INFORMATION SYSTEMS

**A logo with text on it

Description automatically generated**

**Course: Business Data Analytics BIS - 581**

**Prof:** Vishal Shah

**Final Project Report: Cleaning, Visualization, Modeling and Analysis of a Dataset using RStudio**

**Professor: Vishal Shah**

**Author: Harinath Bhagya Sree Tiruvallur**

**E-Mail: tiruv1h@cmich.edu**

**Phone: (989)506-3993**

**CMICH ID#: 894338**

**Agenda**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Topic** | **Page No.** |
|  | Introduction | 1 |
|  | Dataset Description | 1-2 |
|  | Data Identification & Data Collection | 2 |
|  | Cleaning the data | 2-3 |
|  | Data Visualization | 3-5 |
|  | Data Modelling | 5-6 |
|  | Conclusion | 6 |
|  | Reference | 7 |

**Introduction:**

The dataset Naukri is taken from Kaggle website which provides a detailed job listing, encompassing a wide variety of roles across different industries.It explores India's leading job portal, to uncover current job market trends, employer ratings, and the skills most in demand across various industries. Through this investigation, we aim to provide actionable insights for job seekers, employers, and policymakers to better understand the dynamics of the Indian job market.

**Objective:**

* To evaluate employer reputations through available ratings and reviews.

**Dataset Description:**

15,217 job postings total from the Naukri dataset, which has been organized into 19 different columns in an organized manner. Every item in the collection offers comprehensive details about specific job advertisements. Below is an analysis of some of the dataset's most important columns:

* **Title**: The job title, which succinctly describes the nature of the job.
* **Title href**: A URL link to the detailed job listing on Naukri.com.
* **SubTitle**: The name of the company or employer posting the job.
* **SubTitle href**: A URL link that typically directs to the company’s profile or additional listings from the same employer.
* **StarRating**: Numerical employer ratings provided by users, indicative of employer reputation.
* **ReviewsCount**: The number of reviews the company has received, complementing the star ratings.
* **Skills and Tags**: Various columns like **fleft**, **fleft 2**, etc., contain categorized data that may represent required skills, job roles, or technological expertise needed for the job.

**Data Characteristics:**

* **Categorical Data**: Most of the dataset is categorical, including job titles, company names, and skills/tags which classify the jobs into various categories.
* **Numerical Data**: Includes columns like StarRating, which are crucial for quantitative analysis of employer reputation.
* **Missing Values**: Some columns, particularly those related to ratings and reviews, contain missing data, indicating not all listings are rated by users.

**Importance of the Analysis:**

Company analysis is done in the dataset to explore the ratings and reviews of companies to understand their reputation and popularity among job seekers. This dataset not only helps in understanding what types of jobs are currently available but also assists in analysing trends over time, such as changes in employer ratings.

This analysis is pivotal for job seekers who wish to align their career paths with market demands and for employers who aim to attract top talent by understanding the competitive landscape and improving their visibility and reputation in the market.

In the following sections, we will dive deeper into how the data was collected, cleaned, and prepared for this comprehensive analysis, ensuring the reliability and accuracy of our insights.

**Data Identification:**

The Naukri dataset's comprehensive and up-to-date listings of job openings across various sectors in India made it an important source for employment trend analysis. The dataset is ideal for investigating the dynamics of the employment sector in an in-depth manner due to its wide range of job titles, company information, and related skills.

**Source of the Data:**

The source of the data is Naukri.com, one of India's biggest employment websites with an extensive collection of job postings and employer details. The platform is a useful tool for real-time employment market analysis as it acts as a hub to link companies and potential employees.

**Data Collection:**

* **Extraction Technique**: The dataset was compiled using a method where data provided directly by Naukri.com under a partnership of kaggle. The availability and accessibility of data in an organized manner that is suitable for analysis would determine the precise technique of data collection.
* **Data Format**: The dataset was obtained in CSV format, facilitating easy import into data analysis tools like R studio.

The process involved in gathering the Naukri dataset was meticulously planned to ensure the data was representative of the current job market. This phase provided the foundation for a thorough analysis that aimed to extract accurate and significant insights from the data.

In the next section, we will delve into the data cleaning process, describing the steps taken to prepare the data for thorough and effective analysis, ensuring that the insights generated are based on accurate and reliable data.

