Chocolate

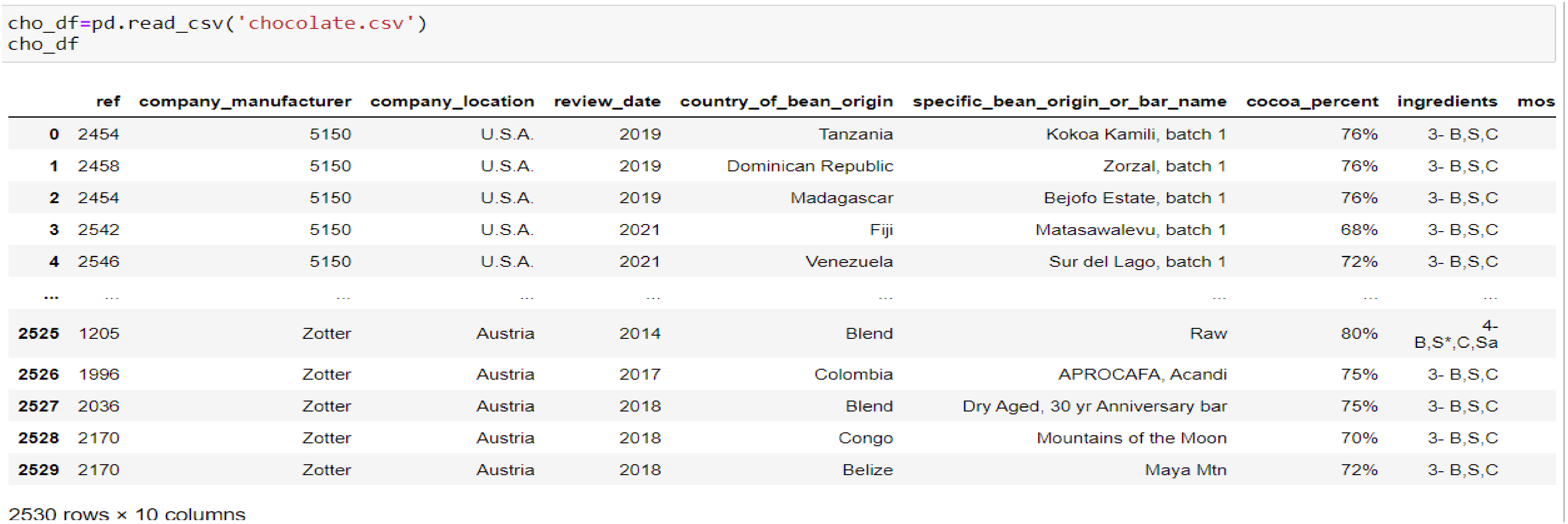
DATA ANAYTICS

Project

PREPARED BY

Rajshree Shah

Knowing basic composition of data



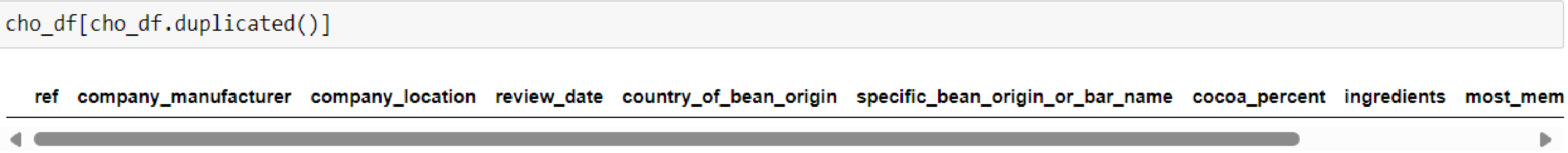
```
cho_df=pd.read_csv('chocolate.csv')
cho_df
```

	ref	company_manufacturer	company_location	review_date	country_of_bean_origin	specific_bean_origin_or_bar_name	cocoa_percent	ingredients	mos
0	2454	5150	U.S.A.	2019	Tanzania	Kokoa Kamili, batch 1	76%	3- B,S,C	
1	2458	5150	U.S.A.	2019	Dominican Republic	Zorzal, batch 1	76%	3- B,S,C	
2	2454	5150	U.S.A.	2019	Madagascar	Bejofo Estate, batch 1	76%	3- B,S,C	
3	2542	5150	U.S.A.	2021	Fiji	Matasawalevu, batch 1	68%	3- B,S,C	
4	2546	5150	U.S.A.	2021	Venezuela	Sur del Lago, batch 1	72%	3- B,S,C	
...	...	...	...	...	...	...	...	...	
2525	1205	Zotter	Austria	2014	Blend	Raw	80%	4- B,S*,C,Sa	
2526	1996	Zotter	Austria	2017	Colombia	APROCAFA, Acandi	75%	3- B,S,C	
2527	2036	Zotter	Austria	2018	Blend	Dry Aged, 30 yr Anniversary bar	75%	3- B,S,C	
2528	2170	Zotter	Austria	2018	Congo	Mountains of the Moon	70%	3- B,S,C	
2529	2170	Zotter	Austria	2018	Belize	Maya Mtn	72%	3- B,S,C	

2530 rows × 10 columns

The given data includes information about chocolate, such as the company manufacturer, company location, review date, country of bean origin, specific bean origin or bar name, cocoa percentage, ingredients, most memorable characteristics, and rating. chocolate market, including its size, growth rate, and segmentation by product type, distribution channel, and geography. The market is driven by factors such as the increasing demand for high-quality and organic chocolate products, the popularity of dark chocolate due to its health benefits, and the rise of bean-to-bar chocolate makers. The market is also segmented by distribution channel, with supermarkets and hypermarkets leading the way.

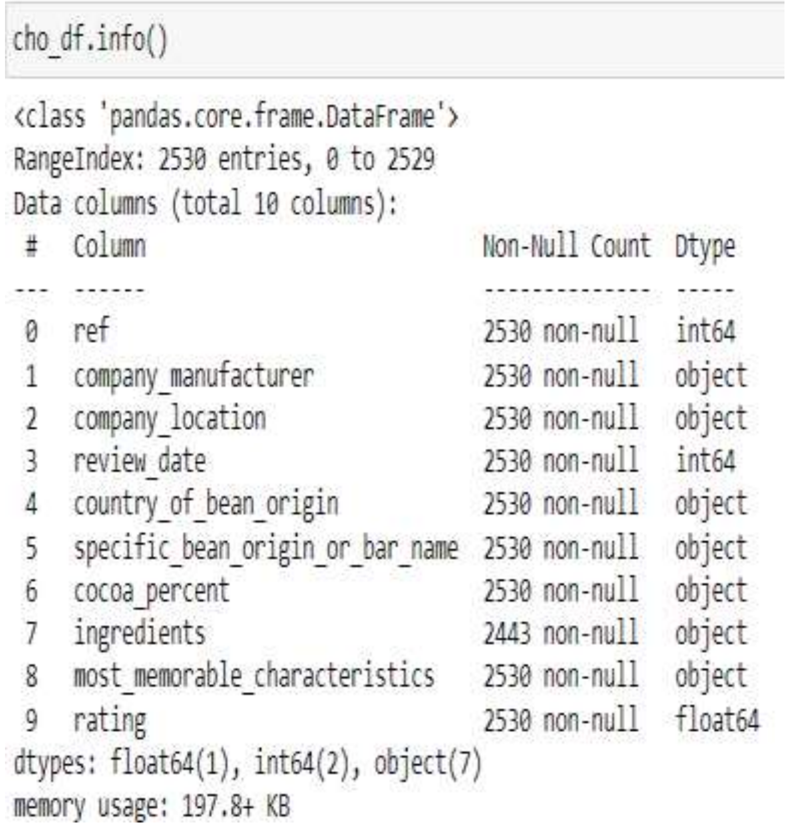
Remove Duplicate



```
cho_df[cho_df.duplicated()]
```

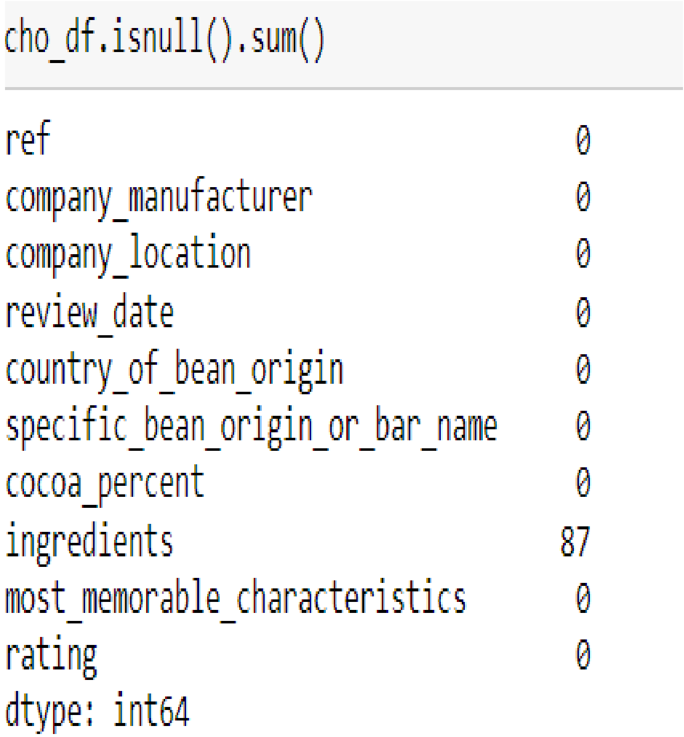
ref	company_manufacturer	company_location	review_date	country_of_bean_origin	specific_bean_origin_or_bar_name	cocoa_percent	ingredients	most_mem

