Diabetes in the US

Education and Preventative Measures

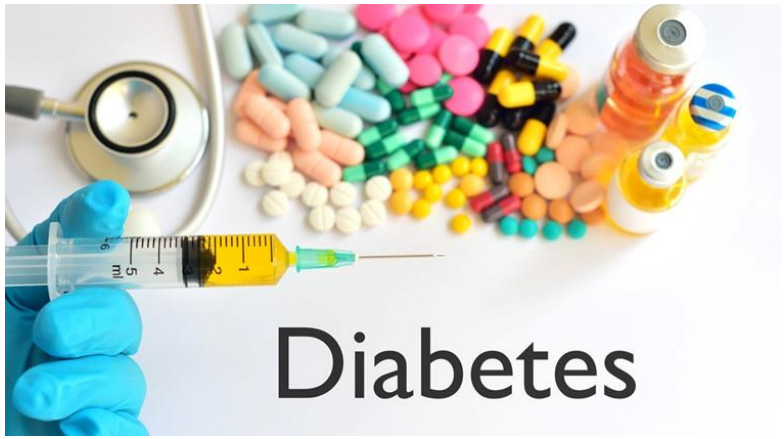
Chris Johnson, August 2018
UCLA- Data Science (COM SCI X450.1)

Agenda

- Problem & Proposed Solution
- Choosing MSAs
- Finding & Educating High-Risk Patients
- Appendix

Problem: Diabetes Prevalence in the US

- 100+ million Americans have diabetes or prediabetes, leading to significant health issues
- Diabetics spend 2.3x more than non-diabetics on healthcare
- ***However, 90%+ of diabetes is Type II and is preventable***



Source: American Diabetes Association

Solution: Focus Education Efforts

1. **Find Target MSAs** *(Google dataset)*
   - Identify MSAs with the lowest recent “diabetes index” scores

2. **Understand Key Demographics** *(NHANES dataset)*
   - Use health data to find risk factors and demographics of diabetes patients

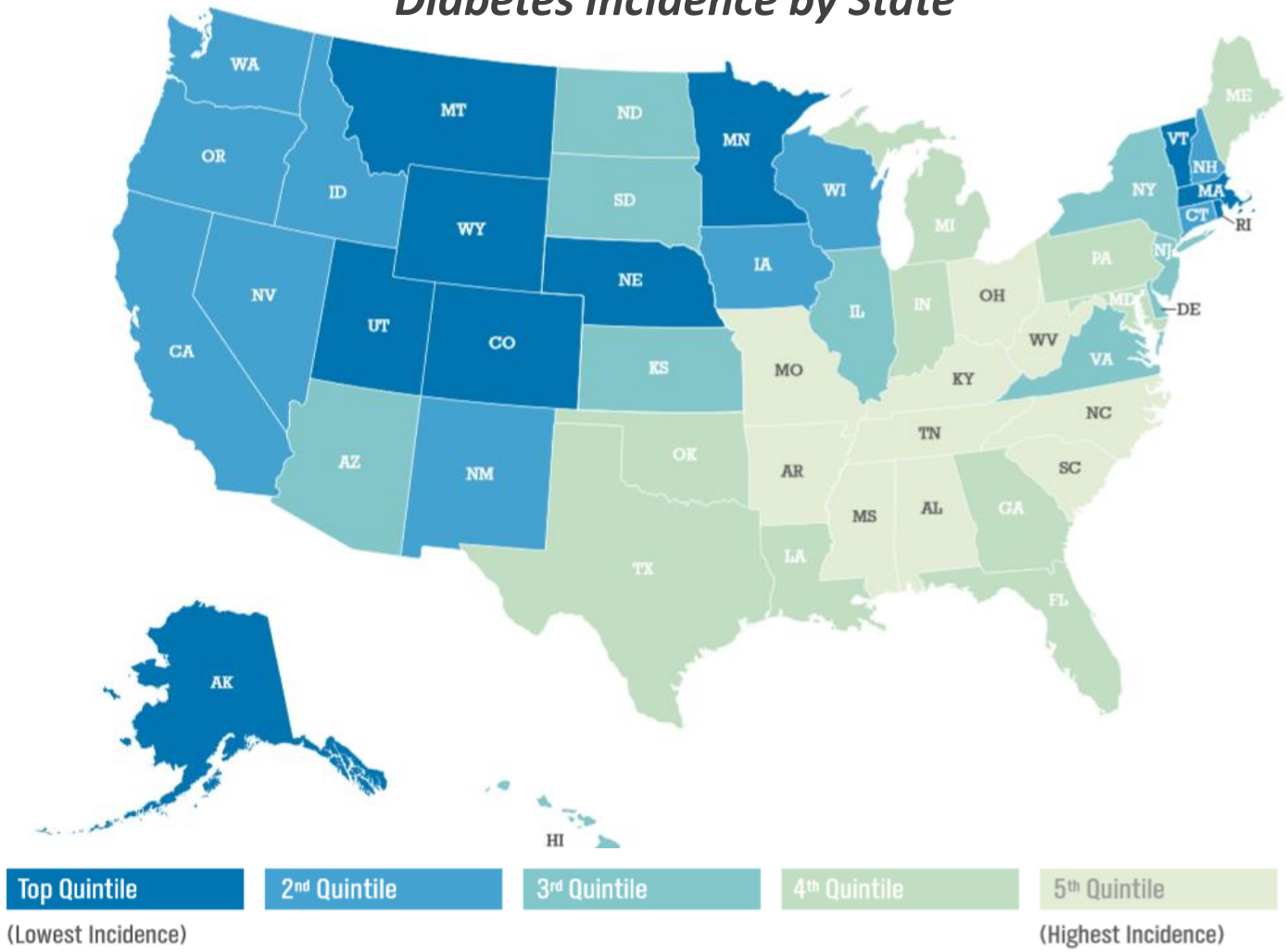
3. **Tailor Education Plans to At-Risk Individuals in Target MSAs**

