```
# import python packages
import pandas as pd
print("import package libraries")
```

```
import package libraries
```

```
# load dataset
tree_census = pd.read_csv('trees.csv')
print("load dataset may take long to load")
```

```
load dataset may take long to load
```

```
# look at the first five rows
tree_census.head()
```

	created_at	tree_id	block_id	the_geom	tree_dbh	stump_diam	curb_loc	status	health	spc_latin	...	st_assem	st_senat
0	8/27/2015	180,683	348,711	POINT (-73.84421521958048 40.723091773924274)	3	0	OnCurb	Alive	Fair	Acer rubrum	...	28.0	16.(
1	9/3/2015	200,540	315,986	POINT (-73.81867945834878 40.79411066708779)	21	0	OnCurb	Alive	Fair	Quercus palustris	...	27.0	11.(
2	9/5/2015	204,026	218,365	POINT (-73.93660770459083 40.717580740099116)	3	0	OnCurb	Alive	Good	Gleditsia triacanthos var. inermis	...	50.0	18.(
3	9/5/2015	204,337	217,969	POINT (-73.93445615919741 40.713537494833226)	10	0	OnCurb	Alive	Good	Gleditsia triacanthos var. inermis	...	53.0	18.(
4	8/30/2015	189,565	223,043	POINT (-73.97597938483258 40.66677775537875)	21	0	OnCurb	Alive	Good	Tilia americana	...	44.0	21.(

5 rows × 42 columns

```
# look at the last five rows
tree_census.tail()
```

	created_at	tree_id	block_id	the_geom	tree_dbh	stump_diam	curb_lo
27164	9/13/2015	220,123	340,784	POINT (-73.87119241987155 40.75029650437607)	2	0	OnCur
27165	9/17/2015	232,789	515,680	POINT (-73.85265467406566 40.895612923851026)	18	0	OnCur
27166	9/11/2015	215,243	515,114	POINT (-73.84831727384334 40.89490581253175)	13	0	OnCur
27167	9/14/2015	222,536	314,720	POINT (-73.84218839558766 40.78471185843157)	9	0	OnCur
27168	9/14/2015	222,170	229,849	POINT (-73.95796197977943 40.62209316291482)	2	0	OnCur

5 rows × 42 columns

```
# list of column names
tree_census.columns
```

```
Index(['created_at', 'tree_id', 'block_id', 'the_geom', 'tree_dbh',
       'stump_diam', 'curb_loc', 'status', 'health', 'spc_latin', 'spc_common',
       'steward', 'guards', 'sidewalk', 'user_type', 'problems', 'root_stone',
       'root_grate', 'root_other', 'trnk_wire', 'trnk_light', 'trnk_other',
       'brnch_ligh', 'brnch_shoe', 'brnch_othe', 'address', 'zipcode',
       'zip_city', 'cb_num', 'borocode', 'boroname', 'cncldist', 'st_assem',
```

```
       'st_senate', 'nta', 'nta_name', 'boro_ct', 'state', 'Latitude',
       'longitude', 'x_sp', 'y_sp'],
      dtype='object')
```

```
# identify the size, number of rows and columns in the dataset
tree_census.shape
```

```
(27169, 42)
```

```
# summary of the dataset
tree_census.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27169 entries, 0 to 27168
Data columns (total 42 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   created_at  27169 non-null  object
 1   tree_id     27169 non-null  object
 2   block_id    27169 non-null  object
 3   the_geom    27169 non-null  object
 4   tree_dbh    27169 non-null  int64
 5   stump_diam  27169 non-null  int64
 6   curb_loc    27169 non-null  object
 7   status      27169 non-null  object
 8   health      25810 non-null  object
 9   spc_latin   25809 non-null  object
 10  spc_common  25809 non-null  object
 11  steward     8432 non-null   object
 12  guards      4323 non-null   object
 13  sidewalk    25809 non-null  object
 14  user_type   27168 non-null  object
 15  problems    11338 non-null  object
 16  root_stone  27168 non-null  object
 17  root_grate  27168 non-null  object
 18  root_other  27168 non-null  object
 19  trnk_wire   27168 non-null  object
 20  trnk_light  27168 non-null  object
 21  trnk_other  27168 non-null  object
 22  brnch_ligh  27168 non-null  object
 23  brnch_shoe  27168 non-null  object
 24  brnch_othe  27168 non-null  object
 25  address     27168 non-null  object
 26  zipcode     27168 non-null  float64
 27  zip_city    27168 non-null  object
 28  cb_num      27168 non-null  float64
 29  borocode    27168 non-null  float64
 30  boroname    27168 non-null  object
 31  cncldist    27168 non-null  float64
 32  st_assem    27168 non-null  float64
 33  st_senate   27168 non-null  float64
 34  nta         27168 non-null  object
 35  nta_name    27168 non-null  object
 36  boro_ct     27168 non-null  float64
 37  state       27168 non-null  object
 38  Latitude    27168 non-null  float64
 39  longitude   27168 non-null  float64
 40  x_sp        27168 non-null  object
 41  y_sp        27168 non-null  object
dtypes: float64(9), int64(2), object(31)
memory usage: 8.7+ MB
```

```
# health status of trees
tree_census.status.value_counts(dropna=False)
```

```
status
Alive    25810
Stump      787
Dead       572
Name: count, dtype: int64
```

```
# get status on the trees
tree_census.status.value_counts(dropna=False)
```

```
status
Alive    25810
Stump      787
Dead       572
Name: count, dtype: int64
```

```
# subset of the original, removed columns not interested in
trees_subset = tree_census
[['tree_id', 'tree_dbh',
            'stump_diam', 'curb_loc', 'status', 'health', 'spc_latin', 'spc_common',
            'steward', 'guards', 'sidewalk', 'user_type', 'problems', 'root_stone',
            'root_grate', 'root_other', 'trink_wire', 'trnk_light', 'trnk_other',
            'brnch_light', 'brnch_shoe', 'brnch_othe']]
```

```
  # list the first 5 rows of the new subset
tree_census.head()
```

