New York City Demographic Dataset (Prediction of Gender Ratio)

R Code

```
#### Setting The Working Directory ####
setwd("D:/01. B.TECH 3rd YEAR/04. Mini Project/MINI")

#### Reding The Dataset ####
ny <- read.delim("New_York.dat")

#### EXPLORATORY DATA ANALYSIS ####
# Count of NA values
sapply(ny, function(x) sum(is.na(x)))
# Count of empty strings
sapply(ny, function(x) length(which(x=='')))
# Counting the Count of Number of Unique values in every Column
sapply(ny, function(x) length(unique(x)))

summary(ny)
# Here we find that the median population is 4013 but Max population is coming very high i.e 7322563.
# Therefore we can sense that there might be outliers present in this column.
# Now we'll see which place has such a high value
new_york[which(new_york$TOT_POP==7322564),]
# Hence we come to know that this population is of New York City
# After Cross Checking on the Internet, in 2015 which was when the book in which this dataset is given was published,
# the population of New York City was 8.2 million but since here we have the population as 7.32 million which
# was the population of New York City in 1990. So maybe this dataset is of 1990.

# install.packages("ggplot2")
library(ggplot2)
ggplot(data = ny, aes(y=ny$TOT_POP)) + geom_boxplot() + ggtitle("Boxplot of TOT_POP")
```

```
ggplot(data = ny, aes(y=ny$PCT_U18)) + geom_boxplot() + ggtitle("Boxplot of PCT_U18")

ggplot(data = ny, aes(y=ny$PC_18_65)) + geom_boxplot() + ggtitle("Boxplot of PC_18_65")

ggplot(data = ny, aes(y=ny$PCT_O65)) + geom_boxplot() + ggtitle("Boxplot of PCT_O65")

ggplot(data = ny, aes(y=ny$MALE_FEM)) + geom_boxplot() + ggtitle("Boxplot of MALE_FEM")
```

```
# install.packages("psych")

library(psych)

pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],method = "pearson",lm=TRUE,
ellipses = FALSE)
```

```
#### Data Cleaning ####

# TOT_POP

boxplot(ny$TOT_POP,main="TOT_POP")$stats[c(1,5),] #1000 #19750

length(which(ny$TOT_POP>19750)) #81

length(which(ny$TOT_POP<1000))  #0

ny <- ny[-which(ny$TOT_POP>19750),]

summary(ny$TOT_POP)
```

```
# PCT_U18

boxplot(ny$PCT_U18,main="PCT_U18")$stats[c(1,5),]

length(which(ny$PCT_U18>33.7)) #9

length(which(ny$PCT_U18<15.1)) #23

ny <- ny[-which(ny$PCT_U18>33.7),]

ny <- ny[-which(ny$PCT_U18<15.1),]
```

```
# PC_18_65

boxplot(ny$PC_18_65,main="PC_18_65")$stats[c(1,5),]

length(which(ny$PC_18_65>72.4)) #4

length(which(ny$PC_18_65<49.5)) #2

ny <- ny[-which(ny$PC_18_65>72.4),]

ny <- ny[-which(ny$PC_18_65<49.5),]
```

```
# PCT_O65

boxplot(ny$PCT_O65,main="PCT_O65")$stats[c(1,5),]
```

```
length(which(ny$PCT_O65>26.2)) #10
length(which(ny$PCT_O65<2.7))
ny <- ny[-which(ny$PCT_O65>26.2)]
```

 # MALE_FEM

```
boxplot(ny$MALE_FEM,main="MALE_FEM")$stats[c(1,5),]
length(which(ny$MALE_FEM<69.7)) #3
length(which(ny$MALE_FEM>107.4)) #5
ny <- ny[-which(ny$MALE_FEM<69.7),]
ny <- ny[-which(ny$MALE_FEM>107.4),]
```

 # Scatterplot Matrix

```
pairs(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")])
```

 # Correlation Matrix (Pearson)

```
cor(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")])

pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],method = "pearson",lm=TRUE)

ny$TOT_POP <- log(ny$TOT_POP)

pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],method = "pearson",lm=TRUE,
ellipses = FALSE)
```

Splitting The Dataset

```
ny1 <- ny[,c(6,2,3,4,5)]
```

#install.packages('caTools')

```
library(caTools)
set.seed(123)
```

#Split of dataset into training dataset and test dataset

```
split <- sample.split(ny1$MALE_FEM,SplitRatio = 0.8)
training_set <- subset(ny1, split==TRUE)
```

```
test_set <- subset(ny1, split == FALSE)

#Scaling
#For putting variables in same scale
training_set[,2:5] <- scale(training_set[,2:5])
test_set[,2:5] <- scale(test_set[,2:5])

#### Model Building ####

# Using all variables
model1 <- lm(MALE_FEM ~ TOT_POP + PCT_U18 + PC_18_65 + PCT_O65,data = training_set)
summary(model1)
y_pred1 = predict(model1,newdata=test_set)
cor(y_pred1,test_set$MALE_FEM)*cor(y_pred1,test_set$MALE_FEM)
# 48.08

# Eliminating PCT_O65
## FINAL MODEL
model2 <- lm(MALE_FEM ~ TOT_POP + PCT_U18 + PC_18_65,data = training_set)
summary(model2)
y_pred2 = predict(model2,newdata=test_set)
cor(y_pred2,test_set$MALE_FEM)*cor(y_pred2,test_set$MALE_FEM)
# 48.48

# Removing the PCT_U18 in model2
model22 <- lm(MALE_FEM ~ PCT_U18 + PC_18_65,data = training_set)
summary(model22)

# Calculating Variance Inflation Factor
library(caret)
varImp(model1)

# Error Rate
sigma(model2)/mean(training_set$MALE_FEM)*100
```

```
# Calculating the No of rows in training set
nrow(training_set)



# Eliminating PC_18_65
model3 <- lm(MALE_FEM ~ TOT_POP + PCT_U18 + PCT_O65,data = training_set)
summary(model3)
y_pred3 = predict(model3,newdata=test_set)
cor(y_pred3,test_set$MALE_FEM)*cor(y_pred3,test_set$MALE_FEM)
# 48.24

# Eliminating PC_18_65 and PCT_O65
model4 <- lm(MALE_FEM ~ TOT_POP + PCT_U18,data = training_set)
summary(model4)
y_pred4 = predict(model4,newdata=test_set)
cor(y_pred4,test_set$MALE_FEM)*cor(y_pred4,test_set$MALE_FEM)
# 0.1

#### Different Plots ####
# Histogram of residuals
resid = test_set$MALE_FEM-y_pred2

ggplot() + aes(resid)+ geom_histogram(binwidth=1, colour="black", fill="white") +
  geom_density(aes(y=1*..count..)) +
  ggtitle("Overlay Histogram of Residuals") +
  xlab("Residuals") +
  ylab("")

# Homoscedasticity
ggplot(data=NULL,aes(x=y_pred2,y=resid))+geom_point() +
  ggtitle("Predicted Values VS Residuals") +
  ylab("Residuals") +
  xlab("Predicted Values")
```

ny

Filter

	PLACE	TOT_POP	PCT_U18	PC_18_65	PCT_O65	MALE_FEM
1	Adams village	7.469084	25.3	56.2	18.5	79.0
2	Adams Center CDP	7.423568	30.4	58.7	10.9	87.0
3	Addison village	7.518607	28.8	56.0	15.2	84.0
4	Airmont CDP	8.966356	25.0	65.3	9.7	89.0
5	Akron village	7.974533	24.7	57.1	18.2	81.1
7	Albertson CDP	8.549854	18.9	61.5	19.6	91.8
8	Albion village	8.676417	26.1	56.8	17.1	85.2
9	Alden village	7.806696	25.0	60.6	14.4	88.3
10	Alexandria Bay village	7.085064	20.8	56.2	23.0	83.3
12	Allegany village	7.590852	20.8	62.7	16.5	89.1
13	Altamont village	7.325808	28.1	62.1	9.8	89.9
15	Amenia CDP	6.963190	24.8	60.0	15.2	86.2
16	Amityville village	9.136263	20.3	61.8	17.9	87.3
18	Andover village	7.025538	27.8	56.4	15.8	88.4
19	Angola village	7.710205	27.8	58.9	13.3	86.3
20	Angola on the Lake CDP	7.449498	24.4	62.5	13.0	91.0
21	Apalachin CDP	7.096721	28.6	61.8	9.5	97.3
22	Aquebogue CDP	7.630461	23.3	59.5	17.2	94.7
23	Arcade village	7.640604	31.3	57.0	11.7	87.2
24	Ardsley village	8.359837	22.4	62.3	15.4	84.4
25	Arlington CDP	9.388319	16.6	71.4	12.1	87.1
26	Armonk CDP	7.917536	24.1	65.1	10.7	97.1
27	Athens village	7.443078	25.5	57.3	17.2	88.3
28	Atlantic Beach village	7.566828	19.3	65.6	15.1	92.8
29	Attica village	7.874739	26.2	59.1	14.7	87.7
31	Averill Park CDP	7.412160	25.9	64.8	9.3	95.1
32	Avoca village	6.940222	30.2	55.3	14.5	86.3
33	Avon village	8.004700	26.1	59.7	14.2	84.2

Showing 1 to 29 of 663 entries