**Capestone**

Sougata

2023-01-11

install.packages(“tidyverse”) install.packages(“ggplot2”) install.packages(“dplyr”)

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
library(dplyr)

Jan <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202101-divvy-tripdata.csv")  
Feb <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202102-divvy-tripdata.csv")  
Mar <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202103-divvy-tripdata.csv")  
Apr <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202104-divvy-tripdata.csv")  
May <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202105-divvy-tripdata.csv")  
Jun <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202106-divvy-tripdata.csv")  
Jul <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202107-divvy-tripdata.csv")  
Aug <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202108-divvy-tripdata.csv")  
Sep <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202109-divvy-tripdata.csv")  
Oct <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202110-divvy-tripdata.csv")  
Nov <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202111-divvy-tripdata.csv")  
Dec <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202112-divvy-tripdata.csv")  
Jan22 <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202201-divvy-tripdata.csv")  
Feb22 <- read.csv("C:/Users/Sougata/Desktop/data\_trip/New folder/202202-divvy-tripdata.csv")

triprawdata <- rbind(Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov,Dec)

nrow(triprawdata) *#number of rows*

## [1] 5595063

ncol(triprawdata) *#number of columns*

## [1] 13

head(triprawdata) *#see the first 6 rows of the data frame*

## ride\_id rideable\_type started\_at ended\_at  
## 1 E19E6F1B8D4C42ED electric\_bike 2021-01-23 16:14:19 2021-01-23 16:24:44  
## 2 DC88F20C2C55F27F electric\_bike 2021-01-27 18:43:08 2021-01-27 18:47:12  
## 3 EC45C94683FE3F27 electric\_bike 2021-01-21 22:35:54 2021-01-21 22:37:14  
## 4 4FA453A75AE377DB electric\_bike 2021-01-07 13:31:13 2021-01-07 13:42:55  
## 5 BE5E8EB4E7263A0B electric\_bike 2021-01-23 02:24:02 2021-01-23 02:24:45  
## 6 5D8969F88C773979 electric\_bike 2021-01-09 14:24:07 2021-01-09 15:17:54  
## start\_station\_name start\_station\_id end\_station\_name end\_station\_id  
## 1 California Ave & Cortez St 17660   
## 2 California Ave & Cortez St 17660   
## 3 California Ave & Cortez St 17660   
## 4 California Ave & Cortez St 17660   
## 5 California Ave & Cortez St 17660   
## 6 California Ave & Cortez St 17660   
## start\_lat start\_lng end\_lat end\_lng member\_casual  
## 1 41.90034 -87.69674 41.89 -87.72 member  
## 2 41.90033 -87.69671 41.90 -87.69 member  
## 3 41.90031 -87.69664 41.90 -87.70 member  
## 4 41.90040 -87.69666 41.92 -87.69 member  
## 5 41.90033 -87.69670 41.90 -87.70 casual  
## 6 41.90041 -87.69676 41.94 -87.71 casual

tail(triprawdata) *#see the last 6 rows of the data frame*

## ride\_id rideable\_type started\_at ended\_at  
## 5595058 92BBAB97D1683D69 electric\_bike 2021-12-24 15:42:09 2021-12-24 19:29:35  
## 5595059 847431F3D5353AB7 electric\_bike 2021-12-12 13:36:55 2021-12-12 13:56:08  
## 5595060 CF407BBC3B9FAD63 electric\_bike 2021-12-06 19:37:50 2021-12-06 19:44:51  
## 5595061 60BB69EBF5440E92 electric\_bike 2021-12-02 08:57:04 2021-12-02 09:05:21  
## 5595062 C414F654A28635B8 electric\_bike 2021-12-13 09:00:26 2021-12-13 09:14:39  
## 5595063 37AC57E34B2E7E97 classic\_bike 2021-12-13 08:45:32 2021-12-13 08:49:09  
## start\_station\_name start\_station\_id end\_station\_name  
## 5595058 Canal St & Madison St 13341   
## 5595059 Canal St & Madison St 13341   
## 5595060 Canal St & Madison St 13341 Kingsbury St & Kinzie St  
## 5595061 Canal St & Madison St 13341 Dearborn St & Monroe St  
## 5595062 Lawndale Ave & 16th St 362.0   
## 5595063 Michigan Ave & Jackson Blvd TA1309000002 Dearborn St & Monroe St  
## end\_station\_id start\_lat start\_lng end\_lat end\_lng member\_casual  
## 5595058 41.88180 -87.63997 41.88000 -87.64000 casual  
## 5595059 41.88229 -87.63975 41.89000 -87.61000 casual  
## 5595060 KA1503000043 41.88212 -87.64005 41.88911 -87.63886 member  
## 5595061 TA1305000006 41.88196 -87.63995 41.88025 -87.62960 member  
## 5595062 41.86000 -87.72000 41.85000 -87.71000 member  
## 5595063 TA1305000006 41.87785 -87.62408 41.88132 -87.62952 member

