DQR of Credit Card Applications

Peipei Han 1424165189

March 3, 2017

File description

data file name is application.csv. It is a subset of credit card application dataset containing 100000 rows and 9 columns. In total, there are 8 categoriacal variables and only 1 numeric variable that is record. There is no missing value in the dataset so the population of all the variables are 100%.

```
## [1] "record"    "date"      "ssn"       "firstname" "lastname"  "address"  
## [7] "zip"       "dob"       "homephone"
```

```
##      record            date               ssn                firstname    
##  Min.   :     1   Min.   :20150101   Min.   :     2503   EAMSTRMT : 1258  
##  1st Qu.: 25001   1st Qu.:20150401   1st Qu.:255816942   TXEMXZZM : 1032  
##  Median : 50001   Median :20150701   Median :509886303   UXXJJZTUZ: 1018  
##  Mean   : 50001   Mean   :20150667   Mean   :504629765   UJSRSMUEZ:  991  
##  3rd Qu.: 75000   3rd Qu.:20150930   3rd Qu.:745870823   SREZUJMJU:  987  
##  Max.   :100000   Max.   :20151231   Max.   :999993079   EASEXMJAT:  745  
##                                                          (Other)  :93969  
##      lastname                 address            zip       
##  ERJSAXA :  829   2602 AJTJ AVE :  117   Min.   :    2  
##  UMXUUUSE:  703   7433 RAEZA ST :   13   1st Qu.:25036  
##  UMARRMA :  642   1775 XJXE LN  :    9   Median :50405  
##  MEAXJUX :  539   426 XUAXZ BLVD:    9   Mean   :50105  
##  XMERRR  :  523   8911 MZSU DR  :    9   3rd Qu.:74514  
##  SXZXJRJT:  439   4907 RRAAU DR :    8   Max.   :99999  
##  (Other) :96325   (Other)       :99835                  
##       dob              homephone        
##  Min.   :19000101   Min.   :6.354e+05  
##  1st Qu.:19161129   1st Qu.:2.675e+09  
##  Median :19500920   Median :5.413e+09  
##  Mean   :19516527   Mean   :5.303e+09  
##  3rd Qu.:19821108   3rd Qu.:8.128e+09  
##  Max.   :20161031   Max.   :9.997e+09  
## 
```

```
##    record      date       ssn firstname  lastname   address       zip 
##         0         0         0         0         0         0         0 
##       dob homephone 
##         0         0
```

unique value ratios are

```
##            [,1]     [,2]   [,3]    [,4]        [,5]       [,6]     
## names.df.  "record" "date" "ssn"   "firstname" "lastname" "address"
## percent.a. "100.0%" "0.4%" "96.5%" "16.6%"     "36.3%"    "97.6%"  
##            [,7]    [,8]    [,9]       
## names.df.  "zip"   "dob"   "homephone"
## percent.a. "16.5%" "36.8%" "22.2%"
```

