

Diabetes Prediction

Alexander Wen, Raymond Wong, Michael Eirikson

Table of contents

Summary	1
Introduction	1
Methods	2
EDA	2
Modeling	11
Discussion	11
References	11

Summary

In this project we attempt to build a model to predict diabetes disease. We compared a decision tree model and naive bayes model and found the decision tree is stronger in this context. We used f2-score as our scoring function because detecting diabetes is the priority: a false negative could be much worse than a false positive.

In the test dataset: the decision tree model correctly detected 8283 of 10604 positive cases (recall rate is about 78%). This result does come at a fairly significant cost in terms of false positives (precision rate is about 30%) with 19650 false positives. Depending on the actual cost of false positive this may need significant improvement to be a viable screening model.

Introduction

In Canada and the USA approximately 10% of people are living with diabetes. In Canada in 2023 approximately 3.7 million people were living with diabetes and in the USA in 2021 approximately 38.4 million people were living with diabetes. (“Snapshot of Diabetes in Canada, 2023” (2023)) In the USA it is the 8th leading cause of death. (Rios et al. (2017)) Globally an estimated 44% of people living with diabetes are undiagnosed. (Stafford et al. (2025))

In this project we try to predicted diabetes disease based on common health factors. A reliable model could help to prescreen people and recommend following up with a physician for people who are at risk. Given the large number of people living with undiagnosed diabetes this could potentially have a significant positive impact of world health.

The analysis uses the American CDC Behavioural Risk Factor Surveillance System (BRFSS) 2015 Diabetes Health Indicators dataset (UCI ID 891), containing 253,680 survey responses with 21 health-related features and a binary diabetes outcome (0 = no diabetes/pre-diabetes, 1 = diabetes). (Dane and Teboul (2021))

No missing values were present and all features were already encoded numerically. The target classes is imbalanced (86% non-diabetic, 14% diabetic).

Methods

This analysis was performed in Python 3.11.6 (*Python 3.11.6 Documentation* 2021-2025). Additionally, here is a list of the Python packages used within the analysis with brief explanation:

Package	Version	Use case	Reference
numpy	1.26.4	General analysis use	<i>NumPy Documentation</i> (2008-2022)
pandas	2.1.2	Data management/processing	team (2020), McKinney (2010)
pandera	0.27.0	Data valiadion	Bantilan (2020)
altair	5.1.2	Generating plots	VanderPlas et al. (2018), Satyanarayan et al. (2017)
scikit-learn	1.3.2	Model creation and evaluation	Pedregosa et al. (2011)
ucimlrepo	0.0.7	Data extraction	Kelly, Longjohn, and Nottingham (2021)
deepchecks	0.18.1	Data validation	Chorev et al. (2022)
click	8.3.1	Script tool	Pallets (2020)
quarto	1.8.26	Report creation	Allaire et al. (2025)
tabulate	0.9.0	Table formatting	Astanin (2025)

Modeling Approach

Results

EDA

Data Summary

Table 2: Description of the training data.

HighBP	177576	0.4293	0.495	0	0	0	1	1
HighChol	177576	0.4228	0.494	0	0	0	1	1
CholCheck	177576	0.9627	0.1896	0	1	1	1	1
BMI	177576	28.385	6.5916	12	24	27	31	98
Smoker	177576	0.4433	0.4968	0	0	0	1	1
Stroke	177576	0.0407	0.1976	0	0	0	0	1
HeartDiseaseorAttack	177576	0.094	0.2919	0	0	0	0	1
PhysActivity	177576	0.7563	0.4293	0	1	1	1	1
Fruits	177576	0.6351	0.4814	0	0	1	1	1
Veggies	177576	0.8122	0.3905	0	1	1	1	1
HvyAlcoholConsump	177576	0.0563	0.2305	0	0	0	0	1
AnyHealthcare	177576	0.9513	0.2153	0	1	1	1	1
NoDocbcCost	177576	0.0845	0.2782	0	0	0	0	1
GenHlth	177576	2.5128	1.0685	1	2	2	3	5
MentHlth	177576	3.1833	7.4056	0	0	0	2	30
PhysHlth	177576	4.2542	8.7248	0	0	0	3	30
DiffWalk	177576	0.1682	0.374	0	0	0	0	1
Sex	177576	0.4408	0.4965	0	0	0	1	1
Age	177576	8.0321	3.0558	1	6	8	10	13
Education	177576	5.049	0.9863	1	4	5	6	6
Income	177576	6.0514	2.0729	1	5	7	8	8
diabetes	177576	0.1393	0.3463	0	0	0	0	1

