Dominicks data exploration

Dieudonne Ouedraogo

3/13/2018

Data Exploration

Part1 : SALES

Loading libraries and Shampoo dataset

```
library(dplyr)
library(knitr)
library(igraph)
library(pander)
library(ggplot2)
shadata <- read.csv("~/shadata.csv")
kable(head(shadata))
```

X	WEEK	Quarter	Year	UPC	perm1	size	unit	price	unit_price	sales	quantity_sold	unit_
1	128	1	1992	521328700	NA	6	OZ	1.690000	0.2816667	1.69	1	
2	128	1	1992	1150900201	NA	4	OZ	5.996897	1.4992241	173.91	29	
3	128	1	1992	1150900273	NA	1	CT	5.790000	5.7900000	86.85	15	
4	128	1	1992	1150900275	NA	1	CT	5.838889	5.8388889	262.75	45	
5	128	1	1992	1150900277	NA	1	CT	5.796000	5.7960000	289.80	50	
6	128	1	1992	1150900280	NA	1	CT	5.690000	5.6900000	39.83	7	

```
#pander(head(shadata))
#length(shadata)
#nrow(shadata)
```

Summary

```
total.sales <- sum(shadata$sales)
kable(paste('Total sales is $',round(total.sales)))
```

Total sales is $ 27228593

```
sales.UPC <- shadata %>% group_by(UPC) %>% summarise(value = sum(sales)) %>% filter(value==max(value))
d=as.character(sales.UPC)
names(d)=c("Best UPC","Total Sales")
kable(d)
```

Best UPC	3700000089
Total Sales	354104.03

```
prof.prod <- shadata %>% group_by(Year) %>% summarise(value = sum(sales)) %>% filter(value==max(value))
kable(caption="Best Year",prof.prod)
```

Table 4: Best Year

Year	value
1996	6613329

We can group the data by UPC ONLY

We can then summarize the data by average sales,quantities and by total sales

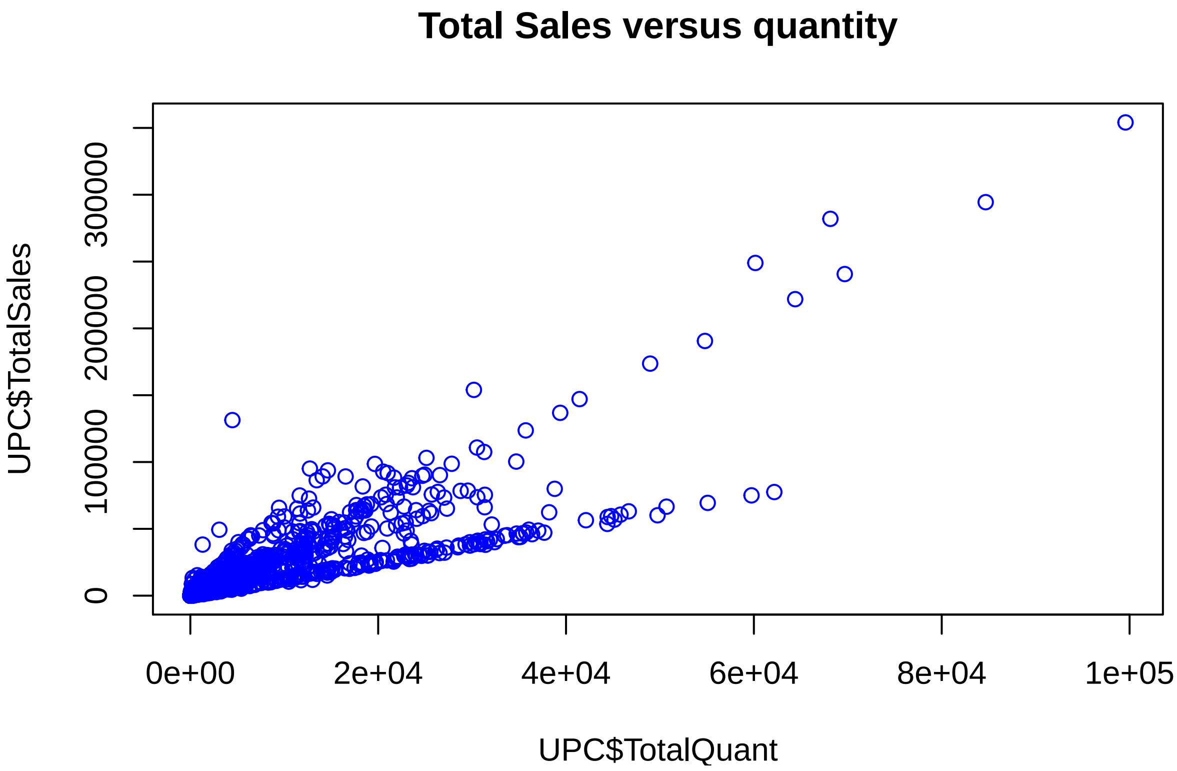
```
byUPC<-shadata%>%group_by(UPC)
summarize(shadata, AvgSales= mean(sales, na.rm = T),AvgQuantity= mean(quantity_sold, na.rm = T))
```

```
##   AvgSales AvgQuantity
## 1 125.4323     47.0339
```

