Data Quality Report

1. **Data Description**
   The data contains individuals personal identifying information (PII) that is used to identify potential credit card application fraud. It contains records from Jaunuary 1st 2017 to December 31st 2017. There are 10 fields and 1,000,000 records available.

2. **Summary Tables**
   a. Numerical Table

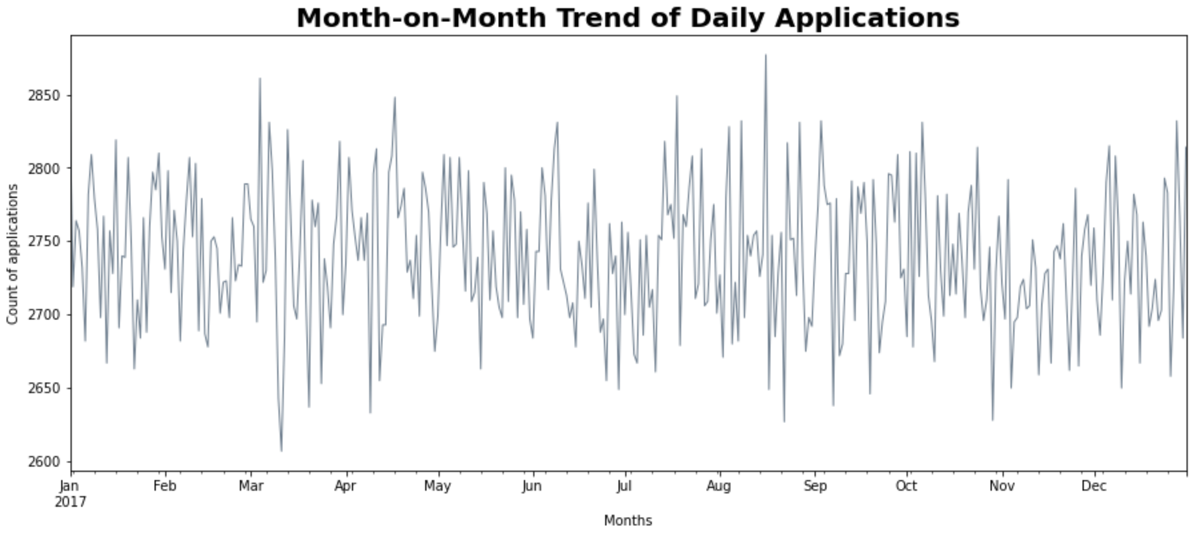
Field Name	% Populated	Min	Max	Std Dev	% Zero
Date	100 %	2017-01-01	2017-12-31	/	0 %
D.o.b	100 %	1900-01-01	2016-10-31	/	0 %

   b. Categorical Table

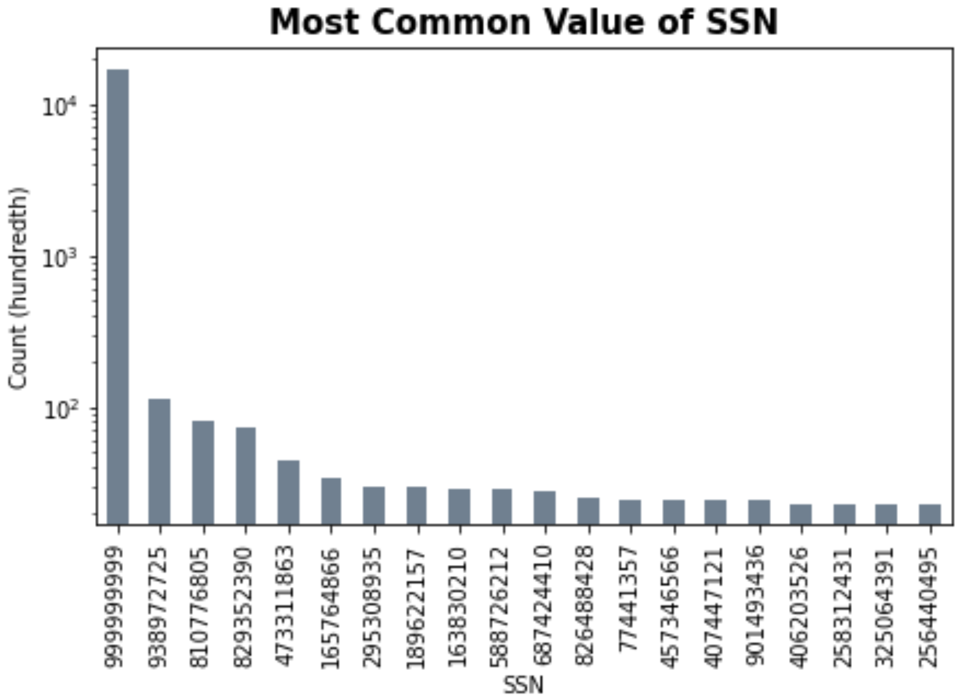
Field Name	% Populated	# Unique Values	Most Common Value
Record	100 %	1000000	/
SSN	100 %	835819	999999999
Firstname	100 %	78136	EAMSTRMT
Lastname	100 %	177001	ERJSAXA
Address	100 %	828774	123 MAIN ST
Zip5	100 %	26370	68138
Homephone	100 %	28244	9999999999
Fraud_label	100 %	2	0