str(triprawdata) *#see list of columns and data types*

## 'data.frame': 5595063 obs. of 13 variables:  
## $ ride\_id : chr "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...  
## $ rideable\_type : chr "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : chr "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 22:35:54" "2021-01-07 13:31:13" ...  
## $ ended\_at : chr "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-01-07 13:42:55" ...  
## $ start\_station\_name: chr "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" ...  
## $ start\_station\_id : chr "17660" "17660" "17660" "17660" ...  
## $ end\_station\_name : chr "" "" "" "" ...  
## $ end\_station\_id : chr "" "" "" "" ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.7 -87.7 -87.7 -87.7 -87.7 ...  
## $ end\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num -87.7 -87.7 -87.7 -87.7 -87.7 ...  
## $ member\_casual : chr "member" "member" "member" "member" ...

summary(triprawdata) *#statistical summary of data*

## ride\_id rideable\_type started\_at ended\_at   
## Length:5595063 Length:5595063 Length:5595063 Length:5595063   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## start\_station\_name start\_station\_id end\_station\_name end\_station\_id   
## Length:5595063 Length:5595063 Length:5595063 Length:5595063   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## start\_lat start\_lng end\_lat end\_lng   
## Min. :41.64 Min. :-87.84 Min. :41.39 Min. :-88.97   
## 1st Qu.:41.88 1st Qu.:-87.66 1st Qu.:41.88 1st Qu.:-87.66   
## Median :41.90 Median :-87.64 Median :41.90 Median :-87.64   
## Mean :41.90 Mean :-87.65 Mean :41.90 Mean :-87.65   
## 3rd Qu.:41.93 3rd Qu.:-87.63 3rd Qu.:41.93 3rd Qu.:-87.63   
## Max. :42.07 Max. :-87.52 Max. :42.17 Max. :-87.49   
## NA's :4771 NA's :4771   
## member\_casual   
## Length:5595063   
## Class :character   
## Mode :character   
##   
##   
##   
##

colnames(triprawdata) *#list of column names*

## [1] "ride\_id" "rideable\_type" "started\_at"   
## [4] "ended\_at" "start\_station\_name" "start\_station\_id"   
## [7] "end\_station\_name" "end\_station\_id" "start\_lat"   
## [10] "start\_lng" "end\_lat" "end\_lng"   
## [13] "member\_casual"

#find the na data and remove it

sum(is.na(triprawdata))

## [1] 9542

triprawdata<- na.omit(triprawdata)

triprawdata$date <- as.Date(triprawdata$started\_at) *#The defalt format is yyyy-mm-dd*  
 triprawdata$month <- format(as.Date(triprawdata$date), "%b")  
 triprawdata$day <- format(as.Date(triprawdata$date), "%d")  
 triprawdata$year <- format(as.Date(triprawdata$date),"%Y")  
 triprawdata$day\_of\_week <- format(as.Date(triprawdata$date),"%A")  
 triprawdata$ride\_length <- difftime(triprawdata$ended\_at,triprawdata$started\_at)  
 glimpse(triprawdata)

