Deliverable 1

Introduction

In Deliverable 1, 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. **Pre-process the data**
2. **Answer question: Are there industry differences in representation in advertisements?**
3. **Answer question: Is there a change in representation in advertisements over time?**
4. **Wrap up everything to a report and submit to the repo**

Pre-process the data

The general idea in this step is to first **normalize** the data and columns in each individual dataset (Mars2021_Data.xlsx and Mars2020_Data.xlsx) according to the codebooks, and **merge** them into one (Merged_Preprocessed_Mars2020_2021.csv).
We did this step-by-step:

1. Process Mars2021_Data.xlsx:
    - Process “Q4_Gender”, “Q7_Race_Ethnicity” to make its values match the codebook.
    - Merge “Q8_Physical_Disability”, “Q9_Cognitive_Disability”, and “Q10_Communication_Disability” into one since we think it is redundant to have “disability” distributed in three columns.
    - Rearrange the rest of the columns and drop unrelated ones.
    - For the “Year_Aired” column, we only keep values that are equal to 2020 (according to Meredith Conroy).

- For “Segment” column, rename all “Pet” to “Petcare”.
- Save the results to a csv file.

2. Process Mars2020_Data.xlsx:
    - Steps are basically the same as above.
    - For the “Year_Aired” column, we only keep values that are equal to 2019 (according to Meredith Conroy).

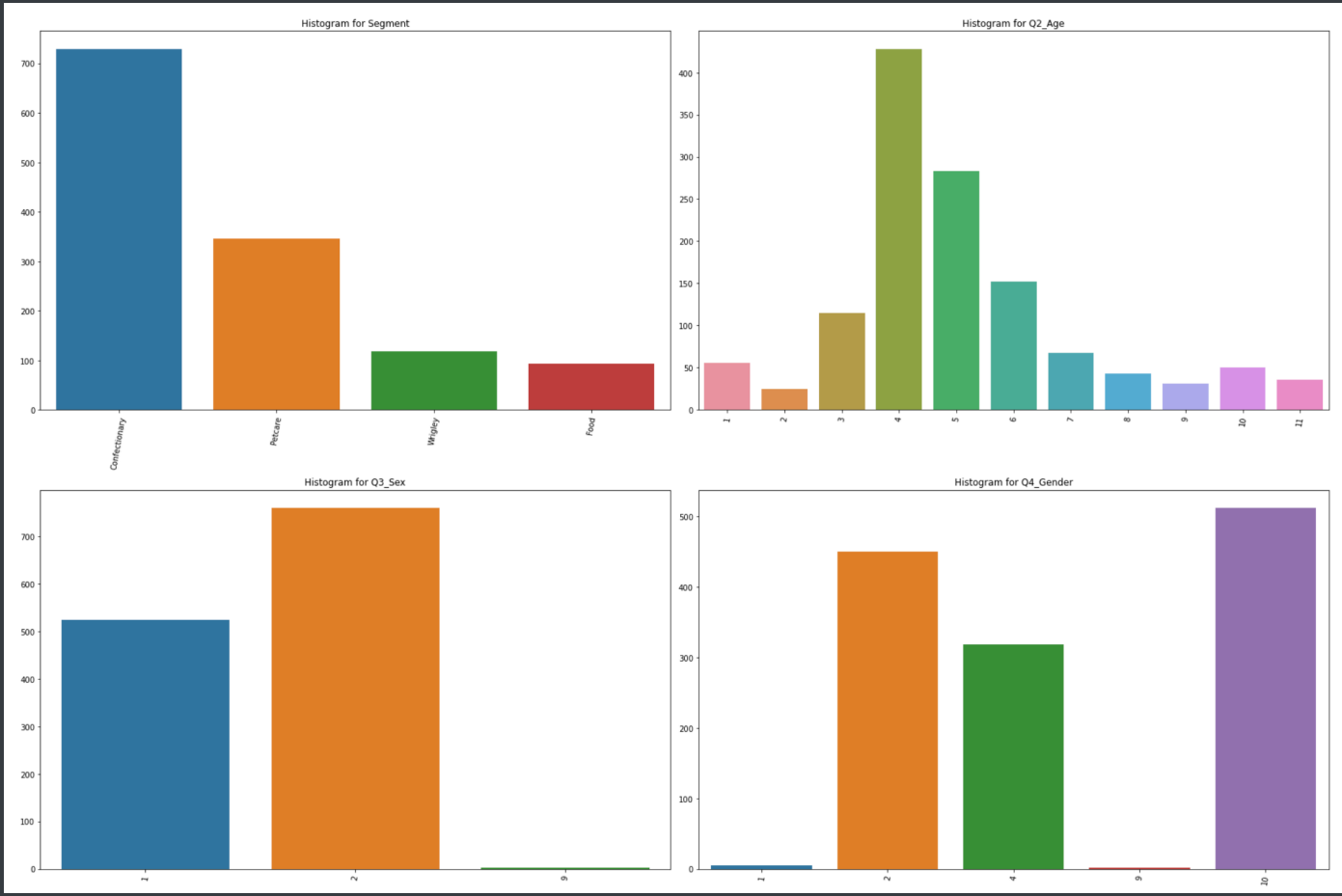
3. **Merge** the two refined datasets and make sure everything is **normalized** in the final dataset, namely Merged_Preprocessed_Mars2020_2021.csv.

