

New York Institute of Technology

## **Spring 2020 - Introduction to Data Mining - CSCI\_415**

### **Explore 10 data sets**

Members & ID:

<b>An Zifan</b>	<b>1199548</b>
<b>Yang Ziqing</b>	<b>1250876</b>
<b>Fan Zehao</b>	<b>1199729</b>
<b>Jing Chengye</b>	<b>1199581</b>

Instructor:

**Huanying Gu**

Data:

**2020/2/26**

## I. Sources & Contribution

The 10 data sets were retrieved from **NYC OpenData**, the **U.S. Government's open data** and **Sinaapp**. The specific division of labor is shown in the table below:

Name:	Data Sets:
An Zifan	1. Accidental Drug Related Deaths 2012-2018
	2. RDU Local Weather Archive
	3. NYPD Shooting Incident Data (Historic)
Yang Ziqing	4. China AQI(Air Quality Index) Data
	<b>5. NYPD Complaint Data</b>
Fan Zehao	<b>6. 2018_Central_Park_Squirrel_Census_-_Squirrel_Data</b>
	<b>7. Citywide Payroll Data</b>
Jing Chengye	8. Bridge Data
	9. 2013-2019_Attendance_Results
	10. Energy_Incentive_Program

Tabel 1: The data for NYC OpenData has been bolded

Each person is responsible for the visualization of the corresponding data. We used an online tool called **Hanabi** for data visualization. (<http://hanabi.data-viz.cn/>)

## II. Data Sets & Visualization

### 1. Accidental Drug Related Deaths 2012-2018

*Metadata Updated: May 8, 2019*

The data was retrieved from the U.S. Government's open data

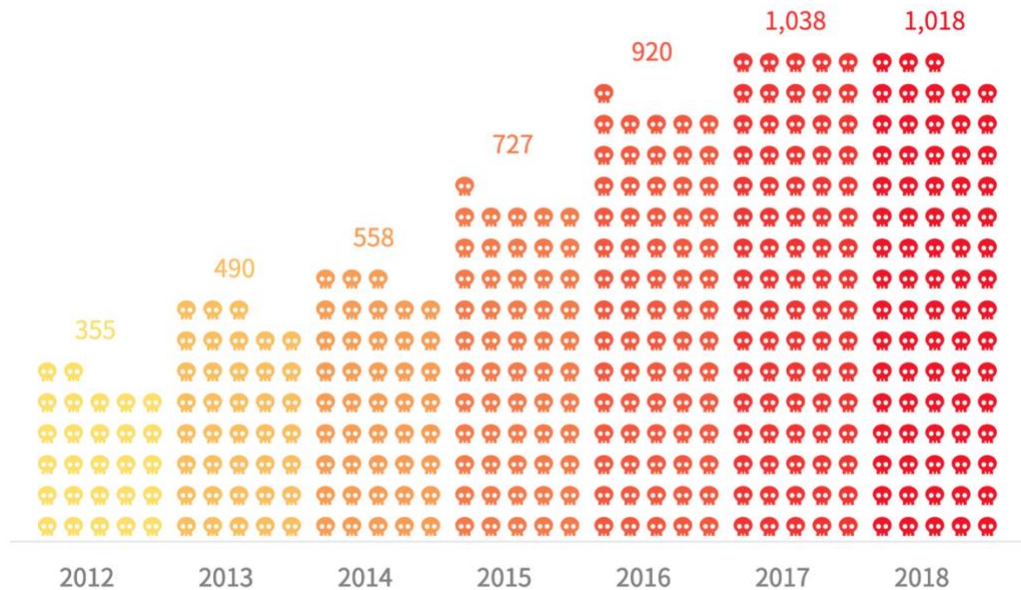
This data set is a listing of each accidental death associated with drug overdose from 2012 to 2018 in Connecticut. It details what kind of drugs the victim died from.

The data set contains information about the age, sex, race, location of death and type of drug overdose.