Agenda

- Problem & Proposed Solution
- Choosing MSAs
- Finding & Educating High-Risk Patients
- Appendix

Where to Focus Our Efforts?

Diabetes Incidence by State



1. **States with lots of diabetics**
2. **MSAs where education will make an impact**

- *Problem: How do we find where education will make an impact?*

Map Source: Gallup 2015

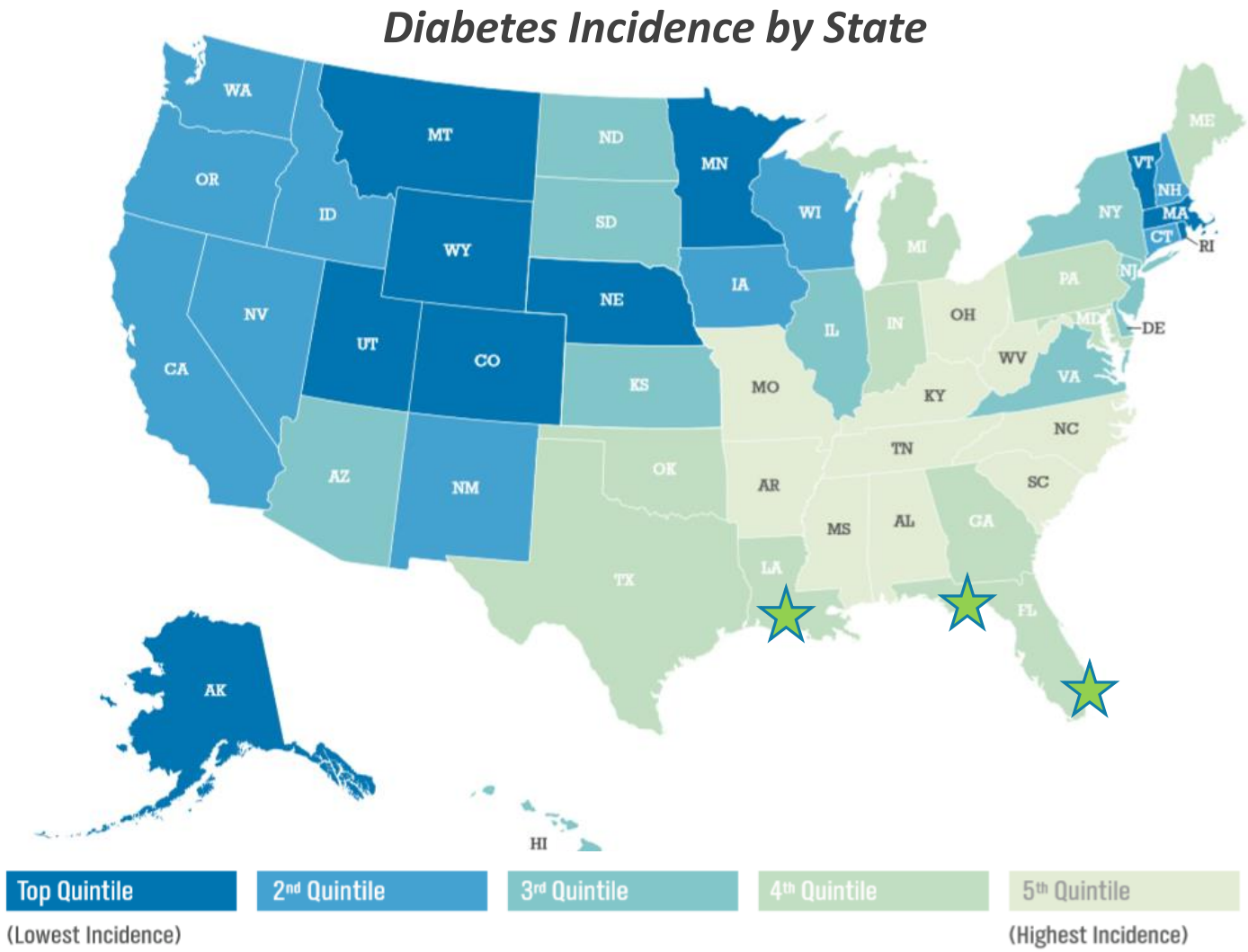
Using Google Search Data to Screen MSAs

- **Google "Diabetes Index" dataset**
  - Index compares Google search interest vs. incidence of that condition by MSA

- **A low score could mean those MSAs are not as proactive on seeking diabetes info**
  - Opportunity for education efforts!

