What is the Relationship Between Fast Food and Diabetes?

Group 6
Johanna Pina,
Giulia Tasca,
Rachel Fletcher,

The Business Case: Omar Nasir

We will be presenting on the relationship between the number of fast food restaurants and diabetes cases in the U.S..
My name is Giulia Tasca and my group members presenting this project are Rachel Fletcher, and Johanna Pina, and Omar Nasir will be presenting the Business Case.

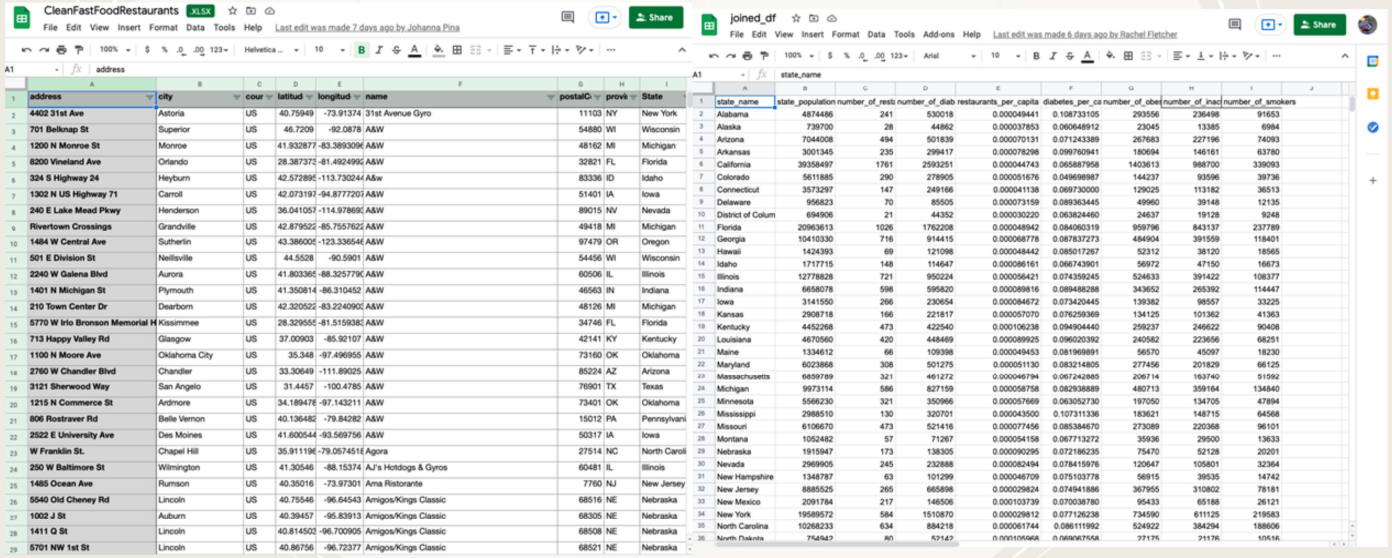
here is a quick layout of how the presentation will go

Introduction and Initial Thoughts

- Exploring the relationship between number of fast food restaurants and the number of Diabetes cases in each state
- Topic selected based on common interests among team members
- Various datasets on Kaggle regarding fast food restaurants

- We considered various topics and wanted to do a project relating to predicting and discovering hidden relationships between prevalent topics in our country; We decided that fast food restaurants and diabetes are two prevalent topics in our country today and decided to investigate the possible relationship.
- We found various fast food restaurant datasets on Kaggle, then retrieved our diabetes data from the CDC website, and our population data from census.gov.

