Price Analysis of Luxury Watches: Trends and Determinants

Statistical Analysis of Key Features Influencing Luxury Watch Pricing

Abey Thomas s4059720,Ananya George Martin s4109526,SamuelJohn s4119690

Last updated: 24 October, 2024

Introduction

Luxury timepieces are more than just instruments for keeping time; they embody craftsmanship, tradition, and status.Luxury watches are renowned for their craftsmanship and status, yet the factors driving their prices remain complex and multifaceted.This research seeks to dissect the myriad factors influencing the cost of luxury watches, focusing on materials, movement types, and additional features. Utilizing statistical tools, we aim to provide a comprehensive understanding of these influences.The overarching objective is to aid retailers and consumers in making informed decisions in the luxury watch market by highlighting the significance of various features, with Price as the primary focus. Other factors such as case dimensions, strap material, and movement type are equally critical in shaping the value of luxury watches.We will examine these variables using rigorous statistical tests to ensure robust insights.

This study investigates the relationship between key variables—such as strap material, case dimensions, and movement type—and the cost of luxury watches. By analyzing descriptive statistics, conducting hypothesis testing, and exploring correlations, the research seeks to demystify the pricing mechanisms and provide actionable insights for stakeholders in the luxury watch industry.

- **Statistical Approach**:
  - ***Descriptive Statistics****: Analyze mean, median, and price range.*
  - ***Hypothesis Testing***:
    - T-tests for price differences based on most common strap materials
    - Chi-square tests for categorical associations (movement type vs. strap material).
  - ***Correlation****: Study for Price and various features.*

Data

Data Collection Method

The dataset was downloaded directly from Kaggle.

Source: Rattanaporn K (n.d.) *Luxury Watches Price Dataset*, Kaggle website, accessed 19 October 2024. https://www.kaggle.com/datasets/rkiattisak/luxury-watches-price-dataset/data

Data Description

The *Luxury Watches Price Dataset* provides a comprehensive collection of luxury watch listings, detailing essential attributes such as brand, model, price, materials, and features. Here are the variables included in the dataset:

- **Brand**: Indicates the luxury watch brand.
- **Strap Material**:Indicates the type of material used for strap material
- **Case Material**: Specifies the material of the watch case.
- **Movement Type**: Indicates the type of movement (e.g., Automatic, Quartz).
- **Dial Color**: Specifies the color of the watch dial.
- **Case Diameter (mm)**: Represents the diameter of the watch case in millimeters.
- **Case Thickness (mm)**: Represents the thickness of the watch case in millimeters.
- **Band Width (mm)**: Indicates the width of the watch band in millimeters.
- **Price (USD)**: Represents the price of the watch in US dollars.

Preprocessing Steps

- **Subsetted Data**: Dropped variables "Complications" and "Power Reserve" as they were assumed to be less relevant for our analysis.
- **Filtered Data**: Focused on the most common strap materials, Leather and Stainless Steel, to streamline the analysis. These materials were understood to be common choices among the dataset.
- **Variable Data Type Conversions**: Converted Water Resistance into an ordered factor.
- **Converted Categorical Variables**: Transformed other categorical variables into factors, for instance, converting Dial Color into a factor.
- **Other pre processing steps**: We also scanned for missing values, outliers and did some tranformations for Price variable.

Lets check for the pre processing steps in detail; We used readr package to import the csv file.

```
#Reading the file using read_csv
watch <- read_csv("C:/Users/ANANYA/OneDrive/Desktop/Luxury watch.csv")
```

Lets have a look on the variable names

```
#display the names of the variable
colnames(watch)
```

```
##  [1] "Brand"               "Model"               "Case Material"
##  [4] "Strap Material"      "Movement Type"       "Water Resistance"
##  [7] "Case Diameter (mm)"  "Case Thickness (mm)" "Band Width (mm)"
## [10] "Dial Color"          "Crystal Material"    "Complications"
## [13] "Power Reserve"       "Price (USD)"
```

Preprocessing Steps Cont.

