705604096_stats101a_hw1

Jade Gregory

2023-04-05

Jade Gregory 705604096

Question 1

```
df <- read_csv('chicagotaxiraw.csv')
```

```
## Rows: 100000 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (11): Trip ID, Taxi ID, Fare, Tips, Tolls, Extras, Trip Total, Payment ...
## dbl  (10): Trip Seconds, Trip Miles, Pickup Census Tract, Dropoff Census Tra...
## dttm  (2): Trip Start Timestamp, Trip End Timestamp
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
df
```

```
## # A tibble: 100,000 x 23
##    ‘Trip ID‘     Taxi ~1 Trip Start Timest~2 Trip End Timestam~3 Trip ~4 Trip ~5
##    <chr>         <chr>   <dttm>              <dttm>                <dbl>   <dbl>
##  1 2f946572a1f4~ 626cdd~ 2014-11-13 11:00:00 2014-11-13 11:00:00     240     0
##  2 340c309a4375~ da1882~ 2014-05-15 06:45:00 2014-05-15 07:00:00     480     2.2
##  3 3a1a1f336266~ db337a~ 2016-04-17 03:00:00 2016-04-17 03:15:00     480     2.3
##  4 39018c6704e7~ ef4143~ 2014-11-05 07:45:00 2014-11-05 08:00:00     780     2.8
##  5 40d518fdeedc~ dcb626~ 2014-10-14 21:30:00 2014-10-14 21:45:00    1140     5.7
##  6 3825c467c1b4~ af3b4b~ 2013-04-10 06:15:00 2013-04-10 06:45:00    1380    18.2
##  7 3df87a82d469~ 386ace~ 2015-11-24 19:00:00 2015-11-24 19:30:00    1200     8.6
##  8 3270aa9ee4ac~ 94cb43~ 2014-06-21 04:15:00 2014-06-21 04:30:00     720     5.7
##  9 2a61add9211e~ 589abe~ 2014-10-29 22:00:00 2014-10-29 22:00:00     420     1.5
## 10 29d3a51c5422~ 429edc~ 2015-11-04 20:45:00 2015-11-04 21:00:00    1020     6
## # ... with 99,990 more rows, 17 more variables: ‘Pickup Census Tract‘ <dbl>,
## #   ‘Dropoff Census Tract‘ <dbl>, ‘Pickup Community Area‘ <dbl>,
## #   ‘Dropoff Community Area‘ <dbl>, Fare <chr>, Tips <chr>, Tolls <chr>,
## #   Extras <chr>, ‘Trip Total‘ <chr>, ‘Payment Type‘ <chr>, Company <chr>,
## #   ‘Pickup Centroid Latitude‘ <dbl>, ‘Pickup Centroid Longitude‘ <dbl>,
## #   ‘Pickup Centroid Location‘ <chr>, ‘Dropoff Centroid Latitude‘ <dbl>,
## #   ‘Dropoff Centroid Longitude‘ <dbl>, ‘Dropoff Centroid  Location‘ <chr>, ...
```

```
df_new <- df[,c(1:6,11:17)]
df_new
```

```
## # A tibble: 100,000 x 13
##    Trip ~1 Taxi ~2 Trip Start Timest~3 Trip End Timestam~4 Trip ~5 Trip ~6 Fare
##    <chr>   <chr>   <dttm>              <dttm>                <dbl>   <dbl> <chr>
##  1 2f9465~ 626cdd~ 2014-11-13 11:00:00 2014-11-13 11:00:00     240     0   $4.65
##  2 340c30~ da1882~ 2014-05-15 06:45:00 2014-05-15 07:00:00     480     2.2 $7.65
##  3 3a1a1f~ db337a~ 2016-04-17 03:00:00 2016-04-17 03:15:00     480     2.3 $9.50
##  4 39018c~ ef4143~ 2014-11-05 07:45:00 2014-11-05 08:00:00     780     2.8 $10.~
##  5 40d518~ dcb626~ 2014-10-14 21:30:00 2014-10-14 21:45:00    1140     5.7 $15.~
##  6 3825c4~ af3b4b~ 2013-04-10 06:15:00 2013-04-10 06:45:00    1380    18.2 $36.