Deliverable 3

Introduction

In Deliverable 3, our group follows the checklist in Navigating Spark Projects **strictly**.

After thorough discussions, our group has decided to break down the checklist into the following steps:

1. **Merge and normalize the new dataset & update the codebook**
2. **Refine and amplify the question analysis of Deliverable 2**
3. **Finish the final project draft**
4. **Refine project scope**
5. **Wrap up everything to a report and submit to the repo**

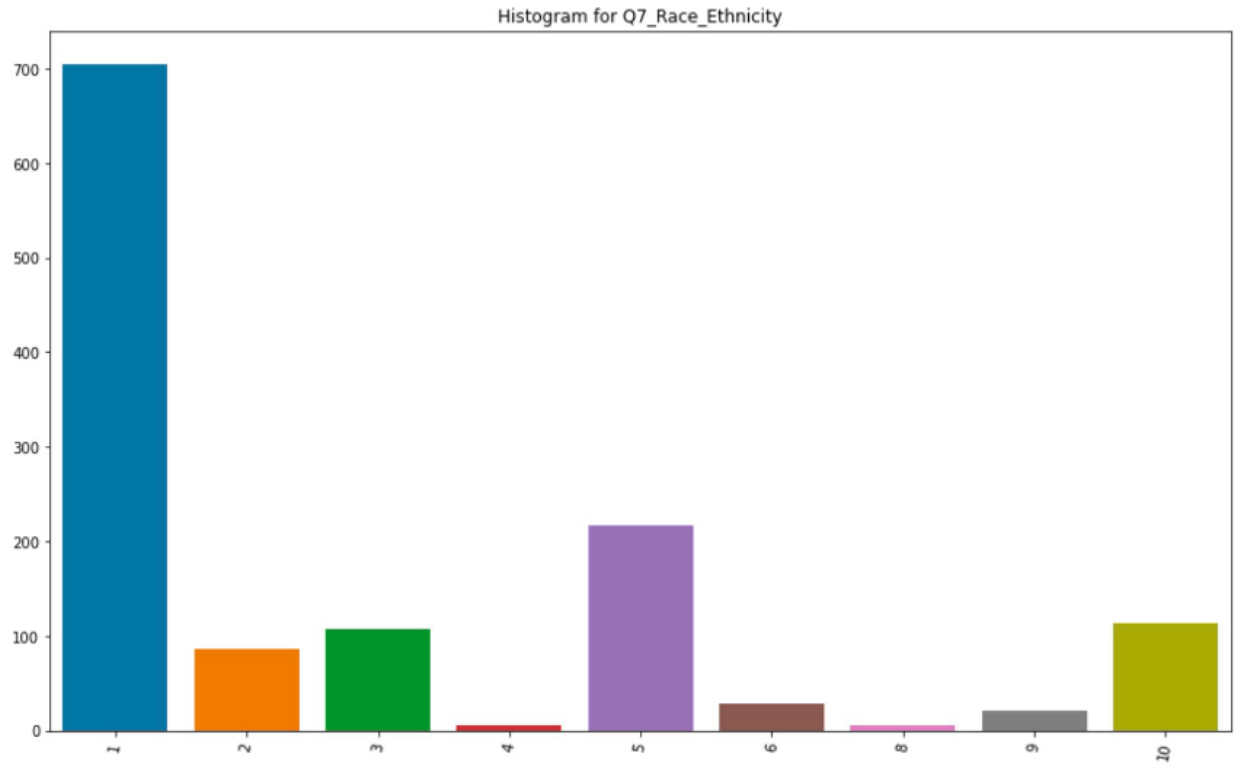
Merge and normalize the new dataset

Last week, after finishing almost all the work for the previous dataset, the clients have given us a new dataset. Thus, we have done another round of merging and normalizing on the new dataset. We also updated a new codebook for the processed dataset.

Refine and amplify the question analysis of Deliverable 2

More plots and detailed analysis are shown in code.

In Deliverable 1, we have answered the question "Are there industry differences in representation in advertisements?" with regards to columns "Q2_Age", "Q3_Sex", and "Q4_Gender". This time, we have shifted our focus on another important column, namely "Q7_Race_Ethnicity"



For the Q7_Race_Ethnicity column, we find that value "1"("White") is the largest, the second one and third one are "5"("Asian/Asian American") and "10"("Not Applicable"). We can find that the number of whites is greater than the sum of the numbers of all other races.

Industry differences of the Q7_Race_Ethnicity column

We can find that among the Q7_Race_Ethnicity = 1("White"), the number of "Confectionary" is 310 and "Petcare " is 263 which are far more than "Wrigley" and "Food" industry. Majority of "White" work in these two industries.

We can find that among the Q7_Race_Ethnicity = 2("Hispanic/Latino"), the "Confectionary" is much higher than other industries, each industry has "Hispanic/Latino", most of "Hispanic" are in "Confectionary" industry. We can find that among the Q7_Race_Ethnicity = 3("Black"), the "Confectionary" is also higher than other industries, each industry has "Black". The portion of "Petcare" for "Black" is the highest compared with other ethnicities.

We can find that among the Q7_Race_Ethnicity = 4(Native American/Hawaiian/Alaskan/Pacific Islander), there is a small amount of data. We can find that among the Q7_Race_Ethnicity = 5("Asian/Asian American"), there are a very high number of people working in the "Confectionary" industry.

We can find that among the Q7_Race_Ethnicity = 6("Middle Eastern"), no one is in the "Food" industry. The ratio of "Confectionary" among industries for "Asian" is the highest compared with other ethnicities.

We can find that among the Q7_Race_Ethnicity = 8("Mixed Race"), there are small amount of data, the number of "Confectionary" is 3, for "Petcare" is 2, for "Wrigley" is 1, no one is in "Food" industry.

We can find that among the Q7_Race_Ethnicity = 9("Can't tell"), the number of data is relatively small, the number of people for "Confectionary" and "Food" is the same.

We can find that among the Q7_Race_Ethnicity = 10("Not Applicable"), the "Confectionary" is higher than other industries, people are only in the "Confectionary" and "Food" industry.

By the statistics, we observe that the largest amount of data whatever their race/ethnicity are in the "Confectionary" industry. And many of people are "Not Applicable" of Q7_Race_Ethnicity in "Confectionary" and "Food".

Finish the final project draft

Our team has reviewd the whole semester work and design the final project draft. Below is the skeleton of our draft (please visit our repo to see the **whole** report):

- Background
- Motivation
- Goal
- Exploration
- Analysis

Refine project scope

- For the new dataset, the current questions are not applicable. We want to clarify that **what analysis should we do for the new dataset.**

P.S. Already got reply from clients: "The main goal is to merge and normalize and if you have more time you can also answer the key questions based on the new dataset."

Wrap up everything to a report and submit to the repo

The final submission includes:

- **this report**,
- **a folder called "Cannes", containing the merged datasets, codebooks, and the code**,
- **a folder called "Refine Mars 2020-2021, containing our work for refining the analysis in Deliverable2**,
- **the pdf for the final report draft**,
- **a readme file containing the links to the merged dataset and codebook for the previous datasets (i.e. Mars) and links to the merged dataset and codebook for the new datasets (i.e. Cannes). (So 4 links in total)**.

They all locate in **Deliverables/Deliverable 3** folder of our repo.