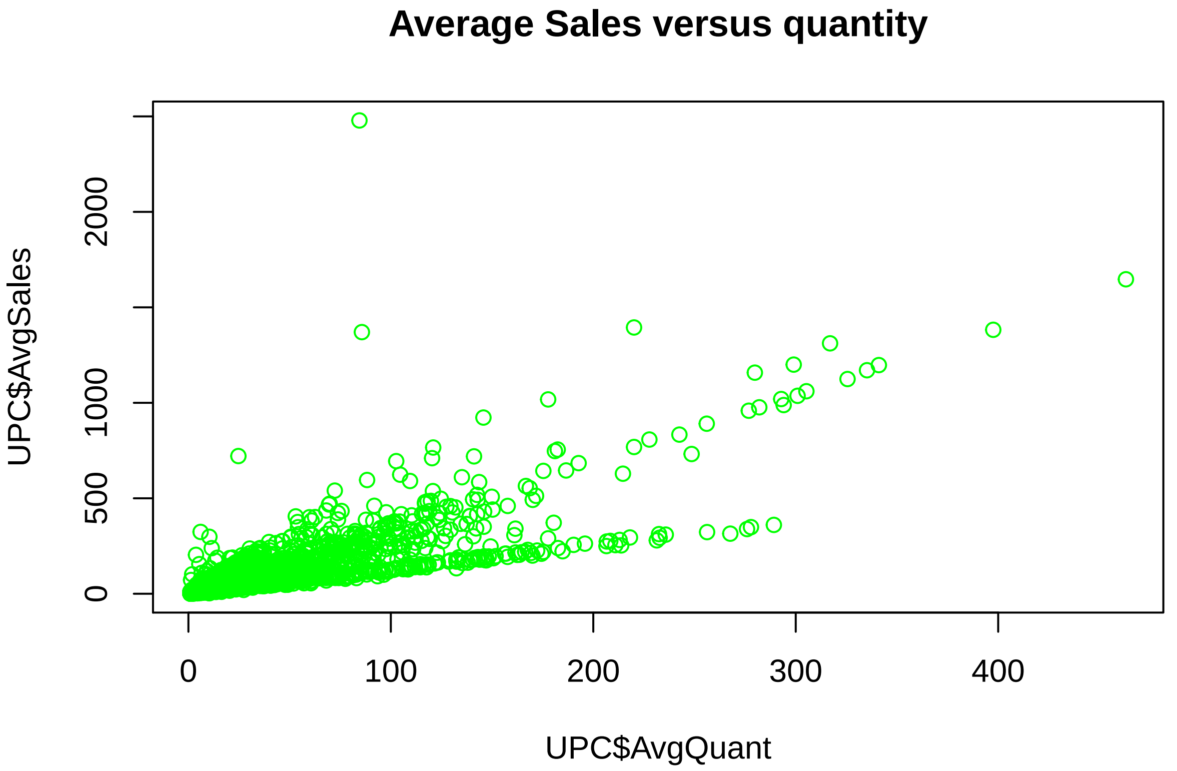
```
UPC<- summarize(byUPC, count = n(), AvgSales = mean(sales, na.rm = F),TotalSales=sum(sales, na.rm = F),
kable(head(UPC,10))
```

