Midterm 2

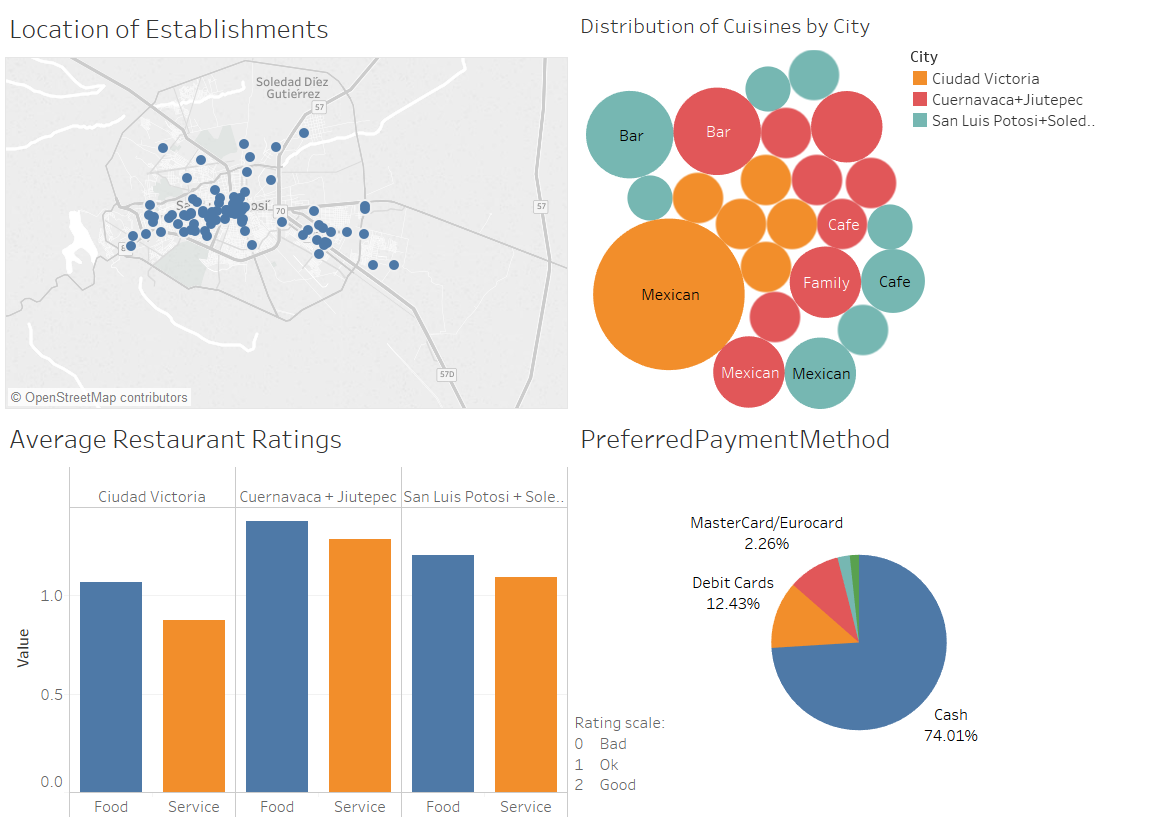
MG-GY 9753 Business Analytics

Priyanandana Varma Mudunuri, pvm255@nyu.edu, N16372290)

Sharang Biradar, sdb418@nyu.edu, N11407631)

Thorsteinn Nielsen, tjn259@nyu.edu, N10717700)

Problem 1



On the top left we have a map showing the locations of the restaurants. This image is showing restaurants in San Luis Potosi and Soledad.

On the top right, there is a bubble chart showing the highest represented cuisines in each city. The size of the bubbles indicates the ratio of certain types of restaurants to the total number of restaurants. The colors represent the cities. Only cuisines with 5% or more are shown. This chart tells us that the proportion of Mexican restaurants in Ciudad Victoria is fairly high compared to the other two cities. For the sake of not putting all our eggs in the same basket and appealing to a broader clientele, MNC might consider diversifying its business in Ciudad Victoria.

The pie chart on the bottom right shows that most of the customers prefer paying with cash so MNC has to make sure that they always have enough change available in the registers. MNC also has to make sure that the restaurants have POS systems so their customers can pay with cards. Rougly 1 in 4 customer prefers paying with a card.

The side-by-side bar chart in the lower left corner shows the average food and service rating of restaurants by city. MNC is obviously not happy with the ratings in Ciudad Victoria. Ratings in Cuernavaca+Jiutepec and San Luis Postoli + Soledad are substantially better and MNC should look further into that.

Problem 2

Reading the data and creating a linear regression model with Total Spend as a dependent variable and Gender, Age and Source as independent variables.

problem2 <- read.csv("C:\\Users\\niels\\OneDrive\\NYU\\Business Analytics\\Exam2\\Problem2\\CasinoA.csv", header = TRUE)  
casinomodel <- lm(Total.Spend ~ Gender + Age + Source, data = problem2)  
summary(casinomodel)

##   
## Call:  
## lm(formula = Total.Spend ~ Gender + Age + Source, data = problem2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1845.1 -995.1 -605.6 17.9 14216.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 144.258 137.492 1.049 0.294   
## GenderMale -42.839 62.153 -0.689 0.491   
## Age 14.884 2.511 5.928 3.28e-09 \*\*\*  
## SourceWALK -52.297 90.060 -0.581 0.561   
## SourceWEB 595.856 77.724 7.666 2.11e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2197 on 4995 degrees of freedom  
## Multiple R-squared: 0.02685, Adjusted R-squared: 0.02607   
## F-statistic: 34.45 on 4 and 4995 DF, p-value: < 2.2e-16

The significant variables are Age and Source, more specifically Walk-ins. The R-squared adjusted is 0.026 which means that our model can only explain 2.6% of the variance in the data which is not ideal. We would want the R-squared adjusted to be at least 0.65.

