Data Analysis for Linkedin Data

GITHUB LINK

NAME: ARPIT PANWAR

ROLL NUMBER: 045013

SECTION: H

Submitted to - Dr. Amarnath Mitra

Project Objective: Designing a LinkedIn Web Scraper for Data Extraction and Analysis

Introduction:

In today's data-driven environment, significant insights can be gleaned from a diverse set of web sources. As one of the major professional networking sites, LinkedIn provides a plethora of information about individuals, businesses, job markets, and trends. The goal of this project is to create and construct a web scraper to extract data from LinkedIn profiles and then analyse this data thoroughly to generate relevant reports. This project aims to provide consumers with actionable information by leveraging Python and data analysis tools.

Project Objectives:

1. **Web Scraping**: The primary purpose of this project is to create a basic web scraping program capable of extracting data from LinkedIn profiles. Extracting information such as user profiles, connections, work experiences, and abilities falls under this category.

2. **Cleaning and Transformation of Data**: Once the data has been obtained, it will be thoroughly cleansed and transformed. Because web data can be untidy and inconsistent, this phase is critical to ensuring the dataset's accuracy and consistency.

3. **Data Storage**: For further examination, the extracted and processed data will be saved and organised in CSV files. To enable quick querying and reporting, proper data indexing and structuring will be prioritised

4. **Data Analysis**: The goal of this project is to analyse the collected LinkedIn data in order to get insights into many topics. Analysis may include: - determining how gender influences a person's career - determining the average age and gender ratio

- Recognising how frequently people change occupations
- What organisations do they work for, and how many followers do they have?

Expected Outcomes:

1. A fully functional basic LinkedIn web scraper capable of extracting diverse data from profiles.
2. A clean and organised dataset suitable for analysis.
3. Insights and reports that provide valuable information for various stakeholders, including job seekers, recruiters, and businesses.

1. **Positions_in_Previous_Tenure**: This column represents the number of positions or job roles the individual had in their previous tenures or employment history. It provides insight into the person's career trajectory and job mobility.

2. **Total_YOE (Total Years of Experience):** This column is a numeric value representing the total number of years of professional experience the individual has accumulated. It's a key indicator of their overall experience level.

3. **Org_Name (Organization Name)**: This column contains the names of the organizations where the individual has worked. It provides information about the companies the person has been associated with in their career.

4. **No_Of_Positions_Total**: This numeric column represents the total number of different job positions or roles the individual has held throughout their career. It's a measure of their job diversity.

5. **No_of_Orgs_Worked_For:** This numeric column represents the total number of different organizations or companies the individual has worked for during their career. It provides insight into the person's career mobility and adaptability.

6. **Current_Tenure_Length**: This column represents the length of time (in years or months) the individual has spent in their current job or position. It can help gauge their stability in their current role.

7. **Age**: This column contains the age of the individual. Age is a significant factor in understanding their career progression and potential.

8. **Gender**: This column indicates the gender of the individual. It's important to note that gender information should be handled with care and in compliance with privacy and ethical considerations.

9. **Followers**: This column represents the number of followers or connections the individual has on LinkedIn. It can be an indicator of their professional network and influence on the platform.

Summary:
The dataset includes various attributes related to individuals' professional backgrounds and LinkedIn profiles, such as career history, total years of experience, organizations worked for, current job tenure, age, gender, and LinkedIn network size (followers). Analyzing this data can provide valuable insights into the career paths and characteristics of the individuals in the dataset, which can be useful for various purposes, such as talent acquisition, career counseling, or workforce analytics.

```
print(df.head())
```

```
   Positions_in_Previous_Tenure  Total_YOE                    Org_Name  \
0                             2       3.67                          TD
1                             2       2.46  Light Up The World (LUTW)
2                             1       1.83                     Glacier
3                             1       1.54                  Sprout App
4                             1       1.30                College Pro

   No_Of_Positions_Total  No_of_Orgs_Worked_For  Current_Tenure_Length  Age  \
0                      1                      1                   1.25   37
1                      1                      2                   0.58   37
2                      1                      3                   0.67   37
3                      1                      4                   0.34   37
4                      1                      5                   0.67   37

  Gender  Followers
0   Male        420
1   Male        420
2   Male        420
3   Male        420
4   Male        420
```

We use df.head() to show the top 5 rows for a quick overview of the DataFrame's contents.

