SIGNIFICANT IMPACT OF SOCIAL MEDIA ON HUMANS-
A MACHINE LEARNING APPROACH

A PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

BY

D.HANEESH
(Roll No: 18331A0528)
D.NANDINI
(Roll No: 18331A0537)

B.G.G.S.PRASAD
(Roll No: 18331A0514)
D.YASASWI
(Roll No: 18331A0531)

Under the Supervision of

Mr. P RAMA SANTOSH NAIDU

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MVGR COLLEGE OF ENGINEERING (Autonomous)

VIZIANAGARAM-535005, AP (INDIA)

(Accredited by NBA, NAAC, and Permanently Affiliated to Jawaharlal Nehru Technological University Kakinada)

JUNE, 2022

SIGNIFICANT IMPACT OF SOCIAL MEDIA ON HUMANS- A MACHINE LEARNING APPROACH

A PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

BY

D.HANEESH
(Roll No: 18331A0528)

B.G.G.S.PRASAD
(Roll No: 18331A0514)

D.NANDINI
(Roll No: 18331A0537)

D.YASASWI
(Roll No: 18331A0531)

Under the Supervision of

Mr/Ms/Dr/Prof. P RAMA SANTOSH NAIDU

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MVGR COLLEGE OF ENGINEERING (Autonomous)

VIZIANAGARAM-535005, AP (INDIA)

(Accredited by NBA, NAAC, and Permanently Affiliated to Jawaharlal Nehru Technological University Kakinada)

JUNE, 2022

Maharaj Vijayaram Gajapathi Raj (MVGR) College of Engineering(A)

Vizianagaram

CERTIFICATE

This is to certify that the project report entitled "**SIGNIFCANT IMPACT OF SOCIAL MEDIA ON HUMANS-A MACHINE LEARNING APPROACH**" being submitted by **D.HANEESH, B.G.G.S.PRASAD, D.NANDINI, D.YASASWI** bearing registered numbers **18331A0528, 18331A0514, 18331A0537, 183310A0531** respectively, in partial fulfillment for the award of the degree of "**Bachelor of Technology" in Computer Science and Engineering** is a record of bonafide work done by them under my supervision during the academic year 2018-2022.

Dr. P. RAVI KIRAN VARMA
Head of the Department,
Department of CSE,
MVGR College of Engineering,
Vizianagaram.

Mr. P. R. S. NAIDU
Assistant professor,
Department of CSE,
MVGR College of Engineering,
Vizianagaram.

External Examiner

ACKNOWLEDGEMENTS

We would like to express our indebtedness and sincere thanks to our supervisor **Mr. P. R. S. Naidu** for his suggestions and support from an early stage of this project and providing us experiences throughout the work. Above all, his meticulous supervision at each and every stage of the project inspired us in many ways specially acknowledge him for his supervision, and the vital contribution as required during the research. His involvement with originality has nourished our intellectual maturity that will help us through out the project and a long time to come. we are proud to record that we had the opportunity to work with an experienced Professor like him.

We also thank **Dr. K. V. L. Raju**, Principal, and **Dr. P. Ravi Kiran Varma**, Head of the Department, for their support and cooperation in providing all the tools for us to successfully complete the project. We sincerely thank everyone from the Department of Computer Science & Engineering for their help in our pursuits. We thank all those who contributed directly or indirectly in successfully carrying out this project.

D.HANEESH
B.G.G.S.PRASAD
D.NANDINI
D.YASASWI

ABSTRACT

Social media has become a better part of our lives in today's world. Most of us are so involved in social media apps like Facebook, WhatsApp, Snapchat etc. The more people are involved the more they face the risk of exposure to the world. Results of many researches have proven that social media has a higher impact and influence on people. Social influence plays a prominent role in shaping people’s behavior now-a-days and affecting human decisions in various domains. People's mental health could have both positive and negative impact due to social media, some of the negatives might be depression, anxiety, self-harm whereas the positives might be a feeling that we are not alone in this busy world. Social media plays an important role in influencing people in their opinions on topics like the environment, politics, and society. One of the main negatives of social media is cyberbullying, lack of privacy and security. The main objective of this project is to show the impact or influence of social media in the form of statistics by comparing different quantities like different social media sites, number of people on it, amount of time spent, etc with the help of data visualization and machine learning techniques.

CONTENTS

List of Abbreviations

Osn	-online social network
Ml	-machine learning
Slr	-simple linear regression
Rvalue	-The correlation between the predictor variable and the response variable
PValue	-is a statistical test that determines the probability of extreme results of the statistical hypothesis test, taking the Null Hypothesis to be correct
Pd	-pandas
Np	-NumPy
Df	-data frame
R	-Squared-is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model
LE	-linear equation
SM	-social media
FB	-Facebook
WAPP	-WhatsApp
IG	-Instagram
SC	-snapchat
TWTR	-twitter
YT	-YouTube

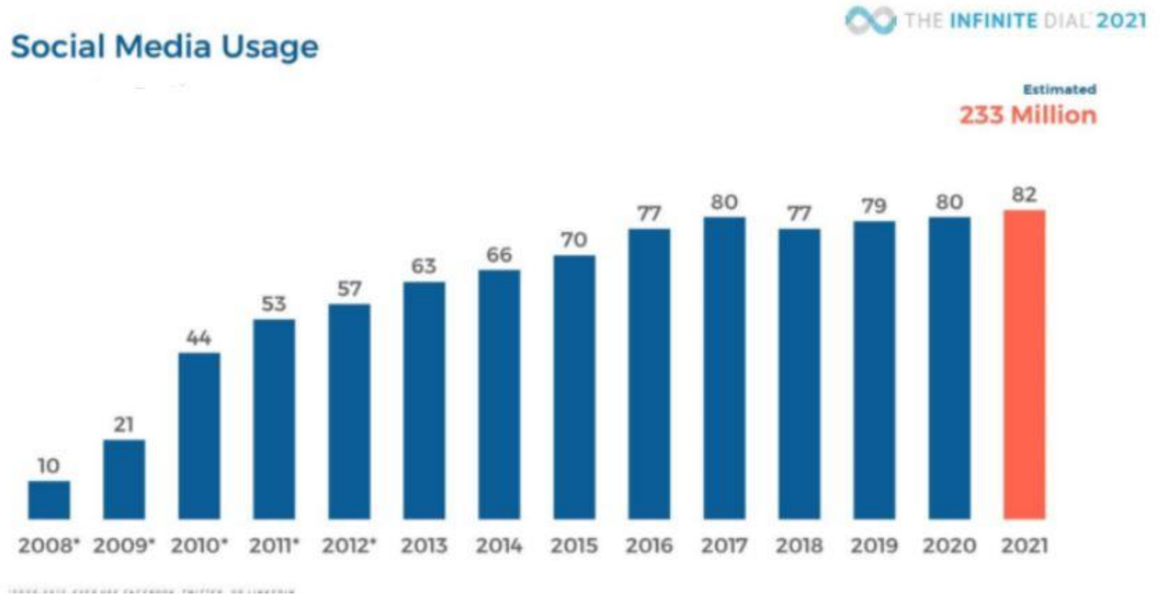
List of Figures

CHAPTER-1

INTRODUCTION

We recognize nearly 80% of the world's populace is at the social media, from this we can set up that we aren't alone on this world. Social media interest has an effect on each thing of the society like social or affairs of state may be mentioned on this platform without bias. The Cyberbullying Institute's 2019 survey of U.S. observed that over 36% of college students proceedings are approximately those cybercrimes and almost 11% of them admitted to had been bullying humans over the internet, young adults can misuse social media systems to unfold rumors and to blackmail others.

Social media has grown ingrained in our everyday lives and has a big influence on our actions. However, how has social media influenced our everyday lives, and are there any discernible patterns? In this group assignment, we will answer four questions based on data from a survey of people. To begin, on which social media sites do individuals spend the most time? Second, how does the amount of time spend on social media affect people's physical activity? Third, does the amount of time people spend on social media affect their risk of being a victim of cybercrime? Finally, what impact does communication type choice have in people's time allocation to social media? How does social media impact our lives?



1.1 Figure: Social media usage over the years

1.1 Identification of seriousness of the problem:

Cybercrimes is the most common issue that public is facing right now .most of the people getting targeted are students and teenagers. people spending more time on social media rather than spending on physical activities. The impact of people spending more of their time on social media affects everyone with many issues. cybercrime is the most common issue that public is facing right now. People that are mostly targeted are students and teenagers. The statistics concerning these issues are increasing by the day. People these days are more accepting towards indirect communications like social media rather that face to face is creating more opportunities for cyberbullies to extract personal information on people.