Table 2 displays a numerical distribution of the dataset. Note that this and Table 3 are transposed tables for more syntactical presentation. All features of the dataset are numeric, and further EDA shows there are no null values in the dataset. Most features are binary in nature, with the exceptions of To better visualize we then plotted frequency counts of the target labels:

Table 3: First few rows of the training data as example observations.

HighBP	0	0	1	0	1
HighChol	1	0	1	0	0
CholCheck	1	1	1	1	1
BMI	23	25	28	25	30
Smoker	0	0	0	0	1
Stroke	0	0	0	0	0
HeartDiseaseorAttack	0	0	0	0	0
PhysActivity	1	1	1	0	1
Fruits	1	1	0	1	1
Veggies	1	1	1	1	1
HvyAlcoholConsump	0	0	0	0	0
AnyHealthcare	1	1	1	1	1
NoDocbcCost	0	0	0	0	1
GenHlth	1	3	2	2	4
MentHlth	0	0	15	0	30
PhysHlth	0	30	2	0	15
DiffWalk	0	0	0	0	0
Sex	0	0	1	0	0
Age	10	12	6	8	8
Education	6	6	5	5	4
Income	8	7	6	7	4
diabetes	0	0	0	0	0

Here are some examples of our data.

Visualizations

Count of Diabetes vs Non-Diabetes Records in Dataset

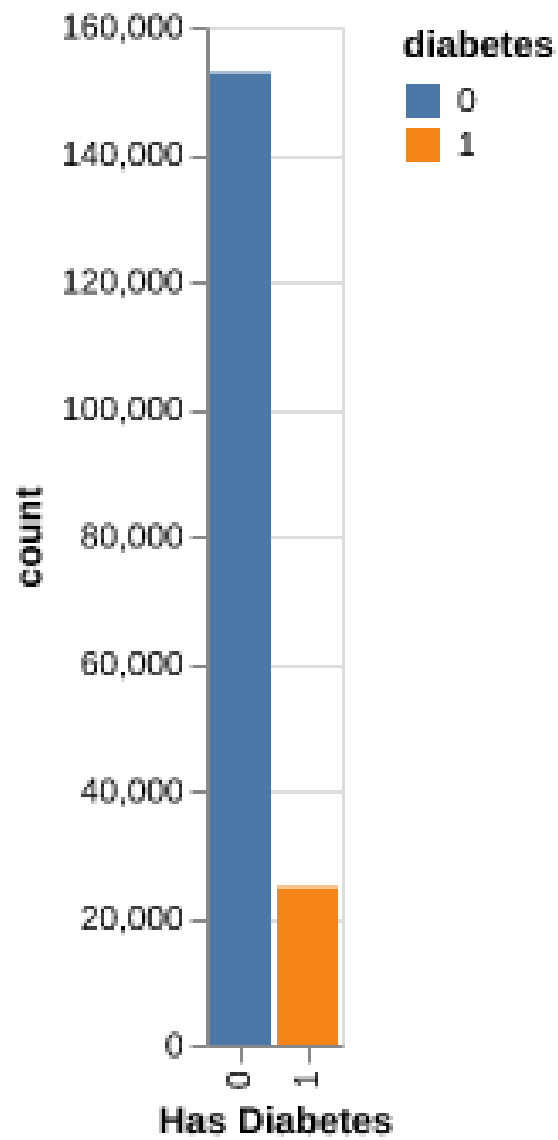


Figure 1: Frequency bar graph of the labels.

