ecommerce

Oppy

2024-12-31

The e-commerce dataset, sourced from Kaggle, provides a robust foundation for exploring various statistical and analytical techniques. It consists of 3,660 rows and 8 columns, capturing information such as product categories, prices, discounts, final prices, payment methods, and purchase dates. The dataset was initially loaded and pre-processed by selecting relevant columns and renaming them for clarity and usability. Preliminary analysis, including summary statistics, revealed that product prices range from $10.09 to $499.96, while discounts span from 0% to 50%, resulting in final prices ranging from $5.89 to $496.82. Importantly, a check for missing values showed no null entries, indicating data completeness. This dataset is particularly suitable for examining pricing dynamics, category-specific trends, and payment method preferences within the e-commerce domain, forming a basis for further statistical modelling and hypothesis testing.

```
tinytex::install_tinytex(force = TRUE)
```

```
## tlmgr install tlgpg

## tlmgr update --self

## tlmgr install tlgpg

## tlmgr --repository http://www.preining.info/tlgpg/ install tlgpg

## tlmgr option repository "https://au.mirrors.cicku.me/ctan/systems/texlive/tlnet"

## tlmgr update --list
```

```
tinytex::tlmgr_update()
```

load tidyverse

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("tidyverse")
```

```
## Installing package into 'C:/Users/fadar/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\fadar\AppData\Local\Temp\RtmpqIe516\downloaded_packages
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2
```

```
## -- 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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

import data set

```
ecommerce <- read.csv("ecommerce_dataset_updated.csv")
```

Inspect the data

```
View(ecommerce)
```

```
head(ecommerce)
```

```
##    User_ID Product_ID Category Price..Rs.. Discount.... Final_Price.Rs..
## 1 337c166f f414122f-e   Sports       36.53           15            31.05
## 2 d38a19bf fde50f9c-5 Clothing      232.79           20           186.23
## 3 d7f5f0b0 0d96fc90-3   Sports      317.02           25           237.76
## 4 395d4994 964fc44b-d     Toys      173.19           25           129.89
## 5 a83c145c d70e2fc6-e   Beauty      244.80           20           195.84
## 6 3fdcdae8 0816ee12-5    Books      241.86           50           120.93
##   Payment_Method Purchase_Date
## 1    Net Banking    12-11-2024
## 2    Net Banking    09-02-2024
## 3    Credit Card    01-09-2024
## 4            UPI    01-04-2024
## 5    Net Banking    27-09-2024
## 6            UPI    08-08-2024
```

```
colnames(ecommerce)
```

```
## [1] "User_ID"          "Product_ID"       "Category"         "Price..Rs.."
## [5] "Discount...."     "Final_Price.Rs.." "Payment_Method"   "Purchase_Date"
```

```
dim(ecommerce)
```

```
## [1] 3660    8
```

select relevant columns

```
ecommerce_2 <- subset(ecommerce, select = c("Product_ID", "Category", "Price..Rs..",
                                            "Discount....","Final_Price.Rs..",
                                            "Payment_Method","Purchase_Date"))
```

rename columns

```
ecommerce_3 <- (rename(ecommerce_2, "product_id" = "Product_ID",
                     "category" = "Category",
                     "price" = "Price..Rs..",
                      "discount" = "Discount....",
                     "final_price" = "Final_Price.Rs..",
                      "method" = "Payment_Method",
                     "date" = "Purchase_Date"))
```

```
dim(ecommerce_2)
```

```
## [1] 3660    7
```

#check for null values

```r
sum(is.na(ecommerce_2))
```

```
## [1] 0
```

```r
View(ecommerce_3)
```

measure of central tendency

mean

```r
mean(ecommerce_3$price)
```

```
## [1] 254.8007
```

```r
ecommerce_3 %>%
  select("price", "discount", "final_price") %>%
           summary()
```

```
##      price           discount      final_price
##  Min.   : 10.09   Min.   : 0.00   Min.   :  5.89
##  1st Qu.:134.01   1st Qu.: 5.00   1st Qu.:104.51
##  Median :253.84   Median :15.00   Median :199.19
##  Mean   :254.80   Mean   :18.83   Mean   :206.91
##  3rd Qu.:377.60   3rd Qu.:25.00   3rd Qu.:304.12
##  Max.   :499.96   Max.   :50.00   Max.   :496.82
```

mode

```r
mode("ecommerce_3$price")
```

```
## [1] "character"
```

```r
mode("ecommerce_3$final_price")
```

```
## [1] "character"
```

```r
ecommerce_3 %>%
  select("price", "final_price")%>%
  mode()
```

```
## [1] "list"
```

```r
get_mode <- function(x) {
  uniq_x <- unique(x)
  freq_x <- tabulate(match(x, uniq_x))
  modes <- uniq_x[freq_x == max(freq_x)]
  return(modes)
}
ecommerce_3 %>%
  select(price, final_price)%>%
  summarise(across(everything(), get_mode))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
##    price final_price
## 1 286.09       259.1
## 2 410.07       259.1
## 3 335.09       259.1
## 4 185.53       259.1
```

measure of central tendency | range variance SD IQR

range

```r
range(ecommerce_3$price)
```

```
## [1]  10.09 499.96
```

```r
range(ecommerce_3$final_price)
```

```
## [1]   5.89 496.82
```

```r
ecommerce_3 %>%
  select("price") %>%
  var()
```

```
##          price
## price 20073.97
```

```r
ecommerce_3 %>%
  select("final_price") %>%
  var()
```

```
##             final_price
## final_price    15052.31
```

```
sd <- sd(ecommerce_3$price)
print(sd)
```

```
## [1] 141.6826
```

measure of shapes | skewness & kurtosis

```
install.packages("e1071")
```

```
## Installing package into 'C:/Users/fadar/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'e1071' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\fadar\AppData\Local\Temp\RtmpqIe516\downloaded_packages
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.4.2
```

```
kurtosis(ecommerce_3$price, type = 3)
```

```
## [1] -1.195883
```

```
kurtosis(ecommerce_3$final_price, type = 3)
```

```
## [1] -0.9476906
```

```
kurtosis(ecommerce_3$discount, type = 3)
```

```
## [1] -0.1329752
```

```
skewness(ecommerce_3$price, type = 3)
```

```
## [1] -0.005346183
```

```
skewness(ecommerce_3$final_price, type = 3)
```

```
## [1] 0.2374094
```

```r
skewness(ecommerce_3$discount, type = 3)
```

```
## [1] 0.7654854
```

The kurtosis and skewness values provide insights into the distribution shapes of the price, final price, and discount variables. The kurtosis values for price (-1.20), final price (-0.95), and discount (-0.13) indicate platykurtic distributions, where the data have lighter tails than a normal distribution. The skewness values for price (-0.005) and final price (0.24) suggest near-symmetric distributions, while the discount (0.77) shows a moderate positive skew, indicating a longer tail to the right. These measures highlight that while price and final price are relatively symmetric, discounts tend to cluster towards lower values with a few higher outliers.

method

```r
methods_group <- ecommerce_3 %>%
            group_by(method)%>%
            summarise(final_price = mean(final_price))%>%
            arrange(final_price)
View(methods_group)
```

category

```r
category_group <- ecommerce_3 %>%
            group_by(category) %>%
            summarise(final_price = mean(final_price))%>%
            arrange(final_price)
View(category_group)
```

