Crime Rate and Weather Analysis in San Francisco
By: Cassandra Loh (USC ID 5653-6631-08)

Motivation

My main motivation behind this project is to see if a change in weather conditions would also result in a change in crime rates. Ever since the pandemic and lockdown, we've seen a steep increase in crime throughout the years and after reading up more on the heat hypothesis which mentions that higher temperatures could bring out aggressive behaviors in people, I wanted to see if a change in temperature would in any way affect the crime rate. Specifically, I've chosen San Francisco for its unique climate as it is much cooler and rains a lot more compared to a majority of California.

Description of Datasets

For the purpose of the project, I have chosen three different datasets which contribute to the final analysis:

1. **San Francisco Historical Weather Data (National Oceanic and Atmospheric Administration)**
    a. This dataset is provided as a csv file and the data ranges from January 2019 to January 2021.
    b. This dataset contains weather data obtained from the **San Francisco Downtown** Weather Station and will give insight into the minimum, maximum, and average temperatures, as well as precipitation levels.
2. **San Francisco Historical Weather Data (Wunderground)**
    a. This dataset is scraped from the website and the data ranges from January 2019 to January 2021.
    b. This dataset contains weather data obtained from the **South San Francisco** Weather Station and will give insight into the minimum, maximum, and average temperatures, as well as precipitation and wind speed.
3. **San Francisco Police Department Crime Incident Reports (US Government Open Data)**
    a. This dataset is obtained through the external API provided on the US Government's Open Data platform and ranges from January 2019 to January 2023.
    b. This dataset provides us with valuable information about the crime reports made between the specified date range, neighborhood of crime, zip code, type of crime and the time the incident occurred.
4. **Combined Dataset (For Analysis)**
    a. This dataset is made by combining certain columns from each of the datasets mentioned above and the key column between all the datasets is the date.
    b. Columns:
        i. **Date**

ii. **RAIN** – Taken from the NOAA.csv dataset
iii. **Average Temp (NOAA)** – Taken from the NOAA.csv dataset
iv. **Average Temp (Wunderland)** – Taken from the weather.csv dataset
v. **Number of Crimes** – Taken from the incidents.csv dataset
vi. **Average Temperature (Combined)** – Made by averaging out the values from column iii & iv.

**The reasoning behind using two sources for historical weather data is because one station is located up North and the other is down South, so by taking the average from the two stations and averaging them out again, we hope to obtain a more accurate representation of the average temperature.

Analysis and Results

I came into this project with 2 hypotheses in mind:

1. As the temperature increases, the rate of crime should also increase.
2. There will be less crime on rainy days compared to sunny days because there is less foot traffic outside on rainy days.

To see if my hypotheses were correct, I employed the use of:

- Two tailed t-test
- Pearson's Correlation
- Kendall's Correlation
- Spearman's Correlation

The aim was to see if there would be a positive correlation between temperature and crime rate and a negative correlation between rain and crime rate as well as its statistical significance.

- **Two Tailed t-test**
  - Temperature and Crime Rate
    - The analysis was carried out using the pandas and scipy libraries in Python.
    - The data was split into two by assigning the temperature threshold as 65 degrees Fahrenheit. Anything above 65 degrees would be considered hot and anything below would be considered cold.
    -

Temperature (Low vs. High) and Crime Rate	
t-statistic	1.24120
***p*-value**	0.21491

    - The t-statistic above indicates a positive correlation between our variables; however if we take a closer look at our *p*-value, since it is greater than 0.05, we aren't able to say that this relationship is statistically significant.

        - With the insufficient evidence, we are not able to reject the null hypothesis (H_0) that there is no relationship between temperature and crime rate.
    - Rain and Crime Rate
        - The analysis was carried out using the pandas and scipy libraries in Python.
        - Since the variables in the dataset were categorical for rain (Yes/No), we assigned the value of Yes = 1 and No = 0 and conducted the analysis.
        -

Rain and Crime Rate	
t-statistic	-0.39181
***p*-value**	0.69556

        - The t-statistics above indicates a negative correlation between rain and crime rate; however if we look at our *p*-value of 0.69556, since it is greater than 0.05, we aren't able to say that this relationship is statistically significant.
        - Our t-statistic also indicates that the mean number of crimes on rainy days are less when compared to non-rainy days.
        - With the insufficient evidence, we are not able to reject the null hypothesis (H_0) that there is no relationship between rain and crime rate.

