EDA-810 team project

Yixuan Wang

2/26/2021

```
library(data.table)
library(ggplot2)
library(ggcorrplot)
```

Exploratory Data Analysis (EDA)

Load data set into R

```
mobile_price <- fread("/Users/wangyixuan/Desktop/BA810 Supervised machine learning/tr
ain.csv")
```

```
head(mobile_price, 5)
```

```
##    battery_power blue clock_speed dual_sim fc four_g int_memory m_dep mobile_wt
## 1:           842    0         2.2        0  1      0          7   0.6       188
## 2:          1021    1         0.5        1  0      1         53   0.7       136
## 3:           563    1         0.5        1  2      1         41   0.9       145
## 4:           615    1         2.5        0  0      0         10   0.8       131
## 5:          1821    1         1.2        0 13      1         44   0.6       141
##    n_cores pc px_height px_width  ram sc_h sc_w talk_time three_g touch_screen
## 1:       2  2        20      756 2549    9    7        19       0            0
## 2:       3  6       905     1988 2631   17    3         7       1            1
## 3:       5  6      1263     1716 2603   11    2         9       1            1
## 4:       6  9      1216     1786 2769   16    8        11       1            0
## 5:       2 14      1208     1212 1411    8    2        15       1            1
##    wifi price_range
## 1:    1           1
## 2:    0           2
## 3:    0           2
## 4:    0           2
## 5:    0           1
```

Observing the data structure

```
dim(mobile_price)
```

```
## [1] 2000   21
```

The data set has 2000 rows with 21 variables.

Then, we explore the data types of each variable/column.

```
cat_vars <- names(mobile_price)[which(sapply(mobile_price, is.character))]
cat_vars
```

```
## character(0)
```

```
numeric_vars <- names(mobile_price)[which(sapply(mobile_price, is.numeric))]
numeric_vars
```

```
##  [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"
```

In our data set, all variables are numeric.

Checking for missing values

```
colSums(sapply(mobile_price, is.na))
```

```
## battery_power          blue   clock_speed      dual_sim            fc 
##             0             0             0             0             0 
##        four_g    int_memory         m_dep     mobile_wt       n_cores 
##             0             0             0             0             0 
##            pc     px_height      px_width           ram          sc_h 
##             0             0             0             0             0 
##          sc_w     talk_time       three_g  touch_screen          wifi 
##             0             0             0             0             0 
##   price_range 
##             0
```

There is no missing value in our data set, so we can proceed to analysis without worrying about missing values.

