VN\_ecommerce

NguyenLSCMFTU2

10/03/2020

#Introduction

This project is intended to compare the number of searching times based on keywords these are Tiki, Sendo, Shopee,Lazada, then understand more and more e-commerce users activites within a **week** (7 day)

#Set up the packages and standardize datasets to visualize

##Load packages

suppressMessages(library(tidyverse))  
suppressMessages(library(viridis))  
suppressMessages(library(ggsci))  
suppressMessages(library(lubridate))  
suppressMessages(library(stringr))  
suppressMessages(library(stringi))  
suppressMessages(library(explore))  
suppressMessages(library(DT))  
suppressMessages(library(Amelia))  
suppressMessages(library(gridExtra))  
  
theme\_set(theme\_minimal())

##Load datasets

These concepts to manipulate datasets are create a list then use the *lapply* to replicate to optimize time process

dataset<- list()  
dataset<- lapply(Sys.glob("\*.csv"),read.csv)  
  
"Built function to standardize"

## [1] "Built function to standardize"

clean\_function <- function(dt) {  
   
 dt <- dt %>% mutate(date = as.POSIXct(str\_replace\_all(date,"T"," "), format = "%Y-%m-%d %H"))  
 }  
dataset<-lapply(dataset, clean\_function)  
  
#Check na-value in dataset  
  
##Final datasets  
lazada<-as.data.frame(dataset[1])%>% mutate(hour=factor(hour(date)))%>% mutate(wday=factor(wday(date)))  
sendo<-as.data.frame(dataset[2])%>% mutate(hour=factor(hour(date)))%>% mutate(wday=factor(wday(date)))  
shopee<-as.data.frame(dataset[3])%>% mutate(hour=factor(hour(date)))%>% mutate(wday=factor(wday(date)))  
tiki<-as.data.frame(dataset[4])%>% mutate(hour=factor(hour(date)))%>% mutate(wday=factor(wday(date)))

## Let have a look at sample Lazada dataset right now

knitr::kable(head(lazada))

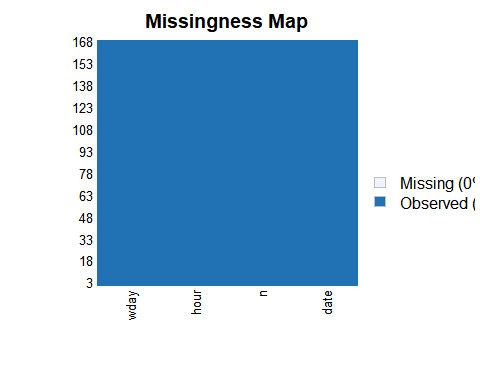
|  |  |  |  |
| --- | --- | --- | --- |
| date | n | hour | wday |
| 2020-03-03 09:00:00 | 86 | 9 | 3 |
| 2020-03-03 10:00:00 | 87 | 10 | 3 |
| 2020-03-03 11:00:00 | 92 | 11 | 3 |
| 2020-03-03 12:00:00 | 81 | 12 | 3 |
| 2020-03-03 13:00:00 | 84 | 13 | 3 |
| 2020-03-03 14:00:00 | 86 | 14 | 3 |

From these table, you can easily see the main components are time and n (n is the total of searching times in each record (row))

I also standardize the datasets and extract the weekdays and hour for these below purposes

##Check NA values Once more time, be sure that you are comfortable without NA values

#Test na in datasets  
missmap(tiki)  
missmap(lazada)



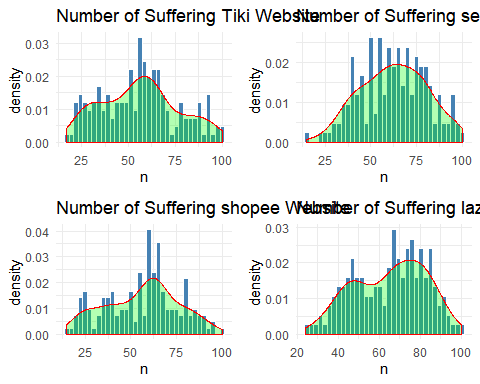
missmap(sendo)  
missmap(shopee)

