Cyclistic Bike Share Analysis

Jake Isaacs

1/10/2022

Analysis

```
library(tidyverse)
```

Loading the necessary packages

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: ’janitor’

## The following objects are masked from ’package:stats’:
##
##     chisq.test, fisher.test
```

```
full_year <- read_csv("fullyear-tripdata-R-v3.csv")
```

Loading the data

```
## New names:
## * ‘‘ -> ...1
## Rows: 12 Columns: 16

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): ...1
## dbl (15): max_ride_length, avg_ride_sun, avg_ride_mon, avg_ride_tues, avg_ri...

##
## 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.
```

```
head(full_year)
```

Quick review of the data

```
## # A tibble: 6 x 16
##   ...1   max_ride_length avg_ride_sun avg_ride_mon avg_ride_tues avg_ride_wed
##   <chr>            <dbl>        <dbl>        <dbl>         <dbl>        <dbl>
## 1 20-Dec           9741.         22.4         14.4          13.8         14.6
## 2 21-Jan          19826.         17.2         14.4          13.2         15.4
## 3 21-Feb          30129.         27.1         21.9          22.2         19.8
## 4 21-Mar          31682.         28.4         24.5          20.4         17.4
## 5 21-Apr          47777.         29.5         22.8          24.2         21.4
## 6 21-May          53922.         34.5         25.0          19.2         20.2
## # ... with 10 more variables: avg_ride_thurs <dbl>, avg_ride_fri <dbl>,
## #   avg_ride_sat <dbl>, mean_ride_length <dbl>, mode_day_of_week <dbl>,
## #   count_rides_casual <dbl>, count_rides_member <dbl>, total_rides <dbl>,
## #   avg_ride_casual <dbl>, avg_ride_member <dbl>
```

```
colnames(full_year)
```

```
##  [1] "...1"               "max_ride_length"    "avg_ride_sun"
##  [4] "avg_ride_mon"       "avg_ride_tues"      "avg_ride_wed"
##  [7] "avg_ride_thurs"     "avg_ride_fri"       "avg_ride_sat"
## [10] "mean_ride_length"   "mode_day_of_week"   "count_rides_casual"
## [13] "count_rides_member" "total_rides"        "avg_ride_casual"
## [16] "avg_ride_member"
```

```
skim_without_charts(full_year)
```

Table 1: Data summary

Name	full_year
Number of rows	12
Number of columns	16
Column type frequency:	
character	1

Table 1: Data summary

numeric	15
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
...1	0	1	6	6	0	12	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
max_ride_length	0	1	37359.82	13738.47	9740.98	31293.55	37851.37	48109.31	55944.15
avg_ride_sun	0	1	26.50	5.00	17.23	25.06	26.73	29.33	34.47
avg_ride_mon	0	1	20.13	4.28	14.03	15.97	21.20	23.20	25.62
avg_ride_tues	0	1	18.17	3.97	12.42	14.78	18.78	20.86	24.25
avg_ride_wed	0	1	17.95	3.06	12.75	15.33	17.94	20.22	22.73
avg_ride_thurs	0	1	17.50	3.20	13.23	15.27	16.78	19.75	23.60
avg_ride_fri	0	1	20.21	4.41	14.10	17.00	20.31	23.79	26.02
avg_ride_sat	0	1	25.53	5.29	17.63	22.14	25.98	29.00	33.87
mean_ride_length	0	1	21.25	4.11	14.82	18.32	22.25	24.27	26.08
mode_day_of_week	0	1	5.83	2.04	1.00	5.50	7.00	7.00	7.00
count_rides_casual	0	1	207438.67	162320.32	10131.00	70524.00	196758.50	365587.75	442056.00
count_rides_member	0	1	249116.50	132984.63	39491.00	133632.75	263883.00	375576.50	392257.00
total_rides	0	1	456555.17	291037.52	49622.00	204156.75	445805.50	736233.00	822410.00
avg_ride_casual	0	1	32.88	7.49	23.12	27.58	30.78	38.06	49.37
avg_ride_member	0	1	13.96	1.65	11.30	12.84	14.04	14.64	18.02

