Using Data to Improve a Marketing Promotion

Matt Jackson

December 10, 2019

Load Required Libraries

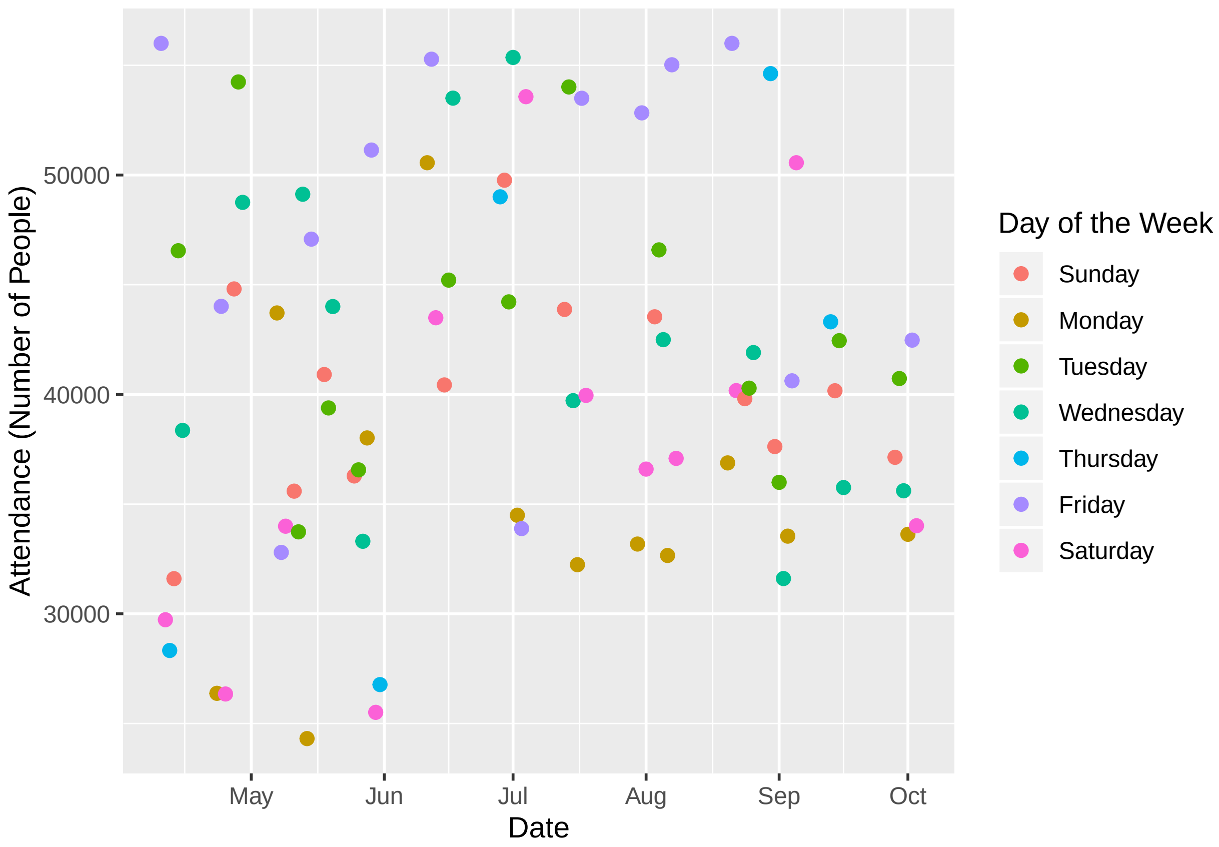
```r
library("ggplot2")
library("Hmisc")
library("dplyr")
library("lubridate")
```