Our analysis will primarily focus on the prices of watches, specifically examining the most common strap materials.

```
#Check for number of unique values of strap material
watch$`Strap Material` %>% table()
```

We will focus on Leather and Stainless Steel straps, removing columns like Complications and Power Reserve as less relevant. Water Resistance will be converted to ordered factors, along with other character variables. These are our preprocessing steps.

```
# Filter the data based on most common strap materials
watch2 <- watch %>% filter(`Strap Material` %in% c("Leather", "Stainless Steel"))
#ignoring 2 columns
watch2 <- watch2[, !names(watch2) %in% c("Complications", "Power Reserve")]
```

```
#Checking the unique values in `Water Resistance` and the possibility of conversion to ordered factor
watch2$`Water Resistance` %>% table()
```

```
# Convert Water Resistance column to numeric
watch2$`Water Resistance` <- as.numeric(gsub("[^0-9]", "", watch2$`Water Resistance`))
# Define the categories based on the unique water resistance values
watch2 <- watch2 %>%mutate(`Water Resistance Category` = cut(`Water Resistance`,breaks =
c(-Inf, 60, 150, 300,Inf),labels = c("Low", "Medium", "High", "Very High"),right = TRUE))
```

```
# Convert Water Resistance Category to an ordered  factor
watch2$`Water Resistance Category` <- factor(watch2$`Water Resistance Category`, levels = c("Low", "Medium", "High",
"Very High"), ordered = TRUE)
#Dropping the old variable
watch2 <- watch2[, !names(watch2) %in% "Water Resistance"]
```

```
#Other data type conversions
watch2$`Case Material` = as.factor(watch2$`Case Material`)
watch2$`Strap Material` = as.factor(watch2$`Strap Material`)
watch2$`Movement Type` = as.factor(watch2$`Movement Type`)
watch2$`Dial Color` = as.factor(watch2$`Dial Color`)
watch2$Brand = as.factor(watch2$Brand)
```

We verified all the type conversions that are required and now we have factors, numeric and characters(Brand and Model).

Descriptive Statistics and Visualisation

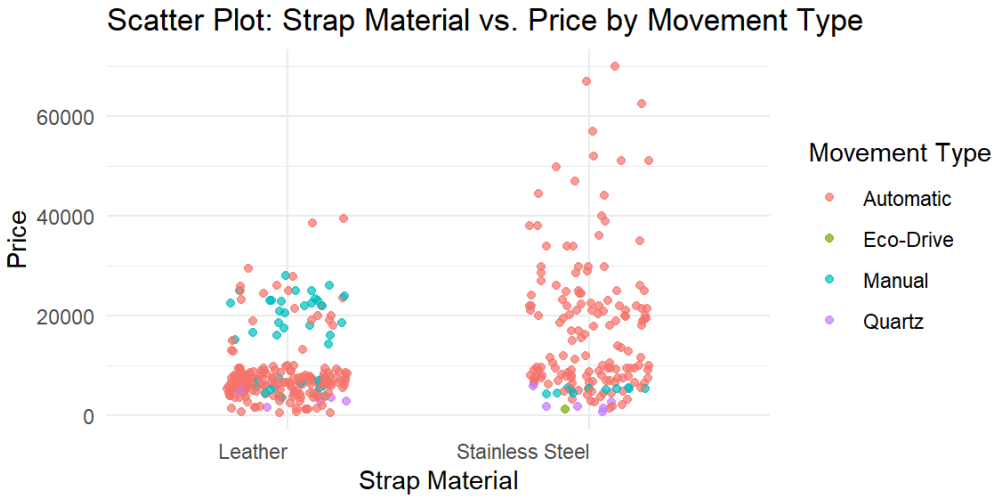
Price is our key variable in our analysis. To provide a comprehensive understanding of watch prices, we summarized the descriptive statistics for two common strap materials: Leather and Stainless Steel. By calculating key metrics like minimum, quartiles, median, maximum, and mean, we gain insights into the price distribution across these categories. Here's the code used to achieve this:

```
        # Group data by Strap Material and calculate descriptive statistics for Price (USD)
        price_desc_stat <- watch2 %>%
          group_by(`Strap Material`) %>%
          summarise(Min = min(`Price (USD)`, na.rm = TRUE),Q1 = quantile(`Price (USD)`, 0.25, na.rm = TRUE),Median =
median(`Price (USD)`, na.rm = TRUE),Q3 = quantile(`Price (USD)`, 0.75, na.rm = TRUE),Max = max(`Price (USD)`, na.rm = TRUE),Mean =
mean(`Price (USD)`, na.rm = TRUE))

        #Print the result
        print(price_desc_stat)
```

```
## # A tibble: 2 × 7
##   `Strap Material`    Min    Q1 Median    Q3   Max   Mean
##   <fct>             <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>
## 1 Leather             495  5200   6900  9500 39500  9402.
## 2 Stainless Steel     650  6500  10550 22000 70000 16595.
```

```
        #scatter plot to visualize the strap materials and price, color-coded by movement type.
        ggplot(watch2, aes(x = `Strap Material`, y = `Price (USD)`, color = `Movement Type`)) +
          geom_jitter(width = 0.2, alpha = 0.7) +
          labs(title = "Scatter Plot: Strap Material vs. Price by Movement Type", x = "Strap Material", y = "Price") +
          theme_minimal() +
          theme(axis.text.x = element_text(hjust = 1))
```



Our dataset features over 35 luxury watch brands, with leather and stainless steel as the most popular strap materials. Water resistance varies from 30 to 300 meters, and seven shades of black highlight color diversity. Prices range from $485 to $70,000, covering a broad economic spectrum. Automatic movement watches dominate, with stainless steel straps being more common than leather. Lets check for the data quality issues in the coming slides.

