Machine Learning and the Faud Credit Card

Henri Makika

7/29/2019

Import packages

```
library(readxl)
library(knitr)
library(lubridate)
library(dplyr)
library(ggplot2)
```

Part 1: File Description

File name: Application dataset Source (url, client…): Number of records: 100,000 Number of fields: 9 fields, 1 index, 2 date, 6 categorial Time frame: 01/01/2015 - 12/31/2015

Import data

```
data = read_excel("~/Videos/Credits Fauds.R/applications100k.xlsx")
summary(data[1:4]) %>% kable(digits = 0)
```

record #	date	ssn	firstname
Min. : 1	Min. :20150101	Min. : 2503	Length:100000
1st Qu.: 25001	1st Qu.:20150401	1st Qu.:255816942	Class :character
Median : 50000	Median :20150701	Median :509886303	Mode :character
Mean : 50000	Mean :20150667	Mean :504629765	NA
3rd Qu.: 75000	3rd Qu.:20150930	3rd Qu.:745870823	NA
Max. :100000	Max. :20151231	Max. :999993079	NA

```
data = read_excel("~/Videos/Credits Fauds.R/applications100k.xlsx")
summary(data[5:9]) %>% kable(digits = 0)
```

lastname	address	zip5	dob	homephone
Length:100000	Length:100000	Min. : 2	Min. :19000101	Min. :6.354e+05
Class :character	Class :character	1st Qu.:25036	1st Qu.:19161129	1st Qu.:2.675e+09
Mode :character	Mode :character	Median :50405	Median :19500920	Median :5.413e+09
NA	NA	Mean :50105	Mean :19516527	Mean :5.303e+09
NA	NA	3rd Qu.:74514	3rd Qu.:19821108	3rd Qu.:8.128e+09
NA	NA	Max. :99999	Max. :20161031	Max. :9.997e+09

Part 2: List of info for each field

Description (continuous, categorical with metric, categorical no metric) % populated

unique values

Min, max, mean, median, mode, standard deviation for continuous Picture (either a distribution, histogram, table…)

Field 1: record

Description: Index of record Number of unique values: 100,000, start from 1, interval is 1, no missing value

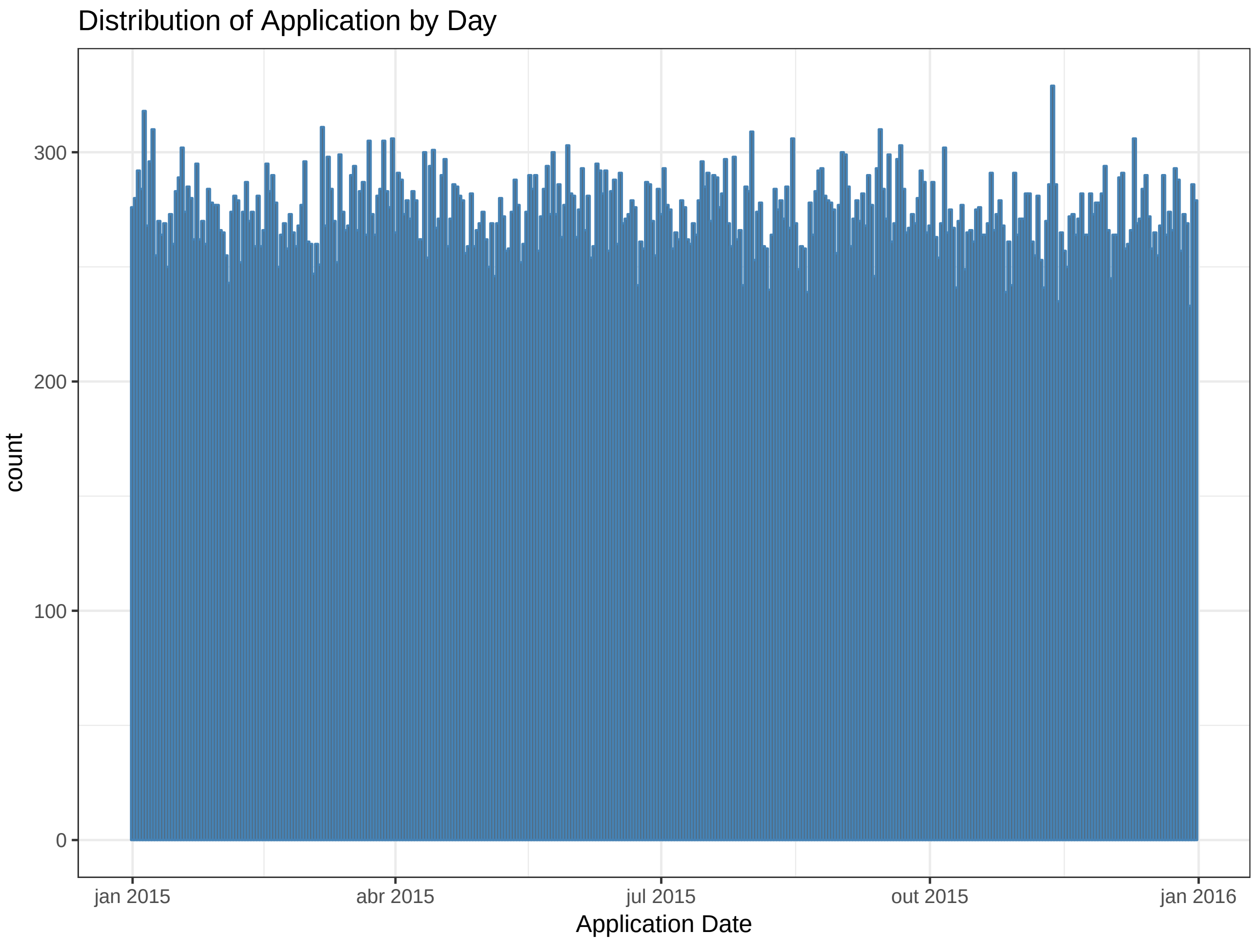
Field 2: date

Description: date of each application record, categorical with metric Percent of Populated: 100%, no missing values Number of unique values: 365, from 01/01/2015 to 12/31/2015

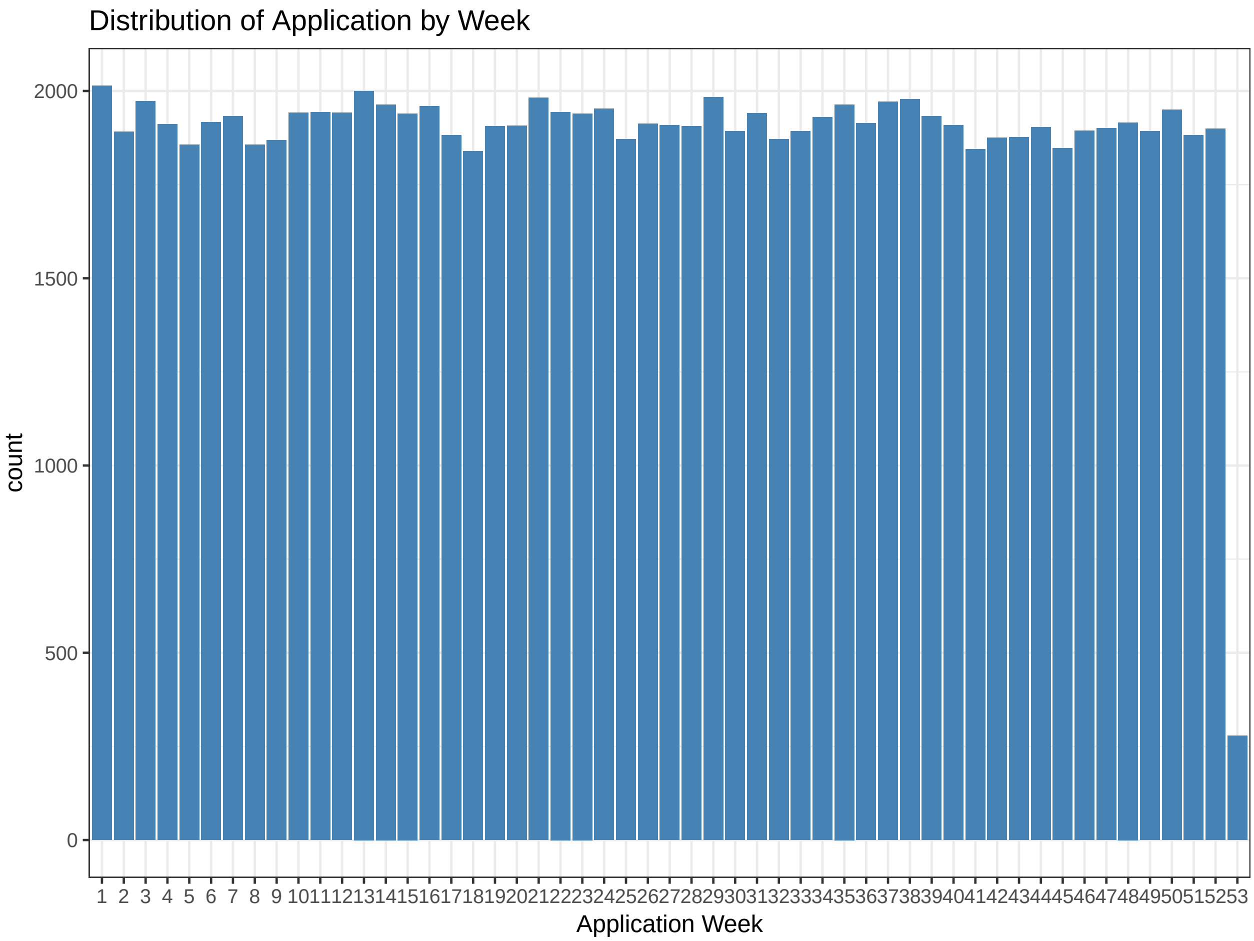
```r
# transfer to standard date style
data$date = ymd(data$date)
# Number of unique values
length(unique(data$date))
```

```
## [1] 365
```

