WADEPS_Tut_Public_Classification

2024-04-14

Public Safety Question

Suppose you are moving to, or living in WA, and you want to ensure that the location you live has low crime. Perhaps you're question is, which counties in WA are most likely to contain crime? We can answer this question using logistic regression. Before diving into this, lets first upload the data. A given data observation depicts the counts of different crime-types in each county of WA in a given year.

```
#install.packages("readxl")
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```
directory="C:/Users/eliza/OneDrive/Documents/WADEPS/Weekly notes an ideas/Classification_SVM_Example_04
data=read_excel(directory)
data
```

```
## # A tibble: 2,907 x 34
##    IndexYear county Location    Population Total  Rate PrsnTotal PrsnRate Murder
##    <chr>     <chr>  <chr>            <dbl> <dbl> <dbl>     <dbl>    <dbl>  <dbl>
##  1 2012      ADAMS  Adams Coun~       9860   620  62.9       145     14.7      0
##  2 2013      ADAMS  Adams Coun~       9935   693  69.8       130     13.1      2
##  3 2014      ADAMS  Adams Coun~      10025   688  68.6       165     16.5      1
##  4 2015      ADAMS  Adams Coun~       9960   685  68.8       139     14.0      0
##  5 2016      ADAMS  Adams Coun~       9975   571  57.2       166     16.6      1
##  6 2017      ADAMS  Adams Coun~      10035   498  49.6       133     13.3      1
##  7 2018      ADAMS  Adams Coun~      10090   456  45.2       142     14.1      0
##  8 2019      ADAMS  Adams Coun~      10145   389  38.3       118     11.6      0
##  9 2020      ADAMS  Adams Coun~      10275   438  42.6       129     12.6      0
## 10 2021      ADAMS  Adams Coun~      10410   677  65.0       181     17.4      2
## # i 2,897 more rows
## # i 25 more variables: Manslaughter <dbl>, Forcible_Sex <dbl>, Assault <dbl>,
## #   Non_Forcible_Sex <dbl>, Kidnapping_Abduction <dbl>,
## #   Human_Trafficking <dbl>, Viol_Of_No_Contact <dbl>, PrprtyTotal <dbl>,
## #   PrprtyRate <dbl>, Arson <dbl>, Bribery <dbl>, Burglary <dbl>,
## #   Counterfeiting_Forgery <dbl>, Destruction_of_Property <dbl>,
## #   Extortion_Blackmail <dbl>, Robbery <dbl>, Theft <dbl>, SctyTotal <dbl>, ...
```

Clean Your Data

The first step to any data analysis is to clean the data. That is, we need to remove any observations with empty entries. These correspond to rows with "NA" in them. Let's check if there are any such rows.

```
NA_row_numbers=!complete.cases(data) #identify row numbers with empty entries
NA_rows=data[NA_row_numbers,] #identify all rows in the data with empty entries
dim(NA_rows)[1] #count the number of rows with empty entries
```

```
## [1] 27
```

As we can see, there are 27 rows with empty entries. To convince you, let's visualize one of these rows.

```
NA_rows[1,]
```

```
## # A tibble: 1 x 34
##   IndexYear county Location     Population Total  Rate PrsnTotal PrsnRate Murder
##   <chr>     <chr>  <chr>             <dbl> <dbl> <dbl>     <dbl>    <dbl>  <dbl>
## 1 2012      CLARK  WSU Vancouv~         NA     9    NA         1       NA      0
## # i 25 more variables: Manslaughter <dbl>, Forcible_Sex <dbl>, Assault <dbl>,
## #   Non_Forcible_Sex <dbl>, Kidnapping_Abduction <dbl>,
## #   Human_Trafficking <dbl>, Viol_Of_No_Contact <dbl>, PrprtyTotal <dbl>,
## #   PrprtyRate <dbl>, Arson <dbl>, Bribery <dbl>, Burglary <dbl>,
## #   Counterfeiting_Forgery <dbl>, Destruction_of_Property <dbl>,
## #   Extortion_Blackmail <dbl>, Robbery <dbl>, Theft <dbl>, SctyTotal <dbl>,
## #   SctyRate <dbl>, Drug_Violations <dbl>, Gambling_Violations <dbl>, ...
```

The first observation from Vancouver WA contains an empty entry "NA" in the population column. We want to remove all such rows from our data. Let's do this, and visualize a few of the remaining observations.

```
data_removed_NAs=data[!NA_row_numbers,]
data_removed_NAs
```

```
## # A tibble: 2,880 x 34
##    IndexYear county Location    Population Total  Rate PrsnTotal PrsnRate Murder
##    <chr>     <chr>  <chr>            <dbl> <dbl> <dbl>     <dbl>    <dbl>  <dbl>
##  1 2012      ADAMS  Adams Coun~       9860   620  62.9       145     14.7      0
##  2 2013      ADAMS  Adams Coun~       9935   693  69.8       130     13.1      2
##  3 2014      ADAMS  Adams Coun~      10025   688  68.6       165     16.5      1
##  4 2015      ADAMS  Adams Coun~       9960   685  68.8       139     14.0      0
##  5 2016      ADAMS  Adams Coun~       9975   571  57.2       166     16.6      1
##  6 2017      ADAMS  Adams Coun~      10035   498  49.6       133     13.3      1
##  7 2018      ADAMS  Adams Coun~      10090   456  45.2       142     14.1      0
##  8 2019      ADAMS  Adams Coun~      10145   389  38.3       118     11.6      0
##  9 2020      ADAMS  Adams Coun~      10275   438  42.6       129     12.6      0
## 10 2021      ADAMS  Adams Coun~      10410   677  65.0       181     17.4      2
## # i 2,870 more rows
## # i 25 more variables: Manslaughter <dbl>, Forcible_Sex <dbl>, Assault <dbl>,
## #   Non_Forcible_Sex <dbl>, Kidnapping_Abduction <dbl>,
## #   Human_Trafficking <dbl>, Viol_Of_No_Contact <dbl>, PrprtyTotal <dbl>,
## #   PrprtyRate <dbl>, Arson <dbl>, Bribery <dbl>, Burglary <dbl>,
## #   Counterfeiting_Forgery <dbl>, Destruction_of_Property <dbl>,
## #   Extortion_Blackmail <dbl>, Robbery <dbl>, Theft <dbl>, SctyTotal <dbl>, ...
```