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62709 entries, 0 to 62708
Data columns (total 9 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Positions_in_Previous_Tenure  62709 non-null  int64
 1   Total_YOE                     62709 non-null  float64
 2   Org_Name                      62703 non-null  object
 3   No_Of_Positions_Total         62709 non-null  int64
 4   No_of_Orgs_Worked_For         62709 non-null  int64
 5   Current_Tenure_Length         62709 non-null  float64
 6   Age                           62709 non-null  int64
 7   Gender                        62709 non-null  object
 8   Followers                     62709 non-null  int64
dtypes: float64(2), int64(5), object(2)
memory usage: 4.3+ MB
None
```

df.info() is used to provide a concise summary of the DataFrame's metadata. It includes information such as the number of non-null values and column types

```
print(df.describe())
```

```
       Positions_in_Previous_Tenure    Total_YOE   No_Of_Positions_Total  \
count                  62709.000000  62709.000000           62709.000000
mean                       1.191567      3.015914               1.243506
std                        0.556407      2.700674               0.756841
min                        1.000000      0.000000               1.000000
25%                        1.000000      1.470000               1.000000
50%                        1.000000      2.420000               1.000000
75%                        1.000000      3.840000               1.000000
max                       15.000000    108.990000              12.000000

       No_of_Orgs_Worked_For  Current_Tenure_Length           Age  \
count           62709.000000           62709.000000  62709.000000
mean                3.498653               2.637674     44.048717
std                 3.112945               2.981699     10.683842
min                 1.000000              -0.330000      1.000000
25%                 2.000000               0.830000     37.000000
50%                 3.000000               1.750000     44.000000
75%                 4.000000               3.330000     51.000000
max                49.000000              59.960000     80.000000

