Diabetes Prevalence in the U.S.

We are interested in factors affecting the prevalence of diabetes in the US, specifically Social Vulnerability Index (SVI), Average Income, Physician's Rate, Walkability, Gender, Education Level, and Smoking history.

BACKGROUND

- *High prevalence of diabetes: 38.4 million of Americans, which was 11.6% of the population, had diabetes in 2021 (CDC).*
- *High mortality rate: the eighth leading cause of death in the United States.*
- *Significant economic burden: the total estimated cost of diagnosed diabetes in the U.S. in 2022 is $412.9 billions (Parker).*

RESEARCH QUESTIONS

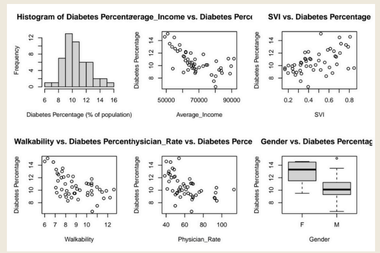
- *Do states with higher average income have a lower prevalence of diabetes than those with lower average income?*
- *Do states with a higher physician rate have a lower prevalence of diabetes than those with less?*
- *Do environmental factors, such as higher walkability, have a lower prevalence of diabetes than states with little walkability?*

01. Data Summary and EDA

- **Response Variable:** Diabetes Prevalence in US adults since 2021
- **Data**: Obtained from Social Explorer, CDC, and UVA Policy Map, all with data collecting from year 2021, stratified by 50 states
- **Manipulation**: 1). To match for other explanatory variables, the SVI for each county was added together and then averaged, rounded to four decimal places. 2). We changed the data of gender, smoking, and education to a qualitative variable by classifying them by if they met a certain threshold.

Data Summary Table

Name	Abbreviation	Description	Unit
Diabetes Percentage	Diabetes_Level	The percent of the population in the state that have Diabetes (response variable).	Percentage (%)
Average Income	Average_Income	The average income per state.	Dollars ($)
Social Vulnerability Index	SVI	The negative effects on communities caused by external stressors on human health	Scale from 0-1, where 0 indicates low vulnerability and 1 is high vulnerability
Population Density	Population_Density	The average number of people per square mile per state	mile²
Average National Walkability Index	Walkability	Ranks the census block groups on whether the primary mode of transportation is walking based on various environmental factors.	Scale of 1-20
Fast Food Restaurant	Fast_Food	The average of the average number of fast food restaurants in each county of a U.S. state.	Number of Fast Food Restaurants
Average Physician Rate	Physician_Rate	The number of primary care physicians per 100,000 people in each county, then averaged and grouped by state	Percentage (%)
Gender Proportion	Gender	The "majority" gender proportion that have diabetes per state	Two levels: male and female
Smoker	Smokers	If 20% of the state population over 18 years-old are smokers.	Two levels: yes and no
Education Level	Education_Level	If more than 10% of the state population have not earned a high-school degree or equivalent.	Two levels: yes and no



02. Model Modeling

Stage One: Quantitative Variables:
Initial:$Diabetes Level = \beta 0 + \beta 1 Average\_Income + \beta 2 SVI + \beta 3 Walkability + \beta 4 Population\_Density + \beta 5\ FastFood + \beta 6\ Physician\_Rate$
Final:$Diabetes Level = \beta 0 + \beta 1 Average\_Income + \beta 2 SVI + \beta 3 Walkability + \beta 4 Population\_Density$

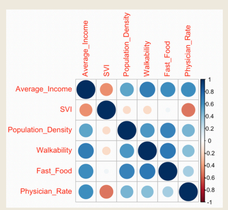
Stage Two: Qualitative Variables:
$Initial: Diabetes Level = \beta 0 + \beta 1 Average\_Income + \beta 2 SVI + \beta 3 Walkability + \beta 4 Population\_Density + \beta 5 DummyEducation\_Level + \beta 6 DummySmoking + \beta 7 DummyGender$

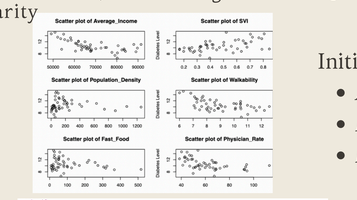
$Final: Diabetes Level = \beta 0 + \beta 1 Average\_Income + \beta 2 SVI + \beta 3 Walkability + \beta 4 Population\_Density + \beta 5 DummySmoking$

where dummyEducation level = {1 if yes, 0 if no}, dummySmoking={1 if yes, 0 if no} dummyGender={1 if male, 0 if female }

Stage Three: Interaction:
There were no interactions to add after second step.

