NYC Forestry Service Request Analysis

Abhay Mourya

Dec 2017

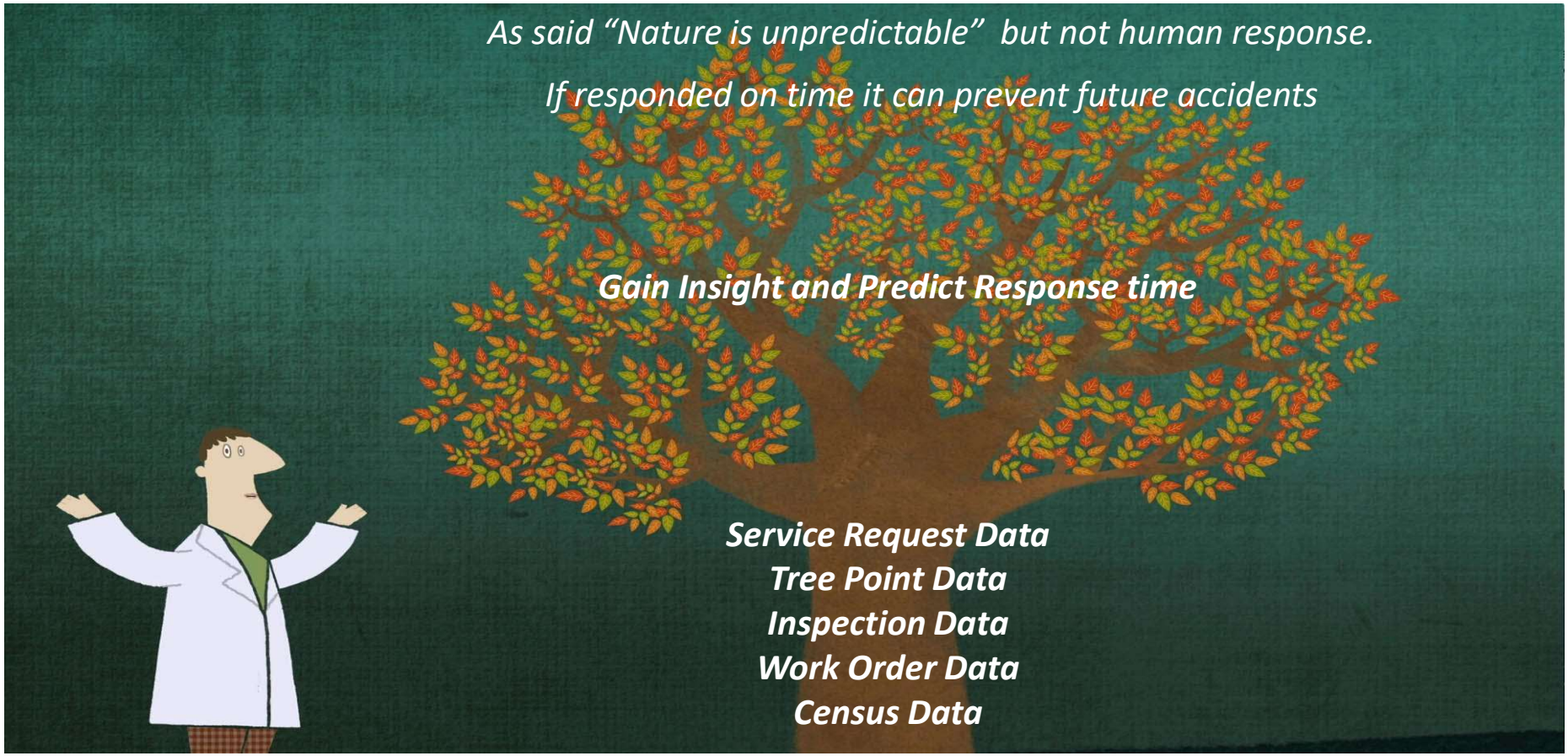
Background

"Neglected, Rotting Trees Turn Deadly"
There are roughly 2.5 million trees in the city's parks and on its street. Forestry department has to prioritize care for more than 70,000 trees a year with procedures. " Nature is unpredictable and limbs can fall even from healthy and well pruned tress"
NY Times May 13,2012

