CMPT 353 D100
Summer 2024
Greg Baker
August 2, 2024

Insights from the Great American Coffee Taste Test

Analyzing Preferences and Trends in Coffee Consumption Across the United States

Hamoudi Saleh Baratta
SFU ID 301540229
mba177@sfu.ca

Brayden Sue
SFU ID 301434449
bm8@sfu.ca

Introduction

Overview: The Great American Coffee Taste Test analyzed coffee preferences and trends across the United States. The study involved more than 4,000 respondents who were shipped four different coffees: Coffees A, B, C, and D.

Participants: The study included 4,042 participants from various demographics, spanning urban, suburban, and rural areas across the U.S. The majority (over 75%) of participants were aged between 25 and 44 years old, with 72% male, 24% female, 3% non-binary, and 1% declining to answer.

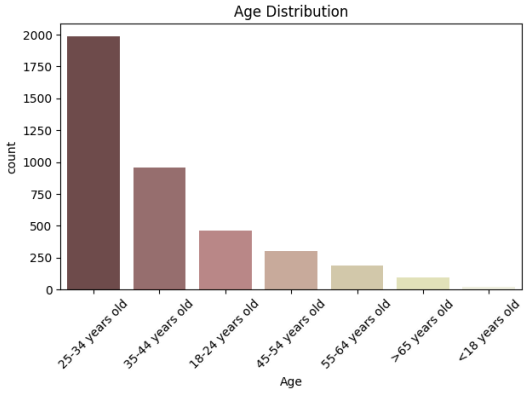

Figure 1: Age Distribution

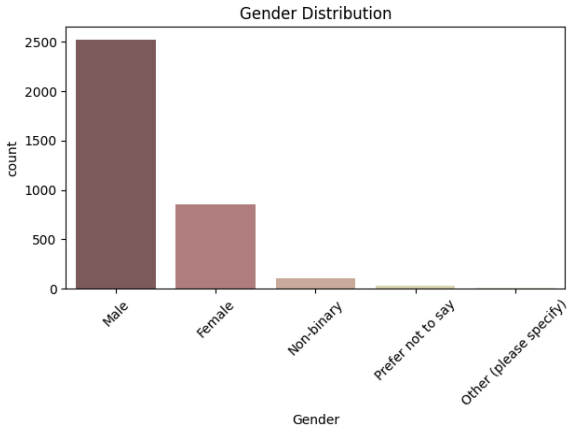

Figure 2: Gender Distribution

Problem Statement:

This project aims to identify varying coffee preferences across different demographics and understand the factors influencing these preferences.

Methodology

Coffee Selection: Coffee samples represented a range of flavors and brewing methods, including different roast levels (light: coffee A, medium: coffee B, dark: coffee C) and fermentation methods (coffee D).

Survey Process: Participants rated each coffee sample based on bitterness, acidity, flavor notes, and overall enjoyment through a structured survey.

Data Collection and Cleaning: Data was gathered through an online survey that took place during a Youtube Livestream by coffee enthusiast James Hoffmann. The data was then anonymized and made public. We downloaded the raw data and cleaned it to remove incomplete or inconsistent responses. The preprocessing of the data included renaming lengthy field names, enforcing required fields, imputing certain values, and encoding ordinal data in numerical form to aid statistical testing. Because the data includes columns for each boolean survey question and compiles the True values in a list, imputation was beneficial to fill in any fields where data was missing but could still be calculated. Another case was in the 'number of children' field, where a missing entry was assigned a value of 0 children. This was done primarily because of the main age range of participants which made it a reasonable assumption, although prone to misrepresenting data in other age ranges. As such, a comment was added to highlight the imputation and suggest removal if necessary.