```
glimpse(full_year)
```

```
## Rows: 12
## Columns: 16
## $ ...1               <chr> "20-Dec", "21-Jan", "21-Feb", "21-Mar", "21-Apr", "~
## $ max_ride_length    <dbl> 9740.98, 19825.92, 30129.23, 31681.65, 47776.70, 53~
## $ avg_ride_sun       <dbl> 22.35, 17.23, 27.10, 28.43, 29.52, 34.47, 32.10, 29~
## $ avg_ride_mon       <dbl> 14.40, 14.37, 21.90, 24.47, 22.78, 24.98, 21.17, 25~
## $ avg_ride_tues      <dbl> 13.77, 13.22, 22.17, 20.42, 24.25, 19.15, 22.62, 20~
## $ avg_ride_wed       <dbl> 14.62, 15.37, 19.85, 17.37, 21.37, 20.15, 22.73, 20~
## $ avg_ride_thurs     <dbl> 15.37, 13.80, 16.37, 16.73, 16.83, 20.80, 23.60, 21~
## $ avg_ride_fri       <dbl> 14.83, 14.27, 25.67, 17.73, 24.77, 23.47, 26.02, 22~
## $ avg_ride_sat       <dbl> 17.63, 18.02, 33.87, 28.77, 26.80, 29.67, 31.75, 27~
## $ mean_ride_length   <dbl> 15.97, 15.27, 24.42, 22.87, 24.13, 26.03, 26.08, 24~
## $ mode_day_of_week   <dbl> 4, 7, 7, 7, 6, 7, 7, 7, 1, 7, 7, 3
## $ count_rides_casual <dbl> 29997, 18117, 10131, 84033, 136601, 256916, 370681,~
## $ count_rides_member <dbl> 101142, 78717, 39491, 144463, 200629, 274717, 35891~
## $ total_rides        <dbl> 131139, 96834, 49622, 228496, 337230, 531633, 72959~
## $ avg_ride_casual    <dbl> 26.85, 25.68, 49.37, 38.17, 38.02, 38.23, 37.12, 32~
## $ avg_ride_member    <dbl> 12.75, 12.87, 18.02, 13.97, 14.68, 14.63, 14.68, 14~
```

```
clean_names(full_year)
```

Making sure that the column names are unique and consistent

```
## # A tibble: 12 x 16
##    x1     max_ride_length avg_ride_sun avg_ride_mon avg_ride_tues avg_ride_wed
##    <chr>            <dbl>        <dbl>        <dbl>         <dbl>        <dbl>
##  1 20-Dec           9741.         22.4         14.4          13.8         14.6
##  2 21-Jan          19826.         17.2         14.4          13.2         15.4
##  3 21-Feb          30129.         27.1         21.9          22.2         19.8
##  4 21-Mar          31682.         28.4         24.5          20.4         17.4
##  5 21-Apr          47777.         29.5         22.8          24.2         21.4
##  6 21-May          53922.         34.5         25.0          19.2         20.2
##  7 21-Jun          55944.         32.1         21.2          22.6         22.7
##  8 21-Jul          49107.         29.3         25.6          20.2         20.4
##  9 21-Aug          41629.         26.2         20.2          18.4         18.5
## 10 21-Sep          32859.         26.4         21.2          16.2         17.0
## 11 21-Oct          40705.         26.0         16.5          15.1         15.2
## 12 21-Nov          34998.         19.0         14.0          12.4         12.8
## # ... with 10 more variables: avg_ride_thurs <dbl>, avg_ride_fri <dbl>,
## #   avg_ride_sat <dbl>, mean_ride_length <dbl>, mode_day_of_week <dbl>,
## #   count_rides_casual <dbl>, count_rides_member <dbl>, total_rides <dbl>,
## #   avg_ride_casual <dbl>, avg_ride_member <dbl>
```

```
full_year %>%
  rename(months = ...1)
```

Renaming the first column for better clarity

