Washington D.C. Demographic Trends

# Introduction

## Background

Washington D.C has changed dramatically over the past 40 years. Neighborhoods that were once considered rough places to live have turned into high demand neighborhoods due to an improving economy, driven by both the federal and private sector jobs. Understanding how demographics and social economic factors have changed over the past few years can help determine how which neighborhoods of Washington DC are best.

## Problem

Data that may contribute to determining the best neighborhoods of Washington DC could be population change over time, average property value, crime rate, and average household income. It would also be interesting out the relationship between these data points and ethnicity.

## Interest

This data would be useful to political campaigns, people trying to determine where to open a business, potential home buyers, real-estate investors, and city planners.

# Data acquisition and cleaning

## Data source

The primary dataset I will use will be from http://data.codefordc.org containing demographic information by neighborhood from 1980 to 2017. The dataset contains several socio-economic measures clustered by DC neighborhood. The dataset includes latitude and longitude values for each neighborhood, making it easy to plot data points by geographic location. Some of the demographic measures include:

% low weight births -1998 – 2011

% births to teen mothers 1998 – 2011

Poverty rate by decade

Unemployment Rate

Median Family Income

% persons receive food stamps

Violent Crimes per 1000 people

Property Crimes per 1000 people

Sale of single family homes

Median home sale price of single family homes

% subprime loans

Home foreclosure notice rate.

## 2.2 Data Cleansing and Feature Selection

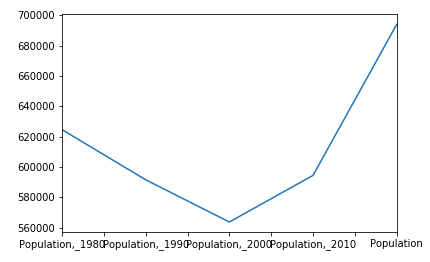
The data set contains 39 rows (including headers) and 416 columns. The data is organized pretty well, so the data downloaded into a single dataframe. The neighborhood name was set as the index. Because the dataset contains over 400 dimensions, I chose subsets of the larger dataframe to create visualizations and do analysis. For some subsets of the data, I added either a total or average row to determine change in metric over time. I did not drop any values from the master table.

# Exploratory Data Analysis

## Population Change by Neighborhood 1980 – 2017

To give the data some context, it is good to explore how total population has changed in Washington D.C since 1980. Population in the district declined 11% between 1980 and 2000. From 2000 to 2010 the population increased 5% then an estimated increase of 93% from 2010 to 2017. The drastic increase in population was due to an increase in jobs in both the federal and private sector in the area.

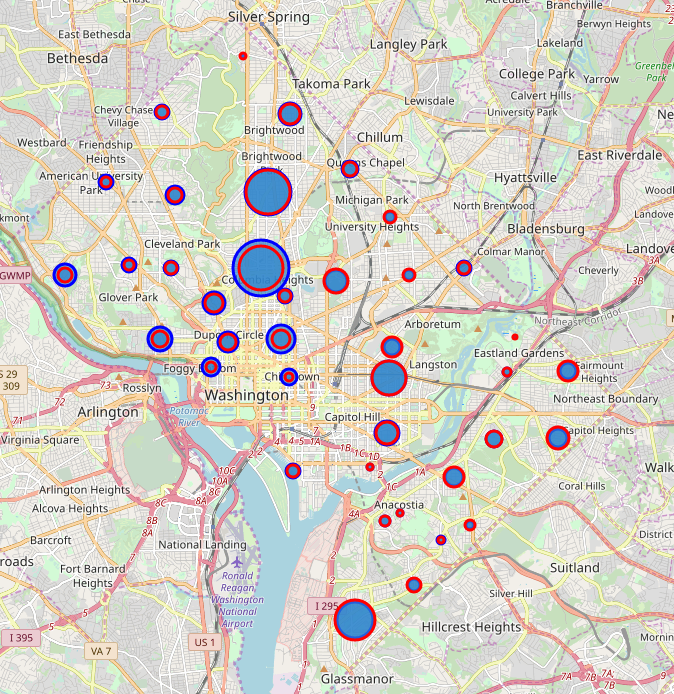
**Washington DC Population Change Since 1980**



### 2017 estimated population change by Neighborhood

From 1980 to 2010, the greatest increase in % population change was the Downtown neighborhood, with a 78% increase in population. The greatest decrease in population was Capitol View, with a 32% decrease in population. See the Jupyter Notebook for details.

