1. Introduction
    a. “As a result of this trend, there have been real, and negative, consequences…”
        i. Too many commas
        ii. Here are grammatically correct alternatives:
            1. “There have been real and negative consequences resulting from this trend…”
            2. “There have been real, negative consequences resulting from this trend…”
            3. “Real, negative consequences have resulted from this trend…”
    b. “So much so that working families are currently spending more of their budget on transportation than housing.”
        i. This is hard to believe without citation. You are telling the reader that most households spend less than or equal to 20% of their income on housing?
    c. “This statistic, coupled with the fact that Americans now walk the least of any industrialized nation in the world [5] indicate a growing health problem due in part to a lack of physical activity.”
        i. You cite correlations, but you conclude causation. Are you sure your sources draw these same, strong conclusions about causation?
    d. “drive-till-you-qualify model”
        i. While I understand what this phrase implies, much of your audience may not. Consider further explanation, i.e. “…traveling further and further from the city center until finding affordable housing options.”
    e. “The US is experiencing an increase in the number of pedestrian fatalities, reaching a 25-year high in 2017, with nearly 6,000 fatalities…”
        i. The U.S. population is increasing over that same time period. How does your reader know you’re not cherry picking statistics? Try controlling for population growth AND urbanization. Something like “fatalities per capita per unit of area”.
    f. “Midwest identify fatal occurrences: [12] "An”
        i. Your opening quotation mark (after the ‘[12]’) prints backward in the PDF version of your paper
    g. "An uptick in pedestrians being hit by cars in the Cincinnati and Northern Kentucky area has officials sounding the alarm.”
        i. Is this a statistically significant uptick? Or is it random variation (i.e. irreducible noise)? It feels like you are making your case on the back of anecdotes, not hard, scientifically rigorous evidence.
    h. “which launched in Feb-2018”
        i. This is definitely not the right date format for a formal paper.
        ii. I would think an exact date (i.e. February 1, 2018) or a month and year (“February of 2018”) would be more appropriate
    i. Your introduction is incomplete
        i. Engels provided a solid template for authors to follow (“How to Write a Research Paper”). I would suggest you attempt to follow it unless you have a compelling reason not to:
            1. 1 Paragraph Motivation (sets the general problem domain)
            2. 1 Paragraph Problem Statement (specific problem solved by the work)
            3. 2-3 Paragraphs on solution MISSING
            4. 1 Paragraph on main results (plural) MISSING
            5. 1 Paragraph on main conclusions (plural) MISSING

6. 1 Paragraph on paper organization MISSING

2. Pedestrian Safety
    a. Table 1
        i. I think you are citing the wrong document. I pulled the document you cite (https://c.ymcdn.com/sites/www.safestates.org/resource/resmgr/evaluation_resources_webpage/estimating-the-costs-of-unin.pdf) and it lists much lower costs. However, when I pull the 2015 NSC report, it matches the number you cite.
        ii. You should probably provide a bit more context to these costs. For instance, these are not the average of realized costs ("economic cost"), but rather "virtual" costs associated with what people are willing to pay for quality of living improvements
        iii. You citation is screwed up. You say it is the 2012 report in both the in text reference and the title of the citation in your reference section, yet the link you provided (which doesn't work) says it is the 2014 report.
    b. You need to put more space between the two columns in Table 2. In its current state, it appears that the Description begins "Im-Measures"
    c. Actual Cost
        i. How are you getting these values? Aren't you actually using modeled values? There is no way for you to know what the actual economic and quality-of-living costs are for each incident, right? This is pretty misleading.
3. Data Sets
    a. "Feb 2018 to April 2018"
        i. You are being inconsistent in your date formatting. Why is "Feb" abbreviated but "April" is not?
    b. "The user then selected a neighborhood, and selects"
        i. You are changing tense mid-sentence. Keep it consistent: "selected" or "selects"
    c. Table 3
        i. "…Data is not granular"
            1. To which row does this statement belong? Improve your table formatting so this is clearer.
    d. "…City of Cincinnati are listed in Table 3. Datasets that originated from the Open Data Cincinnati"
        i. What is the difference between these two sources? You write about them as though the reader has fore-knowledge about them. S/he doesn't.
    e. Walk Score
        i. I don't fully understand how this is relevant or how it relates to pedestrian safety? Is it a proxy for how much walking occupants of a given area are assumed to be doing? The website says it is a measure of how many amenities are within walkable distance.
4. Methods and Experiments
    a. "4.1 Data Ingestion and Grid-Cell Aggregation" and "4.2 Random Forest - Binary Model" are presented as headers but there is no text following them. These headers are repeated in section 5 ("Results").
    b. "…must lend itself to the assumptions of a parametric model."
        i. Different parametric models have different assumptions, and you are using a particular parametric model (linear regression), so why are you using the

