Interview Case Study Write-up

Haoxuan Xu

Problem 2: Identify any deficiencies or errors in the dataset and explain what issues they may lead to if unresolved.

The major deficiency of this dataset is about the inconsistent cell types of key values. What we can see from the current dataset is that two of the three key variables: “Customer Number” and “Item Number” are formatted erroneously as numeric values in the “Invoice Data” sheet. Moreover, the data types of variables of all three key variables: “Customer Number”, “Product Line” and “Item Number” in the “Customer Names”, “Product Line Names”, and “Item Unit Costs” sheets were also formatted erroneously as numeric values.

This deficiency in the dataset can be damaging when we attempt to incorporate different fields from other data tables. Relational database establishes relationships with different tables through having unique “key” values, the “key” values in this case being “Customer Number”, “Product Line” and “Item Number”. These numbers have to be unique in the dataset to allow them to establish relationships across different tables. In another word, the nature of those key values have to be categorical even though they may be comprised of number. For example, if the “Item Number” variable is structured as numeric rather than as characters, the two item numbers “01234” and “1234” would be identical even though they are not. Therefore, you will likely to lose the “uniqueness” of your key values, and this deficiency in the dataset would be likely to result in wrong values being joined or in missing values.

The second deficiency in the raw dataset is the existence of missing values. The original “Invoice Data” contains 10 missing values, 8 in the “Product Line” variable and 2 in the “Item Number” variable (this statistic can be found in the “Invoice Data Joined no NA” sheet of the new Excel file). The existence of missing data could cause bias in the resulting revenue, cost and margin analysis depending on the nature of the missing values. If values are missing at random, then they would not affect the results of the certain analyses (Average Costs per customers, Average Costs per Product Line, Average Revenue per Product Line, Margin per Customer) dramatically. However, if the data are missing in a systematic way (an example will be that a specific product line has more missing values or products to a specific customer has more missing values), then the data is biased, and it will bias the result of the revenue and cost analysis we run subsequently.

Problem 4: Exclude any transactions that do not have a product cost available. Explain why this is important.

Excluding transactions that do not have product costs available are important in calculating the margin percent for 2015 and 2016. This is because margin percent is calculated by dividing the gross profit with revenue, and the gross profit is calculated by subtracting cost from revenue. If the transactions that do not have product costs are not removed from the analysis, then you will have multiple transactions in which the product costs are treated as 0. This will inflate the gross profit and result in a margin percent that is higher than it actually should be.

Problem 5: What next steps would you take in your analysis based on the trends you see in the data so far?

First Part: What are the trends?

Based on the revenue analysis, the total annual revenue from the three customers are $6322575.41. 73% of the revenue is from Acme Corporation, with 18% from Wonka Industries and 9% from Dunder Mifflin (This result can be seen on the sheet "Pivot Table (Revenue)").

The cost analysis shows the average product costs for products delivered to the three customers in 2015 and 2016 (The result can be seen on the sheet "Pivot Table (Product Costs)"). There are two insights that can be gleaned from this result. The first insight is that the average products have increased between 2015 and 2016 for all three customers. The second insight is that products delivered to Acme Corporation have the highest average product costs, followed by products delivered to Dunder Mifflin, and at last by products delivered to Wonka Industries. This insights have interesting implications to the margin percent, especially for products delivered to Acme Corporation, since those products generate the highest total revenue but also have the highest costs, they may not have the highest margin percent.

The Margin percent analysis shows the total product margin percent for products delivered to the three customers in 2015 and 2016. Similarly, there are two insights that can be drawn from this analysis. The first insight is that the margin percent for products delivered to the three customers have decreased from 2015 to 2016. This result is not surprising since that the average product costs of the products delivered to the three customers have increased around the same time. The second insight from the result is that Dunder Mifflin has the highest margin percent in both 2015 and 2016, followed by Acme Corporation and Wonka Industries. This insights is interesting because Dunder Mifflin counts only for 9% of the revenue source, the lowest among all three customers. While the margin percent are not that far apart among the three customers (33% to 40.7% in 2015 and 31% to 33.6% in 2016), it requires us to conduct a more in-depth analysis around the product lines of the products sold to the three customers, as well as on the breakdown of the cost-increases along the product line.

Second Part: What are the next steps?

