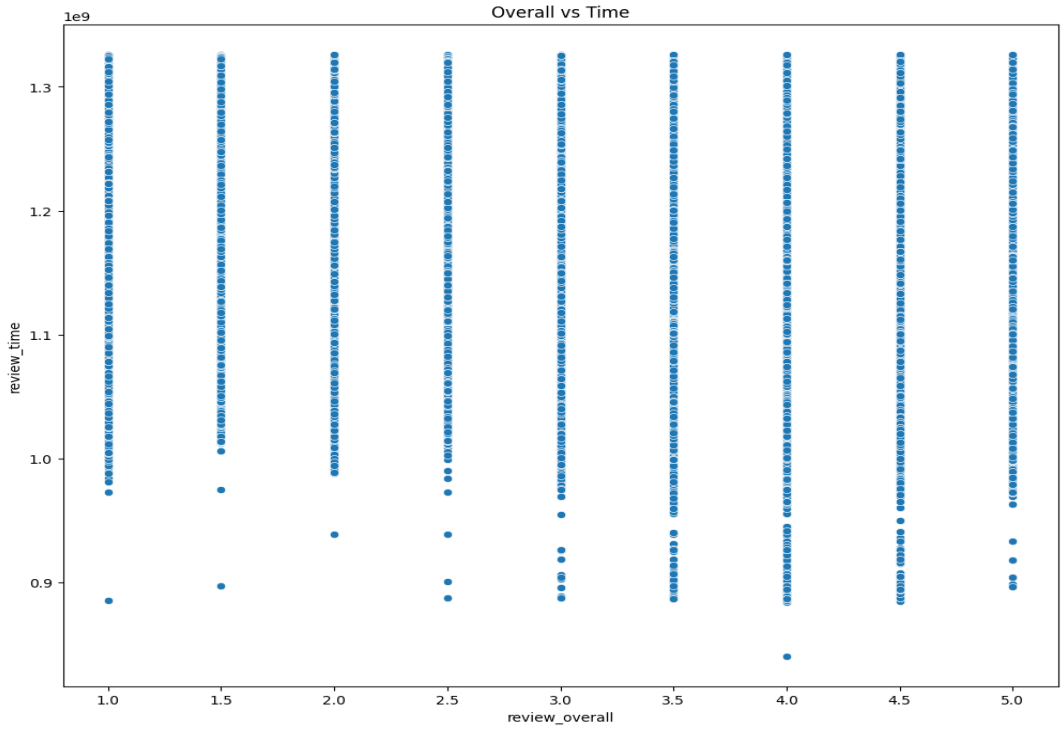
Hypothesis 03: Stronger beers (above 10% abv) are rated 5 for overall characteristics

It is found that for beers with abv>10, 90% of the data has overall ratings below 5. Thus, hypothesis 3 is rejected.

```
ABV:
 {'Min': 10.01, '10%ile': 10.3, 'Q1': 10.5, 'Q2 (Median)': 11.03, 'Q3': 12.0, '90%ile': 14.5, '95%ile': 17.5, '99%ile': 18.2,
'Max': 57.7}

 Overall:
 {'Min': 1.0, '10%ile': 3.0, 'Q1': 3.5, 'Q2 (Median)': 4.0, 'Q3': 4.5, '90%ile': 4.5, '95%ile': 5.0, '99%ile': 5.0, 'Max': 5.0}
```

Hypothesis 04: More the review time, better is the rating i.e beers rated 4, 5 for overall characteristics has taken more time for review



This shows that there is no relationship between the time taken for review and the overall quality. Hypothesis 04 is not true

Hypothesis 05: Atleast 80% of the beers have abv between 4% - 7%

From data, we found that Less than 75% of the beer has abv < 7%, thus, hypothesis 5 is rejected too

```
summary_stats(db["beer_abv"])
```

```
{'Min': 0.01,
 '10%ile': 4.8,
 'Q1': 5.3,
 'Q2 (Median)': 6.5,
 'Q3': 8.4,
 '90%ile': 10.0,
 '95%ile': 11.0,
 '99%ile': 14.0,
 'Max': 57.7}
```