~
##  7 3df87a~ 386ace~ 2015-11-24 19:00:00 2015-11-24 19:30:00    1200     8.6 $20.~
##  8 3270aa~ 94cb43~ 2014-06-21 04:15:00 2014-06-21 04:30:00     720     5.7 $14.~
##  9 2a61ad~ 589abe~ 2014-10-29 22:00:00 2014-10-29 22:00:00     420     1.5 $6.65
## 10 29d3a5~ 429edc~ 2015-11-04 20:45:00 2015-11-04 21:00:00    1020     6   $15.~
## # ... with 99,990 more rows, 6 more variables: Tips <chr>, Tolls <chr>,
## #   Extras <chr>, ‘Trip Total‘ <chr>, ‘Payment Type‘ <chr>, Company <chr>, and
## #   abbreviated variable names 1: ‘Trip ID‘, 2: ‘Taxi ID‘,
## #   3: ‘Trip Start Timestamp‘, 4: ‘Trip End Timestamp‘, 5: ‘Trip Seconds‘,
## #   6: ‘Trip Miles‘
```

```
df_final <- rename(df_new, c(trip_id = 'Trip ID', taxi_id = 'Taxi ID',
                             trip_start_timestamp = 'Trip Start Timestamp',
                             trip_end_timestamp = 'Trip End Timestamp',
                             trip_seconds = 'Trip Seconds', trip_miles = 'Trip Miles',
                             fare = Fare, tips = Tips, tolls = Tolls, extras = Extras,
                             trip_total = 'Trip Total', payment_type = 'Payment Type',
                             company = Company))
df_final
```

```
## # A tibble: 100,000 x 13
##    trip_id taxi_id trip_start_timest~1 trip_end_timestamp  trip_~2 trip_~3 fare
##    <chr>   <chr>   <dttm>              <dttm>                <dbl>   <dbl> <chr>
##  1 2f9465~ 626cdd~ 2014-11-13 11:00:00 2014-11-13 11:00:00     240     0   $4.65
##  2 340c30~ da1882~ 2014-05-15 06:45:00 2014-05-15 07:00:00     480     2.2 $7.65
##  3 3a1a1f~ db337a~ 2016-04-17 03:00:00 2016-04-17 03:15:00     480     2.3 $9.50
##  4 39018c~ ef4143~ 2014-11-05 07:45:00 2014-11-05 08:00:00     780     2.8 $10.~
##  5 40d518~ dcb626~ 2014-10-14 21:30:00 2014-10-14 21:45:00    1140     5.7 $15.~
##  6 3825c4~ af3b4b~ 2013-04-10 06:15:00 2013-04-10 06:45:00    1380    18.2 $36.~
##  7 3df87a~ 386ace~ 2015-11-24 19:00:00 2015-11-24 19:30:00    1200     8.6 $20.~
##  8 3270aa~ 94cb43~ 2014-06-21 04:15:00 2014-06-21 04:30:00     720     5.7 $14.~
##  9 2a61ad~ 589abe~ 2014-10-29 22:00:00 2014-10-29 22:00:00     420     1.5 $6.65
## 10 29d3a5~ 429edc~ 2015-11-04 20:45:00 2015-11-04 21:00:00    1020     6   $15.~
## # ... with 99,990 more rows, 6 more variables: tips <chr>, tolls <chr>,
## #   extras <chr>, trip_total <chr>, payment_type <chr>, company <chr>, and
## #   abbreviated variable names 1: trip_start_timestamp, 2: trip_seconds,
## #   3: trip_miles
```

Question 2

```
min(df_final$trip_start_timestamp)
```

```
## [1] "2013-01-01 00:30:00 UTC"
```

```
max(df_final$trip_end_timestamp, na.rm = TRUE)
```

```
## [1] "2017-05-31 23:30:00 UTC"
```

This data ranges from 1/1/2013 to 5/31/2017. The date of the first pickup is 1/1/2013 and the date of the last drop-off is 5/31/2017.

