Etapa 3 - Análise de dados

Leitura das bases do star-schema e respostas das perguntas elaboradas no objetivo

```
#bibliotecas
library(dplyr)
library(SparkR)
```

```
#leitura das bases
fato <- read.df("/amazon_star_schema/fato.parquet", "parquet")
users <- read.df("/amazon_star_schema/users.parquet", "parquet")
products <- read.df("/amazon_star_schema/products.parquet", "parquet")
reviews <- read.df("/amazon_star_schema/reviews.parquet", "parquet")
```

```
# transformação do DF Spark para R
fato <- as.data.frame(fato)
users <- as.data.frame(users)
products <- as.data.frame(products)
reviews <- as.data.frame(reviews)
```

1) Existem usuários que compraram mais de um produto?

```
#usuarios que deram mais de um review para produtos (ou seja, compraram produtos diferentes)

usercounted <- fato %>% dplyr::group_by(user_id) %>% dplyr::count()
user_fil <- usercounted[usercounted$n>1,]
print(paste(nrow(user_fil)," usuários escreveram reviews para mais de um produto"))
```

```
[1] "172  usuários escreveram reviews para mais de um produto"
```

2) As categorias dos usuários que compraram o mesmo produto são as mesmas?

```
# usuarios que compraram mais de um produto:
df <- fato %>% dplyr::filter(user_id %in% user_fil$user_id)

df <- merge(df,products,by="product_id", all.x=T)
df_gp <- df %>% dplyr::group_by(user_id,category) %>% dplyr::count()
df_gp <- df_gp %>% dplyr::group_by(user_id,n) %>% dplyr::count()

x1 <- round(sum(df_gp$n>=2)/nrow(df_gp),4)*100 # compraram na mesma categoria
x2 <- round(sum(df_gp$n<2)/nrow(df_gp),4)*100 # compraram em categorias diferentes

print(paste(x1,"% compraram na mesma categoria e ",x2,"% em categorias diferentes."))
```

```
Storing counts in `nn`, as `n` already present in input
ℹ Use `name = "new_name"` to pick a new name.
[1] "55.03 % compraram na mesma categoria e  44.97 % em categorias diferentes."
```

3) Qual a categoria de produtos mais presente?

```
df <- products %>% dplyr::group_by(category) %>% dplyr::count()
dplyr::arrange(df, desc(n))
```

```
# A tibble: 39 × 2
# Groups:   category [39]
   category                      n
   <chr>                     <int>
 1 "Electronics"               403
 2 "Home&Kitchen"              403
 3 "Computers&Accessories"     343
 4 "OfficeProducts"             26
 5 " AI Voice Assistance"        4
 6 " 60 Sports Modes"            3
 7 " Real Heart Rate Monitor"    3
```

```
 8 "120+ Sports Modes"            3
 9 " 150 Watch Faces"             2
10 " Aluminium Alloy Body"        2
# … with 29 more rows
```

O mercado é dominado pelos produtos de Casa e Cozinha, Eletronicos e Computadores.

4) Qual a média, mediana, mínima e máxima do preço dos produtos das principais categorias?

```
df <- products %>%
  dplyr::group_by(category) %>%
  dplyr::summarise(media = mean(actual_price), mediana=median(actual_price),minimo = min(actual_price),maximo = max(actual_price), n = dplyr::n())
  head(dplyr::arrange(df, desc(n)),4)
```

```
# A tibble: 4 × 6
  category              media mediana minimo maximo     n
  <chr>                 <dbl>   <dbl>  <dbl>  <dbl> <int>
1 Electronics           690.    167.   11.5   9373.   403
2 Home&Kitchen          283.    134.    5.29  5091.   403
3 Computers&Accessories 108.     66.9   2.61  2546.   343
4 OfficeProducts         28.2    13.1   3.35   201.    26
```

O maior preço foi da categoria de Electronics, por 9373 reais, e o menor da categoria de computadores por 2,61 reais.

Lembrete: Os valores originais da base estavam em rúpias indianas e foram convertidos para Reais com base na cotação do dia 01/07/2024 no notebook create_schema.

5) Qual a porcentagem de produtos com nota de avaliação suficientemente confiável?

Pelo TLC, um tamanho de amostra 30 é o suficiente para a aproximação da distribuição normal. Aqui pode-se imaginar que se o rating_count for maior que 30 as notas são confiáveis.

```
df <- products %>%
  dplyr::mutate(nota_maior_30 = ifelse(rating_count>=30,T,F)) %>%
  dplyr::group_by(nota_maior_30) %>% dplyr::count()

df$porcentagem <- round((df$n/sum(df$n))*100,2)
display(df)
```

Table

	nota_maior_30	n	porcentagem
1	false	32	2.61
2	true	1150	93.8
3	null	44	3.59

3 rows

Quase 94% dos produtos possuem nota com quantidade suficiente de avaliações para confiabilidade da nota