These result is quite good

#Visualize the datasets

##Histogram in data set

tiki\_plot<-tiki%>% ggplot(aes(x=n,y=..density..))+geom\_histogram(bins=35,fill="steelblue",col="white")+geom\_density(fill="green",alpha=0.3,col="red")+ggtitle(label="Number of Suffering Tiki Website")  
  
lazada\_plot<-lazada%>% ggplot(aes(x=n,y=..density..))+geom\_histogram(bins=35,fill="steelblue",col="white")+geom\_density(fill="green",alpha=0.3,col="red")+ggtitle(label="Number of Suffering lazada Website")  
  
sendo\_plot<-sendo%>% ggplot(aes(x=n,y=..density..))+geom\_histogram(bins=35,fill="steelblue",col="white")+geom\_density(fill="green",alpha=0.3,col="red")+ggtitle(label="Number of Suffering sendo Website")  
  
shopee\_plot<-shopee%>% ggplot(aes(x=n,y=..density..))+geom\_histogram(bins=35,fill="steelblue",col="white")+geom\_density(fill="green",alpha=0.3,col="red")+ggtitle(label="Number of Suffering shopee Website")  
  
 grid.arrange(tiki\_plot,sendo\_plot,shopee\_plot,lazada\_plot)

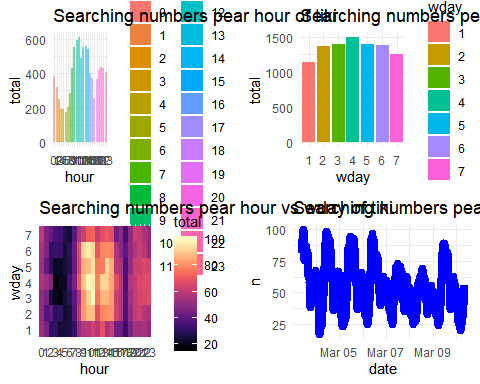


From these plot I can conclude that Shopee has the largest number of searching keyword. But bins are not quite same. Likewise, Lazada frequency clearly separates to 2 parts. Tiki and Sendo are quite same day to day

##Hour vs Weekday

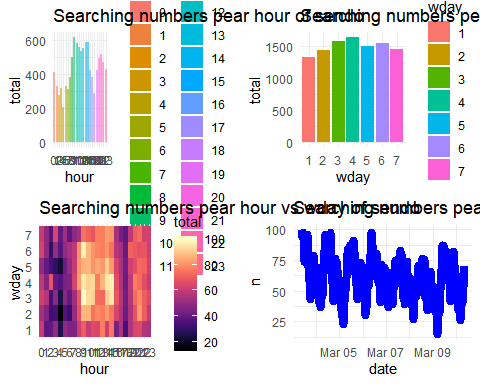
###Tiki

plot1<- tiki%>% group\_by(hour)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=total,fill=hour,alpha=0.5))+geom\_col()+labs(title="Searching numbers pear hour of tiki",xlabs="Hour",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot2<-tiki%>% group\_by(wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=wday,y=total,fill=wday))+geom\_col()+labs(title="Searching numbers pear wday of tiki",xlabs="Wday",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot3<- tiki%>%group\_by(hour,wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=wday))+geom\_tile(aes(fill=total))+scale\_fill\_viridis(option="A")+labs(title="Searching numbers pear hour vs wday of tiki",xlabs="Hour",ylabs="Wday")  
  
plot4<- tiki%>% ggplot(aes(x=date,y=n))+geom\_line(aes(group=1),color="blue",lwd=3)+labs(title="Searching numbers pear hour vs wday of tiki",xlabs="Date",ylabs="Frequency")  
  
 grid.arrange(plot1,plot2,plot3,plot4,nrow=2,ncol=2)



