```
## Start by installing accessing important packages in R

install.packages("tidyverse")
library(tidyverse)
install.packages("lubridate")
library(lubridate)
install.packages("ggplot2")
library(ggplot2)
install.packages("janitor")
library(janitor)
install.packages("skimr")
library(skimr)
library(dplyr)

## Import case study files

Trips_Nov21<- read_csv("202111-divvy-tripdata.csv")
Trips_OCT21<- read_csv("202110-divvy-tripdata.csv")
Trips_SEP21<- read_csv("202109-divvy-tripdata.csv")
Trips_AUG21<- read_csv("202108-divvy-tripdata.csv")
Trips_JUL21<- read_csv("202107-divvy-tripdata.csv")
Trips_JUN21<- read_csv("202106-divvy-tripdata.csv")
Trips_MAY21<- read_csv("202105-divvy-tripdata.csv")
Trips_APR21<- read_csv("202104-divvy-tripdata.csv")
Trips_MAR21<- read_csv("202103-divvy-tripdata.csv")
Trips_FEB21<- read_csv("202102-divvy-tripdata.csv")
Trips_JAN21<- read_csv("202101-divvy-tripdata.csv")
Trips_DEC20<- read_csv("202012-divvy-tripdata.csv")

##Check file structures
```

```
str(Trips_Nov21)


head(Trips_Nov21)
str(Trips_Nov21)
glimpse(Trips_Nov21)


## I checked for file structures for each file
str(Trips_Nov21)
str(Trips_OCT21)
str(Trips_SEP21)
str(Trips_AUG21)
str(Trips_JUL21)
str(Trips_JUN21)
str(Trips_MAY21)
str(Trips_APR21)
str(Trips_MAR21)
str(Trips_FEB21)
str(Trips_JAN21)
str(Trips_DEC20)

## As an extra measure I compared column datatype across all files to check for inconsistencies
compare_df_cols(Trips_Nov21,Trips_OCT21,Trips_SEP21,Trips_AUG21,Trips_JUL21,Trips_JUN21,Trips_MAY21,Trips_APR21,Trips_MAR21,Trips_FEB21,Trips_JAN21,Trips_DEC20, return = "mismatch")


## I then joined all csv files into 1 big file for my data analysis work
all_trips <-
bind_rows(Trips_Nov21,Trips_OCT21,Trips_SEP21,Trips_AUG21,Trips_JUL21,Trips_JUN21,Trips_MAY21,Trips_APR21,Trips_MAR21,Trips_FEB21,Trips_JAN21,Trips_DEC20)
```

```
##I went on to delete some of the columns which I will not be using
all_trips<- all_trips%>%
  select(-c(start_lat, start_lng, end_lat, end_lng))

##I added columns in the all trip file to get the ride time and weekday for each ride
all_trips$day_of_week<-format(as.Date(all_trips$date), "%A")
all_trips$ride_length<-difftime(all_trips$ended_at,all_trips$started_at)

##I then renamed columns for ease of use
all_trips<- all_trips%>%
  rename(
    trip_id=ride_id,
    ride_type=rideable_type,
    start_time=started_at,
    end_time=ended_at,
    usertype=member_casual)
  )

## I rechecked if the structure is still correct
str(all_trips)

##I then tried to get the year, month, date and day from date fields
all_trips$dated<- as.Date(all_trips$start_time)
all_trips$month<- format(as.Date(all_trips$dated), "%m")
all_trips$date<- format(as.Date(all_trips$dated), "%d")
all_trips$year<- format(as.Date(all_trips$dated), "%Y")
all_trips$day<- format(as.Date(all_trips$dated), "%A")

##I tried to get the duration of ride
all_trips$ride_length<- difftime(all_trips$end_time,all_trips$start_time)
```

```
is.factor(all_trips$ride_length)

all_trips$ride_length<- as.numeric(as.character(all_trips$ride_length))

is.numeric(all_trips$ride_length)


##When performing data analysis, I noticed that there are some rows where data was negative. So I deleted the data

all_trips_clean<- all_trips[!(all_trips$ride_length<0),]


## Lastly I got various stats from the ride_length data

summary(all_trips_clean$ride_length)


##Finally I exported the clean file via csv

write.csv(all_trips_clean, "Cyclistic.csv")
```