Data Cleaning and Manipulation

Anita Dash

2022-05-23

Data Cleaning

Previous 12 Months data of Cyclistics's Bike Trips were downloaded.(May 2021-April 2022)

The monthly data were cleaned individually.They were done so to avoid if there were any date mismatches in a particular month, or if a particular month showed some certain patterns of error.

All the Month data were cleaned based on the following order

1. Column names were read and converted to upper camel case.
2. Duplicate Entries were removed.
3. Rows with empty data in one of the columns- RideId, RideableType, StartedAt, EndedAt, MemberCasual were dropped.
4. Checking for mistyped entries in RideableType and MemberCasual.
5. Entries in the StartedAt, EndedAt column were split into StartDate, StartTime, EndDate, EndTime.
6. Errors in Date were checked, if there was any date that was before or after the month of the data, or if start date was after the end date.
7. New column- RideLength that specified the duration of the ride was added.
8. New columns- StartDay, EndDay were added to mention the day of the week the ride took place.
9. A new data set which was the subset of the months dataset was made, it included the columns- RideId, RideableType, MemberCasual, StartDate, EndDate, StartTime, EndTime, RideLength, StartDay, EndDay

(Even though information regarding the stations would have been useful in the analysis, a lot of rows did not have station related data and we could not get the details regarding the stations, thus the station related columns were dropped)

Packages and library used:

```
install.packages("lubridate")
install.packages("hms")
install.packages("janitor")
install.packages("tidyverse")
library("tidyverse")
library("tidyr")
library("readr")
library("dplyr")
library("janitor")
library("lubridate")
library("hms")
```

Month 1 Cleaning

```
#Getting a rough Metadata about the monthly data
month_1 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202105-divvy-tripdata.csv")
glimpse(month_1)
month_1 <- clean_names(month_1,"upper_camel")
colnames(month_1)

#Removing Duplicate Entries
month_1_dist <- distinct(month_1)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_1_dist <- month_1_dist %>% drop_na(RideId)
month_1_dist <- month_1_dist %>% drop_na(RideableType)
month_1_dist <- month_1_dist %>% drop_na(StartedAt)
month_1_dist <- month_1_dist %>% drop_na(EndedAt)
month_1_dist <- month_1_dist %>% drop_na(MemberCasual)
View(month_1_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_1_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_1_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_1_dist$StartDate <- as.Date(month_1_dist$StartedAt)
month_1_dist$StartTime <- format(as.POSIXct(month_1_dist$StartedAt),format =
"%H:%M:%S")
month_1_dist$EndDate <- as.Date(month_1_dist$EndedAt)
month_1_dist$EndTime <- format(as.POSIXct(month_1_dist$EndedAt),format =
"%H:%M:%S")
```

```
#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_1_dist %>% filter(StartDate > '2021-05-31' | StartDate< '2021-05-01')
month_1_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_1_dist <- mutate(month_1_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_1_dist$RideLength <- as_hms(month_1_dist$RideLength)

#Finding which day it was from the date
month_1_dist$StartDay <- weekdays(as.Date(month_1_dist$StartDate))
month_1_dist$EndDay <- weekdays(as.Date(month_1_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_1_clean_analysis <- subset(month_1_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_1_clean_analysis)
```

