Analysis of the Correlation Between Crime Rate and Housing Price in Washington, D.C., USA Based on Big Data

Hanshu Yang[1], Meili Liu[2], Jeng-Eng Lin[3], and Chun-Te Lee[4](✉)

[1] School of Mathematical Sciences, College of Science and Technology,
Wenzhou-Kean University, Wenzhou, China
[2] Institute of Artificial Intelligence, Midea Group, Shenzhen, China
[3] Department of Mathematical Sciences, George Mason University, Washington DC, USA
jelin@gmu.edu
[4] School of Mathematical Sciences, College of Science and Technology,
Wenzhou-Kean University, Wenzhou, China
chulee@kean.edu

Abstract. Knowledge of what happens to housing values is limited when properties are near high crime density areas. Big data analysis has become one of the tools for effective crime prevention and can be used as an effective reference when buying house. In this article, we analyzed crime data from 2017 to 2021 in Washington, D.C., and a data set of housing sales information in Washington, D.C. in 2018, which includes crime rates for nine different crime types, as well as internal and external information. The configuration of sold houses in the DC area uses a naive Bayes model to predict the ward where the next crime will occur, and uses XGBoost to explore the housing characteristics of the housing price. The results show that the crime rate of burglary is the highest among all crime types, while the crime rate of ward2 is the highest and the housing price is relatively low. We also created a multiple regression model to predict housing prices based on many numerical and categorical variables provided by the data set. After several cycles of processing and optimization, the most useful parameters for predicting the sales price of houses are determined as the forecasting tool for future housing prices. The results show that the three areas with the highest housing prices are Southwest First Street/Canal Street, Southwest Third Street/Southwest D Street, and Rhode Island Avenue Northwest/Northwest 8th Street. In addition, the regional crime rate is also related to housing prices.

Keywords: Crime rate · house price · random forest · linear regression · XGBoost

1 Introduction

Regional crime rates and housing prices are generally considered to be related. Since crime is closely related to social structure, crime types and crime rates change with

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 1165–1175, 2023.
https://doi.org/10.2991/978-94-6463-034-3_120

economic and social changes. Researchers often try to measure this indirect crime cost in terms of the impact on housing value. For example, it is observed that areas with higher housing prices seem to be affected by fewer crimes [1] This relationship may be correct in many regions, but in the case of metropolitan areas, it is not so straightforward. What distinguishes Washington, DC from other cities is that it has a geographically restrictive border. The limited land area of this city results in a higher cost of living and population density than other cities with wider borders.

XGBoost, which stands for "extreme gradient boosting", is an optimized distributed gradient boosting library that solves many data science problems quickly and accurately. The same code runs in major distributed environments (Hadoop, SGE, MPI) and can solve problems with over billions of examples.[3]Prices can be predicted with our regression model within a certain margin of error. The prediction model can be used to predict house prices for one purpose and see if the house value is correctly assessed for another purpose. Data on the lot size, number of rooms, and year of construction of the home will power the prediction algorithm. As we begin to discover this project, we expect that variables such as land area and number of rooms will be the most important. Based on correlation analysis, correlations between different characteristics and house prices are explained. The goals of this project include predicting the location of the next crime versus the type of crime using Naive Bayesian Model. XGBoost determines the correlation between different house neighborhood feature types and house prices.

2 Sorting and Analyzing Data

2.1 Data Collating

The data is divided into two parts The sources for the first part are primarily the Open Data Center and the Washington, DC Police Department. All crime, justice, census, demographic, crime type-based crime statistics, and zoning data were obtained from these sources in the form of CSVs or shapefiles. All types of crime data are compiled as separate shapefiles by location for each year. The Simple Data Management Tool compiles all crime data for the five years from 2017 to 2021 into one data. We firstly import the excel file into the Jupyter notebook and get the data summarization. We have such large data; We processed the original dataset by transforming the START_DATE variable into separate variables for START_YEAR, START_MONTH,START_DAY, and START_WEEK. We found that START_DATE is not very meaningful because each date is unique and scattered, and it is difficult to analyze the accuracy of the model with thousands of p-values. The original START_DATE column remains in the dataset, but it is not used in the analysis. The OBJECTID, OCTO_RECORD_ID in the data is not strongly associated with the overall data subsequent data will be removed. OFFENSE is transformed into the numeric from 1 to 9 in preparation for exploring the WARD (District is divided into eight wards, each with approximately 75,000 residents) variables correlation with the OFFENSE variables.