- **Correlational Testing**
    -

Temperature and Crime Rate			
	Pearson's Correlation	**Spearman's Correlation**	**Kendall's Correlation**
Correlation Coefficient (*r*)	0.03827	0.02995	0.02183
***p*-value**	0.29135	0.40910	0.36883

        - Looking at the results above from the three different correlation coefficients, we are able to see that all of them show a positive correlation between temperature and crime rate. However, the *p*-value for all three of the coefficients were greater than our cutoff of $p > 0.05$ so we aren't able to conclude that this correlation is statistically significant thus we cannot confidently reject the null hypothesis (H_0) that there is no relationship between temperature and crime rate.

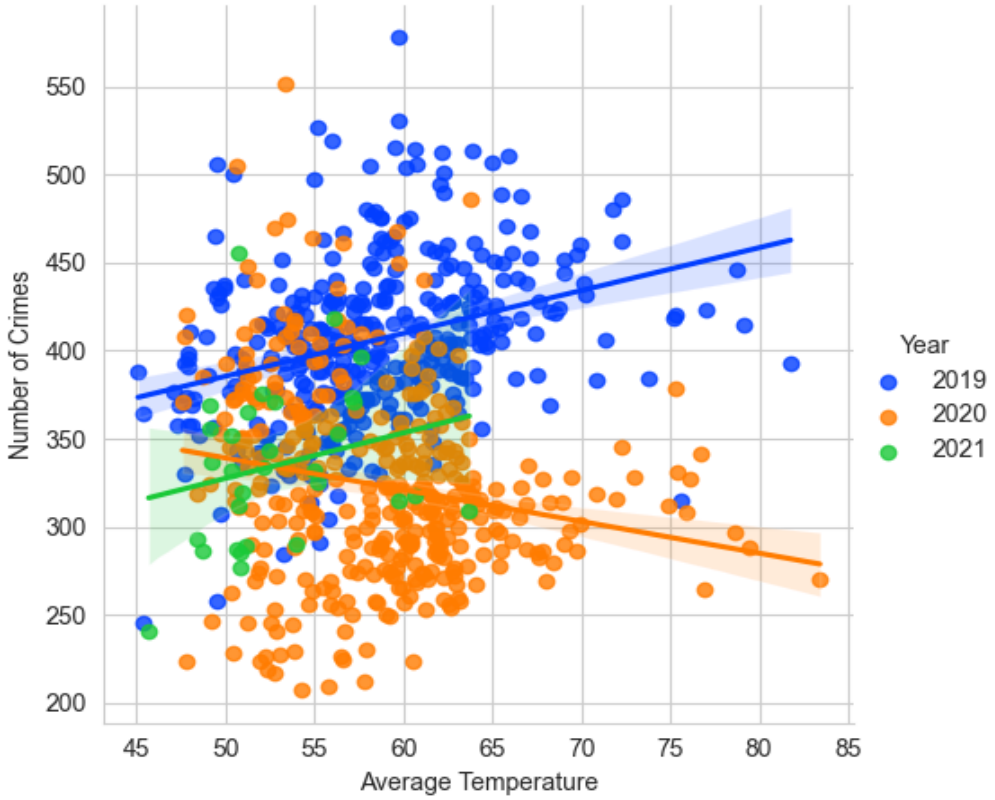
    -

Rain and Crime Rate			
	Pearson's Correlation	**Spearman's Correlation**	**Kendall's Correlation**
Correlation Coefficient (*r*)	-0.01360	-0.00147	-0.00121
***p*-value**	0.70777	0.96759	0.96756

- Referencing the results above from the three different correlation coefficients, we are able to see that all of them show a negative correlation between rain and crime rate whereby there is less crime on rainy days compared to non-rainy days. However, the *p*-value for all three of the coefficients were greater than our cutoff of $p > 0.05$ so we aren't able to conclude that this correlation is statistically significant thus we cannot confidently reject the null hypothesis (H_0) that there is no relationship between rain and the rate of crime.

