Data Science Challenge

bochao

2022/1/17

1)

```
dt <- read.csv("2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv")

avg_wrong <- sum(dt[,4])/length(dt[,5]) ### Which is wrong

avg_right <- sum(dt[,4])/sum(dt[,5]) ### which is right

avg_wrong
```

```
## [1] 3145.128
```

```
avg_right
```

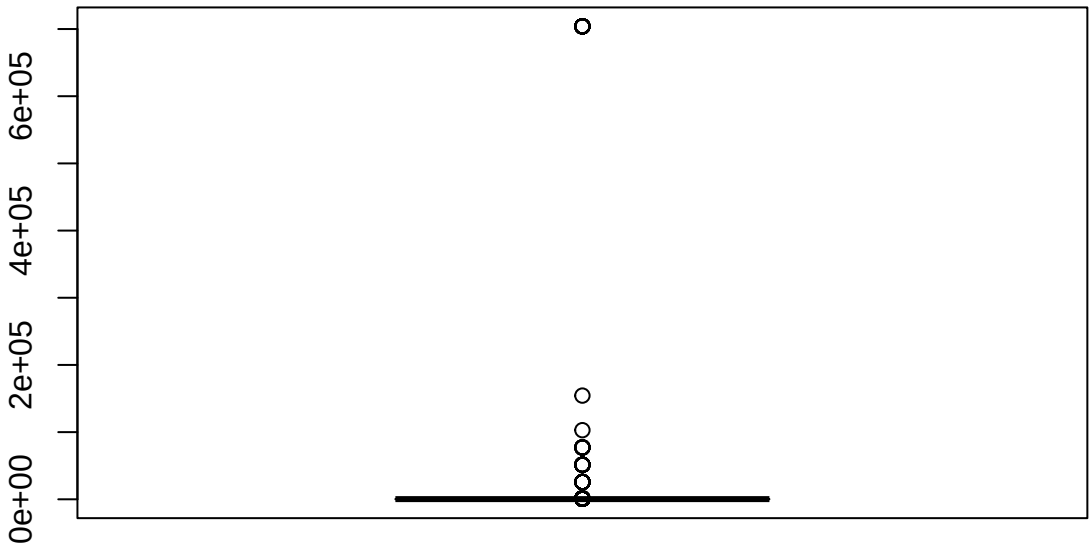
```
## [1] 357.9215
```

As the above result shown, AOV of 3145.128 is calculated using the number of data records, not the number of orders. The correct AOV is 357.9251.

2) I want to know how many records is considered as outliers in the order_amount.

3) The anwser is, there are 141 outliers.

```
plt <- boxplot(dt[,4])
```




```
length(plt$out)
```

```
## [1] 141
```