Fields Description

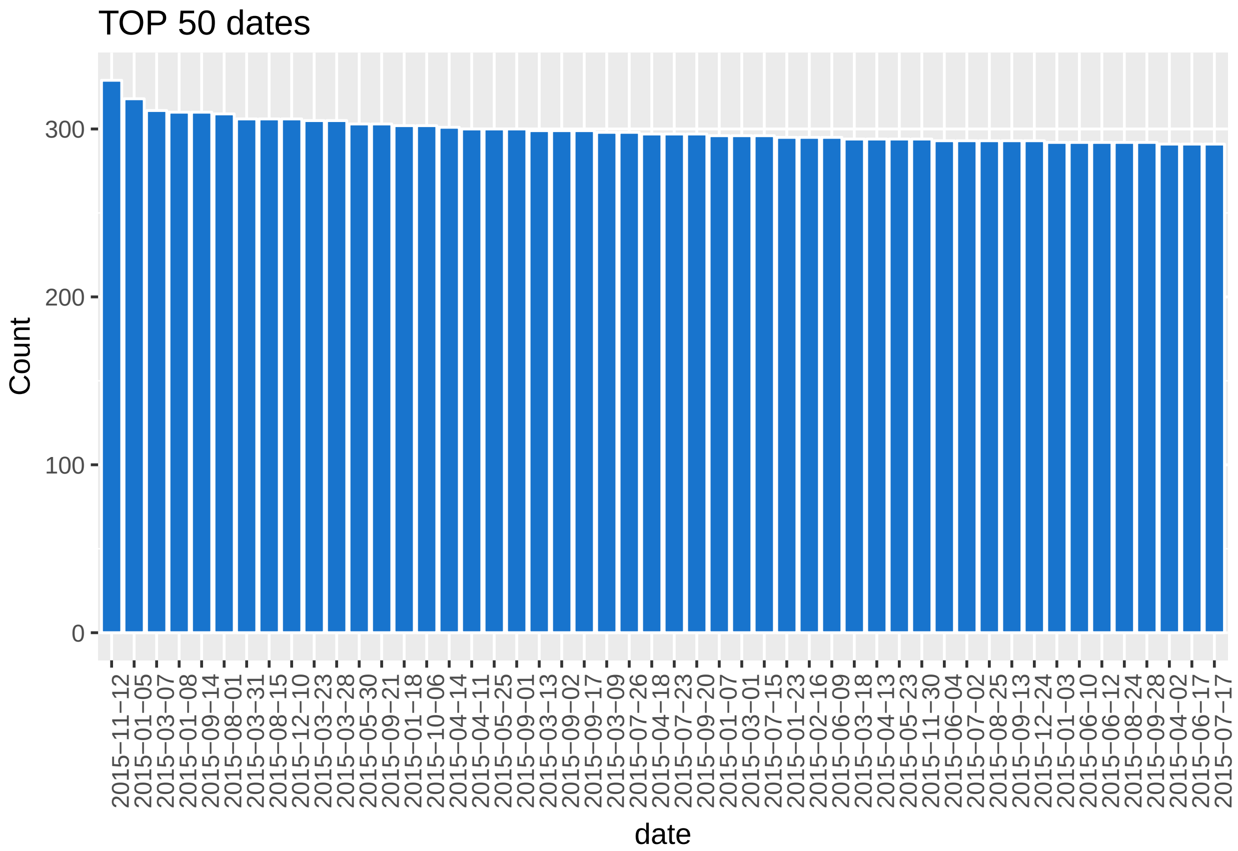
record

record is 100% unique and this is useless field for analysis.

date

date is Categorical variable and the date format is yyyymmdd.

