Patient Panel Survey

Raaga Likhitha

required

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(writexl)
```

```
## Warning: package 'writexl' was built under R version 4.3.3
```

```r
# Load the dataset
file_path <- "C:/Users/drraa/Downloads/Patient_Engagement_Assignment_2 (1) (2)-1.xlsx"
data <- read_excel(file_path, skip = 1)
head(data)
```

```
## # A tibble: 6 x 9
##   Name     `Date Follow-up Due`      `Actual Contact Dates` `Type of Contact`
##   <chr>    <chr>                     <chr>                  <chr>
## 1 Joe Test 2-week follow-up schedule <NA>                   <NA>
## 2 Joe Test 42987                     42987                  Digital
## 3 Joe Test 43001                     43001                  Digital
## 4 Joe Test 43015                     43015                  Digital
## 5 Joe Test 43029                     43029                  Phone
## 6 Joe Test 43043                     43043                  In person at clinic
## # i 5 more variables: `PHQ-9 Score` <chr>,
## #   `Patient Activation Measure (PAM)` <chr>, `Doses Per Day (3)` <chr>,
## #   `Self reported Medication Adherence` <chr>, `Avg Steps per day` <chr>
```

```r
print(colnames(data))
```

```
## [1] "Name"                      "Date Follow-up Due"
## [3] "Actual Contact Dates"      "Type of Contact"
## [5] "PHQ-9 Score"               "Patient Activation Measure (PAM)"
## [7] "Doses Per Day (3)"         "Self reported Medication Adherence"
## [9] "Avg Steps per day"
```

```r
# Assuming the column names based on the initial row and correcting them
colnames(data) <- c("Name", "Date_Follow_up_Due", "Actual_Contact_Dates", "Type_of_Contact", "PHQ_9_Sco

# Convert data types if necessary, for example:
data$PHQ_9_Score <- as.numeric(data$PHQ_9_Score)
```

```
## Warning: NAs introduced by coercion
```

```r
data$PAM <- as.numeric(data$PAM)
```

```
## Warning: NAs introduced by coercion
```

```r
data$Avg_Steps_per_day <- as.numeric(data$Avg_Steps_per_day)
```

```
## Warning: NAs introduced by coercion
```

```r
summary(data$PHQ_9_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.00   11.00   15.00   14.49   18.00   22.00      37
```

```r
summary(data$PAM)
```
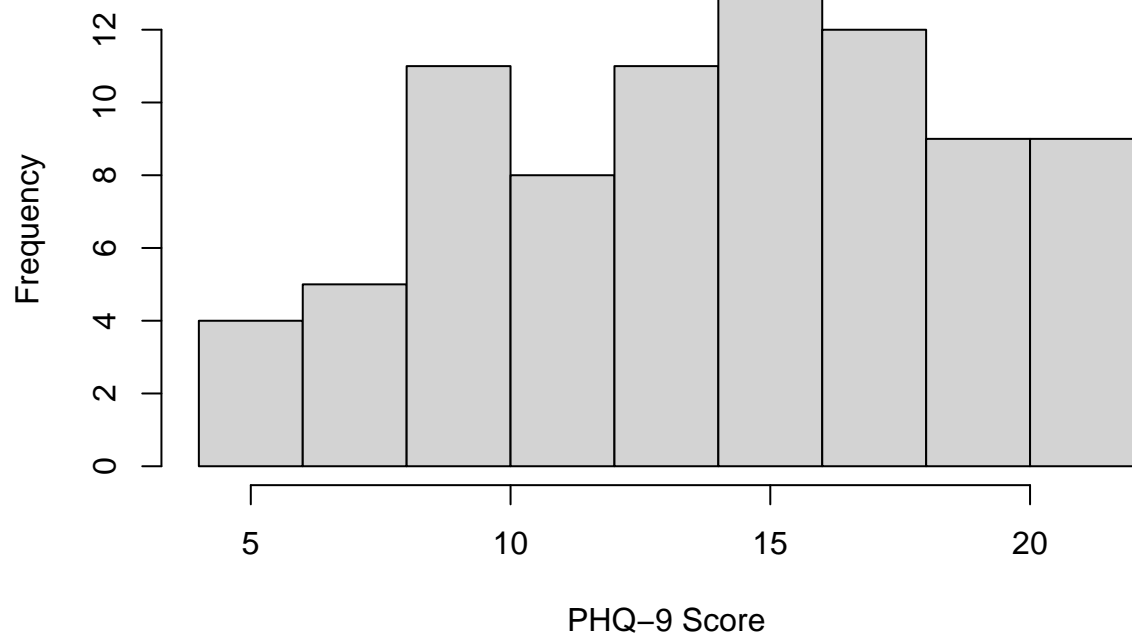
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   2.268   3.000   4.000      37
```

```r
summary(data$Avg_Steps_per_day)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3453    8996   10727   10312   11999   14760      37
```
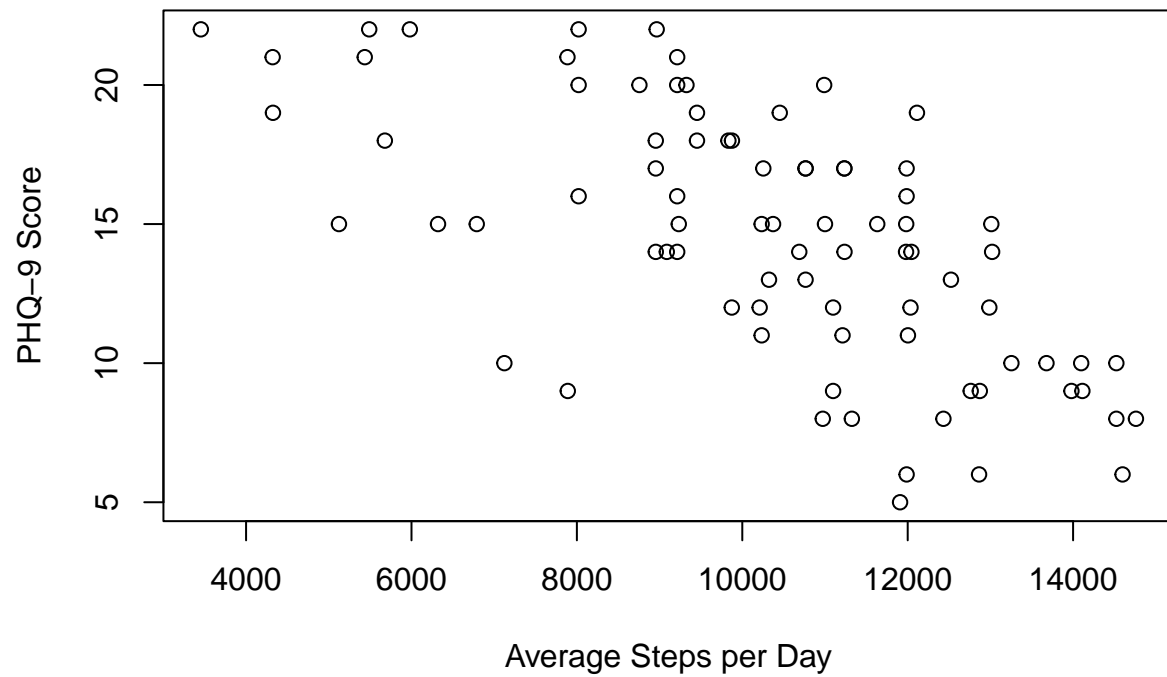
```r
hist(data$PHQ_9_Score, main = "Histogram of PHQ-9 Scores", xlab = "PHQ-9 Score")
```
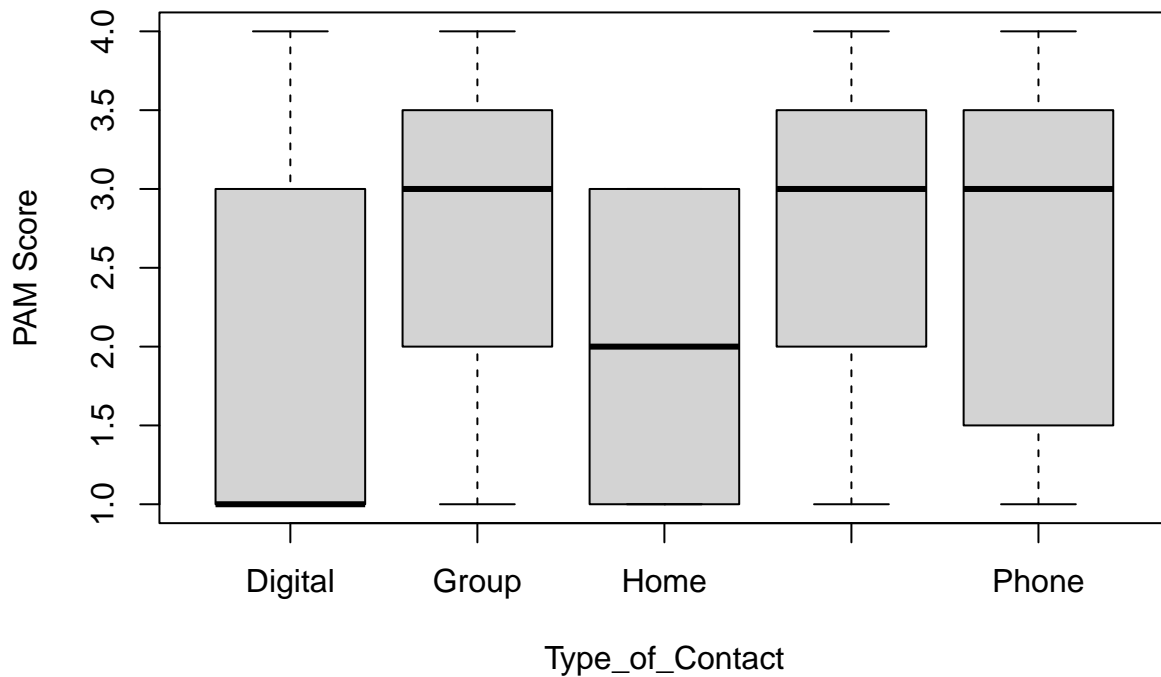
Histogram of PHQ–9 Scores


```r
plot(data$Avg_Steps_per_day, data$PHQ_9_Score, main = "PHQ-9 Score vs. Avg Steps per Day", xlab = "Avera
```