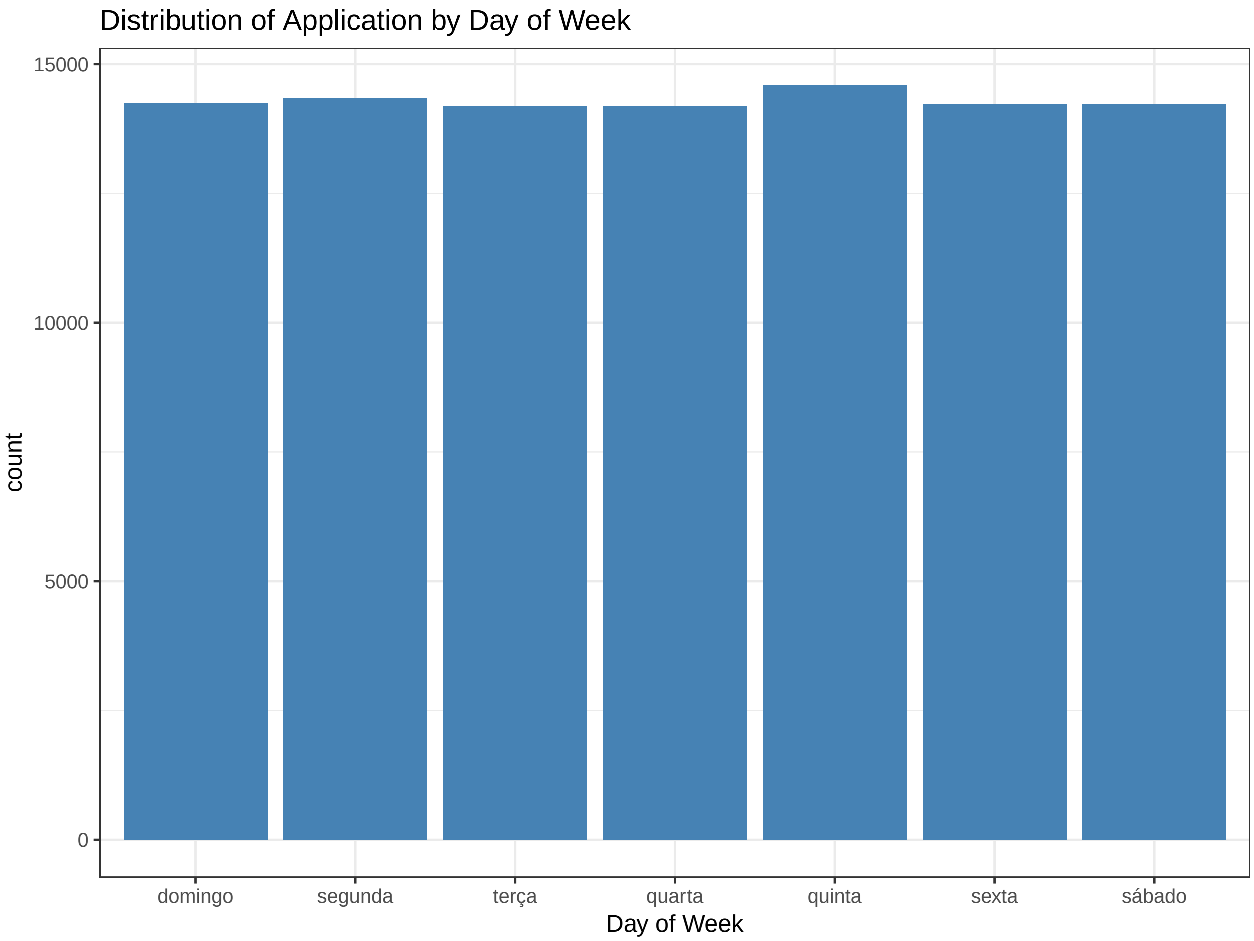
```r
# plot the histogram by day
ggplot(data, aes(x = date)) +
  geom_histogram(bins = 365, color = "steelblue") +
  ggtitle("Distribution of Application by Day") +
  theme_bw() + xlab("Application Date")
```



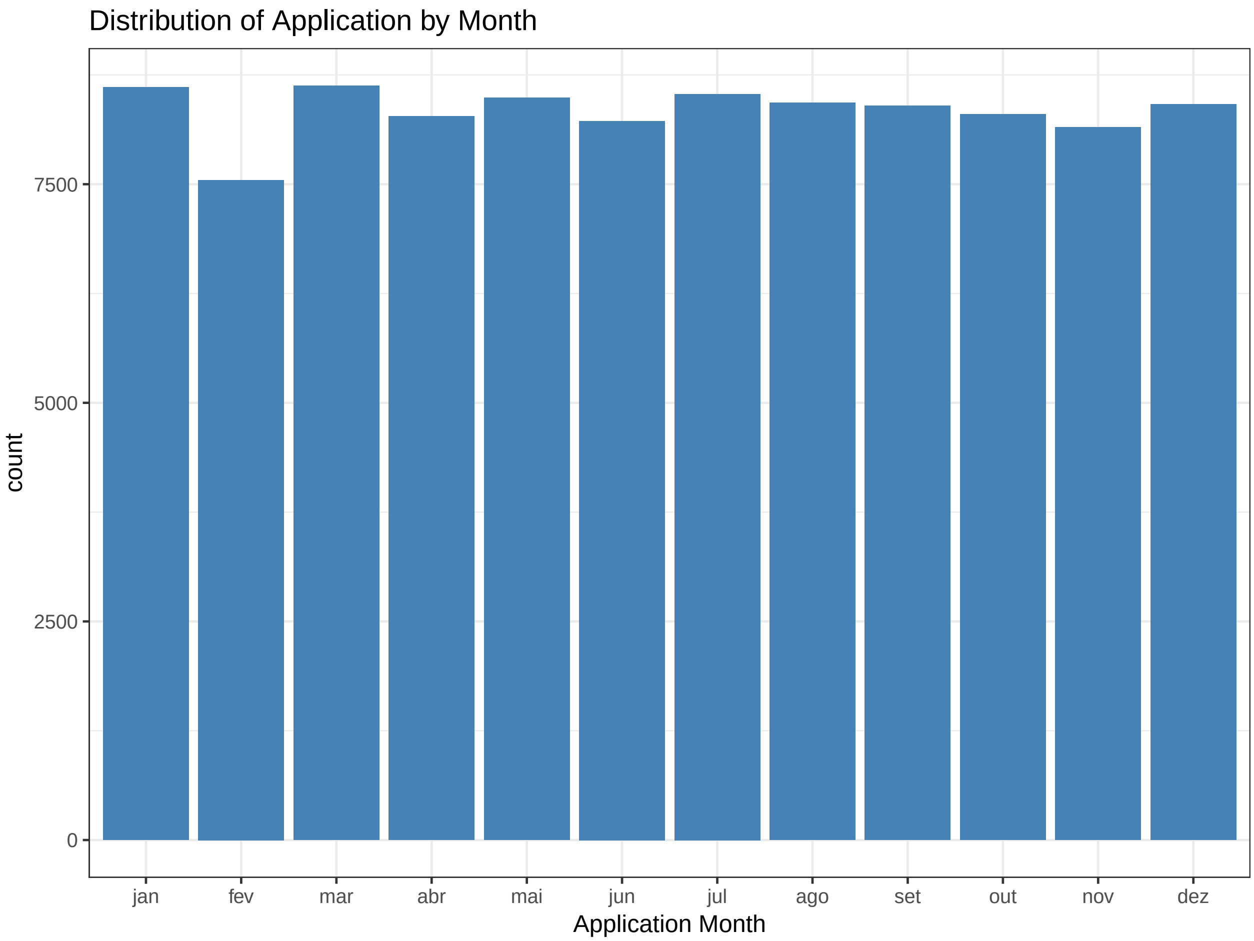
```
# plot the histogram by week
date0 <- data.frame(as.factor(week(data$date)))
names(date0) = c("date0")
ggplot(date0, aes(x = date0)) + geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Application by Week") +
  theme_bw() +
  xlab("Application Week")
```



```r
# plot the histogram by day of week (on average)
date1 = data.frame(as.factor(wday(data$date, label = T, abbr = F)))
names(date1) = c("date1")

ggplot(date1, aes(x = date1)) + geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Application by Day of Week") +
  theme_bw() + xlab("Day of Week")
```



```r
# plot the histogram by month
date2 = data.frame(as.factor(month(data$date, label = T, abbr = T)))
names(date2) = c("date2")

ggplot(date2, aes(x = date2)) + geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Application by Month") +
  theme_bw() + xlab("Application Month")
```