indefinite article “a”, and not the definite article “the”. This whole clause feels superfluous. Consider deleting.

c. “…notably including the logarithmic transformation.”
    i. Why is this noteworthy?
d. “In order to meet the assumption of normally distributed data, the data is subset to exclude locations where there is no data, and the few events resulting in extreme damage of $5 million or more. This data subset dramatically reduces the skewness present in the dataset, and allows for linear regression modelling to proceed.”
    i. So you are excluding data exclusively for the purposes of forcing it to meet the assumptions of a linear model? This feels dangerously close to some kind of data dredging. It is very high bar that must be cleared in order to exclude data, and it doesn’t appear as though you’ve met it. Why wouldn’t you first seek to find a more appropriate model?
e. “The 10-fold cross validation splits the data into 10 pieces”
    i. Why wouldn’t you call them “folds” or “splits” since you already use those in this same sentence and they are the more generally accepted terms with respect to C.V. partitions?
f. “In each of the 10 iterations, one is selected to be the testing dataset, and the other 9 are used to train the model.”
    i. “One” what? This is bad grammar and results in a confusing sentence.
g. “10 trials are averaged”
    i. These are not trials. Trials has a specific meaning in statistics. These are iterations in a cross validation scheme.
h. “parametric assumptions in linear regression”
    i. Which assumption? Are you sure it is a severe enough violation to warrant eliminating the variable
i. Why bother with an procedural method for variable selection/elimination if you are going to then manually override it?
    i. I’m not suggesting it shouldn’t ever be done, but your approach doesn’t appear to be very systematic or theoretically grounded
j. “p-value, and the typical value used is .05.”
    i. Used for what? This is a fundamental concept of statistical hypothesis testing and you are flubbing its explanation, which doesn’t inspire confidence in your paper overall.
k. “With the reduced model”
    i. What is the reduced model? You haven’t introduced this concept/vocabulary.
    ii. You are using the terms “reduced model” and “full model” in an unusual manner. The reduced model is the simpler model and the full model is the more complex model. So far, so good. However, the reduced model is supposed to correspond to some null hypothesis, and the full model is supposed to correspond to an alternate hypothesis. That is not the case here. Both models correspond to an alternate hypothesis. I would consider changing your y
    iii. Did you perform a test to prove that one model is better than the other (for example, an *F*-test)? Or why is the reader to believe that the model with fewer variables is superior?
l. “Fig. 1: Spread of data visulized using t-sne.”

