Day 1 – Practical Session

Course Code: DSA0612

Course: Data Handling and Visualization

1. Visualize Scatter plot for given dataset.

|  |  |
| --- | --- |
| Weight (kg) | Height (cm) |
| 60 | 170 |
| 65 | 175 |
| 70 | 168 |
| 72 | 180 |
| 75 | 178 |

Sample Code:

# Load necessary library

library(ggplot2)

# Create a data frame

weight\_height\_data <- data.frame(

Weight = c(60, 65, 70, 72, 75),

Height = c(170, 175, 168, 180, 178)

)

# Plot scatter plot

ggplot(weight\_height\_data, aes(x = Weight, y = Height)) +

geom\_point(color = "blue", size = 3) +

labs(title = "Weight vs. Height",

x = "Weight (kg)",

y = "Height (cm)") +

theme\_minimal()

2. Visualize a Line Plot for the Given Population Data

|  |  |
| --- | --- |
| Year | Population (millions) |
| 2010 | 1000 |
| 2011 | 1020 |
| 2012 | 1040 |
| 2013 | 1060 |
| 2014 | 1080 |

# Create a data frame

year\_population\_data <- data.frame(

Year = 2010:2014,

Population = c(1000, 1020, 1040, 1060, 1080)

)

# Plot line plot

ggplot(year\_population\_data, aes(x = Year, y = Population)) +

geom\_line(color = "green", size = 1) +

geom\_point(color = "red", size = 3) +

labs(title = "Year vs. Population",

x = "Year",

y = "Population (millions)") +

theme\_minimal()

3. Bar Plot

Dataset: Sales by Product

|  |  |
| --- | --- |
| Product | Sales |
| A | 300 |
| B | 450 |
| C | 500 |
| D | 350 |
| E | 400 |

# Create a data frame

sales\_data <- data.frame(

Product = c("A", "B", "C", "D", "E"),

Sales = c(300, 450, 500, 350, 400)

)

# Plot bar plot

ggplot(sales\_data, aes(x = Product, y = Sales, fill = Product)) +

geom\_bar(stat = "identity") +

labs(title = "Sales by Product",

x = "Product",

y = "Sales") +

theme\_minimal()

4. Histogram

Dataset: Age Distribution

|  |  |
| --- | --- |
| | Age | | --- | |
| |  | | --- | | 25 | |
| |  | | --- | | 30 | |
| |  | | --- | | 35 | |
| |  | | --- | | 40 | |
| |  | | --- | | 45 | |
| |  | | --- | | 50 | |
| |  | | --- | | 55 | |
| |  | | --- | | 60 | |
| |  | | --- | | 65 | |
| |  | | --- | | 70 | |

# Create a data frame

age\_data <- data.frame(

Age = c(25, 30, 35, 40, 45, 50, 55, 60, 65, 70)

)

# Plot histogram

ggplot(age\_data, aes(x = Age)) +

geom\_histogram(binwidth = 5, fill = "purple", color = "black", alpha = 0.7) +

labs(title = "Age Distribution",

x = "Age",

y = "Frequency") +

theme\_minimal()

5. Pie Chart

Dataset: Market Share by Company

|  |  |
| --- | --- |
| Company | Market Share |
| A | 20% |
| B | 30% |
| C | 25% |
| D | 25% |

# Create a data frame

market\_share\_data <- data.frame(

Company = c("A", "B", "C", "D"),

Market\_Share = c(20, 30, 25, 25)

)

# Plot pie chart

ggplot(market\_share\_data, aes(x = "", y = Market\_Share, fill = Company)) +

geom\_bar(stat = "identity", width = 1) +

coord\_polar("y") +

labs(title = "Market Share by Company",

x = NULL,

y = NULL) +

theme\_void() +

theme(legend.position = "right")

6. Box Plot

Dataset: Exam Scores by Class

|  |  |
| --- | --- |
| Class | Scores |
| A | 85 |
| A | 90 |
| A | 78 |
| A | 92 |
| B | 88 |
| B | 76 |
| B | 80 |
| B | 84 |