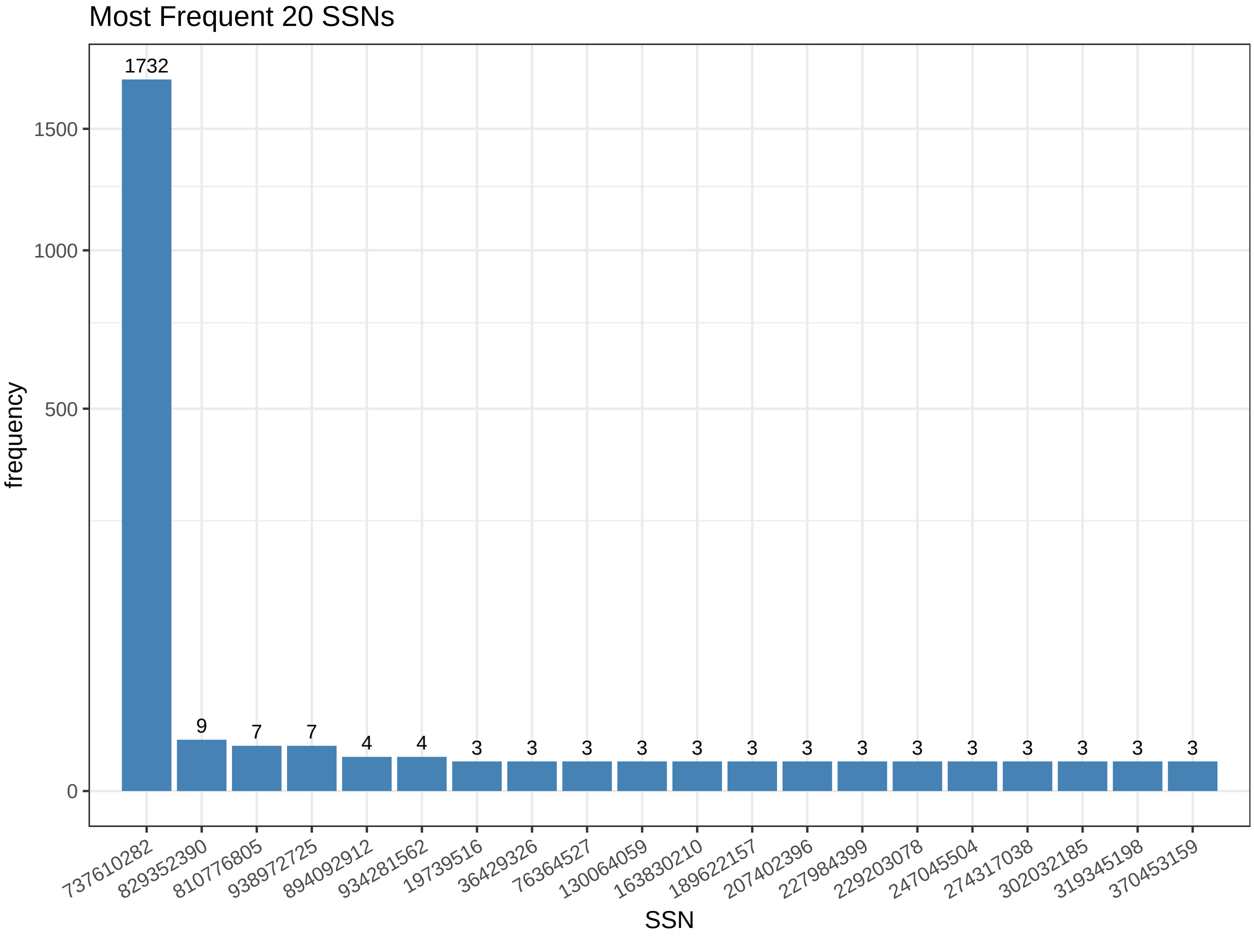
Field 3: ssn

Description: ssn of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 96535

```
# Number of unique values
length(unique(data$ssn))
```

```
## [1] 96535
```

```
# Find the most 20 frequently used ssn
ssn = data %>%
  group_by(ssn) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
ssn1 = ssn[1:20, ]

ggplot(ssn1, aes(x = reorder(ssn, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 SSNs") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("SSN")
```



```
max(data$ssn) ## 9 9999 3079
```

```
## [1] 999993079
```

```
min(data$ssn) ## 2503
```

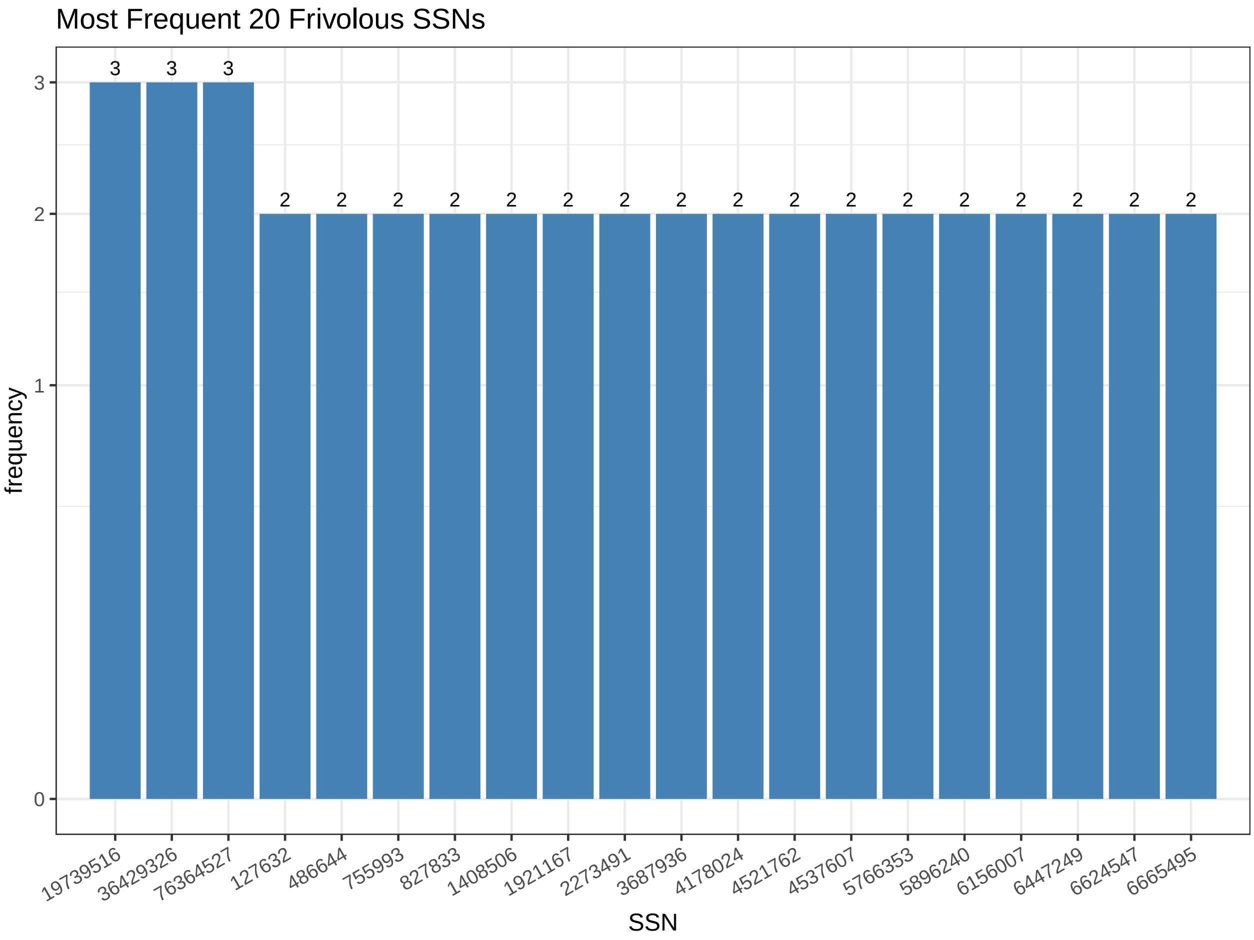
```
## [1] 2503
```

```
ssn_f1 = data.frame(data$ssn)
ssn_f = ssn_f1[ssn_f1 < 1e8]
length(ssn_f)
```

```
## [1] 9805
```

```
ssn_f = data.frame(ssn_f)
names(ssn_f) = "ssn"
View(ssn_f)

ssn_f1 = ssn_f %>%
  group_by(ssn) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
ssn_ff = ssn_f1[1:20, ]
```

```
ggplot(ssn_ff, aes(x = reorder(ssn, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Frivolous SSNs") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("SSN")
```



Field 4: firstname

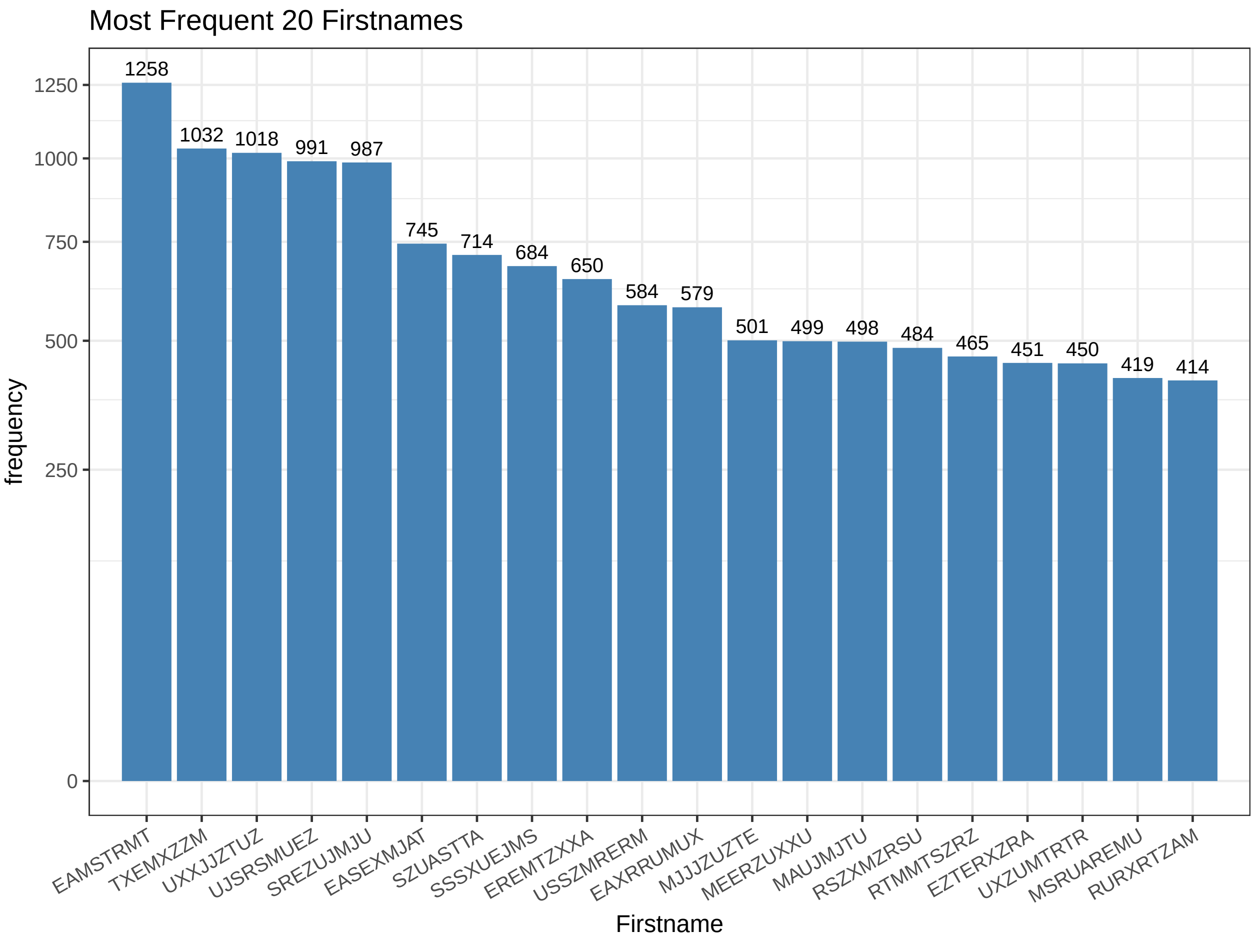
Description: firstname of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 16576

```
# Number of unique values
length(unique(data$firstname))
```

```
## [1] 16576
```

```
# Find the most 20 frequently used firstname
firstname = data %>%
  group_by(firstname) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
```

```
firstname1 = firstname[1:20, ]

ggplot(firstname1, aes(x = reorder(firstname, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Firstnames") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Firstname")
```



