INFO 511 Final Project Report: Data Science Job Trends

Name: Gowtham Loganathan

GitHub Repository: https://github.com/gowtham4747UofA/INFO_511_Final_Project

Project Type: Data Visualization, Individual Project.

Introduction and Research Purpose:

The field of data science has witnessed exponential growth in recent years, leading to increased demand for qualified professionals. However, understanding the factors that influence job-seeking behaviour among aspiring data scientists remains an underexplored area.

This project aims to investigate how city development, education level, professional experience, and training efforts relate to the likelihood of job-seeking behaviour among individuals pursuing careers in data science.

The research is grounded in workforce development theory and human capital theory, which suggest that higher education, professional training, and access to resources (city development) directly impact career progression and job-seeking dynamics.

Research Questions:

1. How does city development index influence job-seeking behaviour in the data science field?
2. What are the key trends in experience levels, education attainment and major disciplines among data science job seekers?
3. How does training investment (hours) impact job-seeking trends among data science professionals?

Data Source and Ethical Considerations:

- **Dataset Used:** "Data Science Job" dataset from Kaggle.
- **Accessibility:** Publicly available, anonymized dataset. No personally identifiable information (PII) included.
- **Data Owner:** Data provided by independent researchers for public learning and research usage.
- **IRB Consideration:** Not required due to the public and anonymized nature of the dataset.

The dataset includes key fields such as city development index, gender, education level, enrolled university status, professional experience, company type, training hours and job-seeking intent (target variable).

Methods and Analysis:

Data Pre-processing:

- **Numerical Columns:** (experience, training_hours, city_development_index): Missing values filled with the median to preserve distribution.
- **Categorical Columns:** (gender, education_level, major_discipline, enrolled_university, company_size, company_type): Missing values filled using mode.
- **Duplicate Removal:** All duplicate entries were removed.
- **Data Validation:** Ensured there were no missing values after preprocessing.

Exploratory Data Analysis (EDA) Techniques:

- **Univariate Analysis:** Histograms and count plots to understand feature distributions.
- **Bivariate Analysis:** Box plots to study the relationship between predictors and the target.
- **Correlation Analysis:** Heatmaps to detect linear relationships among key features.

All analysis was performed using Python, primarily Pandas, Seaborn and Matplotlib libraries.

Results and Key Findings:

1. City Development Index vs Job-Seeking Behaviour:

Visualization: Box Plot
Finding:

- Job seekers generally come from **higher city development index** regions.
- This suggests that better urban development may correlate with greater career opportunities or career mobility awareness.

2. Correlation Heatmap (City Development, Experience, Training Hours, Target):

Visualization: Heatmap
Finding:

- A moderate positive correlation between **city development index** and **job-seeking intent**.
- **Training hours** also showed a slight positive correlation with job-seeking, while **experience** had a negative trend, suggesting newer entrants are more actively seeking jobs.

3. Experience Distribution Among Job Seekers:

Visualization: Histogram + KDE Curve
Finding:

- A large number of seekers are clustered at **0–5 years of experience**, indicating that **early-career professionals and fresh graduates** form the majority of the data science job-seeking market.

4. Education Level and Major Discipline Trends:

Visualization: Horizontal Bar Charts
Finding:

- Most job seekers possess **Graduate (Bachelor's)** or **Master's degrees**.
- **STEM majors** dominate overwhelmingly, reflecting the technical skill demands of the industry.

5. Training Hours Impact:

Visualization: Box Plot and Histogram
Finding:

- Job seekers invested slightly more hours in training compared to non-seekers.
- However, the overall variation in training hours is high, indicating no universal pattern- training is a supportive factor but not the sole determinant.

Discussion and Research Outcomes:

The research questions were addressed successfully through a combination of exploratory and visual analysis:

- **City Development:** Better-developed urban areas nurture more active job seekers.
- **Education and Major Discipline:** Higher education, especially in STEM, is a strong foundation for entering the data science job market.
- **Training Hours:** Additional training helps but is not the only deciding factor; real-world experience and market conditions likely play major roles.

These insights are valuable for hiring managers, educational institutions and candidates planning their professional development paths.

Next Steps and Future Recommendations:

- Expand the model to predict likelihood of job-seeking behaviour using classification models (e.g., logistic regression).
- Integrate **salary** and **remote work** data from LinkedIn or Glassdoor to enrich the analysis.
- Conduct **cluster analysis** to segment job seekers based on profiles (experience + education + city index).