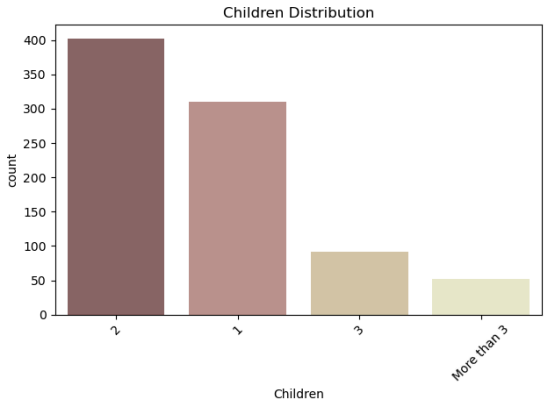

Figure 3: Number of Children Distribution with dropping no answers

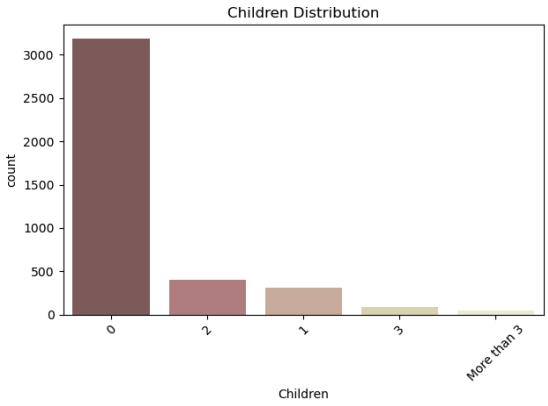

Figure 4: Number of Children Distribution accounting for when participants decided no to answer

Data Analysis Techniques: The data was analyzed using exploratory data analysis (EDA), statistical tests (Mann-Whitney U, Chi-Square), and a machine learning voting classifier to predict coffee preferences. To accomplish this task, Python and specialized data analysis libraries, including NumPy, Pandas, Scikit-Learn, and SciPy, were used.

Due to the categorical and ordinal nature of the data set as well as the uneven partitioning of participant backgrounds, non-parametric statistical tests were selected to determine the dependence between various groupings of the data. The ordinal data in the DataFrames

included the participants' age ranges, number of coffees per day, self-assigned expertise levels, and the ratings given for a variety of metrics during a testing of 4 different coffees (A, B, C, and D). To determine the difference between the distribution of groups in different age, gender, and expertise level demographics when rating the 4 coffees, the Mann-Whitney U test was selected as it is only concerned about sort order and requires no assumption of an underlying normal distribution. The categorical data included fields such as favorite coffee drinks, favorite coffee out of the 4 that were tested, where people drink coffee, and dairy choice. Out of these fields, the favorite coffee drinks and tested coffee were selected to determine if there was significant association between categorical groups.

Mann-Whitney U tests were run on ordinal data and returned p-values for each combination of bitterness, acidity, and personal preference with coffees A, B, C, and D between two groups, before storing the values in a DataFrame. The demographic groups chosen were: *Young-Old*, *Male-Female*, *Male-Other(Misc.)*, *Female-Other*, and *Low-High Expertise*. With a null hypothesis of 'the two groups share the same distribution' and a significance level of $\alpha = 0.05$, the null hypothesis was rejected in many cases - showing different distributions of ratings for different demographics.

Young vs Old

	A	B	C	D
Bi	0.000	0.000	0.000	0.000
Ac	0.000	0.000	0.186	0.000
Pp	0.011	0.003	0.000	0.000

Figure 5: Young-Old Mann-Whitney U Results

Favorite coffee vs Demographic

	Age	Gender	Expertise
Coffee B	0.000	0.000	0.000
Coffee D	0.000	0.000	0.000
Coffee A	0.000	0.000	0.000
Coffee C	0.000	0.000	0.000

Figure 6: Chi-Square Results

For example, as shown in *Figure 5*, male and female participants typically scored each coffee differently, except for the acidity (*Ac*) levels in coffees A and D. Additionally, among all groups except those including 'other' genders, bitterness (*Bi*) and personal preference (*Pp*) consistently rejected the null hypothesis. The reason for a difference in tests including miscellaneous genders was likely a large discrepancy in record counts between groups. As a side note, the p-values were rounded to 3 decimal places, so the values of 0.000 are much smaller than the significance level.

Chi-Square tests were run to determine association between participants' favorite coffees and their ages, genders, as well as expertise levels. The p-values were then stored in a DataFrame and as shown in *Figure 6*, they are all less than the significance level of 0.05. This allows us to conclude that age, gender, and expertise levels all influence and affect a participant's favorite coffee out of the 4 tested.

