# Capstone code

````
---
output:
  pdf_document: default
  html_document: default
---
--
title: "Capstone"
author: "Akintunde Akinsanmi"
date: "2023-03-02"
output:
  pdf_document:
    latex_engine: xelatex
---
#JUNE
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r install knitr}
install.packages("knitr")
```

```{r install packages lubridate, include=FALSE}
install.packages("lubridate")
```

```{r install packages tidyverse, include=FALSE}
install.packages("tidyverse")
```

```{r install packages dplyr, include=FALSE}
install.packages("dplyr")
```

```{r install packages here, include=FALSE}
install.packages("here")
```

```{r install packages skimr}
install.packages("skimr")
````
````

```

```{r install packages janitor}
install.packages("janitor")
```

```{r load packages}
library(lubridate)
library(tidyverse)
library(here)
library(dplyr)
library(skimr)
library(janitor)
library(knitr)
```

```{r}
# Load the data into R
june <- read.csv("Files/202006-divvy-tripdata.csv",stringsAsFactors =
FALSE)
```

```{r}
# Explore the data
str(june)
summary(june)
head(june)
tail(june)
```
```{r}
# Clean the data
# Remove rows with missing data
june <- na.omit(june)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(june)
head(june)
tail(june)

```
```

```{r}
# Check for missing values in the "column_name" column
sum(is.na(june$rideable_type))
sum(is.na(june$started_at))
sum(is.na(june$ended_at))
sum(is.na(june$start_station_name))
sum(is.na(june$end_station_name))
sum(is.na(june$start_station_id))
sum(is.na(june$start_lat))
sum(is.na(june$start_lng))
sum(is.na(june$end_lat))
sum(is.na(june$end_lng))
sum(is.na(june$member_casual))
```

```{r (june) combining the co-ordinates to get the start locations}
june$start_location <- paste(june$start_lat, june$start_lng, sep = ",")
```

```{r (june2)combining the c0-0rdinates to get the end locations}
june$end_location <- paste(june$end_lat, june$end_lng, sep = ",")
```

```{r (june) creating new columns with date}
june$started_at <- ymd_hms(june$started_at)
june$date <- as.Date(june$started_at)
```

```{r (june) creating start time by formatting}
june$start_time<- format(june$started_at, "%H:%M:%S")
```

```{r (june) creating end time by formatting}
june$ended_at <- ymd_hms(june$ended_at)
june$end_time<- format(june$ended_at, "%H:%M:%S")
```

```{r (june) created the duration column by formatting and using difftime}

june$started_at <- ymd_hms(june$started_at)
june$ended_at <- ymd_hms(june$ended_at)
june$duration <- difftime(june$ended_at, june$started_at, units =
"secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
```

```
june <- june[june$duration >= 0, ]

```

```{r (june) excluding unused columns }
june <- subset(june, select = c(  "rideable_type", "start_station_name",
"end_station_name", "start_station_id", "end_station_id",
"member_casual", "start_location", "end_location","date",
"start_time","end_time", "duration"))
```

```{r (june) removing white spaces to create a new df}
june_cleaned <- data.frame(lapply(june, trimws))
```

```{r (june) new csv for export}
# Write the cleaned data to a new file
write.csv(june_cleaned, file = "june_cleaned_data.csv", row.names =
FALSE)
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++--