Data Sets



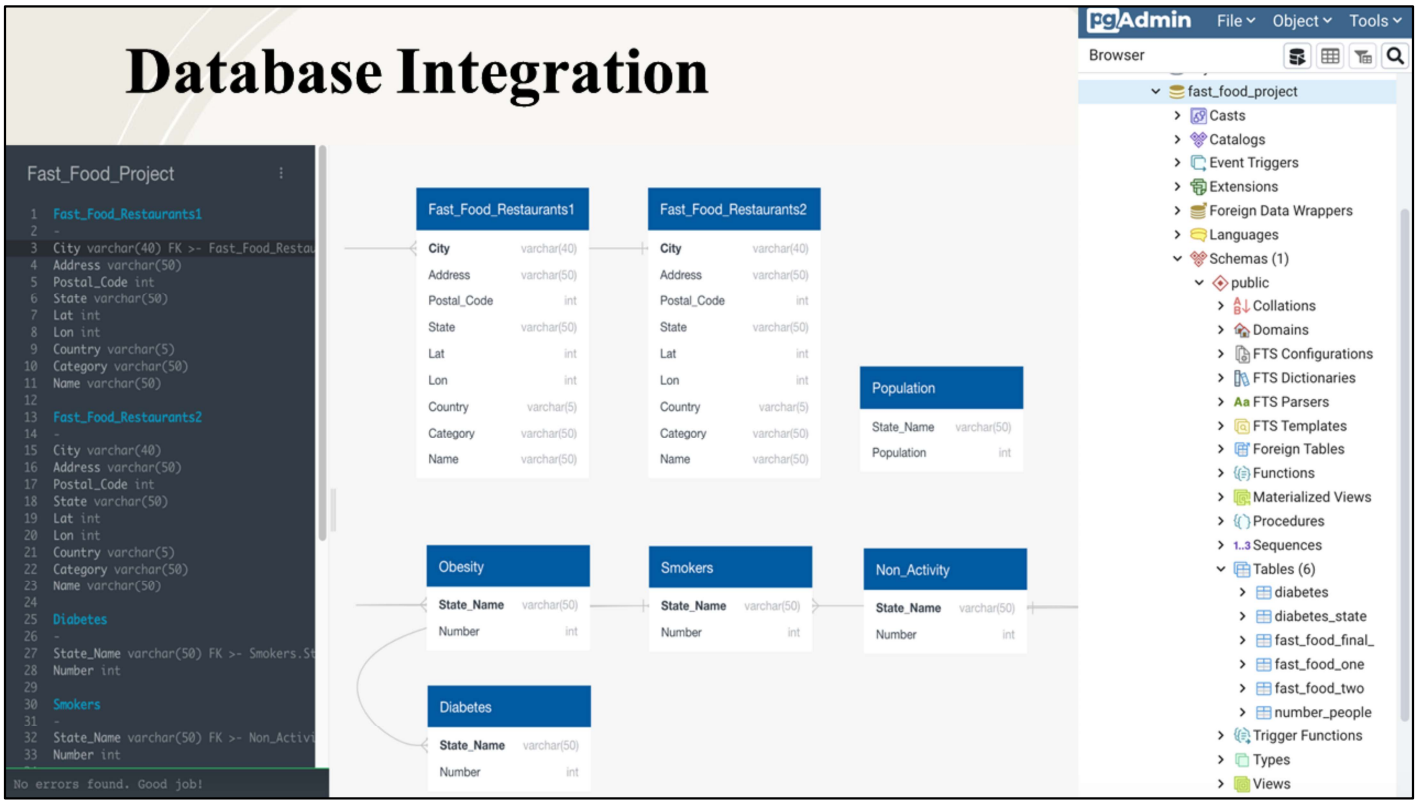
CleanFastFoodRestaurants.XLSX

address	city	cour	latitud	longitud	name	postalC	provi	State
4402 31st Ave	Astoria	US	40.75949	-73.91374	31st Avenue Gyro	11103	NY	New York
701 Belknap St	Superior	US	46.7209	-92.0878	A&W	54880	WI	Wisconsin
1200 N Monroe St	Monroe	US	41.932877	-83.389309	A&W	48162	MI	Michigan
8200 Vineland Ave	Orlando	US	28.387372	-81.492499	A&W	32821	FL	Florida
324 S Highway 24	Heyburn	US	42.572895	-113.730244	A&w	83336	ID	Idaho
1302 N US Highway 71	Carroll	US	42.073197	-94.877720	A&W	51401	IA	Iowa
240 E Lake Mead Pkwy	Henderson	US	36.041057	-114.97869	A&W	89015	NV	Nevada
Rivertown Crossings	Grandville	US	42.879522	-85.755762	A&W	49418	MI	Michigan
1484 W Central Ave	Sutherlin	US	43.386005	-123.33654	A&W	97479	OR	Oregon
501 E Division St	Neillsville	US	44.5528	-90.5901	A&W	54456	WI	Wisconsin
2240 W Galena Blvd	Aurora	US	41.803365	-88.325779	A&W	60506	IL	Illinois
1401 N Michigan St	Plymouth	US	41.350814	-86.310452	A&W	46563	IN	Indiana
210 Town Center Dr	Dearborn	US	42.320522	-83.224090	A&W	48126	MI	Michigan
5770 W Irlo Bronson Memorial H	Kissimmee	US	28.329555	-81.515938	A&W	34746	FL	Florida
713 Happy Valley Rd	Glasgow	US	37.00903	-85.92107	A&W	42141	KY	Kentucky
1100 N Moore Ave	Oklahoma City	US	35.348	-97.496955	A&W	73160	OK	Oklahoma
2760 W Chandler Blvd	Chandler	US	33.30649	-111.89025	A&W	85224	AZ	Arizona
3121 Sherwood Way	San Angelo	US	31.4457	-100.4785	A&W	76901	TX	Texas
1215 N Commerce St	Ardmore	US	34.189478	-97.143211	A&W	73401	OK	Oklahoma
806 Rostraver Rd	Belle Vernon	US	40.136482	-79.84282	A&W	15012	PA	Pennsylvani
2522 E University Ave	Des Moines	US	41.600544	-93.569756	A&W	50317	IA	Iowa
W Franklin St.	Chapel Hill	US	35.911196	-79.057451	Agora	27514	NC	North Caroli
250 W Baltimore St	Wilmington	US	41.30546	-88.15374	AJ's Hotdogs & Gyros	60481	IL	Illinois
1485 Ocean Ave	Rumson	US	40.35016	-73.97301	Ama Ristorante	7760	NJ	New Jersey
5540 Old Cheney Rd	Lincoln	US	40.75546	-96.64543	Amigos/Kings Classic	68516	NE	Nebraska
1002 J St	Auburn	US	40.39457	-95.83913	Amigos/Kings Classic	68305	NE	Nebraska
1411 Q St	Lincoln	US	40.814503	-96.700905	Amigos/Kings Classic	68508	NE	Nebraska
5701 NW 1st St	Lincoln	US	40.86756	-96.72377	Amigos/Kings Classic	68521	NE	Nebraska

joined_df

state_name	state_population	number_of_rests	number_of_diab	restaurants_per_capita	diabetes_per_ca	number_of_obes	number_of_inac	number_of_smokers
Alabama	4874486	241	530018	0.000049441	0.108733105	293556	236498	91653
Alaska	739700	28	44862	0.000037853	0.060648912	23045	13385	6984
Arizona	7044008	494	501839	0.000070131	0.071243389	267683	227196	74093
Arkansas	3001345	235	299417	0.000078298	0.099760941	180694	146161	63780
California	39358497	1761	2593251	0.000044743	0.065887958	1403613	988700	339093
Colorado	5611885	290	278905	0.000051676	0.049698987	144237	93596	39736
Connecticut	3573297	147	249166	0.000041138	0.069730000	129025	113182	36513
Delaware	956823	70	85505	0.000073159	0.089363445	49960	39148	12135
District of Colum	694906	21	44352	0.000030220	0.063824460	24637	19128	9248
Florida	20963613	1026	1762208	0.000048942	0.084060319	959796	843137	237789
Georgia	10410330	716	914415	0.000068778	0.087837273	484904	391559	118401
Hawaii	1424393	69	121098	0.000048442	0.085017267	52312	38120	18565
Idaho	1717715	148	114647	0.000086161	0.066743901	56972	47150	16673
Illinois	12778828	721	950224	0.000056421	0.074359245	524633	391422	108377
Indiana	6658078	598	595820	0.000089816	0.089488288	343652	265392	114447
Iowa	3141550	266	230654	0.000084672	0.073420445	139382	98557	33225
Kansas	2908718	166	221817	0.000057070	0.076259369	134125	101362	41363
Kentucky	4452268	473	422540	0.000106238	0.094904440	259237	246622	90408
Louisiana	4670560	420	448469	0.000089925	0.096020392	240582	223656	68251
Maine	1334612	66	109398	0.000049453	0.081969891	56570	45097	18230
Maryland	6023868	308	501275	0.000051130	0.083214805	277456	201829	66125
Massachusetts	6859789	321	461272	0.000046794	0.067242885	206714	163740	51592
Michigan	9973114	586	827159	0.000058758	0.082938889	480713	359164	134840
Minnesota	5566230	321	350966	0.000057669	0.063052730	197050	134705	47894
Mississippi	2988510	130	320701	0.000043500	0.107311336	183621	148715	64568
Missouri	6106670	473	521416	0.000077456	0.085384670	273089	220368	96101
Montana	1052482	57	71267	0.000054158	0.067713272	35936	29500	13633
Nebraska	1915947	173	138305	0.000090295	0.072186235	75470	52128	20201
Nevada	2969905	245	232888	0.000082494	0.078415976	120647	105801	32364
New Hampshire	1348787	63	101299	0.000046709	0.075103778	56915	39535	14742
New Jersey	8885525	265	665898	0.000029824	0.074941886	367955	310802	78181
New Mexico	2091784	217	146506	0.000103739	0.070038780	95433	65188	26121
New York	19589572	584	1510870	0.000029812	0.077126238	734590	611125	219583
North Carolina	10268233	634	884218	0.000061744	0.086111992	524922	384294	188606
North Dakota	754842	80	52142	0.000105968	0.069067558	27175	21176	10516

- The restaurant data sets were obtained from kaggle.com and was uploaded by Datafiniti's Business Database.
- Additionally, from the CDC we obtained more data sets of other risk factors (i.e. obesity rates, smoking rates and non-activity rates) that may also contribute to number of diabetes cases.
- The last data set was obtained from census.gov to obtain the population from each state to calculate per capita rates.Finally, we did a full outer join on state_name of the joined fast food restaurant table and diabetes risk factor table to create our final joined dataset. We then exported this into pandas for the machine learning work



- We have two sets of restaurant data that were merged using postgress SQL. We joined the two fast food restaurants using an inner join on the city, so that we could have all of the necessary information needed for each city and state.
- We then obtained the four data sets we got from the cdc website, joined those four tables together one at a time with inner joins, to create one giant diabetes table with the different states, numbers, and risk factors
- Finally, we did a full outer join on state_name of the joined fast food restaurant table and diabetes risk factor table to create our final joined dataset. We then exported this into pandas for the machine learning work and onto our shared data source for the visualizations.

Questions We Hoped to Answer

1- Which states have the most and least fast food restaurants?

2- What are the top 10 Fast food restaurants nationally? Which Fast food restaurants have the most locations nationally?

3- What is the number of fast food restaurants per capita by state?

4- Which states have the most cases of diabetes?

5- What is the national number of diabetes cases? Number of diabetes related risk factors (smokers, obesity, non-activity levels)?

6- Does the number of fast food restaurants correlate to the number of diabetes cases by the state?

These are the 6 questions we hoped to answer in our analysis with our machine learning model as well as our visualizations.