1.2 Problem definition:

With the help of this project, we attempt to analyze and find the statistics by comparing different attributes like the most effectively used social media, its impact on the people and whether they are prone to any kind of cybercrime due to over-usage of a particular social media, time spent on physical activity.

1.3 Objective:

We conducted a survey among different age groups in which we explored the following key questions:

- Which social media site or platforms do you prefer or utilize the most?
- How much time do you spend on social media vs exercise?
- Have you been a victim of cybercrime before?
- What is the preferred method of communication?

CHAPTER-2

LITERATURE SURVEY

In [1] they mentioned that the impact of social media is basically not only on human behavior but also has its wide impact on politics. Having higher influence on social media can make a candidate win or lose the elections. A survey has been conducted in the recent times which showed that the politicians in the Netherlands (2010-2011) with higher social media engagement tends to receive higher number of votes among all the political parties. There is a article that studies about how social media helps the political parties to reinvent and also to interact with the common people and find the solutions to their problems.

There's always been a concern that needs to be addressed which is between the social media use and mental health and wellbeing in young people. In [2] they studied about the social media usage and later mental health and wellbeing in adolescents of 12866 young people of 13-16 years of age in England which also concluded that they are effected with cyber bullying, sleep adequacy and physical activity.

Social media has effect on human health in both positive and negative ways. In [3] they studied about social media impact on human health stated that the more time people spend on the social media the more likely that they suffer from mental illness. The primary concerns are decreased self-esteem, eating disorders, anxiety, feelings of inferiority, declined focus in work, etc. The positive thing about social media is it guides in terms of getting health related information, that might be helpful in curing the disease similarly it also has a negative effect that if the information that is mentioned is fake then it might lead to some serious problems or it might create some mental tension in person's head. People might also lose their life if they follow the medication blindly that is available on internet.

In [4] they showed some recent views on social media network that is related to the business sector. other than the positives of sharing information worldwide, social media networks allows people to create false identifications, spread negative rumors or false information, post videos that kills people's reputation, and blackmail others. Finally we can say that since social media networks are still growing and giving much more opportunities to all the people to either become a positive impact on society or a negative one.

In [5] there is a direct relation between social media and human behavior in society. This means that the higher the students use the social media higher is the influence in social activities done together with people from various aspects of the background in the society.

Cybercrime refers to any criminal activity that is done with the help of a computer or internet. Examples of cybercrime include email spamming, identity theft, online child pornography, phishing and sending a virus to other computers. In [6] according to Bromium a cyber security firm, the social media platforms that are used to communicate with friends and family is also a huge global cybercriminal network. Nearly 1 out of 5 organizations worldwide are now infected by malware or virus that are sent through social media platforms according to Bromium company.

In [7] they proved how social self-efficacy, self-esteem, and need to belong somewhere can be used to predict teenager's use of various social media sites. The focus of this paper was particularly on how these social psychological variables together with social media use provides variations in teens' participation in a flash mob – an example for 21st-century collective action. The data is collected from a survey that takes the teens in the USA. Teens' need to belong has positively impacted the amount of time they reported spending on social networking sites, when controlling for gender, race, and household socio-economic activity status.

CHAPTER-3

THEORETICAL BACKGROUND

3.1 MACHINE LEARNING

3.1.1 What is Machine Learning?

One of the most trending technology that is available in the market is Machine learning. It is the study of the computer algorithms that learns through experience and using the data that is available. Machine learning is a part of artificial intelligence. The data that is used in machine learning can be categorized into two categories training data and testing data. Machine learning algorithms helps in building a model that is trained against the training dataset to make predictions or decisions without being explicitly programmed. After training is completed that machine learning model is tested against the testing dataset and then accuracy of the model is measured. Machine learning algorithms has widely used in various number of fields like health department, email filtering, spam filtering, fake profile detection etc. where it is generally not possible to carry the tasks with conventional algorithms.

3.1.2 Why Machine Learning?

Machine learning involves computers doing the task without explicitly programmed and perform tasks using the knowledge they gain through experience. For simpler kind of tasks it is possible to instruct the machines how to perform each and every step but for complicated kind of tasks it is generally difficult to instruct every step the computer needs to understand and learn through experience other wise it is very difficult to perform the task. There are multiple approaches to teach computers to accomplish tasks where there is no guarantee that it is 100% satisfied. In chance there are huge number of answers exist, one approach is to label some of the correct answers as valid. They can be using as training dataset and helps in building the model and then it can tested against the testing dataset. For example consider in order to train a system for the task of digital character recognition, then we need to use the MNIST dataset of handwritten digits.

3.2 DATA VISUALIZATION TECHNIQUES

Data visualization is important in machine learning. Statistics focuses on quantitative descriptions and estimations of data. Data visualization provides an understanding on how important the tools are for gaining a qualitative understanding on our data. This is important in knowing and exploring the dataset and can help with identifying patterns, corrupt data, outliers, missing data. Data visualizations can be useful in expressing the key relationships in plots and charts that hold significance for better understanding of the data.

There are five key plots that you need to know well for basic data visualization. They are:

- Line Plot
- Bar Chart
- Histogram Plot
- Box and Whisker Plot
- Scatter Plot

3.2.1 LINE PLOT

A line plot is used to present attributes collected at regular intervals. The x-axis represents the regular interval, such as time. The y-axis shows the observations, ordered by the x-axis and connected by a line. A line plot can be created by calling the plot() function and passing the x-axis data for the regular interval, and y-axis for the observations. Line plots are useful for presenting time series data as well as any sequence data where there is an ordering between different attributes.

3.2.2 BAR CHART

A bar chart is used to present relative quantities for multiple categories. The x-axis represents the categories and are evenly spaced. The y-axis represents the quantity for each category and is drawn as a bar from the baseline to the appropriate level on the y-axis. A bar chart can be created by calling the bar() function and passing the category names for the x-axis and the quantities for the y-axis. Bar charts can be used to compare multiple point quantities or estimations.

3.2.3 HISTOGRAM PLOT

A histogram plot is used to summarize the distribution of a data sample. The x-axis represents discrete intervals. The y-axis represents the count of observations in the dataset that belong to each interval. A histogram plot can be created by calling the hist() function and passing in a list or array that represents the data sample. Histograms are valuable for summarizing the distribution of data samples.

3.2.4 BOX PLOT

Box plots are useful to summarize the distribution of a data sample as an alternative to the histogram. They can help in getting an idea of the range of common values in the box and in the whisker respectively. Because we are not looking at the shape of the distribution, this method is often used when the data is an unknown or unusual distribution, such as non-

Gaussian. Box plots can be drawn by calling the box plot() function passing in the data sample as an array or list.

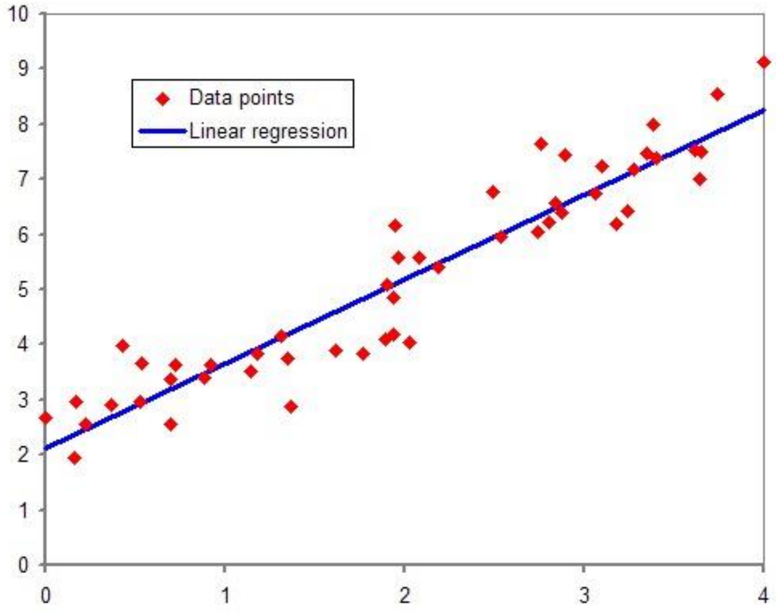
3.2.5 SCATTER PLOT

Scatter plots are used to show correlation between two variables. A correlation can be quantified, such as line of best fit, making the relationship clearer. A data set may have more than two measures for a given observation. A scatter is generally used to summarize the relationship between two paired data samples. Scatter plots can be created by calling the scatter() function and passing the two data sample arrays.

