YipitData Associate Data Consultant - Take Home Exercise 2
March 31, 2022

This document outlines the workflow I used in order to arrive at my estimates for Groupon's 4Q13 Performance.

Exploring the Data

Prior to beginning any kind of leg work on this assessment, I first decided to try and understand the provided datasets at a high level.

Considering the fact that I am tasked with predicting the estimated total billings for Groupon in the 4th Quarter of 2013, I uploaded the datasets to a Jupyter environment.

Upon ingesting the 4Q13 raw dataset and exploring the data, I found out that:

- The dataset pertains to the performance of deals offered by Groupon. In this specific case, the information is restricted to Groupon deals active in the 4th quarter of 2013.
- There is a total of 138,534 samples (rows) and 7 features (columns).
- Each row in the data is uniquely identified by the Deal ID and Deal URL columns. In other words, each row of the data contains one row per deal.
- The Deals in the dataset range from November 11, 2021, to December 31, 2013.
- A quick check for missing values in the dataset yielded no results. That is, this dataset is fully free of missing values.
    - However, this is a misconception. It has been mentioned that information concerning Local Deals that went live between 10/20/2013 to 10/30/2013 is not available due to an outage. I will talk about these missing values and their treatment in a later section.
- Checking for duplicate entries in the dataset did not yield any results, either.

Segmenting the Data

Upon finishing a cursory exploration of the raw 4Q13 data, I divided the data into 3 data frames, separating them by Segment (Local/Goods/Travel).

Splitting the data in this manner allowed me to study each segment on its own, and calculate the estimates for each separately before summing them together to get the overall, final estimates for 4Q13.

After splitting the data, I derived summary statistics pertaining to each. These summary statistics are summarized in the subsequent sections. They can also be seen in the attached Jupyter notebook '*Yipit 4Q13 Groupon Data Analysis.ipynb*'.

Calculation of 4Q13 Estimates

Goods and Travel Segments

The data pertaining to the Goods and Travel segments did not have any missing or duplicated entries, **meaning the estimates for these two segments could be calculated without any prior adjustments to their data**.

The workflow used for this was pretty straightforward:

- I grouped the data by the Deal Start Date to obtain the total units sold and billings for deals that went live on a given date. Then, I calculated the total sum for each of the fields to arrive at the 4Q13 estimates for Billings and Units Sold.
    - Technical Implementation Detail: I was able to do this by performing a group-by operation on the data frame along the Start Date feature and computing the sum.
- The resulting 4Q13 estimates for the Goods and Travel Segment are summarized below:

Metric	Goods	Travel
Total Active Deals	15,234	2724
New Deals Launched	12,749	2,177
Units Sold	10,491,746	378,910
Billings	$282.5 million	$70.55 million

Local Segment

The first step in calculating the estimates for the Local segment was the same as for the Goods or Travel segments: grouping the data by the Deal Start Date to obtain the total units sold and billings for deals that went live on a given date.

However, calculating the 4Q13 estimates for the Local segment, was not as simple. Data for deals going live between 20 October 2013 and 30 October 2013 were missing, owing to a data outage. This in turn meant that these missing values needed to be handled properly before any work relating to estimation could commence.

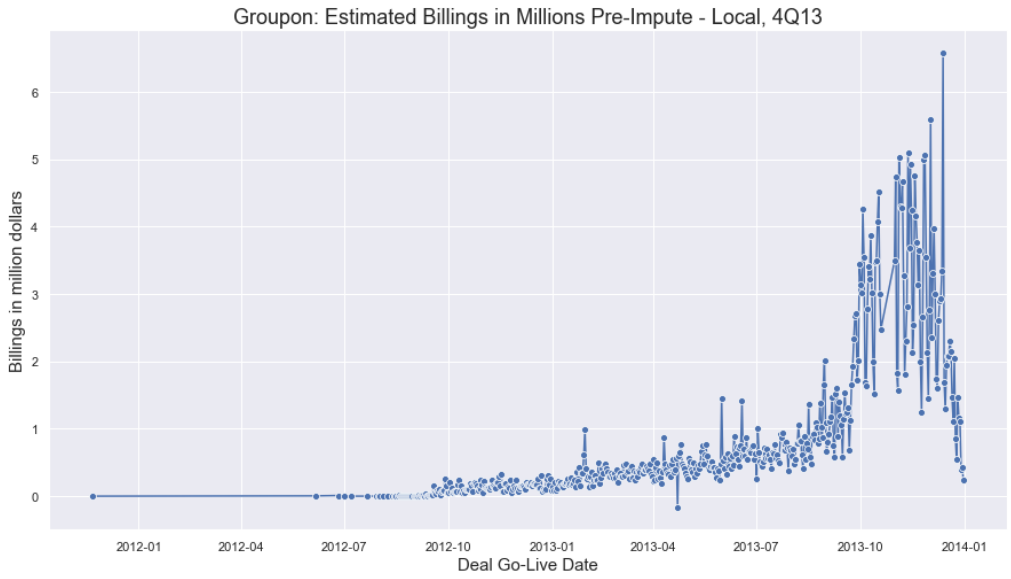
Addressing the Missing Values

The gap in data available for Local deals in 4Q13 meant:

