Guru Gobind Singh Indraprastha University
University School of Information, Communication and Technology

MINOR PROJECT

Understanding LinkedIn User Trends (Using Data Analytics)

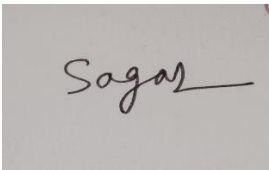Under the supervision of **Dr. Jaspreeti Singh**

(Associate Professor)

SUBMITTED BY

Sagar Haldar	*Prerit*
B.Tech CSE 7th Semester	B.Tech CSE 7th Semester
01616403221	02716403221

STUDENT UNDERTAKING

This is to undertake the LinkedIn User Behaviour Analysis work in this Report as part of the 7th Semester in B.Tech. (Computer Science and Engineering) during August- October, 2024 under the guidance of **Dr. Jaspreeti Singh** is our original work.

Anything that appears in this report which is not our original has been duly and appropriately acknowledged. Any academic misconduct and dishonesty found now or in future regarding above or any other matter about this report shall be solely and entirely the whole team's responsibility. In such a situation, we understand that strict disciplinary action can be undertaken against us by the concerned authorities of the University now or in future and we shall abide by it.

Sagar	*Prerit*
Sagar Haldar *01616403221*	*Prerit* *02716403221*

<u>*Certificate*</u>

This is to certify that the work titled **Understanding *LinkedIn User Trends using Data Analysis*** submitted by **Sagar Haldar and Prerit Masih**, in this project report as part of the 7th Semester in B.Tech. (Computer Science and Engineering) during August – October 2024 was done under my guidance and supervision.

Dr. Jaspreeti Singh
(Associate Professor)

<u>ACKNOWLEDGEMENT</u>

We take this opportunity to thank all the people contributing to the completion of the project.

We would like to sincerely thank *Dr. Jaspreeti Singh (Associate Professor), Department of USICT*. She has helped throughout the project by stimulating discussions and analysing issues. Our weekly discussions with her were quite educational. We would like to extend our profound gratitude for their support, steadfast advice, and belief in us. We owe her a great deal for her insightful advice throughout our endeavour.

We acknowledge our indebtedness to various project coordinators whose valuable supervision throughout the project helped in its completion.

We feel obliged to our friends and family for their inspiration and support in completing our project.

Name: Sagar Haldar
Enrollment No: 01616403221

Name: Prerit Masih
Enrollment No: 02716403221

Understanding LinkedIn User Trends (Using Data Analysis)

Abstract

We are doing data analysis on the LinkedIn database. Data analysis is the process of evaluating data using analytical or statistical tools to discover useful information. There are various tools available for performing data analysis like R, Python, Data Studio, and Microsoft Excel. After data has been gathered and organized using these techniques, judgments are made based on the findings. The final results can be presented in the form of a summary or a visual, such as a graph or chart. Data visualization is the process of displaying data visually. Tools for data visualizations simplify the task. You may get various graphics that can make data come to life using tools like Tableau or Microsoft Power BI. The factors based on which we are evaluating are Company Name, Age, Gender, Ethnicity, and number of followers. In this way, we can find out how these factors affect a person's perspective and chances of getting a job in a good company. There is not much research done on this topic so it is difficult to know other people's perspective towards the topic.

CONTENT

P

1. <u>INTRODUCTION</u>

Numerous social networking websites with both professional and non-professional orientations have been introduced over the past few years. It offered a brand-new, unanticipated method of acquiring valuable, precise, and regularly updated data on customer preferences and interests. One such widely used professional social networking application is LinkedIn. It is the biggest professional social network online. It may help you locate the appropriate job or internship, build professional connections, and acquire the knowledge you need to thrive in your career. Using your experience, abilities, and education, a thorough LinkedIn profile can help you connect with prospects.

It features more than 1 million company sites and more than 50 million professional profiles. Members update their profiles every week with more than 50,000 talents and 5,000 active jobs to highlight their qualifications to potential customers, co-workers, or employers. Significant growth has been seen in both revenue and net profit.

The goal of this project is to identify the behaviour of the users on LinkedIn based on criteria such as gender, age, ethnicity, followers, and companies. We concentrate on analysing the employment opportunities the platform offers based on these criteria using this identification. Data for the same is gathered from Kaggle. We use the collected data as the key data for our research.

1.1 <u>Problem Statement & Objectives</u>

<u>Problem Statement</u>

The project aims to gather the LinkedIn data of users to ascertain information about the activities that the user performed, how many followers they have, how many people work for a well-known company, what age group of people are most active on LinkedIn, how gender- balanced LinkedIn is in terms of providing job opportunities, and how impartial LinkedIn is in terms of ethnicity. The acquired dataset would next be examined to discover trends and connections between these qualities.

Objective

1. How many LinkedIn members work for a reputable company? How many people are a part of respectable organizations?

2. Age group of LinkedIn users who are most active on the site or who use LinkedIn most frequently

3. We can also make an assumption about a person's level of experience based on the aforementioned goal.

4. What is the predominant ethnicity of employees in various companies? What type of ethnicity is most prominently found among people working in big companies

5. Are there equal prospects for men and women? How gender-neutral is LinkedIn when offering career opportunities?

6. Based on the number of followers, identifying a fake profile

1.2 **Motivation**

The professional culture of the platform may be better understood through analysis of various LinkedIn metrics.

If one wants to connect with people in the machine learning community, for instance, does the majority of the network's members work in a field that is related to machine learning. As a LinkedIn user, one might not be aware of the network's many divisions.

Utilizing the data that is currently accessible, LinkedIn profile might be optimized or job opportunities by evaluating it.

The analysis is carried out using data visualization which could be expanded upon to identify the scams and fraud profiles on the platform.

1.3 Scope and Limitation

Scope

The project can be expanded using software such as the Squid proxy server, Wireshark, OWASP Web scrap, and Burp proxy server {}. These tools can be used to analyze data in order to gain useful insight into a person's psychological process. For instance, a user's behaviour can be examined based on factors such as:

1. What proportion of time do users spend on a specific LinkedIn activity?
2. How many users exhibit comparable behaviour?
3. What is a certain user's typical behavioural pattern?
4. What would the transition graphs look like depending on a user's activities?
5. What time of day are LinkedIn members most active?
6. On LinkedIn, how much content does each user publish?

Using the right tools and technologies, such as machine learning, cybersecurity, or Artificial Intelligence we may use the analysis to identify fraudulent and scam posts on LinkedIn, which are highly common on the platform and using machine learning we can learn from data, identify patterns, and make decisions with minimal human intervention.

