The potential factors that affect the time spent for users on Instagram

Lee Kai Ngai David

Quantitative and Qualitative Research

10/02/2022

Word count: 8061

Acknowledgments

This dissertation is wholeheartedly acknowledged to my class instructor, David Johnson, who has been my source of inspiration and taught me a lot of key concepts and procedures of doing a research. No words can express my gratitude for all your guidance, patience and help. Without you, I would have struggled so much in writing a proper research paper and reporting my research results. You get the biggest thank you to help me complete this dissertation.

To my classmates and my friends, who encouraged me to finish this study, have taken their precious time to be a part of my research sample.

Table of Contents

List of Tables

List of Equation

List of Figures

Chapter 1: Introduction to the Study

1.1-Introduction

Instagram, a kind of popular social media mobile application, has more than a billion users sharing photos and videos (Kemp, 2021). Social media like Instagram attracts people of all ages, especially teenagers while the time spent on the virtual world goes beyond their real life via different kinds of features it offers. Owing to the similar features for every social media, such as photo-liking feature, messaging feature, shopping feature and 'TikTok' feature, this research is focusing on the reasons and underlying features on Instagram that why people are willing to spend their spare time on Instagram. Previous researches about Instagram were all focusing on mental health and Instagram usage. For example, Treitel researched the impact of Instagram and other social media on people's mental health or physical health (2020). Yesilyurt and Solpuk Turhan researched the prediction of the time spent on Instagram by social media addiction and life satisfaction. And also, Olufadi (2015) researched a new instrument for measuring the time spent on social networking sites.

Contrary to the previous paper, My research extends their work on investigating the potential factors of social media that affect the time spent on Instagram.

1.2-Purpose

The aim of this primary quantitative research was to explore the relationship and features importance between time spent and different features on Instagram so as to determine the underlying factors which affect the time spent for people on Instagram.

1.3-Significance of the study

Based on the result of the research, it will bring an enormous contribution on business value that increasing the duration for the users on social media can contribute to a higher revenue gain for a company. Knowing which features would increase social media use, then companies can modify their marketing strategies and direction of promotion on social media to increase their profit.

1.4-Research Questions and Hypothesis

Research Question 1 (RQ1): Is there a correlation between the time spent on Instagram and different kinds of potential features on Instagram?

RQ1.1: Is there a significant correlation between the time spent on Instagram and the average number of likes for the top five posts?

$H_0$1.1 = There is no significant correlation between the time spent on Instagram and the average number of likes for the top five posts.

H_a1.1 = There is a significant correlation between the time spent on Instagram and the average number of likes for the top five posts.

RQ1.2: Is there a significant correlation between the time spent on Instagram and the number of people each Instagram user follows?

$H_0$1.2 = There is no significant correlation between the time spent on Instagram and the number of people follows?

H_a1.2 = There is a significant correlation between the time spent on Instagram and the number of people follows?

RQ1.3: Is there a significant correlation between the time spent Instagram and the extent of using the Instagram story feature?

$H_0$1.3 = There is no significant correlation between the time spent on Instagram and the extent of using the Instagram story feature.

H_a1.3 = There is a significant correlation between the time spent on Instagram and the extent of using the Instagram story feature.

RQ1.4: Is there a significant correlation between the time spent on Instagram and the extent of glancing Instagram stories from other users?

$H_0$1.4 = There is no significant correlation between the time spent on Instagram and the extent of glancing Instagram stories from other users.

H_a1.4 = There is a significant correlation between the time spent on Instagram and the extent of glancing Instagram stories from other users.

RQ1.5: Is there a significant correlation between the time spent on Instagram and the extent of using the Instagram shop feature.

$H_0$1.5 = There is no significant correlation between the time spent on Instagram and the extent of using the Instagram shop feature.

H_a1.5 = There is a significant correlation between the time spent on Instagram and the extent of using the Instagram shop feature.

RQ1.6: Is there a significant correlation between the time spent on Instagram and the extent of chatting with other users on Instagram?

$H_0$1.6 = There is no significant correlation between the time spent on Instagram and the extent of chatting with other users on Instagram.

H_a1.6 = There is a significant correlation between the time spent on Instagram and the extent of chatting with other users on Instagram.

RQ1.7: Is there a significant correlation between the time spent on Instagram and the extent of scrolling through the 'Explore' and 'Reel' features on Instagram?

$H_0$1.7 = There is no significant correlation between the time spent on Instagram and the extent of scrolling through the 'Explore' and 'Reel' features on Instagram.

H_a1.7 = There is a significant correlation between the time spent on Instagram and the extent of scrolling through the 'Explore' and 'Reel' features on Instagram.

RQ1.8: Is there a significant correlation between the time spent on Instagram and the extent of spending time on Instagram advertisement?

$H_0$1.8 = There is no significant correlation between the time spent on Instagram and the extent of spending time on Instagram advertisement.

H_a1.8 = There is a significant correlation between the time spent on Instagram and the extent of spending time on Instagram advertisement.

Research Question 2 (RQ2): Which potential feature has the most statistical influence on affecting the time spent on Instagram?

2-Literature Review

Yesilyurt and Solpuk Turhan (2020) investigate the prediction of the time spent on Instagram by social media addiction and satisfaction. This is a quantitative primary research which determined three main purposes. The first purpose is to determine the correlation between social media addiction and life satisfaction, and they found that there is a weak correlation between these two variables. The second purpose is to find the difference between the university students' social media addiction or level of satisfaction and the time that they spend on Instagram, and the research analysed that there is the difference. Moreover, the last purpose is to build a model which helps to predict the time spent on Instagram in terms of gender, age and social media addiction, and they found that the variable that predicts the time spent on Instagram the most is gender, and it is followed by age and social media addiction. This study gave me some ideas to use the Pearson correlation to evaluate the correlation between two variables. Also, this study provided me with ideas for the metrics which I need to evaluate the model. Although they use the likelihood ratio (LR) test to measure the significance of an independent variable in the model, the LR test usually measures the goodness-of-fit between two models. Rather than the LR test, I would use a popular and advanced metric, SHAP to obtain the feature importance and interpret the model.

Kirik and other researchers (2015) investigate the level of social media addiction among young people in Turkey. This research is a quantitative primary research. The age of the sample students are between 13-19, which is in the stage of teenagers and it is quite similar with the sample which I am examining. In their study, they initially defined and explained the social media and different variety of addiction which caused by social media. Also, their study showed that social media has a significant influence on young people, which causes the addiction level of the young to increase, and the time they spend on these networks to go up. This study gave me some ideas to analyse the quantitative data. As their research mentioned, they used Independent-Samples T-Test and One-Way Analysis of Variance. Besides, they used the addiction factor of "Social Networking Status Scale" as the data collection tool, which was developed by Arslan and Kirik (2013: 223-231) to measure social media addiction of the young people. In contrast, owing to the lack of public research

about time spent on social media caused by different factors, I made my own assumption of different features which may affect the time spent for students on Instagram.

Treitel and Yael (2020) investigate if there is a connection between Instagram usage and self-esteem by looking at the variables of the length of a person's marriage, gender, happiness in marriage, age, and culture. It is a quantitative primary research which concluded that there might not be a causational relationship between Instagram and self-esteem but social media exacerbate feelings people already have. In this research, the researcher set lots of research questions to determine if varying kinds of variables are a significant predictor of self-esteem when controlling for other variables. However, there are some flaws to make these hypotheses that it can be more than a variable affecting the target variable. Moreover, they use multiple regression to analyse the variable. Similarly, I will use relevant models to fit on the data and obtain the most significant feature among different variables. To take into account the survey, this research gave me some ideas to include demographic information, such as time spent on Instagram.

