Product Applications Data

Data Quality Report

03/02/2017

**Summary**

**File Description:**

Production Application Data is a dataset containing the records of 100,000 product applications. It includes information about the date the application made, and the SSN number, name, address, zip code, data of birth and home phone number that each applicant self-reported.

**File Name:**

Applications 100k.xlsx

**Data Source:**

This dataset is the simulated data based on real product application records.

**Number of Records:**

100,000 records

**Number of Fields:**

9 variables in total – 6 categorical variables, 1 text variable, 2 date variables

|  |  |
| --- | --- |
| **Field Name** | **Data Type** |
| record | Categorical variable |
| date | Date variable |
| ssn | Categorical variable |
| firstname | Categorical variable |
| lastname | Categorical variable |
| address | Text variable |
| zip5 | Categorical variable |
| dob | Date variable |
| homephone | Categorical variable |

**Time of Records:**

Jan 1st 2015 – Dec 31st 2015

**Fields Explanation**

**Field 1**

**Field Name:** record

**Description:**

“record” is a categorical variable. It works as the ordinal reference number for each application record.

**Unique Values:**

100% populated with 100,000 unique values, ranging from 1 to 100,000. No repeated values or missing values exist.

**Field 2**

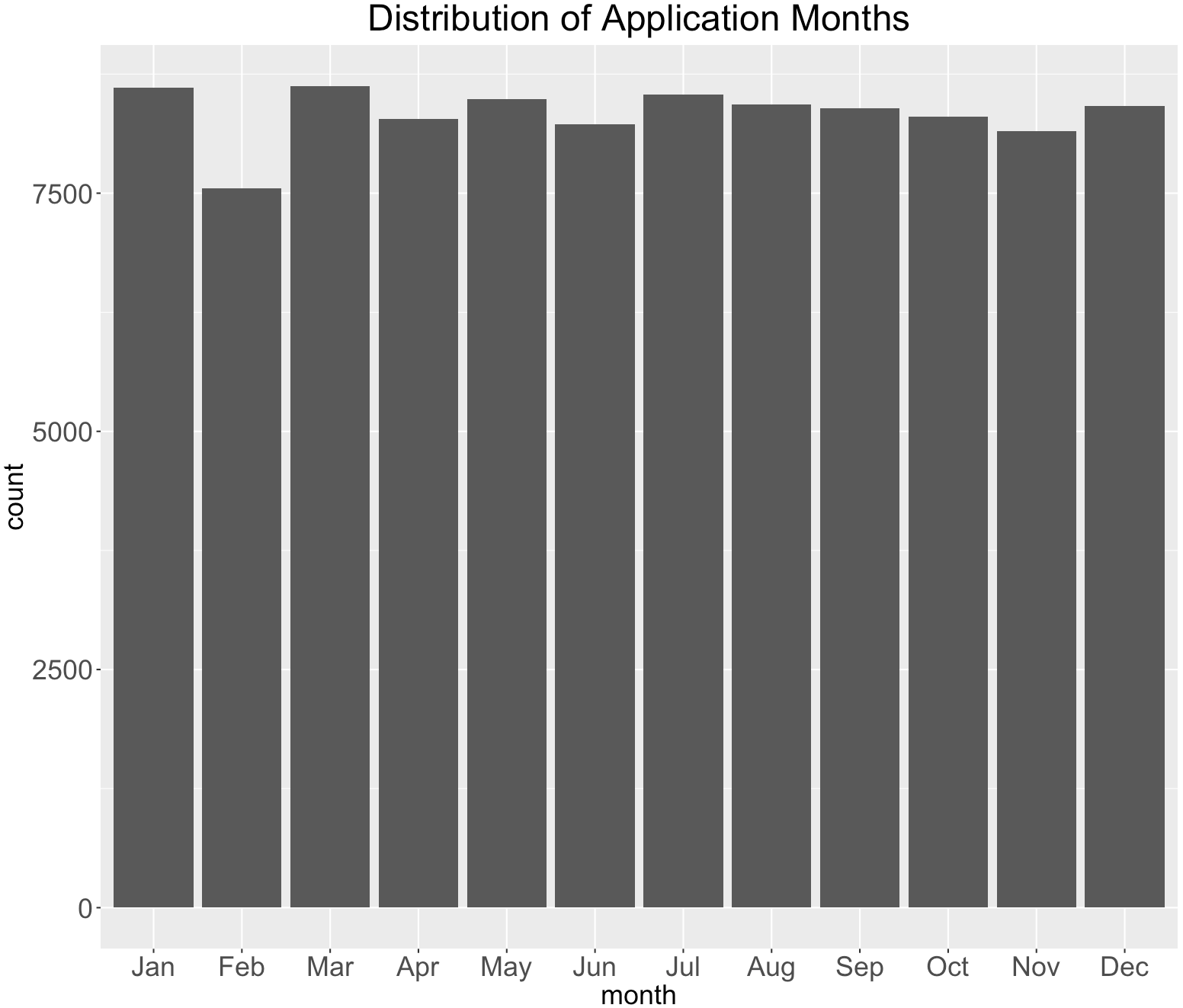
**Field Name:** date

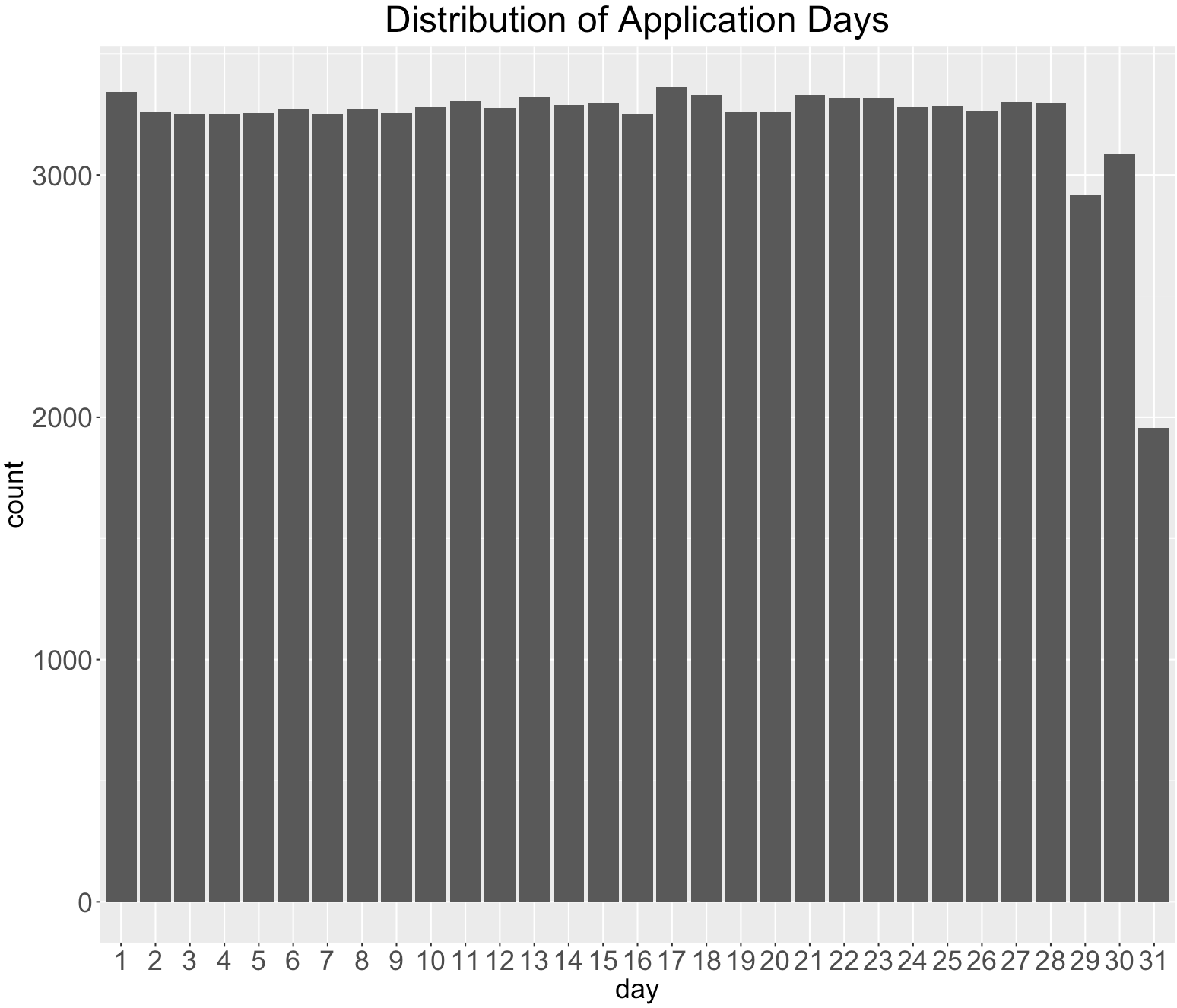
**Description:**

“date” is a date variable, indicating the date that the application was made.

