Bellabeat Case Study

Brennan Grout

2022-06-02

Ask

Analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. Select one Bellabeat product to apply thse insights.

Questions

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How coud these trends help influence Bellabeat marketing strategy?

Deliverables

1. A clear summary of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulating of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top high-level content recommendations based on your analysis.

Guiding questions

- **What is the problem you are trying to solve?** Gain insight into how consumers use non-Bellabeat smart devices, then select one Bellabeat product to apply the insights to in a presentation.
- **How can your insights drive business decisions?** Insights can drive business decisions because we can find how people are actually interacting with their smart devices, so we can gain insight on how consumers interatct with their personal products. Allowing Bellabeat to take advantage of this insight and implement these insights into their own products.

Key tasks

1. **Identify the business task** Unlock new growth opportunities for the company by gaining insight into how consumers are using their smart devices. This will help guide a marketing startegy for the company.
2. **Consider stakeholders**

- Urska Srse: Cofounder and Chief Creative Officer - Primary stakeholder
- Sando Mur: Mathmetician and cofounder - Primary stakeholder
- Ballabeat marketing analytics team - Secondary stakeholders

Deliverable

- A clear statement of the business task.

Prepare

Guiding questions

- **Where is your data stored?** Locally in project filepath.
- **How is the data organized? Is it long or wide format?** Data is organized with multiple csv spread sheets. The data is in long format.
- **Are there issues with bias or credibility in this data?** The bias would be people who could afford fitbit trackers. The data is credible because it is cited and the data was collected during a 2 month period which is an acceptaple amount of time. In addition, the sample size was 30, which is the analytical recommended minimum data sample size where an average result of a sample starts to represent the average result of a population.
- **How are you addressing licensing, privacy, security, and accessibility?** Going to cite the license from the Kaggle data source. For secutiry measures I will ensure no personal information will be published with this notebook. Accessibility issues may be encountered so I will ensure any images used will have alt tags, chart plots will be descriptive and I will not use colors that can't be seen by people with color blindness.
- **How did you veify the data's integrity?** I verified the data's integrity by reviewing the data entity - making sure each table has a primary key value, filtered columns to check for blank values, and no duplicate sources of data.
- **How does it help to answer your question?** Ensures the data is credible and valid so we can have a high confidence level.
- **Are there any problems with the data?** I don't see any problems with the data, except there are alot of rows.

Key tasks

1. **Download data and store it appropriately.** Downloaded to my local machine and stored in a directory.
2. **Identify how it's organized** The data is organized by seperate csv files that can be queried using the primary keys of each table.
3. **Sort and filter the data.** I turned some of the date time stamps to just the date. Then sorted the data by date oldest to newest ascended in the spreadsheets. I also normalized all of the column names to lowercase and underscore _ spaces fillers.
4. **Determine the credibility of the data.** The data is credible because it is cited and the table's structure are normalized.

Deliverables

- **A description of all data sources used** I chose the following data sources because I wanted to analyze how people use their fitbit to track their daily steps to see if they sleep longer at night.

1. dailySteps_merged.csv
2. sleepDay_merged.csv

Process

Process data for analysis

Guiding questions

- **What tools are you choosing and why?** I am using excel spreadsheets, SQL, Rstudio, and Tableau. I am using excel to clean the data in the spreadsheets. Then I will use SQL to join the data I need from the tables. Then I will tidy the data in Rstudio and create a viz using ggplot2. Along with creating a r markdown notebook, this notebook actually… Lastly I will upload the data to Tableau and create viz's in Tableau public.
- **Have you ensured your data's integrity?** Yes.
- **What steps have you taken to ensure your data is clean?** I filtered all of the data values in a Excel and made sure there were no blank/null values.
- **How can you verify that your data is clean and ready to analyze?** I loaded my data sets into Rstudio and checked data for irregulatrities, unique values, and missing values. I converted the dates from strings to date data types so it will be easy to query in SQL and will be acrrate during analysis.
- **Have you documented your cleaning process so you can review and share those results?** Yes, I added the cleaning log to this notebook and added the script name for people to review how I cleaned this data set in Rstudio.

