Data Scientist job change analysis

Overview

This report details a data analysis process using Python, Pandas, MySQL, and popular data visualization libraries such as Matplotlib and Seaborn. The dataset used, enrollee_data, contains information about enrollees, which is retrieved from a MySQL database and analyzed for patterns related to education level, job change, experience, and other demographic factors.

Steps Performed

Connecting to MySQL Database

The MySQL connector is used to establish a connection to the local MySQL server. A SQL query is executed to fetch data from the table enrollee_data:

```
conn = mysql.connector.connect(
   host='localhost',
   user='root',
   password='#MYSql@1',
   database='SQL_PROJECT'
)
query = "SELECT * FROM enrollee_data;"
df = pd.read_sql(query, conn)
conn.close()
```

1. The connection is closed after fetching the data.
2. **Exploratory Data Analysis (EDA)**
    - **Shape and Size**: The dataset has 38,316 entries and 14 columns. This gives us a total of 536,424 data points.
    - **Data Information**: All 14 columns are non-null, with various data types like integers, floats, and objects.
    - **Unique Values**: The dataset contains a mix of categorical and numerical variables. For instance:
        - There are 93 unique city development indexes.
        - Gender has four categories: Male, Female, Other, and Not Known.
        - The target variable, which indicates job change, has two values (0 or 1).
3. **Descriptive Statistics** Using df.describe(), basic statistics such as mean, standard deviation, and quartiles for numerical columns were computed.

Data Preprocessing and Cleaning The dataset has no missing values, confirmed through:
df.isna().sum()

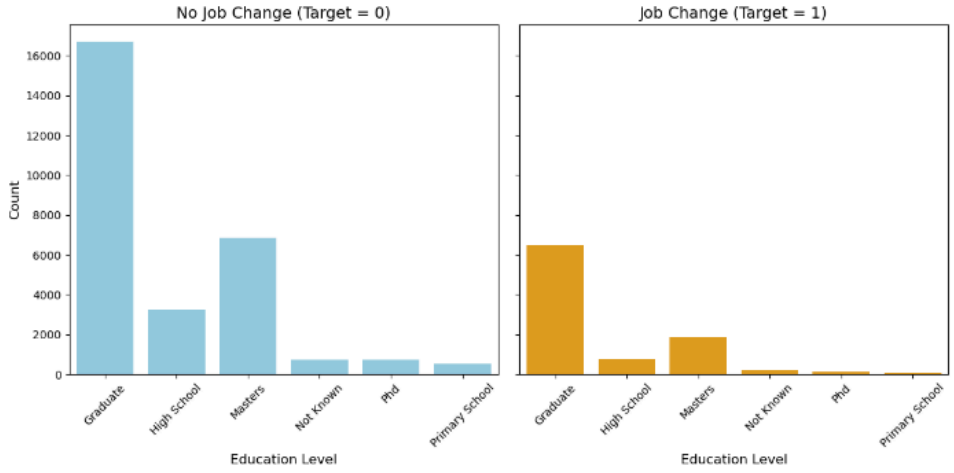
4. Therefore, no additional cleaning steps were needed.

Analysis and Visualizations

1. Education Level vs Job Change

Using the groupby function, the relationship between education level and job change (target) was explored:

```
education_target_counts = df.groupby(['education_level', 'target']).size().reset_index(name='count')
```



