Final Project

MATH 40024/50024: Computational Statistics

October 19, 2023

ACADEMIC INTEGRITY: Every student should complete the project by their own. A project report having high degree of similarity with work by any other student, or with any other document (e.g., found online) is considered plagiarism, and will not be accepted. The minimal consequence is that the student will receive the project score of 0, and the best possible overall course grade will be D. Additional consequences are described at http://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism and will be strictly enforced.

Instruction

Goal: The goal of the project is to go through the complete data analysis workflow to answer questions about your chosen topic using a real-life dataset. You will need to acquire the data, munge and explore the data, perform statistical analysis, and communicate the results.

Report: Use this Rmd file as a template. Edit the file by adding your project title in the YAML, and including necessary information in the four sections: (1) Introduction, (2) Computational Methods, (3) Data Analysis and Results, and (4) Conclusion.

Submission: Please submit your project report as a PDF file (8-10 pages, flexible) to Canvas by **11:59 p.m. on December 10**. The PDF file should be generated by "knitting" the Rmd file. You may choose to first generate an HTML file (by changing the output format in the YAML to `output: html_document`) and then convert it to PDF. **20 points will be deducted if the submitted files are in wrong format.**

Grade: The project will be graded based on your ability to (1) recognize and define research questions suitable for data-driven, computational approaches, (2) use computational methods to analyze data, (3) appropriately document the process (with R code) and clearly present the results, and (4) draw valid conclusions supported by the data analysis.

Example topics:

- Post-Hurricane Vital Statistics
- Tidy Tuesday

Datasets: I suggest to work on a dataset with at least thousands of observations and dozens of variables. You may consider (but are not restricted) to use the following data repositories: Data.gov, Kaggle, FiveThirtyEight, ProPublica, and UCI Machine Learning Repository

Introduction [15 points]

- ### What research question(s) would you like to answer?

1)What have been the main sales trends over the years? Are there any specific products, categories, or geographical areas where sales are significantly increasing or decreasing?

2)Is it possible to spot trends in consumer behaviour, such as recurrent purchases or brand loyalty?

3)What is the impact of discounts on overall sales and profitability? Are certain products or customer segments more responsive to discounts?

- ### Why a data-driven, computational approach may be useful to answer the questions?

Because of the complexity of the Global Superstore Dataset many variables and large amounts of data a data-driven, computational method is required to extract subtle insights. The size and complexity of the large dataset make manual examination impractical. Statistical techniques and machine learning algorithms in particular are computationally intensive and are essential for detecting hidden patterns, identifying complex relationships, and performing predictive modelling. A thorough investigation of sales trends, consumer behaviour, and profitability dynamics is ensured by the computational power required to navigate and process the data effectively when multiple variables are integrated. Automation makes reproducibility easier, and scalability makes it possible to adjust to big datasets. The computational method offers a comprehensive understanding of the complex interactions within the Global Superstore Dataset by facilitating the synthesis of multifaceted information and speeding up the analysis.

- ### Describe the dataset that you choose.

```r
# choosen dataset
library(readr)
superstore <- read_csv("C:/Users/Sreedhar Jhansy/Desktop/superstore.csv")
```

```
## Rows: 51290 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (16): Category, City, Country, Customer.ID, Customer.Name, Market, Orde...
## dbl   (8): Discount, Profit, Quantity, Row.ID, Sales, Shipping.Cost, Year, w...
## time  (2): Order.Date, Ship.Date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
View(superstore)
```

category: The category of products sold in the superstore. city: The city where the order was placed. country: The country in which the superstore is located. customer_id: A unique identifier for each customer. customer_name: The name of the customer who placed the order. discount: The discount applied to the order. market: The market or region where the superstore operates.

order_date: The date when the order was placed. order_id: A unique identifier for each order. order_priority: The priority level of the order. product_id: A unique identifier for each product. product_name: The name of the product. profit: The profit generated from the order. quantity: The quantity of products ordered. region: The region where the order was placed. row_id: A unique identifier for each row in the dataset. sales: The total sales amount for the order. segment: The customer segment (e.g., consumer, corporate, or home office). ship_date: The date when the order was shipped. ship_mode: The shipping mode used for the order. shipping_cost: The cost of shipping for the order. state: The state or region within the country. sub_category: The subcategory of products within the main category. year: The year in which the order was placed. market2: Another column related to market information. weeknum: The week number when the order was placed.

enteries:51290 column:26 This dataset provides a comprehensive snapshot for diverse data analysis tasks in the context of the global superstore’s operations.

Computational Methods [30 points]

-

For the choosen dataset, what are the necessary data wrangling steps to make the data ready for subsequent analyses?