Answer question: Are there industry differences in representation in advertisements?

Note: More plots and detailed analysis are shown on code

First we have an observation for entire dataset

1. For Segment column, we can find the 'confectionary' accounts for half of the total data volume of Segment column
2. For Q2_Age column, most of values are Q2_Age = 4, which is the age of 20-29 year olds
3. For Q3_Sex column, number of male is more than female, and there is a little people don't tell their sex
4. For Q4_Gender column, we find that value 'Not Applicable' is the largest, and then is the 'Masculine' and 'Feminine'. Also, there are a few 'Hyper-Masciline' and 'Can't Tell'



Since “industry” refers to the “Segment” column, and “representation” refers to all the question columns. We find the industry differences among 'Q2_Age', 'Q3_Sex' and 'Q4_Gender' columns, since they are common statistical variables

Observation of industry differences for Q2_Age:

1. We find that among the Q2_Age=1, the value of Petcare industry is the **highest**
2. With the increasing of the age, more people are in the "Confectionary" industry, but when Q2_Age = 8 the Petcare is **greater** than Confectionary
3. By the statistics, we observe that the largest amount of data are in the "Confectionary" industry
4. And someone don't tell or are unwilling to disclose the information of age

Observation of industry differences for Q3_Sex:

1. We find that among the Q3_Sex = 1 which is "Female", the value of "Confectionary" is the **highest**, and the "Food" is the **smallest**
2. We find that among the Q3_Sex = 2 which is "Male", the value of "Confectionary" is the **highest**, and the "Food" is the **smallest**

3. We find that among the Q3_Sex = 9 which is "Can't tell", there is only Petcare
4. By the statistics, we observe that the **largest** amount of data are in the "Confectionary" industry
5. And some of people can not tell their sex

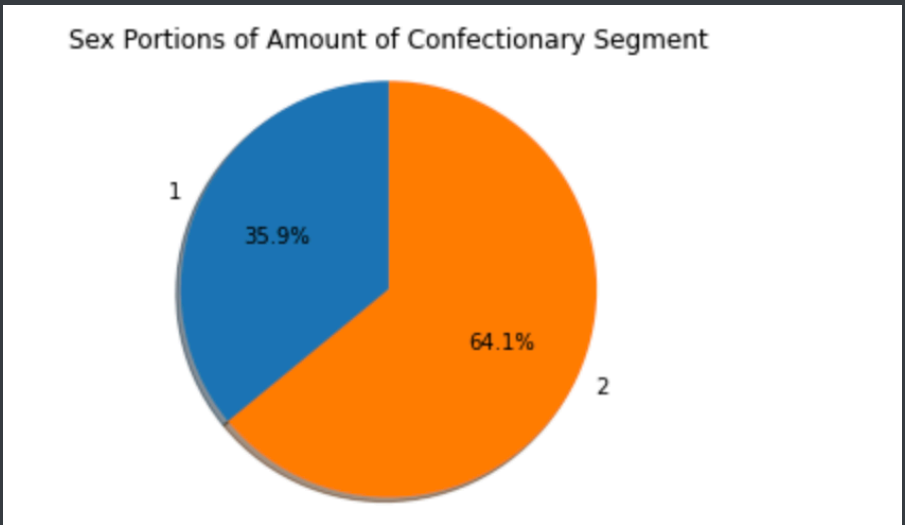
Observation of industry differences for Q4_Gender:

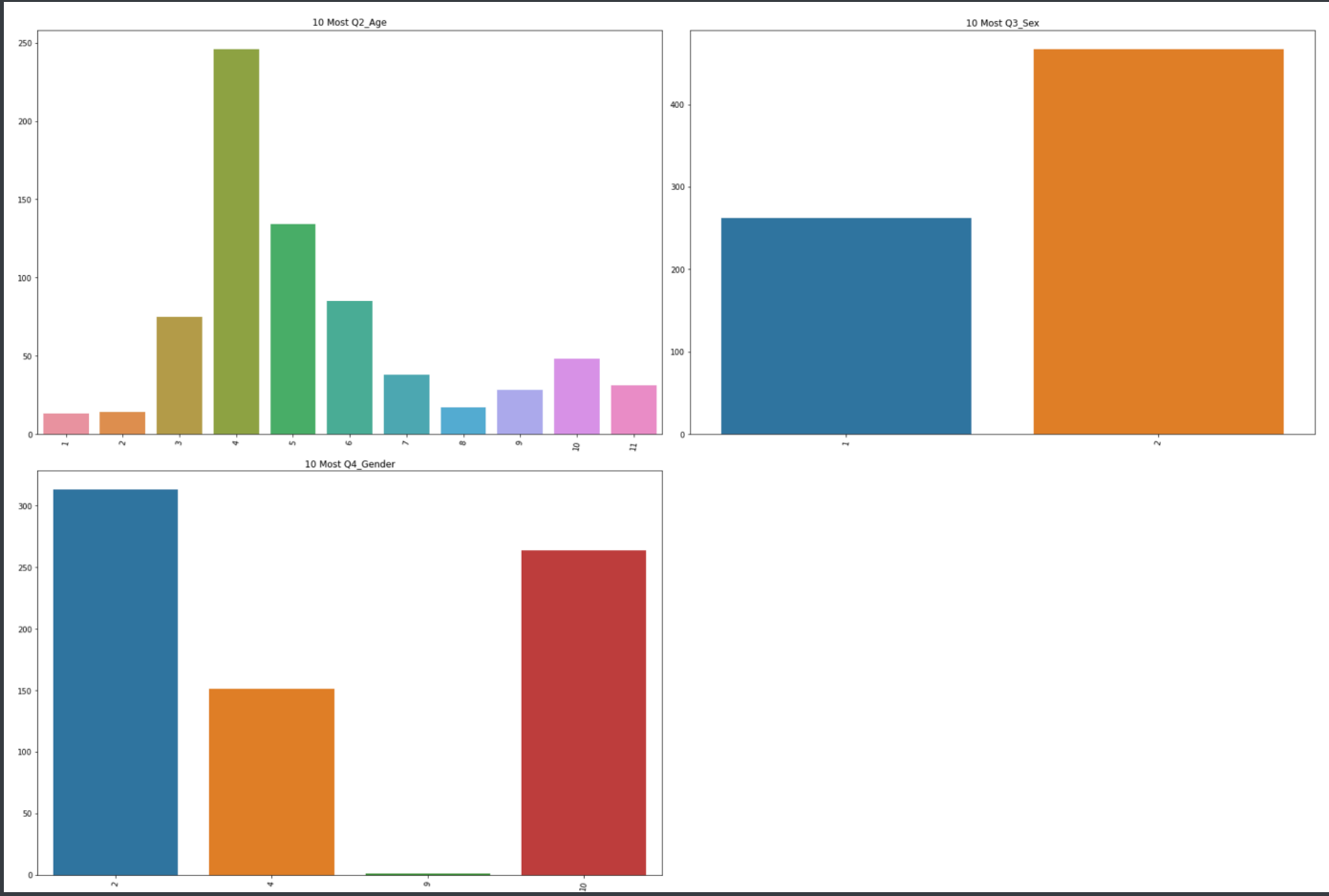
1. We find that among the Q4_Gender = 1 which is "Hyper-Masculine", is in either Petcare or Wrigley
2. We find that among the Q4_Gender = 2 which is "Masculine", the "Confectionary" is much higher than other industries, each industry has "Masculine"
3. We find that among the Q4_Gender = 4 which is "Feminine", the "Confectionary" is also higher than other industries, each industry has "Feminine"
4. By the statistics, we observe that the largest amount of data whatever their gender are in the "Confectionary" industry
5. Large amount of people are "Not Applicable" of Q4_gender in "Confectionary", "Petcare" and "Food"

Then try different way to answer the question: separating four different segments to different dataframes and analyze

Observation of industry differences for confectionary dataframe:

1. We find that for Confectionary, the 20-29 age group(4) has the largest proportion, the second one is group(5)
2. Through Q3_Sex and Q4_Gender, we find that the number of male is definitely greater than female and lots of people choose "Not Applicable" on Q4_gender
3. From barplot we find that for Confectionary segment, the amount of "Male(2)" is 64.1%, which is more than "Female(1)" (35.9%)
4. From four barplots for sex portions of different industry, we conclude that Confectionary has the largest difference in the number of male and female
5. We get the average value of Q2_Age for the Confectionary industry is 5.36(around 34 years old),the std value is 2.33
6. We can conclude that the the average age of **the Confectionary industry is the highest**