3.3 LINEAR REGRESSION

Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y).



3.1 Figure: Linear Regression Graph

CHAPTER-4

APPROACH DESCRIPTION

4.1 APPROACH FLOW

- First we collect data by conducting survey.
- In the next step we read the input data from the survey conducted and based on the responses we received, we particularly target the youth below age 21.
- Analysing data by identifying and handling missing values from the input file.
- By the process of vectorization, we convert the textual data into numerical data.
- We need to choose the right visual representation so we can find the relations between the different features.
- Display the output to view the impact of social media on people.

CHAPTER-5

DATA EXPLORATION

The data that we handle in this project is taken from the survey form that is conducted and collected from various different colleges. That information is used as dataset. The dataset consists of questionnaire where people answer to those questions irrespective of their age group which helps to find out the answer to the following questions and also find the relations between multiple attributes

- Which social media platform/s do you like the most or use the most?
- How much time do you spend on social media in a day?
- How much time do you spend on physical activities in a day?
- Have you ever been a victim of any of these cyber crimes?
- Which type of communication do you generally prefer?

5.1 Youth Data

The survey is sent to all people irrespective of age group. The responses we received are mostly from 18-21 age group. We targeted them in particular So that, we can focus on one specific age group. We are finding out all different possible relations that we can obtain from the attributes.

5.1.1 Data Manipulation

Data manipulation refers to the process of adjusting data to make it organized and easier to read. It allows us to update, modify and delete and also perform crud operations in the database. This enables us to expand our scope in business decisions and also there is a high chance that the data we obtained consists of fake, erroneous or unwanted information which must be handled in-order to have a model with good accuracy. The data in general can be of different formats like .xlsv, .csv, .txt. so in-order to maintain some consistency we follow standard format that is the data should be converted into .csv format

5.1.2 Data preparation

Generally, if a dataset is available in the internet it consists of many unnecessary fields, erroneous data but since we are considering our own dataset which we obtained from the survey. So we collect only data that is required to us. hence our dataset consists of required fields only and we must also check if there are any missing values.

5.1.3 Missing Values:

Some users might not fill all the fields, the fields which respondents do not answer are considered as missing values and more no of missing values leads to less accuracy so the missing values must be handled properly.

Handling of missing values is generally of 2 types:

- Filling the missing values
- Ignoring the entries

If we fill the missing values with random values this might again lead to less accuracy. one of the efficient ways to fill these missing values is by finding the correlation between the other attributes and fill those values and also ignoring all the enteries might be bad idea if more than half of our dataset consists fields of missing values.

CHAPTER-6

DATA ANALYSIS

6.1 Data visualization

Data Visualization is the graphical representation of information and data in a pictorial or graphical format. Data visualization tools provide an easy way to see and understand trends, patterns in data, and outliers. Data visualization tools and technologies are essential to analyzing massive amounts of information and making data-driven decisions. The concept of using pictures is to understand data that has been used for centuries. General types of data visualization are Charts, Tables, Graphs, Maps, Dashboards.

6.1.1 Age Factor

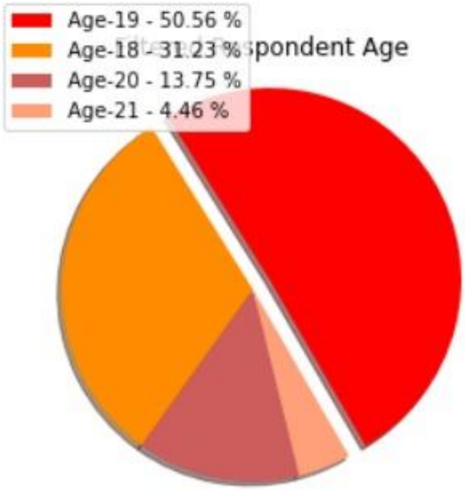


Figure 6.1 : pie chart representing respondent ages

This is a piechart representing the ages of the respondents, as we can see 50.56% of them are 19 year olds who use social media the most.

6.1.2 Time spent on social media in a day

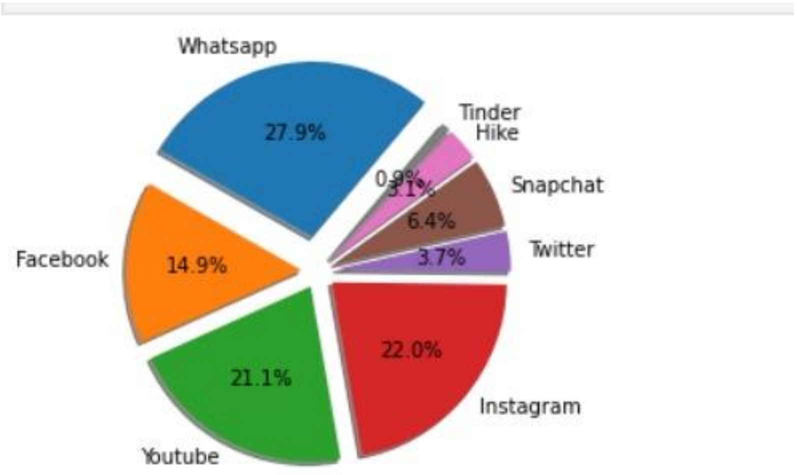


Figure 6.2 : Preferred social media sites

This represents the percentages of different social media sites that are used by the age group of 18 to 21. The most popular social media site from the pie chart is Whatsapp followed by Instagram, YouTube and Facebook.

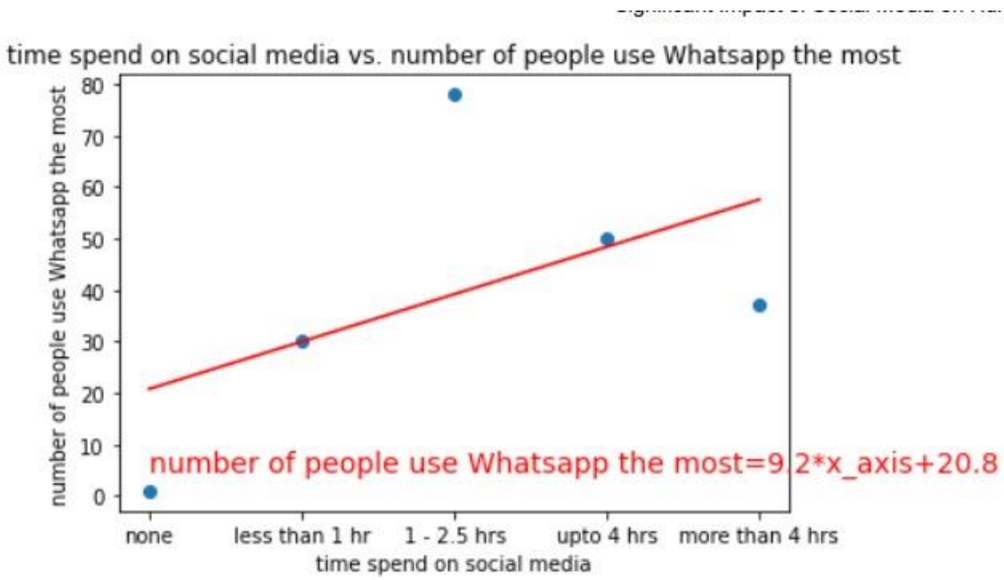


Figure 6.3 : Number of people using Whatsapp

This linear regression model assumes a linear relation between time spent on social media (x) and people using Whatsapp (y). The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). The linear equation is 9.2*x_axis+20.8.

