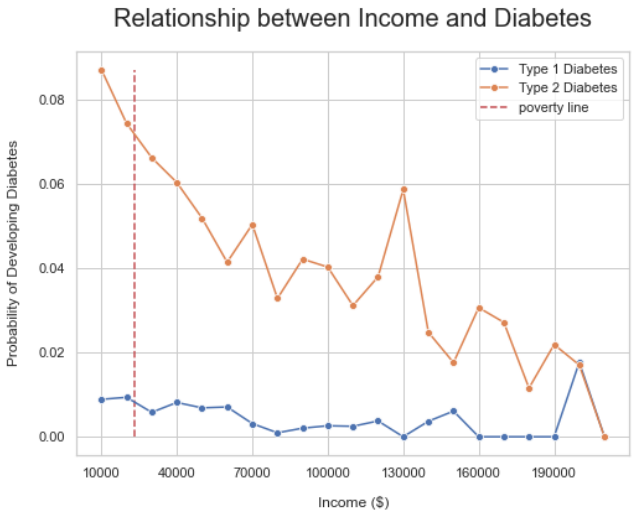
Dollar-betes

A statistical analysis of socio-economic factors and diabetes prevalence in the United States

West Coast Regional Datathon Report (Team 11)
Adamya Prakash Srivastava, Samarth Goel, Ming Fong, Max Vaysburd
UC Berkeley
9 April 2022

I. Executive Summary

Our team explored the effects of wealth and social inequality on diabetes rates in the United States, looking into household income, education level, and race, amongst many other factors. We found a statistically significant inverse relationship between the variables we explored and the probability of having diabetes.

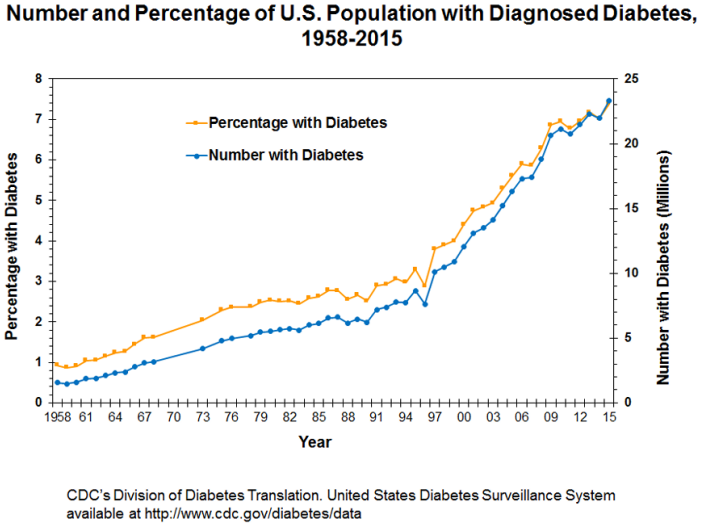


We made linear models to check whether diabetes had correlations (or effects, if our controls were deemed strong enough) on family income and individual employment outcomes. Our findings broadly include that having diabetes negatively correlates with those factors, *except* for type 1 diabetes on family income, perhaps due to the tendency of early diagnoses of the condition, helping prevent the loss of future incomes.

Delving deeper, we unearthed that this relationship arises from type II diabetes specifically, while the relationship with type I diabetes is not statistically significant, as shown in Figure 1. This suggests that the source of the correlation is very much related to the different lifestyles of high income and low income families, since type I diabetes is genetic while type II diabetes typically arises due to personal health decisions and environmental factors. We found that in the United States, a person with a family income below $25,000 is twice as likely to have type II diabetes than someone with a family income above $100,000.

Those who make lower incomes have access to worse healthcare and are more restricted in their diets with a tendency towards unhealthier foods, both of which can be correlated with diabetes. For example, processed food tends to be cheaper and more accessible to poorer families, but these foods are also higher in added sugars.

With individuals near or below the poverty line at a much higher risk of diabetes, increasing wealth inequality in the United States means



that more individuals will be forced into these unhealthy lifestyles. As a result, the proportion of the United States population with diagnosed diabetes will keep on increasing as the economic problems facing our nation such as record-breaking inflation only get worse in the future. Past data has shown us that even with normal levels of inflation, stagnant wages have meant that the number of people with diabetes in the population has only increased with time (Figure 2), leading experts to label diabetes as a true endemic.

II. Technical Exposition

Dataset and Data Processing

Our team's interest was in the socio-economic factors affecting diabetes rates, and the standard dataset provided did not contain the information we were looking for. Instead, we turned to the US National Health Interview Survey, which is a census-like representative random survey of American households that records demographic and health information. This dataset gave us income, employment, education, demographic, and diabetes data at an anonymized individual-level (n = 31949).

In our analysis, we removed all respondents who refused to respond to the diabetes question (YES/NO answer). Since the survey randomly sampled the US population, we are confident that the data is representative.

