Split Data

Split data into training, validation, and testing datasets.

Training - 70% of records

Validation - 15% of records

Testing - 15% of records

Will not be shuffling records since the data is fairly random in value distribution as-is.

Imports

In [1]:
```
import numpy as np
import pandas as pd
```

Load Data

In [2]:
```
data = None
dir = "data/"
with open("%sVulnerabilityData_csv.csv"%dir) as file:
    data = pd.read_csv(file)
```

Split Data

In [3]:
```
l = len(data)
l1 = int(l*0.7)
l2 = l1 + int(l*0.15)
print("Indicies: \nTrain:\t%d\t%d\nVal:\t%d\t%d\nTest:\t%d\t%d\n"%(0, l1, l1,
l2, l2, l))

train = data.iloc[:l1]
val = data.iloc[l1:l2]
test = data.iloc[l2:]
```

```
Indicies:
Train:  0       21000
Val:    21000   25500
Test:   25500   30000
```

In [4]:
```
train
```

Out[4]:

	Property	Occupancy	Construction	YearBuilt	Floor	SquareFootage	Windspeed	Los
0	1	MultiFamily	Masonry	1995-2005	1	1700+	1	0.3061
1	2	SingleFamily	Concrete	2005+	1	<1500	3	0.1102
2	3	SingleFamily	Masonry	<=1995	0	<1500	2	0.2125
3	4	SingleFamily	Masonry	2005+	1	1500-1700	3	0.1679
4	5	MultiFamily	Frame	<=1995	1	1500-1700	1	0.3401
...	...	...	...	...	...	...	...	
20995	20996	MultiFamily	Frame	<=1995	2+	1500-1700	3	0.5248
20996	20997	MultiFamily	Frame	2005+	2+	1500-1700	2	0.2834
20997	20998	SingleFamily	Frame	<=1995	1	<1500	2	0.2361
20998	20999	SingleFamily	Frame	1995-2005	1	1500-1700	1	0.2550
20999	21000	MultiFamily	Frame	2005+	0	<1500	2	0.1913

21000 rows × 8 columns

In [5]:
```
val
```

Out[5]:

	Property	Occupancy	Construction	YearBuilt	Floor	SquareFootage	Windspeed	Los
21000	21001	SingleFamily	Concrete	<=1995	1	<1500	5	0.2268
21001	21002	SingleFamily	Masonry	2005+	1	<1500	3	0.1259
21002	21003	SingleFamily	Frame	<=1995	1	1700+	4	0.4860
21003	21004	SingleFamily	Concrete	1995-2005	2+	1500-1700	3	0.2755
21004	21005	SingleFamily	Frame	2005+	2+	1500-1700	2	0.2361
...	...	...	...	...	...	...	...	
25495	25496	SingleFamily	Frame	<=1995	2+	1500-1700	1	0.3542
25496	25497	SingleFamily	Frame	2005+	2+	<1500	2	0.1771
25497	25498	MultiFamily	Concrete	2005+	1	1500-1700	1	0.1428
25498	25499	SingleFamily	Frame	1995-2005	2+	1500-1700	1	0.3188
25499	25500	MultiFamily	Masonry	1995-2005	0	1500-1700	3	0.3401

4500 rows × 8 columns

In [6]:
```
test
```

Out[6]:

	Property	Occupancy	Construction	YearBuilt	Floor	SquareFootage	Windspeed	Los
25500	25501	SingleFamily	Frame	<=1995	0	1500-1700	4	0.4374(
25501	25502	SingleFamily	Frame	<=1995	2+	1500-1700	1	0.3542(
25502	25503	MultiFamily	Frame	<=1995	1	<1500	1	0.2550(
25503	25504	MultiFamily	Concrete	2005+	1	<1500	3	0.1322(
25504	25505	SingleFamily	Frame	2005+	2+	1500-1700	2	0.2361(
...	...	...	...	...	...	...	...	
29995	29996	MultiFamily	Concrete	<=1995	1	1700+	4	0.4082(
29996	29997	MultiFamily	Frame	1995-2005	1	<1500	1	0.2295(
29997	29998	SingleFamily	Frame	2005+	2+	1700+	1	0.2657(
29998	29999	MultiFamily	Frame	1995-2005	1	<1500	2	0.2550(
29999	30000	SingleFamily	Concrete	1995-2005	1	1500-1700	3	0.2204(

4500 rows × 8 columns

Save Data

In [8]:
```
train.to_csv("%strain.csv"%dir)
val.to_csv("%sval.csv"%dir)
test.to_csv("%stest.csv"%dir)
```