MT 5762 Individual Project

Joshua Arrabaca (ID: 220029955)

Executive Summary

St Andrews Ice Cream Shop is doing well, but it can do better. Generally, sales will only dip below 200 sales units 18.4% of the time, and for ice cream sales, 35% of the time. However, for ice cream sales specifically, this number can range from 26.4% to 44.5%, meaning there may be days where ice cream sales will be below 200 units 44.5% of the time.

When looking at specific factors and correlations, ice cream sales increase as hot drink sales decrease, and ice cream sales also increase as temperatures increase (both as expected). These products do sell better on school holidays. In addition, weekends have a noticeable impact on average sales compared to weekdays; the average sales from Monday to Friday is only 324 whereas on weekends it can be as high as 600.

The shop should also stock more products in April (to coincide with Easter holidays) and in July and August (during peak summer months); in April, sales can reach 557 on average, and in July and August, 606 and 536 respectively.

With the current data, some predictions can be made for the shop; for example, ice cream sales can reach up to 580 sales units in 28°C on a school holiday and a weekend in April, while it can go down to as low as 9.6 unit in 12°C on an ordinary weekday in September. However, these predictions should be taken with a grain of salt as the model was found to be deficient in some areas. Still, it can be a useful tool in guiding the shop in the expected number of sales given the factors mentioned above.

Overall, the ice cream shop is performing well in sales, and while there will be days where sales will be lower than 200 units, the shop should take advantage of the school holidays in April and Summer months of July and August to triple those sales up to 600 units.

Introduction

A data set containing 103 rows of ice cream sales and hot drink sales from an ice cream shop was analysed in R. Exploratory data analyses was performed to see any correlations, and it was shown that ice cream sales is conversely related to hot drink sales, and ice cream sales directly relates to temperature. However, sales as a whole are greatly affected by day of the week (weekend vs weekday) and school holidays. In addition, April, July and August are peak months for total sales, and while the ice cream shop needs 200 unit sales to break even, this is achieved 35% of the time for ice cream sales alone, and 18.4% of the time for sales as a whole. Tests for mean difference and power were conducted and showed that weekend sales are significantly different from weekday sales, and that the current sample sizes are reliable.

A linear model was then created to predict ice cream sales given factors such temperature, humidity, windspeed, weekend, bank holidays, school holidays and month, and reinforces the exploratory finding that temperature and day of the week have a heavy influence on sales. However, evaluation of this model showed visibly right skewed residuals, and the AIC suggests removing humidity and month to improve the model's performance.

Despite the poor evaluation results, the model can be used to give a general idea of estimated ice cream sales; for a weekday in May with temperature 18°C, 6% humidity, and 10 km/h windspeed, the predicted ice cream sales is 145.37, whereas, a day on a January weekend that is not a holiday with temperature -2°C, 75% humidity, and 15 km/h windspeed is 104.71 sales units.

Some improvements can be explored to improve the model such as removing the low-scoring variables of month and humidity. In addition, some clustering methods could also be employed to reveal other patterns between hot drinks versus ice cream sales.

Methods

The analysis was done in R, and the following libraries were used: **readr** for loading the dataset into R; **tidyverse** for basic data wrangling; **Hmisc** for computing confidence intervals for binomial probabilities; **epitools** for computing odds ratios, **pwr** for effect size and power computation, and **ggplot2** for visualisations.

Ice cream sales and hot drink sales were also summed up to create an additional column *total_sales* to find the expected proportion of days with fewer than 200 total sales.

For Part 1, pair plots were created using *pairs(sales_data)*, to examine variables of interest. Additional plots were then created using *ggplot()* with *geom_point()* or *geom_boxplot()*.

Proportions and odds ratios were computed using *bincof()* and *oddsratio()* respectively for part 2. For the odds ratios, *method = 'wilson'* is preferred due to the smaller sample sizes according to Wallis (2013); in this dataset, there are 52 sample points for weekday and 51 for weekend.

For Part 3, *t.test()* was used to compute for the statistical difference between weekend and weekday sales. Then, *pwr.t2n.test()* and *pwr.t.test()* were used to find the power of the t-test result using Cohen's d as the effect size, and the sample size for a power size of 90%.

Finally, *lm()* and *predict()* functions were used in Part 4 to create a linear model and predict ice cream sales. To evaluate the model, *plot(model)* was used and a residuals histogram was created to visualize the evaluation result.