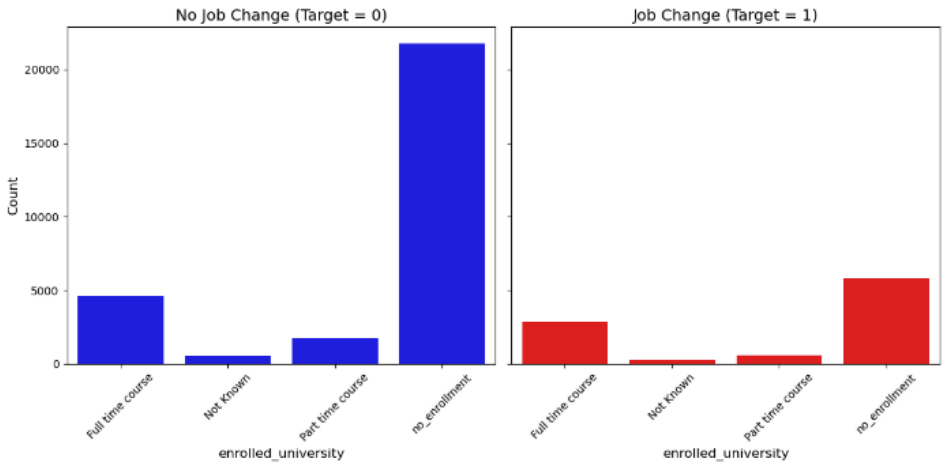
Two bar plots were created to compare the counts of enrollees who changed jobs versus those who didn't, grouped by their education level:

- **Key Insights**:
  - Graduates have the highest number of enrollees for both job change (target=1) and no job change (target=0).
  - Individuals with high school education or less are less likely to change jobs.
  - Enrollees with Master's degrees also form a significant group.

2. Enrolled University vs Job Change

Similarly, the relationship between university enrollment status and job change was examined:

```
university_target_counts = df.groupby(['enrolled_university', 'target']).size().reset_index(name='count')
```

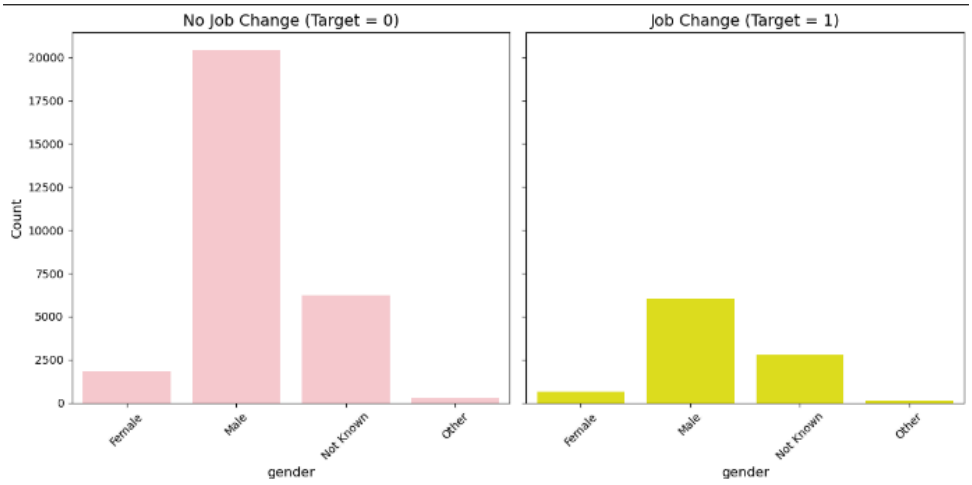


The visualization shows that enrollees without any enrollment had the highest count, especially among those not looking for a job change.

3. Gender vs Job Change

Gender distribution among those who changed jobs and those who didn't:

```
gender_target_counts = df.groupby(['gender', 'target']).size().reset_index(name='count')
```

