

US FIRES DATASET - DATA PRE-PROCESSING

Group 3

27/02/2021

1. Loading packages required for Data Pre-processing

```
# Loading Dplyr and VIM packages
if (require(dplyr,VIM)==FALSE) {
  library("dplyr")
  library("VIM")
}
```

```
## Loading required package: dplyr
```

```
##
```

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

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

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
```

```
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
```

```
##
```

```
##      sleep
```

2. Loading all datasets for the project

```
# Loading datasets
us_fires1 = read.csv("us_fires_7.csv")
us_fires_all = read.csv("all_fires.csv")
us_cities = read.csv("uscities.csv")
```

3. Select the required columns from the dataset

```
# Selecting required columns from dataset
all_fires_us <- us_fires_all %>%
  select(FIRE_YEAR,STATE,STAT_CAUSE_DESCR,FIRE_SIZE,FIRE_SIZE_m2,FIRE_SIZE_ha,
         IGNITION,Wind,NBCD_countrywide_biomass_mosaic,GROUPVEG,EcoArea_km2,
         LATITUDE,LONGITUDE)

#-----
fires_us_sub <- us_fires1 %>%
  select(county,fire_name,nwgc_reporting_unit_name,
         source_reporting_unit,fire_year,discovery_date,
         discovery_doy,stat_cause_code,stat_cause_descr,
         fire_size,fire_size_class,fips_code,fips_name,latitude,
         longitude)

#-----
cities_us_sub <- us_cities %>%
  select(city,state_name,state_id,county_name,
         population,density,id,lat,lng)
```

4. Joining the shortlisted datasets

```
test <- inner_join(all_fires_us,cities_us_sub,by=c("STATE" = "state_id","LATITUDE"="lat"))
dim(test)
```

```
## [1] 12664    20
```

```
sumNa(test)
```

```
## [1] 19
```

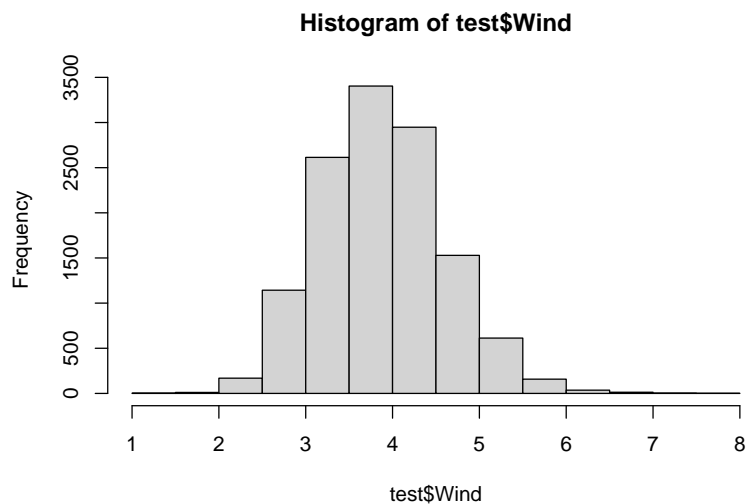
```
apply(test, 2, sumNa)
```

```
##          FIRE_YEAR          STATE
##              0              0
## STAT_CAUSE_DESCR      FIRE_SIZE
##              0              0
##      FIRE_SIZE_m2      FIRE_SIZE_ha
##              0              0
##          IGNITION          Wind
##              0              19
## NBCD_countrywide_biomass_mosaic  GROUPVEG
##              0              0
```

```
##          EcoArea_km2          LATITUDE
##          0          0
##          LONGITUDE          city
##          0          0
##          state_name          county_name
##          0          0
##          population          density
##          0          0
##          id          lng
##          0          0
```

5. Histogram of wind variable

```
hist(test$Wind)
```



6. kNN clustering Imputation method

```
# Rational for this imputation technique is the shape of the distribution which,
# suggests that the mean is a good parametric and validates kNN as an adequate imputation technique.
imputed_test <- VIM::kNN(test,variable = "Wind",numFun = weighted.mean,weightDist = TRUE ,k = 5)

imputed_test$Wind_imp = NULL
apply(imputed_test, 2, sumNa)
```

```
##          FIRE_YEAR          STATE
##          0          0
##          STAT_CAUSE_DESCR          FIRE_SIZE
##          0          0
##          FIRE_SIZE_m2          FIRE_SIZE_ha
##          0          0
##          IGNITION          Wind
```

```
##                                0                                0
## NBCD_countrywide_biomass_mosaic                                GROUPVEG
##                                0                                0
##                                EcoArea_km2                        LATITUDE
##                                0                                0
##                                LONGITUDE                        city
##                                0                                0
##                                state_name                        county_name
##                                0                                0
##                                population                        density
##                                0                                0
##                                id                                lng
##                                0                                0
```

7. Renaming and re-structuring columns

```
# Standardising Column Names
test34 <- imputed_test %>%
  rename(State_id =STATE)
colnames(imputed_test)
```

```
## [1] "FIRE_YEAR"                "STATE"
## [3] "STAT_CAUSE_DESCR"        "FIRE_SIZE"
## [5] "FIRE_SIZE_m2"            "FIRE_SIZE_ha"
## [7] "IGNITION"                "Wind"
## [9] "NBCD_countrywide_biomass_mosaic" "GROUPVEG"
## [11] "EcoArea_km2"             "LATITUDE"
## [13] "LONGITUDE"              "city"
## [15] "state_name"              "county_name"
## [17] "population"              "density"
## [19] "id"                      "lng"
```

```
imputed_test_newcolname <- imputed_test %>%
  rename(Year = FIRE_YEAR, State_name = STATE, Ignition_method = STAT_CAUSE_DESCR,
         Fire_size = FIRE_SIZE, Fire_size_m2 = FIRE_SIZE_m2,
         Fire_size_hectares = FIRE_SIZE_ha, Cause = IGNITION,
         Wind_direction = Wind , Countrywide_biomass = NBCD_countrywide_biomass_mosaic,
         Vegetation_type = GROUPVEG, Eco_areakm2 = EcoArea_km2, Latitude = LATITUDE,
         Longitude = LONGITUDE, City = city,County = county_name,
         Population =population, Pop_density = density, Fire_ID =id, t=lng)
```

```
#-----
# Removing Surplus Columns
final_set <- imputed_test_newcolname[,c(1:19)]
```

```
#-----
# Re-organising column sequence
reorganised_final_set <- imputed_test_newcolname[, c(19, 14, 15, 2 ,16 , 1,
                                                    3, 7, 4, 5, 6, 8, 17, 18,
                                                    9, 10, 11, 12, 13)]
#-----
```

```
# Renaming last columns
```

```
reorganised_final_set <- reorganised_final_set %>%
  rename(State_id=State_name, State_name = state_name)
colnames(reorganised_final_set)
```

```
## [1] "Fire_ID"          "City"              "State_name"
## [4] "State_id"         "County"            "Year"
## [7] "Ignition_method"  "Cause"             "Fire_size"
## [10] "Fire_size_m2"     "Fire_size_hectares" "Wind_direction"
## [13] "Population"       "Pop_density"       "Countrywide_biomass"
## [16] "Vegetation_type"  "Eco_areakm2"       "Latitude"
## [19] "Longitude"
```

```
head(reorganised_final_set)
```

```
##      Fire_ID      City State_name State_id      County Year
## 1 1840023113    Toyah     Texas      TX      Reeves 2005
## 2 1840013053 McCord Bend Missouri      MO      Stone 2006
## 3 1840013053 McCord Bend Missouri      MO      Stone 2006
## 4 1840022159 Livingston    Texas      TX      Polk 2006
## 5 1840028097   Muscoy California    CA San Bernardino 1997
## 6 1840026983     Swink  Oklahoma      OK      Choctaw 1993
## Ignition_method Cause Fire_size Fire_size_m2 Fire_size_hectares
## 1  Debris Burning Human      55.0  222577.300      22.2577300
## 2           Arson Human       1.0   4046.860      0.4046860
## 3           Arson Human       1.5   6070.290      0.6070290
## 4           Arson Human      10.0  40468.600      4.0468600
## 5 Equipment Use Human       0.1    404.686      0.0404686
## 6  Debris Burning Human       3.0  12140.580      1.2140580
## Wind_direction Population Pop_density Countrywide_biomass Vegetation_type
## 1      3.978921      108          25      788.7599 Hardwood-Conifer
## 2      4.443915      299         393      507.9600      Hardwood
## 3      4.443915      299         393      507.9600      Hardwood
## 4      4.322040     5242         231     1007.0100      Riparian
## 5      2.807137     12562        1606       42.8400      Shrubland
## 6      4.112240       50          79     148.1400 Hardwood-Conifer
## Eco_areakm2 Latitude Longitude
## 1  151719.54  31.3125 -94.27083
## 2  106370.52  36.7875 -92.08500
## 3  106370.52  36.7875 -92.08611
## 4  151719.54  30.7100 -95.41083
## 5   20067.56  34.1550 -117.93833
## 6  151719.54  34.0168 -94.70020
```

Creating the final csv file for data visualisation

```
# Create final dataset
```

```
write.csv(reorganised_final_set,"final_us_fires.csv")
```