Results

Part 1 - Exploratory Data Analysis

An interesting pattern can be seen in Figure 1 with the response variables of interest: ice cream sales and hot drink sales. There appears to be 2 clusters of data points; one cluster with lower sales (for hot drinks up to 250 and ice cream up to 400), and another cluster with higher sales (for hot drink from 130 and ice cream from 300). In addition, the second cluster seems to have a negative correlation, i.e. as ice cream sales decrease, hot drink sales increase.

The box plot in Figure 2 shows the mean total sales on holidays vs non holidays, with holidays having 635 average sales compared to non-holidays with only around 284, and the quantiles of the box plots do not overlap each other.

When viewing the scatterplot of ice cream sales and temperature in Figure 3, the latter's effect is more evident as a positive relationship can be seen (as temperature increases, ice cream sales also increase). In addition, the significant effect of holidays over non-holidays can be seen in the clustering behaviour.

A similar relationship can be seen with total sales and temperature in Figure 4, this time with weekdays vs weekends; there are significantly more sales on weekends than on weekdays, and these weekend sales form their own cluster above weekday sales.

The behaviour of total sales on weekends vs weekdays can be more accurately seen in Figure 5. While the weekday upper quartile and weekend lower quartile overlap, the weekend plot extends more than the weekday. Moreover, the average weekday sale is only around 324 sales units, whereas the average weekends sale is at 600 units.

When looking at total sales per month in Figure 6, April has high average sales of 557, with similar sales in July and August (606 and 536 respectively). These months also have the widest quartile ranges, with August extending to beyond 750. Meanwhile, the months with the lowest average sales is March with only around 257 units.