           Followers
count   62709.000000
mean     1225.838173
std      6406.523908
min         0.000000
25%       405.000000
50%       725.000000
75%      1244.000000
max    530566.000000
```

This is a description of our data, with the count, mean, standard deviation, max, min, top 25, 50 and 75 percentiles values

Analysis & Findings

Age

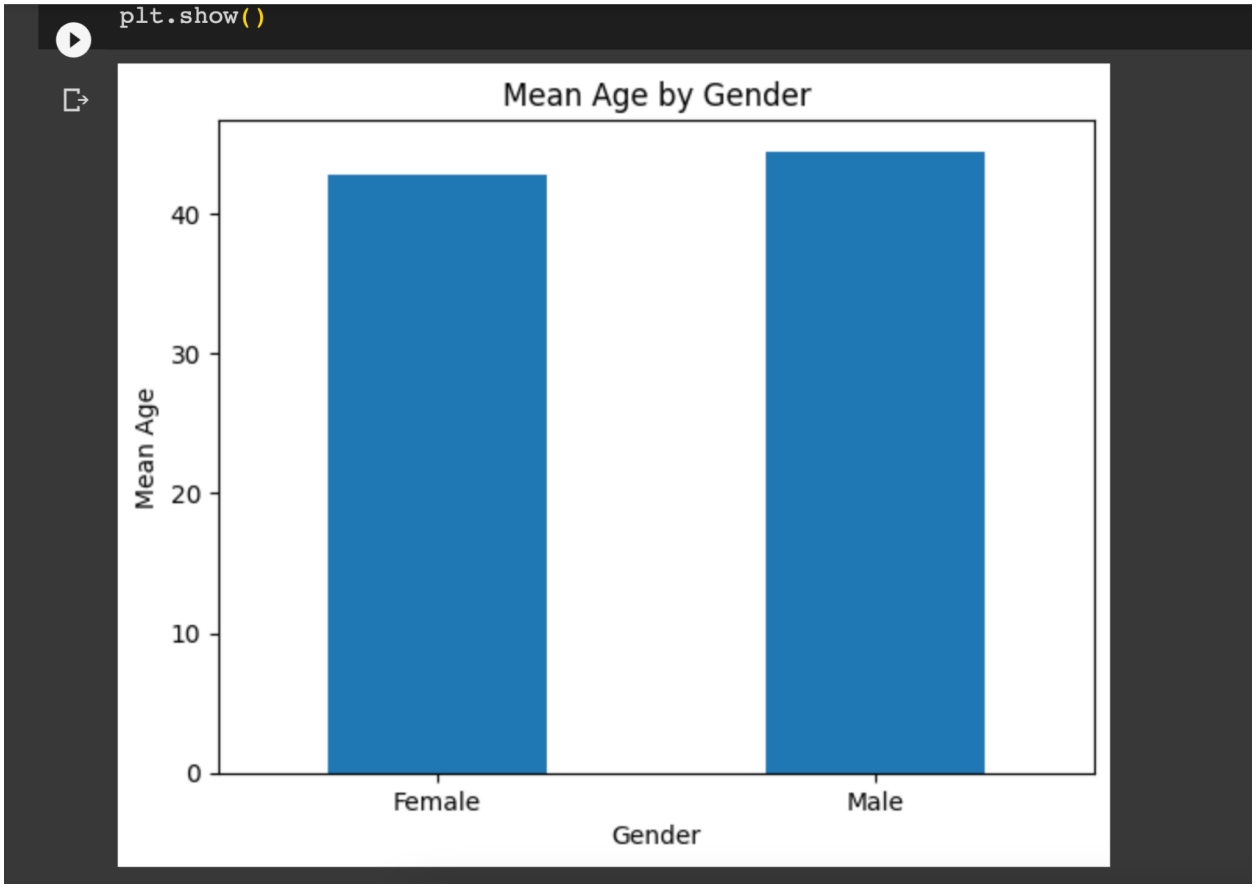
Mean Age

For Females: 42.79 years

For Males: 44.44 years

Insights

The dataset's age distribution displays some intriguing characteristics. The average age for both genders is in their early 40s, with females being slightly younger than males. This shows that the dataset contains a wide range of ages, ranging from 7 years old to 80 years old. The presence of younger people may reflect a growing interest in LinkedIn among professionals at the beginning of their careers. The presence of older people, on the other hand, illustrates the platform's importance to seasoned professionals, especially those approaching retirement age. Understanding the age distribution can aid in the development of marketing and recruitment strategies that cater to various age groups.



Followers

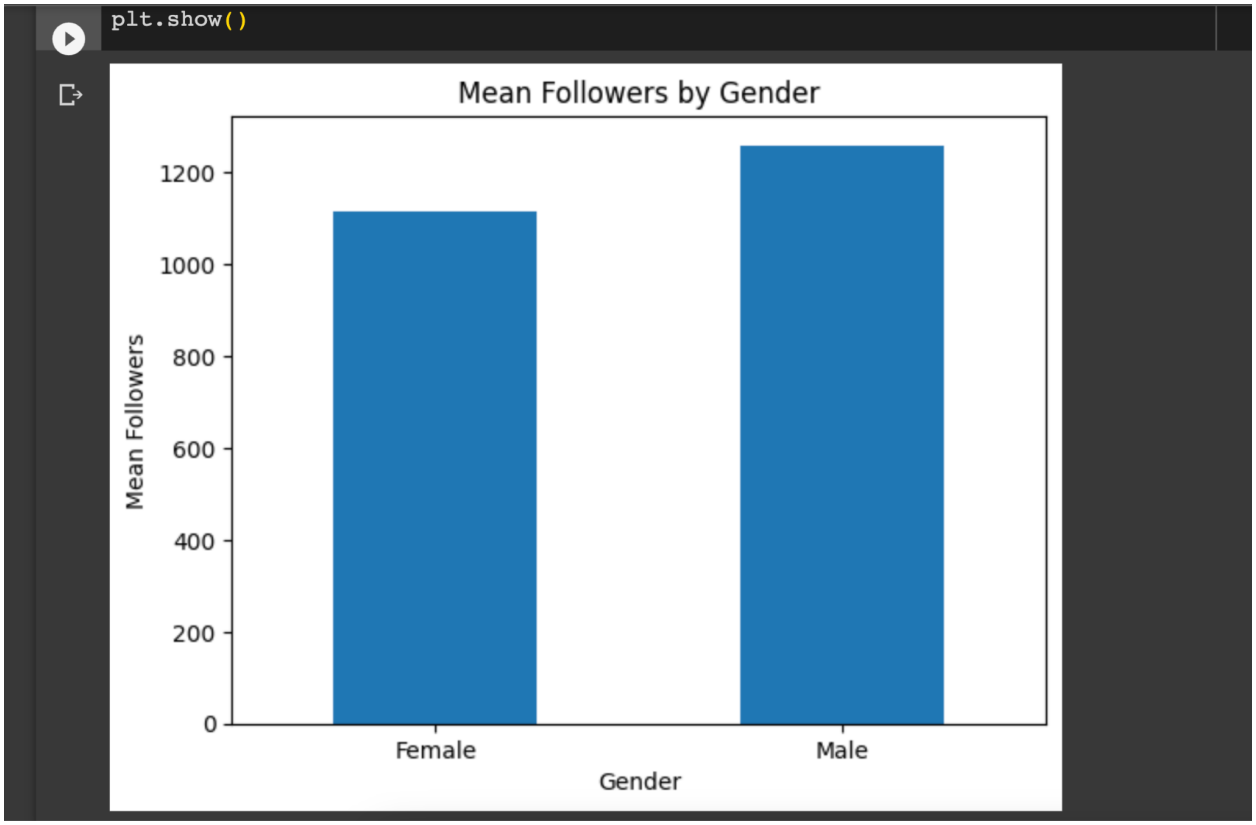
Mean Number of Followers

- For Males: 1,259.81
- For Females: 1,117.08

Insights

It's important to note that both genders have a wide range of follower counts, with some individuals having no followers, while others have significantly high counts. This suggests that there is significant variation in the popularity and influence of individuals on LinkedIn.

Males have a somewhat larger average number of followers than females. The number of LinkedIn followers shows a person's reach and influence within the platform's professional network. Both genders have a wide variety of followers, with girls having roughly 1,100 and guys having around 1,260. These