We also utilized the CDC Diabetes Diagnosis dataset to look at long-term patterns in diabetes rates in America. This dataset contained year-by-year data on diabetes rates in the United States.

Linear Modeling

We explored two main outcomes for labor: (log) family income and recent employment. Recent employment is a yes/no question to whether the survey answerer worked last week and family income is rounded to the nearest $10,000 and capped at $220,000 and includes income from all sources (work, SSI, etc.). The questions our linear models tested were *whether having diabetes (as well as its different types) has a statistically significant correlation on income as well as employment*, after controlling for a variety of factors.

We controlled for years of education, age, age^2 (this term generally has a negative correlation with income in economic models due to ageism-related employment effects), marital status, BMI, BMI^2, race (white or nonwhite), and gender. These variables served as covariates in our models, as shown below. We could have obtained more data to control for more factors (like additional races), but we found these controls strong enough and prevalent in economic literature already. The terms we kept an eye on were the $\boldsymbol{\delta_1, \ldots, \delta_6}$ coefficients on all diabetes variables.

The null and alternative hypotheses for each of these models are that diabetes is uncorrelated with income/employment (null) and that having diabetes is negatively correlated with income/employment (alternative). The models' specifications are summarized below:

$$H_0: \delta_j = 0$$
$$H_1: \delta_j < 0$$

(1) $\boldsymbol{ln(Salary)_i} = const_i + \boldsymbol{\delta_1}\ \mathbf{Diabetes_i} + \beta_1 \text{Age}_i + \beta_2 \text{Age}^2_i + \beta_3 \text{Married}_i + \beta_4 \text{BMI}_i + \beta_5 \text{BMI}_i^2 + \beta_6 \text{Minority}_i + \beta_7 \text{Female}_i + \beta_8 \text{No High School}_i + \beta_9 \text{High School}_i + \beta_{10} \text{Some College}_i + \beta_{11} \text{Bachelors}_i + \beta_{12} \text{Masters}_i + \varepsilon_i$

(2) $\boldsymbol{ln(Salary)_i} = const_i + \boldsymbol{\delta_2}\ \mathbf{Type\ 1\ Diabetes_i} + \beta_1 \text{Age}_i + \beta_2 \text{Age}^2_i + \beta_3 \text{Married}_i + \beta_4 \text{BMI}_i + \beta_5 \text{BMI}_i^2 + \beta_6 \text{Minority}_i + \beta_7 \text{Female}_i + \beta_8 \text{No High School}_i + \beta_9 \text{High School}_i + \beta_{10} \text{Some College}_i + \beta_{11} \text{Bachelors}_i + \beta_{12} \text{Masters}_i + \varepsilon_i$

(3) $\boldsymbol{ln(Salary)_i} = const_i + \boldsymbol{\delta_3}\ \mathbf{Type\ 2\ Diabetes_i} + \beta_1 \text{Age}_i + \beta_2 \text{Age}^2_i + \beta_3 \text{Married}_i + \beta_4 \text{BMI}_i + \beta_5 \text{BMI}_i^2 + \beta_6 \text{Minority}_i + \beta_7 \text{Female}_i + \beta_8 \text{No High School}_i + \beta_9 \text{High School}_i + \beta_{10} \text{Some College}_i + \beta_{11} \text{Bachelors}_i + \beta_{12} \text{Masters}_i + \varepsilon_i$

(4) $\mathbf{Worked\ last\ week_i} = const_i + \boldsymbol{\delta_4}\ \mathbf{Diabetes_i} + \beta_1 \text{Age}_i + \beta_2 \text{Age}^2_i + \beta_3 \text{Married}_i + \beta_4 \text{BMI}_i + \beta_5 \text{BMI}_i^2 + \beta_6 \text{Minority}_i + \beta_7 \text{Female}_i + \beta_8 \text{No High School}_i + \beta_9 \text{High School}_i + \beta_{10} \text{Some College}_i + \beta_{11} \text{Bachelors}_i + \beta_{12} \text{Masters}_i + \varepsilon_i$

(5) $\mathbf{Worked\ last\ week_i} = const_i + \boldsymbol{\delta_5}\ \mathbf{Type\ 1\ Diabetes_i} + \beta_1 \text{Age}_i + \beta_2 \text{Age}^2_i + \beta_3 \text{Married}_i + \beta_4 \text{BMI}_i + \beta_5 \text{BMI}_i^2 + \beta_6 \text{Minority}_i + \beta_7 \text{Female}_i + \beta_8 \text{No High School}_i + \beta_9 \text{High School}_i + \beta_{10} \text{Some College}_i + \beta_{11} \text{Bachelors}_i + \beta_{12} \text{Masters}_i + \varepsilon_i$

