A Data Analytics Perspective on Hospital Readmissions

Project2 Report

Data Analysis and Visualization

AUTHOR
ViZiT

AFFILIATION
College of Information Science, University of Arizona

Summary:

Hospital readmissions have become a major concern in the healthcare industry in recent years. When patients are discharged from the hospital, they are expected to recover fully and not require readmission. However, some patients end up being readmitted, which can be expensive and negatively impact their health. This report will explore the causes and consequences of hospital readmissions and the measures that can be taken to prevent them. A better understanding of the differences between people requiring hospitalization may translate into more effective ways to prevent readmissions.

```
knitr::opts_chunk$set(echo = TRUE)
```

Description of the Dataset

The dataset represents ten years of clinical care at 130 US hospitals and integrated delivery networks.This report analyzes a dataset of 25,000 hospital records to identify risk factors associated with readmission. The study examined 17 variables in each record, with the primary outcome being readmission for any cause.

Information about the variables:

- “age” - age bracket of the patient
- “time_in_hospital” - days (from 1 to 14)
- “n_procedures” - number of procedures performed during the hospital stay
- “n_lab_procedures” - number of laboratory procedures performed during the hospital stay
- “n_medications” - number of medications administered during the hospital stay
- “n_outpatient” - number of outpatient visits in the year before a hospital stay
- “n_inpatient” - number of inpatient visits in the year before the hospital stay
- “n_emergency” - number of visits to the emergency room in the year before the hospital stay
- “medical_specialty” - the specialty of the admitting physician
- “diag_1” - primary diagnosis (Circulatory, Respiratory, Digestive, etc.)
- “diag_2” - secondary diagnosis
- “diag_3” - additional secondary diagnosis

- “glucose_test” - whether the glucose serum came out as high (> 200), normal, or not performed
- “A1Ctest” - whether the A1C level of the patient came out as high (> 7%), normal, or not performed
- “change” - whether there was a change in the diabetes medication (‘yes’ or ‘no’)
- “diabetes_med” - whether a diabetes medication was prescribed (‘yes’ or ‘no’)
- “readmitted” - if the patient was readmitted at the hospital (‘yes’ or ‘no’)

```
df <- read.csv("hospital_readmissions.csv")
head(df)
```

```
      age time_in_hospital n_lab_procedures n_procedures n_medications
1 [70-80)                8               72            1            18
2 [70-80)                3               34            2            13
3 [50-60)                5               45            0            18
4 [70-80)                2               36            0            12
5 [60-70)                1               42            0             7
6 [40-50)                2               51            0            10
  n_outpatient n_inpatient n_emergency medical_specialty      diag_1
1            2           0           0           Missing Circulatory
2            0           0           0             Other       Other
3            0           0           0           Missing Circulatory
4            1           0           0           Missing Circulatory
5            0           0           0  InternalMedicine       Other
6            0           0           0           Missing       Other
       diag_2      diag_3 glucose_test A1Ctest change diabetes_med readmitted
1 Respiratory       Other           no      no     no          yes         no
2       Other       Other           no      no     no          yes         no
3 Circulatory Circulatory           no      no    yes          yes        yes
4       Other    Diabetes           no      no    yes          yes        yes
5 Circulatory Respiratory           no      no     no          yes         no
6       Other       Other           no      no     no           no        yes
```

```
dim(df)
```

```
[1] 25000    17
```

Reasons for choosing the dataset

Hospital readmission is a problem in healthcare where patients are discharged from the hospital and then readmitted within a certain period of time, often within 30 days of their initial discharge. This is a costly and preventable problem that can negatively impact patients’ health outcomes and quality of life. The Centers for Medicare and Medicaid Services (CMS) implemented a Hospital Readmissions Reduction Program (HRRP) in 2012, which financially penalizes hospitals with higher-than-expected readmission rates for certain conditions. Causes of readmissions include inadequate care during initial hospitalization and poor discharge planning. Patients with chronic conditions, such as heart failure, diabetes, and respiratory disease, are at a particularly high risk of readmission. To reduce readmissions, interventions such as improved care coordination, enhanced patient

education, and medication management are implemented. Machine learning and artificial intelligence (AI) algorithms are also used to predict which patients are at the highest risk of readmission and enable healthcare providers to intervene proactively to prevent readmissions.

Checking for missing values:

```
        missing_count <- sapply(df, function(col) sum(is.na(col)))

        # Display the result
        missing_count
```

```
             age  time_in_hospital  n_lab_procedures      n_procedures
               0                 0                 0                 0
    n_medications      n_outpatient       n_inpatient       n_emergency
               0                 0                 0                 0
medical_specialty            diag_1            diag_2            diag_3
               0                 0                 0                 0
     glucose_test           A1Ctest            change      diabetes_med
               0                 0                 0                 0
       readmitted
               0
```

Unique Levels in each categorical column:

```
        char_columns <- sapply(df, is.character)


        for (col in names(df)[char_columns]) {
          cat("Unique values in column:", col, "\n")
          print(unique(df[[col]]))
          cat("\n")
        }
```

```
Unique values in column: age
[1] "[70-80)"  "[50-60)"  "[60-70)"  "[40-50)"  "[80-90)"  "[90-100)"

Unique values in column: medical_specialty
[1] "Missing"               "Other"                 "InternalMedicine"
[4] "Family/GeneralPractice" "Cardiology"            "Surgery"
[7] "Emergency/Trauma"

Unique values in column: diag_1
[1] "Circulatory"     "Other"           "Injury"          "Digestive"
[5] "Respiratory"     "Diabetes"        "Musculoskeletal" "Missing"

Unique values in column: diag_2
[1] "Respiratory"     "Other"           "Circulatory"     "Injury"
[5] "Diabetes"        "Digestive"       "Musculoskeletal" "Missing"

Unique values in column: diag_3
[1] "Other"           "Circulatory"     "Diabetes"        "Respiratory"
[5] "Injury"          "Musculoskeletal" "Digestive"       "Missing"
```

```
Unique values in column: glucose_test
[1] "no"     "normal" "high"

Unique values in column: A1Ctest
[1] "no"     "normal" "high"

Unique values in column: change
[1] "no"  "yes"

Unique values in column: diabetes_med
[1] "yes" "no"

Unique values in column: readmitted
[1] "no"  "yes"
```

Q1) What are the key demographic and clinical factors that influence hospital readmission rates?

Objective: To identify the most significant demographic (age, medical specialty etc) and clinical factors (length of stay, number of procedures, etc.) that contribute to hospital readmissions.

Fig 1. Countplot of Categorical Variables

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

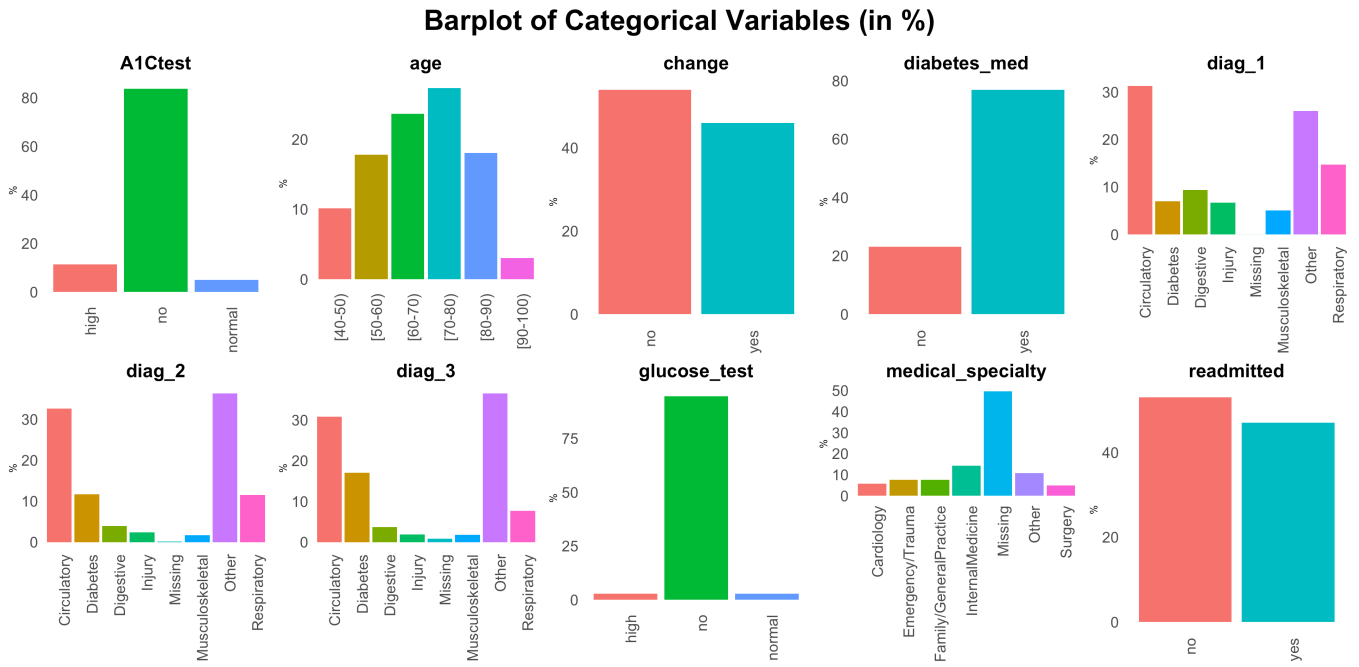
```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggpubr)
library(grid)
library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

```
# Reshape the dataframe to long format and calculate percentages
df_long <- df %>%
  select(where(is.character)) %>%  # Include all character-type columns, inclu
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value"
```

```
  group_by(variable, value) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(variable) %>%
  mutate(percentage = (count / sum(count)) * 100)  # Calculate percentages

# Generate individual plots for each variable
plots <- df_long %>%
  group_by(variable) %>%
  group_split() %>%
  lapply(function(sub_df) {
    ggplot(sub_df, aes(x = value, y = percentage, fill = value)) +
      geom_bar(stat = "identity") +  # Use percentages instead of raw counts
      theme_minimal() +
      theme(
        panel.grid = element_blank(),                # Remove gridlines
        axis.text.x = element_text(angle = 90, hjust = 1, size = 16),  # Rotat
        axis.text.y = element_text(size = 16),       # Enlarge y-axis labels
        legend.position = "none",                   # Remove legend
        plot.title = element_text(                  # Format title
          size = 20,                                # Smaller size
          face = "bold",                            # Bold text
          hjust = 0.5                               # Center title
        )
      ) +
      labs(
        title = unique(sub_df$variable),  # Add centered plot-specific title
        x = NULL,                         # Remove x-axis label
        y = "%"                           # Change y-axis label to %
      ) +
      scale_fill_manual(values = scales::hue_pal()(length(unique(sub_df$value)
  })

# Arrange the plots in a 5x2 grid
combined_plot <- ggarrange(
  plotlist = plots,
  ncol = 5, nrow = 2  # Set number of columns and rows
)

