# Pima Indian Women: Diabetes Risk Analysis

Siddharth Das, Isabelle Berkowitz, Martin Topacio, Carly Schwartberg

STA 135 Multivariate Data Analysis, June 6, 2024

## Contents

# 1 INTRODUCTION

In the 1960s the long-term commitment to study type 2 diabetes and obesity in Pima Indians began to indulge. With their willingness to participate in research exploration, statisticians identified significant findings regarding "the epidemiology, physiology, clinical assessment, and genetics" [1] heavily related to these major health issues. Throughout history, Pima Indians were forced to adapt to new territory due to white settlers, causing them to reset irrigation systems for agriculture to support their food growth. As a result of this displacement and farming issues, the amount of food, physical activity, and the height of their economy was very poor. As a result diabetes and obesity became more prevalent among the Pima Indians. The research collected showed that type 2 diabetes increased during this time for Pima women, likely as a result of increased body mass index. The goal of our project is to visually demonstrate the correlation between our dependent variable against various factors and to perform a more in-depth analysis of this data by applying multivariate techniques taught in the course and understanding why certain methods are or are not significant to our data.

# 2 THE DATA

In this project, we dive into the Pima Indians Diabetes Database to see what factors have higher significance in contributing to the prevalence of type 2 diabetes in female Pima Indians. The data has been collected from the U.S. National Institute of Diabetes and Digestive Kidney Diseases on women 21 $\leq$. We extracted this data in R-Studio using the library package **(mlbench)** containing 768 observations onto 9 variables. [2] Computing the total counts in R, we found that **500** women do not have diabetes, while **268** do.

## 2.1 Description of Data Set

- Pregnancies: Number of times a woman has been pregnant
- Glucose: Plasma Glucose Concentration
- Pressure: Diastolic Blood Pressure (mm Hg)
- Triceps: Triceps Skin Fold Thickness (mm)
- Insulin: 2 Hour Serum Insulin (mu U/ml)
- Mass: Body mass index $\frac{weight(kg)}{height(m)^2}$
- Pedigree: Diabetes Pedigree Function
- Age: The age of the woman
- Diabetes: Represented as a factor level for the test results (2 = positive, 1 = negative)

# 3 Initial Data Visualizations

The following visualizations represent the beginning understanding of the relationships between diabetes prevalence among other factors.

## 3.1 Correlation Matrix Analysis

The first graphs demonstrate the correlation between factor variables and the distributions between the categorical variables, on whether a woman has diabetes or not. We see there is a high correlation between **age** and **pregnancy**, likely because as we get older, pregnancy becomes more common. There is some moderate correlation between **insulin** and **triceps**, implying that women with high insulin levels may have thicker triceps folds (and vice versa).

## 3.2 Box Plot Analysis

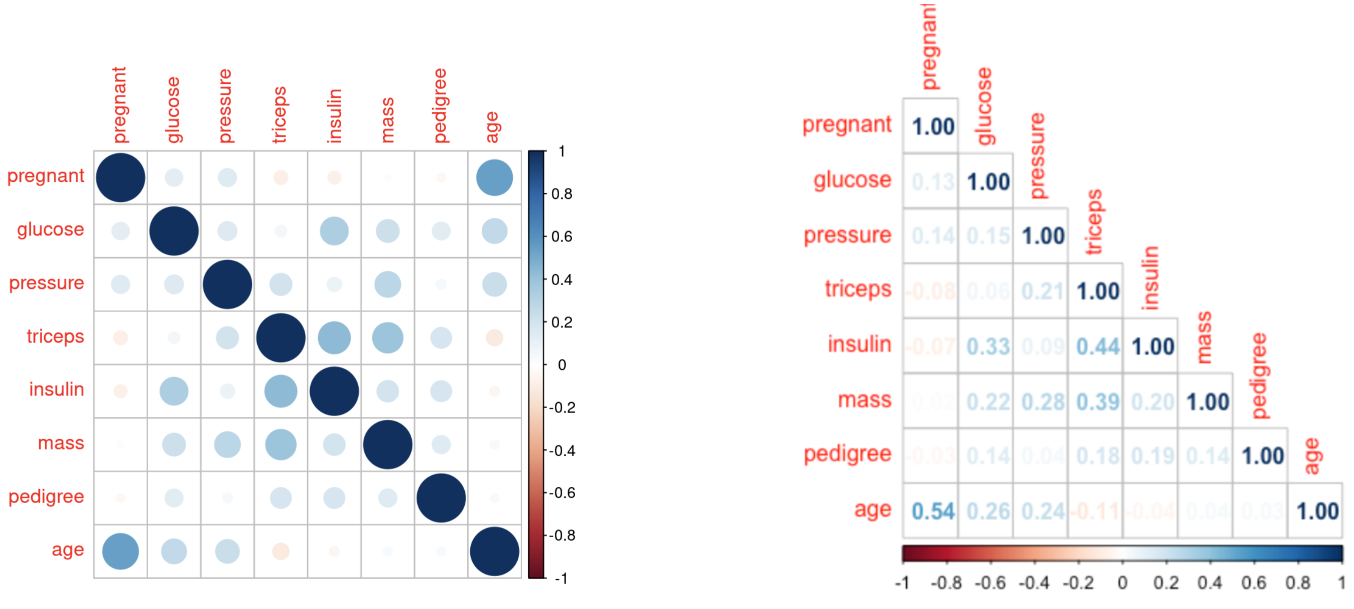
The second set of graphs represents the distribution of diabetes (pos/neg) of individuals among others. A significant finding is that young adults **(ages 20-30)** have a higher prevalence of diabetes-positive cases compared to aged 30 $\leq$. **Higher glucose, BMI, and pedigree function values** associate with a **higher** incidence of diabetes. However, **blood pressure** and **triceps thickness** do **not strongly correlate**.

