Homework 2
Data Quality Report

Name: Aditya Shrikant Chavan
USC ID: 8741411805

1) Data Description

The data is about **product applications**. It comprises **applicant information** for each application. There are **10 attributes** and **one million entries** from the year **2017**.

2) Summary

1. Numerical Table

Field Name	% Populated	Min	Max	Mean	Median	Stddev	% Zero
Date	100	01-01-2017	12-31-2017	-	-	-	0.00

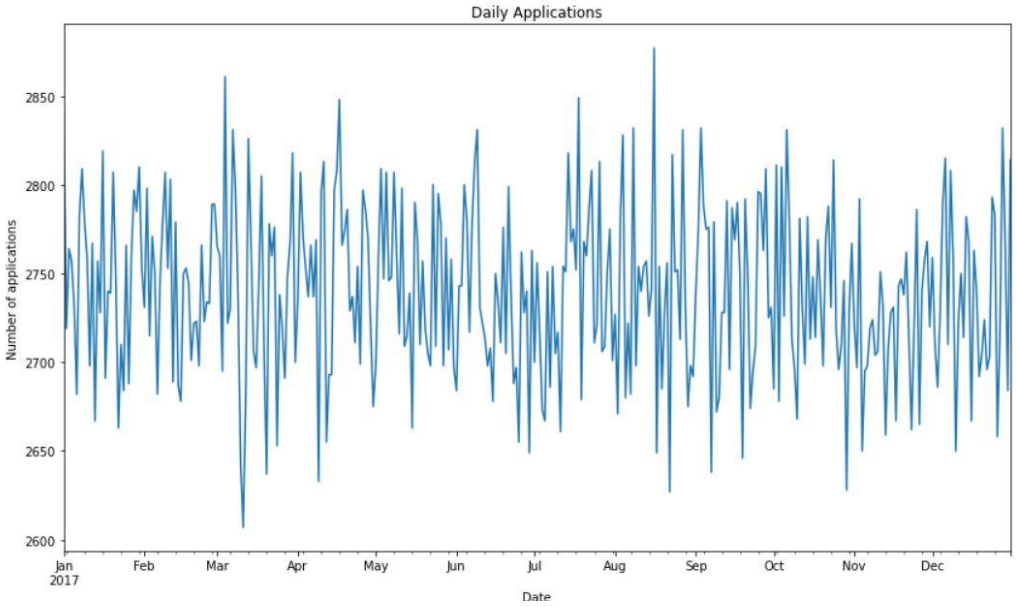
2. Categorical Table

Field Name	Description	%Populated	# Unique Values	Most Common Value
Record	Unique record Number	100.00	1,000,000	N/A
SSN	SSN of the applicant	100.00	835,819	999999999
First Name	First Name of the applicant	100.00	78,136	EAMSTRMT
Last Name	Last Name of the applicant	100.00	177,001	ERJSAXA
Address	Address of the applicant	100.00	828,774	123, MAIN ST
Zip	Zip code of the address	100.00	26,370	68138
DOB	Date of Birth of the applicant	100.00	42,673	6-26-1907
Homephone	Contact number of the applicant	100.00	28,244	9999999999
Fraud Label	Is the application a fraud	100.00	2	0

