Data Analysis using R

Group 10

2025-03-31

1. Load the Dataset

```
# Import dataset
df <- read.csv("D:/c.csv", stringsAsFactors = FALSE)

# Print dataset structure
str(df)
```

```
## 'data.frame':    1094 obs. of  6 variables:
##  $ Sales.Person : chr  "Jehu Rudeforth" "Van Tuxwell" "Gigi Bohling" "Jan Morforth" ...
##  $ Country      : chr  "UK" "India" "India" "Australia" ...
##  $ Product      : chr  "Mint Chip Choco" "85% Dark Bars" "Peanut Butter Cubes" "Peanut Butter Cubes"
##  $ Date         : chr  "04-Jan-22" "01-Aug-22" "07-Jul-22" "27-Apr-22" ...
##  $ Amount       : chr  "$5,320 " "$7,896 " "$4,501 " "$12,726 " ...
##  $ Boxes.Shipped: int  180 94 91 342 184 38 176 73 59 102 ...
```

2. List Variables in the Dataset

```
colnames(df)
```

```
## [1] "Sales.Person"  "Country"       "Product"       "Date"
## [5] "Amount"        "Boxes.Shipped"
```

3. Display the First 15 Rows

```
head(df, 15)
```

```
##           Sales.Person     Country             Product      Date   Amount
## 1       Jehu Rudeforth          UK     Mint Chip Choco 04-Jan-22  $5,320
## 2          Van Tuxwell       India       85% Dark Bars 01-Aug-22  $7,896
## 3         Gigi Bohling       India Peanut Butter Cubes 07-Jul-22  $4,501
## 4         Jan Morforth   Australia Peanut Butter Cubes 27-Apr-22 $12,726
## 5       Jehu Rudeforth          UK Peanut Butter Cubes 24-Feb-22 $13,685
## 6          Van Tuxwell       India  Smooth Sliky Salty 06-Jun-22  $5,376
## 7           Oby Sorrel          UK     99% Dark & Pure 25-Jan-22 $13,685
```

```
## 8      Gunar Cockshoot   Australia         After Nines 24-Mar-22  $3,080
## 9       Jehu Rudeforth New Zealand      50% Dark Bites 20-Apr-22  $3,990
## 10         Brien Boise   Australia     99% Dark & Pure 04-Jul-22  $2,835
## 11 Rafaelita Blaksland          UK  Smooth Sliky Salty 13-Jan-22  $4,704
## 12        Barr Faughny         USA        Orange Choco 10-Mar-22  $3,703
## 13      Mallorie Waber      Canada             Eclairs 13-Jan-22  $1,442
## 14    Karlen McCaffrey New Zealand       Drinking Coco 28-Jul-22    $168
## 15      Marney O'Breen New Zealand Peanut Butter Cubes 03-Aug-22  $8,379
##    Boxes.Shipped
## 1            180
## 2             94
## 3             91
## 4            342
## 5            184
## 6             38
## 7            176
## 8             73
## 9             59
## 10           102
## 11            62
## 12            11
## 13           286
## 14           156
## 15           173
```

4. Create a User-Defined Function

```
total_sales <- function(x) {
  total_sales_made <- sum(x, na.rm = TRUE)
  return(total_sales_made)
}

df$Amount_clean <- gsub("[\\$,]", "", df$Amount, perl = TRUE)
df$Amount_numeric <- as.numeric(df$Amount_clean)
y <- total_sales(df$Amount_numeric)
print(y)
```

```
## [1] 6183625
```

5. Remove Missing Values

```
df <- df %>% drop_na()
print("Data after removing missing values:")
```

```
## [1] "Data after removing missing values:"
```

```
head(df, 10)
```

```
##       Sales.Person     Country             Product      Date   Amount
## 1   Jehu Rudeforth          UK     Mint Chip Choco 04-Jan-22  $5,320
## 2      Van Tuxwell       India       85% Dark Bars 01-Aug-22  $7,896
## 3     Gigi Bohling       India Peanut Butter Cubes 07-Jul-22  $4,501
## 4     Jan Morforth   Australia Peanut Butter Cubes 27-Apr-22 $12,726
## 5   Jehu Rudeforth          UK Peanut Butter Cubes 24-Feb-22 $13,685
## 6      Van Tuxwell       India  Smooth Sliky Salty 06-Jun-22  $5,376
## 7       Oby Sorrel          UK     99% Dark & Pure 25-Jan-22 $13,685
## 8  Gunar Cockshoot   Australia         After Nines 24-Mar-22  $3,080
## 9   Jehu Rudeforth New Zealand      50% Dark Bites 20-Apr-22  $3,990
## 10     Brien Boise   Australia     99% Dark & Pure 04-Jul-22  $2,835
##    Boxes.Shipped Amount_clean Amount_numeric
## 1            180        5320            5320
## 2             94        7896            7896
## 3             91        4501            4501
## 4            342       12726           12726
## 5            184       13685           13685
## 6             38        5376            5376
## 7            176       13685           13685
## 8             73        3080            3080
## 9             59        3990            3990
## 10           102        2835            2835
```

