PROJECT REPORT

ON

HR DATA ANALYSIS

BY

SANTHOSH S

1. **Introduction**

**Background Introduction**

Human Resource (HR) departments play a vital role in managing the most important asset of any organization: its people. The data collected by HR departments can provide valuable insights into employee performance, satisfaction, and overall organizational health. However, raw HR data is often messy and unstructured, containing unnecessary columns, redundant entries, inconsistent data formats, and missing values. This unclean data can lead to inaccurate analyses and poor decision-making.

Data analysis in HR involves scrutinizing various aspects of employee data, including recruitment, training, performance, compensation, and attrition rates. By systematically cleaning and preparing this data, organizations can uncover patterns and trends that help optimize HR strategies, enhance employee engagement, and improve operational efficiency.

**Aim**

The aim of this project is to undertake a comprehensive data cleansing and preparation process on an HR dataset using MySQL. The primary goal is to transform the raw, unstructured data into a clean, structured format that is ready for detailed analysis and reporting. This will involve a series of steps including the removal of unnecessary columns, renaming of columns for clarity, elimination of redundant entries, sanitization of specific columns, and handling of missing values.

**Purpose of the System**

The purpose of this system is multifaceted:

* Accuracy and Reliability: Ensure that the HR data is accurate and reliable, which is crucial for any subsequent analyses or reporting.
* Improved Decision-Making: Enable HR managers and decision-makers to derive meaningful insights from the data, leading to informed decisions regarding recruitment, training, and retention.
* Efficiency: Streamline the data preparation process, making it easier to update and maintain the dataset over time.
* Compliance: Ensure that the data adheres to regulatory and compliance standards, particularly in terms of data privacy and security.
* Usability: Enhance the usability of the data by organizing it in a clear and understandable format, making it accessible for various stakeholders within the organization.

**2. Analysis**

**Hardware and Software Requirements**

* **Hardware System Configuration**
  + **processor - Pentium – IV**
  + **RAM - 4 GB (min)**
  + **Hard Disk - 20 GB**
* **Software System Configuration**
  + **Operating System : Windows 7 or 8**
  + **Software : MySQL Workbench**

**Feasibility Study**

* **Technical Feasibility**

The project is technically feasible as MySQL provides robust tools for data manipulation and cleansing. MySQL Workbench offers an intuitive interface for executing SQL queries and managing databases.

* **Economic Feasibility**

The cost involved is minimal since MySQL is an open-source database management system. The primary investment would be in hardware, which is a standard requirement for most IT departments.

* **Operational Feasibility**

The operational feasibility is high as the project involves standard data cleansing practices which are well-understood and can be executed efficiently with SQL commands.

**3. Design**

**Block Diagram**

**The block diagram below illustrates the flow of the data cleansing process:**

**+---------------------+**

**| Data Import |**

**| (Load Raw HR Data) |**

**+----------+----------+**

**|**

**v**

**+---------------------+**

**| Remove Unnecessary |**

**| Columns |**

**+----------+----------+**

**|**

**v**

**+---------------------+**

**| Rename Columns |**

**| (to Meaningful Names) |**

**+----------+----------+**

**|**

**v**

**+---------------------+**

**| Eliminate Redundant |**

**| Entries |**

**+----------+----------+**

**|**

**v**

**+---------------------+**

**| Sanitize Specific |**

**| Columns |**

**+----------+----------+**

**|**

**v**

**+---------------------+**

**| Handle NaN Values |**

**+---------------------+**

**4. Methodology**

The following steps outline the methodology used for the data cleansing process:

1. **Data Import:** Load the raw HR dataset into a MySQL database.
2. **Remove Unnecessary Columns:** Identify and remove columns that are not needed for analysis.
3. **Rename Columns:** Change column names to more meaningful and user-friendly names.
4. **Eliminate Redundant Entries:** Identify and remove duplicate records to ensure each entry is unique.
5. **Sanitize Specific Columns:** Cleanse specific columns to standardize data formats, remove invalid characters, and correct data entries.
6. **Handle NaN Values:** Identify and handle missing values by either filling them with appropriate data or removing the rows/columns.
7. Convert categorical data to numerical data