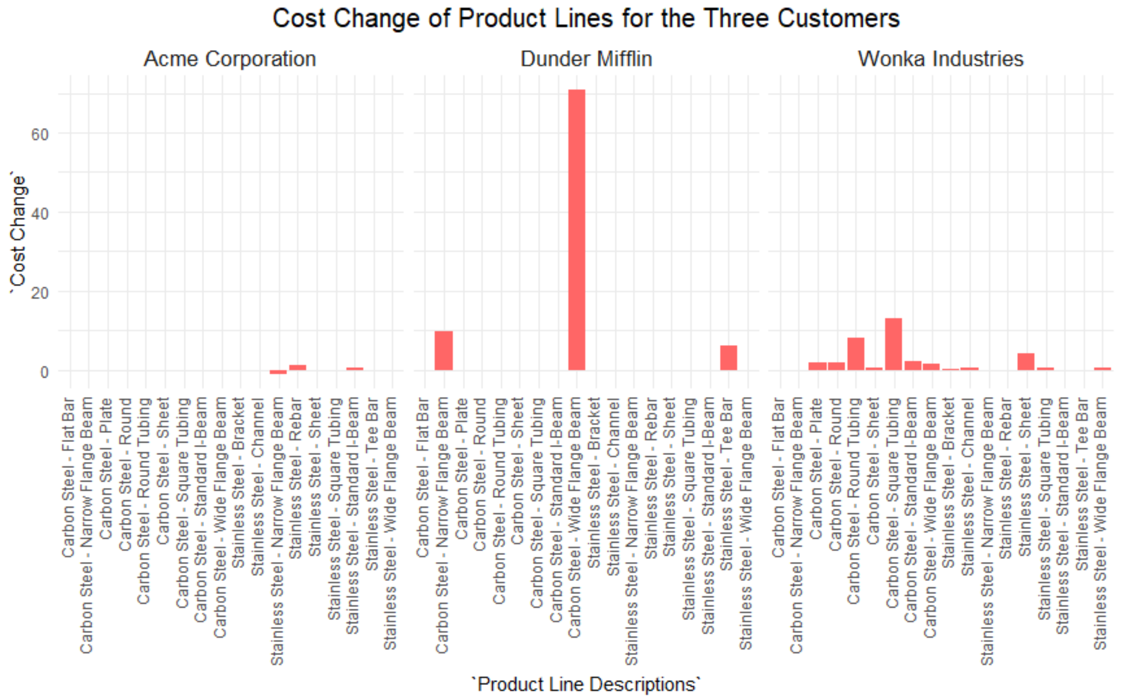
Assuming that the goal of the ABC Co. in this analysis is to find out ways to increase profit and to find out the details behind the cost increase, it makes sense to continue the analysis by finding out which product line is the most profitable, and which customer purchased the most amount from the product line that has the highest profitability.

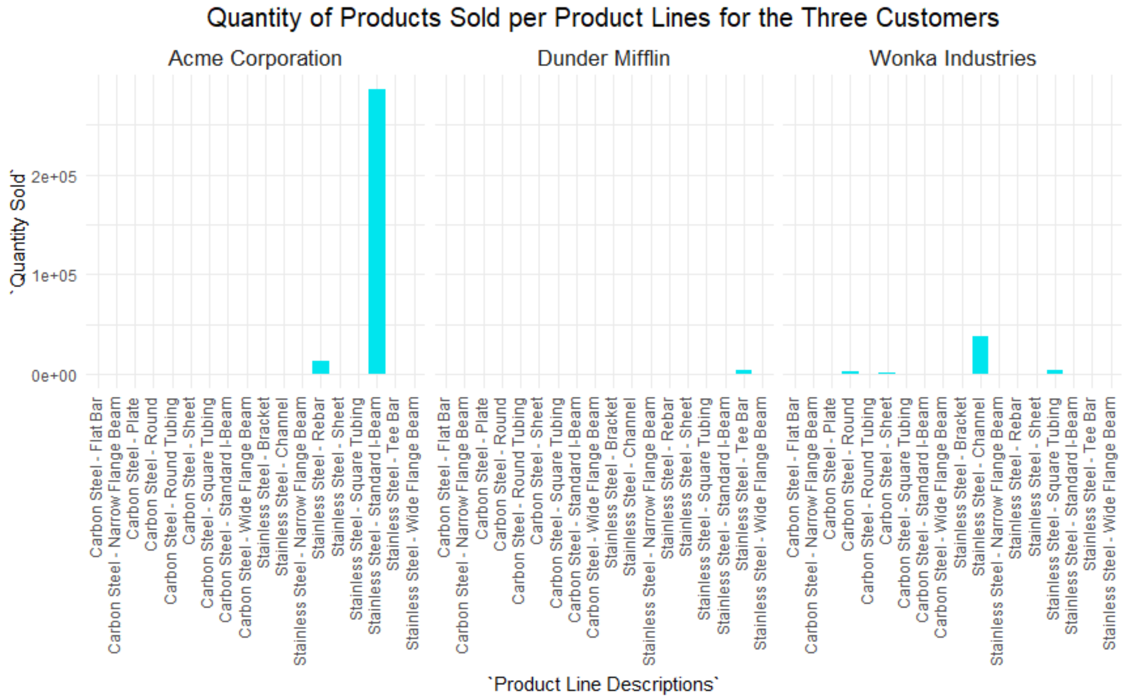
The next step of the analysis will to get a more detailed analysis on values and changes of revenues and profits for each product line between 2015 and 2016, and to evaluate the percentage of products from each product line that were sold to each customers.