Month 2 Cleaning

```
#Getting a rough Metadata about the monthly data
month_2 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202106-divvy-tripdata.csv")
glimpse(month_2)
month_2 <- clean_names(month_2,"upper_camel")
colnames(month_2)

#Removing Duplicate Entries
month_2_dist <- distinct(month_2)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_2_dist <- month_2_dist %>% drop_na(RideId)
month_2_dist <- month_2_dist %>% drop_na(RideableType)
month_2_dist <- month_2_dist %>% drop_na(StartedAt)
month_2_dist <- month_2_dist %>% drop_na(EndedAt)
month_2_dist <- month_2_dist %>% drop_na(MemberCasual)
View(month_2_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_2_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_2_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_2_dist$StartDate <- as.Date(month_2_dist$StartedAt)
month_2_dist$StartTime <- format(as.POSIXct(month_2_dist$StartedAt),format =
"%H:%M:%S")
month_2_dist$EndDate <- as.Date(month_2_dist$EndedAt)
month_2_dist$EndTime <- format(as.POSIXct(month_2_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_2_dist %>% filter(StartDate > '2021-06-30' | StartDate< '2021-06-01')
month_2_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_2_dist <- mutate(month_2_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_2_dist$RideLength <- as_hms(month_2_dist$RideLength)

#Finding which day it was from the date
month_2_dist$StartDay <- weekdays(as.Date(month_2_dist$StartDate))
month_2_dist$EndDay <- weekdays(as.Date(month_2_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_2_clean_analysis <- subset(month_2_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_2_clean_analysis)
```

Month 3 Cleaning

```
#Getting a rough Metadata about the monthly data
month_3 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202107-divvy-tripdata.csv")
glimpse(month_3)
month_3 <- clean_names(month_3,"upper_camel")
colnames(month_3)

#Removing Duplicate Entries
month_3_dist <- distinct(month_3)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_3_dist <- month_3_dist %>% drop_na(RideId)
month_3_dist <- month_3_dist %>% drop_na(RideableType)
month_3_dist <- month_3_dist %>% drop_na(StartedAt)
month_3_dist <- month_3_dist %>% drop_na(EndedAt)
```

```
month_3_dist <- month_3_dist %>% drop_na(MemberCasual)
View(month_3_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_3_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_3_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_3_dist$StartDate <- as.Date(month_3_dist$StartedAt)
month_3_dist$StartTime <- format(as.POSIXct(month_3_dist$StartedAt),format =
"%H:%M:%S")
month_3_dist$EndDate <- as.Date(month_3_dist$EndedAt)
month_3_dist$EndTime <- format(as.POSIXct(month_3_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_3_dist %>% filter(StartDate > '2021-07-31' | StartDate< '2021-07-01')
month_3_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_3_dist <- mutate(month_3_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_3_dist$RideLength <- as_hms(month_3_dist$RideLength)

#Finding which day it was from the date
month_3_dist$StartDay <- weekdays(as.Date(month_3_dist$StartDate))
month_3_dist$EndDay <- weekdays(as.Date(month_3_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_3_clean_analysis <- subset(month_3_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_3_clean_analysis)
```

Month 4 Cleaning

```
#Getting a rough Metadata about the monthly data
month_4 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202108-divvy-tripdata.csv")
glimpse(month_4)
month_4 <- clean_names(month_4,"upper_camel")
```

```
colnames(month_4)

#Removing Duplicate Entries
month_4_dist <- distinct(month_4)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_4_dist <- month_4_dist %>% drop_na(RideId)
month_4_dist <- month_4_dist %>% drop_na(RideableType)
month_4_dist <- month_4_dist %>% drop_na(StartedAt)
month_4_dist <- month_4_dist %>% drop_na(EndedAt)
month_4_dist <- month_4_dist %>% drop_na(MemberCasual)
View(month_4_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_4_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_4_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_4_dist$StartDate <- as.Date(month_4_dist$StartedAt)
month_4_dist$StartTime <- format(as.POSIXct(month_4_dist$StartedAt),format =
"%H:%M:%S")
month_4_dist$EndDate <- as.Date(month_4_dist$EndedAt)
month_4_dist$EndTime <- format(as.POSIXct(month_4_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_4_dist %>% filter(StartDate > '2021-08-31' | StartDate< '2021-08-01')
month_4_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_4_dist <- mutate(month_4_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_4_dist$RideLength <- as_hms(month_4_dist$RideLength)

#Finding which day it was from the date
month_4_dist$StartDay <- weekdays(as.Date(month_4_dist$StartDate))
month_4_dist$EndDay <- weekdays(as.Date(month_4_dist$EndDate))

#making a new dataset with the variables required for the analysis
```

```
month_4_clean_analysis <- subset(month_4_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_4_clean_analysis)
```

