Groupon Assignment Report

1 Introduction

This is the report of Groupon assignment from YipitData. The target is to use 4Q13 data to estimate Groupon's 4Q13 North America gross billings by segments (i.e., Local, Travel, and Goods). Then, I use this estimate to make a buy or sell recommendation for Groupon stock.

In this report, I first observe the entire dataset and calculate the initial statistics. Then, I deep dive into the 4Q13 data and find the correlations between different variables, such as billings and units sold. Next, I use the observed data patterns and correlations to make an accurate estimate across each segment. Finally, I combine the estimate with the history data to make a a buy or sell recommendation for Groupon stock.

All the calculations in this report are done in Python3.

2 Data Observation

2.1 Initial observation

In this section, I first observe the 4Q13 dataset and calculate the basic statistics. For each segment, I compute the billings, units sold, new started deals and total number of deals. The results are presented in Table. 1. According to the problem statement, there was a data outage in Local segment. Thus, all the statistics of the Local segment are underestimated.

Table 1: Basic statistics of 4Q13 dataset for Local, Goods and Travel.

	Billings ($, million)	**Units sold**	**New started deals**	**Total deals**
Local	409.22	13,924,480.25	46,980	120,576
Goods	282.25	10,419,746.30	12,749	2,724
Travel	70.55	378,910.20	2,177	15,234
Total	762.02	24,723,136.76	61,906	138,534

2.2 Further observation

Since the 'Billings' and 'Units sold' are the only two columns that contain numerical values, I start checking from these two quantities. For each segment, I plot the graph with respect to units sold and billings (Fig. 1, Fig. 2, Fig. 3).

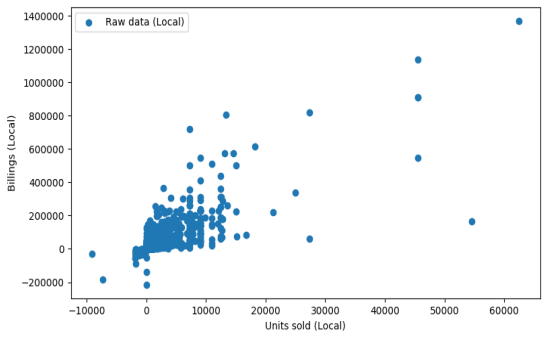


Figure 1: Units sold vs Billings (Local)

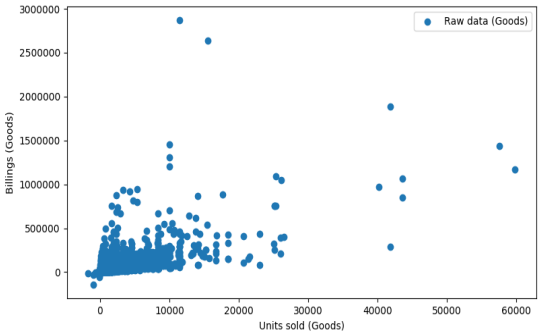


Figure 2: Units sold vs Billings (Goods)

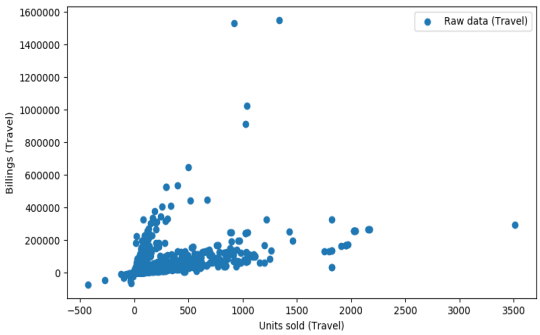


Figure 3: Units sold vs Billings (Travel)

Empirically, if we sell more units, we can earn more money. However, we cannot observe this trend in the three graphs above. Thus, I decide to group the data by Start Date. That is, for each segment, I calculate the sum of billings, units sold and number of deals with respect to each start date. The reason why I process data in this way is that the start date is another attribute of each deal, so this method can group the deals that have the same start date. Moreover, in order to fill up the Local data outage, I will focus on the Local deals in the rest of the report.

