Zappos.com Data Challenge

Dawei Geng

1. Preprocess the dataset

- **Remove rows with column “new_customer” is null;**
  - Reason: these visits are not from real users, but from bots, search engines, staff who are testing the website, etc since they are generating a great number of visits, but end up with no order, no sales at all, bounce rate is higher than normal.
  - Number of rows down to 12802 rows from 21060
- **Remove rows with visits being 0;**
  - Reason: there is no point to have other data without a single visit
  - Number of rows : 10733 rows
- **Remove rows with order being 0, but gross sales being nonzero;**
  - Reason: not make sense to get revenue with no order.
  - Number of rows: 9355 rows
- **Remove Site: Botly, Tabular, Widgetry**
  - Reason: it is impossible in reality to have order every visit and to have users finish the orders after adding to cart (bounce rate=0, conversion_rate=1, add_to_cart_rate=1)

2. Constructed KPI(Key Performance Indicator)

- **Content effectiveness:**
  - Average Page Views Per visit: average number of pages that a visitor views while on your web site
  - Percent of Returning Visitors: the number of returning users divided by the total number of users
  - Page Bounce Rate: the percentage of users who leave immediately after viewing the page
  - Average Searches per Visit: This is the average amount of times that a user uses the search function on your web site.

- **Marketing effectiveness:**
  - Percent Revenue from New Visitors vs Returning Visitors
  - Percent Orders from new Visitors vs Returning Visitors
  - Conversion Rate: Orders/Visits

- **Shopping Cart KPI:**
  - Add To Cart Rate: the number of add to carts by users divided by the number of

visits

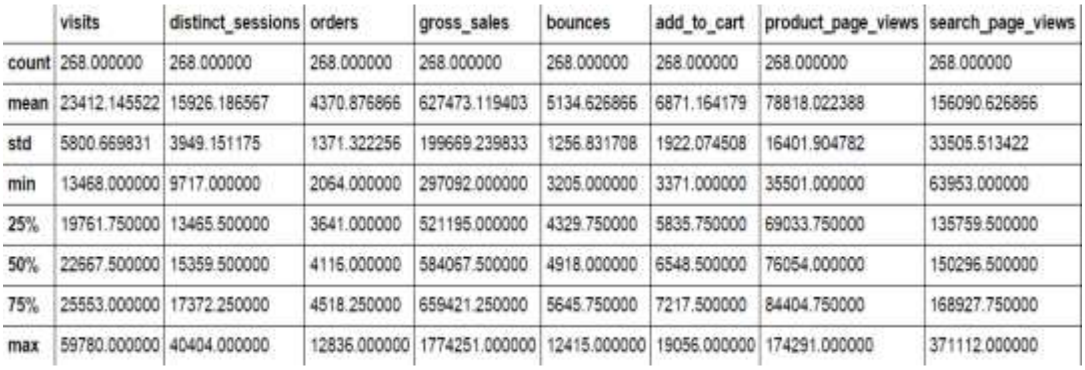
- Cart Finished Rate: the number of orders divided by the number of add to cart by users

Trick：No data between March 2013 and May 2013, between June 10 2013 and June 14 2013.

3. DashBoard

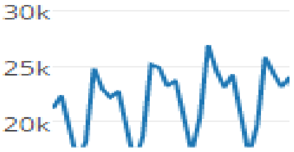
For the Challenge, I build a KPI dashboard using Plotly, a visualization package in Python, you can see the dashboard via this link:
https://dashboards.ly/ua-VEebVnakHP7BZf2c36MtWF

- **Findings from the dashboard:**
  - "Descriptive Statistics for Acme Data" plot is a boxplot for 8 features after grouping data by day for Acme site(since 95% of the visits come from Acme). We could see the distribution of each feature through this plot. We can also see that from table below. From gross sale's boxplot, we see there are some outliers above the box, this make sense because on special sales date like Black Friday, high revenue is understandable.



	visits	distinct_sessions	orders	gross_sales	bounces	add_to_cart	product_page_views	search_page_views
count	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000
mean	23412.145522	15926.186567	4370.876866	627473.119403	5134.626866	6871.164179	78818.022388	156090.626866
std	5800.669831	3949.151175	1371.322256	199669.239833	1256.831708	1922.074508	16401.904782	33505.513422
min	13468.000000	9717.000000	2064.000000	297092.000000	3205.000000	3371.000000	35501.000000	63953.000000
25%	19761.750000	13465.500000	3641.000000	521195.000000	4329.750000	5835.750000	69033.750000	135759.500000
50%	22667.500000	15359.500000	4116.000000	584067.500000	4918.000000	6548.500000	76054.000000	150296.500000
75%	25553.000000	17372.250000	4518.250000	659421.250000	5645.750000	7217.500000	84404.750000	168927.750000
max	59780.000000	40404.000000	12836.000000	1774251.000000	12415.000000	19056.000000	174291.000000	371112.000000

- From the time series plot for Visits, Users, Orders and Sales data, we could see weekly patterns for the data, but for late November and December, because of holiday season, there is an obvious peak in the plot.