UPC	count	AvgSales	TotalSales	AvgQuant	TotalQuant
5690310	1	25.000000	25.00	10.000000	10
5690400	1	15.000000	15.00	6.000000	6
370071913	3	2.386667	7.16	1.333333	4
521328700	12	1.339167	16.07	1.250000	15
521346000	13	2.533077	32.93	2.153846	28
1150900201	204	177.442108	36198.19	26.593137	5425
1150900238	18	11.083889	199.51	1.611111	29
1150900261	117	34.288974	4011.81	5.401709	632
1150900262	119	61.956639	7372.84	9.714286	1156
1150900263	118	59.620424	7035.21	9.279661	1095

```
plot(UPC$TotalQuant,UPC$TotalSales,col="blue",main="Total Sales versus quantity")
```



```r
plot(UPC$AvgQuant,UPC$AvgSales,col="green",main="Average Sales versus quantity")
```



Here we can count the number of sales by UPC and per Year

```
SalesYear<- group_by(shadata, Year, UPC)
per_year <- summarize(SalesYear, number_sales = n(),totalSales=sum(sales))
kable(head(per_year,10))
```

Year	UPC	number_sales	totalSales
1992	370071913	3	7.16
1992	521328700	8	11.81
1992	521346000	9	28.17
1992	1150900201	46	7144.02
1992	1150900273	46	5653.20
1992	1150900275	46	8043.27
1992	1150900277	46	10623.55
1992	1150900280	46	1907.48
1992	1150900291	35	558.14
1992	1150900299	46	4024.29

Visualization

Sales per year

```
data2<- shadata %>% group_by(Year) %>% summarise(value = sum(sales))
kable(data2)
```

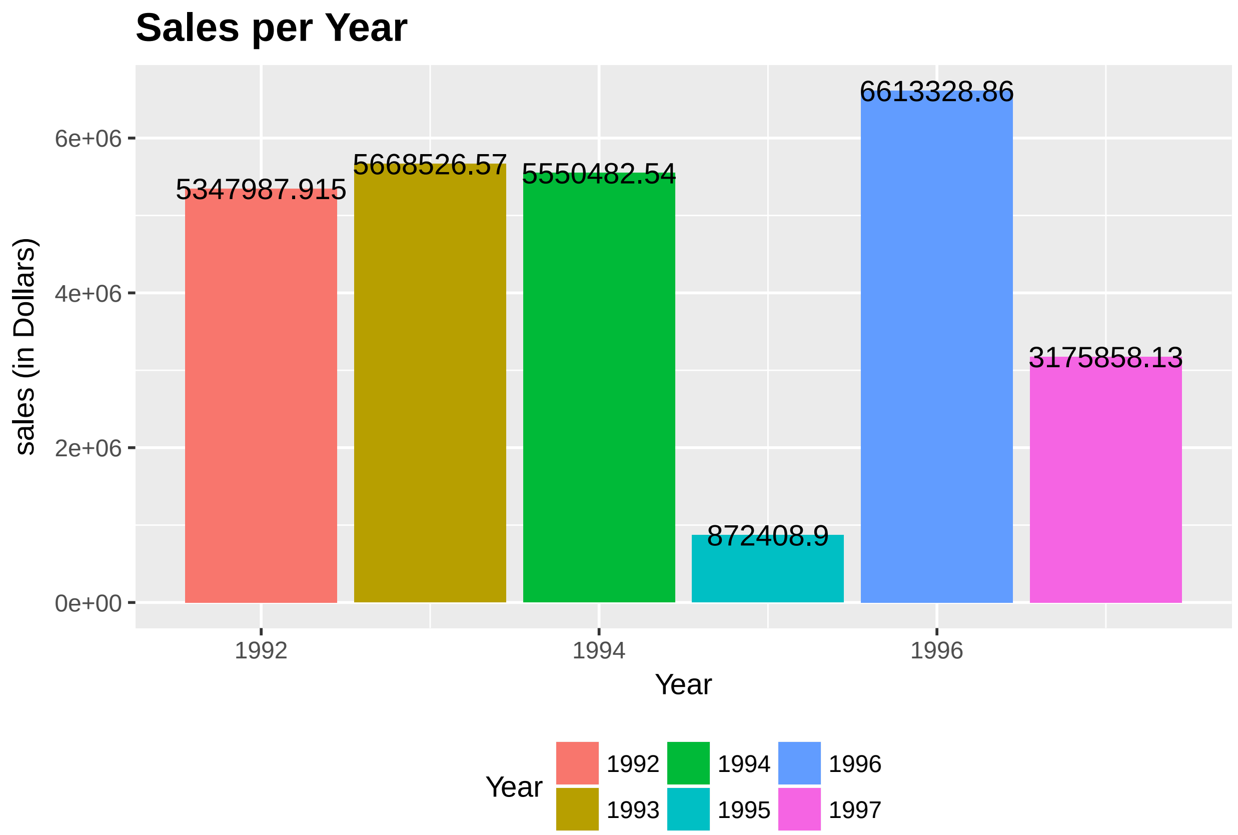
Year	value
1992	5347987.9
1993	5668526.6
1994	5550482.5
1995	872408.9
1996	6613328.9
1997	3175858.1

```
d=arrange(data2,by_group=desc(value))# By ordering
kable(d)
```