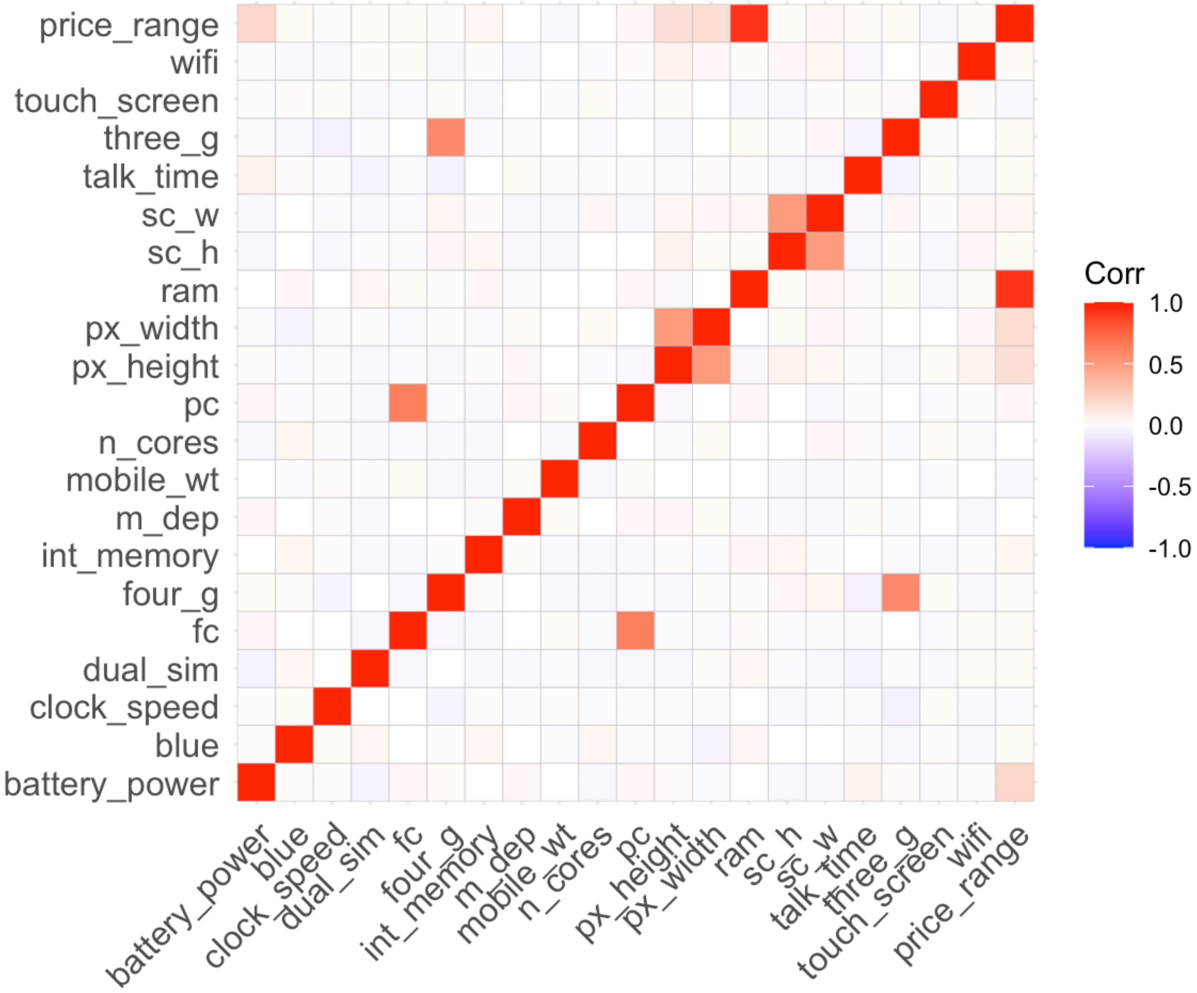
Data summary

```
summary(mobile_price)
```

```
##  battery_power         blue        clock_speed       dual_sim
##  Min.   : 501.0   Min.   :0.000   Min.   :0.500   Min.   :0.0000
##  1st Qu.: 851.8   1st Qu.:0.000   1st Qu.:0.700   1st Qu.:0.0000
##  Median :1226.0   Median :0.000   Median :1.500   Median :1.0000
##  Mean   :1238.5   Mean   :0.495   Mean   :1.522   Mean   :0.5095
##  3rd Qu.:1615.2   3rd Qu.:1.000   3rd Qu.:2.200   3rd Qu.:1.0000
##  Max.   :1998.0   Max.   :1.000   Max.   :3.000   Max.   :1.0000
##        fc             four_g         int_memory        m_dep
##  Min.   : 0.000   Min.   :0.0000   Min.   : 2.00   Min.   :0.1000
##  1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:16.00   1st Qu.:0.2000
##  Median : 3.000   Median :1.0000   Median :32.00   Median :0.5000
##  Mean   : 4.309   Mean   :0.5215   Mean   :32.05   Mean   :0.5018
##  3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.:48.00   3rd Qu.:0.8000
##  Max.   :19.000   Max.   :1.0000   Max.   :64.00   Max.   :1.0000
##    mobile_wt        n_cores            pc           px_height
##  Min.   : 80.0   Min.   :1.000   Min.   : 0.000   Min.   :   0.0
##  1st Qu.:109.0   1st Qu.:3.000   1st Qu.: 5.000   1st Qu.: 282.8
##  Median :141.0   Median :4.000   Median :10.000   Median : 564.0
##  Mean   :140.2   Mean   :4.521   Mean   : 9.916   Mean   : 645.1
##  3rd Qu.:170.0   3rd Qu.:7.000   3rd Qu.:15.000   3rd Qu.: 947.2
##  Max.   :200.0   Max.   :8.000   Max.   :20.000   Max.   :1960.0
##     px_width           ram            sc_h            sc_w
##  Min.   : 500.0   Min.   : 256   Min.   : 5.00   Min.   : 0.000
##  1st Qu.: 874.8   1st Qu.:1208   1st Qu.: 9.00   1st Qu.: 2.000
##  Median :1247.0   Median :2146   Median :12.00   Median : 5.000
##  Mean   :1251.5   Mean   :2124   Mean   :12.31   Mean   : 5.767
##  3rd Qu.:1633.0   3rd Qu.:3064   3rd Qu.:16.00   3rd Qu.: 9.000
##  Max.   :1998.0   Max.   :3998   Max.   :19.00   Max.   :18.000
##    talk_time        three_g        touch_screen        wifi
##  Min.   : 2.00   Min.   :0.0000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 6.00   1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.000
##  Median :11.00   Median :1.0000   Median :1.000   Median :1.000
##  Mean   :11.01   Mean   :0.7615   Mean   :0.503   Mean   :0.507
##  3rd Qu.:16.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :20.00   Max.   :1.0000   Max.   :1.000   Max.   :1.000
##   price_range
##  Min.   :0.00
##  1st Qu.:0.75
##  Median :1.50
##  Mean   :1.50
##  3rd Qu.:2.25
##  Max.   :3.00
```

