Project Methodology - Lukas Coffey, CMDA 3654 SU19

Lukas Coffey

8/9/2019

Brief Overview of Dataset

The data contains information about online super store orders. It includes the date, shipping info, pricing info, and product info. There are 21 columns and 9994 rows of data.

```r
# I have organized and cleaned this data, and lat/lng data has been merged in
# based on postal code, and return data has been merged in as well based on Order ID.
# I am not converting columns to factors here for the sake of space and time.
glimpse(orders)
```

```
## Observations: 9,994
## Variables: 25
## $ `Row ID`        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,...
## $ `Order ID`      <chr> "CA-2016-152156", "CA-2016-152156", "CA-2016-1...
## $ `Order Date`    <date> 2016-11-08, 2016-11-08, 2016-06-12, 2015-10-1...
## $ OrderMonth      <dbl> 11, 11, 6, 10, 10, 6, 6, 6, 6, 6, 6, 6, 4, 12,...
## $ `Ship Date`     <date> 2016-11-11, 2016-11-11, 2016-06-16, 2015-10-1...
## $ `Ship Mode`     <chr> "Second Class", "Second Class", "Second Class"...
## $ `Customer ID`   <chr> "CG-12520", "CG-12520", "DV-13045", "SO-20335"...
## $ `Customer Name` <chr> "Claire Gute", "Claire Gute", "Darrin Van Huff...
## $ Segment         <chr> "Consumer", "Consumer", "Corporate", "Consumer...
## $ Country         <chr> "United States", "United States", "United Stat...
## $ City            <chr> "Henderson", "Henderson", "Los Angeles", "Fort...
## $ State           <chr> "Kentucky", "Kentucky", "California", "Florida...
## $ `Postal Code`   <dbl> 42420, 42420, 90036, 33311, 33311, 90032, 9003...
## $ Region          <chr> "South", "South", "West", "South", "South", "W...
## $ `Product ID`    <chr> "FUR-BO-10001798", "FUR-CH-10000454", "OFF-LA-...
## $ Category        <chr> "Furniture", "Furniture", "Office Supplies", "...
## $ `Sub-Category`  <chr> "Bookcases", "Chairs", "Labels", "Tables", "St...
## $ `Product Name`  <chr> "Bush Somerset Collection Bookcase", "Hon Delu...
## $ Sales           <dbl> 261.9600, 731.9400, 14.6200, 957.5775, 22.3680...
## $ Quantity        <dbl> 2, 3, 2, 5, 2, 7, 4, 6, 3, 5, 9, 4, 3, 3, 5, 3...
## $ Discount        <dbl> 0.00, 0.00, 0.00, 0.45, 0.20, 0.00, 0.00, 0.20...
## $ Profit          <dbl> 41.9136, 219.5820, 6.8714, -383.0310, 2.5164, ...
## $ Returned        <chr> "No", "No", "No", "No", "No", "No", "No", "No"...
## $ Latitude        <dbl> 37.81061, 37.81061, 34.07041, 26.14421, 26.144...
## $ Longitude       <dbl> -87.51500, -87.51500, -118.35041, -80.17279, -...
```

Summary of Questions

I've found through EDA that this dataset is fairly normal with not many obvious trends; however, this could just mean I need to dig deeper. These are some questions I'd like to answer:

1. Given a month of the year and a region, which product category is most likely to be purchased?
2. Given a category of product, what region is it most likely to be ordered from?
3. What is the probability of an item being returned based on different factors?

Mainly, these questions are meant to solve problems that an online business might have, such as:

- Where should we build a distribution center, and what products should we keep in stock more than others?
- What time of the year do we need to focus our marketing efforts in, as well as what region?
- What product categories seem to have the most returns? If there are a lot of returns in one category, why is this so? Is it from product defects, failing manafacturing process, shipping damage, not satisfied by the product, etc.

These are just examples of some of the questions you could answer by statistical analysis like this. I've included some plots at the bottom of the report to illustrate some of the trends, and why I want to ask these questions.