The second part of the dataset, named D.C. Residential Properties, can be seen in Fig. 2 and was obtained on kaggle.com. This dataset shows real estate information, including the most recent sales prices for properties in Washington, D.C. as of July 2018. Some of the data was filtered some irrelevant or useless variables were immediately

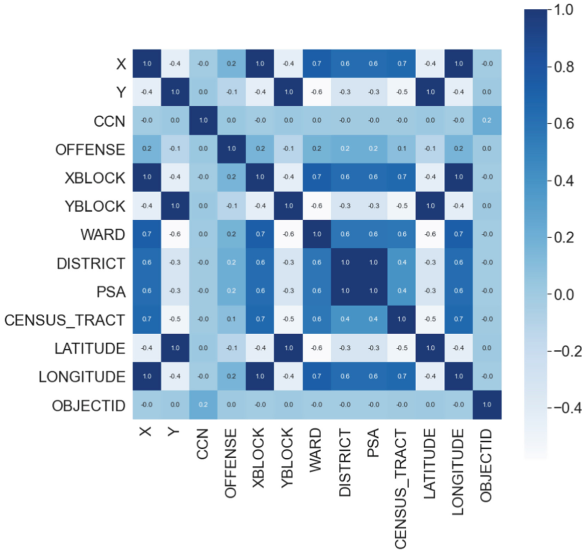


Fig. 1. Correlation coefficient thermal diagram

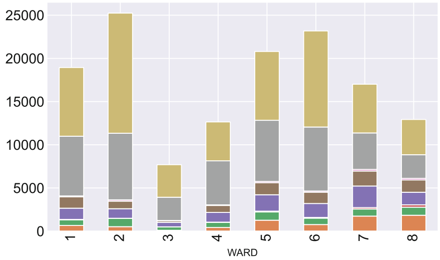


Fig. 2. Histograms in different WARD crime types

removed. In addition, such data types were checked to ensure that some of them could be converted to the correct type. Change the GRADE & CONDITION column to a value with a range (Label Encode). GRADE values range from 1 to 11, CONDITION from 1 to 6 (the larger the number the worse the situation) prepare for the house condition rating.

2.2 Data Prepossessing and Mapping

We have deployed crime data. We first analyzed the correlation between crime types and regions. Here we use heat maps to display the correlation matrix of variables. Blue indicates a positive correlation between variables, and white indicates a negative correlation. The intensity of the color is similar to the intensity of the correlation. In this picture, some correlations seem to be confirmed by intuition. For example, there is some correlation between crime type and area -0.2.

Through the histograms in Fig. 2, we can intuitively see that the number of crimes in high-density areas is higher than in other areas. Ward2 is a high crime area. As can be seen from the above figure, theft, burglary and car are the top three crime types in Washington, DC, USA. One of the best ways to understand data is to visualize it. Since the data may be mainly geographic data, visualizing prices on a map of Washington DC would be a good attempt. By using the color shades in the heat map, we can determine that the darker color area is the area with higher housing prices, and the lighter color area is the area with lower housing prices. The areas without points are mostly hills or other areas that cannot be built. Of course, there are also some expensive areas and some cheap areas. Washington, D.C. is composed of 8 wards, with darker spots in areas 2 and 3, and brighter spots in the rest. Housing prices in areas with high crime rates are generally lower than those in areas with low crime rates.

2.3 Data Visualization of XGBoost Model to Explore the Importance of Each Feature

In order to analyze the factors that affect Washington DC housing prices and build a housing price prediction model, we are trying to use machine learning XGBoost and linear regression algorithms to take into account factors such as the median income of a county, the crime rate of the county, public schools, hospitals, etc. Hospital rating and county unemployment rate. We selected housing information for selected U.S. cities as a reference to reduce errors.