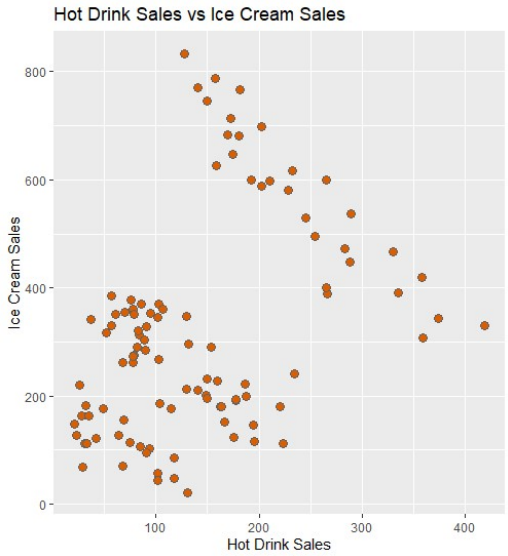

Figure 1 Clustering in Hot Drinks vs Ice Cream Sales

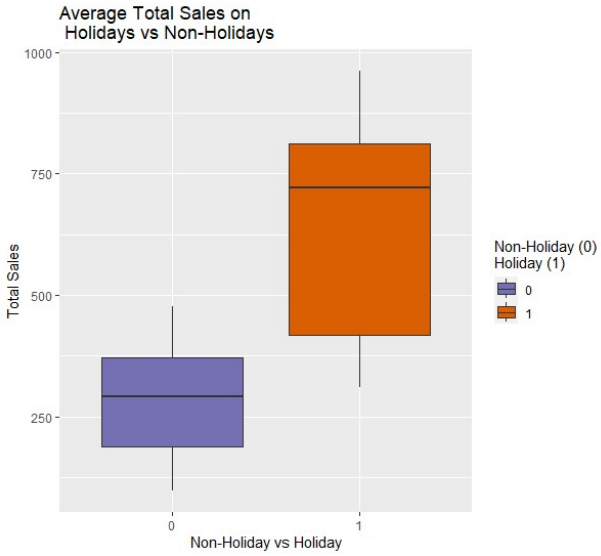

Figure 2 Average Sales on Holidays

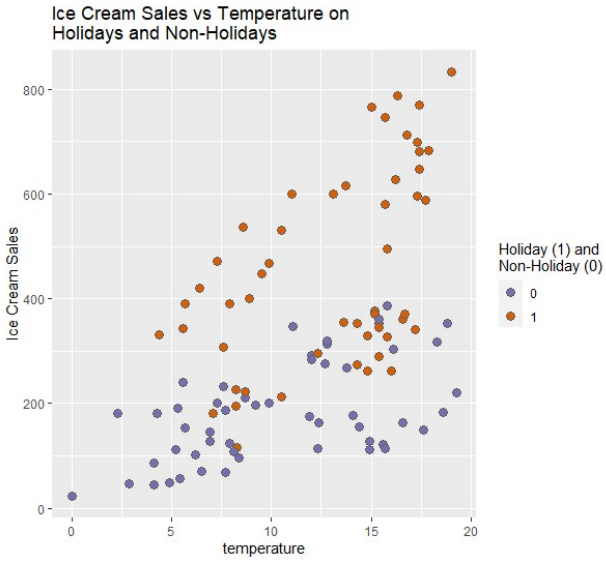

Figure 3 Clustering Effect of Holidays on Sales vs Temperature

Figure 4 Average Sales on Weekends

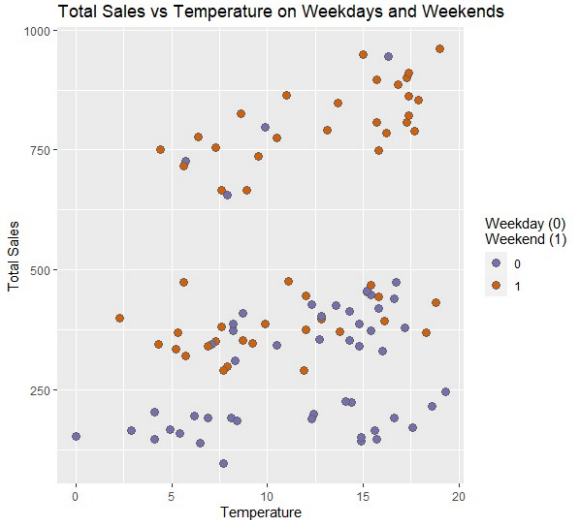

Figure 5 Clustering Effect of Weekends on Sales vs Temperature

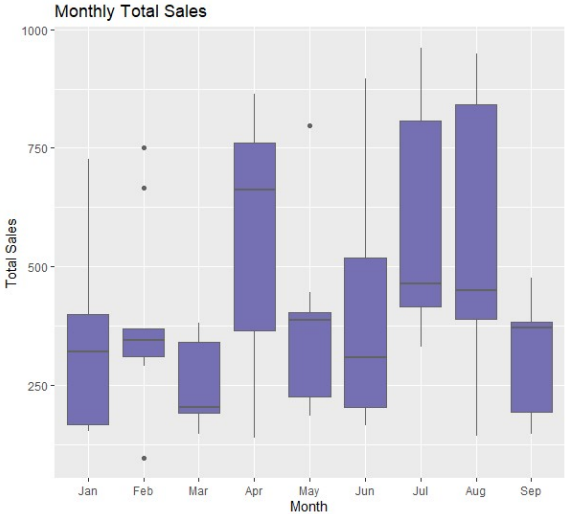

Figure 6 Average Sales by Month

Part 2 – Proportions and Odds Ratios
The expected proportion of days with fewer than 200 **ice cream sales** and a 95% confidence interval is **0.35** with bounds (**0.264**, **0.445**), and the expected proportion of days with fewer than 200 **total sales** and the same CI is **0.184** with bounds (**0.121**, **0.27**).

The odds ratio for a purchase being an ice cream rather than a hot drink in January is **0.21** with bounds (**0.19**, **0.23**) and in August it is **4.74** with upper and lower bounds (**4.33**, **5.20**). Since the 95% confidence intervals of these ratios do not overlap, it can be deduced that there is a significant difference in the odds ratios between January and August.

Part 3 - Power
A t-test was conducted to test the difference in sales on weekdays vs weekends. The resulting t-statistic of **6.81** and p-value of **9.171e-10** show that there is a significant difference between the two periods.

The resulting power of the above is **0.999**. For a power of only 90%, the effect size becomes **0.64**, and for that effect size, the sample size (per group) would be **51.50**.

