Predicting Sales Win or Lose

Herman Toeante

February 12, 2019

How do you know which deals will close? You've worked your territory and leads. Demonstrated the value of your product or service to your customers' businesses and it should be a done deal. But is it? All too often, determining which customers will buy is a guessing game. You have pipeline reports, regional sales figures and wins and losses that you could analyze. Unlock the information in those sources and you'll unlock more revenue and more satisfied customers. Better understanding of sales pipeline can help any sales team organization can expect win or lose based on data. In this project, I am going to be the sales manager at an automotive supply company. Any B2B company like GE, AMAZON, GOOGLE, ADP, etc can use the same approach that I demonstrated on this project to their sales team. As a manager, I'm trying to assess a sales execution issue. We have not been able to convert enough opportunities lately. As a start, I load the required packages and data for this project.

```
library(tidyverse)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.2
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.5.2
```

```
sales_win_loss <- read_csv("DATA/WA_Fn-UseC_-Sales-Win-Loss.csv")
```

Data for this project is publicly available from Wattson Analytic sample data. (https://community.watsonanalytics.com/wp-content/uploads/2015/04/WA_Fn-UseC_-Sales-Win-Loss.csv) The csv file contains 78K row of data and 19 columns. Each row is assigned a unique sales opportunity ID. The dependent variable is 'Opportunity Result' column with values of either 'Won' or 'Loss' There are several independent variables from the sample data that I can use such as: 'Supplies Group', 'Region', 'Route To Market', 'Elapsed Days In Sales Stage', 'Opportunity Result', 'Sales Stage Change Count', 'Total Days Identified Through Closing', 'Total Days Identified Through Qualified', 'Opportunity Amount USD', 'Client Size By Revenue', 'Client Size By Employee Count', 'Revenue From Client Past Two Years', and 'Competitor Type.' I use 'glimpse' and 'head' to understand the data structure of my sample data.

```
glimpse(sales_win_loss, give.attr = FALSE)
```

```
## Observations: 78,025
## Variables: 19
## $ `Opportunity Number`                    <int> 1641984, 1658010, 16...
## $ `Supplies Subgroup`                     <chr> "Exterior Accessorie...
## $ `Supplies Group`                        <chr> "Car Accessories", "...
## $ Region                                  <chr> "Northwest", "Pacifi...
## $ `Route To Market`                       <chr> "Fields Sales", "Res...
## $ `Elapsed Days In Sales Stage`           <int> 76, 63, 24, 16, 69, ...
## $ `Opportunity Result`                    <chr> "Won", "Loss", "Won"...
## $ `Sales Stage Change Count`              <int> 13, 2, 7, 5, 11, 3, ...
```

```
## $ `Total Days Identified Through Closing`   <int> 104, 163, 82, 124, 9...
## $ `Total Days Identified Through Qualified` <int> 101, 163, 82, 124, 1...
## $ `Opportunity Amount USD`                  <int> 0, 0, 7750, 0, 69756...
## $ `Client Size By Revenue`                  <int> 5, 3, 1, 1, 1, 5, 4,...
## $ `Client Size By Employee Count`           <int> 5, 5, 1, 1, 1, 1, 5,...
## $ `Revenue From Client Past Two Years`      <int> 0, 0, 0, 0, 0, 0, 0,...
## $ `Competitor Type`                         <chr> "Unknown", "Unknown"...
## $ `Ratio Days Identified To Total Days`     <dbl> 0.696360, 0.000000, ...
## $ `Ratio Days Validated To Total Days`      <dbl> 0.113985, 1.000000, ...
## $ `Ratio Days Qualified To Total Days`      <dbl> 0.154215, 0.000000, ...
## $ `Deal Size Category`                      <int> 1, 1, 1, 1, 4, 5, 2,...
```

```r
summary(sales_win_loss)
```

```
##  Opportunity Number Supplies Subgroup  Supplies Group
##  Min.   : 1641984   Length:78025       Length:78025
##  1st Qu.: 6900423   Class :character   Class :character
##  Median : 7545569   Mode  :character   Mode  :character
##  Mean   : 7653429
##  3rd Qu.: 8228329
##  Max.   :10094266
##     Region          Route To Market    Elapsed Days In Sales Stage
##  Length:78025       Length:78025       Min.   :  0.0
##  Class :character   Class :character   1st Qu.: 19.0
##  Mode  :character   Mode  :character   Median : 43.0
##                                        Mean   : 43.6
##                                        3rd Qu.: 65.0
##                                        Max.   :210.0
##  Opportunity Result Sales Stage Change Count
##  Length:78025       Min.   : 1.000
##  Class :character   1st Qu.: 2.000
##  Mode  :character   Median : 3.000
##                     Mean   : 2.956
##                     3rd Qu.: 3.000
##                     Max.   :23.000
##  Total Days Identified Through Closing
##  Min.   :  0.00
##  1st Qu.:  4.00
##  Median : 12.00
##  Mean   : 16.73
##  3rd Qu.: 24.00
##  Max.   :208.00
##  Total Days Identified Through Qualified Opportunity Amount USD
##  Min.   :  0.00                          Min.   :      0
##  1st Qu.:  4.00                          1st Qu.:  15000
##  Median : 12.00                          Median :  49000
##  Mean   : 16.31                          Mean   :  91637
##  3rd Qu.: 24.00                          3rd Qu.: 105099
##  Max.   :208.00                          Max.   :1000000
##  Client Size By Revenue Client Size By Employee Count
##  Min.   :1.00           Min.   :1.000
##  1st Qu.:1.00           1st Qu.:1.000
##  Median :1.00           Median :1.000
##  Mean   :1.62           Mean   :1.604
```

```
##  3rd Qu.:1.00           3rd Qu.:1.000
##  Max.   :5.00           Max.   :5.000
##  Revenue From Client Past Two Years Competitor Type
##  Min.   :0.0000                     Length:78025
##  1st Qu.:0.0000                     Class :character
##  Median :0.0000                     Mode  :character
##  Mean   :0.3033
##  3rd Qu.:0.0000
##  Max.   :4.0000
##  Ratio Days Identified To Total Days Ratio Days Validated To Total Days
##  Min.   :0.0000                      Min.   :0.0000
##  1st Qu.:0.0000                      1st Qu.:0.0000
##  Median :0.0000                      Median :0.4480
##  Mean   :0.2031                      Mean   :0.4883
##  3rd Qu.:0.1972                      3rd Qu.:1.0000
##  Max.   :1.0000                      Max.   :1.0000
##  Ratio Days Qualified To Total Days Deal Size Category
##  Min.   :0.0000                     Min.   :1.000
##  1st Qu.:0.0000                     1st Qu.:2.000
##  Median :0.0000                     Median :3.000
##  Mean   :0.1850                     Mean   :3.437
##  3rd Qu.:0.1886                     3rd Qu.:5.000
##  Max.   :1.0000                     Max.   :7.000
```

```
head(sales_win_loss[, 1:6])
```

```
## # A tibble: 6 x 6
##   `Opportunity Nu~ `Supplies Subgr~ `Supplies Group` Region
##              <int> <chr>            <chr>            <chr>
## 1          1641984 Exterior Access~ Car Accessories  North~
## 2          1658010 Exterior Access~ Car Accessories  Pacif~
## 3          1674737 Motorcycle Parts Performance & N~ Pacif~
## 4          1675224 Shelters & RV    Performance & N~ Midwe~
## 5          1689785 Exterior Access~ Car Accessories  Pacif~
## 6          1692390 Shelters & RV    Performance & N~ Pacif~
## # ... with 2 more variables: `Route To Market` <chr>, `Elapsed Days In
## #   Sales Stage` <int>
```

```
head(sales_win_loss[, 7:13])
```

```
## # A tibble: 6 x 7
##   `Opportunity Re~ `Sales Stage Ch~ `Total Days Ide~ `Total Days Ide~
##   <chr>                       <int>            <int>            <int>
## 1 Won                            13              104              101
## 2 Loss                            2              163              163
## 3 Won                             7               82               82
## 4 Loss                            5              124              124
## 5 Loss                           11               91               13
## 6 Loss                            3              114                0
## # ... with 3 more variables: `Opportunity Amount USD` <int>, `Client Size
## #   By Revenue` <int>, `Client Size By Employee Count` <int>
```

```
head(sales_win_loss[, 14:19])
```

```
## # A tibble: 6 x 6
##   `Revenue From C~ `Competitor Typ~ `Ratio Days Ide~ `Ratio Days Val~
##              <int> <chr>                       <dbl>            <dbl>
## 1                0 Unknown                     0.696         0.114
## 2                0 Unknown                     0             1
## 3                0 Unknown                     1             0
## 4                0 Known                       1             0
## 5                0 Unknown                     0             0.141
## 6                0 Unknown                     0             0.000877
## # ... with 2 more variables: `Ratio Days Qualified To Total Days` <dbl>,
## #   `Deal Size Category` <int>
```

