Bellabeat Case Study Report

Aron Rana

2024-10-24

Contents

Introduction

Bellabeat is a high-tech company that manufactures health-focused, wearable smart technology for women. I have been tasked with analyzing smart device data from non-Bellabeat device owners to understand how they currently use their devices. The insights gained from this analysis would then be applied to Bellabeat products and used to suggest informed marketing strategies to the executive team members at Bellabeat. These informed marketing strategies, if approved, would then be implemented to promote company growth.

This report will include the steps taken to prepare datasets and the insights gained from analysis.

RStudio Desktop (Version 2024.09.0+375) for macOS Ventura 13.7 (22H123) was used for this project.

Step 1: Load the required R packages

The following R packages were installed and loaded:

1. `tidyverse`
2. `skimr`
3. `vctrs`
4. `janitor`
5. `lubridate`
6. `knitr`

Step 2: Import the required datasets

Data from FitBit fitness trackers was obtained from https://www.kaggle.com/datasets/arashnic/fitbit. These datasets were for the period 12 March - 12 May, 2016 and was downloaded as a single ZIP file named "archive.zip". Once decompressed, the ZIP file was a main folder named "archive" that contained two sub-folders: one for datasets for the period 12 March - 11 April, 2016 and the second for datasets for the period 12 April - 12 May, 2016. The folder for the 12 March - 11 April, 2016 dataset contained 11 CSV files with various categories of health data. The folder for the 12 April - 12 May, 2016 dataset contained 18 CSV files also with various categories of health data. The two folders both contained files with identical naming conventions. In addition to this, both folders had CSV files for identical health data categories with identical column names.

The downloaded CSV files were stored in a sub-folder on my password protected personal computer. This sub-folder, for CSV files only, was contained in a main folder that was used for exclusively for this project. Backups of the CSV and project files were stored on an external flash drive in addition to my personal Google Drive account. Other than myself, no one had access to the datasets or their backups.

Out of the possible 29 CSV files downloaded, I gauged 15 of them to be necessary for my project and imported them into the global environment on RStudio Desktop using the code below:

```r
ma_dailyActivity_merged <- read.csv("dailyActivity_merged.csv")
ma_heartrate_seconds_merged <- read.csv("heartrate_seconds_merged.csv")
ma_hourlyCalories_merged <- read.csv("hourlyCalories_merged.csv")
ma_hourlyIntensities_merged <- read.csv("hourlyIntensities_merged.csv")
ma_hourlySteps_merged <- read.csv("hourlySteps_merged.csv")
ma_minuteMETsNarrow_merged <- read.csv("minuteMETsNarrow_merged.csv")
ma_minuteSleep_merged <- read.csv("minuteSleep_merged.csv")
am_dailyActivity_merged <- read.csv("am_dailyActivity_merged.csv")
am_heartrate_seconds_merged <- read.csv("am_heartrate_seconds_merged.csv")
am_hourlyCalories_merged <- read.csv("am_hourlyCalories_merged.csv")
am_hourlyIntensities_merged <- read.csv("am_hourlyIntensities_merged.csv")
am_hourlySteps_merged <- read.csv("am_hourlySteps_merged.csv")
am_minuteMETsNarrow_merged <- read.csv("am_minuteMETsNarrow_merged.csv")
am_minuteSleep_merged <- read.csv("am_minuteSleep_merged.csv")
am_sleepDay_merged <- read.csv("am_sleepDay_merged.csv")
```

NOTE:

1. Dataframes with the prefix "ma_" are for the period 12 March - 11 April, 2016.
2. Dataframes with the prefix "am_" are for the period 12 April - 12 May, 2016.

Step 3: Clean the imported datasets

All imported datasets were previewed using the `skim_without_charts()` and `View()` functions. The `n_unique()` function was used to determine the number of unique id's in each dataframe and thus gave an idea of how many users shared their health data for that particular category of data. I also made use of the `vec_duplicate_any()` and `duplicated()` functions to determine if any of the dataframes had duplicate records. I analyzed the dataframes that had duplicate records and found that these records were acceptable due to the nature of the data being collected, for example the per second heart rate reading of each individual for each day of the month. Each user would have multiple rows for the same ID and date, but different entries for time and heart rate. Records that were complete duplicates were removed using the `distinct()` function. I also made use of the sort and filter features of R Studio Desktop to look for irregularities in each of the imported datasets.

I made the following observations after my initial preview of the imported datasets:

1. The data was outdated.
2. The data was limited to only 3 months of 2016.
3. Certain dataframes had column names that were vague and I could not make use of much of the data from that dataframe.

4. Certain records from a few dataframes contained data that did not make sense. For example, records on intensity and MET's contained values that seemed practically impossible.
5. The `skim_without_charts()` function was used to generate comprehensive summaries of all imported datasets. No dataframe had missing or N/A values and no cell values had white spaces.
6. The naming conventions for columns were not entirely consistent.
7. The datasets had personally identifiable information already removed.
8. The datasets would be more insightful if they contained the age of users in addition to data about when the users wore and removed their wearable health devices. Units of measurement for relevant columns would also be helpful.

The imported datasets were cleaned in the following and respective manner:

```
# dailyActivity 12 March - 11 April
ma_dailyActivity_merged$ActivityDate <- mdy(ma_dailyActivity_merged$ActivityDate)
ma_dailyActivity_merged$LoggedActivitiesDistance <- round(ma_dailyActivity_merged$LoggedActivitiesDistance, digits = 2)
ma_dailyActivity_merged <- clean_names(ma_dailyActivity_merged)

ma_dailyActivity_merged <- ma_dailyActivity_merged %>%
  rename(lightly_active_distance = light_active_distance) %>%
  rename(sedentary_distance = sedentary_active_distance)

# heartrate_seconds 12 March - 11 April
ma_heartrate_seconds_merged$Time <- mdy_hms(ma_heartrate_seconds_merged$Time)
ma_heartrate_seconds_merged <- clean_names(ma_heartrate_seconds_merged)

ma_heartrate_seconds_merged <- ma_heartrate_seconds_merged %>%
  rename(activity_date_time = time) %>%
  rename(heart_rate = value)

# hourlyCalories 12 March - 11 April
ma_hourlyCalories_merged$ActivityHour <- mdy_hms(ma_hourlyCalories_merged$ActivityHour)
ma_hourlyCalories_merged <- clean_names(ma_hourlyCalories_merged)

ma_hourlyCalories_merged <- ma_hourlyCalories_merged %>%
  rename(activity_date_time = activity_hour)

# hourlyIntensities 12 March - 11 April
ma_hourlyIntensities_merged$ActivityHour <- mdy_hms(ma_hourlyIntensities_merged$ActivityHour)
ma_hourlyIntensities_merged$AverageIntensity <- round(ma_hourlyIntensities_merged$AverageIntensity, digits = 2)
ma_hourlyIntensities_merged <- clean_names(ma_hourlyIntensities_merged)
```

```r
ma_hourlyIntensities_merged <- ma_hourlyIntensities_merged %>%
  rename(activity_date_time = activity_hour)

# hourlySteps 12 March - 11 April
ma_hourlySteps_merged$ActivityHour <- mdy_hms(ma_hourlySteps_merged$ActivityHour)
ma_hourlySteps_merged <- clean_names(ma_hourlySteps_merged)

ma_hourlySteps_merged <- ma_hourlySteps_merged %>%
  rename(activity_date_time = activity_hour) %>%
  rename(total_steps = step_total)

# minuteMETsNarrow 12 March - 11 April
ma_minuteMETsNarrow_merged$ActivityMinute <- mdy_hms(ma_minuteMETsNarrow_merged$ActivityMinute)
ma_minuteMETsNarrow_merged <- clean_names(ma_minuteMETsNarrow_merged)

ma_minuteMETsNarrow_merged <- ma_minuteMETsNarrow_merged %>%
  rename(activity_date_time = activity_minute) %>%
  rename(METs = me_ts)

# minuteSleep 12 March - 11 April
ma_minuteSleep_merged$date <- mdy_hms(ma_minuteSleep_merged$date)
ma_minuteSleep_merged <- clean_names(ma_minuteSleep_merged)

ma_minuteSleep_merged <- ma_minuteSleep_merged %>%
  rename(activity_date_time = date)

# dailyActivity 12 April - 12 May
am_dailyActivity_merged$ActivityDate <- mdy(am_dailyActivity_merged$ActivityDate)
am_dailyActivity_merged$LoggedActivitiesDistance <- round(am_dailyActivity_merged$LoggedActivitiesDistance, digits = 2)
am_dailyActivity_merged <- clean_names(am_dailyActivity_merged)

am_dailyActivity_merged <- am_dailyActivity_merged %>%
  rename(lightly_active_distance = light_active_distance) %>%
  rename(sedentary_distance = sedentary_active_distance)

# heartrate_seconds 12 April - 12 May
am_heartrate_seconds_merged$Time <- mdy_hms(am_heartrate_seconds_merged$Time)
am_heartrate_seconds_merged <- clean_names(am_heartrate_seconds_merged)
```

```r
am_heartrate_seconds_merged <- am_heartrate_seconds_merged %>%
  rename(activity_date_time = time) %>%
  rename(heart_rate = value)

# hourlyCalories 12 April - 12 May
am_hourlyCalories_merged$ActivityHour <- mdy_hms(am_hourlyCalories_merged$ActivityHour)
am_hourlyCalories_merged <- clean_names(am_hourlyCalories_merged)

am_hourlyCalories_merged <- am_hourlyCalories_merged %>%
  rename(activity_date_time = activity_hour)

# hourlyIntensities 12 April - 12 May
am_hourlyIntensities_merged$ActivityHour <- mdy_hms(am_hourlyIntensities_merged$ActivityHour)
am_hourlyIntensities_merged$AverageIntensity <- round(am_hourlyIntensities_merged$AverageIntensity, digits = 2)
am_hourlyIntensities_merged <- clean_names(am_hourlyIntensities_merged)

am_hourlyIntensities_merged <- am_hourlyIntensities_merged %>%
  rename(activity_date_time = activity_hour)

# hourlySteps 12 April - 12 May
am_hourlySteps_merged$ActivityHour <- mdy_hms(am_hourlySteps_merged$ActivityHour)
am_hourlySteps_merged <- clean_names(am_hourlySteps_merged)

am_hourlySteps_merged <- am_hourlySteps_merged %>%
  rename(activity_date_time = activity_hour) %>%
  rename(total_steps = step_total)

# minuteMETsNarrow 12 April - 12 May
am_minuteMETsNarrow_merged$ActivityMinute <- mdy_hms(am_minuteMETsNarrow_merged$ActivityMinute)
am_minuteMETsNarrow_merged <- clean_names(am_minuteMETsNarrow_merged)

am_minuteMETsNarrow_merged <- am_minuteMETsNarrow_merged %>%
  rename(activity_date_time = activity_minute) %>%
  rename(METs = me_ts)

# minuteSleep 12 April - 12 May
am_minuteSleep_merged$date <- mdy_hms(am_minuteSleep_merged$date)
am_minuteSleep_merged <- clean_names(am_minuteSleep_merged)
```

```r
am_minuteSleep_merged <- am_minuteSleep_merged %>%
  rename(activity_date_time = date)

# SleepDay 12 April - 12 May
am_sleepDay_merged$SleepDay <- mdy_hms(am_sleepDay_merged$SleepDay)
am_sleepDay_merged <- clean_names(am_sleepDay_merged)

am_sleepDay_merged <- am_sleepDay_merged %>%
  rename(activity_date_time = sleep_day) %>%
  rename(total_minutes_in_bed = total_time_in_bed) %>%
  distinct()

am_sleepDay_merged$time_awake_in_bed <- am_sleepDay_merged$total_minutes_in_bed - am_sleepDay_merged$total_minutes_asleep
am_sleepDay_merged$weekday <- wday(am_sleepDay_merged$activity_date_time, label=TRUE)
am_sleepDay_merged$month <- month(am_sleepDay_merged$activity_date_time, label=TRUE)
```

Step 4: Merge relevant dataframes

The cleaned datasets were merged accordingly to form one collective dataframe for each health data category with all the required data in one place. Dataframes were merged accordingly using the respective code below:

```r
dailyActivity_merged <- rbind(ma_dailyActivity_merged, am_dailyActivity_merged)
heartrate_seconds_merged <- rbind(ma_heartrate_seconds_merged, am_heartrate_seconds_merged)
hourlyCalories_merged <- rbind(ma_hourlyCalories_merged, am_hourlyCalories_merged)
hourlyIntensities_merged <- rbind(ma_hourlyIntensities_merged, am_hourlyIntensities_merged)
hourlySteps_merged <- rbind(ma_hourlySteps_merged, am_hourlySteps_merged)
minuteMETsNarrow_merged <- rbind(ma_minuteMETsNarrow_merged, am_minuteMETsNarrow_merged)
minuteSleep_merged <- rbind(ma_minuteSleep_merged, am_minuteSleep_merged)
```

Step 5: Verify data cleaning procedures

I verified my data cleaning process in a number of ways:

1. I inspected each dataframe after running code chunks to ensure that the desired effect had taken place.

2. Each individual dataframe was cleaned according to its specific needs; the dataframes were not cleaned with a common set of code.
3. I used the sort and filter options to look for inconsistencies in the data.
4. I double checked for duplicate records and made sure that the number of rows in each dataframe did not change unexpectedly.
5. I monitored the environment pane as I made changes to dataframes; the environment pane displays the number of rows and number of variables for each dataframe and can be useful in tracking the data cleaning process.
6. I double checked every line of code to ensure accuracy.

Step 6: Create summary statistics

The following summary statistics were made based on the imported and cleaned datasets:

1. Heart rate tracked per second (every 5-10 seconds)

1(a). Statistical summary of heart rate values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.0    64.0    74.0    78.1    88.0   203.0
```

1(b). Data on records with the lowest heart rate

```
##            id  activity_date_time heart_rate
## 1  5577150313 2016-04-01 03:58:20         39
## 2  5577150313 2016-04-01 03:58:30         38
## 3  5577150313 2016-04-01 03:58:45         38
## 4  5577150313 2016-04-01 03:58:55         37
## 5  5577150313 2016-04-01 03:59:00         37
## 6  5577150313 2016-04-01 03:59:15         37
## 7  5577150313 2016-04-10 06:23:00         39
## 8  5577150313 2016-04-10 06:23:10         38
## 9  5577150313 2016-04-10 06:23:20         37
## 10 5577150313 2016-04-10 06:23:30         36
## 11 5577150313 2016-04-10 06:23:45         36
## 12 5577150313 2016-04-10 06:23:55         37
## 13 5577150313 2016-04-10 06:24:00         38
## 14 5577150313 2016-04-10 06:24:05         39
```

```
## 15 2022484408 2016-04-27 13:47:00         38
## 16 2022484408 2016-04-27 13:47:15         38
## 17 2022484408 2016-04-27 13:47:20         39
## 18 2022484408 2016-04-27 13:47:35         39
## 19 2022484408 2016-04-27 13:47:50         39
## 20 2022484408 2016-04-27 13:48:35         39
## 21 4388161847 2016-05-01 04:09:30         39
## 22 5577150313 2016-04-15 03:57:40         39
## 23 5577150313 2016-04-15 03:57:50         39
## 24 5577150313 2016-04-15 03:58:05         39
## 25 5577150313 2016-04-15 03:58:20         39
## 26 5577150313 2016-04-15 03:58:35         39
## 27 5577150313 2016-04-15 03:59:10         39
## 28 5577150313 2016-05-04 02:00:00         38
## 29 5577150313 2016-05-04 02:00:10         36
## 30 5577150313 2016-05-04 02:00:20         36
## 31 5577150313 2016-05-04 02:00:35         37
## 32 5577150313 2016-05-04 02:00:50         37
## 33 5577150313 2016-05-04 07:41:10         39
## 34 5577150313 2016-05-04 07:41:20         38
## 35 5577150313 2016-05-04 07:41:30         39
## 36 5577150313 2016-05-04 07:41:45         39
## 37 5577150313 2016-05-04 18:48:20         39
```

1(c) Data on records with the highest heart rate

```
##            id  activity_date_time heart_rate
## 1  2022484408 2016-04-21 16:31:20        200
## 2  2022484408 2016-04-21 16:31:30        202
## 3  2022484408 2016-04-21 16:31:40        203
## 4  2022484408 2016-04-21 16:31:50        202
## 5  2022484408 2016-04-21 16:32:00        203
## 6  2022484408 2016-04-21 16:32:10        203
## 7  2022484408 2016-04-21 16:32:20        203
## 8  2022484408 2016-04-21 16:32:35        203
## 9  2022484408 2016-04-21 16:32:40        201
## 10 2022484408 2016-04-21 16:32:50        200
## 11 2022484408 2016-04-21 16:33:05        200
```

```
## 12 2022484408 2016-04-21 16:33:10        199
## 13 2022484408 2016-04-21 16:33:25        199
## 14 2022484408 2016-04-21 16:33:40        199
## 15 2022484408 2016-04-21 16:33:50        198
## 16 2022484408 2016-04-21 17:05:40        202
## 17 2022484408 2016-04-21 17:05:50        203
## 18 2022484408 2016-04-21 17:06:05        203
## 19 2022484408 2016-04-21 17:06:20        203
## 20 2022484408 2016-04-21 17:06:30        202
## 21 2022484408 2016-04-21 17:06:40        200
## 22 2022484408 2016-04-21 17:06:50        199
## 23 4558609924 2016-05-08 13:34:35        198
## 24 4558609924 2016-05-08 13:34:50        198
## 25 4558609924 2016-05-08 13:34:55        199
## 26 4558609924 2016-05-08 13:35:00        198
## 27 4558609924 2016-05-08 13:35:10        199
## 28 4558609924 2016-05-08 13:35:20        199
## 29 4558609924 2016-05-08 13:35:30        198
## 30 4558609924 2016-05-08 13:35:35        199
## 31 4558609924 2016-05-08 13:36:45        198
## 32 4558609924 2016-05-08 13:36:50        198
```

1(d)(i). Summary of number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month       n
##   <ord>   <int>
## 1 Mar     45531
## 2 Apr   2699396
## 3 May    893412
```

1(d)(ii). Summary of number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday      n
##   <ord>    <int>
```

```
## 1 Sun     464795
## 2 Mon     474370
## 3 Tue     584000
## 4 Wed     540515
## 5 Thu     475835
## 6 Fri     565747
## 7 Sat     533077
```

2. MET (metabolic equivalent of task) tracked per the minute

2(a). Statistical summary of MET values

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   10.00   14.45   11.00  189.00
```

2(b). Data on records with the lowest MET

2(c). Data on records with the highest MET

```
##            id  activity_date_time METs
## 1  2022484408 2016-03-16 16:33:00  148
## 2  2873212765 2016-03-12 07:49:00  143
## 3  3372868164 2016-03-21 11:22:00  159
## 4  3372868164 2016-03-29 10:37:00  145
## 5  5577150313 2016-03-19 06:57:00  144
## 6  5577150313 2016-03-19 07:01:00  144
## 7  5577150313 2016-03-23 16:33:00  144
## 8  5577150313 2016-03-23 16:34:00  146
## 9  8053475328 2016-03-19 14:03:00  189
## 10 2022484408 2016-04-21 16:33:00  146
## 11 2022484408 2016-04-21 16:34:00  144
## 12 2022484408 2016-04-21 17:07:00  144
## 13 2873212765 2016-04-23 08:07:00  149
## 14 2873212765 2016-05-07 07:43:00  157
## 15 2873212765 2016-05-07 07:44:00  153
## 16 4558609924 2016-05-08 13:36:00  144
```

2(d)(i). Summary of number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month       n
##   <ord>   <int>
## 1 Mar    950400
## 2 Apr   1371420
## 3 May    448800
```

2(d)(ii). Summary of number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday      n
##   <ord>    <int>
## 1 Sun     399540
## 2 Mon     394860
## 3 Tue     415860
## 4 Wed     401460
## 5 Thu     385260
## 6 Fri     368040
## 7 Sat     405600
```

3. Daily activities

3(a)(i). Summary of number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month     n
##   <ord> <int>
## 1 Mar      74
## 2 Apr     994
## 3 May     329
```

3(a)(ii). Summary of number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday     n
##   <ord>   <int>
## 1 Sun       193
## 2 Mon       188
## 3 Tue       225
## 4 Wed       198
## 5 Thu       195
## 6 Fri       199
## 7 Sat       199
```

3(b). Statistical summary of number of daily total steps

1. Minimum
2. Maximum
3. Average
4. Sum

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

```
## [1] 36019
```

```
## [1] 7280.898
```

```
## [1] 10171415
```

3(c). Statistical summary of daily total distance

1. Minimum
2. Maximum
3. Average
4. Sum

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

```
## [1] 28.03
```

```
## [1] 5.219434
```

```
## [1] 7291.55
```

3(d). Statistical summary of logged activity distance

1. Minimum
2. Maximum
3. Average
4. Sum

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

```
## [1] 6.73
```

```
## [1] 0.1314603
```

```
## [1] 183.65
```

3(e). Statistical summary of daily calorie burn

1. Minimum
2. Maximum
3. Average
4. Sum

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

```
## [1] 4900
```

```
## [1] 2266.266
```

```
## [1] 3165973
```

3(f). Data for records with highest number of daily total steps

1. Average very active minutes
2. Average fairly active minutes
3. Average lightly active minutes
4. Average sedentary minutes

```
## [1] 114.8
```

```
## [1] 26.2
```

```
## [1] 261.5
```

```
## [1] 1037.5
```

3(g). Data for records with lowest number of daily total steps

1. Average lightly active minutes
2. Average sedentary minutes

```
## [1] 10.27273
```

```
## [1] 1197.636
```

3(h). Data for records with highest logged activity distance

1. Minimum total steps
2. Maximum total steps
3. Average total steps
4. Average very active minutes
5. Average fairly active minutes
6. Average lightly active minutes
7. Average sedentary minutes

```
## [1] 9766
```

```
## [1] 20067
```

```
## [1] 14409.67
```

```
## [1] 48.75
```

```
## [1] 24.41667
```

```
## [1] 306.5833
```

```
## [1] 1021.167
```

3(i). Data for records with highest daily calorie burn

1. Average very active minutes
2. Average fairly active minutes
3. Average lightly active minutes
4. Average sedentary minutes
5. Minimum total steps
6. Maximum total steps
7. Average total steps
8. Sum total steps

```
## [1] 129.6364
```

```
## [1] 104.1818
```

```
## [1] 199.1818
```

```
## [1] 739.5455
```

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

```
## [1] 29326
```

```
## [1] 16641.55
```

```
## [1] 183057
```

4. Hourly calorie burn

4(a)(i). Summary of number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month     n
##   <ord> <int>
## 1 Mar   15840
## 2 Apr   22857
## 3 May    7486
```

4(a)(ii). Summary of number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday     n
##   <ord>   <int>
## 1 Sun      6659
## 2 Mon      6581
## 3 Tue      6931
## 4 Wed      6691
## 5 Thu      6427
## 6 Fri      6134
## 7 Sat      6760
```

4(b). Statistical summary of hourly calorie burn

1. Minimum hourly calorie burn
2. Maximum hourly calorie burn
3. Average hourly calorie burn

```
## [1] 42
```

```
## [1] 948
```

```
## [1] 95.75967
```

5. Hourly intensity

5(a)(i). Summary of number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month     n
##   <ord> <int>
## 1 Mar   15840
## 2 Apr   22857
## 3 May    7486
```

5(a)(ii). Summary of number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday     n
##   <ord>   <int>
## 1 Sun      6659
## 2 Mon      6581
## 3 Tue      6931
## 4 Wed      6691
## 5 Thu      6427
## 6 Fri      6134
## 7 Sat      6760
```

