FA5 DSC1105

Lindsay Faith Bazar

May 01, 2025

Load and Explore the Data

```
data <- read.csv("store_sales_data.csv")
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
```

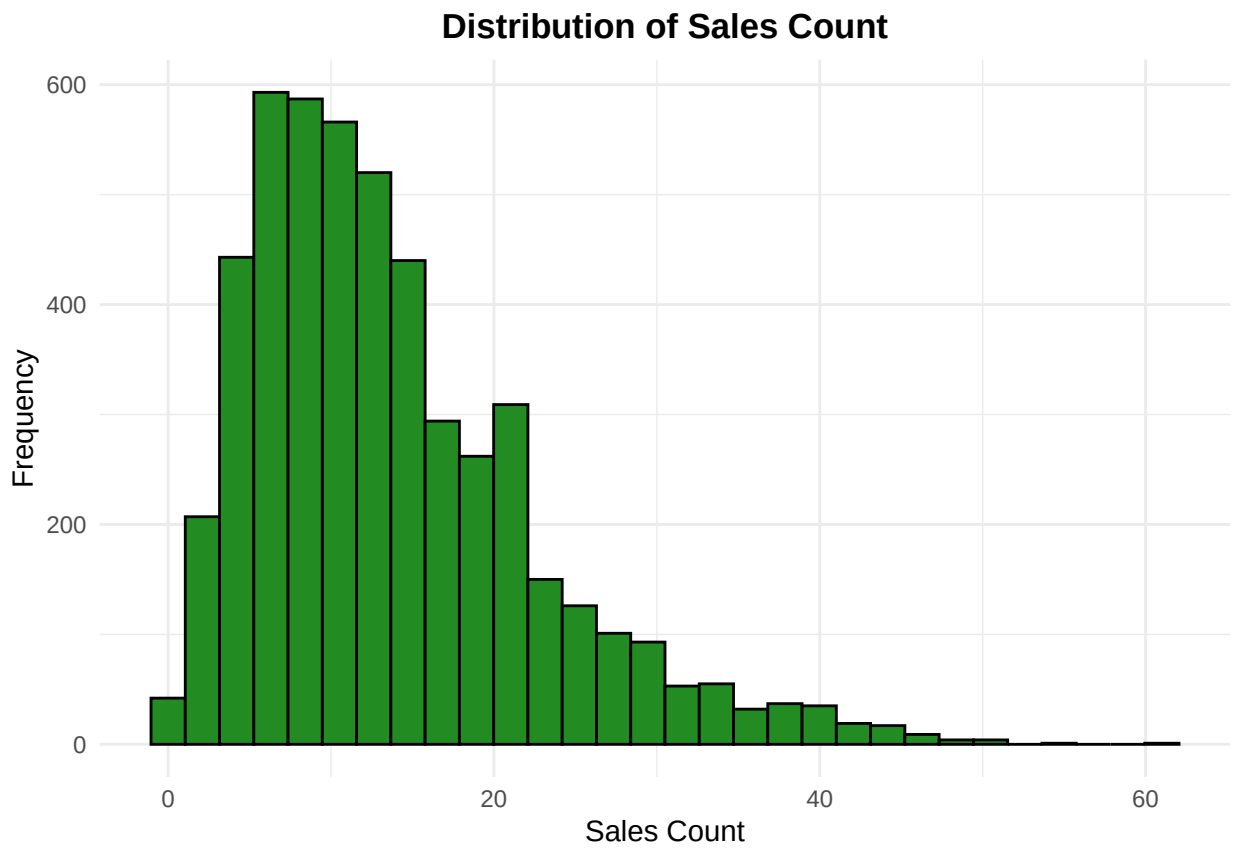
```
str(data)
```

```
## '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 ...
```

