Appendix

Statistics Final project

2023-11-27

```r
mobile_dataset<-read.csv("MobilePrice.csv",header = TRUE)
names(mobile_dataset)
```

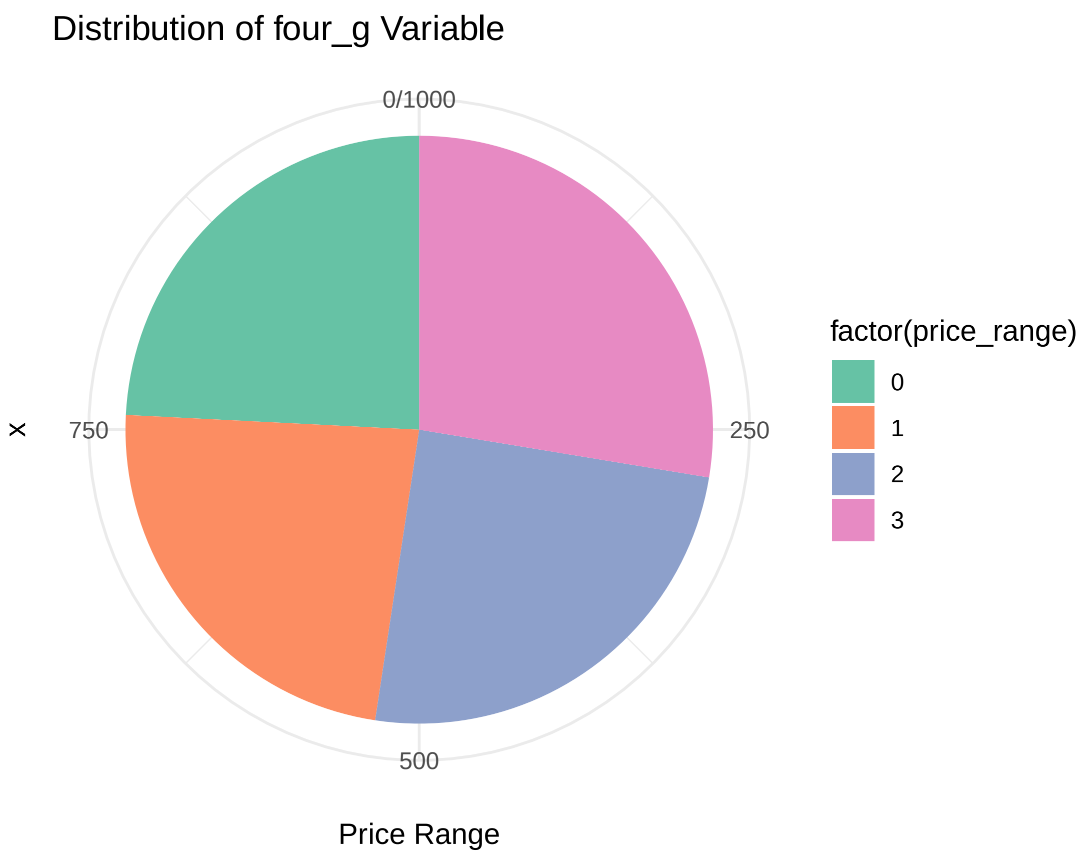
```
##  [1] "battery_power" "blue"          "clock_speed"   "dual_sim"
##  [5] "fc"            "four_g"        "int_memory"    "m_dep"
##  [9] "mobile_wt"     "n_cores"       "pc"            "px_height"
## [13] "px_width"      "ram"           "sc_h"          "sc_w"
## [17] "talk_time"     "three_g"       "touch_screen"  "wifi"
## [21] "price_range"
```