Limitation

The majority of the dataset's objects are too close to one another to be displayed separately on the screen. Loss of information: Reducing the number of visible data sets might be useful, but also causes information to be lost. There are restrictions on the use of the LinkedIn API, including a lifetime user cap of 100,000, a ban on content storing, and a ban on using it for academic research. You could also get the data from your network and work with that csv output if you had a big enough network. In essence, you need some computational expertise to gather and utilise the LinkedIn data, and even then, the kind of research you can conduct would be constrained.

The sample size for this study is one of its drawbacks. The sample size was constrained by time, which makes it difficult to obtain results that are statistically significant. Furthermore, the sample is skewed. A sample with an even distribution could not be obtained.

1.4 <u>Organization of Report</u>

The following subjects are covered in this study:

<u>Section 1</u>

A detailed introduction to our subject is covered in the first section.

The purpose of this project is to analyse user behaviour on LinkedIn using variables like gender, age, ethnicity, followers, and companies. Using this identification, we concentrate on analysing the employment prospects the platform provides based on these characteristics. The objectives that correspond to the problem description have been made clear. The motivation is further explained. The project's scope and restrictions are exclusively stated sequentially in this section.

<u>Section 2</u>

This section summarises the relevant research that has already been done on our subject. Every relevant piece of writing has been carefully examined, and similarities have been analysed. Regarding our study, all the various technologies used in these investigations and the variations in data collecting are described. Finally, the various strategies are contrasted, and research gaps are identified.

<u>Section 3</u>

This section provides a brief explanation of the project's methodology. Initiating with a quick explanation of what the area is about, the entire procedure is described in full.

This section includes a summary of the data need gathering, data collecting, data cleaning, data interpretation, data pre-processing, data analysis, and data visualization processes. Furthermore, the data has been thoroughly defined, including a breakdown of all its properties. Lastly, this specific section also includes an overview of the Python code used.

<u>Section 4</u>

This section briefly outlines the approaches employed as well as the results and observations that have been attained. All the visualizations of the obtained results are attached. The study can be read using these graphic representations. Conclusion may be drawn from these observations.

Section 5

This section details the project's conclusion and its anticipated scope.

We tried to outline how the analysis may be extended and used in future research in the same field, for example, to identify the often-created false postings on LinkedIn.

Section 6

In our last section, we review the report's originality by including the similarity report.

1.5 Dataset Description

The dataset was gathered from Kaggle's official website which has compiled data from LinkedIn profiles including information from almost 15000 profiles. To do the study, the dataset has total entries and includes variables such firm name, gender, age, ethnicity, and followers.

Each row includes information about:
- Gender
- Age
- Ethnicity
- Company Name
- Followers

LinkedIn is the social network that offers the most to business executives. LinkedIn is a potent platform, whether you are a recruiter seeking for top candidates, a content producer trying to create an article, or a marketer endeavoring to reach the most significant audience.

Discovering about the LinkedIn statistics might affect the efforts before starting any social media campaign.

This dataset has been used to address some factors like:

In terms of work prospects, how gender neutral is LinkedIn? How many people work for legitimate companies instead of scam ones? Do People with more followers are more likely to succeed? Does a person's ethnicity have an impact on their job chances on LinkedIn? Which age group on LinkedIn has the highest level of activity and career success?

Table 3.1.1 is an example describing the dataset through counts of some key entities involved inthe dataset:

Details	Count
Number of Entries	313,550
Number of columns	5
Number of Rows	62,710

Table 3.1.1 : Details of the dataset

Data attributes are another component of any dataset. Data attributes are described in Table 3.1.2.

Details	Count
No. of companies	16
Gender count considered	2(Male and Female)
Types of ethnicities considered	3(White, Asian, and Black)
Age (including repetition)	60
Followers(including repetition)	62,710

Table 3.1.2 : Details of Data Attributes

1.6 Data Pre-processing

It is necessary to arrange or process the acquired data before conducting an analysis. This involves organizing the data in the manner necessary for the pertinent Analysis Tools, which may be done using Excel. Data processing is the action being taken here. The structured and processed data might then be inaccurate, duplicate, or incomplete. It's time to tidy up the

information we've gathered because not all of it will be helpful.

Monitor for errors: Keep a log and look for patterns in the sources of the majority of errors.

Verify Accuracy: Investigate purchasing data solutions that let you instantly clean your data.

Inspect for Duplicate Data: Find them and get rid of them to speed up analysis.

Remove all Formatting: Standardize the appearance of your data by getting rid of all formatting elements.

There are several ways for data to be duplicated or incorrectly categorized when merging different data sources. Even though they may appear to be right, results and algorithms are unreliable if data is erroneous.
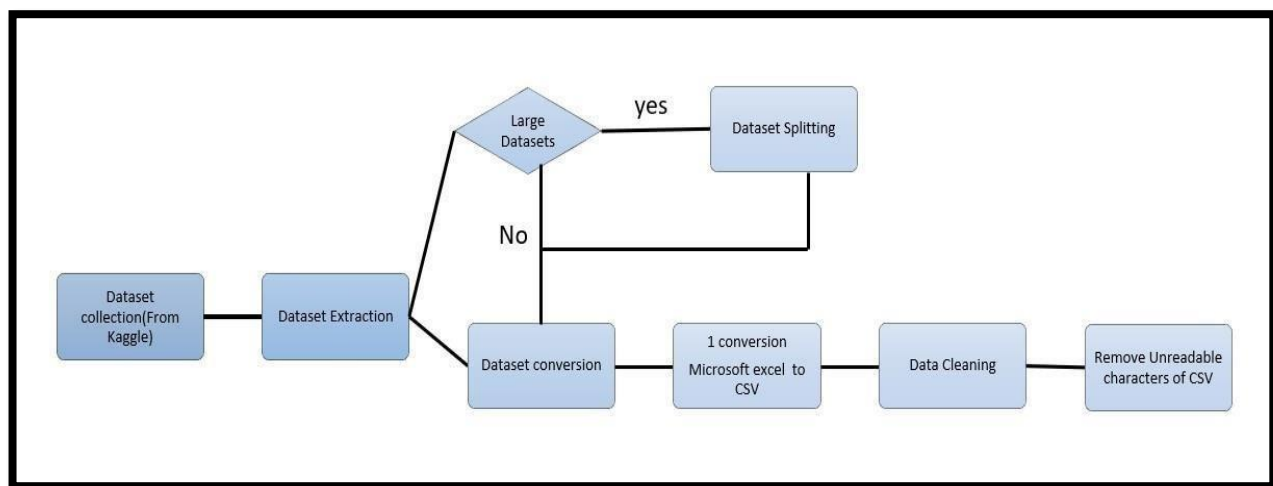
Fig 1.6.1 : Data Pre-Processing

2.Methodology

2.1 Description

Data analysis is the process of developing answers to questions through the examination and interpretation of data.

