Data Visualization Report

CS5803

Shivangi Dubey

2039934

2021-2022

**INTRODUCTION**

DATA

The original dataset consists of 253681 observations and 22 attributes.

Only the attributes in the following table have been retained for the final

Question formulation and visualization.

Changes that were made are included in the table in the description column.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| GenHlth | 1 Excellent  2 Very good  3 Good  4 Fair  5 Poor | Factor |
| Smoker | 0 No  1 Yes | Factor |
| Days of PhysActivity | Scale 1-30  Number of Days exercised in the past month  0 means not exercised | Numeric |
| Fruits | 0: No Consumption during last 30 days  1: Consumption during last 30 days | Factor |
| Veggies | 0: No Consumption during last 30 days  1: Consumption during last 30 days | Factor |
| Days of Bad Mental Heath | Numeric  0 None  1-30  During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual  activities, such as self  -care, work, or recreation?’ | Numeric |
| Sex | 0 Female  1 Male | Factor |
| Age (Ordinal) | 18y-24y  25y-29y  30y-34y  35y-39y  40y-44y  45y-49y  50y-54y  55y-59y  60y-64y  65y-69y  70y-74y  75y-79y  80y or older | Factor |
| BMI | Percentage of Body Mass Index | Numeric |
| Diabetes Binary | 0 Not Diagnosed  1 Diagnosed  Edited to No: 0  And Yes: 1 | Factor |
| HighBP | High Blood Pressure  0: No High BP.  1: Yes High BP.  Edited to No and yes. | Factor |
| High Chol | High Cholesterol  Edited to  0: No  1: Yes | Factor |
| CholCheck | Regular Cholesterol Check-up  Edited to  0: No  1: Yes | Factor |
| Stroke | Any history of Stroke.  Edited to  0: No  1: Yes | Factor |
| HeartDiseaseorAttack | Any history of heart disease or attack. Edited to  0: No  1: Yes | Factor |
| Heavy Alcohol Consumption | If alcohol has been abused in the last 30 days.  Edited to  0: No  1: Yes | Factor |
| AnyHeathCare | If the person has any healthcare plan. Edited to  0: No  1: Yes | Factor |
| NoDocbcCost | If the person hasn’t visited the doctor because of cost concerns  Edited to  0: No  1: Yes | Factor |
| Difficulty.in.Walking | If the person has witnessed any walking difficulties in the past.  0: No  1: Yes | Factor |
| Education | Did not Graduate High School  Graduated High School  Attended College or Technical School  Graduated from College or Technical School  Attempted Higher Studies after college  Completed Higher Studies after College | Factor |
| Income | <10000$  <15000$  <20000$  <25000$  <35000$  <50000$  <75000$  >75000$ | Factor |

The dataset has been taken from Kaggle data repository and the link has been included in the references.

PERSONA

The user is an endocrinologist who is interested in understanding the sample data for a better understanding of diabetic population and their lifestyle and health factors that might or might not affect their diagnosis.

