**Capstone Project Documentation: Web Scraping, Cleaning, SQL Integration, and Visualization of Naukri Job Data**

**S. KARUPPASAMY**

**1. Introduction**

**Project Overview**

This project automates the extraction, cleaning, and analysis of job listings from the Naukri website. It involves web scraping, data cleaning, storing data in a SQL database, and visualizing insights in Power BI to help with decision-making.

**Problem Statement**

The Naukri website contains extensive job data scattered across multiple pages, including job titles, companies, salaries, and locations. Manually processing this data is time-consuming, making automation for extraction, cleaning, and analysis essential for efficiency.

**Objectives**

* **Web Scraping:** Automate the extraction of job listings from Naukri.com using Python and web scraping techniques.
* **Data Cleaning:** Clean and preprocess the scraped data to ensure consistency and usability.
* **Data Integration:** Transfer the cleaned data into a SQL database for efficient storage and management.
* **Data Visualization:** Use Power BI to create interactive and insightful dashboards for visualization of job market trends.

**2. Methodology**

**Web Scraping**

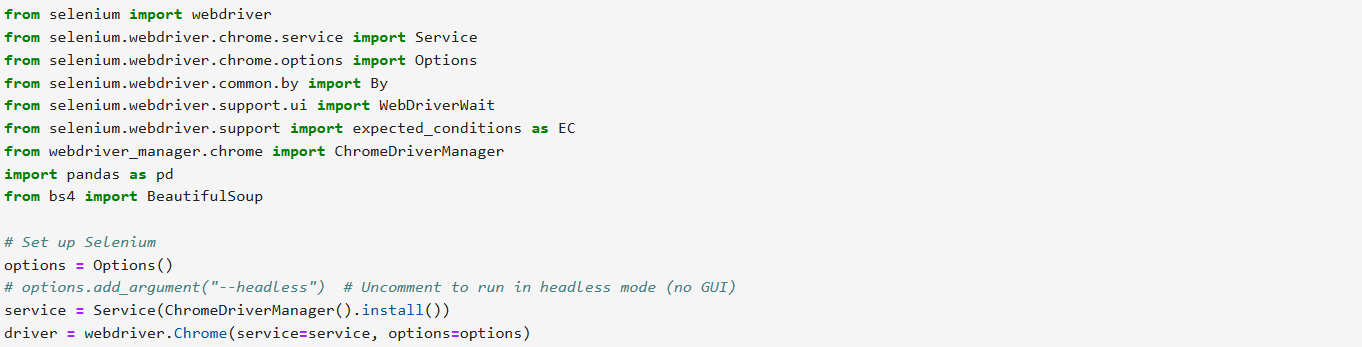
The first step of the project involved scraping job data from the Naukri website. The scraping process was done using Python libraries such as BeautifulSoup and Requests.

**Web Scraping with Python**

* **BeautifulSoup:** This library was used to parse HTML data and extract relevant information from the Naukri job listings pages.
* **Requests:** The Requests library helped in sending HTTP requests to the Naukri website and retrieving HTML content of the pages.

**Steps in Scraping:**

The scraping process involved identifying key data points like job titles, company names, and salary. I used the Requests library to retrieve HTML content and parsed it with BeautifulSoup to extract the relevant data. The scraped data was then stored in a Python dictionary or DataFrame for further processing.

**A close-up of a computer screen

Description automatically generated**

**Data Cleaning**

Once the data was scraped, it needed cleaning to remove inconsistencies, handle missing values, and transform it into a structured format. I used Python's Pandas library for this task.

**Data Preprocessing**

* **Unwanted Rows**: I cleaned the dataset by removing irrelevant or unnecessary rows that did not contribute meaningful data to the analysis.
* **Data Transformation**: Unwanted words and characters were removed from the job title and location fields to ensure consistency and clarity.
* **Duplication Removal**: Duplicate job listings were removed by checking for identical rows based on job title and company name.

**A screenshot of a computer

Description automatically generated**

**SQL Integration**

The cleaned data was then transferred into a SQL database for efficient storage, querying, and management.

