Does your Taste in Coffee and the Way you Brew it Change as you Age?

Final Project Report

CMPT 353, Fall 2024
Sanjit Mann - 301425396,
Yat Tsun Curtis Pu - 301571702

Introduction

While looking for ideas for our project, we ran across one of James Hoffmann's videos, who is a world-renowned barista. In his video titled, "Surprising And Fascinating Results From The Taste Test," he explains how he conducted a large-scale coffee preference experiment involving 4,000 participants across the U.S. Its aim was to explore different potential preferences for coffee roast levels and processing methods of coffee. After watching this video and confining in ourselves that we were both at least decently enthusiastic about coffee, we decided to take on a question that we both had *"Does your taste in coffee and the way you brew it change as you age?"*

Data Collection and Preparation

Firstly, we had to better define the question we had formulated. What is taste? After having a quick skim through the columns of the data we decided that there were 3 main columns that define taste, "Before today's tasting, which of the following best described what kind of coffee you like?, "How strong do you like your coffee?," and "What roast level of coffee do you prefer?." Having multiple different descriptions for taste was quite tough to manage, so we decided to create an overall score by combining them all together and calling it the Coffee Score™. We assigned numerical values to all the different types of descriptors, giving the sweeter and gentler descriptors a lower score and giving the stronger and bolder descriptors a higher score. Then we added up each individual's scores from the 3 columns and averaged them. That way we know on average which age group likes stronger or sweeter coffees.

For our portion about whether or not your age affects your brewing methods, we did not have too much because the column "How do you brew coffee at home?," gave us a compiled list of all the different methods in one column, making it much easier to extract.

Thankfully, the dataset in its entirety was provided for free in the description of the video, albeit with a couple issues. Although not disclosed, we believe that this survey was conducted through some sort of online form service such as Google or Microsoft Forms where not much care was given to whether certain fields were left blank or not, leaving numerous NULL values. This then became our second step, to remove missing values in the relevant columns and aggregating them for analysis.

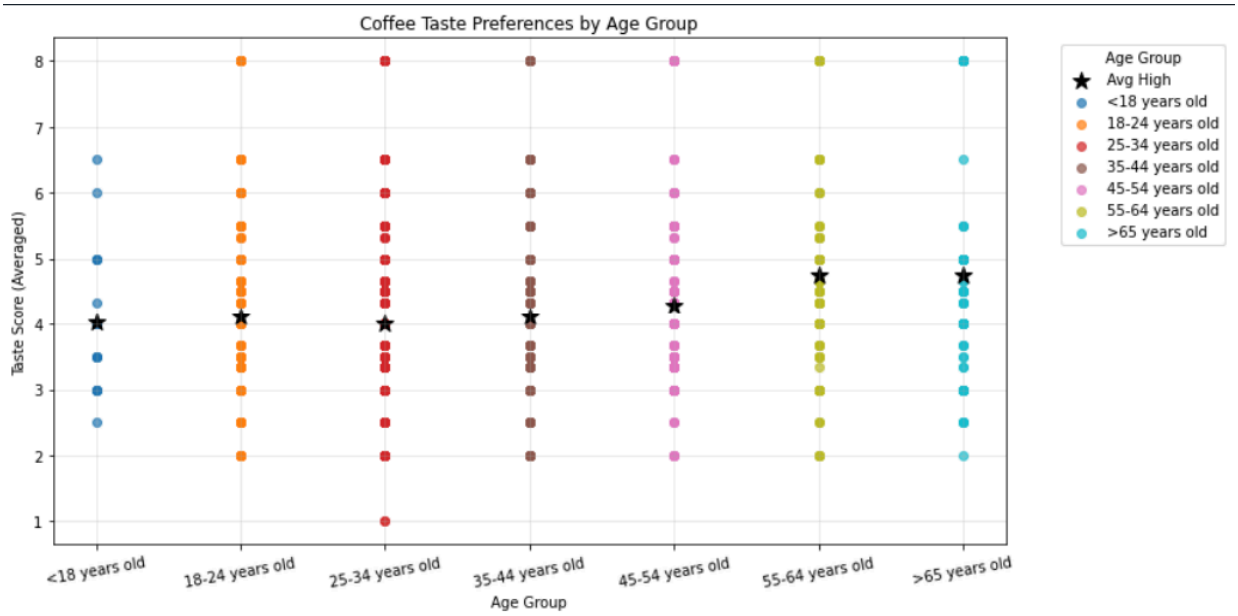
Analysis Techniques Used

We used a few different methods in our analysis of the data. To check whether there is a significant statistical connection between our coffee score and age groups we used Chi-Tests as well as machine learning to predict whether or not these trends would continue if we introduced new values. We mostly used Chi-test and the graphical analysis for our testing of the way you brew coffee vs your age. We heavily took advantage of many of the libraries discussed in class

