Data Analyst Take-Home Assignment

Yijing Cai

Libraries or packages used: ggplot2, gridExtra for plot, caret for evaluate models.

Data Description:

As the output below shows, the copackager table has 1200 observations and 10 columns, which means 1200 packaging jobs. Among them, 600 jobs' customer are Procter & Gamble and 600 jobs' customer are Unilever. 866 jobs are On-Time In-Full (OTIF) while 334 jobs are not OTIF.

Dimensions of data and descriptions of columns:

```
## 'data.frame':    1200 obs. of  10 variables:
##  $ job.id                       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ purchase.order.received.date: Factor w/ 630 levels "17-01-01 17:22",..: 629 588 5..
##  $ materials.availablity.date  : Factor w/ 629 levels "17-01-03 3:01",..: 629 586 50..
##  $ production.started.date     : Factor w/ 629 levels "17-01-03 13:37",..: 629 589 5..
##  $ production.completed.date   : Factor w/ 629 levels "17-01-03 23:47",..: 629 587 5..
##  $ quantity.produced           : int  110 102 81 429 489 761 699 862 94 773 ...
##  $ unit.of.measure             : Factor w/ 3 levels "cases","eaches",..: 3 3 3 1 1 2..
##  $ shipment.shipped.date       : Factor w/ 629 levels "17-01-05 8:45",..: 629 587 50..
##  $ OTIF                        : int  0 1 0 0 1 0 1 0 0 1 ...
##  $ customer                    : Factor w/ 2 levels "Procter & Gamble",..: 2 2 1 2 1..
```

Counts by OTIF or job customer:

```
##  OTIF.factor             customer
##  0:334       Procter & Gamble:600
##  1:866       Unilever        :600
```

Cleaning data:

1) I checked for missing value and there is none.
2) The values in columns where dates and times are stored are converted to date/time objects for calculation purpose. The values in OTIF column are converted from integer to factors to be treated as categorical variable.
3) Erroneous values were corrected by removing all the observations with purchase order received date later than materials availability date, or materials availability date later than production started date, or production started date later or equal to production completed date, or production completed date later or equal to shipment shipped date. After cleaning, there are 1101 observations left.

Dimensions of cleaned data:

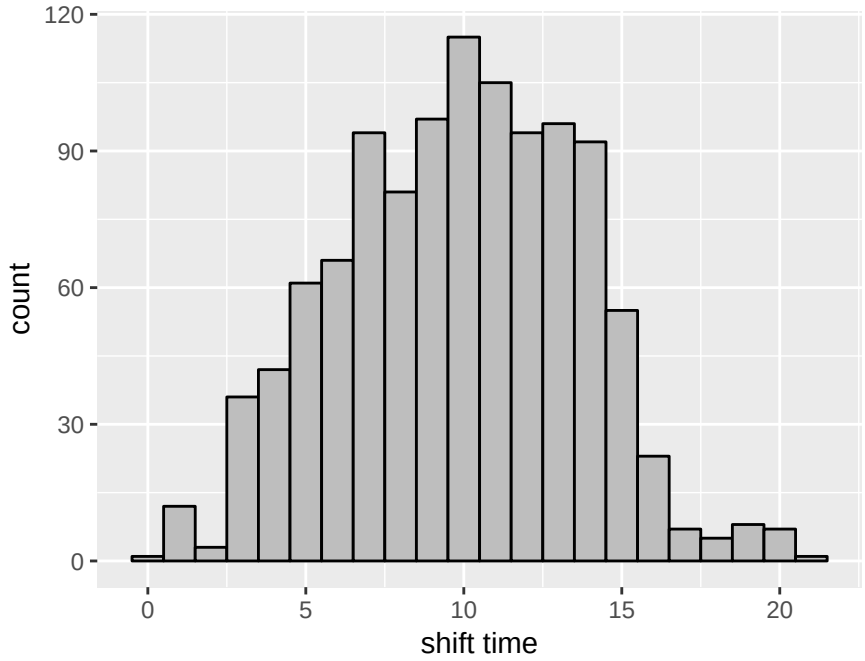
```
## [1] 1101   11
```

Counts by OTIF or job customer for cleaned data:

```
##  OTIF.factor             customer
##  0:319       Procter & Gamble:547
##  1:782       Unilever        :554
```

Q1: What is the average shift length?

Histogram for Shift Time:



Code output for mean shift time:

```
## [1] 9.866409
```

Answer: 9.866409 ≈ 10hrs

Q2.1: What is the change in probability of OTIF 3 days after receiving the PO vs 4 days?

Assumption: 1) Based on description file, we can assume that the job is always completed in full. 2) Observations are independent. 3) Effects of factors other than the time from receiving the PO to shipment shipped are not considered.