#JULY

```{r}
# Load the data into R
july <- read.csv("Files/202007-divvy-tripdata.csv",stringsAsFactors =
FALSE)
```

```{r}
# Explore the data
str(july)
summary(july)
head(july)
tail(july)
```
```{r}
# Clean the data
# Remove rows with missing data
july <- na.omit(july)
```

```
# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(july)
head(july)
tail(july)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(july$rideable_type))
sum(is.na(july$started_at))
sum(is.na(july$ended_at))
sum(is.na(july$start_station_name))
sum(is.na(july$end_station_name))
sum(is.na(july$start_station_id))
sum(is.na(july$start_lat))
sum(is.na(july$start_lng))
sum(is.na(july$end_lat))
sum(is.na(july$end_lng))
sum(is.na(july$member_casual))
```

```{r (july) combining the co-ordinates to get the start locations}
july$start_location <- paste(july$start_lat, july$start_lng, sep = ",")
```

```{r (july2) combining the c0-0rdinates to get the end locations}
july$end_location <- paste(july$end_lat, july$end_lng, sep = ",")
```

```{r (july) creating new columns with date}
july$started_at <- ymd_hms(july$started_at)
july$date <- as.Date(july$started_at)
```

```{r (july) creating start time by formatting}
july$start_time<- format(july$started_at, "%H:%M:%S")
```

```{r (july) creating end time by formatting}
```

```
july$ended_at <- ymd_hms(july$ended_at)
july$end_time<- format(july$ended_at, "%H:%M:%S")
```

```{r (july) created the duration column by formatting and using difftime}

july$started_at <- ymd_hms(july$started_at)
july$ended_at <- ymd_hms(july$ended_at)
july$duration <- difftime(july$ended_at, july$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
july <- july[july$duration >= 0, ]

```

```{r (july) excluding unused columns and removing white spaces to create a new df}
july <- subset(july, select = c("rideable_type", "start_station_name", "end_station_name", "start_station_id", "end_station_id", "member_casual", "start_location", "end_location","date", "start_time","end_time", "duration"))
july_cleaned <- data.frame(lapply(july, trimws))
```

```{r (july) new csv}
# Write the cleaned data to a new file
write.csv(july_cleaned, file = "july_cleaned_data.csv", row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#AUGUST

```{r}
# Load the data into R
august <- read.csv("Files/202008-divvy-tripdata.csv",stringsAsFactors =
```

```
FALSE)
```

```{r}
# Explore the data
str(august)
summary(august)
head(august)
tail(august)
```
```{r}
# Clean the data
# Remove rows with missing data
august <- na.omit(august)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(august)
head(august)
tail(august)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(august$rideable_type))
sum(is.na(august$started_at))
sum(is.na(august$ended_at))
sum(is.na(august$start_station_name))
sum(is.na(august$station_name))
sum(is.na(august$start_station_id))
sum(is.na(august$start_lat))
sum(is.na(august$start_lng))
sum(is.na(august$end_lat))
sum(is.na(august$end_lng))
sum(is.na(august$member_casual))
```

```{r (august) combining the co-ordinates to get the start locations}
august$start_location <- paste(august$start_lat, august$start_lng, sep =
",")
```

```{r (august) combining the c0-0rdinates to get the end locations}
august$end_location <- paste(august$end_lat, august$end_lng, sep =
",")
```

```{r (august) creating new columns with date}
august$started_at <- ymd_hms(august$started_at)
august$date <- as.Date(august$started_at)
```

```{r (august) creating start time by formatting}
august$start_time<- format(august$started_at, "%H:%M:%S")
```

```{r (august) creating end time by formatting}
august$ended_at <- ymd_hms(august$ended_at)
august$end_time<- format(august$ended_at, "%H:%M:%S")
```

```{r (august) created the duration column by formatting and using
difftime}

august$started_at <- ymd_hms(august$started_at)
august$ended_at <- ymd_hms(august$ended_at)
august$duration <- difftime(august$ended_at, august$started_at, units =
"secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
august <- august[august$duration >= 0, ]

```

```{r (august) excluding unused columns and removing white spaces to
create a new df}
august <- subset(august, select = c("rideable_type",
"start_station_name", "end_station_name", "start_station_id",
"end_station_id", "member_casual", "start_location",
"end_location","date", "start_time","end_time", "duration"))
august_cleaned <- data.frame(lapply(august, trimws))
```

```{r (august) new csv}
# Write the cleaned data to a new file
write.csv(august_cleaned, file = "august_cleaned_data.csv", row.names =
```

FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#SEPTEMBER

```{r}
# Load the data into R
september <- read.csv("Files/202009-divvy-tripdata.csv",stringsAsFactors = FALSE)
```

```{r}
# Explore the data
str(september)
summary(september)
head(september)
tail(september)
```
```{r}
# Clean the data
# Remove rows with missing data
september <- na.omit(september)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value", #data$column_name)

```