Part 4 – Linear Model
A linear model was created with ice cream sales as the response variable and the following as the predictors: temperature, humidity, windspeed, weekend, bank holiday, school holiday, and month. The resulting formula becomes:

$$\begin{aligned} Estimate(Ice\ Cream) \\ &= -51.583 + 12.174\ temperature + 0.150\ humidity - 3.151\ windspeed \\ &+ 216.654\ weekend + 213.396\ bank\_holiday\ + 224.738\ school\_holidays\ + month \end{aligned}$$

Where month is treated as a factor, and ranges from -58.915 to 50.412 depending on the month. Assuming the above model is reliable, predictions can be made for ice cream sales with 95% confidence interval bounds. For a weekday in May with temperature 18°C, 6% humidity, and 10 km/h windspeed, the predicted ice cream sales is **145.37**, with lower and upper bounds of (**51.90**, **238.83**).

For a school holiday on a weekend in April with temperature 28°C, 35% humidity, and 5 km/h windspeed, the predicted value is **702.30**, with lower and upper values (**520.02**, **884.57**).

For a week day in September with temperature 12°C, 90% humidity, and 35 km/h windspeed, the prediction is **9.63**, and bounds (**-110.04**, **129.30**).

And finally, a day on a January weekend that is not a holiday with temperature -2°C, 75% humidity, and 15 km/h windspeed: **104.71**, and bounds (**13.04**, **196.37**).

Discussion

Part 1 - Exploratory Data Analysis
A preliminary look at ice cream sales vs ice cream sales shows two interesting observations. First, there seems to be at least two different clusters with one group clustering around the lower total sales, and the other group around higher total sales. Second, the second cluster shows a negative correlation in that higher ice cream sales relates to lower hot drink sales. This second observation is made more plausible in relation to the other plots where an increase in temperature shows an increase in ice cream sales.

In addition to the positive correlation between ice cream sales and temperature, there is also a significant effect from school holidays and weekends. The average total sales more than doubled on school holidays, with weekends also having a similar result compared to weekdays. Certain months are also of interest, with the highest sales in April (around Easter season) as well as in July and August (in the Summer months.)

Part 2 – Proportions and Odds Ratios

Although the probability of total sales only reaching below 200 units is 18.4%, the 95% CI is wide, ranging from 12.1% to 27%. This could mean that total sales could dip below 200 units 1/3 of the time. When looking at ice cream sales, the figures are more concerning; ice cream sales should reach at least 200 units 35% of the time, but the 95% confidence interval also shows a wide range from 26.4% to 45.5%. This means that ice cream sales could dip below 200 units 45.5% of the time, which should be a concern for the shop owner.

The above proportions do not consider the time of the year, however. For example, while ice cream sales might be below 200 units around 35% of the time, it the odds ratio of ice cream sales to hot drinks is considerably much lower in January than in August. Thus, the current month (as well as other factors discussed in the linear model) should be taken into account in estimating when sales will dip below 200.

Part 3 – Power

The results of the t-test with p-value of 9.171e-10 show that weekends sales are significantly different than weekday sales, and the associated power of this (which is the probability of avoiding a Type II error) of 0.99 shows that this result is reliable.

For a lower effect size of 0.64, only 51.49 samples are needed per group, which makes sense since the current sample sizes are 52 and 51 for weekday and weekend respectively, and a lower effect size would require only a smaller sample size (Verma & Verma, 2020).

These results definitively show that ice cream sales are much different depending on the day of the week, and the earlier analyses show weekends typically result in higher sales overall.

Part 4 - Linear Model

The predictions confirm the general positive effect of temperature and holidays; the above predictions that have higher temperatures or have holidays have higher predictions than those without. On assessing the model however, the residuals histogram showed a right skewed distribution, which could mean that the model is underpredicting.

In addition, running the AIC (both directions) showed that the current model's score is 894.47, and could be improved by removing humidity and month_name to arrive at an improved score of 887.71.

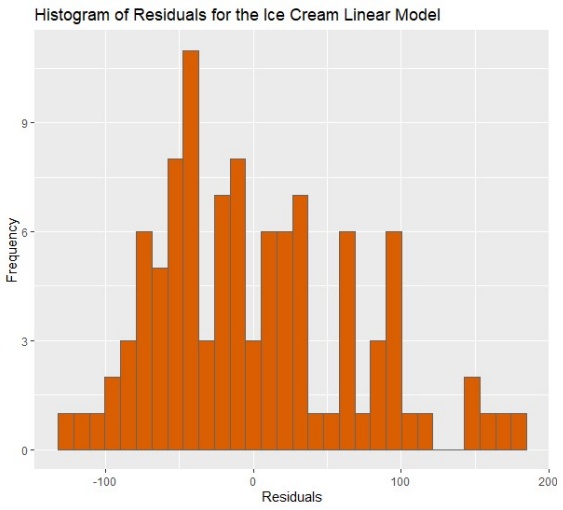

Figure 7 Residuals Histogram

General Recommendations

The below recommendations are from the above analyses and modelling. While they may seem obvious, they still bear repeating for St Andrews Ice Cream shop:

- Stock more ice cream on warmer days, and hot drinks on cooler days;
- Stock both ice cream and hot drinks on school holidays; and
- Stock both ice cream and hot drinks on weekends

In addition, the above plots show some clustering between hot drinks versus ice cream sales, so this could be explored more in the future. Improving the linear model to improve the variance in residuals will also help in providing predictions for the ice cream shop.

References

Heumann, C. (2022). *Introduction to statistics and Data Analysis: With exercises, solutions and applications in R*. Spring.

Verma, J. P., & Verma, P. (2020). Chapter 2. In *Determining sample size and power in research studies: A manual for researchers*. essay, Springer Nature Singapore Pte Ltd.

Wallis, S. (2013). Binomial confidence intervals and contingency tests: Mathematical Fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, *20*(3), 178–208. https://doi.org/10.1080/09296174.2013.799918

Appendix – R Code

```
# MT5762 Project
# Joshua Arrabaca (ID: 220029955)

# Load libraries
library(readr)
library(tidyverse)
library(Hmisc)
library(epitools)
library(pwr)

# Load the dataset
sales_data <- read_csv("sales_data.csv")

# Change month column to a factor
sales_data <- sales_data %>%
  mutate(month_name = factor(month_name,
                             levels = c("Jan", "Feb", "Mar",
                                        "Apr", "May", "Jun",
                                        "Jul", "Aug", "Sep")))

# Add a column for total sales (icecream_sales + hotdrink_sales)
sales_data <- sales_data %>%
  add_column(total_sales = sales_data$hotdrink_sales + sales_data$icecream_sales)


# Part 1 - Exploratory Data Analysis

# Pairs plots to examine possible correlations
pairs(sales_data)

# Scatterplot ice cream sales vs hot drink sales
plt.ice.hot <- ggplot(data=sales_data,
                      aes(x=hotdrink_sales, y=icecream_sales)) +
  geom_point(colour="#636363", fill= "#d95f02", pch=21, size = 3) +
  xlab("Hot Drink Sales") + ylab("Ice Cream Sales") +
  ggtitle("Hot Drink Sales vs Ice Cream Sales")

plt.ice.hot

# Scatterplot ice cream sales vs temperature, with school holidays as the color fill
plt.ice.schholi <- ggplot(data=sales_data,
                          aes(x=temperature,
                              y=icecream_sales,
                              fill = factor(school_holidays))) +
```

```
  geom_point(colour="#636363",pch=21, size=3) +
  scale_fill_manual(values = c("#7570b3", "#d95f02")) +
  xlab("temperature") + ylab("Ice Cream Sales") +
  labs(fill = "Holiday (1) and \nNon-Holiday (0)") +
  ggtitle("Ice Cream Sales vs Temperature on Holidays and Non-Holidays")

plt.ice.schholi

# Find the means of sales on holidays vs non-holidays
holi.means <- sales_data %>% group_by(school_holidays) %>%
  summarise("Holiday Means" = round(mean(total_sales), 2))

holi.means

# Boxplot of total sales on holidays vs non-holidays
plt.total.box.holi <- ggplot (sales_data) +
  geom_boxplot (aes(x = factor(school_holidays),
                    y = total_sales,
                    fill = factor(school_holidays))) +
  scale_fill_manual(values = c("#7570b3", "#d95f02")) +
  xlab("Non-Holiday vs Holiday") + ylab("Total Sales") +
  labs(fill = "Non-Holiday (0)\nHoliday (1)") +
  ggtitle("Average Total Sales on Holidays vs Non-Holidays")

plt.total.box.holi

# Scatterplot of total sales vs temperature, with weekends/weekdays as the color fill
plt.total.wkend <- ggplot(data=sales_data,
                          aes(x=temperature,
                              y=total_sales,
                              fill = factor(weekend))) +
  geom_point(colour="#636363", pch=21, size=3) +
  scale_fill_manual(values = c("#7570b3", "#d95f02")) +
  xlab("Temperature") + ylab("Total Sales") +
  labs(fill = "Weekday (0)\nWeekend (1)") +
  ggtitle("Total Sales vs Temperature on Weekdays and Weekends")

plt.total.wkend

# Boxplot of total sales on weekends vs weekdays
plt.weekend <- ggplot (sales_data) +
  geom_boxplot (aes(x = factor(weekend),