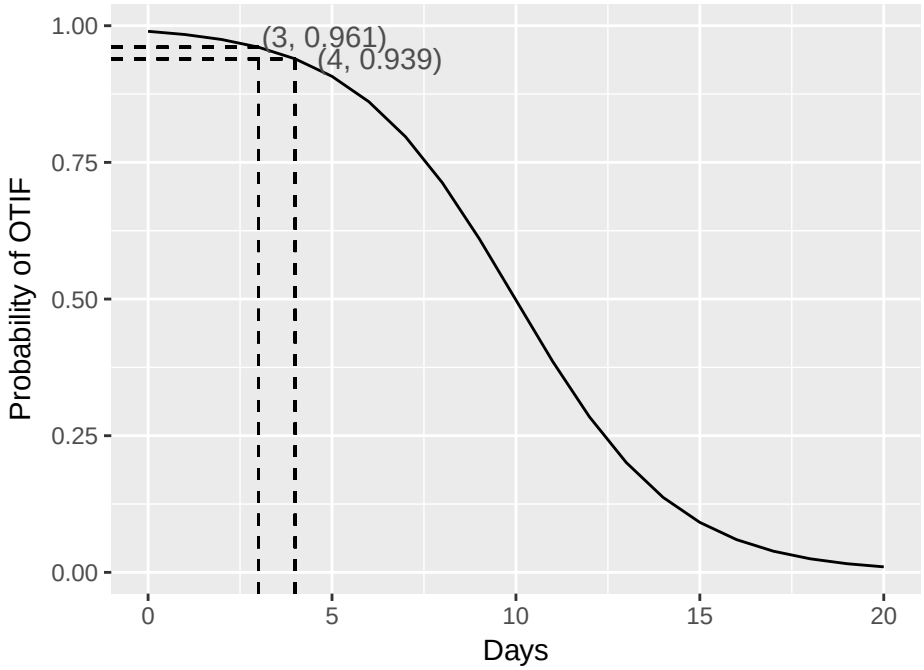
Model for predicting OTIF based on the time from receiving the PO to shipment shipped:

```
##
## Call:
## glm(formula = OTIF.factor ~ cycle.time, family = binomial(logit),
##     data = cleanCopackager)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4944  -1.0100   0.5907   0.8160   1.5366
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.57082    0.35235   12.97   <2e-16 ***
## cycle.time  -0.45780    0.04162  -11.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1325.4  on 1100  degrees of freedom
## Residual deviance: 1179.6  on 1099  degrees of freedom
## AIC: 1183.6
##
## Number of Fisher Scoring iterations: 4
```

Residual deviance for the model with predictors is samller than deviance for the null model. Likelihood ratio test p-value less than 0.001 tells us that our model as a whole fits significantly better than an empty model.

```
## [1] 145.7893
```

```
## [1] "p-value:1.44348747189234e-33"
```



Code output for probability of OTIF 3 days after minus probability of 4 days after:

```
## [1] 0.02141281
```

Answer: 2%

Q2.2: How many days can the supplier afford to wait after receiving the PO to start production if they hope to be OTIF?

Assumption: 1) Based on description file, we can assume that the job is always completed in full. 2) Observations are independent. 3) The effects of factors other than time from receiving the PO to start production are not considered.

```
##
## Call:
## glm(formula = OTIF.factor ~ wait.toproduct, family = binomial(logit),
##     data = cleanCopackager)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8310  -0.8543   0.5343   0.7834   1.9336
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.72475    0.32433   14.57   <2e-16 ***
## wait.toproduct -0.69390    0.05499  -12.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1325.4  on 1100  degrees of freedom
## Residual deviance: 1113.7  on 1099  degrees of freedom
## AIC: 1117.7
##
## Number of Fisher Scoring iterations: 4
```

Residual deviance for the model with predictors is samller than deviance for the null model. Likelihood ratio test p-value less than 0.001 tells us that our model as a whole fits significantly better than an empty model.

```
## [1] 211.745
```