statistics emphasise the need to cultivate an interesting LinkedIn profile. A high number of followers can suggest thought leadership, whilst a low number can indicate a more private or less active LinkedIn presence. It is important to note, however, that follower numbers alone do not represent the complete picture of an individual's impact, as the quality of engagement and network linkages also contribute.



Number of Organizations Worked For (No_of_Orgs_Worked_For)

Mean Number of Organizations Worked For

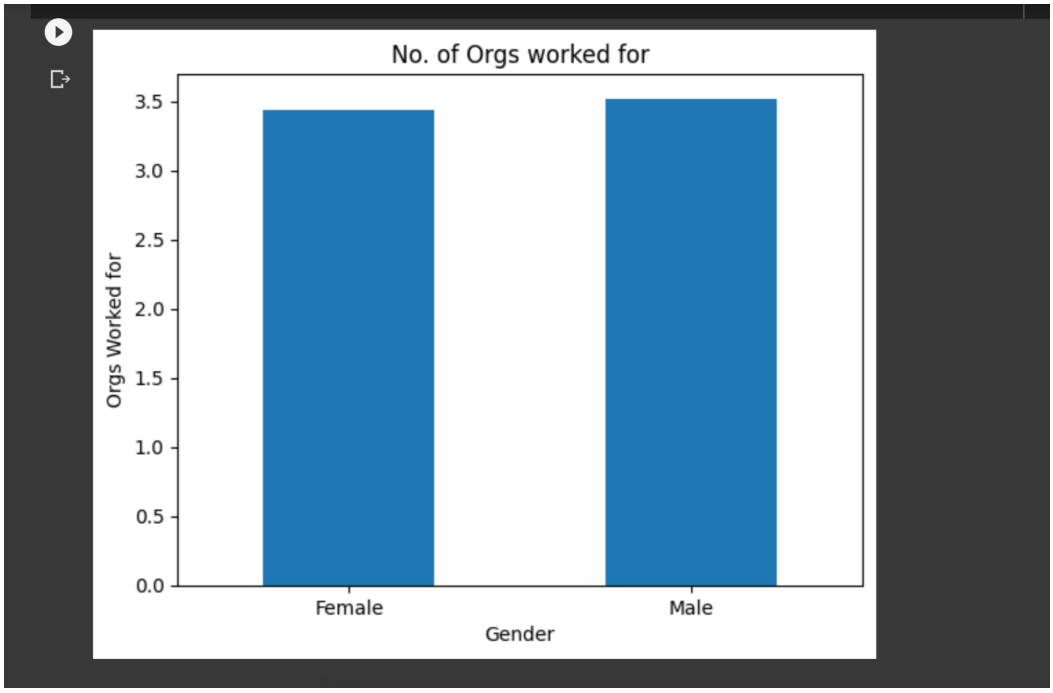
- For Females: 3.43
- For Males: 3.52

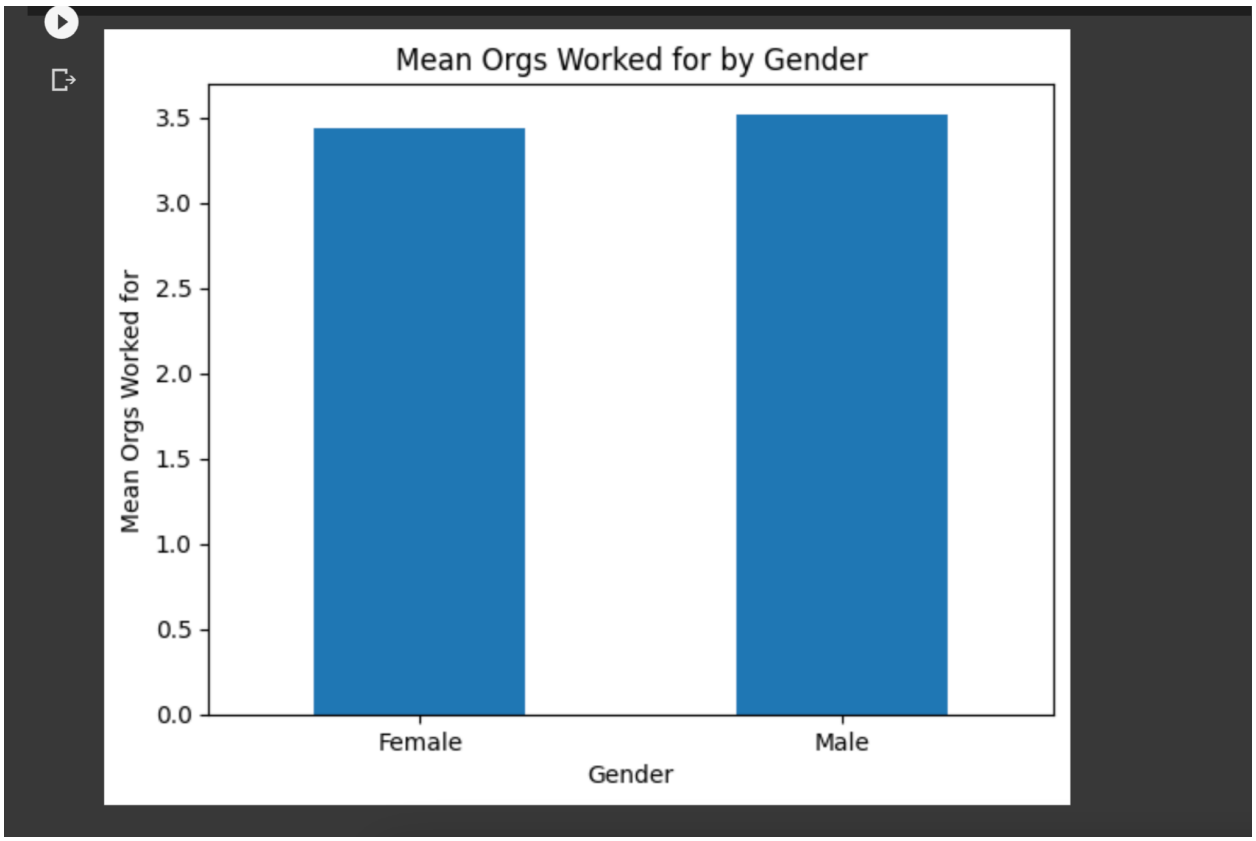
Insights:

On average, both genders have worked for a similar number of organisations.

The average number of organisations for which both genders have worked is 3.43 and 3.52, respectively. This means that, on average, individuals in the sample have worked at a similar number of organisations. The variety is notable, with some individuals having worked for a single organisation their entire careers, while others have changed jobs frequently. This variation reflects the diverse professional routes that people take, with some preferring stability and devotion to one organisation while others preferring a broader range of experiences.

The range of organizations worked for is relatively broad, with individuals having worked for as few as 1 organization to as many as 46 or 49 organizations. This indicates diverse career experiences among the individuals in the dataset.


No. of Orgs worked for
Orgs Worked for
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Female
Male
Gender


Mean Orgs Worked for by Gender
Mean Orgs Worked for
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Female
Male
Gender

Positions in Previous Tenure (Positions_in_Previous_Tenure)

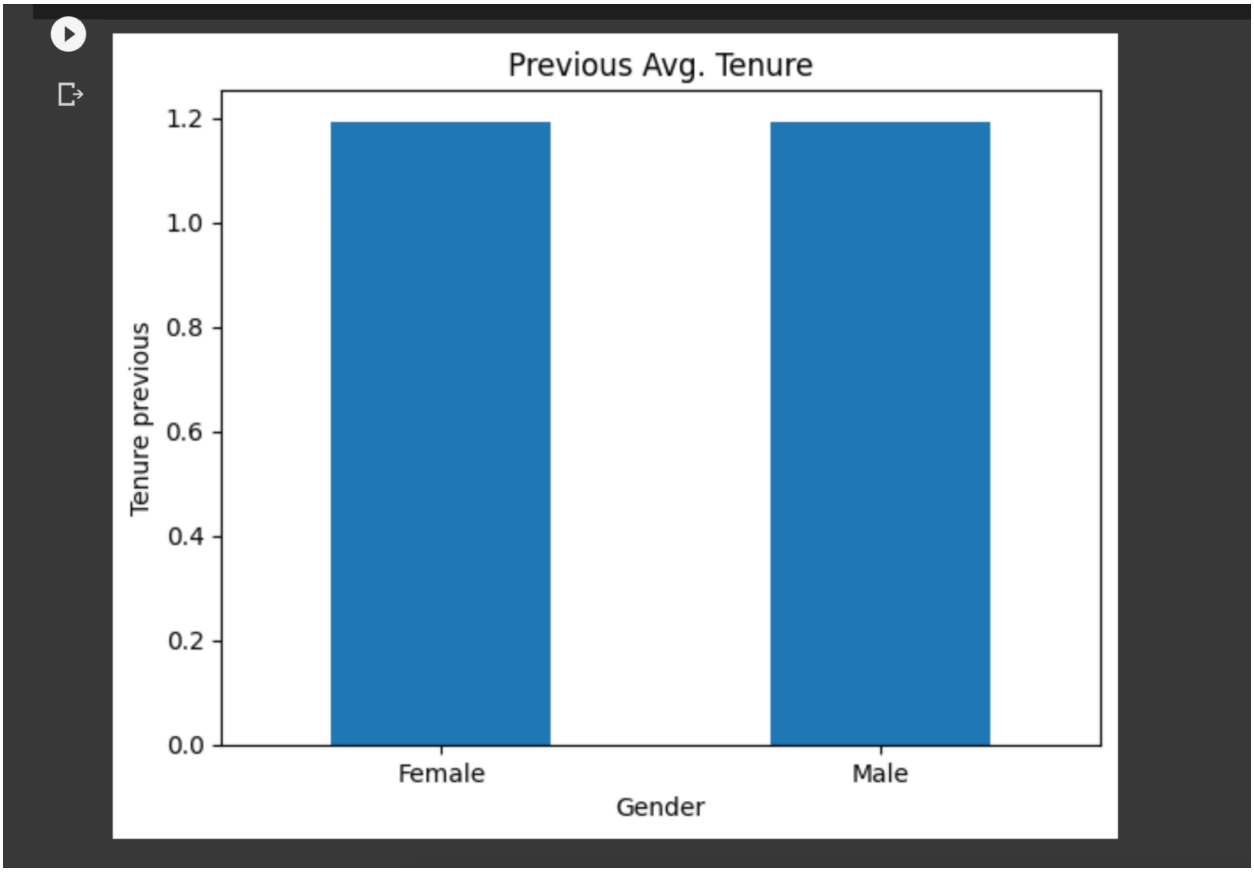
