Winter Challenge

Michael He

9/20/2021

Task 1

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

Answers 1A

Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

A quick glimpse of the dataset shows the bulk orders, which can significantly distort the average order value.

```
library(googlesheets4)
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
# downloading the spreadsheet completely
gs4_deauth()
sneakershop <- read_sheet("16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM",
sheet="Sheet1",)
```

```
## v Reading from "2019 Winter Data Science Intern Challenge Data Set".
```

```
## v Range ''Sheet1''.
```

```
#[,1:4] # skip = 1, keeping only some columns, skip row 1

# a quick ordering of order in descending dollar amount
sneakershop %>%
  arrange(desc(order_amount))
```

```
## # A tibble: 5,000 x 7
##    order_id shop_id user_id order_amount total_items payment_method
##       <dbl>   <dbl>   <dbl>        <dbl>       <dbl> <chr>
##  1       16      42     607       704000        2000 credit_card
```

```
##  2       61      42     607       704000        2000 credit_card
##  3      521      42     607       704000        2000 credit_card
##  4     1105      42     607       704000        2000 credit_card
##  5     1363      42     607       704000        2000 credit_card
##  6     1437      42     607       704000        2000 credit_card
##  7     1563      42     607       704000        2000 credit_card
##  8     1603      42     607       704000        2000 credit_card
##  9     2154      42     607       704000        2000 credit_card
## 10     2298      42     607       704000        2000 credit_card
## # ... with 4,990 more rows, and 1 more variable: created_at <dttm>
```

```
sneakershop %>%
  filter(order_amount < 100000) %>%
  summarize(AOV = mean(order_amount))
```

```
## # A tibble: 1 x 1
##     AOV
##   <dbl>
## 1  703.
```

Part 1B

What metric would you report for this dataset? What is its value?

```
sneakershop %>%
  summarize(MedOrd = median(order_amount))
```

```
## # A tibble: 1 x 1
##   MedOrd
##    <dbl>
## 1    284
```

Part 2

For this question you'll need to use SQL. Follow this link to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL

Answer Part 2A

How many orders were shipped by Speedy Express in total?

Answer Part 2B

What is the last name of the employee with the most orders?

Answer Part 2C

What product was ordered the most by customers in Germany?