Chapter 3: Research Method

This quantitative study was to determine the relationship between the time spent on Instagram and different potential factors by looking at the variables of the average number of likes for the posts, number of followings, extent of using or glancing Instagram story, shop, message, 'explore' and 'reel' features and Instagram advertisement.

3.1-Population and sampling

For the potential features which would be investigated via the survey, the questionnaire will be designed to collect students' Instagram account ID because some information such as number of followers, number of followings and the number of posts can be collected from students' Instagram profile directly. Other potential factors such as the extent for the user to spend time on different Instagram advertisements and the target variable which is time spent on Instagram can be investigated from the questionnaire.

The target population for this study is the people who has Instagram account around the world. The vast majority of samples are surveyed on Instagram and the sampling method that I have used was convenience sampling due to the short period time of the research project.

The process of sampling and collecting the data was divided into two parts. In the first part, the majority of samples are collected on Instagram by crawling the name of each instagram user and then a message is sent to everyone by building an automated program (See appendix 1). The process of automation in the program is mainly achieved by the Selenium web driver Python package which helps to control the web browsers and perform browser automation. However, owing to the stringent policy on Instagram which limits the new message to be sent daily, it is not as efficient as expected to spread the survey only via this method. Alternatively, it is discovered that since Instagram has not revealed a limitation or a guideline on the number of posts which users can post daily, it can be considered as another approach to spread the survey by automatically tagging other Instagram users in each post (See appendix 2). Additionally, in the second part, different information and features, such as the average number of likes for the top five posts, the number of followings and the number of followers should be scraped from the Instagram users who finished the survey (See appendix 3). All of these requesting and scraping processes are using five Instagram accounts.

The population are people who use Instagram. Around 40,000 potential respondents (N=40,000) were contacted and 859 responded (n=859).

3.2-Instrumentation

This research conducted a survey which collected quantitative data. The survey consists of 8 questions, including 1 interval question and 6 ordinal question, and the remaining question is asked for the Instagram name. Each question in the survey represents a variable apart from the first question which requested the Instagram username. There are totally 8 independent variables and a dependent variable. Two of the variables (the average number of likes for top 5 posts and the number of followers) is directly scraped from samples' Instagram profile. A copy of the survey has been put to the appendix (See appendix 4).

Table 1.0 Variables for the research questions

	Research Questions	Independent Variable	Dependent Variable	Controlled Variables
1.1	Is there a significant correlation between the time spent on Instagram and the average number of likes for the top five posts?	The average number of likes for the top five posts.	Time spent for Instagram users on Instagram.	All features apart from the average number of likes for the top five posts.
1.2	Is there a significant correlation between the time spent on Instagram and the number of people each Instagram user follows?	The number of people each Instagram user follows?	Time spent for Instagram users on Instagram.	All features apart from the number of people each Instagram users follows.

	Research Questions	Independent Variable	Dependent Variable	Controlled Variables
1.3	Is there a significant correlation between the time spent on Instagram and the extent of using the Instagram story feature?	The extent of using the Instagram story feature.	Time spent for Instagram users on Instagram.	All features apart from the extent of using the Instagram story feature.
1.4	Is there a significant correlation between the time spent on Instagram and the extent of glancing Instagram stories from other users?	The extent of glancing Instagram stories from other users.	Time spent for Instagram users on Instagram.	All features apart from the extent of glancing Instagram stories from other users.
1.5	Is there a significant correlation between the time spent on Instagram and the extent of using the Instagram shop feature.	The extent of using the Instagram shop feature.	Time spent for Instagram users on Instagram.	All features apart from the extent of using the Instagram shop feature.
1.6	Is there a significant correlation between the time spent on Instagram and the extent of chatting with other users on Instagram?	The extent of chatting with other users on Instagram.	Time spent for Instagram users on Instagram.	All features apart from the extent of using chatting with other users on Instagram.
1.7	Is there a significant correlation between the time spent on Instagram and the extent of scrolling through the 'Explore' and 'Reel' features on Instagram?	The extent of scrolling through the 'Explore' and 'Reel' features on Instagram.	Time spent for Instagram users on Instagram.	All features apart form the extent of scrolling through the 'Explore' and 'Reel' features on Instagram.
1.8	Is there a significant correlation between the time spent on Instagram and the extent of spending time on Instagram advertisement?	The extent of spending time on Instagram advertisement.	Time spent for Instagram users on Instagram.	All features apart from the extent of spending time on Instagram advertisement.

	Research Questions	Independent Variable	Dependent Variable	Controlled Variables
2	Which potential feature has the most statistical influence on affecting the time spent on Instagram for Instagram users?	All the independent variables in the research question 1.	Time spent for Instagram users on Instagram.	The number of sample

3.3-Procedure and time frame

The research began on 16th January, 2022 and ended on 1st February, 2022. After all the raw datasets are collected, it is necessary to clean the raw datasets and preprocess them. First of all, to consider that some people may do the survey twice, the duplicate Instagram name should be deleted. Also, people who did not do it seriously and filled the unrealistic information have been considered as outliers and have been also deleted. Besides, the features such as number of posts, timestamp and Instagram name can be dropped since these features are no longer helpful for the research analysis and cannot give any insides. After that, the information such as number of followers and posts which are scraped on the users' profile pages can be merged with the main dataset. The merging method which I used is inner merge. Finally, to take into account the data of the 'time spent' feature, people typed the time spent on Instagram in different formats, such as '2h3m1s', '30minutes3seconds' and '2hrs44mins'. Owing to dealing with the different string format of this, I build an algorithm to transform these formats into minutes (See appendix 4). The remaining seconds are considered to be ignored since 'seconds' has a smaller interpretation than 'minutes'. Last but not least, it is more convenient to handle different features by converting the column name to the following symbols.

Table 2.0 Symbols to represent different variables

The average number of likes for first 5 posts	a
The number of followings	b
The extent of using the Instagram story feature.	c
The extent of glancing Instagram stories from other users.	d
The extent of using the Instagram shop feature.	e

The extent of chatting with other users on Instagram.	f
The extent of scrolling through the 'Explore' and 'Reel' features on Instagram.	g
The extent of spending time on Instagram advertisement.	h
The time spent for Instagram users on Instagram.	time spent

3.4-Analysis plan

To determine the relationship between the time spent for Instagram users on Instagram and different kinds of potential features on Instagram, it is recommended to do an exploratory data analysis for the dataset. First of all, the skewness and kurtosis will be analysed. For the research questions, the correlation of numerical features with time spent on Instagram can be visualized as a heat map with Pearson correlation coefficient. Also the correlation of categorical features with time spent on Instagram can be visualized as a box-whisker diagram. The statistical test will be used is student's t-test for correlation. The critical α level is specified as 0.05.

