Slide 1 : **Kevin** Introductions

Good evening everyone! We are the Drops of jupyter group since our analysis is on diabetes and socioeconomic factors. (get it? Drops like the drop of blood you need for a glucometer???).

Slide 2 : **Manny** Why we chose: Make it personal

Slide 3: **Elina**:

These articles informed our objective, we learned that The world bank, the diabetes journal, and several medical studies and journals have identified a link between low socioeconomic status and/or racial/ethnic minorities having worse health outcomes.

Slide 4: **Kevin** : question we started with:

The questions we started with are on

* the demographics and socioeconomic factors correlated with diabetes
  + Who would most likely get diabetes
  + And if we could successfully predict diabetes based on demographics and socioeconomic status
* Finally what changes/solutions could be made to positively impact this issue?

Slide 5: **Elina**; Technologies we used

Slide 6: **Amy**: Data source: 2 sentences

We downloaded NHIS survey data for 2021. This data came from a survey in which individuals self identified their status. Some questions were subjective, some were completely unrelated to diabetes. We had over 29,000 rows and 600+ columns. We cleaned this initial data, and loaded into our database. We attempted to use matplotlib but since so many of the variable were categorical we decided the visualizations would be more interesting in tableau.

Now we will take you to our live dashboard to demonstrate how it works and the rest of our analysis.

(7min)

Open the dashboard –

**Manny**, first, give us a tour of the dashboard. Explain how we can navigate through sections with the links in the navigation bar. Then go into our data sample that is meant to be filtered for user to explore. However, one roadblock we encountered was not being able to get the data to filter with this tool.

* Click on data
* Manny tell us about the filter (maybe? Leave out in interest of time)
* Scroll up to dashboard

**Kevin** will present the tableau and explain our exploration process and the visualizations we created. (Explanation of how data correlates etc.)

We want to help bring more awareness to those that weren’t included in the study

-All the charts in this dashboard are scaled to percentage totals which means they are comparable.

Dashboard 1

In this pie chart we have the total sample shown as diabetic and non diabetics. The blue color scheme is non diabetics and the brown color scheme are diabetics which is consistent with the other visualizations in this storyboard. We had a sample of about 30,000 participants where about 10% are diabetic.

**Ourworldindata.org**, has records of a 10.7% population rate of diabetes which is consistent with our sample.

Now onto BMI, BMI is divided into 4 groups (under, healthy, over, obese). Since underweight had a very small size, it is almost negligible. So the other 3 almost equally take up ⅓ of the chart.

This bar chart shows the correlation between BMI and Diabetes. It is a linear relationship, so as the BMI increases, the incidence of diabetes increases.

Dashboard 2

AIAN is an American Indian/Alaskan native.

In this dashboard we can make arguments that people with less access have higher incidences of diabetes. The races with higher incidences are Hispanic, black/African American, and AIAN, listed in increasing order.

Dashboard 3

On the top of this section, we have education correlation with diabetes. Here as the level of completed education increases, the incidence of diabetes decreases.

On the bottom of this section, we have the correlation of the poverty ratio with diabetes.

The poverty ratio is the **income divided by the poverty threshold**, so the higher the poverty ratio, the wealthier you are.

So Looking at this chart, we can see that as you increase in wealth, the incidence of diabetes decreases.

Dashboard 4

This dashboard takes in 3 variables, education, poverty ratio and diabetes.

The size of the circle relates to the number of observations falling into these categories.

The takeaway from this dashboard is that as education and poverty ratio increases, diabetes incidences decrease.

—----------first explain machine learning models that the app relies on (predictor)

Go Back to view Slide 7

**Amy** will explain the machine learning model we’ve selected and why she/we chose it. Please tell us about the other machine learning models you tried. What struggles did you experience?

* Accuracy, precision, and recall scores.
* After training the model, we created binary files that our app can use to predict the probability of some new individual having diabetes.

Now go back to the dashboard to demonstrate the app –

**Elina** – talks about how app works. Display Terminal and Dashboard side by side so we can see the user view and the back end view simultaneously.

This app is a solution that helps individuals take matters into their own hands. The app cannot replace visiting your primary care doctor. However, if you we are unable to see a doctor, this tool allows you to measure your probability of having diabetes. It aims to help people get diagnosed as early as possible, so people can live longer with better health outcomes.

Conclusion: What we would do differently in the future.

We were hyperfocused on the SES variables and so overlooked the opportunity to run a correlation function on all 600 columns of our original data set. In the future we’d likely run that first to help us identify important variables

1. Our model has a very large number of categorical variables which can affect the predictive power of our model. We would like to test SVM and RFC models for future versions of our app since those can sometime handle categorical variables more reliably.
2. We have built a framework here that would allow us to integrate a CDC API in the future and load new data into our database when ever CDC releases it and then further train our data on many more observations. This will further improve our model’s scores.