Pie Chart

```r
library(ggplot2)
```

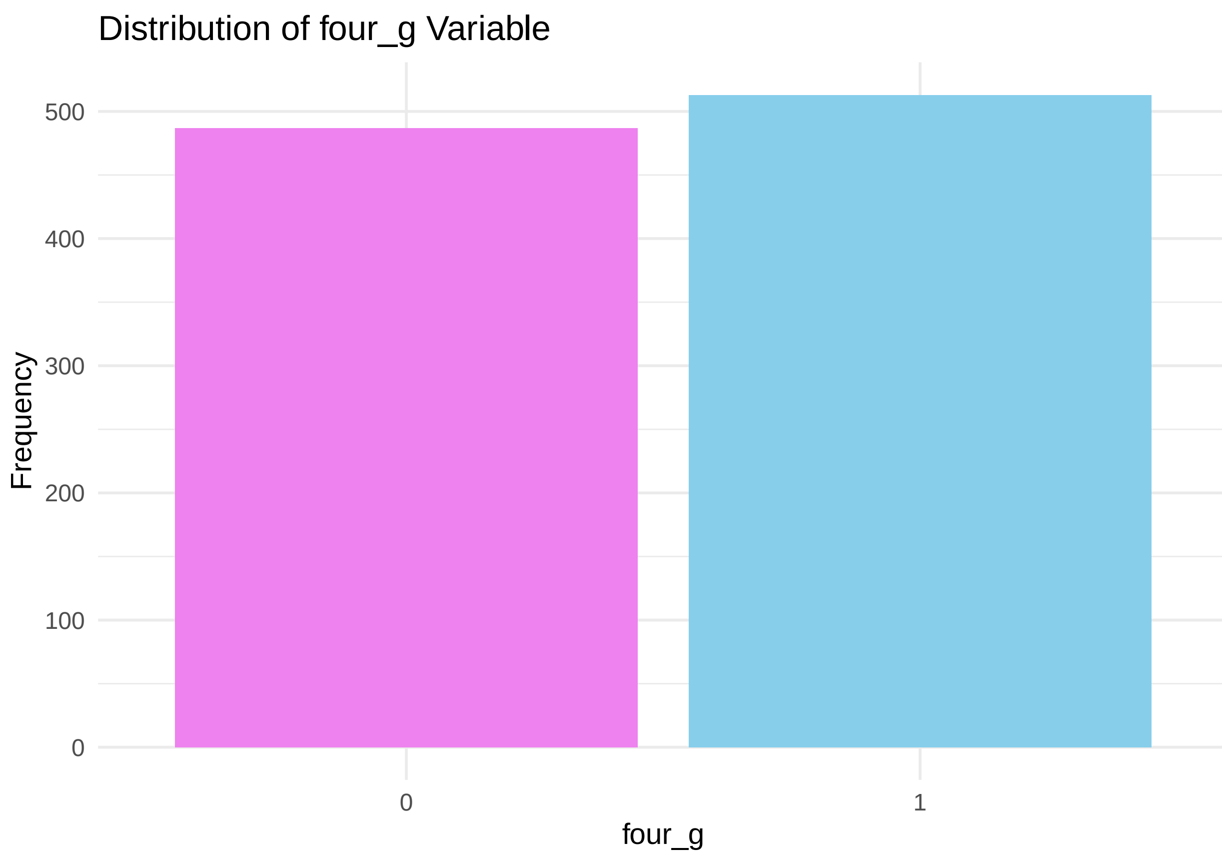
```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```r
ggplot(mobile_dataset, aes(x = "", fill = factor(price_range))) +
  geom_bar(width = 1, stat = "count") +
  coord_polar("y") +
  labs(title = "Distribution of Price Ranges") +
  theme_void() +
  scale_fill_manual(values = c("#66c2a5", "#fc8d62", "#8da0cb", "#e78ac3"))+
  labs(title = "Distribution of four_g Variable") +
  ylab("Price Range") +
  theme_minimal()
```



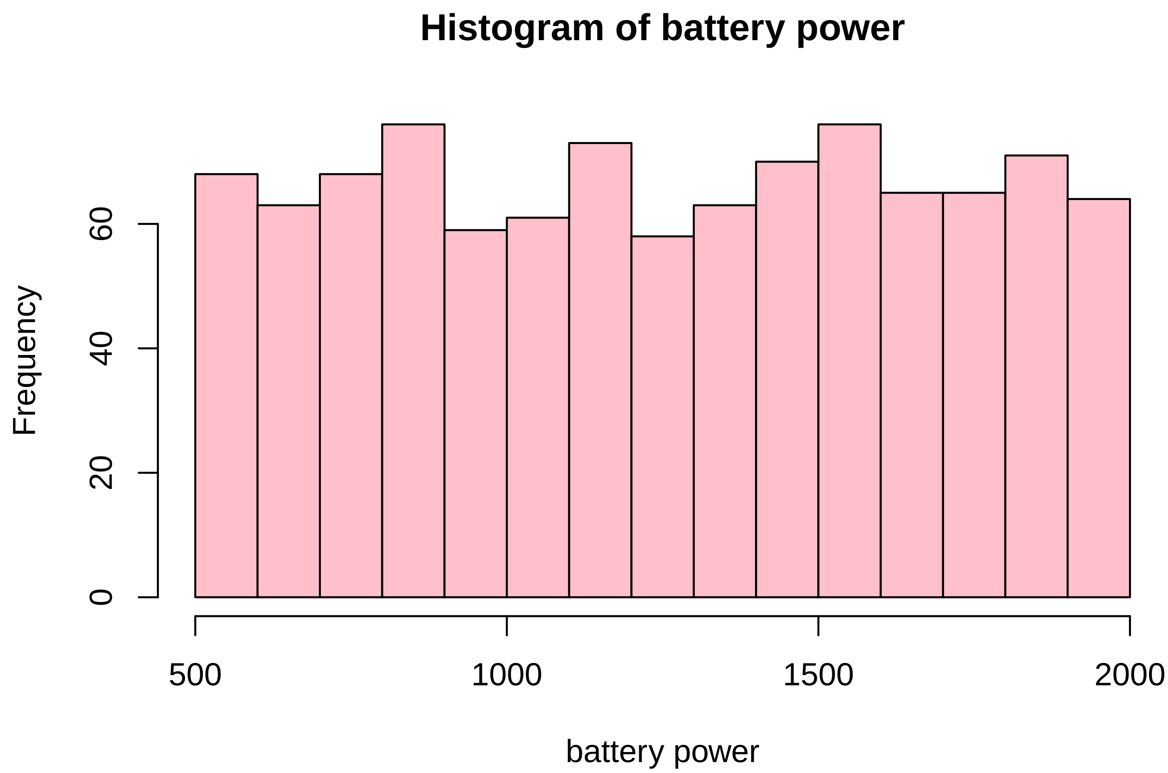
BAR CHART:

```
ggplot(mobile_dataset, aes(x = factor(four_g))) +
  geom_bar(fill = c("violet", "skyblue")) +  # Set colors for 1 and 0
  labs(title = "Distribution of four_g Variable") +
  xlab("four_g") +
  ylab("Frequency") +
  theme_minimal()
```