Multicolinearity:
Average VIF for all variables <10, and average VIF is 2.512 (<3), indicting no strong concern for multicolinearity






```
> pricevif
Average_Income            SVI Population_Density      Walkability        Fast_Food     Physician_Rate
         3.369          1.931           2.825           2.595           3.241           1.911
> mean(pricevif)
[1] 2.512
```

Initial Model:

- *Adjusted R^2:0.684*
- *MSE*: 0.8811471
- *Pvalue*: 7.289e-10

03. Analysis

Variable Screening

Stepwise Selection

p_ent = 0.05 and p_rem = 0.05

Stepwise Summary

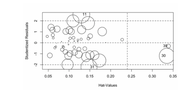
Step	Variable	Addition/Removal	AIC	SBC	SBIC	R2	Adj. R2
1	Average_Income	Addition	177.255	182.931	NA	0.40909	0.39652
2	Population_Density	Addition	168.314	175.881	NA	0.52735	0.50680
3	Walkability	Addition	162.376	171.835	NA	0.59804	0.57124
4	SVI	Addition	153.880	165.231	NA	0.67554	0.64604

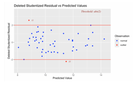
- Average_Income, Population_Density, Walkability, SVI were selected and added to the model
- Fast food and physician rate were not added because they weren't significant

Outliers and Influential Observations

	StudRes <dbl>	Hat <dbl>	CookD <dbl>
11	2.5674070	0.1295170	0.145027140
30	-1.1907946	0.3349374	0.117900959
31	-2.2279370	0.1479356	0.131763417
39	-0.3019042	0.3388558	0.007950057

- Observations 11, 30, 31 are influential because they exceed Cook's Distance threshold of 0.08
- Observation 11 and 31 are outliers in the y-direction since studentized residual threshold is exceeded
- Observation 30 and 39 are outliers in the x-direction because they exceed the leverage threshold of 0.2
- Removed observation 11 and 31





04. Final Model

$$Diabetes Level = 17.34 - (7.966e-5) Average\_Income + 0.003068 Population\_Density^2 - 0.3729 Walkability + 2.905 SVI + 0.8216 Smokers$$

where dummySmoking={1 if yes, 0 if no}

Adjusted R^2:0.7604

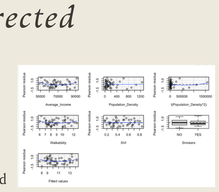
MSE: 0.899

P-value: 4.919e-13

Interpretation: Our final model has a higher adjusted R^2 than our initial model, indicating it accounts for more of the variance in our model. The lower MSE and pvalue indicate the model is more accurate.

05. Assumptions Corrected

- To correct for the assumptions violated in our initial analysis, we have added a higher order term to population density to correct for the quadratic pattern. By doing so our model now does not violate any of the assumptions.
- By removing the influential observations and outliers, specifically numbers 11 and 31, our model is better at predicting diabetes prevalence.
- These observations were removed due to their cooks distance value, indicating them as influential points, as well as their studentized residual value, indicating them as outliers in the y direction, both of which were above the threshold. These observations were likely skewing our model and making it less accurate.



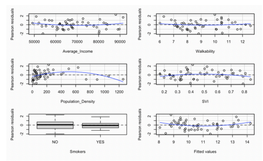
06. External Model Validation: Jackknifing

Original Model:
- Adjusted R^2: 0.6603
- MSE: 1.1456

Jackknifing:
- Adjusted R^2:0.6609
- MSE: 1.1424

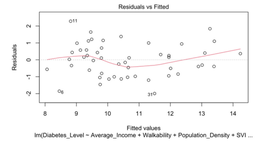
Interpretation: These jackknife measures give a more conservative assessment of the ability of the model to predict future observations. While it is typical for the adjusted R^2 to be smaller and the MSE to be larger, the opposite is true in this model. These slight variations are likely due to either the original model overfitting the data or containing influential observations that are skewing the data. However, the values are very close to one another, suggesting stability and reliability of the model.

Assumptions

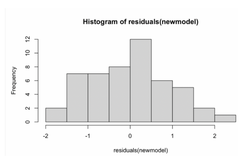


1. **Mean Zero**
   - There is a violation of this assumption for population density due to a curvilinear trend indicating a lack of fit. This will be corrected later with transformation by refitting with higher-order terms.

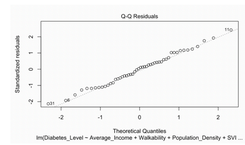
2. **Homoscedasticity**
   - No violation of this assumption because residuals are evenly spread. There is no "fanning out" pattern.



