21bcs022-ex5

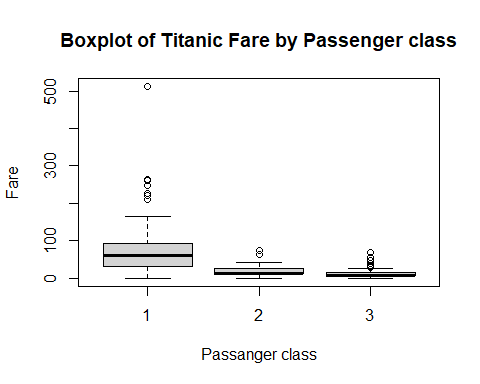
Nagajothi

2024-03-13

## R Markdown

**6.Consider a Titanic dataset and compare metric values across #different subgroups of the data. Also assume you have a greater number of groups, which visualization method do you prefer over a column chart.**

library(readr)  
  
 Titanic<-read.csv("E:/jothi/Titanic (1).csv")   
  
boxplot(Fare ~ Pclass,data=Titanic,main="Boxplot of Titanic Fare by Passenger class",xlab="Passanger class",ylab="Fare")



**7.Consider the below given Data-set and plot different charts.**

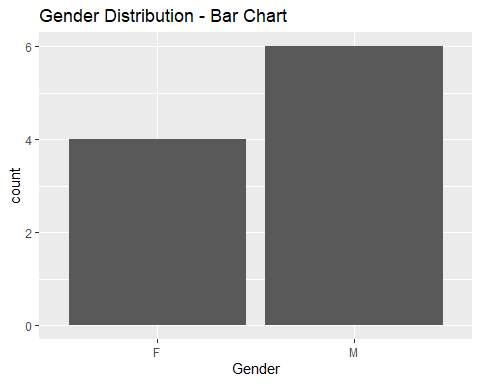
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.3

df<-data.frame(  
 EMPID=paste0("EMP",1:10),  
 Gender=c("M","F","F","M","F","M","M","F","M","M"),  
 Age=c(34,40,37,30,44,36,32,26,32,36),  
 Sales=c(123,114,135,139,117,121,133,140,133,133),  
 BMI=c("Normal","OverWeight","Obesity","UnderWeight",  
 "UnderWeight","Normal","Obesity","Normal","Normal","UnderWeight"),  
 Income=c(350,450,169,189,183,80,166,120,75,40)  
)

**#Create Bar Chart:**

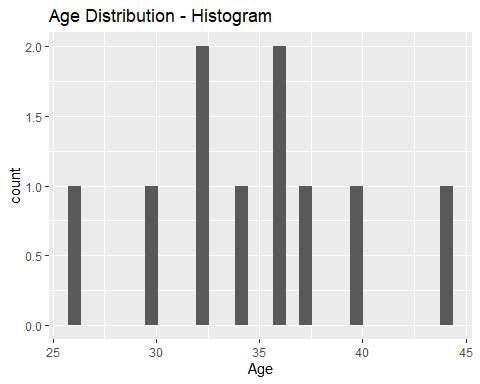
ggplot(df,aes(x=Gender)) + geom\_bar()+ggtitle("Gender Distribution - Bar Chart")



**#Create Histogram:**

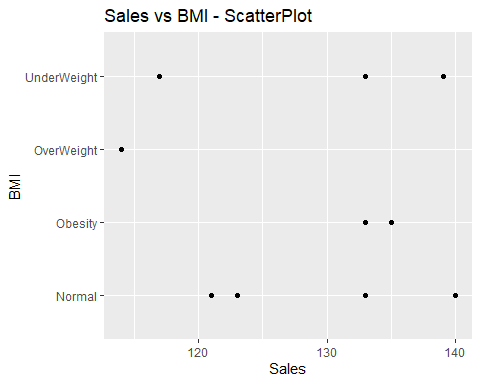
ggplot(df,aes(x=Age)) + geom\_histogram()+ggtitle("Age Distribution - Histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