Figure 1

Bar Plots of Binary Features

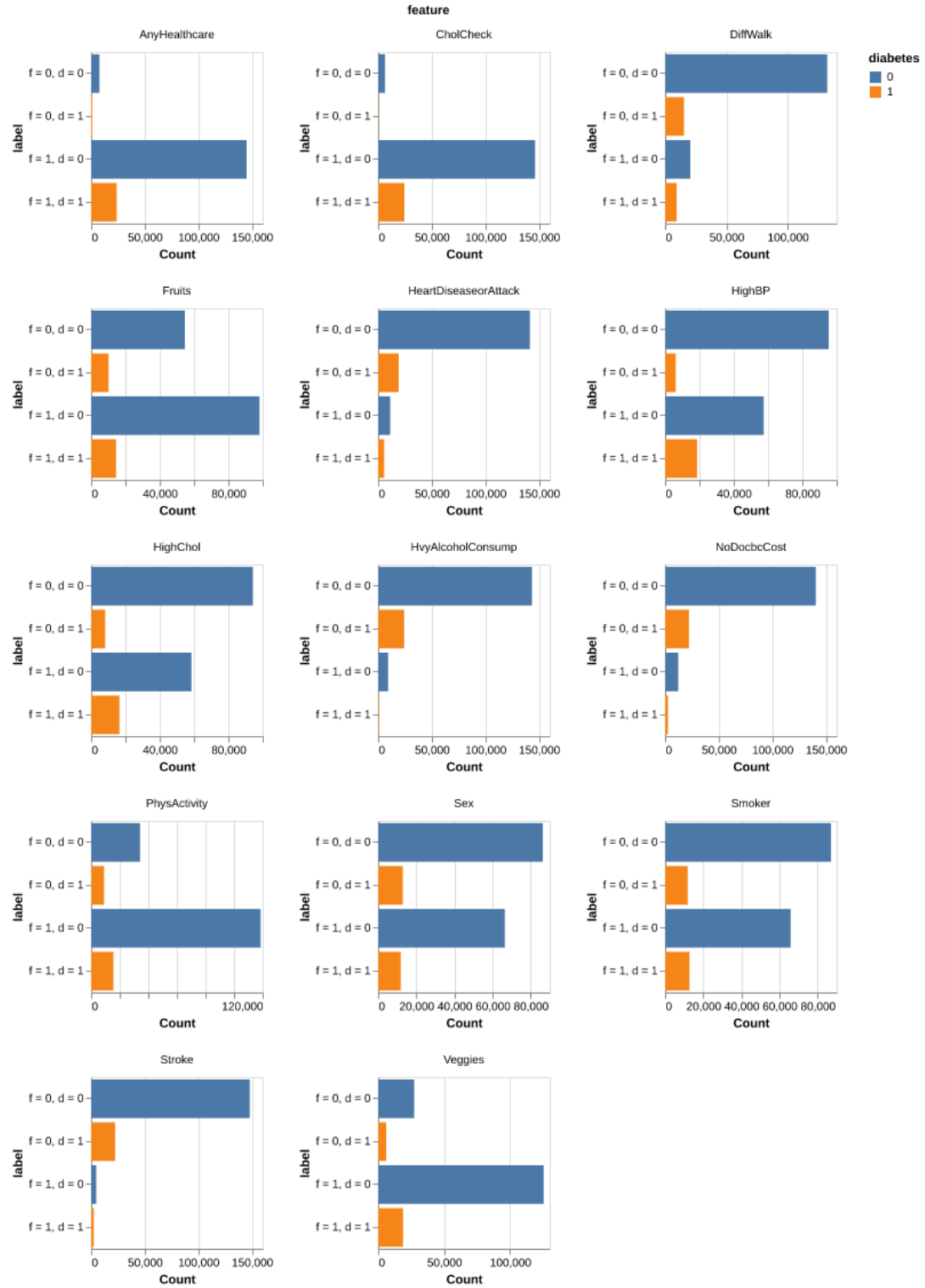


Figure 2: Frequency bar graphs of the binary features.