We now have 2880 observations out of the original 2907. Now that we have cleaned our data, lets familiarize ourselves with logistic regression.

Logistic Regression

Logistic regression is a method of predicting the probability (or odds) of an event occurring, or not occurring, given a set of data. Think of it as this: you plug in a set of data to a predictive function you are building, your function learns the patterns of yes/no events of a given outcome variable of interest, and then uses what it learned to predict the amount of influence each variable in the data has on the odds of this outcome. This function returns this magnitude of influence as a number between -1 and 1. This number is called the correlation coefficient, as it describes how correlated each variable in the data is to the outcome variable in

terms of probability. A coefficient of 0 indicates that a unit increase in a given variable has no influence on the odds of the outcome variable, while coefficients near plus (minus) 1 indicate that a unit increase in a given variable has perfect positive (negative) influence on the odds of the outcome variable.

Like any function, we have inputs and outputs. Our inputs in this case are continuous (numerical) or binary (yes/no) variables, and our output is a binary variable. Numerically, this means none of our variables can have 3 or more discrete, whole-numbers, as responses. For instance, if one input variable "Race" has 4 categories "white", "black", "asian", or "other", then we would need to convert it into four binary input variables "white", "black", "asian", and "other". Our output is a binary variable. For instance, the variable "WhiteSuspect" with response 1 if the subject is white and 0 if not could be an outcome variable.

Our Public Safety Question

Because our research question is related to crime, lets first visualize all crime-related variables in our data set.

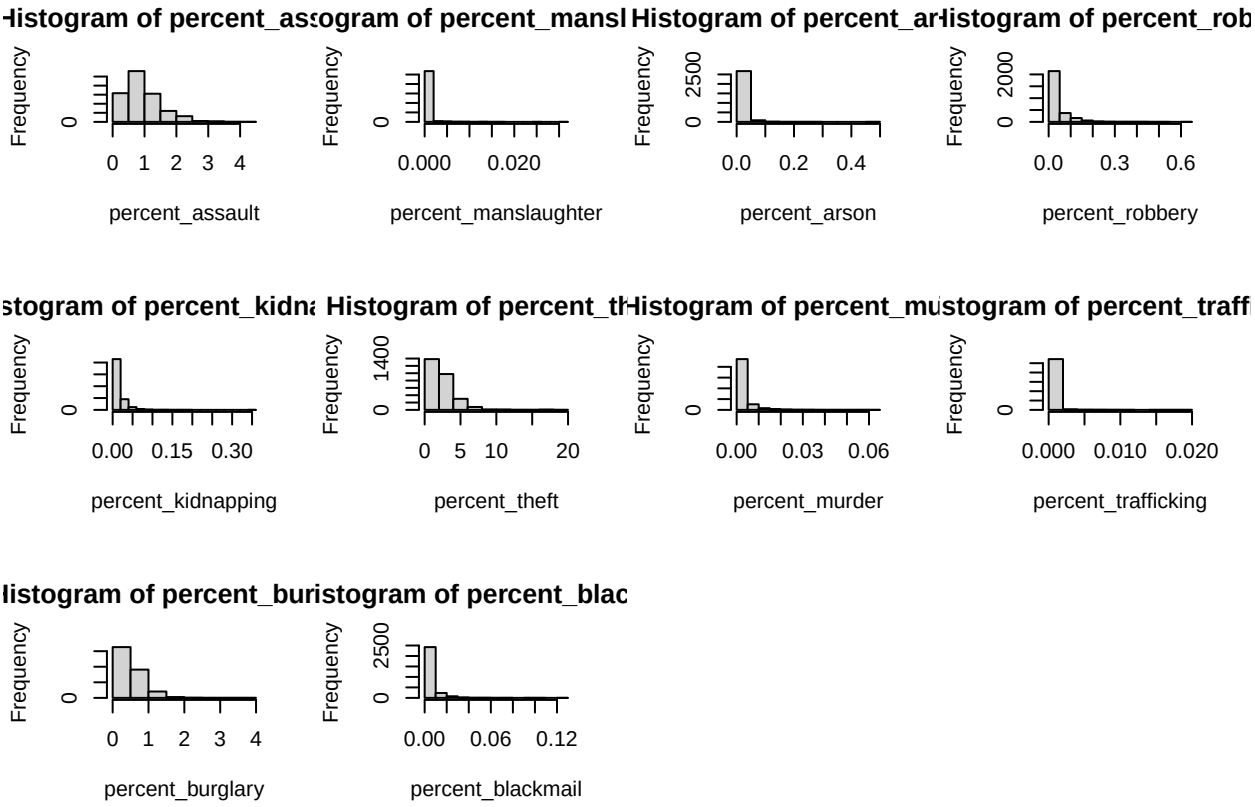
```
colnames(data_removed_NAs)
```

```
##  [1] "IndexYear"                "county"
##  [3] "Location"                 "Population"
##  [5] "Total"                    "Rate"
##  [7] "PrsnTotal"                "PrsnRate"
##  [9] "Murder"                   "Manslaughter"
## [11] "Forcible_Sex"             "Assault"
## [13] "Non_Forcible_Sex"         "Kidnapping_Abduction"
## [15] "Human_Trafficking"        "Viol_Of_No_Contact"
## [17] "PrprtyTotal"              "PrprtyRate"
## [19] "Arson"                    "Bribery"
## [21] "Burglary"                 "Counterfeiting_Forgery"
## [23] "Destruction_of_Property" "Extortion_Blackmail"
## [25] "Robbery"                  "Theft"
## [27] "SctyTotal"                "SctyRate"
## [29] "Drug_Violations"          "Gambling_Violations"
## [31] "Pornography"              "Prostitution"
## [33] "Weapon_Law_Violation"     "Animal_Cruelty"
```