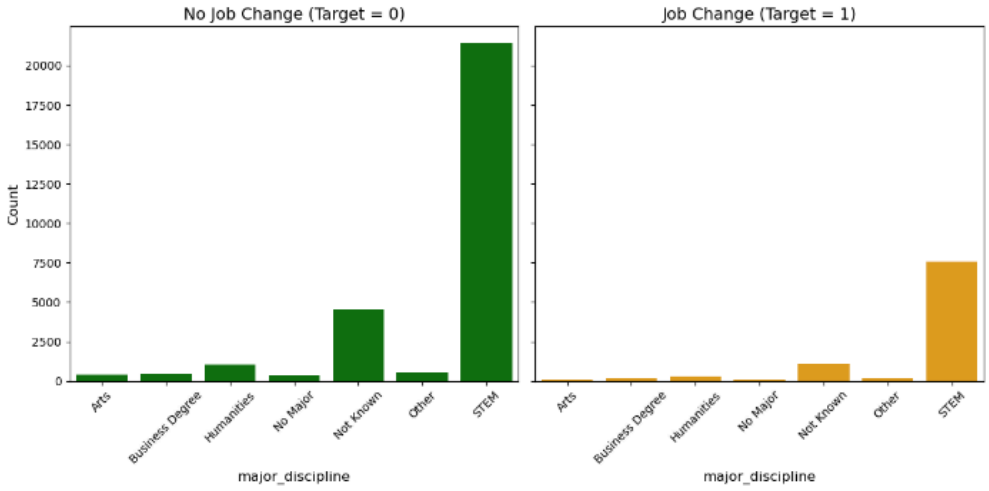


- **Key Insights**:
  - The majority of the dataset consists of male enrollees, and they are more likely to have no job change.
  - A significant portion of enrollees are classified as "Not Known" in terms of gender.

4. Major Discipline vs Job Change

Exploring the distribution of major disciplines among enrollees, split by job change status:

major_discipline_target_counts = df.groupby(['major_discipline', 'target']).size().reset_index(name='count')

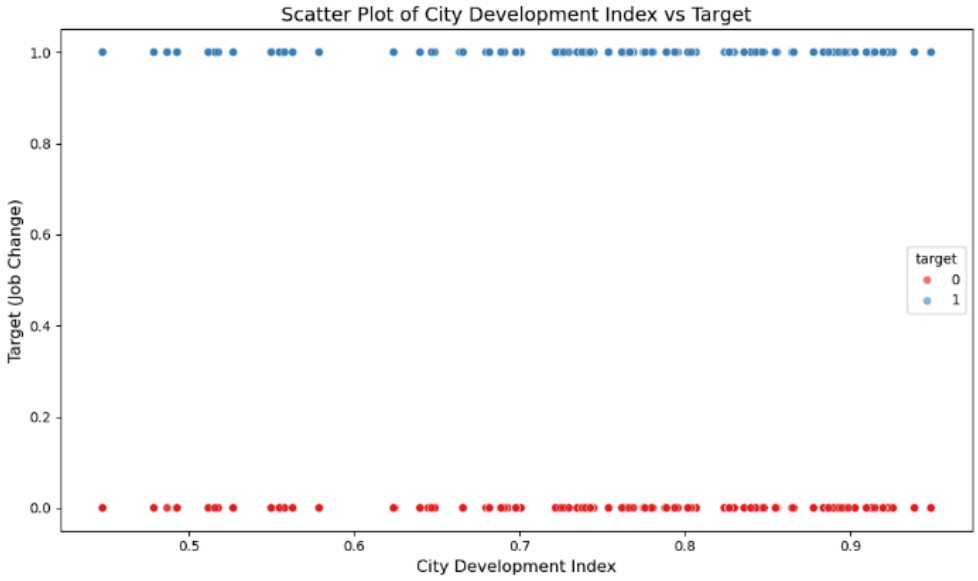


- **Key Insights**:
  - The majority of enrollees come from a STEM background, and they dominate both job change and no job change categories.

5. City Development Index and Job Change (Scatter Plot)

A scatter plot was created to visualize the relationship between city_development_index and target:

sns.scatterplot(x='city_development_index', y='target', data=df, hue='target', palette='Set1')



