Analyzing Natesh’s Coffee Consumption

New York City | Team #1

Rasheed Sabar, Sham Mustafa, Haris Jaliawala, John Benjamin

# Business Problem

Correlation One is a company that assesses, connects, and trains advanced data scientists across the globe. It is crucial for Correlation One to gain a better understanding of Natesh’s coffee consumption in order to deliver better outcomes to its students. In particular, we are interested in understanding **what are patterns in Natesh’s coffee consumption**, and ultimately, **what levels of coffee provide optimal lecture delivery?**

# Business Impact

With optimal coffee consumption, experts at Correlation One estimate that Natesh can speak up to 5x faster, thereby delivering 5x more jokes and content, on average. Experts say Natesh’s jokes are expected to increase student satisfaction and participation by 20%, on average. Delivering 5x more content will also allow Natesh to review previous concepts more frequently and cover more advanced topics. Therefore, understanding Natesh’s coffee consumption and predicting the optimal amount prior to each lecture could have a substantial impact on student outcomes.

# Data

The dataset consists of four data points with the following fields:

date (DateTime): date of lecture

num\_coffee (Int): number of coffee cups Natesh consumed prior to lecture

avg\_wpm (Float): average number of words/minute during lecture

live\_chat (Dict[String]): a dictionary of all the student chats in #live-chat during that lecture

One advantage of this dataset is that it provides rich information on student chats for every single lecture. One disadvantage of this dataset is that it contains relatively few - only four - data points, which is a very small sample size. Another complication is that #live-chat was created during the fourth lecture, which means that information is missing for the previous lectures. This means that the advantage of the dataset actually isn’t an advantage at all, and this is just all in all a pretty mediocre dataset.

# Methods

**Visualizations**

One key component of understanding the patterns in Natesh’s coffee consumption is providing high-quality visualizations of his historical coffee consumption. Here are some of the static and interactive visualizations we will provide:

* Natesh’s coffee consumption over time (univariate)
* Student outcomes of interest over time (univariate)
* Correlation plot of Natesh’s coffee consumption vs. student outcomes
* Heatmap of where Natesh’s coffee was sourced from

**Models**

Another key component of our project is determining the association between Natesh’s coffee consumption and key student outcomes (satisfaction, learning, engagement… etc.). Ideally, we will be able to identify the optimal amount of coffee to maximize these outcomes of interest. These are some of the methods we will be exploring:

* **polynomial regression**: since we expect a non-linear relationship between coffee consumption and student outcomes (i.e., too much coffee might impact Natesh’s ability to focus).
* **heterogeneous treatment effect**: since different students may respond to Natesh’s coffee consumption in different ways, we want to understand both the outcomes on average, but also within different subpopulations of students.
* **neural network:** eh, why not.

# Interface

The final front-end product will feature two landing pages: an **Analytics** page with visualizations of the historical data, and a **Recommendation** page, where the results are summarized with a recommendation for the number of coffee cups Natesh should consume prior to lecture. The interface will allow for interactive visualizations of the historical data, so that users can click on a particular lecture day and see the results for that day, or all outcomes over time. If time permits, we will also try to include a feature in the **Recommendation** section where a user can see what the effect is predicted to be if Natesh consumes X% more or less coffee than recommended.



This is an example of what our dashboard could look like: (1) a panel with key metrics on student engagement, (2) a plot of Natesh’s coffee consumption over time, (3) a plot summarizing the metrics of student engagement for each lecture day, (4) plots highlighting key features of messages in the #live-chat channel.

# Milestones

In this section, we provide details on the milestones we intend to achieve in our project. In particular, we have outlined four different versions: we expect to finish Version 1 with 100% probability, Version 2 with ~70% probability, Version 3 with ~20% probability, and Version 4 with ~5% probability (if things go extremely well). We have color-coded our versions as follows: data, analysis, and visualizations, to make it clear that our versions advance on all levels.

Version 1: Build simple dashboard (2 static plots, 2 interactive plots) to understand Natesh’s historical coffee consumption and its correlation with 3 student outcomes of interest (engagement, satisfaction, and performance)

Version 2: Build prediction model (+ 1 static plot, 1 interactive plot) and determine what level of coffee consumption is optimal

Version 3: Build appropriate causal inference model to assess the strength of the causal relationship between the factors of interest (include key assumptions in the dashboard)

Version 4: Use additional data from Natesh’s classes at Harvard and/or DS4A 2019 to stress-test the model

# Timeline

|  |  |  |
| --- | --- | --- |
| **Date** | **Deliverable** | **Details** |
| **Week 1** | Team formation  Environment setup |  |
| **Week 2** | One-page summary  Workflow setup | *This includes getting data access, e.g. NDAs.* |
| **Week 3** | Scoping document  Data access | *The scoping document is the skeleton of the Final Report.* |
| **Week 4** | Data cleaning  Initial data exploration | *Update Final Report to include EDA results.* |
| **Week 5** | Continue data exploration | *Get initial code reviewed by TA as well.* |
| **Week 6** | Advanced data exploration  Initial modeling | *Begin front-end visualization.* |
| **Week 7** | Continue modeling  Application on cloud | *Update Final Report to include initial modeling results.* |
| **Week 8** | Front-end complete  Advanced modeling | *Update Final Report to include final modeling results.* |
| **Week 9** | Fine-tune modeling  Fine-tune application | *Write Conclusions section of Final Report.* |
| **Week 10** | Finalize presentation, report, and application |  |

# Concerns

The primary concerns with our project are (1) that some of our team members do not know anything about data analysis, and (2) that we have basically no data.