The steps for datawrangling to the dataset is to load the dataset identify the missing values and remove the missing values,identify and remove the duplicate values,convert the datatypes if needed.

```r
#identify missing values
missing_values = colSums(is.na(superstore))
#printing the missing values
missing_values
```

```
##       Category           City        Country    Customer.ID  Customer.Name 
##              0              0              0              0              0 
##       Discount         Market     Order.Date       Order.ID Order.Priority 
##              0              0              0              0              0 
##     Product.ID   Product.Name         Profit       Quantity         Region 
##              0              0              0              0              0 
##         Row.ID          Sales        Segment      Ship.Date      Ship.Mode 
##              0              0              0              0              0 
##  Shipping.Cost          State   Sub.Category           Year        Market2 
##              0              0              0              0              0 
##        weeknum 
##              0
```

```r
#removing the missing values
superstore = na.omit(superstore)
superstore
```

```
## # A tibble: 51,290 x 26
##    Category   City  Country Customer.ID Customer.Name Discount Market Order.Date
##    <chr>      <chr> <chr>   <chr>       <chr>            <dbl> <chr>  <time>    
##  1 Office Su~ Los ~ United~ LS-172304   Lycoris Saun~        0 US     00'00"    
##  2 Office Su~ Los ~ United~ MV-174854   Mark Van Huff        0 US     00'00"    
##  3 Office Su~ Los ~ United~ CS-121304   Chad Sievert         0 US     00'00"    
##  4 Office Su~ Los ~ United~ CS-121304   Chad Sievert         0 US     00'00"    
##  5 Office Su~ Los ~ United~ AP-109154   Arthur Prich~        0 US     00'00"    
##  6 Office Su~ Los ~ United~ JF-154904   Jeremy Farry         0 US     00'00"    
##  7 Office Su~ Los ~ United~ WB-218504   William Brown        0 US     00'00"    
##  8 Office Su~ Los ~ United~ JA-159704   Joseph Airdo         0 US     00'00"    
##  9 Office Su~ Los ~ United~ SP-209204   Susan Pistek         0 US     00'00"    
## 10 Office Su~ Los ~ United~ RL-196154   Rob Lucas            0 US     00'00"    
## # i 51,280 more rows
## # i 18 more variables: Order.ID <chr>, Order.Priority <chr>, Product.ID <chr>,
## #   Product.Name <chr>, Profit <dbl>, Quantity <dbl>, Region <chr>,
## #   Row.ID <dbl>, Sales <dbl>, Segment <chr>, Ship.Date <time>,
## #   Ship.Mode <chr>, Shipping.Cost <dbl>, State <chr>, Sub.Category <chr>,
## #   Year <dbl>, Market2 <chr>, weeknum <dbl>
```

```r
summary(superstore)
```

```
##    Category             City             Country           Customer.ID       
##  Length:51290       Length:51290       Length:51290       Length:51290      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  Customer.Name         Discount         Market           Order.Date      
##  Length:51290       Min.   :0.0000   Length:51290       Length:51290     
##  Class :character   1st Qu.:0.0000   Class :character   Class1:hms       
##  Mode  :character   Median :0.0000   Mode  :character   Class2:difftime  
##                     Mean   :0.1429                      Mode  :numeric   
##                     3rd Qu.:0.2000                                       
##                     Max.   :0.8500                                       
##    Order.ID         Order.Priority      Product.ID        Product.Name      
##  Length:51290       Length:51290       Length:51290       Length:51290      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      Profit            Quantity         Region              Row.ID     
##  Min.   :-6599.98   Min.   : 1.000   Length:51290       Min.   :    1  
##  1st Qu.:    0.00   1st Qu.: 2.000   Class :character   1st Qu.:12823  
##  Median :    9.24   Median : 3.000   Mode  :character   Median :25646  
##  Mean   :   28.61   Mean   : 3.477                      Mean   :25646  
##  3rd Qu.:   36.81   3rd Qu.: 5.000                      3rd Qu.:38468  
##  Max.   : 8399.98   Max.   :14.000                      Max.   :51290  
##      Sales          Segment           Ship.Date         Ship.Mode        
##  Min.   :    0.0   Length:51290       Length:51290      Length:51290      
##  1st Qu.:   31.0   Class :character   Class1:hms        Class :character  
##  Median :   85.0   Mode  :character   Class2:difftime   Mode  :character  
##  Mean   :  246.5                      Mode  :numeric                      
##  3rd Qu.:  251.0                                                          
##  Max.   :22638.0                                                          
##  Shipping.Cost        State           Sub.Category            Year     
##  Min.   :  0.002   Length:51290       Length:51290       Min.   :2011  
##  1st Qu.:  2.610   Class :character   Class :character   1st Qu.:2012  
##  Median :  7.790   Mode  :character   Mode  :character   Median :2013  
##  Mean   : 26.376                                          Mean   :2013  
##  3rd Qu.: 24.450                                          3rd Qu.:2014  
##  Max.   :933.570                                          Max.   :2014  
##    Market2              weeknum     
##  Length:51290       Min.   : 1.00
##  Class :character   1st Qu.:20.00
##  Mode  :character   Median :33.00
##                     Mean   :31.29
##                     3rd Qu.:44.00
##                     Max.   :53.00
```

```
#identify duplicate rows
duplicate_rows = superstore[duplicated(superstore), ]
duplicate_rows
```

```
## # A tibble: 0 x 26
## # i 26 variables: Category <chr>, City <chr>, Country <chr>, Customer.ID <chr>,
## #   Customer.Name <chr>, Discount <dbl>, Market <chr>, Order.Date <time>,
## #   Order.ID <chr>, Order.Priority <chr>, Product.ID <chr>, Product.Name <chr>,
## #   Profit <dbl>, Quantity <dbl>, Region <chr>, Row.ID <dbl>, Sales <dbl>,
## #   Segment <chr>, Ship.Date <time>, Ship.Mode <chr>, Shipping.Cost <dbl>,
## #   State <chr>, Sub.Category <chr>, Year <dbl>, Market2 <chr>, weeknum <dbl>
```

```
#removing duplicate rows
superstore = unique(superstore)
superstore
```

```
## # A tibble: 51,290 x 26
##    Category   City  Country Customer.ID Customer.Name Discount Market Order.Date
##    <chr>      <chr> <chr>   <chr>       <chr>            <dbl> <chr>  <time>
##  1 Office Su~ Los ~ United~ LS-172304   Lycoris Saun~        0 US     00’00"
##  2 Office Su~ Los ~ United~ MV-174854   Mark Van Huff        0 US     00’00"
##  3 Office Su~ Los ~ United~ CS-121304   Chad Sievert         0 US     00’00"
##  4 Office Su~ Los ~ United~ CS-121304   Chad Sievert         0 US     00’00"
##  5 Office Su~ Los ~ United~ AP-109154   Arthur Prich~        0 US     00’00"
##  6 Office Su~ Los ~ United~ JF-154904   Jeremy Farry         0 US     00’00"
##  7 Office Su~ Los ~ United~ WB-218504   William Brown        0 US     00’00"
##  8 Office Su~ Los ~ United~ JA-159704   Joseph Airdo         0 US     00’00"
##  9 Office Su~ Los ~ United~ SP-209204   Susan Pistek         0 US     00’00"
## 10 Office Su~ Los ~ United~ RL-196154   Rob Lucas            0 US     00’00"
## # i 51,280 more rows
## # i 18 more variables: Order.ID <chr>, Order.Priority <chr>, Product.ID <chr>,
## #   Product.Name <chr>, Profit <dbl>, Quantity <dbl>, Region <chr>,
## #   Row.ID <dbl>, Sales <dbl>, Segment <chr>, Ship.Date <time>,
## #   Ship.Mode <chr>, Shipping.Cost <dbl>, State <chr>, Sub.Category <chr>,
## #   Year <dbl>, Market2 <chr>, weeknum <dbl>
```

-

What exploratory analyses and modeling techniques can be used to answer the research questions?

