STATISTICAL ANALYSIS OF DIABETES IN PIMA INDIANS

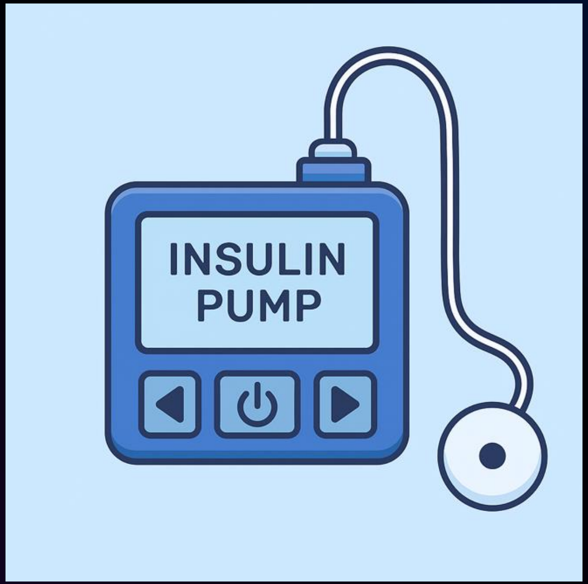
PRESENTATION

AGENDA

- Introduction
- Data exploration & descriptive statistics
- Inferential statistics
- Summary of analysis
- Recommendations

INTRODUCTION TO DIABETES

- Diabetes affects 590 million adults globally (International Diabetes Federation, 2025).
- By 2050, it is projected that 853 million adults will be diagnosed with the condition (International Diabetes Federation, 2025).
- Pima Indian population has one of the highest recorded rates of diabetes (Narayan *et al.*, 2021)



OpenAI (2025)

OVERVIEW OF DATASET

- Diabetes dataset analysed – descriptive and inferential statistics were generated
- All participants are women of Pima Indian heritage, older than 21 years
- R used to perform analysis – ease and simplicity of use

DATA EXPLORATION AND DESCRIPTIVE STATISTICS

DATA EXPLORATION

Total Sample	**768**
% of people diagnosed with diabetes	34.90
Age distribution	Mean age – 33.24; median age - 29
Pregnant women	657
Non-pregnant women	111

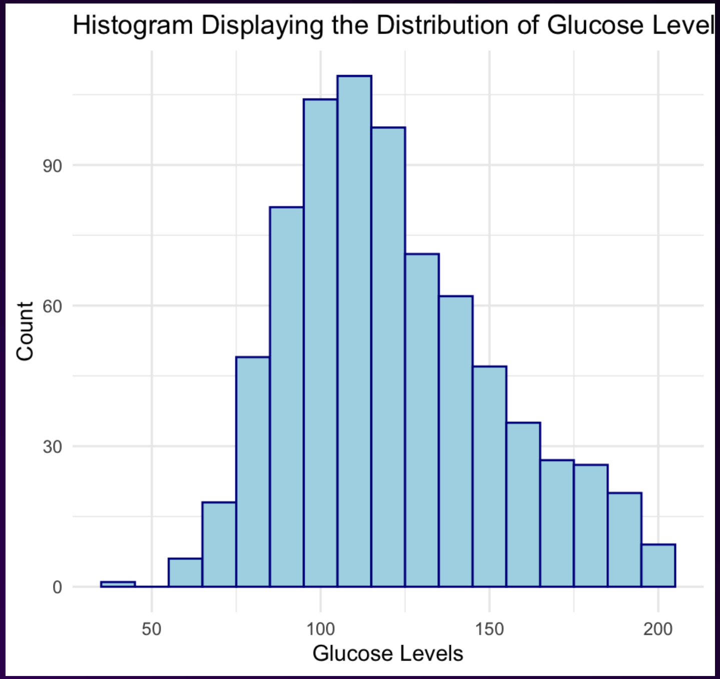
- Percentage of sample with diabetes higher than global population
- Younger group of women in sample
- Vast majority of women have been pregnant

DESCRIPTIVE STATISTICS

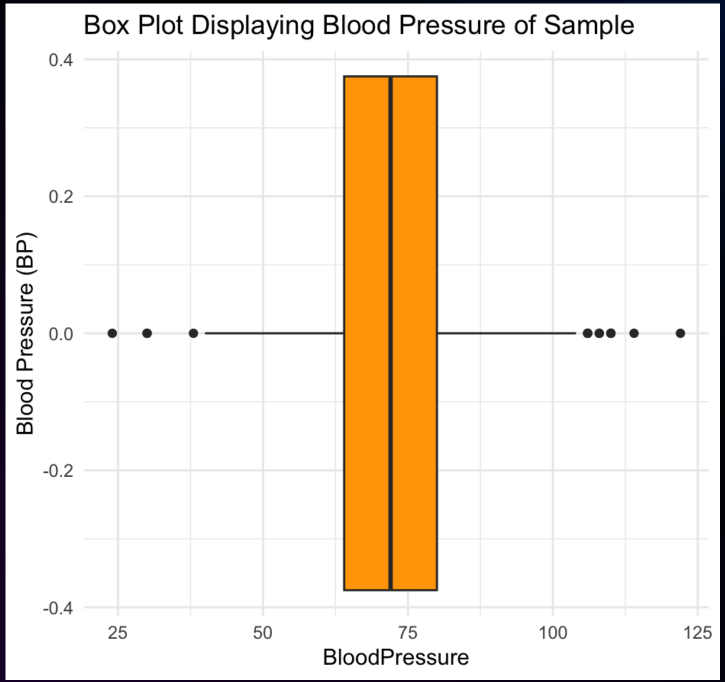
	Age	BMI	Glucose	Blood Pressure	Number of Pregnancies
Mean	33.24	32.46	121.69	72.41	3.58
Median	29.00	32.30	117.00	72.00	3.00
Mode	22	32	99	70	1
Minimum	21.00	18.20	44.00	24.00	0.00
Maximum	81.00	67.10	199.00	122.00	17.00
Range	60.00	48.90	155.00	98.00	17.00
Standard Deviation	11.67	6.92	30.54	12.38	3.70

