Shop Sales for Common Mobile In Egypt

> Investigated a dataset using R and exploratory data analysis techniques, exploring both single variables and relationships between variables. I analyze shop sales for mobile phone in Egypt containing the prices,sellout transaction and other attributes of almost 200 famous Mobile that sold in 1700 in Jan 2018 shop

Univariate Plots Section

```
dim(mob)
```

```
## [1] 26553    17
```

```
str(mob)
```

```
## 'data.frame':    26553 obs. of  17 variables:
##  $ model_id       : int  76732 7760 7869 8314 6708 8397 8785 8126 8561 8904 ...
##  $ model_name     : Factor w/ 201 levels "2","3","5","6",..: 53 56 59 63 50 54 62 57 64 55 ...
##  $ brand_name     : Factor w/ 12 levels "Alcatel","Apple",..: 11 11 11 11 11 11 11 11 11 11 ...
##  $ released_year  : int  2015 2016 2016 2016 2014 2016 2017 2016 2017 2018 ...
##  $ released_month : int  10 1 4 9 10 12 7 6 7 1 ...
##  $ connection_band: Factor w/ 3 levels "3G","4G","Wifi": 2 2 2 2 2 1 2 2 2 2 ...
##  $ size_in_inch   : num  4.3 5 5.2 5.5 5 4 5.5 5 5.5 5 ...
##  $ camera_mb      : num  5 8 13 13 8 5 13 8 13 8 ...
##  $ storage        : int  4 8 16 16 8 8 16 16 32 16 ...
##  $ ram            : num  0.512 1.5 2 3 1 1 2 2 3 1.5 ...
##  $ sellout        : int  11 8 3 4 47 27 17 5 4 5 ...
##  $ price          : int  1639 2289 3449 4499 1867 1353 3499 3449 5599 2499 ...
##  $ shop_id        : int  5257 5257 5257 5257 5257 5257 5257 5257 5257 5257 ...
##  $ channel        : Factor w/ 3 levels "Chain store",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ governorate    : Factor w/ 26 levels "Alexandria","Aswan",..: 16 16 16 16 16 16 16 16 16 16 ...
##  $ latitude       : num  30.3 30.3 30.3 30.3 30.3 ...
##  $ longitude      : num  31 31 31 31 31 ...
```

Our dataset consists of 17 variables, with almost 26,000 observations.

```
levels(mob$channel)
```

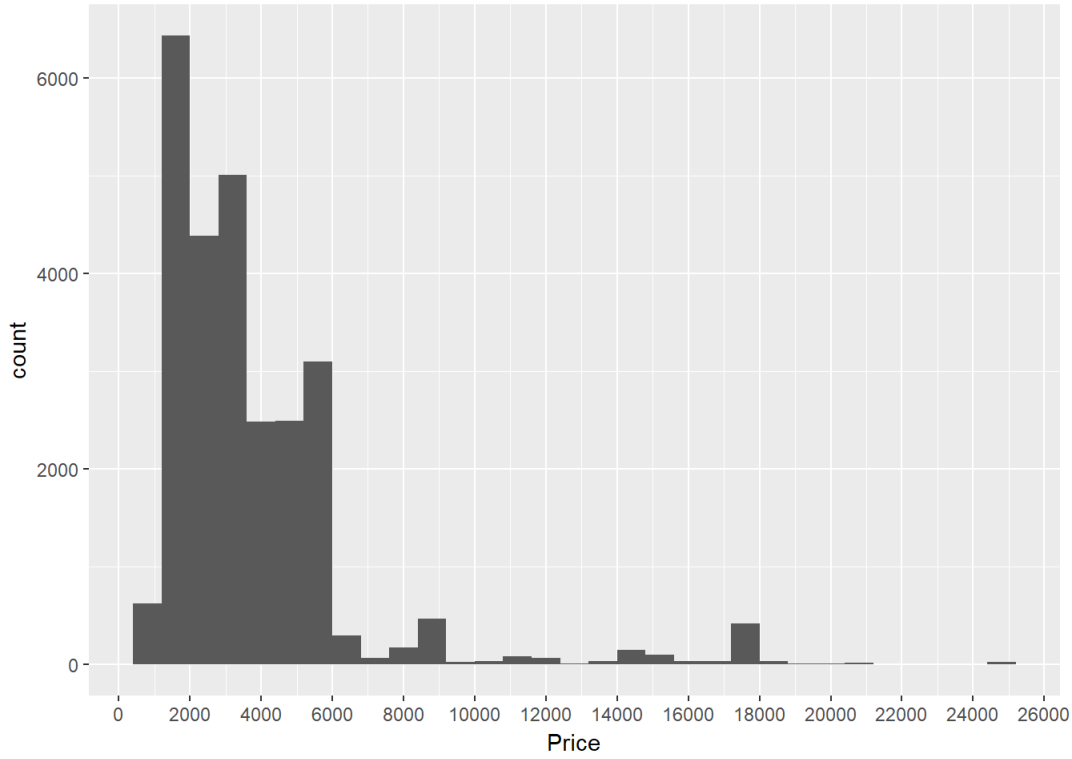
```
## [1] "Chain store"      "Hyper Market"     "Independent Shop"
```

```
mob$channel <- factor(mob$channel,
                      levels = c('Hyper Market','Chain store',
                                 'Independent Shop'))

levels(mob$channel)
```

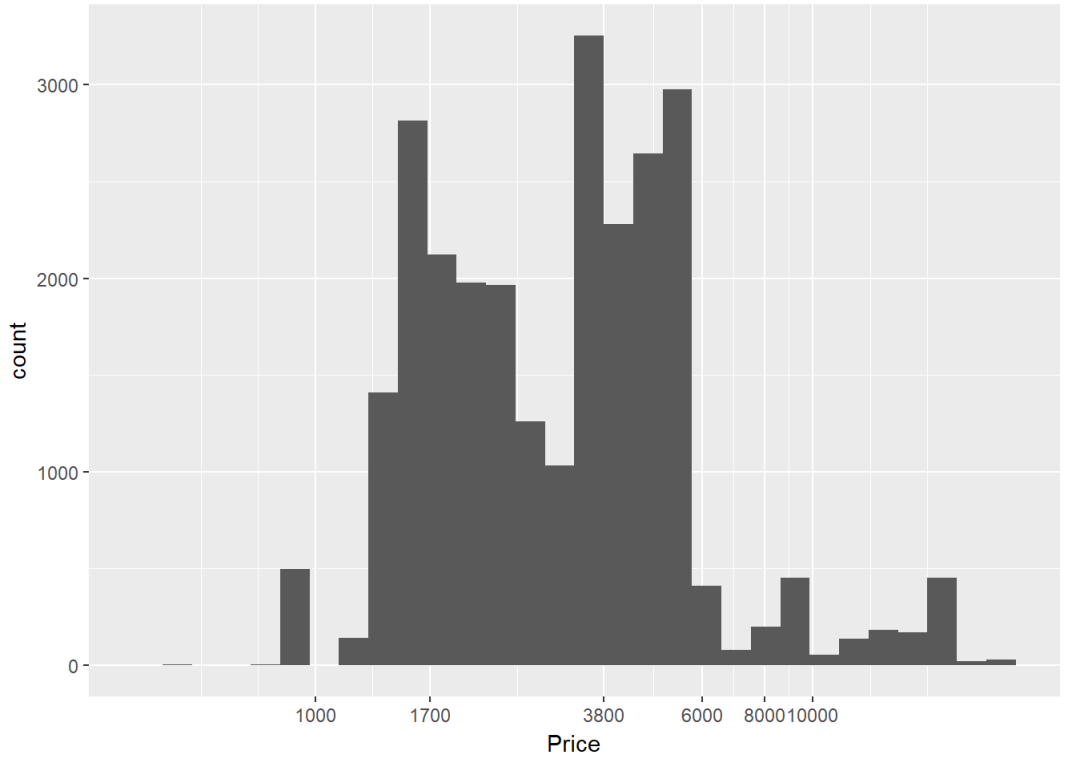
```
## [1] "Hyper Market"     "Chain store"      "Independent Shop"
```

Reorder the levels of the channel from big (Hyper) to small (Independent)

```
ggplot(data = mob , aes(x = price))+
  geom_histogram(binwidth = 800)+
  scale_x_continuous(breaks = seq(0,27000,2000))+
  xlab('Price')
```




```
ggplot(data = mob , aes(x = price))+
  geom_histogram(bins  = 30)+
  scale_x_log10(breaks=c(1000,1700,3800,6000,8000,10000))+
  xlab('Price')
```

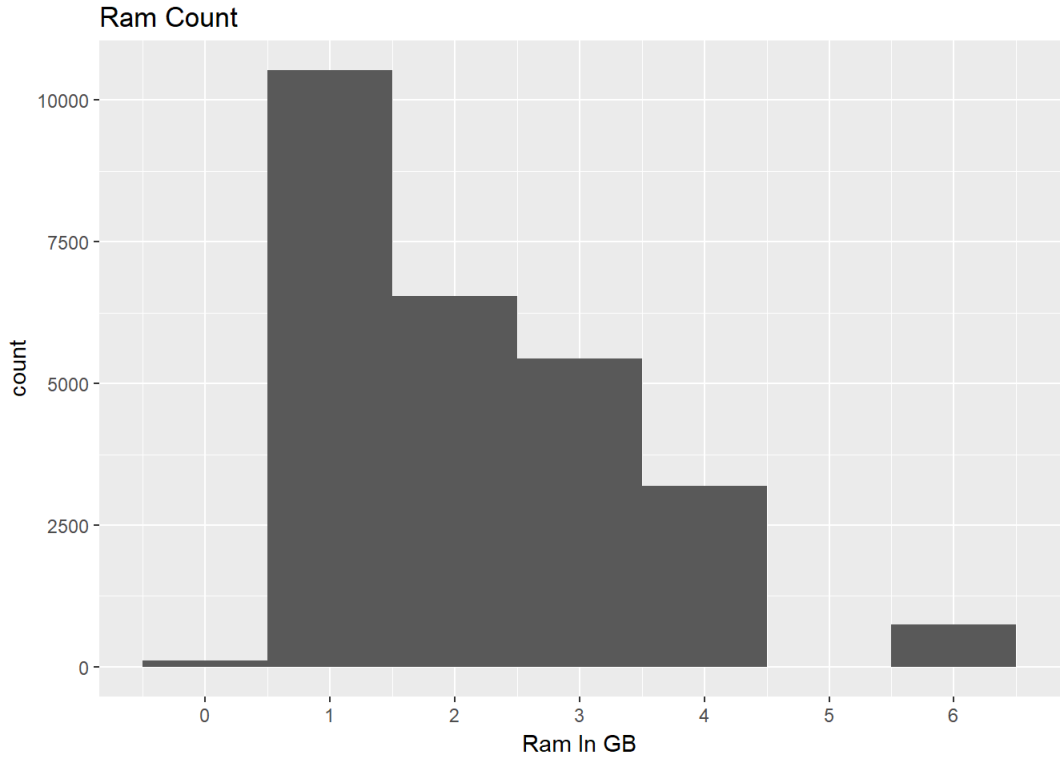



```
summary(mob$price)
```

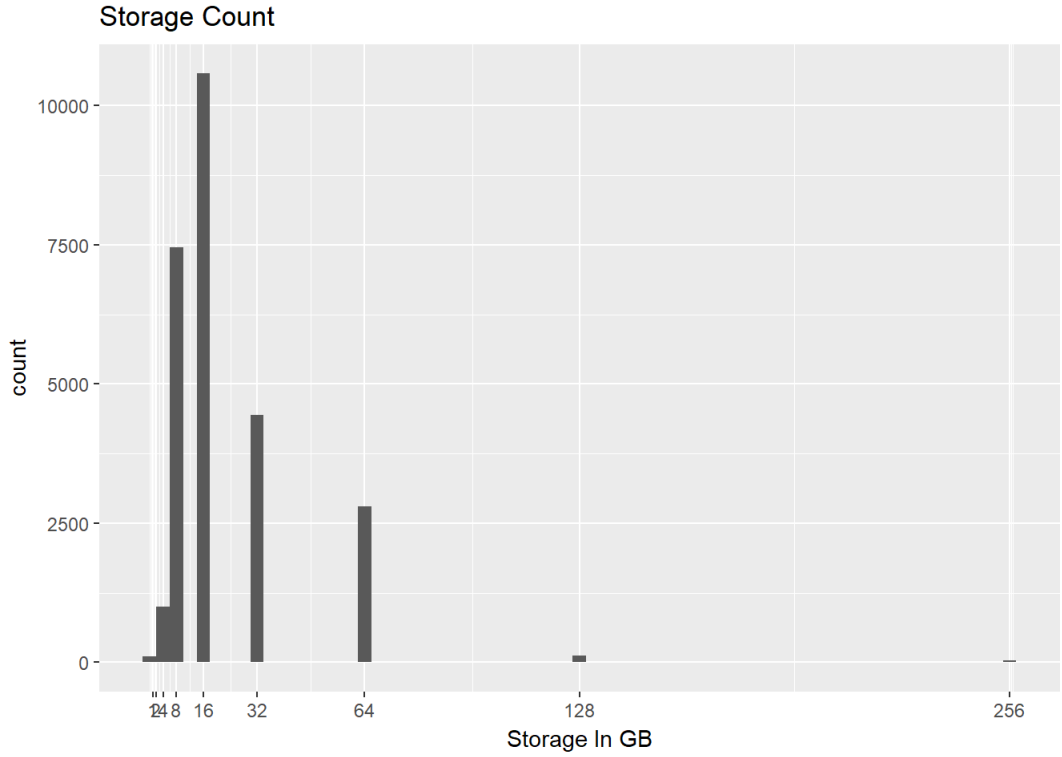
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     475    1867    3339    3797    4499   24700
```

The tranformed price distribution appears bimodal with the price peaking around 1700 or so and again at 3800 I wonder what this plot looks like across the variables of camera_mb, size_in_inch, ram, and storage.

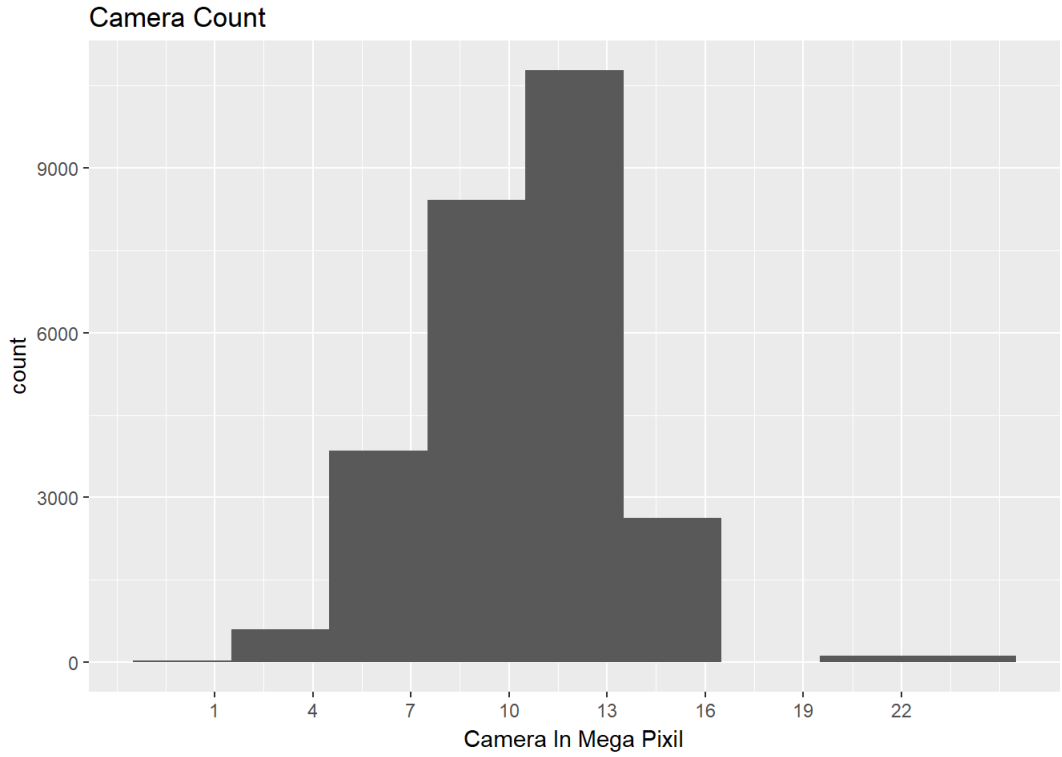
```
ggplot(data = mob , aes(x = ram))+
  geom_histogram(binwidth = 1)+
  scale_x_continuous(breaks = seq(0,6,1))+
  xlab('Ram In GB')+
  ggtitle('Ram Count')
```



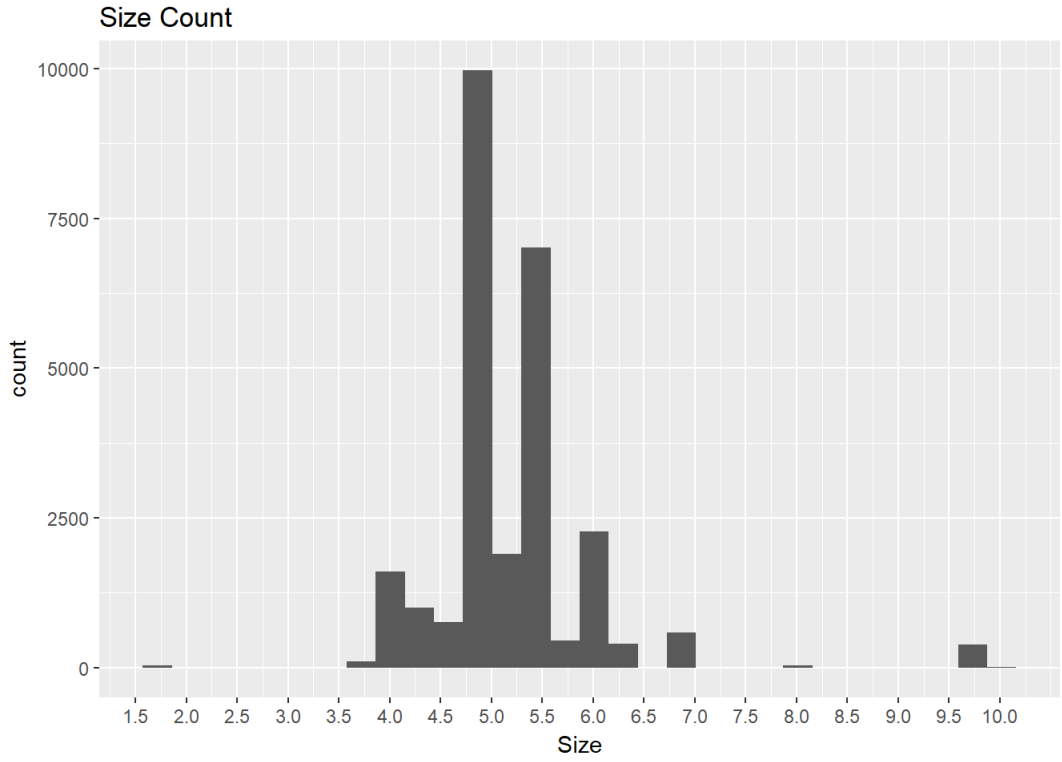