Figure 2

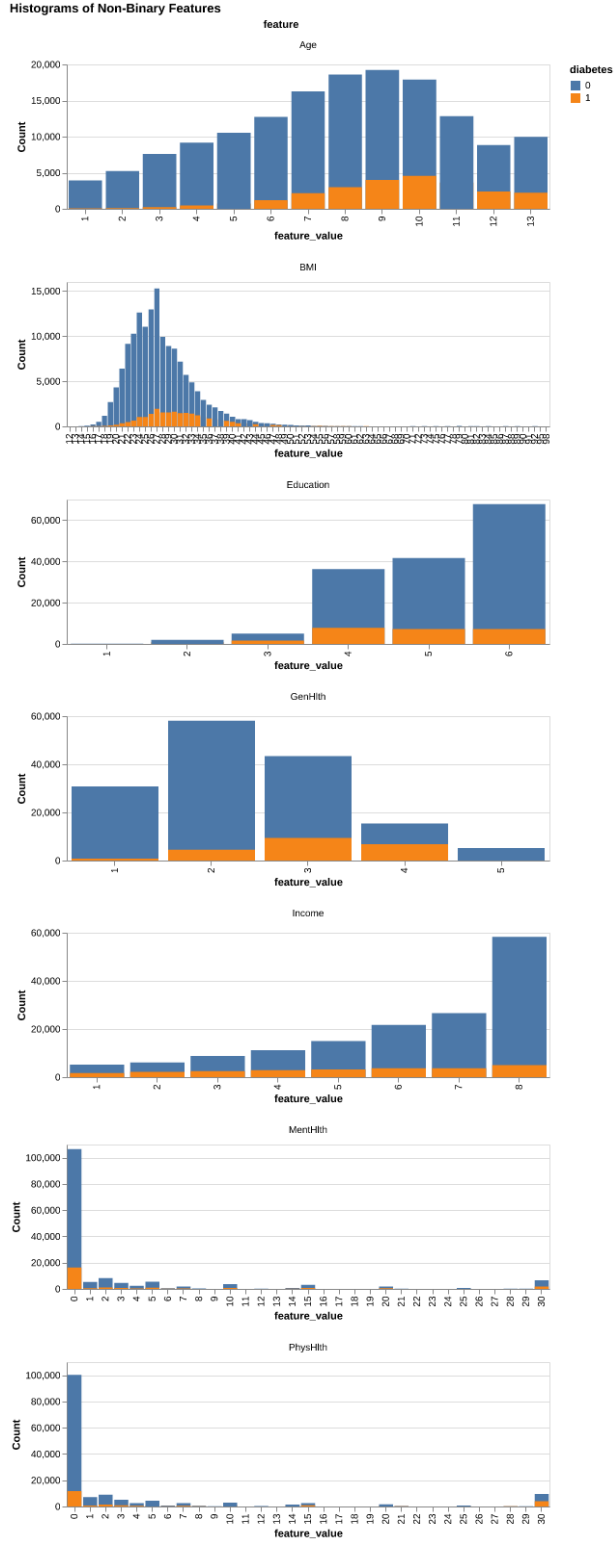


Figure 3: Histograms of the non-binary numeric features.

Figure 3

Boxplots for Non-Binary Features (Sample size n = 1000)