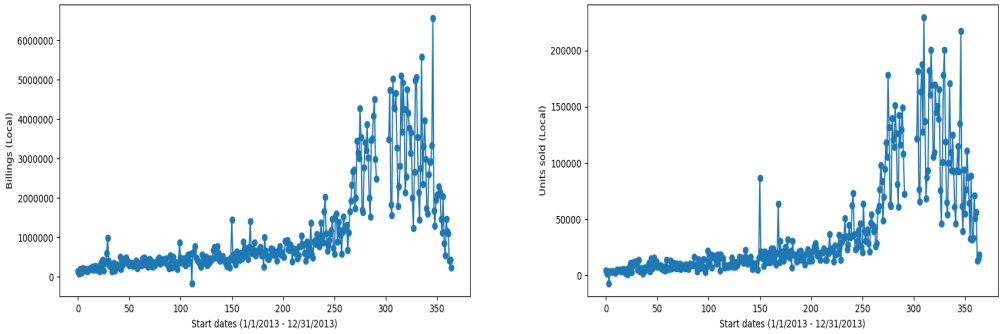
Before grouping the data, I notice that some start dates only have limited number of deals (for example, the local deals that started in August 2012). These data points may be biased. Thus, I use the local deals starting from 1/1/2013 to 12/31/2013.

The grouping results are presented in Fig. 4 and Fig. 5. There are several important data patterns found in Fig. 4:

- There exits a similar trend between the grouped billings and the grouped units sold with respect to start date.
- There is a big increase in billings and sold units when the start date gets close to Oct 1st, 2013.
- There is a decrease in billings and sold units when the start date gets close to the end of 2013.

Accordingly, we can summarize the reasons behind those patterns:

- There is a linear correlation between the grouped billings and the grouped units sold, as is shown in Fig. 5. That is, the unit price is similar among different groups of deals.
- The new started deals in Q4-2013 are more active than the previous deals.
- There is not enough time for deals whose start dates are close to the end of 2013.



(a) Grouped Local billings vs start date. (b) Grouped Local units sold vs start date.

Figure 4: Grouped Local deals with respect to start date.

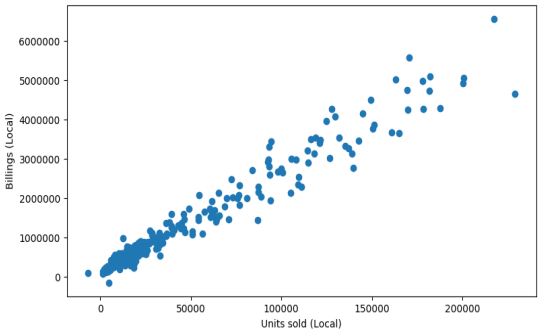


Figure 5: Units sold vs Billings for grouped Local deals.

3 Methodology and Results

3.1 Data adjustment strategy

Based on the observations in Section 2, I come up with the following work flow to adjust the Local deals:

1) Correlate the grouped billings and grouped units in Fig. 5 using Linear Regression model.

2) Fill up the missing units in Fig. 4b using high-order Polynomial Regression model.

3) Get the missing billings using the linear model in step 1 and units in step 2 .

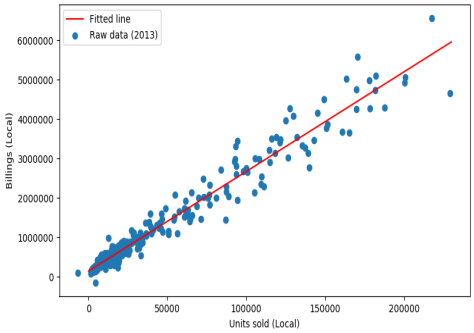
3.2 Correlate the grouped billings and the grouped units

In this section, I use Linear Regression model to correlate the grouped billings and grouped units for Local deals.

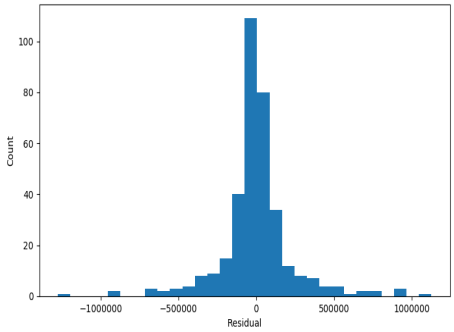
Here I use the same data points in Fig. 5. There are 354 points in total. First, I shuffle all the points to avoid bias. Then I train the Linear Regression Model using 'sklearn' package in Python3. The resulting model is given in Fig. 6a, and its mathematical expression is:

$$billings = 25.28 * units + 144066.24 \tag{1}$$

Next, I evaluate the the model with R^2 score and residuals (R^2 score is a statistical measure of how close the data are to the fitted regression line. The closer it is to 1, the better the model fits the data.). The resulting R^2 score is **0.96**, and the residual distribution is given in Fig. 6b. We can see that the residual is almost symmetrically distributed and most of the predictions are located around the zero residual. Thus, it can be concluded that the linear model is reliable.



