Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6010 - Probability Theory and Introductory Statistics**

**Assignment:**

MODULE 3 PRACTICE ASSIGNMENT 3

**Submitted on:**

November 20, 2022

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI NIKSHITA RANGANATHAN

**INTRODUCTION**

Amazon is an eCommerce company founded by Jeff Bezos in 1994. The business offers a range of goods for sale, including fashion, gadgets, music, and books. Along with other services, Amazon also provides cloud computing, internet storage, and video streaming. Among these, there are many reasons why people prefer Amazon books. The reason being they can be read on devices such as the Kindle, which can be helpful when traveling. Amazon books can also be borrowed from the library, which is helpful for people who do not own a Kindle. Additionally, Amazon books are often cheaper.

Due to the rise in online book purchases, book sales analysis is important for Amazon. Perhaps the most obvious reason is that it can help authors and publishers understand how well their books are selling. Making a decision for their upcoming book projects will be simpler with this data. Amazon book analytics has a service that allows authors to track the sales and readership of their books on Amazon over time. It provides detailed data on how many copies have been sold, where they were sold, and how many people have read them. Authors can use this data to improve their writing and marketing strategies.

The Amazon top 50 book sales list is a compilation of the best-selling books on Amazon.com. It is updated hourly and includes both fiction and non-fiction books. The list is a great resource for finding new books to read, as well as for keeping track of the best-selling books on Amazon.

**About this Dataset:** This project aims to have an exploratory data analysis on Amazon’s top 50 selling books from 2009 to 2019 and the dataset considered is from the [Kaggle dataset](https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019). This Dataset comprises of 550 rows that contain books of fiction & non-fiction categories. The attribute of this Dataset includes 7 items - Book name, Author, User Rating, Reviews, Price, Year & Genre. The data from the dataset is imported into R and subjected to data analysis to understand the various insights of the Top 10 Amazon Book Sales.

Below are the data descriptions of each variable of the data that briefly describe the contents of the Amazon Book Sales data set. The features of the dataset are as follows:

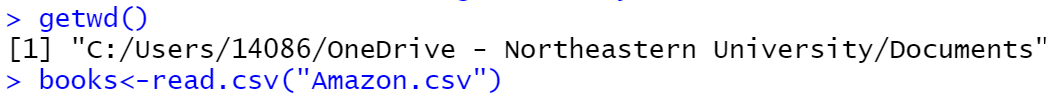
|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | Name | Name of the Book |
| 2 | Author | Author of the Book |
| 3 | User Rating | User Rating on Amazon |
| 4 | Reviews | Written review numbers on Amazon |
| 5 | Price | Price of the book |
| 6 | Year | Year(s) it ranked on the bestseller |
| 7 | Genre | Whether the book is fiction or non-fiction |

*Table 1: Features of the Amazon Book Sales dataset with their dictionary*

**ANALYSIS & INTERPRETATION**

* Importing the Amazon Bestselling Books CSV file

The information about the Top 50 books from 2009 to 2019 is present in the <books> vector.

vector.

**Figure 1-read.csv()**

* Data Cleaning and Manipulation
* Visualization of <NA> values

vis\_dat() and vis\_miss() both return the plot with the missing data.

The naniar package's gg\_miss\_which function is used to determine which column has empty values.

There are no NA values in the books dataset.

