JDBC Comparison

Wendy L.

December 11, 2017

This is an exercise to validate the results of EDW data extract using JDBC. We compared the Week 44 SOTC OTS results using both ODBC and JDBC methods. The objective is to make sure JDBC extraction generates the same results as ODBC so we can adopt JDBC method in the future as it's faster and more efficient.

Load libraries and saved objects

```
suppressMessages(library(ggplot2))
suppressMessages(library(dplyr))
suppressMessages(library(scales))

load("OTS_Master_JDBC_wk44.rds")
load("OTS_Master_object.rds")
load("SOT_Master_JDBC_wk44.rds")
load("SOT_Master_object.rds")
```

SOT_Master

1.Compare dimensions of the two SOT_Master dfs

```
dim(SOT_Master)
```

```
## [1] 1374154      43
```

```
dim(SOT_Master_JDBC)
```

```
## [1] 1374154      43
```

2. Compare summary statistics of the numeric attributes

```
summary(SOT_Master[c('Units','DAYS_LATE')])
```

```
##      Units              DAYS_LATE     
##  Min.   :      0.0   Min.   :-3653.00  
##  1st Qu.:     80.0   1st Qu.:    0.00  
##  Median :    292.0   Median :    0.00  
##  Mean   :    842.3   Mean   :    1.29  
##  3rd Qu.:    854.0   3rd Qu.:    1.00  
##  Max.   :1200000.0   Max.   :  738.00  
##                      NA's   :95405
```

```
summary(SOT_Master_JDBC[c('Units','DAYS_LATE')])
```

```
##      Units              DAYS_LATE     
##  Min.   :      0.0   Min.   :-3653.00
##  1st Qu.:     80.0   1st Qu.:   0.00  
##  Median :    292.0   Median :   0.00  
##  Mean   :    842.3   Mean   :   1.29  
##  3rd Qu.:    854.0   3rd Qu.:   1.00  
##  Max.   :1200000.0   Max.   : 738.00  
##                      NA's   :95405
```

3.Compare aggregated SOT results by brand by category

```r
SOT_Master_JDBC$ReportingBrand <- as.factor(SOT_Master_JDBC$ReportingBrand)
SOT_Master_JDBC$Category <- as.factor(SOT_Master_JDBC$Category)
SOTbyBrandODBC <- SOT_Master %>%
  group_by(ReportingBrand, Category) %>%
  summarise(n = n(), sumUnits <- sum(Units)) %>%
  arrange(n)
SOTbyBrandODBC
```

```
## # A tibble: 101 x 4
## # Groups:   ReportingBrand [14]
##    ReportingBrand                Category     n `sumUnits <- sum(Units)`
##            <fctr>                  <fctr> <int>                    <dbl>
##  1          GO NA          Category Other     4                     2500
##  2      PIPERLIME          Category Other     5                      576
##  3        GO INTL                3P & Lic    14                     7728
##  4        ATHLETA Denim and Woven Bottoms    22                    70968
##  5        ON INTL                3P & Lic    66                    53754
##  6  GAP FRANCHISE                3P & Lic    79                    18348
##  7   BR FRANCHISE                3P & Lic   103                     2063
##  8          BR NA                      IP   103                    13874
##  9        BRFS NA                3P & Lic   127                   145460
## 10   ON FRANCHISE                Sweaters   192                    10371
## # ... with 91 more rows
```

```r
SOTbyBrandJDBC <- SOT_Master_JDBC %>%
  group_by(ReportingBrand, Category) %>%
  summarise(n = n(), sumUnits <- sum(Units)) %>%
  arrange(n)
SOTbyBrandJDBC
```

```
## # A tibble: 101 x 4
## # Groups:   ReportingBrand [14]
##    ReportingBrand                Category     n `sumUnits <- sum(Units)`
##            <fctr>                  <fctr> <int>                    <dbl>
##  1          GO NA          Category Other     4                     2500
##  2      PIPERLIME          Category Other     5                      576
##  3        GO INTL                3P & Lic    14                     7728
##  4        ATHLETA Denim and Woven Bottoms    22                    70968
##  5        ON INTL                3P & Lic    66                    53754
##  6  GAP FRANCHISE                3P & Lic    79                    18348
##  7   BR FRANCHISE                3P & Lic   103                     2063
##  8          BR NA                      IP   103                    13874
##  9        BRFS NA                3P & Lic   127                   145460
## 10   ON FRANCHISE                Sweaters   192                    10371
## # ... with 91 more rows
```