Mean Number of Positions in Previous Tenure:

- For Females: 1.19
- For Males: 1.19

Insights

Individuals in the sample have held an average of roughly 1.19 posts throughout their past tenures, demonstrating remarkable constancy across genders. This research suggests that LinkedIn users have generally steady career pathways with few job changes in their past employment. However, it's important to note that outliers exist, with some people holding up to 15 spots. This shows that, whereas most professionals have relatively stable career paths, others have had more dramatic employment shifts.

The range of past roles held is rather narrow, with most individuals holding between 1 and 15 positions. This implies that individuals have a generally consistent professional path with only a few position changes in past roles.



Accounting for Human/Computer Error

It is critical to use data quality checks and validation methods to ensure the data's dependability and accuracy. Outliers, anomalies, and potential data entry errors should be recognised and resolved to ensure the dataset's integrity. Furthermore, it is critical to validate gender classification accuracy, as this information might have a substantial impact on subsequent studies and interpretations. Addressing data mistakes and inconsistencies is critical for gaining useful and trustworthy insights.

Summary of Analysis and Insights

Finally, the deep analysis of the dataset provides a full perspective of the demographics and job characteristics of LinkedIn members. It emphasises the platform's attractiveness to a wide range of age groups, the value of developing an engaged network, and the variety of career choices

pursued by professionals. This information can be used to create focused marketing efforts, personalised career counselling, and educated decisions in a variety of professional domains.