******Chart, bar chart

Description automatically generated**Chart

Description automatically generated with low confidence



A picture containing chart

Description automatically generated

Logo

Description automatically generated with low confidence

**Figure 2-<NA> values**

* Eliminating the <NA>s

Text

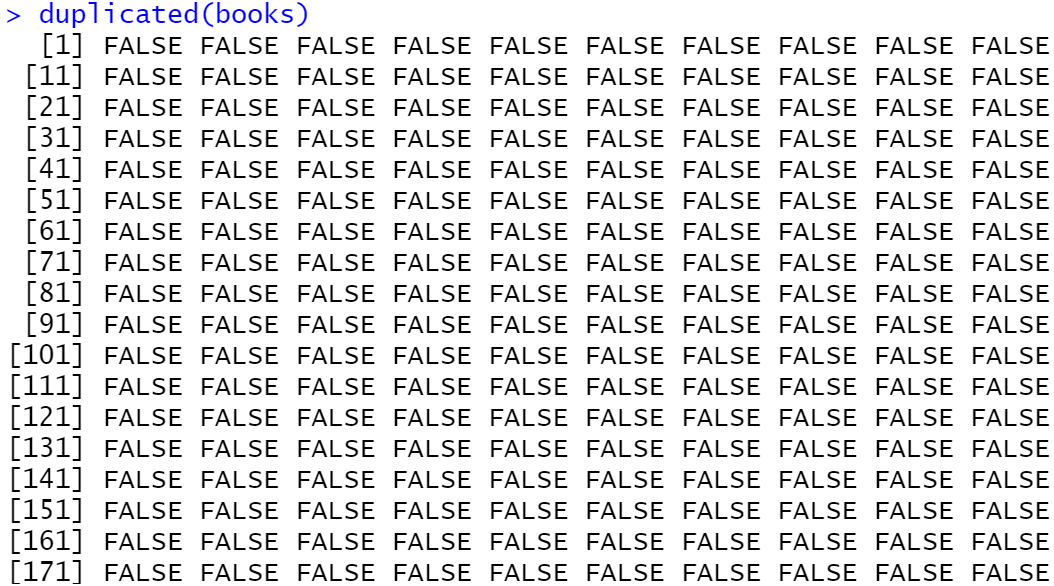
Description automatically generatedThe number of rows reduced from 119390 to 85518.



**Figure 3 - Code for obtaining clean Netflix dataset without <NA> values**

* Duplication

I searched the dataset for duplicate rows. Using anyDuplicated(), we can verify that there are no duplicate rows.



Graphical user interface

Description automatically generated with medium confidence

**Figure 3 – duplicated()and anyDuplicated()**

* Updating the column names



**Figure 4-Changing column names**

* Changing datatypes

The data type of the columns is changed from one datatype to a factor by using the as.factor() function.



**Figure 5 – as.factor()**

* Removing outliers using Boxplots

Values that deviate from the pattern the rest of the dataset follows are known as outliers.

Boxplots are frequently used to identify outliers. I used which() to eliminate all the extreme outliers from the columns.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

The blue-colored dots in below boxplots represent the mean values

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**Figure 6 – Outliers and Boxplots**

* Understanding the Books dataset
* The str() function shows the different variable datatypes. The variables in the dataset include the Title of books, Name of the Author, Book Rating and Reviews, Price, Year, and Genre.
* There are 208 observations under the Fiction category and 272 observations under the Non-Fiction category.

Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

**Figure 7-str() and summary()**

* The dataset consists of 480 observations and 7 attributes. Before data cleaning, the dataset had 550 rows and 7 columns.

A picture containing logo

Description automatically generated

**Figure 8-dim()**

* glimpse() is one of the functions from the dplyr package. headTail function returns the first 4 and last 4 records of “books”.

Text, letter

Description automatically generated

Text

Description automatically generated

**Figure 9-glimpse() and headTail()**

* describe()

