# Case Study 2: Google Data Analytics

Ricardo Valins

2023-05-20

## Roteiro do estudo de caso - Preparar

### ● Onde seus dados são armazenados?

O dataset foi baixado da comunidade Kaggle no seguinte endereço: https://www.kaggle.com/datasets/arashnic/fitbit (https://www.kaggle.com/datasets/arashnic/fitbit)

O ficheiro foi gravado e extraido localmente.

O ficheiro comprimido continha vários ficheiros em csv, a saber:

- dailyActivity_merged.csv
- dailyCalories_merged.csv
- dailyIntensities_merged.csv
- dailySteps_merged.csv
- heartrate_seconds_merged.csv
- hourlyCalories_merged.csv
- hourlyIntensities_merged.csv
- hourlySteps_merged.csv
- minuteCaloriesNarrow_merged.csv
- minuteCaloriesWide_merged.csv
- minuteIntensitiesNarrow_merged.csv
- minuteIntensitiesWide_merged.csv
- minuteMETsNarrow_merged.csv
- minuteSleep_merged.csv
- minuteStepsNarrow_merged.csv
- minuteStepsWide_merged.csv
- sleepDay_merged.csv
- weightLogInfo_merged.csv

### ● Como os dados são organizados? No formato longo ou largo?

Como os próprios nomes dos ficheiros sugere, há uma granularidade temporal nos mesmos com base no tempo de apresentação das informações, por exemplo, em dias, horas e minutos.

Para o propósito de nossa análise, é suficiente a abertura destes dados por dia, razão pela qual não iremos explorar as outras granularidades salvo se julgarmos relevantes um maior detalhamento.

Dentre os ficheiros com indicação de dados diários, exploramos a possibilidade de se criar um modelo de dados. Contudo, percebemos que para este problema de negócio tal situação não seria necessária, bastando apenas utilizar o ficheiro `dailyActivity_merged` .

Conforme sugerido pela formação, iremos utilizar o MS Excel para fazer uma análise preliminar dos dados antes de trabalharmos os mesmos no `R Studio`.

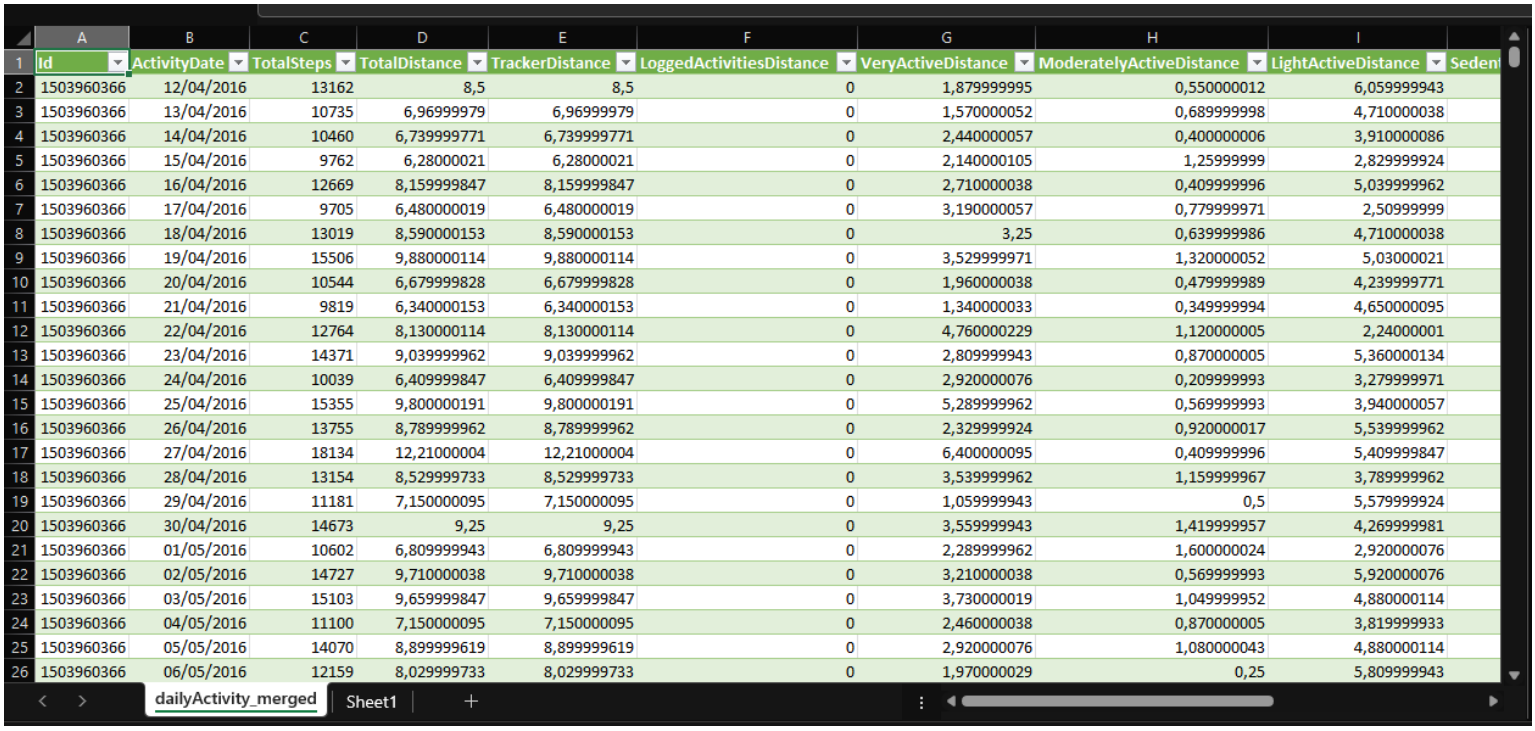
Preparamos o ficheiro em Excel para importar os dados com o locale `"English (United States)"` em Regional Settings.