5(b). Statistical summary for total intensity values

1. Minimum total intensity
2. Maximum total intensity
3. Average total intensity

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

```
## [1] 180
```

```
## [1] 11.40485
```

6. Hourly steps

6(a)(i). Summary of number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month     n
##   <ord> <int>
## 1 Mar   15840
## 2 Apr   22857
## 3 May    7486
```

6(a)(ii). Summary of number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday     n
##   <ord>   <int>
## 1 Sun      6659
## 2 Mon      6581
## 3 Tue      6931
## 4 Wed      6691
## 5 Thu      6427
## 6 Fri      6134
## 7 Sat      6760
```

6(b) Statistical summary for total hourly steps

1. Minimum total hourly steps
2. Maximum total hourly steps
3. Average total hourly steps

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

```
## [1] 10565
```

```
## [1] 302.463
```

7. Sleep tracking by minute

7(a)(i). Summary on number of records by month

```
## # A tibble: 3 x 2
## # Groups:   month [3]
##   month      n
##   <ord>  <int>
## 1 Mar   124075
## 2 Apr   196327
## 3 May    66678
```

7(a)(ii). Summary on number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday     n
##   <ord>   <int>
## 1 Sun     61910
## 2 Mon     52425
## 3 Tue     55049
## 4 Wed     57698
## 5 Thu     51688
## 6 Fri     47808
## 7 Sat     60502
```

8. Sleep habits by day (12 April - 12 May, 2016)

8(a)(i). Summary on number of records by month

```
## # A tibble: 2 x 2
## # Groups:   month [2]
##   month     n
##   <ord> <int>
## 1 Apr     264
## 2 May     146
```

8(a)(ii). Summary on number of records by weekday

```
## # A tibble: 7 x 2
## # Groups:   weekday [7]
##   weekday     n
##   <ord>   <int>
## 1 Sun        55
## 2 Mon        46
## 3 Tue        65
## 4 Wed        66
## 5 Thu        64
## 6 Fri        57
## 7 Sat        57
```

8(b). Number of entries with 1 sleep record

```
##     n
## 1 364
```

8(c). Number of entries with 2 sleep record

```
##    n
## 1 43
```

8(d). Number of entries with 3 sleep record

```
##   n
## 1 3
```

8(e). Statistical summary for total minutes in bed

1. Minimum total minutes in bed
2. Maximum total minutes in bed
3. Average total minutes in bed

```
## [1] 61
```

```
## [1] 961
```

```
## [1] 458.4829
```

8(f). Data for records with highest total minutes in bed

1. Average total minutes in bed

```
## [1] 620.2258
```

8(g). Data for records with lowest total minutes in bed

1. Average total minutes in bed