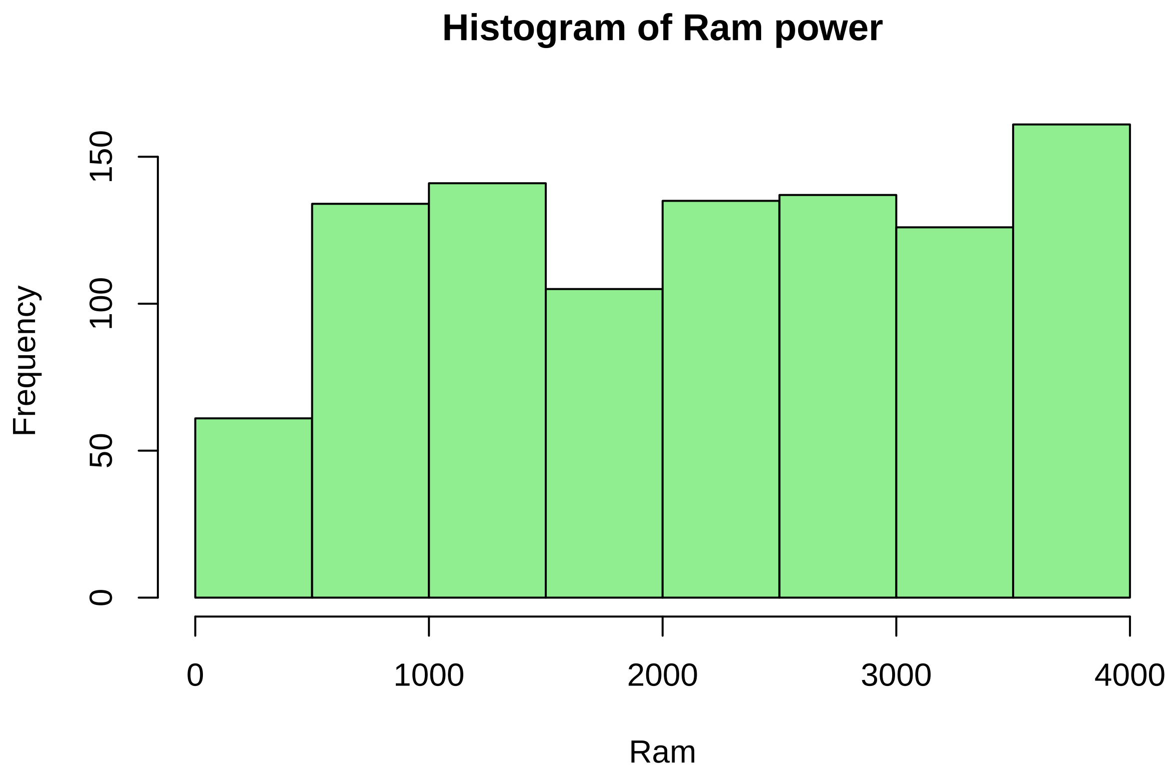


Histogram:

```
hist(mobile_dataset$battery_power,main = "Histogram of battery power",
     xlab = "battery power",col ="pink")
```

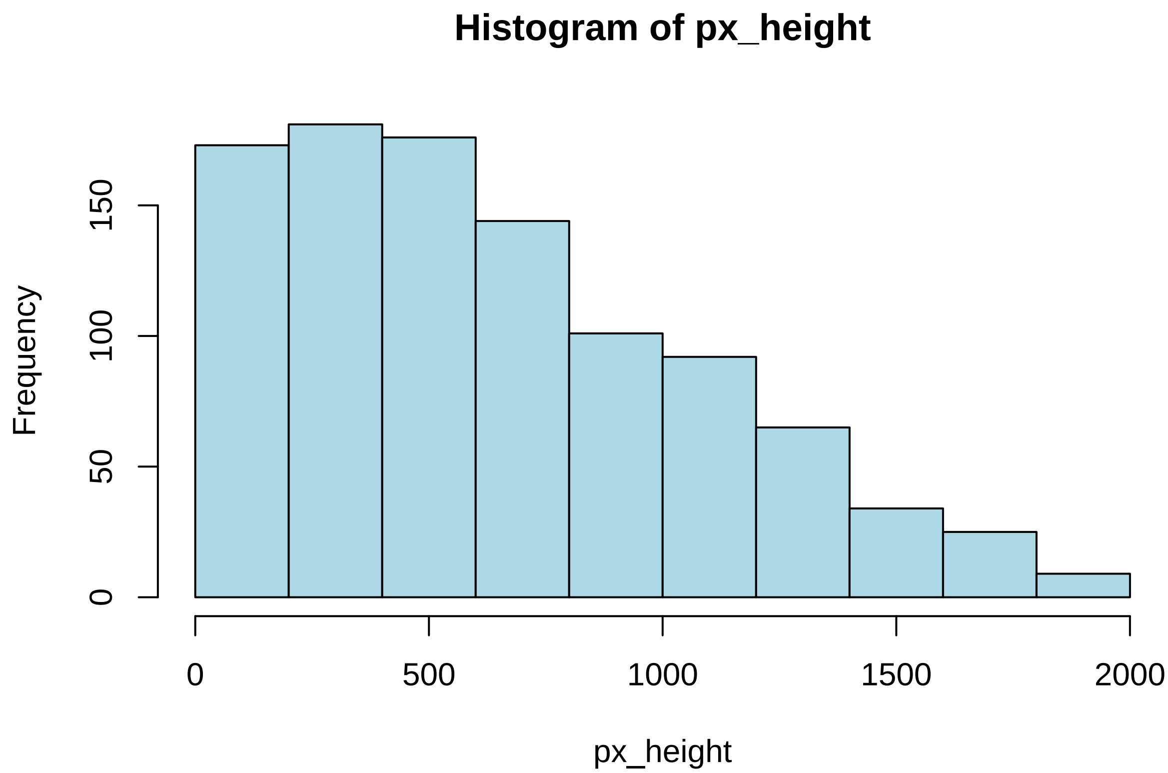



```
hist(mobile_dataset$ram,main = "Histogram of Ram power",xlab = "Ram",
     col="lightgreen")
```