I checked for missing values of my dataset

```
map_dbl(sales_win_loss, ~sum(is.na(.)))
```

```
##                      Opportunity Number
##                                       0
##                       Supplies Subgroup
##                                       0
##                          Supplies Group
##                                       0
##                                  Region
##                                       0
##                         Route To Market
##                                       0
##             Elapsed Days In Sales Stage
##                                       0
##                      Opportunity Result
##                                       0
##                Sales Stage Change Count
##                                       0
##   Total Days Identified Through Closing
##                                       0
## Total Days Identified Through Qualified
##                                       0
##                   Opportunity Amount USD
##                                       0
##                   Client Size By Revenue
##                                       0
##           Client Size By Employee Count
##                                       0
##      Revenue From Client Past Two Years
##                                       0
##                          Competitor Type
##                                       0
##     Ratio Days Identified To Total Days
##                                       0
##      Ratio Days Validated To Total Days
##                                       0
```

```
##      Ratio Days Qualified To Total Days
##                                       0
##                       Deal Size Category
##                                       0
```

The next step is setting the standard theme for the charts to theme_minimal with legend set at the bottom of the chart.

```
theme_set(theme_minimal() + theme(legend.position = "bottom"))
```

For better data visualization on the chart, I rename the columns with long name.

```
colnames(sales_win_loss) <- c("ID","SuppliesSubgroup","SuppliesGroup","Region", "Route",
                              "ElapsedDays", "Result","SalesStageCount",
                              "TotalDaysClosing","TotalDaysQualified",
                              "Opportunity","ClientSizeRev","ClientSizeCount",
                              "Revenue","Competitor","RDaysIdentified",
                              "RDaysValidated","RDaysQualified",
                              "DealSize")
```

Moreover, I made several assumptions to translate the categorical columns into meaningful information.

```
  sales_win_loss <- sales_win_loss %>%
    mutate(ClientSizeRev2 = case_when(
      ClientSizeRev == 1 ~ "ClientRev<$1M",
      ClientSizeRev == 2 ~ "$1M<=ClientRev<$10M",
      ClientSizeRev == 3 ~ "$10M<=ClientRev<$50M",
      ClientSizeRev == 4 ~ "$50M<=ClientRev<$100M",
      ClientSizeRev == 5 ~ "ClientRev>=$100M"))

  sales_win_loss <- sales_win_loss %>%
    mutate(ClientSizeCount2 = case_when(
      ClientSizeCount == 1 ~ "Count<1K",
      ClientSizeCount == 2 ~ "1K<=Count<5K",
      ClientSizeCount == 3 ~ "5K<=Count<10K",
      ClientSizeCount == 4 ~ "10K<=Count<30K",
      ClientSizeCount == 5 ~ "Count>=30K"))

  sales_win_loss <- sales_win_loss %>%
    mutate(Revenue2 = case_when(
      Revenue == 0 ~ "Rev=$0",
      Revenue == 1 ~ "$1<=Rev<$50K",
      Revenue == 2 ~ "$50K<=Rev<$400K",
      Revenue == 3 ~ "$400K<=Rev<$1.5M",
      Revenue == 4 ~ "Rev>=$1.5M"))
```

There are several values in the dataset where the sales opportunity resulting in 'Won' but the revenue is '$0' I am going to exclude these data from the dataset

sales_win_loss %>% sales_win_loss %>% (filter(Result == "Won" & Revenue == 0)) – Error WHY?

Revenue description is it before, after closing

Data Dictionary

```
var_descriptions <- c(
  "A random number assigned to the opportunity",
  "Supplies Subgroup",
  "Supplies Group",
  "Region",
  "Route to market",
  "The number of days between the change in sales stages",
  "A closed opportunity. Values is either won or loss",
  "A count of number of times an opportunity changes sales stages",
  "Total days from Identified to Gained Agreement/closing",
  "Total days from Identified to Qualified Agreement",
  "Sum of line item revenue estimates",
  "Client size based on annual revenue",
  "Client size based on number of employees",
  "Revenue from client the past two years",
  "An indicator whether or not competitor has been identified",
  "Ratio of Identified/Validating over total days",
  "Ratio of Qualified/Gaining Agreement over total days",
  "Ratio of Validated/Qualifying over total days",
  "Categorical grouping of the opportunity amount"
)

var <- colnames(sales_win_loss)
var_type <- unlist(map(sales_win_loss, class))
as_tibble(cbind(c(var, var_type, var_descriptions)))
```

```
## # A tibble: 63 x 1
##    V1
##    <chr>
##  1 ID
##  2 SuppliesSubgroup
##  3 SuppliesGroup
##  4 Region
##  5 Route
##  6 ElapsedDays
##  7 Result
##  8 SalesStageCount
##  9 TotalDaysClosing
## 10 TotalDaysQualified
## # ... with 53 more rows
```

```
as_data_frame(cbind(c(1:length(var)), var, var_type, var_descriptions))
```

```
## Warning in cbind(c(1:length(var)), var, var_type, var_descriptions): number
## of rows of result is not a multiple of vector length (arg 4)
```

```
## # A tibble: 22 x 4
```

```
##    V1    var           var_type var_descriptions
##    <chr> <chr>         <chr>    <chr>
##  1 1     ID            integer  A random number assigned to the opportuni~
##  2 2     SuppliesSubg~ charact~ Supplies Subgroup
##  3 3     SuppliesGroup charact~ Supplies Group
##  4 4     Region        charact~ Region
##  5 5     Route         charact~ Route to market
##  6 6     ElapsedDays   integer  The number of days between the change in ~
##  7 7     Result        charact~ A closed opportunity. Values is either wo~
##  8 8     SalesStageCo~ integer  A count of number of times an opportunity~
##  9 9     TotalDaysClo~ integer  Total days from Identified to Gained Agre~
## 10 10    TotalDaysQua~ integer  Total days from Identified to Qualified A~
## # ... with 12 more rows
```

