**Capstone Project: H1B Visa Data Analysis Using CRISP-DM Framework**

The goal of this project is to provide students with a real-world data analysis experience by analyzing the H1B visa dataset. The project integrates database management, data visualization, Python-based data cleaning, and storytelling to present findings. The steps align with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework.

1. **Project Overview:**

**Objective**: Understand the context and business problem. Analyze the H1B visa data to provide insights for international students seeking employment in the U.S.

* + 1. Key questions to address: Which employers sponsor the most H1B visas?
    2. What are the top-paying roles and industries?
    3. Are job types geographically concentrated?
    4. What strategies maximize an applicant’s chances of sponsorship?

**2. Data Description**

* + Dataset: ~500,000 records with 90+ columns in CSV format.
  + Contains LCA details such as job title, employer name, state, industry, salary, etc.
  + Tasks: Explore the dataset’s structure, key attributes, and missing values.

**3. Tools and Technologies Used**

* + MySQL
  + Power BI and Python for visualisations and analysis.

**4. Data Cleaning & Preprocessing**

There were many duplicates and issues with data in the wrong columns and wrong column type. There were also outliers that needed changing. These were changed in MySQL. Fact and Dimension tables were created as specified in the Project Overview.

**5. Key Insights**

**Employer sponsorship trends:** I used worksite state as best indicator of location. I didn’t feel cities was fair, would cluster towards very large cities or states with large cities. I found California, Texas and New York State having the largest numbers of visas issued. I found Amazon, Microsoft and Cognizant as the biggest employers of staff needing visas. I found Software roles were the most common roles visas were issued for.

**Salary distributions by role and location**: I produced various graphics by job title and sum or average of salary, as well as some by location. I also produced a python boxplot to see the distribution of salary values.

. Key questions to address: Which employers sponsor the most H1B visas?

Amazon, Microsoft and Cognizant

. What are the top-paying roles and industries? # **For full visibility see Power BI dashboard #**

Top paying roles: Chief of Breast Imaging, Chief Financial Officer and Executive Director, MicroVascular/Head and Neck Physician.

. Are job types geographically concentrated?

Yes, 40% of roles are in California, Texas or New York State.

. What strategies maximize an applicant’s chances of sponsorship?

Being qualified in software/computing and applying to companies that issue visas are sensible strategies to maximise the chance of getting a visa. However, many different roles have Visas issued for them.

I noticed software engineering dominates the highest salaried roles for visas by sum of salaries. Software Engineer is first, Software developer 2nd, ‘Software Engineering’ 3rd, ‘Senior Software Engineer’ 4th and ‘Software Development Engineer’ 5th.

**6. Visualisations / Dashboards**

The link to the Power BI dashboard is in the Readme file.

**7. Conclusions / Recommendations**

For students looking to get a visa in the US, working in software as one of the roles stated is obviously going to give you a much better chance. Also, applying to the employers who give out the most visas seems a sensible strategy. Attention should be paid to the non software roles as well, roles like Assistant Professor still have a lot of visas issued and other computing roles do as well.