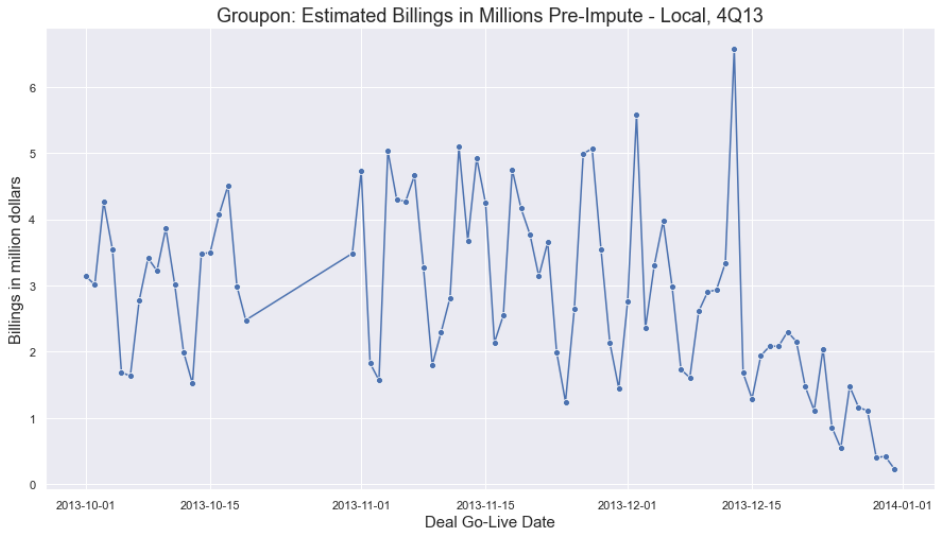
- No information regarding deals starting, units sold, or billings for this period was available. (*Note: Other information, such as that of deal IDs and deal URLs weren't available as well. The categorical nature of these variables does not influence the calculation of billing estimates, because of which their absence can be safely ignored.)*

Could I just remove these dates and proceed with calculating estimates? No. Removing this information meant removing 11 days' worth of sales data out of 90 for the highest-selling segment of Groupon's business. This would definitely distort the estimates by a considerable margin.

Removal not being a feasible option meant the only option to move forward would be through the imputation of the missing data. I decided to plot the billings by Local deal go-live date over the course of 4Q13 to deduce the best imputation approach:



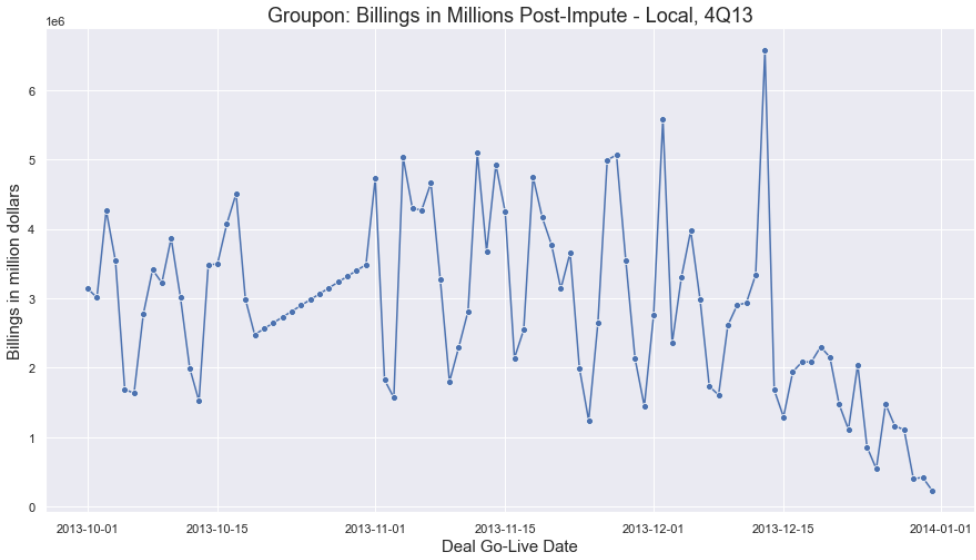
The impact of deals that went live before the beginning of 4Q13 seemed minuscule and caused a significant skew in the distribution, because of which I decided to extract data pertaining only to deals that went live in 4Q13. The plot of the latter is as shown below:



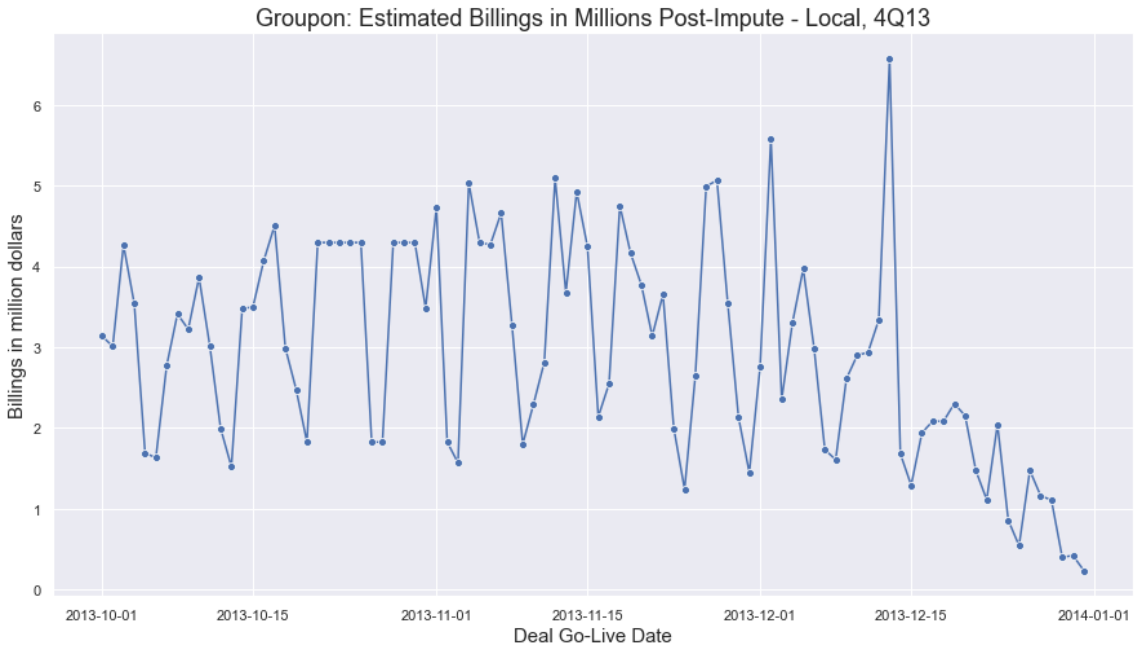
From the above plot, the disruption caused by the outage is clearly visible in the straight-line portion between 2013-10-15 and 2013-11-01 with no data markers on it. However, it also

highlights the absence of any seasonality or definite trend - meaning mean/median imputation could be the best possible imputation strategy that can be used.

However, I did try applying a linear imputation as well, though it didn't yield the result I was looking for, as the missing values were equidistantly populated along the straight line as shown below:



The median imputation seemed to yield results for the outage period similar to the rest of the quarter. Upon taking a closer look, I found that data for 8 weekdays and 3 weekend days was missing. I computed the median Units Sold and Billings for 8 weekdays (4 each from before and after the outage) and populated the missing weekday data with it. I repeated the same approach for the 3 weekend days, and below shown is the result:



While not the perfect imputation, it yielded a result that fits much better than the one obtained by linear interpolation.

I followed a similar approach to impute the number of new deals launched in the Local segment during the outage. The output of that imputation can be found in the attached Jupyter notebook titled *Yipit 4Q13 Groupon Data Analysis*.

These imputations gave me a complete dataset for the Local segment, and I just summed the Units Sold and Billings features to arrive at the below-shown estimates for the Local segment in 4Q13:

Metric	Local Segment 4Q13 Estimate
Active Deals	120,576
New Deals Launched	53,712
Units Sold	15.3 million
Billings	$449.08 million

Calculating the Final 4Q13 Estimate

Making all metrics' 4Q13 data available across all the different segments made it possible to calculate Groupon's company-wide Q4 estimates. They are summarized below:

	Active Deals	**New Deals Launched**	**Units Sold**	**Billings**
Local	120,576	53,712	15.3 million	$449.08 million
Goods	15,234	12,749	10.4 million	$282.25 million
Travel	2,724	2,177	0.38 million	$70.55 million

Total	**71,670**	**68,638**	**26.14 million**	**$801.88 million**