(6) $\mathbf{Worked\ last\ week_i} = const_i + \boldsymbol{\delta_6}\ \mathbf{Type\ 2\ Diabetes_i} + \beta_1 \text{Age}_i + \beta_2 \text{Age}^2_i + \beta_3 \text{Married}_i + \beta_4 \text{BMI}_i + \beta_5 \text{BMI}_i^2 + \beta_6 \text{Minority}_i + \beta_7 \text{Female}_i + \beta_8 \text{No High School}_i + \beta_9 \text{High School}_i + \beta_{10} \text{Some College}_i + \beta_{11} \text{Bachelors}_i + \beta_{12} \text{Masters}_i + \varepsilon_i$

In this section we will justify why we chose the control variables in these models we did. Education plays a large explanatory role in diabetes, as shown in Figure 4 (US 2019 NHIS Diabetes vs Education levels) and Figure 5 (US diabetes levels over time for differing levels of education). BMI is an excellent predictor of life expectancy and diabetes, so it as well as its square were added as controls. The same applies for age, and the combination of age, age squared, and education tend to predict income well in economic models. Race and marital status were included to control for other demographic characteristics.

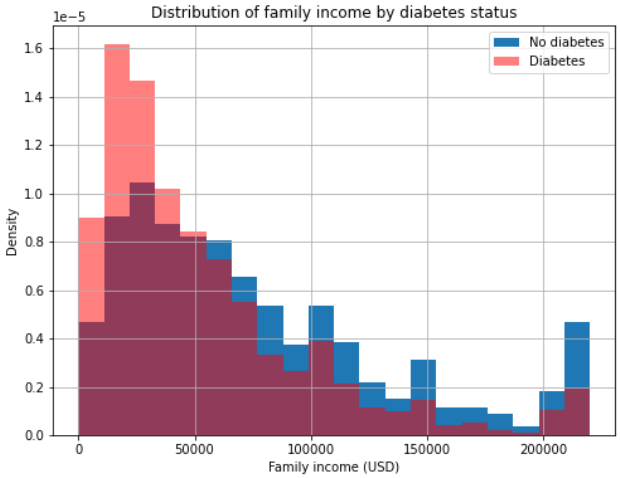
We chose to stratify the type (any, type 1, or type 2) of diabetes as the independent variable because early explorations showed Type 2 diabetes having a stronger correlation with negative income and labor outcomes (Figure 6 and Figure 7 respectively).

Linear Model Findings

Our findings from these linear models are broadly summarized by diabetes having strong negative effects (P-values < 0.001) on family income and probability of having worked recently, **except** for one key finding - that having type 1 diabetes is not statistically significant (at the α=0.01 level) with log family income after controlling for the factors discussed above (Figure 12). Examples of the broad findings include the fact that having type 2 diabetes is correlated with a 10% decrease to the probability of having worked last week (Figure 13), and correlates to a -0.14 log-point loss in family income (Figure 10).

The implications of the type 1 diabetes non-discovery are that type 1 diabetes is genetic, with earlier discoveries and treatment available. Therefore patients may be able to make up for the otherwise lost family income the counterfactual patients with type 2 diabetes lost through earlier diagnoses. It still comes as a surprise since families tend to share type 1 diabetes outcomes (having or not having it).

Hypothesis Testing



We decided to test our main hypothesis on the relationship between family income and the probability of developing diabetes directly using a pooled variance two-sample t-test. We defined our null and alternative hypothesis as follows.

H_0: *There is no relationship between income and having diabetes*

H_1: *There is a significant relationship between income and having diabetes*

We split our sample into those with and without diabetes, then ran a two-sample t-test, where we found a p-value of 1.042×10^{-104}, which is much lower than a standard significance level of $\alpha = 0.05$. Thus, we're able to clearly reject our null hypothesis and dive deeper into the consequences of an income disparity related to diabetes.

T-Test Assumptions

In order for our t-test to be valid, we had to assume normality and equal variance. Due to the large sample size of our dataset, we used the tendency of the t-curve to approach a normal curve in backing our normality assumption. We also checked the standard deviations for both groups, and found that they were 57,271 and 46,842, and thus close enough to use a pooled variance t-test. Without using pooled variance we got a p-value of 9.57621149e-135, even more significant than our original pooled variance finding.