(a) Units sold vs Billings for grouped Local deals and fitted linear model.



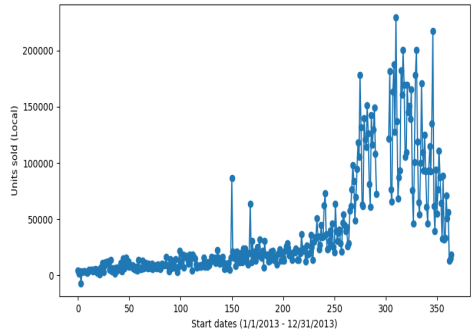
(b) Residual histogram.

Figure 6: Linear regression model and residual distribution.

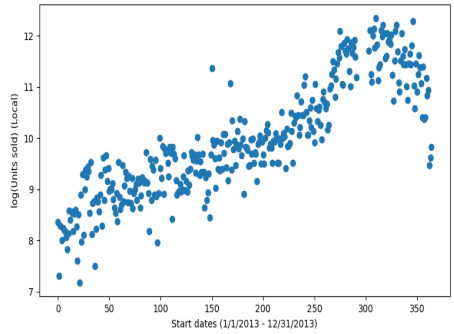
3.3 Fill up the missing units

In this section, I use Polynomial Regression model to correlate the grouped units and start dates.

Instead of build the model directly, I notice there exists noise in the grouped units data (Fig. 7a). Thus, in order to remove the noise to some extent, I transform the grouped units with a log function. As is shown in Fig. 7b, the grouped units first slowly increase with the start date, and then go through a large increase as the start date getting close to Oct 1st (2013). Finally, it drops down as the start date getting close the end of year 2013.



(a) Before transformation.



(b) After transformation.

Figure 7: Transformation of the grouped units sold (Local).

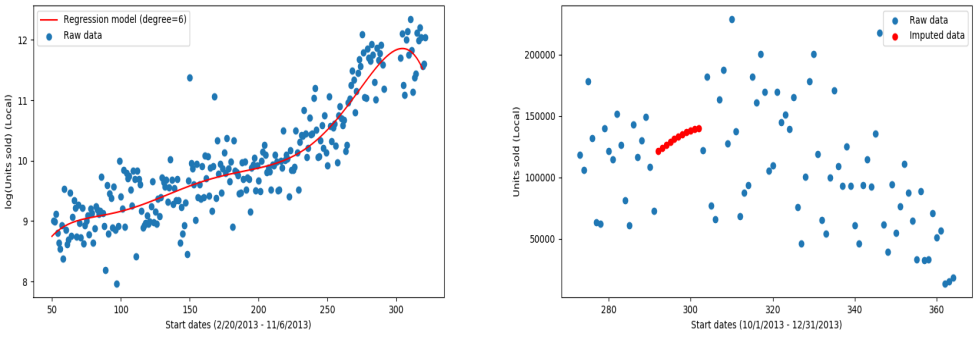
Before building the regression model, I cut off some leading points and trailing points in Fig. 7b to make the training easier. Since I only need to simulate the trend close to the gap, the points that are far from the gap may lead to inaccuracy during the training process. Thus, there are 260 points in total and I randomly split 208 points to the training set, 52 points to the test set for model evaluation.

During the training process, I build the Polynomial Regression model with several different degrees. Then I evaluate each one using RMSE (i.e., Root Mean Squared Error) and R^2 score to select the most reliable model. The results are given in Table. 2.

Table 2: Training and test results of polynomial regression model.