### Population Change by Neighborhood: 1980 - 2010



1980

2010

## Violent Crimes Trended

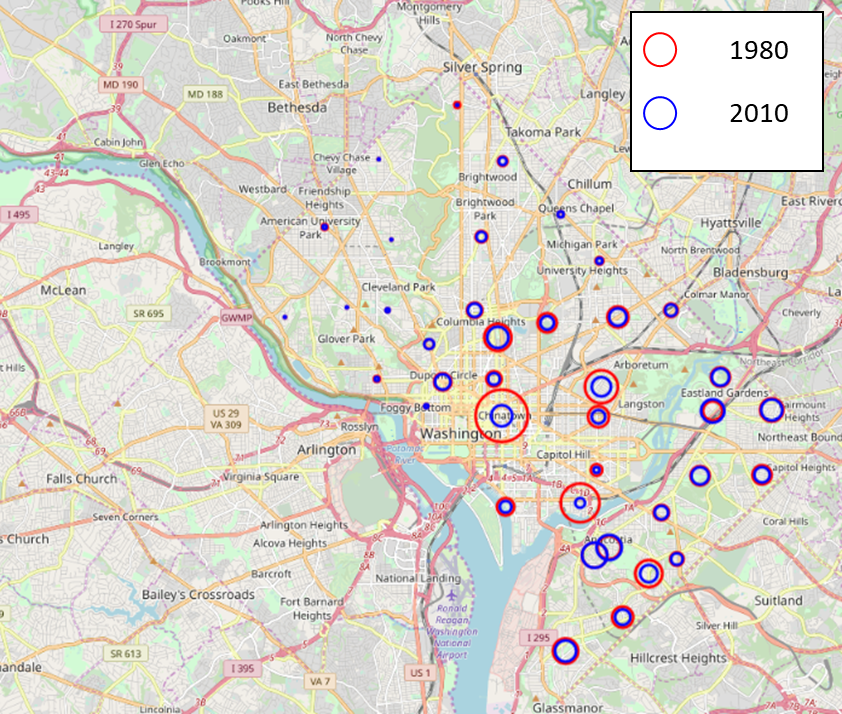
Part of the reason population has increased in Washington DC since 2000 is because the crime rate has steadily declined. After September 11th, 2001, police presence in DC increased, contributing the decrease in violent crime.

## Violent Crime Rate per 1000 People – 2000 to 2010

## 

Crime dropped most dramatically in the China Town and Navy Yard neighborhoods. Those neighborhoods are home to the Washington Wizards/Capitals and the Washington Nationals respectively. Both neighborhoods have had stadiums built and an effort has been made to clean up the areas.