# Create a data frame

scores\_data <- data.frame(

Class = c("A", "A", "A", "A", "B", "B", "B", "B"),

Scores = c(85, 90, 78, 92, 88, 76, 80, 84)

)

# Plot box plot

ggplot(scores\_data, aes(x = Class, y = Scores, fill = Class)) +

geom\_boxplot() +

labs(title = "Exam Scores by Class",

x = "Class",

y = "Scores") +

theme\_minimal()

7. Density Plot

Dataset: Weights of Individuals

|  |  |
| --- | --- |
| | Weight (kg) | | --- | |
| |  | | --- | | 60 | |
| |  | | --- | | 65 | |
| |  | | --- | | 70 | |
| |  | | --- | | 75 | |
| |  | | --- | | 80 | |
| |  | | --- | | 85 | |
| |  | | --- | | 90 | |
| |  | | --- | | 95 | |
| |  | | --- | | 100 | |

# Create a data frame

weights\_data <- data.frame(

Weight = c(60, 65, 70, 75, 80, 85, 90, 95, 100)

)

# Plot density plot

ggplot(weights\_data, aes(x = Weight)) +

geom\_density(fill = "lightblue") +

labs(title = "Density Plot of Weights",

x = "Weight (kg)",

y = "Density") +

theme\_minimal()

8. Violin Plot

Dataset: Scores by Group

|  |  |
| --- | --- |
| Group | Score |
| X | 80 |
| X | 85 |
| X | 78 |
| X | 92 |
| Y | 88 |
| Y | 76 |
| Y | 80 |
| Y | 84 |

# Create a data frame

scores\_group\_data <- data.frame(

Group = c("X", "X", "X", "X", "Y", "Y", "Y", "Y"),

Score = c(80, 85, 78, 92, 88, 76, 80, 84)

)

# Plot violin plot

ggplot(scores\_group\_data, aes(x = Group, y = Score, fill = Group)) +

geom\_violin() +

labs(title = "Scores by Group",

x = "Group",

y = "Score") +

theme\_minimal()

9. Heatmap

Dataset: Monthly Temperatures

|  |  |  |  |
| --- | --- | --- | --- |
| Month | City A | City B | City C |
| Jan | 5 | 10 | 15 |
| Feb | 6 | 11 | 16 |
| Mar | 7 | 12 | 17 |
| Apr | 8 | 13 | 18 |
| May | 9 | 14 | 19 |

# Create a data frame

temperature\_data <- data.frame(

Month = c("Jan", "Feb", "Mar", "Apr", "May"),

City\_A = c(5, 6, 7, 8, 9),

City\_B = c(10, 11, 12, 13, 14),

City\_C = c(15, 16, 17, 18, 19)

)

# Reshape data for heatmap

library(reshape2)

melted\_data <- melt(temperature\_data, id.vars = "Month")

# Plot heatmap

ggplot(melted\_data, aes(x = Month, y = variable, fill = value)) +

geom\_tile() +

scale\_fill\_gradient(low = "white", high = "red") +

labs(title = "Monthly Temperatures",

x = "Month",

y = "City") +

theme\_minimal()

10. Facet Plot

Dataset: Sales by Month and Region

|  |  |  |
| --- | --- | --- |
| Month | Region | Sales |
| Jan | East | 200 |
| Jan | West | 150 |
| Feb | East | 220 |
| Feb | West | 170 |
| Mar | East | 210 |
| Mar | West | 160 |

# Create a data frame

sales\_region\_data <- data.frame(

Month = c("Jan", "Jan", "Feb", "Feb", "Mar", "Mar"),

Region = c("East", "West", "East", "West", "East", "West"),

Sales = c(200, 150, 220, 170, 210, 160)

)

# Plot facet plot

ggplot(sales\_region\_data, aes(x = Month, y = Sales, fill = Region)) +

geom\_bar(stat = "identity", position = "dodge") +

facet\_wrap(~ Region) +

labs(title = "Sales by Month and Region",

x = "Month",

y = "Sales") +

theme\_minimal()

