EDA\_team1

Bryton Levalley, Deac Long, Emma Johnson, Jessican Hosler, Minju Lee

3/25/2020

# Team1 EDA

Our team will be looking at fictitious customer data provided by our mentor from SEL company. Missing data and outliers are dealt with in excel. Steps: 1. Import data using CSV 2. change customer keys in each table to match all across the tables. 3. Join necessary tables

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

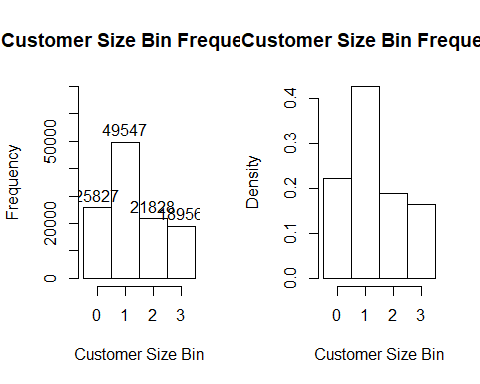
Summaries of each table. We can see that there are some missing data in table\_d and table\_e. We will omit those, since the amount of NA values are insignificant. We also see the table\_d contains 6 tuples with year 1916, this we will also exclude these values for analysis purpose.

Created new joined tables for analysis purpose.

# Data Overview & Visualization

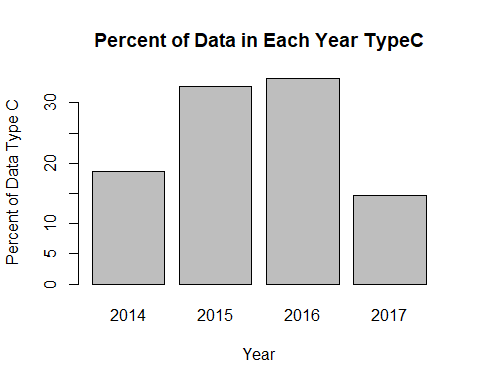
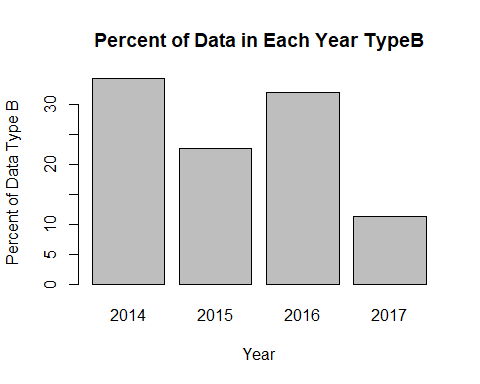
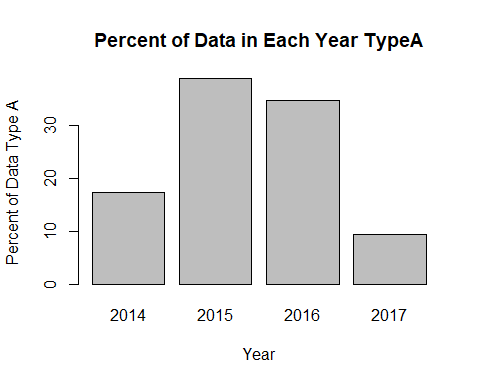
### 1. Customer\_info

The histogram shows frequency and density of customer size bin in Customer\_Info table. there are customer size 0,1,2 and 3. Most customers belong to customer size 1 and the least amount of customers belong to customer size 3.



### 2. Frequench of Yearly Data by each Type

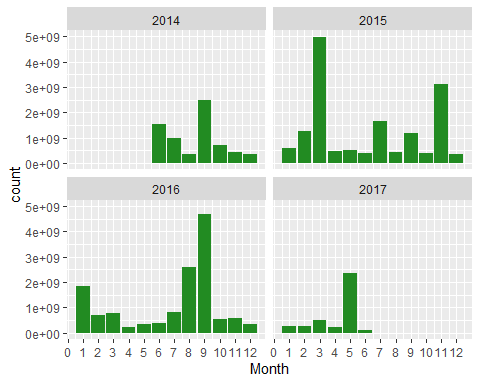
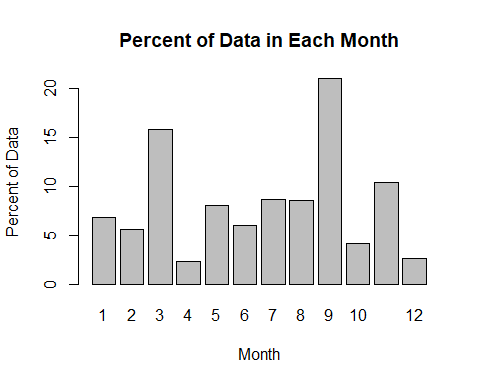
The given data starts mid 2014 and ends mid 2017.



Type A data contains most data in 2015 and 2016. Type B data contains most data in 2014 and 2016. Type C data contains most data in 2015 and 2016.

