Data Quality Report

Title: Credit Card Application Data

Description: This data is a representation of 94866 credit card applications made during 2016. A typical observation contains information about the personal details of each applicant such as name, social security number, phone number, date of application, date of birth and address. A binary fraud marker is used to indicate whether the application was an act of fraud or not.

Number of Records: 94866

Number of Variables: 10

Part I: Summary statistics

Variable	Description	Percent Populated	Number of Unique Values
record	Identifier	100%	94866
date	Date	100%	365
ssn	Social Security Number	100%	86771
firstname	First Name	100%	14626
lastname	Last Name	100%	31513
address	Address	100%	88167
zip5	5-digit ZIP	100%	15855
dob	Date of birth	100%	30599
homephone	Home Phone Number	100%	20762
fraud	Fraud or Not	100%	2

Part II: Description and Visualization

1. **RECORD** is a categorical variable. It works as the ordinal reference number for each property record. There are 94866 records overall. Each row is a unique number/identifier and hence, a visualization is not required.

2. **DATE** refers to the date on which a given applicant made the credit card application. From Figure 1, the number of applications appears to increase linearly over time. Figure 4 shows that the applications slowed down during the first four months of 2016.

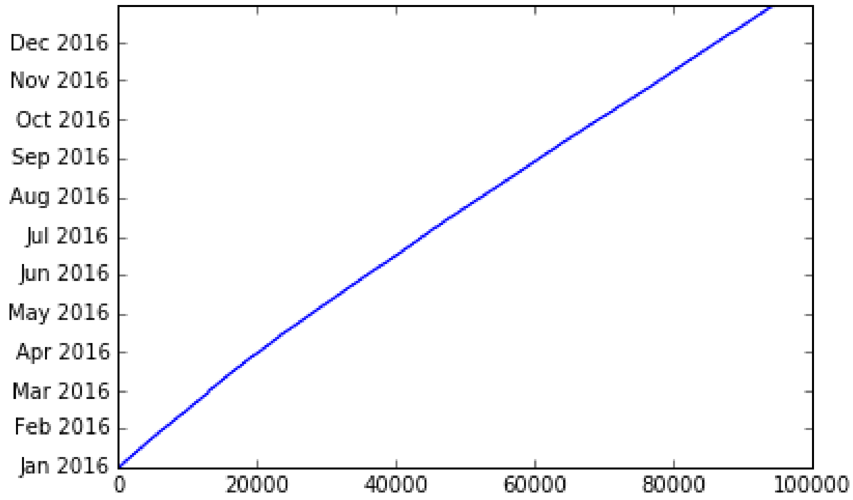

Figure 1: Cumulative Number of Applications Over the Year

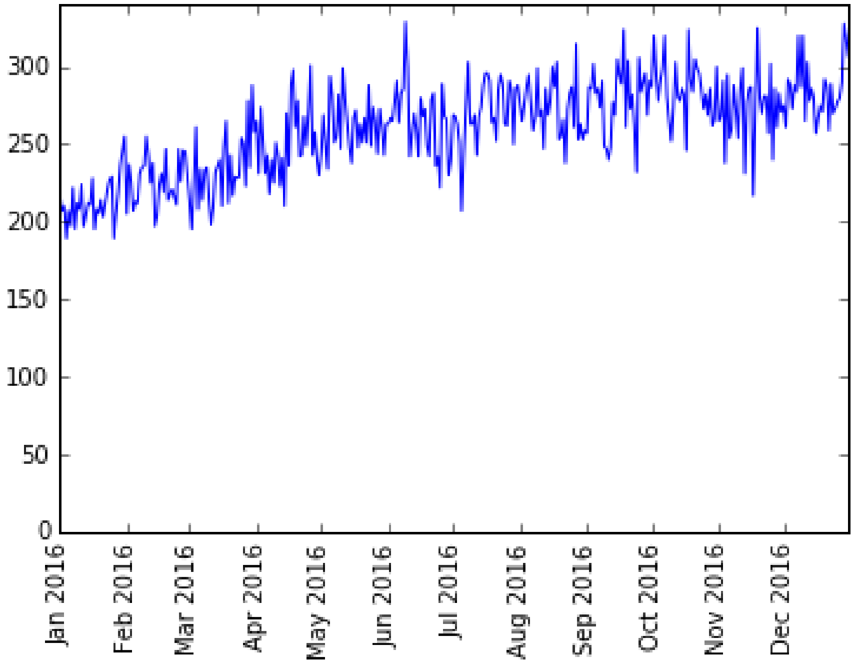

Figure 2: Daily frequency of applications over the year

There was a dip in frequency of applications at the end of 2016 (as can be seen in Figure 3 and 4), which can be attributed to the sudden change in week/month at the end of 2016 (given that 2017 data is not available).

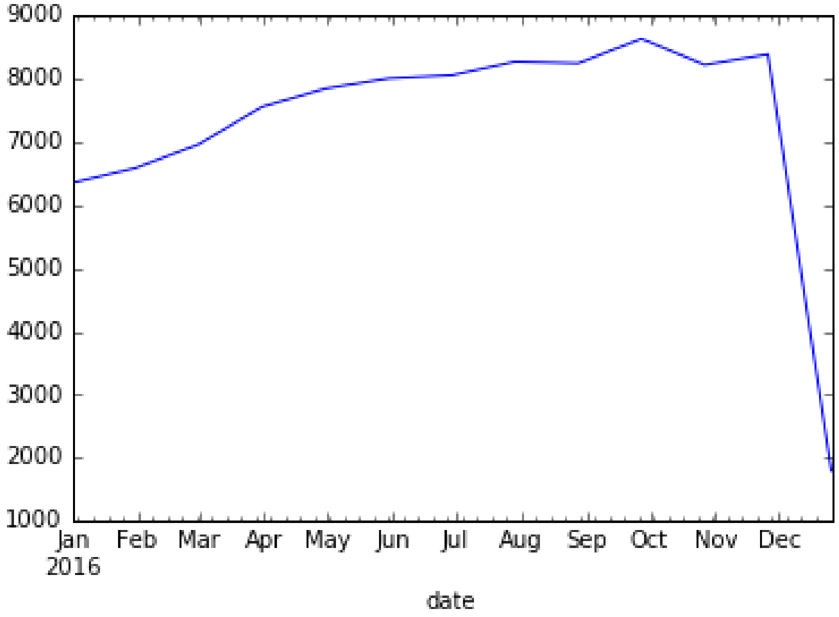

Figure 3: Monthly frequency of applications over the year

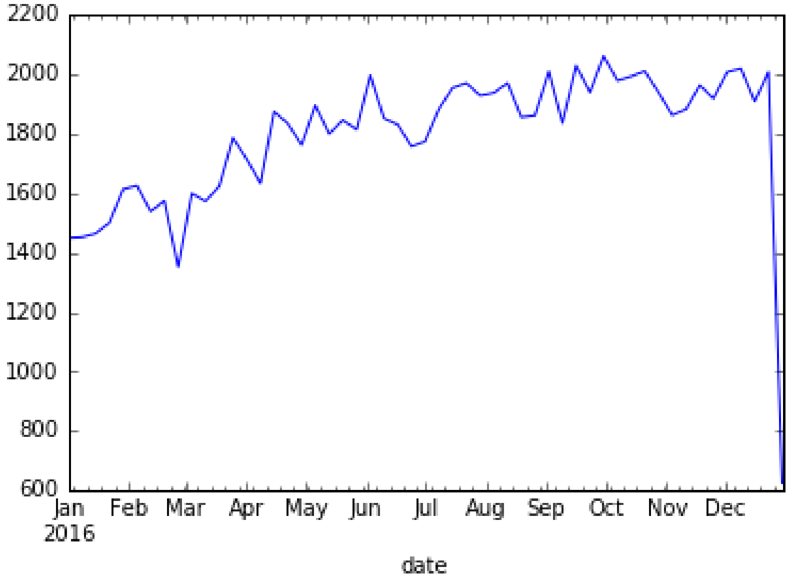

Figure 4: Weekly frequency of applications over the year

3. **SSN** contains information about each applicant's Social Security Number. As can be seen in Figure 3, there are few SSNs with higher frequencies than usual; the tallest bar corresponds to 737610282, whose frequency was 1478.
   For consistency, leading zeros were added such that the phone numbers always showed 9 digits.

