2023 St. Andrews Ice Cream Shop Analysis Report

Jim YANG
October 2023

MT5762: Introductory Data Analysis
Final Project

University of
St Andrews

The University of St Andrews
St Andrews

1 Introduction

Based on the given data collecting in 2023 from an ice-cream shop in St.Andrews Scotland, this report is aimed to improve the efficiency of stock ordering by making better predictions. The data set includes those different types of variables: the number of ice cream sold and hot drinks sold; the average temperature, humidity and wind-speed; the months in 2023; holidays (includes weekend, bank holidays and school holidays). This report primarily centers its attention on and thoroughly addresses the following subjects: (1) Estimate low sales days and odds ratios for ice cream purchases in January and August; (2) Test sales differences between weekdays and weekends, calculating power and sample size for effect size; (3) Examine how temperature, humidity, and other factors impact ice cream sales, providing sales predictions for specific weather conditions.

2 Method

To analyze the relationship between the number of ice cream or hot drinks with temperature, the graph of 'Temperature distribution' is drawn to help more intuitively see the frequency of temperatures throughout the year in St.Andrews 1.

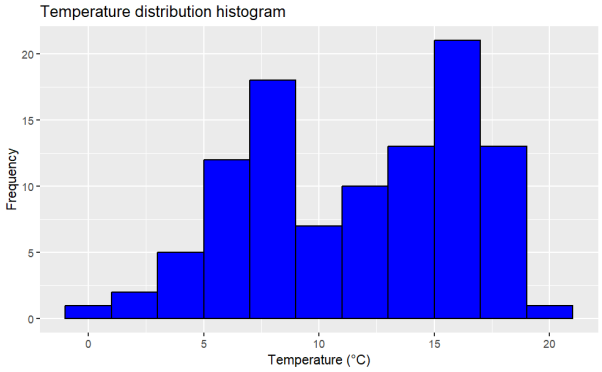


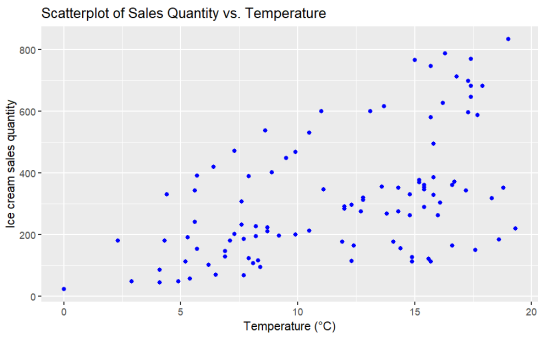
Figure 1: Temperature Distribution

The scatter plots 2 show the relationship between the sales quantities with temperature, which also show the positive relationship between temperature and ice cream sales, and the negative relationship between temperature and hot drinks sales.

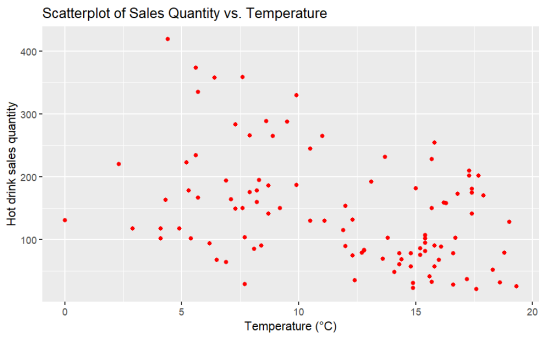
After a general analysis of temperature, it is necessary to analyze the specific subject that mentioned above. An description of exploratory data analysis (EDA) are listed below respectively.

2.1 Part 1: Analyzing Sales Trends in January and August

This section mainly includes 4 parts. Firstly, data filtering and analysis. It is necessary to substitute the 'empty data' in the data set. Also, we need to calculate the proportion



(a) ice cream



(b) hot drink

Figure 2: Scatter plots of sales and temperature

of days with fewer than 200 ice cream sales and total sales. Secondly, using appropriate statistical methods to compute a 95% confidence interval for the aforementioned proportions. Thirdly, it is also important to calculate the odds ratio so that the difference between different months can be shown. Lastly, we need to test for significant difference in odds ratios.