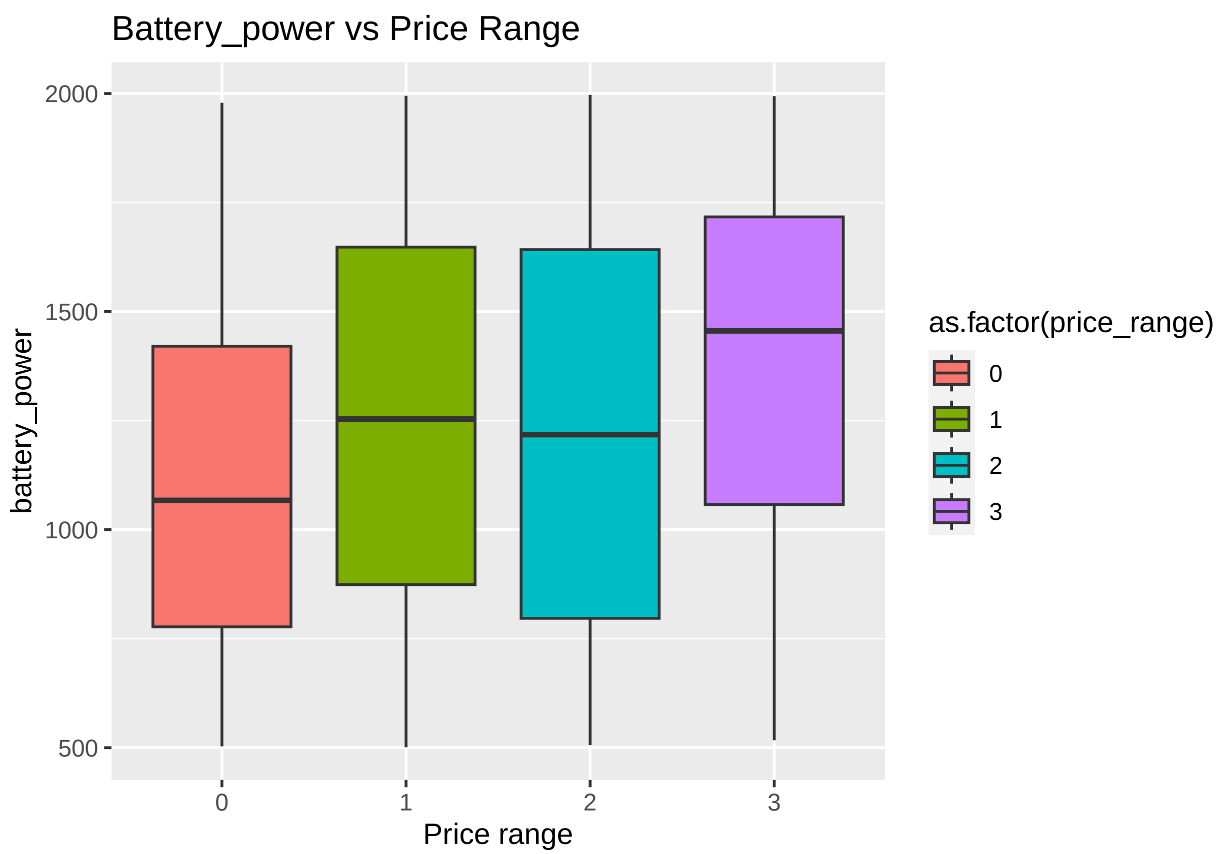

```
hist(mobile_dataset$px_width,main = "Histogram of px_width",
     xlab = "px_width",col="violet")
```


```
hist(mobile_dataset$px_height,main = "Histogram of px_height",
     xlab = "px_height",col="lightblue")
```