```
ggplot(data = mob , aes(x = storage))+
  geom_histogram(binwidth = 4)+
  scale_x_continuous(breaks = c(1,2, 4, 8,16,32,64,128,256))+
  xlab('Storage In GB')+
  ggtitle('Storage Count')
```



```
ggplot(data = mob , aes(x = camera_mb))+
  geom_histogram(binwidth = 3)+
  scale_x_continuous(breaks = seq(1,23,3))+
  xlab('Camera In Mega Pixil')+
  ggtitle('Camera Count')
```

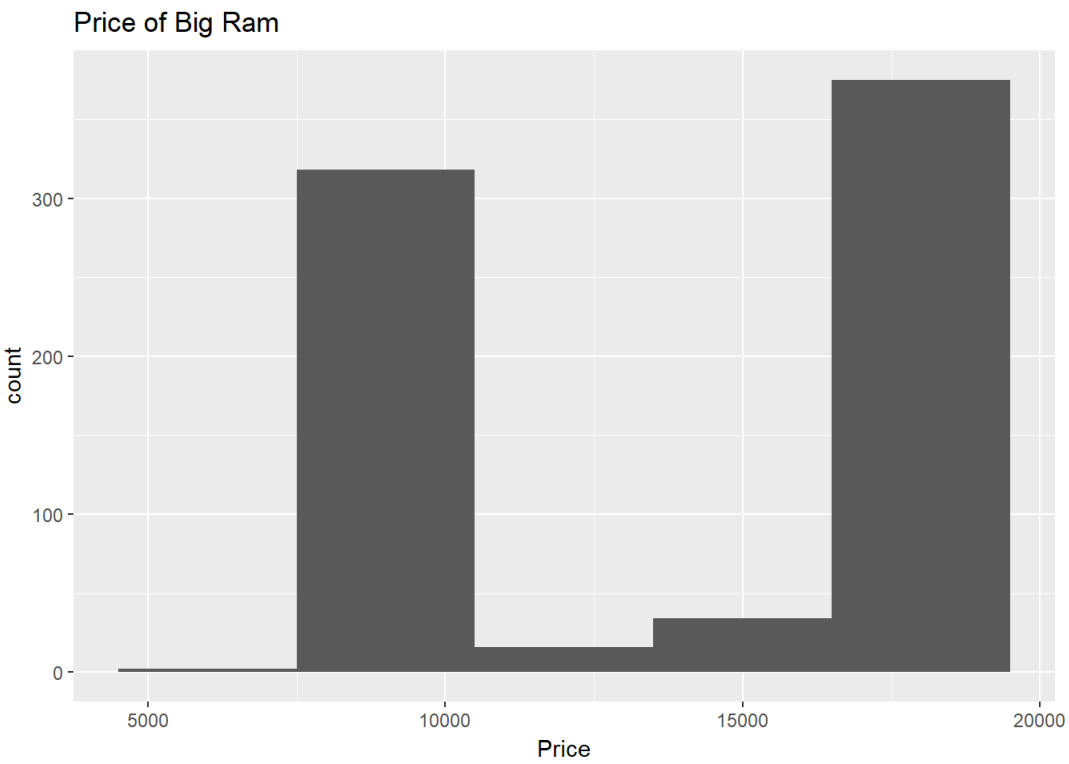



```
ggplot(data = mob , aes(x = size_in_inch))+
  geom_histogram()+
  scale_x_continuous(breaks = seq(0,10,.5))+
  xlab('Size')+ggtitle('Size Count')
```



Ram size play a big part in the price for each device

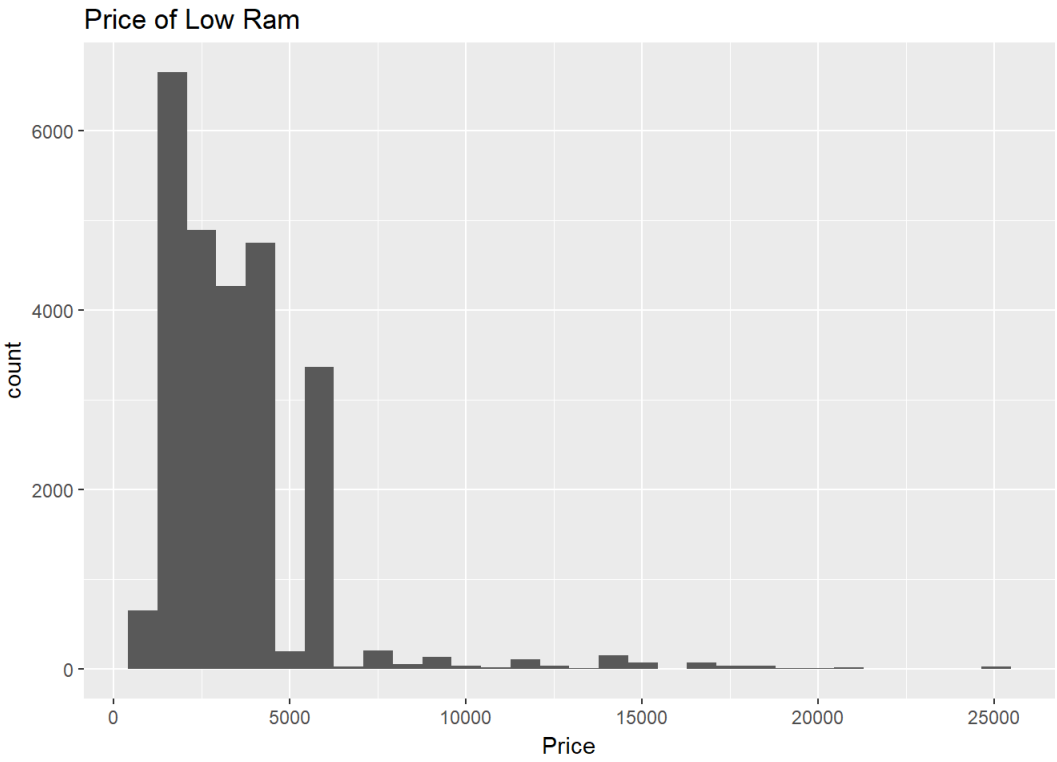
```
ggplot(data = subset(mob,ram > 4) , aes(x = price))+
  geom_histogram(binwidth = 3000)+
  scale_x_continuous()+
  xlab('Price')+ggtitle('Price of Big Ram')
```




```
summary(subset(mob,ram > 4 ,select=c(price)))
```

```
##      price
##  Min.   : 6960
##  1st Qu.: 8999
##  Median :17700
##  Mean   :13739
##  3rd Qu.:17700
##  Max.   :17700
```

```
ggplot(data = subset(mob,ram <= 4) , aes(x = price))+
  geom_histogram()+
  scale_x_continuous()+
  xlab('Price')+ggtitle('Price of Low Ram')
```

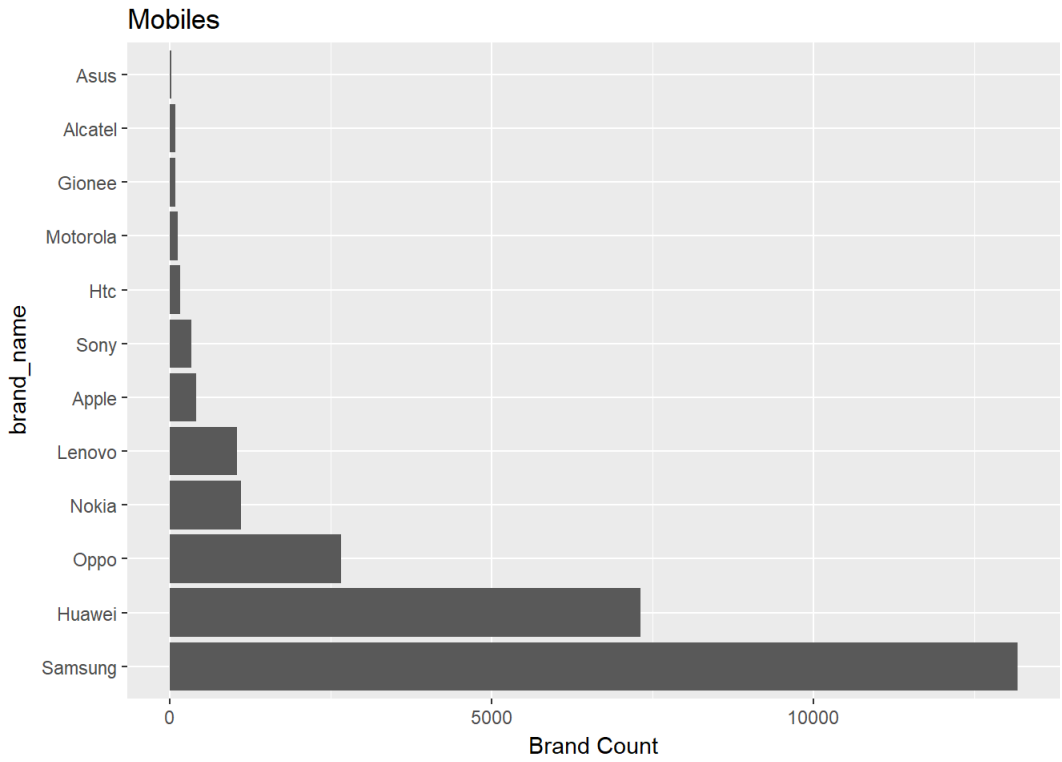



```
summary(subset(mob,ram <= 4 ,select=c(price)))
```

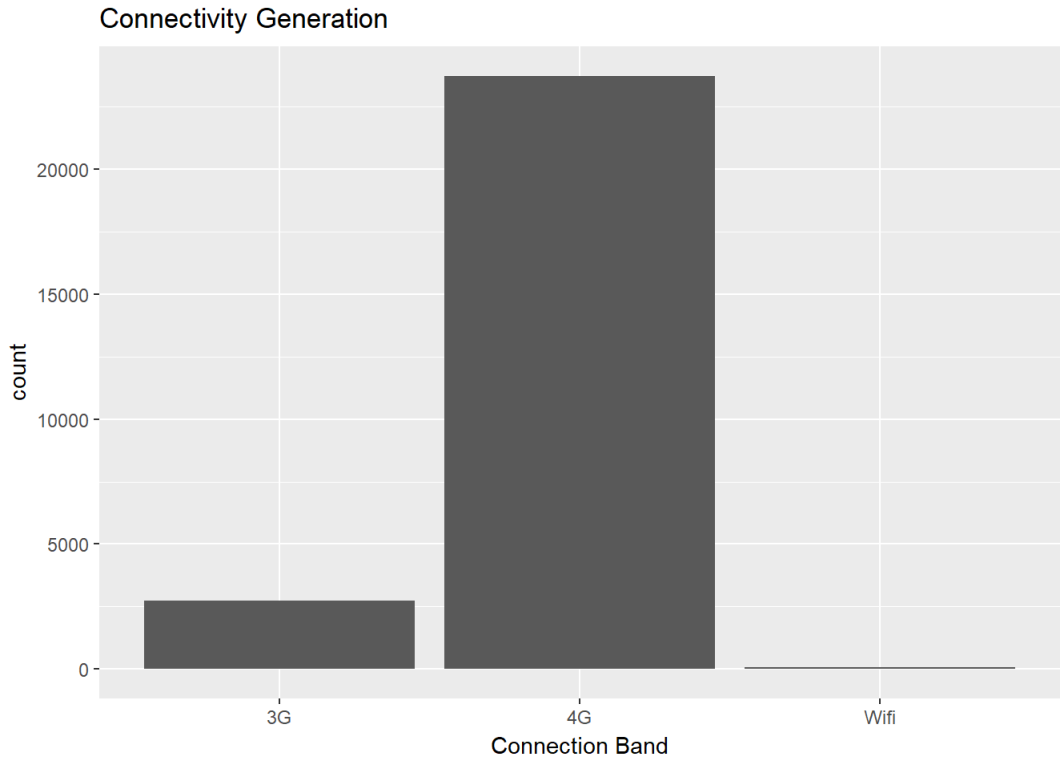
```
##      price
##  Min.   :  475
##  1st Qu.: 1867
##  Median : 3199
##  Mean   : 3510
##  3rd Qu.: 4450
##  Max.   :24700
```

As we can see the price range for the ram exeed 4 GB and below it

```
# sort brand before histogram
brand_count = sort(table(mob$brand_name), decreasing = T)

mob$brand_name = ordered(mob$brand_name ,levels = names(brand_count))

ggplot(data = mob , aes(brand_name))+
  geom_histogram(stat = 'count')+ coord_flip() +
  ylab('Brand Count')+
  ggtitle('Mobiles')
```




```
ggplot(data = mob , aes(connection_band))+
  geom_histogram(stat="count")+
  ggtitle('Connectivity Generation')+
  xlab('Connection Band')
```



In the mobile brand graph we can see that most of the sales quantity came from samsung then huawei

```
names(mob)
```

```
##  [1] "model_id"        "model_name"      "brand_name"
##  [4] "released_year"   "released_month"  "connection_band"
##  [7] "size_in_inch"    "camera_mb"       "storage"
## [10] "ram"             "sellout"         "price"
## [13] "shop_id"         "channel"         "governorate"
## [16] "latitude"        "longitude"
```

```
shop_data <- mob %>%
  group_by(shop_id,governorate,latitude,longitude,channel,brand_name)  %>%
  summarise(ttl_quantity = sum(sellout),
            ttl_amount = sum(sellout * price)) %>%
  arrange(shop_id)



shop_data_qnt_wide = dcast(shop_data, shop_id + governorate + latitude +
                         longitude + channel ~ brand_name,
                       value.var="ttl_quantity", fun.aggregate=sum)

shop_data_subtotal = mob %>%
  group_by(shop_id,governorate,latitude,longitude,channel)  %>%
  summarise(ttl_quantity = sum(sellout),
            ttl_amount = sum(sellout * price)) %>%
  arrange(governorate,shop_id)

str(shop_data, give.attr = FALSE)
```

```
## Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame':  5905 obs. of  8 variables:
##  $ shop_id     : int  17 17 17 17 19 19 19 20 20 20 ...
##  $ governorate : Factor w/ 26 levels "Alexandria","Aswan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ latitude    : num  31.2 31.2 31.2 31.2 31.2 ...
##  $ longitude   : num  30 30 30 30 29.9 ...
##  $ channel     : Factor w/ 3 levels "Hyper Market",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ brand_name  : Ord.factor w/ 12 levels "Samsung"<"Huawei"<..: 1 2 3 4 1 2 6 1 2 6 ...
##  $ ttl_quantity: int  278 116 26 11 14 17 5 36 19 3 ...
##  $ ttl_amount  : int  966968 292814 120100 27596 152593 80812 79896 130944 61101 59647 ...
```