Key tasks

- **Check the data for errors.** I checked the data for errors and they are free of errors.
- **Choose your tools.** I am using Rstudio, Excel, and Tableau.
- **Transform the data so you can work with it effectively.** Selected the columns I will need.
- **Document the cleaning prcoess.** I added my cleaning log to this notebook.

```
library(tidyverse)
library(readr)
library(lubridate)
library(dplyr)

## Clean and verify data is cleaned using the tidyverse library and dplyr.
## Load the data frames and bind it to a variable data frame.
daily_steps_df <- read_csv("dailySteps_merged.csv")

## Familiarize yourself with the data set by viewing the column names and structure.
## Check for structural errors (column names are normalized, data types are correct, mislabeled variabl
head(daily_steps_df)
```

```
## # A tibble: 6 x 3
##           id date       step_total
##        <dbl> <date>          <dbl>
## 1 1503960366 2016-04-12      13162
## 2 1624580081 2016-04-12       8163
## 3 1644430081 2016-04-12      10694
## 4 1844505072 2016-04-12       6697
## 5 1927972279 2016-04-12        678
## 6 2022484408 2016-04-12      11875
```

```
colnames(daily_steps_df)
```

```
## [1] "id"         "date"       "step_total"
```

```r
str(daily_steps_df)
```

```
## spec_tbl_df [940 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id        : num [1:940] 1.50e+09 1.62e+09 1.64e+09 1.84e+09 1.93e+09 ...
##  $ date      : Date[1:940], format: "2016-04-12" "2016-04-12" ...
##  $ step_total: num [1:940] 13162 8163 10694 6697 678 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   date = col_date(format = ""),
##   ..   step_total = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
typeof(daily_steps_df$date)
```

```
## [1] "double"
```

```r
## Check unique values and for missing values.
unique(daily_steps_df$id)
```

```
##  [1] 1503960366 1624580081 1644430081 1844505072 1927972279 2022484408
##  [7] 2026352035 2320127002 2347167796 2873212765 3372868164 3977333714
## [13] 4020332650 4057192912 4319703577 4388161847 4445114986 4558609924
## [19] 4702921684 5553957443 5577150313 6117666160 6290855005 6775888955
## [25] 6962181067 7007744171 7086361926 8053475328 8253242879 8378563200
## [31] 8583815059 8792009665 8877689391
```

```r
unique(daily_steps_df$date)
```

```
##  [1] "2016-04-12" "2016-04-13" "2016-04-14" "2016-04-15" "2016-04-16"
##  [6] "2016-04-17" "2016-04-18" "2016-04-19" "2016-04-20" "2016-04-21"
## [11] "2016-04-22" "2016-04-23" "2016-04-24" "2016-04-25" "2016-04-26"
## [16] "2016-04-27" "2016-04-28" "2016-04-29" "2016-04-30" "2016-05-01"
## [21] "2016-05-02" "2016-05-03" "2016-05-04" "2016-05-05" "2016-05-06"
## [26] "2016-05-07" "2016-05-08" "2016-05-09" "2016-05-10" "2016-05-11"
## [31] "2016-05-12"
```

```r
unique(daily_steps_df$step_total)
```

```
##   [1] 13162  8163 10694  6697   678 11875  4414 10725 10113  8796  4747  8856
##  [13]  8539  5394  7753 10122  3276  5135  7213 11596  8135     0  4562 10199
##  [25] 14172 11317 18060  9033  7626  5014  2564 23186 10735  7007  8001  4929
##  [37]   356 12024  4993  7275 10352  7618  9715 10035  5974  8204 10993  2961
##  [49]  4978  6877  4832  5077  7142  4053  5652 12862  5813 16433  8053 12386
##  [61]  5571  1320 15337 10460  9107 11037  7937  2163 10690  3335  3973 10129
##  [73]  7910  8844  7641   108 10210  8863  3974  6799  7860 17022  8596  7671
##  [85]  5162  1551 11179  9123 20159  5234 13318  3135  1219 21129  9762  1510
##  [97]  5263  3844   980 11034  3821  5205 10465  8482  7451  9010  1882  