Death in the United States

1. Introduction

Every year the **CDC** releases the country's most detailed report on death in the United States under the **National Vital Statistics Systems**. This mortality dataset is a record of every death in the country for the year 2014.

Mortality data from the NVSS are a fundamental source of demographic, geographic, and cause-of-death information. This is one of the few sources of health-related data that are comparable for small geographic areas and are available for a long time period in the United States.

Analyzing mortality data is essential to understanding the complex circumstances of death across the country. The US Government uses this data to determine life expectancy and understand how death in the U.S. differs from the rest of the world.

2. Overview on Dataset:

This dataset is a collection of tables and is available in CSV. Each row in the DeathRecords table is an individual death record. Each death record has a one-to-many relationship with the EntityAxisConditions and RecordAxisConditions tables via a DeathRecordId key. Both of these conditions tables contain ICD-10 codes that indicate cause of death for each person.

DeathRecords file

Primary table containing a single row per death record with these columns:

- **Id** (*integer primary key*) - Main identifier, used for joining with DeathRecordId in EntityAxisConditions and RecordAxisConditions tables.
- **ResidentStatus** (*integer*) - (e.g. 1 = Residents, 2 = Intrastate resident, etc)
- **Education1989Revision** (*integer*) - Years of education using the 1989 revision format (e.g. 8 = 8 years of elementary education)
- **Education2003Revision** (*integer*) - Years of education using the 2003 revision code (e.g. 8 = Doctorate or professional degree)
- **EducationReportingFlag** (*integer*) - (0 = 1989 revision was used on death certificate, 1 = 2003 revision was used)
- **MonthOfDeath** (*integer*) - Month of death (e.g. 1 = January, 12 = December)
- **Sex** (*text*) - (M = Male, F = Female)
- **AgeType** (*integer*) - Units for the **Age** column (e.g. 1 = Years, 2 = Months)
- **Age** (*integer*) - Age at death (in **AgeType** units)
- **AgeSubstitutionFlag** (*integer*) - (1 = Calculated age is substituted for reported age)
- **AgeRecode52** (*integer*) - **Age** recoded into 52 bins (e.g. 1 = Under 1 hour)
- **AgeRecode27** (*integer*) - **Age** recoded into 27 bins (e.g. 1 = Under 1 month)
- **AgeRecode12** (*integer*) - **Age** recoded into 12 bins (e.g. 1 = Under 1 year)

- **InfantAgeRecode22** (*integer*) - In the event of an infant, **Age** recoded into 22 bins (e.g. 1 = Under 1 hour)
- **PlaceOfDeathAndDecedentsStatus** (*integer*) - (e.g. 6 = Nursing home/long term care)
- **MaritalStatus** (*text*) - (e.g. M = married, D = divorced, W = widowed)
- **DayOfWeekOfDeath** (*text*) - (e.g. 1 = Sunday, 7 = Saturday)
- **CurrentDataYear** (*text*) - Year on death record. Always 2014 for this dataset.
- **InjuryAtWork** (*text*) - Was the person injured at work? (Y = yes, N = no, U = unknown)
- **MannerOfDeath** (*integer*) - (e.g. 1 = Accident, 2 = Suicides)
- **MethodOfDisposition** (*text*) - (e.g. B = burial, C = cremation)
- **Autopsy** (*text*) - Was an autopsy performed? (Y = Yes, N = No, U = Unknown)
- **ActivityCode** (*integer*) - (e.g. 0 = While engaged in sports activity, 1 = While engaged in leisure activity)
- **PlaceOfInjury** (*integer*) - (e.g. 0 = Home, 1 = Residential institution)
- **Icd10Code** (*text*) - ICD-10 code for the underlying cause of death (e.g. I251 = Atherosclerotic heart disease)
- **CauseRecode358** (*integer*) - Cause of death recoded into 358 bins
- **CauseRecode113** (*integer*) - Cause of death recoded into 113 bins
- **InfantCauseRecode130** (*integer*) - Infant cause of death recoded into 130 bins
- **CauseRecode39** (*integer*) - Cause of death recoded into 39 bins
- **NumberOfEntityAxisConditions** (*integer*) - Number of entries for this death record in the EntityAxisConditions table
- **NumberOfRecordAxisConditions** (*integer*) - Number of entries for this death record in the RecordAxisConditions table
- **Race** (*integer*) - Reported race (e.g. 1 = White, 2 = Black)
- **BridgedRaceFlag** (*integer*) - (e.g. 1 = Race is bridged)
- **RaceImputationFlag** (*integer*) - (e.g. 1 = Unknown race is imputed)
- **RaceRecode3** (*integer*) - **Race** recoded into 3 bins (e.g. 2 = Races other than White or Black)
- **RaceRecode5** (*integer*) - **Race** recoded into 5 bins (e.g. 4 = Asian or Pacific Islander)
- **HispanicOrigin** (*integer*) - (e.g. 220 = Central and South American)
- **HispanicOriginRaceRecode** (*integer*) - **HispanicOrigin** / **Race** recoded (e.g. 1 = Mexican)