- Range: indicates diversity across sample
- BMI: more individuals are obese
- Glucose: levels across the group are higher
- Blood pressure: average reading is within normal range
- No. pregnancies: most women have only experienced 1 pregnancy

PLOTS FOR GLUCOSE AND BLOOD PRESSURE SCORES

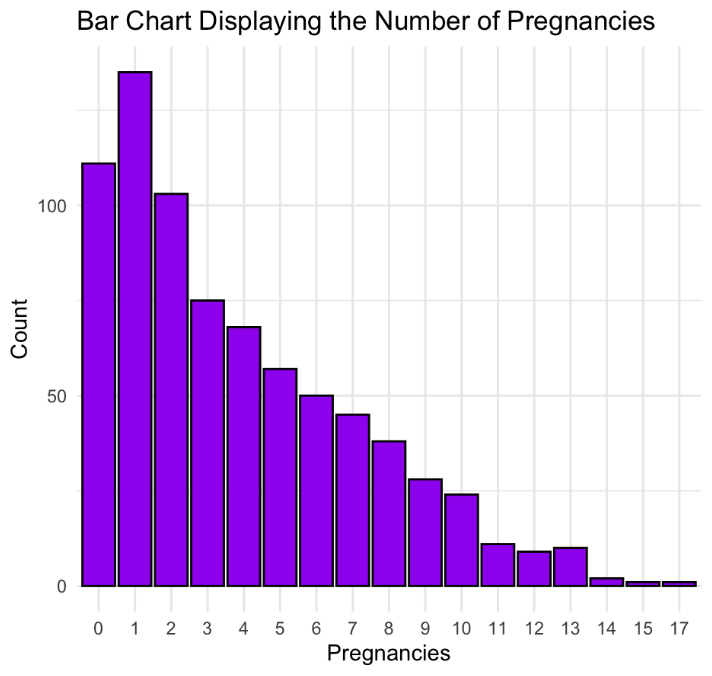


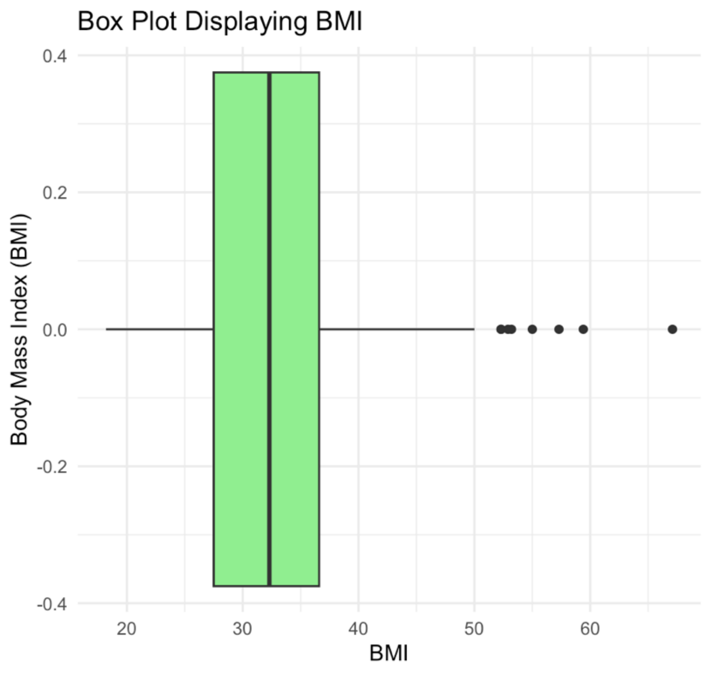
Distribution of glucose levels is quite normal

