


# Understanding Diabetes Risk Factors

This presentation explores the associations between various health indicators and the prevalence of diabetes/prediabetes in U.S. adults. We will delve into the study design, data analysis, and the insights gained from our logistic regression model, highlighting key risk factors and their impact.

 by Ahmed Tarek



# Scientific Question and Study Design

Our core scientific question investigates the associations between risk factors like high blood pressure, BMI, and smoking, and the prevalence of diabetes/prediabetes in U.S. adults. We also examine how age and sex influence these associations.

## Study Design

We conducted a cross-sectional study, collecting data on diabetes status, risk factors, and demographics at a single time point. This approach allows us to estimate prevalence and associations, though it does not establish causality. The 50/50 split in the dataset biases prevalence but maintains valid odds ratios.

# Data Preprocessing and Visualization

Our data preprocessing pipeline involved meticulously loading the raw dataset and recoding key categorical variables into more descriptive factors, ensuring data quality and readability. Following this, we generated a series of insightful graphs to visualize crucial distributions and relationships within the data, providing a foundational understanding before advanced modeling.

## Load and Preprocess Data

Initial data handling included converting numerical codes into descriptive factors for essential variables such as ``Diabetes_binary`` (diabetes/prediabetes status),...

## Visualize Diabetes Status Distribution

A bar plot was generated to clearly illustrate the 50/50 split of the dataset based on diabetes status, confirming the deliberate balancing of the two groups for...

## Analyze BMI and Age Distributions

Histograms and comparative boxplots were used to visualize the distributions of Body Mass Index (BMI) and age. These revealed that individuals in the...

## Examine HighBP and Sex Proportions

A stacked bar plot was created to vividly illustrate the prevalence of high blood pressure (HighBP) across different sexes and diabetes statuses, highlighting ho...

