MT5762 Introductory Data Analysis – Final Project Report

230012908

Abstract

This report provides valuable insights for the ice cream shop owner aiming to optimise their business operations. The analysis reveals that a significant portion of days, approximately 35%, experience lower ice cream sales, making them potentially less profitable for the shop. Moreover, about 18% of the days result in total sales (combining ice cream and hot drinks) below 200 units. This finding underscores the importance of assessing overall sales performance, as nearly one-fifth of days may not be profitable when considering total sales. While the analysis for January indicates a slight preference for hot drinks, August shows a strong inclination towards ice cream. The test confirms the significant seasonal difference in preferences, suggesting that customer choices between ice cream and hot drinks are not independent of the month. The test provides robust evidence that there is a substantial difference in sales between weekdays and weekends. Weekend sales significantly outperform weekday sales, with a power analysis confirming the test's reliability and sensitivity to detecting real-world differences. The results from the multiple regression models offer specific sales predictions under varying conditions. Notably, even during less conventional times, such as weekdays in September or weekends in January, there is a demand for ice cream. These findings allow for more effective inventory and staffing planning, ensuring that customer preferences and behaviour are taken into account. Overall, this analysis equips the ice cream shop owner with essential insights into sales patterns influenced by factors like weather, holidays, and weekends. By understanding these dynamics, the owner can optimise inventory management, staffing, and marketing strategies.

Introduction

The ice cream shop operates in an environment influenced by a host of factors, including weather conditions, day of the week, and special occasions like holidays and school breaks. Understanding the shop's sales patterns and factors affecting them is crucial for effective business planning and decision-making. This report details the methods used to analyse the shop's sales data and presents the findings that can inform these decisions.

Methods

The ice cream shop sales data is stored in a CSV file and includes various variables such as ice cream sales, hot drink sales, temperature, humidity, windspeed, and categorical variables like weekend, month names, school holidays, and bank holidays.

To determine the proportion of days with fewer than 200 ice cream sales, first the number of days meeting this criterion was calculated with the help of 'Hmisc::binconf' function and then using the asymptotic method computation of a 95% confidence interval was performed. Similarly, calculation of the proportion of days with total sales (combining ice cream and hot drinks) below 200 was done and a 95% confidence interval obtained. By finding a standard error, using the 'sqrt' function the odds ratios for purchasing ice cream versus hot drinks in January and August were calculated. This allows one to explore the preference for ice cream during these months. With the help of 'chisq.test', the odds ratios were compared to determine if there is a significant difference between January and August in terms of ice cream versus hot drink purchases.

The data was divided into two subsets, weekdays and weekends, and a two-sample t-test was conducted to examine whether there is a significant difference in sales between these groups. This test aimed to provide insights into the impact of the day of the week on sales of ice cream and hot drinks. To assess the power of the test performed, a power analysis was conducted using the 'rnorm' and 't.test' functions, assuming that the observed difference in sales is the true effect size. The effect size required to achieve a power level of 90% for the observed sample size was determined with the help of the 'pwr' package. Additionally, the sample size needed to achieve a power of 90% for a given effect size was calculated.

A multiple linear regression model was constructed to predict ice cream sales on a weekday in May based on temperature, humidity, windspeed, and other variables such bank and school holidays. Similarly, a regression model was built to predict ice cream sales on a weekend in April, considering school holidays. Furthermore, ice cream sales on a weekday in September, accounting for variables like temperature, humidity, and windspeed were estimated. Finally, a regression model was created to estimate ice cream sales on a weekend in January that is not a holiday. Expected sales and a 95% prediction interval were reported for each multiple regression model.

Results

The analysis revealed that approximately 35% of the days had fewer than 200 ice cream sales, with a 95% confidence interval ranging from 26% to 45%. Approximately 18% of the days had total sales (ice cream and hot drinks combined) below 200, and the 95% confidence interval extended from 11% to 26%. The odds ratios of 0.868 for January and 4.12 for August were obtained. The 'chisq.test' result is that there is a sufficient difference between the odds ratios for ice cream and hot drinks sales in January and August.