PHQ−9 Score vs. Avg Steps per Day


```r
boxplot(PAM ~ Type_of_Contact, data = data, main = "PAM Scores by Type of Contact", ylab = "PAM Score")
```

PAM Scores by Type of Contact


```r
library(readxl)
library(dplyr)
library(ggplot2)
# data-preparation
data <- data %>%
  rename(Name = `Name`,
         DateFollowUpDue = `Date_Follow_up_Due`,
         ActualContactDates = `Actual_Contact_Dates`,
         TypeOfContact = `Type_of_Contact`,
         PHQ9Score = `PHQ_9_Score`) %>%
  mutate(PHQ9Score = as.numeric(PHQ9Score)) %>% # Convert to numeric if not already
  filter(!is.na(PHQ9Score) & !is.na(TypeOfContact))

# Now let's calculate the average PHQ-9 Score by type of contact- 'score-analysis'
average_scores_by_contact <- data %>%
  group_by(TypeOfContact) %>%
  summarise(AveragePHQ9 = mean(PHQ9Score, na.rm = TRUE))

# To see the list of average scores by contact type
print(average_scores_by_contact)
```

```
## # A tibble: 5 x 2
##   TypeOfContact      AveragePHQ9
##   <chr>                    <dbl>
## 1 Digital                   17.0
```

```
## 2 Group                       9.67
## 3 Home                         16
## 4 In person at clinic         10.9
## 5 Phone                       11.8
```

```
#'score-visualization'
```

```
ggplot(average_scores_by_contact, aes(x = TypeOfContact, y = AveragePHQ9, fill = TypeOfContact)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Average PHQ-9 Score by Type of Contact", x = "Type of Contact", y = "Average PHQ-9 Score
```