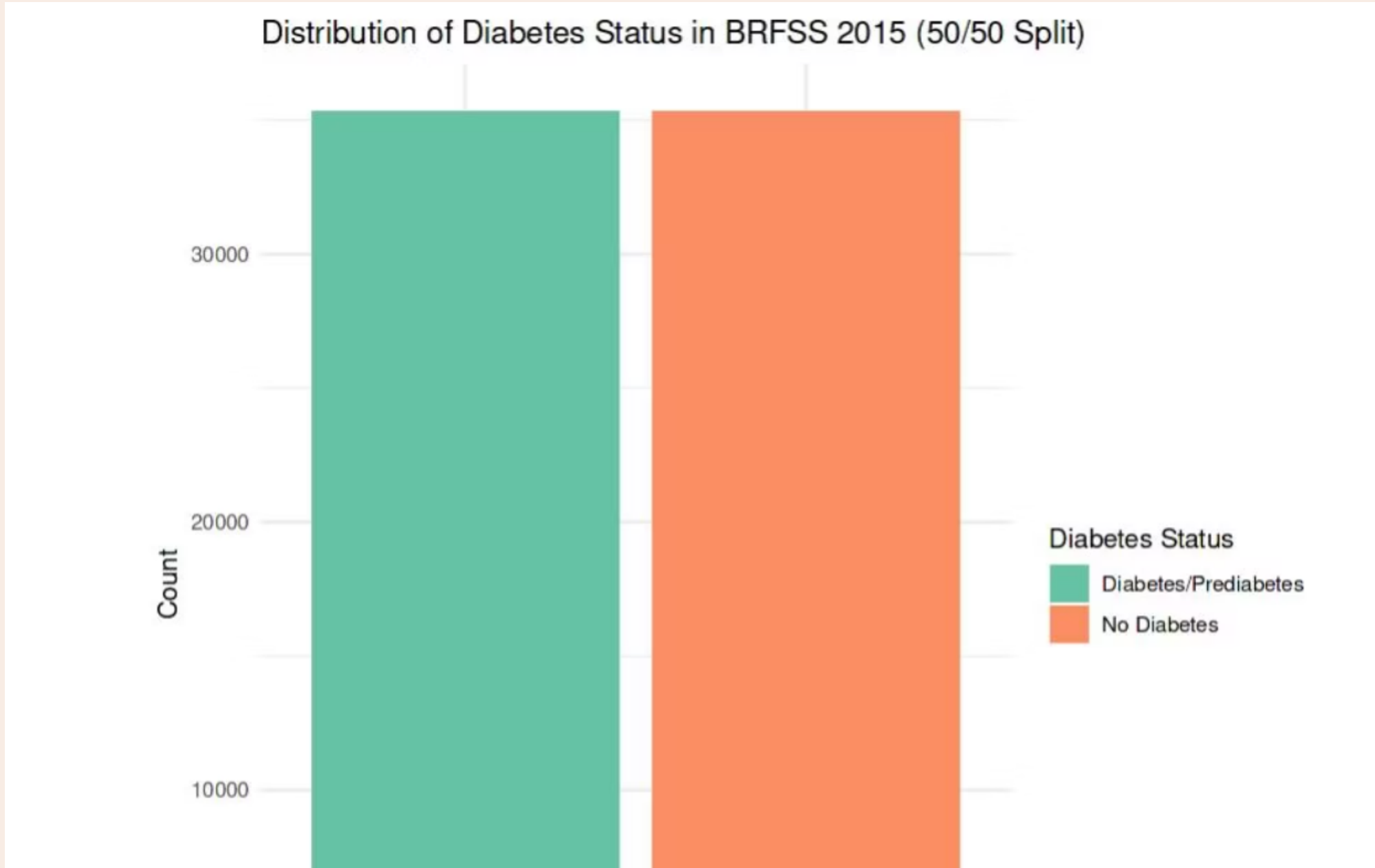
# Load and Preprocess Data

Initial data handling included converting numerical codes into descriptive factors for essential variables such as `Diabetes\_binary` (diabetes/prediabetes status), `Sex`, `HighBP` (high blood pressure), and `HighChol` (high cholesterol), preparing them for analysis.

```
# Step 2: Preprocess the data for visualization
cat("\n=== Data Preprocessing ===\n")
data <- data %>% mutate(  Diabetes_binary = as.factor(dplyr::recode(as.character(Diabetes_binary),
"0" = "No Diabetes",
"1" = "Diabetes/Prediabetes")),
Sex = as.factor(ifelse(Sex == 0, "Female", "Male")),
HighBP = as.factor(ifelse(HighBP == 0, "No High BP", "High BP")),
HighChol = as.factor(ifelse(HighChol == 0, "No High Chol", "High Chol"))  )
```

