Data Science Intern Assessment (2025)

Instructions

- *You can choose **one** of the two core tasks (Task 1 or 2) and welcome to complete Task 3.*
- *You can use either R or Python to tackle these problems.*
- *Feel free to share visualizations, conclusions, challenges, and your thought process.*
- *The submission should be a Jupyter or R Notebook for Tasks 1-2, and a separate Jupyter/R Notebook for Task 3.*
- *You are encouraged to use any tools at your disposal, including GenAI platforms, but you must document how and why you used them.*

Task 1: Property Lease-Up Analysis with Feature Engineering

Using the provided real estate dataset:

1. Identify properties delivered since April 2008 in the two markets.
2. Calculate the average lease-up time for these markets.
3. Determine which properties had negative effective rent growth during lease-up.
4. **Feature Engineering Challenge:** Create at least 5 additional features from this dataset that could improve predictive modeling of lease-up time. Explain your reasoning for each feature.
5. **Embedding Application:** Use an embedding model to group similar properties based on their characteristics. Visualize these clusters and explain what insights they provide.

Definitions:

- *"Delivered" means the property's first recorded monthly status is either LU (lease up) or UC/LU (Under construction/Lease up).*
- *"Lease-up time" is the number of months from delivery to market until the first month when the property reached 90% occupancy.*

Task 2: Migration Pattern Analysis with Data Integration

Using the provided migration dataset:

1. Identify the county pairs with the most interactions (by number of exemptions).
2. Determine which counties show the highest in-migration and out-migration.
3. Calculate net migration for the New York metropolitan area.
4. **Data Enrichment Challenge:** Using publicly available GenAI tools, augment this migration data with at least three additional contextual data points (economic indicators, demographic information, etc.) that help explain the migration patterns. Document your process.
5. **NLP Analysis:** Extract and analyze common themes or factors driving migration by scraping relevant news articles from the top migration counties.

Task 3: GenAI-Enhanced Interactive Dashboard

Using data from your completed task (1 or 2):

1. Create an interactive dashboard that presents key findings and allows for exploration.
2. Incorporate at least one GenAI-powered feature in your dashboard (e.g., natural language query interface, automated insight generation, anomaly explanation).
3. Include documentation on how you built the AI components and any prompt engineering techniques used.
4. Host the application on a public server and share the link.
5. **Reflection:** Write a brief analysis (maximum 500 words) on the advantages and limitations of your GenAI approach, including potential ethical considerations.

Evaluation Criteria

We will evaluate your submission based on:

1. ***Technical Execution:*** *Code quality, analytical approach, and correct implementation*
2. ***GenAI Integration:*** *Appropriate and innovative use of generative AI tools*
3. ***Critical Thinking:*** *Your ability to interpret results and identify limitations*
4. ***Communication:*** *Clear explanations of your methodology, findings, and visualizations*
5. ***Creativity:*** *Novel approaches to problem-solving and feature engineering*
6. ***Documentation:*** *Thorough explanation of your process, including when and how you used GenAI tools*

Good luck and have fun!