Problem Set 8 Solutions

CSCI E-5a: Programming in R

Let's clear the global computing environment:

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

Problem 1: Data Frame Reporter

Throughout this course, we've written what I call "reporter" functions, which take an R object and print out a range of information about its properties. Reporter functions are extremely useful when we first read in a data set, because they can quickly summarize a number of features of the data.

For this problem, you must construct a reporter function for a data frame. The reporter function takes one argument, a data frame, and then prints out a range of information:

- First, it reports the total number of rows and columns in the data frame.
- Then the reporter function prints out information for each column:
  - First, it prints out the name of the column.
  - Next, it prints out the class of the column, either `numeric`, `logical`, or `character`.
  - If the column is a numeric vector, then the reporter function prints out the sample mean, sample standard deviation, sample minimum, and sample maximum.
  - If the column is a logical vector, then the reporter function prints out the proportion of elements that are true.
  - If the column is a character string vector, then the reporter function prints out the number of unique values, as well as the first 3 unique values.
  - Finally, for each column, the reporter function prints out the number of missing items.

Part (a): Data frame reporter function

Write a data frame reporter function according the specifications above. Make the output nice by using newlines and tab characters to space out the display. There's nothing to report here, but write you code clearly so the TAs can understand what you're doing.

Solution

Part (b): Testing the data frame reporter

Read in the data from the file "Problem 1 Data.csv" located in the "Problem Set 8 Data" folder and run your data frame reporter on it.

Solution

End of problem 1

Problem 2: Find Your Own Dataset

This problem is essentially the entire point of the course.

In this problem, we want you to find your own dataset to study and practice with.

Your dataset should be in .csv or tab-delimited text form.

Find something that interests you. As always, you'll get a lot more out of the exercise if you engage with it, and if you care about the subject matter that makes it easier for you to maintain energy.

Part (a): Dataset description

Tell us a little bit about your dataset. Write a few sentences to give us some background information about the subject. Also, tell us where you got the dataset from. When you're all done, read in the data file and store the resulting data frame in a variable.

Solution

Part (b): Number of rows

How many rows does your dataset have? Determine this using R, and report your result with one sentence.

Solution

Part (c): Number of variables

How many variables does your dataset have? Determine this using R, and report your result with one sentence.

Solution

Part (d): `NA` values

Are there any `NA` values in your dataset? Tell us with one or two sentences.

Solution

Part (e): Ask a question

Pose an interesting question about your subject matter, and then use the data in your data frame to answer this question. You should calculate some sort of numerical value such as a sample mean or a sample proportion, and report this to us. Also, you should make some sort of visualization, although the precise graph is up to you.

What's important is that the numerical result and the visualization should be relevant to answering your question.

Hint: it's probably easier to find a good quality dataset, and then think of a question that you can answer with the data, rather than to think of the question first and then try to dig up some data to answer it.

You'll get full credit as long as you do *something*, and the TAs won't grade you based on the content of your analysis. However, I think you'll get more out of this exercise if you dig in and engage with it. Once again, we come back to: find something that you care about, and then you'll find it interesting to do the analysis.

Remember: all I can do is to demonstrate the tools, and show you some basic applications. Ultimately, it's up to you to integrate this knowledge into your practice, and only you can do this.

Finally, do a nice job, and present your results in a finished manner. Amaze and delight the TAs with your imagination, creativity, precision, and attention to detail.

Solution

End of problem 2

Problem 3: Loan Amortization

In our course we usually use data frames as a way to store data. But they can also be used to construct tabular displays, and this application is sometimes underappreciated.

In this problem, we're going to use R to display all of the steps involved in constructing a loan amortization schedule.

Part (a): Payment amount

Taylor takes out a loan for $200,000, to be paid back in 10 annual installments. The interest rate for the loan is 2.8%.

Calculate the annual payment amount that Taylor needs to make in order to fully pay off this loan with 10 annual installments. Report your result using a `cat()` statement, displaying this value with 2 decimal places.

Solution

Part (b): Loan amortization schedule

Construct a report that displays the loan amortization schedule for all 10 years.

To construct this report, first create three storage vectors, each of length 10:

- The first storage vector holds the current loan balance at the start of the year.
- The second storage vector holds the current loan balance at the end of the year, before the payment.
- The third storage vector holds the current loan balance at the end of the year, after the payment.

We've already seen the algorithm for constructing a loan amortization schedule.

