Assignment 5

Multiple Views

IAT 355 - Introduction to Visual Analytics

Samantha Chung, 301426865, D100
Elina Mokhammad, 301416708, D100

Observable Notebook

Link to the Notebook: **https://observablehq.com/d/ac8f1f9714f7840e**

Scope

Our research aims to find the unique factors influencing coffee bean ratings, focusing on key elements such as origin, harvest method, flavour profile, and pricing.

Audience

Tailored for professionals in product research, operations, and marketing, our platform addresses individuals with diverse data literacy levels, offering advanced data analysis tools and providing factual information about coffee beans in an accessible format. Users benefit from a versatile solution that empowers them to interpret datasets, extract insights, and navigate the complexities of the coffee bean industry with ease, ensuring comprehensive access to both advanced analytics and foundational industry knowledge.

Describe Datasets

a. The first dataset, "simplified_coffee@4.csv," encompasses a collection of coffee products described by their names, roasters, roast levels, location countries, origins, and pricing per 100 grams (100g_USD). It also includes a quality rating and review date for each coffee, accompanied by a textual review. These dimensions are crucial for graphically analyzing trends in coffee quality and price, as well as understanding the geographical distribution and roast preferences in the coffee market.

b. The second dataset, "coffee_df.csv," details coffee reviews with a focus on geographical and sensory attributes. Each record includes the review's location, coffee origin, roast level, estimated price, review date, and Agtron numbers (which measure roast degree). Sensory ratings for aroma, acidity, body, flavor, aftertaste, and performance with milk are also provided. Descriptive text fields offer insights into coffee profiles and production details. The dataset's comprehensive sensory and descriptive dimensions enable rich visualizations of coffee characteristics and their relationships to consumer ratings and pricing.

c.The third dataset, "arabica_data_cleaned.csv," is an extensive compilation of Arabica coffee bean evaluations, featuring 44 dimensions including owner, country of origin, farm name, altitude, and various aspects of coffee quality such as aroma, flavor, acidity, body, and cupper points. Importantly, it covers the bean's processing method and the presence of defects, which are critical for assessing coffee quality. These attributes are essential for crafting nuanced visualizations that explore the relationships between coffee bean characteristics, their cultivation environment, and the resulting quality ratings.

User testing

When conducting user testing with our peers, we encountered a readability issue on one of our scatterplot graphs that plots ratings against prices. The difficulty arose when applying filters to isolate data points from a single country of origin; the points became challenging to distinguish. To overcome this, we adjusted the graph's scale and modified the opacity—diminishing it for non-selected points and amplifying it for the selected ones. Additionally, we enlarged the points in response to feedback that it would enhance the graph's clarity. For color representation, we included a legend delineating which colors correspond to which countries, coupled with a tooltip feature to further aid in user understanding.

We also encountered a visual discernibility problem with the choropleth map intended to display the distribution of coffee bean prices by country. The use of a red color gradient to represent coffee prices draws undue attention to Australia. Its placement at the edge of the map, combined with a lack of surrounding country outlines, makes it appear disproportionately larger. This visual prominence overshadows England, which, despite having the highest coffee bean prices and being depicted by a more saturated shade, is less visually striking due to its smaller size and the presence of surrounding countries with varying colors. To address this issue, we shifted the color scheme to shades of blue, leveraging the insight from our lecture that people can distinguish more shades of blue than other colors.

Questions

Visualization 1: Is there a relationship between the rating of coffee beans and their price, with a focus on whether beans with higher ratings tend to be priced higher? How does the geographical location of coffee bean producers impact the pricing of coffee beans?

