Deliverable 2

> ## Introduction
>
> In Deliverable 2, 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. **Update the codebook for the new merged dataset**
> 2. **Refine and amplify the question analysis of Deliverable 1**
> 3. **Answer the final key question: What are the trends?**
> 4. **Refine project scope**
> 5. **Wrap up everything to a report and submit to the repo**

Update the codebook for the new merged dataset

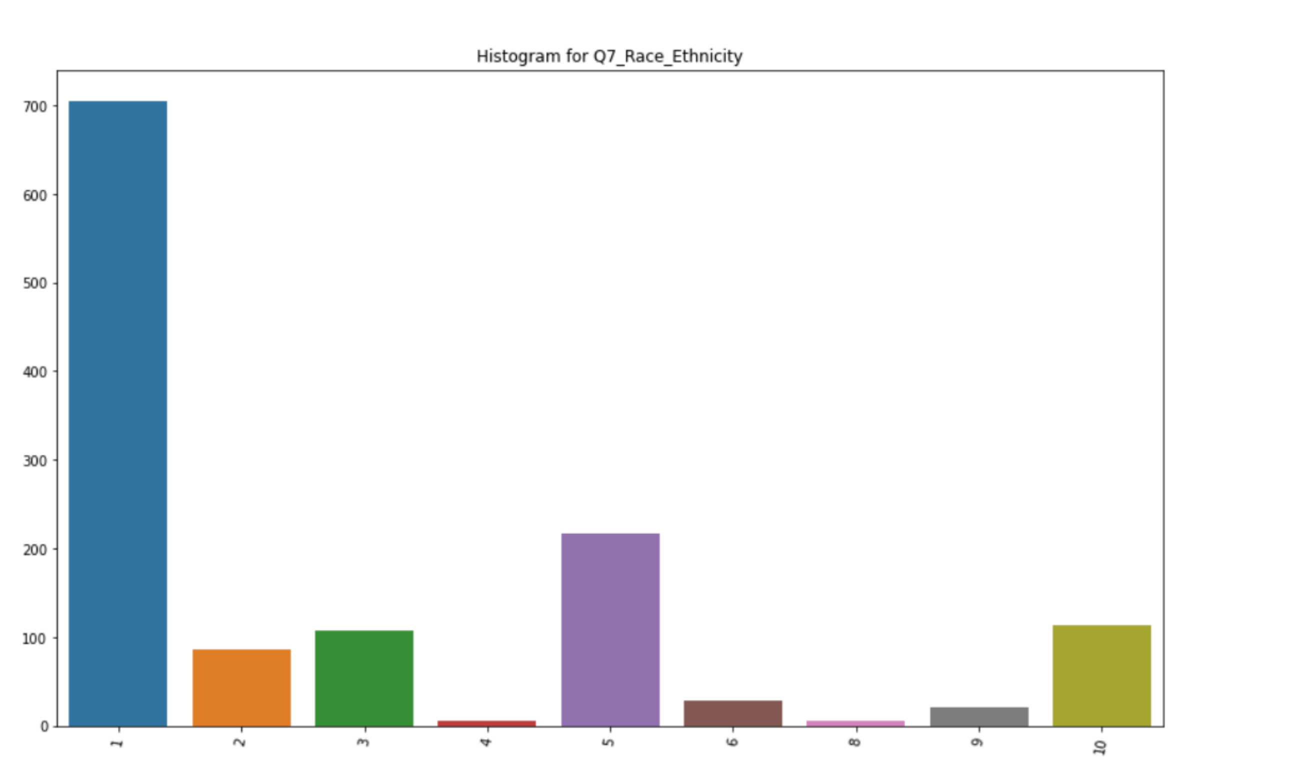
Since we have merged and normalized the two datasets into one during Deliverable 1, our group realized that the codebook should also be updated to adapt to the new dataset. Thus, in Deliverable 2, we have "normalized" the two codebooks into one, which contains **tailor-made** explanations of the meanings of each column and its values in the new dataset.

Refine and amplify the question analysis of Deliverable 1

Note: More plots and detailed analysis are shown on code

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

First we have an observation for the new column



1. For **Q7_Race_Ethnicity** column, we find that value "1"("White") is the largest, the second one and thrid one are "5"("Asian/Asian American") and "10"("Not Applicable").