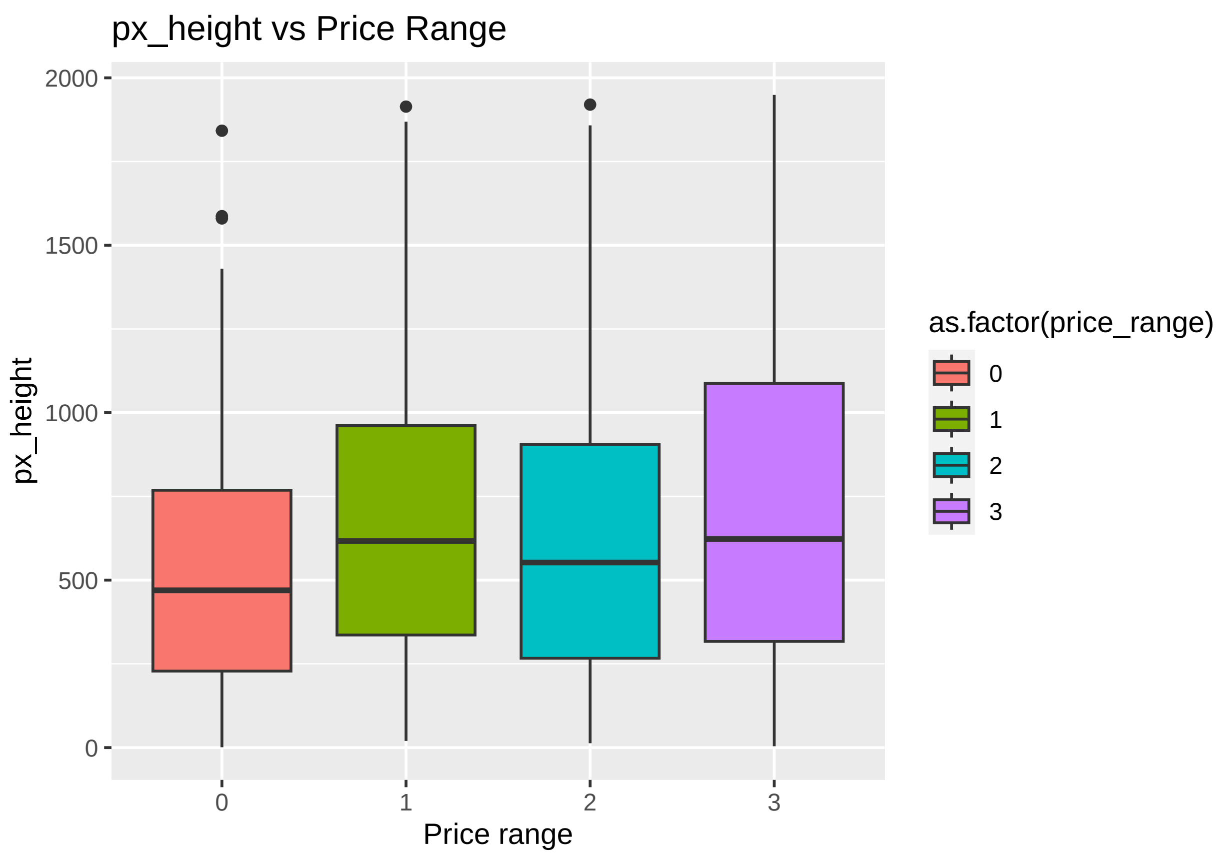


boxplot

```
library(ggplot2)
par(mfrow = c(2, 2))
ggplot(mobile_dataset,aes(x=as.factor(price_range), y=battery_power,
                          fill=as.factor(price_range))) + geom_boxplot()+ labs(
    x = "Price range",
    y = "battery_power",
    title = "Battery_power vs Price Range"
  )
```

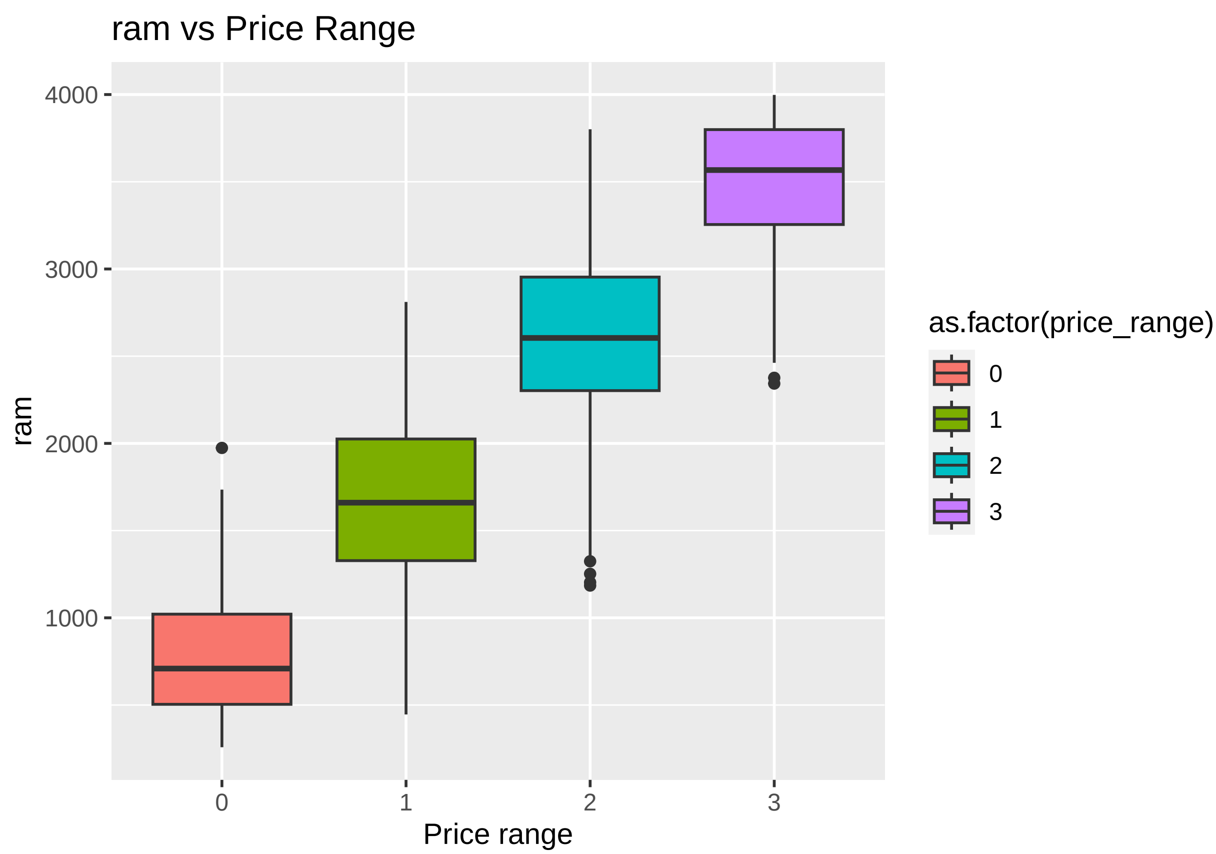



```r
ggplot(mobile_dataset,aes(x=as.factor(price_range), y=px_height,
                          fill=as.factor(price_range))) + geom_boxplot()+ labs(
    x = "Price range",
    y = "px_height",
    title = "px_height vs Price Range"
  )
```