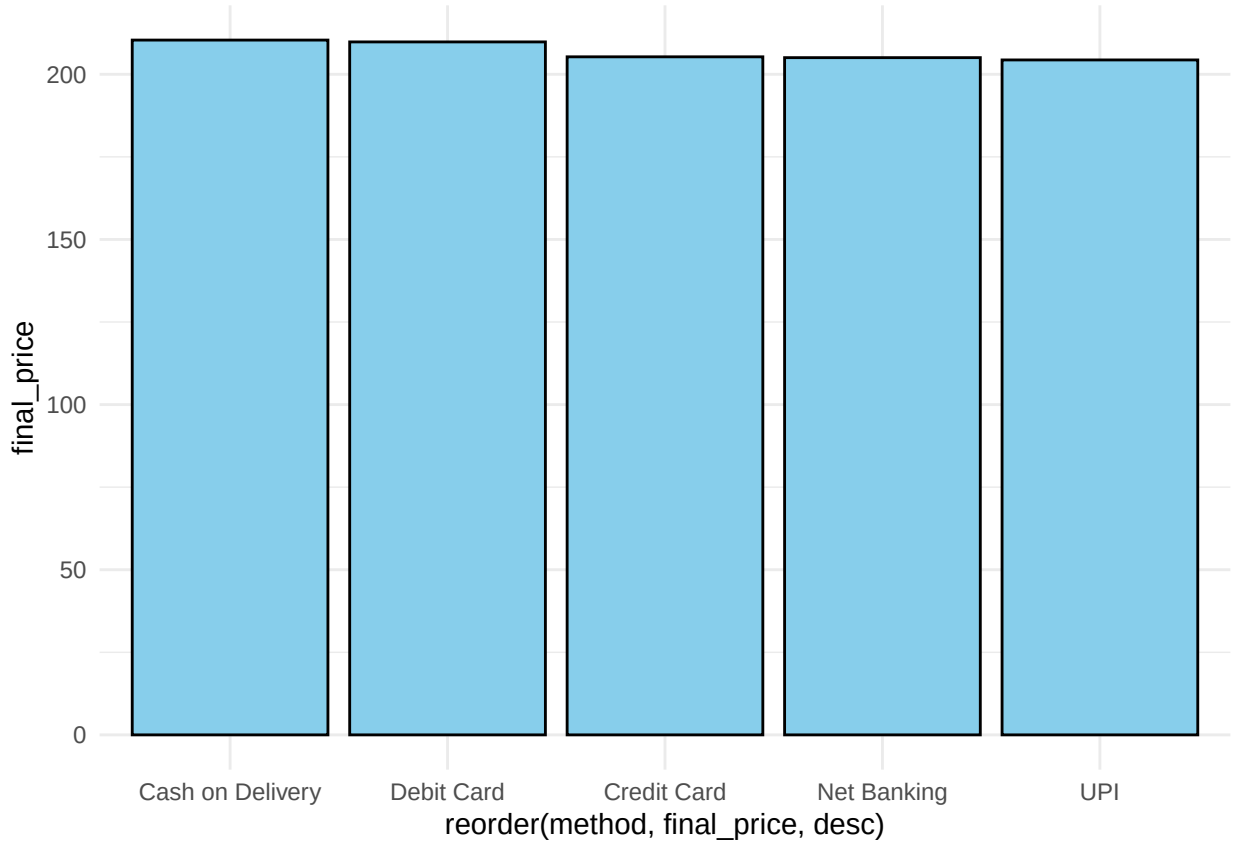
graphical representation

bar chat histogram |scatter plot | box plot | QQ plot

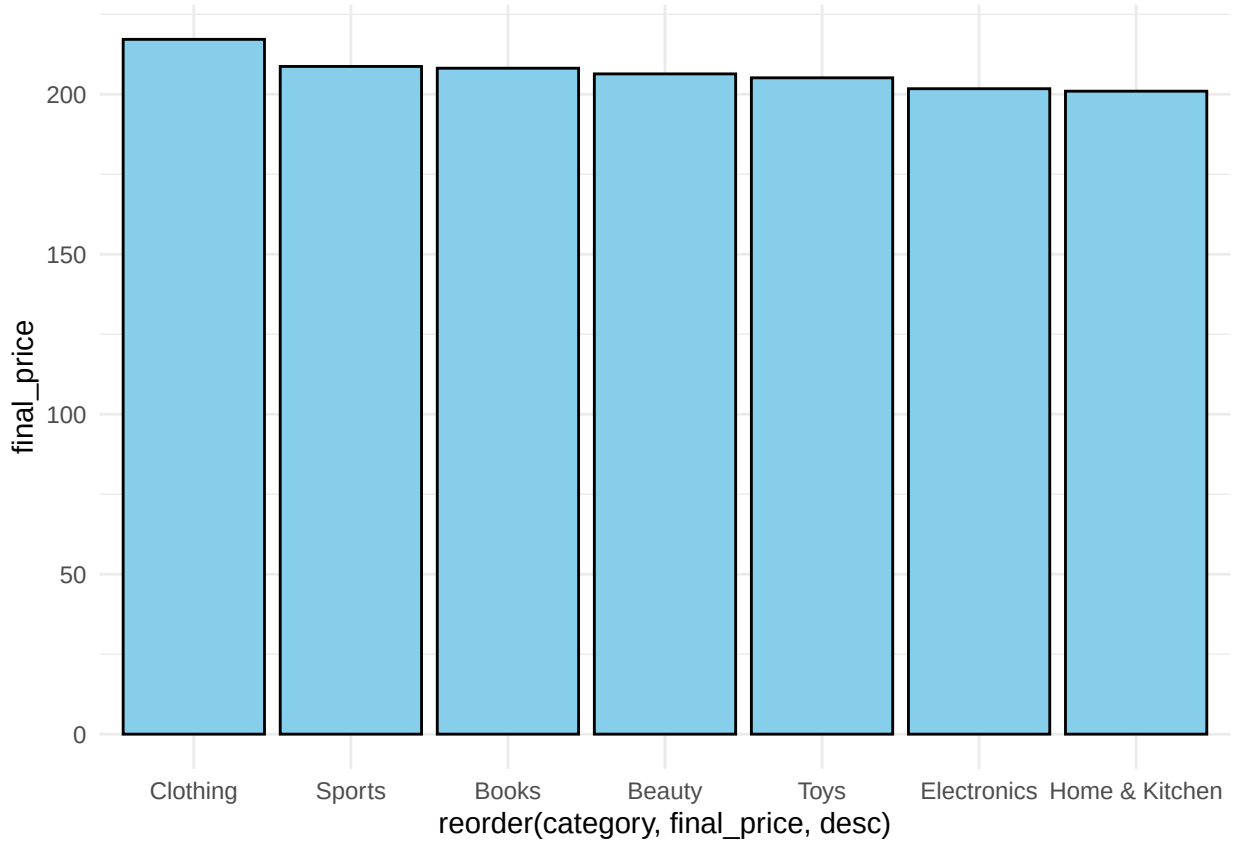
```r
unique(ecommerce_3$category)
```

```
## [1] "Sports"         "Clothing"       "Toys"           "Beauty"
## [5] "Books"          "Home & Kitchen" "Electronics"
```

```
ggplot(data = methods_group,
         mapping = aes (x = reorder(method, final_price, desc), y = final_price)) +
          geom_bar(stat = "identity", fill = "skyblue", color = "black")+
          theme_minimal()
```

