Shopify Challenge

Yongpeng Fu

15/05/2022

Question 1:

Given some sample data, write a program to answer the following: click **here** to access the required data set.

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.

- (a) Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
- (b) What metric would you report for this dataset?
- (c) What is its value?

Load data

```
setwd("/Users/yongpengfu/Google Drive/Job Search in Canada/2022/Job Looking/Shopify")
challenge <- as.tibble(read.csv("2019 Winter Data Science Intern Challenge Data Set.csv"))
```

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
```

```
str(challenge)
```

```
## tibble [5,000 x 7] (S3: tbl_df/tbl/data.frame)
##  $ order_id      : int [1:5000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ shop_id       : int [1:5000] 53 92 44 18 18 58 87 22 64 52 ...
##  $ user_id       : int [1:5000] 746 925 861 935 883 882 915 761 914 788 ...
##  $ order_amount  : int [1:5000] 224 90 144 156 156 138 149 292 266 146 ...
##  $ total_items   : int [1:5000] 2 1 1 1 1 1 1 2 2 1 ...
##  $ payment_method: chr [1:5000] "cash" "cash" "cash" "credit_card" ...
##  $ created_at    : chr [1:5000] "2017-03-13 12:36:56" "2017-03-03 17:38:52" "2017-03-14 4:23:56" "201
```

From the structure of the data, I noticed that order_id, shop_id, and user_id are integer. They are actually character variable. And created_at is character type, while it should be datetime. I will change all of them.

```
challenge$order_id <- as.character(challenge$order_id)
challenge$shop_id <- as.character(challenge$shop_id)
challenge$user_id <- as.character(challenge$user_id)
challenge$created_at <- as.POSIXct(challenge$created_at)
str(challenge)
```

```
## tibble [5,000 x 7] (S3: tbl_df/tbl/data.frame)
```

```
##  $ order_id      : chr [1:5000] "1" "2" "3" "4" ...
##  $ shop_id       : chr [1:5000] "53" "92" "44" "18" ...
##  $ user_id       : chr [1:5000] "746" "925" "861" "935" ...
##  $ order_amount  : int [1:5000] 224 90 144 156 156 138 149 292 266 146 ...
##  $ total_items   : int [1:5000] 2 1 1 1 1 1 1 2 2 1 ...
##  $ payment_method: chr [1:5000] "cash" "cash" "cash" "credit_card" ...
##  $ created_at    : POSIXct[1:5000], format: "2017-03-13" "2017-03-03" ...
```

(a) Show statistics of the dataframe.