Missing Value Treatment

We need to check for missing values in any of the columns.

```
#Verifying the missing values in individual column
colSums(is.na(watch2))
```

```
##                      Brand                      Model            Case Material
##                          0                          0                        0
##            Strap Material             Movement Type        Case Diameter (mm)
##                          0                          0                        0
##       Case Thickness (mm)           Band Width (mm)                Dial Color
##                          0                          0                        0
##          Crystal Material              Price (USD) Water Resistance Category
##                          0                          0                        0
```

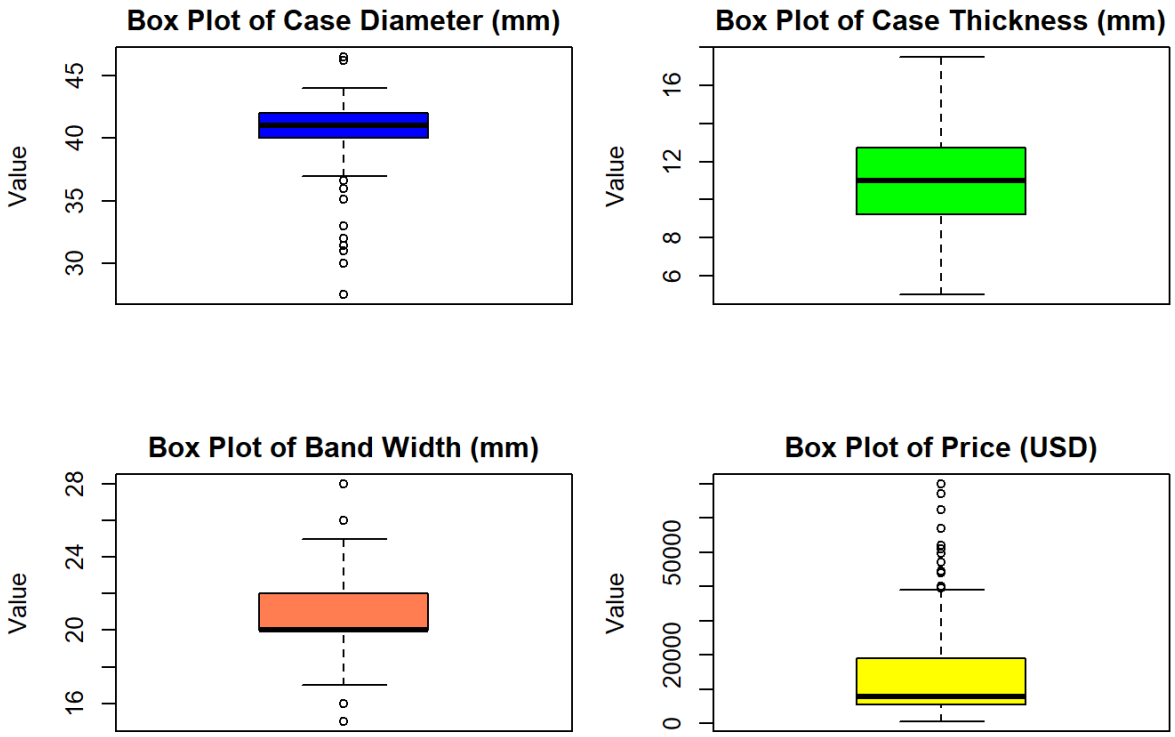
No missing values are found in any of the columns.Now lets check for outliers.

Outliers

We have four numerical columns: Case Diameter (mm), Case Thickness (mm), BandWidth (mm), and Price (USD). We need to check for outliers before analysis.

```
        # Set up the plotting area for 2x2 grid
        par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))

        # Box plots and capturing outliers
        Out1 <- boxplot(watch2$`Case Diameter (mm)`, main = "Box Plot of Case Diameter (mm)", ylab = "Value", col = "blue")$out
        Out2 <- boxplot(watch2$`Case Thickness (mm)`, main = "Box Plot of Case Thickness (mm)", ylab = "Value", col =
"green")$out
        Out3 <- boxplot(watch2$`Band Width (mm)`, main = "Box Plot of Band Width (mm)", ylab = "Value", col = "coral")$out
        Out4 <- boxplot(watch2$`Price (USD)`, main = "Box Plot of Price (USD)", ylab = "Value", col = "yellow")$out
```



We identified outliers in Case Diameter, Band Width, and Price (USD). To maintain the integrity of our analysis, we retained the outliers for Case Diameter and Price since they represent genuine variations in luxury watch data and are crucial for determining market prices. Instead of removing these outliers, we applied transformations to better understand their distribution and impact.