Average PHQ−9 Score by Type of Contact
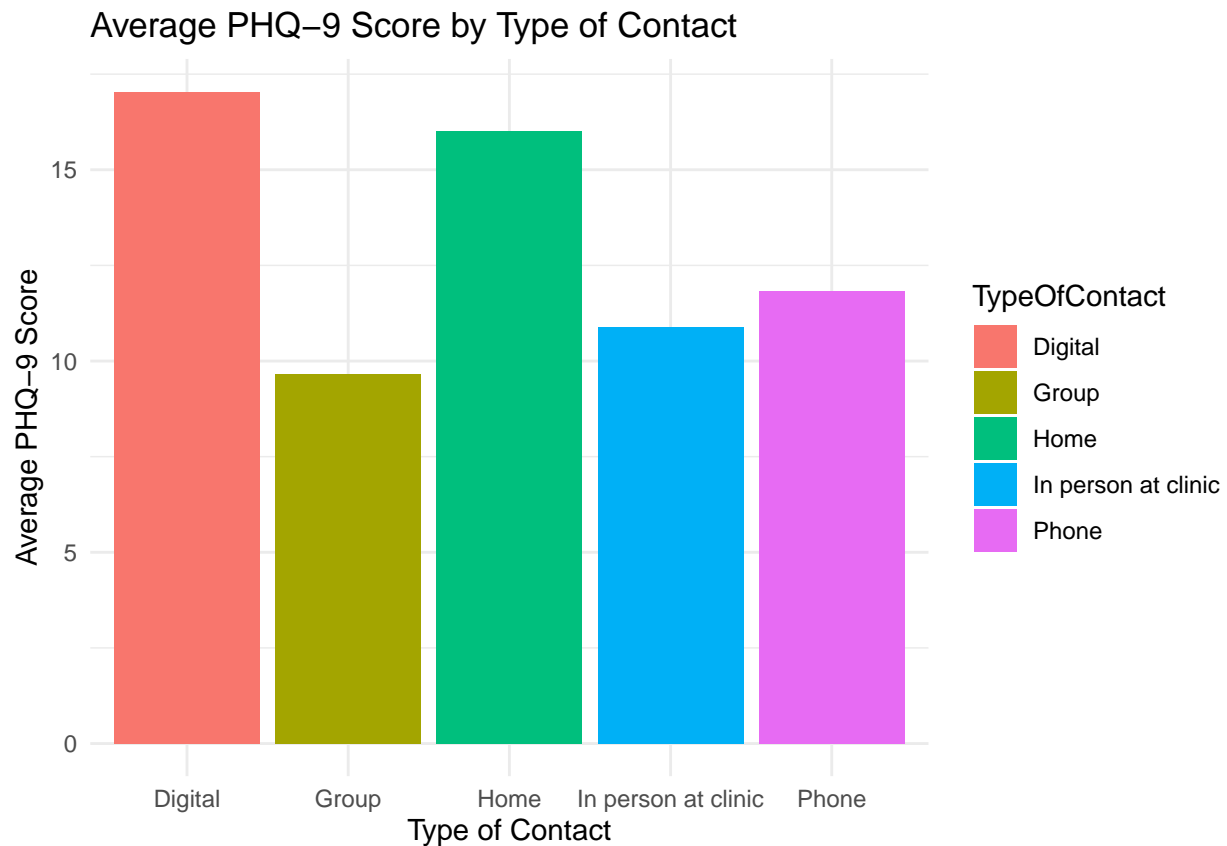

```
# Convert PHQ-9 Score to numeric in case it's read as a character
data$PHQ9Score <- as.numeric(as.character(data$PHQ9Score))
```

Looking at the PHQ-9 scores by the "Type of Contact" we get that:

In the digital,There were 45 entries with a variety of scores, where the most frequent score was 15, occurring 6 times among different participants. Group: Only 3 entries were recorded with each having a unique PHQ-9 score, one of which was 8. Home: There were 2 entries with distinct scores, one of them being 15. In person at clinic: 20 entries were noted, with scores ranging across 11 unique values. The most common score was 9, observed 3 times. Missed: There were 18 entries all marked as "Missed," indicating no score was recorded. Phone: 12 entries were made, with 6 different scores noted. The score of 12 was the most frequent, occurring 3 times.

From these insights, it seems that:

Digital and in-person clinic contacts have a broader distribution of PHQ-9 scores, indicating a wider range of depression severity among these groups. There's a significant number of missed scores, which could potentially skew understanding of overall depression severity if not accounted for. The smaller counts for "Group" and "Home" types might not provide a strong basis for general conclusions but do suggest that such contacts are less frequent or less commonly used for this patient group.
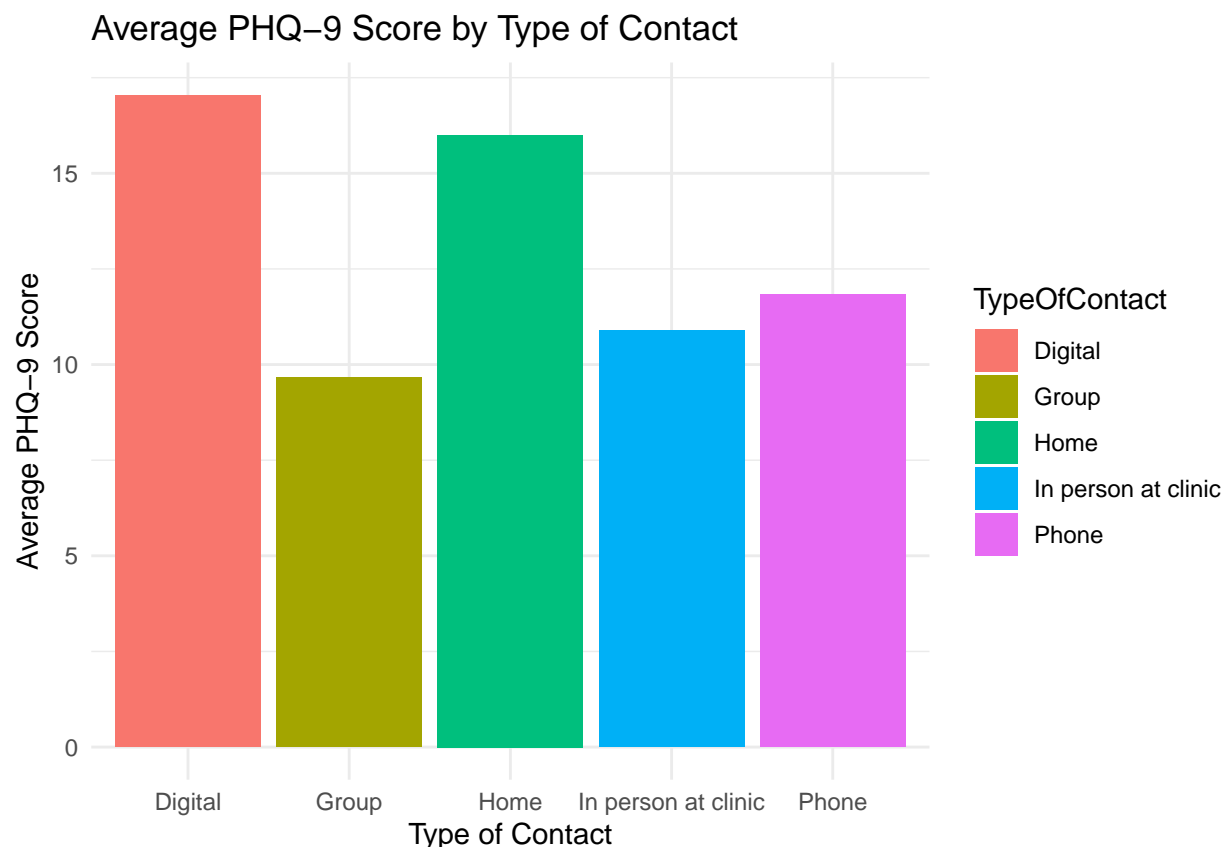
The average PHQ-9 scores by type of contact are:

