Seattle Energy Analysis

April 22, 2019

1 Seattle Energy Analysis

This data comes from the City of Seattle's Open Data program that hosts a large number of free datasets for analytic use. The description of this specific program is as follows:

Seattle's Building Energy Benchmarking Program (SMC 22.920) requires owners of non-residential and multifamily buildings (20,000 square feet or larger) to track energy performance and annually report to the City of Seattle. Annual benchmarking, reporting and disclosing of building performance are foundational elements of creating more market value for energy efficiency.

Per Ordinance (125000), starting with 2015 energy use performance reporting, the City of Seattle will make the data for all building 20,000 SF and larger available annually. This dataset contains all 2017 buildings required to report.

For our purposes, we will use this as an exercise for cleaning and preparing data for visual analysis, as well as for building machine learning models.

1.1 Initial reading and exploration of data

```
In [13]: data<- read.csv("/Users/alecduggan/2017_Building_Energy_Benchmarking.csv", header = T
         options(warn = -1)
```

```
In [14]: head(data)
```

OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName
1	2017	NonResidential	Hotel	Mayflower park hotel
2	2017	NonResidential	Hotel	Paramount Hotel
3	2017	Campus	Hotel	84SC9-The Westin Seattle
5	2017	NonResidential	Hotel	HOTEL MAX
8	2017	NonResidential	Hotel	WARWICK SEATTLE HOTEL
9	2017	Nonresidential COS	Other	West Precinct

```
In [15]: str(data)
```

```
'data.frame':        3461 obs. of  45 variables:
 $ OSEBuildingID                : int  1 2 3 5 8 9 10 11 12 13 ...
 $ DataYear                     : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
 $ BuildingType                 : Factor w/ 8 levels "Campus","Multifamily HR (10+)",..: 5 5
 $ PrimaryPropertyType          : Factor w/ 24 levels "","Distribution Center",..: 5 5 5 5 5
 $ PropertyName                 : Factor w/ 3428 levels "","(71367A) SEATTLE Macy's",..: 191
 $ Address                        : Factor w/ 3437 levels "10 Harrison St.",..: 2092 3045 1073
 $ City                           : Factor w/ 14 levels "Ballard","King",..: 9 9 9 11 9 9 9 9
 $ State                          : Factor w/ 6 levels "CA","CO","WA",..: 3 3 3 3 3 3 3 3 3 3
 $ ZipCode                        : int  98101 98101 98101 98101 98121 98101 98101 98101 98104
 $ TaxParcelIdentificationNumber  : Factor w/ 3340 levels "0000000000","0001800021",..: 172 176
 $ CouncilDistrictCode            : int  7 7 7 7 7 7 7 7 7 7 ...
 $ Neighborhood                   : Factor w/ 20 levels "","Ballard","BALLARD",..: 8 8 8 8 8 8
 $ Latitude                       : num  47.6 47.6 47.6 47.6 47.6 ...
 $ Longitude                      : num  -122 -122 -122 -122 -122 ...
 $ YearBuilt                      : int  1927 1996 1969 1926 1980 1999 1926 1926 1904 1910 ...
 $ NumberofBuildings              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ NumberofFloors                 : int  12 11 41 10 18 2 11 8 15 6 ...
 $ PropertyGFATotal               : int  88434 103566 956110 61320 175580 97288 83008 102761 16
 $ PropertyGFAParking             : int  0 15064 196718 0 62000 37198 0 0 0 1496 ...
 $ PropertyGFABuilding.s.         : int  88434 88502 759392 61320 113580 60090 83008 102761 163
 $ ListOfAllPropertyUseTypes      : Factor w/ 480 levels "","Adult Education",..: 162 169 174
 $ LargestPropertyUseType         : Factor w/ 57 levels "","Adult Educa",..: 16 16 16 16 16 42
 $ LargestPropertyUseTypeGFA      : num  88434 83880 756493 61320 123445 ...
 $ SecondLargestPropertyUseType   : Factor w/ 51 levels "","Adult Education",..: 1 36 36 1 36
 $ SecondLargestPropertyUseTypeGFA: int  NA 15064 138635 NA 68009 40971 NA NA NA NA ...
 $ ThirdLargestPropertyUseType    : Factor w/ 48 levels "","Bank Branch",..: 1 40 46 1 46 1 1
 $ ThirdLargestPropertyUseTypeGFA : int  NA 4622 0 NA 0 NA NA NA NA NA ...
 $ YearsENERGYSTARCertified       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ ENERGYSTARScore                : int  63 72 48 51 78 NA 33 NA 44 2 ...
 $ SiteEUI.kBtu.sf.               : num  83.2 88.2 98.4 120.2 116.1 ...
 $ SiteEUIWN.kBtu.sf.             : num  82.3 86.8 98.2 119 114.1 ...
 $ SourceEUI.kBtu.sf.             : num  184 164 243 234 210 ...
 $ SourceEUIWN.kBtu.sf.           : num  182 160 243 230 206 ...
 $ SiteEnergyUse.kBtu.            : num  7361655 7804844 74470328 7372222 14335778 ...
 $ SiteEnergyUseWN.kBtu.          : num  7274452 7678810 74311368 7294312 14081251 ...
 $ SteamUse.kBtu.                 : num  2122836 NA 24313482 2228120 NA ...
 $ Electricity.kWh.               : num  1157783 884161 14276917 881745 1523506 ...
 $ Electricity.kBtu.              : num  3950356 3016757 48712841 3008514 5198202 ...
 $ NaturalGas.therms.             : num  12885 47881 14440 21356 91376 ...
 $ NaturalGas.kBtu.               : num  1288463 4788087 1444000 2135588 9137576 ...
 $ TotalGHGEmissions              : num  198 267 1571 244 507 ...
 $ GHGEmissionsIntensity          : num  2.23 2.58 1.64 3.98 2.89 ...
 $ DefaultData                    : Factor w/ 2 levels "N","Y": 1 1 1 1 1 2 1 1 1 1 ...
 $ ComplianceStatus               : Factor w/ 2 levels "Compliant","Not Compliant": 1 1 1 1 1
 $ Outlier                        : Factor w/ 3 levels "","High outlier",..: 1 1 1 1 1 1 1 1 1
```

```
In [16]: summary(data)
```

```
 OSEBuildingID      DataYear                  BuildingType
 Min.   :    1   Min.   :2017   NonResidential      :1480
 1st Qu.:20033   1st Qu.:2017   Multifamily LR (1-4):1042
 Median :23212   Median :2017   Multifamily MR (5-9): 620
```

```
Mean   :21719   Mean   :2017   Multifamily HR (10+): 112
3rd Qu.:26147   3rd Qu.:2017   SPS-District K-12   :  99
Max.   :50289   Max.   :2017   Nonresidential COS  :  67
                               (Other)             :  41
                 PrimaryPropertyType           PropertyName
Low-Rise Multifamily       :1011                    :  21
Mid-Rise Multifamily       : 607     Northgate Plaza :   3
Small- and Mid-Sized Office: 297     Airport Way     :   2
Other                      : 247     Bayview Building:   2
Large Office               : 180     Canal Building  :   2
Warehouse                  : 175     Central Park    :   2
(Other)                    : 944     (Other)         :3429
                       Address           City      State
2203 Airport Way S         :   4   Seattle    :3209   CA:   1
2600 SW Barton St          :   4   SEATTLE    : 185   CO:   1
309 South Cloverdale Street:   4   seattle    :  45   WA:3453
2400 11th Ave East         :   3   Seatle     :   5   WI:   1
100 West Harrison          :   2   Seatt;e    :   4   WQ:   4
10510 5th Ave NE           :   2   Seattle, WA:   3   WV:   1
(Other)                    :3442   (Other)    :  10
   ZipCode      TaxParcelIdentificationNumber CouncilDistrictCode
Min.   :98006   1625049001:  21               Min.   :1.000
1st Qu.:98105   0925049346:   6               1st Qu.:3.000
Median :98115   0002400002:   5               Median :4.000
Mean   :98117   3224049012:   5               Mean   :4.411
3rd Qu.:98122   3624039009:   4               3rd Qu.:7.000
Max.   :98272   7666203240:   4               Max.   :7.000
NA's   :21      (Other)   :3416               NA's   :17
               Neighborhood     Latitude       Longitude        YearBuilt
DOWNTOWN             : 586   Min.   :47.50   Min.   :-122.4   Min.   :1900
EAST                 : 449   1st Qu.:47.60   1st Qu.:-122.4   1st Qu.:1949
MAGNOLIA / QUEEN ANNE: 429   Median :47.62   Median :-122.3   Median :1976
GREATER DUWAMISH     : 368   Mean   :47.62   Mean   :-122.3   Mean   :1970
NORTHEAST            : 300   3rd Qu.:47.66   3rd Qu.:-122.3   3rd Qu.:1998
LAKE UNION           : 264   Max.   :47.73   Max.   :-122.3   Max.   :2017
(Other)              :1065   NA's   :17      NA's   :17       NA's   :1
NumberofBuildings NumberofFloors   PropertyGFATotal  PropertyGFAParking
Min.   :  0.000   Min.   : 0.000   Min.   :  20000   Min.   :     0
1st Qu.:  1.000   1st Qu.: 2.000   1st Qu.:  28800   1st Qu.:     0
Median :  1.000   Median : 4.000   Median :  45000   Median :     0
Mean   :  1.131   Mean   : 4.759   Mean   :  97555   Mean   : 16514
3rd Qu.:  1.000   3rd Qu.: 5.000   3rd Qu.:  93191   3rd Qu.:  6567
Max.   :111.000   Max.   :76.000   Max.   :9320156   Max.   :686750
NA's   :21        NA's   :7                          NA's   :1317
PropertyGFABuilding.s.                ListOfAllPropertyUseTypes
Min.   :   3636       Multifamily Housing         : 864
1st Qu.:  28302       Multifamily Housing, Parking: 513
Median :  47996       Office                      : 144
```

```
Mean   : 100814        K-12 School                 : 138
3rd Qu.:  95898        Office, Parking             : 127
Max.   :9320156        Non-Refrigerated Warehouse  :  94
NA's   :1317           (Other)                     :1581
LargestPropertyUseType LargestPropertyUseTypeGFA SecondLargestPropertyUseType
Multifamily:1739       Min.   :   5656                       :1715
Office     : 509       1st Qu.:  25368           Parking     :1053
Non-Refrige: 187       Median :  40361           Office      : 209
K-12 School: 143       Mean   :  80277           Retail Store: 153
Other      :  93       3rd Qu.:  78440           Other       :  55
Retail Stor:  93       Max.   :9236849           Restaurant  :  41
(Other)    : 697       NA's   :22                (Other)     : 235
SecondLargestPropertyUseTypeGFA ThirdLargestPropertyUseType
Min.   :     0                              :2851
1st Qu.:  5000                  Retail Store: 111
Median : 10840                  Office      : 106
Mean   : 28909                  Parking     :  70
3rd Qu.: 27414                  Restaurant  :  61
Max.   :686750                  Other       :  49
NA's   :1715                    (Other)     : 213
ThirdLargestPropertyUseTypeGFA YearsENERGYSTARCertified ENERGYSTARScore
Min.   :     0                 Min.   :2.007e+03        Min.   :  1.00
1st Qu.:  2392                 1st Qu.:2.018e+03        1st Qu.: 57.00
Median :  5178                 Median :2.018e+07        Median : 80.00
Mean   : 11757                 Mean   :1.382e+65        Mean   : 71.33
3rd Qu.: 10120                 3rd Qu.:2.018e+15        3rd Qu.: 92.00
Max.   :459748                 Max.   :2.018e+67        Max.   :100.00
NA's   :2851                   NA's   :3315             NA's   :1006
SiteEUI.kBtu.sf.   SiteEUIWN.kBtu.sf. SourceEUI.kBtu.sf. SourceEUIWN.kBtu.sf.
Min.   :     0.0   Min.   :     0.0   Min.   :     0.0   Min.   :     0.0
1st Qu.:    28.9   1st Qu.:    28.5   1st Qu.:    77.7   1st Qu.:    76.6
Median :    40.4   Median :    39.8   Median :   100.2   Median :    98.7
Mean   :   278.3   Mean   :   275.2   Mean   :   359.2   Mean   :   355.3
3rd Qu.:    63.7   3rd Qu.:    62.8   3rd Qu.:   147.1   3rd Qu.:   144.6
Max.   :757644.5   Max.   :746358.1   Max.   :757968.5   Max.   :746681.0
NA's   :33         NA's   :46         NA's   :33         NA's   :46
SiteEnergyUse.kBtu. SiteEnergyUseWN.kBtu. SteamUse.kBtu.
Min.   :0.000e+00   Min.   :0.000e+00     Min.   :    83081
1st Qu.:9.814e+05   1st Qu.:9.672e+05     1st Qu.:  1166578
Median :1.938e+06   Median :1.902e+06     Median :  2744007
Mean   :5.608e+07   Mean   :5.546e+07     Mean   : 10885517
3rd Qu.:4.451e+06   3rd Qu.:4.395e+06     3rd Qu.:  7811140
Max.   :1.725e+11   Max.   :1.699e+11     Max.   :257991360
NA's   :33          NA's   :46            NA's   :3324
Electricity.kWh.    Electricity.kBtu.   NaturalGas.therms. NaturalGas.kBtu.
Min.   :   -36727   Min.   :  -125314   Min.   :       0   Min.   :        0
1st Qu.:   194413   1st Qu.:   663336   1st Qu.:    4268   1st Qu.:   426845
Median :   353149   Median :  1204945   Median :   10107   Median :  1010714
```

```
Mean   :  1106094   Mean   :  3773992   Mean   :  24502    Mean   :  2450212
3rd Qu.:   857211   3rd Qu.:  2924805   3rd Qu.:  22910    3rd Qu.:  2290958
Max.   :196026272   Max.   :668841640   Max.   :4169035    Max.   :416903500
NA's   :27          NA's   :27          NA's   :1271       NA's   :1271
TotalGHGEmissions  GHGEmissionsIntensity DefaultData      ComplianceStatus
Min.   :   -0.52   Min.   :-0.010        N:3234      Compliant    :3191
1st Qu.:    6.17   1st Qu.: 0.129        Y: 227      Not Compliant: 270
Median :   32.20   Median : 0.572
Mean   :  120.74   Mean   : 1.169
3rd Qu.:   96.24   3rd Qu.: 1.384
Max.   :22813.61   Max.   :50.139

        Outlier
            :3429
High outlier:   9
Low outlier :  23
```

From str() and summary() we see this data is not very clean. There are significant NA values, mispelling of text strings (City), incorrect entries (State), and some unnecessary columns.

