Analysis on Consumer Preference Data for Black Coffee - Prabakaran Chandran - UChicago - MS ADS

Further Information on Data Available : https://github.com/prabakaranc98/ILoveCoffee/

Information and Description of Data :

The dataset provided here contains physical, chemical, and consumer preference data for 118 individual consumers who each tasted 27 distinct coffees prepared with precisely controlled brewing conditions, yielding a total of 3,186 individual tastings of black coffee. Physical measurements include the temperature, total dissolved solids, and percent extraction of the brewed coffee; chemical measurements include the pH and titratable acidity of the brewed coffee; and the consumer preference measurements include hedonic liking, just-about-right assessments of beverage temperature, flavor intensity, mouthfeel, and acidity, and purchase intent.

```
#importing the neccessary Libraries

import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
pd.set_option('display.max_columns', 100)

from IPython import display

from plotly.subplots import make_subplots
import plotly.graph_objects as go
```

```
# read the cotter's coffee preference dataset

coffee_testing = pd.read_csv("/content/cotter_dataset.csv")

coffee_testing.head()
```

	Judge	Cluster	Week	Session Number	Position	Brew	Temp.x	TDS.x	PE.x	Dose	Setting	Grind	Empty Carafe	Full Carafe	Brew Mass	TDS__
0	1	1	3	13	3	87-1.0-16	Low	Low	Low	176.1	1SM	5	1769.3	4598.4	2829.1	1.0
1	1	1	2	10	2	87-1.0-20	Low	Low	Medium	145.0	6LG	3	1764.3	4670.2	2905.9	0.9
2	1	1	3	13	9	87-1.0-24	Low	Low	High	119.0	4LG	3	1791.0	4675.1	2884.1	0.9
3	1	1	2	10	9	87-1.25-16	Low	Medium	Low	215.3	12LG	5	1775.1	4586.2	2811.1	1.0
4	1	1	3	13	10	87-1.25-20	Low	Medium	Medium	176.1	12LG	3	1782.1	4651.3	2869.2	1.2

```
# coffee_testing.info() --> as there are major null values
missing_data_columns = coffee_testing.isnull().sum()
null_columns = missing_data_columns[missing_data_columns > 0]

display.display(null_columns)
```

	0
Titration pH	84

dtype: int64

Only the Titration pH is having 84 null value columns , we shall not worry about this as this number is less than 3-5% and also in one particular columnand also , the primary analysis that we are going to focus on is

- How an Individual's decision on preferrence be impacted by various factors staring from relationship between sensory attributes and preference/intent
- How sensory attributes get impacted by the chemical and brewing properties , then focus on the way it is being served / how the diminishing temperature makes the preferrence impacted - I like to drink black coffees to be cooler than piping hot temperature - Let'see