1. Bar chart -- Count of Restaurants
2. Pie chart -- Top 10 Restaurants from Sample data set
3. Map chart -- Restaurants per Capita
4. Heat Map -- Heat map of diabetes by State and Year
5. Text blurbs listing # of Diabetes in the US. Risk factors for Type 2 Diabetes.. also Map that shows Diabetes per Capita
6. Answered on machine learning slide

Dashboard

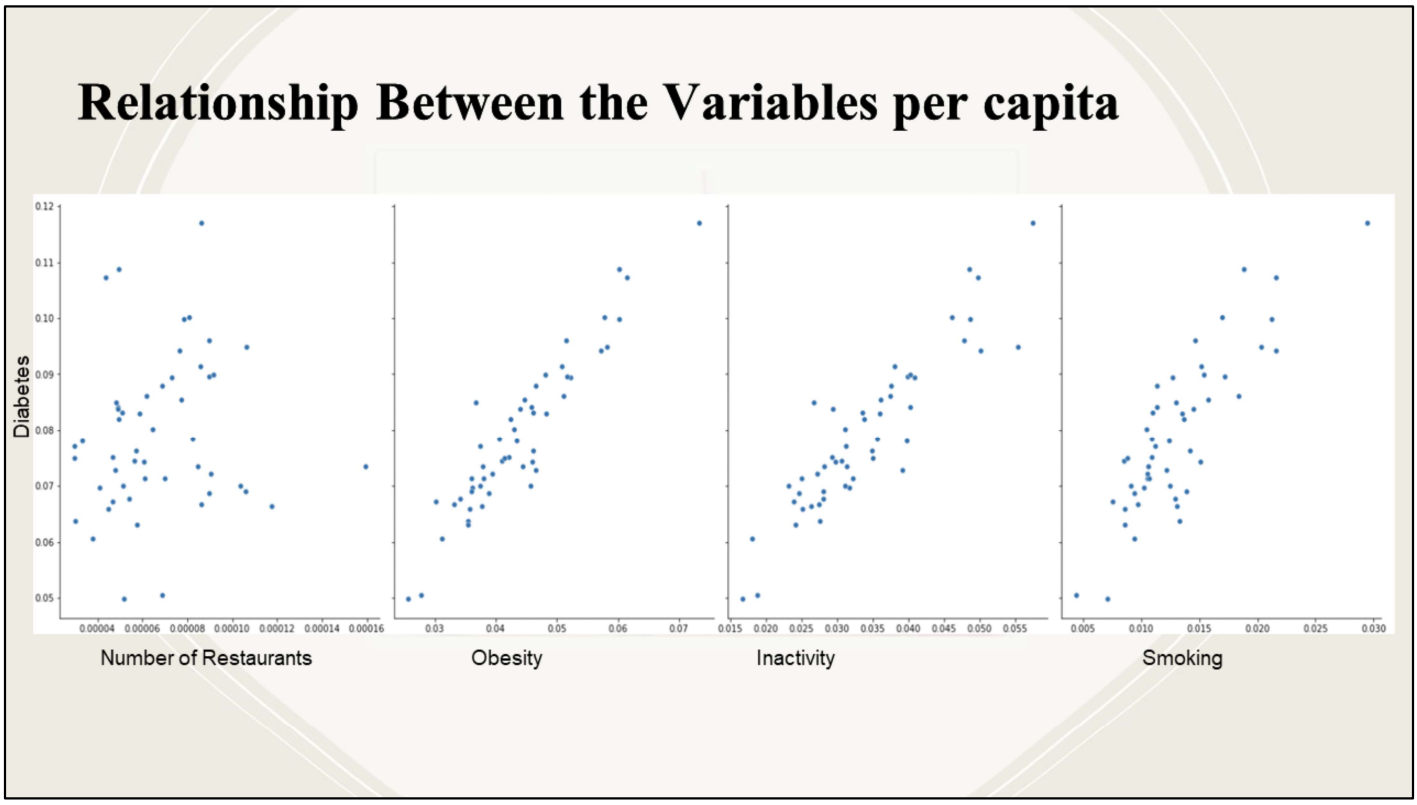
Link to Dashboard

Link to Storyboard

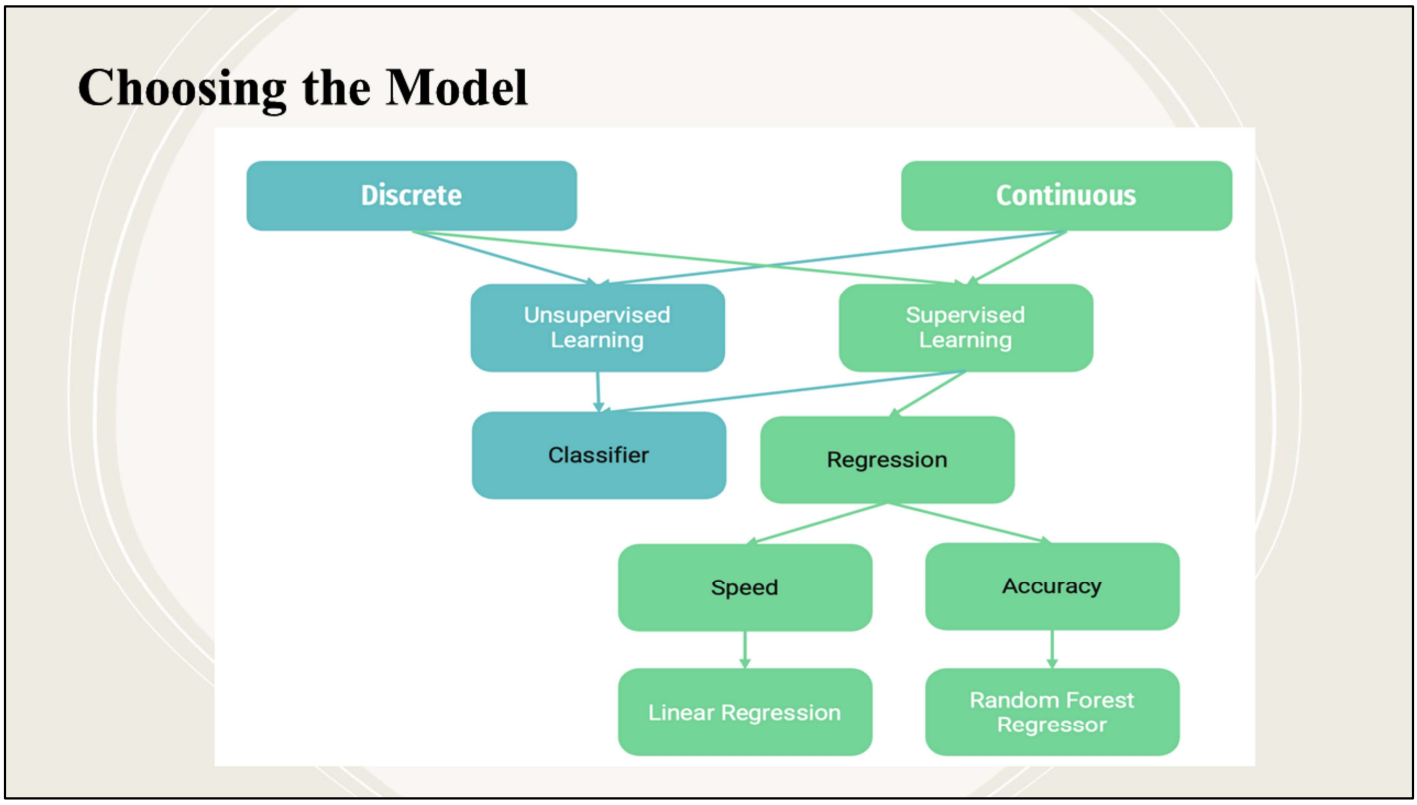
Dashboard consists of:

- # of Diabetes cases in the US
  - The merged dataset that was used to build this map was from the “joined_df” data set.
  - Field used SUMNumberofDiabetesCases was dropped in the Text field under “Marks”
- Risk Factors for Type 2 diabetes
  - The merged dataset that was used to build this map was from the “joined_df” data set.
  - Field used SUMNumberofInactiveAdultsCases was dropped in the Text field under “Marks”
  - Field used SUMNumberofObesityCases was dropped in the Text field under “Marks”
  - Field used SUMNumberofSmokers was dropped in the Text field under “Marks”
- Heat Map of Diabetes by State for 2017 and 2018
  - The dataset that was used to build this heat map was the DiabetesAtlasData dataset.
  - Columns included the Year field, Rows was the State Field, SUM(Number) was dropped in Color and Detail under MARKS
- Top 10 restaurants for sample of dataset
  - The merged dataset that was used to build this map was from the “CleanFastFoodRestaurant_df” data set.