Dealing with missing values



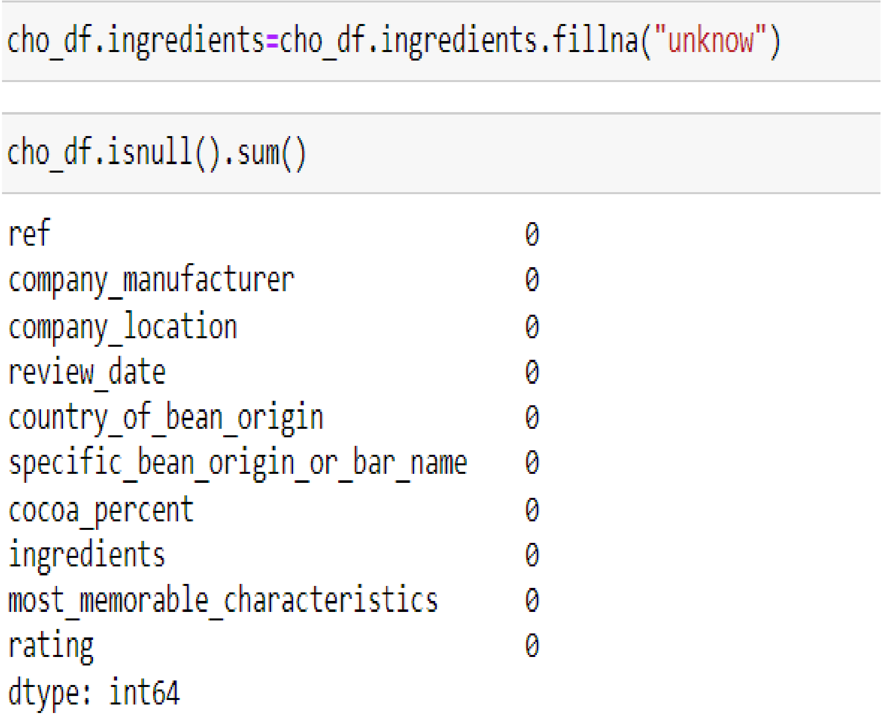
```
cho_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2530 entries, 0 to 2529
Data columns (total 10 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   ref                               2530 non-null   int64  
 1   company_manufacturer              2530 non-null   object 
 2   company_location                  2530 non-null   object 
 3   review_date                       2530 non-null   int64  
 4   country_of_bean_origin            2530 non-null   object 
 5   specific_bean_origin_or_bar_name  2530 non-null   object 
 6   cocoa_percent                     2530 non-null   object 
 7   ingredients                       2443 non-null   object 
 8   most_memorable_characteristics    2530 non-null   object 
 9   rating                            2530 non-null   float64
dtypes: float64(1), int64(2), object(7)
memory usage: 197.8+ KB
```



```
cho_df.isnull().sum()
```

```
ref                                  0
company_manufacturer                 0
company_location                     0
review_date                          0
country_of_bean_origin               0
specific_bean_origin_or_bar_name     0
cocoa_percent                        0
ingredients                         87
most_memorable_characteristics       0
rating                               0
dtype: int64
```

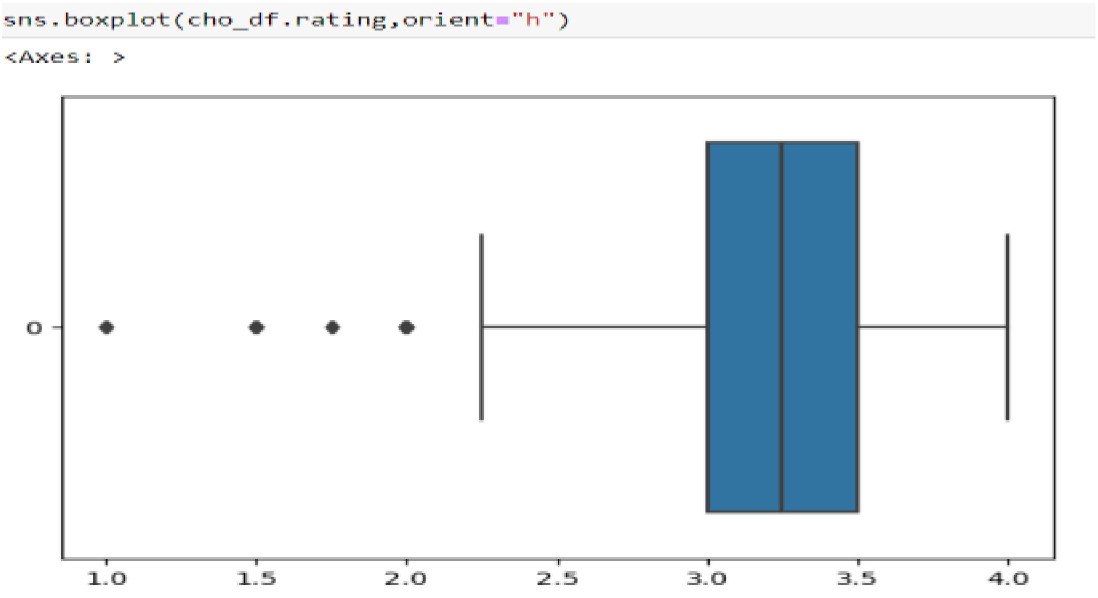


```
cho_df.ingredients=cho_df.ingredients.fillna("unknow")
```

```
cho_df.isnull().sum()
```

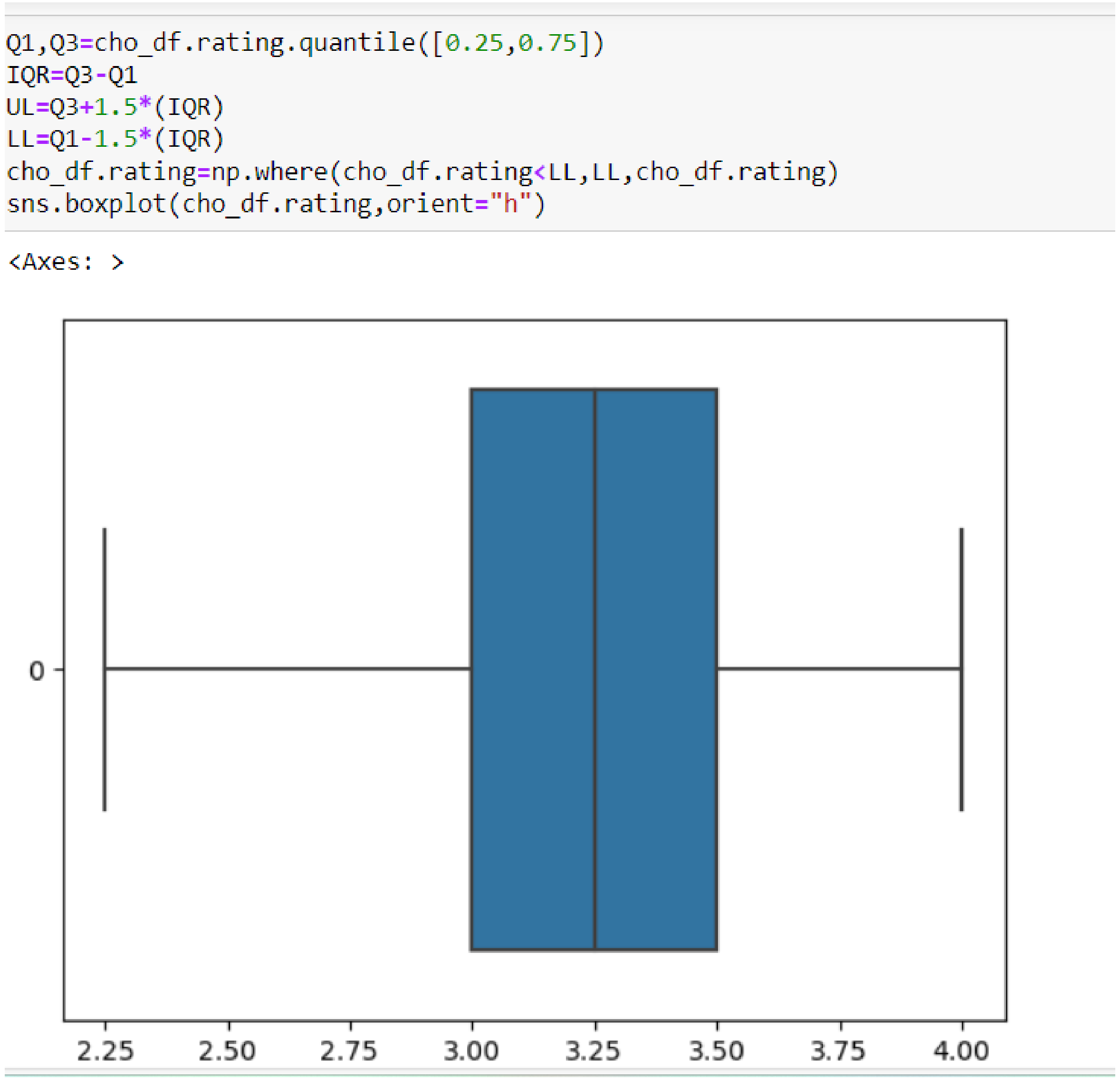
```
ref                                 0
company_manufacturer                0
company_location                    0
review_date                         0
country_of_bean_origin              0
specific_bean_origin_or_bar_name    0
cocoa_percent                       0
ingredients                         0
most_memorable_characteristics      0
rating                              0
dtype: int64
```