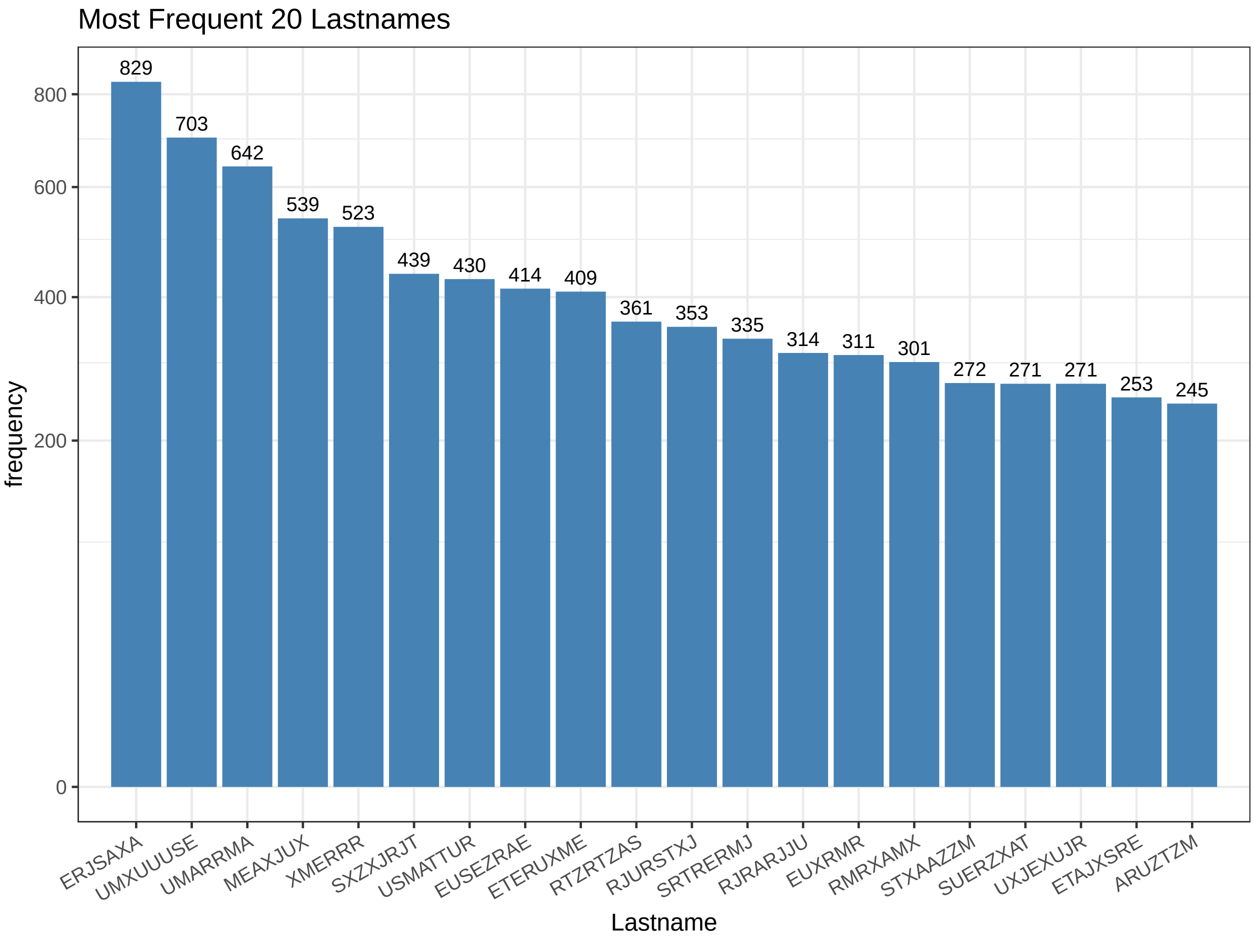
Field 5: lastname

Description: lastname of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 36312

```
# Number of unique values
length(unique(data$lastname))
```

```
## [1] 36312
```

```
# Find the most 20 frequently used lastname
lastname = data %>%
  group_by(lastname) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
lastname1 = lastname[1:20, ]

ggplot(lastname1, aes(x = reorder(lastname, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Lastnames") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Lastname")
```



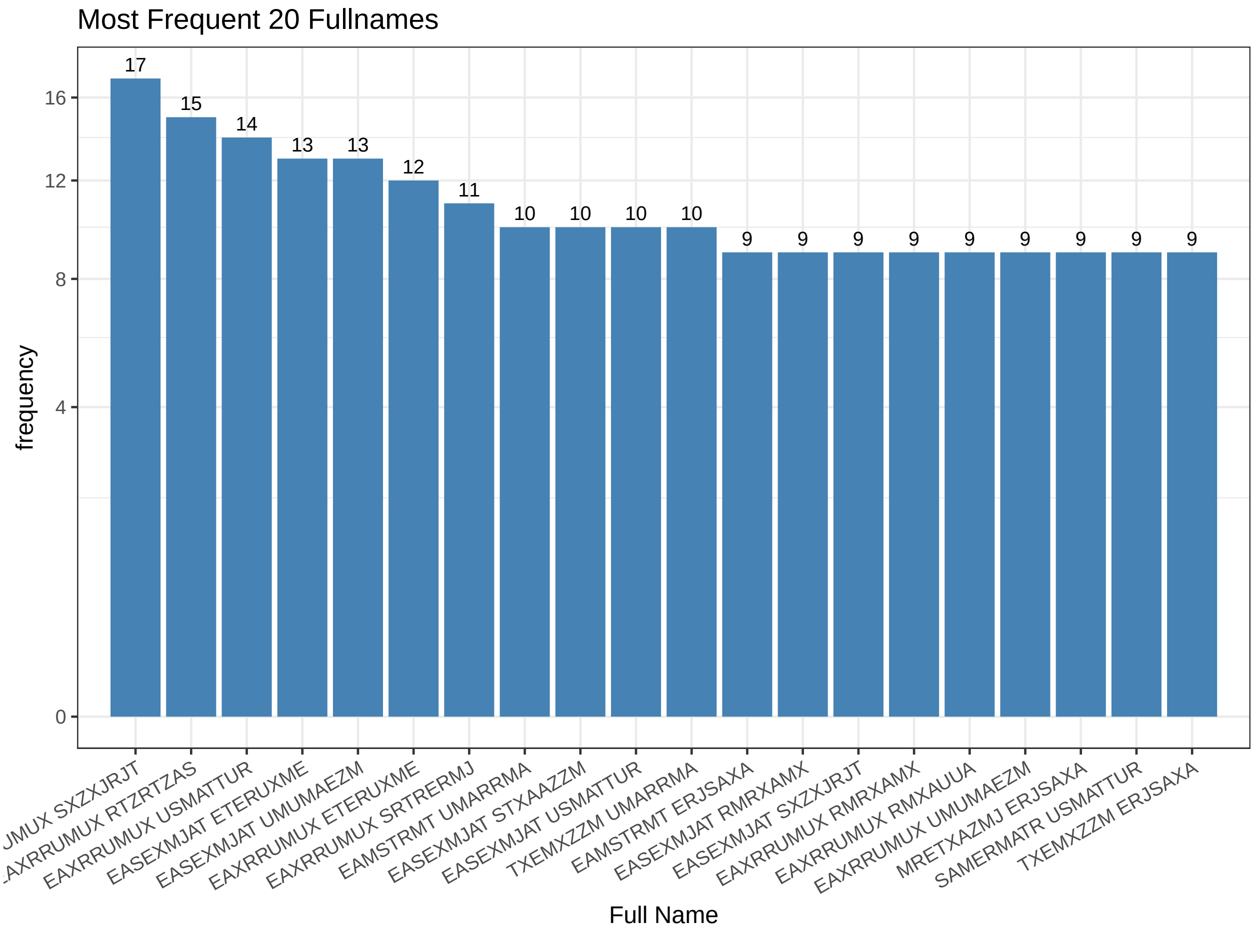
Field 5* : Fullname

Description: Fullname of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 93726

```
## Plot the most 20 frequent name (first + last)
name = data.frame(paste(data$firstname, data$lastname, sep = " "))
names(name) = c("Name")
# Number of unique values
length(unique(name$Name))
```

```
## [1] 93726
```

```
Fullname = name %>%
  group_by(Name) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
Fullname1 = Fullname[1:20, ]

ggplot(Fullname1, aes(x = reorder(Name, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Fullnames") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Full Name")
```