Revised model where the Gender variable has been removed since it's not statistically significant.

casinomodel <- lm(Total.Spend ~ Age + Source, data = problem2)  
summary(casinomodel)

##   
## Call:  
## lm(formula = Total.Spend ~ Age + Source, data = problem2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1823.6 -990.8 -604.0 20.6 14238.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 122.577 133.838 0.916 0.360   
## Age 14.884 2.511 5.928 3.28e-09 \*\*\*  
## SourceWALK -52.835 90.052 -0.587 0.557   
## SourceWEB 596.021 77.720 7.669 2.07e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2197 on 4996 degrees of freedom  
## Multiple R-squared: 0.02675, Adjusted R-squared: 0.02617   
## F-statistic: 45.78 on 3 and 4996 DF, p-value: < 2.2e-16

In this revised model the R-squared adjusted has not changed much and is still around 0.026. Age and WEB are still significant.

The model equation is:

Total Spend = 122.577 + X1\**14.884 + x2*\*-52.835 + x3\*596.021

Where:

x1=Age

x2=WALK customer (1/0)

x3=WEB customer (1/0)

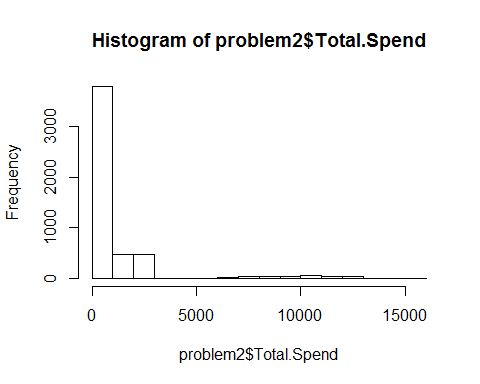
The third source variable for AAA is included in the intercept (122.577)

This means that for every unit increase in the variables, we can expect the total spend to increase/decrease by the value of the coefficients. For example, if the customer is WEB, it means that the total spending will increase by 596.021.

For the sake of creating a better model, let's try to focus only on high rollers

A secondary model only focusing on high rollers.

hist(problem2$Total.Spend)



The distribution of total spending doesn't seem to be normally distributed. By focusing in customers spending more that 5000, we may have more luck.

highrollers <- subset(problem2, Total.Spend >=5000)  
casinomodel <- lm(Total.Spend ~ Age + Source, data = highrollers)  
summary(casinomodel)

##   
## Call:  
## lm(formula = Total.Spend ~ Age + Source, data = highrollers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4659.9 -1524.0 17.5 1267.3 5507.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10798.59 1005.41 10.741 <2e-16 \*\*\*  
## Age -17.86 18.48 -0.967 0.335   
## SourceWALK -61.83 571.68 -0.108 0.914   
## SourceWEB 26.53 424.00 0.063 0.950   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1895 on 258 degrees of freedom  
## Multiple R-squared: 0.00373, Adjusted R-squared: -0.007854   
## F-statistic: 0.322 on 3 and 258 DF, p-value: 0.8095

R-squared has not improved and none of the variables show statistical significance. We will stick with the previous model.

Business meeting model presentation:

This model is used to predict the spending of a certain type of customer in the Casino. We are using customer information that tell us the customers age, their gender and their customer type category (from the AAA travel, hotel guests or walk-ins). These are all parameters we have taken into account when developing this prediction model. What we have determined is that based on this information, we can not make an accurate prediction on the total customer spending. What we can do is tell you that to impact the average customer spending, and in turn your revenue, you should focus on hotel guests and the customers age. For every hotel guest you can expect the spending to increase by $596 from the base spending of $123. With every year of age, you can expect a further $15. To summarize, focus on on older overnight guests.

The model is not very reliable. The low R-squared is far from being ideal so when we make predictions using this model there is going to be a lot of error (big prediction interval). However, the variables Age and WEB are very significant which means that they provide good information about the response even if we have high unexplained variability. We can say with confidence that the afformentioned variables have an impact on the total spending.

To increase the Casino's revenue, the executives should increase the number/portion of older overnights guests. If they decide to steer their marketing efforts to focus on that customer segment they will see an increase in total customer spending, and subsequently, their revenue.

Problem 3

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tm)

## Loading required package: NLP

library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':  
##   
## annotate

library(stm)

## stm v1.1.3 (2016-01-14) successfully loaded. See ?stm for help.

library(tm)

Load .csv files with the discussions

precorpus<-read.csv("MyGivingStory.csv",header=TRUE, stringsAsFactors=FALSE)  
attach(precorpus)  
names(precorpus)

## [1] "Inclusion" "votes\_total" "votes\_gallery\_.Calc."  
## [4] "votes\_local\_network" "createdate" "ipaddress"   
## [7] "EIN" "Photo" "URL\_in\_Story\_.Calc."   
## [10] "matched\_zip" "Story\_Text" "give\_reason"   
## [13] "give\_method"

Passing Study Text to a variable corpus

corpus<-precorpus$Story\_Text

Cleaning corpus

require(quanteda)