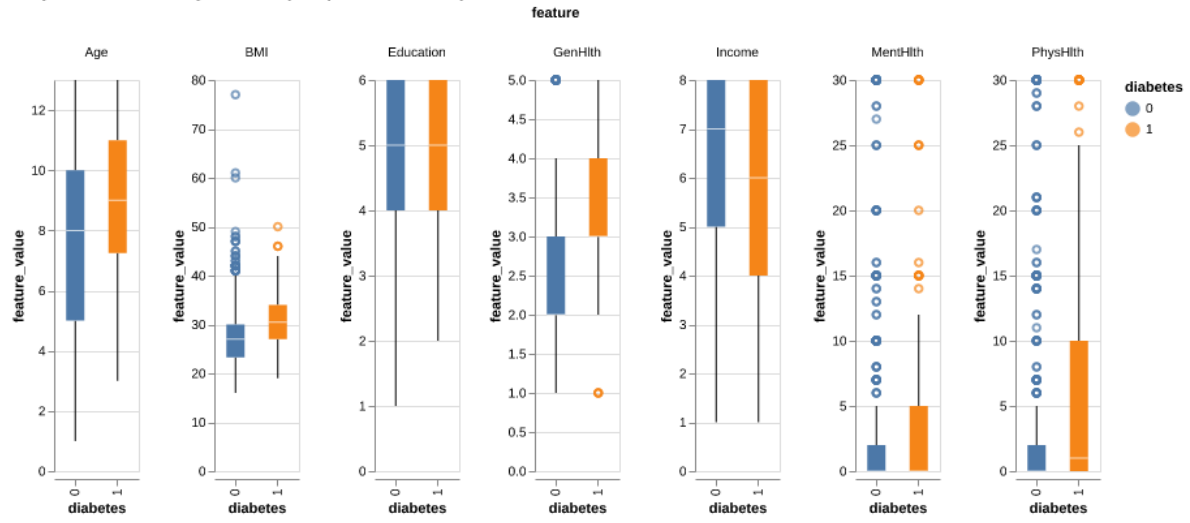


Figure 4: Boxplots of the non-binary numeric features.

