1. Summarize this dataset. The actual analytics and metrics are left up to your discretion. This problem is left intentionally unstructured, so just include a file or notebook that describes quantitatively what this dataset contains.

Answer:

The notebook file: StreetTreeCensus.ipynb

The output of the file running in the Jupyter Notebook is given below:

```
Total Number Of Trees: 683788

Mean of Breast Height of Tree, Stump Diameter:  11.27978701000895, 0.43246298
560372515
Standard Deviation of Breast Height of Tree, Stump Diameter: 8.72304226854944
4, 3.2902407401961704

Total Tree On Curb, Total Tree Offset From Curb:  656896, 26892

Dead Tree, Alive Tree, Stump Tree:  13961, 652173, 17654

Good Tree, Fair Tree, Poor Tree:  528850, 96504, 26818
Top 10 borough: Trees with 'Good' health
[Stage 24:====================================================> (195 + 4) / 2
00]+-------------+------+
|     boroname| count|
+-------------+------+
|       Queens|194008|
|     Brooklyn|138212|
|Staten Island| 82669|
|        Bronx| 66603|
|    Manhattan| 47358|
+-------------+------+

Top 10 borough: Trees with 'Poor' health
+-------------+-----+
|     boroname|count|
+-------------+-----+
|       Queens| 9417|
|     Brooklyn| 6459|
|Staten Island| 4238|
|    Manhattan| 3609|
|        Bronx| 3095|
+-------------+-----+
Percentage of Trees Having Good Health :  77.34%
Percentage of Trees Having Poor Health :  3.92%
Top 20 Tree Name by Thier Count
+-----------------+-----+
|        Tree Name|Count|
+-----------------+-----+
| London planetree|87014|
|      honeylocust|64264|
|     Callery pear|58931|
|          pin oak|53185|
|     Norway maple|34189|
|littleleaf linden|29742|
|           cherry|29279|
| Japanese zelkova|29258|
|           ginkgo|21024|
|          Sophora|19338|
|        red maple|17246|
|        green ash|16251|
|  American linden|13530|
|     silver maple|12277|
|         sweetgum|10657|
| northern red oak| 8400|
|    silver linden| 7995|
|     American elm| 7975|
|            maple| 7080|
| purple-leaf plum| 6879|
+-----------------+-----+
only showing top 20 rows

Total Gaurds , Helpful Gaurds , Harmful Gaurds , Niether Helpful nor Harmful,
Unsure: 652172, 51866, 20252, 572306, 7748
Percentage of Helpful Gaurd  :  7.95%
Percentage of Harmful Gaurd  :  3.11%
Percentage of Neither Helpful not Harmful Gaurd  :  87.75%
Percentable of Tree With Root Problem:  20.47
Percentable of Tree With Trunk Problem:  6.86
Percentable of Tree With Branch Problem:  12.74
Top 20 Tree Count by zip_city
+-------------------+------+
|           Zip City| Count|
+-------------------+------+
|           Brooklyn|177300|
|      Staten Island|105318|
|              Bronx| 85203|
|           New York| 64488|
|            Jamaica| 26028|
|           Flushing| 23389|
|          Ridgewood| 10937|
|      Fresh Meadows| 10441|
|     Queens Village| 10127|
|            Astoria| 10007|
|         Whitestone|  9449|
|            Bayside|  8679|
|Springfield Gardens|  7470|
|        Little Neck|  7280|
|       Forest Hills|  7059|
|    Oakland Gardens|  7054|
|       Far Rockaway|  6887|
|      East Elmhurst|  6475|
|           Rosedale|  6324|
|           Woodside|  5651|
+-------------------+------+
only showing top 20 rows

Bottom 20 Tree Count by zip_city
+-------------------+-----+
|           Zip City|Count|
+-------------------+-----+
|             Inwood|    9|
|       Breezy Point|   30|
|      New Hyde Park|  865|
|       Central Park|  935|
|        Floral Park| 1539|
|          Sunnyside| 1664|
|        Kew Gardens| 1743|
|            Arverne| 2013|
|South Richmond Hill| 2805|
|          Woodhaven| 2855|
|          Rego Park| 3084|
|      College Point| 3099|
|    Cambria Heights| 3229|
|    Jackson Heights| 3295|
|      Richmond Hill| 3391|
|   Long Island City| 3479|
|      Rockaway Park| 3572|
|             Hollis| 3591|
|            Maspeth| 4033|
|          Glen Oaks| 4130|
+-------------------+-----+
only showing top 20 rows

Tree Count by borough
+-------------+------+
|      borough| Count|
+-------------+------+
|       Queens|250551|
|     Brooklyn|177293|
|Staten Island|105318|
|        Bronx| 85203|
|    Manhattan| 65423|
+-------------+------+
```