Degree	Training RMSE	Training R^2	Test RMSE	Test R^2
5	0.388	0.823	0.413	0.817
6	0.383	0.834	0.430	0.801
7	0.383	0.834	0.428	0.803
8	0.383	0.834	0.425	0.803

Since the model with higher degree will cause overfitting, the model with degree 6 is good enough to simulate the trend. The resulting model is presented in Fig. 8a and the imputed units data in 4Q13 is given in Fig. 8b. We can see that the polynomial model can adjust the missing data appropriately.



(a) Polynomial regression model. (b) Imputed data.

Figure 8: Polynomial regression model and imputed data (Local).

3.4 Adjustment for missing deals

In this section, I adjust the missing billings and the number of new started deals for the Local segment.

Since we have imputed the missing units, we can adjust the missing billings using the linear model in Section 3.2. Fig. 9 presents the grouped Local billings in 4Q13 after adjustment. **The sum of the imputed Local billings is $ 38.34 million.**

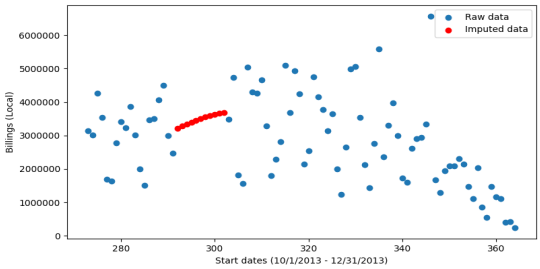
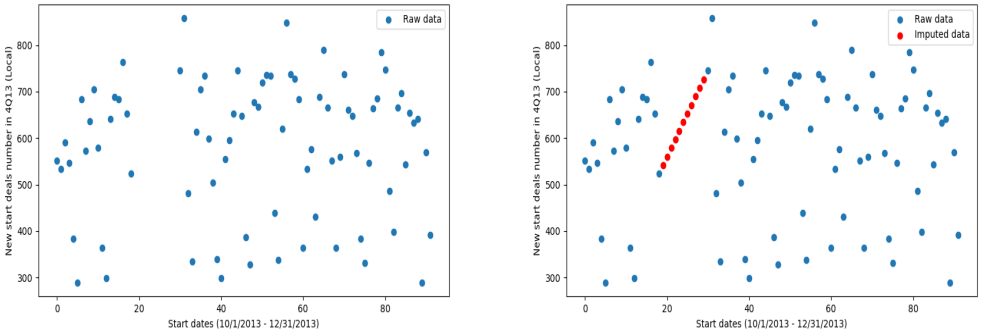


Figure 9: Imputed billings for grouped Local deals.

Next, I adjust the number of new started deals in 4Q13. Since there is no obvious trend between new deals and start date (Fig. 10a), I implement linear interpolation to fill up the gap. Fig. 10b presents the results after adjustment. The sum of the imputed Local new deals is 6,974.



(a) New deals in 4Q13. (b) Imputed new deals.

Figure 10: The number of new deals in 4Q13.

In summary, the gross billings estimates of Groupon's 4Q13 North America are $ 447.56 million (Local), $ 282.25 million (Goods), $ 70.55 million (Travel). Other estimates, such as units sold and new started deals, are presented in Table. 3.

Table 3: Statistics of 4Q13 dataset for Local, Goods and Travel after adjustment.

	Billings ($, million)	**Units sold**	**New started deals**
Local	447.56	15,378,408.78	53,954
Goods	282.25	10,419,746.30	12,749
Travel	70.55	378,910.20	2,177
Total	800.36	26,177,065.29	68,880

4 Recommendation

4.1 History data observation

Based on the estimates in Section 3, I combine them with the history data. Fig. 11, 12, 13 present the history of billings, units sold and new started deals with respect to Local, Goods and Travel segment, respectively. Fig. 14 shows the total history of each measurement.

For the Local segment, both billings and units sold go through a rise-and-fall process before 4Q13. However, there is a recovery in the latest quarter. Moreover, the new started deals undergo a big increase in 4Q13, compared to the previous five quarters. One possible reason is that fourth quarter is the holiday season and Groupon made a big sales promotion during that period. Another possible reason that can explain this fluctuation is seasonality.