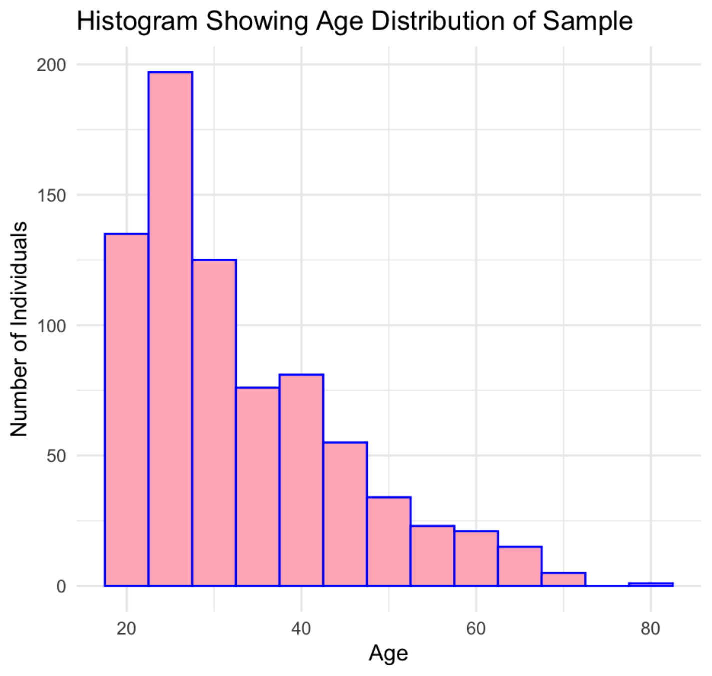


Distribution of blood pressure readings is quite normal

PLOTS FOR AGE, BMI AND PREGNANCIES







Number of pregnancies, BMI, and age all skewed right

OUTLIERS

- Interquartile range (IQR) method used to identify any outliers for the age, BMI, glucose levels, blood pressure, and no. pregnancies.
- Outliers identified for age, BMI, blood pressure, and pregnancies.
- BMI outlier scores an occurrence (Flegal, Kit and Graubard, 2014)
- Low scores for blood pressure unrealistically low (Kelly et al, 2022) – these values were excluded from dataset.
- Hugh blood pressure scores retained due to hypertension coexisting with diabetes (Przezak, Bielka and Pawlik, 2022)
- High pregnancy scores retained as it is a possibility (Tamir et al, 2025)

Age	BMI	Glucose Levels	Blood Pressure	Pregnancies
9	8	0	14	4

Variable	Outlier values
Age	67, 67, 67, 68, 69, 69, 70, 72, 81
BMI	52.3, 52.3, 52.9, 53.2, 55, 57.3, 59.4, 67.1
Blood Pressure	24, 30, 30, 38, 106, 106, 106, 108, 108, 110, 110, 110, 114, 122
Pregnancies	14, 14, 15, 17

RATE OF DIABETES ACROSS AGE GROUP, BMI CATEGORY, AND PREGNANCY GROUP

Age Group	Diabetes Rate (%)
<30	20.4
30-39	45.2
40-49	54.9
50+	48.9

Body Mass Index (BMI) Category	Diabetes Rate (%)
Underweight	0
Normal	7.29
Overweight	22.2
Obese	45.2
NA	25

Pregnancy Group	Diabetes Rate (%)
0	31.7
1-2	19.9
3-4	34.3
5+	47.9

- Age group: <30 group had lowest rate of diabetes
- BMI Category: diabetes rate increases as BMI increases
- Pregnancy group: diabetes rate highest in 5+ pregnancies group

INFERENTIAL STATISTICS

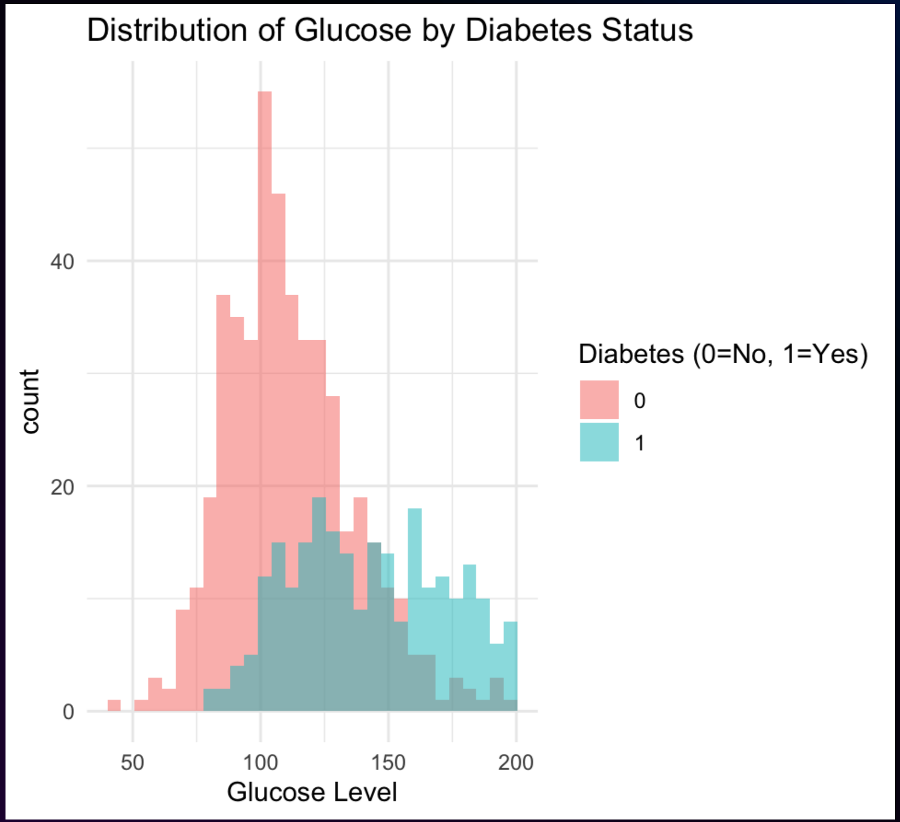
IS THERE A SIGNIFICANT DIFFERENCE IN GLUCOSE LEVELS BETWEEN THOSE WITH AND WITHOUT DIABETES?

- Glucose levels – a continuous variable
- Diabetes diagnosis – a categorical variable

Test of normality of glucose levels:

- Shapiro-Wilk:

Group	W	P-value
With Diabetes	0.97	p<0.001
Without diabetes	0.97	p<0.001



IS THERE A SIGNIFICANT DIFFERENCE IN GLUCOSE LEVELS BETWEEN THOSE WITH AND WITHOUT DIABETES?