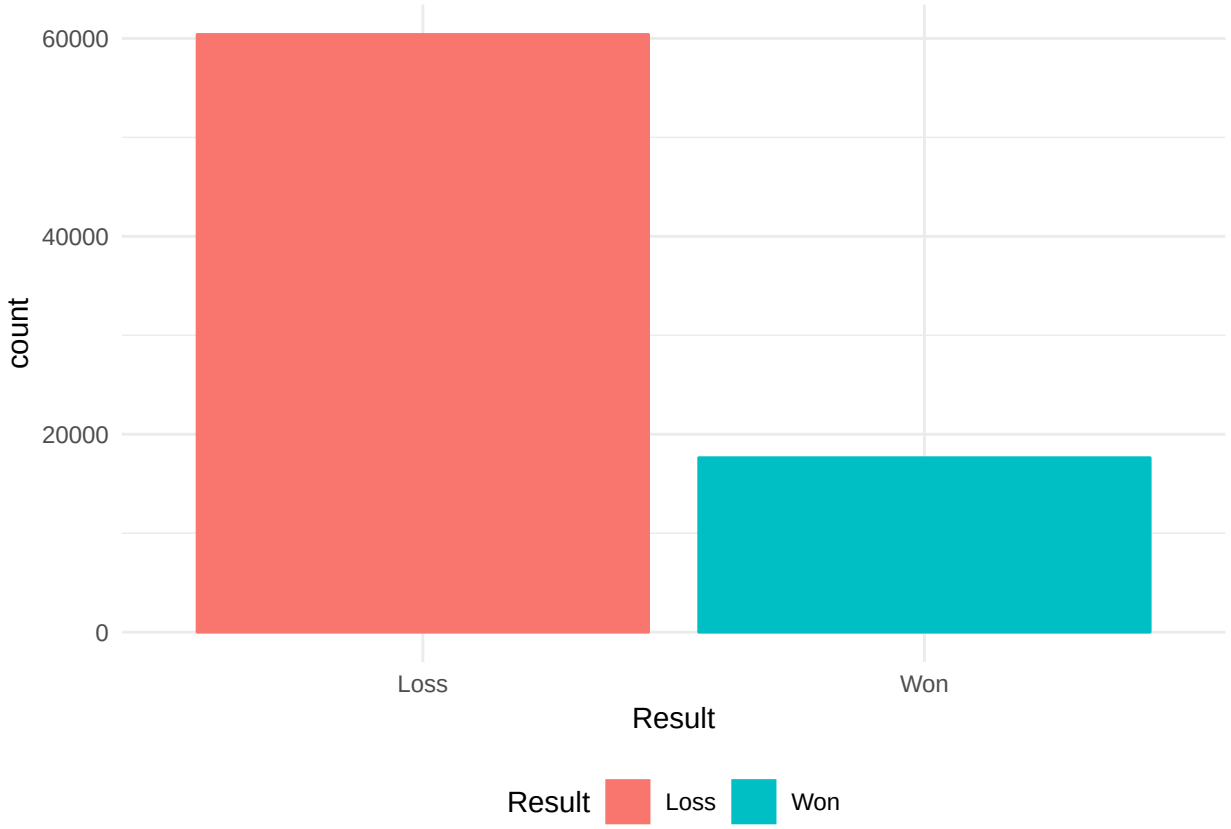
```
as_tibble(sales_win_loss)
```

```
## # A tibble: 78,025 x 22
##        ID SuppliesSubgroup SuppliesGroup Region Route ElapsedDays Result
##     <int> <chr>            <chr>         <chr>  <chr>       <int> <chr>
##  1 1.64e6 Exterior Access~ Car Accessor~ North~ Fiel~          76 Won
##  2 1.66e6 Exterior Access~ Car Accessor~ Pacif~ Rese~          63 Loss
##  3 1.67e6 Motorcycle Parts Performance ~ Pacif~ Rese~          24 Won
##  4 1.68e6 Shelters & RV    Performance ~ Midwe~ Rese~          16 Loss
##  5 1.69e6 Exterior Access~ Car Accessor~ Pacif~ Rese~          69 Loss
##  6 1.69e6 Shelters & RV    Performance ~ Pacif~ Rese~          89 Loss
##  7 1.94e6 Garage & Car Ca~ Car Accessor~ Pacif~ Fiel~         111 Won
##  8 1.95e6 Exterior Access~ Car Accessor~ Pacif~ Fiel~          82 Loss
##  9 2.00e6 Batteries & Acc~ Car Accessor~ North~ Fiel~          68 Loss
## 10 2.05e6 Exterior Access~ Car Accessor~ Pacif~ Rese~          18 Loss
## # ... with 78,015 more rows, and 15 more variables: SalesStageCount <int>,
## #   TotalDaysClosing <int>, TotalDaysQualified <int>, Opportunity <int>,
## #   ClientSizeRev <int>, ClientSizeCount <int>, Revenue <int>,
## #   Competitor <chr>, RDaysIdentified <dbl>, RDaysValidated <dbl>,
## #   RDaysQualified <dbl>, DealSize <int>, ClientSizeRev2 <chr>,
## #   ClientSizeCount2 <chr>, Revenue2 <chr>
```

Data Exploration

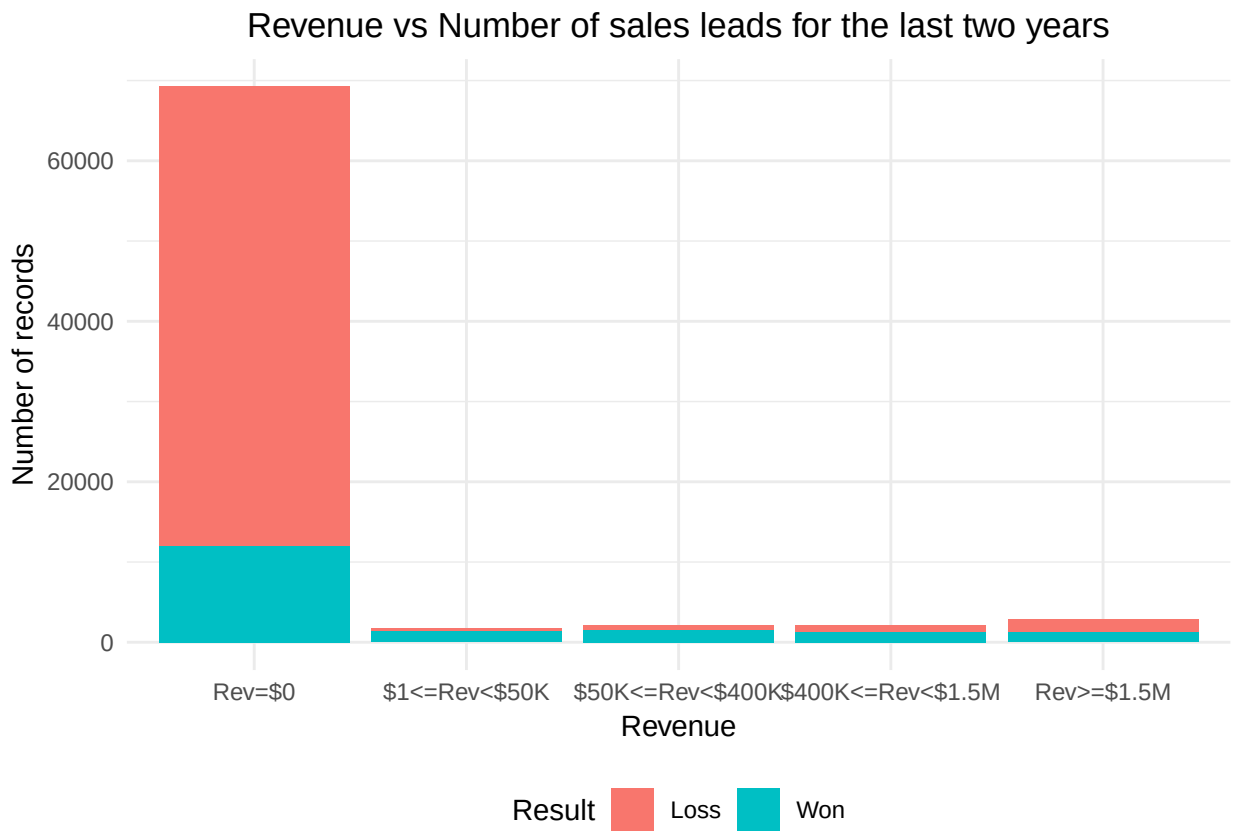
The first chart is to understand the number of sales leads that the company won versus loss in respect to revenue for the last two years.

```
ggplot(sales_win_loss, aes(x = Result, color = Result, fill = Result)) +
  geom_bar()
```