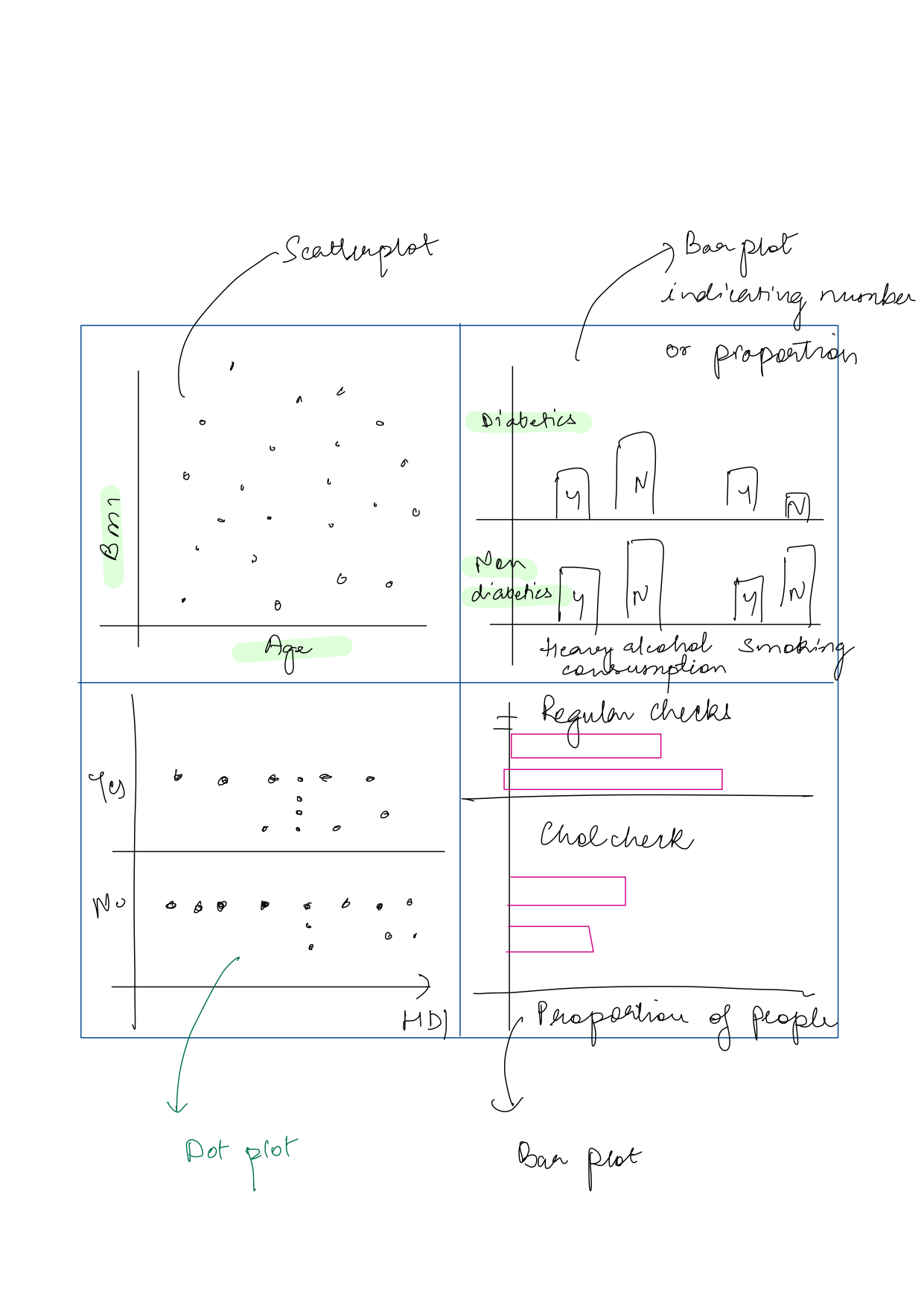
QUESTIONS

* Is there a correlation between self-care habits, or lifestyle factors with the diagnosis of diabetes in a patient.
* Is HDI something that can be correlated with diabetes diagnosis of a patient.

REQUIREMENTS

The following relationships were viewed, and their requirements have been mentioned alongside them.

* Distribution of HDI values amongst diabetic and non-diabetic population.
  + The requirement for this is a density plot as it gives a clear picture of the distribution of HDI values across categories of diabetes variables.
* Distribution of BMI across different age intervals.
  + The best plot for this was found to be a box plot for each of the age intervals.
  + It helps analyze the distribution of BMI values across age categories.
* Distribution of instances of people across categories of heavy alcohol consumption and smoking practice.
  + The requirement is to study the proportion of total diabetics and non-diabetics individually and the categories they belong to.