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 comparing with other ethnicities

We can find that among the Q7_Race_Ethnicity = 4(Native American/Hawaiian/Alaskan/Pacific Islander), there are small amount of data

We can find that among the Q7_Race_Ethnicity = 5("Asian/Asian American"), there are very high number of people working in "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 comparing 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"), number of data is relatively small, the number of people for "Confectionary" and "Food" is same.

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

By the statistics, we observe that the largest amount of data whatever their race/ethnicity are in the "Confectionary" industry

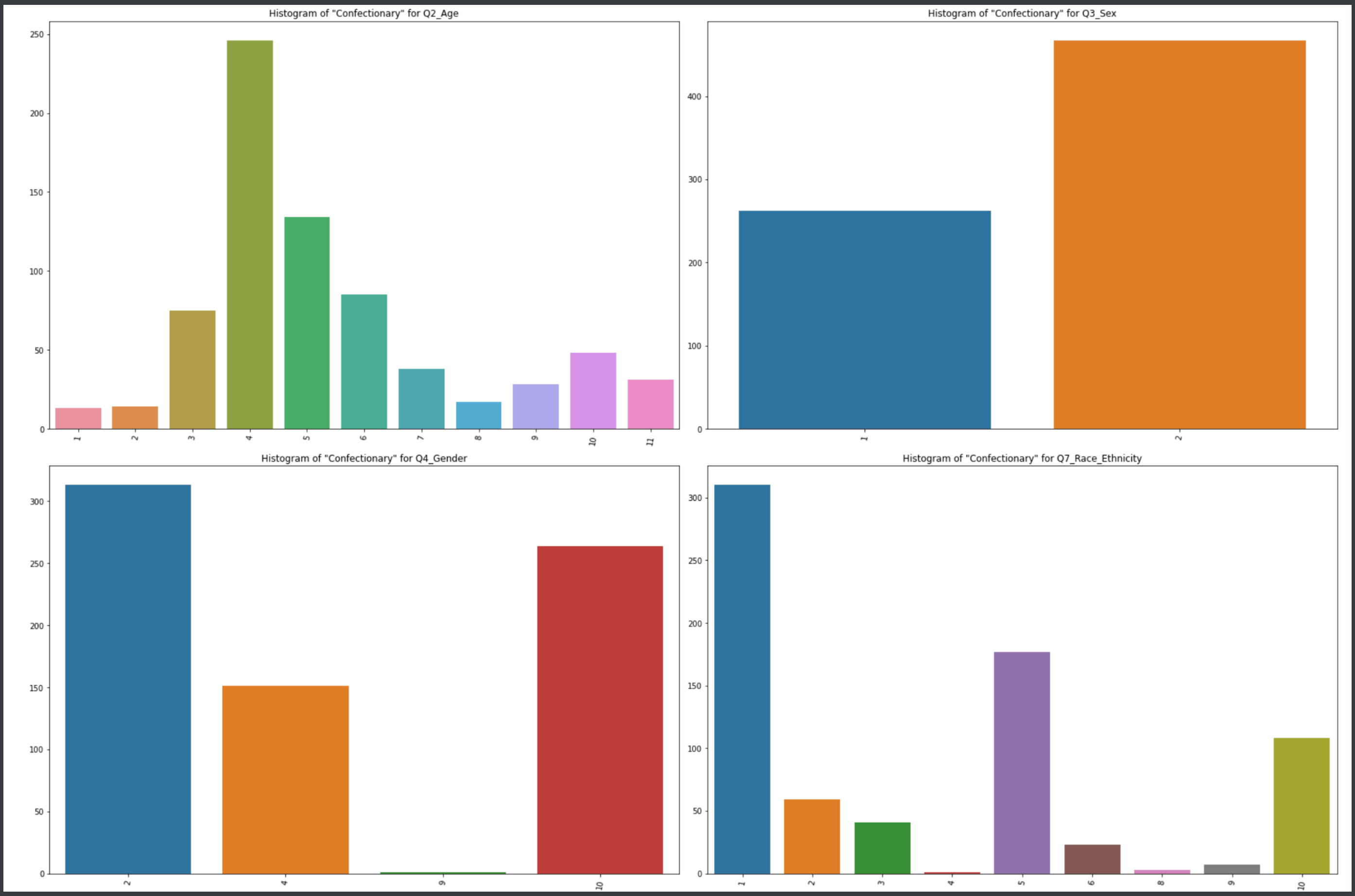
And many of people are "Not Applicable" of Q7_Race_Ethnicity in "Confectionary" and "Food"

Observation of industry differences for confectionary dataframe:

We can find that for Confectionary, the 20-29 age group(4) has the largest proportion, the second one is group(5)