Utilizamos o código a seguir para importar os dados:

## dailyActivity_merged

```
let
    Source = Csv.Document(File.Contents("C:\_google data analytics\08 Projeto final de Data Analytics do Google\Fitabase Data 4.12.16-5.12.16\dailyActivity_merged.csv"),[Delimiter=",", Columns=15, Encoding=1252, QuoteStyle=QuoteStyle.None]),
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"Id", Int64.Type}, {"ActivityDate", type date}, {"TotalSteps", Int64.Type}, {"TotalDistance", type number}, {"TrackerDistance", type number}, {"LoggedActivitiesDistance", type number}, {"VeryActiveDistance", type number}, {"ModeratelyActiveDistance", type number}, {"LightActiveDistance", type number}, {"SedentaryActiveDistance", type number}, {"VeryActiveMinutes", Int64.Type}, {"FairlyActiveMinutes", Int64.Type}, {"LightlyActiveMinutes", Int64.Type}, {"SedentaryMinutes", Int64.Type}, {"Calories", Int64.Type}}),
    #"Filtered Rows" = Table.SelectRows(#"Changed Type", each true)
in
    #"Filtered Rows"
```

Com isto, teremos os seguintes dados carregados:



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Id | ActivityDate | TotalSteps | TotalDistance | TrackerDistance | LoggedActivitiesDistance | VeryActiveDistance | ModeratelyActiveDistance | LightActiveDistance | Sedent |
| 2 | 1503960366 | 12/04/2016 | 13162 | 8,5 | 8,5 | 0 | 1,879999995 | 0,550000012 | 6,059999943 | |
| 3 | 1503960366 | 13/04/2016 | 10735 | 6,96999979 | 6,96999979 | 0 | 1,570000052 | 0,689999998 | 4,710000038 | |
| 4 | 1503960366 | 14/04/2016 | 10460 | 6,739999771 | 6,739999771 | 0 | 2,440000057 | 0,400000006 | 3,910000086 | |
| 5 | 1503960366 | 15/04/2016 | 9762 | 6,28000021 | 6,28000021 | 0 | 2,140000105 | 1,25999999 | 2,829999924 | |
| 6 | 1503960366 | 16/04/2016 | 12669 | 8,159999847 | 8,159999847 | 0 | 2,710000038 | 0,409999996 | 5,039999962 | |
| 7 | 1503960366 | 17/04/2016 | 9705 | 6,480000019 | 6,480000019 | 0 | 3,190000057 | 0,779999971 | 2,50999999 | |
| 8 | 1503960366 | 18/04/2016 | 13019 | 8,590000153 | 8,590000153 | 0 | 3,25 | 0,639999986 | 4,710000038 | |
| 9 | 1503960366 | 19/04/2016 | 15506 | 9,880000114 | 9,880000114 | 0 | 3,529999971 | 1,320000052 | 5,03000021 | |
| 10 | 1503960366 | 20/04/2016 | 10544 | 6,679999828 | 6,679999828 | 0 | 1,960000038 | 0,479999989 | 4,239999771 | |
| 11 | 1503960366 | 21/04/2016 | 9819 | 6,340000153 | 6,340000153 | 0 | 1,340000033 | 0,349999994 | 4,650000095 | |
| 12 | 1503960366 | 22/04/2016 | 12764 | 8,130000114 | 8,130000114 | 0 | 4,760000229 | 1,120000005 | 2,24000001 | |
| 13 | 1503960366 | 23/04/2016 | 14371 | 9,039999962 | 9,039999962 | 0 | 2,809999943 | 0,870000005 | 5,360000134 | |
| 14 | 1503960366 | 24/04/2016 | 10039 | 6,409999847 | 6,409999847 | 0 | 2,920000076 | 0,209999993 | 3,279999971 | |
| 15 | 1503960366 | 25/04/2016 | 15355 | 9,800000191 | 9,800000191 | 0 | 5,289999962 | 0,569999993 | 3,940000057 | |
| 16 | 1503960366 | 26/04/2016 | 13755 | 8,789999962 | 8,789999962 | 0 | 2,329999924 | 0,920000017 | 5,539999962 | |
| 17 | 1503960366 | 27/04/2016 | 18134 | 12,21000004 | 12,21000004 | 0 | 6,400000095 | 0,409999996 | 5,409999847 | |
| 18 | 1503960366 | 28/04/2016 | 13154 | 8,529999733 | 8,529999733 | 0 | 3,539999962 | 1,159999967 | 3,789999962 | |