The main principles of XGBoost are briefly described below. The base model in XGBoost is a CART tree, which here simply means that the tree in the base model is a binary tree. Assuming that the CART tree generated for the kth time (which can also be called a residual tree) is f_k, , the final model prediction for sample i after T rounds (there are T trees in total) is (x_i is the input value of the i th sample and T represents the number of trees). The final result is a summation over all CART trees. Xgboost is an incremental learning method, i.e., each tree must be generated after the previous one before it can continue to be obtained, so the code here is designed to be executed serially. The goal of each train is to make the predicted value closest to the true value (i.e., minimizing the loss function). The loss function is often expressed by the following equation

$$\sum\nolimits_{i=1}^{n} l\left(y_i, \hat{y}_i\right) \tag{1}$$

Notice that instead of directly minimizing the above loss function as the goal of train, xgboost has to add the tree complexity to the above formula. This is to avoid overfitting (good performance on the training set, but poor performance on the test set or on the data to be predicted). So, the final objective function is:

$$obj = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \tag{2}$$

Therefore, after getting the t th tree, the total loss value becomes:

$$obj^t = \sum\nolimits_{i=1}^{n} l(y_i,\ \hat{y}_i^t) + \sum\nolimits_{k=1}^{t} \Omega(f_k) \tag{3}$$

The value of the total complexity of the model at the t th tree is the complexity of the previous t-1 trees plus the complexity of the t-th tree.

$$\sum_{k=1}^{t} \Omega(f_k) = \left(\sum_{k=1}^{t-1} \Omega(f_k)\right) + \Omega(f_t) = \text{const} + \Omega(f_t) \tag{4}$$

The objective function becomes:

$$obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \text{const} + \Omega(f_t) \tag{5}$$

and the Loss function part of the objective function (first part). Second-order Taylor formula:

$$f(x + \Delta x) \approx f(x) + f'(x)\,\Delta x + \frac{1}{2} f''(x) \Delta x^2 \tag{6}$$

We take $\hat{y}_i^{t-1}$ as x and $f_t(x_i)$ as Δx. Then the first part of the loss function becomes:
$\sum_{i=1}^{n} l(y_i, \hat{y}_i^t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i))$
$\approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{t-1}) + \frac{\partial l(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}} f_t(x_i) + \frac{1}{2} \frac{\partial^2 l(y_i, \hat{y}_i^{t-1})}{\partial (\hat{y}_i^{t-1})^2} (f_t(x_i))^2 \right]$
$\approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i (f_t(x_i))^2 \right]$

In a classification task, assuming that the first 5 trees are known, now to determine the 6th tree, the loss function is:

$$L(\theta) = \sum_i \left[y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i}) \right] \tag{7}$$

There is a sample with a true label of 1, but the first 5 trees predict a value of -1. Since $y_i = 1$ the loss function for this sample becomes: $\ln(1 + e^{-\hat{y}_i})$. Here g_i should then be the resultant value obtained by bringing the total predicted value of the first 5 trees -1 into the derivative of the above equation for $\hat{y}_i$.

The derivation yields:

$$\frac{1}{(1 + e^{-\hat{y}_i})} * (e^{-\hat{y}_i}) * (-1) = \frac{-e^{-\hat{y}_i}}{1 + e^{-\hat{y}_i}} \tag{8}$$

From the above analysis, it can be seen that.for N samples, then there will be N with the required requirements. For each sample, there will be one vs. one, so here the sums corresponding to each sample can be obtained in a parallel manner (this is one of the reasons why XGBoost is fast). For each tree, the process of finding involves the quadratic derivation of the loss function, so XGBoost can customize the loss function, and only the loss function is quadratically differentiable. At this point, the objective function becomes:

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i (f_t(x_i))^2 \right] + \text{const} + \Omega(f_t) \tag{9}$$

Tree complexity in objective function, regular term(second part).

