Final Project - First Analysis

Programming for Business Analytics (11410ISS 406600)

Group 10

2025-11-03

Group Member

Choonhanunt Amnuaychaikij 111006224 Panuvit Tuicharoen 111006105 Napat Leesaksakul 112006208 Thanutcha Amnajcharoensak 112006412 Nantachaporn Sudjai 112006431

In this report, we will analyze some of the interested business question by using the analysis techniques.

1. Does setting a higher price actually reduce the number of units people buy (Price Elasticity) by using correlation between Price vs Quantity.
2. Do highly-rated products generate more money per transaction? by using the linear regression between Rating vs Revenue.
3. Which product category is statistically the most consistent in quality? by using the confidence intervals.

Library Setup

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stringr)
library(dplyr)
library(ggplot2)
library(lubridate)
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.5.2
```

Data Importing

```
ecommerce <- read.csv('ecommerce_dataset_10000.csv')

head(ecommerce)
```

```
##   customer_id  first_name last_name gender age_group signup_date country
## 1    CUST2353       Erica    Oliver Female Teenagers  2022-06-29  Canada
## 2    CUST4463 Christopher     White   Male    Adults  2023-08-24   China
## 3    CUST4512     Spencer    Foster   Male    Senior  2023-07-18 Germany
## 4    CUST5711     Jessica    Harris   Male Teenagers  2025-08-22  France
## 5    CUST1296         Amy   Johnson Female Teenagers  2021-03-23  Brazil
## 6    CUST2790      Shelby    Sutton  Other    Adults  2025-07-18  Canada
##   product_id         product_name       category quantity unit_price order_id
## 1    PROD108       Fitbit Versa 3    Electronics        3        229 ORD10000
## 2    PROD103         Levi's Jeans        Apparel        4         59 ORD10001
## 3    PROD111   Lego Star Wars Set           Toys        2         59 ORD10002
## 4    PROD107         Dyson Vacuum Home & Kitchen        4        399 ORD10003
## 5    PROD105 Adidas Running Shoes        Apparel        1        110 ORD10004
## 6    PROD108       Fitbit Versa 3    Electronics        5        229 ORD10005
##   order_date order_status   payment_method rating review_text review_id
## 1 2023-07-13      Pending      Credit Card      2        good  REV20000
## 2 2024-08-12      Pending           PayPal      2     average  REV20001
## 3 2024-08-04    Delivered Cash on Delivery      5        good  REV20002
## 4 2025-05-23    Delivered Cash on Delivery      2   very good  REV20003
## 5 2023-07-02     Returned Cash on Delivery      1   very good  REV20004
## 6 2023-04-13     Returned           PayPal      3   very good  REV20005
##   review_date
## 1  2025-06-06
## 2  2023-08-05
## 3  2023-01-03
## 4  2023-03-14
## 5  2023-10-18
## 6  2023-02-14
```

Data Cleaning

```
ecommerce <- ecommerce %>%
  mutate(
    order_year = year(ymd(order_date))
  )
```

Data Preparation