```
print(df.describe())
```

```
       Positions_in_Previous_Tenure    Total_YOE   No_Of_Positions_Total  \
count                  62709.000000  62709.000000           62709.000000
mean                       1.191567      3.015914               1.243506
std                        0.556407      2.700674               0.756841
min                        1.000000      0.000000               1.000000
25%                        1.000000      1.470000               1.000000
50%                        1.000000      2.420000               1.000000
75%                        1.000000      3.840000               1.000000
max                       15.000000    108.990000              12.000000

       No_of_Orgs_Worked_For  Current_Tenure_Length           Age  \
count           62709.000000           62709.000000  62709.000000
mean                3.498653               2.637674     44.048717
std                 3.112945               2.981699     10.683842
min                 1.000000              -0.330000      1.000000
25%                 2.000000               0.830000     37.000000
50%                 3.000000               1.750000     44.000000
75%                 4.000000               3.330000     51.000000
max                49.000000              59.960000     80.000000

           Followers
count   62709.000000
mean     1225.838173
std      6406.523908
min         0.000000
25%       405.000000
50%       725.000000
75%      1244.000000
max    530566.000000
```

```
print(df.info())
```

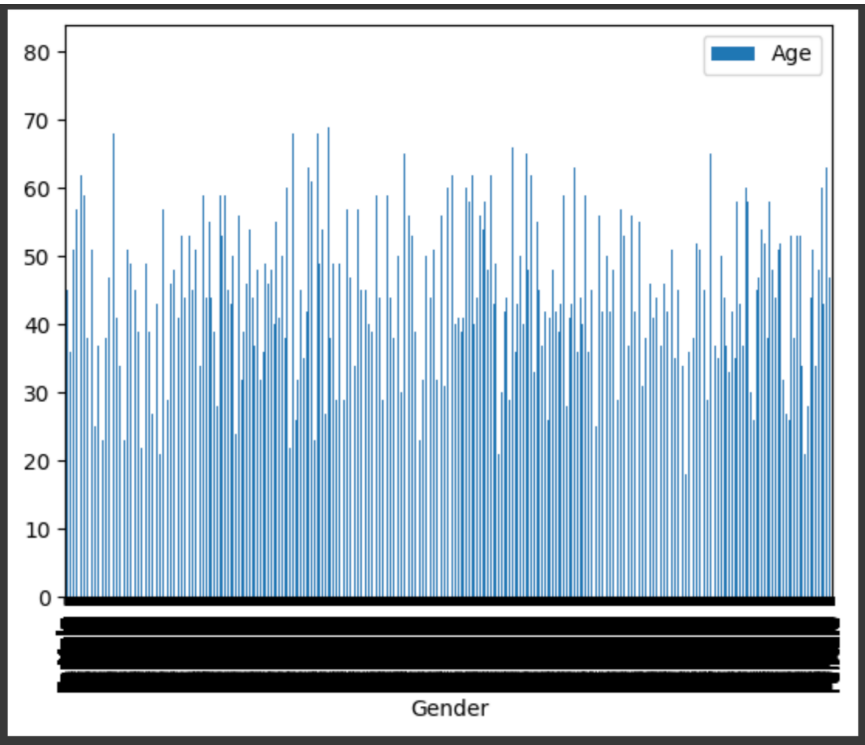
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62709 entries, 0 to 62708
Data columns (total 9 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Positions_in_Previous_Tenure  62709 non-null  int64
 1   Total_YOE                    62709 non-null  float64
 2   Org_Name                     62703 non-null  object
 3   No_Of_Positions_Total        62709 non-null  int64
 4   No_of_Orgs_Worked_For        62709 non-null  int64
 5   Current_Tenure_Length        62709 non-null  float64
 6   Age                          62709 non-null  int64
 7   Gender                       62709 non-null  object
 8   Followers                    62709 non-null  int64
dtypes: float64(2), int64(5), object(2)
memory usage: 4.3+ MB
None
```