Hypotheses:

- H_0 – There is no significant difference in glucose levels between individuals with and without diabetes.
- H_1 - There is a significant difference in glucose levels between individuals with and without diabetes.

Group	Mean Glucose	T(df)	95% CI	P-value
Non-diabetics (0)	111.13			
Diabetics (1)	142.76	t(433.19) = -14.28	(-35.98, -27.27)	p<0.001

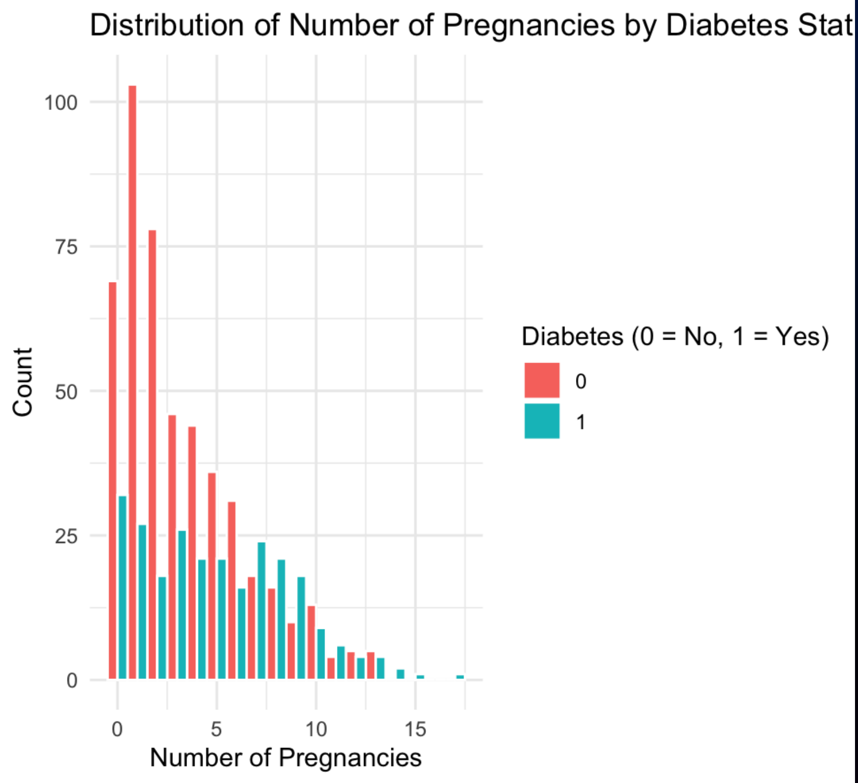
IS THERE A SIGNIFICANT DIFFERENCE IN THE NUMBER OF PREGNANCIES BETWEEN THOSE WITH AND WITHOUT DIABETES?

- Number of pregnancies– a continuous variable
- Diabetes diagnosis – a categorical variable

Test of normality of number of pregnancies:

- Shapiro-Wilk:

Group	W	P-value
With Diabetes	0.95	p<0.001
Without diabetes	0.88	p<0.001



IS THERE A SIGNIFICANT DIFFERENCE IN THE NUMBER OF PREGNANCIES BETWEEN THOSE WITH AND WITHOUT DIABETES?

Hypotheses:

- H_0 – There is no significant difference in the number of pregnancies between individuals with and without diabetes.
- H_1 - There is a significant difference in the number of pregnancies between individuals with and without diabetes.

Test statistic (W)	P-value
44550	$p<0.001$

CORRELATIONS BETWEEN ALL CONTINUOUS VARIABLES

- Correlation analysis conducted to assess relationship between continuous variables
- Coefficients calculated to determine strength & direction and significance of relationship

	No. Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes pedigree function	Age
No. Pregnancies		0.13 p<0.001	0.21 p<0.001	0.10 p=0.02	0.07 p=0.12	0.02 p=0.60	-0.03 p=0.40	0.56 p<0.001
Glucose	0.13 p<0.001		0.22 p<0.001	0.23 p<0.001	0.57 p<0.001	0.23 p<0.001	0.13 p<0.001	0.26 p<0.001
Blood pressure	0.21 p<0.001	0.22 p<0.001		0.24 p<0.001	0.08 p=0.11	0.32 p<0.001	-0.01 p=0.87	0.33 p<0.001
Skin Thickness	0.10 p=0.02	0.23 p<0.001	0.24 p<0.001		0.19 p<0.001	0.65 p<0.001	0.18 p=0.01	0.17 p<0.001
Insulin	0.08 p=0.12	0.58 p<0.001	0.08 p=0.11	0.19 p<0.001		0.24 p<0.001	0.13 p=0.01	0.22 p<0.001
BMI	0.02 p=0.60	0.23 p<0.001	0.32 p<0.001	0.65 p<0.001	0.24 p<0.001		0.16 p<0.001	0.03 p=0.49
Diabetes pedigree function	-0.03 p=0.40	0.13 p<0.001	-0.01 p=0.87	0.18 p=0.01	0.13 p=0.01	0.16 p<0.001		0.02 p=0.59
Age	0.56 p<0.001	0.26 p<0.001	0.33 p<0.001	0.17 p<0.001	0.22 p<0.001	0.03 p=0.49	0.02 p=0.59	

THE ASSOCIATION BETWEEN DIABETES AND BMI CATEGORIES AND AGE GROUPS

- BMI categories created in line with the World Health Organisation's (2025) categories:
  - Underweight - <18.5; Normal – 18.5– 24.9; Overweight – 25 – 29.9; Obese – 30+
- Age groups split by decade
  - <30; 30-39; 40-49; 50+
- Chi-squares test of association found a statistically significant relationship between diabetes diagnosis status and both BMI categories and age groups.