**Unique Values:**

100% populated with 365 unique values, i.e. every day in year 2015. Plot the distribution of the months and dates as below, we can see that the two distributions follow the natural occurrence distribution.





**Field 3**

**Field Name:** ssn

**Description:**

“ssn” is a categorical variable, indicating the applicant’s ssn number he or she self-reported.

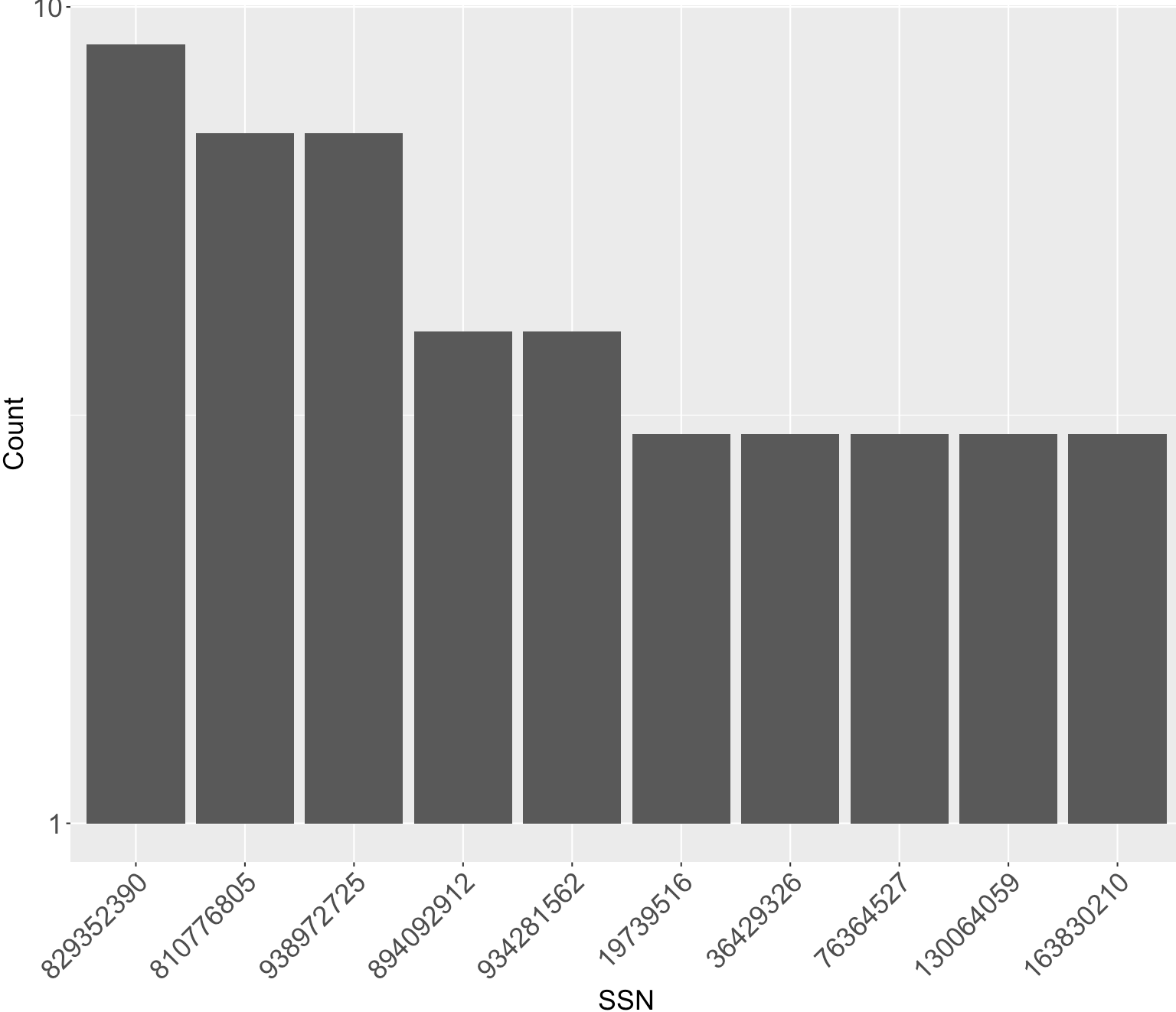
**Unique Values:**

100% populated with 96,535 unique values. Values in this field have lengths from 4 to 9. A ssn shorter than 9 digits indicates there are leading 0s in the ssn record. For example, ssn record “2503” is actually “000002503”. The top 10 frequent appeared ssn records are shown below. It is obvious that the ssn record “737610282” is a frivolous record, which appears 173 times more than the second highest appeared record.

