CEBD 1260 - BIG DATA ANALYTICS

SPRING 2019

Project Swarm:

WebApp for determining the quarantine period due to a Bed bug infestation

June 14, 2019

Luigi Clerici & Ilia Kassianenko

Agenda

- Problem Definition
- Declarations Trend and Analysis
- Dataset Description and Cleaning
- Model
- Results and Suggested improvements
- Q&A

Problem Definition

Context

- Our analysis centers around the bedbug infestations that affecting the boroughs formed by the City of Montreal.
- Some boroughs and neighborhoods have been impacted considerably more than others. While some occupants escape this nightmare, others have to live through the chaotic turmoil of an extermination procedure.
- An infestation can turn a relatively tranquil existence upside down and impact one's life temporarily through the tedious and time consuming process required to eradicate these bugs from your home and the costs associated to it. The tracking of the dataset declarations to the city began in 2011.

Source

- City of Montreal Open Data portal. Report including 33,365 declarations/rows for bed bug exterminations.

Target audience for WebApp:

- General public, potential dwelling occupants, buyers, and investors.

Desired output

- Expected number of days until Infestation Free (or quarantine period), based on dwelling location and extermination history and proximity to three infestation epicenter locations.

Declarations: Trend

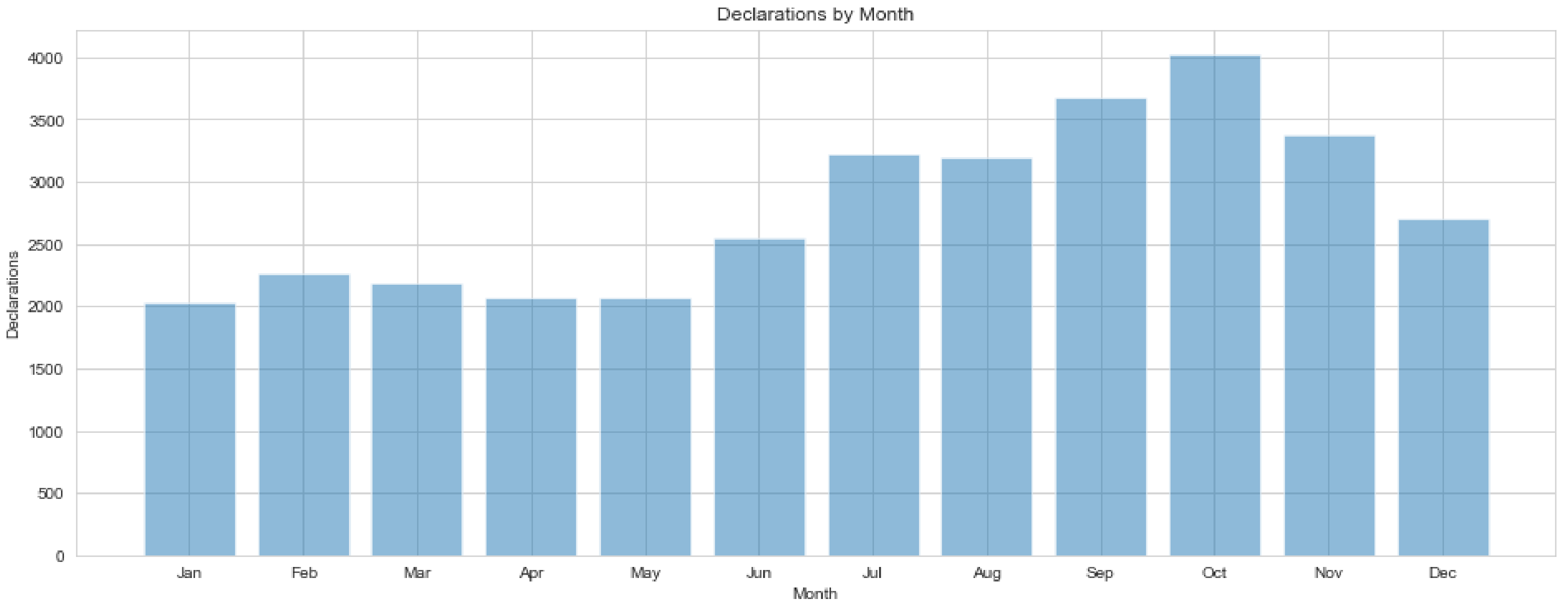
The top 5 boroughs with the highest count of declarations

- Plateau-Mont-Royal, Mercier-Hochelaga-Maisonneuve, Rosemont-La Petite Patrie, Ville-Marie, and Villeray-Saint-Michel-Parc-Extension.

