FA5_EDA

ABLIAN

2025-05-09

```
data <- read.csv("D:/FEU/3RD YR 2ND SEM/EDA/store_sales_data.csv")
library(ggplot2)
library(dplyr)
```

Load and Explore the Data (10 points)

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
head(data)
```

```
##   day_of_week promo holiday store_size sales_count
## 1           6     0       0     medium          18
## 2           3     0       0     medium          13
## 3           4     0       0      large          24
## 4           6     1       0      small          16
## 5           2     0       0     medium          11
## 6           4     0       1     medium          13
```

```
summary(data)
```

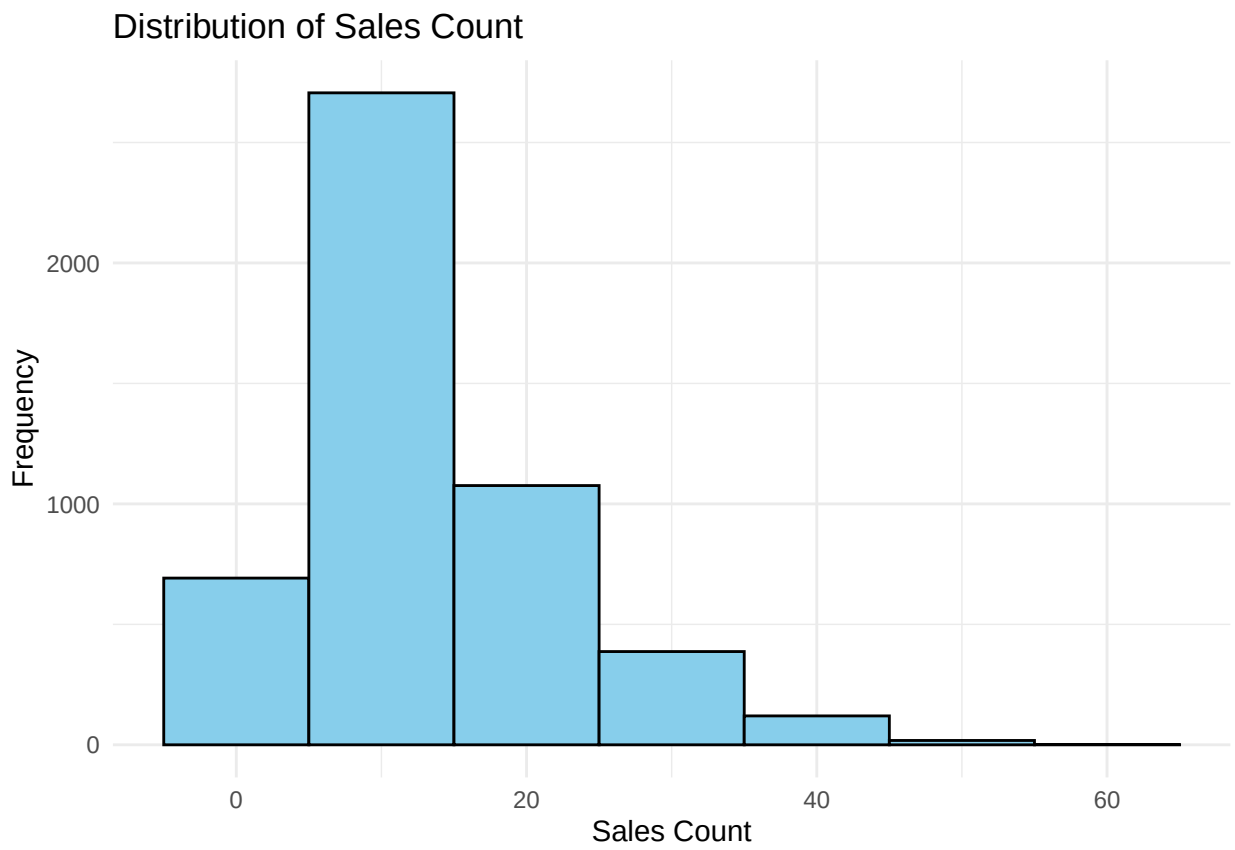
```
##   day_of_week        promo           holiday        store_size
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Length:5000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :3.000   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :2.985   Mean   :0.3012   Mean   :0.0956
##  3rd Qu.:5.000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :6.000   Max.   :1.0000   Max.   :1.0000
##   sales_count
```

```
##  Min.   : 0.00
##  1st Qu.: 7.00
##  Median :12.00
##  Mean   :13.73
##  3rd Qu.:18.00
##  Max.   :61.00
```