Year	value
1996	6613328.9
1993	5668526.6
1994	5550482.5
1992	5347987.9
1997	3175858.1
1995	872408.9

```
g=ggplot(data = data2, aes(x=Year, y=value, fill=factor(Year))) +
  geom_bar(position = "dodge", stat = "identity") + ylab("sales (in Dollars)") +
  xlab("Year") + theme(legend.position="bottom" ,plot.title = element_text(size=15, face="bold")) +
  ggtitle("Sales per Year") + labs(fill = "Year")+geom_text(aes(label=value))
```

```
g
```



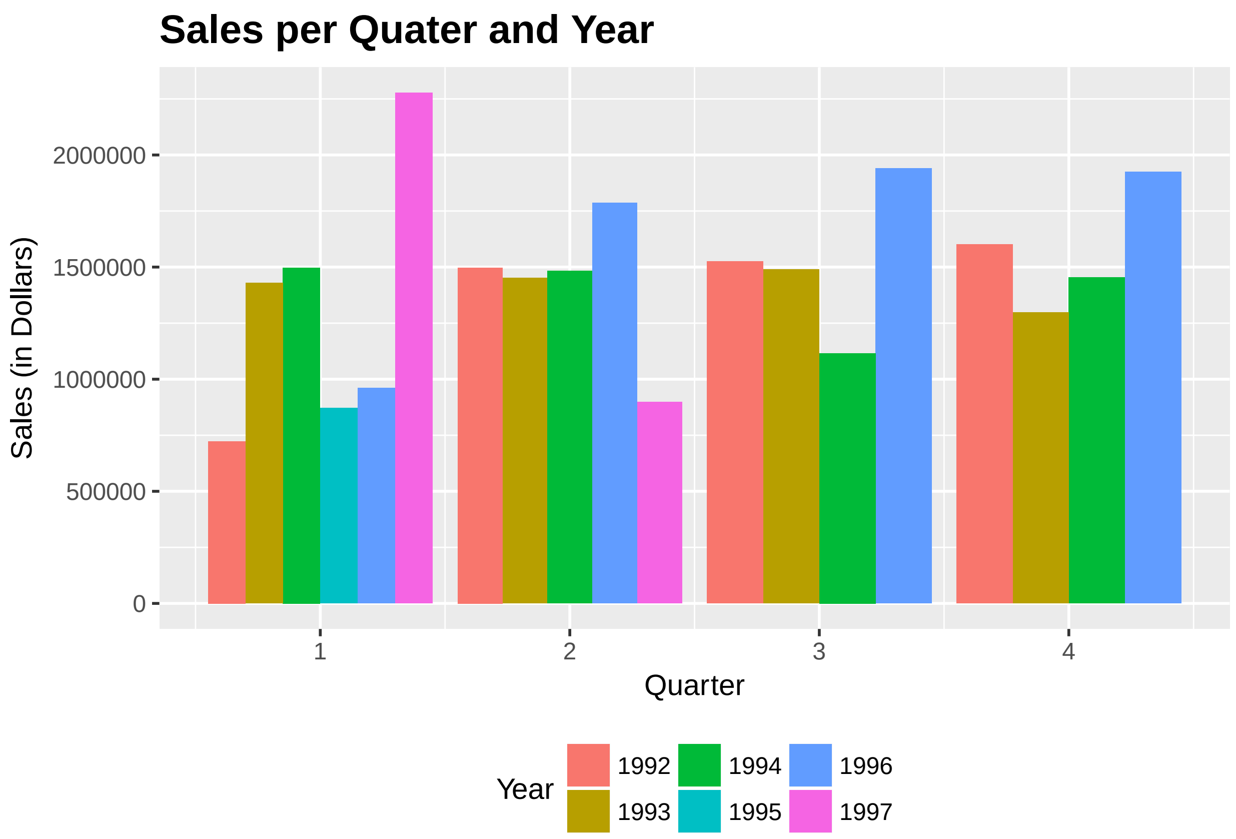
Sales per Quarter

One way to look at it

```
data3<- shadata %>% group_by(Quarter,Year) %>% summarise(value = sum(sales))
kable(head(data3))# We look at the first rows of the dataset.
```

Quarter	Year	value
1	1992	723135.4
1	1993	1430168.4
1	1994	1497399.3
1	1995	872408.9
1	1996	960527.1
1	1997	2277609.4

```
g=ggplot(data = data3, aes(x=Quarter, y=value, fill=factor(Year))) +
  geom_bar(position = "dodge", stat = "identity") + ylab("Sales (in Dollars)") +
  xlab("Quarter") + theme(legend.position="bottom" ,plot.title = element_text(size=15, face="bold")) +
  ggtitle("Sales per Quater and Year") + labs(fill = "Year")
g
```