The results of the Welch Two Sample t-test provide strong evidence that there is a significant difference in the expected number of sales between weekdays and weekends, 324.4038 and 600.0588 respectively. Specifically, weekend sales are substantially higher on average compared to weekday sales. The power analysis revealed the statistical power of the test – 0.99999. To achieve a power of 90%, an effect size of approximately 0.65 is required. A sample size of around 13 observations is needed for the specified effect size (1.3) and power level (0.99999).

The multiple regression model predicted that, on a weekday in May with a temperature of 18°C, 6% humidity, and 10 km/h windspeed, the expected ice cream sales will approximately be 145 with a 95% prediction interval of [-25.36 – 316.09]. The model estimated that on a weekend in April with temperature 28°C, 35% humidity, and 5 km/h windspeed during school holidays, the expected ice cream sales will approximately be 702, with a 95% prediction interval of [470.7 - 933.89]. The model predicted that on a weekday in September with temperature 12°C, 90% humidity, and 35 km/h windspeed, the expected ice cream sales will approximately be 10, with a 95% prediction interval of [-176.73 - 196]. The model estimated that on a weekend in January that is not a holiday with temperature -2°C, 75% humidity, and 15 km/h windspeed, the expected ice cream sales will approximately be 105, with a 95% prediction interval of [-65.04 - 274.45].

Discussion

The results of the analysis indicate that, on average, a substantial portion of the days (approximately 35%) may experience lower ice cream sales, specifically fewer than 200 units meaning that 35% of the days are not profitable for the ice cream shop owner.

Furthermore, the analysis results reveal that, on average, approximately 18% of the days may have total sales below 200 units when considering both ice cream and hot drinks combined.

This finding provides valuable insights into the overall sales performance of the ice cream shop. Particularly, that about 18% of the days are not profitable for the ice cream shop owner when taking into account total sales.

The odds ratios provide valuable insights into the preferences of customers when it comes to purchasing ice cream versus hot drinks during January and August. For January, an odds ratio of 0.868 was obtained. This odds ratio indicates that during January, there is a slightly lower likelihood of customers choosing ice cream over hot drinks. While this suggests a preference for hot drinks during January, it's important to note that the effect size is relatively modest. In contrast, for August, the odds ratio of 4.12 suggests a significantly higher preference for ice cream over hot drinks. During August, customers are approximately 4.12 times more likely to choose ice cream than hot drinks. This substantial odds ratio implies a strong preference for ice cream during this summer month. The result of the chi-squared test indicates that there is a significant difference between the odds ratios for ice cream and hot drink sales in January and August. This finding holds important implications for understanding how customer preferences shift between these two months.

The chi-squared test is commonly used to assess the independence of two categorical variables, in this case, the preference for ice cream or hot drinks during January and August. The result suggests that the odds ratios for these two months are not independent, meaning that there is a clear and statistically significant association between the month and the choice of products. In practical terms, this means that there is a distinct seasonal pattern in customer preferences. The stronger preference for ice cream in August and a more balanced choice during January are not random occurrences but are statistically significant.

The results of the Welch Two Sample t-test indicate a substantial and statistically significant difference in the expected number of sales between weekdays and weekends. With average sales of 324.4038 on weekdays and 600.0588 on weekends, the findings suggest that weekend sales are notably higher, providing strong evidence that the day of the week significantly influences sales patterns. The substantial difference in average sales between weekdays and weekends has practical implications for the ice cream shop's business operations. It implies that there is a clear and consistent trend of increased customer activity on weekends. This could be due to various factors such as people having more leisure time, family outings, or special promotions offered during weekends. Understanding this pattern allows for more effective resource allocation, such as staffing, inventory management, and marketing efforts.

The power analysis results are quite remarkable, revealing a statistical power of 0.99999. This high statistical power signifies an extremely robust and reliable test, making it almost certain that the test can correctly identify a difference in sales between weekdays and weekends if such a difference truly exists. A statistical power of 0.99999 means that there is only a 0.001% chance of making a Type II error, which is failing to detect a true effect. In practical terms, this implies that the test is exceptionally sensitive and capable of accurately detecting even subtle differences in sales patterns between weekdays and weekends. The high power level adds a significant level of confidence to the results of the test, suggesting that the observed difference in sales between weekdays and weekends is not a random occurrence but a genuine and substantial pattern. It also implies that the test is well-equipped to identify real-world effects, making it a reliable tool for informing business decisions.