Distribution of Sales Count

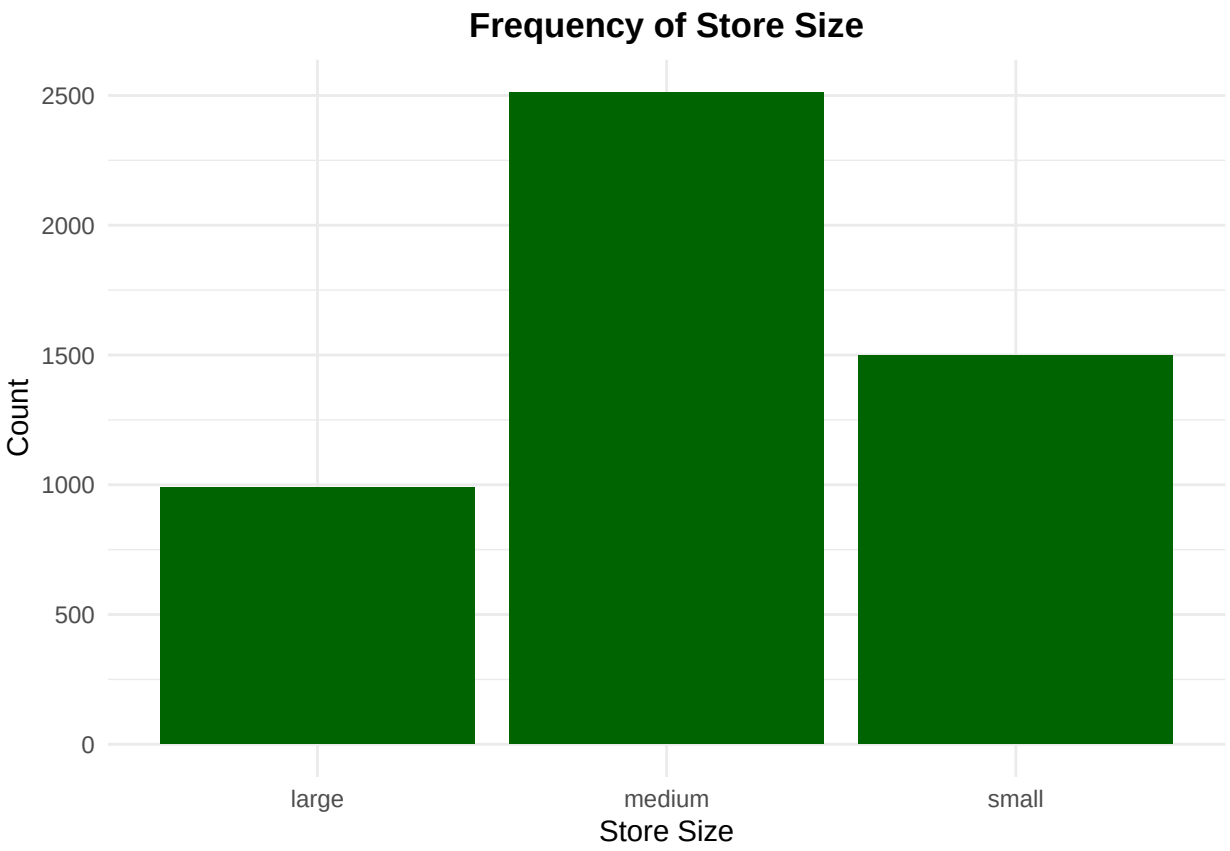
```
ggplot(data, aes(x = sales_count)) +
  geom_histogram(fill = "forestgreen", color = "black") +
  labs(title = "Distribution of Sales Count", x="Sales Count", y="Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



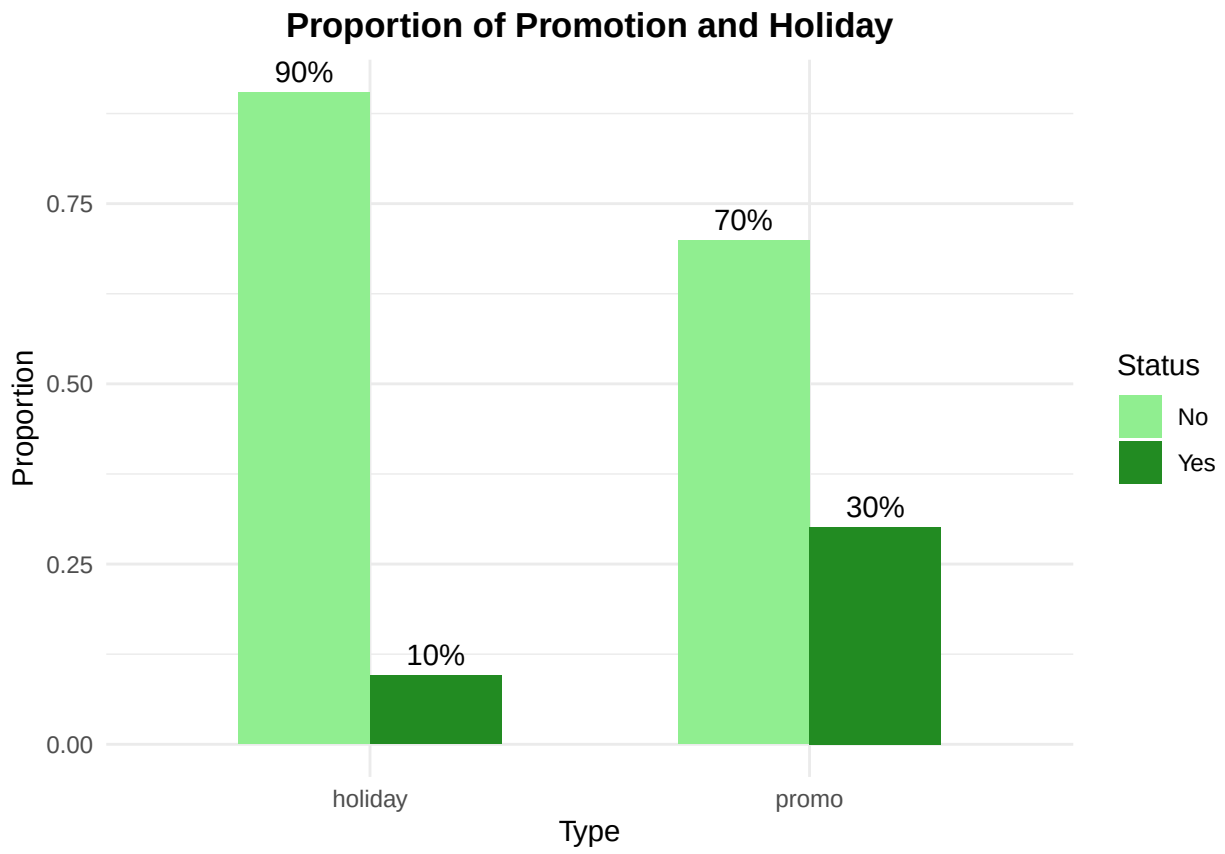
Frequency of each Store Size