- Note: Declarations are not available for Boroughs not part of the City of Montreal, such as Montreal-Est.

Monthly trend

- For grouped intersections that are viewed on a monthly scale, we can visualise the months with the highest count of declarations: the period from June through October is the most problematic. (*with 2126 declarations removed)



Declarations: Analysis

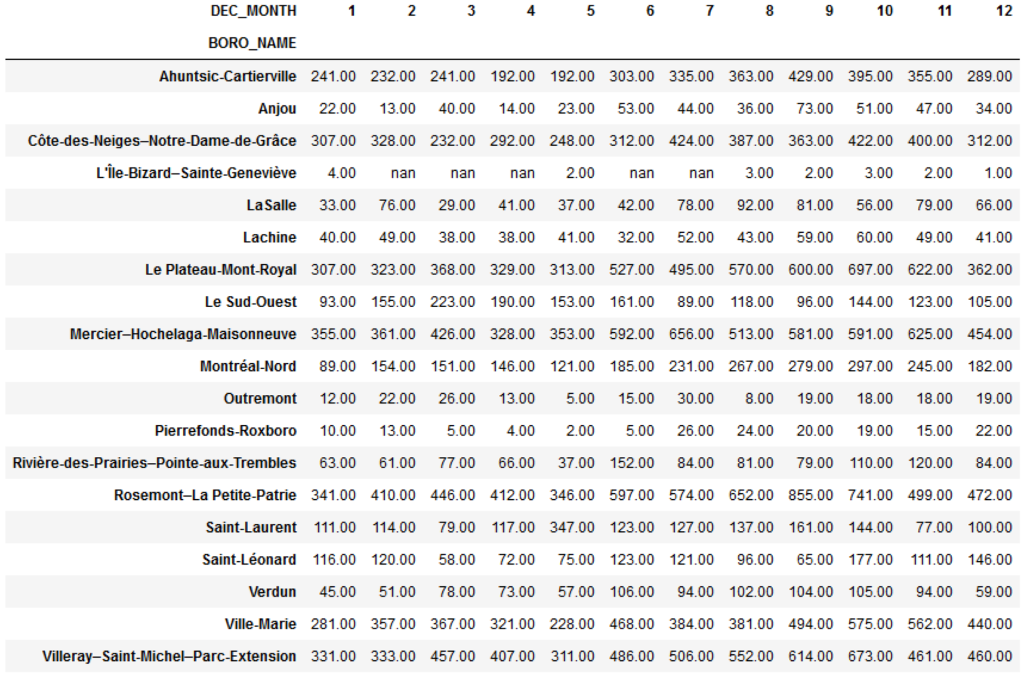
Intersections

- Borough and Neighborhood labels are not used in the regression model, since an intersection can be shared by two neighborhoods.

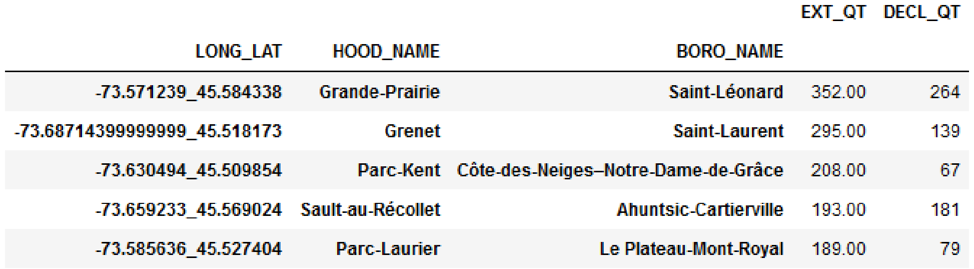
Counts

- The highest count of declarations does signify the highest count of exterminations.
- Boroughs with the highest count of declaration may not contain the most affected intersections.