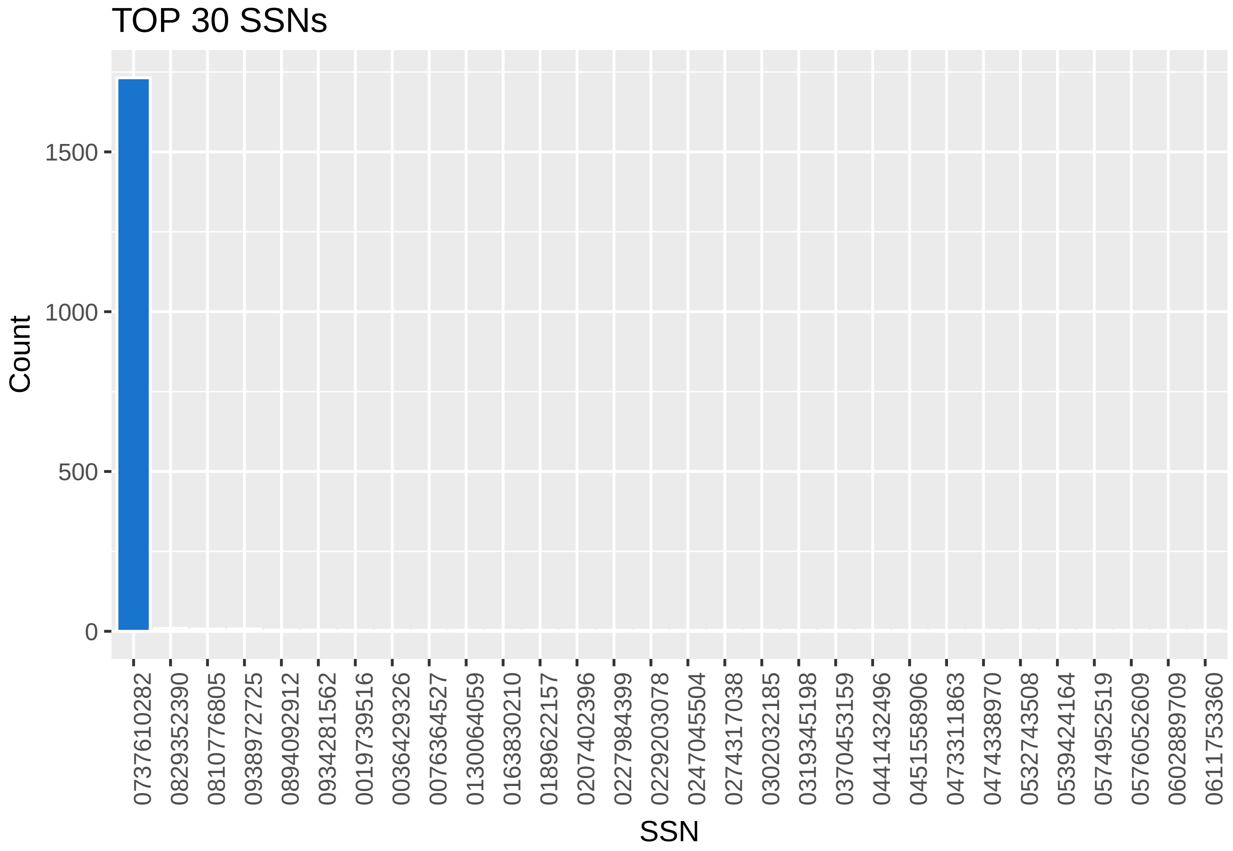
```
df %>% group_by(date)   %>%
    summarize(cnt = n())    %>%
    arrange(desc(cnt) )  %>%
    slice(1:50)    %>%
    ggplot(aes( x =  reorder(as.factor(date),-cnt), y = cnt) )+
    geom_bar(stat = "identity", color = "white" , fill = "dodgerblue3")+
    xlab("date")+
    ylab("Count")+
    ggtitle("TOP 50 dates")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



ssn