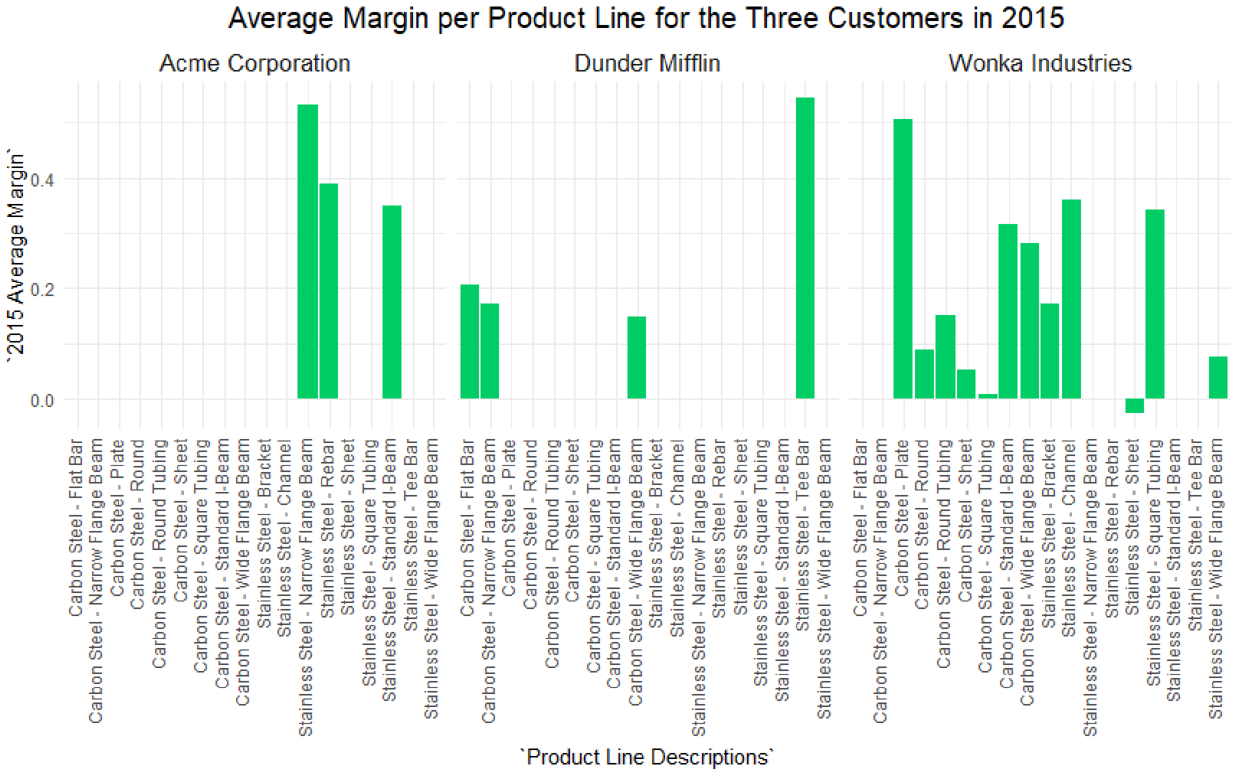
The detailed values needed to be drawn from the existing dataset are such: the quantity sold to each customer per product line, the standard costs for product lines in which products were sold to each customers, change in standard costs between 2015 and 2016 per product line, and the revenue of a single product for different product lines.

