APS Odd Semester 2024 Coding Problem Set-2

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

# Load the house price dataset  
hData = read.csv('houseprices.csv', header = TRUE, stringsAsFactors = FALSE, na.strings = c("", "NA", "Not Available", "not available"))  
str(hData)

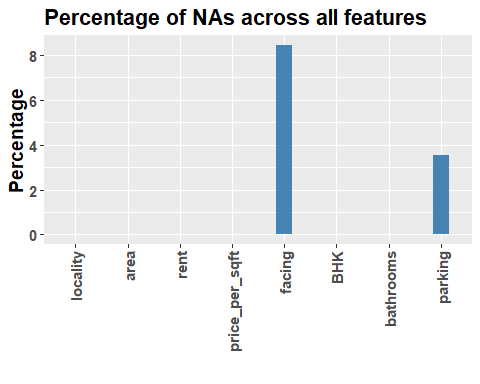
## 'data.frame': 225 obs. of 8 variables:  
## $ locality : chr "BTM Layout" "BTM Layout" "BTM Layout" "BTM Layout" ...  
## $ area : int 565 1837 1280 2220 1113 1332 1815 1400 3006 1600 ...  
## $ rent : int 20060 97434 54448 117000 34388 36394 112000 41266 129000 92849 ...  
## $ price\_per\_sqft: int 6195 9254 7422 9234 5391 4767 10744 5143 7485 10125 ...  
## $ facing : chr "North-West" "East" "East" "North" ...  
## $ BHK : int 1 3 2 3 2 2 3 2 4 3 ...  
## $ bathrooms : int 1 3 2 3 2 2 2 2 5 2 ...  
## $ parking : chr "Bike" "Bike and Car" "Car" "Bike and Car" ...

# Convert 'locality', 'facing' and 'parking' columns to factors  
categorical\_cols = c('locality', 'facing', 'parking')  
hData[categorical\_cols] = lapply(hData[categorical\_cols], as.factor)  
str(hData)

## 'data.frame': 225 obs. of 8 variables:  
## $ locality : Factor w/ 9 levels "Attibele","BTM Layout",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ area : int 565 1837 1280 2220 1113 1332 1815 1400 3006 1600 ...  
## $ rent : int 20060 97434 54448 117000 34388 36394 112000 41266 129000 92849 ...  
## $ price\_per\_sqft: int 6195 9254 7422 9234 5391 4767 10744 5143 7485 10125 ...  
## $ facing : Factor w/ 7 levels "East","North",..: 4 1 1 2 1 7 3 6 1 5 ...  
## $ BHK : int 1 3 2 3 2 2 3 2 4 3 ...  
## $ bathrooms : int 1 3 2 3 2 2 2 2 5 2 ...  
## $ parking : Factor w/ 3 levels "Bike","Bike and Car",..: 1 2 3 2 2 2 3 2 2 2 ...

# Continuous columns  
continuous\_cols = setdiff(colnames(hData), categorical\_cols)

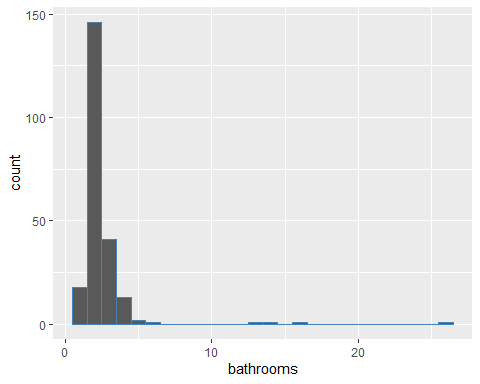
# Plot percentage of NAs in each column of the data frame  
hData\_NA = setNames(stack(sapply(hData, function(x){(sum(is.na(x))/length(x))\*100}))[2:1], c('Feature','Value'))  
p = ggplot(data = hData\_NA, aes(x = Feature, y = Value)) +  
 geom\_bar(stat = 'identity', fill = 'steelblue', width = 0.3) +  
 theme(text = element\_text(size = 14, face = 'bold'),  
 axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5)) +  
 xlab('') + ylab('Percentage') +  
 ggtitle('Percentage of NAs across all features')  
p



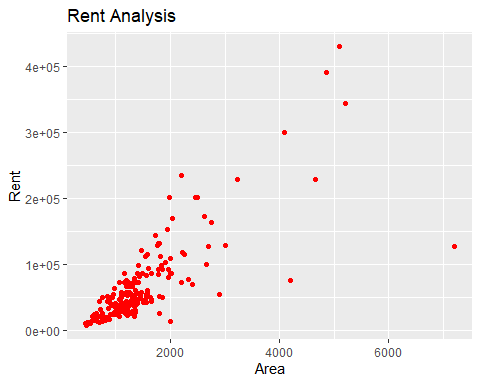
# Add NA as a factor level for categorical columns   
hData[categorical\_cols] = lapply(hData[categorical\_cols], addNA)  
str(hData)

## 'data.frame': 225 obs. of 8 variables:  
## $ locality : Factor w/ 10 levels "Attibele","BTM Layout",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ area : int 565 1837 1280 2220 1113 1332 1815 1400 3006 1600 ...  
## $ rent : int 20060 97434 54448 117000 34388 36394 112000 41266 129000 92849 ...  
## $ price\_per\_sqft: int 6195 9254 7422 9234 5391 4767 10744 5143 7485 10125 ...  
## $ facing : Factor w/ 8 levels "East","North",..: 4 1 1 2 1 7 3 6 1 5 ...  
## $ BHK : int 1 3 2 3 2 2 3 2 4 3 ...  
## $ bathrooms : int 1 3 2 3 2 2 2 2 5 2 ...  
## $ parking : Factor w/ 4 levels "Bike","Bike and Car",..: 1 2 3 2 2 2 3 2 2 2 ...