**11. Area Plot**

**Dataset: Yearly Revenue**

|  |  |
| --- | --- |
| **Year** | **Revenue (in million)** |
| 2015 | 30 |
| 2016 | 35 |
| 2017 | 40 |
| 2018 | 45 |
| 2019 | 50 |

# Create a data frame

revenue\_data <- data.frame(

Year = 2015:2019,

Revenue = c(30, 35, 40, 45, 50)

)

# Plot area plot

ggplot(revenue\_data, aes(x = Year, y = Revenue)) +

geom\_area(fill = "skyblue", alpha = 0.5) +

labs(title = "Yearly Revenue",

x = "Year",

y = "Revenue (in million)") +

theme\_minimal()

**12. Step Plot**

**Dataset: Cumulative Sales**

|  |  |
| --- | --- |
| **Month** | **Cumulative Sales** |
| Jan | 100 |
| Feb | 200 |
| Mar | 300 |
| Apr | 400 |
| May | 500 |

# Load necessary library

library(ggplot2)

# Create a data frame

cumulative\_sales\_data <- data.frame(

Month = factor(c("Jan", "Feb", "Mar", "Apr", "May"), levels = c("Jan", "Feb", "Mar", "Apr", "May")),

Cumulative\_Sales = c(100, 200, 300, 400, 500)

)

# Create the step plot

ggplot(cumulative\_sales\_data, aes(x = Month, y = Cumulative\_Sales, group = 1)) +

geom\_step() +

labs(title = "Cumulative Sales Over Months",

x = "Month",

y = "Cumulative Sales") +

theme\_minimal()

**13. Ridgeline Plot**

**Dataset: Temperature Distributions**

|  |  |
| --- | --- |
| **City** | **Temperature** |
| City1 | 20 |
| City1 | 21 |
| City1 | 19 |
| City2 | 22 |
| City2 | 23 |
| City2 | 24 |

# Load necessary library

library(ggridges)

# Create a data frame

temperature\_data <- data.frame(

City = c("City1", "City1", "City1", "City2", "City2", "City2"),

Temperature = c(20, 21, 19, 22, 23, 24)

)

# Plot ridgeline plot

ggplot(temperature\_data, aes(x = Temperature, y = City, fill = City)) +

geom\_density\_ridges() +

labs(title = "Temperature Distributions",

x = "Temperature",

y = "City") +

theme\_minimal()

**14. Dumbbell Plot**

**Dataset: Sales Before and After Campaign**

|  |  |  |
| --- | --- | --- |
| **Product** | **Before Campaign** | **After Campaign** |
| A | 200 | 250 |
| B | 300 | 350 |
| C | 400 | 450 |
| D | 500 | 550 |
| E | 600 | 650 |

# Load necessary library

library(ggalt)

# Create a data frame

sales\_campaign\_data <- data.frame(

Product = c("A", "B", "C"),

Before = c(200, 300, 400),

After = c(250, 350, 450)

)

# Plot dumbbell plot

ggplot(sales\_campaign\_data) +

geom\_dumbbell(aes(x = Before, xend = After, y = Product), size = 3, color = "gray", size\_x = 3, size\_xend = 3) +

labs(title = "Sales Before and After Campaign",

x = "Sales",

y = "Product") +

theme\_minimal()

**15. Lollipop Plot**

**Dataset: Sales by Region**

|  |  |
| --- | --- |
| **Region** | **Sales** |
| North | 150 |
| South | 200 |
| East | 180 |
| West | 210 |

# Create a data frame

sales\_region\_data <- data.frame(

Region = c("North", "South", "East", "West"),

Sales = c(150, 200, 180, 210)

)

# Plot lollipop plot

ggplot(sales\_region\_data, aes(x = Region, y = Sales)) +

geom\_segment(aes(x = Region, xend = Region, y = 0, yend = Sales), color = "gray") +

geom\_point(color = "blue", size = 5) +

labs(title = "Sales by Region",

x = "Region",

y = "Sales") +

theme\_minimal()