Month 5 Cleaning

```
#Getting a rough Metadata about the monthly data
month_5 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202109-divvy-tripdata.csv")
glimpse(month_5)
month_5 <- clean_names(month_5,"upper_camel")
colnames(month_5)

#Removing Duplicate Entries
month_5_dist <- distinct(month_5)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_5_dist <- month_5_dist %>% drop_na(RideId)
month_5_dist <- month_5_dist %>% drop_na(RideableType)
month_5_dist <- month_5_dist %>% drop_na(StartedAt)
month_5_dist <- month_5_dist %>% drop_na(EndedAt)
month_5_dist <- month_5_dist %>% drop_na(MemberCasual)
View(month_5_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_5_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_5_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_5_dist$StartDate <- as.Date(month_5_dist$StartedAt)
month_5_dist$StartTime <- format(as.POSIXct(month_5_dist$StartedAt),format =
"%H:%M:%S")
month_5_dist$EndDate <- as.Date(month_5_dist$EndedAt)
month_5_dist$EndTime <- format(as.POSIXct(month_5_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_5_dist %>% filter(StartDate > '2021-09-30' | StartDate< '2021-09-01')
month_5_dist %>% filter(StartDate > EndDate)
#no errors
```

```
#Making a new column to find duration of ride, format; H:M:S
month_5_dist <- mutate(month_5_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_5_dist$RideLength <- as_hms(month_5_dist$RideLength)

#Finding which day it was from the date
month_5_dist$StartDay <- weekdays(as.Date(month_5_dist$StartDate))
month_5_dist$EndDay <- weekdays(as.Date(month_5_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_5_clean_analysis <- subset(month_5_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_5_clean_analysis)
```

Month 6 Cleaning

```
#Getting a rough Metadata about the monthly data
month_6 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202110-divvy-tripdata.csv")
glimpse(month_6)
month_6 <- clean_names(month_6,"upper_camel")
colnames(month_6)

#Removing Duplicate Entries
month_6_dist <- distinct(month_6)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_6_dist <- month_6_dist %>% drop_na(RideId)
month_6_dist <- month_6_dist %>% drop_na(RideableType)
month_6_dist <- month_6_dist %>% drop_na(StartedAt)
month_6_dist <- month_6_dist %>% drop_na(EndedAt)
month_6_dist <- month_6_dist %>% drop_na(MemberCasual)
View(month_6_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_6_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_6_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_6_dist$StartDate <- as.Date(month_6_dist$StartedAt)
```

```
month_6_dist$StartTime <- format(as.POSIXct(month_6_dist$StartedAt),format =
"%H:%M:%S")
month_6_dist$EndDate <- as.Date(month_6_dist$EndedAt)
month_6_dist$EndTime <- format(as.POSIXct(month_6_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_6_dist %>% filter(StartDate > '2021-10-31' | StartDate< '2021-10-01')
month_6_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_6_dist <- mutate(month_6_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_6_dist$RideLength <- as_hms(month_6_dist$RideLength)

#Finding which day it was from the date
month_6_dist$StartDay <- weekdays(as.Date(month_6_dist$StartDate))
month_6_dist$EndDay <- weekdays(as.Date(month_6_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_6_clean_analysis <- subset(month_6_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_6_clean_analysis)
```