3984
## [109]  5664  8758  7198  7795  6506 16556 12087 14019  9501  1282  5563  5273
## [121]  8585 20669  2672 14461  3430  2483 13422 12669  5370 15300  3414 10100
## [133]  2547  5057 22244  9685  6905 13459  1982  4744  6580  3945  7289 11140
## [145]  5771 14269 14450  8301  4732 13217  4631    31 14549  9256 11207  5319
## [157]   244 29326  9705  6175  8757  4525 15112   838  6198  5472  2524  8199
## [169] 10415    16    29  4660  2268  9634 12692   655 12231  7150  7851  2497
## [181] 10145  8059 18827 10204  2132  3008 15118 13019 10536  7132  4597 14131
## [193]  3325  6559  8247  7762  6798 11663    62  2276 11009  6155  8940  9105
## [205]  3727  9893  5153  6885  8294 11404 14816  9827 17076  5151 13630  3864
## [217] 11423 15506  2916 11256   197 11548  2424  5997  6711  7948  7711 12414
## [229]  8925 10181  2064  5401  6708 15482 12574 11135 10742 14194 10688 15929
## [241]  4212 13070  5697 18785 10544  4974  2436     8  7222  7192 10999  9202
## [253]  4880 11658  8954 10553  2072  4803  8793  2713  8330 10449  6361 10771
## [265] 13928 15566 14365 15108  6466  9388  3147 19948  9819  6349  1223  8054
## [277] 12453  2467  3404 10080  8859  8857  6093  3702 10055  3809 13743  6530
## [289] 12346 10830 19542 11835 13744  9469 16057 11268 15148  8538   144 19377
## [301] 12764  4026  3673  5372   149 12954  2915  5583  7804  7286  3843  8911
## [313]  4500 12139  6831  9601  1664 11682  9172  8206  6238   637 15299  9753
## [325] 10520  2824 12200  8687  4068 18258 14371  6637  3570  2945  6001 12357
## [337]  5079 16901  9317  7396 12058  4935 13236  4363  6890 15126  4112  7638
## [349] 11495 20031  8093  2817 22359  9282  5709  9423  5245 11200 10039  6076
## [361]  3321  2090 13481  3490  4165  9471  6873  6731 14112  4081 10243  5002
## [373]  8563 15050  1807 15764  7623  5896  2153  5029 11085  3520 22988  8905
## [385]  3703  8286   400 16674 15355  6497  3580   152 11369  6017  3588  9482
## [397]  7373  5995 11177  9259 12961  3385  8095  9167 10946  6393  7802  6474
## [409] 13239 18229 10091 20500  6829 12405  4503 12986 13755  2826  9919  3761
## [421] 10119  5933  3409  5980  8242  8283 11388  9899  9461  6326  9148  6108
## [433] 11886  5325  9543  7091 10433 15090 10387 12685 16208 10499  1321 11101
## [445] 18134  8367  3032     4 10159  6088  1715  3516  7904  7193 10780 11193
## [457]  7243  9557  7047 10538  6805  9411  5565 10320 13541 11107 12422 10232
## [469]  7359 12474  1758 23629 13154  2759  9405  6907  1675 10140  6375  1532
## [481]  5439  7913  5512  7114 10817 10074  4493  9451  9023 11393  9841  3403
## [493]  5731   703 12627 15128 11584 15447  2718  5417  6174  6157 14890 11181
## [505]  2390  3176  4920 10245  7604   924    42  7365  9135 10645  7990  9232
## [517]  4676  7833  9930  7924  9592 10762 20067  7881 12315  6260 15168  8360
## [529]  9733 14673 18213  4014 18387  4729  4571  8452  5250 13238  8221 12533
## [541]  6222 10319 10144  1202 12363  6987  6744  2503 10081 14560  7135  2946
## [553] 10085  7174 27745 10602 36019  6132  2573  2704  3609   772  7399  3077
## [565] 10414  1251 10255  5232  3428  5164 13368  8915  9837  2487  5454  5600
## [577] 12390  1170 11419  4512  1619 10930 14727  7155  3758  3790 10379  7018
## [589]  3634  7525 16520   475  9261 10096  6910  7891  7245  9769  7439  4933
## [601]  6781 12912 13041 10052  1969  6064  8469  1831  4790 15103  2100 12850
## [613]  4059  1326 12183  5992  7443  7412 14335  4496  9648 12727  7502  5267
## [625]  9454 12848 11045  6047     9 12109 14510 10288 15484  