Understanding results from the surveys, administrative, and pilot study results, providing information on data gaps, designing and redesigning surveys, scheduling new statistical tasks, and developing quality targets all depend on data analysis.

We are conducting data analysis on Understanding LinkedIn user trends using data analysis, where we will examine various LinkedIn features that aid new graduates and regular people in finding better employment chances. Before applying for a job, an individual could perhaps consider several factors, such as how many LinkedIn users are hired by reputable businesses, how many skilled people work in the IT industry and business, and whether or not a person's gender or ethnicity has an impact on job opportunities. To answer these questions, data analysis is used for collecting, transforming, cleaning, and modelling data to discover the required information.

To start with this project, we should know about the process involved in data analysis. The process of data analysis consists of:

1. **Data Requirement Gathering**: This involves asking questions like why we are doing this analysis, what type of data we want to use, and what data we plan to analyze. Data may be numerical or categorical.

   To gain a better knowledge of how various characteristics affect career opportunities, we are conducting a LinkedIn analysis. We'll take into account things like the company name, target demographic, gender, race, and number of followers.

2. **Data Collection**: Guided by our identified requirements, it is time to collect the data from various sources. Case studies, surveys, questionnaires, interviews, focus groups, and direct observation are examples of sources. Data collection makes sure the information acquired is correct so that judgments made in connection with it are valid. Data Collection provides both a baseline to measure and a target to improve. The resulting data could be unstructured and could include irrelevant information. Hence, the collected data is required to be subjected to Data Processing and Data Cleaning.

From Kaggle, we obtained the user database for LinkedIn. Kaggle is an online community platform that allows users to find and publish data sets and explore and build models in a web-based data-science environment. Presently, the data that we have obtained from Kaggle is rather large and contains some inaccuracies. Better results can be obtained by cleaning the data. Thus, data cleaning would be the next step.

3. **Data Cleaning**: The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools and this can be done using Excel. This process is called data processing. Then the processed and organized data may be incomplete, contain duplicates, or contain errors. Not all the data that we have collected will be useful, so it is time to clean it up. Cleaning is the process of preventing and correcting these errors. This is the process where we remove white spaces, duplicate records, and basic errors. Data cleaning is mandatory before sending the information on for analysis.
The data we have gathered from Kaggle is enormous, has some inaccuracies, and contains certain attributes (or columns) that are not necessary for our project. So now we will execute data cleaning on the dataset using Excel, in which we will delete all the unnecessary columns and database entries. We shall then properly organize our data after that. After doing data processing and data cleaning on our dataset, then we will further move into our next step of data analysis.

4. **Data Analysis**: Here is where we use data analysis software and other tools to help interpret and understand the data and arrive at conclusions. Data analysis tools include Excel, Python, R, Looker, Rapid Miner, Chartio, Metabase, Redash, and Microsoft Power BI.
We are utilising Python, Excel, and Looker as our tools. Python is a powerful high-level programming language that supports both structured and functional programming methods. Its extensive collection of libraries makes it very useful in data analysis. Excel enables users to format, organize and calculate data in a spreadsheet. Looker helps us explore, share, and visualize a company's data so that we can make better business decisions.

5. **Data Interpretation**: Now that we have our results, we need to interpret them and come up with the best courses of action based on our findings.

6. **Data Visualization**: Data visualization is a fancy way of saying, "graphically show your information in a way that people can read and understand it." We have a variety of tools at our disposal, such as graphs, maps, charts, and bullet points. Visualization helps us derive valuable insights by helping us compare datasets and observe relationships. Sometimes data is presented using data visualization to make it simpler to spot patterns in the data that can be beneficial.

The dataset will be represented in the form of graphs and through this, we can observe various patterns in our dataset.

DATA ANALYSIS PROCESS:

Define the problem

↓

Collect data

↓

Clean the data

↓

Analyse the data

↓

Visualize and share your findings
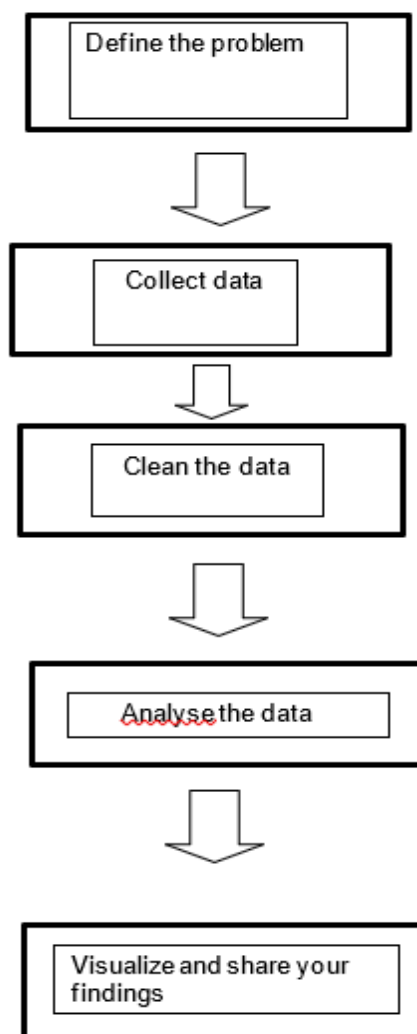
2.2 **Proposed Methodology**

We come up with an approach to analyze the data using visual techniques. To check assumptions with the help of graphical representations.

It involves generating a summary of the given dataset and creating various graphical representations to make the understating of data better

We have used python and its libraries for data visualization. To ensure the perfect representation is shown we have extended the use of looker studio for visual representation.

Pandas is the prime library along which we have used NumPy and matplotlib.

Numpy and pandas are used for cleaning and analysis while matplotlib is used for statistical analysis by data visualization.

We tend to analyze the LinkedIn user based on the company they are working for, their age, the number of followers they have, and their ethnicity.

In order to help a person who wants to enter the platform and wants to know what the platform contains.

For instance, if I wanted to create an account on LinkedIn, I would check at the ratio of male to female users to get a sense of the demographics. Is there a specific percentage of women on the site that would make me feel secure using it?

