1 Classification of User Errors

Total entries from first attempt (raw_user_address): 21,927
Total entries from second attempt (raw_user_address_2): 1,265

2 Successful Entries

The first step is to remove all of entries that successfully returned an address (ward != blank). Overall, there were 5,475 successes on the first attempt and additional 302 successes produced on the second attempt.

There were multiple ways the ward numbers were produced on the first try (either correctly or incorrectly):

- Full address (minimum of street number + name): 1,531
- Partial address (zip code): 816
- Partial address (city name or some other identifier): 1,683
- Erroneous input (uninformative number, random text/punctuation): 1,694
- Miscellaneous: 53

3 Categorizing Errors

After removing entries that successfully generated a ward number, while leaving those that generated a match, but only after the second try, there are 16,506 errors in raw_user_address to code. I currently use three broad categorizing errors: 1) incomplete entries, 2) incorrect entries, and 3) extraneous entires.

Incomplete entries These are entries where individuals appear to have made an effort to submit location information, but did not submit enough information to produce a successful match. Within this category, I create multiple subcategories of incomplete entries:

- Ward only: only a ward number was entered (i.e. "Ward 3")
- Zip code only: only a zip code was entered (a four digit number)
- Street number only: only a street number was entered (¡ than a four digit number)
- Zip code + other location info: a zip code was entered with other location information (i.e. a city or village name)
- Street name + incomplete location: An entry must have a minimum of three words after the street number (street names must have two words [i.e. main street, mandela road] and a city name. Thus, any correctly formatted entry with a street number but less than three words is included in this category)

This category currently accounts for 6,885 of 16,506 errors (42%).

Incorrect entries These are entries where individuals appear to have made an effort to submit some type of information, but may have submitted erroneous information or committed a user error. Within this category, I create multiple subcategories of incorrect entries:

- Phone number: the user entered a phone number (a ten digit number)
- Random number: the user entered a number that could not be considered a phone number, zip code, or street number.
- Unrecognized address: the user submitted an address in the correct format with a street number and enough words to form a full address. However, the address was not recognized. This can be attributed to misspellings, unconventional abbreviations, and unofficial or fake addresses.
- P.O Box: only a P.O. Box was entered or a P.O. Box was entered instead of a street number.

This category currently accounts for 4,319 of 16,506 errors (26%).

Extraneous entries These are entries where individuals make no effort enter useful information. Within this category, I create multiple subcategories of incorrect entries:

- Party name: the user entered a party name or statement of support for a party (includes the ANC, EFF, and DA)
- Expletives: the user entered an expletive.
- Single character: the user entered only a single letter or punctuation character.
- Personal responses: the user entered a brief, random reply (yes, no, okay, maybe, hi, hello)
- Other responses: the user entered a brief reply that could be interpreted as a response to other questions on the survey (yes, no, okay, maybe, hi, hello)

This category currently accounts for 1,335 of 16,506 errors (8%).