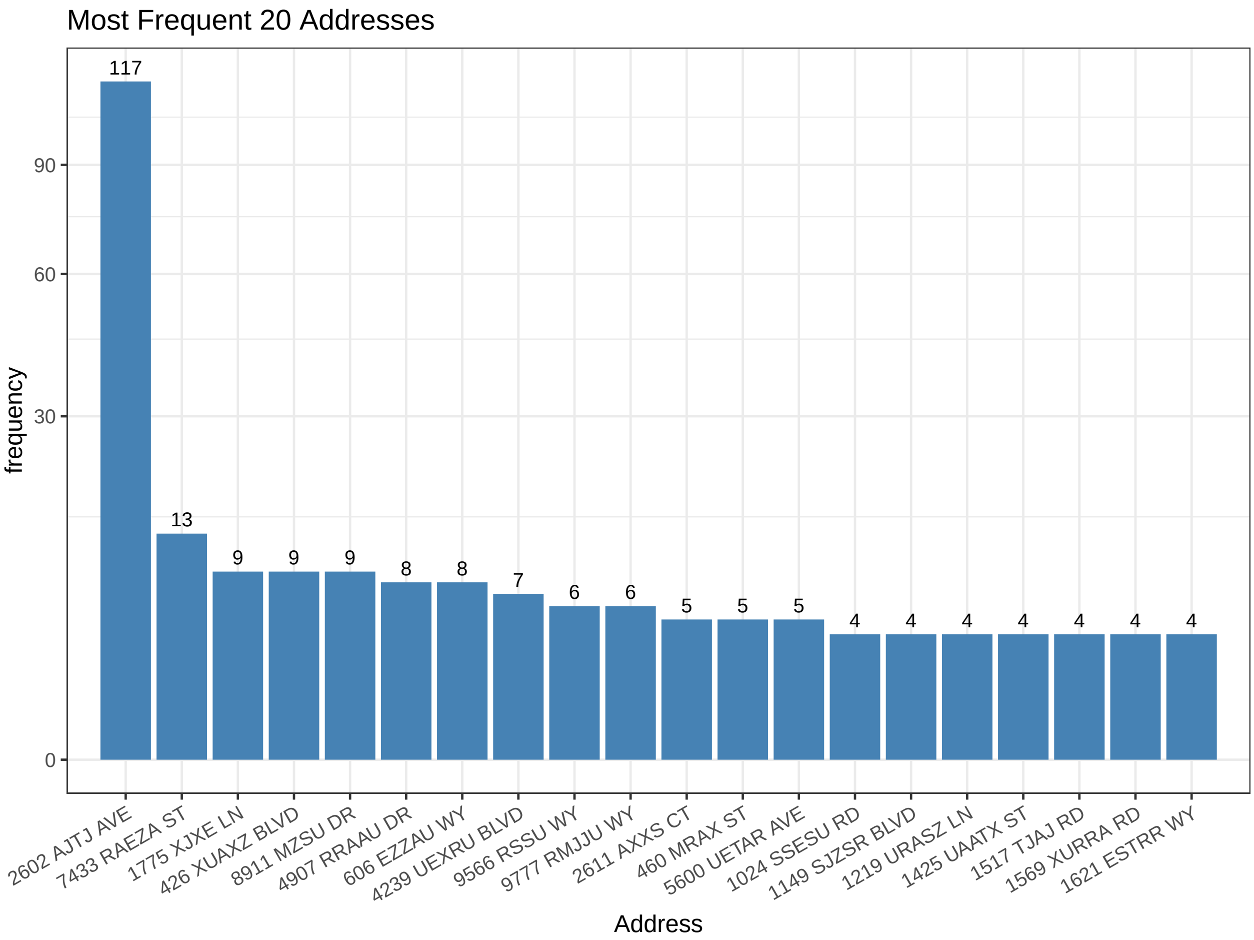
Field 6: address

Description: address of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 97563 Frivolous item: 2602 AJTJ AVE

```
# Number of unique values
length(unique(data$address))
```

```
## [1] 97563
```

```r
# Find the most 20 frequently used address
address = data %>%
  group_by(address) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
address1 = address[1:20, ]

ggplot(address1, aes(x = reorder(address, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Addresses") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Address")
```



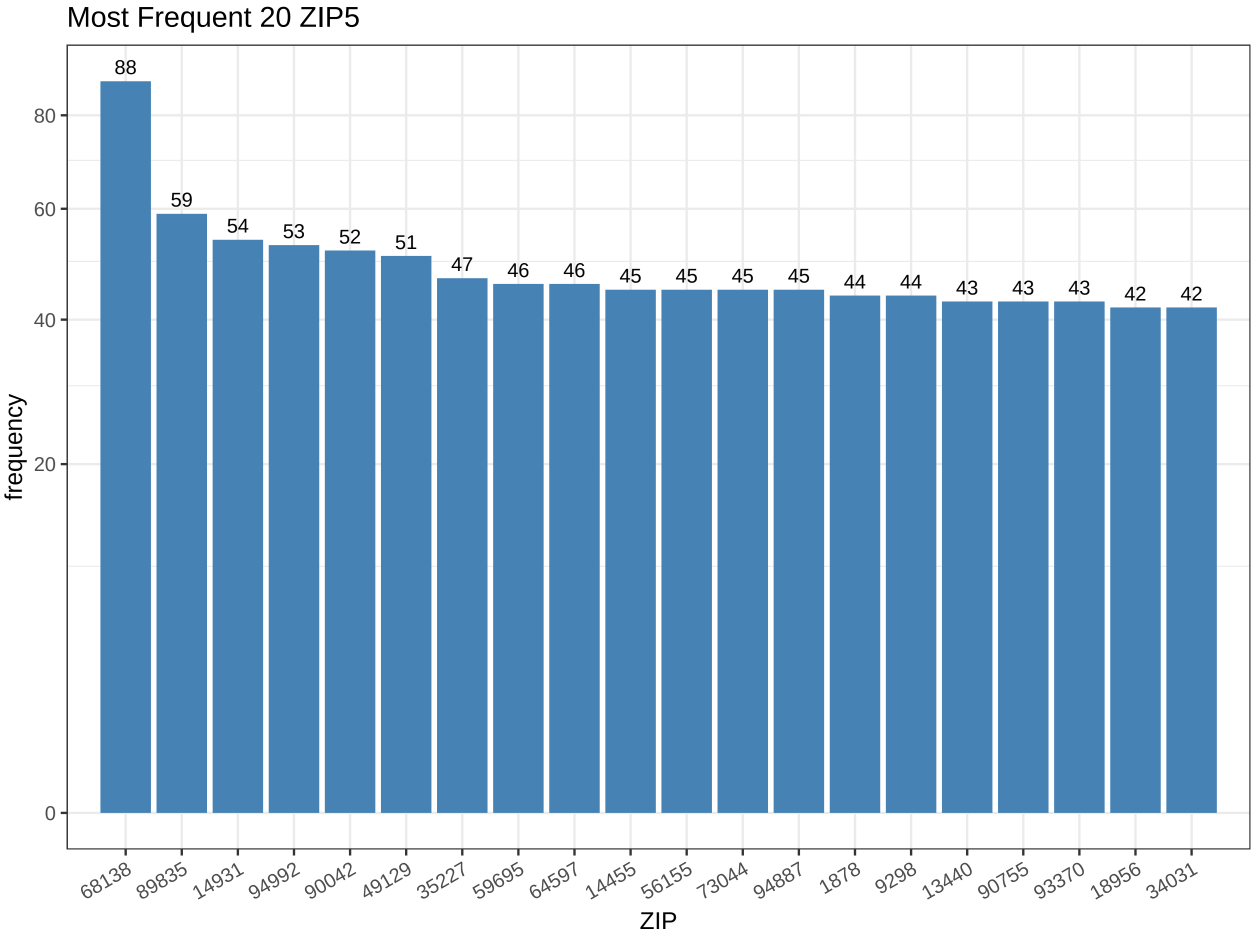
Field 7: zip5

Description: zip5 of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 16547

```r
# Number of unique values
length(unique(data$zip5))
```

```
## [1] 16547
```

```r
# Find the most 20 frequently used zip5
zip5 = data %>%
  group_by(zip5) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
zip51 = zip5[1:20, ]
ggplot(zip51, aes(x = reorder(zip5, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 ZIP5") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("ZIP")
```



```r
max(data$zip5) ## 99999
```

```
## [1] 99999
```