In the above equation, $f_t(x)$ represents a tree model. Imagine that a tree model is given an input sample x_i, which after going through the model will be divided into some leaf nodes and will give the predicted value of that leaf node. So $f_t(x)$ can be represented by the following equation:

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d \to \{1,\ 2\ \cdot\cdot\cdot,\ T \tag{10}$$

XGBoost defines the canonical term as: $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$.
By analyzing the two parts, we obtain

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{n}\left[g_i w_{q(x_i)} + \tfrac{1}{2}h_i w_{q(x_i)}^2\right] + \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &\approx \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \tfrac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T \end{aligned} \tag{11}$$

Another upgrade of the above objective function yields,

$$\text{obj}^{(t)} = \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2\right] + \gamma T \tag{12}$$

Our final goal is to minimize the above objective function, then the derivative of the above equation (the above equation is partial derivative for the variable w_j and make it 0,

$$G_j + (H_j + \lambda)w_j = 0 \tag{13}$$

we can get

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \lambda T \tag{14}$$

With the above equation, it is then possible to find the specific value of each leaf node of the tree, in addition to requiring exactly which partition point of each tree should be divided by which attribute (feature). XGBoost is to select the attribute and the partition point that makes the greatest loss reduction (Gain of loss value) after partitioning. Since the model sums the results of each prediction, it means that each prediction is a regression value, not a category value, otherwise the summation is meaningless.

We can also see the importance of each feature by calling plot_importance inside XGBoost. The importance of a feature is calculated as the decrease in the impurity of the node and weighted by the probability of reaching that node. The higher the value, the more important the feature is.

Figure 3, f0 is the number of schools, f1 is the number of hospitals, f2 is the average rating of hospitals, f3 is the unemployment rate, f4 is the crime rate, and f5 is the median income. As can be seen from the above graph, unemployment rate has the highest importance, followed by crime rate, median income, number of schools, and then average rating of hospitals. Figure 4, ward2 is ranked seventh in terms of their overall price impact on housing combined. Generally, there is a correlation between crime rate and house price, but it is not the most important influencing factor.

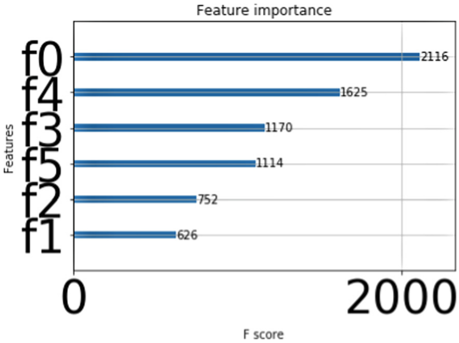


Fig. 3. Feature importance with XGBoost

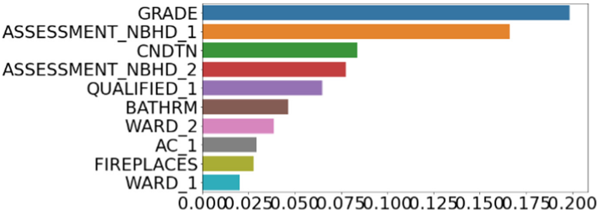


Fig. 4. Feature importance of Washington D.C House price

3 Data Visualization of The Forecasting Model For Washington D.C House Price

3.1 predictive Modeling-Find the Best Algorithm

With housing prices as the goal of analysis and prediction, looking at the dependent variables also helps to understand the independent variables, because the goal is to define the relationship between them. The histogram gives the basic concept of price distribution. There are deviations in the model and statistics of the housing price data. Most houses are within a certain range, and some are expensive. The histogram reflects limited information. Statistics show that the minimum value of $10 is meaningless, but the price is biased towards very high values. In addition, buying a house at a median price in Washington, DC will be very difficult.