Question 3

```
df_final$weekday <- wday(df_final$trip_start_timestamp, week_start = 1)
df_final
```

```
## # A tibble: 100,000 x 14
##    trip_id taxi_id trip_start_timest~1 trip_end_timestamp  trip_~2 trip_~3 fare 
##    <chr>   <chr>   <dttm>              <dttm>                <dbl>   <dbl> <chr>
##  1 2f9465~ 626cdd~ 2014-11-13 11:00:00 2014-11-13 11:00:00     240     0   $4.65
##  2 340c30~ da1882~ 2014-05-15 06:45:00 2014-05-15 07:00:00     480     2.2 $7.65
##  3 3a1a1f~ db337a~ 2016-04-17 03:00:00 2016-04-17 03:15:00     480     2.3 $9.50
##  4 39018c~ ef4143~ 2014-11-05 07:45:00 2014-11-05 08:00:00     780     2.8 $10.~
##  5 40d518~ dcb626~ 2014-10-14 21:30:00 2014-10-14 21:45:00    1140     5.7 $15.~
##  6 3825c4~ af3b4b~ 2013-04-10 06:15:00 2013-04-10 06:45:00    1380    18.2 $36.~
##  7 3df87a~ 386ace~ 2015-11-24 19:00:00 2015-11-24 19:30:00    1200     8.6 $20.~
##  8 3270aa~ 94cb43~ 2014-06-21 04:15:00 2014-06-21 04:30:00     720     5.7 $14.~
##  9 2a61ad~ 589abe~ 2014-10-29 22:00:00 2014-10-29 22:00:00     420     1.5 $6.65
## 10 29d3a5~ 429edc~ 2015-11-04 20:45:00 2015-11-04 21:00:00    1020     6   $15.~
## # ... with 99,990 more rows, 7 more variables: tips <chr>, tolls <chr>,
## #   extras <chr>, trip_total <chr>, payment_type <chr>, company <chr>,
## #   weekday <dbl>, and abbreviated variable names 1: trip_start_timestamp,
## #   2: trip_seconds, 3: trip_miles
```

```
df_mon <- filter(df_final, weekday == 1)
length(df_mon$weekday)
```

```
## [1] 12406
```

```
df_tue <- filter(df_final, weekday == 2)
length(df_tue$weekday)
```

```
## [1] 13196
```

```
df_wed <- filter(df_final, weekday == 3)
length(df_wed$weekday)
```

```
## [1] 13896
```

```
df_thur <- filter(df_final, weekday == 4)
length(df_thur$weekday)
```

```
## [1] 14873
```

```
df_fri <- filter(df_final, weekday == 5)
length(df_fri$weekday)
```

```
## [1] 16767
```

```
df_sat <- filter(df_final, weekday == 6)
length(df_sat$weekday)
```

```
## [1] 15859
```

```
df_sun <- filter(df_final, weekday == 7)
length(df_sun$weekday)
```

```
## [1] 13003
```

From our output, we can see that Friday is the day of the week to have the highest number of trips begin with a total of 16767 trips began throughout this data set. Monday is the day of the week with the least amount of trips started with a total of 12406 trips started throughout this data set.

Question 4

```
tips <- df_final$tips
tips <- gsub("\\$", "", tips)
tips <- as.numeric(tips)
mean(tips, na.rm = TRUE)
```

```
## [1] 1.154516
```

```
median(tips)
```

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

The typical value for the tip is $0. The mean of the tip amount is better to use since it is an average of how much people tip the taxis.