Salary Data Analysis Documentation (2021)

Project: Milestone Project - 1
Author: CS Sundhar
Course: DADS March-25
Year: 2021

Table of Contents

Project Overview

This project focuses on comprehensive salary data analysis for the year 2021, encompassing data cleaning, database implementation, statistical analysis, and visualization. The analysis covers multiple dimensions including industry sectors, geographic locations, education levels, gender distribution, and experience levels to provide actionable insights for workforce planning and compensation strategies.

Objectives

Primary Objectives

- **Data Quality Assurance**: Remove duplicates and detect outliers
- **Data Standardization**: Clean and format data for consistency
- **Data Categorization**: Organize columns into meaningful categories
- **Database Integration**: Convert CSV data to MySQL database
- **Query Execution**: Perform comprehensive data analysis through SQL
- **Data Export**: Export processed tables for further analysis
- **Visualization**: Create pivot tables and charts for insights
- **Dashboard Development**: Build interactive dashboard for stakeholders
- **User Experience**: Ensure seamless workflow and engagement
- **Interactive Analysis**: Implement slicer integration for dynamic filtering

---

Dataset Description

The dataset contains 19 key variables covering demographic, professional, and compensation information:

Demographic Variables

- **Age Range**: Grouped age brackets of employees
- **Gender**: Gender identity classification (Male, Female, Others)
- **Country**: Nation of employment
- **Region**: Larger geographical area within country
- **State**: Sub-regional or state-level location
- **City**: Specific city of employment

Professional Variables

- **Industry**: General work sector (Technology, Finance, etc.)
- **Category of Industry**: Specific industry subcategory
- **Job Title**: Position or role designation
- **Years of Professional Experience Overall**: Total career experience
- **Overall Level**: General experience classification (Junior, Mid, Senior)
- **Years of Professional Experience in Field**: Domain-specific experience
- **Field Level**: Skill level within current field specialty

Education and Compensation

- **Highest Level of Education Completed**: Educational qualification
- **Annual Salary**: Base yearly compensation
- **Additional Monetary Compensation**: Bonuses and extra income
- **Currency**: Monetary unit used for compensation
- **Final Salary**: Total compensation (salary + additional compensation)
- **Salary Category**: Grouped salary ranges (50k-1L, 1L-5L, etc.)

Data Processing Procedures

Phase 1: Data Import and Initial Assessment

1.1 File Import Procedure

1. **Source File Verification**
    - Verify CSV file format and encoding (UTF-8 recommended)
    - Check file size and estimated record count
    - Validate column headers and data types
2. **Initial Data Loading**
    - Import CSV file into Excel/data processing tool
    - Perform preliminary data scan for obvious errors
    - Document original dataset dimensions and characteristics

1.2 Data Quality Assessment

1. **Completeness Check**
    - Identify missing values in each column
    - Calculate completion rates for critical fields
    - Flag columns with excessive missing data
2. **Data Type Validation**
    - Verify numeric fields contain valid numbers
    - Check date fields for proper formatting
    - Validate categorical fields for consistency

Phase 2: Data Cleaning and Transformation

2.1 Duplicate Removal Procedure

1. **Duplicate Identification**
    - Define duplicate criteria (exact match vs. fuzzy matching)
    - Use conditional formatting to highlight potential duplicates
    - Cross-reference key fields (name, email, employee ID if available)
2. **Duplicate Resolution**
    - Review flagged duplicates manually for verification
    - Establish priority rules (most recent, most complete record)

- Remove confirmed duplicates and document removal count

2.2 Outlier Detection and Treatment

1. **Statistical Outlier Analysis**
    - Calculate quartiles and interquartile range (IQR) for salary fields
    - Identify values beyond 1.5 * IQR threshold
    - Create box plots to visualize outlier distribution
2. **Outlier Validation**
    - Cross-reference outliers with job titles and industries
    - Verify high salaries against executive or specialized roles
    - Flag unrealistic values for removal or correction
3. **Outlier Treatment**
    - Remove obvious data entry errors
    - Cap extreme values at reasonable thresholds
    - Document all outlier treatment decisions

