Final Project (Spring 2021)

**Dr. Dave Wanik - Statistics in Business Analytics - University of Connecticut**

This is your final project - a group effort that will comprise 40% of your grade. It will be cumulative and include all aspects of the class - enjoy it! Work hard on this so you can show it off on your CV/resume.

****

**Figure: Logo for the PLACES dataset (replacement to 500 cities)**

‘PLACES’ is the new state-of-the-art health dataset funded by the Centers for Disease Control (CDC) and the Robert Wood Johnson Foundation. This dataset only use to exist for the 500 largest cities in the United States, but recently has been expanded to county resolution. The purpose of the dataset is to provide estimates of chronic disease risk factors, health outcomes, and clinical preventive service use in the United States. These estimates will allow cities and local health departments to better understand the burden and geographic distribution of health-related variables in their jurisdictions and assist them in planning public health interventions.

# Background

This is a group project! You are all talented graduate students - please pull your weight and contribute to your team’s success. At the end of the notebook, please specify who contributed (%) to each section.

* **Source for the original data:** <https://www.cdc.gov/places/index.html>
* Link to the clean dataset (cleaned and curated by Dr. Dave for you - you are welcome!): <https://drive.google.com/file/d/1MGKHrtBYTzWsUZQ8lkumayQhpzTohHT_/view?usp=sharing>
* **Data dictionary:** The dataset includes estimates for 27 measures: 5 chronic disease-related unhealthy behaviors, 13 health outcomes, and 9 on use of preventive services. They are listed in this data dictionary: <https://docs.google.com/spreadsheets/d/1mjOOrLq90Uo8b2z9Dhn03912AjsNxbzzJ4Tj-xAzCGk/edit?usp=sharing>

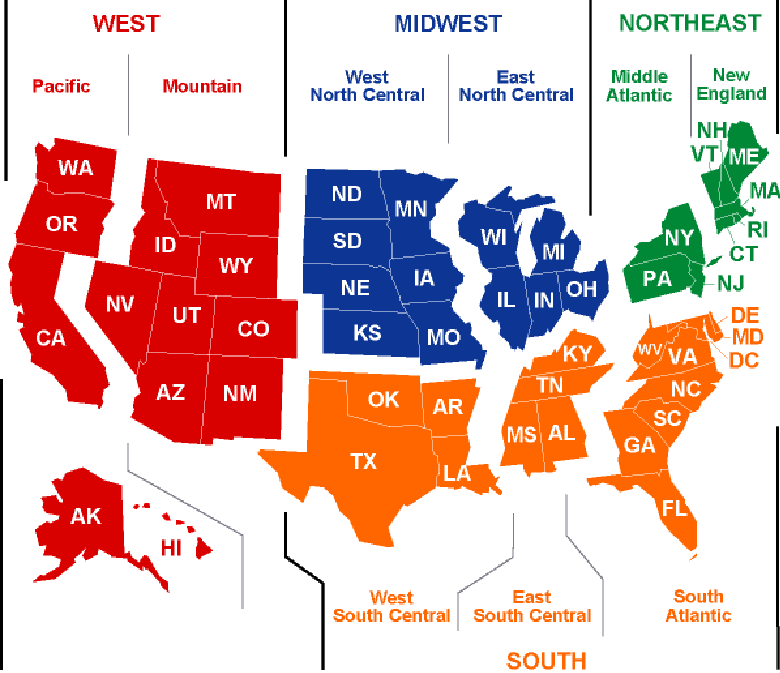
You are going to explore data, aggregate data, run a few hypothesis tests (one for a proportion, one for a mean), and run four different regressions (one full model, one reduced model - for linear and logistic regression). More details below.

# Rubric

## External Data Joining (2.5 pts)

Download and join **this datafile** to your dataset. This is to help you perform your EDA - don’t use these geography variables for modeling, please.

* List of regions in the United States
  + <https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv>



**Figure:** Regions vs. divisions in the United States

## Exploratory Data Analysis (15 pts)

Explore the data by geography!

(5 pts) The PLACES cities data encompasses 1) ‘Use of preventative services’, 2) ‘unhealthy behaviors’ and 3) ‘health outcomes’ and you now have data on regions and divisions. **Make five interesting tables and five interesting plots**. Make sure each plot and table is LABELED and has a descriptive caption.