1)Data Visualisation: To identify particular products, categories, or geographic areas with notable sales variations,create visualisations using tools like line charts, bar graphs, and heatmaps.

2)Customer Segmentation: To identify distinct consumer segments with distinctive purchasing patterns, apply clustering techniques.By identifying correlations between products that are frequently bought together, association rule mining can provide valuable insights into consumer preferences and brand loyalty.

3)Customer Segmentation: Using segmentation based on demographics or past purchases, examine whether specific customer segments react better to discounts. Profitability Modelling: Using factors like product category and customer segments, develop predictive models to estimate how discounts affect overall profitability.

-

What metrics will be used to evaluate the quality of the data analysis?

1)To identify particular regions of notable growth or decline, analyse the sales distribution across various products, categories, or geographic locations. Metric: Sales Distribution by Product/Category/Geographical Area

2)explanation: This metric measures the proportion of customers who, over a given time period, make repeat purchases, indicating brand loyalty and recurring consumer behaviour. Metric: Customer Retention Rate

3)To determine how discounts affect financial metrics, consider the percentage change in overall sales and profitability after applying discounts. Metric: Sales and Profitability Change as a Percentage

Data Analysis and Results [40 points]

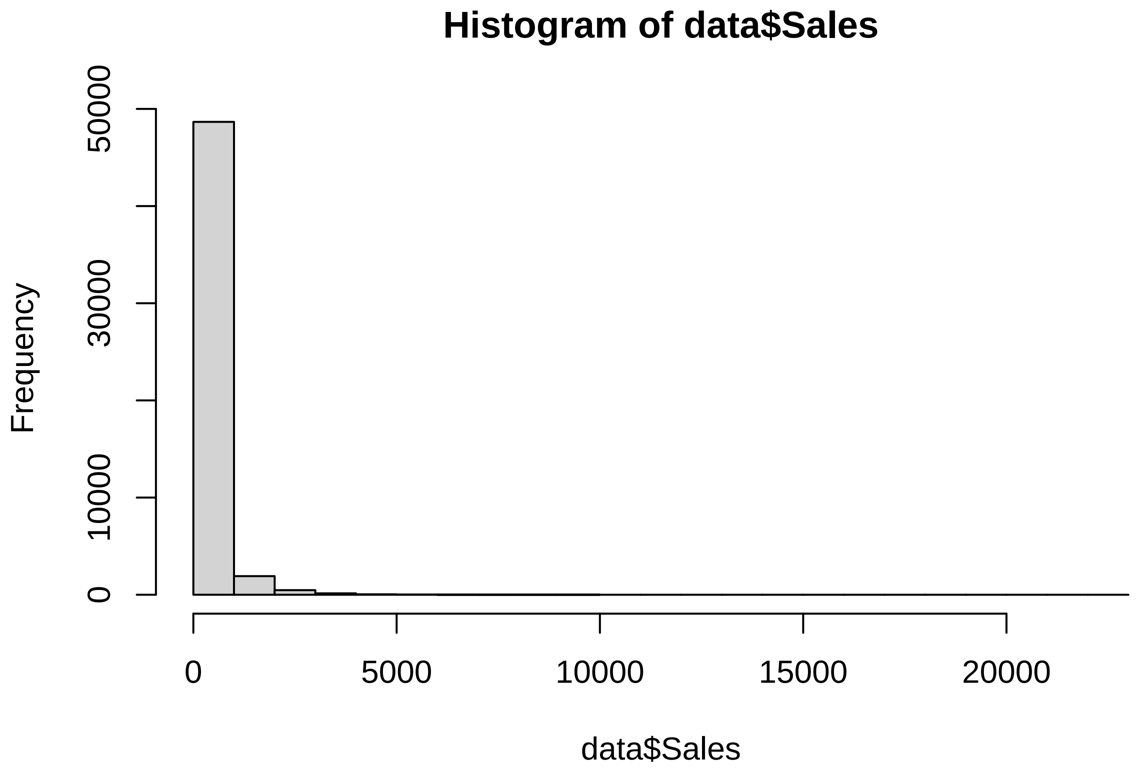
-

Perform data analysis, document the analysis procedure, and evaluate the outcomes.

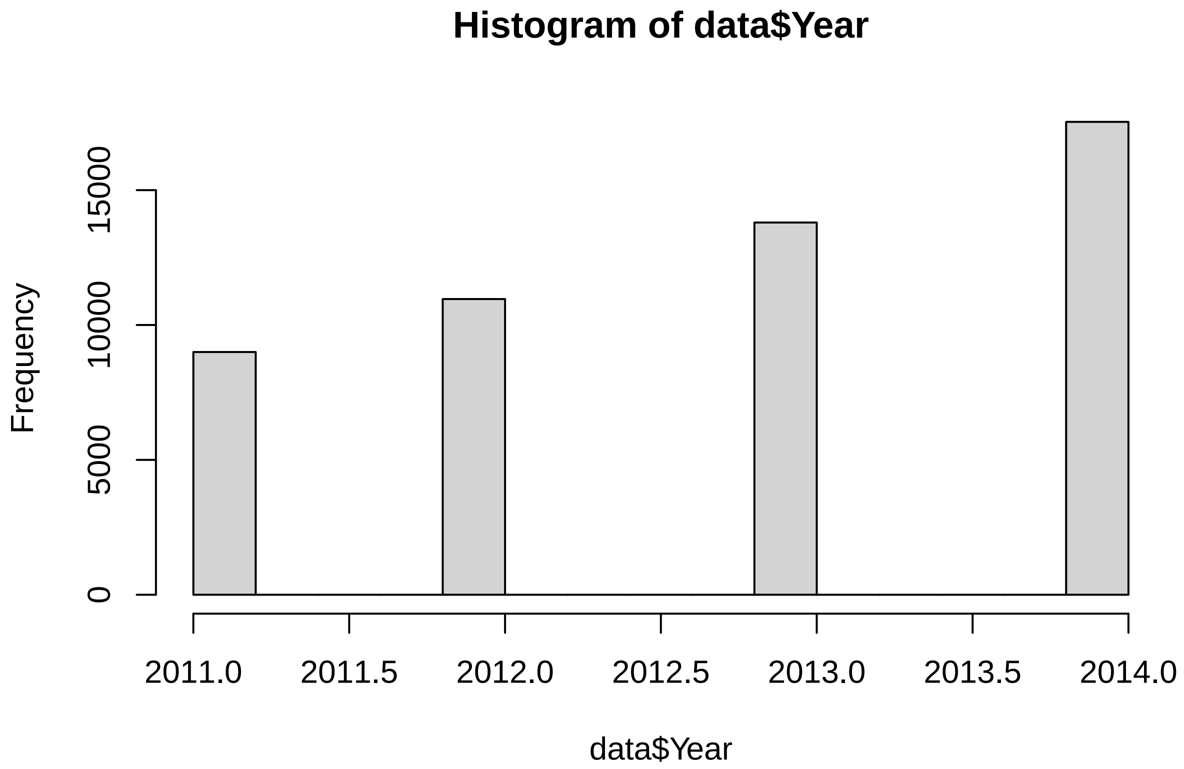
```r
#build the histogram
data = as.data.frame(sapply(superstore, as.numeric))
```

```
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
```

```r
hist(data$Sales)
```