- **Sorted MSAs by lowest 2014-2017 avg. scores**
  - Filtered out MSAs with >avg. 2017 scores & 1 outlier
  - Selected first 3 MSAs located in top 2 quintile states for diabetes incidence in the Gallup map

The Result: Our Top 3 MSA Candidates



- Our top 3 candidates, based on our diabetes incidence and impact criteria are:

1. **Lafayette, LA**
2. **Miami-Ft. Lauderdale, FL**
3. **Panama City, FL**

Map Source: Gallup 2015

Agenda

- Problem & Proposed Solution
- MSAs to Focus On
- Finding & Educating High-Risk Patients
- Appendix

Identifying the Right Patient Demographics

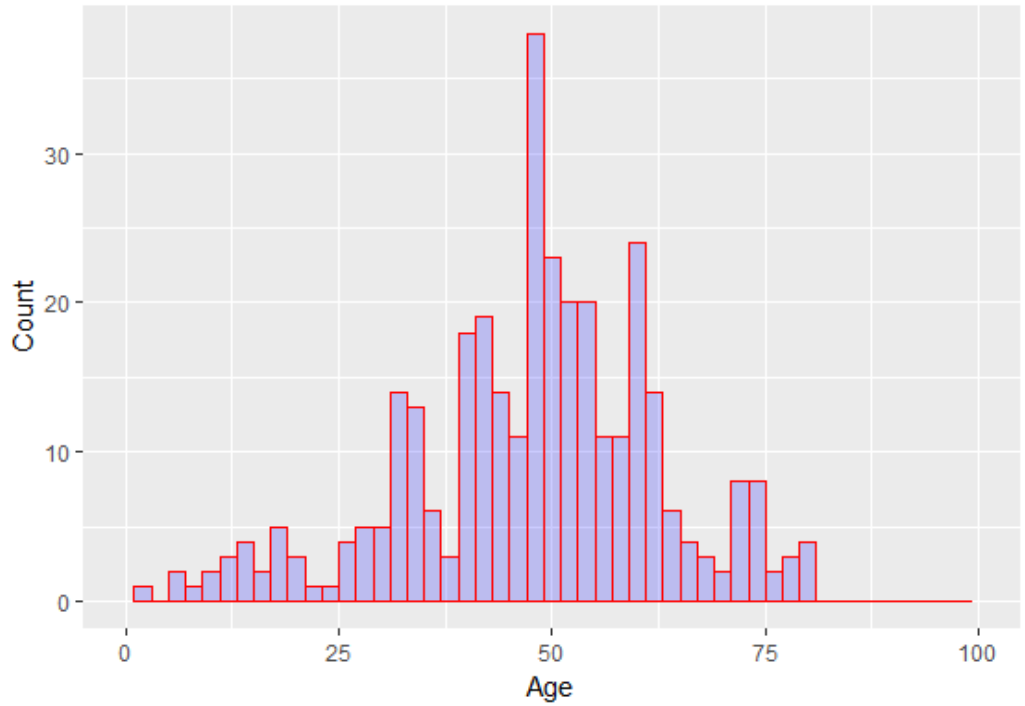
- **CDC's "NHANES" dataset can provide clues**
  - Survey responses on 75 health variables from 10,000 patients
  - Sampled to reflect the US population
- **Utilize R to analyze data on diabetic patients**
  - Look for major risk factors and characteristics

Diabetes is Usually Diagnosed Later in Life

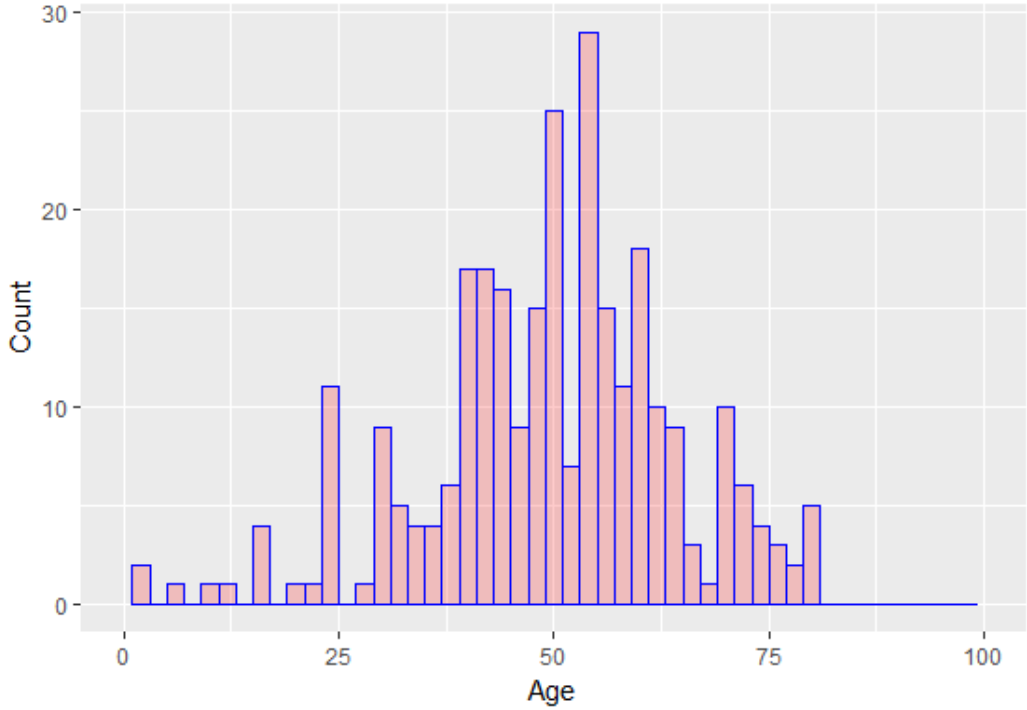
Our goal is to educate years before the average diagnosis