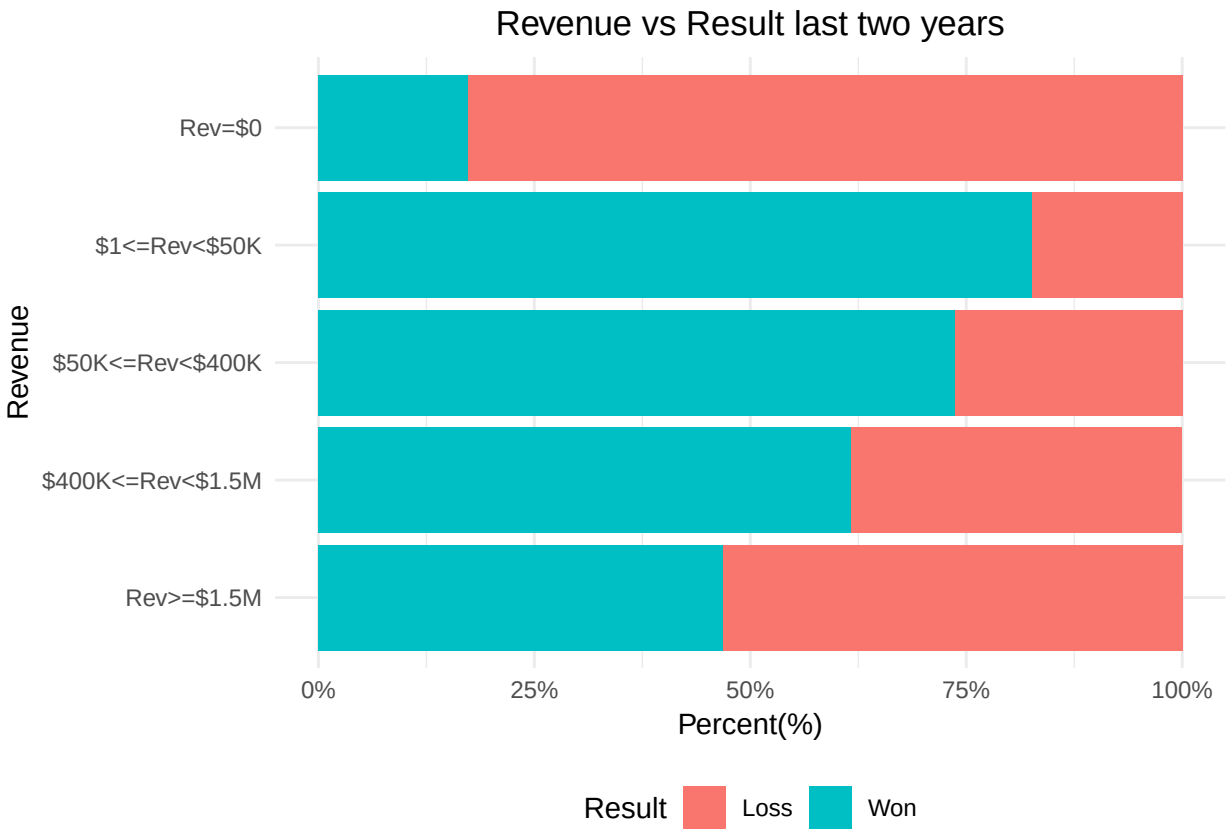
From the chart, we have ~18K records of 'Won' sales leads and ~60K 'Loss" sales leads for the last two years. Using this data, it looks the high number of loss opprotunity could be attributed to business condition or we can explore deeper into the data set if there are variables that affecting the loss opportunity.

```
position <- c("Rev=$0", "$1<=Rev<$50K", "$50K<=Rev<$400K", "$400K<=Rev<$1.5M", "Rev>=$1.5M")
 ggplot(sales_win_loss, aes(x = Revenue2, fill = Result)) +
  geom_bar() +
  scale_x_discrete(limits = position) +
  xlab("Revenue") +
  ylab("Number of records") +
  ggtitle("Revenue vs Number of sales leads for the last two years") +
  theme(plot.title = element_text(hjust = 0.5))
```



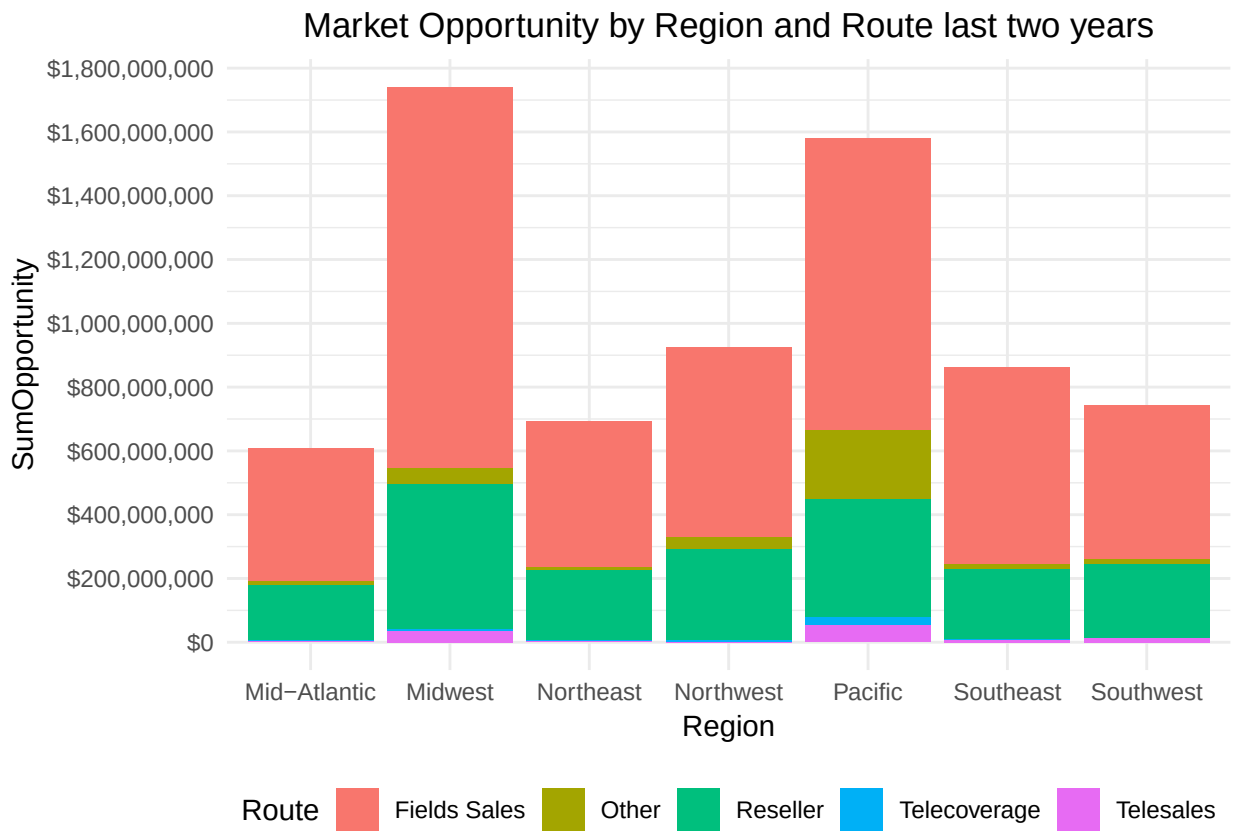
From the chart, we can see for the last two years, there are plenty of sales leads opportunity A lot sales leads opportunity that resulted in $0 revenue for the last two years that the sales team can learn or put more effort

```
position <- c("Rev>=$1.5M","$400K<=Rev<$1.5M","$50K<=Rev<$400K","$1<=Rev<$50K","Rev=$0")
ggplot(sales_win_loss) +
  geom_bar(aes(x = Revenue2, fill = Result), position = "fill") +
  scale_x_discrete(limits = position) +
  scale_y_continuous(labels = scales::percent_format()) +
  coord_flip() +
  ggtitle("Revenue vs Result last two years") +
  xlab("Revenue") +
  ylab("Percent(%)") +
  theme(plot.title = element_text(hjust = 0.5))
```



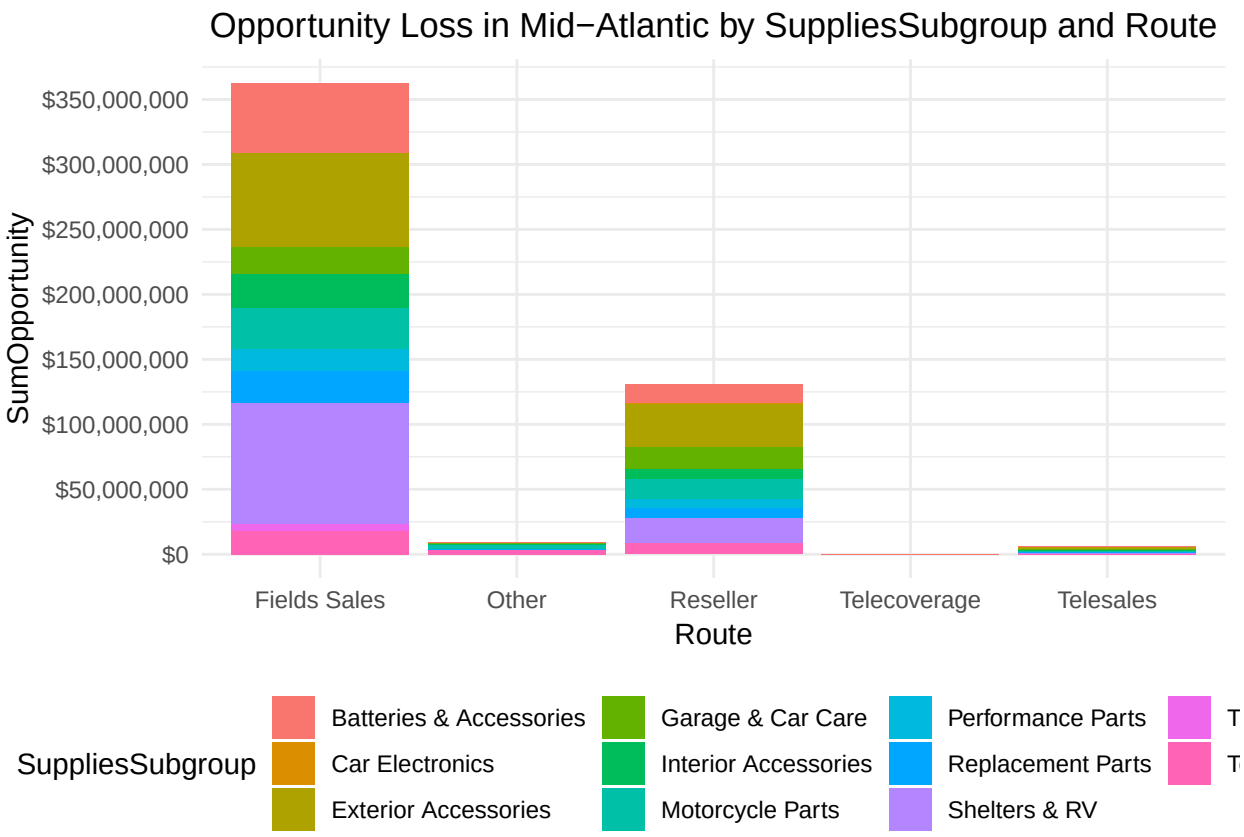
I can see that the probability of loss opportunity is higher if customer didn't buy anything in the last two years. If client purchase in the last two years, the chance of win decreases as sales deals rises

```
sales_win_loss %>%
  group_by(Region, Route) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Region, y = SumOpportunity, fill = Route)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0,1e+11, 2e+08), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Market Opportunity by Region and Route last two years") +
  theme(plot.title = element_text(hjust = 0.5))
```