Digital: 17.04 Group: 9.67 Home: 16.00 In person at clinic: 10.90 Phone: 11.83

This suggests that encounters through digital means and home visits are associated with higher depression screening scores, indicating potentially more severe symptoms, whereas group sessions and in-person clinic visits show lower scores on average.

```
average_scores_by_contact <- data %>%
  group_by(TypeOfContact) %>%
  summarise(AveragePHQ9 = mean(PHQ9Score, na.rm = TRUE))
```

```
ggplot(average_scores_by_contact, aes(x = TypeOfContact, y = AveragePHQ9, fill = TypeOfContact)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Average PHQ-9 Score by Type of Contact", x = "Type of Contact", y = "Average PHQ-9 Score
```



```
# Plotting trends for each patient
patient_trends <- data %>%
  group_by(Name, TypeOfContact) %>%
  summarise(AveragePHQ9 = mean(PHQ9Score, na.rm = TRUE)) %>%
```

```r
  ungroup() %>%
  mutate(Name = factor(Name, levels = unique(Name))) # this ensures consistency in the ordering of the
```

```
## `summarise()` has grouped output by 'Name'. You can override using the
## `.groups` argument.
```

```r
ggplot(patient_trends, aes(x = Name, y = AveragePHQ9, fill = TypeOfContact)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Trends of PHQ-9 Score per Patient by Type of Contact", x = "Patient", y = "Average PHQ-
```


Trends of PHQ−9 Score per Patient by Type of Contact
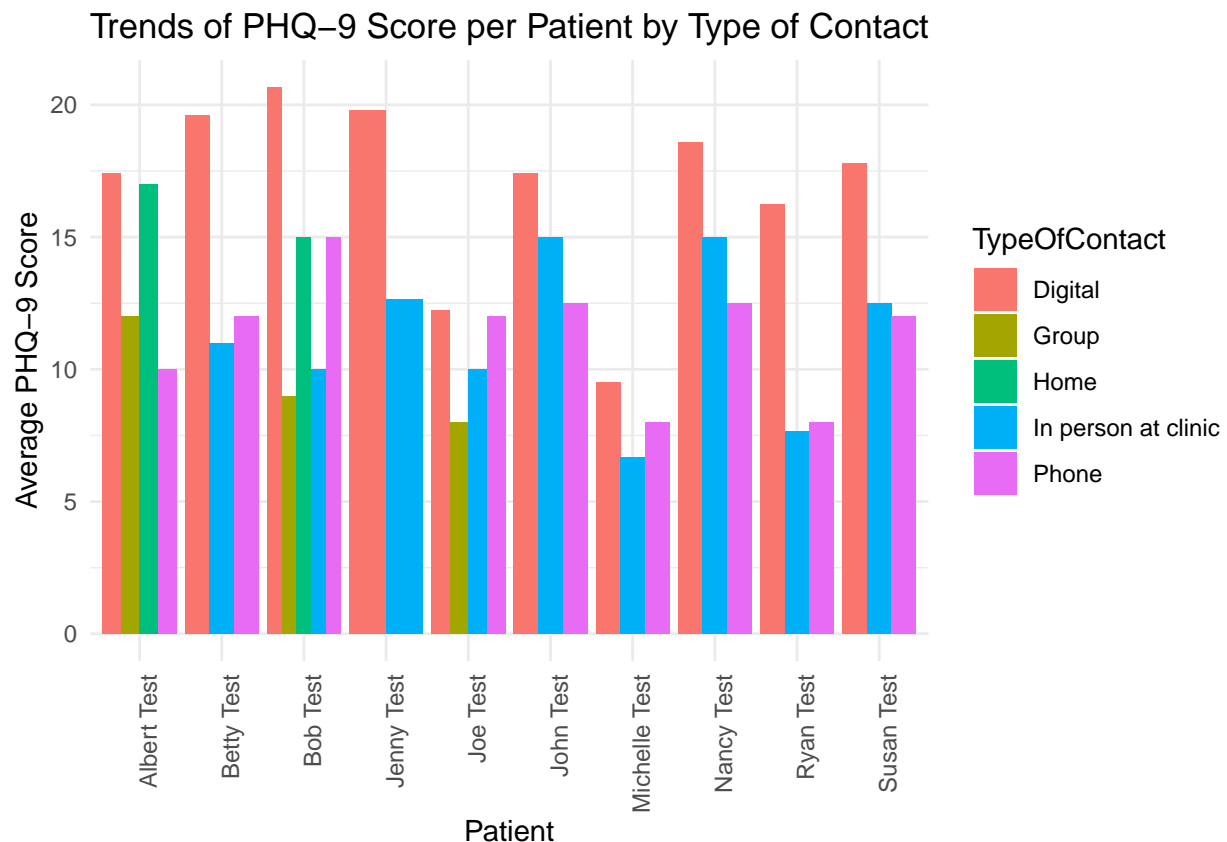
```r
# Summarize the average PHQ-9 scores by type of contact
average_scores_by_contact <- data %>%
  group_by(TypeOfContact) %>%
  summarise(AveragePHQ9 = mean(PHQ9Score, na.rm = TRUE))

# Print the summarized data frame to see the list
print(average_scores_by_contact)
```

```
## # A tibble: 5 x 2
##   TypeOfContact      AveragePHQ9
##   <chr>                    <dbl>
## 1 Digital                   17.0
```

```
## 2 Group                     9.67
## 3 Home                      16
## 4 In person at clinic       10.9
## 5 Phone                     11.8
```

These averages suggest that digital and home contacts are associated with higher PHQ-9 scores, indicating more severe symptoms of depression among patients engaged through these methods. In contrast, group and in-person clinic contacts show lower average scores, suggesting less severe symptoms among these patients.

