Leadership

R Programming

Descriptive analytics & Pre-Processing

Naveen Kumar

ChooseMIS
Major in the Future

Agenda

- Week 4 Summary and Business Apps Presentation
- Review
- Getting Familiar with Datasets
- Character Strings and Business Applications
- Questions and Summary

GETTING FAMILIAR WITH DATASETS

Manipulating the Dataset

Deleting columns using the function "subset“

•subset() function can also be used to remove columns from the dataset

•**What is a subset?**

A is said to be subset of B, if elements(rows and columns) in A are contained in B.

```
med.data <- subset(med.data, select = -c ("ChargesInDollars", "VisitTimeInMin"))
```

Manipulating the Dataset

Displaying and changing column names

colnames(): Returns the names of columns in the object (matrix, dataframe etc)

#Display column names

colnames(med.data)

```
> colnames(med.data)
[1] "ID"          "AgeInYears"  "Gender"      "Insurance"   "PriorVisits" "Date"
```

#Displaying the first column name

colnames(med.data)[1]

```
> colnames(med.data)[1]
[1] "ID"
```

#Setting the 4th column name to “InsuranceUpdated”

colnames(med.data)[4] <- "InsuranceUpdated"

ID	AgeInYears	Gender	Insurance

ID	AgeInYears	Gender	InsuranceUpdated

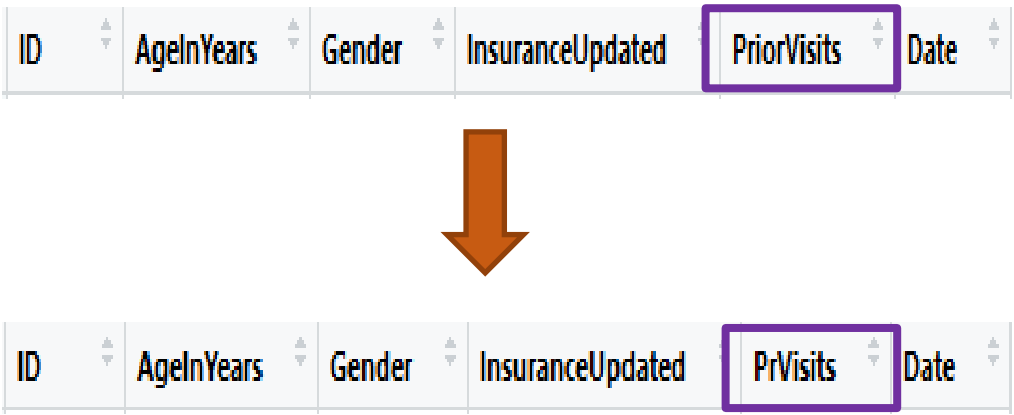
Manipulating the Dataset

Displaying and Changing column names

#Change column name (from "PriorVisits" to "PrVisits")

Here, we are updating the “PriorVisits” column name to “PrVisits” using which() function. Using which() function we are trying to identify the column with name “PriorVisits“. Later we are renaming it as “PrVisits”

```
colnames(med.data)[which(colnames(med.data) == "PriorVisits")] <-
"PrVisits"
```



Filtering the Data

#Filtering data by one condition

Below statement returns only those rows with PrVisits greater than or equal to 70.

med.data[med.data$PrVisits >= 70,]

	ID	AgeInYears	Gender	InsuranceUpdated	PrVisits	Date
16	16	64	M	<NA>	72	4/25/2014
7	7	62	F	BCBS	120	2/14/2014
13	13	61	M	Self Pay	70	4/18/2014
9	9	60	F	Private	107	3/25/2014
3	3	56	F	Medicaid	81	1/15/2014
15	15	54	M	Self Pay	111	4/21/2014
19	19	35	M	Private	107	5/15/2014
20	20	32	F	Private	101	5/25/2014
14	14	28	F	Self Pay	109	4/20/2014

Filtering the Data

#Filtering data by more than one condition

Below statement returns only those rows with AgeInYears greater than or equal to 40 and less than 70.

```
med.data[(med.data$AgeInYears >= 40) & (med.data$AgeInYears < 70),]
```

