Door Dash: Case Study

Part 1

Attached here is a month's worth of sample delivery data (glossary here) for our New Verticals businesses (Grocery, Convenience, Alcohol and DashMart). This isn't our complete dataset, but it's comprehensive enough for you to make observations on our business. The time zone in the dataset is UTC; ensure you are conducting the analysis in the appropriate time zone. The task is to analyze the dataset, and to provide a clear set of recommendations for our business. Please clearly state any assumptions you make throughout your analyses.

Here's what we are assessing:

Overall Deliverable

- 4-6 page written deliverable in PDF format, with strong attention to detail and putting forth a strong strategic point of view
- We do not internally use Powerpoints/Slides, so please submit a written word document in a PDF format (no slides)
- Appendix is optional
- Graphs/visualizations to show supporting analyses

Data Analysis

- Your work should demonstrate that you successfully cleaned the data and used it to analyze our three-sided marketplace
- Descriptive statistics/visualizations that support your recommendations

Recommendations

- Actionable and realistic recommendations, with a path to execution
- Scrappy ideas while also thinking long term and strategically
- Recommendations align with the complexity of our New Verticals business, while taking into account the three-sided marketplace

Below is a glossary of definitions for the variables you will find in the dataset:

- **DELIVERY_UUID:** Unique Identifier for the delivery
- **DELIV_CREATED_AT:** Timestamp the delivery was created at
- **DELIV_STORE_NAME:** Store the delivery was placed at
- **DELIV_DASHER_ID:** Unique identifier for the dasher
- **DELIV_SUBMARKET:** Location the delivery took place in

- **DELIV_D2R:** How long it took the dasher to get from their starting location to the store, in minutes
- **DELIV_IS_20_MIN_LATE (TRUE/FALSE):** field denoting whether or not the Dx was late by 20 minutes in delivering the items
- **DELIV_CLAT:** How many minutes elapsed between when the order was placed, and when the ultimate dasher accepted the delivery
- **DELIV_CANCELLED_AT:** Timestamp for when the delivery was canceled, if at all
- **DELIV_MISSING_INCORRECT_REPORT (TRUE/FALSE):** flag for whether the consumer submitted a complaint saying that their order had a missing or incorrect item. Flagged at the Delivery level.
- **WAS_REQUESTED:** [0 or 1] for whether this row is for an item that was originally ordered by the customer
- **WAS_MISSING:** [0 or 1] = 1 if this row is for an item that was not found by the dasher
- **WAS_SUBBED:** [0 or 1] for whether this row is for an item that was substituted
- **WAS_FOUND:** [0 or 1] = 1 if this row is for an item that was found by the dasher
- **ITEM_NAME:** Name of the originally requested item
- **ITEM_PRICE_CENTS:** Price of the item in cents
- **ITEM_CATEGORY:** Category of the original item
- **SUBSTITUTE_ITEM_NAME:** If there was a substitute, the name of that substitute item
- **SUBSTITUTE_ITEM_CATEGORY:** If there was a substitute, the category of that substitute

Part 1 and Part 2 work shown **below** *(cont.)*

Part 1: Analysis

Intro:

This analysis involves data from DoorDash that illustrates various metrics in the process of a delivery: the grocery logistics/inventory management, Door Dasher performance, and customer satisfaction. With these metrics in mind, I set out my analysis with the goal of looking at information on the 'Grocery Store' level, since this role at Door Dash is centered around logistics management for Dasher stores, therefore I thought it relevant to pursue the analysis in this manner. While there are many other avenues to explore, this will be the approach and aspect that I will conduct the analysis under.

1a) Cleaning the Data

I like to start my analysis by cleaning the data, as part of a standard procedure I run. For this analysis, I started by looking for missing/NULL/Nan values in the data and deciding on keeping or deleting them. 'DELEV_CANCELLED_AT', 'SUBSTITUTE_ITEM_NAME' and 'SUBSTITUTE_ITEM_CATEGORY' each had blank values, but I kept the corresponding rows seeing as logically, some rows would not require these columns' data filled in if they were not canceled or required substitutes. I did update the 'DELIV_CREATED_AT' column to 'DELIV_CREATED_AT_PST' to reflect the correct time zone in PST and did the same for 'DELEV_CANCELLED_AT'. For 'DELIV_D2R' and 'DELIV_CSAT', I found that under 0.5% of the data in these columns had what I would consider being outlier data in the top percentile of data, however I kept the data seeing as the other columns had information that was valuable to the rest of the analysis.