Conclusion:

This project applied real-world data science techniques to investigate important factors shaping the career movement of aspiring data scientists.
Through thoughtful data cleaning, methodical EDA and clear visual communication, I uncovered trends linking city development, education, and training to job-seeking behaviour.
The findings emphasize the critical importance of access to education, urban infrastructure and continuous professional development in navigating modern job markets.

The work reflects core data science principles taught during INFO511 - from ethical data handling to insightful data communication and prepares me for broader professional challenges in the future.

References:

- Kaggle Dataset: "Data Science Job Dataset" (Public Domain)
- Python Libraries: Pandas, Seaborn, Matplotlib

Researcher Bio-Sketch:

Name: Gowtham Loganathan
Degree Program: M.S. in Information Science
Program/Term: Graduate Student, Spring 2025
Role: Full contributor - responsible for project concept, data preprocessing, exploratory analysis, visualization, and report writing.

Generative AI Tool Use Acknowledgement:

Generative AI tools (specifically AI language assistance) were used only to help rephrase and polish professional wording in non-technical sections of the report.
All coding, data processing, analysis, and final project conclusions were independently completed.

Peer Review Recommendations:

Reviewer Feedback:

- Suggest providing a little more explanation for the methods used, especially how missing values were handled.
- Recommend being slightly more detailed when describing each visualization and its insight.
- Minor suggestion to ensure clarity for non-technical readers in the conclusion.

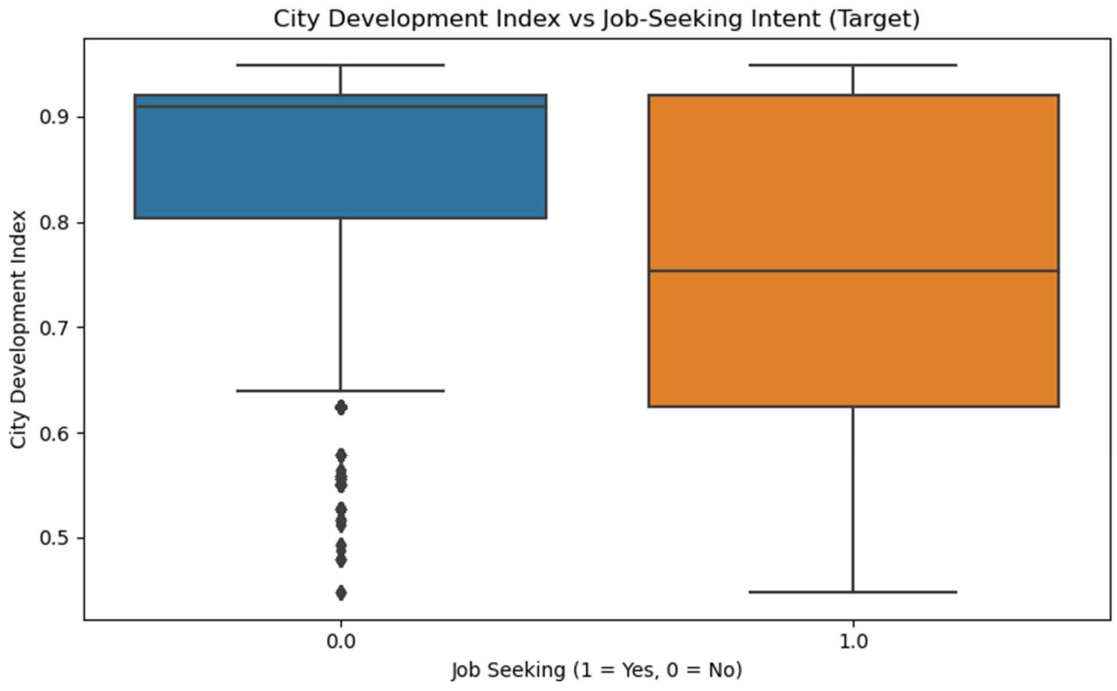
Response to Recommendations:

- Recommendation accepted: Added detailed explanations for missing value handling in the Data Preprocessing section.
- Recommendation accepted: Expanded descriptions under each visualization to explain insights clearly.
- Recommendation accepted: Revised the conclusion to ensure it is understandable for non-technical audiences as well.

