Housing & Residential Safety: The Impact of Crime on Property Value and Taxes in Pierce County

Preslav Angelov, Huy Le, and Evin Tolentino

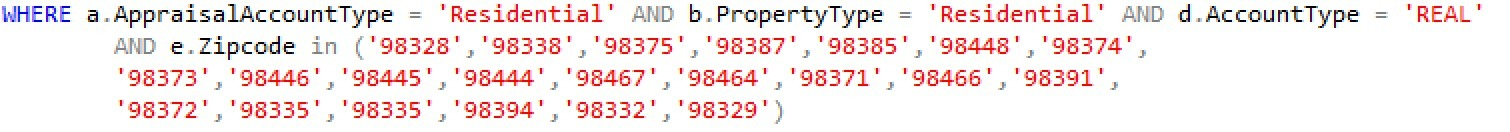
14 July 2019

BUAN 5510: Capstone Project in Business Analytics

**Data Preprocessing:**

Null Values

The null value ratio for the columns that we chose to integrate is extremely low, as shown in the data dictionary in the Appendix. However, we took further steps to ensure we eliminated any possible null values in some of the more critical columns that we ended up choosing. For example, any null values in the AppraisalAccountType column are eliminated by filtering for only rows having ‘Residential’ values with our WHERE clause. The same principle was applied to the PropertyType and AccountType columns. Furthermore, we chose to focus our research only on specific zip codes due to the availability of our crime data. This naturally took out any null values from our Zipcode column because we used a WHERE clause for a range of given Zipcodes. The full SQL query can be found in the Appendix, but this is an excerpt of the WHERE statement:



Data Integration

The first step to integrate the data was to see which tables from the original dataset will actually be useful for our research. We looked at each table and picked out the columns that would be most relevant to our problem statement. The table below summarizes all the tables that we integrated, along with the columns names that we took from each of them.

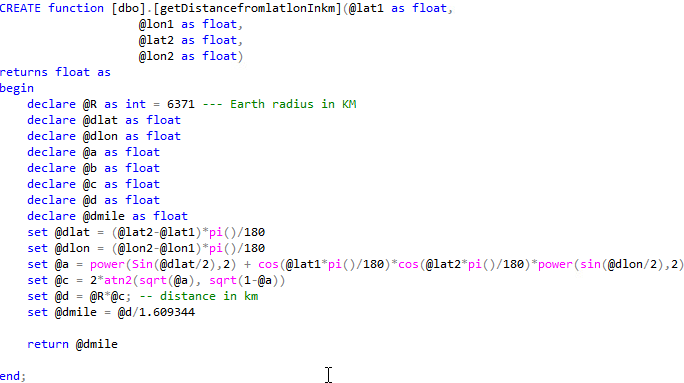


Since ParcelNumber was the unique identification of each property in the majority of the tables, we used a LEFT JOIN query to join the tables on the Parcel Number. However, CrimeData does not have a ParcelNumber; in fact, that dataset has no property identification number. It only has the x and y coordinates to identify where the crime occurred. For this reason, we created a function (see full code in Appendix) which takes the longitude and latitude data from the house and crime datasets and produces a calculated distance value (in miles). The distance values tell us how far away from the ParcelNumber a specific crime occurred.

Data Integration Query



Coordinate Distance Function



Correlation Analysis

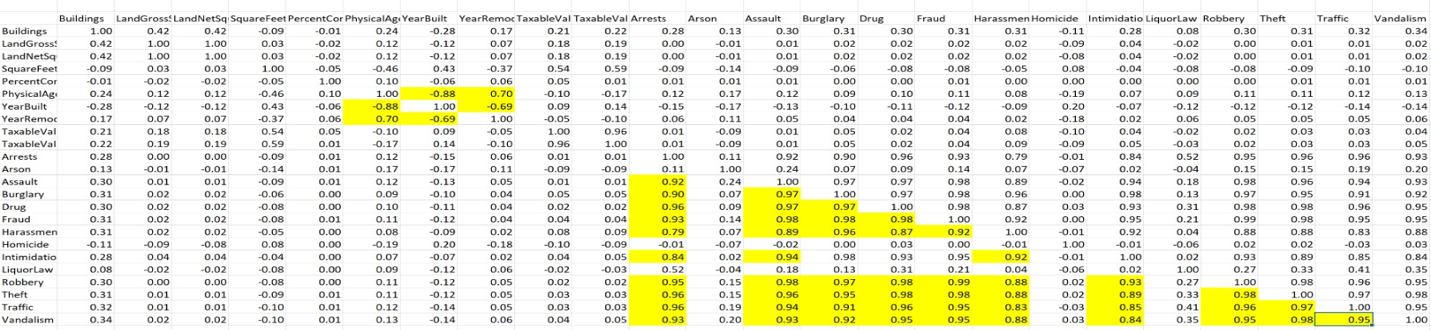
To determine which attributes in the dataset are correlated among themselves we conduct an analysis in R. After integrating the compiled data with Parcels and crime information in the previous step, we only select those crimes which occurred within the range 2 miles from the house location. In this part, we only use sample of 10% parcel data and crime to conduct the analysis.

We grouped the crime into 14 categories and aggregate the total number of crimes occurred within the house’s range. We divide the variables into two type for analyzing correlation:

* Norminal variables: Condition, and Quality
* Numerical Variables: Buildings, LandGrossSquareFeet ,LandNetSquareFeet, SquareFeet, PercentComplete, PhysicalAge, YearBuilt, YearRemodeled, TaxableValuePriorYear, TaxableValueCurrentYear, Arrests, Arson, Assault, Burglary, Drug, Fraud, Harassment, Homicide, Intimidation, LiquorLaw, Robbery, Theft, Traffic, Vandalism.

We used function cor() for calculating the correlation coefficient and cor.test() for performing correlation test. The results show that Condition and Quality have a high correlation coefficient. Among the House’s attribute only PhysicalAge, YearBuilt and YearRemodeled are highly correlated with each other. Another interesting finding is that type of crimes has a high correlation coefficient among themselves (see full correlation matrix in appendix). Taxable value of house current year is correlated with crimes except Arrest and Assault. Taxable value of house prior year is also correlated with crime except Arrest, Assault, and Drug. However, the correlation coefficients are quite low.

Correlation matrix:



**Appendices**

Exploratory Data Analysis (R Code)

> #eda of appraisal account table

> summary(appraisal)

ParcelNumber AppraisalAccountType

Min. :1.901e+07 Residential:286531

1st Qu.:6.153e+08 Commercial : 16252

Median :5.000e+09 Mobile Home: 12512

Mean :4.313e+09 Condominium: 5287

3rd Qu.:6.440e+09 Reference : 3155

Max. :9.900e+09 Industrial : 3133

(Other) : 4472

BusinessName ValueAreaID

SANS SOUCI MHP : 253 PI0: 71

LIPOMA FIRS NORTH PDD PH 2 : 213 PI1:40022

BOATHOUSE TACOMA YACHT CLUB BLDG ONLY: 206 PI2:67341

CRYSTAL POINTE CONDO : 206 PI3:58201

NORTH TAPPS ESTATES : 202 PI4:58902

(Other) : 40761 PI5:47707

NA's :289501 PI6:59098

LandEconomicArea Buildings GroupAccountNumber

070201 : 6403 Min. : 0.000 Min. : 2

120605 : 5867 1st Qu.: 1.000 1st Qu.: 3941

070101 : 5070 Median : 1.000 Median : 32600

090601 : 4709 Mean : 1.045 Mean : 8247579

060701 : 3985 3rd Qu.: 1.000 3rd Qu.: 64589

(Other):305298 Max. :110.000 Max. :830209752

NA's : 10 NA's :2171 NA's :296854

LandGrossAcres LandNetAcres LandGrossSquareFeet

Min. : 0.0 Min. : 0.000 Min. :0.000e+00

1st Qu.: 0.1 1st Qu.: 0.147 1st Qu.:6.020e+03

Median : 0.2 Median : 0.239 Median :1.029e+04

Mean : 9.5 Mean : 2.333 Mean :4.148e+05

3rd Qu.: 0.6 3rd Qu.: 0.530 3rd Qu.:2.469e+04

Max. :2242517.0 Max. :3033.400 Max. :9.768e+10

NA's :16966

LandNetSquareFeet LandGrossFrontFeet LandWidth

Min. : 0 Min. : 0.00 Min. : 0.000

1st Qu.: 6404 1st Qu.: 0.00 1st Qu.: 0.000

Median : 10405 Median : 0.00 Median : 0.000

Mean : 101615 Mean : 28.06 Mean : 5.438

3rd Qu.: 23087 3rd Qu.: 0.00 3rd Qu.: 0.000

Max. :132134904 Max. :144330.00 Max. :23232.000

NA's :16966 NA's :1642

LandDepth SubmergedAreaSqrFeet AppraisalDate

Min. : 0.000 Min. : 0 Min. :2001-09-12

1st Qu.: 0.000 1st Qu.: 0 1st Qu.:2015-04-02

Median : 0.000 Median : 0 Median :2017-03-20

Mean : 8.254 Mean : 1258 Mean :2016-10-09

3rd Qu.: 0.000 3rd Qu.: 0 3rd Qu.:2018-05-23

Max. :3200.000 Max. :10545254 Max. :3016-10-11

NA's :275920 NA's :1296

WaterfrontType ViewQuality UtilityElectric

WF Lake : 4682 View Lim : 5204 POWER AVAILABLE : 27556

WF River : 1128 View Lim -: 3997 POWER INSTALLED :300932

WF Salt : 5257 View Avg : 3966 POWER NO - COMMENT: 2854

WF Stream/Creek: 3207 View Good : 2541

NA's :317068 View Lim +: 2484

(Other) : 2280

NA's :310870

UtilitySewer UtilityWater

SEWER/SEPTIC AVAIL : 7247 WATER AVAILABLE: 19234

SEWER/SEPTIC INSTALLED:301889 WATER INSTALLED:299673

SEWER/SEPTIC NO : 21898 WATER NO : 12435

SEWER/SEPTIC NO PERC : 308

StreetType Latitude Longtitude

PAVED :307896 Min. :46.74 Min. :-122.8

STREET NO ROAD: 2527 1st Qu.:47.12 1st Qu.:-122.5

STREET UNPAVED: 20919 Median :47.18 Median :-122.4

Mean :47.18 Mean :-122.4

3rd Qu.:47.24 3rd Qu.:-122.3

Max. :47.40 Max. :-121.5

NA's :1414 NA's :1414

> names(appraisal)

[1] "ParcelNumber" "AppraisalAccountType" "BusinessName"

[4] "ValueAreaID" "LandEconomicArea" "Buildings"

[7] "GroupAccountNumber" "LandGrossAcres" "LandNetAcres"

[10] "LandGrossSquareFeet" "LandNetSquareFeet" "LandGrossFrontFeet"

[13] "LandWidth" "LandDepth" "SubmergedAreaSqrFeet"

[16] "AppraisalDate" "WaterfrontType" "ViewQuality"

[19] "UtilityElectric" "UtilitySewer" "UtilityWater"

[22] "StreetType" "Latitude" "Longtitude"

> str(appraisal)

'data.frame': 331342 obs. of 24 variables:

$ ParcelNumber : num 19012000 19014000 19021000 19121002 19121003 ...