6. Remove Duplicates

```
df <- df %>% distinct()
print("Data after removing duplicates:")
```

```
## [1] "Data after removing duplicates:"
```

```
head(df, 10)
```

```
##       Sales.Person     Country             Product      Date   Amount
## 1   Jehu Rudeforth          UK     Mint Chip Choco 04-Jan-22  $5,320
## 2      Van Tuxwell       India       85% Dark Bars 01-Aug-22  $7,896
## 3     Gigi Bohling       India Peanut Butter Cubes 07-Jul-22  $4,501
## 4     Jan Morforth   Australia Peanut Butter Cubes 27-Apr-22 $12,726
## 5   Jehu Rudeforth          UK Peanut Butter Cubes 24-Feb-22 $13,685
## 6      Van Tuxwell       India  Smooth Sliky Salty 06-Jun-22  $5,376
## 7       Oby Sorrel          UK     99% Dark & Pure 25-Jan-22 $13,685
## 8  Gunar Cockshoot   Australia         After Nines 24-Mar-22  $3,080
## 9   Jehu Rudeforth New Zealand      50% Dark Bites 20-Apr-22  $3,990
## 10     Brien Boise   Australia     99% Dark & Pure 04-Jul-22  $2,835
##    Boxes.Shipped Amount_clean Amount_numeric
## 1            180        5320            5320
## 2             94        7896            7896
## 3             91        4501            4501
## 4            342       12726           12726
## 5            184       13685           13685
## 6             38        5376            5376
## 7            176       13685           13685
```

```
## 8             73        3080            3080
## 9             59        3990            3990
## 10           102        2835            2835
```

7. Filter Data (Amount > 5000)

```
df_filtered <- df %>% filter(Amount_numeric > 5000)
print("Filtered data (Amount > 5000):")
```

```
## [1] "Filtered data (Amount > 5000):"
```

```
head(df_filtered, 10)
```

```
##      Sales.Person     Country             Product      Date   Amount
## 1  Jehu Rudeforth          UK     Mint Chip Choco 04-Jan-22  $5,320
## 2     Van Tuxwell       India       85% Dark Bars 01-Aug-22  $7,896
## 3    Jan Morforth   Australia Peanut Butter Cubes 27-Apr-22 $12,726
## 4  Jehu Rudeforth          UK Peanut Butter Cubes 24-Feb-22 $13,685
## 5     Van Tuxwell       India  Smooth Sliky Salty 06-Jun-22  $5,376
## 6      Oby Sorrel          UK     99% Dark & Pure 25-Jan-22 $13,685
## 7  Marney O'Breen New Zealand Peanut Butter Cubes 03-Aug-22  $8,379
## 8  Beverie Moffet   Australia Organic Choco Syrup 26-Jan-22  $6,790
## 9  Beverie Moffet      Canada           Milk Bars 16-Feb-22  $8,799
## 10    Brien Boise   Australia             Eclairs 27-Jun-22  $6,888
##    Boxes.Shipped Amount_clean Amount_numeric
## 1            180        5320            5320
## 2             94        7896            7896
## 3            342       12726           12726
## 4            184       13685           13685
## 5             38        5376            5376
## 6            176       13685           13685
## 7            173        8379            8379
## 8            356        6790            6790
## 9            250        8799            8799
## 10            88        6888            6888
```