Explore the correlation

```
correlation <- cor(mobile_price)
ggcorrplot(correlation)
```



Data visualization

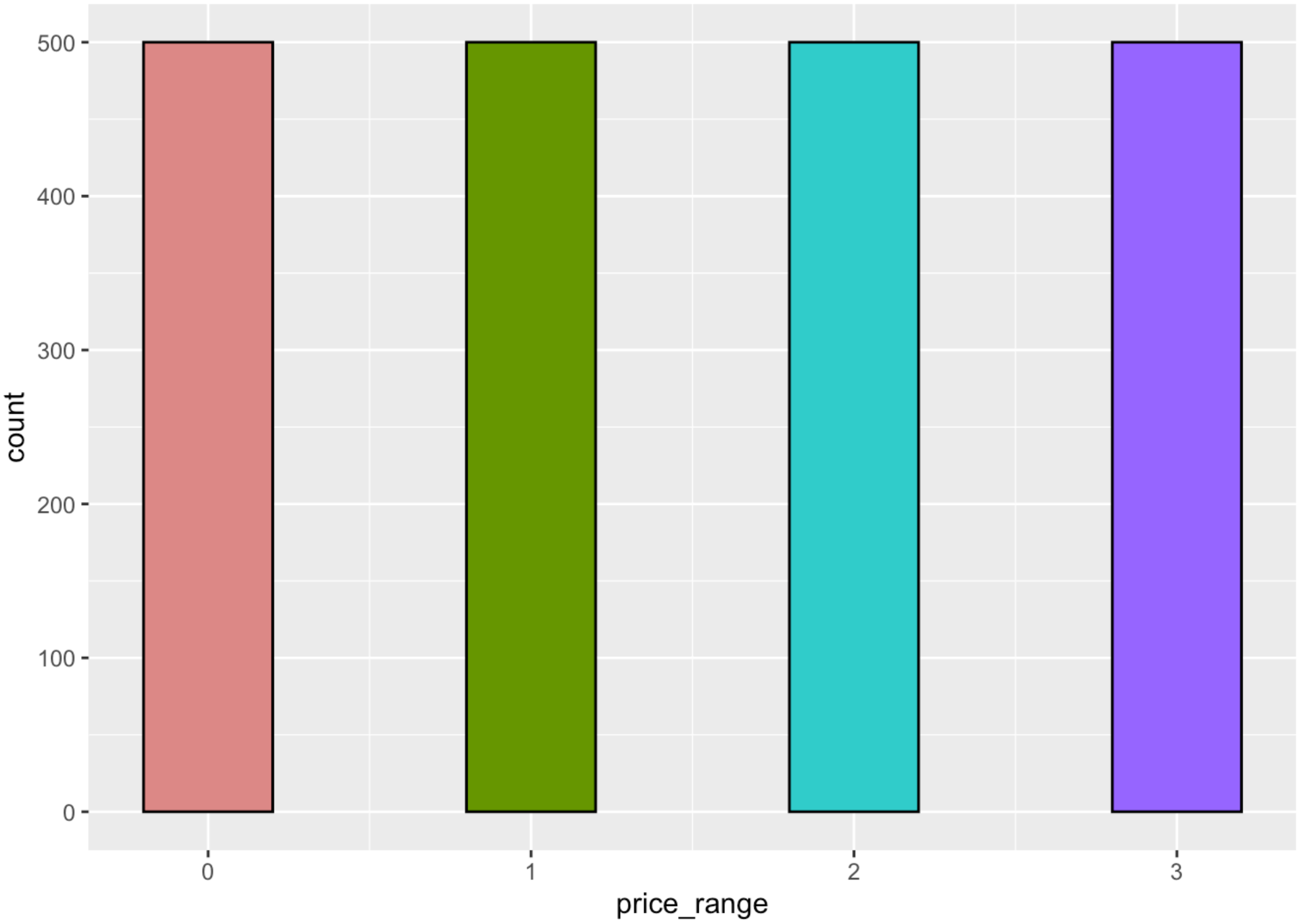
Price Range:

```
table(mobile_price$price_range)
```

```
##
##   0   1   2   3
## 500 500 500 500
```

The mobile phone's price range is equally distributed.