However, for this problem, instead of printing out the values using a `cat()` statement, you should now store the result each individual calculation in one of the storage vectors:

- For year 1:
  - At the beginning of year 1, the current loan balance is just the initial loan amount. Store this in location 1 of the storage vector for the loan balance at the start of the year.
  - At the end of year 1 before the payment, the current loan balance has accrued interest. Determine the loan balance at the end of the year before the payment, and store this value in location 1 of the storage vector for the loan balance at the end of the year before the payment.
  - At the end of year 1 after the payment, the current loan balance has decreased by the payment amount. Determine this value, and store it in location 1 of the storage vector for the loan balance at the end of the year after the payment.
- Now we move on to year 2:
  - Determine the loan balance at the beginning of the year, and store this value in location 2 of the appropriate storage vector.
  - Determine the loan balance at the end of the year before the payment, and store this value in location 2 of the appropriate storage vector.

– Determine the loan balance at the end of the year after the payment, and store this value in location 2 of the appropriate storage vector.

Continue in this way: calculate each of the three loan balances for year i, storing them in location i of the appropriate storage vector.

At the end of this calculation you will have three storage vectors, each of length 10.

When you've finished calculating all the values, combine the three storage vectors into a data frame.

Finally, display the data frame directly.

Remember that you can use any string you like as a name for the data frame columns as long as you enclose the characters in quotes.

Solution

End of problem 3

Problem 4: Kepler's Third Law

This is one of my favorite problems in the whole course, and I'm excited to present it to you. Kepler's laws of planetary motion were some of the most important discoveries of early modern science, and the final result is easy to visualize. But we're studying this problem in our course because it also involves many of the R techniques that we've studied so far, and so it's a great example of combining multiple methods to obtain a solution.

Before we get started, you might find it useful to go back and review Section 2 from Week 3 Module 6: Stripcharts, which gives some background for this problem.

In the year 1619, Johannes Kepler published his Third Law of Planetary Motion.

The *period* of a planet's orbit, denoted T, is the time required to make one full orbit around the sun.

By definition, the Earth has a period of 1 year.

Mars has a period of 1.88 years, while Neptune has a period of 165 years.

The mean distance from the sun is denoted by R.

Then Kepler's third law states that for any planet the square of the period T is proportional to the cube of the mean distance R:

$$T^2 \propto R^3$$

If we take the logarithm of both sides, we obtain:

$$\log T \;=\; \frac{3}{2}\log R + c$$

In other words, when we take logarithms of both sides, we have a linear relationship between the logarithm of the period of the orbit and the logarithm of the radius of the orbit. (The constant c is also very interesting, but for our purposes we'll just ignore it.)

Let X denote the logarithm of the radius:

$$X \;=\; \log R$$

Let Y denote the logarithm of the period:

$$Y \;=\; \log T$$

Then we can write Kepler's third law as:

$$Y \;=\; \frac{3}{2}X + c$$

Part (a): Read in the data

The file "Problem 8 Data.csv" contains data on the mean distance from the sun and the period for all 8 planets (sorry, Pluto).

Read this data in, and store it in a data frame variable.

Then directly display the entire data frame.

Solution

Part (b): Log radius

Construct a vector of the logarithms of the mean radius from the sun, and store it in a variable.

Directly display the first 6 values of this vector, formatting the values with 2 decimal places.

Solution

Part (c): Log period

Construct a vector of the logarithms of the periods, and store it in a variable.

Directly display the first 6 values of this vector, formatting the values with 2 decimal places.

Solution

Part (d): Scatter plot

Construct a scatter plot of the data, using x-axis to display the logarithm of the radius and the y-axis to display the logarithm of the period.

For this scatter plot, the x-axis should range from 24 to 30, and the y-axis should range from -2 to 6.

You only have 8 points for this scatter plot, so you should make them larger than normal.

Kepler's third law of planetary motion states that the relationship between the log of the radius of the orbit and the log of the period is a straight line.

Looking at your scatter plot, do you think that it supports Kepler's third law of planetary motion?

Solution

Part (e): Linear model

Kepler's third law actually says something more than the relationship between the log of the radius and the log of the period is a straight line.

It also makes a precise prediction about the slope of this straight line: it must be 3/2, or 1.5.

Construct a linear model, where the log of the radius is the predictor variable and the log of the period is the outcome variable.

You can obtain the slope of the regression line by using the `summary()` function, or by directly extracting the coefficient from the linear model object.

Determine the slope of the least-squares regression line. Is it close to 1.5?

Solution

Part (f): Graphing the line

Copy your code over from part (d). Then draw the least-squares regression line from part (e) by using the `abline()` function.