3. **Data Visualizations**
   a. Field Name: Record
   Description: Contains record number of ordinal positive numbers that ranges from 1 to 1,000,000.

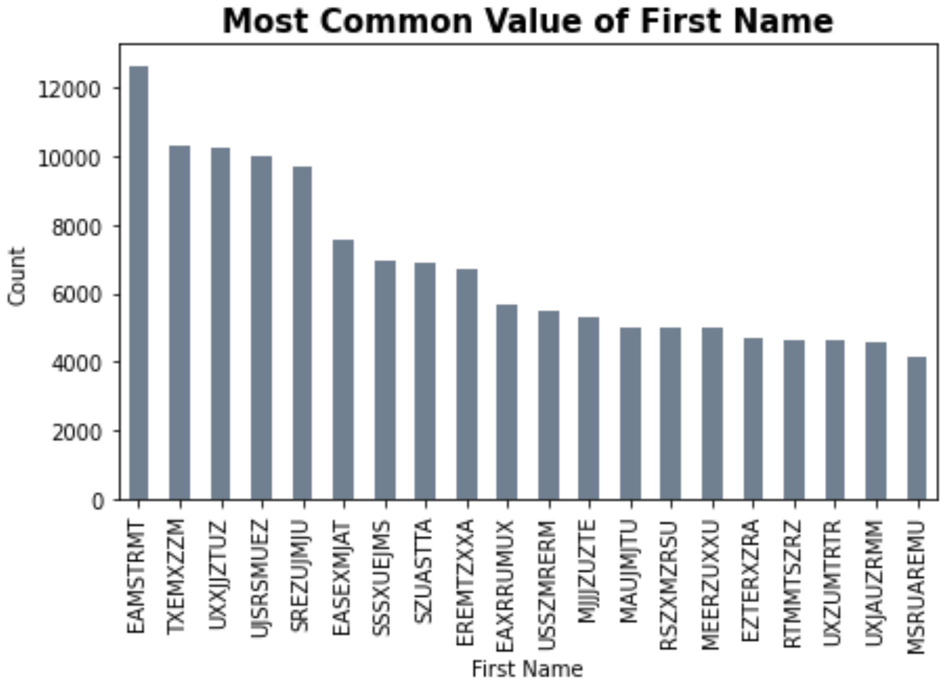
   b. Field Name: Date
   Description: The trend of daily applications received across months of 2017.