	created_at	tree_id	block_id	the_geom	tree_dbh	stump_diam	curb_loc	st
0	8/27/2015	180,683	348,711	POINT (-73.84421521958048 40.723091773924274)	3	0	OnCurb	
1	9/3/2015	200,540	315,986	POINT (-73.81867945834878 40.79411066708779)	21	0	OnCurb	
2	9/5/2015	204,026	218,365	POINT (-73.93660770459083 40.717580740099116)	3	0	OnCurb	
3	9/5/2015	204,337	217,969	POINT (-73.93445615919741 40.713537494833226)	10	0	OnCurb	
4	8/30/2015	189,565	223,043	POINT (-73.97597938483258 40.66677775537875)	21	0	OnCurb	

5 rows × 42 columns

```
  # check for any null values
trees_subset.isna().sum()
```

```
created_at        0
tree_id           0
block_id          0
the_geom          0
tree_dbh          0
stump_diam        0
curb_loc          0
status            0
health         1359
spc_latin      1360
spc_common     1360
steward       18737
guards        22846
sidewalk       1360
user_type         1
problems      15831
root_stone        1
root_grate        1
root_other        1
trnk_wire         1
trnk_light        1
trnk_other        1
brnch_ligh        1
brnch_shoe        1
brnch_othe        1
address           1
zipcode           1
zip_city          1
cb_num            1
borocode          1
boroname          1
cncldist          1
st_assem          1
st_senate         1
nta               1
nta_name          1
boro_ct           1
state             1
Latitude          1
longitude         1
```

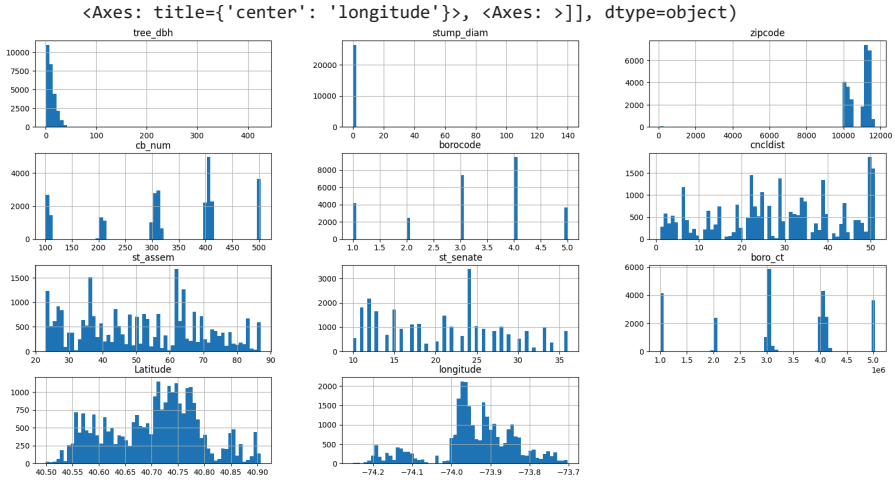
```
x_sp              1
y_sp              1
dtype: int64
```

```
# show all that are none values in health, alot of missing values NaN
tree_census.describe()
```