2.3 Data Standardization

1. **Text Field Cleaning**
    - Standardize capitalization (proper case for names, titles)
    - Remove extra spaces and special characters
    - Harmonize industry and job title naming conventions
2. **Geographic Data Standardization**
    - Validate country names against standard ISO codes
    - Standardize state/region abbreviations
    - Correct common misspellings in city names
3. **Categorical Data Harmonization**
    - Group similar job titles under standard categories
    - Standardize education level classifications
    - Create consistent experience level groupings

Phase 3: Column Categorization and Enhancement

3.1 Derived Field Creation

1. **Salary Categorization**
    - Create salary range buckets (50k-100k, 100k-500k, etc.)
    - Calculate total compensation (base + additional)
    - Develop experience-to-salary ratios
2. **Geographic Grouping**
    - Group cities by metropolitan areas
    - Create regional classifications
    - Develop cost-of-living adjustments if needed

3.2 Data Validation and Quality Control

1. **Cross-Field Validation**
    - Verify salary ranges align with job titles and experience
    - Check geographic consistency (city-state-country alignment)
    - Validate education levels against professional roles
2. **Final Quality Check**
    - Run comprehensive data validation rules
    - Generate data quality report
    - Document all cleaning and transformation steps

Database Implementation

CSV to MySQL Migration Process

4.1 Database Setup

1. **Database Creation**
2. CREATE DATABASE salary_data;
3. USE salary_data;
4. **Table Structure Design**
5. CREATE TABLE salaries (
6. id INT AUTO_INCREMENT PRIMARY KEY,

```
    age_range VARCHAR(50),
    industry VARCHAR(100),
    industry_category VARCHAR(100),
    job_title VARCHAR(200),
    annual_salary DECIMAL(15,2),
    additional_compensation DECIMAL(15,2),
    currency VARCHAR(10),
    country VARCHAR(100),
    region VARCHAR(100),
    state VARCHAR(100),
    city VARCHAR(100),
    years_experience_overall INT,
    overall_level VARCHAR(50),
    years_experience_field INT,
    field_level VARCHAR(50),
    education_level VARCHAR(100),
    gender VARCHAR(20),
    final_salary DECIMAL(15,2),
    salary_category VARCHAR(50)
);
```

4.2 Data Import Procedure

1. **Import Wizard Configuration**
    - Use MySQL Table Import Wizard
    - Map CSV columns to database fields
    - Set appropriate data types and constraints
    - Configure character encoding (UTF-8)
2. **Import Validation**
    - Verify record count matches source file
    - Check for import errors or data truncation

- Validate sample records for accuracy

4.3 Post-Import Processing

1. **Index Creation**
2. CREATE INDEX idx_industry ON salaries(industry);
3. CREATE INDEX idx_job_title ON salaries(job_title);
4. CREATE INDEX idx_country ON salaries(country);
5. CREATE INDEX idx_salary ON salaries(final_salary);
6. **Data Integrity Checks**
    - Run constraint validation queries
    - Check for null values in critical fields
    - Verify referential integrity

SQL Queries and Analysis

Query 1: Average Salary by Industry and Gender

```
SELECT
    industry,
    gender,
    AVG(final_salary) AS avg_salary,
    COUNT(*) AS employee_count
FROM salaries
GROUP BY industry, gender
ORDER BY industry, avg_salary DESC;
```

Query 2: Total Compensation by Job Title

```
SELECT
    job_title,
    industry,
    SUM(IFNULL(annual_salary, 0) + IFNULL(additional_compensation, 0)) AS
total_compensation,
    AVG(IFNULL(annual_salary, 0) + IFNULL(additional_compensation, 0)) AS
avg_compensation
```

```
FROM salaries
GROUP BY job_title, industry
ORDER BY total_compensation DESC;
```