such as numpy, pandas, seaborn, matplotlib, and sklearn. Numpy was used for things like generating evenly spaced color gradients for visualizations and averaging our CoffeeScores. We used Pandas for its data manipulation capabilities, particularly in working with DataFrames. Seaborn and matplotlib were used to create clear and visually appealing graphs. Sklearn was used for its machine learning capabilities but more specifically to use the RandomForest model. We also used many different types of graphs to visualize this data such as heatmaps, scatterplots, and histograms.

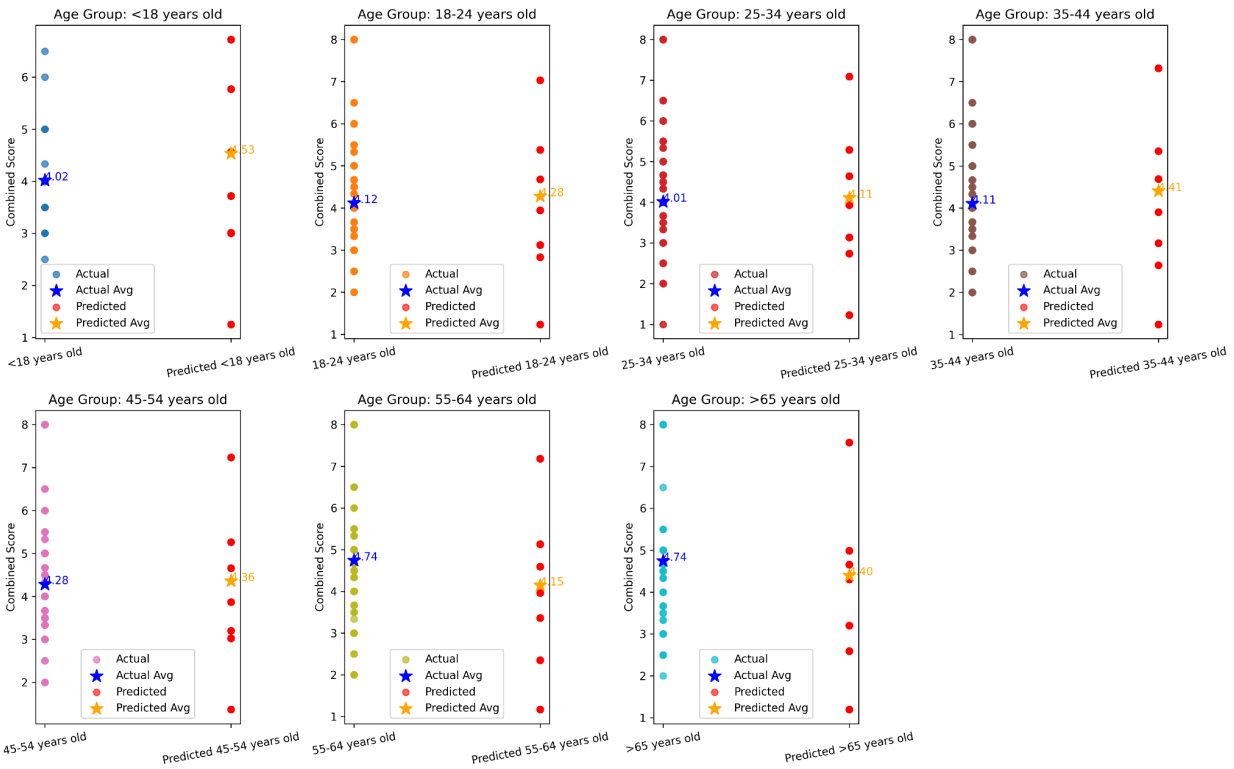
Results, Findings and Conclusions

Due to not knowing the underlying distribution of our data, we decided to run a Chi-test. After running the initial Chi-square test on our CoffeeScore vs the age group, we got a p-value of 2.3014375347269343e-22 which tells us that there is a significant relationship (p-value < 0.05) between age groups and coffee preference.



As you can see the scatter plot above supports our Chi-square test conclusion. Although it is a little difficult to see, the star, which represents the average coffee score among each group, is different for each group. With people that are in the >18 and 25-34 year category having overall sweeter tastes in coffee than those in the 55+ categories.

