**PROJECT ON**

**EXPLORATORY DATA ANALYSIS OF ZOMATO’S RESTAURANT DATASET**

**A comprehensive study**

**on**

****

Submitted To:

**Ms. Mrinali Mam**

Submitted in Partial Fulfilment of the Requirement for the Degree of

**Bachelor of Science in Computer Science**

Submitted by

**Yash Kumar (Roll No. 20/49032)**



Session 2020-2022

**Dyal Singh College (University of Delhi)**

**Department of Computer Science**

**About Dataset**

[*https://www.kaggle.com/shrutimehta/zomato-restaurants-data*](https://www.kaggle.com/shrutimehta/zomato-restaurants-data)

**1st Dataset**

Data has been collected from the zomato.csv file using above link.

**Every Restaurant contains the following variables:**

• Restaurant Id: Unique id of every restaurant across various cities of the world

• Restaurant Name: Name of the restaurant

• Country Code: Country in which restaurant is located

• City: City in which restaurant is located

• Address: Address of the restaurant

• Locality: Location in the city

• Locality Verbose: Detailed description of the locality

• Longitude: Longitude coordinate of the restaurant’s location

• Latitude: Latitude coordinate of the restaurant’s location

• Cuisines: Cuisines offered by the restaurant

• Average Cost for two: Cost for two people in different currencies 👫

• Currency: Currency of the country

• Has Table booking: yes/no

• Has Online delivery: yes/ no

• Is delivering: yes/ no

• Switch to order menu: yes/no

• Price range: range of price of food

• Aggregate Rating: Average rating out of 5

• Rating colour: depending upon the average rating colour

• Rating text: text-based on the rating of rating

• Votes: Number of ratings cast by people

**2nd Dataset**

Data has been collected from the Country-Code.xlsx file using above link.

**Each country has a unique country code and country name.**

* Country Code: Unique Id of the country
* Country: Name of the country

**Installing Packages:**

install.packages("dplyr")

install.packages("xlsx")

install.packages("ggplot2")

**Including Libraries:**

library("dplyr")

library("xlsx")

library("ggplot2")

getwd()

setwd("C:\\Users\\Yash Kumar\\Desktop\\R\_language\_tutorial")

**#Read the First File**

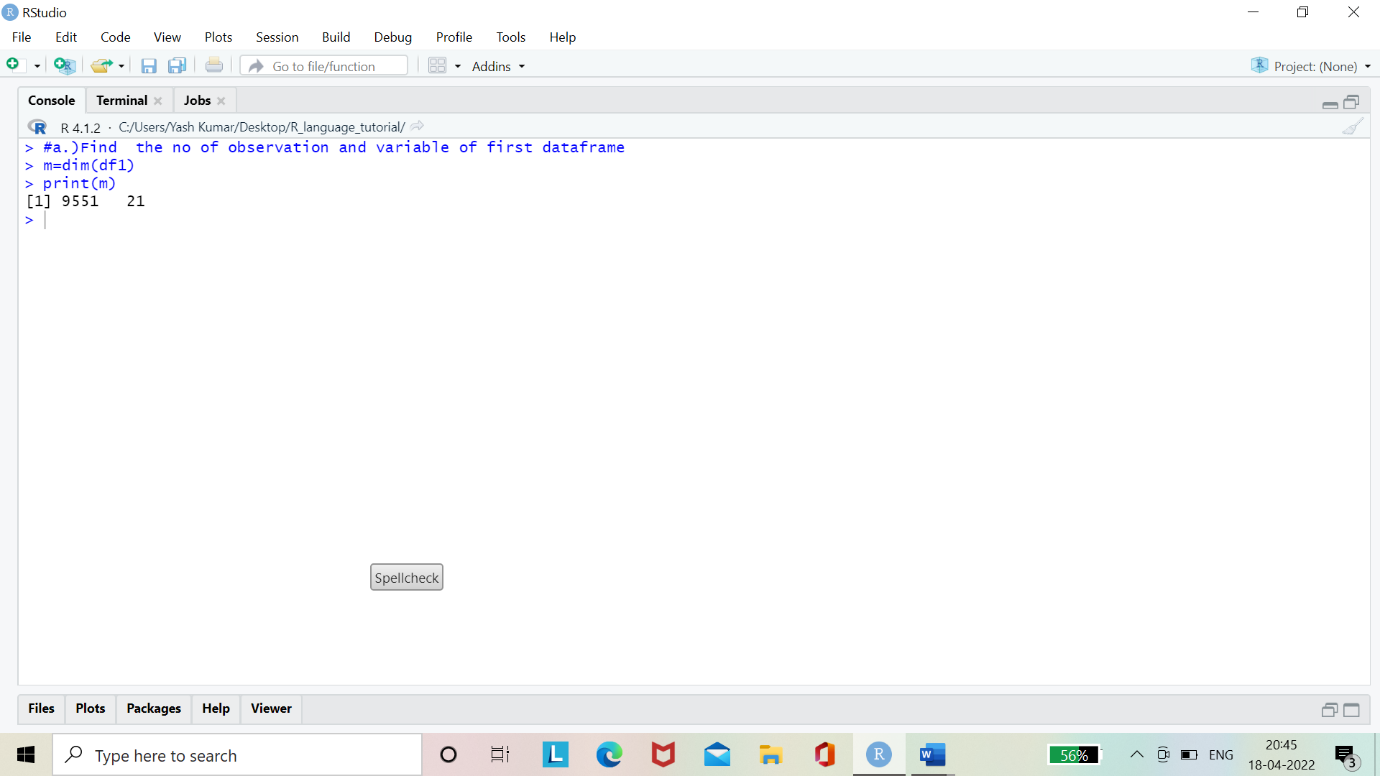
df1<-read.csv("zomato.csv")

df1

**#a.) Find no. of observation and variable of first Dataframe**

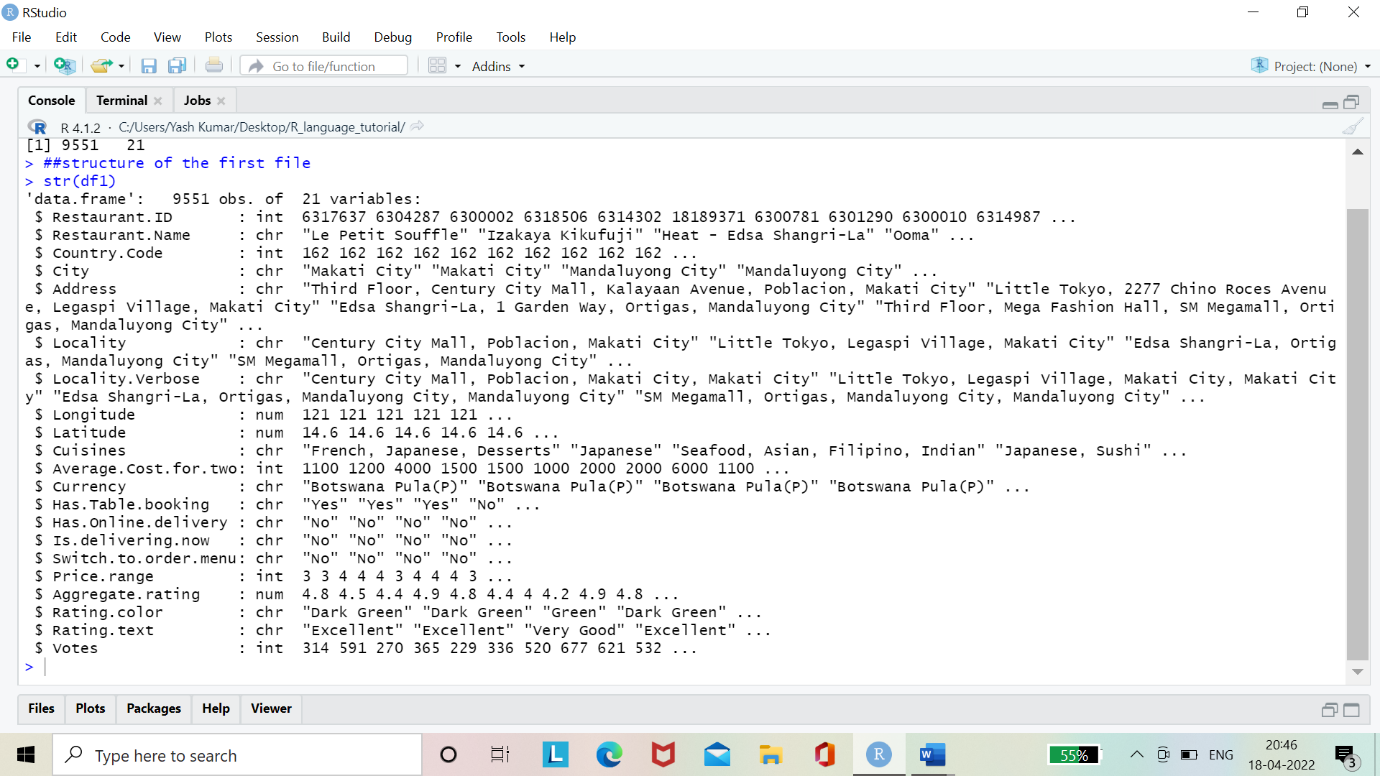
m=dim(df1)

print(m)