## Rows: 5,590,292  
## Columns: 19  
## $ ride\_id <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683…  
## $ rideable\_type <chr> "electric\_bike", "electric\_bike", "electric\_bike", …  
## $ started\_at <chr> "2021-01-23 16:14:19", "2021-01-27 18:43:08", "2021…  
## $ ended\_at <chr> "2021-01-23 16:24:44", "2021-01-27 18:47:12", "2021…  
## $ start\_station\_name <chr> "California Ave & Cortez St", "California Ave & Cor…  
## $ start\_station\_id <chr> "17660", "17660", "17660", "17660", "17660", "17660…  
## $ end\_station\_name <chr> "", "", "", "", "", "", "", "", "", "Wood St & Augu…  
## $ end\_station\_id <chr> "", "", "", "", "", "", "", "", "", "657", "13258",…  
## $ start\_lat <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4…  
## $ start\_lng <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696…  
## $ end\_lat <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4…  
## $ end\_lng <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700…  
## $ member\_casual <chr> "member", "member", "member", "member", "casual", "…  
## $ date <date> 2021-01-23, 2021-01-27, 2021-01-21, 2021-01-07, 20…  
## $ month <chr> "Jan", "Jan", "Jan", "Jan", "Jan", "Jan", "Jan", "J…  
## $ day <chr> "23", "27", "21", "07", "23", "09", "04", "14", "09…  
## $ year <chr> "2021", "2021", "2021", "2021", "2021", "2021", "20…  
## $ day\_of\_week <chr> "Saturday", "Wednesday", "Thursday", "Thursday", "S…  
## $ ride\_length <drtn> 625 secs, 244 secs, 80 secs, 702 secs, 43 secs, 32…

triprawdata <- triprawdata %>%  
 select(-c(start\_lat,start\_lng,end\_lat,end\_lng)) *#remove the unnecessary columns*

install.packages(“magrittr”) # package installations are only needed the first time you use it install.packages(“dplyr”) # alternative installation of the %>%

library(magrittr) *# needs to be run every time you start R and want to use %>%*

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(dplyr) *# alternatively, this also loads %>%*  
 summary(triprawdata)

## ride\_id rideable\_type started\_at ended\_at   
## Length:5590292 Length:5590292 Length:5590292 Length:5590292   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## start\_station\_name start\_station\_id end\_station\_name end\_station\_id   
## Length:5590292 Length:5590292 Length:5590292 Length:5590292   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## member\_casual date month day   
## Length:5590292 Min. :2021-01-01 Length:5590292 Length:5590292   
## Class :character 1st Qu.:2021-06-07 Class :character Class :character   
## Mode :character Median :2021-08-01 Mode :character Mode :character   
## Mean :2021-07-28   
## 3rd Qu.:2021-09-24   
## Max. :2021-12-31   
## year day\_of\_week ride\_length   
## Length:5590292 Length:5590292 Length:5590292   
## Class :character Class :character Class :difftime   
## Mode :character Mode :character Mode :numeric   
##   
##   
##

View(triprawdata)

triprawdata$ride\_length <- as.numeric(as.character(triprawdata$ride\_length))  
 is.numeric(triprawdata$ride\_length) *#This code was for checking if the data was converted correctly.*

## [1] TRUE

triprawdata\_v2 <- triprawdata  
View(triprawdata\_v2)  
mean(triprawdata\_v2$ride\_length)

## [1] 1258.87

median(triprawdata\_v2$ride\_length)

## [1] 719

max(triprawdata\_v2$ride\_length)

## [1] 3356649

min(triprawdata\_v2$ride\_length)

## [1] -3482

aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual, FUN = mean)

## triprawdata\_v2$member\_casual triprawdata\_v2$ride\_length  
## 1 casual 1814.2083  
## 2 member 801.2464

aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual, FUN = median)

## triprawdata\_v2$member\_casual triprawdata\_v2$ride\_length  
## 1 casual 957  
## 2 member 576

aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual, FUN = max)

## triprawdata\_v2$member\_casual triprawdata\_v2$ride\_length  
## 1 casual 3356649  
## 2 member 89996

aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual, FUN = min)

## triprawdata\_v2$member\_casual triprawdata\_v2$ride\_length  
## 1 casual -3482  
## 2 member -3245

aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual + triprawdata\_v2$day\_of\_week, FUN = mean)