| 19 | 1503960366 | 29/04/2016 | 11181 | 7,150000095 | 7,150000095 | 0 | 1,059999943 | 0,5 | 5,579999924 | |
| 20 | 1503960366 | 30/04/2016 | 14673 | 9,25 | 9,25 | 0 | 3,559999943 | 1,419999957 | 4,269999981 | |
| 21 | 1503960366 | 01/05/2016 | 10602 | 6,809999943 | 6,809999943 | 0 | 2,289999962 | 1,600000024 | 2,920000076 | |
| 22 | 1503960366 | 02/05/2016 | 14727 | 9,710000038 | 9,710000038 | 0 | 3,210000038 | 0,569999993 | 5,920000076 | |
| 23 | 1503960366 | 03/05/2016 | 15103 | 9,659999847 | 9,659999847 | 0 | 3,730000019 | 1,049999952 | 4,880000114 | |
| 24 | 1503960366 | 04/05/2016 | 11100 | 7,150000095 | 7,150000095 | 0 | 2,460000038 | 0,870000005 | 3,819999933 | |
| 25 | 1503960366 | 05/05/2016 | 14070 | 8,899999619 | 8,899999619 | 0 | 2,920000076 | 1,080000043 | 4,880000114 | |
| 26 | 1503960366 | 06/05/2016 | 12159 | 8,029999733 | 8,029999733 | 0 | 1,970000029 | 0,25 | 5,809999943 | |

dailyActivity_merged | Sheet1

Como podemos observar, os dados estão em `formato largo (wide-format)`

## ● Existem problemas com viés ou credibilidade nesses dados? Seus dados são confiáveis, originais, abrangentes, atuais e incluem a fonte?

Cada linha do ficheiro representa uma captura de informação do dia de um utilizador (campo `id`).

Verificamos através de uma tabela dinâmica que temos 33 utilizadores distintos.

Nota: a descrição dos dados do Kaggle indicam existir 30 utilizadores e como vimos temos dados de 33.

Ao fazer uma tabela dinâmica, percebemos também, que os dados foram obtidos no período de `12/04/2016` até `12/05/2016`. Não há dias em que não houve captura.

Notamos que nem todos os utilizadores utilizaram o aplicativo todos os dias (isso implicaria termos 31 dias x 33 ids = 1023 rows, quando temos 940 rows total).

O utilizador que menos dias teve de captura de informação foi o id `4057192912` que utilizou o app apenas 4 dias,

- 12/04/2016
- 13/04/2016
- 14/04/2016
- 15/04/2016

Não há evidências de que este id teve seus dados capturados em outro id, já que todos os outros também tem dados no período acima, o que nos faz acreditar que este utilizador somente utilizou neste período acima.

Os demais ids tiveram um nível de utilização razoável para o período, sendo o segundo id com menos utilização o id `2347167796` com 18 dias. Temos 30 ids com 20 ou mais dias de captura de informação.

A princípio, não temos evidência da existência de algum tipo de viés, até mesmo porque os dados são descaracterizados.

Podemos também assumir que os mesmos são confiáveis e originais, pois o Kaggle geralmente é criterioso em ter este tipo de coleções em sua comunidade.

Como indicado anteriormente, os dados são de 2016, e portanto não são tão recentes. Contudo, dado o carater das informações (dados de atividade e consumo calórico), não é um critério que possa a vir a prejudicar a análise.

Os dados incluem a fonte, contudo, como vimos anteriormente, a abrangência é um pouco limitada (pouco mais de 30 utilizadores por um período de 31 dias). Se pensarmos na população potencial e/ou utilizadora de apps de monitoração é um período muito curto e com poucas pessoas participantes.