The top 10 most frequent records are listed below:

|  |  |
| --- | --- |
| **SSN** | **Frequency** |
| 737610282 | 1.73% |
| 938972725 | 0.01% |
| 810776805 | 0.01% |
| 829352390 | 0.01% |
| 473311863 | 0.00% |
| 189622157 | 0.00% |
| 163830210 | 0.00% |
| 407447121 | 0.00% |
| 118692079 | 0.00% |
| 849295926 | 0.00% |

The distribution of the top 10 records (excluding the most frequent record “737610282”):



**Field 4**

**Field Name:** firstname

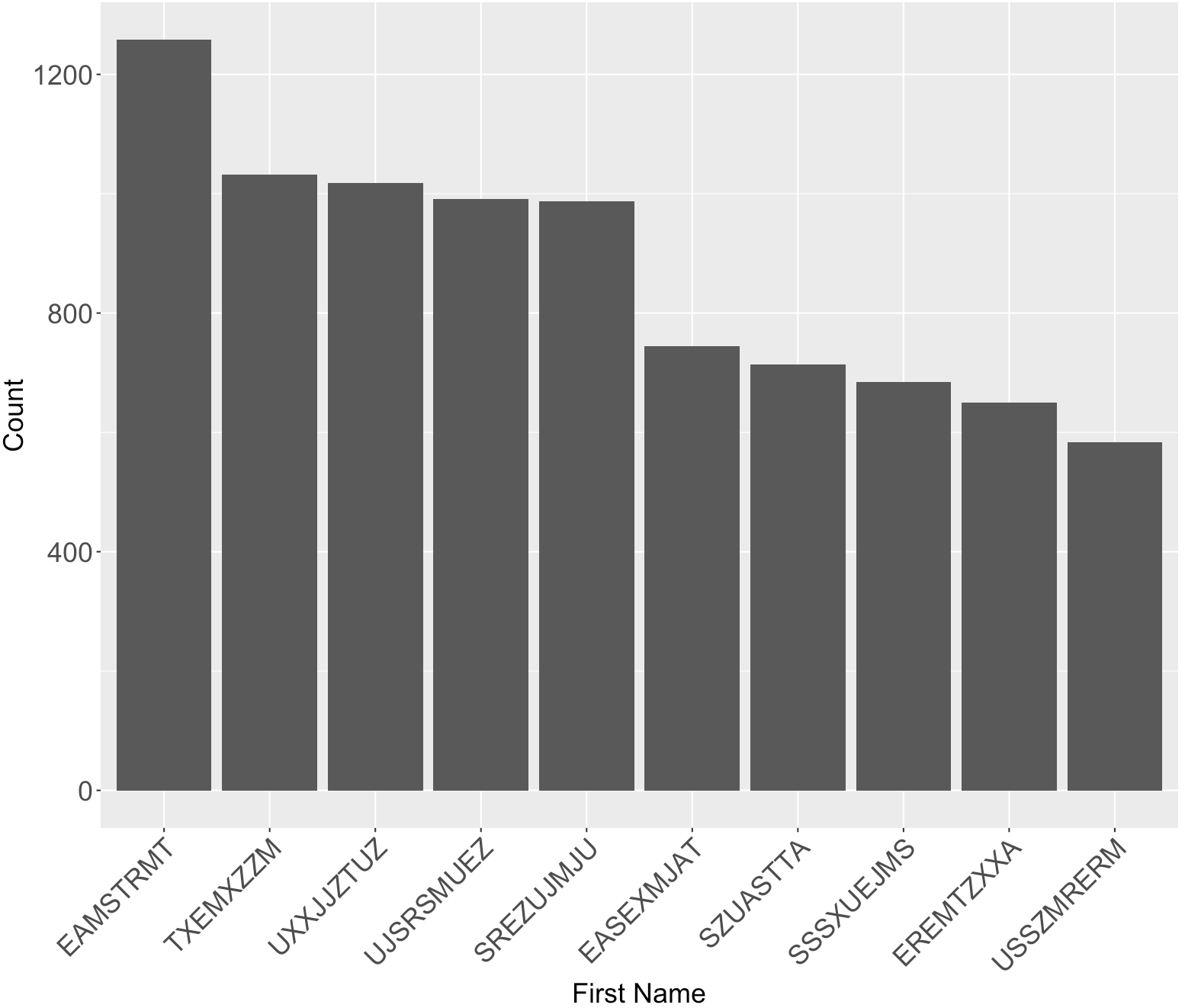
**Description:**

“firstname” is a categorical variable, indicating the applicant’s first name he or she self-reported.

