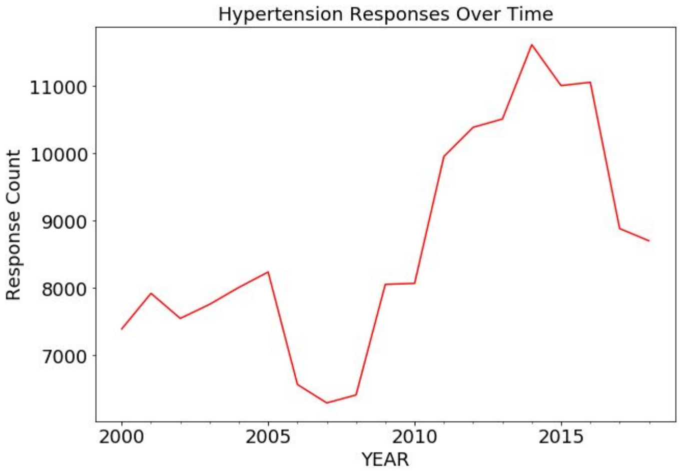
Alexia Marcous
Springboard Data Science Career Track
August 5th2019 Cohort

Capstone Project 1 – Predicting Hypertension
Data Story

Data science offers the exciting possibility of prediction, and with this comes the risk of predicting just for the sake of exploiting this powerful concept. So with any data science project, the first question should be: **is this relevant**? High blood pressure, known as hypertension, has been associated with increased risk of all-cause mortality, meaning *any* cause of death, and is the major risk factor for cardiovascular disease, the leading cause of death worldwide.[1] However, hypertension was declared an epidemic over twenty years ago.[2] Is it still an issue, or is our blood pressure now under control?

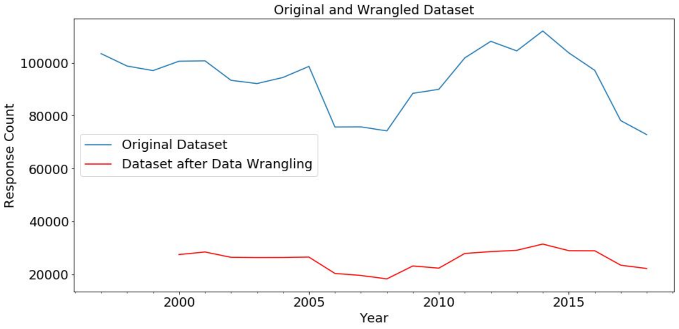
To answer this question, we can look at the trend of hypertension in the dataset prepared for this project. The data contains results from National Health Interview Surveys taken between 1997 and 2017, with 39 features selected based on data availability and relevance to hypertension, and 484,750 surveys selected for highest prevalence of non-null data. The trend of hypertension over survey years appears as follows:



This is extremely strange! Was there a miracle drug released to market in 2005 and taken off the market a couple of years later? What's causing the recent precipitous drop?