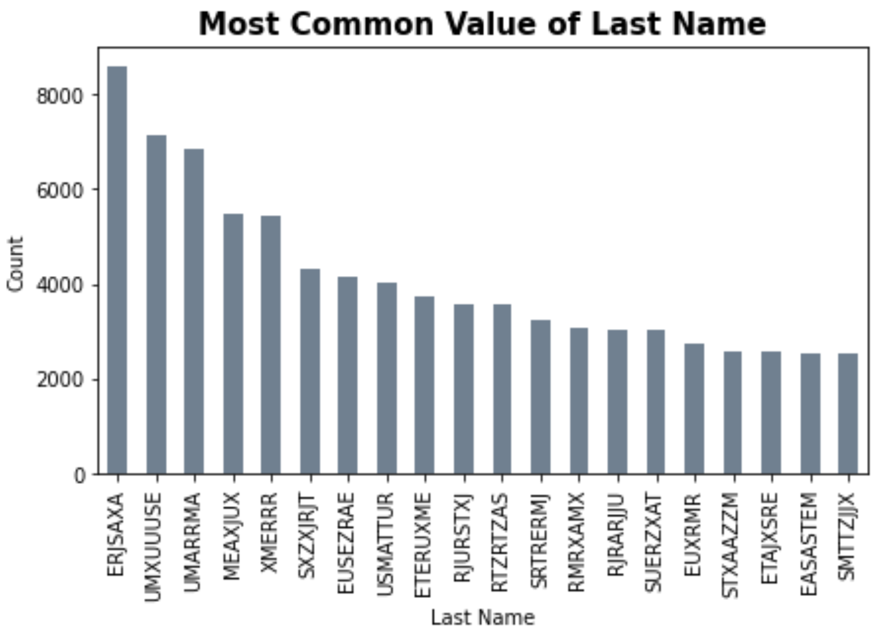
c. Field Name: SSN
   Description: The top 20 most common value of SSN appearing in the dataset, where there are 835,819 unique SSN value, and the SSN value of '999999999' is the most common value at 16,935 records.



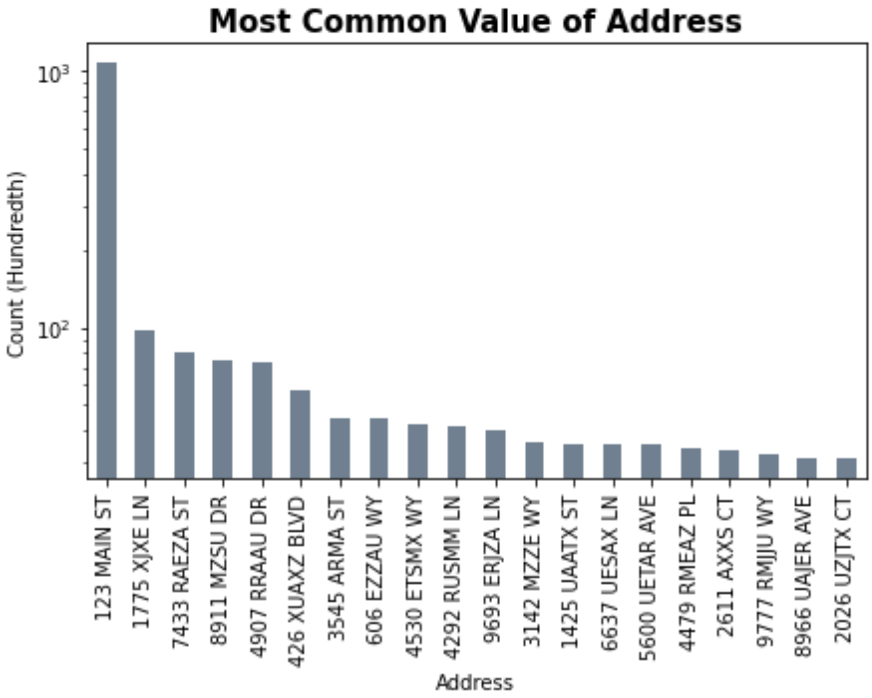
d. Field Name: Firstname
   Description: The top 20 most common first name appearing in the dataset, where there are 78,136 unique first names, and the first name 'EAMSTRMT' is the most common value at 12,658 records.