We will load dplyr to help us with our processing

```
In [17]: library(dplyr)
```

```
In [18]: names(data)
```

1. 'OSEBuildingID' 2. 'DataYear' 3. 'BuildingType' 4. 'PrimaryPropertyType' 5. 'PropertyName' 6. 'Address' 7. 'City' 8. 'State' 9. 'ZipCode' 10. 'TaxParcelIdentificationNumber' 11. 'CouncilDistrictCode' 12. 'Neighborhood' 13. 'Latitude' 14. 'Longitude' 15. 'YearBuilt' 16. 'NumberofBuildings' 17. 'NumberofFloors' 18. 'PropertyGFATotal' 19. 'PropertyGFAParking' 20. 'PropertyGFABuilding.s.' 21. 'ListOfAllPropertyUseTypes' 22. 'LargestPropertyUseType' 23. 'LargestPropertyUseTypeGFA' 24. 'SecondLargestPropertyUseType' 25. 'SecondLargestPropertyUseTypeGFA' 26. 'ThirdLargestPropertyUseType' 27. 'ThirdLargestPropertyUseTypeGFA' 28. 'YearsENERGYSTARCertified' 29. 'ENERGYSTARScore' 30. 'SiteEUI.kBtu.sf.' 31. 'SiteEUIWN.kBtu.sf.' 32. 'SourceEUI.kBtu.sf.' 33. 'SourceEUIWN.kBtu.sf.' 34. 'SiteEnergyUse.kBtu.' 35. 'SiteEnergyUseWN.kBtu.' 36. 'SteamUse.kBtu.' 37. 'Electricity.kWh.' 38. 'Electricity.kBtu.' 39. 'NaturalGas.therms.' 40. 'NaturalGas.kBtu.' 41. 'TotalGHGEmissions' 42. 'GHGEmissionsIntensity' 43. 'DefaultData' 44. 'ComplianceStatus' 45. 'Outlier'

We will begin by subsetting the data into columns that should reduce redundancy. Columns like City and State will be removed due to all of the building being located in Seattle, WA.

```
In [19]: small <- c("OSEBuildingID", "BuildingType", "PrimaryPropertyType", "PropertyName", "Z
         data1 <- data[small]
```

When viewing the Neighborhood column, we see this one has multiple issues

```
In [20]: data1 %>% group_by(Neighborhood) %>% summarise(no_rows = length(Neighborhood), perc =
```

Neighborhood	no_rows	perc
	17	0.4911875
Ballard	7	0.2022537
BALLARD	134	3.8717134
Central	28	0.8090147
CENTRAL	110	3.1782722
Delridge	4	0.1155735
DELRIDGE	80	2.3114707
DOWNTOWN	586	16.9315227
EAST	449	12.9731292
GREATER DUWAMISH	368	10.6327651
LAKE UNION	264	7.6278532
MAGNOLIA / QUEEN ANNE	429	12.3952615
North	43	1.2424155
NORTH	146	4.2184340
NORTHEAST	300	8.6680150
Northwest	11	0.3178272
NORTHWEST	214	6.1831841
SOUTHEAST	96	2.7737648
SOUTHWEST	171	4.9407686
water	4	0.1155735

We will cast the values all to lower case to solve the double value issue. We will also change it to a factor variable

```
In [21]: data1$Neighborhood <- tolower(as.character(data1$Neighborhood))
         data1$Neighborhood <- as.factor(data1$Neighborhood)

         data1 %>% group_by(Neighborhood) %>% summarise(no_rows = length(Neighborhood), percen
```

Neighborhood	no_rows	percentage
	17	0.4911875
ballard	141	4.0739671
central	138	3.9872869
delridge	84	2.4270442
downtown	586	16.9315227
east	449	12.9731292
greater duwamish	368	10.6327651
lake union	264	7.6278532
magnolia / queen anne	429	12.3952615
north	189	5.4608495
northeast	300	8.6680150
northwest	225	6.5010113
southeast	96	2.7737648
southwest	171	4.9407686
water	4	0.1155735

There are still some missing values as blanks so we will explore those.

```
In [22]: data1[which(data1$Neighborhood == ""),]
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName
488	649	NonResidential	Small- and Mid-Sized Office	INScape
628	839	NonResidential	Hotel	Silver Cloud Inn - Broadway
919	20210	Multifamily MR (5-9)	Mid-Rise Multifamily	Brix Condominium
1262	21364	NonResidential	Other	Pacific Northwest Research
1383	21656	NonResidential	Worship Facility	Japanese Baptist Church
1387	21662	Multifamily LR (1-4)	Low-Rise Multifamily	Cal Anderson House
2108	24430	Multifamily MR (5-9)	Mid-Rise Multifamily	Hollywood Lofts
2358	25325	NonResidential	Worship Facility	All Pilgrims Christian Chur
2503	25798	Multifamily MR (5-9)	Mid-Rise Multifamily	Capitol Building
2833	27059	Multifamily LR (1-4)	Low-Rise Multifamily	700 Broadway
2927	29069	Multifamily LR (1-4)	Senior Care Community	Lakeshore
3017	27825	NonResidential	Other	AKER'S VOLKS-PORSCHE
3109	29170	NonResidential	Mixed Use Property	Chief Seattle Club/Montere
3166	42067	Multifamily MR (5-9)	Mixed Use Property	Broadway Building
3195	49710	Multifamily MR (5-9)	Mid-Rise Multifamily	Lyric
3197	49714	Multifamily MR (5-9)	Mid-Rise Multifamily	Barclay Broadway
3276	49911	NonResidential	Other	Educare Early Learning Cen

These 17 values are also missing Lat and Long, but have a zipcode so we can generalize them. We'll store the zipcodes into a vector for our imputation.

```
In [28]: miss_zip <- data1[which(is.na(data1$Latitude)),]$ZipCode
         miss_zip
```

1. 98134 2. 98122 3. 98102 4. 98122 5. 98122 6. 98122 7. 98102 8. 98102 9. 98102 10. 98102 11. 98178 12. 98122 13. 98104 14. 98122 15. 98102 16. 98122 17. 98146

Instead of searching each individual location, we can use the zipcode package which gives an approximate lat/long for the zipcode.

```
In [30]: library(zipcode)
         data(zipcode)
         head(zipcode)
```

zip	city	state	latitude	longitude
00210	Portsmouth	NH	43.0059	-71.0132
00211	Portsmouth	NH	43.0059	-71.0132
00212	Portsmouth	NH	43.0059	-71.0132
00213	Portsmouth	NH	43.0059	-71.0132
00214	Portsmouth	NH	43.0059	-71.0132
00215	Portsmouth	NH	43.0059	-71.0132

We will rename the above data frame so that it matches with our Seattle City data

```
In [32]: colnames(zipcode)[1] <- "ZipCode"
         colnames(zipcode)[4] <- "Latitude"
         colnames(zipcode)[5] <- "Longitude"
         zipcode$ZipCode <- as.numeric(as.character(zipcode$ZipCode))
```

```
In [33]: data1[which(is.na(data1$Latitude)),]

         zipcode[which(zipcode$ZipCode %in% miss_zip),]
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName
488	649	NonResidential	Small- and Mid-Sized Office	INScape
628	839	NonResidential	Hotel	Silver Cloud Inn - Broadwa
919	20210	Multifamily MR (5-9)	Mid-Rise Multifamily	Brix Condominium
1262	21364	NonResidential	Other	Pacific Northwest Research
1383	21656	NonResidential	Worship Facility	Japanese Baptist Church
1387	21662	Multifamily LR (1-4)	Low-Rise Multifamily	Cal Anderson House
2108	24430	Multifamily MR (5-9)	Mid-Rise Multifamily	Hollywood Lofts
2358	25325	NonResidential	Worship Facility	All Pilgrims Christian Chur
2503	25798	Multifamily MR (5-9)	Mid-Rise Multifamily	Capitol Building
2833	27059	Multifamily LR (1-4)	Low-Rise Multifamily	700 Broadway
2927	29069	Multifamily LR (1-4)	Senior Care Community	Lakeshore
3017	27825	NonResidential	Other	AKER’S VOLKS-PORSCHE
3109	29170	NonResidential	Mixed Use Property	Chief Seattle Club/Montere
3166	42067	Multifamily MR (5-9)	Mixed Use Property	Broadway Building
3195	49710	Multifamily MR (5-9)	Mid-Rise Multifamily	Lyric
3197	49714	Multifamily MR (5-9)	Mid-Rise Multifamily	Barclay Broadway
3276	49911	NonResidential	Other	Educare Early Learning Cen

	ZipCode	city	state	Latitude	Longitude
43376	98102	Seattle	WA	47.63287	-122.3225
43378	98104	Seattle	WA	47.60252	-122.3286
43395	98122	Seattle	WA	47.61157	-122.3041
43405	98134	Seattle	WA	47.57867	-122.3344
43413	98146	Seattle	WA	47.50175	-122.3569
43431	98178	Seattle	WA	47.49797	-122.2466

We now have the latitude and longitude associated with the zipcode values in our Seattle data. Now we need to use a combination of joins and loops to get a fixed zipcode set.

```
In [34]: data_subset <- left_join(data1[which(is.na(data1$Latitude)),], zipcode[which(zipcode$Z
         data1_subset <- data_subset[,c(-7,-8,-18,-19)]
         head(data1_subset)
```

OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName	Zip
649	NonResidential	Small- and Mid-Sized Office	INScape	9813
839	NonResidential	Hotel	Silver Cloud Inn - Broadway	9812
20210	Multifamily MR (5-9)	Mid-Rise Multifamily	Brix Condominium	9810
21364	NonResidential	Other	Pacific Northwest Research	9812
21656	NonResidential	Worship Facility	Japanese Baptist Church	9812
21662	Multifamily LR (1-4)	Low-Rise Multifamily	Cal Anderson House	9812

1. ’OSEBuildingID’ 2. ’BuildingType’ 3. ’PrimaryPropertyType’ 4. ’PropertyName’ 5. ’ZipCode’ 6. ’Neighborhood’ 7. ’YearBuilt’ 8. ’NumberofFloors’ 9. ’ENERGYSTARScore’ 10. ’SiteEnergyUse.kBtu.’ 11. ’SteamUse.kBtu.’ 12. ’Electricity.kWh.’ 13. ’NaturalGas.therms.’ 14. ’TotalGHGEmissions’ 15. ’ComplianceStatus’ 16. ’Latitude’ 17. ’Longitude’

This join has placed the lat/long columns on the end. So we will rename them and join them back into the larger set.

```
In [35]: colnames(data1_subset)[16] <- "Latitude"
         colnames(data1_subset)[17] <- "Longitude"
         names(data1_subset)
```

1. 'OSEBuildingID' 2. 'BuildingType' 3. 'PrimaryPropertyType' 4. 'PropertyName' 5. 'ZipCode' 6. 'Neighborhood' 7. 'YearBuilt' 8. 'NumberofFloors' 9. 'ENERGYSTARScore' 10. 'SiteEnergyUse.kBtu.' 11. 'SteamUse.kBtu.' 12. 'Electricity.kWh.' 13. 'NaturalGas.therms.' 14. 'TotalGHGEmissions' 15. 'ComplianceStatus' 16. 'Latitude' 17. 'Longitude'

```
In [36]: data1_subset %>% select(OSEBuildingID, BuildingType, PrimaryPropertyType, PropertyNam


         ##Loops to add lat/long values to the larger data set
         for(i in data2_subset[,1]) {
           data1[which(data1$OSEBuildingID == i),7] <- data2_subset[which(data2_subset$OSEBuil
         }

         for(i in data2_subset[,1]) {
           data1[which(data1$OSEBuildingID == i),8] <- data2_subset[which(data2_subset$OSEBuil
         }

         summary(data1)
```

```
 OSEBuildingID                 BuildingType
 Min.   :    1   NonResidential      :1480
 1st Qu.:20033   Multifamily LR (1-4):1042
 Median :23212   Multifamily MR (5-9): 620
 Mean   :21719   Multifamily HR (10+): 112
 3rd Qu.:26147   SPS-District K-12   :  99
 Max.   :50289   Nonresidential COS  :  67
                 (Other)             :  41
                  PrimaryPropertyType           PropertyName     ZipCode
 Low-Rise Multifamily       :1011                     :  21   Min.   :98006
 Mid-Rise Multifamily       : 607     Northgate Plaza :   3   1st Qu.:98105
 Small- and Mid-Sized Office: 297     Airport Way     :   2   Median :98115
 Other                      : 247     Bayview Building:   2   Mean   :98117
 Large Office               : 180     Canal Building  :   2   3rd Qu.:98122
 Warehouse                  : 175     Central Park    :   2   Max.   :98272
 (Other)                    : 944     (Other)         :3429   NA's   :21
                Neighborhood     Latitude       Longitude        YearBuilt
 downtown             : 586   Min.   :47.50   Min.   :-122.4   Min.   :1900
 east                 : 449   1st Qu.:47.60   1st Qu.:-122.4   1st Qu.:1949
 magnolia / queen anne: 429   Median :47.62   Median :-122.3   Median :1976
 greater duwamish     : 368   Mean   :47.62   Mean   :-122.3   Mean   :1970
 northeast            : 300   3rd Qu.:47.66   3rd Qu.:-122.3   3rd Qu.:1998
 lake union           : 264   Max.   :47.73   Max.   :-122.2   Max.   :2017
 (Other)              :1065                                    NA's   :1
 NumberofFloors   ENERGYSTARScore  SiteEnergyUse.kBtu. SteamUse.kBtu.
```

```
Min.   : 0.000   Min.   :  1.00   Min.   :0.000e+00   Min.   :    83081
1st Qu.: 2.000   1st Qu.: 57.00   1st Qu.:9.814e+05   1st Qu.:  1166578
Median : 4.000   Median : 80.00   Median :1.938e+06   Median :  2744007
Mean   : 4.759   Mean   : 71.33   Mean   :5.608e+07   Mean   : 10885517
3rd Qu.: 5.000   3rd Qu.: 92.00   3rd Qu.:4.451e+06   3rd Qu.:  7811140
Max.   :76.000   Max.   :100.00   Max.   :1.725e+11   Max.   :257991360
NA's   :7        NA's   :1006     NA's   :33          NA's   :3324
Electricity.kWh.    NaturalGas.therms. TotalGHGEmissions
Min.   :   -36727   Min.   :      0    Min.   :   -0.52
1st Qu.:   194413   1st Qu.:   4268    1st Qu.:    6.17
Median :   353149   Median :  10107    Median :   32.20
Mean   :  1106094   Mean   :  24502    Mean   :  120.74
3rd Qu.:   857211   3rd Qu.:  22910    3rd Qu.:   96.24
Max.   :196026272   Max.   :4169035    Max.   :22813.61
NA's   :27          NA's   :1271
     ComplianceStatus
Compliant    :3191
Not Compliant: 270
```

Now all the lat/long values are fixed.