Suppose we are interested in whether or not a given observation reveals that over 50% of the total population contained a particular crime of interest in a given year in a given county. Our binary outcome variable can then be defined using the ratio of the crime type to "Population" in our data set. Let's define this variable as "percent_crime", and observe its distribution for each crime.

```
percent_assault=(data_removed_NAs$Assault/data_removed_NAs$Population)*100
percent_manslaughter=(data_removed_NAs$Manslaughter/data_removed_NAs$Population)*100
percent_arson=(data_removed_NAs$Arson/data_removed_NAs$Population)*100
percent_robbery=(data_removed_NAs$Robbery/data_removed_NAs$Population)*100
percent_kidnapping=(data_removed_NAs$Kidnapping_Abduction/data_removed_NAs$Population)*100
percent_theft=(data_removed_NAs$Theft/data_removed_NAs$Population)*100
percent_murder=(data_removed_NAs$Murder/data_removed_NAs$Population)*100
percent_trafficking=(data_removed_NAs$Human_Trafficking/data_removed_NAs$Population)*100
percent_burglary=(data_removed_NAs$Burglary/data_removed_NAs$Population)*100
percent_blackmail=(data_removed_NAs$Extortion_Blackmail/data_removed_NAs$Population)*100

par(mfrow=c(3,4))
hist(percent_assault); hist(percent_manslaughter); hist(percent_arson)
hist(percent_robbery); hist(percent_kidnapping); hist(percent_theft)
hist(percent_murder); hist(percent_trafficking); hist(percent_burglary)
hist(percent_blackmail)
```



It appears that the most-occuring crimes in WA are assault, theft, and burglary. Theft occurs the most according to its histogram, so lets select this crime to construct our outcome variable. We define our outcome variable as "large_theft_rate", where its response is 1 if there are greater than 3% of the population committed theft in a given year in a given county, and 0 otherwise.

First, notice that some of our observations have "Population"=0. We of course are only interested in studying percent theft in counties with population greater than 0. So let's remove these from our data set and check there are no observations with 0 population.

```
population_zero_row_numbers=which(data_removed_NAs$Population==0)
data_removed_NAs_new=data_removed_NAs[-population_zero_row_numbers,]
length(which(data_removed_NAs_new$Population==0))
```

```
## [1] 0
```

Now we must re-define our percent theft variable with this updated data set.

```
percent_theft=(data_removed_NAs_new$Theft/data_removed_NAs_new$Population)*100
```

```
large_theft_rate=c()
for (observation in 1:dim(data_removed_NAs_new)[1]){
  if (percent_theft[observation]>=3){
    large_theft_rate[observation]=1}
  else{large_theft_rate[observation]=0}
}
```

Let's now observe the counts of observations depicting at least 5% of the population committing theft and those that don't. Observe we have 2558 observations for which less than 5% of the population committed

murder, and 217 for which at least 5% of the population did in a given county of WA.

```
length(which(large_theft_rate==0));length(which(large_theft_rate==1))
```

```
## [1] 2013
```

```
## [1] 762
```

Now we have our binary outcome variable. With this, we can evaluate the influence of our input variables in the data on large theft rates in each county. To do this, we can first break up our data set into observations based on which county they come from. This information comes from the "Location" column.

Influence of Variables on Large Theft Rates by City

Lets observe the different counties represented in our data:

```
unique(data_removed_NAs_new$county)
```

```
##  [1] "ADAMS"         "ASOTIN"        "BENTON"        "CHELAN"        "CLALLAM"
##  [6] "CLARK"         "COLUMBIA"      "COWLITZ"       "DOUGLAS"       "FERRY"
## [11] "FRANKLIN"      "GARFIELD"      "GRANT"         "GRAYS HARBOR"  "ISLAND"
## [16] "JEFFERSON"     "KING"          "KITSAP"        "KITTITAS"      "KLICKITAT"
## [21] "LEWIS"         "LINCOLN"       "MASON"         "OKANOGAN"      "PACIFIC"
## [26] "PEND OREILLE"  "PIERCE"        "SAN JUAN"      "SKAGIT"        "SKAMANIA"
## [31] "SNOHOMISH"     "SPOKANE"       "STATEWIDE"     "STEVENS"       "THURSTON"
## [36] "WAHKIAKUM"     "WALLA WALLA"   "WHATCOM"       "WHITMAN"       "YAKIMA"
```

