Exploratory Data Analysis

- StudentID: 21800272

- Name: Park Sang Beom

- 1st Major: Korean Law

- 2nd Major: Data Science

Brief summary of your proposed project idea.

We analysed the data by considering the main audience for this data.

1. Account transaction data: bank
    1. Analysed the trend of deposits and withdrawals
2. Card data: card companies
    1. Bank information by card company
    2. Percentage of customers who matured by card company
3. Employees and detailed data: company owners
    1. Wage gap by region
    2. Analysis of the causes of turnover
4. Sales data: sales outlets
    1. Calculate the profit margin percentage for each item sold
    2. What are the top selling items in each country

1. Data overview

1. **Data size**
    1. Commonality
        1. 5000000 rows exist
    2. Differences (variables)
        1. BT_data : 5 columns
        2. CC_data: 11 columns
        3. HR_data: 37 columns
        4. HRA_data: 35 columns
        5. S_data: 14 columns
2. **BT_data**
    1. Introduction: Data on account deposits and withdrawals by transaction date
    2. Detailed variable introduction

        1. Convert Date column to Datetime because it is an Object type The year is from 2020 to 2155, but there is no description of the data, so it is applied as it is.
        2. Deposits and Withdrawals are also of Object type, so we removed unnecessary characters and converted them to float.
        3. Subtract Year from Date to create a Year column to analyse the trend of deposits and withdrawals by year

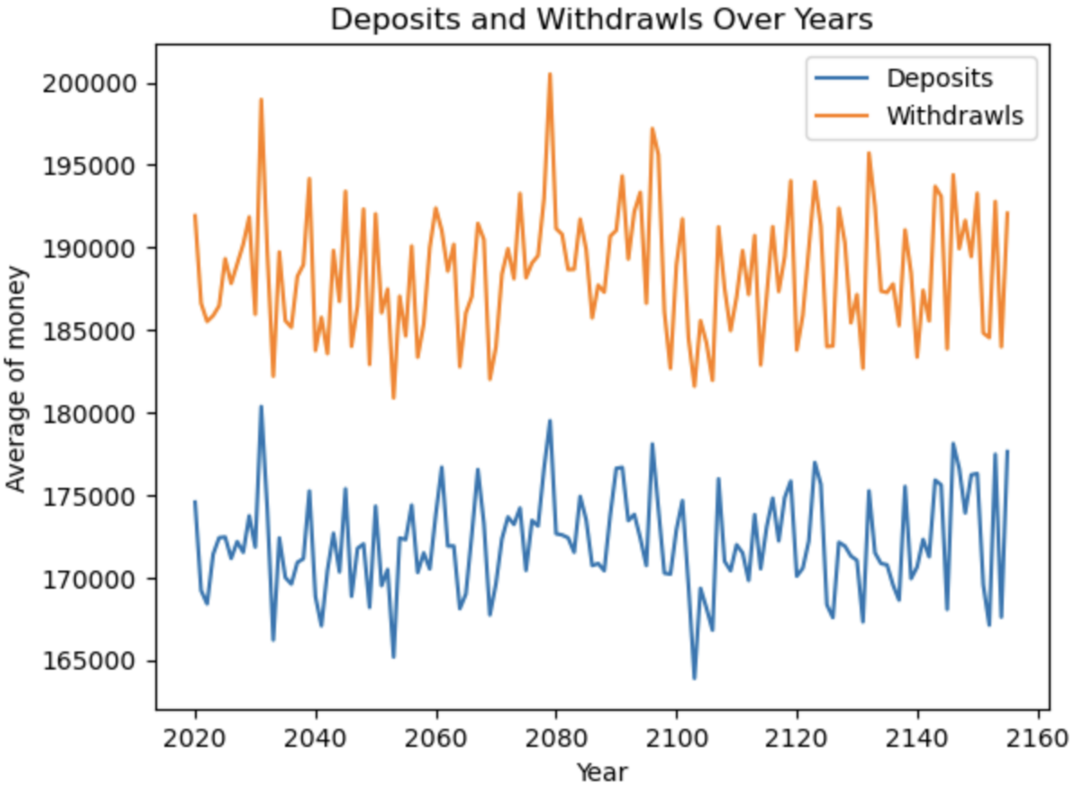
3. **CC_data**
    1. Introduction: data containing customer's card information
    2. Detailed variable introduction
        1. Convert Issue_date and Expiry_date to Datetime as they are Object types
        2. Create Time remaining until reissue by (Expiry_date - current date) to check whether the card is expired for each customer.
        3. Using the column of Time remaining until reissue, divide the value of Time remaining until reissue by less than 0 to expired customers, and greater than 0 to healthy customers, but less than 60 days away from expiry date to replace the card.
4. **HR_data**
    1. Introduction. : Data about employees' information
    2. Detailed variable introduction
        1. For Salary column, the minimum value is 4.000000e+04 and maximum value is 2.000000e+05
5. **HRA_data**
    1. Introduction: data about employees' details
    2. Introduction of detailed variables
        1. For NumCompaniesWorked, a variable that deals with the number of job changes from a minimum of 1 to a maximum of 8
        2. For JobSatisfaction, a variable that ranges from 1 to 4 in terms of job satisfaction
        3. For RelationshipSatisfaction, a variable with a satisfaction level from 1 to 4
        4. For JobInvolvement, a variable with a satisfaction level from 1 to 4
6. **S_data**
    1. Introduction: data containing detailed information about the items sold by country
    2. Detailed variable introduction
        1. Profit Margin Ratio is created through Total Profit / Total Revenue to distinguish items with good net margin ratio