```
str(shop_data_subtotal, give.attr = FALSE)
```

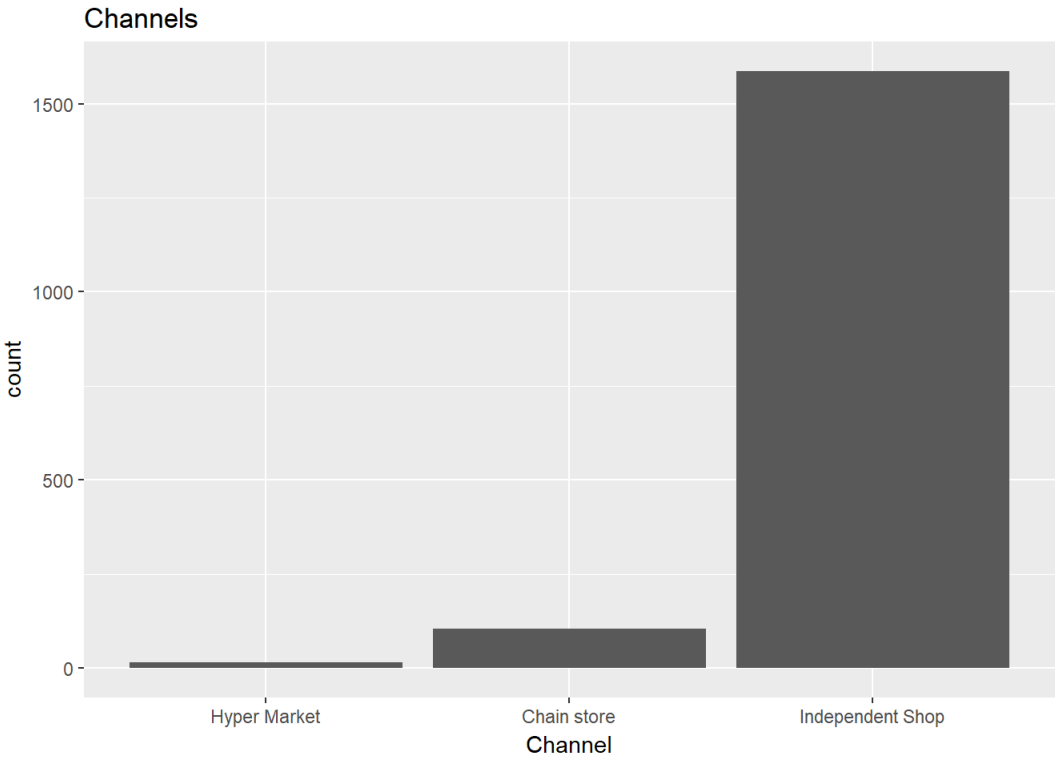
```
## Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame':  1706 obs. of  7 variables:
##  $ shop_id     : int  17 19 20 21 25 27 28 30 31 38 ...
##  $ governorate : Factor w/ 26 levels "Alexandria","Aswan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ latitude    : num  31.2 31.2 31.2 31.2 31.2 ...
##  $ longitude   : num  30 29.9 30 29.9 29.9 ...
##  $ channel     : Factor w/ 3 levels "Hyper Market",..: 3 3 3 2 3 1 1 1 3 1 ...
##  $ ttl_quantity: int  431 36 58 42 68 22 40 15 60 3179 ...
##  $ ttl_amount  : int  1407478 313301 251692 313839 165722 56562 112563 36577 189299 9996980 ...
```

```
str(shop_data_qnt_wide, give.attr = FALSE)
```

```
## 'data.frame':    1706 obs. of  17 variables:
##  $ shop_id    : int  17 19 20 21 25 27 28 29 30 31 ...
##  $ governorate: Factor w/ 26 levels "Alexandria","Aswan",..: 1 1 1 1 1 1 1 15 1 1 ...
##  $ latitude   : num  31.2 31.2 31.2 31.2 31.2 ...
##  $ longitude  : num  30 29.9 30 29.9 29.9 ...
##  $ channel    : Factor w/ 3 levels "Hyper Market",..: 3 3 3 2 3 1 1 1 1 3 ...
##  $ Samsung    : int  278 14 36 42 68 19 31 5 8 40 ...
##  $ Huawei     : int  116 17 19 0 0 1 7 1 4 20 ...
##  $ Oppo       : int  26 0 0 0 0 0 0 0 0 0 ...
##  $ Nokia      : int  11 0 0 0 0 0 0 0 0 0 ...
##  $ Lenovo     : int  0 0 0 0 0 2 2 1 0 0 ...
##  $ Apple      : int  0 5 3 0 0 0 0 0 0 0 ...
##  $ Sony       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Htc        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Motorola   : int  0 0 0 0 0 0 0 2 3 0 ...
##  $ Gionee     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Alcatel    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Asus       : int  0 0 0 0 0 0 0 0 0 0 ...
```

Our Shops are devided to 3 main Channels

```
ggplot(data = shop_data_subtotal , aes(x = channel))+
  geom_histogram(stat="count",binwidth = 1)+
  ggtitle('Channels')+
  xlab('Channel')
```

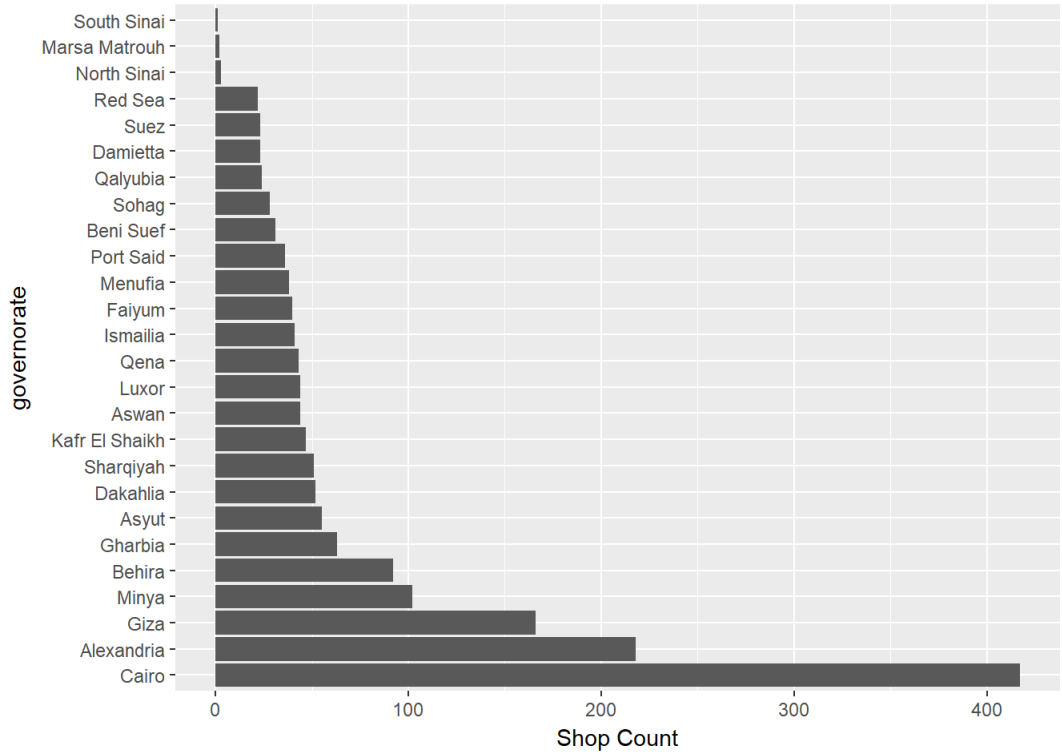



```
table(shop_data_subtotal$channel)
```

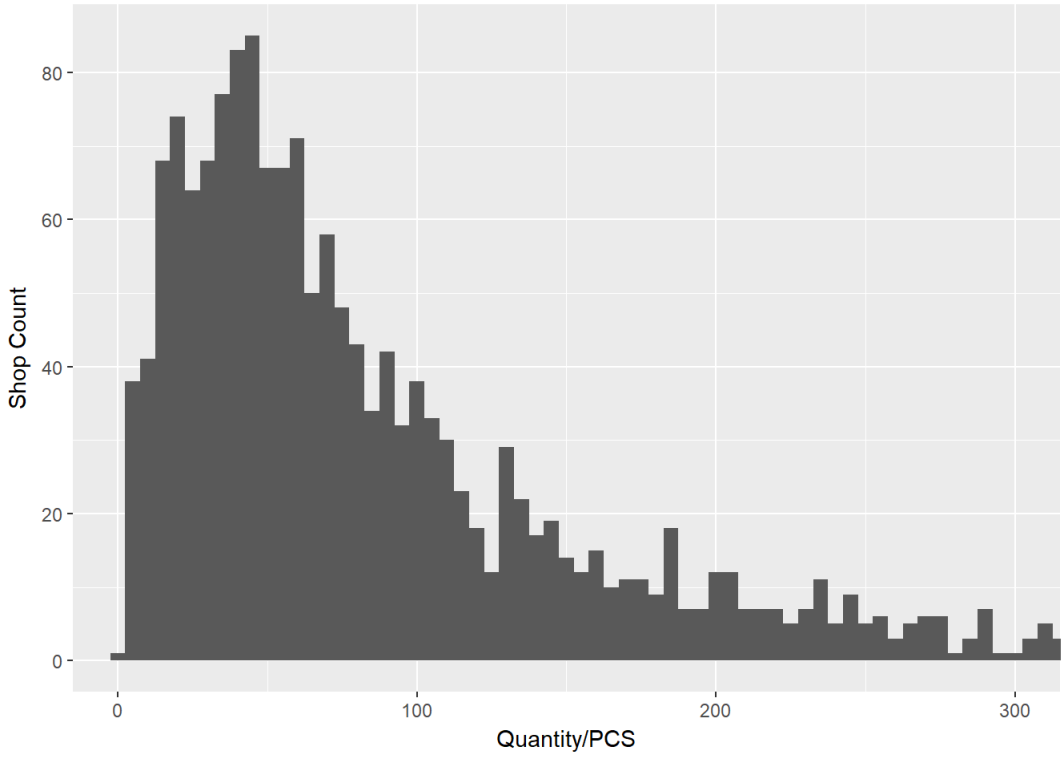
```
##
##     Hyper Market      Chain store Independent Shop
##               15              104             1587
```

As we want to summarize the transactions to each shop we will summary the transactions for each shop

```
# sort governorate before histogram
gov_count = sort(table(shop_data_subtotal$governorate), decreasing = T)

shop_data_subtotal$governorate = ordered(shop_data_subtotal$governorate ,
                                         levels = names(gov_count))

ggplot(data = shop_data_subtotal ,
       aes(x = governorate ))+
  geom_histogram(stat="count") + coord_flip() +
  ylab('Shop Count')
```



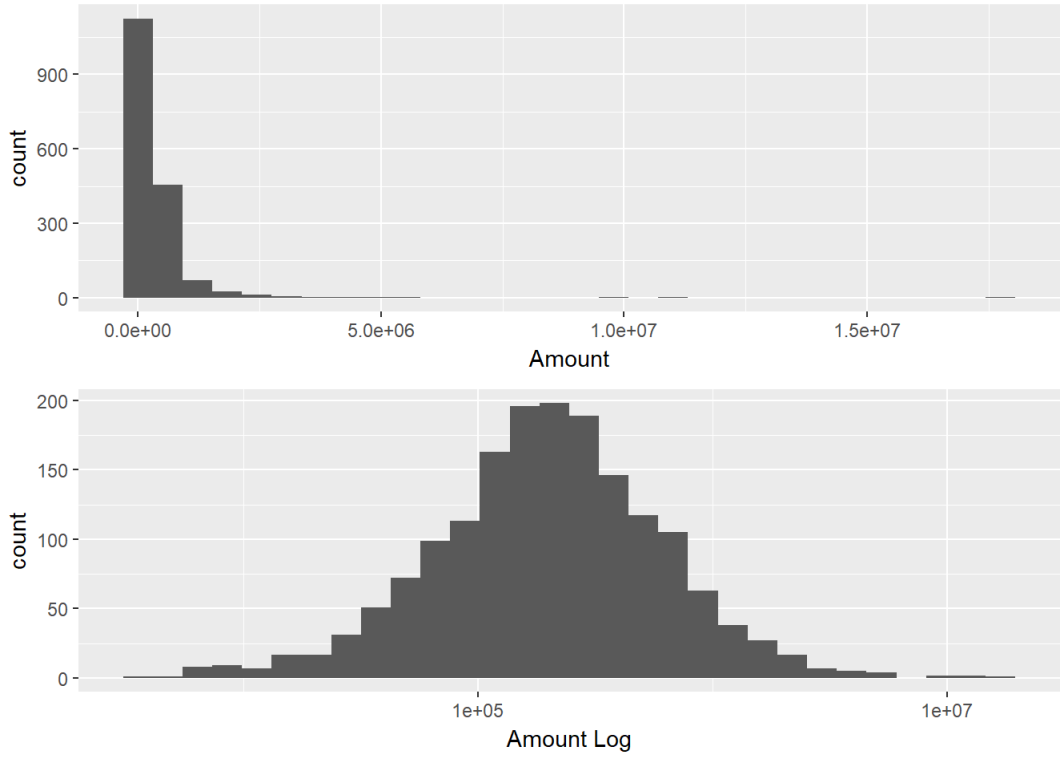

```
ggplot(data = shop_data_subtotal , aes(x = ttl_quantity))+
  geom_histogram(binwidth = 5)+
  coord_cartesian(xlim = c(0,300))+
  ylab('Shop Count') + xlab('Quantity/PCS')
```




```
summary(shop_data_subtotal$ttl_quantity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    37.0    67.0   121.2   130.0  5704.0
```

```
amount_p =ggplot(data = shop_data_subtotal , aes(x = ttl_amount))+
  geom_histogram() + xlab('Amount')
  # coord_cartesian(xlim = c(0,3000000))
amount_log_p = amount_p + scale_x_log10() + xlab('Amount Log')

