

Project title

Brilliant Cassowary - Report

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

Introduction

One of the most widely known beverages in today's age is coffee. It is present in many different settings, ranging from students and employees to casual coffee shop frequenters and critical coffee enthusiasts. Our current research questions are: - How does self-perceived coffee experience affect actual and stated preferences? - How does coffee preference and consumption habits differ by demographic attributes such as age, gender, education level, race, and political identity?

Data description

What are the observations (rows) and the attributes (columns)?

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.

Why was this dataset created?

To understand the general public's preferences as consumers for coffee. Additionally, since Cometeer funded the creation of this dataset, they may be interested to know people's preferences in order to make their coffee capsules more appealing to a wider market.

Who funded the creation of the dataset?

World champion barista James Hoffmann and Cometeer – a subscription service that makes flash-frozen coffee capsules.

What processes might have influenced what data was observed and recorded and what was not?

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. One example is that because participants were following a livestream to demonstrate how to do their taste test, their coffees may have been out long enough to have cooled, which could be another unaccounted variable that affected their taste preferences.

What preprocessing was done, and how did the data come to be in the form that you are using?

Zip codes and geographic data seemed to have been removed. Participants were anonymized to protect their privacy. 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.

If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

People were involved, and they were made aware of the data collection by a YouTube video stating that the purpose of this taste test was to understand coffee preferences in the USA. The participants had to order the coffee tasting kit on their own in order to participate, showing their willingness to accept these terms. Hoffmann also made his intentions clear in his video with why he wanted to collect the data, and that he was planning to publicize the raw data later on.

Was there a specific gap that needed to be filled? Please provide a description.

Are relationships between individual instances made explicit? There are no explicit relationships between individual instances because the data provided is it’s sole instance. *Are there recommended data splits?* No there are no recommended Data splits in this collection. *Are there any errors, sources of noise, or redundancies in the dataset?* There are instances of *Does the dataset contain data that might be considered confidential?* None *Does the dataset identify any subpopulations (e.g., by age, gender)?* Yes, the great America Coffee taste separated entries into sub-populations at the discretion of the participant *Is it possible to identify individuals, either directly or indirectly from the dataset?* No the identity of each participant was anonymized. *Does the dataset contain data that might be considered sensitive in any way What mechanisms or procedures were used to collect the data?* Participants who participated of the great American Coffee test received an event kit includes four different coffees from four different roasters that will be chemically identical if used with distilled water at the same ratio. Participants will need a weighing scale, a way of heating hot water, and cups that can accommodate about 250 mLs or eight ounces. The event is only available in the US, but the organizers are interested in conducting similar events in other regions. *Were the individuals in question notified about the data collection?* The data was self reported and as a result consensual to each participant. *Was any preprocessing/cleaning/labeling of the data done* Yes data cleaning was done to make the

variables and fields more readable and humanly understandably. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data* Yes the raw data is stored in our repository and can still be accessed through this link publicly (<https://bit.ly/gacttCSV+>) *Is there anything about the composition of the dataset or the way it was collected and leaned/labeled that might impact future uses?* The data was collected by individuals all over the US. Therefore there may be inconsistencies with recording the data since there were no qualifications or standardized way of verifying results, except from participants practices.