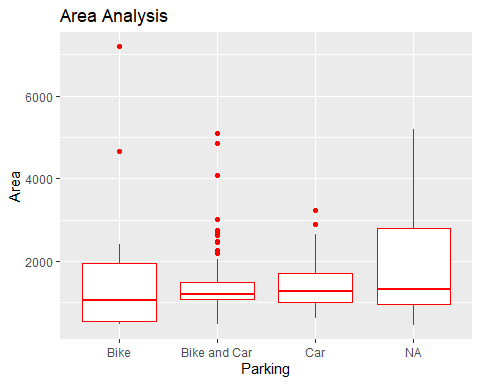
p = ggplot(data = hData) +  
 geom\_histogram(aes(x = bathrooms, y= after\_stat(count)), binwidth = 1, color = 'steelblue')  
p



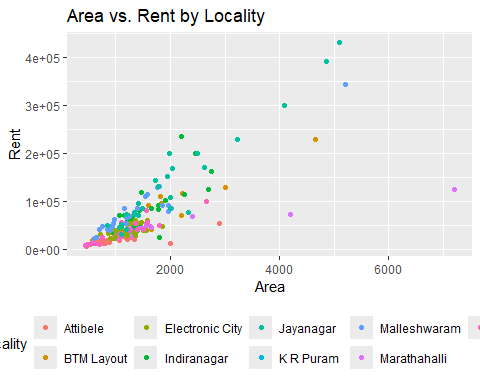
#Rent distribution by area  
p = ggplot(data = hData) +  
 geom\_point(aes(x = area, y = rent), color = 'red') +  
 labs(x = 'Area', y = 'Rent', title = 'Rent Analysis')  
p



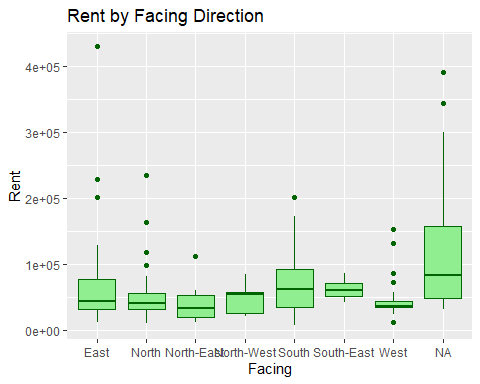
# Area Distribution by Parking Availability using Box Plot:  
p = ggplot(data = hData) +  
 geom\_boxplot(aes(x = parking, y = area), color = 'red') +  
 labs(x = 'Parking', y = 'Area', title = 'Area Analysis')  
p



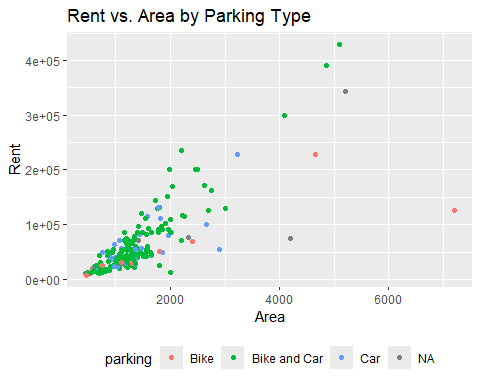
p = ggplot(data = hData, aes(x = area, y = rent, color = locality)) +  
 geom\_point() +  
 labs(x = 'Area', y = 'Rent', title = 'Area vs. Rent by Locality') +  
 theme(legend.position = "bottom")  
p



#Rent distribution by Facing Direction  
p = ggplot(data = hData, aes(x = facing, y = rent)) +  
 geom\_boxplot(fill = 'lightgreen', color = 'darkgreen') +  
 labs(x = 'Facing', y = 'Rent', title = 'Rent by Facing Direction')  
p



#Rent VS Area by parking type  
p = ggplot(data = hData, aes(x = area, y = rent, color = parking)) +  
 geom\_point() +  
 labs(x = 'Area', y = 'Rent', title = 'Rent vs. Area by Parking Type') +  
 theme(legend.position = "bottom")  
p



# Plot rent per square foot by locality  
p = ggplot(data = hData, aes(x = locality, y = price\_per\_sqft)) +  
 geom\_boxplot(fill = 'green', color = 'darkred') +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 labs(x = 'Locality', y = 'Rent per Square Foot', title = 'Rent per Square Foot by Locality')  
p