To create a box plot



```
sns.boxplot(cho_df.rating,orient="h")
```

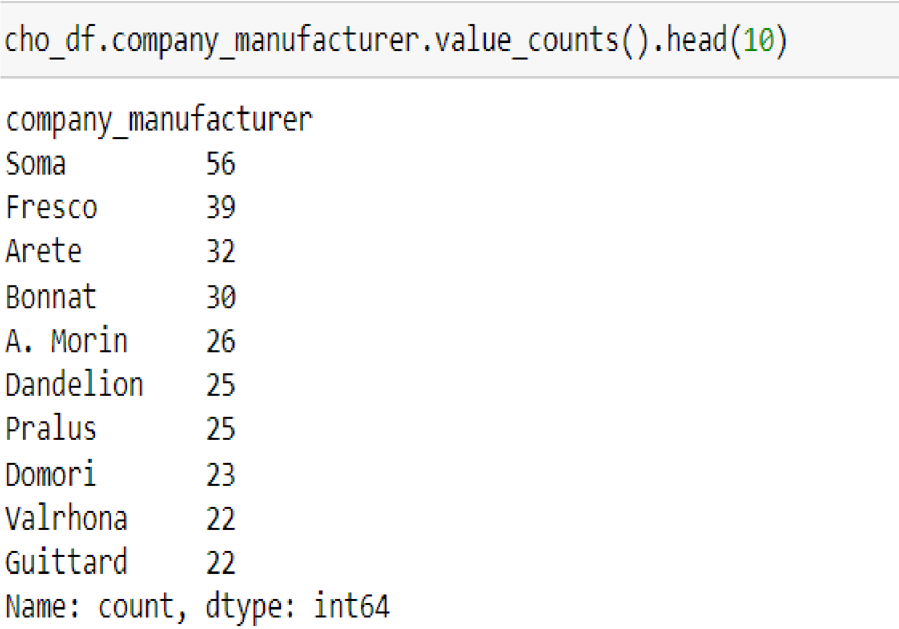
```
<Axes: >
```



```
Q1,Q3=cho_df.rating.quantile([0.25,0.75])
IQR=Q3-Q1
UL=Q3+1.5*(IQR)
LL=Q1-1.5*(IQR)
cho_df.rating=np.where(cho_df.rating<LL,LL,cho_df.rating)
sns.boxplot(cho_df.rating,orient="h")
```

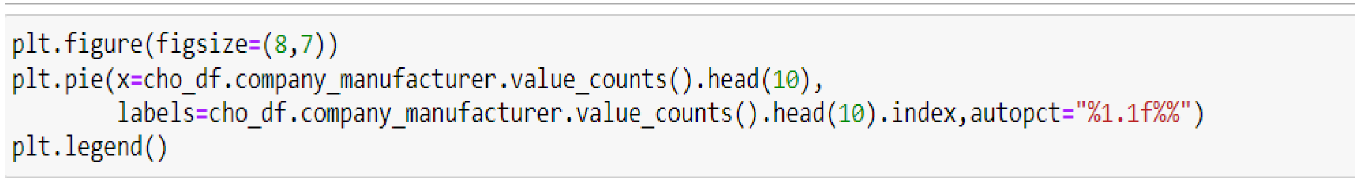
```
<Axes: >
```

To create a top 10 company_manufacturer frequency

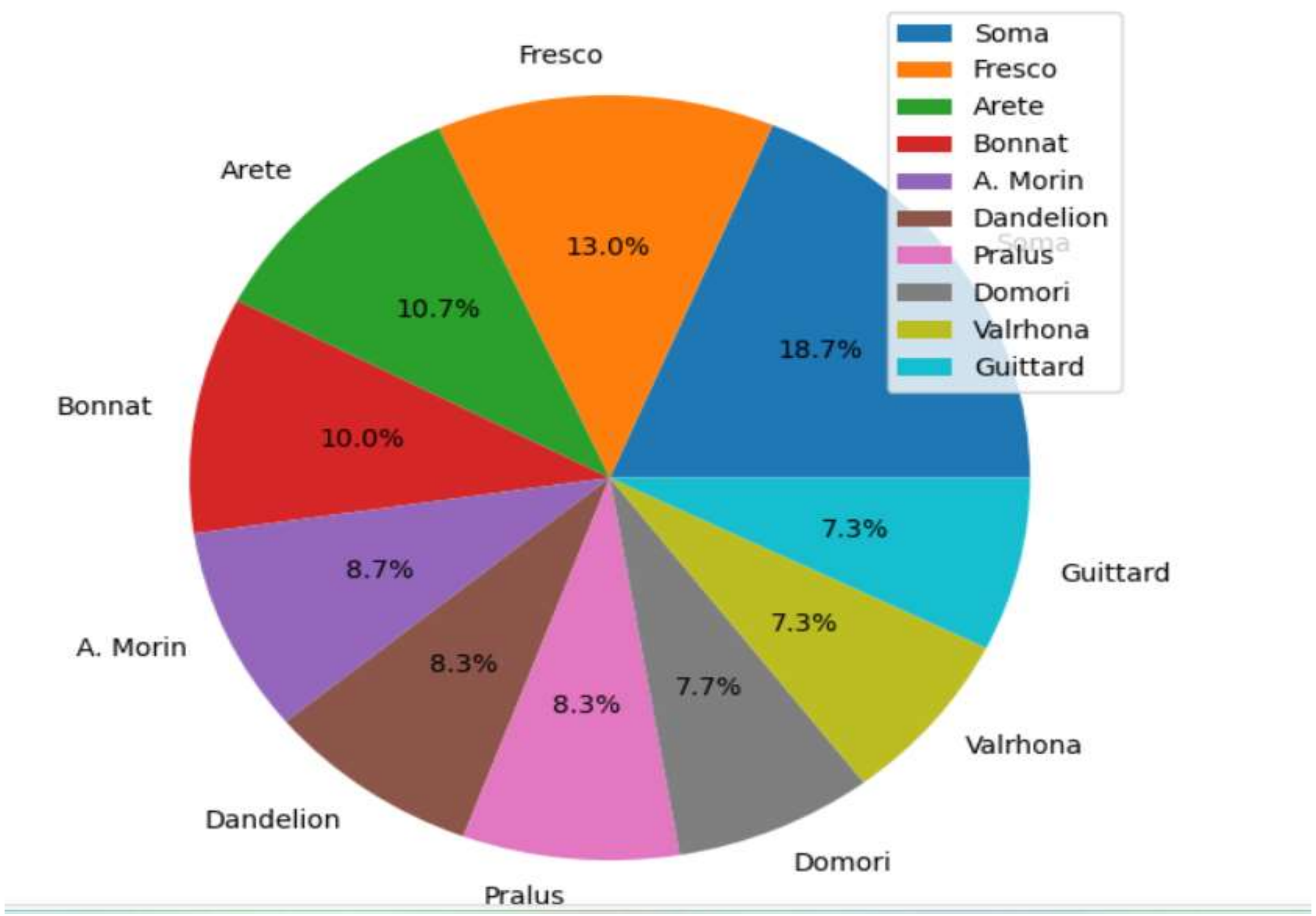


```
cho_df.company_manufacturer.value_counts().head(10)
```

