**Diabetes Data Set**

**Exploratory Analysis**

Allie Baker, abaker8@bellarmine.edu

Nikola Mrdak, nmrdak@bellarmine.edu

1. **INTRODUCTION**

Short description of the data set including a reference to where it can be found and why you chose it.

This dataset provides an extensive overview of various types of diabetes. Also included are different attributes such as medical, genetic, and lifestyle. This helps to understand the important factors contributing to diabetes. Some key features include the target, genetic markers, autoantibodies, insulin levels, age, and dietary habits. This data set can be found on Kaggle following this link: <https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset>. Neither of us were very knowledgeable on this topic, so this gave us the opportunity to learn more about this disease as it is one of the leading causes of death in America. This data set also provided us with a large range of options on how we wanted to present our data since it is such a big data set.

1. **DATA SET DESCRIPTION**

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

This data is 70,000 rows and 34 columns. A completes listing of the data is shown in Table 1. Our data types consisted of object and int64.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| V1: Target | Object | 0% |
| V2: Genetic Markets | Object | 0% |
| V3: Autoantibodies | Object | 0% |
| V4: Family History | Object | 0% |
| V5: Environmental Factors | Object | 0% |
| V6: Insulin Levels | Int64 | 0% |
| V7: Age | Int64 | 0% |
| V8: BMI | Int64 | 0% |
| V9: Physical Activity | Object | 0% |
| V10: Dietary Habits | Object | 0% |
| V11: Blood Pressure | Int64 | 0% |
| V12: Cholesterol Levels | Int64 | 0% |
| V13: Waist Circumference | Int64 | 0% |
| V14: Blood Glucose Levels | Int64 | 0% |
| V15: Ethnicity | Object | 0% |
| V16: Socioeconomic Factors | Object | 0% |
| V17: Smoking Status | Object | 0% |
| V18: Alcohol Consumption | Object | 0% |
| V19: Glucose Tolerance Test | Object | 0% |
| V20: History of PCOS | Object | 0% |
| V21: Previous Gestational Diabetes | Object | 0% |
| V22: Pregnancy History | Object | 0% |
| V23: Weight Gain During Pregnancy | Int64 | 0% |
| V24: Pancreatic Health | Int64 | 0% |
| V25: Pulmonary Function | Int64 | 0% |
| V26: Cystic Fibrosis Diagnosis | Object | 0% |
| V27: Steroid Use History | Object | 0% |
| V28: Genetic Testing | Object | 0% |
| V29: Neurological Assessments | Int64 | 0% |
| V30: Liver Function Tests | Object | 0% |
| V31: Digestive Enzyme Levels | Int64 | 0% |
| V32: Urine Test | Object | 0% |
| V33: Birth Weight | Int64 | 0% |
| V34: Early Onset Symptoms | Object | 0% |

1. **Data Set Summary Statistics**

Narrative introduction to the section.

**Table 2: Summary Statistics for Diabetes Data Set**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| Insulin Levels | 70,000.00 | 21.607443 | 10.785852 | 5.00 | 13.00 | 19.00 | 28.00 | 49.00 |
| Age | 70,000.00 | 32.020700 | 21.043173 | 0 | 14.00 | 31.00 | 49.00 | 79.00 |
| BMI | 70,000.00 | 24.782943 | 6.014236 | 12.00 | 20.00 | 25.00 | 29.00 | 39.00 |
| Blood Pressure | 70,000.00 | 111.339543 | 19.945000 | 60.00 | 99.00 | 113.00 | 125.00 | 149.00 |
| Cholesterol Levels | 70,000.00 | 194.867200 | 44.532466 | 100.00 | 163.00 | 191.00 | 225.00 | 299.00 |
| Waist Circumference | 70,000.00 | 35.051657 | 6.803461 | 20.00 | 30.00 | 34.00 | 39.00 | 54.00 |
| Blood Glucose Levels | 70,000.00 | 160.701657 | 48.165547 | 80.00 | 121.00 | 152.00 | 194.00 | 299.00 |
| Weight Gain During Pregnancy | 70,000.00 | 15.496414 | 9.633096 | 0 | 7.00 | 16.00 | 22.00 | 39.00 |
| Pancreatic Health | 70,000.00 | 47.564243 | 19.984683 | 10.00 | 32.00 | 46.00 | 64.00 | 99.00 |
| Pulmonary Function | 70,000.00 | 70.264671 | 11.965600 | 30.00 | 63.00 | 72.00 | 79.00 | 89.00 |
| Neurological Assessments | 70,000.00 | 1.804157 | 0.680154 | 1.00 | 1.00 | 2.00 | 2.00 | 3.00 |
| Digestive Enzyme Levels | 70,000.00 | 46.420529 | 19.391089 | 10.00 | 31.00 | 48.00 | 61.00 | 99.00 |
| Birth Weight | 70,000.00 | 3097.061071 | 713.837300 | 1500.00 | 2629.00 | 3103.00 | 3656.25 | 4499.00 |