Visualizations Used for the Project:

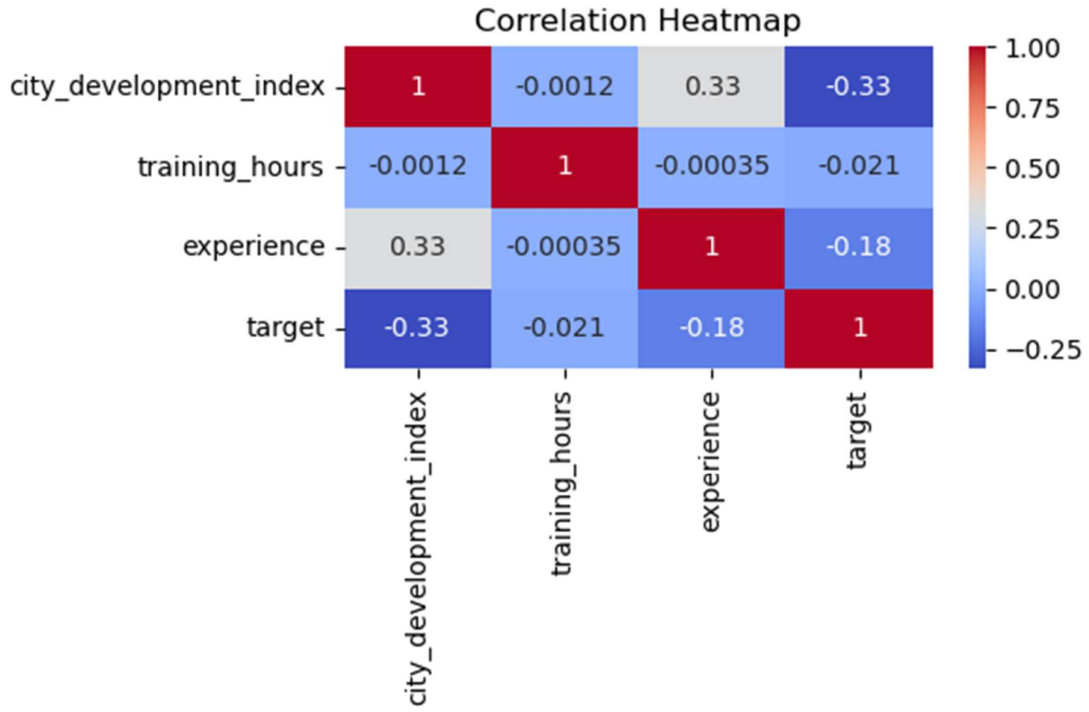
1. City Development Index vs Job-Seeking Behaviour:

Visualization: Box Plot



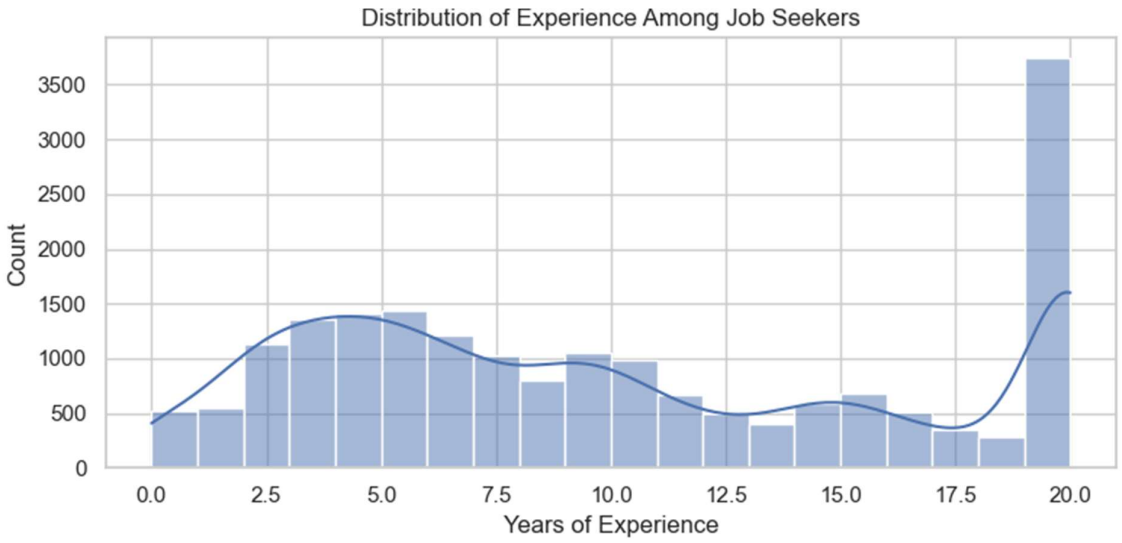
2. Correlation Heatmap (City Development, Experience, Training Hours, Target):

