MGMT 6570 ADVANCED DATA RESOURCE MANAGEMENT FINAL PROJECT

Employment Database & Salary Analysis

Abstract

The report includes the project introduction, the data processing, the basic data statistic, and the salary analysis with few focused topics which will be discussed in the last section. The code and sheets will be attached with this report for your further review.

Kaiyue Zeng

RIN: 662013556

1. Project and Data Description

1.1 Introduction

The project works on the data selected from Thiri Yadana's GitHub dataset sample for data scientists, and the original source contains three separate txt. files. These files record the salary data of 1,000 employees in North America, which include their basic contact information, company information, job titles, and the location information. In this project, the salary analysis will be conducted with the database '*Employment*' combining these three data files, where primary key and foreign keys will be defined in the following sections.

1.2 Tables and Columns

1.2.1 Table *company_divisions*

This table records the relationship of employees' working departments and divisions.

- *department:* Employees are in different departments depending on their responsible works, and it also related to the products and service lines provided by their companies.
- *company_division:* There are 8 different divisions in this dataset, and the divisions refer to the industries and markets where their companies focus on.

1.2.2 Table *company_region*

This table records the locations of employees' companies, and a primary key *region_id* is also included in this table, which will be used in the table merging procedure.

- *region_id:* This is the primary key in this table.
- *company_regions:* The U.S. is divided into 4 regions and Canada is divided into 3 regions.
- *country:* Two countries are included in this dataset, which are the U.S. and Canada.

1.2.3 Table *staff*

This table has 9 columns recording the employees' basic information.

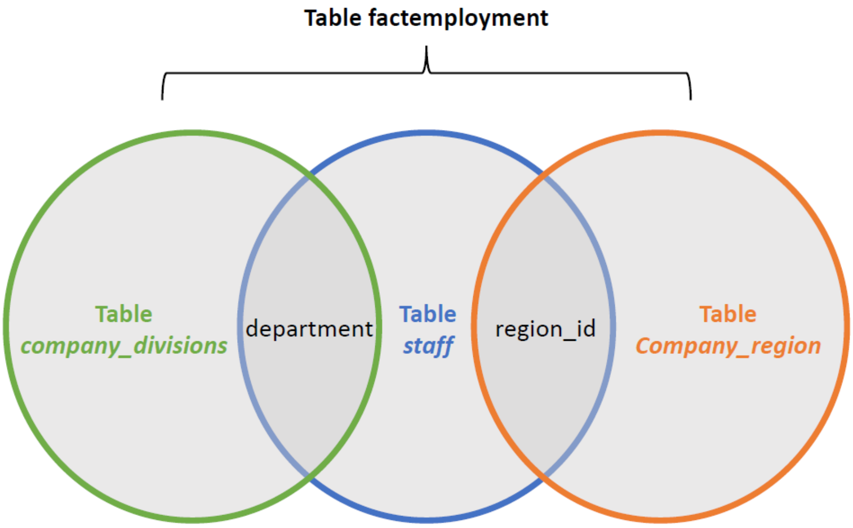
- *id:* The primary key in this table, each employee has a unique ID number.
- *last_name:* The name of the employee.
- *email:* The email address of the employee.
- *gender:* The gender of the employee.
- *department:* The meaning of this column is the same as the one in the table *company_dicisions*.
- *start_date:* The start date of the specific employment, and it can indicate how long an employee has been in the firm.
- *salary:* The annual salary of the employee.
- *job_title:* The recent job title of the employee.
- *region_id:* The meaning of this column is the same as the one in the table *company_region* which indicate the location of employees' companies, but it is a foreign key in this table.

2. Data Processing

2.1 Join Tables

In order to create a single dataset to easily analyze all of the data from separate three tables discussed above, the full joint method is utilized here with the *region_id* and the *department* as references. As a result, the table *factemployment* is established containing all columns from the table *staff*, the division information from the table *company_divisions*, and the company location information form the table company_region. Additionally, the primary key of this newly created table is the employee ID referring to the *id* column in the table *staff*.

Additional note: the book department does not have a corresponding division; thus, the division data for the employees who work in book department appears to be Null.



2.2 Dimension Tables

When considering to convert the characteristics into standard numerical indexes, the dimension tables are created with following thoughts.

