Problem Set 9

CSCI 5a: Programming in R

Let's clear the global computing environment:

```r
rm( list = ls() )
```

Problem 1: Stock Data

Part (a): Reading in the stock data

First, read in the data from the file "Problem 1 Stock Data.csv", which is contained in the folder "Problem Set 9 Data". Save this data in a variable, and display the first six rows by using a `head()` command.

Solution

Part (b): Defining the VWAP function

In Problem Set 7, Problem 5, part (a), you defined a function to calculate the volume-weighted average price (VWAP), given a vector of price per share data and a vector of number of shares sold data. Copy that function, and paste it into this file; if you want to, you can use my solution, but you'll get more out of the problem if you use your own code.

Solution

Part (c): Selecting rows

Using the data frame from part (a), select the rows corresponding to sales for Global Widget. Save this data frame in a variable, and display the first 6 rows using a `head()` statement.

Solution

Part (d): Selecting columns

Using the Global Widget data frame you created in part (c), select the columns for price per share data and number of shares sold, and use your function from part (b) to calculate the VWAP for Global Widget. Report your result using a `cat()` statement, displaying this value to 2 decimal places.

Solution

Part (e): Writing a function

Write a function that takes a character string identifier for a company name, and returns the VWAP for the company:

- First, using the stock data frame from part (a), the function selects all transactions for the specified company.
- Next, the function computes the VWAP for this company using the function definition form part (b) and the appropriate rows from the data frame, and then returns this value.

There's nothing to report for this part, but write your code clearly so the TAs can understand what you're doing.

Solution

Part (f): WiDgT VWAP

Use your vwap database calculator function to determine the VWAP for WiDgT. Report this value using a `cat()` statement, displaying this value with 2 decimal places.

Solution

End of problem 1

Problem 2: Baseball Database Reporter

In this problem, we're going to pull together a number of ideas from the course to make a baseball database reporter function. The actual new code that you have to write is not all that much.

Our goal in this problem is to create a baseball reporter system that can take the name of a player, query a database for the annual batting data for that player, and then produce a summary report.

We'll start by working through a single case, and then we'll write a function to generalize this method.

Part (a): Read in the data

First, read in the data from the file "Baseball Batting Database.csv", which is contained in the folder "Problem Set 9 Data".

Solution

Part (b): Selecting the rows for Willie Mays

We're going to start by learning how to handle a special case, which for this problem will be Willie Mays.

Select the rows from the baseball database corresponding to Willie Mays. Save this data frame in a variable, and display the first 5 rows directly.

Solution

Part (c): Willie Mays baseball report

In Problem Set 7, we constructed a baseball reporter function that took baseball batting data for a player and generated a batting summary report.

Copy the code for this baseball reporter function from your Problem Set 7 solutions, along with any additional functions that you need. (If you want to, you can use the code from the Problem Set 7 solutions.)

Then call the baseball reporter function using the data from part (b) to generate a baseball batting report on Willie Mays.

Solution

Part (e): Player reporter function

Write a function that takes a player's name, and then generates a baseball report:

- First, the function selects the rows from the baseball database corresponding to that player.
- Then the function calls the baseball reporter function, using the data from the selected rows.

There's nothing to report here, but write your code clearly so the TAs can understand what you're doing.

Solution

Part (f): Roberto Clemente batting report

Use the baseball player reporter function you defined in part (e) to generate a report for Roberto Clemente.

Solution

End of problem 2

Problem 3: Sales Data

Sales data for the months of June, July, and August are contained in these files:

- “Problem 3 June Data.csv”
- “Problem 3 July Data.csv”
- “Problem 3 August Data.csv”

These files are contained in the folder “Problem 3” in the “Problem Set 9 Data” folder.

Part (a): Reading the June data

Read in the file “Problem 3 June Data.csv” and store the data frame in a variable. Then display the first 8 rows using a `head()` statement.

Solution

Part (b): Reading the July data

Read in the file “Problem 3 July Data.csv” and store the data frame in a variable. Then display the first 8 rows using a `head()` statement.

Solution

Part (c): Reading in the August data

Read in the file “Problem 3 August Data.csv” and store the data frame in a variable. Then display the first 8 rows using a `head()` statement.

Solution

Part (d): Appending rows

Combine these three monthly data sets together by appending rows into a single data frame and store it in a variable. You might have to make some adjustments before you’re able to do this. The column names for this combined data frame should be:

- Location
- Revenues
- Costs

Once you’ve created the combined data frame, display the first 8 rows using a `head()` statement.