Text

Description automatically generated with medium confidence

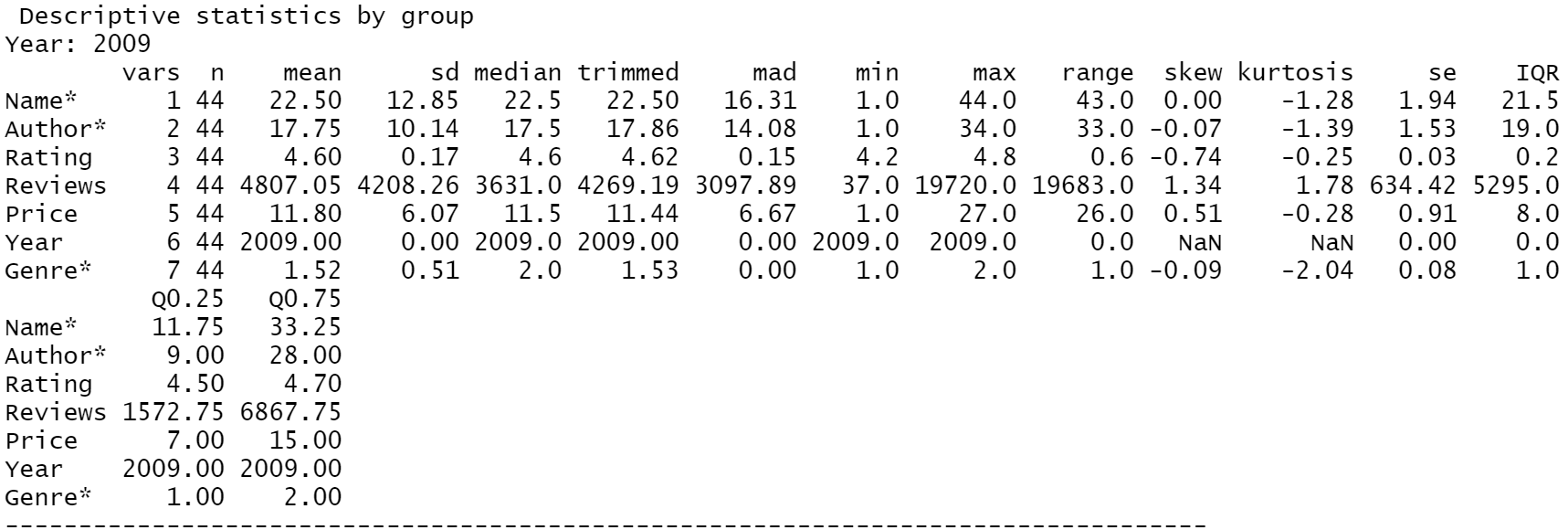
**CONCLUSION**

* A screenshot of a computer

  Description automatically generated with low confidencedescribeBy() for different groups (Genre and Year)

A picture containing text

Description automatically generated



A picture containing text, receipt

Description automatically generated

A screenshot of a computer

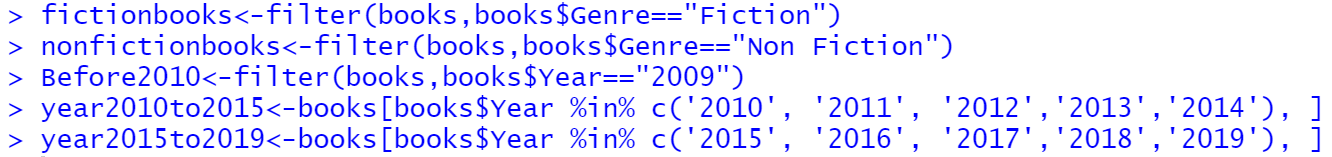
Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

**Figure 10-describe() and describeBy()**

* I created different groups of Fiction and Non-fiction books and also grouped the dataset by year(Before the 2010s, Years 2010 to 2015, and Years 2015 to 2019)



**Figure 11-filter()**

* Histograms for Price, Ratings, and Reviews column

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

**Chart, histogram

Description automatically generated**

**Figure 12-histograms using ggplot**

* Checking Normality

**Chart, line chart, histogram

Description automatically generatedChart, line chart

Description automatically generated**Chart, line chart

Description automatically generated

**Figure 13-QQ plots**

The above QQ plots represent non-normal distribution.

* Checking the Normality of Non-Fiction and Fiction groups

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart

Description automatically generatedDiagram

Description automatically generatedChart, line chart

Description automatically generated**

**Figure 14-QQ plots of groups**

* Density Plots

Features shown in the density graphs are not normally distributed.

**Chart

Description automatically generated with medium confidence**

**Figure 15-Density plots**

* Shapiro tests

To determine if a sample of data has a normal distribution, a statistical technique known as the Shapiro-Wilk test is utilized. The normality of the characteristics is rejected since the p-value is below 0.05.

**Graphical user interface, text, application

Description automatically generated**

**Figure 16-shapiro.test()**

* One-Sample T-tests
  + Mean of Price feature of Fiction books

**Null Hypothesis H0:** True Mean is less than equal to 10

**Alternate Hypothesis H1:** True Mean is greater than 10

**Text, letter

Description automatically generated**

**Figure 17-t.test() of Fiction books (Price feature)**

**Degree of Freedom: 207**

**Confidence Interval: 95% (Default)**

**P value: 0.9769879**

We fail to reject the test's null hypothesis since the p-value is more than the significance level(α) of 0.05. There is insufficient evidence to disprove Ho.

* + Mean of Reviews feature of Fiction books

**Null Hypothesis H0:** True Mean is greater than equal to 14000

**Alternate Hypothesis H1:** True Mean is less than 14000

**Text

Description automatically generated**

**Figure 18-t.test() of Fiction books (Reviews feature)**

**Degree of Freedom: 207**

**Confidence Interval: 95% (Default)**

**P value: 0.01316908**

We reject the null hypothesis of the test because the p-value is less than the significance level of 0.05.There is sufficient evidence to disprove Ho.