```
## [1] 100.619
```

8(h). Time in bed sleeping as a percent of time in bed

1. Minimum percent_time_sleeping
2. Maximum percent_time_sleeping
3. Average percent_time_sleeping

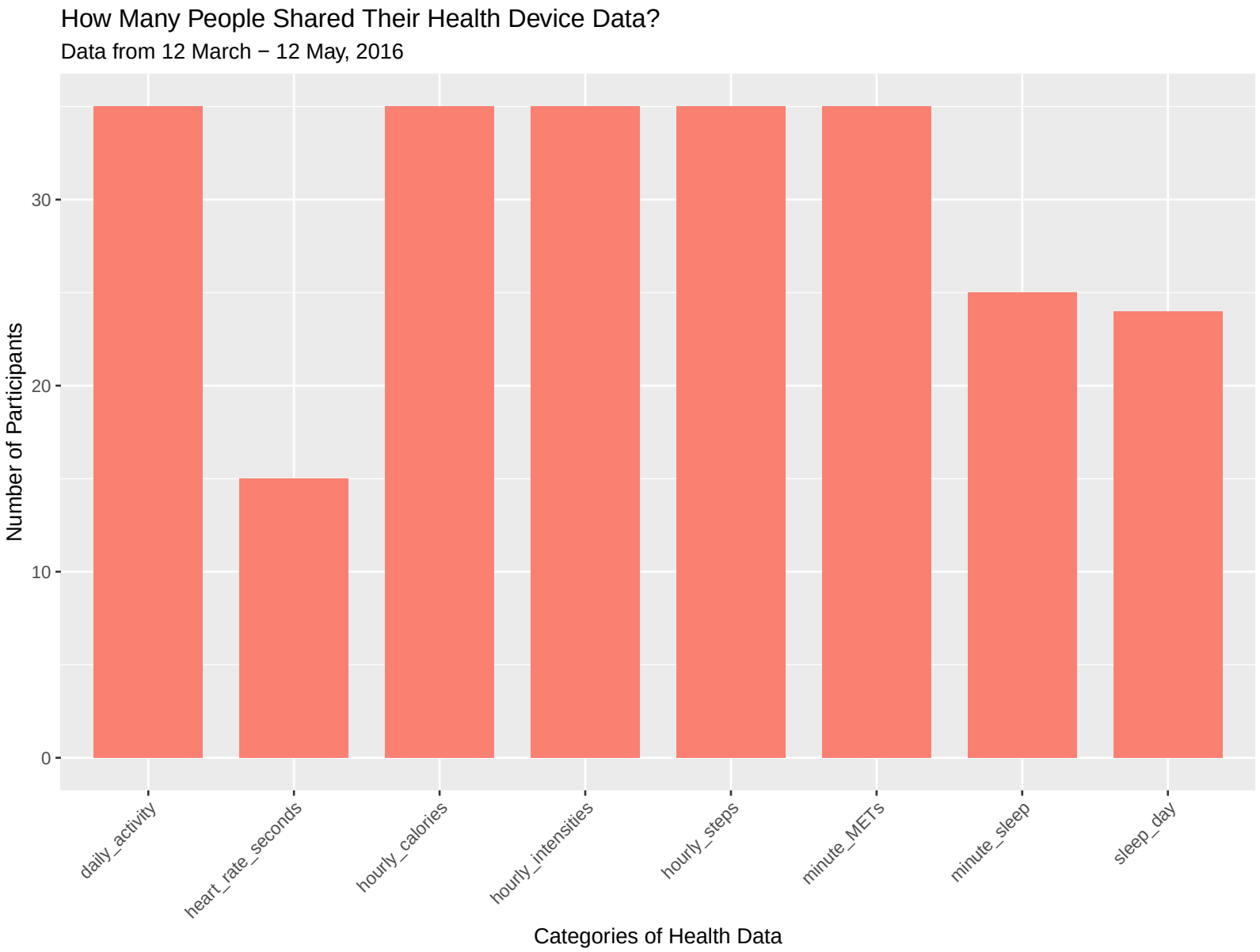
```
## [1] 49.84
```

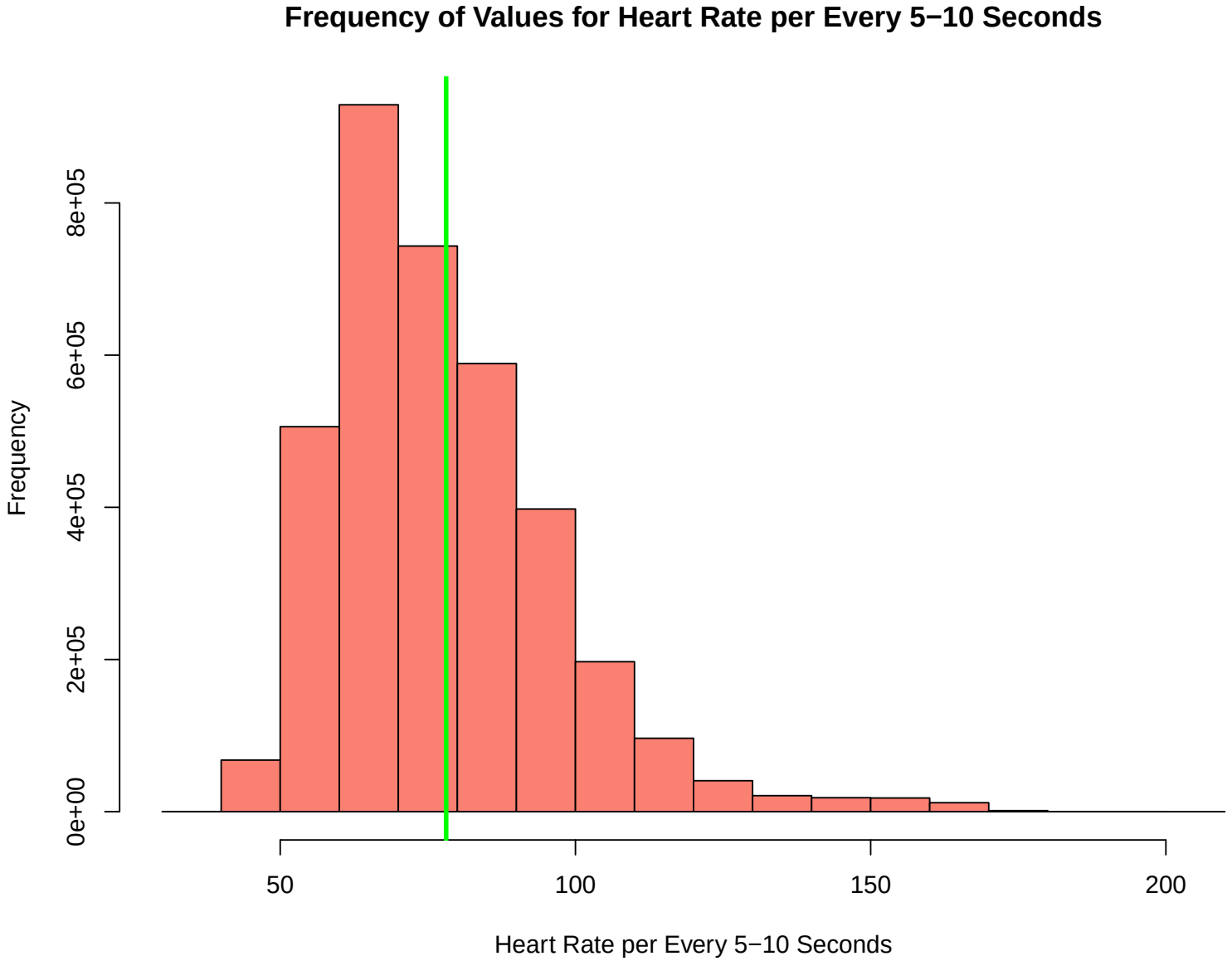
```
## [1] 100
```

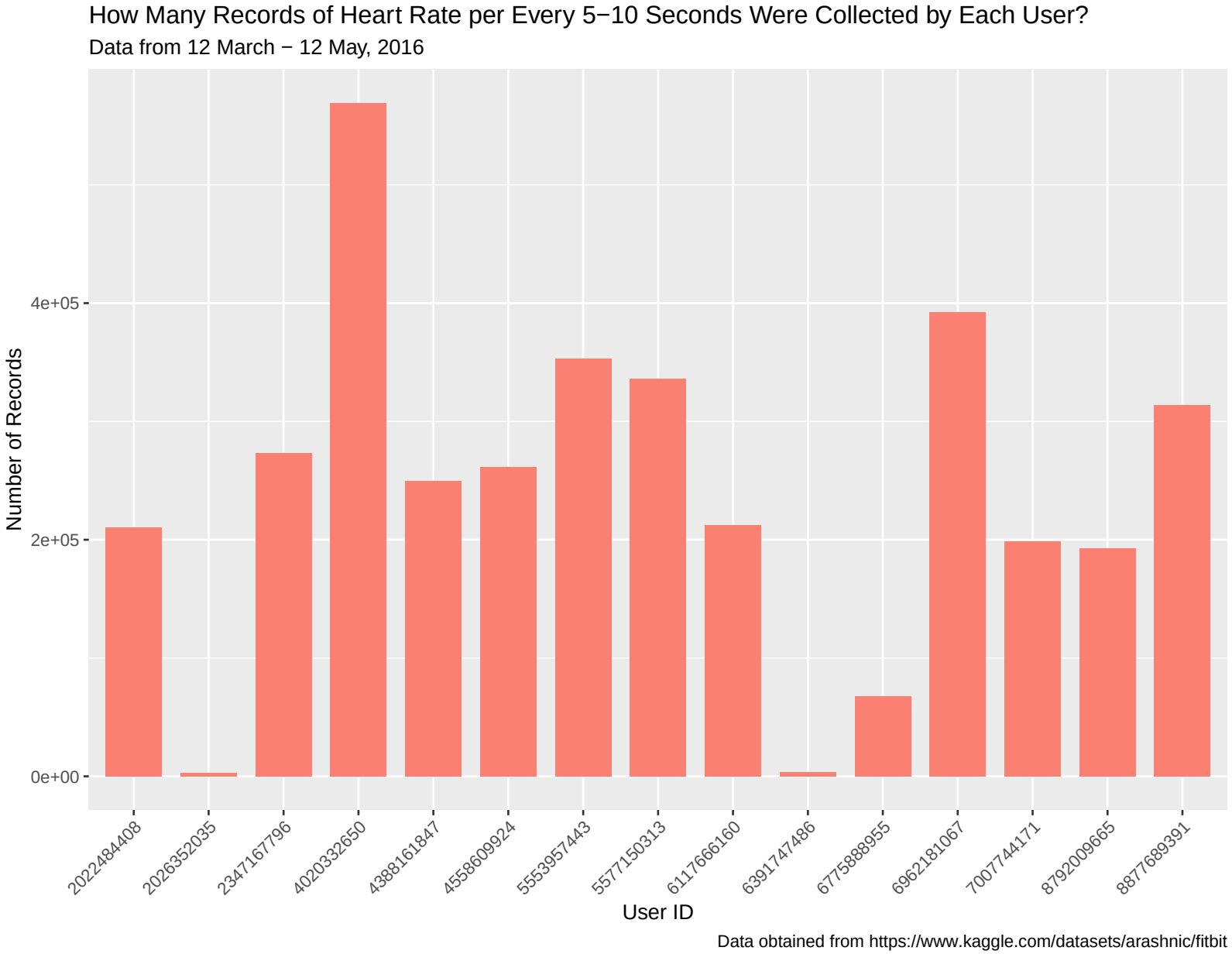
```
## [1] 91.64676
```

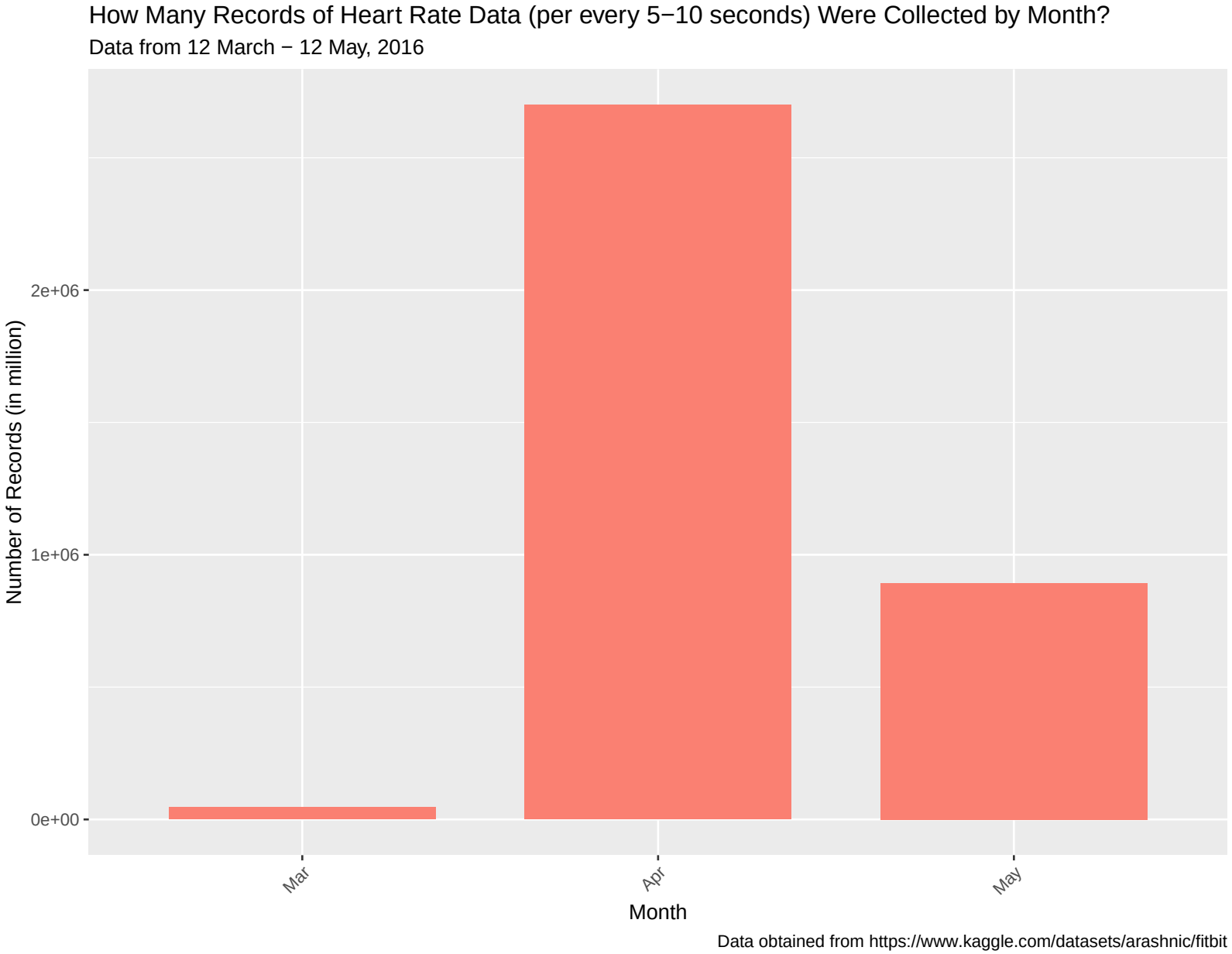
Step 7: Create data visualizations

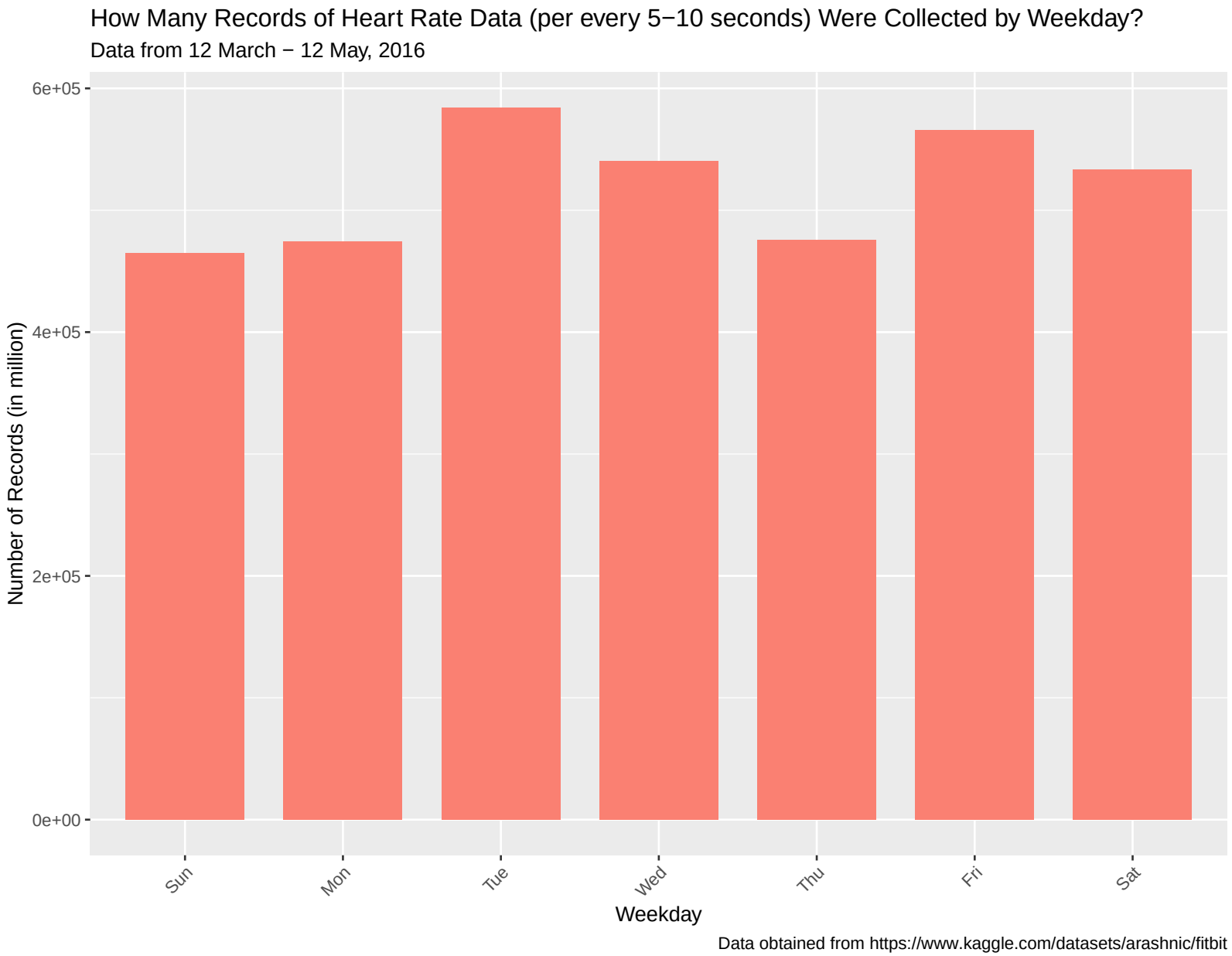
The following data visualizations were created based on the imported and cleaned datasets:

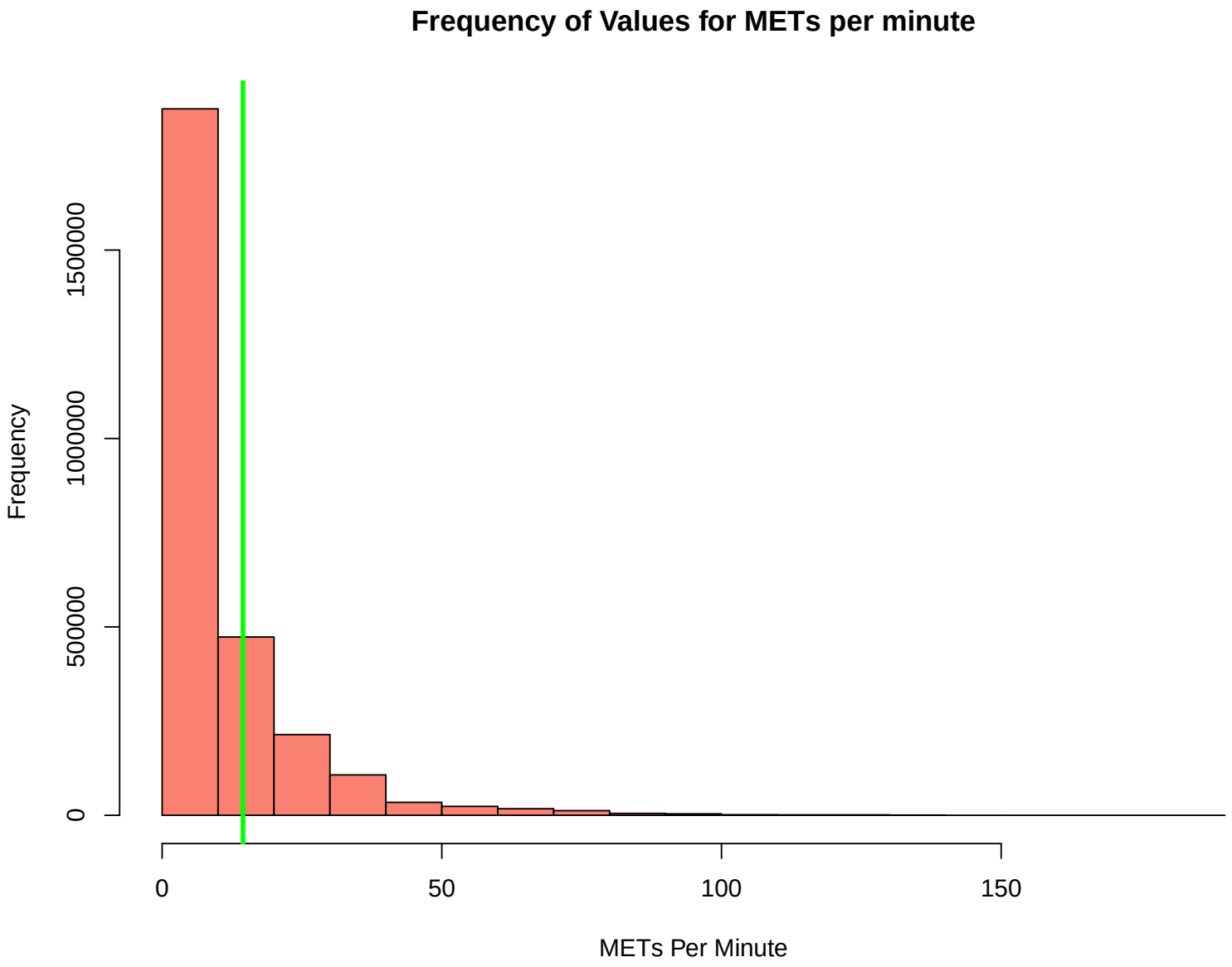

How Many People Shared Their Health Device Data?
Data from 12 March − 12 May, 2016
Number of Participants
0
10
20
30
daily_activity
heart_rate_seconds
hourly_calories
hourly_intensities
hourly_steps
minute_METs
minute_sleep
sleep_day
Categories of Health Data
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

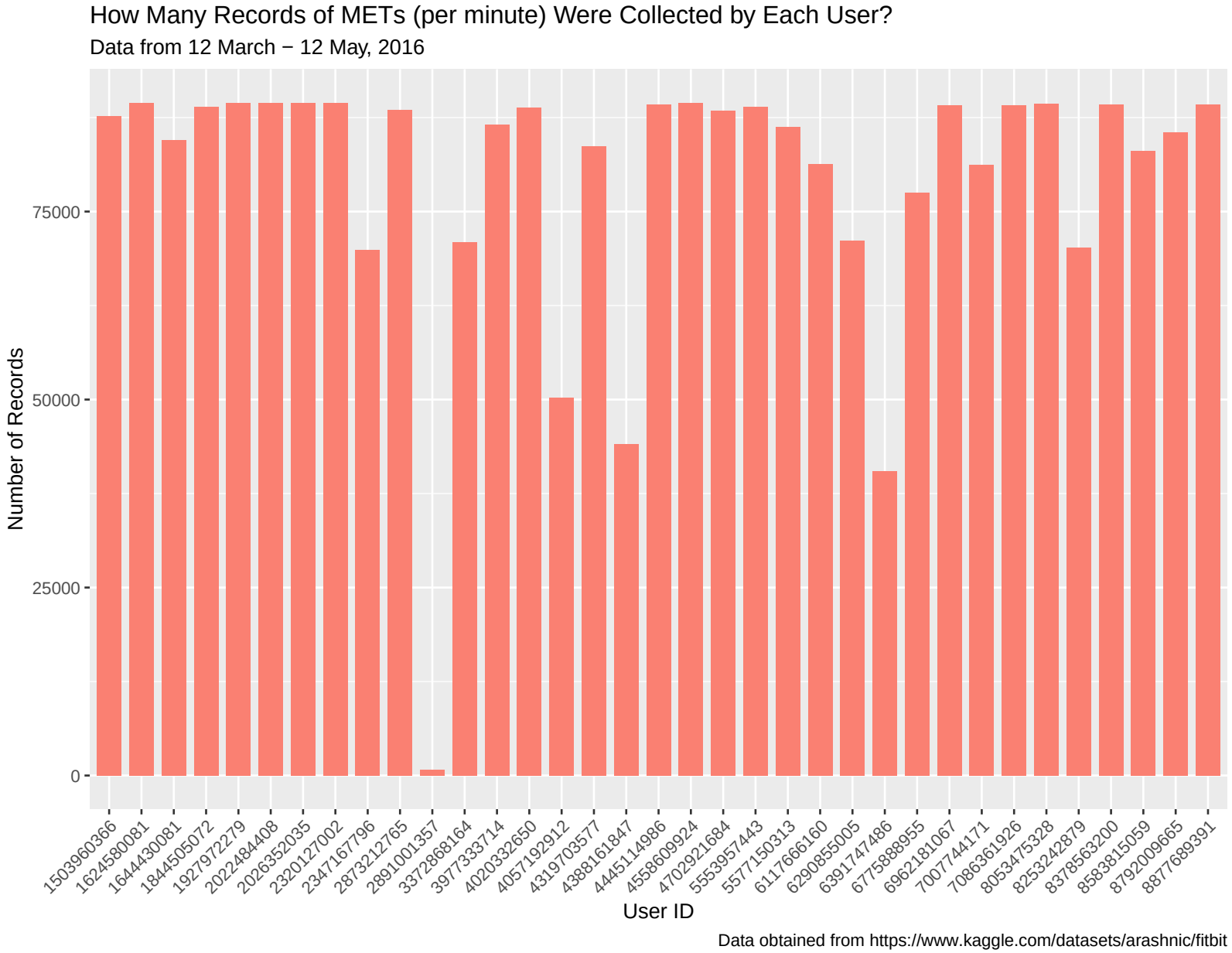

Frequency of Values for Heart Rate per Every 5−10 Seconds
Frequency
0e+00
2e+05
4e+05
6e+05
8e+05
50
100
150
200
Heart Rate per Every 5−10 Seconds

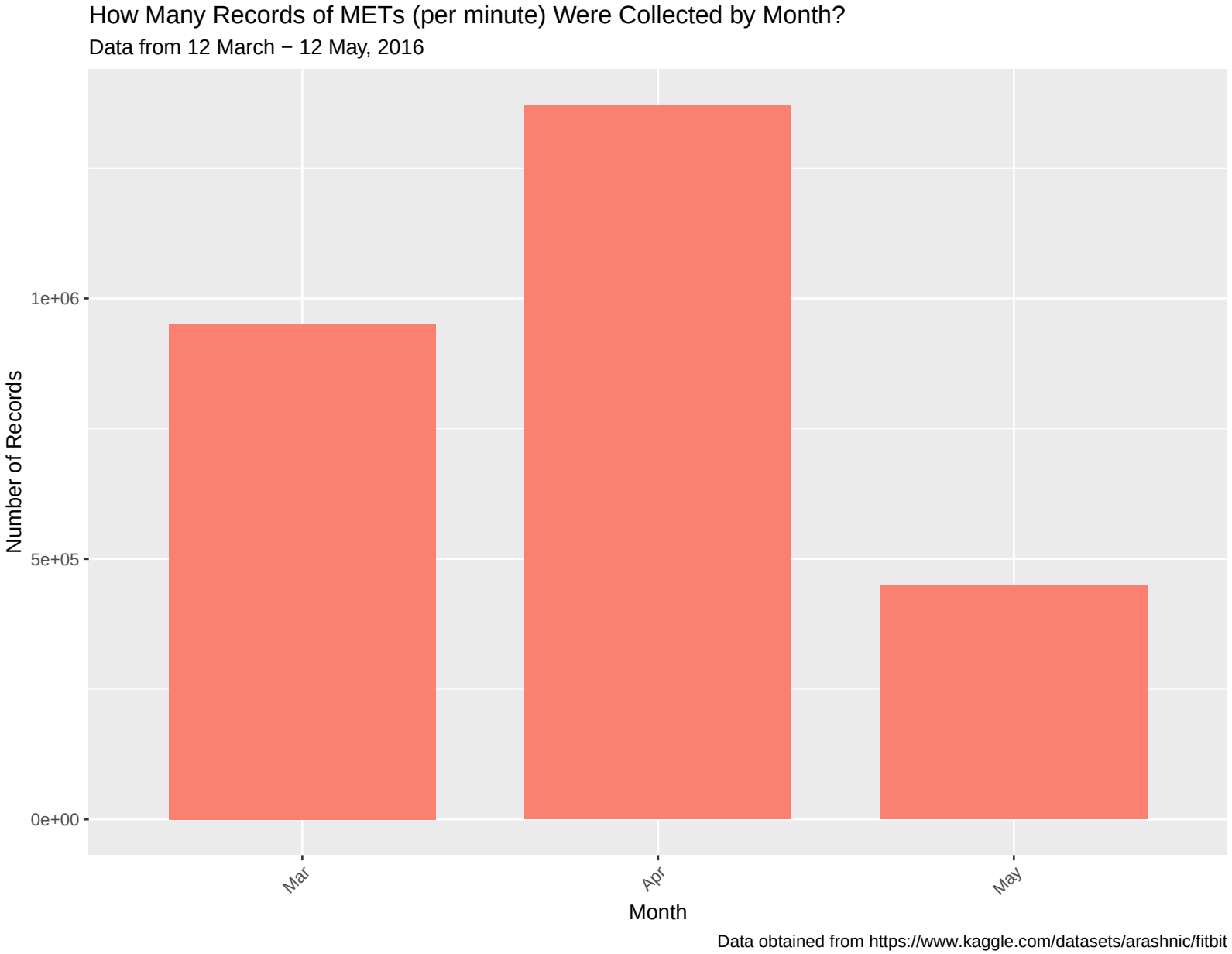

How Many Records of Heart Rate per Every 5−10 Seconds Were Collected by Each User?
Data from 12 March − 12 May, 2016
Number of Records
4e+05
2e+05
0e+00
2022484408
2026352035
2347167796
4020332650
4388161847
4558609924
5553957443
5577150313
6117666160
6391747486
6775888955
6962181067
7007744171
8792009665
8877689391
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

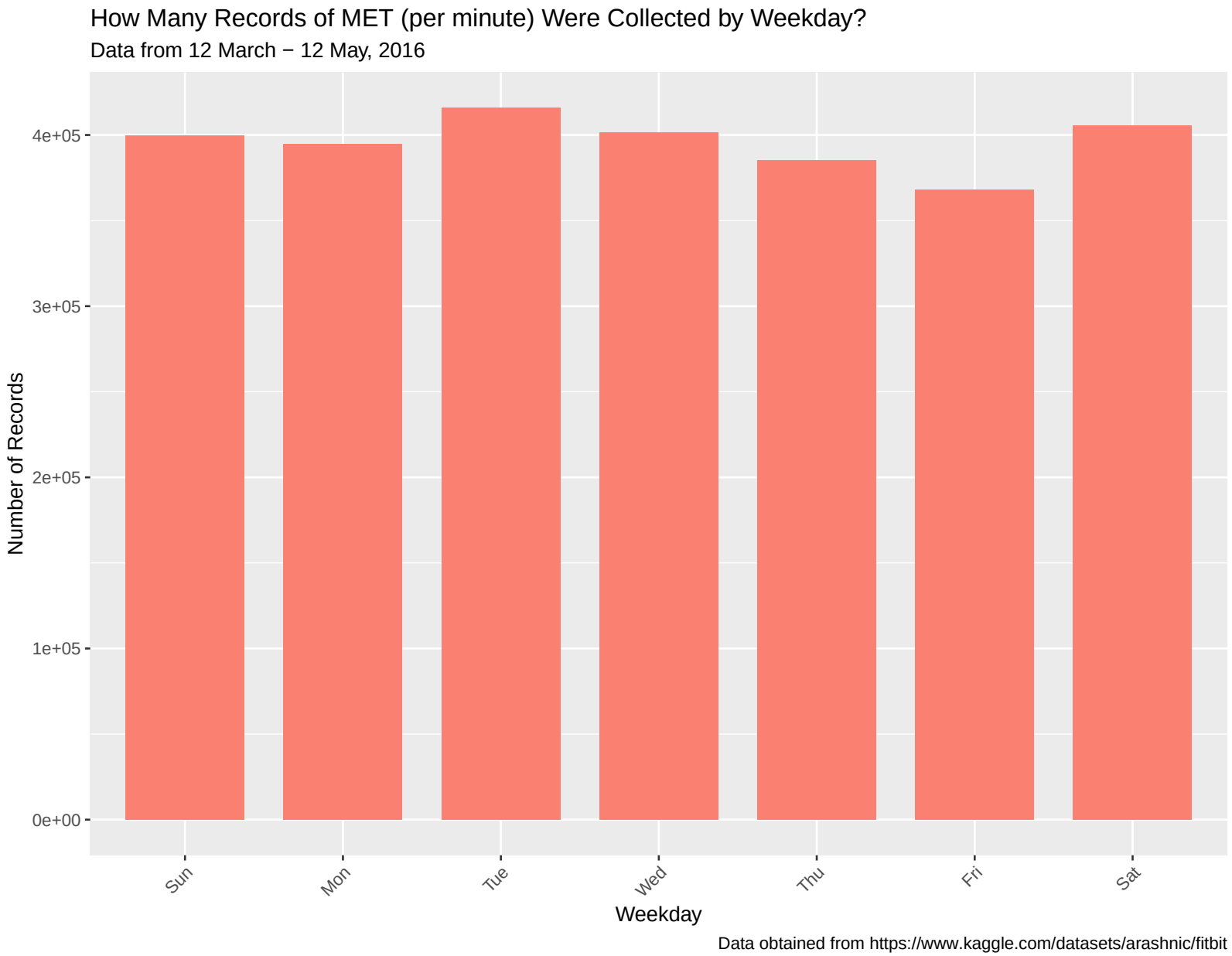

How Many Records of Heart Rate Data (per every 5−10 seconds) Were Collected by Month?
Data from 12 March − 12 May, 2016
Number of Records (in million)
2e+06
1e+06
0e+00
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

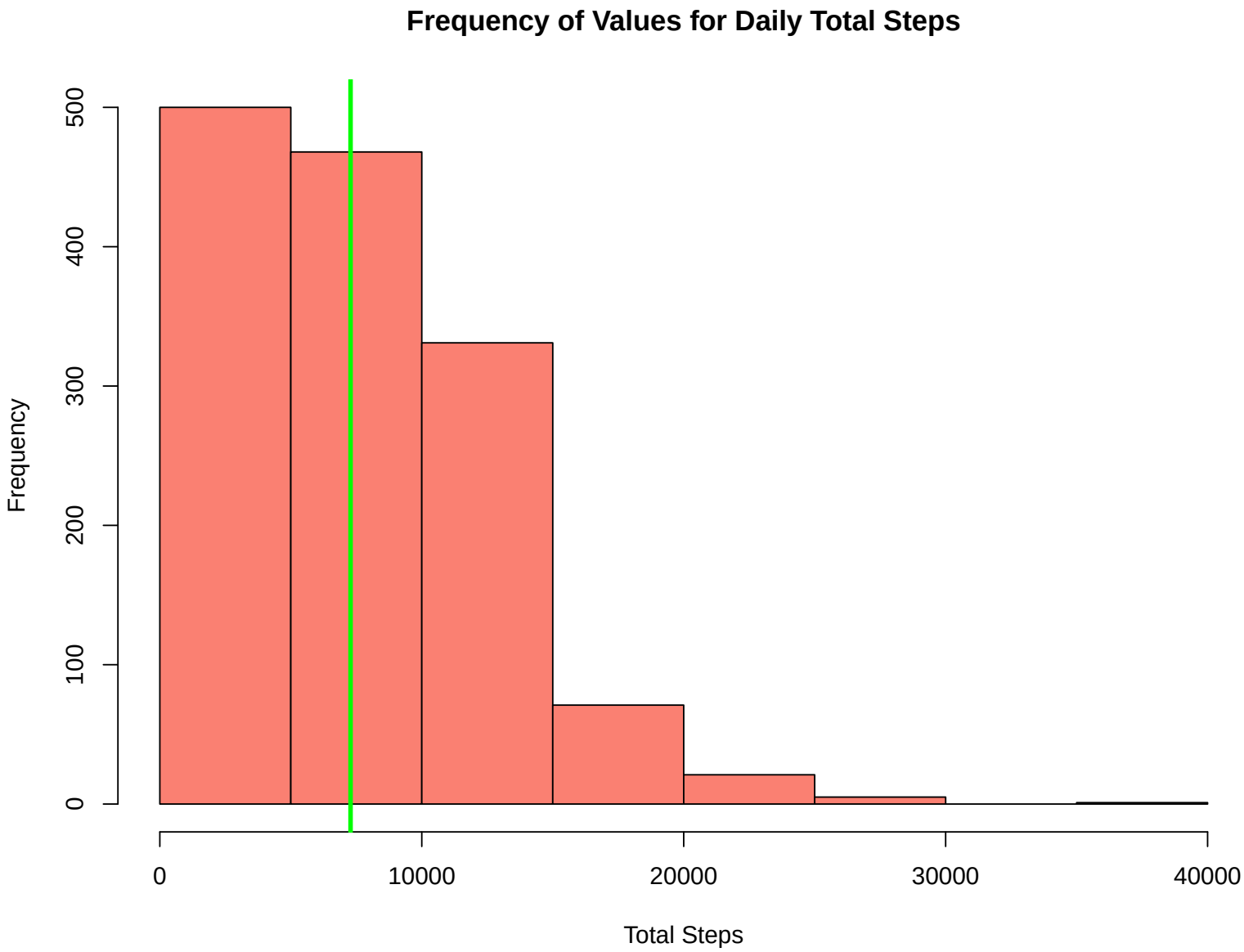

How Many Records of Heart Rate Data (per every 5−10 seconds) Were Collected by Weekday?
Data from 12 March − 12 May, 2016
6e+05
4e+05
2e+05
0e+00
Number of Records (in million)
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

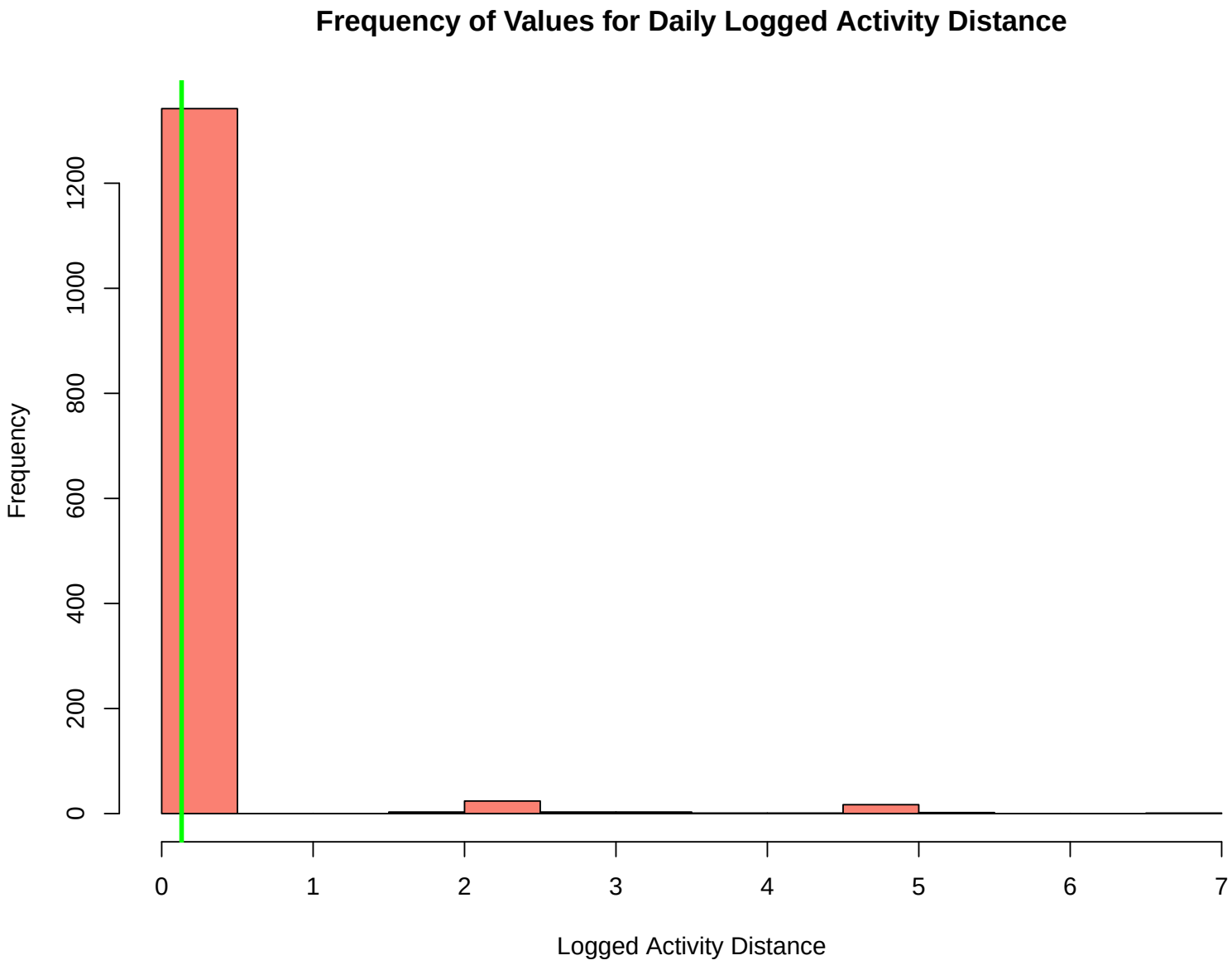

Frequency of Values for METs per minute
Frequency
0
500000
1000000
1500000
0
50
100
150
METs Per Minute

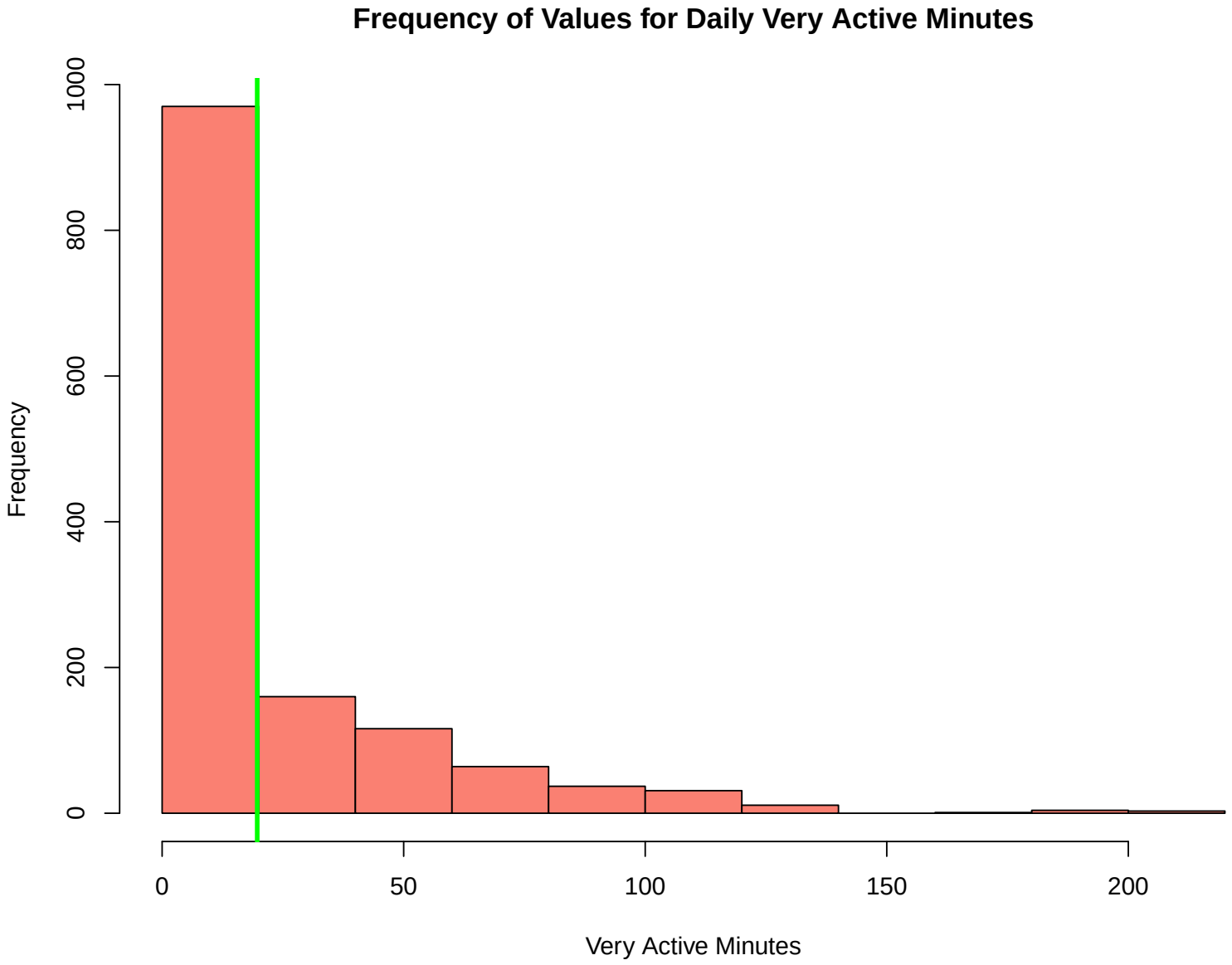

How Many Records of METs (per minute) Were Collected by Each User?
Data from 12 March − 12 May, 2016
Number of Records
75000
50000
25000
0
1503960366
1624580081
1644430081
1844505072
1927972279
2022484408
2026352035
2320127002
2347167796
2873212765
2891001357
3372868164
3977333714
4020332650
4057192912
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6290855005
6391747486
6775888955
6962181067
7007744171
7086361926
8053475328
8253242879
8378563200
8583815059
8792009665
8877689391
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

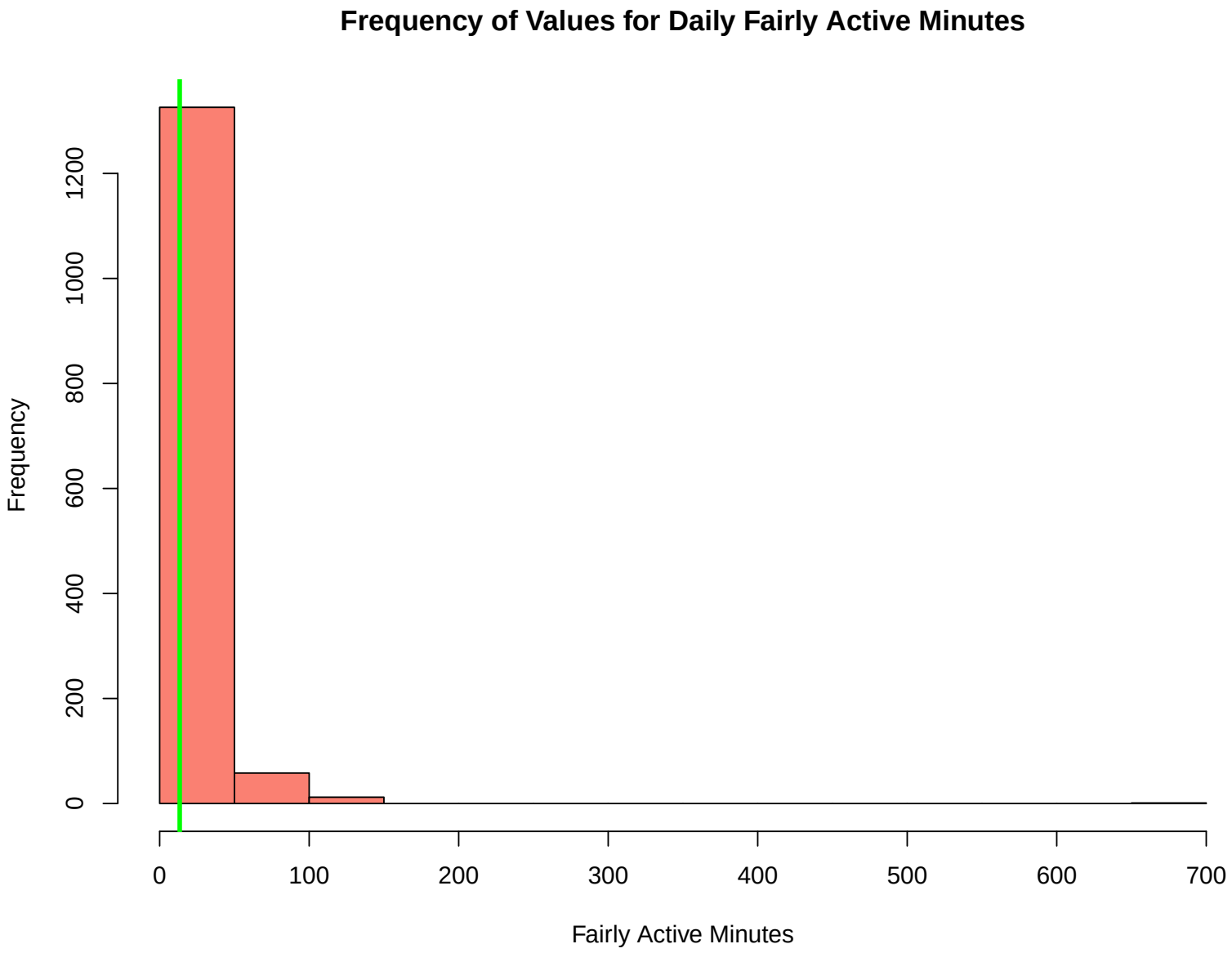