# Add a common title with spacing using gridExtra
final_plot <- grid.arrange(
  textGrob("Barplot of Categorical Variables (in %)",
           gp = gpar(fontface = "bold", fontsize = 30),  # Common title format
           hjust = 0.5),                                 # Center the title
  combined_plot,
  heights = c(0.1, 1)                                   # Adjust height ratio f
)
```




```
        # Display the final plot
        final_plot
```

```
TableGrob (2 x 1) "arrange": 2 grobs
  z     cells    name                grob
1 1 (1-1,1-1) arrange text[GRID.text.264]
2 2 (2-2,1-1) arrange      gtable[layout]
```

Important findings from Fig 1 :

1. Variables unrelated to diabetes:
    1. The most frequent primary, secondary, and other diagnoses were circulatory and other, with notable mentions being respiratory for primary diagnoses and diabetes and respiratory for secondary and tertiary diagnoses.
    2. The majority of patients were between 60 and 90 years old.
    3. Patient management was primarily categorized as missing, with internal medicine and other categories following in frequency.
    4. Around 47% of the patients are readmitted.

Variables related to diabetes:

1. Most A1C tests performed showed elevated values.
2. Few blood glucose tests were performed, with a similar proportion of elevated vs. normal results.
3. Nearly half of the patients had their diabetes medication changed.
4. Approximately 75% of patients were prescribed medication classified for diabetes.

Table 1. Summary of numerical variables:

```r
# Load required libraries
library(dplyr)
library(tidyr)
library(gt)

# Calculate summary statistics for all numeric columns in df
summary_table <- df %>%
  select(where(is.numeric)) %>%  # Select all numeric columns
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value"
  group_by(Variable) %>%  # Group by variable name
  summarise(
    mean = mean(Value, na.rm = TRUE),        # Calculate mean
    std = sd(Value, na.rm = TRUE),          # Calculate standard deviation
    min = min(Value, na.rm = TRUE),         # Minimum value
    `25%` = quantile(Value, 0.25, na.rm = TRUE), # 25th percentile
    `50%` = median(Value, na.rm = TRUE),    # Median (50th percentile)
    `75%` = quantile(Value, 0.75, na.rm = TRUE), # 75th percentile
    max = max(Value, na.rm = TRUE)          # Maximum value
  )

# Create a styled table using gt
styled_table <- summary_table %>%
  gt() %>%
  tab_header(
    title = "Summary Statistics for Numeric Columns",
    subtitle = "This table summarizes key statistics for each numeric variable
  ) %>%
  fmt_number(
    columns = c(mean, std, min, `25%`, `50%`, `75%`, max),
    decimals = 2  # Format numbers with 2 decimal places
  ) %>%
  cols_label(
    Variable = "Numeric Variable",
    mean = "Mean",
    std = "Std Dev",
    min = "Minimum",
    `25%` = "25th Percentile",
    `50%` = "Median",
    `75%` = "75th Percentile",
    max = "Maximum"
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_column_labels(everything())  # Bold column labels
  )

# Print the table
styled_table
```

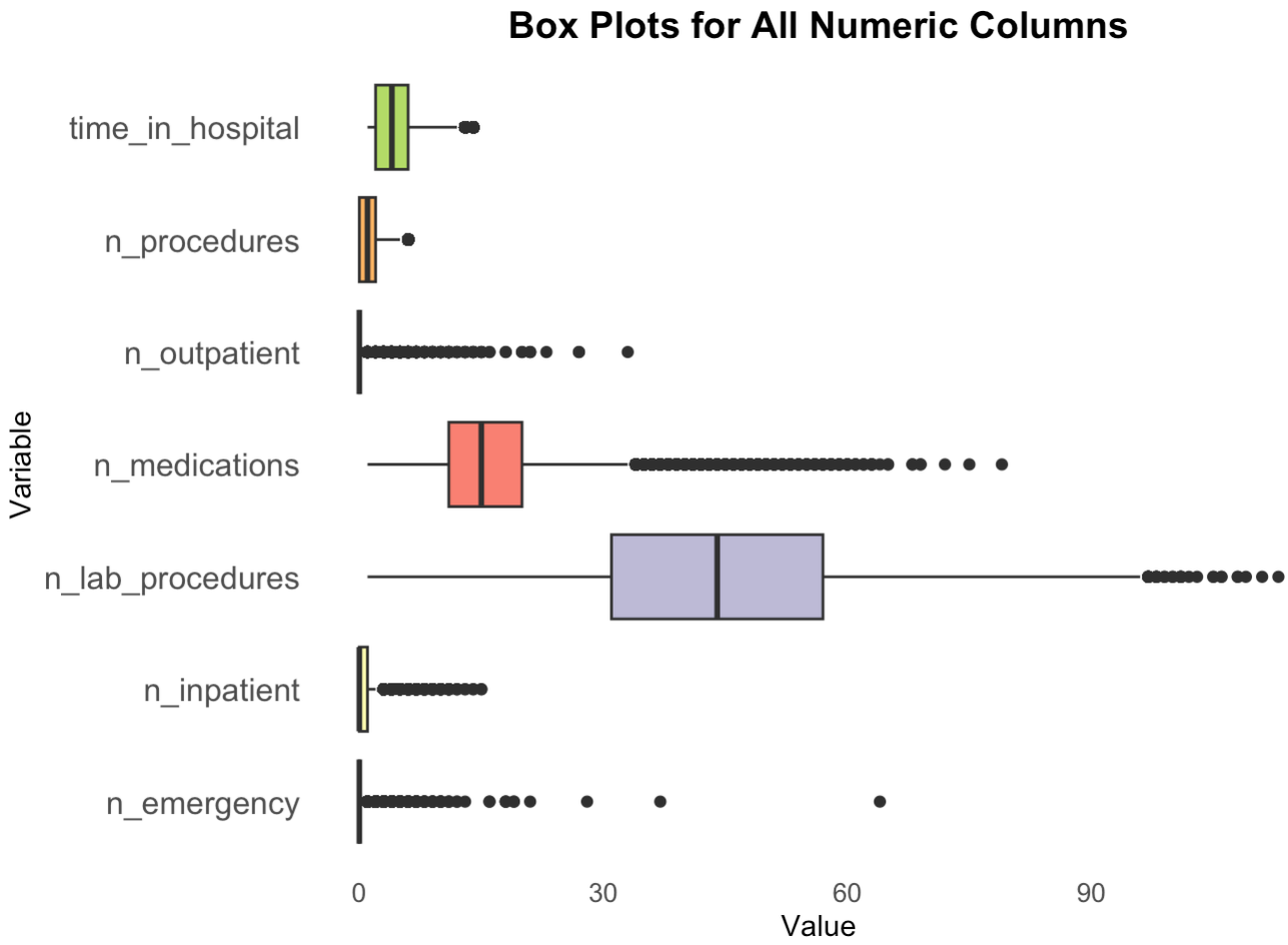
Summary Statistics for Numeric Columns

This table summarizes key statistics for each numeric variable in the dataset

Numeric Variable	Mean	Std Dev	Minimum	25th Percentile	Median	75th Percentile	Maximum
n_emergency	0.19	0.89	0.00	0.00	0.00	0.00	64.00
n_inpatient	0.62	1.18	0.00	0.00	0.00	1.00	15.00
n_lab_procedures	43.24	19.82	1.00	31.00	44.00	57.00	113.00
n_medications	16.25	8.06	1.00	11.00	15.00	20.00	79.00
n_outpatient	0.37	1.20	0.00	0.00	0.00	0.00	33.00
n_procedures	1.35	1.72	0.00	0.00	1.00	2.00	6.00
time_in_hospital	4.45	3.00	1.00	2.00	4.00	6.00	14.00

Fig 2. Boxplot of numerical variables:

```
# Load required libraries
library(ggplot2)
library(dplyr)
library(tidyr)

# Select numeric columns and reshape the dataframe to long format
df_long_numeric <- df %>%
  select(where(is.numeric)) %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value"