Results

Overall Preferences

- **Most Popular Coffee:** Coffee D (unique fermentation process) was the overall favorite with 37% of the total vote.
- **Least Popular Coffee:** Coffee B (medium roast) received 21% of the votes, similar to Coffees A and C.
- Most people (90%) drink coffee at home *Figure 7* shows how many people choose both column and row. So 30% of people who drink coffee at home also drink it at the office. The blank spaces are below 0.5%

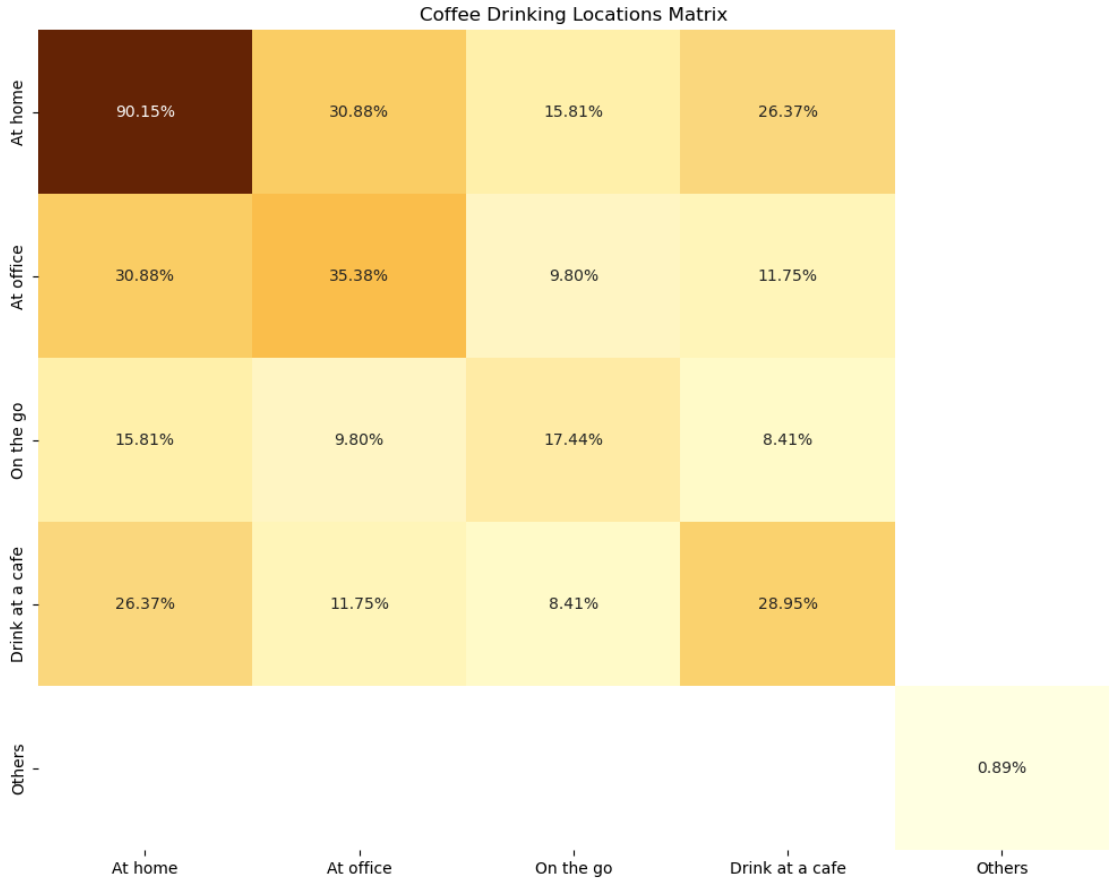

Figure 7: Coffee Drinking Locations Matrix

Demographic Insights

- **Age:** Younger participants (18-34) preferred lighter roasts and fermented coffees, while older participants (50+) favored darker roasts.
- **Gender:** Coffee D (fermented coffee) was more popular among men and less popular among women. Women showed a higher preference for light and medium roasts compared to men.
- **Brewing Method:** Pour over drinkers preferred lighter roasted and fermented coffees. Espresso and milk-based drinkers had a slight preference towards lighter roasts but a more balanced distribution overall.
- **Descriptive Preferences:** Participants who liked fruity descriptors preferred Coffees D and A. Participants who liked chocolaty descriptors preferred medium and dark roasts (Coffees B and C).

