DSE\_Capstone\_EDA\_Canfield

Ryan Canfield

2024-11-10

## R Markdown

library(tidyverse) # For data manipulation.

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2) # For data visualizations.   
library(RColorBrewer) # For coloring visuals.

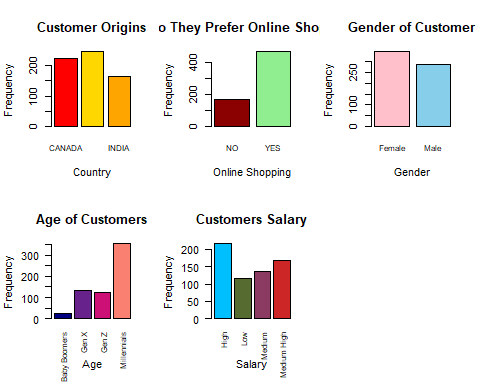
# Reading the CSV file into a data frame  
df <- read.csv("cleaned\_data.csv")  
  
# Display a preview of the data frame.  
head(df)

# Visualizations:

## Frequency Bar Charts of Demographic Variables:

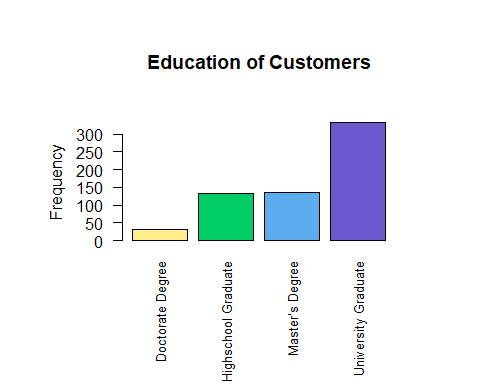
### This table will present frequencies and percentages of key demographic variables, such as country, age, gender, education, and online service preferences.

# Specify colors for each level   
country\_colors <- c("red", "gold", "orange")   
age\_colors <- c("navy", "darkorchid4", "deeppink3", "salmon")  
gender\_colors <- c("pink", "skyblue")   
education\_colors <- c("lightgoldenrod1", "springgreen3", "steelblue2", "slateblue3")   
OnlineS\_colors <- c("darkred", "lightgreen")   
salary\_colors <- c("deepskyblue", "darkolivegreen", "hotpink4", "firebrick3")  
  
  
  
par(mfrow=c(2,3))  
  
# Plot 1: Country   
barplot(table(df$Country),   
 col = country\_colors,   
 xlab = "Country",   
 ylab = "Frequency",  
 main = "Customer Origins",   
 cex.names = 0.8)   
  
# Plot 2: Online Service Preference  
barplot(table(df$Online\_Service\_Preference),   
 col = OnlineS\_colors,   
 xlab = "Online Shopping",   
 ylab = "Frequency",  
 main = "Do They Prefer Online Shopping",   
 cex.names = 0.8)  
  
# Plot 3: Gender   
barplot(table(df$Gender),   
 col = gender\_colors,   
 xlab = "Gender",   
 ylab = "Frequency",  
 main = "Gender of Customer",   
 cex.names = 0.8)  
  
# Plot 4: Age   
barplot(table(df$Age),   
 col = age\_colors,   
 xlab = "Age",   
 ylab = "Frequency",  
 main = "Age of Customers",   
 las = 2,   
 cex.names = 0.8)  
   