Query 3: Salary Range by Education Level

```
SELECT
    education_level,
    MIN(final_salary) AS min_salary,
    MAX(final_salary) AS max_salary,
    AVG(final_salary) AS avg_salary,
    COUNT(*) AS employee_count
FROM salaries
WHERE final_salary IS NOT NULL
GROUP BY education_level
ORDER BY avg_salary DESC;
```

Query 4: Employee Count by Industry and Experience

```
SELECT
    industry,
    overall_level,
    COUNT(*) AS employee_count,
    AVG(final_salary) AS avg_salary
FROM salaries
GROUP BY industry, overall_level
ORDER BY industry, employee_count DESC;
```

Query 5: Median Salary by Age and Gender

```
SELECT
    age_range,
    gender,
    SUBSTRING_INDEX(
        SUBSTRING_INDEX(
```

```
        GROUP_CONCAT(final_salary ORDER BY final_salary SEPARATOR ','),
        ',',
        FLOOR((COUNT(*) + 1) / 2)
    ),
    ',',
    -1
  ) AS median_salary
FROM salaries
WHERE final_salary IS NOT NULL
GROUP BY age_range, gender
ORDER BY age_range, gender;
```

Query 6: Highest Salary Job in Each Country

```
SELECT
  s1.country,
  s1.job_title,
  s1.final_salary,
  s1.industry
FROM salaries s1
INNER JOIN (
  SELECT
    country,
    MAX(final_salary) AS max_salary
  FROM salaries
  GROUP BY country
) s2 ON s1.country = s2.country AND s1.final_salary = s2.max_salary
ORDER BY s1.final_salary DESC;
```

Query 7: Average Salary by City and Industry

```
SELECT
  city,
```

```
    industry,
    AVG(final_salary) AS avg_salary,
    COUNT(*) AS employee_count
FROM salaries
GROUP BY city, industry
HAVING COUNT(*) >= 5
ORDER BY avg_salary DESC;
```

Query 8: Percentage of Employees with Extra Compensation by Gender

```
SELECT
    gender,
    COUNT(*) AS total_employees,
    SUM(CASE WHEN additional_compensation > 0 THEN 1 ELSE 0 END) AS employees_with_extra,
    ROUND(
        (SUM(CASE WHEN additional_compensation > 0 THEN 1 ELSE 0 END) * 100.0 / COUNT(*)), 2
    ) AS percentage_with_extra
FROM salaries
GROUP BY gender
ORDER BY percentage_with_extra DESC;
```

Query 9: Total Compensation by Job and Experience

```
SELECT
    job_title,
    overall_level,
    SUM(final_salary) AS total_compensation,
    AVG(final_salary) AS avg_compensation,
    COUNT(*) AS employee_count
FROM salaries
GROUP BY job_title, overall_level
ORDER BY total_compensation DESC;
```

Query 10: Average Salary by Industry, Gender, and Education

```
SELECT
    industry,
    gender,
    education_level,
    AVG(final_salary) AS avg_salary,
    COUNT(*) AS employee_count
FROM salaries
GROUP BY industry, gender, education_level
HAVING COUNT(*) >= 3
ORDER BY industry, gender, avg_salary DESC;
```

Visualization and Dashboard

Pivot Tables and Charts

Industry-wise Salary Insights

- **Visualization Type**: Stacked bar charts and pivot tables
- **Metrics**: Average and total compensation across industries
- **Breakdown**: Salary plus additional compensation analysis
- **Key Finding**: Technology and Business sectors show highest compensation

Education vs Salary Analysis

- **Visualization Type**: Box plots and scatter charts
- **Comparison**: Salary distribution by education levels
- **Insights**: Higher education correlates with increased compensation
- **Recommendation**: Investment in education yields salary returns