Lookup Table

There are many columns in the DeathRecords table that contain various codes (e.g. "Education1989Revision", "Education2003Revision", etc.). I have provided lookup tables which I used in analysis.

- MannerOfDeath
- Education1989Revision
- Education2003Revision
- MaritalStatus
- RaceRecode3
- Race
- Icd10Code

3. Download and read the dataset:

Data is downloaded from **here.** The dataset contains many files in the '.csv' format. It also has a SQLLite format of data. I am using '.csv' format for my analysis. Below are the list of few files which I am reading into RStudio.
"DeathRecords.csv" has 2.6 million observation, so while reading the file I am using "fread" function from data.table package to load the data fast.

```
Death_US <- fread("DeathRecords.csv", header = T)
MannerOfDeath <- read.csv("MannerOfDeath.csv", header = T)
Edu_1989 <- read.csv("Education1989Revision.csv", header = T)
Edu_2003 <- read.csv("Education2003Revision.csv", header = T)
MaritalStatus <- read.csv("MaritalStatus.csv", header = T)
Race3 <- read.csv("RaceRecode3.csv", header = T)
RaceAll <- read.csv("Race.csv", header = T)
icd10_code <- fread("Icd10Code.csv")
```

4. Key Finding of Exploratory Data Analysis:

Performed EDA on the dataset, below are the key finding from the analysis. The detailed report on EDA can be fount **here**.

1. Life expectancy of the population is 73.4 years.
2. Female population group has more life expectancy (76.6 years) than the male population group (70.2 years)
3. As per research from 1999 through 2014, the age-adjusted suicide rate in the United States increased 25%, from 10.5 to 13.5 per 100,000 population, with the pace of increase greater after 2006.
4. Male and Female Suicidal Death Ration :: Majority of suicidal cases belong to male, they represent the 77.34% of the total suicidal cases.
5. 50% of suicidal cases are be discharging the firearm.
6. In the given dataset, 85% of the death observations belong to white, about 12% are black, and the other races each account for under %1 of the total deaths recorded.
7. By analyzing the Homicide cases wrt to Race, found out that homicide death count in Black race is 6 times more than the homicide death count in White race.
8. Death rate per month analysis gave a result that, there are seasonal fluctuations in U.S. deaths. One's chances of dying in the winter months are significantly greater than in the summer. This is a statistical fact. It is generally true regardless of where one lives in the U.S.
9. 69% of total accidental death happens in "AgeGroup > 40". But "AgeGroup > 40" has a very large population. So after performing analysis of accidental death percentage for each age group, we can say younger group show the more accidental death rate(40%).

