Herman Huff

Springboard Capstone Project I

Capstone Project I – Milestone Report

Introduction

Problem Statement: Predicting heart disease prevalence in a US zip code using physiological and socioeconomic features.

Background

Heart disease is the leading cause of death in the United States, causing more than 600,000 deaths annually. This project will examine various risk factors for heart disease and seek to determine which combination of factors optimally predict heart disease prevalence per zip code. The project will focus on zip codes within the largest 500 cities in the US, as health data is available for these zip codes through the CDC's 500 Cities project. Additionally, socioeconomic and fast food restaurant datasets will be incorporated.

Impact

This model can help to increase preventative and interventional efforts by identifying new socioeconomic features of heart disease and quantifying the impact of previously known socioeconomic features. This is beneficial to legislature and organizations seeking to decrease the prevalence of heart disease by giving them new targets to work on and helping to quantify their efforts.

Datasets

a. 500 Cities: Local Data for Better Health

This dataset comes from the CDC's collaborative project that provides small area estimates for disease risk factors, outcomes and preventative service use. The project delivers data for the 497 largest cities in America and represents approximately one-hundred million people, approximately 33% of the US population. There are 810103 entries, each representing a measure of health data for a census tract. There are 24 columns specifying features of these measures. Some useful features include geographic area information and population counts.

b. HUD-USPS ZIP Crosswalk Files

This is a dataset maintained by the US Department of Housing and Development. It contains census tracts and their corresponding zip codes. It is created using the 2010 census data along with USPS Vacancy Data, which is updated quarterly. It should be noted that some census tracts are within multiple zip codes and this will be addressed during data wrangling.

c. fastfoodmaps.com

This is a dataset that provides locations of fast food restaurants all over the US. It was built in 2007 and authored by Phil Dhingra, using website scrapers to extract information. It provides geographic information for 50,000 fast food restaurants. We make the assumption that the number and location of these restaurants has not changed enough over the years to significantly impact the results.

d. Uszipcode

This is a zip code database authored by Sanhe Hu. It provides a rich set of data for US zip codes including geographic, demographic, real estate, socioeconomic and education information. Scrapers are used to retrieve from multiple active data sources to maintain this database.

Data Wrangling

500 Cities Dataset

The 500 cities dataset is loaded into a Panda's dataframe to reveal 810103 entries and 24 columns. Many of the columns contain redundant information: 'StateAbbr (two letter state abbreviation) and 'StateDesc' (full state name), 'Category' (health category abbreviation) and 'Short_Question_Text' (health category unabbreviated). Additionally, the columns 'DataSource', 'UniqueID', 'DataValueTypeID', and 'Data_Value_Footnote_Symbol' contain information that won't be useful for our analysis. These columns are dropped. Finally, some rows do not contain our metric of interest (heart disease prevalence) due to population size being too small, so these rows are dropped along with the 'Data_Value_Footnote' column which contains a note mentioning this.

Each row contains an estimate for some health metric at the city or census tract level. For our analysis, we are interested in census tract rows, so the 'GeographicLevel' column is filtered to show only rows at the census tract level.

We filter to show only rows related to our metric of interest. To do this, the 'Measure' column is filtered to contain only heart disease prevalence and the prevalences of risk factors for heart disease. These include hypertension, high cholesterol, smoking, diabetes, sedentarism and obesity.

HUD-USPS ZIP Crosswalk Files

The columns 'zip' (containing the zip code), 'tract' (containing the census tract) and 'res_ratio' (containing the percentage of census tract residents who reside in this zip code) are selected and stored. The resulting dataframe is grouped by tract and the row with the maximum value of 'res_ratio' is selected for each tract. This is then merged back in with the original crosswalk df so that only one zip code, where most of the tract's population resides, is represented for each tract.

Merging 500 Cities and Zip Crosswalk

The crosswalk dataframe is merged with each of the health prevalences individually. A weighted mean prevalence is obtained by grouping by zip code and weighting each tract's prevalence by its population. This is applied to heart disease and all its risk factors individually. Finally, all risk factor prevalences are merged into a single dataframe along with heart disease prevalence.

Fastfoodmaps.com

This dataset contains a row for each fast food restaurant and 8 unnamed columns containing restaurant name, address, city, state, zip code, phone number, longitude, latitude. The columns are renamed. A new column is created which fills in the number one for each row. This will be used to sum a count of restaurants for each zip code. The zip code column contains entries with trailing digits beyond the 5th digit. These trailing digits are removed. The dataframe is grouped by zip code and a sum of restaurants is calculated. Finally, this dataframe is merged with the working health dataframe, and zeroes are filled in for zip codes that are missing restaurant counts. Here we assumed that these zip codes have no fast food restaurants.