Dataset view by Month and Borough since 2011 (*with 2126 declarations removed)



DEC_MONTH	1	2	3	4	5	6	7	8	9	10	11	12
BORO_NAME												
Ahuntsic-Cartierville	241.00	232.00	241.00	192.00	192.00	303.00	335.00	363.00	429.00	395.00	355.00	289.00
Anjou	22.00	13.00	40.00	14.00	23.00	53.00	44.00	36.00	73.00	51.00	47.00	34.00
Côte-des-Neiges–Notre-Dame-de-Grâce	307.00	328.00	232.00	292.00	248.00	312.00	424.00	387.00	363.00	422.00	400.00	312.00
L'Île-Bizard–Sainte-Geneviève	4.00	nan	nan	nan	2.00	nan	nan	3.00	2.00	3.00	2.00	1.00
LaSalle	33.00	76.00	29.00	41.00	37.00	42.00	78.00	92.00	81.00	56.00	79.00	66.00
Lachine	40.00	49.00	38.00	38.00	41.00	32.00	52.00	43.00	59.00	60.00	49.00	41.00
Le Plateau-Mont-Royal	307.00	323.00	368.00	329.00	313.00	527.00	495.00	570.00	600.00	697.00	622.00	362.00
Le Sud-Ouest	93.00	155.00	223.00	190.00	153.00	161.00	89.00	118.00	96.00	144.00	123.00	105.00
Mercier–Hochelaga-Maisonneuve	355.00	361.00	426.00	328.00	353.00	592.00	656.00	513.00	581.00	591.00	625.00	454.00
Montréal-Nord	89.00	154.00	151.00	146.00	121.00	185.00	231.00	267.00	279.00	297.00	245.00	182.00
Outremont	12.00	22.00	26.00	13.00	5.00	15.00	30.00	8.00	19.00	18.00	18.00	19.00
Pierrefonds-Roxboro	10.00	13.00	5.00	4.00	2.00	5.00	26.00	24.00	20.00	19.00	15.00	22.00
Rivière-des-Prairies–Pointe-aux-Trembles	63.00	61.00	77.00	66.00	37.00	152.00	84.00	81.00	79.00	110.00	120.00	84.00
Rosemont–La Petite-Patrie	341.00	410.00	446.00	412.00	346.00	597.00	574.00	652.00	855.00	741.00	499.00	472.00
Saint-Laurent	111.00	114.00	79.00	117.00	347.00	123.00	127.00	137.00	161.00	144.00	77.00	100.00
Saint-Léonard	116.00	120.00	58.00	72.00	75.00	123.00	121.00	96.00	65.00	177.00	111.00	146.00
Verdun	45.00	51.00	78.00	73.00	57.00	106.00	94.00	102.00	104.00	105.00	94.00	59.00
Ville-Marie	281.00	357.00	367.00	321.00	228.00	468.00	384.00	381.00	494.00	575.00	562.00	440.00
Villeray–Saint-Michel–Parc-Extension	331.00	333.00	457.00	407.00	311.00	486.00	506.00	552.00	614.00	673.00	461.00	460.00



			EXT_QT	DECL_QT
LONG_LAT	HOOD_NAME	BORO_NAME		
-73.571239_45.584338	Grande-Prairie	Saint-Léonard	352.00	264
-73.68714399999999_45.518173	Grenet	Saint-Laurent	295.00	139
-73.630494_45.509854	Parc-Kent	Côte-des-Neiges–Notre-Dame-de-Grâce	208.00	67
-73.659233_45.569024	Sault-au-Récollet	Ahuntsic-Cartierville	193.00	181
-73.585636_45.527404	Parc-Laurier	Le Plateau-Mont-Royal	189.00	79

Recall: the top 5 boroughs with the highest count of declarations

- Plateau-Mont-Royal, Mercier-Hochelaga-Maisonneuve, Rosemont-La Petite Patrie, Ville-Marie, and Villeray-Saint-Michel-Parc-Extension.

Dataset Description

Source

- Declaration report tracking for our dataset began on the 5th of July 2011, as per the data dictionary.

- The data dictionary 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.

Composition

- Included in the dataset are 13 columns detailing the specific dates the infestations were declared along with unique declaration ID's per entry.

- Each record 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 to a maximum of 4.

- Columns:
  NO_DECLARATION, DATE_DECLARATION, DATE_INSP_VISPRE, NBR_EXTERMIN, DATE_DEBUTTRAIT, DATE_FINTRAIT, No_QR, NOM_QR, NOM_ARROND, COORD_X, COORD_Y, LONGITUDE, LATITUDE

Dataset Cleaning

Null values

- Data dictionary was updated following a confirmation from the City of Montreal that null values (2124 cases) are not removed from the dataset, as they may be populated a later time.