**16. Spaghetti Plot**

**Dataset: Sales over Time by Product**

|  |  |  |
| --- | --- | --- |
| **Month** | **Product** | **Sales** |
| Jan | A | 100 |
| Feb | A | 150 |
| Mar | A | 130 |
| Jan | B | 120 |
| Feb | B | 140 |
| Mar | B | 160 |

# Create a data frame

sales\_time\_data <- data.frame(

Month = c("Jan", "Feb", "Mar", "Jan", "Feb", "Mar"),

Product = c("A", "A", "A", "B", "B", "B"),

Sales = c(100, 150, 130, 120, 140, 160)

)

# Plot spaghetti plot

ggplot(sales\_time\_data, aes(x = Month, y = Sales, group = Product, color = Product)) +

geom\_line(size = 1) +

geom\_point(size = 3) +

labs(title = "Sales over Time by Product",

x = "Month",

y = "Sales") +

theme\_minimal()

**17. Waterfall Plot**

**Dataset: Profit and Loss by Month**

|  |  |
| --- | --- |
| **Month** | **Amount** |
| Jan | 100 |
| Feb | -20 |
| Mar | 50 |
| Apr | -10 |
| May | 80 |

# Load necessary library

library(waterfalls)

# Create a data frame with the appropriate structure

profit\_loss\_data <- data.frame(

labels = c("Jan", "Feb", "Mar", "Apr", "May"),

values = c(100, -20, 50, -10, 80)

)

# Plot waterfall plot

waterfall(profit\_loss\_data, fill\_by\_sign = TRUE)

**18. Hexbin Plot**

**Dataset: X and Y Coordinates**

|  |  |
| --- | --- |
| **X** | **Y** |
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |
| 5 | 6 |
| 6 | 7 |
| 7 | 8 |
| 8 | 9 |

# Load necessary library

library(hexbin)

# Create a data frame

xy\_data <- data.frame(

X = c(1, 2, 3, 4, 5, 6, 7, 8),

Y = c(2, 3, 4, 5, 6, 7, 8, 9)

)

# Plot hexbin plot

ggplot(xy\_data, aes(x = X, y = Y)) +

geom\_hex() +

labs(title = "Hexbin Plot",

x = "X",

y = "Y") +

theme\_minimal()

**19. Chord Diagram**

**Dataset: Connections between Categories**

|  |  |  |
| --- | --- | --- |
| **From** | **To** | **Value** |
| A | B | 10 |
| A | C | 20 |
| B | C | 15 |
| B | D | 25 |
| C | D | 30 |

# Load necessary library

library(circlize)

# Create a data frame

chord\_data <- data.frame(

From = c("A", "A", "B", "B", "C"),

To = c("B", "C", "C", "D", "D"),

Value = c(10, 20, 15, 25, 30)

)

# Plot chord diagram

chordDiagram(chord\_data)

**20. Calendar Heatmap**

**Dataset: Daily Activity Count**

|  |  |
| --- | --- |
| **Date** | **Count** |
| 2023-01-01 | 10 |
| 2023-01-02 | 12 |
| 2023-01-03 | 15 |
| 2023-01-04 | 8 |
| 2023-01-05 | 20 |

# Load necessary library

library(lubridate)

library(ggplot2)

# Create a data frame

activity\_data <- data.frame(

Date = as.Date(c("2023-01-01", "2023-01-02", "2023-01-03", "2023-01-04", "2023-01-05")),

Count = c(10, 12, 15, 8, 20)

)

# Add weekday and week of year columns

activity\_data$Weekday <- wday(activity\_data$Date, label = TRUE)

activity\_data$Week <- week(activity\_data$Date)

# Plot calendar heatmap

ggplot(activity\_data, aes(x = Weekday, y = Week, fill = Count)) +

geom\_tile(color = "white") +

scale\_fill\_gradient(low = "white", high = "blue") +

labs(title = "Daily Activity Count",

x = "Weekday",

y = "Week",

fill = "Count") +

theme\_minimal()