Outliers Cont.

For Band Width, however, we capped the outliers to balance preserving essential information with enhancing the reliability of our analysis. This approach allowed us to address data anomalies without compromising the key insights from our dataset. Lets check for the codes for capping Band width

```
# Set the percentile thresholds for capping
lower_bound <- 0.06  # 1st percentile
upper_bound <- 0.95  # 99th percentile
```

```
#create a function to apply capping
cap_outliers <- function(x, lower_bound, upper_bound) {
  # Compute the percentile values
  lower_val <- quantile(x, probs = lower_bound, na.rm = TRUE)
  upper_val <- quantile(x, probs = upper_bound, na.rm = TRUE)

  # Cap the values below the lower bound and above the upper bound
  x[x < lower_val] <- lower_val
  x[x > upper_val] <- upper_val
  return(x)}
# Apply the capping to 'Band Width (mm)'
watch2$`Band Width (mm)` <- cap_outliers(watch2$`Band Width (mm)`, lower_bound, upper_bound)
```

```
# Create the boxplot for band width after capping
out_ct2 <- boxplot.stats(watch2$`Band Width (mm)`)
outliers <- out_ct2$out

# Verify the outliers
outliers
```

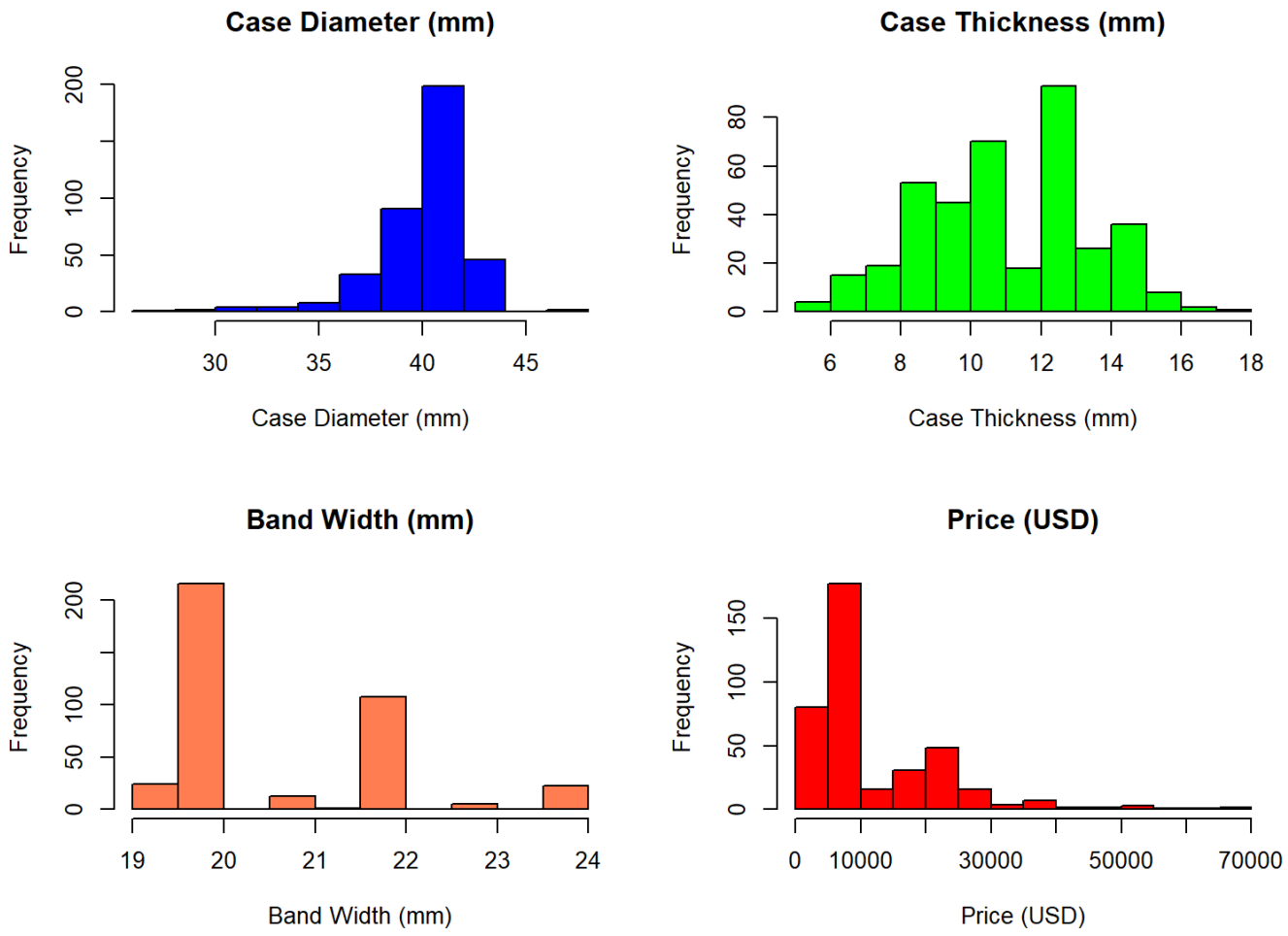
```
## numeric(0)
```

Thus,by capping we eliminated outliers in Band Width (mm).

Transformation

Now lets look for the distribution of our numeric variables.

```
# Set up the plotting area to have 2 rows and 2 columns
par(mfrow = c(2, 2))
# Plot each histogram
hist(watch2$`Case Diameter (mm)`, main = "Case Diameter (mm)", xlab = "Case Diameter (mm)",col = "blue")
hist(watch2$`Case Thickness (mm)`, main = "Case Thickness (mm)", xlab = "Case Thickness (mm)",col = "green")
hist(watch2$`Band Width (mm)`, main = "Band Width (mm)", xlab = "Band Width (mm)",col = "coral")
hist(watch2$`Price (USD)`, main = "Price (USD)", xlab = "Price (USD)",col = "red")
```

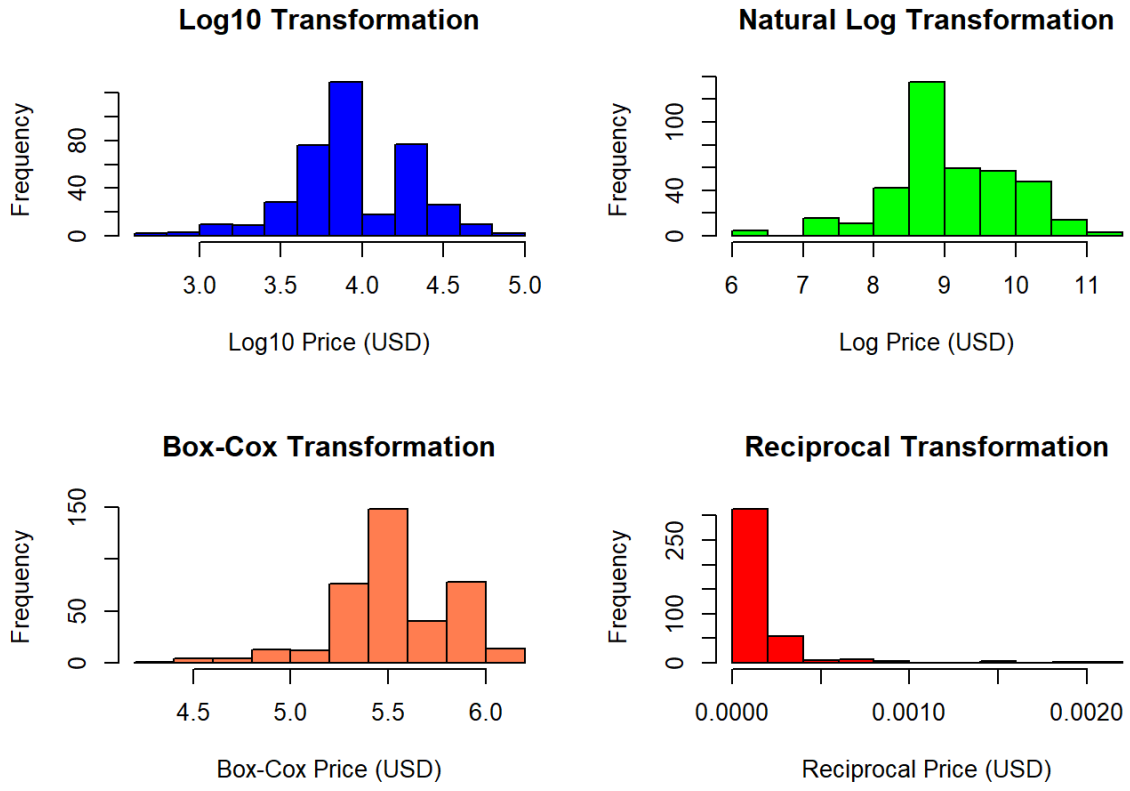


Transformation Cont.