# Create a combined horizontal box plot
ggplot(df_long_numeric, aes(x = value, y = variable, fill = variable)) +
  geom_boxplot() +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 12),          # Adjust y-axis label size
    axis.text.x = element_text(size = 10),          # Adjust x-axis label size
    plot.title = element_text(size = 14,            # Title formatting
                              face = "bold",
                              hjust = 0.5),         # Center the title
    panel.grid.major = element_blank(),             # Remove major grid lines
    panel.grid.minor = element_blank(),             # Remove minor grid lines
    legend.position = "none"                        # Remove the legend
  ) +
  labs(
    title = "Box Plots for All Numeric Columns",    # Centered and bold title
    x = "Value",
    y = "Variable"
  ) +
  scale_fill_brewer(palette = "Set3")             # Add colors to boxes
```



Important findings from Table 1 & Fig 2:

1. Most variables showed outliers, with a higher frequency in medical services (inpatient, outpatient, emergency) and number of medications.
2. The average length of hospital stay was 4.45 days.
3. The average number of laboratory procedures was 43.24.
4. The average number of medications used was 16.25.

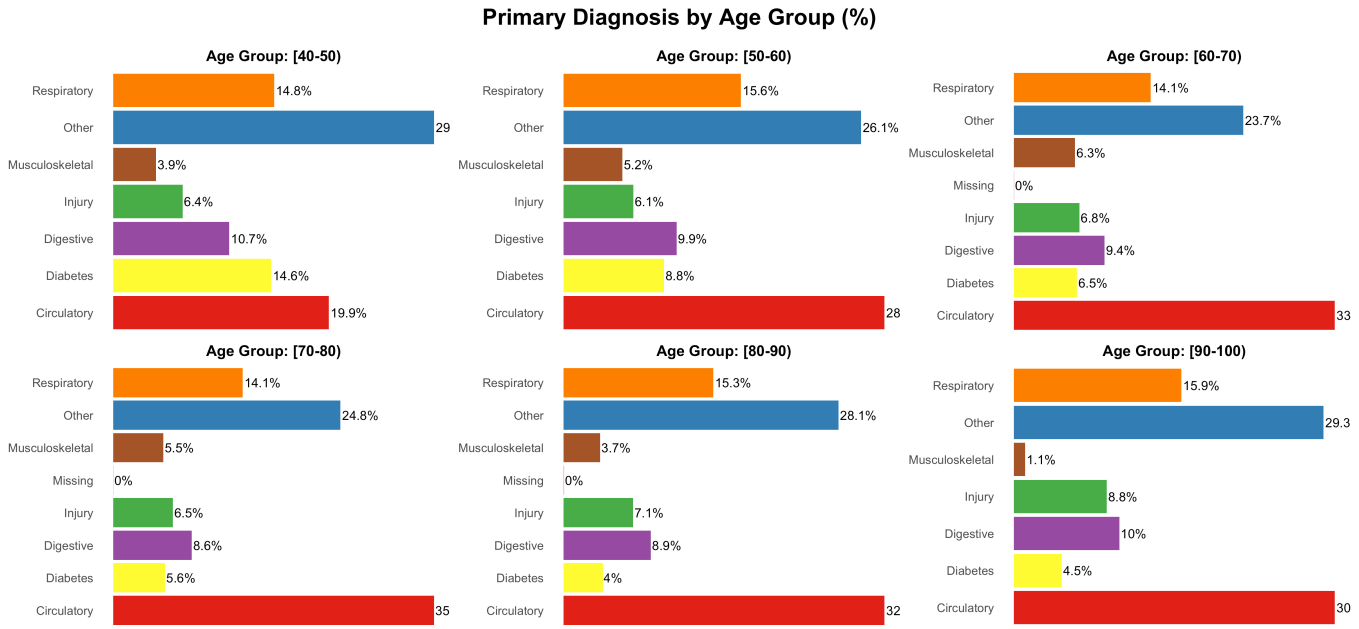
Fig 3. Primary Diagnosis by age group

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(gridExtra)

# Define a fixed color palette for diag_1 categories
unique_diag_1 <- unique(df$diag_1)  # Get unique categories of diag_1
color_palette <- setNames(RColorBrewer::brewer.pal(length(unique_diag_1), "Set

# Calculate percentages of diag_1 categories by age
df_percentage <- df %>%
  group_by(age, diag_1) %>%  # Group by age and diag_1
  summarise(count = n(), .groups = "drop") %>%  # Count occurrences
  group_by(age) %>%
  mutate(percentage = count / sum(count) * 100)  # Calculate percentage
```

```
# Create individual plots for each age group
plots <- df_percentage %>%
  group_split(age) %>%
  lapply(function(sub_df) {
    ggplot(sub_df, aes(x = percentage, y = diag_1, fill = diag_1)) +
      geom_bar(stat = "identity") +
      geom_text(aes(label = paste0(round(percentage, 1), "%")), hjust = -0.05,
      theme_minimal() +
      theme(
        panel.grid = element_blank(),                # Remove gridlines
        axis.text.x = element_blank(),               # Remove x-axis labels
        axis.text.y = element_text(size = 16),       # Enlarge y-axis labels
        axis.ticks.x = element_blank(),              # Remove x-axis ticks
        legend.position = "none",                   # Remove legend
        plot.title = element_text(                  # Format title
          size = 20,                                # Adjust size
          face = "bold",                            # Bold text
          hjust = 0.5                               # Center title
        )
      ) +
      labs(
        title = paste("Age Group:", unique(sub_df$age)),  # Add centered plot-
        x = NULL,                           # Remove x-axis label
        y = NULL                            # Remove y-axis label
      ) +
      scale_fill_manual(values = color_palette)  # Use consistent colors
  })
```

```
Warning: ... is ignored in group_split(<grouped_df>), please use group_by(..., .add =
TRUE) %>% group_split()
```

```
# Arrange the plots in a 4x2 grid without displaying
combined_plot <- arrangeGrob(
  grobs = plots,
  ncol = 3, nrow = 2  # Set number of columns and rows
)

# Add a common title with spacing (display only final plot)
final_plot <- grid.arrange(
  textGrob(
    "Primary Diagnosis by Age Group (%)",
    gp = gpar(fontface = "bold", fontsize = 30),  # Common title format
    hjust = 0.5                                  # Center the title
  ),
  combined_plot,
  heights = c(0.1, 1)                            # Adjust height ratio for tit
)
```




```
# Display only the final plot
final_plot
```

```
TableGrob (2 x 1) "arrange": 2 grobs
  z     cells    name                grob
1 1 (1-1,1-1) arrange text[GRID.text.512]
2 2 (2-2,1-1) arrange     gtable[arrange]
```

Important findings from Fig 3:

It is shown that the main diagnoses for all age groups are circulatory and others. We can classify the prevalence (in order) of diagnoses into three groups as follows:

- 40-50: other, circulatory, respiratory, diabetes, digestive.
- 50-60: circulatory, other, respiratory, digestive, diabetes.
- 60-100: circulatory, other, respiratory, digestive, trauma (wounds).

Conclusion:

We have a predominantly elderly population who are mostly admitted for circulatory and unspecified (other) causes. The majority of the classification regarding care was not recorded (missing), and the vast majority did not have glucose or hemoglobin A1C tests performed. The average number of laboratory procedures and medications per patient is high relative to the length of hospital stay. All of this may indicate that our population has high comorbidity and polypharmacy, which may justify the high readmission rates.

Q2) Some doctors believe diabetes might play a central role in readmission. Explore the effect of a diabetes diagnosis on readmission rates. Objective: To validate the hypothesis that a diabetes diagnosis significantly impacts readmission rates.

This question requires us to check if the doctors assumption that diabetes plays a central role in readmission is true. We will be running a hypothesis test to see if there is a relationship between a diabetes diagnosis and readmission rates.

Since the variables we are going to be comparing are categorical variables, the test we will be using will be a chi-square test for independence. Also, there are 3 diagnosis columns, so we will be conducting 3 seperate test to determine which of them has a significant effect on readmission rates.

Let's start by definig our **hypotheses:**

- **Null hypothesis (H0): There is no relationship between diabetes diagnosis and the readmission rates.**
- **Alternative hypothesis (Ha): There is a significant relationship between diabetes diagnosis and the readmission rates.**

Now, that we have defined the hypotheses, we will use the columns "diag_1", "diag_2", "diag_3" to create new columns called diabetes_diag1,diabetes_diag2 and diabetes_diag3 respectively to have "Diabetes" and "Not diabetes" values in them.

Lastly, let's set a significance level of 5%.

- If the p-value is < or = the significance level, we reject the null hypothesis and adopt the alternative hypothesis;

If the p-value is > the significance level, we retain the null hypothesis.

Table 2. Primary diagnosis vs readmission rates

```
library(dplyr)
library(gt)

# Step 1: Create the new column diabetes_diag1
df <- df %>%
  mutate(diabetes_diag1 = ifelse(diag_1 == "Diabetes", "Diabetes", "No Diabete

# Step 2: Compute counts for the table
summary_table <- df %>%
  group_by(diabetes_diag1, readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  pivot_wider(
    names_from = readmitted,
    values_from = count,
    names_prefix = "Readmitted: "
  ) %>%
  rename(
    `Diabetes Diagnosis` = diabetes_diag1,
    `Has Been Readmitted` = `Readmitted: yes`,
    `Not Been Readmitted` = `Readmitted: no`
  ) %>%
  replace(is.na(.), 0)  # Replace NAs with 0

# Step 3: Create a table using gt
```

```
gt_table <- summary_table %>%
  gt() %>%
  tab_header(
    title = "Counts of Patients by Primary Diabetes Diagnosis and Readmission"
    subtitle = "Readmission Status"
  ) %>%
  fmt_number(
    columns = c(`Has Been Readmitted`, `Not Been Readmitted`),
    decimals = 0  # Format numbers as integers
  ) %>%
  cols_label(
    `Diabetes Diagnosis` = "Diabetes or Not",
    `Has Been Readmitted` = "Has Been Readmitted",
    `Not Been Readmitted` = "Not Been Readmitted"
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_column_labels(everything())  # Bold column labels
  ) %>%
  tab_style(
    style = cell_text(weight = "bold", size = px(16)), # Make subtitle bold ar
    locations = cells_title(groups = "subtitle")      # Target the subtitle
  ) %>%
  tab_style(
    style = cell_text(size = px(16)),                 # Ensure column labels m
    locations = cells_column_labels(everything())    # Target all column label
  )

# Display the gt table
gt_table
```

Counts of Patients by Primary Diabetes Diagnosis and Readmission

	Readmission Status	
Diabetes or Not	**Not Been Readmitted**	**Has Been Readmitted**
Diabetes	810	937
No Diabetes	12,436	10,817

Chi- Square Test:

```
# Create a contingency table for diabetes_diag1 and readmitted
contingency_table <- table(df$diabetes_diag1, df$readmitted)

# Perform the chi-square test of independence
chi_square_result <- chisq.test(contingency_table)
```

```
# Print the results of the chi-square test
print("Chi-Square Test Result:")
```

```
[1] "Chi-Square Test Result:"
```

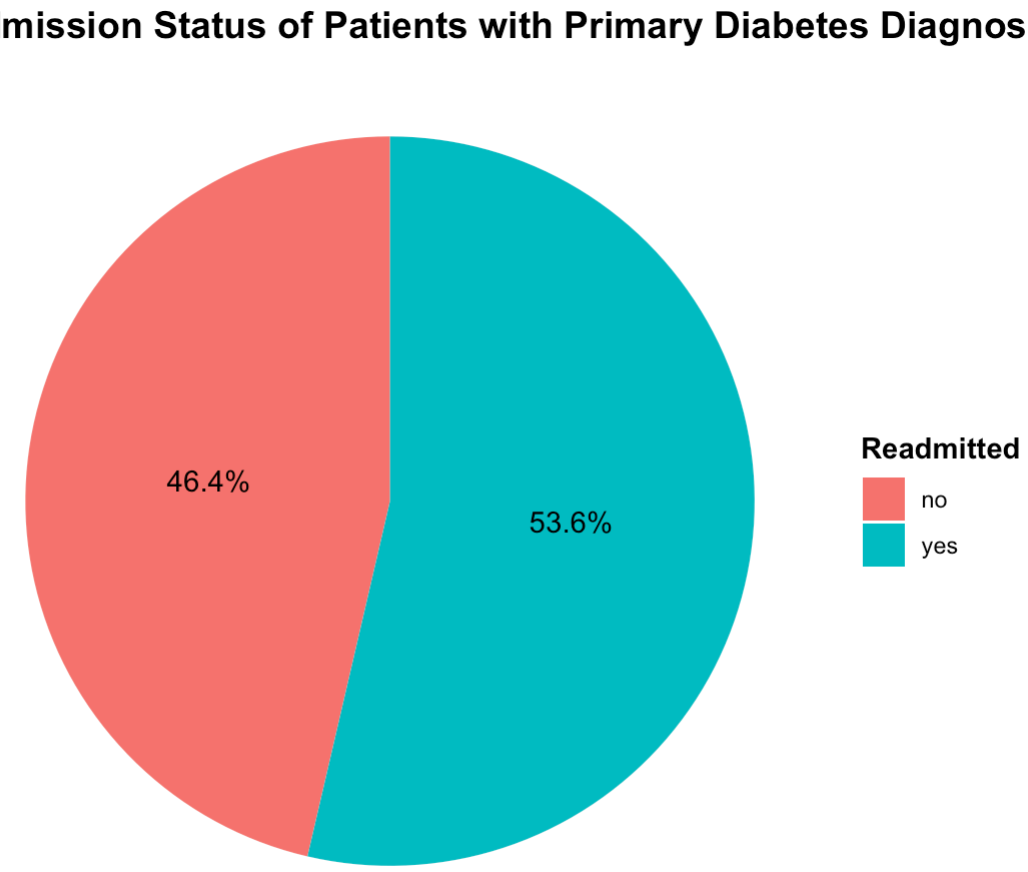
```
print(chi_square_result)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 32.746, df = 1, p-value = 1.05e-08
```

Here the p-value <0.05 and so we have to reject the null hypothesis and take on the alternative hypothesis. In this case, we can come to a conclusion that there is indeed a statistically significant relationship between diabetes diagnosis 1 and readmissions.