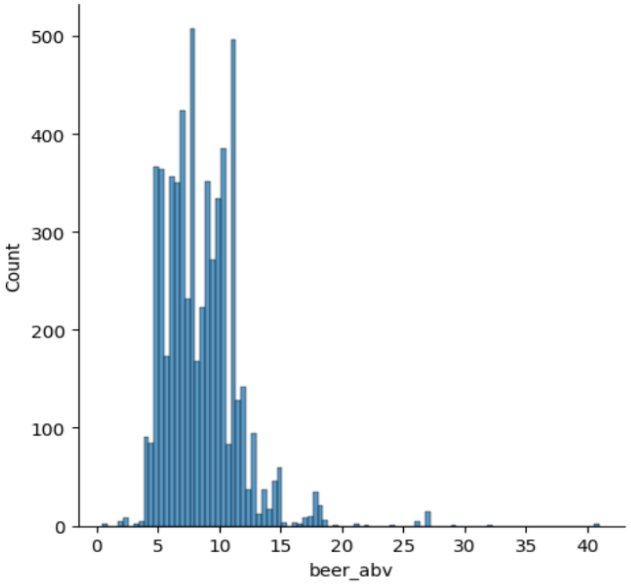
Feature Engineering:

The new rating system is dervied from all 5 ratings - aroma, appearance, taste, palate and overall. It is taken as weighted sum of each variable.

Previous rating	Weights
Appearance	10%
Palate	10%
Aroma	20%
Taste	20%
Overall	40%

Source: https://www.ratebeer.com/our-scores

Based on the rating, I found 5967 beers (0.4%) having an exact rating of 5. The mean abv is 8.1 with maximum of 41 and minimum of 0.5



Weird Beer:

With these wide range of beers having abv value from 0.5 – 41, it is weird that they all are rated 5. It means they all share an overall great personal experience in terms of all 4 character – aroma, taste, palate (mouthfeel) and appearance. Thus, I would suggest trying beers from highest abv to lowest abv and at last, a beer with exact average of both. Presenting:

Brewery	Beer style	Name	abv
BrewDog	American Double / Imperial IPA	Sink The Bismarck!	41
Erdinger Weissbräu	Low Alcohol Beer	Erdinger Weissbier Alkoholfrei	0.5
Cerveceria Vegana, S.A.	Low Alcohol Beer	Malta India	0.5
Boston Beer Company (Samuel Adams)	American Strong Ale	Samuel Adams Millennium	21

3 Beer I would recommend

For someone I know, I would recommend having a beer with unique and rare beer style. From the whole dataset, I found 4 styles which have just a single count out of 5967 rows. This could also be interpreted as this style is quite rare too. The good part is that it is rated 5. Thus, we can confirm the quality and experience of the beer would be great.

Thus, I recommend any 3 beer out of:

Brewery	Beer style	Name	Abv
Brauerei Tucher Brau	Kristalweizen	Kristall Weizen	5.1
Arbor Brewing Company	English Pale Mild Ale	Arbor Brewing Big Ben House Mild	3.5
Belhaven Brewery Company Ltd.	English Stout	Belhaven Scottish Stout	7
Mahrs-Bräu	Keller Bier / Zwickel Bier	Mahr's Ungespundet-hefetrüb	5.2

KPI 1

KPI_01 = Change in L1

$$L1 = \frac{\text{no.of beers with (rating=5)}}{\text{total no.of beer}} \text{ X 100; at Brewery level, filtered by abv}$$

Abv can be bucketed into 3 categories as low(0 – 5%); medium (5 – 14%); high(14 – 60%). The value must be checked for each of the 3 categories. Values of L1 ranges from 0.02 – 100. This will acts as the first level of analytics. A deviation of even 1% will act as a trigger to check if there has been surge or decrease in the number of finest beers produced by the competitors

But it can also trigger a false alarm if no. of beers reviewed are quite less and they are of good quality. For eg: If brewery A reviews 3 beers and all 3 beers are rated 5, then L1 = 100%.

A condition of minimum total no. of beers > 741 must be applied to avoid such false alarm. 741 represent 10%ile value of the count of brewery with rating = 5

KPI 2

KPI_02 = Change in L2

$$L2 = \frac{\text{no.of beers with (rating=5)}}{\text{total no.of beer}} \text{ X 100; at beer style level, filtered by abv}$$

Abv can be bucketed into 3 categories as low(0 – 5%); medium (5 – 14%); high(14 – 60%). The value must be checked for each of the 3 categories. Values of L2 ranges from 0.03 – 1.73. This will acts as the first level of analytics.

A deviation of even 0.1% will act as a trigger to indirectly check if there has been surge or decrease in the beer styles accepted by the customers in the market. If there is an increase in beer styles, then it indicates users are accepting new style of beers. If there is a decrease, it indicates users are happy with the current available styles.

Best Brewery

For the best brewery, I picked **Brouwerij Westvleteren (Sint-Sixtusabdij van Westvleteren).** It has the highest L1 = 7.91% with considerable number of total beer reviews (171 out of 2378).

---------- Thank you----------