Solution

Part (e): Calculating profits

Create a new column in the combined data frame from part (d) named “Profits”, defined as the Revenues minus the Costs. Then display the first 8 rows using a `head()` statement.

Solution

Part (f): Histogram of profits

Using the profit data from part (e), construct a histogram of the profits for the months June, July, and August.

Solution

End of problem 3

Problem 4: Cereal Data

Part (a): Reading the brand data

Read in the file "Problem 4 Brand Data.csv", which is located in the folder "Problem 4". Save this data frame in a variable, and display the first 8 rows using a `cat()` statement.

Solution

Part (b): Reading the sales data

Read in the file "Problem 4 Sales Data.csv", which is located in the folder "Problem 4". Save this data frame in a variable, and display the first 8 rows using a `cat()` statement.

Solution

Part (c): Stripping the id numbers

We would like to merge the two data files together, but unfortunately the id numbers for the transactions have extra characters which will prevent this.

Create a new column in the brand data frame which consists of just the transaction id numbers. To do this, use the `substr()` function to strip out the extra characters.

When you're done, display the first 8 rows of the brand data frame, including the column with the transaction id numbers.

Solution

Part (d): Stripping the id numbers

We would like to merge the two data files together, but unfortunately the id numbers for the transactions have extra characters which will prevent this.

Create a new column in the sales data frame which consists of just the transaction id numbers. To do this, use the `substr()` function to strip out the extra characters.

When you're done, display the first 8 rows of the sales data frame, including the column with the transaction id numbers.

Solution

Part (e): Merging the data frames

Merge the two data frames from parts (c) and (d) together. Save the resulting data frame in a variable, and display the first 8 rows using a `head()` statement.

Solution

Part (f): Total sales amount

Using your merged data frame, calculate the total sales amount for all the transactions.

Use the familiar table of cereal brand prices:

Brand	Price
SBZ	2.99
KYM	3.49
HKT	7.99

Report your result using a `cat()` statement, rounding to 2 decimal places.

Solution

End of problem 4

Problem 5: Stratified Charts

Part (a): Stratified boxplot

Construct a stratified boxplot across species for the `Petal.Width` values in the `iris` data frame. Include a main title, and titles for the x- and y-axes. Align the boxes vertically, and specify a color for each one.

Solution

Part (b): Stratified stripchart

Construct a stratified stripchart across species for the `Petal.Width` values in the `iris` data frame. Include a main title, and a title for the x-axis. Align the individual stripcharts horizontally, and specify the point shape, size, and color. Finally, be sure to use jitter on the points.

Solution

End of problem 5

Problem 6: Likert Scale

The marketing team at WiDgT has conducted a survey to determine consumer interest in widgets.

The survey consists of 5 statements:

1. "I think widgets are important and valuable in modern society."
2. "I enjoy watching videos about my favorite celebrities and their expensive designer widgets."
3. "I dislike widgets intensely."
4. "I am a morally superior person because I possess many widgets."
5. "Only the dissolute and wicked dabble in widgets."

For each statement, participants can respond with either "Agree", "No Opinion", or "Disagree".

Responses to items 1, 2, and 4 are converted to a numeric score using this table:

Response	Score
"Agree"	3
"No Opinion"	2
"Disagree"	1

Responses to items 3 and 5 are converted to a numeric score using this table:

Response	Score
"Agree"	1
"No Opinion"	2
"Disagree"	3

In addition, participants are invited to join the WiDgT TikTok challenge group, and their response is scored using this table:

Response	Score
"Yes"	2
"No"	1

The scores on all items are added together to obtain a final score.

Subjects are then categorized into 3 groups based on an interval table:

Range	Interest Level
Final Score < 12	Low Interest
12 <= Final Score < 15	Medium Interest
15 <= Final Score	High Interest

Part (a): Data Analysis

The data for this survey is contained in the file "Problem 6 Data.csv".

Your job is to produce three summaries of this data:

- A table of the relative proportions of the three groups of interest levels, with the values displayed with 2 decimal places.
- A barplot of the relative proportions of the three groups of interest levels.
- A pie chart of the relative proportions of the three groups of interest levels.

You're on your own for this one! You'll have to decide how to structure this analysis, and there are many different approaches that are all valid.

Break your computation into multiple steps, using the features of R notebooks to document your work.

Choose your variable names carefully, and write your code clearly so the TAs can understand what you're doing.

You'll have to determine how display your results clearly; don't make the TAs hunt around for your answer.

In the end we just want to see the table of relative proportions, the barplot, and the pie chart.

Good luck!!

Solution