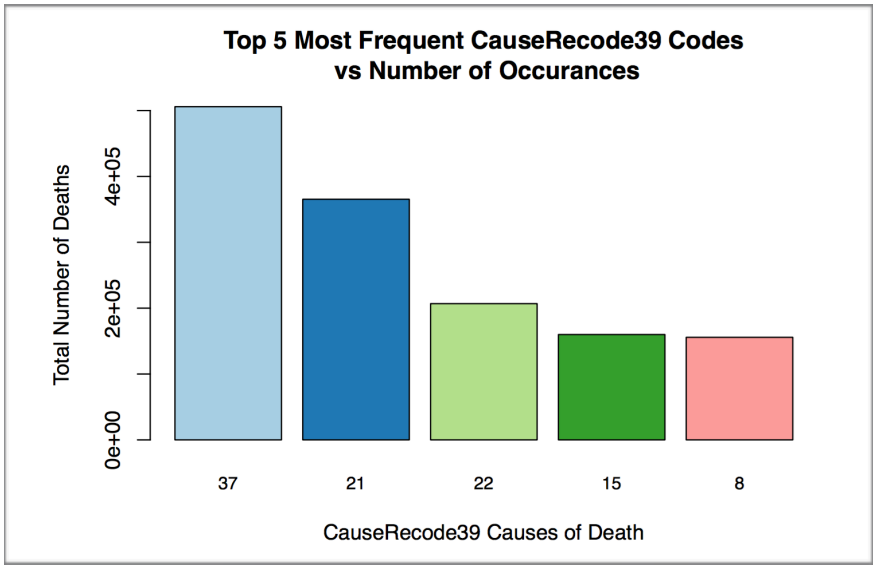
10. **ICD-10** is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD). Below is the list of Top10 ICD10 codes which caused the death in U.S.

Icd10Code	Description
I251	Atherosclerotic heart disease
C349	Malignant neoplasm: Bronchus or lung, unspecified
F03	Unspecified dementia
I219	Acute myocardial infarction, unspecified
J449	Chronic obstructive pulmonary disease, unspecified
G309	Alzheimer disease, unspecified
I64	Stroke, not specified as haemorrhage or infarction
I250	Atherosclerotic cardiovascular disease, so described
I500	Congestive heart failure
J189	Pneumonia, unspecified

5. Feature Selection for the prediction model:

Predicting the "CauseRecode39" which is the cause of death recoded into 39 bins. before selecting the input variable, lets check the Top5 CauseRecode39 codes in the given dataset. All the code and description can be found **here**.

NOTE: Considering the only Natural death records for the model.