  + A bar plot works best for proportion analysis.
* Again, a proportion analysis of instances of diabetic and non-diabetic population.
  + Requirement for this is also a bar plot for comparative analysis.
* All the above requirements have been studied from lectures and labs. More extension work is by Stephen Few (2012).

**DESIGN**

**IMPLEMENTATION**

The steps involved in the implementation of the dashboard design are

* First the data was read in RStudio for proper data cleaning.
* The datatypes were transformed as per the attribute.
* The labels were changed for better visual readability in plots.
* The original labels in the dataset for factors were 0 and 1 which could induce unclear conclusions. This issue was resolved by giving them more obvious labels for better understanding.
* The HDI was calculated using the age, income and education columns. The details are mentioned below
  + Dimension index formula= (Actual value-min value)/(max value-min value)

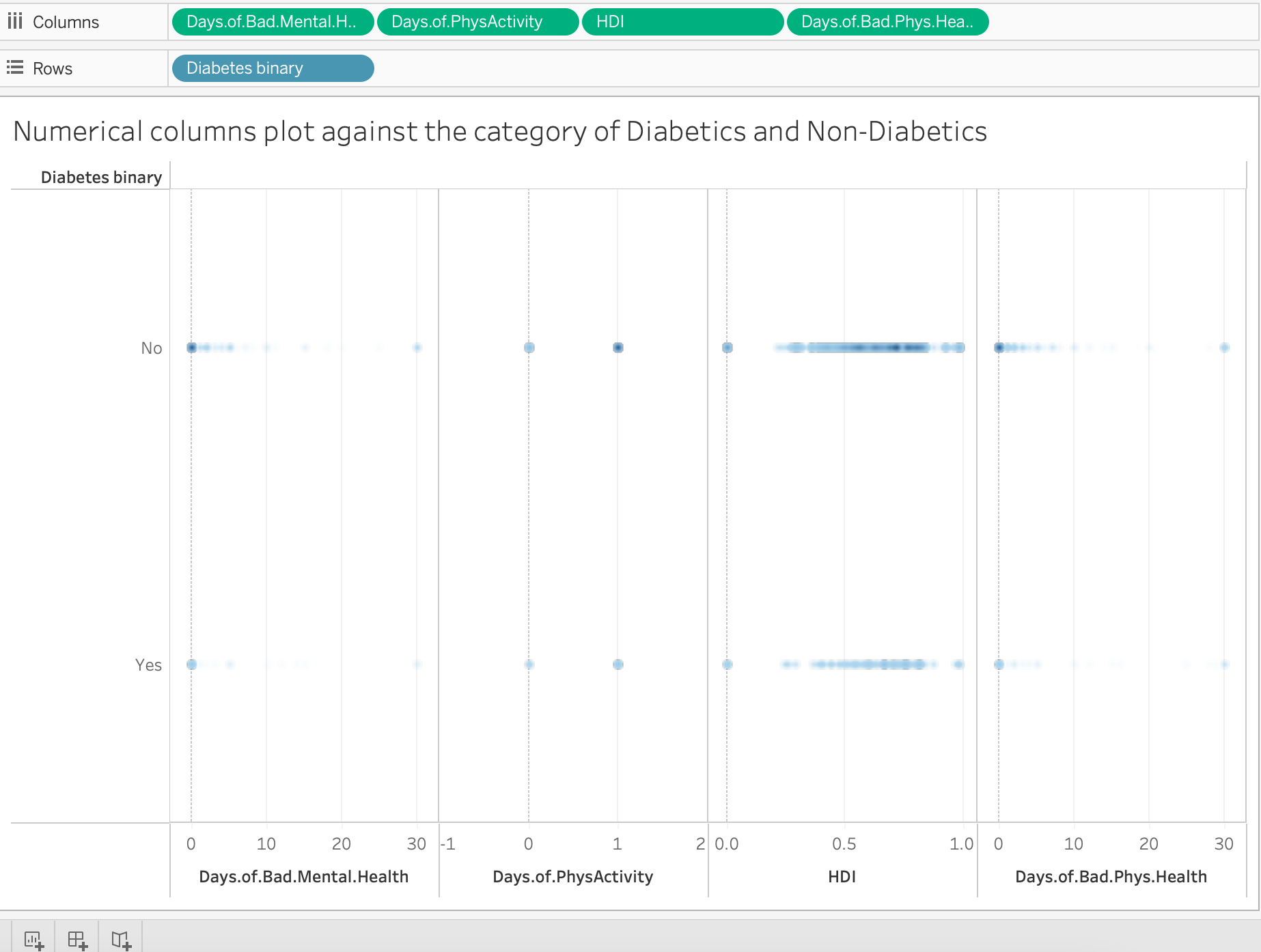
(Roser, 2022)