Using the exploratory data analysis you’ve conducted, **select ONE ‘health outcome’** (like ALL TEETH LOST, ARTHRITIS… or STROKE). You can find a list of health outcomes in the data dictionary: <https://docs.google.com/spreadsheets/d/1mjOOrLq90Uo8b2z9Dhn03912AjsNxbzzJ4Tj-xAzCGk/edit?usp=sharing>

## Literature Review (10 pts)

Go read at least 10 scientific articles, blogs, news stories etc. and say something about the ‘health outcome’ you have selected - give a high-level overview in the form of a literature review. What are the current trends? Is it preventable? Does it affect a certain community/demographic? Is there new medication available? Has the pandemic affected the incidence?

Your literature review should just be a part of your notebook and citations should be referenced in the literature like this [1]. You should have a works cited list at the end of your notebook - needs to have enough info for me to check it out (no specific formatting required - just don’t be sloppy - make sure I can find what you read).

Now that you’ve done your lit review, feel free to go find more datasets if you think it will help your model. You can join on the State or County field.

## Hypothesis Testing (20 pts)

Conceptualize and design a) one sample and b) two-sample hypothesis test. Compute by hand like we do in class (with a graph of the test statistic and critical value on a graph), write LaTeX formulas for the test statistic, calculate the p-value and say what the results of the test mean in plain English. 10 pts for each test.

## Modeling and Error Metrics (40 pts)

10 pts for each model and associated metrics.

I want you to fit two models (linear, logistic) with a full and reduced set of variables. Four models total. I want you to compare the model performance of the full and reduced models for each type. Do not split your data into train and test partitions. Do not use the caret() package for this project.

**Regression**

* Linear - full variables
  + Use all the variables that make sense to predict your health column (don’t use the FIPS column or ID columns!) Just use the health outcomes, prevention and unhealthy behaviors data.
  + Store your predictions in a vector for use later.
  + Calculate AIC, RMSE, MAE, scatterplot of actual vs. predicted with nice titles/labels, and any other metrics that you want to analyze.
* Linear - reduced variables
  + Use stepwise forward, backwards regression or p-hacking.
  + Store predictions in a vector for use later.
  + Calculate AIC, RMSE, MAE, scatterplot of actual vs. predicted, and any other metrics.

**Classification**

Now, create a new column called ‘Flag’ for your ONE health outcome variable. Make it a ‘1’ if the row has a value **GREATER than the median**. Make it a ‘0’ if it is less than or equal to the median value. Now you will fit a LOGISTIC REGRESSION for your health outcome.

* Logistic - full variables
  + Use all the variables that make sense. Don’t use your original health outcome to predict the flag variable - this is data leakage!
  + Store your 0/1 predictions in a vector for use later.
  + Calculate AIC, confusion matrix, accuracy, TPR, TNR, FPR, FNR
* Logistic - reduced variables
  + Use stepwise forward or backwards regression.
  + Store 0/1 predictions in a vector for use later.
  + Calculate AIC, confusion matrix, accuracy, TPR, TNR, FPR, FNR

## Discussion/Analysis (10 pts)

The reason why I wanted you to store your predictions is so you could analyze the data by State, division or region. In addition to the results I had you calculate above, now I want you to be more specific. **Be thorough** and tell me something about how your model does! Is it particularly good or bad in a region? Make at least two plots and two tables that help tell your story.

**Hint:** I love to see boxplots of mean absolute error by region and division with nice labels and titles.

## Conclusion (2.5 pts)

Wrap it up and say something about health in the United States and any next steps you would be interested in pursuing in the future. Be creative and generalize about your conclusions, talk about the value of your analysis. Make sure your ‘works cited’ page is at the end of your document.

Overall, high blood pressure is a disease that has affected millions in the US and worldwide. Some groups, like Hispanic blacks, are more susceptible to the disease than others. However, there are some healthy lifestyle changes that one can practice to reduce the chances of developing high blood pressure. Further, do not rely on a healthy diet and lifestyle changes only, it is important to consult a doctor when required. Consulting a doctor and getting medication at the early stage of Hypertension can reduce the chances of making the situation worst for ourselves. We also concluded from Hypothesis testing that Obesity is equally fatal in Mississippi as compared to High Blood Pressure.

<br>

In the future, it would be interesting to further analyze the differentiators of the Western Region in the US as compared to the rest parts. Is this low statistical data of HBP due to healthy diet choices? Regular exercise? Healthy lifestyle decisions? If students, or researchers, could replicate the exact reason or series of researches for this clear difference in the Western region. This analysis can help us to understand how we can improve the situation of High Blood Pressure in other parts of the United State as well as other countries facing the same issues.