grid.arrange(amount_p,amount_log_p,ncol = 1)
```




```
summary(shop_data_subtotal$ttl_amount)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     3734   101951   202223   383337   397831 17745933
```

Univariate Analysis

What is the structure of your dataset?

This dataset is about most usable mobile phone in Egypt containing the prices, sellout transaction and other attributes of almost 200 famous Mobile that sold in 1700 shop

(Biggest) ——> (Smallest) Channel : (Hyper Market (bigger), Chain store, Independent Shop (Small))

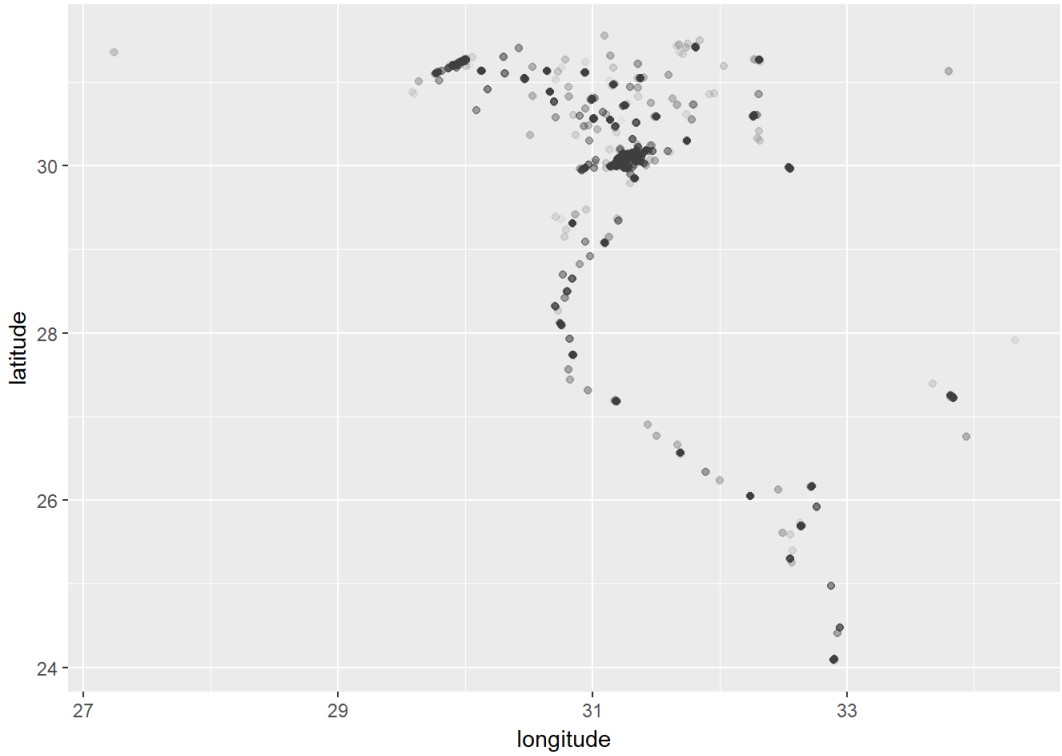
What is/are the main feature(s) of interest in your dataset?

The main Features in this data are : 1-The Amount of the sales and quantity in each shop. I'd like to compare each geogrfic location with its sales 2-The best selling mobile phone and what is the price segment that make more sales 3-What could effict the price of each model ### What other features in the dataset do you think will help support your
investigation into your feature(s) of interest? 1-The channel type of each shop and the location of the shop whether its on capital cities or rural 2-The ram,storage,screen size will help to determine the best mobile ### Did you create any new variables from existing variables in the dataset? I added the amount of each transaction as the quantity soldout * the price ### Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this? I Created two main sets from this data 1-The shop and its detail with total quantity of sales and the amount 2-The mobile sales with its detail

Bivariate Plots Section

We have the longitude and latitude for our sample Shops

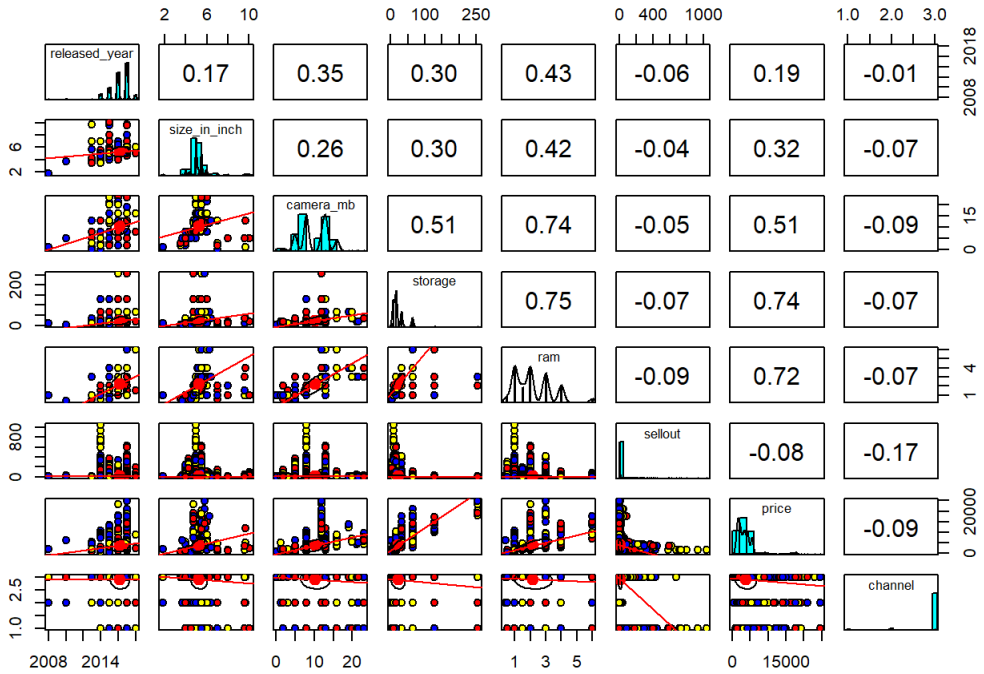
```
ggplot(aes(x=longitude, y=latitude), data=mob) +
    geom_point(alpha = .008)
```

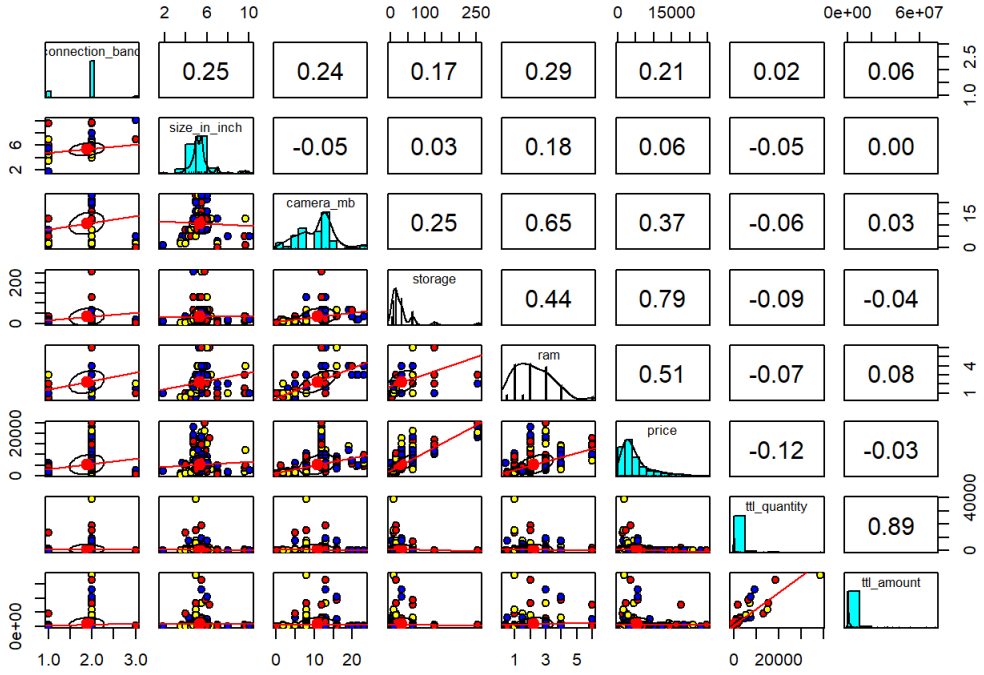


1. The more point is darker the more transaction happens
2. Darker point are centralized cities or capital

```
mobile_data <- mob %>%
  group_by(model_id,model_name,brand_name,released_year,released_month,
           connection_band,size_in_inch,camera_mb,storage,ram,price)  %>%
  summarise(ttl_quantity = sum(sellout),
            n = n()) %>%
  arrange(brand_name,model_name)
mobile_data$ttl_amount <- (mobile_data$ttl_quantity * mobile_data$price)
str(mobile_data,give.attr = FALSE)
```

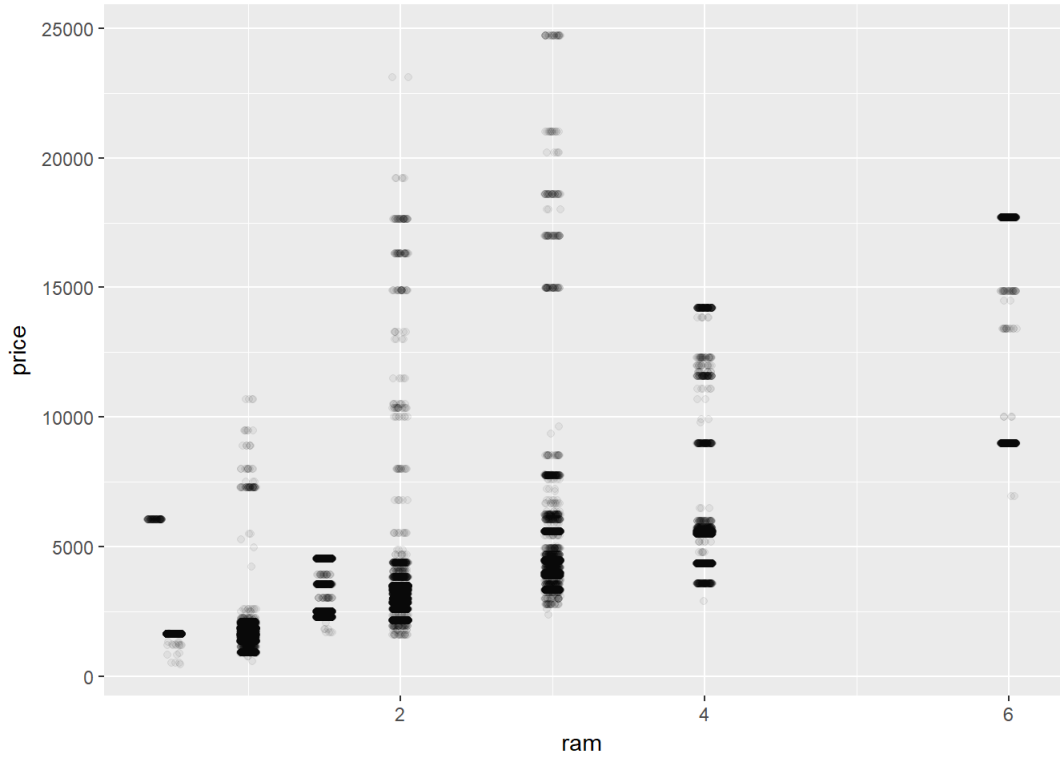
```
## Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame':  224 obs. of  14 variables:
##  $ model_id       : int  3294 7789 8494 6763 7759 8335 8790 6478 6708 6907 ...
##  $ model_name     : Factor w/ 201 levels "2","3","5","6",..: 42 43 44 45 46 47 48 49 50 51 ...
##  $ brand_name     : Ord.factor w/ 12 levels "Samsung"<"Huawei"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ released_year  : int  2010 2015 2017 2015 2015 2017 2018 2014 2014 2015 ...
##  $ released_month : int  5 12 1 2 12 1 1 8 10 2 ...
##  $ connection_band: Factor w/ 3 levels "3G","4G","Wifi": 1 2 2 2 2 2 2 2 2 1 ...
##  $ size_in_inch   : num  3.7 5.2 5.2 5.5 5.5 5.7 6 4 5 4.3 ...
##  $ camera_mb      : num  5 13 16 13 13 16 16 5 8 5 ...
##  $ storage        : int  1 16 32 16 16 32 64 4 8 4 ...
##  $ ram            : num  0.384 2 3 3 3 3 6 1 1 0.512 ...
##  $ price          : int  6059 6816 6249 6050 8533 7749 8999 783 1867 1261 ...
##  $ ttl_quantity   : int  268 15 300 228 32 324 1147 1 38871 2 ...
##  $ n              : int  105 6 69 76 22 138 314 1 1634 1 ...
##  $ ttl_amount     : int  1623812 102240 1874700 1379400 273056 2510676 10321853 783 72572157 2522 ...
```





From this data we notice that correlations between the price and the ram , storage are near to be strong while its mid in the camera quality

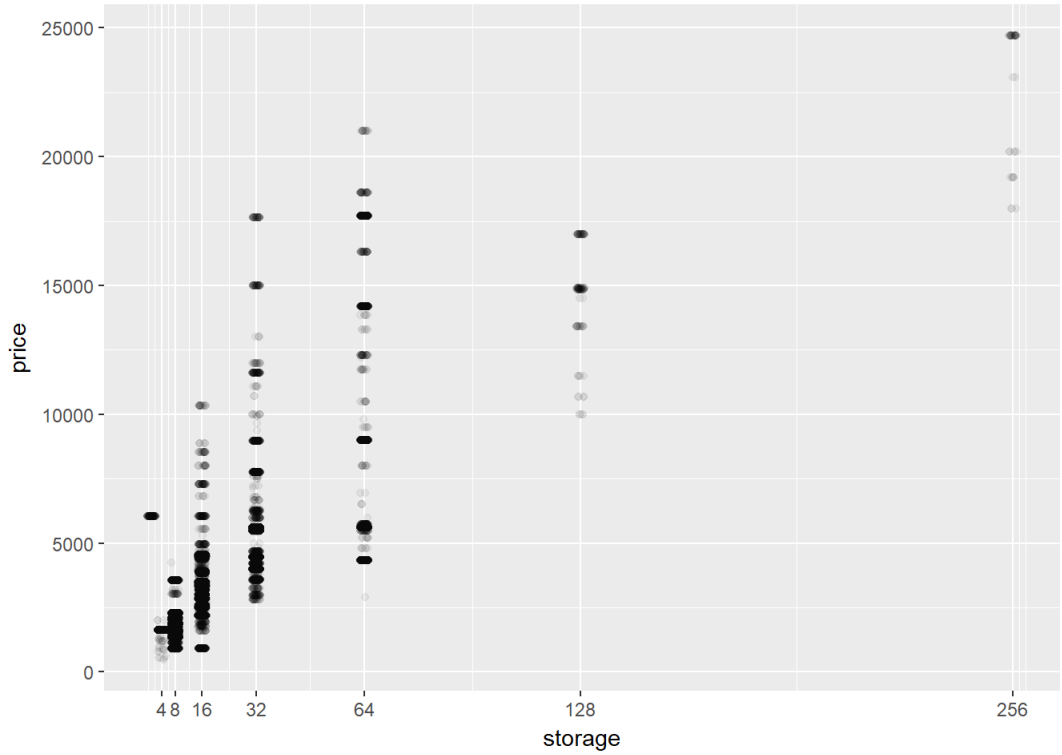
```
ggplot(data = mob,
       aes(x = ram ,y = price ))+
  geom_point(alpha = .05, position = position_jitter(h = 0))
```




```
cor.test(mob$ram, mob$price)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mob$ram and mob$price
## t = 167.2, df = 26551, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7102559 0.7219747
## sample estimates:
##       cor
## 0.7161658
```

```
ggplot(data = mob,
       aes(x = storage ,y = price ))+
  geom_point(alpha = .05, position = position_jitter(h = 0))+
  scale_x_continuous(breaks = c(4,8,16,32,64,128,256))
```

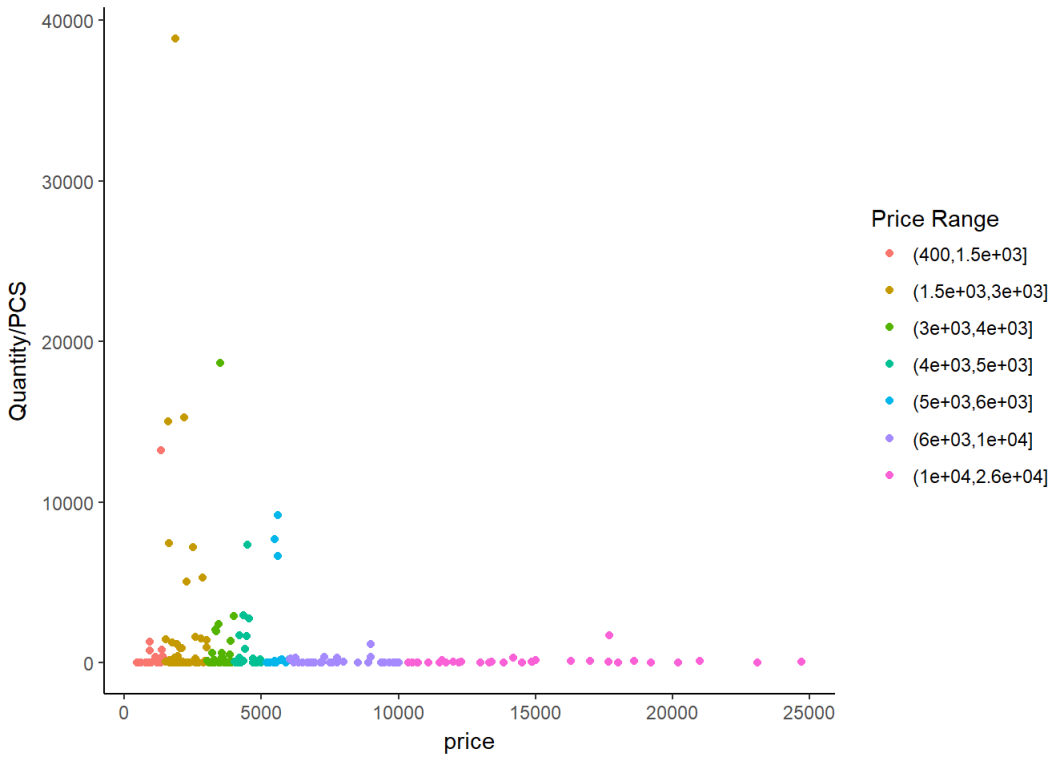