```
ggplot(data, aes(x = store_size)) +
  geom_bar(fill = "darkgreen") +
  labs(title = "Frequency of Store Size", x = "Store Size", y= "Count") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Proportion of days with Promotion and Holiday

```
data %>%
  pivot_longer(c(promo, holiday), names_to = "type", values_to = "flag") %>%
  count(type, flag) %>%
  group_by(type) %>%
  mutate(proprtn = n / sum(n)) %>%
  ggplot(aes(x = type, y = proprtn, fill = as.factor(flag))) +
  geom_col(position = "dodge", width = 0.6) +
  geom_text(aes(label = scales::percent(proprtn, accuracy = 1)),
            position = position_dodge(width = 0.6), vjust = -0.5, size = 4) +
  labs(title = "Proportion of Promotion and Holiday",
       y ="Proportion", x ="Type", fill = "Status") +
  scale_fill_manual(values = c("0" = "lightgreen", "1" = "forestgreen"),
                    labels = c("0" = "No", "1" = "Yes")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Fit a Poisson Regression Mode

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

summary(model)
```

```
##
## 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
```

What happens to expected sales when there's a promotion?

```
exp(coef(model)["promo"])*100
```

```
##    promo
## 150.8089
```

When there is a promotion, the expected sales count increases by approximately 50.8% compared to when there is no promotion. This effect is statistically significant ($p < 0.001$), which means that promotions have a strong positive effect on sales counts.

How does store size affect expected sales?

```
exp(coef(model)["store_sizemedium"])*100
```

```
## store_sizemedium
##         49.80335
```

```
exp(coef(model)["store_sizesmall"])*100
```

```
## store_sizesmall
##        24.76932
```

The model compares the expected sales of medium and small stores relative to large stores. Medium stores have 49.8% of the sales of large stores (a 50.2% decrease), which means that if a large store expects to sell 100 items on a typical day, a medium store would only sell about 50. While small stores have 24.77% of the sales of large stores (a 75.23% decrease), which means that if a large store sells 100 items, a small store would only sell around 25. The expected sales drop significantly as store size decreases.

Assess Model Fit

```
deviance(model) / df.residual(model)
```

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

Since 1.0298 is very close to 1 and much less than 1.5, there's no evidence of overdispersion in the model. The poisson regression model seems appropriate for the data, which means that there's no remedy or model comparison needed.

Make Predictions

```
predict_sales <- function(day_of_week, promo, holiday, store_size) {
  new_data <- data.frame(
    day_of_week = day_of_week,
    promo = promo,
    holiday = holiday,
    store_size = store_size
  )

  predicted_sales <- predict(model, newdata = new_data, type = "response")

  return(predicted_sales)
}
```

Predict sales for a medium store on a Monday with a promotion and no holiday

```
predict_sales(1, 1, 0, "medium")
```

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

On a normal Monday with a promotion, a medium store is expected to make about 16 sales.

Predict sales for a large store on a Sunday with no promotion and a holiday

```
predict_sales(7, 0, 1, "large")
```

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

On a holiday Sunday without a promotion, a large store is expected to make about 21 sales.

Overall, sale are affected by store size, promotions, holidays, and the day of the week, with larger stores, promotions, and holidays leading to higher sales.

Reflection

This project analyzed store sales data using Poisson regression to understand how factors like store size, day of the week, promotions, and holidays influence daily sales counts. A Poisson regression model was fit using sales_count as the outcome variable, including all specified predictors. The model summary showed that all predictors were statistically significant and contributed meaningfully to explaining sales variation. Store size had the strongest effect, with larger stores significantly increasing expected sales. Promotions also raised sales, while day of the week and holiday had moderate but meaningful impacts. Model diagnostics showed no overdispersion, validating the use of the Poisson model without the need for alternatives like the negative binomial. However, one limitation of using this model in a real-world setting is that it relies only on the variables in the dataset. Important outside influences—like weather, local events, or changes in customer habits—aren't included, so the model might miss key factors that affect sales in practice.