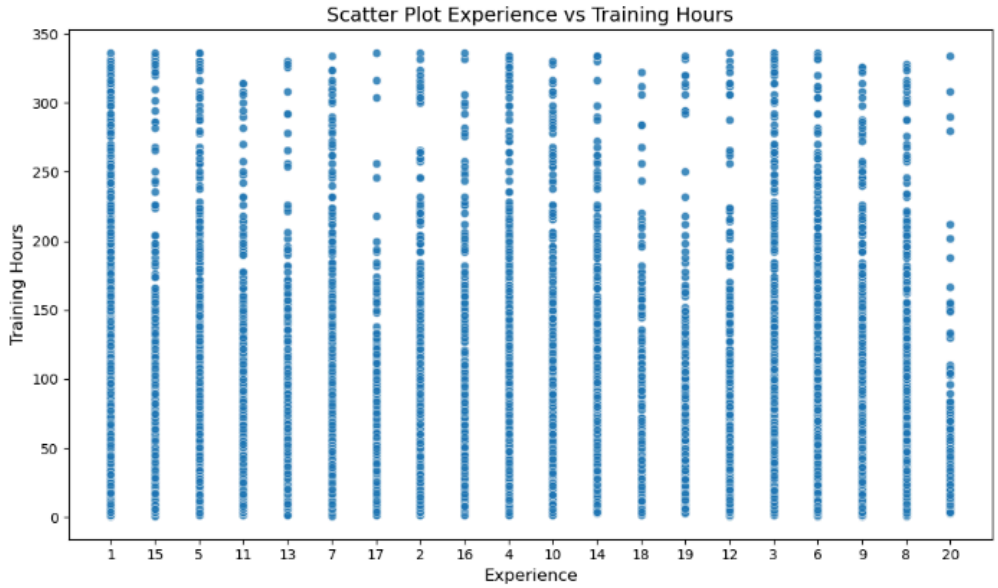
- **Key Insights**:

- There is no clear linear relationship between the city development index and job change.
- Cities with higher development indexes seem to have a mix of individuals who did or did not change jobs.

6. Experience and Training Hours (Scatter Plot)

Another scatter plot was generated to see if there's a relationship between experience and training_hours:

```
sns.scatterplot(x='experience', y='training_hours', data=df)
```

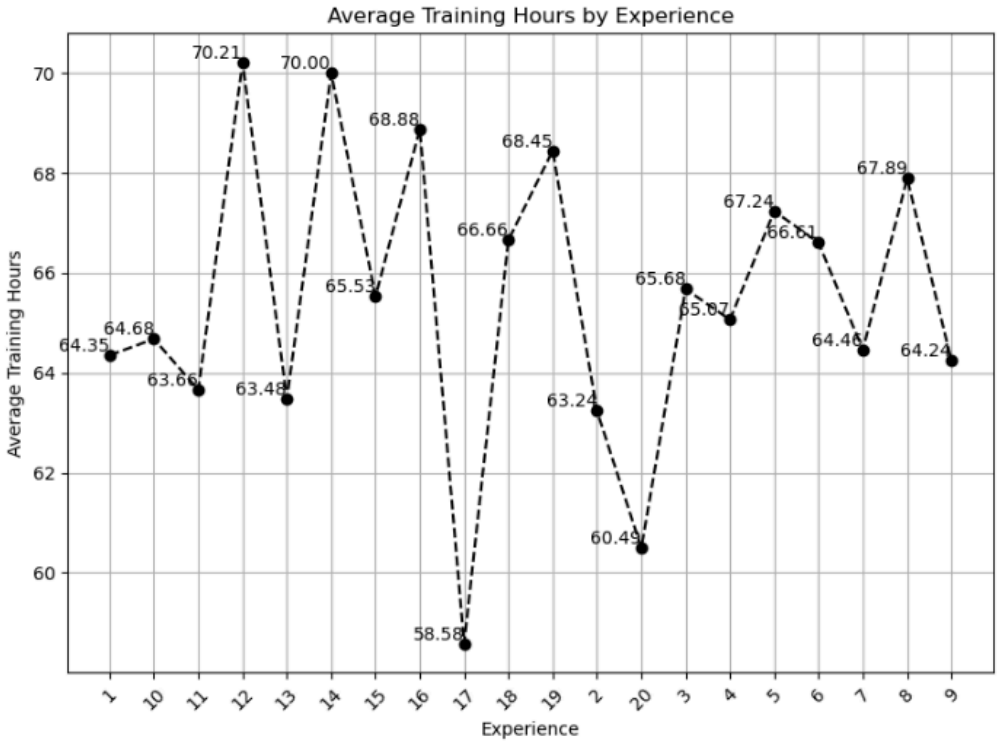


- **Key Insights**:
  - Individuals with more experience do not necessarily have more training hours. The data shows a dispersed pattern.

