DA5020.A3.Hsiao-Yu.Peng

Hsiao-Yu Peng

2023-09-24

Q1 See the attached certificate.

Q2

```
# load NYC Green Taxi Trip Record
# tripdata_df <- read_csv("~/Desktop/2023Fall_Syllabus/DA5020/week3/2018_Green_Taxi_Trip_Data.csv")
tripdata_df<-read.delim2("~/Desktop/2023Fall_Syllabus/DA5020/week3/2018_Green_Taxi_Trip_Data.csv", head

# dataset overview
dim(tripdata_df)
```

```
## [1] 1048575      19
```

```
glimpse(tripdata_df)
```

```
## Rows: 1,048,575
## Columns: 19
## $ VendorID              <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, ~
## $ lpep_pickup_datetime  <chr> "1/1/2018 0:18", "1/1/2018 0:30", "1/1/2018 0:07~
## $ lpep_dropoff_datetime <chr> "1/1/2018 0:24", "1/1/2018 0:46", "1/1/2018 0:19~
## $ store_and_fwd_flag    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"~
## $ RatecodeID            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ PULocationID          <int> 236, 43, 74, 255, 255, 255, 189, 189, 129, 226, ~
## $ DOLocationID          <int> 236, 42, 152, 255, 255, 161, 65, 225, 82, 7, 129~
## $ passenger_count       <int> 5, 5, 1, 1, 1, 1, 5, 5, 1, 1, 2, 2, 1, 1, 2, 1, ~
## $ trip_distance         <chr> "0.7", "3.5", "2.14", "0.03", "0.03", "5.63", "1~
## $ fare_amount           <chr> "6", "14.5", "10", "-3", "3", "21", "8.5", "14.5~
## $ extra                 <chr> "0.5", "0.5", "0.5", "-0.5", "0.5", "0.5", "0.5"~
## $ mta_tax               <chr> "0.5", "0.5", "0.5", "-0.5", "0.5", "0.5", "0.5"~
## $ tip_amount            <chr> "0", "0", "0", "0", "0", "0", "0", "3.16", "0", ~
## $ tolls_amount          <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
## $ ehail_fee             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ improvement_surcharge <chr> "0.3", "0.3", "0.3", "-0.3", "0.3", "0.3", "0.3"~
## $ total_amount          <chr> "7.3", "15.8", "11.3", "-4.3", "4.3", "22.3", "9~
## $ payment_type          <int> 2, 2, 2, 3, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, ~
## $ trip_type             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

```
summary(tripdata_df)
```

```
##     VendorID     lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag
##  Min.   :1.000   Length:1048575       Length:1048575        Length:1048575
##  1st Qu.:2.000   Class :character     Class :character      Class :character
##  Median :2.000   Mode  :character     Mode  :character      Mode  :character
##  Mean   :1.827
##  3rd Qu.:2.000
##  Max.   :2.000
##
##    RatecodeID      PULocationID  DOLocationID   passenger_count
##  Min.   : 1.000   Min.   :  1   Min.   :  1.0   Min.   :0.000
##  1st Qu.: 1.000   1st Qu.: 49   1st Qu.: 61.0   1st Qu.:1.000
##  Median : 1.000   Median : 82   Median :129.0   Median :1.000
##  Mean   : 1.072   Mean   :110   Mean   :128.5   Mean   :1.359
##  3rd Qu.: 1.000   3rd Qu.:166   3rd Qu.:191.0   3rd Qu.:1.000
##  Max.   :99.000   Max.   :265   Max.   :265.0   Max.   :9.000
##
##  trip_distance      fare_amount          extra             mta_tax
##  Length:1048575     Length:1048575     Length:1048575     Length:1048575
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   tip_amount        tolls_amount       ehail_fee      improvement_surcharge
##  Length:1048575     Length:1048575     Mode:logical   Length:1048575
##  Class :character   Class :character   NA's:1048575   Class :character
##  Mode  :character   Mode  :character                  Mode  :character
##
##
##
##
##  total_amount        payment_type     trip_type
##  Length:1048575     Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:1.000   1st Qu.:1.000
##  Mode  :character   Median :1.000   Median :1.000
##                     Mean   :1.471   Mean   :1.018
##                     3rd Qu.:2.000   3rd Qu.:1.000
##                     Max.   :5.000   Max.   :2.000
##                                     NA's   :3
```