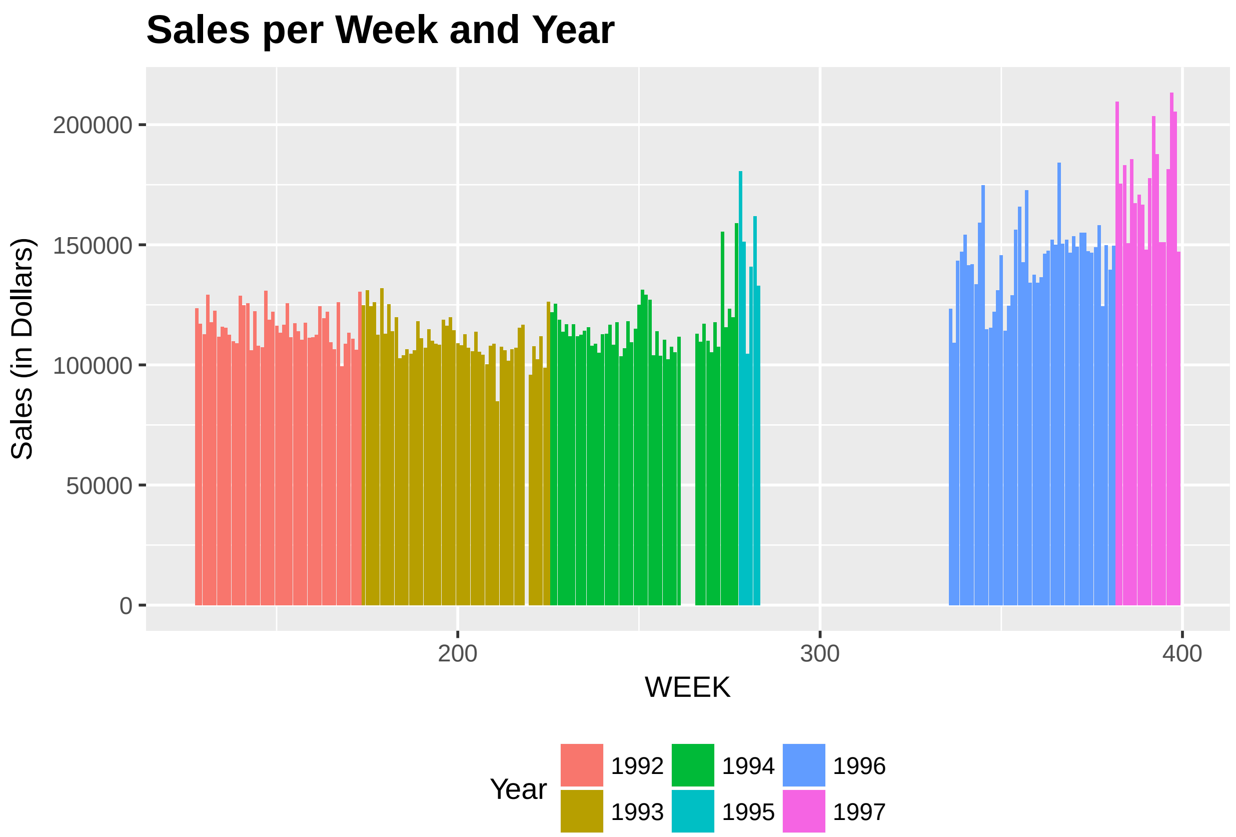


Sales By WEEK

```r
data4<- shadata %>% group_by(WEEK,Year) %>% summarise(value = sum(sales))
kable(head(data4,10))
```

WEEK	Year	value
128	1992	123687.6
129	1992	117067.1
130	1992	112779.8
131	1992	129243.5
132	1992	117808.5
133	1992	122548.8
134	1992	111820.8
135	1992	115866.3
136	1992	115577.9
137	1992	112473.1

```r
g=ggplot(data = data4, aes(x=WEEK, y=value, fill=factor(Year))) +
  geom_bar(position = "dodge", stat = "identity") + ylab("Sales (in Dollars)") +
  xlab("WEEK") + theme(legend.position="bottom" ,plot.title = element_text(size=15, face="bold")) +
  ggtitle("Sales per Week and Year") + labs(fill = "Year")
g
```



Saving datasets into csv files which can be used in Gephi

```
#write.csv(data4, file = "Weekly.csv")
#write.csv(UPC, file = "UPC.csv")
```

Network Science with igraph

We can transform some of the newly created datasets into network entities for further insights For example the total weekly sales dataset name data4 can be converted