**#Structure of the first file**

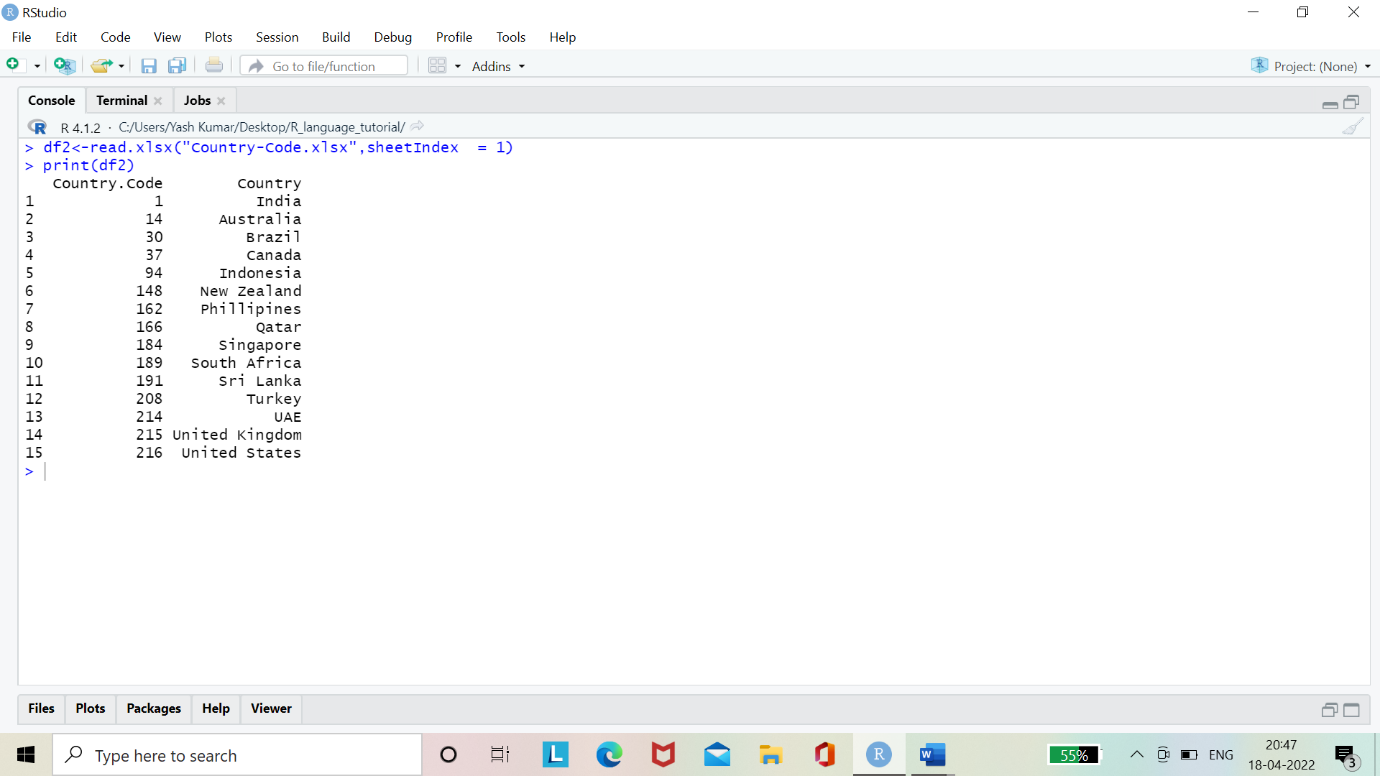
str(df1)



**#Read the Second File Country Code One**

df2<-read.xlsx("Country-Code.xlsx",sheetIndex = 1)

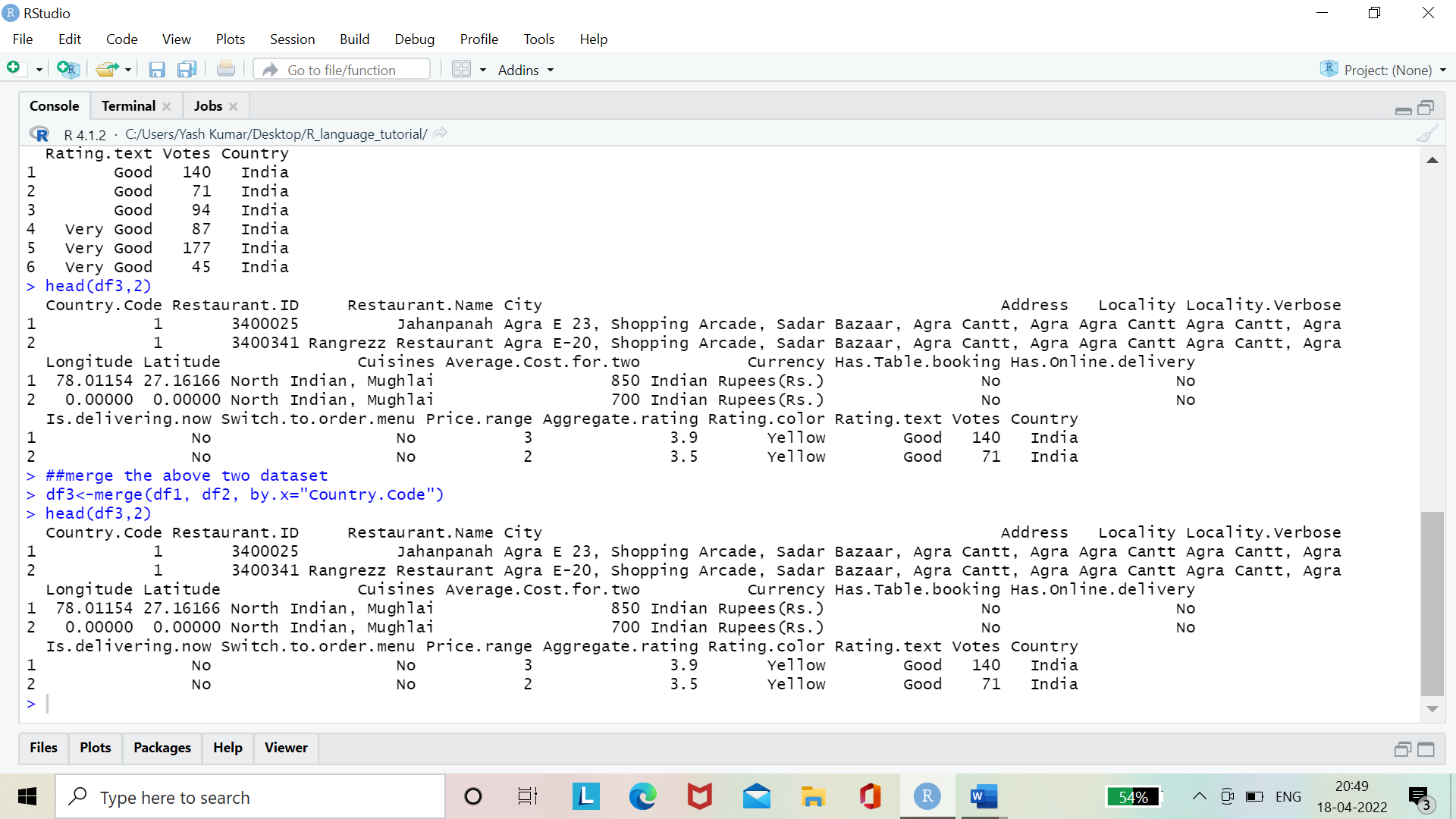
print(df2)