## Loading required package: quanteda

## quanteda version 0.9.8.5

##   
## Attaching package: 'quanteda'

## The following objects are masked from 'package:tm':  
##   
## as.DocumentTermMatrix, stopwords

## The following object is masked from 'package:NLP':  
##   
## ngrams

## The following object is masked from 'package:base':  
##   
## sample

Additional junk words showing in the data

stop\_words <- stopwords("SMART")  
stop\_words <- c(stop\_words,"a", "about", "above", "above", "across", "after", "afterwards", "again", "against", "all", "almost", "alone", "along", "already", "also","although","always","am","among", "amongst", "amoungst", "amount", "an", "and", "another", "any","anyhow","anyone","anything","anyway", "anywhere", "are", "around", "as", "at", "back","be","became", "because","become","becomes", "becoming", "been", "before", "beforehand", "behind", "being", "below", "beside", "besides", "between", "beyond", "bill", "both", "bottom","but", "by", "call", "can", "cannot", "cant", "co", "con", "could", "couldnt", "cry", "de", "describe", "detail", "do", "done", "down", "due", "during", "each", "eg", "eight", "either", "eleven","else", "elsewhere", "empty", "enough", "etc", "even", "ever", "every", "everyone", "everything", "everywhere", "except", "few", "fifteen", "fify", "fill", "find", "fire", "first", "five", "for", "former", "formerly", "forty", "found", "four", "from", "front", "full", "further", "get", "give", "go", "had", "has", "hasnt", "have", "he", "hence", "her", "here", "hereafter", "hereby", "herein", "hereupon", "hers", "herself", "him", "himself", "his", "how", "however", "hundred", "ie", "if", "in", "inc", "indeed", "interest", "into", "is", "it", "its", "itself", "keep", "last", "latter", "latterly", "least", "less", "ltd", "made", "many", "may", "me", "meanwhile", "might", "mill", "mine", "more", "moreover", "most", "mostly", "move", "much", "must", "my", "myself", "name", "namely", "neither", "never", "nevertheless", "next", "nine", "no", "nobody", "none", "noone", "nor", "not", "nothing", "now", "nowhere", "of", "off", "often", "on", "once", "one", "only", "onto", "or", "other", "others", "otherwise", "our", "ours", "ourselves", "out", "over", "own","part", "per", "perhaps", "please", "put", "rather", "re", "same", "see", "seem", "seemed", "seeming", "seems", "serious", "several", "she", "should", "show", "side", "since", "sincere", "six", "sixty", "so", "some", "somehow", "someone", "something", "sometime", "sometimes", "somewhere", "still", "such", "system", "take", "ten", "than", "that", "the", "their", "them", "themselves", "then", "thence", "there", "thereafter", "thereby", "therefore", "therein", "thereupon", "these", "they", "thickv", "thin", "third", "this", "those", "though", "three", "through", "throughout", "thru", "thus", "to", "together", "too", "top", "toward", "towards", "twelve", "twenty", "two", "un", "under", "until", "up", "upon", "us", "very", "via", "was", "we", "well", "were", "what", "whatever", "when", "whence", "whenever", "where", "whereafter", "whereas", "whereby", "wherein", "whereupon", "wherever", "whether", "which", "while", "whither", "who", "whoever", "whole", "whom", "whose", "why", "will", "with", "within", "without", "would", "yet", "you", "your", "yours", "yourself", "yourselves", "the")

Removing Apostrophes, Punctuations with space, Control Characters with space, Whitespace at the begining & end of Document, Alowing only letters, Forces Lowercase

stop\_words <- tolower(stop\_words)  
corpus <- gsub("'", "", corpus)  
corpus <- gsub("[[:punct:]]", " ", corpus)  
corpus <- gsub("[[:cntrl:]]", " ", corpus)  
corpus <- gsub("^[[:space:]]+", "", corpus)   
corpus <- gsub("[[:space:]]+$", "", corpus)   
corpus <- gsub("[^a-zA-Z -]", " ", corpus)   
corpus <- tolower(corpus)

Get rid of Blank docs

corpus <- corpus[corpus != ""]

Tokenize on space and output as a list

doc.list <- strsplit(corpus, "[[:space:]]+")

Compute the table of terms

term.table <- table(unlist(doc.list))  
term.table <- sort(term.table, decreasing = TRUE)

remove terms that are stop words or occur fewer than 5 times:

del <- names(term.table) %in% stop\_words | term.table < 5  
term.table <- term.table[!del]  
term.table <- term.table[names(term.table) != ""]  
vocab <- names(term.table)

Now put the documents into the format required by the lda package

get.terms <- function(x) {  
 index <- match(x, vocab)  
 index <- index[!is.na(index)]  
 rbind(as.integer(index - 1), as.integer(rep(1, length(index))))  
 }  
documents <- lapply(doc.list, get.terms)

Compute some statistics related to the data set:

D <- length(documents)  
W <- length(vocab)   
doc.length <- sapply(documents, function(x) sum(x[2, ]))  
N <- sum(doc.length)   
term.frequency <- as.integer(term.table)

MCMC and model tuning parameters:

K <- 10  
G <- 5000  
alpha <- 0.02  
eta <- 0.02

Fit the model:

library(lda)  
set.seed(357)  
t1 <- Sys.time()  
fit <- lda.collapsed.gibbs.sampler(documents = documents, K = K, vocab = vocab,   
 num.iterations = G, alpha = alpha,   
 eta = eta, initial = NULL, burnin = 0,  
 compute.log.likelihood = TRUE)  
t2 <- Sys.time()  
  
t2 - t1

## Time difference of 2.808144 mins

theta <- t(apply(fit$document\_sums + alpha, 2, function(x) x/sum(x)))  
phi <- t(apply(t(fit$topics) + eta, 2, function(x) x/sum(x)))  
   
news\_for\_LDA <- list(phi = phi,  
 theta = theta,  
 doc.length = doc.length,  
 vocab = vocab,  
 term.frequency = term.frequency)