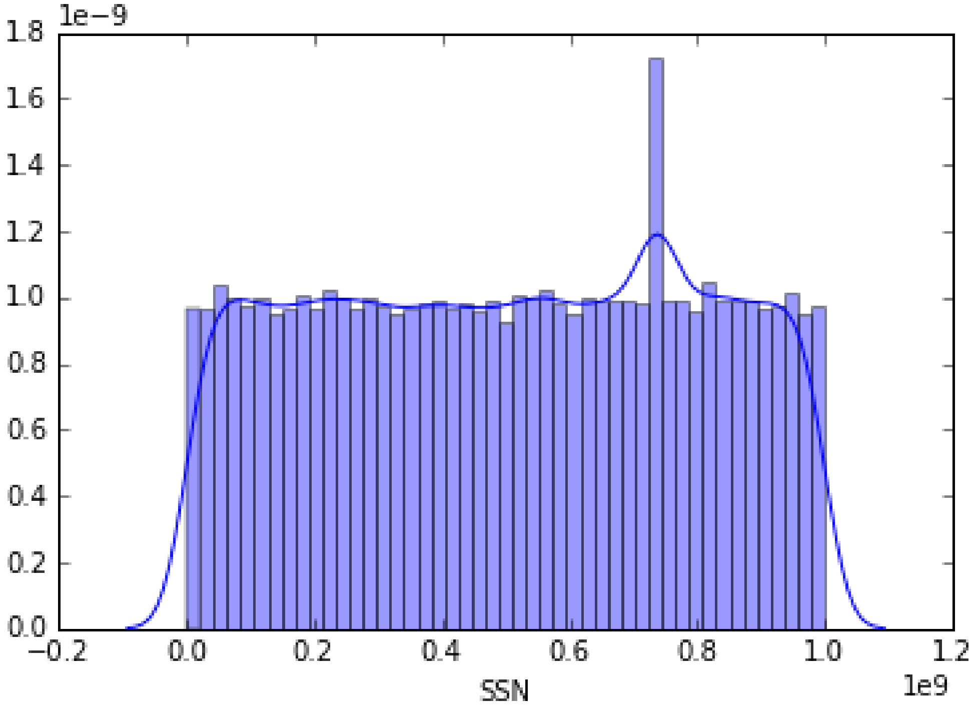

Figure 5: Distribution Plot of SSN of Applicants

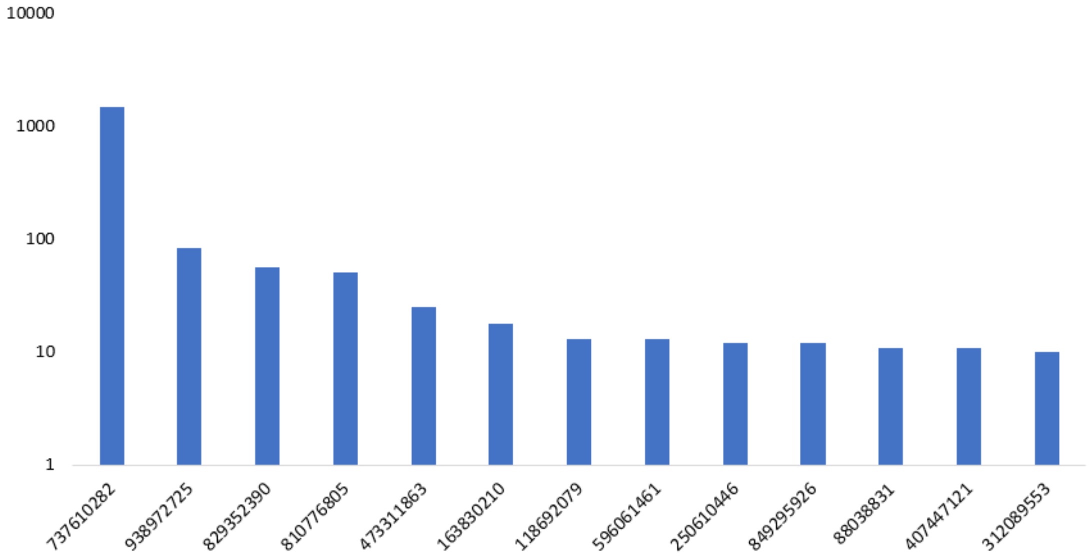

Figure 6: Log scaled frequency plot of SSNs with at least 10 applications