Month 7 Cleaning

```
#Getting a rough Metadata about the monthly data
month_7 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202111-divvy-tripdata.csv")
glimpse(month_7)
month_7 <- clean_names(month_7,"upper_camel")
colnames(month_7)

#Removing Duplicate Entries
month_7_dist <- distinct(month_7)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_7_dist <- month_7_dist %>% drop_na(RideId)
month_7_dist <- month_7_dist %>% drop_na(RideableType)
month_7_dist <- month_7_dist %>% drop_na(StartedAt)
month_7_dist <- month_7_dist %>% drop_na(EndedAt)
month_7_dist <- month_7_dist %>% drop_na(MemberCasual)
View(month_7_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
```

```
errors
month_7_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_7_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_7_dist$StartDate <- as.Date(month_7_dist$StartedAt)
month_7_dist$StartTime <- format(as.POSIXct(month_7_dist$StartedAt),format =
"%H:%M:%S")
month_7_dist$EndDate <- as.Date(month_7_dist$EndedAt)
month_7_dist$EndTime <- format(as.POSIXct(month_7_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_7_dist %>% filter(StartDate > '2021-11-30' | StartDate< '2021-11-01')
month_7_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_7_dist <- mutate(month_7_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_7_dist$RideLength <- as_hms(month_7_dist$RideLength)

#Finding which day it was from the date
month_7_dist$StartDay <- weekdays(as.Date(month_7_dist$StartDate))
month_7_dist$EndDay <- weekdays(as.Date(month_7_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_7_clean_analysis <- subset(month_7_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_7_clean_analysis)
```

Month 8 Cleaning

```
#Getting a rough Metadata about the monthly data
month_8 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202112-divvy-tripdata.csv")
glimpse(month_8)
month_8 <- clean_names(month_8,"upper_camel")
colnames(month_8)

#Removing Duplicate Entries
month_8_dist <- distinct(month_8)
```

```
#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_8_dist <- month_8_dist %>% drop_na(RideId)
month_8_dist <- month_8_dist %>% drop_na(RideableType)
month_8_dist <- month_8_dist %>% drop_na(StartedAt)
month_8_dist <- month_8_dist %>% drop_na(EndedAt)
month_8_dist <- month_8_dist %>% drop_na(MemberCasual)
View(month_8_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_8_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_8_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_8_dist$StartDate <- as.Date(month_8_dist$StartedAt)
month_8_dist$StartTime <- format(as.POSIXct(month_8_dist$StartedAt),format =
"%H:%M:%S")
month_8_dist$EndDate <- as.Date(month_8_dist$EndedAt)
month_8_dist$EndTime <- format(as.POSIXct(month_8_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_8_dist %>% filter(StartDate > '2021-12-31' | StartDate< '2021-12-01')
month_8_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_8_dist <- mutate(month_8_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_8_dist$RideLength <- as_hms(month_8_dist$RideLength)

#Finding which day it was from the date
month_8_dist$StartDay <- weekdays(as.Date(month_8_dist$StartDate))
month_8_dist$EndDay <- weekdays(as.Date(month_8_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_8_clean_analysis <- subset(month_8_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_8_clean_analysis)
```

Month 9 Cleaning

```
#Getting a rough Metadata about the monthly data
month_9 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202201-divvy-tripdata.csv")
glimpse(month_9)
month_9 <- clean_names(month_9,"upper_camel")
colnames(month_9)

#Removing Duplicate Entries
month_9_dist <- distinct(month_9)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_9_dist <- month_9_dist %>% drop_na(RideId)
month_9_dist <- month_9_dist %>% drop_na(RideableType)
month_9_dist <- month_9_dist %>% drop_na(StartedAt)
month_9_dist <- month_9_dist %>% drop_na(EndedAt)
month_9_dist <- month_9_dist %>% drop_na(MemberCasual)
View(month_9_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_9_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_9_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_9_dist$StartDate <- as.Date(month_9_dist$StartedAt)
month_9_dist$StartTime <- format(as.POSIXct(month_9_dist$StartedAt),format =
"%H:%M:%S")
month_9_dist$EndDate <- as.Date(month_9_dist$EndedAt)
month_9_dist$EndTime <- format(as.POSIXct(month_9_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_9_dist %>% filter(StartDate > '2022-01-31' | StartDate< '2022-01-01')
month_9_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_9_dist <- mutate(month_9_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_9_dist$RideLength <- as_hms(month_9_dist$RideLength)
```

```
#Finding which day it was from the date
month_9_dist$StartDay <- weekdays(as.Date(month_9_dist$StartDate))
month_9_dist$EndDay <- weekdays(as.Date(month_9_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_9_clean_analysis <- subset(month_9_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_9_clean_analysis)
```