2. Univariate analysis

1. **BT_data**

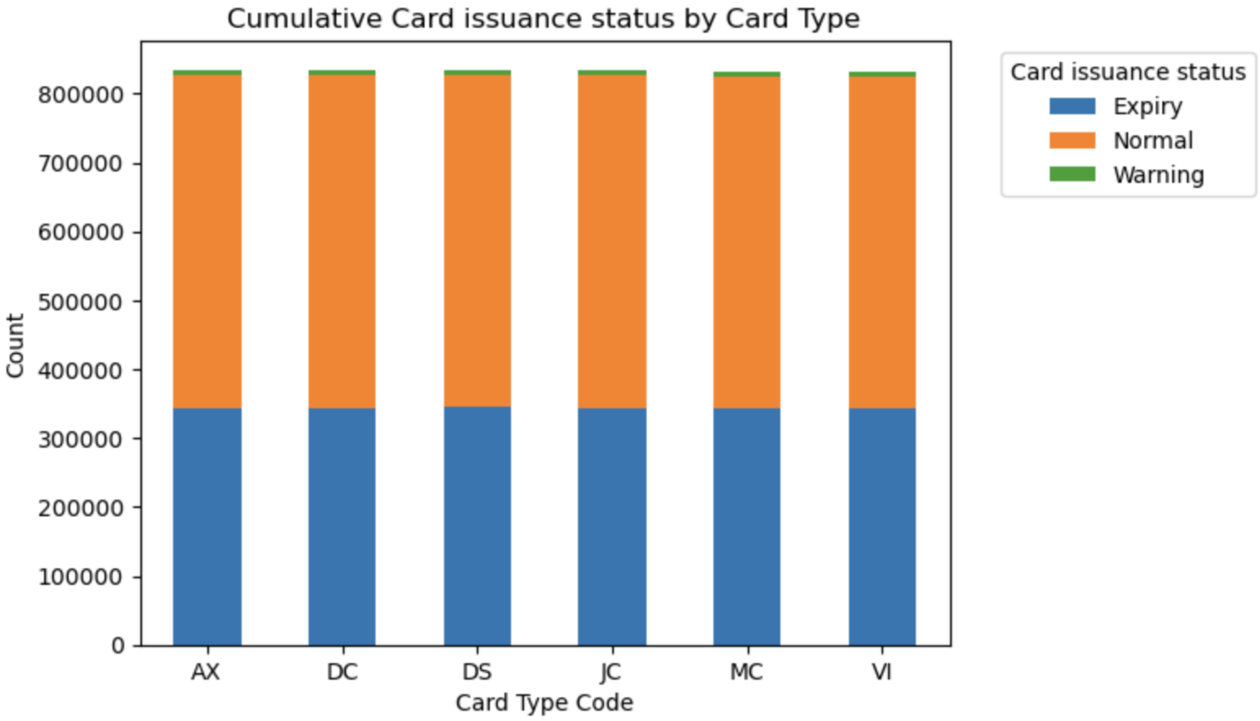
    1. Title : Deposits and Withdrawals Over Years
    2. Description : We can see that the amount of withdrawals is larger than the amount of deposits every year. However, you can see that the peaks and troughs of deposits coincide with the peaks and troughs of withdrawals, which suggests that when more money comes in, more

spending occurs, and vice versa.