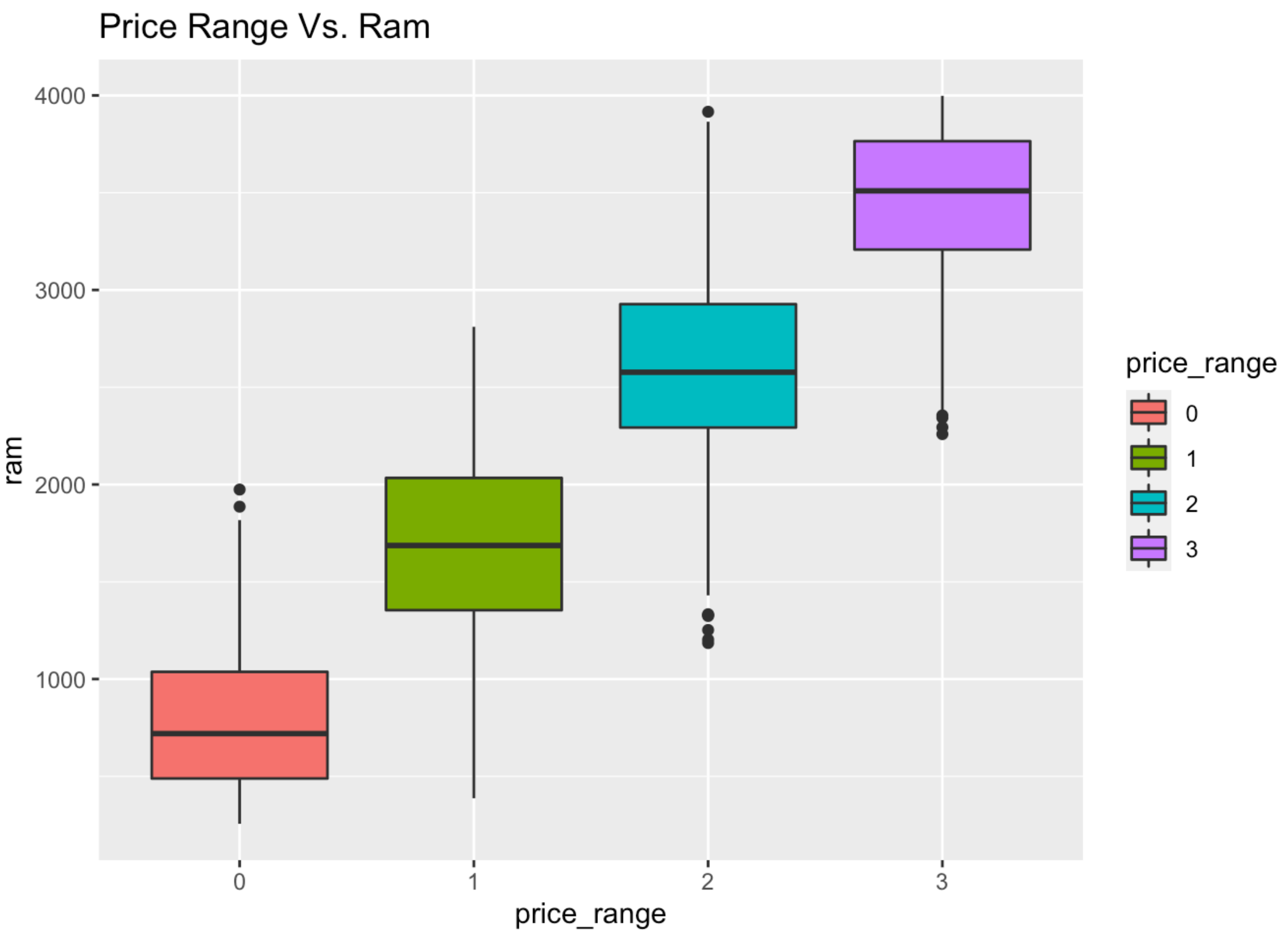
```
ggplot(data=mobile_price) +
geom_bar(mapping = aes(x=price_range), width = 0.4,colour="black", fill=c("#DD8888",
"#669900", "#33CCCC","#9966FF" ))
```



```
mobile_price$price_range <- as.factor(mobile_price$price_range)
```