```
cor.test(mob$storage, mob$price)
```

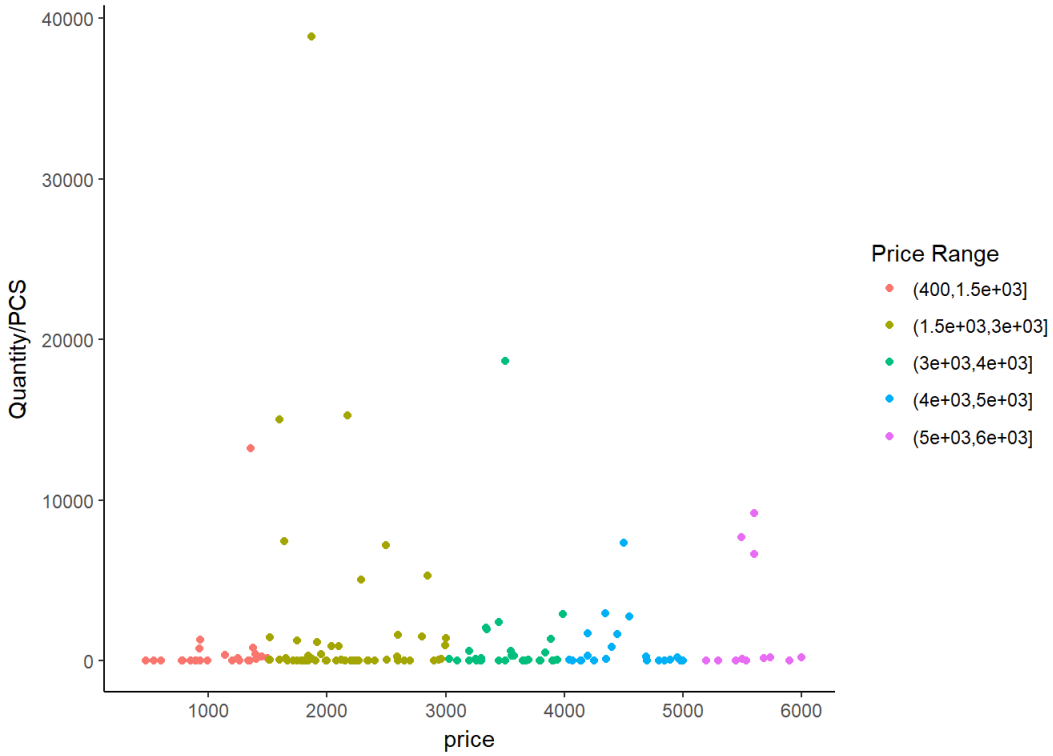
```
##
##  Pearson's product-moment correlation
##
## data:  mob$storage and mob$price
## t = 179.2, df = 26551, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7343652 0.7452541
## sample estimates:
##       cor
## 0.7398581
```

The tall vertical strips indicate Storage and Ram values are mostly integers. Adding jitter, transparency, and changing the plot limits lets us see the slight correlation between table and price.

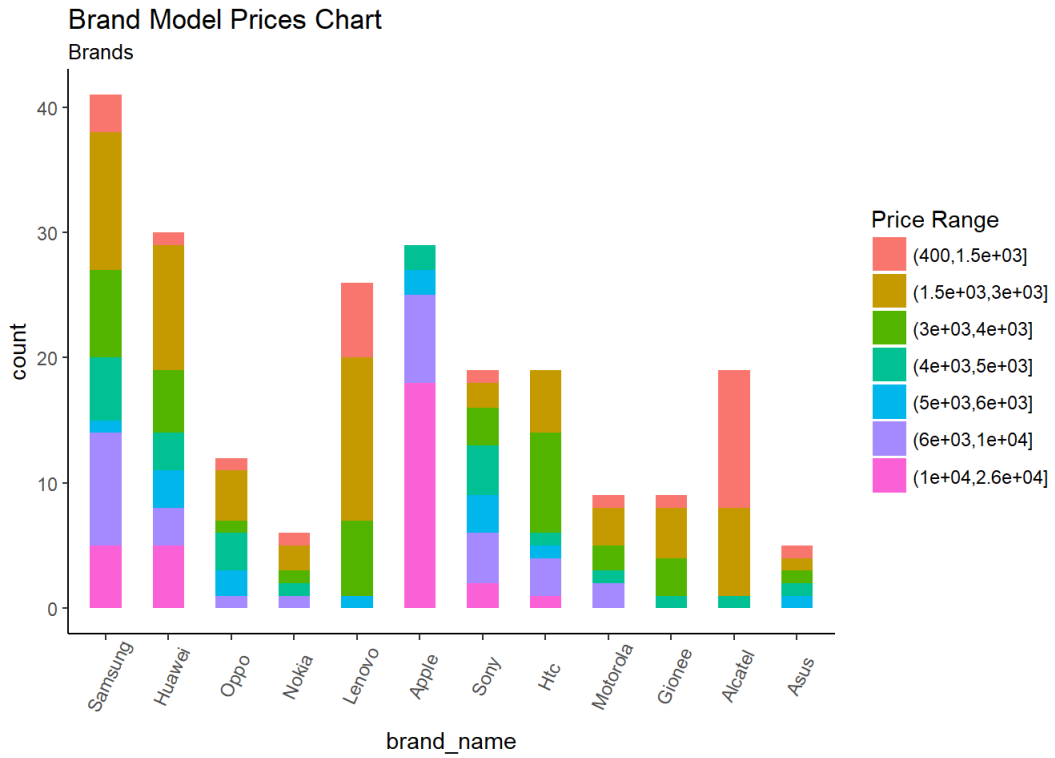
```
Palette <- c('#003399','#f02f53','#ff664f','#b2ae85','#66CCFF','#4c4c4c',
             '#ffef96','#00a86b','#CCFFCC','#c2e8c8','#b61e98','#ffbcc8')
theme_set(theme_classic())

mobile_data$price.bucket <- cut(mobile_data$price,breaks =
                                  c(400,1500,3000,4000,5000,6000,10000,26000))

ggplot(data = mobile_data,
       aes(x = price ,y = ttl_quantity ))+
  geom_point(aes(color = price.bucket), stat = 'summary' , fun.y = sum )+
  guides(size = F, fill = guide_legend(override.aes = list(size=3)))+
  ylab('Quantity/PCS')+
  labs(color = "Price Range")
```



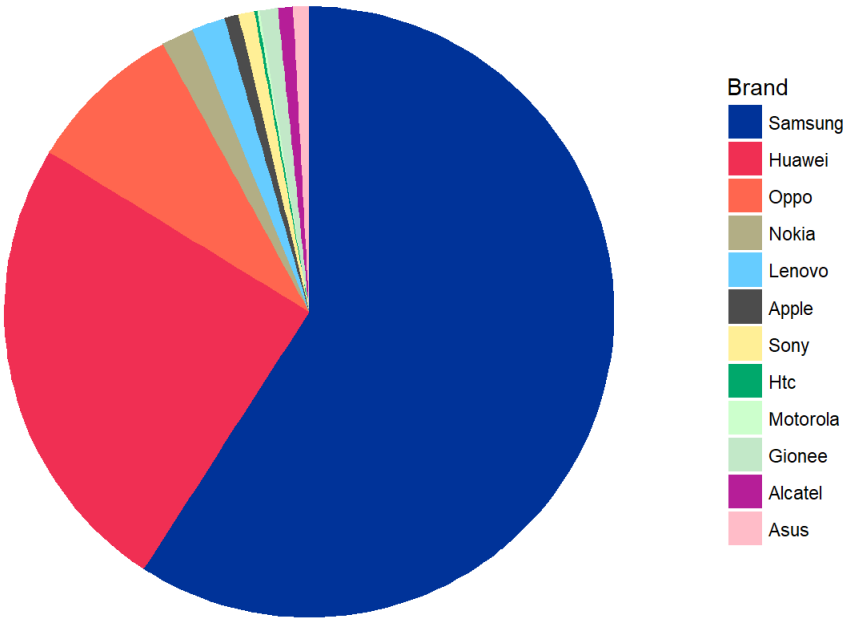

```
ggplot(data = mobile_data,
       aes(x = price ,y = ttl_quantity ))+
  geom_point(aes(color = price.bucket), stat = 'summary' , fun.y = sum )+
  ylab('Quantity/PCS')+
  scale_x_continuous(breaks = seq(0,6000,1000) , limits = c(400,6000))+
  guides(size = F, fill = guide_legend(override.aes = list(size=3)))+
  labs(color = "Price Range")
```




```
ggplot(mobile_data, aes(brand_name))+
  geom_bar(aes(fill=price.bucket), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title ="Brand Model Prices Chart",
       subtitle ="Brands",
       fill = "Price Range")
```




```
ggplot(mobile_data, aes(x = "", y = ttl_quantity,
                        fill = fct_inorder(brand_name))) +
       geom_bar(width = 1, stat = "identity") +
       coord_polar("y", start = 0, direction = -1) + theme_void()+
       geom_label_repel(aes(label = ''),
                        size=5, show.legend = F, nudge_x = 5) +
       guides(fill = guide_legend(title = "Brand"))+

  scale_fill_manual(values= Palette )
```

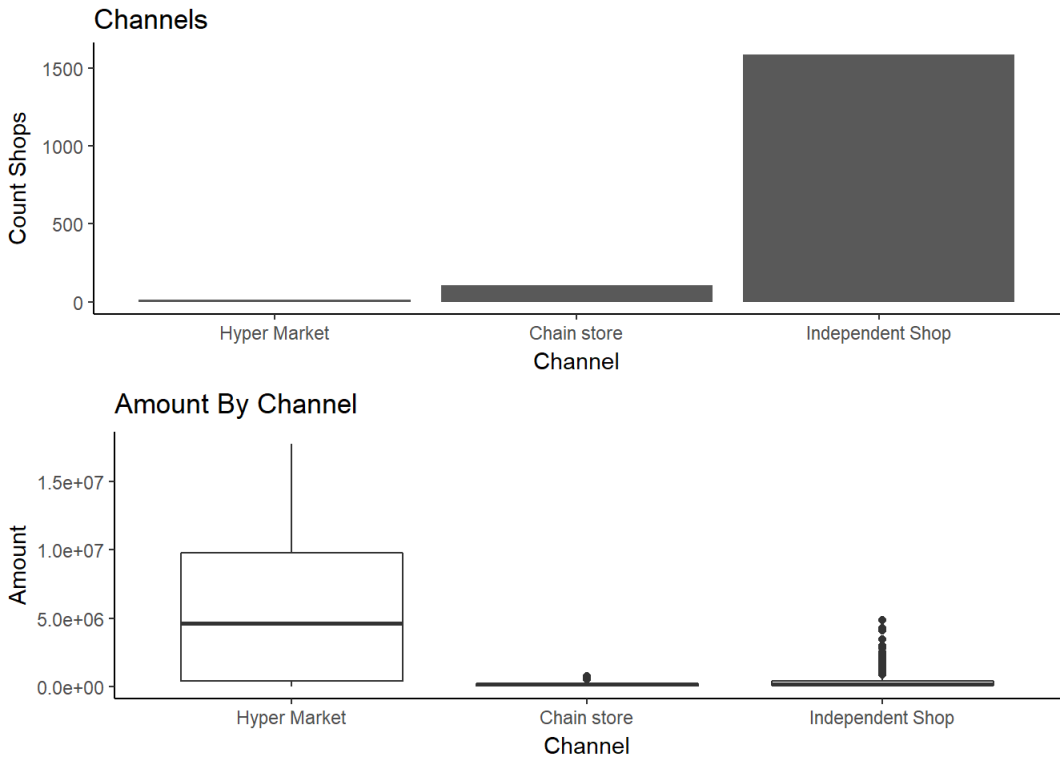


```
# ggplot(data = mobile_data)+
#   geom_bar(aes(x = "", y = ttl_quantity, fill = brand_name),
#            stat ="identity", width = 1)+
#   coord_polar("y", start=0, direction = -1)+
#   theme_void()+
#        guides(fill = guide_legend(title = "Brand"))+
#   scale_fill_manual(values = Palette )
```

As we can see from this figure that the most of the sales are for the mobiles below 6000 pound and most of the sales came from Samsung then Huawei

Next, I will look at how is the sales on these shops and where it is on geographic map and how its distributed through different channels and which brand has the most sales.

```
p_channel_cnt = ggplot(data = shop_data_subtotal , aes(x = channel))+
  geom_histogram(stat="count",binwidth = 1)+
  ggtitle('Channels')+xlab('Channel')+ylab('Count Shops')

p_amount_chnl = ggplot(data = shop_data_subtotal ,
                       aes(x =channel ,y = ttl_amount))+
  geom_boxplot()+
  ggtitle('Amount By Channel')+xlab('Channel')+ylab('Amount')

grid.arrange(p_channel_cnt,p_amount_chnl,ncol = 1)
```



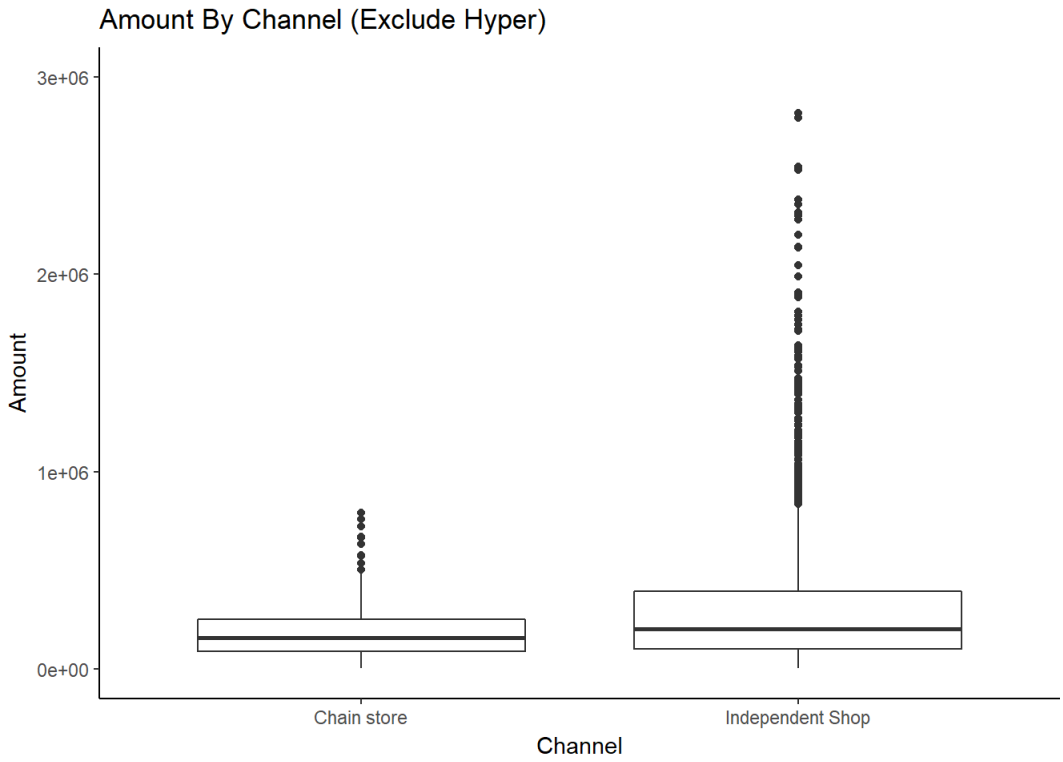

```
prop.table(table(shop_data_subtotal$channel))
```

```
##
##     Hyper Market      Chain store Independent Shop
##      0.008792497      0.060961313      0.930246190
```

```
sum(subset(shop_data_subtotal,channel != 'Hyper Market' ,
           select = c('ttl_amount')))/sum(shop_data_subtotal$ttl_amount)
```

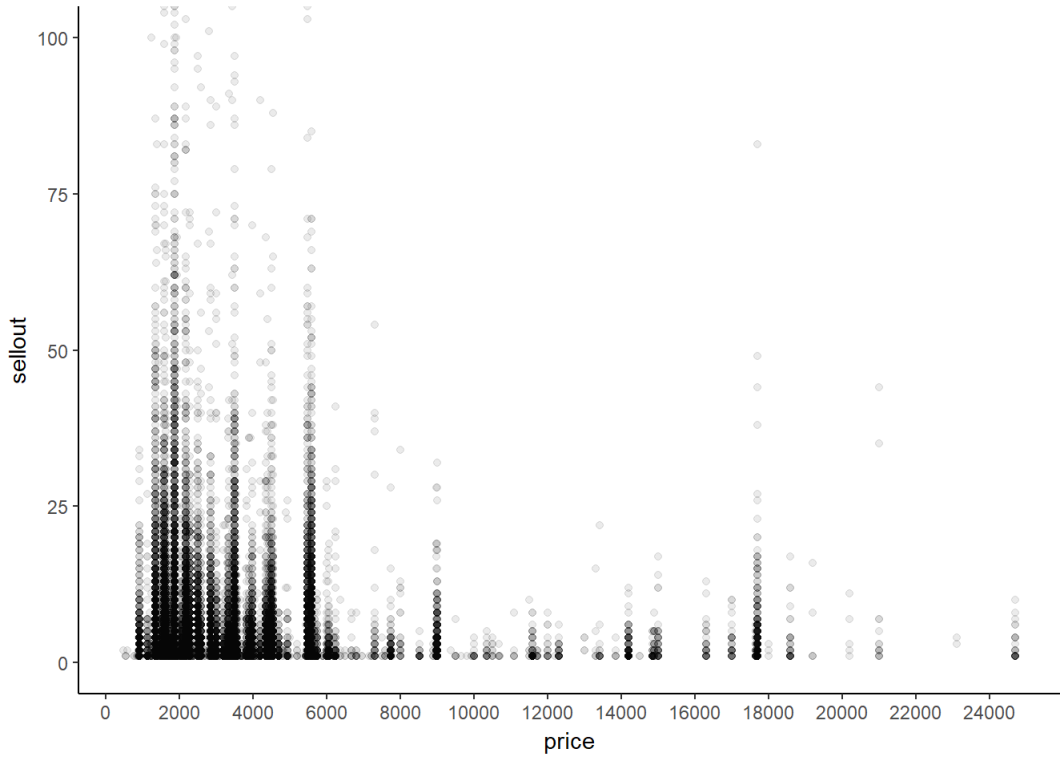
```
## [1] 0.8780856
```