The result indicating that an effect size of approximately 0.65 is required to achieve a power of 90% is a key finding in the power analysis. This suggests that in order to detect a significant difference in sales between weekdays and weekends with a 90% probability (power), the effect size, or the magnitude of the difference, needs to be around 0.65. In practical terms, this means that for the test to reliably identify a difference in sales between these two time periods, the change in sales must be substantial, about 0.65 times the standard deviation of the data.

The result indicating that a sample size of around 13 observations is needed to achieve a specified effect size (1.3) and power level (0.99999) is quite informative. It demonstrates that with a relatively small sample size, the test is highly effective at detecting even substantial differences in sales patterns between weekdays and weekends. This is a remarkable level of confidence, as it means one does not need a large dataset to make informed decisions about their business strategy.

The result from the multiple regression model is valuable for the ice cream shop's decision-making process. It provides a specific sales prediction for a particular scenario, which in this case is a weekday in May with certain weather conditions. According to the model, one can expect to sell approximately 145 ice creams in these conditions, and this prediction is accompanied by a 95% prediction interval of [-25.36 – 316.09].

The result from the multiple regression model offers valuable insights for the ice cream shop's operations. It predicts that on a weekend in April, during school holidays, with specific weather conditions (temperature at 28°C, 35% humidity, and 5 km/h windspeed), the expected ice cream sales are estimated to be around 702. This prediction is accompanied by a 95% prediction interval of [470.7 - 933.89]. It allows the owner to anticipate and plan for higher demand during weekends in April when the weather is warm and school holidays are in session.

The prediction from the multiple regression model for ice cream sales on a weekday in September, with specific weather conditions (temperature at 12°C, 90% humidity, and 35 km/h windspeed), suggests that the expected sales will be around 10. This estimate is accompanied by a 95% prediction interval of [-176.73 - 196].The predicted value of 10 indicates that under the specified weather conditions, you can anticipate relatively low ice cream sales on weekdays in September. Factors such as lower temperature and high humidity may contribute to decreased customer demand during this time.

The prediction from the multiple regression model for ice cream sales on a weekend in January, specifically when it's not a holiday, and under particular weather conditions (temperature at -2°C, 75% humidity, and 15 km/h windspeed), suggests that the expected sales will be around 105. This estimate is accompanied by a 95% prediction interval of [-65.04 - 274.45]. The predicted value of 105 indicates that even during the winter month of January, you can expect some level of ice cream sales on weekends, particularly when it's not a holiday. It's interesting to note that certain customers still prefer ice cream under cold weather conditions, possibly due to personal preferences or special occasions. This information can be valuable in planning for inventory and staffing during weekends in January.

Conclusion

To sum it all up, the analysis conducted in this report equips the ice cream shop with crucial information for efficient inventory management, staffing, and marketing strategies. The findings highlight the variability in sales under different conditions and emphasize the importance of accounting for weather, holidays, and weekends when making business decisions. Further research could focus on refining predictive models and exploring additional factors that influence ice cream sales, contributing to enhanced decision-making capabilities for the ice cream shop.

Appendix

1. Tables with outputs/answers

Part 2		
Task	Condition	Answer
A	The expected proportion of days with fewer than 200 ice cream sales and a 95% confidence interval	0.35 [0.257 , 0.442]
B	The expected proportion of days with fewer than 200 total sales (ice cream and hot drinks) and a 95% confidence interval	0.184 [0.11 , 0.259]
C	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.	0.868 4.12
D	Is there a significant difference in odds ratios between January and August?	2.2e-16

Part 3		
Task	Condition	Answer
A	Testing whether there is a difference between the expected number of sales on week days (Mon-Fri) and weekends. Interpret and explain your results.	324.4038 600.0588
B	Computing the power of the above test, assuming that the true difference is the one observed.	0.99999
C	For the observed sample size, what effect size (i.e., difference between the expected values) would be required to obtain a power of 90%?	0.65
D	For the given effect size, what sample size would be required to obtain a power of 90%?	13