Through Q3_Sex and Q4_Gender, we can find that the number of male is definitely greater than female and lots of people choose "Not Applicable" on Q4_gender

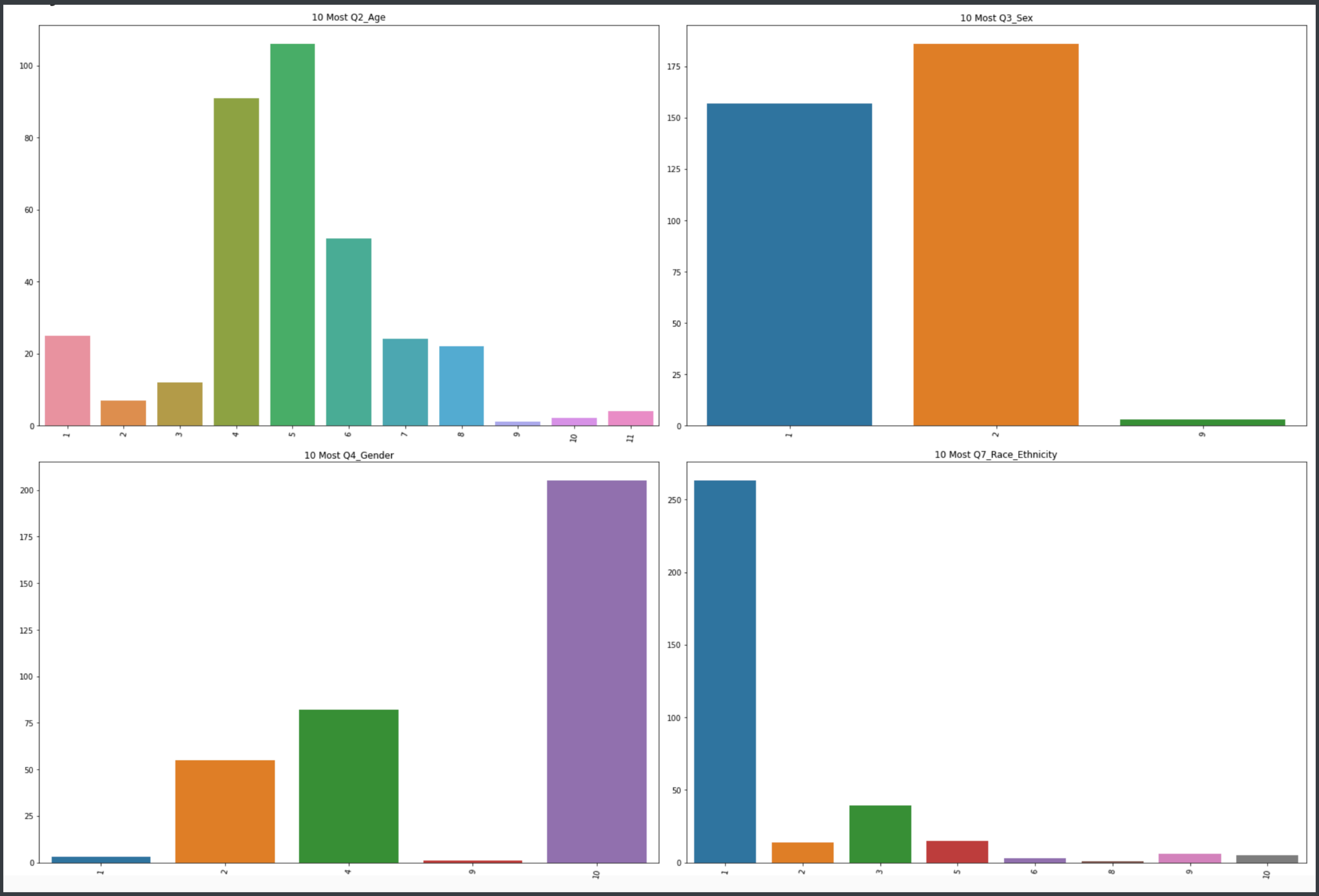
From the histogram, we observe that "White" has the largest number of any race in "Confectionary" industry, and there are large amount of "Black" and "Not Applicable". Such a data distribution is almost consistent with the proportion of racial populations in the United States



Observation of industry differences for Petcare dataframe:

We can find that for "Petcare", the 30s age group(5) has the largest proportion, the second one is group(4)