ssn is a categorical variable and “737610282” might be a frivolous value in ssn field.

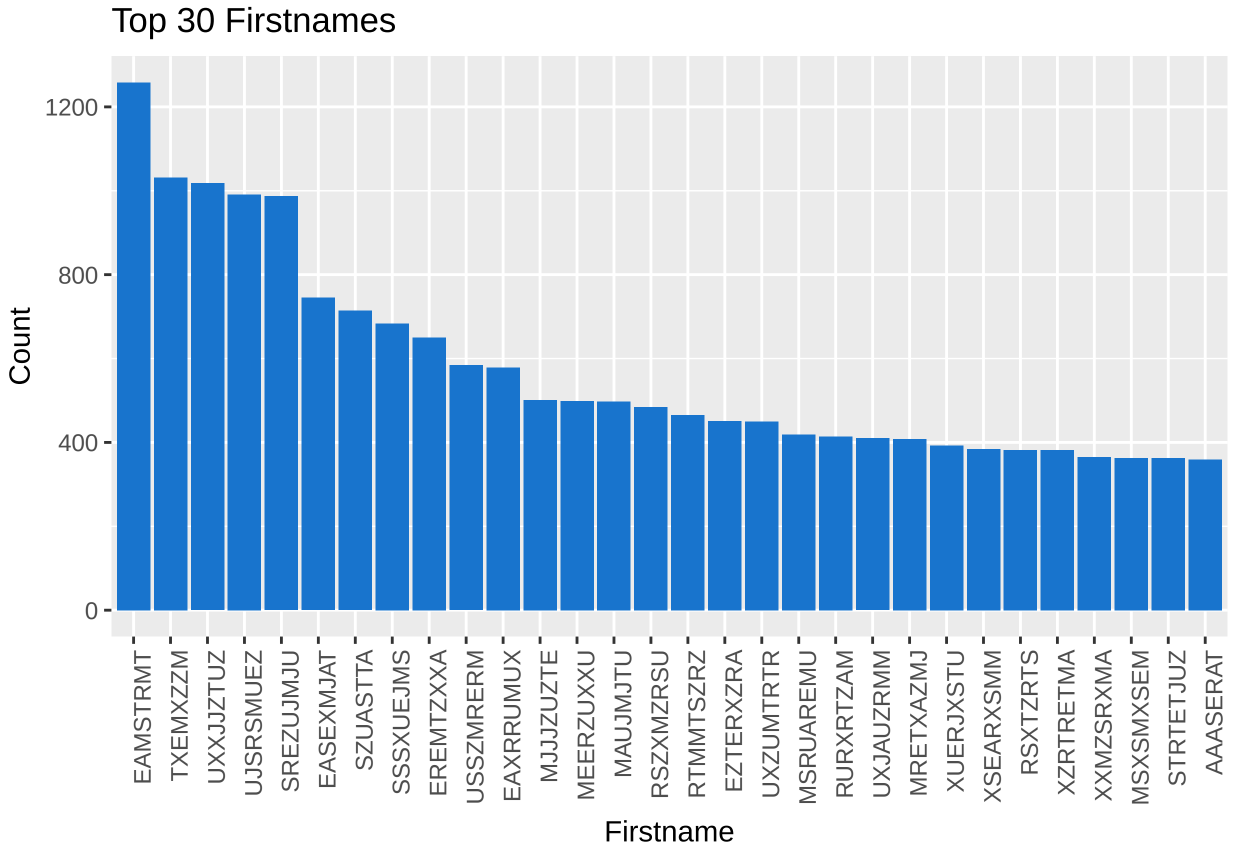
```
df %>% group_by(ssn)   %>%
    summarize(cnt = n())    %>%
    arrange(desc(cnt) )  %>%
    slice(1:30)    %>%
    ggplot(aes( x =  reorder(as.factor(ssn),-cnt), y = cnt) )+
    geom_bar(stat = "identity", color = "white" , fill = "dodgerblue3")+
    xlab("SSN")+
    ylab("Count")+
    ggtitle("TOP 30 SSNs")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