* + Mean of Price feature of Non-Fiction books

**Null Hypothesis H0:** True Mean is greater than equal to 15

**Alternate Hypothesis H1:** True Mean is less than 15

**Text

Description automatically generated**

**Figure 19-t.test() of Non-Fiction books (Price feature)**

**Degree of Freedom: 271**

**Confidence Interval: 95% (Default)**

**P value: 7.176864e-19**

Since the p-value is under the significance level of 0.05, we reject the test's null hypothesis. There is sufficient evidence to disprove Ho.

* + Mean of Reviews feature of Non-Fiction books

**Null Hypothesis H0:** True Mean is less than equal to 10000

**Alternate Hypothesis H1:** True Mean is greater than 10000

**Text, letter

Description automatically generated**

**Figure 20-t.test() of Non-Fiction books (Reviews feature)**

**Degree of Freedom: 271**

**Confidence Interval: 95% (Default)**

**P value: 0.9512266**

We fail to reject the test's null hypothesis since the p-value is more than the significance level(α) of 0.05. There is insufficient evidence to disprove Ho.

* Two-Sample T-test

**Null Hypothesis H0:** The difference in means between group Fiction and Non-fiction is equal to 0

**Alternate Hypothesis H1:** The difference in means between group Fiction and Non-fiction is not equal to 0

**Text, letter

Description automatically generated**

**Figure 21- Welch Two sample test of books by Genre (Rating feature)**

**Degree of Freedom: 407.76**

**Confidence Interval: 90 %**

**P value: 1.459616e-05**

We reject the null hypothesis of the test because the p-value is less than the significance level of 0.05. There is sufficient evidence to disprove Ho.

* Hypothesis testing for p-value
  + Mean of Price feature

**Null Hypothesis H0:** The sample mean (Price Before Year 2010s) is greater than or equal to the population mean

**Alternate Hypothesis H1:** The sample mean (Price Before Year 2010s) is less than the population mean

**Graphical user interface, text

Description automatically generated**

**Figure 22- P value calculation (Price feature)**

**Degree of Freedom: 43**

**T score: 1.190715, P value: 0.8798513**

**Lower.Tail=TRUE**

Due to the p-value above the 0.05 level of significance, we are unable to reject the null hypothesis of the test. There is not enough evidence to disprove Ho.

* + Mean of the Reviews feature

**Null Hypothesis H0:** The sample mean (Reviews Before Year 2010s) is greater than or equal to the population mean

**Alternate Hypothesis H1:** The sample mean (Reviews Before Year 2010s) is less than the population mean

**Graphical user interface, text

Description automatically generated**

**Figure 22- P value calculation (Reviews feature)**

**Degree of Freedom: 43**

**T score: -9.317699, P value: 3.524633e-12**

**Lower.Tail=TRUE**

We reject the test's null hypothesis since the p-value is below the significance level of 0.05. There is sufficient evidence to disprove Ho.

**CONCLUSION**

The dataset of 'Top 50 Amazon Book Sales' has provided various insights about the Book types, Prices, User Ratings & Reviews, and also shows variations in the patterns between them. In this project, I performed initial data analysis, exploratory data analysis, and calculations based on statistics and plotted a few data visualization graphs to understand the analysis. I also performed normality tests and various hypothesis testing based on the data. The inferences & findings of this analysis are presented below:

* Density plots can be used to assess the visual normality of the characteristics It can be seen that none of the features ( Price, Reviews, and Ratings ) follow a normal distribution.
* It was possible to learn more about the normality of the attributes using Shapiro-Wilks Tests and QQ plots too.
* One sample t-test compares the sample's mean value to a known value It is decided whether to accept or reject the assumed null hypothesis.
* A two-sample t-test was carried out to ascertain the relationship between two separate samples. It can be concluded that the means of the two categories (fiction books and non-fiction books) are not equal.
* Hypothesis testing with P-value helped us understand if the mean population is greater, equal, or less than the mean of the samples.