As Price (USD) is our key variable and we could see skewness in the data, we will transform this using log base 10, natural log, box cox, and reciprocal methods.

As we have log10 transformation gave the best distribution we will do our analysis based on this price.Now we will start our statistical analysis in the coming sections.

```
# Set up the 2x2 plotting layout
par(mfrow = c(2, 2))
# Transformation using base 10 log
log_price <- log10(watch2$`Price (USD)`)
#Histogram of transformed data
hist(log_price,main = "Log10 Transformation", xlab = "Log10 Price (USD)",col = "blue")
# Transformation using base 10 log
log_price2 <- log(watch2$`Price (USD)`)
#Histogram of transformed data
hist(log_price2,main = "Natural Log Transformation", xlab = "Log Price (USD)",col = "green")
# Transformation using Box Cox
BoxCox_price<- BoxCox(watch2$`Price (USD)`,lambda = "auto")
#Histogram of transformed data
hist(BoxCox_price,main = "Box-Cox Transformation", xlab = "Box-Cox Price (USD)",col = "coral")
# Transformation using Reciprocal
reci_price <- 1/(watch2$`Price (USD)`)
#Histogram of transformed data
hist(reci_price, main = "Reciprocal Transformation", xlab = "Reciprocal Price (USD)",col = "red")
```



Hypothesis Testing and Confidence Interval

This analysis examines whether there's a significant price difference between watches with Leather and Stainless Steel straps using a two-sample independent t-test.

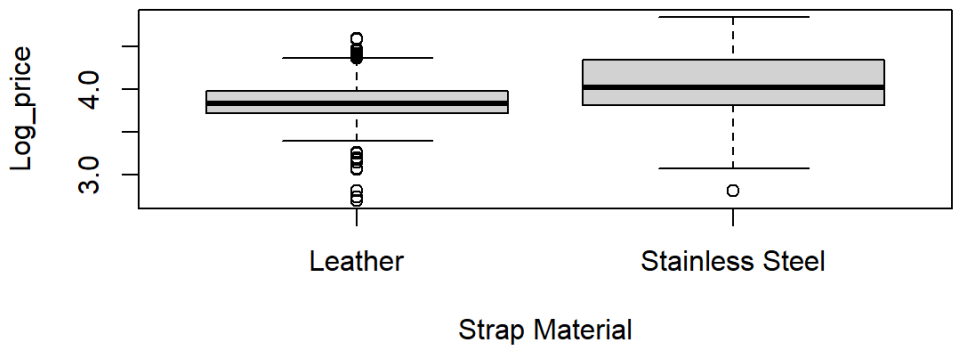
Null Hypothesis (H0): No significant difference in mean price between Leather and Stainless Steel straps.

Alternative Hypothesis (H1): Significant difference in mean price between Leather and Stainless Steel straps.

The t-test assumes independent populations, equal variance, and normal distribution, which must be verified.

```
# Creating a new column as the log10 of Price (USD)
watch2$log_price <- log10(watch2$`Price (USD)`)
```

```
# Box plot for Log Price and Strap Material
watch2 %>% boxplot(`log_price` ~ `Strap Material`, data = ., ylab = "Log_price")
```



While it's close, Stainless Steel watches appear to have higher prices. The two-sample t-test will determine if this difference is statistically significant.