In order to determine the most substantial impact which affects the time spent for Instagram users on Instagram compared with different kinds of variables, we should split the data into the training set and testing set in order to train an optimal machine learning model which can well fit the dataset and it can be better to interpret the features' relationship. In this study, some popular machine learning models which are widely used across data science competitions will be trained and compared. For example, Random Forest Regression model (Breiman, 2001) is a supervised learning algorithm that uses ensemble learning for regression. Also, XGBoost model (Chen and Guestrin, 2016) is another powerful model which is used for regression predictive modelling.

To select the model with the best performance on the testing set, the metric which is widely used to evaluate the models is coefficient of determination (R2 score). R2 score can extract more information than MAE, RMSE and MSE in regression analysis. The R2 score is

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1.1}$$

where :

$R^2 = coefficient\ of\ determination$

$RSS = sum\ of\ squares\ of\ residuals$

$TSS = Total\ sum\ of\ squares$

According to Yin and Fan (2001), the adjusted R2 score has the capability to decrease with the addition of less significant variables which is a more reliable evaluation. The adjusted R2 score which I would use is

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{1.2}$$

where :

$\overline{R}^2 = adjusted\ R^2$

$n = sample\ size$

$p = number\ of\ independent\ variables$

Based on the model with the highest performance, it is decided to evaluate the highest importance of feature using model-agnostic methods. One of the methods is SHapley Additive exPlanations (SHAP) which was introduced by Lindberg and Lee (2017). SHAP is based on the game theoretically optimal Shapley values which is the average marginal contribution of a feature value across all possible coalitions. Based on the equations of Shapley values,

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\}) - val(S)) \tag{2.1}$$

where :

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X)) \tag{2.2}$$

S is a subset of the features used in the model, x is the vector of the feature values of instances and p is the number of features. $val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S.

With the SHAP values, the feature importance can be visualized and analysed using Decision plot and Beeswarm plot.

All the analysis and visualization in this research will be determined using Python with relevant programming packages: Pandas, Seaborn, Matplotlib, Scikit Learn, SciPy and Tensorflow.

3.5-Validity and reliability

Due to the nature of self-administrated surveys, the participants are all expected to be as truthful as possible. Even though the declaration about adults agreement are indicated in the survey, there is no way to prove if the participants are over 18 years old. There is always a concern that people would lie to show themselves while filling the surveys. To combat this, the researcher stressed that all information the participants given are confidential as they are not putting their full names on the survey and only the Instagram user names are collected.

To take into account the reliability of the study, the questions asked in the survey are not complicated and do not require the knowledge of any realms. Furthermore, all the questions are related to the usage of social media, which is very consistent.

3.6-Assumptions

I made several assumptions in this study. First, I assumed that everyone who participated in the study gave truthful answers when filling out the survey. Also, it is assumed that there is no technical mistake on the program while collecting and merging the data.

3.7-Scope and limitations

The scope of the study was that the participants were required to be 18 or older. I chose the study of potential factors which affect the time spent for Instagram users because this area has not been studied in the past or has not been able to be publicized by Instagram. Due to the short period of time for this research, it is not feasible to have a holistic exploration of all underlying factors on Instagram which affect the time spent for users. Besides, to take into account the complexity and specialization of this dissertation, it is impossible to make an in-depth analysis under the limitation of word count.

Chapter 4: Results

In this chapter, the results which include analysis findings, research questions and hypothesis testing. After a brief exploratory data analysis on data responses and demographics, the two research questions will be discussed.

4.1-Data responses and demographics

For 859 samples in the dataset, skewness and kurtosis were used to describe the shape of the distribution, whether normal or abnormal shape for different variables. Once the kurtosis had been reviewed, the heaviness of distribution tails can be determined. Figures 1 to 4 show the distributions of frequencies of each variables.

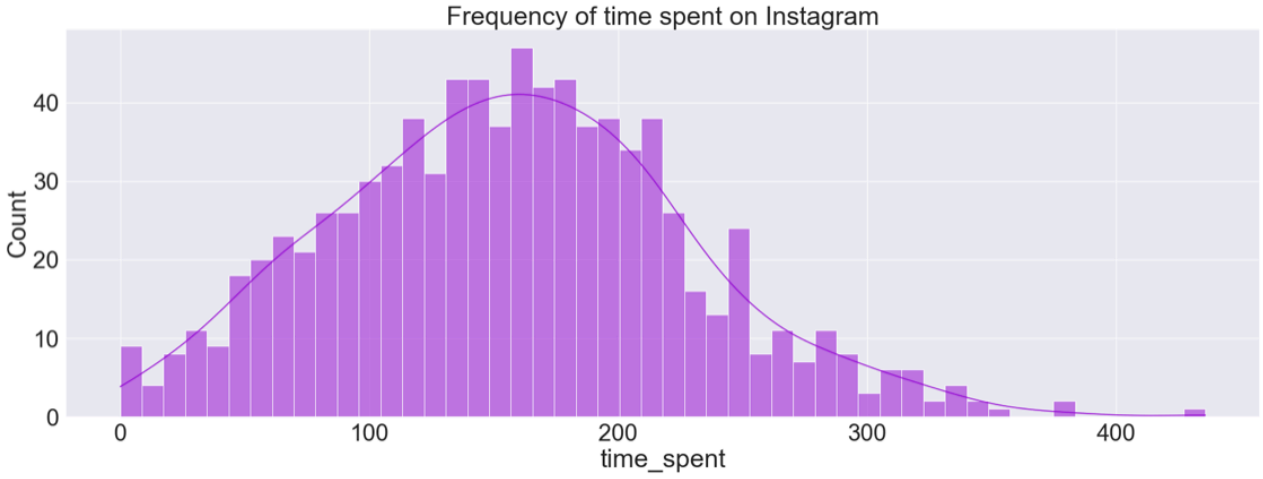


Figure 1. Frequency distribution of the time spent for Instagram users on Instagram

The skewness of time spent was 0.2565. Figure 1 shows the symmetrical distribution since the value of skewness is between -0.5 to 0.5. Even though the right-hand tail looks longer than the left-hand tail and it can be misclassified as positive skewness, the value of skewness illustrates that the distribution is fairly symmetrical.

The kurtosis in Figure 1 was 0.02617. Since the kurtosis is smaller than 3, the distribution for time spent is called platykurtic distribution.

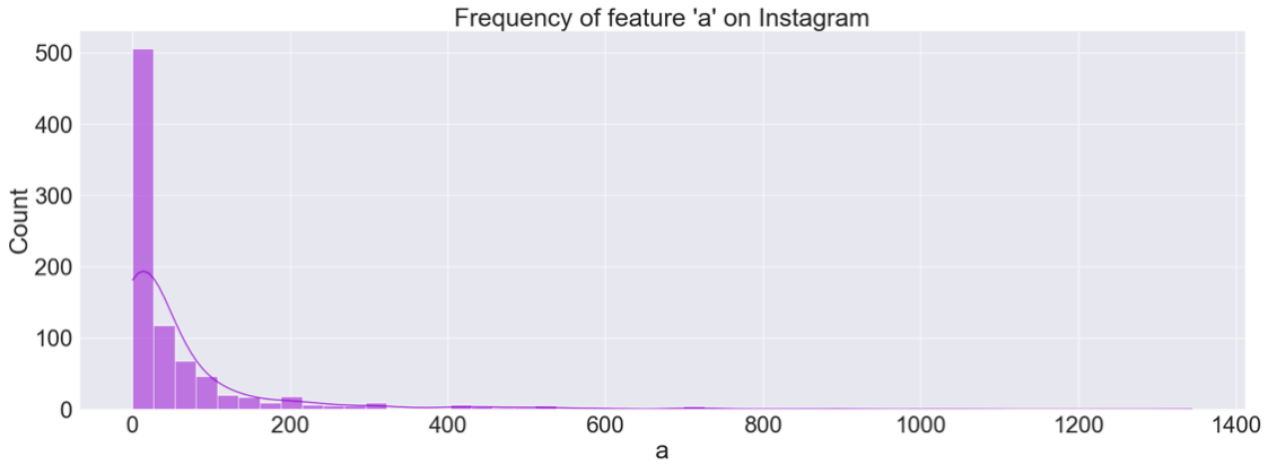


Figure 2. Histogram of the feature 'a' on Instagram

The feature 'a', according to Table 2, is the average number of likes for first 5 posts for each sample. The skewness of the feature 'a' is 4.437. Figure 2 shows the highly positive skewness since the value of skewness is larger than 1. The kurtosis was 26.45. Since the kurtosis is much greater than 3, the distribution for feature 'a' is called leptokurtic distribution.

