Homework 12

Code ▾

6210422036: Tanat Iempreedee

Hide

```
library(dplyr)
library(ggplot2)
library(caret)
library(e1071)
library(reshape2)

d <- read.table('shipment.csv', sep=',', header=TRUE,
                colClasses=c('factor','numeric','numeric','factor'))
str(d)
```

```
'data.frame':   400 obs. of  4 variables:
 $ Late    : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 1 2 1 ...
 $ Distance: num  380 660 800 640 520 760 560 400 540 700 ...
 $ Weight  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
 $ LSPRank : Factor w/ 4 levels "1","2","3","4": 3 3 1 4 4 2 1 2 3 2 ...
```

1) EDA

Among all late shipments, the average distance and weight are higher than on-time shipments.

Hide

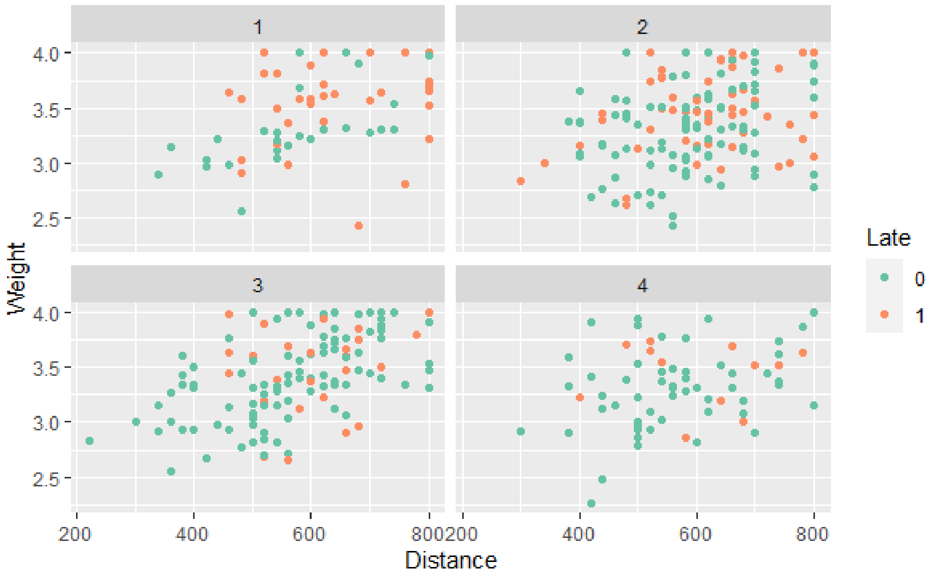
```
d %>% group_by(Late) %>%
  summarise(AvgDistance=mean(Distance), AvgWeight=mean(Weight), n=n())
```

Late	AvgDistance	AvgWeight	n
<fctr>	<dbl>	<dbl>	<int>
0	573.1868	3.343700	273
1	618.8976	3.489213	127

2 rows

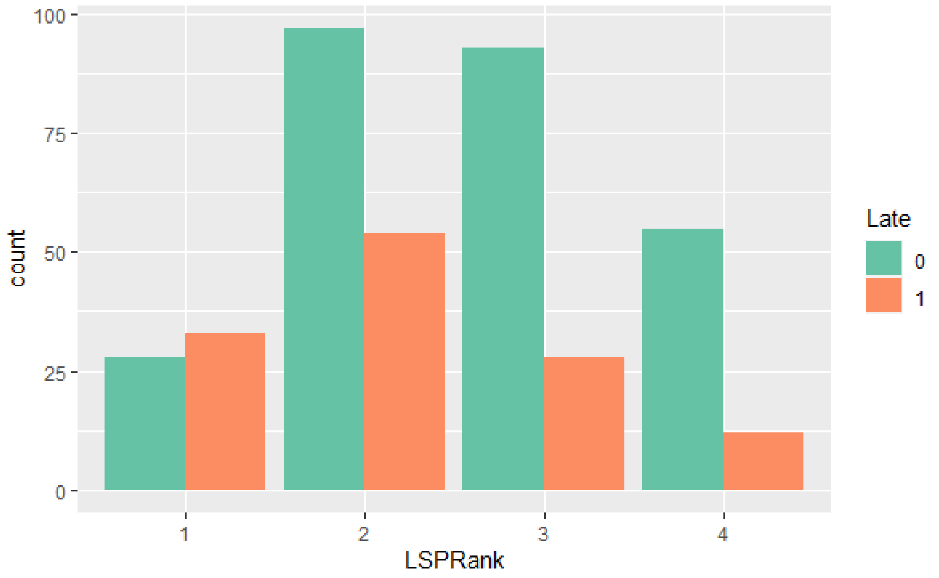
Hide

```
# EDA
ggplot(d, aes(Distance, Weight)) +
  geom_point(aes(color=Late)) +
  facet_wrap(~LSPRank) +
  scale_color_brewer(palette="Set2")
```