For us, a good range is: Men: 34-47; **Women:** 36-47

Males- Age of Diabetes Diagnosis



Females- Age of Diabetes Diagnosis

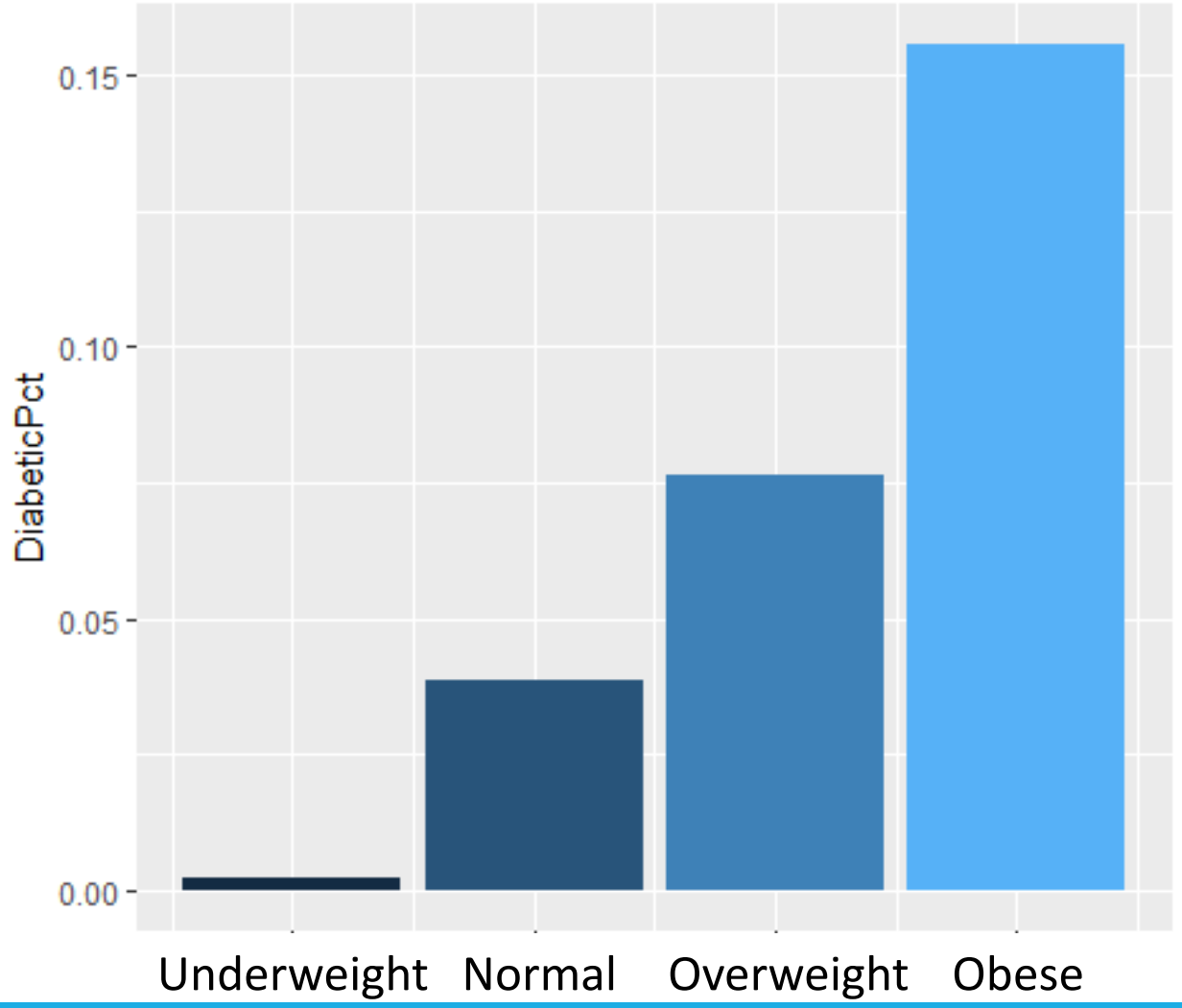


Lower Bound: 3 years before the 20th percentile; Upper Bound: 40th percentile

A High BMI Strongly Increases Diabetes Risk

- **Obese patients have 4x the risk of diabetes than patients with normal BMIs**
- **16% of Obese Individuals Have Diabetes**
- **CDC's BMI Categories:**
  - Underweight: Below 18.5
  - Normal: 18.5-25
  - Overweight: 25-30
  - Obese: 30+