* + For the age index, the actual value is mean of the lower and upper limit of the age-interval. Min value is 0 and maximum was taken as 100. (As per WHO website data about average life-expectancy in U.S.A)
  + For the Education index, following years were taken for the different labels:

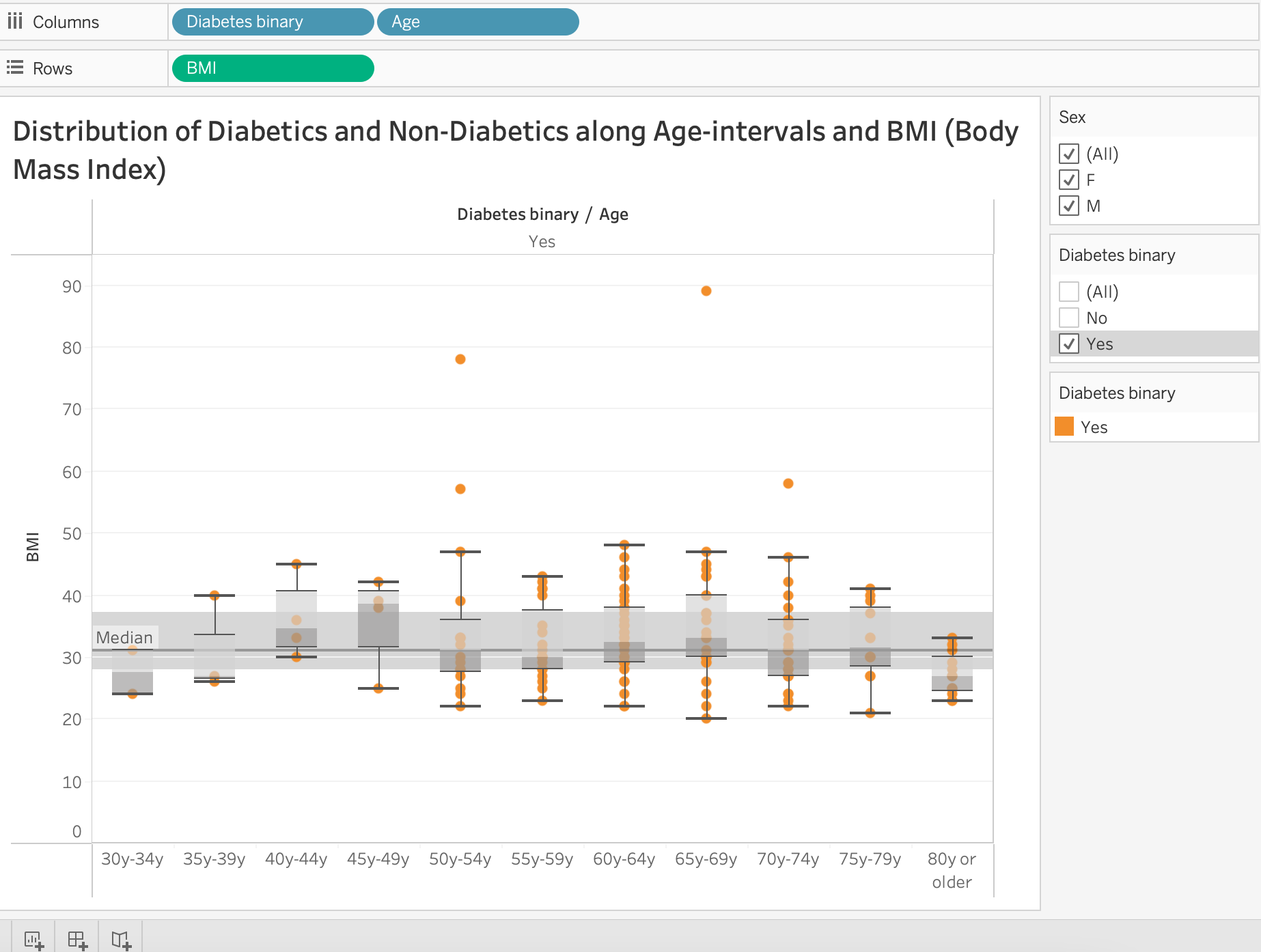
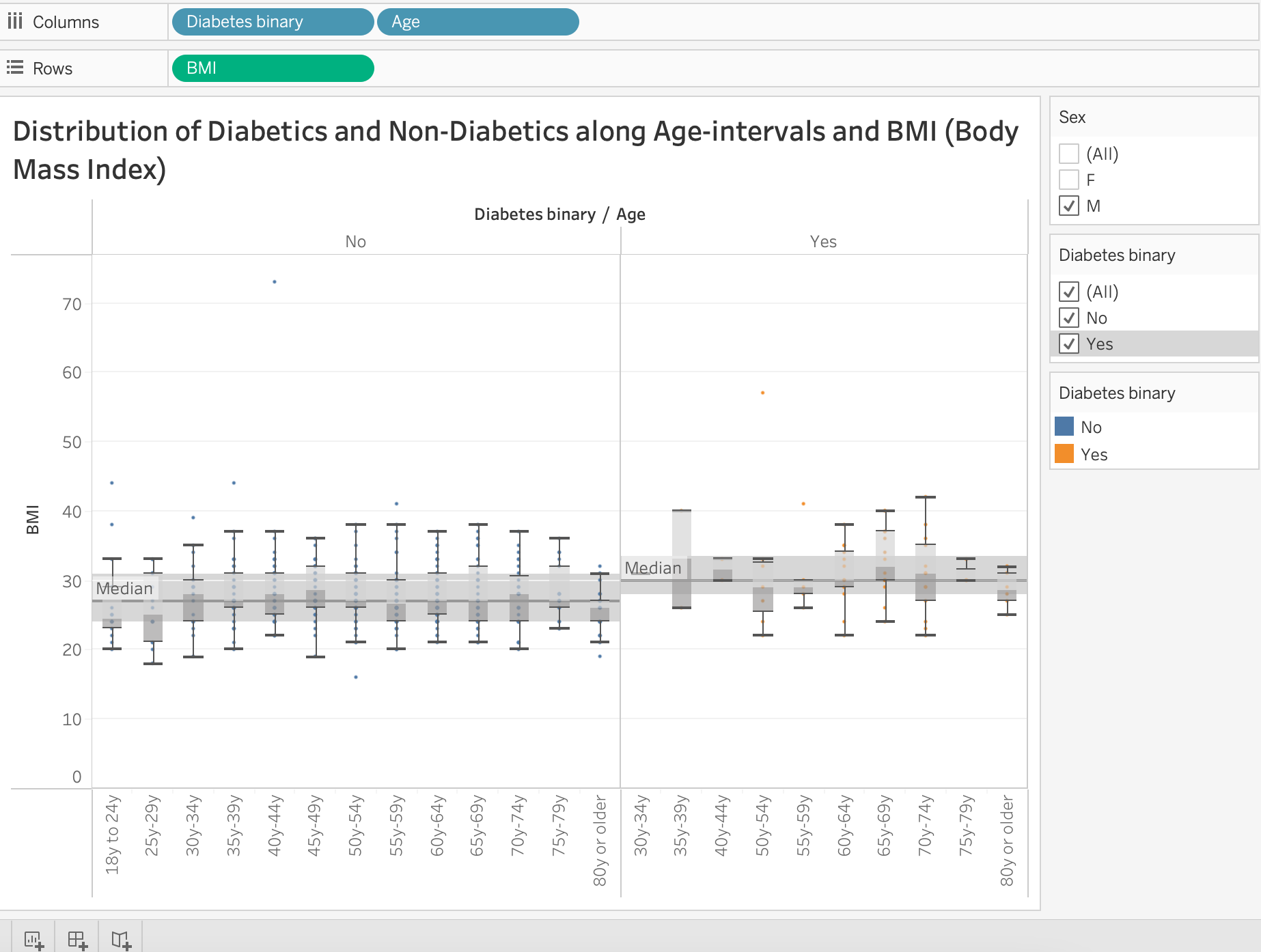
| Label | Actual Value for index (assumed years of Education) |
| --- | --- |
| Did not graduate High School | 12 |
| Graduated High School | 16 |
| Attended College or Technical School | 17 |
| Graduated College or Technical School | 19 |
| Attempted Higher Studies after College | 20 |
| Completed Higher Studies | 21 |