8. Sort Data by Amount (Descending Order)

```
df <- df %>% arrange(desc(Amount_numeric))
print("Sorted Data:")
```

```
## [1] "Sorted Data:"
```

```
head(df, 10)
```

```
##           Sales.Person     Country             Product      Date   Amount
## 1         Ches Bonnell       India Peanut Butter Cubes 27-Jan-22 $22,050
## 2          Van Tuxwell       India Organic Choco Syrup 16-May-22 $19,929
```

```
## 3  Rafaelita Blaksland New Zealand             Eclairs 07-Feb-22 $19,481
## 4           Van Tuxwell   Australia Organic Choco Syrup 10-Aug-22 $19,453
## 5        Curtice Advani       India  Smooth Sliky Salty 19-Apr-22 $19,327
## 6        Marney O'Breen          UK  Smooth Sliky Salty 13-May-22 $18,991
## 7           Kaine Padly          UK         After Nines 21-Jan-22 $18,697
## 8          Jan Morforth New Zealand     Mint Chip Choco 30-Jun-22 $18,340
## 9           Brien Boise       India       85% Dark Bars 09-Aug-22 $18,032
## 10         Jan Morforth   Australia     Mint Chip Choco 22-Feb-22 $17,626
##    Boxes.Shipped Amount_clean Amount_numeric
## 1            208       22050           22050
## 2            174       19929           19929
## 3             51       19481           19481
## 4             14       19453           19453
## 5            135       19327           19327
## 6             88       18991           18991
## 7            176       18697           18697
## 8            285       18340           18340
## 9            205       18032           18032
## 10           103       17626           17626
```

9. Rename Columns

```
df <- df %>% rename(Salesperson = Sales.Person, Shipment_Count = Boxes.Shipped)
print("Renamed Columns:")
```

```
## [1] "Renamed Columns:"
```

```
head(df, 10)
```

```
##            Salesperson     Country             Product      Date   Amount
## 1         Ches Bonnell       India Peanut Butter Cubes 27-Jan-22 $22,050
## 2          Van Tuxwell       India Organic Choco Syrup 16-May-22 $19,929
## 3  Rafaelita Blaksland New Zealand             Eclairs 07-Feb-22 $19,481
## 4          Van Tuxwell   Australia Organic Choco Syrup 10-Aug-22 $19,453
## 5       Curtice Advani       India  Smooth Sliky Salty 19-Apr-22 $19,327
## 6       Marney O'Breen          UK  Smooth Sliky Salty 13-May-22 $18,991
## 7          Kaine Padly          UK         After Nines 21-Jan-22 $18,697
## 8         Jan Morforth New Zealand     Mint Chip Choco 30-Jun-22 $18,340
## 9          Brien Boise       India       85% Dark Bars 09-Aug-22 $18,032
## 10        Jan Morforth   Australia     Mint Chip Choco 22-Feb-22 $17,626
##    Shipment_Count Amount_clean Amount_numeric
## 1             208       22050           22050
## 2             174       19929           19929
## 3              51       19481           19481
## 4              14       19453           19453
## 5             135       19327           19327
## 6              88       18991           18991
## 7             176       18697           18697
## 8             285       18340           18340
## 9             205       18032           18032
## 10            103       17626           17626
```

10. Add a New Column

```
df <- df %>% mutate(Double_Amount = Amount_numeric * 2)
print("New Column Added:")
```

```
## [1] "New Column Added:"
```

```
head(df, 10)
```

```
##            Salesperson     Country             Product      Date   Amount
## 1         Ches Bonnell       India Peanut Butter Cubes 27-Jan-22 $22,050
## 2          Van Tuxwell       India Organic Choco Syrup 16-May-22 $19,929
## 3  Rafaelita Blaksland New Zealand             Eclairs 07-Feb-22 $19,481
## 4          Van Tuxwell   Australia Organic Choco Syrup 10-Aug-22 $19,453
## 5       Curtice Advani       India  Smooth Sliky Salty 19-Apr-22 $19,327
## 6       Marney O'Breen          UK  Smooth Sliky Salty 13-May-22 $18,991
## 7          Kaine Padly          UK         After Nines 21-Jan-22 $18,697
## 8         Jan Morforth New Zealand     Mint Chip Choco 30-Jun-22 $18,340
## 9          Brien Boise       India       85% Dark Bars 09-Aug-22 $18,032
## 10        Jan Morforth   Australia     Mint Chip Choco 22-Feb-22 $17,626
##    Shipment_Count Amount_clean Amount_numeric Double_Amount
## 1             208       22050           22050         44100
## 2             174       19929           19929         39858
## 3              51       19481           19481         38962
## 4              14       19453           19453         38906
## 5             135       19327           19327         38654
## 6              88       18991           18991         37982
## 7             176       18697           18697         37394
## 8             285       18340           18340         36680
## 9             205       18032           18032         36064
## 10            103       17626           17626         35252
```