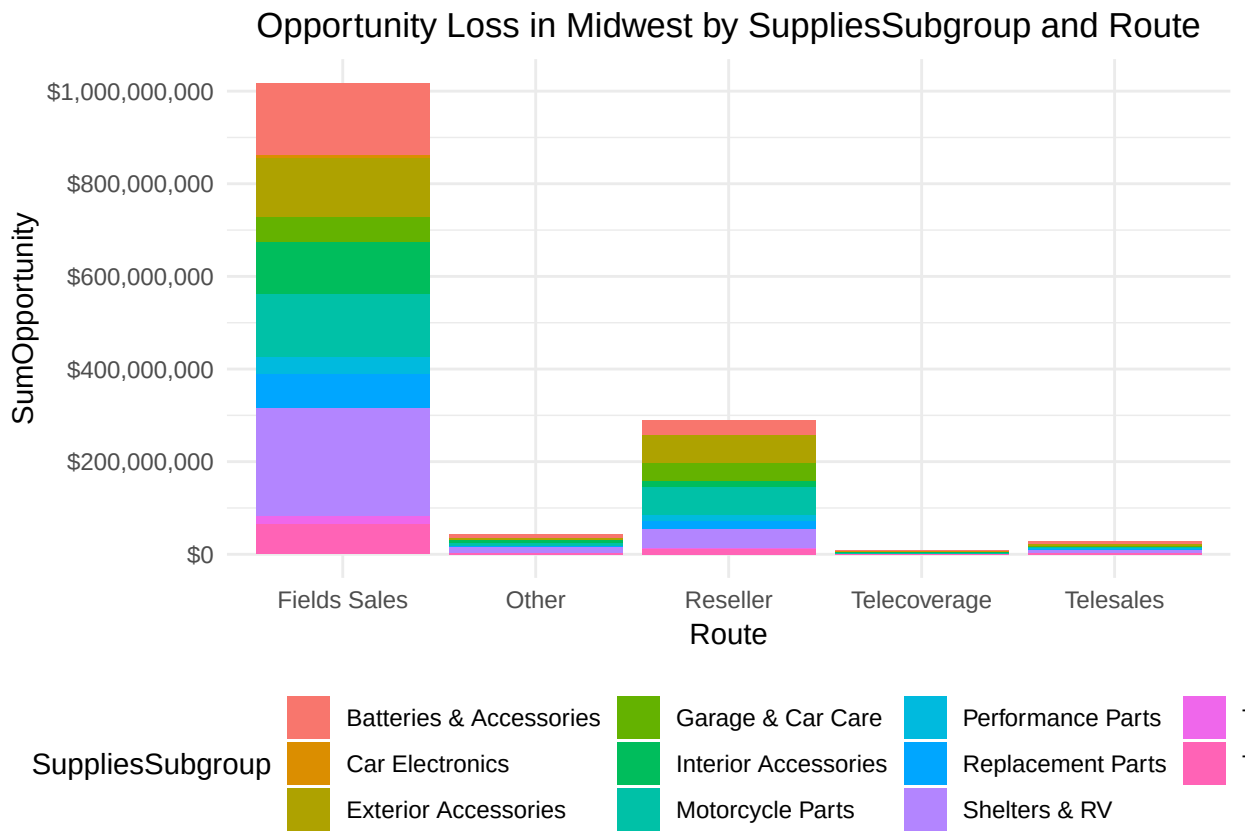
From the chart above, it's clear that the field sales and reseller are the most common channel of sales route to market across all regions. In the pacific region, other sales channel also play some role to bring in revenue in comparison to other regions. I want to see the breakdown of the 'Loss' opportunity by each region to dig deeper. The subsequent charts will show the break down of the supplies subgroup for each region where sales lead is 'Loss'

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Mid-Atlantic") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 4e+08, 5e+07), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Mid-Atlantic by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



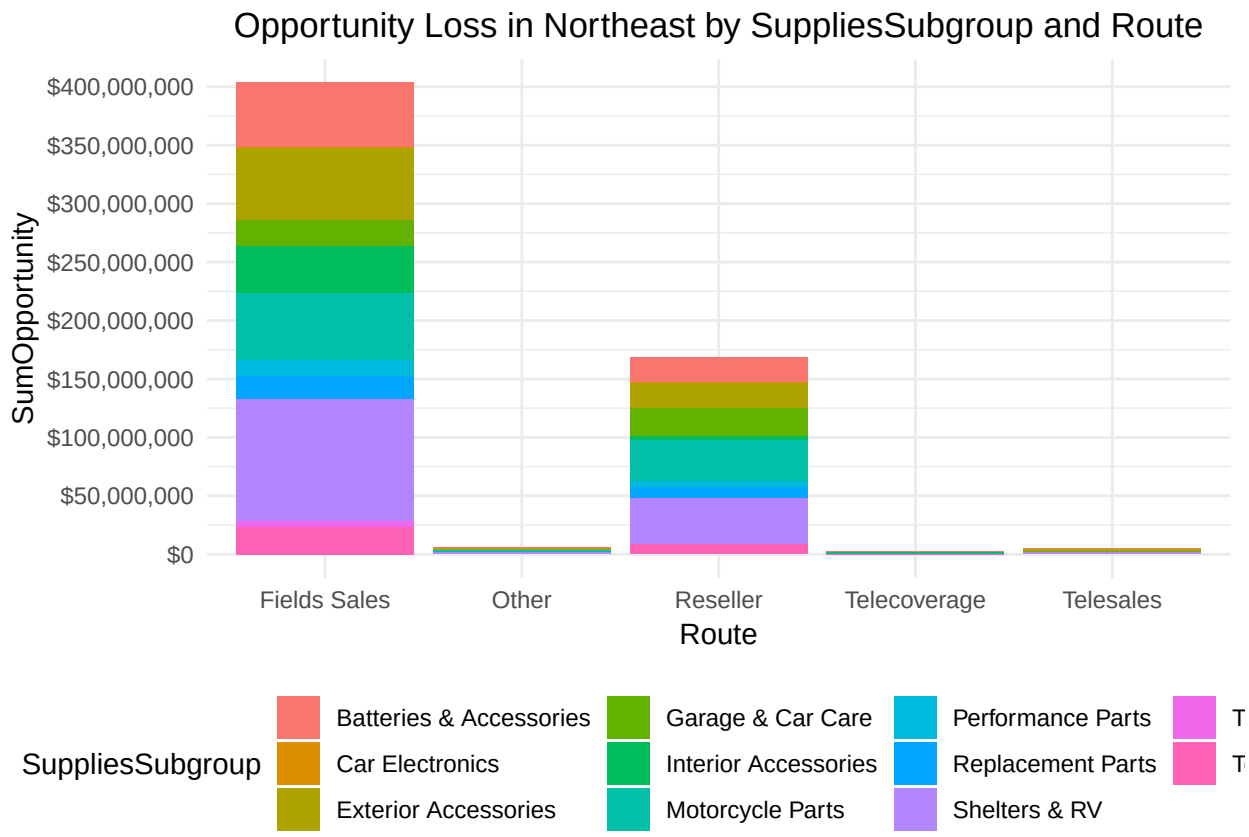
The chart above show the Mid-Atlantic region breakdown of loss opportunity by supplies subgroup. In this region, “Shelters & RV”, Batteries & Accessories“, and”Exterior Accessories" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Midwest") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 1.5e+09, 2e+08), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Midwest by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