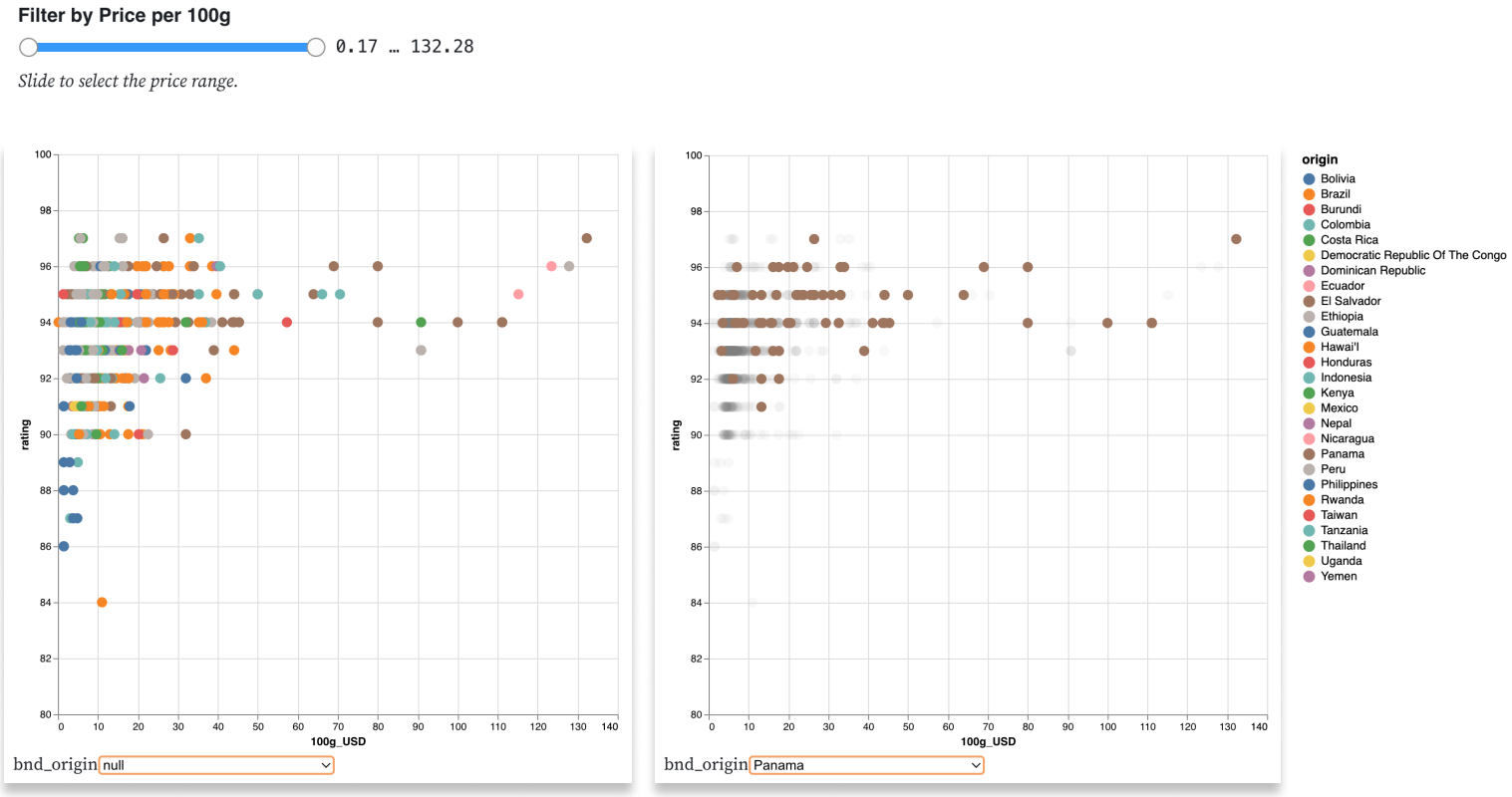
Visualization 2: How does the Agtron value, which represents the roast level, affect the flavor profile of coffee beans? What are the most frequently used flavor descriptors in highly rated coffee beans? Can we identify and visualize these common flavor descriptors in a pie chart?
Visualization 3: What factors influence the rating of high-quality coffee beans, specifically examining their correlation with roast level, origin, harvesting method, and the location of the coffee bean producer?

