Predicting if mental health will affect work performance (in tech)

[1 in 5 adults in the US experience mental health conditions each year](#) (46.6 million people in 2017.) People with mental health conditions may doubt their abilities and have a hard time learning, concentrating and making decisions, reducing overall productivity in the workplace. [Depression alone is estimated to cause 200 million lost workdays each year at a cost to employers of $17 to $44 billion](#). It's clear that a solution needs found to help not only increase productivity for the employer's benefit but to also increase the overall well being of employees.

[Research has found that feeling authentic and open at work leads to better performance, engagement, and overall well being.](#) With this in mind I will be exploring the key predictors that determine if an employee's mental health condition will affect their performance at work (in the employees own subjective perspective), using demographic factors such as age, gender, location, # of employees in company etc., as well as impact of wellness programs and benefits offered on said perception. Once these predictors are identified my goal is to identify specific areas employers can improve on/target to increase feelings of authenticity/openness regarding mental illness for employees, and thereby decrease the likelihood of their illness impacting work.

I will be using data from 'Mental Health in Tech' surveys, combining 4 years of data:
https://www.kaggle.com/osmi/mental-health-in-tech-survey
https://www.kaggle.com/osmi/mental-health-in-tech-2016
https://www.kaggle.com/osmihelp/osmi-mental-health-in-tech-survey-2017
https://www.kaggle.com/osmihelp/osmi-mental-health-in-tech-survey-2018
All respondents are anonymous with no identifiers, so assuming all responses are from unique individuals. While many questions remained the same over the years, there is some variation so I will need to normalize / clean.

To solve this problem I will start by combining the data, cleaning it, and pulling in data from APIs. Once data is cleaned I will model the data using logistic regression to predict the work_performance variable. The result will be 0=No and 1=Yes for if work performance is impacted by mental health.

My deliverables will be a combination of code, Jupyter Notebook, and slide deck.