(<https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015>)

## Accidental Drug Related Deaths 2012-2018

Num of Death



Source: <https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015>

## 2. RDU Local Weather Archive

*Metadata Updated: July 22, 2017*

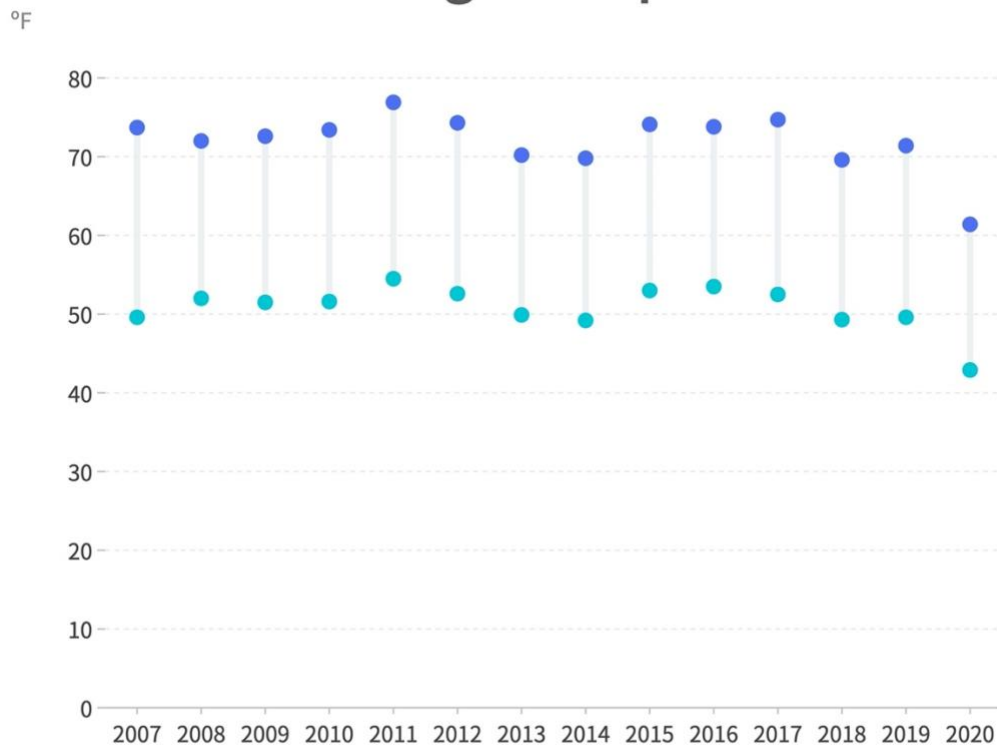
The data was retrieved from the U.S. Government's open data

This data set contains Raleigh Durham International Airport weather data pulled from the NOAA web service.

It contains weather data for RDU International Airport from 2007 to 2020, such as temperature, wind speed, and weather conditions.

(<https://catalog.data.gov/dataset/local-weather-archive>)

### RDU Local Average Temperature



Source: <https://catalog.data.gov/dataset/local-weather-archive>

### 3. NYPD Shooting Incident Data (Historic)

*Metadata Updated: May 21, 2019*

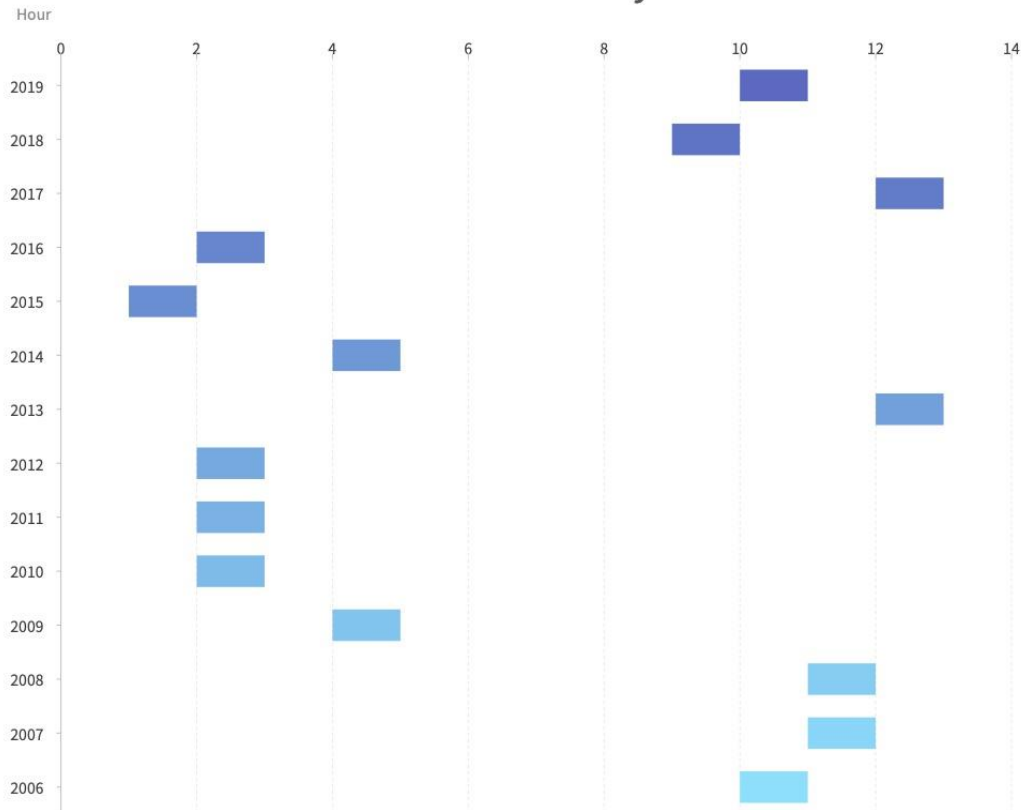
The data was retrieved from the U.S. Government's open data