Uszipcode

This library is imported and a search engine to retrieve all of its rows (one for each zip code) is created and run. The resulting list is transformed to a dictionary and used to create a dataframe. The 'zip code', 'median_household_income', and 'population' columns are selected and stored. The zip code column is changed from object to integer so it may be merged. This dataframe is merged with the working health dataframe. Rows missing values for income or population are dropped.

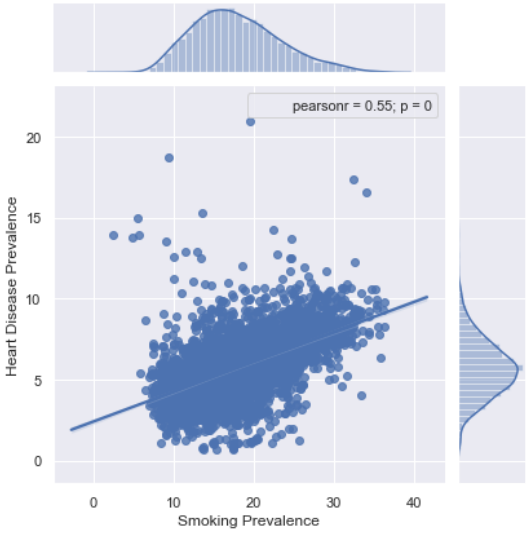
Data Story

Features with positive correlations

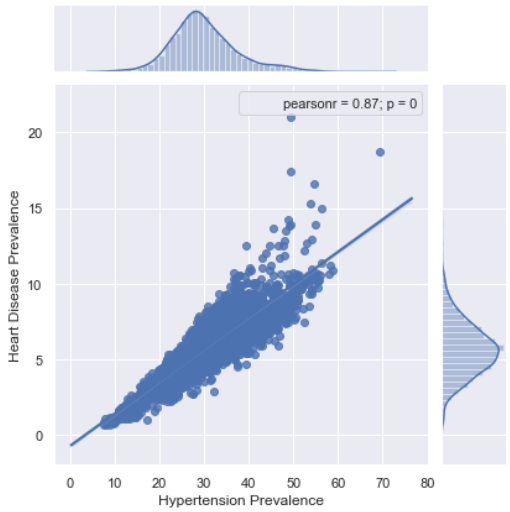
Do the data show established relationships between heart disease and its risk factors?

Below are joint plots of heart disease prevalence and positively correlated features: smoking, hypertension, obesity, sedentarism, high cholesterol, and diabetes.