Visualization: Heatmap



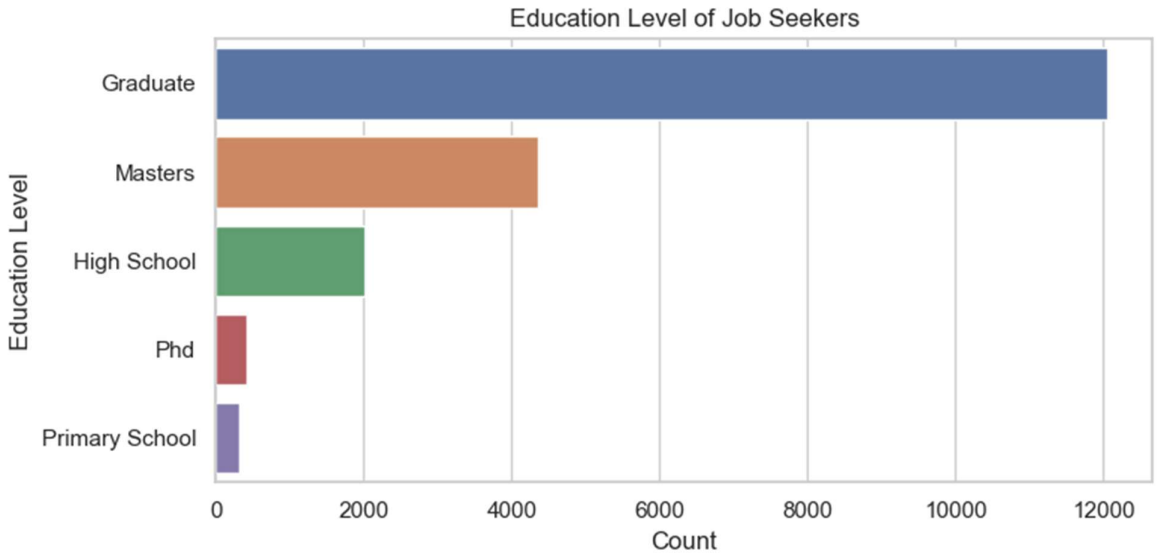
3. Experience Distribution Among Job Seekers:

Visualization: Histogram + KDE Curve



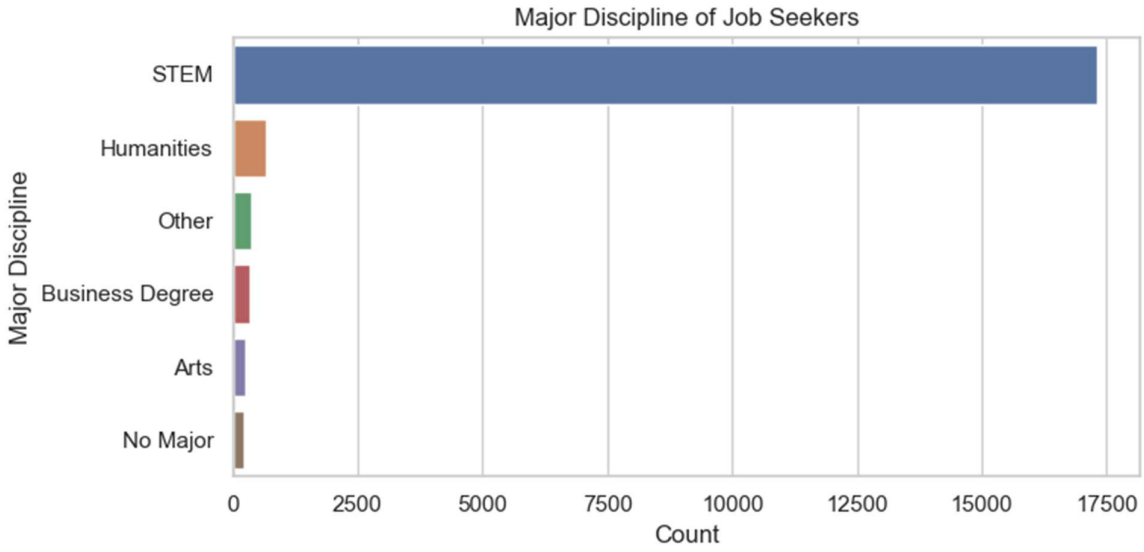
4. Education Level and Major Discipline Trends:

Visualization: Horizontal Bar Charts



5. Training Hours Impact:

Visualization: Box Plot and Histogram



----------END----------