4. **FIRSTNAME** denotes the first name of applicants as indicated in the application.

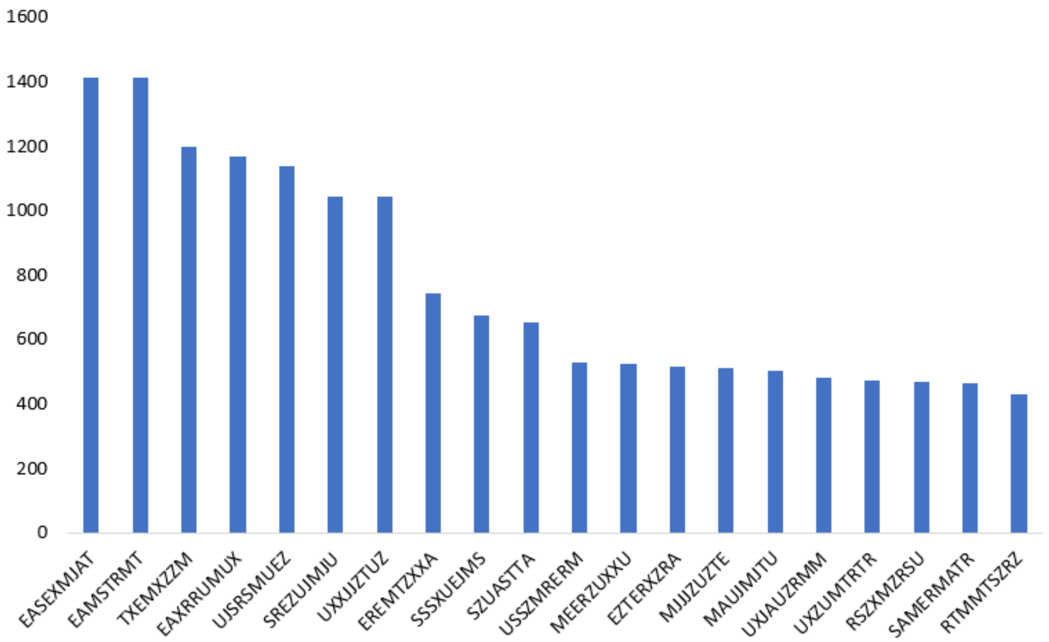

Figure 7: Frequency plot of 20 most recurring first names

5. **LASTNAME** denotes the last name of applicants as indicated in the application.