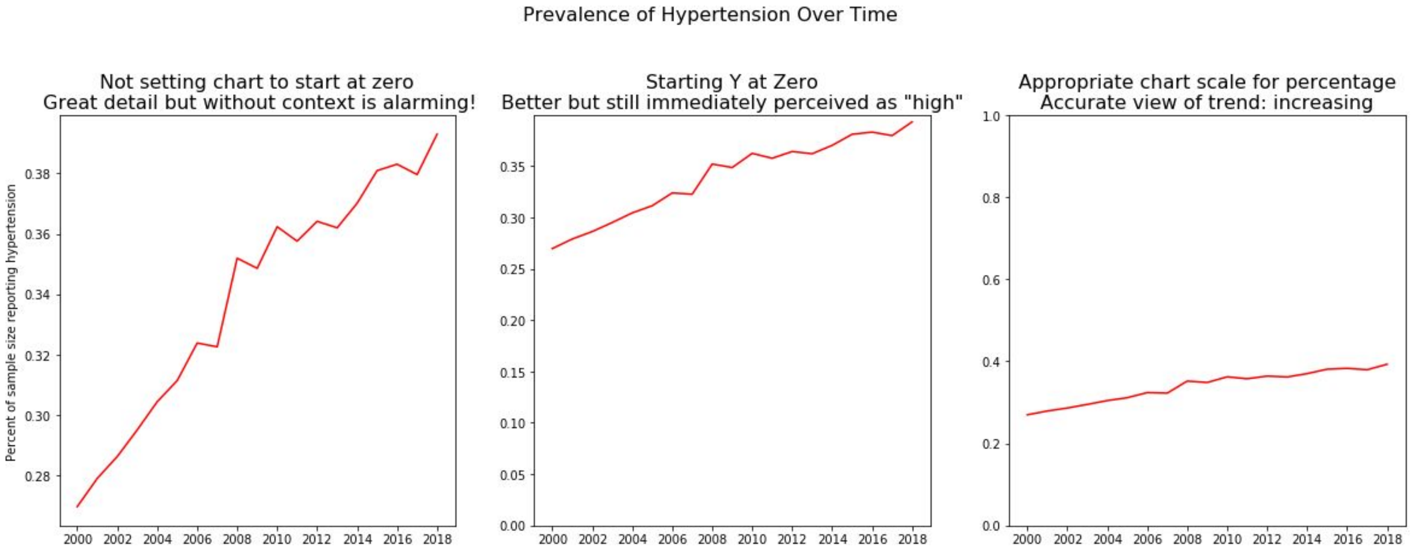
Fortunately, the foundation of data science is good data, so that's the first place to look. As it turns out, budget cuts were causing the effect. For example, there was a deliberate reduction in the number of surveys completed between 2006-2008 in attempt "to achieve cost savings".[3]

To confirm whether variability was also introduced by data wrangling, sample sizes of the raw data were compared to the final dataset prepared for this project. Fortunately the final dataset appears to not only mirror but smooth out the variability in sample sizes. However, this means time series analysis will have to be normalized over sample sizes to produce accurate results.



This chart also revealed that the first three years of data were lost after wrangling, reducing the overall period available for time series analysis.

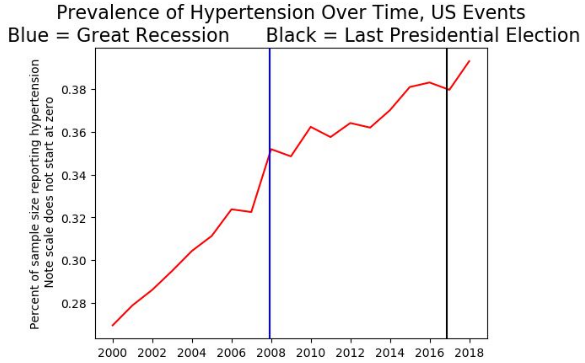
To adjust for variable numbers of surveys conducted each year, positive hypertension responses were plotted as a percentage of each year size, yielding a more accurate visual for the **hypertension trend**. Graphing this data proved to be an excellent example for the importance of proper scaling:

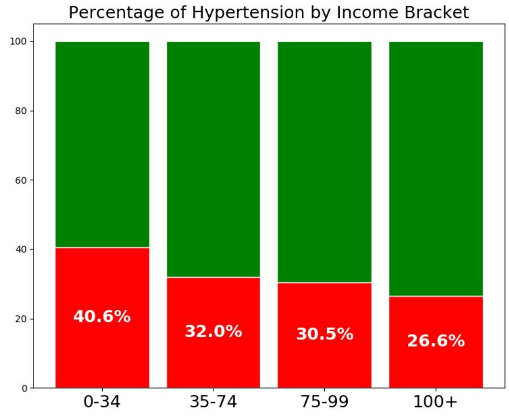


It never hurts to validate results, and a recent study published in the Journal of the American Heart Association confirmed that the absolute burden of hypertension has consistently increased, from 87.0 million in 1999–2000 to 108.2 million in 2015–2016.[4]

This trend answers the question about relevance and shows that hypertension is clearly an ongoing, increasing problem, and predicting hypertension would be a valuable contribution to healthcare.