**Unique Values:**

100% populated with 16,576 unique values. The top 10 frequent appeared first name records are shown below.

|  |  |
| --- | --- |
| **First Name** | **Frequency** |
| EAMSTRMT | 1.26% |
| TXEMXZZM | 1.03% |
| UXXJJZTUZ | 1.02% |
| UJSRSMUEZ | 1.00% |
| SREZUJMJU | 1.00% |
| EASEXMJAT | 0.08% |
| SZUASTTA | 0.07% |
| SSSXUEJMS | 0.07% |
| EREMTZXXA | 0.07% |
| USSZMRERM | 0.06% |



**Field 5**

**Field Name:** lastname

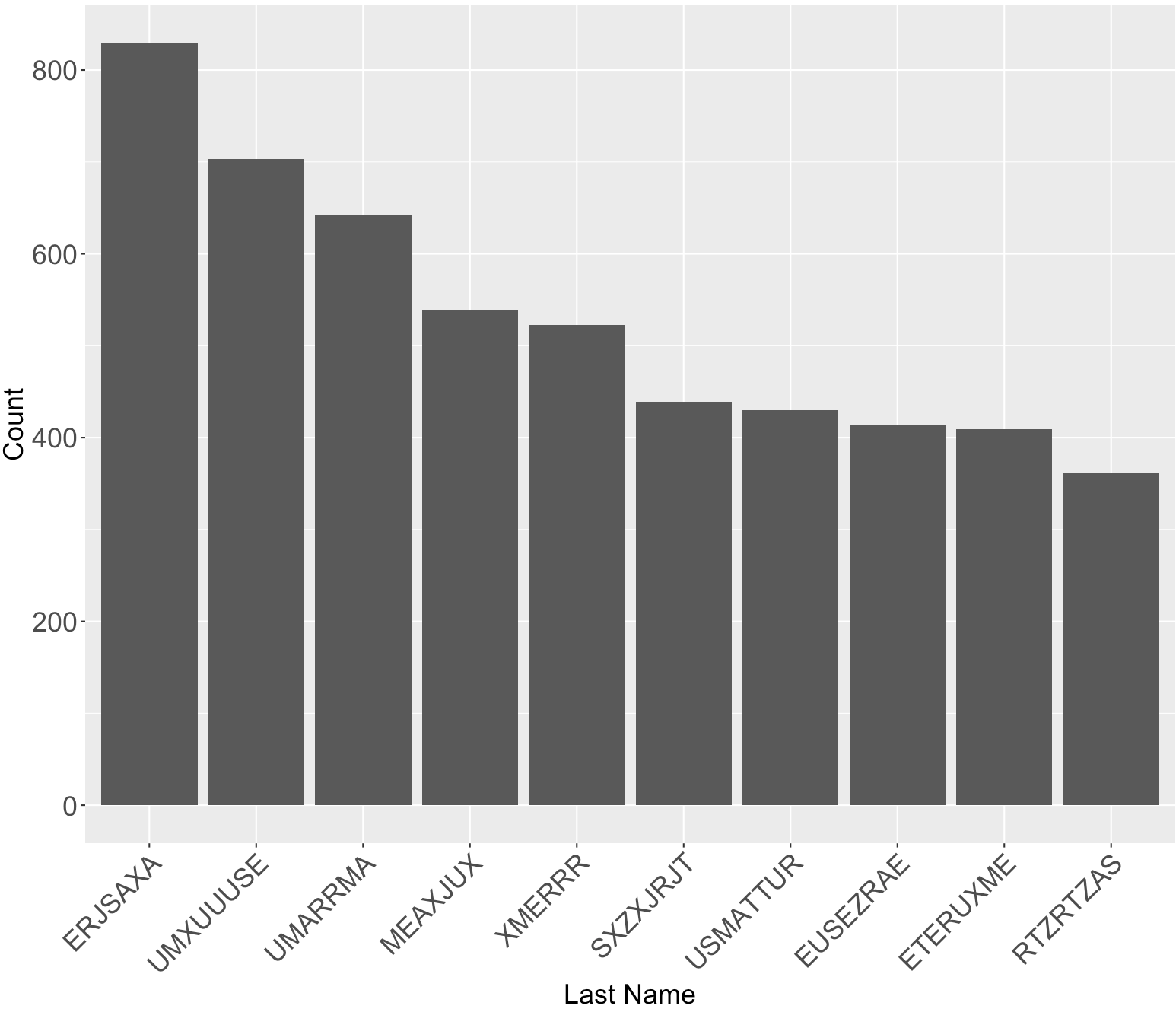
**Description:**

“lastname” is a categorical variable, indicating the applicant’s last name he or she self-reported.