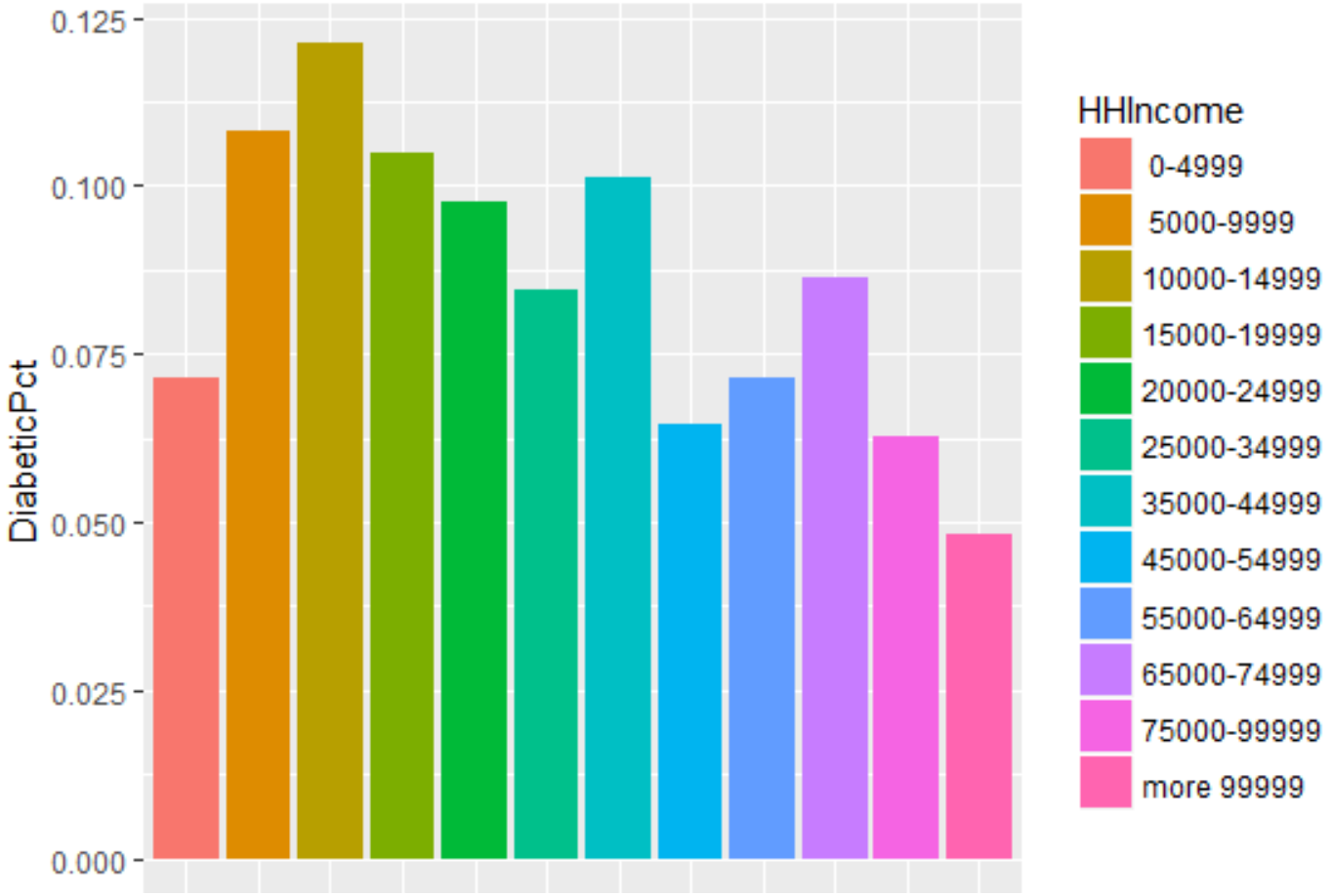
% of Patients with Diabetes by BMI Category



What About Income or Gender?

- **Diabetes is much more common in lower income households**
  - Highest for the $5k-$45k range
  - Adjusted for inflation that's about $6k-$52k in 2018 USD
- **Diabetes was prevalent for both genders**
  - 7-8% of both males and females have diabetes

% of Patients with Diabetes by Household Income ($ USD 2009-2012)

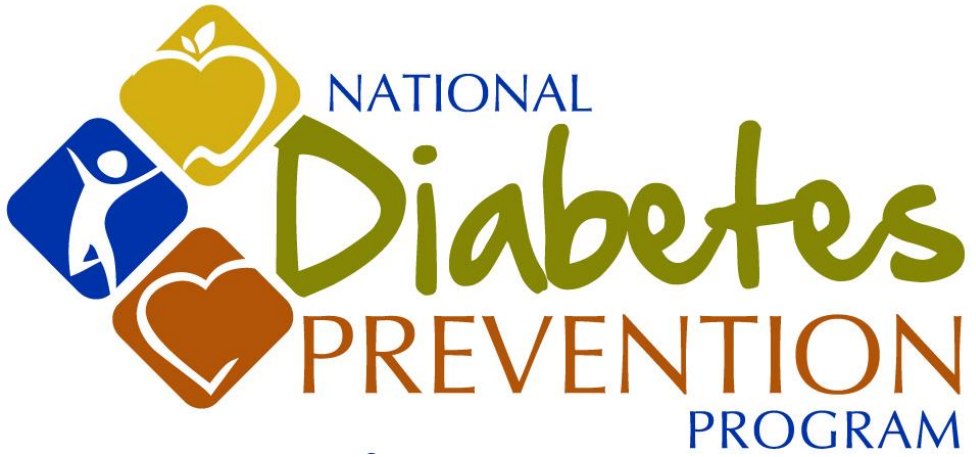


Based on Our Analysis, Our Focus Group Is:

- **Located In:**
  - Lafayette, LA
  - Miami-Ft. Lauderdale, FL
  - Panama City, FL
- **Gender:**
  - Males & Females
- **Ages:**
  - 34-47 for Males, 36-47 for Females
- **BMI Levels:**
  - Over 25- Obese/Overweight
- **Household Incomes Between:**
  - $6k-52k

The CDC's National Diabetes Prevention Program Offers a Promising Solution

- **NDPP Program Overview**
  - Low-cost lifestyle program-> initial goal is to lose 5-7% of weight
  - Often community-based through churches, schools, etc.
  - Target audience of adults who have a BMI >25
- **Characteristics**
  - Marketed through word of mouth & referrals
  - 12-month program, starts out 1x / week
  - Lessons, handouts, lifestyle coach, and more
- **Evidence-Based Outcomes**
  - Participants have reduced risk of diabetes by up to 58%
  - Support groups built on healthy habits



Further Resources to Explore

- **CDC's National Diabetes Prevention Program**
  - https://www.cdc.gov/diabetes/prevention/index.html

- **Community Intervention in Diabetes Care in Low Income Populations by Yvonne Greer, MPH, RD, CD**
  - https://professional.diabetes.org/sites/professional.diabetes.org/files/media/greer-community_interventions.pdf

- **Map of Medicare Diabetes Prevention Program (MDPP) Resources**
  - https://innovation.cms.gov/initiatives/medicare-diabetes-prevention-program/mdpp-map.html?dist=100

Agenda

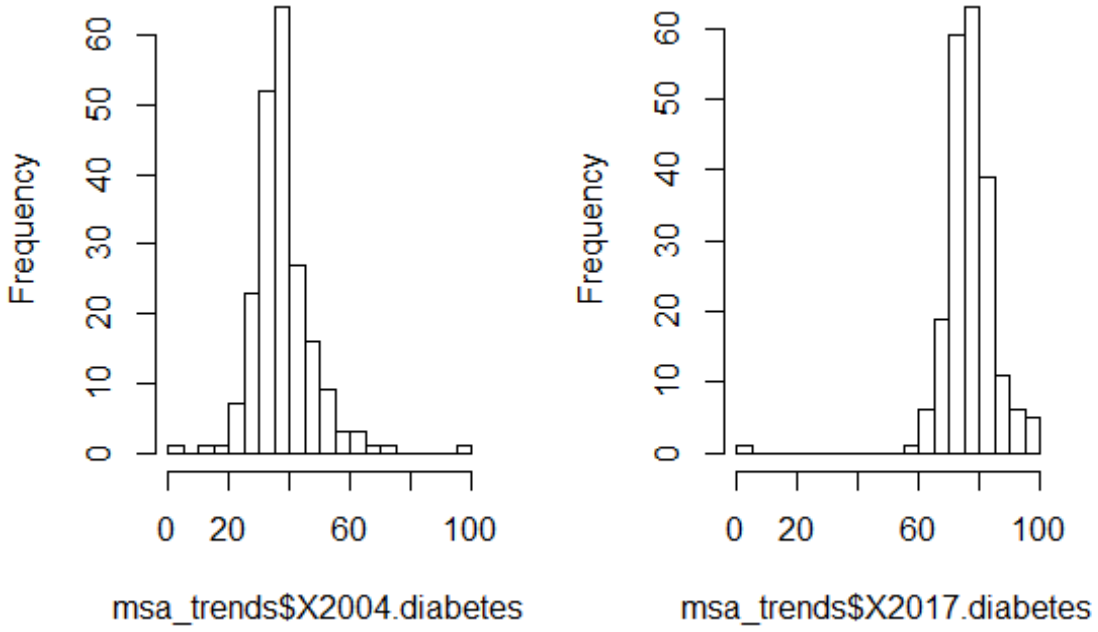
- Problem & Proposed Solution
- MSAs to Focus On
- Finding & Educating High-Risk Patients
- Appendix

Data Source 1: Google Search Interest

- **Data Overview**
  - Compares Google searches vs. the incidence of 9 health conditions for 210 MSAs from 2004-2017

- **Dataset Stats**
  - Scale of 0-100
  - 210 rows x 128 columns
  - Format: .csv file

Google "Diabetes Index"
Searches vs. Incidences
2004 & 2017



Data Source 2: CDC Dataset

- **Data Overview**
  - The dataset (NHANES) is health survey data collected by the CDC's US National Center for Health Statistics (NCHS) from 2009-2012

- **75 variables for 10,000 patients- sampled to represent the US population**

- **Dataset Stats**
  - 10,000 rows x 76 columns
  - Format: R Library "NHANES"