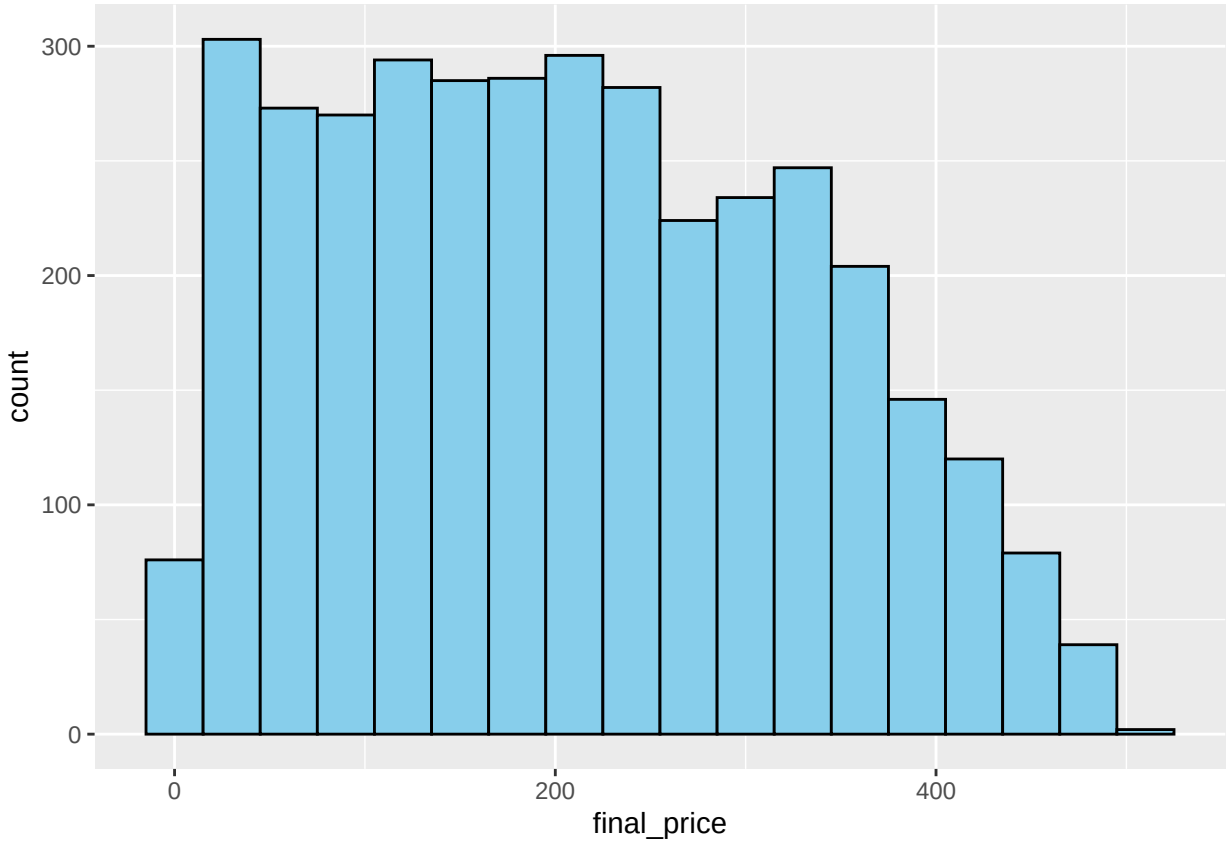



```
ggplot(data = category_group,
       mapping = aes(x=reorder(category,final_price, desc), y=final_price))+
  geom_bar(stat = "identity", fill = "skyblue", color = "black")+
  theme_minimal()
```