Lets also observe the different cities within each county represented in our data.

```
unique(data_removed_NAs_new$Location)
```

```
##   [1] "Adams County Sheriffs Office"
##   [2] "COUNTY TOTAL"
##   [3] "Othello Police Department"
##   [4] "Ritzville Police Department"
##   [5] "Asotin County Sheriffs Office"
##   [6] "Asotin Police Department"
##   [7] "Clarkston Police Department"
##   [8] "Benton County Sheriffs Office"
##   [9] "Kennewick Police Department"
##  [10] "Prosser Police Department"
##  [11] "Richland Police Department"
##  [12] "West Richland Police Department"
##  [13] "Chelan County Sheriffs Office"
##  [14] "Wenatchee Police Department"
##  [15] "Clallam County Sheriffs Office"
##  [16] "Forks Police Department"
##  [17] "Port Angeles Police Department"
##  [18] "Sequim Police Department"
##  [19] "Battle Ground Police Department"
##  [20] "Camas Police Department"
##  [21] "Clark County Sheriffs Office"
##  [22] "La Center Police Department"
##  [23] "Ridgefield Police Department"
##  [24] "Vancouver Police Department"
##  [25] "Washougal Police Department"
##  [26] "Columbia County Sheriffs Office"
```

```
##  [27] "Castle Rock Police Department"
##  [28] "Cowlitz County Sheriffs Office"
##  [29] "Kalama Police Department"
##  [30] "Kelso Police Department"
##  [31] "Longview Police Department"
##  [32] "Woodland Police Department"
##  [33] "Douglas County Sheriffs Office"
##  [34] "East Wenatchee Police Department"
##  [35] "Ferry County Sheriffs Office"
##  [36] "Republic Police Department"
##  [37] "Connell Police Department"
##  [38] "Franklin County Sheriffs Office"
##  [39] "Pasco Police Department"
##  [40] "Garfield County Sheriffs Office"
##  [41] "Ephrata Police Department"
##  [42] "Grand Coulee Police Department"
##  [43] "Grant County Sheriffs Office"
##  [44] "Mattawa Police Department"
##  [45] "Moses Lake Police Department"
##  [46] "Quincy Police Department"
##  [47] "Royal City Police Department"
##  [48] "Soap Lake Police Department"
##  [49] "Warden Police Department"
##  [50] "Aberdeen Police Department"
##  [51] "Cosmopolis Police Department"
##  [52] "Elma Police Department"
##  [53] "Grays Harbor County Sheriffs Office"
##  [54] "Hoquiam Police Department"
##  [55] "McCleary Police Department"
##  [56] "Montesano Police Department"
##  [57] "Oakville Police Department"
##  [58] "Ocean Shores Police Department"
##  [59] "Westport Police Department"
##  [60] "Coupeville Police Department"
##  [61] "Island County Sheriffs Office"
##  [62] "Langley Police Department"
##  [63] "Oak Harbor Police Department"
##  [64] "Jefferson County Sheriffs Office"
##  [65] "Port Townsend Police Department"
##  [66] "Algona Police Department"
##  [67] "Auburn Police Department"
##  [68] "Beaux Arts Police Department"
##  [69] "Bellevue Police Department"
##  [70] "Black Diamond Police Department"
##  [71] "Bothell Police Department"
##  [72] "Burien Police Department"
##  [73] "Carnation Police Department"
##  [74] "Clyde Hill Police Department"
##  [75] "Covington Police Department"
##  [76] "Des Moines Police Department"
##  [77] "Duvall Police Department"
##  [78] "Enumclaw Police Department"
##  [79] "Federal Way Police Department"
##  [80] "Issaquah Police Department"
```

```
##  [81] "Kenmore Police Department"
##  [82] "Kent Police Department"
##  [83] "King County Sheriffs Office"
##  [84] "Kirkland Police Department"
##  [85] "Lake Forest Park Police Department"
##  [86] "Maple Valley Police Department"
##  [87] "Medina Police Department"
##  [88] "Mercer Island Police Department"
##  [89] "Newcastle Police Department"
##  [90] "Normandy Park Police Department"
##  [91] "North Bend Police Department"
##  [92] "Pacific Police Department"
##  [93] "Redmond Police Department"
##  [94] "Renton Police Department"
##  [95] "Sammamish Police Department"
##  [96] "Seatac Police Department"
##  [97] "Seattle Police Department"
##  [98] "Shoreline Police Department"
##  [99] "Skykomish Police Department"
## [100] "Snoqualmie Police Department"
## [101] "Tukwila Police Department"
## [102] "Woodinville Police Department"
## [103] "Yarrow Point Police Department"
## [104] "Bainbridge Island Police Department"
## [105] "Bremerton Police Department"
## [106] "Kitsap County Sheriffs Office"
## [107] "Kitsap Police Department"
## [108] "Port Orchard Police Department"
## [109] "Poulsbo Police Department"
## [110] "Cle Elum -Roslyn Police Department"
## [111] "Cle Elum Police Department"
## [112] "Ellensburg Police Department"
## [113] "Kittitas County Sheriffs Office"
## [114] "Kittitas Police Department"
## [115] "Bingen Police Department"
## [116] "Goldendale Police Department"
## [117] "Klickitat County Sheriffs Office"
## [118] "White Salmon Police Department"
## [119] "Centralia Police Department"
## [120] "Chehalis Police Department"
## [121] "Lewis County Sheriffs Office"
## [122] "Morton Police Department"
## [123] "Mossyrock Police Department"
## [124] "Napavine Police Department"
## [125] "Pe Ell Police Department"
## [126] "Toledo Police Department"
## [127] "Vader Police Department"
## [128] "Winlock Police Department"
## [129] "Lincoln County Sheriffs Office"
## [130] "Odessa Police Department"
## [131] "Reardan Police Department"
## [132] "Wilbur Police Department"
## [133] "Mason County Sheriffs Office"
## [134] "Shelton Police Department"
```

