Data Project EDA - CMDA 3654 SU19

Lukas Coffey

7/26/2019

Note

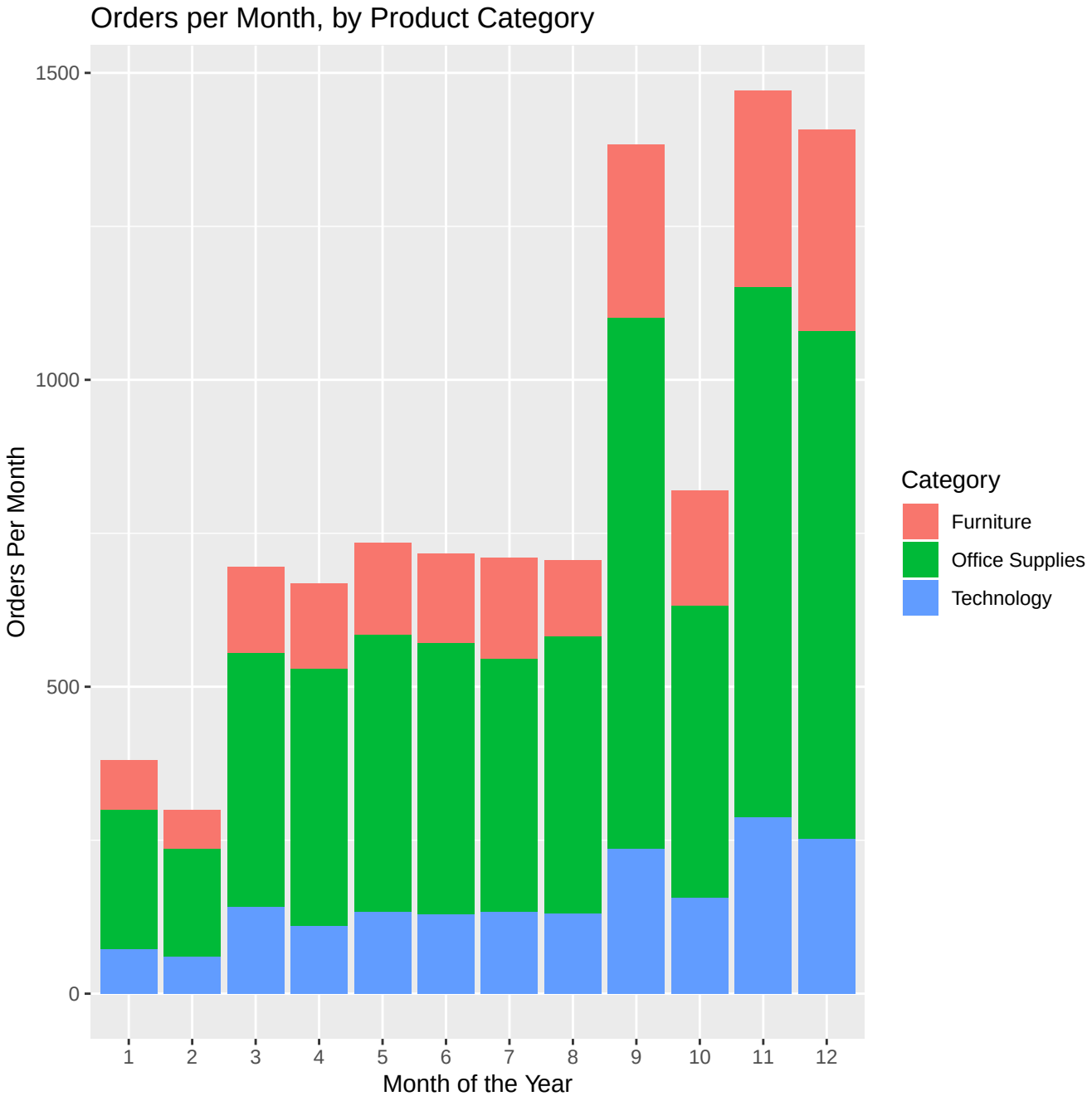
I was unsure whether or not to include my code, so I did include it here.

```
# Sales Data
orders <- read_csv("SampleSuperstoreData/Orders-Table1.csv")
returns <- read_csv("SampleSuperstoreData/Returns-Table2.csv")
# This is official census data put into comma-delimited format, you can find it here: https://gist.github.com/
zipcodes <- read_csv("ExtraData/zipcodes.txt",
      col_types = list(
        "Postal Code" = col_number(),
        "Latitude" = col_double(),
        "Longitude" = col_double()
      ),
      col_names = c("Postal Code", "Latitude", "Longitude"),
      skip = 1
      )

order_data <- left_join(orders, returns)
order_data$Returned[is.na(order_data$Returned)] <- "No"
unique(order_data$Returned) # Successful join and clean of Returned column

# Transform to factors
order_data <- order_data %>%
  mutate_each(funs(as.factor), c(5,8:11,13,15:16))

# Split date into years/months
date_col_names <- c("OrderMonth", "OrderDay", "OrderYear")
order_data <- order_data %>%
  separate(`Order Date`,
           into = date_col_names,
           sep = "/",
           remove = F,
           convert = T)
# Same for shipping date
ship_col_names <- c("ShipMonth", "ShipDay", "ShipYear")
order_data <- order_data %>%
  separate(`Ship Date`,
           into = ship_col_names,
           sep = "/",
           remove = F,
           convert = T)
# Convert OrderDate and ShipDate to date type
order_data <- order_data %>%
  mutate_at(.funs = as.Date,
            format = "%m/%d/%y",
            .vars = c("Order Date", "Ship Date"))

# Join zip code data
order_data <- left_join(order_data, zipcodes)

# Remove excess columns
order_data <- order_data %>%
```

```
  select(-OrderDay, -OrderYear, -ShipDay, -ShipYear, -ShipMonth)
```