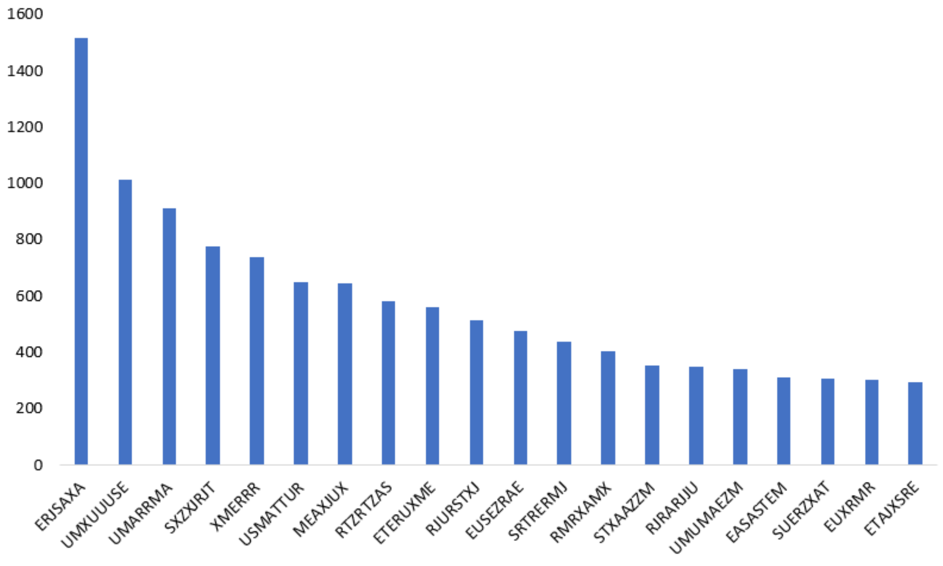

Figure 8: Frequency plot of 20 most recurring last names

6. **ADDRESS** denotes address of applicants as indicated in the application.

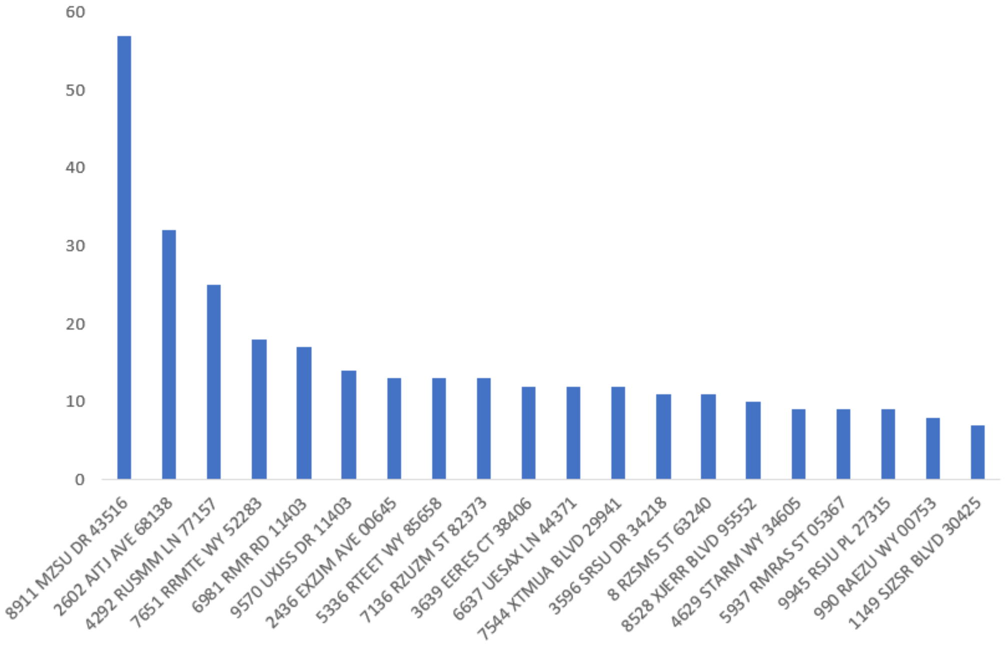

Figure 9: Frequency plot of 20 most recurring addresses

7. **ZIP5** denotes the 5-digit ZIP code of the locality of the applicant as indicated in the application. As can be seen from Figure 8, applicants from certain localities make more applications than others. Several reasons such as affluence of the locality, activity of fraud rings, etc., can be attributed to this.
   For consistency, leading zeros were added such that the ZIP codes always showed 5 digits.