```{r}
# Verify the data
summary(september)
head(september)
tail(september)

```
```{r}
# Check for missing values in the "column_name" column

```
sum(is.na(september$rideable_type))
sum(is.na(september$started_at))
sum(is.na(september$ended_at))
sum(is.na(september$start_station_name))
sum(is.na(september$station_name))
sum(is.na(september$start_station_id))
sum(is.na(september$start_lat))
sum(is.na(september$start_lng))
sum(is.na(september$end_lat))
sum(is.na(september$end_lng))
sum(is.na(september$member_casual))
```

```{r (september) combining the co-ordinates to get the start locations}
september$start_location <- paste(september$start_lat,
september$start_lng, sep = ",")
```

```{r (september) combining the c0-0rdinates to get the end locations}
september$end_location <- paste(september$end_lat,
september$end_lng, sep = ",")
```

```{r (september) creating new columns with date}
september$started_at <- ymd_hms(september$started_at)
september$date <- as.Date(september$started_at)
```

```{r (september) creating start time by formatting}
september$start_time<- format(september$started_at, "%H:%M:%S")
```

```{r (september) creating end time by formatting}
september$ended_at <- ymd_hms(september$ended_at)
september$end_time<- format(september$ended_at, "%H:%M:%S")
```

```{r (september) created the duration column by formatting and using
difftime}

september$started_at <- ymd_hms(september$started_at)
september$ended_at <- ymd_hms(september$ended_at)
september$duration <- difftime(september$ended_at,
september$started_at, units = "secs")
```

```{r}
```

```
# Subsetting to remove rows with negative values in the "duration" column
september <- september[september$duration >= 0, ]

```

```{r (september) excluding unused columns and removing white spaces to create a new df}
september <- subset(september, select = c("rideable_type", "start_station_name", "end_station_name", "start_station_id", "end_station_id", "member_casual", "start_location", "end_location","date", "start_time","end_time", "duration"))
september_cleaned <- data.frame(lapply(september, trimws))
```

```{r (september) new csv}
# Write the cleaned data to a new file
write.csv(september_cleaned, file = "september_cleaned_data.csv", row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#OCTOBER

```{r}
# Load the data into R
october <- read.csv("Files/202010-divvy-tripdata.csv",stringsAsFactors = FALSE)
```

```{r}
# Explore the data
str(october)
summary(october)
head(october)
tail(october)
```
```{r}
# Clean the data
```

```
# Remove rows with missing data
october <- na.omit(october)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(october)
head(october)
tail(october)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(october$rideable_type))
sum(is.na(october$started_at))
sum(is.na(october$ended_at))
sum(is.na(october$start_station_name))
sum(is.na(october$station_name))
sum(is.na(october$start_station_id))
sum(is.na(october$start_lat))
sum(is.na(october$start_lng))
sum(is.na(october$end_lat))
sum(is.na(october$end_lng))
sum(is.na(october$member_casual))
```

```{r (october) combining the co-ordinates to get the start locations}
october$start_location <- paste(october$start_lat, october$start_lng,
sep = ",")
```

```{r (october) combining the c0-0rdinates to get the end locations}
october$end_location <- paste(october$end_lat, october$end_lng, sep =
",")
```

```{r (october) creating new columns with date}
october$started_at <- ymd_hms(october$started_at)
october$date <- as.Date(october$started_at)
```

```{r (october) creating start time by formatting}
```

```
october$start_time<- format(october$started_at, "%H:%M:%S")
```

```{r (october) creating end time by formatting}
october$ended_at <- ymd_hms(october$ended_at)
october$end_time<- format(october$ended_at, "%H:%M:%S")
```

```{r (october) created the duration column by formatting and using difftime}

october$started_at <- ymd_hms(october$started_at)
october$ended_at <- ymd_hms(october$ended_at)
october$duration <- difftime(october$ended_at, october$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
october <- october[october$duration >= 0, ]

```

```{r (october) excluding unused columns and removing white spaces to create a new df}
october <- subset(october, select = c("rideable_type", "start_station_name", "end_station_name", "start_station_id", "end_station_id", "member_casual", "start_location", "end_location","date", "start_time","end_time", "duration"))
october_cleaned <- data.frame(lapply(october, trimws))
```