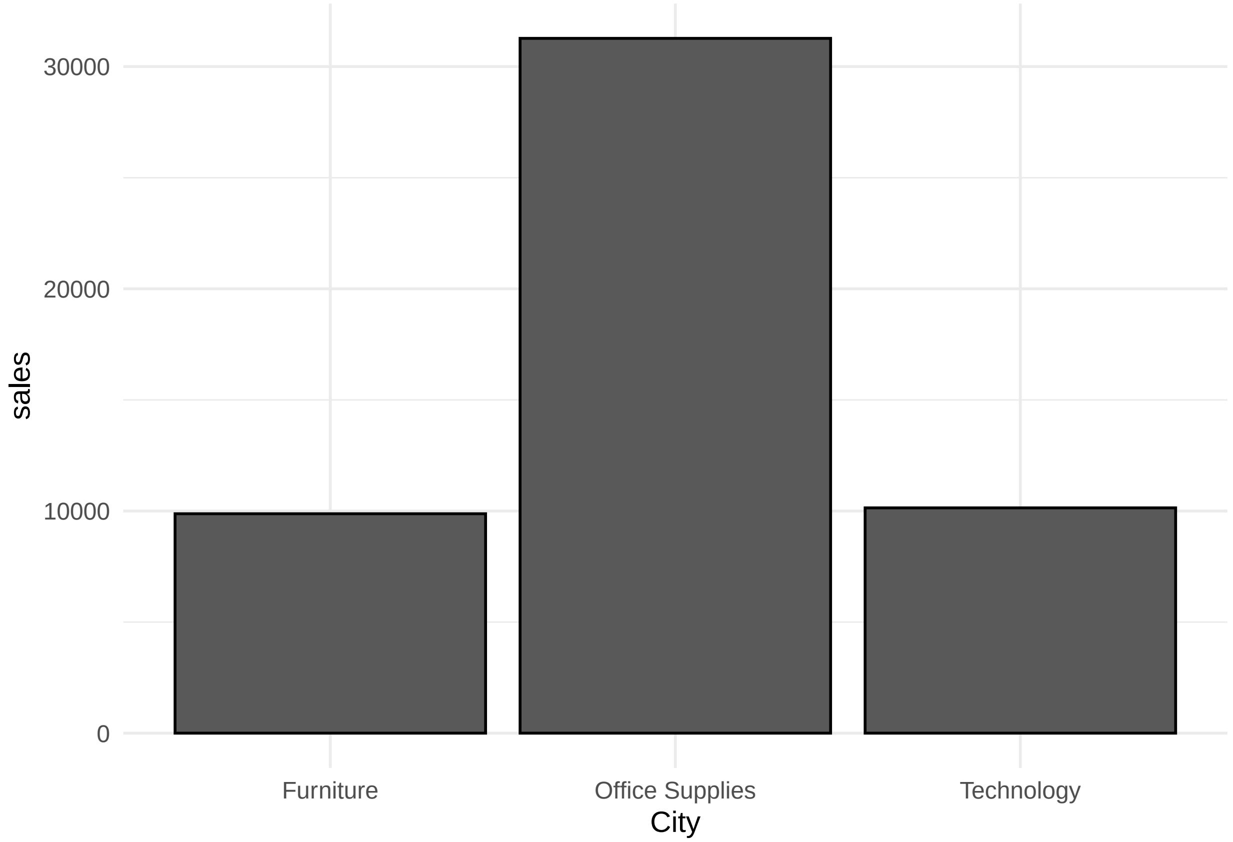


```r
hist(data$Year)

library(ggplot2)
```



```
ggplot(superstore, aes(x = superstore$Category, fill = Sales)) +
  geom_bar(position = "dodge", stat = "count", color = "black") +
  labs(x = "City", y = "sales") +
  theme_minimal()
```

```
## Warning: Use of ‘superstore$Category‘ is discouraged.
## i Use ‘Category‘ instead.

## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



```
#finding the correlations
numerical_columns = sapply(superstore, is.numeric)
numerical_data = superstore[, numerical_columns]
correlation = cor(numerical_data)

#finding the chi square
contingency_table1 = table(superstore$Category, superstore$Sales)
chi_square1 = chisq.test(contingency_table1)
```

```
## Warning in chisq.test(contingency_table1): Chi-squared approximation may be
## incorrect
```

```
contingency_table2 = table(superstore$Sales, superstore$City)
chi_square2 = chisq.test(contingency_table2)
```

```
## Warning in chisq.test(contingency_table2): Chi-squared approximation may be
## incorrect
```

```
contingency_table3 = table(superstore$Customer.Name, superstore$Sales)
chi_square3 = chisq.test(contingency_table3)
```

```
## Warning in chisq.test(contingency_table3): Chi-squared approximation may be
## incorrect
```

```
contingency_table4 = table(superstore$Country, superstore$Sales)
chi_square4 = chisq.test(contingency_table4)
```

```
## Warning in chisq.test(contingency_table4): Chi-squared approximation may be
## incorrect
```

```
contingency_table5 = table(superstore$Category, superstore$Sales)
chi_square5 = chisq.test(contingency_table5)
```

```
## Warning in chisq.test(contingency_table5): Chi-squared approximation may be
## incorrect
```

```
#linearmodel
superstore$Sales = as.factor(superstore$Sales)
str(superstore)
```