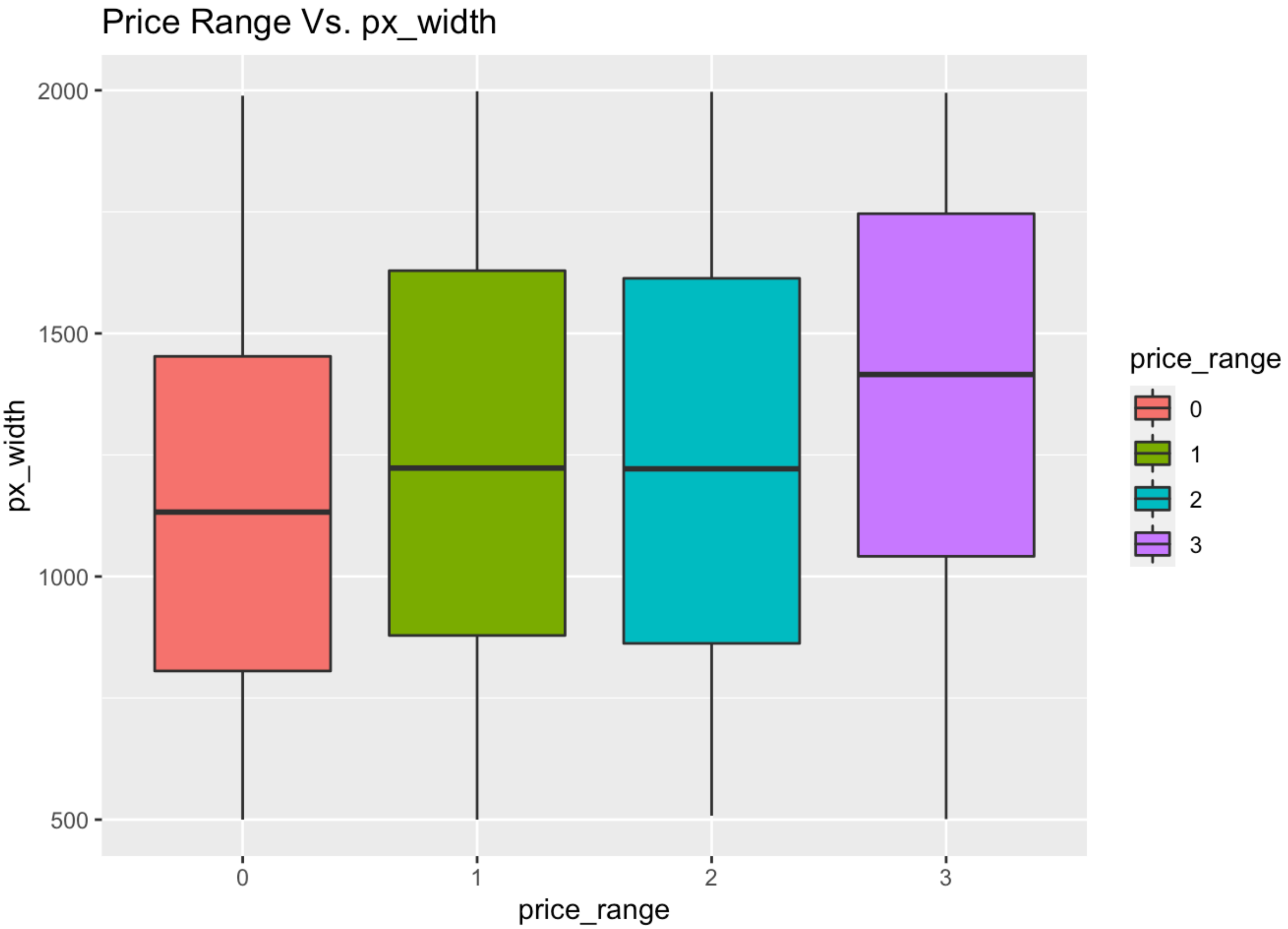
Price range Vs. Ram

```
ggplot(mobile_price, aes(x=price_range, y=ram, fill=price_range)) +
geom_boxplot() +
ggtitle("Price Range Vs. Ram")
```