## 3.3 Box Plot Analysis

The box plots represent the distribution of values for each factor on whether a person has diabetes(pos/neg), by identifying outliers. Showing **age** and **high glucose levels** are the most influential factors in diabetes prevalence over time. **Pregnancy** and **BMI** suggest significant differences among others, implying **lifestyle** heavily contributes to diabetes risk. In conclusion, these graphs confirm the well-known risk factors for diabetes outlining a clear comparison of these metrics between those with and without diabetes.

[1] Dataset source: High-Risk Populations: The Pimas of Arizona and Mexico

[2] Dataset source: Pima Indians Diabetes Data



(a) Correlation Matrix Plots

(b) Numeric Correlation Matrix

Figure 1: Correlation Matrix Plots to Understand Relationships between factors

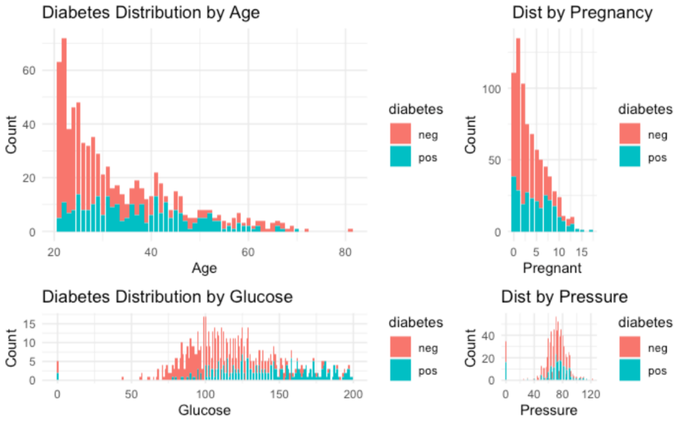


Figure 2: Distribution of first 4 Categorical variables

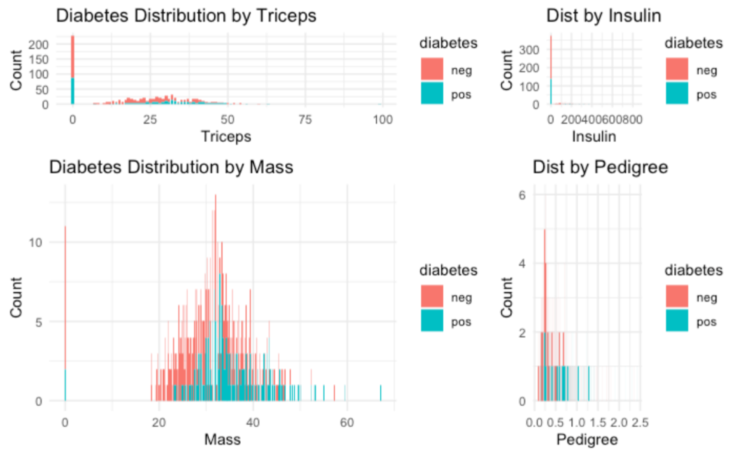
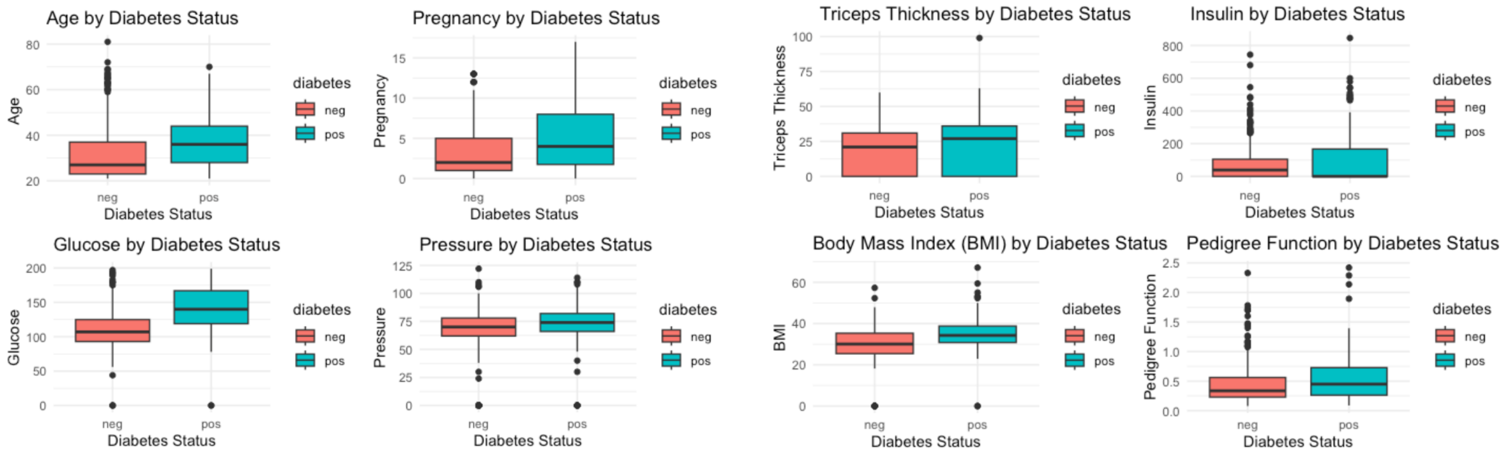


Figure 3: Distribution of last 4 Categorical variables



(a) Caption for Figure 1

(b) Caption for Figure 2

Figure 4: Overall Caption

# 4 Methodology

**1. Box's M Test:**
For the transitions of univariate to multivariate we conduct this test to check for the variance between the covariance matrices. This test can help statisticians determine if LDA or QDA applies better to specific data through hypothesis testing. When the **p-value** is <0.05 we **reject the null hypothesis** stating that variance-covariance is the same. When the p-value is close to zero you will want to continue with QDA. This test assumes $x_1$,$x_2$,..., $x_{n1}$˜ $N_p(\mu_x$,$\Sigma_1)$ and $y_1$,$y_2$,..., $y_{n2}$˜ $N_p(\mu_y$,$\Sigma_2)$. These assumptions guide us into the computation of M:

