CPE695 Mid Stage Report

*Note: Sub-titles are not captured in Xplore and should not be used

1st Junyan Yang
Department of Computer Science
Stevens Institute of technology
Hoboken, USA
jyang68@stevens.edu

2nd Xiaohan Hou
Department of Computer Science
Stevens Institute of technology
Hoboken, USA
xhou9@stevens.edu

3rd Hongyi Zhang
Department of Computer Science
Stevens Institute of technology
Hoboken, USA
hzhan29@stevens.edu

***Abstract*—In this project, we are going to look into the The Big Cities Health Inventory dataset. We are going to analyze the dataset, and perform some machine learning technology to see if the result can be accurately predicted.**

***Index Terms*—machine Learning**

I. Introduction

The Big Cities Health Inventory dataset includes 569 health datasheet records including more than 50 indicators that describe health status, death rates, and other socio-economic and demographic factors. The purpose of the dataset was to identify the relationship between leading causes of morbidity and mortality in the United States. The 9 key leading causes of morbidity included: Behavioral Health and Substance Abuse, Cancer, Chronic Disease, Environmental Health, Food Safety, HIV/AIDs, Infectious Disease, Injury and Violence, and Maternal and Child Health.

	Indicator Category	Indicator	Year	Sex	Race/Ethnicity	Value	Place	BCHC Requested Methodology	Source	Methods	Notes	90% Confidence Level - Low	90% Confidence Level - High	95% Confidence Level - High	95% Confidence Level - Low	95% Confidence Level - High
0	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	Black	1.8	U.S. Total, U.S. Total	Age-Adjusted rate of opioid-related mortality ...	CDC WONDER	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	1.7	1.9
1	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	American Indian/Alaska Native	6.8	U.S. Total, U.S. Total	Age-Adjusted rate of opioid-related mortality ...	CDC WONDER	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	5.8	7.9
2	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	All	4.4	U.S. Total, U.S. Total	Age-Adjusted rate of opioid-related mortality ...	CDC WONDER	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	4.4	4.5
3	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	White	3.3	Fort Worth (Tarrant County), TX	Age-adjusted rate of opioid-related mortality ...	National Center for Health Statistics	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	2.4	4.9
4	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	Other	NaN	Fort Worth (Tarrant County), TX	Age-adjusted rate of opioid-related mortality ...	National Center for Health Statistics	NaN	Records where the value is blank in the data t...	NaN	NaN	NaN	NaN	NaN

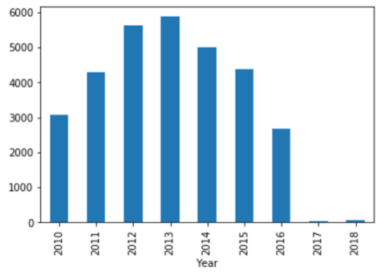
After reviewing the data set, we decided to define the attribute "Value" as the dependent variable y as it represented the number of deaths per 100,000 people. We found this to be appropriate given that we were interested in understanding what factors and attributes drive the death rate in our population. After defining "Value" as the dependent variable y, we did some exploration to find and define the independent variables which would have a possible correlation to y. The goals and our next steps would revolve around finding these variables, creating a model based on algorithms we learnt, which then would allow us to predict "Value" (dependent variable) given inputs we enter. Below is a breakdown of the parameters we assessed and our conclusions at the midstage.

Identify applicable funding agency here. If none, delete this.

II. Data Description / Analysis

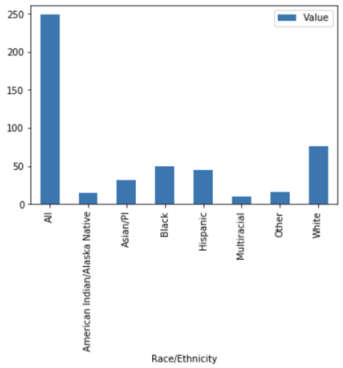
The below summarizes our results from our analysis of the data parameters in the dataset and how they provide meaningful information or potential trends. We utilized bar charts and graphs to plot out various aspects of our data to identify trends and to support the inclusion of certain parameters as independent variables and for our algorithms.