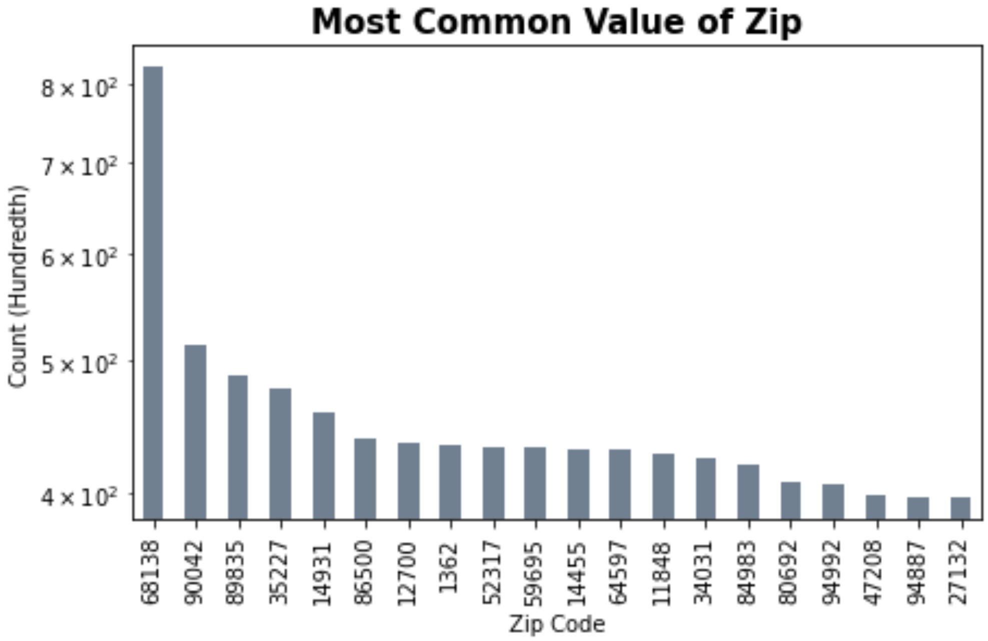
e. Field Name: Lastname
   Description: The top 20 most common last name appearing in the dataset, where there are 177,001 unique last names, and the last name ‘ERJSAXA’ is the most common value at 8,580 records.



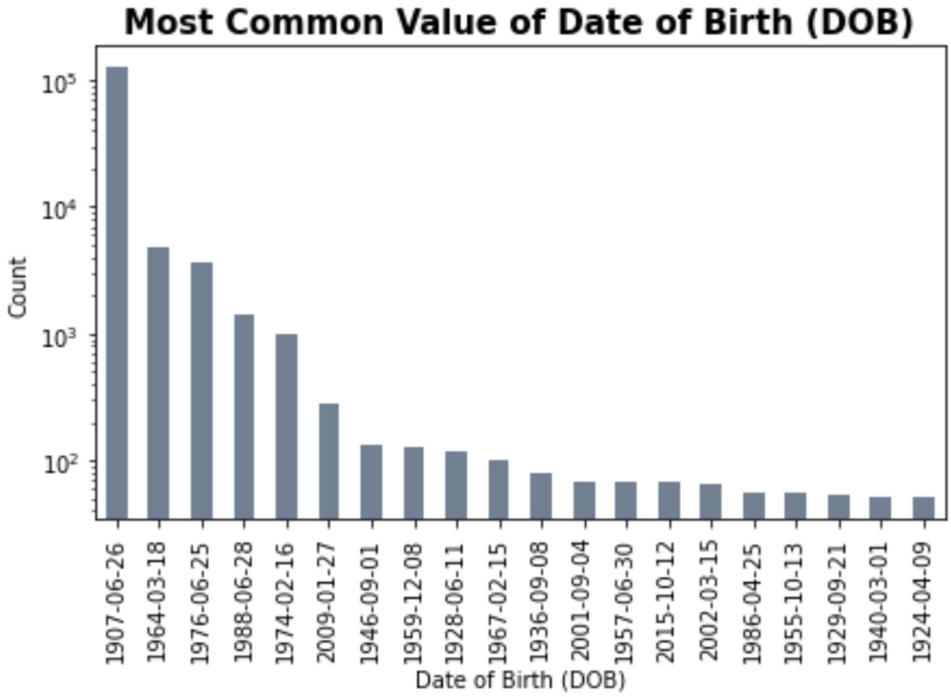
f. Field Name: Address
   Description: The top 20 most common addresses appearing in the dataset, where there are 828,774 unique addresses, and the address ‘123 MAIN ST’ is the most common value at 1,079 records.