```{r (october) new csv}
# Write the cleaned data to a new file
write.csv(october_cleaned, file = "october_cleaned_data.csv", row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#NOVEMBER

```{r}
# Load the data into R
november <- read.csv("Files/202011-divvy-tripdata.csv",stringsAsFactors
= FALSE)
```

```{r}
# Explore the data
str(november)
summary(november)
head(november)
tail(november)
```
```{r}
# Clean the data
# Remove rows with missing data
november <- na.omit(november)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(november)
head(november)
tail(november)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(november$rideable_type))
sum(is.na(november$started_at))
sum(is.na(november$ended_at))
sum(is.na(november$start_station_name))
sum(is.na(november$station_name))
sum(is.na(november$start_station_id))
sum(is.na(november$start_lat))
sum(is.na(november$start_lng))
sum(is.na(november$end_lat))
sum(is.na(november$end_lng))
sum(is.na(november$member_casual))
```

```{r (november) combining the co-ordinates to get the start locations}
november$start_location <- paste(november$start_lat,
november$start_lng, sep = ",")
```

```{r (november) combining the c0-0rdinates to get the end locations}
november$end_location <- paste(november$end_lat, november$end_lng,
sep = ",")
```

```{r (november) creating new columns with date}
november$started_at <- ymd_hms(november$started_at)
november$date <- as.Date(november$started_at)
```

```{r (november) creating start time by formatting}
november$start_time<- format(november$started_at, "%H:%M:%S")
```

```{r (november) creating end time by formatting}
november$ended_at <- ymd_hms(november$ended_at)
november$end_time<- format(november$ended_at, "%H:%M:%S")
```

```{r (november) created the duration column by formatting and using
difftime}

november$started_at <- ymd_hms(november$started_at)
november$ended_at <- ymd_hms(november$ended_at)
november$duration <- difftime(november$ended_at,
november$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
november <- november[november$duration >= 0, ]

```

```{r (november) excluding unused columns and removing white spaces to
create a new df}
november <- subset(november, select = c("rideable_type",
"start_station_name", "end_station_name", "start_station_id",
"end_station_id", "member_casual", "start_location",
"end_location","date", "start_time","end_time", "duration"))
november_cleaned <- data.frame(lapply(november, trimws))
```

```

```{r (november) new csv}
# Write the cleaned data to a new file
write.csv(november_cleaned, file = "november_cleaned_data.csv",
row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#DECEMBER

```{r}
# Load the data into R
december <- read.csv("Files/202012-divvy-
tripdata.csv",stringsAsFactors = FALSE)
```

```{r}
# Explore the data
str(december)
summary(december)
head(december)
tail(december)
```
```{r}
# Clean the data
# Remove rows with missing data
december <- na.omit(december)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(december)
head(december)
```
```

tail(december)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(december$rideable_type))
sum(is.na(december$started_at))
sum(is.na(december$ended_at))
sum(is.na(december$start_station_name))
sum(is.na(december$station_name))
sum(is.na(december$start_station_id))
sum(is.na(december$start_lat))
sum(is.na(december$start_lng))
sum(is.na(december$end_lat))
sum(is.na(december$end_lng))
sum(is.na(december$member_casual))
```

```{r (december) combining the co-ordinates to get the start locations}
december$start_location <- paste(december$start_lat,
december$start_lng, sep = ",")
```

```{r (december) combining the c0-0rdinates to get the end locations}
december$end_location <- paste(december$end_lat, december$end_lng,
sep = ",")
```

```{r (december) creating new columns with date}
december$started_at <- ymd_hms(december$started_at)
december$date <- as.Date(december$started_at)
```

```{r (december) creating start time by formatting}
december$start_time<- format(december$started_at, "%H:%M:%S")
```

```{r (december) creating end time by formatting}
december$ended_at <- ymd_hms(december$ended_at)
december$end_time<- format(december$ended_at, "%H:%M:%S")
```

```{r (december) created the duration column by formatting and using
difftime}

december$started_at <- ymd_hms(december$started_at)
december$ended_at <- ymd_hms(december$ended_at)

```
december$duration <- difftime(december$ended_at,
december$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
december <- december[december$duration >= 0, ]

```