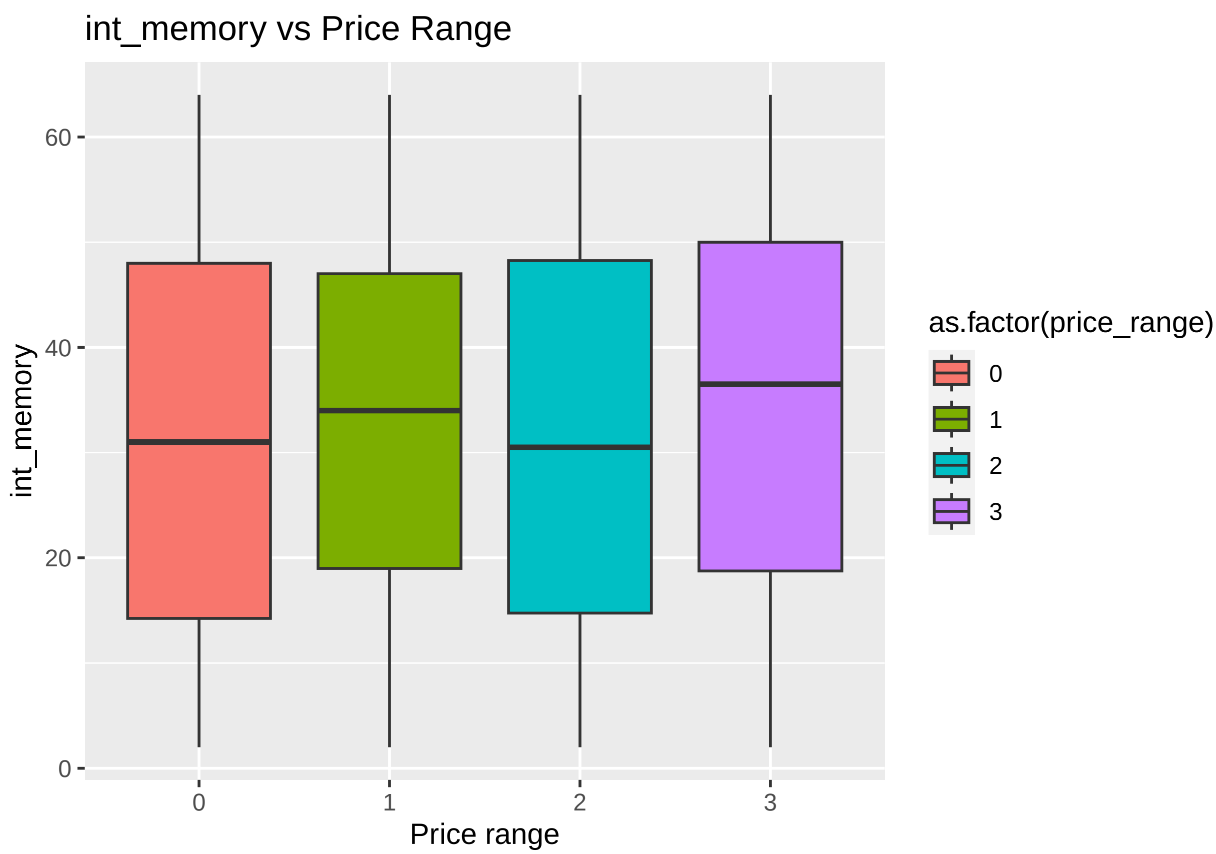



```r
ggplot(mobile_dataset,aes(x=as.factor(price_range), y=px_width,
                          fill=as.factor(price_range))) + geom_boxplot()+labs(
    x = "Price range",
    y = "px_width",
    title = "px_width vs Price Range"
  )
```