11. Create a Training Set

```
set.seed(123)
train_index <- sample(1:nrow(df), 0.7 * nrow(df))
train_set <- df[train_index, ]
test_set <- df[-train_index, ]

print("Training Set (First 10 rows):")
```

```
## [1] "Training Set (First 10 rows):"
```

```
head(train_set, 10)
```

```
##          Salesperson     Country             Product      Date  Amount
## 415   Camilla Castle   Australia      Raspberry Choco 23-Aug-22 $6,342
## 463  Gunar Cockshoot          UK Choco Coated Almonds 30-Jun-22 $5,775
```

```
## 179   Camilla Castle         USA           White Choc 18-Aug-22 $9,681
## 526      Brien Boise   Australia Choco Coated Almonds 25-May-22 $5,124
## 195  Roddy Speechley      Canada   Smooth Sliky Salty 07-Mar-22 $9,338
## 938    Kelci Walkden         USA         Orange Choco 03-Feb-22 $1,379
## 1038    Husein Augar New Zealand Caramel Stuffed Bars 27-Jan-22   $497
## 665   Camilla Castle   Australia        85% Dark Bars 19-May-22 $3,654
## 602   Marney O'Breen       India Caramel Stuffed Bars 16-Mar-22 $4,361
## 709   Andria Kimpton New Zealand       50% Dark Bites 02-Mar-22 $3,374
##      Shipment_Count Amount_clean Amount_numeric Double_Amount
## 415             178        6342            6342         12684
## 463             135        5775            5775         11550
## 179              24        9681            9681         19362
## 526              62        5124            5124         10248
## 195              11        9338            9338         18676
## 938             138        1379            1379          2758
## 1038            475         497             497           994
## 665              14        3654            3654          7308
## 602              81        4361            4361          8722
## 709             202        3374            3374          6748
```

```
print("Test Set (First 10 rows):")
```

```
## [1] "Test Set (First 10 rows):"
```

```
head(test_set, 10)
```

```
##            Salesperson     Country              Product      Date   Amount
## 1         Ches Bonnell       India  Peanut Butter Cubes 27-Jan-22 $22,050
## 3  Rafaelita Blaksland New Zealand              Eclairs 07-Feb-22 $19,481
## 7          Kaine Padly          UK          After Nines 21-Jan-22 $18,697
## 9          Brien Boise       India        85% Dark Bars 09-Aug-22 $18,032
## 12       Kelci Walkden         USA   Manuka Honey Choco 16-Feb-22 $17,318
## 14         Brien Boise      Canada      99% Dark & Pure 18-May-22 $16,793
## 15       Kelci Walkden      Canada          After Nines 13-Jan-22 $16,702
## 22       Kelci Walkden New Zealand        Drinking Coco 10-Mar-22 $15,855
## 25        Ches Bonnell      Canada Choco Coated Almonds 24-Aug-22 $15,547
## 27 Rafaelita Blaksland       India      Mint Chip Choco 26-Jan-22 $15,491
##    Shipment_Count Amount_clean Amount_numeric Double_Amount
## 1             208       22050           22050         44100
## 3              51       19481           19481         38962
## 7             176       18697           18697         37394
## 9             205       18032           18032         36064
## 12             87       17318           17318         34636
## 14            416       16793           16793         33586
## 15            198       16702           16702         33404
## 22            111       15855           15855         31710
## 25            269       15547           15547         31094
## 27             85       15491           15491         30982
```

12. Summary Statistics

```r
summary(df)
```

```
##  Salesperson          Country             Product              Date
##  Length:1094        Length:1094        Length:1094        Length:1094
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Amount          Shipment_Count  Amount_clean       Amount_numeric
##  Length:1094        Min.   :  1.0   Length:1094        Min.   :    7
##  Class :character   1st Qu.: 70.0   Class :character   1st Qu.: 2390
##  Mode  :character   Median :135.0   Mode  :character   Median : 4868
##                     Mean   :161.8                      Mean   : 5652
##                     3rd Qu.:228.8                      3rd Qu.: 8027
##                     Max.   :709.0                      Max.   :22050
##  Double_Amount
##  Min.   :   14
##  1st Qu.: 4781
##  Median : 9737
##  Mean   :11305
##  3rd Qu.:16054
##  Max.   :44100
```