## triprawdata\_v2$member\_casual triprawdata\_v2$day\_of\_week  
## 1 casual Friday  
## 2 member Friday  
## 3 casual Monday  
## 4 member Monday  
## 5 casual Saturday  
## 6 member Saturday  
## 7 casual Sunday  
## 8 member Sunday  
## 9 casual Thursday  
## 10 member Thursday  
## 11 casual Tuesday  
## 12 member Tuesday  
## 13 casual Wednesday  
## 14 member Wednesday  
## triprawdata\_v2$ride\_length  
## 1 1716.5180  
## 2 783.1153  
## 3 1817.0877  
## 4 776.4844  
## 5 1966.2261  
## 6 895.4045  
## 7 2119.6339  
## 8 914.9310  
## 9 1562.9715  
## 10 751.6393  
## 11 1602.7887  
## 12 754.2399  
## 13 1573.6698  
## 14 757.5636

triprawdata\_v2$day\_of\_week <- ordered(triprawdata\_v2$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))  
  
aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual + triprawdata\_v2$day\_of\_week, FUN = mean)

## triprawdata\_v2$member\_casual triprawdata\_v2$day\_of\_week  
## 1 casual Sunday  
## 2 member Sunday  
## 3 casual Monday  
## 4 member Monday  
## 5 casual Tuesday  
## 6 member Tuesday  
## 7 casual Wednesday  
## 8 member Wednesday  
## 9 casual Thursday  
## 10 member Thursday  
## 11 casual Friday  
## 12 member Friday  
## 13 casual Saturday  
## 14 member Saturday  
## triprawdata\_v2$ride\_length  
## 1 2119.6339  
## 2 914.9310  
## 3 1817.0877  
## 4 776.4844  
## 5 1602.7887  
## 6 754.2399  
## 7 1573.6698  
## 8 757.5636  
## 9 1562.9715  
## 10 751.6393  
## 11 1716.5180  
## 12 783.1153  
## 13 1966.2261  
## 14 895.4045

}  
triprawdata\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>% #creates weekday field using wday()  
 group\_by(member\_casual, weekday) %>% #groups by usertype and weekday  
 summarise(number\_of\_rides = n(), average\_duration = mean(ride\_length)) %>% # calculates the average duration  
 arrange(member\_casual, weekday)  
  
counts <- aggregate(triprawdata\_v2$ride\_length ~ triprawdata\_v2$member\_casual + triprawdata\_v2$day\_of\_week, FUN = mean)

***##Ride type Vs. Number of trips***  
triprawdata\_v2 %>%  
 group\_by(rideable\_type, member\_casual) %>%  
 summarise(number\_of\_rides = n()) %>%   
 ggplot(aes(x= rideable\_type, y=number\_of\_rides, fill= member\_casual))+  
 geom\_bar(stat='identity') +  
 scale\_y\_continuous(labels = **function**(x) format(x, scientific = FALSE)) +  
 labs(title ="Ride type Vs. Number of trips")

## `summarise()` has grouped output by 'rideable\_type'. You can override using the  
## `.groups` argument.



***##Average trip duration by customer type Vs. rideable\_type***  
 triprawdata\_v2 %>%   
 group\_by(member\_casual, rideable\_type) %>%   
 summarise(number\_of\_rides = n()  
 ,average\_duration =mean(ride\_length)) %>%   
 arrange(member\_casual, rideable\_type) %>%   
 ggplot(aes(x = rideable\_type, y = average\_duration, fill = member\_casual)) +  
 geom\_col(position = "dodge")

## `summarise()` has grouped output by 'member\_casual'. You can override using the  
## `.groups` argument.



labs(title ="Average trip duration by customer type Vs. rideable\_type")

## $title  
## [1] "Average trip duration by customer type Vs. rideable\_type"  
##   
## attr(,"class")  
## [1] "labels"

*#total trips by customer type vs month*  
 triprawdata\_v2 %>%   
 group\_by(member\_casual,month) %>%   
 summarise(number\_of\_rides = n()  
 ) %>%   
 arrange(member\_casual, month) %>%   
 ggplot(aes(x = month, y = number\_of\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge")

## `summarise()` has grouped output by 'member\_casual'. You can override using the  
## `.groups` argument.



labs(title ="total trips by customer type vs month")

## $title  
## [1] "total trips by customer type vs month"  
##   
## attr(,"class")  
## [1] "labels"