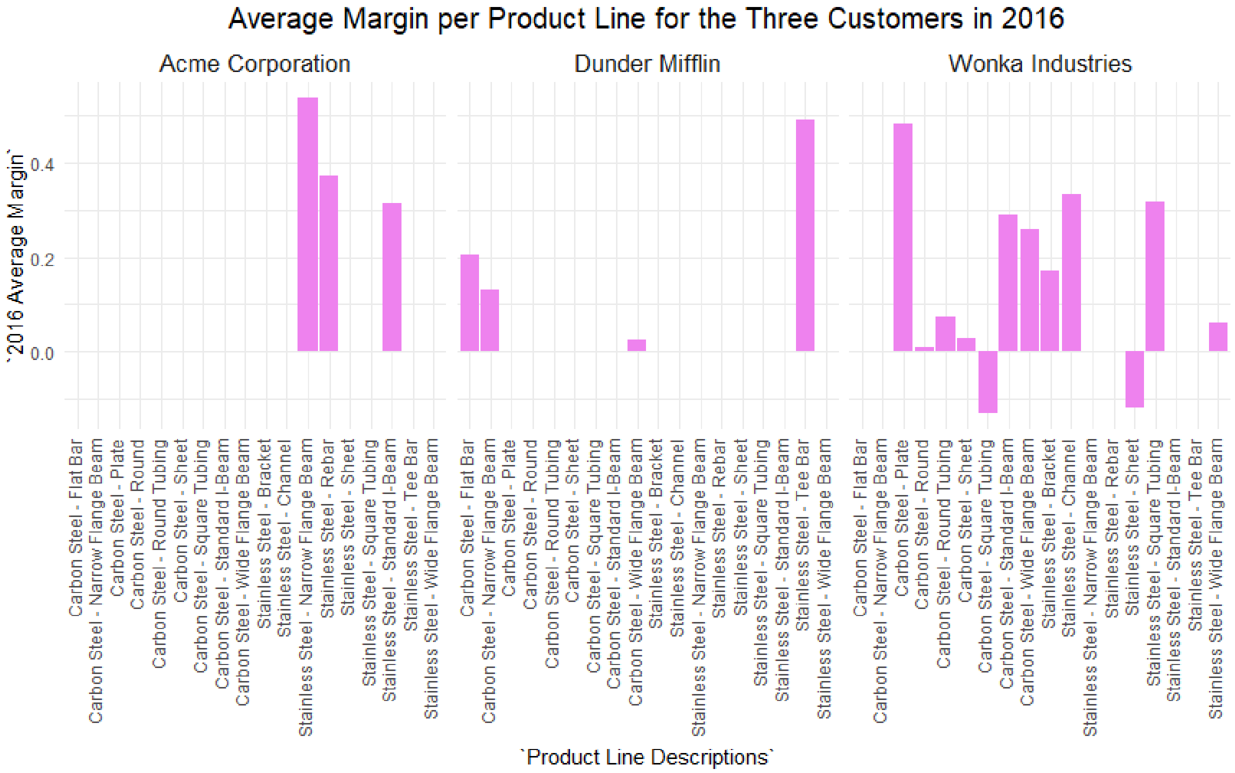
(Those data have been selected and are in the “Data Summary” sheet)

Optional: Preliminary Visualization of the Additional Values Proposed in Question 5






Average Margin per Product Line for the Three Customers in 2015
Acme Corporation
Dunder Mifflin
Wonka Industries
`2015 Average Margin`
0.0
0.2
0.4
Carbon Steel - Flat Bar
Carbon Steel - Narrow Flange Beam
Carbon Steel - Plate
Carbon Steel - Round
Carbon Steel - Round Tubing
Carbon Steel - Sheet
Carbon Steel - Square Tubing
Carbon Steel - Standard I-Beam
Carbon Steel - Wide Flange Beam
Stainless Steel - Bracket
Stainless Steel - Channel
Stainless Steel - Narrow Flange Beam
Stainless Steel - Rebar
Stainless Steel - Sheet
Stainless Steel - Square Tubing
Stainless Steel - Standard I-Beam
Stainless Steel - Tee Bar
Stainless Steel - Wide Flange Beam
`Product Line Descriptions`