firstname

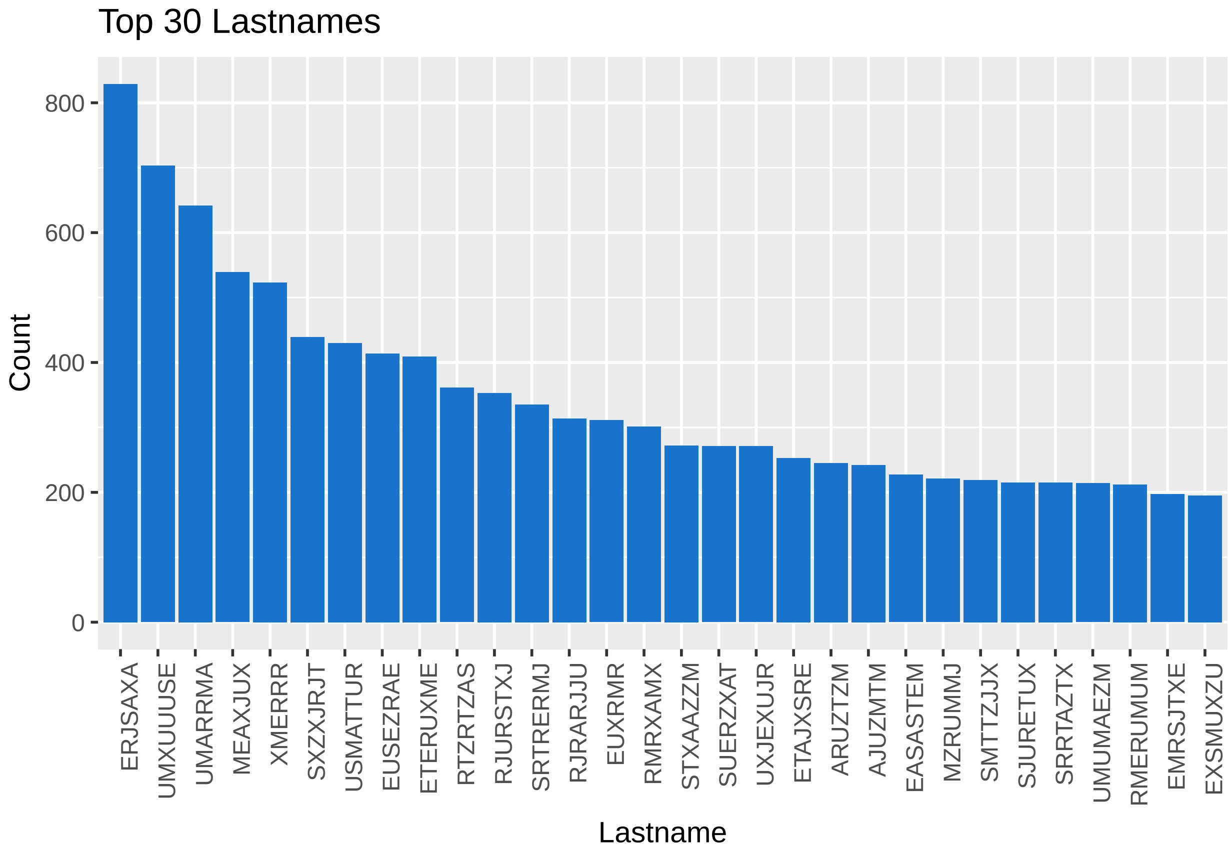
firstname is a categorical varible. “EAMSTRMT” might be a frivolous value in firstname field.

```
df%>%
  group_by(firstname) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt) )  %>%
  slice(1:30)  %>%
  ggplot( aes( x = reorder(firstname,-cnt), y =  cnt) )+
    geom_bar(stat = "identity",fill = "dodgerblue3")+
    xlab("Firstname")+
    ylab("Count")+
    ggtitle("Top 30 Firstnames")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



lastname

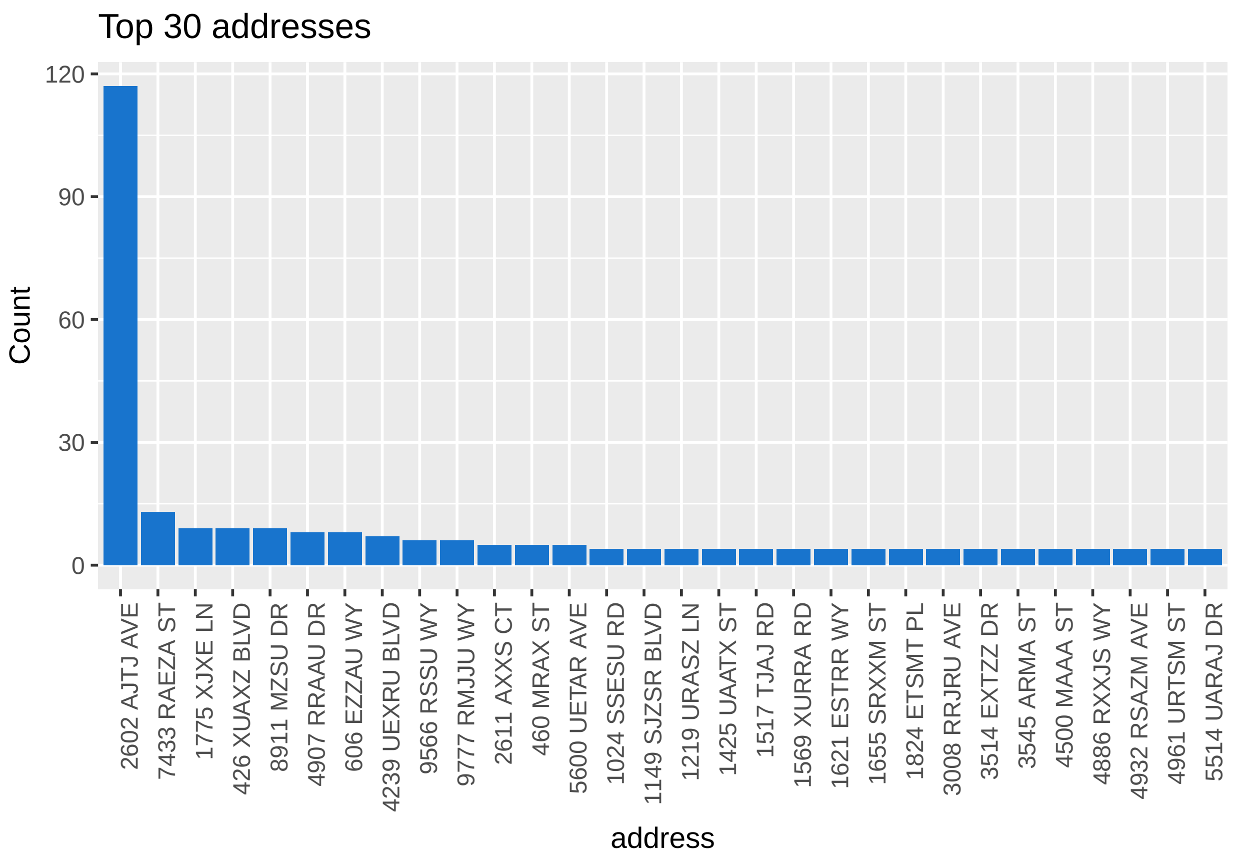