This data set is a list of every shooting incident that occurred in NYC going back to 2006 through the end of 2019.

Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included, such as gender, age and race.

(<https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic>)

#### The Deadliest Period in New York City



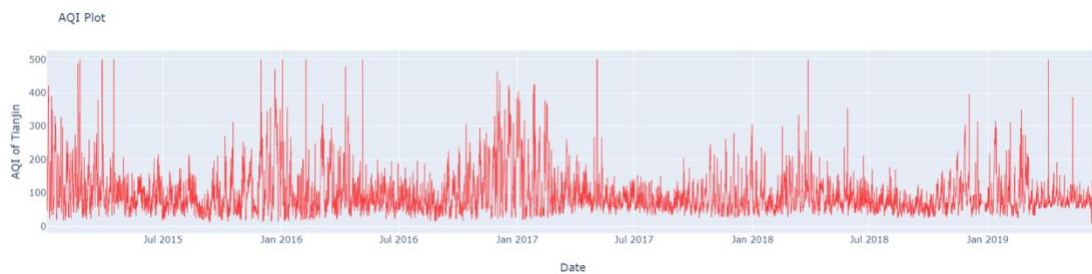
Source: <https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic>

#### 4. China AQI(Air Quality Index) Data

*Metadata Updated: Feb 23, 2020*

This dataset presents the AQI of each city of China and those attributes affect AQI of that city, such as, PM2.5, PM10, SO2, O3 and CO. The time range of this dataset is from 2015/01/01 to 2020/02/23. It contains 340 cities of China. The data is collected hourly, which means each row of the data set is the hourly value of that specific attributes of the specific city.

(<http://beijingair.sinaapp.com/>)



#### 5. NYPD Complaint Data

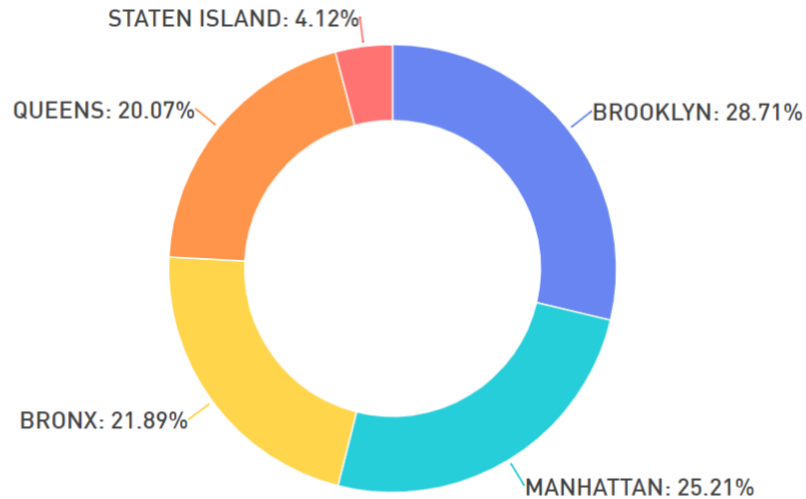
*Metadata Updated: February 7, 2020*

This dataset includes all kinds of violation of crimes which reported to the NYPD (New York City Police Department). The time range of this dataset is about from 1974 - 2019.