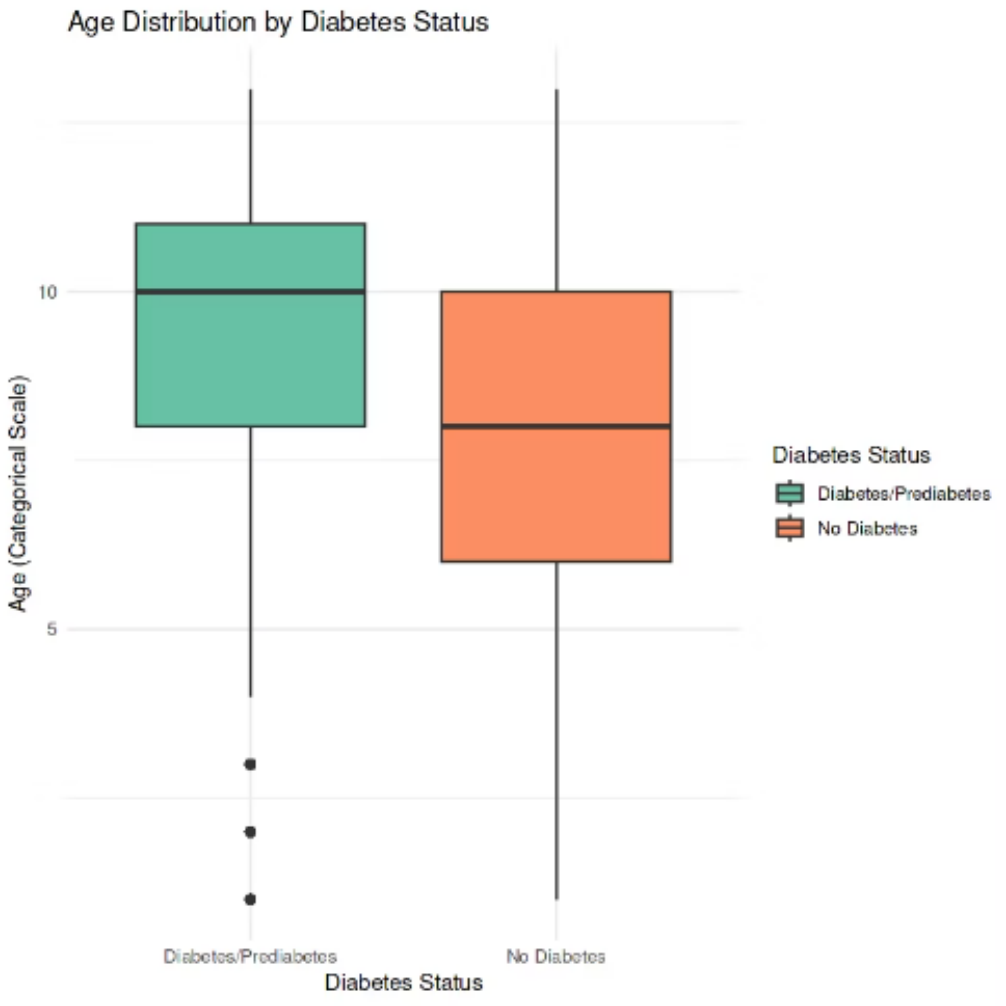
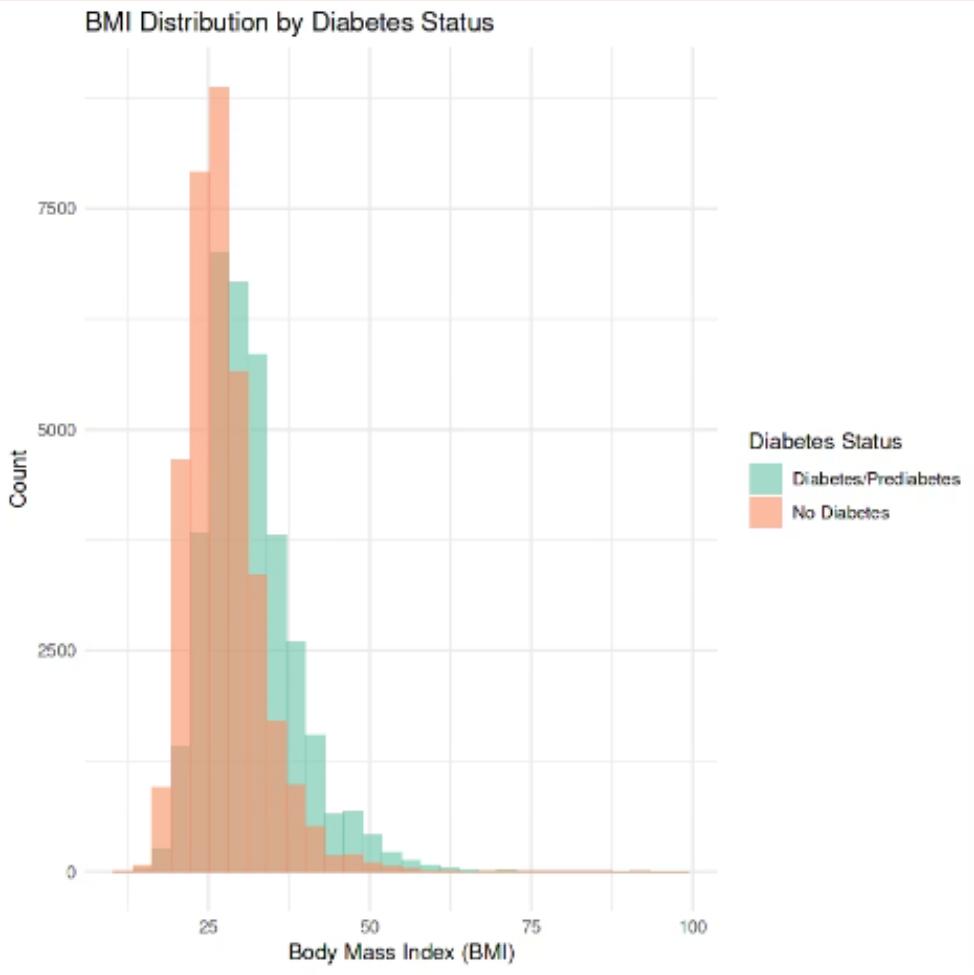
# Visualize Diabetes Status Distribution