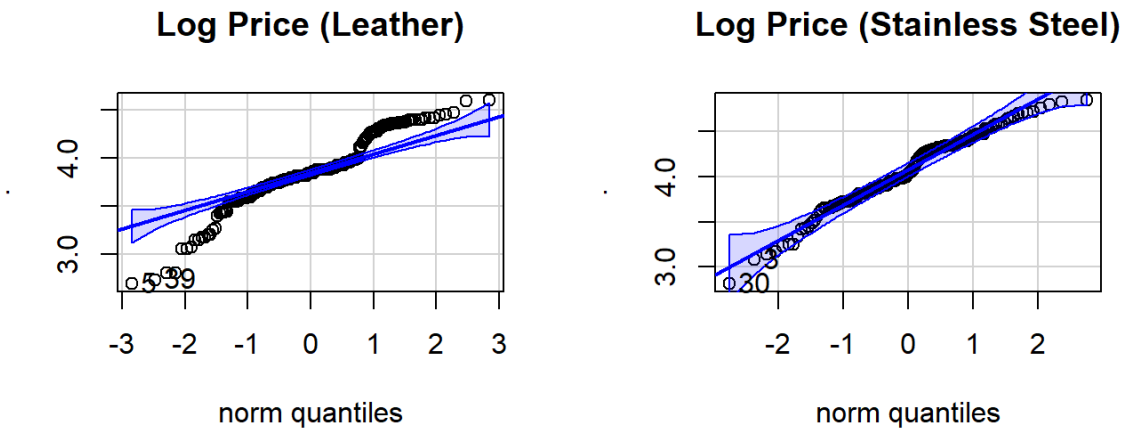
Testing the Assumption of Normality

We need to check each category of strap material, i.e, Leather and Stainless Steel follow normal distribution .

```
# Set up a 1x2 plotting layout to display both QQ plots side by side
par(mfrow = c(1, 2))
# Filtering the Leather data
log_price_leather <- watch2$log_price[watch2$`Strap Material` == "Leather"]
# Filtering the Stainless Steel data
log_price_steel <-watch2$log_price[watch2$`Strap Material` == "Stainless Steel"]
# Plot QQ Plot for Leather
log_price_leather %>% qqPlot(dist = "norm", main = "Log Price (Leather)")
```

```
## [1]  5 39
```

```
# Plot QQ Plot for Stainless Steel
log_price_steel %>% qqPlot(dist = "norm", main = "Log Price (Stainless Steel)")
```




```
## [1] 30  3
```

Points outside the distribution's tails indicate heavier-than-expected data, suggesting non-normality. However, due to the large sample size (n > 30), we can assume normality using the Central Limit Theorem. Therefore, we will proceed with the two-sample independent t-test.

Homogeneity of Variance

We need to verify the equality of variance before proceeding with our analysis. Homogeneity of variance is tested using Levene's test. The Levene's test has the following statistical hypotheses:

Null Hypothesis (H0): The variances are equal

Alternative Hypothesis (HA): The variances are not equal

This test ensures that the assumptions of equal variance are met, allowing us to proceed with more reliable statistical comparisons.

```
# Applying  Levene's test
leveneTest(log_price ~ `Strap Material`, data = watch2)
```

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	1	11.60658	0.0007259226
	388	*NA*	*NA*

2 rows

Levene's Test for Homogeneity of Variance shows a p-value of 0.0007259, rejecting the null hypothesis of equal variances ($p < 0.05$).

Welch's t-test

Since the variances are not equal, we will use Welch's t-test instead of the standard independent two-sample t-test.

```
# testing two independent samples using Welch's t-test with unequal variance and assume normality
t.test(log_price ~ `Strap Material`,data = watch2,var.equal = FALSE,alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  log_price by Strap Material
## t = -5.716, df = 326.67, p-value = 2.459e-08
## alternative hypothesis: true difference in means between group Leather and group Stainless Steel
is not equal to 0
## 95 percent confidence interval:
##  -0.2882736 -0.1406527
## sample estimates:
##         mean in group Leather mean in group Stainless Steel
##                      3.853550                      4.068013
```

The p-value strongly rejects the null hypothesis, indicating a significant difference in mean `log_price` between Leather and Stainless Steel strap watches (p = 2.459e-08). The t-value of -5.716 shows that Leather straps have a lower mean `log_price`. The 95% confidence interval [-0.288, -0.141] excludes 0, confirming the significant difference.

Welch's t-test Cont.

We can exponentiate the mean log_price(back transform) values to interpret the results in terms of the original price.

```
#store the log price man values in variables
mean_log_leather <- 3.853550
mean_log_steel <- 4.068013
#exponentiate the mean log_price (back trasformation)
mean_price_leather <- 10^mean_log_leather
mean_price_steel <- 10^mean_log_steel
#Print the results
mean_price_leather
```

```
## [1] 7137.564
```

```
mean_price_steel
```

```
## [1] 11695.34
```

Leather strap watches average $7,137.5, while Stainless Steel watches average $11,695.3

The significant p-value from Welch's t-test shows this price difference is statistically significant, with Stainless Steel watches priced higher than Leather.

Categorical association- Chi-Square Test

For the Chi-Square test, observations must be independent, and each cell in the contingency table should have an expected frequency of at least 5; if many cells fall below this threshold, the test may not be valid. Additionally, both variables should be categorical.