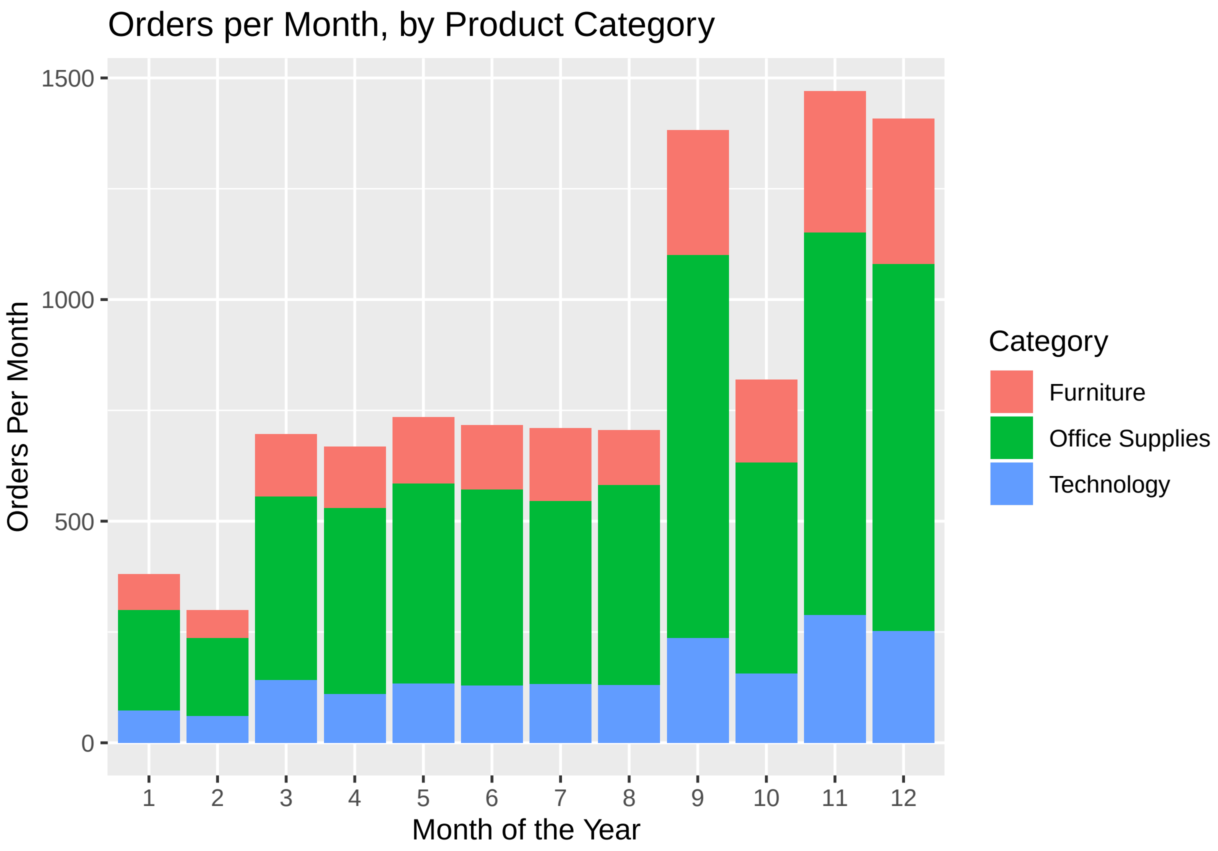
Summary of Statistical Methods

In this case, supervised learning methods seem the most appropriate, because we already know the categories we are trying to answer questions about. I plan to use:

- **Classification using LDA, QDA, or SVM** for the region/category questions,
- **Logistic Regression** for finding the probability of returns.

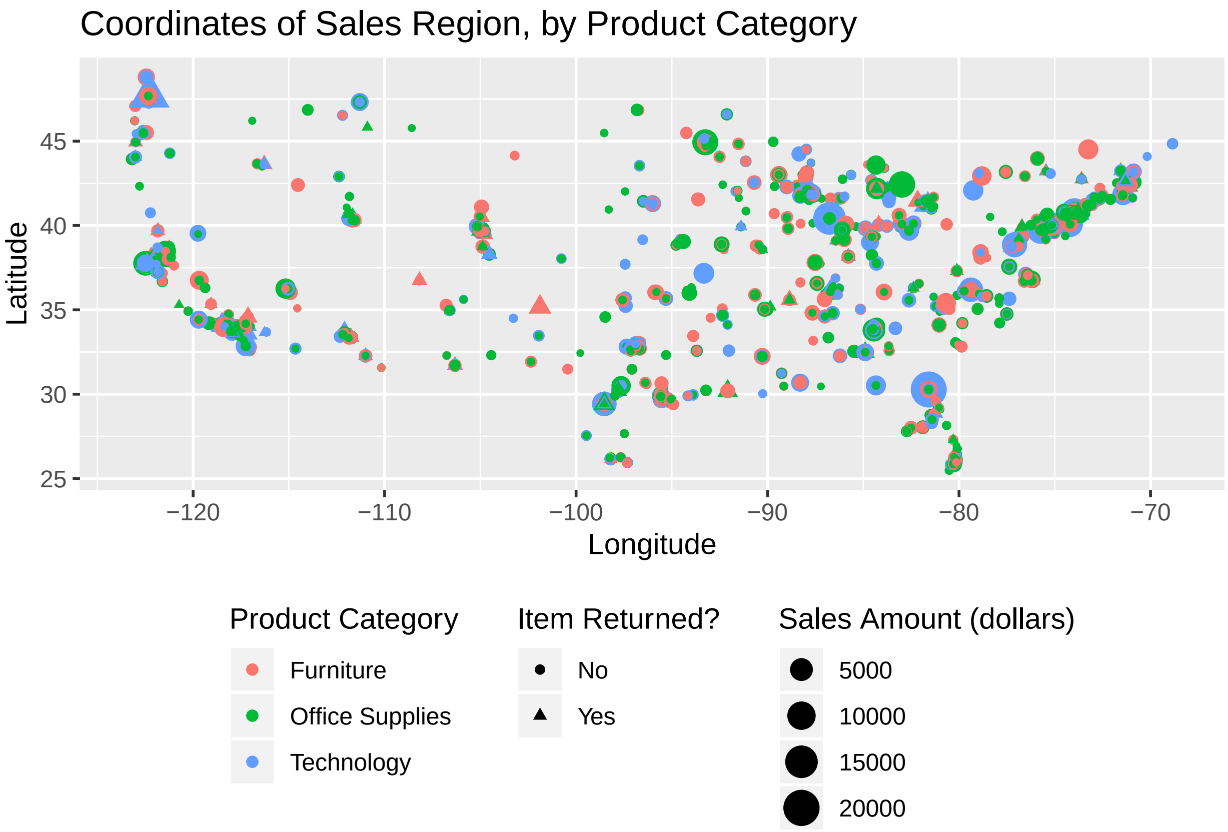
Included Plots

```r
ggplot(orders,
       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")
```



You can see by this map how the orders are distributed across the U.S. I would like to find any patterns in these areas.

```r
sales.map <- ggplot(na.omit(orders),
       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")
sales.map
```


Coordinates of Sales Region, by Product Category
Latitude
45
40
35
30
25
Longitude
−120
−110
−100
−90
−80
−70
Product Category
Furniture
Office Supplies
Technology
Item Returned?
No
Yes
Sales Amount (dollars)
5000
10000
15000
20000