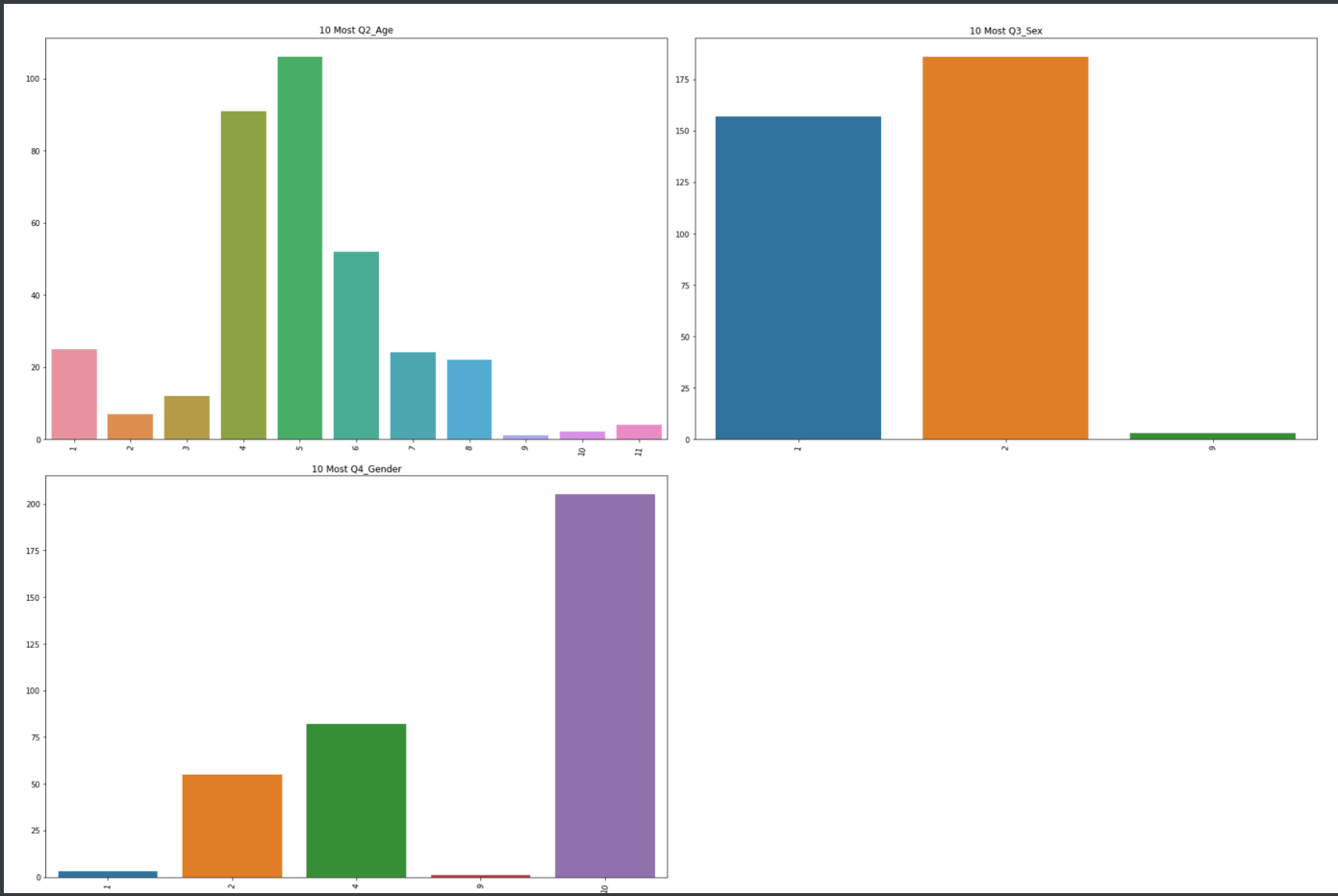




Observation of industry differences for Petcare dataframe:

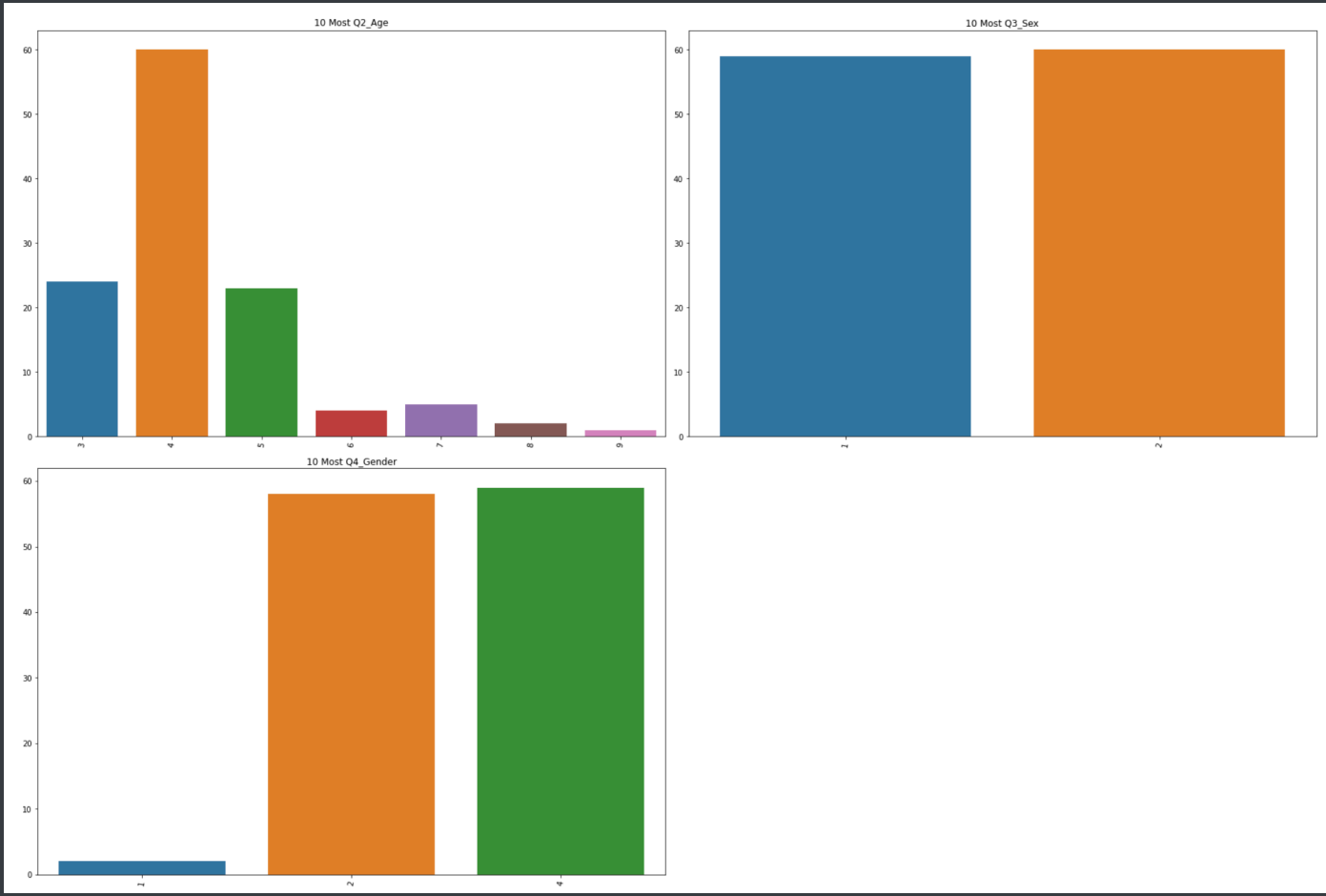
1. We find that for Petcare, the 30s age group(5) has the largest proportion, the second one is group(4)
2. From barplot we find that for Petcare, the amount of "Male(2)" is 53.8% which is the largest one, the amount of "Female(1)" is 45.4% and the amount of "Can't tell(9)" is 0.9%
3. We get the average value of Q2_Age for the Petcare industry is 4.91(around 30 years