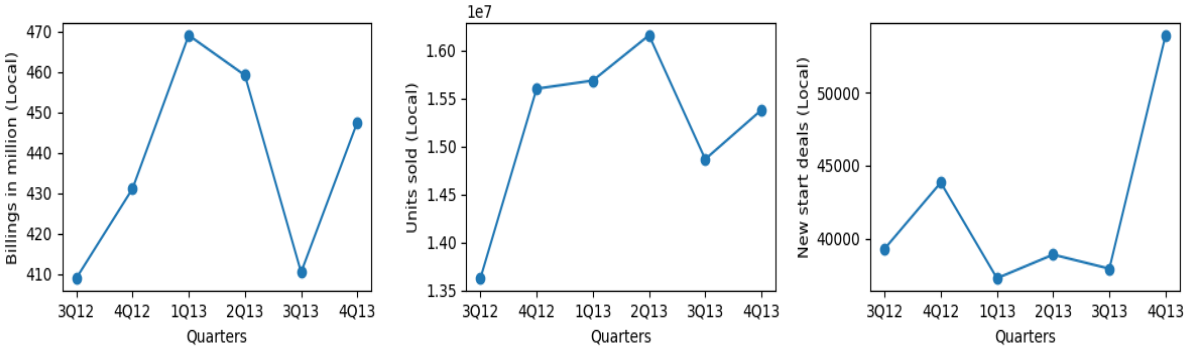


Figure 11: Local deals history.

For the Goods segment, all the three quantities go through a little fluctuation in the previous quarters, but the overall trend is going upward. This reveals that Groupon is putting more emphasis on the Goods segment, including improving service quality, developing advanced technology on their shopping platform, etc. As a result, customers and merchants become more active in this segment.

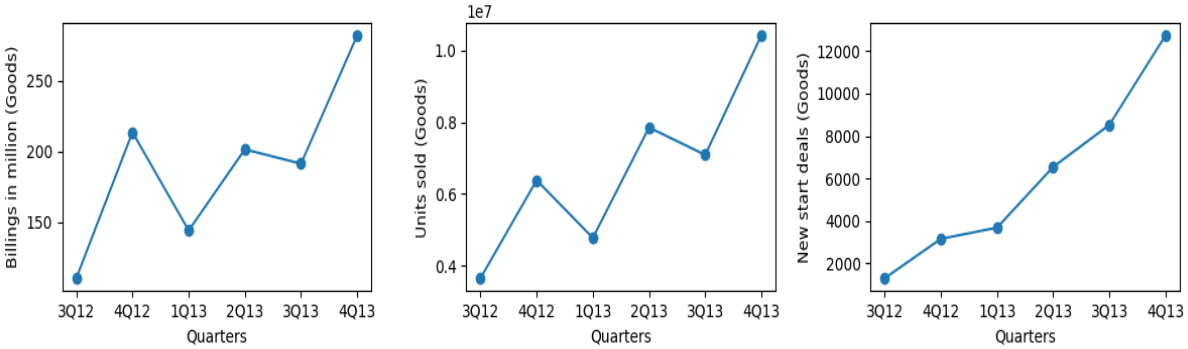


Figure 12: Goods deals history.

For the Travel segment, the overall trend of all the quantities is also rising, although there is a peak in units sold in 2Q13. This reveals that there is an increasing need for traveling and more customers are becoming active on the Groupon's Travel segment, such as reserving hotels, booking airline tickets, etc.

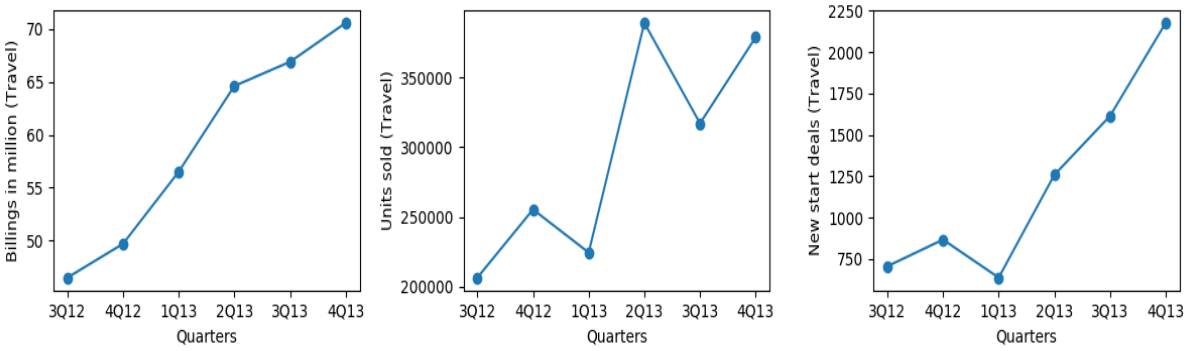


Figure 13: Travel deals history.