    - Name count field was put in the Color field of Marks, CNT(Name) was put in Angle field of Marks, Max(Name) was put in Text field of Marks, CNT(Name) was put in Text field of Marks.
    - A filter was done to limit the # of restaurants to the Top 10
- Restaurants per capita using sample set
    - The merged dataset that was used to build this map was from the “CleanFastFoodRestaurant_df” data set.
    - Longitude was dropped in the Columns field, Latitude was dropped in the Rows field, SUM(Restaurants Per Capita) was dropped in the Color field under Marks, and State Name was dropped in the Detail field under Marks
- Diabetes cases per capita
    - The merged dataset that was used to build this map was from the “joined_df” data set.
    - Longitude was dropped in the Columns field, Latitude was dropped in the Rows field, SUM(Diabetes Per Capita) was dropped in the Color field under Marks, and State Name was dropped in the Detail field under Marks
- Count of fast food restaurants by state using a sample from a data set
    - The merged dataset that was used to build this map was from the “CleanFastFoodRestaurant_df” data set.
    - CNT(Name) was dropped in Columns field, State was dropped in Rows field



In this visualization we plotted the y and the various x features. Here we can see that Obesity, inactivity and smoking appear to have a strong positive correlation with diabetes. Whereas, number of restaurants per capita appears more scattered. At this point we could drop number of restaurant as a irrelevant feature (but we decided to still run the model, since this is what we wanted to evaluate).



Continuous: Our predictions would result in continuous data, meaning the data itself would not be classified/discrete. Instead the predictions would be similar values to what we have (which can be any number/value).
Supervised: The machine will not need to independently find which groups or clusters to put the values. The data is already separated by its labels.
Regression: What is the value of y when X is given
Speed: Linear Regression is a simple model to decide if there is a correlation between the x & y variables. This model performs quickly with large data sets.
Accuracy: Random Forest Regressor splits the data into trees, each tree is formed from random rows. Each tree then provides its own prediction and the final result is the average of all of the predictions, making it more accurate by reducing overfitting (models the training data too well).

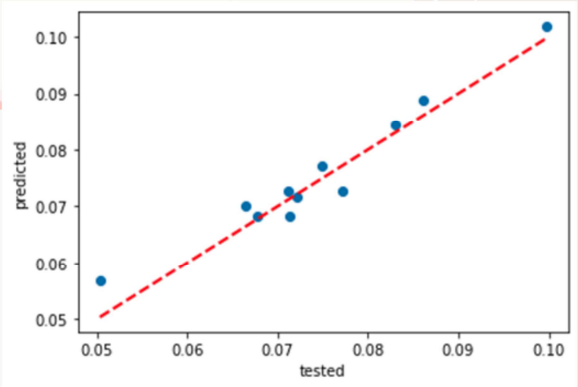
Multiple Linear Regression Model

```
#define the variables
features=['restaurants_per_capita', 'obesity_per_capita',
          'inactivity_per_capita', 'smokers_per_capita']
X = final_df[features].values.reshape(-1, len(features))
y = final_df['diabetes_per_capita'].values

# split data using 80/20 ratio
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1, test_size=0.2)

# instantiate and fit
model = LinearRegression()
model.fit(X_train, y_train)

# predict
y_pred = model.predict(X_test)
```



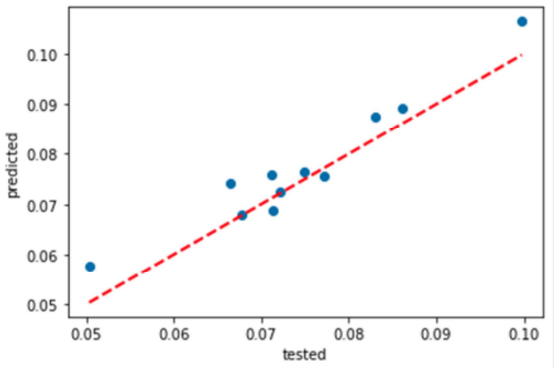

```
Linear Regression mean square error is: 9.501711071642502e-06
Linear Regression R squared is: 0.8891662637660523
Linear Regression pearson coefficient is: (0.9709414518591231, 6.82305350681304e-07)
```

- Linear regression model is desirable for its speed with large data sets.
- The mean square error: is the variance around the fitted regression line. The lower the number, the smaller the “errors/ variance”.
- R squared: measures the strength of the relationship between the model and the rates of diabetes. Generally, the higher the better, but if it is too close to 100%, then there could be overfitting.
- P-value: The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. The p-value is the probability that the null hypothesis is true.
- Pearson Correlation Coefficient: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

Random Forest Regressor Model

```
# instantiate and fit
rf_model = RandomForestRegressor(n_estimators=128)
rf_model = rf_model.fit(X_train, y_train)

# Making predictions using the testing data
predictions = rf_model.predict(X_test)
```




```
Random Forest mean square error is: 1.5383019058075547e-05
Random Forest R squared is: 0.9600584924033808
Random Forest pearson coefficient is: (0.971580926292777, 6.178638237088701e-07)
```

- The RFR model shows a high positive correlation between the y and the X features.
- The model accurately predicts y when X is given. The model is more accurate than the linear regression model.

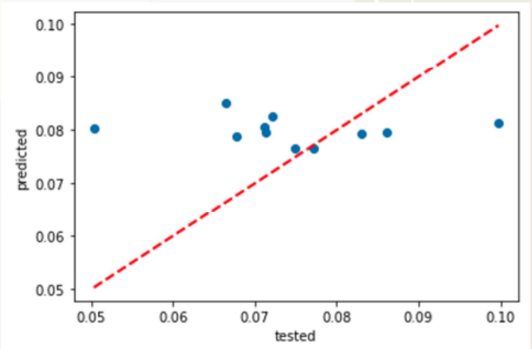
Linear Regression Model with Re-defined Variables

```
X2 = final_df['restaurants_per_capita'].values.reshape(-1, 1)
y2 = final_df['diabetes_per_capita'].values

# split data using 80/20 ratio
X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2, random_state=1, test_size=0.2)

# instantiate and fit
model = LinearRegression()
model.fit(X2_train, y2_train)

# predict
y2_pred = model.predict(X2_test)
```




```
Linear Regression mean square error is: 0.00018518260603333042
Linear Regression R squared is: 0.007241588114186404
Linear Regression pearson coefficient is: (-0.11802692440597003, 0.7296281036952373)
```

- Here we run the linear regression machine learning model again but this time we only give it the values of restaurants per capita as X. Y remains the same= diabetes per capita.
- There appears to be no correlation between the variables.

