**Data Perspective**

**One Variable**

**Categorical**

There are a few categorical values that we can generate utilizing our data. This would be counts based on neighborhood (NOM\_QR), borough (NOM\_ARROND), declaration date (DATE\_DECLARATION) and location (COORD\_X x COORD\_Y). This would give us different insights on the infestations declared in the various areas of the city.

**Numerical**

Based on the categorical values we generated above, we can then establish Min, Max, Mean values based on those Counts. We can also generate Sums based on the number of visits by location or by time of year (based on declaration dates) to possibly gain insight on variables such as temperature or moving period (seasonal).

**Multiple Variables**

Multi variable statistics can also be generated based on our dataset to try and establish trends and gain further insight. Examples of such include:

* Counts on borough (NOM\_ARROND) and neighborhood (NOM\_QR) by date of declaration
* Number of days between DATE\_INSP\_VISPRE and DATE\_DECLARATION.

(Long waiting period can cause infestations to spread.)

* Number of days between DATE\_DECLARATION - DATE\_DEBUTTRAIT
* (Long waiting period can cause infestations to spread.)

**Business Perspective**

**Descriptive Data Analysis**

The project we have undertaken will look at the City of Montreal bedbug infestation declarations dataset available on the city website.  The dataset contains a list of over 33000 declarations of confirmed infestations located in the various boroughs.  Included in the dataset are 13 columns detailing the specific dates the infestations were declared along with unique declaration numbers per entry.  Each record also identifies the neighborhood and borough.  The specific locations are solely identified by coordinates to ensure the privacy of the residents.  The dataset also includes inspection dates along with the start and end periods for exterminations including the number of visits per household.

The city clarifies some information concerning the data in the following manner.  The declarations that are not associated to an extermination have been removed from the dataset.  Therefore, the dataset pertains to households that have been treated.  Declarations linked to non-residential housing or out of the territory of the 19 boroughs have been removed and finally to keep the confidentiality of the affected household, the closest intersection to the household was used to not pinpoint the actual location.  The infestation reports were submitted since the 5th of July 2011.  The city website indicates that the data has a low degree of reliability due to the concern that they were entered manually by a third party, pest control managers, and are not validated by the City of Montreal.

While analysing the data we found numerous entries, approx. 2100, with a lack of inspection dates as well as missing start and end dates for the extermination along with missing values for number of visits required.  Since the city indicated that all declarations not associated with an extermination are removed from the dataset, we can conclude that a blank visits field is at minimum equal to 1.  As for missing dates, we will have to fill in the gaps, if required, by using the Declaration and Inspection dates as approximates to feed the Start and End dates.

**Diagnostic Data Analysis**

Bedbugs can spread and infest an area in numerous ways.  Based on government reports we know infestations can spread by human contact (clothing), unsanitary environments, along with used furniture and mattresses to name a few.  This exercise in data mining has been undertaken to potentially uncover flash points, periods and/or locations that can possibly bring to light problematic scenarios that we can rectify and decrease, if not eliminate, future infestations.

Validation Examples:

1. Number of days between Inspection/Declaration and Start-End of extermination.

(Given a lengthy waiting period we could have a potential opportunity for spreading)

1. Identify the most problematic boroughs

(What is the difference between worst and best ranked)

1. Identify the most problematic neighborhood

(What is the difference between worst and best ranked)

1. Based on 2 and 3, is the issue borough wide or at neighborhood level?

(Management issue)

1. When was the declaration made?

(Is the problem based on a specific time of year? Heat? Moving period?)

1. Is a problem more relevant to a particular block?

(COORD\_X x COORD\_Y)

1. Nbr\_Exterminations x COORD\_X x COORD\_Y x Nom\_QR

(Can the issue be subject to extermination mismanagement?)

<http://donnees.ville.montreal.qc.ca/dataset/declarations-exterminations-punaises-de-lit>

<http://www.environnement.gouv.qc.ca/pesticides/permis-en/code-gestion-en/BedBugControl.pdf>

<https://www.mayoclinic.org/diseases-conditions/bedbugs/symptoms-causes/syc-20370001>