Coffee Preferences

Brilliant Cassowary - Report

Nidhi Soma (ns848) Joice Chen (jc3528) Jinpeng Li (jl3496)
Stephen Syl-Akinwale (sis33)

Introduction

As our group was sitting in a cafe, deciding on what to research, we noticed that one of the most frequently bought drinks on campus is coffee. But it is present in many more settings, ranging from casual coffee shop frequenters and critical coffee enthusiasts. We looked through many sources that collect data, and found one dataset from Data is Plural courtesy of James Hoffmann that seemed promising. The source had links to Hoffmann's own video about the dataset, and we found that he seemed to be a creditable source, so we were fairly confident about trusting the data that he collected.

Our question was how does coffee preference differ by demographic attributes such as age, gender, education level, race, and political identity? To examine this question in a narrower scope, we looked at certain consumption habits based on age, gender, and education level. We found that there is an association between roast preference and age, and males on average consume more coffee per day than females. We were not able to make conclusions about how education level affects the amount of sweetness preferred. With these results, we now have some information about the consumers, which can allow for tailored marketing strategies.

Data description

Our dataset comes from survey responses that world champion barista James Hoffmann made available to the public after conducting his "Great American Coffee Taste Test". This datset was also partially funded by Cometeer – a subscription service that makes flash-frozen coffee capsules - who may be interested to know people's preferences in order to make their coffee capsules more appealing to a wider market. The observations represent an individual respondent to the survey. There are 4,042 rows. The columns are questions that they answered, ranging from demographic data to coffee preferences. There are 98 of these columns. This dataset was created to understand the general public's preferences as consumers for coffee.

The survey quickly was sold out, and Hoffman's audience in general is coffee specialists. That will likely skew the population surveyed to be people who likely prefer specialty coffee, so it may be a biased sample. Additionally, this survey was conducted through people ordering tasting kits online, which were then sent to the participants to prepare and complete voluntarily, so there may have been differences in that.

There was some preprocessing in that zip codes and geographic data have been removed and participants were anonymized to protect their privacy, so it doesn't seem possible to re-identify individuals. The dataset contains information that may be confidential and identifies subpopulations, such as political affiliation, race, age, gender, employment status, number of children, and education level. It wasn't disclosed how Hoffmann and his team collected all the taste test results that participants filled out, but once they got that data, they made it into a spreadsheet to be shared with the public. People were made aware of the data collection by a YouTube video. Hoffmann made his intentions for the survey clear in his video, and that he was planning to publicize the raw data later on. The participants had to order the coffee tasting kit on their own in order to participate, implying their consent to these terms. There are no explicit relationships between individual instances because the data was anonymized, and there are also no recommended data splits in this collection.

Some data cleaning was done to rename the variables and remove columns with little information, such as "flavorings" and "Gender (please specify)". The raw data is stored in our repository and can still be accessed through this link publicly (https://bit.ly/gacttCSV+).

Detailed description on task and methods used to collect the data:

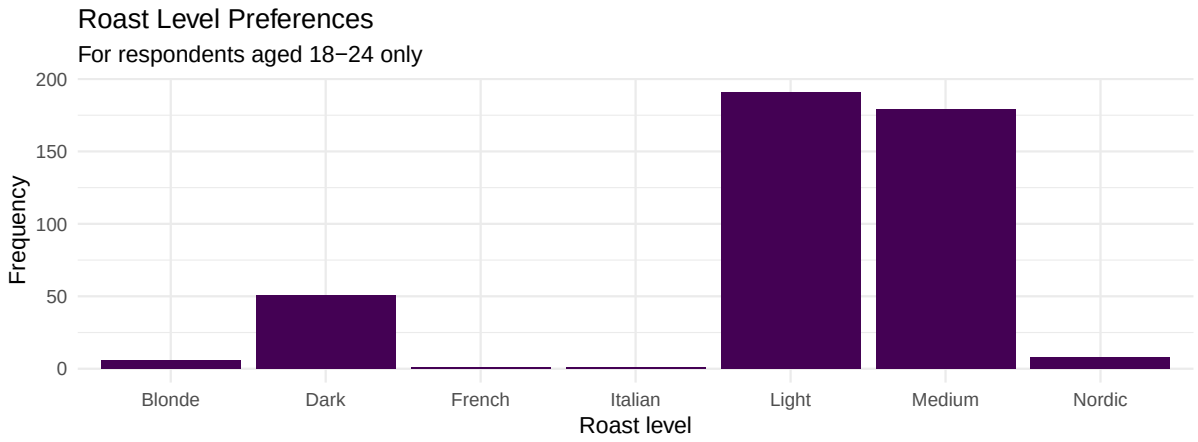
(https://www.youtube.com/watch?v=1fN_z4-EcOU)

Data analysis

We will look into several factors such as age, gender, education level, race, and political identity. To do this we cleaned the data in a manner that preserves variables related to coffee preference, consumption habits and demographic indicators. This Data Analysis attempts to visually represents each variables and their relations to coffee.

