Test Paper Set 2

25 Marks

DATA AVAILABLE -

This book has the following sheets:

- Customer Acquisition: At the time of card issuing, company maintains the details of customers.
- Spend (Transaction data): Credit card spend for each customer
- Repayment: Credit card Payment done by customer

Create graphs for

a. Monthly comparison of total spends, city wise
b. Comparison of yearly spend on air tickets
c. Comparison of monthly spend for each product (look for any seasonality that exists in terms of spend)

Write user defined R function to perform the following analysis:

You need to find top 10 customers for each city in terms of their repayment amount by different products and by different time periods i.e. year or month. The user should be able to specify the product (Gold/Silver/Platinum) and time period (yearly or monthly) and the function should automatically take these inputs while identifying the top 10 customers.

2. From the above dataset create the following summaries:
   a. How many distinct customers exist?
   b. How many distinct categories exist?
   c. What is the average monthly spend by customers?
   d. What is the average monthly repayment by customers?
   e. If the **monthly** rate of interest is 2.9%, what is the profit for the bank for each month? (Profit is defined as interest earned on Monthly Profit. Monthly Profit = Monthly repayment – Monthly spend. Interest is earned only on positive profits and not on negative amounts)
   f. What are the top 5 product types?
   g. Which city is having maximum spend?
   h. Which age group is spending more money?
   i. Who are the top 10 customers in terms of repayment?

3. Calculate the city wise spend on each product on yearly basis. Also include a graphical representation for the same.

25 Marks

In this project we will be working with the UCI adult dataset. We will be attempting to predict if people in the data set belong in a certain class by salary, either making <=50k or >50k per year.

Instructions

Just complete the tasks outlined below.

Get the Data

Q1. Read in the adult_sal.csv file and set it to a data frame called adult.

Q2. Check the head of adult

Q3. Check the head,str, and summary of the data now.

Q4. Use table() to check out the frequency of the type_employer column.

Q5. How many Null values are there for type_employer? What are the two smallest groups?

Q6. Combine these two smallest groups into a single group called "Unemployed". There are lots of ways to do this, so feel free to get creative. Hint: It may be helpful to convert these objects into character data types (as.character() and then use sapply with a custom function)

Q7. What other columns are suitable for combining? Combine State and Local gov jobs into a category called SL-gov and combine self-employed jobs into a category called self-emp.

Q8. Use table() to look at the marital column

Q9. Check the country column using table()

Q10. Group these countries together however you see fit. You have flexibility here because there is no right/wrong way to do this, possibly group by continents. You should be able to reduce the number of groups here significantly though.

Q11. Use table() to confirm the groupings

Q12. Check the str() of adult again. Make sure any of the columns we changed have factor levels with factor()

Q13. Use ggplot2 to create a histogram of ages, colored by income.

Q14. Plot a histogram of hours worked per week

Q15. Rename the country column to region column to better reflect the factor levels.

Q16 Create a barplot of region with the fill color defined by income class. Optional: Figure out how rotate the x axis text for readability

Q17. Split the data into a train and test set using the caTools library

Q18. Train the model using glm() function