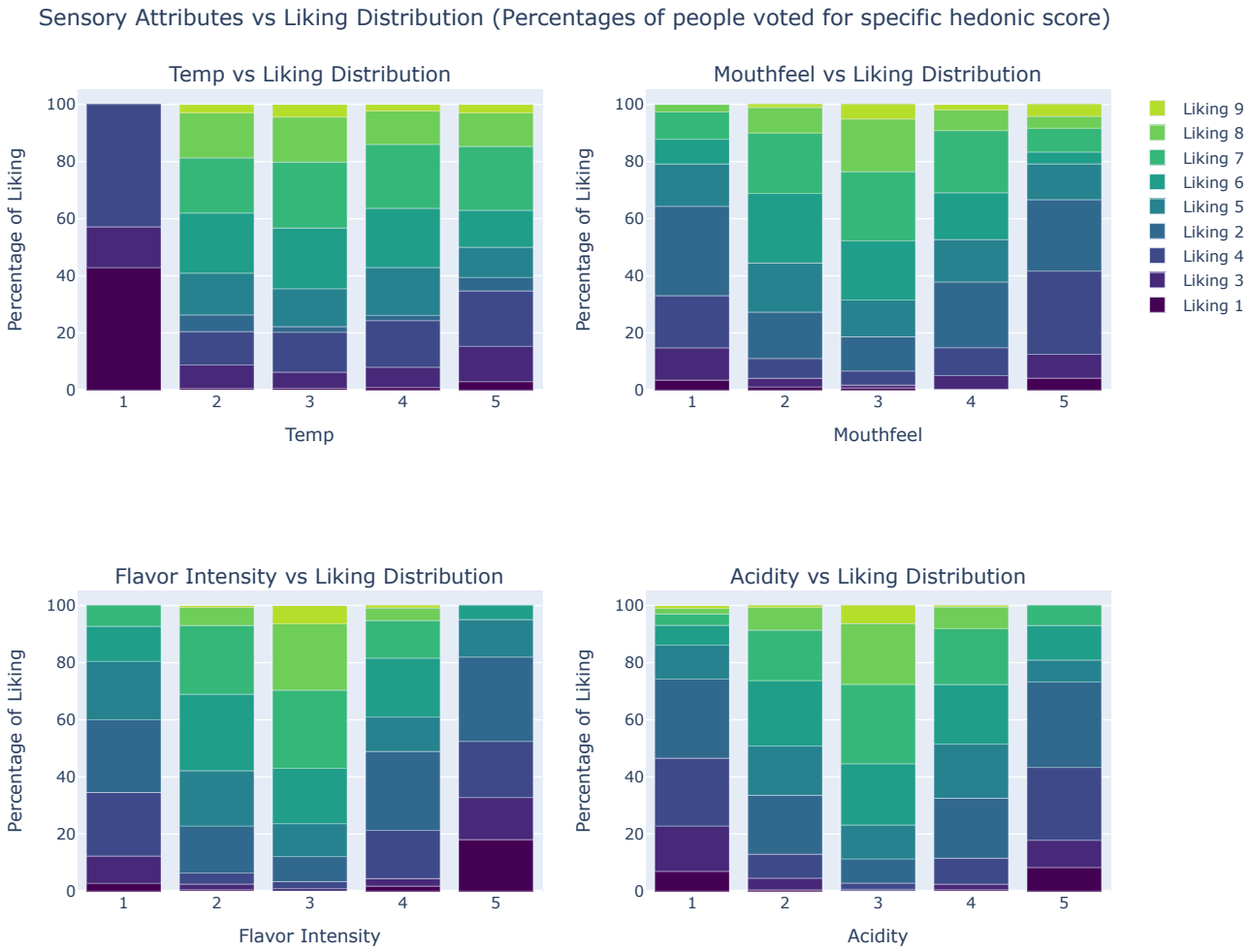
Irrespective of the brew parameters (if we consider brew and grinding parameters are the instrument variable on the sensory attributes) , the sensory attributes shall show the variation in liking / purchace intent

```
identifier_variables = ["Judge","Week","Session Number","Brew"]
sensory_attributes = ["Temp","Flavor.intensity","Acidity","Mouthfeel"]

impact_variables = ["Liking","Purchase.intent"]
coffee_testing_sensory = coffee_testing[identifier_variables+sensory_attributes+impact_variables]
```

```
### Let's calculate for each attribute variation , how does the liking percentage changes - unweighted ( we are not weightir
def calculate_percentage(df, attribute):
    grouped = df.groupby([attribute, "Liking"]).size().reset_index(name="Count")
    total_counts = grouped.groupby(attribute)["Count"].transform("sum")
    grouped["Percentage"] = (grouped["Count"] / total_counts) * 100
    return grouped

# Prepare data for each sensory attribute
temp_data = calculate_percentage(coffee_testing_sensory, "Temp")
mouthfeel_data = calculate_percentage(coffee_testing_sensory, "Mouthfeel")
flavor_data = calculate_percentage(coffee_testing_sensory, "Flavor.intensity")
acidity_data = calculate_percentage(coffee_testing_sensory, "Acidity")
```

```
fig = make_subplots(
    rows=2, cols=2,
    subplot_titles=["Temp vs Liking Distribution", "Mouthfeel vs Liking Distribution",
                    "Flavor Intensity vs Liking Distribution", "Acidity vs Liking Distribution"]
)
color_scale = px.colors.sequential.Viridis

# Temp vs Liking
for idx, liking in enumerate(temp_data["Liking"].unique()):
    temp_subset = temp_data[temp_data["Liking"] == liking]
    fig.add_trace(
        go.Bar(x=temp_subset["Temp"], y=temp_subset["Percentage"],
               name=f"Liking {liking}", legendgroup=f"Liking",
               marker_color=color_scale[idx]),
        row=1, col=1
    )

# Mouthfeel vs Liking
for idx, liking in enumerate(mouthfeel_data["Liking"].unique()):
    mouthfeel_subset = mouthfeel_data[mouthfeel_data["Liking"] == liking]
    fig.add_trace(
        go.Bar(x=mouthfeel_subset["Mouthfeel"], y=mouthfeel_subset["Percentage"],
               name=f"Liking {liking}", legendgroup=f"Liking", showlegend=False,
               marker_color=color_scale[idx]),
        row=1, col=2
    )

# Flavor Intensity vs Liking
for idx, liking in enumerate(flavor_data["Liking"].unique()):
    flavor_subset = flavor_data[flavor_data["Liking"] == liking]