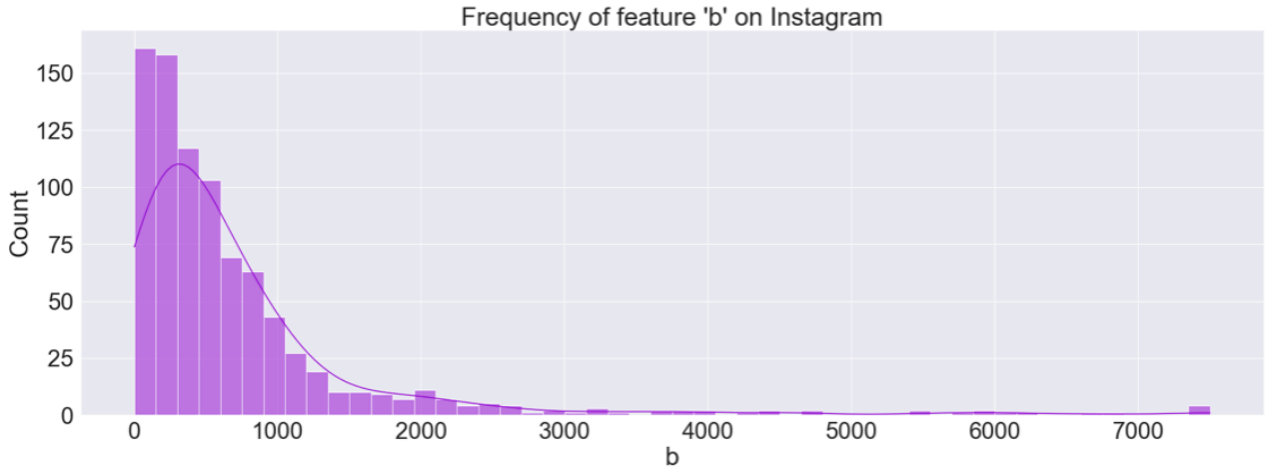


Figure 3. Histogram of the feature 'b' on Instagram

The feature 'b', according to Table 2, is the number of followings. The skewness of the feature 'b' is 3.814. Figure 3 shows the highly positive skewness since the value of skewness is larger than 1. The kurtosis was 18.13. Since the kurtosis is much greater than 3, the distribution for feature 'b' is called leptokurtic distribution.

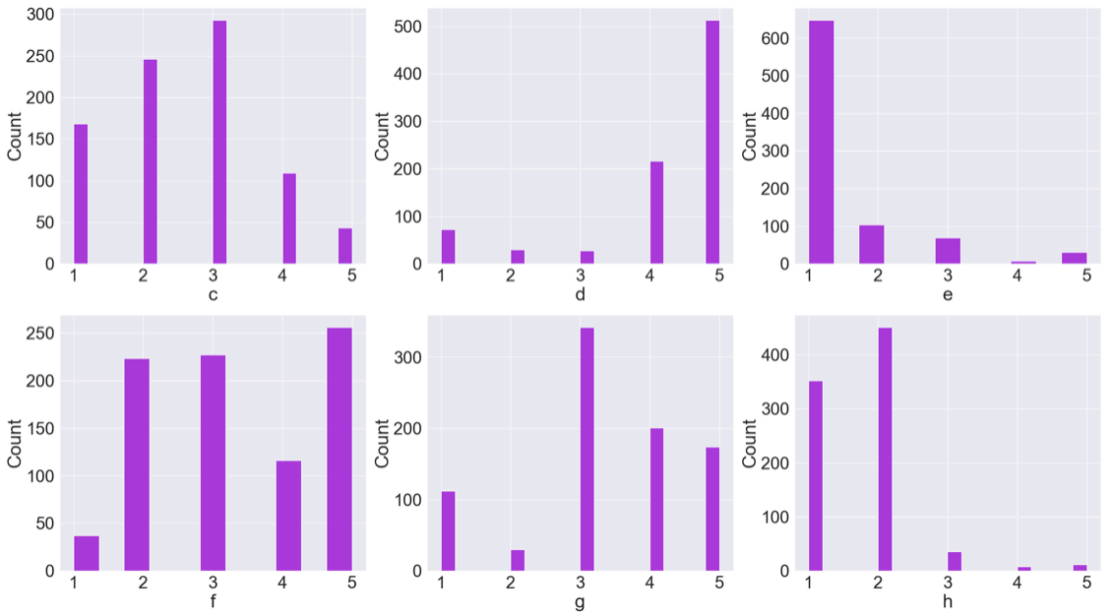


Figure 4. Frequency distribution of the categorical variables from 'c' to 'h'

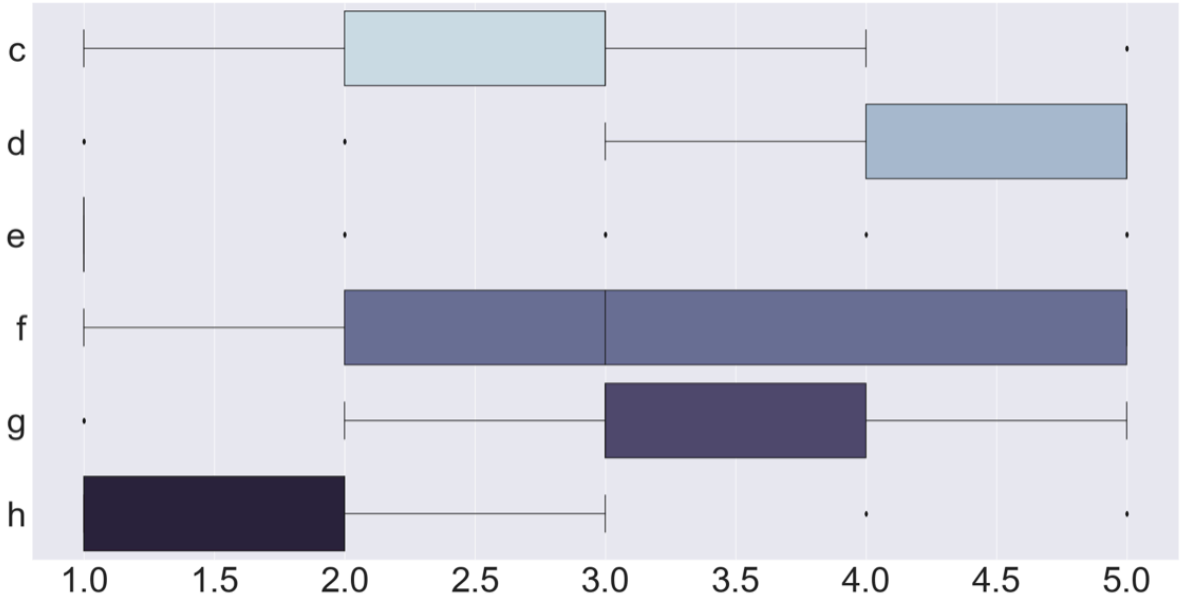


Figure 5. Box and Whisker plot of the categorical variables from 'c' to 'h'

Figure 4 shows the histogram of features 'c' to 'h' while Figure 5 shows the box and whisker diagrams of features 'c' to 'h'. The extent of features from 'c' to 'h' represent as '1'= 'never' to '5'= 'always'.