Machine Learning Model

The machine learning model used was a voting classifier, which combines the predictions of several models to form a final prediction. Classification was chosen due to a lack of quantitative data, which removes the possibility of regression. The models included in the classifier were Gaussian Naive Bayes, K-Neighbors, Decision Tree, Random Forest and Gradient Boosting classifiers, and although the data doesn't follow a normal distribution the GaussianNB model improves the accuracy score, possibly due to the CLT and having many data points. The question given to the model was: "Based on participants' demographics and their ratings for bitterness, acidity, and personal preference, what is their favorite coffee overall?", which was reflected in the selected feature and target columns. The data was then split into training and validation data sets, where the validation data set comprised 20% of all records, before being fed into the model and achieving an accuracy score of 0.848. *Figure 8* shows the confusion matrix for the predictive model.

Prior to the use of the voting classifier, the list of feature and target columns was much more extensive and rather than a person's favorite coffee, the model was trying to predict what kind of coffee drink (e.g. pour over, latte, espresso) was a person's favorite. At first, Gradient Boosting was selected as a model, but only provided a maximum accuracy score of 0.322. This was partly due to very imbalanced partitions, with the pour over method having approximately twice the amount of users than the next most popular, espresso. Therefore, the data was partitioned into different groupings of drink types based on level of complexity, such as basic, intermediate, and specialty. Subsequently, a MLP classifier was used on the new partitions, which achieved an accuracy score of 0.545.

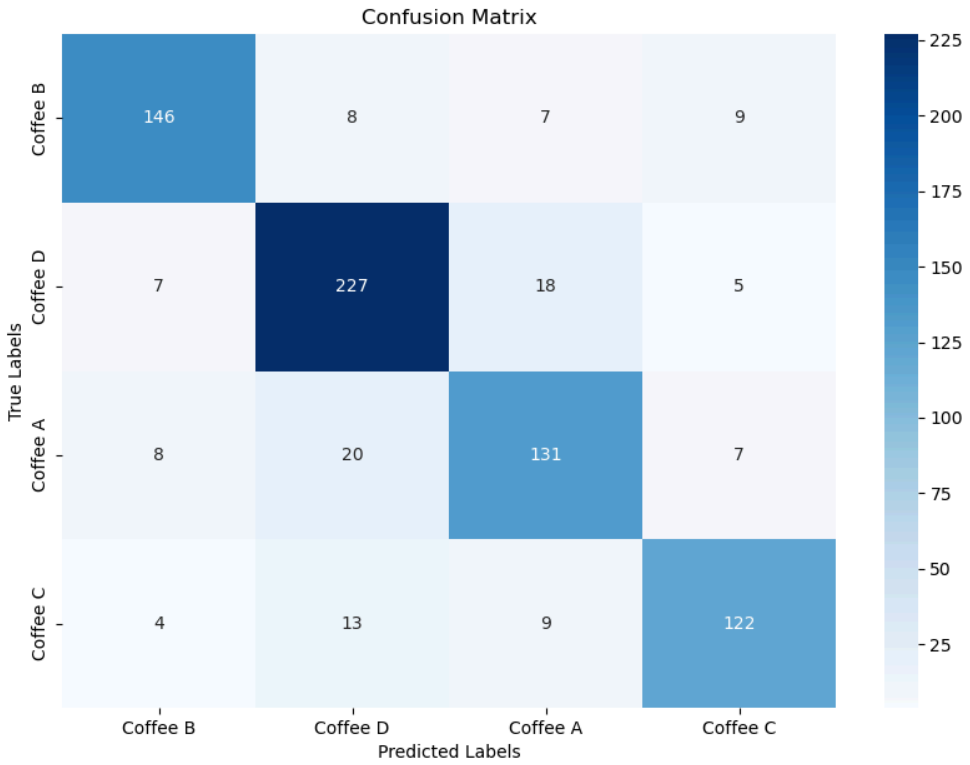

Figure 8: Machine Learning Classifier Confusion Matrix

Discussion

Key Findings: The data suggest that the trend is a growing popularity of light roasted and fermented coffees.

Implications: With these insights, coffee producers and retailers can make data-driven decisions about how to best tailor their products to different demographic groups.

Limitations

- **Self-reported data:** Subject to bias. For example, the distribution of self-assigned expertise scores in *Figure 9* are right-skewed, suggesting that the majority of participants consider themselves to know more than the average person about coffee. As a result, the findings may not be representative of the entire population.

- **Online survey:** May exclude certain demographic groups. As shown in *Figure 10*, White/Caucasian participants outnumbered the next largest ethnic group of participants, Asian/Pacific Islander, by over 500%. Furthermore, the distribution of participants skews heavily to males in the age range of 25-34 years old. As discovered through statistical tests, the demographics of participants significantly affects their taste opinions for coffee, so the skewed nature of the sample group will again introduce bias into the data.