1.2 Year Built Missing

```
In [38]: data1[which(is.na(data1$YearBuilt)),] #Clark hall at UW. Website says 1896 is the yea

         data1[which(is.na(data1$YearBuilt)),9] <- 1896

         summary(data1)
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName	ZipCode	Neighborh
3301	49971	NonResidential	University	Clark Hall	98195	northeast

```
OSEBuildingID                 BuildingType
Min.   :    1   NonResidential      :1480
1st Qu.:20033   Multifamily LR (1-4):1042
Median :23212   Multifamily MR (5-9): 620
Mean   :21719   Multifamily HR (10+): 112
3rd Qu.:26147   SPS-District K-12   :  99
Max.   :50289   Nonresidential COS  :  67
                (Other)             :  41
                 PrimaryPropertyType            PropertyName      ZipCode
Low-Rise Multifamily       :1011                        :  21   Min.   :98006
Mid-Rise Multifamily       : 607     Northgate Plaza :   3   1st Qu.:98105
```

```
Small- and Mid-Sized Office: 297     Airport Way     :   2   Median :98115
Other                      : 247     Bayview Building:   2   Mean   :98117
Large Office               : 180     Canal Building  :   2   3rd Qu.:98122
Warehouse                  : 175     Central Park    :   2   Max.   :98272
(Other)                    : 944     (Other)         :3429   NA's   :21
               Neighborhood     Latitude       Longitude        YearBuilt
downtown             : 586   Min.   :47.50   Min.   :-122.4   Min.   :1896
east                 : 449   1st Qu.:47.60   1st Qu.:-122.4   1st Qu.:1949
magnolia / queen anne: 429   Median :47.62   Median :-122.3   Median :1976
greater duwamish     : 368   Mean   :47.62   Mean   :-122.3   Mean   :1969
northeast            : 300   3rd Qu.:47.66   3rd Qu.:-122.3   3rd Qu.:1998
lake union           : 264   Max.   :47.73   Max.   :-122.2   Max.   :2017
(Other)              :1065
NumberofFloors   ENERGYSTARScore  SiteEnergyUse.kBtu. SteamUse.kBtu.
Min.   : 0.000   Min.   :  1.00   Min.   :0.000e+00   Min.   :    83081
1st Qu.: 2.000   1st Qu.: 57.00   1st Qu.:9.814e+05   1st Qu.:  1166578
Median : 4.000   Median : 80.00   Median :1.938e+06   Median :  2744007
Mean   : 4.759   Mean   : 71.33   Mean   :5.608e+07   Mean   : 10885517
3rd Qu.: 5.000   3rd Qu.: 92.00   3rd Qu.:4.451e+06   3rd Qu.:  7811140
Max.   :76.000   Max.   :100.00   Max.   :1.725e+11   Max.   :257991360
NA's   :7        NA's   :1006     NA's   :33          NA's   :3324
Electricity.kWh.    NaturalGas.therms. TotalGHGEmissions
Min.   :   -36727   Min.   :      0    Min.   :   -0.52
1st Qu.:   194413   1st Qu.:   4268    1st Qu.:    6.17
Median :   353149   Median :  10107    Median :   32.20
Mean   :  1106094   Mean   :  24502    Mean   :  120.74
3rd Qu.:   857211   3rd Qu.:  22910    3rd Qu.:   96.24
Max.   :196026272   Max.   :4169035    Max.   :22813.61
NA's   :27          NA's   :1271
     ComplianceStatus
Compliant    :3191
Not Compliant: 270
```

1.3 Missing Electricity, Natural Gas, Steam data

```
In [39]: head(data1 %>% arrange(desc(SteamUse.kBtu.)))
```

OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName	Z
49967	Campus	University	University of Washington - Seattle Campus	9
828	Campus	Hospital	Swedish First Hill	9
276	Campus	Hospital	Harborview Medical Center	9
49975	NonResidential	University	Health Sciences K-Wing	9
49973	NonResidential	University	Foege Bldg	9
49982	NonResidential	University	Physics Astronomy Bldg	9

It is generally unclear why there are missing values for NaturalGas, Electricity, and Steam for various rows in the data set. Using knowledge of Seattle buildings, we can make the assumption that these NA values should be zero. UW for instance, relies heavily on Steam power instead of Natural Gas, so the NA value in Natural Gas should be zero.

```
In [42]: data1[which(is.na(data1$NaturalGas.therms.)),]$NaturalGas.therms. <- 0
         summary(data1$NaturalGas.therms.)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0       0    3642   15504   13520 4169035 
```

Now we will do the same for Steam.

```
In [43]: data1[which(is.na(data1$SteamUse.kBtu.)),]$SteamUse.kBtu. <- 0
         summary(data1$SteamUse.kBtu.)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
        0         0         0    430892         0 257991360 
```

The missing values for Electricity also contain missing values for the other energy types so we will remove these rows from our set.

```
In [44]: data1[which(is.na(data1$Electricity.kWh.)),]
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName
166	266	Multifamily LR (1-4)	Low-Rise Multifamily	
177	283	NonResidential	Small- and Mid-Sized Office	
193	302	NonResidential	Other	Seattle ReadCntr (50/50)
287	413	NonResidential	Large Office	
582	773	NonResidential	Small- and Mid-Sized Office	Seattle Building / Old Char
753	19776	NonResidential	Other	
811	19892	NonResidential		
840	20198	Multifamily LR (1-4)	Low-Rise Multifamily	
852	19990	Multifamily MR (5-9)	Mid-Rise Multifamily	
878	20367	Multifamily LR (1-4)	Low-Rise Multifamily	City Lights on Harbor
1388	25995	Multifamily MR (5-9)	Mid-Rise Multifamily	
1404	21689	NonResidential	Small- and Mid-Sized Office	Dept of Social & Health Serv
1524	22139	NonResidential	Retail Store	
1993	24030	NonResidential	Retail Store	
2003	24068	Multifamily LR (1-4)	Low-Rise Multifamily	
2036	24162	Multifamily LR (1-4)	Low-Rise Multifamily	
2116	25752	Multifamily LR (1-4)	Low-Rise Multifamily	
2428	25553	NonResidential	Hotel	J & M HOTEL BUILDING (
2445	25617	Multifamily MR (5-9)	Mid-Rise Multifamily	
2456	25655	Multifamily MR (5-9)	Mid-Rise Multifamily	The Seattle Quilt Building
2618	26218	NonResidential	Hotel	
2696	26583	Multifamily MR (5-9)	Mid-Rise Multifamily	
3014	50195	NonResidential		
3184	49693	Multifamily MR (5-9)	Mid-Rise Multifamily	
3408	50150	NonResidential		
3409	50152	Multifamily LR (1-4)		
3457	50265	Multifamily LR (1-4)		

```
In [45]: data2 <- data1[-which(is.na(data1$Electricity.kWh.)),]
```

```
In [46]: summary(data2)
```

```
OSEBuildingID                 BuildingType
Min.   :    1   NonResidential      :1467
1st Qu.:20039   Multifamily LR (1-4):1034
Median :23200   Multifamily MR (5-9): 614
Mean   :21700   Multifamily HR (10+): 112
3rd Qu.:26147   SPS-District K-12   :  99
Max.   :50289   Nonresidential COS  :  67
                (Other)             :  41
                 PrimaryPropertyType              PropertyName
Low-Rise Multifamily       :1005     Northgate Plaza     :   3
Mid-Rise Multifamily       : 601     Airport Way         :   2
Small- and Mid-Sized Office: 294     Bayview Building    :   2
Other                      : 245     Canal Building      :   2
Large Office               : 179     Central Park        :   2
Warehouse                  : 175     Crestview Apartments:   2
(Other)                    : 935     (Other)             :3421
```

```
    ZipCode                     Neighborhood     Latitude       Longitude     
 Min.   :98006   downtown                 : 581   Min.   :47.50   Min.   :-122.4  
 1st Qu.:98105   east                     : 448   1st Qu.:47.60   1st Qu.:-122.4  
 Median :98115   magnolia / queen anne: 426   Median :47.62   Median :-122.3  
 Mean   :98117   greater duwamish     : 365   Mean   :47.62   Mean   :-122.3  
 3rd Qu.:98122   northeast            : 299   3rd Qu.:47.66   3rd Qu.:-122.3  
 Max.   :98272   lake union           : 263   Max.   :47.73   Max.   :-122.2  
                 (Other)              :1052                                   
   YearBuilt    NumberofFloors    ENERGYSTARScore  SiteEnergyUse.kBtu.
 Min.   :1896   Min.   : 0.000   Min.   :  1.00   Min.   :0.000e+00  
 1st Qu.:1949   1st Qu.: 2.000   1st Qu.: 57.00   1st Qu.:9.814e+05  
 Median :1976   Median : 4.000   Median : 80.00   Median :1.938e+06  
 Mean   :1969   Mean   : 4.764   Mean   : 71.33   Mean   :5.608e+07  
 3rd Qu.:1998   3rd Qu.: 5.000   3rd Qu.: 92.00   3rd Qu.:4.451e+06  
 Max.   :2017   Max.   :76.000   Max.   :100.00   Max.   :1.725e+11  
                NA's   :7        NA's   :979      NA's   :6          
 SteamUse.kBtu.      Electricity.kWh.    NaturalGas.therms. TotalGHGEmissions
 Min.   :        0   Min.   :   -36727   Min.   :      0    Min.   :   -0.520  
 1st Qu.:        0   1st Qu.:   194413   1st Qu.:      0    1st Qu.:    6.455  
 Median :        0   Median :   353149   Median :   3765    Median :   32.820  
 Mean   :   434280   Mean   :  1106094   Mean   :  15626    Mean   :  121.690  
 3rd Qu.:        0   3rd Qu.:   857211   3rd Qu.:  13618    3rd Qu.:   97.338  
 Max.   :257991360   Max.   :196026272   Max.   :4169035    Max.   :22813.610  

     ComplianceStatus
 Compliant    :3170  
 Not Compliant: 264  
```

We still have Number of Floors, ENERGYSTARScore, and SiteEnergyUse.kBtu. for missing values.

1.4 Site Energy Use

```
In [48]: data2[which(is.na(data2$SiteEnergyUse.kBtu.)),]
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName
154	244	NonResidential	Small- and Mid-Sized Office	Washington Park Building
743	19753	Multifamily LR (1-4)	Low-Rise Multifamily	New Pacific Apartments
767	19801	NonResidential	Other	APEX BELLTOWN COOP
1181	21175	Multifamily LR (1-4)	Mixed Use Property	41 Dravus St
1323	21507	Multifamily LR (1-4)	Low-Rise Multifamily	Lewiston Apartments
2368	25354	Multifamily HR (10+)	High-Rise Multifamily	One Pacific Towers

Interestingly, these buildings are all listed as Not Compliant, which could be useful for predictions later on, so we will consider SiteEnergyUse.kBtu. as the sum of all listed energy values for that row.

```
In [49]: data2[which(is.na(data2$SiteEnergyUse.kBtu.)),] %>% mutate(SiteEnergyUse.kBtu. = Stea
```

Add these values back into the larger data set.

```
In [50]: for(i in data3_subset[,1]) {
           data2[which(data2$OSEBuildingID == i),12] <- data3_subset[which(data3_subset$OSEBui
         }
```

```
In [51]: summary(data2)
```

```
 OSEBuildingID                 BuildingType
 Min.   :    1   NonResidential      :1467
 1st Qu.:20039   Multifamily LR (1-4):1034
 Median :23200   Multifamily MR (5-9): 614
 Mean   :21700   Multifamily HR (10+): 112
 3rd Qu.:26147   SPS-District K-12   :  99
 Max.   :50289   Nonresidential COS  :  67
                 (Other)             :  41
                  PrimaryPropertyType               PropertyName
 Low-Rise Multifamily       :1005     Northgate Plaza     :   3
 Mid-Rise Multifamily       : 601     Airport Way         :   2
 Small- and Mid-Sized Office: 294     Bayview Building    :   2
 Other                      : 245     Canal Building      :   2
 Large Office               : 179     Central Park        :   2
 Warehouse                  : 175     Crestview Apartments:   2
 (Other)                    : 935     (Other)             :3421
    ZipCode                      Neighborhood     Latitude       Longitude
 Min.   :98006   downtown              : 581   Min.   :47.50   Min.   :-122.4
 1st Qu.:98105   east                  : 448   1st Qu.:47.60   1st Qu.:-122.4
 Median :98115   magnolia / queen anne: 426   Median :47.62   Median :-122.3
 Mean   :98117   greater duwamish      : 365   Mean   :47.62   Mean   :-122.3
 3rd Qu.:98122   northeast             : 299   3rd Qu.:47.66   3rd Qu.:-122.3
 Max.   :98272   lake union            : 263   Max.   :47.73   Max.   :-122.2
                 (Other)               :1052
   YearBuilt    NumberofFloors    ENERGYSTARScore  SiteEnergyUse.kBtu.
 Min.   :1896   Min.   : 0.000   Min.   :  1.00   Min.   :0.000e+00
 1st Qu.:1949   1st Qu.: 2.000   1st Qu.: 57.00   1st Qu.:9.803e+05
 Median :1976   Median : 4.000   Median : 80.00   Median :1.934e+06
 Mean   :1969   Mean   : 4.764   Mean   : 71.33   Mean   :5.599e+07
 3rd Qu.:1998   3rd Qu.: 5.000   3rd Qu.: 92.00   3rd Qu.:4.444e+06
 Max.   :2017   Max.   :76.000   Max.   :100.00   Max.   :1.725e+11
                NA's   :7        NA's   :979
 SteamUse.kBtu.      Electricity.kWh.    NaturalGas.therms. TotalGHGEmissions
 Min.   :        0   Min.   :   -36727   Min.   :        0   Min.   :   -0.520
 1st Qu.:        0   1st Qu.:   194413   1st Qu.:        0   1st Qu.:    6.455
 Median :        0   Median :   353149   Median :   3765    Median :   32.820  
 Mean   :   434280   Mean   :  1106094   Mean   :  15626    Mean   :  121.690  
 3rd Qu.:        0   3rd Qu.:   857211   3rd Qu.:  13618    3rd Qu.:   97.338  
 Max.   :257991360   Max.   :196026272   Max.   :4169035    Max.   :22813.610  
                                                                               
      ComplianceStatus
 Compliant    :3170   
 Not Compliant: 264   
```

We will explore where the SiteEnergyUse.kBtu. is zero to see if they are useful observations.