Load Transformed Dodgers Dataset

```r
# Load Dataset
dodgers <- read.csv("../dodgers_transformed.csv")

# Convert Variables to the Right Types
dodgers$date <- as.Date(dodgers$date)
dodgers$month <- month(dodgers$date)
dodgers$cap <- as.logical(dodgers$cap)
dodgers$shirt <- as.logical(dodgers$shirt)
dodgers$fireworks <- as.logical(dodgers$fireworks)
dodgers$bobblehead <- as.logical(dodgers$bobblehead)

# Reorder Days of the Week
dodgers$day_of_week <- factor(dodgers$day_of_week, levels = c("Sunday", "Monday", "Tuesday", "Wednesday

# Create Promo Column
dodgers$promo <- FALSE
dodgers$promo[dodgers$cap==TRUE] <- TRUE
dodgers$promo[dodgers$shirt==TRUE] <- TRUE
dodgers$promo[dodgers$fireworks==TRUE] <- TRUE
dodgers$promo[dodgers$bobblehead==TRUE] <- TRUE
```

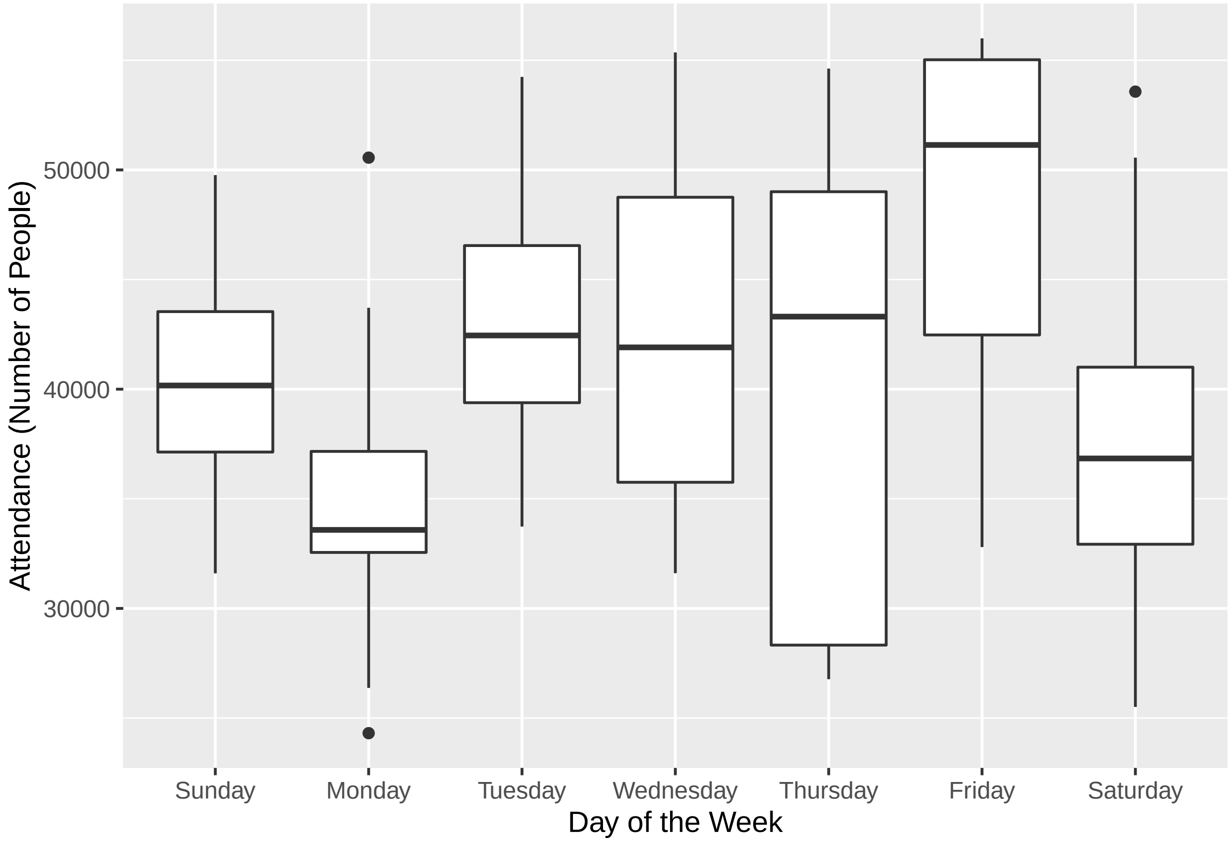
Plot Scatterplot for Attendance over Time with Days of the Week Colored

```r
ggplot(dodgers, aes(x=date, y=attend, color=day_of_week)) +
  geom_point(size=2) +
  xlab("Date") +
  ylab("Attendance (Number of People)") +
  labs(color="Day of the Week")
```