Viewing the trend line also raises questions about what could be causing the spikes. Two major events occurred around points of increase: the great recession and the most recent presidential election. It would be very interesting to look deeper into the correlation of financial/political events and blood pressure. With the current dataset, we can ask **is there a relationship between hypertension and income?**

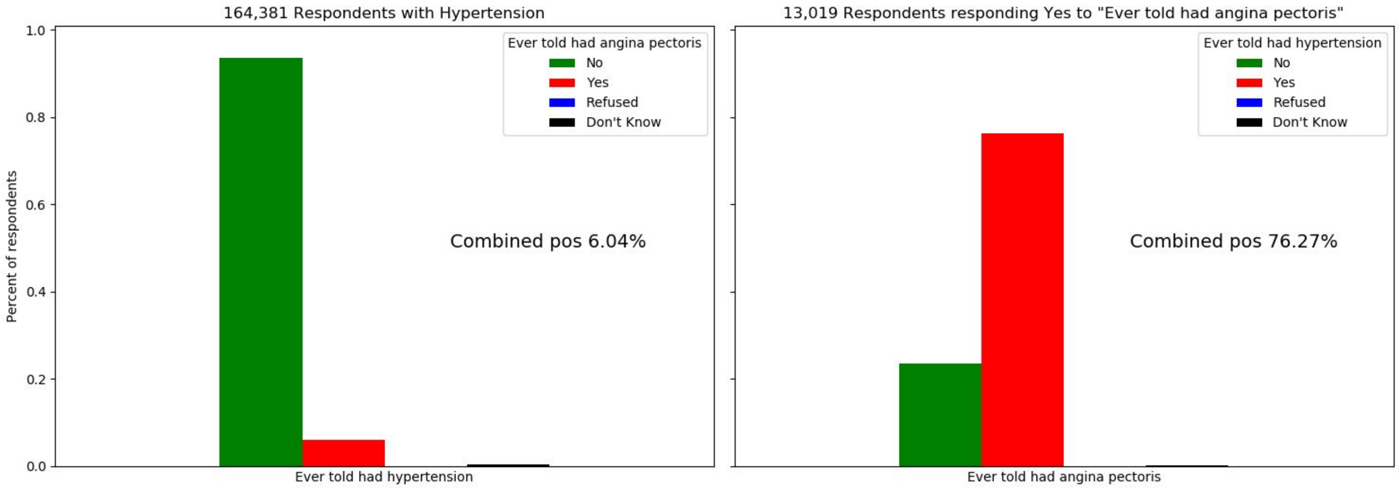




Looking at the percentage of hypertension by income bracket, we can see that there is a decreasing trend in prevalence rates across income brackets.

Stress has not been definitively confirmed to cause hypertension,[5] but has been linked to risk factors for high blood pressure.[6] Stress from financial instability at the low end of the spectrum and the possible compounding of inadequate access to healthcare, nutrition and other socioeconomic factors could increase the vulnerability of this population to hypertension.

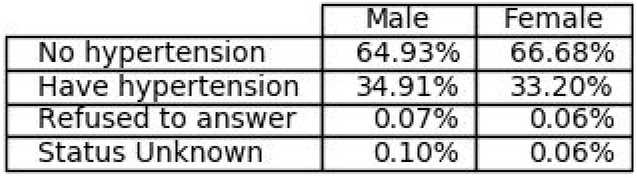
In considering factors that compound hypertension, we can look at the relationship between **hypertension and other chronic conditions**. Here a very interesting phenomenon appeared. When looking at the data from the perspective of respondents with hypertension, there was a low percentage of these respondents with other chronic conditions. However, when looking at the data from the perspective of respondents with a chronic condition, there was a high percentage of these respondents with hypertension. For example, as we see on the left, 6% of people with hypertension said they had experienced chest pain (angina pectoris), but in those with chest pain, 76% had hypertension.



