Diogo de Lima R. C. da Silva
Simplify the model.

Some of the variable are not very useful because they depend of each other, for instance the icms's variables or any other “tax” variable (there is a colinearity between revenues and taxes) .

Others variables have not any distinct information so they not will used, these variableas are:

- cnpj
- fone
- Bairro
- xLgr
- xMun
- Pais
- uf
-xFant
-natOp
-InfCpl (may this could be useful, but I will ignore rigth away based on the fact that in this type of restaurant (kilo), usually the customers gets their food before choose their table).

If we wish to forecast the revenue of the restaurant, the simplest approach is to use the consumption of the products. So we will focus our analysis in the column **dets** which contains the client purchases on a given moment.

We will retreat this data, to do that we will create a data frame a variable called **my_data** that contains:

- a quantity column for each products
- the day of the purchases
- the ticket price

to do that we can use the following R code:

```
#read json file
library("jsonlite")
json_file <- "~/Documents/TOTVS/sample.json"
json_data <- fromJSON(json_file)

# lets find all the products available on the restaurant

products = vector()
for (i in 1:nrow(json_data)){
  a<-do.call("rbind",json_data$dets[i])
  #print( a$prod$xProd)
  products = c(products,a$prod$xProd)
}
```

```
products<- unique(products)
prices <- paste(products,"Prices",sep="_")
quantity <- paste(products,"quantity",sep="_")

vector_names<-c(quantity,"dia","mês","ano","nweek","valor_total")
size_row<-nrow(json_data)
size_name<-length(vector_names)

# lets create the new data frame my_data

my_data<- data.frame(matrix(0,size_row,size_name))
names(my_data)<-vector_names

for (i in 1:nrow(json_data)){
  a<-do.call("rbind",json_data$dets[i])

  for( j in 1: nrow(a))
  {
    #my_data[i,match(pastes(a$prod$xProd[j],"Prices"),vector_names)]=a$prod$vUnCom[j]
    my_data[i,match(paste(a$prod$xProd[j],"quantity",sep="_"),vector_names)]=a$prod$qCom[j]
  }
  my_data[i,match("dia",vector_names)]=as.numeric( substr(json_data$ide$dhEmi$`$date`[i],9,10) )
  my_data[i,match("mês",vector_names)]=as.numeric( substr(json_data$ide$dhEmi$`$date`[i],6,7) )
  my_data[i,match("ano",vector_names)]=as.numeric( substr(json_data$ide$dhEmi$`$date`[i],1,4) )
  my_data[i,match("nweek",vector_names)]=my_data[i,match("dia",vector_names)]%/%7+1
  my_data[i,match("valor_total",vector_names)]=json_data$complemento$valorTotal[i]
}
```

We can see from the data that we have restaurant activity on the january 2016.

This information is interesting but we want to understand the dynamic of the restaurant. To do that we are going to regroup the purchases according to its day of the month, we can do this using the following command

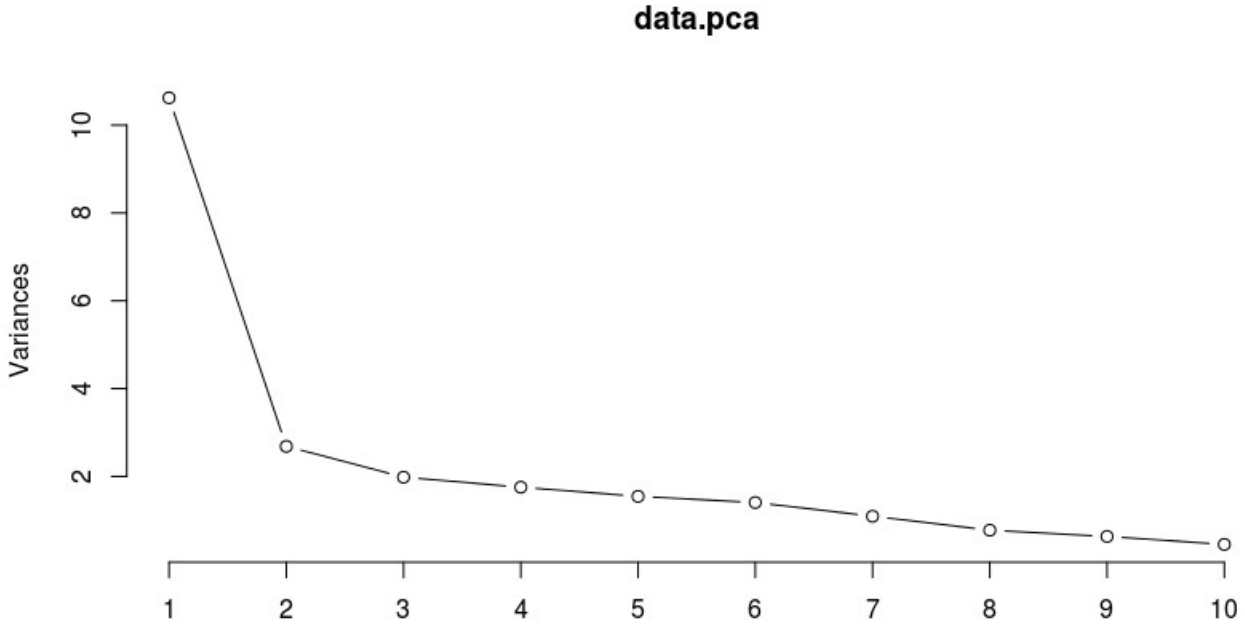
---

```
my_aggregated_data <-aggregate(.~my_data$dia,my_data,sum)
```