```
## tibble [51,290 x 26] (S3: tbl_df/tbl/data.frame)
##  $ Category      : chr [1:51290] "Office Supplies" "Office Supplies" "Office Supplies" "Offic
##  $ City          : chr [1:51290] "Los Angeles" "Los Angeles" "Los Angeles" "Los Angeles" ...
##  $ Country       : chr [1:51290] "United States" "United States" "United States" "United Stat
##  $ Customer.ID   : chr [1:51290] "LS-172304" "MV-174854" "CS-121304" "CS-121304" ...
##  $ Customer.Name : chr [1:51290] "Lycoris Saunders" "Mark Van Huff" "Chad Sievert" "Chad Siev
##  $ Discount      : num [1:51290] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Market        : chr [1:51290] "US" "US" "US" "US" ...
##  $ Order.Date    : 'hms' num [1:51290] 00:00:00 00:00:00 00:00:00 00:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ Order.ID      : chr [1:51290] "CA-2011-130813" "CA-2011-148614" "CA-2011-118962" "CA-2011-
##  $ Order.Priority: chr [1:51290] "High" "Medium" "Medium" "Medium" ...
##  $ Product.ID    : chr [1:51290] "OFF-PA-10002005" "OFF-PA-10002893" "OFF-PA-10000659" "OFF-P
##  $ Product.Name  : chr [1:51290] "Xerox 225" "Wirebound Service Call Books, 5 1/2\" x 4\"" "A
##  $ Profit        : num [1:51290] 9.33 9.29 9.84 53.26 3.11 ...
##  $ Quantity      : num [1:51290] 3 2 3 2 1 3 3 2 9 4 ...
##  $ Region        : chr [1:51290] "West" "West" "West" "West" ...
##  $ Row.ID        : num [1:51290] 36624 37033 31468 31469 32440 ...
##  $ Sales         : Factor w/ 2246 levels "0","1","2","3",..: 20 20 22 112 7 14 20 13 55 50 ..
##  $ Segment       : chr [1:51290] "Consumer" "Consumer" "Consumer" "Consumer" ...
##  $ Ship.Date     : 'hms' num [1:51290] 00:00:00 00:00:00 00:00:00 00:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ Ship.Mode     : chr [1:51290] "Second Class" "Standard Class" "Standard Class" "Standard C
##  $ Shipping.Cost : num [1:51290] 4.37 0.94 1.81 4.59 1.32 2.39 1.15 0.92 6.98 1.99 ...
##  $ State         : chr [1:51290] "California" "California" "California" "California" ...
##  $ Sub.Category  : chr [1:51290] "Paper" "Paper" "Paper" "Paper" ...
##  $ Year          : num [1:51290] 2011 2011 2011 2011 2011 ...
##  $ Market2       : chr [1:51290] "North America" "North America" "North America" "North Ameri
##  $ weeknum       : num [1:51290] 2 4 32 32 40 43 45 46 48 50 ...
```

```
superstore_subset = superstore[sample(nrow(superstore), 10000), ]

# Fit the model on the subset


index=sample(2,nrow(superstore_subset),replace = TRUE,prob =c(.80,.20))
train_set = superstore_subset[index == 1,]
test_set = superstore_subset[index== 2, ]




train_set
```

```
## # A tibble: 8,001 x 26
##    Category   City  Country Customer.ID Customer.Name Discount Market Order.Date
##    <chr>      <chr> <chr>   <chr>       <chr>            <dbl> <chr>  <time>
##  1 Technology Melb~ Austra~ MJ-177401   Max Jones          0.1 APAC   00’00"
##  2 Furniture  Derby United~ PM-191352   Peter McVee        0   EU     00’00"
##  3 Technology Murc~ Spain   EL-137352   Ed Ludwig          0   EU     00’00"
##  4 Office Su~ Dover United~ EP-139154   Emily Phan         0   US     00’00"
##  5 Furniture  Manz~ Cuba    RA-192853   Ralph Arnett       0   LATAM  00’00"
##  6 Office Su~ Trier Germany BP-110952   Bart Pistole       0.1 EU     00’00"
##  7 Furniture  Midd~ United~ AA-104804   Andrew Allen       0   US     00’00"
##  8 Office Su~ Canb~ Austra~ TZ-214451   Tom Zandusky       0.4 APAC   00’00"
##  9 Furniture  Dewas India   MM-179201   Michael Moore      0   APAC   00’00"
## 10 Office Su~ Karb~ Iraq    TS-110853   Thais Sissman      0   EMEA   00’00"
## # i 7,991 more rows
## # i 18 more variables: Order.ID <chr>, Order.Priority <chr>, Product.ID <chr>,
## #   Product.Name <chr>, Profit <dbl>, Quantity <dbl>, Region <chr>,
## #   Row.ID <dbl>, Sales <fct>, Segment <chr>, Ship.Date <time>,
## #   Ship.Mode <chr>, Shipping.Cost <dbl>, State <chr>, Sub.Category <chr>,
## #   Year <dbl>, Market2 <chr>, weeknum <dbl>
```