1b) Top Level Grocery Data

With cleaning done, I started my analysis by looking at a pivot table summarizing metrics in the data as it related to each store location. Here is the first part of the summarizing of grocery data:

Row Labels	Average of DELIV_CLAT	Average of DELIV_D2R	Count of DELIV_IS_20_MIN_LATE	Sum of ITEM_PRICE
DashMart1	4.570054104	2.905169184	34349	164504
Grocery1	4.215352151	4.789527604	17726	91377
Grocery2	5.259216523	8.134489363	7534	39041
Grocery3	3.519706596	7.374725082	974	6135
(blank)				
Grand Total	**4.535177863**	**4.178952494**	**60583**	**301057.22**

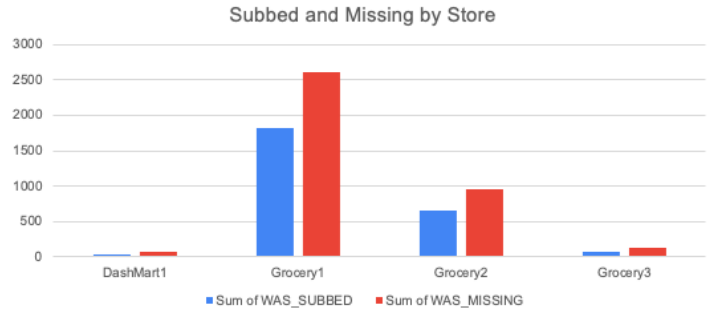
This pivot table above shows us the average time it takes for Dashers to accept a delivery 'DELIV_CLAT' (2nd column), how long on average it takes for Dashers to arrive at the store for pickup 'DELIV_D2R' (3rd column), whether the order was over 20 minutes late 'DELIV_IS_20_MIN_LATE' (4th column), and finally the total cost of all the items delivered in the month long time span of this dataset 'ITEM_PRICE' (last column).

'Grocery 2' stands out for its higher-than-average CLAT and DELIV_D2R times, indicating that the location could be furthest away physically due to the higher D2R time, which could lead to Dashers being more hesitant to accept the order seeing the higher CLAT time as well. Luckily this store is 3rd in our priority list of importance when accounting for money spent, DashMart1 being the 1st, and Grocery 1 being 2nd.

Row Labels	Count of DELIV_MISSING_INCORRECT_REPORT2	Sum of WAS_REQUESTED	Sum of WAS_SUBBED	Sum of WAS_MISSING	Sum of WAS_FOUND
DashMart1	34349	34349	38	78	34145
Grocery1	17726	17726	1817	2620	14977
Grocery2	7534	7534	657	962	6517
Grocery3	974	974	76	126	848
(blank)					
Grand Total	**60583**	**60583**	**2588**	**3786**	**56487**

Moving on we see more information on these stores: the count of deliveries marked as being incorrect (2nd column), the total amount of requested item reports (3rd column), the number of items substituted (4th column), the number of items marked missing (5th column), and finally the number of items marked found (last column). While the last pivot table gave us insights into the actual delivery of the food items, this chart gives us information into the supply chain of the various stores.

DashMart1 has many more incorrect reports than any other store, making up 57% of the total complaints. Despite this, DashMart1 has the lowest 'SUBBED' and 'WAS_MISSING' categories, with nearly the same level of 'WAS_FOUND' as the original number of complaints. Given that complained items where not readily replaced or found missing, potentially indicated a labeling issue for this store on their goods. Dashers are capable of reporting 'SUBBED' and 'MISSING' items as seen in 'Grocery 1', however we're not seeing that consistent pattern in DashMart1, indicating that potential for store item labeling being the issue, especially if customers and not dashers are reporting the issue.



As seen above, Grocery 1 also raises concerns but for different reasons. The number of customer complaint reports on deliveries is less at 29%, however the amount of 'SUBBED' and 'MISSING' amounts for this store is the highest at 70% and 69% respectively. With this high amount of reporting for those 2 columns, this would potentially indicate that Grocery1 has the most significant inventory issue (as opposed to inventory *tracking* in DashMart1), seeing as many items are not available.