```
ecommerce <- ecommerce %>%
  mutate(
    total_amount = quantity * unit_price
  )
```

Question 1: Correlation Analysis

We want to determine if a higher unit price relates to lower quantity purchased.

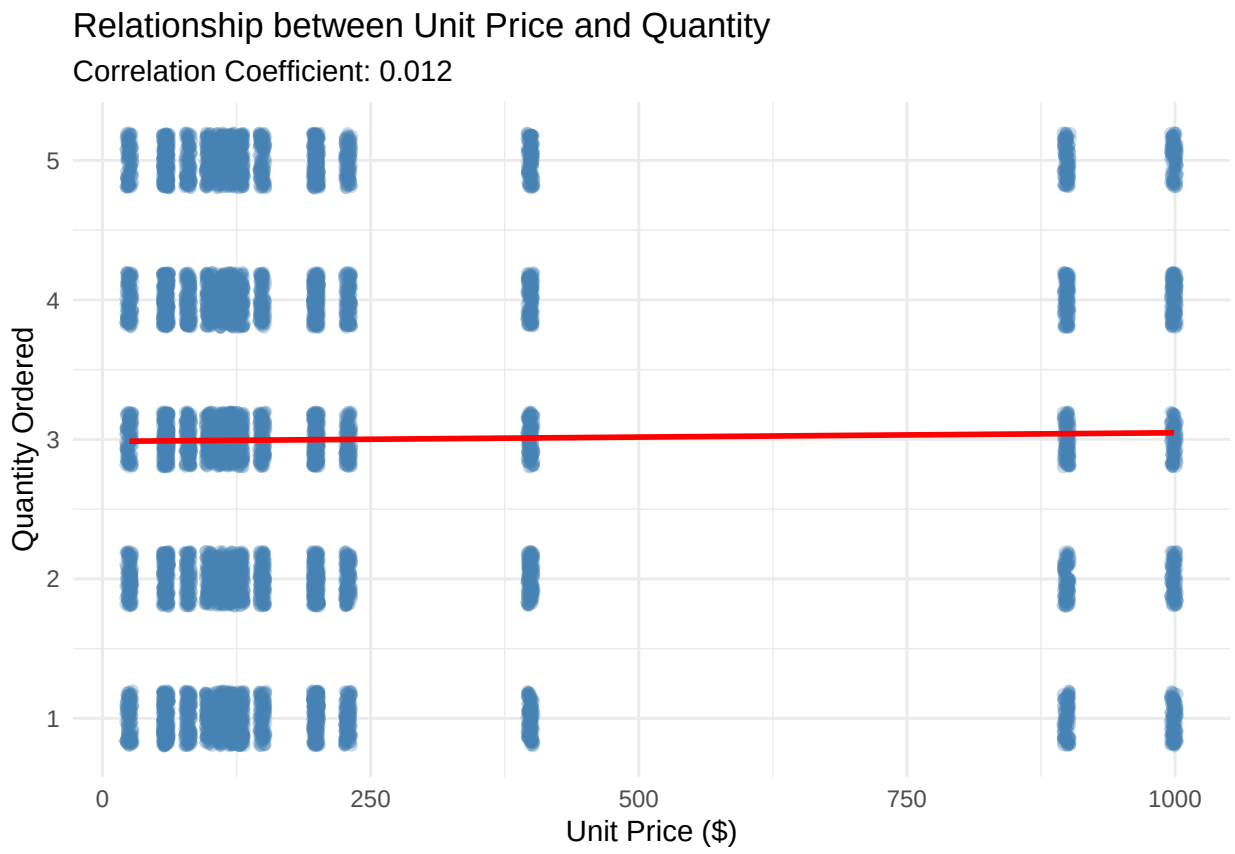
```
cor_test <- cor.test(ecommerce$unit_price, ecommerce$quantity)

cor_test
```

```
##
##  Pearson's product-moment correlation
##
## data:  ecommerce$unit_price and ecommerce$quantity
## t = 1.2335, df = 9998, p-value = 0.2174
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.007266491  0.031927687
## sample estimates:
##        cor
## 0.01233534
```

We also make the visualization for the correlation

```
ggplot(ecommerce, aes(x = unit_price, y = quantity)) +
  geom_jitter(alpha = 0.3, height = 0.2, color = "steelblue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(
    title = "Relationship between Unit Price and Quantity",
    subtitle = paste("Correlation Coefficient:", round(cor_test$estimate, 3)),
    x = "Unit Price ($)",
    y = "Quantity Ordered"
  ) + theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Question 2: Linear Regression

We want to predict that if Customer Rating influences the Total Sales Amount.

```
lm_model <- lm(total_amount ~ rating, data = ecommerce)

summary(lm_model)
```