Figure 4

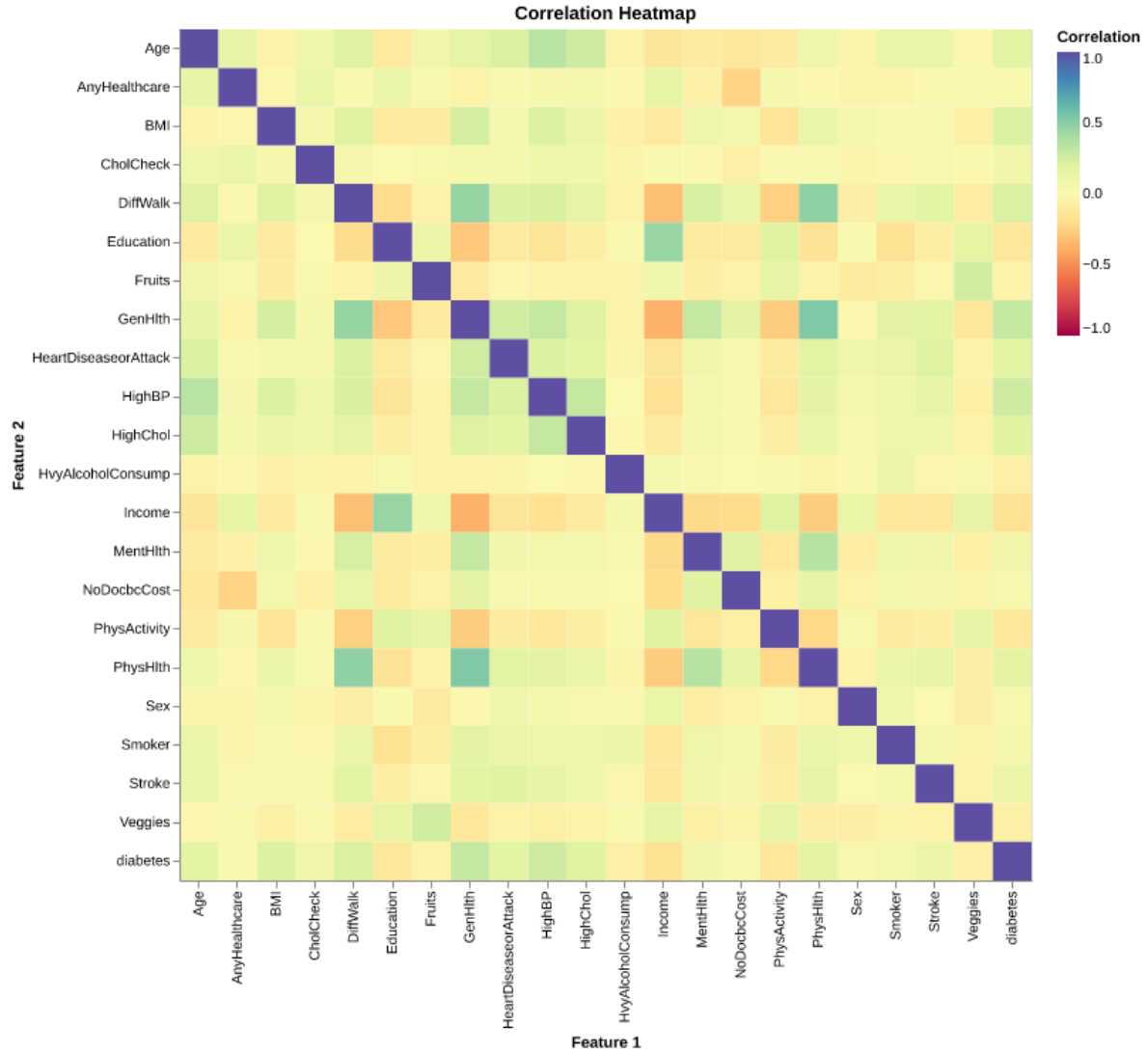


Figure 5: Feature-feature correlation plot of all features.

Figure 5

Modeling

Classification Analysis

Result Visualizations

Discussion

References

- Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, Christophe Dervieux, and Gordon Woodhull. 2025. “Quarto.” <https://doi.org/10.5281/zenodo.5960048>.
- Astanin, Sergeio. 2025. “Tabulate: Pretty-Print Tabular Data in Python.” <https://github.com/astanin/python-tabulate>.
- Bantilan, Niels. 2020. “Pandera: Statistical Data Validation of Pandas Dataframes.” In *Proceedings of the 19th Python in Science Conference*, edited by Meghann Agarwal, Chris Calloway, Dillon Niederhut, and David Shupe, 116–24. <https://doi.org/10.25080/Majora-342d178e-010>.
- Chorev, Shir, Philip Tannor, Dan Ben Israel, Noam Bressler, Itay Gabbay, Nir Hutnik, Jonatan Liberman, Matan Perlmutter, Yurii Romanyshyn, and Lior Rokach. 2022. “Deepchecks: A Library for Testing and Validating Machine Learning Models and Data.” <https://arxiv.org/abs/2203.08491>.
- Dane, Sohier, and Alex Teboul. 2021. “Diabetes Health Indicators Dataset.” <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data>; Center of Disease Control.
- Kelly, Markelle, Rachel Longjohn, and Kolby Nottingham. 2021. “The UCI Machine Learning Repository.” University of California, Irvine, School of Information; Computer Sciences. <http://archive.ics.uci.edu/ml>.
- McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. <https://doi.org/10.25080/Majora-92bf1922-00a>.
- NumPy Documentation*. 2008-2022. <https://numpy.org/doc/1.26/>.
- Pallets. 2020. *Click*. <https://click.palletsprojects.com/>.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.
- Python 3.11.6 Documentation*. 2021-2025. <https://docs.python.org/release/3.11.6/reference/index.html>.
- Rios, Nilka Burrows, Isreal Hora, Linda S. Geiss, Edward W. Gregg, and Ann Albright. 2017. “Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico.” *Incidence of End-Stage Renal*

- Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico*, no. 66(43): 1165–70.
- Satyanarayan, Arvind, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. “Vega-Lite: A Grammar of Interactive Graphics.” *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 341–50.
- “Snapshot of Diabetes in Canada, 2023.” 2023. <https://www.canada.ca/en/public-health/services/publications/diseases-conditions/snapshot-diabetes-canada-2023.html>; Public Health Agency of Canada.
- Stafford, Lauryn K, Anna Gage, Yvonne Yiru Xu, Madeleine Conrad, Ismael Barreras Beltran, and Edward J Boyko. 2025. “Global, Regional, and National Cascades of Diabetes Care, 2000–23: A Systematic Review and Modelling Analysis Using Findings from the Global Burden of Disease Study.” *The Lancet Diabetes & Endocrinology*, no. 13(11): 924–34.
- team, The pandas development. 2020. “Pandas-Dev/Pandas: Pandas.” Zenodo. <https://doi.org/10.5281/zenodo.3509134>.
- VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. “Altair: Interactive Statistical Visualizations for Python.” *Journal of Open Source Software* 3 (32): 1057. <https://doi.org/10.21105/joss.01057>.