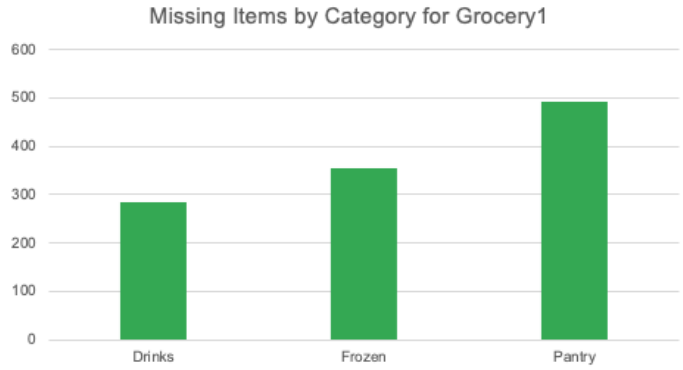
1c) Diving into the Impact from Findings

From our previous findings, I now want to be able to understand the impact of the potential supply chain issues observed in DashMart1 and Grocery 1. First, I want to make sure that the 'MISSING' category of values is statistically significant in the 2 stores, and not a normal missing amount that we would expect to see.

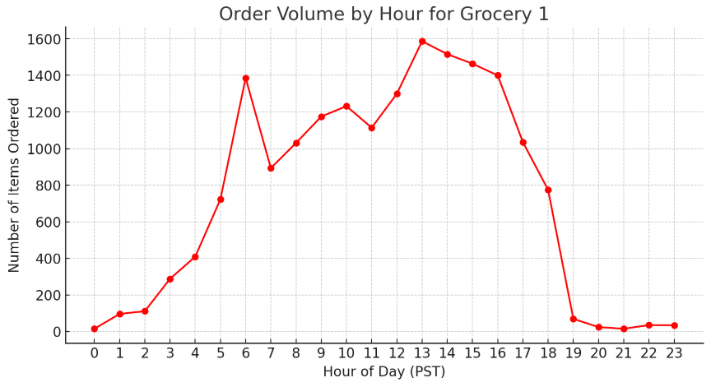
Sum of WAS_MISSING	Expected WAS_MISSING
78	2147
2620	1108
962	471
126	61

To do this, I performed a Chi-squared test to see if the expected value of missing items was statistically significant compared to the actual number of missing items. I got my expected missing values by getting the average missing values as a percentage of total orders, and then using that fraction multiplied to the order amounts for each store, resulting in the 'Expected WAS_MISSING' column seen above. Then I did the Chi-squared test comparing the 2 columns above which yielded a p-value of ~0, suggesting high statistical significance that these missing values are indeed out of the ordinary. With this hypothesis confirmed, we can now move onto assessing the impact from these issues.

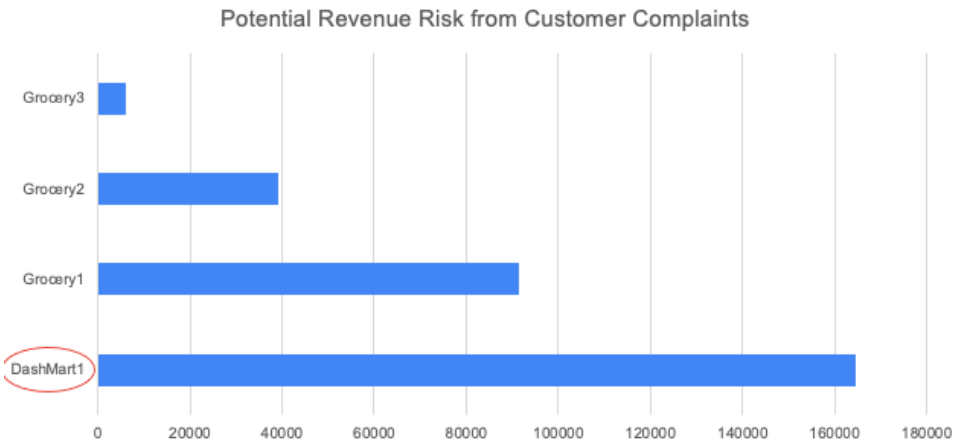
In the figure above, we see the amount of revenue lost due to the number of items marked as missing, clearly in the lead, Grocery 1 has the most revenue lost at $13,506 which is more than 2 times the 2nd highest revenue loss of Grocery 2 for missing items. This signifies the potential savings that could be recouped if DoorDash were to more accurately house inventory in Grocery1 in particular.



For more insight into the types of items that were missing the most, the categories: Pantry, Frozen, and Drinks made up the bulk of the categories of orders that went missing, at 43%. These would be areas of inventory to focus on to recoup Grocery1's losses.