Part 4		
Task	Condition	Answer
A	Estimating the expected number of ice creams that the ice cream	145.37 [-25.36 - 316.09]

	shop can expect to sell on days with the following characteristics, together with 95% bounds, for: a week day in May with temperature 18 C, 6% humidity, and 10 km/h windspeed	
B	a school holiday on a weekend in April with temperature 28 C, 35% humidity, and 5 km/h windspeed	702.3 [470.7 - 933.89]
C	a week day in September with temperature 12 C, 90% humidity, and 35 km/h windspeed	9.63 [-176.73 - 196]
D	a day on a January weekend that is not a holiday with temperature -2 C, 75% humidity, and 15 km/h windspeed	104.71 [-65.04 - 274.45]

2. Code

```
# Introductory data analysis - Final Project Report

library(readr)
sales_data <- read_csv("~/sales_data.csv")
View(sales_data)

library(tidyverse)

## PART 2

### a. The expected proportion of days with fewer than 200 ice
cream sales and a 95% confidence interval

days_below_200 <- sum(sales_data$icecream_sales < 200)
days_below_200

library(Hmisc)
prop_ci <- binconf(x = days_below_200, n =
length(sales_data$icecream_sales), alpha = 0.05, method =
'asymptotic') |>
  round(3)
prop_ci

cat("Expected Proportion of Days with Fewer than 200 Ice Cream
Sales:", prop_ci[1], "\n")
cat("95% Confidence Interval for Days with Fewer than 200 Ice Cream
Sales: [", prop_ci[2], ", ", prop_ci[3], "]\n")
```

```
### b. The expected proportion of days with fewer than 200 total
sales (ice cream and hot drinks) and a 95% confidence interval

total_sales <- sales_data$icecream_sales +
sales_data$hotdrink_sales
total_sales

total_days_below_200 <- sum(total_sales < 200)
total_days_below_200

prop_ci2 <- binconf(x = total_days_below_200, n =
length(total_sales), alpha = 0.05, method = 'asymptotic') |>
  round(3)
prop_ci2

cat("Expected Proportion of Days with Fewer than 200 Total Sales
(ice cream and hoe drinks):", prop_ci2[1], "\n")
cat("95% Confidence Interval: [", prop_ci2[2], ", ", prop_ci2[3],
"]\n")

### c. 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.

january <- sales_data %>%
  filter(month_name == "Jan")
august <- sales_data %>%
  filter(month_name == "Aug")

# Calculating odds ratios for January and August
oddsratio_january <- january %>%
  summarise(icecream = sum(icecream_sales), hotdrinks =
sum(hotdrink_sales)) %>%
  mutate(oddsratio = icecream / hotdrinks)

oddsratio_august <- august %>%
  summarise(icecream = sum(icecream_sales), hotdrinks =
sum(hotdrink_sales)) %>%
  mutate(oddsratio = icecream / hotdrinks)

oddsratio_january$selogOR <- sqrt(1 / oddsratio_january$icecream
+ 1 / oddsratio_january$hotdrinks)
oddsratio_august$selogOR <- sqrt(1 / oddsratio_august$icecream +
1 / oddsratio_august$hotdrinks)

print(oddsratio_january)
print(oddsratio_august)

### d. Is there a significant difference in odds ratios between
January and August?
```

```
oddsratio_january_CI <-
exp(cbind(log(oddsratio_january$oddsratio) - qnorm(0.975) *
oddsratio_january$selogOR,

log(oddsratio_january$oddsratio) + qnorm(0.975) *
oddsratio_january$selogOR))

oddsratio_august_CI <- exp(cbind(log(oddsratio_august$oddsratio)
- qnorm(0.975) * oddsratio_august$selogOR,
                                 log(oddsratio_august$oddsratio)
+ qnorm(0.975) * oddsratio_august$selogOR))

print(oddsratio_january_CI)
print(oddsratio_august_CI)

# Performing a chi-squared test to compare the odds ratios
OR_test <- chisq.test(matrix(c(oddsratio_january$icecream,
oddsratio_august$icecream,
                               oddsratio_january$hotdrinks,
oddsratio_august$hotdrinks), ncol = 2))

# Printing the results
print(OR_test)