```
In [53]: data2[which(data2$SiteEnergyUse.kBtu. == 0),]
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName	ZipCode	Neig
26	31	NonResidential	Other	Seattle Honda	98101	down
768	19805	Multifamily LR (1-4)	Low-Rise Multifamily	Ara Vita	98199	magn
982	20396	NonResidential	Warehouse	Meaves Building	98101	down

They are also Not Compliant but have zero values for all energy values so we will remove them.

```
In [101]: data3 <- data2[-which(data2$SiteEnergyUse.kBtu. == 0),]
          summary(data3)
```

```
 OSEBuildingID                   BuildingType 
 Min.   :    1   NonResidential      :1465   
 1st Qu.:20046   Multifamily LR (1-4):1033   
 Median :23212   Multifamily MR (5-9): 614   
 Mean   :21707   Multifamily HR (10+): 112   
 3rd Qu.:26148   SPS-District K-12   :  99   
 Max.   :50289   Nonresidential COS  :  67   
                 (Other)             :  41   
                 PrimaryPropertyType               PropertyName 
 Low-Rise Multifamily       :1004     Northgate Plaza     :   3  
 Mid-Rise Multifamily       : 601     Airport Way         :   2  
 Small- and Mid-Sized Office: 294     Bayview Building    :   2  
 Other                      : 244     Canal Building      :   2  
 Large Office               : 179     Central Park        :   2  
 Warehouse                  : 174     Crestview Apartments:   2  
 (Other)                    : 935     (Other)             :3418  
    ZipCode                     Neighborhood     Latitude       Longitude     
 Min.   :98006   downtown             : 579   Min.   :47.50   Min.   :-122.4  
 1st Qu.:98105   east                 : 448   1st Qu.:47.60   1st Qu.:-122.4  
 Median :98115   magnolia / queen anne: 425   Median :47.62   Median :-122.3  
 Mean   :98117   greater duwamish     : 365   Mean   :47.62   Mean   :-122.3
 3rd Qu.:98122   northeast            : 299   3rd Qu.:47.66   3rd Qu.:-122.3
 Max.   :98272   lake union           : 263   Max.   :47.73   Max.   :-122.2
                 (Other)              :1052
   YearBuilt    NumberofFloors   ENERGYSTARScore  SiteEnergyUse.kBtu.
 Min.   :1896   Min.   : 0.000   Min.   :  1.00   Min.   :1.448e+04
 1st Qu.:1949   1st Qu.: 2.000   1st Qu.: 57.00   1st Qu.:9.812e+05
 Median :1976   Median : 4.000   Median : 80.00   Median :1.938e+06
 Mean   :1969   Mean   : 4.765   Mean   : 71.31   Mean   :5.603e+07
 3rd Qu.:1998   3rd Qu.: 5.000   3rd Qu.: 92.00   3rd Qu.:4.448e+06
 Max.   :2017   Max.   :76.000   Max.   :100.00   Max.   :1.725e+11
                NA's   :7        NA's   :977
 SteamUse.kBtu.       Electricity.kWh.    NaturalGas.therms.  TotalGHGEmissions
 Min.   :        0   Min.   :   -36727   Min.   :      0    Min.   :   -0.52
 1st Qu.:        0   1st Qu.:   194716   1st Qu.:      0    1st Qu.:    6.47
 Median :        0   Median :   353291   Median :   3771    Median :   32.86
 Mean   :   434659   Mean   :  1107061   Mean   :  15640    Mean   :  121.80
 3rd Qu.:        0   3rd Qu.:   858760   3rd Qu.:  13632    3rd Qu.:   97.42
 Max.   :257991360   Max.   :196026272   Max.   :4169035    Max.   :22813.61

      ComplianceStatus
 Compliant    :3170
 Not Compliant: 261
```

1.5 Number of Floors

```
In [56]: data3[which(is.na(data3$NumberofFloors)),]
```

	OSEBuildingID	BuildingType	PrimaryPropertyType	PropertyName
3153	40031	NonResidential	Medical Office	Sandpoint #25
3154	40034	NonResidential	Small- and Mid-Sized Office	Sandpoint #29
3296	49966	NonResidential	Other	Smilow Rainier Vista Boys & Girls
3300	49970	NonResidential	Residence Hall	Cedar Hall
3309	49979	NonResidential	Residence Hall	Lander Hall
3310	49980	NonResidential	Residence Hall	Mercer Court
3313	49983	NonResidential	Residence Hall	Poplar Hall

These rows contain useful information so we will research the buildings to gather the general number of floors for each one.

The following values were obtained through image searches and knowledge of UW residence halls. ##3309, Lander floors == 6 ##3296, Boys and girls club == 3 ##3300, Cedar Hall == 6 ##3310, Mercer Court == 5 ##3313, Poplar == 7 ##3154 Sandpoint #29 == 3 ##3153 Sandpoint #25 == 3

```
In [102]: data3[which(data3$OSEBuildingID == 40031),10] <- 3
```

```
          data3[which(data3$OSEBuildingID == 40034),10] <- 3
          data3[which(data3$OSEBuildingID == 49966),10] <- 3
          data3[which(data3$OSEBuildingID == 49970),10] <- 6
          data3[which(data3$OSEBuildingID == 49979),10] <- 6
          data3[which(data3$OSEBuildingID == 49980),10] <- 5
          data3[which(data3$OSEBuildingID == 49983),10] <- 7
```

1.6 Factor and variable changes

We will change the remaining variables to their appropriate factor type.

```
In [103]: data3$NumberofFloors <- as.factor(data3$NumberofFloors)
          data3$ZipCode <- as.factor(data3$ZipCode)
          data3$YearBuilt <- as.factor(data3$YearBuilt)
```

We will also create a new variable, called BuildingAge for a continuous type option

```
In [99]: #data3 %>% dplyr::select(BuildingAge = YearBuilt, everything()) -> data3

         #data3 %>% mutate(BuildingAge = 2019 - BuildingAge) -> data3
```

ENERGYSTARScore still has a large number of missing values, so we will subset the non-missing values for visualization purposes.

```
In [116]: miss_ind <- which(is.na(data3$ENERGYSTARScore))

          data4 <- data3[-miss_ind,]
```

```
In [106]: #data4$ENERGYSTARScore <- as.factor(data4$ENERGYSTARScore)
```

1.7 Visualization

```
In [65]: library(ggmap)
         library(ggplot2)
         library(viridis)
```

```
Loading required package: ggplot2
Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
Please cite ggmap if you use it! See citation("ggmap") for details.
Loading required package: viridisLite
```

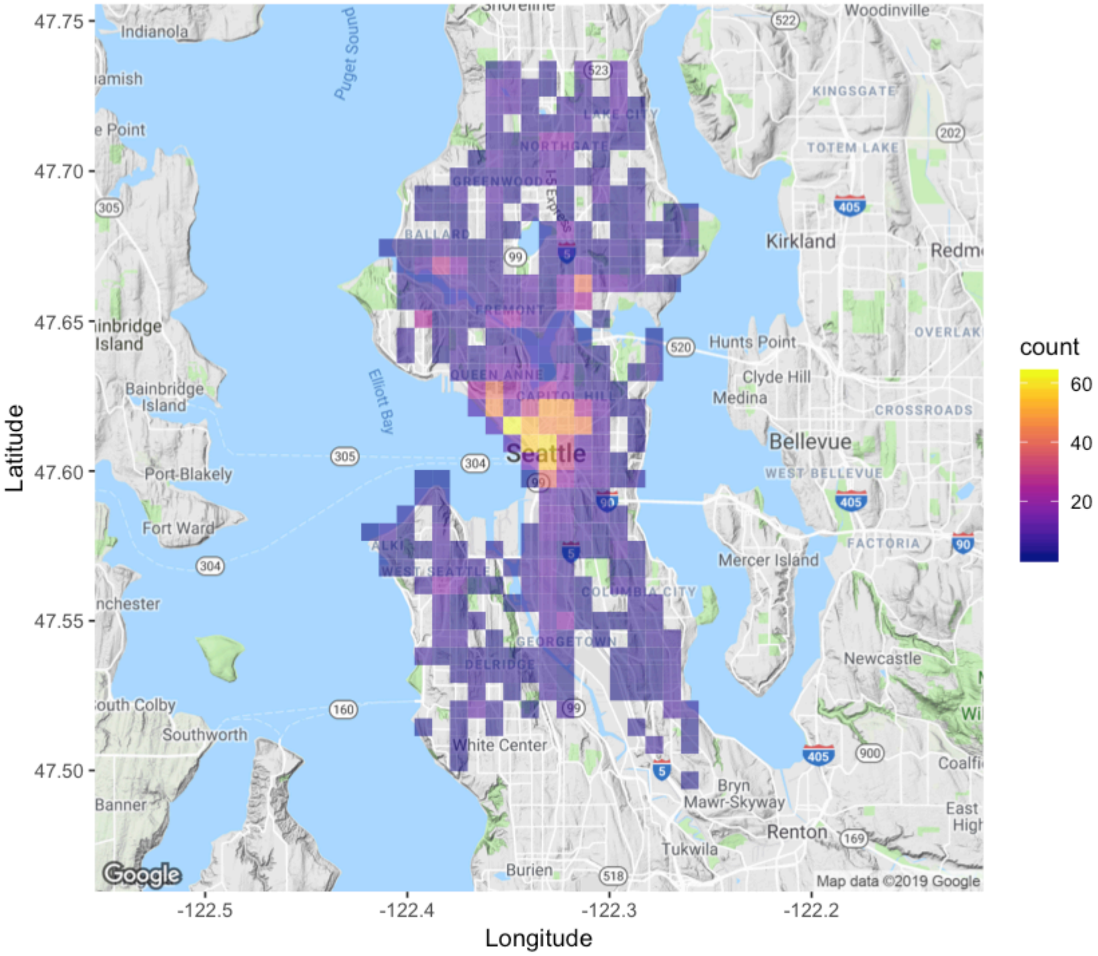
We will use ggmap, ggplot2, and viridis packages to visualize Seattle and its energy production by location

```
In [72]: ggmap::register_google(key = "AIzaSyAF8kUo2fAa-5oAoEvEBa60wgbStKyjMxs")

         p <- ggmap(get_googlemap(center = c(lon = -122.335167, lat = 47.608013),
                                  zoom = 11, scale = 2,
                                  maptype ='terrain',
                                  color = 'color'))
```

```
Source : https://maps.googleapis.com/maps/api/staticmap?center=47.608013,-122.335167&zoom=11&s
```

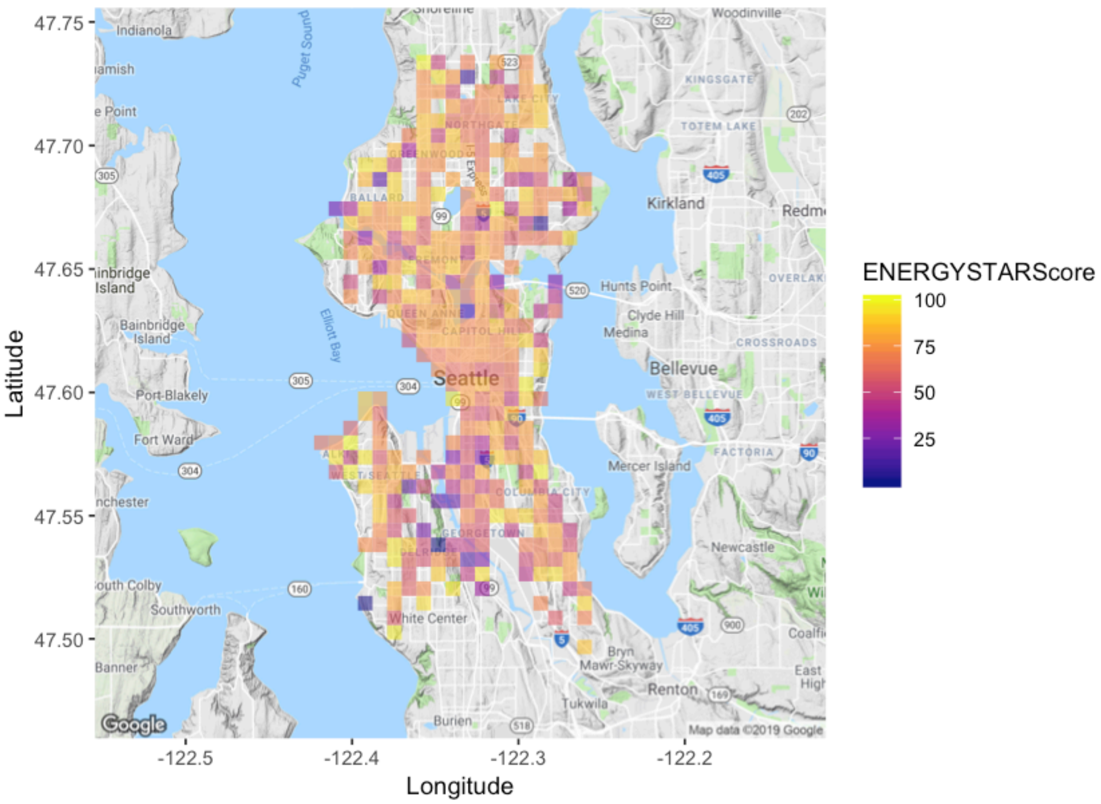
This first plot will show the number of entries by location.

```
In [124]: p + scale_fill_viridis(option = 'plasma') +
            geom_bin2d(mapping = aes(x = Longitude, y = Latitude), data = data4, bins = 50, al
            labs(x = 'Longitude', y = 'Latitude')
```



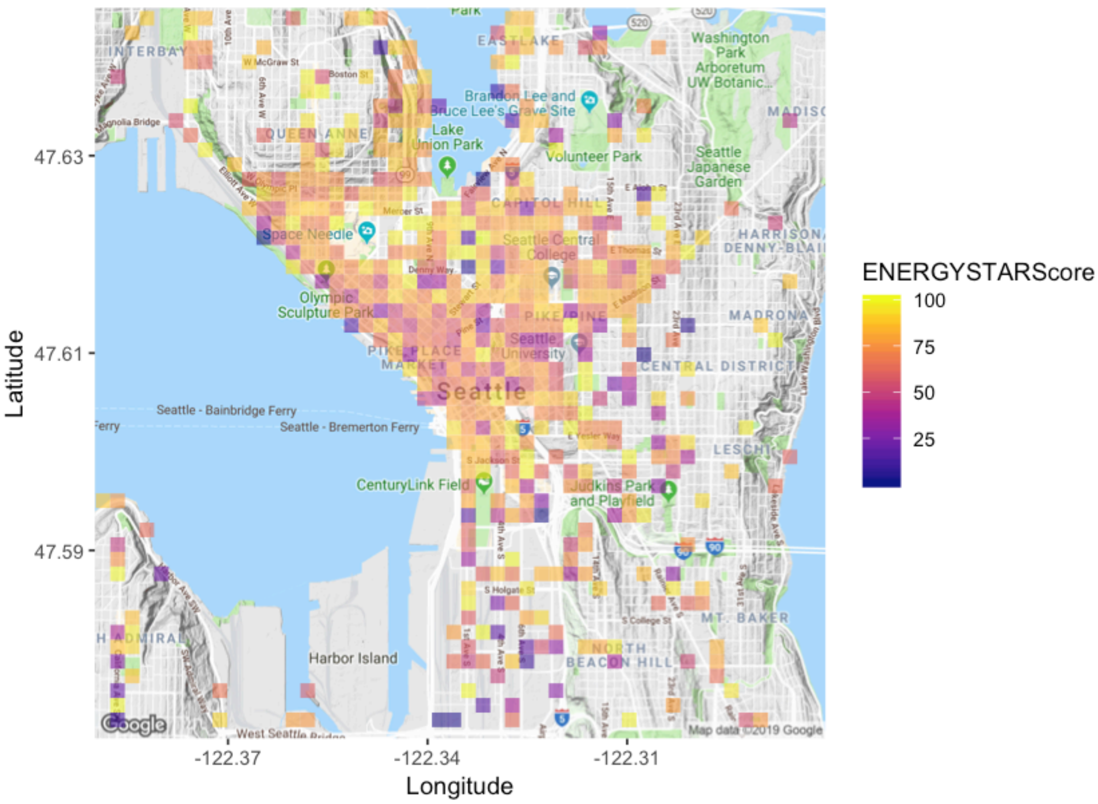
Downtown Seattle seems to have the most entries, followed by University District and Ballard
Now we will visualized the Energystar Score for each entry.

```
In [123]: p + scale_fill_viridis(option = 'plasma') +
            stat_summary_2d(mapping = aes(x = Longitude, y = Latitude, z = ENERGYSTARScore), d
            labs(x = 'Longitude', y = 'Latitude', fill = "ENERGYSTARScore")
```