**#Merge the Above Two Dataset**

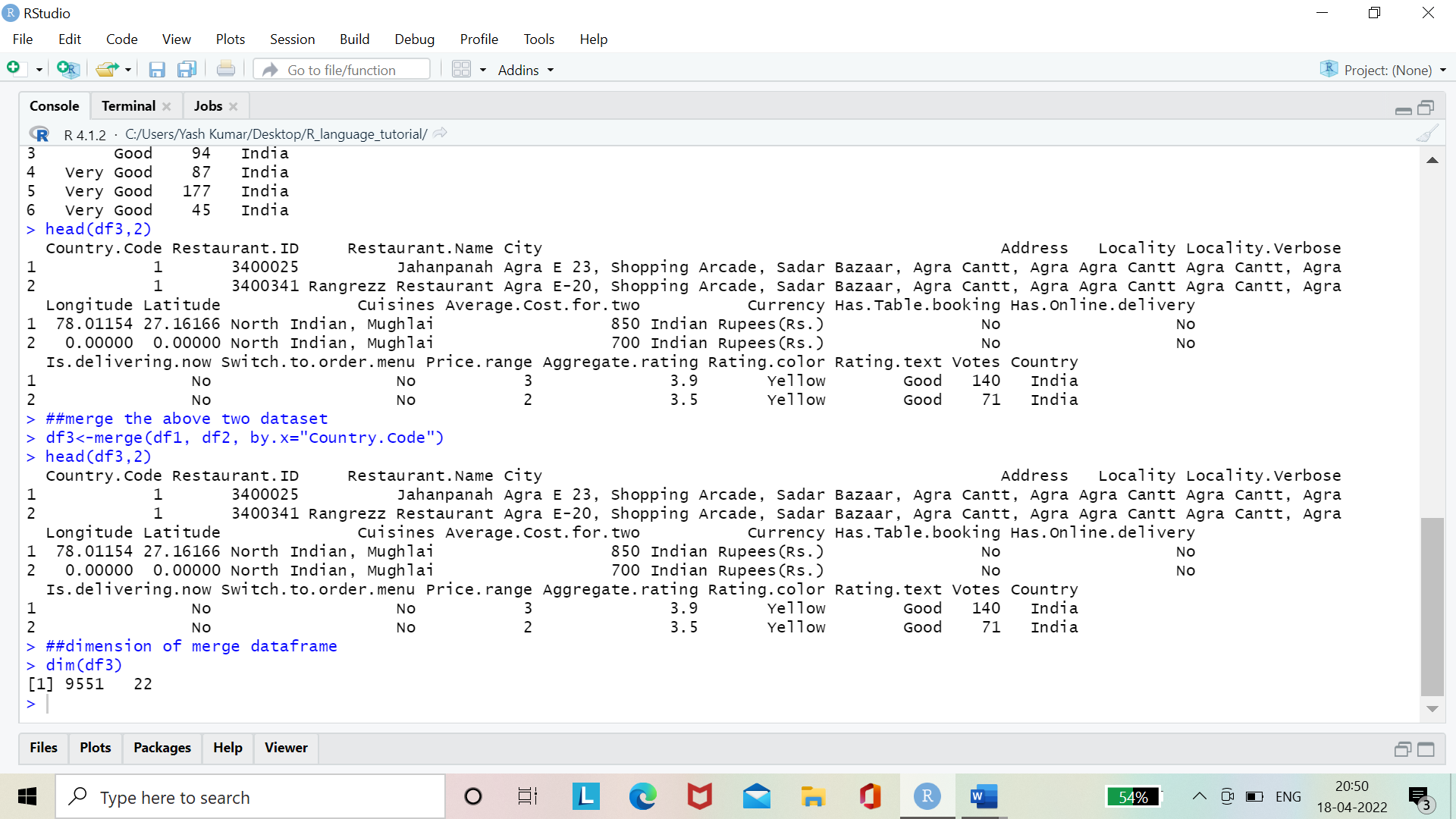
df3<-merge(df1, df2, by.x="Country.Code")

head(df3)



**#Dimension of Resultant Data Frame(after merging)**

dim(df3)

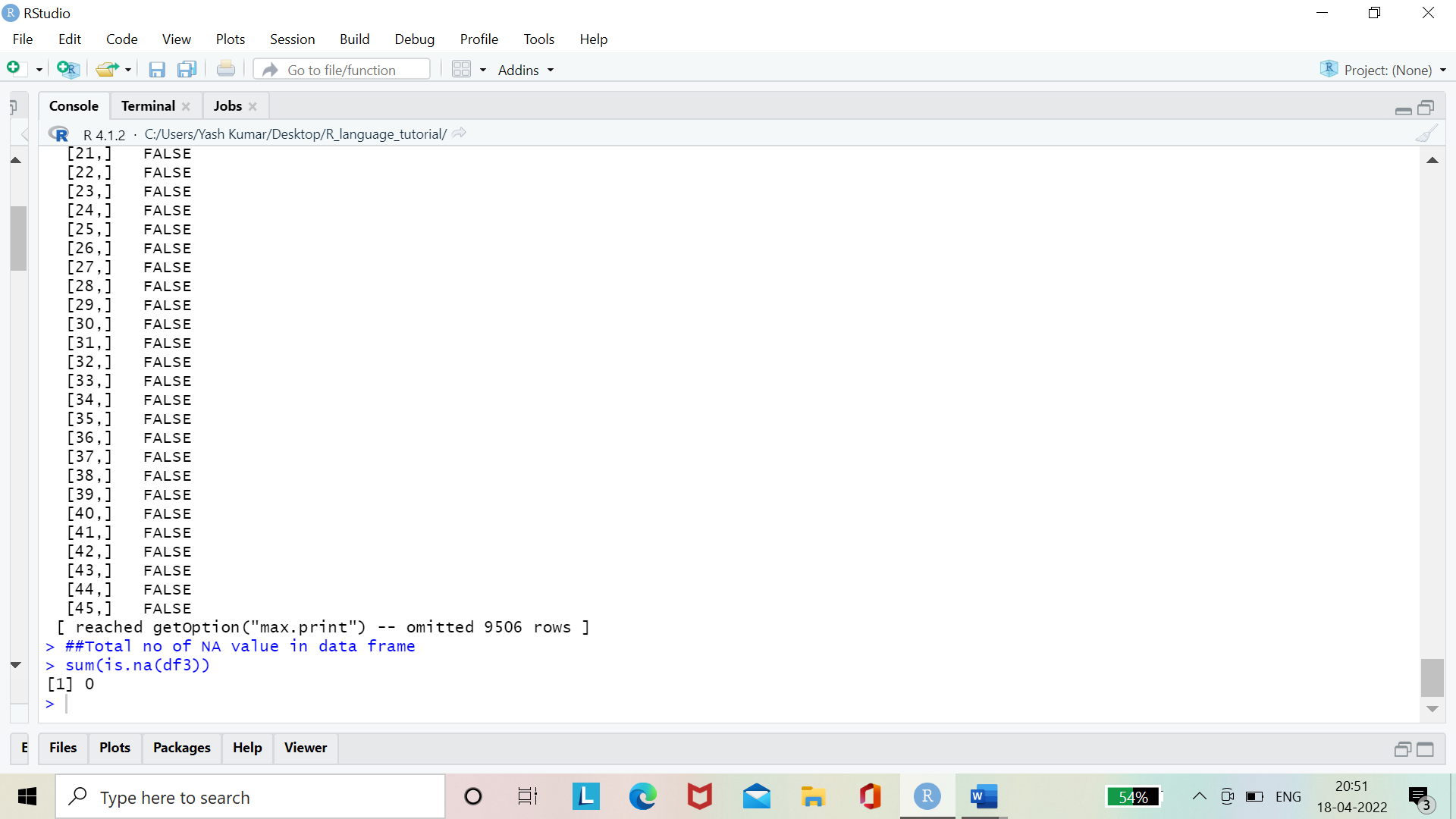


**#Check the NA value in data frame**

is.na(df3)