For "Percare", compared with other races, "White" is the largest

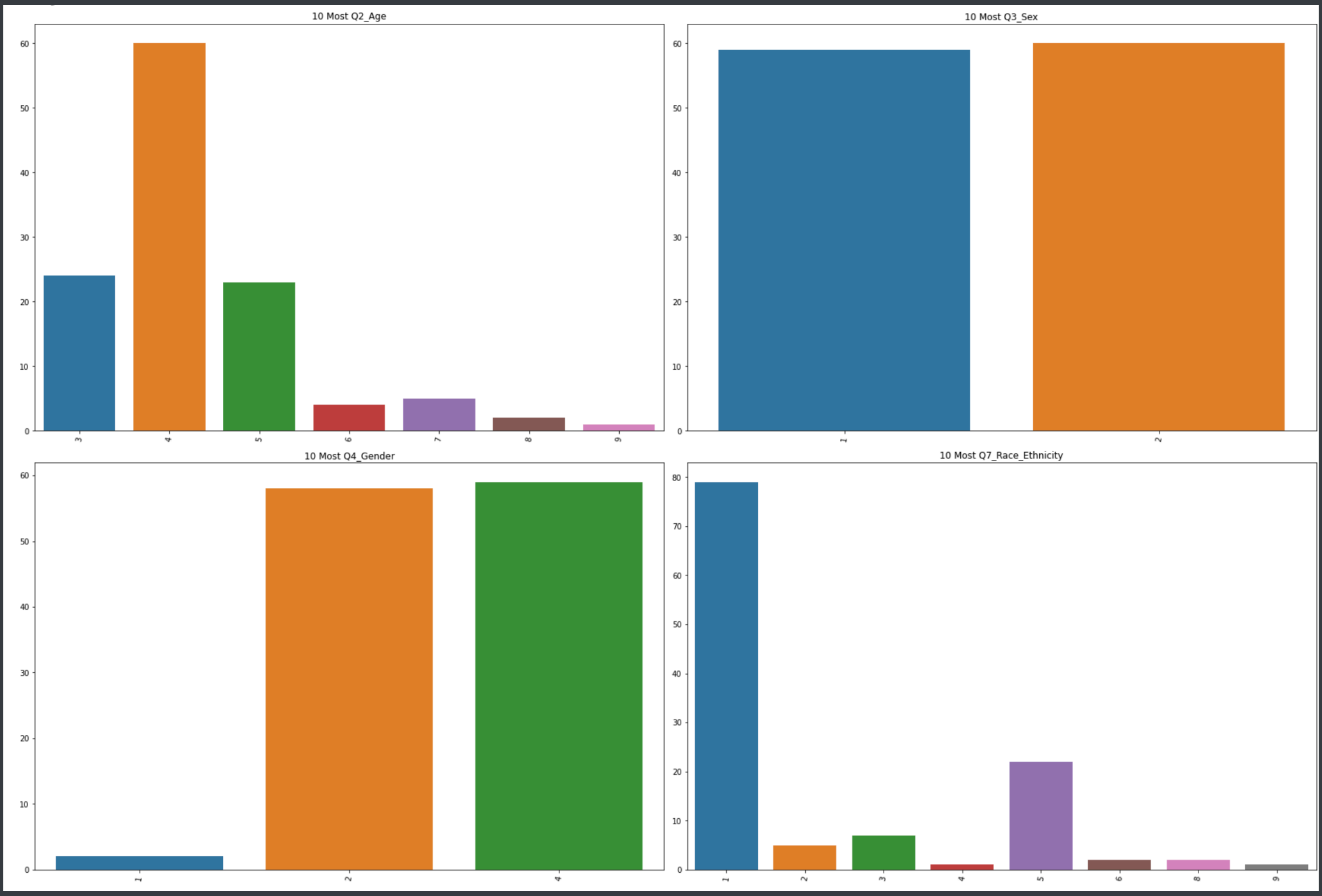


Observation of industry differences for Wrigley dataframe:

We can find that for Petcare, the 20s age group(4) has the largest proportion, the second one is group(3), but number of group(5) is close to group(3)

For the gender column of Wrigley, there are small portion of "Hyper-Masculine(1)", the number of "Masculine(4)" and "Feminine(2)" is close, which is same as sex column