Demo Video

Video Link: **https://youtu.be/WN0-p0w-uc4**

Visualization 1

coffee beans price and rating



Interaction

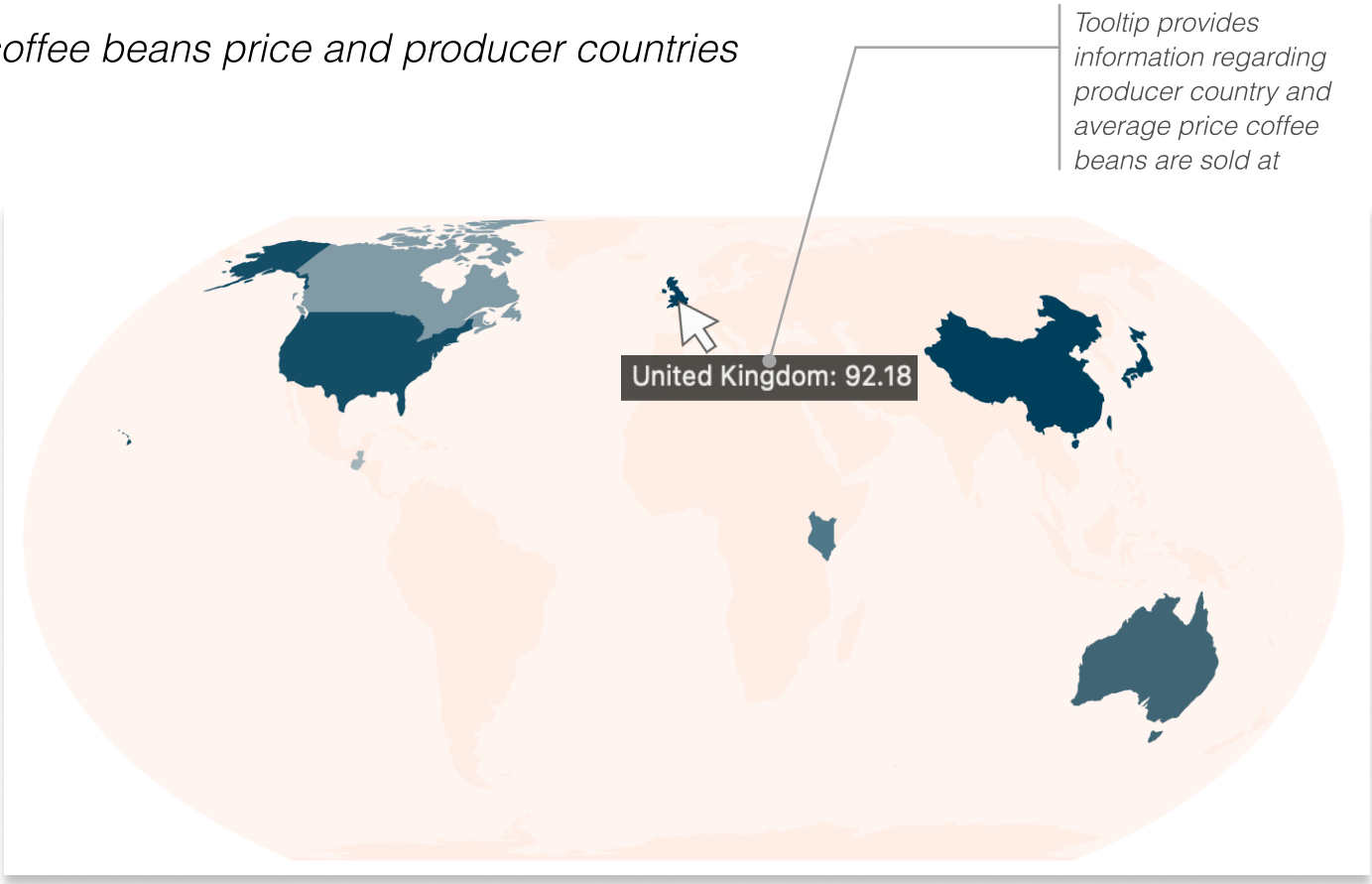
This scatterplot provides a user-friendly tool for professionals in the coffee industry to scrutinize and evaluate market data. The "Filter by Price per 100g" slider empowers users to refine the dataset within a range of prices, resulting in a focused view that highlights the relationship between the cost of coffee and its quality rating. As the slider is adjusted, the scatterplot updates in real time, showcasing only the data points that fall within the specified price bracket.

Visually, the data points are colored according to their country of origin, offering an instant, comparative view of how ratings and prices are distributed globally. This color-coding becomes particularly effective when a specific country is selected from the dropdown menu; the chosen country's data points remain brightly colored while others fade into the background with reduced opacity. This contrast facilitates a direct comparison and a more straightforward interpretation of how a country's coffee ranks in the global market.

The tooltip feature further adds to the scatterplot's utility by revealing detailed information for each data point upon mouseover, including price, rating, and origin. This function is crucial for buyers and sellers in the coffee bean industry, where swift access to detailed data can inform purchasing decisions, pricing strategies, and market positioning.