2.2 Part 2: Weekday vs. Weekend Ice Cream Sales Analysis

In this part, we also have 4 steps to finish this process. Firstly, we need to re-group the data into 'weekend' and 'weekdays' lists, which is convenient to check the difference. Secondly, we need to test for differences in sales. Making hypothesis and using suitable statistical test to find out if there's a significant difference in sales between weekdays and weekends. Next, it is necessary to compute the power of the above test (assuming that the true difference is the one observed). Lastly, the power analysis is also considered in this report.

2.3 Part 3: Predicting Sales Based on Weather Conditions

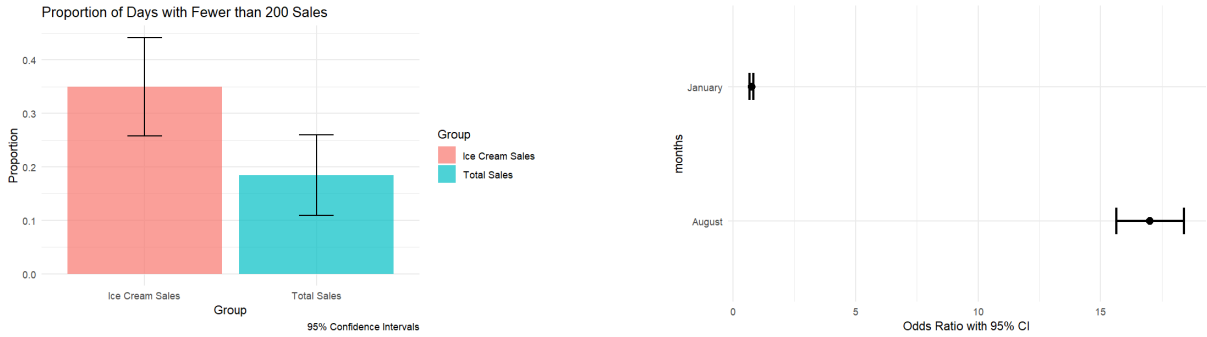
In the last section, there are 3 parts involved to finish the examination. Firstly, we consider to utilize a multiple linear regression model to determine how the variables like temperature, humidity, wind-speed, weekends, bank holidays, and school holidays affect the number of ice cream sales. Secondly, we consider about the interpretation of parameters. More exactly, it is interpreting the estimated parameters from the model to understand how each variable affects the sales. Thirdly, we make a sales prediction to finish the final goal in this report.

3 Results

3.1 Results of Part 1

Based on the calculation, it indicates that 34.95% of the days have ice cream sales less than 200, which means there are over 1/3 of the days that the shop cannot be profitable by selling ice cream. A 95% confident interval of proportion of days with fewer than 200 sales of ice cream and total sales is determined 3a:

The odds ratio analysis for ice cream vs. hot drink purchases shows a significant sales difference in January or August, while the OR is 0.753 in January but 16.953 in August 3b. This indicates that the odds of purchasing an ice cream over a hot drink in August are approximately 17 times higher than in January. This huge difference probably states that people much prefer ice cream in summer than hot drink.



(a) CI of days with fewer than 200 sales

(b) CI of ratio

Figure 3: Confident Intervals

3.2 Results of Part 2

-Null Hypothesis(H_0):*There is no difference between the expected number of sales on weekdays and weekends.*
-Alternative Hypothesis (H_1): *There is a difference between the expected number of sales on weekdays and weekends.*

Upon hypothesizing, we verified the data's normality for t-test applicability using a Q-Q plot 4. The plot revealed substantial normality, despite slight skewness at the extremes. This confirmation justified utilizing an independent two-sample t-test, yielding our results.

From this table 1, a small p-value and a negative t-value led us to reject the null hypothesis, indicating significantly higher sales on weekends compared to weekdays. In fact, average weekend sales are roughly double those on weekdays.