**REFERENCES**

Saalu, Sooter. (2020, October). Amazon Top 50 Bestselling Books 2009 - 2019. Kaggle. Retrieved November 16, 2022, from https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

**APPENDIX: CODE**

#---------------------- Week\_3\_Module\_3 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Week 3 Assignment - Module 3 R Practice")

print("Course Name - ALY6010: Probability Theory and Introductory Statistics")

# Loading the packages

library(dplyr)

library(visdat)

library(psych)

library(ggplot2)

library(naniar)

library(skimr)

library(moments)

library(magrittr)

# Importing the dataset of student performance

# Check the current working directory

getwd()

books<-read.csv("Amazon.csv")

#------------------- Data Cleaning -------------------#

# Checking NA values

vis\_dat(books)

gg\_miss\_which(books)

vis\_miss(books)

sum(is.na(books))

sum(is.null(books))

# Checking for duplication

duplicated(books)

anyDuplicated(books)

sum(duplicated(books))

# Changing column names

names(books)<-c('Name','Author','Rating','Reviews','Price','Year','Genre')

View(books)

# Changing datatypes

books$Genre=as.factor(books$Genre)

# Boxplots

boxplot(books$Reviews)

outlier\_1<-boxplot(books$Reviews,plot = FALSE)$out

outlier\_1

books<-books[-which(books$Reviews %in% outlier\_1),]

boxplot(books$Reviews~books$Genre,main = "Boxplot A",xlab = "Genre",ylab = "Reviews",col = c("#FFE0B2", "#F57C00"))

means <- tapply(books$Reviews,books$Genre, mean)

points(means, pch=20,col="royalblue1")

boxplot(books$Rating)

outlier\_2<-boxplot(books$Rating,plot = FALSE)$out

outlier\_2

books<-books[-which(books$Rating %in% outlier\_2),]

boxplot(books$Rating~books$Genre,main = "Boxplot B",xlab = "Genre",ylab = "Rating",col = c("#FFE0B2", "#F57C00"))

means <- tapply(books$Rating, books$Genre, mean)

points(means, pch=20,col="royalblue1")

boxplot(books$Price)

outlier\_3<-boxplot(books$Price,plot = FALSE)$out

outlier\_3

books<-books[-which(books$Price %in% outlier\_3),]

boxplot(books$Price~books$Genre,main = "Boxplot C",xlab = "Genre",ylab = "Price",col = c("#FFE0B2", "#F57C00"))

means <- tapply(books$Price, books$Genre, mean)

points(means, pch=20,col="royalblue1")

#------------------- Exploratory Data Analysis -------------------#

# Analysis

str(books)

summary(books)

glimpse(books)

dim(books)

headTail(books)

describe(books,quant = c(0.25, 0.75),IQR = T)

describeBy(books,group="Genre",quant = c(0.25, 0.75), IQR = T)

describeBy(books,group="Year",quant = c(0.25, 0.75), IQR = T)

# Creating groups

fictionbooks<-filter(books,books$Genre=="Fiction")

nonfictionbooks<-filter(books,books$Genre=="Non Fiction")

Before2010<-filter(books,books$Year=="2009")

year2010to2015<-books[books$Year %in% c('2010', '2011', '2012','2013','2014'), ]

year2015to2019<-books[books$Year %in% c('2015', '2016', '2017','2018','2019'), ]

# Histograms of Features

library(ggplot2)

library(viridis)

ggplot(books,aes(`Price`))+geom\_histogram(bins=20,aes(fill=..count..))+labs(title="Histogram for Price")+xlim(0,120)

ggplot(books,aes(`Rating`))+geom\_histogram(bins=20,aes(fill=..count..))+labs(title="Histogram for Ratings")+xlim(0,6)+scale\_fill\_distiller(type = "div")

ggplot(books,aes(`Reviews`))+geom\_histogram(aes(fill=..count..))+labs(title="Histogram for Reviews")+xlim(0,20000)+scale\_fill\_viridis()