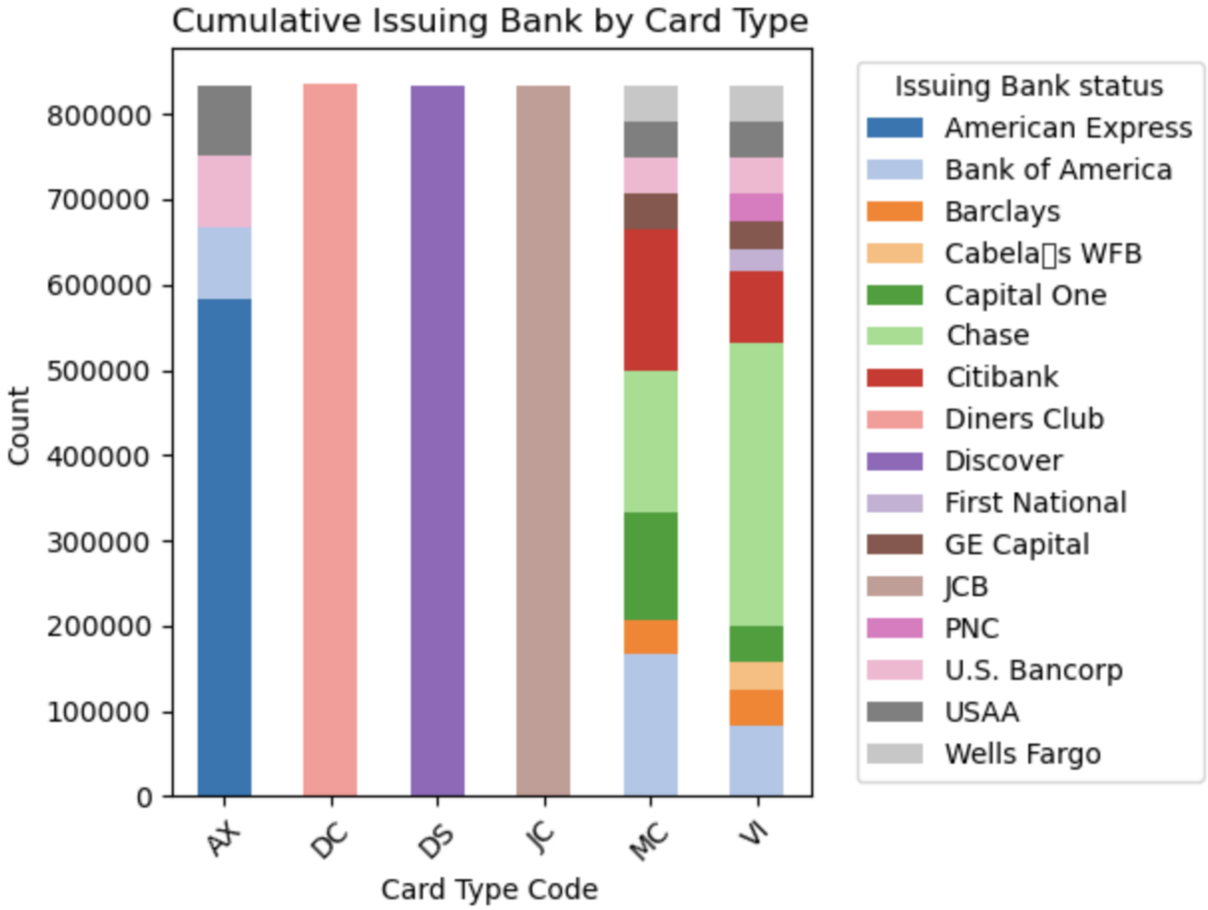
2. **CC_data**

    1. Title : Cumulative Card issuance status by Card Type
    2. Description : This graph shows the percentage of expired customers, the percentage of customers who need to replace the card, and the percentage of people who are normal for each card company. Most of the expired customers and low-quality customers are similar, but since there are many card issuers for all expired customers, it is necessary to make a sales strategy to regain customers.