Workforce Distribution

- **Visualization Type**: Pie charts and horizontal bar charts
- **Analysis**: Employee count across different industries
- **Geographic Breakdown**: Distribution by country and state
- **Demographic Split**: Analysis by age groups and gender

Dashboard Components

Interactive Elements

- **Slicers**: Industry, Country, Gender, Education Level
- **Filters**: Salary range, Experience level, Age group
- **Dynamic Charts**: Real-time updates based on slicer selection
- **KPI Cards**: Total employees, Average salary, Top industry

User Experience Features

- **Seamless Navigation**: Intuitive layout with clear sections
- **Responsive Design**: Adapts to different screen sizes
- **Export Functionality**: Download charts and data tables
- **Drill-down Capability**: Click-through from summary to detail

---

Key Insights

Workforce Demographics

- **Dominant Age Group**: 25-34 years (early-to-mid career professionals)
- **Experience Level**: Majority hold Associate-level positions
- **Gender Distribution**: Higher proportion of female employees
- **Education**: College degree as minimum qualification standard

Compensation Patterns

- **Top Contributors**: Associates contribute highest total salary sums
- **Highest Paying Role**: Software Engineers across all job titles
- **Gender Pay**: Analysis reveals compensation gaps and advancement patterns
- **Geographic Premium**: United States, particularly California, shows highest pay

Industry Trends

- **Leading Sectors**: Technology and Business show superior growth
- **Emerging Fields**: Business Analytics and Data Science in high demand
- **Traditional Industries**: Logistics, FMCG, and Farming show lower growth
- **Career Progression**: Strong advancement opportunities in tech sector

Geographic Insights

- **Top Country**: United States leads in employee compensation
- **Premium Locations**: California shows highest salaries and education levels
- **Regional Variations**: Significant compensation differences by location
- **Market Opportunities**: U.S. market presents expansion potential

---

Recommendations

Talent Development

1. **Associate-Level Focus**
    - Implement comprehensive upskilling programs
    - Create leadership development tracks
    - Establish mentorship programs for career progression
2. **Female Career Advancement**
    - Develop gender-inclusive leadership policies
    - Create women's advancement initiatives
    - Address pay equity gaps systematically

Industry Investment

1. **Technology and Business Focus**
    - Prioritize training in Business Analytics and Data Science
    - Invest in Software Development capabilities
    - Align skill development with market demand
2. **Geographic Strategy**
    - Focus expansion efforts on U.S. market
    - Target California for high-skill talent acquisition
    - Develop location-based compensation strategies

Workforce Planning

1. **Early-Career Talent**
    - Target 25-34 age demographic for recruitment
    - Develop comprehensive career path planning
    - Create learning and development initiatives

2. **Education and Skills**
    - Promote continuous learning programs
    - Align training with industry requirements
    - Support advanced degree attainment

Conclusion

Core Workforce Profile

- **Primary Demographics**: 25-34 years, Associate-level professionals
- **Gender Representation**: Strong female presence with growth potential
- **Geographic Concentration**: United States, particularly California
- **Education Foundation**: College degree as baseline requirement

High-Opportunity Sectors

- **Technology**: Highest salary potential and growth opportunities
- **Business**: Strong compensation and advancement prospects
- **Emerging Fields**: Data Science and Analytics showing rapid growth

Strategic Priorities

1. **Talent Investment**: Focus on Associate-level development and upskilling
2. **Diversity and Inclusion**: Support diverse talent and gender-inclusive policies
3. **Market Alignment**: Align training and development with Technology and Business demands
4. **Geographic Focus**: Leverage U.S. market opportunities, especially California

Success Metrics

- **Retention**: Improved Associate-level employee retention
- **Advancement**: Increased female representation in leadership
- **Skills**: Enhanced Technology and Business competencies
- **Compensation**: Competitive positioning in key markets

This comprehensive analysis provides a foundation for strategic workforce planning, compensation strategy development, and talent management initiatives based on data-driven insights from the 2021 salary analysis.