```
# Check any NA values in the dataset
# sapply(tripdata_df, function(x) sum(is.na(x)))
```

This dataset has 19 variables and 1,048,575 observations, meaning it has 1,048,575 rows and 19 columns.

Additionally, it includes logical data for "ehai_fee," although "ehai_fee" has more than a 90% missing value. "total_amount" and "trip_type" have fewer missing values, and we will address these missing values in question 7.

Q3.

```r
#Convert character to date.time type
tripdata_df <- tripdata_df %>%
  mutate(
    across(
      c(lpep_pickup_datetime, lpep_dropoff_datetime),
      ~ as.POSIXct(., format = "%m/%d/%Y %H:%M")
    )
  )

summary(tripdata_df)
```

```
##     VendorID     lpep_pickup_datetime
##  Min.   :1.000   Min.   :2009-01-01 00:02:00.00
##  1st Qu.:2.000   1st Qu.:2018-01-12 04:21:30.00
##  Median :2.000   Median :2018-01-21 18:47:00.00
##  Mean   :1.827   Mean   :2018-01-21 17:48:24.42
##  3rd Qu.:2.000   3rd Qu.:2018-01-31 18:58:00.00
##  Max.   :2.000   Max.   :2018-04-05 04:11:00.00
##
##  lpep_dropoff_datetime            store_and_fwd_flag   RatecodeID
##  Min.   :2009-01-01 11:16:00.00   Length:1048575     Min.   : 1.000
##  1st Qu.:2018-01-12 05:11:00.00   Class :character   1st Qu.: 1.000
##  Median :2018-01-21 19:07:00.00   Mode  :character   Median : 1.000
##  Mean   :2018-01-21 18:08:24.00                      Mean   : 1.072
##  3rd Qu.:2018-01-31 19:14:00.00                      3rd Qu.: 1.000
##  Max.   :2018-04-05 04:25:00.00                      Max.   :99.000
##
##   PULocationID  DOLocationID   passenger_count trip_distance
##  Min.   :  1   Min.   :  1.0   Min.   :0.000   Length:1048575
##  1st Qu.: 49   1st Qu.: 61.0   1st Qu.:1.000   Class :character
##  Median : 82   Median :129.0   Median :1.000   Mode  :character
##  Mean   :110   Mean   :128.5   Mean   :1.359
##  3rd Qu.:166   3rd Qu.:191.0   3rd Qu.:1.000
##  Max.   :265   Max.   :265.0   Max.   :9.000
##
##  fare_amount          extra              mta_tax           tip_amount
##  Length:1048575     Length:1048575     Length:1048575     Length:1048575
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  tolls_amount       ehail_fee      improvement_surcharge total_amount
##  Length:1048575     Mode:logical   Length:1048575        Length:1048575
##  Class :character   NA's:1048575   Class :character      Class :character
##  Mode  :character                  Mode  :character      Mode  :character
##
##
##
##
##   payment_type     trip_type
##  Min.   :1.000   Min.   :1.000
```

