Candidate: CHU WANG

Summer 2022 Data Science Intern Challenge:

Question1:

a.

The business problem from the given datasets is to analyze the AOV of 100 sneaker shops, each of them only sell one mode. Calculating AOV directly using the 100 shops total sales divided by total order number may suffer from Outliers or Some shops have a very high sales which will increase the overall average level. After carefully exploration and verification, I found there are two shops (id=42, id=78) have a very high possibility with outliers, which results in the overall AOV higher.

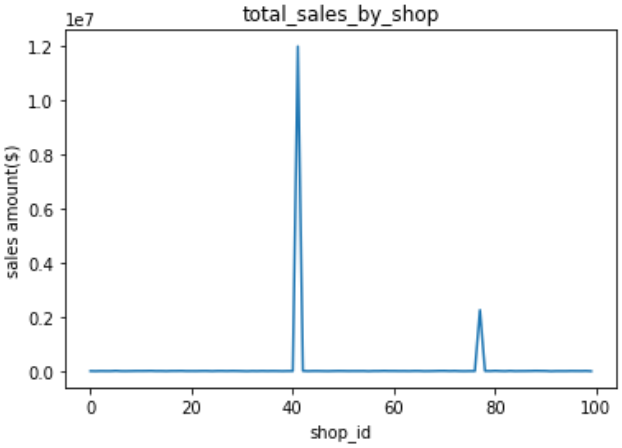
My detailed analysis process are as follows:

1.Download the datasets, read through PySpark with defined schema and conducted data exploration

Findings:

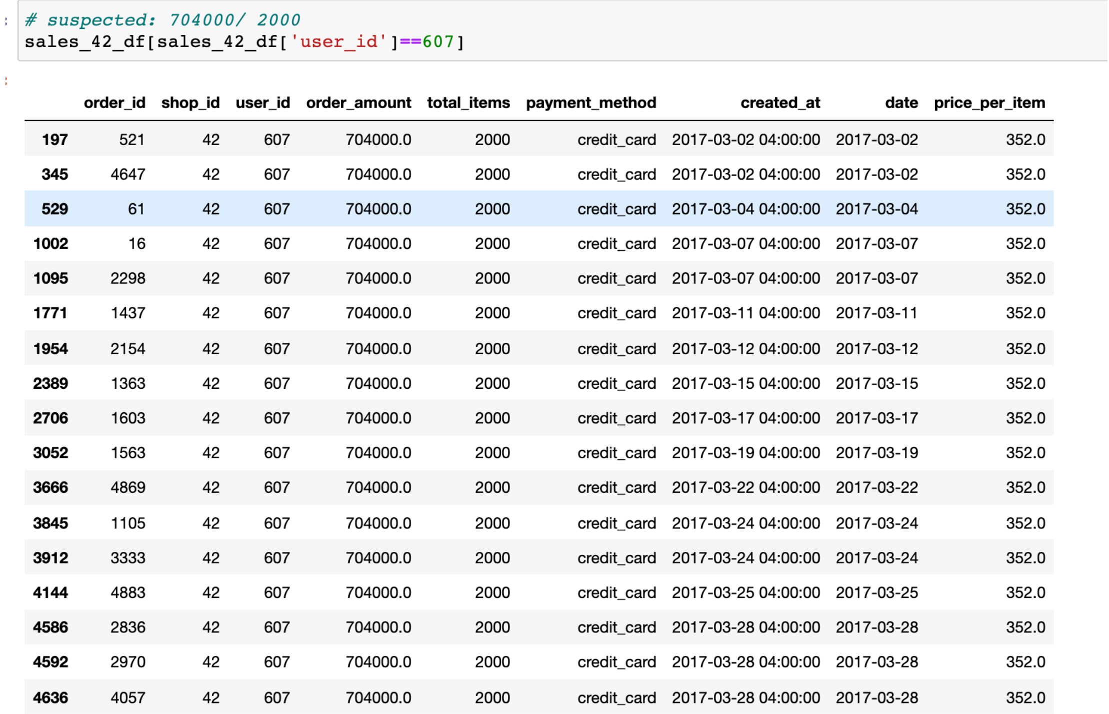
- 5000 records all happen between 2017-03-01 to 2017-03-31
- Each shop's shoe price is around **$90-350**, except **shop 78** which is **$25725.**
- After checking all shop 78 ordering history, the buying behavior is normal except for the high price. Normally a shoe can't be $25725, which directly results the total sales amount for shop78 takes up over **14%** of overall 100 shops sales amount.

2. Plot each shop's **total sales amount** over the 30 days, check for whether there would be outliers



Findings:

- From above plot, except from shop78, **shop42** also has a very suspicious sales amount, which is way much higher than most stores.
- After check shop42 records, I found there are some orders incurring $704000 amount for 2000 items, all from the same **user id 607 (**As shown in following picture**)**



```
# suspected: 704000/ 2000
sales_42_df[sales_42_df['user_id']==607]
```