Fig 4. Readmission Status of Patients with Primary Diabetes Diagnosis

```
library(ggplot2)
library(dplyr)

# Filter data for diabetes_diag1 = "Diabetes"
diabetes_data <- df %>%
  filter(diabetes_diag1 == "Diabetes") %>%
  group_by(readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(percentage = count / sum(count) * 100)

# Create a pie chart
ggplot(diabetes_data, aes(x = "", y = percentage, fill = readmitted)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +  # Transform to pie chart
  labs(
    title = "Readmission Status of Patients with Primary Diabetes Diagnosis",
    fill = "Readmitted"
  ) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5)) +  # Add percentage labels
  theme_void() +  # Remove unnecessary plot elements
  theme(
    plot.title = element_text(hjust = 0.1, size = 14, face = "bold"),  # Bold
    legend.title = element_text(face = "bold")
  )
```

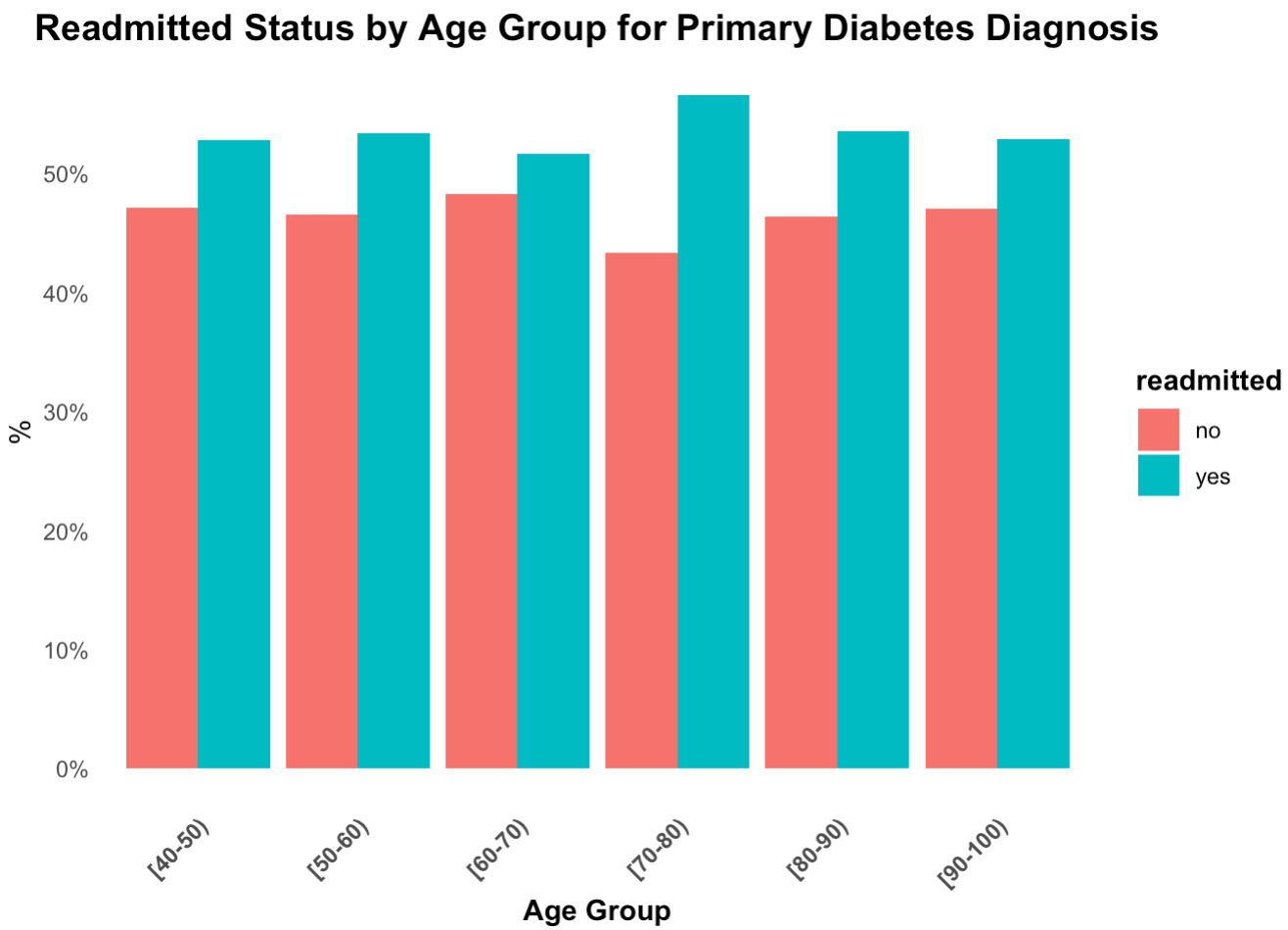


Around 53.6 % of patients with primary diabetes diagnosis are readmitted.

Fig 5. Readmitted Status by Age Group for Primary Diabetes Diagnosis

```
library(ggplot2)
library(dplyr)

# Filter data for diabetes_diag1 = "Diabetes"
bar_data <- df %>%
  filter(diabetes_diag1 == "Diabetes") %>%
  group_by(age, readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(age) %>%
  mutate(percentage = count / sum(count) * 100)

# Create the bar chart
ggplot(bar_data, aes(x = age, y = percentage, fill = readmitted)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(
    breaks = seq(0, 100, by = 10),  # Set Y-axis breaks at intervals of 10
    labels = function(x) paste0(x, "%")  # Add '%' symbol to labels
  ) +
  labs(
    title = "Readmitted Status by Age Group for Primary Diabetes Diagnosis",
    x = "Age Group",
    y = "%"
  ) +
  theme_minimal() +
```

```
  theme(
    panel.grid = element_blank(),  # Remove grid lines
    axis.line = element_blank(),   # Remove x and y axis lines
    axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),  # Bold
    axis.title.x = element_text(face = "bold"),  # Bold x-axis title
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # Cente
    legend.title = element_text(face = "bold")
  ) +
  scale_fill_manual(values = c("no" = "#F8766D", "yes" = "#00BFC4"))  # Defaul
```



The 70–80 age group with a primary diagnosis of diabetes appears to have the highest rate of hospital readmissions.

Table 3. Secondary diagnosis vs readmission rates

```
library(dplyr)
library(gt)

# Step 1: Create the new column diabetes_diag2
df <- df %>%
  mutate(diabetes_diag2 = ifelse(diag_2 == "Diabetes", "Diabetes", "No Diabete

# Step 2: Compute counts for the table
summary_table <- df %>%
  group_by(diabetes_diag2, readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  pivot_wider(
```

```
    names_from = readmitted,
    values_from = count,
    names_prefix = "Readmitted: "
  ) %>%
  rename(
    `Diabetes Diagnosis` = diabetes_diag2,
    `Has Been Readmitted` = `Readmitted: yes`,
    `Not Been Readmitted` = `Readmitted: no`
  ) %>%
  replace(is.na(.), 0)  # Replace NAs with 0

# Step 3: Create a table using gt
gt_table <- summary_table %>%
  gt() %>%
  tab_header(
    title = "Counts of Patients by Secondary Diabetes Diagnosis and Readmissic
    subtitle = "Readmission Status"
  ) %>%
  fmt_number(
    columns = c(`Has Been Readmitted`, `Not Been Readmitted`),
    decimals = 0  # Format numbers as integers
  ) %>%
  cols_label(
    `Diabetes Diagnosis` = "Diabetes or Not",
    `Has Been Readmitted` = "Has Been Readmitted",
    `Not Been Readmitted` = "Not Been Readmitted"
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_column_labels(everything())  # Bold column labels
  ) %>%
  tab_style(
    style = cell_text(weight = "bold", size = px(16)), # Make subtitle bold ar
    locations = cells_title(groups = "subtitle")      # Target the subtitle
  ) %>%
  tab_style(
    style = cell_text(size = px(16)),                 # Ensure column labels m
    locations = cells_column_labels(everything())    # Target all column label
  )

# Display the gt table
gt_table
```

Counts of Patients by Secondary Diabetes Diagnosis and Readmission

	Readmission Status	
Diabetes or Not	**Not Been Readmitted**	**Has Been Readmitted**
Diabetes	1,623	1,283

Counts of Patients by Secondary Diabetes Diagnosis and Readmission

	Readmission Status	
Diabetes or Not	**Not Been Readmitted**	**Has Been Readmitted**
No Diabetes	11,623	10,471

Chi- Square Test:

```
# Create the diabetes_diag2 column
df <- df %>%
  mutate(diabetes_diag2 = ifelse(diag_2 == "Diabetes", "Diabetes", "No Diabete

# Create a contingency table for diabetes_diag2 and readmitted
contingency_table_diag2 <- table(df$diabetes_diag2, df$readmitted)

# Perform the chi-square test of independence
chi_square_result_diag2 <- chisq.test(contingency_table_diag2)


# Print the results of the chi-square test
print("Chi-Square Test Result for diabetes_diag2 and readmitted:")
```

```
[1] "Chi-Square Test Result for diabetes_diag2 and readmitted:"
```

```
print(chi_square_result_diag2)
```