```
## [135] "Brewster Police Department"
## [136] "Coulee Dam Police Department"
## [137] "Okanogan County Sheriffs Office"
## [138] "Omak Police Department"
## [139] "Oroville Police Department"
## [140] "Tonasket Police Department"
## [141] "Twisp Police Department"
## [142] "Winthrop Marshal"
## [143] "Winthrop Police Department"
## [144] "Ilwaco Police Department"
## [145] "Long Beach Police Department"
## [146] "Pacific County Sheriffs Office"
## [147] "Raymond Police Department"
## [148] "South Bend Police Department"
## [149] "Newport Police Department"
## [150] "Pend Oreille County Sheriffs Office"
## [151] "Bonney Lake Police Department"
## [152] "Buckley Police Department"
## [153] "DuPont Police Department"
## [154] "Dupont Police Department"
## [155] "Eatonville Police Department"
## [156] "Edgewood Police Department"
## [157] "Fife Police Department"
## [158] "Fircrest Police Department"
## [159] "Gig Harbor Police Department"
## [160] "Lakewood Police Department"
## [161] "Milton Police Department"
## [162] "Orting Police Department"
## [163] "Pierce County Sheriffs Office"
## [164] "Puyallup Police Department"
## [165] "Roy Police Department"
## [166] "Ruston Police Department"
## [167] "Steilacoom Police Department"
## [168] "Sumner Police Department"
## [169] "Tacoma Police Department"
## [170] "University Place Police Department"
## [171] "San Juan County Sheriffs Office"
## [172] "San Juan Police Department"
## [173] "Anacortes Police Department"
## [174] "Burlington Police Department"
## [175] "Mount Vernon Police Department"
## [176] "Sedro Woolley Police Department"
## [177] "Skagit County Sheriffs Office"
## [178] "Skamania County Sheriffs Office"
## [179] "Arilington Police Department"
## [180] "Arlington Police Department"
## [181] "Brier Police Department"
## [182] "Darrington Police Department"
## [183] "Edmonds Police Department"
## [184] "Everett Police Department"
## [185] "Gold Bar Police Department"
## [186] "Granite Falls Police Department"
## [187] "Index Police Department"
## [188] "Lake Stevens Police Department"
```

```
## [189] "Lynnwood Police Department"
## [190] "Marysville Police Department"
## [191] "Mill Creek Police Department"
## [192] "Monroe Police Department"
## [193] "Mountlake Terrace Police Department"
## [194] "Mukilteo Police Department"
## [195] "Snohomish County Sheriffs Office"
## [196] "Snohomish Police Department"
## [197] "Stanwood Police Department"
## [198] "Sultan Police Department"
## [199] "Woodway Police Department"
## [200] "Airway Heights Police Department"
## [201] "Cheney Police Department"
## [202] "Liberty Lake Police Department"
## [203] "Spokane County Sheriffs Office"
## [204] "Spokane Police Department"
## [205] "Spokane Valley Police Department"
## [206] "STATEWIDE TOTALS"
## [207] "Chewelah Police Department"
## [208] "Colville Police Department"
## [209] "Kettle Falls Police Department"
## [210] "Springdale Police Department"
## [211] "Stevens County Sheriffs Office"
## [212] "Lacey Police Department"
## [213] "Olympia Police Department"
## [214] "Rainier Police Department"
## [215] "Tenino Police Department"
## [216] "Thurston County Sheriffs Office"
## [217] "Tumwater Police Department"
## [218] "Yelm Police Department"
## [219] "Wahkiakum County Sheriffs Office"
## [220] "College Place Police Department"
## [221] "Walla Walla County Sheriffs Office"
## [222] "Walla Walla Police Department"
## [223] "Bellingham Police Department"
## [224] "Blaine Police Department"
## [225] "Everson Police Department"
## [226] "Ferndale Police Department"
## [227] "Lynden Police Department"
## [228] "Sumas Police Department"
## [229] "Whatcom County Sheriffs Office"
## [230] "Colfax Police Department"
## [231] "Colton Police Department"
## [232] "Garfield Police Department"
## [233] "Palouse Police Department"
## [234] "Pullman Police Department"
## [235] "Rosalia Police Department"
## [236] "Whitman County Sheriffs Office"
## [237] "Grandview Police Department"
## [238] "Granger Police Department"
## [239] "Mabton Police Department"
## [240] "Moxee City Police Department"
## [241] "Selah Police Department"
## [242] "Sunnyside Police Department"
```

```
## [243] "Tieton Police Department"
## [244] "Toppenish Police Department"
## [245] "Union Gap Police Department"
## [246] "Wapato Police Department"
## [247] "Yakima County Sheriffs Office"
## [248] "Yakima Police Department"
## [249] "Zillah Police Department"
```