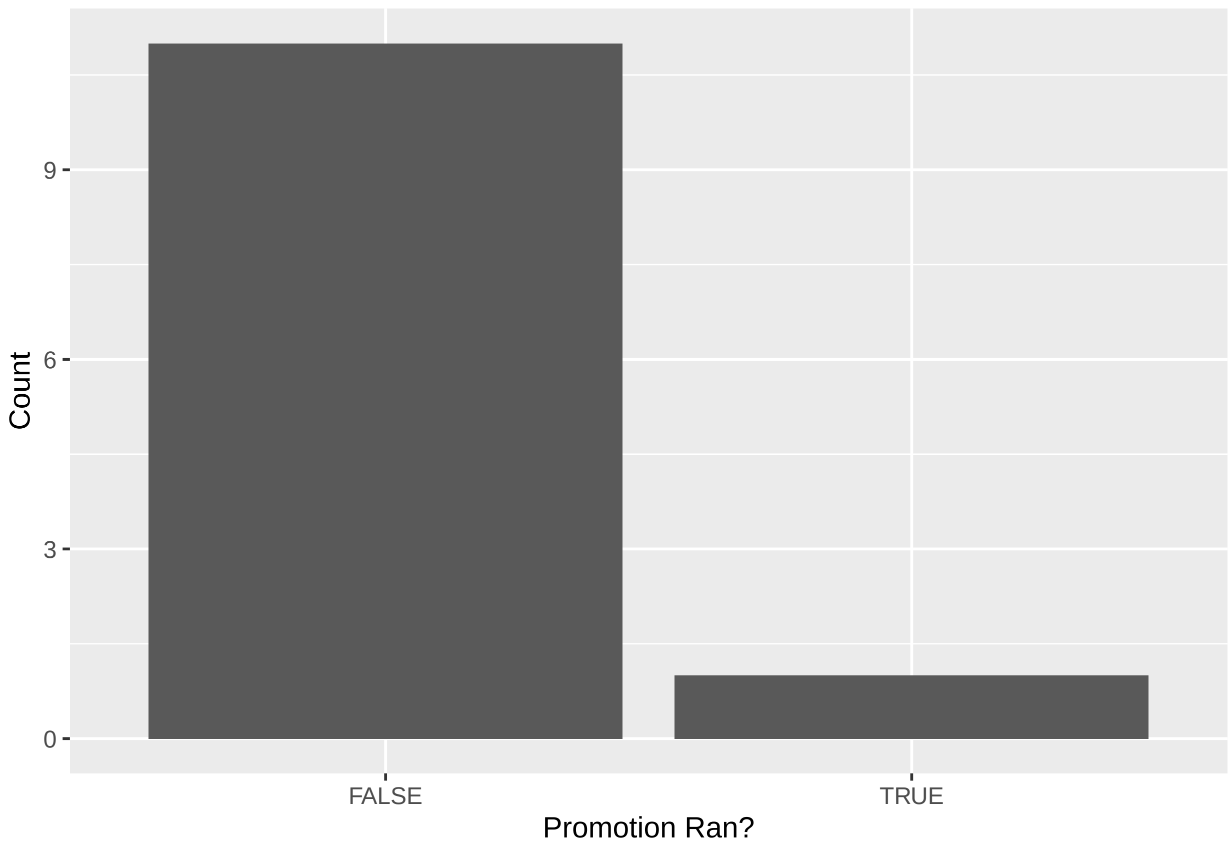
Plot Boxplots for Each Day of the Week

```
ggplot(dodgers, aes(x=day_of_week, y=attend)) +
  geom_boxplot() +
  xlab("Day of the Week") +
  ylab("Attendance (Number of People)")
```