A bar plot was generated to clearly illustrate the 50/50 split of the dataset based on diabetes status, confirming the deliberate balancing of the two groups for effective model training.



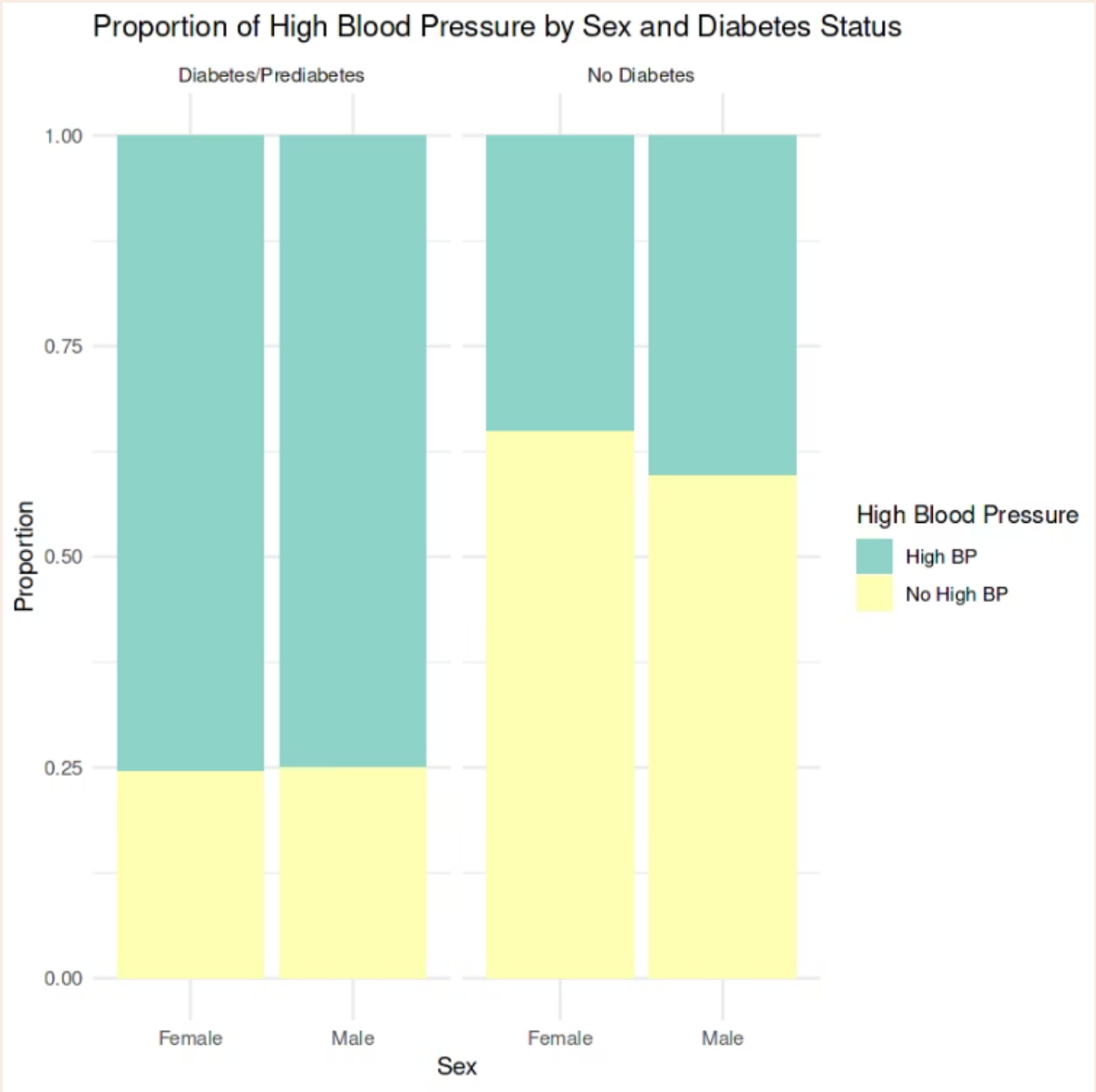
# Analyze BMI and Age Distributions

Histograms and comparative boxplots were used to visualize the distributions of Body Mass Index (BMI) and age. These revealed that individuals in the diabetes/prediabetes group generally exhibit higher BMI values and are older compared to the non-diabetes group.



# Examine HighBP and Sex Proportions

A stacked bar plot was created to vividly illustrate the prevalence of high blood pressure (HighBP) across different sexes and diabetes statuses, highlighting how this significant risk factor varies within demographic segments.



# Data Loading and Initial Exploration

This section details the foundational steps of our study, including the ingestion of the dataset and a preliminary examination of its structure and content.

## R Code for Data Loading and Summary

```
# Step 1: Load the dataset
data <- read_csv("/kaggle/input/diabetes-health-indicators-
dataset/diabetes_binary_5050split_health_indicators_BRFSS2015.csv",
show_col_types = FALSE)

# Step 2: Explore and preprocess the data
cat("\n=== Data Summary ===\n")

# Check for missing values
cat("Missing Values Check:\n")
print(colSums(is.na(data)))
# No missing values expected in this dataset

# Convert categorical variables to factors
data <- data %>%
  mutate(
    Diabetes_binary = as.factor(Diabetes_binary),
    Sex = as.factor(Sex),    Smoker = as.factor(Smoker),
    HighBP = as.factor(HighBP),    HighChol = as.factor(HighChol),
    PhysActivity = as.factor(PhysActivity)  )

# Summary statistics
cat("\nDescriptive Statistics of Variables:\n")
print(summary(data))
```

## Output: Data Summary and Statistics

```
=== Data Summary ===
Missing Values Check:
Diabetes_binary      HighBP      HighChol
      0              0              0