Lastname is a Categorical variable. “ERJSAXA” might be a frivolous value in Lastname field.



address

address is a categorical variable.“2602 AJTJ AVE” might be a frivolous value in address.

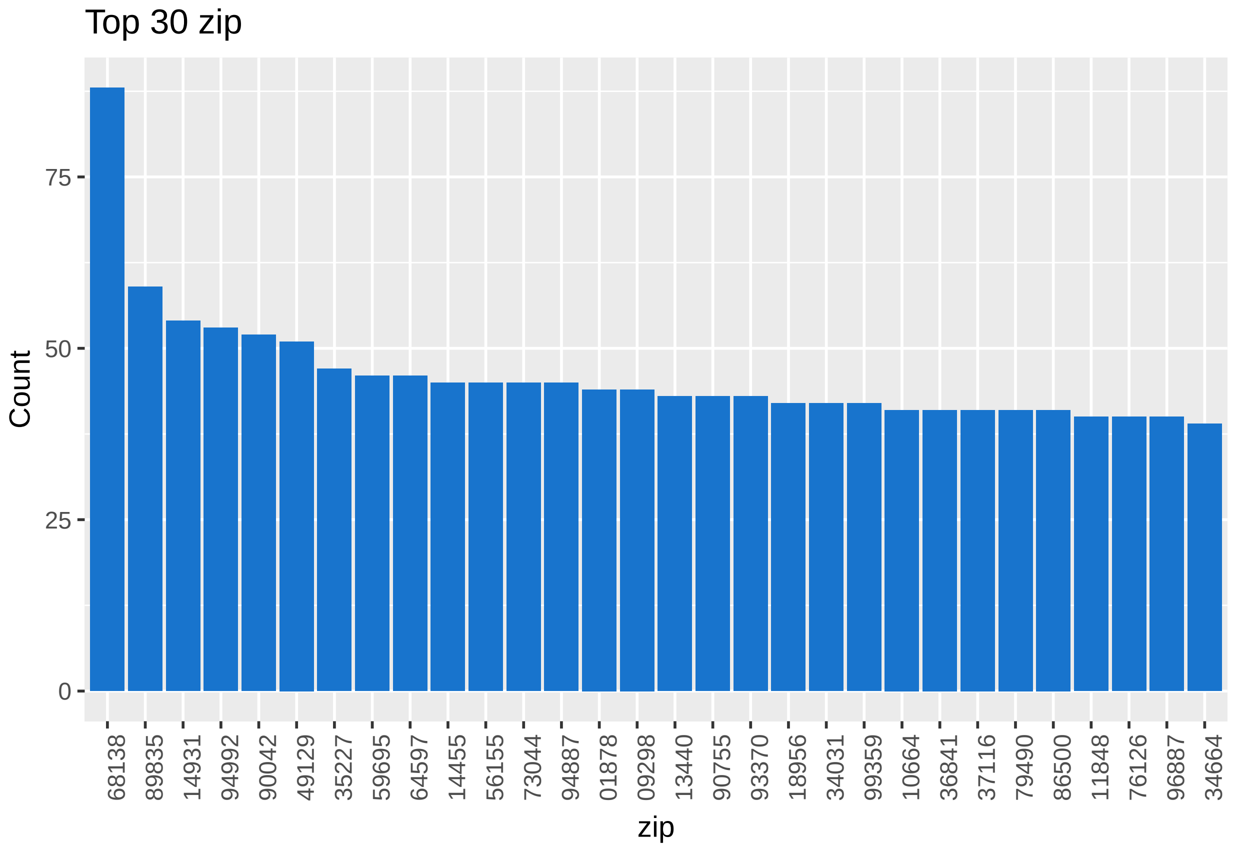
```
df%>%
  group_by(address) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt) )  %>%
  slice(1:30) %>%
  ggplot( aes( x = reorder(address,-cnt), y =  cnt) )+
    geom_bar(stat = "identity",fill = "dodgerblue3")+
    xlab("address")+
    ylab("Count")+
    ggtitle("Top 30 addresses")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