Create the JSON object to feed the visualization:

library(LDAvis)  
library(servr)  
  
json <- createJSON(phi = news\_for\_LDA$phi,   
 theta = news\_for\_LDA$theta,   
 doc.length = news\_for\_LDA$doc.length,   
 vocab = news\_for\_LDA$vocab,   
 term.frequency = news\_for\_LDA$term.frequency)  
library(gistr)

##   
## Attaching package: 'gistr'

## The following objects are masked from 'package:stats':  
##   
## embed, update

serVis(json, out.dir = 'new2', open.browser = FALSE)

Link to LDAvis: <http://bl.ocks.org/sdb418/raw/803ab4bc4c19b829ee281eb6eafdeac5/>

Below are examples of how the Bill and Melinda Gates foundation can implement the results of the topic analysis for further improving the #MyGivingStory2016 campaign. The marketing mission for the campaign is to try to involve and engage as many people as possible and in return convincing people to submit their stories. One way for the foundation to achieve that goal is to focus their marketing campaign in the direction that wins the broadest appeal of its audience. By utilizing the topic analysis, they will be able to figure out what topics most people relate to.

Looking at topic number one with words like people, life, community, support, family and time we see that the reason people give is rather primal. They are concerned with things closest to them like their family or community. The marketing campaign should be constructed around these topics that are common throughout the stories.

People who share their stories and share photos on Facebook and Twitter tend to get more “Likes” which suggest photos are more appealing to people and as they give more exposure to campaign Mr. Gates should get more active on photo sharing apps like Instagram and Snapchat.

One way of leveling the playing field would be to introduce story categories that had individual awards. The topics could be used to create the categories. Examples of categories could be Women’s Education (Topic 6) with frequent occurring word like girls, women, education, Wildlife preservation (Topic 10) with frequent occurring words like wildlife, forest and Animal compassion (Topic 7) with words like dogs, animals, rescue, foster.