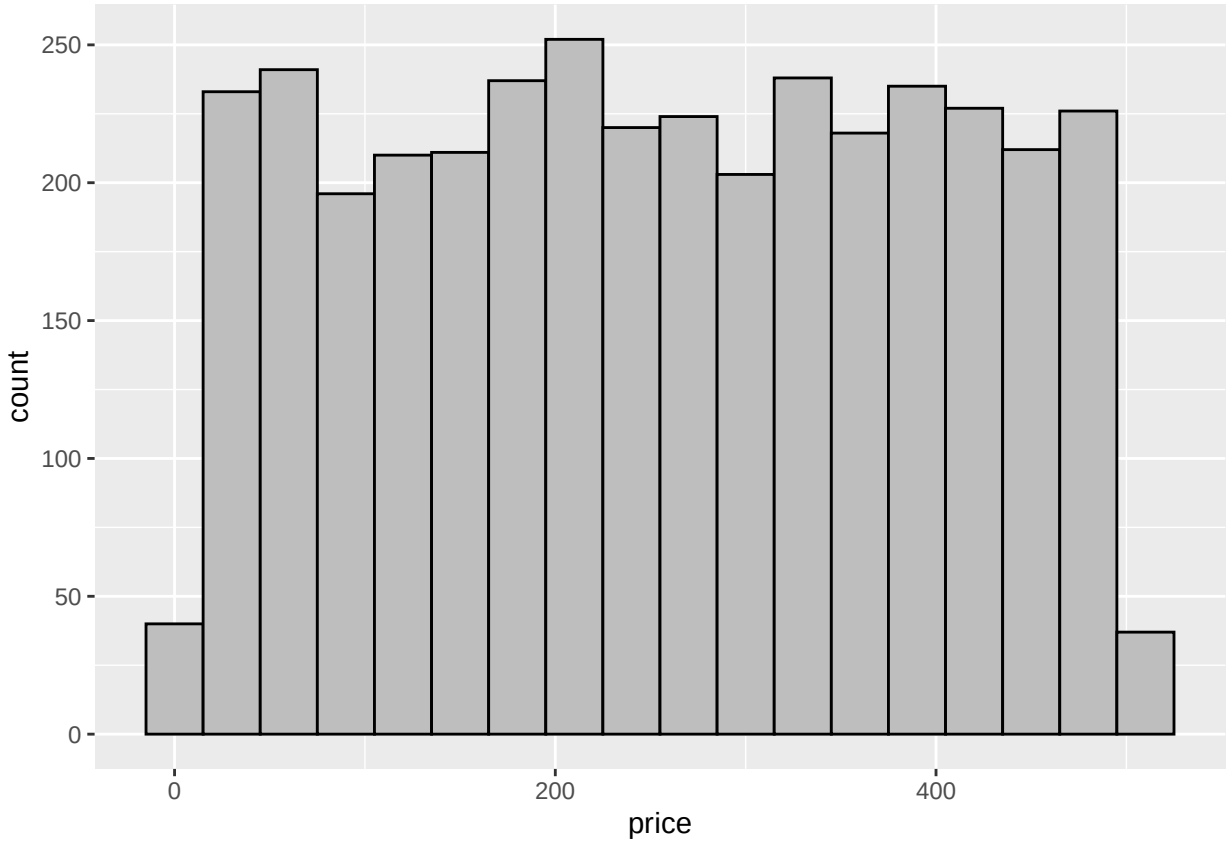


histogram

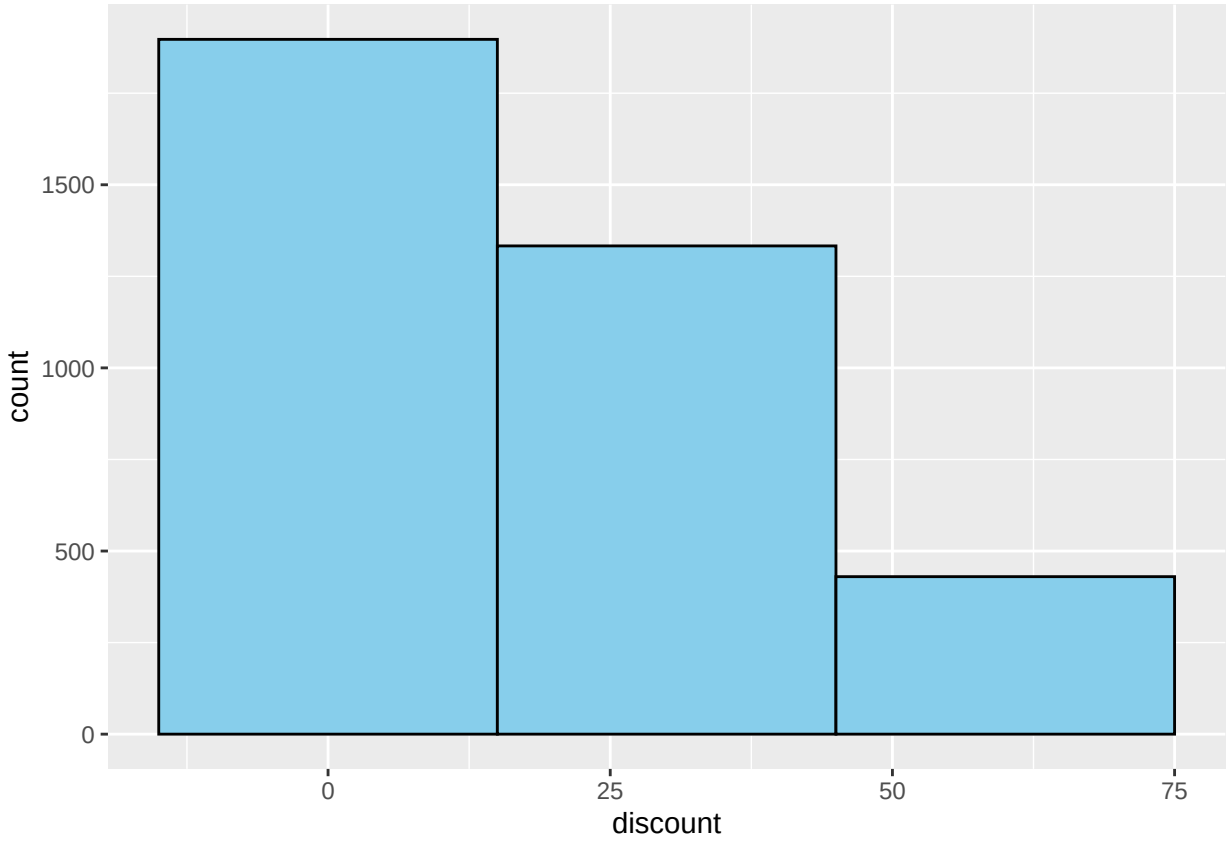
```
ggplot(data = ecommerce_3,
       mapping = aes(x = final_price)) +
  geom_histogram(binwidth = 30, fill = "skyblue", color = "black")
```




```
ggplot(data = ecommerce_3,
       mapping = aes(x=price))+
      geom_histogram(binwidth = 30, fill = "grey", color = "black")
```

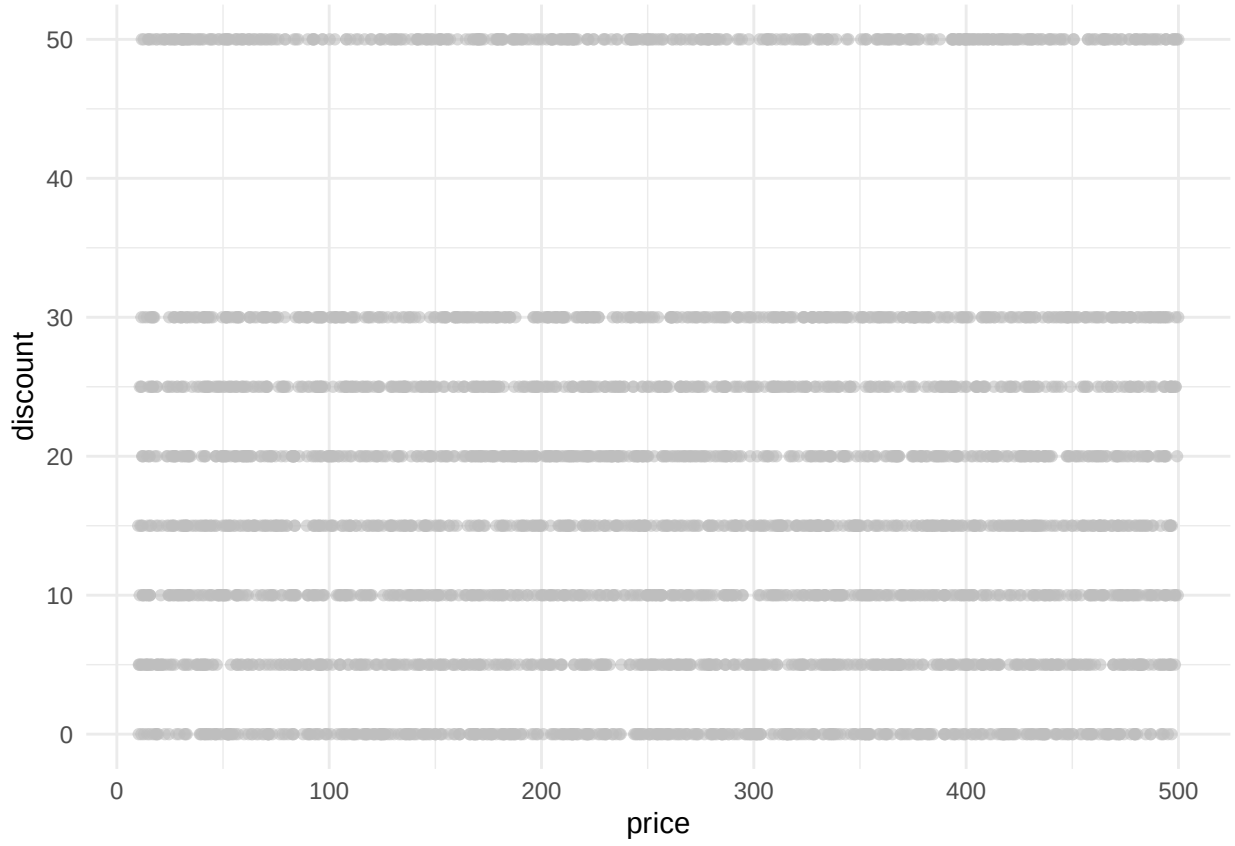



```
ggplot( data = ecommerce_3,
        mapping =  aes(x=discount))+
        geom_histogram(binwidth =  30, fill = "skyblue", color = "black")
```