This made us curious about the results and we decided to introduce machine learning into trying to help us answer this question. We decided to pick a machine learning model, in our case we picked RandomForest, and trained it on the CoffeeScore data that we collected to see whether or not the trend continued for each age group if we provided it with 200 new random values. Unsurprisingly, the trend did continue.

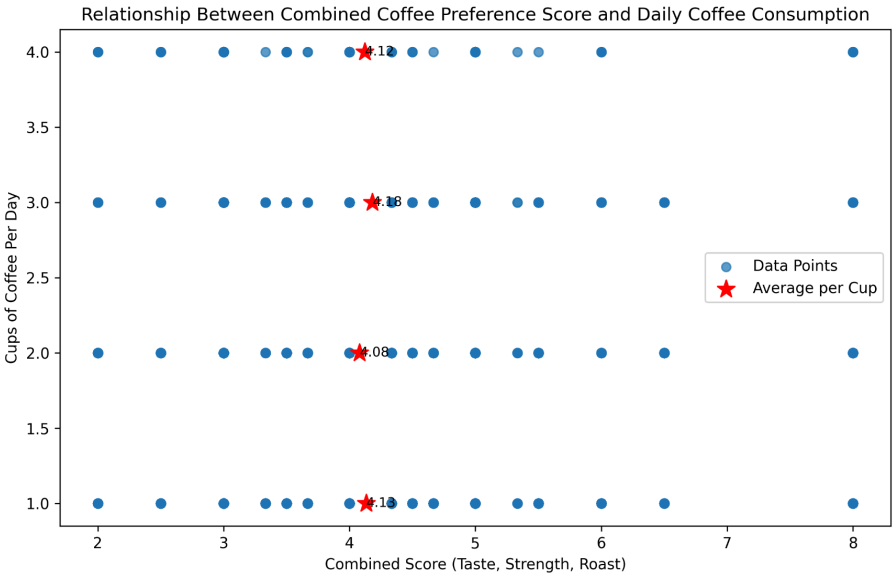


The graph above is the results that we got by using the RandomForest model. The training score that our model achieved was about 83% making it a fairly decent predictor. On the left side of each window are the actual values and averages that are carried over from the previous graph and on the right side of the window are the values and averages predicted with 200 random values for each age group, using the RandomForest model. As you can see visually, the trends more or less remain the same for each age group. We believe that the model would have given an even more accurate prediction if we had more precisely calculated the number of random values given. For example, there are less than 200 responses for the >65 year old category so it does not have as much information to train on as the other groups and that is why the predicted average is much further off. Overall, this machine learning test also confirmed our initial Chi-test as well as our scatterplot.

In the terms of our experiment and the question we sent out to answer, these results do show us that as you age your taste in coffee typically changes. Although we would like to mention that there could be other factors at play here such as differing societal coffee tastes at different periods of time. For example, someone who is in their 50s or 60s might not necessarily like stronger coffee. It's just the fact that when they started drinking coffee many years ago, coffee was just brewed stronger than it is now and those habits that they made then never changed.

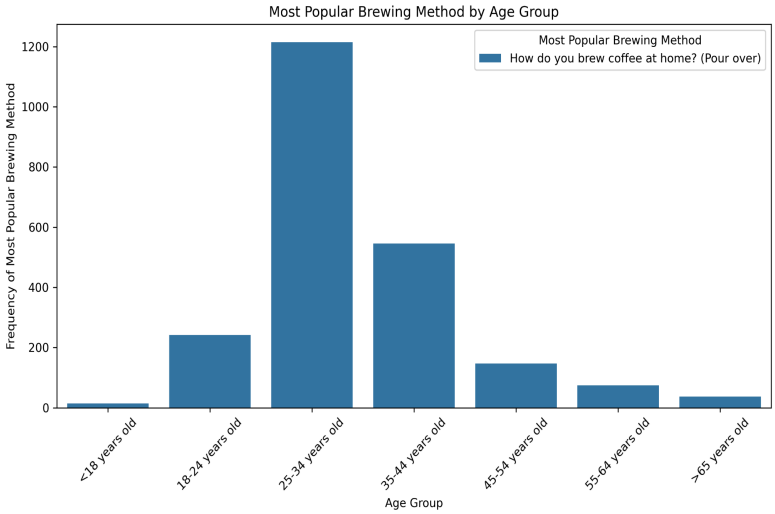
After reaching a conclusion about this we decided to go on a bit of an aside based on the information that we had just learned and try to answer the question, *"Do people who on average drink stronger coffee, also drink less coffee?"*