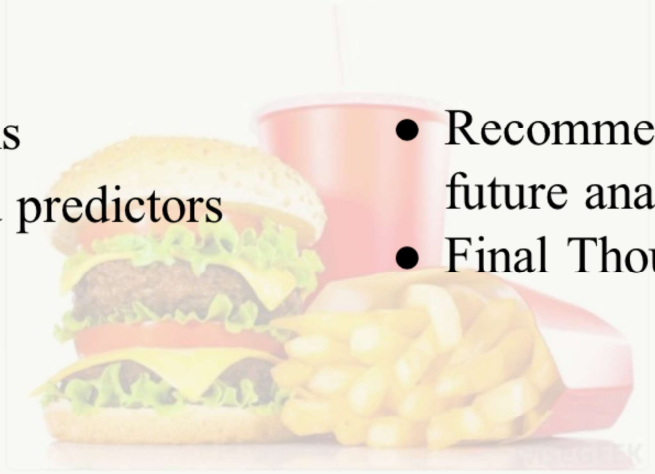
Result of Analysis

- The scatterplots show a strong positive correlation between obesity, smoking and inactivity with diabetes
- We found an extremely weak (~0.0) correlation between number of fast food restaurants and diabetes from our sample
- The models predict the number of diabetes cases when the relevant features/predictor values are given
- The model is able to calculate the needed feature values for preferred diabetes outcomes

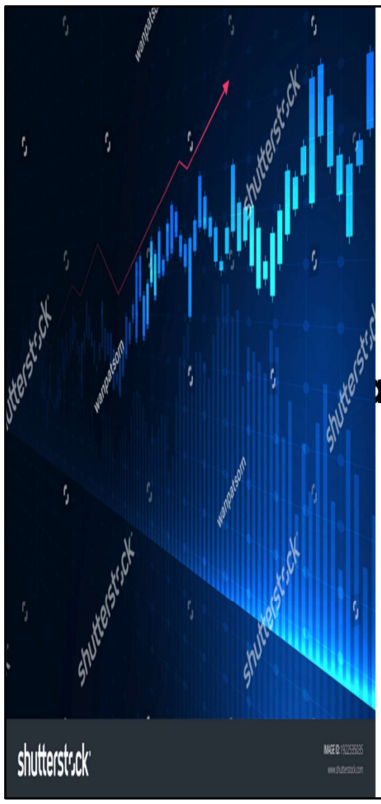
- There is a strong positive correlation between diabetes and obesity, smoking and inactivity.
- No correlation was found between diabetes and number of restaurants
- One possible use case for this model includes predicting the number of diabetes cases for a given population when the values of obesity, smoking and inactivity are provided.
- Another use case could be to lower the number of diabetes to a predetermined level, by lowering the cases of obesity, smoking, and inactivity to a matching level.

Changes During the Project and Reflection

- Questions
- Data and predictors
- Models
- Recommendations for future analysis
- Final Thoughts



- What we would have done differently is pretty much our recommendations so—Given more time and resources for our analysis, we would recommend further research into predictors of diabetes and include those variables in our data frames and machine learning models. Obviously there are various factors that contribute to diabetes in the U.S., like location, genes, health, physical habits, and more. Including as many X variables as possible will allow you to create a sufficient model that can accurately predict diabetes and portray the strength of each correlation between the variables. In addition, we would recommend finding more data sets, ones that could more accurately represent the population of our country, as well as the datasets throughout different years. We used a sample data which is good for an initial analysis but in order to strengthen our argument, more data is needed. A final recommendation would be to broaden our search and compare with restaurants with healthy options, look at the areas with healthier restaurants and see their diabetes cases.
- Although we weren't able to find a correlation between the number of fast food restaurants and number of diabetes cases, the three of us had a successful project.
- We learned how to work as a team and contribute to an analysis where factors were constantly changing but managed to focus on the questions we hoped to answer; The number of fast food restaurants and diabetes cases in our country continue to grow, and we think it is important to dive into the underlying relationship these topics have with one another. Now Omar will briefly present on his project: The Business Case.



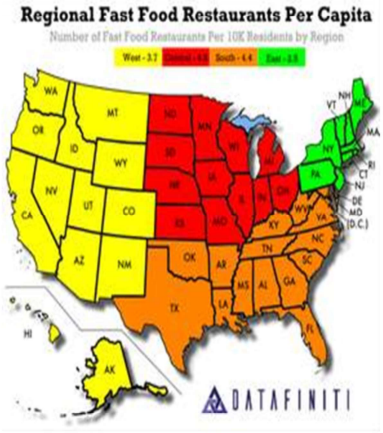
ast Food Industry : The Business Case

By

Omar J. Nasir, MBA

The Growth and Investment Opportunity

- Number of Fast Food Restaurants=
- 196,000
- Income=$200Billion +
- Employes=4.5 Million
- Growth: 2200 per month
- Franchise Costs= $100,000- $ 1 Million
- Franchise Owner Makes= $80.000+
- Worker: $15/hr. so dissatisfaction is high
- MBA Salaries = $100,000 +

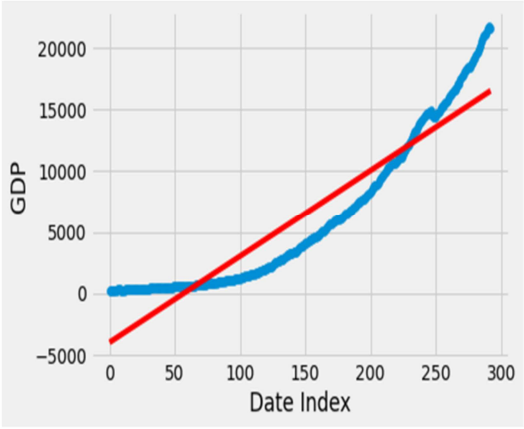


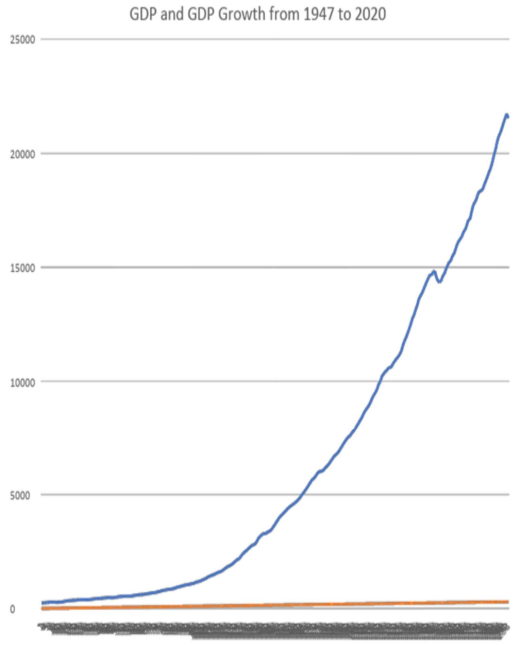
The Growth Rate of Fast Food Industry

- The growth of the Fast Food industry is predicated on:
- Affordability
- Service Time-Drive Ins
- Income Levels patronizing Fast Food Industry:
- $50,000 to $199,000- Household Income
- $12.50 per hour to $50 per hour
- **This is the American Middle Class**
- As more people join the middle class, the growth of the industry becomes certain. Therefore. we study **GDP and Income Distribution through Machine Learning and Mathematical Models**

Predicting GDP Growth

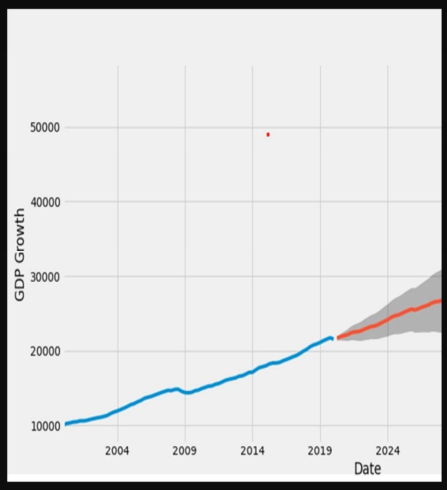
- Supervised Machine Learning
- The Linear Regression Model is checked against existing data.
- RED LINE: Regression Model
- Blue Line: Actual Results