Group	x^2	df	P-value
BMI category	69.57	3	$p<0.001$
Age group	67.79	3	$p<0.001$

COMPARISON OF MEAN GLUCOSE LEVELS PER AGE GROUP

- Age groups split by decade
- Normality of glucose levels per age group tested using Shapiro-Wilk normality test
- Homogeneity of variances tested using Bartlett test
  - $K^2 = 7.15$, df = 3, p=0.07
- Hypotheses
  - H_0 – there is no significant difference in the distributions of glucose levels per age group
  - H_1 – there is a significant difference in the distributions of glucose levels per age group
- Kruskal-Wallis test showed a significant difference in distribution of glucose levels

Age Group	W	P-value
<30	0.95	P<0.001
30-39	0.97	p=0.003
40-49	0.98	p=0.20
50+	0.98	p=0.18

Shapiro-Wilk normality test

Age Group	Mean Glucose	Chi-squared	df	P-value
<30	115.06			
30-39	126.25			
40-49	125.16			
50+	139.78	54.74	3	p<0.001

Kruskal-Wallis test

WHICH VARIABLES PREDICT GLUCOSE LEVELS?

Aim: predict glucose levels based on Age, BMI, Pregnancies, Blood Pressure, Skin Thickness, Insulin, and Diabetes Pedigree Function.

- Outcome variable – glucose levels
- Predictor variables - Age, BMI, Pregnancies, Blood Pressure, Skin Thickness, Insulin, and Diabetes Pedigree Function

Hypotheses:

- H_0 – there is no significant relationship between predictor variables and glucose levels
- H_1 – there is a significant relationship between the predictor variables and glucose levels.

Model Summary

- Multiple linear regression model developed is statistically significant overall ($F(7, 380) = 35.84$, $p<0.001$)
- Most significant variables? Age ($\beta = 0.59$, $p<0.001$), Insulin ($\beta = 0.13$, $p<0.001$).
- 38.1% of the variance in glucose levels explained; 61.9% not explained

	F-statistic	df	R^2	P-value
Multiple Linear regression model	0.59	7, 380	0.39	p<0.001

Significance of multiple linear regression model

Predictor	Coefficient (β)	P-value
Age	0.59	p<0.001
BMI	0.18	0.47
Pregnancies	0.07	0.90
Blood pressure	0.19	0.11
Skin thickness	0.06	0.67
Insulin	0.13	p<0.001
Diabetes pedigree function	4.01	0.27

CAN AGE, BMI, AND GLUCOSE LEVELS PREDICT DIABETES?

- Aim: predict diabetes status based on age, BMI, and glucose levels
  - Outcome variable – diabetes outcome
  - Predictor variables - Age, BMI, glucose levels
- Hypotheses:
  - H_0 – there is no significant relationship between age, BMI, and glucose levels and diabetes outcome
  - H_1 – there is a significant relationship between age, BMI and glucose levels, and diabetes outcome
- Model Summary
  - Logistic regression model created
  - All three predictor variables significantly predict diabetes outcome: age (β = 0.03, $p<0.001$), BMI (β = 0.09, $p<0.001$), and glucose levels (β = 0.03, $p<0.001$)
  - 38.1% of the variance in glucose levels explained; 61.9% not explained

Predictor	Coefficient (β)	P-value
Age	0.03	$p<0.001$
BMI	0.09	$p<0.001$
Glucose levels	0.03	$p<0.001$

MODEL EVALUATION

- Regression model evaluated using the Hosmer-Lemeshow Goodness of Fit test and classification accuracy, sensitivity, and specificity.

	Chi-squared	df	P-value
Hosmer-Lemeshow test	8.73	8	0.37

Metric	Value
Accuracy	0.77
Sensitivity (recall)	0.57
Specificity	0.88
Precision (positive predictor value)	0.71

DOES AN AGE X BMI INTERACTION IMPROVE THE PREDICTION OF DIABETES RISK?

- Aim: determine whether age x BMI interaction improves prediction of diabetes risk beyond main effects of age, BMI, glucose levels.
- Two logistic regression models created:
  - Model 1: main effects (i.e., age, BMI, glucose levels, no. pregnancies)
  - Model 2: main effects + age x BMI interaction
- Both models compared using likelihood ratio test
- Analysis summary:
  - Model 2 with age x BMI interaction did not provide significantly better fit than model 1 without interaction
  - Age x BMI interaction was not statistically significant - effect of BMI on diabetes risk does not increase or decrease with age
  - Glucose levels and no. pregnancies are statistically significant predictors of diabetes risk

	Chi-squared	df	P-value
Analysis of deviance	0.75	1	0.39

Results on likelihood ratio test

	Odds ratio	95% CI (lower-upper)	p-value
Age	0.98	0.90-1.06	0.62
BMI	1.05	0.96-1.15	0.27
Glucose levels	1.04	1.03-1.04	p<0.001
No. pregnancies	1.11	1.04-1.20	0.001
Age x BMI	1.00	1.00-1.00	0.39

Odds ratios

SUMMARY AND RECOMMENDATIONS

SUMMARY OF FINDINGS

- A relationship exists between diabetes and BMI category, and diabetes and age group
- A relationship exists between glucose levels and age group
- Age and insulin are strong predictors of glucose levels
- Age, BMI, and glucose levels are strong predictors of diabetes status
- Risk of developing diabetes based on BMI doesn't increase or decrease with age
- Women with a higher number of pregnancies have a heightened risk of diabetes

RECOMMENDATIONS

- Future analysis could cover a similar range of variables for male Pima Indians

- Future research could also investigate the relationship between the variables in this analysis and the different types of diabetes (type 1, type 2, gestational)

Diabetes Dataset (no date). Available at: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset (Accessed: 20 October 2025).