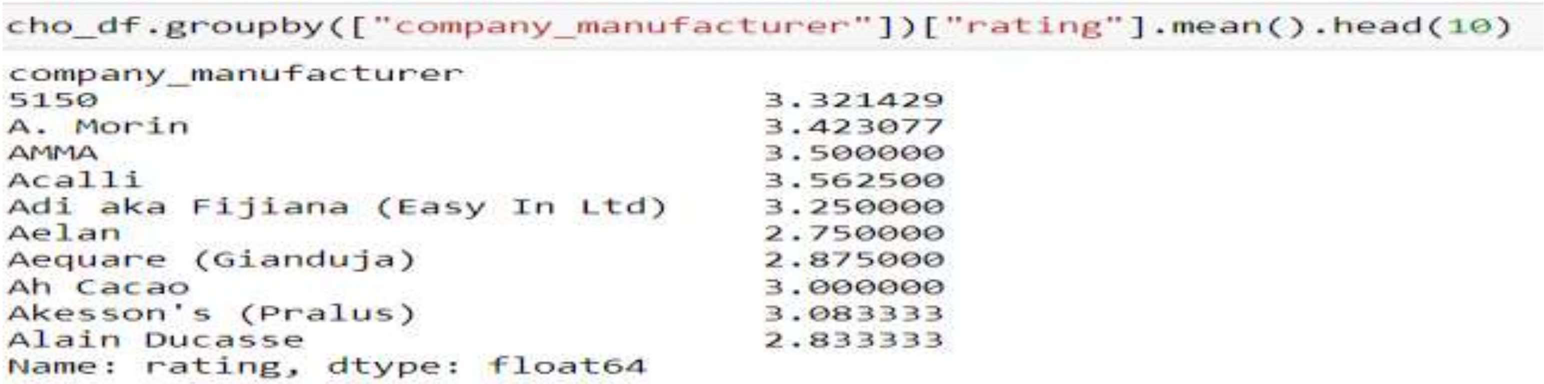
```
company_manufacturer
Soma         56
Fresco       39
Arete        32
Bonnat       30
A. Morin     26
Dandelion    25
Pralus       25
Domori       23
Valrhona     22
Guittard     22
Name: count, dtype: int64
```



```
plt.figure(figsize=(8,7))
plt.pie(x=cho_df.company_manufacturer.value_counts().head(10),
        labels=cho_df.company_manufacturer.value_counts().head(10).index,autopct="%1.1f%%")
plt.legend()
```



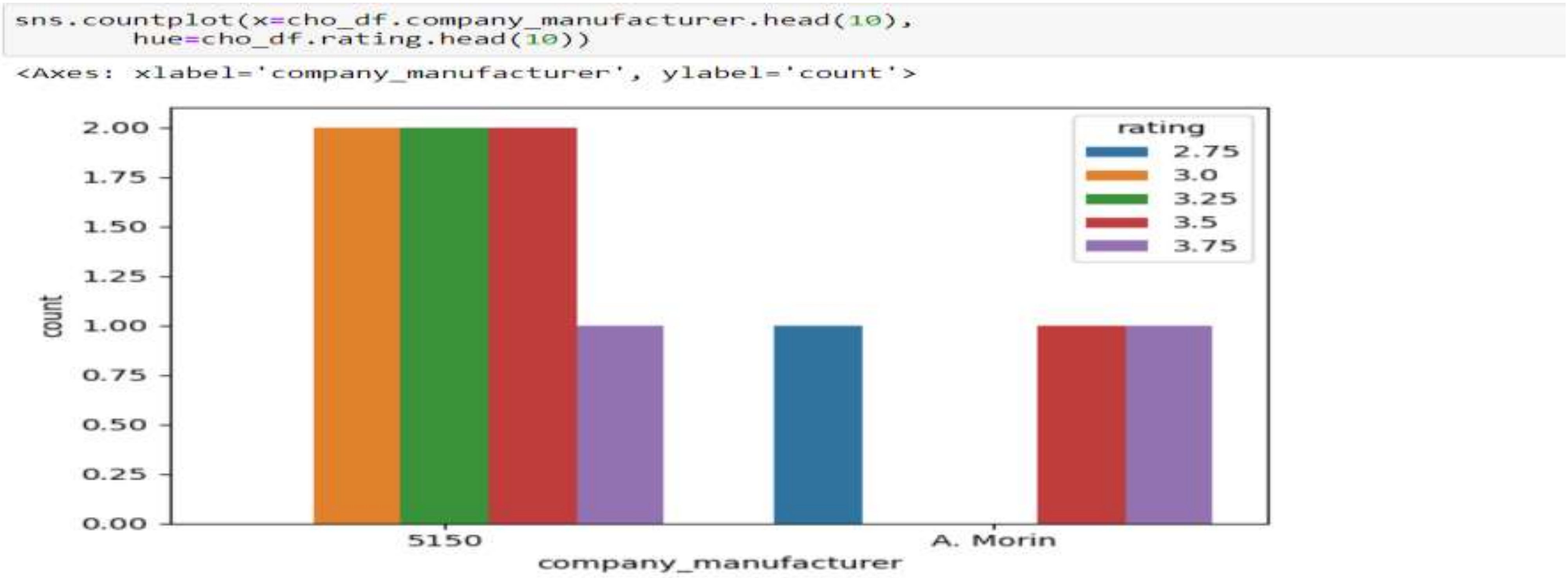
To find the rating distribution based on the company_manufacturer column



```
cho_df.groupby(["company_manufacturer"])["rating"].mean().head(10)
```

```
company_manufacturer
5150                                3.321429
A. Morin                            3.423077
AMMA                                3.500000
Acalli                              3.562500
Adi aka Fijiana (Easy In Ltd)       3.250000
Aelan                               2.750000
Aequare (Gianduja)                  2.875000
Ah Cacao                            3.000000
Akesson's (Pralus)                  3.083333
Alain Ducasse                       2.833333
Name: rating, dtype: float64
```

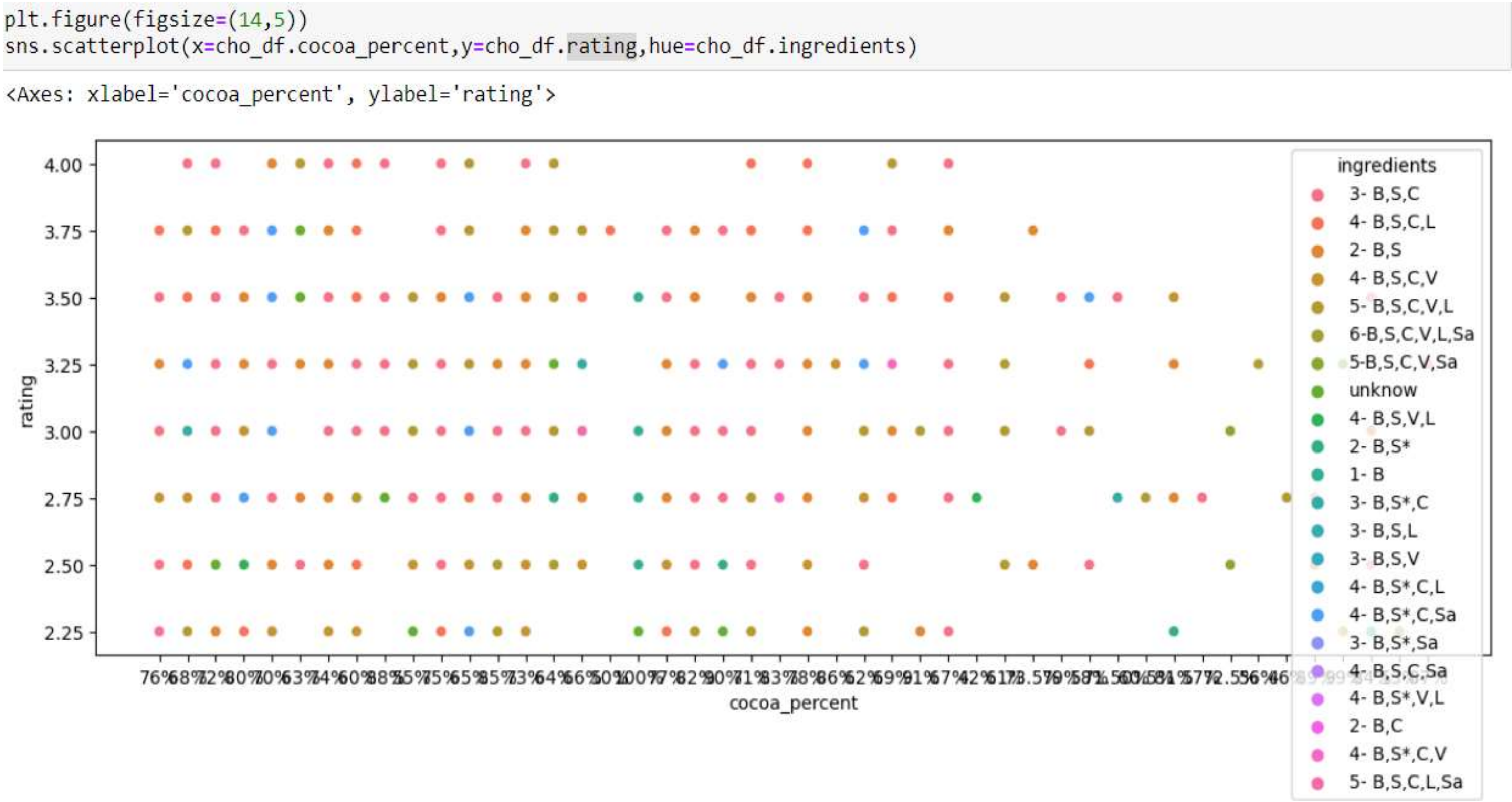
To find the top 10 company manufacturers and their corresponding average ratings



```
sns.countplot(x=cho_df.company_manufacturer.head(10),
       hue=cho_df.rating.head(10))
```

```
<Axes: xlabel='company_manufacturer', ylabel='count'>
```