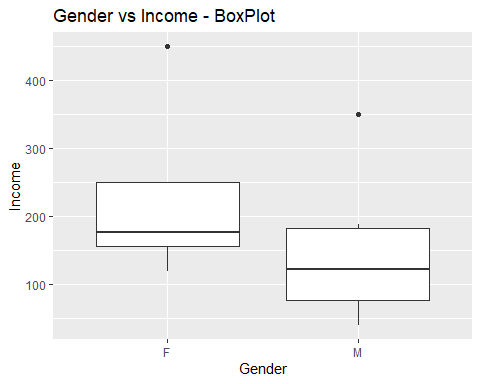
**#Create Scatterplot:**

ggplot(df,aes(x=Sales,y=BMI)) + geom\_point()+ggtitle("Sales vs BMI - ScatterPlot")



**#Create Boxplot:**

ggplot(df,aes(x=Gender,y=Income)) + geom\_boxplot()+ggtitle("Gender vs Income - BoxPlot")



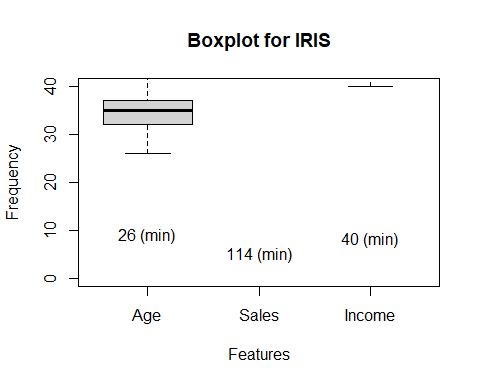
**#8.Consider the above dataset and draw the box plot for the**

**#statistical data based on the minimum, #first quartile, median, third quartile, and maximum**

library(ggplot2)  
b<-subset(df,select=c(-EMPID,-Gender,-BMI))  
b<-boxplot(b,ylim=c(0,40),main="Boxplot for IRIS",xlab="Features",ylab="Frequency")  
b

## $stats  
## [,1] [,2] [,3]  
## [1,] 26 114 40.0  
## [2,] 32 121 80.0  
## [3,] 35 133 167.5  
## [4,] 37 135 189.0  
## [5,] 44 140 350.0  
##   
## $n  
## [1] 10 10 10  
##   
## $conf  
## [,1] [,2] [,3]  
## [1,] 32.5018 126.005 113.0393  
## [2,] 37.4982 139.995 221.9607  
##   
## $out  
## [1] 450  
##   
## $group  
## [1] 3  
##   
## $names  
## [1] "Age" "Sales" "Income"

text(x=1,y=80,paste(max(df$Age),"(max)"))  
text(x=2,y=180,paste(max(df$Sales),"(max)"))  
text(x=3,y=380,paste(max(df$Income),"(max)"))  
text(x=1,y=9,paste(min(df$Age),"(min)"))  
text(x=2,y=5,paste(min(df$Sales),"(min)"))  
text(x=3,y=8,paste(min(df$Income),"(min)"))  
text(x=1,y=60,paste(median(df$Age),"(median)"))  
text(x=2,y=160,paste(median(df$Sales),"(median)"))  
text(x=3,y=360,paste(median(df$Income),"(median)"))  
text(x=1,y=100,paste(quantile(df$Age,0.25),"(First Quantile)"))  
text(x=2,y=200,paste(quantile(df$Sales,0.25),"(First Quantile)"))  
text(x=3,y=400,paste(quantile(df$Income,0.25),"(First Quantile)"))  
text(x=1,y=120,paste(quantile(df$Age,0.75),"(Third Quantile)"))  
text(x=2,y=220,paste(quantile(df$Sales,0.75),"(Third Quantile)"))  
text(x=3,y=340,paste(quantile(df$Income,0.75),"(Third Quantile)"))

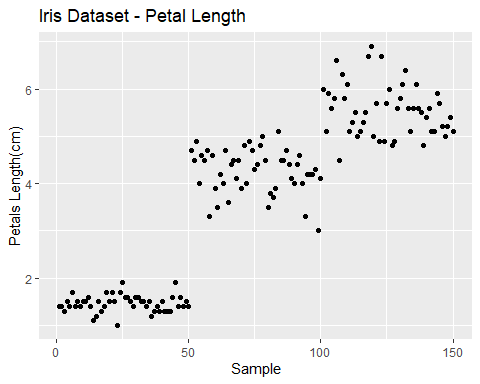


**9.Make a scatterplot for the features in the Iris dataset.**

library(ggplot2)  
install.packages("datasets")