How Many Records of METs (per minute) Were Collected by Month?
Data from 12 March − 12 May, 2016
Number of Records
1e+06
5e+05
0e+00
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

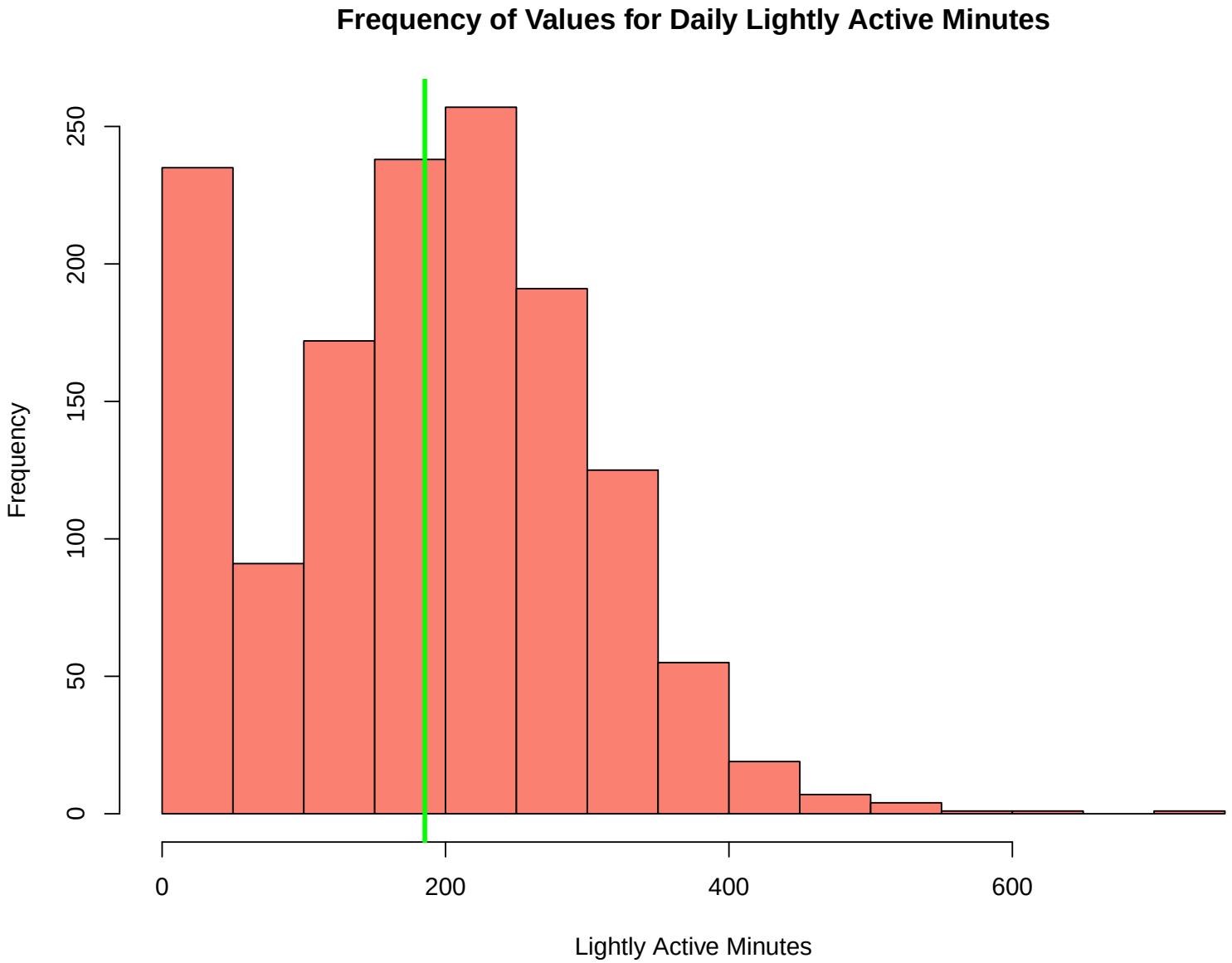

How Many Records of MET (per minute) Were Collected by Weekday?
Data from 12 March − 12 May, 2016
Number of Records
4e+05
3e+05
2e+05
1e+05
0e+00
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

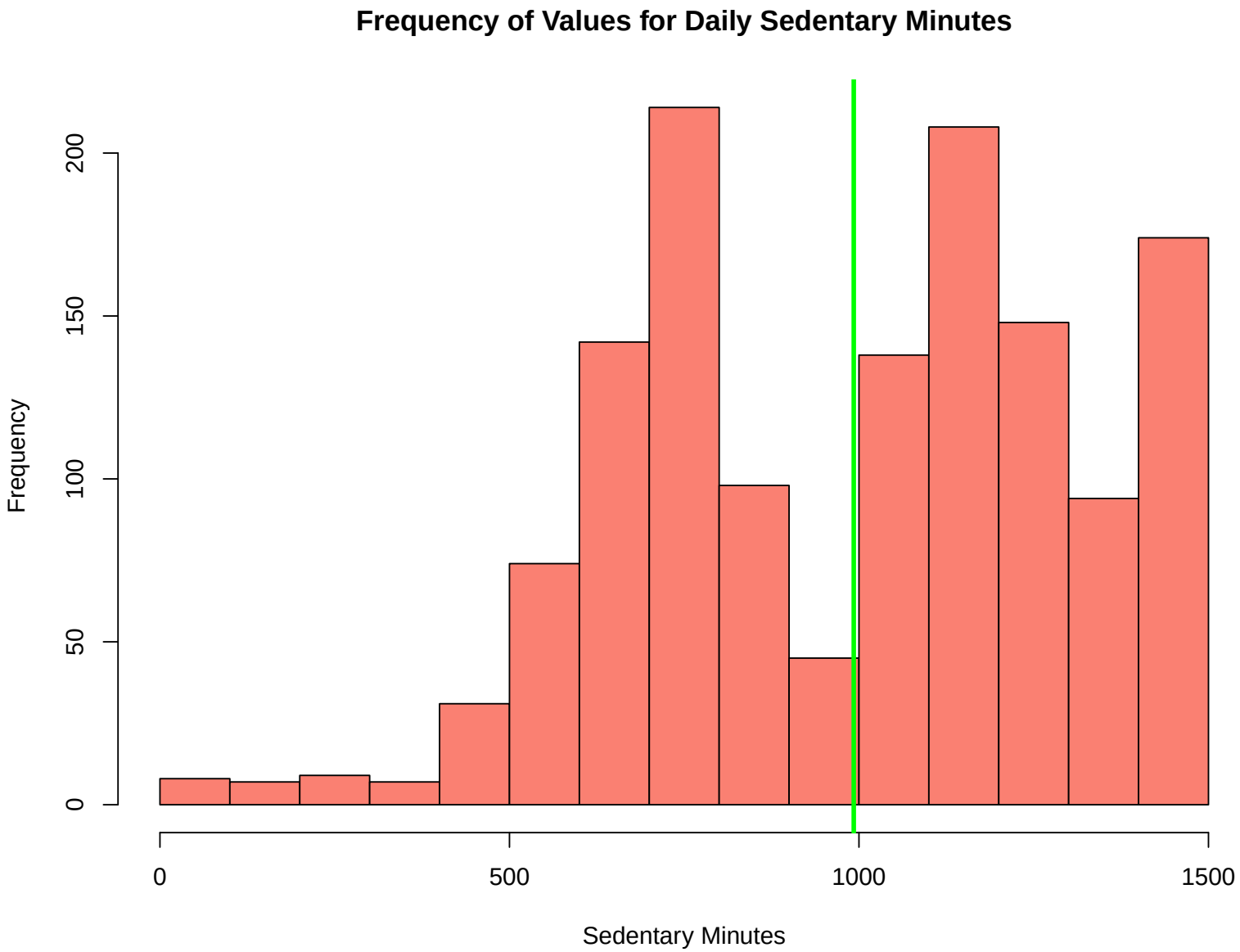

Frequency of Values for Daily Total Steps
Frequency
0
100
200
300
400
500
0
10000
20000
30000
40000
Total Steps

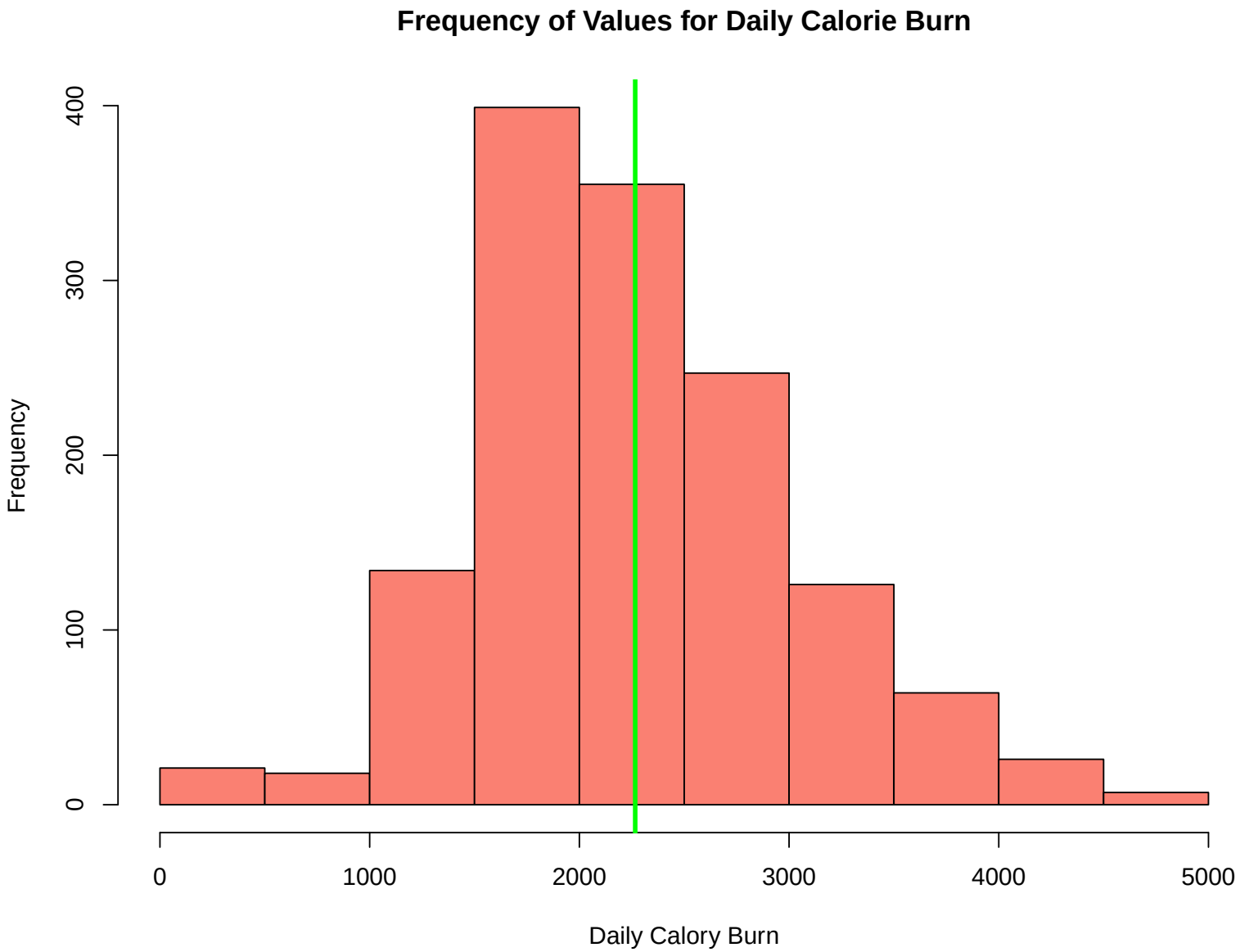

Frequency of Values for Daily Logged Activity Distance
Frequency
0
200
400
600
800
1000
1200
0
1
2
3
4
5
6
7
Logged Activity Distance

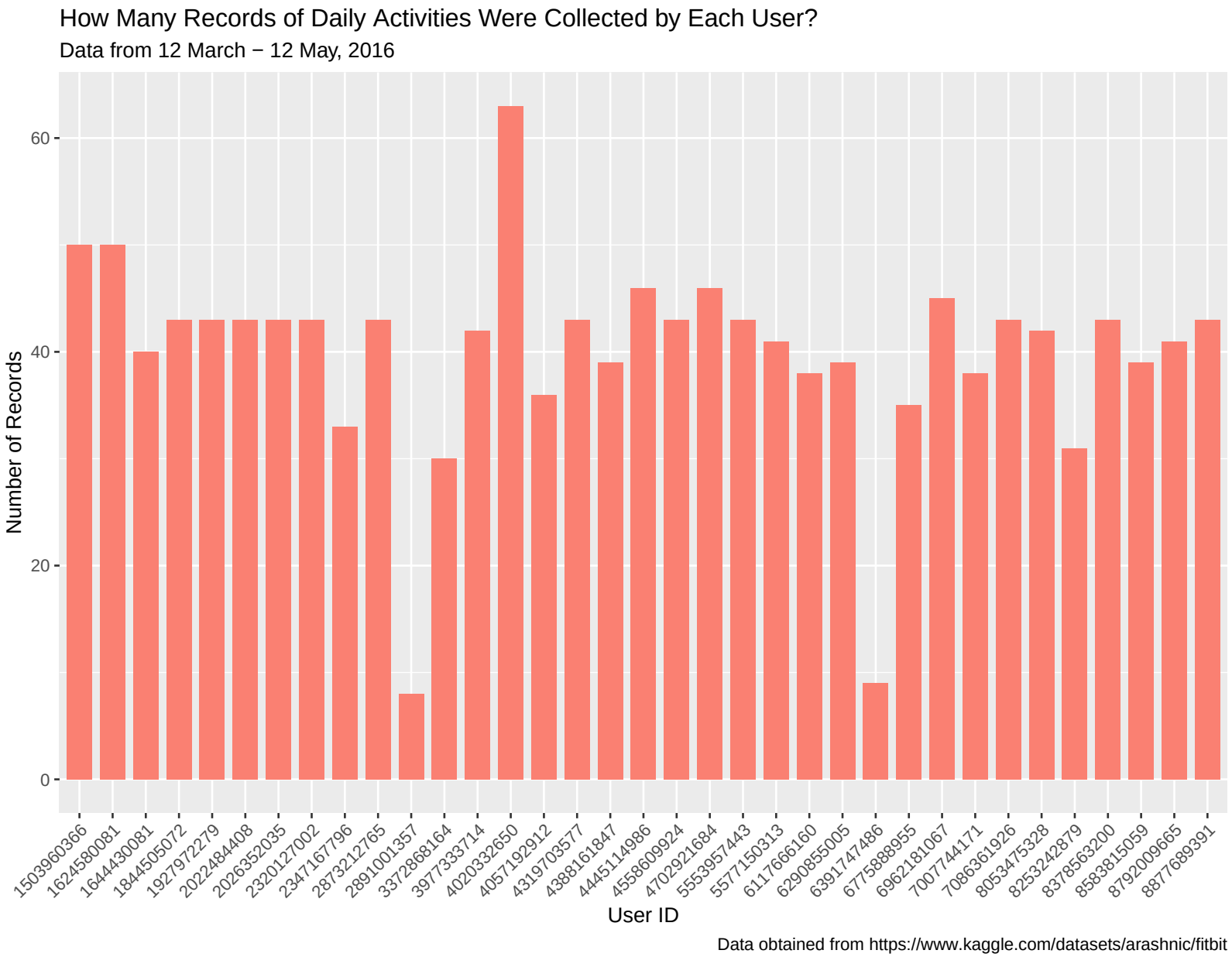

Frequency of Values for Daily Very Active Minutes
Frequency
0
200
400
600
800
1000
0
50
100
150
200
Very Active Minutes

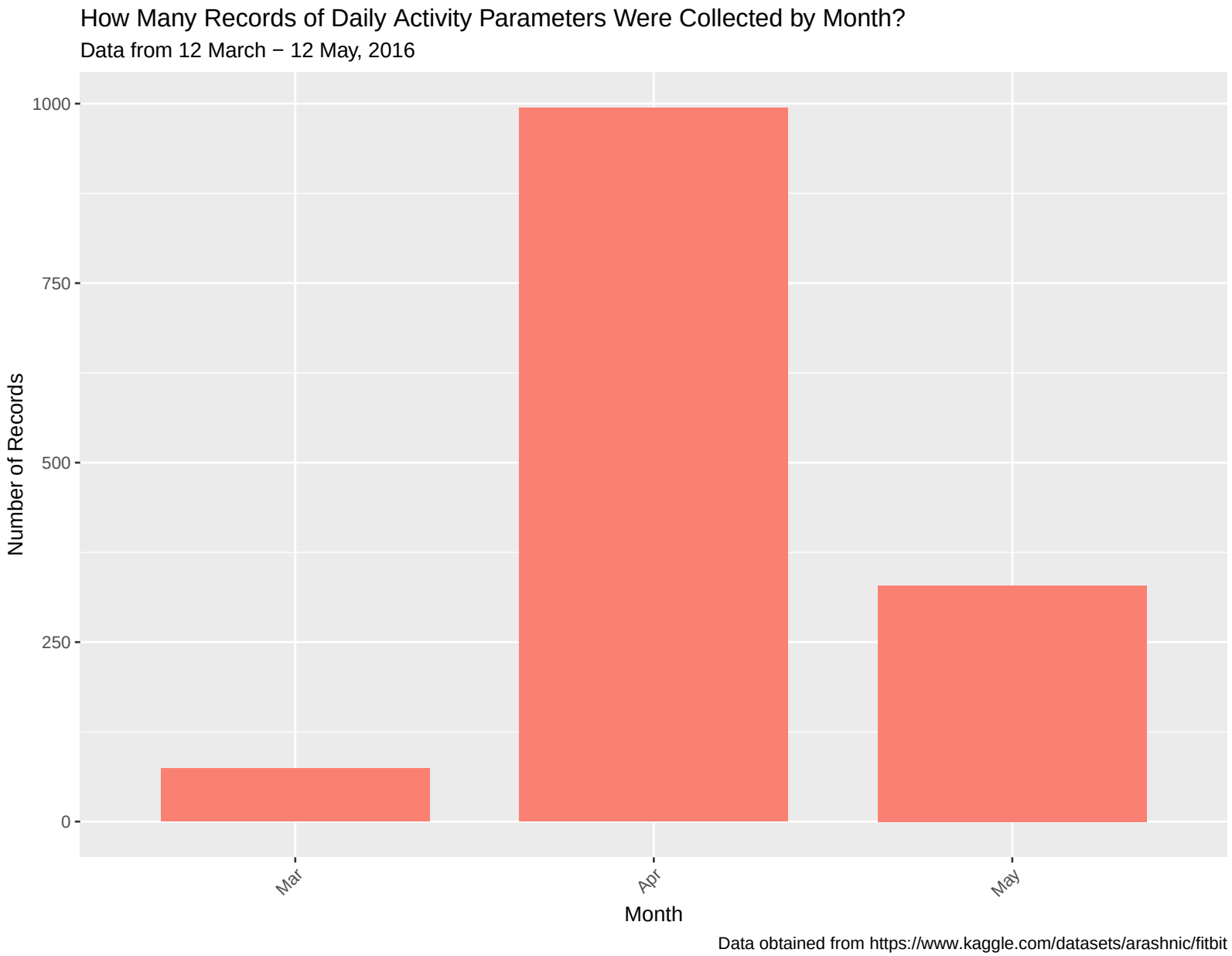

Frequency of Values for Daily Fairly Active Minutes
Frequency
0
200
400
600
800
1000
1200
0
100
200
300
400
500
600
700
Fairly Active Minutes

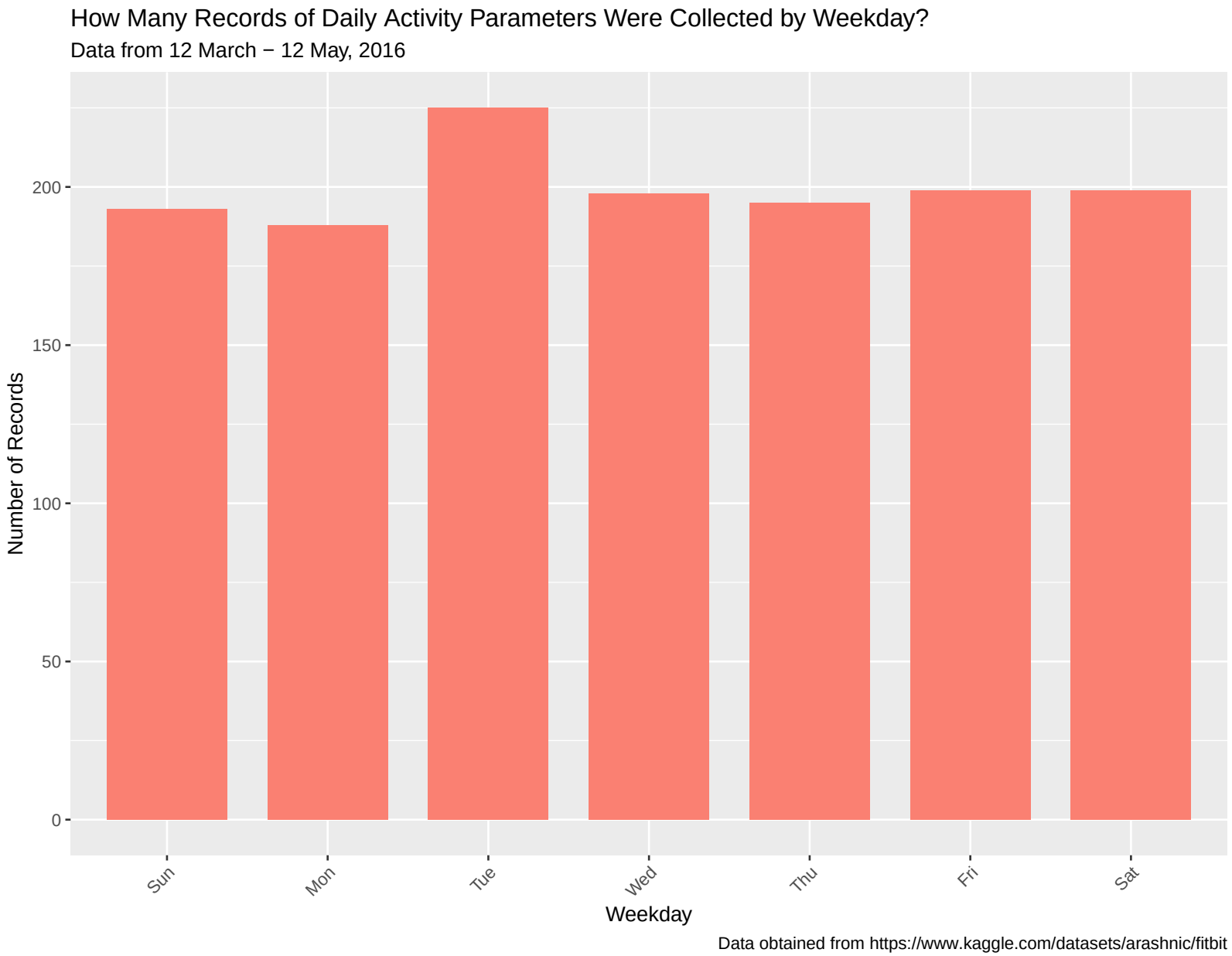

Frequency of Values for Daily Lightly Active Minutes
Frequency
0
50
100
150
200
250
0
200
400
600
Lightly Active Minutes

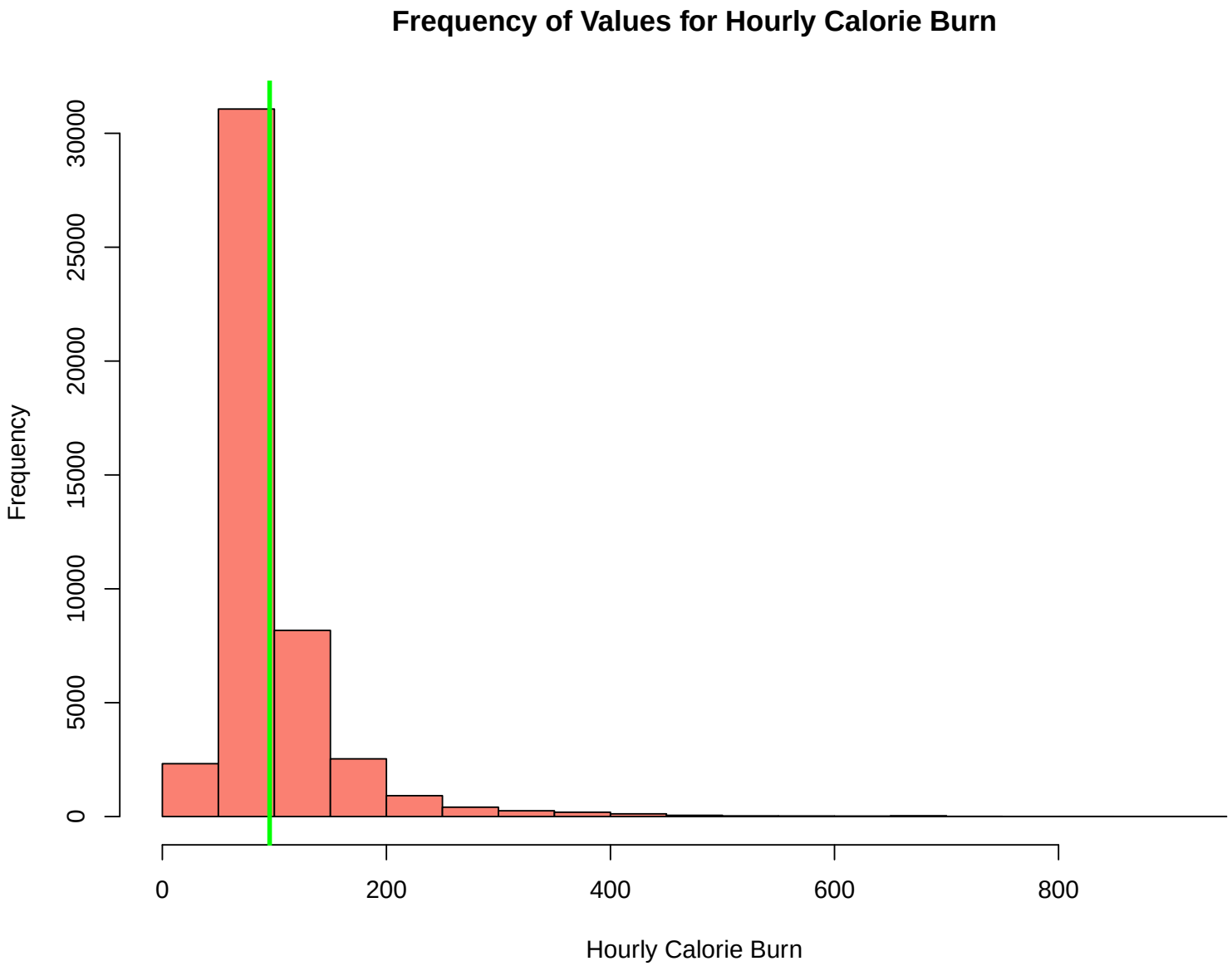

Frequency of Values for Daily Sedentary Minutes
Frequency
0
50
100
150
200
0
500
1000
1500
Sedentary Minutes

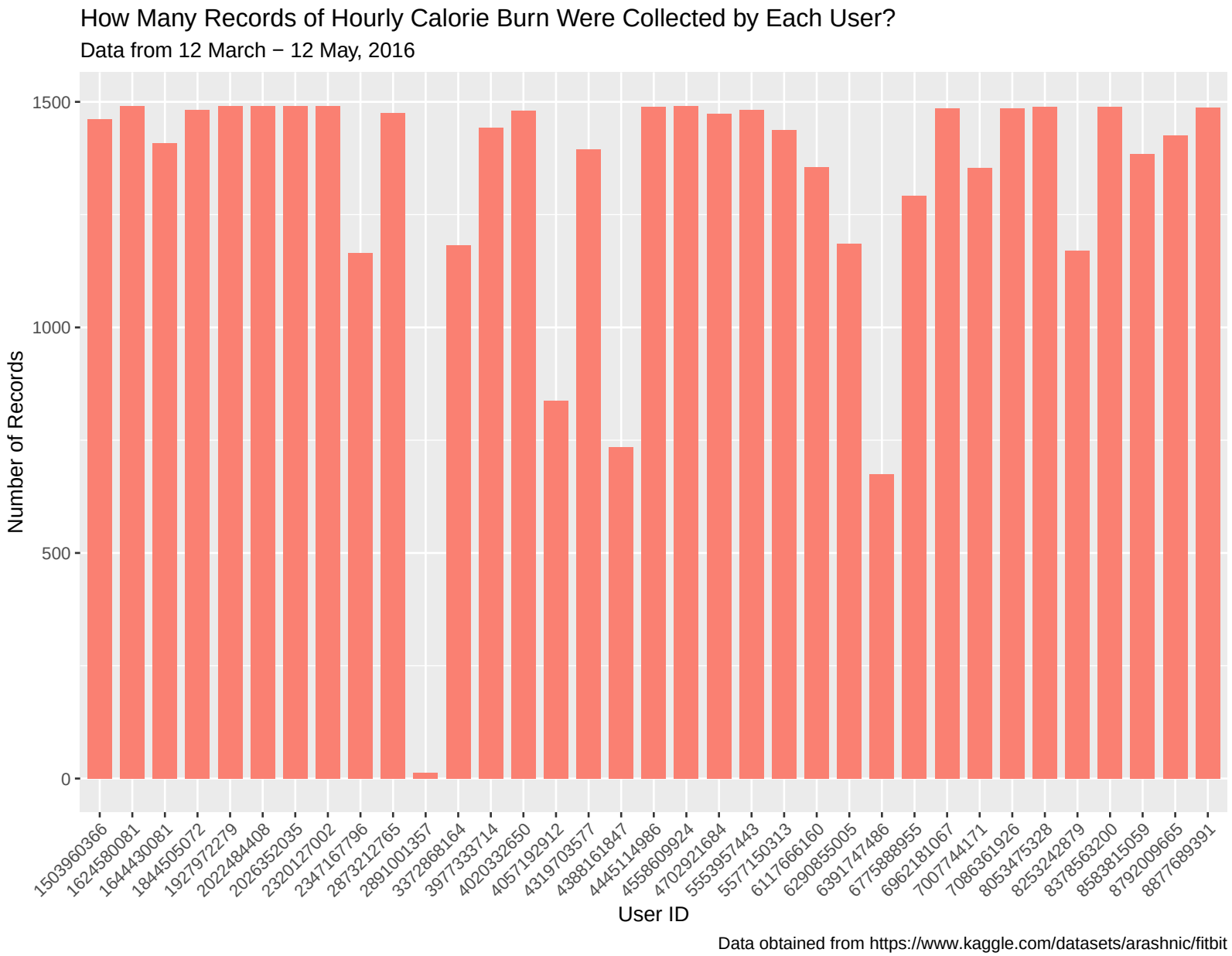

Frequency of Values for Daily Calorie Burn
Frequency
0
100
200
300
400
0
1000
2000
3000
4000
5000
Daily Calory Burn

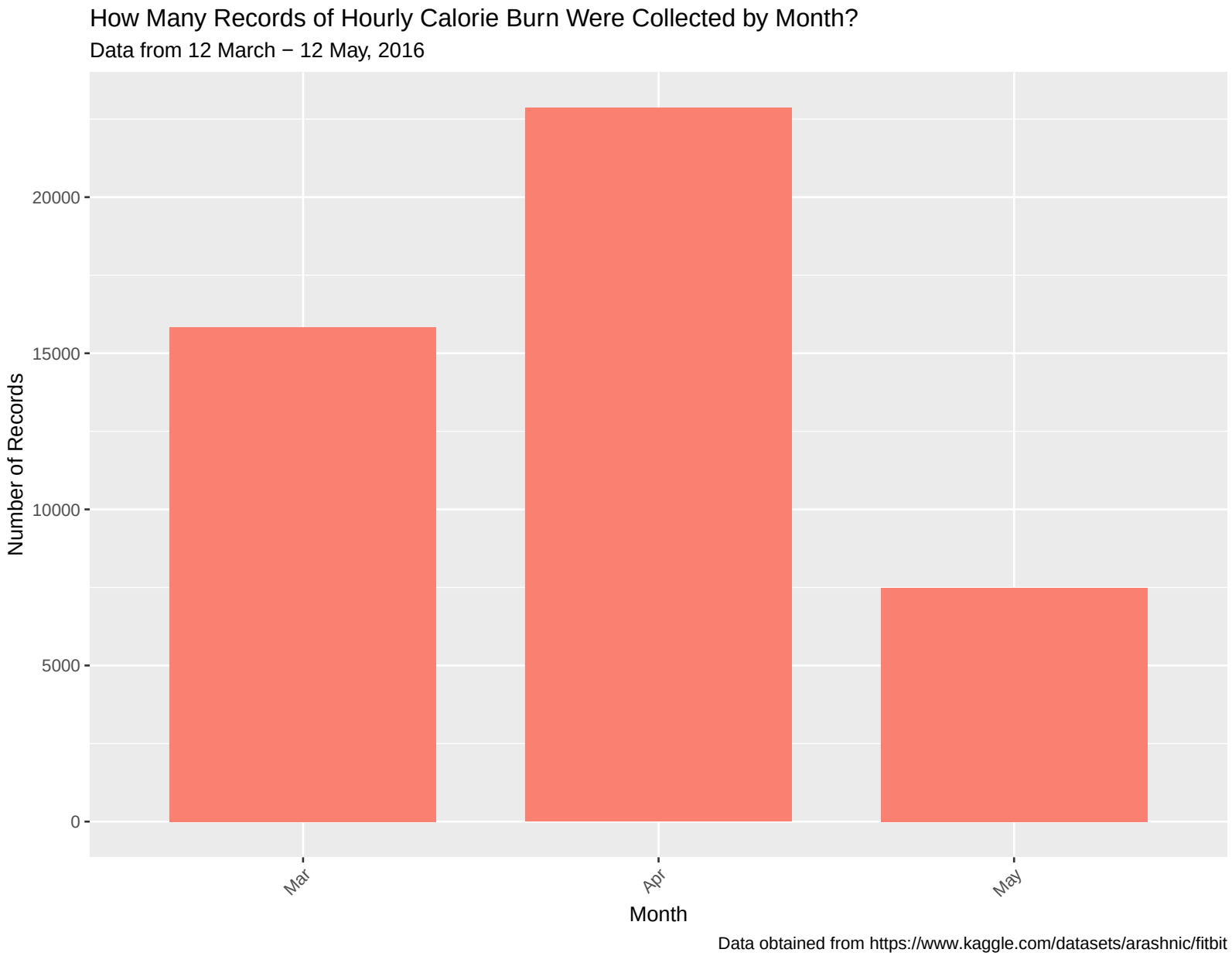

How Many Records of Daily Activities Were Collected by Each User?
Data from 12 March − 12 May, 2016
Number of Records
0
20
40
60
1503960366
1624580081
1644430081
1844505072
1927972279
2022484408
2026352035
2320127002
2347167796
2873212765
2891001357
3372868164
3977333714
4020332650
4057192912
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6290855005
6391747486
6775888955
6962181067
7007744171
7086361926
8053475328
8253242879
8378563200
8583815059
8792009665
8877689391
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

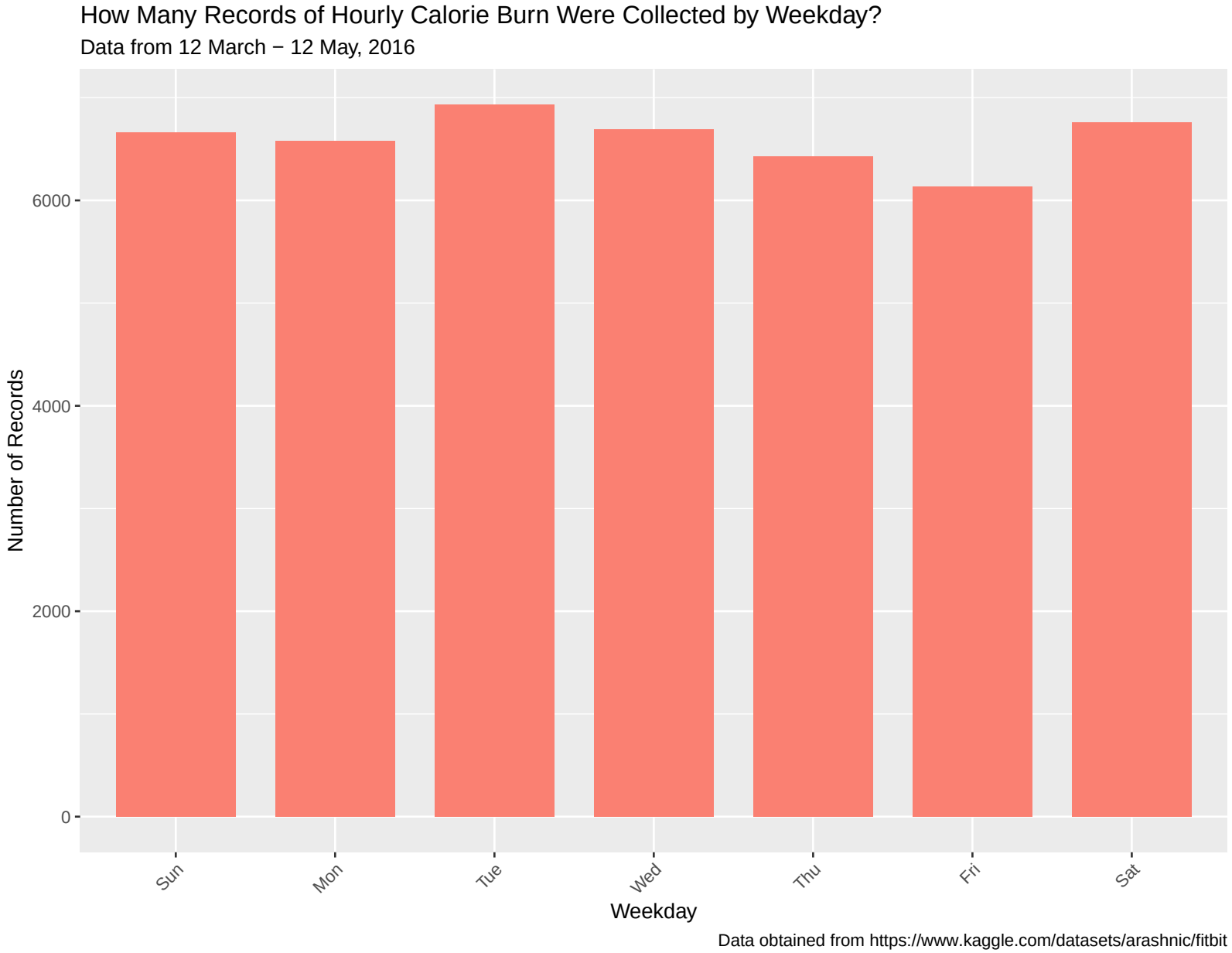

How Many Records of Daily Activity Parameters Were Collected by Month?
Data from 12 March − 12 May, 2016
1000
750
500
250
0
Number of Records
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

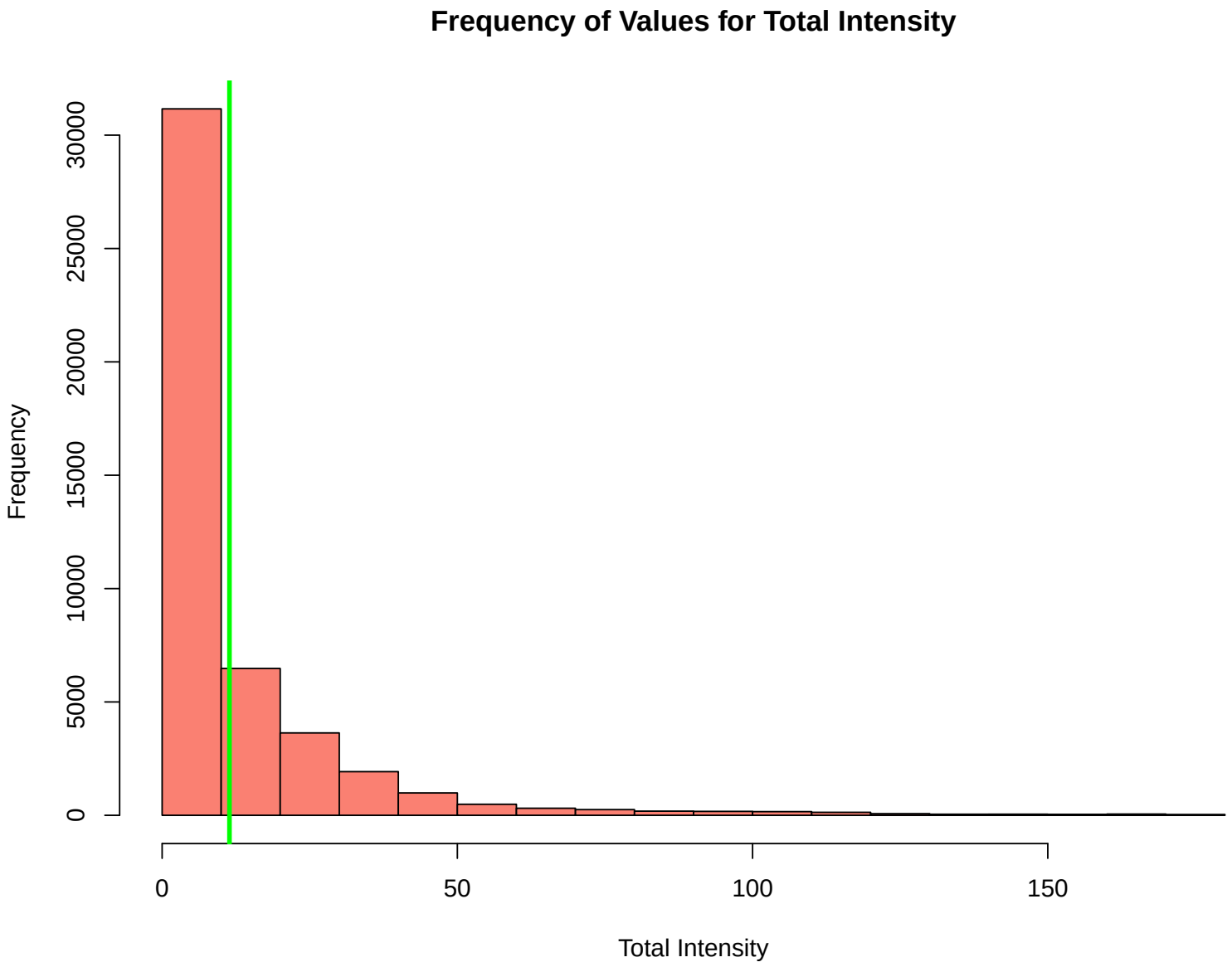

How Many Records of Daily Activity Parameters Were Collected by Weekday?
Data from 12 March − 12 May, 2016
Number of Records
200
150
100
50
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

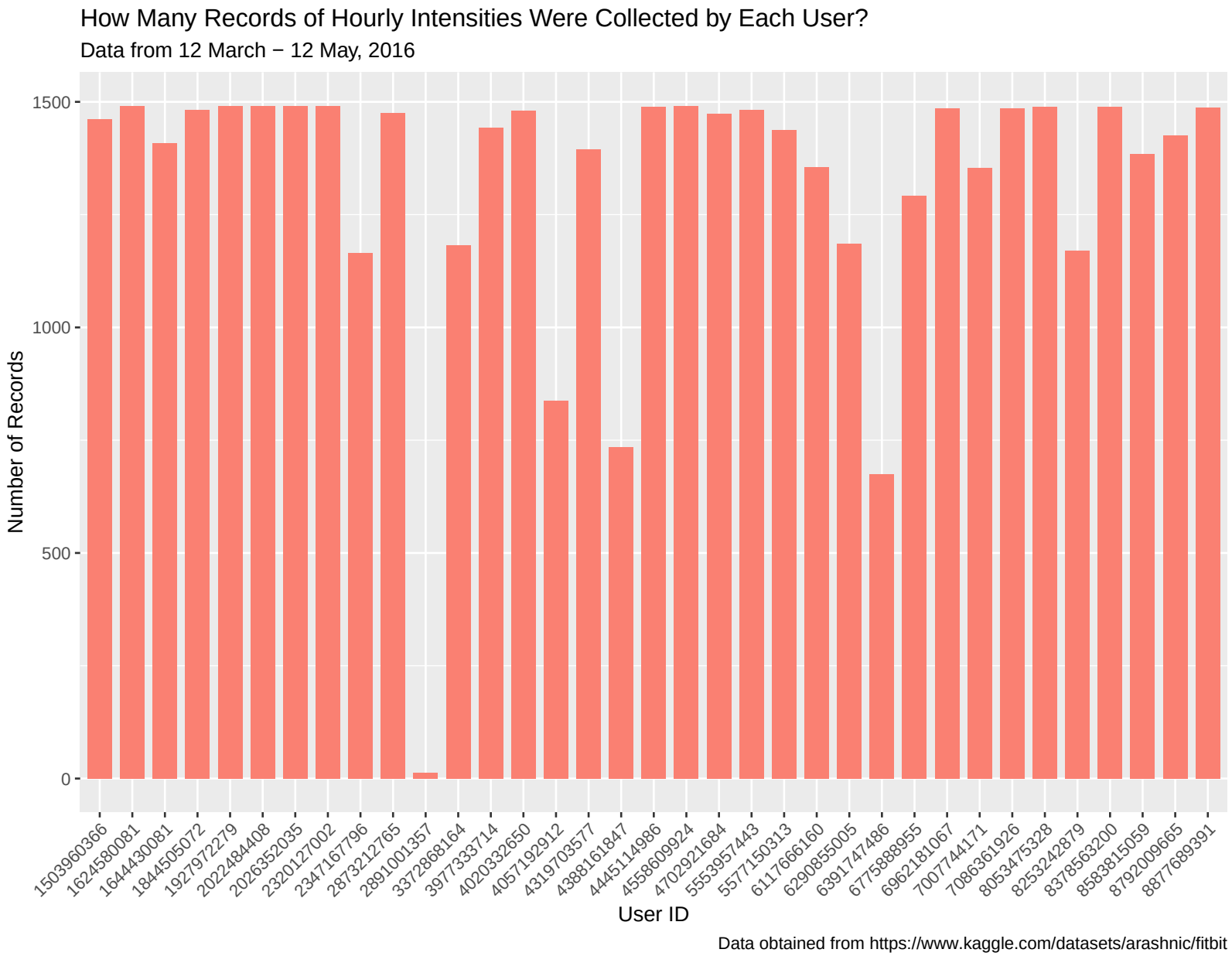

Frequency of Values for Hourly Calorie Burn
Frequency
0
5000
10000
15000
20000
25000
30000
0
200
400
600
800
Hourly Calorie Burn

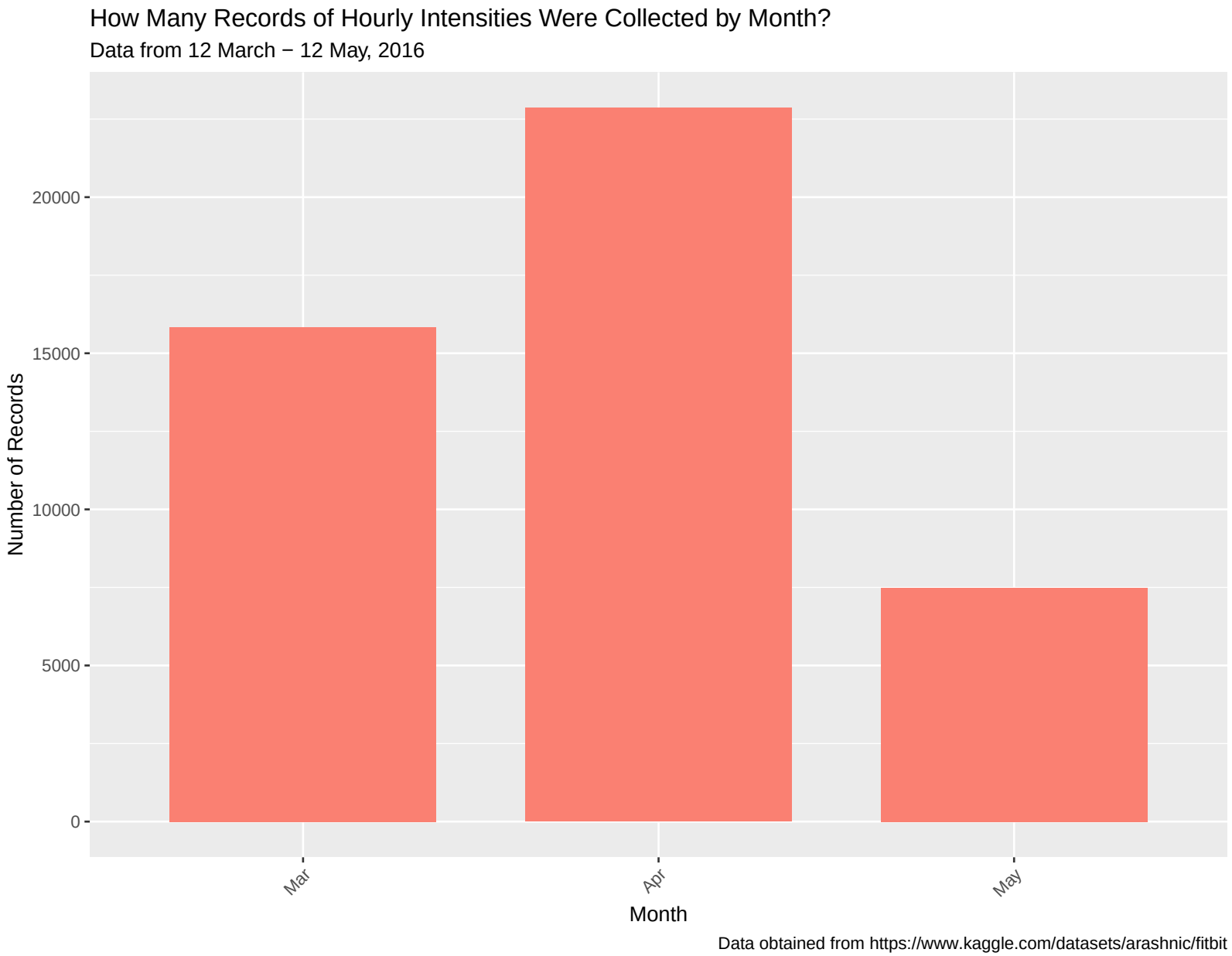

How Many Records of Hourly Calorie Burn Were Collected by Each User?