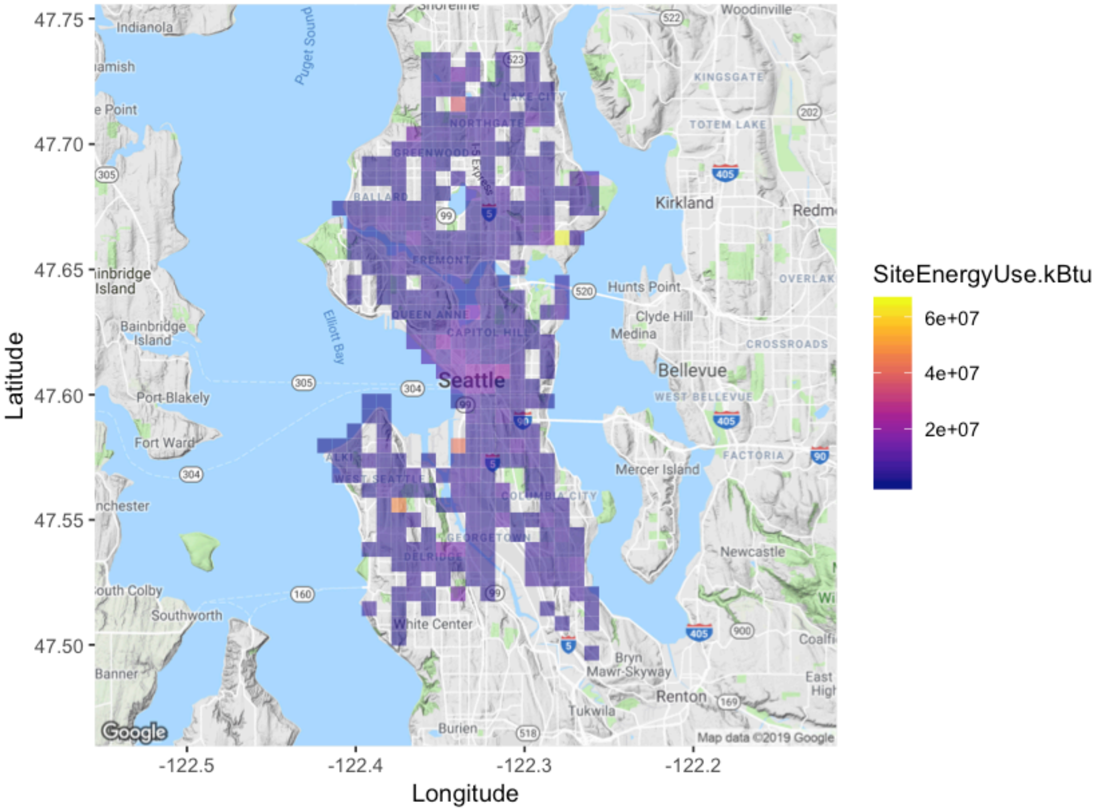
Generally, most of these entries are above 50 and are distributed throughout the city. We can zoom in on Downtown and see how the building in that location perform.

```
In [125]: p1 <- ggmap(get_googlemap(center = c(lon = -122.335167, lat = 47.608013),
                                    zoom = 13, scale = 2,
                                    maptype ='terrain',
                                    color = 'color'))

Source : https://maps.googleapis.com/maps/api/staticmap?center=47.608013,-122.335167&zoom=13&s


In [127]: p1 + scale_fill_viridis(option = 'plasma') +
              stat_summary_2d(mapping = aes(x = Longitude, y = Latitude, z = ENERGYSTARScore), da
              labs(x = 'Longitude', y = 'Latitude', fill = 'ENERGYSTARScore')
```



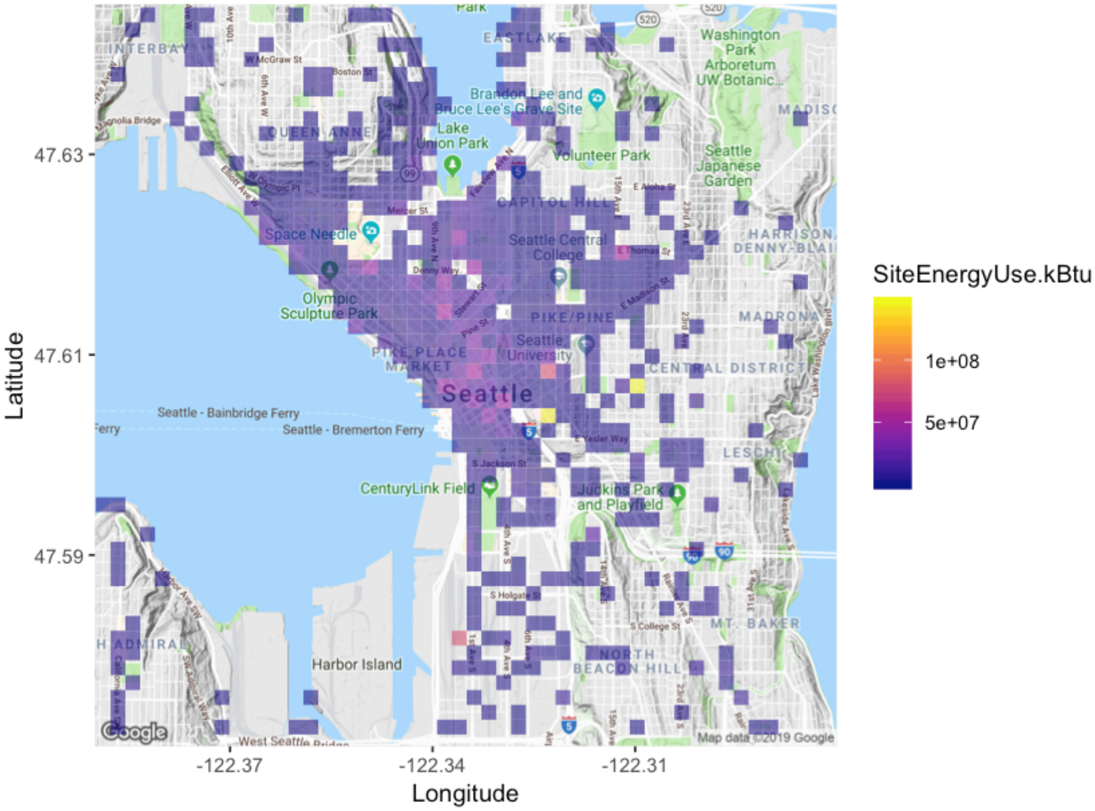
Now we will plot by SiteEnergyUse.kBtu.

```
In [130]: p + scale_fill_viridis(option = 'plasma') +
            stat_summary_2d(mapping = aes(x = Longitude, y = Latitude, z = SiteEnergyUse.kBtu.)
            labs(x = 'Longitude', y = 'Latitude', fill = "SiteEnergyUse.kBtu")
```



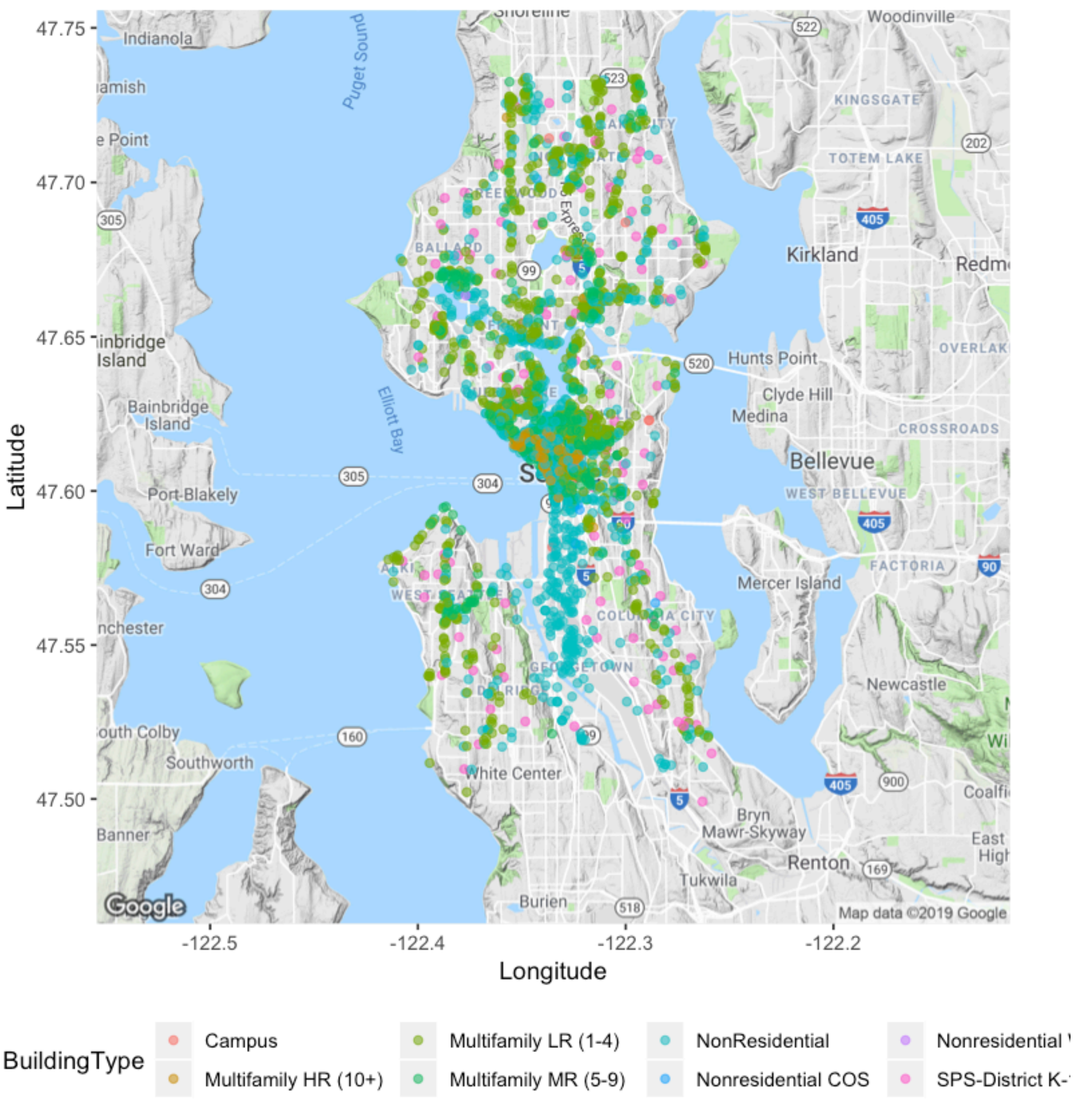
And again by Downtown

```
In [133]: p1 + scale_fill_viridis(option = 'plasma') +
            stat_summary_2d(mapping = aes(x = Longitude, y = Latitude, z = SiteEnergyUse.kBtu.)
            labs(x = 'Longitude', y = 'Latitude', fill = "SiteEnergyUse.kBtu")
```



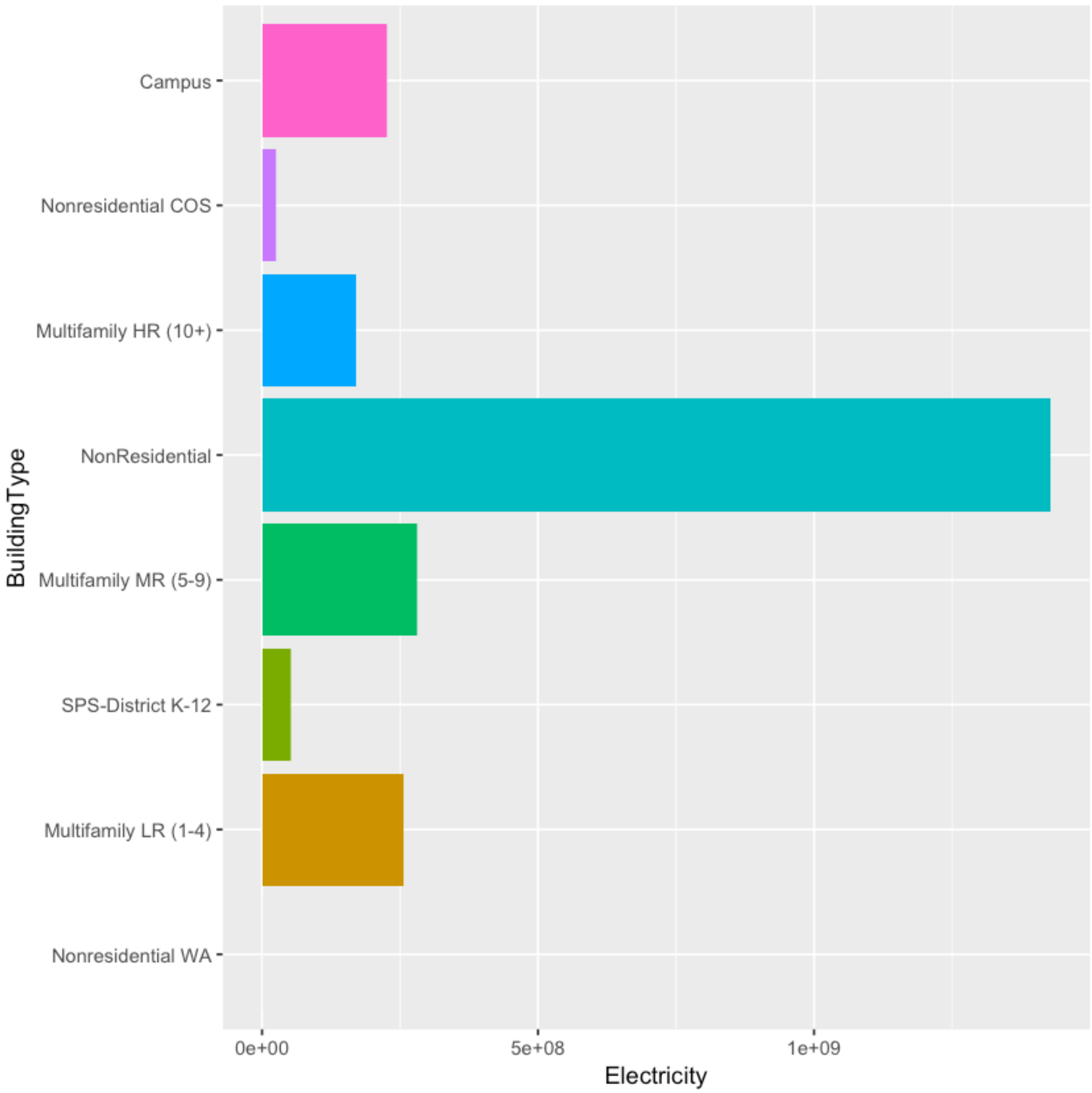
We can also draw by building type. The following plot shows that most of the multifamily highrise apartments with 10+ floors are located Downtown.

```
In [138]: p + geom_point(aes(x = Longitude, y = Latitude,  colour = BuildingType), data = data
            theme(legend.position="bottom") +
            labs(x = "Longitude", y = "Latitude")
```