Feature 'c', according to Table 2, is the extent of using the Instagram story feature. The skewness of the feature 'c' is 0.2922. Figure 4 shows that it is symmetrical distribution since the value of skewness is between -0.5 and 0.5. The kurtosis of feature 'c' is -0.5120.

Since the kurtosis is lower than 3, the distribution for feature 'c' is called platykurtic. Figure 5 shows that the 25th and 75th percentile are at options 2 and 3.

Feature 'd', according to Table 2, is the extent of glancing Instagram stories from other users. The skewness of the feature 'd' is -1.730. Figure 4 shows that it is highly negative skewed since the value of skewness is smaller than -1. The kurtosis of feature 'd' is 1.881. Since the kurtosis is lower than 3, the distribution for feature 'd' is called platykurtic. Figure 5 shows that most samples are clustered at options 4 and 5.

Feature 'e', according to Table 2, is the extent of using the Instagram shop feature. The skewness of the feature 'e' is 2.384. Figure 4 shows that it is highly positive skewed since the value of skewness is larger than 1. The kurtosis of feature 'e' is 5.339. Since the kurtosis is larger than 3, the distribution for feature 'e' is called leptokurtic distribution. Figure 5 shows that nearly all samples are clustered at option 1.

Feature 'f', according to Table 2, is the extent of chatting with other users on Instagram. The skewness of the feature 'f' is -0.001948. Figure 4 shows that it is fairly symmetrical since the value of skewness is between -0.5 and 0.5. The kurtosis of feature 'f' is -1.304. Since the kurtosis is lower than 3, the distribution for feature 'f' is called platykurtic. Figure 5 shows that data is more distributed between options 2 to 5 and the median is option 3.

Feature 'g', according to Table 2, is the extent of scrolling through the 'Explore' and 'Reel' features on Instagram. The skewness of the feature 'g' is -0.4392. Figure 4 shows that it is fairly symmetrical since the value of skewness is between -0.5 and 0.5. The kurtosis of feature 'g' is -4392. Since the kurtosis is smaller than 3, the distribution for feature 'f' is called platykurtic. Figure 5 shows that the 25th and 75th percentile are options 3 and 4.

Feature 'h', according to Table 2, is The extent of spending time on Instagram advertisement. The skewness of the feature 'h' is 1.582. Figure 4 shows that it is highly positive skewed since the value of skewness larger than 1. The kurtosis of feature 'h' is 5.006. Since the kurtosis is larger than 3, the distribution for feature 'h' is called leptokurtic distribution. Figure 5 shows that most samples are clustered at options 1 and 2.

4.2-First research question: Is there a correlation between the time spent for Instagram users on Instagram and different kinds of potential features on Instagram?

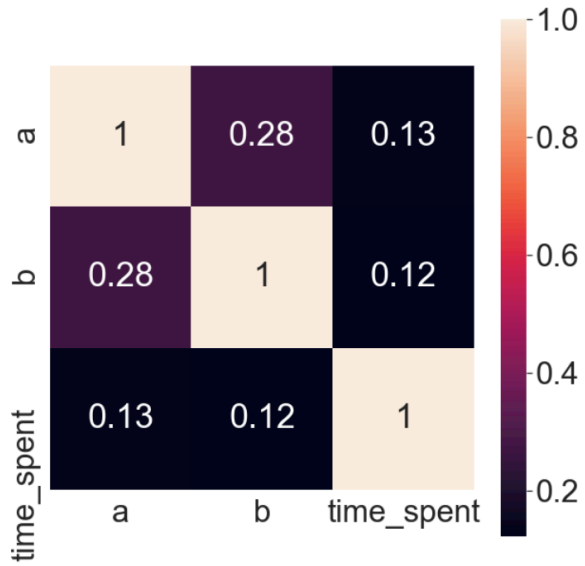


Figure 6. Heat map of continuous variables

The first research question will be analysed by evaluating Pearson correlation coefficient. The Figure 6 shows the heat map of correlation coefficient of continuous variables. The correlation coefficient of feature 'a' with time spent is 0.13 and feature 'b' with time spent is 0.12. Also, the correlation coefficient of feature 'a' and 'b' is 0.28.

For the first research question, students t-test is used to test the hypothesis test. The critical α level is specified as 0.05. For the RQ1.1, the p-value for the association between feature 'a' and time spent is 7.606e-5 which means that p<0.05. So the H_0 for the RQ1.1 can be rejected in favor of the H_a. Therefore, there is a significant correlation between the time spent for Instagram users on Instagram and the average number of likes for posts for each Instagram user.

For the RQ1.2, the p-value for the association between feature 'b' and time spent is 0.00032619 which means that p<0.05. So the H_0 for the RQ1.2 can be rejected in favor of the H_a. Therefore, there is a significant correlation between the time spent on Instagram and the number of people follows.