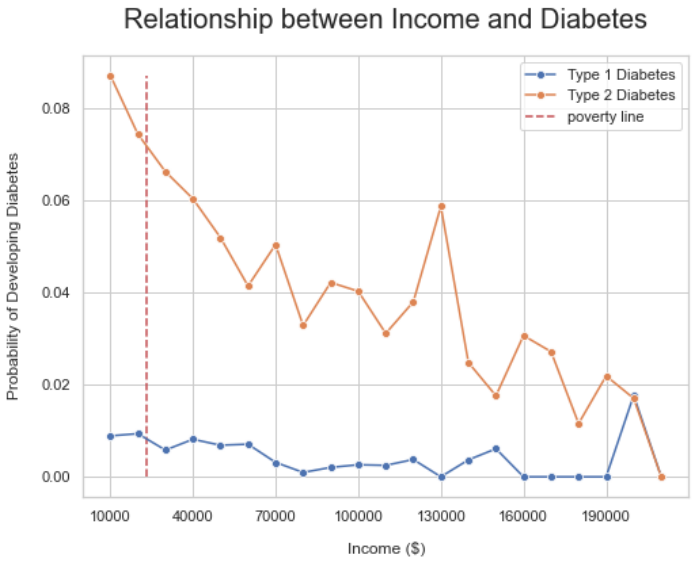


Figure 1

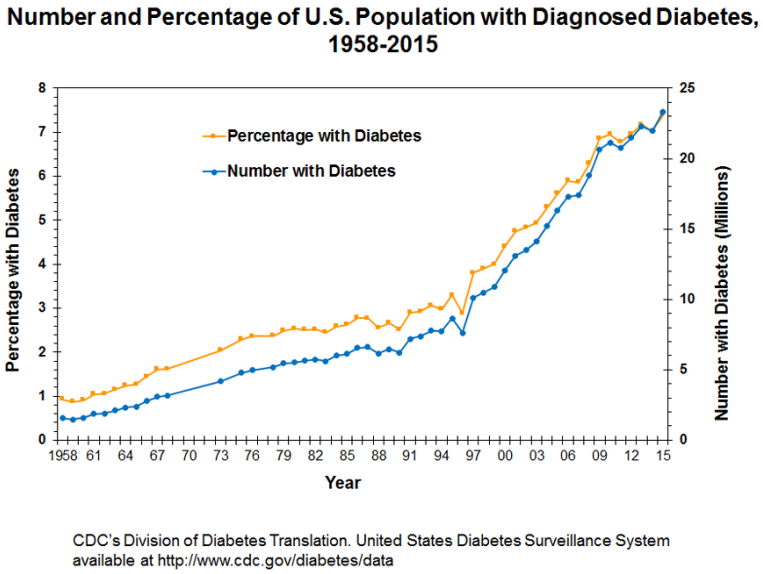


Figure 2

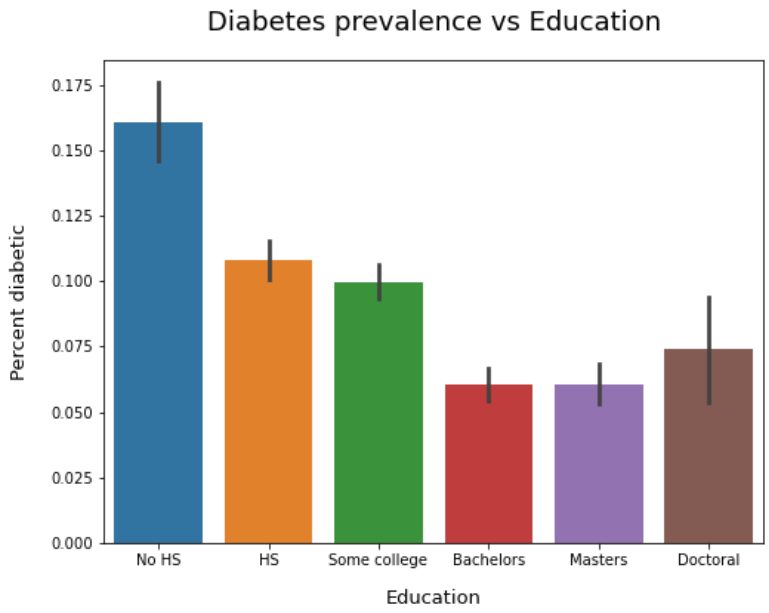


Figure 4

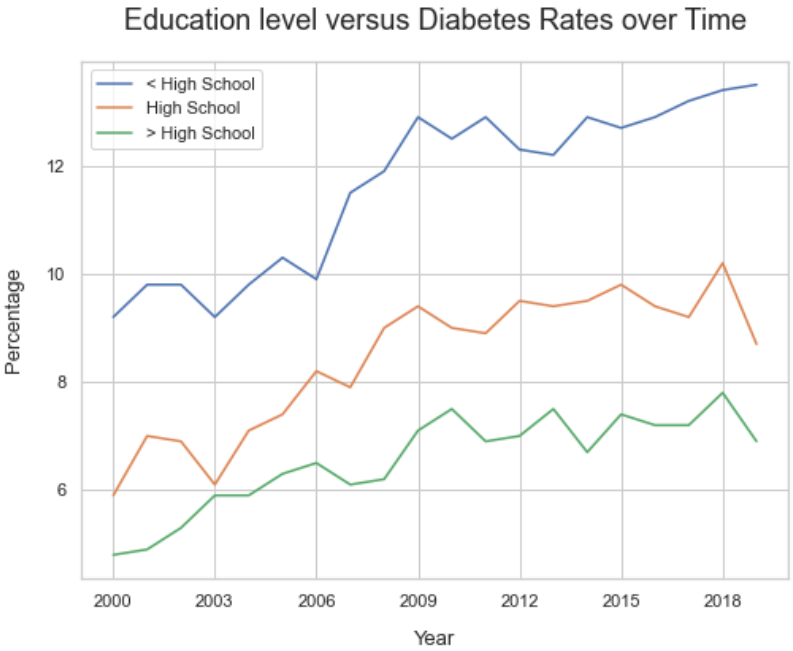


Figure 5

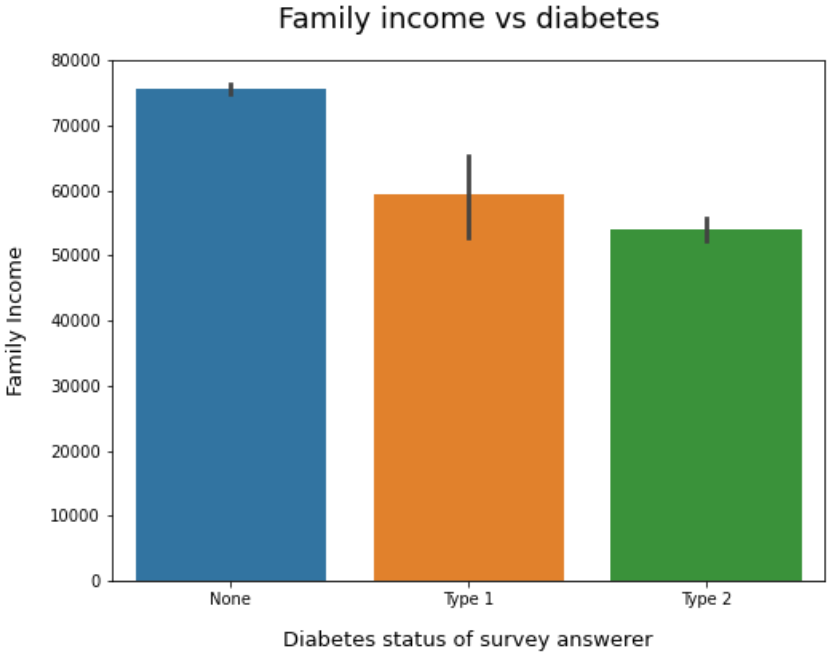


Figure 6

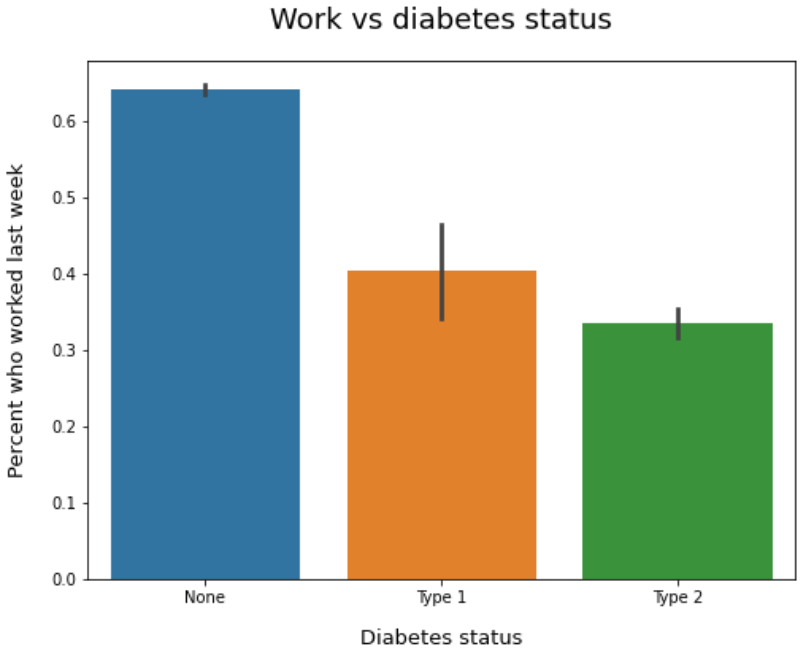


Figure 7

(1)

OLS Regression Results

Dep. Variable:	log family income	R-squared:	0.308
Model:	OLS	Adj. R-squared:	0.307
Method:	Least Squares	F-statistic:	888.3
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	0.00
Time:	14:48:47	Log-Likelihood:	-32642.
No. Observations:	28010	AIC:	6.531e+04
Df Residuals:	27995	BIC:	6.544e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8292	0.110	98.464	0.000	10.614	11.045
age	0.0170	0.002	10.796	0.000	0.014	0.020
age squared	-0.0002	1.5e-05	-13.516	0.000	-0.000	-0.000
0.0	-1.1745	0.036	-32.841	0.000	-1.245	-1.104
12.0	-0.8225	0.033	-24.742	0.000	-0.888	-0.757
14.0	-0.6377	0.033	-19.288	0.000	-0.702	-0.573
16.0	-0.3148	0.033	-9.440	0.000	-0.380	-0.249
18.0	-0.1112	0.035	-3.220	0.001	-0.179	-0.043
married	0.6062	0.010	62.684	0.000	0.587	0.625
diabetic	-0.1415	0.018	-8.080	0.000	-0.176	-0.107
type1	0.0152	0.052	0.294	0.769	-0.086	0.116
bmi	13.1941	4.883	2.702	0.007	3.624	22.764
bmi squared	-209.9795	56.981	-3.685	0.000	-321.666	-98.293
female	-0.1151	0.010	-12.104	0.000	-0.134	-0.096
minority	-0.1958	0.011	-18.640	0.000	-0.216	-0.175

