Script for presentation:

We are Erika, Mohamed and Sri, and our project is examining the risk factors strongly….

For our presentation, we will go through some background of the project along with our research question, some explanation about the data, and then we’ll get through the EDA (which will show the data summary, some graphs as well as chi square tests – maybe don’t need to say this) and then we’ll conclude with some notes for project 2.

So, a little background – according to the CDC, diabetes is a common chronic condition causing low insulin level in humans– that’s the hormone that helps us break down glucose(sugar) into energy. This cause diabetics to have high level of sugar in their blood. Many complications can result from diabetes, including heart disease, vision loss, kidney disease, limb amputation, etc.

This affects millions of people, particularly in the united states, where 37.3m people have diabetes and 96m adults have prediabetes (where your blood sugar level is high but not enough to be diabetic). There is no cure. Treatment to help control diabetes include lifestyle changes (diet & exercise) and medication (pills, insulin injections). Identifying which factors (age, lifestyle, etc) correlate with diabetes can help with diagnosis and treatment. So, our research question for the project is: which risk….

I want to talk a bit about our data. Every year since 1985, the CDC conducts…

The data we use came from original survey from 2015, which contains….

Our data set was subsetted and cleaned by …, which we got from Kaggle. It has…. To subset…

*Next part*

As mentioned above, since the data mostly consists of categorical variables, bar charts and histograms were used to best describe the distribution of the variables. In most cases, participants with AND without pre/diabetes were plotted together for easier comparisons.

First, we looked at the bio-demographic variables.

Age and education level might be correlated – usually older people tend to have been in school for longer

*(remember to rephrase the variable scales)*

*(talk about the distribution if continuous)*

*Notes for p2*

All the variables we looked at have correlation to diabetes and are worth looking at when we build our models for project 2

We could also compare our results with existing literature to see if they match up

Before the model building, we still need to test whether the independent variables are correlated to each other to decide to include or exclude

Since diabetes is binary, and many variables are binary/categorical, we probably would have to log transform