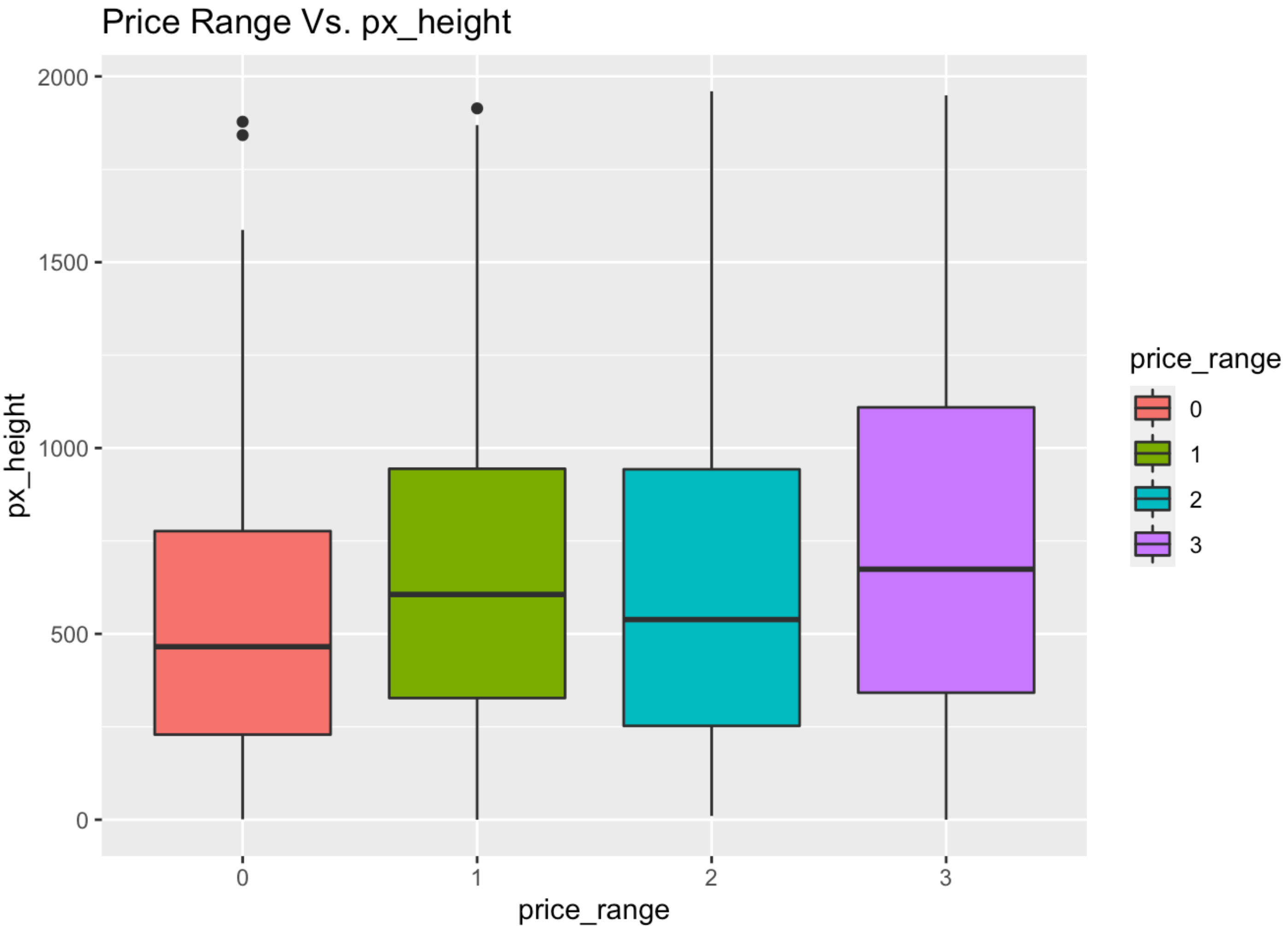
Price range Vs. px_width

```
ggplot(mobile_price, aes(x=price_range, y=px_width, fill=price_range)) +
geom_boxplot()+
ggtitle("Price Range Vs. px_width")
```