A. Year vs. Number of recorded values



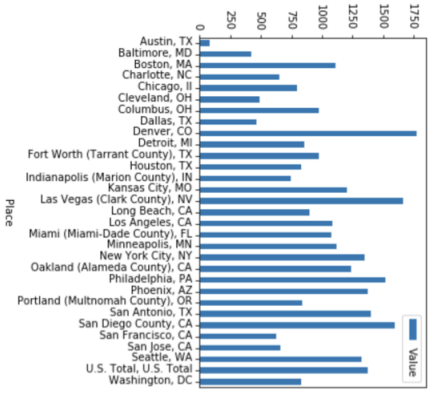
Plotting the total number of recorded samples against the parameter year from 2010 to 2018 allowed us to identify a high level trend, that the number of recorded values was increasing rapidly from 2010 to 2013, peaking in 2013. Beginning 2014, the number of recorded values began to steadily decrease to 2016. The drop off noted in 2017 and 2018 is likely due to a lack of available data.

B. Race/Ethnicity



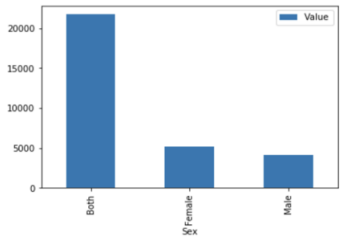
This chart displays the number of recorded values across each race/ethnicity. Reviewing the graph, the results are rather inconclusive as approximately 14,000 of the data points assigned a race/ethnicity value of "All" indicating that the samples were in majority not race specific. The remaining split was relatively even.

C. Place/Location



Using a bar chart to plot the number of recorded values against the city/location we are able to identify where the majority of recorded values occur and whether living in a specific location may increase the likelihood of morbidity. The leaders of the recorded values from our dataset were Denver, CO, followed by Las Vegas and San Diego County. By contrast, the recorded values in the location with the fewest recorded values was Austin, TX with a count of 80 drastically less compared with Denver.

D. Gender



The following chart shows the recorded values distribution for the different genders. Similar with our findings from our analysis of Race, many data entries did not assign a specific gender (Both). As a result, given the fairly equal distribution, we are unable to conclude how gender currently impacts the number of recorded values.

III. Current Status

A. Defining Key Independent Variables

Having reviewed the list of indicators in the data set and analyzed the plot diagrams, we decided to pinpoint the key indicators to focus on and further assess, while dropping other attributes we did not deem material or data that was unavailable(missing values).

Indicator Category	Indicator	Year	Sex	Race/Ethnicity	Value	Place	BCHC Requested Methodology	Source	Methods	Notes

Our team designated Race, Area, and Sex to be the key drivers for determining the death rate given the following:

- Race, location and sex are attributes that are more readily available and obtainable in comparison to other indicators such as indicator category (Reason for death)
- The indicator category (Reason for death) was not selected as a key indicator as we believe the correlation between this independent variable and the dependent variable (number of deaths) would be too high and prevalent and might outweigh the impact of other variables.
- Additionally, when attempting to apply machine learning and create a model for estimating the number of deaths, the reason or cause of death might not be available.

B. Pre-processing Data

As many of the indicator values in the data set were not numerical but strings, we were unable to directly analyze or quantify the correlation between the independent variables and dependent variable. As a result, with each key indicator noted above, we transformed their categorical features into numerical features by implementing one-hot coding illustrated below.

This would allow our dataset to be analyzed deeper.

```
In [17]: df_categorical_col = data.select_dtypes(exclude=np.number).columns
         df_categorical_col
Out[17]: Index(['Sex', 'Race/Ethnicity', 'Place'], dtype='object')

In [18]: df_numeric_col = data.select_dtypes(include=np.number).columns
         df_numeric_col
Out[18]: Index(['Value'], dtype='object')

In [19]: df_onehot = pd.get_dummies(data[df_categorical_col])
         df_onehot.head(5)
Out[19]:
```

	Sex_Both	Sex_Female	Sex_Male	Race/Ethnicity_All	Race/Ethnicity_All	Race/Ethnicity_American Indian/Alaska Native	Race/Ethnicity_Asian/Pi	Race/Ethnicity_Black	Race/Ethnicity
0	1	0	0	1	0	0	0	0	
1	1	0	0	1	0	0	0	0	
2	1	0	0	1	0	0	0	0	
3	1	0	0	1	0	0	0	0	
4	1	0	0	1	0	0	0	0	

5 rows × 43 columns

C. Divided training data vs test data

To begin training our model, we divided our population data set into two categories, a) Training Data - Data used to train our model and b) Test Data - Remaining data population to later assess the accuracy of our model and the residual error.