Flegal, K.M., Kit, B.K. and Graubard, B.I. (2014) ‘Body Mass Index Categories in Observational Studies of Weight and Risk of Death’, *American Journal of Epidemiology*, 180(3), pp. 288–296. Available at: https://doi.org/10.1093/aje/kwu111.

International Diabetes Federation (2025) *Diabetes facts and figures*. Available at: https://idf.org/about-diabetes/diabetes-facts-figures/ (Accessed: 18 October 2025)

Kelly, Tanika.N. *et al.* (2022) *Insights From a Large-Scale Whole-Genome Sequencing Study of Systolic Blood Pressure, Diastolic Blood Pressure, and Hypertension | Hypertension*. Available at: https://www.ahajournals.org/doi/full/10.1161/HYPERTENSIONAHA.122.19324 (Accessed: 18 October 2025).

Li, Y. *et al.* (2020) ‘Maternal age and the risk of gestational diabetes mellitus: A systematic review and meta-analysis of over 120 million participants’, *Diabetes Research and Clinical Practice*, 162. Available at: https://doi.org/10.1016/j.diabres.2020.108044.

Narayan, K.M.V. *et al.* (2021) ‘Incidence of diabetes in South Asian young adults compared to Pima Indians’, *BMJ Open Diabetes Research & Care*, 9(1), p. e001988. Available at: https://doi.org/10.1136/bmjdrc-2020-001988.

Przezak, A., Bielka, W. and Pawlik, A. (2022) ‘Hypertension and Type 2 Diabetes—The Novel Treatment Possibilities’, *International Journal of Molecular Sciences*, 23(12), p. 6500. Available at: https://doi.org/10.3390/ijms23126500.

Tamir, T.T. *et al.* (2025) ‘Magnitude, distribution and determinants of non-utilization of antenatal care services among women in low- and middle-income countries: Insights for implementation of WHO recommendations’, *PLOS ONE*, 20(8), p. e0330596. Available at: https://doi.org/10.1371/journal.pone.0330596.

The Lancet Diabetes & Endocrinology (2025) ‘Diabetes and frailty in an ageing world’, *The Lancet Diabetes & Endocrinology*, 13(5), p. 355. Available at: https://doi.org/10.1016/S2213-8587(25)00094-4.