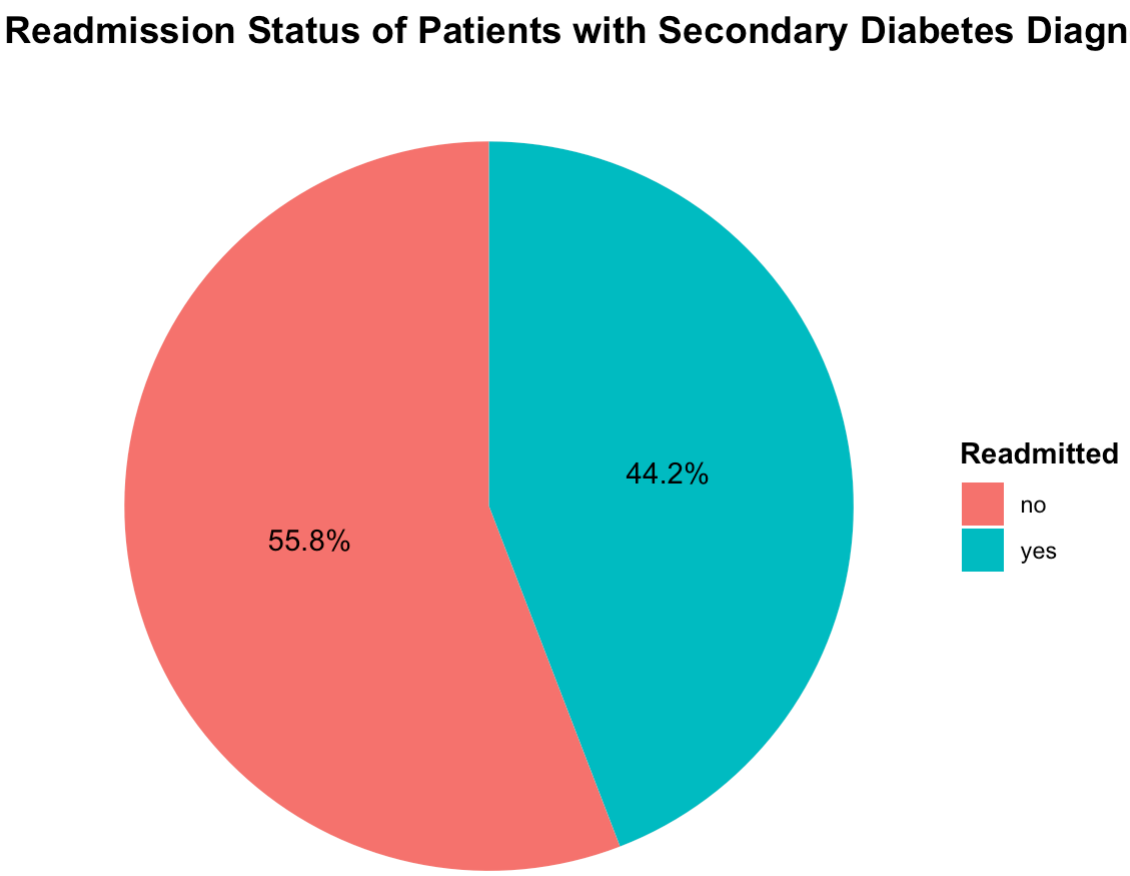
```
    Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table_diag2
X-squared = 10.712, df = 1, p-value = 0.001064
```

Here the p-value <0.05 and so we have to reject the null hypothesis and take on the alternative hypothesis. In this case, we can come to a conclusion that there is indeed a statistically significant relationship between diabetes diagnosis 2 and readmissions.

Fig 6. Readmission Status of Patients with Secondary Diabetes Diagnosis

```
library(ggplot2)
library(dplyr)

# Filter data for diabetes_diag2 = "Diabetes"
diabetes_data <- df %>%
  filter(diabetes_diag2 == "Diabetes") %>%
  group_by(readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(percentage = count / sum(count) * 100)

# Create a pie chart
ggplot(diabetes_data, aes(x = "", y = percentage, fill = readmitted)) +
  geom_bar(stat = "identity", width = 1) +
```

```
  coord_polar(theta = "y") +  # Transform to pie chart
  labs(
    title = "Readmission Status of Patients with Secondary Diabetes Diagnosis"
    fill = "Readmitted"
  ) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5)) +  # Add percentage labels
  theme_void() +  # Remove unnecessary plot elements
  theme(
    plot.title = element_text(hjust = 0.1, size = 14, face = "bold"),  # Bold
    legend.title = element_text(face = "bold")
  )
```

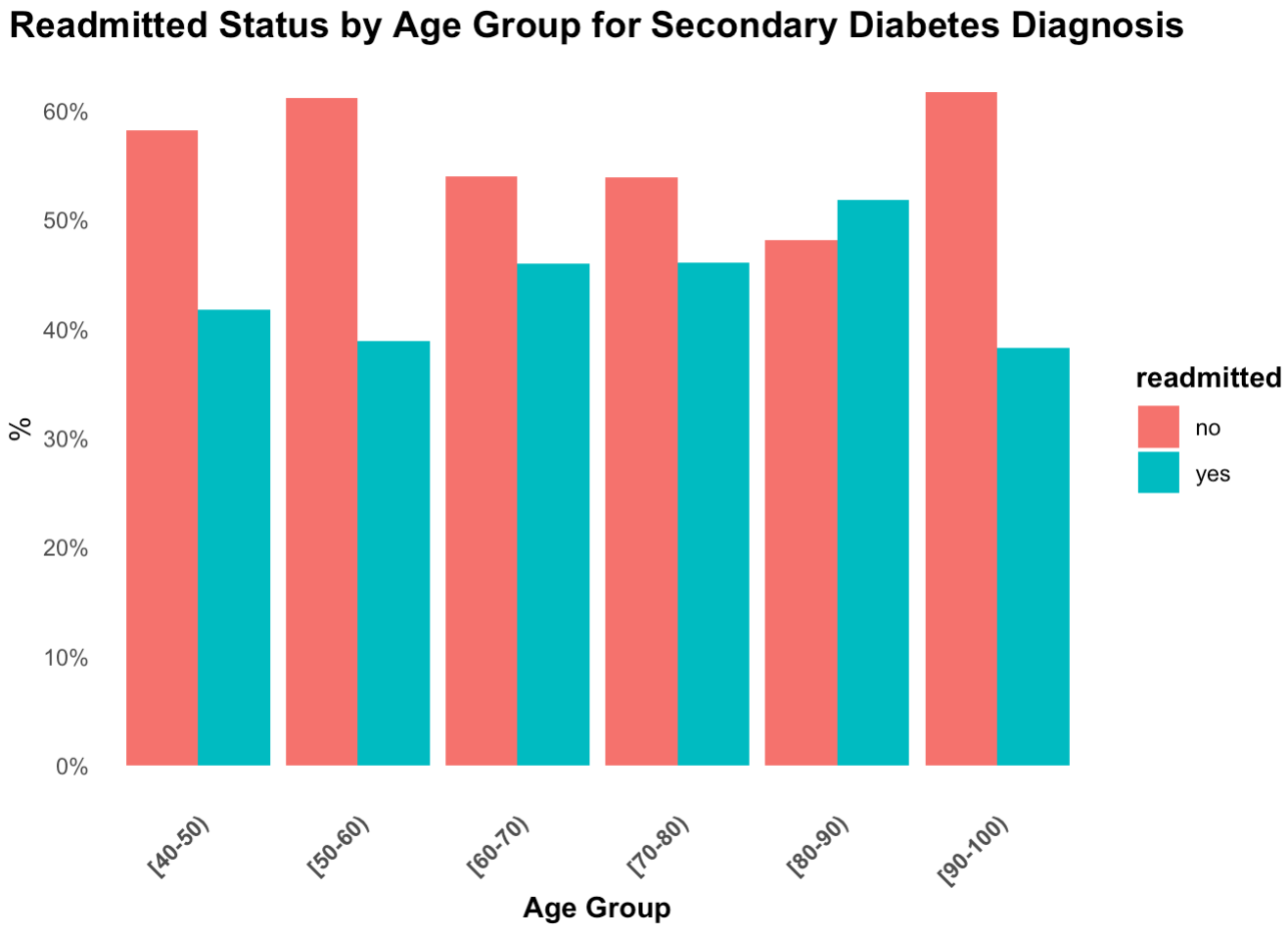


Around 44.2 % of patients with secondary diabetes diagnosis are readmitted which is less than patients with primary diabetes diagnosis.

Fig 7. Readmitted Status by Age Group for Secondary Diabetes Diagnosis

```
library(ggplot2)
library(dplyr)

# Filter data for diabetes_diag2 = "Diabetes"
bar_data <- df %>%
  filter(diabetes_diag2 == "Diabetes") %>%
  group_by(age, readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(age) %>%
```

```
  mutate(percentage = count / sum(count) * 100)

# Create the bar chart
ggplot(bar_data, aes(x = age, y = percentage, fill = readmitted)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(
    breaks = seq(0, 100, by = 10),  # Set Y-axis breaks at intervals of 10
    labels = function(x) paste0(x, "%")  # Add '%' symbol to labels
  ) +
  labs(
    title = "Readmitted Status by Age Group for Secondary Diabetes Diagnosis",
    x = "Age Group",
    y = "%"
  ) +
  theme_minimal() +
  theme(
    panel.grid = element_blank(),  # Remove grid lines
    axis.line = element_blank(),   # Remove x and y axis lines
    axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),  # Bold
    axis.title.x = element_text(face = "bold"),  # Bold x-axis title
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # Cente
    legend.title = element_text(face = "bold")
  ) +
  scale_fill_manual(values = c("no" = "#F8766D", "yes" = "#00BFC4"))  # Defaul
```



The 80–90 age group with a secondary diagnosis of diabetes appears to have the highest rate of hospital readmissions.

Table 4: Tertiary diagnosis and Readmission rates

```
library(dplyr)
library(gt)

# Step 1: Create the new column diabetes_diag3
df <- df %>%
  mutate(diabetes_diag3 = ifelse(diag_3 == "Diabetes", "Diabetes", "No Diabete

# Step 2: Compute counts for the table
summary_table <- df %>%
  group_by(diabetes_diag3, readmitted) %>%
  summarise(count = n(), .groups = "drop") %>%
  pivot_wider(
    names_from = readmitted,
    values_from = count,
    names_prefix = "Readmitted: "
  ) %>%
  rename(
    `Diabetes Diagnosis` = diabetes_diag3,
    `Has Been Readmitted` = `Readmitted: yes`,
    `Not Been Readmitted` = `Readmitted: no`
  ) %>%
  replace(is.na(.), 0)  # Replace NAs with 0

# Step 3: Create a table using gt
gt_table <- summary_table %>%
  gt() %>%
  tab_header(
    title = "Counts of Patients by Tertiary Diabetes Diagnosis and Readmissior
    subtitle = "Readmission Status"
  ) %>%
  fmt_number(
    columns = c(`Has Been Readmitted`, `Not Been Readmitted`),
    decimals = 0  # Format numbers as integers
  ) %>%
  cols_label(
    `Diabetes Diagnosis` = "Diabetes or Not",
    `Has Been Readmitted` = "Has Been Readmitted",
    `Not Been Readmitted` = "Not Been Readmitted"
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_column_labels(everything())  # Bold column labels
  ) %>%
  tab_style(
    style = cell_text(weight = "bold", size = px(16)), # Make subtitle bold ar
    locations = cells_title(groups = "subtitle")      # Target the subtitle
  ) %>%
  tab_style(
    style = cell_text(size = px(16)),                 # Ensure column labels r
    locations = cells_column_labels(everything())    # Target all column label
```

```
  )

# Display the gt table
gt_table
```

Counts of Patients by Tertiary Diabetes Diagnosis and Readmission		
	Readmission Status	
Diabetes or Not	**Not Been Readmitted**	**Has Been Readmitted**
Diabetes	2,314	1,947
No Diabetes	10,932	9,807

Chi-Square test

```
library(dplyr)

# Create the diabetes_diag3 column
df <- df %>%
  mutate(diabetes_diag3 = ifelse(diag_3 == "Diabetes", "Diabetes", "No Diabete

# Create a contingency table for diabetes_diag3 and readmitted
contingency_table_diag3 <- table(df$diabetes_diag3, df$readmitted)

# Perform the chi-square test of independence
chi_square_result_diag3 <- chisq.test(contingency_table_diag3)



# Print the results of the chi-square test
print("Chi-Square Test Result for diabetes_diag3 and readmitted:")
```

```
[1] "Chi-Square Test Result for diabetes_diag3 and readmitted:"
```

```
print(chi_square_result_diag3)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table_diag3
X-squared = 3.5426, df = 1, p-value = 0.05981
```

Here the p-value > 0.05 and so we fail to reject the null hypothesis. In this case, we can come to a conclusion that there is indeed no statistically significant relationship between diabetes diagnosis 3 and readmissions.