## Warning: package 'datasets' is in use and will not be installed

library(datasets)  
feature<-iris$Petal.Length  
ggplot(iris,aes(x=1:nrow(iris),y=feature)) +  
 geom\_point() +  
 xlab("Sample") +  
 ylab("Petals Length(cm)") +  
 ggtitle("Iris Dataset - Petal Length")

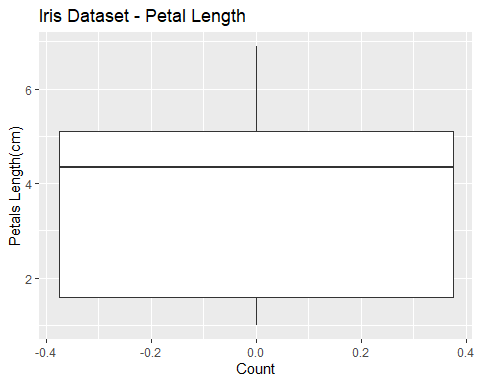


**10.Look at an individual feature in Seaborn through a boxplot.**

install.packages("ggplot2")

## Warning: package 'ggplot2' is in use and will not be installed

library(ggplot2)  
library(datasets)  
feature<-iris$Petal.Length  
ggplot(iris,aes(y=feature)) +  
 geom\_boxplot() +  
 xlab("Count") +  
 ylab("Petals Length(cm)") +  
 ggtitle("Iris Dataset - Petal Length")



**#Create a synthetic dataset of 100 entries (randomly generated) #with the given fields and store it in a .csv file:**

**#Generate synthetic data**

data <- data.frame(  
 StudentId = 1:100,  
 Dept = sample(c('CSE', 'EEE', 'ECE', 'CIV', 'BME'), 100, replace=TRUE),  
 Sem = factor(sample(1:8, 100, replace=TRUE)),  
 GPA = round(runif(100, min=2.0, max=10.0),3)  
)  
# Save data to CSV  
write.csv(data, file = "ex4-studata.csv", row.names = FALSE)  
  
#1. Scatterplot with smooth curve for every semester   
#with Sized points for every Department  
library(ggplot2)  
library("dplyr")

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

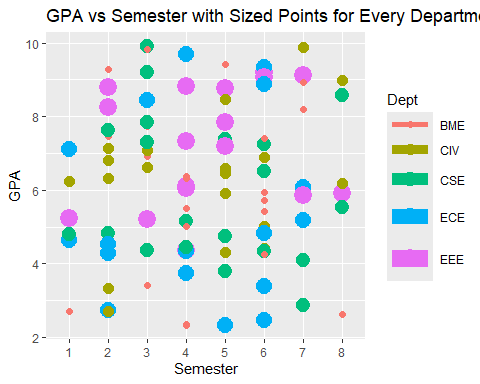
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

ggplot(data, aes(x = Sem, y = GPA, color = Dept, size = Dept)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(title = "GPA vs Semester with Sized Points for Every Department",  
 x = "Semester",  
 y = "GPA")

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

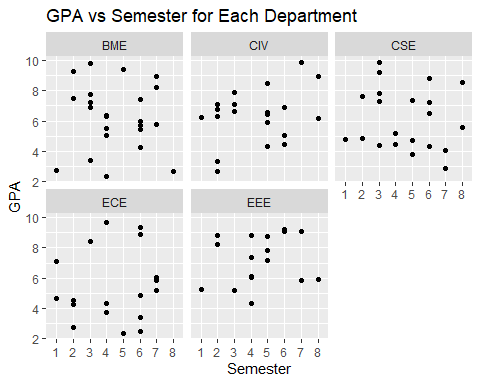
## Warning: Using size for a discrete variable is not advised.

