Testing Trends Using Boundless Analytics and Permutation Tests

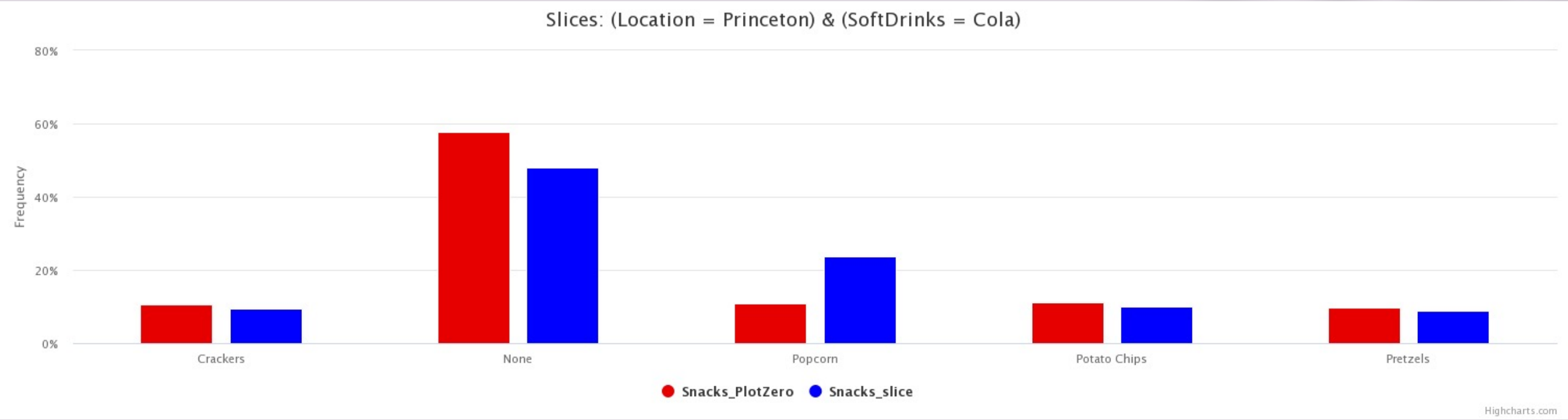
By Fauzan Amjad

The Dataset

- We’re provided a rich dataset with 20,000 columns with 7 columns of categorical variables.
  - Each row represent a buyer
  - The categorical variables are beer, day, location, soft drinks, sweets, wine, and snacks.
- Our goal is to find patterns or trends and then run tests to see if their significant.
  - We’ll first use Boundless Analytics to find these potential trends
  - We will then test to see if these trends are statistically significant

Boundless Analytics Plot

- Attribute -> Snacks
- Blue Slice -> Location = Princeton and SoftDrinks = Cola



Analyzing the Boundless Analytics Plot

- I set the attribute I wanted to analyze to be snacks.
- Boundless analytics made approximately 438 plots; however, we'll predominantly be toning on the data regarding popcorn.
- With the general population (the entire dataset), the frequency of of popcorn is approximately 10.95%.
- When we subset the population to only include people who live in Princeton and to people who got Cola as their soft drink, the frequency of popcorn becomes 23.78%
  - That is more than double the frequency of popcorn of the general population.

Null and Alternate Hypothesis

- Null Hypothesis
  - Buyers of Cola who live in Princeton buy popcorn at the same frequency as the general population.
- Alternate Hypothesis
  - Buyers of Cola who live in Princeton do not buy popcorn at the same frequency as the general population.

Bonferroni Correction

- The issue with using 0.05 as our p-value threshold is that the more hypotheses that we potentially test out, the more of a chance we will come across a rare event.
  - In a scenario like this where we keep 0.05 as our threshold, we could actually incorrectly reject the null hypothesis.
- In order to Bonferroni correct, we have to divide 0.05 by the number of tests we can potentially form.
  - Boundless analytics gives us 438 plots and each plot has 5 graphs .
  - We can run 2190 tests!
- Our new threshold is 2.28 * 10^(-5)
  - 0.05/2190