Null Hypothesis (H0): There is no association between strap material and movement type. **Alternative Hypothesis (H1)**: There is a significant association between strap material and movement type.

```
# Lets check if `Movement Type` is meeting the assumptions
watch2$`Movement Type` %>% table()
```

```
## .
## Automatic Eco-Drive    Manual    Quartz
##       327         1        50        12
```

We need to eliminate the Eco-Drive for meeting the assumptions of the test, hence we exclude that particular row.

```
#exclude the Eco_Drive row
watch3 <- watch2[watch2$`Movement Type` != "Eco-Drive", ]
# Drop unused factor levels
watch3$`Movement Type` <- droplevels(watch3$`Movement Type`)
```

We need to make a contingency table with our interest variables;

```
#contingency table- Strap Material vs Movement Type
chi_table <- table(watch3$`Strap Material`,watch3$`Movement Type`)
```

Now our data is ready for chi square test

```
#apply chisq.test()
chi_square_result <- chisq.test(chi_table)
print(chi_square_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  chi_table
## X-squared = 9.8979, df = 2, p-value = 0.007091
```

A p-value of 0.007 is very small, allowing us to reject the null hypothesis. This indicates a strong connection between strap type and movement type in luxury watches. In other words, the choice of strap (Leather or Stainless Steel) is related to the mechanism (Automatic or Quartz). This significant relationship suggests that certain straps are more likely paired with specific movement types.

Correlation

We will analyze the relationship between the log-transformed price of watches (log_price) versus three numeric variables: Case Diameter, Case Thickness, and Band Width (all in mm). This analysis will reveal how these physical characteristics influence watch prices. Lets check using Pearson Coefficient

```
# Select log_price and numeric variables
watch_data <- watch2[, c("Case Diameter (mm)", "Case Thickness (mm)", "Band Width (mm)", "log_price")]
# Calculate correlations with log_price
correlations <- cor(watch_data, use = "complete.obs")["log_price", ]
correlations <- correlations[names(correlations) != "log_price"]  # Remove the correlation of log_price with itself
# Print correlations
print(correlations)
```

```
##  Case Diameter (mm) Case Thickness (mm)     Band Width (mm)
##         -0.05050820         -0.41134770         -0.07505524
```

```
# Calculate confidence intervals for the correlations
ci_results <- list()
for (variable in names(correlations)) {
  r <- correlations[variable]
  n <- nrow(watch2)
  ci <- CIr(r, n)
  ci_results[[paste("log_price vs", variable)]] <- ci
}
# Print confidence intervals
print(ci_results)
```

```
## $`log_price vs Case Diameter (mm)`
## [1] -0.14906278  0.04903998
##
## $`log_price vs Case Thickness (mm)`
## [1] -0.4906096 -0.3253347
##
## $`log_price vs Band Width (mm)`
## [1] -0.17306756  0.02442905
```

The correlation between Case Diameter and Bandvwidth versus log-transformed price are a weak negative ones. For Case Thickness, the moderate negative correlation of -0.4113, with a confidence interval of [-0.4906, -0.3253], indicates that thicker cases are associated with lower log-transformed prices, suggesting a significant impact.

Conclusion

Findings

This study identified key factors affecting the price of luxury timepieces. The correlation analysis revealed varied relationships between price and specific features. Hypothesis testing demonstrated significant price differences based on strap materials, while Chi-square tests illuminated associations between movement types and strap materials.

Strengths

The strength of this research lies in its comprehensive statistical approach, combining descriptive analysis, hypothesis testing, and correlation studies to deliver a multi-faceted understanding of the luxury watch market.

Limitations

A limitation of this study is the potential for sample bias, given that the data may not be fully representative of the entire luxury watch market. Additionally, the analysis focused on only a few parameters, which may limit the scope of our findings.

Future Directions

Future research could expand on these findings by incorporating a larger and more diverse sample, exploring non-linear relationships, and examining the impact of brand reputation and market trends on watch prices. This expanded scope could yield deeper insights and more robust conclusions.

References

Rattanaporn K (n.d.) *Luxury Watches Price Dataset*, Kaggle website, accessed 19 October 2024. https://www.kaggle.com/datasets/rkiattisak/luxury-watches-price-dataset/data

MyApps Portal (2019a) Instructure.com, https://rmit.instructure.com/courses/124219/pages/week-7-learning-materials-slash-activities?module_item_id=6322124, accessed 23 October 2024.

Hayes A (2023) T-Test: What It Is With Multiple Formulas and When To Use Them, Investopedia, https://www.investopedia.com/terms/t/t-test.asp.

The Chi-Square Test (n.d.) www.jmp.com, https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test.html.

https://plus.google.com/u/0/+Datacamp (2011) Learn R, Python & Data Science Online, Datacamp.com, https://www.datacamp.com/.