## `geom\_smooth()` using formula = 'y ~ x'



**#2. Draw subsets of scatterplot to plot GPA for each department. Scatterplot for GPA vs Semester with subsets for each department**

ggplot(data, aes(x = Sem, y = GPA)) +  
 geom\_point() +  
 facet\_wrap(~ Dept) +  
 labs(title = "GPA vs Semester for Each Department",  
 x = "Semester",  
 y = "GPA")



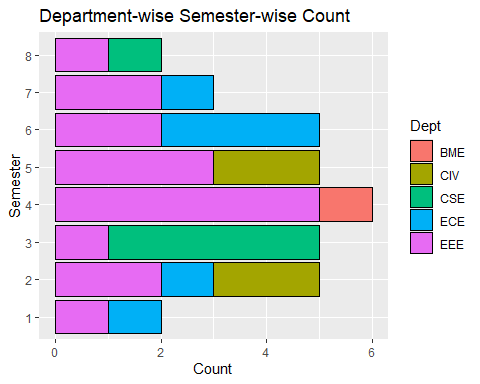
**#3. Overlapping stacked bar chart represents department-wise-semester-wise count. #(Horizontal bar chart)**

# Generate department-wise-semester-wise count data  
count\_data <- data %>% group\_by(Dept, Sem) %>% summarise(count = n()) %>% arrange(Dept, Sem)

## `summarise()` has grouped output by 'Dept'. You can override using the  
## `.groups` argument.

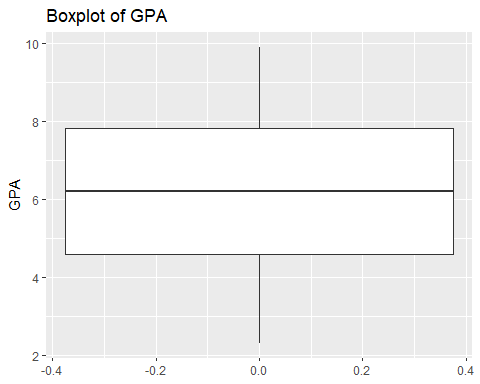
# Create the overlapping stacked bar chart  
ggplot(count\_data, aes(x = count, y = Sem, fill = Dept)) +  
 geom\_bar(stat = "identity", position = "identity", color = "black", horiz=TRUE) +  
 labs(title = "Department-wise Semester-wise Count",  
 x = "Count",  
 y = "Semester")

## Warning in geom\_bar(stat = "identity", position = "identity", color = "black",  
## : Ignoring unknown parameters: `horiz`



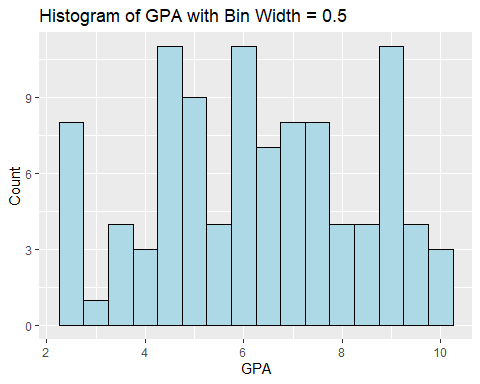
#**4.Identify the outlier’s using boxplot Create a boxplot to identify outliers in GPA**

ggplot(data, aes(y = GPA)) +  
 geom\_boxplot() +  
 labs(title = "Boxplot of GPA",  
 y = "GPA")



**#5. Draw histograms for count of GPA with different bin width & zoom to see any one region**

# To Create histograms for GPA with different bin widths  
ggplot(data, aes(x = GPA)) +  
 geom\_histogram(binwidth = 0.5, fill = "lightblue", color = "black") +  
 labs(title = "Histogram of GPA with Bin Width = 0.5",  
 x = "GPA",  
 y = "Count")



#To Zoom in to see any one region  
ggplot(data, aes(x = GPA)) +  
 geom\_histogram(binwidth = 0.5, fill = "lightblue", color = "black") +  
 xlim(8, 10) +   
 labs(title = "Zoomed-in Histogram of GPA (8 to 10)",  
 x = "GPA",  
 y = "Count")

## Warning: Removed 77 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_bar()`).