$ AppraisalAccountType: Factor w/ 11 levels "Com Condo","Com Leasehold",..: 10 10 10 10 10 10 10 10 10 10 ...

$ BusinessName : Factor w/ 11088 levels "","''","-","--",..: NA NA NA NA NA NA NA NA NA NA ...

$ ValueAreaID : Factor w/ 7 levels "PI0","PI1","PI2",..: 2 6 2 6 6 6 6 6 6 6 ...

$ LandEconomicArea : Factor w/ 441 levels "000001","010201",..: 134 218 134 218 218 218 218 218 218 218 ...

$ Buildings : int 0 0 0 1 1 0 1 0 0 1 ...

$ GroupAccountNumber : int NA NA NA NA NA NA NA NA NA NA ...

$ LandGrossAcres : num 1.48 1.78 72.66 3.17 1.96 ...

$ LandNetAcres : num 1.48 1.78 72.66 3.17 1.96 ...

$ LandGrossSquareFeet : num 64386 77464 3165070 138085 85170 ...

$ LandNetSquareFeet : num 64386 77464 3165070 138085 85170 ...

$ LandGrossFrontFeet : num 200 200 3470 275 235 225 330 220 120 630 ...

$ LandWidth : int 200 200 3470 275 235 225 330 220 130 630 ...

$ LandDepth : int 90 100 590 490 350 155 750 595 700 1000 ...

$ SubmergedAreaSqrFeet: int NA NA NA NA NA NA NA NA NA NA ...

$ AppraisalDate : Date, format: "2014-08-11" "2018-10-04" ...

$ WaterfrontType : Factor w/ 4 levels "WF Lake","WF River",..: 3 3 3 3 3 3 3 3 3 3 ...

$ ViewQuality : Factor w/ 9 levels "View Avg","View Avg +",..: NA NA NA NA NA NA NA NA NA NA ...

$ UtilityElectric : Factor w/ 3 levels "POWER AVAILABLE",..: 3 3 3 2 1 1 2 1 1 2 ...

$ UtilitySewer : Factor w/ 4 levels "SEWER/SEPTIC AVAIL",..: 3 3 3 2 3 3 2 3 3 3 ...

$ UtilityWater : Factor w/ 3 levels "WATER AVAILABLE",..: 3 3 3 2 2 3 2 3 3 2 ...

$ StreetType : Factor w/ 3 levels "PAVED","STREET NO ROAD",..: 2 2 2 3 3 3 3 3 3 3 ...

$ Latitude : num 47.2 47.2 47.2 47.2 47.1 ...

$ Longtitude : num -123 -123 -123 -123 -123 ...

>

> #correlation matrix for appraisal account numerical variables

> appraisal\_num <- appraisal[,c(1,6,7,8,9,10,11,12,13,14,15,23,24)]

> cor(appraisal\_num)

ParcelNumber Buildings GroupAccountNumber

ParcelNumber 1.000000000 NA NA

Buildings NA 1 NA

GroupAccountNumber NA NA 1

LandGrossAcres -0.002990989 NA NA

LandNetAcres NA NA NA

LandGrossSquareFeet -0.002990999 NA NA

LandNetSquareFeet NA NA NA

LandGrossFrontFeet -0.015336243 NA NA

LandWidth NA NA NA

LandDepth -0.095933413 NA NA

SubmergedAreaSqrFeet NA NA NA

Latitude NA NA NA

Longtitude NA NA NA

LandGrossAcres LandNetAcres LandGrossSquareFeet

ParcelNumber -2.990989e-03 NA -2.990999e-03

Buildings NA NA NA

GroupAccountNumber NA NA NA

LandGrossAcres 1.000000e+00 NA 1.000000e+00

LandNetAcres NA 1 NA

LandGrossSquareFeet 1.000000e+00 NA 1.000000e+00

LandNetSquareFeet NA NA NA

LandGrossFrontFeet 1.450425e-05 NA 1.450317e-05

LandWidth NA NA NA

LandDepth -1.058789e-04 NA -1.058802e-04

SubmergedAreaSqrFeet NA NA NA

Latitude NA NA NA

Longtitude NA NA NA

LandNetSquareFeet LandGrossFrontFeet LandWidth

ParcelNumber NA -1.533624e-02 NA

Buildings NA NA NA

GroupAccountNumber NA NA NA

LandGrossAcres NA 1.450425e-05 NA

LandNetAcres NA NA NA

LandGrossSquareFeet NA 1.450317e-05 NA

LandNetSquareFeet 1 NA NA

LandGrossFrontFeet NA 1.000000e+00 NA

LandWidth NA NA 1

LandDepth NA 5.124467e-02 NA

SubmergedAreaSqrFeet NA NA NA

Latitude NA NA NA

Longtitude NA NA NA

LandDepth SubmergedAreaSqrFeet Latitude

ParcelNumber -0.0959334133 NA NA

Buildings NA NA NA

GroupAccountNumber NA NA NA

LandGrossAcres -0.0001058789 NA NA

LandNetAcres NA NA NA

LandGrossSquareFeet -0.0001058802 NA NA

LandNetSquareFeet NA NA NA

LandGrossFrontFeet 0.0512446686 NA NA

LandWidth NA NA NA

LandDepth 1.0000000000 NA NA

SubmergedAreaSqrFeet NA 1 NA

Latitude NA NA 1

Longtitude NA NA NA

Longtitude

ParcelNumber NA

Buildings NA

GroupAccountNumber NA

LandGrossAcres NA

LandNetAcres NA

LandGrossSquareFeet NA

LandNetSquareFeet NA

LandGrossFrontFeet NA

LandWidth NA

LandDepth NA

SubmergedAreaSqrFeet NA

Latitude NA

Longtitude 1

>

> #eda of improvement table

> summary(improvement)

ParcelNumber BuildingID PropertyType

Min. :1.912e+07 Min. : 0.000 Residential :450850

1st Qu.:5.194e+08 1st Qu.: 1.000 Out Building: 98626

Median :4.720e+09 Median : 1.000 Commercial : 57614

Mean :4.262e+09 Mean : 1.387 Mobile Home : 46200

3rd Qu.:6.490e+09 3rd Qu.: 1.000 Townhouse : 16718

Max. :9.900e+09 Max. :822.000 Duplex : 10498

(Other) : 7388

Neighborhood NeighborhoodExtension SquareFeet

070201 : 13158 0 :607822 Min. : 1

120605 : 11516 862 : 7330 1st Qu.: 1046

070101 : 11220 P2 : 6524 Median : 1540

090601 : 10358 PL1 : 6160 Mean : 2266

160702 : 9916 852 : 5392 3rd Qu.: 2130

(Other):631724 (Other): 54664 Max. :1078155

NA's : 2 NA's : 2

NetSquareFeet PercentComplete Condition

Min. : 0.0 Min. :0.0000 Average :642742

1st Qu.: 0.0 1st Qu.:1.0000 Fair : 29298

Median : 0.0 Median :1.0000 Poor : 8188

Mean : 843.2 Mean :0.9978 Very Poor : 3352

3rd Qu.: 0.0 3rd Qu.:1.0000 Extra Poor: 2300

Max. :1523580.0 Max. :1.0040 (Other) : 1372

NA's :6 NA's : 642

Quality PrimaryOccupancyCode

Average :329520 Min. : 26.0

Fair Plus :112090 1st Qu.: 100.0

Fair : 97712 Median : 100.0

Average Plus: 59246 Mean : 174.7

Low : 33408 3rd Qu.: 100.0

(Other) : 55286 Max. :9997.0

NA's : 632 NA's :4194

PrimaryOccupancyDescription MobileHomeSerialNumber

Single Family Residential :446730 0 : 1524

Detached Garage : 69286 FO : 10

Mobile or Manufactured Home: 46212 \* : 6

Storage - Material : 13990 0213 : 6

Townhouse/Condo : 10960 0340 : 6

(Other) : 96522 (Other): 44064

NA's : 4194 NA's :642278

MobileHomeTotalLength MobileHomeMake AtticFinishedSquareFeet

Min. : 0.0 FLTWD : 9068 Min. : 36

1st Qu.:48.0 MODUL : 5028 1st Qu.: 280

Median :56.0 SKY : 3414 Median : 362

Mean :55.6 LIBER : 3412 Mean : 395

3rd Qu.:64.0 REDMN : 2874 3rd Qu.: 472

Max. :84.0 (Other): 21696 Max. :3200

NA's :642368 NA's :642402 NA's :652014

BasementSquareFeet BasementFinishedSquareFeet CarportSquareFeet

Min. : 49 Min. : 1 Min. : 56

1st Qu.: 720 1st Qu.: 656 1st Qu.: 276

Median : 1008 Median : 970 Median : 391

Mean : 1068 Mean : 1160 Mean : 414

3rd Qu.: 1336 3rd Qu.: 1327 3rd Qu.: 510

Max. :38230 Max. :77597 Max. :1224

NA's :602104 NA's :634126 NA's :686492

BalconySquareFeet PorchSquareFeet AttachedGarageSquareFeet

Min. : 30.0 Min. : 0.0 Min. : 1

1st Qu.: 78.0 1st Qu.: 160.0 1st Qu.: 420

Median : 96.0 Median : 342.0 Median : 484

Mean : 267.4 Mean : 477.4 Mean : 532

3rd Qu.: 226.0 3rd Qu.: 621.0 3rd Qu.: 610

Max. :17609.0 Max. :13468.0 Max. :6896

NA's :685044 NA's :263164 NA's :371498

DetachedGarageSquareFeet Fireplaces BasementGarageDoor

Min. : 112.0 Min. :0.00 Min. :-502.0

1st Qu.: 420.0 1st Qu.:1.00 1st Qu.: 1.0

Median : 576.0 Median :1.00 Median : 1.0

Mean : 642.5 Mean :1.14 Mean : 1.5

3rd Qu.: 768.0 3rd Qu.:1.00 3rd Qu.: 2.0

Max. :4608.0 Max. :8.00 Max. : 760.0

NA's :650802 NA's :293072 NA's :671340

> names(improvement)

[1] "ParcelNumber" "BuildingID"

[3] "PropertyType" "Neighborhood"

[5] "NeighborhoodExtension" "SquareFeet"

[7] "NetSquareFeet" "PercentComplete"

[9] "Condition" "Quality"

[11] "PrimaryOccupancyCode" "PrimaryOccupancyDescription"

[13] "MobileHomeSerialNumber" "MobileHomeTotalLength"

[15] "MobileHomeMake" "AtticFinishedSquareFeet"

[17] "BasementSquareFeet" "BasementFinishedSquareFeet"

[19] "CarportSquareFeet" "BalconySquareFeet"

[21] "PorchSquareFeet" "AttachedGarageSquareFeet"

[23] "DetachedGarageSquareFeet" "Fireplaces"

[25] "BasementGarageDoor"

> str(improvement)

'data.frame': 687894 obs. of 25 variables:

$ ParcelNumber : num 7e+09 7e+09 7e+09 7e+09 7e+09 ...

$ BuildingID : int 1 1 1 1 1 1 1 1 1 1 ...

$ PropertyType : Factor w/ 9 levels "Commercial","Duplex",..: 7 7 7 7 7 7 7 7 7 7 ...