```
##  1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :1.000
##  Mean   :1.471   Mean   :1.018
##  3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :5.000   Max.   :2.000
##                  NA's   :3
```

a) Most variables are numerical, but date-related variables, such as "lpep_pickup_datetime" and "lpep_dropoff_datetime", are character type. We convert these character data types to datetime type.

b) Since the dataset contains numerical variables related to taxes, tips, charges, etc., the numbers should be positive or zero. However, some variables have minimum values that are negative, which should be unreasonable. These variables include "fare_amount," "extra," "mta_tax," "tip_amount," "improvement_surcharge," and "total_amount."

c) This is 2018 dataset. But datetime variables has information that was not in 2018, it's inconsistent data.

d) payment_type are ordinal values, so we can consider converting their data type as character type.

Q4.

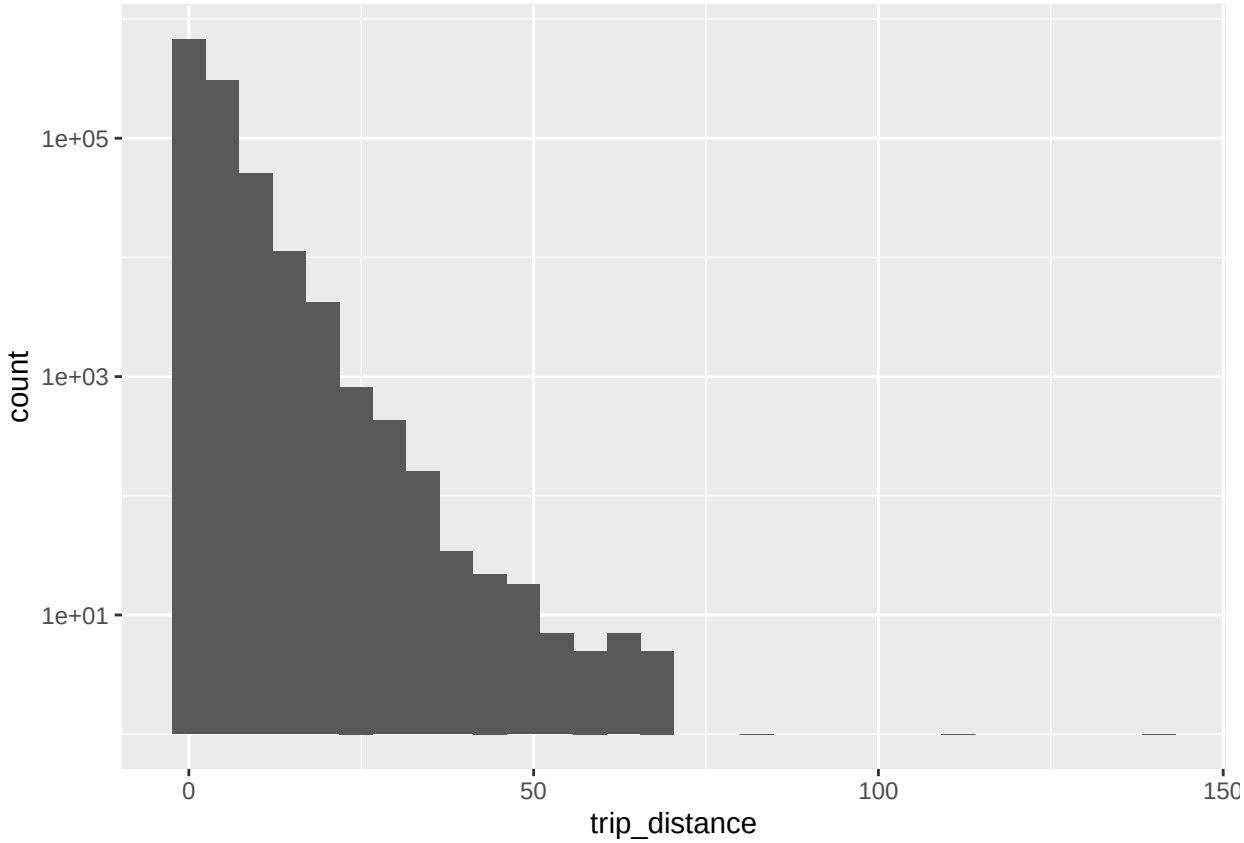
```r
# Convert trip_distance to numeric if needed
tripdata_df$trip_distance <- as.numeric(tripdata_df$trip_distance)

# Create histogram plot
ggplot(tripdata_df, aes(trip_distance)) +
  geom_histogram() +
  scale_y_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 12 rows containing missing values (`geom_bar()`).
```



The histogram displays a right-skewed distribution, with most of the "trip_distance" values being less than 3~5 km (left side). However, it is important to note that having many "0" values for "trip_distance" is not reasonable. The "0" values in the trip_distance data may be invalid.

Q5 remove outlier

```
# Calculate mean and standard deviation of tip_amount and trip_distance
mean_tip_amount <- mean(tripdata_df$tip_amount)
```