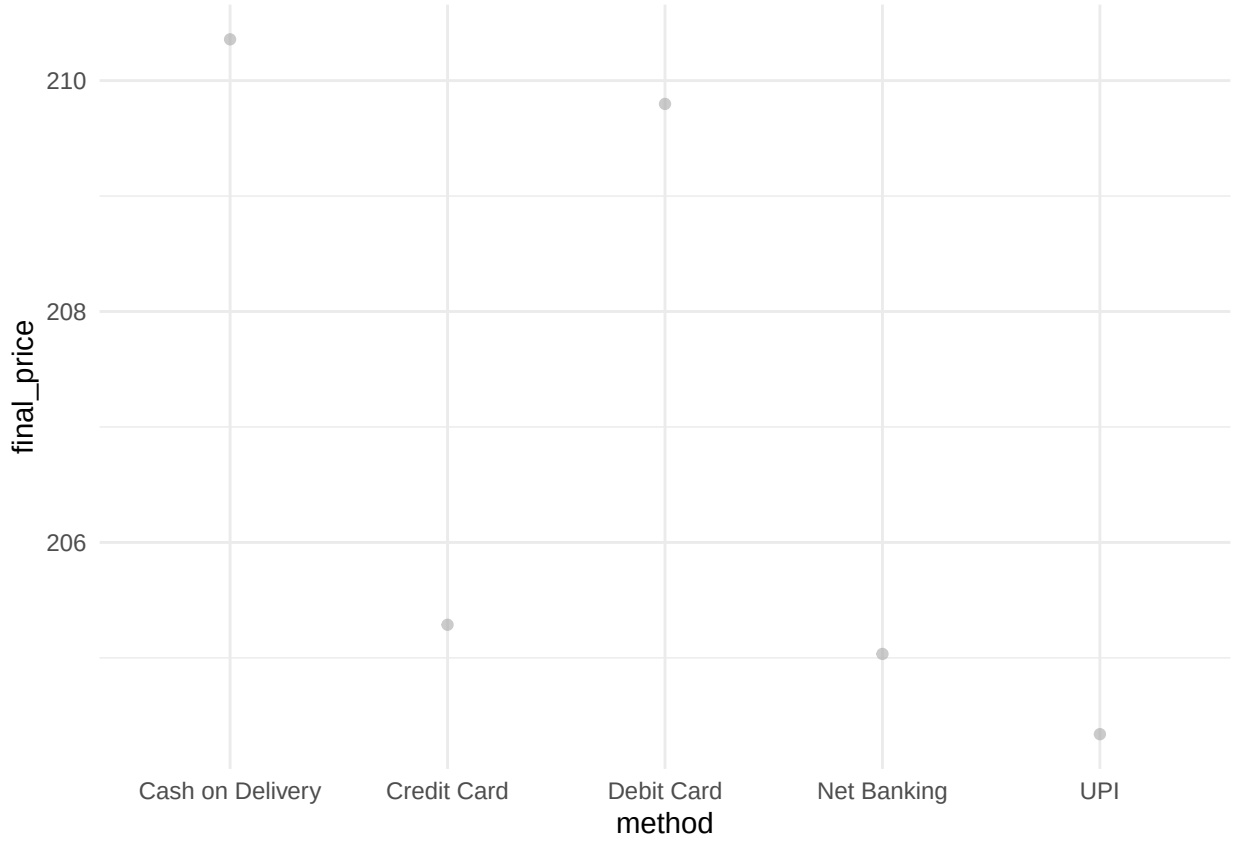


scatterplot

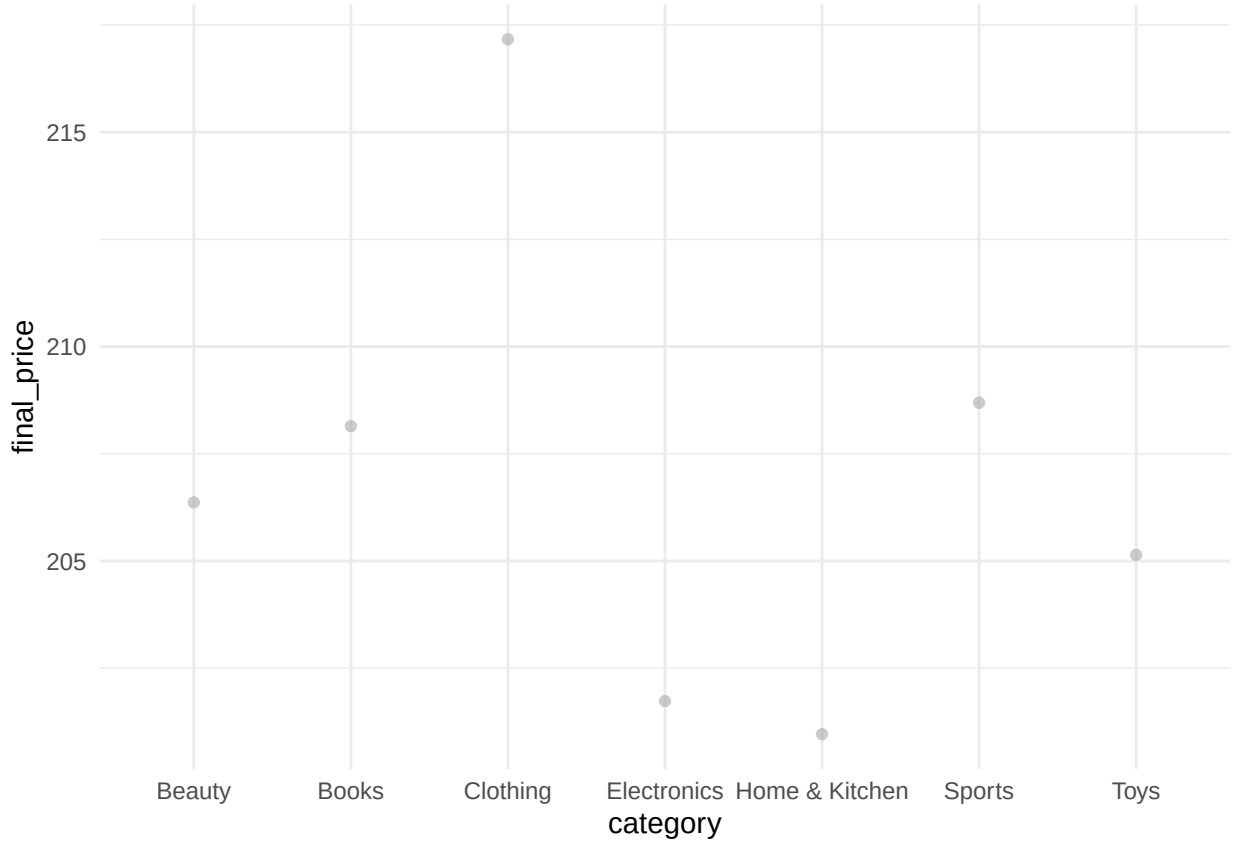
```
ggplot(data = ecommerce_3,
       mapping = aes(price, discount))+
      geom_point(color = "grey", alpha = 0.6)+
      theme_minimal()
```




```
ggplot(data = methods_group,
       mapping = aes(method, final_price))+
  geom_point(color = "darkgrey", alpha = 0.6)+
  theme_minimal()
```




```
ggplot(data = category_group,
       mapping = aes(category, final_price))+
  geom_point(color = "darkgrey", alpha = 0.6)+
  theme_minimal()
```