## ● Como você está lidando com o licenciamento, a privacidade, a segurança e a acessibilidade?

Os dados são públicos e anonimizados, por isto, não é necessário maiores cuidados neste sentido. Por também estarem a ser utilizados em um diretório local e com o propóstio desta análise, tomaremos os dados de segurança e acessibilidade para garantir a integridade destes dados.

## ● Como você verificou a integridade dos dados? Como isso o ajuda a responder à sua pergunta?

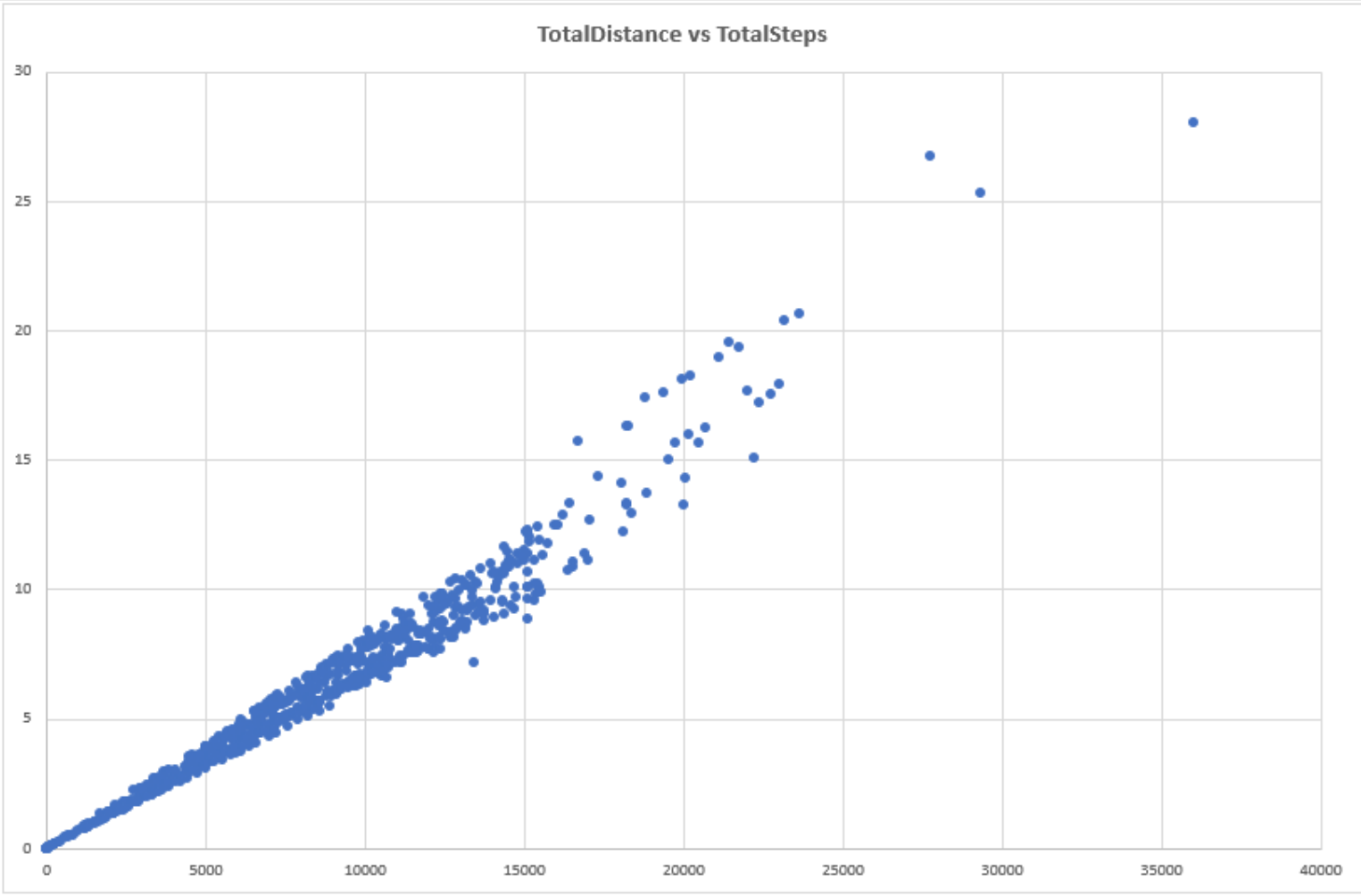
A integridade foi verificada com o uso do Excel para perceber potenciais indicios de problemas nos dados.

O Excel é uma ferramenta prática que com simples funções, como a inclusão de tabelas dinâmicas, classificações e filtros de dados permitem perceber falhas na integridade dos dados.