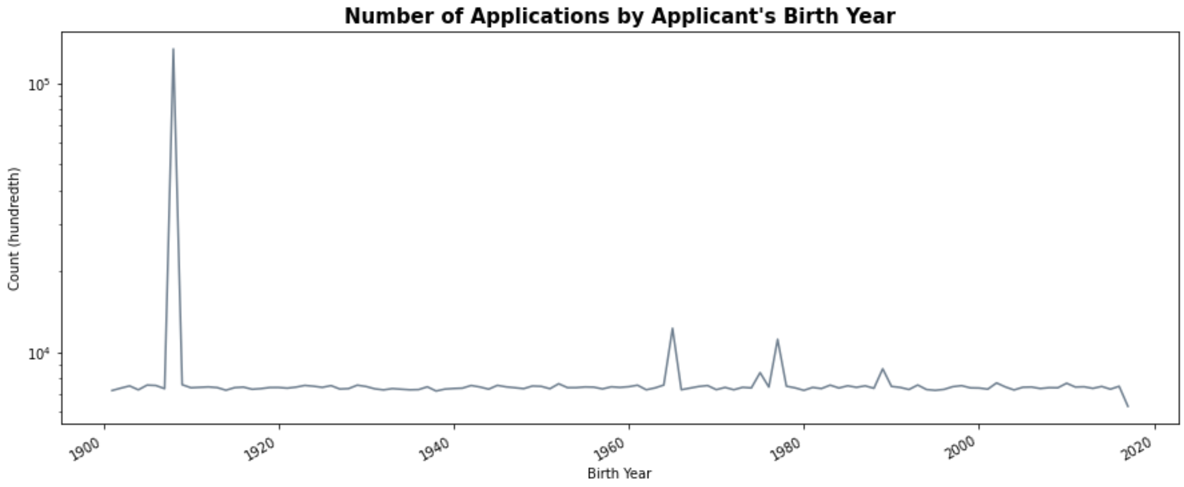


g. Field Name: Zip5
   Description: The top 20 most common values zip codes appearing in the dataset, where there are 26,370 unique zip codes, and the zip code ‘68138’ is the most common value at 823 records.

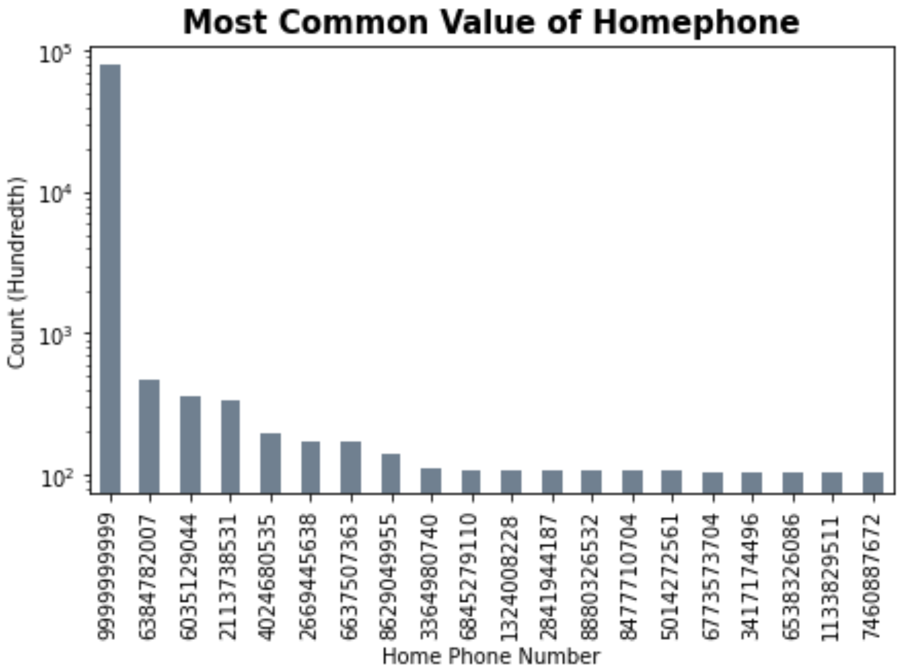


h. Field Name: DOB
   Description: The top 20 most common values date of birth (DOB) appearing in the dataset, where there are 42,673 unique DOB, and the DOB ‘1907-06-26’ is the most common value at 126,568 records.
   The following graph shows the year of birth for applicants ranges from 1900-01-01 to 2016-10-31, with most applicant’s birth year being 1907 at 133,986 count.