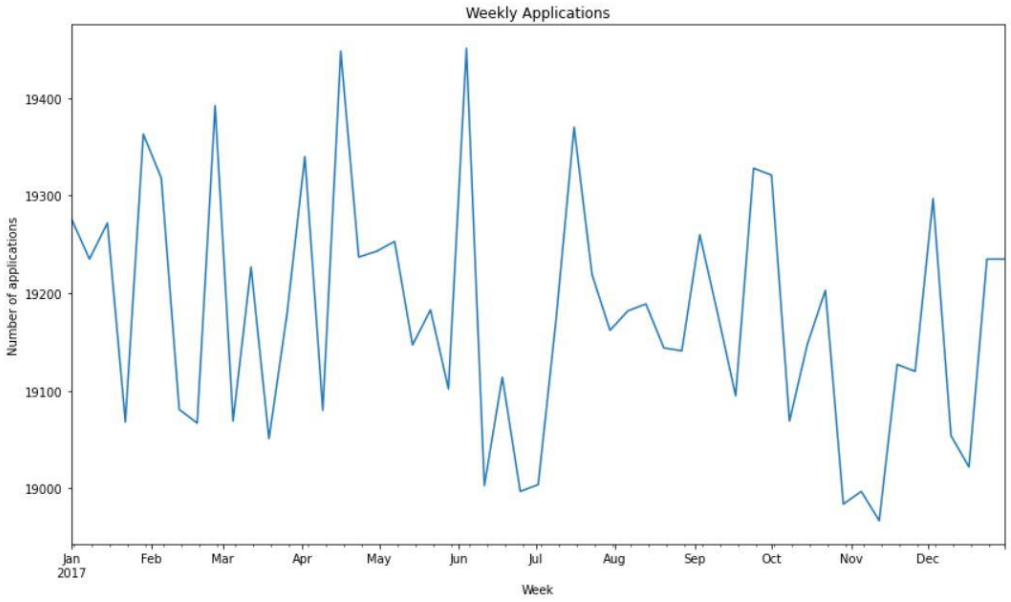
3) Visualization of Each field

1. Field Name: Record
   Field Description: This field is unique for every row. It can be used to identify an application. There are 1,000,000 unique values for the field.

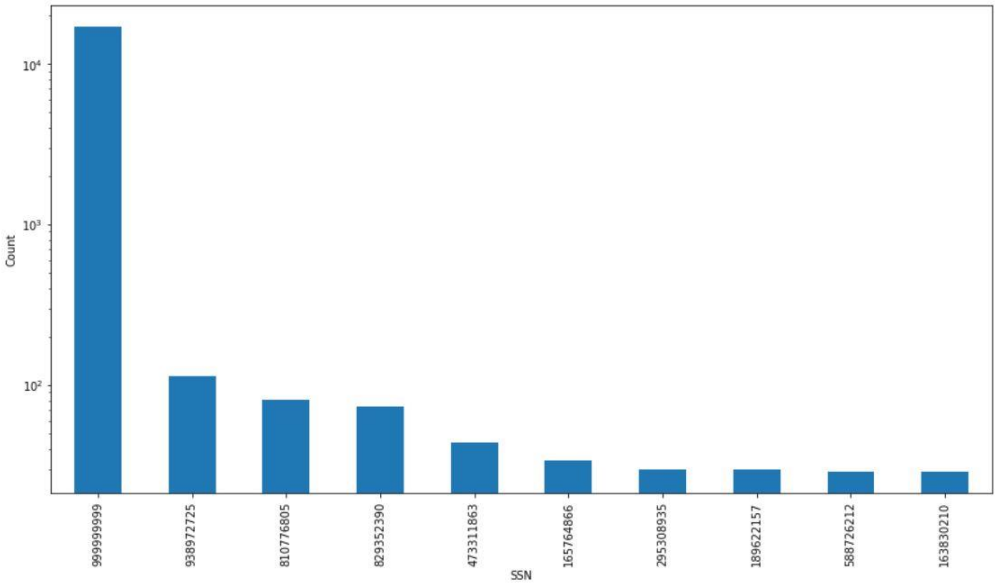
2. Field Name: Date
   Field Description: This field signifies the date when the applicant applied for the product. The graph shows number of applications done on each day.