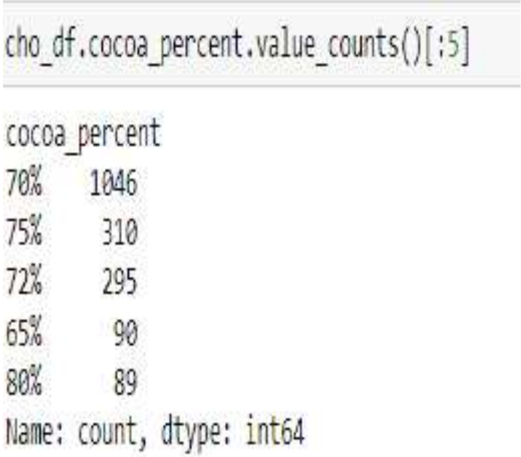
cocoa_percent vs rating



```
plt.figure(figsize=(14,5))
sns.scatterplot(x=cho_df.cocoa_percent,y=cho_df.rating,hue=cho_df.ingredients)
```

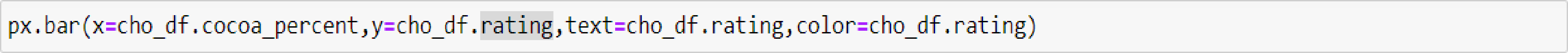
```
<Axes: xlabel='cocoa_percent', ylabel='rating'>
```

cocoa_percent vs rating

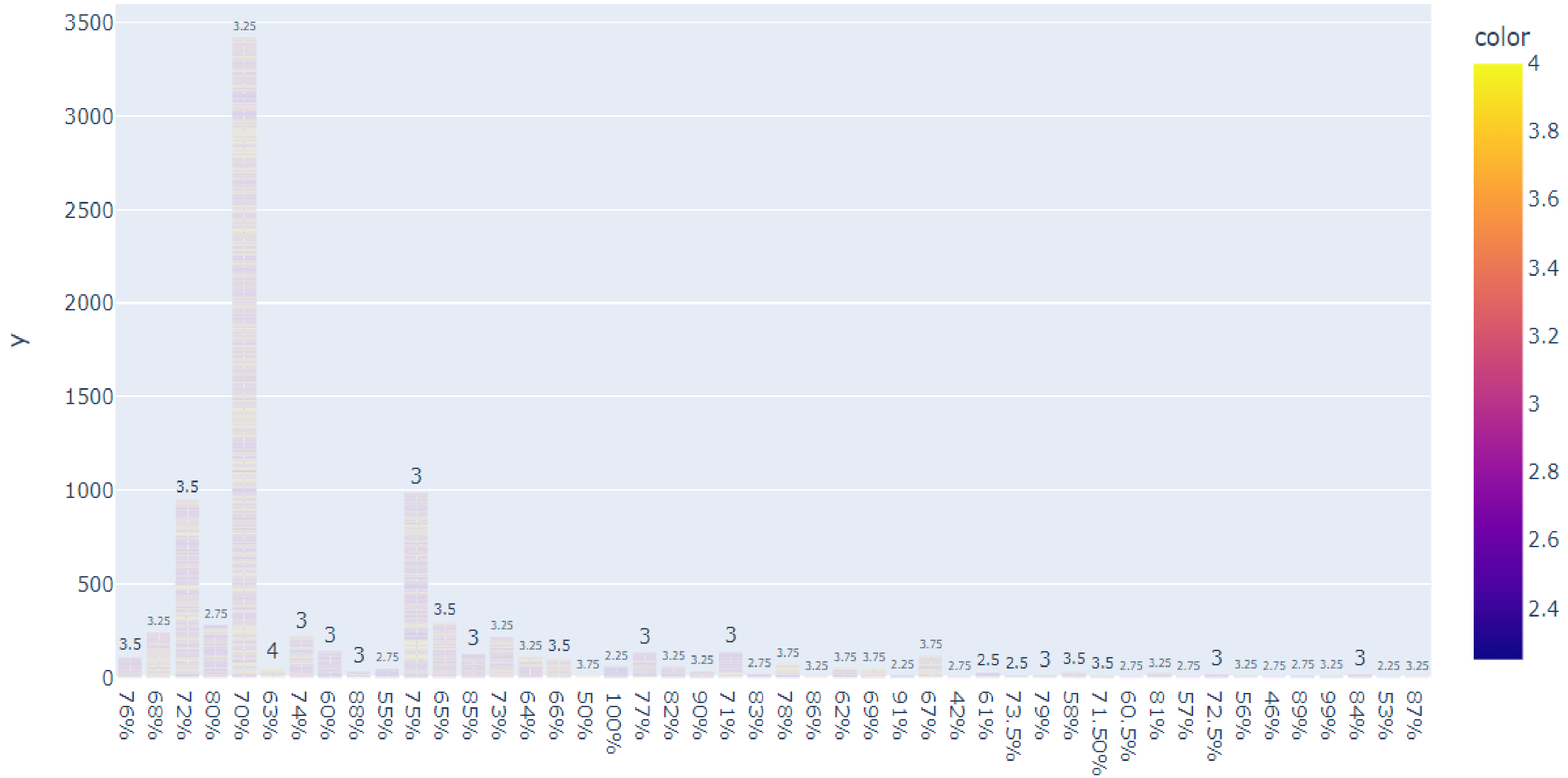


```
cho_df.cocoa_percent.value_counts()[:5]
```

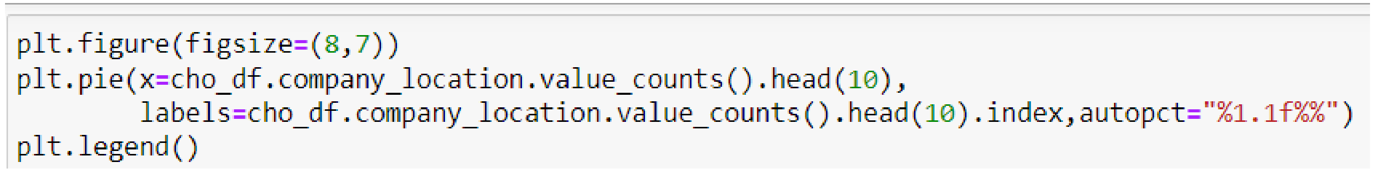
```
cocoa_percent
70%    1046
75%     310
72%     295
65%      90
80%      89
Name: count, dtype: int64
```