Data from 12 March − 12 May, 2016
Number of Records
1500
1000
500
0
1503960366
1624580081
1644430081
1844505072
1927972279
2022484408
2026352035
2320127002
2347167796
2873212765
2891001357
3372868164
3977333714
4020332650
4057192912
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6290855005
6391747486
6775888955
6962181067
7007744171
7086361926
8053475328
8253242879
8378563200
8583815059
8792009665
8877689391
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

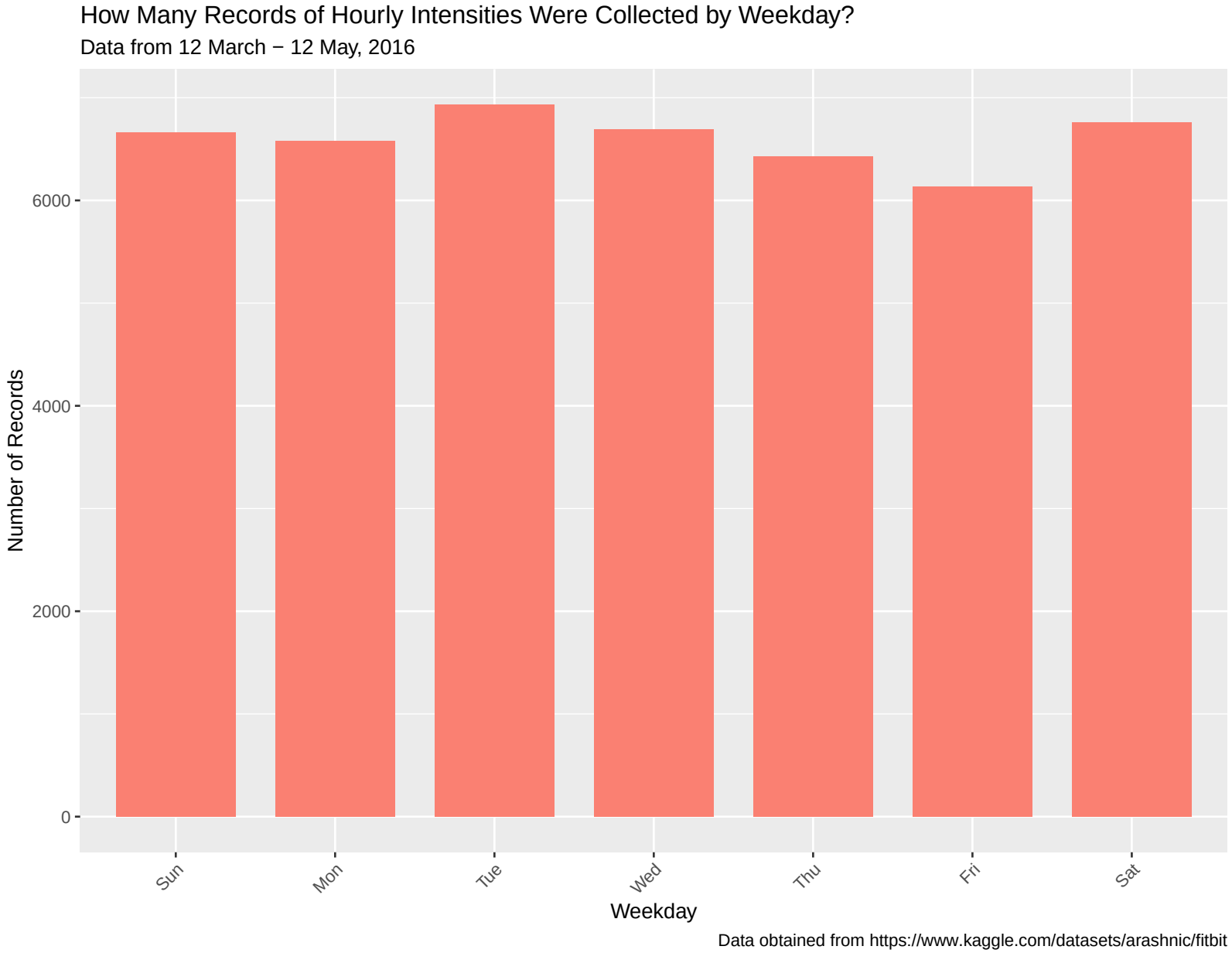

How Many Records of Hourly Calorie Burn Were Collected by Month?
Data from 12 March − 12 May, 2016
Number of Records
20000
15000
10000
5000
0
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

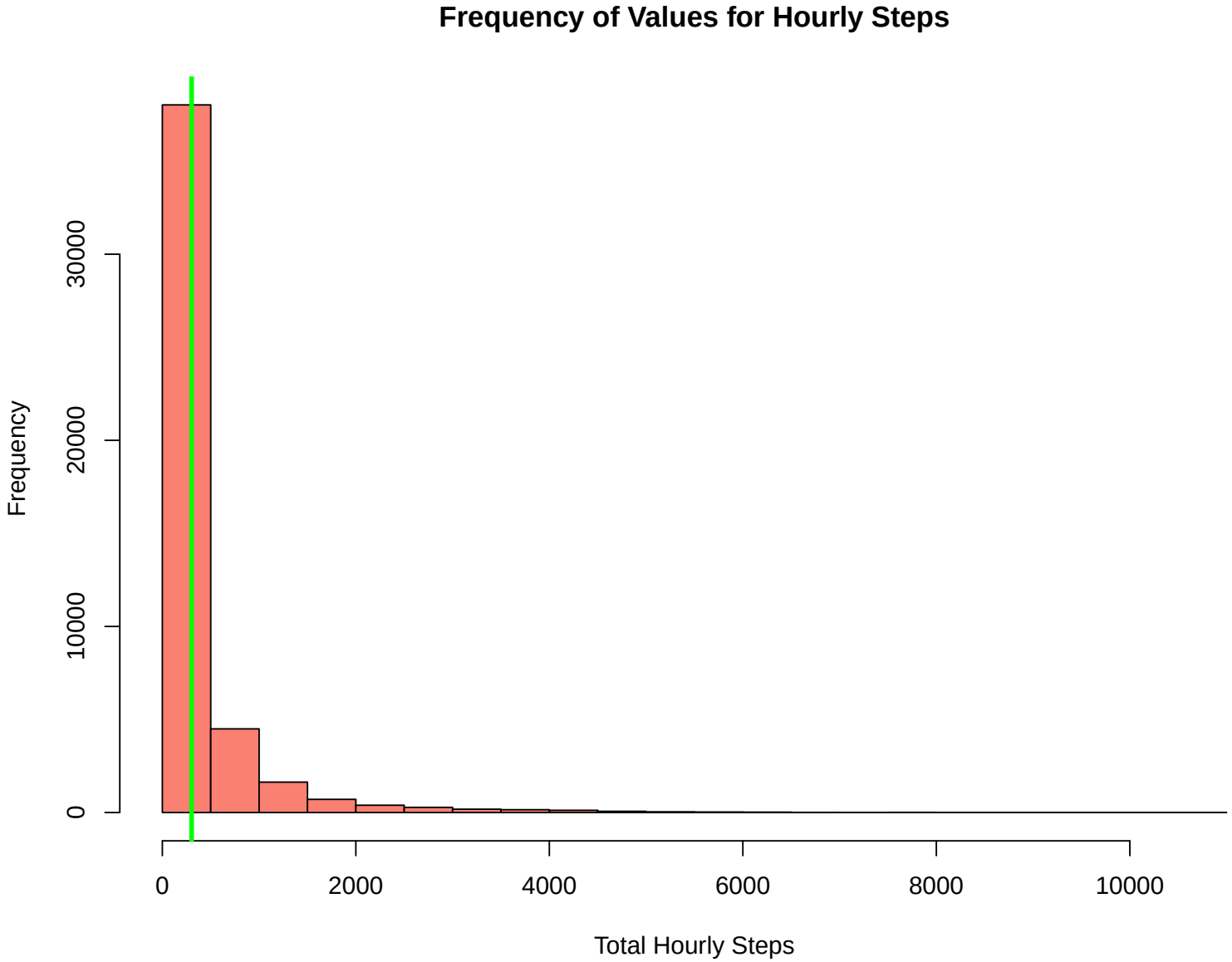

How Many Records of Hourly Calorie Burn Were Collected by Weekday?
Data from 12 March − 12 May, 2016
Number of Records
6000
4000
2000
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

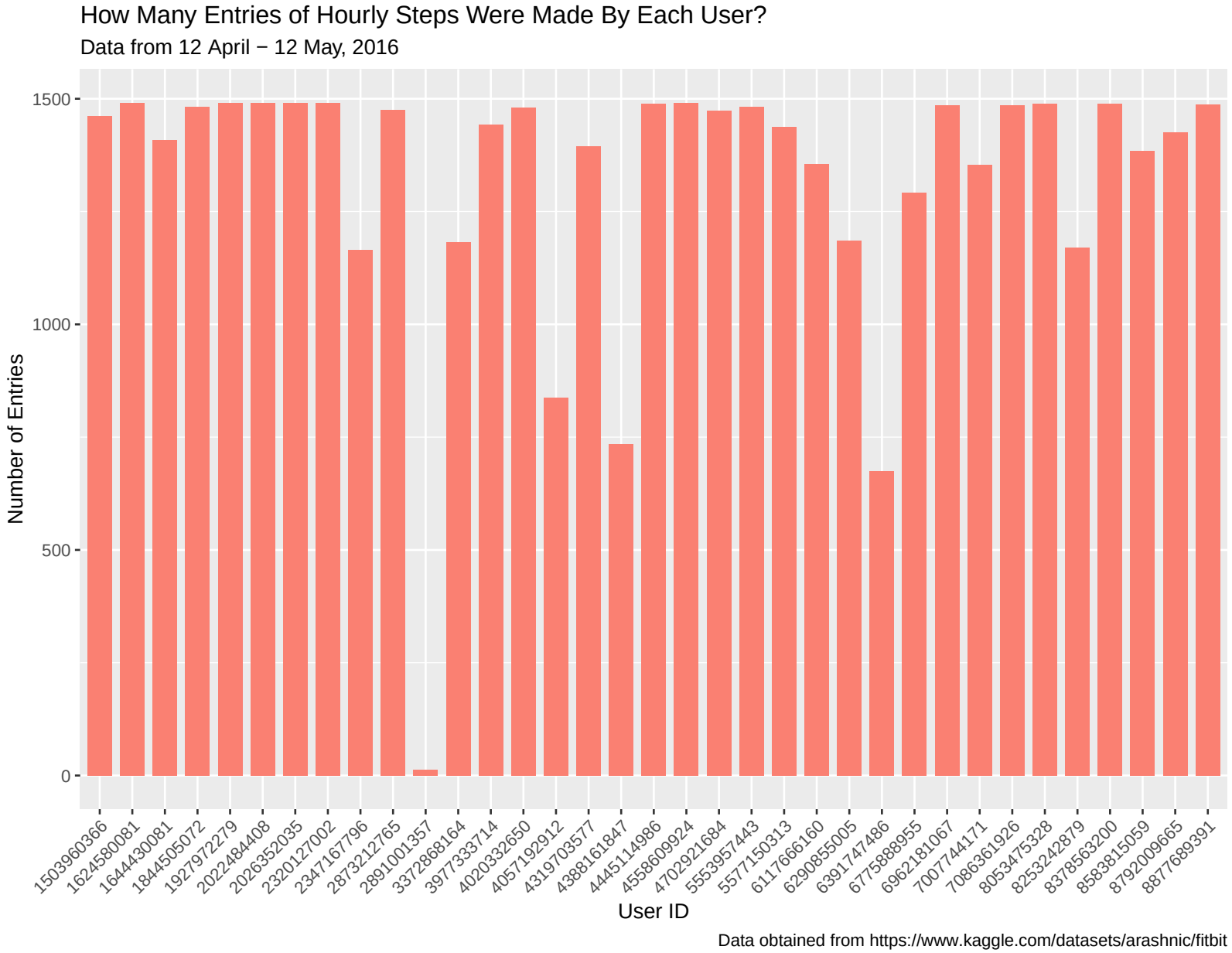

Frequency of Values for Total Intensity
Frequency
0
5000
10000
15000
20000
25000
30000
0
50
100
150
Total Intensity

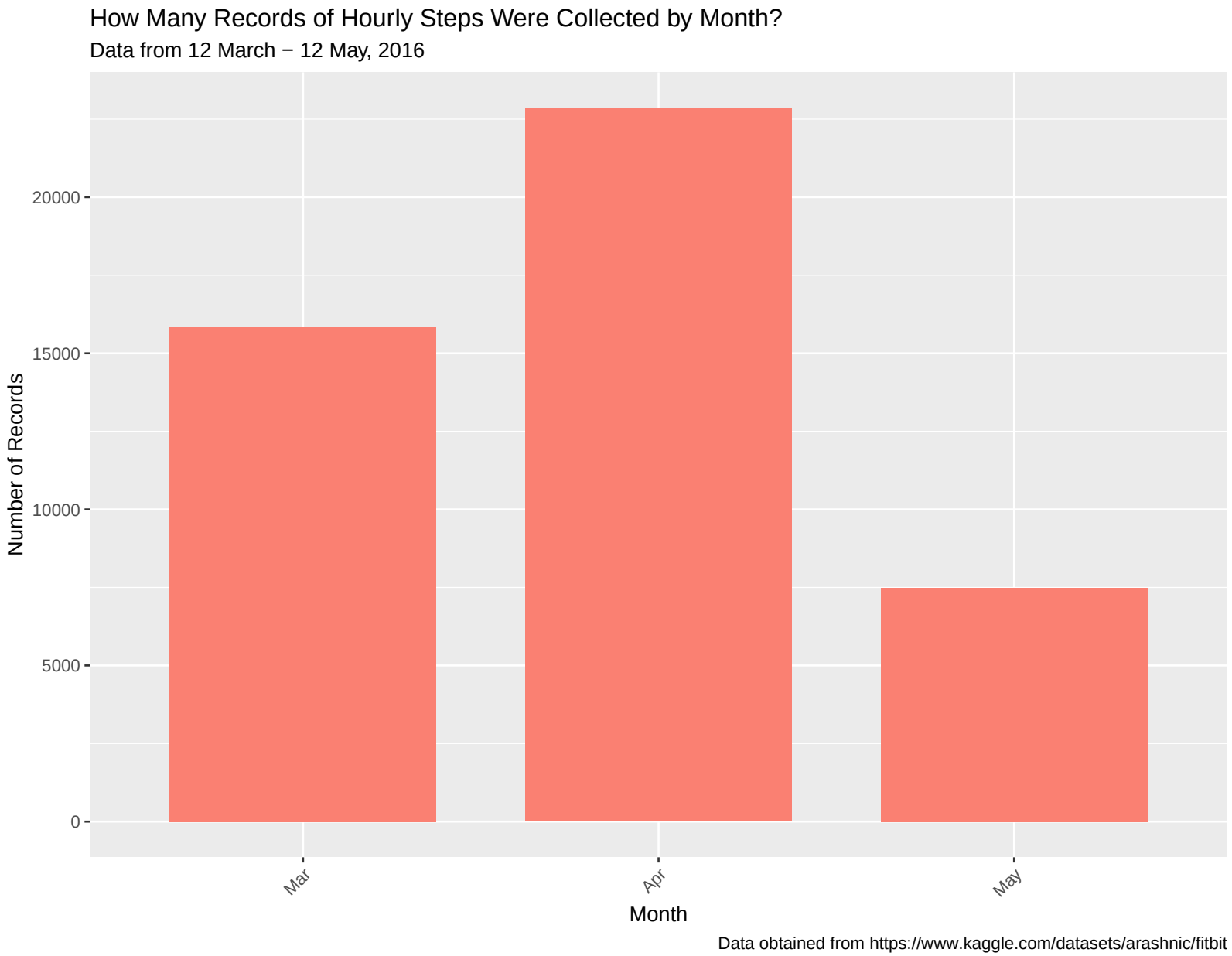

How Many Records of Hourly Intensities Were Collected by Each User?
Data from 12 March − 12 May, 2016
Number of Records
1500
1000
500
0
1503960366
1624580081
1644430081
1844505072
1927972279
2022484408
2026352035
2320127002
2347167796
2873212765
2891001357
3372868164
3977333714
4020332650
4057192912
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6290855005
6391747486
6775888955
6962181067
7007744171
7086361926
8053475328
8253242879
8378563200
8583815059
8792009665
8877689391
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

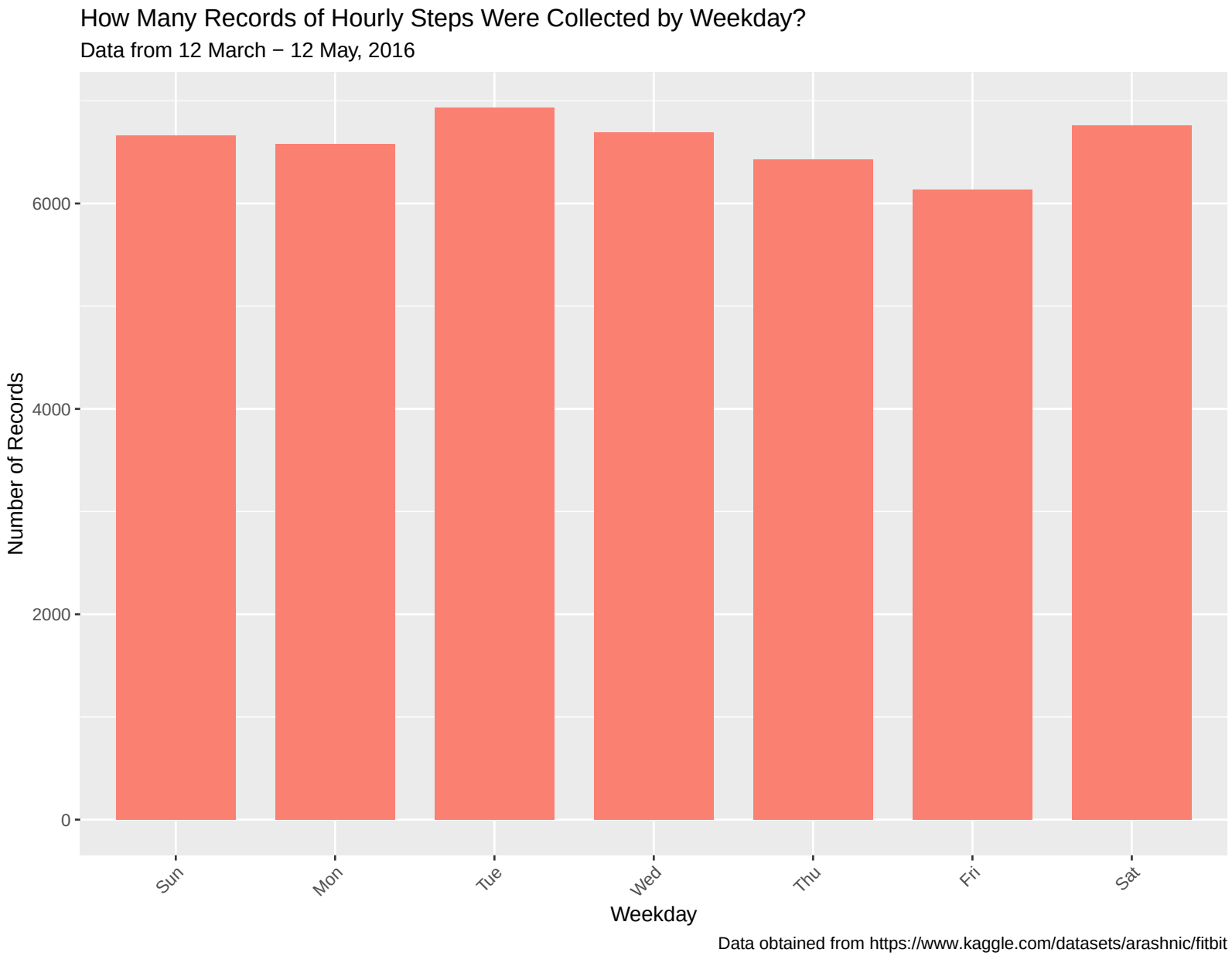

How Many Records of Hourly Intensities Were Collected by Month?
Data from 12 March − 12 May, 2016
Number of Records
20000
15000
10000
5000
0
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

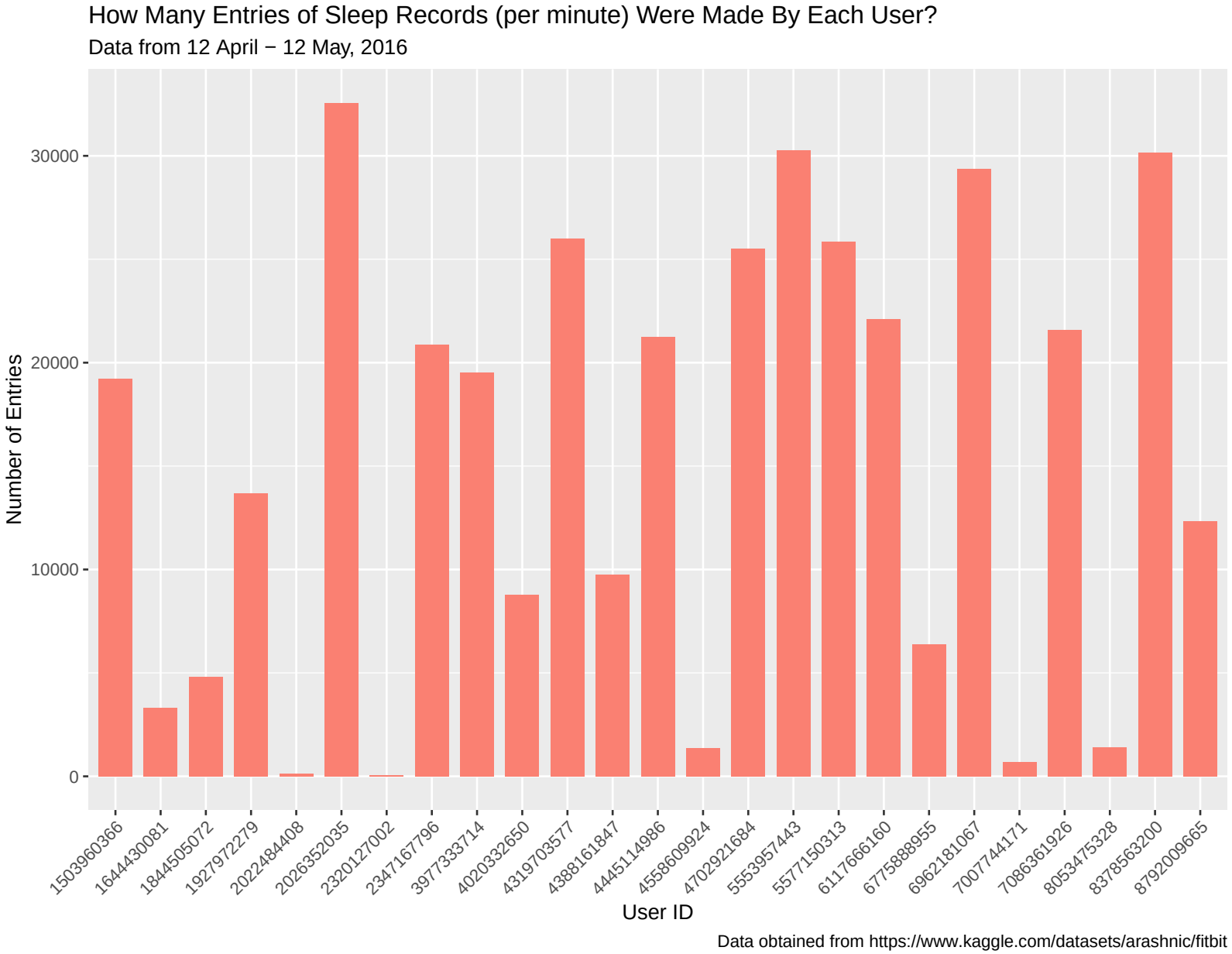

How Many Records of Hourly Intensities Were Collected by Weekday?
Data from 12 March − 12 May, 2016
Number of Records
6000
4000
2000
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

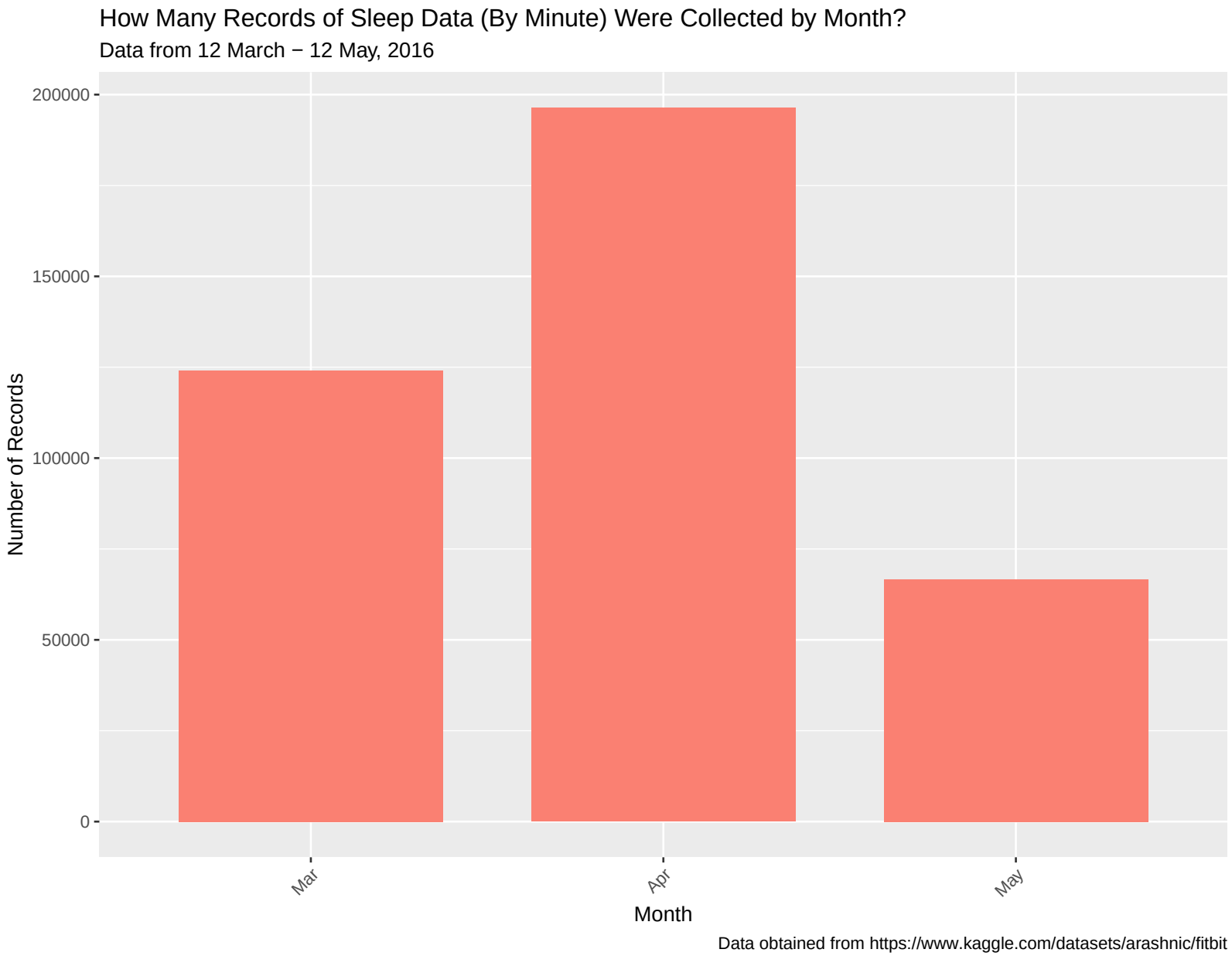

Frequency of Values for Hourly Steps
Frequency
0
10000
20000
30000
0
2000
4000
6000
8000
10000
Total Hourly Steps

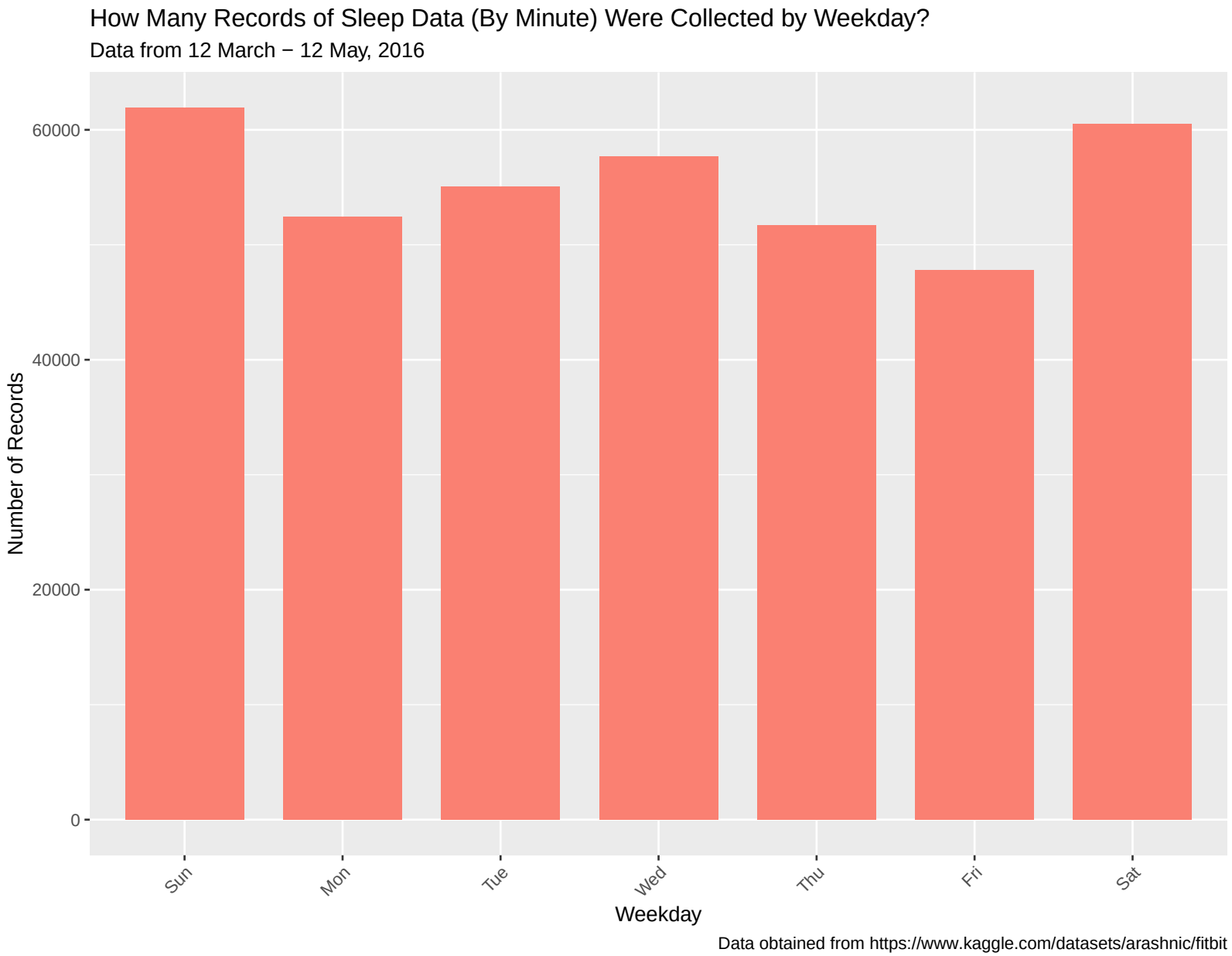

How Many Entries of Hourly Steps Were Made By Each User?
Data from 12 April − 12 May, 2016
Number of Entries
1500
1000
500
0
1503960366
1624580081
1644430081
1844505072
1927972279
2022484408
2026352035
2320127002
2347167796
2873212765
2891001357
3372868164
3977333714
4020332650
4057192912
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6290855005
6391747486
6775888955
6962181067
7007744171
7086361926
8053475328
8253242879
8378563200
8583815059
8792009665
8877689391
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

