summarize in r, when we have a dataset and need to get a clear idea about each parameter then a summary of the data is important. Summarized data will provide a clear idea about the data set.

In this tutorial we are going to talk about summarize () function from dplyr package. Summarizing a data set by group gives better indication on the distribution of the data.

This tutorial you will get the idea about summarise(), group\_by summary and important functions in summarise()

**Load Library**

library(dplyr)

Let’s load iris data set for summarization. Let’s store the iris data set into new variable say **df** for summarize in r.

df<-iris

df1<-summarise(df, mean(Sepal.Length())

Output:-

mean(Sepal.Length)

5.843333

Let’s create mean and sd of Sepal Length.

df2<-summarise(df, Mean=mean(Sepal.Length(),

SD=sd(Sepal.Length())

Output:-

Mean SD

5.843333 0.8280661

Now we try to summarize based on groups.

df3<-summarise(group\_by(df, Species),

Mean=mean(Sepal.Length(),

SD=sd(Sepal.Length())

Output:-

Species Mean SD

1 setosa 5.01 0.352

2 versicolor 5.94 0.516

3 virginica 6.59 0.636

You can make use of pipe operator for summarising the data set.

Pipe operator comes under magrittr package. Let’s load the package.

library(magrittr)

df4<-df %>%

  group\_by(Species) %>%

  summarise(Mean = mean(Sepal.Length),

            SD=sd(Sepal.Length))

Output:-

Species Mean SD

1 setosa 5.01 0.352

2 versicolor 5.94 0.516

3 virginica 6.59 0.636

Based on pipe operator you can easily summarize and plot it with the help of ggplot2.

library(ggplot2)

For plotting the datset we have main four steps

Step 1: Select the appropriate data frame

Step 2: Group the data frame

Step 3: Summarize the data frame

Step 4: Plot the summary statistics based on your requirement

df %>%

  group\_by(Species) %>%

  summarise(Mean = mean(Sepal.Length)) %>%

  ggplot(aes(x = Species, y = Mean, fill = Species)) +

  geom\_bar(stat = "identity") +

  theme\_classic() +

  labs(

    x = "Species",

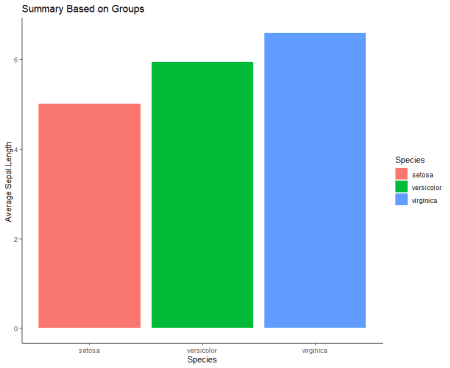
    y = "Average Sepal.Length ",

    title = paste(

      "Summary Based on Groups"

    )

  )



**Sum**

Another useful function to aggregate the variable is sum().

df5<-df %>%

  group\_by(Species) %>%

  summarise(sum = sum(Sepal.Length),

            SD=sd(Sepal.Length))

Output:-

Species sum SD

1 setosa 250 0.352

2 versicolor 297 0.516

3 virginica 329 0.636

**Minimum and maximum**

Find the minimum and the maximum of a vector or variable with the help of function min() and max().

df6<-df %>%

  group\_by(Species) %>%

  summarise(Min = min(Sepal.Length),

            Max=max(Sepal.Length))

Output:-

Species Min Max

1 setosa 4.3 5.8

2 versicolor 4.9 7

3 virginica 4.9 7.9

**Count**

Suppose if you want to count observations by group you can aggregate the number of occurrence with n().

df7<-df %>%

  group\_by(Species) %>%

  summarise(Sepal.Length = n())%>%

            arrange(desc(Sepal.Length))

Output:-

Species Sepal.Length

1 setosa 50

2 versicolor 50

3 virginica 50

**First and Last**

Some cases first cases or position identification is important, then you can make use of first, last or nth position of a group.

df8<-df %>%

  group\_by(Species) %>%

  summarise(First = first(Sepal.Length),

            Last=last(Sepal.Length))

Output:-

Species First Last

1 setosa 5.1 5

2 versicolor 7 5.7

3 virginica 6.3 5.9

The same way you can make use of following functions some of the functions already covered in the tutorial.

You can see the important functions below for summarizing the dataset.

**Mean**

summarise(df,mean = mean(x1))

**Median**

summarise(df,median = median(x1))

**Sum**

summarise(df,sum = sum(x1))

**Standard** **Deviation**

summarise(df,sd = sd(x1))

**Interquartile**

summarise(df,interquartile = IQR(x1))

**Minimum**

summarise(df,minimum = min(x1))

**Maximum**

summarise(df,maximum = max(x1))

**Quantile**

summarise(df,quantile = quantile(x1))

**First** **Observation**

summarise(df,first = first(x1))

**Last** **observation**

summarise(df,last = last(x1))

**nth** **observation**

summarise(df,nth = nth(x1, 2))

**Number** **of** **occurrence**

summarise(df,count = n(x1))

**Number** **of distinct occurrence**

summarise(df,distinct = n\_distinct(x1))