```
## Warning in mean.default(tripdata_df$tip_amount): argument is not numeric or
## logical: returning NA
```

```
sd_tip_amount  <- sd(tripdata_df$tip_amount)

mean_trip_distance <- mean(tripdata_df$trip_distance)
sd_trip_distance  <- sd(tripdata_df$trip_distance)


# Identify outliers of tip_amount
outliers_tip_amount <- tripdata_df$tip_amount > (mean_tip_amount + 3*sd_tip_amount) | tripdata_df$tip_a
sum(tripdata_df$tip_amount > (mean_tip_amount + 3*sd_tip_amount))
```

```
## [1] NA
```

```r
sum(tripdata_df$tip_amount < (mean_tip_amount - 3 * sd_tip_amount))
```

```
## [1] NA
```

```r
# total number of outliers in tip_amount
total_tip_outliers <- sum(tripdata_df$tip_amount > (mean_tip_amount + 3*sd_tip_amount)) +
sum(tripdata_df$tip_amount < (mean_tip_amount - 3 * sd_tip_amount))
cat("The number of outlier in tip_amount is" ,total_tip_outliers)
```

```
## The number of outlier in tip_amount is NA
```

```r
# Identify outliers of trip_distance
outliers_trip_distance <- tripdata_df$trip_distance > (mean_trip_distance + 3*sd_trip_distance) | tripda
sum(tripdata_df$trip_distance > (mean_trip_distance + 3*sd_trip_distance))
```

```
## [1] 20953
```

```r
sum(tripdata_df$trip_distance < (mean_trip_distance - 3 * sd_trip_distance))
```

```
## [1] 0
```

```r
# total number of outliers in tip_amount
total_tripDist_outliers <- sum(tripdata_df$trip_distance > (mean_trip_distance + 3*sd_trip_distance)) +
sum(tripdata_df$trip_distance < (mean_trip_distance - 3 * sd_trip_distance))
cat("The number of outlier in trip_distance is" ,total_tripDist_outliers)
```

```
## The number of outlier in trip_distance is 20953
```

```r
# Remove outlier of tip_amount in the dataset
tip_amount_clean <- tripdata_df[!outliers_tip_amount, ]
head(tip_amount_clean, 3)
```

```
##      VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag
## NA         NA                 <NA>                  <NA>               <NA>
## NA.1       NA                 <NA>                  <NA>               <NA>
## NA.2       NA                 <NA>                  <NA>               <NA>
##      RatecodeID PULocationID DOLocationID passenger_count trip_distance
## NA           NA           NA           NA              NA            NA
## NA.1         NA           NA           NA              NA            NA
## NA.2         NA           NA           NA              NA            NA
##      fare_amount extra mta_tax tip_amount tolls_amount ehail_fee
## NA          <NA>  <NA>    <NA>       <NA>         <NA>        NA
## NA.1        <NA>  <NA>    <NA>       <NA>         <NA>        NA
## NA.2        <NA>  <NA>    <NA>       <NA>         <NA>        NA
##      improvement_surcharge total_amount payment_type trip_type
## NA                    <NA>         <NA>           NA        NA
## NA.1                  <NA>         <NA>           NA        NA
## NA.2                  <NA>         <NA>           NA        NA
```

