Diwakar Sinha R Graded assignment 2

# Libraries used: read\_excel, dplyr, ggplot2

## Q1

# Null Hypotheses: H0: Avg Men Salary = Avg womens Salary

# Alternate Hypotheses: Ha: Avg Men Salary != Avg womens Salary

# Confidence level= 95% = 0.95

# Significance level = 5% = 0.05

*No of observations = 30 each, since population standard deviation is unknown we use a T test*

### 2 sample unpaired T test

t.test(df$`Men Salary`,df$`Women Salary`,conf.level = 0.95)

# p- Value = 0.001494 which is < Significance level

# we have significant evidence to reject the Null Hypothesis H0

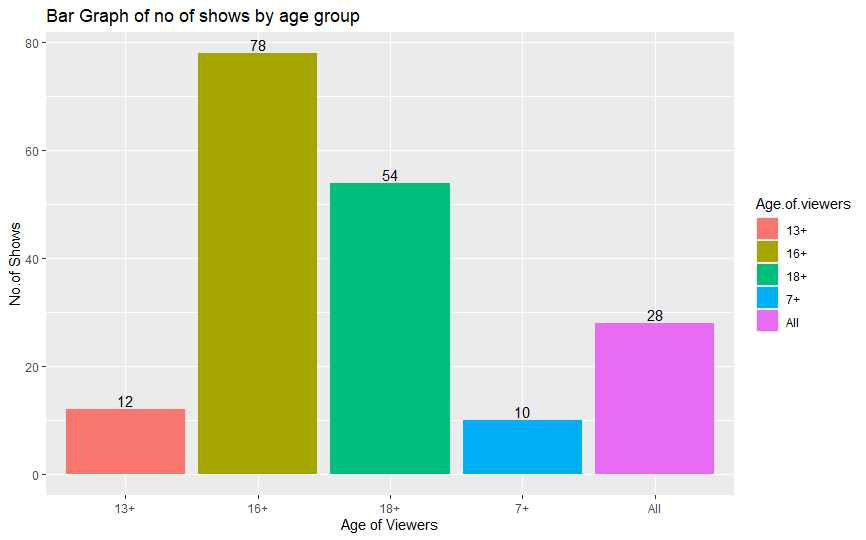
# Business conclusion : there is Salary disparity between men and women in this workplace

The sample estimates show that the mean salary for men is $54413.40, while the mean salary for women is $41256.87. This indicates that, on average, men earn more than women in this sample.

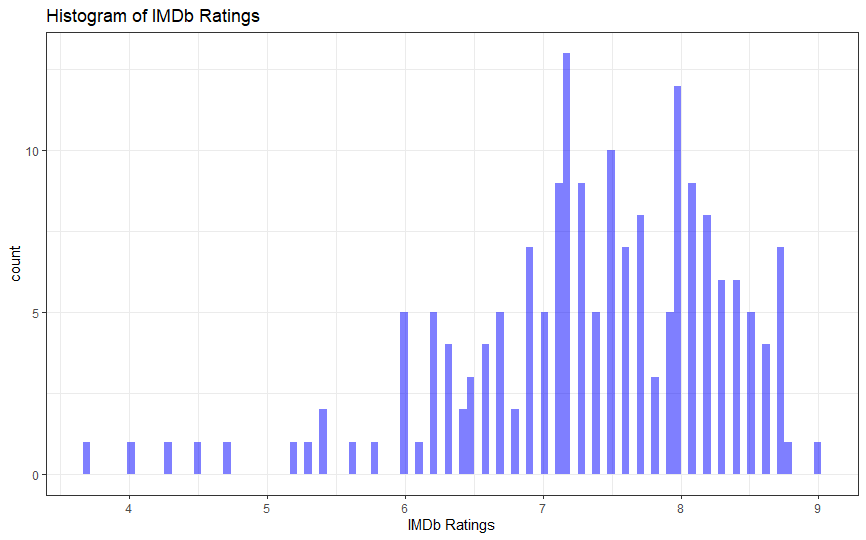
In summary, the t-test results indicate that there is a statistically significant difference in the means of the men's and women's salary in this sample, with men earning more on average.