**Data Cleaning:**

Data cleaning is a crucial step in ensuring the reliability and validity of the analysis. For the Naukri dataset, to correct discrepancies, missing data, duplicate entries, and additional information that can skew the results or provide erroneous insights, several methodical processes were applied in this step.

1. **Initial Data Assessment:** The first thing to do after importing the dataset into R was to do an exploratory data analysis (EDA) to evaluate the general quality and organization of the data:

* Basic summary statistics and visualizations allowed for the identification of distributions, trends, anomalies, and outliers within the dataset.
* Missing values, erroneous data types, and abnormalities in several columns, especially those about company ratings and skills, were discovered during the initial examination.

1. **Handling Missing Values:** Missing data can lead to biased estimates and less generalizable results. Different strategies were employed depending on the nature of the data:

* **Deletion**: Rows with missing values in critical fields, such as job titles or company names, were removed when such absences represented a small portion of the dataset.
* **Imputation**: For numerical fields like ‘CompanyRating’, where missing values were more prevalent, imputation techniques such as the mean or median of the available data were used to fill in gaps.

1. **Correcting Data Formats:** Ensuring that data types are correctly assigned enhances the accuracy of the analysis:

* **Converting Data Types**: Fields such as ‘CompanyRating’ were converted from character to numeric types to perform statistical analysis.
* **Date Formatting**: Any date fields were standardized to ensure consistency across the dataset.

1. **Removing Duplicates and Irrelevant Data**

* **Removing duplicates**: Duplicate job listings were identified by key identifiers like job titles and company names and removed to prevent data redundancy.
* **Filtering Irrelevant Data**: Entries not relevant to the analysis objectives, such as title href

1. **Renaming and Reorganizing Column**: To improve the readability and accessibility of the data:

* **Renaming Columns**: Columns are renamed for clarity and ease of use in R.
* **Reorganizing Data Structure**: Columns were reordered to group similar types of information together, enhancing the workflow in subsequent analyses.

The dataset was strengthened by the data cleaning procedure, making it suitable for in-depth examination. Greater accuracy and confidence may be gained in the later phases of modelling and visualization with a clean and well-structured dataset. The next phase will look at several ways to visualize the cleaned data to get some preliminary insights and get ready for more sophisticated analysis methods.

**Data Visualization:**

Data visualization is a powerful tool to uncover patterns, trends, and anomalies in data. In the analysis of the Naukri dataset, various visualization techniques were employed to provide a graphical representation of the data, making complex data more accessible and understandable using R's powerful ggplot2 package. With a special emphasis on business ratings, reviews, and their consequences, these visualizations seek to identify and draw attention to important patterns, connections, and distributions within the dataset.

1. **Histogram of Company Ratings Distribution:** The first visualization is a histogram that illustrates the frequency distribution of company ratings:

* **Purpose**: To understand how company ratings are distributed across the dataset and identify the commonality of different rating levels.
* **Visualization Details**: The histogram bins company ratings from 0 to 5, using a bin width of 0.5 to clearly delineate rating categories. The aesthetic enhancements include a minimal theme and clearly formatted axes, enhancing readability and focus on the data.

1. **Bar Chart of Top 10 Companies by Average Rating:** Next, I have calculated the average rating for each company and display the top 10 companies with the highest average ratings:

* **Purpose**: To highlight companies that consistently receive high ratings, potentially indicating superior employer reputation or employee satisfaction.
* **Visualization Details**: This bar chart reorders companies based on their average ratings, ensuring that the highest-rated companies are prominently displayed. The chart uses interactive elements from the **plotly** package, making it dynamic and user-friendly.

1. **Scatter Plot of Total Reviews vs. Company Rating:** A scatter plot explores the relationship between the total number of reviews a company has received and its average rating:

* **Purpose**: To investigate if there is any correlation between the volume of reviews and the ratings companies receive, which might suggest that more frequently reviewed companies have better or worse ratings.
* **Visualization Details**: Each point on the scatter plot represents a company, with axes showing ratings and total review counts. This visualization helps identify outliers and trends in how customers rate companies based on their experiences.

1. **Top 10 Companies by Total Reviews:** Following the scatter plot, I have listed the top 10 companies based on the total number of reviews received.