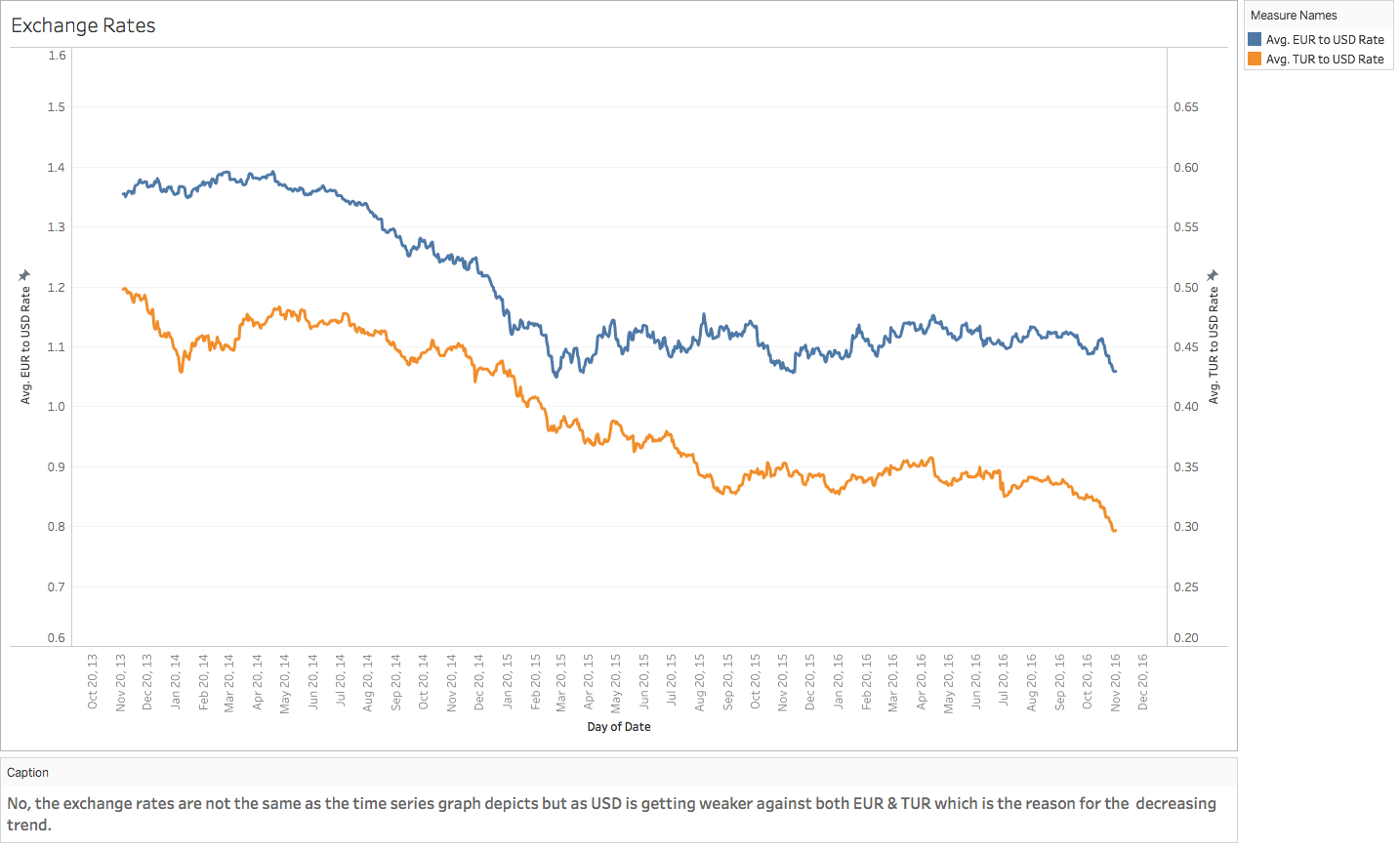
Problem 4

# Time Series Objects

ts.all<-ts(timeseries1)  
ts.1<-ts(timeseries1$EUR.to.USD.Rate)   
ts.2 <- ts(timeseries1$TUR.to.USD.Rate)

ts.month<-ts(timeseries1$EUR.to.USD.Rate, frequency = 12,start=c(2013,11), end=c(2016,11))   
ts.month1<-ts(timeseries1$TUR.to.USD.Rate, frequency = 12,start=c(2013,11), end=c(2016,11))

# Plotting Time Series

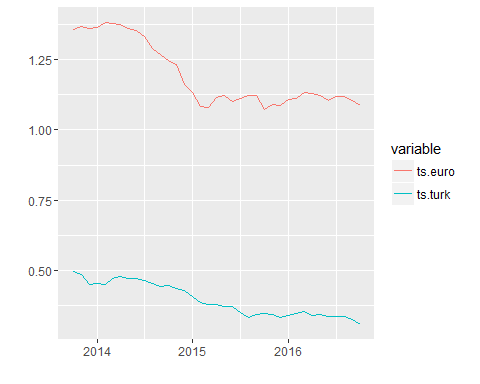


# Quesiton 2 plotted on Tablaeu

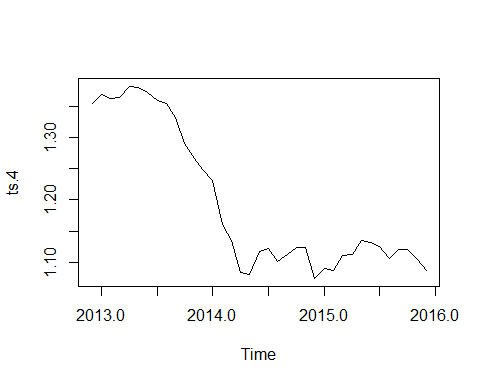
library(ggfortify)

## Loading required package: ggplot2

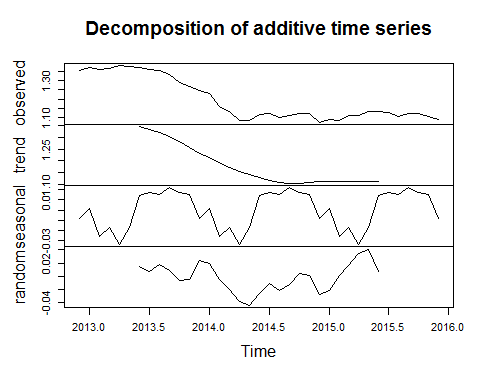
ts.euro<-ts(ts.month)  
ts.turk<-ts(ts.month1)  
autoplot(ts( cbind(ts.euro, ts.turk) , start = c(2013,10),end=c(2016,10), frequency = 12 ),facets = FALSE)

 #Decomposing Time Series

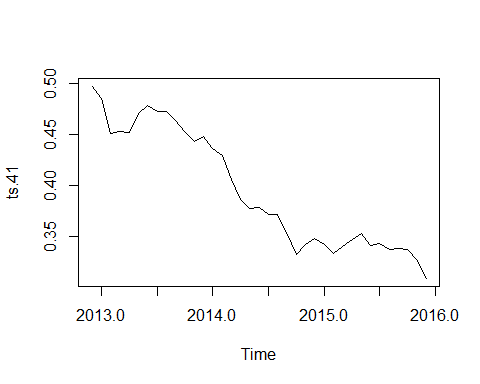
ts.4 <- ts(timeseries1$EUR.to.USD.Rate, frequency = 12,start=c(2013,0), end=c(2016,0))  
plot.ts(ts.4)



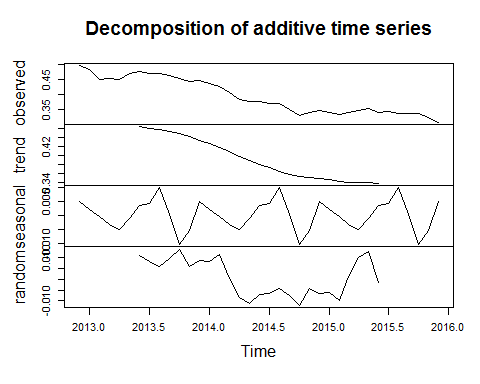
ts.4.d <- decompose(ts.4)  
plot(ts.4.d)



ts.41 <- ts(timeseries1$TUR.to.USD.Rate, frequency = 12,start=c(2013,0), end=c(2016,0))  
plot.ts(ts.41)



ts.41.d <- decompose(ts.41)  
plot(ts.41.d)

 # Question 1 Trends for Euro : According to the data and the decomposition of the time series we can see that there is a decreasing trend in the lira to usd data.