Does this graph support Kepler's third law of planetary motion?

Solution

End of Problem 4

Problem 5: Baseball Data

Whenever we've calculated baseball batting statistics, we've always used the career totals for a player.

For instance, for Babe Ruth, we have:

Statistics	Career Total
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

However, this is not how baseball data is typically reported.

Instead, baseball statistics are typically reported on an annual basis.

That is, the data is presented for each year that a player played in the major leagues.

Babe Ruth's first season in the major leagues was 1914, and his last season was 1935, so he played a total of 22 seasons.

Thus, the data for Babe Ruth consists of 22 rows, one for each season.

The CSV file `Babe Ruth Annual Batting Data.csv` is located in the `Problem Set 8 Data` folder, in a subfolder called `Baseball Data`.

The column names for this data use these abbreviations:

Batting statistic	Abbreviation
At-Bats	AB
Hits	H
Doubles	2B
Triples	3B
Home Runs	HR
Bases on Balls	BB
Hit by a Pitch	HBP
Sacrifice Flies	SF

Part (a): Reading in the data

Read in the data from `Babe Ruth Annual Batting Data.csv` and store it in a variable. (You'll have to use a path for this.) There's nothing to report for this part, but write your code clearly so the TAs can understand what you're doing.

Solution

Part (b): Babe Ruth career hits

Can we use the data in this format to calculate Babe Ruth's career batting average? To do this, we have to know his career total hits and his career total at-bats. In this part, we'll calculate his career total hits.

First, select the vector of hits from the data frame from part (a). Then calculate the sum of the values in this vector,and store it in a variable. Report your result using a `cat()` statement, formatting the value with 0 decimal places.

Solution

Part (b): Babe Ruth career total at-bats

In this part, we'll calculate Babe Ruth's career total at-bats.

First, select the vector of at-bats from the data frame from part (a). Then calculate the sum of the values in this vector,and store it in a variable. Report your result using a `cat()` statement, formatting the value with 0 decimal places.

Solution

Part (d): Babe Ruth career batting average

Use your results from parts (b) and (c) to calculate Babe Ruth's career batting average. Report your result using a `cat()` statement, displaying the value using standard baseball formatting conventions.

Solution

Part (e): Ted Williams career on-base percentage

A CSV file containing Ted Williams' annual baseball batting data is contained in the file named "Ted Williams Annual Baseball Data.csv", located in the `Data` folder, in a subfolder called `Baseball Data`.

Read in Ted Williams' annual baseball batting data from this file, and store it in a variable. Then use this data to calculate Ted Williams' career on-base percentage. Report your result using a `cat()` statement, displaying the value using standard baseball formatting conventions.

Note: sacrifice flies were not recorded before 1954, so we have no data on these for players who played in these years. When computing baseball statistics involving sacrifice flies for players who played during these years, the standard practice is to simply assume a value of 0 sacrifice flies for these years. You'll have to figure out how to deal with this issue in this dataset.

Solution

Part (f): Willie Mays career slugging percentage

A CSV file containing Willie Mays' annual baseball batting data is contained in the file named "Willie Mays Annual Baseball Data.csv", located in the `Problem Set 8 Data` folder, in a subfolder called `Baseball Data`.

Read in Willie Mays' annual baseball batting data from this file, and store it in a variable. Then use this data to calculate Willie Mays' career on-base percentage. Report your result using a `cat()` statement, displaying the value using standard baseball formatting conventions.

Solution

End of Problem 5

Problem 6: Final Grades, Encore

You might have noticed that we still haven't fully solved the problem of calculating final grades for our course.

The remaining issue is that I've always told you what the total problem set scores are, but in reality this is the sum of the 11 individual problem set scores.

The file `Problem 6 Data.csv` contains data for a class of 80 students:

- Registration status
- The final raw score for each of the 11 problem sets
- The raw score for the Midterm Assessment
- The raw score for the Comprehensive Assessment

Use the grading system we discussed in Week 2 Module 4: Your Final Grade. However, for this problem you should use this grading schedule:

Range	Letter Grade
90 <= Score	A
80 <= Score < 90	B
70 <= Score < 80	C
60 <= Score < 70	D
Score < 60	F

Your challenge in this problem is to calculate the final letter grade for each of the students of this class.

When you're all done, directly display a relative proportions table of the letter grades, and then use this table to construct a barplot of the letter grade frequency counts. Remember to include all the required elements of the barplot.

You're on your own for this one! You'll have to decide how to implement this, especially how to obtain the total problem set raw scores.

Solution