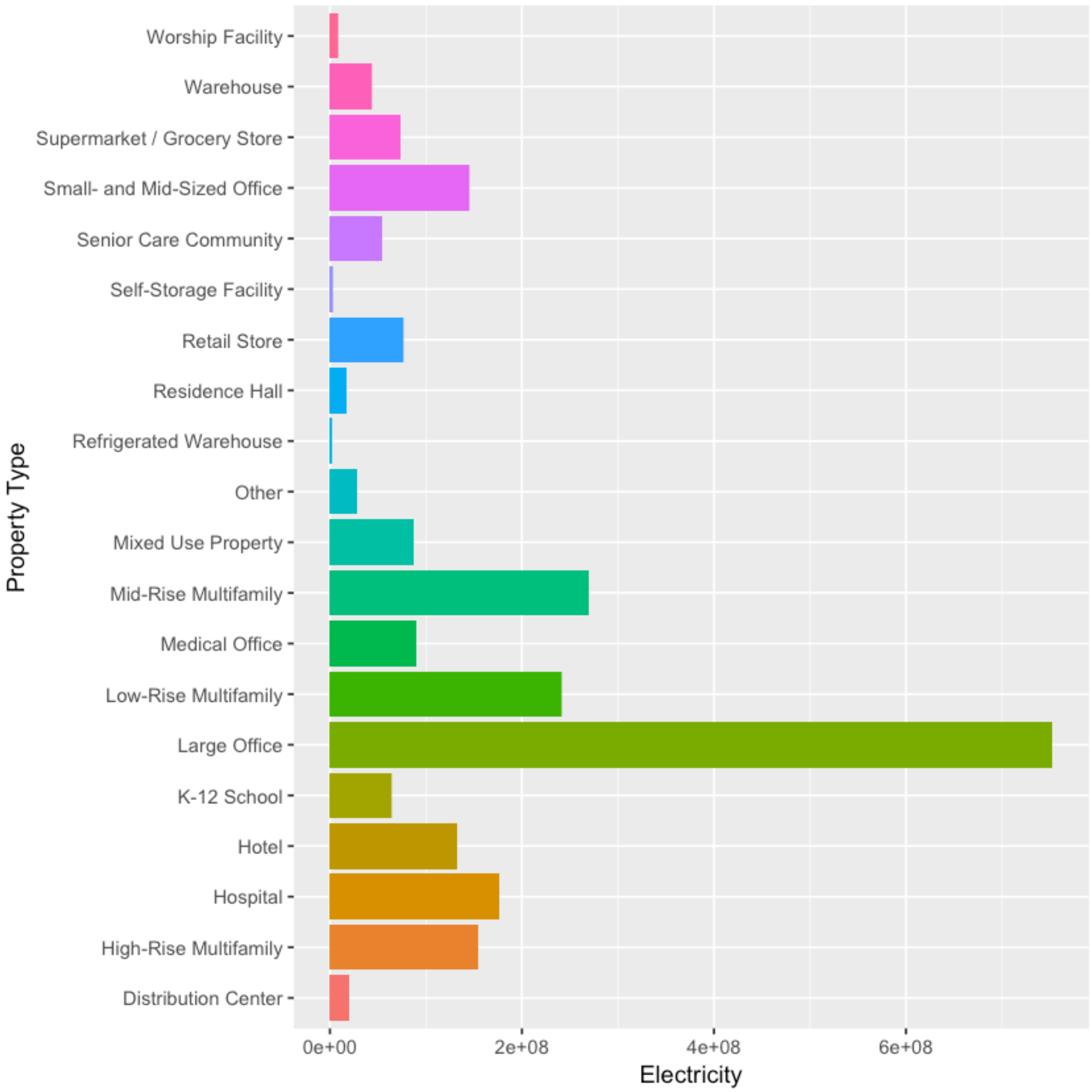
When visualizing Electricity production by Building type, we see that NonResidential produces the most

```
In [160]: data4 %>%
            mutate(BuildingType = reorder(BuildingType,Electricity.kWh.)) %>%
            ggplot(aes(x = BuildingType, y = Electricity.kWh., fill = BuildingType)) +
            geom_col() +
            coord_flip() +
            xlab("BuildingType") +
            ylab("Electricity") +
            theme(legend.position='none')
```



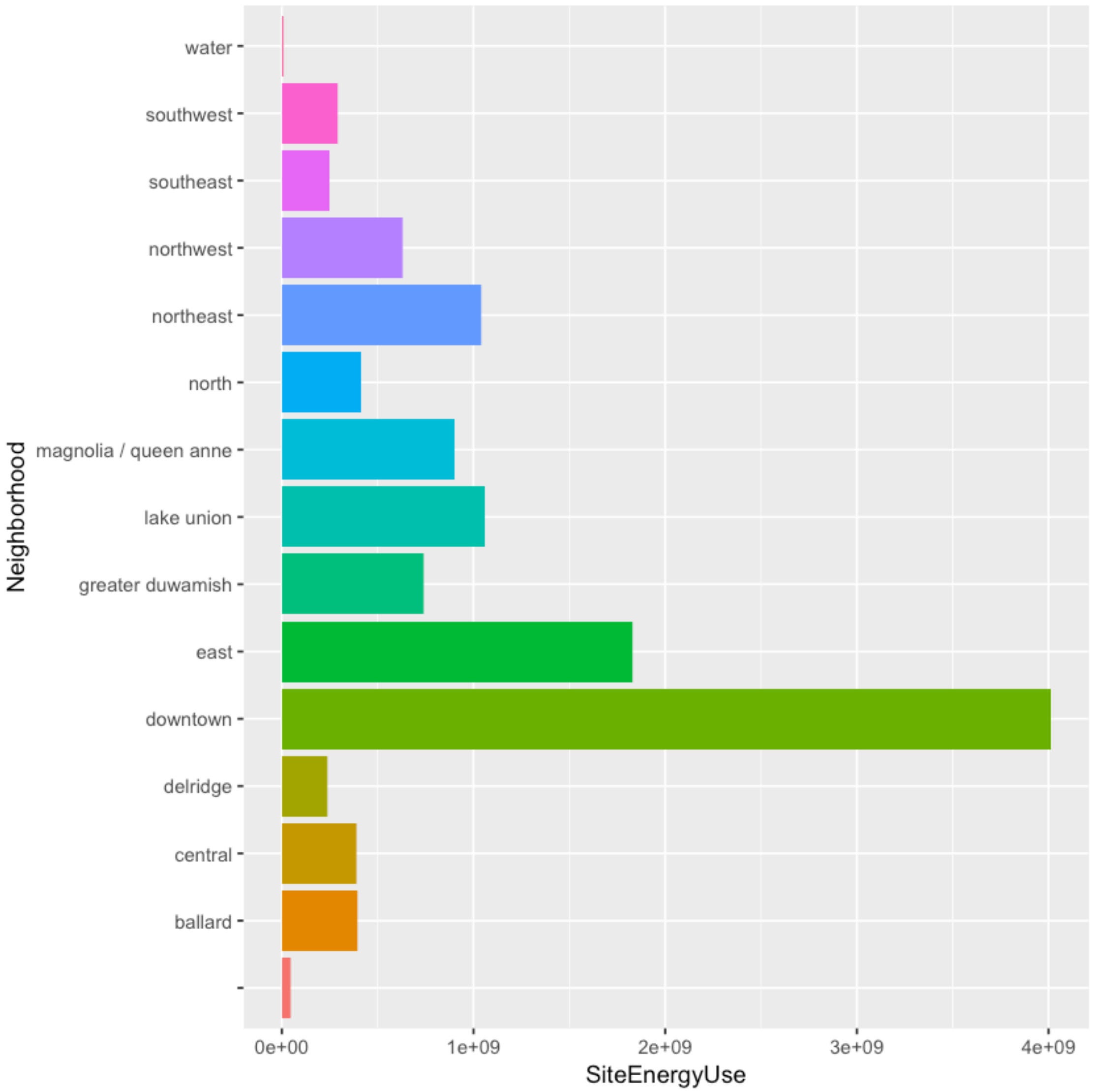
Going one level deeper into PrimaryPropertyType, we then see that Large Offices produce the most electricity of NonResidential buildings

```
In [159]: data4 %>%
            ggplot(aes(x = PrimaryPropertyType, y = Electricity.kWh., fill = PrimaryPropertyTyp
            geom_col() +
            coord_flip() +
            xlab("Property Type") +
            ylab("Electricity") +
            theme(legend.position='none')
```



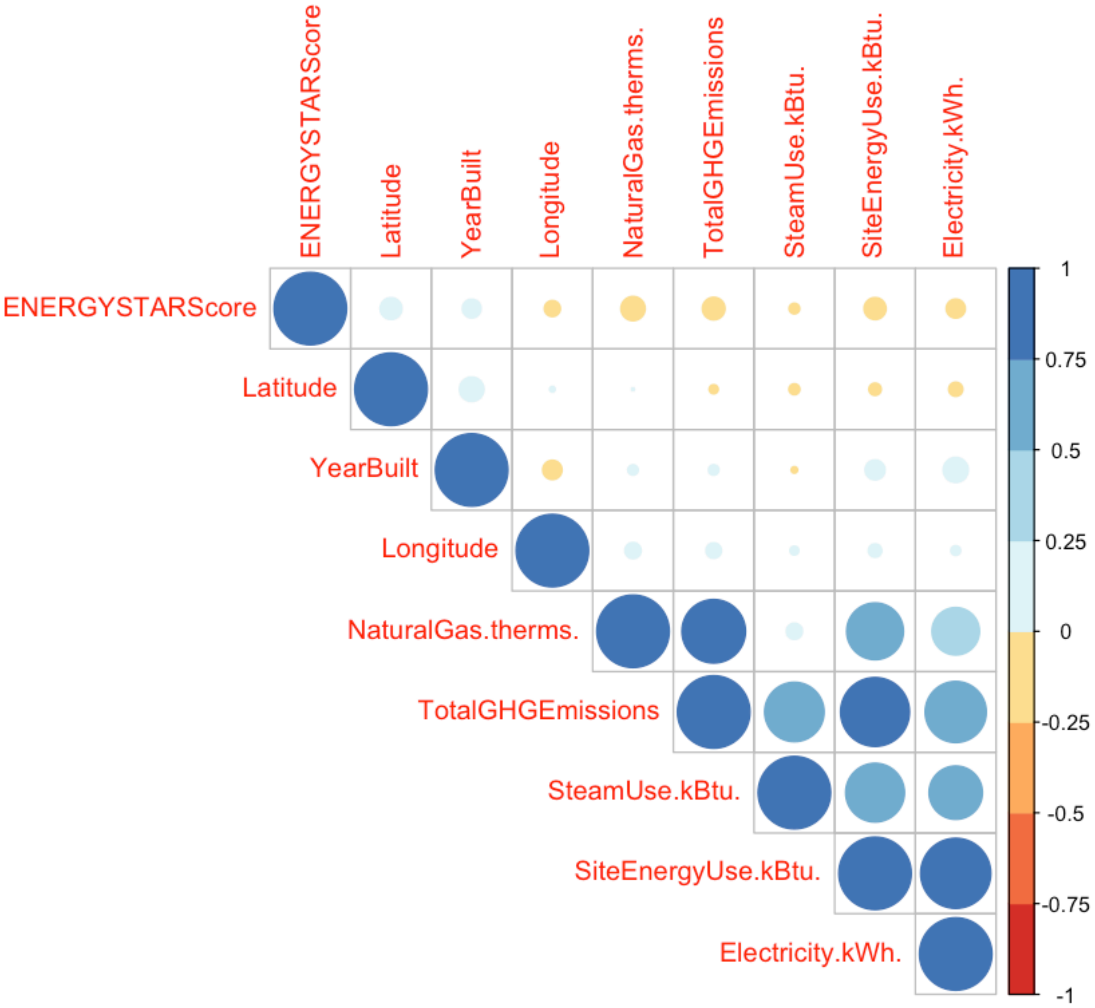
Now to see what Neighborhoods produce the most energy. The chart below confirms our findings from the map above with Downtown producing the most energy.

```
In [158]: data4 %>%
            ggplot(aes(x = Neighborhood, y = SiteEnergyUse.kBtu., fill = Neighborhood)) +
            geom_col() +
            coord_flip() +
            xlab("Neighborhood") +
            ylab("SiteEnergyUse") +
            theme(legend.position='none')
```



We can also use correlation plots to visualize the relationship between factors

```
In [171]: library(corrplot)
          library(RColorBrewer)

          data5 <-data4[,c(-1,-2,-3,-4,-5,-6,-10,-17)] ##Remove string factors
          data5$YearBuilt <- as.numeric(as.character(data5$YearBuilt))
          M <-cor(data5)
          corrplot(M, type="upper", order="hclust",
                   col=brewer.pal(n=8, name="RdYlBu"))
```



Unsurprisingly, TotalGHGEmissions is strongly correlated with SiteEnergyUse. Apart from this, EnergyStarScore and YearBuilt do not correlate with other energy factors.

1.8 Tree Models

We will begin by building a regression Tree model to see what factors are considered important for branching.