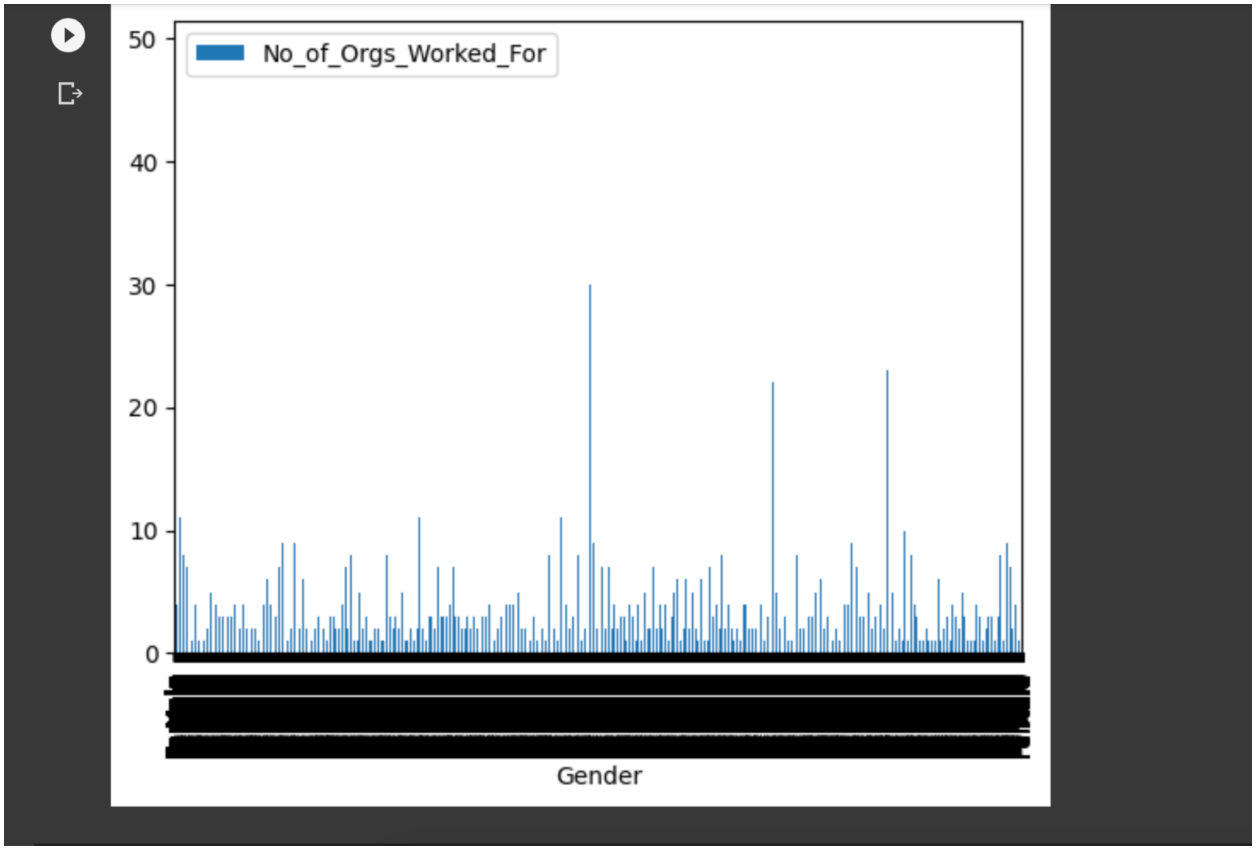
```
print(df.head())
```

```
   Positions_in_Previous_Tenure  Total_YOE                    Org_Name  \
0                             2       3.67                          TD
1                             2       2.46  Light Up The World (LUTW)
2                             1       1.83                     Glacier
3                             1       1.54                  Sprout App
4                             1       1.30                College Pro

   No_Of_Positions_Total  No_of_Orgs_Worked_For  Current_Tenure_Length  Age  \
0                      1                      1                   1.25   37
1                      1                      2                   0.58   37
2                      1                      3                   0.67   37
3                      1                      4                   0.34   37
4                      1                      5                   0.67   37

  Gender  Followers
0   Male        420
1   Male        420
2   Male        420
3   Male        420
4   Male        420
```





The dataset gives useful insights into the demographics and job characteristics of LinkedIn users. While there are some disparities in age, following, and organisations worked for between genders, these variances are minor, indicating that both genders have similar professional profiles.

LinkedIn users have various professional backgrounds and experiences, as seen by the huge differences in follower numbers, organisations worked for, and positions held in past tenures.