Coming to the company, if I am a job seeker I would look for a referral for my dream company, our analysis shows the percentage of people working in a particular company specifically the top 10 MNCs. If would provide me with a surety that I can go to LinkedIn and look for the person who works at that particular company for a referral.

Fig 2.2.1 : Process of Data Analysis
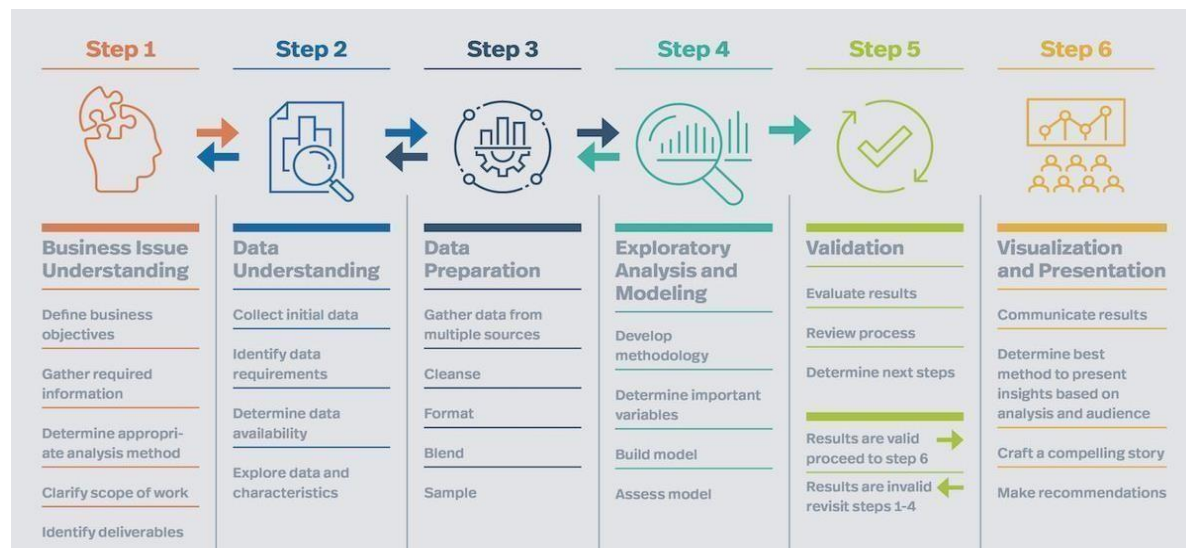
3. <u>Experiment Setup and Results</u>

<u>READING THE DATA</u>

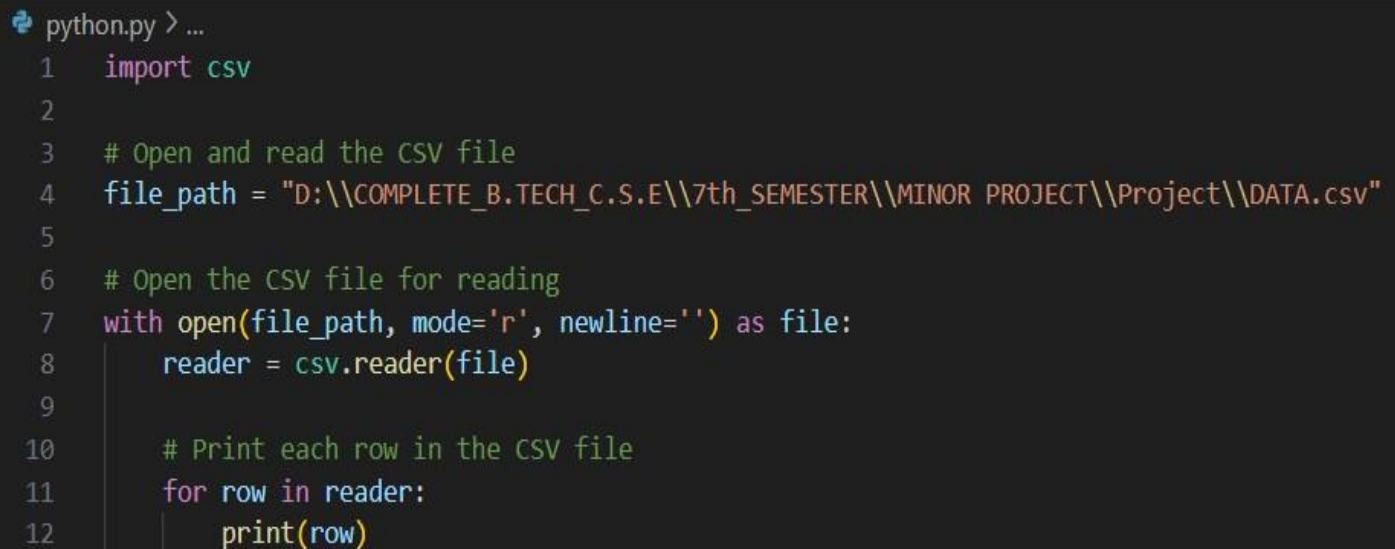**Below is the code snippet for the Experiment setup and Results :-**

```python
import csv

# Open and read the CSV file
file_path = "D:\\COMPLETE_B.TECH_C.S.E\\7th_SEMESTER\\MINOR PROJECT\\Project\\DATA.csv"

# Open the CSV file for reading
with open(file_path, mode='r', newline='') as file:
    reader = csv.reader(file)

    # Print each row in the CSV file
    for row in reader:
        print(row)
```

Below is the Screenshot of the code :-

```python
python.py > ...
1    import csv
2
3    # Open and read the CSV file
4    file_path = "D:\\COMPLETE_B.TECH_C.S.E\\7th_SEMESTER\\MINOR PROJECT\\Project\\DATA.csv"
5
6    # Open the CSV file for reading
7    with open(file_path, mode='r', newline='') as file:
8        reader = csv.reader(file)
9
10       # Print each row in the CSV file
11       for row in reader:
12           print(row)
```