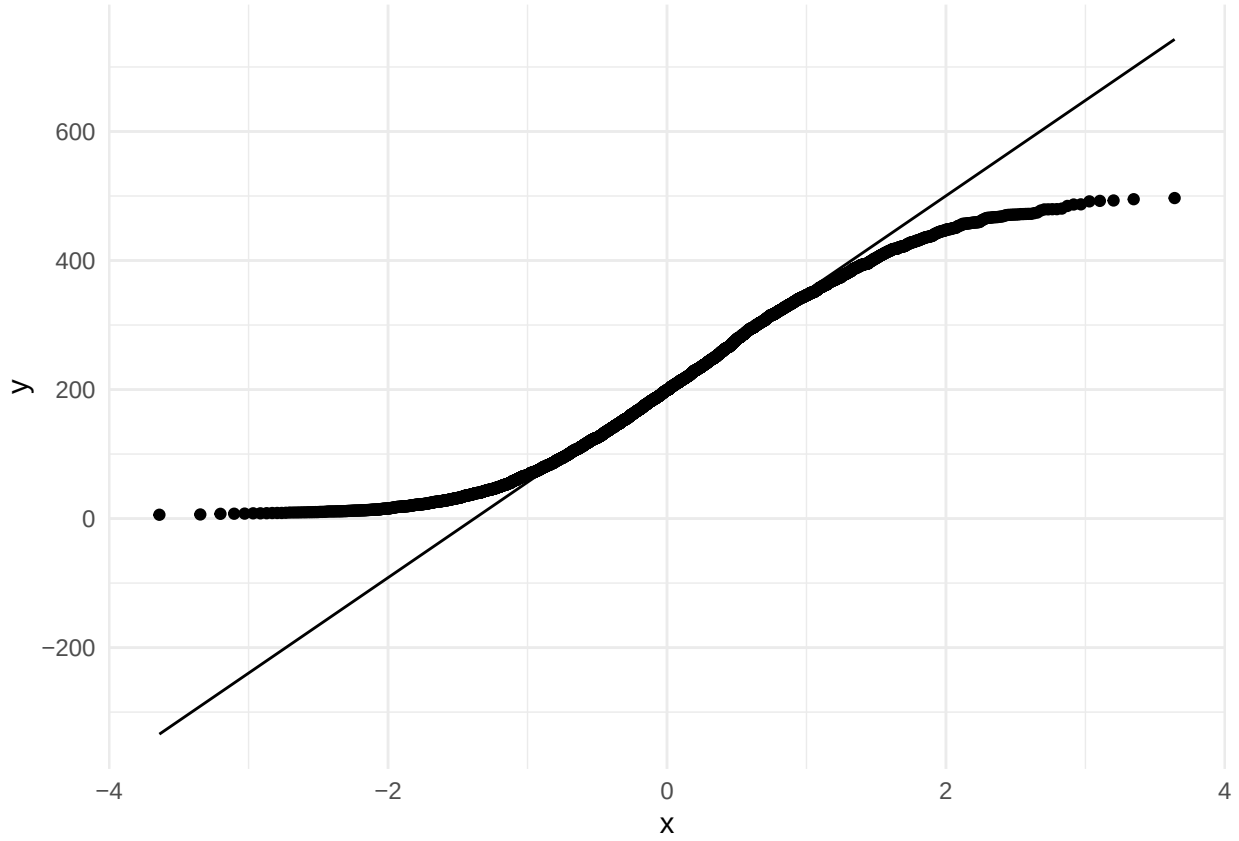


box_plot

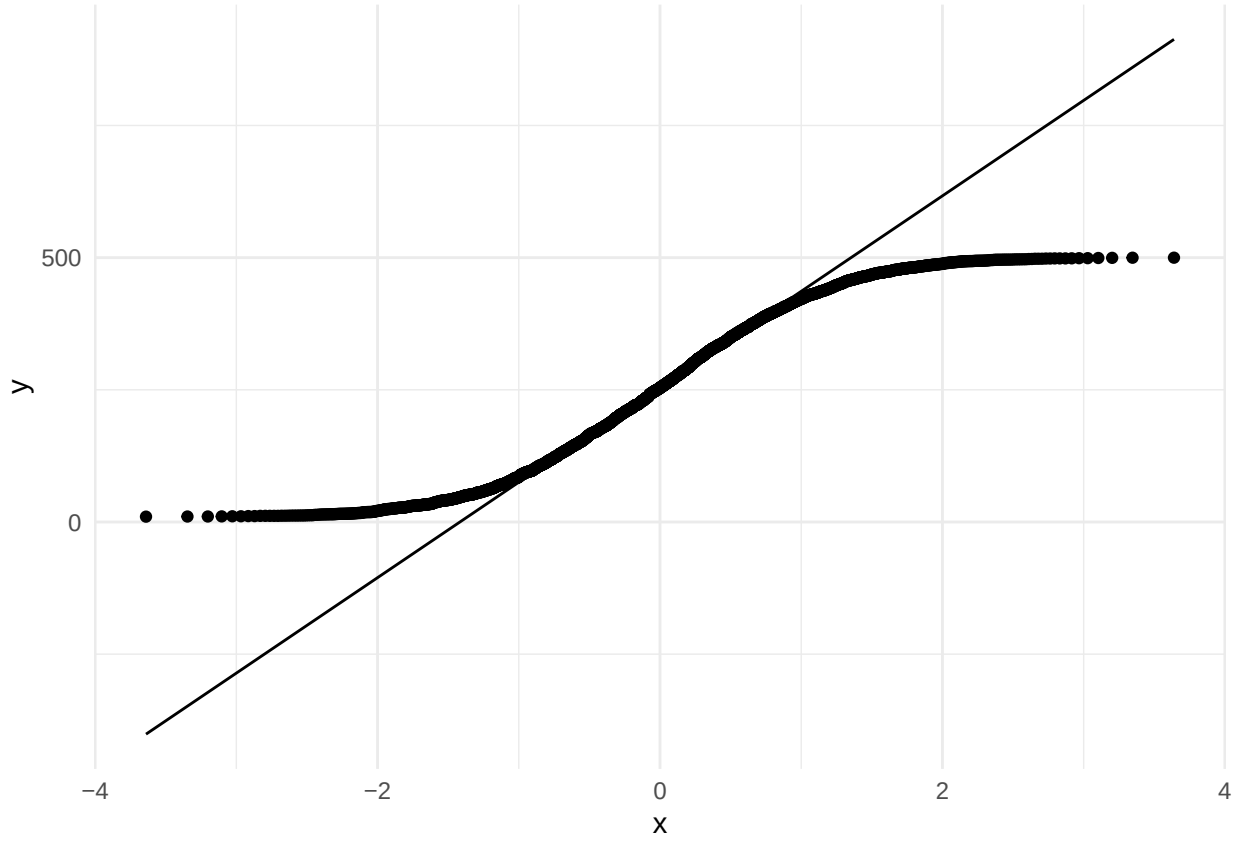
```
ggplot(data = ecommerce_3,
       mapping = aes(method, final_price))+
      geom_boxplot()+
      theme_minimal()
```

qqplot

```
ggplot(data=ecommerce_3,
       mapping = aes(sample = final_price))+
      geom_qq()+
      stat_qq_line()+
      theme_minimal()
```




```
ggplot(data=ecommerce_3,
       mapping = aes(sample = price))+
  geom_qq()+
  stat_qq_line()+
  theme_minimal()
```