CholCheck            BMI      Smoker
      0              0              0
Stroke      HeartDiseaseorAttack      PhysActivity
      0              0              0
Fruits      Veggies      HvyAlcoholConsump
      0              0              0
AnyHealthcare      NoDocbcCost      GenHlth
      0              0              0
MentHlth      PhysHlth      DiffWalk
      0              0              0
Sex      Age      Education
0              0              0
Income
0

Descriptive Statistics of Variables:

Diabetes_binary HighBP      HighChol      CholCheck      BMI      Smoker
0:35346          0:30860    0:33529    Min.   :0.0000    Min.   :12.00    0:37094    1:35346
1:39832    1:37163    1st Qu.:1.0000    1st Qu.:25.00    1:33598
Median :1.0000    Median :29.00
:0.9753    Mean   :29.86
3rd Qu.:33.00      Max.   :1.0000    Max.   :98.00      ...
                                Mean
                                3rd Qu.:1.0000
```



# Logistic Regression Model

To answer our scientific question, we fitted a logistic regression model. This model predicts diabetes status based on age, sex, BMI, high blood pressure, high cholesterol, smoking, and physical activity.

```
# Step 3: Fit logistic regression model
cat("\n=== Logistic Regression Model ===\n")
cat("Model: Predicting Diabetes with Health Indicators\n")

# Predictors chosen based on health indicators relevant to diabetes
model <- glm( Diabetes_binary ~ Age + Sex + BMI + HighBP + HighChol + Smoker +PhysActivity, family = binomial(link = "logit"), data =
data)

# Step 4: Model summary
cat("\n=== Model Coefficients and Odds Ratios ===\n")
print(summary(model))
```

# Model Diagnostics and Performance

We performed several diagnostic checks on our model. Model Coefficients and Odds Ratios, Effect modification for interaction between Age and BMT and The model's predictive performance was evaluated using the Area Under the ROC Curve (AUC), which showed a moderate predictive ability.