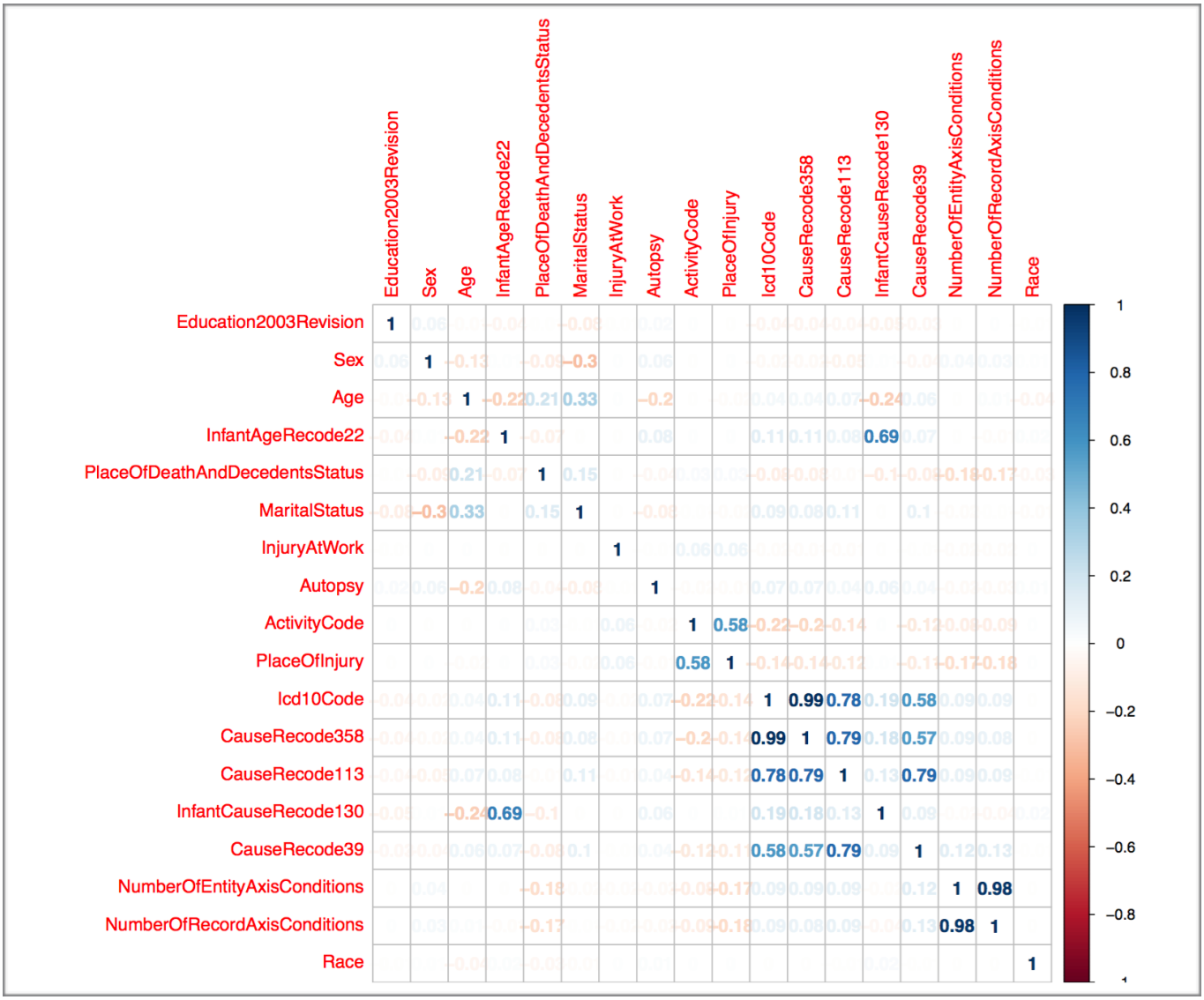
CODES

- **37 ::** All other diseases (Residual) (A00-A09,A20-A49,A54-B19,B25-B99,D00-E07, E15-G25,G31-H93,I80- J06,J20-J39,J60-K22,K29-K66,K71-K72, K75-M99,N10-N15,N20-N23,N28-N98,U04)
- **21 ::** Ischemic heart diseases (I20-I25)

- **22 ::** Other diseases of heart (I00-I09,I26-I51)
- **25 ::** Other malignant neoplasms (C00-C15,C17,C22-C24,C26-C32,C37-C49,C51-C52,C57-C60,C62-C63,C69- C81,C88,C90,C96-C97)
- **8 ::** Malignant neoplasms of trachea, bronchus and lung (C33-C34)

Find the Correlation

As a first step of finding feature variables, lets find the correlation between variables.



As we know “CauseRecord358”, “CauseRecord113”, “CauseRecord39” are the recoded bin of ICD10 code. If we use these variable it will cause multicollinearity. So we can not use “CauseRecord358”, “CauseRecord113” and “ICD10Code” as the model variable.

From the above visual correlation matrix we can say "Age, InfantAgeRecode22, PlaceOfDeathAndDecedentsStatus, MaritalStatus, ActivityCode, PlaceOfInjury,

NumberOfRecordAxisConditions and NumberOfEntityAxisConditions" variables has a correlation with "CauseRecord39", therefore selecting these variables for the model.
Detailed feature selection can be found **here**.

6. Model 1: Decision Tree

Since CauseRecord39 is a categorical variable using decision tree for the model. decision tree support the tree structure unto 32 levels, but CauseRecord39 outcome variable had a 39 levels. Therefore removed 7 levels which had less number of observations (In total removed 1051 observation out of 2059933 observations). Dived the train and test set into 70:30 ratio respectively.