## ● Há algum problema com os dados?

Analisamos por exemplo, se confirmaria-se a hipótese de que nos dados existentes, para maiores distâncias percorridas, maiores teriam sido os passos necessários.

Com isto fizemos o seguinte gráfico no Excel:




Quer dizer: há uma relação positiva e direta entre os dois indicadores, evidenciando uma clara pertinência nos dados existentes.

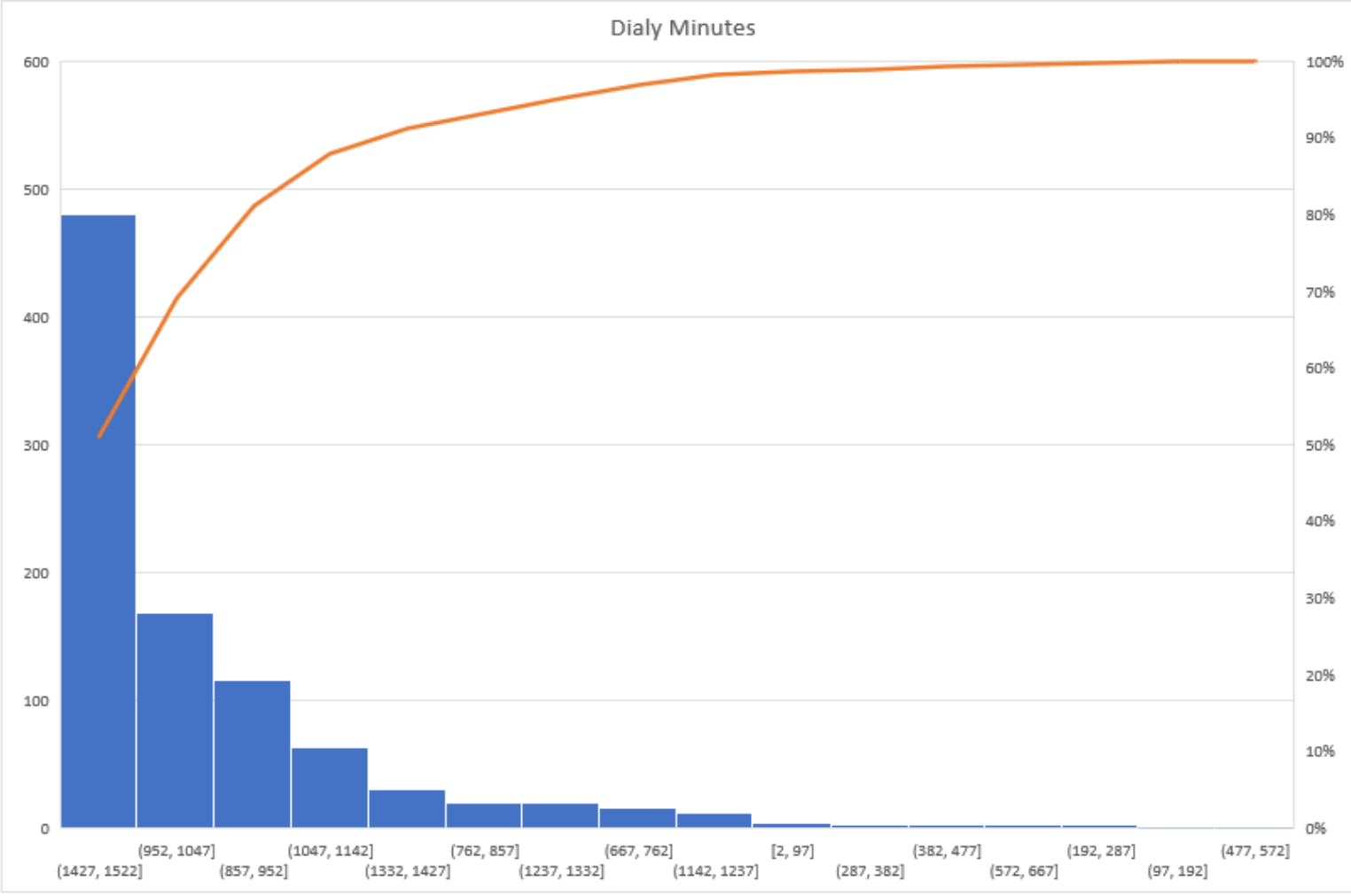
No dataset, temos os seguintes campos:

- TotalDistance (sum: 5160,319995)
- TrackerDistance (Sum: 5146,829994)
- LoggedActivitiesDistance (sum: 101,6806835)
- VeryActiveDistance (sum: 1412,52)
- ModeratelyActiveDistance (Sum: 533,4899983)
- LightActiveDistance (sum: 3140,37)
- SedentaryActiveDistance (Sum: 1,509999979)

Imaginavamos que a soma das distancias segregadas iria ser o mesmo das distancias totais, mas isto não se verificou.