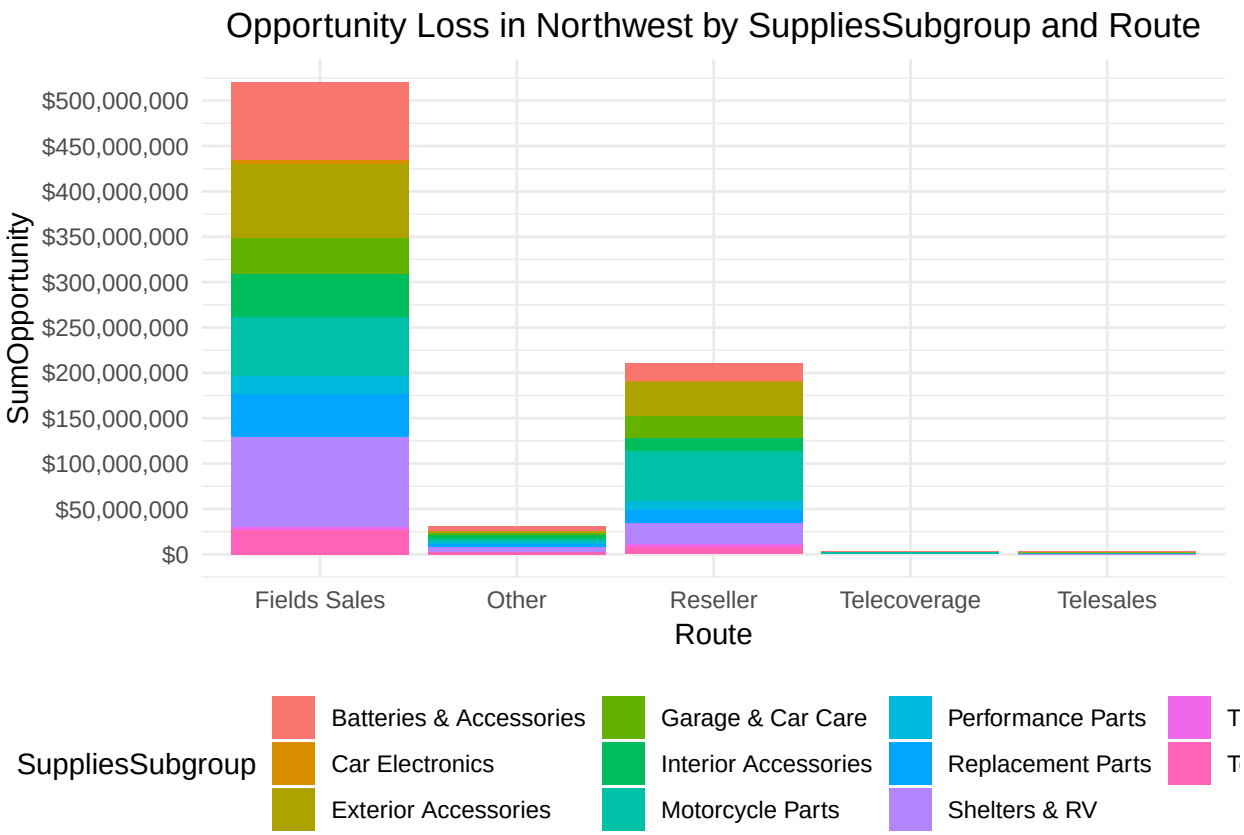
The chart above show the MidWest region breakdown of loss opportunity by supplies subgroup. In this region, “Shelters & RV”, Batteries & Accessories“, and”Exterior Accessories" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Northeast") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 5e+08, 5e+07), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Northeast by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



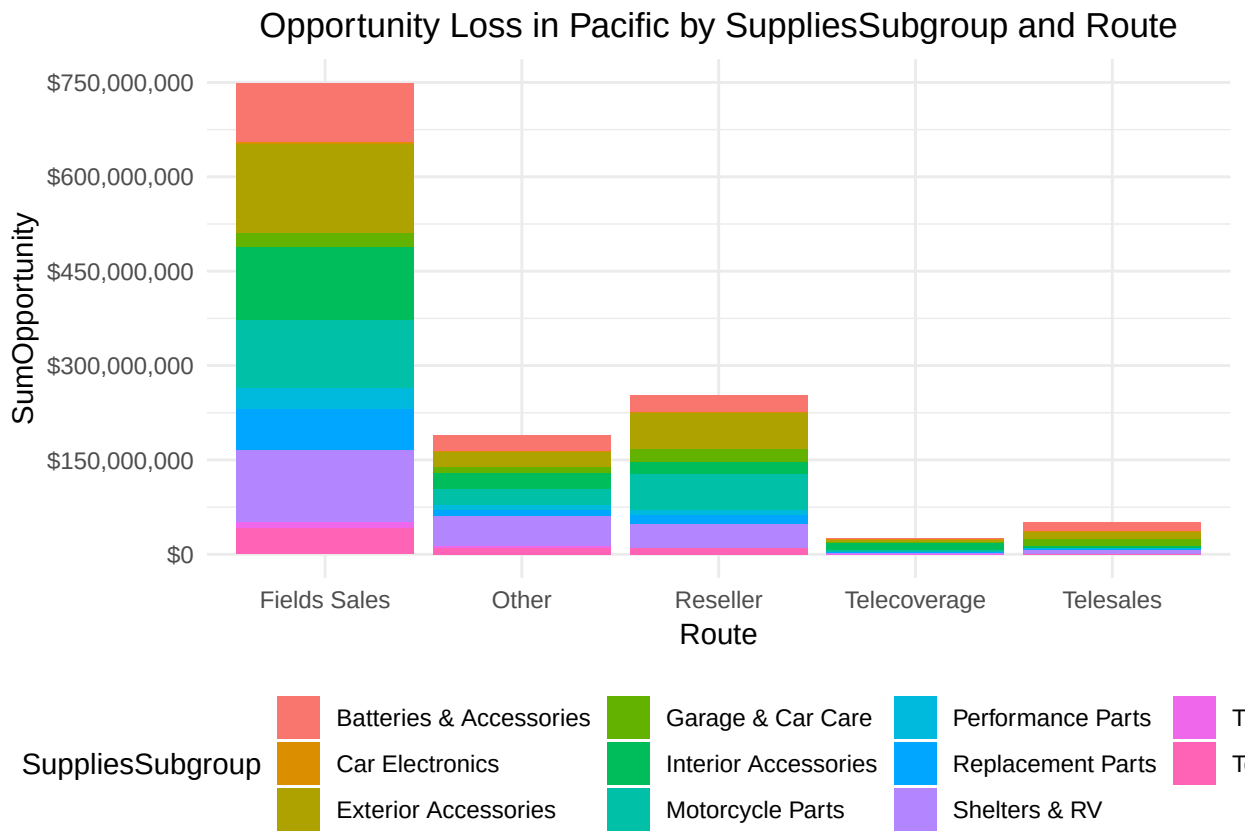
The chart above show the Northeast region breakdown of loss opportunity by supplies subgroup. In this region, "Shelters & RV" and "Motorcycle Parts"" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Northwest") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 1e+11, 5e+07), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Northwest by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