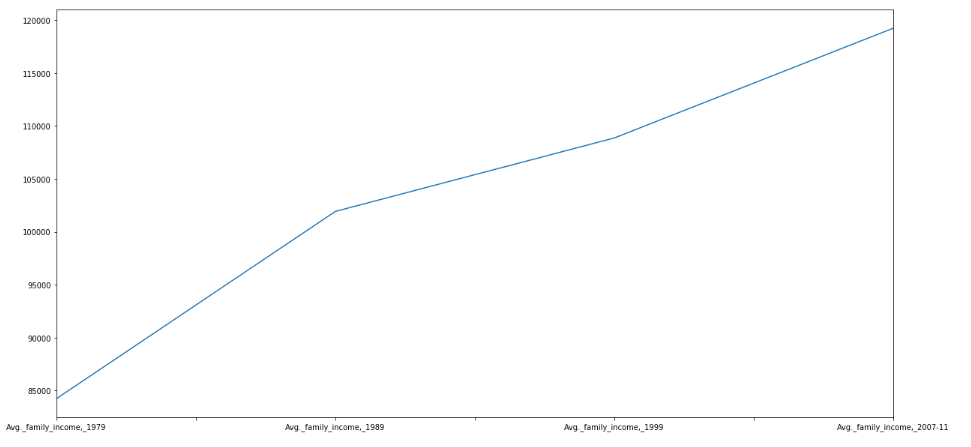
## Violent Crime Rate by Neighborhood – 2000 vs 2010



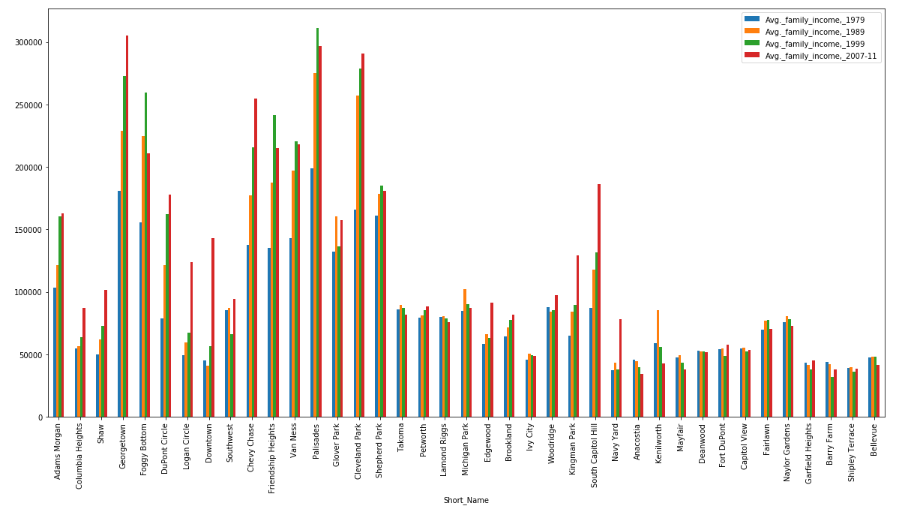
## Median Household Income by Neighborhood

Average family income in Washington DC has increased by 29% from 1979 to 2011 from $84,230 in 1979 to $119,230 in 2011. The greatest increase in that time was the Downtown neighborhood at 219% and the greatest decrease was the Kenilworth neighborhood at -27%.

## Median Household Homcome – 1979 - 2011



## Median Household Income by Neighborhood



# 4. Machine Learning

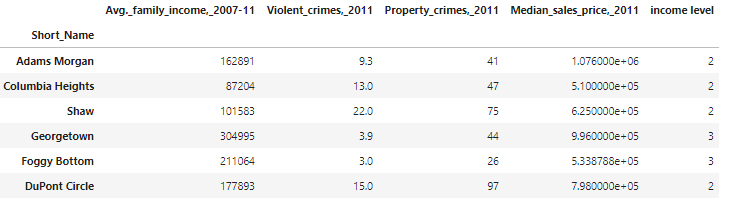
## Using KNN to predict household income level based on Crime rate and Home Sale Price

Based on some of the exploratory data, we realize that there is a relationship between crime levels, home sale price, and household income level. Therefore, we would like to use violent crime, property crime, and home sale price to predict what median income level each neighborhood.