**Unique Values:**

100% populated with 36,312 unique values. The top 10 frequent appeared last name records are shown below.

|  |  |
| --- | --- |
| **Last Name** | **Frequency** |
| ERJSAXA | 0.83% |
| UMXUUUSE | 0.70% |
| UMARRMA | 0.64% |
| MEAXJUX | 0.54% |
| XMERRR | 0.52% |
| SXZXJRJT | 0.44% |
| USMATTUR | 0.43% |
| EUSEZRAE | 0.41% |
| ETERUXME | 0.41% |
| RTZRTZAS | 0.36% |



**Field 6**

**Field Name:** address

**Description:**

“address” is a text variable that contains inputs from applicants of their address.

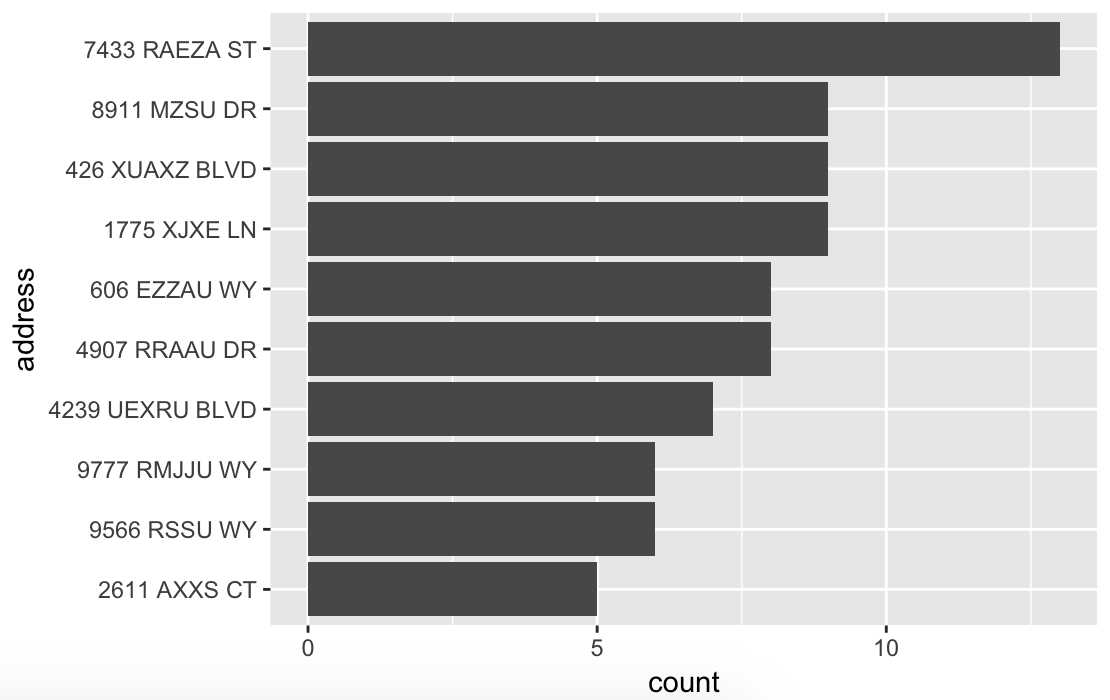
**Unique Values:**

100% populated with 97,563 unique values from 100,000 records. The most frequent address “2602 AJTJ AVE” appeared 117 times, which accounts for 0.12% of all records. This address is likely to be a frivolous address.

The top 10 most frequent records are listed below:

|  |  |
| --- | --- |
| **Address** | **Frequency** |
| 2602 AJTJ AVE | 0.12% |
| 7433 RAEZA ST | 0.01% |
| 1775 XJXE LN | 0.01% |
| 426 XUAXZ BLVD | 0.01% |
| 8911 MZSU DR | 0.01% |
| 4907 RRAAU DR | 0.01% |
| 606 EZZAU WY | 0.01% |
| 4239 UEXRU BLVD | 0.01% |
| 9566 RSSU WY | 0.01% |
| 9777 RMJJU WY | 0.01% |

The distribution of the top 10 records (excluding the most frequent record “2602 AJTJ AVE”):



**Field 7**

**Field Name:** zip5

**Description:**

“zip5” is a categorical variable that contains inputs from applicants of their zip codes.