  
# Plot 5: Annual Salary   
barplot(table(df$Annual\_Salary),   
 col = salary\_colors,   
 xlab = "Salary",   
 ylab = "Frequency",  
 main = "Customers Salary",   
 las = 2,   
 cex.names = 0.8)



par(mar=c(5, 4, 4, 2) + 2.4) # Adjusts the bottom margin

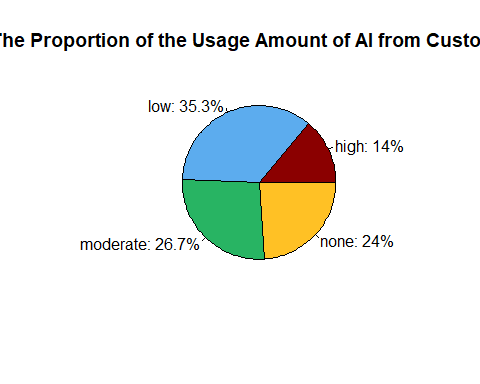
# Plot 6: Education   
barplot(table(df$Education),   
 col = education\_colors,   
 ylab = "Frequency",  
 main = "Education of Customers",   
 las = 2,   
 cex.names = 0.8)



## Pie Charts of AI Tool Usage:

### These pie charts will show the proportion of customers in high, moderate, and low AI usage categories.

# Calculate the percentage of people who use AI  
AI\_usage\_percents <- df %>%  
 count(AI\_Usage) %>%  
 mutate(Percentage = n / sum(n) \* 100)  
  
# Create the pie chart using base R  
pie(AI\_usage\_percents$Percentage,   
 labels = paste0(AI\_usage\_percents$AI\_Usage, ": ", round(AI\_usage\_percents$Percentage, 1), "%"),   
 col = c('darkred', 'steelblue2', '#28B463', 'goldenrod1'), # Customize colors  
 main = "The Proportion of the Usage Amount of AI from Customers")



## Scatterplot of AI Trust vs. AI Satisfaction:

### This scatter plot will visualize the relationship between the newly created AI\_Trust variable and AI satisfaction levels.

unique(df$AI\_Trust)

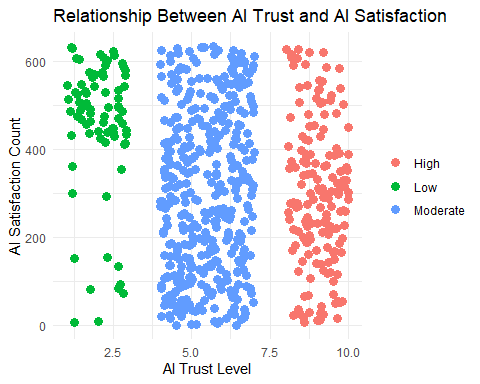
## [1] "Moderate" "Low" "High"

unique(df$AI\_Satisfaction)

## [1] "YES" "NO"

set.seed(3) # For repeated results  
  
# Function to generate random number based on category  
generate\_random <- function(category) {  
 if (category == "Low") {  
 return(round(runif(1, min = 1, max = 3), 2)) # Random number between 1 and 3  
 } else if (category == "Moderate") {  
 return(round(runif(1, min = 4, max = 7), 2)) # Random number between 4 and 7  
 } else if (category == "High") {  
 return(round(runif(1, min = 8, max = 10), 2)) # Random number between 8 and 10  
 }  
}  
  
  
df$Random\_Num\_based\_AITrust <- sapply(df$AI\_Trust, generate\_random)  
  
# View the result  
head(df)

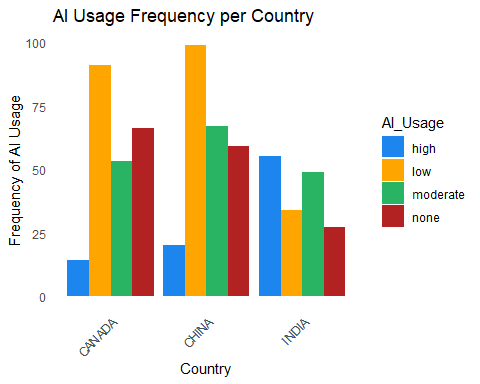
df$Index <- seq\_along(df$AI\_Trust)  
  
  
ggplot(df, aes(x = Random\_Num\_based\_AITrust, y = Index)) +  
 geom\_point(aes(color = AI\_Trust), size = 3) +  
 labs(  
 title = "Relationship Between AI Trust and AI Satisfaction",  
 x = "AI Trust Level",  
 y = "AI Satisfaction Count"  
 ) +  
 theme\_minimal() +  
 theme(legend.title = element\_blank())