7. Average Training Hours by Experience (Line Plot)

The mean training hours for each experience level were calculated and visualized using a line plot:

```
experience_training_hours = df.groupby('experience')['training_hours'].mean()
plt.plot(experience_training_hours.index, experience_training_hours.values, marker='o', linestyle='dashed')
```

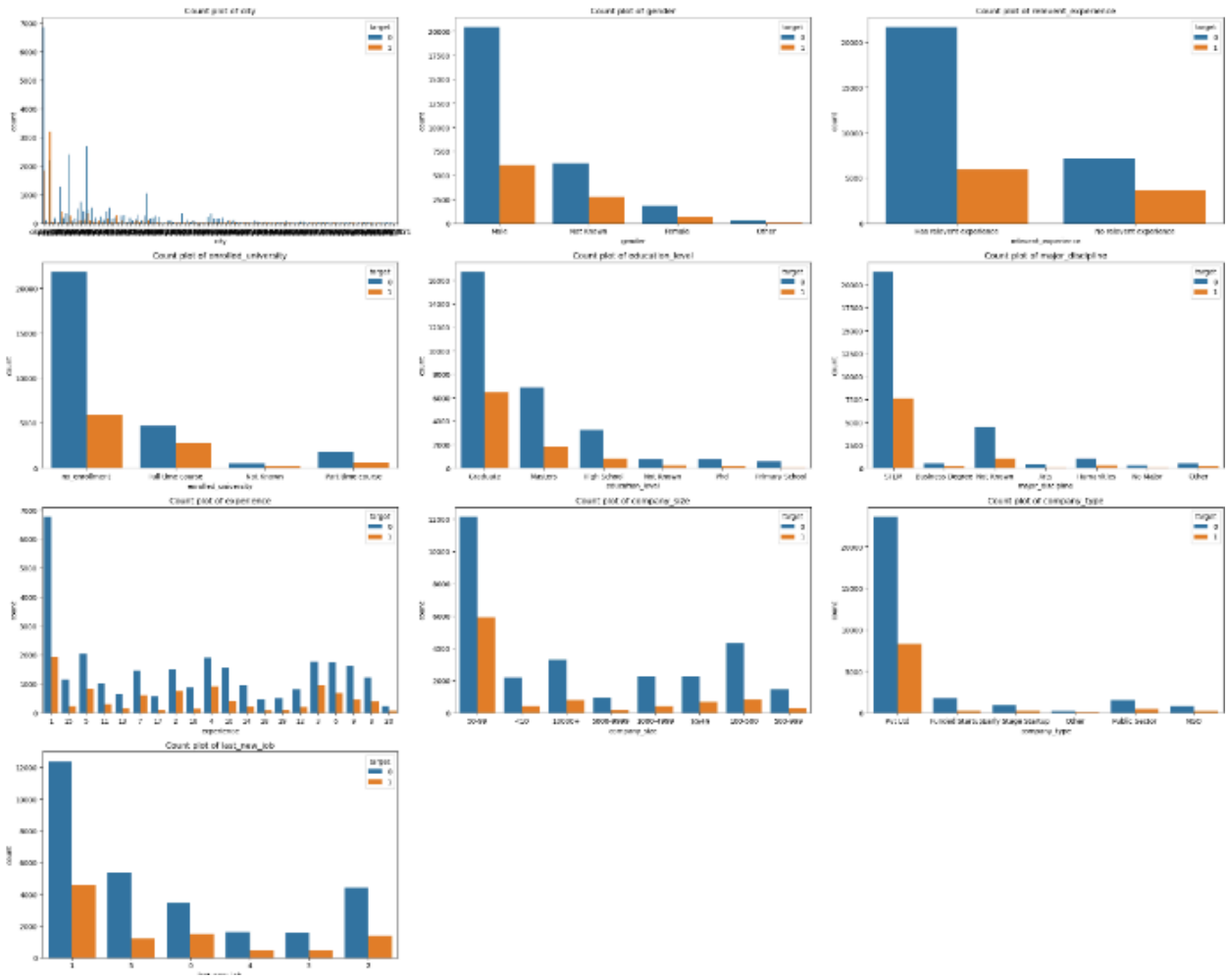


- **Key Insights**:
    - There is a fluctuation in the average training hours across different experience levels, with no clear increasing or decreasing trend.