```
test_set
```

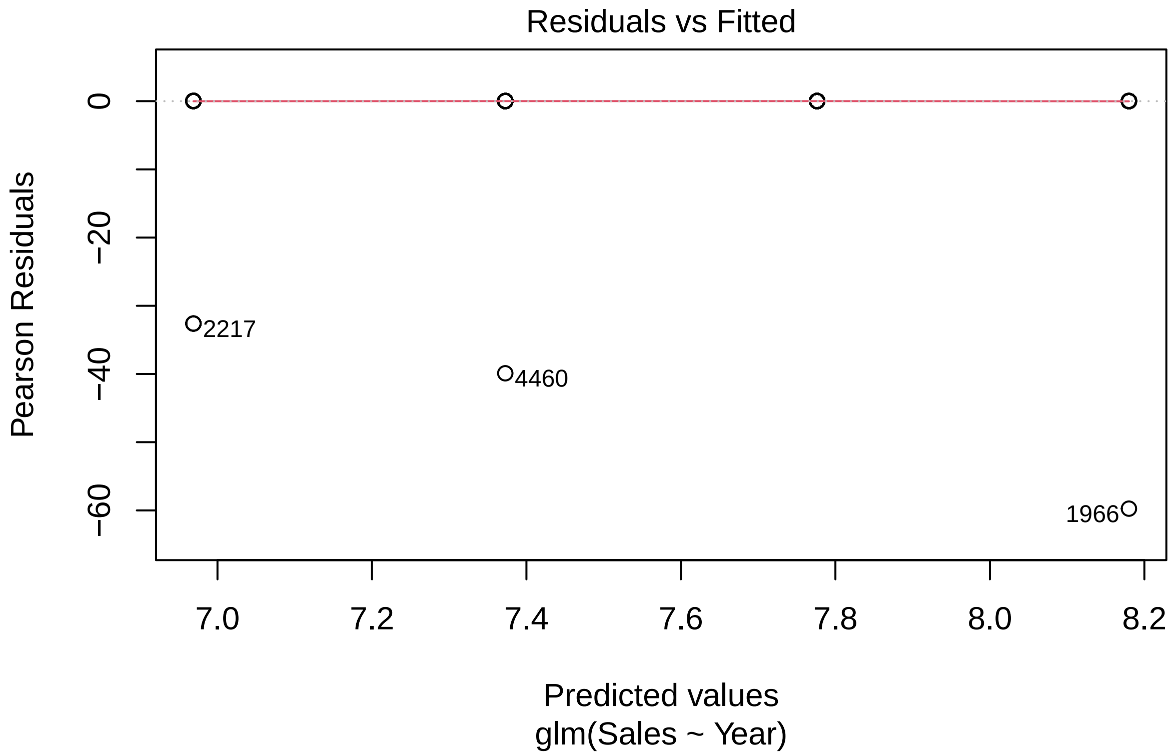
```
## # A tibble: 1,999 x 26
##    Category   City  Country Customer.ID Customer.Name Discount Market Order.Date
##    <chr>      <chr> <chr>   <chr>       <chr>            <dbl> <chr>  <time>
##  1 Office Su~ Plan~ United~ AH-106904   Anna Häberlin      0.7 US     00’00"
##  2 Office Su~ Bras~ Brazil  BG-116953   Brooke Gilli~      0   LATAM  00’00"
##  3 Office Su~ Phil~ United~ CB-124154   Christy Brit~      0.7 US     00’00"
##  4 Furniture  Sant~ Domini~ CK-123253   Christine Ka~      0.2 LATAM  00’00"
##  5 Office Su~ Sain~ France  JB-160002   Joy Bell-          0   EU     00’00"
##  6 Furniture  Cieg~ Cuba    RD-199303   Russell D’As~      0   LATAM  00’00"
##  7 Technology Chap~ United~ NF-183852   Natalie Frit~      0   EU     00’00"
##  8 Office Su~ Dort~ Germany CK-125952   Clytie Kelty       0   EU     00’00"
##  9 Office Su~ Sinc~ Turkey  CA-19652    Carol Adams        0.6 EMEA   00’00"
## 10 Office Su~ Phil~ United~ HD-147854   Harold Dahlen      0.7 US     00’00"
## # i 1,989 more rows
## # i 18 more variables: Order.ID <chr>, Order.Priority <chr>, Product.ID <chr>,
## #   Product.Name <chr>, Profit <dbl>, Quantity <dbl>, Region <chr>,
## #   Row.ID <dbl>, Sales <fct>, Segment <chr>, Ship.Date <time>,
## #   Ship.Mode <chr>, Shipping.Cost <dbl>, State <chr>, Sub.Category <chr>,
## #   Year <dbl>, Market2 <chr>, weeknum <dbl>
```

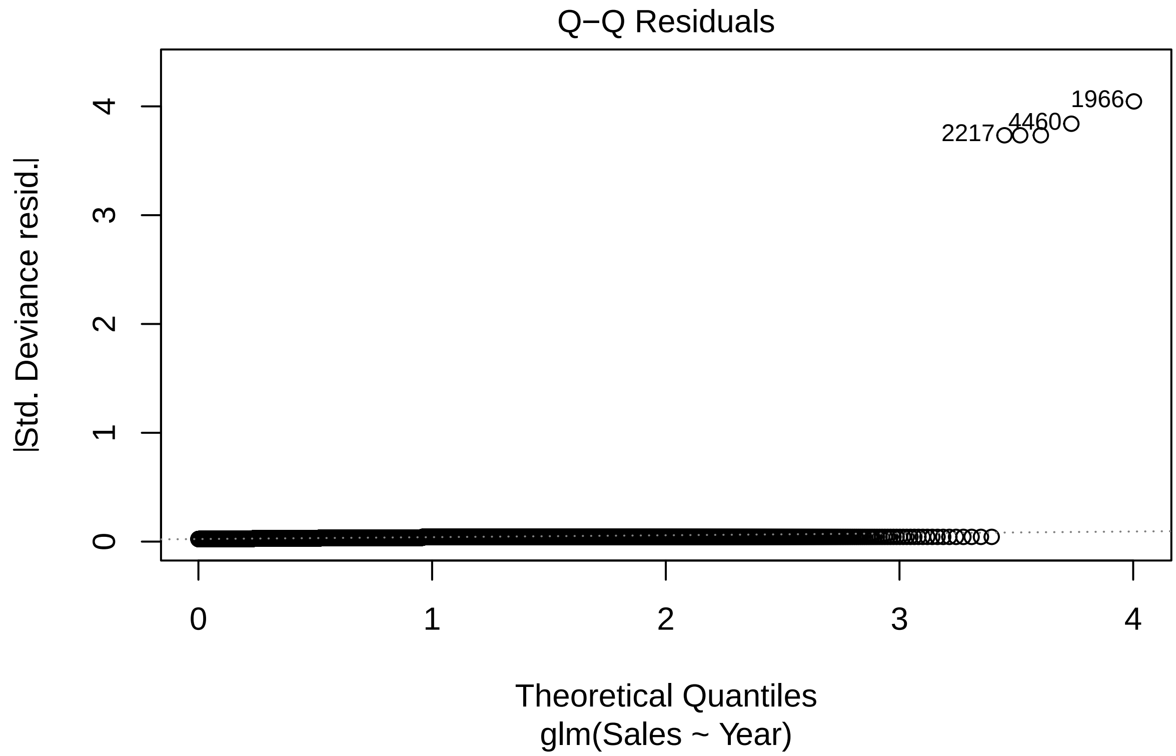
```r
library(ROCR)
library(caTools)
```

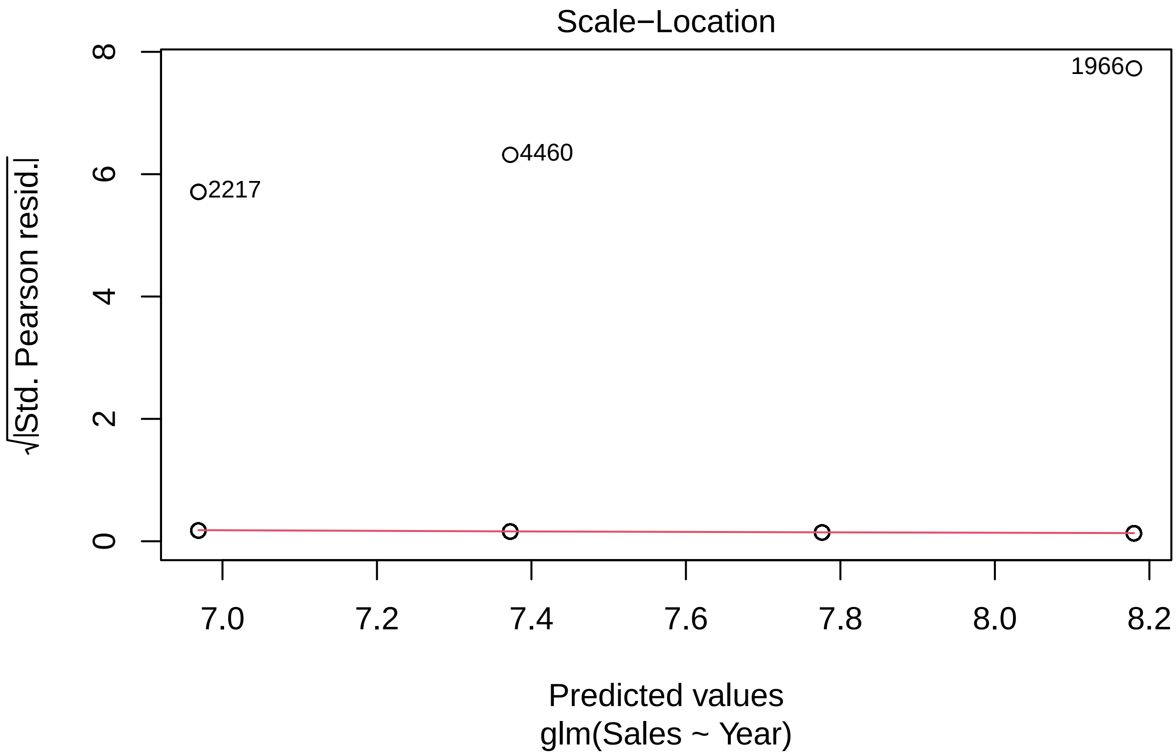
```r
# Fit logistic regression model
log_reg1 = glm(Sales ~ Year,  data = train_set, family = "binomial")


