Introduction to Data Visualization Coursework 2

Thanh-Dat Kevin Trang
Department of Informatics
King's College London
K21206580

I. Part 1: Analytics

A. Research questions

- Question 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 from Keats on the social influence on shopping [1]. We are going to analyze the overall result for each social media and the none part to find trends.
- Question 2: Analyze the effects of background actual situation on the influence of social media on shopping . Does the actual situation and background (family, school, etc…) of the user can affect media influence on his 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 also going to use the data set from Keats but we are going to focus on more specified groups.
- Question 3: Analyze the effects of social media influence on shopping in the university group. Is the social media that has more influence on University is also the one that is the most preferred overall? We are assuming that the social media affect a lot teenagers and young adults. We want to verify if the influence of social media on shopping on the group "university" is the same than the one that the millennial care about the most. To do so, we are going to use the data set provided by Keats and . We have also another data set "Which Social Media Millennial Care About Most" about the "preferred" social media of millennial. We can use it for a comparison between the two data sets.

B. Data sets and data type

- Data set: "Social influence on Shopping".
  For the research question 1, 2 and 3, we are going to use the data set "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 data 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 description of the segment population who were surveyed for each question listed in "Question" above. We also have the answers to the questions which are the social medias 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. To answer question 1,2 and 3, we will need a mix of quantitative data and qualitative nominal data to answer it. For question one, we will take the total quantitative data of each social media to see if a trend can be find. For question 2, we need to filter qualitative data to have specified groups to answer the question. For question 3, we are going to do the total of count for each social media in the group "university". The strength of this data is that it is easy to understand, to analyze and have no null or missing data. The weakness of this data is that we don't what is the date when the data was retrieved which can be an important parameter. The data does not have deeper information about some groups which can limit the analysis and to find patterns. Even tho we know the location of university, it could be interesting to know the location of where the user did the survey from for all other segment type. However, for the purpose that we need, the actual data set can answer our research questions.

- Data set: "Which Social Media Millennial Care About Most":
  For the research question 3, we can use another data set to complete the research question. The data set is "Which Social Media Millennial 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 first data set. This data set has the same segment type and have very similar segment description as the first data set.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 data set. To help answer question 3, we are going to do the total of the count for each social media in

the group ”university”. This is the same step that we did before but it’s just another data set, 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. Also, it does have not the same sample of population as the first data set. We can still use of it to answer research question 3 and to link it with the other data.

C. Correlation

The two data sets 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 data sets that might affects the correlation. For example the data set on influence of social media on shopping has 5 possibles answer which are Facebook, Instagram, Snapchat, Twitter and None. The data set on the social media that millennials care the most has for answers: Facebook, Instagram, Linkedin and Snapchat. The missing possible answer in the second data set which can lead to partial correlation. Another problem is that it does not have the same sample of population. The first data set has 2676 surveyed people against 9141 for the second data set. So it can lead us to inaccurate or misleading results because of the different sample . We can answers question 1 and 2 with the first data set and use the mix of the two data sets for question 3 to have more understanding and better answers.

II. Part 2. Design and Discussion

A. Design and drawing

B. Design description

- 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 chart and each slice is the total count for one social media. So I came with this design in Figure 1 . The pie chart will have 5 slices representing the 4 socials media and the none answer. Inside each slice, we will have the percentage. Under 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, so green. The user will have the possibility of clicking 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 do the total of the count. 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 visualisation that permit an easy comparison of the count of social media between the different groups that we chose. To do this, we are choosing a basic group bar plot. We choose this visualisation because it displays for each group each social media next to each other and it easy to compare. So, I came with this design in Figure 2. The bar in 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 for research question 1 (color of the logo to the corresponding bar). The x axis will have the names of the group and the y axis will have the count number. Under the bar plot, we are putting the legend for each bar. 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 data sets 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.So, I came with this design in Figure 3. 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 data sets 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 preferred social media and the y axis will be the counter for social media influence on shopping. Under the bar plot 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.

III. Implementation

We take the data set ”Social Influence on Shopping”and we manually filter it in different csv file by groups and linked to each button. The groups that we chose to represent background and actual situation are the parent’s earning, student loan, actual situation, private school and employment. In each group, there are subgroups as we can see in Figure 4. We just retrieved the quantitative value for each social media in each subgroup. Then using these filter data, we can create our bar plot.