**5. CODING AND IMPLEMENTATION**

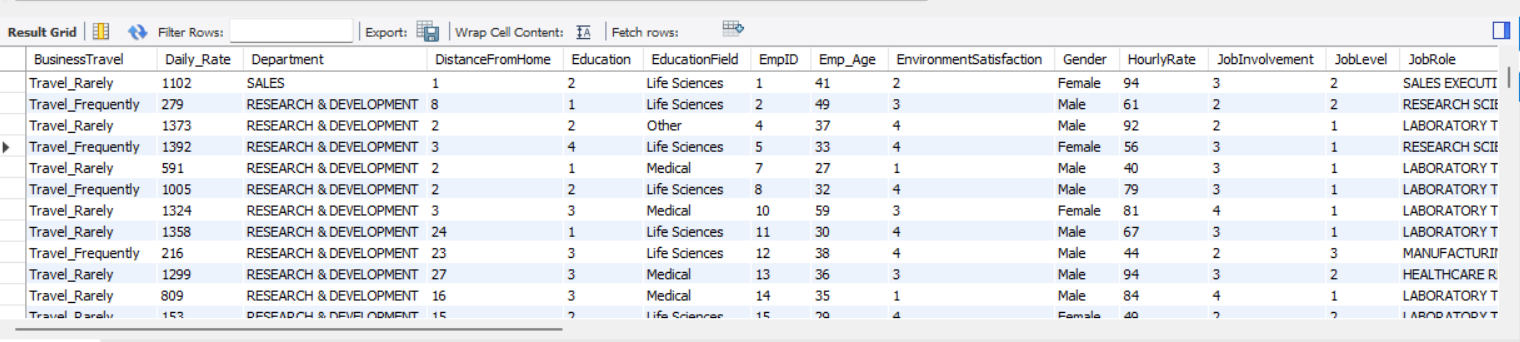
**1. Create Database**

create database hr;

use hr;

**2. Import .CSV file**

select \* from hr\_data;



**3. Remove Unnecessary Columns**

ALTER TABLE hr\_data

DROP COLUMN EmployeeCount,

DROP COLUMN Over18,

DROP COLUMN StandardHours;

**4. Rename Columns**

ALTER TABLE hr\_data

CHANGE COLUMN EmployeeNumber EmpID INT,

CHANGE COLUMN DailyRate Daily\_Rate INT,

CHANGE COLUMN ï»¿Age Emp\_Age INT,

CHANGE COLUMN MonthlyIncome Monthly\_Rate bigint,

CHANGE COLUMN MonthlyRate Monthly\_Income bigint;

SELECT EmpID, COUNT(\*)

FROM hr\_data

GROUP BY EmpID

HAVING COUNT(\*) > 1;

**5. Eliminate Redundant Entries**

CREATE TEMPORARY TABLE TempTable AS

SELECT DISTINCT empid, dept

FROM test;

select \* from temptable;

truncate table test;

INSERT INTO test (empid, dept)

SELECT empid, dept

FROM TempTable;

drop temporary table TempTable;

**6. Sanitize Specific Columns**

UPDATE hr\_data

SET JobRole = TRIM(UPPER(JobRole)),

Department = TRIM(UPPER(Department));

**7. Handle NaN Values**

DELETE FROM hr\_data

WHERE Emp\_Age IS NULL

OR Attrition IS NULL

OR BusinessTravel IS NULL

OR Daily\_Rate IS NULL

OR Department IS NULL

OR DistanceFromHome IS NULL

OR Education IS NULL

OR EducationField IS NULL

OR EmpID IS NULL

OR EnvironmentSatisfaction IS NULL

OR Gender IS NULL

OR HourlyRate IS NULL

OR JobInvolvement IS NULL

OR JobLevel IS NULL

OR JobRole IS NULL

OR JobSatisfaction IS NULL

OR MaritalStatus IS NULL

OR Monthly\_Rate IS NULL

OR Monthly\_Rate IS NULL

OR NumCompaniesWorked IS NULL

OR OverTime IS NULL

OR PercentSalaryHike IS NULL

OR PerformanceRating IS NULL

OR RelationshipSatisfaction IS NULL

OR StockOptionLevel IS NULL

OR TotalWorkingYears IS NULL

OR TrainingTimesLastYear IS NULL

OR WorkLifeBalance IS NULL

OR YearsAtCompany IS NULL

OR YearsInCurrentRole IS NULL

OR YearsSinceLastPromotion IS NULL

OR YearsWithCurrManager IS NULL;

**8. Additional Changes:**

* Convert categorical data to numerical data.

UPDATE hr\_data

SET Attrition\_Num = CASE

WHEN Attrition = 'Yes' THEN 1

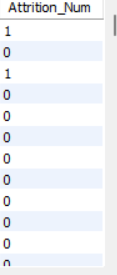
WHEN Attrition = 'No' THEN 0

ELSE NULL

END;

ALTER TABLE hr\_data

DROP COLUMN Attrition;



**5. Discussion**

**Results and Observations**

The data cleansing process was executed successfully, resulting in a clean and reliable HR dataset. Unnecessary columns were removed, columns were renamed for better readability, redundant entries were eliminated, specific columns were sanitized, and missing values were handled appropriately.

**Challenges and Solutions**

* **Challenge:** Handling a large volume of data efficiently.
  + **Solution:** Used MySQL indexing and optimized queries to improve performance.
* **Challenge:** Ensuring data accuracy during the sanitization process.
  + **Solution:** Implemented validation checks and cross-referenced data entries.

**Future Enhancements**

* Automating the data cleansing process using stored procedures.
* Implementing advanced data validation rules.
* Integrating with data visualization tools for better insights.