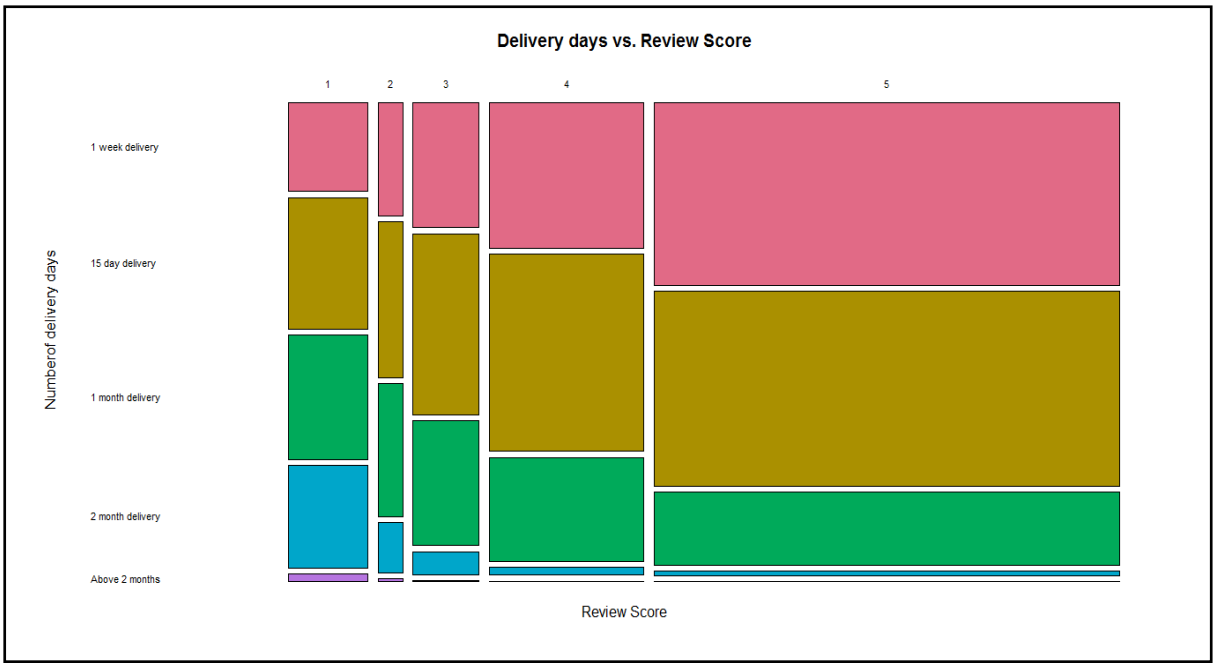
Part 1

The ecommerce dataset has been divided into 5 different groups.

- 1 week delivery
- 15 day delivery
- 1 month delivery
- 2 month delivery
- Above 2 months

The customer has provided reviews on different products on the scale of 1 to 5.
The below table shows the data distribution of the delivery groups against the review scores.

```
> Orders_merged$delivery_days_groups <- ifelse(as.numeric(Orders_merged$`d
elivery days`) <= 7, '1 week delivery', ifelse(as.numeric(Orders_merged$`d
elivery days`) <= 15, '15 day delivery', ifelse(as.numeric(Orders_merged$`
delivery days`) <= 30, '1 month delivery', ifelse(as.numeric(Orders_merged
$`delivery days`) <= 60, '2 month delivery', 'Above 2 months'))))

> order_mosaic= table(Orders_merged$review_score,Orders_merged$delivery_da
ys_groups)

> order_mosaic

    15 day delivery 1 week delivery 2 month delivery 1 month delivery
  1            2794             1894              2178              2652
  2            1028              742               333               870
  3            3185             2199               406              2193
  4            8208             6063               340              4313
  5           24350            22797               573              9064

    Above 2 months
  1            175
  2             19
  3             22
  4             28
  5             44
> order_mosaic = order_mosaic[,c(2,1,4,3,5)]

> mosaicplot(order_mosaic, main = "Delivery days vs. Review Score",col = s
ubs_pal, off = 5, las = 1,xlab="Review Score",ylab="Numberof delivery days
")
```