8712 12015  2421
## [637] 10818 11100  2193  2309  2080  1786 11768  6564  1201  8278 13559 10252
## [649] 10429 12375  2923  8161  4249  5206  2997  5832 10147 10988 14581  7875
## [661]  2283 18193 14070  2470  2237 11895 12167  5202  8314 12312 11728 13658
## [673]  9603  3800 10611  8614 14331  7550  9799  6339 10524 15010  8564 14990
## [685]  8567 12427 14055 12159  1727  9787    44  2091 10227  8198  4878  7063
## [697] 11677  4369  9524 13175  4514  3755  6943  9632  4950  3365  6116  4697
## [709]  5908 11459 12461 13953  7045  5843 21727 11992  2104 13372  4193  7379
## [721]  4940 11550 22770  5183  8237 14370  1868  7336  5510  1967  6815 12827
## [733] 19769  4468  6117 12332 10060  3427  6724  3292  5528  5161  8168 13585
## [745]  5862  3672 17298  7303  6543 12857  6083  7328  7706  4188 10677 22026
## [757]  2943  9217 10686 12022  1732  6643 13379 10685  3090  7726 14687  4556
## [769] 10378 10218  5275 11451  8232 11611  3421  4477  6277 12342 13566 12465
## [781]  8382  9877 20226 12207  2969 12798   254  6227  8275 13072  5546  9487
## [793] 10299  3915  6435 10613 16358  8869 15448 14433 14810  6582  8240 10733
## [805] 12770  3134  1329 13272  8580  6424  6440   746  3689  9129 10201  9108
## [817]  9810  4926  4038  6722  9572 12209  9143  8701 21420  2971  9117  8891
## [829]  2661  7566   590    17  3369   768  6307  2752  3121  3587  3789  4998
## [841]  4561  8064
```

```
## Load the sleepDay csv data frame then bind it to a new data frame variable.
sleep_day_df <- read_csv("sleepDay_merged.csv")

head(sleep_day_df)
```

```
## # A tibble: 6 x 5
##           id date      total_sleep_records total_minutes_asleep total_time_in_b~
##        <dbl> <chr>                   <dbl>                <dbl>            <dbl>
## 1 1503960366 4/12/2016                   1                  327              346
## 2 1927972279 4/12/2016                   3                  750              775
## 3 2026352035 4/12/2016                   1                  503              546
## 4 3977333714 4/12/2016                   1                  274              469
## 5 4020332650 4/12/2016                   1                  501              541
## 6 4445114986 4/12/2016                   2                  429              457
```

```
str(sleep_day_df)
```

```
## spec_tbl_df [413 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id                  : num [1:413] 1.50e+09 1.93e+09 2.03e+09 3.98e+09 4.02e+09 ...
##  $ date                : chr [1:413] "4/12/2016" "4/12/2016" "4/12/2016" "4/12/2016" ...
##  $ total_sleep_records : num [1:413] 1 3 1 1 1 2 1 1 1 1 ...
##  $ total_minutes_asleep: num [1:413] 327 750 503 274 501 429 425 441 419 366 ...
##  $ total_time_in_bed   : num [1:413] 346 775 546 469 541 457 439 464 438 387 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   date = col_character(),
##   ..   total_sleep_records = col_double(),
##   ..   total_minutes_asleep = col_double(),
##   ..   total_time_in_bed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
typeof(sleep_day_df$date)
```

```
## [1] "character"
```

```
## Check unique values and for missing values.
unique(sleep_day_df$id)
```

```
##  [1] 1503960366 1927972279 2026352035 3977333714 4020332650 4445114986
##  [7] 4702921684 5553957443 5577150313 6962181067 7086361926 8378563200
## [13] 8792009665 2347167796 6775888955 4319703577 1844505072 4388161847
## [19] 6117666160 7007744171 8053475328 4558609924 2320127002 1644430081
```

```r
unique(sleep_day_df$date)
```

```