Conclusion:

A connection exists between hospital readmissions and patients with primary and secondary diabetes conditions.

Q3) How do prior healthcare utilization and patient characteristics influence readmission, and which groups should be prioritized for follow-up care?

Objective: To determine which factors (including previous healthcare visits, diagnoses, and treatments) predict hospital readmission, and identify high-risk groups that require focused follow-up.

Dropping diabetes_diag1, diabetes_diag2 & diabetes_diag3 from the main dataframe.

```
df <- df %>%
  select(-c(diabetes_diag1, diabetes_diag2, diabetes_diag3))
```

Converting categorical variables to factors.

```
df <- df %>%
  mutate(across(where(is.character), as.factor))
```

Univariate analysis using Logistic regression to find some risk and protective factors for readmission in our population.

Only significant variables with p< 0.05 with their 95% CI will be used.

```
library(dplyr)
library(broom)

# Step 1: Create an empty dataframe for significant results
df_sig <- data.frame(
  Variable = character(),
  Term = character(),
  OddsRatio = numeric(),
  CI_Lower = numeric(),
  CI_Upper = numeric(),
  PValue = numeric(),
  stringsAsFactors = FALSE
)

# Step 2: Logistic regression for numerical variables
numerical_vars <- df %>%
  select(where(is.numeric)) %>%
  names()

for (var in numerical_vars) {
  formula <- as.formula(paste("readmitted ~", var))
  model <- glm(formula, data = df, family = binomial)
  results <- tidy(model, conf.int = TRUE, exponentiate = TRUE) %>%  # Get odds
    filter(term != "(Intercept)" & p.value < 0.05)  # Exclude intercept and fi
  if (nrow(results) > 0) {
    results <- results %>%
      mutate(Variable = var) %>%
      select(Variable, term, estimate, conf.low, conf.high, p.value) %>%
      rename(
```

```
            Term = term,
            OddsRatio = estimate,
            CI_Lower = conf.low,
            CI_Upper = conf.high,
            PValue = p.value
          )
        df_sig <- bind_rows(df_sig, results)
      }
    }
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
    # Step 3: Logistic regression for categorical variables
    factor_vars <- df %>%
      select(where(is.factor)) %>%
      select(-readmitted) %>%
      names()

    for (var in factor_vars) {
      formula <- as.formula(paste("readmitted ~", var))
      model <- glm(formula, data = df, family = binomial)
      results <- tidy(model, conf.int = TRUE, exponentiate = TRUE) %>%
        filter(term != "(Intercept)" & p.value < 0.05)  # Exclude intercept and fi
      if (nrow(results) > 0) {
        results <- results %>%
          mutate(Variable = var) %>%
          select(Variable, term, estimate, conf.low, conf.high, p.value) %>%
          rename(
            Term = term,
            OddsRatio = estimate,
            CI_Lower = conf.low,
            CI_Upper = conf.high,
            PValue = p.value
          )
        df_sig <- bind_rows(df_sig, results)
      }
    }

    # Display the significant results dataframe
    print("Significant Results:")
```

```
[1] "Significant Results:"
```

```
    print(df_sig)
```

```
           Variable                        Term OddsRatio
1   time_in_hospital          time_in_hospital 1.0292027
2   n_lab_procedures          n_lab_procedures 1.0033436
3       n_procedures              n_procedures 0.9490850
4      n_medications             n_medications 1.0092075
5       n_outpatient              n_outpatient 1.2103202
6        n_inpatient               n_inpatient 1.5558509
7        n_emergency               n_emergency 1.4968294
```

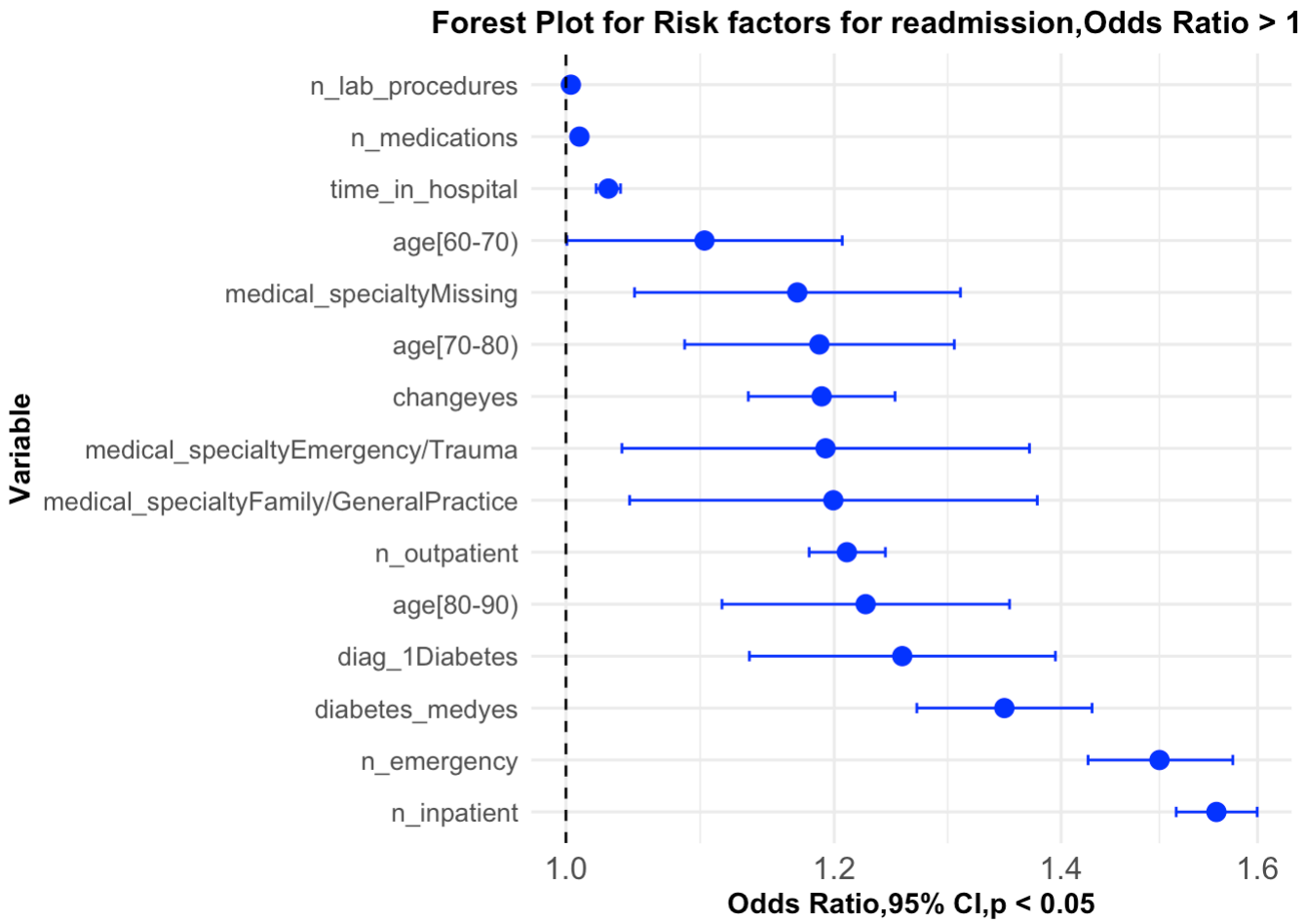
8	age	age[60-70)	1.0987219	
9	age	age[70-80)	1.1879177	
10	age	age[80-90)	1.2258673	
11	medical_specialty	medical_specialtyEmergency/Trauma	1.1929267	
12	medical_specialty	medical_specialtyFamily/GeneralPractice	1.1992363	
13	medical_specialty	medical_specialtyMissing	1.1702243	
14	medical_specialty	medical_specialtyOther	0.8664203	
15	diag_1	diag_1Diabetes	1.2567368	
16	diag_1	diag_1Injury	0.8411212	
17	diag_1	diag_1Musculoskeletal	0.7103937	
18	diag_1	diag_1Other	0.8932487	
19	diag_2	diag_2Diabetes	0.8447937	
20	diag_2	diag_2Digestive	0.8498074	
21	diag_2	diag_2Injury	0.7307127	
22	diag_3	diag_3Diabetes	0.9007878	
23	diag_3	diag_3Injury	0.7899050	
24	diag_3	diag_3Missing	0.4282328	
25	diag_3	diag_3Other	0.9380274	
26	glucose_test	glucose_testno	0.8117375	
27	A1Ctest	A1Ctestnormal	0.8554832	
28	change	changeyes	1.1897702	
29	diabetes_med	diabetes_medyes	1.3470565	

	CI_Lower	CI_Upper	PValue
1	1.0207186	1.0377642	9.543693e-12
2	1.0020851	1.0046048	1.883045e-07
3	0.9353276	0.9630161	2.185854e-12
4	1.0060998	1.0123289	5.839279e-09
5	1.1797175	1.2424331	2.644787e-47
6	1.5139042	1.5995617	1.257761e-217
7	1.4259203	1.5732744	3.440897e-58
8	1.0006615	1.2065629	4.855481e-02
9	1.0840134	1.3019896	2.292214e-04
10	1.1118355	1.3517858	4.400413e-05
11	1.0387754	1.3702241	1.251324e-02
12	1.0442239	1.3775357	1.013560e-02
13	1.0477307	1.3074704	5.388204e-03
14	0.7606695	0.9869921	3.091119e-02
15	1.1327253	1.3945769	1.648140e-05
16	0.7560130	0.9355067	1.451660e-03
17	0.6287549	0.8020113	3.612198e-08
18	0.8361916	0.9541700	7.996064e-04
19	0.7757672	0.9198247	1.036747e-04
20	0.7431831	0.9712016	1.709261e-02
21	0.6160435	0.8653587	2.934977e-04
22	0.8356193	0.9709812	6.368654e-03
23	0.6529771	0.9540072	1.468702e-02
24	0.3108047	0.5821206	1.103796e-07
25	0.8826948	0.9968238	3.916153e-02
26	0.6971553	0.9448708	7.137675e-03
27	0.7472575	0.9789656	2.346428e-02
28	1.1319119	1.2506041	8.455880e-12
29	1.2692804	1.4297716	1.032441e-22