	ID	AgeInYears	Gender	InsuranceUpdated	PrVisits	Date
11	11	65	M	Private	61	4/2/2014
16	16	64	M	<NA>	72	4/25/2014
7	7	62	F	BCBS	120	2/14/2014
10	10	61	F	Private	51	3/28/2014
13	13	61	M	Self Pay	70	4/18/2014
9	9	60	F	Private	107	3/25/2014
12	12	60	F	Medicaid	42	4/8/2014
3	3	56	F	Medicaid	81	1/15/2014
15	15	54	M	Self Pay	111	4/21/2014
4	4	53	M	BCBS	31	1/25/2014
5	5	51	F	Private	48	2/5/2014
17	17	44	F	BCBS	57	4/28/2014

Introduction To Character String And Its Applications

Overview

- How to get a text string
- How to split a given string by space + limitation
- How to manipulate the date

Replace Missing Data When Data Type Is Character

- Create new column having Datatype as character

med.data$Insurance.char <- as.character(med.data$InsuranceUpdated)
typeof(med.data$Insurance.char)

```
> med.data$Insurance.char <- as.character(med.data$InsuranceUpdated)
> typeof(med.data$Insurance.char)
[1] "character"
```

Insurance.char
Private
BCS

- Replace with “BCS”

The generic function is.na() indicates which elements are missing in the specified columns

med.data$Insurance.char[is.na(med.data$Insurance.char)] <- "BCS"

How To Get A Text String

Substring (extract) the last two characters of the string

- substr() function is used to extract a part or specific number of characters from a string
- **Syntax for substr() function:** substr(dataset$columnvalue, startingdigit, endingdigit)

Determining the number of characters nchar()

- nchar returns the size of a character vector
- In the below example, we are trying to identify the number of characters in the value present in 4th row of the Insurance.char column

nchar(med.data$Insurance.char)[4]

```
> nchar(med.data$Insurance.char)[4]
[1] 7
```

How To Get A Text String

Substring (extract) the last two characters of the string

- **Syntax for substr() function:** substr(dataset$columnvalue, startingdigit, endingdigit)

- In the below case, we are trying to identify the last two characters of the value present in 4th row Insurance.char column

substr(med.data$Insurance.char[4],nchar(as.character(med.data$Insurance.char[4]))-1, nchar(as.character(med.data$Insurance.char[4])))

```
> substr(med.data$Insurance.char[4], nchar(as.character(med.data$Insurance.char[4]))-1,
+          nchar(as.character(med.data$Insurance.char[4])))
[1] "te"
```

How To Get A Text String

```
> substr(med.data$Insurance.char[4], nchar(as.character(med.data$Insurance.char[4]))-1,
+          nchar(as.character(med.data$Insurance.char[4])))
[1] "te"
```

Considering the syntax of the substr(),

- Dataset$columnvalue here is med.data$Insurance.char[4] as we are determining the substring for the Insurance.char value "**Private**"

- As we want the last two characters, our starting digit will be last but one character position

- nchar(as.character(med.data$Insurance.char[4]))-1: Considers the number of characters in "**Private**" which is **7** and subtracts **1** from it. So the starting character will be **"t"**

- nchar(as.character(med.data$Insurance.char[4]: Determines the number of characters in "**Private**" which is **7.** Hence the ending digit will be **7** and the character in **7**th place is **"e"**

How To Split A Given String By Space + Limitation

strsplit() function splits the string based on specific condition given. In the below example, strsplit() splits the string value based on the space i.e., “ “

strsplit(med.data$Insurance.char[5], " ")

```
> strsplit(med.data$Insurance.char[5], " ")
[[1]]
[1] "Self" "Pay"
```

- **Limitation**: Sometimes strsplit doesn’t work the way you want it to; the above code splits the string at *every* space.
  - “West Virginia WV” becomes “West” “Virginia” “WV”, instead of “West Virginia” “WV”

Summary and Questions