Task 6.1 – Sourcing Open Data

Data Profile - Uniform Crime Reporting Program (UCR) – Hate Crime Statistics

Data Source:
This is an external data source. The data is provided by the FBI, which is a United States Federal Agency. The Hate Crime Statistics Program of the FBI's Uniform Crime Reporting (UCR) Program collects data regarding criminal offenses that were motivated, in whole or in part, by the offender's bias against a race, gender, gender identity, religion, disability, sexual orientation, or ethnicity, and were committed against persons, property, or society. As the data is provided by the government, we can count on this source to be trustworthy.

Data Collection:
The FBI collects administrative data from over 18,000 federal, state, city, tribal, university and college law enforcement agencies monthly or quarterly. Whilst federal law enforcement agencies are mandated to participate in the UCR program, all other law enforcement agencies participate on a voluntary basis. Data is either submitted through a State's Uniform Crime Reporting Program or directly to the FBI's UCR program. Participating agencies submit hate crime data electronically in a National Incident-Based Reporting System submission, the hate crime record layout, or a Microsoft Excel Workbook Tool.

Data Contents:
The data helps us to understand the occurrence of prosecutable hate crimes by location, date and nature, as well as details of the law enforcement agency that reported the crime. Variables here include location of crime, date of the crime, the name and type of law enforcement agency, type of offender and victim, the number of offenders and victims, the type of offense and the nature of the bias.

Data Limitations and Bias:
The data is limited to a certain extent as it represents an incomplete picture of hate crimes happening in the USA. As most law enforcement agencies are under no obligation to report data to the UCR it is hard to know how comprehensive the data set is. To understand how representative of the whole population of law enforcement agencies this sample is, I would need to research the total number of law enforcement agencies in the United States.

The data collection method is problematic as definitions of what constitutes a hate crime differ from region to region. Whilst the FBI issues standardised guidelines and training to all LEA's on how to classify and score offenses that they report, these definitions often vary from the local or state crime offense definitions which could cause confusion and error in how crimes are reported. The accuracy of the statistics depends upon each LEA's ability to follow the guidelines and standards set out by the FBI.

We should also be careful to note that many hate crimes go unreported and are therefore missing from the data. The data only represents hate crimes that have been deemed to be prosecutable and does not contain information on hate incidents. It is also important to consider the possibility of awareness bias here. A heightened awareness of the issue of hate crimes over time, as more people become aware through media coverage etc, could have led to more crimes being reported.

Data Profile - United States Census Data

Data Source:
This is an external data source. The data is provided by the United States Census Bureau which is the largest statistical agency of the Federal Government. As the data is provided by the government, we can count on this source to be trustworthy.

Data Collection:
This is survey data collected manually by The American Community Survey. Selected households are notified via mail and asked to complete the survey online or by calling a toll-free telephone line. If a household does not respond using these methods a paper survey is then sent to the address. For households that do not respond to the questionnaire, data can be collected in person. Due to budget constraints only around one in three nonresponding households will be contacted in person by The American Community Survey.

Data Contents:
The data describes information about population demographics by State, provided as annual estimates between 2015-2019. Variables here include median income, unemployment rate, poverty rate by race/ethnicity, population distribution by race and citizenship status, indicators for education and the gini index.

Data Limitations and Bias:
The ACS provides both 1-year estimates and 5-year estimates. The date here is from 1-year estimates, which are published for areas that have populations above 65,000 people whereas the 5-year estimates include all areas. The 1-year estimate data is limited as it does not include information about rural areas. As this is manually entered survey data there will always be the chance of erroneous errors.

Why these Data Sets?

I chose these data sets after reading an article published on the website FiveThirtyEight titled "Higher Rates Of Hate Crimes Are Tied To Income Inequality". The article discusses whether it is possible to look to certain socioeconomic factors to understand why some states see more reported hate crimes than others. The analysis was predominately focused on the 2016 Presidential Election. As we seem to be becoming more aware of the issue of hate crimes happening in our communities, I thought it would be interesting to look at what data is available and what we can understand from it.

Cleaning & Understanding the Data

Uniform Crime Reporting Program (UCR) – Hate Crime Statistics

Dataframe shape:

- 209442 rows and 28 variables

Data Types of Columns:

- Found three columns with mixed data types, converted to most appropriate data types.
- Change data type of 'INCIDENT_ID' column to string as does not represent number with statistical importance.

Missing Data:

- The following columns are missing around 80% of values: ‘Adult Victim Count’, ‘Juvenile Victim Count’, ‘Adult Offender Count’ and ‘Juvenile Offender Count’. Will delete columns.
- The following columns are missing around 90% of values: ‘Pub Agency Unit’ and ‘Offender Ethnicity’. Will delete column.
- The following column is missing 92640 values, just under 5% of values: ‘Offender Race’. Will not address missing values as represents under 5% of total values so will not affect analysis.
- The following column is missing 2610 values, just over 1% of values: ‘Total Individual Victims’. Will not address missing values as represents under 5% of total values so will not affect analysis.

Duplicate Values:

- No duplicate rows found.

Descriptive Statistics:

- Mean, Min, Max values seem reasonable and do not suggest errors in data.
- High max value of 99 for offender count variable suggests default value, perhaps entered for riots/protests? Needs further exploration.
- See Jupyter Script for exploration of relevant variables.

United States Census Data - ACS 1-year estimates

I have downloaded multiple csv files for years 2015-2019 from several different tables from the ACS Census Data. I think it will be more efficient to combine all these individual files using Excel before I check their consistency and clean them in Python. I wasn’t sure if this was the right way to approach this additional data?

Questions to Explore:

- Have the number of hate crimes remained the same over time?
- Are the number of hate crimes the same in every state?
- What kind of hate crimes occur most frequently?
- Have the type of hate crimes and their frequency remained the same over time?
- Are most hate crimes committed against individuals or groups?
- Are most hate crimes committed by individuals or groups?
- Can we see a link between certain socioeconomic factors and the occurrence of hate crimes?

I may need to source additional data to answer the following questions:

- How many law enforcement agencies report hate crimes through the UCR?
- How many law enforcement agencies do not report hate crimes through the UCR?
- How many law enforcement agencies report ‘0’ hate crimes through the UCR?