Omnibus:	17133.700	Durbin-Watson:	1.820
Prob(Omnibus):	0.000	Jarque-Bera (JB):	662502.506
Skew:	-2.356	Prob(JB):	0.00
Kurtosis:	26.355	Cond. No.	4.44e+07

Figure 8

(2)

OLS Regression Results

Dep. Variable:	log family income	R-squared:	0.306
Model:	OLS	Adj. R-squared:	0.306
Method:	Least Squares	F-statistic:	949.4
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	0.00
Time:	14:48:47	Log-Likelihood:	-32675.
No. Observations:	28010	AIC:	6.538e+04
Df Residuals:	27996	BIC:	6.549e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8578	0.110	98.663	0.000	10.642	11.074
age	0.0167	0.002	10.585	0.000	0.014	0.020
age squared	-0.0002	1.5e-05	-13.612	0.000	-0.000	-0.000
0.0	-1.1807	0.036	-32.983	0.000	-1.251	-1.111
12.0	-0.8243	0.033	-24.767	0.000	-0.890	-0.759
14.0	-0.6398	0.033	-19.330	0.000	-0.705	-0.575
16.0	-0.3140	0.033	-9.403	0.000	-0.379	-0.249
18.0	-0.1097	0.035	-3.173	0.002	-0.177	-0.042
married	0.6078	0.010	62.786	0.000	0.589	0.627
type1	-0.1103	0.049	-2.241	0.025	-0.207	-0.014
bmi	13.0868	4.888	2.677	0.007	3.506	22.668
bmi squared	-218.5304	57.037	-3.831	0.000	-330.326	-106.735
female	-0.1131	0.010	-11.890	0.000	-0.132	-0.094
minority	-0.2007	0.010	-19.117	0.000	-0.221	-0.180

Omnibus:	17117.835	Durbin-Watson:	1.819
Prob(Omnibus):	0.000	Jarque-Bera (JB):	658195.192
Skew:	-2.355	Prob(JB):	0.00
Kurtosis:	26.276	Cond. No.	4.44e+07

Figure 9

(3)

OLS Regression Results

Dep. Variable:	log family income	R-squared:	0.307
Model:	OLS	Adj. R-squared:	0.307
Method:	Least Squares	F-statistic:	955.9
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	0.00
Time:	14:48:47	Log-Likelihood:	-32645.
No. Observations:	28010	AIC:	6.532e+04
Df Residuals:	27996	BIC:	6.543e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8289	0.110	98.452	0.000	10.613	11.044
age	0.0170	0.002	10.757	0.000	0.014	0.020
age squared	-0.0002	1.5e-05	-13.491	0.000	-0.000	-0.000
0.0	-1.1754	0.036	-32.864	0.000	-1.246	-1.105
12.0	-0.8229	0.033	-24.749	0.000	-0.888	-0.758
14.0	-0.6377	0.033	-19.287	0.000	-0.703	-0.573
16.0	-0.3147	0.033	-9.435	0.000	-0.380	-0.249
18.0	-0.1109	0.035	-3.213	0.001	-0.179	-0.043
married	0.6064	0.010	62.694	0.000	0.587	0.625
type2	-0.1397	0.017	-7.982	0.000	-0.174	-0.105
bmi	13.2694	4.883	2.717	0.007	3.698	22.840
bmi squared	-211.0534	56.986	-3.704	0.000	-322.748	-99.359
female	-0.1149	0.010	-12.087	0.000	-0.134	-0.096
minority	-0.1961	0.011	-18.673	0.000	-0.217	-0.176

Omnibus:	17127.126	Durbin-Watson:	1.819
Prob(Omnibus):	0.000	Jarque-Bera (JB):	661409.475
Skew:	-2.356	Prob(JB):	0.00
Kurtosis:	26.335	Cond. No.	4.44e+07

Figure 10

(4)