```
## # A tibble: 12 x 16
##    months max_ride_length avg_ride_sun avg_ride_mon avg_ride_tues avg_ride_wed
##    <chr>            <dbl>        <dbl>        <dbl>         <dbl>        <dbl>
##  1 20-Dec           9741.         22.4         14.4          13.8         14.6
##  2 21-Jan          19826.         17.2         14.4          13.2         15.4
##  3 21-Feb          30129.         27.1         21.9          22.2         19.8
##  4 21-Mar          31682.         28.4         24.5          20.4         17.4
##  5 21-Apr          47777.         29.5         22.8          24.2         21.4
##  6 21-May          53922.         34.5         25.0          19.2         20.2
##  7 21-Jun          55944.         32.1         21.2          22.6         22.7
##  8 21-Jul          49107.         29.3         25.6          20.2         20.4
##  9 21-Aug          41629.         26.2         20.2          18.4         18.5
## 10 21-Sep          32859.         26.4         21.2          16.2         17.0
## 11 21-Oct          40705.         26.0         16.5          15.1         15.2
## 12 21-Nov          34998.         19.0         14.0          12.4         12.8
## # ... with 10 more variables: avg_ride_thurs <dbl>, avg_ride_fri <dbl>,
## #   avg_ride_sat <dbl>, mean_ride_length <dbl>, mode_day_of_week <dbl>,
## #   count_rides_casual <dbl>, count_rides_member <dbl>, total_rides <dbl>,
## #   avg_ride_casual <dbl>, avg_ride_member <dbl>
```

```
full_year_clean <- full_year %>%
  rename(months = ...1)
```

Saving the updated full_year as a new data frame

```
View(full_year_clean)
```

Taking a quick look at the new data frame

```
full_year_clean %>%
  select(months, mode_day_of_week, mean_ride_length, count_rides_casual, count_rides_member, total_ride
```

Looking at specific columns

```
## # A tibble: 12 x 6
##    months mode_day_of_week mean_ride_length count_rides_casual count_rides_memb~
##    <chr>             <dbl>            <dbl>              <dbl>             <dbl>
##  1 20-Dec                4             16.0              29997            101142
##  2 21-Jan                7             15.3              18117             78717
##  3 21-Feb                7             24.4              10131             39491
##  4 21-Mar                7             22.9              84033            144463
##  5 21-Apr                6             24.1             136601            200629
##  6 21-May                7             26.0             256916            274717
##  7 21-Jun                7             26.1             370681            358914
##  8 21-Jul                7             24.2             442056            380354
##  9 21-Aug                1             21.6             412671            391681
## 10 21-Sep                7             20.5             363890            392257
## 11 21-Oct                7             19.1             257242            373984
## 12 21-Nov                3             14.8             106929            253049
## # ... with 1 more variable: total_rides <dbl>
```

```
full_year_clean %>%
  select(months, mode_day_of_week, mean_ride_length, count_rides_casual, count_rides_member, total_ride
  filter(mean_ride_length > 20.00) %>%
  filter(mean_ride_length != 0.00)
```

Finding the months with mean ride lengths that are longer than 20 minutes

```
## # A tibble: 8 x 6
##   months mode_day_of_week mean_ride_length count_rides_casual count_rides_member
##   <chr>             <dbl>            <dbl>              <dbl>              <dbl>
## 1 21-Feb                7             24.4              10131              39491
## 2 21-Mar                 7             22.9              84033             144463
## 3 21-Apr                 6             24.1             136601             200629
## 4 21-May                 7             26.0             256916             274717
## 5 21-Jun                 7             26.1             370681             358914
## 6 21-Jul                 7             24.2             442056             380354
## 7 21-Aug                 1             21.6             412671             391681
## 8 21-Sep                 7             20.5             363890             392257
## # ... with 1 more variable: total_rides <dbl>
```

```
full_year_clean %>%
  select(months, mode_day_of_week, mean_ride_length, count_rides_casual, count_rides_member, total_ride
  filter(mean_ride_length <= 20.00) %>%
  filter(mean_ride_length != 0.00)
```

Finding the months with mean ride lengths that are 20 minutes or shorter

```
## # A tibble: 4 x 6
##   months mode_day_of_week mean_ride_length count_rides_casual count_rides_member
##   <chr>             <dbl>            <dbl>              <dbl>              <dbl>
## 1 20-Dec                4             16.0              29997             101142
## 2 21-Jan                7             15.3              18117              78717
## 3 21-Oct                7             19.1             257242             373984
## 4 21-Nov                3             14.8             106929             253049
## # ... with 1 more variable: total_rides <dbl>
```