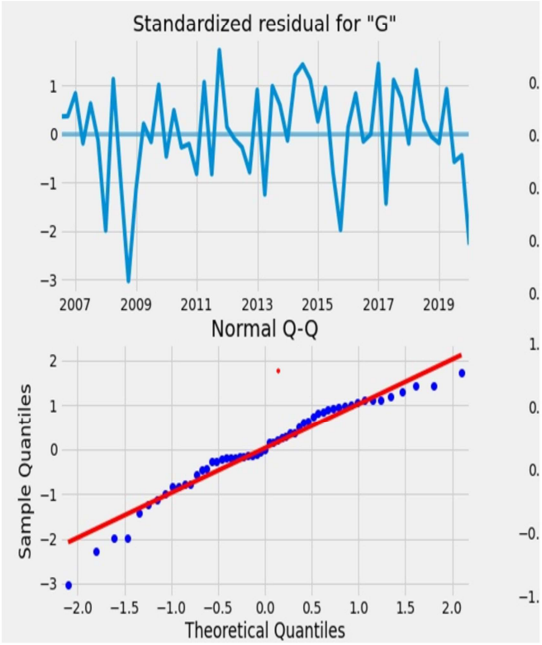


Machine Learning- Python Notebook

Machine Learning Model Prediction



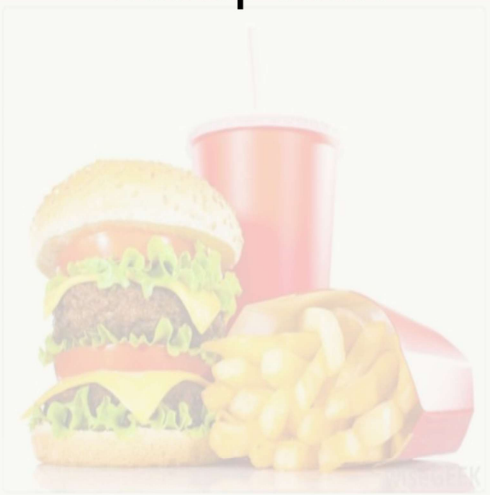
Standardized Residuals and Possibility Plot

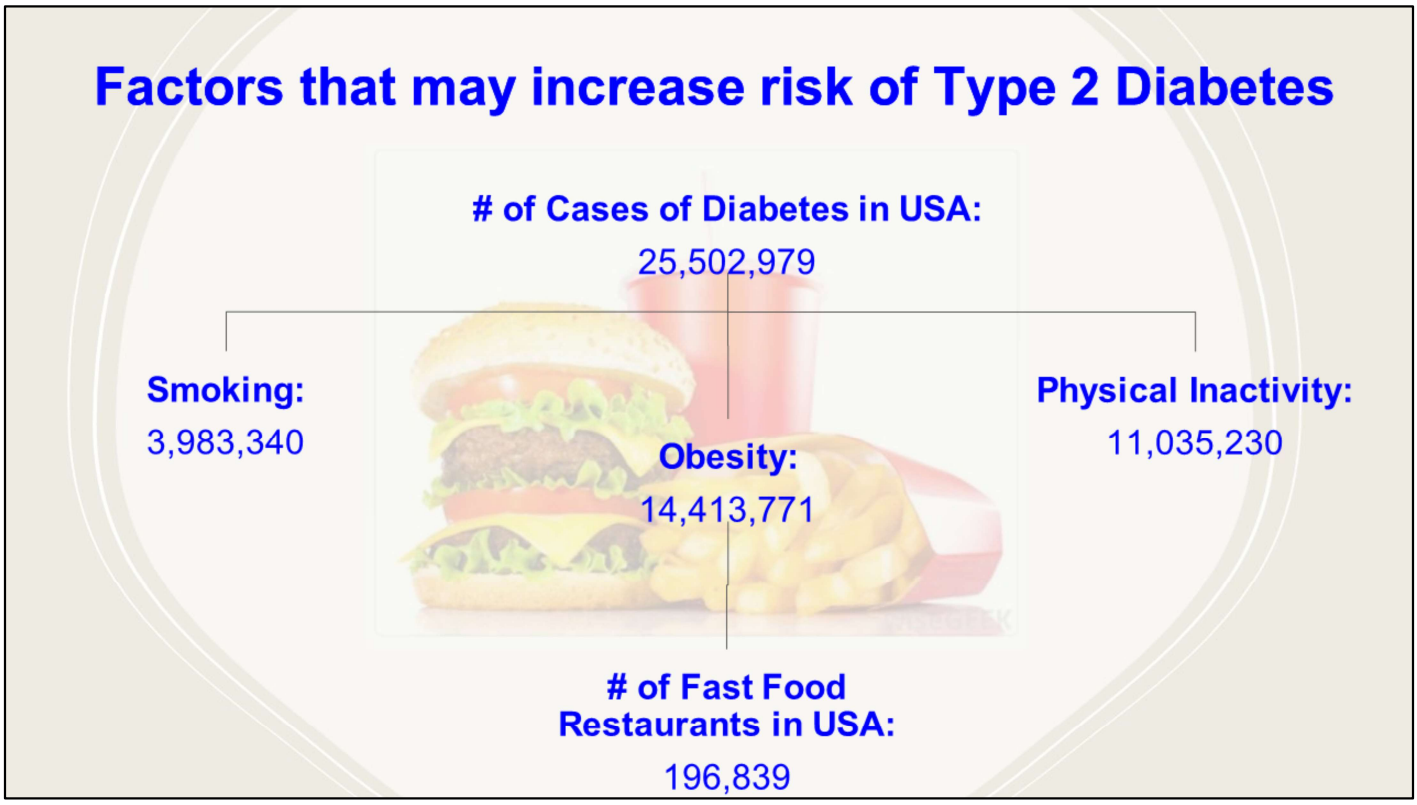


Thank You!