```{r (december) excluding unused columns and removing white spaces to create a new df}
december <- subset(december, select = c("rideable_type",
"start_station_name", "end_station_name", "start_station_id",
"end_station_id", "member_casual", "start_location",
"end_location","date", "start_time","end_time", "duration"))
december_cleaned <- data.frame(lapply(december, trimws))
```

```{r (december) new csv}
# Write the cleaned data to a new file
write.csv(december_cleaned, file = "december_cleaned_data.csv",
row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#JANUARY

```{r}
# Load the data into R
january <- read.csv("Files/202101-divvy-tripdata.csv",stringsAsFactors =
FALSE)
```

```{r}
# Explore the data
str(january)
summary(january)
head(january)
```

```
tail(january)
```
```{r}
# Clean the data
# Remove rows with missing data
january <- na.omit(january)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```


```{r}
# Verify the data
summary(january)
head(january)
tail(january)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(january$rideable_type))
sum(is.na(january$started_at))
sum(is.na(january$ended_at))
sum(is.na(january$start_station_name))
sum(is.na(january$station_name))
sum(is.na(january$start_station_id))
sum(is.na(january$start_lat))
sum(is.na(january$start_lng))
sum(is.na(january$end_lat))
sum(is.na(january$end_lng))
sum(is.na(january$member_casual))
```


```{r (january) combining the co-ordinates to get the start locations}
january$start_location <- paste(january$start_lat, january$start_lng, sep
= ",")
```


```{r (january) combining the c0-0rdinates to get the end locations}
january$end_location <- paste(january$end_lat, january$end_lng, sep =
",")
```


```{r (january) creating new columns with date}
january$started_at <- ymd_hms(january$started_at)
```

```
january$date <- as.Date(january$started_at)
```

```{r (january) creating start time by formatting}
january$start_time<- format(january$started_at, "%H:%M:%S")
```

```{r (january) creating end time by formatting}
january$ended_at <- ymd_hms(january$ended_at)
january$end_time<- format(january$ended_at, "%H:%M:%S")
```

```{r (january) created the duration column by formatting and using difftime}

january$started_at <- ymd_hms(january$started_at)
january$ended_at <- ymd_hms(january$ended_at)
january$duration <- difftime(january$ended_at, january$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
january <- january[january$duration >= 0, ]

```

```{r (january) excluding unused columns and removing white spaces to create a new df}
january <- subset(january, select = c("rideable_type", "start_station_name", "end_station_name", "start_station_id", "end_station_id", "member_casual", "start_location", "end_location","date", "start_time","end_time", "duration"))
january_cleaned <- data.frame(lapply(january, trimws))
```

```{r (january) new csv}
# Write the cleaned data to a new file
write.csv(january_cleaned, file = "january_cleaned_data.csv", row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#FEBRUARY

```{r}
# Load the data into R
february <- read.csv("Files/202102-divvy-tripdata.csv",stringsAsFactors = FALSE)
```

```{r}
# Explore the data
str(february)
summary(february)
head(february)
tail(february)
```
```{r}
# Clean the data
# Remove rows with missing data
february <- na.omit(february)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value", #data$column_name)

```

```{r}
# Verify the data
summary(february)
head(february)
tail(february)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(february$rideable_type))
sum(is.na(february$started_at))
sum(is.na(february$ended_at))
sum(is.na(february$start_station_name))
sum(is.na(february$station_name))
sum(is.na(february$start_station_id))
sum(is.na(february$start_lat))
```

```
sum(is.na(february$start_lng))
sum(is.na(february$end_lat))
sum(is.na(february$end_lng))
sum(is.na(february$member_casual))
```

```{r (february) combining the co-ordinates to get the start locations}
february$start_location <- paste(february$start_lat, february$start_lng, sep = ",")
```

```{r (february) combining the c0-0rdinates to get the end locations}
february$end_location <- paste(february$end_lat, february$end_lng, sep = ",")
```

```{r (february) creating new columns with date}
february$started_at <- ymd_hms(february$started_at)
february$date <- as.Date(february$started_at)
```

```{r (february) creating start time by formatting}
february$start_time<- format(february$started_at, "%H:%M:%S")
```

```{r (february) creating end time by formatting}
february$ended_at <- ymd_hms(february$ended_at)
february$end_time<- format(february$ended_at, "%H:%M:%S")
```

```{r (february) created the duration column by formatting and using difftime}

february$started_at <- ymd_hms(february$started_at)
february$ended_at <- ymd_hms(february$ended_at)
february$duration <- difftime(february$ended_at, february$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
february <- february[february$duration >= 0, ]