trend per patient

```r
print(colnames(data))
```

```
## [1] "Name"               "DateFollowUpDue"      "ActualContactDates"
## [4] "TypeOfContact"      "PHQ9Score"            "PAM"
## [7] "Doses_Per_Day"      "Medication_Adherence" "Avg_Steps_per_day"
```

```r
library(dplyr)
library(ggplot2)

# Calculate the trend of PHQ-9 scores for each patient
trend_per_patient <- data %>%
  group_by(Name) %>%
  summarise(AveragePHQ9 = mean(PHQ9Score, na.rm = TRUE))

# View the trend per patient
print(trend_per_patient)
```

```
## # A tibble: 10 x 2
##    Name          AveragePHQ9
##    <chr>               <dbl>
##  1 Albert Test          15.8
##  2 Betty Test           16.5
##  3 Bob Test             15.8
##  4 Jenny Test           17.1
##  5 Joe Test             11.1
##  6 John Test            15.8
##  7 Michelle Test         8.25
##  8 Nancy Test           16.4
##  9 Ryan Test            12
## 10 Susan Test           15.8
```

```r
# Calculate the average PHQ-9 score for each patient by type of contact
trends_per_patient_by_contact <- data %>%
  group_by(Name, TypeOfContact) %>%
  summarise(AveragePHQ9 = mean(PHQ9Score, na.rm = TRUE)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Name'. You can override using the
## `.groups` argument.
```

```
# View the trends per patient by contact type
print(trends_per_patient_by_contact)
```

```
## # A tibble: 33 x 3
##    Name         TypeOfContact       AveragePHQ9
##    <chr>        <chr>                     <dbl>
##  1 Albert Test  Digital                    17.4
##  2 Albert Test  Group                      12
##  3 Albert Test  Home                       17
##  4 Albert Test  Phone                      10
##  5 Betty Test   Digital                    19.6
##  6 Betty Test   In person at clinic        11
##  7 Betty Test   Phone                      12
##  8 Bob Test     Digital                    20.7
##  9 Bob Test     Group                       9
## 10 Bob Test     Home                       15
## # i 23 more rows
```

```
# Plot the trends per patient by contact type
ggplot(trends_per_patient_by_contact, aes(x = Name, y = AveragePHQ9, fill = TypeOfContact)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Trends of PHQ-9 Score per Patient by Type of Contact", x = "Patient", y = "Average PHQ-9
```
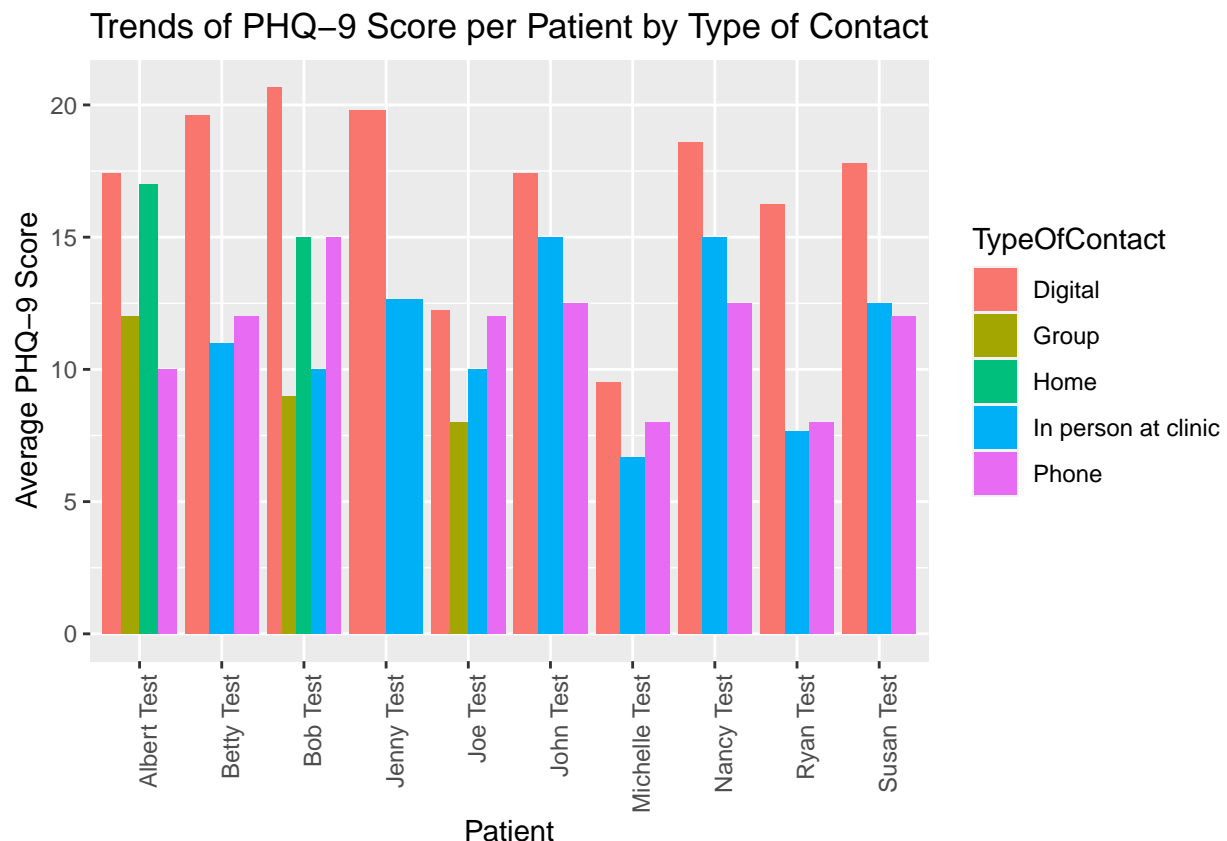

The average scores vary from patient to patient, with some showing higher averages (e.g., Betty Test at 16.50) and others lower (e.g., Michelle Test at 8.25), indicating individual differences in depression severity.

When we delve into deeper analysis: Albert Test had the highest average score with Digital contact (17.4) and the lowest with Phone contact (10.0). Betty Test showed a high average with Digital (19.6) and lower scores for In-person at clinic (11.0) and Phone (12.0). Bob Test's scores were notably higher with Digital contact (20.67) compared to other methods. Michelle Test presented lower averages across the board, with Digital contact being higher (9.5) than In-person at clinic (6.67) and Phone (8.0).

ANOVA

```
# Load necessary library
library(stats)
# First, to  ensure that 'TypeOfContact' is treated as a factor (categorical variable)
data$TypeOfContact <- as.factor(data$TypeOfContact)

# Perform ANOVA
anova_result <- aov(PHQ9Score ~ TypeOfContact, data = data)

# Display the ANOVA table
summary(anova_result)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## TypeOfContact  4   710.4  177.61    14.19 1.02e-08 ***
## Residuals      77  964.0   12.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

there is a statistically significant difference in PHQ-9 scores across different types of contact, as the p-value ($Pr(>F)$) is extremely small (1.02e-08), well below the typical alpha level of 0.05 used to denote statistical significance ANOVA table shows that The analysis involves 4 degrees of freedom for the treatment (TypeOfContact) due to there being 5 groups (the number of groups minus one), and 77 degrees of freedom for the residuals (error term), which corresponds to the number of observations minus the number of groups.

Sum Sq (Sum of Squares) shows the variability due to the TypeOfContact (710.4) and the variability within groups (residuals, 964.0). The mean square is calculated by dividing the sum of squares by the corresponding degrees of freedom. For TypeOfContact, it's 177.61, and for residuals, it's 12.52. The F statistic is 14.19, calculated as the ratio of the mean square for the TypeOfContact to the mean square for the residuals. This value is used to determine the p-value. The p-value is 1.02e-08, indicating that the observed difference in means across the types of contact is extremely unlikely to have occurred by chance.

The significant F-statistic suggests that at least one group's mean PHQ-9 score is significantly different from the others, which implies that the type of contact has a statistically significant effect on the PHQ-9 scores among the groups analyzed.

To identify which specific groups (types of contact) differ from each other, we are now doing a POST HOC ANALYSIS