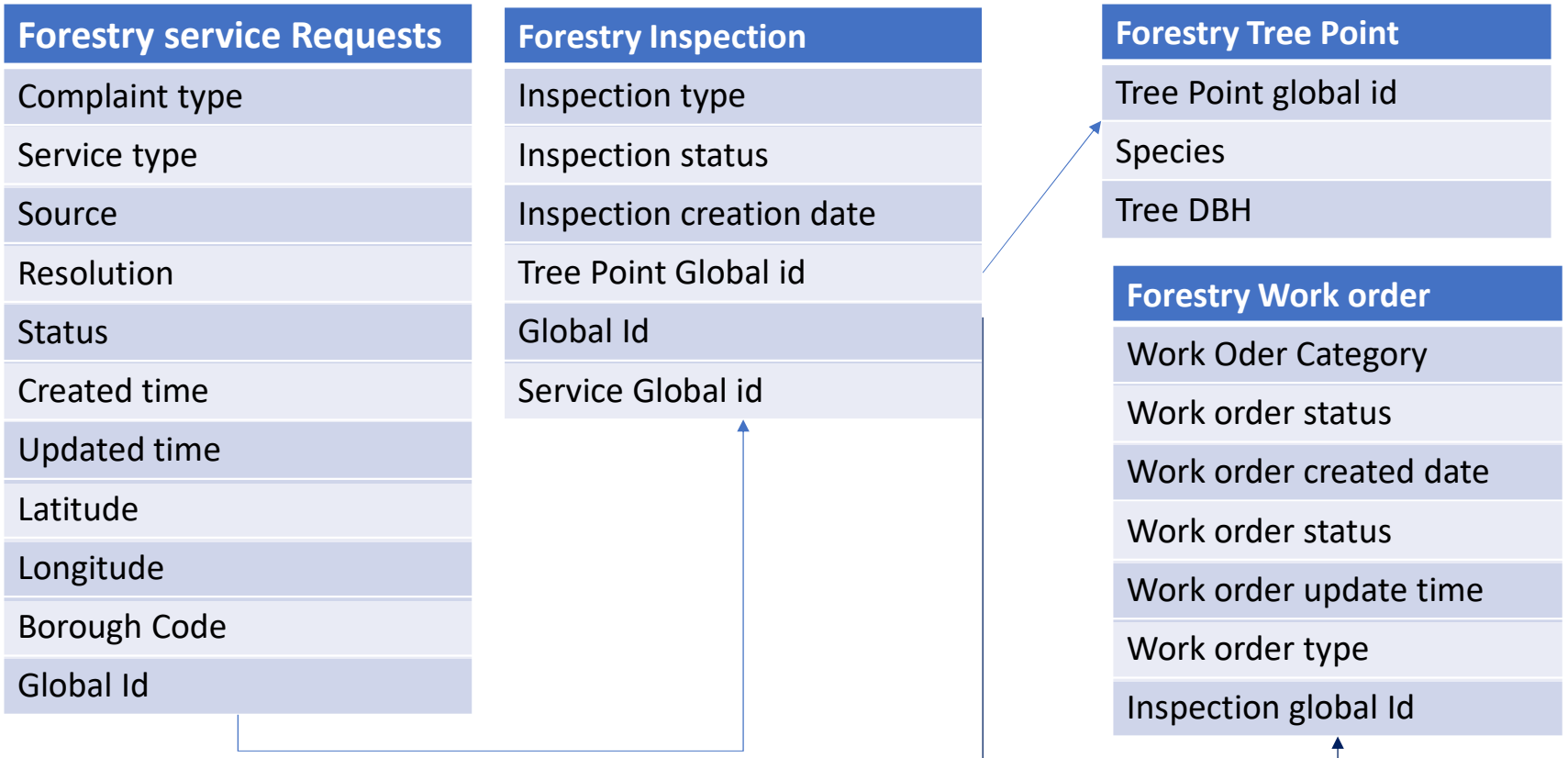
NEW YORK DAILY NEWS September 13, 2017
Between 2011 and 2015, 31 people were injured by falling trees or branches, data from a Freedom of Information Law request shows

Number of notice of claims before a lawsuit is filed against the city by people hit by trees typically tops over 100 each year, data from the city's Controller's Office reveals.

Introduction



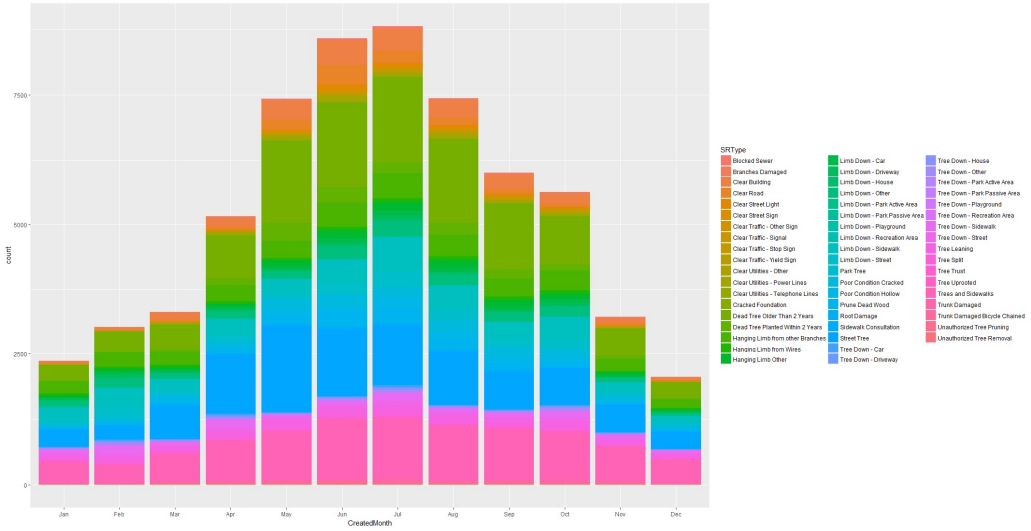
Data Wrangling and Prep

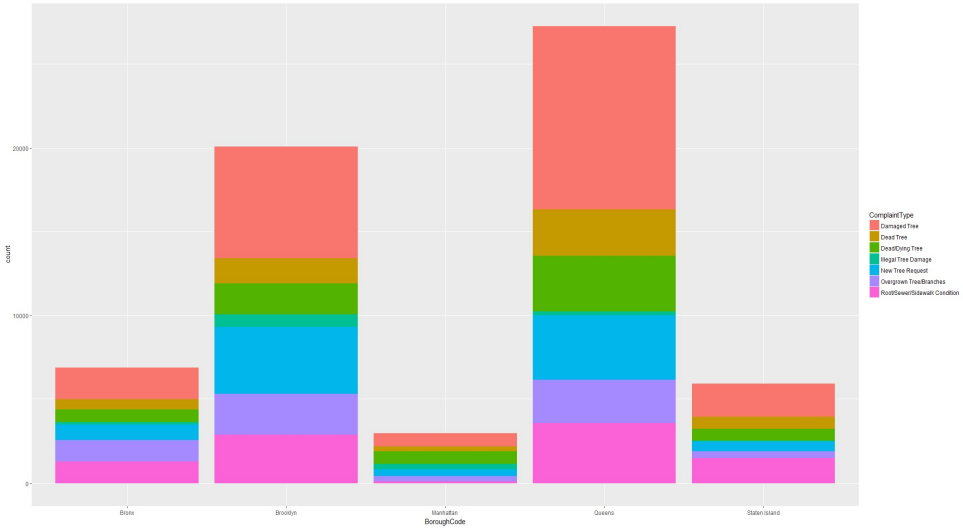


Total Service request: 192K from Year 2015 onwards
Total common data across joined datasets: 43551

Exploratory Analysis

Service Request trends





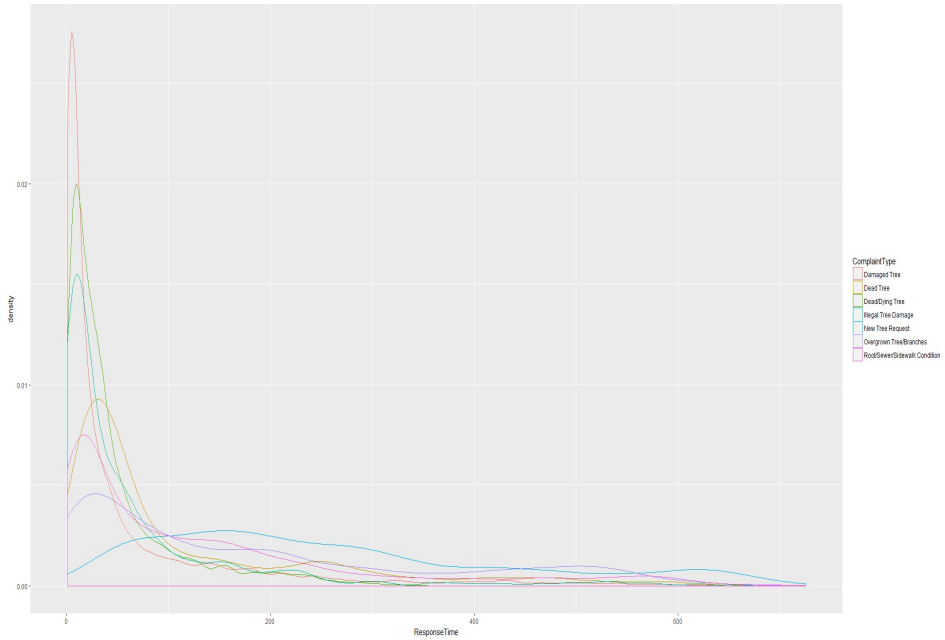
Mapping SR creation month, data shows most of the requests are created in Summer from May to Aug and almost normal distribution. It seems post spring during April to Sep most of the complaints increased.

Queens and Brooklyn lead in most number of service request raised in the dataset.

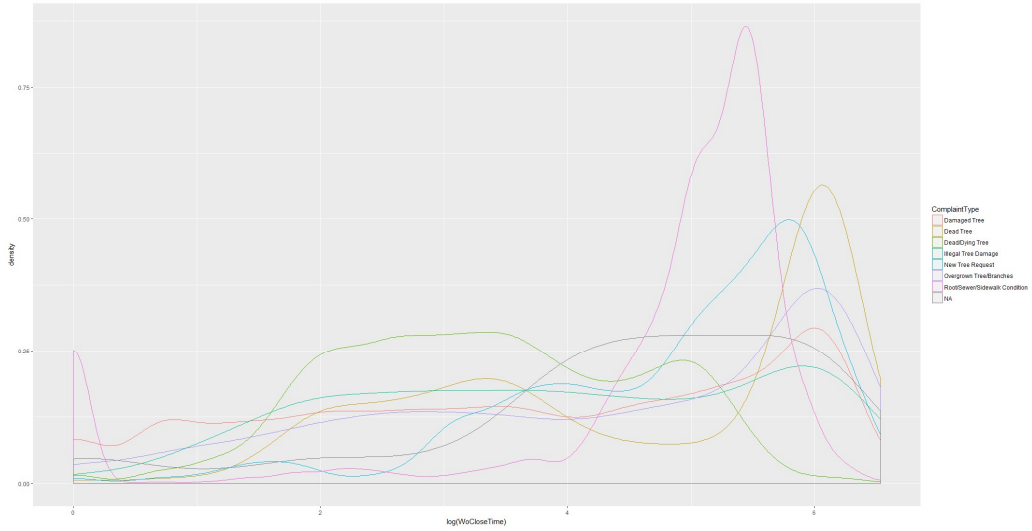
Proportionally Manhattan has around 45 % new tree request

Other boroughs 25 % cases are around sidewalk replairs

Response to Service Request



Response time varies with Complaint type. While Damaged Tree and Dead Tree are responded within 100 days, Root/Sidewalk conditions and overgrown branches have response time in more towards 200 days.

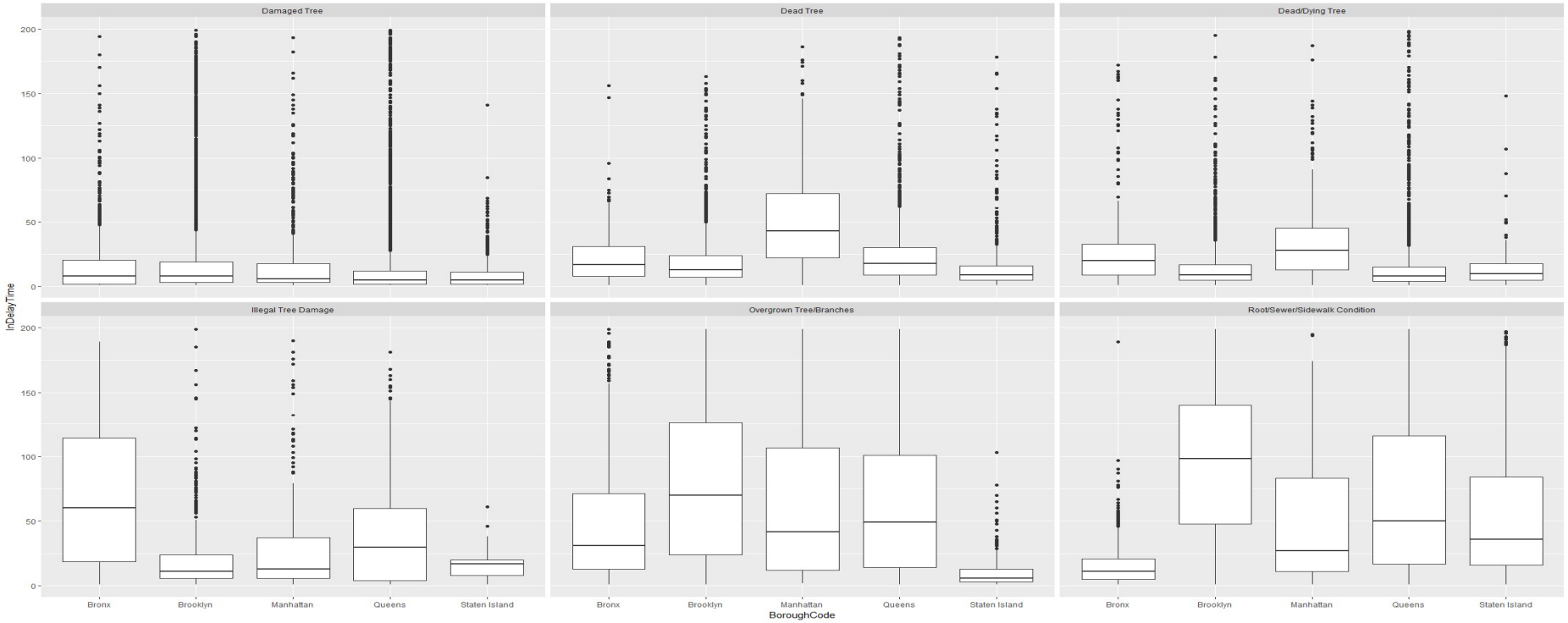


Complaint Type Root/Sewar/Sidewalk and Dead Tree has high probability of high WO closure time.

Curious case of Brooklyn

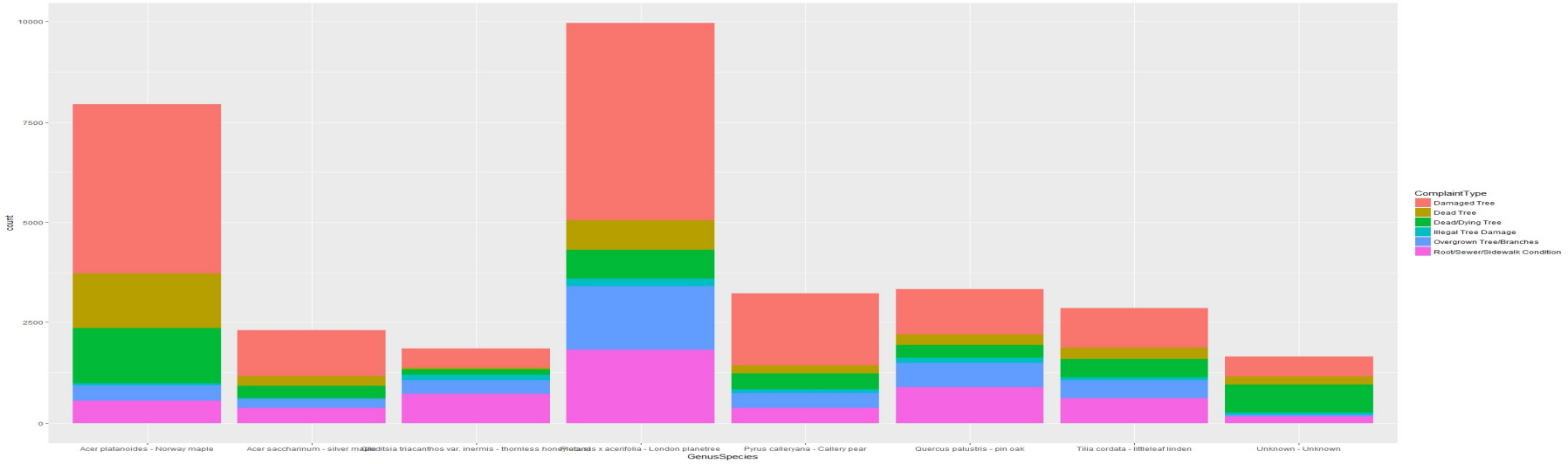
Plotting Borough Code on longitude and latitude with filter on response Time more than 400 days for all complaint types Brooklyn has more presence than any other boroughs even though total number of complaints are less than that of Queens

Inspection Delay time



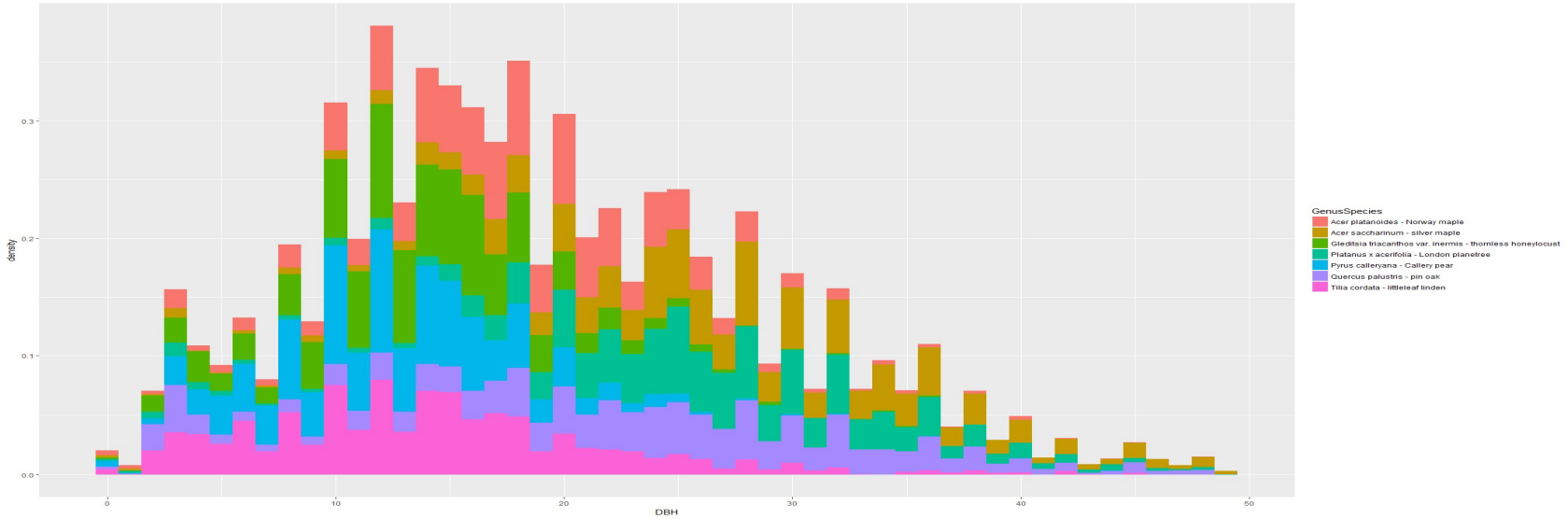
Time interval from service creation to inspection initiation date. We can see that for specific complaints at specific Borough Median and IQR time varies. Due to prioritization of Service request, some SR are prioritized over others for inspection.

Complaints and Tree species



1. London PlaneTree has most cases around 10000 and 50 % of request more of type Damaged tree, Overgrown Branches, Root/Sidewallk problem and has less count on Dead or Dying Tree.
2. Norway Maple has more than 50% cases on Damaged Tree and almost 40 % cases in Dead/Dying Tree.
3. Pin Oak has second most cases of Root/Sidewalk problem and Overgrown Branches and Thornless Honeylocust has similar proportion of requests.
4. Callery Pear has thirdmost Damaged Tree cases after London Planetree and Norway Maple.

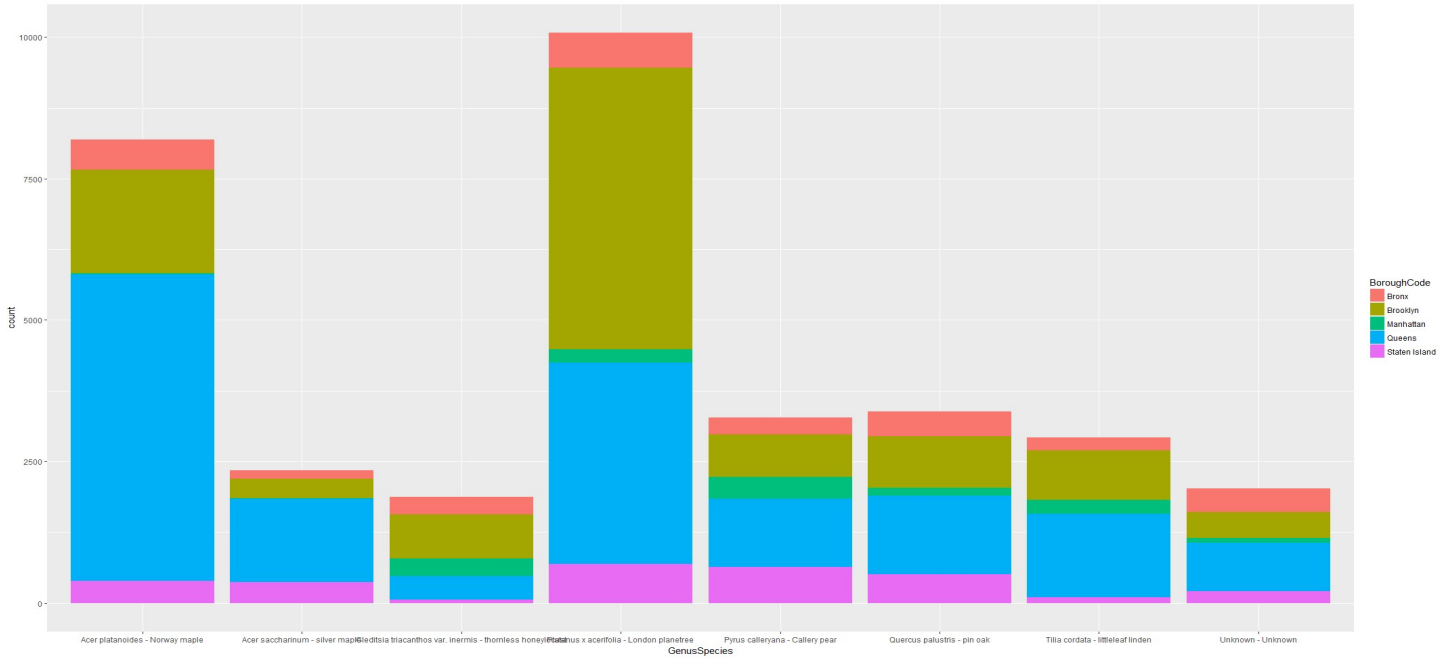
Relationship with Tree Diameter



For each species Diameter is falling under almost normal curve.

1. Most of request for Norway Maple is between DBH 10 to 27 While London plantree most cases when DBH between 15 to 40.
2. Pin Oak has more cases at DBH more than 20 while Pyrus and Tilia as well as Callery Pear has most cases at less than 20 DBH
3. Thornless Honeylocust has more cases in 10-20 DBH

Repeated Requests



1. Queens has most Norway Maple cases which are second most species under complaint and it can inferred that most of these cases are concentrated in Queens.
2. Brooklyn has most or second most presence in all species and comparing with Complaint type they have almost more complaints in each category only second in Tree Removal. Most complaints related to London Planetree are concentrated in Brooklyn which has almost similar proportion of all complaint types.

Analysis Summary

1. Queens has more number of cases, these are more related to one Species Norway maple and Mostly Tree removal which are promptly responded.
2. Brooklyn has more variety of complaints on multiple species which is causing might reason for higher response time.
3. Service request raised from Brooklyn and Adjacent Queens neighborhood chances of update or closure is more than 200 Days and can be 400 days+ as well.
4. Delay in inspection is major cause of delay in closure of service request
5. Brooklyn has most repeated calls for all complaint types.
6. Work Order category on these repeated complaints, it can be inferred that Tree Removal and Tree SideWalk repair as Work Category is likely to be more than 400 days in alignment with overall trend.
7. Root/Sewar/Sidewalk has more cases in higher time range with Brooklyn has higher density than others. Sidewalk Complaint in Brooklyn there are more chances that it will take more time to close than other Borough.
8. Response time on service request is dependent on Inspection delay time which is dependent on Type of request like tree removal is prioritized over others.
9. Inspection time is dependent on Inspection structure like full inspection take more time than partial even for level 1 inspection

Machine Learning Model

Preparation

1. Most of the data in dataset is categorical data.
2. Categorical data is Dummyfies using Caret Package.
3. Dataset after dummyfing has 173 variables.
4. Supervised learning is used by splitting dataset in Training data and test data in 70:30 Ratio.
5. While fitting the model, using LM function system is taking more than 50 Min. We had to use H2o.ai package to speed up.

Model Fitting and Performance

While fitting the model using linear regression, system is taking lot of time due to wide data set. We are using H2o.ai to run Generalized Regression Model, Random Forest and Gradient Boosting method model.

Summary of performance of model on Training and Test Dataset a below:

- ***Generalized Linear Modeling using H20.ai***

Dataset	RMSE	MAE	R2
Training	121.4544	85.53995	0.28528692
Test	121.6521	85.76100	0.2813671

- ***Random forest algorithm using H2o.ai***

Dataset	RMSE	MAE	R2
Training	136.5609	102.6122	0.096437922
Test	136.4425	102.5933	0.09600275

- ***Gradient boosting algorithm using H20.ai***

Dataset	RMSE	MAE	R2
Training	108.0567	73.21972	0.43427122
Test	108.6959	73.73865	0.4262872

On all three models Training and Test data set are giving similar performance.

Model using Random Forest algorithm has worst performance

Out if three models GBM is giving best result.

Data set has lot of noise that's why on 43% of variability can be explained.

Actual vs Predicted values on Test Dataset

Comparing predicted values with actual test set

actual	GLM	Random Forest	GBM
Min. : 0	Min. :-96.25	Min. : 87.75	Min. :-65.76
1st Qu.: 9	1st Qu.: 79.13	1st Qu.:104.51	1st Qu.: 67.42
Median : 48	Median :112.02	Median :118.45	Median :134.71
Mean :112	Mean :119.81	Mean :137.71	Mean :130.80
3rd Qu.:167	3rd Qu.:164.46	3rd Qu.:182.30	3rd Qu.:162.48
Max. :723	Max. :414.29	Max. :206.41	Max. :630.79
NA's :20296			

Even though GLM has R2 of 28 % , predicted values by GLM are much closure to actual values than other two models.

Variable importance

Variable importance as generated by GBM as well as Random forest provide some key insight into importance of variables on the output .

Variable importance Random Forest:

	Variable	relative_importance	scaled_importance	percentage
1	SRCategory.Plant.Tree	14940109824	1.0000000	0.10783034
2	SRType.Street.Tree	13009389568	0.8707693	0.09389536
3	WOCategory.Tree.Planting	12017478656	0.8043769	0.08673623
4	BoroughCode.Brooklyn	8447867392	0.5654488	0.06097254
5	SRResolution.Planting.Dec	5510671872	0.3688508	0.03977331
6	Longitude.x	3422561792	0.2290855	0.02470236

Variable importance GBM:

	Variable	relative_importance	scaled_importance	percentage
1	SRCategory.Plant.Tree	11124196352	1.0000000	0.22275542
2	SRCategory.Prune	5271790080	0.4739030	0.10556446
3	BoroughCode.Brooklyn	3749020928	0.3370150	0.07507191
4	SRResolution.Work.Completed	3221082624	0.2895564	0.0645002
5	SRResolution.Reviewed...Inspection.Assigned		0.2792948	0.06221443
6	BoroughCode.Staten.Island	2283439360	0.2052678	0.04572452

Recommendation

Our analysis of request shows that response to service request related to tree vary a lot from 1 days to 600 days as there are various categories for request. Most of the cases where time period is high due to unavailability of proper information or dependencies on other dept like Sidewalk repair etc.

Recommendations for Forestry department:

- This model can used to set up SLA for type of request and estimated timeline based on dependency.
- Request accurate information to be sufficient to estimate SLA for service request as we have seen Inspection Delay time and Non-availability of information are main cause of delay.
- Learn from the model and provide immediate possibility of solution based in learnings instead of delaying for inspection and then denying the request. While analyzing SR resolution lot of cases are denied after inspection and time.
- Complaint type Street tree and Sidewalk repairs are cases where coordination with other dept. takes lot of time and are mostly delayed. Setup a communication process for joint inspection to reduce delay in work estimation.

Further Work

For future work:

- Analyze relationship between Tree species and type or count of request raised. While analysis we found that specific species are contributing more on specific service request.
- Request related to Street Tree, Pruning and Sidewalk repair can be shared with Traffic dept to prevent accidents related to these tree requests. One can study traffic accidents related to tree service request.
- Refine this model with accurate data and removing some categories which are not contributing much in model. Some cases update time is lower than created time which is not logically possible.
- While analysis we found that Brooklyn cases has higher range of response time but unable to find any correlation with any parameters. Most probable reason we assumed is variation of cases. We can study further to find root cause of variation in response time.
- Also, we have seen there are repeated calls from same location as well as on same tree point which means either response is delayed or solution is not working. Forestry dept can look into cause of repeated calls to prevent expense of effort on these cases.