## Model Coefficients and Odds Ratios

Calculating Coefficients and Odds ratios for all factors in our data

## Effect Modification

An interaction term between Age and BMI was tested, revealing a significant effect modification, suggesting their combined impact on diabetes risk is more than additive.

## ROC Curve and AUC

The ROC curve visually represents the model's performance, and the AUC value of approximately 0.79 indicates a moderate ability to distinguish between individuals with and without diabetes/prediabetes.

# Model Coefficients and Odds Ratios

## R Code

```
cat("\n=== Model Coefficients and Odds Ratios ===\n")
# Calculate odds ratios

odds_ratios <- exp(coef(model))

results <- cbind(odds_ratios)

print(results)
```

## Output

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-4.995821	0.063625	-78.520	< 2e-16	***
Age	0.170507	0.003528	48.327	< 2e-16	***
BMI	0.090052	0.001507	59.756	< 2e-16	***
Sex1	0.180890	0.017612	10.271	< 2e-16	***
HighBP1	0.959728	0.018615	51.556	< 2e-16	***
HighChol1	0.709013	0.017874	39.668	< 2e-16	***
Smoker1	0.135470	0.017635	7.682	1.57e-14	***
PhysActivity1	-0.344072	0.019408	-17.729	< 2e-16	***

=== Odds Ratios ===

	odds_ratios
(Intercept)	0.006766162
Age	1.185905990
BMI	1.094231269
Sex1	1.198282976
HighBP1	2.610986828
HighChol1	2.031985408
Smoker1	1.145074843
PhysActivity1	0.708877866

# Effect Modification

$$OR_{\text{BMI at Age } A} = \exp(\beta_{\text{BMI}} + \beta_{\text{Age:BMI}} \times A)$$

## R Code

```
# Fit logistic regression model with interaction Term

model_interaction <- glm(
  Diabetes_binary ~ Age * BMI + Sex + HighBP + HighChol +
  Smoker + PhysActivity, family = binomial(link = "logit"),
  data = data)

# Display summary to check interaction term

significancesummary(model_interaction)

# Calculate OR for BMI at different ages

beta_bmi <- log(1.094231269) # From main effect

beta_interaction <- 0.0045964 # Hypothetical interaction term

ages <- c(30, 50, 70)

or_bmi <- exp(beta_bmi + beta_interaction * ages)

names(or_bmi) <- paste("Age", ages)

print("Odds Ratios for BMI at different ages:")

print(or_bmi)
```