Hide

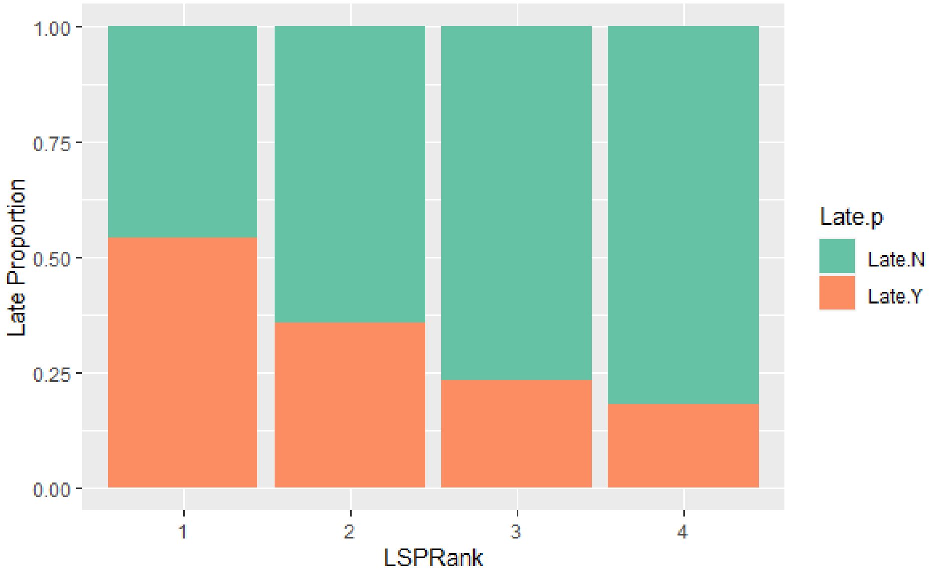
```
# count plot
ggplot(d, aes(fill=Late, x=LSPRank)) +
  geom_bar(position="dodge") +
  scale_fill_brewer(palette="Set2")
```



Hide

```
# normalized count plot
d %>% group_by(LSPRank) %>%
  summarise(Late.N=mean(Late=='0'),Late.Y=mean(Late=='1')) %>%
  melt(variable.name='Late.p') %>%
  ggplot(aes(x=LSPRank, y=value, fill=Late.p)) +
    geom_bar(position='stack', stat='identity') +
    ylab('Late Proportion') +
    scale_fill_brewer(palette="Set2")
```

```
Using LSPRank as id variables
```



2) Fit Logistic Regression

Hide

```
logreg <- glm(Late~Distance + Weight + LSPRank, data=d, family="binomial")
summary(logreg)
```

```

Call:
glm(formula = Late ~ Distance + Weight + LSPRank, family = "binomial",
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
Distance     0.002264   0.001094   2.070 0.038465 *
Weight       0.804038   0.331819   2.423 0.015388 *
LSPRank2    -0.675443   0.316490  -2.134 0.032829 *
LSPRank3    -1.340204   0.345306  -3.881 0.000104 ***
LSPRank4    -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

Coefficient interpretation

- Increases in **Distance** and **Weight** raises the probability of being late. and Weight has a higher effect. -> make sense
- **LSPRank** of 2,3,4 has less probability of being Late compared to LSPRank 1. The higher ranks of LSPRank decreases more probability of being late. This seems to make sense if higher rank means higher quality of a service provider.

3) Predict

Hide

```
newdata <- data.frame(Distance=680,Weight=3.7,LSPRank=factor(2, levels=1:4))
newdata
```

Distance <dbl>	Weight <dbl>	LSPRank <fctr>
680	3.7	2

1 row

Hide

```
y_pred <- predict(logreg, newdata=newdata ,type='response')
y_pred
```

```
        1
0.4624027
```

Hide

```
ifelse(y_pred >= 0.5, 'Late', 'Not Late')
```

```
         1
"Not Late"
```