➤ AGE

Below is the code snippet of the histogram for AGE :-

import pandas as pd
import matplotlib.pyplot as plt

 # Read in the Dataframe
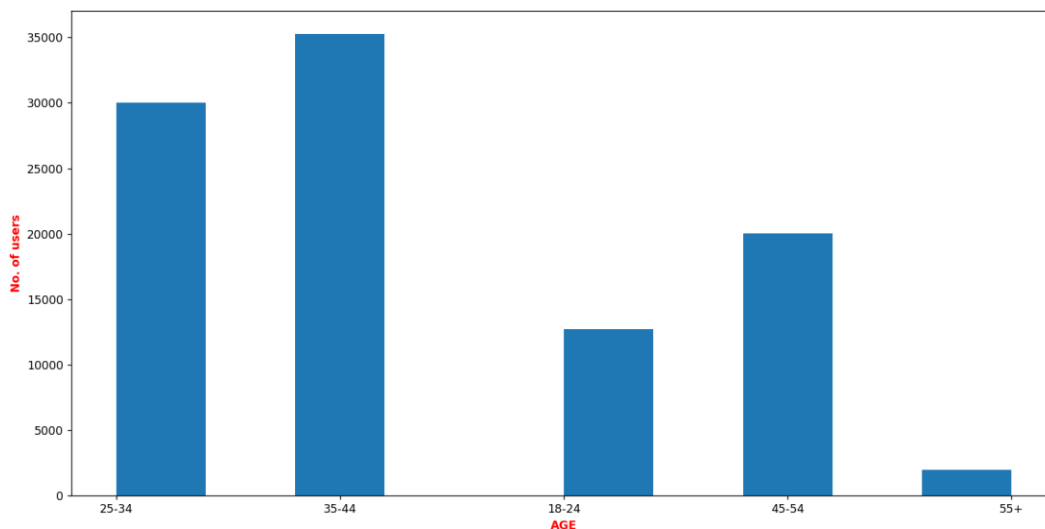**df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_DATA.csv")**

 # Creating a Histogram
plt.hist(df['Age Group'])
plt.xlabel('AGE', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
plt.ylabel('No. of users', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')

plt.show()

Below is the Screenshot of the code :-

```python
ageRange.py > ...
1    import pandas as pd
2    import matplotlib.pyplot as plt
3
4    # Read in the Dataframe
5    df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_DATA.csv")
6
7    # Creating a Histogram
8    plt.hist(df['Age Group'])
9    plt.xlabel('AGE', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
10   plt.ylabel('No. of users', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
11   plt.show()
```
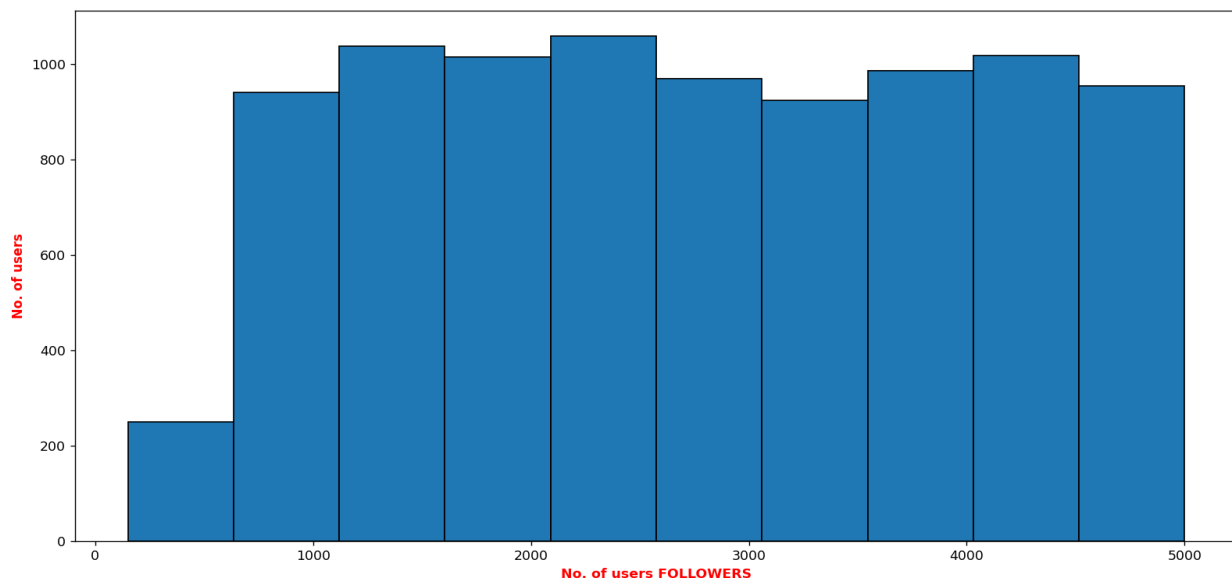
Result :-

➢ **FOLLOWERS**

Below is the code snippet for FOLLOWERS :-

import pandas as pd
import matplotlib.pyplot as plt
 # Read in the Dataframe
df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project
Updated_DATA.csv")
 # Creating a Histogram
plt.hist(df['Followers Count'], edgecolor='black', range=[150, 5000])
plt.xlabel('No. of users FOLLOWERS', fontweight='bold', color = 'red', fontsize='10',
horizontalalignment='center')
plt.ylabel('No. of users', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
plt.show()

Below is the screenshot of the code snippet :-

```
followership.py > ...
1    import pandas as pd
2    import matplotlib.pyplot as plt
3    |
4    # Read in the Dataframe
5    df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_DATA.csv")
6
7    # Creating a Histogram
8    plt.hist(df['Followers Count'], edgecolor='black', range=[150, 5000])
9
10   plt.xlabel('No. of users FOLLOWERS', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
11   plt.ylabel('No. of users', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
12
13   plt.show()
```
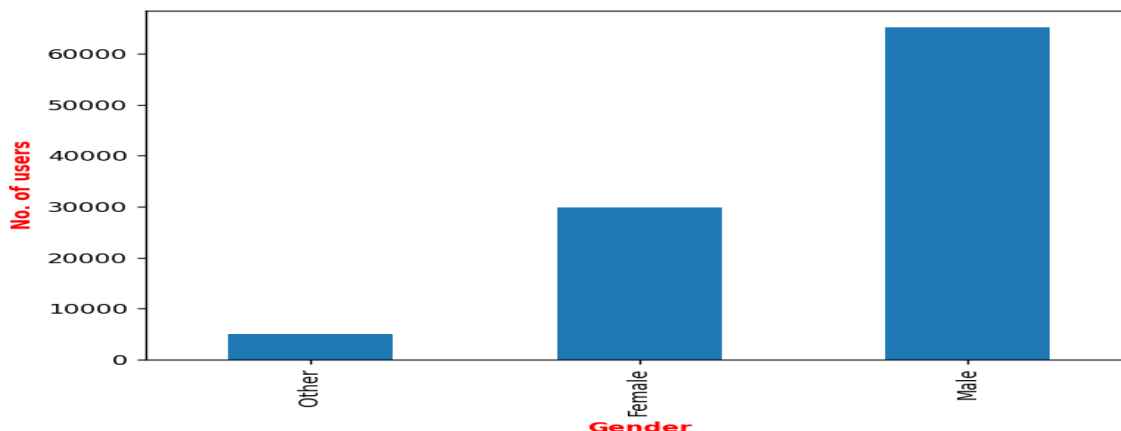
Result :