                    y = total_sales,
                    fill = factor(weekend))) +
  scale_fill_manual(values = c("#7570b3", "#d95f02")) +
  xlab("Weekday vs Weekend") + ylab("Total Sales") +
  labs(fill = "Weekday (0)\nWeekend (1)") +
  ggtitle("Average Total Sales on Weekdays vs Weekends")

plt.weekend

# Compute for the mean total sales of weekdays vs weekends
weekend.means <- sales_data %>% group_by(weekend) %>%
  summarise("Mean Count" = round(mean(total_sales), 2))

weekend.means

# Boxplot of total sales per month
month.plot <- ggplot(sales_data) +
  geom_boxplot(aes (x=month_name, y=total_sales),
               colour="#636363",
               fill = "#7570b3" ) +
```

```r
  xlab ("Month") + ylab("Total Sales") +
  ggtitle("Monthly Total Sales")

month.plot

# Compute for the means per month
month.means <- sales_data %>% group_by(month_name) %>%
  summarise("Mean Count" = round(mean(total_sales), 2))

month.means

# Part 2 - Proportions and Odds Rations
# Q: the expected proportion of days with fewer than 200 ice cream sales and a 95%
confidence interval

# Compute for the aggregate ice cream sales with < 200
icecream.sales.less <- sum (sales_data$icecream_sales < 200)
icecream.sales.greater <- sum (sales_data$icecream_sales >= 200)
icecream.sales.total <- length (sales_data$icecream_sales)

# Use binconf to compute for the odds ratio
binconf(x=icecream.sales.less,
        n=(icecream.sales.total),
        alpha=0.05, method="wilson") |>   round(3)


# Q:the expected proportion of days with fewer than 200 total sales (ice cream and hot
drinks) and a 95% confidence interval

# Compute for the aggregate total sales with < 200
total.sales.less <- sum (sales_data$total_sales < 200)
total.sales.greater <- sum (sales_data$total_sales >= 200)
total.sales.total <- length (sales_data$total_sales)

# Use binconf to compute for the odds ratio
binconf(x=total.sales.less, n=(total.sales.total), alpha=0.05, method="wilson") |>
round(3)

# Q: the odds ratio for a purchase being an ice cream rather than a hot drink in
January and in August and a 95% confidence interval for each.

# Find the sales per product type for January and August
jan.icecream <- sum(filter (sales_data, month_name == "Jan")$icecream_sales)
jan.hotdrink <- sum(filter (sales_data, month_name == "Jan")$hotdrink_sales)
aug.icecream <- sum(filter (sales_data, month_name == "Aug")$icecream_sales)
aug.hotdrink <- sum(filter (sales_data, month_name == "Aug")$hotdrink_sales)

# Create object containing the number of the above
counts.jan <- c(jan.icecream, aug.icecream, jan.hotdrink, aug.hotdrink)
counts.aug <- c(aug.icecream, jan.icecream, aug.hotdrink, jan.hotdrink)

# Convert to a matrix
matcounts.jan <- matrix(counts.jan, nrow=2, byrow=TRUE)
matcounts.aug <- matrix(counts.aug, nrow=2, byrow=TRUE)

# Compute for the odds ratio for January
jan.OR <- oddsratio(matcounts.jan, method='wald')
jan.OR

# Compute for the odds ratio for August
aug.OR <- oddsratio(matcounts.aug, method='wald')
aug.OR
```

```r
# Add row and column names when viewing
dimnames(matcounts.jan) <- list("Food"=c("IceCream","HotDrink"),
                                "Month"=c("Jan","Aug"))
dimnames(matcounts.aug) <- list("Food"=c("IceCream","HotDrink"),
                                "Month"=c("Aug","Jan"))

# Q whether there is a significant difference in odds ratios between January and
August.
# Since the 95% CI do not overlap, we can deduce that the odds ratios are statistically
different.


## Part 3 - Power

# Q: Test whether there is a difference between the expected number of sales on week
days (Mon-Fri) and weekends. Interpret and explain your results.

# Find the means, lengths, and SD for weekend vs weekday, and compute for the mean
sales on weekends vs weekdays
sales.weekday <- (filter (sales_data, weekend == 0)$icecream_sales)
sales.weekend <- (filter (sales_data, weekend == 1)$icecream_sales)

len1 = length(sales.weekday)
len2 = length(sales.weekend)

mean1 = mean(sales.weekday)
mean2 = mean(sales.weekend)