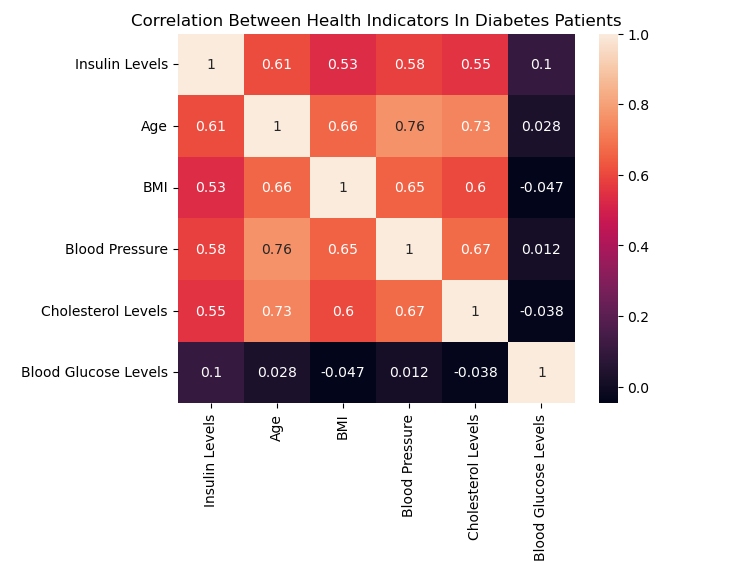
There should be a table for **EACH** categorical variable.

Table 3: Proportions for Patients by Diabetic Condition

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Cystic Fibrosis-Related Diabetes | 5,464 | 7.81% |
| Gestational Diabetes | 5,344 | 7.63% |
| LADA | 5,223 | 7.46% |
| MODY | 5,553 | 7.93% |
| Neonatal Diabetes Mellitus | 5,408 | 7.73% |
| Prediabetic | 5,376 | 7.68% |
| Secondary Diabetes | 5,479 | 7.83% |
| Steroid-Induced Diabetes | 5,275 | 7.54% |
| Type 1 Diabetes | 5,446 | 7.78% |
| Type 2 Diabetes | 5,397 | 7.71% |
| Type 3c Diabetes | 5,320 | 7.60% |
| Wolcott-Rallison Syndrome | 5,400 | 7.71% |
| Wolfram Syndrome | 5,315 | 7.59% |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 4: Correlation Table/Tables