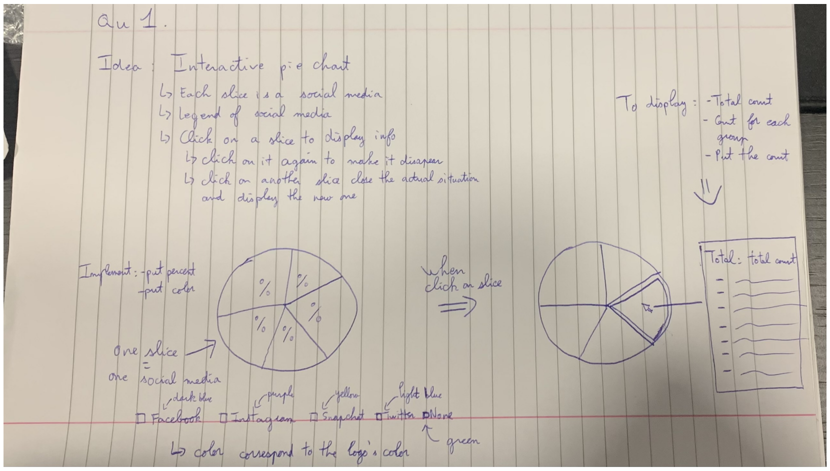
References

[1] A. Halper, “Social influence on shopping - dataset by ahalps,” data.world, 19-May-2017. [Online]. Available: https://data.world/ahalps/social-influence-on-shopping. [Accessed: 01-Apr-2023]

[2] A. Halper, “Which Social Media Millennials Care About most - dataset by ahalps,” data.world, 19-May-2017. [Online]. Available: https://data.world/ahalps/which-social-media-millennials-care-about-most. [Accessed: 01-Apr-2023]

[3] Basic grouped barplot in d3.js. [Online]. Available: https://d3-graph-gallery.com/graph/barplot_grouped_basicWide.html. [Accessed: 1-Apr-2023]

Appendix



Qu 1.

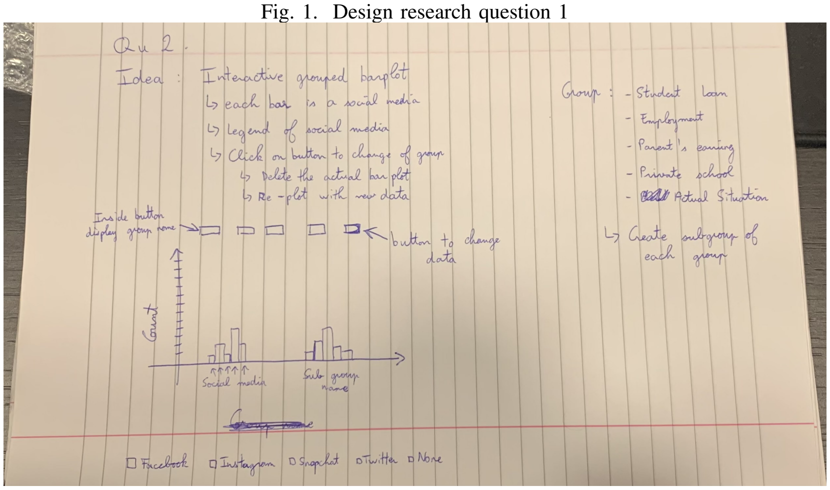
Idea : Interactive pie chart
↳ Each slice is a social media
↳ Legend of social media
↳ Click on a slice to display info
↳ click on it again to make it disapear
↳ click on another slice close the actual situation and display the new one

To display : - Total count
- Cunt for each group
- Put the count

Implement : - put percent
- put color

↳ color correspond to the logo's color

Fig. 1. Design research question 1



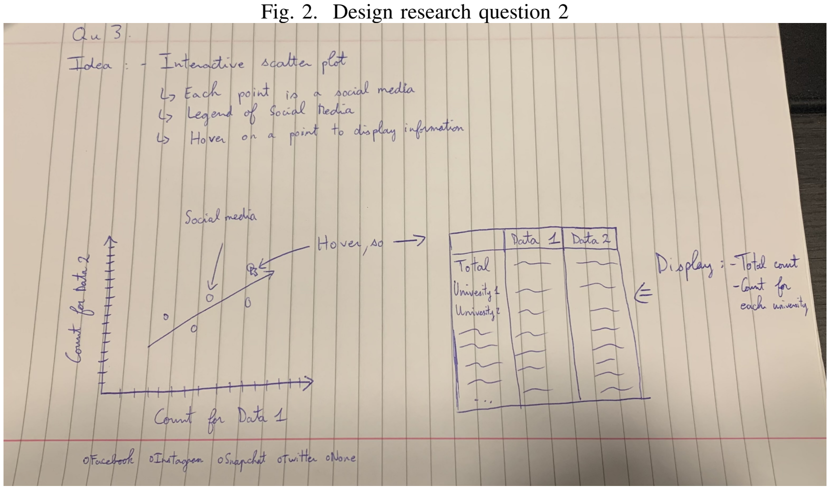
Qu 2.