Are the aggregated results from ODBC and JDBC the same?

```
all.equal(SOTbyBrandODBC, SOTbyBrandJDBC)
```

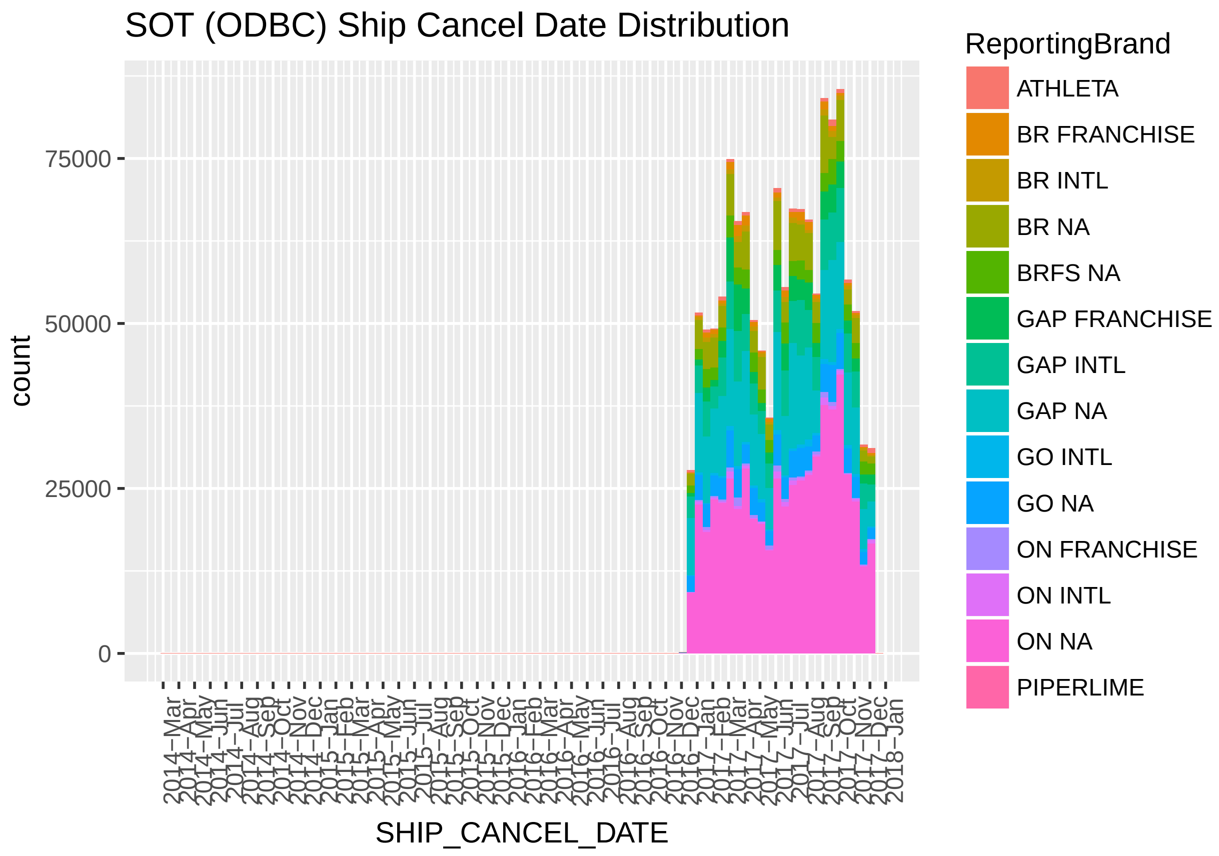
```
## [1] TRUE
```

4.Compare ship cancel date distribution

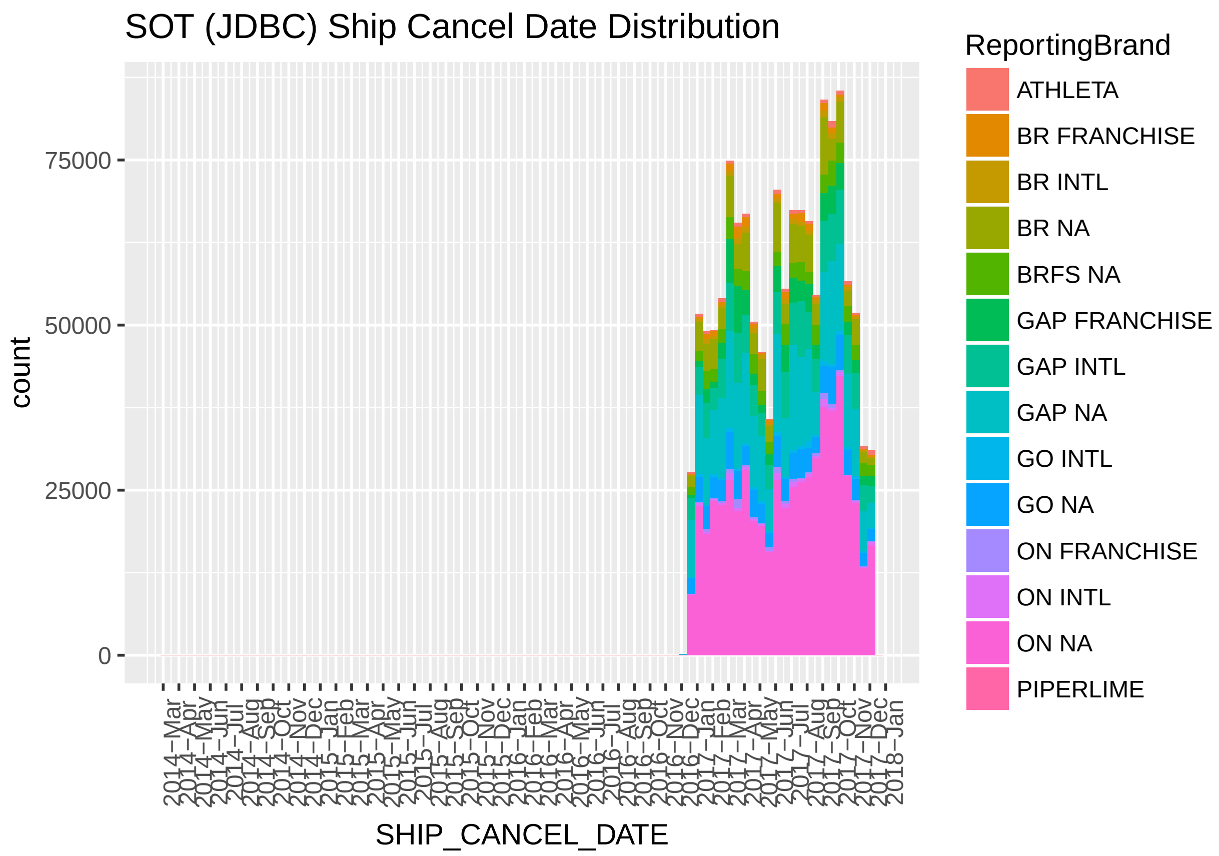
The two histograms show the same Ship Cancel Date distribution between the two dfs.

```
SOT_Master_JDBC$SHIP_CANCEL_DATE <- as.Date(SOT_Master_JDBC$SHIP_CANCEL_DATE)

ggplot(SOT_Master, aes(x = SHIP_CANCEL_DATE, fill=ReportingBrand)) +
  geom_histogram(binwidth = 15) +
  scale_x_date(labels = date_format("%Y-%b"),
               breaks = seq(min(SOT_Master$SHIP_CANCEL_DATE)-5, max(SOT_Master$SHIP_CANCEL_DATE)+5, 30)
  theme(axis.text.x = element_text(angle=90)) +
  ggtitle("SOT (ODBC) Ship Cancel Date Distribution")
```



```
ggplot(SOT_Master_JDBC, aes(x = SHIP_CANCEL_DATE, fill=ReportingBrand)) +
  geom_histogram(binwidth = 15) +
  scale_x_date(labels = date_format("%Y-%b"),
               breaks = seq(min(SOT_Master_JDBC$SHIP_CANCEL_DATE)-5, max(SOT_Master_JDBC$SHIP_CANCEL_DA
  theme(axis.text.x = element_text(angle=90)) +
  ggtitle("SOT (JDBC) Ship Cancel Date Distribution")
```



OTS_Master

1.Compare dimensions of the two SOT_Master dfs

```
dim(OTS_Master)
```

```
## [1] 1374154      34
```

```
dim(OTS_Master_JDBC)
```

```
## [1] 1374154      34
```