Pre-registered hypotheses

Our first pre-registered hypothesis was that younger people 18 - 25 prefer their coffee with lighter roasts.



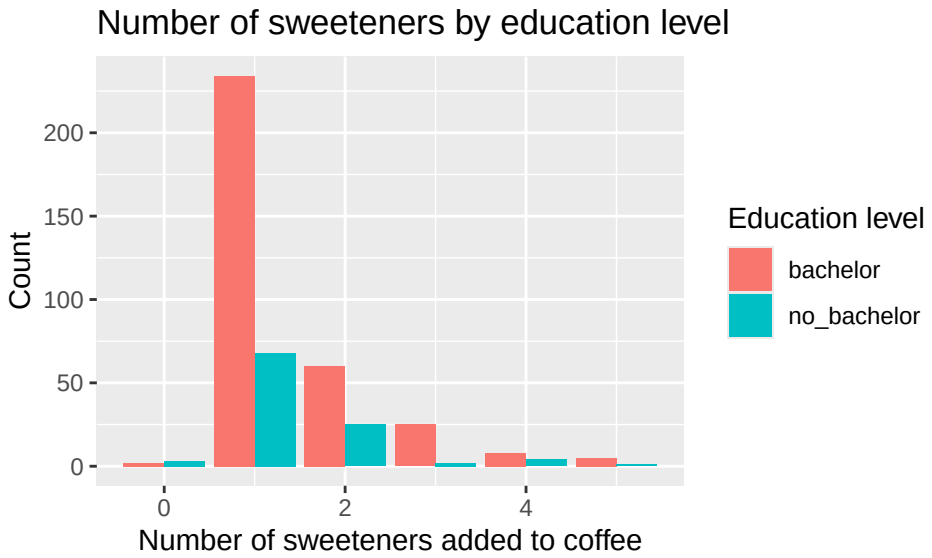
We see that there is a clear peak around the Light and Medium roast levels, suggesting that in this sample, younger people that are 18-24 years old have a strong preference for Light and Medium roast levels. However, there is more of a preference for Dark roasts than Nordic roasts, even though Nordic roasts are also generally considered to be lighter roasts.

This leads to the question of whether the younger age range of 18-24 is the most likely age to prefer lighter roast levels (Light, Medium, or Nordic). We investigate this below by using a linear regression model. In the future, exploring a logistic regression model may also lead to more accurate results, since the data is mostly in a few categories, so it might not be very linear.

```
# A tibble: 7 x 5
  term               estimate std.error statistic  p.value
  <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)           0.842    0.0772    10.9   2.73e-27
2 age18-24 years old    0.0229   0.0789     0.290 7.72e- 1
3 age25-34 years old    0.0519   0.0776     0.669 5.04e- 1
4 age35-44 years old    0.0342   0.0780     0.439 6.61e- 1
5 age45-54 years old   -0.0233   0.0797    -0.293 7.70e- 1
6 age55-64 years old   -0.146    0.0811    -1.81  7.11e- 2
7 age>65 years old     -0.161    0.0847    -1.90  5.70e- 2
```

Each of the coefficients gives the proportion of people choosing lighter roasts for the corresponding age group relative to the reference level of <18 years old. The reference level has a coefficient of 0, meaning its estimated proportion is the intercept, 0.842. Notably, the model predicts that people aged 18-24 years old will have a 0.029 lower proportion of people that prefer lighter roasts than people aged 25-34 years old, on average. So, the model predicts that younger people aged 18-24 are actually not the most likely group to prefer lighter roasts.

Pre-registered Hypothesis 2 Our second pre-registered hypothesis was that people who have higher levels of education are less likely to prefer their coffee with added sweetener. We can visualize this data to see if there seem to be any trends.