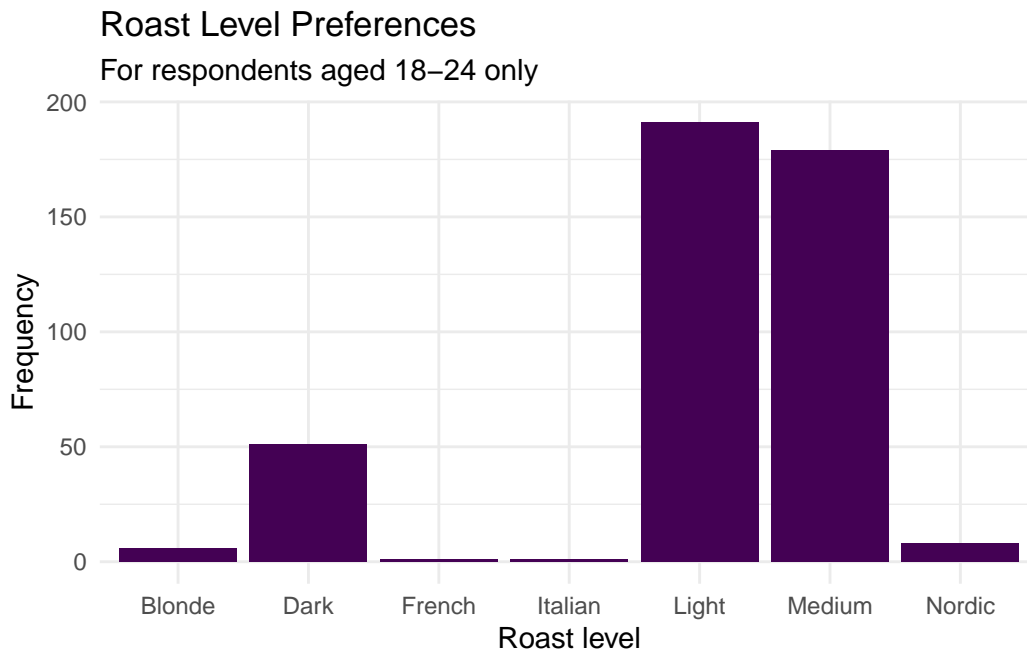
Notes

Coffee A - Light roast, Washed Coffee B - Medium Roast Coffee C - Dark roast Coffee D - Fermented, Natural, Fruity Detailed description on task and methods used to collect the data(https://www.youtube.com/watch?v=1fN_z4-EcOU) *Loading data*