In order to model in this data set, I will try to use several regression models, namely Random Forest Regressor, Lasso, Linear Regression, Lasso, Ridge, and XGBoost Regressor. First use the six algorithms mentioned earlier and the default parameters to find the best algorithm. Then use hyperparameters to predict the best model, and finally use average absolute error, mean square error and root mean square error for evaluation. The target (y) and feature (x) of this modeling. The goal is the price of housing. Here features include quadrant, district, assessment neighborhood, room, bathroom, half-bath, heating, air conditioning, year of remodeling, qualified dwelling, gross floor area, style dwelling, structural dwelling, grade, condition, exterior wall, roof, interior wall, kitchen, fire location.

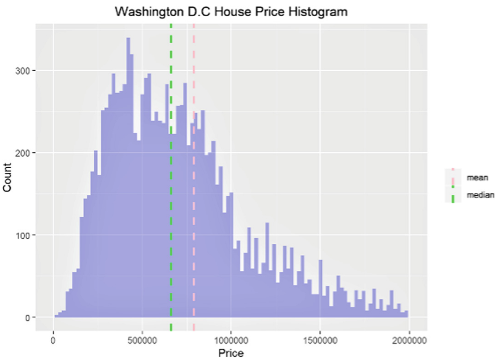
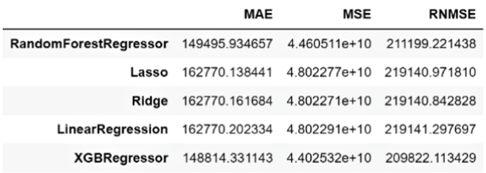


Fig. 5. Histogram of Washington D.C House Price



	MAE	MSE	RNMSE
RandomForestRegressor	149495.934657	4.460511e+10	211199.221438
Lasso	162770.138441	4.802277e+10	219140.971810
Ridge	162770.161684	4.802271e+10	219140.842828
LinearRegression	162770.202334	4.802291e+10	219141.297697
XGBRegressor	148814.331143	4.402532e+10	209822.113429

Fig. 6. the result for modelling with Standard Scaller and Binary Encoding.

Next I try to find the best n_components using principal component analysis (binary encoding and standard scaler) and model with random forest moderator. Try to model with the best n_components = 5.

The result show, Value Mean Absoute error:181593.25, Value Mean squared error:58285660515.74, Value Root Mean Squared Error:241424.23. MAE score increases when the dataset is predicted using principal component analysis (a hot coding and standard scaler), so do not use this algorithm. MAE refers to the mean error. RMSE (root mean square error) is preferred overMSE because RMSE can be interpreted in “y” units. All of these are loss functions, and smaller MAE and RMSE values mean better models. Root mean square error is widely used to evaluate regression models. This is the mean of the difference between the true and predicted values for each observation that ignores the sign.

After a series of steps I found the best algorithm among the six. XGBoost Regressor is the best model for modeling using binary encoding and standard scaller. Optimal hyperparameters for the model (XGBRegressor), max_depth = 5 & learning_rate = 0.1

3.2 Final Modeling- Optimize House Price Forecasts

After finding the optimal hyperparameters, input the hyperparameters into the model to optimize the price prediction. The next step is to find the best hyperparameter score and split the data set into training data and test data. I tried to use two methods-Train Test split and Kfold. Figures 8 and 9 explain the distribution of model predictions relative to the real target (y_test).

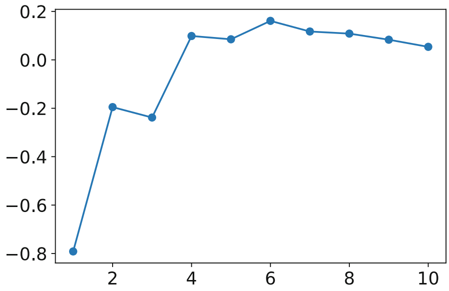


Fig. 7. Random Forest Regressor modelling with Standard Scaller and Binary Encoding.

