

# Pedestrian Safety -- Fundamental to a Walkable City

Joshua Herrera

Patrick McDevitt

Preeti Swaminathan

Raghuram Srinivas

---

## *Review*

---

This review is based on the guidelines provided in the MSDS 6130: Capstone B Course Syllabus and according to the MSDSJournalPaperTemplate.

### **Abstract**

Current Abstract length: 142 words. Consider expanding it to 200 words.

Executive summary: Fine

In the following sentence consider writing out five instead of 5:

“To develop a solution to this problem of identifying likely locations of events, we assemble data sets, primarily from the City of Cincinnati, that include safety reports from a 5 year period, geographic information for these events, a citizen survey that identifies pedestrian reported concerns, and a database of all requests for service for any cause in the city.”

### **Introduction**

Current Introduction length: 1.3 pages. Consider expanding it to 2-4 pages.

Executive summary: The introduction presents and justifies the problem in a well-written form, it provides enough references and literacy which supports the need to attend a significant and public problem. The authors however might consider expanding on the work, methods and solutions developed and presented in their research.

In the following sentence consider changing characteristics:a to characteristics: a

“A walkable neighborhood is a neighborhood with the following characteristics:a center (either a main street or public space), sufficient population density to support local businesses and public transit, affordable housing, public spaces, streets designed for bicyclists and pedestrians, and schools and workplaces within walking distance for residents [3]”.

Footnotes: consider reformatting footnotes according to the guidelines provided in MSDSJournalPaperTemplate.

## Tutorial Sections

**Pedestrian Safety:** domain and variables are well explained. Tables include relevant content in an orderly format. Authors might consider expanding and visualizing context further. Geographical, Demographical, Economical and Historical data pertaining to the city of Cincinnati might help readers to gain comprehensive knowledge of the city and the nature of the problem to be resolved.

**Data Sets:** Consider adding the source of the following data sets “Cincinnati pedestrian safety survey data” , “Pedestrian near miss data” and all of the Cincinnati open data datasets.

**Methods and Experiments:** data collection and ingestion is well explained to meet the assumptions of a parametric model. Outliers were removed to reduce skewness. Dimensionality reduction is also applied. The model selection is well described allowing a knowledgeable person reading the paper to be able to reproduce the results if needed to.

Authors might consider expanding on the deterministic analysis derived from “With the reduced model, the cost of non-fatality incidents to pedestrians is predicted for each event, and the resulting prediction is compared against the actual cost of incident. The difference between the predicted and actual cost is the residual, and the residuals are then mapped back to the Cincinnati map to visualize areas with potential for safety improvement”.

**Non-Supervised Learning - Neighborhood Characterization:** Cluster analysis and selection is well described however authors might consider referencing the code and data set used to plot Fig. 1: Spread of data visualized using t-sne. Consider referencing an Appendix.

Also consider correcting typo in the following sentence: “Fig. 1: Spread of data visualized using t-sne.”

## Results

Data is described including category and feature definition.

Authors might consider adding references to the following visualizations e.g. Appendix to explain how the visualizations were created, also refer to any code, data, platform used.

Fig. 2: Walk Score - Max values

Fig. 3: Public Transportation accessibility

Fig. 4: Number of Traffic Accidents (Non-pedestrian involved)

Authors might consider correcting typo in “Fig. 3: Public Transportation accessibility”.

**Random Forest - Binary Model:** The section describes the purpose of the selected model and its results. Authors might consider justifying however the model of their choice as well as providing references to “Fig. 6: Distribution of random forest model results visualized on city map” which would explain how this visualization was created, e.g. an Appendix including data, code and platform.

**Multi-Variate Linear Regression - Cost Model:** The section describes regression variables and their values. Authors might consider providing further details on the model itself, its analysis and interpretation. Also providing references to “Fig. 8: A map of Cincinnati identifying areas with PSI in the darkest red.” which would explain how this visualization was created, e.g. an Appendix including data, code, and platform.

Non-Supervised Learning - Neighborhood Characterization: The section describes attribute relationship in regards to clusters. It also describes how pedestrian safety issues in each cluster are identified. Authors might consider expanding on how the issues are actually mapped to their locality. Also correcting the typo in the following sentence: “Combining attribute behavior and clusters on the map, pedestrian safety issues in each cluster are identified”. Authors might also consider providing references to “Fig. 9: Cluster results on map of Cincinnati” and “Fig. 10: Sum cost of traffic crashes reported from 2012 to October 2018 involving pedestrians by grid cells.” which would explain how this visualization was created, e.g. an Appendix including data, code, and platform.

## **Analysis**

Random Forest - Binary Model: The analysis seems supported by the presented results. Authors might consider providing reference to the following sentence: “Consistent with the concept as defined and implemented in other works” .

Multi-Variate Regression - Cost Model: The analysis seems supported by the presented results. The scope of the model is limited to non-fatality incidents. Authors might consider expanding the context of statistics and demographics which would assess the significance of the model. The authors might also consider expanding on the following assertion: “These areas within the city in which the model has underestimated the cost the most have the most potential for safety improvement, because all other factors being equal, there is something about that area which is causing a higher amount of damage to pedestrians.”

Non-Supervised Learning - Neighborhood Characterization: The analysis seems supported by the presented results. However the write up seems also identical to section “5.4 Non-Supervised Learning - Neighborhood Characterization” with the exception of a couple of words and punctuation marks. The authors might consider either adding new content to be distinct from section 5.4 or quoting themselves instead.

## **Ethical Discussion**

The ethical discussion is more than one page in length. It also follows the principles stated in the ACM Code of Ethics. It recognizes the risks and possible unintended future use of the results presented as part of the conducted research. It disassociates from any initiatives that may surge from it which are by nature far from intending pedestrian safety, even those directed to achieve individual gains in the form of either private or public sectors.

The authors might consider expanding on the positiveness of the research and the results presented as well. They may also consider expanding on the ethical values that qualify the presented results as means to socially respect, protect, and serve others.

## **Conclusions**

The authors might consider writing up the following paragraph again as it is not clear:

“Our unsupervised learning techniques... Placeholder sentences about how our unsupervised learning techniques have outlined areas which have potential for safety improvement. These areas demonstrate these qualities which are related to pedestrian safety in the city of Cincinnati, Ohio.”

The authors might consider expanding the tutorial sections accordingly to demonstrate:

“For our unsupervised learning we look at the top pedestrian safety issues in each cluster. Vehicles running red lights or stop signs is a concern for the blue cluster neighborhoods, including: Westwood, Mt. Washington and East End. Heavy traffic is a prominent issue in the green cluster and affects Over-The-Rhine, Mt. Auburn and Central Business District neighborhoods. There is a need for increased number of sidewalks in Clifton,

Madisonville, Walnut Hills and Mt. Washington. Over-the-Rhine, Downtown and Clifton need more bike friendly zones.”

The authors might consider revisiting the following paragraph, as it appears to be incorrect:

“Future work to further this project could involve the use of other regression modelling techniques to improve accuracy, such as tobit models. Other improvements which could be made include Stuff goes here from Patrick and Preeti about ways to improve your techniques and models. “

## **Footnotes**

The authors might consider, for the overall document, reformatting footnotes according to the guidelines provided in the MSDSJJournalPaperTemplate.