## PART 3

# Please highlight the code in Part 3 A and run separately from
the rest

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

# Creating two data subsets for weekdays and weekends
weekday_data <- sales_data[sales_data$weekend == 0, ]
weekend_data <- sales_data[sales_data$weekend == 1, ]

sales_data$total_num_sales <- sales_data$hotdrink_sales +
sales_data$icecream_sales

# Performing a two-sample t-test for sales
ttest <- t.test(as.numeric(weekday_data$total_num_sales),
as.numeric(weekend_data$total_num_sales))

# Printing out the results
print("Week days sales vs weekends sales:")
print(ttest)

### b. Computing the power of the above test, assuming that the
true difference is the one observed.
```

```
# Calculating the observed mean difference between weekday and
weekend sales
observed_diff <- mean(weekday_data$total_num_sales) -
mean(weekend_data$total_num_sales)
observed_diff

# Setting the parameters for the simulation
set.seed(123) # Setting a seed for reproducibility
nsim <- 100000 # Number of simulations
value <- rep(0, nsim)

# Defining the critical value based on the significance level
(alpha)
alpha <- 0.05
zcrit <- qnorm(p = 1 - alpha/2) # Two-tailed test

# Beginning the simulation loop
for (i in 1:nsim) {

  # Generating random samples for both weekday and weekend data
  sample_weekday <- rnorm(length(weekday_data$total_num_sales),
mean = mean(weekday_data$total_num_sales), sd =
sd(weekday_data$total_num_sales))

  sample_weekend <- rnorm(length(weekend_data$total_num_sales),
mean = mean(weekend_data$total_num_sales), sd =
sd(weekend_data$total_num_sales))

  # Performing a two-sample t-test
  t_test_result <- t.test(sample_weekday, sample_weekend)

  # Checking if the null hypothesis is rejected (i.e., p-value is
less than alpha/2 for a two-tailed test)
  if (t_test_result$p.value < alpha/2) {
    value[i] <- 1
  }
}
# Calculating the power as the proportion of rejections
power <- sum(value) / nsim |> round(3)

# Printing the power
cat("Power of the test:", power, "\n")

### c. For the observed sample size, what effect size (i.e.,
difference between the expected values) would be required to obtain
a power of 90%?

# Installing and loading the pwr package for power analysis
install.packages("pwr")
library(pwr)

# Setting parameters
alpha <- 0.05 # Significance level
```

```
power_target <- 0.90 # Desired power level
sample_size <- length(sample_weekday) # Observed sample size
# Initialising effect size and power variables
effect_size <- 0.01 # Starting effect size
power <- 0

# Performing a power analysis and iteratively adjusting the effect
size
while (power < power_target) {
  power <- pwr.t.test(d = effect_size, n = sample_size, sig.level
= alpha, power = NULL)$power
  effect_size <- effect_size + 0.01
}

# The loop will stop when the desired power level is reached
cat("Effect size required for a power of 90%:", round(effect_size
- 0.01, 2), "\n")

### d. For the given effect size, what sample size would be
required to obtain a power of 90%?

# Setting the effect size and desired power
effect_size <- 1.3
desired_power <- 0.90

# Calculating the sample size
sample_size <- pwr.t.test(
  d = effect_size,
  power = desired_power,
  sig.level = 0.05, # Significance level (alpha)
  type = "two.sample" # For a two-sample t-test
)$n

# Printing the calculated sample size
cat("Required sample size:", round(sample_size))

## PART 4

### Estimating the expected number of ice creams that the ice cream
shop can expect to sell on days with the following characteristics,
together with 95% bounds, for:

### a. a week day in May with temperature 18 C, 6% humidity, and
10 km/h windspeed

library(dplyr)
library(ggplot2)

