Addressing Instructor Feedback

Jonah Edmundson, Ricky Heinrich,
Noman Mohammad, Avishek Saha

Introduction

- Provide more context on the implications of NAs. Is there a known cause for why a DB would be missing a record? (not enough resources to collect that material, human error, ... etc)
  - See Data Sources section.
- You provided some background into clustering algorithm but surely there must be some information zoning, or disctrict planing, for example, that you may leverage.
  - Zoning is now mentioned in the introduction. We will inquire to our client about this and look into some open-source options.
- how are DM defined? are they all the same size (in terms of area/population)?
  - (Assuming DM = DB,) Added this to the Introduction: "The PMD contains continuous measures for 10 amenities at a 'dissemination block' (DB) level, the most granular area defined by StatCan (Statistics Canada, 2021). In an urban area, a DB corresponds to a city block, whereas in rural areas they are areas "bounded by roads or other natural features" (Alasia et al., 2021). Thus, DBs different broadly in their proximity to these amenities."

Aims and Objectives

- Ill posed. Please read: https://examples.yourdictionary.com/examples-of-good-and-bad-research-questions.html
  - Noted. Research questions heavily triaged down to 2 questions. Please let us know if you think these are more appropriate, and if we could further improve upon them.
- Without any context into what missing values represent and why they are important the first question loses meaning.
  - Removed this research question, as we plan to address this in our EDA.
- how will you determine "best" (in reference to research question 2 and 5).
  - Added a text blurb above the research questions: "The clusters returned by these algorithms will be of varying quality; some will be well-defined, and others may be more muddled. In order to choose one best algorithm, the clusters returned from each algorithm will have to be compared using a validation metric such as the Dunn Index or Silhouette Coefficient. These metrics are generalizable between algorithms because they compare intra- and inter-group variance".
- I don't understand whats meant in the brackets of research question 3.

  - Reworded research question to be more clear.

- question 4 not well-posed. Characteristics in terms of the PMs? in terms of other demographics?

  - Reworded question 4, and added a definition of characteristics in brackets.

- how are you going to choose your different subsets of the data? why is it important to look at subsets?

  - The term 'subsets' was misleading. This was removed.

Dataset

- You mentioned in your presentation that these PMs have been normalized across Canada. That is prudent piece of information is not included in your report

  - This is now mentioned in the Data Sources section.

- is smaller better for PM?

  - This is now mentioned in the Data Sources section: "In this case, a lower proximity measure indicates that the amenity is located farther away from the dissemination block. So, if the proximity measure is low, it means that the amenity is more distant from the dissemination block than if the proximity measure were high".

- what do NAs mean?

  - This is now mentioned in the Data Sources section. NAs are of several different types, and result from missing data in the data sources from which the PMD was constructed.

- how is a DB different from a subdivision... how do you expect to merge these?

  - This is now mentioned in the Data Sources section. Census SubDivisions (CSDs) are larger geographic boundaries that contain DBs. In the PMS dataset, each row is a DB, and each row also has a correpsonding CSD ID that can be used to link other datasets that are recorded at the CSD level, such as the Index of Remoteness (IoR). From our proposal: "the IoR can be linked to the proximity measures dataset by a unique ID that is available in both dataset".

- I'm not sure what is meant by robust. Robust to changes

  - This word was originally mentioned in the Research Questions section, but the concept was moved to the Methodology section. The word "robust" was a bit confusing, so we changed it to "sensitive". This is what we mean by "sensitive": "We will use different approaches to determine the cutoff values or thresholds proximity indices. So, by comparing the cutoff values suggested by each approach, we can assess the sensitivity of the results to the choice of method and determine whether the findings are sensitive (robust) to changes in the method used.".

Deliverables and Timeline

No feedback here.

Style

- parts poorly written
  - Which parts? How so? Hard to fix this without more information.... We tried editing things again but are unsure if it meets your standards.
- numbered lists for one item (limitations section)
  - Fixed.