The dataframe **my_aggregated_data** will be the data that we will use from now on. It contains at given date, the products consumed at the restaurant.

We can try to understand a little more of the restaurant dynamics, our first attempt consist in doing a principal components analysis (PCA). As a results we have the following graphic of the variance of each component

```
# R CODE for PCA Analysis
data.pca <-prcomp(my_aggregated_data[,1:24],center = TRUE,scale=TRUE)
print(data.pca)
plot(c(1:24),data.pca$rotation[,1])
plot(data.pca,type='l')
sort(abs(data.pca$rotation[,1]))
```



We can that the first component explains most of the restaurant's variance. Now lets take a look on the value of each element of the first component.

```
HARUMAKI_quantity       BULE CHA_quantity          VINHO_quantity      SOBREMESA_quantity
               0.01034855              0.02281770              0.06380552              0.06941852
             my_data$dia        SASHIMI_quantity            CHA_quantity     CAIPIRINHA_quantity
               0.11098659              0.11610968              0.13924484              0.15593906
       DOCINHOS_quantity         WHISKY_quantity      YAKISSOBA_quantity SUSHI ESPECIAL_quantity
               0.18495468              0.19191068              0.21609192              0.22166497
        BACARDI_quantity   CERVEJA LATA_quantity       CERVEJA_quantity         TEMAKI_quantity
               0.22673237              0.22746339              0.23286180              0.23346382
       LIMONADA_quantity  CAFE EXPRESSO_quantity          SAKE_quantity           AGUA_quantity
               0.24518782              0.24968260              0.25544249              0.26114962
           SUCO_quantity     CAIPIROSKA_quantity        BUFFET_quantity   REFRIGERANTE_quantity
               0.26162390              0.26350046              0.26618732              0.27854394
```

It is not very clear what we can get from here, we could regroup these values into 3 groups:

- Group1 [0 , 0.07]
- Group2 [0.11 , 0.20]

- Group3 [0.21 , 0.28]
So from this pCA analysis, I would be tempted to do an aggressive simplification and use only 3 variables in the forecast model.

Lets keep that in mind. Now lets look at the correlation matrix of **my_aggregated_data**:

#compute the correlation of my_aggregated_data

correlations_data <-cor(my_aggregated_data)

correlations_data[30,3]

There is a lot of information in correlations_data and I will not paste in this.

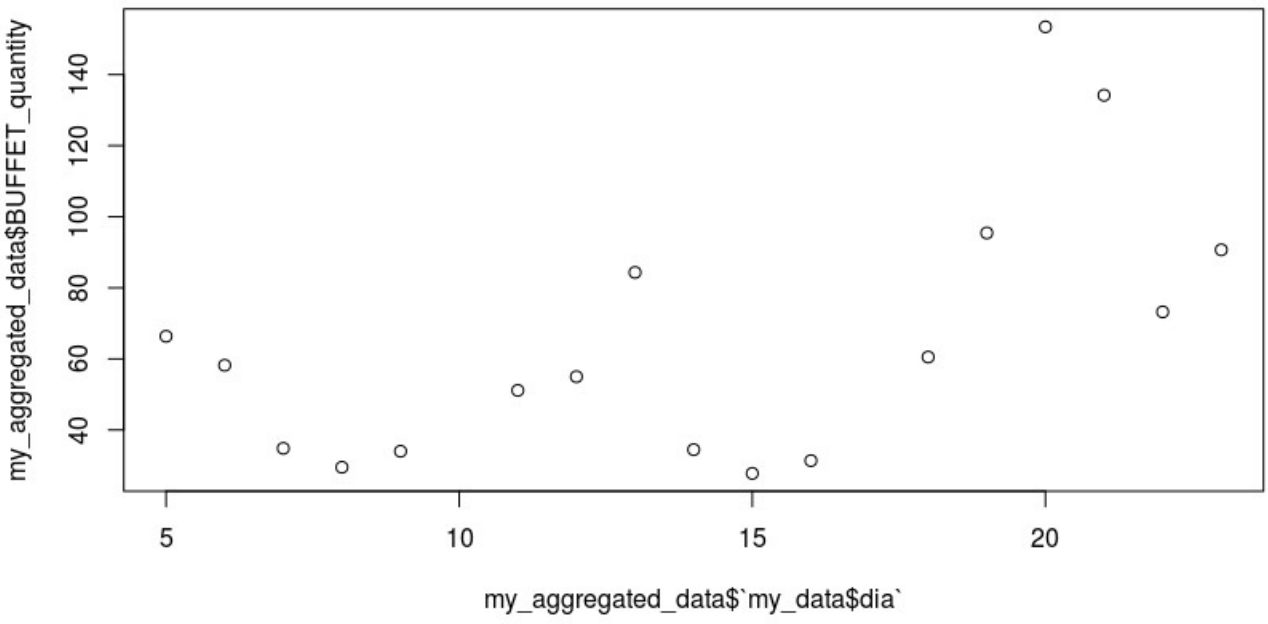
But lets think about the restaurant for a moment, this is a “kilo” restaurant so problably its main source of revenue it is its buffet, the quantity of the buffet consumed in a day would have a lot of influence on the restaurant revenue.We can see that in the correlation matrix.
So we will make another simplification and use the following model to forecast the revenue (valor_total)