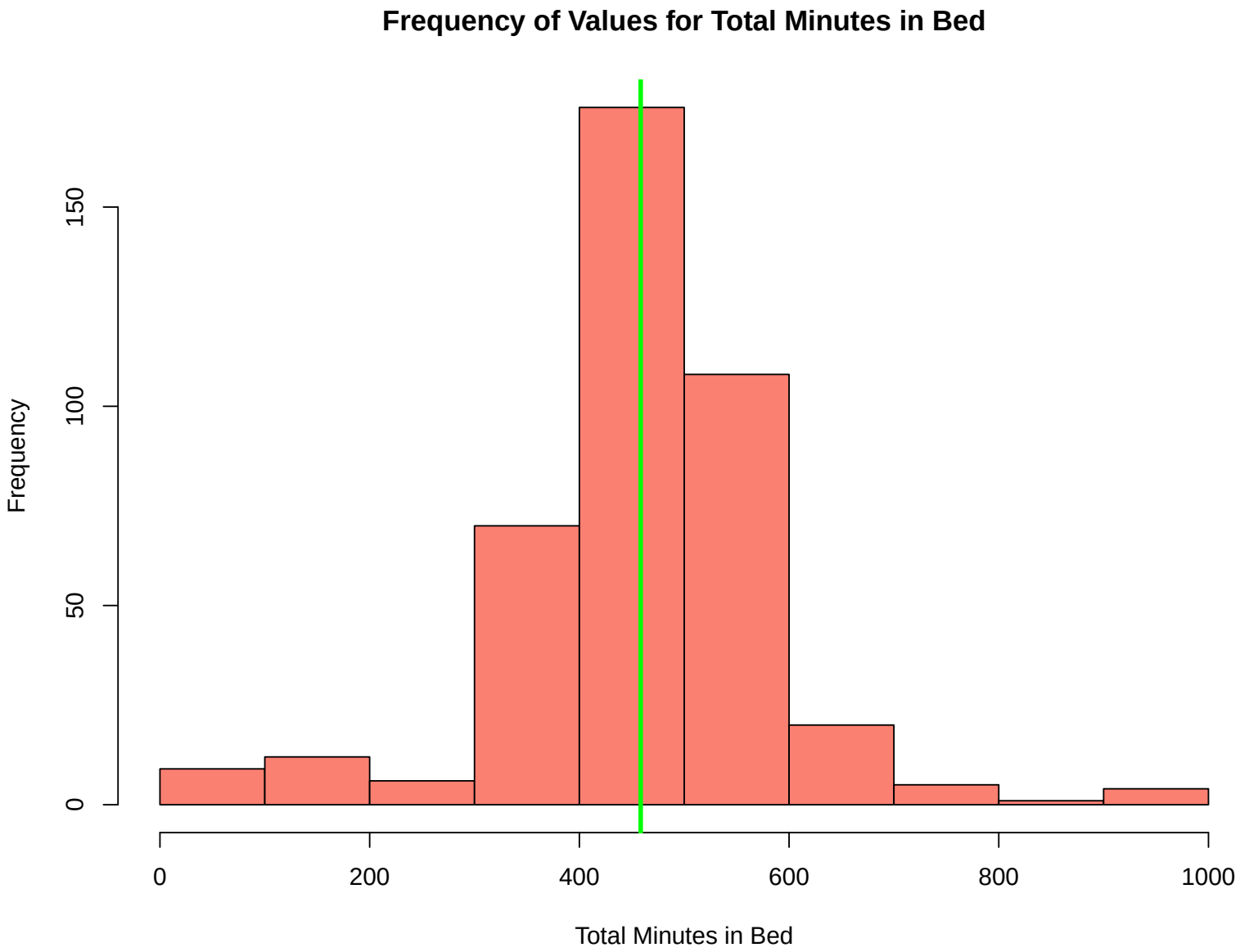

How Many Records of Hourly Steps Were Collected by Month?
Data from 12 March − 12 May, 2016
Number of Records
20000
15000
10000
5000
0
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

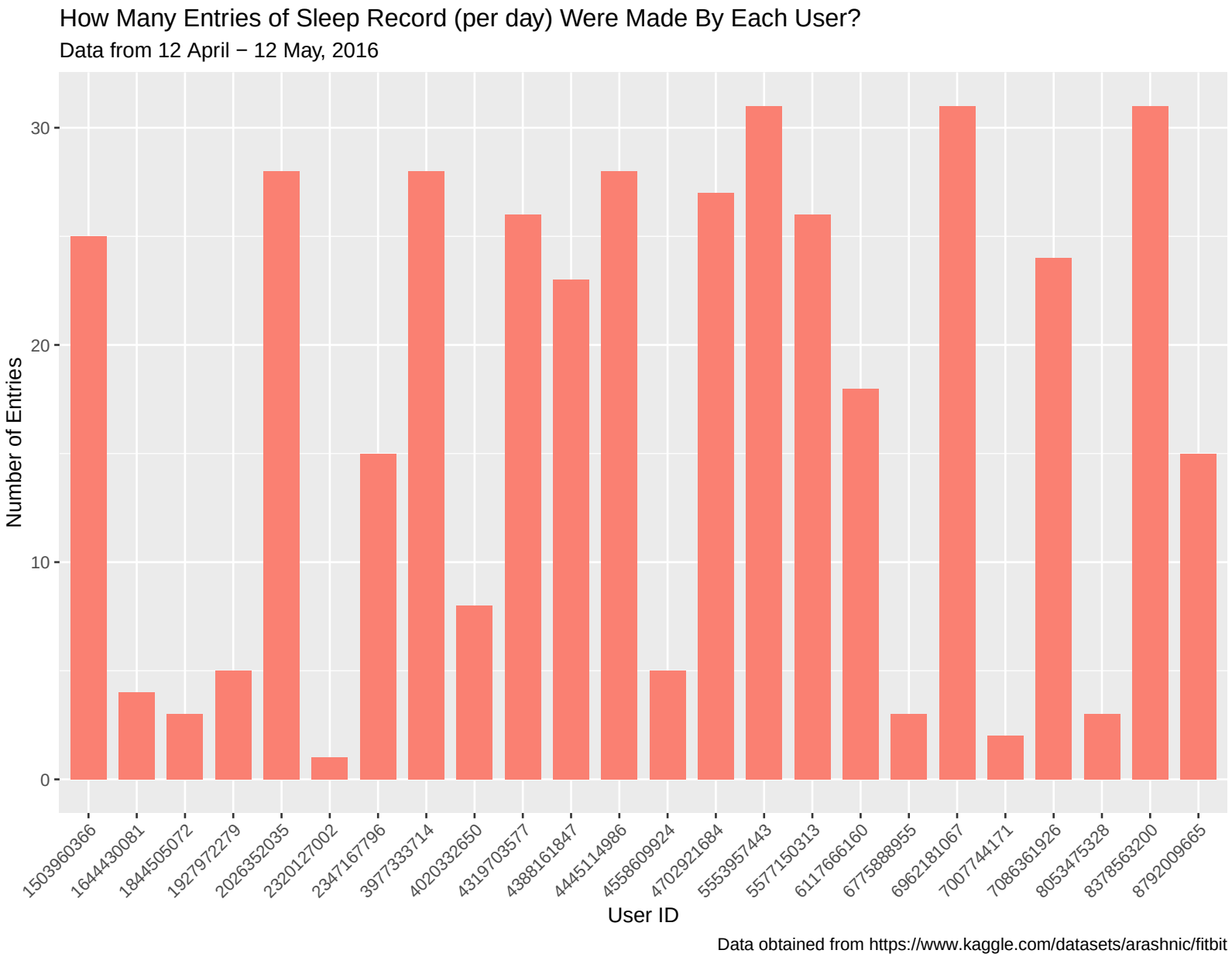

How Many Records of Hourly Steps Were Collected by Weekday?
Data from 12 March − 12 May, 2016
Number of Records
6000
4000
2000
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

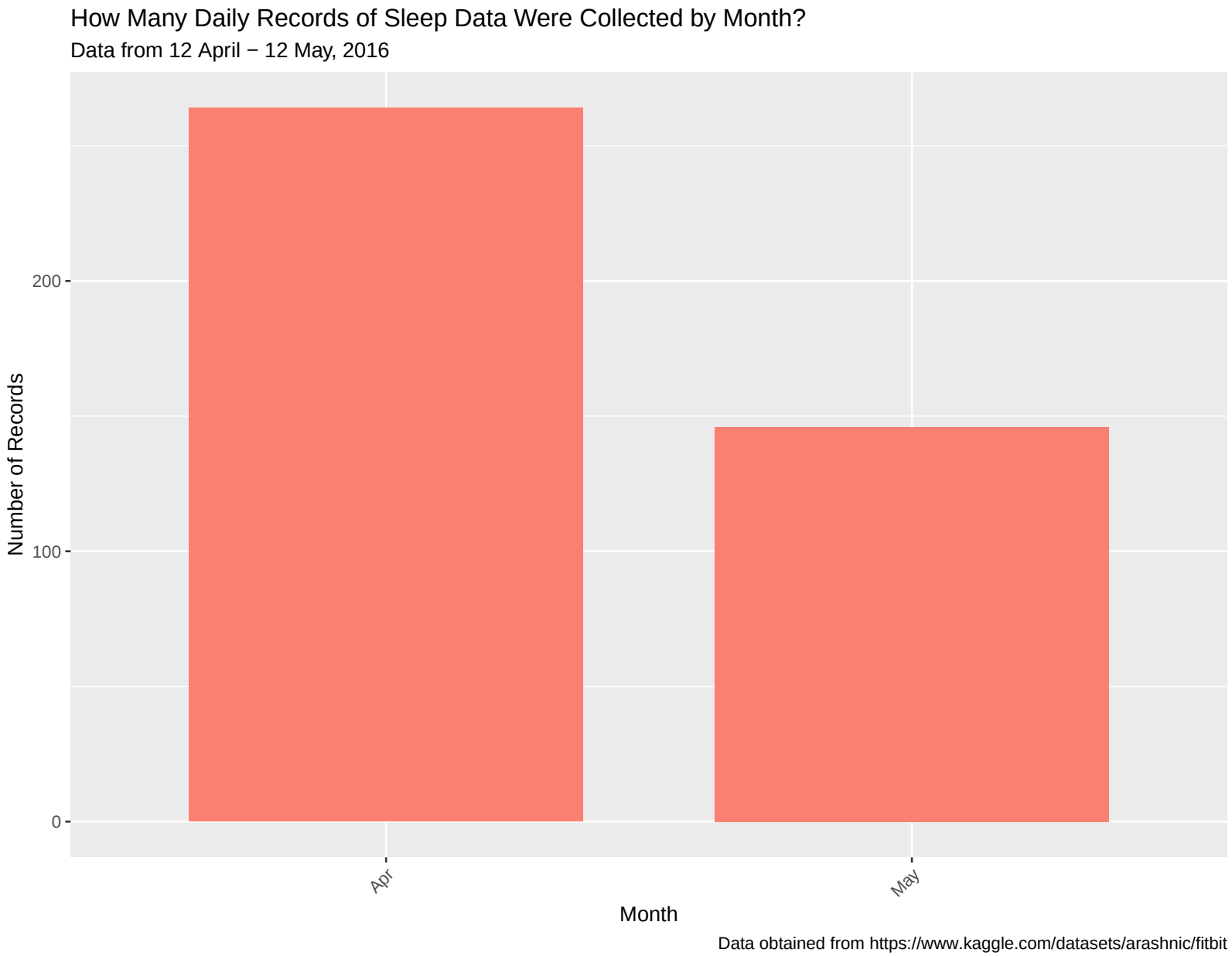

How Many Entries of Sleep Records (per minute) Were Made By Each User?
Data from 12 April − 12 May, 2016
Number of Entries
30000
20000
10000
0
1503960366
1644430081
1844505072
1927972279
2022484408
2026352035
2320127002
2347167796
3977333714
4020332650
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6775888955
6962181067
7007744171
7086361926
8053475328
8378563200
8792009665
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

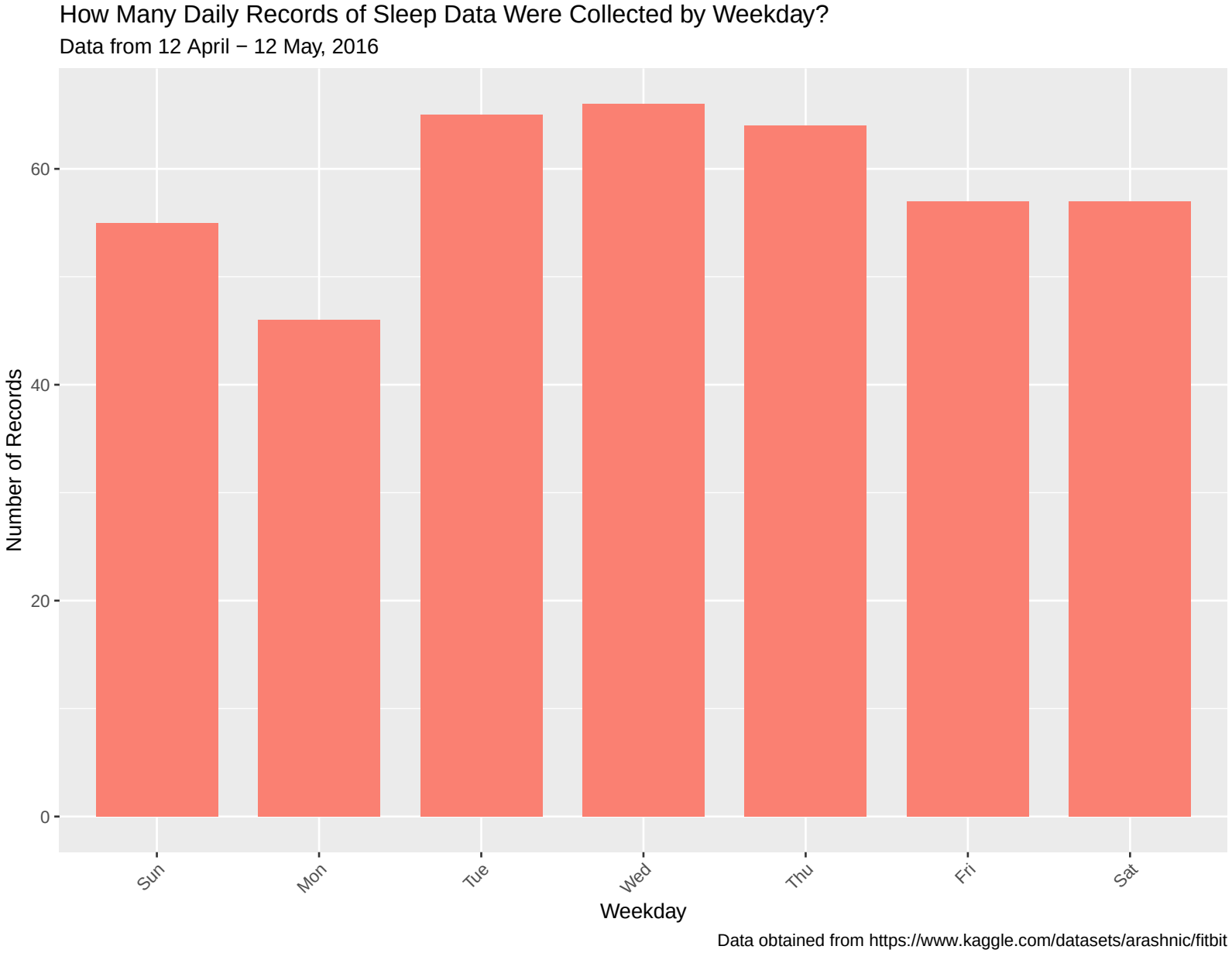

How Many Records of Sleep Data (By Minute) Were Collected by Month?
Data from 12 March − 12 May, 2016
200000
150000
100000
50000
0
Number of Records
Mar
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

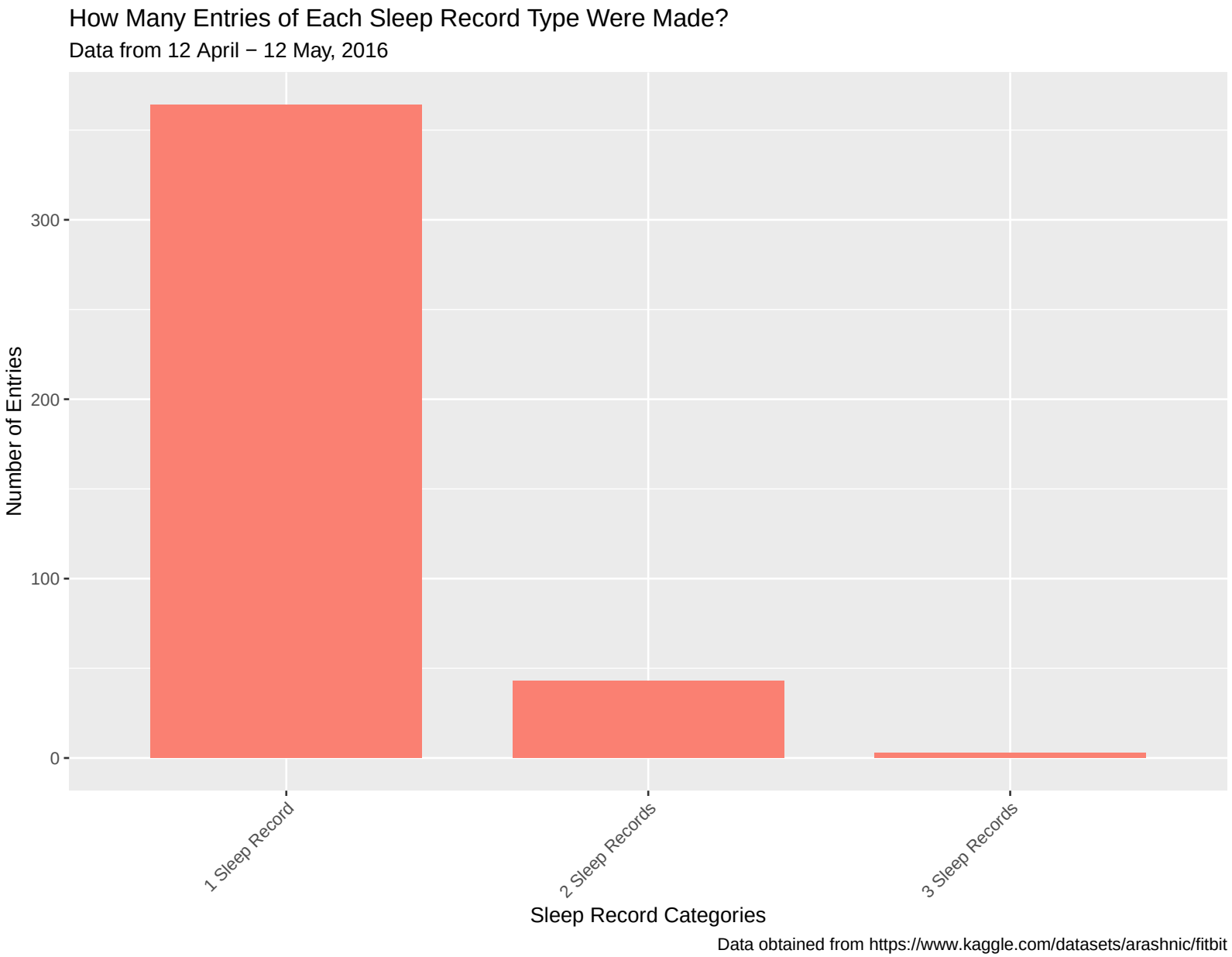

How Many Records of Sleep Data (By Minute) Were Collected by Weekday?
Data from 12 March − 12 May, 2016
60000
40000
20000
0
Number of Records
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit

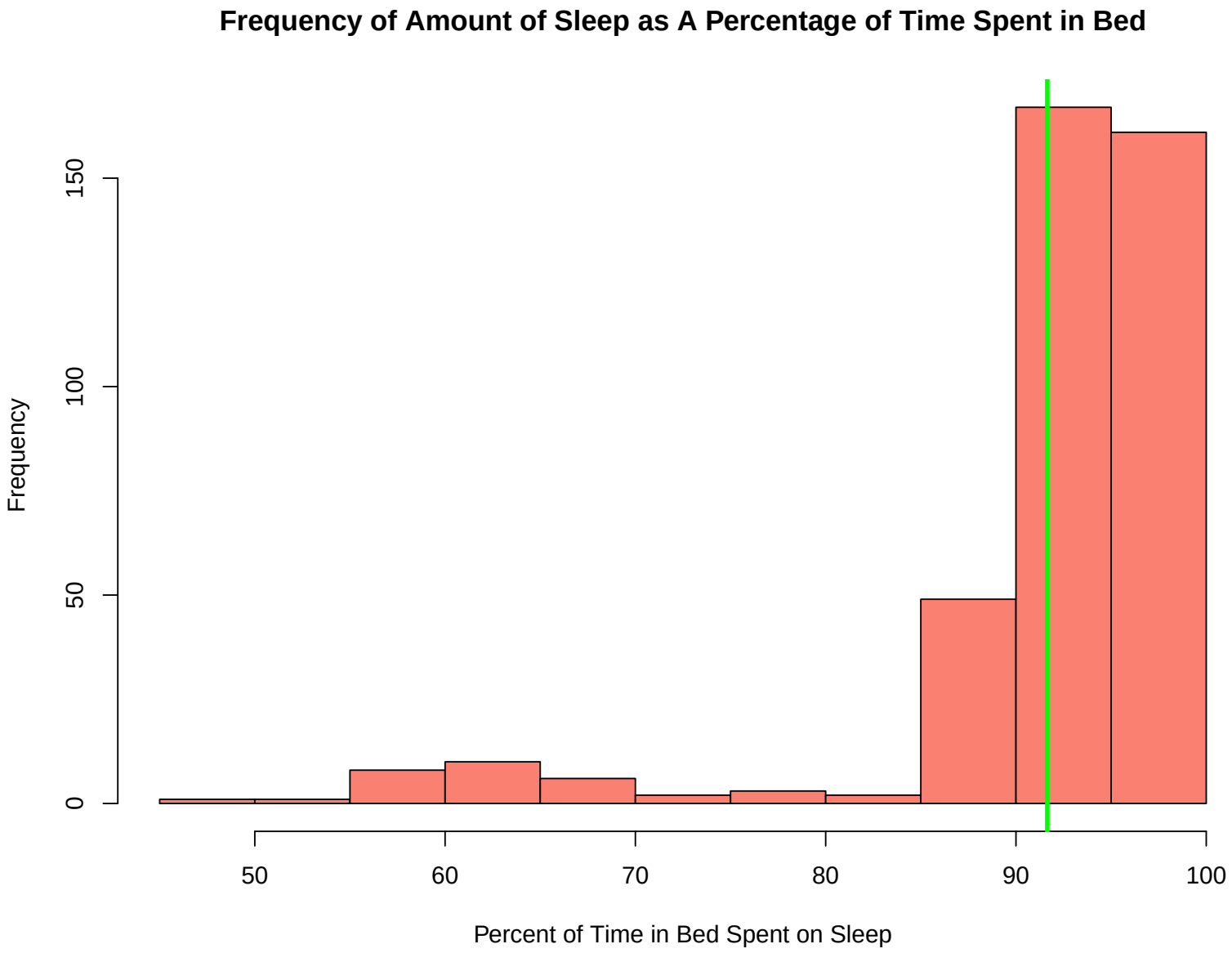

Frequency of Values for Total Minutes in Bed
Frequency
0
50
100
150
0
200
400
600
800
1000
Total Minutes in Bed


How Many Entries of Sleep Record (per day) Were Made By Each User?
Data from 12 April − 12 May, 2016
Number of Entries
30
20
10
0
1503960366
1644430081
1844505072
1927972279
2026352035
2320127002
2347167796
3977333714
4020332650
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6775888955
6962181067
7007744171
7086361926
8053475328
8378563200
8792009665
User ID
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit


How Many Daily Records of Sleep Data Were Collected by Month?
Data from 12 April − 12 May, 2016
Number of Records
200
100
0
Apr
May
Month
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit


How Many Daily Records of Sleep Data Were Collected by Weekday?
Data from 12 April − 12 May, 2016
Number of Records
60
40
20
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit


How Many Entries of Each Sleep Record Type Were Made?
Data from 12 April − 12 May, 2016
Number of Entries
300
200
100
0
1 Sleep Record
2 Sleep Records
3 Sleep Records
Sleep Record Categories
Data obtained from https://www.kaggle.com/datasets/arashnic/fitbit


Frequency of Amount of Sleep as A Percentage of Time Spent in Bed
Frequency
0
50
100
150
50
60
70
80
90
100
Percent of Time in Bed Spent on Sleep

Step 8: Gather insights from summary statistics and data visualizations