13. Statistical Calculations

```r
mean_value <- mean(df$Amount_numeric, na.rm = TRUE)
median_value <- median(df$Amount_numeric, na.rm = TRUE)

mode_func <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
mode_value <- mode_func(df$Amount_numeric)
range_value <- range(df$Amount_numeric, na.rm = TRUE)

print(mean_value)
```

```
## [1] 5652.308
```

```r
print(median_value)
```

```
## [1] 4868.5
```

```r
print(mode_value)
```

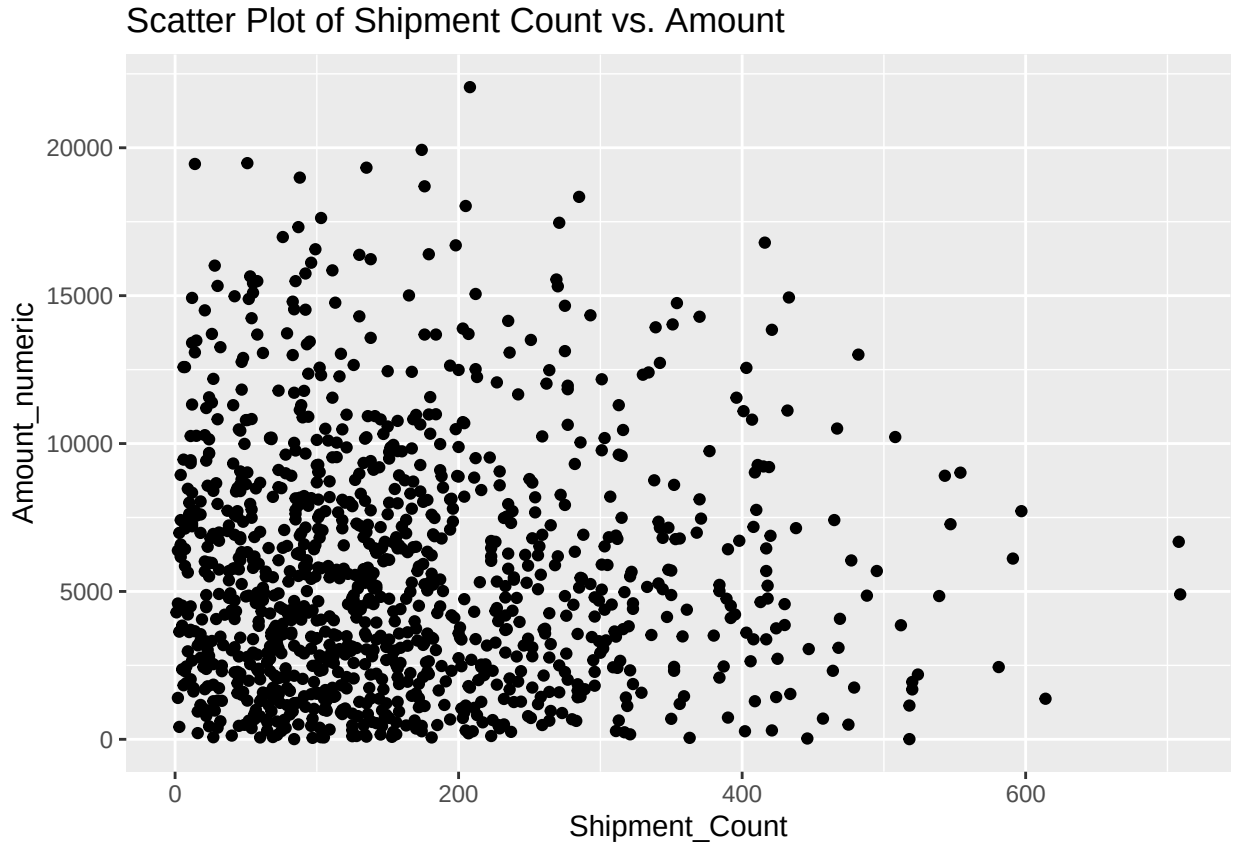
```
## [1] 2317
```

```r
print(range_value)
```

```
## [1]     7 22050
```