```
net=graph_from_data_frame(data4,directed=FALSE)
kable(paste("Mean Distance is :",round(mean_distance(net, directed=F),3)))
```

Mean Distance is : 1.956

```
kable(paste("The graph density is:",round(graph.density(net,loop=FALSE),6)))
```

The graph density is: 0.008844

```
kable(paste("the shortest path is: ",max(shortest.paths(net,mode="all"))))
```

the shortest path is: Inf

```r
#
kable(paste("The maximum Eccentricity",max(eccentricity(net,mode="all"))))
```

The maximum Eccentricity 2

```r
deg <- degree(net, mode="all")
kable(paste("The diameter of this network is",diameter(net)))
```

The diameter of this network is 2

```r
kable(paste("The maximum degree is:",max(deg)))
```

The maximum degree is: 51

```r
kable(paste("The minimum degree is:",min(deg)))
```

The minimum degree is: 1

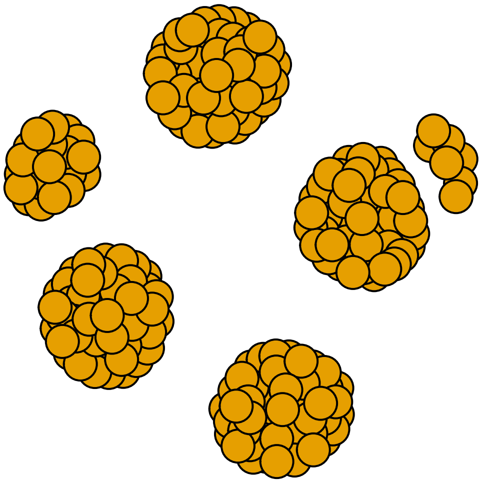
```r
kable(paste("The Average degree is:",round(mean(deg),4)))
```