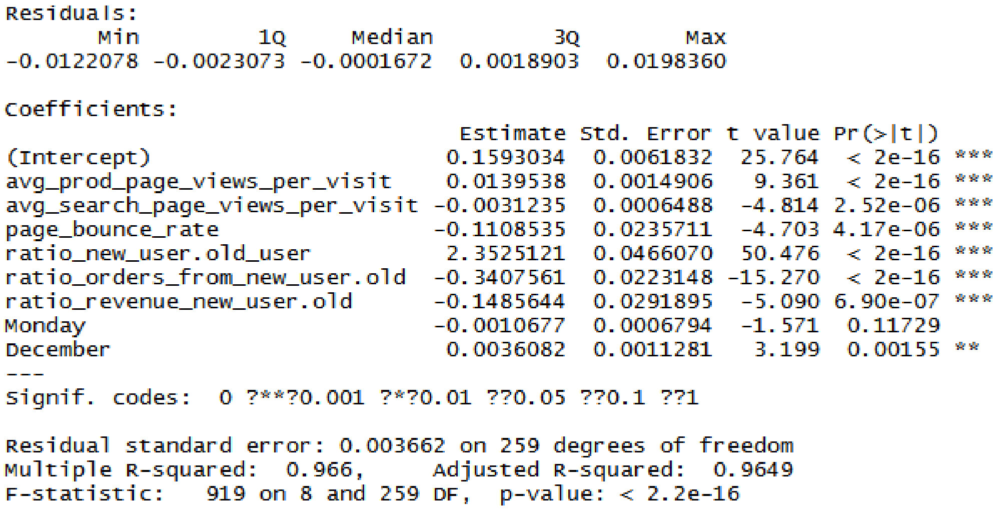
- In terms of Content effectiveness, Acme as my benchmark has highest product page views, highest search page views, lowest bounce rate, meaning customers are intended to visit Acme and they could find products that attracts them after searching just two times on average. For the other two sites, bounce rate is high, search activity is low, which also indicate that these two sites are not popular.
- In terms of Marketing effectiveness, Acme is still the best, customers tend to buy when they visit, and revenue mainly comes from old customers.

- ➢ For shopping cart KPI, still Acme’s customer are more likely to add products to cart every visit and more likely to pay for it after adding to cart.
- ➢ Anomaly: From the above three KPI plots, we do see some strange peaks and trough, such as in late September, cart_finished rate plummet for Sortly. These are the anomaly worth to research on.
- ➢ From month effect, from July to December, we can see upward trend for sales, and out of which users on windows, mac, ios contribute the most. Advertisement would be best to post on these three systems
- ➢ For weekly effect, Sunday, Monday and Tuesday contributes most to the sales. It is best to have sales on these days.

4. Modeling

Because of time limit, here I will only focus on two problems that essential to the business(Acme data are used):

- ➢ **What factors are influencing conversion rate?**
  - ➢ Regression is used for this task:
    - ✓ Dependent variable: Conversion rate
    - ✓ I scan through the features to see which one make more sense to be included. avg_product_page_views_per_visit,avg_search_page_views_per_visit,page_bounce_rate are good metrics since they reflect content effectiveness which is definitely a reason customers make their orders. Ratio of new user by old user is also a good metric because old users are more likely to order every visit. Monday and December are dummy variable I used to capture calendar effect, but add_to_cart_rate, cart_finished_rate are similar to conversion rate because they both reflect customers ‘ desire to shop
    - ✓ Regression:



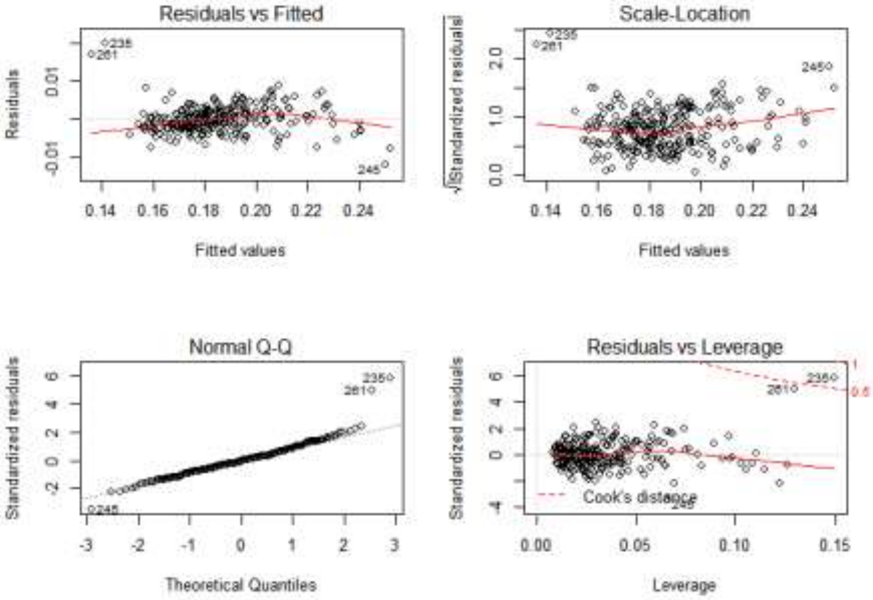
```
Residuals:
      Min         1Q     Median         3Q        Max 
-0.0122078 -0.0023073 -0.0001672  0.0018903  0.0198360 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   0.1593034  0.0061832  25.764  < 2e-16 ***
avg_prod_page_views_per_visit  0.0139538  0.0014906   9.361  < 2e-16 ***
avg_search_page_views_per_visit -0.0031235  0.0006488  -4.814 2.52e-06 ***
page_bounce_rate             -0.1108535  0.0235711  -4.703 4.17e-06 ***
ratio_new_user.old_user       2.3525121  0.0466070  50.476  < 2e-16 ***
ratio_orders_from_new_user.old -0.3407561  0.0223148 -15.270  < 2e-16 ***
ratio_revenue_new_user.old   -0.1485644  0.0291895  -5.090 6.90e-07 ***
Monday                       -0.0010677  0.0006794  -1.571  0.11729    
December                      0.0036082  0.0011281   3.199  0.00155 ** 
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.003662 on 259 degrees of freedom
Multiple R-squared:  0.966,	Adjusted R-squared:  0.9649 
F-statistic:   919 on 8 and 259 DF,  p-value: < 2.2e-16
```

All the variables except Monday are significant, Adjusted R^2 is 96.5%, which is pretty good. All signs are making sense.

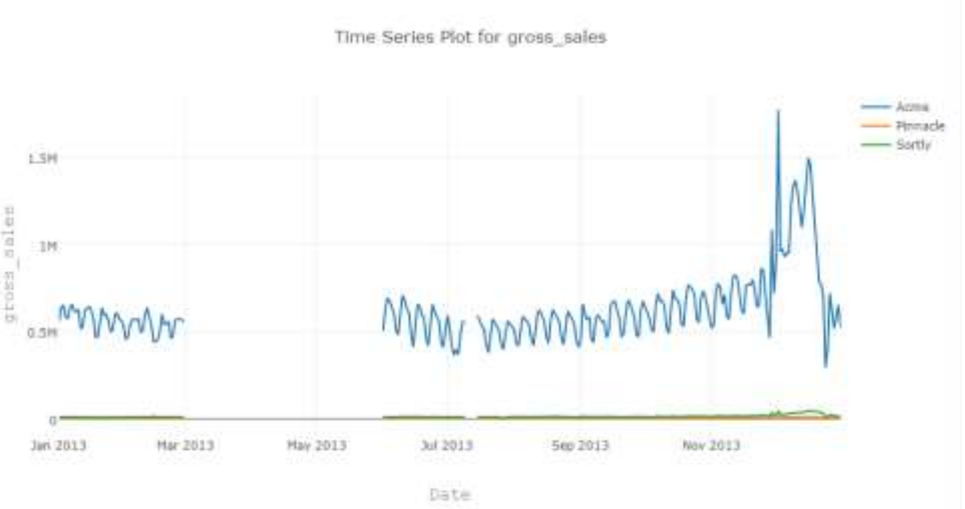
    - ✓ Diagnostic:



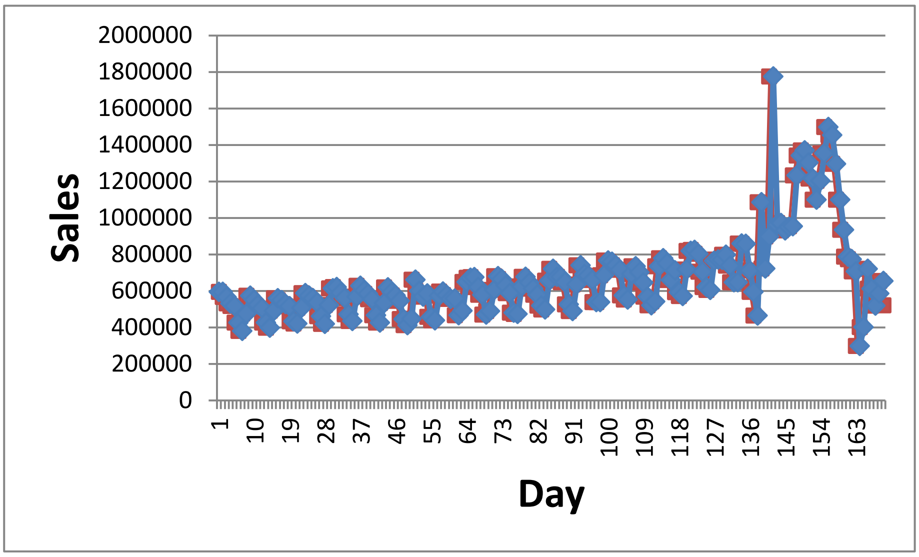
We can see model satisfies its assumption: normality, homoscedasticity.

- ✓ We solve this problem and find all variables except Monday significant, which could explain 96% of the variability of conversion_rate

➢ Can we predict the sales and how accurate could we reach?

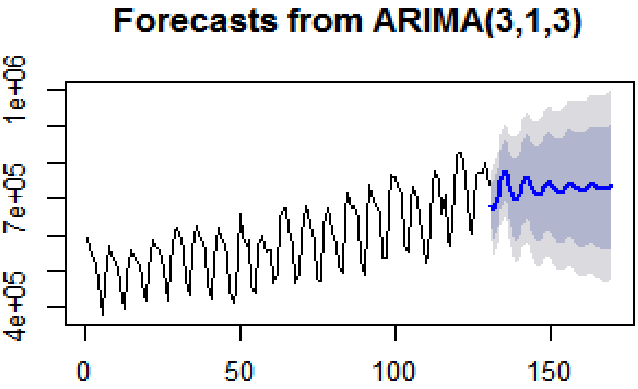


- ✓ Notice that some parts of the time series are missing and it is better to just use the data since July 15.
- ✓ Double Exponential Smoothing is used

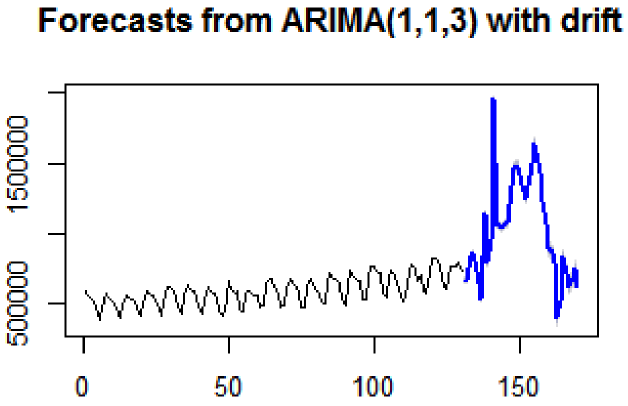


If I use Day 1(July 15) to 100(July 15+100) as training set, 101 (July 15+101) to 138 (July 15+138) as holdout set, training MAE is 55000, holdout MAE is 62594, which is not bad, but if I use 138 to 170 as holdout set, things are quite different since December in the data is more like an anomaly, in this case, holdout MAE is 179348.

- ✓ ARMA could also be used for the first part of the data with regular patterns, but won't work well with December data.



- ✓ Since order definitely has positive correlation with sales, we could run regression between these two variables, and get residual, which are much more stable, then we could run ARMA to get better prediction.



Using this method, holdout MAE is 98384, which is significantly lower than double exponential method and simple ARIMA.