Conclusion: The above mosaic plot shows the data distribution of the review score against the delivery groups.
The plot shows most of the customers has provided a review score of 4 and 5. Majority of these customers have received the delivery within 1 month. Very few customers giving a review score of 5 have received deliveries beyond a month from order date.
The percentage of people assigning a review score of 2 and 3 and less.
The plot clearly shows people receiving the shipment beyond a month generally give a review score of 1. There are quite a few people receiving shipment within 2 months and giving a review score of 1.

Part 2

The dataset has been divided into four different parts.

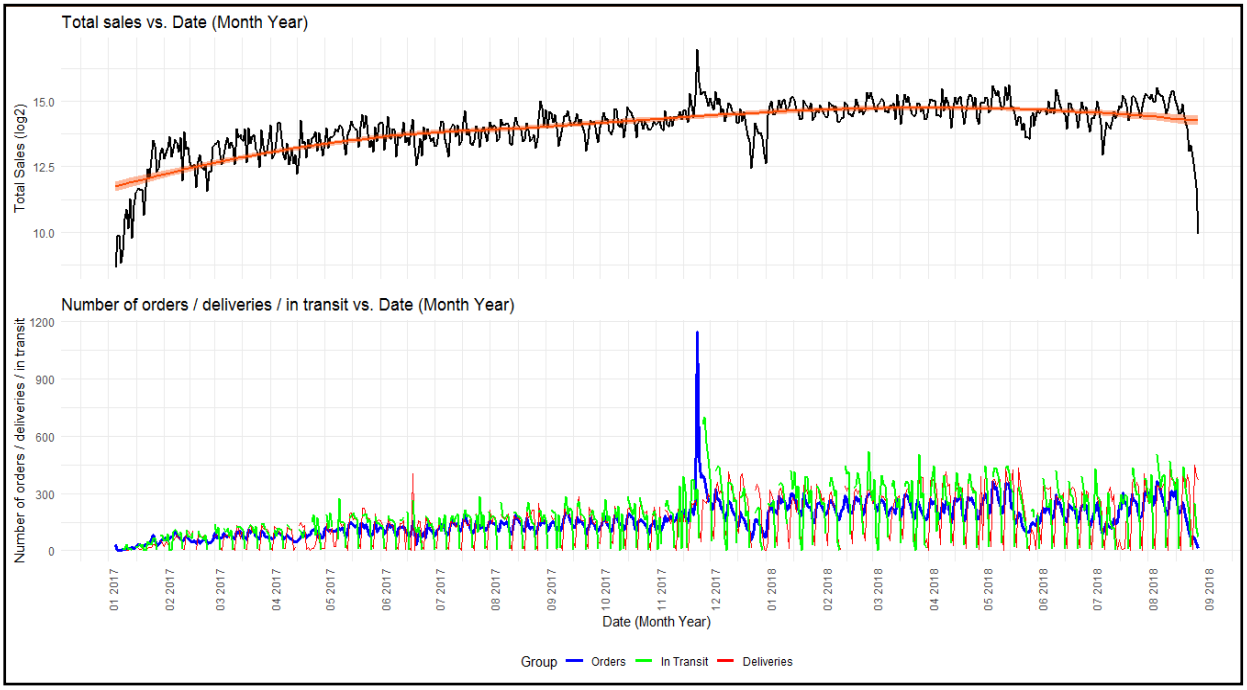
- Total sales on daily basis
- Number of orders on daily basis
- Number of deliveries on daily basis
- Number of shipment (in transit) on daily basis.