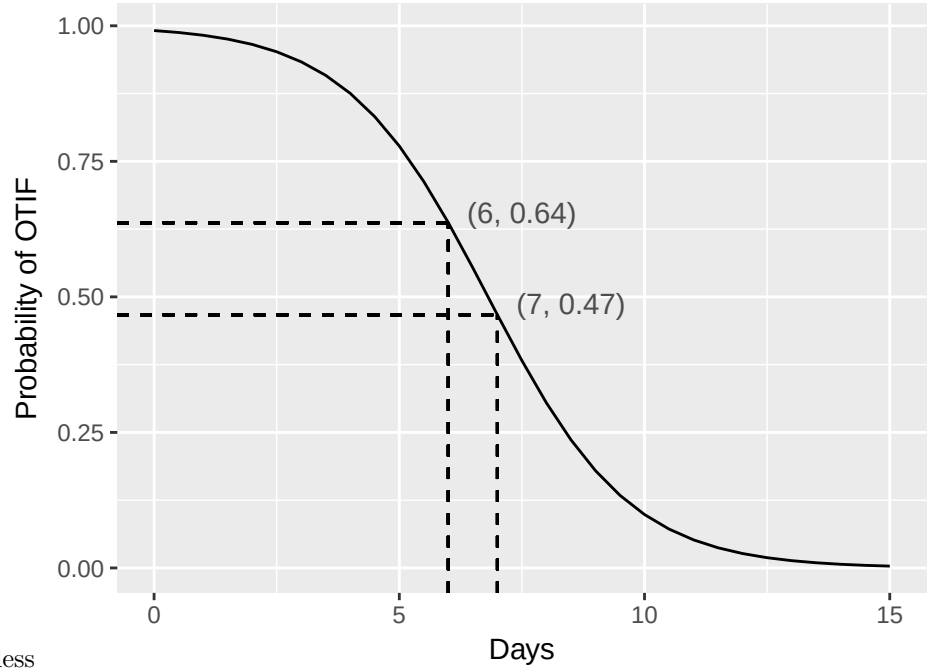
```
## [1] "p-value:5.71689636526554e-48"
```

I built the above model to predict OTIF based on the time packagers wait after receiving the PO to start production. When the input is 6 days, the model predicts true, which means jobs are more likely to be OTIF.

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

When the input is 7 days, the model predicts false, which means jobs are more likely to be not on time.

```
## [1] FALSE
```



Answer: 6 days or less

Additional evaluation for the two models in Q2:

When predictor is the time from receiving the PO to start production, here are some statistics for the model.

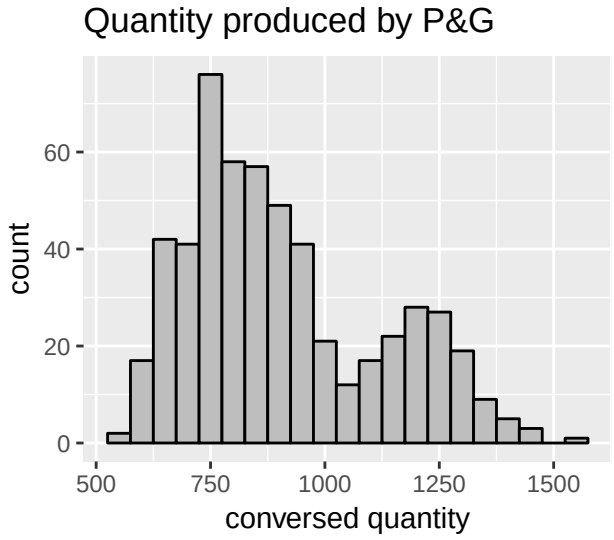
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  31   6
##          1  28 156
##
##                Accuracy : 0.8462
##                  95% CI : (0.7917, 0.891)
##     No Information Rate : 0.733
##     P-Value [Acc > NIR] : 4.358e-05
##
##                   Kappa : 0.5541
##  Mcnemar's Test P-Value : 0.0003164
##
##             Sensitivity : 0.5254
##             Specificity : 0.9630
##          Pos Pred Value : 0.8378
##          Neg Pred Value : 0.8478
##              Prevalence : 0.2670
##          Detection Rate : 0.1403
##    Detection Prevalence : 0.1674
##       Balanced Accuracy : 0.7442
##
##        'Positive' Class : 0
##
```

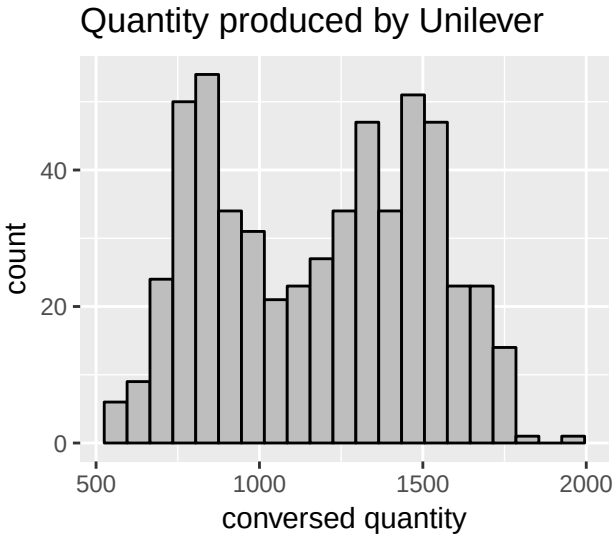
When predictor is the time from receiving the PO to shipment shipped, here are some statistics for the model.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  23  12
##          1  36 150
##
##                Accuracy : 0.7828
##                  95% CI : (0.7226, 0.8353)
##     No Information Rate : 0.733
##     P-Value [Acc > NIR] : 0.0529113
##
##                   Kappa : 0.3627
##  Mcnemar's Test P-Value : 0.0009009
##
##             Sensitivity : 0.3898
##             Specificity : 0.9259
##          Pos Pred Value : 0.6571
##          Neg Pred Value : 0.8065
##              Prevalence : 0.2670
##          Detection Rate : 0.1041
##    Detection Prevalence : 0.1584
##       Balanced Accuracy : 0.6579
##
##        'Positive' Class : 0
##
```