```

```{r (february) excluding unused columns and removing white spaces to create a new df}
```

```
february <- subset(february, select = c("rideable_type",
"start_station_name", "end_station_name", "start_station_id",
"end_station_id", "member_casual", "start_location",
"end_location","date", "start_time","end_time", "duration"))
february_cleaned <- data.frame(lapply(february, trimws))
```

```{r (february) new csv}
# Write the cleaned data to a new file
write.csv(february_cleaned, file = "february_cleaned_data.csv",
row.names = FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#MARCH

```{r}
# Load the data into R
march <- read.csv("Files/202103-divvy-tripdata.csv",stringsAsFactors =
FALSE)
```

```{r}
# Explore the data
str(march)
summary(march)
head(march)
tail(march)
```
```{r}
# Clean the data
# Remove rows with missing data
march <- na.omit(march)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```

```{r}
# Verify the data
summary(march)
head(march)
tail(march)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(march$rideable_type))
sum(is.na(march$started_at))
sum(is.na(march$ended_at))
sum(is.na(march$start_station_name))
sum(is.na(march$station_name))
sum(is.na(march$start_station_id))
sum(is.na(march$start_lat))
sum(is.na(march$start_lng))
sum(is.na(march$end_lat))
sum(is.na(march$end_lng))
sum(is.na(march$member_casual))
```

```{r (march) combining the co-ordinates to get the start locations}
march$start_location <- paste(march$start_lat, march$start_lng, sep =
",")
```

```{r (march) combining the c0-0rdinates to get the end locations}
march$end_location <- paste(march$end_lat, march$end_lng, sep = ",")
```

```{r (march) creating new columns with date}
march$started_at <- ymd_hms(march$started_at)
march$date <- as.Date(march$started_at)
```

```{r (march) creating start time by formatting}
march$start_time<- format(march$started_at, "%H:%M:%S")
```

```{r (march) creating end time by formatting}
march$ended_at <- ymd_hms(march$ended_at)
march$end_time<- format(march$ended_at, "%H:%M:%S")
```

```{r (march) created the duration column by formatting and using
```

```
difftime}

march$started_at <- ymd_hms(march$started_at)
march$ended_at <- ymd_hms(march$ended_at)
march$duration <- difftime(march$ended_at, march$started_at, units =
"secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
march <- march[march$duration >= 0, ]

```

```{r (march) excluding unused columns and removing white spaces to
create a new df}
march <- subset(march, select = c("rideable_type",
"start_station_name", "end_station_name", "start_station_id",
"end_station_id", "member_casual", "start_location",
"end_location","date", "start_time","end_time", "duration"))
march_cleaned <- data.frame(lapply(march, trimws))
```

```{r (march) new csv}
# Write the cleaned data to a new file
write.csv(march_cleaned, file = "march_cleaned_data.csv", row.names =
FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#APRIL

```{r}
# Load the data into R
april <- read.csv("Files/202104-divvy-tripdata.csv",stringsAsFactors =
FALSE)
```

```{r}
# Explore the data
str(april)
summary(april)
head(april)
tail(april)
```
```{r}
# Clean the data
# Remove rows with missing data
april <- na.omit(april)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value",
#data$column_name)

```


```{r}
# Verify the data
summary(april)
head(april)
tail(april)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(april$rideable_type))
sum(is.na(april$started_at))
sum(is.na(april$ended_at))
sum(is.na(april$start_station_name))
sum(is.na(april$station_name))
sum(is.na(april$start_station_id))
sum(is.na(april$start_lat))
sum(is.na(april$start_lng))
sum(is.na(april$end_lat))
sum(is.na(april$end_lng))
sum(is.na(april$member_casual))
```