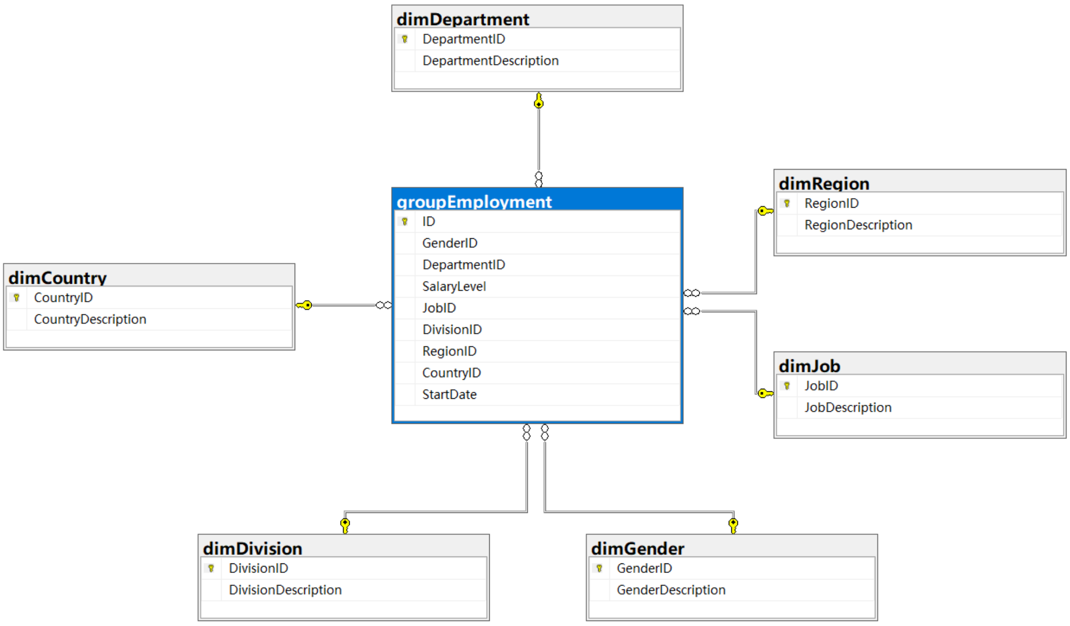
- The employee's name and email information are excluded during this process as they are not as essential as other features when applying a salary analysis.
- Dimension tables are created as *dimGender*, *dimDepartment*, *dimDivision*, *dimJob*, *dimRegion*, and *dimCountry*.
- The ID number for each feature is created as the primary key in each dimension table.

2.3 ID Table

After building the dimension tables, an ID table named *groupEmployment* is created in which the primary key of a dimension table become the foreign key. To be specific, the diagram can be shown as below. Furthermore, besides general data insert for *GenderID*, *DepartmentID*, *DivisionID*, *RegionID*, and *CountryID* with corresponding dimension tables, the salary level and the job ID are standardized using the following ideas.

- *SalaryLevel:* To compare the personal salary with the average salary, the index for this feature is defined as 1 when the specific employee's salary does not reach to the average amount and 2 when it is higher than the average amount.
- *JobID:* As this project wants to find whether engineers tend to earn higher income than other jobs, the data inserted into the table *groupEmployment* appears to be

1 when the employee’s job title does not contain the word ‘engineer’, otherwise it will become 2.