Um outro teste efetuado foi verificar se para cada utilizador / dia, a soma dos tempos segregados era um valor que não ultrapassasse o número de horas que um dia possui, isto é, 1440 minutos.

Obviamente não esperavamos que todos os valores fossem igual a 1440, pois o utilizador poderia nao ter usado o app neste periodo por diversas razões. Aqui, inicialmente, era encontrar valores positivos e menores que 1440.




Daily Minutes

Como podemos observar, há uma integridade nestes dados já que estão no intervalo de [0, 1440]. Através deste visual percebemos que 51.06% da combinação de dias x utilizadores foram acima de 1427 minutos (primeiro nível neste gráfico de Pareto).

# Roteiro do estudo de caso - Processar

## ● Quais ferramentas você está escolhendo e por quê?

Para a etapa de processamento dos dados estamos a utilizar o Excel por ser uma ferramenta simples e prática e que facilmente auxilia nesta tarefa.

## ● Você garantiu a integridade dos seus dados?

Até o momento, identificamos potenciais situações que põe em risco a integridade do nosso dataset. Iremos realizar a limpeza de dados de forma a garantir a completa integridade dos dados durante esta etapa.

## ● Que medidas foram tomadas para garantir que seus dados estejam limpos?

Até o momento, identificamos potenciais situações que põe em risco a integridade do nosso dataset. Iremos realizar a limpeza de dados de forma a garantir a completa integridade dos dados durante esta etapa.

Diante do exposto anteriormente, iremos realizar as seguintes etapas:

- Exclusão do Id `4057192912` que utilizou o app apenas 4 dias,
- Exclusão das colunas `TotalDistance`, `TrackerDistance` e `LoggedActivitiesDistance` (usaremos um somatório das colunas segmentadas como Distancia Total)

- Inclusão de uma coluna com o total dos Minutos de Utilização ( `TotalMinutes = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes + SedentaryMinutes` )

### ● Como você pode verificar se seus dados estão limpos e prontos para análise?

Com a efetivação dos processos de transformação, podemos realizar testes que evidenciam os detalhes que originaram os processos implementados.

Por exemplo: como realizamos a exclusão do utilizador que teve uma pequena participação no processo, basta validar que o mesmo não se encontra mais no dataset final.

### ● Você documentou seu processo de limpeza para poder revisar e compartilhar esses resultados?

Todo o processo de análise da limpeza está a ser documentado em um documento Markdown no `RStudio` .

Como não houve missing data ou necessidades de tratamento de dados, não há etapas a serem incluídas neste processo.

Decidimos fazer alguns processamentos que ficaram gravados no `Power Query` que apresentamos a seguir:

### dailyActivity_merged

```
let
    Source = Csv.Document(File.Contents("C:\_google data analytics\08 Projeto final de Data Analytics d
o Google\Fitabase Data 4.12.16-5.12.16\dailyActivity_merged.csv"),[Delimiter=",", Columns=15, Encoding=
1252, QuoteStyle=QuoteStyle.None]),
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"Id", Int64.Type}, {"ActivityDat
e", type date}, {"TotalSteps", Int64.Type}, {"TotalDistance", type number}, {"TrackerDistance", type nu
mber}, {"LoggedActivitiesDistance", type number}, {"VeryActiveDistance", type number}, {"ModeratelyActi
veDistance", type number}, {"LightActiveDistance", type number}, {"SedentaryActiveDistance", type numbe
r}, {"VeryActiveMinutes", Int64.Type}, {"FairlyActiveMinutes", Int64.Type}, {"LightlyActiveMinutes", In
t64.Type}, {"SedentaryMinutes", Int64.Type}, {"Calories", Int64.Type}}),
    #"Filtered Rows" = Table.SelectRows(#"Changed Type", each [Id] <> 4057192912),
    #"Removed Columns" = Table.RemoveColumns(#"Filtered Rows",{"TotalDistance", "TrackerDistance", "Log
gedActivitiesDistance"}),
    #"Inserted Sum" = Table.AddColumn(#"Removed Columns", "TotalDistance", each List.Sum({[VeryActiveDi
stance], [ModeratelyActiveDistance], [LightActiveDistance], [SedentaryActiveDistance]}), type number),
    #"Inserted Sum1" = Table.AddColumn(#"Inserted Sum", "TotalMinutes", each List.Sum({[VeryActiveMinut
es], [FairlyActiveMinutes], [LightlyActiveMinutes], [SedentaryMinutes]}), Int64.Type)
in
    #"Inserted Sum1"
```