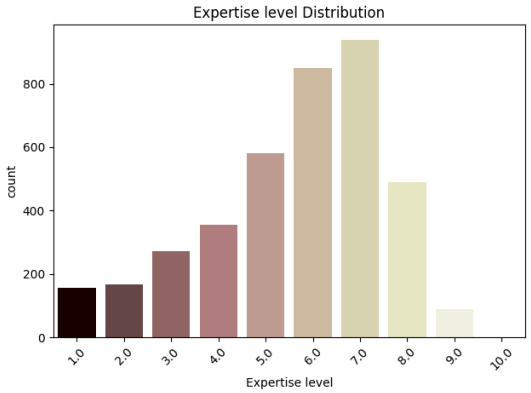

Figure 9: Participant Expertise Level Distribution

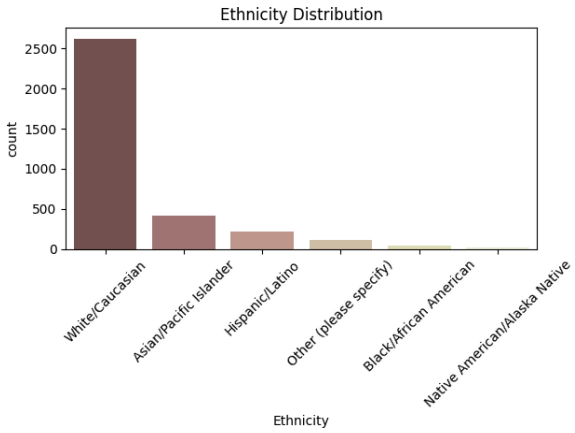

Figure 10: Participant Ethnicity Distribution

Conclusion

Summary: The Great American Coffee Taste Test provided valuable insights into coffee preferences and trends in the U.S, as well as highlighted different demographic preferences within the sample group. The accuracy of the classification model suggests that the ratings provided for different aspects of each coffee can reliably predict a tester's favorite coffee.

Future Research: By reconducting the experiment with random sampling and collecting timestamps, analysis could uncover unbiased long-term trends in coffee consumption behavior. Additionally, with a more representative sample group, the role of social and cultural demographics could also be further explored.

Accomplishment Statements

Brayden: Cleaned and imputed missing data from a coffee testing dataset, performed non-parametric statistical tests, created partitions, and developed a classification model to predict users' favorite coffee types based on their ratings with an accuracy score of 0.848.

Hamoudi: Analyzed world's largest coffee testing dataset through advanced data cleaning and imputation techniques. Conducted comprehensive non-parametric statistical tests leading to the development of a classification model that predicted users' favorite coffee types with an accuracy of 84.8%.

Appendix: Code Results

Below is the full list of summaries and visualizations generated by the project script provided. All can be found at the output directory.

Data Overview

- **Gender Distribution:** Plot generated by `plot_value_counts(df, 'Gender')`
- **Age Distribution:** Plot generated by `plot_value_counts(df, 'Age')`
- **Education Level:** Plot generated by `plot_value_counts(df, 'Education')`
- **Workplace:** Plot generated by `plot_value_counts(df, 'Workplace')`
- **Ethnicity:** Plot generated by `plot_value_counts(df, 'Ethnicity')`
- **Employment:** Plot generated by `plot_value_counts(df, 'Employment')`
- **Children:** Plot generated by `plot_value_counts(df, 'Children')`
- **Political Views:** Plot generated by `plot_value_counts(df, 'Politics')`
- **Coffee Consumption:** Plot generated by `plot_value_counts(df, 'Cups per day')`

Coffee Drinking Locations

- **Bar Plot and Heatmap:** Generated by `create_bar_plot` and `create_heatmap_from_dict` for coffee drinking locations.

At Home Brew Methods

- **Bar Plot and Heatmap:** Generated by `create_bar_plot` and `create_heatmap_from_dict` for brewing methods.

Coffee Purchasing Locations

- **Bar Plot and Heatmap:** Generated by `create_bar_plot` and `create_heatmap_from_dict` for purchasing locations.

Favorite Coffee Drink

- **Plot:** Generated by `plot_value_counts(df, 'Favorite coffee drink')`.