Idea : Interactive grouped barplot
↳ each bar is a social media
↳ Legend of social media
↳ Click on button to change of group
↳ Delete the actual bar plot
↳ Re-plot with new data

Group : - Student Loan
- Employment
- Parent's earning
- Private school
- Actual Situation

↳ Create subgroup of each group

Fig. 2. Design research question 2

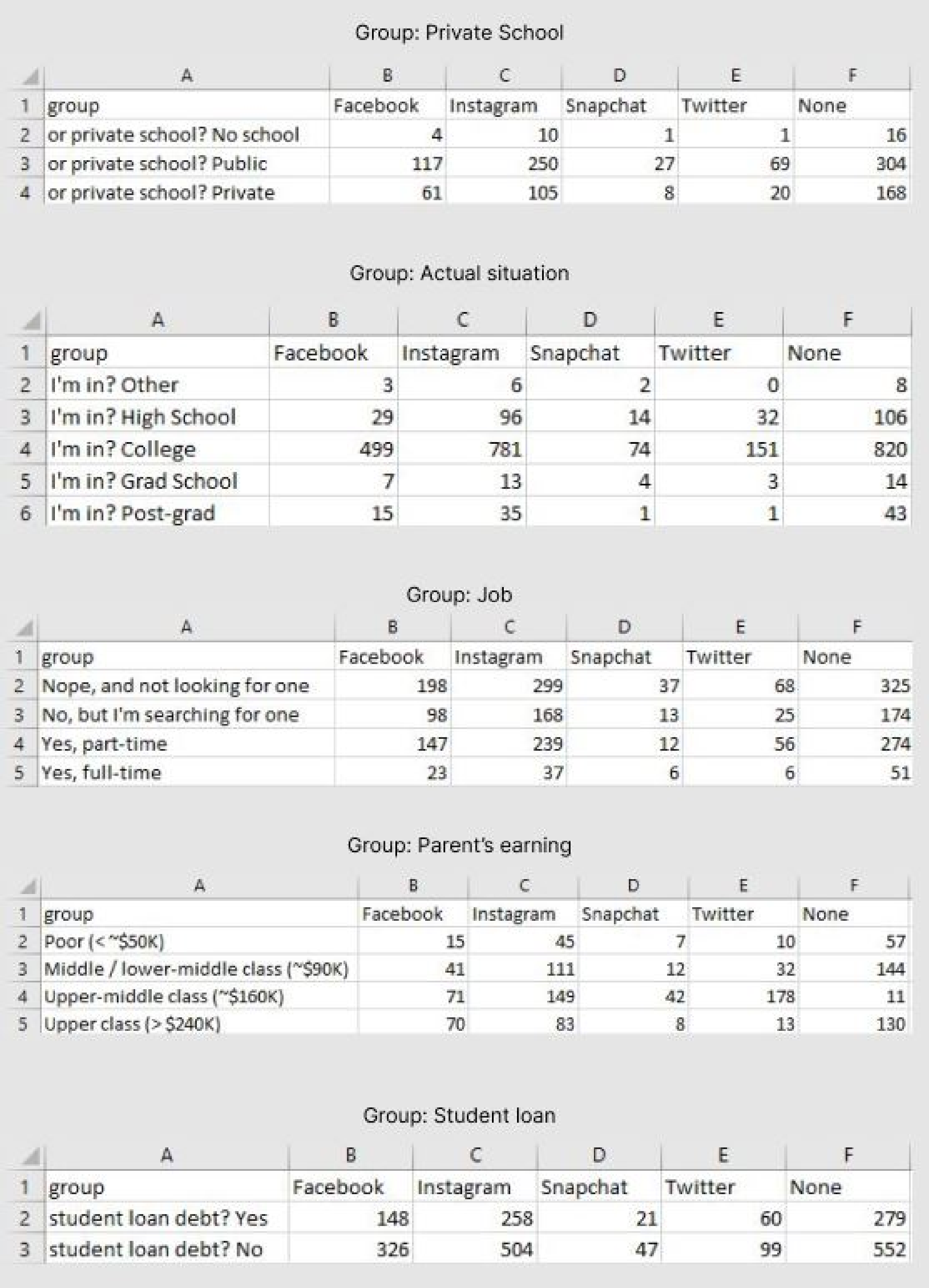


Qu 3.