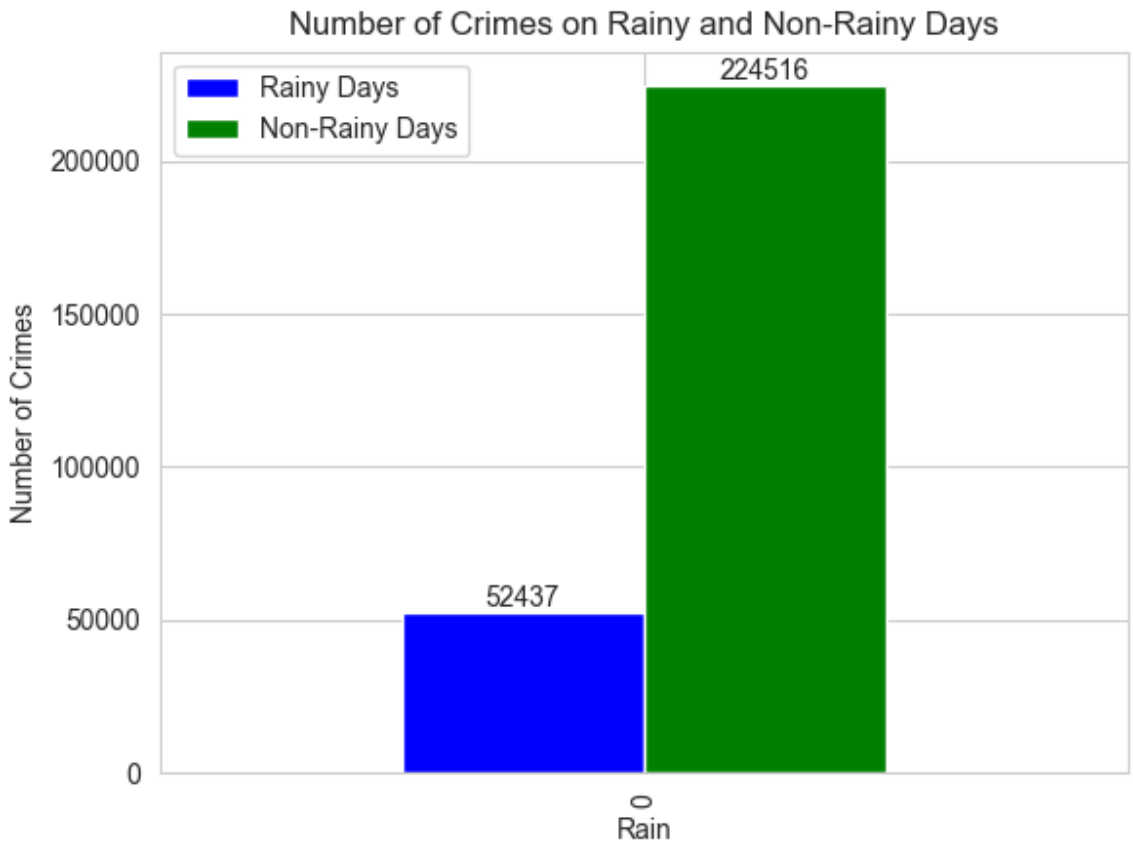
Supporting Visualizations

Scatter Plot of Temperature vs. Crime Rate



This Scatter Plot was created with the help of the pandas, matplotlib and seaborn libraries in Python. It shows our data points from 2019 to 2021 mapped out on a graph and we can see that a majority of the data points are between 50 degrees to 65 degrees Fahrenheit.

Bar Graph of Rain vs. Crime Rate



This Bar Graph was created with the help of the pandas, matplotlib and seaborn libraries in Python. In the graph, we can see that the mean number of crimes on rainy days are almost ¼ of the crimes on non-rainy days and the exact number of crimes labeled on top of each bar.

Conclusion & Recommendations for Future Research

In conclusion, our analysis revealed a positive correlation between temperature and crime rate whereby as the temperature increases, so does the number of crimes. However, we do not have sufficient evidence to reject the null hypothesis. Our analysis also revealed a negative correlation between rain and crime rate where the mean number of crimes on rainy days are less compared to that of non-rain days. However, since our *p*-value was greater than 0.05, we aren't able to conclude that this relationship is statistically significant and therefore cannot confidently reject the null hypothesis.