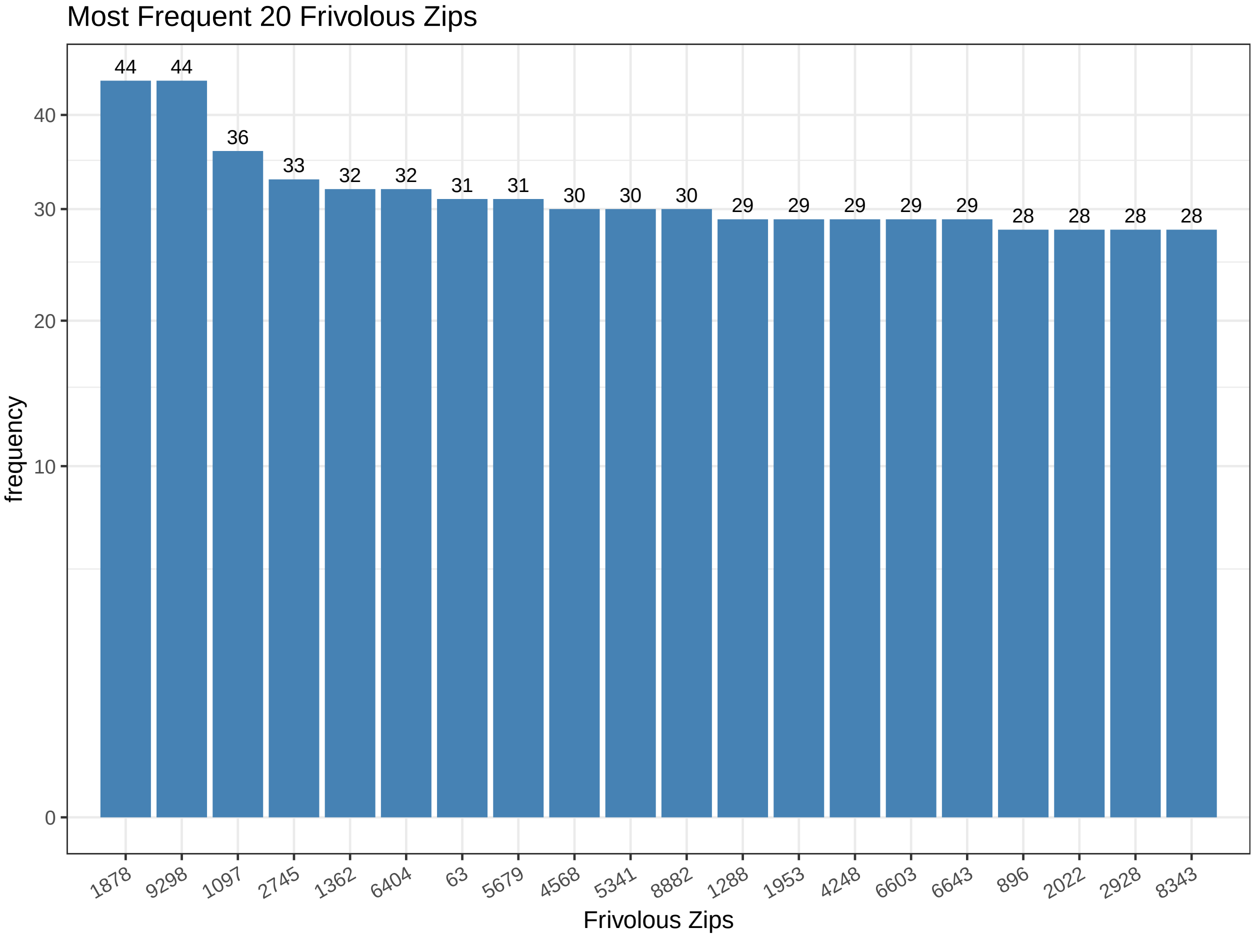
```r
min(data$zip5) ## 2
```

```
## [1] 2
```

```
zip_f1 = data.frame(data$zip5)
zip_f = zip_f1[zip_f1 < 1e4]
length(zip_f)
```

```
## [1] 10360
```

```
zip_f = data.frame(zip_f)
names(zip_f) = "zip"

zip_ff1 = zip_f %>%
  group_by(zip) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
zip_ff = zip_ff1[1:20, ]

ggplot(zip_ff, aes(x = reorder(zip, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Frivolous Zips") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Frivolous Zips")
```



Field 8: dob

Description: birth date of each applicant, categorical with metric Percent of Populated: 100%, no missing values Number of unique values: 36816, from 01/01/1900 to 10/31/2016

```
# transfer to standard date style
data$dob = ymd(data$dob)
# summary of dob
length(unique(data$dob))
```

```
## [1] 36816
```

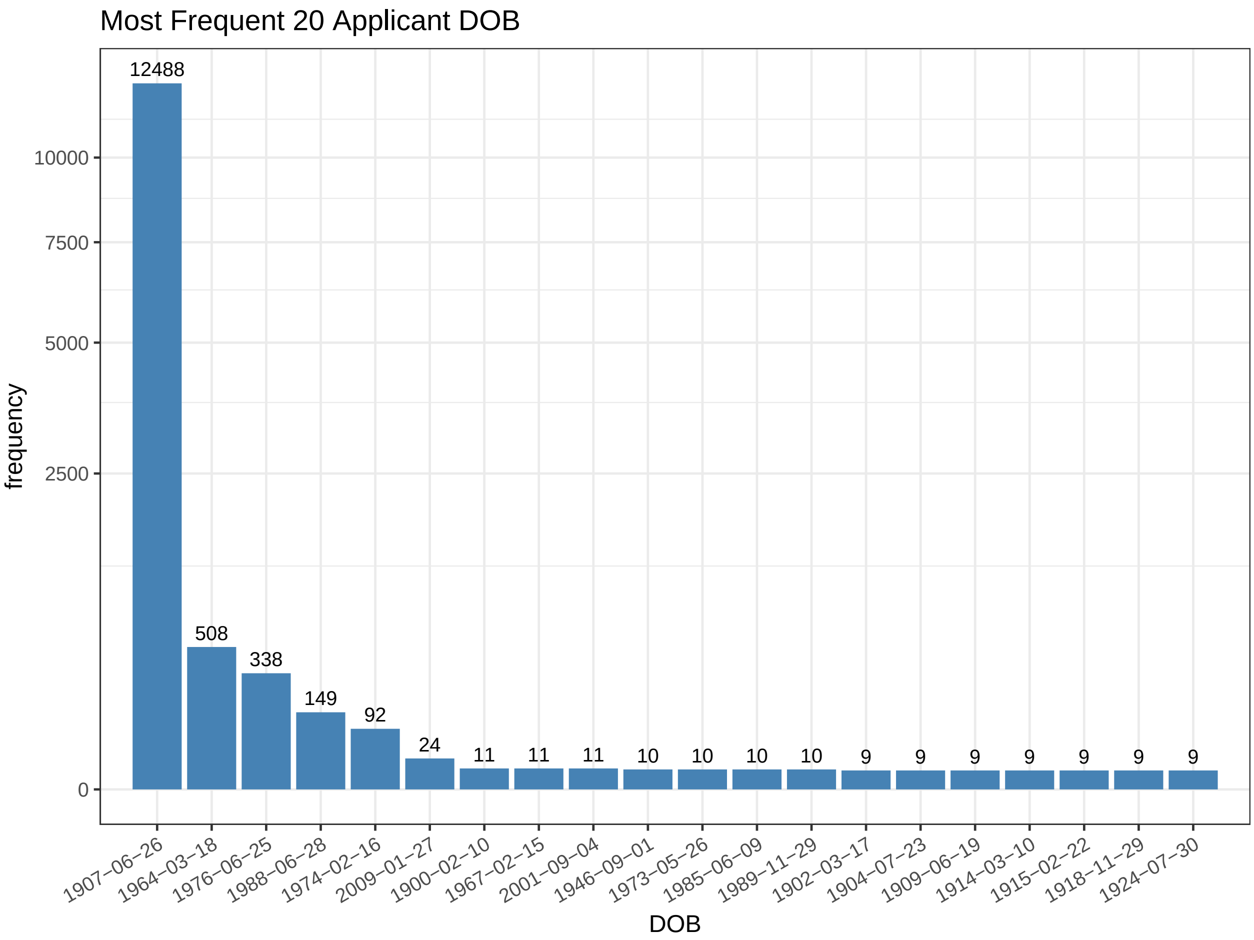
```
max(data$dob)
```

```
## [1] "2016-10-31"
```

```
min(data$dob)
```

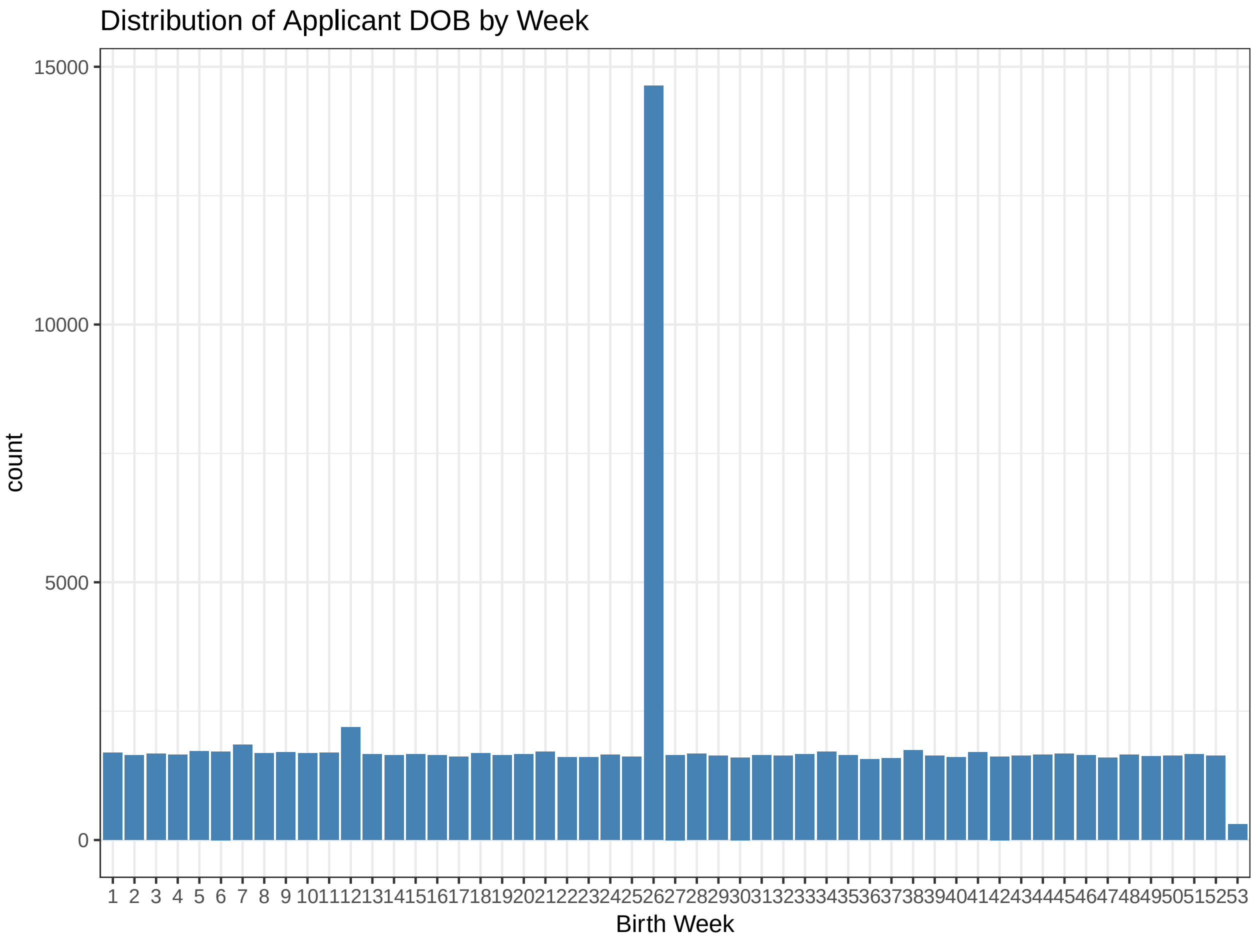
```
## [1] "1900-01-01"
```

```
# Find the most 20 frequently used DOB
dob = data %>%
  group_by(dob) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
dob1 = dob[1:20, ]

ggplot(dob1, aes(x = reorder(dob, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Applicant DOB") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("DOB")
```