8. Categorical Feature Analysis

A set of count plots were generated for each categorical variable, displaying the distribution of the data:

```
for i in range(len(cat_feat)):
    sns.countplot(x=cat_feat[i], data=df, hue='target')
```

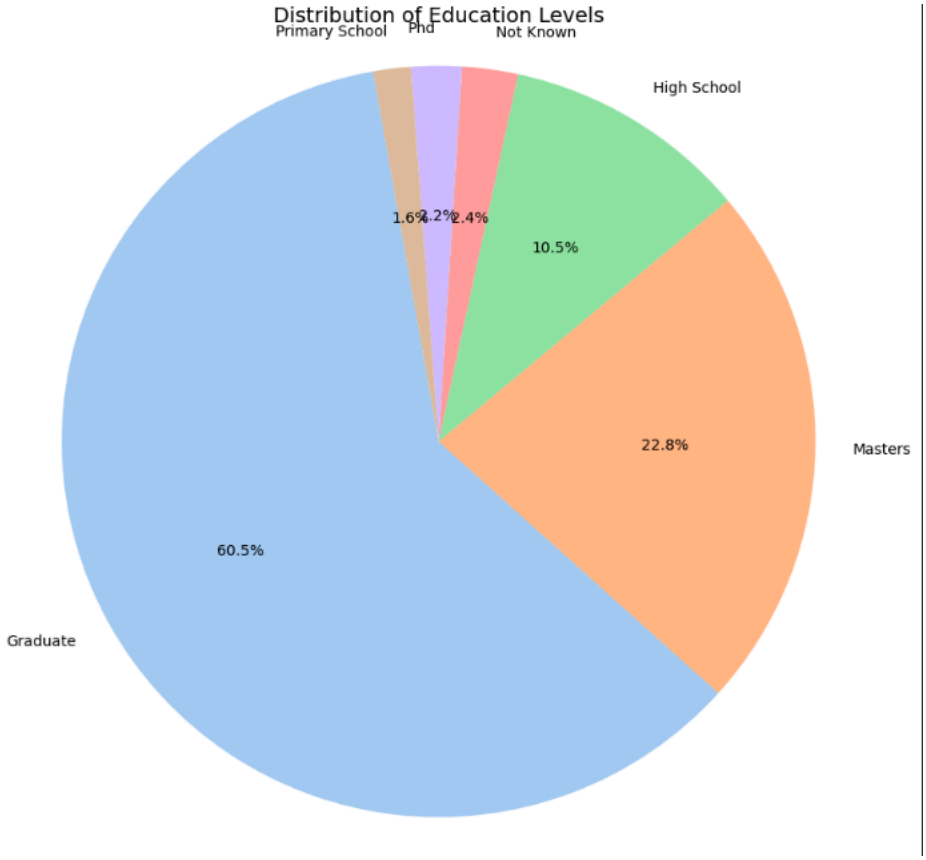


- **Key Insights**:
  - Most categorical features show imbalances in distribution. For example, there are significantly more male enrollees than females.
  - Variables like company type and company size also show skewed distributions.

9. Education Level Distribution

The overall distribution of education levels was visualized using a pie chart:

```
plt.pie(education_counts, labels=education_counts.index, autopct='%1.1f%%')
```



- **Key Insights**:
    - The majority of enrollees have a graduate degree, followed by those with a Master's degree.

Conclusion

The analysis reveals several key patterns:

1. **Education Level**: Graduates are the largest group of enrollees, and individuals with higher education levels are more likely to change jobs.
2. **Gender Distribution**: Male enrollees dominate the dataset, and the gender "Not Known" category is surprisingly large.
3. **Experience and Training Hours**: There is no strong correlation between experience and training hours.
4. **City Development Index**: There is no clear relationship between a city's development index and the likelihood of a job change.