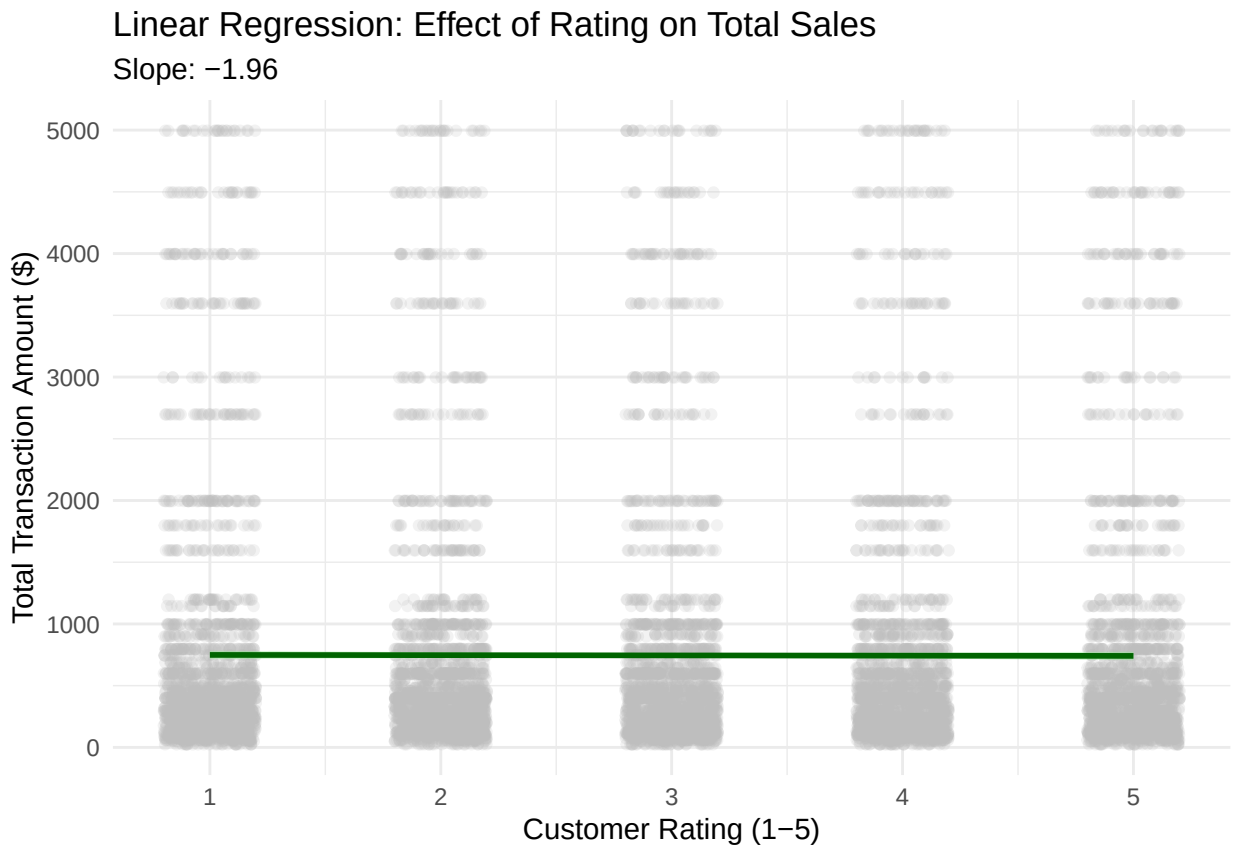
```
##
## Call:
## lm(formula = total_amount ~ rating, data = ecommerce)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -724.0 -548.0 -349.0   50.9 4253.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  750.955     23.751  31.618   <2e-16 ***
## rating        -1.964      7.178  -0.274    0.784
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1012 on 9998 degrees of freedom
```

```
## Multiple R-squared:  7.488e-06,  Adjusted R-squared:  -9.253e-05
## F-statistic: 0.07486 on 1 and 9998 DF,  p-value: 0.7844
```