➢ GENDER

Below is the code snippet for GENDER :-

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_
DATA.csv")
df1 = df['Gender'].value_counts(ascending=True)
plt.xlabel('GENDER', fontweight='bold', color='red', fontsize=10,
horizontalalignment='center')
plt.ylabel('No. of users', fontweight='bold', color='red', fontsize=10,
horizontalalignment='center')
df1.plot.bar()
# Ensure layout fits well and show the plot
plt.tight_layout()
plt.show()
```

Below is the screenshot of the code snippet :-

```python
 gender.py > ...
 1    import pandas as pd
 2    import numpy as np
 3    import matplotlib.pyplot as plt
 4
 5    # Corrected file path with raw string notation
 6    df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_DATA.csv")
 7
 8    # Verify column name and correct if necessary
 9    df1 = df['Gender'].value_counts(ascending=True)
10
11    # Plot settings
12    plt.xlabel('GENDER', fontweight='bold', color='red', fontsize=10, horizontalalignment='center')
13    plt.ylabel('No. of users', fontweight='bold', color='red', fontsize=10, horizontalalignment='center')
14    df1.plot.bar()
15
16    # Ensure layout fits well and show the plot
17    plt.tight_layout()
18    plt.show()
```
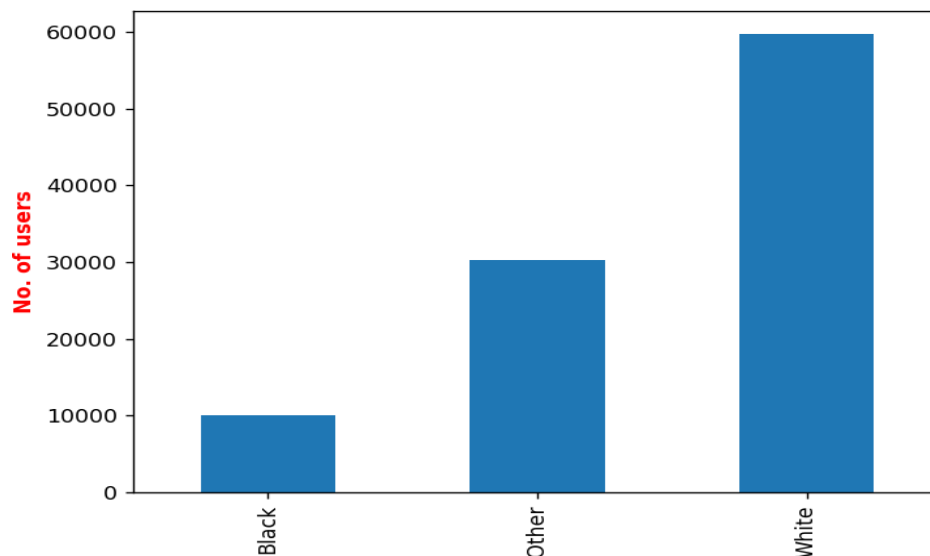
Result :

➢ **ETHNICITY**

Below is the code snippet for ETHNICITY :-

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated _DATA.csv")

df1 = df['Ethnicity'].value_counts(ascending=True)

plt.xlabel('ETHNICITY', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')

plt.ylabel('No. of users', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')

df1.plot.bar()

plt.show()

Below is the screenshot of the snippet:-

```
ethnicity.py > ...
1    import pandas as pd
2    import numpy as np
3    import matplotlib.pyplot as plt
4
5    df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_DATA.csv")
6    df1 = df['Ethnicity'].value_counts(ascending=True)
7
8    plt.xlabel('ETHNICITY', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
9    plt.ylabel('No. of users', fontweight='bold', color = 'red', fontsize='10', horizontalalignment='center')
10
11   df1.plot.bar()
12
13   plt.show()
```
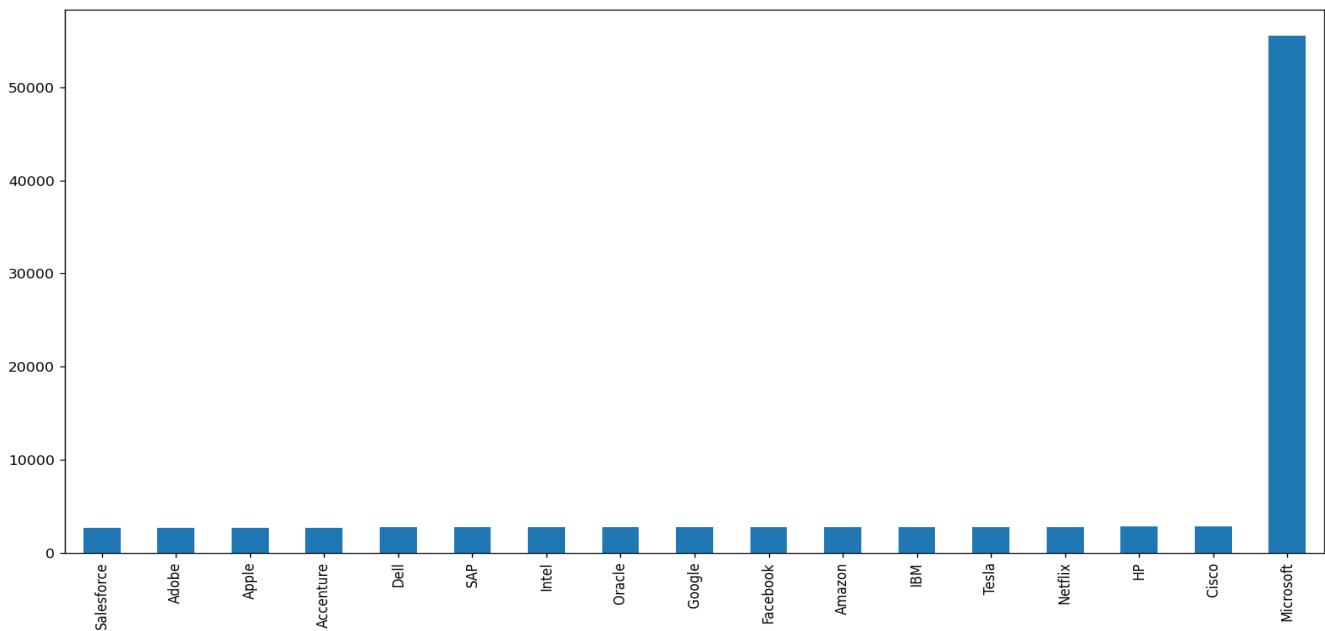
Result :-

➢ COMPANY

Below is the code snippet for COMPANY :-

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# df = pd.read_csv ("C://Users//Downloads//data.csv")
df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_
DATA.csv")
df1 = df['Company Name'].value_counts(ascending=True)
df1.plot.bar()
plt.show()
```