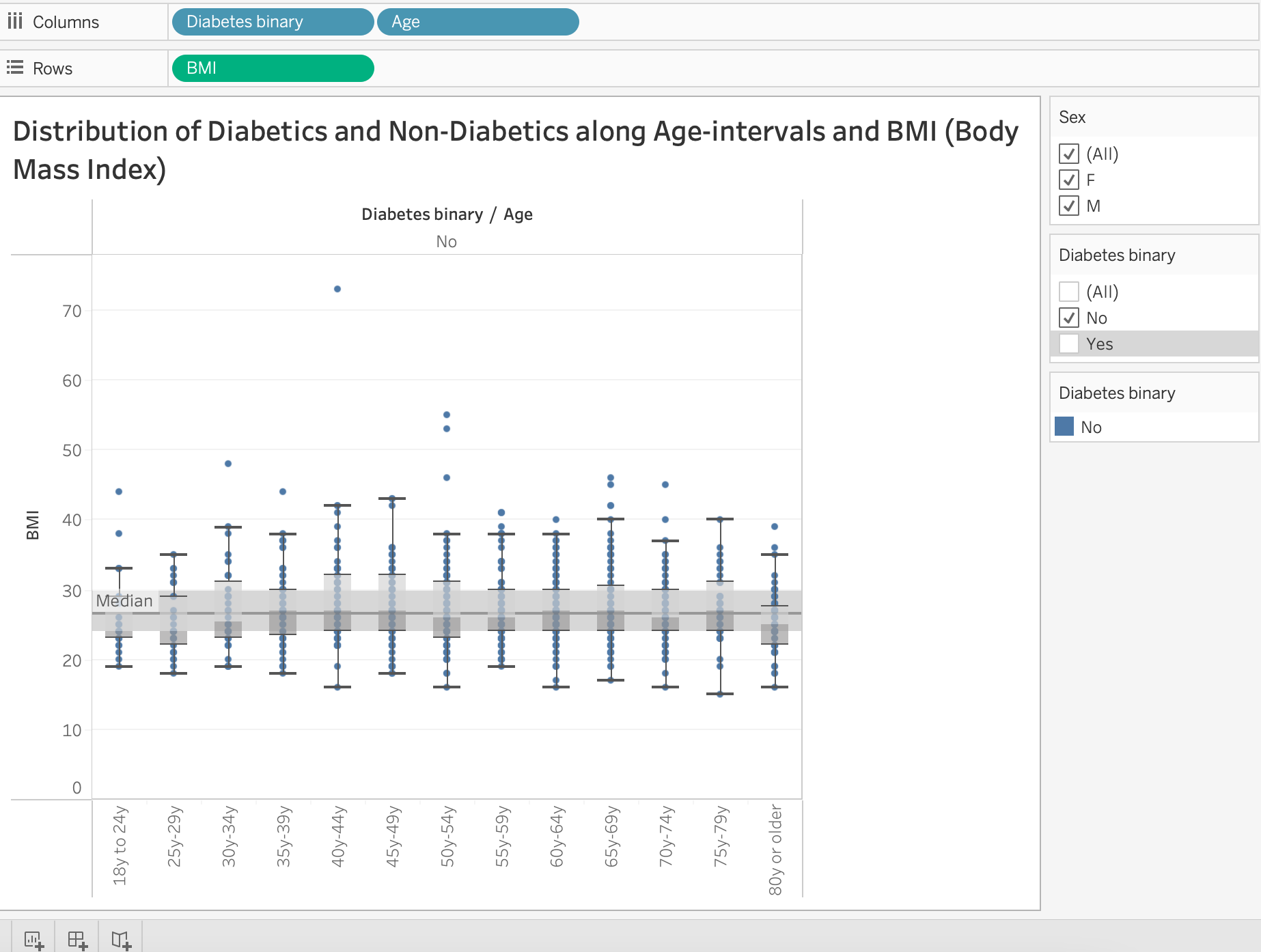
* + For Income Intervals, the average of the upper and lower limit was taken as actual value. The minimum value is 0$ and the maximum was taken as 100,000$.
* The data was then transferred to Tableau Desktop.
* The number of rows were too many (253,681 rows) and hence 1000 rows were sampled, creating a subset from the Data source tab in tableau desktop.
* Henceforth the visualizations were created.
* New column ‘Average HDI per Diabetes diagnoses was calculated using a formula. This basically gives HDI for diabetic and non-diabetic population.
* Filters for gender, diabetic diagnosis were added.
* Diabetes diagnosis was color coded.

