```
#MS 401: Missing Value Imputation
# Daniel Zaremba

#Using the "R" programming language, do the following:
#Explore the data by summarizing it and graphing it

df_old  = read.csv(file="C:/Users/Owner/Documents/Northwestern/
Statistical_Analysis_R/Week8/Bonus_MissingValues/WalkThruVideo/HMEQ.csv",
header=TRUE, sep=",")
df = df_old

print( head( df ) )
print( tail( df ) )
str(df)
summary(df)

boxplot( df$MORTDUE  )

boxplot( df$VALUE  )

boxplot( df$YOJ  )

boxplot( df$DEROG  )

boxplot( df$DELINQ  )

boxplot( df$CLAGE  )

boxplot( df$NINQ  )

boxplot( df$CLNO  )

boxplot( df$DEBTINC  )

#Create a flag variable for each missing numeric variable. The variable name
should begin with "M_" as
#it was in the video.

df$M_MORTDUE = is.na( df$MORTDUE ) + 0
df$M_VALUE = is.na( df$VALUE ) + 0
df$M_YOJ = is.na( df$YOJ ) + 0
df$M_DEROG = is.na( df$DEROG ) + 0
df$M_DELINQ = is.na( df$DELINQ ) + 0
df$M_CLAGE = is.na( df$CLAGE ) + 0
df$M_NINQ = is.na( df$NINQ ) + 0
df$M_CLNO = is.na( df$CLNO ) + 0
df$M_DEBTINC = is.na( df$DEBTINC ) + 0



#Create an imputed variable for each numeric variable that has a missing
value. Fill in the missing
```

```
#value with the mean. The variable name should begin with "IMP_" as it was in
the video.


a <- round(mean( df$MORTDUE, na.rm=TRUE ))
df$IMP_MORTDUE = df$MORTDUE
df$IMP_MORTDUE = ifelse(is.na( df$IMP_MORTDUE ), a, df$IMP_MORTDUE )
head(df)

b <- round(mean( df$VALUE, na.rm=TRUE ))
df$IMP_VALUE = df$VALUE
df$IMP_VALUE = ifelse(is.na( df$IMP_VALUE ), b, df$IMP_VALUE )

c <- round(mean( df$YOJ, na.rm=TRUE ))
df$IMP_YOJ = df$YOJ
df$IMP_YOJ = ifelse(is.na( df$IMP_YOJ ), c, df$IMP_YOJ )

d <- round(mean( df$DEROG, na.rm=TRUE ))
df$IMP_DEROG = df$DEROG
df$IMP_DEROG = ifelse(is.na( df$IMP_DEROG ), d, df$IMP_DEROG )

e <- round(mean( df$DELINQ, na.rm=TRUE ))
df$IMP_DELINQ = df$DELINQ
df$IMP_DELINQ = ifelse(is.na( df$IMP_DELINQ ), e, df$IMP_DELINQ )

f <- round(mean( df$CLAGE, na.rm=TRUE ))
df$IMP_CLAGE = df$CLAGE
df$IMP_CLAGE = ifelse(is.na( df$IMP_CLAGE ), f, df$IMP_CLAGE )

g <- round(mean( df$NINQ, na.rm=TRUE ))
df$IMP_NINQ = df$NINQ
df$IMP_NINQ = ifelse(is.na( df$IMP_NINQ ), g, df$IMP_NINQ )

h <- round(mean( df$CLNO, na.rm=TRUE ))
df$IMP_CLNO = df$CLNO
df$IMP_CLNO = ifelse(is.na( df$IMP_CLNO ), h, df$IMP_CLNO )

i <- round(mean( df$DEBTINC, na.rm=TRUE ))
df$IMP_DEBTINC = df$DEBTINC
df$IMP_DEBTINC = ifelse(is.na( df$IMP_DEBTINC ), i, df$IMP_DEBTINC )



#Create an imputed variable for each categorical variable that has a missing
value. Fill in the
#missing value with the the value "UNKNOWN". The variable name should begin
with "IMP_" as it was
#in the video.
df$IMP_REASON = df$REASON
df$IMP_REASON = ifelse(df$REASON == "", "UNKNOWN", as.character(df$IMP_REASON)
)

df$IMP_JOB = df$JOB
df$IMP_JOB = ifelse(df$JOB == "", "UNKNOWN", as.character(df$IMP_JOB) )
```

```
head(df)

#Identify any outliers and fix them in a method similar to those presented in
the video.
#After variables are fixed, remove the original variables

boxplot( df$IMP_MORTDUE  )
boxplot( df$IMP_VALUE  )
boxplot( df$IMP_YOJ  )
boxplot( df$IMP_DEROG  )
boxplot( df$IMP_DELINQ  )
boxplot( df$IMP_CLAGE  )
boxplot( df$IMP_NINQ  )
boxplot( df$IMP_CLNO  )
boxplot( df$IMP_DEBTINC  )

a1 = max( df$IMP_MORTDUE, na.rm=TRUE )
z1 = min( df$IMP_MORTDUE, na.rm=TRUE )
m1 = mean( df$IMP_MORTDUE, na.rm=TRUE )
s1 = sd( df$IMP_MORTDUE, na.rm=TRUE )

df$IMP_MORTDUE = ifelse( df$IMP_MORTDUE > m1+3*s1, m1+3*s1, df$IMP_MORTDUE )
df$IMP_MORTDUE = ifelse( df$IMP_MORTDUE < m1-3*s1, m1-3*s1, df$IMP_MORTDUE )

a2 = max( df$IMP_VALUE, na.rm=TRUE )
z2 = min( df$IMP_VALUE, na.rm=TRUE )
m2 = mean( df$IMP_VALUE, na.rm=TRUE )
s2 = sd( df$IMP_VALUE, na.rm=TRUE )

df$IMP_VALUE = ifelse( df$IMP_VALUE > m2+3*s2, m2+3*s2, df$IMP_VALUE )
df$IMP_VALUE = ifelse( df$IMP_VALUE < m2-3*s2, m2-3*s2, df$IMP_VALUE )

a3 = max( df$IMP_YOJ, na.rm=TRUE )
z3 = min( df$IMP_YOJ, na.rm=TRUE )
m3 = mean( df$IMP_YOJ, na.rm=TRUE )
s3 = sd( df$IMP_YOJ, na.rm=TRUE )

df$IMP_YOJ = ifelse( df$IMP_YOJ > m3+3*s3, m3+3*s3, df$IMP_YOJ )
df$IMP_YOJ = ifelse( df$IMP_YOJ < m3-3*s3, m3-3*s3, df$IMP_YOJ )

a4 = max( df$IMP_DEROG, na.rm=TRUE )
z4 = min( df$IMP_DEROG, na.rm=TRUE )
m4 = mean( df$IMP_DEROG, na.rm=TRUE )
s4 = sd( df$IMP_DEROG, na.rm=TRUE )

df$IMP_DEROG = ifelse( df$IMP_DEROG > m4+3*s4, m4+3*s4, df$IMP_DEROG )
df$IMP_DEROG = ifelse( df$IMP_DEROG < m4-3*s4, m4-3*s4, df$IMP_DEROG )

a5 = max( df$IMP_DELINQ, na.rm=TRUE )
z5 = min( df$IMP_DELINQ, na.rm=TRUE )
m5 = mean( df$IMP_DELINQ, na.rm=TRUE )
s5 = sd( df$IMP_DELINQ, na.rm=TRUE )
```

```
df$IMP_DELINQ = ifelse( df$IMP_DELINQ > m5+3*s5, m5+3*s5, df$IMP_DELINQ )
df$IMP_DELINQ = ifelse( df$IMP_DELINQ < m5-3*s5, m5-3*s5, df$IMP_DELINQ )

a6 = max( df$IMP_CLAGE, na.rm=TRUE )
z6 = min( df$IMP_CLAGE, na.rm=TRUE )
m6 = mean( df$IMP_CLAGE, na.rm=TRUE )
s6 = sd( df$IMP_CLAGE, na.rm=TRUE )

df$IMP_CLAGE = ifelse( df$IMP_CLAGE > m6+3*s6, m6+3*s6, df$IMP_CLAGE )
df$IMP_CLAGE = ifelse( df$IMP_CLAGE < m6-3*s6, m6-3*s6, df$IMP_CLAGE )

a7 = max( df$IMP_NINQ, na.rm=TRUE )
z7 = min( df$IMP_NINQ, na.rm=TRUE )
m7 = mean( df$IMP_NINQ, na.rm=TRUE )
s7 = sd( df$IMP_NINQ, na.rm=TRUE )

df$IMP_NINQ = ifelse( df$IMP_NINQ > m7+3*s7, m7+3*s7, df$IMP_NINQ )
df$IMP_NINQ = ifelse( df$IMP_NINQ < m7-3*s7, m7-3*s7, df$IMP_NINQ )

a8 = max( df$IMP_CLNO, na.rm=TRUE )
z8 = min( df$IMP_CLNO, na.rm=TRUE )
m8 = mean( df$IMP_CLNO, na.rm=TRUE )
s8 = sd( df$IMP_CLNO, na.rm=TRUE )

df$IMP_CLNO = ifelse( df$IMP_CLNO > m8+3*s8, m8+3*s8, df$IMP_CLNO )
df$IMP_CLNO = ifelse( df$IMP_CLNO < m8-3*s8, m8-3*s8, df$IMP_CLNO )


a9 = max( df$IMP_DEBTINC, na.rm=TRUE )
z9 = min( df$IMP_DEBTINC, na.rm=TRUE )
m9 = mean( df$IMP_DEBTINC, na.rm=TRUE )
s9 = sd( df$IMP_DEBTINC, na.rm=TRUE )

df$IMP_DEBTINC = ifelse( df$IMP_DEBTINC > m9+3*s9, m9+3*s9, df$IMP_DEBTINC )
df$IMP_DEBTINC = ifelse( df$IMP_DEBTINC < m9-3*s9, m9-3*s9, df$IMP_DEBTINC )

summary(df)
# Remove the original fields

df = subset(df, select = -c( MORTDUE, VALUE, REASON, JOB, YOJ, DEROG, DELINQ,
CLAGE, NINQ, CLNO, DEBTINC) )

summary(df)
```