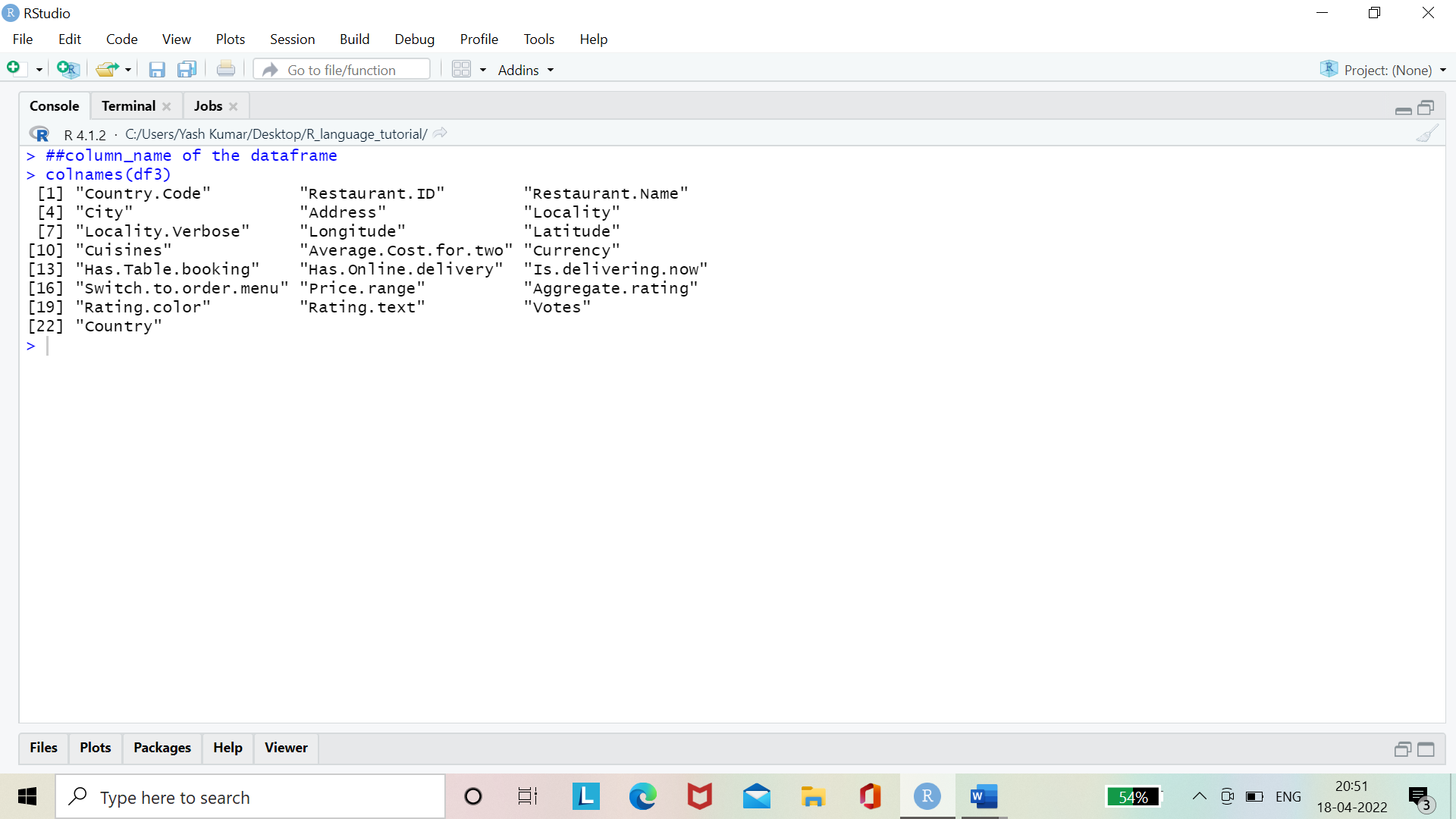
**##Total no of NA value in data frame**

sum(is.na(df3))



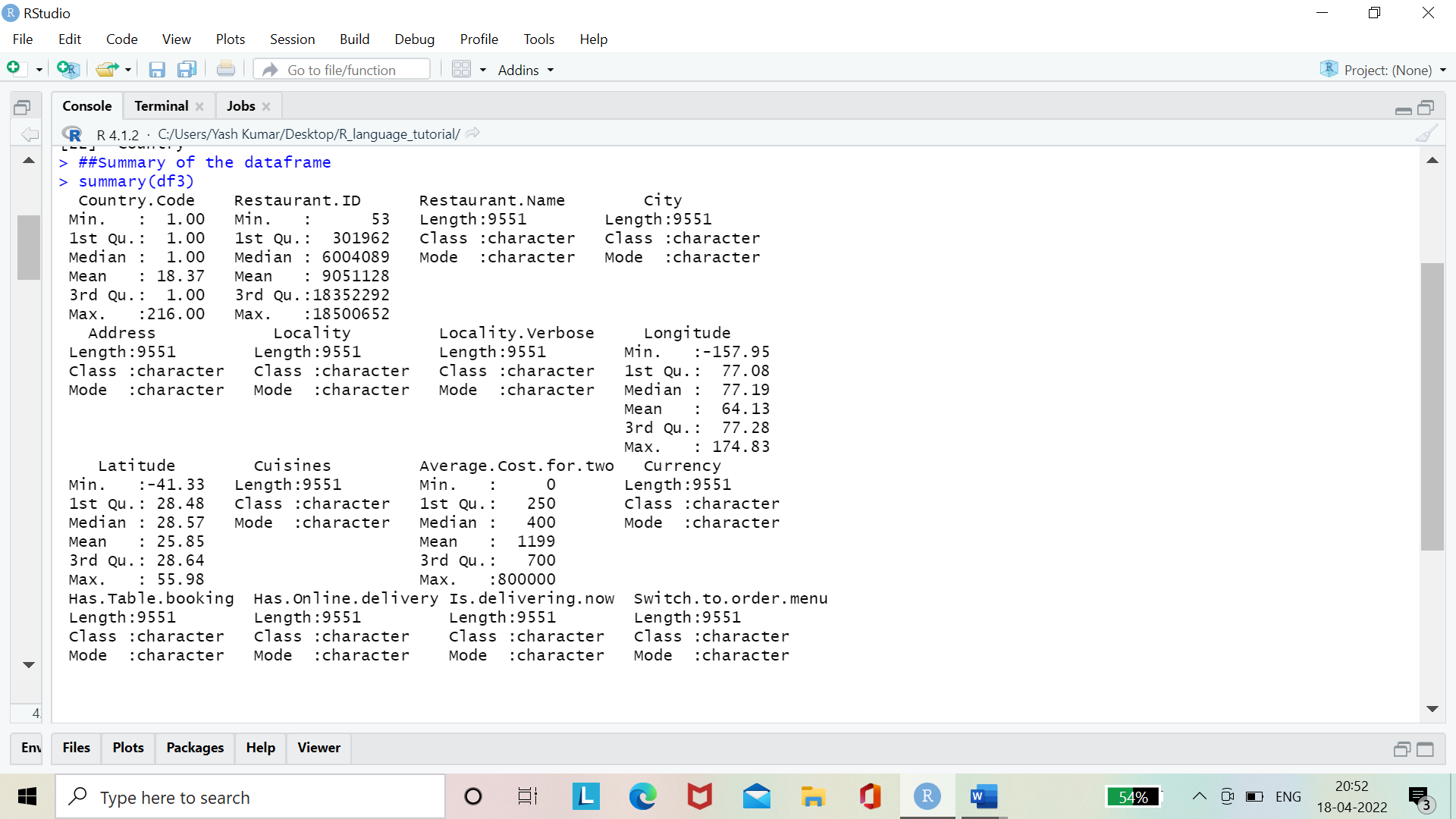
**##Column\_name of the Dataframe**

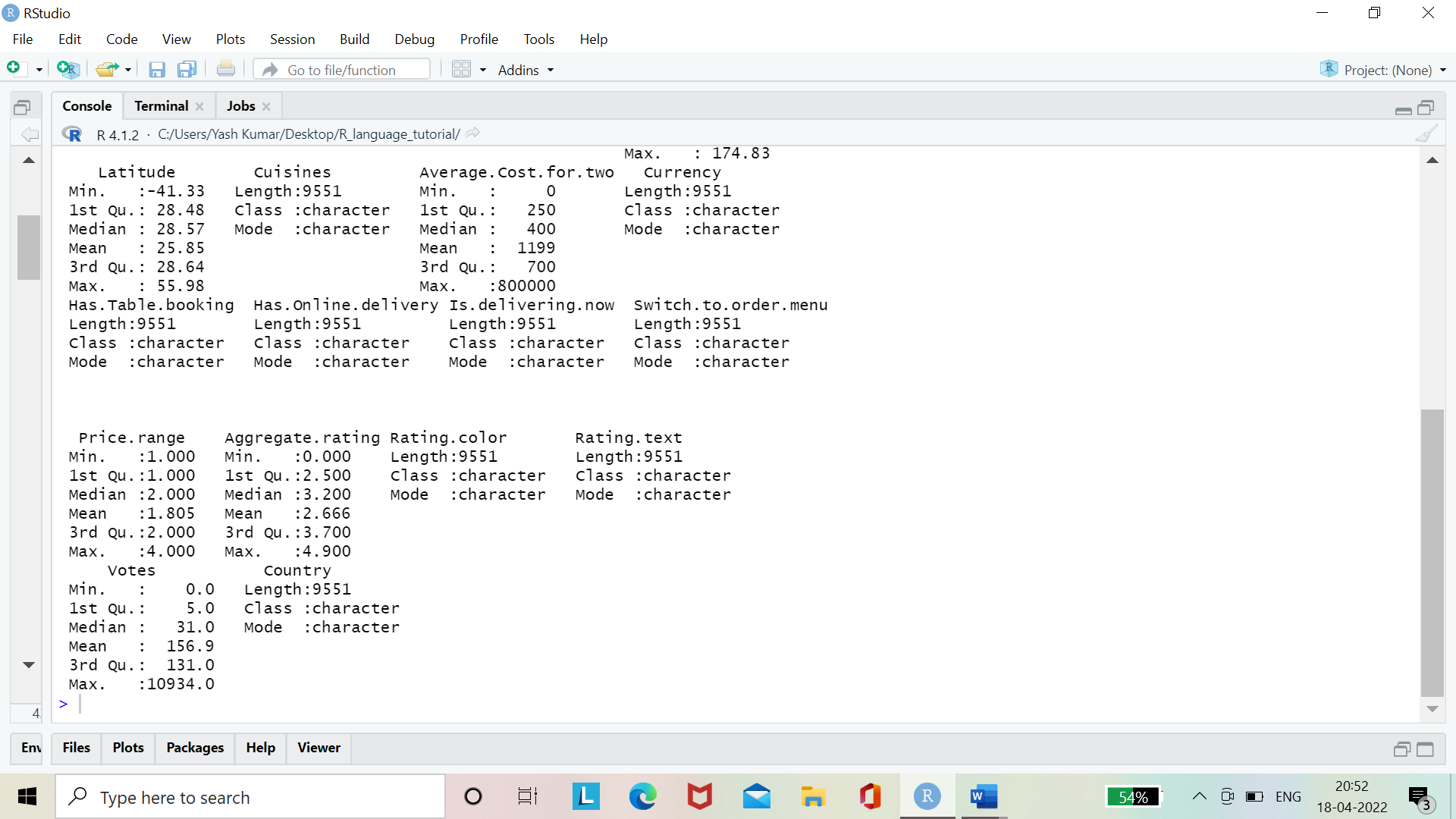
colnames(df3)



**##Summary of the Data frame**

summary(df3)





**##To find the Record with Each Country**

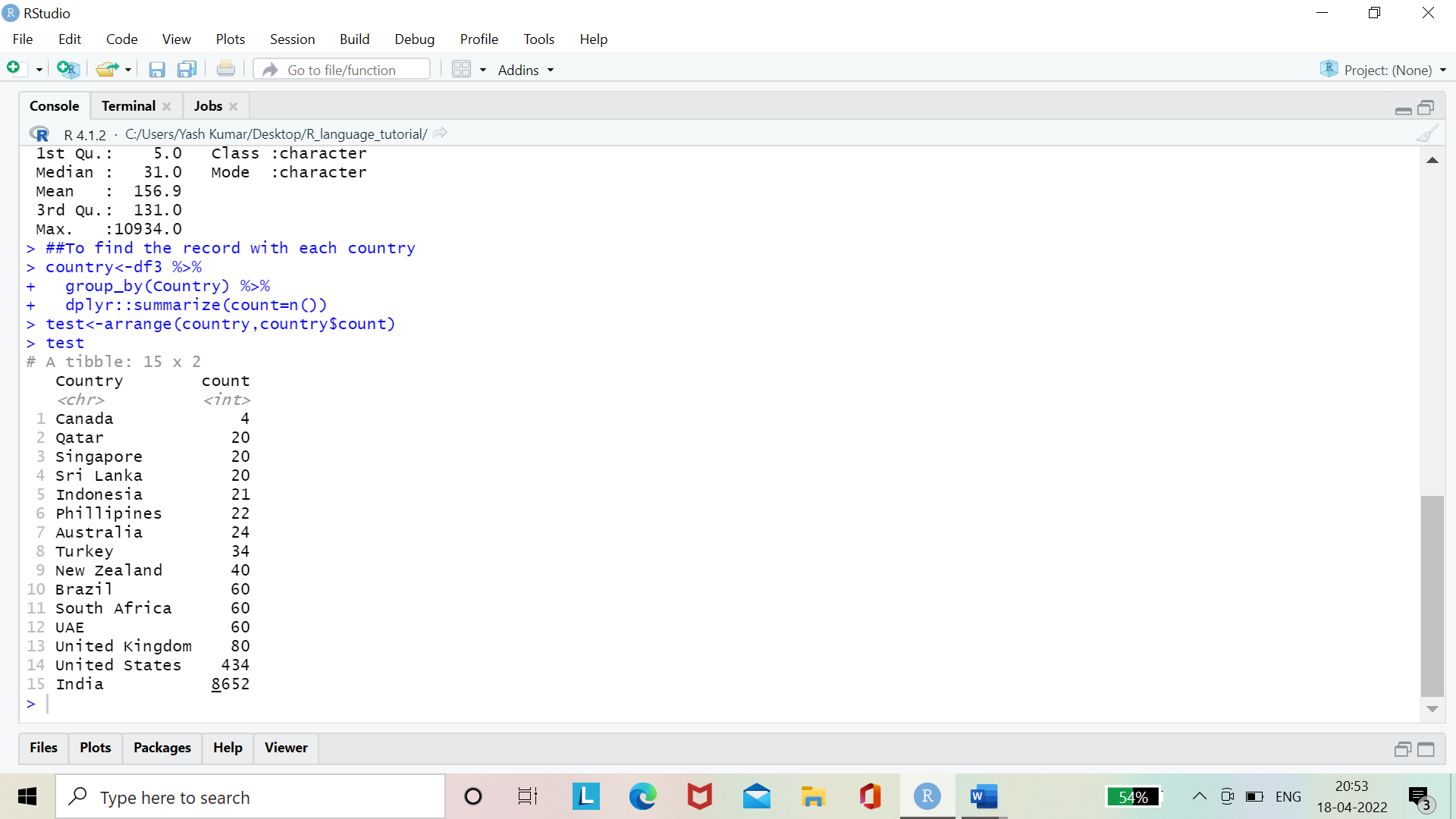
country<-df3 %>%

group\_by(Country) %>%

dplyr::summarize(count=n())

test<-arrange(country,country$count)

test



#Observations:- maximum record are from India and then from USA and then UK

**#Top 5 Indian City with Maximum Number of Zomato Rest**.

india<-filter(df3,df3$Country=="India")

city\_count<-india %>%

group\_by(City) %>%

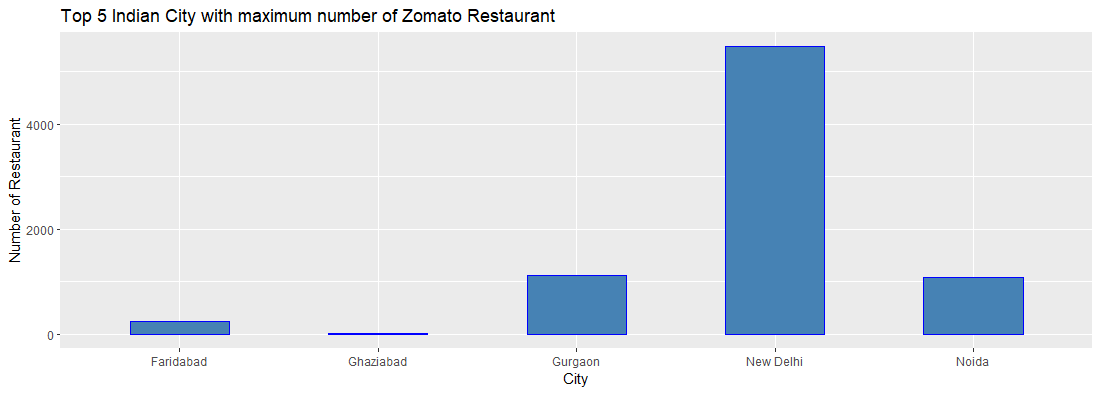
dplyr::summarize(n=n())

five\_city<-tail(arrange(city\_count,city\_count$n),15)

ggplot(data=tail(five\_city,5),aes(x=City,y=n))+

geom\_bar(stat = "identity",position=position\_dodge(), width=0.5, color="blue",fill="steelblue")+

labs(title = "Top 5 Indian City with maximum number of Zomato Restaurant",x="City",y="Number of Restaurant")



#observations:-New Delhi has the maximum number of Restaurant

**##Which Currency is used in which country**

ser<-df3 %>%

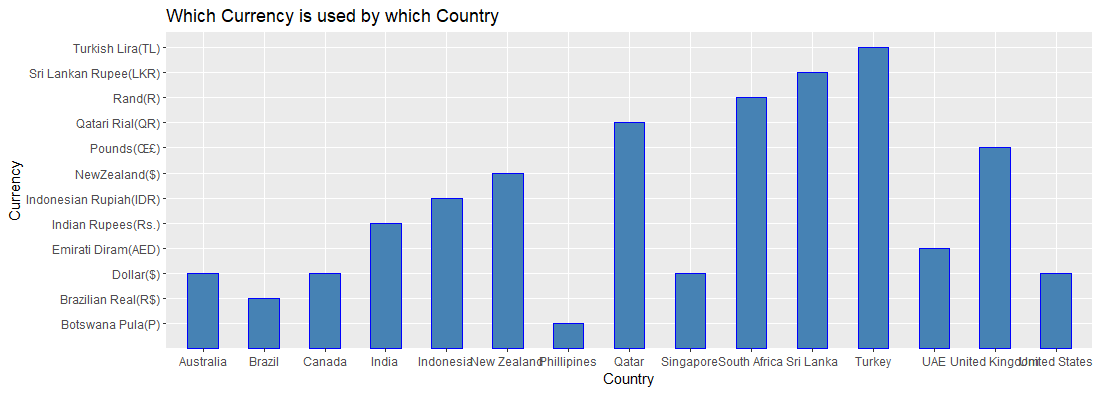
group\_by(Currency,Country) %>%

dplyr::summarize(count=n())

ggplot(data=ser,aes(x=Country,y=Currency))+

geom\_bar(stat = "identity",position=position\_dodge(), width=0.5, color="blue",fill="steelblue")+

labs(title = "Which Currency is used by which Country",x="Country",y="Currency")



**##Which Country offer Online Deliver**

delivery<-df3 %>%

group\_by(Country,Has.Online.delivery) %>%

summarise(count=n())

count1<-which(delivery$Has.Online.delivery=="Yes")

count<-delivery[count1,]

count

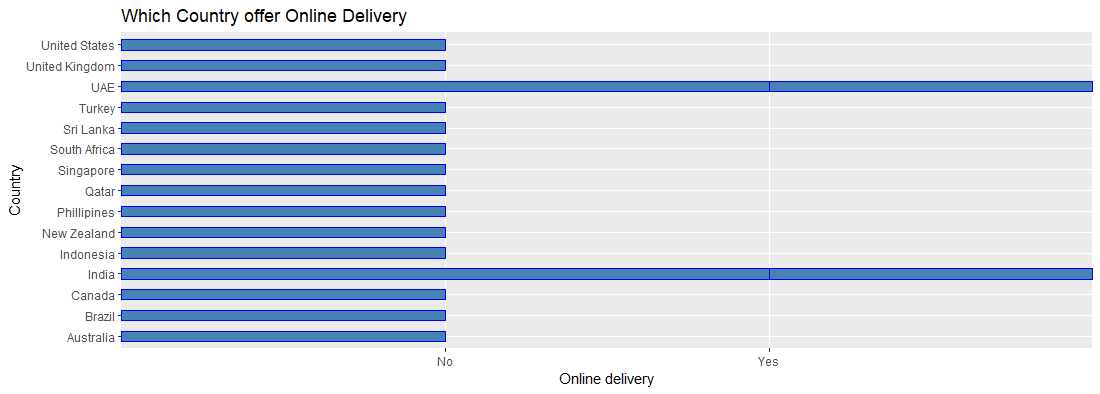
**#Bar Chart Country offer Online Delivry**

p<-ggplot(data=delivery,aes(x=Country,y=Has.Online.delivery))+

geom\_bar(stat = "identity", width=0.5, color="blue",fill="steelblue")+

labs(title = "Which Country offer Online Delivery",x="Country",y="Online delivery")

p+coord\_flip()



#Observations: online delivery is available only in India and UAE .

**##Which Offer Online Deliver Service in India**

deliver<-df3 %>%

filter(Country.Code==1) %>%

unique() %>%

group\_by(City,Has.Online.delivery) %>%

dplyr::summarize(count=n())

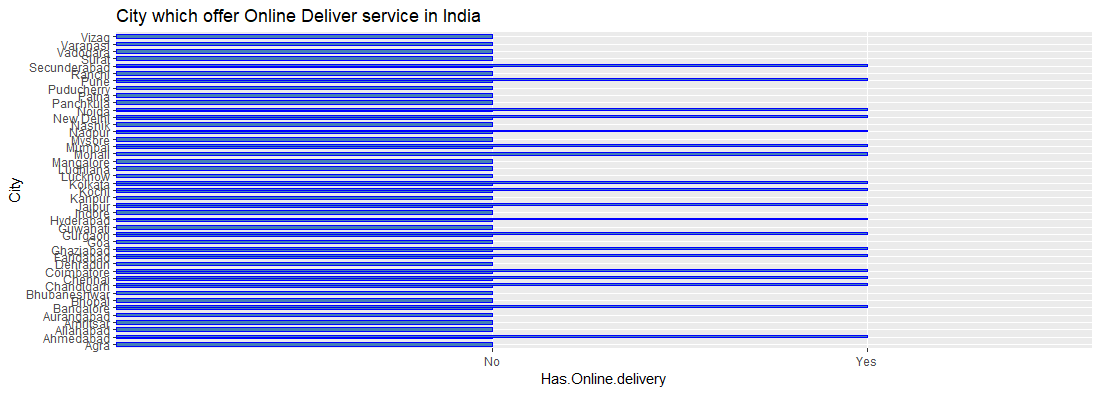
p<-ggplot(data=deliver,aes(x=City,y=Has.Online.delivery))+

geom\_bar(stat = "identity",position=position\_dodge(), width=0.5, color="blue",fill="steelblue")+

labs(title = "City which offer Online Deliver service in India",x="City",y="Has.Online.delivery")

p

p+coord\_flip()



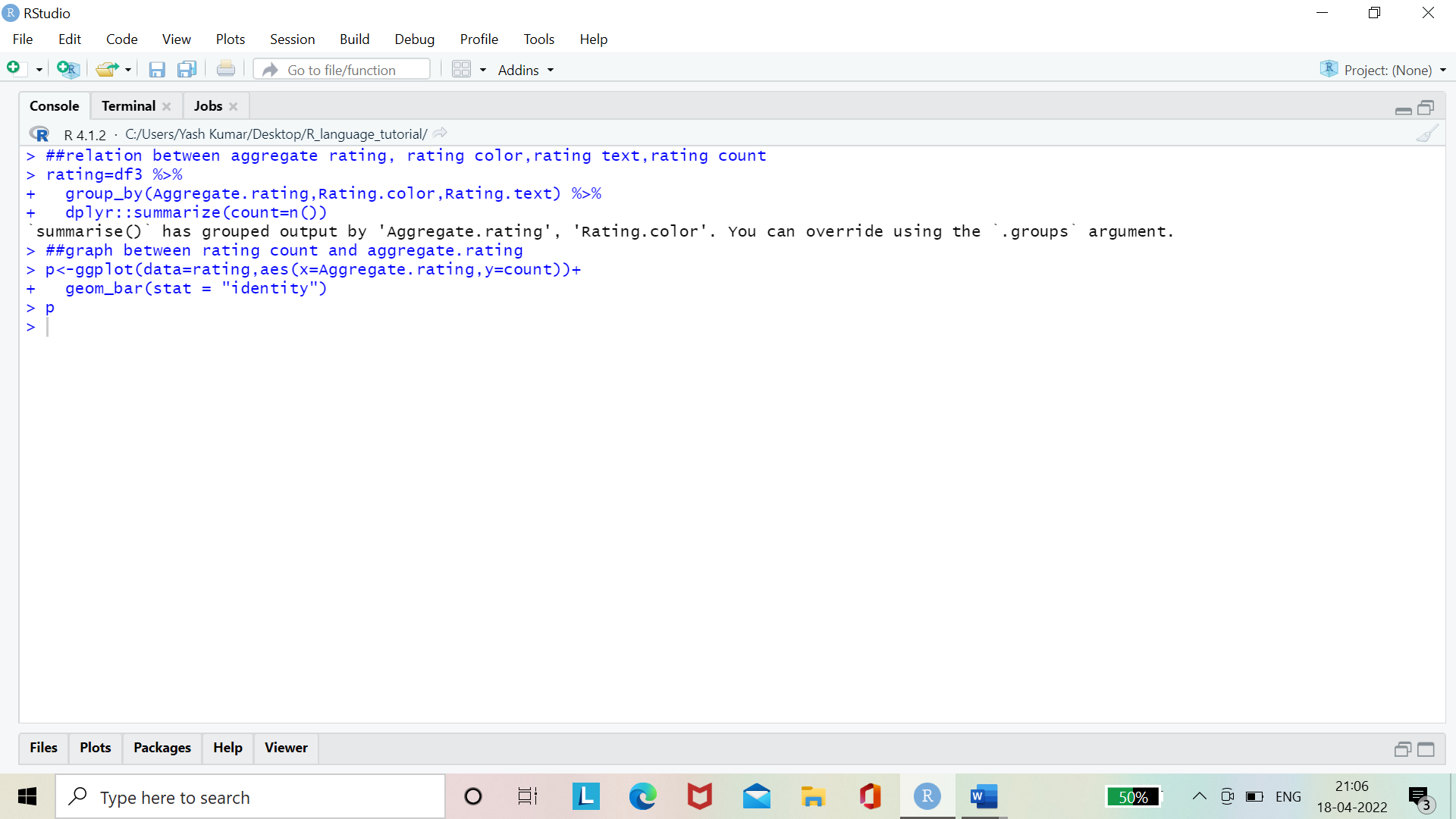
#19 city offer this service

**##Relation Between Aggregate rating, Rating color, Rating Text, Rating Count**

rating=df3 %>%

group\_by(Aggregate.rating,Rating.color,Rating.text) %>%

dplyr::summarize(count=n())

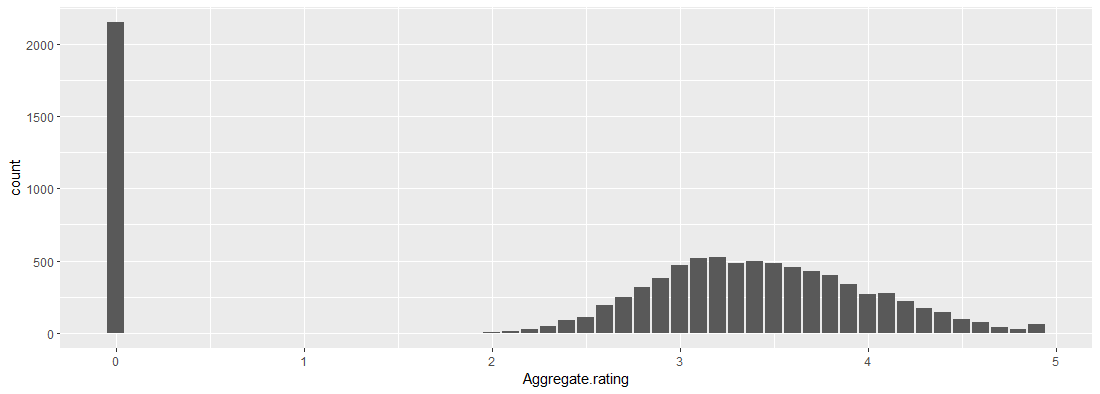


**##Graph Between Rating Count and Aggregate.rating**

p<-ggplot(data=rating,aes(x=Aggregate.rating,y=count))+

geom\_bar(stat = "identity")

p



##1. when rating is between 4.5 to 4.9 -> Excellent

##2. when rating is between 4.0 to 4.4 -> Very Good

##3. when rating is between 3.5 to 3.9 -> Good

##4. when rating is between 2.5 to 3.4 -> Average

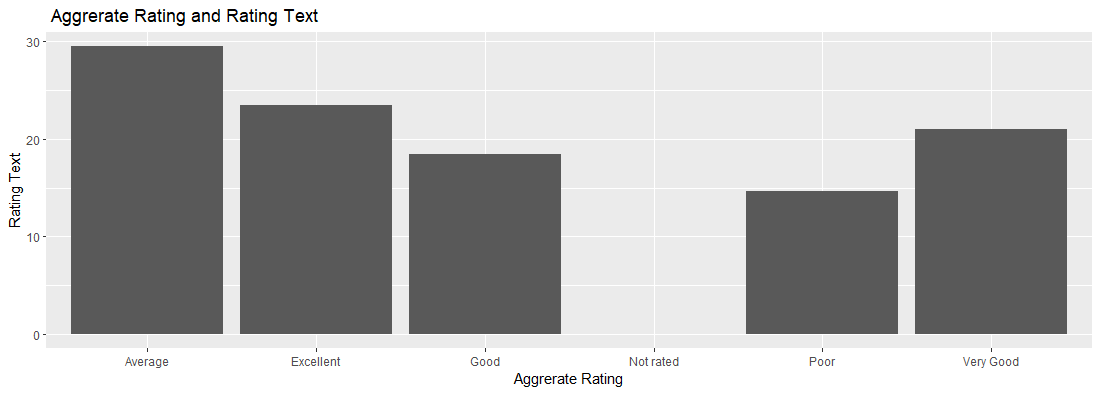
##5. when rating is between 1.0 to 2.4 -> Poor

##graph between

p<-ggplot(data=rating,aes(x=Rating.text,y=Aggregate.rating))+

geom\_bar(stat = "identity")

p



#Observation

#Note: - white is invisible

**##Top 3 Country which has the Cheaper Food**

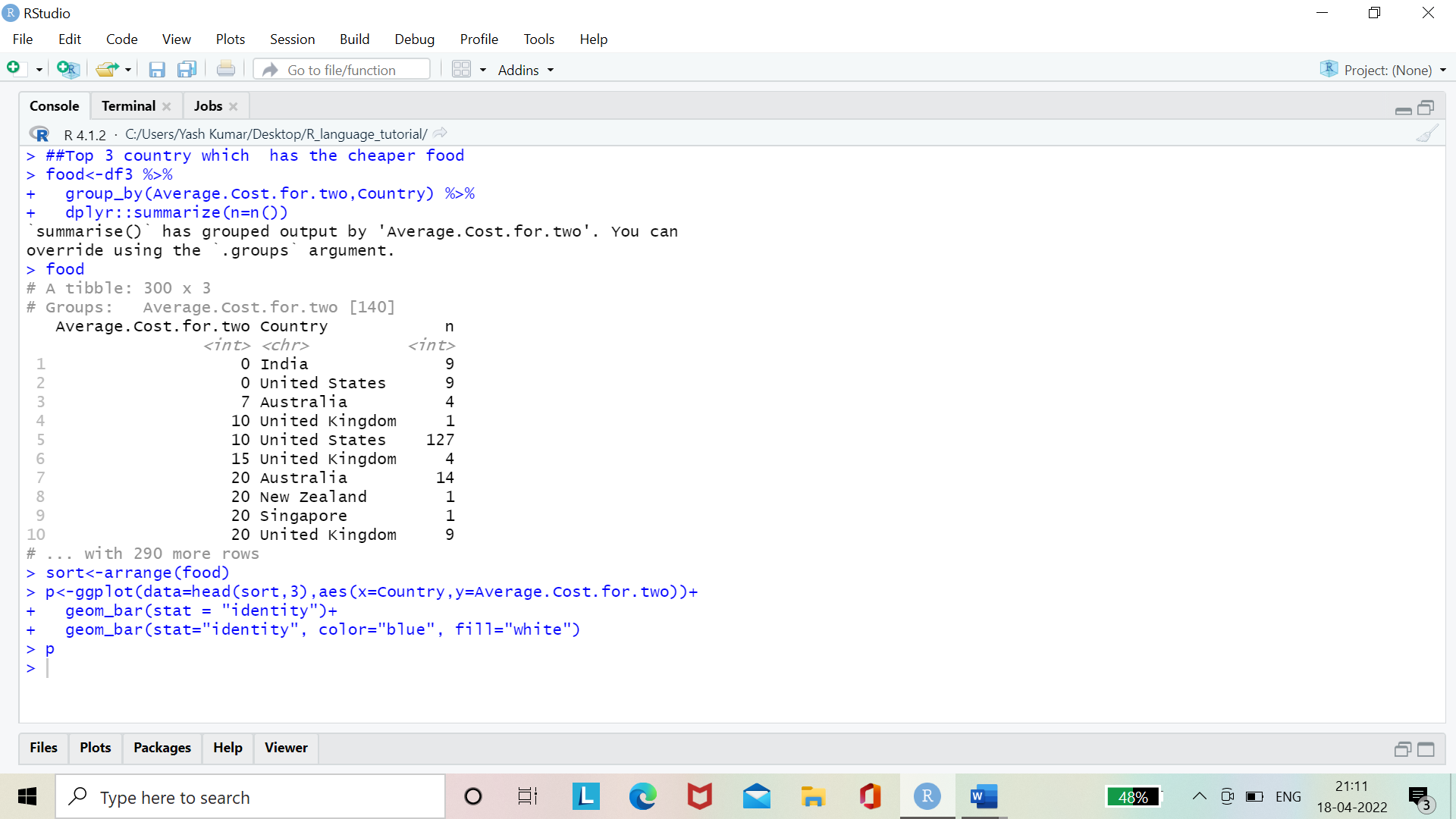
food<-df3 %>%

group\_by(Average.Cost.for.two,Country) %>%

dplyr::summarize(n=n())

food

sort<-arrange(food)