```
summary(challenge)
```

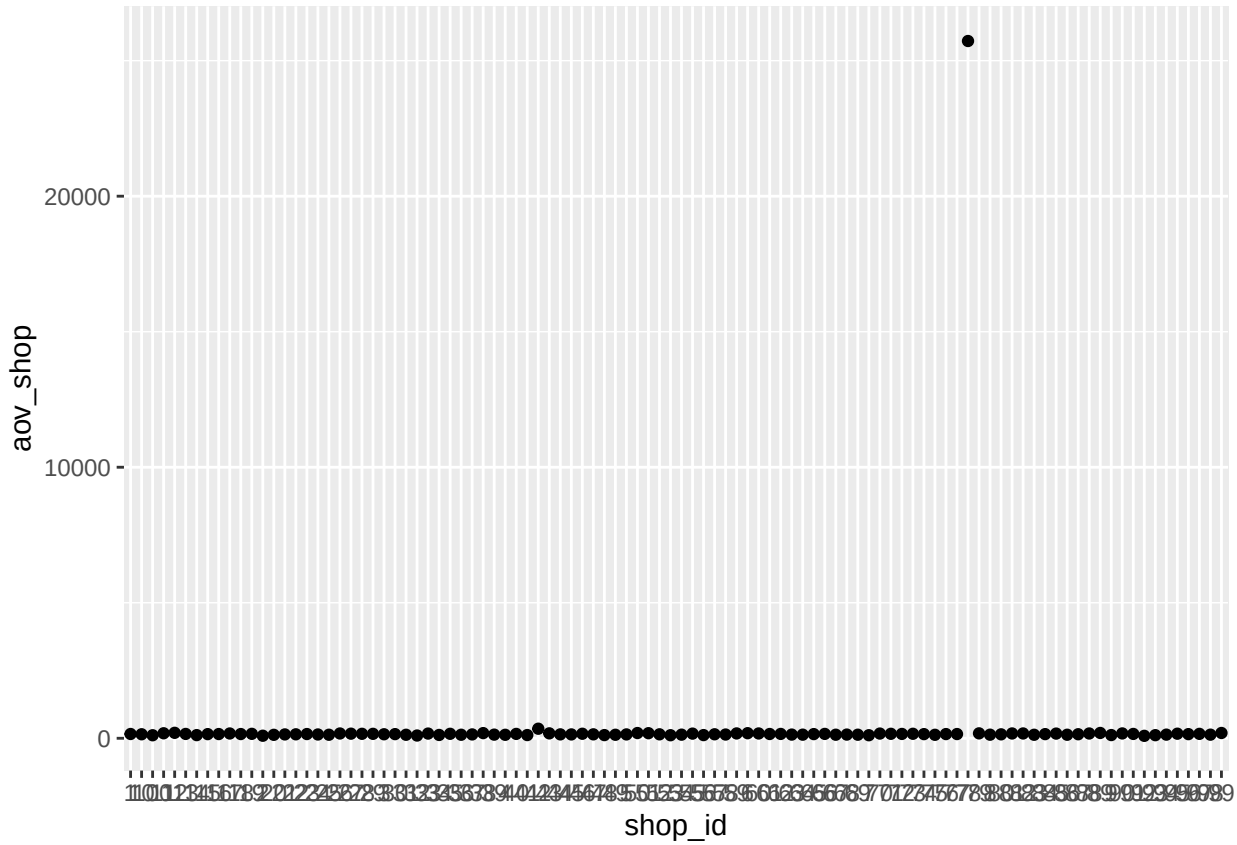
```
##    order_id           shop_id            user_id           order_amount
##  Length:5000        Length:5000        Length:5000        Min.   :    90
##  Class :character   Class :character   Class :character   1st Qu.:   163
##  Mode  :character   Mode  :character   Mode  :character   Median :   284
##                                                           Mean   :  3145
##                                                           3rd Qu.:   390
##                                                           Max.   :704000
##   total_items       payment_method       created_at
##  Min.   :   1.000   Length:5000        Min.   :2017-03-01 00:00:00
##  1st Qu.:   1.000   Class :character   1st Qu.:2017-03-08 00:00:00
##  Median :   2.000   Mode  :character   Median :2017-03-16 00:00:00
##  Mean   :   8.787                      Mean   :2017-03-15 10:50:39
##  3rd Qu.:   3.000                      3rd Qu.:2017-03-23 00:00:00
##  Max.   :2000.000                      Max.   :2017-03-30 00:00:00
```

Answer a: From the above summary statistics, we can see the calculated AOV ($3145) is the mean value of the total order amount for all shops. However, the actual AOV is calculated by dividing total revenue by the number of orders over the 30-day period.

(b) I would perfer to calculating AOV for each sneaker shop.

I will group_by the dataset based on each shop id then divide the total revenue by the total number of orders for each shop.

```
aov_each_shop <- challenge %>% group_by(shop_id) %>% summarise(aov_shop = sum(order_amount)/sum(total_i
ggplot(data = aov_each_shop, aes(x = shop_id, y = aov_shop)) + geom_point()
```



From the above scatter plot, I can see one outlier. I will take a closer look at this particular one.

```
aov_each_shop[which(aov_each_shop$aov_shop == max(aov_each_shop$aov_shop)),]
```

```
## # A tibble: 1 x 2
##   shop_id aov_shop
##   <chr>      <dbl>
## 1 78         25725
```

The outlier is shop with id 78. The AOV for shop 78 is too high to be true, given the averge price for a pair of sneaker is between $70 and $250. The dataset for this outlier shop is as follows:

```
challenge %>% filter(shop_id == 78)
```

```
## # A tibble: 46 x 7
##    order_id shop_id user_id order_amount total_items payment_method
##    <chr>    <chr>   <chr>          <int>       <int> <chr>
##  1 161      78      990            25725           1 credit_card
##  2 491      78      936            51450           2 debit
##  3 494      78      983            51450           2 cash
##  4 512      78      967            51450           2 cash
##  5 618      78      760            51450           2 cash
##  6 692      78      878           154350           6 debit
##  7 1057     78      800            25725           1 debit
##  8 1194     78      944            25725           1 debit
##  9 1205     78      970            25725           1 credit_card
## 10 1260     78      775            77175           3 credit_card
## # ... with 36 more rows, and 1 more variable: created_at <dttm>
```

(c) The corresponding AOV value based on (b)

We see shop 78 was selling one sneaker at price of $25725, which I believe it is a typo. It probably meant $257.25. We will correct this number and generate the AOV again.

```
aov_each_shop[aov_each_shop$shop_id == 78,'aov_shop'] <- 257.25
aov_each_shop %>% head(10)
```

```
## # A tibble: 10 x 2
##    shop_id aov_shop
##    <chr>      <dbl>
##  1 1            158
##  2 10           148
##  3 100          111
##  4 11           184
##  5 12           201
##  6 13           160
##  7 14           116
##  8 15           153
##  9 16           156
## 10 17           176
```