**1: Weather Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Temperature (°C)** | **Humidity (%)** | **Wind Speed (km/h)** |
| 2023-01-01 | 10 | 75 | 15 |
| 2023-01-02 | 12 | 70 | 12 |
| 2023-01-03 | 8 | 80 | 18 |
| 2023-01-04 | 15 | 65 | 20 |
| 2023-01-05 | 14 | 72 | 16 |

1. How does temperature vary with humidity and wind speed over the recorded dates?
2. Can we visualize the relationship between wind speed and humidity, considering temperature as the third dimension?
3. Is there any discernible pattern between temperature, humidity, and wind speed throughout the dataset?
4. Create a 3D surface plot to show how temperature changes with varying levels of humidity and wind speed.
5. Compare the 3D plots of temperature against both humidity and wind speed separately. Are there any notable differences in trends?

**2: Financial Market Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Stock Price ($)** | **Volume Traded (millions)** | **Market Cap ($)** |
| 2023-01-01 | 100 | 2.5 | 500 |
| 2023-01-02 | 102 | 3.0 | 510 |
| 2023-01-03 | 98 | 2.2 | 490 |
| 2023-01-04 | 105 | 2.8 | 525 |
| 2023-01-05 | 108 | 3.5 | 540 |

1. How does stock price relate to both volume traded and market capitalization?
2. Create a 3D scatter plot to visualize the relationship between volume traded, market cap, and stock price.
3. Can we see any clustering or outliers in the 3D plot of stock price, volume traded, and market cap?
4. Generate a 3D surface plot to illustrate how market capitalization changes with variations in stock price and volume traded.
5. Compare the 3D plots of stock price against both volume traded and market cap separately. Are there any noticeable patterns?

**3: Consumer Survey Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Product** | **Price ($)** | **Rating (out of 5)** | **Age Group** |
| A | 50 | 4.2 | 25-35 |
| B | 70 | 3.8 | 35-45 |
| C | 60 | 4.0 | 18-25 |
| D | 45 | 4.5 | 45-55 |
| E | 55 | 3.9 | 25-35 |

1. How do product ratings vary with both price and age group?
2. Create a 3D scatter plot to visualize the relationship between price, rating, and age group.
3. Is there a correlation between age group, product price, and consumer ratings based on the 3D plot?
4. Generate a 3D surface plot to show how product ratings change with variations in both price and age group.
5. Compare the 3D plots of product ratings against both price and age group separately. Are there any significant insights?

**4: Environmental Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Location** | **Temperature (°C)** | **Humidity (%)** | **CO2 Levels (ppm)** |
| A | 15 | 65 | 400 |
| B | 20 | 70 | 450 |
| C | 18 | 68 | 420 |
| D | 12 | 60 | 380 |
| E | 17 | 72 | 430 |

1. How do CO2 levels vary with both temperature and humidity across different locations?
2. Create a 3D scatter plot to visualize the relationship between temperature, humidity, and CO2 levels.
3. Is there a spatial pattern in the 3D plot of temperature, humidity, and CO2 levels among the locations?
4. Generate a 3D surface plot to illustrate how CO2 levels change with variations in both temperature and humidity.
5. Compare the 3D plots of CO2 levels against both temperature and humidity separately. Are there any noticeable trends?

**5: Academic Performance Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student** | **Math Score** | **Science Score** | **Attendance (%)** |
| A | 85 | 78 | 95 |
| B | 72 | 85 | 92 |
| C | 90 | 80 | 98 |
| D | 78 | 75 | 85 |
| E | 88 | 82 | 93 |

1. How do science scores vary with both math scores and attendance percentage?
2. Create a 3D scatter plot to visualize the relationship between math score, science score, and attendance.
3. Is there a correlation between attendance, math scores, and science scores based on the 3D plot?
4. Generate a 3D surface plot to show how science scores change with variations in both math scores and attendance.
5. Compare the 3D plots of science scores against both math scores and attendance separately. Are there any significant patterns or outliers?