Q3: Is the difference in quantity produced between P&G and Unilever statistically significant?

I have converted the quantity to number of eaches based on customer_unit_of_measure_conversions.csv. From histogram and Shapiro-Wilk test, we can learn that the distribution of data is not normal.



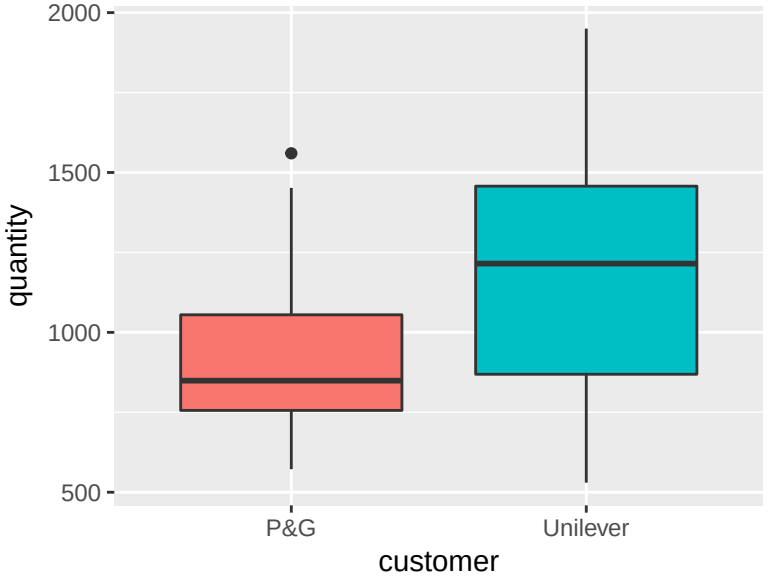



```
##
##  Shapiro-Wilk normality test
##
## data:  q.pg
## W = 0.93621, p-value = 1.533e-14
##
##  Shapiro-Wilk normality test
##
## data:  q.unilever
## W = 0.95333, p-value = 3.141e-12
```

Wilcoxon rank-sum test (also called Mann–Whitney U test or Mann-Whitney-Wilcoxon Test) is conducted and it shows that the difference is statistically significant since p-value is very small. Wilcoxon rank-sum test results:

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  converse.quantity by customer
## W = 76707, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Boxplot for quantities produced by P&G and Unilever:



Answer: Yes

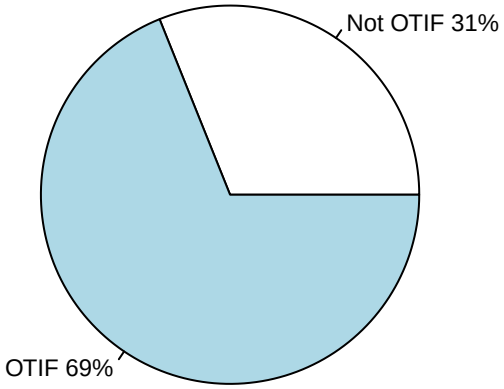
Q4: Assuming everything else is constant, what is the probability of hitting OTIF if the customer was P&G?

Assumption: The effects of factors, like time from receiving the PO to start production, are not considered.

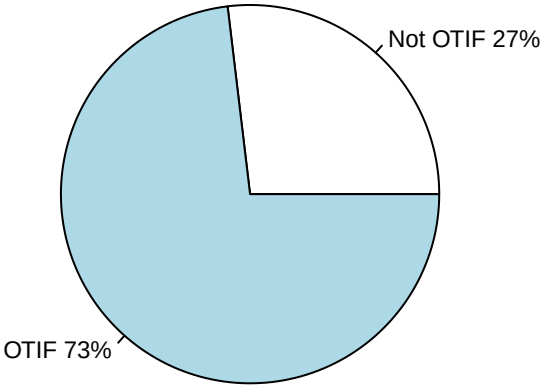
Code output for probability of hitting OTIF if the customer was P&G:

```
## [1] "68.9%"
```

Pie chart for Procter & Gamble:



Pie chart for Unilever:



Answer: 69%