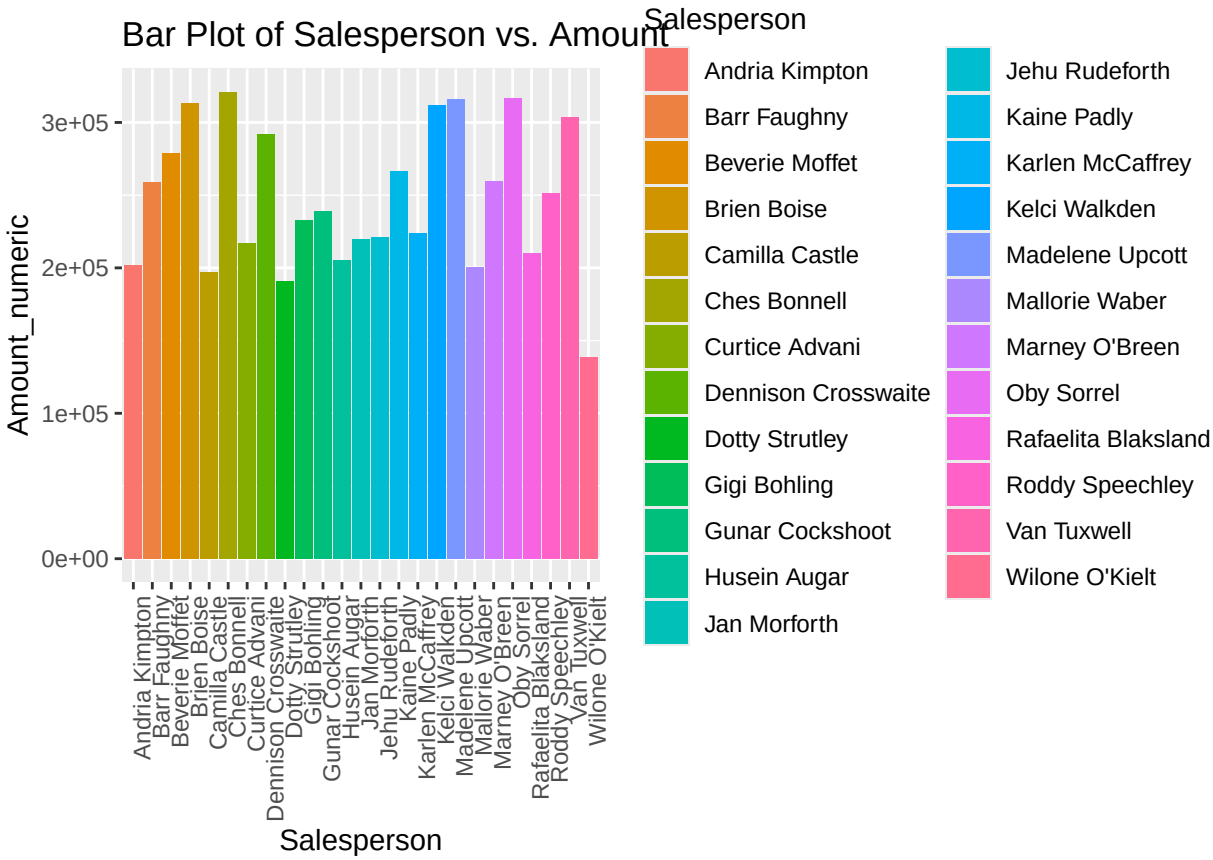
14. Scatter Plot

```r
ggplot(df, aes(x = Shipment_Count, y = Amount_numeric)) +
  geom_point() +
  ggtitle("Scatter Plot of Shipment Count vs. Amount")
```



15. Bar Plot

```r
ggplot(df, aes(x = Salesperson, y = Amount_numeric, fill = Salesperson)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Bar Plot of Salesperson vs. Amount")
```



16. Correlation Calculation

```
correlation <- cor(df$Amount_numeric, df$Shipment_Count, use = "complete.obs", method = "pearson")
print(correlation)
```

```
## [1] -0.01882685
```

GitHub Repository

You can find the full code and dataset on our GitHub Repository, CLICK HERE!.