```
# Load necessary library for post-hoc analysis
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
# Tukey HSD post-hoc test
tukey_result <- TukeyHSD(anova_result)
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = PHQ9Score ~ TypeOfContact, data = data)
##
## $TypeOfContact
##                                 diff       lwr       upr     p adj
## Group-Digital               -7.3777778 -13.271653 -1.483903 0.0068392
## Home-Digital                -1.0444444  -8.187349  6.098460 0.9940346
## In person at clinic-Digital -6.1444444  -8.800778 -3.488111 0.0000001
## Phone-Digital               -5.2111111  -8.422462 -1.999760 0.0001997
## Home-Group                   6.3333333  -2.689783 15.356450 0.2949791
## In person at clinic-Group    1.2333333  -4.886442  7.353109 0.9799702
## Phone-Group                  2.1666667  -4.213640  8.546973 0.8768132
## In person at clinic-Home    -5.1000000 -12.430414  2.230414 0.3036295
## Phone-Home                  -4.1666667 -11.715947  3.382614 0.5389202
## Phone-In person at clinic    0.9333333  -2.675913  4.542580 0.9507119
```

The results from Tukey's Honestly Significant Difference (HSD) test provide detailed comparisons between each pair of types of contact regarding their PHQ-9 scores.

Significant Differences: Group vs. Digital: The mean difference is -7.38, with a p-value of 0.0068392, indicating a significant difference. Group contacts have lower PHQ-9 scores than Digital contacts. In person at clinic vs. Digital: The mean difference is -6.14, with a p-value of 0.0000001, indicating a highly significant difference. In-person contacts at the clinic result in lower PHQ-9 scores compared to Digital. Phone vs. Digital: The mean difference is -5.21, with a p-value of 0.0001997, also indicating a significant difference. Phone contacts have lower PHQ-9 scores than Digital contacts.

Non-Significant Differences: Home vs. Digital: The difference is not statistically significant (p-value of 0.9940346), suggesting similar PHQ-9 scores between these two types of contact. Home vs. Group, In person at clinic vs. Group, Phone vs. Group, In person at clinic vs. Home, Phone vs. Home, and Phone vs. In person at clinic: All these comparisons show non-significant differences in their p-values, indicating that there are no statistically significant differences in PHQ-9 scores between these groups.

The significant results indicate that Digital contacts are associated with higher PHQ-9 scores compared to Group, In-person at clinic, and Phone contacts. This could imply that patients engaging through digital means might report higher levels of depression severity. Conversely, the non-significant differences among Home, Group, In-person at clinic, and Phone suggest similar effectiveness or experiences among these contact types in terms of PHQ-9 scores.

```r
# Load the ggplot2 library for plotting
library(ggplot2)

data$TypeOfContact <- as.factor(data$TypeOfContact)

# Create the boxplot
ggplot(data, aes(x=TypeOfContact, y=PHQ9Score,fill=TypeOfContact)) +
  geom_boxplot() +
  labs(title="Box Plot of PHQ-9 Scores by Type of Contact",
       x="Type of Contact",
       y="PHQ-9 Score") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```
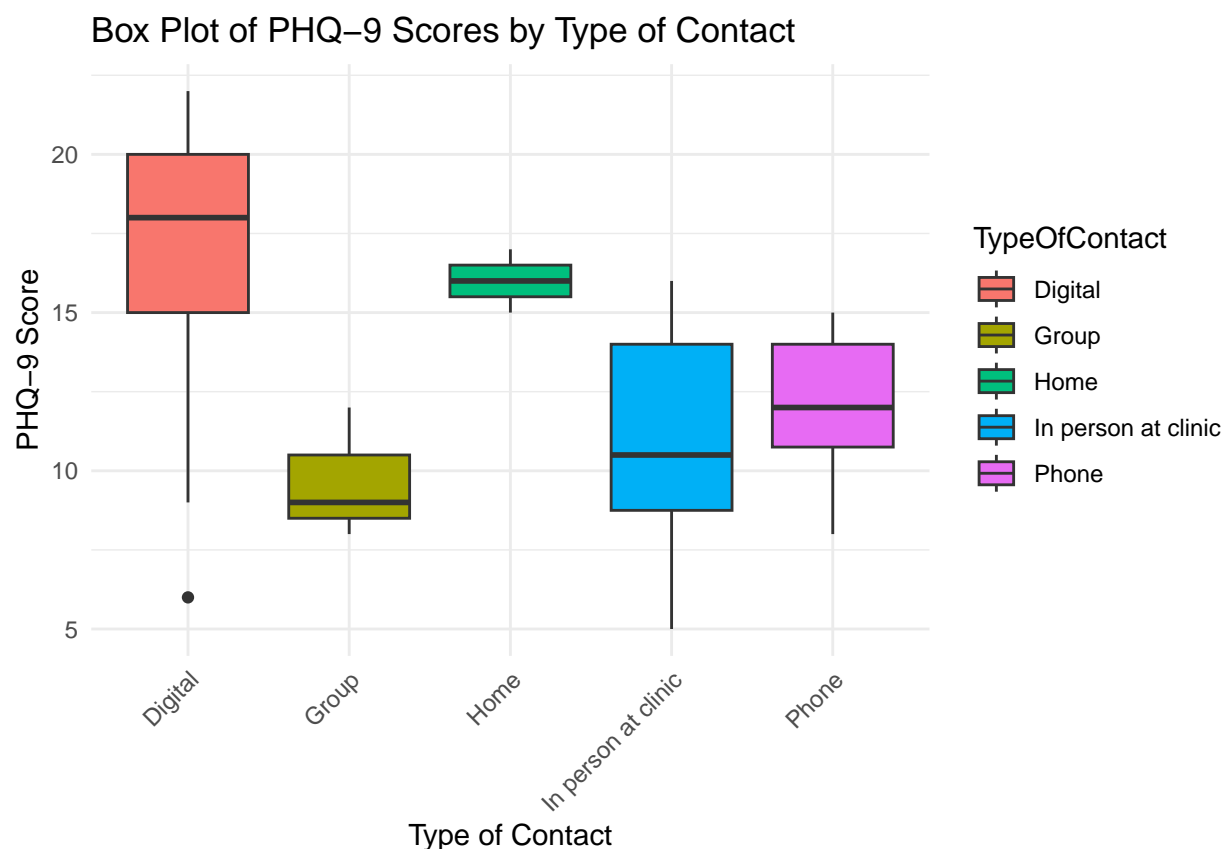

Categories of Contact:

Digital: The red box plot represents PHQ-9 scores for digital contact. The scores here have a wider IQR, suggesting more variability in depression scores, with the median above the medians of other contact types. There's at least one outlier, indicating an unusually low score compared to the rest of the digital data. Group: The light green box plot represents group contact. It has a narrower IQR, indicating less variability. The median is lower than that of digital contact. Home: The green box plot stands for home contact. This

group has an even spread of data with no outliers visible, and the median is lower than digital but similar to group contact. In person at clinic: The blue box plot represents in-person contact at the clinic. The IQR is broader than group but narrower than digital, and the median PHQ-9 score is similar to that of group and home. Phone: The purple box plot corresponds to phone contact. This category's median is similar to that of in-person and group, with a moderately narrow IQR.