Seasonality for Euro :From the limited date we have it looks like its is seasonal but its not logical that currency can be seasonal

Trends for Lira :Simialr to the Euro there is a decereasing trend for the Lira

Seasonality for Lira : For this data it looks like it is seasonal but currency it does not make sesne for currency to be seasonal

library(forecast)

## Loading required package: timeDate

## This is forecast 7.3

##   
## Attaching package: 'forecast'

## The following object is masked from 'package:ggfortify':  
##   
## gglagplot

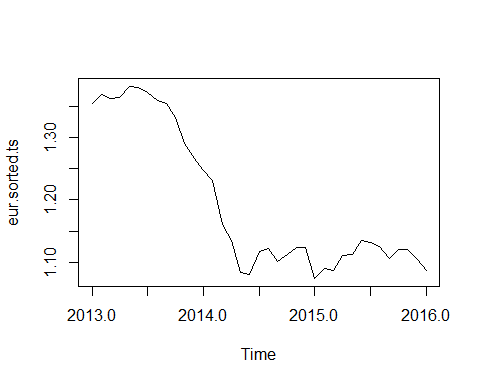
## The following object is masked from 'package:hydroTSM':  
##   
## ma

library(data.table)

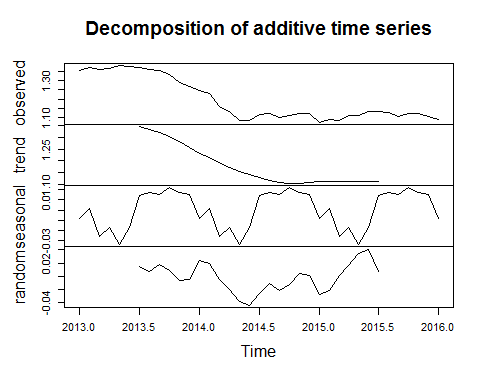
##   
## Attaching package: 'data.table'

## The following object is masked from 'package:xts':  
##   
## last

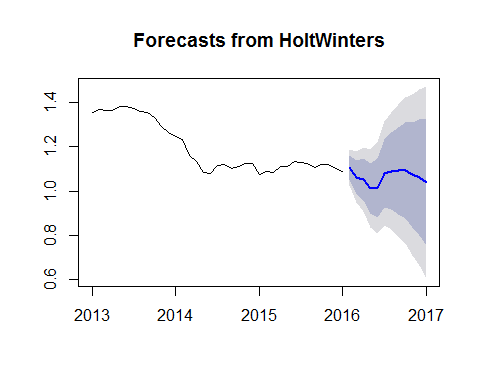
setDT(timeseries1, keep.rownames = TRUE)  
eur.sorted<-timeseries1[order(as.Date(timeseries1$rn, format="%d/%m/%Y")),]  
eur.sorted.ts<-ts(eur.sorted$EUR.to.USD.Rate,start=c(2013),frequency = 12)  
plot(eur.sorted.ts)



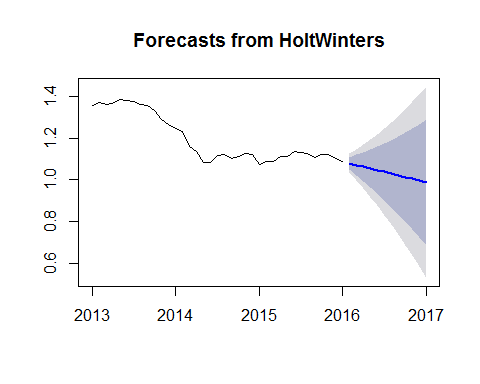
eur.sorted.ts.d<-decompose(eur.sorted.ts)  
plot(eur.sorted.ts.d)



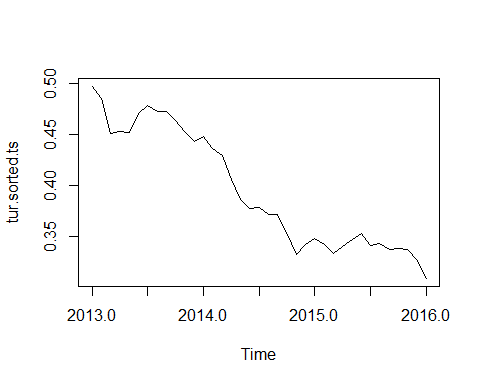
eur.holt.T <- HoltWinters(eur.sorted.ts, gamma=TRUE)  
eur.holt.F <- HoltWinters(eur.sorted.ts, gamma=FALSE)  
eur.forecasts <- forecast.HoltWinters(eur.holt.T, h=12)   
plot.forecast(eur.forecasts)



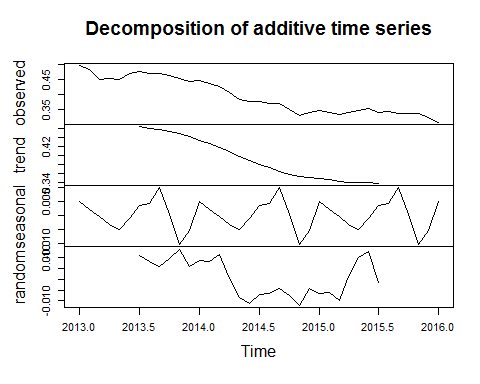
eur.forecasts.F <- forecast.HoltWinters(eur.holt.F, h=12)   
plot.forecast(eur.forecasts.F)