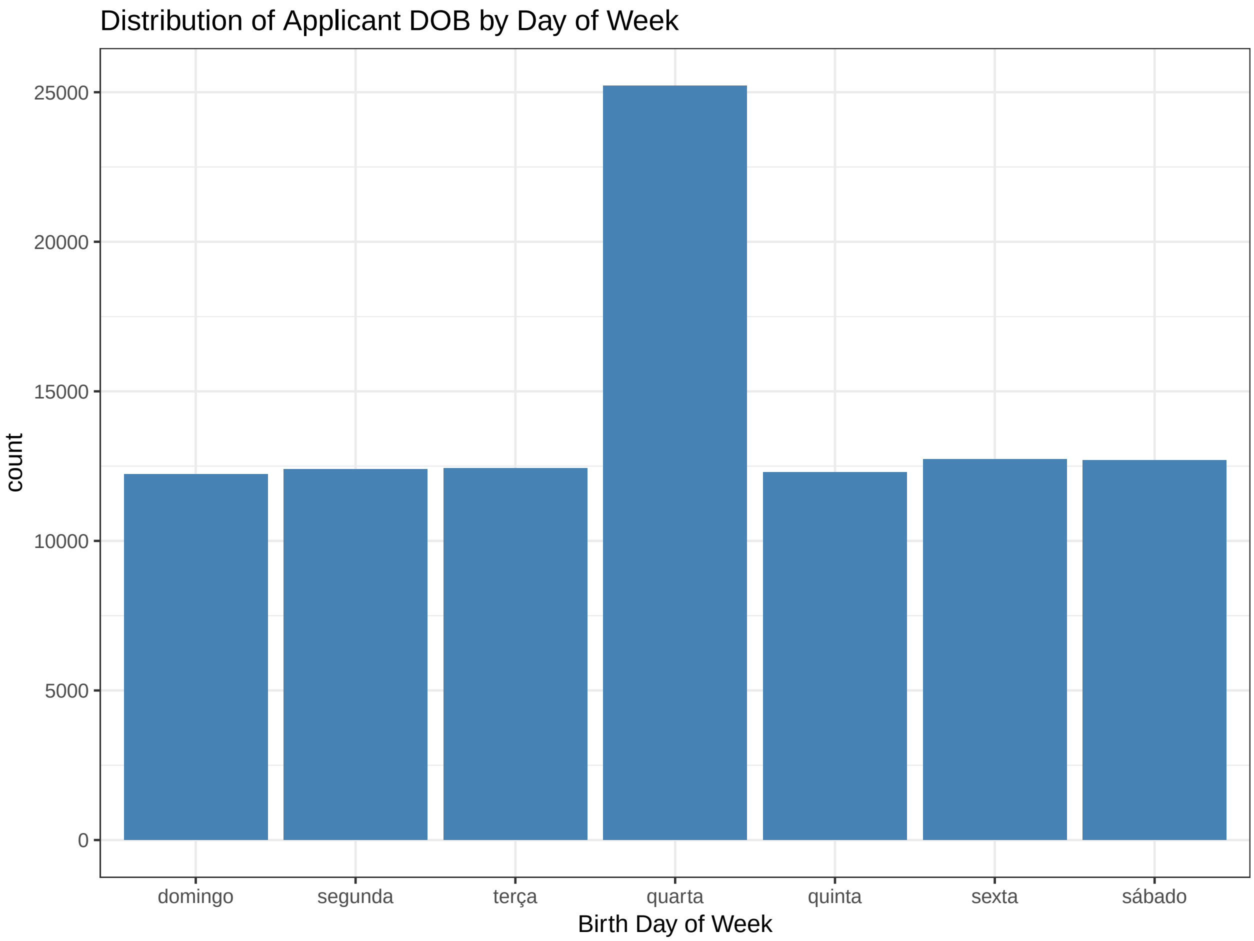


Most Frequent 20 Applicant DOB

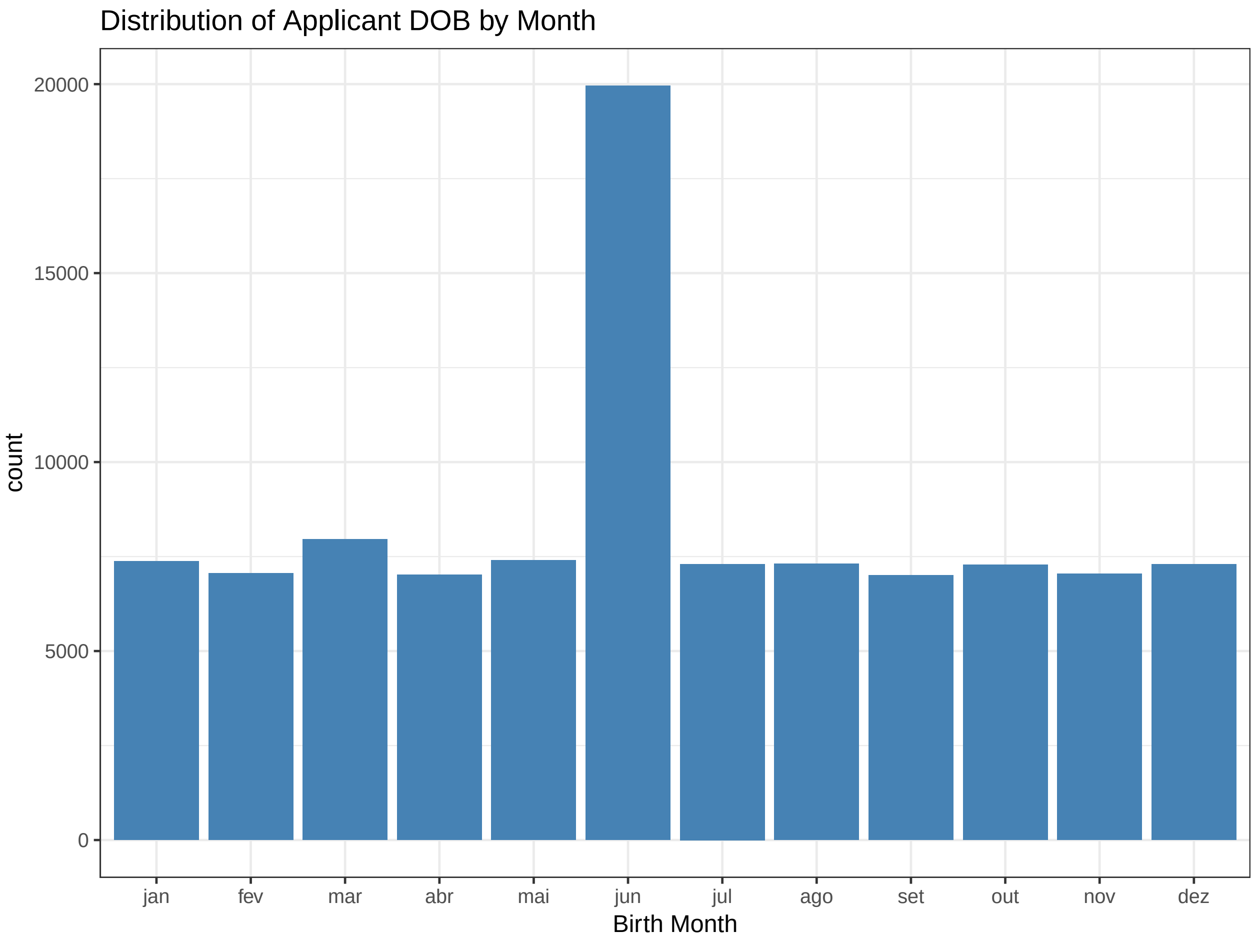
```r
# plot the histogram by week
dob2 = data.frame(as.factor(week(data$dob)))
names(dob2) = c("dob2")

ggplot(dob2, aes(x = dob2)) +
  geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Applicant DOB by Week") +
  theme_bw() +
  xlab("Birth Week")
```



```
# plot the histogram by day of week (on average)
dob3 = data.frame(as.factor(wday(data$dob, label = T, abbr = F)))
names(dob3) = c("dob3")

ggplot(dob3, aes(x = dob3)) +
  geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Applicant DOB by Day of Week") +
  theme_bw() +
  xlab("Birth Day of Week")
```



```
# plot the histogram by month
dob4 = data.frame(as.factor(month(data$dob, label = T, abbr = T)))
names(dob4) = c("dob4")

ggplot(dob4, aes(x = dob4)) +
  geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Applicant DOB by Month") +
  theme_bw() +
  xlab("Birth Month")
```