	tree_dbh	stump_diam	zipcode	cb_num	borocode	cncldist
count	27169.000000	27169.000000	27168.000000	27168.000000	27168.000000	27168.000000
mean	11.186978	0.470610	10874.970370	328.702739	3.220075	28.512036
std	8.695769	3.433636	702.029211	123.049039	1.240908	14.854502
min	0.000000	0.000000	83.000000	101.000000	1.000000	1.000000
25%	4.000000	0.000000	10312.000000	301.000000	3.000000	19.000000
50%	9.000000	0.000000	11211.000000	317.000000	3.000000	29.000000
75%	16.000000	0.000000	11361.000000	408.000000	4.000000	40.000000
max	425.000000	140.000000	11697.000000	503.000000	5.000000	51.000000

```
# generate histogram of data distribution
trees_subset.hist(bins=60, figsize=(20,10))
```

```
array([[<Axes: title={'center': 'tree_dbh'}>,
        <Axes: title={'center': 'stump_diam'}>,
        <Axes: title={'center': 'zipcode'}>],
       [<Axes: title={'center': 'cb_num'}>,
        <Axes: title={'center': 'borocode'}>,
        <Axes: title={'center': 'cncldist'}>],
       [<Axes: title={'center': 'st_assem'}>,
        <Axes: title={'center': 'st_senate'}>,
        <Axes: title={'center': 'boro_ct'}>],
       [<Axes: title={'center': 'Latitude'}>,
        <Axes: title={'center': 'longitude'}>, <Axes: >]], dtype=object)
```



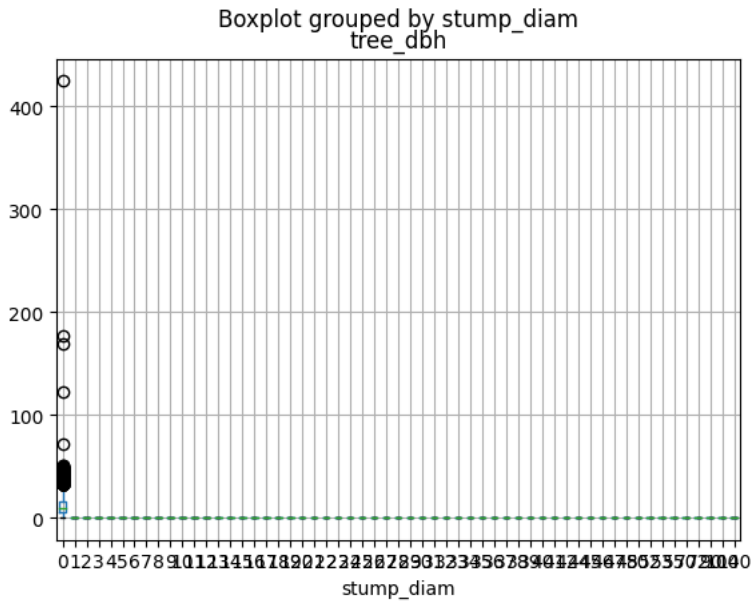
```
# trees larger than 50
big_trees = trees_subset[trees_subset['tree_dbh']>50]
big_trees.head()
```

	created_at	tree_id	block_id	the_geom	tree_dbh	stump_diam	cu
2385	8/23/2015	168,583	226,040	POINT (-73.94693592253036 40.67228657645615)	425	0	
3724	9/3/2015	199,546	315,695	POINT (-73.80329113201749 40.78987155371557)	51	0	
4874	8/12/2015	139,665	409,474	POINT (-74.09171228313842 40.57236260308215)	72	0	OffsetFrc
6711	9/8/2015	209,349	415,127	POINT (-74.11595934608093 40.562379364379545)	122	0	
10053	9/11/2015	215,075	515,054	POINT (-73.84720553041983 40.89488599898038)	169	0	

5 rows × 42 columns

```
# box plot
tree_census.boxplot(column='tree_dbh', by='stump_diam')
```

```
<Axes: title={'center': 'tree_dbh'}, xlabel='stump_diam'>
```




```
# scatter plot
big_trees[['tree_id', 'tree_dbh']].plot(kind='scatter', x='tree_id', y='tree_dbh')
```

```
<Axes: xlabel='tree_id', ylabel='tree_dbh'>
```