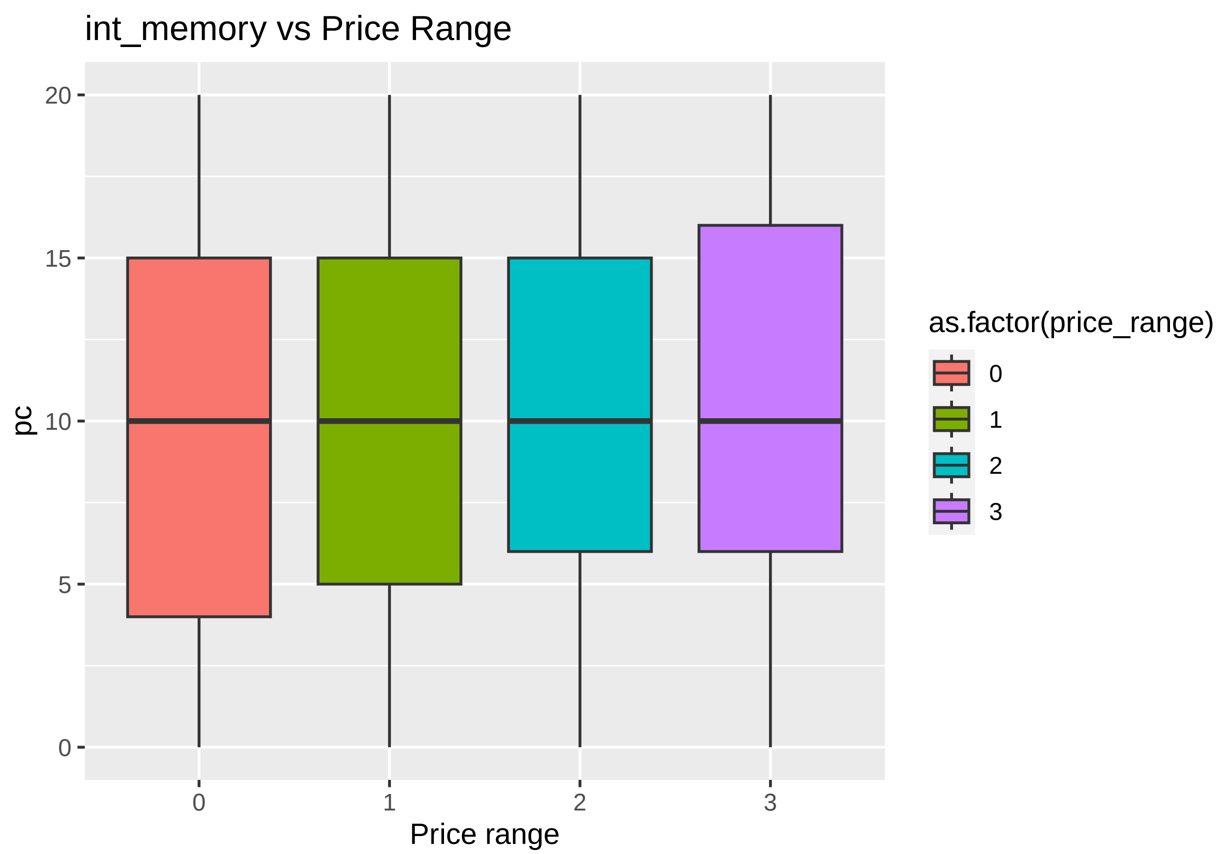

```r
ggplot(mobile_dataset,aes(x=as.factor(price_range), y=ram,
                          fill=as.factor(price_range))) + geom_boxplot()+labs(
    x = "Price range",
    y = "ram",
    title = "ram vs Price Range"
  )
```




```
ggplot(mobile_dataset,aes(x=as.factor(price_range), y=int_memory,
                          fill=as.factor(price_range))) + geom_boxplot()+labs(
    x = "Price range",
    y = "int_memory",
    title = "int_memory vs Price Range"
  )
```




```r
ggplot(mobile_dataset,aes(x=as.factor(price_range), y=pc,
                          fill=as.factor(price_range))) + geom_boxplot()+labs(
    x = "Price range",
    y = "pc",
    title = "int_memory vs Price Range"
  )
```




```
#We can see a huge difference in battery power between the price_range 0 and
#the price_range 3 but not so much between the 1 and 2.
```

Hypothesis testing one-sample hypothesis test: #Average battery power

```
battery_power_mean <- 1238 # average battery_power
t.test(mobile_dataset$battery_power,mu=1238,alternative = "less",
       conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  mobile_dataset$battery_power
## t = 1.0519, df = 999, p-value = 0.8534
## alternative hypothesis: true mean is less than 1238
## 95 percent confidence interval:
##      -Inf 1275.193
## sample estimates:
## mean of x
##  1252.499
```

```
#Do not reject as p value is greater than 0.05
#It means that we dont have enough evidence that the average battery power
#is less than 1238
```

Two sample hypothesis test: Null Hypothesis (H0): There is no difference in the average battery power between phones with and without 4G. Alternative Hypothesis (H1): There is a significant difference in the average battery power between phones with and without 4G.

```r
battery_power_4g <- mobile_dataset$battery_power[mobile_dataset$four_g == 1]
battery_power_no_4g <- mobile_dataset$battery_power[mobile_dataset$four_g == 0]
t_test_result <- t.test(battery_power_4g, battery_power_no_4g)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  battery_power_4g and battery_power_no_4g
## t = 2.2464, df = 996.61, p-value = 0.02489
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    7.815816 115.787844
## sample estimates:
## mean of x mean of y
##  1282.596  1220.795
```

```r
#Reject the hypothesis as there is a significant difference in average battery
# power between phones with and without 4G.
```