For "Wrigley", compared with other races, "White" is the largest

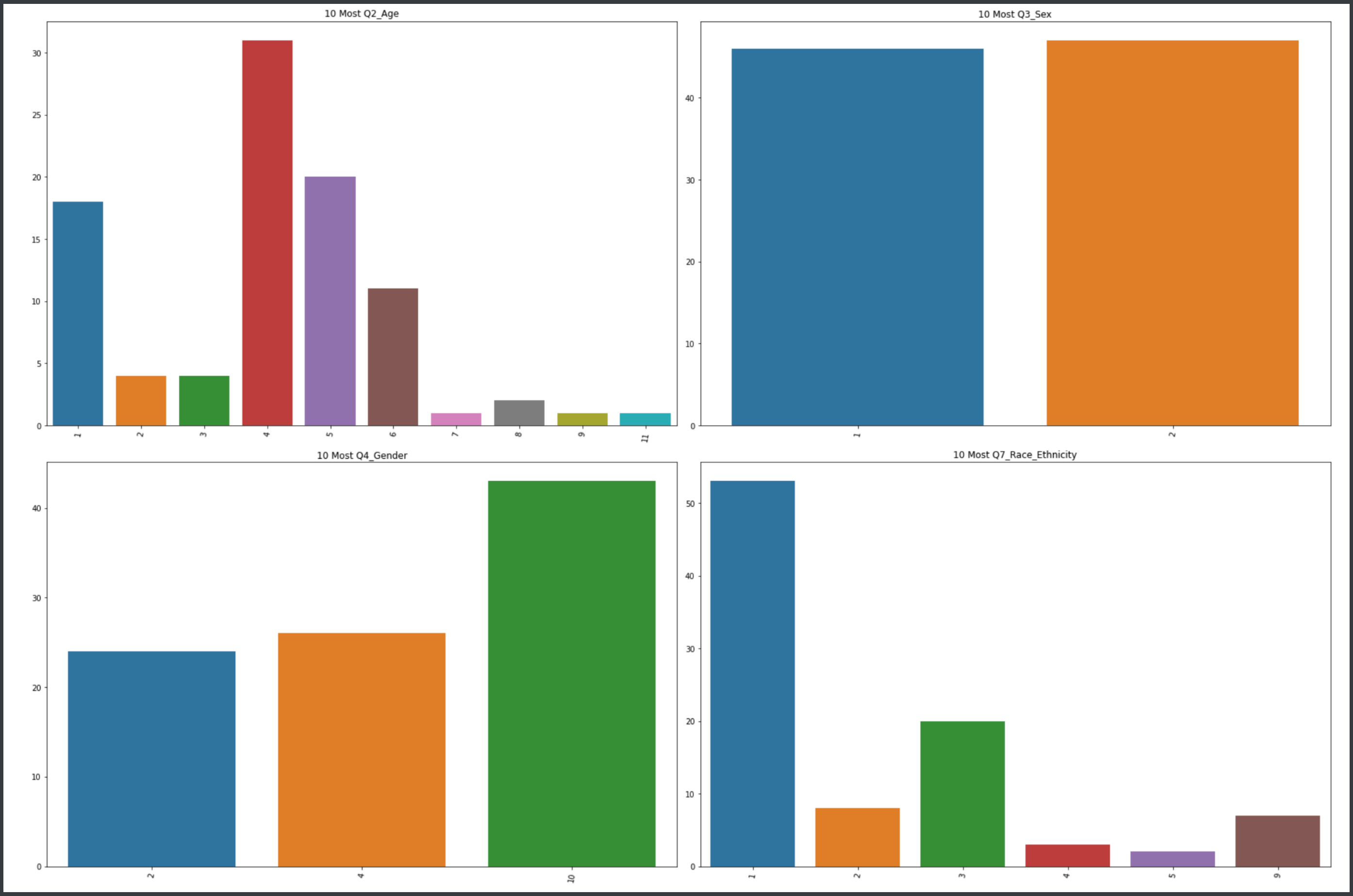


Observation of industry differences for Food dataframe:

We can find that for Petcare, the 20s age group(4) has the largest proportion, the second one is group(5), and the number of group(1) is close to group(5)

For the gender column of Wrigley, there are large portion of "Not Applicable(10)", the number of "Masculine(4)" and "Feminine(2)" is close, which is same as sex column