World Health Organization (2025) *Body mass index (BMI).* Available at: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index (Accessed 18 October 2025)

```
library(readr)
library(ggplot2)
library(dplyr)
library(tidyr)

diabetes<-read_csv("diabetes.csv")
nrow(diabetes)

# % of people with diabetes
prop.table(table(diabetes$Outcome)) * 100

# Age distribution of sample
ggplot(diabetes, aes(x=Age))+
  geom_histogram(binwidth=5,fill="lightpink",color="blue")+
  labs(title="Histogram Showing Age Distribution of Sample",
       x="Age",
       y="Number of Individuals")+
  theme_minimal()

# pregnant vs never pregnant
preg_status<-ifelse(diabetes$Pregnancies==0,"Never been pregnant","Have been pregnant")
table(preg_status)

# Descriptive statistics for Age, BMI, Glucose Levels, BP, No. Pregnancies
## Handle missing values for BMI, Glucose, BP
diabetes$BMI[diabetes$BMI==0]<-NA
diabetes$Glucose[diabetes$Glucose==0]<-NA
diabetes$BloodPressure[diabetes$BloodPressure==0]<-NA
### Calculate descriptive statistics
sapply(diabetes[c("Age","BMI","Glucose","BloodPressure","Pregnancies")],function(x)
  c(
    Mean=mean(x,na.rm=TRUE),
    Median=median(x,na.rm=TRUE),
    SD=sd(x,na.rm=TRUE),
    Min=min(x,na.rm=TRUE),
    Max=max(x,na.rm=TRUE),
    IQR=IQR(x,na.rm=TRUE))
  )
#### Mode of selected variables
mode_age<-as.numeric(names(sort(table(diabetes$Age),decreasing=TRUE)[1]))
show(mode_age)
mode_BMI<-as.numeric(names(sort(table(diabetes$BMI),decreasing=TRUE)[1]))
show(mode_BMI)
mode_glucose<-as.numeric(names(sort(table(diabetes$Glucose),decreasing=TRUE)[1]))
show(mode_glucose)
mode_bp<-as.numeric(names(sort(table(diabetes$BloodPressure),decreasing=TRUE)[1]))
show(mode_bp)
mode_preg<-as.numeric(names(sort(table(diabetes$Pregnancies),decreasing=TRUE)[1]))
show(mode_preg)
```

```
# Create plots for aforementioned variables
## BMI
ggplot(diabetes,aes(x=BMI))+
  geom_boxplot(fill="lightgreen")+
  labs(title="Box Plot Displaying BMI",y="Body Mass Index (BMI)")+
  theme_minimal()
## Glucose Levels
ggplot(diabetes,aes(x=Glucose))+
  geom_histogram(binwidth=10,fill="lightblue",color="darkblue")+
  labs(title="Histogram Displaying the Distribution of Glucose Levels",x="Glucose Levels",y="Count")+
  theme_minimal()
## Blood Pressure
ggplot(diabetes,aes(x=BloodPressure))+
  geom_boxplot(fill="orange")+
  labs(title="Box Plot Displaying Blood Pressure of Sample",y="Blood Pressure (BP)")+
  theme_minimal()
## Number of Pregnancies
ggplot(diabetes,aes(x=factor(Pregnancies)))+
  geom_bar(fill="purple",color="black")+
  labs(title="Bar Chart Displaying the Number of Pregnancies",x="Pregnancies",y="Count")+
  theme_minimal()

# Calculate outliers in variables
detect_outliers<-function(x){
  if(is.numeric(x)){
    Q1<-quantile(x,0.25,na.rm=TRUE)
    Q3<-quantile(x,0.75,na.rm=TRUE)
    IQR_value<-Q3-Q1
    lower<-Q1-1.5*IQR_value
    upper<-Q3+1.5*IQR_value
    sum(x<lower|x>upper,na.rm=TRUE)
  }else{
    NA
  }
}
diabetes_outliers<-sapply(diabetes,detect_outliers)
diabetes_outliers

show_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR_value
  upper <- Q3 + 1.5 * IQR_value
  outliers <- x[x < lower | x > upper]
  return(outliers)
}
outliers_age<-show_outliers(diabetes$Age)
outliers_bmi<-show_outliers(diabetes$BMI)
outliers_bp<-show_outliers(diabetes$BloodPressure)
outliers_glucose<-show_outliers(diabetes$Glucose)
```

```
outliers_preg<-show_outliers(diabetes$Pregnancies)

summary_outliers<-list(
  Age=outliers_age,
  BMI=outliers_bmi,
  Glucose=outliers_glucose,
  BP=outliers_bp,
  Pregnancies=outliers_preg
)%>%
  tibble::enframe(name="Variable",value="OutlierValues")
summary_outliers
summary_outliers_expand<-summary_outliers%>%
  tidyr::unnest(cols=c(OutlierValues))
print(summary_outliers_expand,n=86)

# Remove outlier values
diabetes_clean<-diabetes%>%
  filter(BloodPressure>=40)
summary(diabetes_clean)

# Create age group categories
diabetes_agegroups<-diabetes_clean%>%
  mutate(
    AgeGroup=case_when(
      Age<30~"<30",
      Age>=30&Age<40~"30-39",
      Age>=40&Age<50~"40-49",
      Age>=50~"50+"
    )
  )
# Calculate diabetes rate by age group
diabetes_rate_age <- diabetes_agegroups %>%
  group_by(AgeGroup) %>%
  summarise(
    Total = n(),
    Diabetic = sum(Outcome == 1, na.rm = TRUE),
    DiabetesRate = (Diabetic / Total) * 100
  ) %>%
  arrange(AgeGroup)
show(diabetes_rate_age)

# Create BMI categories
diabetes_bmi<-diabetes_clean%>%
  mutate(
    BMI_Category = case_when(
      BMI < 18.5 ~ "Underweight",
      BMI >= 18.5 & BMI < 25 ~ "Normal",
      BMI >= 25 & BMI < 30 ~ "Overweight",
      BMI >= 30 ~ "Obese"
    )
  )
# Calculate rate by BMI
diabetes_rate_bmi <- diabetes_bmi %>%
  group_by(BMI_Category) %>%
  summarise(
    Total = n(),
    Diabetic = sum(Outcome == 1, na.rm = TRUE),
    DiabetesRate = (Diabetic / Total) * 100
  ) %>%
  arrange(BMI_Category)
show(diabetes_rate_bmi)

# Create categories for no. pregnancies
diabetes_preg <- diabetes_clean %>%
  mutate(
    PregGroup = case_when(
      Pregnancies == 0 ~ "0",
      Pregnancies >= 1 & Pregnancies <= 2 ~ "1-2",
      Pregnancies >= 3 & Pregnancies <= 4 ~ "3-4",
      Pregnancies >= 5 ~ "5+"
    )
  )
# Calculate rate by no. pregnancies
diabetes_rate_preg <- diabetes_preg %>%
  group_by(PregGroup) %>%
  summarise(
    Total = n(),
    Diabetic = sum(Outcome == 1, na.rm = TRUE),
    DiabetesRate = (Diabetic / Total) * 100
  ) %>%
  arrange(PregGroup)
show(diabetes_rate_preg)