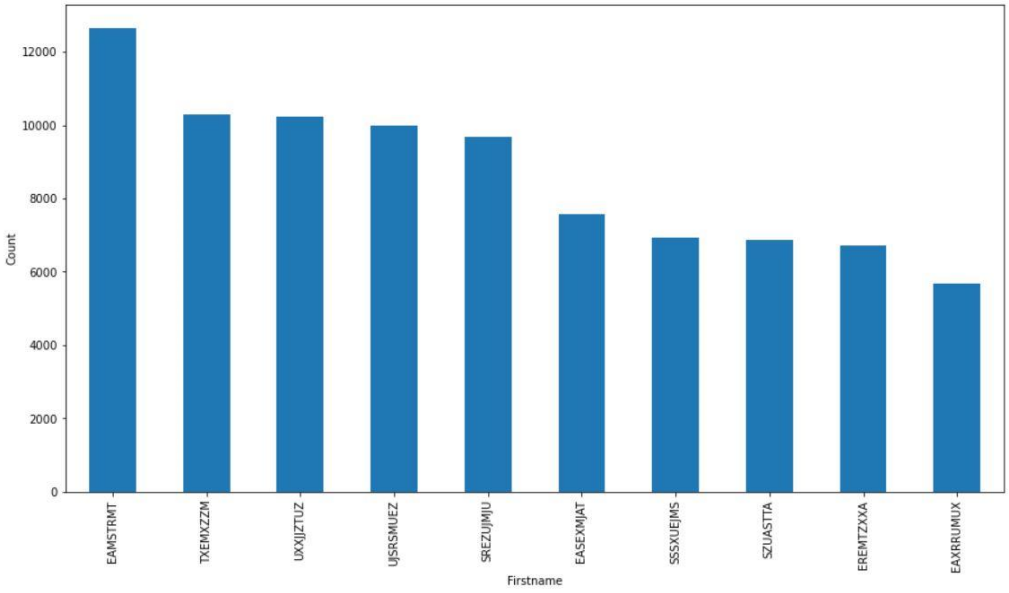
The graph above is too spiky and can be difficult to understand, hence a weekly distribution of applications is provided below.



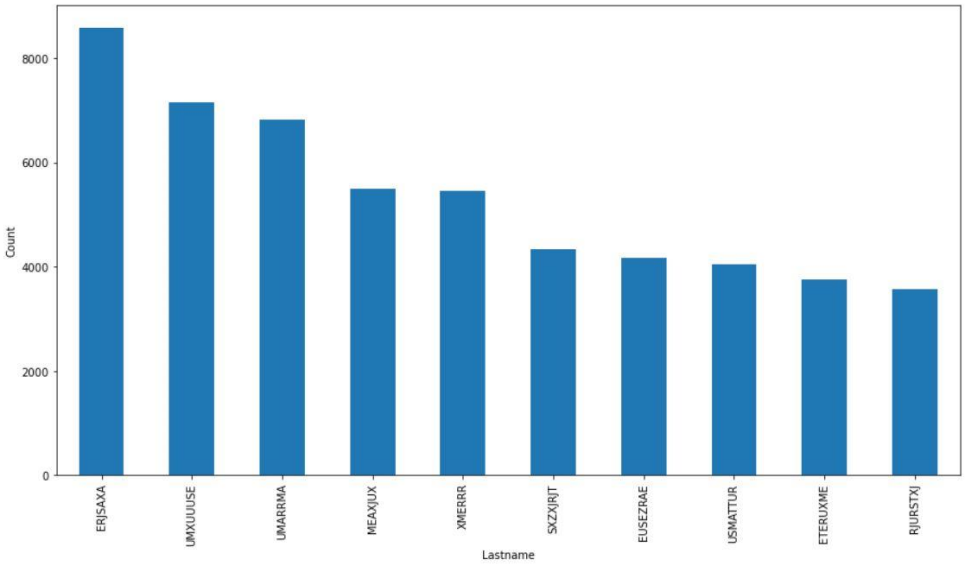
3. Field Name: SSN
   Field Description: SSN is the Social Security Number of the applicant. It is unique to every person. The graph below shows most used SSNs to apply for the product and number of times they were used.