For "Food", compared with other races, **"White" is the largest**



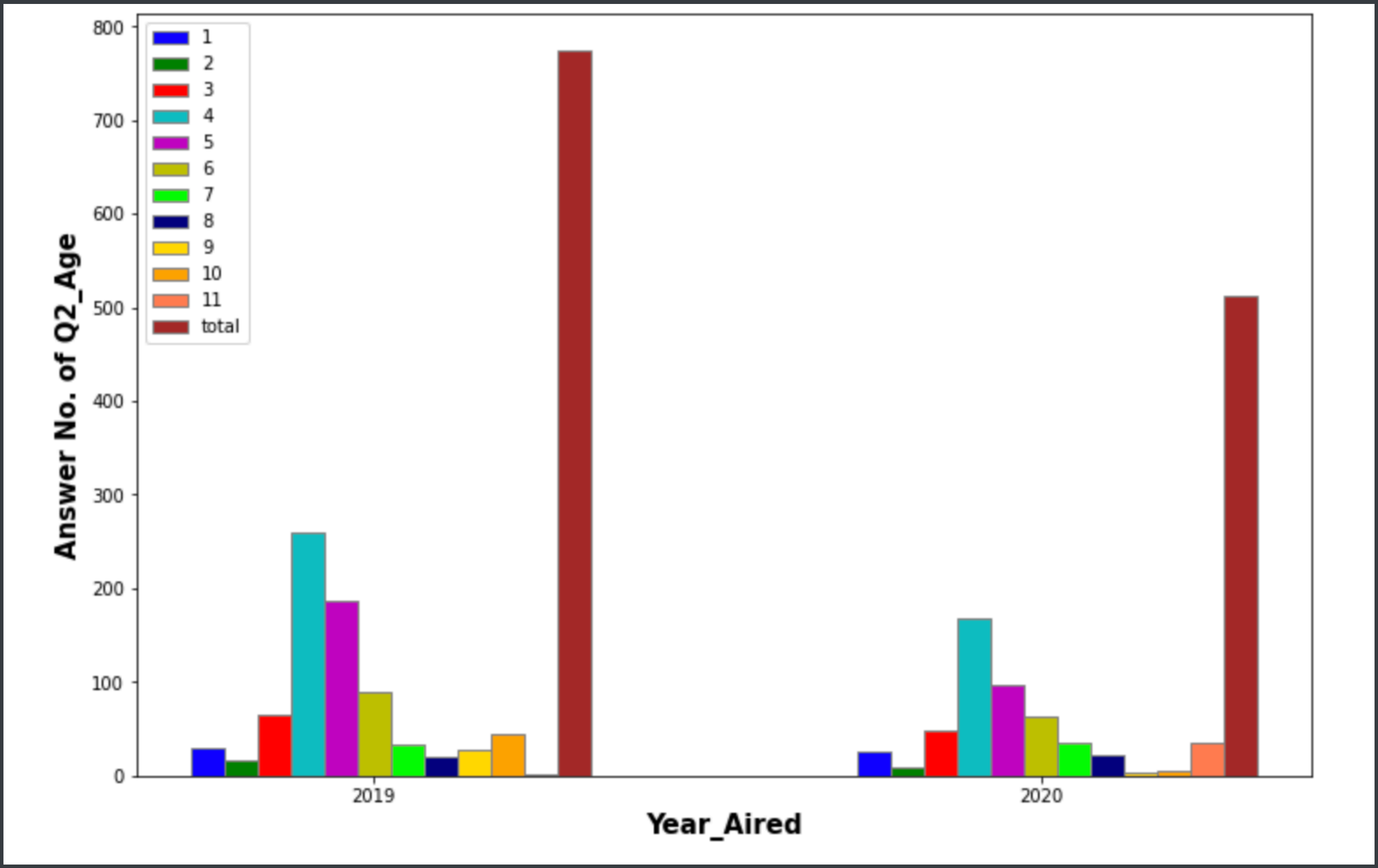
Answer the final key question: What are the trends?

Note: Plots and detailed analysis are showing on code

Our group tackled this question based on our **previous** answer to question "**Is there a change in representation in advertisements over time?**" Our interpretation of "trends" is: **the data moving tendency in terms of the columns (e.g. age, sex) over time (i.e. year_aired).**