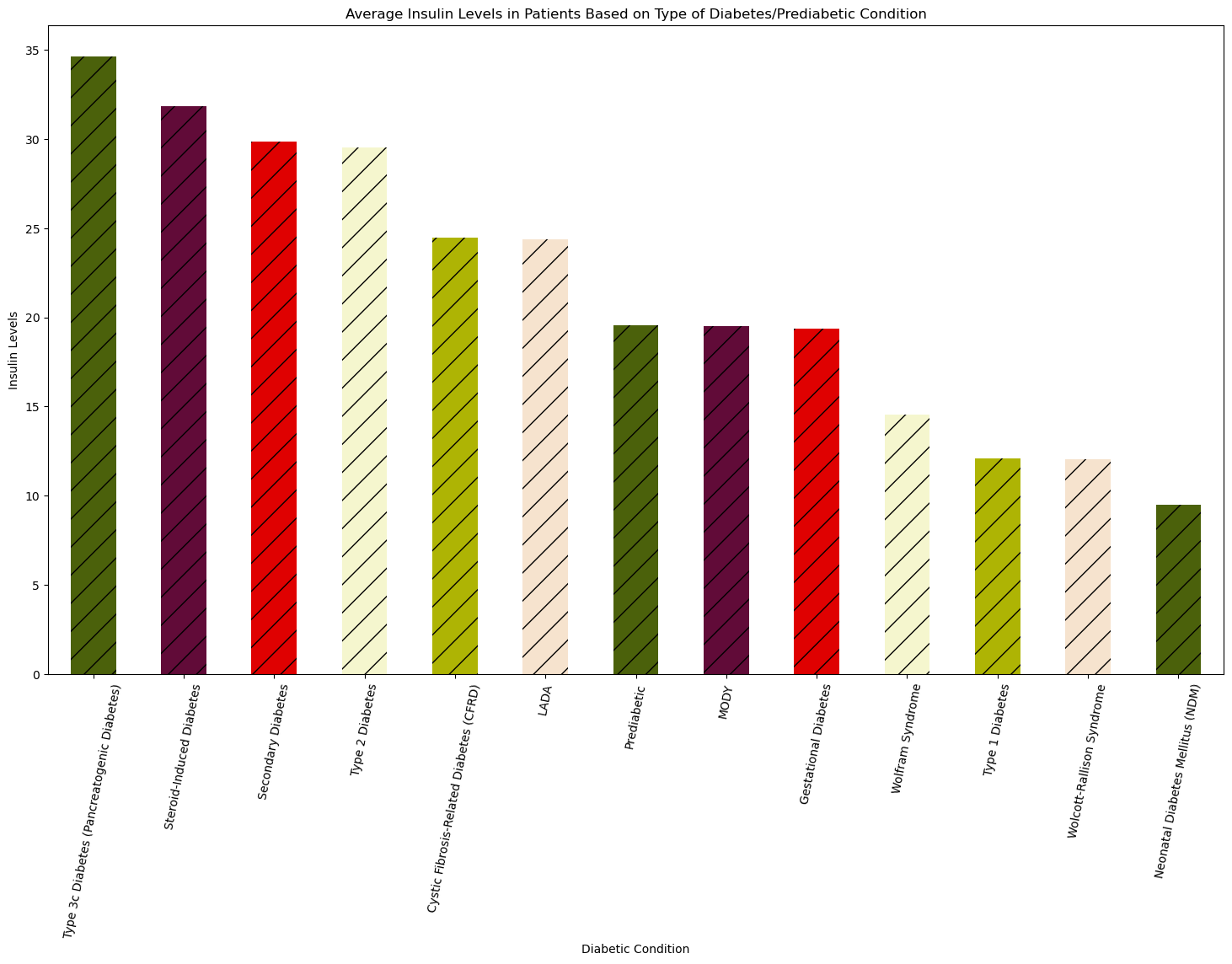
After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

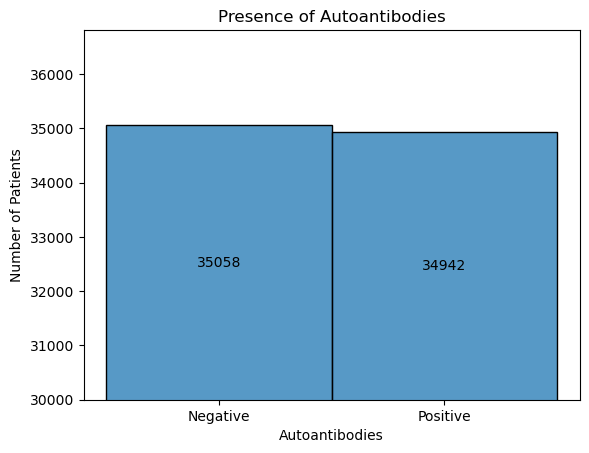
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

* 1. *Distributions*
  2. *ScatterPlots / Pairwise Plots (continuous variables)*
  3. *Barcharts (categorical variables)*
  4. *Other Plots - don’t skimp – there are likely other plots that would be useful that I haven’t already specified. Include those in this section.*

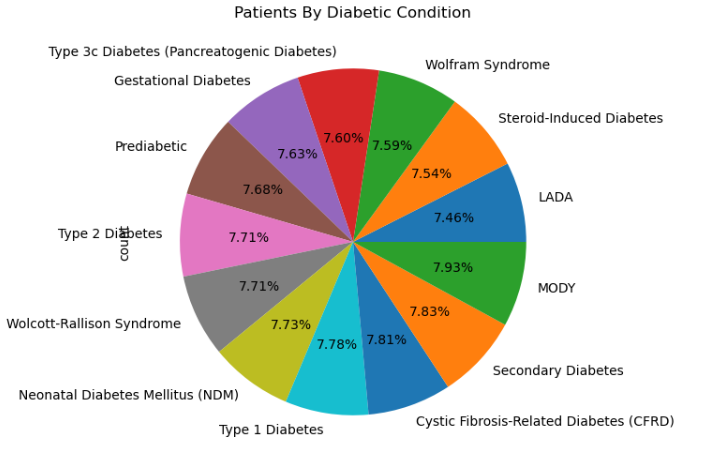
All figures should be cited formatted like this and mentioned in the text.