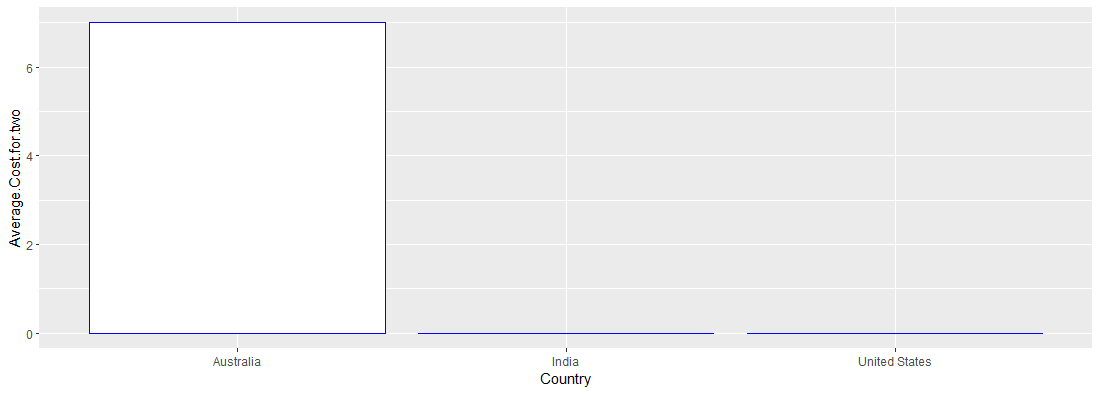
p<-ggplot(data=head(sort,3),aes(x=Country,y=Average.Cost.for.two))+

geom\_bar(stat = "identity")+

geom\_bar(stat="identity", color="blue", fill="white")

p

Graph between the Average.cost.for.two and Country



#Observations:- In india and United States have the cheapest food

**##which Offer Booking Service in India**

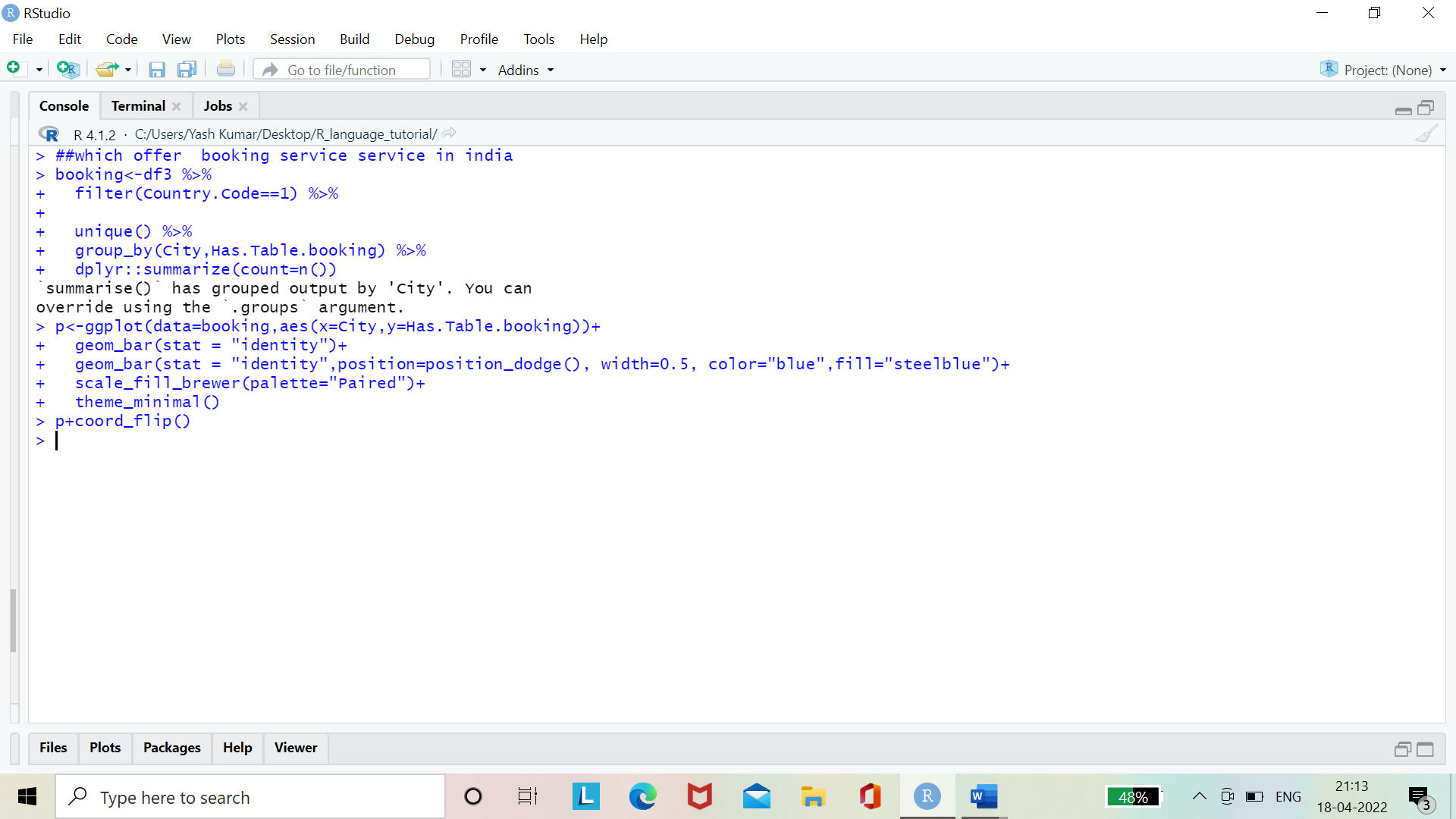
booking<-df3 %>%

filter(Country.Code==1) %>%

unique() %>%

group\_by(City,Has.Table.booking) %>%

dplyr::summarize(count=n())



p<-ggplot(data=booking,aes(x=City,y=Has.Table.booking))+

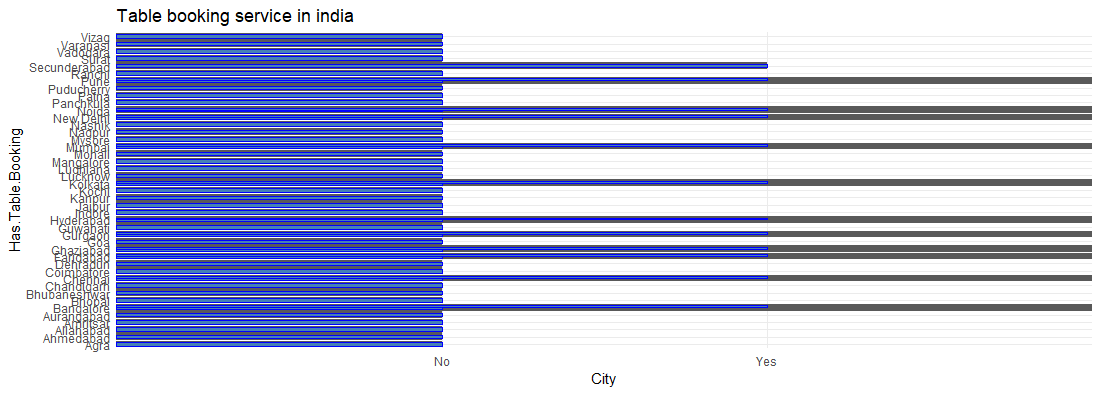
geom\_bar(stat = "identity")+

geom\_bar(stat = "identity",position=position\_dodge(), width=0.5, color="blue",fill="steelblue")+

scale\_fill\_brewer(palette="Paired")+

theme\_minimal()

p+coord\_flip()



#observations:-

# Only 12 coutry offer Table booking service in india