5. Results

a. 5.1 Data Aggregation and Grid Cell Assignment
    i. How is this a result? This looks more like basic EDA. You plotted data onto a map.
    ii. Figure 4
        1. Why do you have a title above and below this figure? The "above" title is not using the correct font.
        2. Same issue w/ figures
b. 5.2 Random Forest
    i. "75% true positive, 20% false positive"
        1. What?? What are the denominators of these percentages? Why don't they sum to 100%? Why don't you just provided a confusion matrix?
    ii. "detection threshold set at 0.20"
        1. Detection threshold: This is usually referred to as a "classification threshold" or simply "cut point".
        2. 0.20 what? The probabilities assigned by your random forest model? Make this clear.
    iii. Figure 5
        1. You should mark the location on the ROC curve corresponding the classification threshold you chose, thereby demonstrating to the reader that you have made the optimal choice.
    iv. Figure 6
        1. Your choice of colors is baffling. Why would you choose two shades of grey for separate categories?? Why would you choose ANY shade of grey when the background of the plot area is grey??
        2. What is the point of this plot? What is it demonstrating to the reader? How is this plot advancing your narrative? How is this a result at all?
c. 5.3 Multi-variate Linear Regression – Cost Model
    i. "Figure 7 is a comparison between the observed and actual cost of a non-fatality incident in Cincinnati."
        1. "Observed" and "actual" are synonymous. I think you mean to write, "observed and estimated"
d. Non-Supervised Learning - Neighborhood Characterization
    i. Figures 8, 9, and 10
        1. These figures are not publication quality. You have informal titles above each plot, all of the font sizes on the plot are way too small, and your legend is inscrutable (both semantically and in terms of size).

6. Analysis
    a. 6.1 Random Forest - Binary Model
        i. I believe the expectation for this draft was a publication ready paper. Better hurry up! You didn't even finish the last sentence of this paragraph.
    b. 6.2 Multi-Variate Regression
        i. "Because the regression model was subset based on the total cost of damage to the pedestrian excluding events resulting in $5 million or more in damages,"
            1. Isn't it "or less"? You previously claim the cost of a fatality is >$10M.
    c. It seems to me that you ought to at least acknowledge methodological error that may be present in your data. Do we have any idea how well your data represent reality? How do we know there isn't significant bias in non-emergency reports (aren't rich people far, far more likely to report quality of life issues than poor people, so wouldn't

poorer neighborhoods be “underreported”?). How do we know there isn’t bias in the surveys? It appears from the article you cited that no sampling frame is employed, responses are anonymous, anyone in the world can fill out the survey, and there is no controlling for duplicate responses. Seems like those data are likely unusable for statistical inference.

7. Ethics
    a. Are there any ethical implications to performing a self-assessment of ethics?
8. Conclusions (and Future Work)
    a. The title of this section is too informal. “Future Work” shouldn’t be in parentheses.
    b. As you are aware, this section is not complete. Hurry up!
    c. Knowing what I know as a data science student, if I were a city council member and you brought this paper to me, I would not accept your conclusions. Your “Actual/Observed” data are in fact not the true costs associated with pedestrian-vehicle incidences as far as I can tell. Also, your methods don’t read as statistically rigorous (exclusion of data, manual removal of variables, comparison of models, etc.). Your analysis section doesn’t really appear to be constructing any sort of narrative to support your conclusions. How are your map plots supporting your conclusions? As far as I can tell, they don’t. Instead, they are superfluous information. What is the point of your clustering exercise? How does that tell the reader about pedestrian safety in Cincinnati? It begins to feel as though you took an approach of using a bunch of different fancy analysis techniques without any overarching plan or intention as to how the results of those techniques might help you test a hypothesis or construct a coherent argument.
    d. Note that I am not saying that you haven’t done good work or made statistically sound conclusions. Rather, I am saying that those things are not apparent in this draft of your paper.
    e. The errors in your citations that I point out previously in this document undermine your credibility.
9. Final Note
    a. My comments in this document may appear harsh or overly critical. My intention is not to insult you or your abilities. Rather, I am attempting to point out as many of the flaws in your paper as possible such that you might correct them prior to publication. I am attempting to provide you the type of feedback that I am hoping to receive on my paper. Most MSDS students have never written a paper for publication in an academic journal, and the training we are provided for doing so in this program is minimal. As such, mistakes and shortcomings are to be expected. The best we can hope for is to learn from our mistakes and from each other. Keep working hard and improving! Your best is yet to come!