Next, we compute the power of the above test, assuming that the true difference is the one observed, and we get:

$power == 0.9604146$

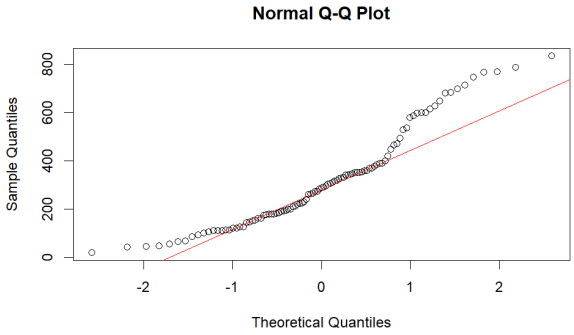


Figure 4: Q-Q plot of ice cream sales

Statistic	Value
Data	weekday_sales and weekend_sales
t-value	-5.3129
Degrees of Freedom (df)	88.233
p-value	8.009×10^{-7}
Alternative Hypothesis	true difference in means is not equal to 0
95% Confidence Interval	[-255.4495, -116.3748]
Sample Estimate (mean of weekday_sales)	225.5192
Sample Estimate (mean of weekend_sales)	411.4314

Table 1: t-test Results for Weekday and Weekend Sales

which implies that our test is highly sensitive to detecting the observed sales difference between weekdays and weekends, bolstering confidence in the t-test results.

If we assume the power is equal to 0.9, we calculate the required effect size and sample size respectively 2. These values help in understanding the sensitivity of our test and inform decisions on sample size requirements for future studies or experiments.

Description	Value
Required effect size for 90% power	0.4606604
Required sample size for 90% power	85.03128

Table 2: Required effect size and sample size for 0.9 power

3.3 Results of Part 3

Using R to help construct and summarize the model, which explains 87.5% of the variance in ice cream sales. We divide all the factors into two categories: Positive factors (ice cream sales will increase if they increase) and Negative factors (sales will decrease if they decrease) 3. Based on different values of those factors, 4 predictions are shown in the table, which includes the predicted sales, lower bounds and upper bounds of their confident intervals 4.

Impact	Factors
Positive	Temperature(°C), humidity(%),Weekend, Bank Holiday, School Holiday
Negative	Windspeed(km/h)

Table 3: Factors Influencing Ice Cream Sales

Scenario	Sales	Lwr (95%)	Upr (95%)
May (18°C, 6%, 10km/h)	185.81	134.26	237.36
Apr (28°C, 35%, 5km/h, School Hol.)	803.97	733.40	874.54
Sep (12°C, 90%, 35km/h)	-12.02	-101.24	77.19
Jan (-2°C, 75%, 15km/h, Weekend)	34.85	-32.07	101.77

Table 4: Predicted Sales and 95% CI for Scenarios

4 Discussion

Although every effort has been made to ensure the accuracy and thoroughness of this report, some findings in this essay herein invite further discussion. For instance, the finding in part 1 could be biased since January and August were not great comparison group. August encompasses school holidays while January is a school term [Uni23]. Shifts in consumer demographics due to holidays can significantly impact comparisons. A further finding of part 3 is that the influences caused by humidity are very small, even can be ignored. However, according to other essays, humidity will have a significant effect on food consumption [CGP+22] [LSYK15], especially under determined weather conditions [WMB03]. The different findings in this report may stem from St Andrews' unique climatic conditions or a limited sample sizes.

5 Summary

In summary, this essay can be divided into 3 parts, which analyzed the ice cream sales based on several factors, notably weather factors. The first part stated that approximately 34.95% of days recorded sales of less than 200 ice creams. A significant difference was shown by odds ratio analysis, it is believed that this huge difference was caused by a common reason that people prefer icy snacks in summer. The second part aimed to discern sales differences between weekdays and weekends. With the tests based on the hypothesis, the results highlighted a notable sales uptick during weekends, which nearly doubled the sales in weekdays. Subsequent power analysis reaffirmed the sensitivity and reliability of the test. In the last part, with utilizing R, a model was constructed explaining 87.5% of the variance in ice cream sales. The given factors were categorized as either positive or negative. Four required predictions were made based on these factors, each accompanied by its respective confidence intervals. In the last section of the content, some improvements and contingent errors were discussed.

References

[CGP+22] Heidy L Contreras, Joaquin Goyret, Clayton T Pierce, Robert A Raguso, and Goggy Davidowitz. Eat, drink, live: Foraging behavior of a nectarivore when relative humidity varies but nectar resources do not. *Journal of Insect Physiology*, 143:104450, 2022.

[LSYK15] Su Jin Lee, Jiyeon Si, Hyun Sun Yun, and GwangPyo Ko. Effect of temperature and relative humidity on the survival of foodborne viruses during food storage. *Applied and environmental microbiology*, 81(6):2075–2081, 2015.