Additionally, Grocery 1 has the above order volume for various times throughout the day. Looking at the timeseries graph above, there's a morning peak around 6AM, followed by a 1PM afternoon peak that is sustained until 4PM, where it falls off quickly for the evening. While this only considers data for the pas ~30 days, looking at this graph would suggest to merchants at Grocery 1 that stocking inventory should be done in evenings the night before the early morning rush hour commute, and in mornings shortly before the lunch break period.



When looking at the potential loss for DashMart1 given the high amount of customer complaints (34,349 instances), we see that when multiplied by the average item cost for this store, yields a potential $164,504 at risk of being incorrect orders. This is an important metric to keep in mind, as the labeling issues with these items in this store that we postulated might be occurring, could results in a large amount of customer churn and revenue reduction given the amount of money associated with customer complaints.

1d) Conclusion:

The Door Dash orders made in this data set helped us find that DashMart1 and Grocery 1 have the most areas to improve on in their logistics management. For DashMart 1, we found the highest amount of customer complaints, with the highest times in the overall delivery process, indicating that the labeling of items in this store could be problematic, misleading dashers to think they are correctly getting items when they are incorrectly listed. For Grocery 1, we found that there were higher 'missing' and 'subbed' items, meaning that items that were desired where not available in this store, potentially increasing customer dissatisfaction and churn. This would be a physical supply chain issue, where Grocery 1 merchants need to keep better inventory levels of its product, specifically in the Snacks, Drinks, and Candy categories.

Recommendation:

I would recommend the following action items in the near-term and long-term. For the near-term, DashMart1 should increase their labeling accuracy by running quality checks for labeling accuracy on their items, therefore decreasing substitutions. This is also a high impact category due to the ~$160,000+ in item cost that is potentially being wasted. In the long-term, I would have Grocery 1 continue to

monitor the stock of their items, particularly in snacks, drinks, and candy, to create a forecast with the historical data. This would be a high impact and high effort category due to the length of time required to build a database of historical information.

There were a lot of areas for further exploration in this data set that I could pursue, such as finding out which specific dashers had the lowest performance in deliveries, what specific food items had the highest customer reported issues, and what patterns where there on the types of substitute items to prioritize certain inventory. However, my goal here was to find the most impactful insights based on how severe issues were for the grocery store and dasher categories. Finding the highest drivers of inefficiencies in these categories and investigating the reasons behind those occurrences was my primary objective.

Part 2: SQL Question Answered

1. During the month of August, how many deliveries did store 123 receive? What was the total revenue from these deliveries?

```
SELECT store_id, COUNT(delivery_id) AS total_deliveries, SUM(order_value) AS total_revenue

FROM DeliveryData

WHERE date BETWEEN '2019-08-01' AND '2019-08-31'

AND store_id = '123'

GROUP BY store_id;
```

2. How many stores are under business id 890? Of these stores, how many of them are on DashPass?

```
SELECT COUNT(store_id) AS total_stores,

    SUM(CASE WHEN dashpass = 'Yes' THEN 1 ELSE 0 END) AS dashpass_stores
```

```
FROM BusinessData
WHERE business_id = '890';
```

3. How many stores do Kevin and Carla manage?

```
SELECT aobm.account_owner, COUNT(DISTINCT bd.store_id) AS total_stores
FROM AccountOwner_Business_Mapping AS aobm
LEFT JOIN BusinessData AS bd
ON aobm.business_id = bd.business_id
WHERE account_owner IN ('Kevin', 'Carla')
GROUP BY account_owner;
```

4. Show monthly sales for stores that Kevin and Carla manage, including a running total of cumulative sales.

```
SELECT
  STRFTIME('%Y-%m', dd.date) AS month,
  aobm.account_owner,
  SUM(dd.order_value) AS monthly_sales,
  SUM(SUM(dd.order_value)) OVER (PARTITION BY aobm.account_owner ORDER BY
  STRFTIME('%Y-%m', dd.date)) AS running_total
FROM AccountOwner_Business_Mapping AS aobm
LEFT JOIN BusinessData AS bd
ON aobm.business_id = bd.business_id
LEFT JOIN DeliveryData AS dd
ON bd.store_id = dd.store_id
WHERE aobm.account_owner IN ('Kevin', 'Carla')
```

```
GROUP BY month, aobm.account_owner
ORDER BY month, aobm.account_owner;
```