Data analysis

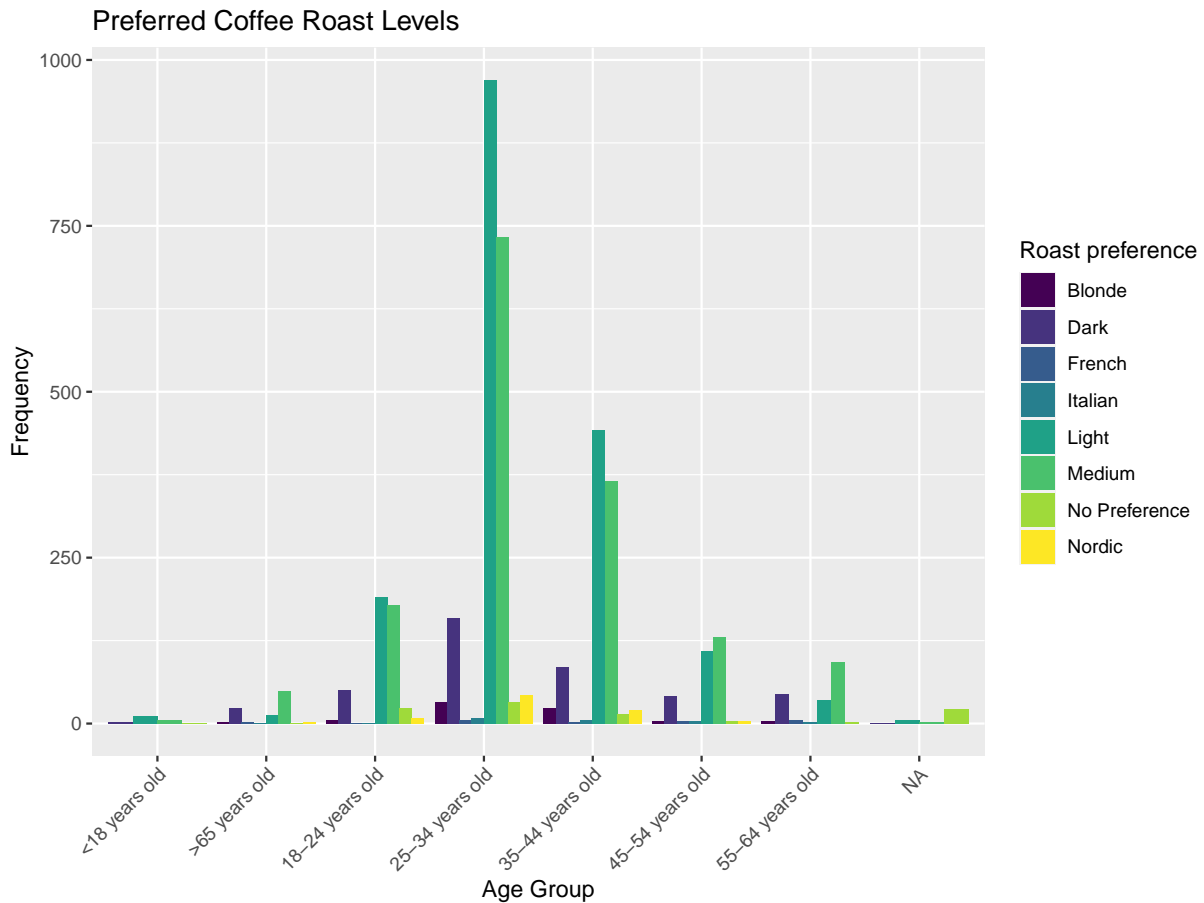
Pre-registered hypotheses

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 first visualizing the preferences of all the age ranges, and then using a linear regression model.