For branching, we are not allowed to use factors that have more than 30+ factors, so zipcode, number of floors, and some of the string variables will be removed. YearBuilt will be returned to a numeric to for inclusion in the model.

```
In [174]: data4$YearBuilt <- as.numeric(as.character(data4$YearBuilt))
```

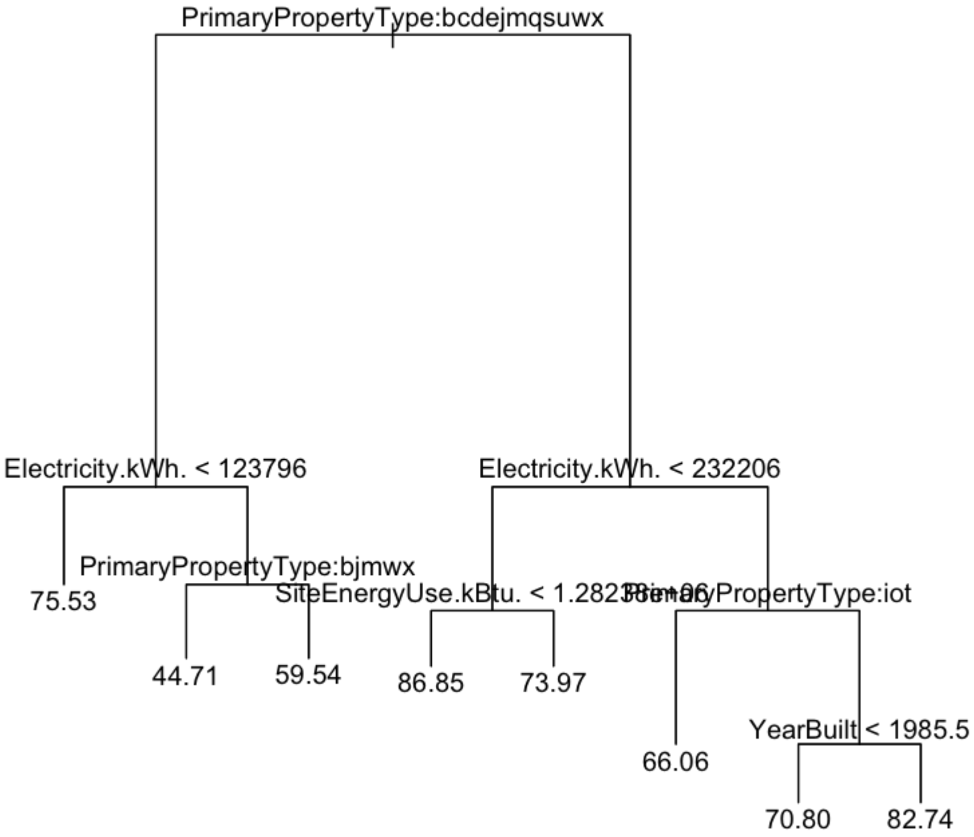
```
In [175]: str(data4)
```

```
'data.frame':        2454 obs. of  17 variables:
 $ OSEBuildingID      : int  1 2 3 5 8 10 12 13 15 16 ...
 $ BuildingType       : Factor w/ 8 levels "Campus","Multifamily HR (10+)",..: 5 5 1 5 5 5 5 4
 $ PrimaryPropertyType: Factor w/ 24 levels "","Distribution Center",..: 5 5 5 5 5 5 5 11 5 5
 $ PropertyName       : Factor w/ 3428 levels "","(71367A) SEATTLE Macy's",..: 1912 2260 405 15
 $ ZipCode            : Factor w/ 57 levels "98006","98011",..: 12 12 12 12 31 12 15 15 12 12
 $ Neighborhood       : Factor w/ 15 levels "","ballard","central",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ Latitude           : num  47.6 47.6 47.6 47.6 47.6 ...
 $ Longitude          : num  -122 -122 -122 -122 -122 ...
 $ YearBuilt          : num  1927 1996 1969 1926 1980 ...
 $ NumberofFloors     : Factor w/ 49 levels "0","1","2","3",..: 13 12 41 11 19 12 16 7 12 26 ..
 $ ENERGYSTARScore    : int  63 72 48 51 78 33 44 2 35 38 ...
 $ SiteEnergyUse.kBtu.: num  7361655 7804844 74470328 7372222 14335778 ...
 $ SteamUse.kBtu.     : num  2122836 0 24313482 2228120 0 ...
 $ Electricity.kWh.   : num  1157783 884161 14276917 881745 1523506 ...
 $ NaturalGas.therms. : num  12885 47881 14440 21356 91376 ...
 $ TotalGHGEmissions  : num  198 267 1571 244 507 ...
 $ ComplianceStatus   : Factor w/ 2 levels "Compliant","Not Compliant": 1 1 1 1 1 1 1 1 1 1 ..
```

```
In [176]: library(tree)
          library(rpart)
          library(rpart.plot)
          tree.data <- tree(ENERGYSTARScore~., data = data4[,c(-1,-4,-5,-10)])
          summary(tree.data)
```

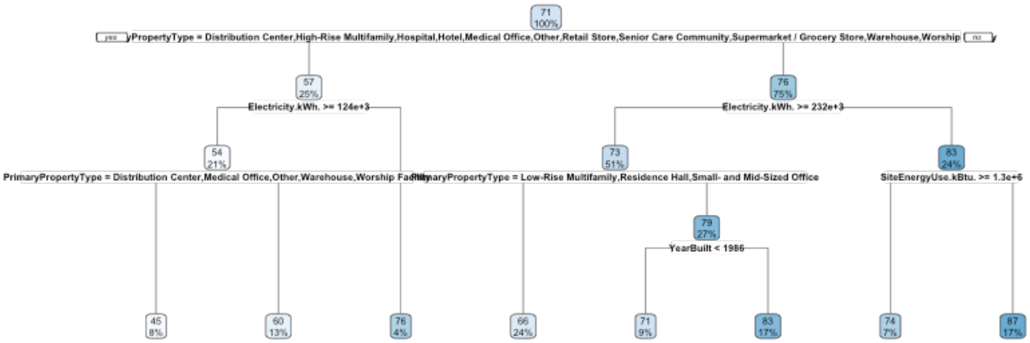
```
Regression tree:
tree(formula = ENERGYSTARScore ~ ., data = data4[, c(-1, -4,
    -5, -10)])
Variables actually used in tree construction:
[1] "PrimaryPropertyType" "Electricity.kWh."    "SiteEnergyUse.kBtu."
[4] "YearBuilt"
Number of terminal nodes:  8
Residual mean deviance:  550.6 = 1347000 / 2446
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -81.74  -11.97    5.94    0.00   15.47   55.29
```

```
In [178]: plot(tree.data)
          text(tree.data)
```


PrimaryPropertyType:bcdejmqsuwx
Electricity.kWh. < 123796
Electricity.kWh. < 232206
PrimaryPropertyType:bjmwx
75.53
SiteEnergyUse.kBtu. < 1.28238e+06
PrimaryPropertyType:iot
44.71
59.54
86.85
73.97
YearBuilt < 1985.5
66.06
70.80
82.74

The base tree model is a little difficult to read so we will try rpart for plotting

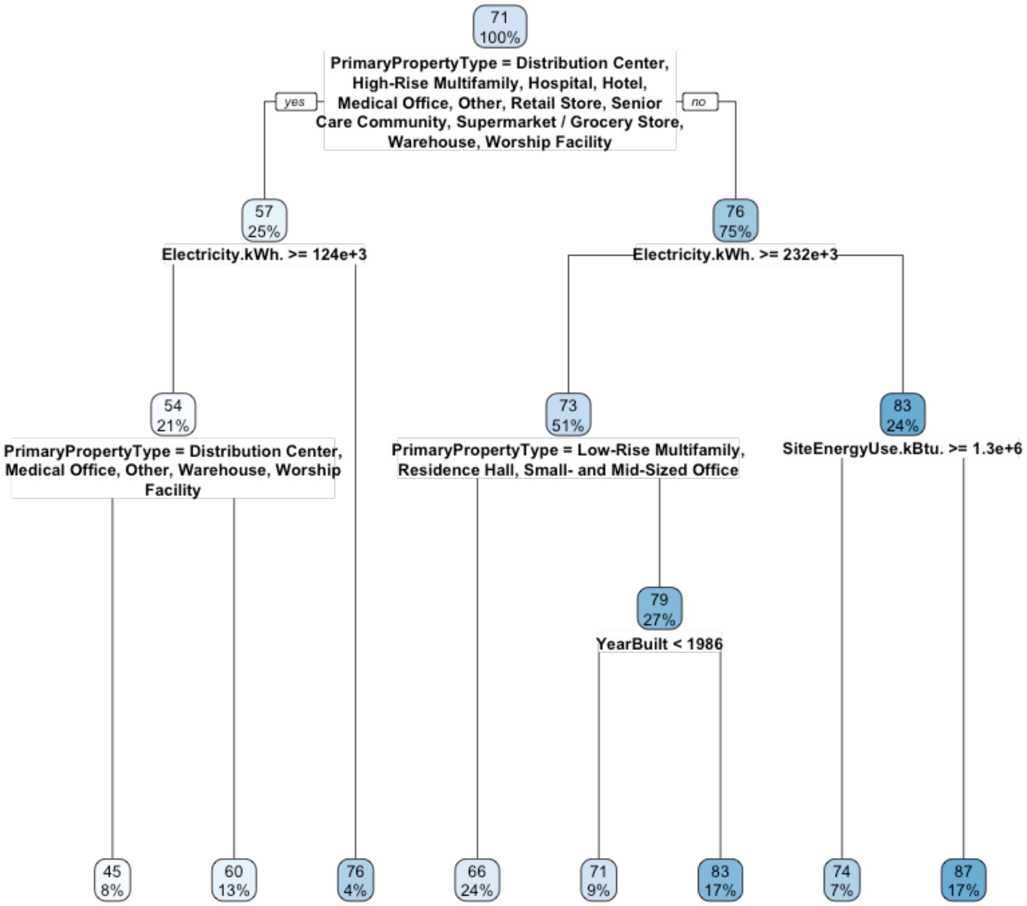
```
In [179]: tree.data1 <- rpart(ENERGYSTARScore~., data = data4[,c(-1,-4,-5,-10)])

In [193]: rpart.plot(tree.data1)
```


71
100%
yes
yPropertyType = Distribution Center,High-Rise Multifamily,Hospital,Hotel,Medical Office,Other,Retail Store,Senior Care Community,Supermarket / Grocery Store,Warehouse,Worship
no
y
57
25%
Electricity.kWh. >= 124e+3
76
75%
Electricity.kWh. >= 232e+3
54
21%
PrimaryPropertyType = Distribution Center,Medical Office,Other,Warehouse,Worship Facility
73
51%
PrimaryPropertyType = Low-Rise Multifamily,Residence Hall,Small- and Mid-Sized Office
83
24%
SiteEnergyUse.kBtu. >= 1.3e+6
79
27%
YearBuilt < 1986
45
8%
60
13%
76
4%
66
24%
71
9%
83
17%
74
7%
87
17%

Readability is still a little difficulty, but we do have the actual text for the factors. We will try the following function to change the text in rpart

```
In [195]: split.fun <- function(x, labs, digits, varlen, faclen)
          {
              # replace commas with spaces (needed for strwrap)
              labs <- gsub(",", ", ", labs)
              for(i in 1:length(labs)) {
                  # split labs[i] into multiple lines
                  labs[i] <- paste(strwrap(labs[i], width=45), collapse="\n")
              }
              labs
          }
```

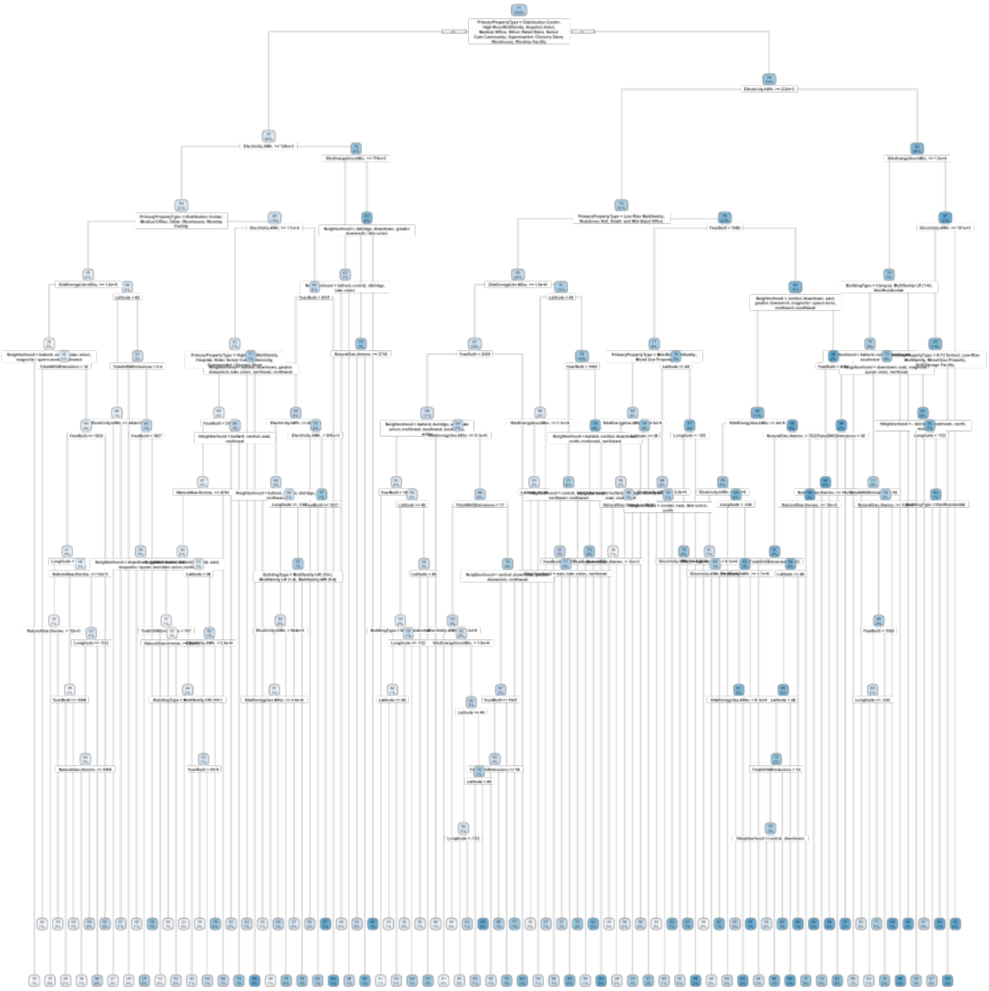
```
rpart.plot(tree.data1, split.fun=split.fun)
```


71
100%
PrimaryPropertyType = Distribution Center, High-Rise Multifamily, Hospital, Hotel, Medical Office, Other, Retail Store, Senior Care Community, Supermarket / Grocery Store, Warehouse, Worship Facility
yes
no
57
25%
Electricity.kWh. >= 124e+3
76
75%
Electricity.kWh. >= 232e+3
54
21%
PrimaryPropertyType = Distribution Center, Medical Office, Other, Warehouse, Worship Facility
73
51%
PrimaryPropertyType = Low-Rise Multifamily, Residence Hall, Small- and Mid-Sized Office
83
24%
SiteEnergyUse.kBtu. >= 1.3e+6
79
27%
YearBuilt < 1986
45
8%
60
13%
76
4%
66
24%
71
9%
83
17%
74
7%
87
17%

This looks much better! Most of the EnergyStar Score prediction are high, with only 3 that are below 70. For the leftmost branch path, it seems Distribution Centers, Warehouses, Medical Offices, and Worship Facilities have a higher chance of a lower Energy score. The building year is only used in one branch in this tree, which is for building built after 1986. There is a 12 score difference in this branch. There might have been some significant update to building codes in the late 1980s that caused newer building to adhere to stricter regulations.