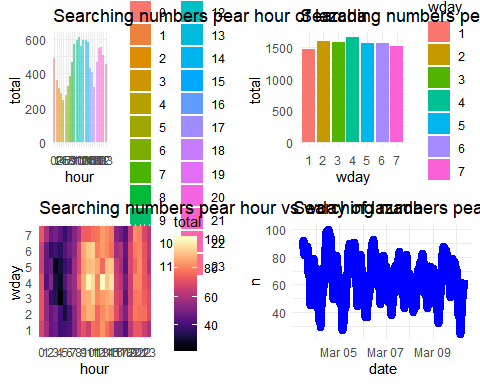
###sendo

plot1<-sendo%>% group\_by(hour)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=total,fill=hour,alpha=0.5))+geom\_col()+labs(title="Searching numbers pear hour of sendo",xlabs="Hour",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot2<-sendo%>% group\_by(wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=wday,y=total,fill=wday))+geom\_col()+labs(title="Searching numbers pear wday of sendo",xlabs="Wday",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot3<- sendo%>%group\_by(hour,wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=wday))+geom\_tile(aes(fill=total))+scale\_fill\_viridis(option="A")+labs(title="Searching numbers pear hour vs wday of sendo",xlabs="Hour",ylabs="Wday")  
  
plot4<- sendo%>% ggplot(aes(x=date,y=n))+geom\_line(aes(group=1),color="blue",lwd=3)+labs(title="Searching numbers pear hour vs wday of sendo",xlabs="Date",ylabs="Frequency")  
  
 grid.arrange(plot1,plot2,plot3,plot4,nrow=2,ncol=2)



###lazada

plot1<-lazada%>% group\_by(hour)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=total,fill=hour,alpha=0.5))+geom\_col()+labs(title="Searching numbers pear hour of lazada",xlabs="Hour",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot2<-lazada%>% group\_by(wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=wday,y=total,fill=wday))+geom\_col()+labs(title="Searching numbers pear wday of lazada",xlabs="Wday",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot3<- lazada%>%group\_by(hour,wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=wday))+geom\_tile(aes(fill=total))+scale\_fill\_viridis(option="A")+labs(title="Searching numbers pear hour vs wday of lazada",xlabs="Hour",ylabs="Wday")  
  
plot4<- lazada%>% ggplot(aes(x=date,y=n))+geom\_line(aes(group=1),color="blue",lwd=3)+labs(title="Searching numbers pear hour vs wday of lazada",xlabs="Date",ylabs="Frequency")  
  
 grid.arrange(plot1,plot2,plot3,plot4,nrow=2,ncol=2)



###shopee

plot1<-shopee%>% group\_by(hour)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=total,fill=hour,alpha=0.5))+geom\_col()+labs(title="Searching numbers pear hour of shopee",xlabs="Hour",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot2<-shopee%>% group\_by(wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=wday,y=total,fill=wday))+geom\_col()+labs(title="Searching numbers pear wday of shopee",xlabs="Wday",ylabs="Frequency")+scale\_color\_viridis(discrete=TRUE,option="A")  
  
plot3<- shopee%>%group\_by(hour,wday)%>%summarise(total=sum(n))%>%ggplot(aes(x=hour,y=wday))+geom\_tile(aes(fill=total))+scale\_fill\_viridis(option="A")+labs(title="Searching numbers pear hour vs wday of shopee",xlabs="Hour",ylabs="Wday")  
  
plot4<- shopee%>%ggplot(aes(x=date,y=n))+geom\_line(aes(group=1),color="blue",lwd=3)+labs(title="Searching numbers pear hour vs wday of shopee",xlabs="Date",ylabs="Frequency")  
  
 grid.arrange(plot1,plot2,plot3,plot4,nrow=2,ncol=2)