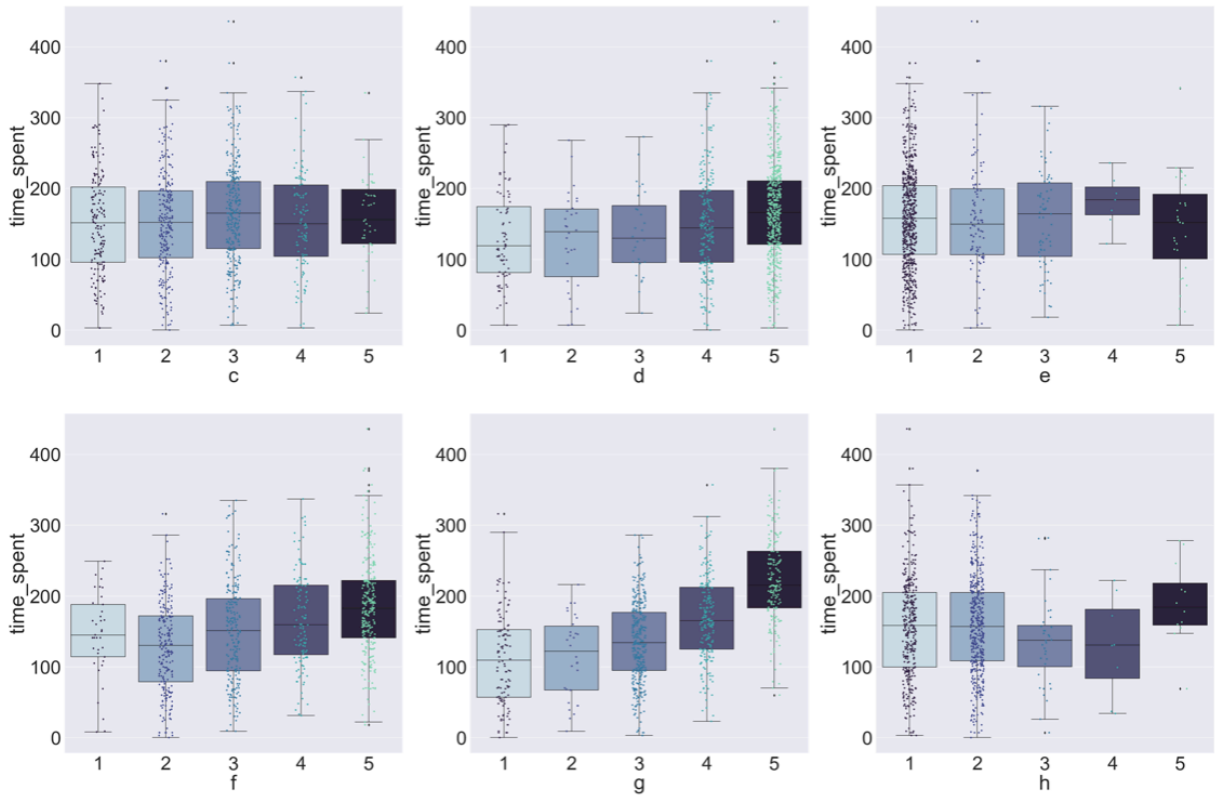


Figure 7. Box and strip plot for the relationship between categorical variables and time spent

Table 3.0 Pearson r and p-value

	Pearson correlation coefficient	**p-value**
c	0.04226	0.2160
d	0.1755	2.272E-07
e	-0.01996	0.5592
f	0.2967	6.535E-19
g	0.4984	3.855E-55
h	0.0005598	0.9869

The Figure 7 combines the box and whisker plot and the strip plot together to make a better interpretation on different categorical variables with time spent. The box and whisker diagrams show the spread of the data and the strip plots show the dots which represent each sample data.

For feature 'c', Figure 7 shows that the sample data is mainly distributed in option 1 to 4 and the interquartile range (IQR) for all options are similar. Table 3.0 shows that the correlation coefficient is 0.04226 for feature 'c' with time spent. The p-value for this

association is 0.2160 which means that p>0.05. So, H_0 for the RQ1.3 fail to be rejected. The t-test shows that there is no significant correlation between the time spent on Instagram and the feature 'c'.

For feature 'd', Figure 7 shows that the sample data is mainly clustered at option 5 and there is an increasing tendency for IQR. Table 3.0 shows that the correlation coefficient is 0.1755 for feature 'd' with time spent. The p-value for this association is 2.272E-07 which means that p<0.05. So, H_0 for the RQ1.4 should be rejected. The t-test shows that there is a significant correlation between the time spent on Instagram and the feature 'd'.

For feature 'e', Figure 7 shows that the sample data is mainly clustered at option 1 and the IQR is for all options order than option 4 are similar. Table 3.0 shows that the correlation coefficient is -0.01996 for feature 'e' with time spent. The p-value for this association is 0.5592 which means that p>0.05. So, H_0 for the RQ1.5 fail to be rejected. The t-test shows that there is no significant correlation between the time spent on Instagram and the feature 'e'.

For feature 'f', Figure 7 shows that the sample data is mainly distributed within these options and the IQR has an increasing trend. Table 3.0 shows that the correlation coefficient is 0.2967 for feature 'f' with time spent. The p-value for this association is 6.535E-19 which means that p<0.05. So, H_0 for the RQ1.6 should be rejected. The t-test shows that there is a significant correlation between the time spent on Instagram and the feature 'f'.

For feature 'g', Figure 7 shows that the sample data is mainly distributed between option 3 and option 5 and the IQR has an obvious increasing trend. Table 3.0 shows that the correlation coefficient is 0.4984 for feature 'g' with time spent. The p-value for this association is 3.855E-55 which means that p<0.05. So, H_0 for the RQ1.6 should be rejected. The t-test shows that there is a significant correlation between the time spent on Instagram and the feature 'g'.

For feature 'h', Figure 7 shows that the sample data is mainly distributed in option 1 and option 2 and the spread of option 1 and option 2 is similar. Table 3.0 shows that the correlation coefficient is 0.0005598 for feature 'h' with time spent. The p-value for this association is 0.9869 which means that p>0.05. So, H_0 for the RQ1.5 fail to be rejected. The t-test shows that there is no significant correlation between the time spent on Instagram and the feature 'h'.

4.3-Second research question: Which potential feature has the most statistical influence on affecting the time spent on Instagram for Instagram users?

To answer this research question, two machine learning models, random forest and XGBoost, will be trained to determine the feature with the most statistically influential. First of all, it is necessary to normalize the features 'a' and 'b' to have similar a range with ordinal features. Next, the original dataset needs to be split into a training set and testing set in order to prevent overfitting and underfitting. The ratio of training set and testing set is 9:1. After finishing training the models, the adjusted R2 score for random forest is -0.05324 while the adjusted R2 score for XGBoost is 0.1074. It shows that XGBoost performs better on fitting the dataset. So, XGBoost will be picked to calculate the SHAP values and evaluate the feature importance.