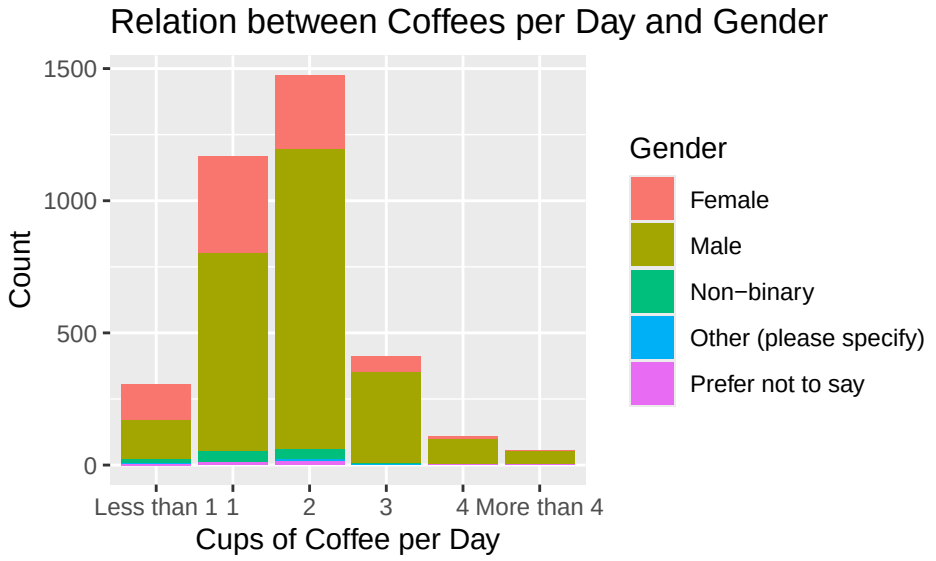


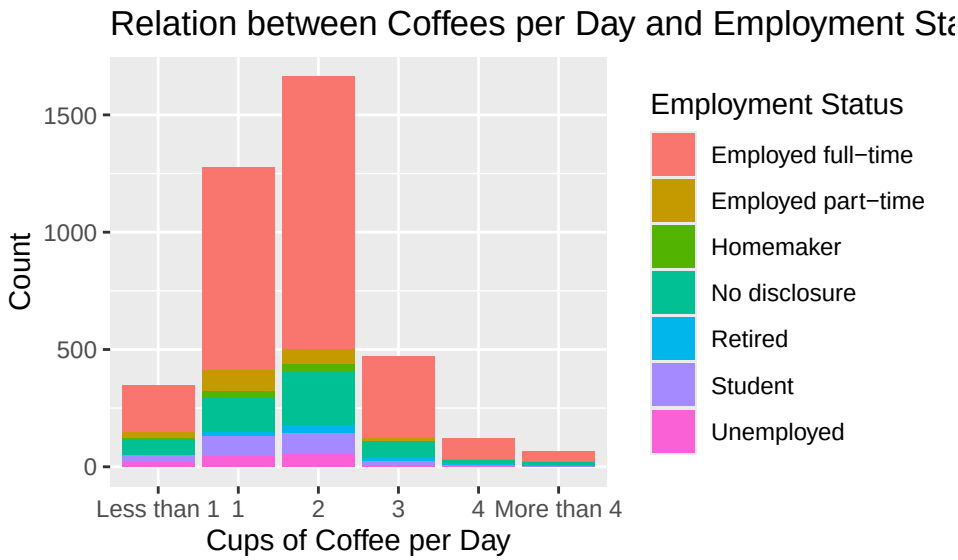
The visualizations for the proportions showed little difference between the distribution of sugar types preferred by those with a Bachelor's or higher degree compared to that preferred by those with less than Bachelor's degree. Instead of looking at the types of sweeteners, we can look at the number of sweeteners that a person uses. Seeing the number of sweeteners a person adds gives an indication to their tolerance/preference for sweetness in their coffee. We can answer the question of how many sweeteners someone uses corresponding to education level through the logistic regression below.

```
# A tibble: 2 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)  -1.08       0.225    -4.81  0.00000153
2 sweet_count  -0.0676     0.137    -0.493 0.622
```

From the results of the logistical regression model, the negative coefficient for sweet_count reflects a lower probability for people with less than a Bachelor's degree choosing more sugar additive types than a person with a Bachelor's degree or higher would.

We explored other demographic attributes below.





In our analysis we can infer several things according to different variables and fields of participants.