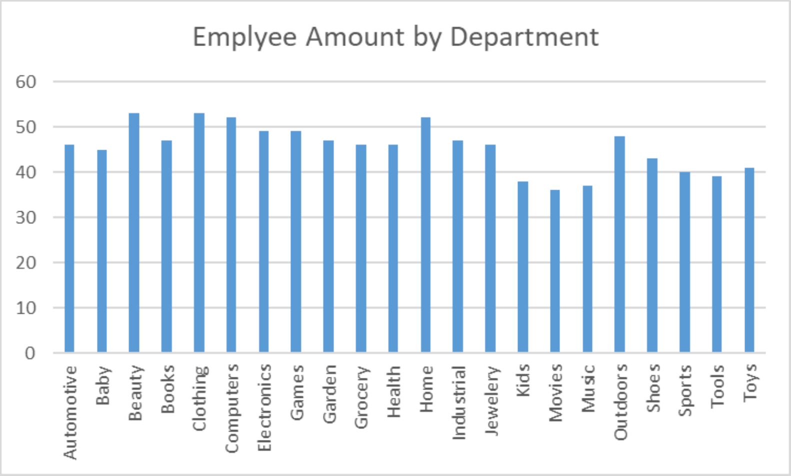


3. Project Analysis

3.1 Exploratory Data Analysis

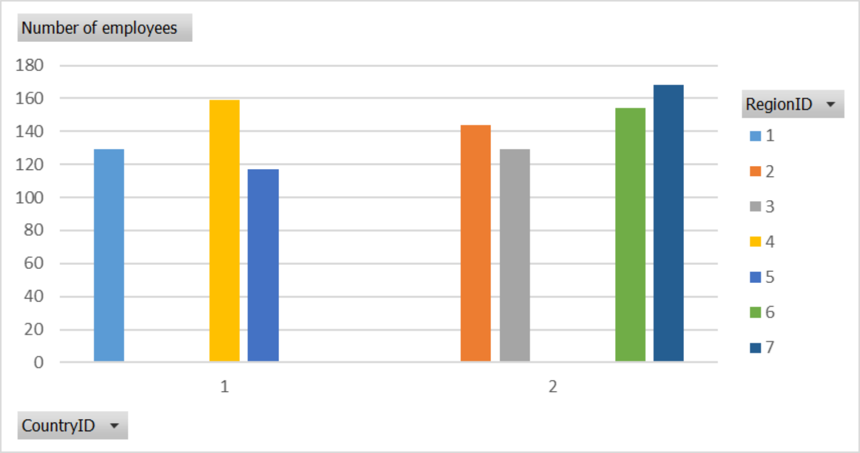
3.1.1 Number of Employee by Department

The total number of employees in each department is calculated using ‘Group by’ function initially and then count the number of employee ID. It shows that beauty department and clothing department have most employees, while least employees work in movies department. Similarly, the departments related to art and kids have less employees; for example, the departments of music, movies, kids, and baby.



3.1.2 Company’s Location

Most employees work in southwest of America, while least employees work in Quebec state in Canada with a possible reason that the weather is too cold to live. Moreover, if an employee in this dataset works in Canada, the Nova Scotia will be a place where most people and company choose to locate.



3.2 Salary Analysis

3.2.1 Basic Statistic

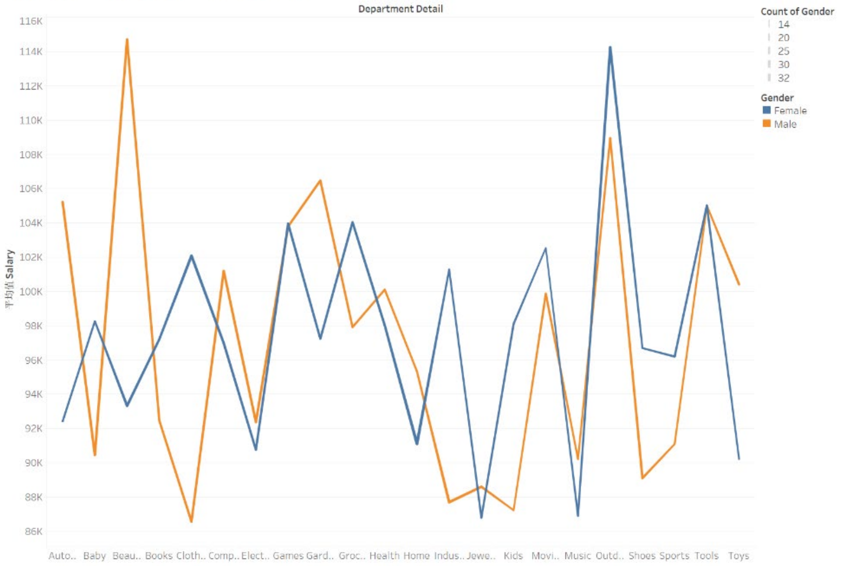
The highest amount of annual income among all of the employees is $149,929, and the lowest amount is $40,138.

3.2.2 Salary Distribution by Gender

The average income in each department is used to be analyzed here. Although in most departments the gender is not an obvious factor which influence the payment amount, several departments are detected to have a large gap between the salaries earned by female and male. For instance, in departments of kids, industrial, grocery, clothing, and baby, the female employees have higher average salary. However, in departments of automotive, beauty, computer, garden, and toys the male employees have higher average salary.

One potential reason for this phenomenon is that the female employees usually are more emotional than the male and have better empathy as well, which allow them to have a strength in departments related to kids and fashion. On the contrary, the male employees tend to have higher salary in the departments related to craftsmanship and delicate tools due to their better ability to get to know and explain these products more clearly.

