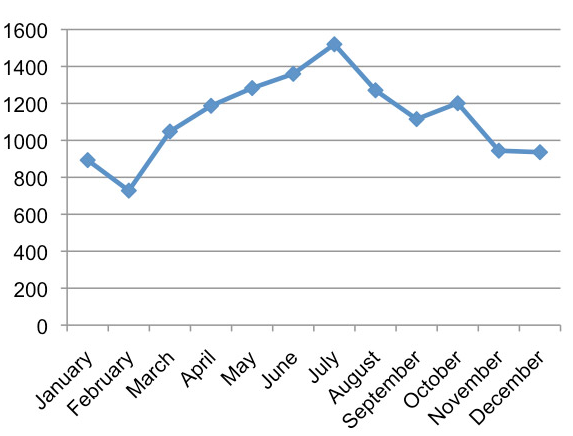
**SUBJECT:** Technical Report: Extract-Transform-Load Methodology Exploring the Relationship Between Crime and Weather

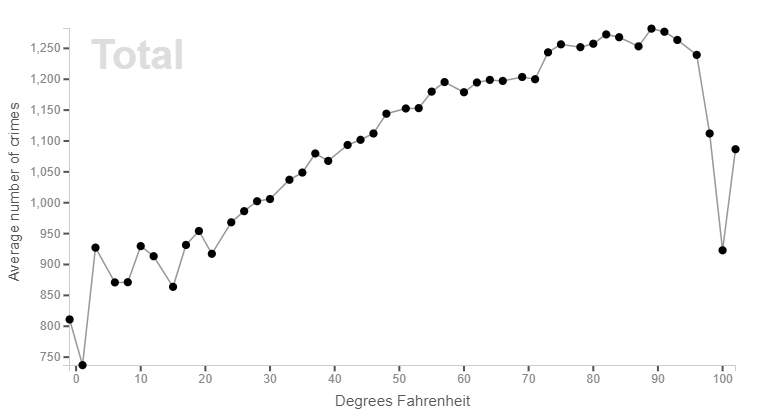


**BACKGROUND:** It’s accepted wisdom that crime gets worse in the spring and summer. But the relationship between crime and temperature is even closer than that, and the questions are many: does heat make you angry and stupid? Is there a point where it gets too hot to commit crime? And what happens after the heat wave?

News stories indicates every year it’s the same trend, and the same conclusion (see Chicago Times summary of heat related violence (<https://www.chicagotribune.com/news/breaking/chi-cops-at-least-half-dozen-wounded-across-city-including-12yearold-boy-20120314-story.html>).

The common pattern throughout the United States shows: violent crimes go up during the spring, peak in the summer, decline during the fall, and reach a bottom during the winter. When it comes to explanations, it gets complicated.

Existing analysis has examined the correlation between **temperature** and **crime rate** in Chicago. This Extract-Transform-Load effort seeks to provide structured data in a format to support follow-on analysis to examine the details of this relationship.



**THE ASK:** Develop a Methodology to Extract-Transform-Load data sources in support of a follow-on analysis to explore the relationship(s) between crime and weather. Analysis will look to identify which categories of crime (i.e. category – Drugs, Assault, Theft, …) are strongly correlated to specific factors of weather (i.e. temperature, precipitation, visibility, …). The hypotheses to be examined:

Ho: Specific Crime Categories = Unique Weather Factors

Ha: Specific Crime Categories < > Unique Weather Factors

**DATA SOURCES:** An initial investigation found multiple data sources in CSV format for crime and weather. The most complete crime data set was a daily crime log for Chicago and the most convenient weather data was available in the form of a weekly aggregate of weather for Chicago.

Weather: The analysis team made use of past weather data from virtualcrossing.com which provides a summary of weekly weather for a specific location.

Crime: The crime dataset at Kaggle’s **BigQuery,** a web service that supports the application of SQL to huge datasets (like the Chicago Crimes Dataset).

**METHODOLOGY:**



**EXTRACT:** The most significant obstacle found during the EXTRACT phase was the difference between DATE fields in the two data sources. Both data sources were read-in to Pandas using pd.read\_csv. The size of both datasets led the group to explore conducting the transform phase of this project on Kaggle’s BigQuery site which supports the user with a unique Kernel (similar to Jupyter Notebook).

**TRANSFORM:** The Weather data set was the easiest to transform by dropping some columns and stripping the DATE field into WEEK MONTH YEAR fields. The Crime data set required a similar action (drop columns and strip DATE). The end state of these actions was a Primary Key in each data set that can be joined.

The most significant transform work was in iterating through the now joined DataFrame using groupby to aggregate each of the Crime categories for each week while retaining details on location (i.e. gas station, warehouse, sidewalk, etc.). This effort expanded an already large data set by an order of magnitude which is more reasonable through BigQuery. The final transformed DataFrame was written to a CSV file using df.to\_csv.

**LOAD:** Since the transformed data set was relational the csv file was then loaded into a Postgres data base (PgAdmin4). This stage of the project relied on an entity relationship diagram (ERD) that was outlined during the extract/transform phase. A schema was developed to support the loading of the csv file into PgAdmin4 and the csv file was imported

**OUTPUT:** The end state for this specific Extract-Transform-Load project is this technical report and the SQL database to support follow-on analysis.