3. **Normality**
   - No violation of this assumption. The histogram of the residuals is unimodal and slightly symmetric. The qq-plot shows clusters in the middle and slight trailing off at the ends, but most of the values are very close to the line. Additionally, regression is robust against this minor violation.



4. **Independence**
   - No violation of this assumption. The data is not time-series, therefore the observations are not dependent on each other.



07. Model Assesement and Conclusion

Increased population density squared, SVI, and smokers increase the diabetes prevalence rate, while increased average income and walkability decrease diabetes prevalence rate in the US.
Example: For the state of Virginia, based on our final model, we would expect to have a diabetes prevalence rate of 9.2, while the actual rate, according to the data from 2021, is 10.7, resulting in a residual of 1.5.
Conclusion: Overall, the model is useful at predicting diabetes prevalence rate in the US based on the factors like average income, population density, walkability, SVI, and smoking or not. This answers our research question that states with higher average income and higher walkability have a lower prevalence diabetes rate than those with lower average income and walkability.
However, regarding the diabetes prevalence and physician rate, there is no significant relationship between these two variables.

08. Improvements/Limitations

Limitations:
- Diabetes level can be influenced by various factors apart from the variables we have analyzed, including genetics which can be difficult to measure. The diabetes we currently focus on are Type II, because Type I is more genetically predisposed

Improvements:
- We may use more qualitative data, such as age, ethnicity, etc. to ensure that the collecting data captures a wide range of demographic variables, which may provide a more comprehensive understanding of diabetes prevalence across different population groups.

09. Works Cited

Background

1. Centers for Disease Control and Prevention. (2023, November 29). National Diabetes Statistics Report. Centers for Disease Control and Prevention. https://www.cdc.gov/ diabetes/data/statistics-report/index.html
2. Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ElSayed, N. A., & Bannuru, R. R. (2023, November 1). Economic costs of diabetes in the U.S. in 2022. American Diabetes Association. https://diabetesjournals.org/care/article/47/1/26/153797/Economic-Costs-of-Diabetes-in-the-U-S-in-2022
3. Bureau, U. C. (2023, November 2). National Diabetes Month: November 2023. Census.gov. https://www.census.gov/newsroom/stories/diabetes-month.html#:~:text=%E2%80%9CDiabetes%20is%20a%20disease%20that,to%20some%20types% 20of%20cancer
4. Hill, J., Nielsen, M., & Fox, M. H. (2013). Understanding the social factors that contribute to diabetes: A means to informing health care and social poli- cies for the chronically ill. The Permanente journal. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3662286/#:~:text=The%20sociobiologic%20cycle%20of%20diabetes.&text=Social%20determinants%20of%20health%20encompass,and%20access%20to%20nutritious%20food.&text=Lifestyle%20factors%20incorporate%20dietary%20choices,to%20primary%20health%20care%20services
5. Silva-Tinoco, R., Cuatecontzi-Xochitiotzi, T., De la Torre-Saldaña, V., Leon-Garcia, E., Serna-Alvarado, J., Orea-Tejeda, A., Castillo-Martínez, L., Gay, J. G., Cantu-de- Leon, D., & Prada, D. (2020, August 26). Influence of social determinants, diabetes knowledge, health behaviors, and glycemic control in type 2 diabetes: An analysis from real-world evidence - BMC endocrine disorders. BioMed Central. https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-020-00604-6

Data

1. Population Density (Per Sq. Mile) [Map]. In SocialExplorer.com. ACS 2016 (5-Year Estimates) Retrieved 24 March 2024, from https://www.socialexplorer.com/ 8e62e93dee/view
2. Average Household Income (In 2021 Inflation Adjusted Dollars) [Map]. In Social- Explorer.com. ACS 2021 (5-Year Estimates) Retrieved 24 March 2024, from https: //www.socialexplorer.com/a9676d974c/view
3. Total Population: [Map]. In SocialExplorer.com. ACS 2021 (5-Year Estimates) Re- trieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view
4. Less than High School [Map]. In SocialExplorer.com. ACS 2021 (5-Year Estimates) Retrieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view
5. Percent of Current Smokers (Persons 18 Years and Over) [Map]. In SocialExplorer.com. Health Data 2023 Release Retrieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view
6. Centers for Disease Control and Prevention. (2021). Social Determinants of health -
7. United States Diabetes Surveillance System. Centers for Disease Control and Prevention. https://gis.cdc.gov/grasp/diabetes/diabetesatlas-sdoh.html#
8. PolicyMap. (n.d.). National walkability index in 2021 [Map based on data from EPA Smart Location Database: Data downloaded from https://www.epa.gov/smartgrowth/ smart-location-mapping#SLD, January 2022]. Retrieved March 24, 2024, from http://www.policymap.com