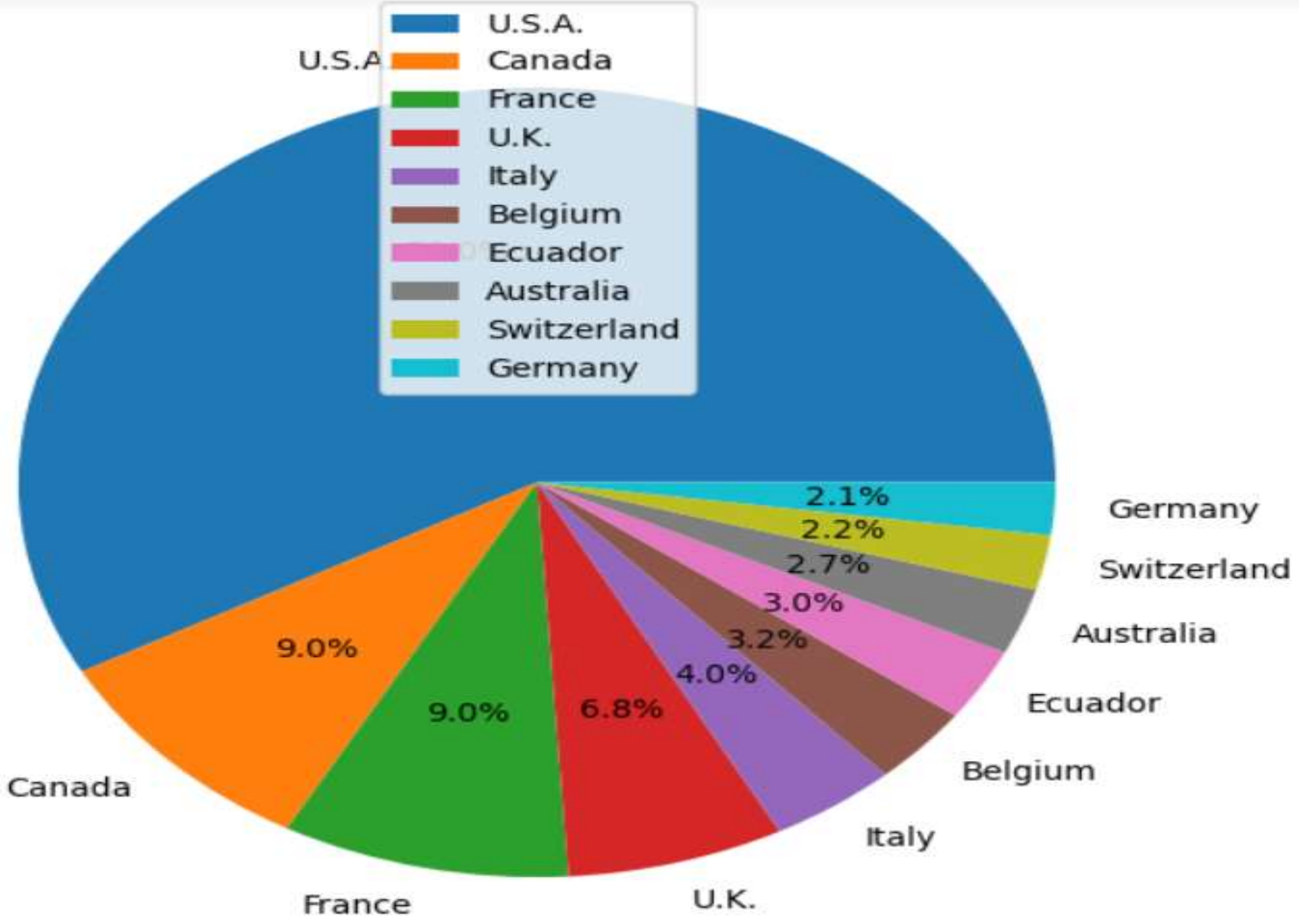
```
px.bar(x=cho_df.cocoa_percent,y=cho_df.rating,text=cho_df.rating,color=cho_df.rating)
```



Top 10 company_location

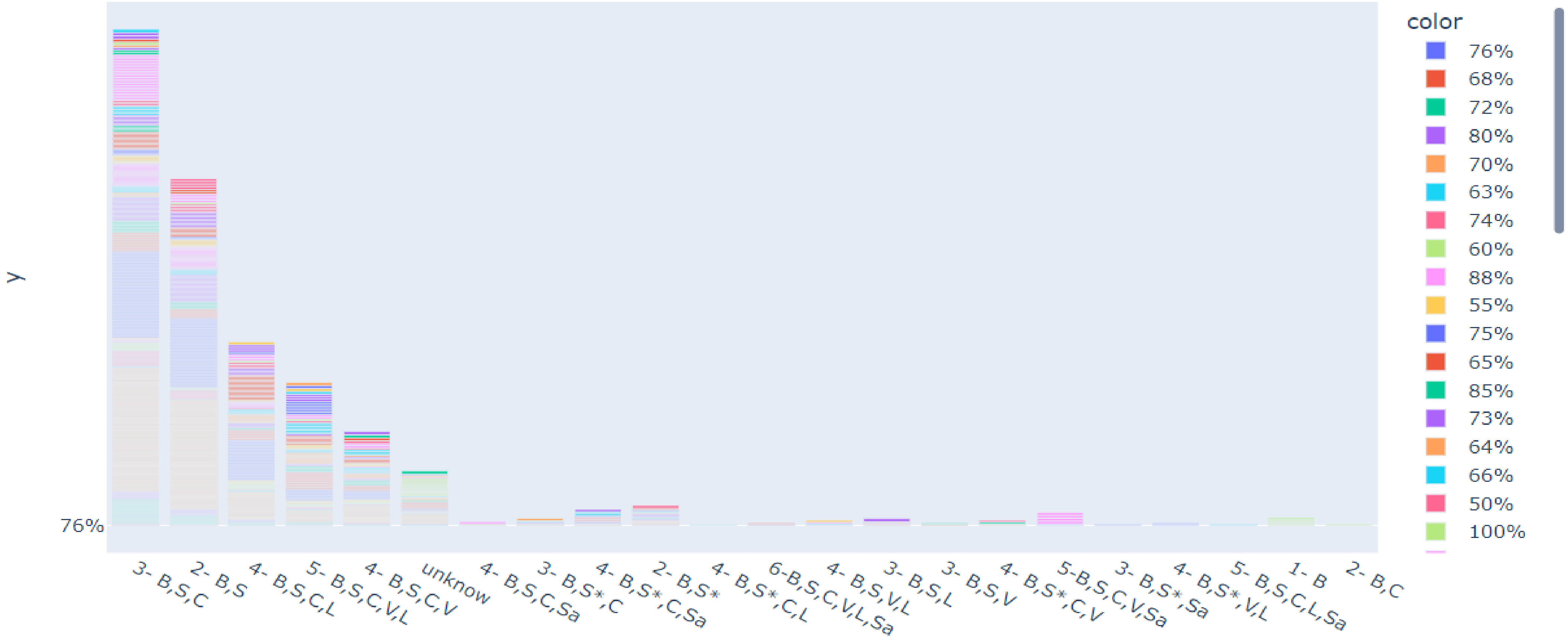


```
plt.figure(figsize=(8,7))
plt.pie(x=cho_df.company_location.value_counts().head(10),
        labels=cho_df.company_location.value_counts().head(10).index,autopct="%1.1f%%")
plt.legend()
```

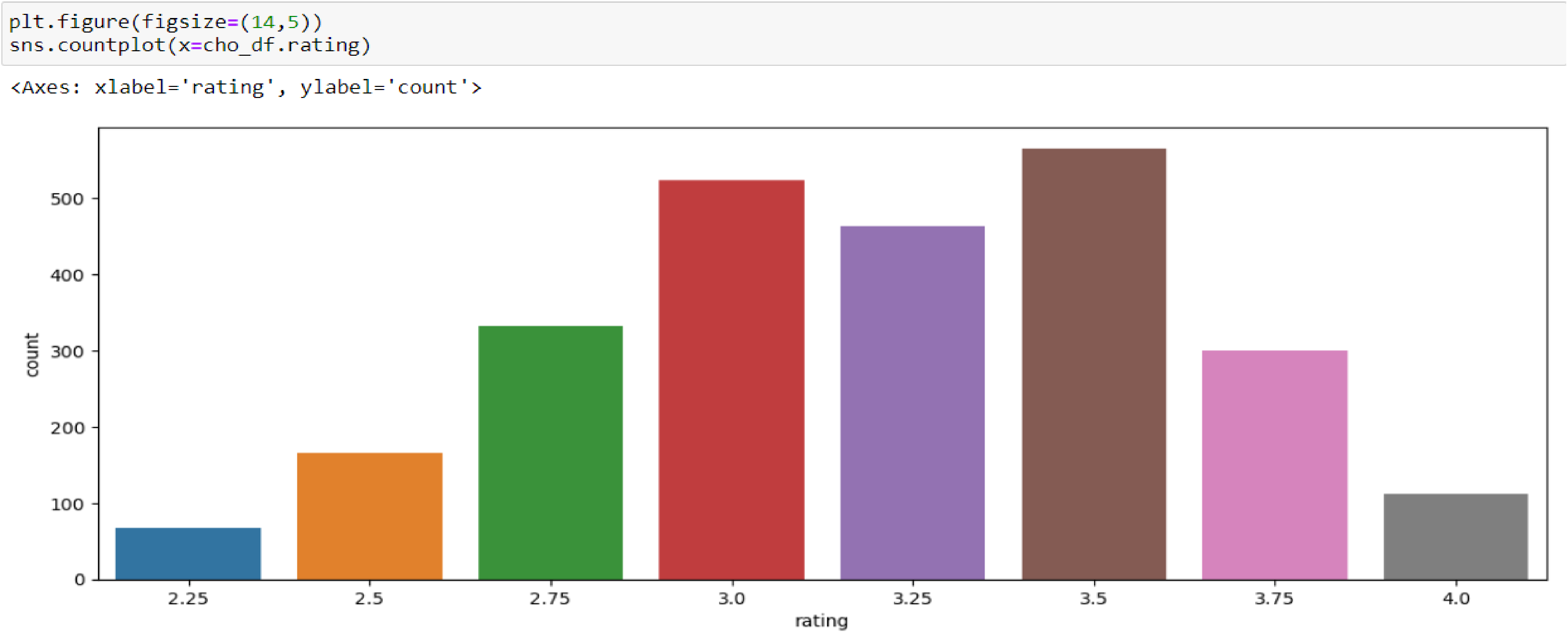


ingredients vs cocoa_percent

```
px.bar(x=cho_df.ingredients,y=cho_df.cocoa_percent,color=cho_df.cocoa_percent)
```