It described the borough where the incident happened, the date and time the incident gets reported, the date and time it gets completed, and etc.

(<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data>)

## Case Reported In Each Borough



Source: NYC OpenData

### 6. 2018\_Central\_Park\_Squirrel\_Census\_-\_Squirrel\_Data

*Metadata Updated: October 18, 2019*

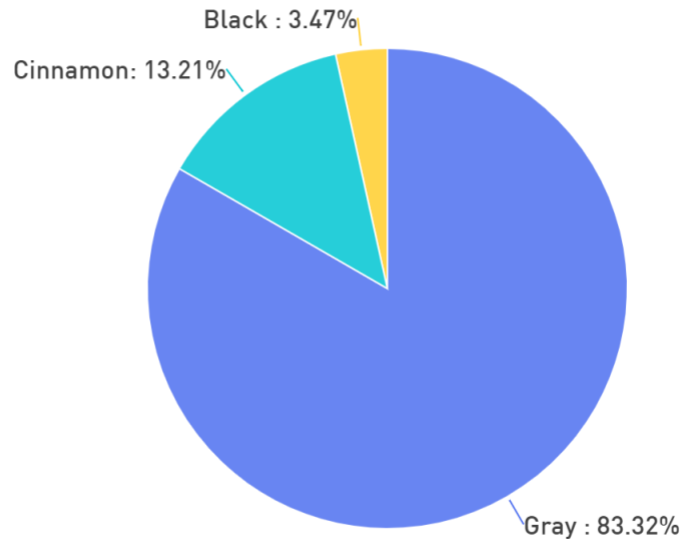
The data was retrieved from NYC OpenData.

The dataset recorded the result of 2018 Central Park Squirrel Census. The Central Park was divided into hectares, and information of conditions of each squirrel sighted in the hectares are recorded. The dataset contains 3023 rows, and each row represents one squirrel.

The information of longitude and latitude coordinate for the sighting point, the hectare from which the squirrel was spotted, date, age, fur color, actions, interactions with other squirrels are included in the dataset.

(<https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Squirrel-Data/vfnx-vebw>)

## Fur Color of Squirrels in the Central Park



Source: <https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Squirrel-Data/vfnx-vebw>

## 7. Citywide Payroll Data

*Metadata Updated: February 7, 2020*

The data was retrieved from NYC OpenData.

This dataset includes yearly final salary of all municipal employee except from some agencies like NYPD and the District Attorneys 'Offices from the year 2014 to 2019.

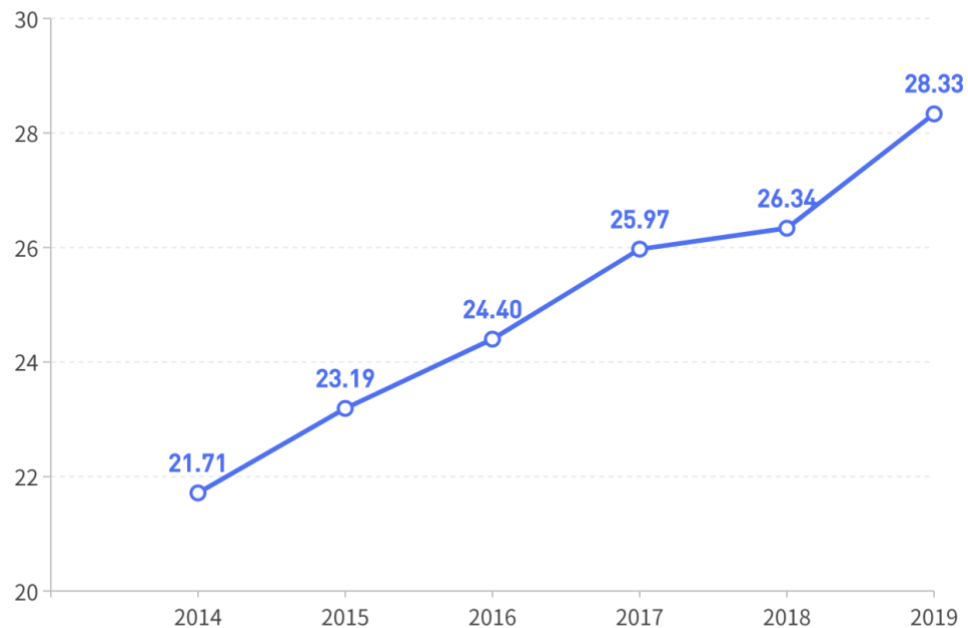
There are 3.33 million records in total, containing agency, names, work location borough, job title description, leave status, base salary, regular gross paid, overtime hours worked, total overtime paid, and total other compensation information of each employee.