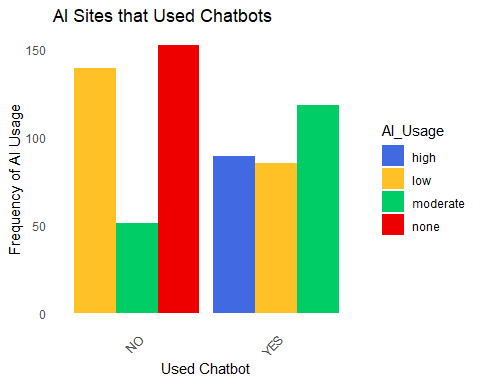
## Histogram of AI Usage Levels:

### This histogram will show the distribution of AI usage levels (high, moderate, low) across the dataset.

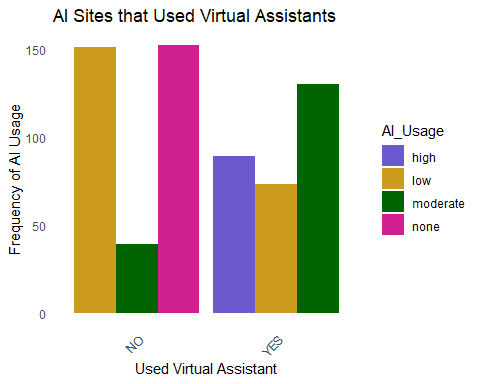
# Summarize the data to get frequencies  
data\_summary <- df %>%  
 count(Country, AI\_Usage)  
  
# Create the bar graph  
ggplot(data\_summary, aes(x = Country, y = n, fill = AI\_Usage)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(x = "Country", y = "Frequency of AI Usage", title = "AI Usage Frequency per Country") +  
 scale\_fill\_manual(values = c("high" = "dodgerblue2", "moderate" = "#28B463", "low" = "orange", "none" = "firebrick")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 panel.grid = element\_blank())



# Summarize the data to get frequencies  
data\_summary <- df %>%  
 count(AI\_Tools\_Used\_Chatbots, AI\_Usage)  
  
# Create the bar graph  
ggplot(data\_summary, aes(x = AI\_Tools\_Used\_Chatbots, y = n, fill = AI\_Usage)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(x = "Used Chatbot", y = "Frequency of AI Usage", title = "AI Sites that Used Chatbots") +  
 scale\_fill\_manual(values = c("high" = "royalblue", "moderate" = "springgreen3", "low" = "goldenrod1", "none" = "red2")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 panel.grid = element\_blank())



# Summarize the data to get frequencies  
data\_summary <- df %>%  
 count(AI\_Tools\_Used\_Virtual\_Assistant, AI\_Usage)  
  
# Create the bar graph  
ggplot(data\_summary, aes(x = AI\_Tools\_Used\_Virtual\_Assistant, y = n, fill = AI\_Usage)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(x = "Used Virtual Assistant", y = "Frequency of AI Usage", title = "AI Sites that Used Virtual Assistants") +  
 scale\_fill\_manual(values = c("high" = "slateblue", "moderate" = "darkgreen", "low" = "goldenrod3", "none" = "violetred")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 panel.grid = element\_blank())



# Summarize the data to get frequencies  
data\_summary <- df %>%  
 count(AI\_Tools\_Used\_Voice\_Photo, AI\_Usage)  
  
# Create the bar graph  
ggplot(data\_summary, aes(x = AI\_Tools\_Used\_Voice\_Photo, y = n, fill = AI\_Usage)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(x = "Used Voice or Photo", y = "Frequency of AI Usage", title = "AI Sites that Used Photos or Voice") +  
 scale\_fill\_manual(values = c("high" = "turquoise2", "moderate" = "yellowgreen", "low" = "khaki", "none" = "tomato")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 panel.grid = element\_blank())