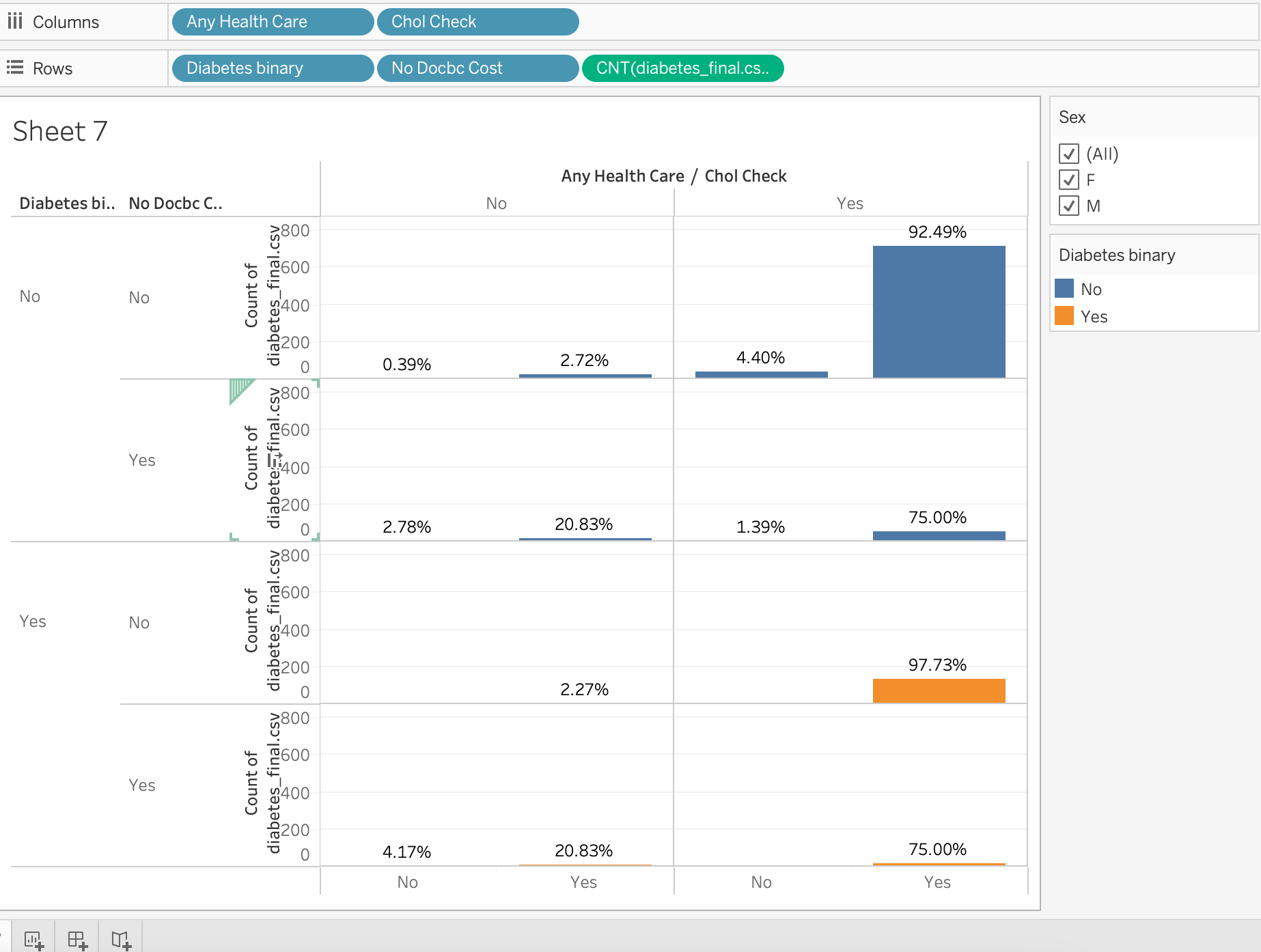
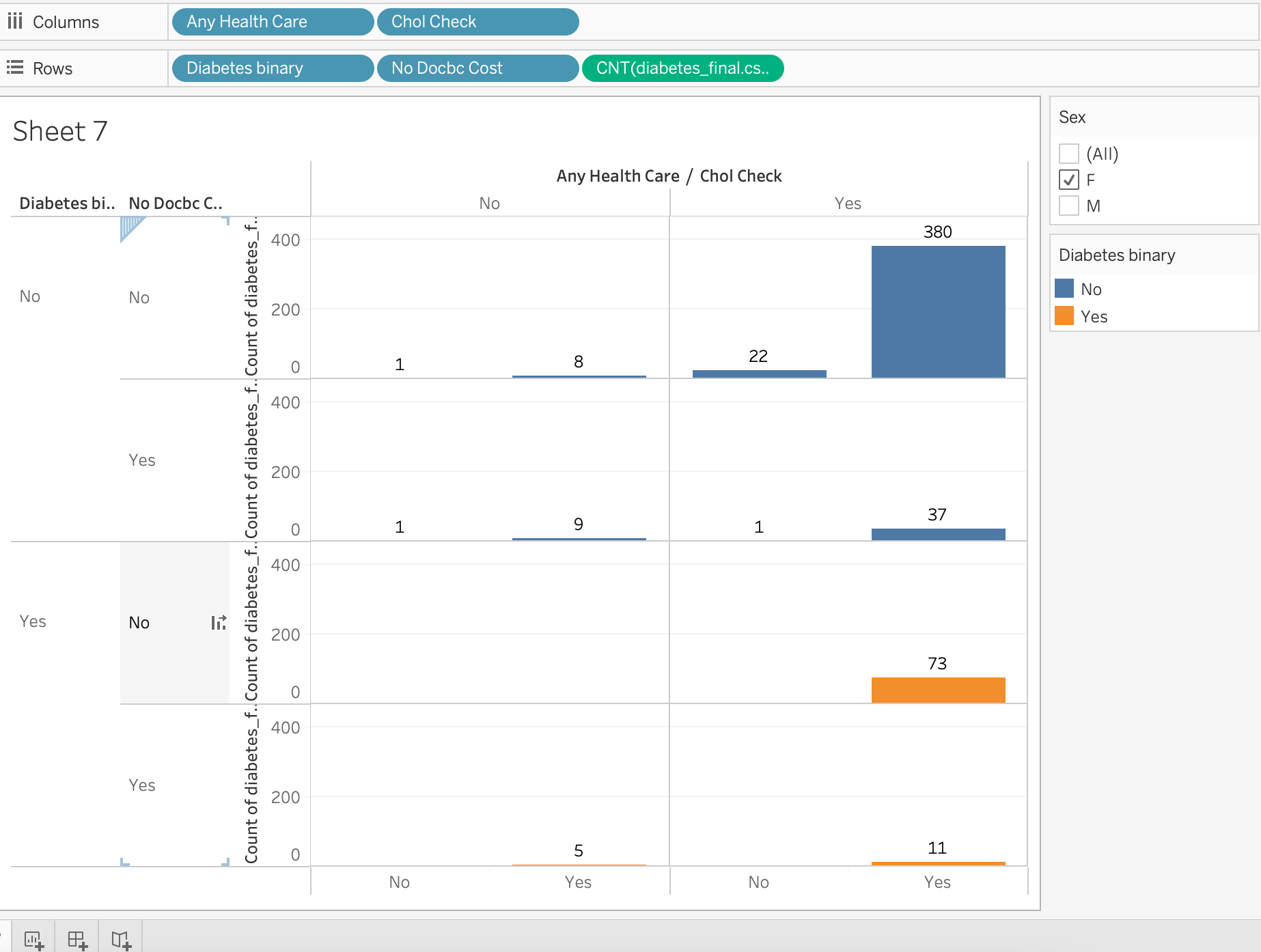
**WALKTHROUGH**