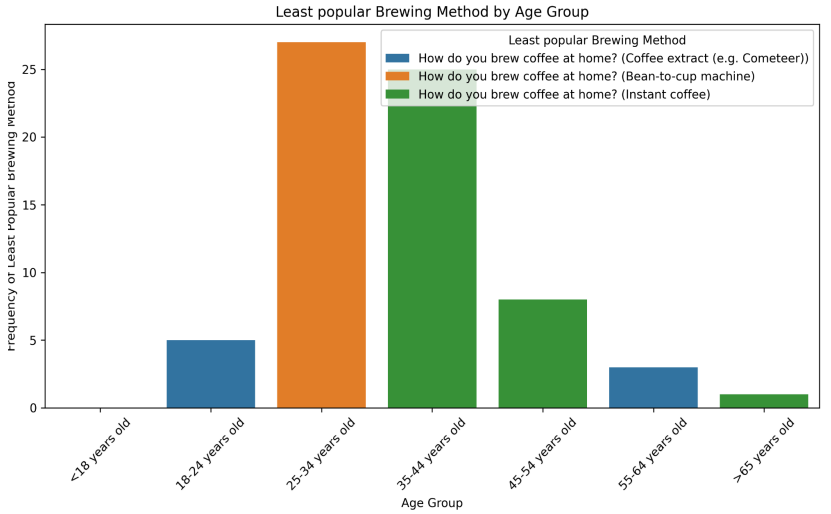
To do this we created another scatter plot comparing the cups of coffee a day vs the coffee score of each individual. Visually you see in the graph that there is a very little difference in a person's coffee score and the amount of coffee they drink per day. That said, there were still 6 responses where the coffee score is 8 (the strongest taste score), and the respondents reported drinking 4 cups of coffee per day but 29 responses where the coffee score is 8, and the respondents reported drinking 1 cup of coffee per day. Meaning that typically, without doing any formal statistical testing, people that drink stronger coffee also tend to drink less coffee.



Due to the same reason as our first analysis, not knowing the underlying distribution of our data, we again decided to run a Chi-test. After running the Chi-square test between brewing methods and age we got a p-value of 4.989986023551648e-27 which also shows a significant relationship (p-value < 0.05).

After aggregating the data further, we found that among all age groups, the most popular way to brew coffee is pour over. And the least popular ways to brew coffee at home are coffee extract (e.g. Cometeer)), Bean-to-cup machine and instant coffee. We think pour over is the most popular way to brew coffee because it balances cost, time and quality. Least popular methods such as coffee extract and bean-to-cup machine face challenges due to their high cost, and instant coffee often has a reputation for inferior quality compared to other methods. So, unlike our first analysis where we found that the significant result highlighted a notable difference across the age





groups, in this analysis the significant result signifies a connection. As highlighted by the graphs above.

Limitations

There were a few limitations that we ran into when we were going through our analysis. While analyzing the dataset, we had to work with categorical age groups instead of precise ages, which really limited the granularity of our analysis. Another thing is that the dataset contained a large proportion of non-numerical responses that required mapping to numerical values for meaningful analysis. We had to accept the fact that this might have introduced potential bias in how descriptors were averaged. The dataset was also very skewed towards younger, individuals and underrepresented older adults which could have also produced a potential bias.

If we had more time, we would try to expand the dataset to include more participants across all age groups, especially underrepresented categories like the >65 years old group. This would provide a more balanced dataset and improve the reliability of machine learning predictions. Or maybe even try to investigate the influence of external factors like cultural or regional coffee preferences, which could provide deeper insights into these trends.

Project Experience Summary

Curtis:

- Conducted statistical tests to analyze the relationship between age, coffee taste preference and brewing methods by using scipy library to assess the significance of differences in tastes across different age groups, discovering significant trends based on age.
- Performed data analysis to find patterns in coffee brewing method and coffee taste preferences by using Matplotlib and Seaborn to create visual graphs, revealed insights that older consumers prefer stronger coffees and the most .
- Co-authored the report summarizing the results by including statistical test outcomes, visualizations and insights

Sanjit:

- Designed a Coffee Scoring System, to measure how strong people like their coffee and how much they drink on average. I set up the system from scratch, including creating a method to collect and analyze the data in a clear and organized way.
- Came Up with an Extra Question for the project, asking if people who prefer stronger coffee drink less overall. This question added depth to the research and gave us more interesting results to analyze.
- Managed the Machine Learning Part of the Project by choosing the best model to use, training it with the data we collected, and testing it to make sure it could accurately predict coffee preferences and brewing habits. I used Python and related tools to carry out this process.
- Co-authored most of the report and was in charge of writing and placement of the different models that we came up with, while most of the editing was left up to Curtis.