Next we will proceed with the visualization,

```
ggplot(ecommerce, aes(x = rating, y = total_amount)) +
  geom_jitter(alpha = 0.2, width = 0.2, color = "gray") +
  geom_smooth(method = "lm", color = "darkgreen", fill = "lightgreen") +
  labs(
    title = "Linear Regression: Effect of Rating on Total Sales",
    subtitle = paste("Slope:", round(coef(lm_model)[2], 2)),
    x = "Customer Rating (1-5)",
    y = "Total Transaction Amount ($)"
  ) + theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Question 3: Confidence Intervals

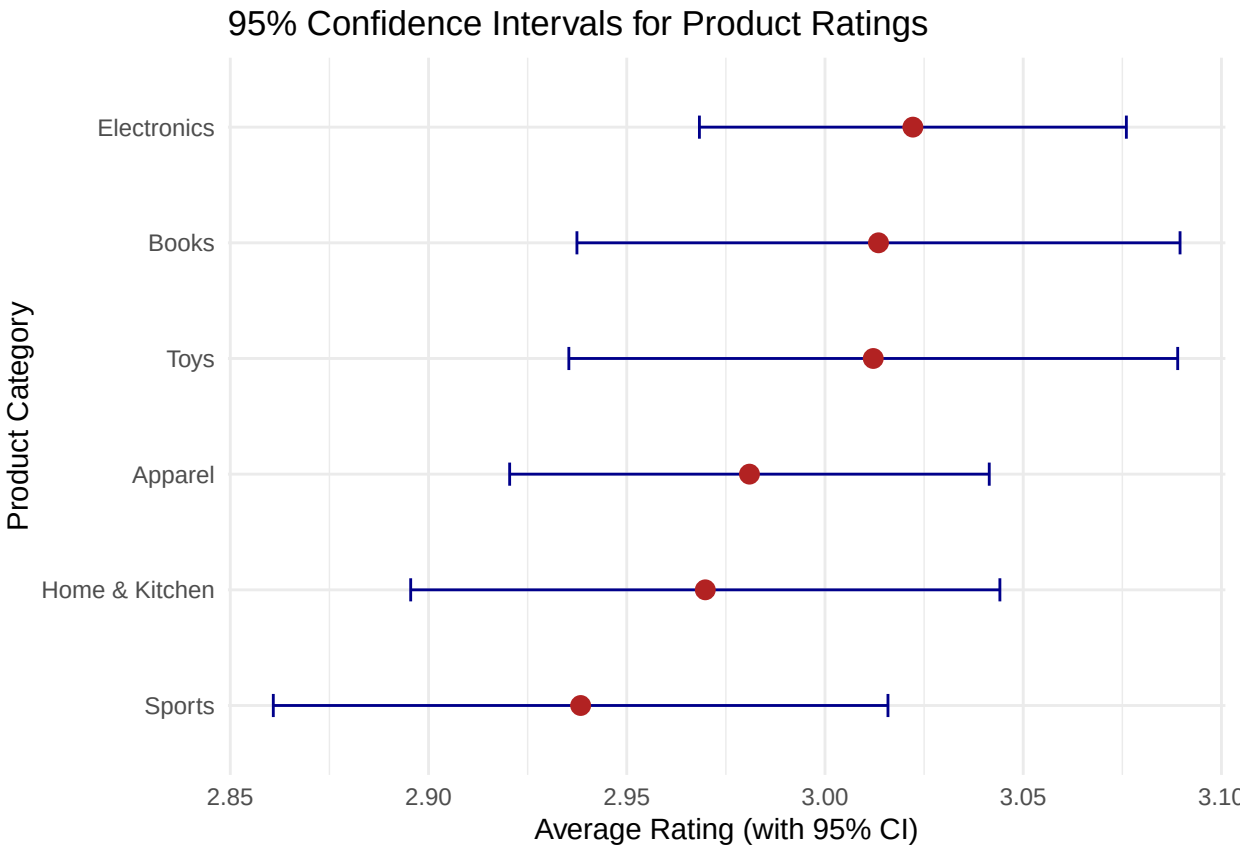
As we want to find the true average rating for each product category with statistical confidence.

```r
category_ci <- ecommerce %>%
  group_by(category) %>%
  summarise(
    avg_rating = mean(rating),
    sd_rating = sd(rating),
    count = n(),
    se = sd_rating / sqrt(count),
    # CI 95% calculation
    ci_lower = avg_rating - (1.96 * se),
    ci_higher = avg_rating + (1.96 * se)
  )

category_ci
```

```
## # A tibble: 6 x 7
##   category       avg_rating sd_rating count     se ci_lower ci_higher
##   <chr>               <dbl>     <dbl> <int>  <dbl>    <dbl>     <dbl>
## 1 Apparel              2.98      1.40  2047 0.0309     2.92      3.04
## 2 Books                3.01      1.42  1334 0.0388     2.94      3.09
## 3 Electronics          3.02      1.41  2616 0.0275     2.97      3.08
## 4 Home & Kitchen       2.97      1.41  1391 0.0379     2.90      3.04
## 5 Sports               2.94      1.42  1298 0.0395     2.86      3.02
## 6 Toys                 3.01      1.42  1314 0.0392     2.94      3.09
```

Next, we will working on the visualization for error bars.

```r
ggplot(category_ci, aes(x = reorder(category, avg_rating), y = avg_rating)) +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_higher), width = 0.2, color = "darkblue") +
  geom_point(size = 3, color = "firebrick") +
  labs(
    title = "95% Confidence Intervals for Product Ratings",
    x = "Product Category",
    y = "Average Rating (with 95% CI)"
  ) + coord_flip() + theme_minimal()
```


95% Confidence Intervals for Product Ratings
Product Category
Electronics
Books
Toys
Apparel
Home & Kitchen
Sports
2.85
2.90
2.95
3.00
3.05
3.10
Average Rating (with 95% CI)