Idea : - Interactive scatter plot
↳ Each point is a social media
↳ Legend of Social Media
↳ Hover on a point to display information

Display : - Total count
- Count for each university

Fig. 3. Design research question 3



Group: Private School

	A	B	C	D	E	F
1	group	Facebook	Instagram	Snapchat	Twitter	None
2	or private school? No school	4	10	1	1	16
3	or private school? Public	117	250	27	69	304
4	or private school? Private	61	105	8	20	168

Group: Actual situation

	A	B	C	D	E	F
1	group	Facebook	Instagram	Snapchat	Twitter	None
2	I'm in? Other	3	6	2	0	8
3	I'm in? High School	29	96	14	32	106
4	I'm in? College	499	781	74	151	820
5	I'm in? Grad School	7	13	4	3	14
6	I'm in? Post-grad	15	35	1	1	43

Group: Job

	A	B	C	D	E	F
1	group	Facebook	Instagram	Snapchat	Twitter	None
2	Nope, and not looking for one	198	299	37	68	325
3	No, but I'm searching for one	98	168	13	25	174
4	Yes, part-time	147	239	12	56	274
5	Yes, full-time	23	37	6	6	51

Group: Parent's earning

	A	B	C	D	E	F
1	group	Facebook	Instagram	Snapchat	Twitter	None
2	Poor (< ~$50K)	15	45	7	10	57
3	Middle / lower-middle class (~$90K)	41	111	12	32	144
4	Upper-middle class (~$160K)	71	149	42	178	11
5	Upper class (> $240K)	70	83	8	13	130

Group: Student loan

	A	B	C	D	E	F
1	group	Facebook	Instagram	Snapchat	Twitter	None
2	student loan debt? Yes	148	258	21	60	279
3	student loan debt? No	326	504	47	99	552

Fig. 4. Processed Data

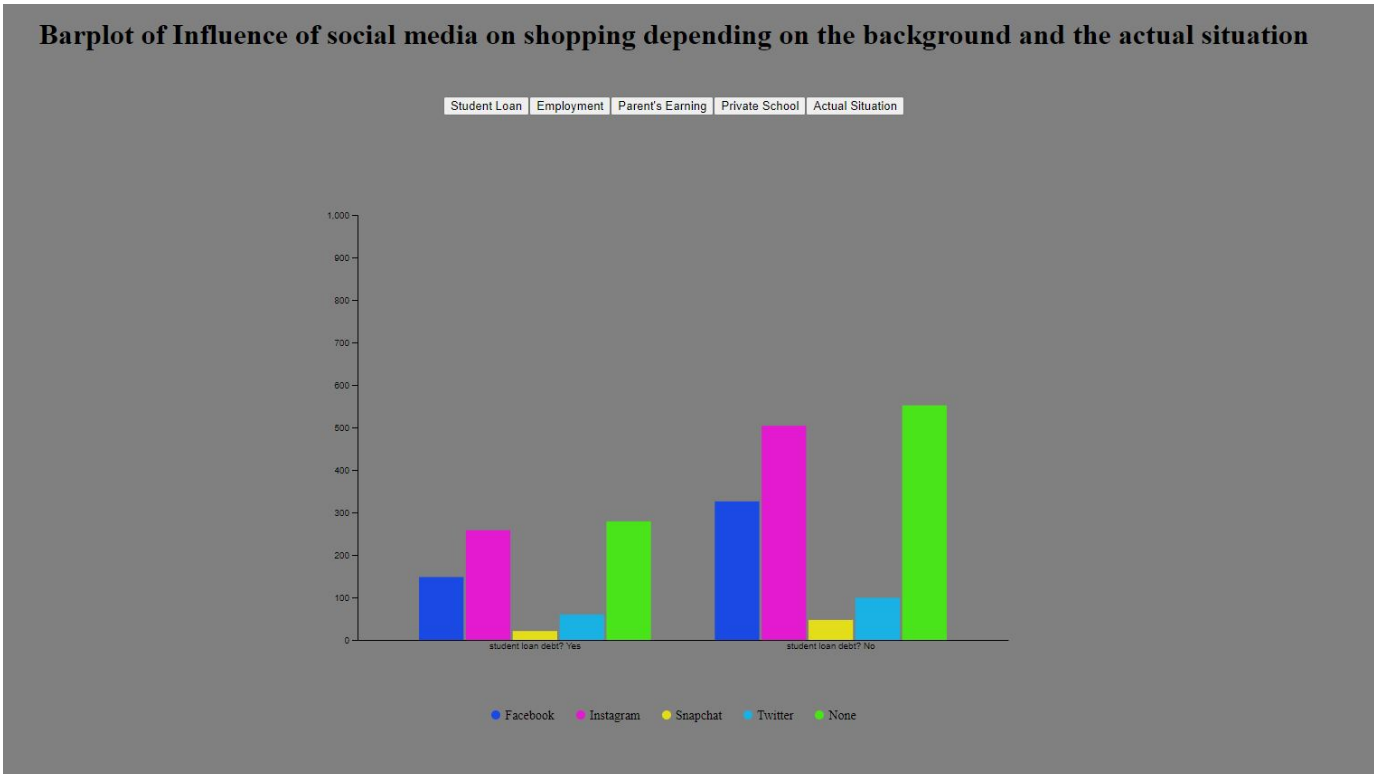


Fig. 5. Grouped bar plot for research question 2