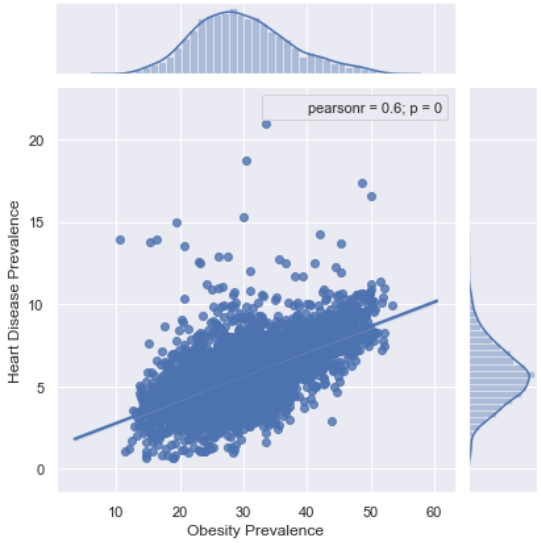
Smoking and heart disease



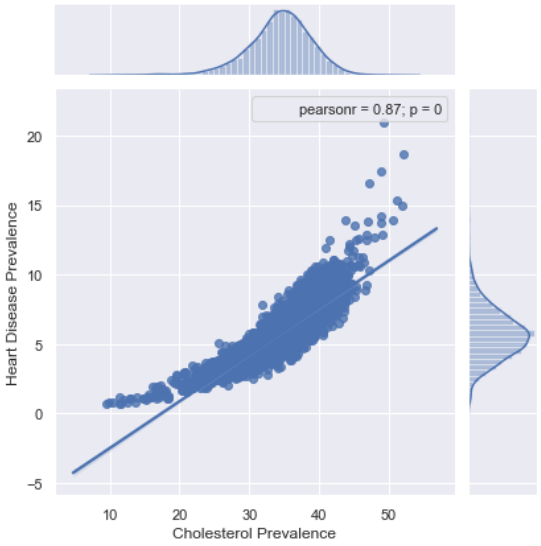
Hypertension and heart disease



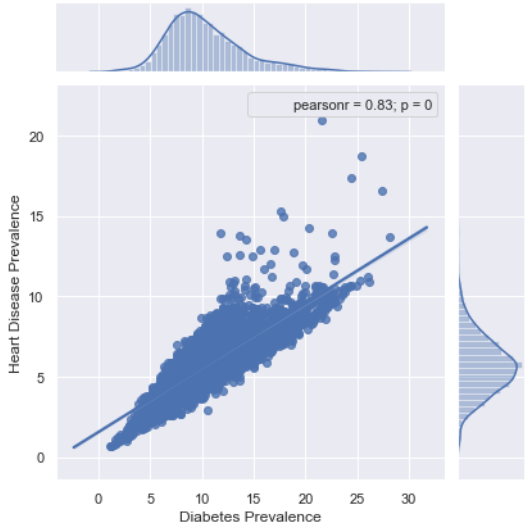
Obesity and heart disease



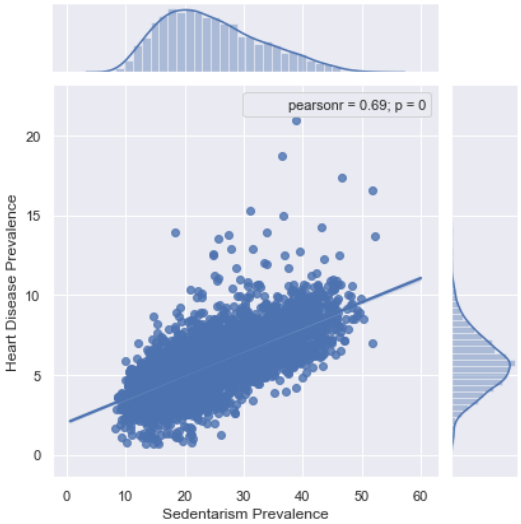
High cholesterol and heart disease



Diabetes and heart disease



Sedentarism and heart disease

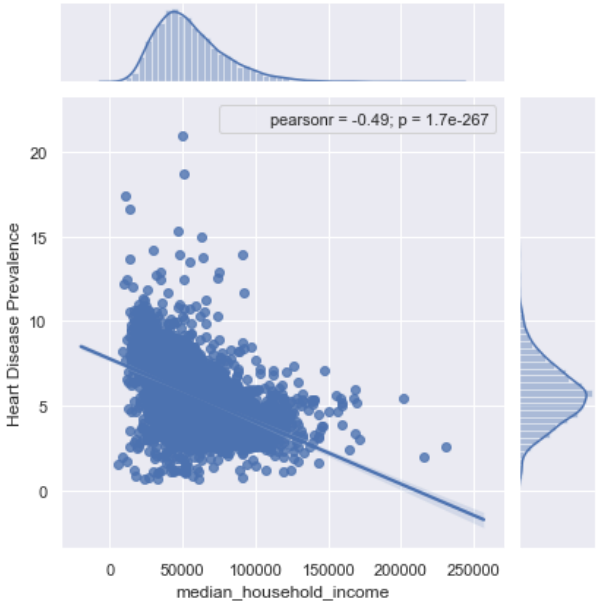


From the joint plots it can be seen that these risk factors are strongly correlated with heart disease prevalence. Hypertension, high cholesterol and diabetes have the strongest correlations.

These are known major risk factors for heart disease so strong correlations are expected.

Heart disease and median household income

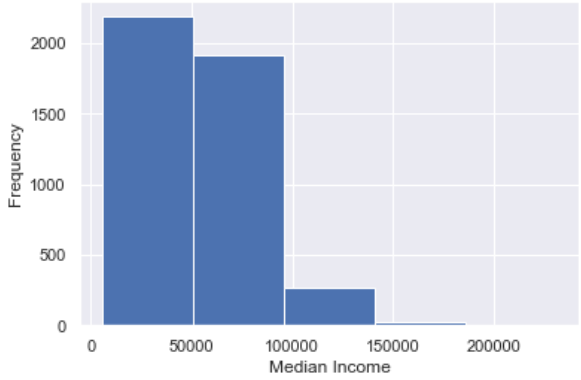
Does median household income, a factor of socioeconomic status, show a relationship to heart disease?¶



We see there is a moderate to strong negative correlation between median household income and heart disease prevalence.