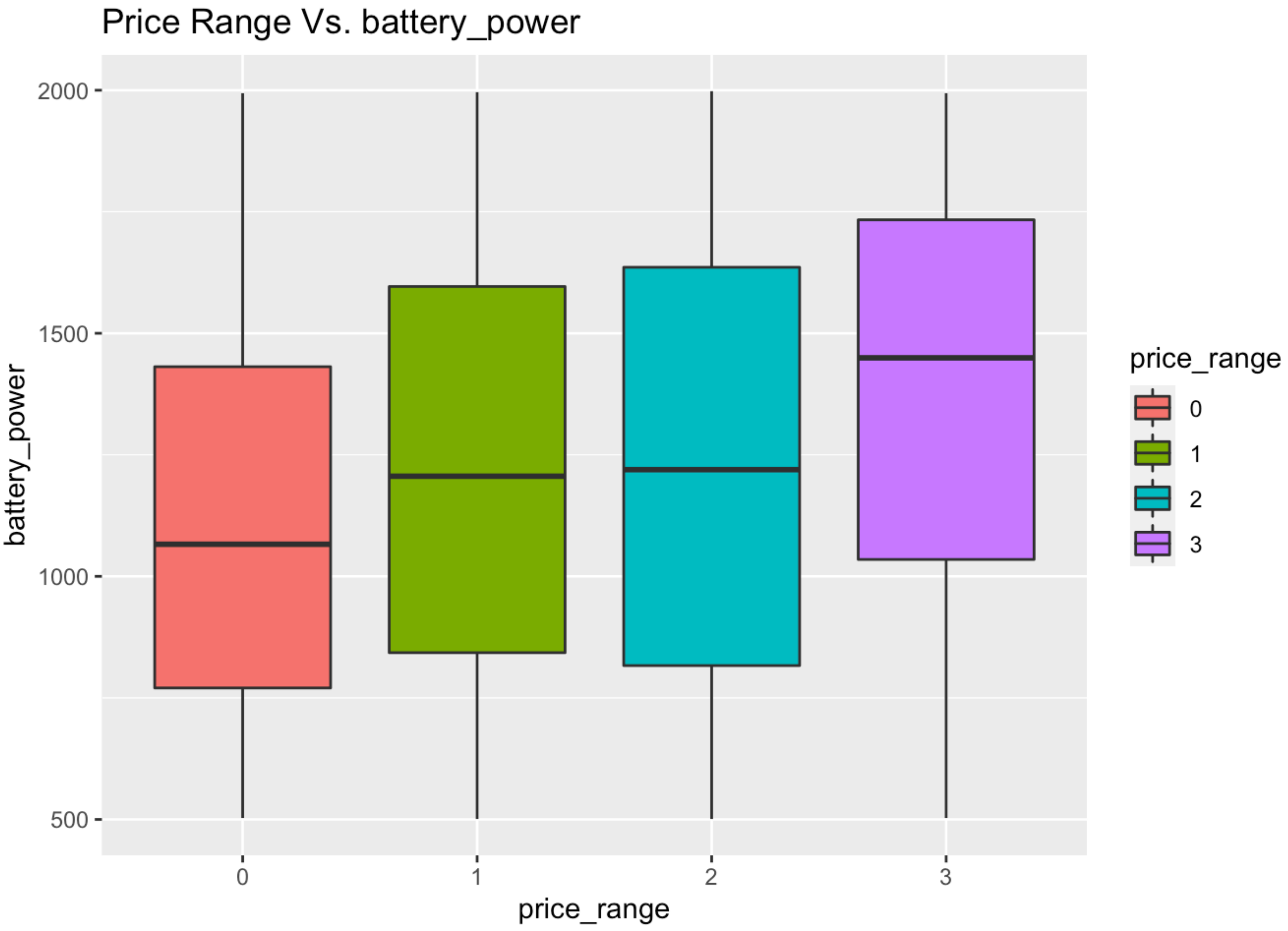
Price range Vs. px_height

```
ggplot(mobile_price, aes(x=price_range, y=px_height, fill=price_range)) +
geom_boxplot()+
ggtitle("Price Range Vs. px_height")
```



Price range Vs. battery_power

```
ggplot(mobile_price, aes(x=price_range, y=battery_power, fill=price_range)) +
geom_boxplot()+
ggtitle("Price Range Vs. battery_power")
```


Price Range Vs. battery_power
battery_power
2000
1500
1000
500
0
1
2
3
price_range
price_range
0
1
2
3