Null Hypothesis (H0): There is no difference in the average RAM between phones with and without dual SIM. Alternative Hypothesis (H1): There is a significant difference in the average RAM between phones with and without dual SIM.

```r
ram_dual_sim <- mobile_dataset$ram[mobile_dataset$dual_sim == 1]
ram_no_dual_sim <- mobile_dataset$ram[mobile_dataset$dual_sim == 0]
t_test_result <- t.test(ram_dual_sim, ram_no_dual_sim)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  ram_dual_sim and ram_no_dual_sim
## t = 2.064, df = 995.8, p-value = 0.03928
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    7.147532 283.120073
## sample estimates:
## mean of x mean of y
##  2254.109  2108.975
```

```r
#reject the hypothesis as there is a significant difference in average ram
# between phones with and without dual sim.
```

ANOVA:

```r
anova_result <- aov(battery_power ~ price_range, data = mobile_dataset)
summary(anova_result)
```

```
##              Df    Sum Sq Mean Sq F value   Pr(>F)    
## price_range   1   8782174 8782174   48.42 6.23e-12 ***
## Residuals   998 181023354  181386                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_result1 <- aov(px_height ~ price_range, data = mobile_dataset)
summary(anova_result1)
```

```
##              Df    Sum Sq Mean Sq F value   Pr(>F)    
## price_range   1   3127063 3127063    16.2 6.12e-05 ***
## Residuals   998 192593648  192980                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_result2 <- aov(ram ~ price_range, data = mobile_dataset)
summary(anova_result2)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)    
## price_range   1 1.050e+09 1.050e+09    5568 <2e-16 ***
## Residuals   998 1.882e+08 1.886e+05                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_result3 <- aov(px_width ~ price_range, data = mobile_dataset)
summary(anova_result3)
```

```
##              Df    Sum Sq Mean Sq F value   Pr(>F)    
## price_range   1   4454605 4454605   25.01 6.74e-07 ***
## Residuals   998 177760259  178116                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_result4 <- aov(int_memory ~ price_range, data = mobile_dataset)
summary(anova_result4)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## price_range   1   1065  1064.6   3.268  0.071 .
## Residuals   998 325156   325.8                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_result5 <- aov(pc ~ price_range, data = mobile_dataset)
summary(anova_result5)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## price_range   1     82   82.00    2.24  0.135
## Residuals   998  36534   36.61               
```

CHI-SQUARE:

```
chi_squared_result_3g <- chisq.test(table(mobile_dataset$three_g,
                                           mobile_dataset$price_range))
chi_squared_result_3g
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mobile_dataset$three_g, mobile_dataset$price_range)
## X-squared = 3.67, df = 3, p-value = 0.2994
```

```
chi_squared_result_sim <- chisq.test(table(mobile_dataset$four_g,
                                            mobile_dataset$price_range))
chi_squared_result_sim
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mobile_dataset$four_g, mobile_dataset$price_range)
## X-squared = 5.007, df = 3, p-value = 0.1713
```

```
#If the p-value is less than the significance level (0.05), we reject the null
#hypothesis, suggesting evidence of an association.
```

shapiro test

```
shapiro_test_result <- shapiro.test(mobile_dataset$ram)
shapiro_test_result
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mobile_dataset$ram
## W = 0.946, p-value < 2.2e-16
```

```
#The data is not normally distributed.
```

Levene's test for homogeneity of variances

```
if (!requireNamespace("car", quietly = TRUE)) {
  install.packages("car")
}
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.2
```

```r
levene_test_result <- leveneTest(mobile_dataset$battery_power,
                                 group = mobile_dataset$price_range)
```

```
## Warning in leveneTest.default(mobile_dataset$battery_power, group =
## mobile_dataset$price_range): mobile_dataset$price_range coerced to factor.
```

```r
levene_test_result
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   3  6.2575 0.0003296 ***
##       996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Variances are significantly different across groups.
```

Kruskal-Wallis test(alternative to Anova test)

```r
#install.packages("coin")
library(coin)
```

```
## Warning: package 'coin' was built under R version 4.3.2
```

```
## Loading required package: survival
```

```r
kw_result1 <- kruskal.test(battery_power ~ price_range, data = mobile_dataset)
kw_result2 <- kruskal.test(px_height ~ price_range, data = mobile_dataset)
kw_result3 <- kruskal.test(ram ~ price_range, data = mobile_dataset)
kw_result4 <- kruskal.test(px_width ~ price_range, data = mobile_dataset)
kw_result5 <- kruskal.test(int_memory ~ price_range, data = mobile_dataset)
kw_result6 <- kruskal.test(pc ~ price_range, data = mobile_dataset)

print(kw_result1)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  battery_power by price_range
## Kruskal-Wallis chi-squared = 52.875, df = 3, p-value = 1.949e-11
```

```r
print(kw_result2)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  px_height by price_range
## Kruskal-Wallis chi-squared = 19.303, df = 3, p-value = 0.0002366
```

```r
print(kw_result3)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ram by price_range
## Kruskal-Wallis chi-squared = 843.2, df = 3, p-value < 2.2e-16
```

```r
print(kw_result4)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  px_width by price_range
## Kruskal-Wallis chi-squared = 29.443, df = 3, p-value = 1.808e-06
```

```r
print(kw_result5)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  int_memory by price_range
## Kruskal-Wallis chi-squared = 5.552, df = 3, p-value = 0.1356
```

```r
print(kw_result6)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  pc by price_range
## Kruskal-Wallis chi-squared = 2.9383, df = 3, p-value = 0.4012
```