The line graph would be plotted with all the datasets sharing the same scale.

```
> Orders_grouped_number = Orders %>%
+   mutate(date_col = as.Date(order_purchase_timestamp)) %>%
+   group_by(date_col) %>%
+   summarize(value = n())

> Orders_grouped = Orders %>%
+   mutate(date_col = as.Date(order_purchase_timestamp)) %>%
+   group_by(date_col) %>%
+   summarize(value = sum(price))

> Orders_grouped_delivery = Orders %>%
+   mutate(date_col = as.Date(order_delivered_customer_date)) %>%
+   group_by(date_col) %>%
+   summarize(value = n())

> Orders_grouped_carrier = Orders %>%
```

```
+   mutate(date_col = as.Date(order_delivered_carrier_date)) %>%
+   group_by(date_col) %>%
+   summarize(value = n())

> Orders_grouped = filter(Orders_grouped, between(date_col, as.Date("2017-
01-01"), as.Date("2019-01-01")))

> Orders_grouped_number = filter(Orders_grouped_number, between(date_col,
as.Date("2017-01-01"), as.Date("2019-01-01")))

> Orders_grouped_delivery = filter(Orders_grouped_delivery,between(date_co
l, as.Date("2017-01-01"), as.Date("2019-01-01")))

> Orders_grouped_carrier = filter(Orders_grouped_carrier,between(date_col,
as.Date("2017-01-01"), as.Date("2019-01-01")))

> df <-
+   dplyr::left_join(Orders_grouped,
+                    Orders_grouped_number,
+                    by = c("date_col" = "date_col"))

> df1 <-
+   dplyr::left_join(df, Orders_grouped_delivery, by = c("date_col" = "dat
e_col"))

> df2 <-
+   dplyr::left_join(df1, Orders_grouped_carrier, by = c("date_col" = "dat
e_col"))

> df2 = df2 %>%
+   rename(
+     total_sales = value.x,
+     num_orders = value.y,
+     num_deliveries = value.x.x,
+     in_transit = value.y.y,
+   )

> p1 <-
+   ggplot(data = df2, aes(x = date_col, y = log2(total_sales))) +
+   geom_line(size = 1) + stat_smooth(color = "#FC4E07",
+                                     fill = "#FC4E07",
+                                     method = "loess") + labs(y = "Total
Sales (log2)", title = "Total sales vs. Date (Month Year)") +
+   theme_minimal() +
+   theme(axis.title.x = element_blank(), axis.text.x =  element_blank())
+ scale_x_date(date_breaks = "1 months", date_labels = "%m %Y")

> p2 <-
+   ggplot() + geom_line(
+     data = df2,
+     aes(x = date_col, y = num_orders,color = "blue"),
+     size = 1.2
+   )  +
+   geom_line(
+     data = df2,
+     aes(x = date_col, y = num_deliveries,color = "red"),
+     size = 0.7
+   )  +
+   geom_line(
+     data = df2,
+     aes(x = date_col, y = in_transit,color = "green"),
+     size = 0.75
+   ) + theme_minimal() + xlab("Date (Month Year)") +
+   labs(y = "Number of orders / deliveries / in transit", title = "Number
of orders / deliveries / in transit vs. Date (Month Year)") +
+   scale_color_manual(name = "Group",values = c(green = "green", red = "r
ed", blue = "blue"),labels = c("Orders", "In Transit", "Deliveries")) +
+   theme(axis.text.x = element_text(angle = 90),legend.position="bottom")
+ scale_x_date(date_breaks = "1 months", date_labels = "%m %Y")
```

```
> grid.newpage()
> grid.draw(rbind(ggplotGrob(p1), ggplotGrob(p2), size = "last"))
```



Conclusion: The above overlapping line graphs are sharing the same scale represent the total sales and number of orders / deliveries / in transit.
The above line graph shows weak but positive total sales on a daily basis from Jan 2017 to Oct 2018. The trend line in orange color highlights the same.
The below line graph shows overlapping line graphs for number of orders, deliveries and in transit packages.
The line (in blue) shows the total number of orders on daily basis ranging from 30 to 1150. The graph clearly shows a spike during the Thanksgiving week. The graph shows a slight downtrend during the Christmas week.
The line (in red and green) shows the number of delivers and shipments for the packages on daily basis. The lines are fluctuating because there are no deliveries / shipments on the weekends.