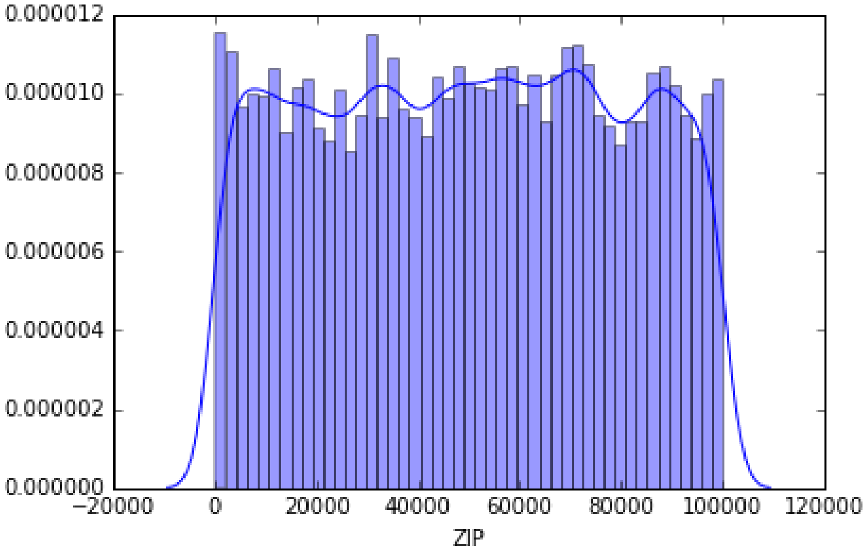

Figure 10: Distribution plot of ZIP codes

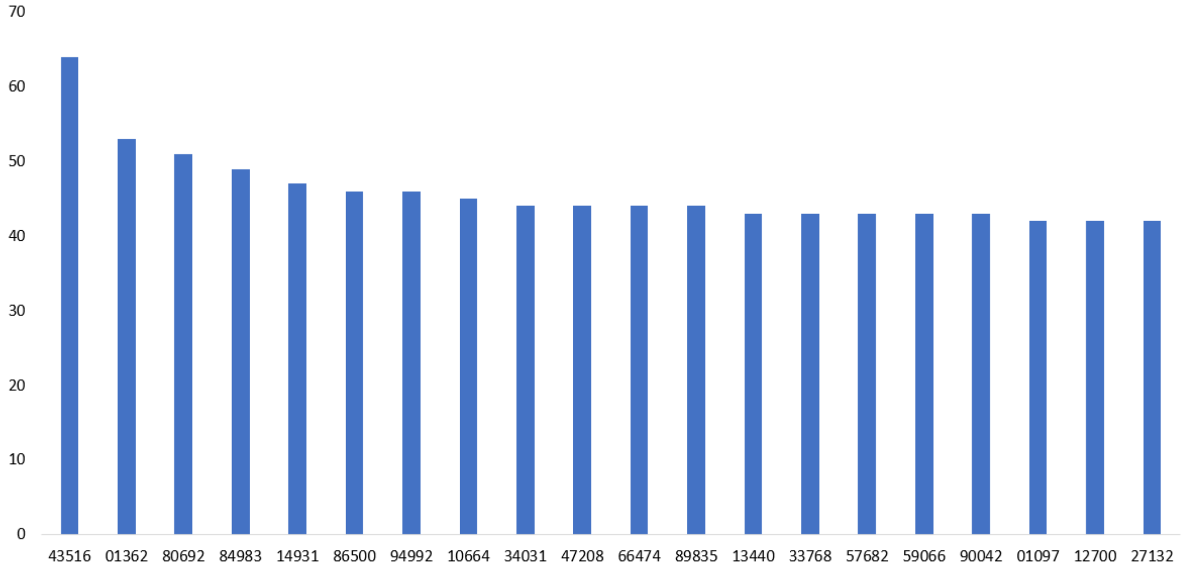

Figure 11: Frequency plot of 20 most recurring ZIP codes

8. **HOMEPHONE** denotes the home phone number of the applicants as indicated in the application. An overall view of the data as shown in Figure 9 shows that there's one home phone number with very high frequency relative to others. Further analysis revealed that this number is (910) 558-0920.
   For consistency, leading zeros were added such that the phone numbers always showed 10 digits.