old), which is smaller than Confectionary industry, the std value is 1.83



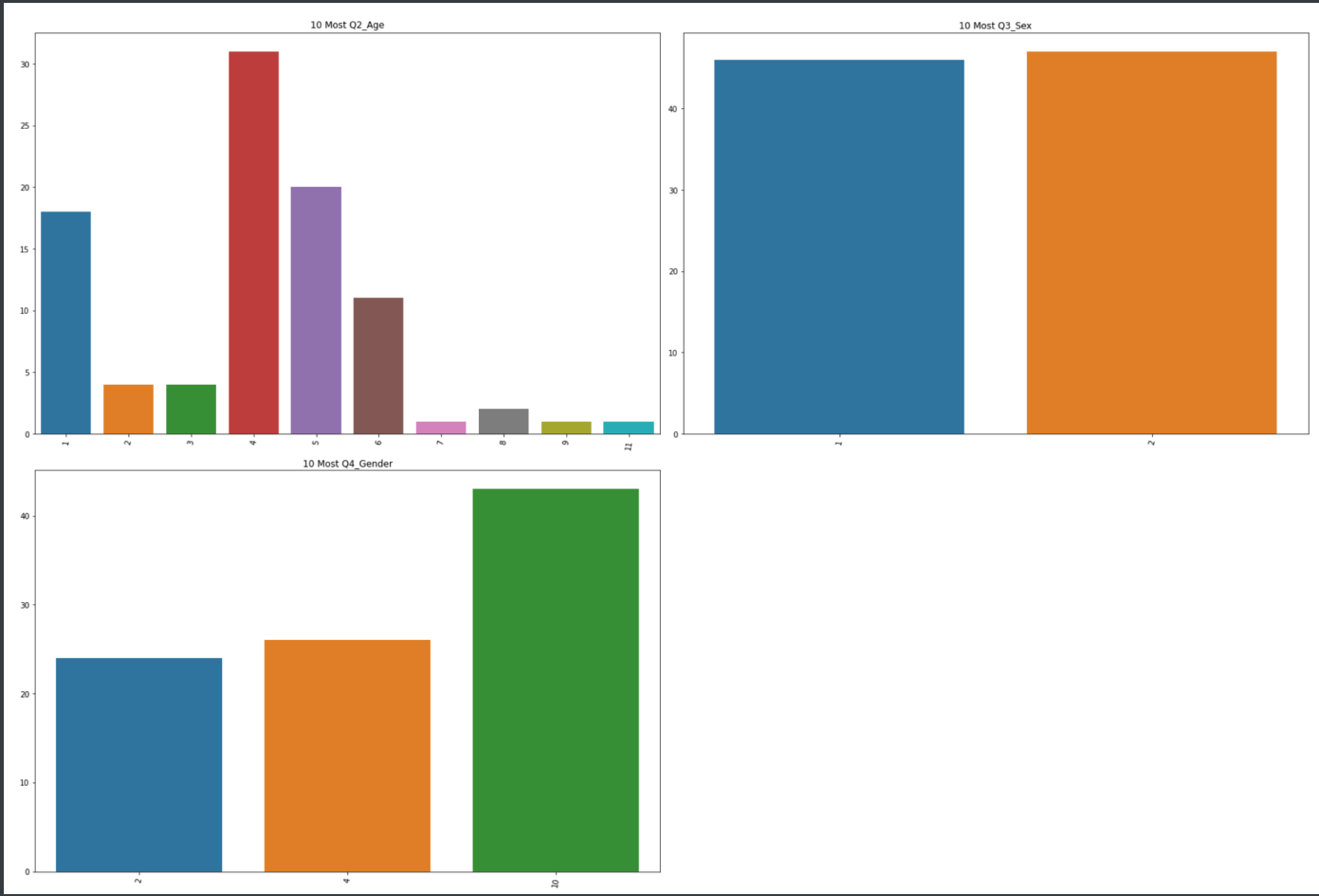
Observation of industry differences for Wrigley dataframe:

1. We 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)
2. 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
3. From barplot we find that for Confectionary segment, the amount of male is 50.4% which is more than female 49.6%
4. By observing barplot, the number of "males(2)" and "females(1)" is basically the same
5. From the four barplot for sex portions of different industry, we can conclude that the Wrigley has the smallest difference in the number of male and female
6. We get the average value of Q2_Age for the Wrigley industry is 4.29(around 25 years old), which is smaller than Confectionary and Petcare and bigger than Food, the std value is 1.15, means the age distribution is relatively concentrated



Observation of industry differences for Food dataframe:

1. We 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)
2. 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
3. From barplot we find that for Confectionary segment, the amount of "Male(2)" is 50.5% which is more than "Female(1)" 49.5%
4. By observing barplot, the number of "Male(2)" and "Female(1)" is basically the same
5. We get the average value of Q2_Age for the Food industry is 3.99(around 20 years old), which is **smallest** among these four industry, the std value is 1.98, means the age distribution is not relatively concentrated
6. We can conclude that the the average age of **the Food industry is the youngest**



Answer question: Is there a change in representation in advertisements over time?

Note: Plots and detailed analysis are showing on code

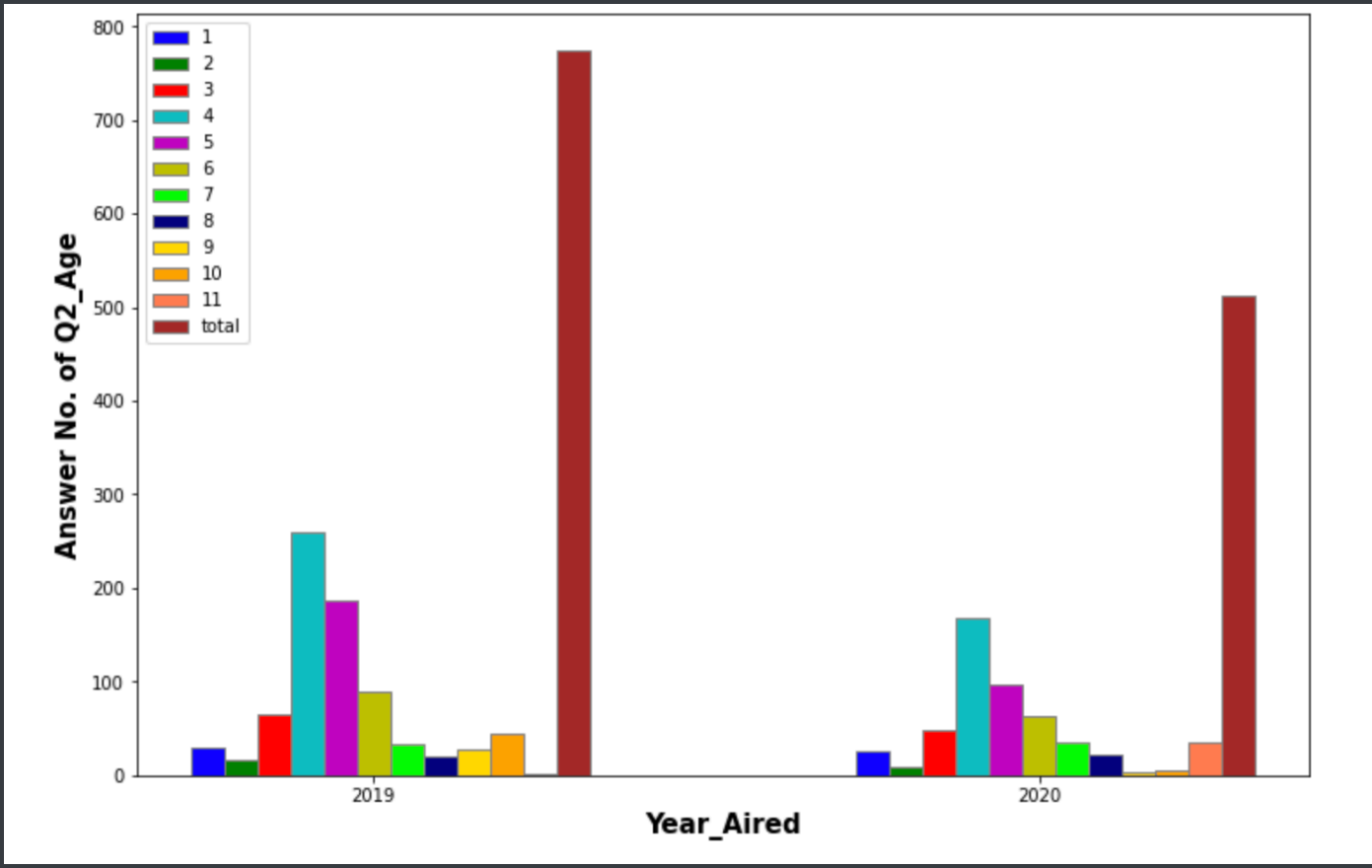
First we have an observation for entire dataset