One hypothesis is residents of zip codes with an average larger income likely have better access to healthcare, exercise and food options.

How is income distributed?



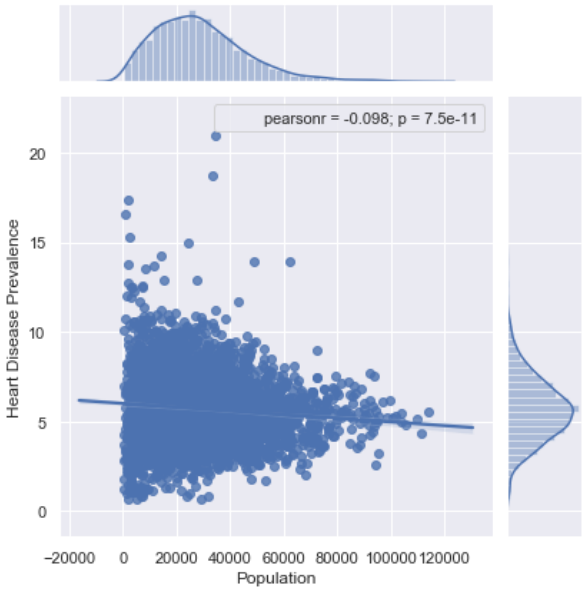
We see that most data points fall above and below 50000. We will separate the data into groups based on this cutoff.

	Income Group	Heart Disease Prevalence
0	<50000	6.540580
1	>=50000	4.959305

We see that the prevalence of heart disease is larger when income is <50000. We will statistically test the significance of the difference between groups.

Heart disease and zip code population

Is there a correlation between zip code population and heart disease?



We see there is a weak negative correlation between zip code population and heart disease prevalence.

This may due to lower incomes in less populated zip codes.

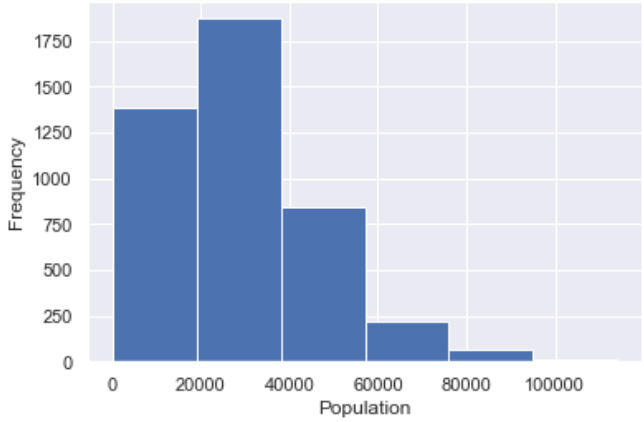
Is the difference in heart disease by population due to income?

Income Group: <50000
---
pearson r and p-value: (-0.16169560730753946, 9.391991673286615e-14)

Income Group: >=50000
---
pearson r and p-value: (-0.04340847959287403, 0.03741615575840685)

We see that when income group is accounted for, there is still a significant negative correlation between population and heart disease prevalence.

How are populations distributed?



We see that most of the data points represent zip codes with populations less than 60000 with a mean of about 30000.

We will refine population groups to <30000 and >=30000 for comparison of low population and high population zip codes.

Does the mean prevalence of heart disease differ when examined by population group?

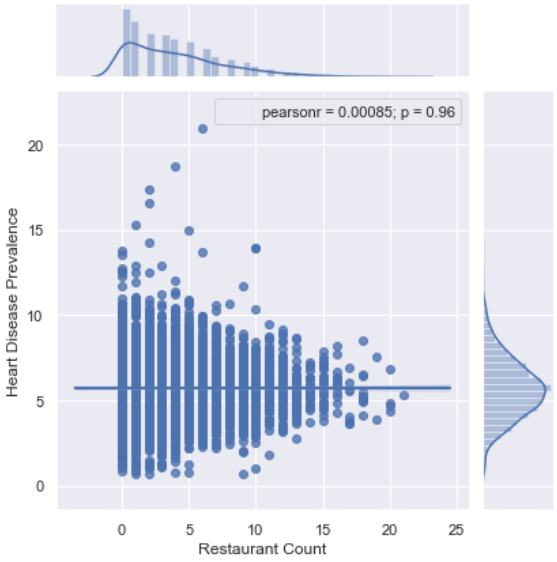
	Population Group	Heart Disease Prevalence
0	<30000	5.843411
1	>=30000	5.535480

We see that there is a difference in the prevalence of heart disease between the two groups.