D. Algorithms

We implemented various algorithm methodologies to assess which would provide us the lowest level of errors and highest levels of correlation between our variables. Below details our results from applying the following methods.

- linear regression
- k Nearest Neighbor (kNN Model)
- Regression Tree
- SVM Algorithm (ongoing)
- ANN
- k-Means Clustering (ongoing)

1) Multi-Linear Regression: Applying Multi-linear regressions first allowed us to identify whether linear relationships existed between the individual independent variables and the dependent variable Value (Death rate).

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{1}$$

We then have the our weights as can be shown in the following image

```
array([[ 2.59874034e+17,  2.59874034e+17,  2.59874034e+17,
        -9.56348050e+16, -9.56348050e+16, -9.56348050e+16,
        -9.56348050e+16, -9.56348050e+16, -9.56348050e+16,
        -9.56348050e+16, -9.56348050e+16, -9.56348050e+16,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17]])
```

After getting the parameters from the training data, we then applied the model from above to the test data which provided the result of $RMSE$ (Root-mean-square deviation)=3331346.43963654.

2) kNN Model: kNN is the simplest supervised algorithm mostly used for classifying data points based on how it's neighbors are classified. In our earlier steps, we transformed all categorical features into numerical discrete attributes 0,1 for the instances, so that we can apply the kNN algorithms here. In order to determine the best value for "k", we assessed the RMSE as we increased the factor of k. The best k is the one that minimizes the prediction error RMSE. Referring to the chart below, we were able to conclude that utilizing a k factor of 3 resulted in the lowest RMSE.

```
Rmse for k = 1 is 3334819.233243835
Rmse for k = 2 is 3334332.7992320075
Rmse for k = 3 is 3334185.2569772894
Rmse for k = 4 is 3334972.892561651
Rmse for k = 5 is 3334744.266393955
Rmse for k = 6 is 3334494.7269203314
Rmse for k = 7 is 3334352.040725512
Rmse for k = 8 is 3334467.092127072
Rmse for k = 9 is 3334915.375326055
Rmse for k = 10 is 3334689.428503498
```

The RMSE for test data when k=3 is the minimal given only 10 iterations (Root-mean-square deviation)=3334185.2569772894, we will do further iterations of the neighbor number to maybe find a smaller RMSE for the model.

3) Regression Tree: Our next method applied involved decision tree learning. Regression tree analysis is when the predicted outcome can be considered a real number(wiki). The goal was to create a model that predicts the value of our target variable based on the several input variables/parameters noted earlier. The predicted outcome can be considered a real number, which in this case would represent the 'value' given our input. We will apply Reduced Error Pruning for the tree in later experiments. The RMSE for test data when k=4 is the minimal given only 10 iterations

```
Rmse for k = 1 is 3331917.708214744
Rmse for k = 2 is 3329179.579061839
Rmse for k = 3 is 3315780.0760172135
Rmse for k = 4 is 3315742.1185617554
Rmse for k = 5 is 3315751.480798816
Rmse for k = 6 is 3315762.327509618
Rmse for k = 7 is 3315762.207724649
Rmse for k = 8 is 3315753.450087217
Rmse for k = 9 is 3315753.2626337833
Rmse for k = 10 is 3315757.476472194
```

(Root-mean-square deviation)=3315742.1185617554, we will do further iterations of the max-node number to maybe find a smaller RMSE for the model.

4) Preliminary result: In the current stage(Midstage), based on our existing findings, having run through three different algorithm approaches, we believe the model using multi-Linear Regression is by far our best model with RMSE=3331346.43963654. However, with more k neighbors in the KNN model, we might get an even smaller RMSE. We might need further tests for the k values in both the knn model and the regression tree model.

Given our current input, the model using kNN has the worst performance with the RMSE 3334185.2569772894 with k=3.

We will apply additional machine learning algorithms which are kMeans Clustering, SVM Algorithm and ANN, so that we can further discuss the model performance.

Moreover, we will use hyper-parameter tuning to determine some very crucial hyper-parameters to avoid overfitting and to increase the accuracy of our models.

REFERENCES

https://www.bigcitieshealth.org/methodology