[Uni23] University of St Andrews. Semester Dates - University of St Andrews. `https://www.st-andrews.ac.uk/semester-dates/`, 2023. [Online; accessed 21-October-2023].

[WMB03] JW West, BG Mullinix, and JK Bernard. Effects of hot, humid weather on milk temperature, dry matter intake, and milk yield of lactating dairy cows. *Journal of dairy science*, 86(1):232–242, 2003.

Appendix: R Codes

```
sales_data <- read.csv("mypath/sales_data.csv")
head(sales_data)
library(ggplot2)
library(readr)
library(tidyverse)
library(effsize)
library(pwr)

summary(sales_data$temperature)
summary(sales_data$icecream_sales)

#Temperature Distri.
ggplot(sales_data, aes(x = temperature)) +
geom_histogram(binwidth = 2, fill = "blue", color = "black") +
labs(title = "Temperature distribution histogram", x = "Temperature (°C)", y = "Freque

#Temp. & ice cream sales scatter plot
ggplot(sales_data, aes(x = temperature, y = icecream_sales)) +
geom_point(color = "blue") +
labs(title = "Scatterplot of Sales Quantity vs. Temperature", x = "Temperature (°C)",

#Temp. & hot drink sales scatter plot
ggplot(sales_data, aes(x = temperature, y = hotdrink_sales)) +
geom_point(color = "red") +
labs(title = "Scatterplot of Sales Quantity vs. Temperature", x = "Temperature (°C)",

# Calculate the proportion of days with fewer than 200 ice cream sales
ice_cream_sales_lt_200 <- sum(sales_data$icecream_sales < 200) / nrow(sales_data)

# Calculate the standard error for proportion
se_proportion <- sqrt(ice_cream_sales_lt_200 * (1 - ice_cream_sales_lt_200) / nrow(sal

# Calculate the margin of error
margin_of_error <- 1.96 * se_proportion

# Calculate the 95% confidence interval
confidence_interval <- c(ice_cream_sales_lt_200 - margin_of_error, ice_cream_sales_lt_

# Print the results
cat("1. Expected proportion of days with fewer than 200 ice cream sales:", ice_cream_s
cat("   95% Confidence Interval:", confidence_interval, "\n")

# Calculate the proportion of days with total sales (ice cream + hot drinks) < 200
total_sales_lt_200 <- sum(sales_data$icecream_sales + sales_data$hotdrink_sales < 200)

# Calculate the standard error for proportion
```

```
se_proportion_total <- sqrt(total_sales_lt_200 * (1 - total_sales_lt_200) / nrow(sales

# Calculate the margin of error
margin_of_error_total <- 1.96 * se_proportion_total

# Calculate the 95% confidence interval
confidence_interval_total <- c(total_sales_lt_200 - margin_of_error_total, total_sales

# Print the results
cat("2. Expected proportion of days with fewer than 200 total sales (ice cream + hot c
cat("   95% Confidence Interval:", confidence_interval_total, "\n")

#Draw CI
confidence_intervals <- data.frame(
  Group = c("Ice Cream Sales", "Total Sales"),
  Proportion = c(0.3495146, 0.184466),
  Lower_CI = c(0.2574296, 0.10956),
  Upper_CI = c(0.4415996, 0.259372)
)

ggplot(confidence_intervals, aes(x = Group, y = Proportion)) +
  geom_bar(stat = "identity", fill = "blue", alpha = 0.7) +
  geom_errorbar(aes(ymin = Lower_CI, ymax = Upper_CI), width = 0.2, position = positio
  labs(title = "Proportion of Days with Fewer than 200 Sales",
       y = "Proportion",
       caption = "95% Confidence Intervals") +
  theme_minimal()

#OR
data_January <- data.frame(
  hotdrink = c(335, 149, 102, 131, 220, 167, 234, 176, 118),
  icecream = c(391, 201, 57, 22, 180, 153, 241, 123, 48)
)

data_August <- data.frame(
  hotdrink = c(91,86,232,182,158,181,175,37,28,79,107,57,31,23),
  icecream = c(328,370,616,767,787,682,647,342,164,352,361,386,112,127)
)

total_January <- sum(data_January$hotdrink) + sum(data_January$icecream)
total_August <- sum(data_August$hotdrink) + sum(data_August$icecream)

matrix_January <- matrix(c(
    sum(data_January$icecream), sum(data_January$hotdrink),