## Q2a

Read dataset > Create a sub dataset using dplyr



## 2b

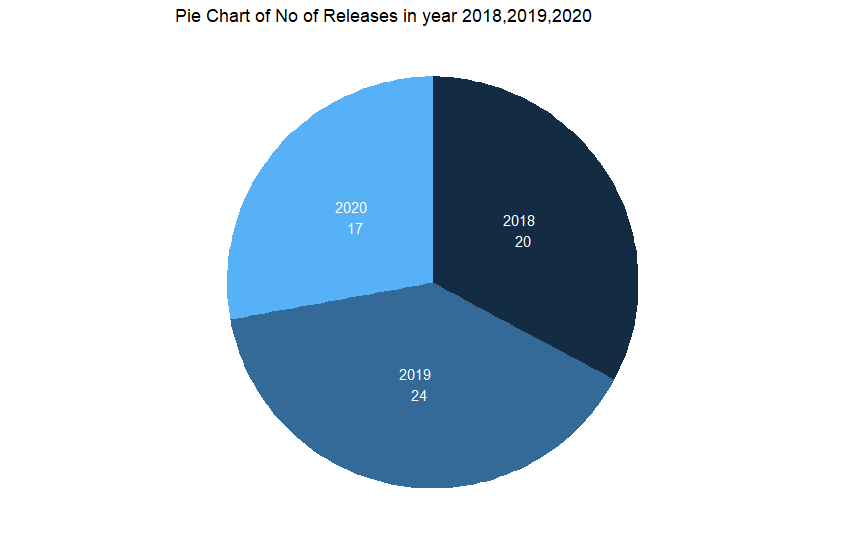
bins is chosen as 100 because there are 100 possible IMDb scores between 0 & 10

## 2c: top 3 shows by Imdb ratings are

Chart, bar chart

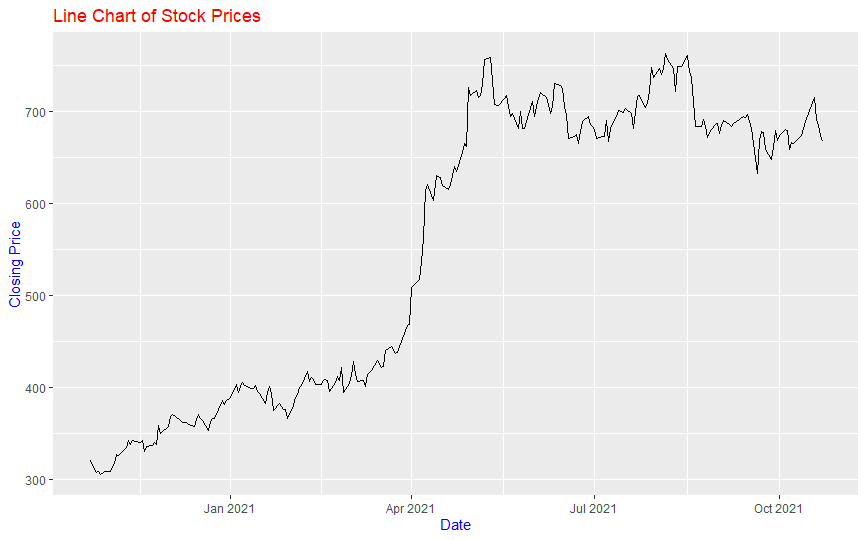
Description automatically generated

## 2d)

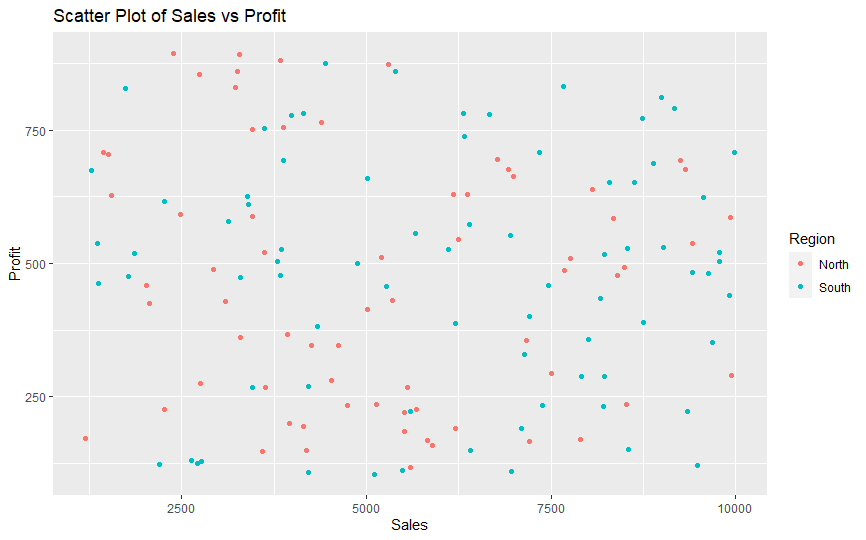
1st filter the shows in our years of interest 2018,’19,’20 then group by to get a count using dplyr

