STA 3100 Programming With Data in R: Data Manipulation

Data Manipulation in R

Key Concepts

Data manipulation is the process of changing data from one form to another. It is a key component of data analysis and can involve transforming, sorting, merging, and summarizing data. Data manipulation in R involves using functions and packages to manipulate data.

Definitions

Data Transformation: The process of changing the values of a variable or set of variables. This can include changing the format, type, or structure of the data.

Data Sorting: The process of arranging data in a specific order. This can be done by sorting on one or more variables.

Data Merging: The process of combining two or more datasets into one dataset. This can be done by joining on one or more variables.

Data Summarization: The process of reducing a dataset to a smaller set of summary statistics. This can be done by calculating the mean, median, mode, etc. of a variable.

Coding Examples

Example 1: Data Transformation

```
Start of Code
```R
# Transform the character vector to a numeric vector
char_vec <- c("1", "2", "3")
num_vec <- as.numeric(char_vec)
# Print the new vector
print(num_vec)
```
End of Code
```

Output:

```
[1] 1 2 3
```

Example 2: Data Sorting

Start of Code

```R
# Create a data frame
df <- data.frame(x = c(3, 2, 1), y = c("b", "a", "c"))
# Sort the data frame by the x column
sorted_df <- df[order(df$x), ]
# Print the sorted data frame
print(sorted_df)
```

End of Code

Output:

```
  x y
3 1 c
2 2 a
1 3 b
```

Example 3: Data Merging

Start of Code

```R
# Create two data frames
df1 <- data.frame(x = c(1, 2, 3), y = c("a", "b", "c"))
df2 <- data.frame(x = c(2, 3, 4), z = c("d", "e", "f"))
# Merge the two data frames on the x column
merged_df <- merge(df1, df2, by = "x")
# Print the merged data frame
print(merged_df)
```

End of Code

Output:

```
  x y z
1 2 b d
2 3 c e
3 4 NA f
```

```

### Example 4: Data Summarization

Start of Code
```R
# Create a data frame
df <- data.frame(x = c(1, 2, 3))

# Calculate the mean of the x column
mean_x <- mean(df$x)

# Print the mean
print(mean_x)
```
End of Code

Output:

```

[1] 2

```

Practice Multiple Choice Questions

Q1. What is data manipulation?

A. The process of changing data from one form to another

B. The process of analyzing data

C. The process of collecting data

D. The process of visualizing data

Answer: A. The process of changing data from one form to another
```