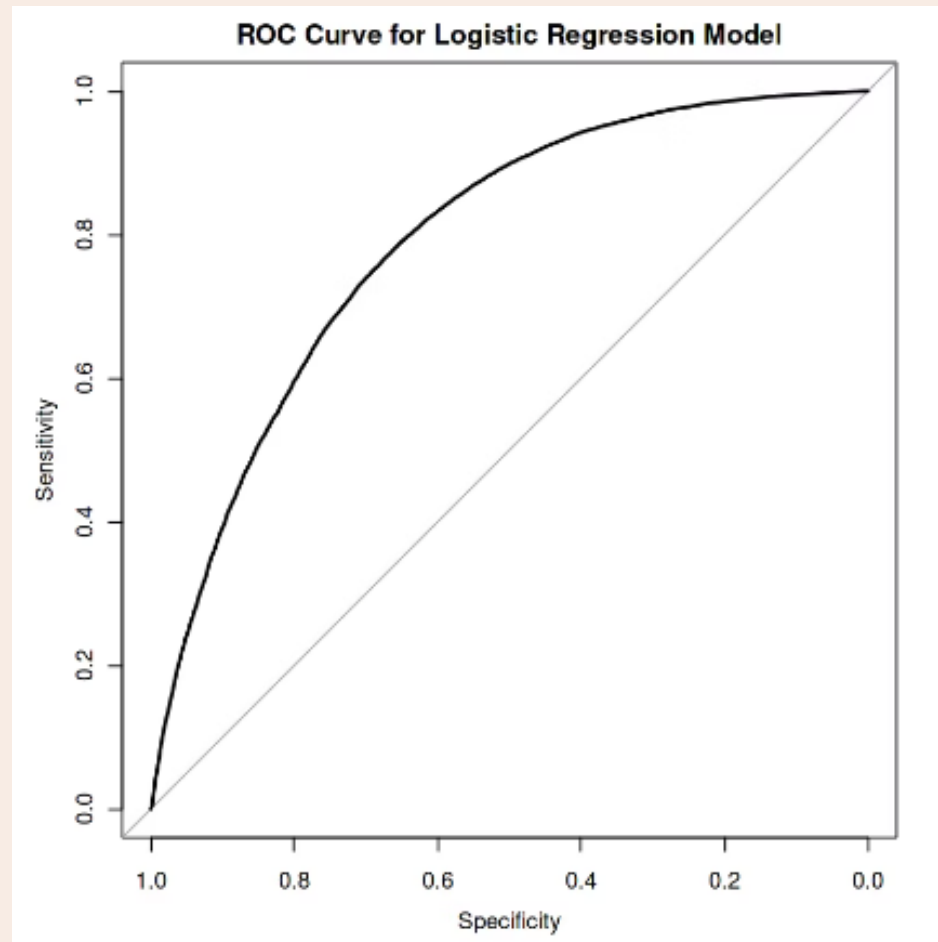
## Output

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.9057326	0.1351352	-28.902	< 2e-16	***
Age	0.0370070	0.0152472	2.427	0.0152	*
BMI	0.0529973	0.0043445	12.199	< 2e-16	***
Sex1	0.1749476	0.0176345	9.921	< 2e-16	***
HighBP1	0.9591559	0.0186338	51.474	< 2e-16	***
HighChol1	0.7084835	0.0178969	39.587	< 2e-16	***
Smoker1	0.1360606	0.0176518	7.708	1.28e-14	***
PhysActivity1	-0.3400024	0.0194415	-17.488	< 2e-16	***
Age:BMI	0.0045964	0.0005126	8.967	< 2e-16	***

```
[1] "Odds Ratios for BMI at different ages:"
   Age 30   Age 50   Age 70
1.256015 1.376952 1.509533
```

# ROC Curve and AUC



The ROC curve visually represents the model's performance, and the AUC value of approximately 0.79 indicates a moderate ability to distinguish between individuals with and without diabetes/prediabetes.

## R Code

```
# Model diagnosticscat  
  
("\n=== Model Performance: ROC and AUC ===\n")  
  
# Predict probabilities  
  
data$predicted_prob <- predict(model, type = "response")  
  
# ROC curve and Area Under the Curve (AUC)  
  
roc_obj <- roc(data$Diabetes_binary, data$predicted_prob)  
  
auc_value <- auc(roc_obj)  
  
cat("Area Under the ROC Curve (AUC):", auc_value, "\n")  
  
plot(roc_obj, main = "ROC Curve for Logistic Regression Model")
```

```
=== Model Performance: ROC and AUC ===  
Setting levels: control = Diabetes/Prediabetes, case = No Diabetes  
  
Setting direction: controls < cases  
  
Area Under the ROC Curve (AUC): 0.7902809
```

# Conclusion and Limitations

Our analysis revealed significant associations between high blood pressure, high cholesterol, and higher BMI with increased diabetes/prediabetes risk. Older age and male sex also correlated with higher prevalence. While the model shows moderate predictive ability, the 50/50 split biases prevalence estimates. Future longitudinal studies are needed to establish causality.

Key Findings	Limitations
Strong link between HighBP, HighChol, BMI and diabetes risk.	50/50 split biases prevalence estimates.
Older age and male sex correlate with higher prevalence.	Sampling weights were not included in the analysis.
Model AUC of ~0.79 indicates moderate predictive ability.	





# Thank You!

We appreciate your attention and engagement throughout this presentation. We hope our insights on diabetes risk factors have been valuable.

## Questions?

Please feel free to ask any questions you may have. We are happy to discuss our findings further.

## Contact Us

For more information or collaboration opportunities, please reach out:

- Email: [ahmedtarek2632@gmail.com](mailto:ahmedtarek2632@gmail.com)