Month 10 Cleaning

```
#Getting a rough Metadata about the monthly data
month_10 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202202-divvy-tripdata.csv")
glimpse(month_10)
month_10 <- clean_names(month_10,"upper_camel")
colnames(month_10)

#Removing Duplicate Entries
month_10_dist <- distinct(month_10)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_10_dist <- month_10_dist %>% drop_na(RideId)
month_10_dist <- month_10_dist %>% drop_na(RideableType)
month_10_dist <- month_10_dist %>% drop_na(StartedAt)
month_10_dist <- month_10_dist %>% drop_na(EndedAt)
month_10_dist <- month_10_dist %>% drop_na(MemberCasual)
View(month_10_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_10_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_10_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_10_dist$StartDate <- as.Date(month_10_dist$StartedAt)
month_10_dist$StartTime <- format(as.POSIXct(month_10_dist$StartedAt),format
= "%H:%M:%S")
month_10_dist$EndDate <- as.Date(month_10_dist$EndedAt)
month_10_dist$EndTime <- format(as.POSIXct(month_10_dist$EndedAt),format =
"%H:%M:%S")
```

```
#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_10_dist %>% filter(StartDate > '2022-02-28' | StartDate< '2022-02-01')
month_10_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_10_dist <- mutate(month_10_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_10_dist$RideLength <- as_hms(month_10_dist$RideLength)

#Finding which day it was from the date
month_10_dist$StartDay <- weekdays(as.Date(month_10_dist$StartDate))
month_10_dist$EndDay <- weekdays(as.Date(month_10_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_10_clean_analysis <- subset(month_10_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_10_clean_analysis)
```

Month 11 Cleaning

```
#Getting a rough Metadata about the monthly data
month_11 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202203-divvy-tripdata.csv")
glimpse(month_11)
month_11 <- clean_names(month_11,"upper_camel")
colnames(month_11)

#Removing Duplicate Entries
month_11_dist <- distinct(month_11)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_11_dist <- month_11_dist %>% drop_na(RideId)
month_11_dist <- month_11_dist %>% drop_na(RideableType)
month_11_dist <- month_11_dist %>% drop_na(StartedAt)
month_11_dist <- month_11_dist %>% drop_na(EndedAt)
month_11_dist <- month_11_dist %>% drop_na(MemberCasual)
View(month_11_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any
errors
month_11_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_11_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_11_dist$StartDate <- as.Date(month_11_dist$StartedAt)
month_11_dist$StartTime <- format(as.POSIXct(month_11_dist$StartedAt),format
= "%H:%M:%S")
month_11_dist$EndDate <- as.Date(month_11_dist$EndedAt)
month_11_dist$EndTime <- format(as.POSIXct(month_11_dist$EndedAt),format =
"%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the
month specified, or if end date is before start date)
month_11_dist %>% filter(StartDate > '2022-03-31' | StartDate< '2022-03-01')
month_11_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_11_dist <- mutate(month_11_dist, RideLength =
difftime(EndedAt,StartedAt,units = "secs"))
month_11_dist$RideLength <- as_hms(month_11_dist$RideLength)

#Finding which day it was from the date
month_11_dist$StartDay <- weekdays(as.Date(month_11_dist$StartDate))
month_11_dist$EndDay <- weekdays(as.Date(month_11_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_11_clean_analysis <- subset(month_11_dist, select =
c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLe
ngth,StartDay,EndDay))
View(month_11_clean_analysis)
```