summary(log_reg1)
```

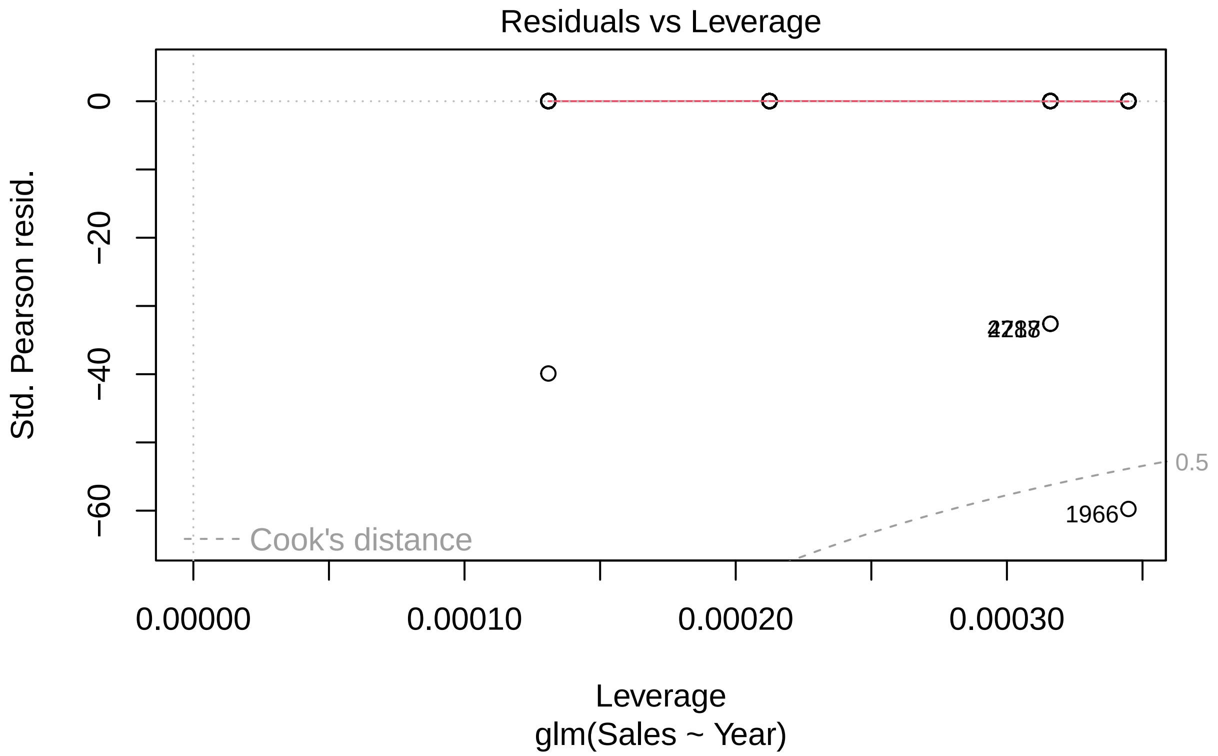
```
##
## Call:
## glm(formula = Sales ~ Year, family = "binomial", data = train_set)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 820.0658   929.2274   0.883    0.377
## Year         -0.4037     0.4616  -0.875    0.382
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.776  on 8000  degrees of freedom
## Residual deviance: 82.922  on 7999  degrees of freedom
## AIC: 86.922
##
## Number of Fisher Scoring iterations: 10
```