```r
dim(tip_amount_clean)
```

```
## [1] 1048575      19
```

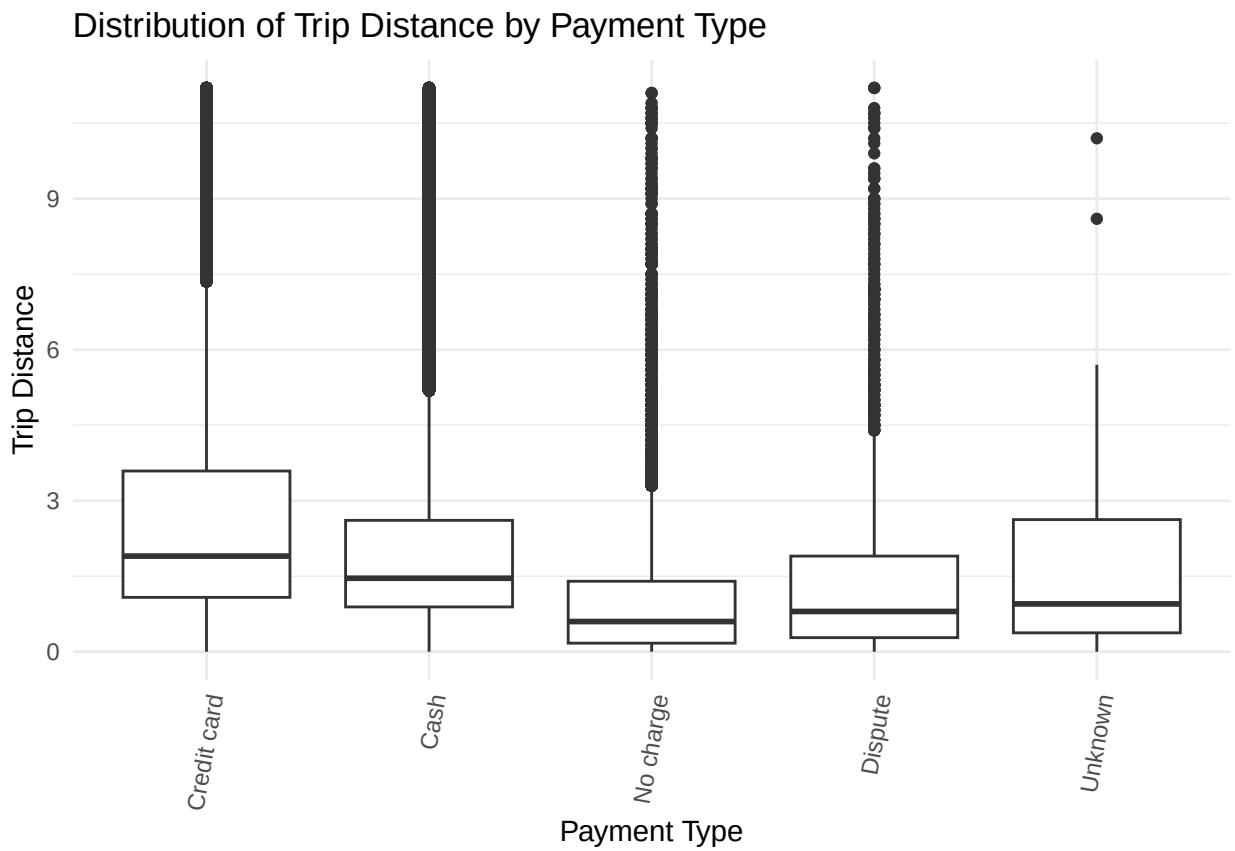
The number of outlier "tip_amount" is 14444, and the number of outlier "trip_distance" variable is 20953. After removing the outlier tip_amount from the data, the dataset dimension becomes 1034131 x 19.

Q6

```r
# Remove outlier trip_distance in the dataset
trip_distance_clean <- tripdata_df[!outliers_trip_distance, ]


# Create a factor variable for payment_type with corresponding labels
trip_distance_clean$payment_type_label <- factor(trip_distance_clean$payment_type,
  levels = c(1, 2, 3, 4, 5, 6),
  labels = c("Credit card", "Cash", "No charge", "Dispute", "Unknown", "Voided trip")
)


# Draw boxplot
ggplot(trip_distance_clean, aes(x = payment_type_label, y = trip_distance)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Trip Distance by Payment Type",
    x = "Payment Type",
    y = "Trip Distance"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 80, hjust = 1))
```



Q7 two methods to handle missing data

If there are any missing values in the dataset, we can choose to remove them, impute them, or encode them to handle the missing data. If the variable has more than 10% missing values, I will consider removing it. If the variable is missing approximately 5% to 10% of its values, I will consider imputing the missing data with the average value.

