

a

by Farley Staines

Submission date: 05-May-2024 04:36PM (UTC+0100)

Submission ID: 232247232

File name: 41624_Farley_Staines_a_565887_30834456.docx (796.78K)

Word count: 964

Character count: 4422

Data analysis

Farley staines

2204098

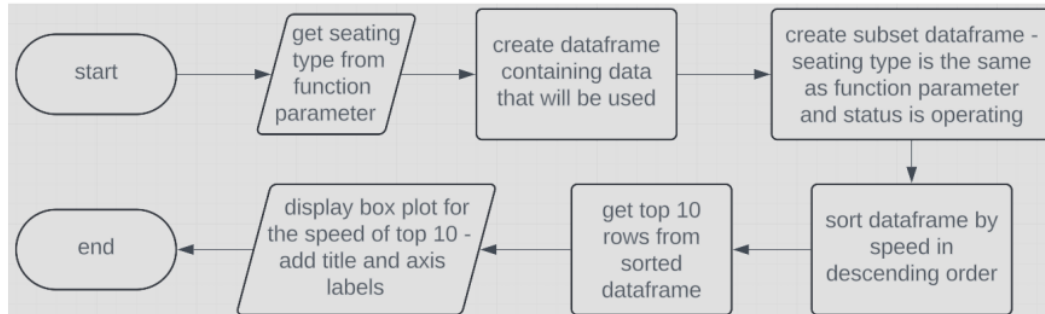
Contents

- 1 Documentation 2
 - 1.1 Question 3 2
 - 1.1.1 Flow diagram 2
 - 1.1.2 How it works 2
 - 1.1.3 Testing 2
 - 1.2 Question 4 3
 - 1.2.1 Flow diagram 3
 - 1.2.2 How it works 3
 - 1.2.3 Testing 3
 - 1.3 Question 5 4
 - 1.3.1 Flow diagram 4
 - 1.3.2 How it works 4
 - 1.3.3 Testing 4
- 2 Reflections 4

1 Documentation

1.1 Question 3

1.1.1 Flow diagram



1.1.2 How it works

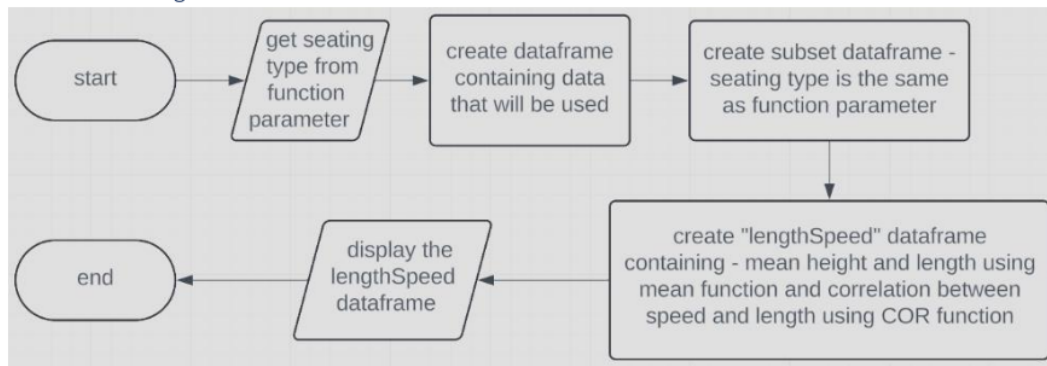
1. A value `seating_type` is passed into the function
2. New dataframe is created containing columns:
 - a. Name
 - b. Park
 - c. `seating_type`
 - d. speed
 - e. status
3. Create a subset of the dataframe where:
 - a. `seating_type` in the dataframe is the same as `seating_type` passed into the function
 - b. status in the dataframe is equal to `'status.operating'`
4. Order the dataframe by speed in descending order
5. Get the top 10 rows in the dataframe and add to it a new dataframe `'fastestRides'`
6. Create a bar chart using data from the dataframe `'fastestRides'`:
 - a. The main values are in the speed column
 - b. The title is `'Top 10 fastest rides'`
 - c. Label each value with the name column
 - d. Label horizontal axis
 - e. Label vertical axis

1.1.3 Testing

To test that the code was working correctly I created an excel sheet containing all the data. I then began filtering each column based on what data was or wasn't needed. I started by filtering the `seating_type` column to values that contained `'inverted'` and filtered status column with the value `'status.operating'`. After filtering the dataset I sorted the speed column in descending order so that the fastest speeds were at the top. I then compared the values in the spreadsheet to those that I was getting from the code, if they were any mismatching values I would adjust the code accordingly.