Fig 8. Forest Plot of Odd Ratio >1

Risk factors for readmission were identified through univariate analysis using logistic regression the higher the factor the greater the likelihood of being readmitted.

```
library(ggplot2)

# Forest Plot for Odds Ratio > 1 (Sorted in Descending Order)
filtered_results_gt1 <- df_sig %>% filter(OddsRatio > 1) %>% arrange(desc(Odds
ggplot(filtered_results_gt1, aes(x = OddsRatio, y = reorder(Term, -OddsRatio))
  geom_point(size = 3, color = "blue") +  # Points are blue
  geom_errorbarh(aes(xmin = CI_Lower, xmax = CI_Upper), height = 0.2, color =
  geom_vline(xintercept = 1, linetype = "dashed", color = "black") +  # Refere
  scale_x_continuous(trans = "log10") +
  labs(
    title = "Forest Plot for Risk factors for readmission,Odds Ratio > 1",
    x = "Odds Ratio,95% CI,p < 0.05",
    y = "Variable"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.9, face = "bold", size = 12),  # Cente
    axis.title = element_text(face = "bold"),  # Bold axis titles
    axis.text.y = element_text(size = 10),  # Adjust size of variable names
    axis.text.x = element_text(size = 12)   # Adjust size of x-axis labels
  )
```



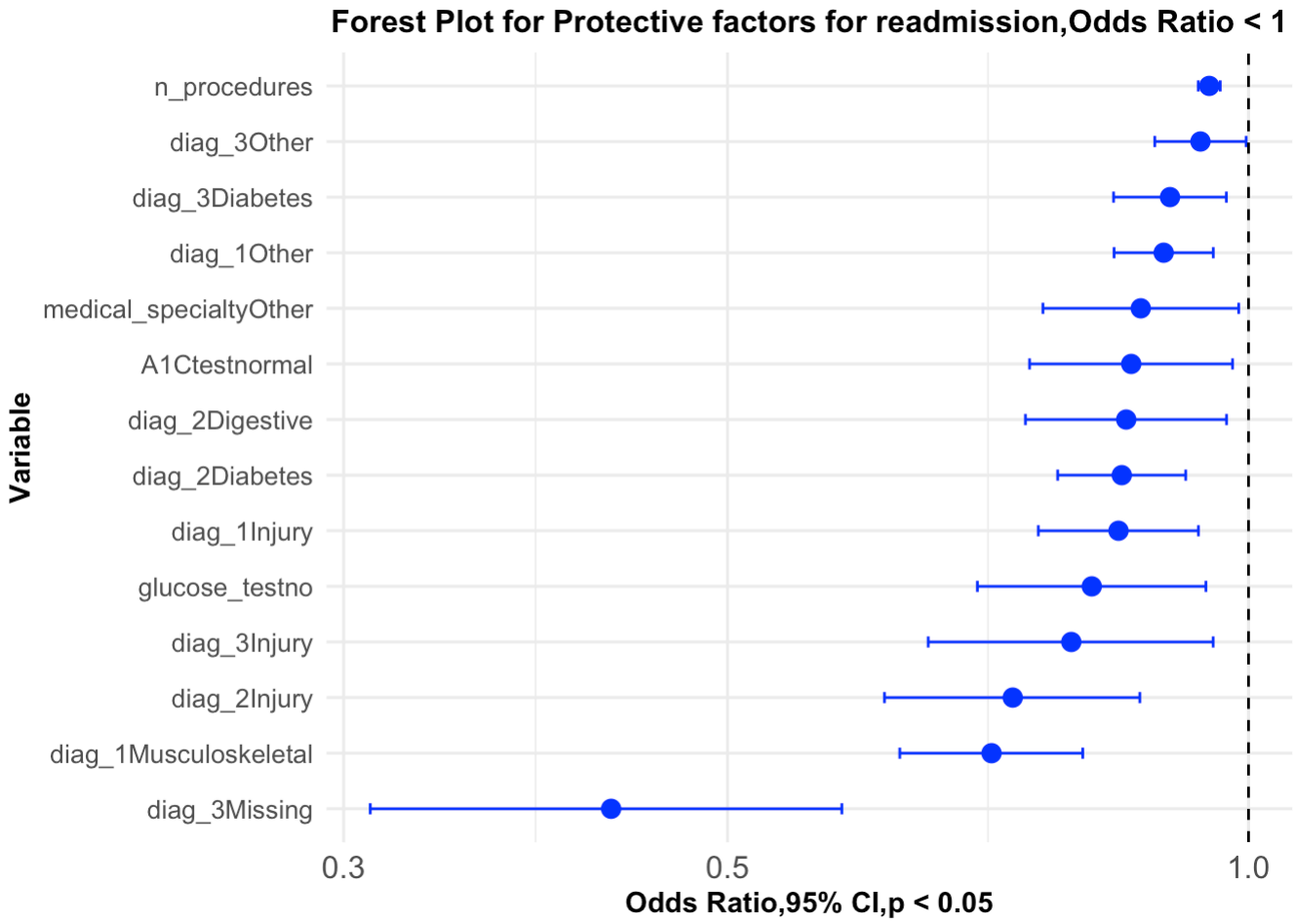
From Fig 8, we find that the use of healthcare services in the previous year, especially previous hospitalizations and number of visits to the emergency room , when a diabetes medication was

prescribed and when the primary diagnosis was diabetes are among the most important risk factors for patients getting readmitted.

Fig 9. Forest Plot of Odd Ratio <1

Protective factors for readmission were identified through univariate analysis using logistic regression the smaller the factor the greater the likelihood of not being readmitted.

```
# Forest Plot for Odds Ratio < 1 (Sorted in Ascending Order)
filtered_results_lt1 <- df_sig %>% filter(OddsRatio < 1) %>% arrange(OddsRatio
ggplot(filtered_results_lt1, aes(x = OddsRatio, y = reorder(Term, OddsRatio))) +
  geom_point(size = 3, color = "blue") +  # Points are blue
  geom_errorbarh(aes(xmin = CI_Lower, xmax = CI_Upper), height = 0.2, color = "
  geom_vline(xintercept = 1, linetype = "dashed", color = "black") +  # Reference
  scale_x_continuous(trans = "log10") +
  labs(
    title = "Forest Plot for Protective factors for readmission,Odds Ratio < 1",
    x = "Odds Ratio,95% CI,p < 0.05",
    y = "Variable"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 12),  # Center
    axis.title = element_text(face = "bold"),  # Bold axis titles
    axis.text.y = element_text(size = 10),  # Adjust size of variable names
    axis.text.x = element_text(size = 12)   # Adjust size of x-axis labels
  )
```



Regarding protective factors (Fig 9), the most notable was a missing tertiary diagnosis, which can be interpreted in two ways:

1. It may highlight a need to improve the classification of patients if the diagnosis was not accurately recorded.
2. Patients with only one or two diagnoses, indicating fewer comorbidities, likely faced a lower risk of readmission.

Other protective factors included having a primary musculoskeletal diagnosis and a secondary injury diagnosis.

Multivariate analysis using Logistic regression to find some risk and protective factors for readmission in our population.

```
df <- df %>%
  mutate(across(where(is.character), as.factor))
str(df)
```

```
'data.frame':   25000 obs. of  17 variables:
 $ age              : Factor w/ 6 levels "[40-50)","[50-60)",..: 4 4 2 4 3 1 2 3 5 4
...
 $ time_in_hospital : int  8 3 5 2 1 2 4 1 4 8 ...
 $ n_lab_procedures : int  72 34 45 36 42 51 44 19 67 37 ...
 $ n_procedures     : int  1 2 0 0 0 0 2 6 3 1 ...
 $ n_medications    : int  18 13 18 12 7 10 21 16 13 18 ...
 $ n_outpatient     : int  2 0 0 1 0 0 0 0 0 0 ...
 $ n_inpatient      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ n_emergency      : int  0 0 0 0 0 0 0 1 0 0 ...
 $ medical_specialty: Factor w/ 7 levels "Cardiology","Emergency/Trauma",..: 5 6 5 5 4
5 5 6 4 3 ...
 $ diag_1           : Factor w/ 8 levels "Circulatory",..: 1 7 1 1 7 7 4 1 3 8 ...
 $ diag_2           : Factor w/ 8 levels "Circulatory",..: 8 7 1 7 1 7 7 7 7 8 ...
 $ diag_3           : Factor w/ 8 levels "Circulatory",..: 7 7 1 2 8 7 7 7 7 7 ...
 $ glucose_test     : Factor w/ 3 levels "high","no","normal": 2 2 2 2 2 2 2 2 2 2 ...
 $ A1Ctest          : Factor w/ 3 levels "high","no","normal": 2 2 2 2 2 2 3 2 2 2 ...
 $ change           : Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 2 1 1 2 ...
 $ diabetes_med     : Factor w/ 2 levels "no","yes": 2 2 2 2 2 1 2 2 1 2 ...
 $ readmitted       : Factor w/ 2 levels "no","yes": 1 1 2 2 1 2 1 2 2 1 ...
```

```
# Convert readmitted to 1 for "yes" and 0 for "no" while keeping it as a factc
df <- df %>%
  mutate(readmitted = factor(ifelse(readmitted == "yes", 1, 0)))

# Verify the changes
cat("Updated readmitted variable:\n")
```

```
Updated readmitted variable:
```

```
print(levels(df$readmitted))
```

```
[1] "0" "1"
```

```
table(df$readmitted)
```

```
    0     1 
13246 11754 
```