That’s it! Hope you enjoy this project. ***No late submissions.***

After exploring the dataset, we have found some interesting relations between the variables. For doing Hypothesis testing and proceeding further with the project it is important to understand the trends that are associated with our target variable. We are going forward with “High Blood Pressure (HBP)” as our target variable to find relevant findings. According to the American heart association, high blood pressure (also referred to as HBP, or hypertension) is the condition when our blood pressure, the force of blood flowing through our blood vessels, is consistently too high [1]. As per the Harvard health survey, for decades, HBP was defined as 140/90(140 over 90). In November 2017, new United States guidelines lowered the threshold for diagnosing the condition. According to new guidelines, anyone with a reading of 130/80 or higher has blood pressure. Based on this new definition, nearly half of Americans now fall into this group [2]. Blood pressure is now categorized as Normal for less than 120/80, Elevated for 120/80 to 129/79, Stage 1 of Hypertension for 130/80 to 139/89, and Stage 2 of Hypertension for more than 140/90 mm Hg. There is one more category for blood pressure higher than 180 as Hypertensive crisis, it is suggested to immediately consider a doctor if someone is having this higher blood pressure [1].

Having hypertension can put an individual at risk for heart disease and stroke, which are leading causes of death in the United States. In 2018, nearly half a million deaths in the United States included hypertension as a primary or contributing cause [3]. Nearly half of adults in the United States (108 million, or 45%) have hypertension defined as an HBP ≥ 130 are taking medication for hypertension. Further, the facts concluded that about 1 in 4 adults (24%) with hypertension in the US have their condition under control. About half of adults (45%) with uncontrolled hypertension have a blood pressure of 140/90 or higher. This includes 37 million U.S. adults.

High blood pressure was a primary cause of death for more than 494,873 people in the United States in 2018 [3]. High blood pressure costs the United States about $131 billion each year, averaged over 12 years from 2003 to 2014.

Despite being one of the most common and widespread diseases in the United States, hypertension is also one of the most preventable. A person's risk of having high blood pressure over their lifespan is influenced by a variety of factors. Our age, a family history of hypertension, and ethnicity are among the hypertension risk factors that are out of our control. When it comes to [preventing high blood pressure](https://www.everydayhealth.com/high-blood-pressure/guide/), the idea is to focus on the risk factors that we can change [5].

To avoid a hypertension diagnosis, one should make the following healthy lifestyle choices:

**Maintain a healthy weight.**

**Eat a balanced diet**

**Cut back on salt.**

**Exercise regularly.**

**Limit alcohol.**

**Manage stress.**

**Monitor your blood pressure.**

Adopting these lifestyle changes can help us to prevent high blood pressure. Making an effort to prevent high blood pressure, will benefit to reduce the risk of heart attack, stroke, and other serious illnesses. One of the major concerns with Hypertension is the formal behavior of an individual towards HBP, if a person is suffering and he/she is not taking the doctor’s advice or following the suggested lifestyle changes, the damages could be life-threatening. With proper treatment and management, one can control blood pressure to live a long and healthy life. The only step we should take is to follow a healthy lifestyle and consulting a doctor at the right time.

It may seem insignificant, but certain individuals are more likely to have a high blood pressure problem than others due to their family background or demographic group.

The Data from the National Health and Nutrition Examination Survey suggests that the prevalence of hypertension was higher among non-Hispanic blacks (57.1%) than among non-Hispanic whites (43.6%) and Hispanics (43.7%) [7]. Among men, the age-adjusted prevalence of hypertension was higher among non-Hispanic blacks (57.2%) than non-Hispanic whites (50.2%) and Hispanic (50.1%) adults. Among women, the age-adjusted prevalence of hypertension was higher among non-Hispanic blacks (56.7%) than non-Hispanic whites (36.7%) and Hispanic (36.8%) adults. This study results can be seen from the graphical description below:

Alone this study was not able to show the proper explanations on why the prevalence of hypertension was higher among non-Hispanic blacks. The research from Daniel T. Lackland addresses certain causes that may be correlated with this racial imbalance, such as salt exposure and obesity; this means that African Americans have a much higher susceptibility to salt consumption (which has a greater impact on Hypertension) and suffer from obesity more than other races, all of which affect blood pressure [8].

We can easily differentiate the level of severity in different states of the United States from the data available on some of the government websites. As per the data of 2019 from the website of “State of childhood obesity”, the most recent Behavioral Risk Factor Surveillance System (BRFSS) finds that nine of the 10 states with the highest rates of hypertension in the United States are in the South. West Virginia has the highest rate at 43.8%, Mississippi being the 2nd state with 43.6%, and Alabama is the 3rd state with 42.5% of people facing HBP [9].