$$M = \frac{|S_1|^{\frac{v_1}{2}} + |S_2|^{\frac{v_2}{2}}}{|S_{pl}|^{\frac{v_1}{2}+\frac{v_2}{2}}}$$

**2. Linear Discriminant Analysis:**
A is a form of statistical classification that identifies patterns and looks for 'linear combinations that separate different factors between classes and data types.'[3] Our model used what is called **fault diagnosis** as we attempted to see which variables were **"good"** = high association and **"bad"** = low effect in diabetes prevalence. Depending on the outcome of the p-value in Box's M test we determine if LDA or QDA is more reasonable. To check assumptions for this model, you can implement Box's M test along with the linear discriminant function:

$$s_k(x) = x^T\Sigma^{-1}\mu_k - 0.5\mu_k^T\Sigma^{-1}\mu_k + log\pi_k$$

**3. Quadratic Discriminant Analysis**
On the other hand, QDA allows us to understand how our specified factors vary when the population covariance matrices are different. In the form of classification, this procedure leads to conclusions as to how variables discriminate between groups of dependent and controlled data. These results tend to be more accurate and credible which is significant in looking at risk factors such as diabetes prevalence among risk and demographic factors. The quadratic discriminant function follows as:

$$S_k(x) = \frac{-1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + log\pi_k$$