```
ggplot(data = subset(shop_data_subtotal,channel != 'Hyper Market') ,
       aes(x =channel ,y = ttl_amount))+
  geom_boxplot()+scale_y_continuous(limits = c(0,3000000))+
  ggtitle('Amount By Channel (Exclude Hyper)')+xlab('Channel')+ylab('Amount')
```

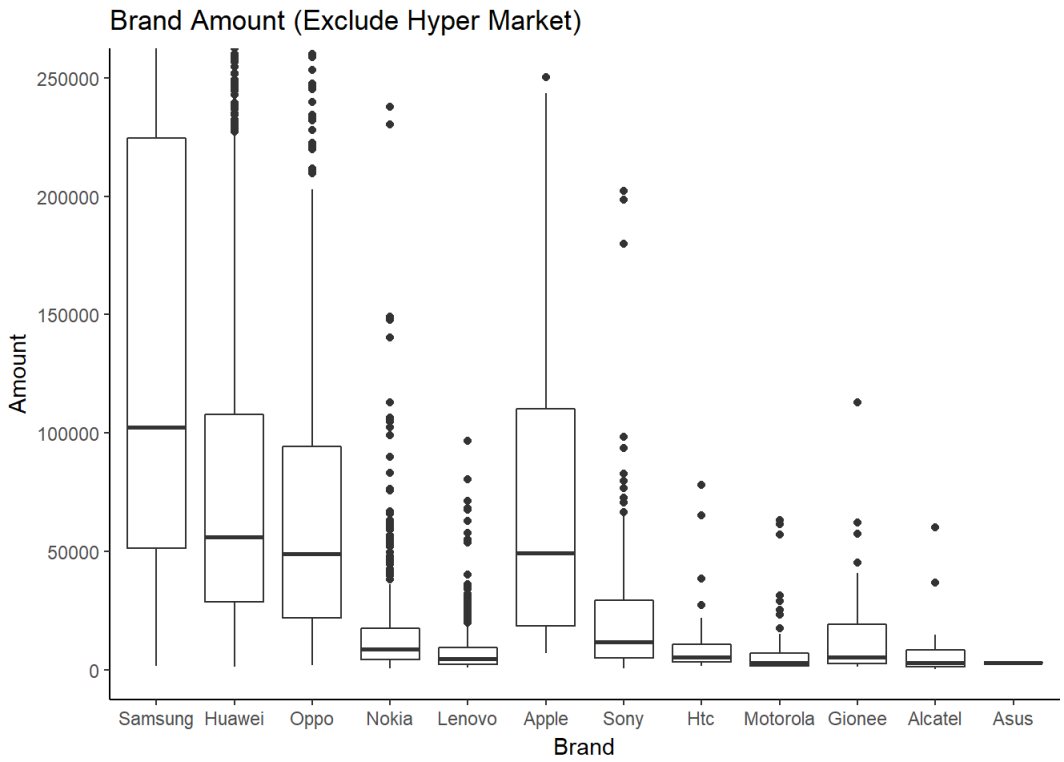


Despite that the hypermarket shops not reach to 1 % but it controls most of sales volume with amount 88%

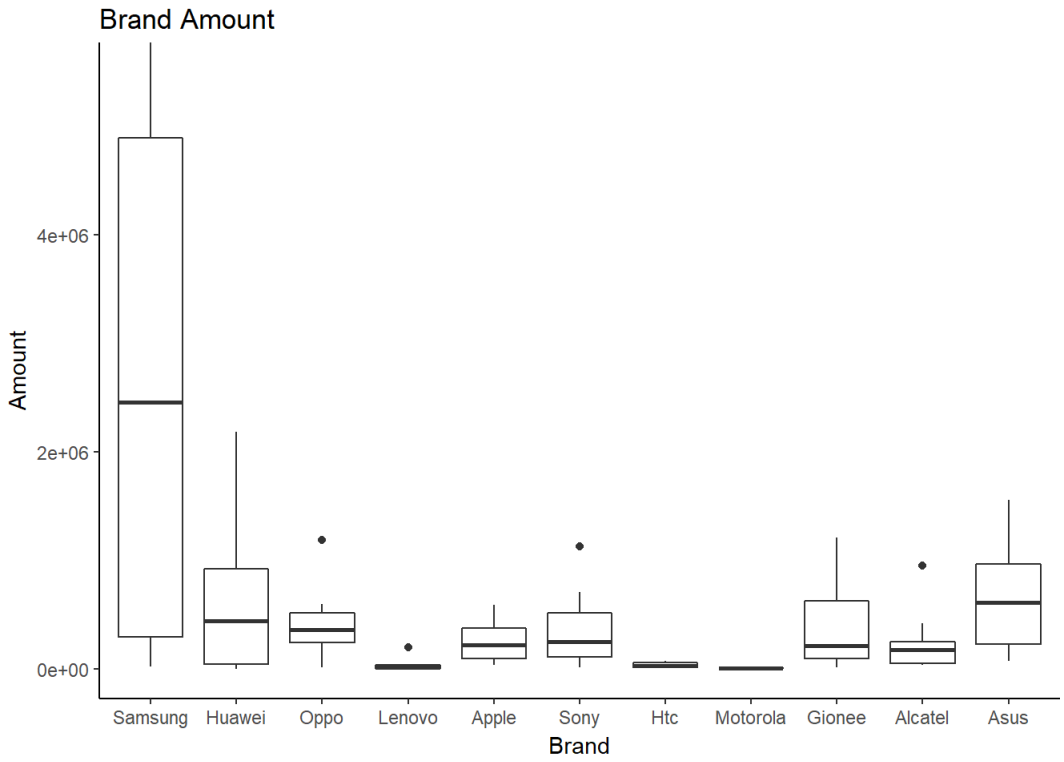
```
ggplot(data = mob,
       aes(x =price ,y =sellout ))+
  geom_point(alpha = .08)+
  scale_x_continuous(breaks = seq(0,26000,2000))+
  coord_cartesian(ylim = c(0,100))
```



```
# ggplot(data = shop_data ,
#        aes(x =brand_name ,y =ttl_amount ))+
#   geom_boxplot()+
#
#   ggtitle('Brand Amount')+
#   ylab('Amount')+xlab('Brand')



ggplot(data = subset(shop_data,channel != 'Hyper Market'),
       aes(x =brand_name ,y =ttl_amount ))+
  geom_boxplot()+
  coord_cartesian(ylim = c(0,250000))+
  ggtitle('Brand Amount (Exclude Hyper Market)')+
  ylab('Amount')+xlab('Brand')
```




```
ggplot(data = subset(shop_data,channel == 'Hyper Market'),
       aes(x =brand_name ,y =ttl_amount ))+
  geom_boxplot()+
  coord_cartesian(ylim = c(0,5500000))+
  ggtitle('Brand Amount')+
  ylab('Amount')+xlab('Brand')
```




```
summary(shop_data[shop_data$channel != 'Hyper Market',c('ttl_amount')])
```

```
##    ttl_amount
##  Min.   :    539
##  1st Qu.:  12480
##  Median :  43620
##  Mean   :  99264
##  3rd Qu.: 106197
##  Max.   :4203798
```

```
summary(shop_data[shop_data$channel == 'Hyper Market',c('ttl_amount')])
```

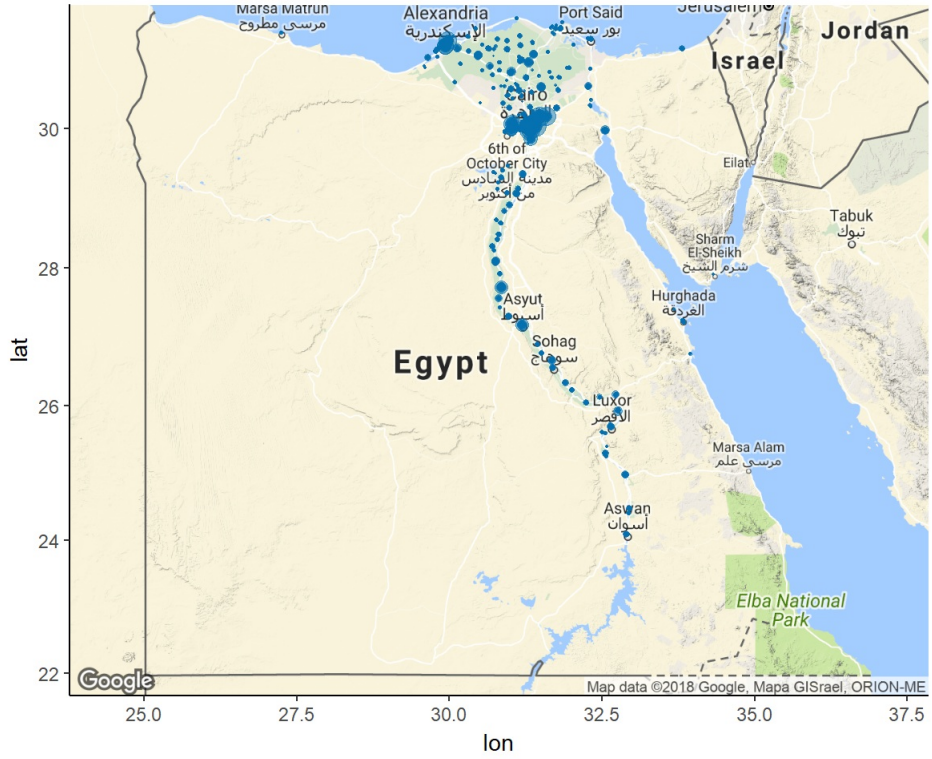
```
##    ttl_amount
##  Min.   :    929
##  1st Qu.:  42647
##  Median : 200851
##  Mean   : 664407
##  3rd Qu.: 592522
##  Max.   :9096018
```

These graphs shows us that the most sales is for the mobiles below 6000 EGP and most of the sales comes from samsung specially from Hyper market as there were a gift for each mobile sold

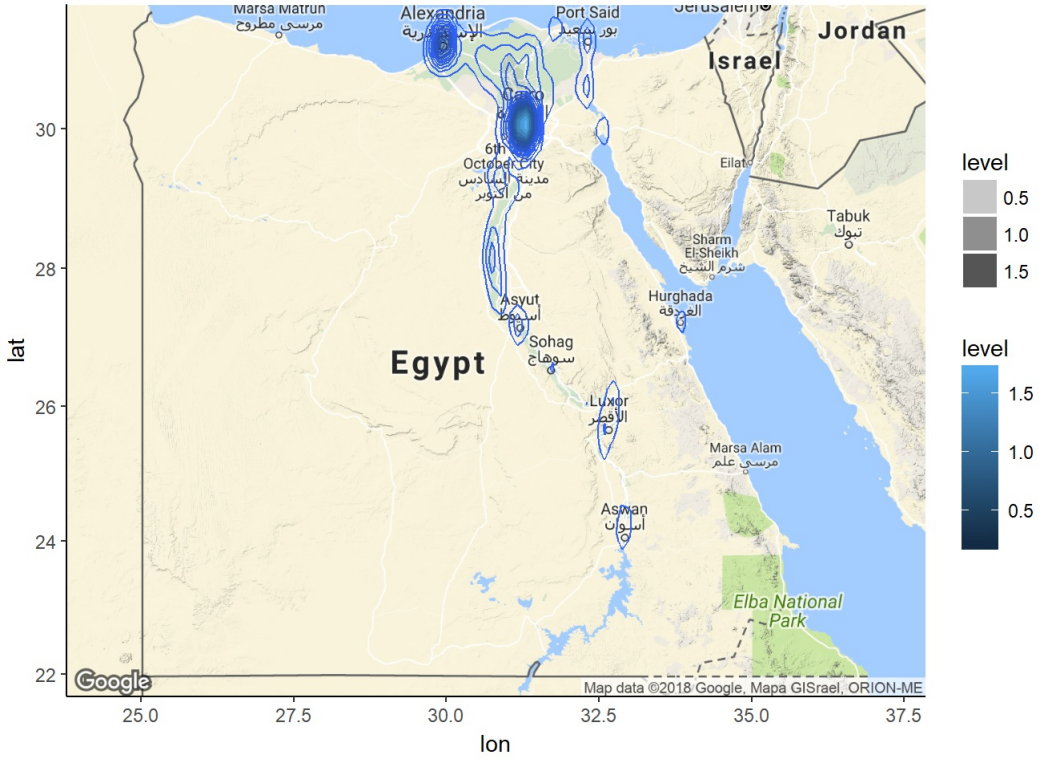
Load the maps

```
egypt <- get_googlemap("Egypt",zoom=6,maptype="terrain",source="google",
                       size = c(640,515),scale = 2)

alex <- get_googlemap("alexandria",zoom=11,maptype="terrain",source="google",
                      size = c(640,440),scale = 2)


cairo <- get_googlemap("cairo",zoom=10,maptype="terrain",source="google",
                       size = c(640,440),scale = 2)
```

```
ggmap(egypt)+
  geom_point(data = mob , aes(x =longitude ,y =latitude ) , color = "#0571b0",
             alpha = .5,size = sqrt(mob$sellout/pi)/4 )
```

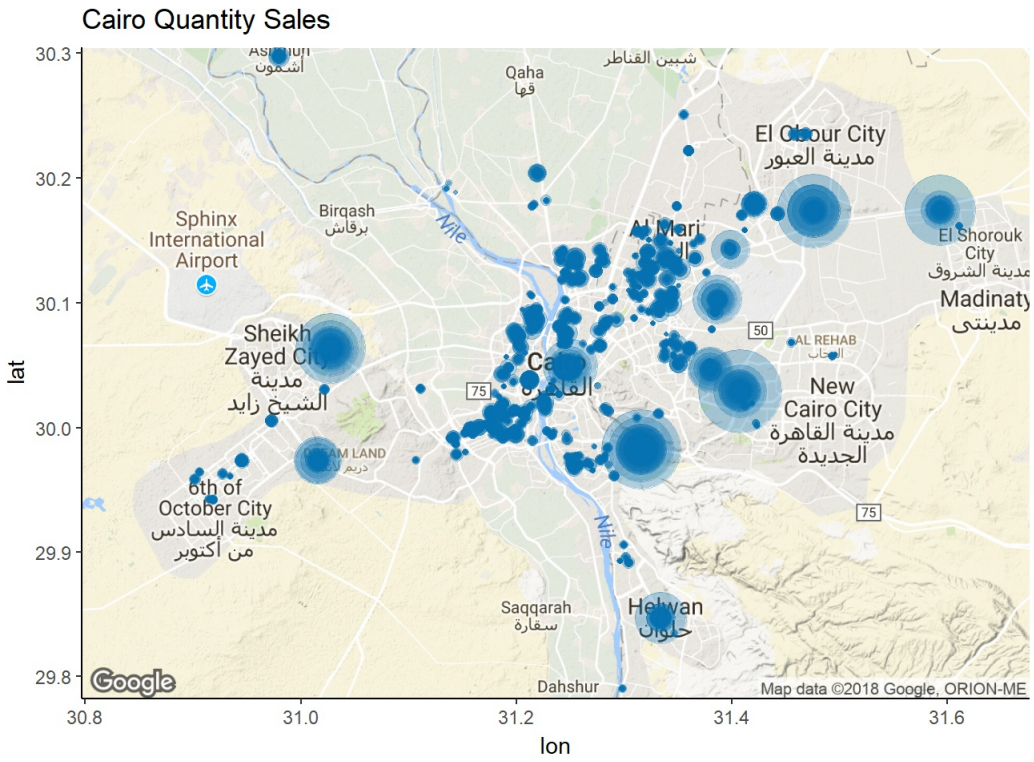



```
ggmap(egypt) +
  geom_density2d(data=shop_data_subtotal,aes(x=longitude,y=latitude), bins=30) +
  stat_density2d(data=shop_data_subtotal,aes(x=longitude,y=latitude,
                                             fill=..level..,alpha=..level..),
                 geom='polygon')
```

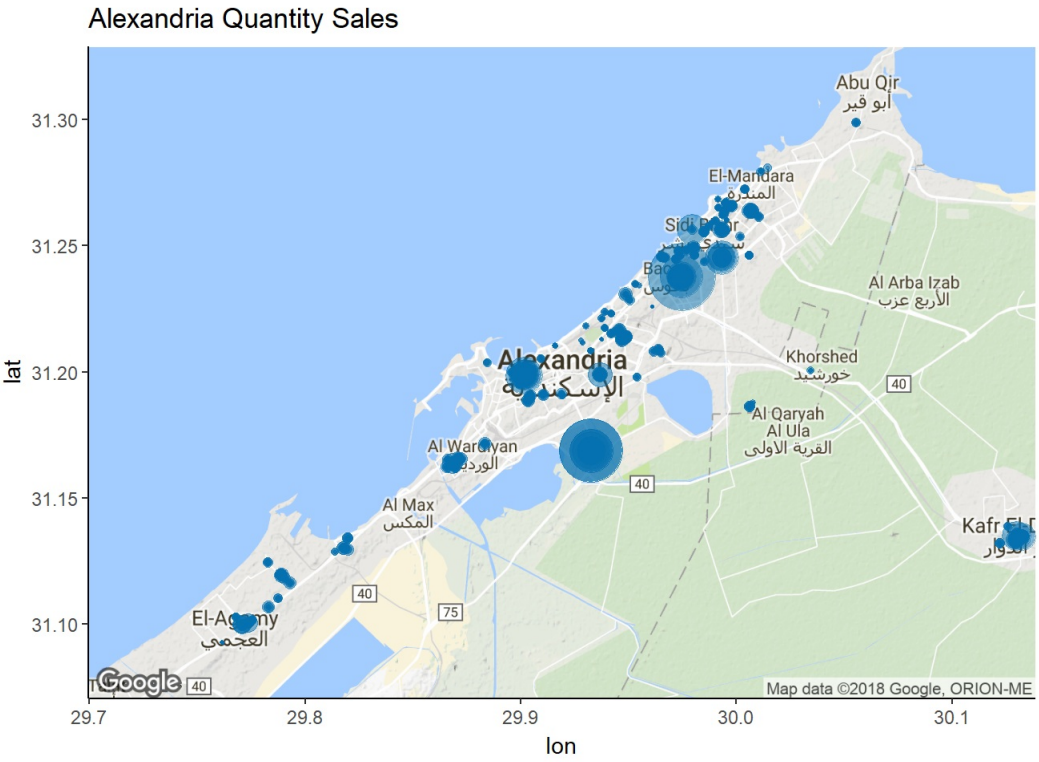


As we see from the above that the most sales are happend in two main cities The biggest one is the Cairo (Capital of Egypt in the center) and Alexandria as these are urban cities and low volume in other cities as they are rural cities