```r
plot(log_reg1, col = "black")
```


Residuals vs Fitted
Pearson Residuals
0
−20
−40
−60
7.0
7.2
7.4
7.6
7.8
8.0
8.2
2217
4460
1966
Predicted values
glm(Sales ~ Year)


Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
1966
4460
2217
0
1
2
3
4
Theoretical Quantiles
glm(Sales ~ Year)


Scale−Location
√|Std. Pearson resid.|
0
2
4
6
8
2217
4460
1966
7.0
7.2
7.4
7.6
7.8
8.0
8.2
Predicted values
glm(Sales ~ Year)



```
# Make predictions on the test set
predict_reg = predict(log_reg1, newdata = test_set, type = "response")

# Threshold for classification
threshold = 0.5

# Convert predicted probabilities to factor with levels 0 and 1
predict_reg = factor(ifelse(predict_reg > threshold, "1", "0"), levels = c("0", "1"))

# Ensure that test_set$Sales is a factor with the same levels
test_set$Sales= factor(test_set$Sales, levels = levels(predict_reg))

# Compare predicted values with actual values
comparison = test_set$Sales == predict_reg

# Ensure that comparison is a logical vector
comparison = as.logical(comparison)

# Create a confusion matrix
conf_matrix = caret::confusionMatrix(data = predict_reg, reference = test_set$Sales)
conf_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction 0 1
##          0 0 0
##          1 0 2
##
##                Accuracy : 1
##                  95% CI : (0.1581, 1)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : NaN
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : NA
##             Specificity :  1
##          Pos Pred Value : NA
##          Neg Pred Value : NA
##              Prevalence :  0
##          Detection Rate :  0
##    Detection Prevalence :  0
##       Balanced Accuracy : NA
##
##        'Positive' Class : 0
##
```

-

Present the data analysis results.

```
correlation
```

```
##                    Discount        Profit     Quantity        Row.ID
## Discount       1.0000000000 -0.3164901718 -0.019874695  0.0875941122
## Profit        -0.3164901718  1.0000000000  0.104365027 -0.0190370508
## Quantity      -0.0198746951  0.1043650272  1.000000000 -0.1734827801
## Row.ID         0.0875941122 -0.0190370508 -0.173482780  1.0000000000
## Sales         -0.0867280172  0.4849231008  0.313579983 -0.0438876324
## Shipping.Cost -0.0790554534  0.3544408272  0.272648732 -0.0390758510
## Year          -0.0058939358  0.0026262698 -0.005049263 -0.0009209292
## weeknum       -0.0002172958 -0.0001818997  0.020839354 -0.0287760217
##                      Sales Shipping.Cost          Year       weeknum
## Discount      -0.086728017  -0.079055453 -0.0058939358 -0.0002172958
## Profit         0.484923101   0.354440827  0.0026262698 -0.0001818997
## Quantity       0.313579983   0.272648732 -0.0050492628  0.0208393539
## Row.ID        -0.043887632  -0.039075851 -0.0009209292 -0.0287760217
## Sales          1.000000000   0.768075073 -0.0029015295  0.0019467553
## Shipping.Cost  0.768075073   1.000000000 -0.0031365502  0.0053466690
## Year          -0.002901529  -0.003136550  1.0000000000 -0.0196274593
## weeknum        0.001946755   0.005346669 -0.0196274593  1.0000000000
```

```
chi_square1
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  contingency_table1
## X-squared = 24871, df = 4490, p-value < 2.2e-16
```

```
chi_square2
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  contingency_table2
## X-squared = 8511860, df = 8160575, p-value < 2.2e-16
```

```
chi_square3
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  contingency_table3
## X-squared = 1780053, df = 1782530, p-value = 0.9053
```

```
chi_square4
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  contingency_table4
## X-squared = 324801, df = 327770, p-value = 0.9999
```

```
chi_square5
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table5
## X-squared = 24871, df = 4490, p-value < 2.2e-16
```

-

Interpret the results in a way to address the research questions.

The histograms give an early look at the sales distributions and the evolution of time, giving a general idea of their patterns. However, the logistic regression model becomes an essential tool to delve deeper into the dynamics of sales trends, particularly with regard to particular products, categories, and geographic areas. We can understand the complex relationship between time and sales and determine whether sales are trending upward or downward by fitting the model with the 'Year' variable. The model summary, which includes the 'Year' coefficient and its importance, gives us the ability to decipher the direction and strength of this time-related impact on sales.The resulting confusion matrix provides useful metrics like precision, recall, and F1 score and acts as a metre for evaluating how well the model classifies sales. It is generated from predictions on the test set.

Conclusion [15 points]

-

Does the analysis answer the research questions?

yes , it gave me the answers. The first research question is partially addressed by the time series analysis in R that is provided. It uses a moving average to smooth out trends and shows overall sales trends over time. Nevertheless, because it lacks particular analyses for identifying trends in products, categories, and geographic areas, it is unable to provide a comprehensive response to the question. Further analyses are needed to address the second research question about consumer behavior, which relates to recurring purchases or brand loyalty, and the third question about the effect of discounts on sales and profitability. A thorough understanding of the dataset and effective answers to all research questions require specific segmentation based on customer behavior, brand loyalty assessments, and in-depth discount impact analyses on products and customer segments.

-

Discuss the scope and generalizability of the analysis.

The scope of the analysis is limited to visualizing overall sales trends over the years and applying a moving average for trend smoothing. It provides a broad understanding of general sales patterns but lacks specificity in exploring trends related to specific products, categories, or geographical areas. The generalizability of the analysis is constrained by its focus on the provided dataset, and the insights derived may not be readily applicable to broader contexts. To enhance the analysis's scope and generalizability, more granular analyses, such as product-level trends and regional variations, along with the incorporation of advanced statistical or machine learning techniques, would be necessary to capture a more comprehensive and transferable understanding of sales dynamics.

-

Discuss potential limitations and possibilities for improvement.

There are various potential limitations to the analysis that should be taken into account. First of all, its granularity is limited as it primarily concentrates on overall sales trends without exploring particular product categories or geographic areas. Furthermore, the analysis may not be able to identify intricate patterns or generate predictions in the absence of statistical models or machine learning methods. Moving averages can be helpful for smoothing trends, but they can also oversimplify dynamics and miss small changes. Additionally, without addressing possible problems with data quality or outliers, the analysis makes the assumption that the dataset is impartial and representative. More sophisticated statistical techniques, in-depth product and regional analyses, and the investigation of predictive modelling could all be used to enhance the analysis.