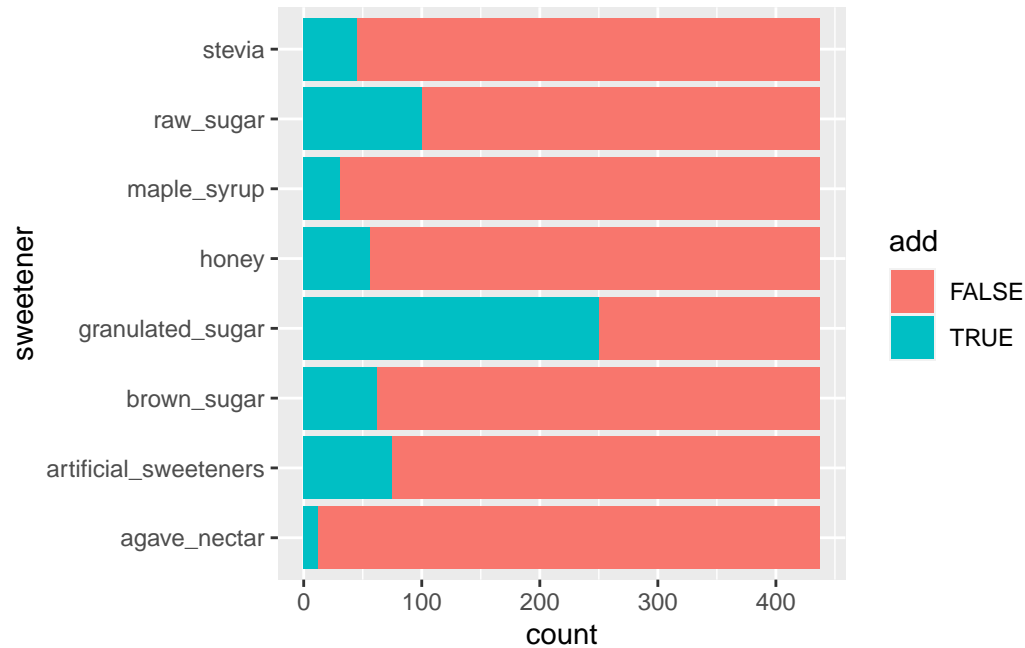


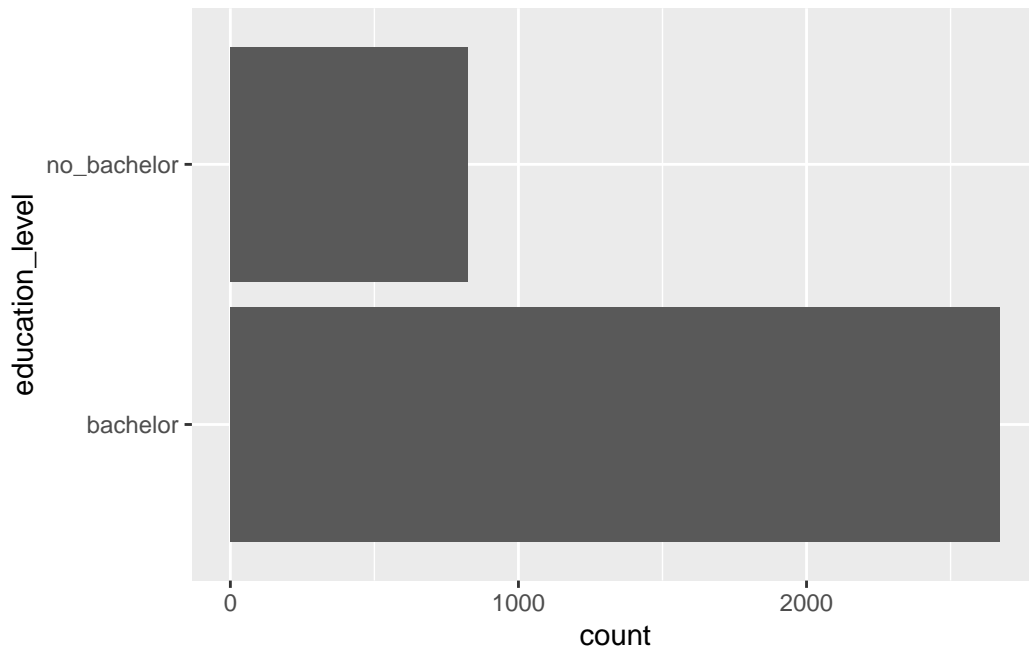
In the future, exploring a logistic regression model may also lead to more accurate results, since the data when visualized doesn't appear to be very linear.

```
# 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
```

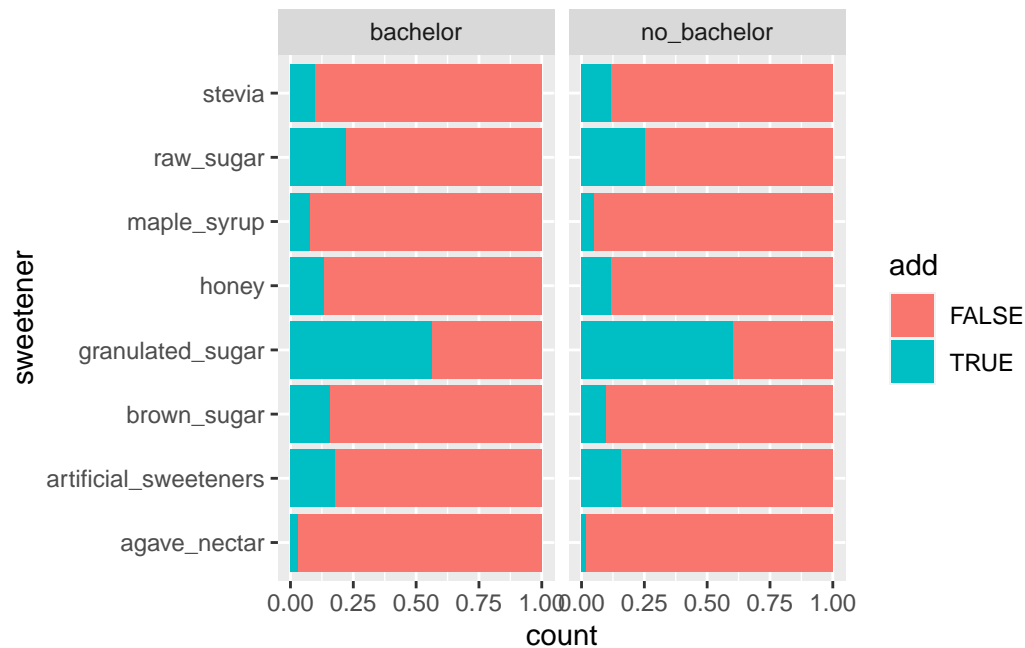
Each of the coefficients gives the proportion of people choosing lighter roasts for the corresponding age group relative to the reference level of <18 years old. The reference level has a coefficient of 0, meaning its estimated proportion is the intercept, 0.842. Notably, the model predicts that people 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.





These graphs show the distributions for each variable, so we can get an idea of the different groups we are working with. We see that granulated_sugar is the most popular sweetener, and people with higher than a Bachelor's degree are more common in this sample, which could potentially skew results. Next, we see how these two variables are related to each other through proportions, where we can more directly compare the visual differences.