1.2 Question 4

1.2.1 Flow diagram



1.2.2 How it works

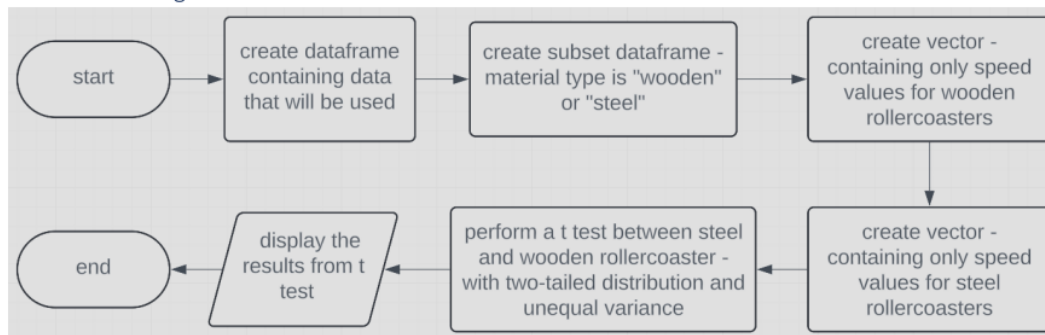
1. At the top a clean dataframe is created that has removed all rows that contains blank cells or NA values
2. A value `seating_type` is passed into the function
3. A new dataframe is created containing columns from the clean dataframe:
 - a. `seating_type`
 - b. `height`
 - c. `length`
 - d. `speed`
4. Create a subset of the dataframe where:
 - a. `seating_type` in the dataframe is the same as `seating_type` passed into the function
5. Create a new dataframe called 'lengthSpeed' containing:
 - a. The `seating_type` pass into the function
 - b. The mean height of the values in the dataframe
 - c. The mean length of the values in the dataframe
 - d. The correlation between the speed and length from values in the dataframe
6. Display the 'lengthSpeed' dataframe

1.2.3 Testing

To test that the code was working correctly I edited the excel sheet containing all the data. I then began by going through each column removing all the blank cells and cells that contains 'na' by filtering them out. I then filtered the `seating_type` column to values that contained 'Sit Down'. I then used the built-in function in excel to calculate the mean and correlation which I then used to compare to the values I got from my code.

1.3 Question 5

1.3.1 Flow diagram



1.3.2 How it works

1. At the top a clean dataframe is created that has removed all rows that contains blank cells or NA values
2. A new dataframe is created containing columns from the clean dataframe:
 - a. Material_type
 - b. speed
3. Create a subset of the dataframe where:
 - a. material_type is 'wooden' or 'steel'
4. Create a vector called 'woodenCoasters' from the dataframe containing:
 - a. Only speed values
 - b. Material_type is 'wooden'
5. Create a vector called 'steelCoasters' from the dataframe containing:
 - a. Only speed values
 - b. Material_type is 'steel'
6. Perform a t test between 'woodenCoasters' and 'steelCoasters':
 - a. The test is a 2-sided distribution
 - b. There is an unequal variance between the vectors
7. Display the results from the t test

1.3.3 Testing

To test that the code was working correctly I edited the excel sheet containing all the data. I then began by going through each column removing all the blank cells and cells that contains 'na' by filtering them out. I then filtered the material_type column to only have 'steel' or 'wooden' values. I then used the built-in function in excel to calculate the t-test with the correct parameters, which I then used to compare to the value I got from my code.

2 Reflections

Overall I didn't find it to challenging as each task was very linear in how it was going to work. The hardest part I found with the tasks was understanding some of the wording and what it was asking me to do. With task 5 I did struggle to perform the t-test as it had been a while since I last used it and I wasn't sure about what parameters I needed to use. I did some looking around for what parameters inputs I needed to use for this test and I couldn't find anything that gave me a definite answer for what I needed, so I asked ChatGPT for some help and it was able to give me a clear and

concise response for what t-test I need to use and its parameters like one or two tailed distributions. I didn't learn much from doing these tasks because most of the knowledge came from doing practise exercises and previous tasks, but I did learn how to perform a t-test. If I were to do this again I probably wouldn't have used function for each task since the code only loops through once and is very linear. If the tasks were more repetitive then I would have used functions more. Besides from not using functions there isn't anything else I would do differently if I had to do it again.

a

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5