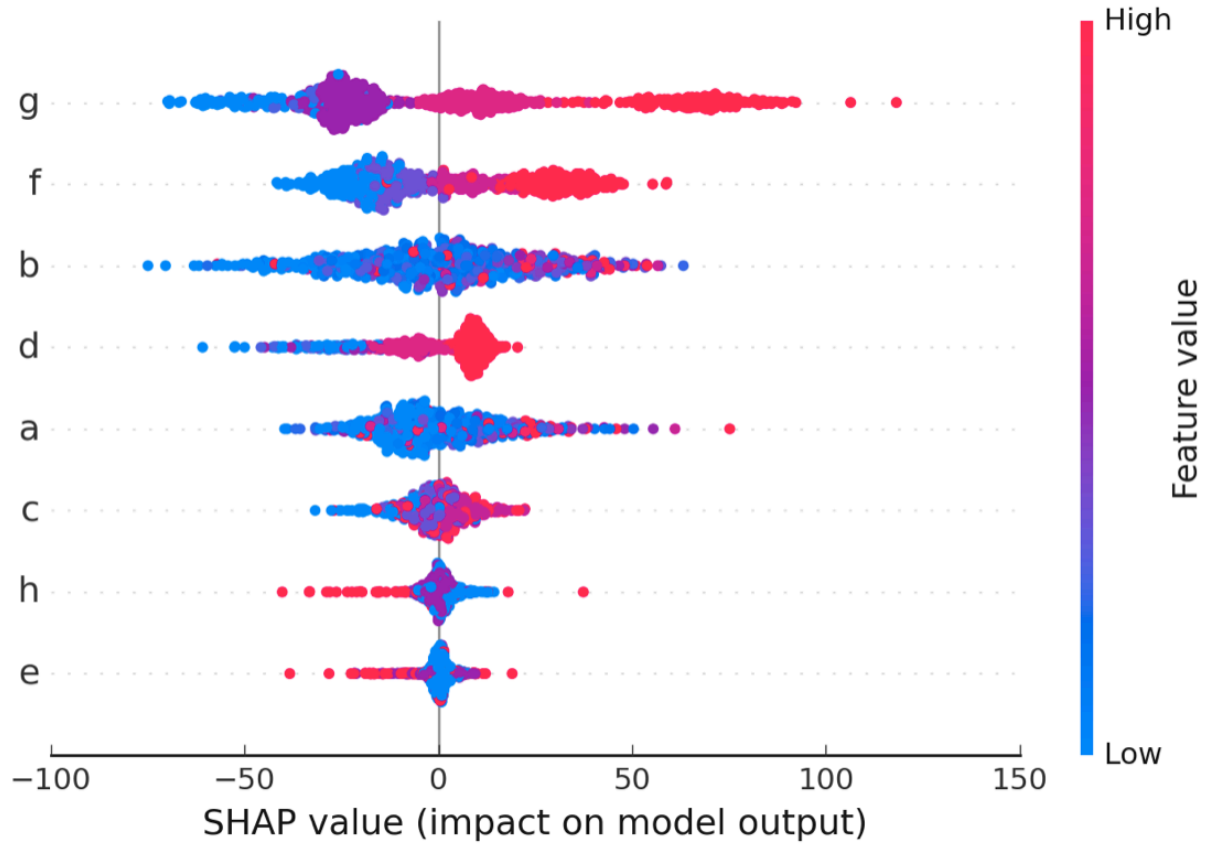
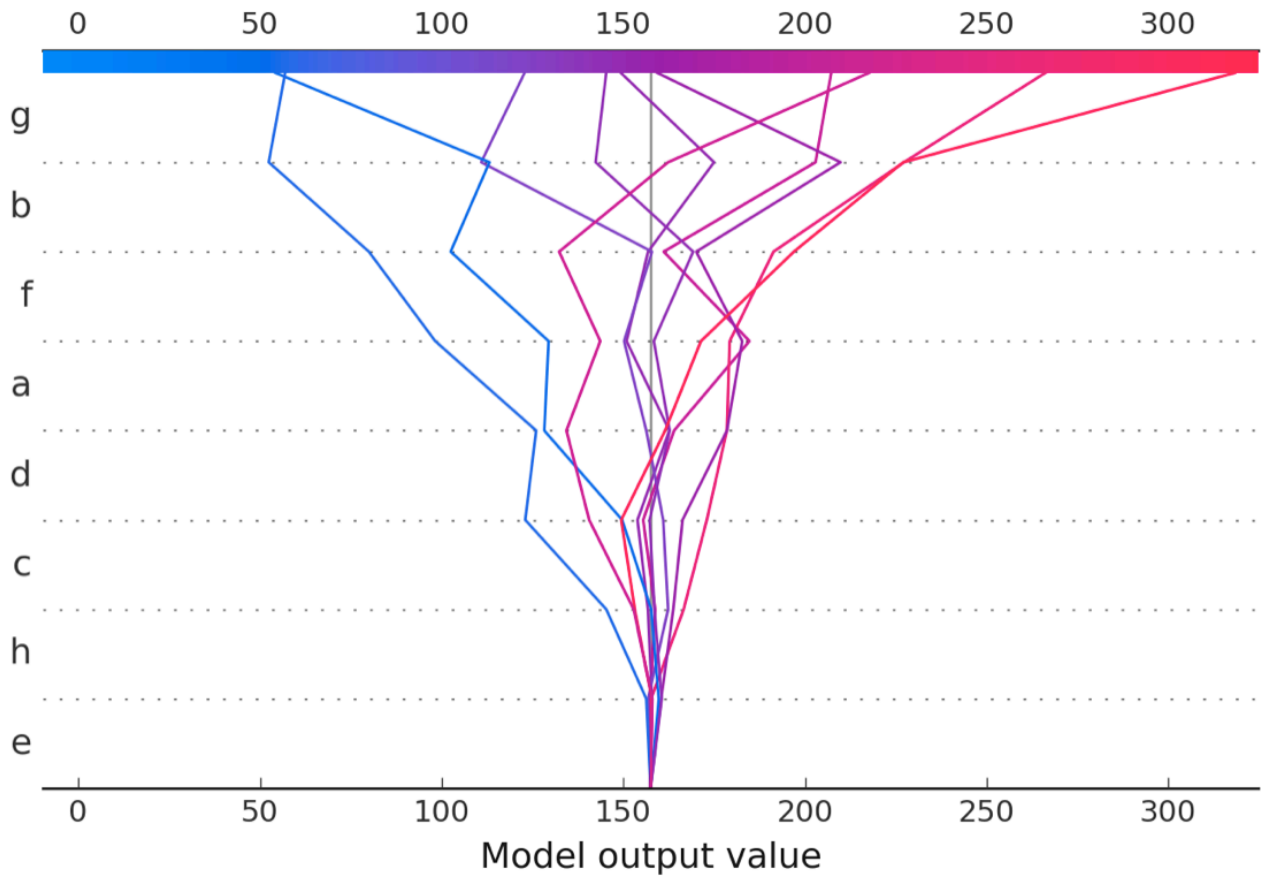


Figure 8. Beeswarm plot for the SHAP values of different features

As seen in Figure 8, this is a Beeswarm plot of all the SHAP values. The values are grouped by the features on the y-axis. For each group, the colour of the point is determined by the value of the same feature. The features are ordered by the mean SHAP values. For example, the feature 'g' has the largest mean SHAP and the feature 'e' has the lowest mean

SHAP out of all the features. In other words, feature 'g' has a significant impact on the model's predictions. In contrast, feature 'e' has a slight impact on the model's predictions. In addition, Figure 8 also reveals that the SHAP value increases when the feature value increases for most of the feature. However, feature 'h' has the opposite relationship. Larger values for this feature are associated with smaller SHAP values.



In Figure 9, there are 10 lines in the plot and each line corresponds to a sample. They all start at the same base value of around 160 and end at their final predicted time spent. As moving up from each feature on the y-axis, the movement on the x-axis is given by the SHAP value for that feature. With only 10 observations, we can see that some of the lines have more zig-zag movement for features at the top of the y-axis than the features at the bottom of the y-axis. All in all, it clearly shows that feature 'g' has the most statistical influence on affecting the time spent on Instagram.

Chapter 5: Discussion

In RQ1.1, the t-test shows that there is a significant correlation between the time spent on Instagram and the average number of likes for the top five posts. It illustrates that this feature has potential influence to affect Instagram users' willingness to spend time to post photos or videos and keep track of the number of likes for each post. Further studies according to Reijmerink, Couquax and Kocijan (n.d.) have also shown that people who do not get as many 'likes' as they expected will have low self-esteem. In order not to feel stressed and anxious, Instagram users may become addicted to spending time to keep track of the 'likes' for the posts.

In RQ1.2, the t-test shows that there is a significant correlation between the time spent on Instagram and the number of people each Instagram user follows. As the number of people each Instagram user follows increases, the user will receive more and more posts. Also, posts are the main approach for everyone to spread information on Instagram. As a result, people would spend time glancing at posts which are posted by other Instagram users.