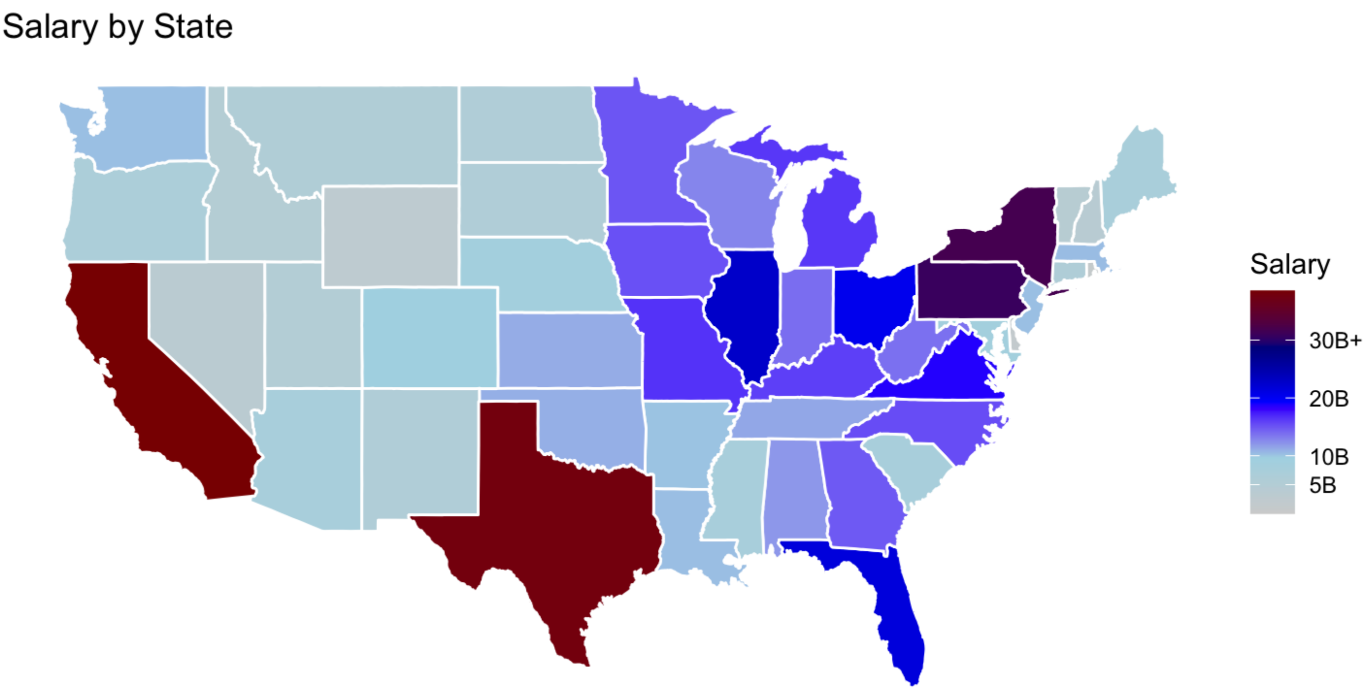
1. Title : Cumulative Issuing Bank by Card Type

2. Description : This graph shows which banks each card company is affiliated with. It can be seen that DC and DS are only partnered with one bank, while famous MC and Visa are partnered with various banks.