2. Compare summary statistics of the numeric attributes

```
summary(OTS_Master[c('Units','FCST_QTY','ACTL_STK_QTY','Days_Late')])
```

```
##      Units              FCST_QTY          ACTL_STK_QTY
##  Min.   :      0.0   Min.   :      0.0   Min.   :     0.0
##  1st Qu.:     80.0   1st Qu.:     79.0   1st Qu.:   109.0
##  Median :    292.0   Median :    286.0   Median :   345.0
##  Mean   :    842.3   Mean   :    852.1   Mean   :   887.7
##  3rd Qu.:    854.0   3rd Qu.:    849.0   3rd Qu.:   968.0
##  Max.   :1200000.0   Max.   :2431569.0   Max.   :237471.0
##                                          NA's   :205716
##    Days_Late
##  Min.   :-72.00
##  1st Qu.: -5.00
##  Median : -2.00
##  Mean   : 11.61
##  3rd Qu.:  3.00
##  Max.   :476.00
##
```

```
summary(OTS_Master_JDBC[c('Units','FCST_QTY','ACTL_STK_QTY','Days_Late')])
```

```
##      Units              FCST_QTY          ACTL_STK_QTY
##  Min.   :      0.0   Min.   :      0.0   Min.   :     0.0
##  1st Qu.:     80.0   1st Qu.:     79.0   1st Qu.:   109.0
##  Median :    292.0   Median :    286.0   Median :   345.0
##  Mean   :    842.3   Mean   :    852.1   Mean   :   887.7
##  3rd Qu.:    854.0   3rd Qu.:    849.0   3rd Qu.:   968.0
##  Max.   :1200000.0   Max.   :2431569.0   Max.   :237471.0
##                                          NA's   :205716
##    Days_Late
##  Min.   :-72.00
##  1st Qu.: -5.00
##  Median : -2.00
##  Mean   : 11.61
##  3rd Qu.:  3.00
##  Max.   :476.00
##
```

3.Compare aggregated results by brand by category

```
OTS_Master_JDBC$ReportingBrand <- as.factor(OTS_Master_JDBC$ReportingBrand)
OTS_Master_JDBC$Category <- as.factor(OTS_Master_JDBC$Category)

OTSbyBrandODBC <- OTS_Master %>%
  group_by(ReportingBrand, Category) %>%
  summarise(n = n(), sumUnits <- sum(Units)) %>%
  arrange(n)
OTSbyBrandODBC
```

```
## # A tibble: 101 x 4
## # Groups:   ReportingBrand [14]
##    ReportingBrand                Category     n `sumUnits <- sum(Units)`
##            <fctr>                  <fctr> <int>                    <dbl>
##  1          GO NA          Category Other     4                     2500
##  2      PIPERLIME          Category Other     5                      576
##  3        GO INTL                3P & Lic    14                     7728
##  4        ATHLETA Denim and Woven Bottoms    22                    70968
##  5        ON INTL                3P & Lic    66                    53754
##  6  GAP FRANCHISE                3P & Lic    79                    18348
##  7   BR FRANCHISE                3P & Lic   103                     2063
##  8          BR NA                      IP   103                    13874
##  9        BRFS NA                3P & Lic   127                   145460
## 10   ON FRANCHISE                Sweaters   192                    10371
## # ... with 91 more rows
```

```
OTSbyBrandJDBC <- OTS_Master_JDBC %>%
  group_by(ReportingBrand, Category) %>%
  summarise(n = n(), sumUnits <- sum(Units)) %>%
  arrange(n)
OTSbyBrandJDBC
```

```
## # A tibble: 101 x 4
## # Groups:   ReportingBrand [14]
##    ReportingBrand                Category     n `sumUnits <- sum(Units)`
##            <fctr>                  <fctr> <int>                    <dbl>
##  1          GO NA          Category Other     4                     2500
##  2      PIPERLIME          Category Other     5                      576
##  3        GO INTL                3P & Lic    14                     7728
##  4        ATHLETA Denim and Woven Bottoms    22                    70968
##  5        ON INTL                3P & Lic    66                    53754
##  6  GAP FRANCHISE                3P & Lic    79                    18348
##  7   BR FRANCHISE                3P & Lic   103                     2063
##  8          BR NA                      IP   103                    13874
##  9        BRFS NA                3P & Lic   127                   145460
## 10   ON FRANCHISE                Sweaters   192                    10371
## # ... with 91 more rows
```

Are the aggregated results from ODBC and JDBC the same?

```
all.equal(OTSbyBrandODBC, OTSbyBrandJDBC)
```