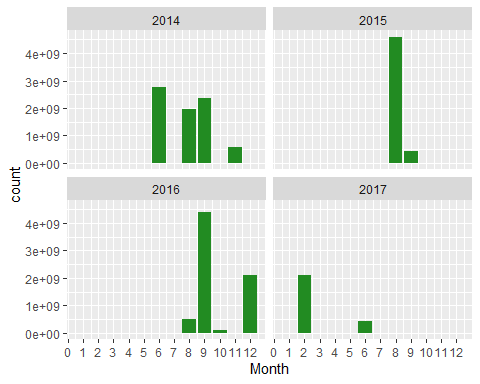
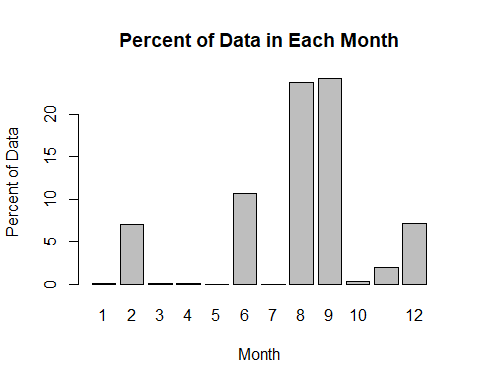
question: Type B does not have similar trend with Type A and Type C. What can be the reason for this? We also know that we only contain half year of data in 2014, however Type B contains most data in 2014.

### 3. Frequench of Monthly Data by each Type.



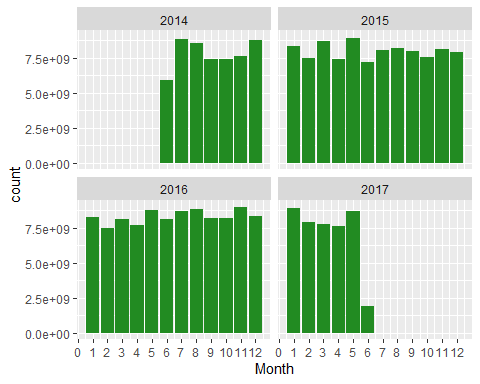
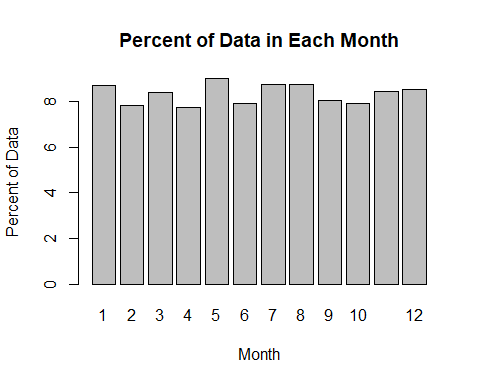
First Bar plot shows percent of Overall Monthly Type A data. Month 9 has the most TypeA data follwed by month3 and month 11

The Second bar plots show monthly typeA data by each year. As I menthiod earlier we can see that 2016 only contains data from May to December and 2017 contains data from January to Jun.



First Bar plot shows percent of Overall Monthly Type B data. Month 9 has the most Type B data follwed by month8 and month 6. Some interestng fouding on Type B is that there are very small amount of TypeB data in month 1,3,4,5,7,and 10

The Second bar plot shows monthly typeB data by each year.

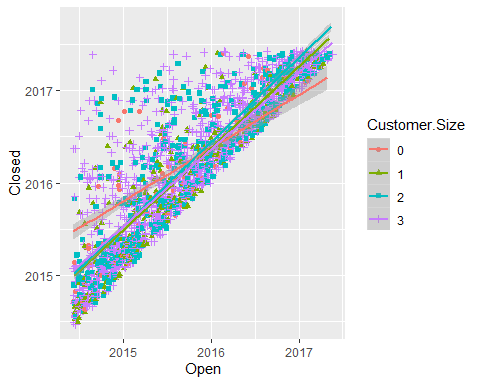


First Bar plot shows percent of Overall Monthly Type C data. Type C data is evenly distributed through out the months.

The Second bar plot shows monthly typeB data by each year. The data is evenly distributed monthly in each year.