Notice that some of these unique locations are coming from county police departments rather than city police departments within each county. Let's remove such locations from our data set. Then lets check that the locations in our new data set are only from cities.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
location_cities_1=data_removed_NAs_new[-grep("COUNTY",data_removed_NAs_new$Location),]
location_cities=location_cities_1[-grep("County",location_cities_1$Location),]
unique(location_cities$Location)
```

```
##   [1] "Othello Police Department"
##   [2] "Ritzville Police Department"
##   [3] "Asotin Police Department"
##   [4] "Clarkston Police Department"
##   [5] "Kennewick Police Department"
##   [6] "Prosser Police Department"
##   [7] "Richland Police Department"
##   [8] "West Richland Police Department"
##   [9] "Wenatchee Police Department"
##  [10] "Forks Police Department"
##  [11] "Port Angeles Police Department"
##  [12] "Sequim Police Department"
##  [13] "Battle Ground Police Department"
##  [14] "Camas Police Department"
##  [15] "La Center Police Department"
##  [16] "Ridgefield Police Department"
##  [17] "Vancouver Police Department"
##  [18] "Washougal Police Department"
##  [19] "Castle Rock Police Department"
##  [20] "Kalama Police Department"
##  [21] "Kelso Police Department"
##  [22] "Longview Police Department"
##  [23] "Woodland Police Department"
##  [24] "East Wenatchee Police Department"
##  [25] "Republic Police Department"
##  [26] "Connell Police Department"
##  [27] "Pasco Police Department"
##  [28] "Ephrata Police Department"
```

```
##  [29] "Grand Coulee Police Department"
##  [30] "Mattawa Police Department"
##  [31] "Moses Lake Police Department"
##  [32] "Quincy Police Department"
##  [33] "Royal City Police Department"
##  [34] "Soap Lake Police Department"
##  [35] "Warden Police Department"
##  [36] "Aberdeen Police Department"
##  [37] "Cosmopolis Police Department"
##  [38] "Elma Police Department"
##  [39] "Hoquiam Police Department"
##  [40] "McCleary Police Department"
##  [41] "Montesano Police Department"
##  [42] "Oakville Police Department"
##  [43] "Ocean Shores Police Department"
##  [44] "Westport Police Department"
##  [45] "Coupeville Police Department"
##  [46] "Langley Police Department"
##  [47] "Oak Harbor Police Department"
##  [48] "Port Townsend Police Department"
##  [49] "Algona Police Department"
##  [50] "Auburn Police Department"
##  [51] "Beaux Arts Police Department"
##  [52] "Bellevue Police Department"
##  [53] "Black Diamond Police Department"
##  [54] "Bothell Police Department"
##  [55] "Burien Police Department"
##  [56] "Carnation Police Department"
##  [57] "Clyde Hill Police Department"
##  [58] "Covington Police Department"
##  [59] "Des Moines Police Department"
##  [60] "Duvall Police Department"
##  [61] "Enumclaw Police Department"
##  [62] "Federal Way Police Department"
##  [63] "Issaquah Police Department"
##  [64] "Kenmore Police Department"
##  [65] "Kent Police Department"
##  [66] "Kirkland Police Department"
##  [67] "Lake Forest Park Police Department"
##  [68] "Maple Valley Police Department"
##  [69] "Medina Police Department"
##  [70] "Mercer Island Police Department"
##  [71] "Newcastle Police Department"
##  [72] "Normandy Park Police Department"
##  [73] "North Bend Police Department"
##  [74] "Pacific Police Department"
##  [75] "Redmond Police Department"
##  [76] "Renton Police Department"
##  [77] "Sammamish Police Department"
##  [78] "Seatac Police Department"
##  [79] "Seattle Police Department"
##  [80] "Shoreline Police Department"
##  [81] "Skykomish Police Department"
##  [82] "Snoqualmie Police Department"
```

```
##  [83] "Tukwila Police Department"
##  [84] "Woodinville Police Department"
##  [85] "Yarrow Point Police Department"
##  [86] "Bainbridge Island Police Department"
##  [87] "Bremerton Police Department"
##  [88] "Kitsap Police Department"
##  [89] "Port Orchard Police Department"
##  [90] "Poulsbo Police Department"
##  [91] "Cle Elum -Roslyn Police Department"
##  [92] "Cle Elum Police Department"
##  [93] "Ellensburg Police Department"
##  [94] "Kittitas Police Department"
##  [95] "Bingen Police Department"
##  [96] "Goldendale Police Department"
##  [97] "White Salmon Police Department"
##  [98] "Centralia Police Department"
##  [99] "Chehalis Police Department"
## [100] "Morton Police Department"
## [101] "Mossyrock Police Department"
## [102] "Napavine Police Department"
## [103] "Pe Ell Police Department"
## [104] "Toledo Police Department"
## [105] "Vader Police Department"
## [106] "Winlock Police Department"
## [107] "Odessa Police Department"
## [108] "Reardan Police Department"
## [109] "Wilbur Police Department"
## [110] "Shelton Police Department"
## [111] "Brewster Police Department"
## [112] "Coulee Dam Police Department"
## [113] "Omak Police Department"
## [114] "Oroville Police Department"
## [115] "Tonasket Police Department"
## [116] "Twisp Police Department"
## [117] "Winthrop Marshal"
## [118] "Winthrop Police Department"
## [119] "Ilwaco Police Department"
## [120] "Long Beach Police Department"
## [121] "Raymond Police Department"
## [122] "South Bend Police Department"
## [123] "Newport Police Department"
## [124] "Bonney Lake Police Department"
## [125] "Buckley Police Department"
## [126] "DuPont Police Department"
## [127] "Dupont Police Department"
## [128] "Eatonville Police Department"
## [129] "Edgewood Police Department"
## [130] "Fife Police Department"
## [131] "Fircrest Police Department"
## [132] "Gig Harbor Police Department"
## [133] "Lakewood Police Department"
## [134] "Milton Police Department"
## [135] "Orting Police Department"
## [136] "Puyallup Police Department"
```

```
## [137] "Roy Police Department"
## [138] "Ruston Police Department"
## [139] "Steilacoom Police Department"
## [140] "Sumner Police Department"
## [141] "Tacoma Police Department"
## [142] "University Place Police Department"
## [143] "San Juan Police Department"
## [144] "Anacortes Police Department"
## [145] "Burlington Police Department"
## [146] "Mount Vernon Police Department"
## [147] "Sedro Woolley Police Department"
## [148] "Arilington Police Department"
## [149] "Arlington Police Department"
## [150] "Brier Police Department"
## [151] "Darrington Police Department"
## [152] "Edmonds Police Department"
## [153] "Everett Police Department"
## [154] "Gold Bar Police Department"
## [155] "Granite Falls Police Department"
## [156] "Index Police Department"
## [157] "Lake Stevens Police Department"
## [158] "Lynnwood Police Department"
## [159] "Marysville Police Department"
## [160] "Mill Creek Police Department"
## [161] "Monroe Police Department"
## [162] "Mountlake Terrace Police Department"
## [163] "Mukilteo Police Department"
## [164] "Snohomish Police Department"
## [165] "Stanwood Police Department"
## [166] "Sultan Police Department"
## [167] "Woodway Police Department"
## [168] "Airway Heights Police Department"
## [169] "Cheney Police Department"
## [170] "Liberty Lake Police Department"
## [171] "Spokane Police Department"
## [172] "Spokane Valley Police Department"
## [173] "STATEWIDE TOTALS"
## [174] "Chewelah Police Department"
## [175] "Colville Police Department"
## [176] "Kettle Falls Police Department"
## [177] "Springdale Police Department"
## [178] "Lacey Police Department"
## [179] "Olympia Police Department"
## [180] "Rainier Police Department"
## [181] "Tenino Police Department"
## [182] "Tumwater Police Department"
## [183] "Yelm Police Department"
## [184] "College Place Police Department"
## [185] "Walla Walla Police Department"
## [186] "Bellingham Police Department"
## [187] "Blaine Police Department"
## [188] "Everson Police Department"
## [189] "Ferndale Police Department"
## [190] "Lynden Police Department"
```

```
## [191] "Sumas Police Department"
## [192] "Colfax Police Department"
## [193] "Colton Police Department"
## [194] "Garfield Police Department"
## [195] "Palouse Police Department"
## [196] "Pullman Police Department"
## [197] "Rosalia Police Department"
## [198] "Grandview Police Department"
## [199] "Granger Police Department"
## [200] "Mabton Police Department"
## [201] "Moxee City Police Department"
## [202] "Selah Police Department"
## [203] "Sunnyside Police Department"
## [204] "Tieton Police Department"
## [205] "Toppenish Police Department"
## [206] "Union Gap Police Department"
## [207] "Wapato Police Department"
## [208] "Yakima Police Department"
## [209] "Zillah Police Department"
```

We can now define subsets of data coming from each of these cities below. As an example, observe that there are only 11 observations from Othello city in Adams County WA. Hence, we don't have enough observations to determine factors that influence high theft rates in each city alone. This is why we compare factors that influence high theft in each county instead.

```
location_cities_split = split(location_cities,location_cities$Location)
location_cities_split$`Othello Police Department`
```

```
## # A tibble: 11 x 34
##    IndexYear county Location    Population Total  Rate PrsnTotal PrsnRate Murder
##    <chr>     <chr>  <chr>            <dbl> <dbl> <dbl>     <dbl>    <dbl>  <dbl>
##  1 2012      ADAMS  Othello Po~       7495  1168 156.        175     23.3      0
##  2 2013      ADAMS  Othello Po~       7565  1142 151.        213     28.2      1
##  3 2014      ADAMS  Othello Po~       7695  1161 151.        241     31.3      0
##  4 2015      ADAMS  Othello Po~       7780   757  97.3       175     22.5      0
##  5 2016      ADAMS  Othello Po~       7875   649  82.4       185     23.5      0
##  6 2017      ADAMS  Othello Po~       8175   623  76.2       163     19.9      0
##  7 2018      ADAMS  Othello Po~       8270   570  68.9       150     18.1      0
##  8 2019      ADAMS  Othello Po~       8345   521  62.4       139     16.7      0
##  9 2020      ADAMS  Othello Po~       8515   516  60.6        92     10.8      0
## 10 2021      ADAMS  Othello Po~       8715   639  73.3       134     15.4      1
## 11 2022      ADAMS  Othello Po~       8920   826  92.6       162     18.2      0
## # i 25 more variables: Manslaughter <dbl>, Forcible_Sex <dbl>, Assault <dbl>,
## #   Non_Forcible_Sex <dbl>, Kidnapping_Abduction <dbl>,
## #   Human_Trafficking <dbl>, Viol_Of_No_Contact <dbl>, PrprtyTotal <dbl>,
## #   PrprtyRate <dbl>, Arson <dbl>, Bribery <dbl>, Burglary <dbl>,
## #   Counterfeiting_Forgery <dbl>, Destruction_of_Property <dbl>,
## #   Extortion_Blackmail <dbl>, Robbery <dbl>, Theft <dbl>, SctyTotal <dbl>,
## #   SctyRate <dbl>, Drug_Violations <dbl>, Gambling_Violations <dbl>, ...
```

Let's instead determine the counts of observations from each county in our data.

```
counties=split(data_removed_NAs_new,data_removed_NAs_new$county) #split data
#observations by county
counties$`ADAMS` #observe one such county subset of 44 observations
```

```
## # A tibble: 44 x 34
##    IndexYear county Location    Population Total  Rate PrsnTotal PrsnRate Murder
##    <chr>     <chr>  <chr>            <dbl> <dbl> <dbl>     <dbl>    <dbl>  <dbl>
##  1 2012      ADAMS  Adams Coun~       9860   620  62.9       145     14.7      0
##  2 2013      ADAMS  Adams Coun~       9935   693  69.8       130     13.1      2
##  3 2014      ADAMS  Adams Coun~      10025   688  68.6       165     16.5      1
##  4 2015      ADAMS  Adams Coun~       9960   685  68.8       139     14.0      0
##  5 2016      ADAMS  Adams Coun~       9975   571  57.2       166     16.6      1
##  6 2017      ADAMS  Adams Coun~      10035   498  49.6       133     13.3      1
##  7 2018      ADAMS  Adams Coun~      10090   456  45.2       142     14.1      0
##  8 2019      ADAMS  Adams Coun~      10145   389  38.3       118     11.6      0
##  9 2020      ADAMS  Adams Coun~      10275   438  42.6       129     12.6      0
## 10 2021      ADAMS  Adams Coun~      10410   677  65.0       181     17.4      2
## # i 34 more rows
## # i 25 more variables: Manslaughter <dbl>, Forcible_Sex <dbl>, Assault <dbl>,
## #   Non_Forcible_Sex <dbl>, Kidnapping_Abduction <dbl>,
## #   Human_Trafficking <dbl>, Viol_Of_No_Contact <dbl>, PrprtyTotal <dbl>,
## #   PrprtyRate <dbl>, Arson <dbl>, Bribery <dbl>, Burglary <dbl>,
## #   Counterfeiting_Forgery <dbl>, Destruction_of_Property <dbl>,
## #   Extortion_Blackmail <dbl>, Robbery <dbl>, Theft <dbl>, SctyTotal <dbl>, ...
```

```
frequencies=sapply(counties,function(counties) nrow(counties)) #obtain number of observations from each
frequencies
```

```
##        ADAMS       ASOTIN       BENTON       CHELAN      CLALLAM        CLARK
##           44           44           65           32           55           88
##     COLUMBIA      COWLITZ      DOUGLAS        FERRY     FRANKLIN     GARFIELD
##           22           77           33           26           44           22
##        GRANT GRAYS HARBOR       ISLAND    JEFFERSON         KING       KITSAP
##          106          110           41           33          340           66
##     KITTITAS    KLICKITAT        LEWIS      LINCOLN        MASON     OKANOGAN
##           51           54           98           44           33           92
##      PACIFIC PEND OREILLE       PIERCE     SAN JUAN       SKAGIT     SKAMANIA
##           66           26          217           22           66           22
##    SNOHOMISH      SPOKANE    STATEWIDE      STEVENS     THURSTON    WAHKIAKUM
##          188           65           11           58           70           22
##  WALLA WALLA      WHATCOM      WHITMAN       YAKIMA
##           43           86           57          136
```

We want to compare theft rates in counties with sufficiently large enough observations. Otherwise, our logistic regression functions won't have enough data to learn the patterns of yes/no responses on, as discussed previously in the Logistic Regression section (see above). For this reason, we choose to compare counties with at least 100 observations each: Grant, Grays Harbor, King, Pierce, and Yakima county.

Build Our Logistic Regression Models

Now that we have our response variable (large_theft_rate), lets re-define our data by adding this response variable to it.

```
data_new=cbind(data_removed_NAs_new,large_theft_rate)
```

Now lets define subsets of data for the five counties we are comparing theft the presence of high theft rates in.

```
data_new_split=split(data_new,data_new$county)
Grant_obs=data_new_split$`GRANT`
GraysHarbor_obs=data_new_split$`GRAYS HARBOR`
```

```
Pierce_obs=data_new_split$`PIERCE`
King_obs=data_new_split$`KING`
Yakima_obs=data_new_split$`YAKIMA`
```

In order for our models to learn patterns of yes/no responses to large_theft_rate, we must split up each of these data sets into training and testing sets. The training set is for our model to learn the patterns, and the test set is for our model to make predictions of future yes/no responses (i.e. if there will be greater than 3% of theft committed in a given county in a given year). We choose to split our data set into 75% of samples for training our models, and the rest for testing our models.

```
#library(readr)
#install.packages("tidymodels")
#library(tidymodels)

#GRANT COUNTY MODEL
#set.seed(300)
#split=initial_split(Grant_obs,prop=0.75,strata=Grant_obs$large_theft_rate)
#train=split %>%
 #   training()
#test=split %>%
 #   testing()

#reference: https://www.datacamp.com/tutorial/logistic-regression-R
```