```
full_year_clean %>%
  select(mode_day_of_week) %>%
  count(mode_day_of_week) %>%
  group_by(mode_day_of_week) %>%
  arrange(-n)
```

Finding the most popular day of the week for all users

```
## # A tibble: 5 x 2
## # Groups:   mode_day_of_week [5]
##   mode_day_of_week     n
##              <dbl> <int>
## 1                7     8
## 2                1     1
## 3                3     1
## 4                4     1
## 5                6     1
```

```
full_year_clean %>%
  select(months, mode_day_of_week, total_rides) %>%
  filter(mode_day_of_week != 1) %>%
  filter(mode_day_of_week != 7)
```

Determining which months had a weekday (not weekend) as the most popular day

```
## # A tibble: 3 x 3
##   months mode_day_of_week total_rides
##   <chr>             <dbl>       <dbl>
## 1 20-Dec                4      131139
## 2 21-Apr                6      337230
## 3 21-Nov                3      359978
```

```
full_year_clean %>%
  select(months, mode_day_of_week, count_rides_casual, count_rides_member, total_rides) %>%
  group_by(mode_day_of_week) %>%
  filter(mode_day_of_week == 1 || mode_day_of_week == 7) %>%
  summarize(sum_casual = sum(count_rides_casual), sum_member = sum(count_rides_member), sum_total = sum
```

Who took more rides on weekends? Casual riders or members?

```
## # A tibble: 2 x 4
##   mode_day_of_week sum_casual sum_member sum_total
##              <dbl>      <dbl>      <dbl>     <dbl>
## 1                1     412671     391681    804352
## 2                7    1803066    2042897   3845963
```

```
full_year_clean %>%
  select(months, mode_day_of_week, count_rides_casual, count_rides_member, total_rides) %>%
  group_by(mode_day_of_week) %>%
  filter(mode_day_of_week != 1 && mode_day_of_week != 7) %>%
  summarize(sum_casual = sum(count_rides_casual), sum_member = sum(count_rides_member), sum_total = sum
```

Who took more rides during the week? Casual riders or members? (Days 2-6)

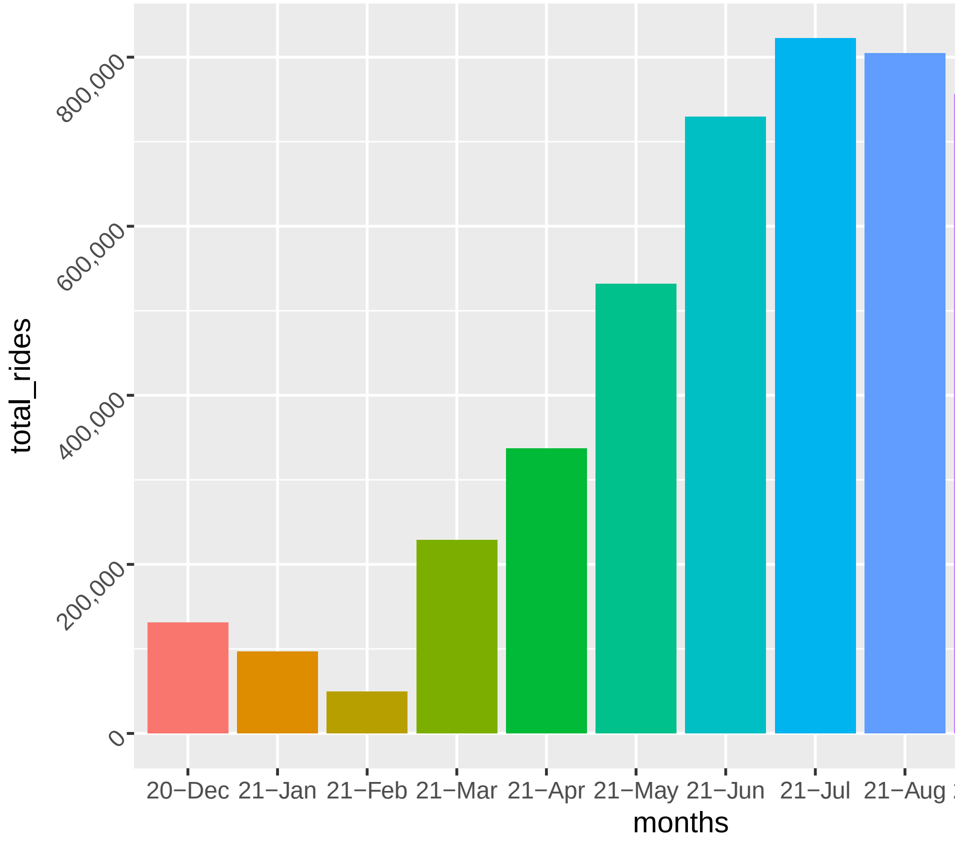
```
## # A tibble: 3 x 4
##   mode_day_of_week sum_casual sum_member sum_total
##              <dbl>      <dbl>      <dbl>     <dbl>
## 1                3     106929     253049    359978
## 2                4      29997     101142    131139
## 3                6     136601     200629    337230
```

Creating the data visualizations