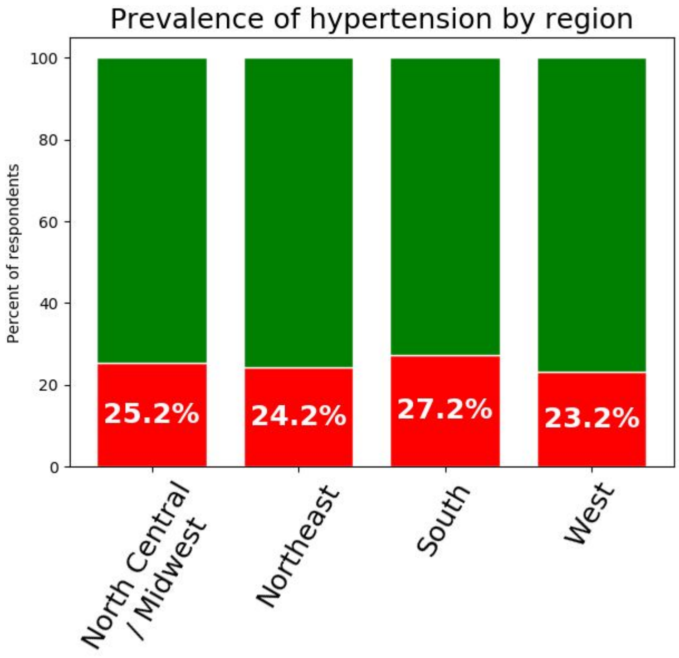
This was also true for cancer, diabetes, emphysema, heart attacks, heart conditions and vision problems. Hopefully this is the result of people being diagnosed and treated for hypertension before it leads to other problems. This would be a great area to explore further.

Given the goal of predicting hypertension we can also ask **do hypertension prevalence rates differ by demographics?** In this dataset, differences were found among gender, region and race features.

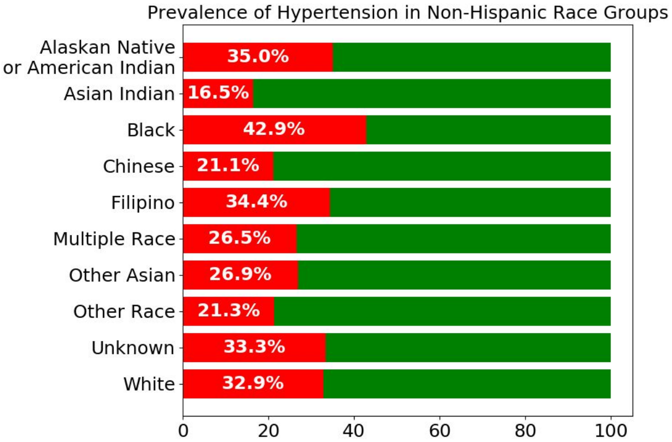
Prevalence rates are higher among men than women, and higher in the southeastern region - not surprising since it was coined the “stroke belt” in the 1940’s and maintains disproportionately high stroke mortality rates compared to the rest of the country.[7]

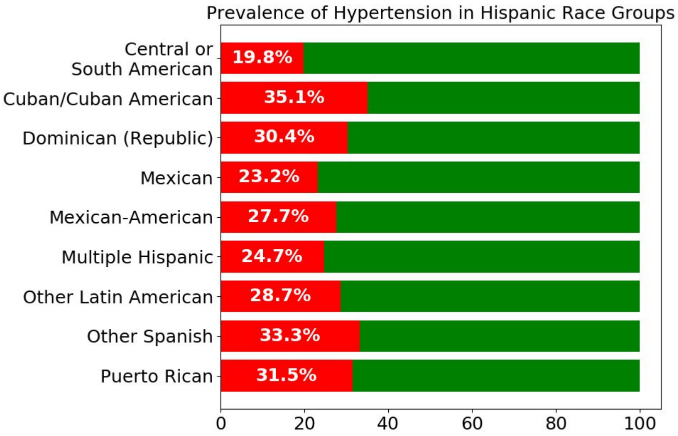


	Male	Female
No hypertension	64.93%	66.68%
Have hypertension	34.91%	33.20%
Refused to answer	0.07%	0.06%
Status Unknown	0.10%	0.06%



Among race groups, people who identified themselves as Black (non-hispanic) and Cuban/Cuban America had the highest prevalence of hypertension.