Por fim, exportamos o dataset transformado com o nome de `dataset.csv`

# Roteiro do estudo de caso - Analisar

### ● Como você deve organizar seus dados para realizar análises sobre eles?

Para esta etapa de análise, iremos usar o `R Studio` . Para isto iremos realizar a importação da biblioteca `tidyverse`

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Faremos a importação para o `dataset`

```
dataset <- read.csv("../dataset/dataset.csv", sep = ";", dec = ",")
```

Aplicaremos o comando para visualização `View`

```
View(dataset)
```

Apresentaremos os primeiros rows com o comando `head`

```
head(dataset)
```

```
##           Id ActivityDate TotalSteps VeryActiveDistance
## 1 1503960366   12/04/2016      13162               1.88
## 2 1503960366   13/04/2016      10735               1.57
## 3 1503960366   14/04/2016      10460               2.44
## 4 1503960366   15/04/2016       9762               2.14
## 5 1503960366   16/04/2016      12669               2.71
## 6 1503960366   17/04/2016       9705               3.19
##   ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
## 1                     0.55               6.06                        0
## 2                     0.69               4.71                        0
## 3                     0.40               3.91                        0
## 4                     1.26               2.83                        0
## 5                     0.41               5.04                        0
## 6                     0.78               2.51                        0
##   VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
## 1                25                  13                  328              728
## 2                21                  19                  217              776
## 3                30                  11                  181             1218
## 4                29                  34                  209              726
## 5                36                  10                  221              773
## 6                38                  20                  164              539
##   Calories TotalDistance TotalMinutes
## 1     1985          8.49         1094
## 2     1797          6.97         1033
## 3     1776          6.75         1440
## 4     1745          6.23          998
## 5     1863          8.16         1040
## 6     1728          6.48          761
```

- Seus dados foram formatados corretamente?

Vamos verificar os formatos do campo com o comando `str`

```
str(dataset)
```

```
## 'data.frame':    936 obs. of  14 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr  "12/04/2016" "13/04/2016" "14/04/2016" "15/04/2016" ...
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
##  $ TotalDistance            : num  8.49 6.97 6.75 6.23 8.16 ...
##  $ TotalMinutes             : int  1094 1033 1440 998 1040 761 1440 1120 1063 1076 ...
```

É preciso converter para data o campo `ActivityDate`

Vamos aproveitar também e converter o campo `Id` para texto

```
dataset$ActivityDate <- as.Date(dataset$ActivityDate, format = "%d/%m/%Y")
dataset$Id <- as.character(dataset$Id)
```

Para confirmar, vamos aplicar novamente o comando `str`

```
str(dataset)
```

```
## 'data.frame':    936 obs. of  14 variables:
##  $ Id                       : chr  "1503960366" "1503960366" "1503960366" "1503960366" ...
##  $ ActivityDate             : Date, format: "2016-04-12" "2016-04-13" ...
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
##  $ TotalDistance            : num  8.49 6.97 6.75 6.23 8.16 ...
##  $ TotalMinutes             : int  1094 1033 1440 998 1040 761 1440 1120 1063 1076 ...
```

Vamos aplicar o comando `summary` para ter uma breve informação do dataset

```
summary(dataset)
```