In RQ1.3, the t-test shows that there is no significant correlation between the time spent on Instagram and the extent of using the Instagram story feature. It is supposed that this feature affects the willingness of users to spend time on posting Instagram stories and keeping track of the Instagram story views. Similar to the RQ1.1, users may care more about how many people have seen their story posts and this feature depends on the number of followers the user has. However, the analysis I did shows the opposite view, which illustrates that there is no significant relationship between these two variables.

In RQ1.4, the t-test shows that there is a significant correlation between the time spent on Instagram and the extent of glancing at Instagram stories from other users. Undeniably, glancing at Instagram stories from other users increases the time spent on Instagram. The stories feature has opened up an entirely separate universe of opportunities on Instagram. Compared to the 'post' feature, 'story' feature is more convenient and flexible to share content and the story will disappear after 24 hours. According to Maya (2021), there are over 500 million users who use stories daily and most users are teenagers who are between 18 and 34. It shed light on the potential factor that Instagram users would receive a large amount of stories from other people every day.

In RQ1.5, the t-test shows that there is no significant correlation between the time spent on Instagram and the extent of using the Instagram shop feature. Even though there is

the 'shop' feature which allows users to buy items on Instagram, most of Instagram users in the sample seldom spend time on this feature.

In RQ1.6, the t-test shows that there is a significant correlation between the time spent on Instagram and the extent of chatting with other users on Instagram. Study has shown that Instagram users who spend time on 'message' feature on Instagram will increase their time spent on Instagram.

In RQ1.7, the t-test shows that there is a significant correlation between the time spent for users on Instagram and the extent of scrolling through the 'Explore' and 'Reel' features on Instagram. As a feature which is similar to 'TikTok', which has around a billion active users monthly (Dean, 2022), it has germane characteristics with 'TikTok' that affect users' mental health and users are addicted in 'TikTok' to spend 52 minutes per day (-, 2020). Also, the 'Explore' feature, which has a similar algorithm for designing the recommendation system, attracts Instagram users to spend time on this feature.

In RQ 1.8, the t-test shows that there is no significant correlation between the time spent for Instagram users on Instagram and the extent of spending time on Instagram advertisements. It means that there are few students spending time on Instagram advertisements or the Instagram advertisements would not affect the time spent on Instagram.

In the second research question, I determined the feature with the most statistical influence on affecting the time spent on Instagram. With the analysis using XGBoost, it shows that the feature 'g' , the extent of scrolling through the 'Explore' and 'Reel' features, has the most statistical influence on affecting the time spent on Instagram. As discussed in RQ1.7, the reasons for this are obvious. Owing to the complex algorithm of recommendation system for 'Explore' and 'Reel', they always recommend the videos and photos which Instagram users are intrigued in. According to Allan (2021), the Instagram 'Reels' algorithm would predict users' preference depending on how long the users stay on a short video or if the users give a like to the short video. Users would not feel bored when glancing through the 'Explore' and 'Reels' since the system would recommend new and different videos. Besides, a statistic by Santora analyzed that these kinds of features are widely applied for Instagram marketing strategy to sell or promote the products. For example, Santora also mentioned that Louis Vuitton used the 'Reels' feature to get an average of 7 millions views and the short videos posted by Red Bull already has a number

of viral Reels with more than 2.4 million views. It shed light that the 'Explore' and 'Reels' feature has a substantial influence on affecting the time spent on Instagram.

Chapter 6: Summary, Conclusion and Recommendation

Among these features, the study concluded that some potential factors, number of likes for the posts, number of followings, the extent of glancing Instagram stories from other users, the extent of chatting with other users on Instagram and the extent of scrolling through the 'Explore' and 'Reel' features on Instagram, have significant influence on affecting the time spent for Instagram users on Instagram. In addition, it has been analysed that the most statistically influential factor on affecting the time spent for Instagram users is the extent of scrolling through the 'Explore' and 'Reel' features on Instagram. And it is the reason why there are a lot of companies use this feature on Instagram to increase their sales and make some promotion.

For further research about this topic, it is recommended that other proper hypothesis testing approaches can be analyzed. Rather than using Student t-test to test the significant difference between correlation coefficient of continuous variable and different categorical variables, it is better to use other approaches, such as logistic regression and Kruskal-Wallis H Test. Also, the assumption on the potential factors which affect the time spent for users can be more holistic.

Reference

Swisseducation.com. 2021. *Education First and Swiss Education Group announce new partnership*. [online] Available at: <https://www.swisseducation.com/en/news/corporate-news/education-first-and-swiss-education-group-announce-new-partnership-5328/> [Accessed 7 February 2022].

Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

Yin, P., & Fan, X. (2001). Estimating *R*2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, *69*(2), 203–224. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.

Olufadi, Y. Social networking time use scale (SONTUS): A new instrument for measuring the time spent on the social networking sites. Telematics and Informatics 33 (2016), 452-471

Reijmerink, A., Couquax, J. and Kocijan, L., n.d. *Why Instagram "likes" effect our self-esteem (and how scrapping them might help)*. [online] The Indigo Project. Available at: <https://theindigoproject.com.au/why-instagram-likes-effect-our-self-esteem-and-how-scrapping-them-might-help/> [Accessed 8 February 2022].

TrueList. 2021. *Instagram Stories Stats - TrueList 2022*. [online] Available at: <https://truelist.co/blog/instagram-stories-stats/> [Accessed 8 February 2022].

Dean, B., 2022. *TikTok User Statistics (2022)*. [online] Backlinko. Available at: <https://backlinko.com/tiktok-users> [Accessed 8 February 2022].

Medium. 2020. *How TikTok Is Addictive*. [online] Available at: <https://medium.com/dataseries/how-tiktok-is-addictive-1e53dec10867> [Accessed 8 February 2022].

Allan, A., 2021. *How to Crack the Instagram Reel Algorithm*. [online] ManyChat Blog. Available at: <https://manychat.com/blog/instagram-reel-algorithm/> [Accessed 8 February 2022].

Santora, J., 2021. *15 Instagram Reels Statistics That Will Blow Your Mind*. [online] Influencer Marketing Hub. Available at: <https://influencermarketinghub.com/instagram-reels-stats/> [Accessed 8 February 2022].

Jennings, R., 2021. *Can social media ever be truly "body positive"?*. [online] Vox. Available at: <https://www.vox.com/the-goods/22226997/body-positivity-instagram- tiktok-fatphobia-social-media> [Accessed 8 February 2022].

Kemp, S., 2021. Instagram stats and trends. [online] DataReportal. Available at: <https://datareportal.com/essential-instagram-stats> [Accessed 18 December 2021].

Treitel, Yael, "The Impact of Instagram Usage and Other Social Factors on Self-Esteem Scores" (2020). *Walden Dissertations and Doctoral Studies*. 7959. https://scholarworks.waldenu.edu/dissertations/7959

Yesilyurt, F. & Solpuk Turhan, N. (2020). Prediction of the time spent on instagram by social media addiction and life satisfaction. Cypriot Journal of Educational Science. 15(2), 208–219. https://doi.org/10.18844/cjes.v15i2.4592

Kirik, A. M., et al. (2015). A Quantitative Research on the Level of Social Media Addiction among Young People in Turkey. 16.

Arslan, A. and Kırık, A.M. (2013). "Sosyal Paylaşım Ağlarında Konum Belirleme Ölçeğinin Geçerlik Ve Güvenirlik Çalışması", Öneri Journal, Vol:10, Issue:40, pp.223-231.