```
ggmap(cairo)+
  geom_point(data = mob , aes(x =longitude ,y =latitude ) , color = "#0571b0",
             alpha = .3,size = sqrt(mob$sellout/pi) )+
  ggtitle('Cairo Quantity Sales')
```




```
ggmap(alex)+
  geom_point(data = mob , aes(x =longitude ,y =latitude ) , color = "#0571b0",
             alpha = .5,size = sqrt(mob$sellout/pi) )+
  ggtitle('Alexandria Quantity Sales')
```



Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

The sales volume increase in Hypermarkets and It controle the market while its not reach 1% of total market

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

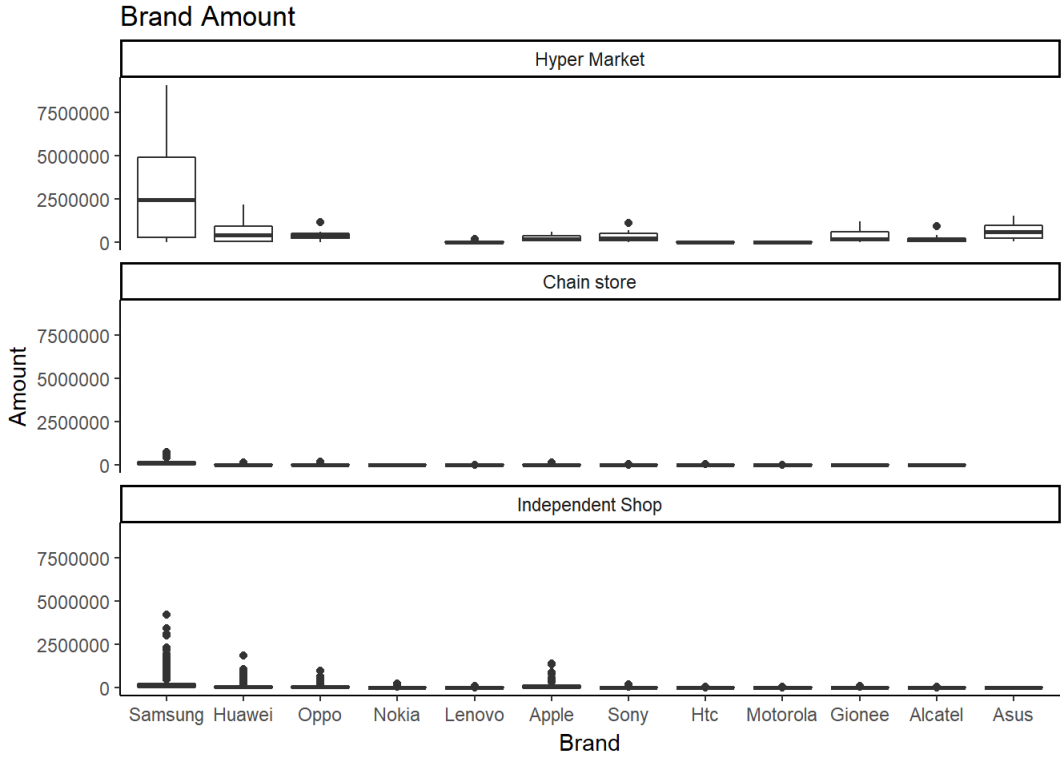
As the Market is bigger the amount and quantity of sales

What was the strongest relationship you found?

The relationship between the ram size and the price the larger ram size the larger price and same for the storage

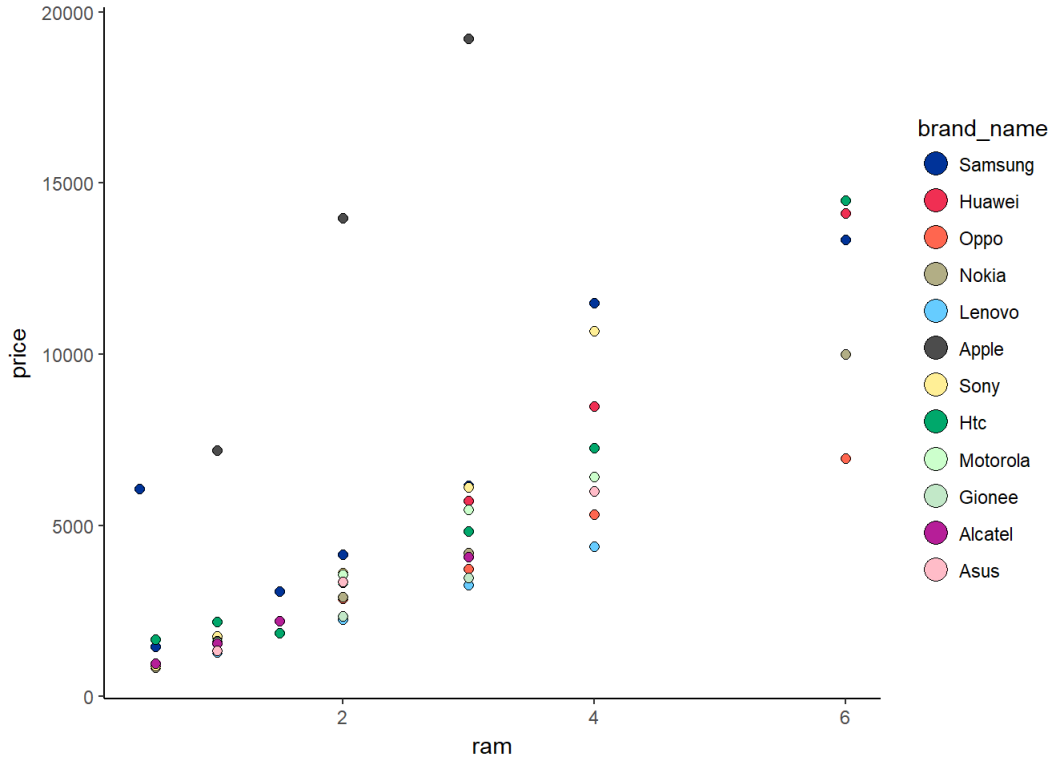
Multivariate Plots Section

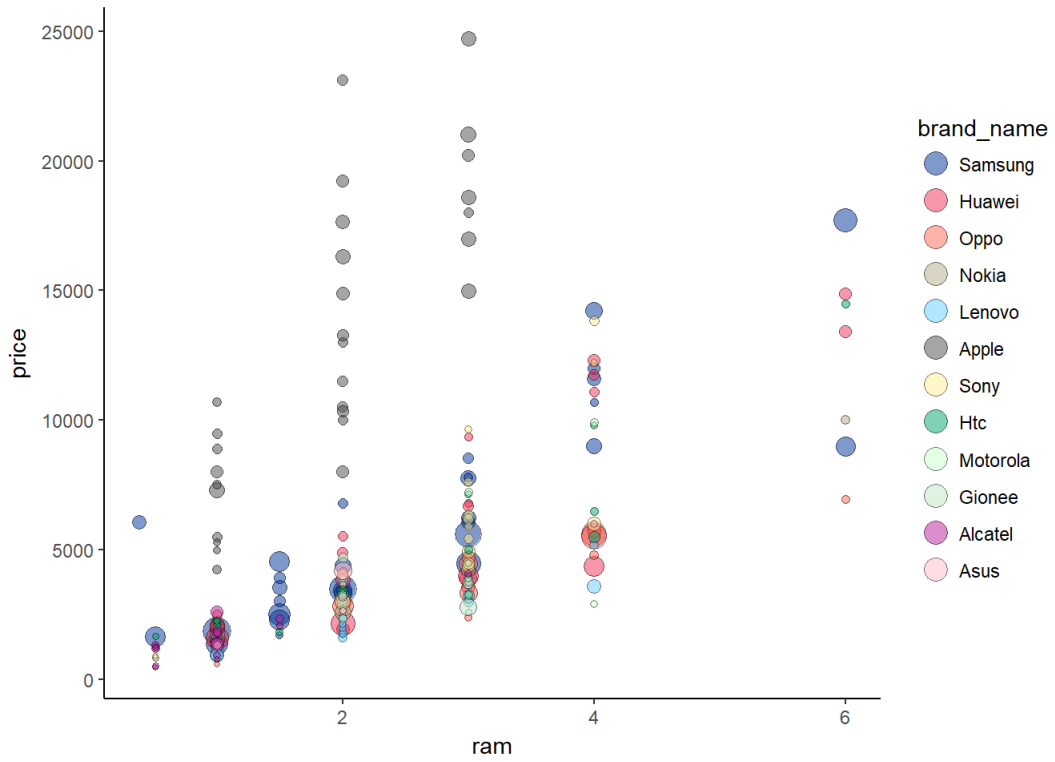
```
ggplot(data = shop_data, aes(x =brand_name ,y =ttl_amount)) +
  geom_boxplot()+
  facet_wrap(~channel , ncol=1)+
  ggtitle('Brand Amount')+
  ylab('Amount')+xlab('Brand')
```

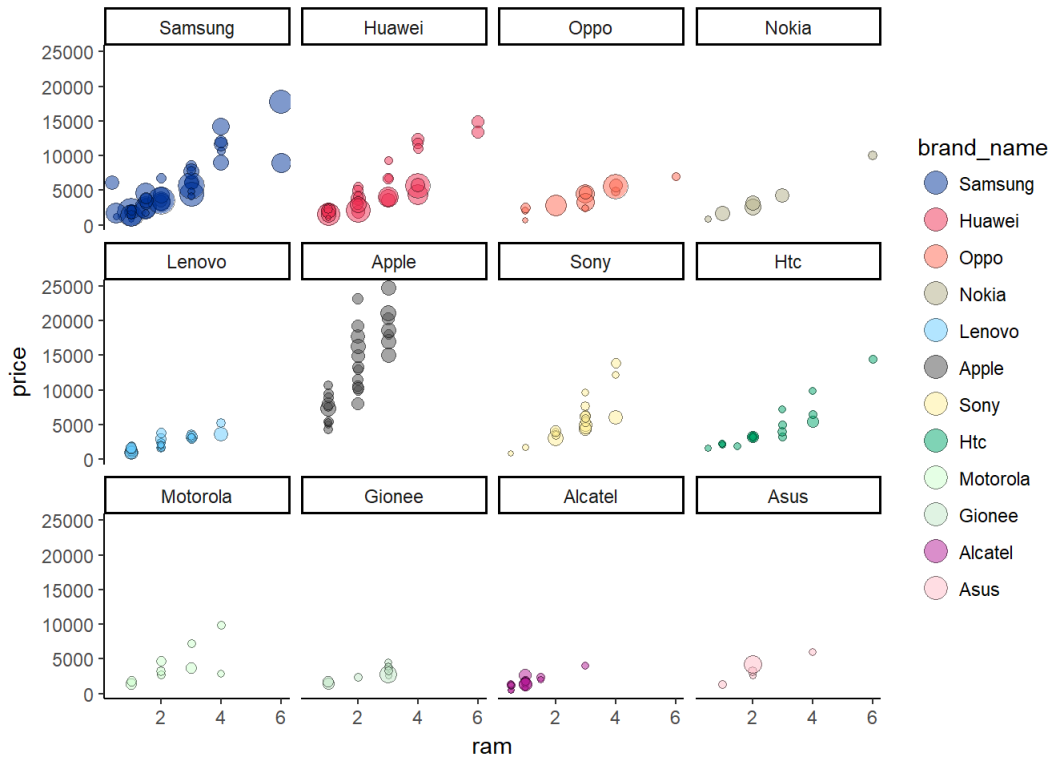


As we can see that sales are booming in hyper market specially for samsung mobiles

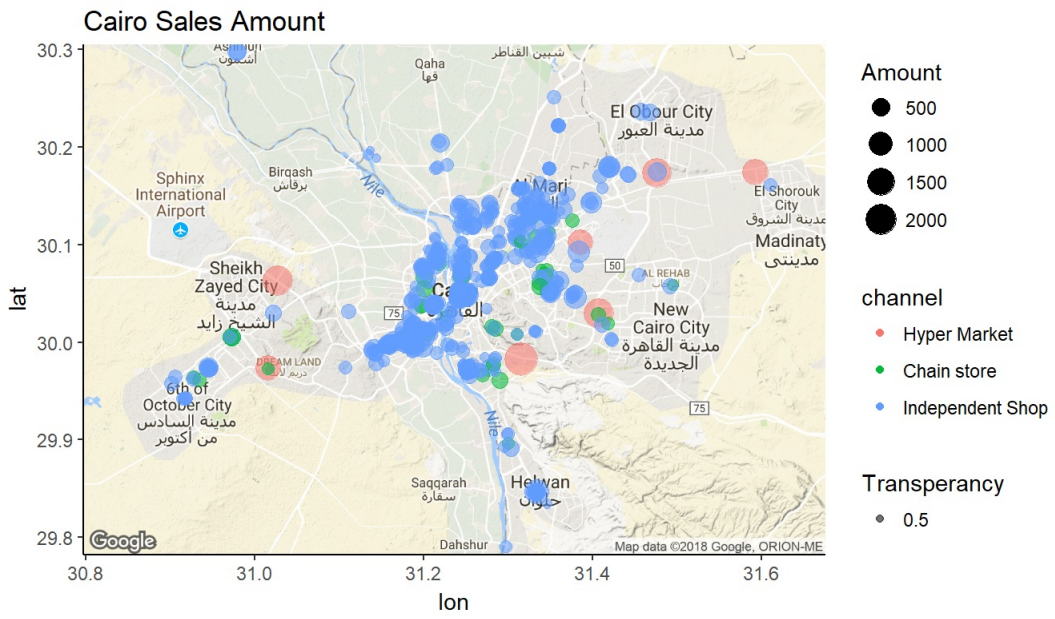
Ram variety for each brand


price
20000
15000
10000
5000
0
2
4
6
ram
brand_name
Samsung
Huawei
Oppo
Nokia
Lenovo
Apple
Sony
Htc
Motorola
Gionee
Alcatel
Asus

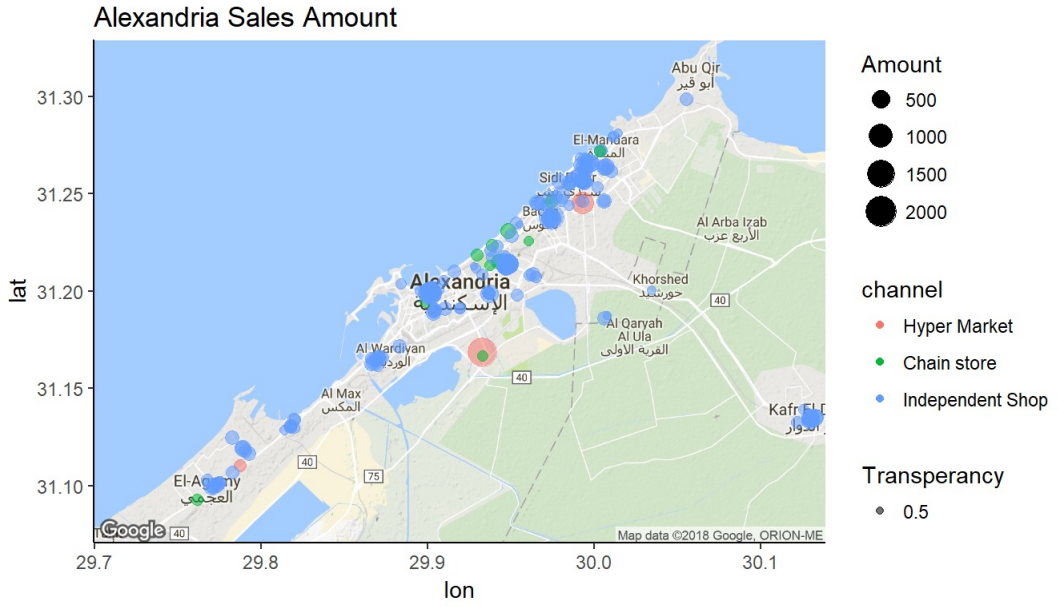

price
25000
20000
15000
10000
5000
0
2
4
6
ram
brand_name
Samsung
Huawei
Oppo
Nokia
Lenovo
Apple
Sony
Htc
Motorola
Gionee
Alcatel
Asus



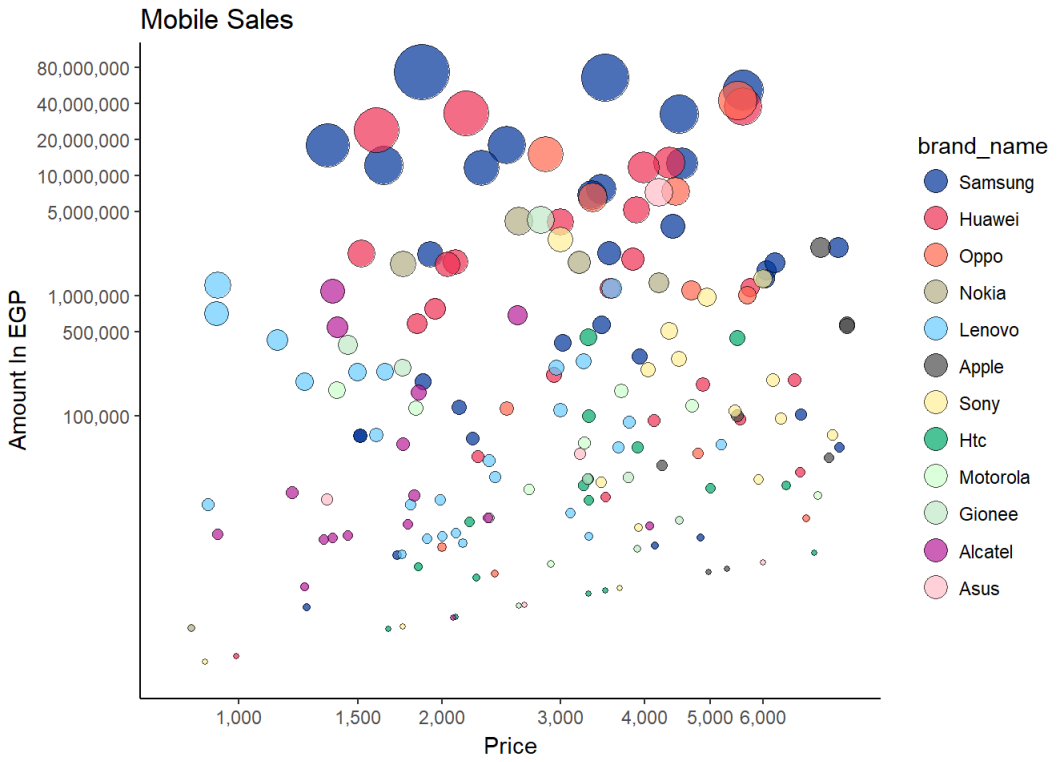

