Problem Set 7

CSCI 5a: Programming in R

Let's clear the global computing environment:

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

```
load( "Problem Set 7 R Objects.Rdata" )
```

Let's review the objects that we just loaded in:

```
ls()
```

```
## [1] "problem.3.number.of.shares.sold.data"
## [2] "problem.3.price.per.share.data"
## [3] "problem.4.data"
## [4] "problem.5.cereal.brand.data"
## [5] "problem.6.number.of.widgets.sold.data"
## [6] "problem.6.region.data"
```

Problem 1: Baseball Functions

Of course I'm going to ask you this.

Part (a): Batting average

Write a function that takes two arguments, the number of at-bats and the number of hits, and returns a numeric value representing the player's batting average.

Then use this function to calculate Ted Williams' batting average. Report your result using a `cat()` statement, displaying this value using the standard baseball convention i.e. no leading 0 and with three digits to the right of the decimal place. (See Exercise 1.1 in Module 1 for how to do this.)

From Lecture 2, Module 3, Exercise 2.1, we know Ted Williams' baseball statistics:

Statistics	Value
Plate appearances	9,792
At-bats	7,706
Hits	2,654
Doubles	525
Triples	71
Home Runs	521
Bases on balls	2,021

Statistics	Value
Hit by a pitch	39
Sacrifice flies	20

Solution

Part (b): On-Base percentage

Write a function that takes five input arguments:

- The number of at-bats
- The number of hits
- The number of bases on balls
- The number of times hit by a pitch
- The number of sacrifice flies

The function then returns a numeric value for the on-base percentage (using the simplified definition for our course).

Then use this function to calculate Willie Mays' on-base percentage. Report your result using a `cat()` statement, displaying this value using the standard baseball convention.

From Problem Set 3, Problem 1, we have Willie Mays' career totals:

Statistics	Value
Plate appearances	12,497
At-bats	10,881
Hits	3,283
Doubles	523
Triples	140
Home Runs	660
Bases on balls	1,464
Hit by a pitch	44
Sacrifice flies	91

Solution

Part (c): Slugging percentage

Write a function that takes five arguments:

- The number of at-bats
- The number of hits
- The number of doubles
- The number of triples

- The number of home runs

The function then returns a numeric value for the slugging percentage.

Then use this function to calculate Babe Ruth's slugging percentage. Report your result using a `cat()` statement, displaying this value using the standard baseball convention.

From Lecture 2, Module 3, Section 5, we have Babe Ruth's baseball statistics:

Statistics	Value
Plate appearances	10,626
At-bats	8,399
Hits	2,873
Doubles	506
Triples	136
Home runs	714
Bases on balls	2,062
Hit by a pitch	43
Sacrifice flies	0

Solution

End of problem 1

Problem 2: Baseball Reporter

Now we can use the functions that you wrote in Problem 1 to help build a baseball reporter.

Part (a): Reporter function

Write a reporter function that takes 9 input arguments:

- The player's name
- The number of at-bats
- The number of hits
- The number of doubles
- The number of triples
- The number of home runs
- The number of bases on balls
- The number of times hit by a pitch
- The number of sacrifice flies

The function then prints out a report on the player:

- The report first lists the player's name.
- The report then displays a tabulation of the player's basic statistics:
  - The number of at-bats
  - The number of outs
  - The number of bases on balls
  - The number of hits
  - The number of singles
  - The number of doubles
  - The number of triples
  - The number of home runs
- Then the report lists the three standard batting performance statistics:
  - Batting average
  - On-base percentage
  - Slugging percentage

All of these perfomance statistics should be reported with a `cat()` statement using the standard baseball convention.

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

Solution

Part (b): Hank Aaron

Use the function you constructed in part (a) to generate a batting performance report for Hank Aaron.

Statistics	Value
Plate appearances	13,941
At-bats	12,364
Hits	3,771
Doubles	624
Triples	98
Home Runs	755
Bases on balls	1,402
Hit by a pitch	32
Sacrifice flies	121

Solution

End of problem 2

Problem 3: VWAP

Part (a): Function definition

Suppose we have data on a number of stock sales for a particular company, and for each sale we know the price per share and the number of shares sold.

Write a function that takes two input arguments:

- The first argument is a numeric vector consisting of the price per share for each transaction.
- The second argument is a numeric vector consisting of the number of shares sold for each transaction.

The function then returns a numeric value for the VWAP.

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

Solution

Part (b): Calculation

The vector `problem.3.price.per.share.data` contain the price per share for 185 stock transactions. The vector `problem.3.number.of.shares.sold.data` contains data on the number of shares sold in each of the corresponding transactions.

Use the function that you defined in part (a) to calculate the VWAP for this stock. Report your result using a `cat()` statement, displaying this value with 2 decimal places.

Solution

End of problem 3

Problem 4: Three Views

In this problem, we'll examine the same set of data using three different visualization methods.

For each graph, produce a finished version with all the features that we've studied e.g. titles, colors, point shapes, jitter, etc.

When you finish the third graph, you should go back and compare all three graphs. Which kinds of information are displayed by each chart? Is there one display that you think is particularly useful? Which is your personal favorite? You don't have to write out answer, but I think you'll get a lot more out of this problem if you engage with it and try to draw some conclusions for yourself.