The Average degree is: 1.9457

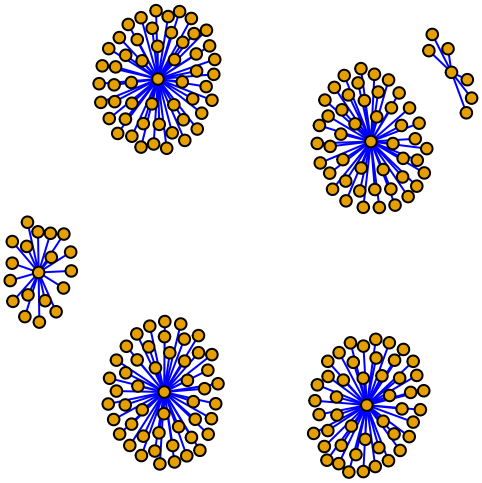
```r
Degree_Correlation=assortativity_degree(net,directed = F)
kable(paste("The degree correlation is:",round(Degree_Correlation,4)))# Degree correlation
```

The degree correlation is: -0.8912

```r
plot.igraph(net,vertex.label=NA)
```



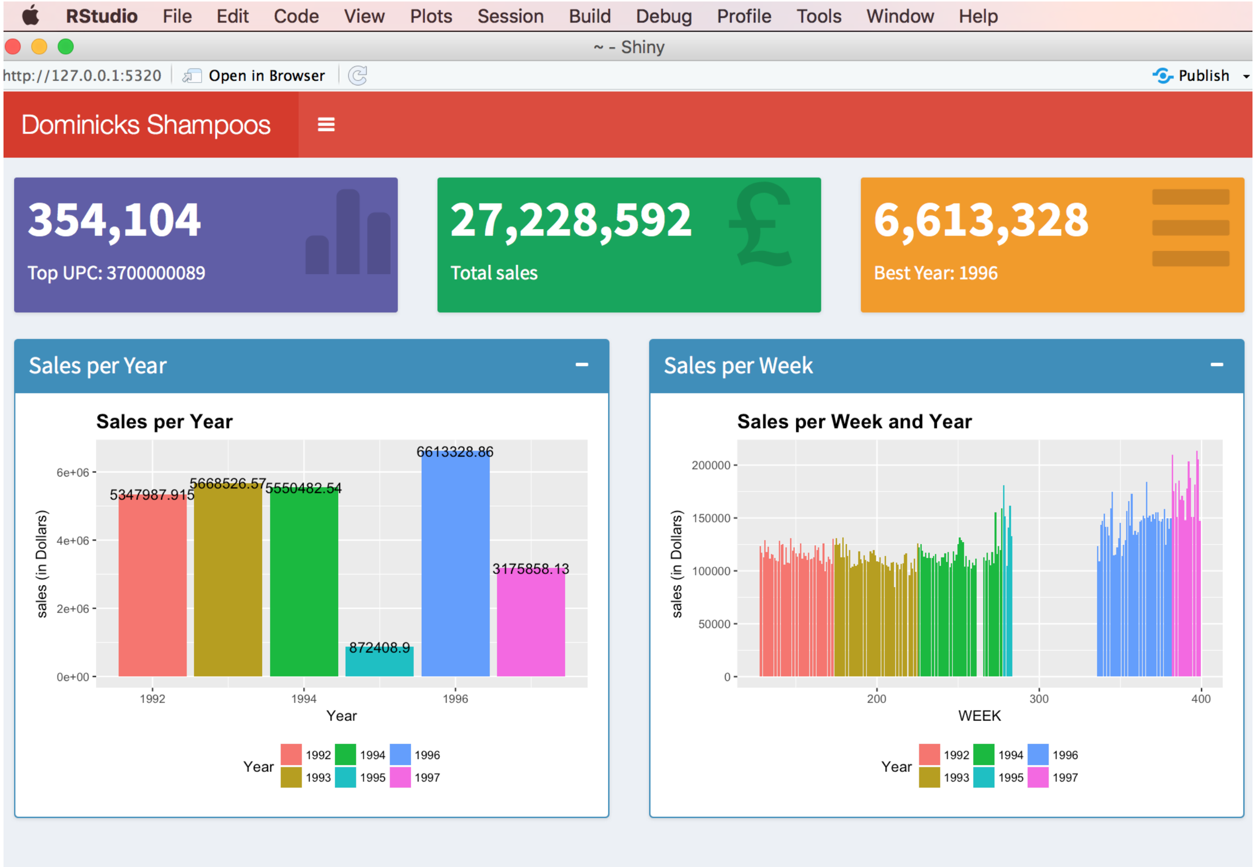
```
##Let's reduce the size of the node to have a better look
plot.igraph(net,vertex.size=5,vertex.label=NA,edge.color="blue")
```