This study can be used to gain a better understanding of the dataset's properties and to influence decisions about talent acquisition, marketing, or any other area where LinkedIn data insights are useful.

Managerial Insights & Implications

Age:

Insights:

- Understanding the age distribution of LinkedIn users can be valuable for marketing and recruitment strategies. For instance, if a company wants to target younger professionals, it can focus on creating content and job postings that resonate with this demographic.
- Recognizing the presence of a diverse age group on LinkedIn suggests that the platform serves professionals at various career stages. This broad user base can be leveraged for talent acquisition, networking, and promoting products or services.

Implications:

- Tailor marketing campaigns to different age groups by considering the content and messaging that resonates with each segment.
- When recruiting, acknowledge the diverse age range and be open to candidates from various experience levels.

Followers:

Managerial Insights:

- High follower counts indicate a strong presence and influence on LinkedIn. Professionals with significant followers can be considered thought leaders in their respective fields.
- The wide range of follower counts highlights the importance of engagement strategies. Companies and individuals should focus on creating valuable and engaging content to grow their follower base.

Managerial Implications:

- Encourage employees to actively engage with LinkedIn to expand their professional networks and influence.
- Recognize and collaborate with LinkedIn influencers to leverage their reach for marketing or branding efforts.

Number of Organizations Worked For

Managerial Insights:

- The similarity in the average number of organizations worked for by both genders suggests that career mobility is relatively consistent.
- The presence of individuals with a single organization on their career history implies that some professionals value long-term commitments, loyalty, and stability.

Managerial Implications:

- When recruiting, consider candidates with diverse career experiences and those who have demonstrated loyalty to a single organization.
- Recognize that individuals with different career trajectories can bring unique perspectives to the workplace.

Positions in Previous Tenure

Managerial Insights:

- The consistency in the average number of positions held in previous tenures indicates that, on average, LinkedIn users have stable career paths with limited job changes.
- The presence of outliers suggests that while most professionals follow a consistent trajectory, some are open to more dynamic career changes.

Managerial Implications:

- When hiring, consider the balance between candidates with stable career paths and those with diverse experiences.
- Promote a workplace culture that values both career stability and adaptability to accommodate the needs and preferences of diverse employees.

Overall Managerial Implications:

- Recognize that LinkedIn offers access to a wide range of professionals at different career stages, from various age groups, and with diverse career experiences.
- Tailor marketing, recruitment, and engagement strategies to accommodate this diversity.
- Encourage employees to actively engage on LinkedIn, fostering both their personal brands and the company's presence.
- Embrace the value of thought leadership and influencers in the LinkedIn ecosystem.
- When making hiring decisions, consider the unique strengths that individuals with different career paths can bring to the organization.

While the managerial insights and implications provided cover the key aspects of the data analysis, there are additional considerations and recommendations that could enhance the decision-making and strategy development for an organization

Industry-Specific Analysis: If the organization operates in a specific industry, consider conducting an industry-specific analysis. This could involve comparing the insights from your dataset with industry benchmarks to gain a competitive edge.

Geographic Analysis: Explore whether there are geographic patterns in the data. This can be valuable for companies with regional or global operations, as it can inform localization strategies and help target specific markets.

Segmentation: Beyond gender, segment the data into other relevant categories, such as job titles, industries, or education levels. This can provide more granular insights and allow for more targeted strategies in marketing, recruitment, and engagement.

Content Strategy: Develop a content strategy based on the insights. Create content that resonates with different segments of the LinkedIn audience. This can include articles, videos, webinars, and more. Tailor the content to address the unique needs and interests of the target groups.

Engagement Metrics: Monitor engagement metrics (likes, comments, shares) on your LinkedIn content. Analyze which types of posts and topics generate the most interaction. Use this data to refine the content strategy and maximize engagement.

Competitor Analysis: Consider conducting a competitive analysis on LinkedIn. Investigate what the competitors are doing on the platform, including the type of content they share, their follower growth strategies, and their engagement rates. Use these insights to refine the approach.

Ethical Considerations: Ensure that the data collection and analysis practices adhere to ethical guidelines and data privacy regulations. Respect user privacy and obtain necessary permissions for data usage.

Continuous Monitoring: LinkedIn, like any social media platform, evolves over time. Continuously monitor changes in user behavior, platform features, and algorithm updates. Adapt the strategies accordingly to stay effective.

Feedback Mechanisms: Establish feedback mechanisms for LinkedIn audience. Encourage them to provide input on the content they find valuable and suggestions for improvement. This can help you tailor the offerings to better meet their needs.

Data Integration: If applicable, integrate LinkedIn data with other data sources, such as CRM systems or website analytics. This holistic view can provide deeper insights into the impact of LinkedIn activities on the overall business.

By considering these additional aspects, we can create a more comprehensive and adaptable LinkedIn strategy that aligns with your organizational goals and responds to the evolving dynamics of the platform and your target audience.

Incorporating these managerial insights and implications into strategies can help an organization make the most of its LinkedIn presence, recruit top talent, and engage effectively with professionals from various backgrounds and career stages.