Fall 2021 Data Science Intern Challenge
Applicant Name: Junduo Dong

Question 1

a. From the given data set, each sneaker shops (shop id) have multiple order id with its order amount, total items, and payment method and created time, and each sneaker shops have multiple user id as well. From the given question description, a metric 'average order value' has been calculated throughout 5000 data records over a 30-day window.

   By critique the calculation. Firstly, we can tell in the dataset, although each sneaker ships only sells one model of shoe. However, using the approach to arbitrarily calculate the average value through the whole dataset is not reasonable since different variables will create variations between different levels. For example:

   - different total items
   - different payment method
   - correlation between different numeric variables
   - create time effect

   Therefore, a better way to evaluate the data, or to calculate the average order value is to make multiple segmentations base on. the variables given above. (I will provide proof of works in section c)

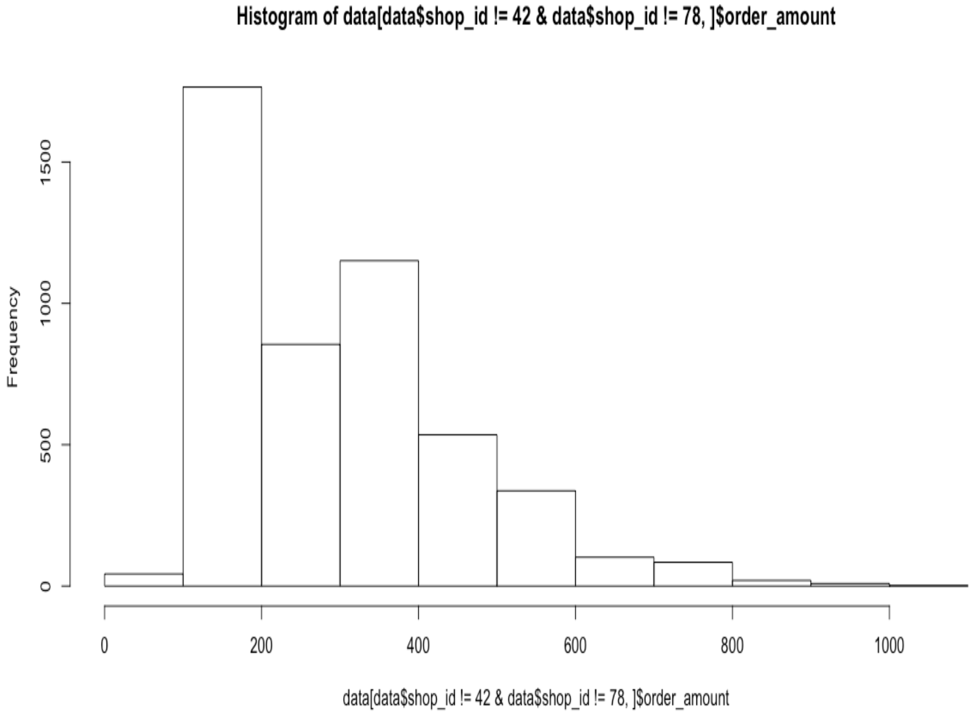
b. I will stay on the average order value (AOV) as the metric since my approach is to segment the given dataset. Therefore, we will be more confident to calculate multiple averages outcomes that can be representative of each group.

   However, because the given dataset is only base on a 30-day window, to be able to receive an expected and more precise average value, I recommend improving the data by increase the time interval to a quarter or a years' sales data. Moreover, after collecting more data, I will suggest using the RFM model to evaluate the term 'order value'. Since the monetary can represent the total revenue of a shoe, however, the other two terms 'recency' and 'frequent' also tells the popularity of a certain type of sneaker. An ideal sneaker shop can have lower recency with higher frequency and monetary.

c. To prove my response to previous questions, I impended different approaches to compare means in order amounts with different variables. I use R to implement multiple analyses of variance for different total items groups and payment methods. Then, I use a regular expression to extract days, hours, and minutes from "created at" and test the Pearson correlation coefficient between all numeric variables to see if there is co-variance between different variables that potentially affect the order amount.

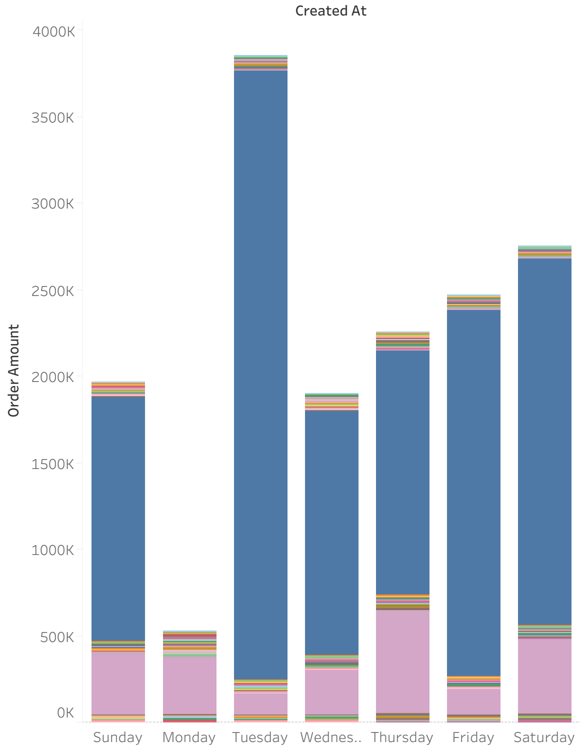
Finally, I use a linear regression model to find out a certain set of variables that have a significant impact to order amounts. I use bullet points to explain the analysis by each step:

