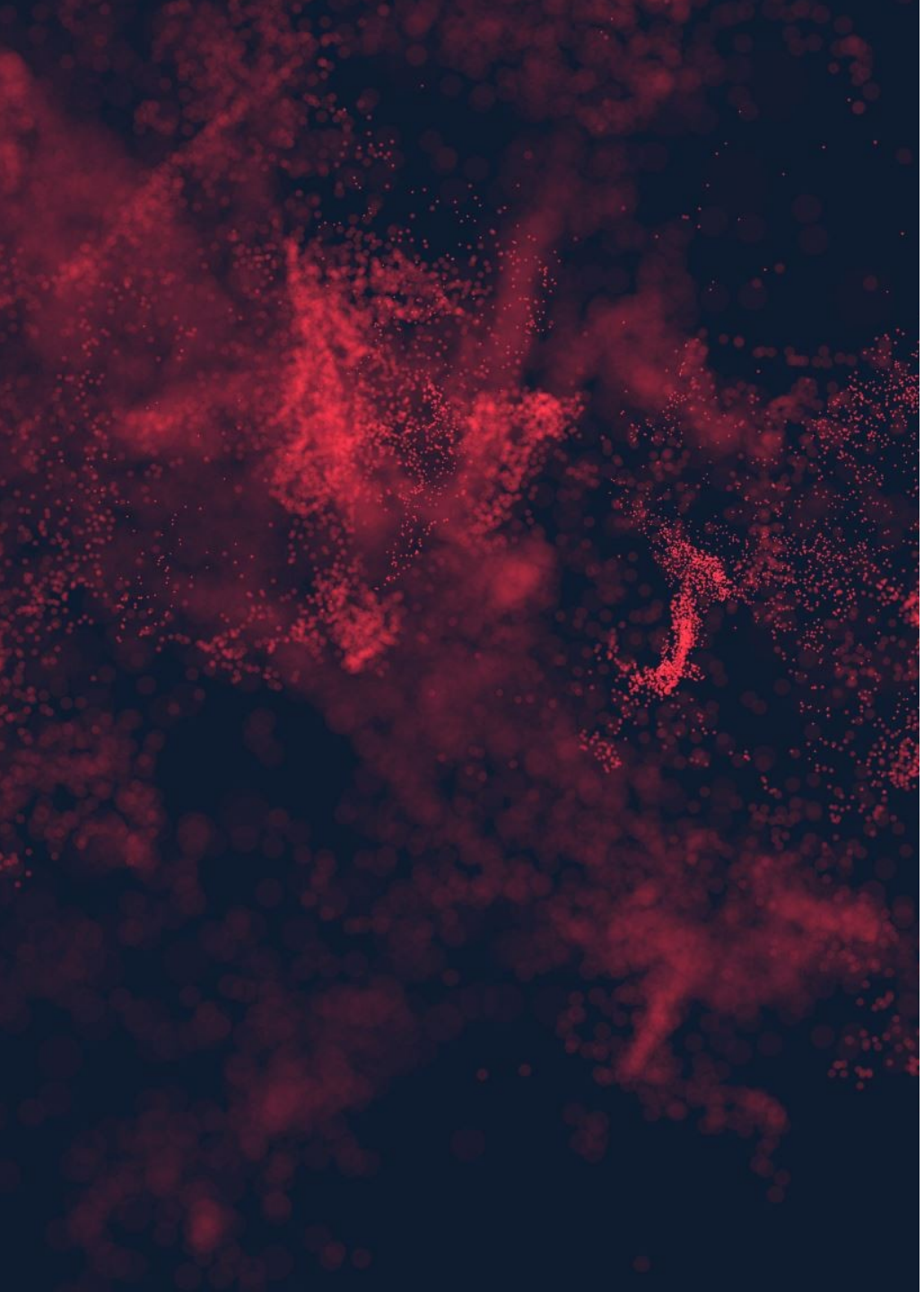
COVID19 DEATH RATE ANALYSIS USING DATA MINING TECHNIQUES

AKSHAY

JOSEPH



BACKGROUND

- **US among worst affected countries.**
- **Jan 21st first case reported.**
- **Mar 13th declared as national emergency.**
- **Lock down and other strict measures imposed.**
- **More than 15 million affected in US today.**

Source : https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165

Source: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1?

DATASET

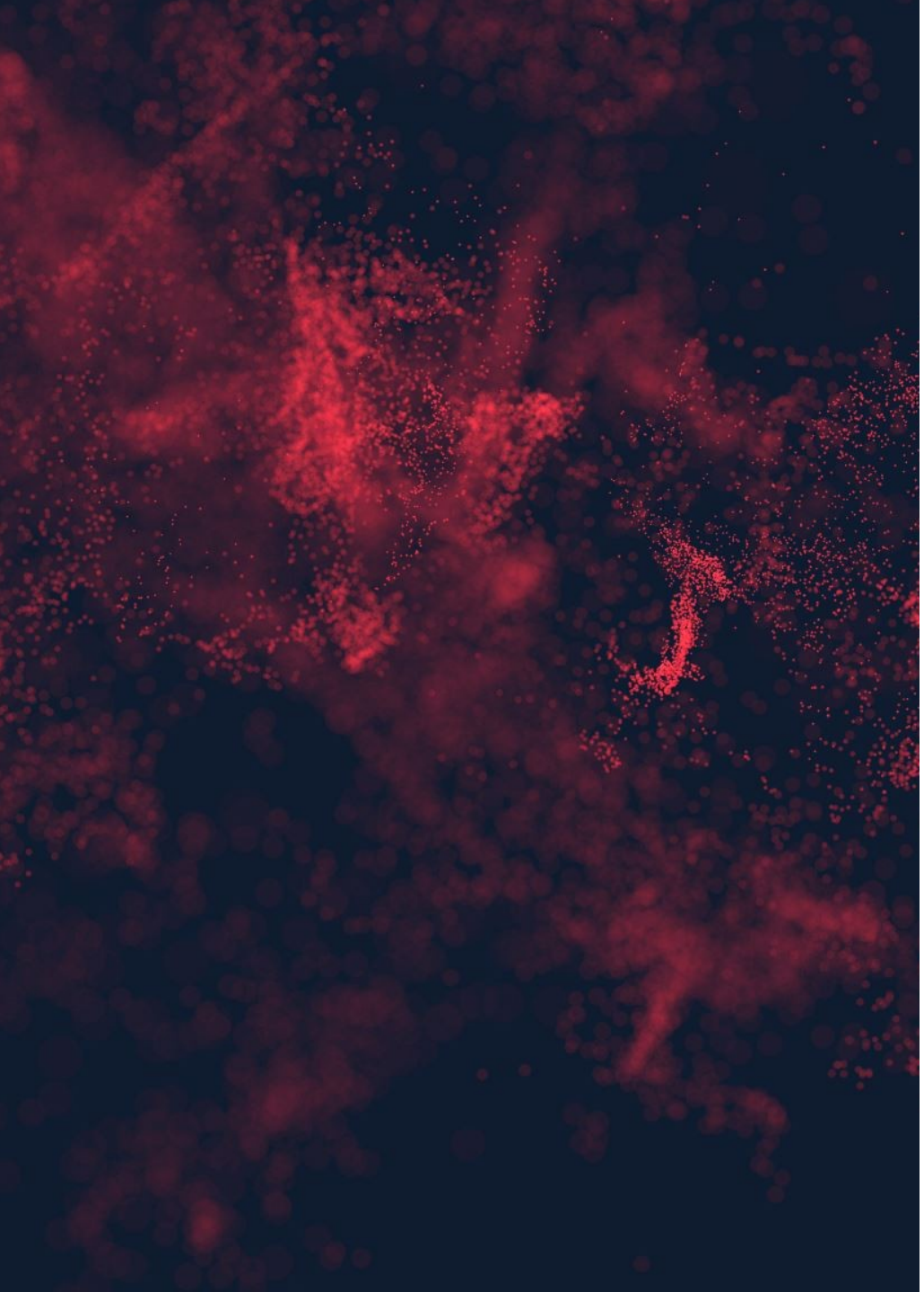
FEATURES

- NCHS urbanization (Noncore, Small Metro, ….., Metropolitan)
- Total Population
- Confirmed cases

RESPONSE VARIABLE

- Death rate (Low, Medium and High)

state	county_name	fips_code	lat	lon	NCHS_urbanization	total_population	confirmed	confirmed_per_100000	deaths	deathrate	classification
Alabama	Autauga	1001	32.53953	-86.64408	Medium metro	55200	2506	4539.86	37	67.030	low
Alabama	Baldwin	1003	30.72775	-87.72207	Small metro	208107	7772	3734.62	84	40.360	low
Alabama	Barbour	1005	31.86826	-85.38713	Non-core	25782	1134	4398.42	9	34.910	low
Alabama	Bibb	1007	32.99642	-87.12511	Large fringe metro	22527	1004	4456.87	17	75.460	medium



MORE ABOUT DATASET

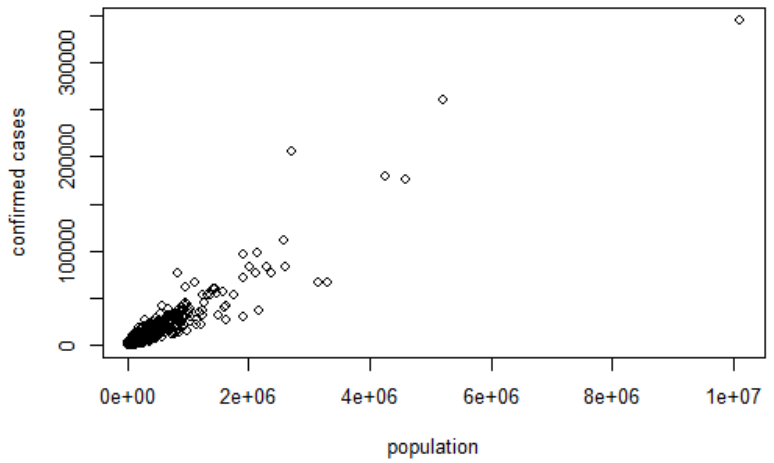
- **Many missing values**
- **Death rate the response variable to be converted into factors**
- **Dataset contains many info that was not necessary for this analysis**

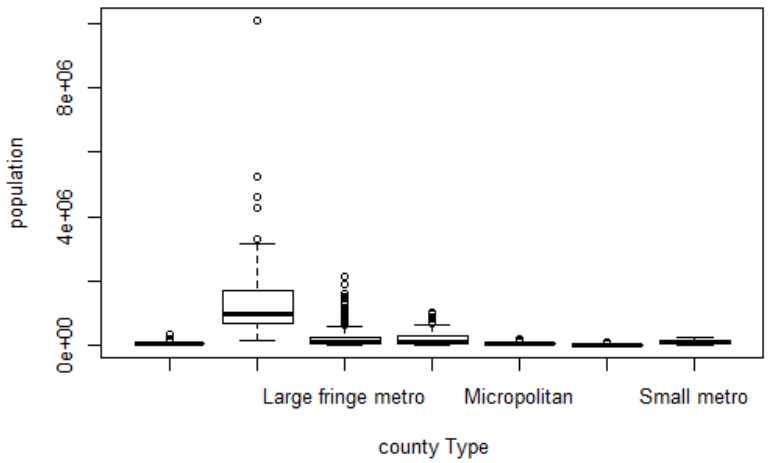
1. C0NFIRMED CASES AND POPULATION PLOT

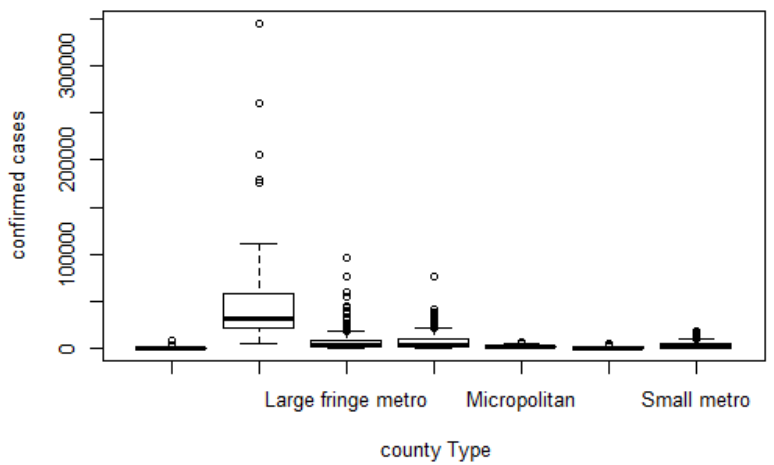
2. BOX PLOT OF POPULATION IN ALL REGION TYPE

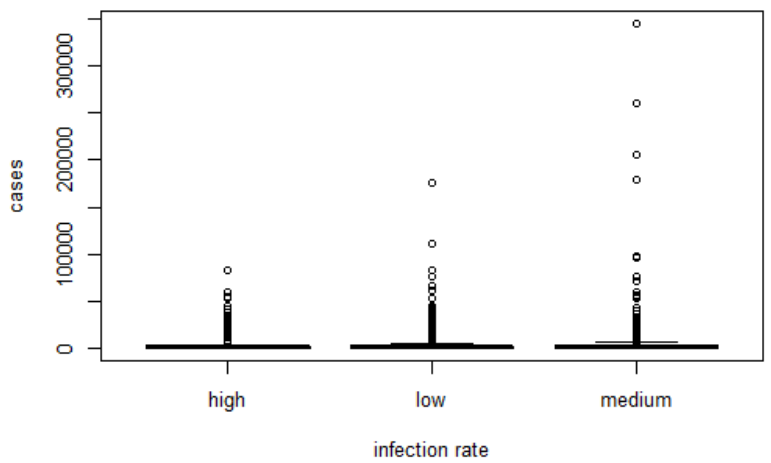
3. BOX PLOT OF CONFIRMED CASES IN ALL REGION TYPE

4. BOX PLOT OF COFIRMED CASES IN ALL CATEGORY OF DEATH RATE

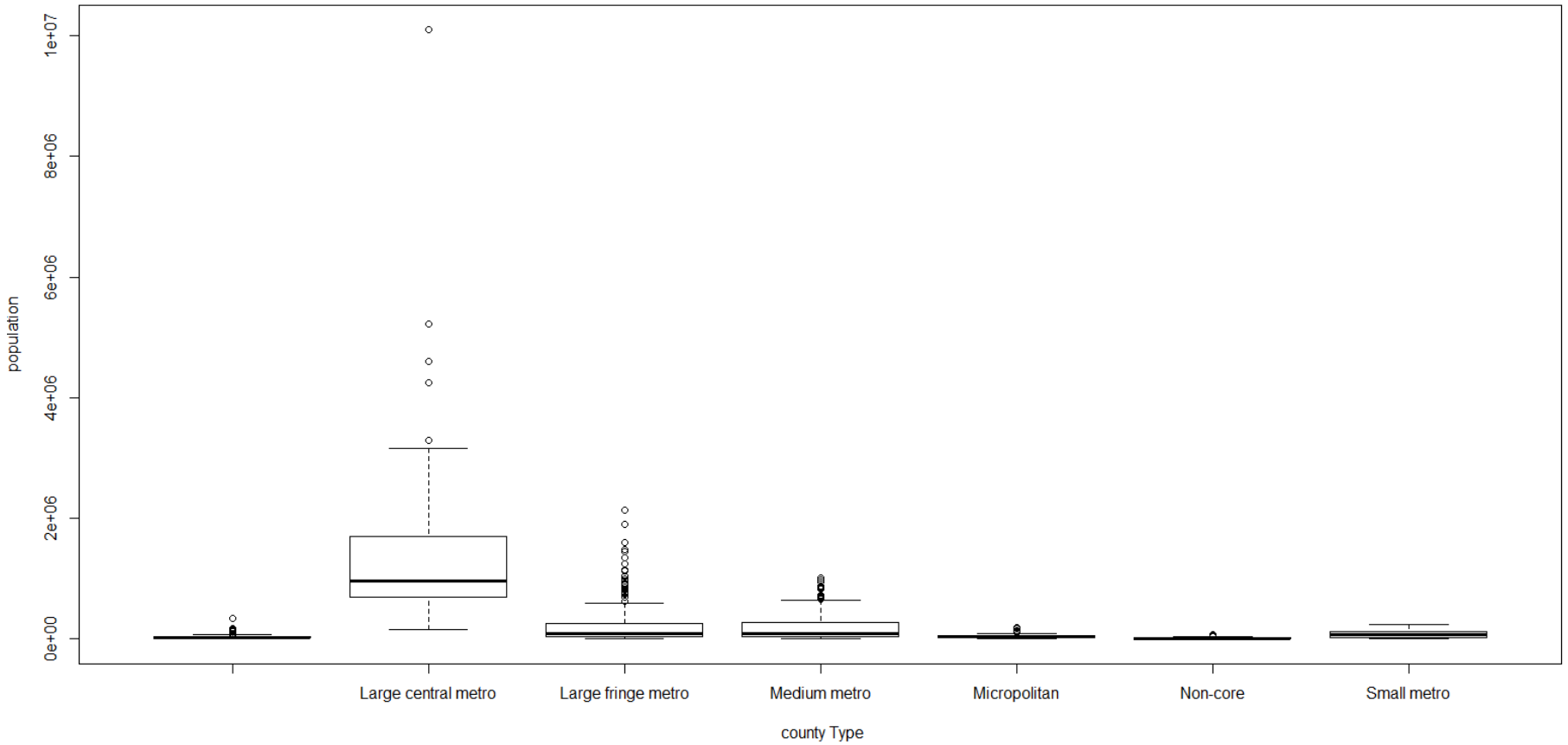




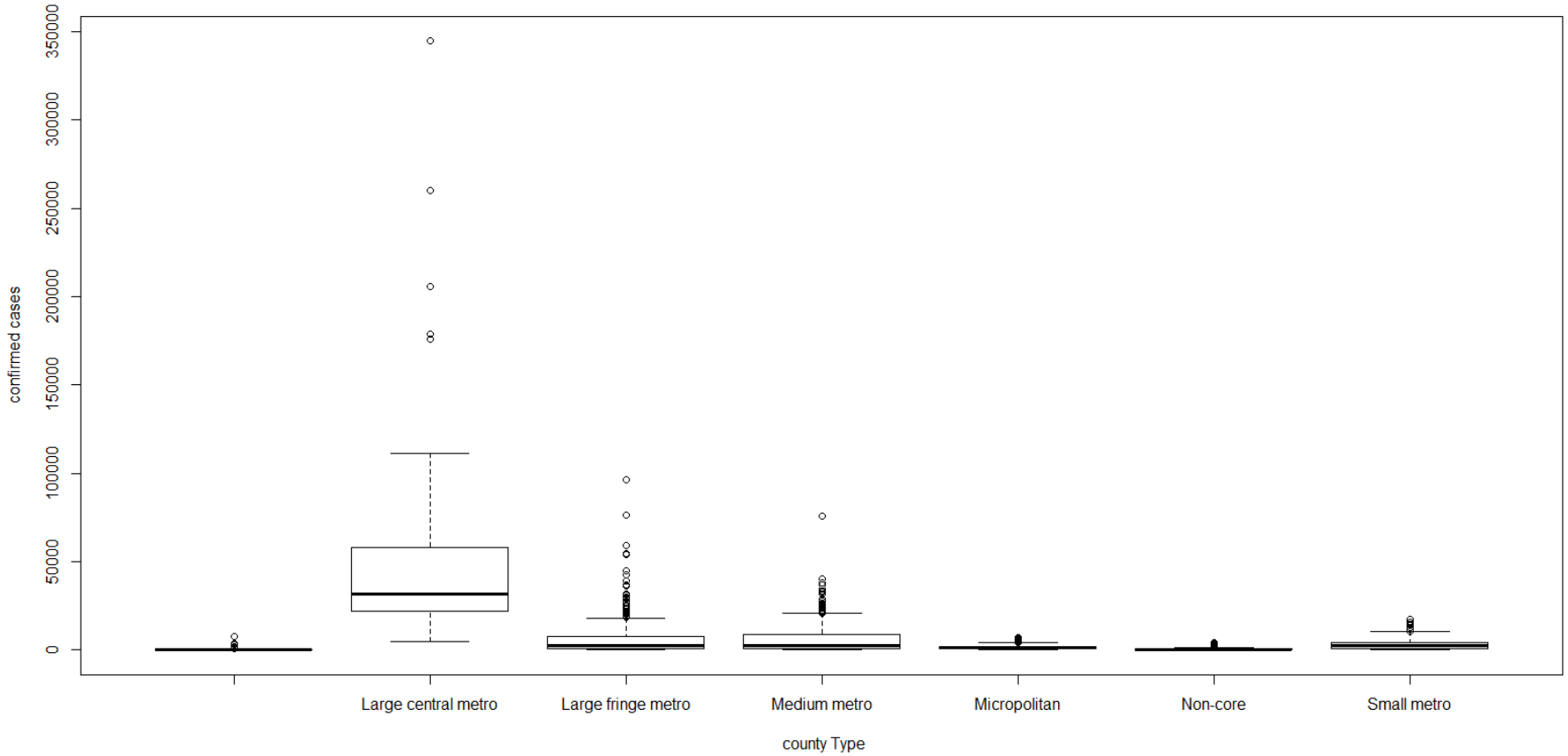




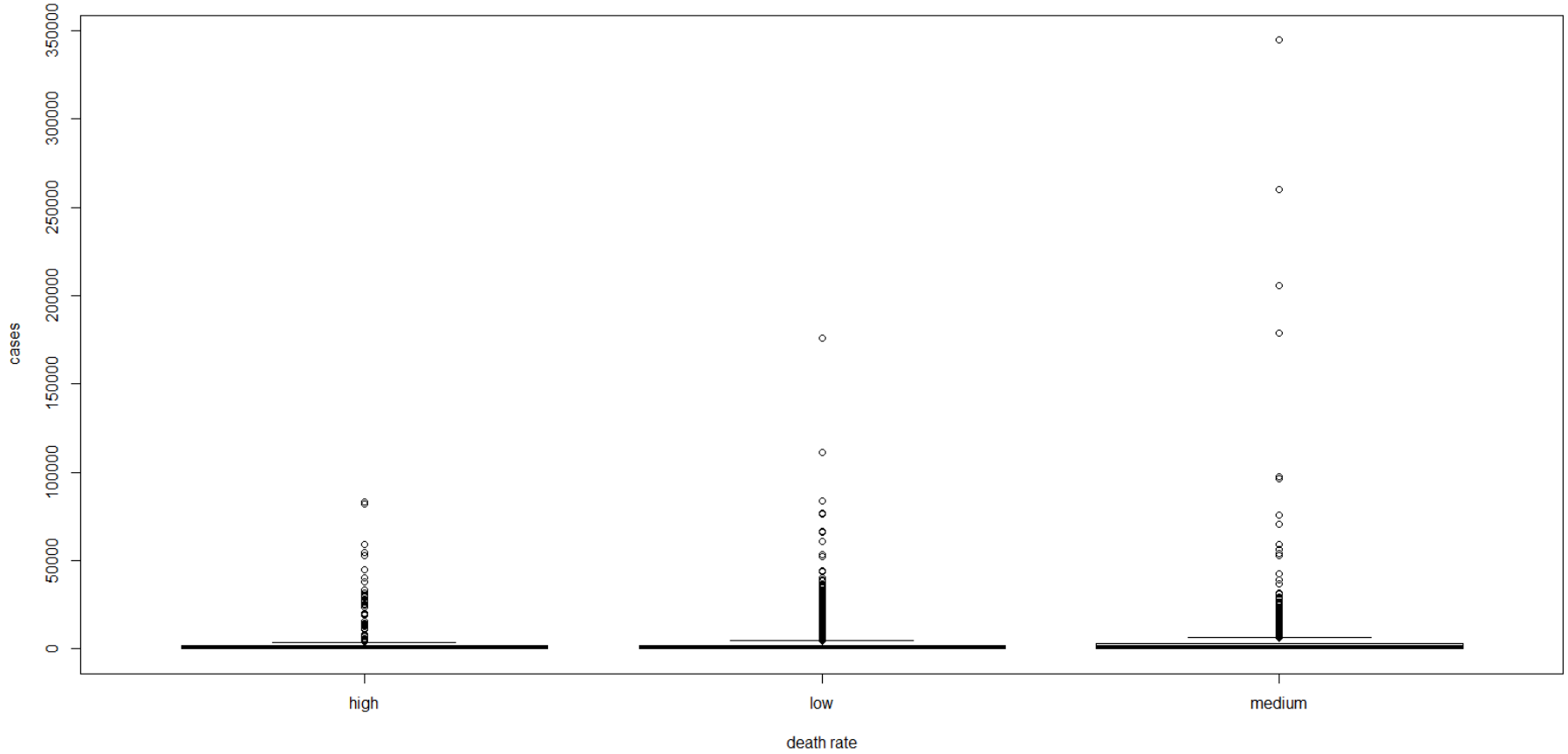
BOX PLOT OF POPULATION RATE IN ALL REGION



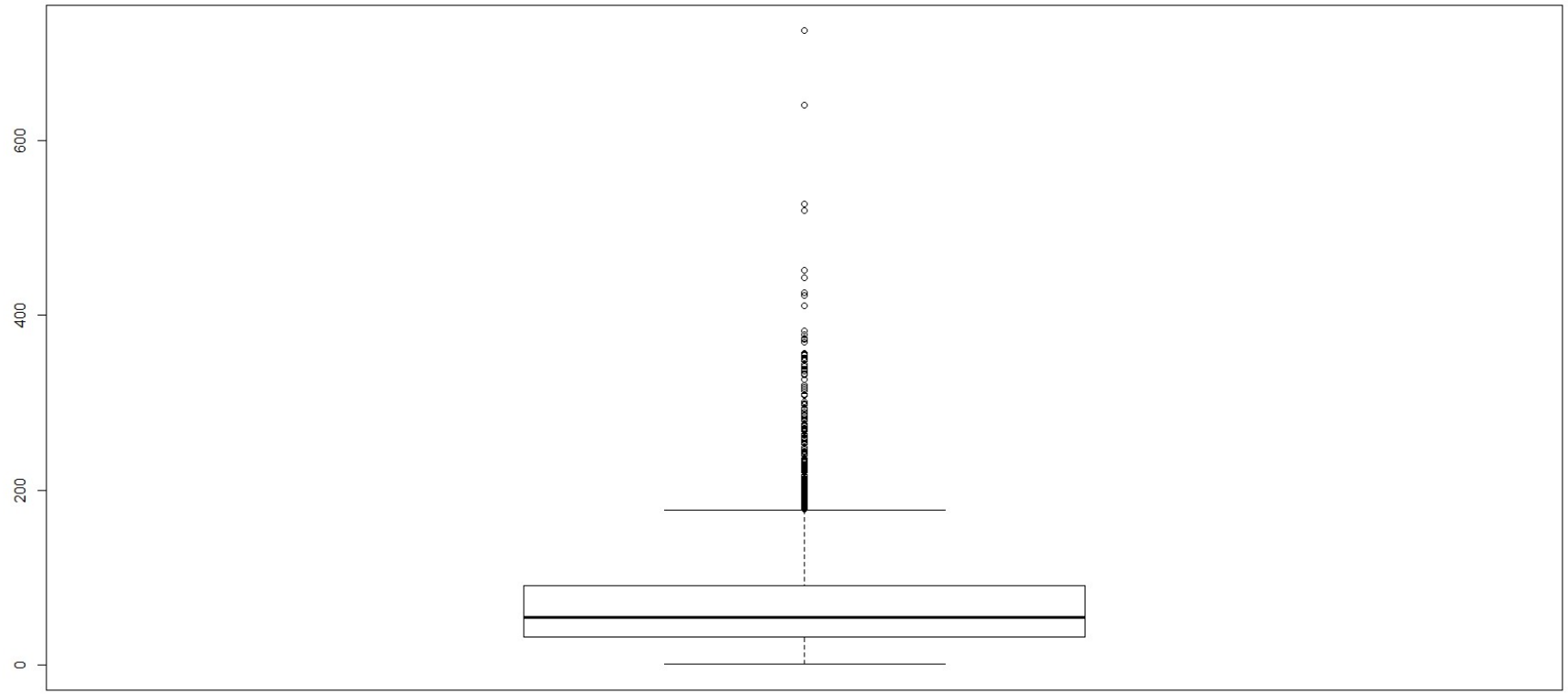
BOX PLOT OF CONFIRMED CASES RATE IN ALL REGION



BOX PLOT OF CONFIRMED CASES RATE AGAINST DEATH RATE



BOX PLOT OF DEATH RATE IN ALL REGION



DATA PREPROCESSING

TAKING CARE OF MISSING VALUE

- Median imputation on missing value
- apply(data, 2, function(x) any(is.na(x)))

state	county_name	fips_code	lat	lon
FALSE	FALSE	FALSE	TRUE	TRUE
NCHS_urbanization	total_population	confirmed	confirmed_per_100000	deaths
FALSE	TRUE	FALSE	TRUE	TRUE
deathrate				
TRUE				

DATA PREPROCESSING

CATEGORIZE NUMERIC RESPONSE VARIABLE

- Using KNN cluster centroid
- Categorized into 3 clusters (Low, Medium and High)
- data$classification = ifelse(data$deathrate<=70,'low',

  ifelse(data$deathrate >71&data$deathrate< 150 ,'medium','high')

)

```
> centroids <- as_tibble(km$centers, rownames = "cluster")
> centroids
# A tibble: 3 x 2
  cluster    V1
  <chr>   <dbl>
1 1       276.
2 2        42.2
3 3       124.
```

RULE BASED CLASSIFIER

- Class decision based on if .. then rule
- IF part of the rule is Rule antecedent
- THEN part of the rule is Rule consequent

- rulesFit <- data2 %>% train(data.classification ~ .,

  method = "PART“, data = .,

  tuneLength = 5,na.action = na.pass,

  trControl = trainControl(method = "cv", indexOut = train)

RULE-BASED CLASSIFIER

```
Rule-Based Classifier

3196 samples
   3 predictor
   3 classes: 'high', 'low', 'medium'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2877, 2876, 2877, 2877, 2875, 2877, ...
Resampling results across tuning parameters:

  threshold  pruned  Accuracy  Kappa
  0.010      yes     0.659     0.1023
  0.010      no      0.654     0.0863
  0.133      yes     0.658     0.1168
  0.133      no      0.654     0.0863
  0.255      yes     0.659     0.1107
  0.255      no      0.654     0.0863
  0.378      yes     0.661     0.1159
  0.378      no      0.654     0.0863
  0.500      yes     0.661     0.1159
  0.500      no      0.654     0.0863

Accuracy was used to select the optimal model using the largest value.
```

CLASSIFICATION WITH DECISION TREE

```
> rulesFit$results
   threshold pruned Accuracy  Kappa AccuracySD KappaSD
1      0.010    yes    0.659 0.1023    0.01565  0.0635
2      0.010     no    0.654 0.0863    0.00861  0.0575
3      0.133    yes    0.658 0.1168    0.00994  0.0519
4      0.133     no    0.654 0.0863    0.00861  0.0575
5      0.255    yes    0.659 0.1107    0.00983  0.0564
6      0.255     no    0.654 0.0863    0.00861  0.0575
7      0.378    yes    0.661 0.1159    0.00941  0.0518
8      0.378     no    0.654 0.0863    0.00861  0.0575
9      0.500    yes    0.661 0.1159    0.00941  0.0518
10     0.500     no    0.654 0.0863    0.00861  0.0575
```

CLASSIFICATION WITH DECISION TREE

- Uses binary recursive partitioning
- Splits data into partitions to create branches
- Consists of Root Nodes, Internal Node and Leaf Node

CLASSIFICATION WITH DECISION TREE

```
> C45Fit
C4.5-like Trees

3196 samples
   3 predictor
   3 classes: 'high', 'low', 'medium'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2876, 2877, 28
Resampling results across tuning parame

  C      M  Accuracy  Kappa
  0.010  1  0.645     0.0000
  0.010  2  0.645     0.0000
  0.010  3  0.645     0.0000
  0.010  4  0.645     0.0000
  0.010  5  0.645     0.0000
  0.133  1  0.655     0.0599
  0.133  2  0.654     0.0583
  0.133  3  0.654     0.0558
  0.133  4  0.652     0.0419
  0.133  5  0.650     0.0438
  0.255  1  0.682     0.2126
  0.255  2  0.679     0.2087
  0.255  3  0.672     0.1722
  0.255  4  0.669     0.1678
  0.255  5  0.664     0.1587
  0.378  1  0.689     0.2456
  0.378  2  0.686     0.2344
  0.378  3  0.680     0.2185
  0.378  4  0.677     0.2140
  0.378  5  0.672     0.2010
  0.500  1  0.691     0.2549
  0.500  2  0.684     0.2407
  0.500  3  0.679     0.2239
  0.500  4  0.674     0.2156
  0.500  5  0.672     0.2059

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.5 and M = 1.
```

RANDOM FOREST

- Create random vectors from Training data
- Build multiple tree using random vector
- Combine the trees
- Might take more processing time

RANDOM FOREST

```
Random Forest

3196 samples
   2 predictor
   3 classes: 'high', 'low', 'medium'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2876, 2876, 2876, 2877, 2877, 2878, ...
Resampling results:

  Accuracy  Kappa
  0.955     0.911

Tuning parameter 'mtry' was held constant at a value of 2
```

MODEL EVALUATION

```
Call:
summary.resamples(object = resamps)

Models: C45, rules, randomForest
Number of resamples: 10

Accuracy
             Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
C45          0.659   0.678  0.683 0.682   0.687 0.703    0
rules        0.646   0.657  0.660 0.661   0.666 0.675    0
randomForest 0.938   0.944  0.956 0.955   0.966 0.972    0

Kappa
               Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
C45          0.1569  0.1910  0.210 0.220   0.249 0.286    0
rules        0.0367  0.0833  0.111 0.116   0.143 0.218    0
randomForest 0.8757  0.8905  0.912 0.911   0.932 0.944    0
```

THANK YOU