The visualizations for the proportions seem to show 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. This can tell us about their taste preferences, and perhaps lead to insights for making a more appealing coffee to that market of users. 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
```

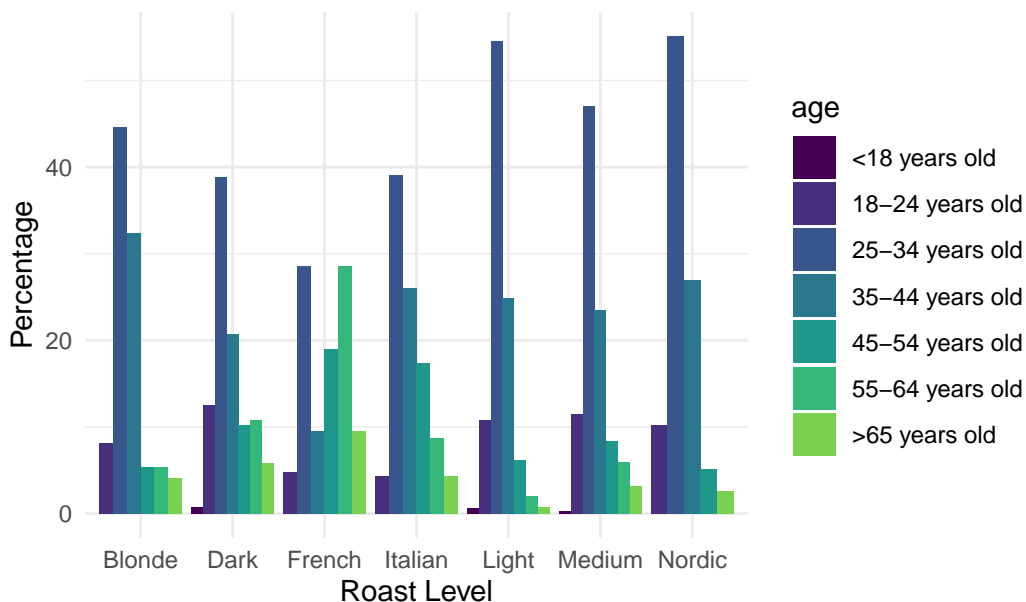
From the results of the logistical regression model, the negative coefficient for `sweet_count` reflects a lower probability for people with less than a Bachelor's degree choosing more sugar additive types than a person with a Bachelor's degree or higher would.

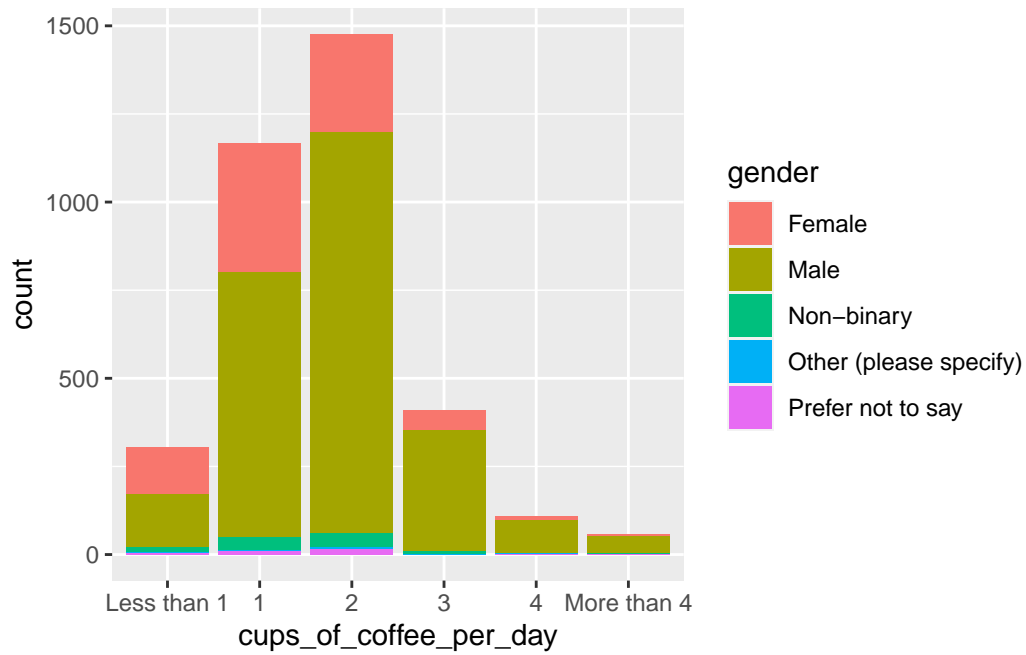
We now address our first question on actual self-perceived coffee experience compared to stated preferences. We use examples of people stating their preferences and look at the difference in data where they have a taste test.