Permutation Test

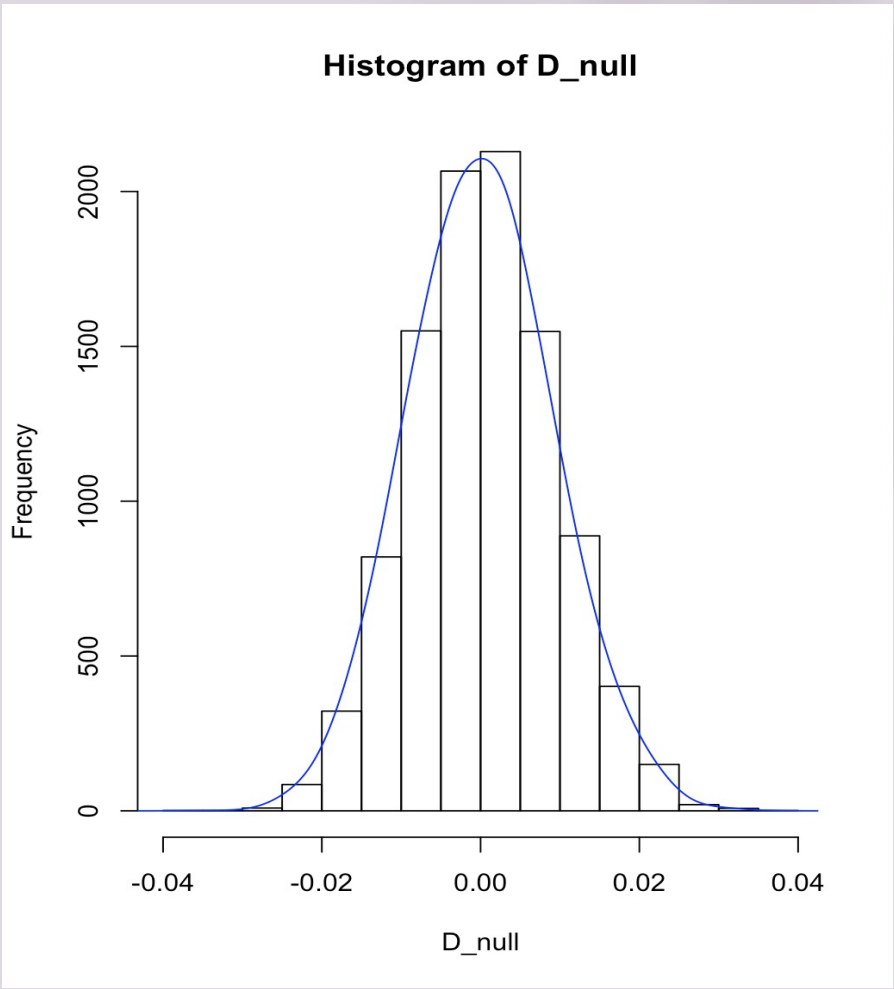
```
> PermutationTestSecond::Permutation(data,"slice", "bar", 10000, "0", "1" )
[1] 0
```

```
library(readr)
data <- read_csv("Desktop/HomeworkMarket.csv")


data$bar <- 0
data$slice <- 0


data[data$Snacks == "Popcorn",]$bar <- 1
data[data$Location == "Princeton" & data$SoftDrinks ==
        "Cola",]$slice <- 1


PermutationTestSecond::Permutation(data,"slice", "bar",
        10000, "0", "1" )
```



Analysis of Results

- From our permutation test, we got a p-value of 0.
- The p-value tell us the likelihood how we could observe the current values under the assumption that the null hypothesis is true.
- Our p-value is under our Bonferroni corrected significance level.
- Conclusion
  - We reject the null hypothesis.
  - We accept the alternate hypothesis: Buyers of Cola who live in Princeton do not buy popcorn at the same frequency as the general population.