### 4. Open\_closed table

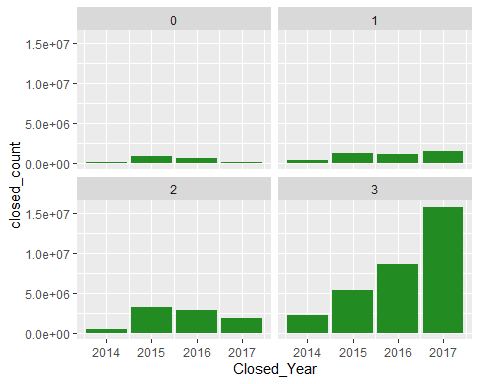
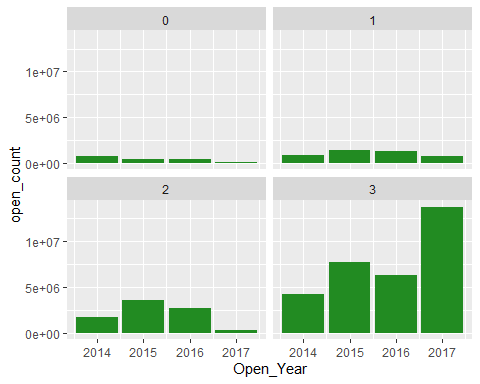
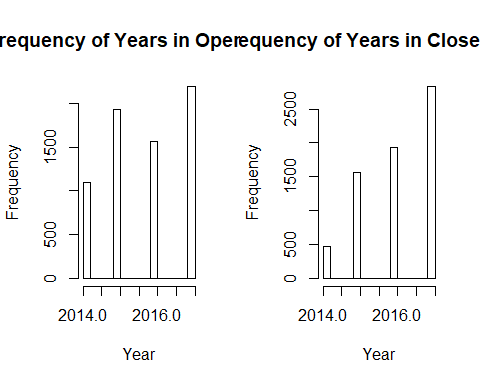
* Overall Trend



The plot above shows that most of the data are concentrated along the line where Open and Closed date is not too apart from each other. However, we have data where closed date is far apart from the open date.

The best fit lines for each Customer.size on its open and closed data is displayed. Customer.Size 0 stands out from the rest. It indicates that around mid 2014 to late 2015 data that belongs to Customer.Size 0 were taking longer to close on average compared to any other Customer.Size and from late 2015 to mid 2017 Customer.Size 0 were closed early on average compared to Others.

* Yearly Trends



First bar graphs show overall open data and closed data by year. When we look at the closed data, it is increasing over time. This can indicate that the company is closing more open data as time goes.

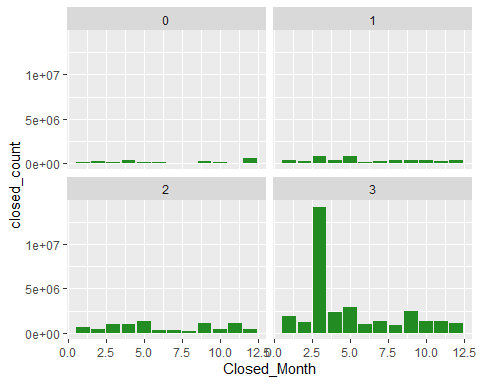
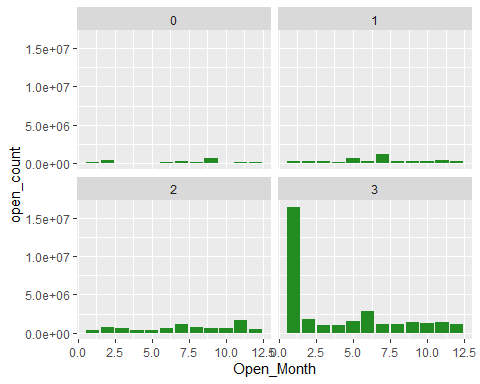
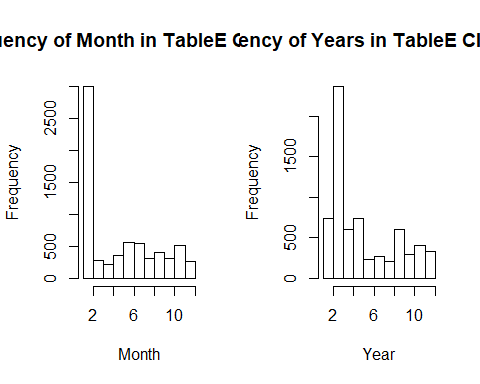
Second bar graphs show open data counts by year on each Customer.Size.

Third bar graphs show closed data counts by year on each Customer.Size.

We can also see that Customer.Size 3 contains a lot more data compared to other Customer.Sizes. However, Customer.Size 3 contains the least amount of Customers compared to others. This can indicate that Customer.Size 3 contains most frequent customers

Possible conclusion on Customer.Size attribute is that it is orderd from least frequent customers to most frequent customers (0 being the least frequent and 3 being the most frequent).

* Monthly Trends



First bar charts show that most data was opened on January and was closed on February. Secone bar charts show open data by Month on each Customer.Size. Secone bar charts show closed data by Month on each Customer.Size.

### 5. More on Type C data

The table below shows number of requested, early,ontime and late data for each Customer Size.

## Customer.Size requested early ontime late  
## 1 0 843276 411323.0 351695.0 80259  
## 2 1 1445088 425490.3 202318.5 817290  
## 3 2 400002 246184.0 112967.0 40851  
## 4 3 1319599 820698.0 340178.0 158724

The Tables blow shows number of early, ontime. late data for each Customer Size. Customer Size 0,2, and 3 shows downward trend indicating most of their requested data is done early. Customer Size 1 has upward trend indicating most of their requested data is done late.