##  [1] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" "4/16/2016" "4/17/2016"
##  [7] "4/18/2016" "4/19/2016" "4/20/2016" "4/21/2016" "4/22/2016" "4/23/2016"
## [13] "4/24/2016" "4/25/2016" "4/26/2016" "4/27/2016" "4/28/2016" "4/29/2016"
## [19] "4/30/2016" "5/1/2016"  "5/2/2016"  "5/3/2016"  "5/4/2016"  "5/5/2016" 
## [25] "5/6/2016"  "5/7/2016"  "5/8/2016"  "5/9/2016"  "5/10/2016" "5/11/2016"
## [31] "5/12/2016"
```

```r
unique(sleep_day_df$total_sleep_records)
```

```
## [1] 1 3 2
```

```r
unique(sleep_day_df$total_minutes_asleep)
```

```
##   [1] 327 750 503 274 501 429 425 441 419 366 514 338 458 384 398 531 467 295
##  [19] 370 400 455 432 235 630 451 447 545 445 291 535 357 477 423 508 472 424
##  [37] 486 412 644 475 523 452 465 499 337 253 377 392 391 513 363 340 524 283
##  [55]  77 506 426 462 382 651 406 380  79 611 700 437 556 381 619  98 591 350
##  [73] 549 336 427 525 500 515  99 293 520 527 493 442 304 498 219 461 329 388
##  [91] 457 449 476 387 360 152 421 439 454 658 474 418 492 528 325 460 332  59
## [109] 436 126 399 414 390 396 405 355 533  82 322 480 428 361 522  61 374 692
## [127] 478 631 331 565 339 430 555 310 552 553 543 353 511 681 277 433 262 488
## [145] 328 446 245 296 250 505 319 103 354 485 550 448 349 286 347 542 469 166
## [163] 408 261 497 393 440 402 341 119 490 411 333 106 171 450 459 600 404 124
## [181] 722 573 237 409 479 775 422 468 343 369 590 383 484 547 115 622 379 507
## [199]  58 796 230 368 502 466 351 415 273 292 417 394 456 516 538 213 603 420
## [217] 247 318 226 471  74 334 323 385 443 401 298 530 594 137 541 259 364 602
## [235] 529 123 568 481  62 435 453 489 359 312 487 416 504 342 285 463 431 302
## [253] 483 438 444 496
```

```r
unique(sleep_day_df$total_time_in_bed)
```

```
##   [1] 346 775 546 469 541 457 439 464 438 387 525 356 493 407 422 565 531 456
##  [19] 406 430 488 458 260 679 465 487 552 568 489 397 557 492 415 418 497 441
##  [37] 535 476 455 503 442 961 499 573 504 556 491 526 379 257 409 413 386 533
##  [55] 377 367 567 510  77 522 448 686 445 398 366  82 689 712 498 602 566 641
##  [73] 107 612 402 583 350 446 591 551 104 312 553 451 320 540 514 395 338 424
##  [91] 410 483 421 305 543 462 468 678 480 502 547 364 484 512  65 137 434 431
## [109] 437 550 411 417 545 461  85 353 435 384 554  69 372 722 501 475 725 539
## [127] 337 360 449 459 595 506 640 615 704 323 471 467 345 380 447 274 315 490
## [145] 371 516 391 121 429 453 500 584 307 374 452 479 600 428 393 178 450 423
## [163] 416 354 127 473 478 108 179 495 485 636 562 425 142 607 382 843 433 555
## [181] 396 626 597 129 575 426  61 527 309 376 542 482 385 296 332 373 560 336
## [199] 536 414 363 470 634 436 477 463 264 248 513  78 548 408 698 333 355 349
## [217] 521 443 530 496 507 334  75 611 154 638 134 608 603 569 606 342 494 399
## [235] 403 517 375 481 306 558 486 321
## Check for data irregularities (invalid values, outliers)
summarize_steps <- daily_steps_df %>%
  summarize(
    min_steps = min(step_total),
    max_steps = max(step_total),
    avg_steps = mean(step_total)
  )
summarize_steps
```

```
## # A tibble: 1 x 3
##   min_steps max_steps avg_steps
##       <dbl>     <dbl>     <dbl>
## 1         0     36019     7638.
```

```
summarize_sleep <- sleep_day_df %>%
  summarize(
    min_sleep = min(total_minutes_asleep),
    max_sleep = max(total_minutes_asleep),
    avg_sleep = mean(total_minutes_asleep)
  )
summarize_sleep
```

```
## # A tibble: 1 x 3
##   min_sleep max_sleep avg_sleep
##       <dbl>     <dbl>     <dbl>
## 1        58       796      419.
```