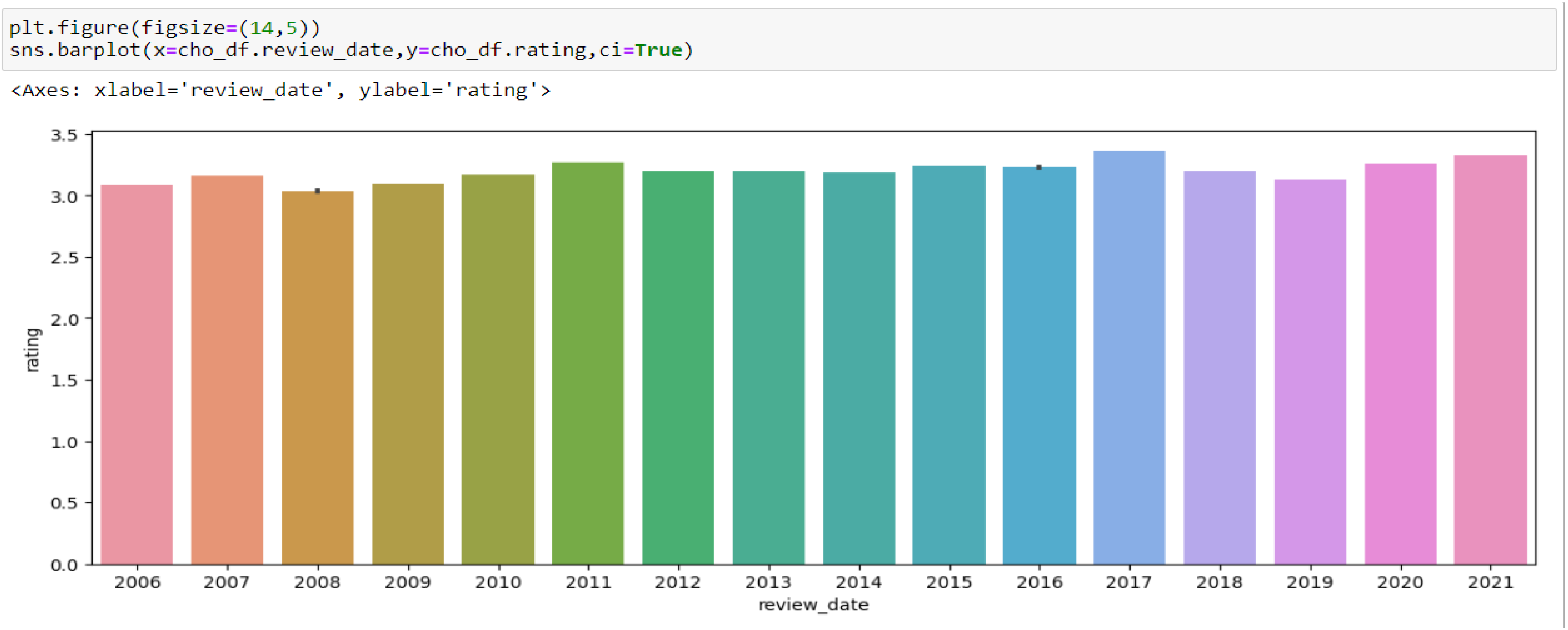
Rating



```
plt.figure(figsize=(14,5))
sns.countplot(x=cho_df.rating)
```

```
<Axes: xlabel='rating', ylabel='count'>
```

review_date vs rating

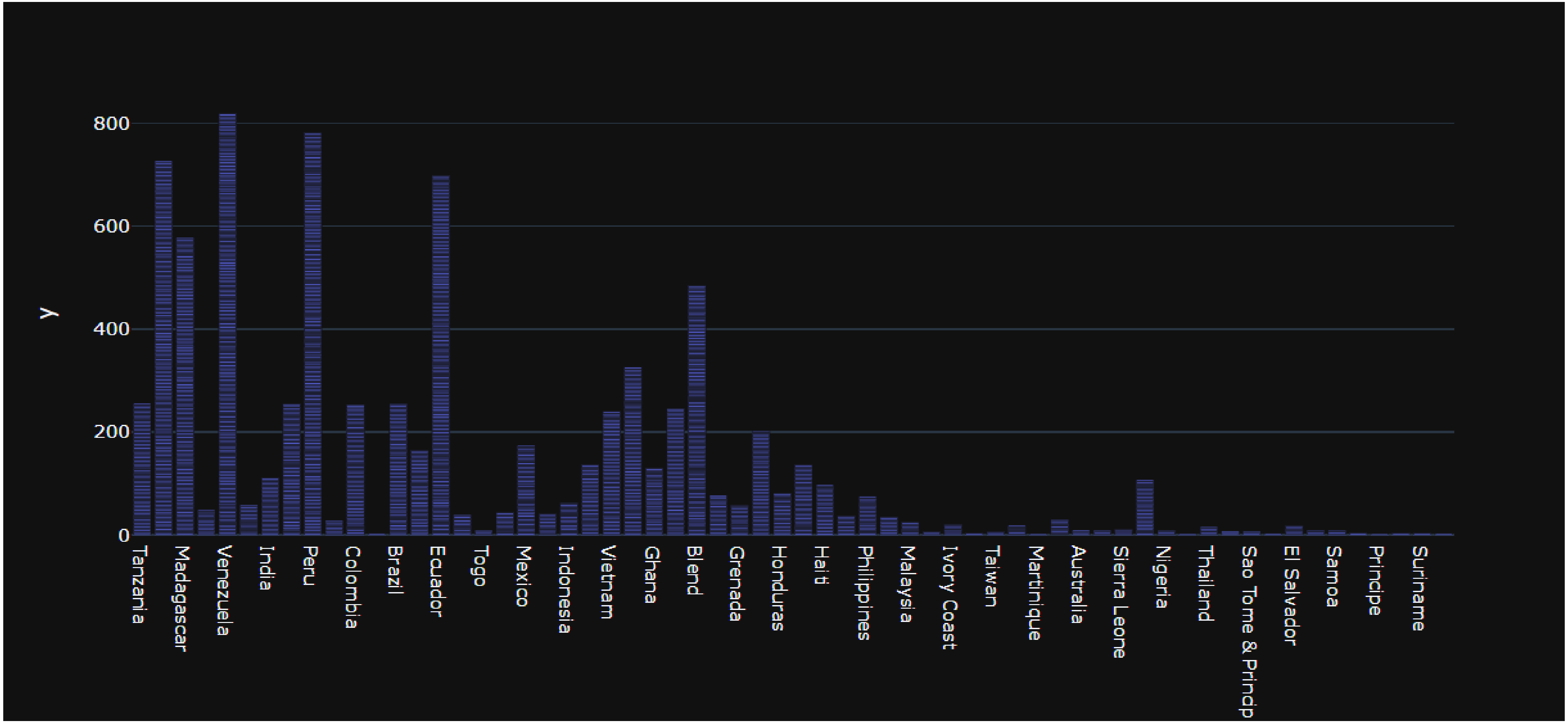


```
plt.figure(figsize=(14,5))
sns.barplot(x=cho_df.review_date,y=cho_df.rating,ci=True)
```