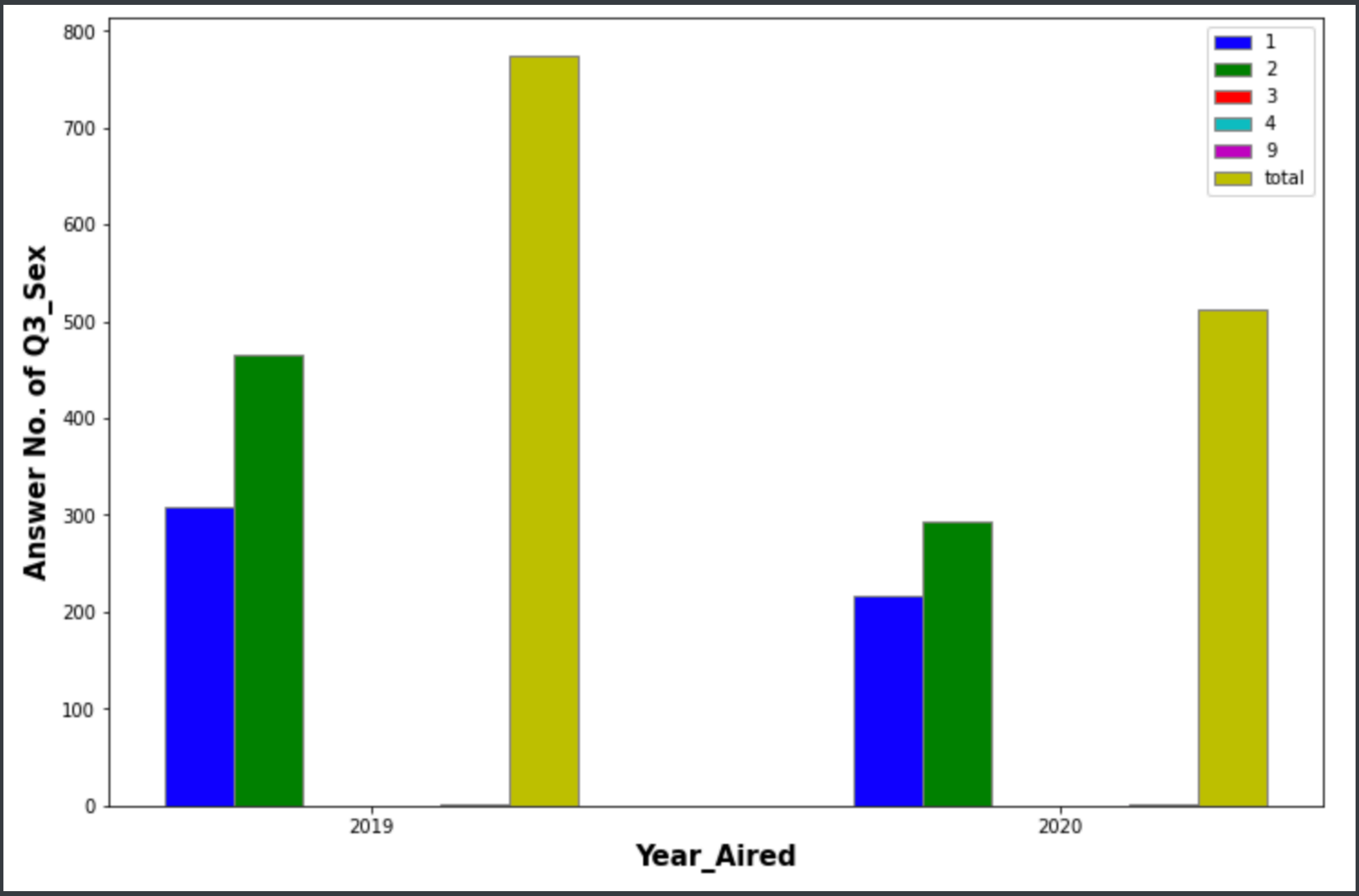
Trends of Q2_Age:

1. The total number **decreases** from 2019 to 2020
2. Q2_Age = 4 is the largest value between two years
3. Someone don't tell or are unwilling to disclose the information of age
4. The distributions of ages between two years are almost the same.