(<https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e>)

## Total Regular Gross Paid + Total Overtime Pay Paid

(billion dollars)



Source: <https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e>

## 8. Bridge Data

Metadata Updated: November 27, 2019

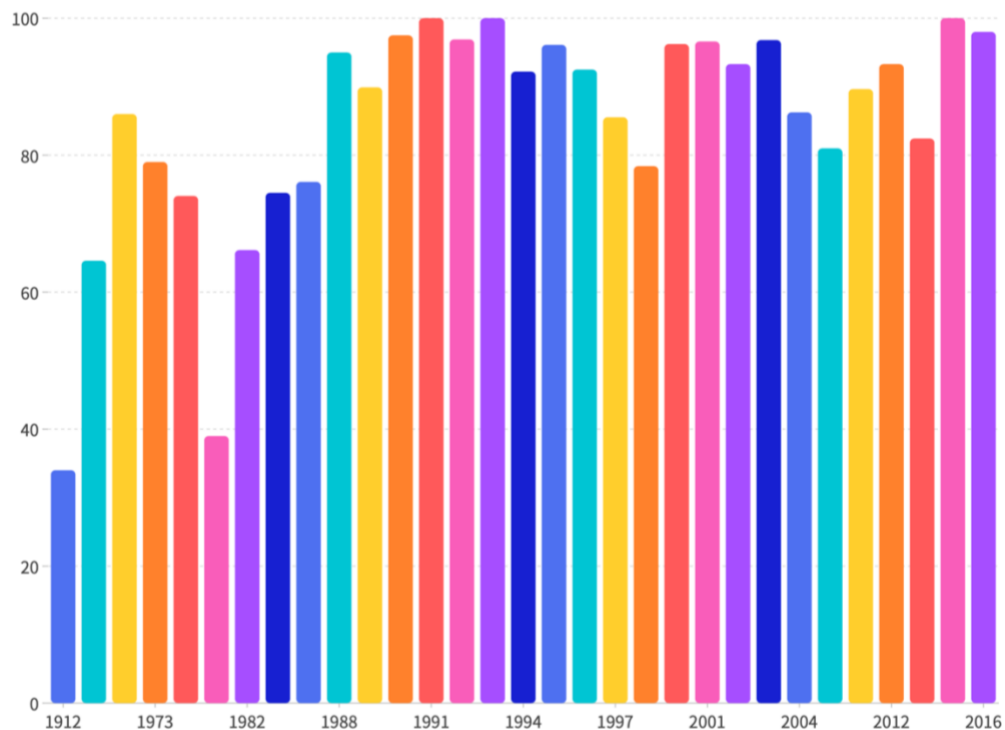
Feature layer containing authoritative bridge points for Sioux Falls, South Dakota.

This data set is about the different bridges. The number of the data is 93 in total.

There are many features of the bridges, such as the location, name, structure number, install year and sufficiency rating. And there is more information of the physical information of those bridges.

(<https://catalog.data.gov/dataset/bridges-f1b2b>)