1. For Year_Aired column, there are only 2 unique values: 2019 and 2020.
2. For Q2_Age column, most of values are Q2_Age = 4, which is the age of 20-29 year olds
3. For Q3_Sex column, number of male is more than female, and there is a little people don't tell their sex
4. For Q7_Race_Ethnicity column, we find that value 1 is the largest, which means white people.

Since “time” refers to the “Year_Aired” column, and “representation” refers to all the question columns. We find the differences among 'Q2_Age', 'Q3_Sex' and 'Q7_Race_Ethnicity' columns over Year_Aired, since they are common statistical variables

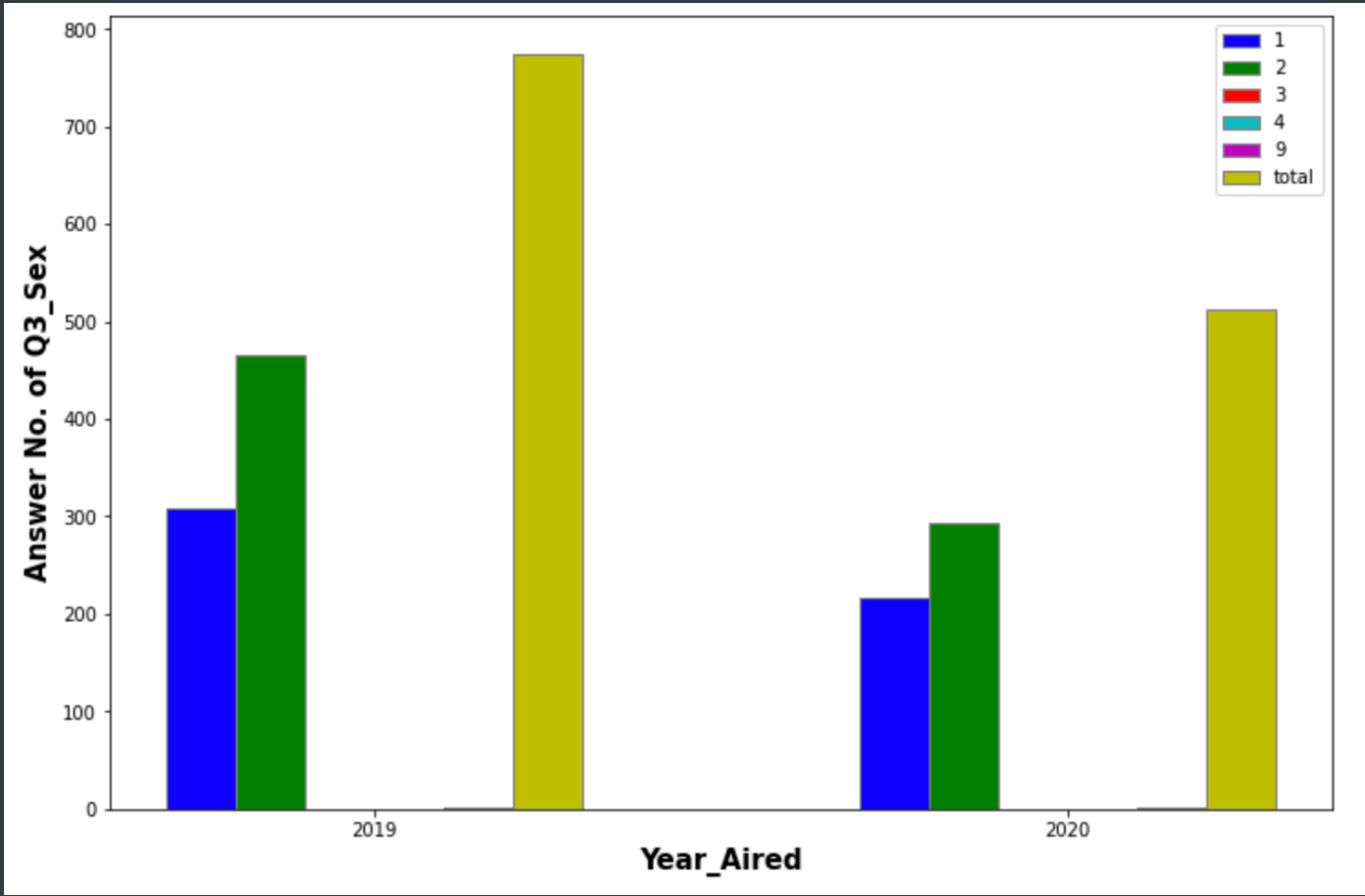
Observation of differences for Q2_Age over time:

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.