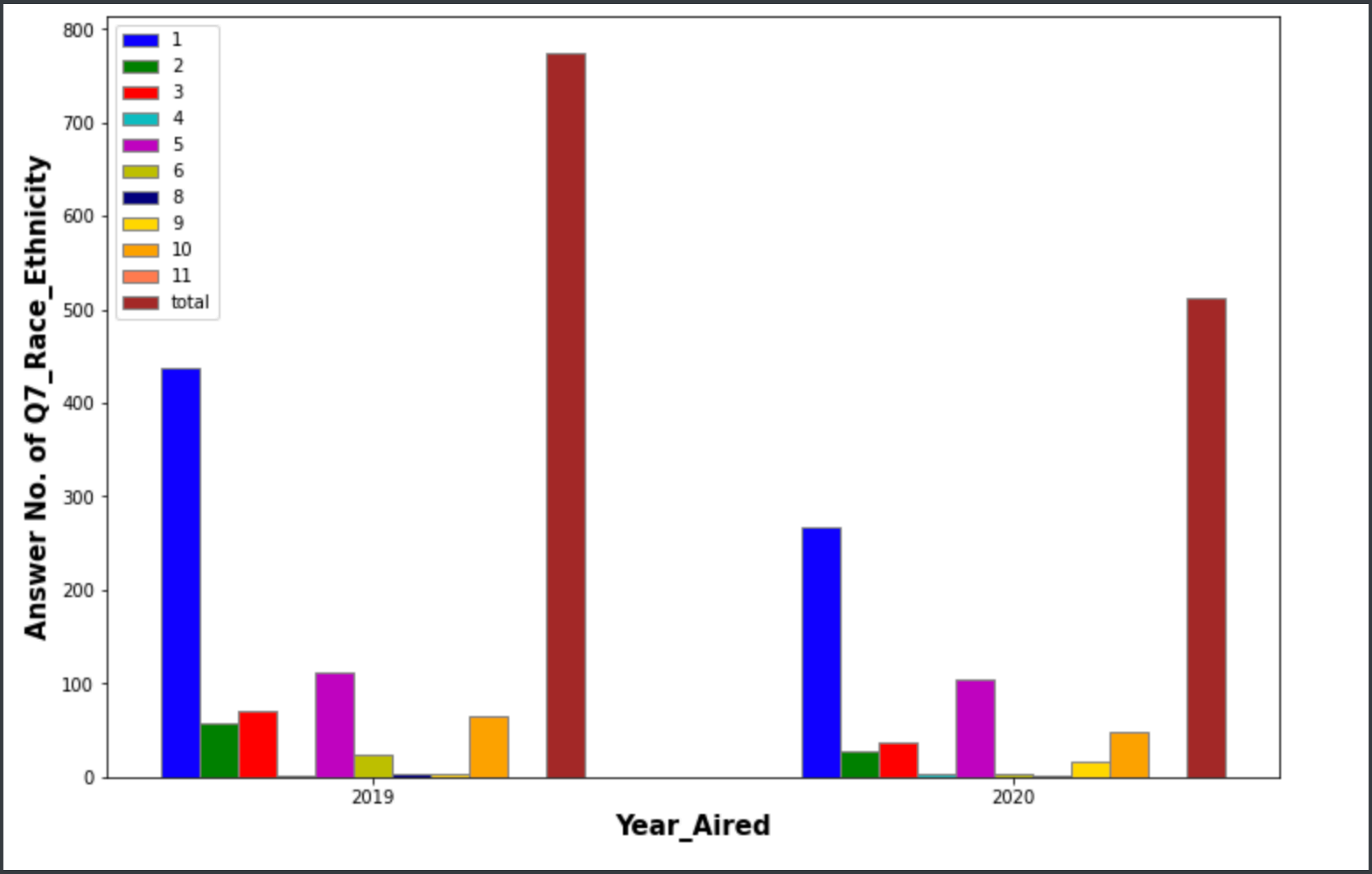
Trends of Q3_Sex:

1. We find that the total number **decreases** from 2019 to 2020
2. Male numbers are **larger** than female numbers in both two years
3. And some of people can not tell their sex



Trends of Q7_Race_Ethnicity:

1. We find that the total number of Q7_Race_Ethnicity decreases from 2019 to 2020
2. Among all different answers, option 1, which refers to "white people", is the largest in both two years
3. The total number of rest options is almost the same as the number of option 1.



Refine project scope

1. We want to further clarify **what are and how many are the questions to be analyzed.**

P.S. Already got reply from clients: "the students can look at whatever variables they want. We tend to focus on gender (differences between men and women) and race (differences between white characters, and characters of color). This isn't a priority for us, we wanted the students to attack this however they want!"

2. Are we going to have more datasets to come and analyze more (key) questions of the project? Because all questions listed in the project proposal will be answered after Deliverable 2.

P.S. Already got reply from clients: "we do have another set of datasets (like this round, its all the same partner, but multiple codebooks and data) we would like normalized. Should we send those, next week?"

Wrap up everything to a report and submit to the repo

The final submission includes **this report**, **Merged_Preprocessed_Mars2020_2021.csv**, **Merged Codebook 2020-2021.docx**, "**Question 1 - refine.ipynb**". They all locate in **Deliverables/Deliverable2** folder of our repo.