Medication:

Several types of research have been done on how to mitigate the risk and effect of High blood pressure on an individual. Fewer than half of those people with high blood pressure have it under control. But when the condition is detected early and treated properly, the outlook is good.

Sometimes high blood pressure can be treated solely through lifestyle changes, which are the first line of defense. In other cases, treatment requires both a healthy lifestyle and medications[10]. Some supplements may help to lower blood pressure, however, more research is needed to determine the potential benefits. These supplements include:

Fiber, such as blond psyllium and wheat bran

Minerals, such as magnesium, calcium and potassium

Folic acid

Supplements or products that increase nitric oxide or widen blood vessels (vasodilators), such as cocoa, coenzyme Q10, L-arginine and garlic

Omega-3 fatty acids, found in fatty fish, high-dose fish oil supplements and flaxseed

Researchers are also studying whether vitamin D can reduce blood pressure[11].

In a recent clinical trial led by researchers from the George Institute for Global Health with branches in the United Kingdom, Australia, and India, they tested the effectiveness and safety of using an innovative drug for the treatment of high blood pressure or hypertension. This drug, nicknamed the “triple pill” by the investigators, combines low doses of three existing drugs for blood pressure [11]. Namely, these are telmisartan (20 milligrams), amlodipine (2.5 milligrams), and chlorthalidone (12.5 milligrams). This new combination drug may have the potential to revolutionize hypertension treatment worldwide, the 1st clinical trial has declared it safe to use and very effective [11].

We are currently in the middle of a pandemic due to the spread of the new coronavirus, researchers are working to learn more about health conditions that may put us at risk for serious illness. One of the conditions being investigated is high blood pressure [12]. A recent study investigated more than 2,800 hospitalized individuals with confirmed COVID-19 in China. Investigators made the following observations related to high blood pressure:

Out of all study participants, 29.5 percent had high blood pressure. Of those with high blood pressure, 83.5 percent were taking medications to manage their condition.

There was a twofold increase in the risk of death due to COVID-19 in people with high blood pressure when compared to those without high blood pressure.

Those with high blood pressure who weren’t taking medications to manage their condition were at a greater risk for death compared to those who took blood pressure medications.

After this research CDC has recently updated its list of factors that put an individual at increased risk for serious illness due to COVID-19 and Hypertension is one among these factors.

Analysis of early data from different countries shows that high blood pressure is the most commonly shared pre-existing condition among those hospitalized, affecting between 30% to 50% of the patients. Other health conditions included cancer, diabetes, or lung disease. A report said that more than 99% of people who had died from the virus had one of these conditions -- and 76% of them had high blood pressure [13]. Other research shows that people with high blood pressure are also slightly more likely to die from coronavirus. Their risk is about twice as high as that of the overall population.

Overall, high blood pressure or hypertension is a huge illness that affects millions of people in the United States and other countries around the world. Some people and states in the US itself are more vulnerable than others, but there are also appropriate approaches to treat the condition. A healthy lifestyle and medication can lower the risk of facing high blood pressure. Due to the pandemic, the situation has gotten worst and it would be interesting to learn more about the impact of HBP and Covid-19 on an individual.

# References

Association, A. H. (Nov 30, 2017). The Facts About High Blood Pressure. From https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure

CDC, A. H. (n.d.). "High blood pressure is a “silent killer”. From https://www.heart.org/en/health-topics/high-blood-pressure/why-high-blood-pressure-is-a-silent-killer

Centers for Disease Control and Prevention, U. D. (n.d.). "Facts about Hypertension in United States". From https://www.cdc.gov/bloodpressure/facts.htm

Harvard Health Survey. (October, 2020). High Blood Pressure (Hypertension). From https://www.health.harvard.edu/a\_to\_z/high-blood-pressure-hypertension-a-to-z

McCoy, K. (n.d.). "7 Ways to prevent Hypertension". *Everyday Health Group*. From https://www.everydayhealth.com/hypertension/preventing.aspx

Shepard, B. (September 09, 2020 ). *"Study shows high blood pressure awareness and control are declining in America".* From https://www.uab.edu/news/research/item/11543-study-shows-high-blood-pressure-awareness-and-control-are-declining-in-america#:~:text=Among%20U.S.%20adults%20who%20were,88%20percent%20in%202017%2D2018.

Yechiam Ostchega, P. R. (April 2020). Hypertension Prevalence Among Adults Aged 18 and Over: United States, 2017–2018. From https://www.cdc.gov/nchs/products/databriefs/db364.htm