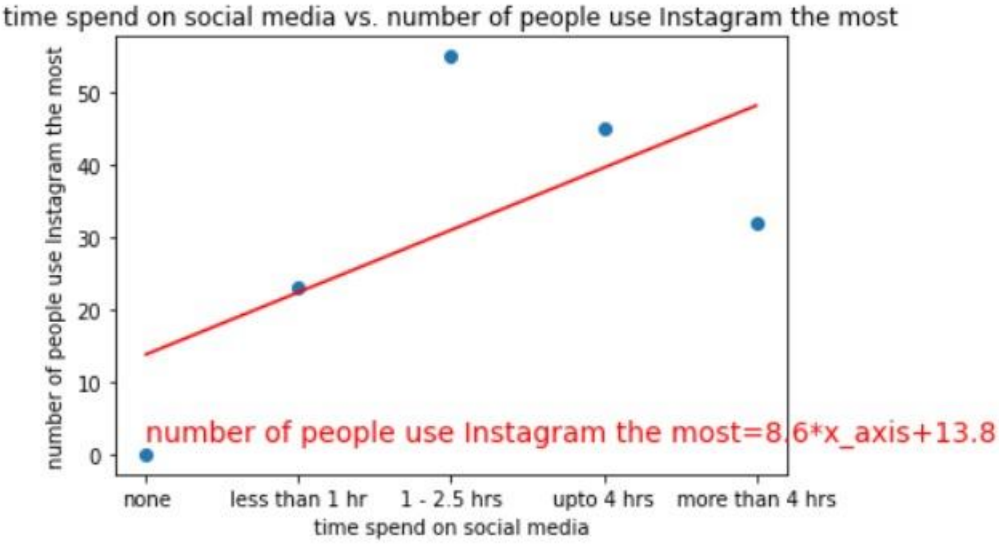


Figure 6.4 : Number of people using Instagram

This linear regression model assumes a linear relation between time spent on social media (x) and people using Instagram (y). The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). The linear equation is 8.6*x_axis+13.8.

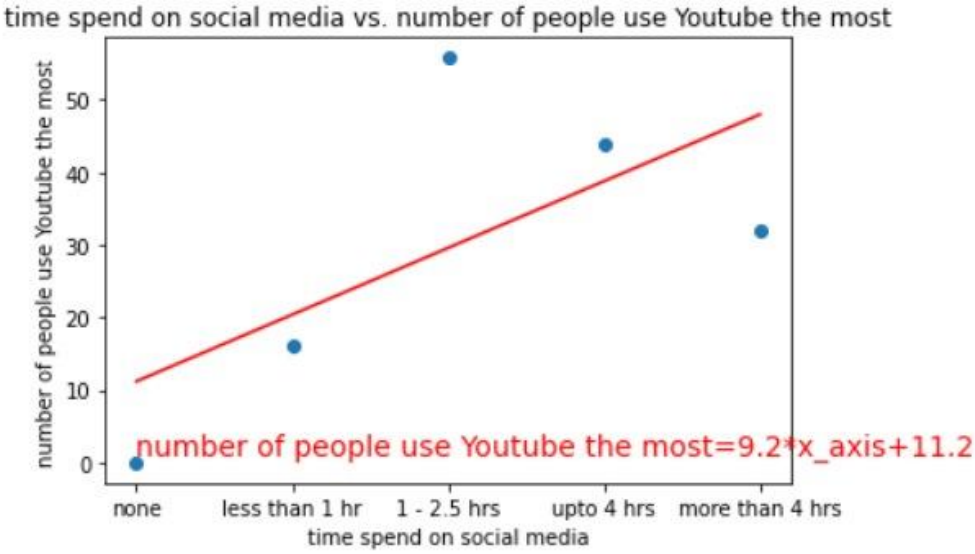


Figure 6.5 : Number of people using Youtube

This linear regression model assumes a linear relation between time spent on social media (x) and people using Youtube (y).The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). The linear equation is 9.2*x_axis+11.2.

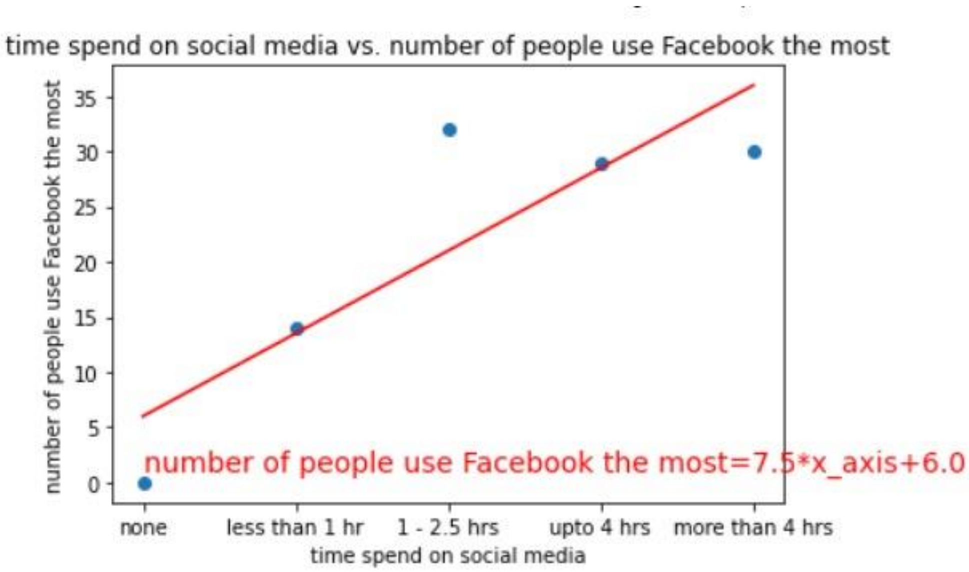


Figure 6.6 : Number of people using Facebook

This linear regression model assumes a linear relation between time spent on social media (x) and people using Facebook (y).The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). The linear equation is 7.5*x_axis+6.0.

6.1.3 Victims of Cybercrimes

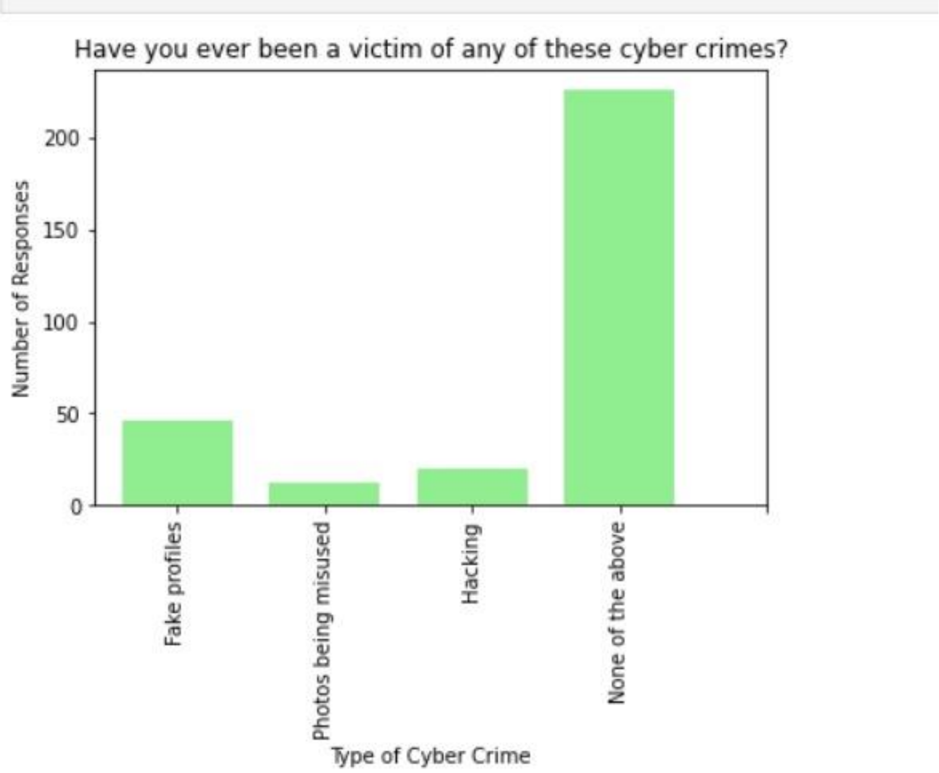


Figure 6.7 : People facing Cybercrimes

As we can see from the bar plot most of the people have not been affected from cybercrimes but the ones that have been victims faced the issue of Fake profiles.