Below is the screenshot of the code snippet :-

```
company.py > ...
1    import pandas as pd
2    import numpy as np
3    import matplotlib.pyplot as plt
4    # df = pd.read_csv ("C://Users//Downloads//data.csv")
5    df = pd.read_csv(r"D:\COMPLETE_B.TECH_C.S.E\7th_SEMESTER\MINOR PROJECT\Project\Updated_DATA.csv")
6    df1 = df['Company Name'].value_counts(ascending=True)
7    df1.plot.bar()
8    plt.show()
```

Result :-

RESULT

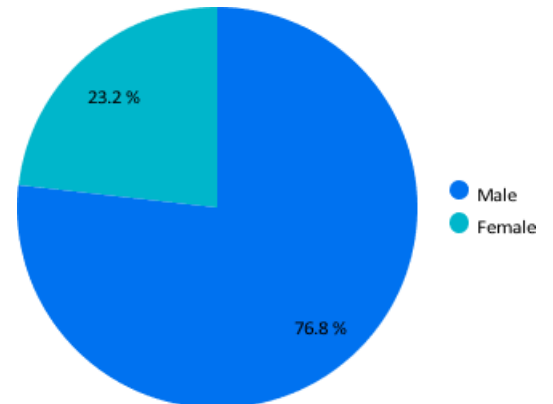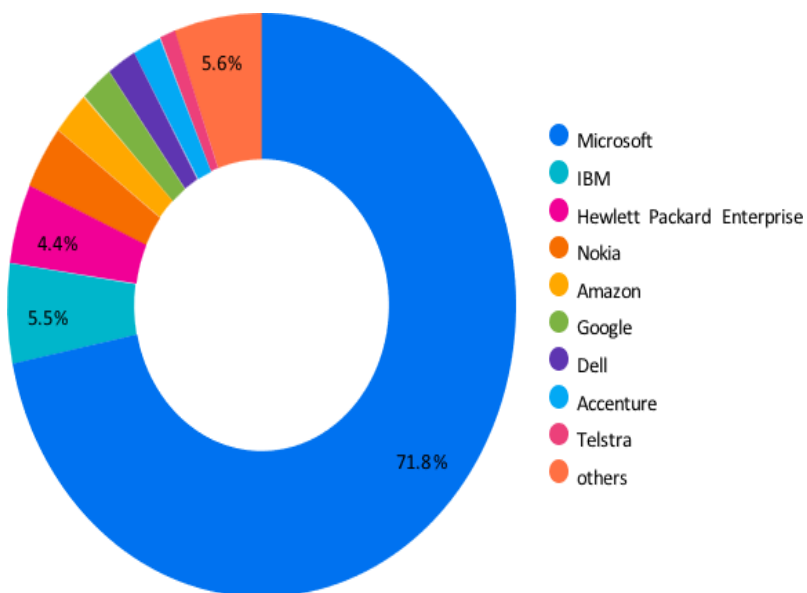The results corresponding to the analysis that was performed are as follows:

    A. Total percentage of female user on LinkedIn = 23.8%

    B. Total percentage of male user on LinkedIn = 76.2%

    A. al number of LinkedIn users who works in top 16 MNC's = 8440

         1. which the percentage of female users is = 23.2%

         2. which the percentage of male users is = 76.8%

    B. er of users with ethnicity (White) = 47,937

    C. er of users with ethnicity (Asian) = 9,878

    D. er of users with ethnicity (Black) = 4,894

The following is the visual representation of the data using Data Studio(Looker) :-
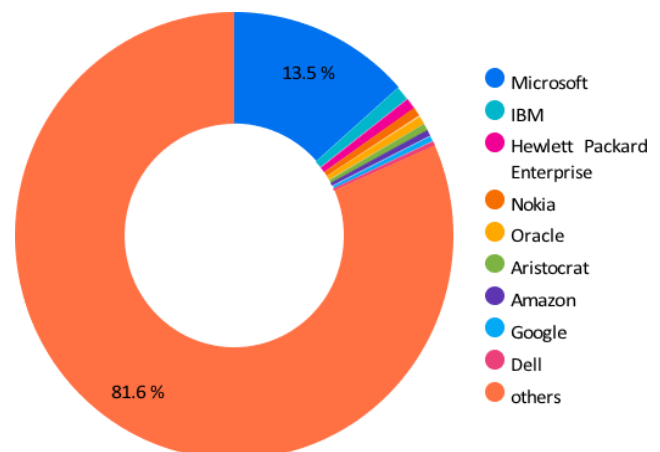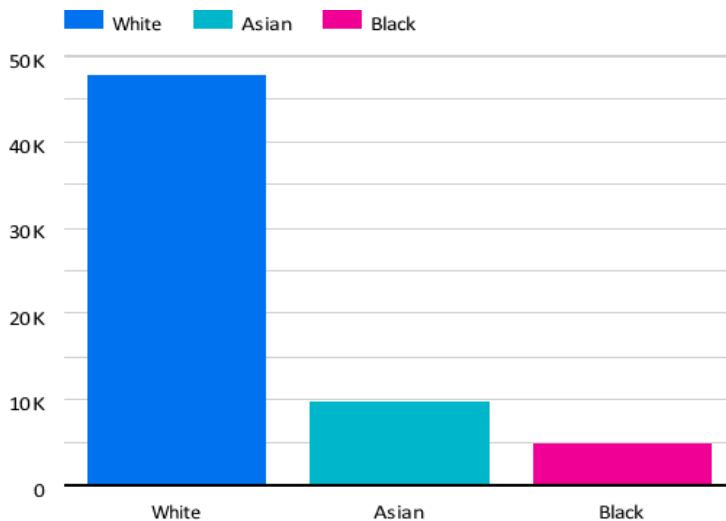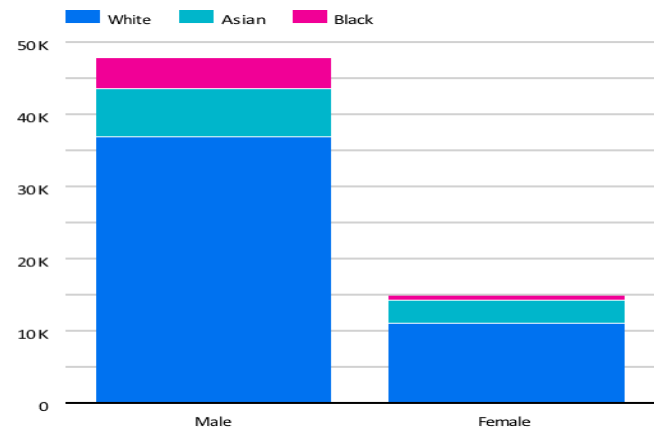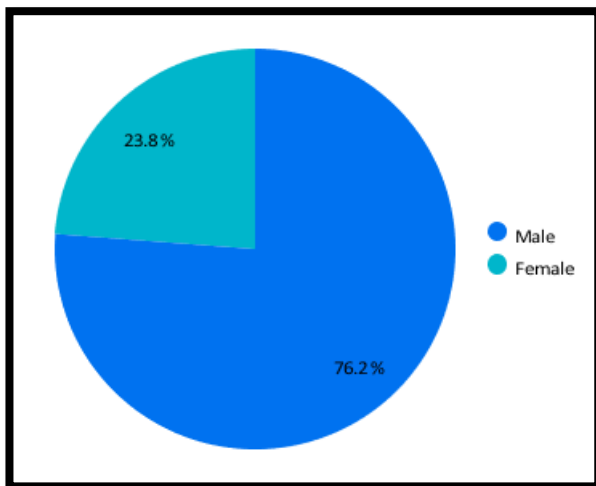
VISUAL REPRESENTATION OF TOP 16 MNC's