    fig.add_trace(
        go.Bar(x=flavor_subset["Flavor.intensity"], y=flavor_subset["Percentage"],
               name=f"Liking {liking}", legendgroup=f"Liking",
               showlegend=False,
               marker_color=color_scale[idx]),
        row=2, col=1
    )

# Acidity vs Liking
for idx, liking in enumerate(acidity_data["Liking"].unique()):
    acidity_subset = acidity_data[acidity_data["Liking"] == liking]
    fig.add_trace(
        go.Bar(x=acidity_subset["Acidity"], y=acidity_subset["Percentage"],
               name=f"Liking {liking}", legendgroup=f"Liking", showlegend=False,
               marker_color=color_scale[idx]),
        row=2, col=2
    )

fig.update_layout(
    title="Sensory Attributes vs Liking Distribution (Percentages of people voted for specific hedonic score)",
```

```
    height=800,
    width=1000,
    barmode="stack",
    showlegend=True,
    xaxis_title="Temp",
    xaxis2_title="Mouthfeel",
    xaxis3_title="Flavor Intensity",
    xaxis4_title="Acidity",
    yaxis_title="Percentage of Liking",
    yaxis2_title="Percentage of Liking",
    yaxis3_title="Percentage of Liking",
    yaxis4_title="Percentage of Liking",
)
fig.show()
```



We could see that , Most people like (hedonic scale > 6) is very evidentt in the medium range of temerature , internsity,acidity and mouth feel

- very significant percentage for score 8 and 9 we could see in value 3 for all the attributes, 8 and 9 denotes they like the product verymuch to strongly as per hedonic score
- are we in simpson paradox , can we just try to divide this further into the brew condition and other chemical parameters

as We could see flavouring attributes and NaOH , pH would also impact / as an instrument variable to the sensory attributes - a good causal discovery problem , this analysis let's focus on the asssociation but to reduce the simpson pardoxial condition

the avaialble brewing condition , for example denotes the target brew conditions. The first number identifies the target brew temperature, the second number identified the target total dissolved solids (TDS), and the third number identifies the target percent extraction (PE). For example, 87-1.0-16 denotes 87 degrees Celsius, 1% TDS, and 16% PE

```
coffee_testing_sensory["Brew"].unique()
```

```
array(['87-1.0-16', '87-1.0-20', '87-1.0-24', '87-1.25-16', '87-1.25-20',
       '87-1.25-24', '87-1.5-16', '87-1.5-20', '87-1.5-24', '90-1.0-16',
```