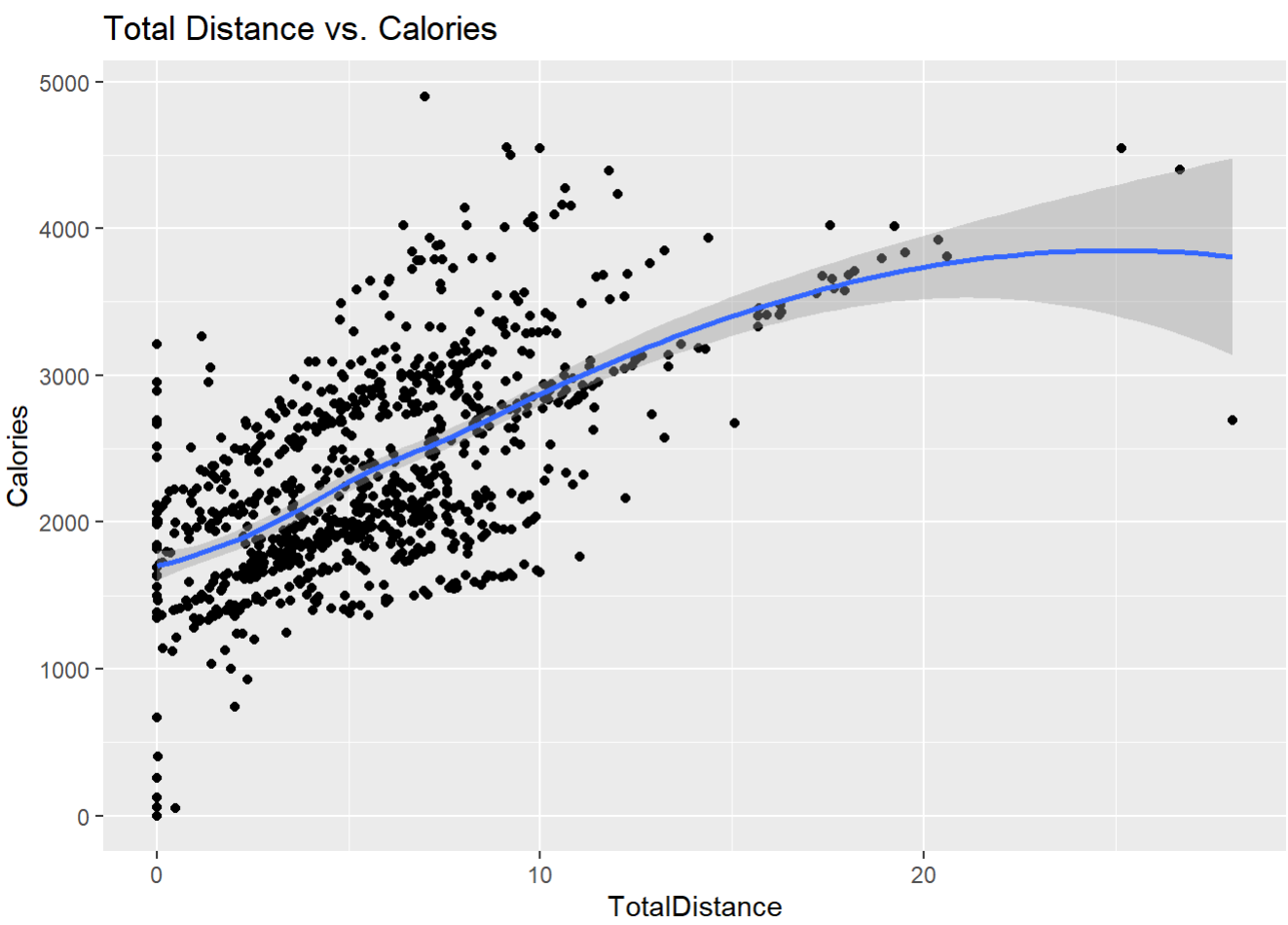
```
##       Id              ActivityDate          TotalSteps    VeryActiveDistance
##  Length:936         Min.   :2016-04-12   Min.   :    0   Min.   : 0.000
##  Class :character   1st Qu.:2016-04-19   1st Qu.: 3790   1st Qu.: 0.000
##  Mode  :character   Median :2016-04-26   Median : 7441   Median : 0.220
##                     Mean   :2016-04-26   Mean   : 7654   Mean   : 1.509
##                     3rd Qu.:2016-05-04   3rd Qu.:10734   3rd Qu.: 2.090
##                     Max.   :2016-05-12   Max.   :36019   Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.945      1st Qu.:0.000000
##  Median :0.2400           Median : 3.365      Median :0.000000
##  Mean   :0.5697           Mean   : 3.344      Mean   :0.001613
##  3rd Qu.:0.8000           3rd Qu.: 4.790      3rd Qu.:0.000000
##  Max.   :6.4800           Max.   :10.710      Max.   :0.110000
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :   0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0        1st Qu.: 729.0
##  Median :  4.00    Median :  7.00      Median :199.0        Median :1057.0
##  Mean   : 21.25    Mean   : 13.62      Mean   :193.2        Mean   : 990.2
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.2        3rd Qu.:1226.8
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
##     Calories    TotalDistance     TotalMinutes
##  Min.   :   0   Min.   : 0.000   Min.   :   2.0
##  1st Qu.:1830   1st Qu.: 2.540   1st Qu.: 989.8
##  Median :2134   Median : 5.190   Median :1440.0
##  Mean   :2305   Mean   : 5.424   Mean   :1218.3
##  3rd Qu.:2794   3rd Qu.: 7.662   3rd Qu.:1440.0
##  Max.   :4900   Max.   :28.040   Max.   :1440.0
```

- Que tendências ou relações você encontrou nos dados?

Vamos verificar a relação entre os campos `TotalDistance` e `Calories`

```
ggplot(data=dataset, aes(x=TotalDistance, y=Calories)) +
  geom_point() +
  geom_smooth() +
  labs(title="Total Distance vs. Calories")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




`geom_smooth()` using method = 'loess' and formula = 'y ~ x'

As próximas etapas serão efetuadas no Tableau.