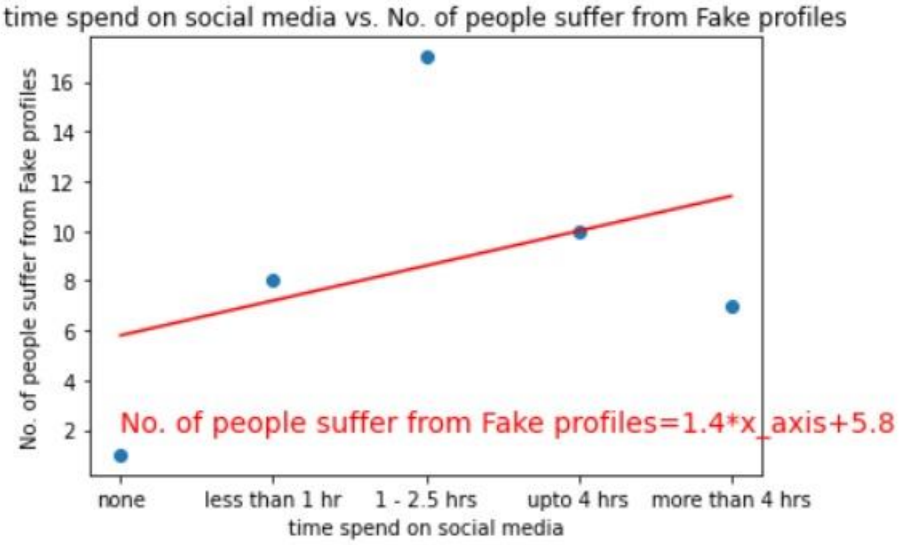


Figure 6.8 : Number of people suffering from fake profiles

This linear regression model assumes a linear relation between time spent on social media (x) and victims of fake profiles (y).The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). The linear equation is 1.4*x_axis+5.8.

6.1.4 Types of Communication generally preferred:

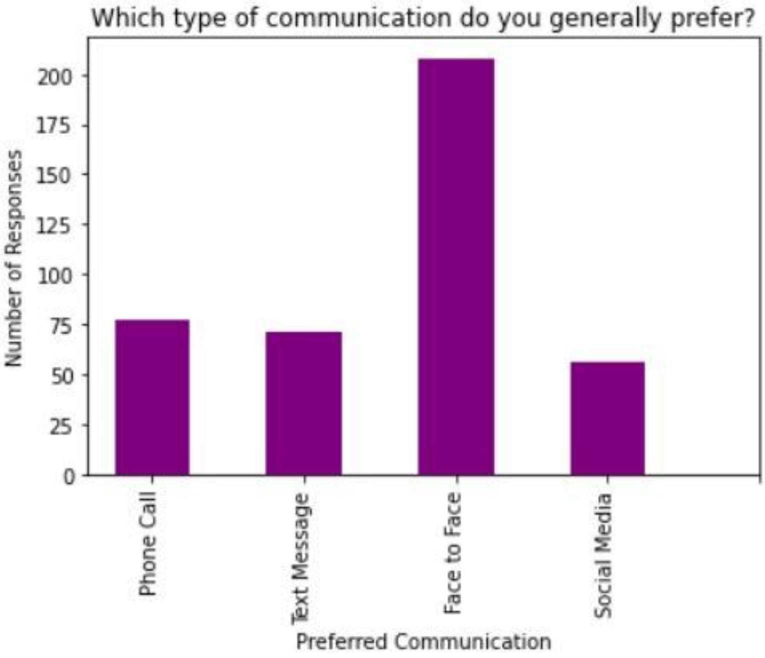


Figure 6.9 : Preferred Communication

From the bar plot we can observe that most people prefer face to face communication rather than phone calls, text messages or through social media sites.

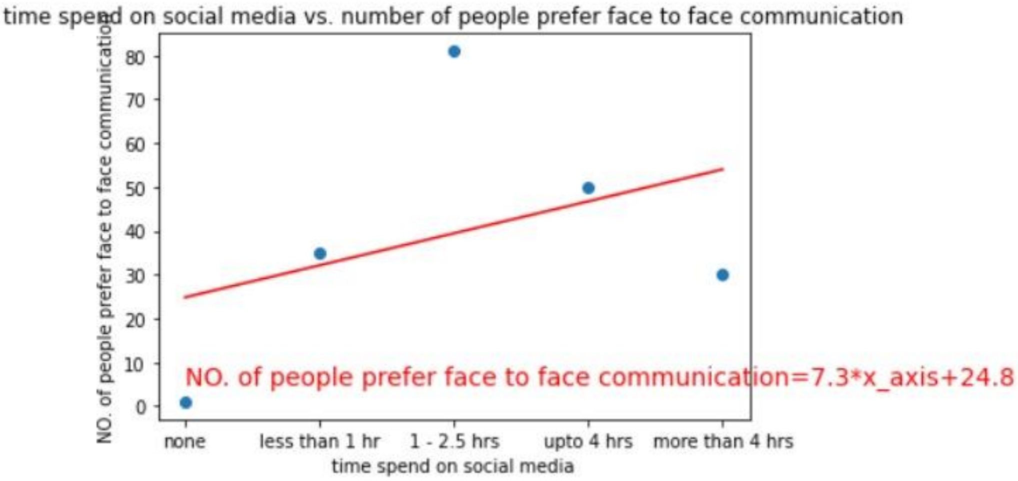


Figure 6.10: people preferring face to face communication

This linear regression model assumes a linear relation between time spent on social media (x) and number of people that prefer face to face communication (y).The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). The linear equation is 7.3*x_axis+24.8.

CHAPTER-7

RESULTS AND CONCLUSIONS

- People prefer to use social media sites like Whatsapp, Instagram more than any of the sites.
- People usually spend between 1-2.5 hours only on Whatsapp.
- 41% of the total respondents spent upto 2.5 hours daily on social media alone.
- Upto 42% of the respondents spent less than an hour on physical activities.
- Most of our respondents have not been affected by any type of Cybercrimes but those who have were victims of fake profiling.
- People that spent from 1 to 2.5 hours daily on social media sites are more likely to be subjected to Cybercrimes more than any other respondents.
- Large number of respondents prefer communication face to face rather that phone calls, text messages or through social media

CHAPTER-8

REFERENCES

[1] Effing, R., van Hillegersberg, J., Huibers, T. (2011). Social Media and Political Participation: Are Facebook, Twitter and YouTube Democratizing Our Political Systems? (https://link.springer.com/chapter/10.1007/978-3-642-23333-3_3)

[2] Ann DeSmet (October 2019), Social media and lifestyles in youth mental health promotion, The Lancet Child & Adolescent Health, Volume 3, Issue 10. (https://www.sciencedirect.com/science/article/abs/pii/S2352464219301865)

[3] Manikant Tripathi, Shiwangi Singh, Soni Ghimire, Seema Shukla, Shailendra Kumar (Feb 2018), Effect of Social Media on Human Health (https://www.researchgate.net/publication/323486379_Effect_of_Social_Media_on_Human_Health)

[4] Evgeny Vasilievich Popov,Using Oxana V. Komarova, V. L. Simonova (Feb 2021), Social Media and Messengers for Social Interactions (https://www.researchgate.net/publication/344955651_The_Influence_of_Social_Networks_on_Human_Society)

[5] Dian Anggraini Kusumajati, Rina Patriana Chairiyani, Nikodemus Thomas Martoredjo (January 2020), The Influence of Social Media on Human Behavior in Adolescents (Case Study of Bina Nusantara University Students), BDET 2020: Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology, Pages 102–105 (https://dl.acm.org/doi/10.1145/3378904.3378917)

[6] Rotimi Onadipe (May 2021), Impact of social media on cyber-crime in today’s digital age (https://www.thecable.ng/impact-of-social-media-on-cyber-crime-in-todays-digital-age)

[7] Jacob Amedie, Santa Clara University (March 2015), The Impact of Social Media on Society,(https://scholar.google.co.in/scholar?q=social+media+impact+on+teenage+behavior+a nd+society+research+paper&hl=en&as_sdt=0&as_vis=1&oi=scholart#d=gs_qabs&u=%23p%3DiHCu6xqL0AEJ)

[8] Sarah M. Coyne, Adam A.Rogers Jessica D.Zurcher, LauraStockdale, McCallBooth Brigham Young University, USA Received 3 July 2019, Revised 2 September 2019, Accepted 6 October 2019, Available online 10 October 2019, Version of Record 31 October 2019. Does time spent using social media impact mental health?: An eight year longitudinal study.