Research Question 1 To answer the question how does coffee preference and consumption habits differ by demographic attributes? 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.

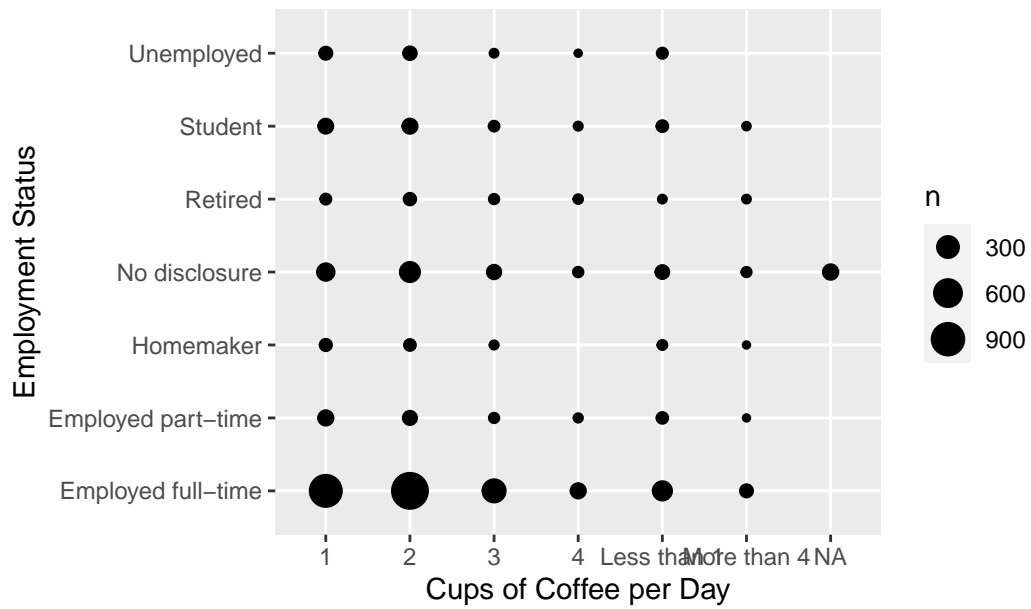
This Descriptive Analysis attempts to visually represents each variables and their relations to coffee.

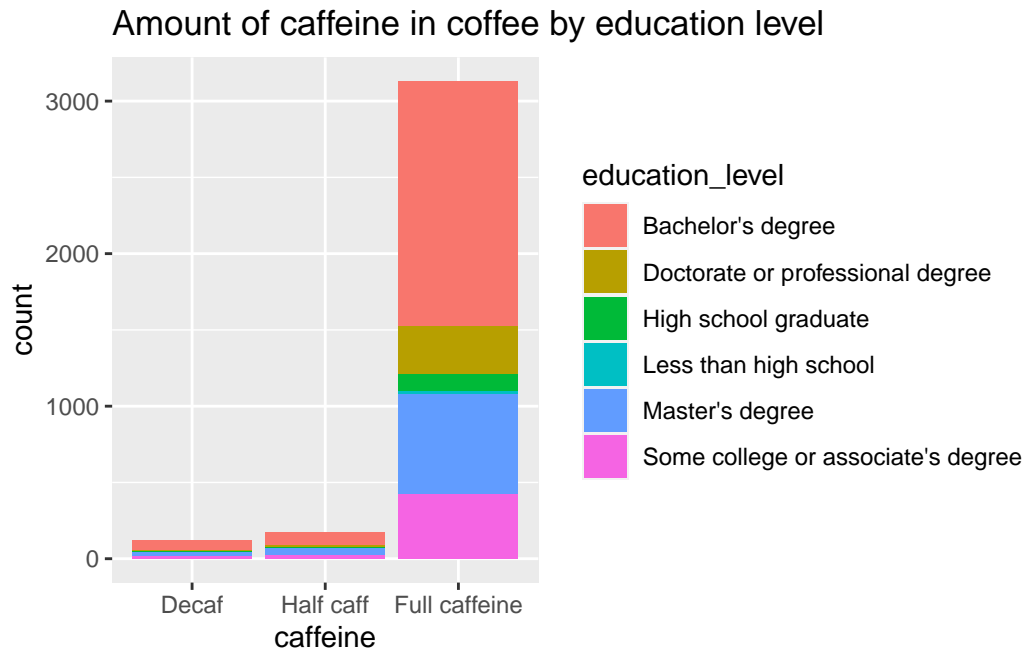
Coffee Roast Level Preferences by Age Group





Correlation between Coffees per Day and Employr

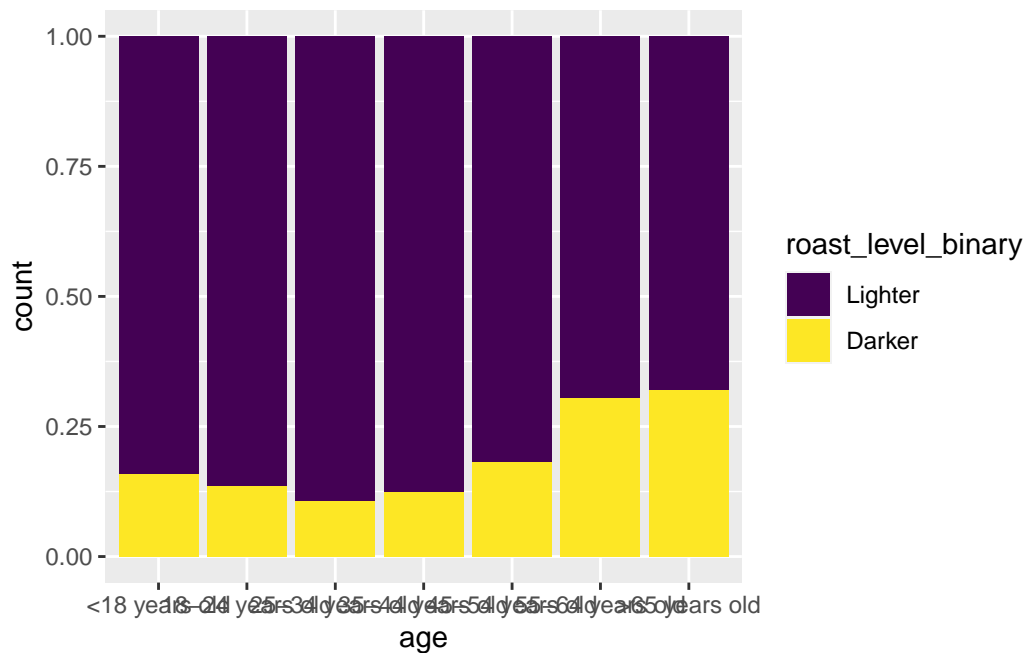




Evaluation of significance

Pre-registered Hypothesis 1

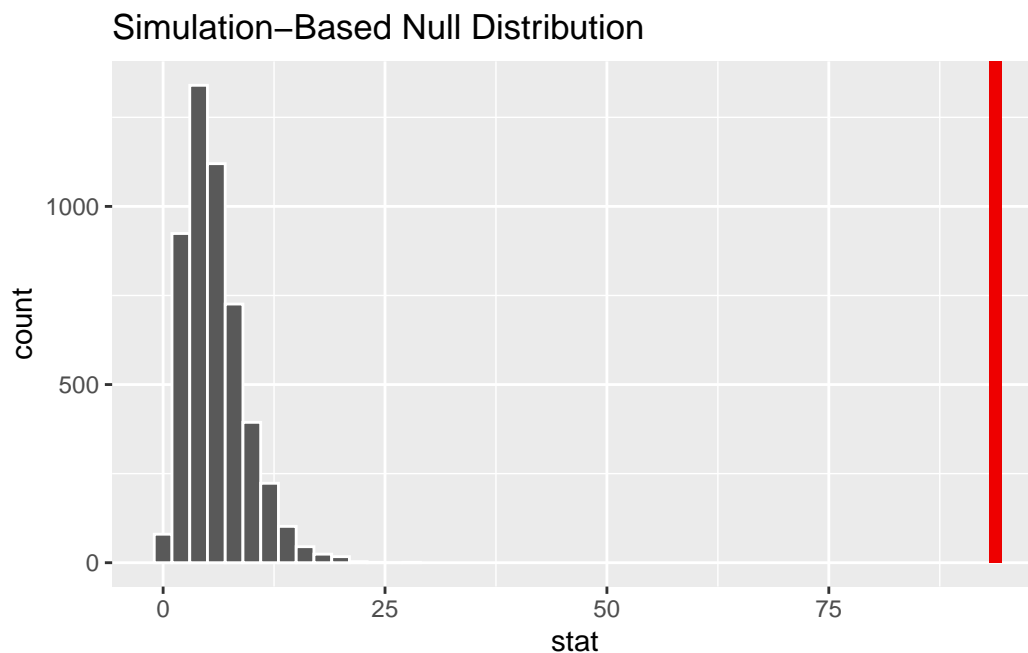