sd1 = sd(sales.weekday)
sd2 = sd(sales.weekend)

# Q: Test whether there is a difference between the expected number of sales on week
days (Mon-Fri) and weekends. Interpret and explain your results.
# Do an F-test
t <- t.test(x = sales.weekday, y = sales.weekend)
t

# Compute for the effect size (Cohen's d)
cohens.d = (mean2 - mean1)/sqrt(((len1-1)*(sd1^2) + (len2-1)*(sd2^2))/
                                (len1+len2-2))
cohens.d

# Q: Compute the power of the above test, assuming that the true difference is the one
observed.
pwr.t2n.test (n1 = len1,
              n2 = len2,
              d = cohens.d,
              sig.level = 0.05,
              alternative = "two.sided")

# Q: For the observed sample size, what effect size (i.e., difference between the
expected values) would be required to obtain a power of 90%?
pwr.t2n.test (n1 = len1,
              n2 = len2,
              power =0.90,
              sig.level = 0.05,
              alternative = "two.sided")

# For the given effect size, what sample size would be required to obtain a power of
90%?
pwr.t.test (d = 0.6449855,
```

```
            sig.level = 0.05,
            power= 0.90,
            alternative = "two.sided")


# Part 4 - Linear Model

# Create a linear model with ice cream sales as the response variable
model <- lm(icecream_sales ~
              temperature +
              humidity +
              windspeed +
              weekend +
              bank_holiday +
              school_holidays +
              month_name,
            data = sales_data)

# View the model
summary(model)

# View the coefficients only
model$coefficients |> round(3) |> data.frame()
#specify confidence interval

# Predict: a weekday in May with temperature 18°C, 6% humidity, and 10 km/h windspeed,
6% humidity, and 10 km/h windspeed
df.a <- data.frame(temperature = 18,
                   humidity = 6,
                   windspeed = 10,
                   weekend = 0,
                   bank_holiday = 0,
                   school_holidays = 0,
                   month_name = "May")

pred.a <- predict(model, newdata = df.a, interval = "confidence", level = 0.95)
pred.a |> round(2)

# Predict: For a school holiday on a weekend in April with temperature 28°C, 35%
humidity, and 5 km/h windspeed, the predictions are:
df.b <- data.frame(temperature = 28,
                   humidity = 35,
                   windspeed = 5,
                   weekend = 1,
                   bank_holiday = 0,
                   school_holidays = 1,
                   month_name = "Apr")

pred.b <- predict(model, newdata = df.b, interval = "confidence", level = 0.95)
pred.b |> round(2)

# Predict: For a week day in September with temperature 12°C, 90% humidity, and 35 km/h
windspeed:
df.c <- data.frame(temperature = 12,
                   humidity = 90,
                   windspeed = 35,
                   weekend = 0,
                   bank_holiday = 0,
                   school_holidays = 0,
                   month_name = "Sep")

pred.c <- predict(model, newdata = df.c, interval = "confidence", level = 0.95)
```

```r
pred.c |> round(2)

# Predict:  a day on a January weekend that is not a holiday with temperature -2°C, 75%
humidity, and 15 km/h windspeed:
df.d <- data.frame(temperature = -2,
                   humidity = 75,
                   windspeed = 15,
                   weekend = 1,
                   bank_holiday = 0,
                   school_holidays = 0,
                   month_name = "Jan")

pred.d <- predict(object = model,
                  newdata = df.d,
                  interval = "confidence",
                  level = 0.95)
pred.d |> round(2)



# Discussion
# View the residuals plots
model
plot(model)

# Create a histogram of the residuals
residuals.plot <- ggplot() +
  geom_histogram(aes(x = model$residuals),
                 fill = '#d95f02',
                 color = "#636363") +
  xlab("Residuals") + ylab("Frequency") +
  ggtitle("Histogram of Residuals for the Ice Cream Linear Model")

residuals.plot

# End of R code
```