ALY 6080:
INTEGRATED EXPERIENTIAL LEARNING

Assignment 3: Group Project- CoverQuick Datasets EDA and Visualization

Submitted To:
Dr. Chinthaka Pathum Dinesh, PhD, Prof. Herath Gedara,
Faculty Lecturer

Submitted By:
Abhilash Dikshit
Kush Patel
Siddharth Alashi

Academic Term: Spring 2023
Graduate Students at Northeastern University, Vancouver, BC, Canada
Master of Professional Studies in Analytics

May 06, 2023

I. Abstract:

Data preparation is an essential step in building a database that can provide useful insights. In this report, we discuss the process of preparing data for a job applicant database. The data were obtained from a job search website and contained information about job seekers' education, work experience, and skills. We used Python programming language and Pandas library to clean, explore, and transform the data. The tasks completed include removing duplicate and null values, standardizing column names, and converting data types. Further tasks include analyzing the data and creating visualizations to understand patterns and trends. We plan to use Jupyter Notebook and Tableau for data analysis and visualization. Finally, we will communicate the results to stakeholders through reports and presentations.

II. Keywords:

Data preparation, data cleaning, data transformation, data analysis, data visualization, job applicant database

III. Introduction:

In today's competitive job market, employers are looking for ways to identify the most qualified candidates for job openings. One way to accomplish this is to create a database of job applicants, which can provide insights into their education, work experience, and skills. However, before building such a database, it is necessary to prepare the data. Data preparation involves cleaning, transforming, and exploring the data to ensure it is accurate, consistent, and meaningful.

IV. Methods:

We obtained the data from our Industry sponsor “CoverQuick”. We used Python programming language and other libraries to clean, explore, and transform the data. The tasks completed include replacing/removing duplicate and null values, standardizing column names, and converting data types. We also created a new column that extracted country names into one field to make it easier to analyze the data by country code.

Display Type, Length, Shape about the dataset:

Type	Length	Shape
<class 'pandas.core.frame.DataFrame'>	13198	(13198, 2)

Display datatypes of respective columns in dataset:

	id	content
0	object	object

Showing max, min length and NA values:

Column	Max Length	Min Length	NA Count
id	25	25	0
content	30743	484	0

How are you addressing data preparation?

We will focus on analyzing the data, identifying patterns and trends, and generating insights based on Country code, Candidate id, geographical location, age, gender, job description and statistics of the applicants applying for the specific job role.

What tasks did you complete?

We have listed specific tasks that has been completed till now. Common tasks in data preparation include removing duplicates, handling missing data, standardizing data formats, and transforming data, including additional columns, segregating, and splitting columns with json values into human-readable format using keys and values which will help us for data visualization.

What tasks are left to be done?

Some common tasks in data analysis that needs to be done include performing statistical tests, visualizing data, and identifying patterns and trends for age and gender applying for specific job role, job description and statistics of the applicants applying based on the skillset and job description and few others where we need to fetch keys and values from the job description, education, experience, certification columns.

What is your plan to complete these tasks?

Once the data cleaning and preprocessing steps is complete, we will use various statistical and visualization tools to analyze the data, identify patterns and trends, and generate insights for future growth and accommodate the requirements

gathered by the dataset to give top in class customer service experience to our users and use the analysis for company's growth.

What are the preferred methods of communicating the results from your initial EDA?

The preferred methods of communicating the results from the initial EDA depends on the nature of the data, the research question, and the target audience. Common methods that we included is data visualizations, descriptive statistics, and reports using multiple libraries available in python.

How do you plan to communicate results of tasks yet to be complete?

I cannot predict the results of tasks that have not been completed. However, once the analysis is complete, the results can be communicated using various methods, including reports, visualizations, and presentations and might use Tableau for the same.

V. Conclusion:

Data preparation is an essential step in building a job applicant database that can provide useful insights. We used Python programming language and Pandas library to clean, transform, and explore the data. The next step is to analyze the data and create visualizations using Jupyter Notebook and Tableau. Finally, we will communicate the results to stakeholders through reports and presentations. By completing these tasks, we will have a comprehensive database of job applicants that can help employers make informed hiring decisions.

VI. References:

People analytics. (n.d.). Deloitte Insights. https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2016/people-analytics-in-hr-analytics-teams.html

Reimagining HR: Insights from people leaders. (2022, March 1). McKinsey & Company. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/reimagining-hr-insights-from-people-leaders

Schoemaker, P. J. H. (1995). Scenario planning: A tool for strategic thinking. Sloan Management Review, 36(2), 25–40. https://sloanreview.mit.edu/article/scenario-planning-a-tool-for-strategic-thinking/