Month 12 Cleaning

```
#Getting a rough Metadata about the monthly data
month_12 <- read_csv("/Users/Anita Dash/Desktop/Google-Capstone/Previous 12
Months/202204-divvy-tripdata.csv")
glimpse(month_12)
month_12 <- clean_names(month_12,"upper_camel")
colnames(month_12)

#Removing Duplicate Entries
month_12_dist <- distinct(month_12)

#Empty entries for columns regarding the station data were not dropped as
These Variables will not be dealt with in the analysis
month_12_dist <- month_12_dist %>% drop_na(RideId)
month_12_dist <- month_12_dist %>% drop_na(RideableType)
month_12_dist <- month_12_dist %>% drop_na(StartedAt)
```

```
month_12_dist <- month_12_dist %>% drop_na(EndedAt)
month_12_dist <- month_12_dist %>% drop_na(MemberCasual)
View(month_12_dist)
#no empty entries, no rows were dropped

#Checking if the data entered for type of ride and membership type had any errors
month_12_dist %>%
  group_by(RideableType) %>%
  summarise(count=n())
month_12_dist %>%
  group_by(MemberCasual) %>%
  summarise(count=n())
#no errors

#making extra two columns for date and time
month_12_dist$StartDate <- as.Date(month_12_dist$StartedAt)
month_12_dist$StartTime <- format(as.POSIXct(month_12_dist$StartedAt),format = "%H:%M:%S")
month_12_dist$EndDate <- as.Date(month_12_dist$EndedAt)
month_12_dist$EndTime <- format(as.POSIXct(month_12_dist$EndedAt),format = "%H:%M:%S")

#Checking for any errors in the date (like if the date is before, after the month specified, or if end date is before start date)
month_12_dist %>% filter(StartDate > '2022-04-30' | StartDate< '2022-04-01')
month_12_dist %>% filter(StartDate > EndDate)
#no errors

#Making a new column to find duration of ride, format; H:M:S
month_12_dist <- mutate(month_12_dist, RideLength = difftime(EndedAt,StartedAt,units = "secs"))
month_12_dist$RideLength <- as_hms(month_12_dist$RideLength)

#Finding which day it was from the date
month_12_dist$StartDay <- weekdays(as.Date(month_12_dist$StartDate))
month_12_dist$EndDay <- weekdays(as.Date(month_12_dist$EndDate))

#making a new dataset with the variables required for the analysis
month_12_clean_analysis <- subset(month_12_dist, select = c(RideId,RideableType,MemberCasual,StartDate,EndDate,StartTime,EndTime,RideLength,StartDay,EndDay))
View(month_12_clean_analysis)
```

Merging past 12 Months Data

After individually cleaning the months data, they were merged and the new dataset was named One_year_data

After merging, the dataset was again cleaned again to remove duplicate entries, just in case new duplicates were formed during merging.

While running through the data there were certain RideLengths that were negative, turns out for some entries, the start time was greater than end time, Rows like these were dropped

Another column-MonthRide that specifies the month the ride took place was added

This final data was saved as One_year_final.csv.

```
#merging past 12 months data
One_year_data <- rbind(month_1_clean_analysis, month_2_clean_analysis,
month_3_clean_analysis, month_4_clean_analysis, month_5_clean_analysis,
month_6_clean_analysis, month_7_clean_analysis, month_8_clean_analysis,
month_9_clean_analysis, month_10_clean_analysis, month_11_clean_analysis,
month_12_clean_analysis)
str(One_year_data)
#in case there were any duplicate entries, only considering unique entries
One_year_data <- distinct(One_year_data)
#adding a column that specifies the month the ride took place
One_year_data <- mutate(One_year_data, MonthRide = months(StartDate))
#removing rows with negative ride length
One_year_data <- One_year_data %>% filter(StartTime<=EndTime)
fwrite(One_year_data,
       file = "/Users/Anita Dash/Desktop/Google-
Capstone/One_year_data_final.csv",
       sep = ",")
```