C4V Hackathon 2020 Track 2 (NGOs)

Team Life Regression

Data Quality Issues

Data quality issues within the data frame that collected the responses to the *Encuesta Nacional de Hospitales* survey revolved around missing values, messy inputs and scales, as well as achievability of recommendations from information provided. Since the recipient of the analysis are NGOs, only 54 variables where utilised. Out of these, approximately 10% of all values were missing. The issues identified are discussed below.

- Missing values

Null observations were prevalent within the data with some columns having as much as 40% of the data entries blank. This may have been a result of unavailability of data, inability of respondents to estimate measures, or oversight during the response process (which per se may be an indication of the ease of the answering process). Missing data impacts data analysis negatively because this warrants the elimination of responses from data sets in extreme cases or imputation of the missing values, either of which distorts data and the reasonableness of assumptions. A basic approach to solving this problem will be incorporating a response category for unknowns (e.g. marked "Unsure", "Unknown", etc.) so to eliminate missing values with its associated ambiguity. An additional solution may be to retrain informants concerning the response process and reiterate the importance and purpose of completeness.

- Messy inputs

This was an obvious result of input method as different inputs providing the same connotation should all be valid. For instance, a numerical response may be written in words using different font cases or as digits, but all inputs will still refer to the same figure. This would usually cause duplication of data and possible difficulties in categorizing responses, which, in turn, requires expensive man hours to clean up. An effective way of overcoming this problem would be to control inputs by specifying ranges and using radio inputs or check boxes. Thus, respondents select pre-selected responses rather than entering unstructured text themselves, minimizing the time for data entry and data analysis.

- Inconsistent scales

Inconsistent scales were present within some parts of the survey. For instance, availability of drugs, performance of specific departments or operability of equipment had differing ranges within different columns. As much as possible, similar indicators should have the same scale to allow for consistency and validity of insights from such analysis.

- Indicators

The survey instrument sought to estimate certain key indicative metrics to provide information to a cross-section of stakeholders, who might have different interests and targets in mind. The main purpose of this metric is estimating the availability of several medical supplies or equipment across Venezuela, which leads to a measure of scarcity. Yet, it would be difficult for a generalist NGO to determine what is essential by evaluating only the supply level in terms of time. Again, availability and distribution of resources seem to be the main variables that can be computed from the data. Other measures that should be assessed include level of demand or urgency for resources, i.e. medicines or equipment, or the volume of supplies available. Defining target variables and success metrics upfront will positively affect the survey responses and the actionability of the insights.