Analyze

Guidng questions

- **How should you organize your data to perform analysis on it?** I sorted the date fields to be the same organized by descending oldest to newest so its easier to validate and join. Then I joined the steps and sleep data frames.
- **Has your data been properly formatted?** I reformatted the date fields from character strings to date type.
- **What surprises did you discover in the data?** Surprises I found was that not everyone who tracked their steps, also tracked their sleep. Which didnt return as many as many results as I thought it would after joined the two tables.
- **What trends or relationships did you find in the data?** Some trends I saw were the more steps people took, the less they slept. This was surprising because my initial hypothesis was the more steps people took during the dat, the longer people slept. However, after analyzing this data, I realized my sleeping pattern when I walk more during the day. I tend to sleep less than when I don't walk as muuch. More steps correlated less sleeping time periods is a correlation, but may not be the causation. This would need to be researched further.
- **How will these insights help answer your business questions?** These insights will help answer my business question by answering the fact that people are in fact using their smart devices to track their steps during the day, then tracking their sleep during the nights they track their steps. The data has shown that the more steps you take during the day, the less sleep you need at night. This is great information to know for people who are on the fence buying a Bellabeat app since it tracks your sleep habits.

Key tasks

1. **Aggregate your data so it's useful and accessible.** I aggregated my data by joining them with MySQL then exported the data via csv so the data is accessible in Rstudio.
2. **Organize and format your data.** I organized and formatted my data by deleting the redundant date columns created from the join, I also deleted the redundant id column that was created after the join. The date column was still formatted correctly after the join.
3. **Perform calculations.** I performed a min and max calculation on the dates to dynamically show the min and max date periods of the data to be displayed on chart plot viz caption.
4. **Identify trends and relationships.** The trends shown from the analysis are the more steps people took during the day, the less amount of sleep they needed at night. I used a scatterplot for this analysis to display each data point.

Delivery

- **A summary of your analysis.** The more steps people take during the day, the less sleep people need at night. This was shown accross all participants of the survey.