Date manipulations

```
glimpse(ecommerce_3)
```

```
## Rows: 3,660
## Columns: 7
## $ product_id  <chr> "f414122f-e", "fde50f9c-5", "0d96fc90-3", "964fc44b-d", "d~
## $ category    <chr> "Sports", "Clothing", "Sports", "Toys", "Beauty", "Books",~
## $ price       <dbl> 36.53, 232.79, 317.02, 173.19, 244.80, 241.86, 76.91, 213.~
## $ discount    <int> 15, 20, 25, 25, 20, 50, 5, 20, 5, 50, 10, 15, 10, 50, 25, ~
## $ final_price <dbl> 31.05, 186.23, 237.76, 129.89, 195.84, 120.93, 73.06, 170.~
## $ method      <chr> "Net Banking", "Net Banking", "Credit Card", "UPI", "Net B~
## $ date        <chr> "12-11-2024", "09-02-2024", "01-09-2024", "01-04-2024", "2~
```

```
ecommerce_3$date <- as.Date(ecommerce_3$date, format = "%d-%m-%Y")
```

```
ecommerce_4 <- ecommerce_3 %>% # split the date vale into day, month, year
  mutate(
    weekday = wday(date, label = TRUE),
    month = month(date, label = TRUE),
    year = year(date)
  )
```

```r
unique(ecommerce_4$month)
```

```
##  [1] Nov Feb Sep Apr Aug Mar May Jan Oct Jun Jul
## 12 Levels: Jan < Feb < Mar < Apr < May < Jun < Jul < Aug < Sep < ... < Dec
```

```r
unique(ecommerce_4$weekday)
```

```
## [1] Tue Fri Sun Mon Thu Wed Sat
## Levels: Sun < Mon < Tue < Wed < Thu < Fri < Sat
```

```r
unique(ecommerce_4$year)
```

```
## [1] 2024
```

```r
clothing <- ecommerce_4 %>%
  filter(category == "Clothing") %>%
  select(category, price, final_price)
```

```r
electronics <- ecommerce_4 %>%
  filter(category == "Electronics")
```

t-tests | one sample t-test

objective: how does the mean price of clothing compare to the market standard.

"For the purpose of this analysis, it is assumed that the prices in the dataset are measured in Australian Dollars (AUD), even though the dataset does not specify the currency."

2023 average annual online cloth shopping fee, Australia $151.00 AUD (www.statista.com)

H (Null Hypothesis): There is no significant difference between the sample mean (mean of your dataset) and the population mean (market standard).

H (Alternative Hypothesis): There is a significant difference between the sample mean (mean of your dataset) and the population mean (market standard).

```
t.test(clothing$price, mu = 151)
```

```
##
##  One Sample t-test
##
## data:  clothing$price
## t = 17.976, df = 530, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 151
## 95 percent confidence interval:
##  250.8637 275.3681
## sample estimates:
## mean of x
##  263.1159
```

The original dataset does not specify the currency for the prices. For the purpose of this analysis, it is assumed that the prices are measured in Australian Dollars (AUD). The one-sample t-test compares the mean price of clothing in the dataset (AUD 263.12) to the market standard of AUD 151. With a t-value of 17.976 and a p-value < 2.2e-16 (significantly less than 0.05), we reject the null hypothesis, indicating a statistically significant difference between the sample mean and the market standard. The 95% confidence interval (250.86, 275.37) further supports that the true mean price of clothing in the dataset is notably higher than AUD 151.

t-test | independent (two) sample t-test

Ho: There is no significant difference between the means of both groups

Ha: There is a significant difference between the means of both groups

```
two_sample <- ecommerce_4 %>%
                filter(category %in% c("Clothing", "Electronics"))
View(two_sample)
```

```
t.test(price ~ category, data = two_sample, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  price by category
## t = 1.309, df = 1027, p-value = 0.1908
## alternative hypothesis: true difference in means between group Clothing and group Electronics is not
## 95 percent confidence interval:
##  -5.723482 28.661574
## sample estimates:
##    mean in group Clothing mean in group Electronics
##                  263.1159                  251.6469
```

```
t.test(price ~ category, data = two_sample, var.equal = FALSE) # going with this as I'm unsure.
```

```
##
##  Welch Two Sample t-test
##
## data:  price by category
```

```
## t = 1.3111, df = 1026.8, p-value = 0.1901
## alternative hypothesis: true difference in means between group Clothing and group Electronics is not
## 95 percent confidence interval:
##  -5.696628 28.634720
## sample estimates:
##    mean in group Clothing mean in group Electronics
##                  263.1159                  251.6469
```

The Welch Two Sample t-test was conducted to compare the mean prices of Clothing and Electronics categories. The test yielded a t-value of 1.3111, with a p-value of 0.1901, indicating no significant difference between the mean prices of the two categories ($p > 0.05$). The 95% confidence interval for the mean difference ranged from -5.70 to 28.63, further supporting the lack of evidence for a significant difference. The assumption of equal variances (var.equal = TRUE) was not used because the homogeneity of variances was uncertain; Welch's t-test is more robust when variances between groups are unequal.

t-test | paired t-test (before and after a treatment) in this case discount

Ho: There is no significant difference between the means before and after the treatment (discount)

Ha: There is a significant difference between the means before and after the treatment (discount)

```
clothing_paired <- t.test(clothing$price, clothing$final_price, paired = TRUE)
clothing_paired
```

```
##
##  Paired t-test
##
## data:  clothing$price and clothing$final_price
## t = 22.855, df = 530, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  42.00098 49.90000
## sample estimates:
## mean difference
##        45.95049
```

A paired t-test was conducted to compare the mean price of clothing items before (original price) and after applying a discount (final price). The test resulted in a t-value of 22.855 with a p-value of less than 2.2e-16, indicating a highly significant difference between the two means ($p < 0.05$). The mean difference between the original price and the final price was 45.95 AUD, with a 95% confidence interval ranging from 42.00 AUD to 49.90 AUD. This suggests that, on average, discounts significantly reduced the price of clothing items in the dataset.

```
mean(clothing$price) - mean(clothing$final_price)
```

```
## [1] 45.95049
```

ANOVA

One-Way ANOVA (To check if there is a statistical difference between the means of the different levels)

Ho: there is no significant difference between the means of the different levels of category

Ha there is a significant difference in at least one mean of the levals in the category

```
one_way_anova_model <- lm(ecommerce_4$final_price ~ ecommerce_4$category, data = ecommerce_4)
anova_output <- anova(one_way_anova_model)
print(anova_output)
```

```
## Analysis of Variance Table
##
## Response: ecommerce_4$final_price
##                          Df   Sum Sq Mean Sq F value Pr(>F)
## ecommerce_4$category      6    92870   15478  1.0284 0.4046
## Residuals              3653 54983521   15052
```

A one-way ANOVA was conducted to determine whether there is a statistically significant difference in the mean final price across the different levels of the "category" variable. The analysis yielded an F-value of 1.0284 with a p-value of 0.4046. Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis (H). This indicates that there is no significant difference in the mean final price among the different product categories in the dataset.

Two-Way ANOVA (To check if there is a significant difference in the mean of the different levels of the two categories or

any interaction between the two categories on the mean final_price)

Main Effect: (category)

Ho: there is no significant difference in the mean of final_price across the different levels of category

Ha: there is a significant difference in the mean of final_price in at least on level of the category variable.

Main Effect: (payment method)

Ho: there is no significant difference in the mean of final_price across the different levels of payment method

Ha: there is a significant difference in the mean of final_price in at least on level of the payment metho variable.

Interaction effect (category and payment method on final_price)

Ho: there is no interaction effect between category and payment method on the mean final_price

Ha: there is interaction effect between category and payment method on the mean final_price

```r
two_way_anova_model <- lm (ecommerce_4$final_price ~ category + method + category:method, data = ecomme
two_way_anove_output <- anova(two_way_anova_model)
print(two_way_anove_output)
```

```
## Analysis of Variance Table
##
## Response: ecommerce_4$final_price
##                    Df   Sum Sq Mean Sq F value  Pr(>F)
## category            6    92870 15478.4  1.0303 0.40336
## method              4    21461  5365.3  0.3571 0.83921
## category:method    24   501017 20875.7  1.3895 0.09797 .
## Residuals        3625 54461042 15023.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the Two-Way ANOVA show that neither the category (p = 0.40336) nor the payment method (p = 0.83921) have a significant main effect on the mean final_price. Additionally, there is no significant interaction effect between category and payment method (p = 0.09797), although it is marginally close to being significant at the 10% level. Therefore, based on the p-values, we fail to reject the null hypotheses (H) for all three effects: there is no significant difference in final_price across the levels of category, payment method, or their interaction.

repeted measure ANOVA is not applicable in the dataset and analsys

correlation

?cor cor(ecommerce_4$price, ecommerce_4$final_price, method = "pearson")

Ho: there is no significant linear correlation between price & final_price.

Ha: there is a significant linear correlation between price & final_price.