```
library(scales)
```

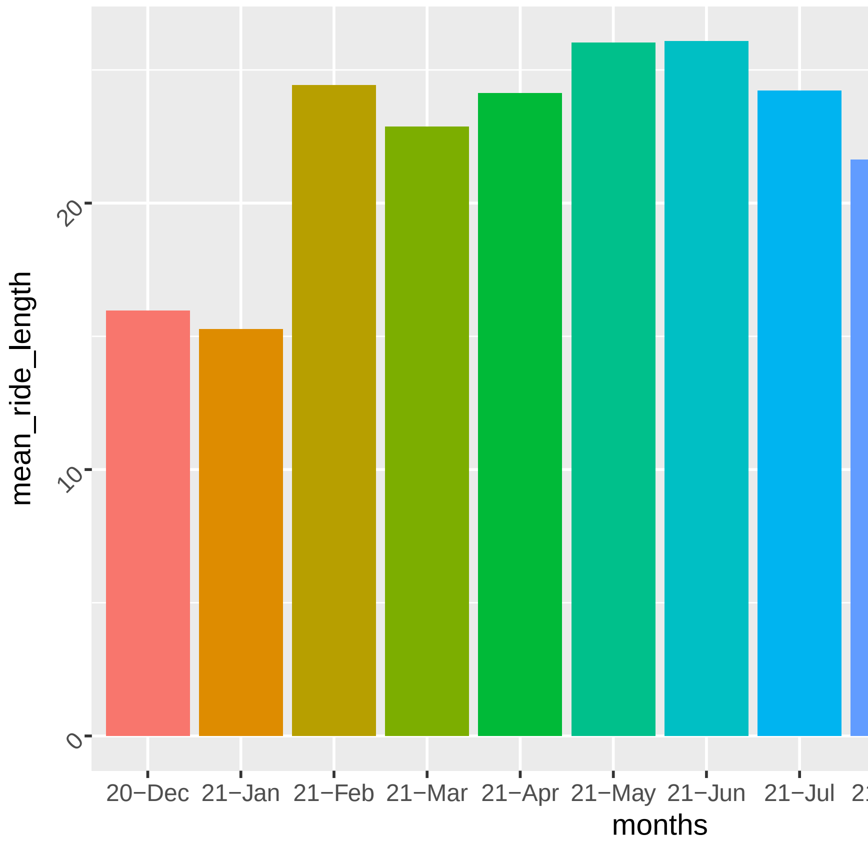
```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
full_year_clean %>%
  mutate(months = factor(months, levels = c(months))) %>%
  ggplot(aes(x = months, y = total_rides, fill = months)) +
    geom_bar(stat = "identity") +
    theme(legend.position = "none") +
    theme(axis.text.y = element_text(angle = 45)) +
    scale_y_continuous(labels = comma)
```



Bar chart for total rides per month

```r
full_year_clean %>%
  mutate(months = factor(months, levels = c(months))) %>%
  ggplot(aes(x = months, y = mean_ride_length, fill = months)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "none") +
  theme(axis.text.y = element_text(angle = 45))
```



Bar chart for mean ride length per month