**SQL Database Creation**

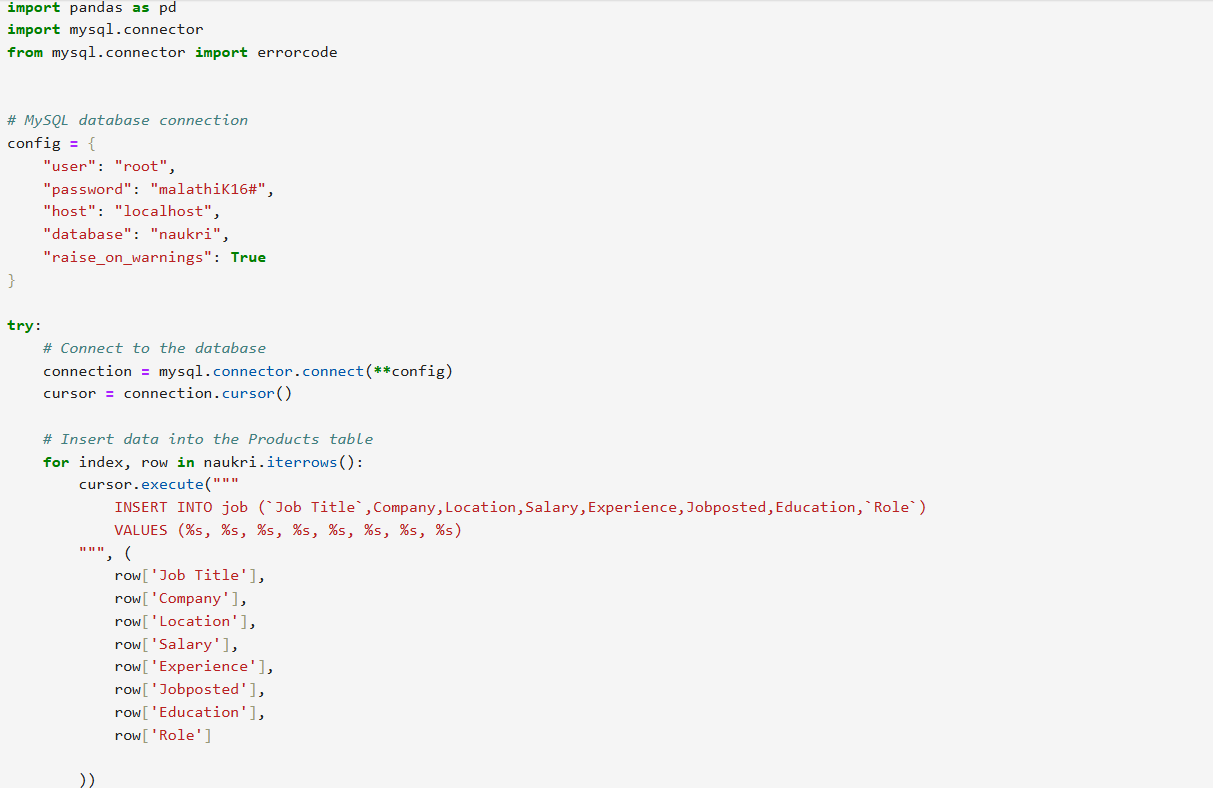
Using MySQL I created a table to store the job listing data. The table had columns for job titles, companies, locations, salary, and other relevant fields.

A screenshot of a computer program

Description automatically generated

**Data Insertion**

Using the MySQL connector, I connected to the database by providing the necessary username and password. Once connected, I inserted the cleaned data into the SQL table using Python.



**A white background with text

Description automatically generated**

**Data Visualization in Power BI**

With the data stored in SQL, I connected Power BI to the database to fetch the data and create interactive dashboards.

**Data Import to Power BI**

In Power BI Desktop, I used the SQL Server connector to link the SQL database. After connecting, I imported the job listings data and prepared it for analysis.

**Data Transformation in Power BI**

Before building the visualizations, I performed additional transformations, such as:

* **Column Splitting**: Some columns, like salary range, were split into lower and upper limits.
* **Data Aggregation:** Aggregated data was generated for the average experience level by company and job title.

**Dashboard Creation**

Using Power BI’s drag-and-drop interface, I created several visuals:

* **Job Titles Count:** A bar chart displaying the count of job listings for each job title.
* **Job Title KPI:** A KPI visualization to highlight the total number of job titles available.
* **Job Posted Duration:** A funnel chart visualizing the duration of job postings (e.g., how long jobs have been listed).
* **Maximum Salary:** A pie chart showing the maximum salary offered across different job titles.
* **Location Slicer:** A slicer to filter and view job listings by specific locations.

**3. Results and Analysis**

The dashboard created in Power BI provided valuable insights into the job market data sourced from the Naukri website. The key findings and visualizations are as follows:

* **Job Titles Count (Bar Chart)**

The bar chart visualization displayed the number of job listings for each job title. This allowed for an immediate understanding of which job roles were most in demand, helping to identify popular industries and job functions.

* **Job Title KPI**

The KPI visualization highlighted the total number of distinct job titles available on the platform. This provided a quick overview of the diversity and range of opportunities available, showcasing the breadth of roles listed across various sectors.

* **Job Posted Duration (Funnel Chart)**

The funnel chart visualized the distribution of job postings based on their duration. This highlighted how long jobs had been listed, allowing us to identify trends in job posting frequency. It also revealed the urgency of job vacancies, with newer postings being more prominently featured.

* **Maximum Salary (Pie Chart)**

The pie chart visualization provided insights into the maximum salary offered across different job titles. It allowed for a clear comparison of salary ranges, helping users understand which job titles offer the highest compensation and assess market trends.

* **Location Slicer**

The slicer allowed users to filter and view job listings by specific locations. This feature gave insights into geographical trends, highlighting which regions or cities have the highest number of job listings. It also helped to identify regional job market disparities.

**4. Conclusion**

This capstone project successfully demonstrated how data scraping, cleaning, SQL integration, and visualization tools like Power BI can be applied to extract meaningful insights from web data. The process from scraping data from Naukri.com to transforming it into an interactive dashboard in Power BI showcases the power of data science and analytics in the job recruitment sector.

**Future Work**

* **Real-Time Data Scraping:** Set up automated daily scraping jobs to fetch new job listings and refresh the Power BI dashboard.
* **Advanced Analytics:** Implement machine learning models to predict job trends or recommend jobs based on a user’s profile.