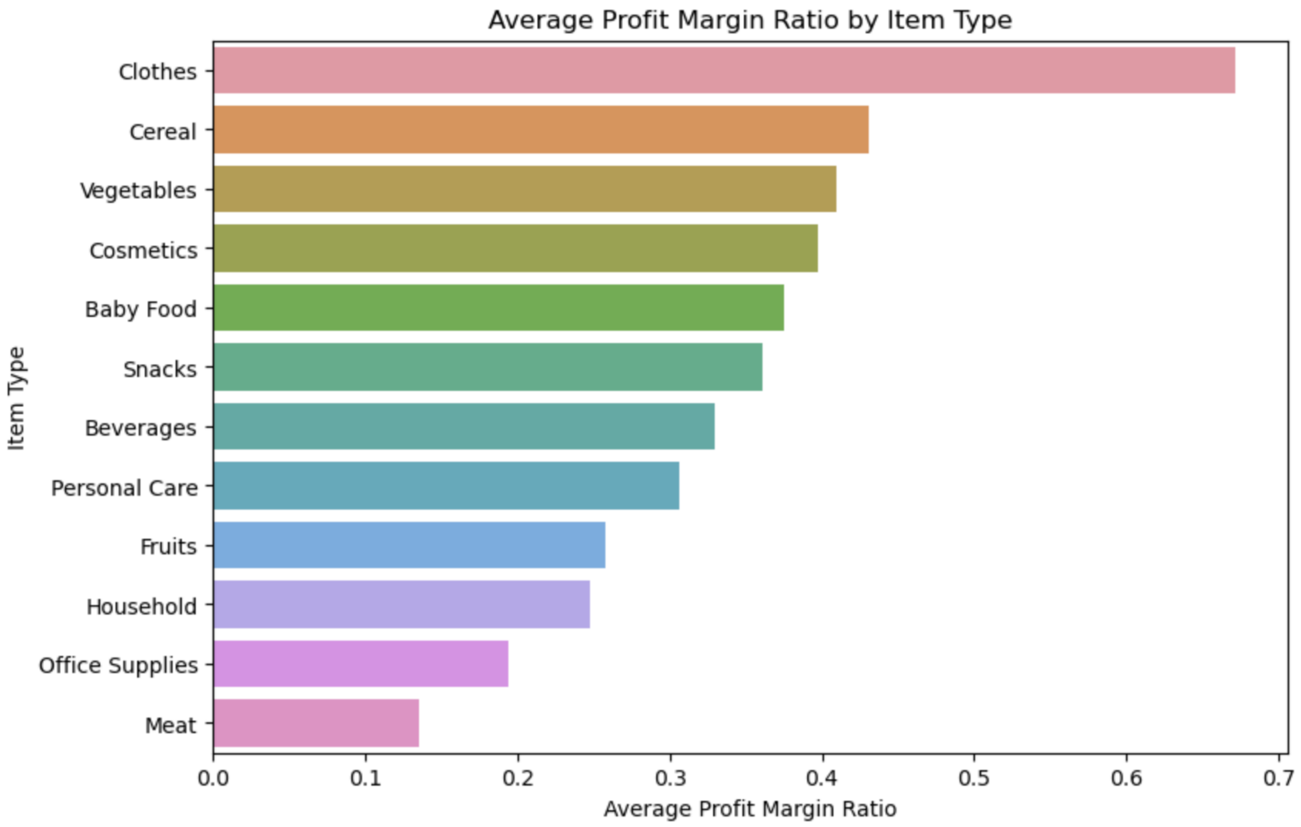


3. **HR_data**

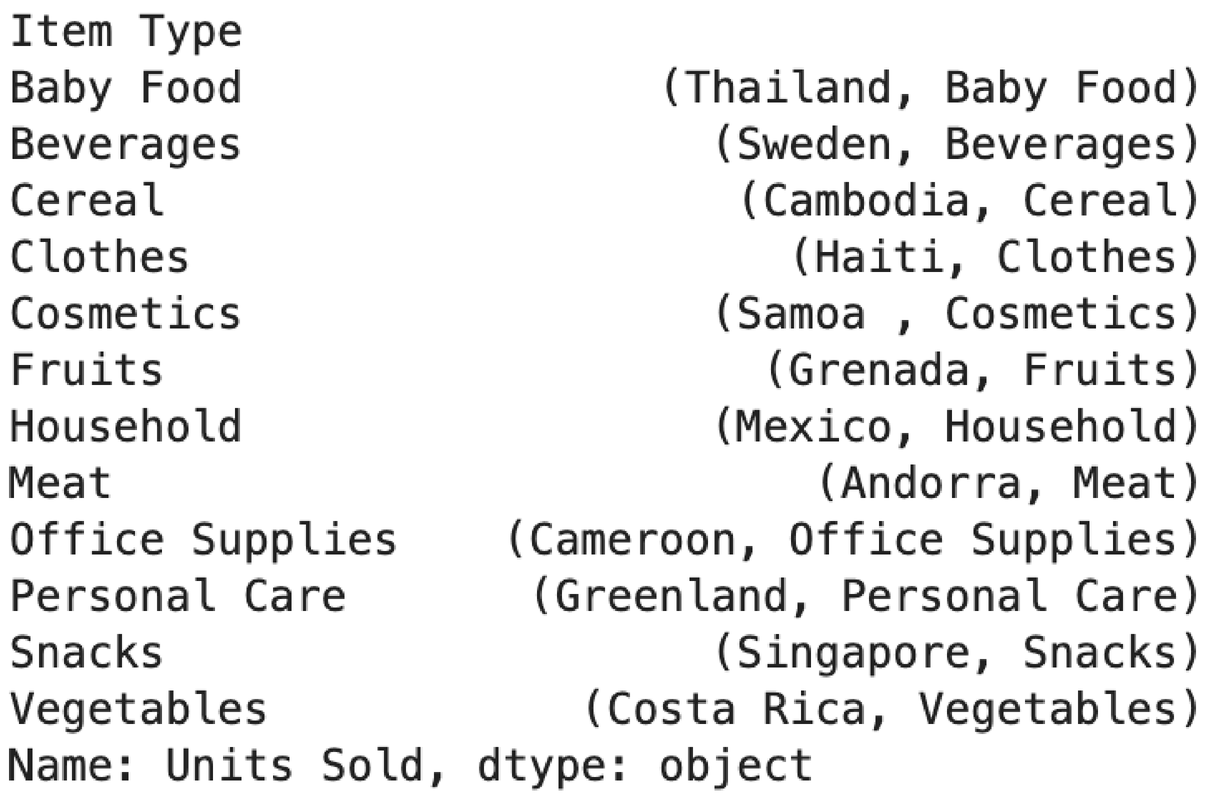
    1. Title : Salary by State
    2. Description : The sum of deposits for each state is calculated and displayed on a map. In the case of this company, you can see that California and Texas pay high wages, and New York also pays high wages.