```
str(data)
```

```
## 'data.frame':    5000 obs. of  5 variables:
##  $ day_of_week: int  6 3 4 6 2 4 4 6 1 2 ...
##  $ promo      : int  0 0 0 1 0 0 0 1 1 1 ...
##  $ holiday    : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ store_size : chr  "medium" "medium" "large" "small" ...
##  $ sales_count: int  18 13 24 16 11 13 12 34 19 8 ...
```

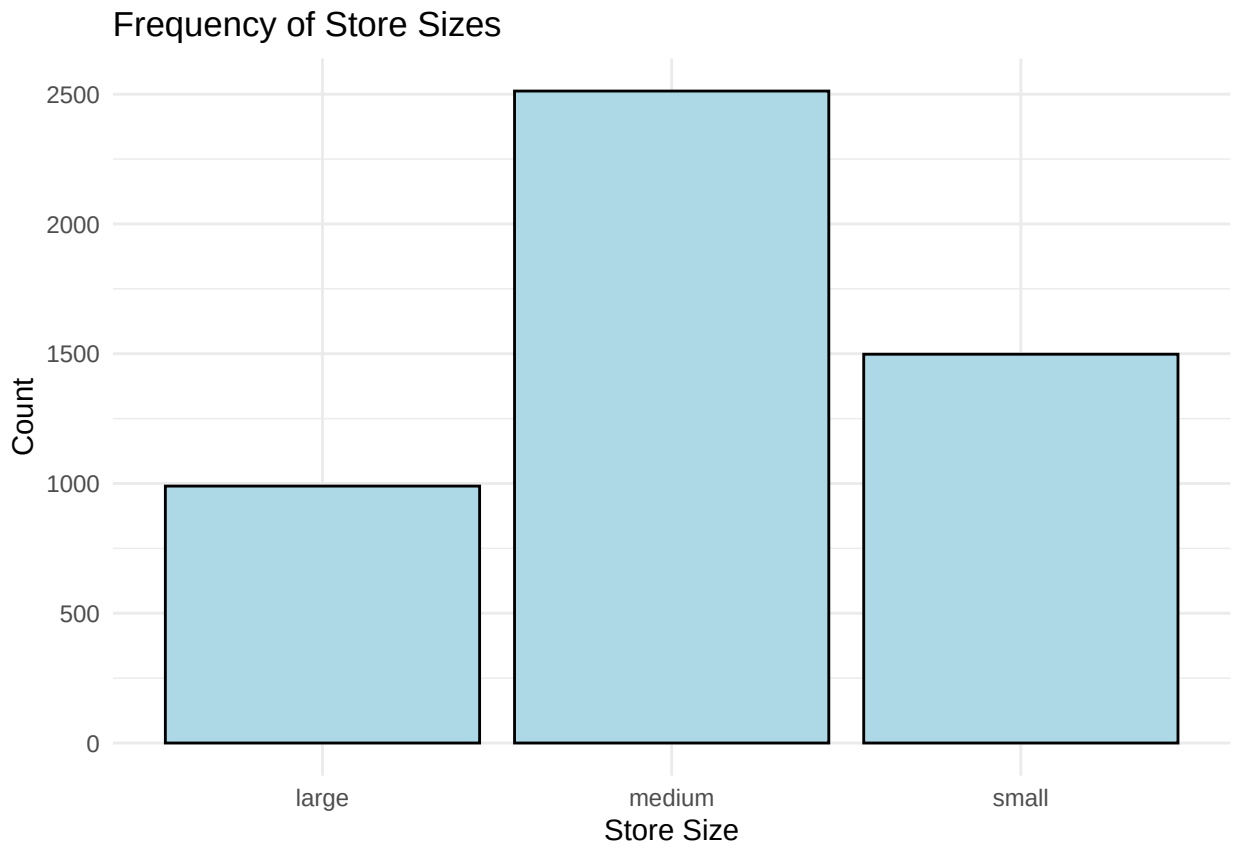
```
ggplot(data, aes(x = sales_count)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Sales Count", x = "Sales Count", y = "Frequency") +
  theme_minimal()
```



Including Plots

You can also embed plots, for example:

```r
ggplot(data, aes(x = store_size)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "Frequency of Store Sizes", x = "Store Size", y = "Count") +
  theme_minimal()
```




```r
data %>%
  count(promo) %>%
  mutate(proportion = n / sum(n))
```

```
##   promo    n proportion
## 1     0 3494     0.6988
## 2     1 1506     0.3012
```

```r
data %>%
  count(holiday) %>%
  mutate(proportion = n / sum(n))
```

```
##   holiday    n proportion
## 1       0 4522     0.9044
## 2       1  478     0.0956
```

the promo column had 69.88% set to have no promos while 30.12% have promosfor holidays, there are 90.44% without holidays while only 9.56% having holidays.