Model code is as below:

```
model_tree <- tree(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                     PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                      PlaceOfInjury + NumberOfRecordAxisConditions +
                      NumberOfEntityAxisConditions, train)
```

Predicted the test dataset:

```
# Predict the test dataset using model
predict_ICD2 <- predict(model_tree, newdata = test, type = "class")
# confusion matrix
conf_matrix2 <- table(predict_ICD2, Cause39)
```

Model Accuracy

```
# accuracy
sum(diag(conf_matrix2)) / nrow(test)
```

```
## [1] 0.2143631
```

As we can see the model accuracy is 21.4% which is very low. Lets check on Random Forest Model.
Detailed modeling can be found **here**.

7. Model 2 : Random Forest

As mentioned earlier in Decision Tree model, removed 7 levels from CauseRecord39 variables and used 70:30 standard train & test dataset ration for model building.

Model code is as below:

```
model_forest1 <- randomForest(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                     PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                       PlaceOfInjury + NumberOfRecordAxisConditions +
                       NumberOfEntityAxisConditions,
                     data = train[1:600000, ], nodesize = 25, ntree = 1501)
```

```
model_forest2 <- randomForest(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfRecordAxisConditions +
                        NumberOfEntityAxisConditions,
                      data = train[600001:1200000, ],
                      nodesize = 25, ntree = 1501)

model_forest <- combine(model_forest1,model_forest2)
```

Predicted the test dataset:

```
# Predict the test dataset using random forest model
predict_forest <- predict(model_forest, newdata = test)
# confusion matrix
conf_matrix <- table(predict_forest, Cause39)
```

Model Accuracy

```
sum(diag(conf_matrix)) / nrow(test)
```

```
## [1] 0.2590312
```

As we can see the model accuracy is 25.9% which is better than Decision Tree model, even though the model accuracy is less. Lets check on Gradient Boot Strapping.
Detailed modeling can be found **here**.

8. Model 3: Gradient Boot Strapping:

As mentioned earlier in Decision Tree model, removed 7 levels from CauseRecord39 variables and used 70:30 standard train & test dataset ration for model building.

Model code is as below:

```
gbm2 <- gbm(as.factor(CauseRecode39) ~ Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfRecordAxisConditions +
                        NumberOfEntityAxisConditions,
                    data = train,
             var.monotone=c(0,0,0,0,0,0,0,0),
                    # +1: monotone increase,
                    #  0: no monotone restrictions
                    distribution="gaussian",     # bernoulli, adaboost, gaussian,
                    # poisson, coxph, and quantile available
                    n.trees=3000,                # number of trees
                    shrinkage=0.005,             # shrinkage or learning rate,
                    # 0.001 to 0.1 usually work
                    interaction.depth=3,         # 1: additive model, 2: two-way interactions, etc.
                    bag.fraction = 0.5,          # subsampling fraction, 0.5 is probably best
                    n.minobsinnode = 10,         # minimum total weight needed in each node
                    cv.folds = 5,                # do 5-fold cross-validation
                    keep.data=TRUE,              # keep a copy of the dataset with the object
                    verbose=T )
```

Predicted the test dataset:

```
data.predict = predict(gbm2, n.trees = best.iter, newdata = test)
# Confusion matrix
conf_matrix <- table(data.predict, Cause39)
```

Accuracy of model and SSE

```
#Accuracy
sum(diag(conf_matrix)) / nrow(test)
```

```
## [1] 4.695087e-05
```

```
# SSE
SSE = sum((Cause39 - data.predict)^2)
print(SSE)
```

```
## [1] 65362296
```

Model accuracy is very less and its showing high SSE, so we cannot consider this prediction model.
Detailed modeling can be found **here**.

9. Tuning the Random Forest Prediction Model.

From the above work we have 2 model with accuracy around 25%. Since Random Forest has comparatively high accuracy, selecting Random Forest over Decision Tree for model tuning.