1. **S_data**
    1. Title: Average Profit Margin Ratio by Item Type
    2. Description: Items with high profit margin rate are followed by clothes and grains



1. Title :top_selling_by_Itemtype
2. Description: For each item, it shows the top selling country by matching it with the top selling country.



```
Item Type
Baby Food                      (Thailand, Baby Food)
Beverages                        (Sweden, Beverages)
Cereal                            (Cambodia, Cereal)
Clothes                             (Haiti, Clothes)
Cosmetics                        (Samoa , Cosmetics)
Fruits                              (Grenada, Fruits)
Household                        (Mexico, Household)
Meat                                  (Andorra, Meat)
Office Supplies        (Cameroon, Office Supplies)
Personal Care           (Greenland, Personal Care)
Snacks                           (Singapore, Snacks)
Vegetables                  (Costa Rica, Vegetables)
Name: Units Sold, dtype: object
```

3. Multivariate analysis

1. **HRA_data**
    1. For people who have changed jobs
        1. Explanation: Job satisfaction and relationship satisfaction are negatively correlated. This means that people with high job satisfaction and low relationship satisfaction may have changed jobs because one or the other is low.

	NumCompaniesWorked	JobSatisfaction	RelationshipSatisfaction	JobInvolvement
NumCompaniesWorked	1.000000	-0.000096	0.001343	0.000725
JobSatisfaction	-0.000096	1.000000	-0.000823	-0.000390
RelationshipSatisfaction	0.001343	-0.000823	1.000000	-0.001209
JobInvolvement	0.000725	-0.000390	-0.001209	1.000000

2. Those who did not change jobs
    1. Explanation: It can be seen that job satisfaction and relationship satisfaction are proportionally high, which means that they are not changing jobs because they are satisfied with their relationship and satisfied with their current job.

	NumCompaniesWorked	JobSatisfaction	RelationshipSatisfaction	JobInvolvement
NumCompaniesWorked	1.000000	-0.000284	-0.000086	0.000105
JobSatisfaction	-0.000284	1.000000	0.000019	-0.000524
RelationshipSatisfaction	-0.000086	0.000019	1.000000	-0.000243
JobInvolvement	0.000105	-0.000524	-0.000243	1.000000

4. Suggestion

1. **BC_data**: By predicting the trend of deposits and withdrawals according to the year, if there are a lot of deposits, there will be a lot of withdrawals, so you can make spending products for customers in relevant years or seasons, or take measures to maintain a sufficient bank balance by predicting the year when there will be a lot of withdrawals so that bank bankruptcy does not occur.

2. **CC_data**: By showing the ratio of expiring customers and card replacement customers for each card company, you can strategise which customers you should pay more attention to at the moment, and by checking the percentage of cooperation with a specific bank, you can customise which banks you should work with more or less.

3. **HR_data**: The distribution of wages in each state shows which states are paying more wages, and the discovery of skewed wages in certain states can be used to analyse why they are paying more wages in those states, so if the wages are skewed for unreasonable reasons, it can be an opportunity for the company to prevent unreasonable expenditure

4. **HRA_data**: Job satisfaction and relationship satisfaction are inversely correlated for those who have changed jobs, and proportionally for those who have not changed jobs.

   This suggests that those who left their jobs may have been driven by either relationship or job satisfaction issues.

   This could lead to further investigation into the reasons for the departure and prevent employees from leaving due to company issues.

5. **S_data**: By checking which items have high margins, you can customise strategies for those items and focus production on them, and you can also check the countries where each item is sold the most to customise strategies for each country.