* **Purpose**: To identify which companies are the most reviewed, suggesting high engagement or interest from job seekers or employees.
* **Visualization Details**: A column chart displays these companies, using color coding to differentiate them and angled text to enhance clarity and legibility of company names.

1. **Boxplot of Company Ratings by Review Groups:** Finally, I have categorized companies into groups based on quartiles of review counts and compare these groups' company ratings through a boxplot:

* **Purpose**: To examine if there is a pattern in ratings across different levels of review frequency, which might indicate how the quantity of feedback correlates with perceived company quality.
* **Visualization Details**: The boxplot categorizes companies into "Very Low," "Low," "Medium," and "High" review groups. It uses different colours for each category, providing a clear visual distinction that helps in quickly assessing the median ratings and the spread of ratings within each group.

These visualizations collectively provide a comprehensive overview of key aspects of the job market captured in the Naukri dataset. By analysing company ratings and review counts, I can infer the visibility and reputation of companies within the job market.

This method helps companies assess their position in the market and helps job seekers make well-informed judgments. Stakeholders may better plan their strategies to increase or sustain their market presence with the help of this information.

**Data Modelling:**

In this phase of the analysis, we apply statistical and machine learning models to predict company ratings using the Naukri dataset. Specifically, we utilize decision trees and linear regression models, each offering unique insights into the dataset's dynamics.

**Decision Tree:**

They are a type of supervised learning algorithm that is used for both classification and regression tasks. They work by splitting the data into branches to form a tree structure. Each internal node of the tree corresponds to a decision rule, and each leaf node represents the outcome.

In the context of predicting company ratings, the decision tree will help identify the combinations of features that lead to different rating levels, providing clear and interpretable results. A decision tree model is constructed using the rpart package, which handles both numerical and categorical inputs effectively.

* **Objective**: To predict company ratings based on diverse predictors from the dataset.
* **Data Preparation**: The dataset is divided into an 80% training set and a 20% testing set.
* The model's summary reveals which factors most significantly impact company ratings.
* I have evaluated the model's accuracy through predictions on the testing set, using the Root Mean Square Error (RMSE) as a metric.

**Linear Regression:**

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The coefficients of the equation are estimated to minimize the difference between the predicted and actual data points. In our analysis, linear regression will quantify how changes in predictors like job role, company size, and industry impact the company rating.

* **Objective**: To quantify the linear relationships between company ratings and predictors.
* **Model Building and Training**: Fits a linear regression model to the training data using all available predictors.
* The model’s summary provides coefficients that indicate how much each predictor is expected to increase or decrease the company rating.
* Evaluation of this model is assessed using RMSE on the testing set, comparing it to the decision tree's performance.

**Visualization of Model Results**

* **Residual Plots and Prediction Accuracy**: For the linear regression model, residual plots can diagnose model fit and visualize prediction accuracy.
* **Actual vs. Predicted Ratings**: Charts comparing actual versus predicted ratings illustrate the effectiveness of both modelling approaches and pinpoint areas for improvement.

This comprehensive modelling stage sheds light on the factors influencing company ratings, utilizing the strengths of both decision trees for their interpretability and linear regression for quantifying predictor impacts. The models provide stakeholders with actionable insights to make informed decisions, guiding future data collection and analysis efforts to enhance the utility of employment market analytics.

**Conclusion:**

In this comprehensive analysis of the Naukri dataset, that navigated through data collection, cleaning, visualization, and predictive modelling to unearth insights into India's job market.

This project sets a foundation for continuous improvement and adaptation in job market analysis, providing actionable insights to help stakeholders make informed decisions in a dynamic employment landscape.

**References:**

1. Introduction to R and RStudio – Introduction to R/tidyverse for Exploratory Data Analysis. (n.d.). Data Carpentry - Introduction to R/Tidyverse for Exploratory Data Analysis. <https://tavareshugo.github.io/r-intro-tidyverse-gapminder/00-starting_your_project/index.html>
2. Grolemund, H. W. M. E. R. A. G. (n.d.). *R for Data Science (2e) - 1  Data visualization*. <https://r4ds.hadley.nz/data-visualize>
3. Serrano, L. (2022, September 30). Splitting Data by Asking Questions: decision trees - Luis Serrano - Medium. Medium. <https://medium.com/@luis.serrano/splitting-data-by-asking-questions-decision-trees-74afed9cd849>