Exploring the Data More

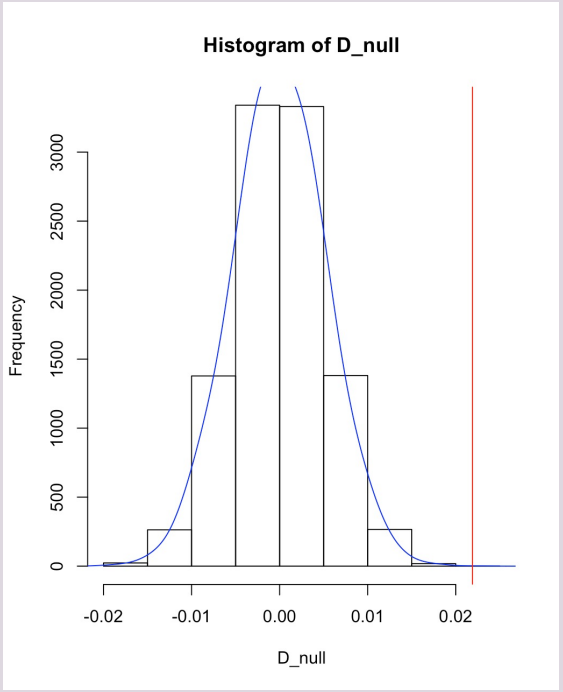
- We just just accepted the alternate hypothesis: Buyers of Cola who live in Princeton do not buy popcorn at the same frequency as the general population.
- What would happen if we would only subset the data with people with people who live in Princeton? What if we did the same with people who buyed Coca-Cola?
- We already know, based on the Boundless Analytics plots, that
- These are tests we can further explore.

New Hypotheses

- Permutation Test #2
  - Null Hypothesis: Buyers who live in Princeton buy popcorn at the same frequency as the general population.
  - Alternate Hypothesis: Buyers who live in Princeton do not buy popcorn at the same frequency as the general population.
- Permutation Test #3
  - Null Hypothesis: Buyers of Cola buy popcorn at the same frequency as the general population.
  - Alternate Hypothesis: Buyers of Cola do not buy popcorn at the same frequency as the general population.
- We'll use the same Bonferroni Correction.

Permutation Test #1

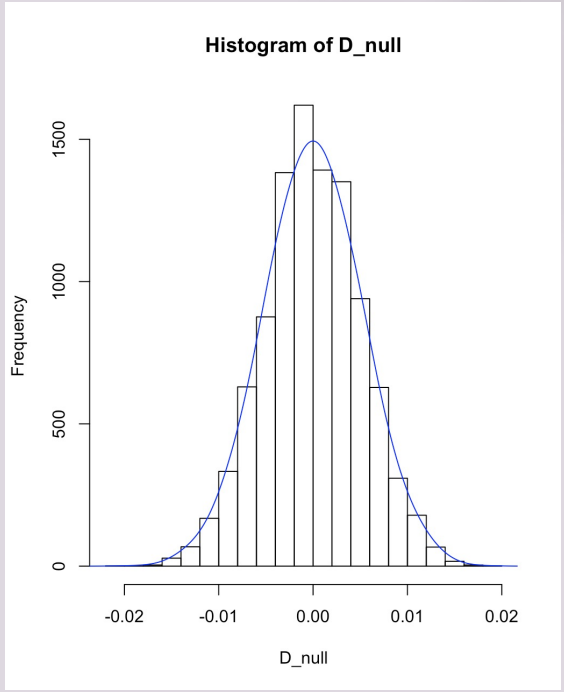
```
data$bar <- 0
data$slice <- 0
data[data$Snacks == "Popcorn",]$bar <- 1
data[data$Location == "Princeton",]$slice <- 1
PermutationTestSecond::Permutation(data,"slice", "bar", 10000, "0", "1" )
```




```
> PermutationTestSecond::Permutation(data,"slice", "bar", 10000, "0", "1" )
[1] 1e-04
```

Permutation Test #2

```
data$bar <- 0
data$slice <- 0
data[data$Snacks == "Popcorn",]$bar <- 1
data[data$SoftDrinks == "Cola",]$slice <- 1
PermutationTestSecond::Permutation(data,"slice", "bar", 10000, "0", "1" )
```




```
> PermutationTestSecond::Permutation(data,"slice", "bar", 10000, "0", "1" )
[1] 0
```

Conclusions From Our Permutation Tests

- Permutation #2
  - The p-value came out to 10^(-4). This is greater than our Bonferroni corrected p-value.
  - We fail to reject the null hypothesis: Buyers who live in Princeton buy popcorn at the same frequency as the general population.
- Permutation #3
  - The p-value came out to 0. This is less than our Bonferroni corrected p-value.
  - We reject the null hypothesis.
  - We accept the alternate hypothesis: Buyers of Cola do not buy popcorn at the same frequency as the general population.

Suggestions Based on Our Tests

- In the Princeton location, if someone buys Cola as their soft drink, a significant portion of them will be getting popcorn with it is as well - a higher frequency than the general population..
- If someone is buying Cola as their soft drink, irrespective of where the location might be, they a significant portion of them will be getting popcorn with it as well – a higher frequency than the general population.