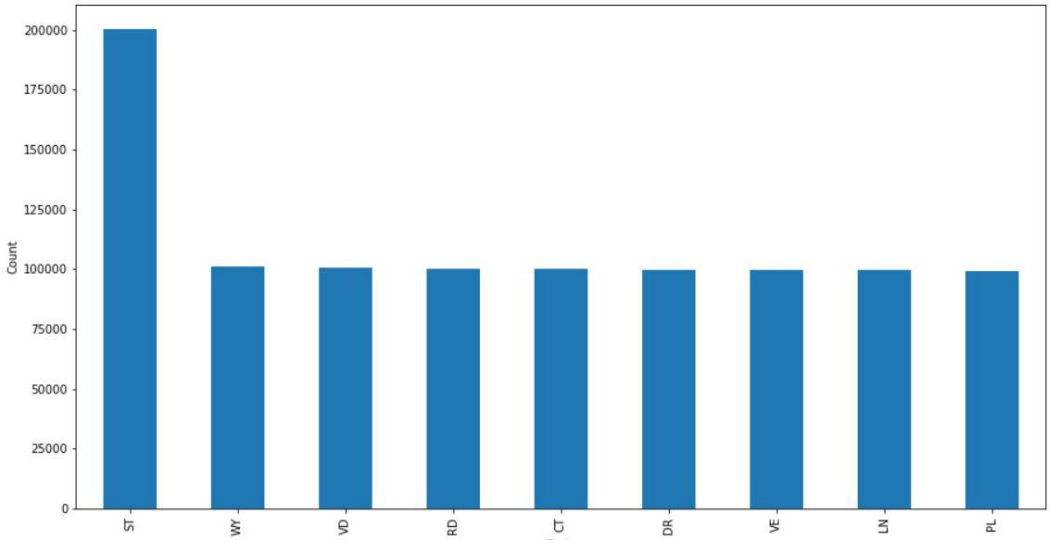
4. Field Name: Firstname
   Field Description: This field is the first name of the applicant. Most used first name is Eamstrmt. It is used over 1200 times.



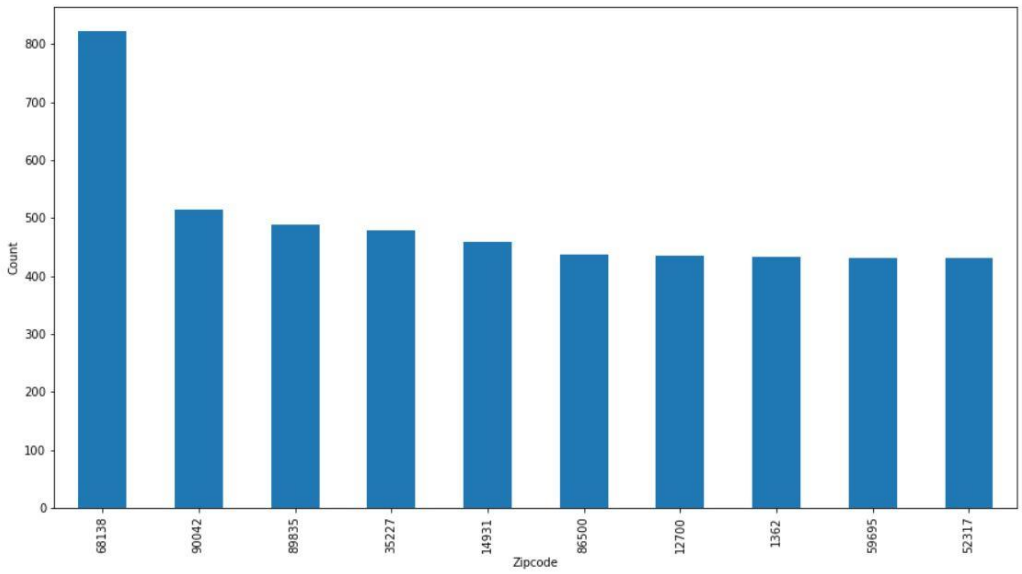
5. Field Name: Lastname
   Field Description: This field is the last name of the applicant. Most used first name is Erjsaxa. It is used over 8000 times.