Part (a): Stripchart

Construct a stripchart for the values in `problem.4.data`.

Solution

Part (b): Histogram

Construct a histogram for the data in `problem.4.data`.

Solution

Part (c): Boxplot

Construct a boxplot for the values in `problem.4.data`.

End of Problem 4

Problem 5: Removing Internal Codes, Part 1

The character string vector `problem.5.cereal.brand.data` contains data on each box of cereal that is sold.

Each item in this vector consists of an internal sales code and an identifier for the cereal brand.

The internal sales code has a fixed length, and the identifier for the cereal brand is always 3 letters.

For this problem, you should construct a frequency count table of the cereal brands and then display this in a barplot with the levels organized in decreasing order.

You're on your own for this one! You'll have to figure out how to do this, and there are alternative approaches. Just make sure that you show us the frequency count table and the barplot with the levels organized in decreasing frequency, and make them look nice so the TAs can grade them.

Solution

End of Problem 5

Problem 6: Removing Internal Codes, Part 2

WiDgT has data on all the transactions for their retail widget sales across four regions: North, East, South, and West.

The character string vector `problem.6.region.data` contains the region for each transaction, but preceded by an internal code of fixed length.

The vector `problem.6.number.of.widgets.sold.data` contains the number of widgets sold in the transaction.

In this problem, you must construct a table of relative proportions for the total number of widgets sold across the four regions, formatting all values with 2 decimal places.

Thus, if there were 1,000 total widgets sold across all 4 regions, and 400 were sold in the North region, then in the table of relative proportions you should display the value 0.40 for North.

In the table, the four regions should be labelled "North", "East", "South", and "West", and should be presented in that order.

Part (a): Stripping out the internal codes

Create a new vector consisting of the values of `problem.6.region.data` with the internal codes stripped out.

The problem statement only specified that the internal codes all have the same fixed length, but did not specify this length, so you'll have to do some exploration to determine how long it is.

Then create a factor from this new vector, with the correct region labels and order, and save this factor in a variable.

Finally, directly display a frequency count table of the regions so the TAs can check your work.

Solution

Part (b): North region total sales

Now let's review how to calculate the total number of widgets sold in a region.

As an example, we'll focus on the North region.

- First, construct a logical indexing vector where an element is `TRUE` if the corresponding element of the stripped region data from Part (a) represents the North region.
- Use this logical indexing vector to select the elements of `problem.6.number.of.widgets.sold.data` that were sold in the North district.
- Add up the values of the North widget sales data to obtain the total sales amount, and store this in a variable.

Report your final result using a `cat()` statement, formatting the value with 0 decimal places.

Solution

Part (c): Total sales function

We could repeat the calculations we just did to obtain the total number of widgets sold for the East, South, and West regions.

That would involve a lot of redundant code.

Instead, let's bundle this computation into a function, which will make our code much cleaner.

Write a function that takes a region name as the input argument and returns the total number of widgets sold for that region.

You can assume that the factor that you created in Part (a) and `problem.6.number.of.widgets.sold.data` are global variables.

There are many ways to do this, but here's one suggestion:

- First construct a logical indexing vector that is TRUE when the corresponding element of `problem.6.region.data` is equal to the input argument region name and FALSE otherwise, and save this in a variable.
- Use this logical indexing vector to select the elements of `problem.6.number.of.widgets.sold.data` that occurred in the specified region, and save this vector in a variable.
  - Add the elements of this filtered vector, and save this in a variable.
  - Return the value of this sum.

There are more compressed ways to do this, but it's fine (and good!) to break the function into a sequence of intermediate steps.

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

Solution

Part (d): North total sales

Use the function you defined in Part (c) to calculate the total sales for the North region. Store your result in a variable and report it using a `cat()` statement, displaying the value with 2 decimal places.

Hint: You should get the same value as you did in Part (b).

Solution

Part (e): East total sales

Use the function you defined in Part (c) to calculate the total sales for the East region. Store your result in a variable and report it using a `cat()` statement, displaying the value with 2 decimal places.

Solution

Part (f): South total sales

Use the function you defined in Part (c) to calculate the total sales for the South region. Store your result in a variable and report it using a `cat()` statement, displaying the value with 2 decimal places.

Solution

Part (g): West total sales

Use the function you defined in Part (c) to calculate the total sales for the West region. Store your result in a variable and report it using a `cat()` statement, displaying the value with 2 decimal places.

Solution

Part (h): Named vector

Construct a named vector where the names are the four regions in correct order, and the corresponding values of the named vector are the regional total number of widgets sold that we've calculated in parts (d) through (g).

Save the named vector in a variable, and display it directly for the TAs to check your work.

Solution

Part (i): Relative proportions table

Finally, we can construct a table of the relative proportions of the total widget sales across the four regions.

Recall that the `proportions()` function takes one input argument, which must be a table.

In part (h), we constructed a named vector, not a table.

Use the `as.table()` function to convert the named vector in part (h) to a table, and then use this table to construct a table of the relative proportions of the total number of widgets sold across the four regions.

Directly display the relative proportions table for the TAs to grade.

Solution