Digital contact has the highest median PHQ-9 scores, which may suggest that patients report higher levels of depression severity through digital means, or it could reflect a group with more variability in scores. Group, Home, In person at clinic, and Phone contacts have similar median scores, which are notably lower than the Digital contact. It suggests that patients contacted through these means have similar levels of reported depression severity. The presence of outliers in the Digital category could indicate special cases that are significantly different from the majority of patients in that group.

Optional

The analysis of average PHQ-9 scores by patient across different types of contact does suggest that the type of contact might matter, but the impact varies among patients. So I think that 1.There's noticeable variability in average PHQ-9 scores across different types of contact for the same patients. For some, digital contact shows higher depression scores, while others have relatively consistent scores across contact types. 2. Patients tend to have higher average scores with digital contact. This could indicate that digital methods might either be used more with patients experiencing severe symptoms, or that this mode of contact is more effective at capturing higher PHQ-9 scores. 3.Some types of contact, like "Group" and "Home," have limited data points, making it difficult to draw broad conclusions for these categories.

higher average scores are seen with digital contact, it might be particularly effective for engaging patients with more severe symptoms. This could be due to the privacy and comfort of engaging through digital means, encouraging more openness about symptoms.

Recommendations : 1.For patients showing higher severity in depression symptoms, expanding digital engagement could be beneficial. 2. Given the variability among patients, a personalized approach in choosing the contact method might be more effective. Patient preferences and the severity of their symptoms should guide the engagement strategy.

```
colnames(data)
```

```
## [1] "Name"                "DateFollowUpDue"      "ActualContactDates"
## [4] "TypeOfContact"       "PHQ9Score"            "PAM"
## [7] "Doses_Per_Day"       "Medication_Adherence" "Avg_Steps_per_day"
```

```
library(dplyr)
library(tidyr)


# Convert relevant columns to numeric, as needed

data$PAM_Score_Numeric <- as.numeric(as.character(data$PAM))
data$Medication_Adherence_Percent <- as.numeric(as.character(data$Medication_Adherence))
data$Avg_Steps_per_day_Numeric <- as.numeric(as.character(data$Avg_Steps_per_day))

# Group by patient to get the latest PHQ-9 and PAM scores, and calculate average medication adherence a
patients_summary <- data %>%
  group_by(Name) %>%
  summarise(
    PHQ_9_Score_Numeric = last(na.omit(PHQ9Score)),
```

```r
    PAM_Score_Numeric = last(na.omit(PAM_Score_Numeric)),
    Medication_Adherence_Percent = mean(na.omit(Medication_Adherence_Percent), na.rm = TRUE),
    Avg_Steps_per_day_Numeric = mean(na.omit(Avg_Steps_per_day_Numeric), na.rm = TRUE)
  ) %>%
  ungroup()

# Define the urgency criteria
calculate_urgency <- function(PHQ_9, PAM, Medication_Adherence, Avg_Steps) {
  urgency_score <- 0
  if(PHQ_9 < 5) { urgency_score <- urgency_score + 1 }
  if(PAM >= 4) { urgency_score <- urgency_score + 1 }
  if(Medication_Adherence >= 75) { urgency_score <- urgency_score + 1 }
  if(Avg_Steps >= 10000) { urgency_score <- urgency_score + 1 }
  return(urgency_score)
}

# Apply the urgency criteria to each patient
patients_summary$Urgency_Score <- mapply(calculate_urgency,
                                         patients_summary$PHQ_9_Score_Numeric,
                                         patients_summary$PAM_Score_Numeric,
                                         patients_summary$Medication_Adherence_Percent,
                                         patients_summary$Avg_Steps_per_day_Numeric)

final_priority_list <- patients_summary %>%
  dplyr::arrange(desc(Urgency_Score)) %>%
  dplyr::select(Name, Urgency_Score) # Use explicit dplyr namespace if necessary

# Display the final list
print(final_priority_list)
```

```
## # A tibble: 10 x 2
##    Name           Urgency_Score
##    <chr>                  <dbl>
##  1 Albert Test               2
##  2 Jenny Test                2
##  3 John Test                 2
##  4 Michelle Test             2
##  5 Ryan Test                 2
##  6 Susan Test                2
##  7 Betty Test                1
##  8 Joe Test                  1
##  9 Bob Test                  0
## 10 Nancy Test                0
```

Based on the urgency criteria derived from the targets provided (PHQ-9 score, PAM score, Medication Adherence, and Steps per day), this is a prioritized list of patients for outreach, from most urgent to least urgent:

Albert Test - Urgency Score: 2 Jenny Test - Urgency Score: 2 John Test - Urgency Score: 2 Michelle Test - Urgency Score: 2 Ryan Test - Urgency Score: 2 Susan Test - Urgency Score: 2 Betty Test - Urgency Score: 1 Joe Test - Urgency Score: 1 Bob Test - Urgency Score: 0 Nancy Test - Urgency Score: 0

The "Urgency Score" indicates the number of target criteria met, with a higher score suggesting a lesser need for immediate outreach based on the targets set (contrary to typical urgency scoring, where a higher

score would indicate greater urgency). Here, patients with higher scores met more of the positive targets (e.g., better medication adherence or higher activity levels), implying they may be managing their conditions more effectively or have less acute needs than others.

I used a simple scoring system that evaluated each patient against four key health and treatment engagement targets.

For this, as per our instructions, For PHQ-9 Score, Target was set for scores less than 5, indicating minimal depression symptoms. This was evaluated at the latest (end) point in the dataset for each patient. Meeting this criterion suggests the patient might not urgently need mental health support, thus not contributing to urgency. For Patient Activation Measure (PAM) Score,a target of 4 or higher was set, with higher scores indicating better patient engagement and understanding of their health condition. This was also evaluated at the latest data point. Meeting or exceeding this target suggests a lower urgency for additional support to improve health engagement.

The medication adherence percentage target was set at 75% or higher, indicating good adherence to prescribed medication regimens. This was calculated as the average over the dataset, considering total doses and self-reported adherence. Higher adherence suggests better management of conditions, reducing the urgency for outreach.

A target of 10,000 steps or higher was set as an indicator of physical activity level, averaged over the treatment period. Meeting this target suggests the patient is active, potentially indicating better physical health and lower urgency for immediate intervention.

Each patient was assessed against these targets, receiving a point for each target met. The total points formed the Urgency Score, with a higher score indicating that a patient met more of the positive health behavior targets.

Tailored Recommendations: Based on the scoring system and the analysis of each patient's data:

Patients with Lower Scores (Higher Urgency): These patients did not meet several of the set targets, suggesting areas where they may need additional support. Recommendations for these patients could include more frequent follow-up calls or visits, targeted interventions to address specific areas of need (such as mental health support for those with higher PHQ-9 scores), and personalized health coaching to improve medication adherence and physical activity levels.

Patients with Higher Scores (Lower Urgency): These patients met most or all of the targets, indicating they are managing their health conditions well. While they might not require immediate intervention, maintaining engagement is important. Recommendations could include regular check-ins to ensure continued adherence to treatment plans, encouragement to maintain or increase physical activity levels, and educational resources to further their health knowledge and self-management capabilities.