```{r (april) combining the co-ordinates to get the start locations}
april$start_location <- paste(april$start_lat, april$start_lng, sep = ",")
```


```{r (april) combining the c0-0rdinates to get the end locations}
april$end_location <- paste(april$end_lat, april$end_lng, sep = ",")
```

```

```{r (april) creating new columns with date}
april$started_at <- ymd_hms(april$started_at)
april$date <- as.Date(april$started_at)
```

```{r (april) creating start time by formatting}
april$start_time<- format(april$started_at, "%H:%M:%S")
```

```{r (april) creating end time by formatting}
april$ended_at <- ymd_hms(april$ended_at)
april$end_time<- format(april$ended_at, "%H:%M:%S")
```

```{r (april) created the duration column by formatting and using difftime}

april$started_at <- ymd_hms(april$started_at)
april$ended_at <- ymd_hms(april$ended_at)
april$duration <- difftime(april$ended_at, april$started_at, units =
"secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
april <- april[april$duration >= 0, ]

```

```{r (april) excluding unused columns and removing white spaces to
create a new df}
april <- subset(april, select = c("rideable_type", "start_station_name",
"end_station_name", "start_station_id", "end_station_id",
"member_casual", "start_location", "end_location","date",
"start_time","end_time", "duration"))
april_cleaned <- data.frame(lapply(april, trimws))
```

```{r (april) new csv}
# Write the cleaned data to a new file
write.csv(april_cleaned, file = "april_cleaned_data.csv", row.names =
FALSE)
```

```{r}
#testbutton
```
```

```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++--

#MAY

```{r}
# Load the data into R
may <- read.csv("Files/202105-divvy-tripdata.csv",stringsAsFactors = FALSE)
```

```{r}
# Explore the data
str(may)
summary(may)
head(may)
tail(may)
```
```{r}
# Clean the data
# Remove rows with missing data
may <- na.omit(may)

# Replace incorrect values
#data$column_name <- gsub("incorrect_value", "correct_value", #data$column_name)

```

```{r}
# Verify the data
summary(may)
head(may)
tail(may)

```
```{r}
# Check for missing values in the "column_name" column
sum(is.na(may$rideable_type))
sum(is.na(may$started_at))
sum(is.na(may$ended_at))
sum(is.na(may$start_station_name))
sum(is.na(may$station_name))
```
```

```
sum(is.na(may$start_station_id))
sum(is.na(may$start_lat))
sum(is.na(may$start_lng))
sum(is.na(may$end_lat))
sum(is.na(may$end_lng))
sum(is.na(may$member_casual))
```

```{r (may) combining the co-ordinates to get the start locations}
may$start_location <- paste(may$start_lat, may$start_lng, sep = ",")
```

```{r (may) combining the c0-0rdinates to get the end locations}
may$end_location <- paste(may$end_lat, may$end_lng, sep = ",")
```

```{r (may) creating new columns with date}
may$started_at <- ymd_hms(may$started_at)
may$date <- as.Date(may$started_at)
```

```{r (may) creating start time by formatting}
may$start_time<- format(may$started_at, "%H:%M:%S")
```

```{r (may) creating end time by formatting}
may$ended_at <- ymd_hms(may$ended_at)
may$end_time<- format(may$ended_at, "%H:%M:%S")
```

```{r (may) created the duration column by formatting and using difftime}

may$started_at <- ymd_hms(may$started_at)
may$ended_at <- ymd_hms(may$ended_at)
may$duration <- difftime(may$ended_at, may$started_at, units = "secs")
```

```{r}
# Subsetting to remove rows with negative values in the "duration" column
may <- may[may$duration >= 0, ]

```

```{r (may) excluding unused columns and removing white spaces to create a new df}
may <- subset(may, select = c("rideable_type", "start_station_name", "end_station_name", "start_station_id", "end_station_id",
```

"member_casual", "start_location", "end_location","date",
"start_time","end_time", "duration"))
may_cleaned <- data.frame(lapply(may, trimws))
```

```{r (may) new csv}
# Write the cleaned data to a new file
write.csv(may_cleaned, file = "may_cleaned_data.csv", row.names =
FALSE)
```

```{r}
#testbutton
```

-- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++--
```