Data Analyst Take-Home Assignment

Yijing Cai

Libraries or packages used: ggplot2 for plot

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.

Dimentions 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 which have dates and times are converted to date/time objects for calculation purpose.
3) Erroneous values were corrected by removing all the observations with purchase order received date later than materials availablity date, or materials availablity 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.

Dimentions 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?

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

Answer: 9.866409 $\approx$ 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 (the packaged quantity is equal to that outlined in the PO). 2) assuming "3 days after" means more than 48 hrs but no more than 72 hrs after, "4 days after" means more than 72 hrs but no more than 96 hrs after.

```
## [1] "15.2%"
```

Answer: 15%

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: The effects of other factors like the time from starting production to shipment shipped, or how long is the time from receiving PO to due date, 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
```

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

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

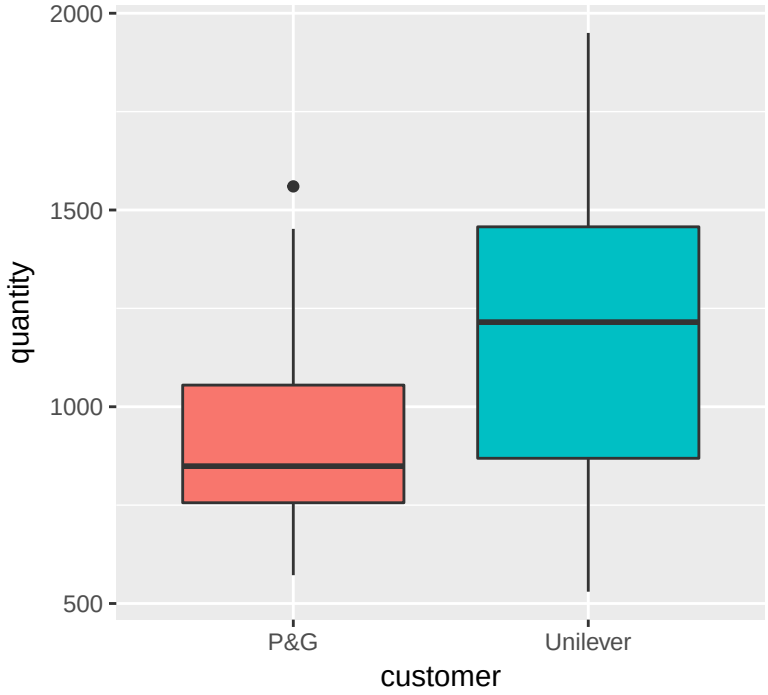
Assumption: Samples come from a distribution that's close to normal.

I have converted the quantity based on customer_unit_of_measure_conversions.csv. Two sample t test is conducted and it shows that the difference is statistically significant since p-value is very small.

T-Test results:

```
##
##  Welch Two Sample t-test
##
## data:  q.pg and q.unilever
## t = -16.701, df = 950.56, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -305.2772 -241.0791
## sample estimates:
## mean of x mean of y
##  909.1865 1182.3646
```

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?

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

Answer: 69%