1. The average hourly heart rate is normal.
2. The highest hourly heart rate was recorded in the afternoon and early evening most likely on Thursday or Sunday.
3. The highest amount of hourly heart rate records were most likely on a Tuesday, Wednesday or Thursday.
4. There were relatively less participants who shared hourly heart rate data in comparison to other health data categories.
5. In comparison with each other, participants who shared hourly heart rate data were not consistent in the amount of entries recorded.
6. The most common values for hourly heart rate fell in the range of 60-80, which is normal.
7. Minute MET values were most frequently in the range 0 - 10.
8. The high average MET indicates strenuous activity.
9. The highest number of MET (by minute) records were in March and April.
10. The highest number of records were recorded on Tuesday's and Friday's.
11. In comparison with each other, participants who shared minute MET data were not consistent in the amount of entries recorded.
12. The highest number of daily activity records were made in April and May.
13. The highest number of daily activity records were made on Tuesday.
14. The high average number of steps, the highest number of steps, the average total distance and average daily calorie burn are all high, indicating that the participants live a active and healthy life.
15. The distance for logged activities were very low.
16. The most frequent values for logged activity distance were in the range 0 - 0.5.
17. The most frequent values for total steps were in the range 0 - 10,000.
18. The most frequent values for very active minutes were in the range 0 - 20.
19. The most frequent values for fairly active minutes were in the range 0 - 50.
20. The most frequent values for lightly active minutes were in the range 150 - 250.
21. The most frequent values for sedentary minutes were in the ranges of 70-80 and 1,010 - 1,020.
22. The most frequent values for daily calorie burn were in the range 1,500 - 2,500.
23. For the top 10 highest number of daily steps, participants had more time in very active and lightly active category.
24. For the lowest number of daily steps, participants spent the most amount of time doing lightly active tasks.
25. For the highest logged activity distances, participants spent the most amount of time doing lightly active and very active tasks.
26. The highest number of logged activity records was in April.
27. For high daily calorie burn, participants spent the most time doing lightly active and very active tasks.
28. In comparison with each other, participants who shared daily activity data were not consistent in the amount of entries recorded.
29. The average hourly calorie burn was a good value, indicating an active lifestyle.
30. The highest number of hourly calorie burn was in March and April.
31. The most frequent values for hourly calorie burn were in the range 50 - 100.