```r
cor_matrix_price <- ecommerce_4 %>%
            select(price, final_price)%>%
            cor()
print(cor_matrix_price)
```

```
##                price final_price
## price       1.0000000   0.9356911
## final_price 0.9356911   1.0000000
```

```
cor.test(ecommerce_4$price, ecommerce_4$final_price, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ecommerce_4$price and ecommerce_4$final_price
## t = 160.4, df = 3658, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9315319 0.9396056
## sample estimates:
##       cor
## 0.9356911
```

There is a strong positive linear correlation (r=0.9357) between price and final_price. The relationship is statistically significant (p-value < 0.05). As price increases, final_price also increases

Ho: there is no significant linear correlation between discount & final_price.

Ha: there is a significant linear correlation between discount & final_price.

```
cor_matrix_discount <- ecommerce_4 %>%
            select(discount, final_price)%>%
            cor()
print(cor_matrix_discount)
```

```
##               discount final_price
## discount     1.0000000  -0.3115149
## final_price -0.3115149   1.0000000
```

```
cor.test(ecommerce$Discount, ecommerce_4$final_price, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ecommerce$Discount and ecommerce_4$final_price
## t = -19.827, df = 3658, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.3404777 -0.2819616
## sample estimates:
##        cor
## -0.3115149
```

There is a weak but statistically significant negative correlation (= −0.3115) etween Discount and final_price. This means as the discount increases, the final price tends to decrease. Given the p-value is much smaller than 0.05, we reject the null hypothesis and conclude that there is evidence of a significant linear correlation between the two variables.

regression

?lm

simple leaner regression | dependent variable = final_price | independent variable = discount

Ho: there is no statistically significant effect of independent variable discount on the dependent variable final_price

Ho: there is a statistically significant effect of independent variable discount on the dependent variable final_price

```
model_SLR <- lm(final_price ~ discount, data = ecommerce_4)
summary(model_SLR)
```

```
##
## Call:
## lm(formula = final_price ~ discount, data = ecommerce_4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -245.417  -96.249   -0.926   95.705  241.073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 255.7467     3.1277   81.77   <2e-16 ***
## discount     -2.5944     0.1308  -19.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 116.6 on 3658 degrees of freedom
## Multiple R-squared:  0.09704,    Adjusted R-squared:  0.09679
## F-statistic: 393.1 on 1 and 3658 DF,  p-value: < 2.2e-16
```

scatter plot of final_price & discount

ggplot(data = ecommerce_4, # main graph mapping = aes(x=discount, y=final_price))+ geom_smooth(method = "lm", col = "blue")

ggplot(data = ecommerce_4, mapping = aes(x=discount, y=final_price))+ geom_point(shape = 3)

ggplot(data = ecommerce_4, mapping = aes(x = discount, y = final_price)) + geom_point(shape = 3) + geom_smooth(method = "lm", col = "blue")

The linear regression model examining the relationship between `final_price` and `discount` reveals that `discount` has a statistically significant negative effect on `final_price`, with a coefficient of -2.5944 (p-value < 2.2e-16), indicating that for each unit increase in `discount`, `final_price` decreases by approximately 2.5944 units. The intercept is 255.7467, representing the expected `final_price` when `discount` is zero. The residuals range widely, suggesting variability in the data, and the multiple R-squared value is 0.09704, indicating that around 9.7% of the variability in `final_price` can be explained by `discount`. The residual standard error is 116.6, which measures the typical size of the residuals, and the high F-statistic (393.1) with a p-value < 2.2e-16 confirms the overall model's statistical significance.

Multiple linear regression model | dependent variable = final_price | independent variable = rice, discount

Ho: There is no statistically significant effect of the independent variables discount and price on the dependent variable final_price.

Ha: There is a statistically significant effect of at least one of the independent variables discount and price on the dependent variable final_price

```
model_MLR <- lm(final_price ~ price + discount, data = ecommerce_4)
summary(model_MLR)
```

```
##
## Call:
## lm(formula = final_price ~ price + discount, data = ecommerce_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.301  -9.215   0.251   9.343  75.702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.022446   0.844516   58.05   <2e-16 ***
## price        0.809320   0.002458  329.32   <2e-16 ***
## discount    -2.567371   0.023636 -108.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.06 on 3657 degrees of freedom
## Multiple R-squared:  0.9705, Adjusted R-squared:  0.9705
## F-statistic: 6.025e+04 on 2 and 3657 DF,  p-value: < 2.2e-16
```

The multiple linear regression model analyzing the relationship between `final_price`, `price`, and `discount` reveals that both `price` and `discount` have statistically significant effects on `final_price`, with p-values less than 2.2e-16. The coefficient for `price` is 0.8093, indicating that for each unit increase in `price`, `final_price` increases by approximately 0.8093 units. Conversely, the coefficient for `discount` is -2.5674, indicating that for each unit increase in `discount`, `final_price` decreases by approximately 2.5674 units. The intercept is 49.0224, representing the expected `final_price` when both `price` and `discount` are zero. The residual standard error of 21.06 and an adjusted R-squared value of 0.9705 suggest that the model explains around 97.05% of the variability in `final_price`, indicating a strong fit. The overall model is statistically significant, as evidenced by the high F-statistic (6.025e+04) and a p-value $<$ 2.2e-16.

simple leaner regression where independent variable is a categorical data

dependent variable = final_price | independent variable = category

Ho: There is no statistically significant effect of the different levels of category on the dependent variable final_price

Ha: There is a statistically significant effect of at least one level of category on the dependent variable final_price

```
model_SLR_Cat <- lm(final_price ~ category, data = ecommerce_4)
summary(model_SLR_Cat)
```

```
##
## Call:
## lm(formula = final_price ~ category, data = ecommerce_4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -209.835 -102.690   -7.443   97.452  295.858
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             206.367      5.459  37.800   <2e-16 ***
## categoryBooks             1.778      7.615   0.234    0.815
## categoryClothing         10.799      7.626   1.416    0.157
## categoryElectronics      -4.635      7.748  -0.598    0.550
## categoryHome & Kitchen   -5.405      7.564  -0.714    0.475
## categorySports            2.323      7.665   0.303    0.762
## categoryToys             -1.224      7.654  -0.160    0.873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.7 on 3653 degrees of freedom
## Multiple R-squared:  0.001686,   Adjusted R-squared:  4.65e-05
## F-statistic: 1.028 on 6 and 3653 DF,  p-value: 0.4046
```

The multiple linear regression model analyzing the `final_price` based on different `category` levels in the e-commerce dataset indicates that the intercept is statistically significant with a p-value less than 2e-16, but the categories themselves do not significantly predict the `final_price`, as shown by their non-significant p-values. The residuals demonstrate variability, with the residual standard error being 122.7 on 3653 degrees of freedom. The model explains a very low proportion of variance in `final_price`, with an R-squared value of 0.001686 and an adjusted R-squared value of 0.0000465. The F-statistic of 1.028 and a p-value of 0.4046 indicate that the overall model is not statistically significant.