We can investigate if the difference in proportions of preferring a lighter roast between age groups is statistically significant or not with a chi-squared test of independence.



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.

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



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

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.

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

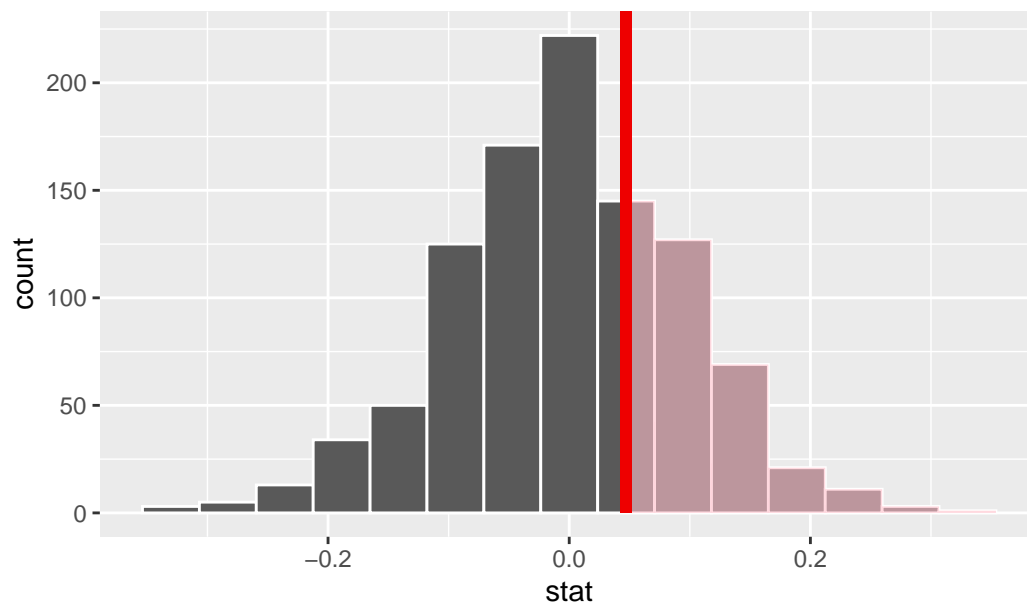
Alternative hypothesis: The true average number of types of sweetener preferred is not the same between people in America with a Bachelor's degree or higher and people without a Bachelor's degree.

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

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

The observed statistic is $1.455 - 1.408 = 0.047$

Simulation-Based Null Distribution



```
# A tibble: 1 x 1
  p_value
  <dbl>
1 0.333
```

Research Question 2

Interpretation and conclusions

Hypothesis 1 From the evaluation of significance, using an alpha value of 0.05, because our p-value of approximately 0 is less than 0.05, we can reject the null hypothesis. There is sufficient evidence that there is an association between age group and preferring lighter roasts of coffee.

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 and reveal something about younger people's coffee preferences.

Hypothesis 2

Limitations

Despite being such a wide variety of people this data is specific to the US and Isolate coffee experiences and preferences common to the US only.

Acknowledgments