**4. Principal Component Analysis**
PCA is a common statistical procedure requiring us to "choose a subspace to maximize the projected variance, or minimize the reconstruction error."[4] Once we remove constants to avoid misleading results, the test can be performed. We are then left with standard deviations on the principal components, variance, and scaled proportions.[5] PCA applies the **Courant-Fischer Theorem** where if **A** is a symmetric matrix, we denote the eigenvalues as $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ where:

$$\lambda_k \min_{\alpha\in_k^T} \cdot \max_{x\in S_k^\alpha} \cdot k(zerovector)\frac{X^TAx}{||x||_2^2}$$

# 5 Results and Discussion

## 5.1 Box's M-test for Homogeneity of Covariance Matrices

We use this to determine whether we should implement QDA or LDA. When we conducted this test in R we got a p-value of < **2.2e-16**, which is so close to zero that we rejected the null hypothesis and continued with performing QDA on the dataset.

## 5.2 Quadratic Discriminant Analysis

By implementing QDA we split the data into training and test sets. This allowed us to perform QDA with k-fold cross-validation, and again without splitting data. We plan to evaluate the accuracy of each QDA process and see which is the best for fitting the data, new and old. From the confusion matrix, we see that it yields **122** true negatives, **28** false negatives, **41** true positives, and **39** false positives. False positives are indicative of Type I error, while false negatives are indicative of Type II error. From our confusion matrix, we see **39** healthy women were diagnosed with onset diabetes, and **28** women with onset diabetes were incorrectly diagnosed as healthy.

[3] Dataset source: How to Use LDA and QDA with Multiple Predictors

[4] Dataset source: Week 8-1_annotated

[5] Dataset source: What is Principal Component Analysis

**a) Confusion Matrix:**
From the confusion matrix, we see that it yielded **122** true negatives and **28** false negatives, **41** true positives, and **39** false positives. False positives are indicative of Type I error, and false negatives are indicative of Type II error. This means from our confusion matrix, **39** healthy women were diagnosed with onset diabetes, and 28 women with onset diabetes were not.

**b) QDA (split) Performance Metrics:**
In this table (outputted in markdown) we identify several key findings. First, we obtain an accuracy rate of **70.87%**, which is the %age of correctly classified instances. We get a precision rate of **75.78%**, which is the %age of identifications that prove to be correct. The recall %age is **81.33%**, the %age of actual positives that are correctly identified. The F1 score, the mean of precision and recall, is 78.46%. Finally, we obtained a misclassification error rate of **29.13%**. For precision, recall, and F1 score, we ideally would desire rates at 80 % or above. Our precision and F1 rates are somewhat close to 80 %, while our recall %age is over 80 %, which is generally good. We want high accuracy, ideally above **90%**, but our accuracy is at **70.87%**, which is not good enough. Our misclassification error rate is mediocre at **29.13%**, with a rate at **40%** or above being considered poor.

**c) QDA with cross-validation - Confusion Matrix**
Examining the confusion matrix from the QDA model with cross-validation, we have **432** true negatives **68** false negatives, **155** true positives, and **113** false positives. This indicates that **113** healthy women were incorrectly diagnosed with diabetes (Type I error), and **68** women with diabetes were incorrectly diagnosed as healthy (Type II error).

**d) QDA with cross-validation - Performance Metrics**
When running a QDA model with cross-validation, we obtain several findings. First, our accuracy is **76.43%**, showing improvement over the prior model **WITH** training and test sets. There are improved precision and recall rates at **79.27%** and **86.4%**, respectively, an improved F1 score of **82.68%**, and a lower misclassification error rate, **23.57%**. Our accuracy rate is about **6%** closer to a "good" rate of 80 %, but not at our desired "great" rate of 90% or higher. Our precision rate is almost at 80 %, considered generally good, and our recall rate is sitting at a good % age. Our F1 score is also now above 80 %, which is good. The misclassification error rate did not drop as much as we hoped, with it still sitting at a somewhat mediocre %age, but it is still beneficial that it lowered at all.

**e) QDA without splitting - Performance Metrics**
Since both the QDA cross-validation model and the QDA model without splitting result in the same accuracy, precision, recall, F1 score, and misclassification error rate, the cross-validation model is better because it has less risk of over fitting, so it will make more accurate predictions on new data compared to the model without splitting data.

## 5.3 Principal Component Analysis:

We implemented PCA in our project which allowed us to explore relationships in a specific subspace to optimize projection variance. Our goal of PCA was to minimize the reconstruction error.