```r
model <- glm(sales_count ~ day_of_week + promo + holiday + store_size,
             family = poisson(link = "log"),
             data = data)

summary(model)
```

Fit a Poisson Regression Model (30 points)

```
##
## Call:
## glm(formula = sales_count ~ day_of_week + promo + holiday + store_size,
##     family = poisson(link = "log"), data = data)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.994849   0.009422  317.86   <2e-16 ***
## day_of_week       0.051115   0.001918   26.65   <2e-16 ***
## promo             0.410843   0.007817   52.55   <2e-16 ***
## holiday          -0.330938   0.014935  -22.16   <2e-16 ***
## store_sizemedium -0.697088   0.008296  -84.03   <2e-16 ***
## store_sizesmall  -1.395564   0.011868 -117.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 25307.2  on 4999  degrees of freedom
## Residual deviance:  5142.7  on 4994  degrees of freedom
## AIC: 26507
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(coef(model))
```

```
##      (Intercept)      day_of_week            promo          holiday
##       19.9823449        1.0524444        1.5080893        0.7182500
## store_sizemedium  store_sizesmall
##        0.4980335        0.2476932
```

What happens to expected sales when there's a promotion? Promo and holiday are binary with 1 being yes and 0 being no, while store size and day of week are categoral with store size having small medium and large and store days being 0-6, we can see that on days of promotion, there is a 50.8% (1.508) sales increase compared to days without it.

How does store size affect expected sales? Store size matters, as smaller stores (24%) have a significantly lower expected sales compared to the medium (49%).

```r
dispersion <- deviance(model) / df.residual(model)
dispersion
```

Assess Model Fit (20 points)

```
## [1] 1.029785
```

```r
model_quasi <- glm(sales_count ~ day_of_week + promo + holiday + store_size,
                   family = quasipoisson(link = "log"),
                   data = data)
summary(model_quasi)
```

```
##
## Call:
## glm(formula = sales_count ~ day_of_week + promo + holiday + store_size,
##     family = quasipoisson(link = "log"), data = data)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.994849   0.009475  316.07   <2e-16 ***
## day_of_week       0.051115   0.001929   26.50   <2e-16 ***
## promo             0.410843   0.007862   52.26   <2e-16 ***
## holiday          -0.330938   0.015020  -22.03   <2e-16 ***
## store_sizemedium -0.697088   0.008343  -83.56   <2e-16 ***
## store_sizesmall  -1.395564   0.011936 -116.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.011374)
##
##     Null deviance: 25307.2  on 4999  degrees of freedom
## Residual deviance:  5142.7  on 4994  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
model_nb <- glm.nb(sales_count ~ day_of_week + promo + holiday + store_size, data = data)
```

```
## Warning in glm.nb(sales_count ~ day_of_week + promo + holiday + store_size, :
## alternation limit reached
```

```r
summary(model_nb)
```

```
##
## Call:
## glm.nb(formula = sales_count ~ day_of_week + promo + holiday +
##     store_size, data = data, init.theta = 891.1483214, link = log)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.994826   0.009525  314.40   <2e-16 ***
## day_of_week       0.051129   0.001937   26.39   <2e-16 ***
## promo             0.410916   0.007903   51.99   <2e-16 ***
## holiday          -0.331041   0.015050  -22.00   <2e-16 ***
## store_sizemedium -0.697132   0.008396  -83.03   <2e-16 ***
## store_sizesmall  -1.395621   0.011952 -116.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(891.1483) family taken to be 1)
##
##     Null deviance: 24892.8  on 4999  degrees of freedom
## Residual deviance:  5065.7  on 4994  degrees of freedom
## AIC: 26508
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  891
##           Std. Err.:  1014
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -26494.11
```

```r
AIC(model)
```

```
## [1] 26506.91
```

```r
AIC(model_nb)
```

```
## [1] 26508.11
```

Since Poisson (first) has a lower AIC, we can use it instead of Negative Binomial(second).

Make Predictions (20 points) Medium store on a Monday with a promotion and no holiday

```r
limit_sale <- tibble(
  day_of_week = 1,
  promo = 1,
  holiday = 0,
  store_size ="medium"
```

```
)

predicted_limit_sale<-predict(model, newdata = limit_sale, type = "response")

predicted_limit_sale
```

```
##        1
## 15.79542
```

Large store on a Sunday with no promotion and a holiday

```
limit_sale_large<-tibble(
  day_of_week = 7,
  promo=0,
  holiday=1,
  store_size="large"
)

predicted_limit_sale_large<-predict(model, newdata = limit_sale_large, type = "response")
predicted_limit_sale_large
```

```
##        1
## 20.52657
```

We can say that a large store size would yield better sales as it had the higher prediction compared to a medium store size. Even with decrease of sales due to holiday on a Sunday, the sales of a large store on a Monday is still lower since since there are other factors of size and day of week. We can say that the size of the store is the biggest factor, as a Sunday without promotion still beat a Monday medium-sized store in sales.

The poisson regression model made a good fit since the overdispersion is only 1.029785 and since it is not over 1.5, there was no need to use another model to compare it to. Again, the biggest factor that contribute to sales is the size of the store, the promo, and then the days as seen in the coefficients of the Poisson model, with medium store size having an increase as it goes from small, medium, to large. The next would be the holidays, and would reflect in real life as seen with holidays but it does not fully reflect the real life settings as there are instances where promos and holidays are on the same day of sales like Christmas where people celebrate the holiday and buy gifts because there are promos.