Task A.

Q.1

Analyze the effects of social media influence on shopping. Are there any detectable trends?” This question was given from the coursework. We are going to use the given database of Whatsgoodly on the social influence on shopping. We are going to analyze the overall result for each social media and the none part.

Analyze the effects of social media influence on shopping . Does the actual situation and background (family, school, etc…) of the user affect media influence and shopping behavior? Social media can influence on shopping but other parameters can also influence it. So, we are going to analyze other parameters which will be the background and the actual situation to see if it has an influence on the shopping. We are going to use the same dataset than the question 1 but we are going to focus on more specified groups.

Is there one social media that has more influence on University? We are assuming that the social media affect more the teenagers and young adults. We want to verify this statement by analyzing the influence of social media on shopping only on university. To do so, we are going to use the same dataset as question 1 and 2. We have also a new dataset “Which Social Media Millenials Care About Most” which is also from Whatsgoodly, about their “preference” of social media. We can use it to compare if the social media that is preferred in university is the same that the one that influence their shopping.

Q.2

For the research question 1, 2 and 3, we are going to use the dataset “Social influence on Shopping” provided by Whatsgoodly, a millennial social polling company. It allows us to access data that they get from their polls. In the data provided, we have 5 columns which are questions, segment type, segment description, answer, count. The question is: “What social platform has influenced your online shopping most? ”. The data has a segment type which show if the individual is part of a group which was surveyed for that specific question. The groups are Mobile, Web, Gender, University and Custom. It comes with a segment description which is a description of the segment population who were surveyed for each question listed in “Question” above. There are 289 unique values. We also have the answers to the questions which are the social media Snapchat, Twitter, Facebook, Instagram or none. And then we have the count and the percentage for each group surveyed. The data types for the segment type, the segment description and answers are qualitative nominal data. The count is a quantitative discrete data and percentage are quantitative continuous data. The strength of this data is that it is easy to understand, to analyze and have null or missing data. The weakness of this data is that we don’t what is the date when the data was retrieved which is an important parameter. It is also has too many segments type, so the data does not have much more and deeper information about some groups. I could be interesting to know the location of where the user did the survey from for all segment type. However, for the purpose that we need, the actual dataset can answer our research questions.

For the research question 3, we can use another dataset to complete the research question. The dataset is “Which Social Media Millennials Care About Most” and is also provided by Whatsgoodly. The question answer in this dataset is “You open ur phone and have a notif badge on instagram, facebook, snapchat, and linkedin...which do you click first?”. It has the same column as the other dataset. To answer this question, the dataset has the same segment type and have very similar segment description as the first dataset. The segment description has 129 unique values. The answer possible are also very similar, it can be Facebook, Instagram, Snapchat and Linkedin. It does not have Twitter and None as answer possible. Then we have the count and Percentage for each segment description. The data type for each column is also the same as the one in the other dataset. The strength of this data is that it is easy to read and to link with the other data. The weakness is that it does not have the location of the user that did the survey and the date of it. Another big problem is that it is only focusing on millennials and so we don’t have data for other group of age. It also have not the same sample of population. We can still use of it to answer research question 3 and to link it with the other data.

Q.3

The two datasets have the same origin, Whatsgoodly. They have also similar columns such as segment type, segment description, answer, count and percentage. So, we can say that it is related in some way. However, there are some problems between these two datasets that might affects the correlation. For example the dataset on influence of social media on shopping has 5 possibles answer which are Facebook, Instagram, Snapchat, Twitter and None. The dataset on the social media that millennials prefer the most has for answers: Facebook, Instagram, Linkedin and Snapchat. The missing possible answer in the second dataset which can lead to incomplete correlation. Another problem is that the second datasets in about millennials and the first dataset is about the anybody in the population. So, we are going to have incomplete correlation because we are missing some age group. We can answers all the questions with the first datasets but use of second datasets can help to have more understanding and deeper answers.

Task 2.

Q1.

For the first research question, we want to create a representation of the overall count for each social media to compare them. So, we decided to choose a pie char and each slice is the total count for one social media. So I came with this design in Figure … . The pie chart will have 5 slices representing the 4 socials media and the none answer. Inside each slice, we will have the percentage. On the right of the pie chart, we will put the legend for each slice. The color used will be the reference of the logo of each social media. For example Facebook will be a dark blue, Instagram in pink, snapchat in yellow and twitter in light blue. For the none part, I just used a color that was no already used. The user will have the possibility of hover the mouse over a slice and the slice will display all the information in terms of count. The scale will be the same for each of the slice as we sum the count for each answer. The idea with this visualisation is to permit to see easily a social media that is standing out, compare each slice to each other and permit to find a trend. This will facilitate analysis and communication due to it easiness to understand and read through and the display of the information of the slice wanted.

For the second research question, we want to create a visualiation that permit an easy comparison of the count of social media between the different groups possible. To do this, we are choosing a basic group barplot. We choose this visualisation because it displays for each group each socia media next to each other. So, I came with this design in Figure … . The bar for each group will be each social media and the none part. The color for each bar will be the same as the one describe in the description of research question 1 ( color of the logo to the corresponsing bar). The x axis will have the names of the group and the y axis will have the number of count. Under the barplot, we are putting the legend for each bar, same principle as the description of research question 1. For the interaction, the user will be able to click on different button that will change the input of the data. The data will be filter data such as parent’s earning or student loan for example. Each button will be link to a csv file that will store each group. The filter and the storage in a csv file will be done manually. It will separate each group and make it easier too choose the visualisation for the data we want. . This will facilitate analysis and communication due to it easiness to understand and read through and the display of the information through the bar plot.

For the third question, we want to create a visualisation that permit to see all the data of the two datasets to easily read through it and analyze it. To answer this question, we are choosing a scatter plot as we want to see if there is a correlation between the favorite social media of millennials and the social media influence on shopping in university. We want to see if the social media that is influencing the shopping is the same as the favorite one of millennial. To do so we are going to manipulate the data first on the two datasets to have the total for each social media in the group university. Using this filtered data, we are going to plot the point for each social media on the scatter plot. Each point will have the color of the social media like the two previous visualisation. The x-axis will be for the counter for preffered social media and the y axis will be the counter for social media influence on shopping. Under the barplot we are going to have the legend for each social media. For the user’s interaction, when the user click on one point, it will display with a pop up page, the details of the total of count with it’s segment description and the count for that group. Then the user would just have to analyze the scatter plot and the information of each point.

Task 3:

* it shall be accompanied by a short description of how data are being processed (and acknowledgement of your data source(s)). [10 marks]

The data is from: <https://data.world/ahalps/social-influence-on-shopping>. The data is manually filter in different csv file by groups and linked to each button. The groups are the parent’s earning, student loan, actual situation, private school and employement. The bar plot will plot for each subgroup the result for each answer possible. The code for the barplot was taken and modified from: https://d3-graph-gallery.com/graph/barplot\_grouped\_basicWide.html.

Changer same millemnials pas me size sample.