Visualization 1

coffee beans price and producer countries



Interaction

This visualization demonstrates the range of prices at which coffee beans are sold by different countries, highlighting a clear disparity where Western countries, notably England, the US, and Canada, exhibit the highest prices. This contrast is visually captured through a color-coded map, where deeper shades of blue signify higher prices, making England the most prominent. Conversely, Kenya is represented with the lightest shade, indicating the lowest prices. The distribution of colors across the map provides an immediate sense of the economic factors at play, suggesting that a country's level of development may influence its coffee pricing.

Interactive features of the map, such as tooltips that display precise prices per 100 grams upon hovering, render the data more accessible and understandable. Where prices are not available, 'N/A' appears, ensuring clarity in data presentation. The gradient of blue tones to represent the price scale, assists in interpreting the map. Such tools are invaluable for students and professionals analyzing the coffee market, offering a clear, concise representation of pricing strategies in relation to economic status.

Visualization 1

Findings

The visualizations highlight significant disparities in the prices at which coffee beans are sold by different countries. England, the United States, and Canada are shown to sell coffee beans at the highest prices, which might be attributed to their higher levels of economic development and welfare standards. In contrast, Kenya is depicted as the lowest-priced seller, indicating a potential link between a country's economic status and its coffee pricing strategies.

The analysis indicates that coffee prices are influenced by more than just bean quality. Factors like a nation's economic status and geographic positioning play significant roles. For example, higher prices in economically developed Western countries contrast with lower prices in markets like Kenya. This variance in pricing, irrespective of quality and origin, points to a complex interaction between global economic forces and local market conditions in determining coffee bean prices.

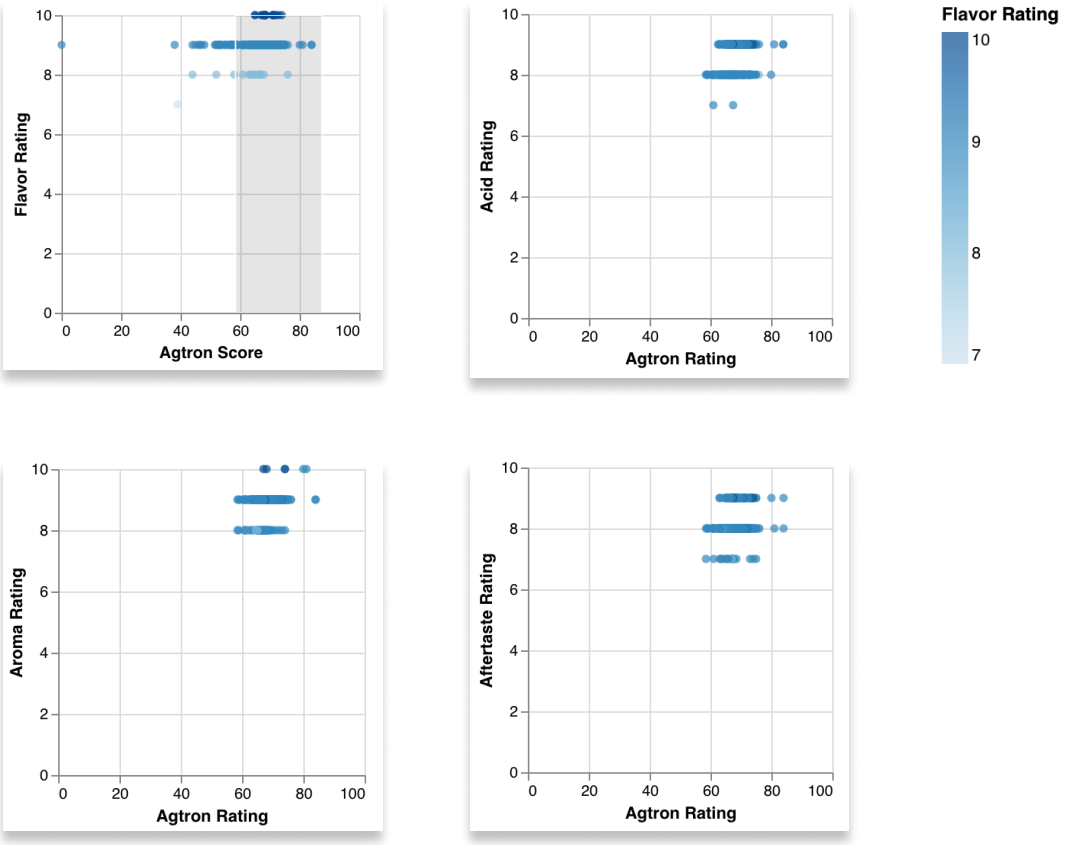
These findings are crucial for professionals in the coffee industry, particularly those involved in data analysis, buying, and selling. They suggest that successful pricing strategies should extend beyond mere quality assessment to include a more holistic understanding of market dynamics, regional economic conditions, and consumer preferences. This approach is vital for developing effective pricing models and competitive strategies.