Using Shiny apps

Below is an app file

```
# app.R
# load the required packages
library(shiny)
require(shinydashboard)
library(ggplot2)
library(dplyr)
shadata <- read.csv("~/shadata.csv")
data2<- shadata %>% group_by(Year) %>% summarise(value = sum(sales))
data4<- shadata %>% group_by(WEEK,Year) %>% summarise(value = sum(sales))
header <- dashboardHeader(title = "Dominicks Shampoos data")
sidebar <- dashboardSidebar(
  sidebarMenu(
    menuItem("Dashboard", tabName = "dashboard", icon = icon("dashboard"))

  )
)


frow1 <- fluidRow(
  valueBoxOutput("value1")
  ,valueBoxOutput("value2")
  ,valueBoxOutput("value3")
)

frow2 <- fluidRow(

  box(
    title = "Sales per Year"
    ,status = "primary"
    ,solidHeader = TRUE
    ,collapsible = TRUE
    ,plotOutput("salesbyYear", height = "300px")
  )

  ,box(
    title = "Sales per Week"
    ,status = "primary"
    ,solidHeader = TRUE
    ,collapsible = TRUE
    ,plotOutput("salesbyWeek", height = "300px")
  )

)
body <- dashboardBody(frow1, frow2)
ui <- dashboardPage(title = 'Dominicks Shampoos category', header, sidebar, body, skin='red')

server <- function(input, output) {
  total.sales <- sum(shadata$sales)
  sales.UPC <- shadata %>% group_by(UPC) %>% summarise(value = sum(sales)) %>% filter(value==max(value)
  prof.prod <- shadata %>% group_by(Year) %>% summarise(value = sum(sales)) %>% filter(value==max(value
  output$value1 <- renderValueBox({
    valueBox(
      formatC(sales.UPC$value, format="d", big.mark=',')
      ,paste('Top UPC:',sales.UPC$UPC)
      ,icon = icon("stats",lib='glyphicon')
      ,color = "purple")
  })
  output$value2 <- renderValueBox({
    valueBox(
      formatC(total.sales, format="d", big.mark=',')
      ,'Total sales'
      ,icon = icon("gbp",lib='glyphicon')
      ,color = "green")
  })
  output$value3 <- renderValueBox({
    valueBox(
      formatC(prof.prod$value, format="d", big.mark=',')
      ,paste('Best Year:',prof.prod$Year)
      ,icon = icon("menu-hamburger",lib='glyphicon')
      ,color = "yellow")
  })
  output$salesbyYear<- renderPlot({

    ggplot(data = data2, aes(x=Year, y=value, fill=factor(Year))) +
      geom_bar(position = "dodge", stat = "identity") + ylab("sales (in Dollars)") +
      xlab("Year") + theme(legend.position="bottom" ,plot.title = element_text(size=15, face="bold")) +
      ggtitle("Sales per Year") + labs(fill = "Year")+geom_text(aes(label=value))


  })
  output$salesbyWeek <- renderPlot({
    ggplot(data = data4, aes(x=WEEK, y=value, fill=factor(Year))) +
      geom_bar(position = "dodge", stat = "identity") + ylab("sales (in Dollars)") +
```



354,104

Top UPC: 3700000089

27,228,592

Total sales

6,613,328

Best Year: 1996

Figure 1: Shampoo data on shinyapp

```
      xlab("WEEK") + theme(legend.position="bottom" ,plot.title = element_text(size=15, face="bold")) +
      ggtitle("Sales per Week and Year") + labs(fill = "Year")
  })
}

shinyApp(ui, server)
```

Shiny applications not supported in static R Markdown documents