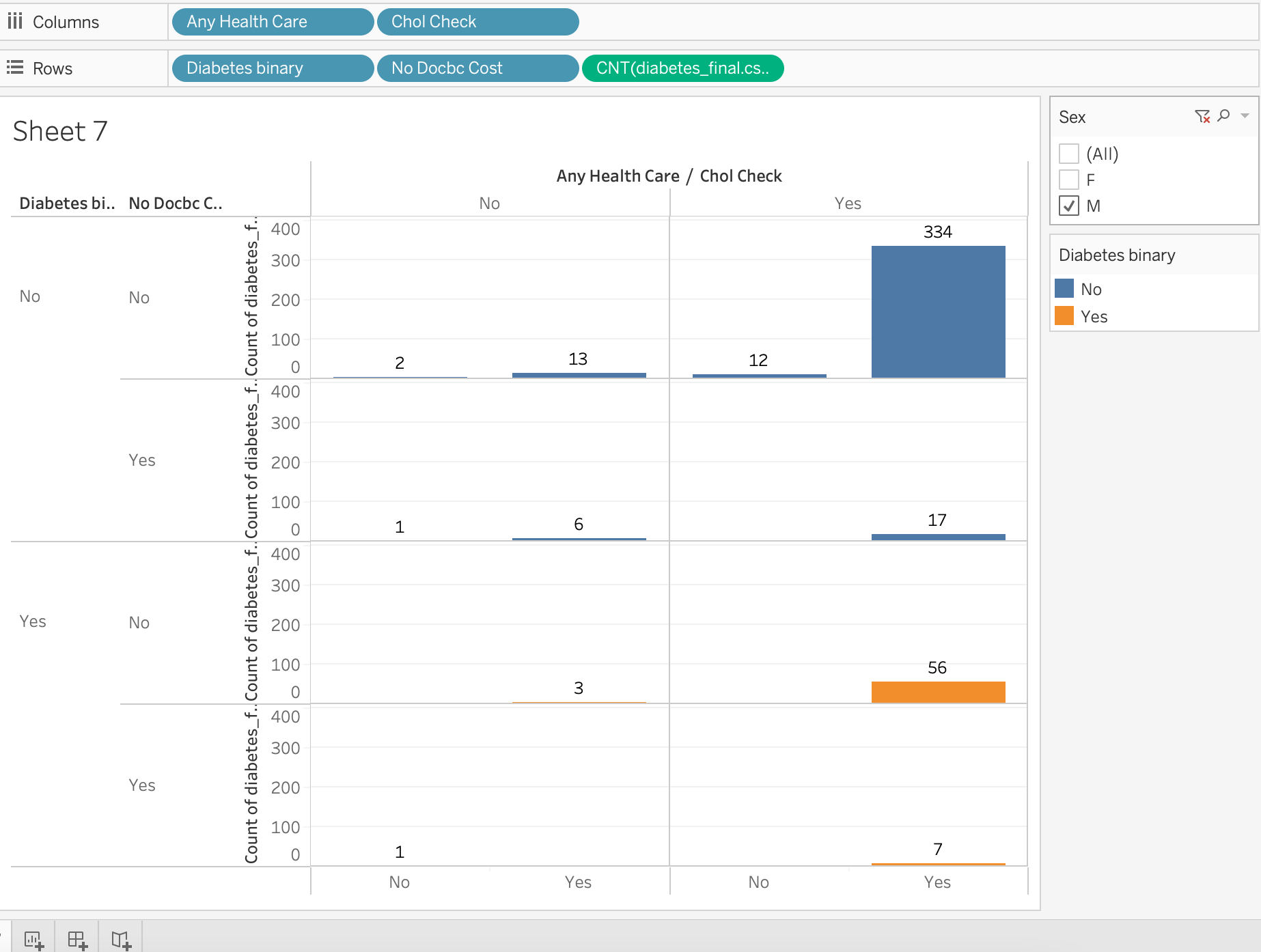
* To combine the effects of age, income and education and study the relationship of these with diabetes diagnosis, HDI is calculated using R Studio.
* First the Numerical columns are plotted against the Diabetes binary column to see any noticeable trends.
* The data for all of them except HDI seems inadequate for analysis.
* A density plot between HDI and Diabetes is further investigated.
* It shows that
  + HDI of people from the survey seems lower than average HDI of United States of America. (The data is from U.S.A. as per the codebook provided with the dataset on Kaggle)
  + For the diabetic population average HDI is lower than for the non-diabetic population.