OLS Regression Results

Dep. Variable:	worked last week	R-squared:	0.354
Model:	OLS	Adj. R-squared:	0.353
Method:	Least Squares	F-statistic:	1178.
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	0.00
Time:	14:48:47	Log-Likelihood:	-13483.
No. Observations:	28010	AIC:	2.699e+04
Df Residuals:	27996	BIC:	2.711e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2228	0.055	4.015	0.000	0.114	0.332
age	0.0263	0.001	33.033	0.000	0.025	0.028
age squared	-0.0004	7.57e-06	-49.865	0.000	-0.000	-0.000
0.0	-0.2584	0.018	-14.319	0.000	-0.294	-0.223
12.0	-0.1838	0.017	-10.955	0.000	-0.217	-0.151
14.0	-0.1540	0.017	-9.230	0.000	-0.187	-0.121
16.0	-0.1002	0.017	-5.956	0.000	-0.133	-0.067
18.0	-0.0723	0.017	-4.148	0.000	-0.106	-0.038
married	0.0082	0.005	1.683	0.092	-0.001	0.018
diabetic	-0.1063	0.008	-12.611	0.000	-0.123	-0.090
bmi	17.3844	2.464	7.056	0.000	12.556	22.213
bmi squared	-196.9983	28.752	-6.852	0.000	-253.353	-140.644
female	-0.0825	0.005	-17.203	0.000	-0.092	-0.073
minority	-0.0137	0.005	-2.580	0.010	-0.024	-0.003

Omnibus:	1712.833	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1921.644
Skew:	-0.620	Prob(JB):	0.00
Kurtosis:	2.670	Cond. No.	4.44e+07

Figure 11

(5)

OLS Regression Results

Dep. Variable:	worked last week	R-squared:	0.351
Model:	OLS	Adj. R-squared:	0.350
Method:	Least Squares	F-statistic:	1163.
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	0.00
Time:	14:48:47	Log-Likelihood:	-13548.
No. Observations:	28010	AIC:	2.712e+04
Df Residuals:	27996	BIC:	2.724e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2445	0.056	4.398	0.000	0.136	0.353
age	0.0261	0.001	32.688	0.000	0.025	0.028
age squared	-0.0004	7.59e-06	-49.929	0.000	-0.000	-0.000
0.0	-0.2628	0.018	-14.531	0.000	-0.298	-0.227
12.0	-0.1850	0.017	-11.002	0.000	-0.218	-0.152
14.0	-0.1555	0.017	-9.303	0.000	-0.188	-0.123
16.0	-0.0996	0.017	-5.906	0.000	-0.133	-0.067
18.0	-0.0712	0.017	-4.078	0.000	-0.105	-0.037
married	0.0093	0.005	1.911	0.056	-0.000	0.019
type1	-0.1330	0.025	-5.350	0.000	-0.182	-0.084
bmi	17.2803	2.469	6.998	0.000	12.440	22.120
bmi squared	-203.1233	28.813	-7.050	0.000	-259.599	-146.648
female	-0.0811	0.005	-16.873	0.000	-0.091	-0.072
minority	-0.0173	0.005	-3.255	0.001	-0.028	-0.007

Omnibus:	1731.742	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1925.307
Skew:	-0.618	Prob(JB):	0.00
Kurtosis:	2.650	Cond. No.	4.44e+07

Figure 12

(6)

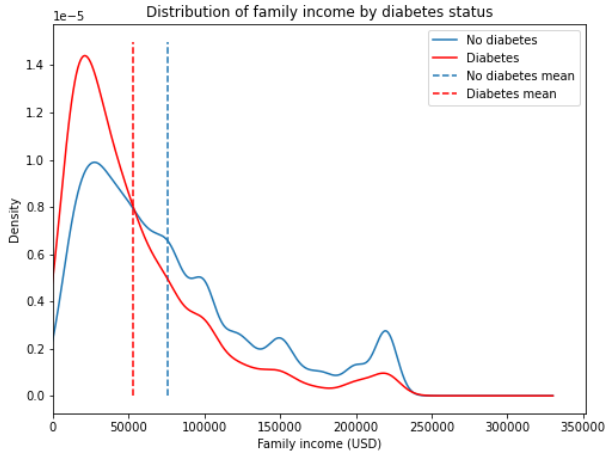
OLS Regression Results

Dep. Variable:	worked last week	R-squared:	0.353
Model:	OLS	Adj. R-squared:	0.353
Method:	Least Squares	F-statistic:	1175.
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	0.00
Time:	14:48:47	Log-Likelihood:	-13498.
No. Observations:	28010	AIC:	2.702e+04
Df Residuals:	27996	BIC:	2.714e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2236	0.056	4.026	0.000	0.115	0.332
age	0.0263	0.001	32.939	0.000	0.025	0.028
age squared	-0.0004	7.57e-06	-49.803	0.000	-0.000	-0.000
0.0	-0.2594	0.018	-14.365	0.000	-0.295	-0.224
12.0	-0.1841	0.017	-10.969	0.000	-0.217	-0.151
14.0	-0.1541	0.017	-9.230	0.000	-0.187	-0.121
16.0	-0.1001	0.017	-5.945	0.000	-0.133	-0.067
18.0	-0.0720	0.017	-4.131	0.000	-0.106	-0.038
married	0.0084	0.005	1.718	0.086	-0.001	0.018
type2	-0.0998	0.009	-11.299	0.000	-0.117	-0.082
bmi	17.4437	2.465	7.077	0.000	12.612	22.275
bmi squared	-198.1944	28.767	-6.890	0.000	-254.579	-141.810
female	-0.0823	0.005	-17.155	0.000	-0.092	-0.073
minority	-0.0141	0.005	-2.666	0.008	-0.025	-0.004