Average Margin per Product Line for the Three Customers in 2016
Acme Corporation
Dunder Mifflin
Wonka Industries
`2016 Average Margin`
0.0
0.2
0.4
Carbon Steel - Flat Bar
Carbon Steel - Narrow Flange Beam
Carbon Steel - Plate
Carbon Steel - Round
Carbon Steel - Round Tubing
Carbon Steel - Sheet
Carbon Steel - Square Tubing
Carbon Steel - Standard I-Beam
Carbon Steel - Wide Flange Beam
Stainless Steel - Bracket
Stainless Steel - Channel
Stainless Steel - Narrow Flange Beam
Stainless Steel - Rebar
Stainless Steel - Sheet
Stainless Steel - Square Tubing
Stainless Steel - Standard I-Beam
Stainless Steel - Tee Bar
Stainless Steel - Wide Flange Beam
`Product Line Descriptions`

R Script

```
rm(list = ls())

setwd("D:/Local Disk E/UChicago/Career Development/Summer Internship/Insight2Profit")



#install.packages("openxlsx")
#install.packages("formattable")
#install.packages("extrafont")

library(readxl)
library(rio)
library(tidyverse)
library(dplyr)
library(openxlsx)
library(forcats)
library(formattable)
library(extrafont)

loadfonts (device = "win")
####Load in the Dataset
Invoice_Data <- read_xlsx("Dynamic Model Case Study Data.xlsx",
              sheet = "Invoice Data")
Customer_Names <- read_xlsx("Dynamic Model Case Study Data.xlsx",
                sheet = "Customer Names")
Product_Line_Names <- read_xlsx("Dynamic Model Case Study Data.xlsx",
                  sheet = "Product Line Names")
Item_Unit_Costs <- read_xlsx("Dynamic Model Case Study Data.xlsx",
                          sheet = "Item Unit Costs")


####Task 1: add new fields to the invoice data by bringing in information from the other datasets
##############(Customer Name, Product Line Names, Product Unit Costs)
Invoice_Data_Joined<- Invoice_Data%>%
  full_join(Customer_Names, by = "Customer Number")%>%
  full_join(Product_Line_Names, by = "Product Line")%>%
  full_join(Item_Unit_Costs, by = "Item Number")


####Task 2: Identify any deficiencies or errors in the dataset and explain what issues they may lead to if left unresolved
lapply(Invoice_Data_Joined, class)   ##Look through the class of variables in the Dataset
sapply(Invoice_Data, class)
sapply(Customer_Names, class)
sapply(Item_Unit_Costs, class)
sapply(Product_Line_Names, class)


#####Find out if there are NA values in the dataset
sum(is.na(Invoice_Data_Joined)) #There is substantial number of NA values--354
sum(is.na(Invoice_Data))    ##The original Invoice Data Contains 10 NA values

na_count_original <-sapply(Invoice_Data, function(y) sum(length(which(is.na(y)))))
na_count_original <- data.frame(na_count_original)
na_count_original <- na_count_original%>%
  rownames_to_column("variable")
```

```
na_count <-sapply(Invoice_Data_Joined, function(y) sum(length(which(is.na(y)))))

na_count <- data.frame(na_count)

na_count <- na_count%>%

rownames_to_column("variable")  ##It seems that majority of the NA values are from "2015 Standard Cost" and "2016 Standard Cost"




Standard_cost_NA <- filter(Invoice_Data_Joined, is.na(`2015 Standard Cost`)==TRUE, is.na(`2015 Standard Cost`))



###Part 3: Compare ABC Co's Revenue, Product Costs and Margin Percent for 2015 and 2016 for each of the three customers


###Calculate the margin for each product

Invoice_Data_Joined_no_NA <- Invoice_Data_Joined%>%

  na.omit()%>%

 mutate(`2015 Total Cost` = `2015 Standard Cost`*`Qty Shipped`,

        `2016 Total Cost` = `2016 Standard Cost`*`Qty Shipped`,

        `2015 Profit` = `Net Sales` - `2015 Total Cost`,

       `2016 Profit` = `Net Sales` - `2016 Total Cost`,

       `2015 Margin` = `2015 Profit`/`Net Sales`,

       `2016 Margin` = `2016 Profit`/`Net Sales`)



###Creating Summary Tables Margin Percent for 2015 and 2016 for each of the three customers

Margin_Percent_per_Customer <- Invoice_Data_Joined_no_NA%>%

  group_by(`Customer Name`)%>%

  summarize(`2015 Margin Percent` = percent(sum(`2015 Profit`)/sum(`Net Sales`)),
        `2016 Margin Percent` = percent(sum(`2016 Profit`)/sum(`Net Sales`)))
```

```
#####Optional: Next Step--Find the following values:
```

```