32. In comparison with each other, participants who shared hourly calorie burn data were not consistent in the amount of entries recorded.
33. The average hourly intensity was a good value, indicating an active lifestyle.
34. The highest number of hourly intensity records was in March and April.
35. The most frequent values for hourly intensity were in the range 0 - 10.
36. In comparison with each other, participants who shared hourly intensity data were not consistent in the amount of entries recorded.

37. The average hourly steps was a good value, indicating an active lifestyle.
38. The highest number of hourly steps records was in March and April.
39. In comparison with each other, participants who shared hourly steps data were not consistent in the amount of entries recorded.
40. The most frequent values for hourly steps were in the range 0 - 500.
41. The highest number of minute sleep entries was in March and April most likely on a Saturday or Sunday.
42. There were relatively less participants who shared minute sleep data in comparison to other health data categories.
43. In comparison with each other, participants who shared minute sleep data were not consistent in the amount of entries recorded.
44. There were relatively less participants who shared day sleep data in comparison to other health data categories.
45. In comparison with each other, participants who shared day sleep data were not consistent in the amount of entries recorded.
46. The highest number of daily sleep entries was in April and most likely on a Tuesday, Wednesday or Thursday.
47. A large majority of 88.78% of daily sleep entries had 1 sleep record, indicating good sleep habits.
48. The longest time in bed most likely occurred on Saturday or Sunday.
49. For the shortest time in bed, participants had only 1 sleep record.
50. The most frequent values for time in bed were in the range 400 - 500.
51. The average amount of time in bed sleeping as a percentage of the time in bed was 91.65%, indicating good sleep habits.
52. The most frequent values for the amount of time in bed sleeping as a percentage of the time in bed were in the range 90% - 100%.

Step 9: Share conclusions

1. There is a smaller number of users who contributed their hourly calorie burn data, minute sleep data and day sleep data when in comparison to the other health data categories.
2. Users recorded an inconsistent amount of data for each health data category.
3. In general, users who contributed sleep data have normal and healthy sleep habits.
4. In general, users who contributed data in all health data categories were active and healthy.
5. Not many users had logged activities. Only 4% of all records for daily activity had data for logged activities. It seems users passively track their daily activities without adding logged activities such as running or swimming.