**Figure 1: Comparison of Diabetic Condition and Insulin Levels**

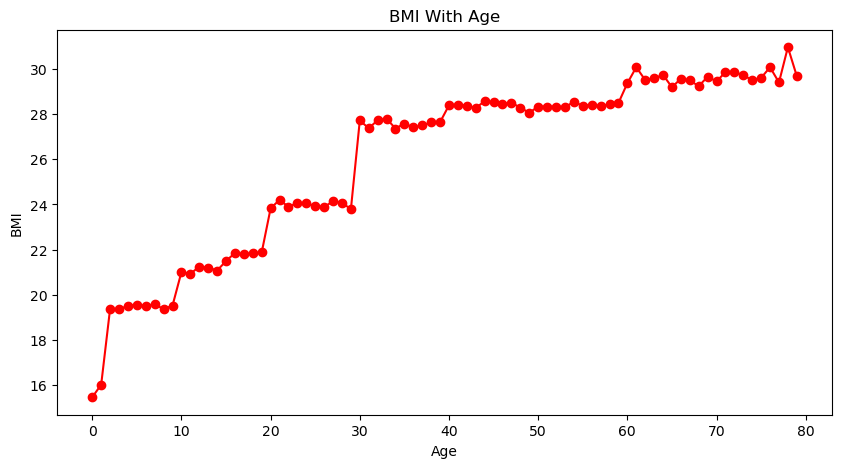


**Figure 2: Autoantibodies**

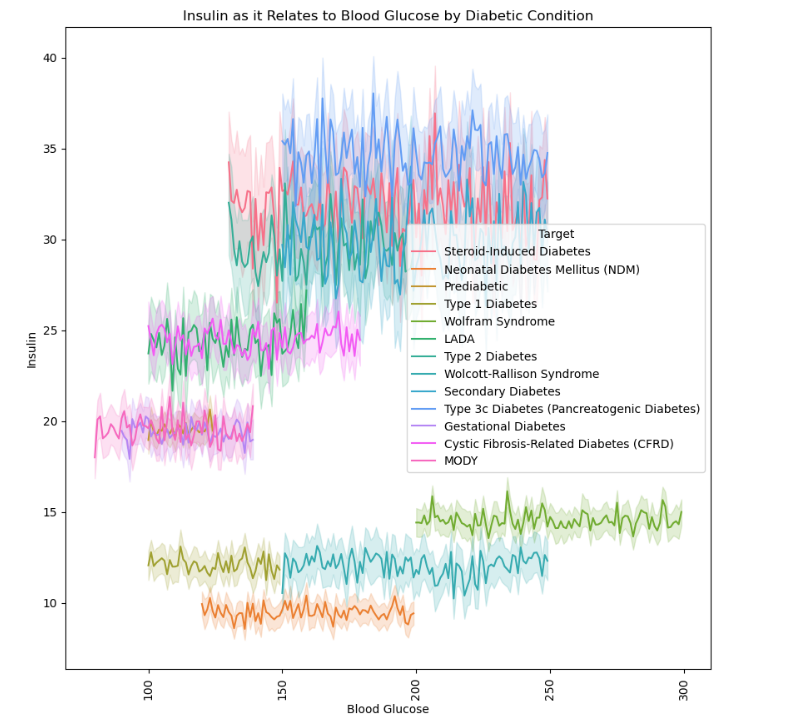


**Figure 3: Percentage of Patients with each Diabetic Condition**

**Figure 4: BMI by Age**



**Figure 5: Insulin as it Relates to Blood Glucose by Diabetic Condition**



1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.

With these analyses, key insights about diabetic condition include age, insulin levels, glucose levels, and bmi being key factors. Some conditions like type 2 involve higher levels of insulin and insulin resistance, while conditions like type 1 involve insulin deficiencies. Overall diabetes is a spectrum of conditions and health factors, but overall exploring this data taught us the significance of viewing data through a bigger lense.

[ ]: