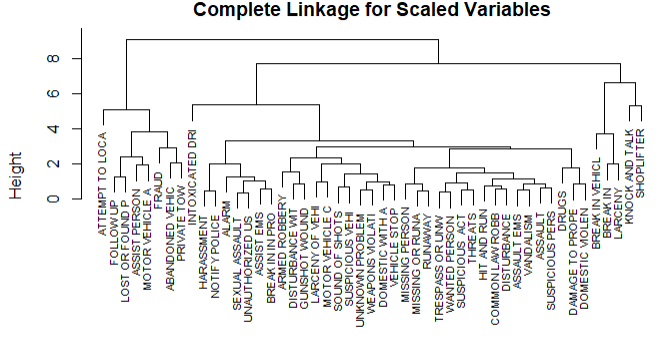
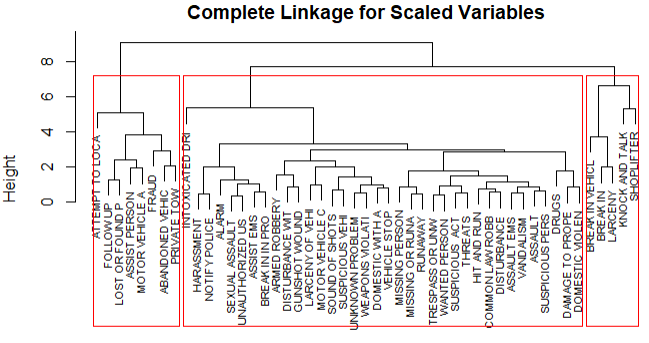
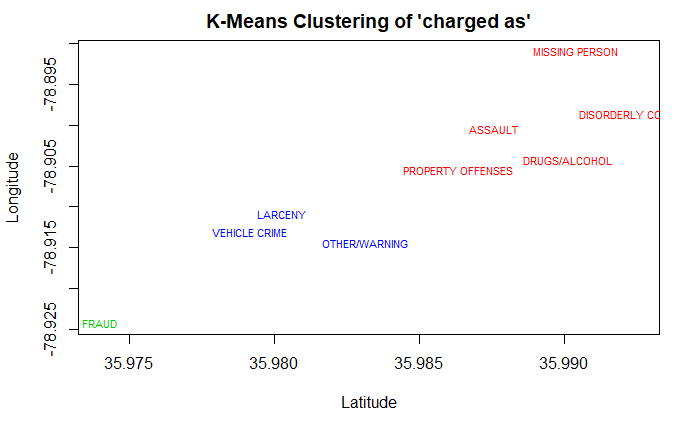
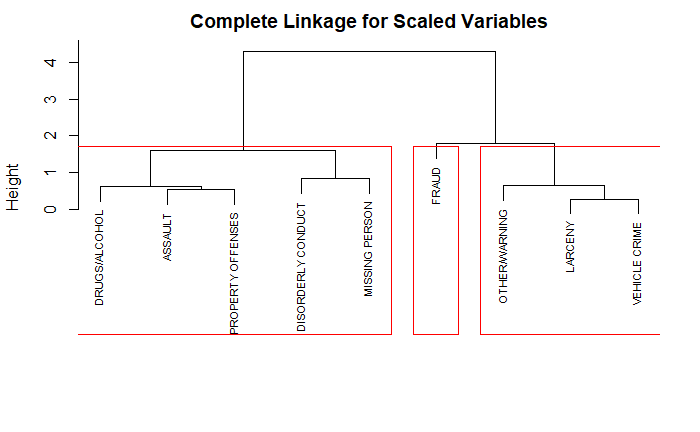
reportedas:

* Hierarchical Clustering
  + Grouped data by “reportedas”, and then calculated summary stats:
    - Counts
    - Latitude mean and sd
    - Longitude mean and sd
    - Hour mean and sd
  + For sake of dendrogram size, only used counts > 150
  + Used complete linkage because distribution is spherical, also used scaled data because variables are measured under different scales (counts vs. latitude/longitude vs hour)
  + 
  + 
  + Findings:
    - Makes sense that home break-ins, vehicle break-ins, larceny, and shoplifting are all clustered together
    - Some other pairs that make sense:
      * Abandoned vehicle/private towing
      * Missing person/runaway
    - Some interesting pairs:
      * Damage to property/domestic violence
      * Sound of shots/suspicious vehicles
* K-means clustering
  + Basically, was almost the exact same dendrogram as hierarchical clustering
  + 98% (48/49) of the crimes were classified into the same cluster in K-means as hierarchical
  + The only difference was knock and talk (residential search request)
    - Make note of this on dendrogram: knock and talk group 3 → 2

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Charged as

* K-Means Clustering
  + Grouped data by “charged as” and then calculated summary stats:
    - Mean latitude
    - Mean longitude
  + k=3
  + 
  + F
* Hierarchical clustering
  + Same graph as K-means; 100% similarity
  + Complete linkage of scaled data
  + 

Performing random forest classifying charge:

Text

Description automatically generated with medium confidence

* Month and hour of occurence have a much higher influence on charge than district and whether or not it occured downtown

a

Performing bagging classifying charge:

Table

Description automatically generated

A picture containing logo

Description automatically generated

l

Using C5.0 classifying charge:

Table

Description automatically generated

C5.0 is an improvement on C4.5. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized [information gain](https://en.wikipedia.org/wiki/Information_gain_in_decision_trees) (difference in [entropy](https://en.wikipedia.org/wiki/Entropy_(information_theory))). The attribute with the highest normalized information gain is chosen to make the decision.

----------------------------------------

Using classification trees to predict downtown without using classifiers like DIST, BIG\_ZONE, or area is basically useless and the tree doesn’t grow past one node: independent variables do not provide enough information to grow the tree.

**Counts:**

Chart, histogram

Description automatically generated

Code used:

Count the number of crimes per hour

```{r}

counts <- table(crimes$hour\_occur)

counts

barplot(counts, main="Crime Distribution by Time", xlab="Time of Day")

```

**Linear Models:**

Code:

Run the linear model

```{r lm}

lm.count <- lm(X ~ hour\_occur, data = crimes)

summary(lm.count)

```

The hour a crime occurs is statistically significant with p=2e-16. Adjusted R^2 = .0001268

Forward Selection:

```{r lm}

lm.count <- lm(X ~ hour\_occur + month\_occur, data = crimes)

summary(lm.count)

```

Hour statistically significant at p=.00191 and month statistically significant at p=2e-16. Adjusted R^2 = 0.01929

```{r lm}

lm.count <- lm(X ~ hour\_occur + month\_occur + year\_occur, data = crimes)

summary(lm.count)

```

Hour statistically significant at p=.05 and month statistically significant at p=2e-16, year statistically significant at p=2e-16. Adjusted R^2 = 0.05559

```{r lm}

lm.count <- lm(X ~ hour\_occur + month\_occur + year\_occur + year\_rept, data = crimes)

summary(lm.count)

```

Hour statistically significant at p=.05 and month statistically significant at p=2e-16, year statistically significant at p=2e-16. Adjusted R^2 = 0.05559

Findings:

-linear model was basically useless provides us with tiny Adjusted R squared.. Is this even worth including?

-the largest adjusted R^2 is only ever around .06

-n>p

**Piecewise**