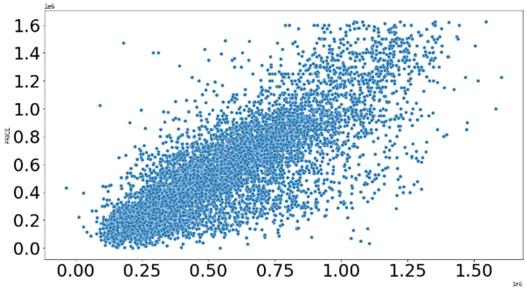


Fig. 8. Train Test split- distribution of prediction from model vs real target (y_test)

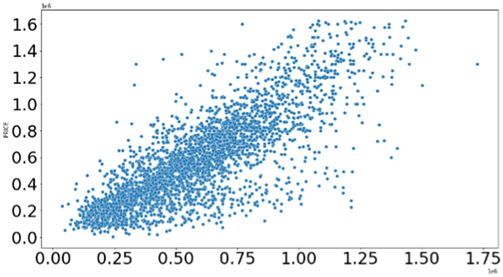


Fig. 9. Kfold-distribution of prediction from model vs real target (y_test)

The prediction model can be used to predict house prices for one purpose and to see if the house values are correctly evaluated for another purpose. The main steps start with splitting the dataset, training a part of it, and then test the model against the rest. However, there may be biases between the train and test sets.

Models usually predict some difference in the price of a particular house. For example, a $500,000 house could be predicted to be $533,000 or $484,000. The graph below shows a plot of the residuals expressed in dollar amounts. There does not appear to be a specific pattern between the residuals. Note that the darker color on the left is simply because there are more observations in that range.

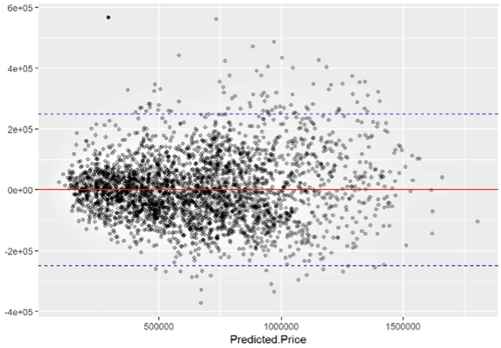


Fig. 10. Residual plot

The final results show that the average and median home prices are $660,583 and $619,900, respectively, and the predicted values are $684,259 and $593,353, respectively. There is a gap of approximately $20,000 between the true statistics and the projections. This is much lower than the overall RMSE, but more meaningful because most observations are around the mean and median, so the new data will be as well. Since Washington, D.C. is the capital of the United States housing is more expensive, even though adding more variables may improve the results further.

4 Conclusion

Crime poses a threat to the residential stability of the community. This project focuses on exploring the impact of community crime on the value of housing. A major disadvantage of these studies is that although crime is undoubtedly endogenous in the property value model due to simultaneity, missing variables or measurement errors, most people regard crime measurement as an exogenous independent variable [2]. Among the nine different types of crimes we investigated, robbery and serious assault crimes had little impact on the value of district housing. With the increase in the number and size of rooms, house prices have become more expensive. Areas such as wards and housing conditions have a greater impact on prices, but construction materials have little impact. Overall, there seems to be a negative correlation between regional crime rates and housing prices. This case study observes that house internal facilities are the main factor affecting house prices, followed by the number of schools. The scatter plot matrix also shows that there is no strong correlation between the total crime rate and house prices.

References

1. Tita, G. E., Petras, T. L., & Greenbaum, R. T. (2006). Crime and residential choice: A neighborhood level analysis of the impact of crime on housing prices. *Journal of Quantitative Criminology, 22*(4), 299-317. doi:https://doi.org/10.1007/s10940-006-9013-z
2. Does crime drive housing sales? Evidence from Los Angeles. (n.d.). Retrieved June 14, 2021, from https://www.tandfonline.com/doi/abs/10.1080/0735648X.2013.812976#.UyBwW84Xe3M

3. XGBoost https://xgboost.readthedocs.io/en/latest/index.html
4. "Open Data DC". Opendata.Dc.Gov, Last modified 2020. https://opendata.dc.gov/.
5. DC Cover photo: https://www.wallpaperflare.com/the-washington-monument-under-blue-sky-washington-dc-wallpaper-204