Traffic Collisions : Montgomery Dataset Analysis

Aasim Inamdar  
*Master of Science in Data Analytics, School of Computing,*   
*National College of Ireland*Dublin, Ireland  
[x23236108@student.ncirl.ie](mailto:x23236108@student.ncirl.ie) Sana Shafiq Jalgaonkar  
*Master of Science in Data Analytics, School of Computing,*   
*National College of Ireland*Dublin, Ireland  
[x22237941@student.ncirl.ie](mailto:x22237941@student.ncirl.ie)

Waleed Bin Umer  
*Master of Science in Data Analytics, School of Computing,*   
*National College of Ireland*Dublin, Ireland  
[x23187956@student.ncirl.ie](mailto:x23187956@student.ncirl.ie)

*Abstract*—Traffic safety is becoming a major concern for all the communities throughout the world, including Montgomery County in Maryland state of the USA. To give a thorough picture of the traffic collisions, the study examines the effectiveness of considering and analyzing three main datasets of an Automated Crash reporting System – ACRS used by Montgomery County to track the details about day-to-day traffic incidents within the county. The study presents an in-depth analysis of traffic collision data of Montgomery County. This analysis uses an Extract, Transform and Load (ETL) pipeline, following which the study does Exploratory Data Analysis and visualizations to inform and conclude urban planners on to develop targeted strategies for minimizing the impact of traffic collisions, and thus enhancing public safety and transportation facilities.

The Integrated analysis of the ACRS datasets utilizes MySQL as the source database, Python Pandas as the Staging area, and MongoDB as the destination. After the ETL pipeline, the study does the visualizations with Python’s pandas’ library and other visualization tools like matplotlib and seaborn. The study aims to find some insights into traffic collisions into aspects such as weather conditions, drivers characteristics and road conditions.

Keywords—Traffic Collisions, Montgomery County, Python, ETL, Visualizations, API, Automation, ACRS, MongoDB, MySQL

# Introduction

Traffic incidents are the major public health and safety concern around the world, resulting in injuries, deaths and financial damages. With developments in vehicle safety and traffic control systems, incidents continue to happen at an disturbing rate. Solving this problem needs an innovative strategy based on thorough data analysis and rational decision-making.

The Montgomery County’s datasets which are obtained from the Automated Crash Reporting System (ACRS) provided by the county’s government offers to be a valuable resource for studying the incidents in-depth. The datasets provides plenty of information on incidents, those involved as well as the contributing factors. With this datasets, studies can obtain the understanding of the occurrence of the collisions and foster specific measures to improve road safety.

To help the study, an ETL meaning Extract, Transform and Load pipeline that uses a relational database- MySQL as the source database followed by Pandas as the staging area for the transformations and MongoDB as the destination database. The overall flow of the pipeline provides a seamless integration of direct data obtained from APIs and direct files, including manipulation of the data ensuring data integrity and consistency throughout the process of analysis.

# Related work

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# methodolgy

## Datasets Selection

This study’s datasets comes from the official government site of Montgomery County [1] which maintains the data into various format like JSON, CSV, APIs that are collected by the Automated Crash Reporting System (ACRS) within the county. These datasets have been selected as it tracks all the information on the collisions in the County and provides the complete information required for the thorough analysis.

## Datasets Descriptions

* Incidents Dataset: Provides general information about each incident as well as traffic information that occurred in Montgomery County, collected via Automated Crash Reporting System (ACRS). The dataset has been extracted from the API provided on the official government site of the county which is powered by Socrata [4] . The dataset identifier for this dataset is Dataset Identifier: bhju-22kf. The total number of rows are 97458 which can be exported as a CSV file without API. As for API extraction via Socrata that rows limit is 1000. Each row is collision record.
* Drivers Dataset: Contains information about the drivers engaged in the traffic incidents on the county and local roadways in the Montgomery County as collected by the Automated Crash Reporting System (ACRS). The dataset has been extracted from the API provided on the official government site of the county which is powered by Socrata [4] .The dataset identifier for this dataset is Dataset Identifier: mmzv-x632. When exported the total number of rows are 172105 and 43 columns into a CSV file. As for API extraction via Socrata that rows limit is 1000. Each row is represented as a driver record.
* Non-Motorist Dataset: Includes information about the non-motorists like pedestrians and bicyclists engaged in the collisions on the county and local roads in the Montgomery County , tracked via Automated Crash Reporting System (ACRS). The dataset is directly exported from the official government site of the Montgomery County. The total number of rows are 5650 with 32 columns, where each row is represented as non-motorist.
* Common Features in all three datasets:

## Tech Stack

* Python: is used as a primary language and a medium for handling an analysing data due to its simplicity and flexibility. The study uses many visualizations and processing libraries of python like NumPy, Pandas, Matplotlib, Seaborn. As for linking with source an target database, the study uses pymongo to establish a connection with MongoDB and mysql-connector-python for getting the language linked to MySQL database.
* MySQL: A relational database link MySQL is the source of data flowing from API and exported data from the official government site of the county as it’s well suited for the applications with complex queries.
* MongoDB: A NoSQL database like MongoDB is the target that is the destination database where the transformed and cleaned data is stored as it is free from any predefined schema requirement.
* Socrata API: an API provided by Socrata, enables developers to interact programmatically with datasets hosted on the platform by the detailed documentation, code samples, and developmental resources to assist users in getting started with the API.
* Luigi: a strong python utility used for creating an organization of data pipeline by defining workflows as tasks.

## Data Architecture

Figure 1. represents the actual lifecycle of the process including the ETL pipeline and visualizations.

* The data for Incidents and Drivers is extracted via Socrata API in the form of JSON which is converted using pandas into a CSV file. As for Non-Motorists, the data is exported into a direct CSV file.
* These 3 CSVs – the original datasets are integrated into MySQL which is the source database in the ETL pipeline.
* Using Python, the connection is established with MySQL to extract data from tables to data frames which acts as a staging area where all the transformation takes place.
* Once the transformations are completed, 3 intermediate JSON files for each of the data frame is created which are loaded as 3 collections into MongoDB which is the target in the established pipeline.
* Finally for getting insights and performing final analysis the cleaned and processed data from MongoDB is fetched into data frames whereby using visualization libraries insights are generated.

## Preprocessing Methods

#### Incidents: jhbnb

#### Drivers: jbn ,.m

#### Non-Motorists: bhbhvbhjsb

# Results and Evaluation

After the text edit has been completed, the paper is ready for the template.

Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. R. Hasarinda, T. Tharuminda, K. Palitharathna and S. Edirisinghe, "Traffic Collision Avoidance with Vehicular Edge Computing," 2023 3rd International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 2023, pp. 316-321, doi: 10.1109/ICARC57651.2023.10145607.
2. Z. Luo, J. Li and M. Zhong, "Prediction of Seasonal Variation in Traffic Collisions on Urban Highway: A Case Study in the Province of British Columbia," 2021 6th International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 2021, pp. 104-109, doi: 10.1109/ICTIS54573.2021.9798674.
3. D. Denti and M. Modica, "Born to Hit and Run? Rugged Individualism, justice quality and non-stopping after traffic collisions in the U.S.," Case Studies on Transport Policy, vol. 15, p. 101137, 2024. [Online]. Available: <https://doi.org/10.1016/j.cstp.2023.101137>.
4. <https://dev.socrata.com/foundry/data.montgomerycountymd.gov/mmzv-x632> for Drivers API
5. <https://dev.socrata.com/foundry/data.montgomerycountymd.gov/bhju-22kf> for Incidents API

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**