Errors and Notes

We conducted a detailed data cleaning process to ensure precision and relevance. Key steps included removing null values to prevent analytical errors and excluding entries like Hawai'i and New Taiwan that don't represent sovereign nations, thereby ensuring geographical accuracy. For visual distinction, we developed arrays to assign specific colors to countries based on their coffee pricing, aiding in the interpretability of the data. Furthermore, we utilized a world map JSON file, from which we extracted country names and codes. This step was crucial for aligning our coffee price data with the correct geographical locations on the map, allowing for an accurate and informative overlay. This rigorous approach to data preparation was essential to create clear and reliable visualizations for our analysis of the global coffee market.

Visualization 2

flavour profiles and rating



Interaction

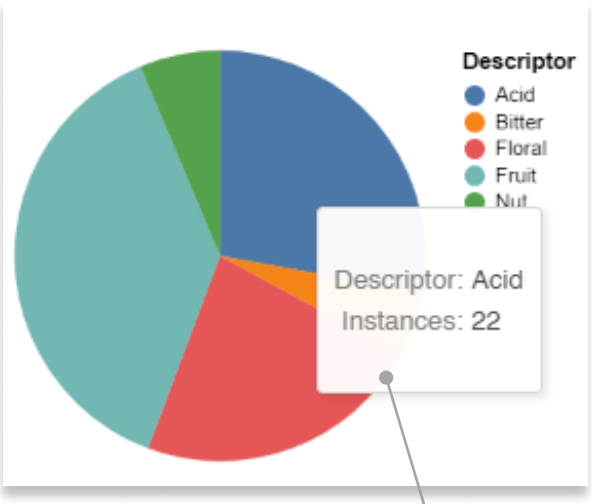
This graph investigates the impact of various roast levels on the overall taste of coffee beans, focusing on unique flavor profiles such as acidity, aroma, and aftertaste. Through a comprehensive analysis of the data, we aim to identify linked characteristics and factors contributing to higher-rated beans.

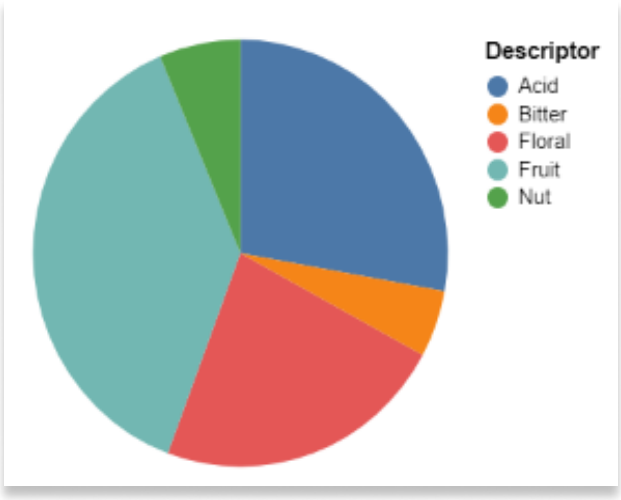
Upon further data refinement, our goal is to delve into the distinct flavor preferences of consumers for top-rated beans. By isolating and filtering out the top-performing beans, we can extract five descriptors that capture the preferred flavors. These descriptors are selected based on a range of flavors that coffee beans can provide, as outlined in https://www.coffeeandhealth.org/coffee-and-the-senses/aroma-and-flavour-descriptors.