The total quantities of all the three segments show that Groupon is still a rising company and it has a potential to reach higher gross billings in the future quarters.

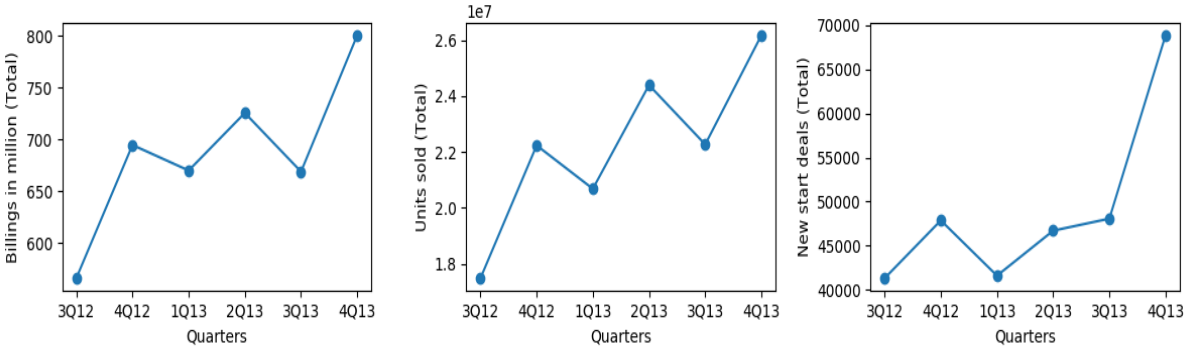


Figure 14: Total deals history.

Lastly, I present a quarter-to-quarter growth report of Groupon's history data (Table. 4). Since there is only a limited number of data, I conduct the comparison between 3Q12 / 3Q13 and 4Q12 / 4Q13. We can observe that there is a positive growth in all the quantities of Goods and Travel segment, especially the new started deals. This is consistent with the analysis above. As for the Local segment, although there is a little drop in sold units (4Q12 / 4Q13) and new started deals (3Q12 / 3Q13), the grossing billings is still in a rising trend.

Table 4: Quarter-to-Quarter growth of Groupon's history data.

	Billings		Units sold		New started deals	
	3Q12 / 3Q13	4Q12 / 4Q13	3Q12 / 3Q13	4Q12 / 4Q13	3Q12 / 3Q13	4Q12 / 4Q13
Local	0.36%	3.82%	9.09%	-1.42%	-3.47%	23.09%
Goods	73.23%	32.07%	94.26%	63.35%	552.63%	303.57%
Travel	43.80%	41.95%	53.58%	48.36%	128.93%	151.38%
Total	18.16%	15.24%	27.39%	17.73%	16.38%	43.92%

4.2 Recommendation

In this section, I first compare my gross billings estimate with the three equity research reports. As is shown in Table. 5, my estimates are closest to Deutsche Bank. Thus, my estimates should be reliable.

Considering all the analysis above, it can be concluded that Groupon is still seeing an accelerating growth across all the three segments and it is on the path to enhance its scale in the mobile commerce market. Therefore, **my final recommendation is BUY**.

Table 5: Comparison of Groupon's 4Q13 gross billings estimates (North America).

$, in million	My estimates	Deutsche Bank	J.P.Morgan	Morgan Stanley
Local	447.56	NA	490.14	508
Goods	282.25	NA	275.72	295
Travel	70.55	NA	71.02	67
Total	800.36	803.2	836.88	870

5 Code Instruction

All code scripts can be reviewed at: **https://github.com/iceljc/Groupon-exercise**.

- 'utils.py': This file contains three functions that read 4Q13 data and history data.
- 'basic.py': This file calculates the basic statistics of the raw 4Q13 dataset.
- 'correlate.py': This file correlates the billings and units sold of the grouped Local deals.
- 'impute.py': This file uses polynomial regression to impute the missing data.
- 'impute_deal.py': This file imputes the number of new started Local deals.
- 'history.py': This file summarizes the history data as well as my estimates.