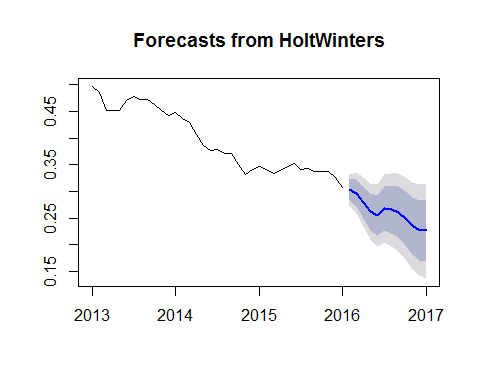
############################################  
#  
library(data.table)  
setDT(timeseries1, keep.rownames = TRUE)  
tur.sorted<-timeseries1[order(as.Date(timeseries1$rn, format="%d/%m/%Y")),]  
tur.sorted.ts<-ts(eur.sorted$TUR.to.USD.Rate,start=c(2013),frequency = 12)  
plot(tur.sorted.ts)



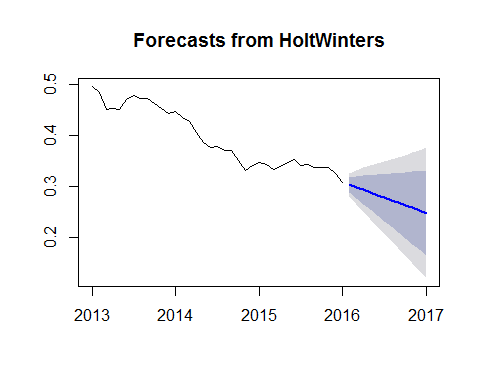
tur.sorted.ts.d<-decompose(tur.sorted.ts)  
plot(tur.sorted.ts.d)



tur.holt.T <- HoltWinters(tur.sorted.ts, gamma=TRUE)  
tur.holt.F <- HoltWinters(tur.sorted.ts, gamma=FALSE)  
tur.forecasts <- forecast.HoltWinters(tur.holt.T, h=12)   
plot.forecast(tur.forecasts)



tur.forecasts.F <- forecast.HoltWinters(tur.holt.F, h=12)   
plot.forecast(tur.forecasts.F)



# Question 3

# Forecasting Using an ARIMA Model

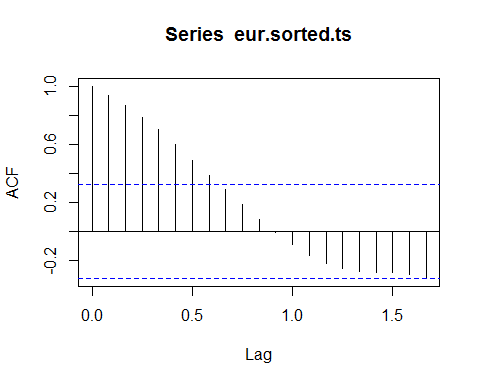
EURO TO $ Time series The first step to performing an arima model is to check if the time series is stationary or not.From the auto correlation function and the partial auto correlation function.By performing the acf and pacf it is clear that there are no significant lags beyond 1.Additionally by performing the L-jung Box test we got a p value that is less than 0.05 so we can conclude that the time series is stationary. The output was ARIMA(0,1,0) with drift for euro to dollar.But by experimentation and our choice of preference we came to the conclusion that ARIMA (3,3,3) is better model

TUR TO $ Time Series The correlograms and the partial correlograms for the TUR TO dollar also have no siginificant lags beyond one.By using the auto arima function R suggest the same the model as that of EUR to dollar.Hence we use the ARIMA(0,1,0) with Drift.After testing out various models we came to a conslusion that ARIMA (3,3,3) is a better model

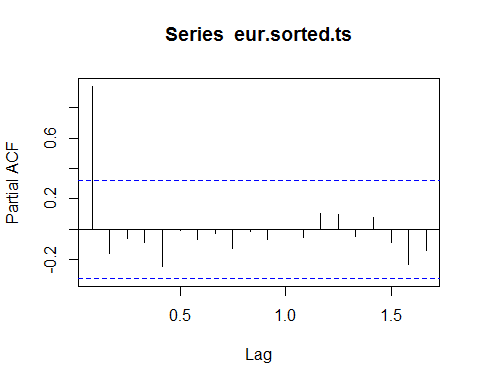
library(forecast)  
auto.arima(eur.sorted.ts)

## Series: eur.sorted.ts   
## ARIMA(0,1,0) with drift   
##   
## Coefficients:  
## drift  
## -0.0075  
## s.e. 0.0037  
##   
## sigma^2 estimated as 0.0005038: log likelihood=86.11  
## AIC=-168.21 AICc=-167.85 BIC=-165.04

acf(eur.sorted.ts, lag.max = 20) # sinece there is no lag



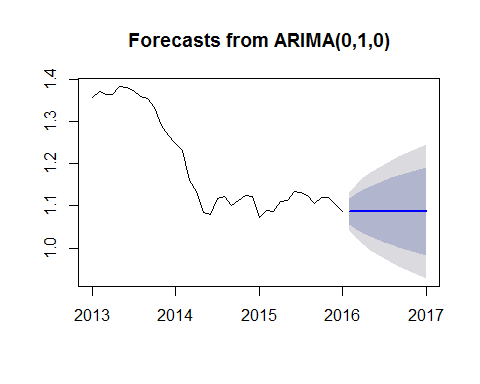
pacf(eur.sorted.ts,lag.max =20)