zip

Zip is a categorical variable and “68138” might be a frivolous value in Zip field.

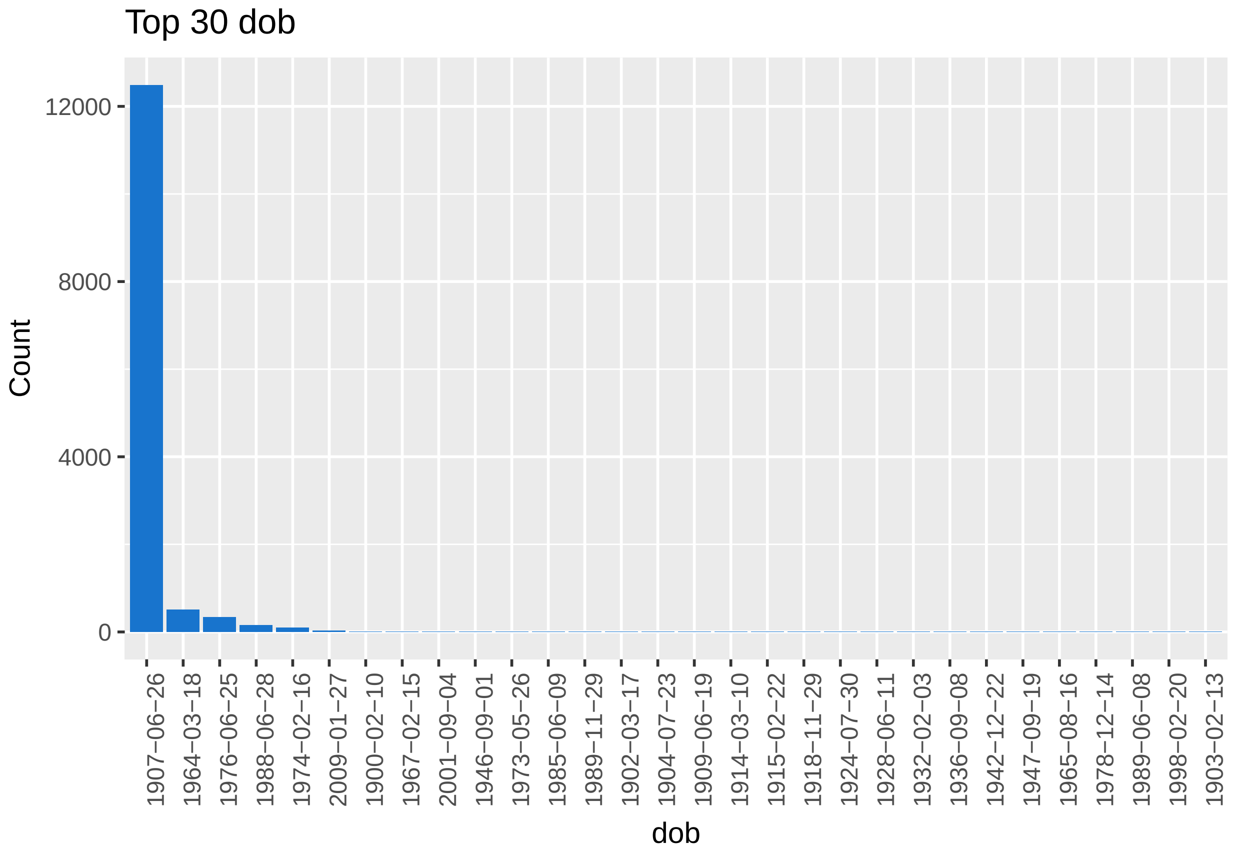
```
df%>%
  group_by(zip) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt) )  %>%
  slice(1:30) %>%
  ggplot( aes( x = reorder(zip,-cnt), y =  cnt) )+
    geom_bar(stat = "identity",fill = "dodgerblue3")+
    xlab("zip")+
    ylab("Count")+
    ggtitle("Top 30 zip")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



dob

dob is a categorical variable.“19070626” might be a frivolous value.

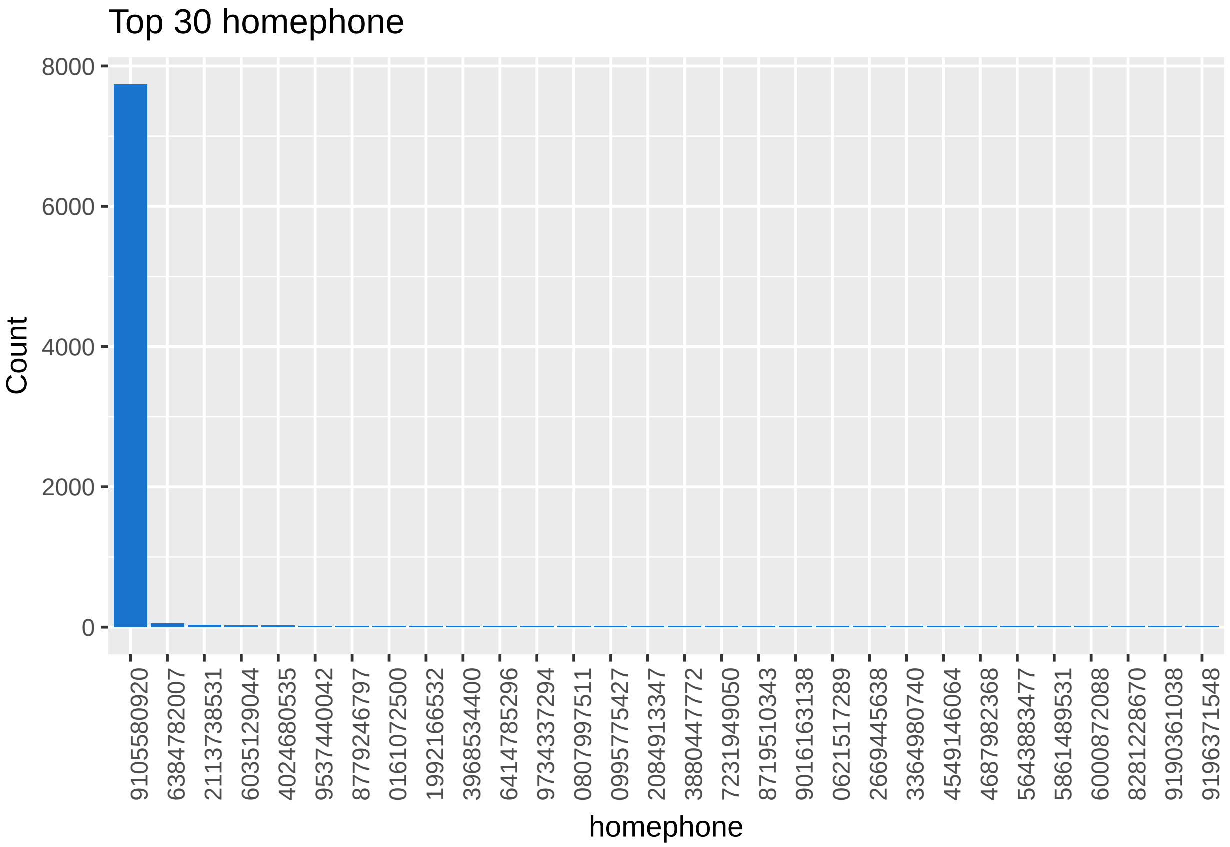
```
df%>%
  group_by(dob) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt) )  %>%
  slice(1:30) %>%
  ggplot( aes( x = reorder(dob,-cnt), y =  cnt) )+
    geom_bar(stat = "identity",fill = "dodgerblue3")+
    xlab("dob")+
    ylab("Count")+
    ggtitle("Top 30 dob")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



homephone

homeophone is a numeric variable.

```
df%>%
  group_by(homephone) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt) )  %>%
  slice(1:30) %>%
  ggplot( aes( x = reorder(homephone,-cnt), y =  cnt) )+
    geom_bar(stat = "identity",fill = "dodgerblue3")+
    xlab("homephone")+
    ylab("Count")+
    ggtitle("Top 30 homephone")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```


Top 30 homephone
Count
8000
6000
4000
2000
0
homephone
9105580920
6384782007
2113738531
6035129044
4024680535
9537440042
8779246797
0161072500
1992166532
3968534400
6414785296
9734337294
0807997511
0995775427
2084913347
3880447772
7231949050
8719510343
9016163138
0621517289
2669445638
3364980740
4549146064
4687982368
5643883477
5861489531
6000872088
8281228670
9190361038
9196371548