	order_id	shop_id	user_id	order_amount	total_items	payment_method	created_at	date	price_per_item
197	521	42	607	704000.0	2000	credit_card	2017-03-02 04:00:00	2017-03-02	352.0
345	4647	42	607	704000.0	2000	credit_card	2017-03-02 04:00:00	2017-03-02	352.0
529	61	42	607	704000.0	2000	credit_card	2017-03-04 04:00:00	2017-03-04	352.0
1002	16	42	607	704000.0	2000	credit_card	2017-03-07 04:00:00	2017-03-07	352.0
1095	2298	42	607	704000.0	2000	credit_card	2017-03-07 04:00:00	2017-03-07	352.0
1771	1437	42	607	704000.0	2000	credit_card	2017-03-11 04:00:00	2017-03-11	352.0
1954	2154	42	607	704000.0	2000	credit_card	2017-03-12 04:00:00	2017-03-12	352.0
2389	1363	42	607	704000.0	2000	credit_card	2017-03-15 04:00:00	2017-03-15	352.0
2706	1603	42	607	704000.0	2000	credit_card	2017-03-17 04:00:00	2017-03-17	352.0
3052	1563	42	607	704000.0	2000	credit_card	2017-03-19 04:00:00	2017-03-19	352.0
3666	4869	42	607	704000.0	2000	credit_card	2017-03-22 04:00:00	2017-03-22	352.0
3845	1105	42	607	704000.0	2000	credit_card	2017-03-24 04:00:00	2017-03-24	352.0
3912	3333	42	607	704000.0	2000	credit_card	2017-03-24 04:00:00	2017-03-24	352.0
4144	4883	42	607	704000.0	2000	credit_card	2017-03-25 04:00:00	2017-03-25	352.0
4586	2836	42	607	704000.0	2000	credit_card	2017-03-28 04:00:00	2017-03-28	352.0
4592	2970	42	607	704000.0	2000	credit_card	2017-03-28 04:00:00	2017-03-28	352.0
4636	4057	42	607	704000.0	2000	credit_card	2017-03-28 04:00:00	2017-03-28	352.0

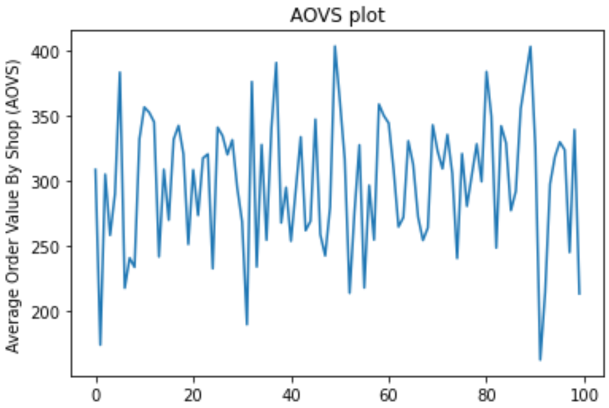
Conclusion:

1. **Shop78's shoe price may have some 'record error' which is an outlier**
2. **Shop42's sales amount also can be counted as an outlier, which drastically increase the AOV. Given the purchase all comes from user 607, there maybe either a 'record error' or 'fraud behavior', which could be tagged for further exploration.**

b.

From above analysis, a much more appropriate metric for given business problem could be **'Average Order Value by Shop' (AOVS)**. By using this measure, we could get a clearer picture of how different shop performs in 30 days and makes **outlier detection** much easier.

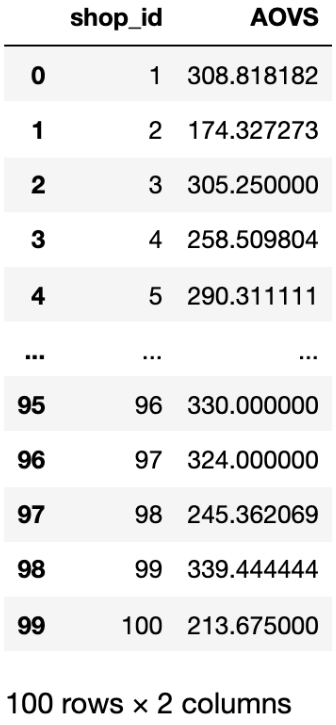
In this case, I did a further exploration to verify my previous thoughts. By calculating every shop's AOVS and filter shop78 and shop42 (Which needs further check), then plot the AOVS for each shop as following:



By filtering outliers, we could see the sales difference among shops, then we can find shops with low sales and further explore solutions for increasing sales.

c.

AOVS value for each shop is saved in the excel file (AOVS.xlsx) and attached screenshot is a slice of the file because of limited space. But, to give a more straightforward sense, I calculated the **average AOVS** (except shop78 and shop42) is **$299.68**, which is way much lower than naively calculated $3145.13.



	shop_id	AOVS
0	1	308.818182
1	2	174.327273
2	3	305.250000
3	4	258.509804
4	5	290.311111
...	...	...
95	96	330.000000
96	97	324.000000
97	98	245.362069
98	99	339.444444
99	100	213.675000

100 rows × 2 columns

Question 2:

a. 54

```
SELECT COUNT(*) FROM Orders
WHERE ShipperID = (
        SELECT ShipperID FROM Shippers
        WHERE ShipperName = 'Speedy Express');
```

b. Peacock

```
SELECT LastName FROM Employees
WHERE EmployeeID = (
        SELECT EmployeeID FROM Orders
        GROUP BY EmployeeID
        ORDER BY COUNT(*) DESC
        LIMIT 1);
```

c. Boston Crab Meat

```
SELECT ProductName FROM Products
WHERE ProductID =
        (SELECT ProductID FROM OrderDetails
        WHERE OrderID in
                (SELECT OrderID FROM Orders
                WHERE CustomerID in (
                        SELECT CustomerID FROM Customers
                        WHERE Country = 'Germany')
                )
        GROUP BY ProductID
        ORDER BY SUM(Quantity) DESC
        LIMIT 1);
```