Coffee Additives

- **Bar Plot and Heatmap:** Generated by `create_bar_plot` and `create_heatmap_from_dict` for coffee additives.

Dairy Added

- **Bar Plot:** Generated by `create_bar_plot(dairy_df, 'Category', 'count', 'Dairy Added')`.

Sweetener Added

- **Bar Plot:** Generated by `create_bar_plot(sweetener_df, 'Category', 'count', 'Sweetener Added')`.

Coffee Preferences and Expertise

- **Plots:** Generated by `plot_value_counts` for coffee preference, strength, roast level, caffeine level, and expertise level.
- **Average Expertise:** `avg_expertise = df['Expertise level'].mean()`.

Reasons for Drinking Coffee

- **Bar Plot and Heatmap:** Generated by `create_bar_plot` and `create_heatmap_from_dict` for reasons for drinking coffee.

Additional Plots

- **Taste of Coffee:** `plot_value_counts(df, 'Like coffee')`
- **Knowledge of Coffee Origins:** `plot_value_counts(df, 'Know coffee origins')`
- **Max Paid for Coffee:** `create_bar_plot(max_paid, 'Max ever paid for a coffee cup', 'count')`
- **Max Willing to Pay:** `create_bar_plot(max_acceptable, 'Max willing to pay for a coffee cup', 'count')`
- **Value for Money at Cafe:** `plot_value_counts(df, 'Good value for money spent at cafe')`
- **Equipment Cost:** `plot_value_counts(df, 'Equipment cost')`
- **Equipment Value:** `plot_value_counts(df, 'Equipment value')`

Coffee Ratings

- **Coffee A:** `coffee_A = calculate_average_ratings(df, 'A')`
- **Coffee B:** `coffee_B = calculate_average_ratings(df, 'B')`
- **Coffee C:** `coffee_C = calculate_average_ratings(df, 'C')`
- **Coffee D:** `coffee_D = calculate_average_ratings(df, 'D')`

Preference Comparisons

- **Preferred Coffee:** `plot_value_counts(df, 'A or B or C')`
- **Coffee A vs D:** `plot_value_counts(df, 'A or D')`
- **Favorite Coffee:** `plot_value_counts(df, 'Favorite coffee')`

Grouped Bar Plots

- **Favorite Coffee by Age:** `create_grouped_bar_plot(df, 'Favorite coffee', 'Age', 'count')`
- **Favorite Coffee by Gender:** `create_grouped_bar_plot(df, 'Favorite coffee', 'Gender', 'count')`
- **Favorite Coffee by Expertise:** `creat0e_grouped_bar_plot(df, 'Favorite coffee', 'Expertise level', 'count')`

Percentage Grouped Bar Plots

- **Age Group Preferences:** `create_grouped_bar_plot(df, 'Favorite coffee', 'Age', '% of Age group', use_percentages=True)`
- **Gender Group Preferences:** `create_grouped_bar_plot(df, 'Favorite coffee', 'Gender', '% of Gender group', use_percentages=True)`
- **Expertise Group Preferences:** `create_grouped_bar_plot(df, 'Favorite coffee', 'Expertise level', '% of Expertise group', width=0.09, use_percentages=True)`

Roast Level Preferences

- **Age Group Roast Preferences:** `create_grouped_bar_plot(df, 'Roast level', 'Age', '% of Age group', use_percentages=True)`
- **Gender Group Roast Preferences:** `create_grouped_bar_plot(df, 'Roast level', 'Gender', '% of Gender group', use_percentages=True)`

- **Expertise Group Roast Preferences:** `create_grouped_bar_plot(df, 'Roast level', 'Expertise level', '% of Expertise group', width=0.09, use_percentages=True)`

Statistical Tests

- **Mann-Whitney U Tests:** Results saved to `output/stat_tests`.
  - Young vs Old: `r_age`
  - Male vs Female: `r_male_female`
  - Male vs Misc Gender: `r_male_other`
  - Female vs Misc Gender: `r_female_other`
  - Low vs High Expertise: `r_expertise`
- **Chi-Square Tests:** Results saved to `output/stat_tests`.
  - Favorite coffee vs Demographic: `coffee_demographics`
  - Favorite drink vs Demographic: `drink_demographics`

Machine Learning Model

- **Model Accuracy:** 84.8%
- **Model Results:** `model, X_valid, y_valid, y_pred` from `run_classification`