valor_total = constant_1 + constant_2 x BUFFET_quantity

We have our forescast model but now we need to forecast make another model for BUFFET_quantity. Let's see how BUFFET_quantity behaves over the time:

---
plot(my_aggregated_data$`my_data$dia`,my_aggregated_data$BUFFET_quantity)
---

by doing this we have the following graphic:



By looking at this graphic, it seems that there is a cyclic behavior during this time, the BUFFET_quantity increases and decreases over the time.
So lets try to fit a polynomial here::

```
fit<- lm(my_aggregated_data$BUFFET_quantity ~ my_aggregated_data$dia,data=  my_aggregated_data)
fit2<- lm(my_aggregated_data$BUFFET_quantity ~
my_aggregated_data$dia+I(my_aggregated_data$`my_data$dia`^2)+I(my_aggregated_data$`my_data$dia`^3),data=
my_aggregated_data)
fit3<- lm(my_aggregated_data$BUFFET_quantity ~
my_aggregated_data$dia+I(my_aggregated_data$`my_data$dia`^2)+I(my_aggregated_data$`my_data$dia`^3)+I(my_aggre
gated_data$`my_data$dia`^4)+I(my_aggregated_data$`my_data$dia`^5),data=  my_aggregated_data)
fit4<- lm(my_aggregated_data$BUFFET_quantity ~
my_aggregated_data$dia+I(my_aggregated_data$`my_data$dia`^2)+I(my_aggregated_data$`my_data$dia`^3)+I(my_aggre
gated_data$`my_data$dia`^4)+I(my_aggregated_data$`my_data$dia`^5)+I(my_aggregated_data$`my_data$dia`^6)+I(my_a
ggregated_data$`my_data$dia`^7),data=  my_aggregated_data)
fit5<- lm(my_aggregated_data$BUFFET_quantity ~
my_aggregated_data$dia+I(my_aggregated_data$`my_data$dia`^2)+I(my_aggregated_data$`my_data$dia`^3)+I(my_aggre
gated_data$`my_data$dia`^4)+I(my_aggregated_data$`my_data$dia`^5)+I(my_aggregated_data$`my_data$dia`^6)+I(my_a
ggregated_data$`my_data$dia`^7)+I(my_aggregated_data$`my_data$dia`^8)+I(my_aggregated_data$`my_data$dia`^9),dat
a=  my_aggregated_data)
fit6<- lm(my_aggregated_data$BUFFET_quantity ~ my_aggregated_data$dia +
my_aggregated_data$SASHIMI_quantity+my_aggregated_data$`BULE CHA_quantity`,data=  my_aggregated_data)
fit7<- lm(my_aggregated_data$valor_total ~ my_aggregated_data$BUFFET_quantity,data=  my_aggregated_data)

summary(fit)
summary(fit2)
summary(fit3)
summary(fit4)
summary(fit5)
summary(fit6)
summary(fit7)
```

By looking at these models the 5 degree polynomial has a good p-value for all its coeficients, The linear modell also has a significant coefficients in terms of p-value. However the first one has a better residual standard error and R-squared.