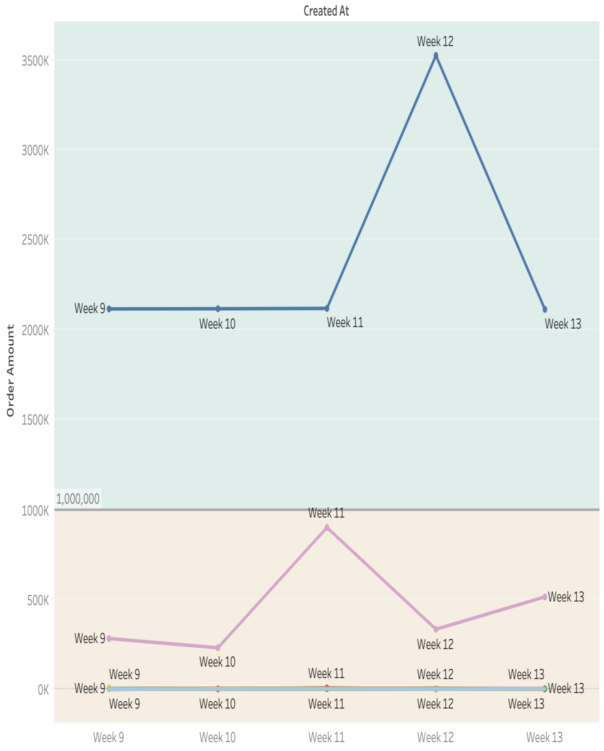
- Firstly, I check the Pearson correlation coefficient between all initial numeric variables. All variables do not have a strong correlation with each other except there is a strong positive correlation between order amount and the total items.
- There are some high leverage points that exists in the dataset, which they exist for sneaker shop 42 and 78. (we will analyze these points later) After eliminating these data points, we can see the distribution of the order amount is a positive-skew distribution, this means a majority of soled shoes are relatively affordable and less expensive shoes in the given dataset.



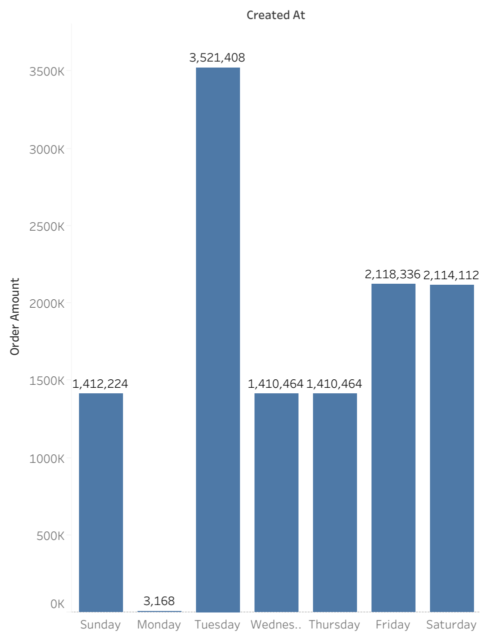
-
- I found there is a data imbalance issue before I use multiple analyses of variance between the number of records for each total item. To be more specific, the total items 1 to 5 have a number of records greater than 75. However, total items 6, 8, and 2000 have less than 20 records. Therefore, I decided to only compare the mean between groups 1 to 5 with random select 75 records within each group. The result indicates there is a significant difference in means between each group except for group 1 and 2, group 5 and 4.
- Then, I use multiple comparisons of means for different payment methods, since they have the relatively same number of records. Therefore, the result shows credit card has a relatively higher means of order amounts than cash and debit.

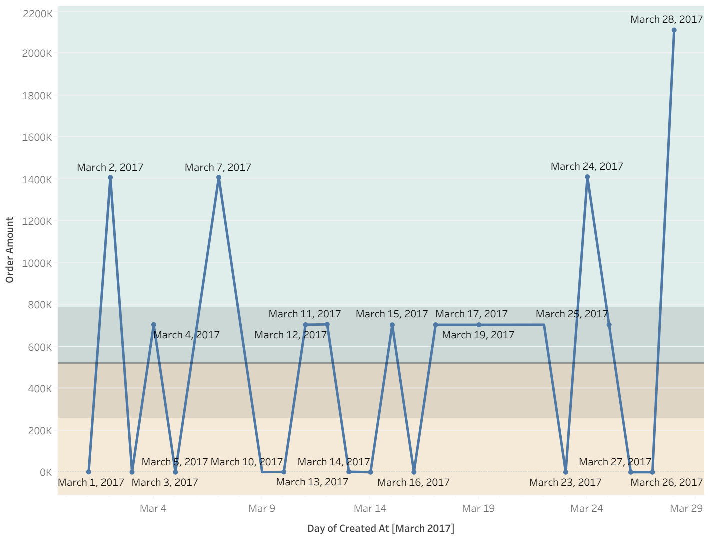
- Lastly, I use linear regression to find out that, except for different sneaker shops and total items, the variable ‘day’ also has a relatively high impact on order amount. So, I break my analysis further with time factors



