Temperature\_Data\_Cleaning\_Project.R

User

2022-02-28

#Loading the data into R  
iot <- read.table("C:\\Users\\User\\Desktop\\IOT-temp.csv", sep=',', strip.white = T)  
  
#Viewing the loaded data  
head (iot)

## V1 V2 V3 V4 V5  
## 1 id room\_id/id noted\_date temp out/in  
## 2 \_\_export\_\_.temp\_log\_196134\_bd201015 Room Admin 8/12/2018 9:30 29 In  
## 3 \_\_export\_\_.temp\_log\_196131\_7bca51bc Room Admin 8/12/2018 9:30 29 In  
## 4 \_\_export\_\_.temp\_log\_196127\_522915e3 Room Admin 8/12/2018 9:29 41 Out  
## 5 \_\_export\_\_.temp\_log\_196128\_be0919cf Room Admin 8/12/2018 9:29 41 Out  
## 6 \_\_export\_\_.temp\_log\_196126\_d30b72fb Room Admin 8/12/2018 9:29 31 In

#Inspecting the Data  
names(iot)

## [1] "V1" "V2" "V3" "V4" "V5"

str(iot)

## 'data.frame': 97607 obs. of 5 variables:  
## $ V1: chr "id" "\_\_export\_\_.temp\_log\_196134\_bd201015" "\_\_export\_\_.temp\_log\_196131\_7bca51bc" "\_\_export\_\_.temp\_log\_196127\_522915e3" ...  
## $ V2: chr "room\_id/id" "Room Admin" "Room Admin" "Room Admin" ...  
## $ V3: chr "noted\_date" "8/12/2018 9:30" "8/12/2018 9:30" "8/12/2018 9:29" ...  
## $ V4: chr "temp" "29" "29" "41" ...  
## $ V5: chr "out/in" "In" "In" "Out" ...

#Appropriately renaming the five variables  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.1

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

iot <- iot%>% rename(log\_id=V1, room\_id=V2, noted\_date=V3,temp=V4, out\_in=V5)  
  
#Deleting the first row  
iot <- subset(iot,iot$log\_id!='id')  
head(iot)

## log\_id room\_id noted\_date temp out\_in  
## 2 \_\_export\_\_.temp\_log\_196134\_bd201015 Room Admin 8/12/2018 9:30 29 In  
## 3 \_\_export\_\_.temp\_log\_196131\_7bca51bc Room Admin 8/12/2018 9:30 29 In  
## 4 \_\_export\_\_.temp\_log\_196127\_522915e3 Room Admin 8/12/2018 9:29 41 Out  
## 5 \_\_export\_\_.temp\_log\_196128\_be0919cf Room Admin 8/12/2018 9:29 41 Out  
## 6 \_\_export\_\_.temp\_log\_196126\_d30b72fb Room Admin 8/12/2018 9:29 31 In  
## 7 \_\_export\_\_.temp\_log\_196125\_b0fa0b41 Room Admin 8/12/2018 9:29 31 In

#Simplifying the id column and removing the room\_id variable because of its irrelevance to future analysis  
  
iot <- iot %>% mutate(log\_id=gsub('\\\_\_export\_\_.temp\_log\_','', log\_id)) %>% select(-room\_id)  
  
#Separating the noted\_date variable into noted\_date and noted\_time  
  
noted\_time <- str\_sub(iot$noted\_date, start=11, end=18)  
  
iot <- iot %>% mutate(noted\_time)  
  
iot <- iot %>% mutate(noted\_date = str\_sub(noted\_date, start=0, end=10))  
  
  
head (iot)

## log\_id noted\_date temp out\_in noted\_time  
## 2 196134\_bd201015 8/12/2018 29 In 9:30  
## 3 196131\_7bca51bc 8/12/2018 29 In 9:30  
## 4 196127\_522915e3 8/12/2018 41 Out 9:29  
## 5 196128\_be0919cf 8/12/2018 41 Out 9:29  
## 6 196126\_d30b72fb 8/12/2018 31 In 9:29  
## 7 196125\_b0fa0b41 8/12/2018 31 In 9:29

#Rectifying data types and making them suitable for analysis  
  
iot <-iot %>% mutate(out\_in = as.factor(out\_in)) %>% mutate(temp = as.numeric(temp))  
  
str(iot)

## 'data.frame': 97606 obs. of 5 variables:  
## $ log\_id : chr "196134\_bd201015" "196131\_7bca51bc" "196127\_522915e3" "196128\_be0919cf" ...  
## $ noted\_date: chr "8/12/2018 " "8/12/2018 " "8/12/2018 " "8/12/2018 " ...  
## $ temp : num 29 29 41 41 31 31 29 29 29 29 ...  
## $ out\_in : Factor w/ 2 levels "In","Out": 1 1 2 2 1 1 1 1 1 1 ...  
## $ noted\_time: chr "9:30" "9:30" "9:29" "9:29" ...

head (iot)