	c_name	Record Count
1.	Microsoft	8,440
2.	IBM	649
3.	Hewlett Packard Enterprise	521
4.	Nokia	417
5.	Amazon	284
6.	Google	242
7.	Dell	221
8.	Accenture	211
9.	Telstra	117
10.	Apple	111
11.	GE	108
12.	Deloitte	107
13.	Facebook	100
14.	Adobe	88
15.	Deutsche Bank	73
16.	Tata Consultancy Services	70

4. <u>Conclusion and Future Work</u>

We aimed to gather the LinkedIn data of users to ascertain information about the activities that the user performed, how many followers they have, how many people work for a well- known company, what age group of people are most active on LinkedIn, how gender-balanced LinkedIn is in terms of providing job opportunities, and how impartial LinkedIn is in terms of ethnicity. The acquired dataset would next be examined to discover trends and connections between these qualities. And hence,

The results corresponding to the analysis that was performed are as follows:

- Total percentage of female users on LinkedIn = 23.8%
- Total percentage of male users on LinkedIn = 76.2%
- Total number of LinkedIn users who work in top 16 MNCs = 8440
- In which the percentage of female users is = 23.2%
- In which the percentage of male users is = 76.8%
- Number of users with ethnicity (White) = 47,937
- Number of users with ethnicity (Asian) = 9,878
- Number of users with ethnicity (Black) = 4,89

<u>Future scope:</u>

The project can be expanded using software such as the Squid proxy server, Wireshark, OWASP Web scrap, and Burp proxy server. These tools can be used to analyze data in order to gain useful insight into a person's psychological process. For instance, a user's behaviour can be examined based on factors such as :

1. What is a certain user's typical behavioural pattern?
2. What would the transition graphs look like depending on a user's activities
3. On LinkedIn, how much content does each user publish?

Using the right tools and technologies, such as machine learning, cybersecurity, or Artificial Intelligence we may use the analysis to identify fraudulent and scam posts on LinkedIn, which are highly common on the platform and using machine learning we can learn from data, identify patterns, and make decisions with minimal human intervention.

5.<u>References</u>

[1] P. Jain, "Social Media: Impact on human behavior and society," LinkedIn, Nov. 20, 2018. [Online]. Available:https://www.linkedin.com/pulse/social-media-impact-human-behavior-society-piyushi-jain/. [Accessed: Nov. 26, 2024].

[2] O. Bondar, "Important LinkedIn Statistics Data & Trends," LinkedIn, Dec. 22, 2023. [Online].

Available: https://www.linkedin.com/pulse/important-linkedin-statistics-data-trends-oleksii-

bondar-pqlie/. [Accessed: Nov. 1, 2024].

[3] A. Nagpal and G. Gabrani, "Python for Data Analytics, Scientific and Technical

Applications," in *Proc. 2019 Amity Int. Conf. on Artificial Intelligence (AICAI)*, Dubai,

UAE, Feb. 2019, doi: 10.1109/AICAI.2019.8701341. [Online]. Available:

https://www.researchgate.net/publication/332760463_Python_for_Data_Analytics_Scientific_a

nd_Technical_Applications. [Accessed: Oct. 26, 2024].

[4] D. Dibekulu, "An Overview of Data Analysis and Interpretations in Research," *Int. J. Acad. Res. Educ. Rev.*, vol. 8, no. 1, pp. 1–10, Jan. 2020, doi: 10.14662/IJARER2020.015. [Online]. Available: https://www.researchgate.net/publication/365668845_An_Overview_of_Data_Analysis_and_Interpretations_in_Research. [Accessed: Oct. 26, 2024].

[5] S. Palmer, "Fake News: A Case Study," LinkedIn, Jan. 28, 2018. [Online]. https://www.linkedin.com/pulse/fake-news-case-study-shelly-palmer/. [Accessed: Nov. 5, 2024].

[6] P. J. Hickey, A. Erfani, and Q. Cui, "Use of LinkedIn Data and Machine Learning to Analyze Gender Differences in Construction Career Paths," *J. Manage. Eng.*, vol. 38, no. 6, Aug. 2022, doi: 10.1061/(ASCE)ME.1943-5479.0001087.[Online].Available:https://www.researchgate.net/publication/362449228_Use_of_LinkedIn_Data_and_Machine_Learning_to_Analyze_Gender_Differences_in_Construction_Career_Paths. [Accessed: Nov. 10, 2024].

[7] geetanjali16, "Power BI - Create Pie and Donut Charts," GeeksforGeeks, Jul. 31, 2024. [Online].https://www.geeksforgeeks.org/power-bi-create-pie-and-donut-charts/. [Accessed: Nov.26, 2024].