If we increase the cp value in the model, which controls the depth of the tree, we see a huge increase in branches. The first mode used a cp values of around 0.01, so reducing by a factor of ten caused a strong impact.

```
In [196]: tree.data2 <- rpart(ENERGYSTARScore~., data = data4[,c(-1,-4,-5,-10)], cp = 0.001)

In [197]: rpart.plot(tree.data2, split.fun=split.fun)
```



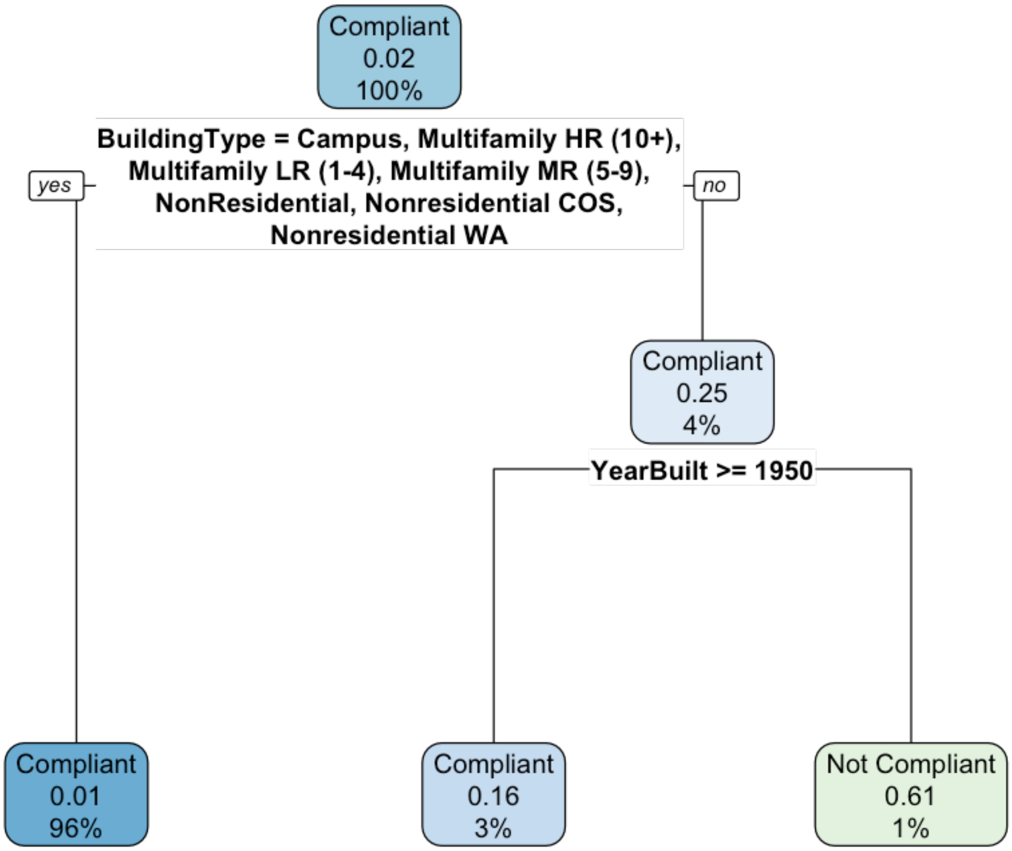
PrimaryPropertyType = Distribution Center, High-Rise Multifamily, Hospital, Hotel, Medical Office, Other, Retail Store, Senior Care Community, Supermarket / Grocery Store, Warehouse, Worship Facility

There are far too many factors in this tree based on the data we have, so it is likely this tree is overfit.

We can also try creating a classification tree based on the Compliance Status variable

```
In [200]: tree.data3 <- rpart(ComplianceStatus~., data = data4[,c(-1,-4,-5,-10)], method = "cla
```

```
In [202]: rpart.plot(tree.data3, split.fun=split.fun)
```


Compliant
0.02
100%
BuildingType = Campus, Multifamily HR (10+), Multifamily LR (1-4), Multifamily MR (5-9), NonResidential, Nonresidential COS, Nonresidential WA
yes
no
Compliant
0.25
4%
YearBuilt >= 1950
Compliant
0.01
96%
Compliant
0.16
3%
Not Compliant
0.61
1%

This model is difficult since there are only 48 out of 2454 rows that are Not Compliant. The first branching path seems to suggest that most apartments and campuses will be compliant with energy codes. Building built before 1950 seem to have a likelihood of failing compliance. This distinction makes sense historically. As new regulation become modernized and advanced, buildings that were built without those qualities will fail to live up to them.

1.9 Predicting EnergyStarScore with Linear Regression

We will now implement linear regression and test its prediction qualities for EnergyStarScore. Since there are a large number of rows without EnergyStarScores, we can build a model based on the rows that do have a score.

```
In [204]: model <- lm(ENERGYSTARScore~., data = data4[,c(-1,-4,-5,-10)])
          summary(model)
```

```
Call:
lm(formula = ENERGYSTARScore ~ ., data = data4[, c(-1, -4, -5,
    -10)])

Residuals:
    Min      1Q  Median      3Q     Max
-85.470 -13.104   6.175  17.081  55.339

Coefficients:
                                                 Estimate Std. Error t value
(Intercept)                                    -4.310e+02  5.303e+03  -0.081
BuildingTypeMultifamily HR (10+)               -2.381e+01  1.484e+01  -1.605
BuildingTypeMultifamily LR (1-4)               -9.317e+00  1.028e+01  -0.906
BuildingTypeMultifamily MR (5-9)               -5.063e+00  1.034e+01  -0.490
BuildingTypeNonResidential                     -1.481e+01  8.042e+00  -1.841
BuildingTypeNonresidential COS                 -2.008e+01  1.146e+01  -1.753
BuildingTypeNonresidential WA                   2.138e+01  2.599e+01   0.823
BuildingTypeSPS-District K-12                   8.002e+00  8.638e+00   0.926
PrimaryPropertyTypeHigh-Rise Multifamily        1.063e+01  1.383e+01   0.769
PrimaryPropertyTypeHospital                     4.470e+01  1.325e+01   3.374
PrimaryPropertyTypeHotel                        1.116e+01  5.338e+00   2.091
PrimaryPropertyTypeK-12 School                  1.007e+01  5.816e+00   1.732
PrimaryPropertyTypeLarge Office                 2.756e+01  5.043e+00   5.466
PrimaryPropertyTypeLow-Rise Multifamily         1.471e+01  7.620e+00   1.931
PrimaryPropertyTypeMedical Office              -9.510e+00  5.962e+00  -1.595
PrimaryPropertyTypeMid-Rise Multifamily         1.631e+01  7.719e+00   2.113
PrimaryPropertyTypeMixed Use Property           2.066e+01  6.299e+00   3.280
PrimaryPropertyTypeOther                       -7.642e-01  8.902e+00  -0.086
PrimaryPropertyTypeRefrigerated Warehouse       3.738e+01  1.164e+01   3.211
PrimaryPropertyTypeResidence Hall               1.619e+01  7.158e+00   2.261
PrimaryPropertyTypeRetail Store                 1.327e+01  5.247e+00   2.529
PrimaryPropertyTypeSelf-Storage Facility        2.423e+01  8.601e+00   2.817
PrimaryPropertyTypeSenior Care Community        9.652e+00  6.482e+00   1.489
PrimaryPropertyTypeSmall- and Mid-Sized Office  1.460e+01  4.586e+00   3.185
PrimaryPropertyTypeSupermarket / Grocery Store  4.618e+00  5.884e+00   0.785
PrimaryPropertyTypeWarehouse                   -2.304e-01  4.531e+00  -0.051
PrimaryPropertyTypeWorship Facility             1.200e+01  5.373e+00   2.234
Neighborhoodballard                            -4.838e+00  8.761e+00  -0.552
Neighborhoodcentral                            -7.044e+00  8.214e+00  -0.858
Neighborhooddelridge                           -1.339e+01  9.241e+00  -1.449
Neighborhooddowntown                           -2.184e+00  8.007e+00  -0.273
Neighborhoodeast                               -2.374e+00  7.921e+00  -0.300
Neighborhoodgreater duwamish                   -7.027e+00  8.364e+00  -0.840
Neighborhoodlake union                          6.544e-01  8.068e+00   0.081
Neighborhoodmagnolia / queen anne              -1.362e+00  8.178e+00  -0.166
Neighborhoodnorth                              -3.727e-01  8.857e+00  -0.042
Neighborhoodnortheast                          -1.230e+00  8.220e+00  -0.150
```

```
Neighborhoodnorthwest                           2.383e+00  8.646e+00   0.276
Neighborhoodsoutheast                          -9.319e+00  8.760e+00  -1.064
Neighborhoodsouthwest                          -4.584e+00  9.111e+00  -0.503
Neighborhoodwater                              -2.042e+01  1.612e+01  -1.266
Latitude                                       -2.833e+01  3.772e+01  -0.751
Longitude                                      -1.485e+01  4.392e+01  -0.338
YearBuilt                                       1.833e-02  1.844e-02   0.994
SiteEnergyUse.kBtu.                            -1.296e-06  4.117e-07  -3.147
SteamUse.kBtu.                                  4.345e-04  9.197e-03   0.047
Electricity.kWh.                                1.186e-04  2.442e-03   0.049
NaturalGas.therms.                              4.336e-02  9.188e-01   0.047
TotalGHGEmissions                              -8.152e+00  1.730e+02  -0.047
ComplianceStatusNot Compliant                   2.586e+00  3.854e+00   0.671
                                               Pr(>|t|)    
(Intercept)                                    0.935234    
BuildingTypeMultifamily HR (10+)               0.108717    
BuildingTypeMultifamily LR (1-4)               0.365061    
BuildingTypeMultifamily MR (5-9)               0.624340    
BuildingTypeNonResidential                     0.065728 .  
BuildingTypeNonresidential COS                 0.079806 .  
BuildingTypeNonresidential WA                  0.410723    
BuildingTypeSPS-District K-12                  0.354374    
PrimaryPropertyTypeHigh-Rise Multifamily       0.442214    
PrimaryPropertyTypeHospital                    0.000752 ***
PrimaryPropertyTypeHotel                       0.036604 *  
PrimaryPropertyTypeK-12 School                 0.083476 .  
PrimaryPropertyTypeLarge Office                5.07e-08 ***
PrimaryPropertyTypeLow-Rise Multifamily        0.053639 .  
PrimaryPropertyTypeMedical Office              0.110801    
PrimaryPropertyTypeMid-Rise Multifamily        0.034705 *  
PrimaryPropertyTypeMixed Use Property          0.001053 ** 
PrimaryPropertyTypeOther                       0.931600    
PrimaryPropertyTypeRefrigerated Warehouse      0.001341 ** 
PrimaryPropertyTypeResidence Hall              0.023830 *  
PrimaryPropertyTypeRetail Store                0.011500 *  
PrimaryPropertyTypeSelf-Storage Facility       0.004884 ** 
PrimaryPropertyTypeSenior Care Community       0.136614    
PrimaryPropertyTypeSmall- and Mid-Sized Office 0.001468 ** 
PrimaryPropertyTypeSupermarket / Grocery Store 0.432547    
PrimaryPropertyTypeWarehouse                   0.959449    
PrimaryPropertyTypeWorship Facility            0.025553 *  
Neighborhoodballard                            0.580808    
Neighborhoodcentral                            0.391183    
Neighborhooddelridge                           0.147584    
Neighborhooddowntown                           0.785094    
Neighborhoodeast                               0.764469    
Neighborhoodgreater duwamish                   0.400932    
Neighborhoodlake union                         0.935368    
```

```
Neighborhoodmagnolia / queen anne              0.867783
Neighborhoodnorth                              0.966440
Neighborhoodnortheast                          0.881033
Neighborhoodnorthwest                          0.782857
Neighborhoodsoutheast                          0.287530
Neighborhoodsouthwest                          0.614895
Neighborhoodwater                              0.205507
Latitude                                       0.452705
Longitude                                      0.735322
YearBuilt                                      0.320214
SiteEnergyUse.kBtu.                            0.001671 **
SteamUse.kBtu.                                 0.962321
Electricity.kWh.                               0.961262
NaturalGas.therms.                             0.962367
TotalGHGEmissions                              0.962423
ComplianceStatusNot Compliant                  0.502260
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 24.43 on 2404 degrees of freedom
Multiple R-squared:  0.1608,Adjusted R-squared:  0.1436
F-statistic: 9.398 on 49 and 2404 DF,  p-value: < 2.2e-16
```

This model has a very low R-squared value so it would not be considered a trustworthy model for prediction

```
In [205]: data_miss <- data4[miss_ind,]
          data_miss_filt <- data_miss[,c(-1,-4,-5,-10)]
```

```
In [206]: test <- data_miss_filt[1,-7]
          pred_model <- predict(model, test)
          pred_model ## Predicted score of 63
```

7: 63.436154911416

```
In [207]: test1 <- data_miss_filt[2,-7]
          pred_model1 <- predict(model, test1)
          pred_model1 ## Predicted score of 79
```

10: 79.5330582768686

```
In [212]: test1 <- data_miss_filt[,-7]
          pred_model1 <- predict(model, test1)
          head(pred_model1, n = 50) ## Predicted scores for the first 50 rows missing ENERGYST
```

7 63.436154911416 **10** 79.5330582768686 **21** 32.3187647991854 **23** 81.5983202889937 **28** 62.046643197868 **31** 84.9952817822952 **38** 88.832608758756 **39** 82.4573465272006 **47**

68.5234379087978 **50** 51.8989303197668 **51** 65.1873780999858 **52** 62.3764472636035 **58**
65.512896174283 **61** 58.1559057426739 **63** 88.6150804882614 **89** 87.7302492039307 **95**
67.5708483944942 **96** 90.4320503350486 **97** 63.7466757183605 **100** 84.7948617330439 **103**
64.5829864168984 **104** 79.2335120356004 **109** 78.0911750814107 **129** 77.9505106027195 **130**
79.1916017447069 **137** 75.8525117398317 **141** 84.9995046472036 **143** 89.621767831762 **148**
84.1447122866903 **150** 89.0821594849828 **151** 65.2008018271486 **152** 87.9247161894575 **168**
37.1954482762885 **169** 79.7992980301178 **170** 82.1284810578212 **174** 78.280467441695 **179**
63.9894336125717 **184** 83.7663102553618 **185** 89.8402303184618 **204** 78.7611139518319 **207**
57.4865219728003 **210** 69.0201016709983 **211** 79.856668492807 **214** 77.8654356835987 **220**
68.3669869584572 **221** 68.3768238030563 **222** 79.6116938466176 **223** 79.9372947814647 **224**
78.5768079130676 **229** 79.912410892627