The interactive feature enables users to filter and explore specific data, observing how different elements interact through both filtering and brushing functions. When focusing on flavor ratings, users can dynamically visualize the corresponding changes in acid, aroma, and aftertaste ratings. This interactive approach facilitates a nuanced exploration of the relationships and variations within the dataset, offering a comprehensive understanding of how these key factors evolve in tandem.

Visualization 2

flavour profiles and rating





The tooltip on the pie chart displays the frequency of each specific flavor descriptor, offering insights into the prevalence of these descriptors.

Findings

Our findings reveal a noteworthy correlation between high Agtron scores, signifying darker roasted beans, and a stronger aroma rating, coupled with a lower aftertaste rating. The acidity level closely aligns with the flavor rating. Additionally, the highest-rated beans predominantly exhibit a fruity flavor, with acidity and floral notes serving as secondary characteristics. Smaller nuances of nuttiness and bitterness are also present.

This refined analysis sets the stage for a clearer understanding of consumer preferences, shedding light on the specific flavors that drive the popularity of top-rated coffee beans in the market.

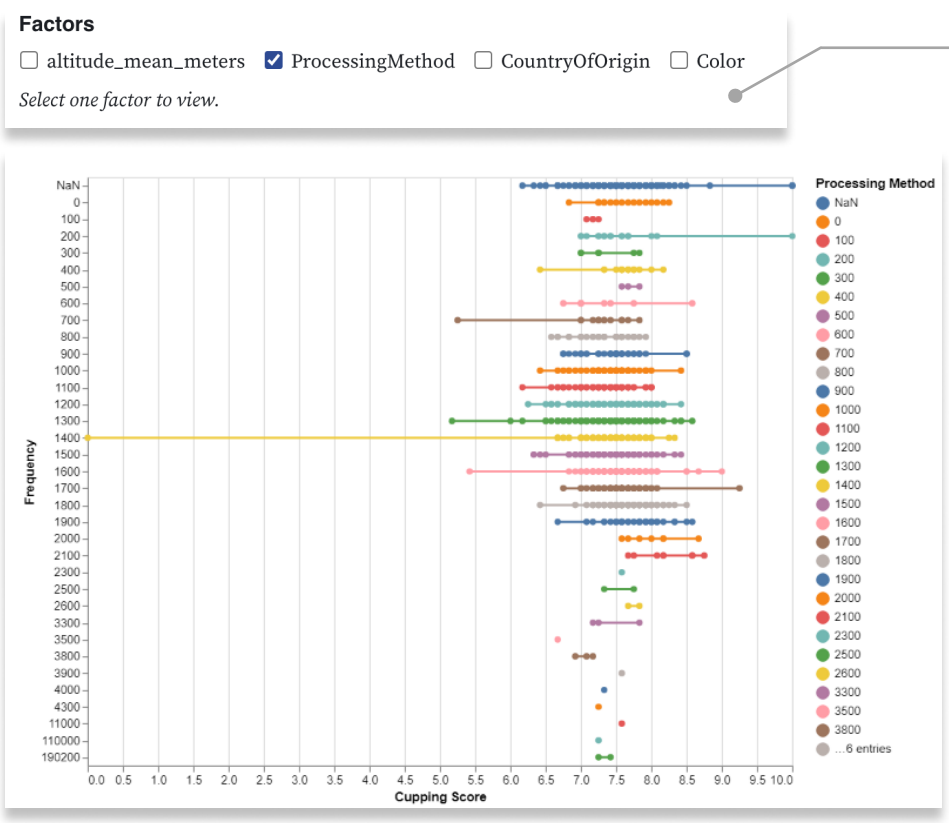
Errors and Notes

To enhance the clarity and comprehension of this dataset, we systematically addressed any instances of null values. This meticulous refinement process aimed to minimize interference and ensure a more coherent analysis.

Navigating the dataset presented a challenge initially, necessitating a thorough exploration of its dimensions to decipher the meaning of each variable. An illustrative example of this process involves the Agtron value, where dedicated research was essential to comprehend its significance and purpose within the context of the dataset. This commitment to understanding each dimension has been crucial in extracting meaningful insights from the data.

Visualization 3

origin country rating and frequency



The four filter options allow sorting by different criteria, providing diverse perspectives on the rating data (cupping score)

Interaction

The primary objective of this graph is to discern the factors influencing the rating of coffee beans. The key parameters under consideration include cupperscore (rating), altitude of the bean, processing method, the country of origin, and the overall color of the bean.