(https://scholar.google.co.in/scholar?q=social+media+impact+on+human+behavior+and+society+research+paper&hl=en&as_sdt=0&as_vis=1&oi=scholart#d=gs_qabs&u=%23p%3DKVNPNAcO9WAJ)

[9] Tracii Ryan, Kelly A. Allen, DeLeon L. Gray and Dennis M. McInerney (May 2017), Published online by Cambridge University Press, How Social Are Social Media? A Review of Online Social Behaviour and Connectedness
(https://www.cambridge.org/core/journals/journal-of-relationships-research/article/how-social-are-social-media-a-review-of-online-social-behaviour-andconnectedness/5F24EBEC0BC036A5B9AF8D4816F05E2E)

[10] Social Media Impact on human behavior and society. School paper published by school ICFAI university
(https://www.coursehero.com/file/77745683/07-Social-Media-Impact-on-human-behavior-and-societydocx/)

[11] Social media use in politics
(https://en.wikipedia.org/wiki/Social_media_use_in_politics)

Appendix: A-packages, tools used and working process

Python programming language:

Python programming language is a high level programming language and also is an interpreter which executes code line by line. Python has a dynamic type feature which makes programmers to write code easily and work without burden of initializing the variables. It supports multiple programming paradigms including object oriented, imperative, functional and procedural.

Python is available in 2 different versions python 2 and python3 current trending version is python 3. Python's source code is available under GNU-GPL. python uses a combination of reference counting and a cycle detecting garbage collector for memory management. It also features dynamic name resolution which binds method and variable names during execution of program. Also we can import multiple packages like matplotlib, NumPy etc. which makes our task easier. These modules has multiple methods which performs certain calculations and help us in fetching the results.

Libraries:

NumPy:

NumPy stands for numerical python is the package for scientific computing with Python. It is open-source so we can use it freely. It contains among other things like multi-dimensional array object, broadcasting functions, tools that help in integrating C/C++ and Fortran code, useful linear algebra, Fourier transform. It has in-built functions for higher mathematical calculations that are required for the execution of machine learning algorithms. NumPy targets the CPython implementation of Python, which is a non-optimizing bytecode interpreter. Mathematical algorithms written for this version of Python run slower than the others. NumPy addresses this problem with speed partly by providing multi-dimensional arrays and functions and operators that provide efficiently on arrays; using these requires reusing some code, mostly inner loops, using NumPy. Using NumPy in Python gives functionality comparable to MATLAB since they are both interpreted, and they both allow the user to write fast programs as long as it works on arrays.

Pandas:

Pandas is an open-source Python package that is mostly used for data analysis and machine learning tasks. This handles big amounts of data with the help of Data Frames. It has an inbuilt method for identifying different formats like CSV, XLSX, HTML. Different machine learning algorithms can be used for representing panda’s data structures and for data analysis. Pandas takes data from different file formats like comma-separated values, JSON, SQL, Microsoft

Excel. Pandas can perform various data manipulative operations like merging, reshaping, selecting, data cleaning, and data wrangling features.

Matplotlib:

Matplotlib is a library that is available in python programming language which is used for plotting the graphs. The plots helps us to understand the trends, patterns and to find the correlation between the attributes. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. It can customize visual style and layout. It can also exported into multiple file formats. Also it uses a rich array of third party packages built on matplotlib. Matplotlib makes easy matters clean and tough matters possible. It is a comprehensive library for creating static, animated and interactive visualizations in python. It also consists of several plots like line, bar, scatter, histogram etc. It can be installed in different operating platforms as well.

SciPy:

SciPy is a NumPy extension of python that contains a library of mathematical methods and convenience functions. It gives the user high-level commands and classes for manipulating and displaying data, giving the interactive Python session a lot more capability. An interactive Python session can be transformed into a data-processing and system-prototyping environment that rivals MATLAB, IDL, Octave, R-Lab, and SciLab. The fact that SciPy is based on Python also means that a powerful programming language is accessible for creating sophisticated programmes and customised applications. SciPy-based scientific applications benefit from the creation of new modules in a variety of software domains by developers all over the world. Subroutines and classes for anything from parallel programming to web and database subroutines and classes have been developed.

Tools Used:

Jupyter Notebooks:

Jupyter Project is a spin-off project from the I-Python project, which initially provided an interface only for the Python language and continues to make available the canonical Python kernel for Jupyter. The name Jupyter itself is derived from the combination of Julia, Python, and R.

A Jupyter Notebook is fundamentally a JSON file with a number of annotations. There are three main parts of the Notebook as follows.

- Metadata: a data dictionary of definitions used to set-up and display the notebook.
- Notebook format: version numbers of the software used to create the notebook. The version number is used for backward compatibility.

- List of cells: there are three different types of cells — markdown (display), code (to excite), and output