* The Distribution of BMI is seen across age-intervals and reveals an interesting insight.
  + Among diabetic population, the median/average of BMI is lower for the male population (=30) and higher for the female population (=33)
  + Among non-diabetics the trend is opposite, i.e., for male population the BMI is higher (=27) and for the female population the BMI is lower. (=26)
  + There is more fluctuation of BMI with increasing age for diabetic people than for non-diabetic people.



* Distribution of people for heavy alcohol consumption and smoker is analyzed using bar chart. The findings are:
  + The population whose survey has been taken consist of more non-diabetics than diabetics.
  + Majority of them have no alcohol consumption issues.
  + Amongst diabetics no clear distinction can be made between smokers or non-smokers.
* Distribution of diabetics and non-diabetics is analyzed as per their self-care habits. It is found that:
  + Majority of the people diabetic or non-diabetic do not avoid doctors due to cost concerns, have a proper healthcare plan and regularly monitor their cholesterol levels.
  + There is no difference in trends between male and female population.
  + There are more females amongst diabetics and non-diabetics who avoid doctor visits due to cost concerns than males amongst non-diabetics and diabetics respectively.



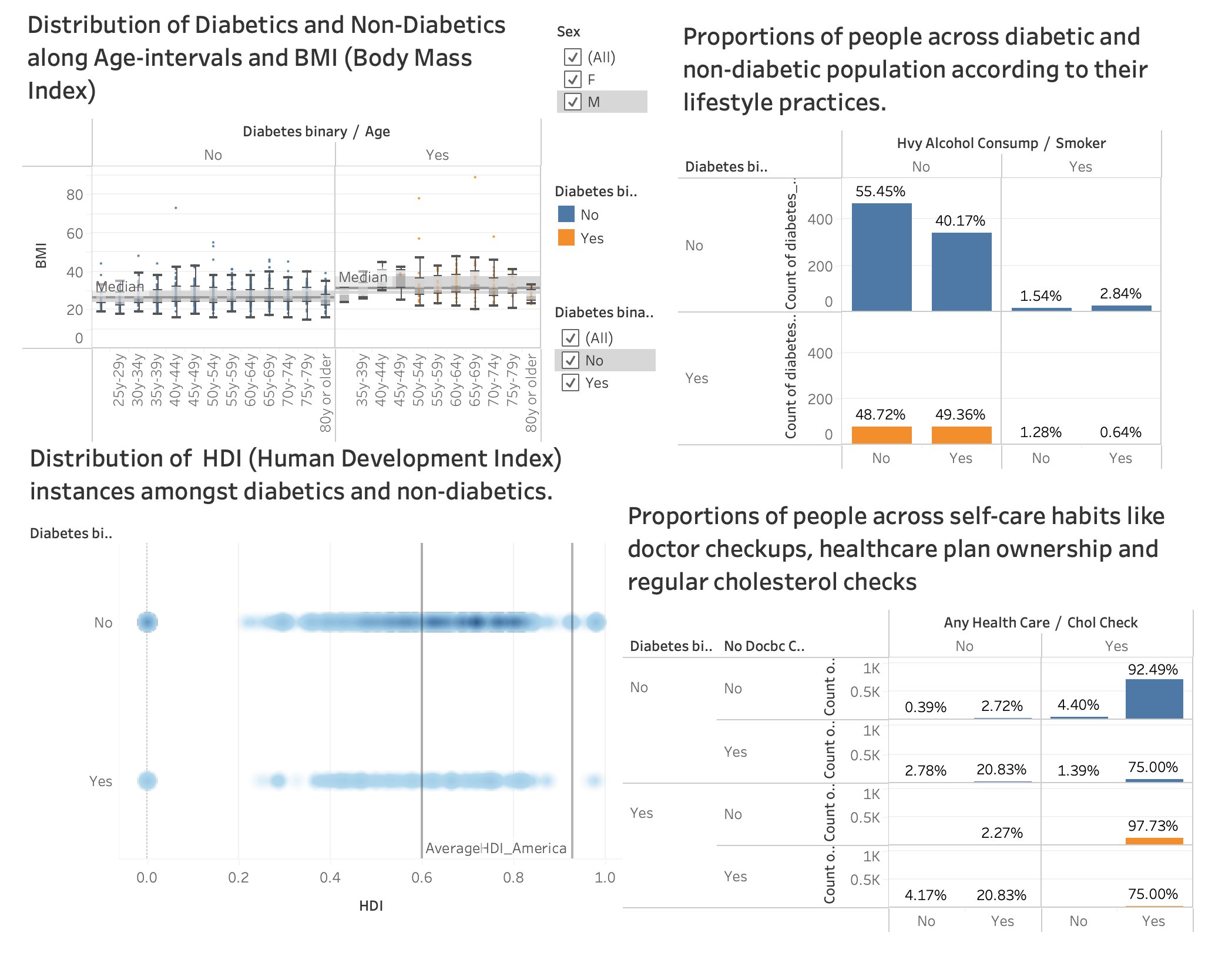
The **major reflections** about diabetics and non-diabetic populations were:

* This population for which the survey is done, has an overall lower HDI than the American population.
* Diabetic population has lower HDI than non-diabetics on an average.
* Females have higher BMI amongst diabetic population than males.
* The fluctuation in BMI with age is greater in diabetic population.
* Heavy alcohol consumption or smoking could not be correlated with the diagnosis of diabetes.
* Approximately 88 % of diabetic males are more vigilant about self-care as opposed to 86% of female diabetics.

Hence **the answers** to the questions specified above are as follows:

* Interesting HDI trends are observed between diabetic and non-diabetic population. Basically, diagnosis of diabetes points towards a lower HDI.
* Self-care practices are more prevalent in diabetics than non-diabetics. Males are more vigilant than females.
* Lifestyle choices of heavy alcohol consumption or smoking do not point towards a substantial correlation. In fact, whether diabetic or non-diabetic, more people opt for a healthier lifestyle than indulging themselves abusively.

The final dashboard has been added on the next page.

****

**REFLECTIVE DISCUSSION**

CRITICAL REFLECTIONS

* The population of diabetics and non-diabetics was not balanced and hence comparison between the two could only be done on a proportion of the total basis.
* Lack of exact numerical data and presence of interval data instead (for income, education and age) is a bit inconvenient for the calculation of dimension indices and for final calculation of Human Development Index (HDI). This kind of calculation can be done in Tableau too.
* The above is not just inconvenient but also more inaccurate than it would have been with exact data.
* HDI is an indicator of life expectancy and standard of living and not of inequality or poverty and hence its correlation with diabetes diagnosis does not imply any lack of financial resources.

STRENGTHS OF TABLEAU

* + Almost every visualization I envisioned was somehow possible in tableau.
  + I found it a very compatible tool for intuitive visualization.
  + New columns can be created without repetitive task of importing and exporting dataset.
  + I was able to adjust my sample size very easily.

WEAKNESSES OF TABLEAU

* + It was quite a task to design formulas for my requirement.
  + I had to seek R studio’s help for some manipulations like relabeling data and checking for nulls.
  + Also preferred to do major calculations regarding HDI via R.

The overall learning experience was quite good. The book by Stephen Few was more than helpful for molding my ideas regarding visualization and insights. I have been always confused about which visualization suits what. This course and its coursework have helped a great deal with that.

I intend to learn more tools for visualizations and practice my skills on multiple projects like these for better practice with data and representation.

**REFERENCES**

* www.kaggle.com. (n.d.). Diabetes Health Indicators Dataset. [online] Available at: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.
* Roser, M., 2022. Human Development Index (HDI). [online] Our World in Data. Available at: <https://ourworldindata.org/human-development-index#:~:text=Aggregating%20the%20four%20metrics%20to%20produce%20the%20HDI&text=The%20HDI%20is%20calculated%20as,and%20expected%20years%20of%20schooling).> [Accessed 15 April 2022].
* Xie, Z., Nikolayeva, O., Luo, J. and Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Preventing Chronic Disease, 16.
* Brath, R. (2003). Paper landscapes: a visualization design methodology. Visualization and Data Analysis 2003.
* Green, M. (1998). Toward a Perceptual Science of Multidimensional Data Visualization: Bertin and Beyond. [online] www.semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Toward-a-Perceptual-Science-of-Multidimensional-%3A-Green/b5fd6166cf2264100a403e1fda019d9e9c5c6303 [Accessed 10 Apr. 2022].
* Stephen Few (2012) Show me the numbers: Designing tables and graphs to enlighten. Oakland, CA: Analytics Press.

‌

‌

‌