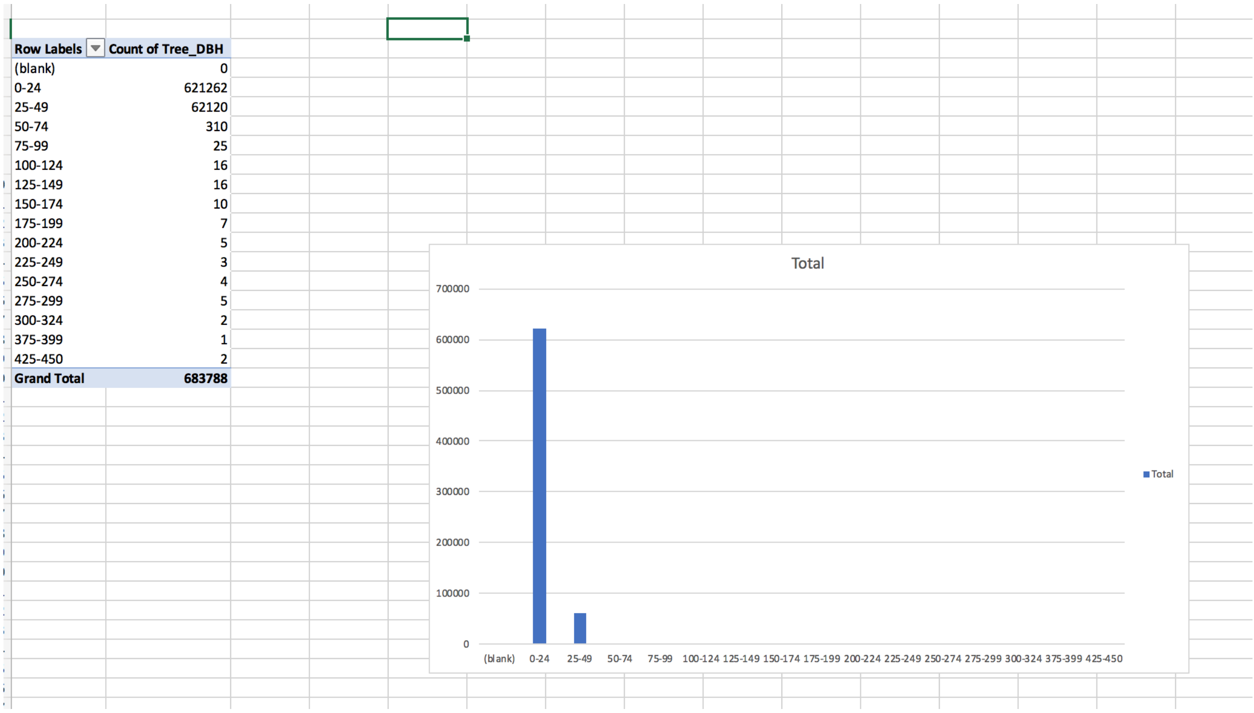
Question 2. Write code that will display the number of alive trees by species name (common) and by borough. Include totals and percent of totals. Usage of Spark or Pandas is acceptable.

The notebook file name is: AliveStreetStat.ipynb

The output of the program is given below:

```
|          spc_common|     boroname|count(status)|
+--------------------+-------------+-------------+
|             NO NAME|        Bronx|            1|
|             NO NAME|       Queens|            4|
|'Schubert' chokec...|        Bronx|          575|
|'Schubert' chokec...|     Brooklyn|         1308|
|'Schubert' chokec...|    Manhattan|          163|
|'Schubert' chokec...|       Queens|         2013|
|'Schubert' chokec...|Staten Island|          829|
|      American beech|        Bronx|           31|
|      American beech|     Brooklyn|           83|
|      American beech|    Manhattan|           22|
|      American beech|       Queens|           88|
|      American beech|Staten Island|           49|
|        American elm|        Bronx|         1471|
|        American elm|     Brooklyn|         2587|
|        American elm|    Manhattan|         1698|
|        American elm|       Queens|         1709|
|        American elm|Staten Island|          510|
|American hophornbeam|        Bronx|          185|
|American hophornbeam|     Brooklyn|          366|
|American hophornbeam|    Manhattan|           84|
+--------------------+-------------+-------------+
only showing top 20 rows

Total Alive Tree: 652173
Total Percentage of Alive Tree: 95.38
```

Question 3. Create a histogram for tree_dbh (diameter of tree). What is the 90% percentile diameter? Create a visual depiction of this histogram.



Row Labels	Count of Tree_DBH
(blank)	0
0-24	621262
25-49	62120
50-74	310
75-99	25
100-124	16
125-149	16
150-174	10
175-199	7
200-224	5
225-249	3
250-274	4
275-299	5
300-324	2
375-399	1
425-450	2
Grand Total	**683788**

The 90% percentile is: 25-49

The spark-shell code is given below:

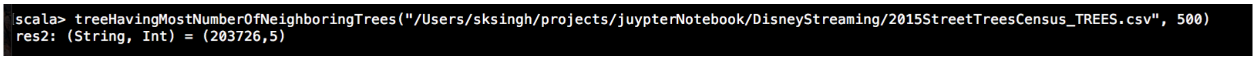
```
// read the file
val treeDF =
spark.read.format("csv").option("header","true").load("/Users/sksingh/projects/juypterNotebook/DisneyStreamin
g/2015StreetTreesCensus_TREES.csv")

// create spark sql view
spark.createOrReplaceTempView("tree_dbh_table")


// select tree_dbh
val treeDbhDF =  spark.sql("select tree_dbh from tree_dbh_table  order by tree_dbh")

// save the result as csv to draw the graph

treeDbhDF.coalesce(1).write.format("csv").save("/Users/sksingh/Downloads/treeDBH.csv")/treeDBH.csv
```

Question 4. Write a program to determine which tree(s) have the most number of neighboring trees within a 500 foot radius? Do not use Pandas, Spark or any high-level library or module. The objective is for you to demonstrate your ability to write an efficient algorithm. What is the complexity of your solution?

The notebook file name is: TreeWithMostNumberOfNeighboringTrees.ipynb

The answer is the tree_id: 203726 with the count 5

The output screen shot:



```
scala> treeHavingMostNumberOfNeighboringTrees("/Users/sksingh/projects/juypterNotebook/DisneyStreaming/2015StreetTreesCensus_TREES.csv", 500)
res2: (String, Int) = (203726,5)
```

The complexity – Big(O) – is n^2