    total_January - sum(data_January$icecream),
    total_January - sum(data_January$hotdrink)
), ncol=2)
```

```
matrix_August <- matrix(c(
    sum(data_August$icecream), sum(data_August$hotdrink),
    total_August - sum(data_August$icecream),
    total_August - sum(data_August$hotdrink)
), ncol=2)

result_January <- fisher.test(matrix_January, alternative="two.sided", conf.level=0.95
result_August <- fisher.test(matrix_August, alternative="two.sided", conf.level=0.95)

print(paste("January OR (using Fisher's Exact Test) CI: ", round(result_January$conf.i
print(paste("August OR (using Fisher's Exact Test) CI: ", round(result_August$conf.int

OR_January <- result_January$estimate
OR_August <- result_August$estimate

print(paste("January Odds Ratio (OR): ", round(OR_January, 3)))
print(paste("August Odds Ratio (OR): ", round(OR_August, 3)))

# Data
months <- c("January", "August")
or_values <- c(0.757, 17.016) # using the midpoint of your CI for OR value
lower <- c(0.68, 15.631)
upper <- c(0.834, 18.402)

data <- data.frame(months, or_values, lower, upper)

#Plot
ggplot(data, aes(x = months, y = or_values)) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.2, size = 1) +
  coord_flip() +
  ylab("Odds Ratio with 95% CI") +
  theme_minimal()

#Test QQ & t-test
qqnorm(sales_data$icecream_sales)
qqline(sales_data$icecream_sales, col="red")

weekday_sales <- sales_data[sales_data$weekend == 0, "icecream_sales"]

weekend_sales <- sales_data[sales_data$weekend == 1, "icecream_sales"]

t_test_result <- t.test(weekday_sales, weekend_sales)
print(t_test_result)

#Power
t_value <- -5.3129
n1 <- 45
```

```
n2 <- 45
d <- t_value / sqrt((n1 + n2) / 2)

power_result <- pwr.t2n.test(n1 = n1, n2 = n2, d = d, sig.level = 0.05, alternative =
print(power_result$power)

# Parameters
n = 100  # Replace 100 with your actual sample size.
alpha = 0.05
power = 0.9

# Calculate effect size
result = pwr.t.test(n=n, d=NULL, sig.level=alpha, power=power, type="two.sample", alte

# Print the effect size
result$d

# Parameters
d = 0.5  # Replace 0.5 with your actual effect size or the one from the previous calcu
alpha = 0.05
power = 0.9

# Calculate required sample size
result = pwr.t.test(n=NULL, d=d, sig.level=alpha, power=power, type="two.sample", alte

# Print the required sample size
result$n

#Model
model <- lm(icecream_sales ~ temperature + humidity + windspeed + weekend + bank_holic

summary(model)

#Predictions
# Scenario 1: A week day in May with temperature 18°C, 6% humidity, and 10 km/h windsp
scenario1 <- data.frame(temperature = 18, humidity = 0.06, windspeed = 10,
                        weekend = 0, bank_holiday = 0, school_holidays = 0)

# Scenario 2: A school holiday on a weekend in April with temperature 28°C, 35% humidi
scenario2 <- data.frame(temperature = 28, humidity = 0.35, windspeed = 5,
                        weekend = 1, bank_holiday = 0, school_holidays = 1)

# Scenario 3: A week day in September with temperature 12°C, 90% humidity, and 35 km/h
scenario3 <- data.frame(temperature = 12, humidity = 0.90, windspeed = 35,
                        weekend = 0, bank_holiday = 0, school_holidays = 0)

# Scenario 4: A day on a January weekend that is not a holiday with temperature -2°C,
scenario4 <- data.frame(temperature = -2, humidity = 0.75, windspeed = 15,
```

```
                        weekend = 1, bank_holiday = 0, school_holidays = 0)

predictions1 <- predict(model, newdata = scenario1, interval = "confidence", level = (
predictions2 <- predict(model, newdata = scenario2, interval = "confidence", level = (
predictions3 <- predict(model, newdata = scenario3, interval = "confidence", level = (
predictions4 <- predict(model, newdata = scenario4, interval = "confidence", level = (
```