The interactive functionality allows users to select and filter specific factors, observing distinct clusters of points corresponding to different ratings. This dynamic feature enables an exploration of prevalent characteristics associated with certain ratings, shedding light on factors that contribute to both higher and lower ratings.

By systematically navigating through the filters, users can gain valuable insights into the patterns and trends within the data, facilitating a nuanced understanding of what attributes commonly result in a higher-rated coffee bean.

Visualization 3

origin country rating and frequency

Findings

The analysis reveals that a substantial number of beans consistently receive ratings in the 7-8 point range, regardless of the specific factor under consideration. Notably, green beans are more prevalent, but those with a blue/green hue tend to receive higher ratings. In terms of country of origin, Tanzanian beans consistently achieve higher ratings, while Honduran beans exhibit a broader range of ratings. Guatemalan beans maintain a stable rating around 7.5 points.

Altitude appears to influence ratings significantly. Beans within the altitude range of 1000-1500 meters, particularly around 1400 meters, exhibit a wide range of ratings. Notably, beans grown at altitudes exceeding 1500 meters tend to receive higher ratings on average.

This synthesis of factors provides a comprehensive overview, showcasing the nuanced relationships between various parameters and the resulting cupperscores. The observed patterns contribute valuable insights into the characteristics that commonly lead to higher or lower ratings for coffee beans.

Errors and Notes

In managing our dataset, addressing a notable prevalence of null values and tackling substantial variances was crucial. To enhance the dataset's quality, several measures were implemented. Null values were systematically handled, and data refinement efforts were undertaken to manage the considerable variances.

Rounding altitudes was employed to bring about uniformity in representation, while the processing method strings underwent meticulous splitting and trimming for standardized and clearer data. These actions collectively contributed to a more consistent and reliable dataset.

Despite the initial challenges, the dataset proved to be highly informative, offering a wealth of valuable information. The structured approach to managing null values and refining data has strengthened the dataset's integrity, providing a solid foundation for meaningful analysis and insights.

Further Work

Initially, our research aimed at conducting an in-depth analysis of importing and exporting data to discern the factors influencing the overall price of coffee beans. However, recognizing the dynamic nature of the dataset and the potential for more impactful insights, we strategically shifted our focus towards investigating what influences the overall rating of coffee beans. This decision allowed us to uncover nuanced patterns and relationships between various factors such as origin, altitude, and processing method, providing a more comprehensive understanding of the elements that contribute to the perceived quality of coffee beans in the market.

Our initial intention involved implementing a more interactive map that would allow users to zoom, move, and explore additional attributes of each country in the dataset. However, due to unforeseen technical constraints, we were unable to integrate this feature. While the original presentation effectively conveyed the available data, the envisioned interactive map would have provided an enhanced user experience, allowing for a deeper exploration of country-specific attributes. While improving this feature remains an option for future iterations, the existing representation still offers valuable insights into the relationships between countries and their respective coffee bean attributes.

In the next version, an enhanced approach to processing method analysis could involve a more detailed categorization or even the creation of a separate variable that captures nuanced variations within each method. This could include considering factors such as fermentation duration, drying techniques, or other specific steps involved in the processing. Additionally, improving the interactions of the visualization with a more diverse selection would contribute to a more nuanced and accurate depiction of the relationship between processing methods and bean ratings.

Conclusion

In this research, we conducted a thorough analysis of coffee data to discern the factors influencing bean ratings. Despite challenges such as null values and variances, systematic data refinement was employed. The findings revealed a prevalence of beans in the 7-8 point range, with color, origin, and altitude impacting ratings. Tanzanian beans consistently scored higher, and processing method and altitude emerged as influential factors. This research offers valuable insights into the complex dynamics shaping coffee bean quality and perception, contributing to informed decision-making in the coffee industry.

References

GitHub. (2020). Official repo for the #tidytuesday project. Retrieved from https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-01-07

Kaggle. (2022). Coffee_Data_CoffeeReview. Retrieved from https://www.kaggle.com/datasets/hanifalirsyad/coffee-scrap-coffeereview/versions/2

Kaggle. (2022). Coffee Reviews Dataset. Retrieved from https://www.kaggle.com/datasets/schmoyote/coffee-reviews-dataset?select=coffee_analysis.csv