Share

Guidng questions

- **Were you able to answer the business question?** Yes, I was able to answer the business question.
- **What story does your data tell?** There is a correlation between the more people walk during the day, the less sleep they need overall. This gives participants a better quality of life during the day since it allows them more time during the day to complete their tasks and potentially gives them more time for personal lives as well.
- **How do your findings relate to your original question?** My findings relate to how consumers use their smart devices in their daily lives, by showing people use their devices to track their daily steps and they use them to track their nightly sleep periods.
- **Who is your audience? What is the best way to communicate with them?** Our audience is people in the workforce looking to be more efficient in their daily lives while living a healthier lifestyle. The best way to communicate with them is online via social media channels and search ads.
- **Can data visualization help you share your findings?** Yes, data viz's can help me share my findings by showing a scatter plot of all consumer data points and showing a smooth line trend to identify total sleeping minutes and total steps.
- **Is your presentation accessible to your audience?** Yes, we could use the viz's to show remarketing audiences on how tracking your steps and sleep informs people how much sleep they get after tracking their daily steps.

Key tasks

1. **Dtermine the best way to share your findings.** The best way to share my findings is through power point.
2. **Create effective data visualizations.** I created a ggplot scatter plot with a smooth line to display data point trents.
3. **Present your findings.** I will present my findings to someone who is not involved with the project and who doesn't know the data.
4. **Ensure your work is accessible.**

Deliverable

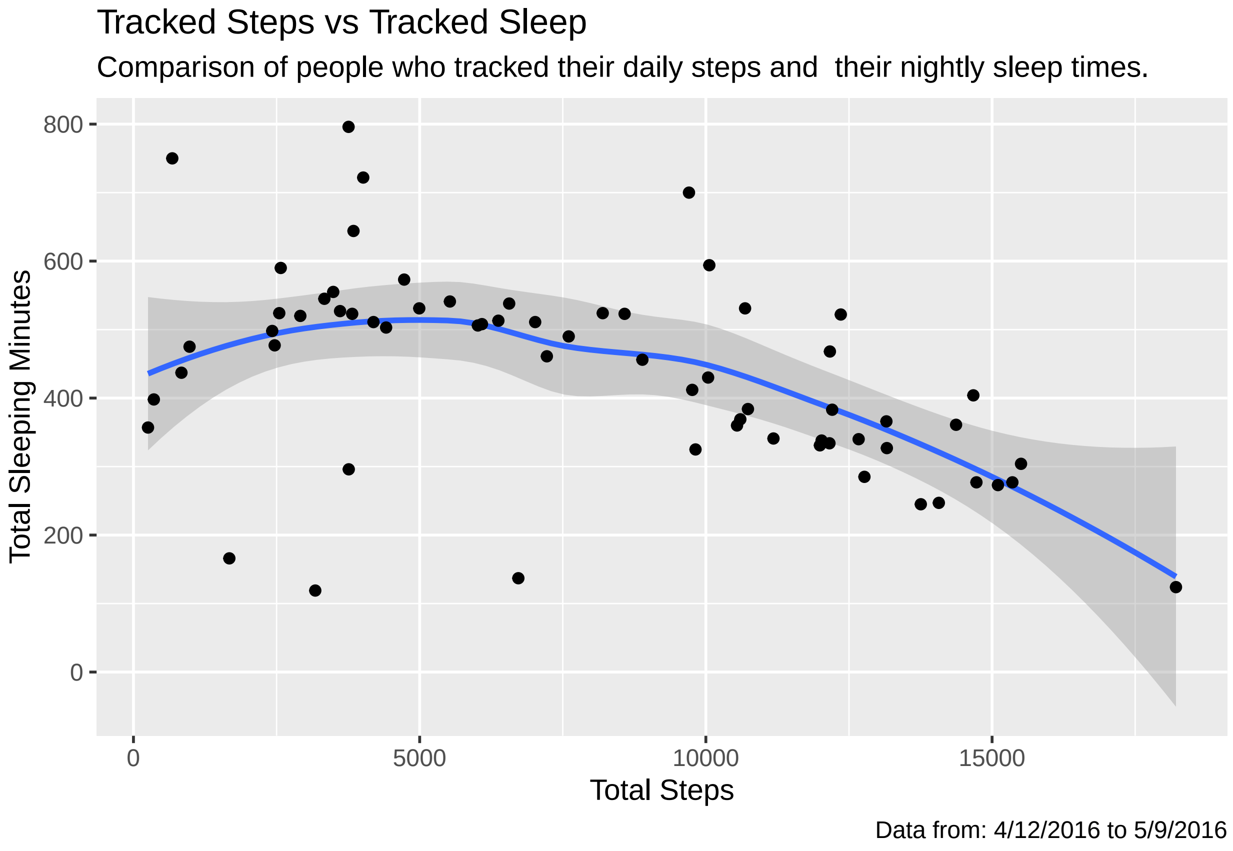
- **Supporting visualizations and key findings.**

```
## Analyze data
library(tidyverse)
library(readr)
library(dplyr)
library(ggplot2)

## Import data frames used for analysis.
steps_sleep_df <- read_csv("joined_steps_days_20220603_v03.csv")

## Use min and max to find data date ranges.
mindate <- min(steps_sleep_df$date)
maxdate <- max(steps_sleep_df$date)

## Use ggplot scatter plot to chart cleaned data.
ggplot(data = steps_sleep_df) +
  geom_smooth(mapping = aes(x=step_total, y=total_minutes_asleep)) +
  geom_point(mapping = aes(x=step_total, y=total_minutes_asleep)) +
  labs(title = "Tracked Steps vs Tracked Sleep",
       subtitle = "Comparison of people who tracked their daily steps and  their nightly sleep times.",
       caption = paste0("Data from: ", mindate, " to ", maxdate),
       x="Total Steps",
       y="Total Sleeping Minutes")
```