Additional Data Points Needed: Social Determinants of Health (SDOH): Incorporating data on social determinants such as housing stability, access to nutritious food, and social support can provide a fuller picture of patient needs and barriers to health.

Clinical History: More detailed clinical histories, including diagnosis dates, treatment changes, and hospitalization records, can help in understanding the patient's health journey and potential risks.

Behavioral Health Data: Beyond PHQ-9, incorporating data from other behavioral health assessments can offer insights into anxiety, stress, and other mental health concerns.

Patient Feedback: Data on patient satisfaction, preferences for types of contact, and feedback on interventions could guide personalized care and engagement strategies.

Engagement Metrics: Detailed metrics on patient engagement with digital tools, response times, and interaction frequencies could help in refining outreach efforts.

Tools for Enrichment: Digital Health Platforms: Implementing or integrating with digital health platforms that offer telehealth, remote monitoring, and health education could enhance patient engagement and data collection.

Mobile Health Applications: Apps that patients can use to track their health metrics, receive reminders, and access health information can support self-management and provide valuable data back to healthcare providers.

Wearable Devices: Encouraging the use of wearable devices for tracking physical activity, sleep patterns, and even physiological metrics like heart rate could enrich the dataset with real-time health data.

AI and Machine Learning Models: Utilizing AI to analyze the enriched dataset can uncover patterns, predict risks, and personalize patient outreach efforts more effectively.

Other Considerations: Privacy and Security: Any expansion of data collection and new tools must prioritize patient privacy and data security, adhering to regulations such as HIPAA.

Interoperability: Ensuring that new tools and data sources are interoperable with existing systems is crucial for a seamless integration and for maintaining data integrity.

User Experience: Both for patients and healthcare providers, the ease of use and accessibility of the dashboard and associated tools should be a key consideration to encourage adoption and ongoing engagement.

Evaluation and Adaptation: Regular evaluation of the dashboard's impact on patient outcomes and healthcare operations should inform continuous improvement and adaptation of the toolset and strategies.

###Predictive modeling

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```r
library(dplyr)

# Assuming 'TypeOfContact', 'Medication_Adherence_Percent', and 'Avg_Steps_per_day_Numeric' as predicto
# Ensuring outcome 'PHQ9Score' and predictors are correctly formatted as numeric
data$PHQ9Score <- as.numeric(as.character(data$PHQ9Score))
data$Medication_Adherence_Percent <- as.numeric(data$Medication_Adherence_Percent)
data$Avg_Steps_per_day_Numeric <- as.numeric(data$Avg_Steps_per_day_Numeric)

# Prepare data for modeling
set.seed(123) # For reproducibility
trainingIndex <- createDataPartition(data$PHQ9Score, p = .8, list = FALSE, times = 1)
trainingData <- data[trainingIndex, ]
testData <- data[-trainingIndex, ]

# Fit a linear model
model <- train(PHQ9Score ~ TypeOfContact + Medication_Adherence_Percent + Avg_Steps_per_day_Numeric, dat
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```r
# Summarize the model fit
summary(model)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min     1Q  Median     3Q     Max
## -9.3821 -2.0346  0.4107  1.6662  4.3341
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    25.2051380  1.5135938  16.653  < 2e-16 ***
## TypeOfContactGroup             -6.1400425  1.6458103  -3.731 0.000426 ***
## TypeOfContactHome              -2.4715106  1.9929673  -1.240 0.219760
## 'TypeOfContactIn person at clinic' -5.0500861  0.8610010  -5.865 2.06e-07 ***
## TypeOfContactPhone             -4.6016428  1.0157003  -4.531 2.86e-05 ***
## Medication_Adherence_Percent    0.6729738  0.4943595   1.361 0.178508
## Avg_Steps_per_day_Numeric      -0.0009879  0.0001592  -6.206 5.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.725 on 60 degrees of freedom
## Multiple R-squared:  0.6795, Adjusted R-squared:  0.6475
## F-statistic: 21.21 on 6 and 60 DF,  p-value: 3.519e-13
```

```r
# Predict on new data
predictions <- predict(model, testData)

# Compare predicted vs actual values
comparison <- data.frame(Actual = testData$PHQ9Score, Predicted = predictions)
head(comparison)
```

```
##   Actual Predicted
## 1     15  15.68414
## 2      9  11.21021
## 3     15  14.35895
## 4     18  16.79553
## 5     21  18.08750
## 6     11  10.76460
```

Coefficients: The intercept and coefficients for each type of contact (except for "Home," which is not statistically significant with a p-value of 0.219760) are statistically significant. This suggests that, relative to the

baseline category of "Digital" (implied due to its absence in the coefficients list), all other types of contact are associated with a decrease in PHQ-9 score, indicating lower levels of reported depression symptoms. Group Contact reduces the PHQ-9 score by approximately 6.14 points. In-person at Clinic reduces the PHQ-9 score by about 5.05 points. Phone Contact reduces the PHQ-9 score by around 4.60 points. Medication Adherence Percent has a positive coefficient, suggesting that higher medication adherence is associated with higher PHQ-9 scores, although this relationship is not statistically significant (p-value: 0.178508). Average Steps per Day: There's a negative association with PHQ-9 scores, significantly indicating that more physical activity (as measured by steps) is associated with lower depression scores.

Model Diagnostics: Residuals: The residuals' range and interquartile range suggest some variability around the predicted values but no extreme outliers. Multiple R-squared (0.6795) and Adjusted R-squared (0.6475): These indicate that approximately 67.95% of the variability in PHQ-9 scores is explained by your model, which is a strong model fit for social science data. F-statistic: The very low p-value (3.519e-13) for the F-statistic indicates that your model is statistically significant and that at least one of the predictors has a non-zero coefficient.

Predictions: The actual vs. predicted values table shows that the model's predictions are reasonably close to the actual PHQ-9 scores, with some expected variability. This demonstrates the model's usefulness in estimating the severity of depression symptoms based on the type of contact, medication adherence, and physical activity levels.

Conclusions and Recommendations:

Type of Contact: Digital methods, being the baseline for comparison, appear to be associated with higher PHQ-9 scores. This means we have to consider strategies to improve digital engagement or provide additional support for patients engaged through digital means.

Physical Activity: Encouraging physical activity, as evidenced by the negative association with PHQ-9 scores, can be beneficial in managing depression symptoms.

Medication Adherence: While not statistically significant, the positive direction of the coefficient suggests exploring deeper into how medication adherence interacts with depression severity.

Future Directions: Further research could explore the non-significant relationship between medication adherence percentages and PHQ-9 scores. Additionally, investigating the reasons behind the effectiveness of non-digital contact methods might provide insights into improving patient engagement strategies.

This analysis underlines the importance of personalized patient engagement strategies, considering the significant impact of different engagement methods on depression symptom severity.

————————————————END————————————————————————