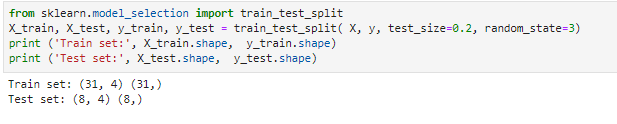
To set up the data, we pulled 2011 property crime, 2011 violent crime, 2011 home sale price, and 2011 median household income into a dataframe. I created a target field called **income level** based on the 2011 median household income. The target field categorizes median income into three levels:

|  |  |
| --- | --- |
| Median income | Income Level |
| $200000 and above | 3 |
| $60001 to $199999 | 2 |
| $60000 and below | 1 |

Here is what the data set looks like after adding the classifier:



After the dataframe is created, we normalize the data, the train, test and split the dataset to improve out-of-sample accuracy.



### KNN Classification

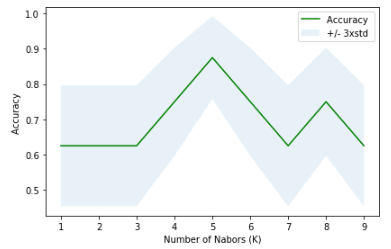
For our first training algorithm we set k to 4 which resulted in the following predicted test array and accuracy level:





After the initial results we would like to find which k is the most accurate. To do so, we loop down from K = 10 and plot the results:

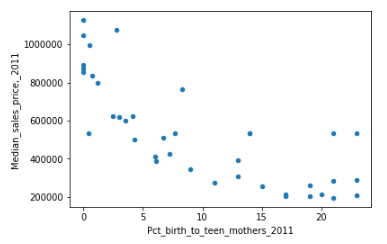






## Using Multiple Linear Regression to predict Home Prices

There are several factors plotted in this dataset that relate to home prices that may not be completely obvious. For example, there is a slight negative correlation between median home sale price and percent birth to teen mothers:



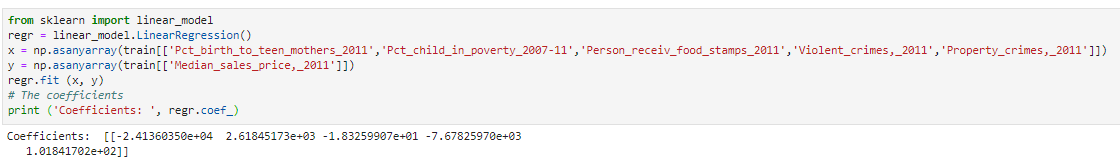
After exploring the dataset, we would like to use percent birth to teen mothers, percent children in poverty, person receiving food stamps, violent crimes, and property crimes, to predict median sales price of a single family home.

First get a random sample of the data set and calculate the coefficients of the parameters used to predict the target variable

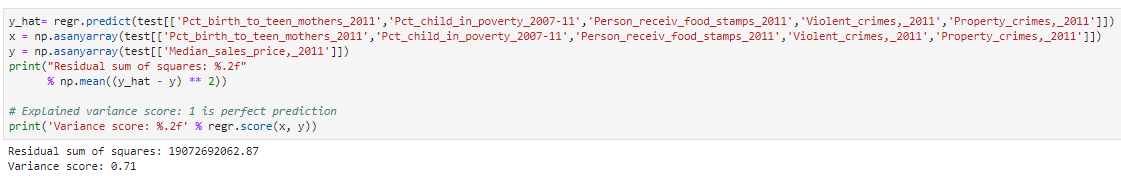
Generating Random Sample:



Calculating Coefficients:



After calculating the coefficients, we can use ordinary least squares to calculate the residual sum of squares and variance score:



# Conclusion

In this study, I explored how different neighborhoods in Washington DC changed over time and the relationships between different demographic metrics. I was able to identify which neighborhoods changed the most in terms of population, income level and home value. I performed a KNN clustering model that was able to accurately predict median income level based on other environmental factors and I was also able to use multiple linear regression to predict home value.

The master dataset has several parameters that were not used in this report that could potentially be used to show how metrics are changing over time, or the relationship between two metrics. For example if teen birth rates, correlate with violent crime, social workers could use that data help neighborhood conditions.