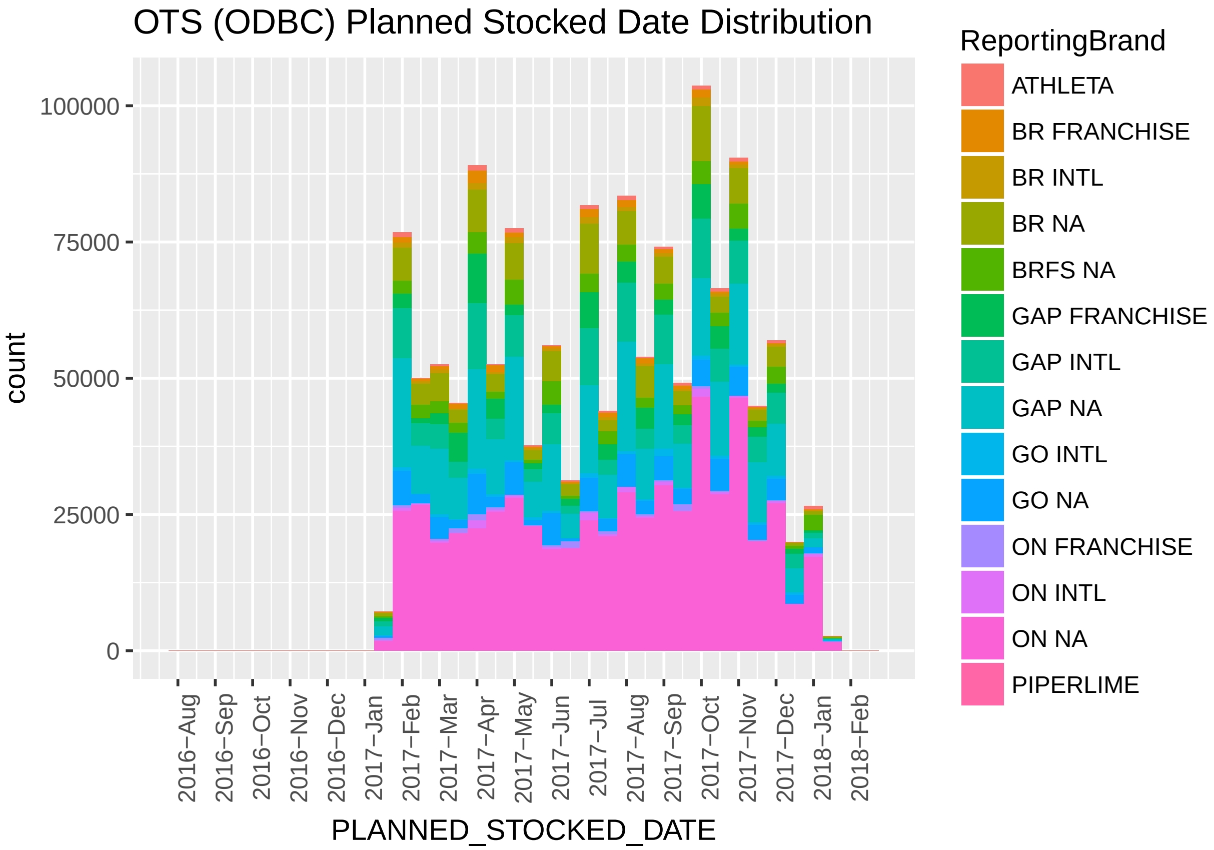
```
## [1] TRUE
```

4.Compare planned stocked date distribution

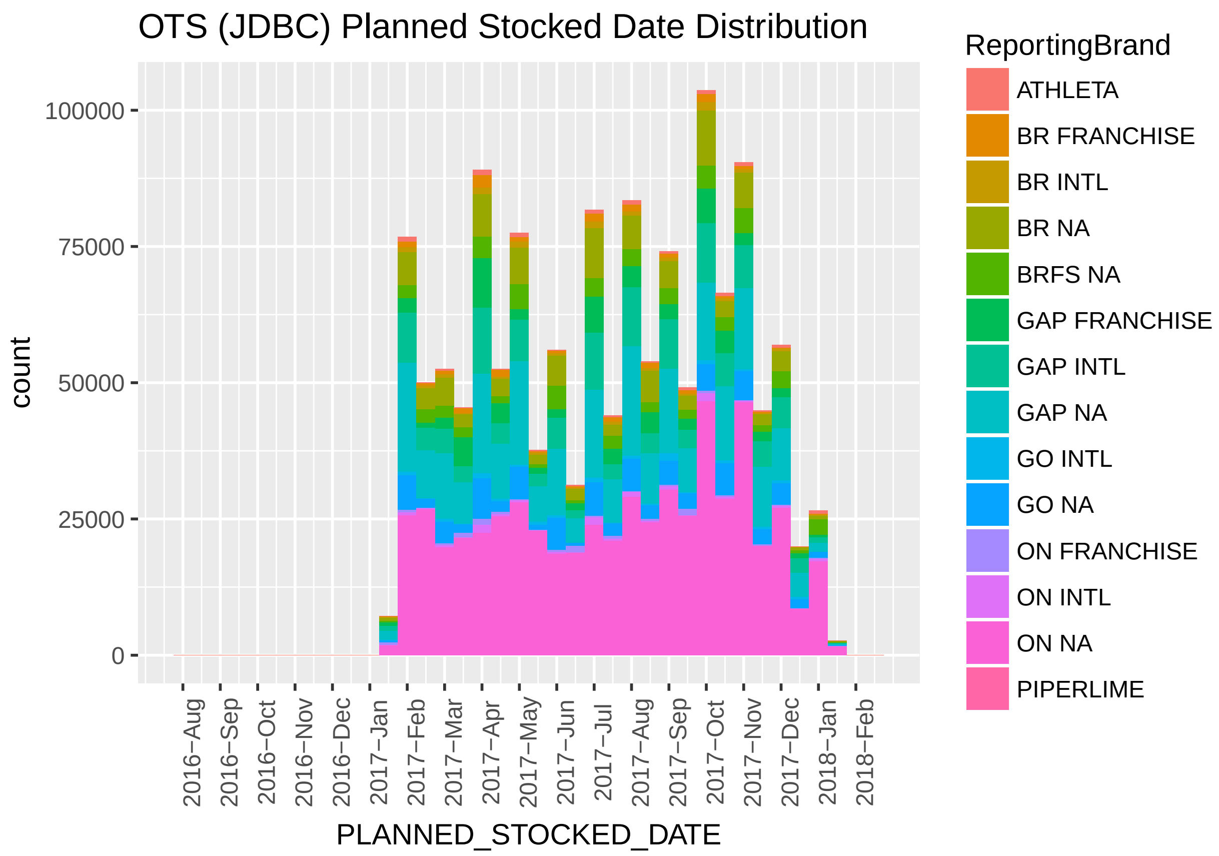
The two histograms show the same Planned Stocked Date distribution between the two dfs.

```
OTS_Master_JDBC$PLANNED_STOCKED_DATE <- as.Date(OTS_Master_JDBC$PLANNED_STOCKED_DATE)

ggplot(OTS_Master, aes(x = PLANNED_STOCKED_DATE, fill=ReportingBrand)) +
  geom_histogram(binwidth = 15) +
  scale_x_date(labels = date_format("%Y-%b"),
               breaks = seq(min(OTS_Master$PLANNED_STOCKED_DATE)-5, max(OTS_Master$PLANNED_STOCKED_DATE
  theme(axis.text.x = element_text(angle=90)) +
  ggtitle("OTS (ODBC) Planned Stocked Date Distribution")
```



```r
ggplot(OTS_Master_JDBC, aes(x = PLANNED_STOCKED_DATE, fill=ReportingBrand)) +
  geom_histogram(binwidth = 15) +
  scale_x_date(labels = date_format("%Y-%b"),
               breaks = seq(min(OTS_Master_JDBC$PLANNED_STOCKED_DATE)-5, max(OTS_Master_JDBC$PLANNED_STO
  theme(axis.text.x = element_text(angle=90)) +
  ggtitle("OTS (JDBC) Planned Stocked Date Distribution")
```


OTS (JDBC) Planned Stocked Date Distribution
count
100000
75000
50000
25000
0
2016−Aug
2016−Sep
2016−Oct
2016−Nov
2016−Dec
2017−Jan
2017−Feb
2017−Mar
2017−Apr
2017−May
2017−Jun
2017−Jul
2017−Aug
2017−Sep
2017−Oct
2017−Nov
2017−Dec
2018−Jan
2018−Feb
PLANNED_STOCKED_DATE
ReportingBrand
ATHLETA
BR FRANCHISE
BR INTL
BR NA
BRFS NA
GAP FRANCHISE
GAP INTL
GAP NA
GO INTL
GO NA
ON FRANCHISE
ON INTL
ON NA
PIPERLIME