Direction for future research:

- Group the crimes by types (e.g., burglary, theft, arson etc.) and run a multiple regression to see if a specific type of crime is correlated with a change in temperature.
- Look into temperature on a more granular level (by zip code) to see if there is a correlation between the crime and weather in a specific geolocation. San Francisco has unique microclimates all over the city so it might be 50 degrees at the Pier but if you take the bus downtown 10 minutes away, the temperature might be 70 degrees there.

Although our results are currently statistically insignificant, it might prove worthwhile to look into it on a deeper level starting with the two recommendations above as we can use the results to make recommendations for a change in policy to the relevant departments to potentially reduce the rate of crime in the city.

Brief Python Script Description (See README.txt for more in-depth explanations)

- This main.py script can be run in four different modes:
    - **Default mode (Command - python3 main.py)**
        - In default mode, the script will fetch data from the API, scrape data from a website as well as load data from a csv file. Once that is completed, it will ask for user input:
            - **Would you like to combine the dataset? (y/n)**
                - If the user answers yes, it will combine the dataset to create combined.csv and move onto the second user input:
                    - **Would you like to run analysis on the dataset? (y/n)**
                        - If the user answers yes, it will run the analysis programmed and print the outputs respectively in the terminal.
                - If the user answers no, the code will break and stop after scraping/fetching/loading the datasets.
    - **Scrape mode (Command - python3 main.py –scrape)**
        - In scrape mode, the script will perform the scraping and fetching of datasets from the API and print out the 10 rows along with the column header from the dataset. **No analysis will be performed in scrape mode, it will just gather the dataset.**
    - **Static mode**
        - In static mode, there are four commands that can be used to print out a portion of each dataset:
            - python3 main.py –static ./data/noaa.csv
            - python3 main.py –static ./data/weather.csv
            - python3 main.py –static ./data/incidents.csv
            - python3 main.py –static ./data/combined.csv
    - **Analyze mode (python3 main.py –analyze)**

- In analyze mode, the pre-programmed analysis (t-test, Pearson, Kendall's and Spearman's Correlation) will be performed on the dataset combined.csv.
- The results of each test will be printed in the terminal.

Extensibility

This Python script has several components and is used for scraping and fetching specific datasets. It is designed to fetch data from the Weather Underground and San Francisco Police Department (SFPD) incident datasets. The script provides several functionalities, such as loading and saving datasets from/to different file formats, displaying datasets in a table format, and fetching data from APIs or by scraping websites. This script also has the capability to create a new dataset by taking the information from specific columns and writing it into a new file. Correlational analysis will then be performed on the file and supporting visualizations will be generated and saved.

Maintainability

In the event that the website or REST API changes in the future, the Python script will need to be updated accordingly to adapt to the new structure or endpoint. The following are some general guidelines on how to modify the code:

1. Update the URLs and endpoints: If the URLs of the websites being scraped or the endpoints of the REST APIs have changed, you will need to update the corresponding variables in the script. For example, if the URL of Weather Underground's historical data changes, you should modify the `URL_WUG_MONTHLY` variable accordingly.
2. Update the parsing logic: If the structure of the website's HTML or the JSON data returned by the API changes, you will need to update the parsing logic. For instance, if the Weather Underground website changes the way it organizes its tables or adds new columns, you may need to modify the `WUGWeatherDataset.fetch_data` method and adjust the BeautifulSoup selectors to extract the correct data.
3. Update the data processing and transformation: If the data format or schema changes, you might need to update the data processing and transformation logic. For example, if the incident dataset starts returning additional fields, you may need to update the `IncidentDataset.fetch_data` method and adjust the DataFrame processing steps.
4. Handle new or deprecated features: If the website or REST API introduces new features or deprecates existing ones, you may need to update the script to incorporate or remove the related functionality. For instance, if the incident dataset API starts requiring authentication, you might need to add code to handle authentication and update the API request logic accordingly.
5. Update error handling: When the website or REST API changes, you may need to revise the error handling mechanisms to account for new error codes or messages. This could involve updating the conditions in exception handling blocks or adding new ones to handle specific errors.

6. Keep documentation up-to-date: It is crucial to keep the documentation up-to-date with the changes in the code. This helps users understand the new functionality, and it also serves as a reference for future developers who might need to maintain or modify the code.

When a website or REST API changes, it is important to thoroughly review the code and make necessary adjustments to ensure the script continues to function correctly. By following these guidelines, you can effectively adapt the code to handle changes and maintain the script's functionality.