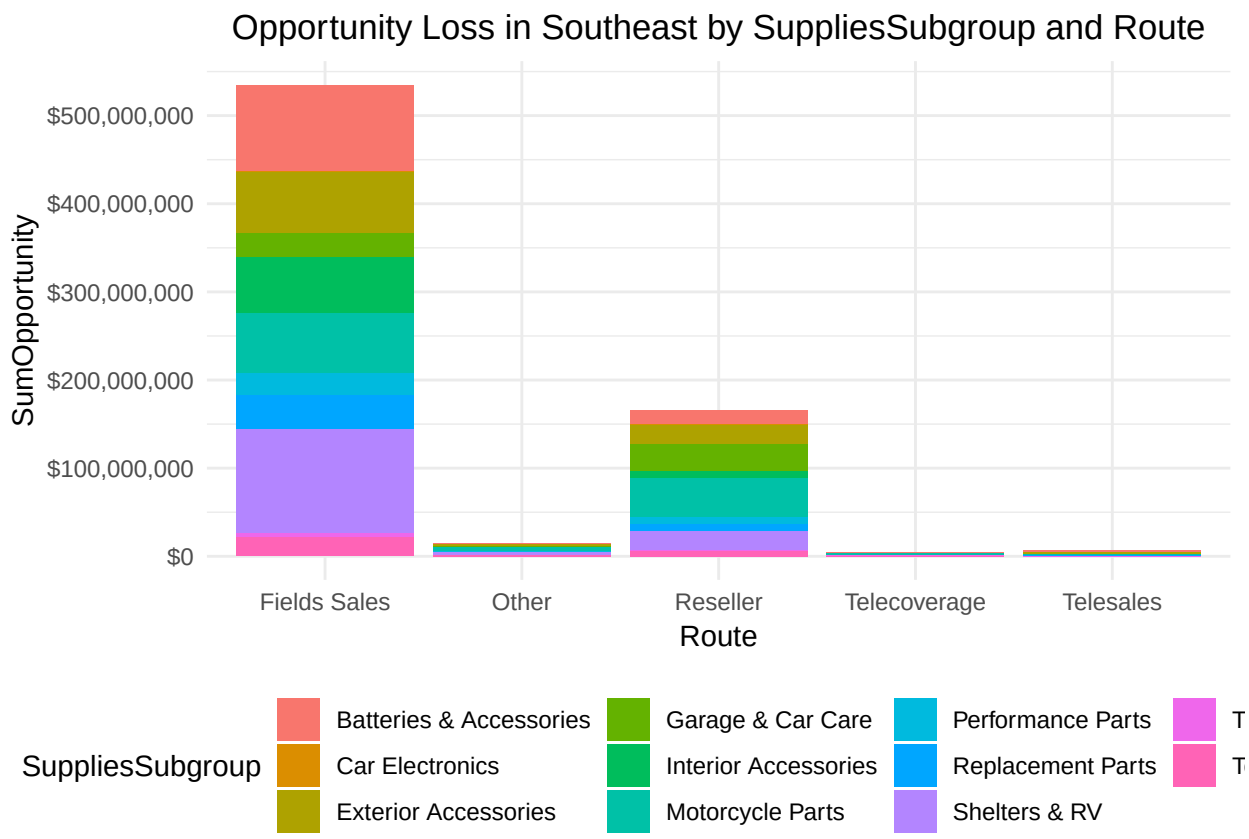
The chart above show the Northwest region breakdown of loss opportunity by supplies subgroup. In this region, “Shelters & RV”, Batteries & Accessories“, and”Exterior Accessories" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Pacific") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 1e+11, 1.5e+08), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Pacific by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



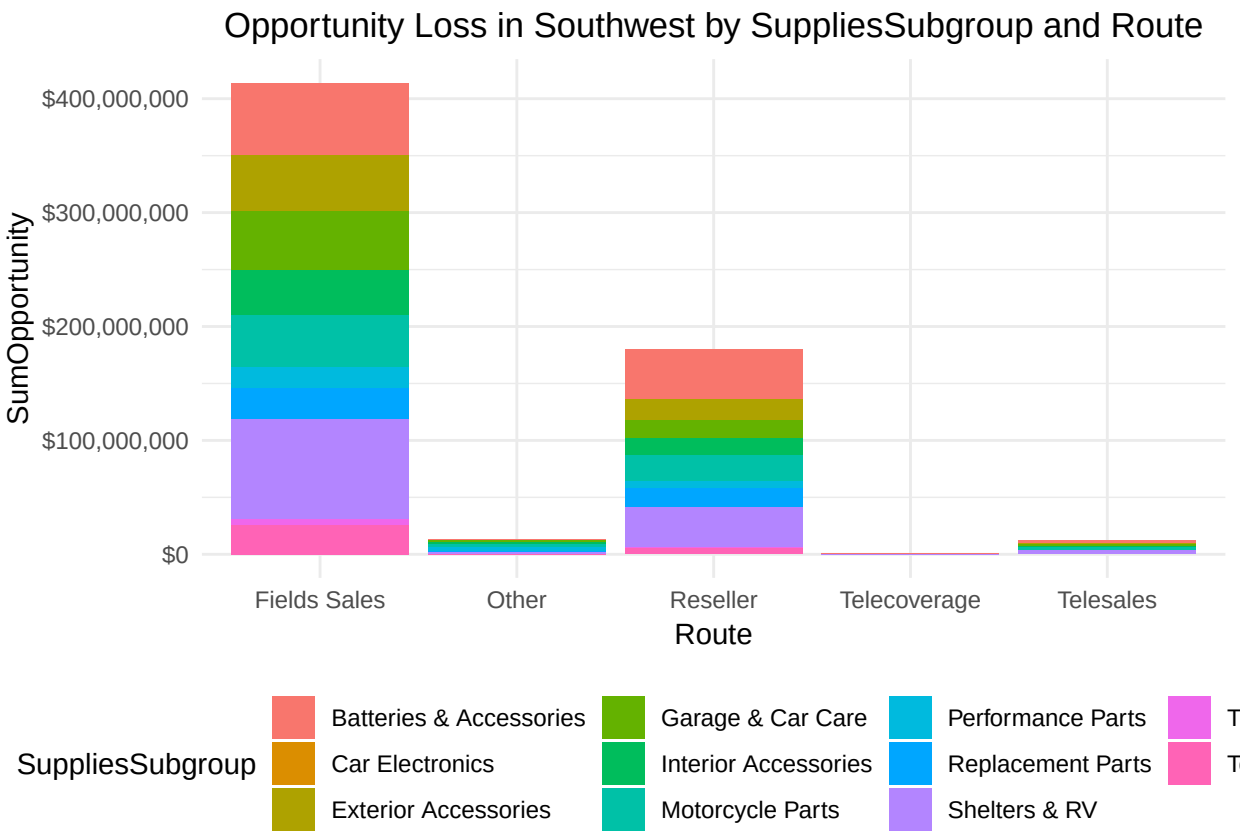
The chart above show the Pacific region breakdown of loss opportunity by supplies subgroup. In this region, “Shelters & RV”, Batteries & Accessories“, and”Exterior Accessories" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Southeast") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 1e+11, 1e+08), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Southeast by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