We can perform statistical testing to examine whether this difference is due to chance.

Heart disease and restaurant count

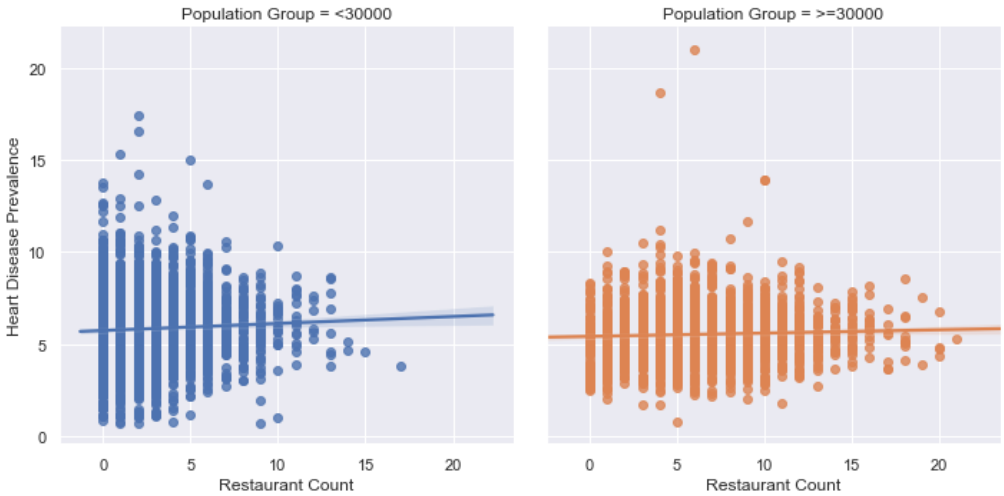
Is the number of fast food restaurants within a zip code correlated to heart disease?



From the plot we see that when viewing all zip codes, there is likely no correlation between a zip code's restaurant count and its heart disease prevalence.

Is the number of fast food restaurants correlated to heart disease when examined by population size?

In the following plots we break down the zip codes by population size and plot restuarant count against heart disease prevalence.



Population Group: <30000
---
pearson r and p-value: (0.050389916369143375, 0.011039806489306121)

Population Group: >=30000
---
pearson r and p-value: (0.045230669577913954, 0.051570248314069915)

We see in populations < 30000 there is a very weak positive correlation.

There is likely no correlation between restaurant count and heart disease in populations >= 30000.

Inferential Statistics

Income groups and heart disease

	Income Group	Heart Disease Prevalence
0	<50000	6.540580
1	>=50000	4.959305

We saw that the mean prevalence of heart disease is higher in zip codes with a median household income below 50000.

Is the observed difference in the samples due to chance?

- **Null Hypothesis**: There is no difference in the mean heart disease prevalence between groups.
- **Alternative Hypothesis**: The mean heart disease prevalence is higher when median household income is <50000.

alpha = 0.05

To test the hypotheses we draw bootstrap samples of both groups and compare the bootstrap difference in means to the observed difference in means.

```
p-value = 0.0
```

A p-value close to 0 indicates we reject the null hypothesis. There is a difference in the means of both groups. The alternative hypothesis suggests that the mean heart disease prevalence is higher when median household income is below 50000 in a zip code.

Population groups and heart disease

	Population Group	Heart Disease Prevalence
0	<30000	5.843411
1	>=30000	5.535480

We saw that the mean prevalence of heart disease is higher in zip codes with a population below 10000.

Is the observed difference in the samples due to chance?

- **Null Hypothesis**: There is no difference in the mean heart disease prevalence between groups.
- **Alternative Hypothesis**: The mean heart disease prevalence is higher when population <10000.

alpha = 0.05

To test the hypotheses we draw bootstrap samples of both groups and compare the bootstrap difference in means to the observed difference in means.

```
p-value = 0.0
```

A p-value close to 0 indicates we reject the null hypothesis. There is a difference in the means of both groups. The alternative hypothesis suggests that the mean heart disease prevalence is higher when zip code population is below 30000.

Summary

The following features have moderate to strong correlations with heart disease prevalence and will be useful for our model:

1. High cholesterol prevalence
2. Hypertension prevalence
3. Diabetes prevalence
4. Sedentarism prevalence
5. Obesity prevalence
6. Smoking prevalence
7. Median household income
8. Population

Restaurant count per zip code may not be helpful for our model. It has a very weak correlation with heart disease prevalence in populations < 30000.