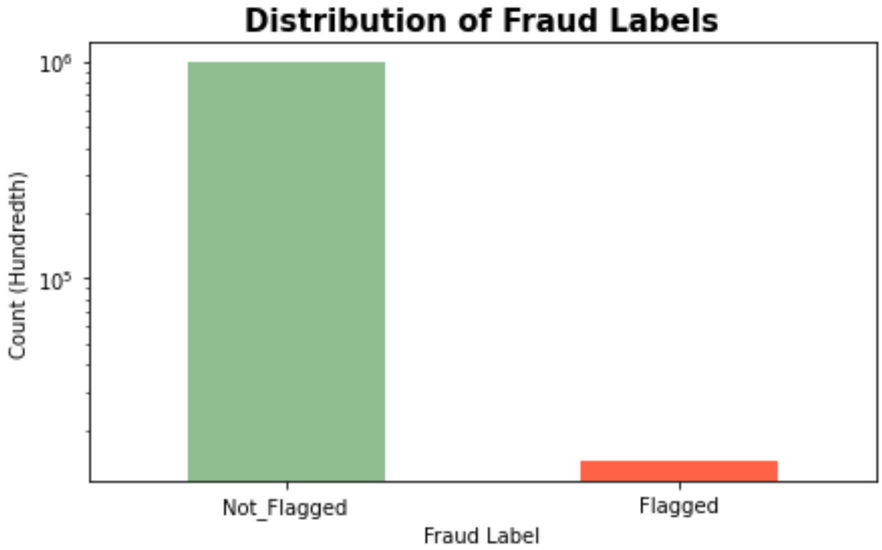


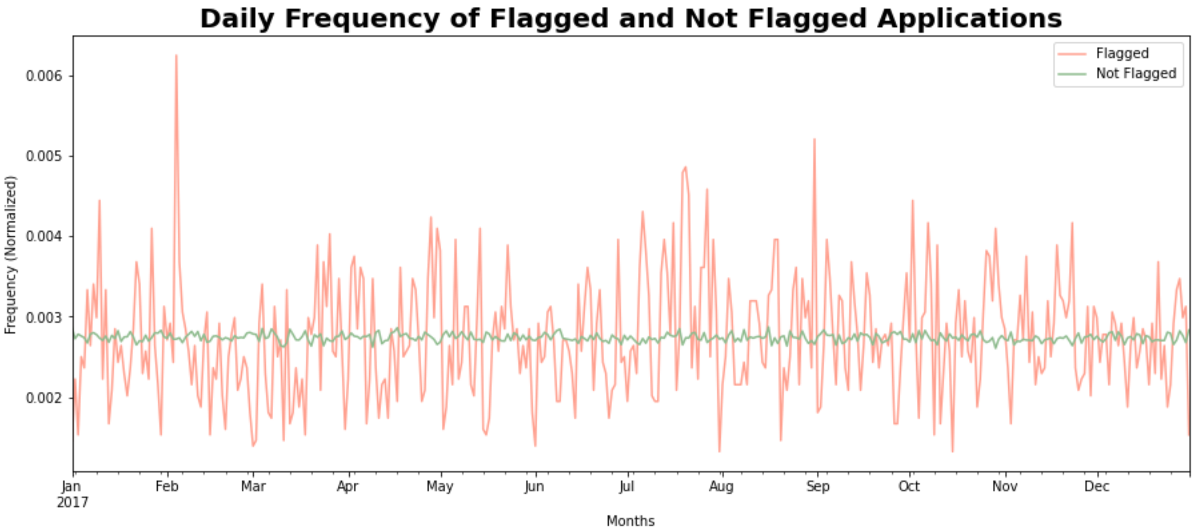


i. Field Name: Homephone
Description: The top 20 most common home phone numbers appearing in the dataset, where there are 28,244 unique home phone numbers, and the number ‘9999999999’ is the most common value at 78,512 records.



j. Field Name: Fraud_label
   Description: The fraud labels in the dataset are represented by a Boolean field, with 0 indicating that the application is not flagged for potential fraud and 1 indicating that it is flagged. Out of 1,000,000 data points, 98.5% (985,607) were not flagged and 1.4% (14,393) were flagged for potential fraud.
   The following graph shows the frequency of flagged and not flagged applications in year 2017. Applications that are flagged has higher amplitude and greater variability across days and months, where else applications that are not flagged has lesser variability and amplitude.


Distribution of Fraud Labels
10^6
10^5
Count (Hundredth)
Not_Flagged
Flagged
Fraud Label


Daily Frequency of Flagged and Not Flagged Applications
Flagged
Not Flagged
Frequency (Normalized)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months