## 3a)

linegraph with title and axes labels colors changed

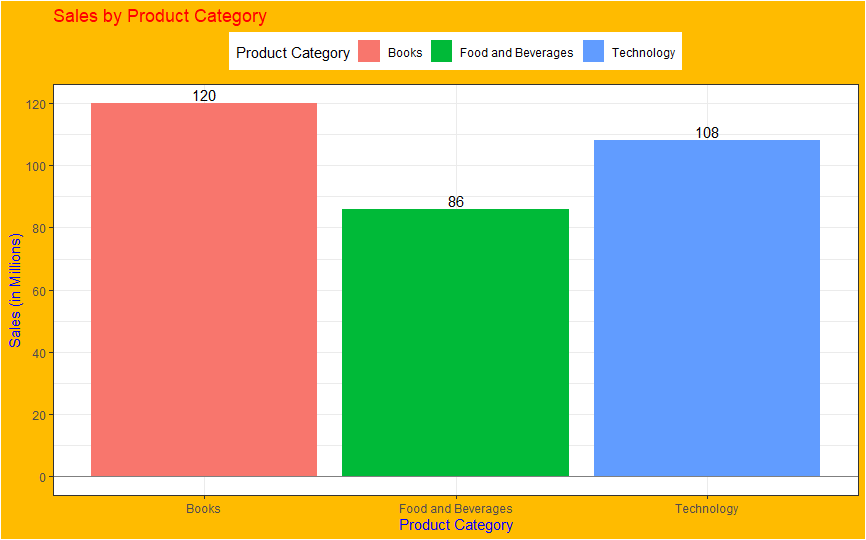


## 3b scatter plot



## 3c) layers added:

|  |  |  |
| --- | --- | --- |
| 1. Different colors for bars | 1. Grid lines with tick size of 20 | 1. Axis labels |
| 1. Red axis labels | 1. Plot title | 1. Blue plot title |
| 1. Orange background | 1. Legend | 1. Value labels on bars |



## 4a)

Import data and build linear regression model . summary results in R File :

Call:

lm(formula = Sales ~ TV + Radio + Newspaper, data = data2)

Residuals:

Min 1Q Median 3Q Max

-7.3034 -0.8244 -0.0008 0.8976 3.7473

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.6251241 0.3075012 15.041 <2e-16 \*\*\*

TV 0.0544458 0.0013752 39.592 <2e-16 \*\*\*

Radio 0.1070012 0.0084896 12.604 <2e-16 \*\*\*

Newspaper 0.0003357 0.0057881 0.058 0.954

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.662 on 196 degrees of freedom

Multiple R-squared: 0.9026, Adjusted R-squared: 0.9011

F-statistic: 605.4 on 3 and 196 DF, p-value: < 2.2e-16

## 4b)

From the above summary we can infer that The dependent variable, Sales, is modeled as a linear function of three predictor variables: TV, Radio, and Newspaper. The coefficients for TV and Radio are both significantly different from zero at the 0.001 level, indicating that both of these variables are important predictors of Sales. The coefficient for Newspaper, however, is not significantly different from zero at the 0.05 level, indicating that it may not be a useful predictor of Sales.

The overall model has a good fit, with a **multiple R-squared value of 0.9026 and an adjusted R-squared value of 0.9011. This means that the model explains about 90% of the variance in the data.** The F-statistic is also large and significant, indicating that the model as a whole is a good fit to the data.