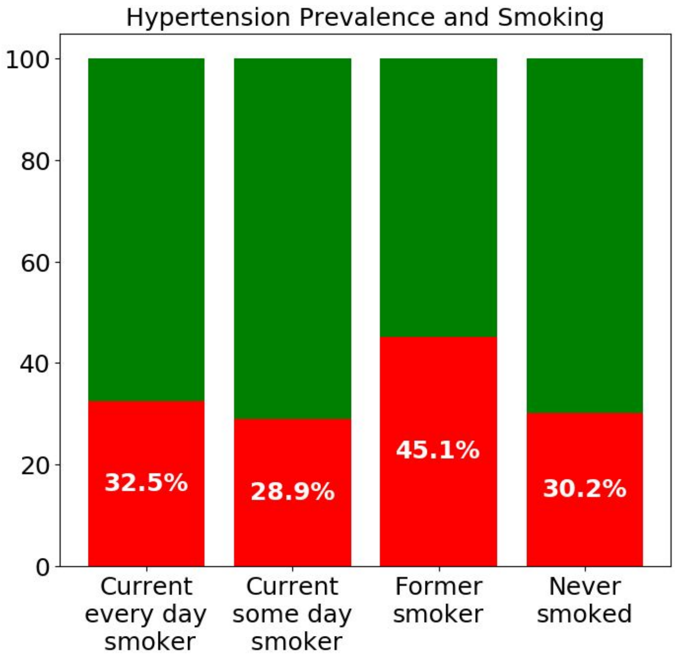




Having reviewed hypertension prevalence trends over time, and across income, health and demographic features, a final area to explore are behavioral features, given that lifestyle choices can increase risk for hypertension[8]. This dataset contained information on smoking, allowing for asking the question **are hypertension prevalence rates different for smokers?** The results showed an interesting finding.

Surprisingly, people who currently smoke some but not all days of the week have lower prevalence rates of hypertension than people who have never smoked. Using data from this same datasource, a Jan 2019 study found that occasional smokers have a 72% higher risk of death from cancer, heart disease and respiratory disease.[8] Also, the 45% prevalence rate of hypertension among former smokers should serve as a warning for smokers - a history of smoking will likely catch up with you.

Other behavioral features such as diet, exercise and alcohol consumption have also been identified as risk factors for hypertension,[8] but unfortunately were not present in this dataset. Including those features from another dataset could help improve prediction accuracy.



This analysis revealed that there are many features in the dataset with relationships to hypertension, and the hypothesis for this capstone project is that these features can be used to predict hypertension. The analysis also revealed that many relationships can be explored further, indicating that other datasets with additional features describing those relationships could be useful for more accurate predictions.

References

1. Zhou D, Xi B, Zhao M, Wang L, Veeranki SP. Uncontrolled hypertension increases risk of all-cause and cardiovascular disease mortality in US adults: the NHANES III Linked Mortality Study. *Sci Rep*. 2018;8(1):9418. Published 2018 Jun 20. doi:10.1038/s41598-018-27377-2
2. Chockalingam A, Campbell NR, Fodor JG. Worldwide epidemic of hypertension. *Can J Cardiol*. 2006;22(7):553–555. doi:10.1016/s0828-282x(06)70275-6
3. 2008 National Health Interview Survey (NHIS) Public Use Data Release. Available here: ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2008/srvydesc.pdf
4. Dorans KS, Mills KT, Liu Y, He J. Trends in Prevalence and Control of Hypertension According to the 2017 American College of Cardiology/American Heart Association (ACC/AHA) Guideline. *J Am Heart Assoc*. 2018;7(11):e008888. Published 2018 Jun 1. doi:10.1161/JAHA.118.008888
5. Mei-Yan Liu, Na Li, William A. Li & Hajra Khan (2017) Association between psychosocial stress and hypertension: a systematic review and meta-analysis, Neurological Research, 39:6, 573-580, DOI: 10.1080/01616412.2017.1317904
6. American Heart Association, “Managing Stress to Control High Blood Pressure”, available here: https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/managing-stress-to-control-high-blood-pressure
7. Karp DN, Wolff CS, Wiebe DJ, Branas CC, Carr BG, Mullen MT. Reassessing the Stroke Belt: Using Small Area Spatial Statistics to Identify Clusters of High Stroke Mortality in the United States. *Stroke*. 2016;47(7):1939–1942. doi:10.1161/STROKEAHA.116.012997
8. CDC, “Behaviors That Increase Risk for High Blood Pressure”, available here:https://www.cdc.gov/bloodpressure/behavior.htm
9. Non-Daily Cigarette Smokers: Mortality Risks in the U.S.Inoue-Choi, Maki et al. American Journal of Preventive Medicine, Volume 56, Issue 1, 27 - 37