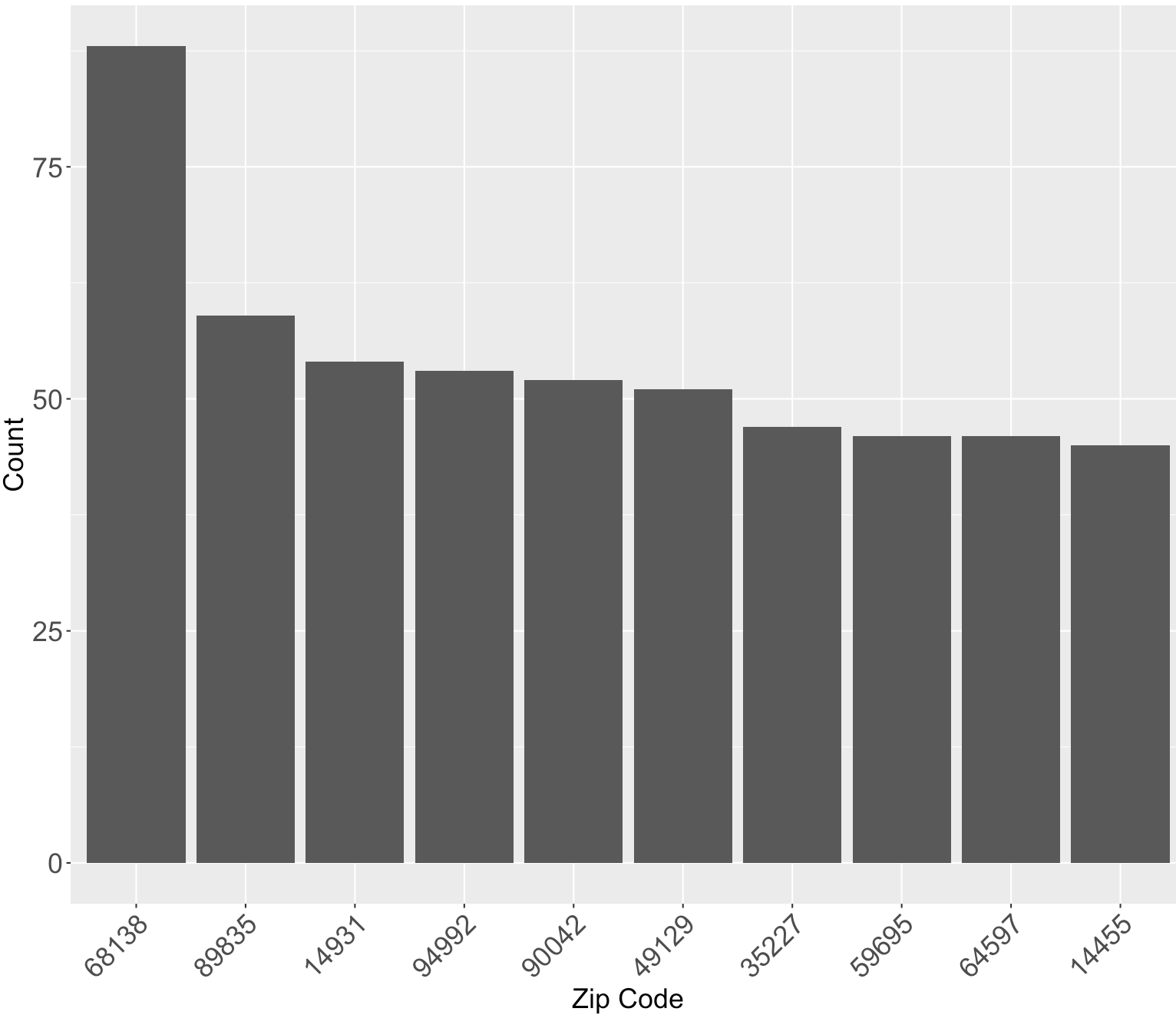
**Unique Values:**

100% populated with 16,547 unique values from 100,000 records. No missing values. The most frequent zip code “68138” appeared 823 times, which accounts for 0.09% of all records.