## The sufficiency rating of bridges in different years



<https://catalog.data.gov/dataset/bridges-f1b2b>

## 9. 2013-2019\_Attendance\_Results

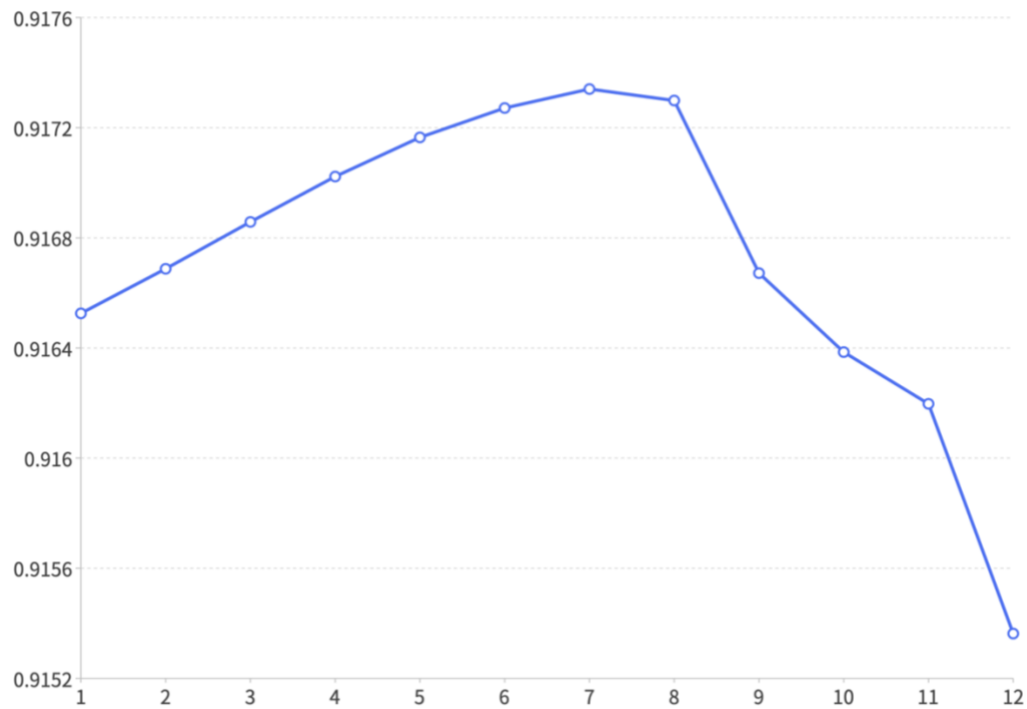
*Metadata Updated: February 19, 2020*

Attendance data includes students in district 1-32, 75 (Special Education), district 79 (Alternative Schools & Programs), charter schools, home schooling.

This data is about the attendance of students in 1 to 12 grades. We can see the total days, absent days, and different kinds of absence of the students. The information is also including the schools of the students.

(<https://catalog.data.gov/dataset/2013-2019-attendance-results-citywide>)

### Attendance of different grades students



<https://catalog.data.gov/dataset/2013-2019-attendance-results-citywide>

### 10. Energy\_Incentive\_Program

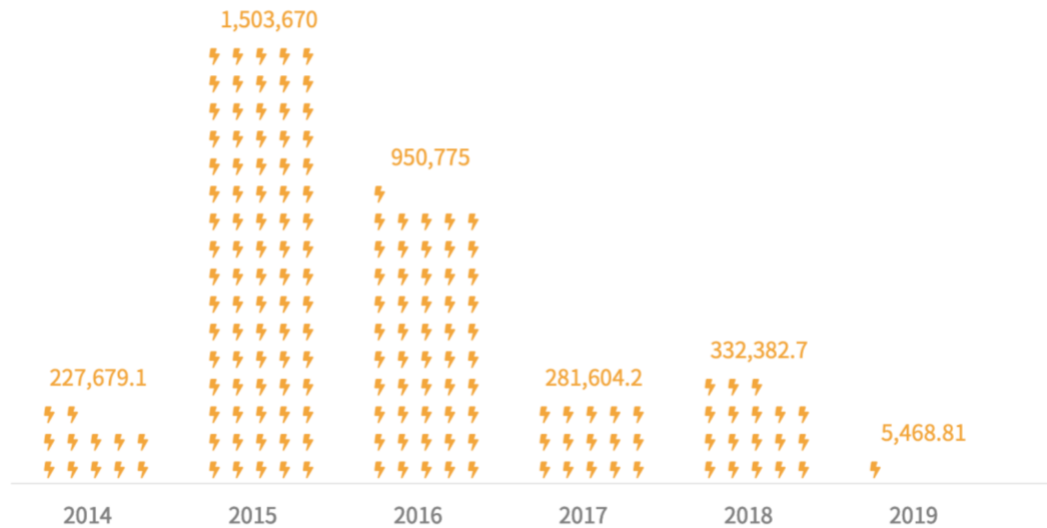
*Metadata Updated: January 16, 2020*

In this data set we can find the data of Oregon Department of Energy. We can find the energy saved in different programs. And we can find the years, name, city address and kWh Savings of the programs. So from the information, we can know the energy saving in different years.

(<https://catalog.data.gov/dataset/energy-incentive-program-eip-fy-2015-2019>)

## Different Years of Energy Saved

Million Btu



Source: <https://catalog.data.gov/dataset/energy-incentive-program-eip-fy-2015-2019>