Act

Guiding questions

- **What is your final conclusion based on your analytics?** The final conclusion is that people are wearing smart devices to track their daily steps and to track their nightly sleep periods. The more steps people take during the day, th eless amount of time people need to sleep for. This is helpful for people who are looking to make their daily lives more efficient.
- **How could your team and business apply your insights?** My team and business could apply these insights by marketing the Bellabeat app to people who are interested in having a more healthy and efficient lifestyle. We could do this by explaining that the more steps you take, the less sleep you actually need at night. This gives you more awake hours during the day to get things done or gives you more time to yourself.
- **What next steps would you or your stakeholders take based on your insights?** The next steps would be to research the correlation between daily steps and sleeping minutes to find if there is a causation between the two. If there is, then we could market the app to people looking to have a more efficient lifestyle with more steps during the day and less needed sleep at night.
- **Is there additional data you could use to expand your findings?** Yes, I think comparing heart rate and calories burned to steps and minutes sleeping to find if there is even more correlation between the data sets. Along with finding consumer's ages to see if they just sleep less with or without more steps. There may be some underlying factors that are causing less sleep that aren't correlated to more steps that we need to investigate further.

Key tasks

1. **Create your portfolio.**
2. **Add your case study.**
3. **Practice prsenting your case study to a friend or family member.**

Deliverable

- **Your top high-level insights based on your analysis.**

Cleaning Log

`cleaning_data_capstone_project_20220603_v01.R`

2022-06-02

1. Changed column names to all lower case and added under score where there were spaces.
2. Split time from dates and changed column name to date.
3. Filtered data to check for null values, found none.

2022-06-03

1. Imported data into Rstudio and viewed data.
2. Converted date data types from string characters to date in Excel.
3. Checked for data irregularities, unique values, and missing values.
4. Joined the sleep and steps tables on their id and dates.
5. Removed the id and date duplicates.