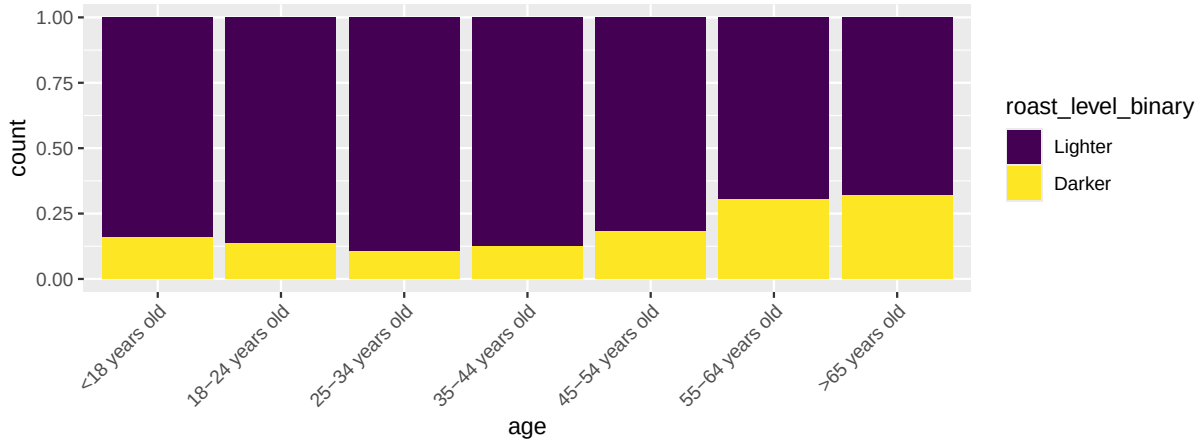
Cups of Coffee per day by Gender: The visuals suggests that people that identify as males are the highest counts of coffee drinkers per day. In all counts they drink less than 1 and more than 4.

Cups of Coffee per day by employment: Employed fulltime individuals are the most frequent coffee drinkers, with at least 2 cups per day for 900 counts of employed individuals. On the opposite side homemakers are recorded to have the least counts of coffee cups.

Evaluation of significance

Pre-registered Hypothesis 1

We can investigate if the difference in proportions of preferring a lighter roast between age groups is statistically significant or not with a chi-squared test of independence.



If there were no relationship between these two variables, we would expect to see the purple bars reaching to the same height, regardless of age group. We need to conduct a chi-squared test of independence to see if the differences we see here are just random noise or a meaningful relationship.

Null hypothesis: There is no association between age group and a lighter roast level preference.

$$H_0 : \textit{Age group and lighter roast level preference are independent.}$$

Alternative hypothesis: There is an association between age group and a lighter roast level preference.

$$H_a : \textit{Age group and lighter roast level preference are not independent.}$$

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

We can reject the null hypothesis in favor of the alternative because our p-value of approximately 0 is less than the alpha value of 0.05.

Pre-registered Hypothesis 2

To evaluate whether the number of types of sweetener accepted differ between people with different education degrees, we will conduct a two-side hypothesis test

Null hypothesis: The true average number of types of sweetener preferred is the same between people in America with a Bachelor's degree or higher and people without a Bachelor's degree.

$$H_0 : \mu_{\text{bach}} - \mu_{\text{no bach}} = 0$$

Alternative hypothesis: The true average number of types of sweetener preferred is not the same between people in America with a Bachelor's degree or higher and people without a Bachelor's degree.

$$H_a : \mu_{\text{bach}} - \mu_{\text{no bach}} \neq 0$$

```
# A tibble: 2 x 2
  education_level  mean
  <fct>           <dbl>
1 bachelor         1.46
2 no_bachelor      1.41
```

The observed statistic is 1.455-1.408 = 0.047

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.349
```

With a p-value of 0.35, which is larger than the critical value of 0.05, we do not have sufficient evidence to reject the null hypothesis.

Research Question

Narrowing to Coffee Consumption by Gender

Research Focus: Another common demographic feature is gender, and we can see if it is related to the amount of coffee someone drinks - another type of coffee preference. We saw in the data analysis section that there seems to be a clear relationship between gender and cups of coffee consumed per day. We can focus on the genders with the most data for better results - male and females. To assess whether there's a significant difference in the number of cups of coffee consumed per day across these genders, we can conduct an independent two sample t-test using the group means. This is an appropriate test since we want an average for two groups with different variances (male data is larger than female data)

Null Hypothesis (H0): Males and Females consume the same average number of cups of coffee per day.

$$H_0 : \mu_{\text{Male}} - \mu_{\text{Female}} = 0$$

Alternative Hypothesis (Ha): There is a difference in the average number of cups of coffee consumed per day between Male and Female.

$$H_0 : \mu_{\text{Male}} - \mu_{\text{Female}} \neq 0$$

We'll proceed with an independent two-sample t-test to compare the average number of cups of coffee consumed per day between male and female groups.