Data Analysis Caveats

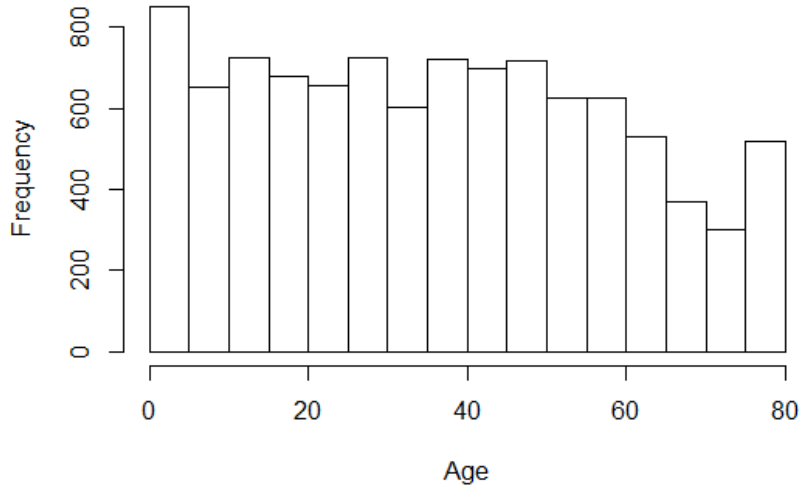
- **Google Dataset**
  - ***Slide 7: MSAs-*** Removed 1 MSA with NA values
    - 209/210 remain (99%+)

- **NHANES Dataset**
  - ***All outputs from NHANES dataset***
    - Removed the 142 samples where Diabetes = NA
    - 9,858/10,000 = 99% remain

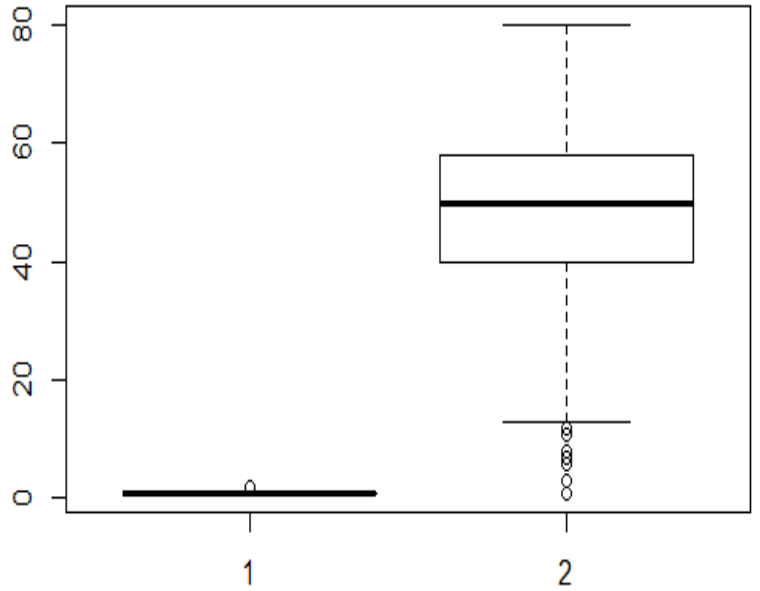
  - ***Slide 11: Age-*** No additional samples omitted
  - ***Slide 12: BMI-*** Removed the additional 229 samples where BMI = NA
    - 9,629/10,000 = 96% remain
  - ***Slide 13: Gender-*** No additional samples omitted
  - ***Slide 13: Income-*** Removed the 795 additional samples where HHIncome = NA
    - 9,063/10,000 = 91% remain