**a) PCA Plots (PC1 vs. PC2):** Plot 1 shows the correlation between PC1 and PC2. From this(in markdown), we know PC1 is heavily driven by insulin levels and PC2 is driven by glucose. The cluster at the origin suggests, that most people have average levels of both. Those straying right on the x-axis indicate higher insulin and those straying up on the y-axis indicate higher glucose levels. From this plot, we conclude that individuals who are at risk for both factors are the most likely to have the risk of diabetes.

**b) Contribution Ratio:** The contribution ratio plot shows us how many principal components are needed to get a majority of the variance. We can see a fast incline from the first 2 points and relative steadiness among other components, not too much fluctuation. Showing us that PC1 and PC2 are the main factors for the variance. Overall from this plot, we can see that after the first 6 components, the plot flattens out.

**c) Overall Threshold:** For the overall threshold, we interpret the eigenvalues and see there is a large drop from PC1 to PC2. The line represents the mean and those above it have significance to the variance. The graph shows only PC1 is above the line, and in far second place PC2 is closest to the line out of the remaining components.

**d)Scree Plot:** The scree plot shows us a very similar thing to the overall threshold plot. Instead of looking at the big drop in eigenvalues to interpret the number of components needed, the Scree plot gives us a bigger overview of the eigenvalues for our dataset. This shows us the rate of variation as you go from one component to the next. As we can see here the output is generally the same showing the eigenvalue for PC1 to be by far the highest.