The top 10 most frequent records are listed below:

|  |  |
| --- | --- |
| **Zip5** | **Frequency** |
| 68138 | 0.09% |
| 89835 | 0.06% |
| 14931 | 0.05% |
| 94992 | 0.05% |
| 90042 | 0.05% |
| 49129 | 0.05% |
| 35227 | 0.05% |
| 59695 | 0.05% |
| 64597 | 0.05% |
| 14455 | 0.05% |

The distribution of the top 10 records:



**Field 8**

**Field Name:** dob

**Description:**

“dob” is a date variable that contains inputs from applicants of their date of birth.

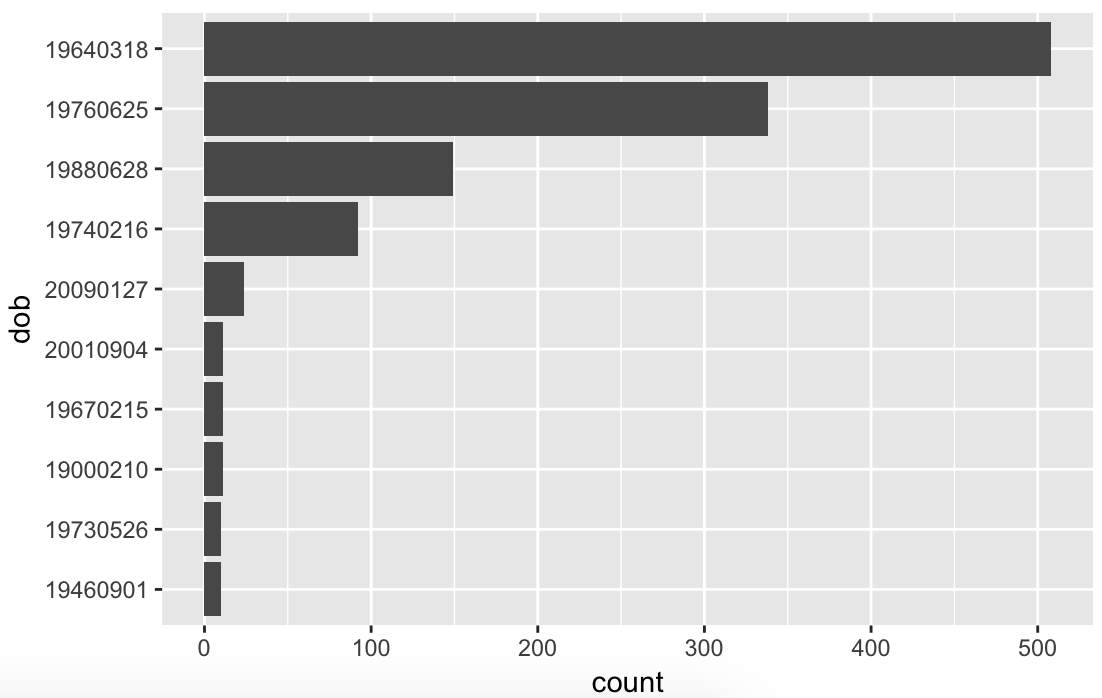
**Unique Values:**

100% populated with 36,816 unique values from 100,000 records. No missing values. The most frequent date of birth “19070626” appeared 12,488 times, which accounts for 12.49% of all records. This date of birth is likely to be a frivolous value.

The top 10 most frequent records are listed below:

|  |  |
| --- | --- |
| **dob** | **Frequency** |
| 19070626 | 12.49% |
| 19640318 | 0.51% |
| 19760625 | 0.34% |
| 19880628 | 0.15% |
| 19740216 | 0.09% |
| 20090127 | 0.02% |
| 19000210 | 0.01% |
| 19670215 | 0.01% |
| 20010904 | 0.01% |
| 19460901 | 0.01% |

The distribution of the top 10 records (excluding the most frequent record “19070626”):



**Field 9**

**Field Name:** homephone

**Description:**

“homephone” is a categorical variable that contains inputs from applicants of their home phone number.

**Unique Values:**

100 populated with 22,181 unique values from 100,000 records. No missing values. The most frequent home phone number “9105580920” appeared 7735 times, which accounts for 7.74% of all records. This homephone number is likely to be a frivolous value.

The top 10 most frequent records are listed below:

|  |  |
| --- | --- |
| **homephone** | **Frequency** |
| 9105580920 | 7.74% |
| 6384782007 | 0.05% |
| 2113738531 | 0.03% |
| 6035129044 | 0.03% |
| 4024680535 | 0.03% |
| 9537440042 | 0.02% |
| 8779246797 | 0.02% |
| 161072500 | 0.02% |
| 1992166532 | 0.02% |
| 3968534400 | 0.02% |

The distribution of the top 10 records (excluding the most frequent record “9105580920”):