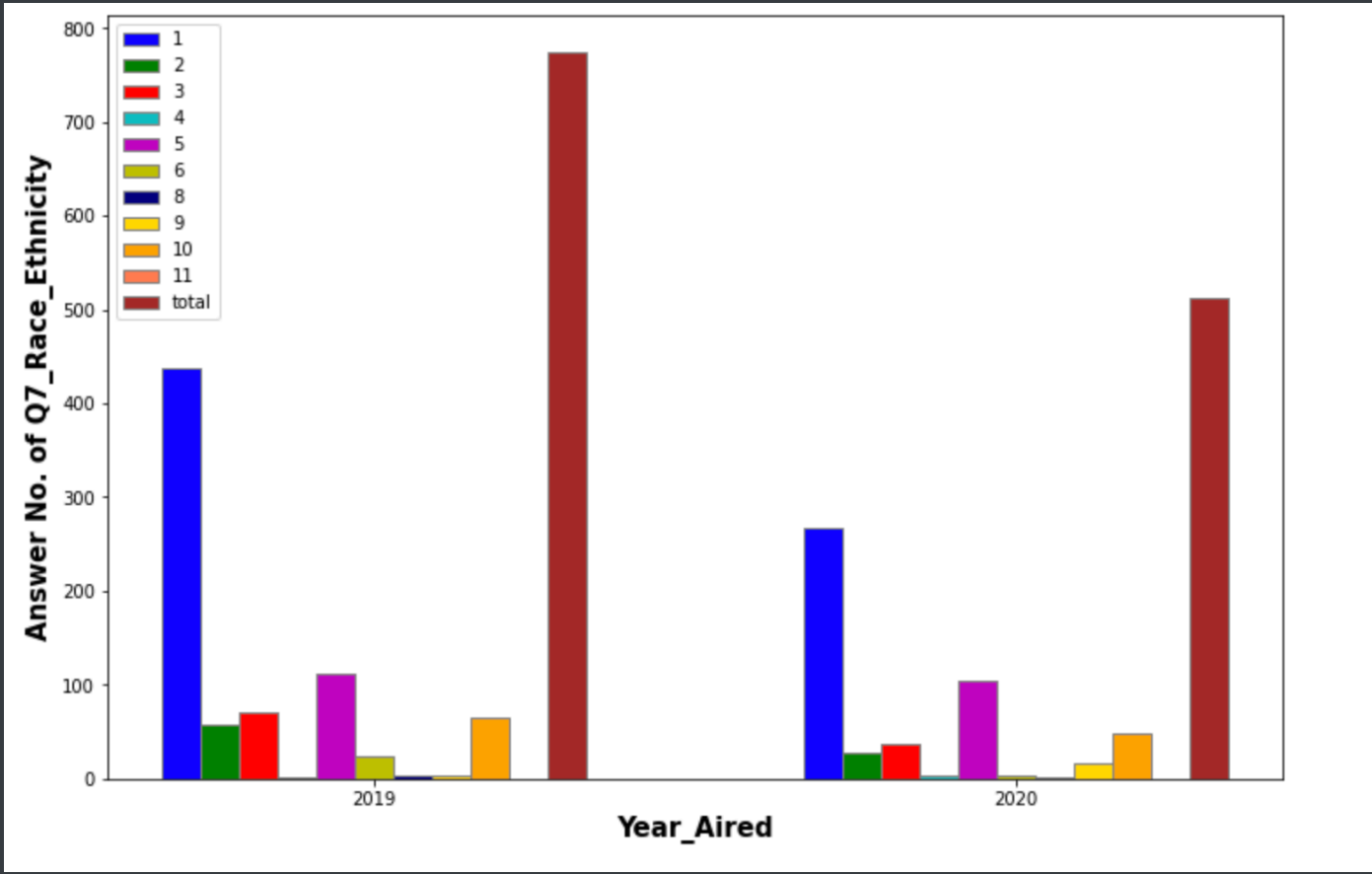
Observation of industry differences for Q3_Sex:

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



Observation of industry differences for 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.



Wrap up everything to a report and submit to the repo

The final submission includes **this report**, **Merged_Preprocessed_Mars2020_2021.csv**, **Python code for the two questions and preprocessing the data (so THREE in total, all in .ipynb format)**. They all locate in **Deliverables/Deliverable1** folder of our repo.