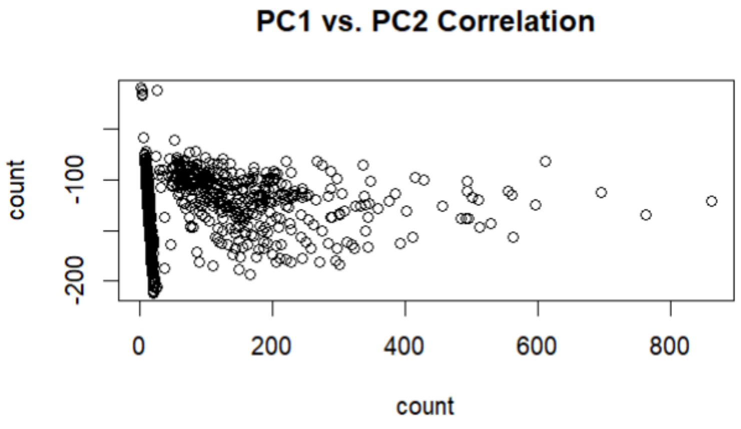


Figure 5: a) PC1, PC2 Correlation

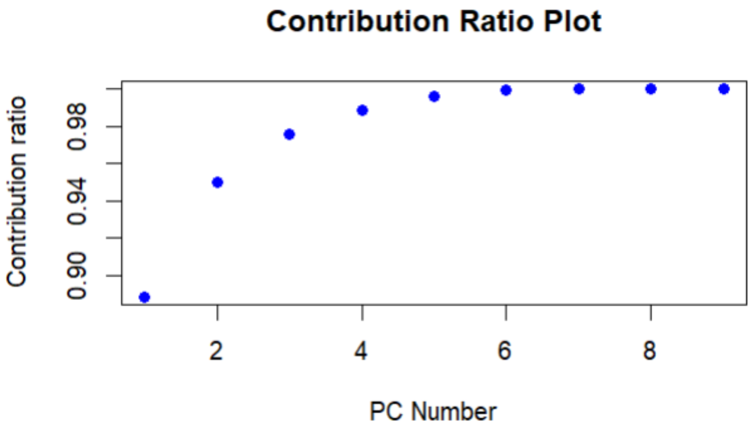


Figure 6: b) PC Relation to Variance

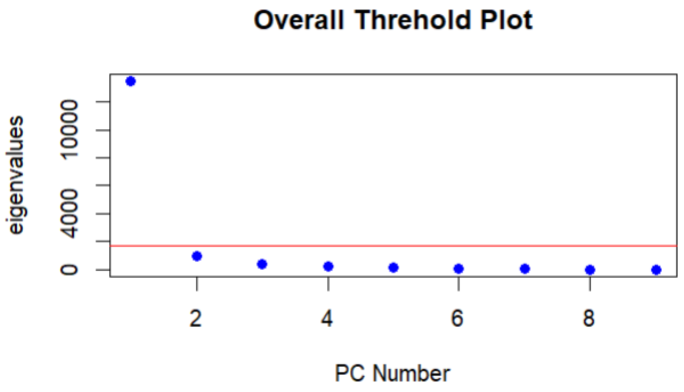


Figure 7: c) Comparison of Trends in Eigenvalues

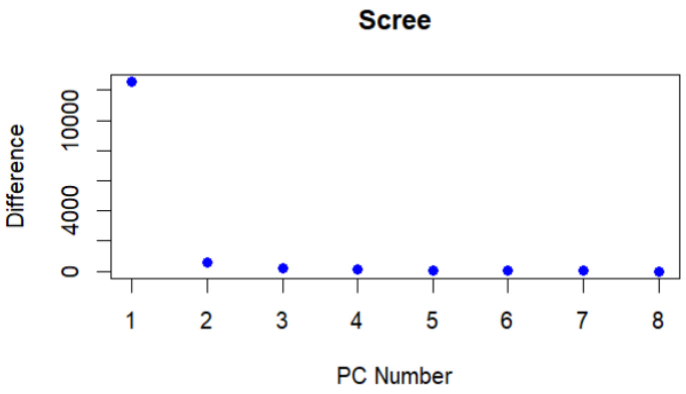


Figure 8: d) Overview of the Eigenvalues

# 6 Conclusion

In the 1960s the long term-commitment to study type 2 diabetes and obesity in women Pima Indians began to indulge. With their homes taken over by American settlers throughout history, their access to food, levels of physical activity, and economy was poor. Our project visually demonstrated the correlation between our dependent variable (diabetes) against our various factors and showed a more in-depth analysis on this data by applying multivariate techniques learned in the course and understanding why certain methods are or are not significant to our data. We accessed our data using the library package (mlbench) which containing data on 768 women with 9 observations. Our initial exploratory data analysis showed us that correlations exist between age and number of pregnancies, and insulin levels and triceps fold thickness. Furthermore, we noticed an increased occurrence of diabetes in young adults between 20 and 30 years old. In terms of multivariate methodology, we utilized the Box's M Test to learn whether the covariance matrices are the same. Since the covariance matrices were different, we developed three QDA models and assessed their performance metrics. Finally, we implemented PCA and found that insulin and glucose, or the first 2 principal components, explained a majority of the variance in the data. One key finding from our QDA models was that regardless of k-fold cross validation, our predictions were almost identical in terms of accuracy. Despite that, we prefer the k-fold cross validation model because it generally performs better while accounting for overfitting and evaluation bias. All in all, this study is beneficial for understanding the risk factors for type 2 diabetes among Pima Indian Women. This knowledge can benefit society by guiding health policies and developing more effective diabetes-prevention strategies. In the future, we intend to further research how health policies can be improved to be more inclusive of minority groups.

**Note:** We utilized Professor Ding's Lecture Notes to help with the methodology, as well as Chat GPT to help with our testing and visualizations. Finally, we used outside resources to provide a more thorough explanation on our testing procedures.

## 7 Refrences

- ChatGPT (For assistance in modeling graphs/errors in LaTeX/some Wording, June 4, 2024).

- Ding Xiuxai (2024) Week 7-1 [HTML Lecture Slides] University of California, Davis. https://canvas.ucdavis.edu/courses/877214/files/folder/Lecture%20notes?preview=24340030. Accessed 6 June 2024.

- Ding Xiuxai (2024) Week 8-1 [HTML Lecture Slides] University of California, Davis. https://canvas.ucdavis.edu/courses/877214/files/folder/Lecture%20notes?preview=24492190. Accessed 6 June 2024.

- “Pima Indians Diabetes Data.” R, search.r-project.org/CRAN/refmans/pdp/html/pima.html. Accessed 11 June 2024.

- Renzo, Charles Di. “How to Use LDA and QDA with Multiple Predictors.” LinkedIn, 29 Feb. 2024, www.linkedin.com/pulse/how-use-lda-qda-multiple-predictors-charles-di-renzo-djrgc/.

- Schulz, Leslie O, and Lisa S Chaudhari. “High-Risk Populations: The Pimas of Arizona and Mexico.” Current Obesity Reports, U.S. National Library of Medicine, Mar. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/. Accessed 8 June, 2024

- “What Is Principal Component Analysis (PCA) and How It Is Used?” Sartorius, 8 Aug. 2020, www.sartorius.com/en/knowledge/scien
snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186. Accessed 7 June 2024.