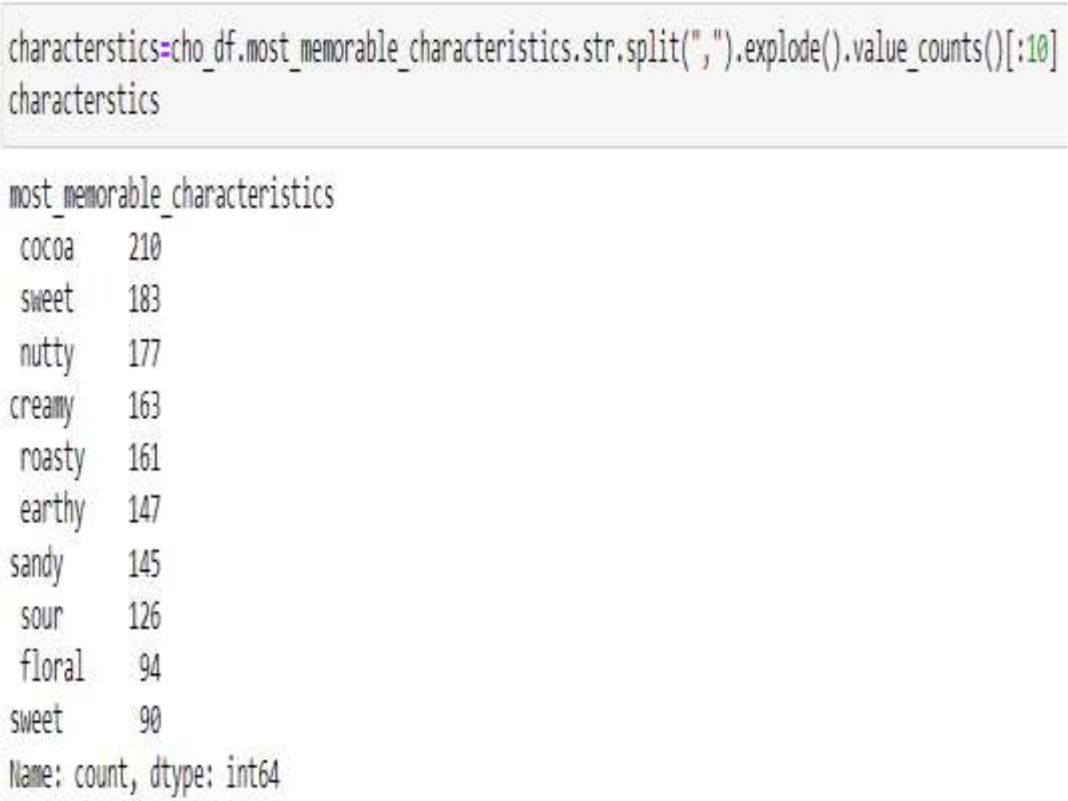
```
<Axes: xlabel='review_date', ylabel='rating'>
```

country_of_bean_origin vs rating

```
px.bar(x=cho_df.country_of_bean_origin,y=cho_df.rating,template="plotly_dark")
```

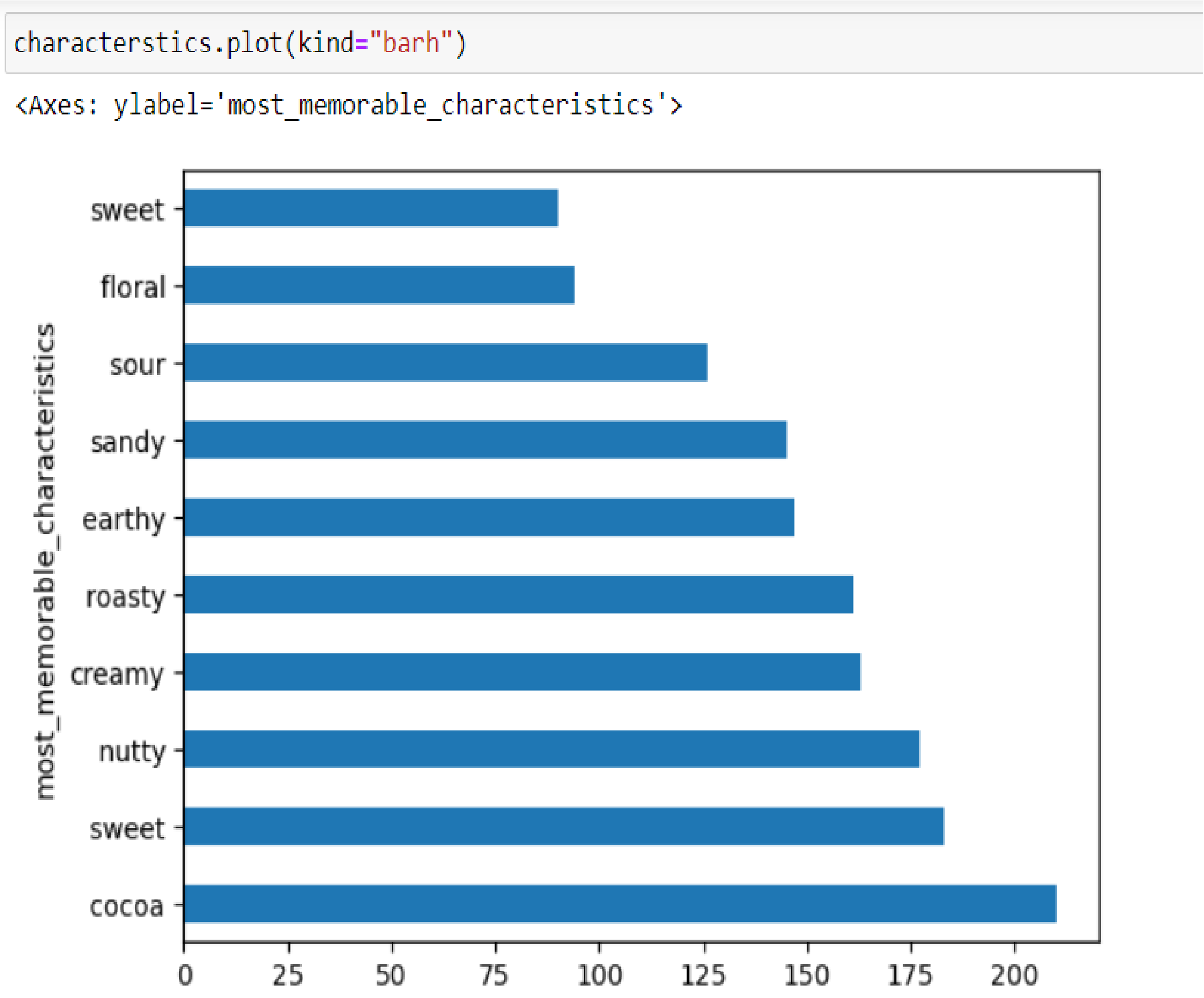


Investigate whether certain types of content are more popular



```
characterstics=cho_df.most_memorable_characteristics.str.split(",").explode().value_counts()[:10]
characterstics
```

```
most_memorable_characteristics
 cocoa     210
 sweet     183
 nutty     177
creamy     163
 roasty    161
 earthy    147
sandy      145
 sour      126
 floral     94
sweet       90
Name: count, dtype: int64
```



```
characterstics.plot(kind="barh")
```

```
<Axes: ylabel='most_memorable_characteristics'>
```

Conclusion:

1. **Cocoa Percentage vs. Rating (Scatter Plot):** The scatter plot analysis revealed a positive correlation between cocoa percentage and rating. Higher cocoa percentages tend to result in higher ratings, indicating a preference among consumers for chocolates with a richer cocoa content.
2. **Top 10 Company Locations (Pie Plot):** The pie plot illustrated the distribution of chocolate manufacturers across the top 10 company locations. This insight provides a glimpse into the geographical concentration of chocolate production, highlighting regions with significant industry presence.
3. **Ingredients vs. Cocoa Percentage (Bar Plot):** The bar plot analysis depicted the relationship between ingredients and cocoa percentage in chocolate products. It showed how different ingredients are utilized across various cocoa percentage ranges, providing insights into formulation strategies employed by manufacturers.
4. **Rating Distribution (Bar Plot):** The bar plot of ratings displayed the distribution of ratings across chocolate products. This insight helps manufacturers gauge the overall performance of their products in the market and identify areas for improvement.
5. **Country of Bean Origin vs. Rating (Bar Plot):** The bar plot analysis showcased the relationship between the country of bean origin and ratings of chocolate products. It revealed potential preferences among consumers for chocolates made from beans sourced from certain regions, influencing product ratings.

Recommendations for Company/Manufacturer Owners:

1. **Optimize Cocoa Percentage:** Based on the positive correlation between cocoa percentage and rating, consider offering a range of chocolate products with varying cocoa percentages to cater to diverse consumer preferences.
2. **Geographical Expansion:** Utilize insights from the top 10 company locations analysis to identify regions with growth potential and consider expanding operations or distribution networks in these areas.
3. **Ingredient Selection:** Leverage insights from the ingredients vs. cocoa percentage analysis to formulate products with high-quality ingredients that complement different cocoa percentages, ensuring superior taste and texture.
4. **Continuous Improvement:** Monitor rating distributions and consumer feedback regularly to identify areas for product enhancement and innovation. Invest in research and development to stay competitive in the market.

Recommendations for Chocolate Users:

1. **Explore Cocoa Varieties:** Experiment with chocolate products featuring different cocoa percentages to discover personal preferences in terms of flavor intensity and richness.
2. **Consider Origin:** Pay attention to the country of bean origin when selecting chocolate products, as this may influence flavor profiles and overall quality. Support manufacturers that prioritize transparent sourcing practices and ethical production.
3. **Check Ingredients:** Review ingredient lists to ensure transparency and choose chocolates made from high-quality ingredients for an enhanced taste experience.
4. **Provide Feedback:** Share feedback and ratings for chocolate products to help manufacturers improve their offerings and meet consumer expectations. Engage with brands that value consumer input and demonstrate a commitment to quality and innovation.

By implementing these recommendations, both chocolate manufacturers and users can contribute to the success and sustainability of the chocolate industry, fostering a culture of quality, transparency, and consumer satisfaction.