APPENDIX: B-SOURCE CODE

```
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as sts
import numpy as np
responses_df = pd.read_csv('responses.csv')
responses_df = responses_df[['What is your age?',
 'Which social media platform/s do you like the most or use the most?',
 'How much time do you spend on social media in a day?',
 'How much time do you spend on physical activities in a day?',
 'How much do you feel that you are exposed to inappropriate content on these platforms (out of 10)?',
 'Have you ever been a victim of any of these cyber crimes?',
 'Which type of communication do you generally prefer?']]
responses_df.head()
# Creating dataframe only for 18 to 21(young population)
response_1 = responses_df.loc[(responses_df["What is your age?"] >= 18) &
(responses_df["What is your age?"] <= 21)]
response_1.head(5)
social_media=["none","less than 1 hr","1 - 2.5 hrs","upto 4 hrs","more than 4 hrs"]
social_media_index=[]
physical=["none","less than 1 hr","1 - 2.5 hrs","upto 4 hrs","more than 4 hrs"]
physical_index=[]
for index , row in response_1.iterrows():
 if row['How much time do you spend on social media in a day?']=="none":
        social_media_index.append(0)
 elif row['How much time do you spend on social media in a day?']=="less than 1 hr":
        social_media_index.append(1)
 elif row['How much time do you spend on social media in a day?']=="1 - 2.5 hrs":
        social_media_index.append(2)
 elif row['How much time do you spend on social media in a day?']=="upto 4 hrs":
        social_media_index.append(3)
 else:
```

```
          social_media_index.append(4)
 if row['How much time do you spend on physical activities in a day?']=="none":
          physical_index.append(0)
 elif row['How much time do you spend on physical activities in a day?']=="less than 1 hr":
          physical_index.append(1)
 elif row['How much time do you spend on physical activities in a day?']=="1 - 2.5 hrs":
          physical_index.append(2)
 elif row['How much time do you spend on physical activities in a day?']=="upto 4 hrs":
          physical_index.append(3)
 else:
          physical_index.append(4)
plt.scatter(physical_index,social_media_index)
plt.xlabel("time spend on physical activities")
plt.ylabel("time spend on social media")
plt.title("time spend on physical activities vs. time spend on social media")
x_axis = [v for v in range(len(physical))]
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations,physical )
y_axis = [v for v in range(len(social_media))]
tick_locations = [value for value in y_axis]
plt.yticks(tick_locations,social_media )
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(physical_index,social_media_index)
linear_equation="social media time="+str(round(slope,2))+"*"+"physical activity
time"+"+"+str(round(intercept,2))
data=pd.DataFrame({"physical activity": physical_index,
 "social media activity":social_media_index})
reg_value=slope*data[["physical activity"]]+intercept
plt.plot(data[["physical activity"]],reg_value,"r-")
print(f"The r-squared is: {rvalue}")
plt.annotate(linear_equation,(0,0.5),color="red",fontsize=14)
plt.savefig('figures/1.png')
platform_choice = responses_df['Which social media platform/s do you like the most or use
the most?']
facebook,whatsapp,Instagram,twitter,youtube, snapchat = 0,0,0,0,0,0
```

```
other = 0
#We count the respective type in the dataframe ( platform_choice )
platform_choice_count = {"Facebook" : facebook, "Whatsapp" : whatsapp, "Instagram" :
instagram, "Twitter" : twitter,
 Youtube" : youtube, Snapchat" : snapchat,"Other" : other}
crime_df = responses_df['Have you ever been a victim of any of these cyber crimes?']
crime_df = crime_df.replace(to_replace ="Fake profiles, None of the above", value ="Fake
profiles")
fake, photo_misuse, hacking, none = 0,0,0,0
#We count the respective type in the dataframe ( crime_df )
preferred_com = responses_df['Which type of communication do you generally prefer?']
face, text, phone, social = 0,0,0,0
#We count the respective type in the data frame ( preferred_com )
preferred_com_count = {"Face to face" : face,
 "Text message" : text,
 "through phone" : phone,
 "through social media" : social}
responses_df1=responses_df[["Which social media platform/s do you like the most or use the
most?","How much time do you spend on s
responses_df1["Whatsapp"],responses_df1["Facebook"],responses_df1["Youtube"],responses
_df1["Instagram"],responses_df1["Twitter"],responses_df1["Snapchat"],responses_df1["Hike
"],responses_df1["Tinder"]="","","","","","","",""
#Place ‘1’ in the desired columns which is created above if the required value is found in that
row else place ‘0’
responses_df2=responses_df1[["How much time do you spend on social media in a
day?","Whatsapp","Facebook","Youtube","Instagram","T
responses_df2_g=responses_df2.groupby("How much time do you spend on social media in a
day?")
responses_df2_groupby=responses_df2_g.sum()
responses_11=response_1[["Which social media platform/s do you like the most or use the
most?","How much time do you spend on soci
responses_11["Whatsapp"],responses_11["Facebook"],responses_11["Youtube"],responses_1
1["Instagram"],responses_11["Twitter"],responses_11["Snapchat"],responses_11["Hike"],res
ponses_11["Tinder"]="","","","","","","",""
```

```
#Place ‘1’ in the desired columns which is created above if the required value is found in that row else place ‘0’
responses_2=responses_11[["How much time do you spend on social media in a day?","Whatsapp","Facebook","Youtube","Instagram","Twit
responses_2_g=responses_2.groupby("How much time do you spend on social media in a day?")
responses_2_groupby=responses_2_g.sum()
print(responses_2_groupby.index)
responses_2_groupby=responses_2_groupby.reindex(["none", "less than 1 hr", "1 - 2.5 hrs","upto 4 hrs","more than 4 hrs"])
responses_2_groupby[["Whatsapp"]].sum()[0]
#Plot time spend on social media vs number of people use Whatsapp the most
x_axis = np.arange(len(responses_df2_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df2_groupby[["Whatsapp"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Whatsapp the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Whatsapp"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in responses_2_groupby[["Whatsapp"]].to_numpy().tolist()]
linear_equation="number of people use Whatsapp the most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,5),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Facebook the most
x_axis = np.arange(len(responses_df2_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df2_groupby[["Facebook"]].to_numpy().tolist()])
```

```
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Facebook the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Facebook"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Facebook"]].to_numpy().tolist()]
linear_equation="number of people use Facebook the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Youtube the most
x_axis = np.arange(len(responses_df2_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df2_groupby[["Youtube"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Youtube the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Youtube"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Youtube"]].to_numpy().tolist()])
linear_equation="number of people use Youtube the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Instagram the most
x_axis = np.arange(len(responses_df2_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in
responses_df2_groupby[["Instagram"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Instagram the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Instagram"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Instagram"]].to_numpy().tolist()
linear_equation="number of people use Instagram the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,2),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Twitter the most
x_axis = np.arange(len(responses_df2_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in
responses_df2_groupby[["Twitter"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Twitter the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Twitter"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
```

```
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Twitter"]].to_numpy().tolist()])
linear_equation="number of people use Twitter the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Snapchat the most
x_axis = np.arange(len(responses_df2_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in
responses_df2_groupby[["Snapchat"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Snapchat the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Snapchat"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Snapchat"]].to_numpy().tolist()]
linear_equation="number of people use Snapchat the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Hike the most
x_axis = np.arange(len(responses_df2_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df2_groupby[["Hike"]].to_numpy().tolist()])
```

```
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Hike the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Hike"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Hike"]].to_numpy().tolist()])
linear_equation="number of people use Hike the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Create table counting mention of prefered communication type for 18 to 21(young
population)
responses_3=response_1[["Which type of communication do you generally prefer?","How
much time do you spend on social media in a
day
responses_3["through phone"],responses_3["Text message"],responses_3["face to
face"],responses_3["through social media"]="","","",""
#Place ‘1’ in the desired columns which is created above if required value is found in that row
else place ‘0’
#Rearrange index
responses_4=responses_3[["How much time do you spend on social media in a day?",
"through phone","Text message","face to face","through social media"]]
responses_4_g=responses_4.groupby("How much time do you spend on social media in a
day?")
responses_4_groupby=responses_4_g.sum()
print(responses_4_groupby.index)
```

```
responses_4_groupby=responses_4_groupby.reindex(["none", "less than 1 hr", "1 - 2.5
hrs","upto 4 hrs","more than 4 hrs"])
responses_4_groupby
#Plot time spend on social media vs number of people use Tinder the most with regression
x_axis = np.arange(len(responses_2_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_2_groupby[["Tinder"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_2_groupby[["Tinder"]].to_numpy().tolist()])
linear_equation="number of people use Tinder the
most="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,3),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs number of people use Tinder the most
x_axis = np.arange(len(responses_df2_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df2_groupby[["Tinder"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df2_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
responses_21_groupby=responses_2_groupby.div(responses_2_groupby.sum(axis=1),
axis=0)
#Compare NO. of people with different platform preference base on hours spend on social
media per platform_stack=responses_2_groupby.plot.bar(stacked=True,
figsize=(10,7),title="NO. of people with different social media plaforms
platform_stack.set_xlabel("time spend on social media") platform_stack.set_ylabel("NO. of
people based on social media plaforms preference") plt.savefig('figures/20.png')
responses_df3=responses_df[["Which type of communication do you generally
prefer?","How much time do you spend on social media in
responses_df3["through phone"],responses_df3["Text message"],responses_df3["face to
face"],responses_df3["through social media"]="","","",""
```

```
#Place ‘1’ in the desired columns which is created above if required value is found in that row
else place ‘0’
#Create table counting mention of prefered communication type for all data
responses_df4=responses_df3[["How much time do you spend on social media in a
day?","through phone","Text message","face to face",
responses_df4_g=responses_df4.groupby("How much time do you spend on social media in a