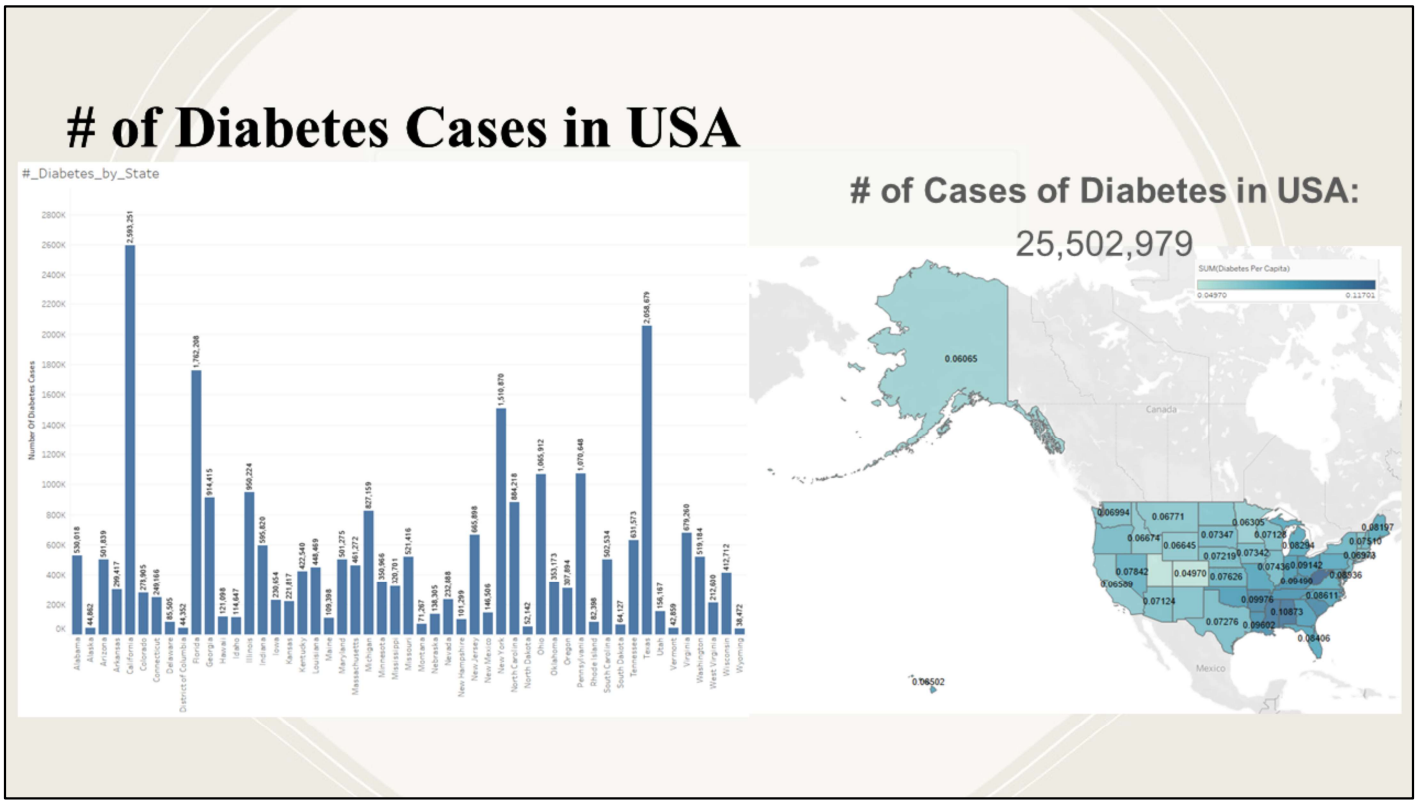
Q & A

Backup Slides

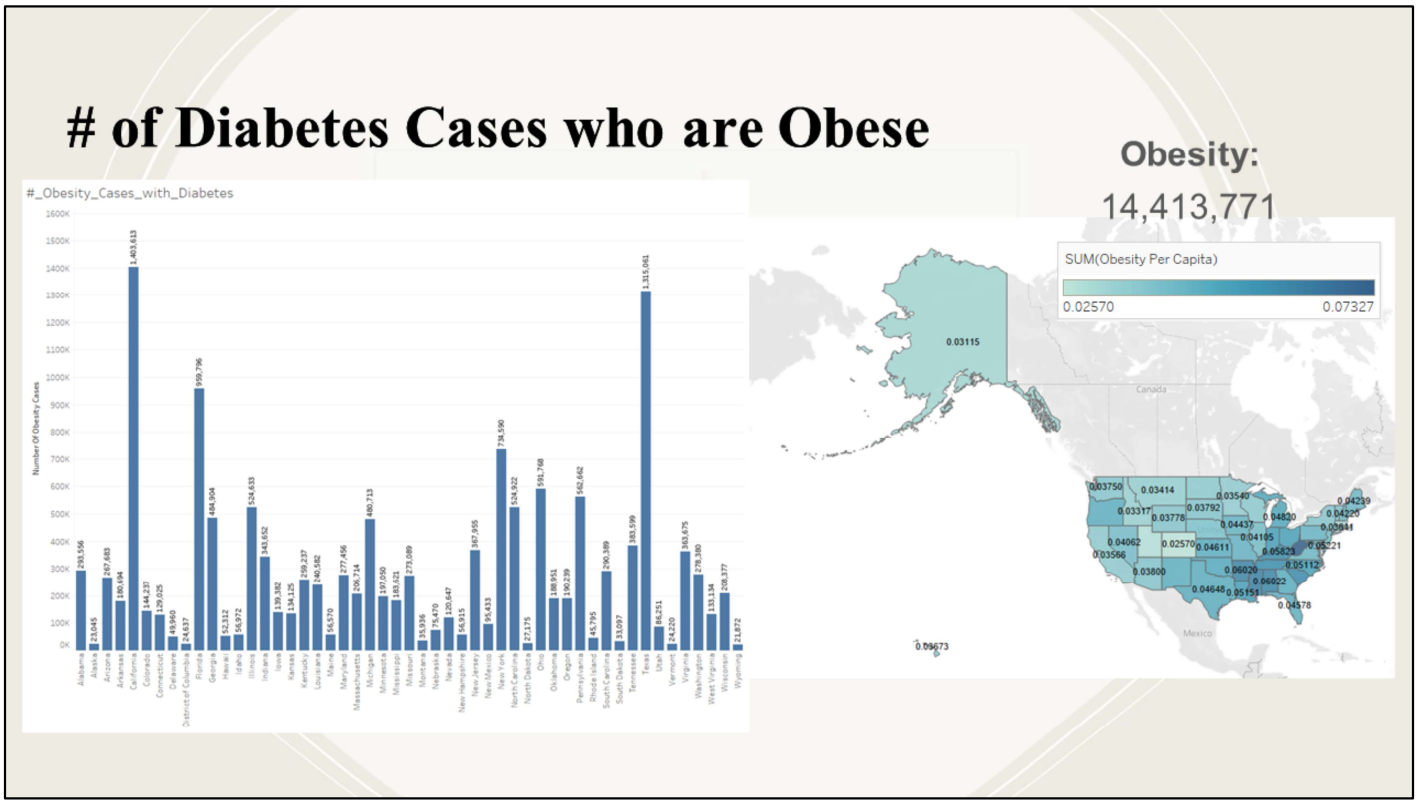




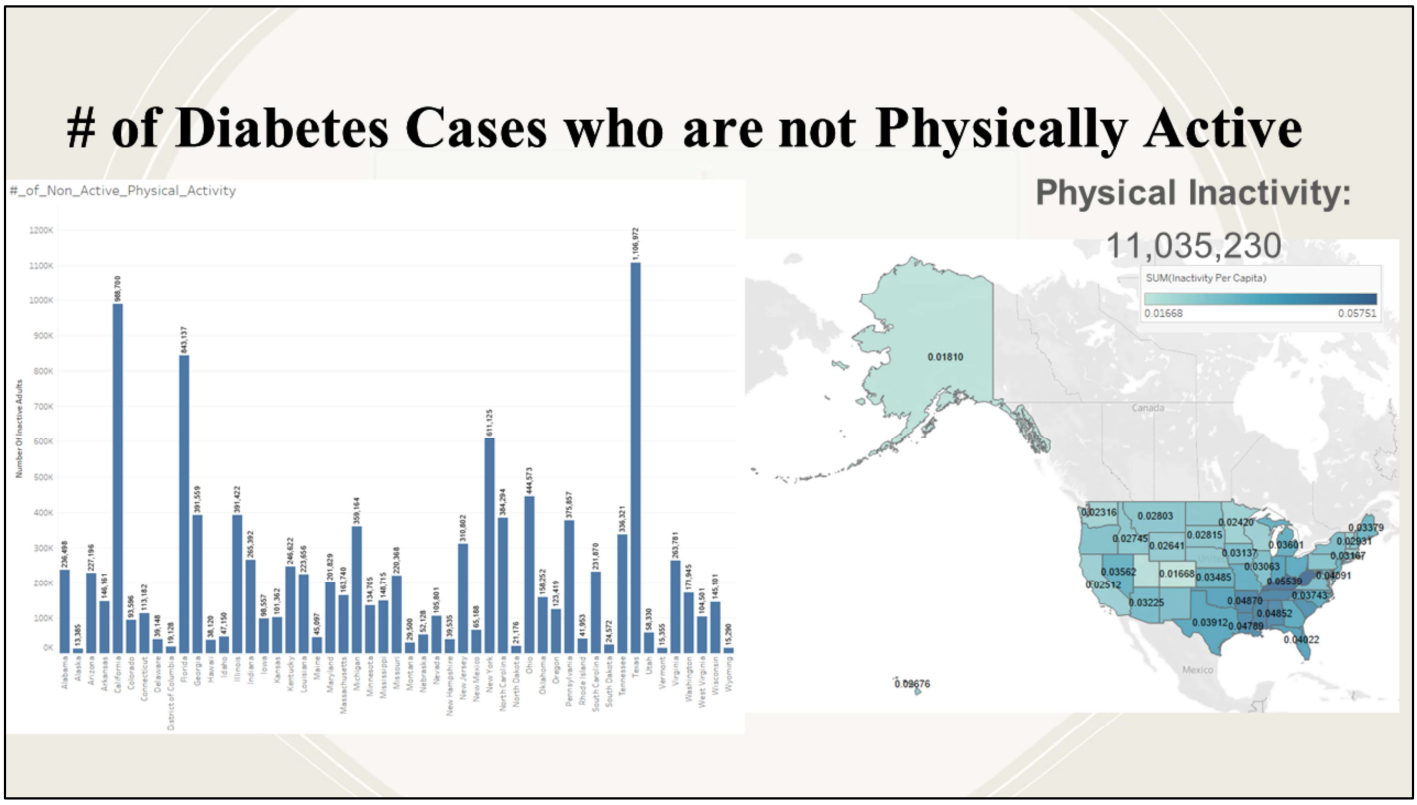
Fast Food has a relation to only Type 2 Diabetes because Fast Food is associated with obesity if eaten in large quantities and often. If a person has particular bad eating habits by eating extremely unhealthy food, which causes obesity, doesn't exercise, and smokes will have a increased danger of developing type 2 diabetes.