Omnibus:	1717.627	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1925.218
Skew:	-0.620	Prob(JB):	0.00
Kurtosis:	2.667	Cond. No.	4.44e+07

Figure 13

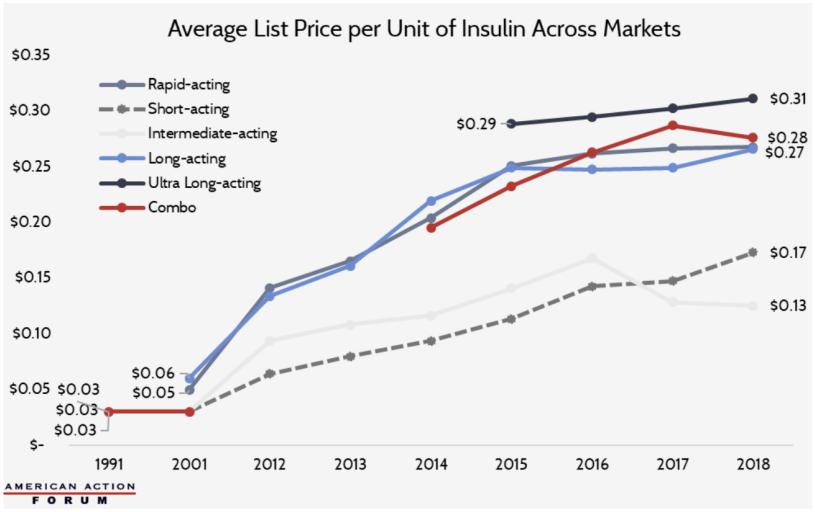


KDE plot of income distribution by diabetes diagnosis

Conclusions

Our main findings confirm that socio-economic factors have a very tangible relationship to diabetes rates. Broadly, having diabetes (as well as its specific types) correlates to a strong decrease in income and employment outcomes. There is one specific case where this isn't true - type 1 diabetes is not correlated with family income after controlling for external factors. We hypothesize this may be due to early diagnoses preventing poor salary outcomes. Furthermore, we find an inverse relationship between type II diabetes (the variant caused by lifestyle choices and the environment) and family income. We confirmed via t-test that diabetes diagnosis rates were higher for low-income families compared to high-income ones.

Our findings suggest that low income and disadvantaged individuals are the most at risk to developing diabetes. And as the groups with the lowest access to healthcare and treatment, this poses a massive problem for public health officials.



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2743027/

[Figure 2]

The price of insulin is also increasing, which also puts even more disproportionate pressure on lower income families. Our findings suggest that it may be necessary for the government to expand health programs for lower income families.

III. Appendix

Caveats

Our analysis shows that there is a correlation between people with low income and diabetes. However, while it might be logical that low income is the causal factor for higher risk of diabetes, it could also be the other way around. People with diabetes may have a harder time working, or be affected by high medical costs. This is mitigated, though, by our finding that type I diabetes is not correlated with low income.

Future Research Areas

Most of our analysis was done over one year, 2019. It would be interesting to conduct further research on how the correlated factors changed over the past decade. This could answer such questions as, is income becoming a larger or smaller factor in risk for diabetes? This could shed light on whether the rising income inequality could be contributing to the rise of diabetes in the United States.

Another avenue to explore would be attempting to isolate income and race separately as correlation factors with diabetes. There is some correlation between race and income. Thus, analyzing data which includes income and diabetes stats for just one race would provide a better picture of how race specifically affects risk for diabetes. Similarly, examining discrepancies in

diabetes rates between races in just one income range could give a more accurate idea of the relationship between race and diabetes.

Unsuccessful Analysis Pathways

The main approaches we tried were trying to predict readmission rates and length of hospital stay. We decided to not use the medicines provided as features, since our knowledge in that area was not sufficient, and we wouldn't be able to interpret the results without extensive research into what each medicine is. However, our initial findings didn't seem very promising, as everything could be explained by a correlation with severity in condition. The worse a patient's condition, the longer they have to stay in the hospital, the more procedures they will get, the more likely they are to go through emergency/urgent care, and the more likely they are to be readmitted later.

There was interesting data related to readmission rate, however, concerning the medical specialty the patient was seen by. There was a significant difference in readmission rates between different specialties, even excluding the ones with low sample size. Given more time, this may have been a good place for further investigation.

Dataset Links

National Health Interview Survey 2019
https://www.cdc.gov/nchs/nhis/2019nhis.htm

CDC Diabetes Diagnosis
https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html

American Action Forum
https://www.americanactionforum.org/research/insulin-cost-and-pricing-trends/