Plot of Promos Being Ran on Mondays

```
dogers_monday <- subset(dodgers, day_of_week=="Monday")

ggplot(dogers_monday, aes(x=promo)) +
  geom_bar() +
  xlab("Promotion Ran?") +
  ylab("Count")
```



Simple Regression Model with Just Day of the Week

```
simple <- lm(attend ~ day_of_week, data=dodgers)
summary(simple)
```

```
##
## Call:
## lm(formula = attend ~ day_of_week, data = dodgers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14942.2  -3909.8   -472.7   4690.1  15984.8
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           40116.9     2102.3  19.082   <2e-16 ***
## day_of_weekMonday     -5151.3     3034.4  -1.698   0.0938 .
## day_of_weekTuesday     2956.0     2973.1   0.994   0.3233
## day_of_weekWednesday   2151.9     2973.1   0.724   0.4715
## day_of_weekThursday     290.5     3988.8   0.073   0.9421
## day_of_weekFriday      7624.3     2973.1   2.564   0.0124 *
## day_of_weekSaturday   -2531.8     3034.4  -0.834   0.4068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7580 on 74 degrees of freedom
## Multiple R-squared:  0.2281, Adjusted R-squared:  0.1655
## F-statistic: 3.644 on 6 and 74 DF,  p-value: 0.003185
```

Create DF to Predict New Attendance Values

```
days <- data.frame("day_of_week" = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
days$day_of_week <- factor(days$day_of_week, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thu
```

Predict Values with Simple Regression Model

```
predict(simple, days)
```

```
##        1        2        3        4        5        6        7
## 40116.92 34965.67 43072.92 42268.85 40407.40 47741.23 37585.17
```

Multiple Regression Model with Day of the Week and Promo

```
multiple <- lm(attend ~ day_of_week + promo, data=dodgers)
summary(multiple)
```

```
##
## Call:
## lm(formula = attend ~ day_of_week + promo, data = dodgers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17898.2  -4090.3     50.1   3753.5  14724.3
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             29611       2748  10.774  < 2e-16 ***
## day_of_weekMonday        4480       3233   1.386 0.170115
## day_of_weekTuesday      11846       3106   3.813 0.000284 ***
## day_of_weekWednesday    10234       3020   3.388 0.001137 **
## day_of_weekThursday      6594       3664   1.800 0.076026 .
## day_of_weekFriday       11665       2690   4.336 4.57e-05 ***
## day_of_weekSaturday      7099       3233   2.196 0.031292 *
## promoTRUE               10506       2061   5.097 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6554 on 73 degrees of freedom
## Multiple R-squared:  0.4307, Adjusted R-squared:  0.3761
## F-statistic:  7.89 on 7 and 73 DF,  p-value: 4.254e-07
```

Add Column for Promo (No Promos Being Run)

```
days$promo <- c(FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE)
```

Predict Values with Multiple Regression Model (No Promos)

```
days_with_no_promos <- predict(multiple, days)
```

Add Column for Promo (Promos Being Run)

```
days$promo <- c(TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE)
```

Predict Values with Multiple Regression Model (Promos)

```
days_with_promos <- predict(multiple, days)
```

Calculate Difference

```
days_with_promos - days_with_no_promos
```

```
##        1        2        3        4        5        6        7
## 10506.47 10506.47 10506.47 10506.47 10506.47 10506.47 10506.47
```

Conclusion

When the scatterplot and boxplots were produced, it is clear that Monday has the lowest attendance. After applying a simple and multiple linear regression model, that still reamined true somewhat (in the mutliple model Sunday had slightly lower attendance). Running a multiple linear regression model taking into account how promos affect attendance, on average the attendance lift is about 10,600 people.