# Fitting a multiple linear regression model
model1 <- lm(icecream_sales ~ temperature + humidity + windspeed
+ weekend + month_name + bank_holiday + school_holidays, data =
sales_data)
```

```
summary(model1)
AIC(model1)

new_data1 <- data.frame(
  temperature = 18,
  humidity = 6,
  windspeed = 10,
  weekend = 0,
  month_name = 'May',
  bank_holiday = 0,
  school_holidays = 0
)

# Predicting ice cream sales for the new data point
predicted_sales1 <- predict(model1, newdata = new_data1, interval
= "prediction")
predicted_sales1

# Extracting the mean and 95% prediction intervals
mean_sales1 <- predicted_sales1[1]
lower_bound1 <- predicted_sales1[2]
upper_bound1 <- predicted_sales1[3]

cat("Expected Ice Cream Sales on a Weekday in May with the Given
Characteristics: ", round(mean_sales1, 2), "\n")
cat("95% Prediction Interval: [", round(lower_bound1, 2), " - ",
round(upper_bound1, 2), "]\n")

### b. a school holiday on a weekend in April with temperature 28
C, 35% humidity, and 5 km/h windspeed

model2 <- lm(icecream_sales ~ temperature + humidity + windspeed
+ weekend + month_name + bank_holiday + school_holidays, data =
sales_data)

summary(model2)
AIC(model2)

new_data2 <- data.frame(
  temperature = 28,
  humidity = 35,
  windspeed = 5,
  weekend = 1,
  month_name = 'Apr',
  bank_holiday = 0,
  school_holidays = 1
)

predicted_sales2 <- predict(model2, newdata = new_data2, interval
= "prediction")
predicted_sales2

# Extracting the mean and 95% prediction intervals
```

```
mean_sales2 <- predicted_sales2[1]
lower_bound2 <- predicted_sales2[2]
upper_bound2 <- predicted_sales2[3]

cat("Expected Ice Cream Sales on a weekend in April with the given
characteristics: ", round(mean_sales2, 2), "\n")
cat("95% Prediction Interval: [", round(lower_bound2, 2), " - ",
round(upper_bound2, 2), "]\n")

### c. a week day in September with temperature 12 C, 90% humidity,
and 35 km/h windspeed

model3 <- lm(icecream_sales ~ temperature + humidity + windspeed
+ weekend + month_name + bank_holiday + school_holidays, data =
sales_data)

summary(model3)
AIC(model3)

new_data3 <- data.frame(
  temperature = 12,
  humidity = 90,
  windspeed = 35,
  weekend = 0,
  month_name = 'Sep',
  bank_holiday = 0,
  school_holidays = 0
)

predicted_sales3 <- predict(model3, newdata = new_data3, interval
= "prediction")
predicted_sales3

# Extracting the mean and 95% prediction intervals
mean_sales3 <- predicted_sales3[1]
lower_bound3 <- predicted_sales3[2]
upper_bound3 <- predicted_sales3[3]

cat("Expected Ice Cream Sales on a week day in September with the
given characteristics: ", round(mean_sales3, 2), "\n")
cat("95% Prediction Interval: [", round(lower_bound3, 2), " - ",
round(upper_bound3, 2), "]\n")

### d. a day on a January weekend that is not a holiday with
temperature -2 C, 75% humidity, and 15 km/h windspeed

model4 <- lm(icecream_sales ~ temperature + humidity + windspeed
+ weekend + month_name + bank_holiday + school_holidays, data =
sales_data)
summary(model4)
AIC(model4)

new_data4 <- data.frame(
  temperature = -2,
  humidity = 75,
  windspeed = 15,
  weekend = 1,
  month_name = 'Jan',
  bank_holiday = 0,
  school_holidays = 0
)

predicted_sales4 <- predict(model4, newdata = new_data4, interval
= "prediction")
predicted_sales4

# Extracting the mean and 95% prediction intervals
mean_sales4 <- predicted_sales4[1]
lower_bound4 <- predicted_sales4[2]
upper_bound4 <- predicted_sales4[3]

cat("Expected Ice Cream Sales on a weekend in January with the
given characteristics: ", round(mean_sales4, 2), "\n")
cat("95% Prediction Interval: [", round(lower_bound4, 2), " - ",
round(upper_bound4, 2), "]\n")
```