Below are the points considered for model tuning:

1. Increasing the count of observations in training data set, accuracy of the model will improve as count of train dataset observation increases. Using 89:11 ration for test and train data set respectively.
2. Not removing the 7 levels from CauseRecord39 variable.
3. Dividing the CauseRecord39 variable into 3 subgroup.
4. Converted all the input variables into factors.
5. Do not filter only natural deaths, take complete death record dataset for modeling.

```
> table(factor(natural_sub$CauseRecode39))

     1      2      3      5      6      7      8      9     10     11     12     13     14
   366     37   5617   9053  43839  33846 133409  34620  23359  23421  25734  17115  19670
    15     16     17     20     21     22     23     24     25     26     27     28     29
133271  63719  75547  37411 310804 175736  23701 111658   5426  16550  45795 125748   2519
    30     31     32     33     34     35     36     37     38     39     40     41     42
 31593  41366   1000   9779   8074    414  23026 433056    212  13088      8      5      9
```

Check the above result of table:

- Few levels are having very less number of observations (ex: 37,38,39..) i.e less than 10,000.
- Couple of levels are having very large number of observations (ex: 8,15,21) i.e more than 100,000
- remains levels are having observations between 10,000 to 100,000 (ex: 6,7,9)

So, dividing the CauseRecord39 variables into groups based on the number of observations belong to each levels.

Detailed code can be found **here**.

Group 1 :: Model for the data with CauseRecode39 having entries less than 10000

```
model_forest <- randomForest(as.factor(CauseRecode39) ~  Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfEntityAxisConditions ,
                      data = train,
                      nodesize = 25, ntree = 1501)
```

Predicted the test dataset, confusion matrix & Accuracy:

```
> # Predict the test dataset using random forest model
> predict_forest <- predict(model_forest, newdata = test)
> # confusion matrix
> conf_matrix <- table(predict_forest, Cause39)
> conf_matrix
              Cause39
predict_forest    1    2    3    5   25   29   32   33   34   35   38   40   41   42
            1     0    0    0    0    0    0    0    0    0    0    0    0    0    0
            2     0    0    0    0    0    0    0    0    0    0    0    0    0    0
            3    17    1  545  130   23   71   44    4  185    0    0    2   14   18
            5    22    1  137  939  241  144   32    0  113    0    0    1    5    7
            25    5    3   11  129  404   34    0    0   29    0    0    0    2    0
            29    8    0    5   30   28   70    0    0   10    0    0    0    0    0
            32    0    0    2    4    0    1   18    0    3    0    0    0    0    1
            33    0    0    0    0    0    0    0 1222  337    3    0    0    0    0
            34    3    0   41   15    4   12   29   68  348   24    0    0    5   18
            35    0    0    0    0    0    0    0    2   22  143    0    0    0    0
            38    0    0    0    0    0    0    0    0    0    0 3826    9   23    2
            40    0    0    0    0    0    0    0    0    0    0   47 4371  932  407
            41    0    0    0    0    0    0    0    0    1    0   43  320  756   69
            42    0    0    2    0    0    3    1    0    8    0    8   13   10   44
> # accuracy
> sum(diag(conf_matrix)) / nrow(test)
[1] 0.7603692
```

Group 2 :: Model for the data with CauseRecode39 having entries more than 10000 and less than 100000

```
model_forest <- randomForest(as.factor(CauseRecode39) ~  Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfEntityAxisConditions ,
                      data = train,
                      nodesize = 25, ntree = 1501, na.action = na.omit)
```

Predicted the test dataset, confusion matrix & Accuracy:

```
> # Predict the test dataset using random forest model
> predict_forest <- predict(model_forest, newdata = test)
> # confusion matrix
> conf_matrix <- table(predict_forest, Cause39)
> conf_matrix
              Cause39
predict_forest    6    7    9   10   11   12   13   14   16   17   20   23   26   27   30   31   36   39
            6   931  741  894  620  207  376  170  173  109  107  306   37   30   63  331  221  286    1
            7   593  730  430  377  394  401  196  220   60  209  124   43   41   49  157  156  115    2
            9   120   79  179   97   17   50   33   42    7    2   19    1   13   57   54   29   69    0
            10    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
            11    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
            12    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
            13    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
            14    2    0    0    1    0    4    9   45    9    0    3    0    7   27    1    6    4    8
            16 1059  720  784  503  611  659  414  430 4602  859 1285 1061  566  750 1241  873   54  219
            17 1766 1270 1239  758 1355 1208  727  720 1830 8560 1658 1445  525 1337  376 1875 1658   48
            20  189  126  141  127   40   77   35   50  291   39  715  145  109   90  213  117  244   19
            23    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
            26    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
            27  632  481  401  291  413  384  435  618  809  488  333  454  643 3043  396 1550  414   78
            30  340  235  408  220   66  175  201  242  637   16  116  125  173  618 1380  437   46  109
            31    0    0    0    0    0    0    0    2    0    0    0    1    0    0    0    3    0    0
            36  121   70  114   95   17   26   26   42   68    8  195   15    7   51   59   31  498    2
            39    0    0    0    0    1    0    1    1    3    4    4    0    2    0    1    1    0 10917
> # accuracy
> sum(diag(conf_matrix)) / nrow(test)
[1] 0.3571001
```

Group 3 :: Model for the data with CauseRecode39 having entries more than 100000

```
model_forest1 <- randomForest(as.factor(CauseRecode39) ~  Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfEntityAxisConditions ,
                      data = train[1:600000, ],
                      nodesize = 25, ntree = 1501, na.action = na.omit)

model_forest2 <- randomForest(as.factor(CauseRecode39) ~  Age + InfantAgeRecode22 +
                      PlaceOfDeathAndDecedentsStatus + MaritalStatus + ActivityCode +
                        PlaceOfInjury + NumberOfEntityAxisConditions ,
                      data = train[600001:1200000, ],
                      nodesize = 25, ntree = 1501, na.action = na.omit)

model_forest <- combine(model_forest1,model_forest2)
```

Predicted the test dataset, confusion matrix & Accuracy:

```
> sum(diag(conf_matrix)) / nrow(test)
[1] 0.3214047
> predict_forest <- predict(model_forest_12, newdata = test)
> # confusion matrix
> conf_matrix <- table(predict_forest, Cause39)
> conf_matrix
               Cause39
predict_forest     8    15    21    22    24    28    37
            8   2283  1417  1026   449   351   859  1126
            15  1569  2697   726   584   291   364  1277
            21  4998  4321 17643  6270  2152  5218  9450
            22     3     9    44    84    19    13    86
            24     0     0     0     0     0     0     0
            28     0     0     0     0     0     0     0
            37  8279  9150 20754 15368 11851  9736 43712
> # accuracy
> sum(diag(conf_matrix)) / nrow(test)
[1] 0.360622
```

10. Final Result of Tuned RandomForest and Discussion

MODEL	SUBGROUP OF causeRecord39	ACCURACY
Model for the data with CauseRecode39 having entries less than 10000	Group 1	76.0%
Model for the data with CauseRecode39 having entries more than 10000 and less than 100000	Group 2	35.7%
Model for the data with CauseRecode39 having entries less than 100000	Group 3	36.0%

- From the above result we can say that model accuracy increased after tuning the model.
- In the first Random forest model few levels of causeRecord39 variable had more observations and few levels had very less observations. Splitting the data by considering the number of observations in each levels as a base line helps to find out better accuracy for data.
- Finally we have 3 models to use as a prediction model.

11. Future work

The focus of this project was not only to predict the Cause of the death of 39 bins, but also to understand the death patterns in U.S.

1. The result of the final model accuracy is not high. Will continue research on collecting more data which helps to increase the accuracy of the model of finding the cause of death in U.S
2. In the current project worked on death record of year 2014. Next will be collecting the data from 2013 & 2015 to find patterns between the years, like increase/decrease in suicidal or accidental rates, finding top cause of death in sequential years.
3. Perform research on hospital mortality data.
4. Research on child mortality deaths.

12. Acknowledgements:

I would like to thank Matt Fornito for his words of encouragement and guidance through out my project work.

13. REFERENCES

1. https://www.kaggle.com/cdc/mortality
2. https://www.cdc.gov/nchs/data/dvs/Record_Layout_2010.pdf
3. http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram
4. https://simba.isr.umich.edu/restricted/docs/Mortality/codedcauses_readfirst.pdf
5. https://www.kaggle.com/c/yelp-recruiting/forums/t/4166/running-regression-tree-model-with-more-than-32-factors