<Average Salary by Department & Gender>



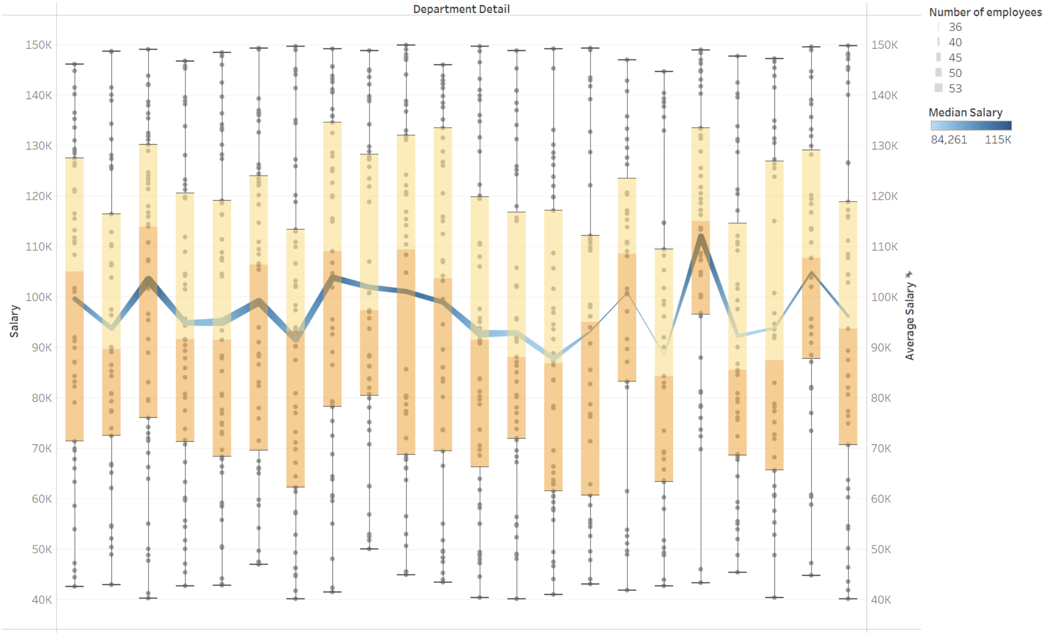
The trend of average of Salary for Department Detail. Color shows details about Gender. Size shows count of Gender.

3.2.3 Salary Distribution by Department

The box-and-whisker plots are drawn here, where each small spot indicates the average salary amount of the employees having the same job titles. Additionally, each box states the salary distribution of a department. For instance, most employees in the outdoors department earn an annual salary equal or above the average salary level in this department, which shows that it is the most concentrated department.

This analysis offers an opportunity for new starters to have an insight in a particular department for how much money they are most likely to get when they first step into their career path with a beginner job title, and how the salary amount will improve once they get senior level position in this department.

<Salary Distribution (Jobs & Department)>

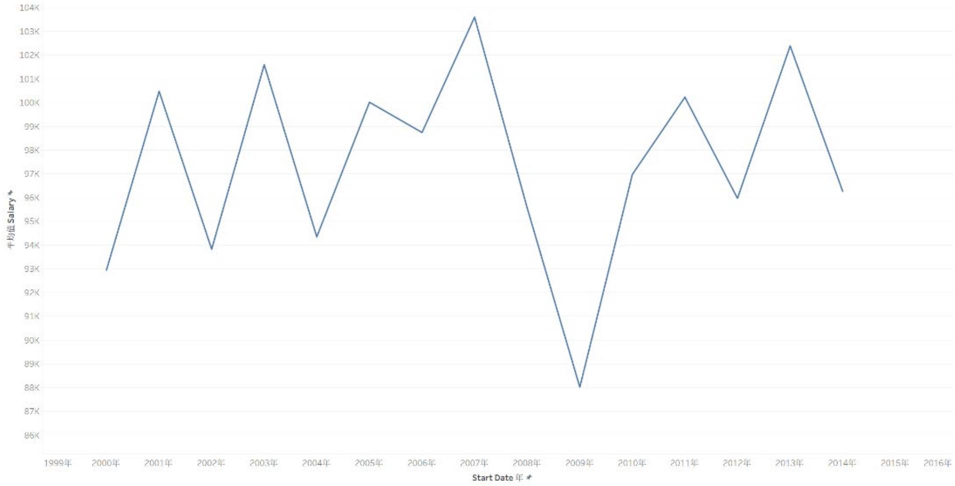


The trends of Salary and average of Salary for Department Detail. For pane Salary: Details are shown for Job title. For pane Average of Salary: Color shows median of Salary. Size shows count of Employee ID. The view is filtered on Salary, which includes values greater than or equal to 40138.

3.2.4 Time-series Analysis

Based on the time series analysis, it indicates that a deep decrease occurs in 2008 and 2009 becomes the year which has the lowest average salary, and a possible reason behind this trend should be the world-wide financial crisis happening in 2008.

<Average Salary Time-series Analysis>



The trend of average of Salary for Start Date Year. The data is filtered on Department Detail, which keeps 22 of 22 members.