# Normal QQ plots

qqnorm(books$Price, pch = 1, frame = FALSE,main="Q-Q Plot (Price)",ylim = c(0, 120),xlim = c(-4, 4))

qqline(books$Price, col = "tomato", lwd = 2)

qqnorm(books$Rating,pch = 1, frame = FALSE,main="Q-Q Plot (Rating)",ylim = c(0, 6),xlim = c(-4, 4))

qqline(books$Rating, col = "tomato", lwd = 2)

qqnorm(books$Reviews,pch = 1, frame = FALSE,main="Q-Q Plot (Reviews)",xlim = c(-4, 4),ylim = c(0, 90000))

qqline(books$Reviews, col = "tomato", lwd = 2)

library("ggpubr")

ggqqplot(nonfictionbooks$Rating,main="Q-Q Plot (Non-Fiction Rating)")

ggqqplot(nonfictionbooks$Reviews,main="Q-Q Plot (Non-Fiction Reviews)")

ggqqplot(nonfictionbooks$Price,main="Q-Q Plot (Non-Fiction Price)")

ggqqplot(fictionbooks$Rating,main="Q-Q Plot (Fiction Rating)")

ggqqplot(fictionbooks$Reviews,main="Q-Q Plot (Fiction Reviews)")

ggqqplot(fictionbooks$Price,main="Q-Q Plot (Fiction Price)")

# Density Plots

library(gridExtra)

library(grid)

library(ggplot2)

library(wesanderson)

density1<-ggplot(books, aes(x = `Price`, colour = `Genre`)) +geom\_density()+ scale\_color\_manual(values = c('#EEB422', '#238E68'))

density2<-ggplot(books, aes(x = `Rating`, colour = `Genre`)) +geom\_density()+ scale\_color\_manual(values = c('#EEB422', '#238E68'))

density3<-ggplot(books, aes(x = `Reviews`, colour = `Genre`)) +geom\_density()+ scale\_color\_manual(values = c('#EEB422', '#238E68'))

grid.arrange(density1,density2,density3)

# Shapiro test

shapiro.test(books$Reviews)

shapiro.test(books$Rating)

shapiro.test(books$Price)

# One Sample T tests

# One-sample t test for mean of Price-Fiction Books

# H0: mu <= 10

# H1: mu > 10

# 95% confidence level for mu

ttest1<-t.test(fictionbooks$Price,mu=10,alt="greater")

ttest1

ttest1$p.value

# One-sample t test for mean of Reviews

# H0: mu >= 14000

# H1: mu < 14000

# 95% confidence level for mu

ttest2<-t.test(fictionbooks$Reviews,mu=14000,alt="less")

ttest2

ttest2$p.value

# One-sample t test for mean of Price

# H0: mu >= 15

# H1: mu < 15

# 95% confidence level for mu

ttest3<-t.test(nonfictionbooks$Price,mu=15,alt="less")

ttest3

ttest3$p.value

# One-sample t test for mean of Reviews

# H0: mu <= 10000

# H1: mu > 10000

# 95% confidence level for mu

ttest4<-t.test(nonfictionbooks$Reviews,mu=10000,alt="greater")

ttest4

ttest4$p.value

# Welch two sample t-test for mean of User Rating

# H0: Difference in means between group Fiction and Non-fiction is equal to 0

# H1: Difference in means between group Fiction and Non-fiction is not equal to 0

ttest5<-t.test(`Rating`~ `Genre`,data=books,conf=0.90)

ttest5

ttest5$p.value

# P value calculation

# Sample 1 - Before 2010 - Price

# H0: The sample mean >= to population mean

# H1: The sample mean < population mean

# Sample Mean

Samplemean<-mean(Before2010$Price)

Samplemean

# Population mean

Populationmean<-mean(books$Price)

Populationmean

# Sample size(n)=50

n<-nrow(Before2010)

degreeoffreedom<-n-1

# Standard Deviation of the sample (SD)

sd<-sd(Before2010$Price)

# T score

tscore<-(Samplemean-Populationmean)/(sd/sqrt(n))

tscore

# P value

pt(tscore,degreeoffreedom,lower.tail = TRUE )

# Sample 2 - Before 2010 - Reviews

# H0: The sample mean >= to population mean

# H1: The sample mean < population mean

# Sample Mean

Samplemean2<-mean(Before2010$Reviews)

Samplemean2

# Population mean

Populationmean2<-mean(books$Reviews)

Populationmean2

# Sample size(n)=50

n2<-nrow(Before2010)

degreeoffreedom2<-n2-1

# Standard Deviation of the sample (SD)

sd2<-sd(Before2010$Reviews)

# T score

tscore2<-(Samplemean2-Populationmean2)/(sd2/sqrt(n2))

tscore2

# P value

pt(tscore2,degreeoffreedom2,lower.tail = TRUE)