$ Neighborhood : Factor w/ 301 levels "010201","010202",..: 65 65 65 65 65 65 65 65 65 65 ...

$ NeighborhoodExtension : Factor w/ 57 levels "0","304","338",..: 1 1 1 1 1 1 1 1 1 1 ...

$ SquareFeet : int 1092 1430 1220 1430 1220 1430 1430 1430 1226 1142 ...

$ NetSquareFeet : int 0 0 0 0 0 0 0 0 0 0 ...

$ PercentComplete : num 1 1 1 1 1 1 1 1 1 1 ...

$ Condition : Factor w/ 9 levels "Average","Avg",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Quality : Factor w/ 11 levels "Average","Average Plus",..: 1 1 1 1 1 1 1 1 1 1 ...

$ PrimaryOccupancyCode : int 100 100 100 100 100 100 100 100 100 100 ...

$ PrimaryOccupancyDescription: Factor w/ 189 levels "Addon Only Comm",..: 165 165 165 165 165 165 165 165 165 165 ...

$ MobileHomeSerialNumber : Factor w/ 21585 levels " 02830364M"," 06910205V",..: NA NA NA NA NA NA NA NA NA NA ...

$ MobileHomeTotalLength : int NA NA NA NA NA NA NA NA NA NA ...

$ MobileHomeMake : Factor w/ 183 levels "ABC","ABCSU",..: NA NA NA NA NA NA NA NA NA NA ...

$ AtticFinishedSquareFeet : int NA NA NA NA NA NA NA NA NA NA ...

$ BasementSquareFeet : int NA NA NA NA NA NA NA NA NA NA ...

$ BasementFinishedSquareFeet : int NA NA NA NA NA NA NA NA NA NA ...

$ CarportSquareFeet : int NA NA NA NA NA NA NA NA NA NA ...

$ BalconySquareFeet : int NA NA NA NA NA NA NA NA NA NA ...

$ PorchSquareFeet : int NA 440 NA 98 NA NA 95 95 NA NA ...

$ AttachedGarageSquareFeet : int 400 400 508 400 508 400 400 400 502 530 ...

$ DetachedGarageSquareFeet : int NA NA NA NA NA NA NA NA NA NA ...

$ Fireplaces : int 1 1 1 1 1 1 1 1 1 1 ...

$ BasementGarageDoor : int NA NA NA NA NA NA NA NA NA NA ...

>

> #correlation matrix for improvement numerical variables

> improvement\_num <- improvement[,c(1,2,6,7,8,11,14,16,17,18,19,20,21,22,23,24,25)]

> cor(improvement\_num)

ParcelNumber BuildingID SquareFeet

ParcelNumber 1.00000000 -0.057913113 -0.04903733

BuildingID -0.05791311 1.000000000 0.03403978

SquareFeet -0.04903733 0.034039776 1.00000000

NetSquareFeet NA NA NA

PercentComplete 0.01057825 -0.001110265 -0.03173867

PrimaryOccupancyCode NA NA NA

MobileHomeTotalLength NA NA NA

AtticFinishedSquareFeet NA NA NA

BasementSquareFeet NA NA NA

BasementFinishedSquareFeet NA NA NA

CarportSquareFeet NA NA NA

BalconySquareFeet NA NA NA

PorchSquareFeet NA NA NA

AttachedGarageSquareFeet NA NA NA

DetachedGarageSquareFeet NA NA NA

Fireplaces NA NA NA

BasementGarageDoor NA NA NA

NetSquareFeet PercentComplete

ParcelNumber NA 0.010578246

BuildingID NA -0.001110265

SquareFeet NA -0.031738666

NetSquareFeet 1 NA

PercentComplete NA 1.000000000

PrimaryOccupancyCode NA NA

MobileHomeTotalLength NA NA

AtticFinishedSquareFeet NA NA

BasementSquareFeet NA NA

BasementFinishedSquareFeet NA NA

CarportSquareFeet NA NA

BalconySquareFeet NA NA

PorchSquareFeet NA NA

AttachedGarageSquareFeet NA NA

DetachedGarageSquareFeet NA NA

Fireplaces NA NA

BasementGarageDoor NA NA

PrimaryOccupancyCode MobileHomeTotalLength

ParcelNumber NA NA

BuildingID NA NA

SquareFeet NA NA

NetSquareFeet NA NA

PercentComplete NA NA

PrimaryOccupancyCode 1 NA

MobileHomeTotalLength NA 1

AtticFinishedSquareFeet NA NA

BasementSquareFeet NA NA

BasementFinishedSquareFeet NA NA

CarportSquareFeet NA NA

BalconySquareFeet NA NA

PorchSquareFeet NA NA

AttachedGarageSquareFeet NA NA

DetachedGarageSquareFeet NA NA

Fireplaces NA NA

BasementGarageDoor NA NA

AtticFinishedSquareFeet BasementSquareFeet

ParcelNumber NA NA

BuildingID NA NA

SquareFeet NA NA

NetSquareFeet NA NA

PercentComplete NA NA

PrimaryOccupancyCode NA NA

MobileHomeTotalLength NA NA

AtticFinishedSquareFeet 1 NA

BasementSquareFeet NA 1

BasementFinishedSquareFeet NA NA

CarportSquareFeet NA NA

BalconySquareFeet NA NA

PorchSquareFeet NA NA

AttachedGarageSquareFeet NA NA

DetachedGarageSquareFeet NA NA

Fireplaces NA NA

BasementGarageDoor NA NA

BasementFinishedSquareFeet CarportSquareFeet

ParcelNumber NA NA

BuildingID NA NA

SquareFeet NA NA

NetSquareFeet NA NA

PercentComplete NA NA

PrimaryOccupancyCode NA NA

MobileHomeTotalLength NA NA

AtticFinishedSquareFeet NA NA

BasementSquareFeet NA NA

BasementFinishedSquareFeet 1 NA

CarportSquareFeet NA 1

BalconySquareFeet NA NA

PorchSquareFeet NA NA

AttachedGarageSquareFeet NA NA

DetachedGarageSquareFeet NA NA

Fireplaces NA NA

BasementGarageDoor NA NA

BalconySquareFeet PorchSquareFeet

ParcelNumber NA NA

BuildingID NA NA

SquareFeet NA NA

NetSquareFeet NA NA

PercentComplete NA NA

PrimaryOccupancyCode NA NA

MobileHomeTotalLength NA NA

AtticFinishedSquareFeet NA NA

BasementSquareFeet NA NA

BasementFinishedSquareFeet NA NA

CarportSquareFeet NA NA

BalconySquareFeet 1 NA

PorchSquareFeet NA 1

AttachedGarageSquareFeet NA NA

DetachedGarageSquareFeet NA NA

Fireplaces NA NA

BasementGarageDoor NA NA

AttachedGarageSquareFeet

ParcelNumber NA

BuildingID NA

SquareFeet NA

NetSquareFeet NA

PercentComplete NA

PrimaryOccupancyCode NA

MobileHomeTotalLength NA

AtticFinishedSquareFeet NA

BasementSquareFeet NA

BasementFinishedSquareFeet NA

CarportSquareFeet NA

BalconySquareFeet NA

PorchSquareFeet NA

AttachedGarageSquareFeet 1

DetachedGarageSquareFeet NA

Fireplaces NA

BasementGarageDoor NA

DetachedGarageSquareFeet Fireplaces

ParcelNumber NA NA

BuildingID NA NA

SquareFeet NA NA

NetSquareFeet NA NA

PercentComplete NA NA

PrimaryOccupancyCode NA NA

MobileHomeTotalLength NA NA

AtticFinishedSquareFeet NA NA

BasementSquareFeet NA NA

BasementFinishedSquareFeet NA NA

CarportSquareFeet NA NA

BalconySquareFeet NA NA

PorchSquareFeet NA NA

AttachedGarageSquareFeet NA NA

DetachedGarageSquareFeet 1 NA

Fireplaces NA 1

BasementGarageDoor NA NA

BasementGarageDoor

ParcelNumber NA

BuildingID NA

SquareFeet NA

NetSquareFeet NA

PercentComplete NA

PrimaryOccupancyCode NA

MobileHomeTotalLength NA

AtticFinishedSquareFeet NA

BasementSquareFeet NA

BasementFinishedSquareFeet NA

CarportSquareFeet NA

BalconySquareFeet NA

PorchSquareFeet NA

AttachedGarageSquareFeet NA

DetachedGarageSquareFeet NA

Fireplaces NA

BasementGarageDoor 1

> #eda of improvement builtas table

> summary(improvement\_builtas)

ParcelNumber BuildingID BuiltAsNumber BuiltAsID

Min. :1.912e+07 Min. : 0.00 Min. :1.000 Min. : 1.00

1st Qu.:5.193e+08 1st Qu.: 1.00 1st Qu.:1.000 1st Qu.: 1.00

Median :4.715e+09 Median : 1.00 Median :1.000 Median : 8.00

Mean :4.251e+09 Mean : 1.39 Mean :1.009 Mean : 70.18

3rd Qu.:6.460e+09 3rd Qu.: 1.00 3rd Qu.:1.000 3rd Qu.: 68.00

Max. :9.900e+09 Max. :822.00 Max. :7.000 Max. :1499.00

BuiltAsDescription BuiltAsSquareFeet HVAC

1 Story :221900 Min. : 0 Min. : NA

2 Story :149918 1st Qu.: 1047 1st Qu.: NA

Detached Garage: 69418 Median : 1542 Median : NA

1 1/2 Story Fin: 33882 Mean : 2248 Mean :NaN

Double Wide : 32548 3rd Qu.: 2136 3rd Qu.: NA

Split Entry : 26514 Max. :1016109 Max. : NA

(Other) :159164 NA's :693344

HVACDescription Exterior

Forced Air :352818 Frame Siding :475420

None :111796 Frame Vinyl : 53918

Electric Baseboard : 92564 Hardboard Sheet : 27454

Heat Pump : 66428 Ribbed Aluminum : 11248

Warm and Cool Air Zone: 23026 Masonry Common Brick: 8074

(Other) : 46070 (Other) : 21344

NA's : 642 NA's : 95886

Interior Stories StoryHeight

Drywall :506608 Min. : 0.000 Min. : 0.000

Paneling: 21474 1st Qu.: 1.000 1st Qu.: 8.000

NA's :165262 Median : 1.000 Median : 8.000

Mean : 1.296 Mean : 8.412

3rd Qu.: 2.000 3rd Qu.: 8.000

Max. :135.000 Max. :5010.000

NA's :14 NA's :4

SprinklerSquareFeet RoofCover Bedrooms

Min. : 0.0 Composition Shingle:437354 Min. : 0.000

1st Qu.: 0.0 Shingle Comp : 35072 1st Qu.: 1.000

Median : 0.0 Concrete Tile : 12420 Median : 3.000

Mean : 470.3 Wood Shake : 11568 Mean : 2.384

3rd Qu.: 0.0 Metal Ribbed : 11142 3rd Qu.: 3.000

Max. :1029110.0 (Other) : 19098 Max. :21.000

NA's :166690 NA's :874

Bathrooms Units ClassCode

Min. : 0.000 Min. : 0.000 A : 1204

1st Qu.: 1.000 1st Qu.: 1.000 B : 1048

Median : 1.750 Median : 1.000 C : 14850

Mean : 1.526 Mean : 1.265 D : 48154

3rd Qu.: 2.500 3rd Qu.: 1.000 P : 17088

Max. :2075.000 Max. :2319.000 S : 5192

NA's :866 NA's :12 NA's:605808

ClassDescription YearBuilt YearRemodeled

Fireproof Steel : 1204 Min. : 0 Min. : 0

Masonry : 14850 1st Qu.:1960 1st Qu.: 0

Metal Frame : 5192 Median :1982 Median :1976

Pole : 17088 Mean :1976 Mean :1237

Reinforced Concrete: 1048 3rd Qu.:1999 3rd Qu.:1991

Wood Frame : 48154 Max. :2019 Max. :2019

NA's :605808 NA's :1686

AdjustedYearBuilt PhysicalAge BuiltAsLength BuiltAsWidth

Min. : 0 Min. : 0.00 Min. : 0.000 Min. : 0.00

1st Qu.:1981 1st Qu.: 19.00 1st Qu.: 0.000 1st Qu.: 0.00

Median :1991 Median : 28.00 Median : 0.000 Median : 0.00

Mean :1987 Mean : 29.17 Mean : 3.713 Mean : 1.54

3rd Qu.:2000 3rd Qu.: 38.00 3rd Qu.: 0.000 3rd Qu.: 0.00

Max. :2019 Max. :134.00 Max. :100.000 Max. :80.00

NA's :2 NA's :2 NA's :1706 NA's :1706

MobileHomeModel

ARDMORE : 1404

FLEETWOOD : 982

GREEN HILL: 966

SKYLINE : 876

BARRINGTON: 864

(Other) : 41114

NA's :647138

> names(improvement\_builtas)