```
    Welch Two Sample t-test

data:  male_coffee and female_coffee
t = 9.796, df = 1303.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2439571 0.3661370
sample estimates:
mean of x mean of y
 1.905887  1.600840
```

The p-value is less than 2.2e-16, which is essentially zero for practical purposes. We have strong evidence against the null hypothesis.

Interpretation and conclusions

Hypothesis 1

According to the evaluation of significance, we reject the null hypothesis because our p-value of approximately 0 is less than the alpha value of 0.05. Thus, there is sufficient evidence that age group and preferring lighter roasts of coffee are not independent of each other. A possible explanation for this phenomenon is that certain age groups are more prone to enjoying a certain taste, while others prefer the caffeine. In the future, we could analyze the the columns specifying people's reason for enjoying coffee versus their age group to explore this hypothesis.

In the data analysis, a possible explanation for younger people showing more of a preference for Dark roasts than Nordic roasts, despite showing an overwhelming preference for Light and

Medium roasts, is that more people may be unfamiliar with what a Nordic roast is. For people with general coffee knowledge, seeing light, medium, and dark roasts tends to be more common than Nordic; however, we should still consider that this sample likely has many respondents that are knowledgeable about coffee. So, these results still may have another underlying cause.

Hypothesis 2

The hypothesis test resulted in a p-value of 0.35, which is larger than the critical value of 0.05. This high p-value indicates that we do not have sufficient evidence to reject the null hypothesis. Therefore, we conclude that the observed difference in the number of sweeteners used by individuals with and without a Bachelor's degree is not statistically significant. It is possible for the lower education level group to have a more skewed average because of their much smaller sample size, and thus larger variability.

Given the lack of statistical significance, any observed differences in sweetener use between the two educated groups are likely attributable to random variability rather than the true underlying effect of education on sweetener diversity. This finding suggests that factors other than education may play a more critical role in determining the type of sweetener an individual chooses to use.

Research Question 1

The calculated t-statistic is 9.796, which is a measure of the difference between the two sample means relative to the variation in the samples. A higher t-value indicates a greater degree of difference. The p-value is less than 2.2e-16, which is essentially zero for practical purposes. This is far below the commonly used significance level of 0.05, indicating strong evidence against the null hypothesis.

We reject the null hypothesis that there is no difference in the average number of cups of coffee consumed per day between males and females. The data provides strong evidence that males, on average, consume more cups of coffee per day than females. The result is statistically significant with a high degree of confidence. We should consider, however, that this dataset contains mostly males, so the mean for females may not be as representative.

Overall, for this data, we can say that there is an association between roast preference and age, and males on average, consume more coffee per day than females. This gives us some specific insights into consumption habits based on these two specific demographic features. This dataset has many interesting relationships to explore that were out of scope for this project, but in the future we would like to be able to generalize our results - so try to eliminate biases - and perform more text analysis to determine people's sentiments about coffee.

Limitations

As mentioned in previous sections, a big limitation of the data is that it's likely not representative of the population of the US, and because it was only done in the US, this isolates coffee experiences and preferences common to the US only. The study relied on data from survey respondents who likely follow James Hoffmann and thus have a bias towards an interest in coffee, potentially skewing the results toward those who are more engaged with their dietary choices and creating voluntary or nonresponse bias. This means people who don't like coffee won't have as much as an input, so coffee taste going forward may not be tailored to them, making them dislike it even more. Most respondents also tended to be male, white, and middle aged, which is not representative, so this data cannot be generalized to the US population, let alone any consumer outside the US. Additionally, since the data was self-reported, there could be inaccurate values that lead to inaccurate inferences. Our inferences assume that this data is correct and representative if they were to relate to any population beyond this sample. If a coffee company like Cometeer wanted to adjust their coffee based on our concluded trends, they may waste money and time making a new recipe when the actual market doesn't favor that recipe.

Acknowledgments

We thank all participants and James Hoffmann for their contributions to this study and acknowledge the efforts of our research team in collecting and analyzing the data. Their hard work has been instrumental in enhancing our understanding of consumer behavior related to coffee consumption. We also thank our project mentor, Breanna Green, for her feedback and suggestions throughout multiple phases of this project.

Source of data:

https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid=