Find the complete codes relating to data prep practice assignment

Dummy Variables

```
setwd("/Users/lalitsachan/Desktop/March onwards/CBAP with R/Data/")
# You'll have to chose path accroding to location of file in your machine

d=read.csv("census_income.csv",stringsAsFactors = F)
library(dplyr)

for(i in 1:ncol(d)){
  if(class(d[,i])=="character"){
    if(names(d)[i]!="Y"){
    message=paste("Number of categories in ",names(d)[i]," : ")
    num.cat=length(unique(d[,i]))
    print(paste0(message,num.cat))
    }
  }
}
```

```
## [1] "Number of categories in  workclass  : 9"
## [1] "Number of categories in  education  : 16"
## [1] "Number of categories in  marital.status  : 7"
## [1] "Number of categories in  occupation  : 15"
## [1] "Number of categories in  relationship  : 6"
## [1] "Number of categories in  race  : 5"
## [1] "Number of categories in  sex  : 2"
## [1] "Number of categories in  native.country  : 42"
```

```
table(d$race)
```

```
##
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black
##                 311                1039                3124
##               Other               White
##                 271               27816
```

```
d=d%>%
  mutate(race_AIE=as.numeric(race==" Amer-Indian-Eskimo"),
         race_API=as.numeric(race==" Asican-Pac-Islander"),
         race_Black=as.numeric(race==" Black"),
         race_White=as.numeric(race==" White")) %>%
  select(-race)
# we ignored the category which had least frequency
```

```
table(d$sex)
```

```
##
##  Female    Male
##   10771   21790
```

```
d=d %>%
  mutate(sex_M=as.numeric(sex==" Male")) %>%
  select(-sex)
table(d$relationship)
```

```
##
##         Husband   Not-in-family  Other-relative       Own-child
##           13193            8305             981            5068
##       Unmarried            Wife
##            3446            1568
```

```
d=d %>%
  mutate(rel_h=as.numeric(relationship==" Husband"),
         rel_nif=as.numeric(relationship==" Not-in-family"),
         rel_oc=as.numeric(relationship==" Own-child"),
         rel_um=as.numeric(relationship==" Unmarried"),
         rel_w=as.numeric(relationship==" Wife")) %>%
  select(-relationship)
```

Combining Similar Categories

Note: Grouping is done on the basis of similar behaviour across classes of target [which is Y in this case]

```
round(prop.table(table(d$workclass,d$Y),1),1)
```

```
##
##                     <=50K  >50K
##    ?                  0.9   0.1
##    Federal-gov        0.6   0.4
##    Local-gov          0.7   0.3
##    Never-worked       1.0   0.0
##    Private            0.8   0.2
##    Self-emp-inc       0.4   0.6
##    Self-emp-not-inc   0.7   0.3
##    State-gov          0.7   0.3
##    Without-pay        1.0   0.0
```

```
# you can take any category [ after grouping ] as base [the one to ignore]
d=d %>%
  mutate(wc_1=as.numeric(workclass==" Self-emp-inc"),
         wc_2=as.numeric(workclass==" Federal-gov"),
         wc_3=as.numeric(workclass %in% c(" Local-gov"," Self-emp-not-inc"," State-gov")),
         wc_4=as.numeric(workclass==" Private"),
         wc_5=as.numeric(workclass==" ?")) %>%
  select(-workclass)
```

```
round(prop.table(table(d$education,d$Y),1),1)
```

```
##
##                  <=50K  >50K
##    10th            0.9   0.1
##    11th            0.9   0.1
##    12th            0.9   0.1
##    1st-4th         1.0   0.0
##    5th-6th         1.0   0.0
##    7th-8th         0.9   0.1
##    9th             0.9   0.1
##    Assoc-acdm      0.8   0.2
##    Assoc-voc       0.7   0.3
##    Bachelors       0.6   0.4
##    Doctorate       0.3   0.7
##    HS-grad         0.8   0.2
##    Masters         0.4   0.6
##    Preschool       1.0   0.0
##    Prof-school     0.3   0.7
##    Some-college    0.8   0.2
```

```
d=d %>%
  mutate(edu_1=as.numeric(education %in% c(" 10th"," 11th"," 12th"," 7th-8th"," 9th")),
         edu_2=as.numeric(education %in% c(" 1st-4th"," 5th-6th"," Preschool")),
         edu_3=as.numeric(education %in% c(" Assoc-acdm"," HS-grad"," Some-college")),
         edu_4=as.numeric(education ==" Assoc-voc"),
         edu_5=as.numeric(education==" Bachelors"),
         edu_6=as.numeric(education==" Masters")) %>%
  select(-education)
round(prop.table(table(d$marital.status,d$Y),1),1)
```

```
##
##                           <=50K  >50K
##    Divorced                 0.9   0.1
##    Married-AF-spouse        0.6   0.4
##    Married-civ-spouse       0.6   0.4
##    Married-spouse-absent    0.9   0.1
##    Never-married            1.0   0.0
##    Separated                0.9   0.1
##    Widowed                  0.9   0.1
```

```
d=d %>%
  mutate(ms_1=as.numeric(marital.status==" Never-married"),
         ms_2=as.numeric(marital.status %in% c(" Married-AF-spouse"," Married-civ-spouse"))) %>%
  select(-marital.status)
round(prop.table(table(d$occupation,d$Y),1),1)
```

```
##
##                       <=50K  >50K
##    ?                    0.9   0.1
##    Adm-clerical         0.9   0.1
##    Armed-Forces         0.9   0.1
##    Craft-repair         0.8   0.2
##    Exec-managerial      0.5   0.5
##    Farming-fishing      0.9   0.1
##    Handlers-cleaners    0.9   0.1
##    Machine-op-inspct    0.9   0.1
##    Other-service        1.0   0.0
##    Priv-house-serv      1.0   0.0
##    Prof-specialty       0.6   0.4
##    Protective-serv      0.7   0.3
##    Sales                0.7   0.3
##    Tech-support         0.7   0.3
##    Transport-moving     0.8   0.2
```

```
d=d %>%
  mutate(oc_1=as.numeric(occupation==" Exec-managerial"),
         oc_2=as.numeric(occupation==" Prof-specialty"),
         oc_3=as.numeric(occupation %in% c(" Protective-serv"," Sales"," Tech-support")),
         oc_4=as.numeric(occupation %in% c(" Craft-repair"," Transport-moving")),
         oc_5=as.numeric(occupation %in% c(" Priv-house-serv"," Other-service"))) %>%
  select(-occupation)
k=round(prop.table(table(d$native.country,d$Y),1),1)
sort(k[,1])
```

```
##                     Cambodia                      France
##                          0.6                         0.6
##                        India                        Iran
##                          0.6                         0.6
##                        Japan                      Taiwan
##                          0.6                         0.6
##                   Yugoslavia                           ?
##                          0.6                         0.7
##                       Canada                       China
##                          0.7                         0.7
##                         Cuba                     England
##                          0.7                         0.7
##                      Germany                      Greece
##                          0.7                         0.7
##                         Hong                       Italy
##                          0.7                         0.7
##                  Philippines                     Hungary
##                          0.7                         0.8
##                      Ireland                      Poland
##                          0.8                         0.8
##                     Scotland                       South
##                          0.8                         0.8
##                     Thailand               United-States
##                          0.8                         0.8
##                      Ecuador                 El-Salvador
##                          0.9                         0.9
##                        Haiti                    Honduras
##                          0.9                         0.9
##                      Jamaica                        Laos
##                          0.9                         0.9
##                       Mexico                   Nicaragua
##                          0.9                         0.9
##                         Peru                    Portugal
##                          0.9                         0.9
##                 Puerto-Rico             Trinadad&Tobago
##                         0.9                         0.9
##                     Vietnam                    Columbia
##                         0.9                         1.0
##          Dominican-Republic                   Guatemala
##                         1.0                         1.0
##          Holand-Netherlands  Outlying-US(Guam-USVI-etc)
##                         1.0                         1.0
```

```
d=d %>%
  mutate(nc_1=as.numeric(native.country %in% c(" Cambodia"," France"," India",
                                               " Iran"," Japan"," Taiwan"," Yugoslavia")),
         nc_2=as.numeric(native.country %in% c(" ?"," Canada"," China"," Cuba"," England",
                                               " Germany"," Greece"," Hong"," Italy",
                                               " Philippines")),
         nc_3=as.numeric(native.country %in% c(" Hungary"," Ireland"," Poland"," Scotland",
                                               " South"," Thailand"," United-States")),
         nc_4=as.numeric(native.country %in% c(" Columbia"," Dominican-Republic",
                                               " Guatemala"," Holand-Netherlands",
                                               " Outlying-US(Guam-USVI-etc)"))) %>%
  select(-native.country)
```

Flag variables

```
# this will give % of observations where capital.gain is 0
sum(d$capital.gain==0)/nrow(d)
```

```
## [1] 0.9167102
```

More than 90% values are 0 , lets go ahead create a flag variable for this

```
d=d %>%
  mutate(cg_flag0=as.numeric(capital.gain==0))
```

```
sum(d$capital.loss==0)/nrow(d)
```

```
## [1] 0.9533491
```

```
d=d %>%
  mutate(cl_flag0=as.numeric(capital.loss==0))
```

Converting the target

```
d$Y=as.numeric(d$Y==" >50K")
```

Save this code for data prep that you have written . We'll be using this prepared data in our exercise in logistic regression module.