[1] "ParcelNumber" "BuildingID" "BuiltAsNumber"

[4] "BuiltAsID" "BuiltAsDescription" "BuiltAsSquareFeet"

[7] "HVAC" "HVACDescription" "Exterior"

[10] "Interior" "Stories" "StoryHeight"

[13] "SprinklerSquareFeet" "RoofCover" "Bedrooms"

[16] "Bathrooms" "Units" "ClassCode"

[19] "ClassDescription" "YearBuilt" "YearRemodeled"

[22] "AdjustedYearBuilt" "PhysicalAge" "BuiltAsLength"

[25] "BuiltAsWidth" "MobileHomeModel"

> str(improvement\_builtas)

'data.frame': 693344 obs. of 26 variables:

$ ParcelNumber : num 4.91e+09 4.91e+09 4.91e+09 4.91e+09 4.91e+09 ...

$ BuildingID : int 1 1 1 1 1 1 1 1 1 1 ...

$ BuiltAsNumber : int 1 1 1 1 1 1 1 1 1 1 ...

$ BuiltAsID : int 1 7 1 7 4 7 4 7 4 7 ...

$ BuiltAsDescription : Factor w/ 269 levels "1 1/2 Story Fin",..: 2 227 2 227 226 227 226 227 226 227 ...

$ BuiltAsSquareFeet : int 1466 1449 1424 1482 2266 1408 2197 1426 1596 1450 ...

$ HVAC : int NA NA NA NA NA NA NA NA NA NA ...

$ HVACDescription : Factor w/ 24 levels "Central Air to Air",..: 10 9 9 9 9 4 9 9 9 9 ...

$ Exterior : Factor w/ 25 levels "Cedar A-Frame",..: 10 10 10 12 10 12 10 10 10 10 ...

$ Interior : Factor w/ 2 levels "Drywall","Paneling": 1 1 1 1 1 1 1 1 1 1 ...

$ Stories : num 1 2 1 2 2 2 2 2 2 2 ...

$ StoryHeight : int 8 8 8 8 8 8 8 8 8 8 ...

$ SprinklerSquareFeet: int 0 0 0 0 0 0 0 0 0 0 ...

$ RoofCover : Factor w/ 12 levels "Built Up Rock",..: 4 4 4 4 4 5 4 4 4 4 ...

$ Bedrooms : int 3 3 3 3 3 3 4 3 3 3 ...

$ Bathrooms : num 1.75 2.25 1.75 2.25 1.75 2.25 2.25 2.5 1.75 2.25 ...

$ Units : int 1 1 1 1 1 1 1 1 1 1 ...

$ ClassCode : Factor w/ 6 levels "A","B","C","D",..: NA NA NA NA NA NA NA NA NA NA ...

$ ClassDescription : Factor w/ 6 levels "Fireproof Steel",..: NA NA NA NA NA NA NA NA NA NA ...

$ YearBuilt : int 1985 1985 1985 1987 1987 1987 1987 1987 1986 1991 ...

$ YearRemodeled : int 1990 1990 1990 1991 2002 1991 1991 1991 1991 1993 ...

$ AdjustedYearBuilt : int 1990 1990 1990 1991 1992 1991 1991 1991 1991 1993 ...

$ PhysicalAge : int 29 29 29 28 27 28 28 28 28 26 ...

$ BuiltAsLength : int 0 0 0 0 0 0 0 0 0 0 ...

$ BuiltAsWidth : int 0 0 0 0 0 0 0 0 0 0 ...

$ MobileHomeModel : Factor w/ 830 levels "ABC","ABCSU",..: NA NA NA NA NA NA NA NA NA NA ...

>

> #correlation matrix for improvement builtas numerical variables

> improvement\_builtas\_num <- improvement\_builtas[,c(1,2,3,4,6,7,11,12,13,15,16,17,20,21,22,23,24,25)]

> cor(improvement\_builtas\_num)

ParcelNumber BuildingID BuiltAsNumber BuiltAsID

ParcelNumber 1.00000000 -0.058084832 -0.039685172 -0.12890557

BuildingID -0.05808483 1.000000000 0.005755478 0.09384225

BuiltAsNumber -0.03968517 0.005755478 1.000000000 0.16246770

BuiltAsID -0.12890557 0.093842248 0.162467699 1.00000000

BuiltAsSquareFeet -0.04992151 0.035284585 0.031523644 0.15490562

HVAC NA NA NA NA

Stories NA NA NA NA

StoryHeight NA NA NA NA

SprinklerSquareFeet -0.02970332 0.036489453 0.026219848 0.13020338

Bedrooms NA NA NA NA

Bathrooms NA NA NA NA

Units NA NA NA NA

YearBuilt 0.01198000 0.008810514 0.001753829 0.07253372

YearRemodeled NA NA NA NA

AdjustedYearBuilt NA NA NA NA

PhysicalAge NA NA NA NA

BuiltAsLength NA NA NA NA

BuiltAsWidth NA NA NA NA

BuiltAsSquareFeet HVAC Stories StoryHeight

ParcelNumber -0.04992151 NA NA NA

BuildingID 0.03528459 NA NA NA

BuiltAsNumber 0.03152364 NA NA NA

BuiltAsID 0.15490562 NA NA NA

BuiltAsSquareFeet 1.00000000 NA NA NA

HVAC NA 1 NA NA

Stories NA NA 1 NA

StoryHeight NA NA NA 1

SprinklerSquareFeet 0.91746804 NA NA NA

Bedrooms NA NA NA NA

Bathrooms NA NA NA NA

Units NA NA NA NA

YearBuilt 0.05566378 NA NA NA

YearRemodeled NA NA NA NA

AdjustedYearBuilt NA NA NA NA

PhysicalAge NA NA NA NA

BuiltAsLength NA NA NA NA

BuiltAsWidth NA NA NA NA

SprinklerSquareFeet Bedrooms Bathrooms Units

ParcelNumber -0.02970332 NA NA NA

BuildingID 0.03648945 NA NA NA

BuiltAsNumber 0.02621985 NA NA NA

BuiltAsID 0.13020338 NA NA NA

BuiltAsSquareFeet 0.91746804 NA NA NA

HVAC NA NA NA NA

Stories NA NA NA NA

StoryHeight NA NA NA NA

SprinklerSquareFeet 1.00000000 NA NA NA

Bedrooms NA 1 NA NA

Bathrooms NA NA 1 NA

Units NA NA NA 1

YearBuilt 0.03347728 NA NA NA

YearRemodeled NA NA NA NA

AdjustedYearBuilt NA NA NA NA

PhysicalAge NA NA NA NA

BuiltAsLength NA NA NA NA

BuiltAsWidth NA NA NA NA

YearBuilt YearRemodeled AdjustedYearBuilt

ParcelNumber 0.011980004 NA NA

BuildingID 0.008810514 NA NA

BuiltAsNumber 0.001753829 NA NA

BuiltAsID 0.072533718 NA NA

BuiltAsSquareFeet 0.055663776 NA NA

HVAC NA NA NA

Stories NA NA NA

StoryHeight NA NA NA

SprinklerSquareFeet 0.033477284 NA NA

Bedrooms NA NA NA

Bathrooms NA NA NA

Units NA NA NA

YearBuilt 1.000000000 NA NA

YearRemodeled NA 1 NA

AdjustedYearBuilt NA NA 1

PhysicalAge NA NA NA

BuiltAsLength NA NA NA

BuiltAsWidth NA NA NA

PhysicalAge BuiltAsLength BuiltAsWidth

ParcelNumber NA NA NA

BuildingID NA NA NA

BuiltAsNumber NA NA NA

BuiltAsID NA NA NA

BuiltAsSquareFeet NA NA NA

HVAC NA NA NA

Stories NA NA NA

StoryHeight NA NA NA

SprinklerSquareFeet NA NA NA

Bedrooms NA NA NA

Bathrooms NA NA NA

Units NA NA NA

YearBuilt NA NA NA

YearRemodeled NA NA NA

AdjustedYearBuilt NA NA NA

PhysicalAge 1 NA NA

BuiltAsLength NA 1 NA

BuiltAsWidth NA NA 1

>

> #eda of improvement detail table

> summary(improvement\_detail)

ParcelNumber BuildingID DetailType

Min. :1.912e+07 Min. : 0.000 Fixture :762467

1st Qu.:2.206e+09 1st Qu.: 1.000 Appliance:464828

Median :5.001e+09 Median : 1.000 Porch :351133

Mean :4.457e+09 Mean : 1.066 Rough In :261717

3rd Qu.:6.490e+09 3rd Qu.: 1.000 Garage :186091

Max. :9.900e+09 Max. :819.000 Add On :134209

(Other) :126001

DetailDescription Units

Allowance :524127 Min. : -512

Laundry Facility:261164 1st Qu.: 1

Bath 3 Fixture :250950 Median : 1

PreFab/Stoves :121668 Mean : 325

Wood Deck :111742 3rd Qu.: 162

Attached :111254 Max. :5000000

(Other) :905541

> names(improvement\_detail)

[1] "ParcelNumber" "BuildingID" "DetailType"

[4] "DetailDescription" "Units"

> str(improvement\_detail)

'data.frame': 2286446 obs. of 5 variables:

$ ParcelNumber : num 20164001 20164001 20164001 20164001 20211007 ...

$ BuildingID : int 9 10 10 10 1 1 1 1 1 1 ...

$ DetailType : Factor w/ 12 levels "Add On","Appliance",..: 10 7 7 10 2 2 4 4 4 7 ...

$ DetailDescription: Factor w/ 420 levels "","1000 SF Cooler 32? to 60?",..: 115 15 43 115 15 169 56 162 282 15 ...

$ Units : num 72 1 2 72 1 ...

>

> #correlation matrix for improvement detail numerical variables

> improvement\_detail\_num <- improvement\_detail[,c(1,2,5)]

> cor(improvement\_detail\_num)

ParcelNumber BuildingID Units

ParcelNumber 1.00000000 -0.033518010 -0.014825142

BuildingID -0.03351801 1.000000000 0.006715357

Units -0.01482514 0.006715357 1.000000000

>

> #eda of land attribute table

> summary(land\_attribute)

ParcelNumber Attribute

Min. :1.901e+07 R AMENITIES :118740

1st Qu.:4.192e+08 R FUNCTIONAL:106061

Median :4.002e+09 R UTILITIES : 83239

Mean :3.760e+09 R ECONOMIC : 39782

3rd Qu.:6.026e+09 R STREETS : 31812

Max. :9.900e+09 R WATERFRONT: 28028

(Other) :145581

AttributeDescription

PLATTED AMENITIES : 62481

PLATTED AMENITIES LIMITED: 35205

POWER AVAILABLE : 27556

SEWER/SEPTIC NO : 21898

STREET UNPAVED : 20931

WATER AVAILABLE : 19234

(Other) :365938

> names(land\_attribute)

[1] "ParcelNumber" "Attribute" "AttributeDescription"

> str(land\_attribute)

'data.frame': 553243 obs. of 3 variables:

$ ParcelNumber : num 1.9e+07 1.9e+07 1.9e+07 1.9e+07 1.9e+07 ...

$ Attribute : Factor w/ 31 levels "C AMENITIES",..: 23 25 26 26 26 28 21 23 25 25 ...

$ AttributeDescription: Factor w/ 268 levels "1 HIGH DENSITY",..: 79 218 179 210 265 202 212 79 65 218 ...

>

> #no correlation matrix for land attribute since it only has one numerical variable

>

> #eda of sale table

> summary(sale)

ENT ParcelCount ParcelNumber

4256595 : 591 Min. : 1.000 Min. :1.901e+07

4246057 : 568 1st Qu.: 1.000 1st Qu.:2.745e+09

1054878 : 423 Median : 1.000 Median :5.004e+09

4246056 : 401 Mean : 6.973 Mean :4.680e+09

4435443 : 309 3rd Qu.: 1.000 3rd Qu.:6.825e+09

4102507C: 273 Max. :591.000 Max. :9.900e+09

(Other) :521664

SaleDate SalePrice

Min. :1997-01-01 Min. : 0

1st Qu.:2003-04-02 1st Qu.: 130000

Median :2007-05-24 Median : 207000

Mean :2008-05-15 Mean : 833401

3rd Qu.:2014-02-07 3rd Qu.: 325000

Max. :2019-06-04 Max. :368586669

DeedType

Statutory Warranty Deed :402204

Trustee Deed (Foreclosure) : 38488

Bargain & Sale Deed : 18576

Special Warranty Deed : 18184

Quit Claim Deed : 13222

Mobile Home Excise Affidavit: 13165

(Other) : 20390

Grantor

NORTHWEST TRUSTEE SERVICES INC: 11872

FEDERAL NATIONAL MORTGAGE ASSO: 4838

QUALITY LOAN SERVICE CORP OF W: 4761

SSHI LLC : 4229

SECRETARY OF HOUSING & URBAN D: 2754

(Other) :481830

NA's : 13945

Grantee ValidInvalid

FEDERAL NATIONAL MORTGAGE ASSOCIATION : 4574 Min. :0.0000

SSHI LLC : 3422 1st Qu.:0.0000

SECRETARY OF HOUSING & URBAN DEVELOPMENT: 2544 Median :1.0000

FEDERAL HOME LOAN MORTGAGE CORPORATION : 1885 Mean :0.6157

WELLS FARGO BANK : 1674 3rd Qu.:1.0000

(Other) :496546 Max. :1.0000

NA's : 13584

ConfirmedUnconfirmed ExcludeReason

Min. :0.0000 MultiParcel Sale one time only: 33579

1st Qu.:0.0000 Improved after sale : 19427

Median :0.0000 Foreclosure Sale : 18120

Mean :0.4739 Estate sale : 11908

3rd Qu.:1.0000 Exempt for taxatn Gov nonprof : 11036

Max. :1.0000 (Other) : 53489

NA's :376670

ImprovedVacant

Min. :0.0000

1st Qu.:1.0000

Median :1.0000

Mean :0.8413

3rd Qu.:1.0000

Max. :1.0000

> names(sale)

[1] "ENT" "ParcelCount" "ParcelNumber"

[4] "SaleDate" "SalePrice" "DeedType"

[7] "Grantor" "Grantee" "ValidInvalid"

[10] "ConfirmedUnconfirmed" "ExcludeReason" "ImprovedVacant"

> str(sale)

'data.frame': 524229 obs. of 12 variables:

$ ENT : Factor w/ 457392 levels "0000001","0094166",..: 196844 196845 196846 196847 196848 196849 196850 196851 196852 196853 ...

$ ParcelCount : int 1 1 1 1 1 1 1 1 1 1 ...

$ ParcelNumber : num 5.00e+09 9.00e+09 2.17e+08 5.02e+09 2.32e+09 ...

$ SaleDate : Date, format: "2006-03-31" "2006-03-21" ...

$ SalePrice : num 29000 252500 364105 149900 220000 ...

$ DeedType : Factor w/ 31 levels "Administrative Deed",..: 14 26 26 26 26 26 26 26 26 26 ...

$ Grantor : Factor w/ 263544 levels "& PORTER RALPH E",..: 125747 168727 165541 161350 13307 83084 175186 40220 259753 21246 ...

$ Grantee : Factor w/ 345482 levels "'ELAN ST HELENS PARTNERS LLC",..: 269593 284399 11219 116483 170309 235960 273798 179324 49737 191405 ...

$ ValidInvalid : int 1 1 1 1 1 1 1 1 1 1 ...

$ ConfirmedUnconfirmed: int 1 1 1 0 1 1 0 0 0 0 ...

$ ExcludeReason : Factor w/ 63 levels "Air Rights only UC9770 Vac Lnd",..: NA NA 15 NA NA NA NA NA 49 NA ...

$ ImprovedVacant : int 1 1 1 1 1 1 1 1 1 1 ...

>

> #correlation matrix for sale numerical variables

> sale\_num <- sale[,c(2,3,5,9,10,12)]

> cor(sale\_num)

ParcelCount ParcelNumber SalePrice ValidInvalid

ParcelCount 1.00000000 0.03732804 0.72551146 -0.16951349

ParcelNumber 0.03732804 1.00000000 -0.02588050 0.05848443

SalePrice 0.72551146 -0.02588050 1.00000000 -0.12662609

ValidInvalid -0.16951349 0.05848443 -0.12662609 1.00000000

ConfirmedUnconfirmed -0.03914719 -0.01574178 -0.03306629 0.01640046

ImprovedVacant -0.29682513 0.10112958 -0.21238256 0.34755507

ConfirmedUnconfirmed ImprovedVacant

ParcelCount -0.03914719 -0.2968251

ParcelNumber -0.01574178 0.1011296

SalePrice -0.03306629 -0.2123826

ValidInvalid 0.01640046 0.3475551

ConfirmedUnconfirmed 1.00000000 -0.1234460

ImprovedVacant -0.12344601 1.0000000

>

> #eda of seg merge table

> summary(segmerge)

SegMergeNumber ParentChildIndicator ParcelNumber

2012-0331: 423 C:126610 Min. :1.013e+07

2013-0001: 422 P: 43829 1st Qu.:5.192e+08

2018-0012: 324 Median :5.003e+09

2005-0719: 320 Mean :4.360e+09

2006-1194: 319 3rd Qu.:6.470e+09

2005-1191: 303 Max. :1.000e+10

(Other) :168328

ContinuedIndicator CompletedDate TaxYear

N: 42202 Min. :1991-07-17 Min. :1992

Y:128237 1st Qu.:1996-09-16 1st Qu.:1997

Median :2003-03-28 Median :2003

Mean :2003-02-17 Mean :2003

3rd Qu.:2007-11-09 3rd Qu.:2008

Max. :2019-06-12 Max. :2019

NA's :13

> names(segmerge)

[1] "SegMergeNumber" "ParentChildIndicator" "ParcelNumber"

[4] "ContinuedIndicator" "CompletedDate" "TaxYear"

> str(segmerge)

'data.frame': 170439 obs. of 6 variables:

$ SegMergeNumber : Factor w/ 21945 levels "2004-000002",..: 1 1 1 2 2 2 3 3 3 4 ...

$ ParentChildIndicator: Factor w/ 2 levels "C","P": 1 2 2 1 1 2 1 2 2 1 ...

$ ParcelNumber : num 2.95e+09 2.95e+09 7.38e+09 5.19e+08 5.19e+08 ...

$ ContinuedIndicator : Factor w/ 2 levels "N","Y": 2 1 1 2 2 1 2 1 1 2 ...

$ CompletedDate : Date, format: "2003-08-01" "2003-08-01" ...

$ TaxYear : int 2003 2003 2003 2003 2003 2003 2003 2003 2003 2004 ...

>

> #correlation matrix for seg merge numerical variables

> segmerge\_num <- segmerge[,c(3,6)]

> cor(segmerge\_num)

ParcelNumber TaxYear

ParcelNumber 1 NA

TaxYear NA 1

>

> #eda of tax account table

> summary(tax\_account)

ParcelNumber AccountType PropertyType

Min. :1.850e+05 MOBIL: 12512 ASIMP: 5

1st Qu.:6.192e+08 PERS : 11114 LNDIM:317506

Median :4.695e+09 REAL :317634 MBLHM: 12512

Mean :4.214e+09 STRUC: 2379 SAPP : 1000

3rd Qu.:6.385e+09 SARP : 123

Max. :9.995e+09 STRUC: 1379

NA's : 11114

SiteAddress UseCode

XXX Undetermined Situs: 1916 Min. : 0

TRACTS : 1154 1st Qu.:1101

OP PROP : 1123 Median :1101

Undetermined Situs : 1006 Mean :2391

REFERENCE : 809 3rd Qu.:1305

XXX ORVILLE RD E : 172 Max. :9900

(Other) :337459 NA's :886

UseDescription TaxYearPrior TaxCodeAreaPrior

SINGLE FAMILY DWELLING :220028 Min. :2018 Min. : 5.0

VACANT LAND UNDEVELOPED: 23805 1st Qu.:2018 1st Qu.: 45.0

MOBILE/MFG HOME : 15482 Median :2018 Median :250.0

MH TITLE ELIM : 9349 Mean :2018 Mean :325.1

DUPLEX 2 UNITS : 5608 3rd Qu.:2018 3rd Qu.:578.0

(Other) : 68481 Max. :2018 Max. :999.0

NA's : 886 NA's :3518

ExemptionTypePriorYear

Senior/Disabled A : 7715

Municipal Corp and Misc Taxing Districts: 4249

Reference Parcels : 3114

Senior/Disabled B : 1789

County Owned Property : 1631

(Other) : 10052

NA's :315089

CurrentUseCodePriorYear LandValuePriorYear ImprovementValuePriorYear

AGRI : 1260 Min. : 0 Min. : 0

FORDG: 2885 1st Qu.: 57600 1st Qu.: 76800

OPBRS: 681 Median : 83900 Median : 155800

OPEN : 216 Mean : 123919 Mean : 218756

NA's :338597 3rd Qu.: 117000 3rd Qu.: 217800

Max. :61605900 Max. :200520800

NA's :14881 NA's :13955

TotalMarketValuePriorYear TaxableValuePriorYear TaxYearCurrent

Min. : 0 Min. : 0 Min. :2019

1st Qu.: 145800 1st Qu.: 112000 1st Qu.:2019

Median : 240000 Median : 231800 Median :2019

Mean : 337970 Mean : 295969 Mean :2019

3rd Qu.: 327300 3rd Qu.: 319600 3rd Qu.:2019

Max. :220839400 Max. :220839400 Max. :2019

NA's :3551 NA's :3551

TaxCodeAreaCurrentYear

Min. : 5.0

1st Qu.: 45.0

Median :250.0

Mean :324.9

3rd Qu.:578.0

Max. :999.0

NA's :463

ExemptionTypeCurrentYear

Senior/Disabled A : 7862

Municipal Corp and Misc Taxing Districts: 4271

Reference Parcels : 3157

Senior/Disabled B : 1848

County Owned Property : 1646

(Other) : 9798

NA's :315057

CurrentUseCodeCurrentYear LandValueCurrentYear

AGRI : 1274 Min. : 0

FORDG: 2905 1st Qu.: 72500

OPBRS: 697 Median : 101100

OPEN : 210 Mean : 147202

NA's :338553 3rd Qu.: 139900

Max. :67142800

NA's :12343

ImprovementValueCurrentYear TotalMarketValueCurrentYear

Min. : 0 Min. : 0

1st Qu.: 87900 1st Qu.: 173300

Median : 170500 Median : 272100

Mean : 236934 Mean : 378027

3rd Qu.: 236800 3rd Qu.: 367000

Max. :209019000 Max. :222953800

NA's :11219 NA's :535

TaxableValueCurrentYear Range Township

Min. : 0 Min. : 0.000 Min. : 0.00

1st Qu.: 133700 1st Qu.: 2.000 1st Qu.:19.00

Median : 263700 Median : 3.000 Median :20.00

Mean : 332097 Mean : 3.123 Mean :19.58

3rd Qu.: 358400 3rd Qu.: 4.000 3rd Qu.:20.00

Max. :222953800 Max. :11.000 Max. :22.00

NA's :535 NA's :24782 NA's :24782

Section QuarterSection SubdivisionName

05 : 13361 22 : 21102 2ND SCHOOL LD ADD: 2205

28 : 12506 41 : 21033 TAC LAND CO 6TH : 1738

03 : 11974 33 : 20884 1ST SCHOOL LD ADD: 1538

32 : 11402 44 : 20709 INDIAN ADD : 1190

04 : 11246 11 : 20487 LAKE PARK : 971

(Other):258369 (Other):214642 (Other) :219011

NA's : 24781 NA's : 24782 NA's :116986

LocatedOnParcel

| :320721

0419104001|: 240

0419203047|: 182

0320312076|: 168

0420223061|: 159

0319281056|: 143

(Other) : 22026

> names(tax\_account)

[1] "ParcelNumber" "AccountType"

[3] "PropertyType" "SiteAddress"

[5] "UseCode" "UseDescription"

[7] "TaxYearPrior" "TaxCodeAreaPrior"

[9] "ExemptionTypePriorYear" "CurrentUseCodePriorYear"

[11] "LandValuePriorYear" "ImprovementValuePriorYear"

[13] "TotalMarketValuePriorYear" "TaxableValuePriorYear"

[15] "TaxYearCurrent" "TaxCodeAreaCurrentYear"

[17] "ExemptionTypeCurrentYear" "CurrentUseCodeCurrentYear"

[19] "LandValueCurrentYear" "ImprovementValueCurrentYear"

[21] "TotalMarketValueCurrentYear" "TaxableValueCurrentYear"

[23] "Range" "Township"

[25] "Section" "QuarterSection"

[27] "SubdivisionName" "LocatedOnParcel"

> str(tax\_account)

'data.frame': 343639 obs. of 28 variables:

$ ParcelNumber : num 2.01e+09 2.01e+09 2.01e+09 2.01e+09 2.01e+09 ...

$ AccountType : Factor w/ 4 levels "MOBIL","PERS",..: 3 3 3 3 3 2 2 3 3 3 ...

$ PropertyType : Factor w/ 6 levels "ASIMP","LNDIM",..: 2 2 2 2 2 NA NA 2 2 2 ...

$ SiteAddress : Factor w/ 308649 levels "1 & 2 SILVER SKIS CHALET",..: 272120 272919 274294 275588 72451 185079 218553 270919 271986 273581 ...

$ UseCode : int 1101 1101 1101 1101 1202 6500 6100 1101 1101 1101 ...

$ UseDescription : Factor w/ 186 levels "AG NOT CURRENT USE",..: 156 156 156 156 40 128 107 156 156 156 ...

$ TaxYearPrior : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...

$ TaxCodeAreaPrior : int 5 5 5 5 5 5 105 5 5 5 ...

$ ExemptionTypePriorYear : Factor w/ 42 levels "0-Unknown","Cemetery",..: NA NA NA NA NA NA NA NA NA NA ...

$ CurrentUseCodePriorYear : Factor w/ 4 levels "AGRI","FORDG",..: NA NA NA NA NA NA NA NA NA NA ...

$ LandValuePriorYear : int 52000 48800 61600 56500 45400 NA NA 58400 58400 52000 ...

$ ImprovementValuePriorYear : int 133400 133900 251400 176200 119100 NA NA 122800 103300 154400 ...

$ TotalMarketValuePriorYear : int 185400 182700 313000 232700 164500 45507 8744 181200 161700 206400 ...

$ TaxableValuePriorYear : int 185400 182700 313000 232700 164500 45507 8744 181200 161700 206400 ...

$ TaxYearCurrent : int 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...

$ TaxCodeAreaCurrentYear : int 5 5 5 5 5 5 105 5 5 5 ...

$ ExemptionTypeCurrentYear : Factor w/ 41 levels "0-Unknown","Cemetery",..: NA NA NA NA NA NA NA 32 NA NA ...

$ CurrentUseCodeCurrentYear : Factor w/ 4 levels "AGRI","FORDG",..: NA NA NA NA NA NA NA NA NA NA ...

$ LandValueCurrentYear : int 73200 69300 84600 78500 65200 NA NA 80800 80800 73200 ...

$ ImprovementValueCurrentYear: int 153000 171300 295100 198000 143700 NA NA 141700 118900 174000 ...

$ TotalMarketValueCurrentYear: int 226200 240600 379700 276500 208900 40322 6722 222500 199700 247200 ...

$ TaxableValueCurrentYear : int 226200 240600 379700 276500 208900 40322 6722 222500 199700 247200 ...

$ Range : int 3 3 3 3 3 NA NA 3 3 3 ...

$ Township : int 20 20 20 20 20 NA NA 20 20 20 ...

$ Section : Factor w/ 39 levels "00","01","02",..: 6 6 6 6 6 NA NA 6 6 6 ...

$ QuarterSection : Factor w/ 25 levels "00","03","10",..: 10 10 10 11 11 NA NA 10 10 10 ...

$ SubdivisionName : Factor w/ 6843 levels "\"A\" STREET CONDOS",..: NA NA NA NA NA NA NA NA NA NA ...

$ LocatedOnParcel : Factor w/ 10761 levels "|","0020011004|",..: 1 1 1 1 1 7434 7441 1 1 1 ...

>

> #correlation matrix for tax account numerical variables

> tax\_account\_num <- tax\_account[,c(1,5,7,8,11,12,13,14,15,16,19,20,21,22,23,24)]

> cor(tax\_account\_num) #essentially all NAs throughout the correlation matrix

ParcelNumber UseCode TaxYearPrior

ParcelNumber 1 NA NA

UseCode NA 1 NA

TaxYearPrior NA NA 1

TaxCodeAreaPrior NA NA NA

LandValuePriorYear NA NA NA

ImprovementValuePriorYear NA NA NA

TotalMarketValuePriorYear NA NA NA

TaxableValuePriorYear NA NA NA

TaxYearCurrent NA NA NA

TaxCodeAreaCurrentYear NA NA NA

LandValueCurrentYear NA NA NA

ImprovementValueCurrentYear NA NA NA

TotalMarketValueCurrentYear NA NA NA

TaxableValueCurrentYear NA NA NA

Range NA NA NA

Township NA NA NA

TaxCodeAreaPrior LandValuePriorYear

ParcelNumber NA NA

UseCode NA NA

TaxYearPrior NA NA

TaxCodeAreaPrior 1 NA

LandValuePriorYear NA 1

ImprovementValuePriorYear NA NA

TotalMarketValuePriorYear NA NA

TaxableValuePriorYear NA NA

TaxYearCurrent NA NA

TaxCodeAreaCurrentYear NA NA

LandValueCurrentYear NA NA

ImprovementValueCurrentYear NA NA

TotalMarketValueCurrentYear NA NA

TaxableValueCurrentYear NA NA

Range NA NA

Township NA NA

ImprovementValuePriorYear

ParcelNumber NA

UseCode NA

TaxYearPrior NA

TaxCodeAreaPrior NA

LandValuePriorYear NA

ImprovementValuePriorYear 1

TotalMarketValuePriorYear NA

TaxableValuePriorYear NA

TaxYearCurrent NA

TaxCodeAreaCurrentYear NA

LandValueCurrentYear NA

ImprovementValueCurrentYear NA

TotalMarketValueCurrentYear NA

TaxableValueCurrentYear NA

Range NA

Township NA

TotalMarketValuePriorYear

ParcelNumber NA

UseCode NA

TaxYearPrior NA

TaxCodeAreaPrior NA

LandValuePriorYear NA

ImprovementValuePriorYear NA

TotalMarketValuePriorYear 1

TaxableValuePriorYear NA

TaxYearCurrent NA

TaxCodeAreaCurrentYear NA

LandValueCurrentYear NA

ImprovementValueCurrentYear NA

TotalMarketValueCurrentYear NA

TaxableValueCurrentYear NA

Range NA

Township NA

TaxableValuePriorYear TaxYearCurrent

ParcelNumber NA NA

UseCode NA NA

TaxYearPrior NA NA

TaxCodeAreaPrior NA NA

LandValuePriorYear NA NA

ImprovementValuePriorYear NA NA

TotalMarketValuePriorYear NA NA

TaxableValuePriorYear 1 NA

TaxYearCurrent NA 1

TaxCodeAreaCurrentYear NA NA

LandValueCurrentYear NA NA

ImprovementValueCurrentYear NA NA

TotalMarketValueCurrentYear NA NA

TaxableValueCurrentYear NA NA

Range NA NA

Township NA NA

TaxCodeAreaCurrentYear LandValueCurrentYear

ParcelNumber NA NA

UseCode NA NA

TaxYearPrior NA NA

TaxCodeAreaPrior NA NA

LandValuePriorYear NA NA

ImprovementValuePriorYear NA NA

TotalMarketValuePriorYear NA NA

TaxableValuePriorYear NA NA

TaxYearCurrent NA NA

TaxCodeAreaCurrentYear 1 NA

LandValueCurrentYear NA 1

ImprovementValueCurrentYear NA NA

TotalMarketValueCurrentYear NA NA

TaxableValueCurrentYear NA NA

Range NA NA

Township NA NA

ImprovementValueCurrentYear

ParcelNumber NA

UseCode NA

TaxYearPrior NA

TaxCodeAreaPrior NA

LandValuePriorYear NA

ImprovementValuePriorYear NA

TotalMarketValuePriorYear NA

TaxableValuePriorYear NA

TaxYearCurrent NA

TaxCodeAreaCurrentYear NA

LandValueCurrentYear NA

ImprovementValueCurrentYear 1

TotalMarketValueCurrentYear NA

TaxableValueCurrentYear NA

Range NA

Township NA

TotalMarketValueCurrentYear

ParcelNumber NA

UseCode NA

TaxYearPrior NA

TaxCodeAreaPrior NA

LandValuePriorYear NA

ImprovementValuePriorYear NA

TotalMarketValuePriorYear NA

TaxableValuePriorYear NA

TaxYearCurrent NA

TaxCodeAreaCurrentYear NA

LandValueCurrentYear NA

ImprovementValueCurrentYear NA

TotalMarketValueCurrentYear 1

TaxableValueCurrentYear NA

Range NA

Township NA

TaxableValueCurrentYear Range Township

ParcelNumber NA NA NA

UseCode NA NA NA

TaxYearPrior NA NA NA

TaxCodeAreaPrior NA NA NA

LandValuePriorYear NA NA NA

ImprovementValuePriorYear NA NA NA

TotalMarketValuePriorYear NA NA NA

TaxableValuePriorYear NA NA NA

TaxYearCurrent NA NA NA

TaxCodeAreaCurrentYear NA NA NA

LandValueCurrentYear NA NA NA

ImprovementValueCurrentYear NA NA NA

TotalMarketValueCurrentYear NA NA NA

TaxableValueCurrentYear 1 NA NA

Range NA 1 NA

Township NA NA 1

Warning message:

In cor(tax\_account\_num) : the standard deviation is zero

>

> #eda of tax description table

> summary(tax\_description)

ParcelNumber LineNumber

Min. :1.901e+07 Min. : 1.000

1st Qu.:4.151e+08 1st Qu.: 1.000

Median :3.375e+09 Median : 3.000

Mean :3.500e+09 Mean : 4.623

3rd Qu.:6.024e+09 3rd Qu.: 6.000

Max. :9.995e+09 Max. :102.000

TaxDescriptionLine

EASE OF RECORD| : 6245

TOG/W EASE & RESTRICTIONS OF REC| : 2419

NEW TACOMA| : 1780

TAC LAND CO 6TH| : 1273

RECORD| : 1262

EASE & RESTRICTIONS OF REC OUT OF|: 1236

(Other) :1179349

> names(tax\_description)

[1] "ParcelNumber" "LineNumber" "TaxDescriptionLine"

> str(tax\_description)

'data.frame': 1193564 obs. of 3 variables:

$ ParcelNumber : num 5e+09 5e+09 5e+09 5e+09 5e+09 ...

$ LineNumber : int 5 6 1 1 2 3 4 5 6 1 ...

$ TaxDescriptionLine: Factor w/ 733935 levels "'05 TAX YR|",..: 119497 50316 577727 578976 136182 669744 432126 119497 50316 577728 ...

>

> #correlation matrix for tax description numerical variables

> tax\_description\_num <- tax\_description[,c(1,2)]

> cor(tax\_description\_num)

ParcelNumber LineNumber

ParcelNumber 1.0000000 -0.2809248

LineNumber -0.2809248 1.0000000

>

> #eda of complied data table

> summary(compiled)

ParcelNumber AppraisalAccountType Buildings

Min. :2.101e+07 Residential:127232 Min. : 0.000

1st Qu.:5.202e+08 1st Qu.: 1.000

Median :5.003e+09 Median : 1.000

Mean :4.334e+09 Mean : 1.199

3rd Qu.:6.026e+09 3rd Qu.: 1.000

Max. :9.900e+09 Max. :10.000

LandGrossSquareFeet LandNetSquareFeet AppraisalDate

Min. : 0 Min. : 510 Min. :2006-04-19

1st Qu.: 7778 1st Qu.: 7781 1st Qu.:2014-12-29

Median : 12000 Median : 12000 Median :2015-11-18

Mean : 37781 Mean : 37809 Mean :2016-04-24

3rd Qu.: 22576 3rd Qu.: 22575 3rd Qu.:2018-04-26

Max. :13939200 Max. :13939200 Max. :2019-06-25

NA's :2

Latitude Longtitude BuildingID PropertyType

Min. :46.77 Min. :-122.8 Min. : 1.000 Residential:127232

1st Qu.:47.11 1st Qu.:-122.5 1st Qu.: 1.000

Median :47.16 Median :-122.3 Median : 1.000

Mean :47.17 Mean :-122.4 Mean : 1.045

3rd Qu.:47.22 3rd Qu.:-122.3 3rd Qu.: 1.000

Max. :47.40 Max. :-122.1 Max. :11.000

SquareFeet PercentComplete Condition

Min. : 1 Min. :0.0000 Average :123188

1st Qu.: 1419 1st Qu.:1.0000 Extra Poor : 167

Median : 1822 Median :1.0000 Fair : 2807

Mean : 1920 Mean :0.9976 Good : 19

3rd Qu.: 2320 3rd Qu.:1.0000 Poor : 656

Max. :13004 Max. :1.0000 Uninhabitable: 173

Very Poor : 222

Quality PhysicalAge YearBuilt YearRemodeled

Average :70678 Min. : 0.0 Min. :1872 Min. : 0

Fair Plus :20038 1st Qu.: 16.0 1st Qu.:1971 1st Qu.: 0

Average Plus:16041 Median : 25.0 Median :1990 Median :1980

Fair : 7932 Mean : 25.4 Mean :1984 Mean :1207

Good : 6197 3rd Qu.: 33.0 3rd Qu.:2002 3rd Qu.:1992

Good Plus : 2606 Max. :119.0 Max. :2019 Max. :2019

(Other) : 3740 NA's :368

AccountType TaxableValuePriorYear TaxableValueCurrentYear

REAL:127232 Min. : 0 Min. : 0

1st Qu.: 229100 1st Qu.: 261000

Median : 282800 Median : 319300

Mean : 322957 Mean : 364307

3rd Qu.: 368500 3rd Qu.: 413900

Max. :4182900 Max. :4600100

NA's :202

Zipcode

Min. :98328

1st Qu.:98371

Median :98375

Mean :98386

3rd Qu.:98391

Max. :98467

> names(compiled)

[1] "ParcelNumber" "AppraisalAccountType"

[3] "Buildings" "LandGrossSquareFeet"

[5] "LandNetSquareFeet" "AppraisalDate"

[7] "Latitude" "Longtitude"

[9] "BuildingID" "PropertyType"

[11] "SquareFeet" "PercentComplete"

[13] "Condition" "Quality"

[15] "PhysicalAge" "YearBuilt"

[17] "YearRemodeled" "AccountType"

[19] "TaxableValuePriorYear" "TaxableValueCurrentYear"

[21] "Zipcode"

> str(compiled)

'data.frame': 127232 obs. of 21 variables:

$ ParcelNumber : num 1.21e+08 1.21e+08 1.21e+08 1.21e+08 1.21e+08 ...

$ AppraisalAccountType : Factor w/ 1 level "Residential": 1 1 1 1 1 1 1 1 1 1 ...

$ Buildings : int 2 1 1 1 2 2 2 2 2 2 ...

$ LandGrossSquareFeet : num 213008 77972 43875 47916 145926 ...

$ LandNetSquareFeet : num 213008 77972 43875 47916 145926 ...

$ AppraisalDate : Date, format: "2013-11-19" "2018-07-10" ...

$ Latitude : num 47.3 47.3 47.3 47.3 47.3 ...

$ Longtitude : num -123 -123 -123 -123 -123 ...

$ BuildingID : int 1 1 1 1 1 1 2 2 1 1 ...

$ PropertyType : Factor w/ 1 level "Residential": 1 1 1 1 1 1 1 1 1 1 ...

$ SquareFeet : int 2082 2948 2235 2909 3921 3921 1315 1315 1568 2812 ...

$ PercentComplete : num 1 0.45 1 1 1 1 1 1 1 1 ...

$ Condition : Factor w/ 7 levels "Average","Extra Poor",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Quality : Factor w/ 11 levels "Average","Average Plus",..: 1 2 6 2 2 2 1 1 5 1 ...

$ PhysicalAge : int 31 1 21 25 24 24 18 18 53 31 ...

$ YearBuilt : int 1979 2018 1998 1992 1994 1994 2001 2001 1935 1979 ...

$ YearRemodeled : int 1988 0 0 1994 1995 1995 0 0 1966 1988 ...

$ AccountType : Factor w/ 1 level "REAL": 1 1 1 1 1 1 1 1 1 1 ...

$ TaxableValuePriorYear : int 463600 147500 495700 501200 745300 745300 745300 745300 125040 406000 ...

$ TaxableValueCurrentYear: int 514200 394000 537700 543400 805200 805200 805200 805200 125040 444300 ...

$ Zipcode : int 98332 98332 98332 98332 98332 98332 98332 98332 98332 98332 ...

>

> #correlation matrix for compiled data numerical variables

> compiled\_num <- compiled[,c(1,3,4,5,7,8,9,11,12,15,16,17,19,20,21)]

> cor(compiled\_num)

ParcelNumber Buildings LandGrossSquareFeet

ParcelNumber 1.00000000 -0.29339362 -0.28493003

Buildings -0.29339362 1.00000000 0.34134390

LandGrossSquareFeet -0.28493003 0.34134390 1.00000000

LandNetSquareFeet -0.28341102 0.34014279 0.99908256

Latitude -0.10424802 -0.02157849 -0.12341222

Longtitude 0.19712970 -0.03731793 -0.02480770

BuildingID -0.13217896 0.48857436 0.16256949

SquareFeet 0.01771059 -0.10326012 0.03383448

PercentComplete 0.02005352 -0.01171366 -0.01182218

PhysicalAge -0.18904639 0.25386564 0.10395845

YearBuilt 0.19844067 -0.29314434 -0.09735845

YearRemodeled NA NA NA

TaxableValuePriorYear NA NA NA

TaxableValueCurrentYear -0.18871943 0.23213511 0.17504089

Zipcode 0.18037079 -0.04615203 -0.17498898

LandNetSquareFeet Latitude Longtitude

ParcelNumber -0.28341102 -0.104248018 0.197129698

Buildings 0.34014279 -0.021578492 -0.037317931

LandGrossSquareFeet 0.99908256 -0.123412218 -0.024807698

LandNetSquareFeet 1.00000000 -0.123139475 -0.024358594

Latitude -0.12313948 1.000000000 -0.582725867

Longtitude -0.02435859 -0.582725867 1.000000000

BuildingID 0.16155109 -0.003673808 -0.019124490

SquareFeet 0.03364196 0.079997420 0.021199825

PercentComplete -0.01171713 -0.022040344 0.014107207

PhysicalAge 0.10375887 0.092976807 -0.137286226

YearBuilt -0.09714429 -0.102987039 0.137868880

YearRemodeled NA NA NA

TaxableValuePriorYear NA NA NA

TaxableValueCurrentYear 0.17397277 0.320104417 -0.161622098

Zipcode -0.17411871 -0.099166989 0.008609365

BuildingID SquareFeet PercentComplete

ParcelNumber -0.132178963 0.01771059 0.02005352

Buildings 0.488574358 -0.10326012 -0.01171366

LandGrossSquareFeet 0.162569490 0.03383448 -0.01182218

LandNetSquareFeet 0.161551088 0.03364196 -0.01171713

Latitude -0.003673808 0.07999742 -0.02204034

Longtitude -0.019124490 0.02119983 0.01410721

BuildingID 1.000000000 -0.12379183 -0.05168209

SquareFeet -0.123791832 1.00000000 -0.05270669

PercentComplete -0.051682090 -0.05270669 1.00000000

PhysicalAge 0.036200896 -0.45194407 0.09199819

YearBuilt -0.053243725 0.43245275 -0.05264774

YearRemodeled NA NA NA

TaxableValuePriorYear NA NA NA

TaxableValueCurrentYear 0.175990471 0.56295022 0.01476502

Zipcode -0.027774302 -0.15364540 0.02022101

PhysicalAge YearBuilt YearRemodeled

ParcelNumber -0.18904639 0.19844067 NA

Buildings 0.25386564 -0.29314434 NA

LandGrossSquareFeet 0.10395845 -0.09735845 NA

LandNetSquareFeet 0.10375887 -0.09714429 NA

Latitude 0.09297681 -0.10298704 NA

Longtitude -0.13728623 0.13786888 NA

BuildingID 0.03620090 -0.05324372 NA

SquareFeet -0.45194407 0.43245275 NA

PercentComplete 0.09199819 -0.05264774 NA

PhysicalAge 1.00000000 -0.87715287 NA

YearBuilt -0.87715287 1.00000000 NA

YearRemodeled NA NA 1

TaxableValuePriorYear NA NA NA

TaxableValueCurrentYear -0.17422112 0.14870015 NA

Zipcode 0.18477227 -0.20185964 NA

TaxableValuePriorYear TaxableValueCurrentYear

ParcelNumber NA -0.18871943

Buildings NA 0.23213511

LandGrossSquareFeet NA 0.17504089

LandNetSquareFeet NA 0.17397277

Latitude NA 0.32010442

Longtitude NA -0.16162210

BuildingID NA 0.17599047

SquareFeet NA 0.56295022

PercentComplete NA 0.01476502

PhysicalAge NA -0.17422112

YearBuilt NA 0.14870015

YearRemodeled NA NA

TaxableValuePriorYear 1 NA

TaxableValueCurrentYear NA 1.00000000

Zipcode NA -0.22050978

Zipcode

ParcelNumber 0.180370793

Buildings -0.046152025

LandGrossSquareFeet -0.174988982

LandNetSquareFeet -0.174118705

Latitude -0.099166989

Longtitude 0.008609365

BuildingID -0.027774302

SquareFeet -0.153645401

PercentComplete 0.020221005

PhysicalAge 0.184772272

YearBuilt -0.201859640

YearRemodeled NA

TaxableValuePriorYear NA

TaxableValueCurrentYear -0.220509785

Zipcode 1.000000000

Code example:

# Nominal data:

dfnorminal <- df %>% select(Condition, Quality)

# Builds a contingency table

nominaltable <- table(dfnorminal$Condition,dfnorminal$Quality)

# conduct Chi-square test

chisq.test(nominaltable)

Chi-squared approximation may be incorrect

Pearson's Chi-squared test

data: nominaltable

X-squared = 1823.2, df = 60, p-value < 2.2e-16

## P-VALUE < 2.2e-16, We reject the null hypothesis. We can say that Condition and Quality have a high correlation

# numerical data:

# Numerical data

dfnumerical <- df %>% select("Buildings","LandGrossSquareFeet","LandNetSquareFeet","SquareFeet","PercentComplete","PhysicalAge","YearBuilt","YearRemodeled","TaxableValuePriorYear","TaxableValueCurrentYear","Arrests","Arson","Assault","Burglary","Drug","Fraud","Harassment","Homicide","Intimidation","LiquorLaw","Robbery","Theft","Traffic","Vandalism")

#correlation matrix

cor(dfnumerical,use ='complete.obs')

# Overall Parcel attribute do not have much correlation with each others, except, Year Build, Year remodel and Physical Age

# Number of type of crime occured have strong correlation wtich each other

# House value has low correlation coefficient with crime variable

# function to test correlation between variable in the dataset.

funccortest <- function(listx, df){

#y = enquo(yaxis)

for(i in 1:length(df)){

for(j in 2:length(df)){

cortest<- cor.test(df[,i], df[,j])

p <- cortest$p.value

result <- ifelse(p<= 0.05, "are highly correlated", "are independent")

print(paste('p-value =',p,',' ,listx[i],'and',listx[j],result, sep =' '))

}

}

}

# testing correlation all numerical variable

funccortest(colnames(dfnumerical),dfnumerical)

Correlation Analysis R Code:

---

title: "Capstone Project"

author:

date: "`r format(Sys.time(), '%B %d, %Y')`"

output: html\_notebook

---

Commented header

=========================================

```{r echo = TRUE}

# Course: Capstone

# Title:

# Purpose: Correlation Analysis

# Date:

# Author:

```

Loading library

===============================

```{r}

#library

library(tidyverse)

library(RODBC)

library(stringr)

library(psych)

```

Creating sql server connection

===============================

```{r}

conn <- odbcConnect("oitap22")

```

Loading data from sql server

===============================

```{r}

# get data sample from sql server

data <- sqlQuery(conn,"select \* from [19su5510\_lehuy].[dbo].[sampledata]")

summary(data)

```

Data preprocessing

===============================

```{r}

#

str(data)

# grouping crime type and spread data to column

theftstr <- c('theft', 'stolen')

arsonstr <- c('Arson - Non-residential','Arson - Residential')

assaultstr <- c('Assault - Aggravated','Assault - Simple')

drugstr <- c('Drug Sale/Manufacture (Methamphetamine)','Drug Possession (Methamphetamine)','Drug Sale/Manufacture (Other)','Drug Possession (Other)')

data1 <- data %>% mutate(CrimeType = ifelse(str\_detect(tolower(Public\_Nam),'vandalism'),'Vandalism',

ifelse(str\_detect(tolower(Public\_Nam),'robbery'),'Robbery',

ifelse(str\_detect(tolower(Public\_Nam),'burglary'),'Burglary',

ifelse(str\_detect(tolower(Public\_Nam),'drug'),'Drug',

ifelse(str\_detect(tolower(Public\_Nam),'assault'),'Assault',

ifelse(str\_detect(tolower(Public\_Nam),'arrests'),'Arrests',

ifelse(str\_detect(tolower(Public\_Nam),'arson'),'Arson',

ifelse(str\_detect(tolower(Public\_Nam),'traffic'),'Traffic',

ifelse(str\_detect(tolower(Public\_Nam),'fraud'),'Fraud',

ifelse(str\_detect(tolower(Public\_Nam),'homicide'),'Homicide',

ifelse(str\_detect(tolower(Public\_Nam),'harassment'),'Harassment',

ifelse(str\_detect(tolower(Public\_Nam),'liquor'),'LiquorLaw',

ifelse(Public\_Nam =='Intimidation','Intimidation','Theft'))))))))))))))

data1crime <- data1 %>% select(ParcelNumber,CrimeType,NoofCrime) %>%

group\_by(ParcelNumber,CrimeType) %>%

summarise(countCrime = sum(NoofCrime)) %>%

spread(CrimeType,countCrime,fill = 0)

df <- data1 %>% select(-22,-23,-24) %>% distinct() %>%

inner\_join(data1crime, by ='ParcelNumber')

```

Correlation analysis

==================================

```{r}

# Norminal data

dfnorminal <- df %>% select(Condition, Quality)

# Builds a contingency table

nominaltable <- table(dfnorminal$Condition,dfnorminal$Quality)

chisq.test(nominaltable)

```

- P-VALUE < 2.2e-16 => Condition and Quality have a strong correlation

```{r}

# Numerical data

dfnumerical <- df %>% select("Buildings","LandGrossSquareFeet","LandNetSquareFeet","SquareFeet","PercentComplete","PhysicalAge","YearBuilt","YearRemodeled","TaxableValuePriorYear","TaxableValueCurrentYear","Arrests","Arson","Assault","Burglary","Drug","Fraud","Harassment","Homicide","Intimidation","LiquorLaw","Robbery","Theft","Traffic","Vandalism")

#correlation matrix

cor(dfnumerical,use ='complete.obs')

# Overall Parcel attribute do not have much correlation with each others, except, Year Build, Year remodel and Physical Age

# Number of type of crime occured have strong correlation wtich each other

# House value has low correlation coefficient with crime variable

# function to test correlation between variable in the dataset.

funccortest <- function(listx, df){

#y = enquo(yaxis)

for(i in 1:length(df)){

for(j in 2:length(df)){

cortest<- cor.test(df[,i], df[,j])

p <- cortest$p.value

result <- ifelse(p<= 0.05, "are highly correlated", "are independent")

print(paste('p-value =',p,',' ,listx[i],'and',listx[j],result, sep =' '))

}

}

}

# testing correlation all numerical variable

funccortest(colnames(dfnumerical),dfnumerical)

```

```{r}

# house attribute

dfyear <- df %>% select("PhysicalAge","YearBuilt","YearRemodeled")

cor(dfyear,use ='complete.obs')

# Perform correlation test

cor.test(dfyear$PhysicalAge, dfyear$YearBuilt)

cor.test(dfyear$PhysicalAge, dfyear$YearRemodeled)

cor.test(dfyear$YearRemodeled, dfyear$YearBuilt)

# taxable current year and crime

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Arrests) # p-value > 0.05

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Arson) # p< 0.05

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Assault) # p>0.05

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Burglary) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Drug) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Fraud) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Harassment) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Homicide) #<

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Intimidation) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$LiquorLaw) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Robbery) # <

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Theft) #

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Traffic) #

cor.test(dfnumerical$TaxableValueCurrentYear, dfnumerical$Vandalism) # <

# Taxable prior year and crime

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Arrests) # >

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Arson)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Assault) # >

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Burglary)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Drug) #>

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Fraud)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Harassment)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Homicide)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Intimidation)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$LiquorLaw)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Robbery) # >

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Theft)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Traffic)

cor.test(dfnumerical$TaxableValuePriorYear, dfnumerical$Vandalism)

```

- taxable value of house current and prior year have strong correlation with number of crime occured except Arrest, Assault and Drug

```{r}

#crime attribute

dfcrim <- df %>% select("Arrests","Arson","Assault","Burglary","Drug","Fraud","Harassment","Homicide","Intimidation","LiquorLaw","Robbery","Theft","Traffic","Vandalism")

#correlation matrix of crime attribute

cor(dfcrim,use ='complete.obs')

# Type of crime

crimetype <- c("Arrests","Arson","Assault","Burglary","Drug","Fraud","Harassment","Homicide","Intimidation","LiquorLaw","Robbery","Theft","Traffic","Vandalism")

```