## log\_id noted\_date temp out\_in noted\_time  
## 2 196134\_bd201015 8/12/2018 29 In 9:30  
## 3 196131\_7bca51bc 8/12/2018 29 In 9:30  
## 4 196127\_522915e3 8/12/2018 41 Out 9:29  
## 5 196128\_be0919cf 8/12/2018 41 Out 9:29  
## 6 196126\_d30b72fb 8/12/2018 31 In 9:29  
## 7 196125\_b0fa0b41 8/12/2018 31 In 9:29

#Removing duplicates  
  
iot <- iot %>% distinct(log\_id, .keep\_all = T)  
  
head (iot)

## log\_id noted\_date temp out\_in noted\_time  
## 1 196134\_bd201015 8/12/2018 29 In 9:30  
## 2 196131\_7bca51bc 8/12/2018 29 In 9:30  
## 3 196127\_522915e3 8/12/2018 41 Out 9:29  
## 4 196128\_be0919cf 8/12/2018 41 Out 9:29  
## 5 196126\_d30b72fb 8/12/2018 31 In 9:29  
## 6 196125\_b0fa0b41 8/12/2018 31 In 9:29

#Checking invalid values  
  
summary (iot)

## log\_id noted\_date temp out\_in   
## Length:97605 Length:97605 Min. :21.00 In :20345   
## Class :character Class :character 1st Qu.:30.00 Out:77260   
## Mode :character Mode :character Median :35.00   
## Mean :35.05   
## 3rd Qu.:40.00   
## Max. :51.00   
## noted\_time   
## Length:97605   
## Class :character   
## Mode :character   
##   
##   
##

#Checking for missing values  
  
sum(is.na(iot))

## [1] 0

apply (iot, 2, function(temp) sum(is.na(temp))/length(temp))

## log\_id noted\_date temp out\_in noted\_time   
## 0 0 0 0 0

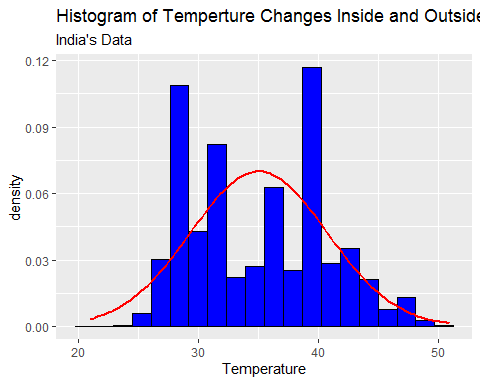
#The iot data is now clean and ready for analysis. Writing it into a transferable .csv file  
  
#Importing the cleaned iot data set for analysis  
  
iotc <- read.table("C:\\Users\\User\\Desktop\\iot.csv", header=T, sep=',',strip.white=T)  
  
head(iotc)

## log\_id noted\_date temp out\_in noted\_time  
## 1 196134\_bd201015 8/12/2018 29 In 9:30  
## 2 196131\_7bca51bc 8/12/2018 29 In 9:30  
## 3 196127\_522915e3 8/12/2018 41 Out 9:29  
## 4 196128\_be0919cf 8/12/2018 41 Out 9:29  
## 5 196126\_d30b72fb 8/12/2018 31 In 9:29  
## 6 196125\_b0fa0b41 8/12/2018 31 In 9:29

#Obtaining the data set's descriptives  
  
summary (iotc)

## log\_id noted\_date temp out\_in   
## Length:97605 Length:97605 Min. :21.00 Length:97605   
## Class :character Class :character 1st Qu.:30.00 Class :character   
## Mode :character Mode :character Median :35.00 Mode :character   
## Mean :35.05   
## 3rd Qu.:40.00   
## Max. :51.00   
## noted\_time   
## Length:97605   
## Class :character   
## Mode :character   
##   
##   
##

#Checking the data set for normality via a histogram of temperature  
  
library(ggplot2)  
  
  
plt <- ggplot(data=iotc, aes(x=temp)) +  
   
 geom\_histogram(aes(y=..density..), bins=20, col='black', fill= 'blue') +   
   
 stat\_function(fun=dnorm, args=list(mean= mean(iotc$temp), sd=sd(iotc$temp)), col='red', lwd=1) +  
   
 labs(title='Histogram of Temperture Changes Inside and Outside a Room Over Time', subtitle="India's Data", x='Temperature')  
  
#Generating a box plot for temperature comparisons in and outside the room  
  
plt\_2 <- ggplot(data=iotc, aes(x= out\_in, y=temp)) +  
   
 geom\_boxplot(varwidth =T, aes(fill= out\_in, alpha=0.5)) +  
   
 labs(title='A Boxplot for Temperature Comparisons In and Outside Rooms in India Over Time', subtitle ='Indian Bureau Data', x='Room-Relative Collecion Point', y= 'Tempmerature', colour='Collecion Point')  
   
#Rendering both plots  
  
par(mfrow=c(2,1))  
  
plt



plt\_2