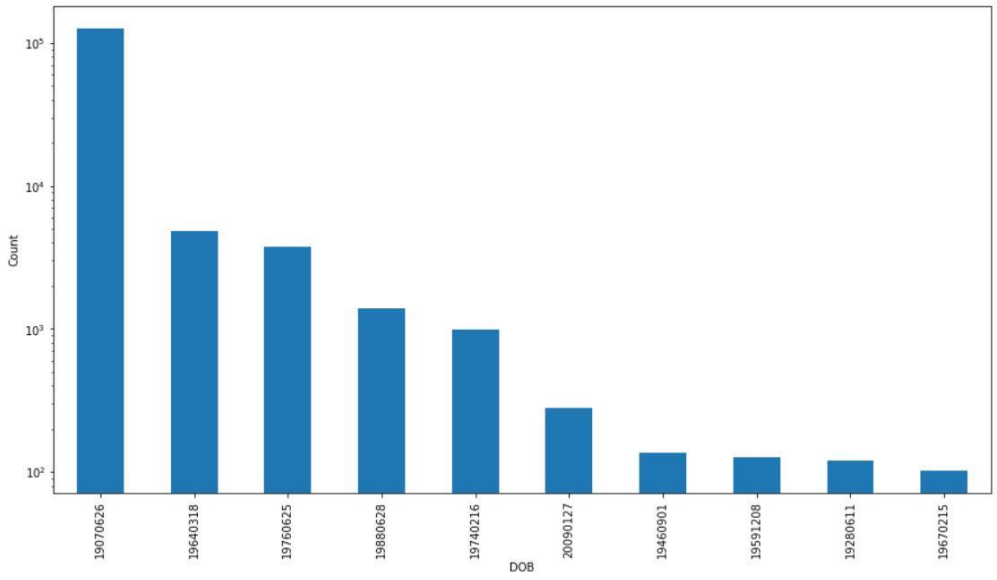
6. Field Name: Address
   Field Description: Address of the applicant. The graph below shows the distribution of application over different states.



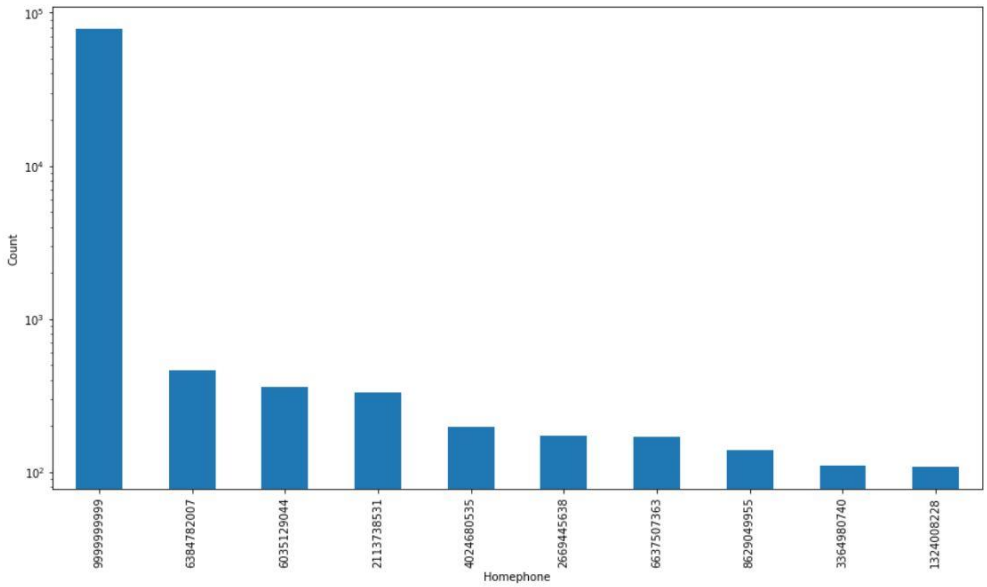
7. Field Name: Zipcode
   Field Description: The field is the zipcode used by the applicant to apply for the product. This signifies the locality where they stay.



8. Field Name: DOB
   Field Description: Date of birth of the applicant. In the data 6/26/1907 is the most seen date of birth.



9. Field Name: Homephone
   Field Description: This is the phone number of the applicant. The most common value is 9999999999.



10. Field Name: Fraud Label

    Field Description: This tells us if the application was a fraud or not. In the data there are significantly less frauds than genuine applications.