Data_Summary <- Invoice_Data_Joined_no_NA%>%
  group_by(`Customer Name`, `Product Line Descriptions`)%>%
  summarize(`Average price` = sum(`Net Sales`)/sum(`Qty Shipped`), ###Find the price for a single product per product line
        `Quantity Sold` = sum(`Qty Shipped`),     ###Find the quantity sold to each customer per product line
        `2015 Average Cost per Product Line` = sum(`2015 Total Cost`)/sum(`Qty Shipped`),
        `2016 Average Cost per Product Line` = sum(`2016 Total Cost`)/sum(`Qty Shipped`),
        `Cost Change` = `2016 Average Cost per Product Line` - `2015 Average Cost per Product Line`,
        `2015 Average Profit` = `Average price` - `2015 Average Cost per Product Line`,
        `2016 Average Profit` = `Average price` - `2016 Average Cost per Product Line`,
        `2016 Average Margin` = `2016 Average Profit`/`Average price`,
        `2015 Average Margin` = `2015 Average Profit`/`Average price`)%>%
        ungroup()
```

```
 ####Optional: Visualization of these values
Data_Summary%>%
  ggplot(aes(x = `Product Line Descriptions`, y = `Cost Change`)) +
  geom_col(fill = "#FF6666") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.2),
        plot.title = element_text(hjust = 0.5, size = 15),
        strip.text = element_text(size = 13))+
  labs(title = "Cost Change of Product Lines for the Three Customers")+
  facet_wrap(~`Customer Name`)
```

```
Data_Summary%>%
  ggplot(aes(x = `Product Line Descriptions`, y = `Quantity Sold`)) +
  geom_col(fill = "turquoise2") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.2),
        plot.title = element_text(hjust = 0.5, size = 15),
        axis.text.y = element_text(family = "Bookman"),
        strip.text = element_text(size = 13))+
  labs(title = "Quantity of Products Sold per Product Lines for the Three Customers")+
  facet_wrap(~`Customer Name`)
```

```
Data_Summary%>%
  ggplot(aes(x = `Product Line Descriptions`, y = `Quantity Sold`)) +
  geom_col(fill = "turquoise2") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.2),
        plot.title = element_text(hjust = 0.5, size = 15),
        axis.text.y = element_text(family = "Bookman"),
        strip.text = element_text(size = 13))+
  labs(title = "Quantity of Products Sold per Product Lines for the Three Customers")+
  facet_wrap(~`Customer Name`)



Data_Summary%>%
  ggplot(aes(x = `Product Line Descriptions`, y = `2015 Average Margin`)) +
  geom_col(fill = "springgreen3") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.2),
        plot.title = element_text(hjust = 0.5, size = 15),
        strip.text = element_text(size = 13))+
  labs(title = "Average Margin per Product Line for the Three Customers in 2015")+
  facet_wrap(~`Customer Name`)


Data_Summary%>%
  ggplot(aes(x = `Product Line Descriptions`, y = `2016 Average Margin`)) +
  geom_col(fill = "violet") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.2),
        plot.title = element_text(hjust = 0.5, size = 15),
        strip.text = element_text(size = 13))+
  labs(title = "Average Margin per Product Line for the Three Customers in 2016")+
  facet_wrap(~`Customer Name`)

#####Exxport the dataset first into a new Excel file
write.xlsx(list(Invoice_Data_Joined_no_NA, Customer_Names, Product_Line_Names, Item_Unit_Costs),
           file = "Dynamic Model Case Study Data Analysis.xlsx",
    sheetName = c("Invoice Data Joined no NA", "Customer Names", "Product Line Names", "Item Unit Costs"))

WorkBook <- loadWorkbook(file = "Dynamic Model Case Study Data Analysis.xlsx")
addWorksheet(WorkBook, sheet = "Data Summary")
writeData(WorkBook, na_count, sheet = "Invoice Data Joined no NA",
    startCol = 20, startRow = 3)
writeData(WorkBook, na_count_original, sheet = "Invoice Data Joined no NA",
    startCol = 24, startRow = 3)
writeData(WorkBook, Data_Summary, sheet = "Data Summary",
    startCol = 1, startRow = 1)
writeData(WorkBook, Margin_Percent_per_Customer, sheet = "Data Summary",
    startCol = 18, startRow = 1)

saveWorkbook(WorkBook, "Dynamic Model Case Study Data Analysis.xlsx", overwrite = TRUE)
```