```
# Check any NA values in the dataset
sapply(tripdata_df, function(x) sum(is.na(x)))
```

```
##              VendorID  lpep_pickup_datetime lpep_dropoff_datetime
##                     0                     0                     0
##    store_and_fwd_flag            RatecodeID          PULocationID
##                     0                     0                     0
##          DOLocationID       passenger_count         trip_distance
##                     0                     0                     0
##           fare_amount                 extra               mta_tax
##                     0                     0                     0
##            tip_amount          tolls_amount             ehail_fee
##                     0                     0               1048575
## improvement_surcharge          total_amount          payment_type
##                     0                     0                     0
##             trip_type
##                     3
```

There are 3 variables, "total_amount", "ehail_fee", and "trip_type", have missing values. "total_amount" and "trip_type" have less than 5% missing values. We can impute it with mean, median value or zero. Most values in "ehail_fee" variables are missing, so we can remove it.

```
# impute mean value in "total_amount" column
tripdata_df <- tripdata_df %>%
  replace_na(list(total_amount = mean(.$total_amount, na.rm = TRUE)))
```

```
## Warning in mean.default(.$total_amount, na.rm = TRUE): argument is not numeric
## or logical: returning NA
```

```
# impute mean value in "trip_type" column
tripdata_df <- tripdata_df %>%
  replace_na(list(trip_type = median(.$trip_type, na.rm = TRUE)))


# As 90% missing value in "ehail_fee", it's hard to impute and encode values in this column. So we igno
tripdata_df <- tripdata_df %>%
  select(-ehail_fee)


# Check the dataset again
summary(tripdata_df)
```

```
##     VendorID     lpep_pickup_datetime
##  Min.   :1.000   Min.   :2009-01-01 00:02:00.00
##  1st Qu.:2.000   1st Qu.:2018-01-12 04:21:30.00
##  Median :2.000   Median :2018-01-21 18:47:00.00
##  Mean   :1.827   Mean   :2018-01-21 17:48:24.42
##  3rd Qu.:2.000   3rd Qu.:2018-01-31 18:58:00.00
##  Max.   :2.000   Max.   :2018-04-05 04:11:00.00
##  lpep_dropoff_datetime           store_and_fwd_flag   RatecodeID
##  Min.   :2009-01-01 11:16:00.00   Length:1048575     Min.   : 1.000
##  1st Qu.:2018-01-12 05:11:00.00   Class :character   1st Qu.: 1.000
##  Median :2018-01-21 19:07:00.00   Mode  :character   Median : 1.000
##  Mean   :2018-01-21 18:08:24.00                      Mean   : 1.072
##  3rd Qu.:2018-01-31 19:14:00.00                      3rd Qu.: 1.000
##  Max.   :2018-04-05 04:25:00.00                      Max.   :99.000
##   PULocationID  DOLocationID   passenger_count trip_distance
##  Min.   :  1   Min.   :  1.0   Min.   :0.000   Min.   :  0.000
##  1st Qu.: 49   1st Qu.: 61.0   1st Qu.:1.000   1st Qu.:  0.990
##  Median : 82   Median :129.0   Median :1.000   Median :  1.700
##  Mean   :110   Mean   :128.5   Mean   :1.359   Mean   :  2.662
##  3rd Qu.:166   3rd Qu.:191.0   3rd Qu.:1.000   3rd Qu.:  3.260
##  Max.   :265   Max.   :265.0   Max.   :9.000   Max.   :140.620
##  fare_amount          extra             mta_tax          tip_amount
##  Length:1048575     Length:1048575     Length:1048575     Length:1048575
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  tolls_amount       improvement_surcharge total_amount        payment_type
```

```
##  Length:1048575     Length:1048575        Length:1048575     Min.   :1.000
##  Class :character   Class :character      Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character      Mode  :character   Median :1.000
##                                                              Mean   :1.471
##                                                              3rd Qu.:2.000
##                                                              Max.   :5.000
##    trip_type
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.018
##  3rd Qu.:1.000
##  Max.   :2.000
```

```
dim(tripdata_df)
```

```
## [1] 1048575      18
```

We imputed missing values in the 'total_amount' and 'trip_type' variables. Additionally, we removed '-ehail_fee,' which had 90% missing values. In summary, there are no remaining missing values, and the dataset now consists of 18 variables.