NHANES Initial Exploratory Data Analysis

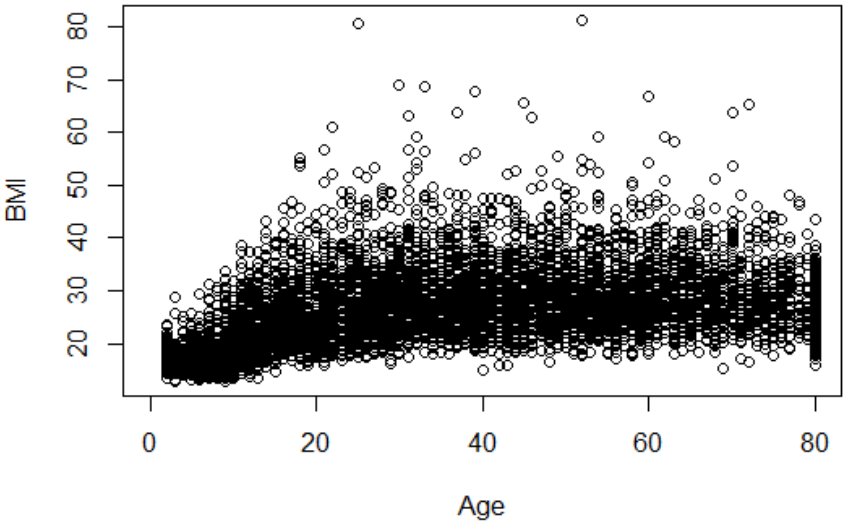
Histogram of Population Age



Age of Diabetes Diagnosis



Population BMI vs. Age



8% of the patients in the sample have diabetes

In-line with national average of 9%

Financial Projections and Assumptions

Exhibit 1: Financial Projections

	Year 0	Year 1	Year 2	Year 3
Number of Transactions		30	35	40
Average Order Size		$399	$411	$423
Revenue		**$11,970**	**$14,178**	**$16,794**
% Growth			*18%*	*18%*
Less: COGS		-$1,197	-$1,418	-$1,679
Less: Other Variable Costs		(599)	(709)	(840)
Total Variable Costs		**-$1,796**	**-$2,127**	**-$2,519**
% Margin		*15%*	*15%*	*15%*
Gross Profit		**$10,175**	**$12,052**	**$14,275**
% Margin		*85%*	*85%*	*85%*
Less: Upfront Costs (Investment)	-$7,250			
Less: Ongoing Maintenance/G&A		(5,000)	(5,000)	(5,000)
Less: Depreciation Expense		-	-	-
Net Income		**$5,175**	**$7,052**	**$9,275**
Less: Taxes		(1,811)	(2,468)	(3,246)
Net Income after Tax		**$3,363**	**$4,584**	**$6,029**
Plus: Depreciation (Non-Cash)		-	-	-
Free Cashflow After Tax	**-$7,250**	**$3,363**	**$4,584**	**$6,029**
% Margin		28%	32%	36%

Business Model

- The anticipated business model is to sell the report to healthcare firms and government entities for a set fee via the web.

Key Assumptions

- Initial Project Costs: $7,500
  - Labor cost of $5,250: 60 hours x $75/hour ($150k annual fully-loaded cost for data scientist in Los Angeles)
  - Additional upfront costs of $2,000
- Reports sold at $399 each
  - Brief market research overviews sell from $100-1,000 per blog.marketresearch.com
- 30 transactions Year 1, 15% volume growth, 3% price increases

Exhibit 2: ROI

	Year 0	Year 1	Year 2	Year 3
Net Cash Flows	-$7,250	$3,363	$4,584	$6,029
Net Present Value	$3,698			
Cumulative Cash Flows	**-$7,250**	**-$3,887**	**$697**	**$6,726**

Discount Rate	12.0%
Tax Rate	35.0%
Internal Rate of Return	**36.9%**

Google Dataset Variables

```
> names(msa_trends)
  [1] "ï..dma"              "geoCode"             "X2004.cancer"          "X2004.cardiovascular"
  [5] "X2004.stroke"        "X2004.depression"    "X2004.rehab"           "X2004.vaccine"
  [9] "X2004.diarrhea"      "X2004.obesity"       "X2004.diabetes"        "X2005.cancer"
 [13] "X2005.cardiovascular" "X2005.stroke"        "X2005.depression"      "X2005.rehab"
 [17] "X2005.vaccine"       "X2005.diarrhea"      "X2005.obesity"         "X2005.diabetes"
 [21] "X2006.cancer"        "X2006.cardiovascular" "X2006.stroke"         "X2006.depression"
 [25] "X2006.rehab"         "X2006.vaccine"       "X2006.diarrhea"        "X2006.obesity"
 [29] "X2006.diabetes"      "X2007.cancer"        "X2007.cardiovascular"  "X2007.stroke"
 [33] "X2007.depression"    "X2007.rehab"         "X2007.vaccine"         "X2007.diarrhea"
 [37] "X2007.obesity"       "X2007.diabetes"      "X2008.cancer"          "X2008.cardiovascular"
 ….
# Repeats through 2017
```

https://www.kaggle.com/shaunmgbray/health-searches-by-us-metropolitan-area-2004-2017

NHANES Dataset Variables

```
> names(NHANES)
[1] "ID" "SurveyYr" "Gender" "Age" "AgeDecade“
[6] "AgeMonths" "Race1" "Race3" "Education" "MaritalStatus“
[11] "HHIncome" "HHIncomeMid" "Poverty" "HomeRooms" "HomeOwn“
[16] "Work" "Weight" "Length" "HeadCirc" "Height“
[21] "BMI" "BMICatUnder20yrs" "BMI_WHO" "Pulse" "BPSysAve“
[26] "BPDiaAve" "BPSys1" "BPDia1" "BPSys2" "BPDia2“
[31] "BPSys3" "BPDia3" "Testosterone" "DirectChol" "TotChol“
[36] "UrineVol1" "UrineFlow1" "UrineVol2" "UrineFlow2" "Diabetes“
[41] "DiabetesAge" "HealthGen" "DaysPhysHlthBad" "DaysMentHlthBad" "LittleInterest“
[46] "Depressed" "nPregnancies" "nBabies" "Age1stBaby" "SleepHrsNight“
[51] "SleepTrouble" "PhysActive" "PhysActiveDays" "TVHrsDay" "CompHrsDay“
[56] "TVHrsDayChild" "CompHrsDayChild" "Alcohol12PlusYr" "AlcoholDay" "AlcoholYear“
[61] "SmokeNow" "Smoke100" "Smoke100n" "SmokeAge" "Marijuana“
[66] "AgeFirstMarij" "RegularMarij" "AgeRegMarij" "HardDrugs" "SexEver“
[71] "SexAge" "SexNumPartnLife" "SexNumPartYear" "SameSex" "SexOrientation“
[76] "PregnantNow"
```

https://cran.r-project.org/web/packages/NHANES/NHANES.pdf