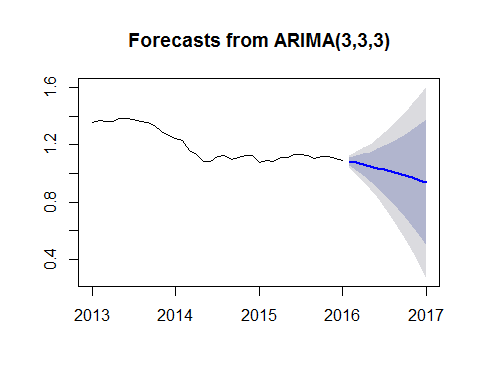
Box.test(eur.sorted.ts,type = "Ljung-Box") #p-value less than 0.05 do stationary

##   
## Box-Ljung test  
##   
## data: eur.sorted.ts  
## X-squared = 35.421, df = 1, p-value = 2.657e-09

eur.arima<-arima(eur.sorted.ts, c(0,1,0))   
eur.arima.forecasts <- forecast.Arima(eur.arima, h=12)  
plot(eur.arima.forecasts)



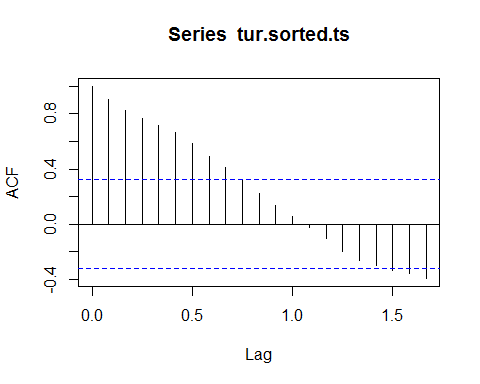
eur.arima<-arima(eur.sorted.ts, c(3,3,3))   
eur.arima.forecasts <- forecast.Arima(eur.arima, h=12)  
plot(eur.arima.forecasts)



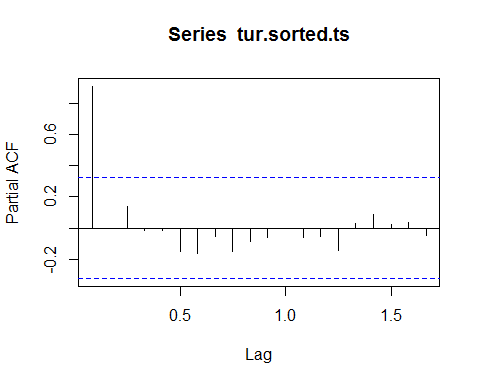
########  
  
auto.arima(tur.sorted.ts)

## Series: tur.sorted.ts   
## ARIMA(0,1,0) with drift   
##   
## Coefficients:  
## drift  
## -0.0052  
## s.e. 0.0018  
##   
## sigma^2 estimated as 0.0001151: log likelihood=112.69  
## AIC=-221.38 AICc=-221.01 BIC=-218.21

acf(tur.sorted.ts, lag.max = 20)



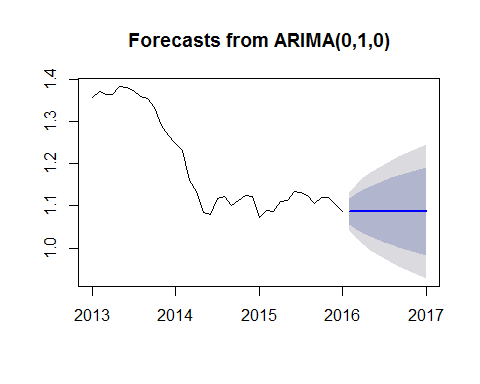
pacf(tur.sorted.ts,lag.max =20)



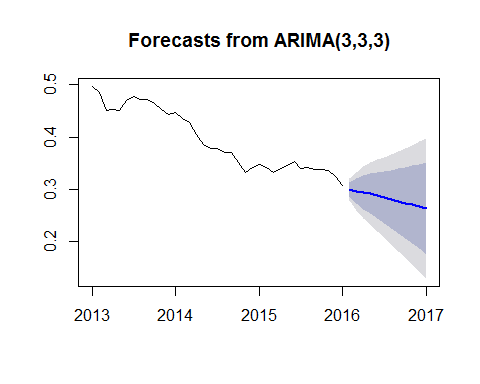
Box.test(tur.sorted.ts,type = "Ljung-Box")

##   
## Box-Ljung test  
##   
## data: tur.sorted.ts  
## X-squared = 32.985, df = 1, p-value = 9.288e-09

eur.arima<-arima(eur.sorted.ts, c(0,1,0))   
eur.arima.forecasts <- forecast.Arima(eur.arima, h=12)  
plot(eur.arima.forecasts)



tur.arima<-arima(tur.sorted.ts, c(3,3,3))   
tur.arima.forecasts <- forecast.Arima(tur.arima, h=12)  
plot(tur.arima.forecasts)

 # Interpretations from the ARIMA models

EUR TO USD : Our model shows that the dollar is becoming weaker when compared to the euro in 2017,for US based companies that generate the majority of their revenue in the US but pay their expenses in Euro's our advice would be to invest more now when the dollar is stronger so they suffer less in the future. On the other hand, US companies draw in a significant amount of revenue in euros, but pay their employees and other expenses in US dollars would fare better so we would advise them to invest less. Additionally, the political situation in the US could also be another reason for the weaker dollar

LIRA TO USD : The ARIMA model shows that the LIRA is growing stronger when compared to the dollar in 2017. Our advice for a US based firm would be similar to the situation in Europe if the company generates majority of the income in the US but pay their expenses in Lira's we would advise the company to invest more now. On the other hand if the company generate majority of the revenue in Lira's but pays in dollars we would advise them to invest less.