- Out of 33,365 declarations there are 2,124 null entries, which lack inspection dates, extermination start and end dates, and the number of exterminations. These declarations are removed from the train/test model.

Extermination date preceding an inspection date

- The City of Montreal communicated that an extermination and inspection can possibly occur on the same date. Although, an inspection needs to occur prior to an extermination. 726 declarations have an extermination date preceding an inspection. These declarations are not removed from the model, pending feedback from the City.

- The inspection date and declaration may be several days apart, as per a communication from the City of Montreal. 22,608 declarations have an inspection date preceding a declaration date by more than 7 days. These declarations are not removed from the model, pending feedback from the City.

Model

Model

- Machine learning, Supervised learning, Regression, Random Forest (10)
- Heroku URL: https://houseproj.herokuapp.com/
- Random Forest was used due to it’s flexibility for both regression and classification problems.

  Ref.: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Model Inputs:

- Month of the last extermination (1 through 12) at the reference dwelling
- Year of the last extermination (2018 through 2019) at the reference dwelling
- Distance for reference dwelling from intersection (-73.585636, 45.527404) in Parc-Laurier, Le Plateau-Mont-Royal (in 2 KM increments. from 2 to 14)
- Distance for reference dwelling from intersection (-73.571239, 45.584338) in Grande-Prairie, Saint-Léonard
- Distance for reference dwelling from intersection (-73.68714399999999, 45.518173) in Grenet, Saint-Laurent
- These distances are taken from 3 intersections out of the 5 top intersections previously shown, which are ranked based on the count of exterminations per intersection.

Output

- Expected number of days until Infestation Free (or quarantine period), based on dwelling location and extermination history

- The lead time is an indication if the extermination was successful. Based on the assumption that all infestation cases are reported, for a short lead time, the extermination are assumed to be successful, since subsequent visits did not occur.

Discussion: Results

Model

- Random Forest (10) was chosen Random Forest (100) due to runtime, as perceived in Jupiter Notebook. Testing of runtime did not occur on the application.

Location

- The reference dwelling may not exactly be located within three distances indicated.

Retroactive results

- When using three distances (2 KM each) for the reference dwelling and January 2018 as inputs, the output may indicate a quarantine period of a year. Although, by using the same three distances and January 2019 as inputs, the output may indicate half a year.

- These outputs show that the results cannot be interpreted retroactively. Specifically, the user would need to stay away for a year from the reference dwelling, if the application is run at the start of 2018. Thereafter, before choosing to occupy the reference dwelling, the user would need to stay away another half a year, if the application is run at the start of 2019.

- The model did not produce an output of a year and half immediately. Running the app from the perspective of the past may result in non-cumulative output.

Discussion: Suggested Improvements

Location references

- Distance from Epicenters are input using an interval (2-4 KM) rather than a single value in the dropdown menu of the WebApp. The model would need to be converted to use distance interval as Getdummies variables.
    - Using Getdummies variables for Longitude and Latitude created runtime challenges for the python model.

Model

- The minimum quarantine duration is not considered in the model, as a forced constraint for the model.
- Runtime needs to be tested for the application at Random Forest (100).
- Additional input fields can be added for the app for public officials, such as the lead time between the inspection and extermination dates. Several input fields were removed for public users.

- An alternative regression would entail in determining the number of declarations, although the number of exterminations is capped at 4 for each declaration. The relationship between declarations and intersections is unknown. It is possible that exterminations are split over several declarations for the same location for the same period of time. This analysis was not performed.

- A classification model can be created to determine whether the extermination occurred successfully or not with False Positives and True Negatives. This approach would indicate that an inspection is needed to prevent a future occurrence and schedule a pre-emptive extermination.

Q&A

References

- https://www.epa.gov/bedbugs/do-it-yourself-bed-bug-control
- https://www.canada.ca/en/health-canada/services/pest-control-tips/bedbugs-how-do-i-get-rid-them.html
- https://en.wikipedia.org/wiki/Boroughs_of_Montreal
- https://www.webmd.com/skin-problems-and-treatments/guide/bedbugs-infestation#1