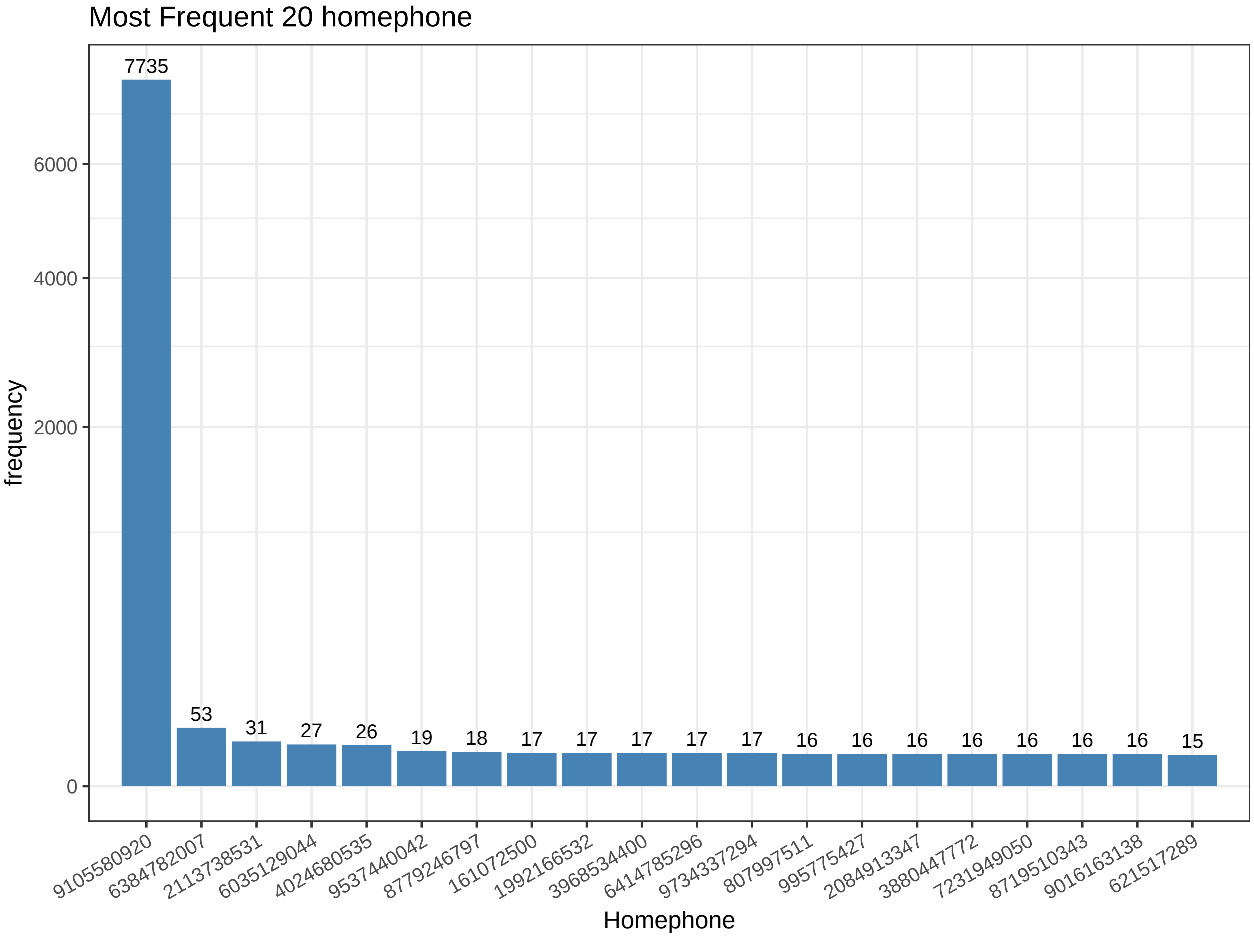
Field 9: homephone

Description: homephone of each application record, categorical with no metric Percent of Populated: 100%, no missing values Number of unique values: 22181

```
# Number of unique values
length(unique(data$homephone))
```

```
## [1] 22181
```

```
# Find the most 20 frequently used homephone
homephone = data %>%
  group_by(homephone) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
homephone1 = homephone[1:20, ]

ggplot(homephone1, aes(x = reorder(homephone, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 homephone") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Homephone")
```



```
max(data$homephone)
```

```
## [1] 9996906703
```

```
min(data$homephone)
```

```
## [1] 635392
```

```
max(data$homephone) ## 99999
```

```
## [1] 9996906703
```

```
min(data$homephone) ## 2
```

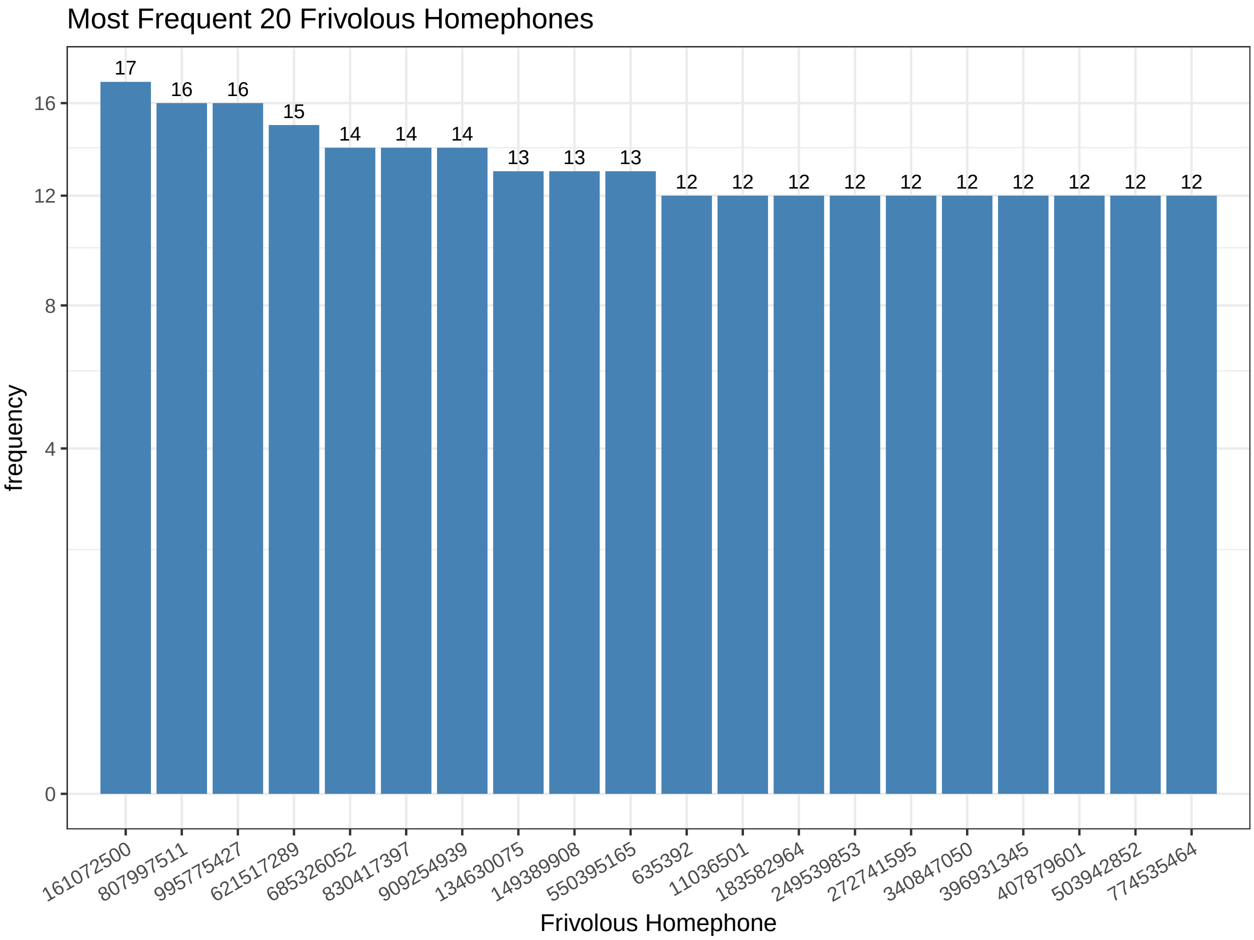
```
## [1] 635392
```

```
homephone_f1 = data.frame(data$homephone)
homephone_f = homephone_f1[homephone_f1 < 1e9]
length(homephone_f)
```

```
## [1] 9221
```

```
homephone_f = data.frame(homephone_f)
names(homephone_f) = "homephone"
```

```
homephone_ff1 = homephone_f %>%
  group_by(homephone) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency))
homephone_ff = homephone_ff1[1:20, ]

ggplot(homephone_ff, aes(x = reorder(homephone, -frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  coord_trans(y = "sqrt") +
  ggtitle("Most Frequent 20 Frivolous Homephones") +
  geom_text(aes(label = frequency, y = frequency), size = 3, vjust = -0.5) +
  xlab("Frivolous Homephone")
```


Most Frequent 20 Frivolous Homephones
frequency
Frivolous Homephone