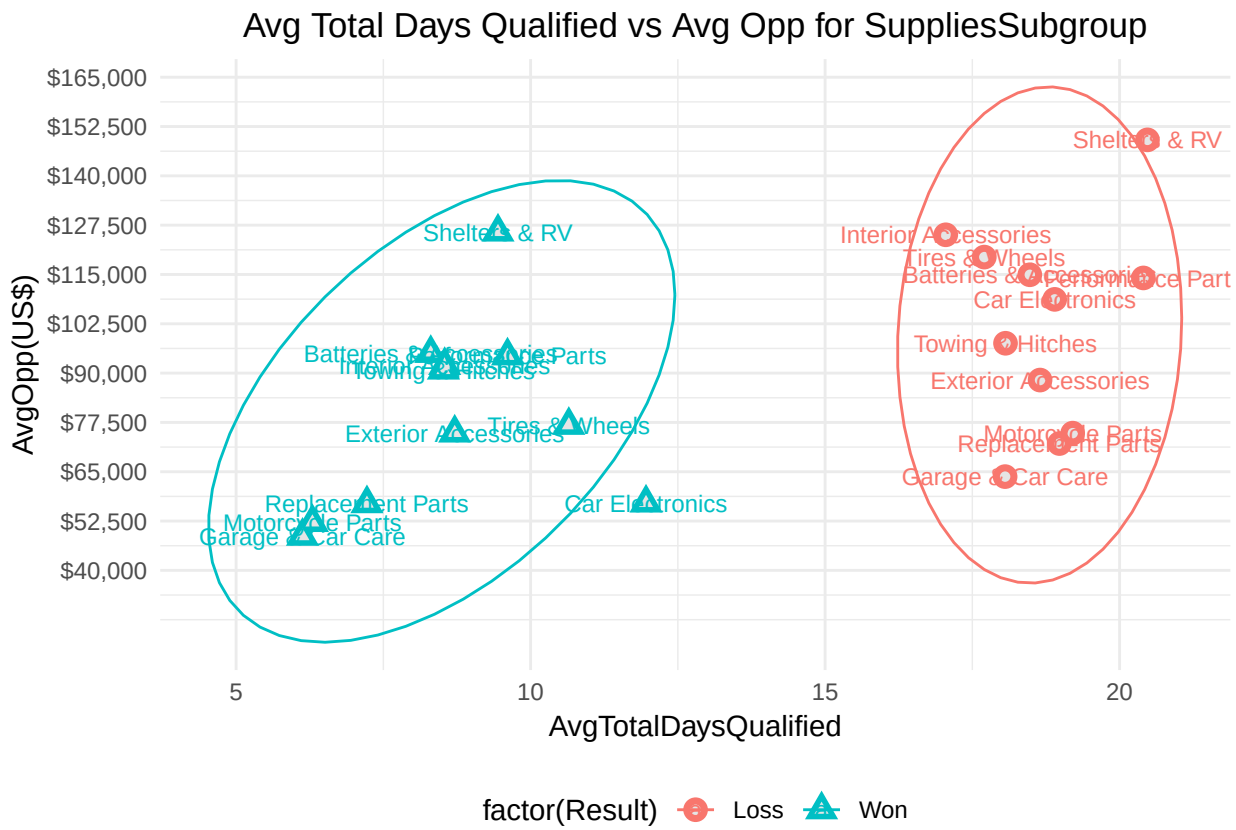
The chart above show the MidWest region breakdown of loss opportunity by supplies subgroup. In this region, “Shelters & RV”, Batteries & Accessories“, and”Exterior Accessories" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>% filter(Result == "Loss" & Region == "Southwest") %>%
  group_by(Route, SuppliesSubgroup) %>%
  summarise(SumOpportunity = sum(Opportunity)) %>%
  ggplot(aes(x = Route, y = SumOpportunity, fill = SuppliesSubgroup)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 1e+11, 1e+08), labels = scales::dollar_format(prefix = "$")) +
  ggtitle("Opportunity Loss in Southwest by SuppliesSubgroup and Route") +
  theme(plot.title = element_text(hjust = 0.5))
```



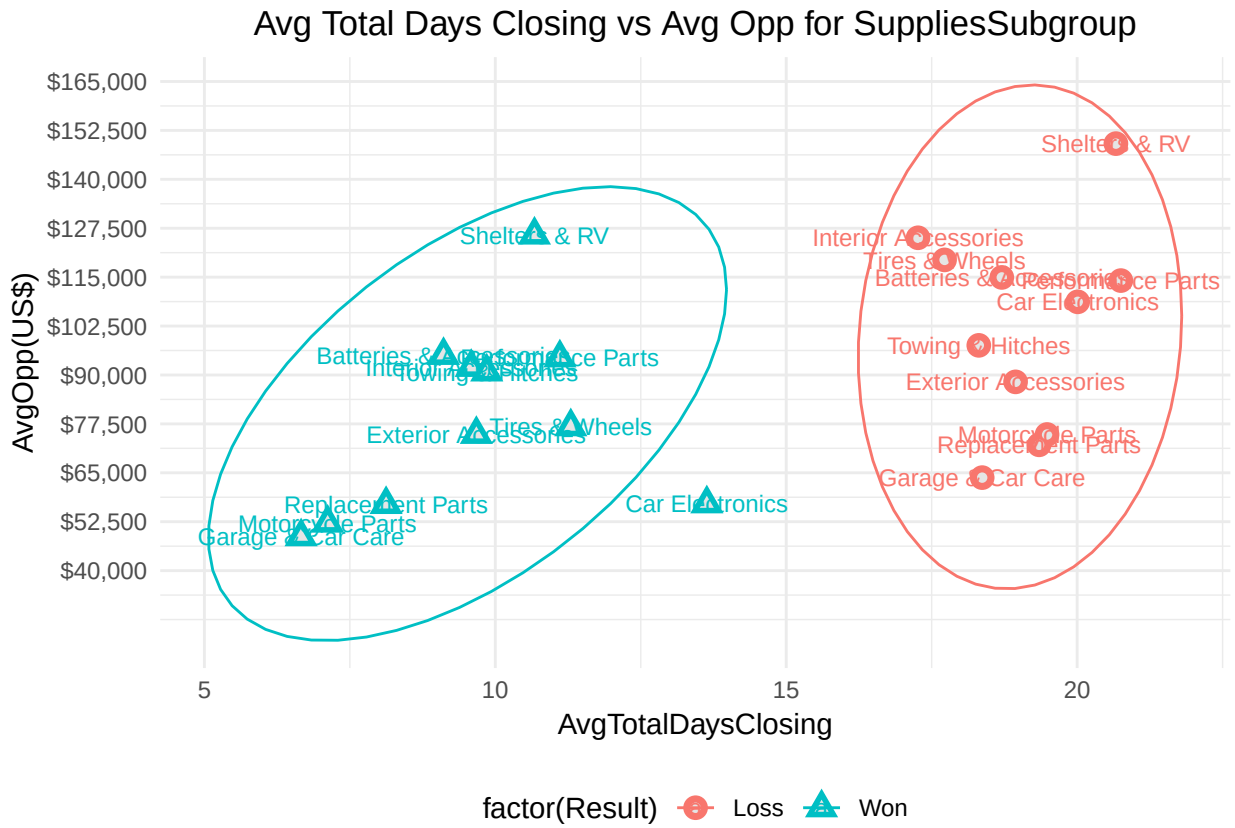
The chart above show the SouthWest region breakdown of loss opportunity by supplies subgroup. In this region, “Shelters & RV”, Batteries & Accessories“, and”Exterior Accessories" are the main supplies subgroup that contribute to the loss opportunity.

```
sales_win_loss %>%
  group_by(Result, SuppliesSubgroup) %>%
  summarise(AvgOpp = mean(Opportunity), AvgQual = mean(TotalDaysQualified )) %>%
  ggplot(aes(x = AvgQual, y = AvgOpp, shape = factor(Result), label = SuppliesSubgroup)) +
  geom_point(aes(colour = factor(Result)), size = 4) +
  geom_point(colour = "grey90", size = 1.5) + xlab("AvgTotalDaysQualified") + ylab("AvgOpp(US$)") +
  scale_x_continuous(breaks = seq(0, 30, 5)) +
  scale_y_continuous(breaks = seq(40000, 175000, 12500), labels = scales::dollar_format(prefix = "$")) +
  geom_text(aes(label = SuppliesSubgroup,  color = Result), size = 3) +
  stat_ellipse(aes(color = Result), type = "t") +
  ggtitle("Avg Total Days Qualified vs Avg Opp for SuppliesSubgroup") +
  theme(plot.title = element_text(hjust = 0.5))
```



Looking from left to right, this Scatter chart shows that irrespective of opportunity amounts, we start losing deals as they stay longer in the pipeline. This could help formulate threshold levels for each supplier based on how many days a deal is in the pipeline and create alert mechanisms to expedite its progression.

```
sales_win_loss %>%
  group_by(Result, SuppliesSubgroup) %>%
  summarise(AvgOpp = mean(Opportunity), AvgQual = mean(TotalDaysClosing )) %>%
  ggplot(aes(x = AvgQual, y = AvgOpp, shape = factor(Result), label = SuppliesSubgroup)) +
  geom_point(aes(colour = factor(Result)), size = 4) +
  geom_point(colour = "grey90", size = 1.5) + xlab("AvgTotalDaysClosing") + ylab("AvgOpp(US$)") +
  scale_x_continuous(breaks = seq(0, 30, 5)) +
  scale_y_continuous(breaks = seq(40000, 175000, 12500), labels = scales::dollar_format(prefix = "$")) +
  geom_text(aes(label = SuppliesSubgroup,  color = Result), size = 3) +
  stat_ellipse(aes(color = Result), type = "t") +
  ggtitle("Avg Total Days Closing vs Avg Opp for SuppliesSubgroup") +
  theme(plot.title = element_text(hjust = 0.5))
```



Revenue calculations: is it before or after closing

Clustering, select independent variables, don't select the dependent variable

Run clustering algorithm

its going to return a vector given name cluster_id added to data frame

convert cluster_id column to factor using the factor function

Modeling

Create modeling data

How to setup logistic regression, and random forest for classification problems?

How to setup clustering k-means for this problem?