Orders Per Month

```
ggplot(order_data,
       aes(x = as.factor(OrderMonth),
           fill = Category)) +
  geom_bar() +
  labs(title = "Orders per Month, by Product Category",
       y = "Orders Per Month",
       x = "Month of the Year")
```

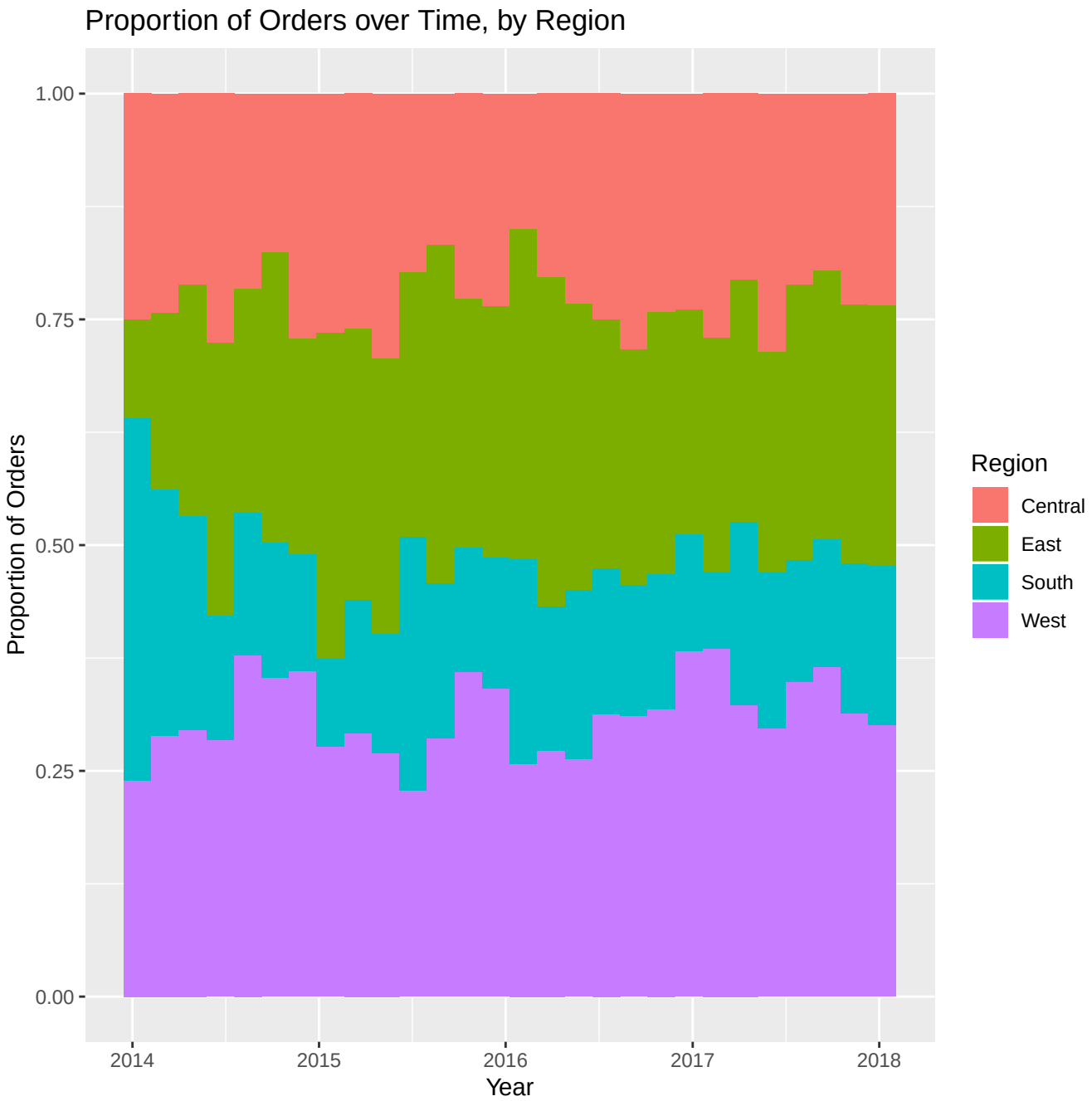


For this visual, I wanted to see which months of the year produced the most orders. I also wanted to know which categories were doing the best duing these months. We can see a clear upward trend at the end each year, and then a "restart" at the beginning of each year. Order volume is fairly consistent throughout the middle 4-5 months. This end of year spike would most likely correspond to Christmas sales, but I find it interesting that the proportion of categories does not vary much during

those months.

Orders Over Time

```
ggplot(order_data,
       aes(x = `Order Date`,
           fill = Region)) +
  geom_histogram(bins = 28,
                 position = "fill") +
  labs(title = "Proportion of Orders over Time, by Region",
       y = "Proportion of Orders",
       x = "Year")
```



Here I'm trying to visualize the proportion of orders going to each region. We don't see a sloping trend in one direction or another, but we do see consistency in order volume. This visual would lead us to believe that we should focus our shipping efforts in the East and West regions mainly, since they hold the largest proportion of orders. This also means that we could push our marketing and sales efforts harder in the Central and South regions; however, this might be difficult since the central

U.S. is much less densely populated.

Orders by Geographic Location

```
ggplot(na.omit(order_data),
       aes(y = Latitude,
           x = Longitude,
           color = Category)) +
  geom_point(aes(size = Sales,
                 shape = Returned)) +
  labs(title = "Coordinates of Sales Region, by Product Category",
       y = "Latitude",
       x = "Longitude",
       size = "Sales Amount (dollars)",
       color = "Product Category",
       shape = "Item Returned?") +
  theme(legend.position = "bottom",
        legend.direction = "vertical")
```

In order to make this plot, I found a data set of zip codes and their corresponding latitude/longitude, and joined it by the postal code column. When looking at this plot, we can see a bulk of orders going to the Northeastern U.S. Most of these orders appear to be in the Office Supplies category.

The biggest sales amounts come from orders in what look like big cities scattered across the U.S. I also joined the other datasheet of return data, mapped by Order ID so we can see where the most returns are coming from.

This may be too much data/variables for one plot, but I was curious to see if there were any trends that would not be visible otherwise.