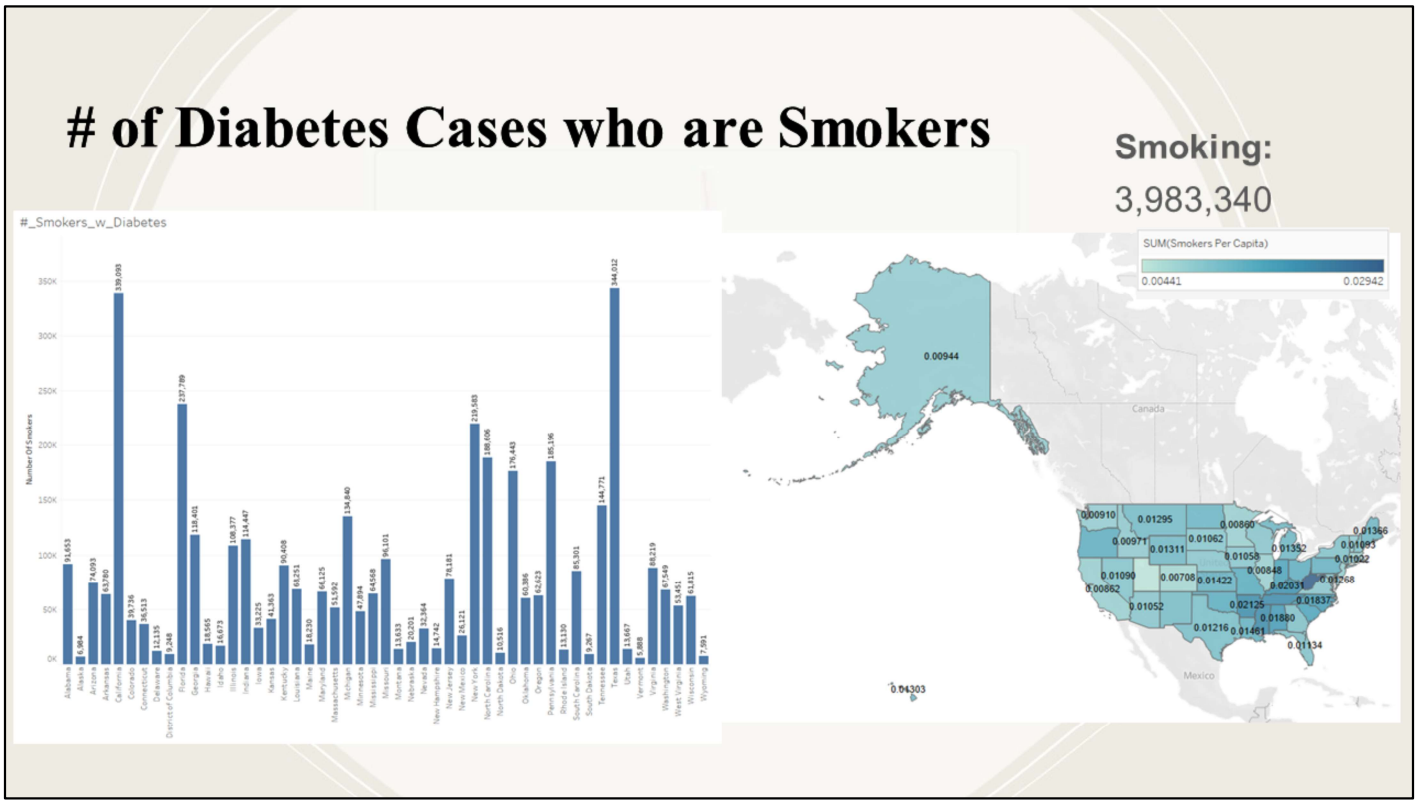
- State with Minimum cases: Wyoming 38,472
- State with minimum cases per capita: Colorado 0.04970
- State with maximum cases: California 2,593,251
- State with maximum cases per capita: West Virginia 0.11701
- Data set was pulled from CDC website: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html
- Data was merged creating a csv file that included data that was pulled regarding diabetes cases in USA, how many cases are obese, how many cases are smokers, and how many cases do not do physical activity.
- The merged dataset that was used to build this map was from the “joined_df” data set.
- Longitude was dropped in the Columns field, Latitude was dropped in the Rows field, SUM(Number of Diabetes Cases) was dropped in the Color field under Marks, and State Name was dropped in the detail field under Marks



- State with Minimum cases: Wyoming 21,872
- State with minimum cases per capita: Colorado 0.02570
- State with maximum cases: California 1,403,613
- State with maximum cases per capita: West Virginia 0.07327
- Data set was pulled from CDC website: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html
- Data was merged creating a cdv file that included data that was pulled regarding diabetes cases in USA, how many cases are obese, how many cases are smokers, and how many cases do not do physical activity.
- The merged dataset that was used to build this map was from the “joined_df” data set.
- Longitude was dropped in the Columns field, Latitude was dropped in the Rows field, SUM(Number of Obesity Cases) was dropped in the Color field under Marks, and State Name was dropped in the detail field under Marks



- State with Minimum cases: Alaska 13,385
- State with minimum cases per capita: Colorado 0.01668
- State with maximum cases: Texas 1,106,972
- State with maximum cases per capita: West Virginia 0.05751
- Data set was pulled from CDC website: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html
- Data was merged creating a cdv file that included data that was pulled regarding diabetes cases in USA, how many cases are obese, how many cases are smokers, and how many cases do not do physical activity.
- The merged dataset that was used to build this map was from the “joined_df” data set.
- Longitude was dropped in the Columns field, Latitude was dropped in the Rows field, SUM(Number of Inactive Adults) was dropped in the Color field under Marks, and State Name was dropped in the detail field under Marks



- State with Minimum cases: Vermont 5,888
- State with minimum cases per capita: Utah 0.00441
- State with maximum cases: Texas 344,012
- State with maximum cases per capita: West Virginia 0.02942
- Data set was pulled from CDC website: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html
- Data was merged creating a cdv file that included data that was pulled regarding diabetes cases in USA, how many cases are obese, how many cases are smokers, and how many cases do not do physical activity.
- The merged dataset that was used to build this map was from the “joined_df” data set.
- Longitude was dropped in the Columns field, Latitude was dropped in the Rows field, SUM(Number of Smokers) was dropped in the Color field under Marks, and State Name was dropped in the detail field under Marks