day?")
responses_df4_groupby=responses_df4_g.sum()
responses_df4_groupby=responses_df4_groupby.reindex(["none", "less than 1 hr", "1 - 2.5
hrs","upto 4 hrs","more than 4 hrs"])
responses_df4_groupby
#Plot time spend on social media vs No. of people prefer communication through phone
x_axis = np.arange(len(responses_4_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_4_groupby[["through phone"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people prefer communication through phone with
regression
x_axis = np.arange(len(responses_4_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_4_groupby[["through phone"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_4_groupby[["through phone"]].to_numpy().toli
linear_equation="NO. of people prefer phone
communication="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,2),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people prefer communication through text
message
```

```
x_axis = np.arange(len(responses_df4_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in responses_df4_groupby[["Text
message"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
plt.figure(figsize=(10,10))
#Plot time spend on social media vs No. of people prefer communication through text
message with regression
x_axis = np.arange(len(responses_4_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_4_groupby[["Text message"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_4_groupby[["Text message"]].to_numpy().tolis
linear_equation="NO. of people prefer Text message
communication="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people prefer face to face communication
x_axis = np.arange(len(responses_df4_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in responses_df4_groupby[["face to
face"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
plt.figure(figsize=(10,7))
#Plot time spend on social media vs No. of people prefer face to face communication with
regression
x_axis = np.arange(len(responses_4_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_4_groupby[["face to face"]].to_numpy().tolist()])
```

```
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_4_groupby[["face to face"]].to_numpy().tolis
linear_equation="NO. of people prefer face to face
communication="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,5),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people prefer communication through social
media
x_axis = np.arange(len(responses_df4_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in responses_df4_groupby[["through social
media"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
plt.figure(figsize=(10,7))
#Plot time spend on social media vs No. of people prefer through social media
communication with regression
x_axis = np.arange(len(responses_4_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_4_groupby[["through social
media"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_4_groupby[["through social media"]].to_numpy
linear_equation="NO. of people prefer through social media
communication="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercep
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(1,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_4_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
responses_41_groupby=responses_4_groupby.div(responses_4_groupby .sum(axis=1),
axis=0)
#Compare NO. of people perfer certain communication base on hours spend on social media
per day
platform_stack=responses_4_groupby.plot.bar(stacked=True, figsize=(10,7),title="time spend
on social media vs. NO. of people perfer certain communication")
platform_stack.set_xlabel("time spend on social media")
platform_stack.set_ylabel("NO. of people perfer certain communication")
plt.savefig('figures/30.png')
responses_df5=responses_df[["Have you ever been a victim of any of these cyber
crimes?","How much time do you spend on social medi
responses_df5["Fake profiles"],responses_df5["Photos being
misused"],responses_df5["Hacking"],responses_df5["None of the above"]="","","",""
#Place ‘1’ in the desired columns which is created above if required value is found in that row
else place ‘0’
responses_df6=responses_df5[["How much time do you spend on social media in a
day?","Fake profiles","Photos being misused","Hackin
responses_df6_g=responses_df6.groupby("How much time do you spend on social media in a
day?")
responses_df6_groupby=responses_df6_g.sum()
responses_df6_groupby=responses_df6_groupby.reindex(["none", "less than 1 hr", "1 - 2.5
hrs","upto 4 hrs","more than 4 hrs"])
responses_5=response_1[["Have you ever been a victim of any of
these cyber crimes?","How much time do you spend on social media in
responses_5["Fake profiles"]=""
responses_5["Photos being misused"]=""
responses_5["Hacking"]=""
responses_5["None of the above"]=""
#Place ‘1’ in the desired columns which is created above if required value is found in that row
else place ‘0’
responses_6=responses_5[["How much time do you spend on social media in a day?","Fake
profiles","Photos being misused","Hacking","
```

```
responses_6_g=responses_6.groupby("How much time do you spend
on social media in a day?")
responses_6_groupby=responses_6_g.sum()
responses_6_groupby=responses_6_groupby.reindex(["none", "less than 1 hr", "1 - 2.5
hrs","upto 4 hrs","more than 4 hrs"])
#Plot time spend on social media vs No. of people suffer from Fake profiles
x_axis = np.arange(len(responses_df6_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df6_groupby[["Fake profiles"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people suffer from Fake profiles with regression
x_axis = np.arange(len(responses_6_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_6_groupby[["Fake profiles"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_6_groupby[["Fake profiles"]].to_numpy().toli
linear_equation="No. of people suffer from Fake
profiles="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,2),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people suffer from Photos being misused
x_axis = np.arange(len(responses_df6_groupby.index))
plt.bar(x_axis,[v[0] for v in responses_df6_groupby[["Photos being
misused"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
plt.figure(figsize=(10,7))
```

```
#Plot time spend on social media vs No. of people suffer from Photos being misused with
regression
x_axis = np.arange(len(responses_6_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_6_groupby[["Photos being
misused"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_6_groupby[["Photos being misused"]].to_numpy
linear_equation="No. of people suffer from Photos being
misused="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,0),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people suffer from Hacking
x_axis = np.arange(len(responses_df6_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in
responses_df6_groupby[["Hacking"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people suffer from Hacking with regression
x_axis = np.arange(len(responses_6_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_6_groupby[["Hacking"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_6_groupby[["Hacking"]].to_numpy().tolist()])
linear_equation="No. of people suffer from Hacking="+
str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,0),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Plot time spend on social media vs No. of people suffer from other cyber crimes
x_axis = np.arange(len(responses_df6_groupby.index))
Youtube_plot=plt.bar(x_axis,[v[0] for v in responses_df6_groupby[["None of the
above"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, [j for j in responses_df6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
# Plot time spend on social media vs No. of people suffer from other cyber crimes with
regression
x_axis = np.arange(len(responses_6_groupby.index))
plt.scatter(x_axis,[v[0] for v in responses_6_groupby[["None of the
above"]].to_numpy().tolist()])
tick_locations = [value for value in x_axis]
#Plot regression line
(slope,intercept,rvalue,pvalue,stderr)=sts.linregress(x_axis,[v[0] for v in
responses_6_groupby[["None of the above"]].to_numpy().
linear_equation="No. of people suffer from other cyber
crimes="+str(round(slope,2))+"*"+"x_axis"+"+"+str(round(intercept,2))
reg_value=slope*x_axis+intercept
plt.plot(x_axis,reg_value,"r-")
plt.annotate(linear_equation,(0,1),color="red",fontsize=14)
plt.xticks(tick_locations, [j for j in responses_6_groupby.index])
#Display x-label, y-label, title for above displayed graph and save Img
#Share of type of cyber crime faced base on people's usual hour spend on social media per
day
responses_61_groupby= responses_6_groupby.div(responses_6_groupby.sum(axis=1),
axis=0)
def my_autopct(pct):
 return ('%1.1f%%'% pct) if pct > 0 else ''
fig, axs = plt.subplots(nrows=responses_61_groupby.index.size, ncols=1, figsize=(50,50))
fig.subplots_adjust(hspace=0.5, wspace=0.05)
```

```
i=0
for row in range(responses_61_groupby.index.size ):
 count_list=[]
 name_list=[]
[count_list.append(responses_61_groupby.loc[responses_61_groupby.index[row],:][i]) for i in
range(len(responses_61_groupby.loc
[name_list.append(responses_61_groupby.loc[responses_61_groupby.index[row],:].index[i])
for i in range(len(responses_61_groupb
 fig.add_subplot(axs[row] )
 plt.pie(count_list, labels=name_list,autopct=my_autopct,
explode=[0.2]+[0.1]*(len(name_list)-1),shadow=True, startangle=90)
 plt.axis('off')
 plt.title(responses_61_groupby.index[i])
 i=i+1
plt.savefig('figures/39.png')
#bar chart of respondents exposure to crime, what type of crime
crime_type=['Fake profiles','Photos being misused','Hacking','None of the above']
numb_resps=[responses_df5['Fake profiles'].sum(),responses_df5['Photos being
misused'].sum(),responses_df5['Hacking'].sum(),respon
plt.bar(crime_type,numb_resps, color="lightgreen", align="center", width = 0.75)
tick_locations = [value for value in x_axis]
plt.xticks(ticks=tick_locations, label=list(crime_type), rotation="vertical")
plt.show()
#Display x-label, y-label, title for above displayed graph and save Img
# Creating a pie chart considering respondent age (18 to 21)
age_details = ["Age-19", "Age-18", "Age-20", "Age-21"]
count = [136,84,37, 12]
colors = ["red", "darkorange", "indianred", "lightsalmon"]
explode = (0.1, 0, 0, 0)
b = sum(count)
percent = [100*y/b for y in count]
labels = ['{0} - {1:1.2f} %'.format(i,j) for i,j in zip(age_details, percent)]
plt.title("Filtered Respondent Age")
plt.pie(count, explode=explode, colors=colors, shadow=True, startangle=300)
plt.axis("equal")
plt.legend(labels=labels, loc="center left", bbox_to_anchor=(0.1, 1.))
plt.savefig('figures/44.png')
plt.figure(figsize=(60,100))
plt.show()
# Creating a pie chart considering respondent age (18 to 21) for time spent online on social
media
time_spend_online_1 = response_1['How much time do you spend on social media in a day?']
time_details_1 = ["Less than 1 hr", "1 to 2.5 hrs", "Upto 4 hrs", "More than 4 hrs", "None"]
count = [43,110,65,50,1]
colors = ["red", "salmon", "darkorange", "indianred", "lightsalmon"]
explode = (0, 0.1, 0, 0,0)
plt.title("Time spend on social media")
plt.pie(count, explode=explode, labels=time_details_1, colors=colors,
 autopct="%1.0f%%", shadow=True, startangle=100)
plt.axis("equal")
plt.legend(loc="center left", bbox_to_anchor=(0.1, 1.))
plt.savefig('figures/45.png')
plt.figure(figsize=(40,40))
plt.show()
# Creating a pie chart for time spent offline on physical activity
time_spend_online_1 = response_1['How much time do you spend on physical activities in a
day?']
time_details_1 = ["Less than 1 hr", "1 to 2.5 hrs", "Upto 4 hrs", "More than 4 hrs", "None"]
count = [112,95,15,5,42]
colors = ["red", "salmon", "darkorange", "indianred", "lightsalmon"]
explode = (0.1, 0.1, 0, 0,0)
plt.title("Time spend on physical activities")
plt.pie(count, explode=explode, labels=time_details_1, colors=colors,
 autopct="%1.0f%%", shadow=True, startangle=100)
plt.axis("equal")
plt.legend(loc="center left", bbox_to_anchor=(0.1, 1.))
plt.savefig('figures/46.png')
plt.figure(figsize=(40,40)), plt.show()
```