# Determine whether significant difference exists betw. glucose levels & diabetes diagnosus
## Determine normality for continuous variable (glucose)
glucose_no_diabetes<-diabetes_clean$Glucose[diabetes_clean$Outcome==0]
glucose_diabetes<-diabetes_clean$Glucose[diabetes_clean$Outcome==1]
shapiro.test(glucose_no_diabetes)
shapiro.test(glucose_diabetes)

# Histogram to show distribution
ggplot(diabetes_clean, aes(x = Glucose, fill = as.factor(Outcome))) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 30) +
  labs(title = "Distribution of Glucose by Diabetes Status",
       x = "Glucose Level", fill = "Diabetes (0=No, 1=Yes)") +
  theme_minimal()

# Conduct t-test
t_test_glucose<-t.test(Glucose~Outcome,data=diabetes_clean)
show(t_test_glucose)

# Determine whether significant difference exists in no. pregnancies betw. diabetics & non-diabetics
## Determine normality for continuous variable (no. pregnancies)
```

```r
preg_no_diabetes<-diabetes_clean$Pregnancies[diabetes_clean$Outcome==0]
preg_diabetes<-diabetes_clean$Pregnancies[diabetes_clean$Outcome==1]
shapiro.test(preg_no_diabetes)
shapiro.test(preg_diabetes)

# Histogram to show distribution
ggplot(diabetes_clean, aes(x = Pregnancies, fill = factor(Outcome))) +
  geom_histogram(binwidth = 1, position = "dodge", color = "white") +
  labs(
    title = "Distribution of Number of Pregnancies by Diabetes Status",
    x = "Number of Pregnancies",
    y = "Count",
    fill = "Diabetes (0 = No, 1 = Yes)"
  ) +
  theme_minimal()

# Conduct Mann-Whitney U test
wilcox.test(Pregnancies~Outcome,data=diabetes_clean)

# Determine correlations between continuous variables
## Replace zeros with NA for SkinThickness, Insulin variables
diabetes_clean<-diabetes_clean %>%
  mutate(
    SkinThickness=na_if(SkinThickness,0),
    Insulin=na_if(Insulin, 0),
  )

continuous_vars <- diabetes_clean[, c("Pregnancies", "Glucose", "BloodPressure",
                                   "SkinThickness", "Insulin", "BMI",
                                   "DiabetesPedigreeFunction", "Age")]
r_matrix <- cor(continuous_vars, use = "pairwise.complete.obs", method = "pearson")
n_matrix <- outer(
  colnames(continuous_vars),
  colnames(continuous_vars),
  Vectorize(function(x, y) sum(complete.cases(continuous_vars[, c(x, y)])))
)
dimnames(n_matrix) <- list(colnames(continuous_vars), colnames(continuous_vars))
t_matrix <- r_matrix * sqrt((n_matrix - 2) / (1 - r_matrix^2))
p_matrix <- 2 * pt(-abs(t_matrix), df = n_matrix - 2)
p_matrix <- round(p_matrix, 4)

r_p_table <- matrix(
  paste0(round(r_matrix, 2), " (p=", p_matrix, ")"),
  nrow = nrow(r_matrix)
)
rownames(r_p_table) <- rownames(r_matrix)
colnames(r_p_table) <- colnames(r_matrix)

r_p_table

# Test association between diabetes & Age Groups and BMI categories
```

```r
diabetes_clean$AgeGroup<-cut(
  diabetes_clean$Age,
  breaks = c(-Inf, 29, 39, 49, Inf),
  labels = c("<30","30-39","40-49","50+")
)
diabetes_clean$BMICategory <- cut(
  diabetes_clean$BMI,
  breaks = c(-Inf, 18.5, 24.9, 29.9, Inf),
  labels = c("Underweight", "Normal", "Overweight", "Obese")
)
table_age <- table(diabetes_clean$Outcome, diabetes_clean$AgeGroup)
show(table_age)
table_bmi <- table(diabetes_clean$Outcome, diabetes_clean$BMICategory)
show(table_bmi)

chisq_age <- chisq.test(table_age)
show(chisq_age)
chisq_bmi <- chisq.test(table_bmi)
show(chisq_bmi)

# Comparison of mean glucose scores across groups
## View mean glucose levels per group
aggregate(Glucose ~ AgeGroup, data = diabetes_clean, mean, na.rm = TRUE)
## Test normality of glucose levels
by(diabetes_clean$Glucose, diabetes_clean$AgeGroup, shapiro.test)
## Test homogeneity of variance
bartlett.test(Glucose ~ AgeGroup, data = diabetes_clean)
## Run Kruskal-Willis test
kruskal.test(Glucose ~ AgeGroup, data = diabetes_clean)

# Create multiple linear regression of glucose & predictor variables
model_glucose <- lm(Glucose ~ Age + BMI + Pregnancies + BloodPressure +
                      SkinThickness + Insulin + DiabetesPedigreeFunction,
                    data = diabetes_clean)
summary(model_glucose)

# Create logistic regression of diabetes & predictor variables
## Create a clean dataset with only complete cases
model_diabetes_clean <- na.omit(diabetes_clean[, c("Outcome", "BMI", "Age", "Glucose")])
model_log<- glm(Outcome~BMI+Age+Glucose,
                data=model_diabetes_clean,
                family=binomial)
summary(model_log)

library(ResourceSelection)
hoslem.test(model_diabetes_clean$Outcome,fitted(model_log),g=10)

## Classification Performance
model_diabetes_clean$pred_prob <- predict(model_log, type = "response")
model_diabetes_clean$pred_class <- ifelse(model_diabetes_clean$pred_prob >= 0.5, 1, 0)

install.packages("caret")
library(caret)
```

```r
conf_matrix <- confusionMatrix(
  factor(model_diabetes_clean$pred_class),
  factor(model_diabetes_clean$Outcome),
  positive = "1"
)

conf_matrix

# Investigate if there are any significant interactions betw. BMI
& age when predicting dibaetes risk
## Log regression model w/o interaction - main effects only
model_maineff <- glm(Outcome ~ BMI + Age + Glucose +
Pregnancies,
                 data = diabetes_clean,
                 family = binomial)
model_interaction <- glm(Outcome ~ BMI * Age + Glucose +
Pregnancies,
                     data = diabetes_clean,
                     family = binomial)
## Likelihood ratio test & odds ratio
anova(model_maineff, model_interaction, test = "LRT")
summary(model_interaction)
exp(cbind(OR = coef(model_interaction),
confint(model_interaction)))
```