Multivariate logistic regression

```
library(broom)

# Ensure `readmitted` is coded correctly as a binary factor
df <- df %>%
  mutate(readmitted = factor(readmitted, levels = c(0, 1)))  # 0 for no, 1 for

# Fit the multiple logistic regression model
model <- glm(readmitted ~ ., data = df, family = binomial)

# Summarize the model
summary(model)
```

```
Call:
glm(formula = readmitted ~ ., family = binomial, data = df)

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -0.6959890  0.1243553  -5.597 2.18e-08
age[50-60)                               0.0282194  0.0528270   0.534 0.593214
age[60-70)                               0.1375938  0.0508485   2.706 0.006811
age[70-80)                               0.2084921  0.0501655   4.156 3.24e-05
age[80-90)                               0.2140844  0.0537393   3.984 6.78e-05
age[90-100)                             -0.0686333  0.0884838  -0.776 0.437950
time_in_hospital                         0.0188921  0.0052373   3.607 0.000309
n_lab_procedures                         0.0016871  0.0007701   2.191 0.028465
n_procedures                            -0.0420680  0.0089816  -4.684 2.82e-06
n_medications                            0.0015125  0.0021306   0.710 0.477788
n_outpatient                             0.1162880  0.0132262   8.792  < 2e-16
n_inpatient                              0.3844127  0.0144816  26.545  < 2e-16
n_emergency                              0.2156650  0.0253555   8.506  < 2e-16
medical_specialtyEmergency/Trauma        0.0457440  0.0754141   0.607 0.544135
medical_specialtyFamily/GeneralPractice  0.0339589  0.0751408   0.452 0.651314
medical_specialtyInternalMedicine       -0.1659060  0.0677442  -2.449 0.014325
medical_specialtyMissing                 0.0286694  0.0604106   0.475 0.635090
medical_specialtyOther                  -0.1525642  0.0719832  -2.119 0.034053
medical_specialtySurgery                -0.2085732  0.0835378  -2.497 0.012534
diag_1Diabetes                           0.1703880  0.0589531   2.890 0.003850
diag_1Digestive                         -0.0063191  0.0523705  -0.121 0.903958
diag_1Injury                            -0.1821992  0.0587529  -3.101 0.001928
diag_1Missing                            0.1028459  1.0191327   0.101 0.919618
diag_1Musculoskeletal                   -0.1665560  0.0683757  -2.436 0.014855
diag_1Other                             -0.1543533  0.0383352  -4.026 5.66e-05
diag_1Respiratory                       -0.0414262  0.0435392  -0.951 0.341366
diag_2Diabetes                          -0.0326347  0.0475856  -0.686 0.492833
diag_2Digestive                         -0.1593199  0.0745184  -2.138 0.032517
```

```
diag_2Injury                                -0.1811418  0.0909207  -1.992 0.046338
diag_2Missing                                0.1808029  0.3352550   0.539 0.589680
diag_2Musculoskeletal                       -0.0157449  0.1065304  -0.148 0.882503
diag_2Other                                 -0.0793915  0.0338336  -2.347 0.018949
diag_2Respiratory                           -0.0686083  0.0460299  -1.491 0.136089
diag_3Diabetes                              -0.0361742  0.0407770  -0.887 0.375013
diag_3Digestive                             -0.0108316  0.0748685  -0.145 0.884968
diag_3Injury                                -0.1226001  0.1008465  -1.216 0.224096
diag_3Missing                               -0.6042856  0.1696425  -3.562 0.000368
diag_3Musculoskeletal                       -0.0808097  0.1021463  -0.791 0.428875
diag_3Other                                 -0.0837535  0.0329876  -2.539 0.011119
diag_3Respiratory                           -0.0077083  0.0533567  -0.144 0.885132
glucose_testno                              -0.0470871  0.0821866  -0.573 0.566693
glucose_testnormal                          -0.0301003  0.1126668  -0.267 0.789344
A1Ctestno                                    0.0644164  0.0435206   1.480 0.138838
A1Ctestnormal                               -0.1117039  0.0713567  -1.565 0.117482
changeyes                                    0.0289969  0.0313779   0.924 0.355423
diabetes_medyes                              0.2397106  0.0366213   6.546 5.92e-11

(Intercept)                                 ***
age[50-60)
age[60-70)                                  **
age[70-80)                                  ***
age[80-90)                                  ***
age[90-100)
time_in_hospital                            ***
n_lab_procedures                            *
n_procedures                                ***
n_medications
n_outpatient                                ***
n_inpatient                                 ***
n_emergency                                 ***
medical_specialtyEmergency/Trauma
medical_specialtyFamily/GeneralPractice
medical_specialtyInternalMedicine           *
medical_specialtyMissing
medical_specialtyOther                      *
medical_specialtySurgery                    *
diag_1Diabetes                              **
diag_1Digestive
diag_1Injury                                **
diag_1Missing
diag_1Musculoskeletal                       *
diag_1Other                                 ***
diag_1Respiratory
diag_2Diabetes
diag_2Digestive                             *
diag_2Injury                                *
diag_2Missing
diag_2Musculoskeletal
diag_2Other                                 *
diag_2Respiratory
diag_3Diabetes
diag_3Digestive
```

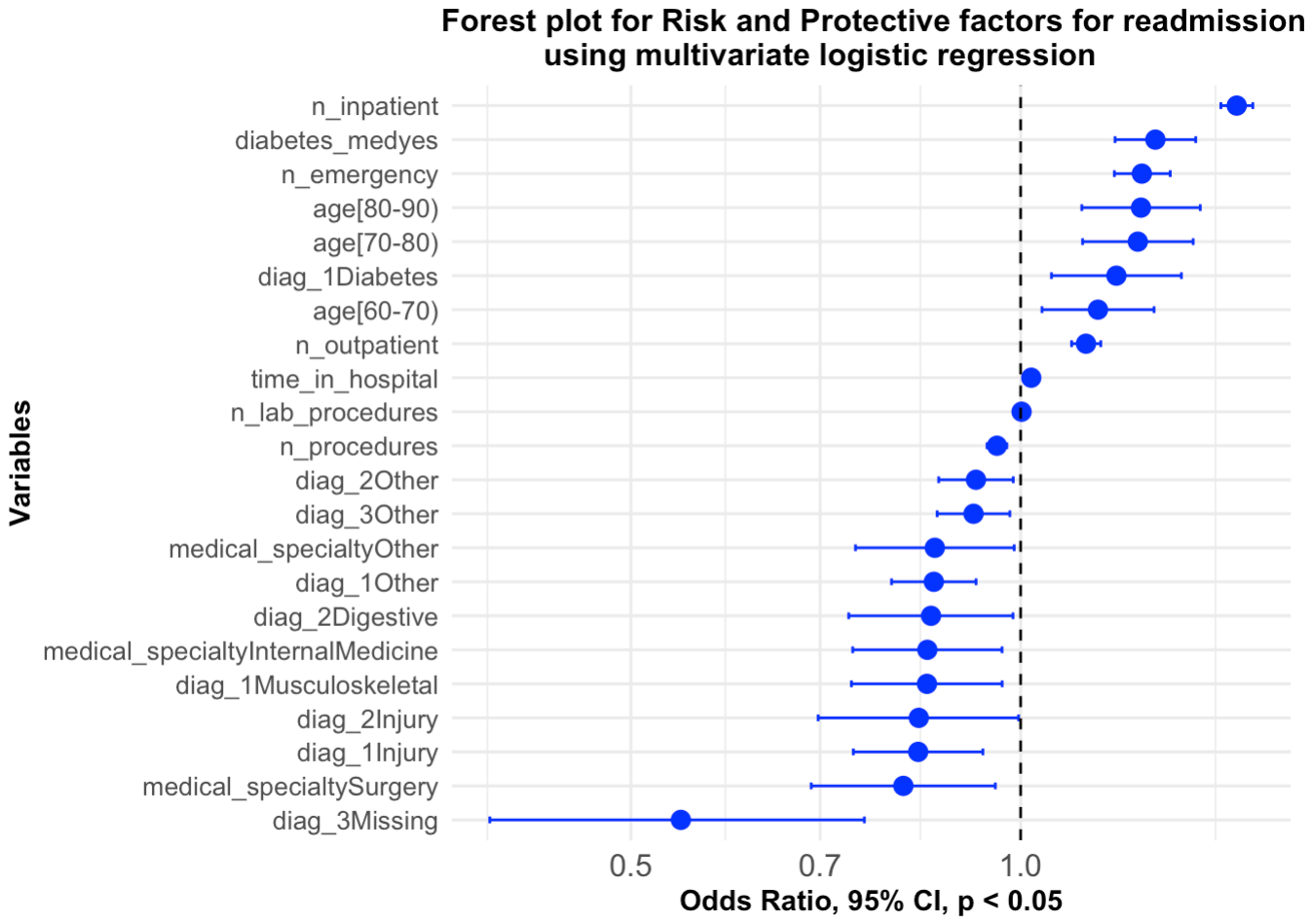
```
diag_3Injury
diag_3Missing                          ***
diag_3Musculoskeletal
diag_3Other                            *
diag_3Respiratory
glucose_testno
glucose_testnormal
A1Ctestno
A1Ctestnormal
changeyes
diabetes_medyes                        ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34568  on 24999  degrees of freedom
Residual deviance: 32772  on 24954  degrees of freedom
AIC: 32864

Number of Fisher Scoring iterations: 4
```

```
# Extract coefficients, p-values, odds ratios, and confidence intervals
results <- tidy(model, conf.int = TRUE, exponentiate = TRUE) %>%
  rename(
    Variable = term,
    OddsRatio = estimate,
    CI_Lower = conf.low,
    CI_Upper = conf.high,
    PValue = p.value
  )

# Display significant results (p < 0.05)
significant_results <- results %>% filter(PValue < 0.05)

#cat("Significant Variables (p < 0.05):\n")
#print(significant_results)
```

Fig 10. Forest plot for Risk and Protective factors for readmission using multivariate logistic regression

```
library(ggplot2)
library(dplyr)


# Check if there are significant variables
if (nrow(significant_results) == 0) {
  stop("No significant variables (p < 0.05) to plot.")
}

# Reorder variables for better visualization (sorted by Odds Ratio)
significant_results <- significant_results %>%
  arrange(OddsRatio) %>%
  mutate(Variable = reorder(Variable, OddsRatio))
```

```
# Create the forest plot
# Exclude the intercept from the significant results
filtered_results <- significant_results %>% filter(Variable != "(Intercept)")

# Create the forest plot
ggplot(filtered_results, aes(x = OddsRatio, y = Variable)) +
  geom_point(size = 3, color = "blue") +  # Points for odds ratios
  geom_errorbarh(aes(xmin = CI_Lower, xmax = CI_Upper), height = 0.2, color =
  geom_vline(xintercept = 1, linetype = "dashed", color = "black") +  # Refere
  scale_x_continuous(trans = "log10") +  # Log scale for odds ratios
  labs(
    title = "Forest plot for Risk and Protective factors for readmission \n us
    x = "Odds Ratio, 95% CI, p < 0.05",
    y = "Variables"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.3, face = "bold", size = 12),  # Cente
    axis.title = element_text(face = "bold"),  # Bold axis titles
    axis.text.y = element_text(size = 10),  # Adjust size of variable names
    axis.text.x = element_text(size = 12)   # Adjust size of x-axis labels
  )
```



The main **risk factors** are confirmed to be the use of healthcare services in the previous year, primary diagnosis of diabetes, and prescription of medications for diabetes.Ages 70-90 seem to be at a greater risk.

For the **protective factors**, we found the missing tertiary diagnosis (with the same considerations mentioned above) and the specialty of general surgery.

Conclusion:

- Readmission rate: 11,754 people (47.01%).
- The top 5 primary diagnoses for each age group can be classified by frequency in the following order:
    - (40-50): other, circulatory, respiratory, diabetes, digestive.
    - (50-60): circulatory, other, respiratory, digestive, diabetes.
    - (60-100): circulatory, other, respiratory, digestive, trauma (wounds).
- Statistically significant relationship between primary diabetes diagnosis and readmissions with 53.6% of people with primary diabetes diagnosis being readmitted.
- Statistically significant relationship between secondary diabetes diagnosis and readmissions with 44.2% of people with secondary diabetes diagnosis being readmitted.
- Patients with the following characteristics are more likely to be readmitted compared to the general population:
    - Patients with visits in the previous year to inpatient and emergency.
    - Patients with a primary diagnosis of diabetes and prescribed medication for diabetes at hospital discharge.
    - Ages 70-90 group; the ages 70-80 have a higher proportion of readmission given they have diabetes.
- Patients with the following characteristics are less likely to be readmitted:
    - Patients whose tertiary diagnosis is not there which might mean mean they have less co-morbidities.
    - Treated by internal medicine, surgery or other services.
    - Patients with a primary diagnosis of others, injury or musculoskeletal.