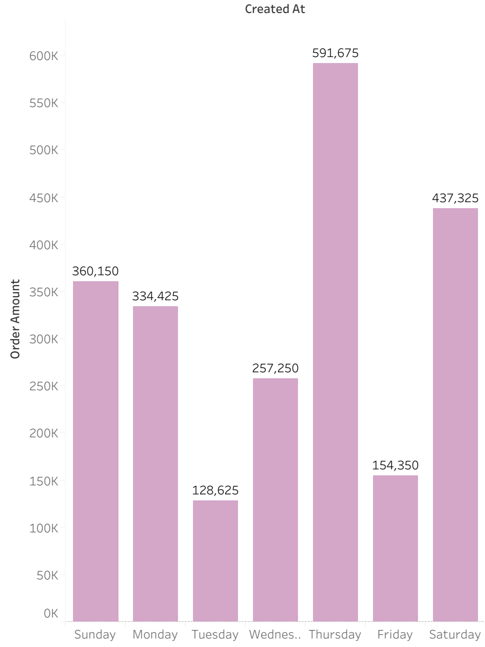


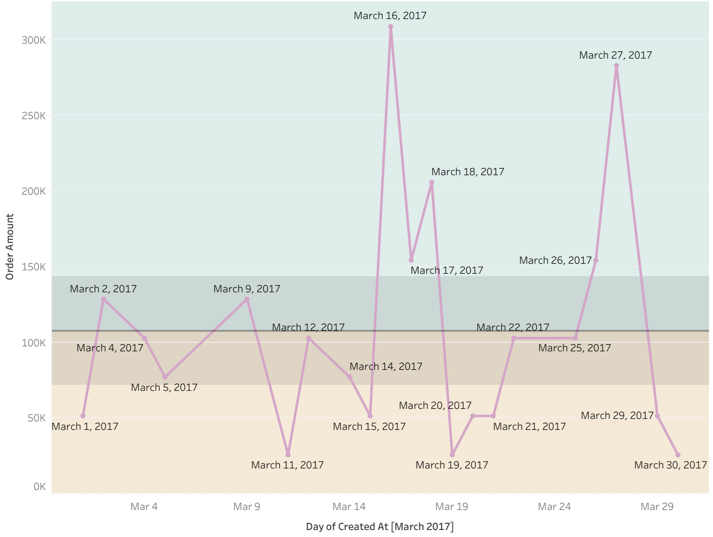
- From the perspective of weekdays, all orders are likely happening except on Monday, sustainable growth from Wednesday Saturday with a little cool down on Sunday, and highest on Tuesday. And from the perspective of week numbers, orders are much more likely to happen in weeks 11 and 12, which are the second half of March.
- If we take a deep dive for stores 42 and 78, we will find out that for store 78, Tuesday is the highest sales day, and Friday and Saturday stand the second-highest sales day and the rest of the weekday are relatively lower. If we level down to the daily level of detail, we will see the order amount are the highest on Thursday and Tuesday.





- For store 42, highest sales day is on Thursday and second highest day are Saturday, Sunday and Monday. If we break to the level of detail of each day, these highest sales days are on Thursday, Saturday and Monday.





- Some key takeaways are:
  - To calculate average order value, we can segment data into 6 groups when total purchased items equal to 1 or 2, 3, 4,5 or 6, 8 and 2000.
  - Also apply segmentation base on different payment type, use credit card as a group, debit and cash a group.

- Analyze order value separately for the first half of March and second half of March. For store 78, analyze order value separately between Tuesday and weekends with other days. Fore store 42, analyze order value separately between Thursday and weekends with other days.

Question 2

```
# How many orders were shipped by Speedy Express in total?

SELECT o.ShipperID,s.ShipperName,COUNT(OrderID) as count
FROM Orders o join Shippers  s
ON o.ShipperID = s.ShipperID
GROUP BY 1
HAVING o.ShipperID  = 1;
```

Answer: 54

```
# What is the last name of the employee with the most orders?
SELECT e.LastName, COUNT(o.OrderID)
FROM Orders o join Employees e
ON o.EmployeeID = e.EmployeeID
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

Answer: Peacock, 40

```
# What product was ordered the most by customers in Germany?
SELECT P.ProductID,P.ProductName,SUM(OD.Quantity) as Quantity
FROM Customers C JOIN Orders O
ON C.CustomerID = O.CustomerID
JOIN OrderDetails OD
ON OD.OrderID = O.OrderID
JOIN Products P
ON OD.ProductID = P.ProductID
GROUP BY 1,2
HAVING C.Country = 'Germany'
ORDER BY 3 DESC;
```

Answer: Boston Crab Meat, 256