```
ggmap(cairo)+
  geom_point(data = shop_data_subtotal , aes(x =longitude ,y =latitude  ,
             colour = channel,
             alpha = .5,size = (sqrt(shop_data_subtotal$ttl_amount/pi))) )+
  labs(size="Amount" , alpha = 'Transperancy' , channel = 'Channel')+
  scale_size_continuous(range=c(1,7)) +
  ggtitle('Cairo Sales Amount')
```




```
ggmap(alex)+
  geom_point(data = shop_data_subtotal , aes(x =longitude ,y =latitude  ,
             colour = channel,
             alpha = .5,size = (sqrt(shop_data_subtotal$ttl_amount/pi))) )+
  labs(size="Amount" , alpha = 'Transperancy' , channel = 'Channel')+
  scale_size_continuous(range=c(1,7)) +
  ggtitle('Alexandria Sales Amount')
```



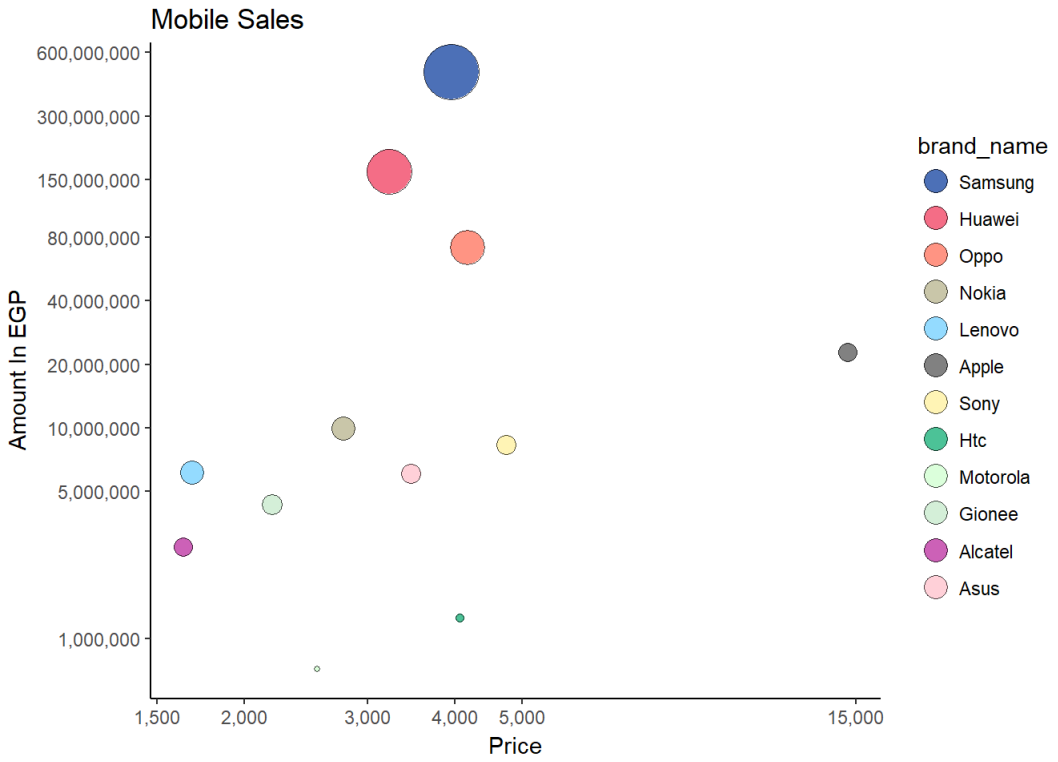

```
xbreaks = c(1000, 1500, 2000, 3000, 4000, 5000, 6000, 10000,15000,28000)
ybreaks = c(100000, 500000, 1000000, 5000000, 10000000, 20000000, 40000000,
            80000000)
ggplot(data = mobile_data,
       aes(x = price, y = ttl_amount)) +
  geom_point(aes(fill = brand_name, size = sqrt(ttl_quantity/pi)),
             pch = 21 , alpha = .7) +
  ggtitle("Mobile Sales") +
  xlab("Price") +
  ylab("Amount In EGP")+
  scale_x_log10(breaks = xbreaks,
                labels = comma(xbreaks) , limits = c(800,8000)) +
  scale_y_log10(breaks = ybreaks,
                labels = comma(ybreaks)) +

  scale_size_continuous(range=c(1,12)) +
  guides(size = F, fill = guide_legend(override.aes = list(size=5))) +
  scale_fill_manual(values= Palette)
```




```
brand_summary <- mob %>%
  group_by(brand_name)  %>%
  summarise(ttl_quantity = sum(sellout),
            avg_price = mean(price),
            n = n()) %>%
  arrange(brand_name)
brand_summary$ttl_amount <- (brand_summary$ttl_quantity *
                               brand_summary$avg_price)
str(brand_summary,give.attr = FALSE)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    12 obs. of  5 variables:
##  $ brand_name  : Ord.factor w/ 12 levels "Samsung"<"Huawei"<..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ ttl_quantity: int  122212 50938 17110 3578 3626 1547 1752 310 283 1967 ...
##  $ avg_price   : num  3965 3227 4178 2772 1686 ...
##  $ n           : int  13170 7316 2662 1114 1051 410 338 162 130 92 ...
##  $ ttl_amount  : num  4.85e+08 1.64e+08 7.15e+07 9.92e+06 6.11e+06 ...
```

```
xbreaks = c(1000, 1500, 2000, 3000, 4000, 5000,15000)
ybreaks = c(100000, 500000, 1000000, 5000000, 10000000, 20000000, 40000000,
            80000000,150000000,300000000,600000000)
ggplot(data = brand_summary,
       aes(x = avg_price, y = ttl_amount)) +
  geom_point(aes(fill = brand_name, size = sqrt(ttl_quantity/pi)),
             pch = 21 , alpha = .7) +
  ggtitle("Mobile Sales") +
  xlab("Price") +
  ylab("Amount In EGP")+
  scale_x_log10(breaks = xbreaks,
                labels = comma(xbreaks) ) +
  scale_y_log10(breaks = ybreaks,
                labels = comma(ybreaks)) +

  scale_size_continuous(range=c(1,12)) +
  guides(size = F, fill = guide_legend(override.aes = list(size=5))) +
  scale_fill_manual(values= Palette)
```



If we take the average price for each brand and measuere the sales amount for each brand we could see that samsung has the lead of sales amount and huawei and oppo try to follow up

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

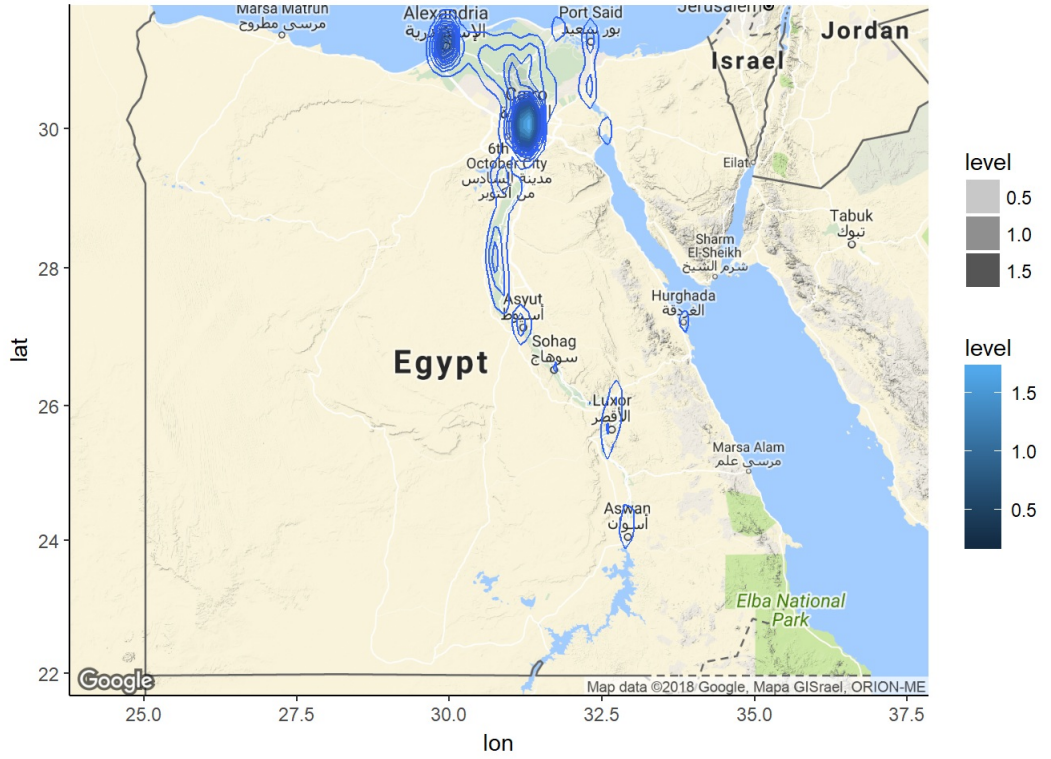
yes having a varaities of models for each price segment leads to better sales and win bigger market share

Were there any interesting or surprising interactions between features?

yes all the mobiles feature with the same ram and storage have almost same price except for apple that plays in a vary big price range comparing to others ### OPTIONAL: Did you create any models with your dataset? Discuss the
strengths and limitations of your model.

Final Plots and Summary

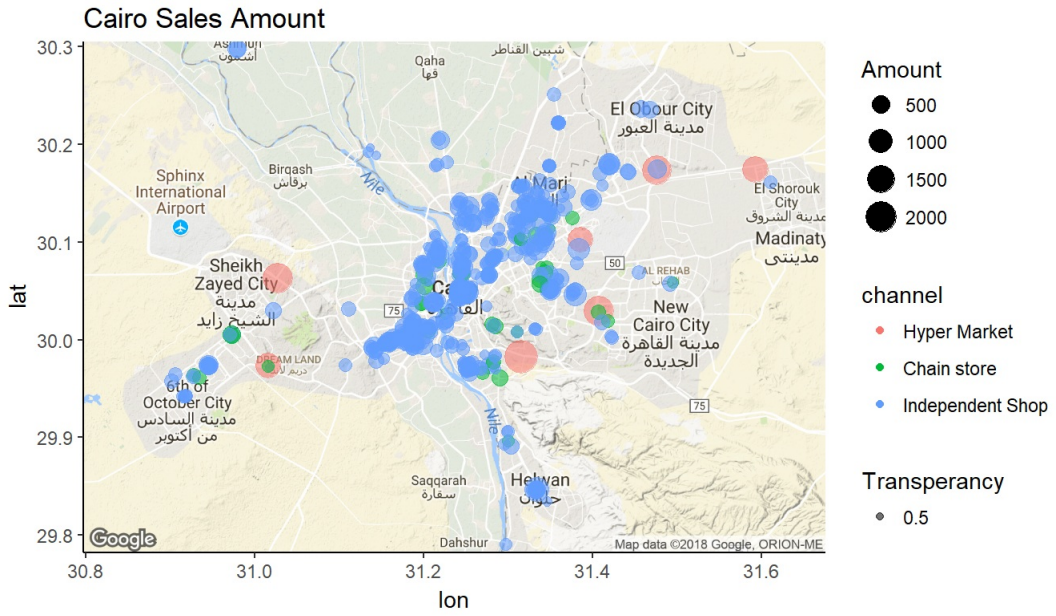
Plot One



Description One

The sales are centralized in the capital urban cities In Cairo and Alexandria while in rural area are small at the same time the income level on the urban and the culture for them deffer in the indevedual behaviour of consuming

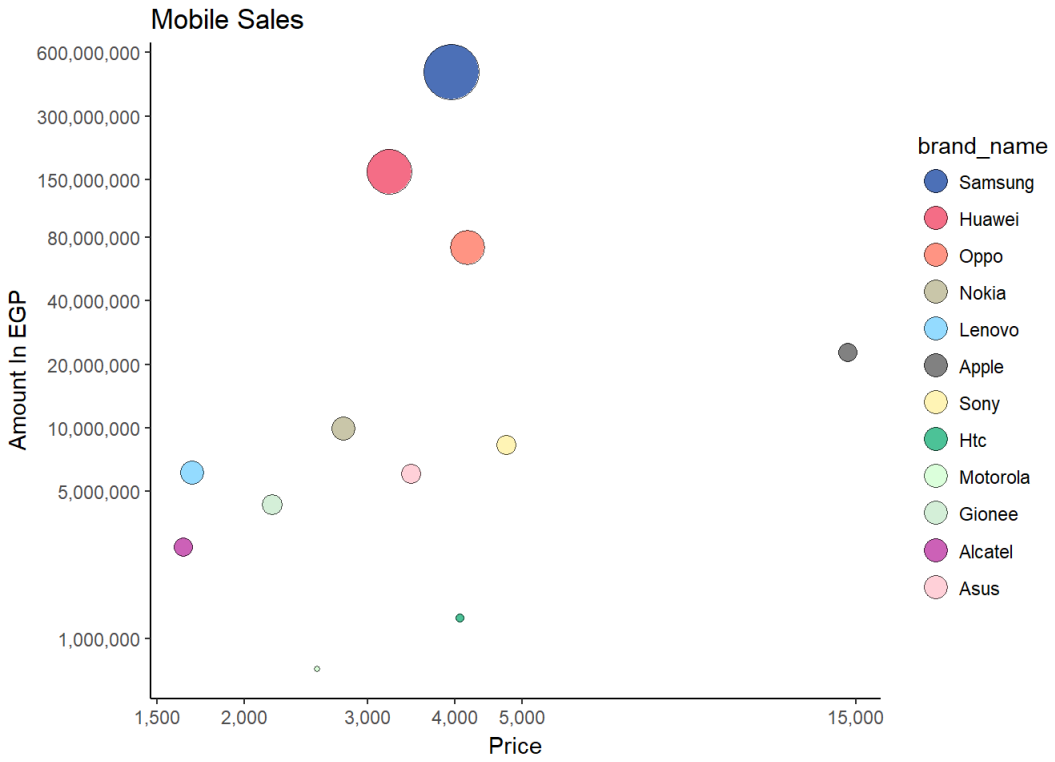
Plot Two



Description Two

Most of the sales quantity are in the Hyper big market as it attract the consumer with its discount specially in January the sales season (Black friday) and it has the variaities that satisfiy each consumer price segment

Plot Three



Description Three

compete in each price segment and with good advertising makes samsung lead the sales and have an average price of 4000 and good reputation

---

Reflection

It was a great time work on this project I used some wrangling and scraping skills to get the mobile specs and info from http://gsmarena.com and match each mobile specs with this data set, but I miss one big variable that could clear more the idea and help in this analysis and that's the processor type and speed as I think the processor and ram play a big part in the Mobile price which will effect at the end the sales portion for each shop