Therefore, I would recommend the use of the following model to explain BUFFET quantity :

```
Call:
lm(formula = my_aggregated_data$BUFFET_quantity ~ my_aggregated_data$dia +
     I(my_aggregated_data$`my_data$dia`^2) + I(my_aggregated_data$`my_data$dia`^3) +
     I(my_aggregated_data$`my_data$dia`^4) + I(my_aggregated_data$`my_data$dia`^5),
     data = my_aggregated_data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.078   -2.481   -0.140    1.415   12.921

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             122.028123  21.963505   5.556 0.000171 ***
my_aggregated_data$dia                    0.063592   0.006256  10.165 6.27e-07 ***
I(my_aggregated_data$`my_data$dia`^2)    -6.757847   1.611945  -4.192 0.001505 **
I(my_aggregated_data$`my_data$dia`^3)     0.980790   0.240074   4.085 0.001804 **
I(my_aggregated_data$`my_data$dia`^4)    -0.052762   0.012744  -4.140 0.001644 **
```

```
I(my_aggregated_data$`my_data$dia`^5)   0.000971   0.000230   4.222 0.001431 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.939 on 11 degrees of freedom
Multiple R-squared:  0.9157,    Adjusted R-squared:  0.8773
F-statistic: 23.88 on 5 and 11 DF,  p-value: 1.462e-05
```

Here i

Now we can create a model for valor_total and this model is:

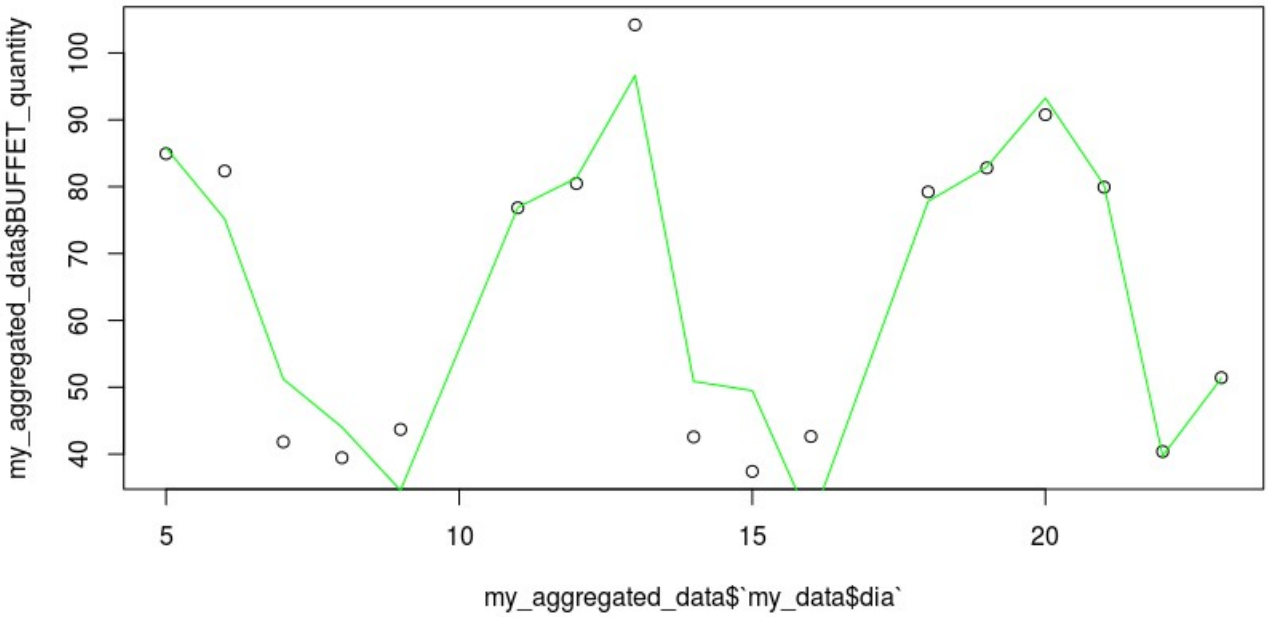
```
Call:
lm(formula = my_aggregated_data$valor_total ~ my_aggregated_data$BUFFET_quantity,
     data = my_aggregated_data)

Residuals:
     Min      1Q  Median      3Q     Max
-648.73 -387.61    9.66  130.49  858.71

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                          -826.778    341.905  -2.418   0.0288 *
my_aggregated_data$BUFFET_quantity   95.189      4.999  19.042 6.42e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 453.2 on 15 degrees of freedom
Multiple R-squared:  0.9603,    Adjusted R-squared:  0.9576
F-statistic: 362.6 on 1 and 15 DF,  p-value: 6.417e-12
```

So we to make the forecast we can use these two models. First we estimate BUFFET_quantity and the we would proceed to compute valor_total



The advantage of this model is that it tries to capture the cyclical behavior that custmoers must have in going to restaurant, ie, they do not got to the same restaurant everyday. And it can help to predict this change on volume and as consequences in

its revenues.