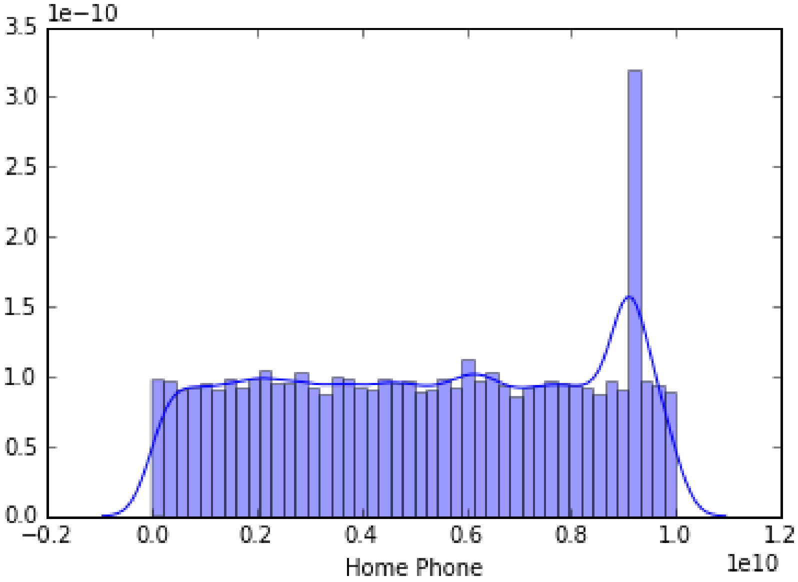

Figure 12: Distribution plot of home phone numbers

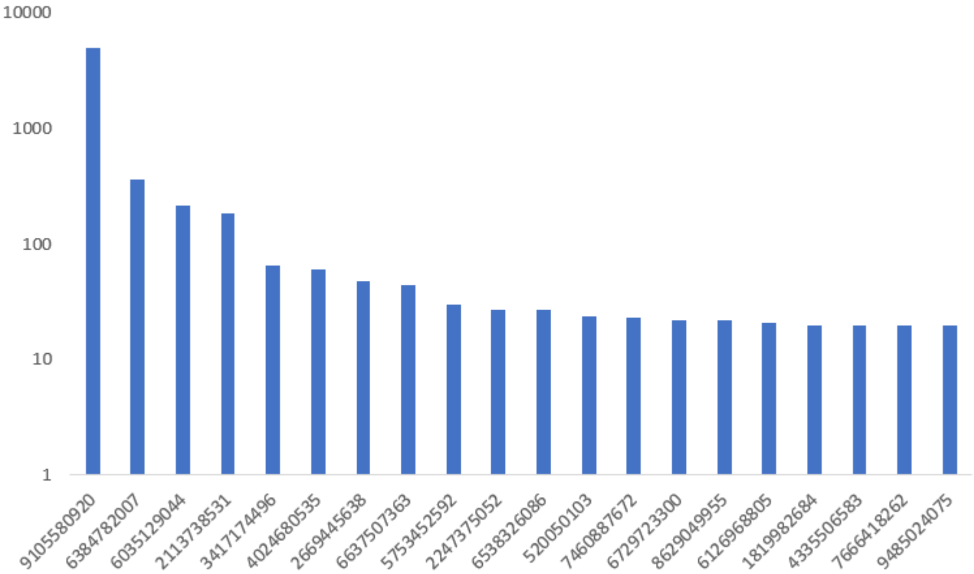

Figure 13: Log scaled frequency plot of 20 most recurring home phone numbers

9. **DOB** denotes the date of birth of applicants as indicated in the applications. Since, the minimum age for credit card holders is 18, it is fair to assume that all applicants (for the application year, 2016) were born in the 20th century.

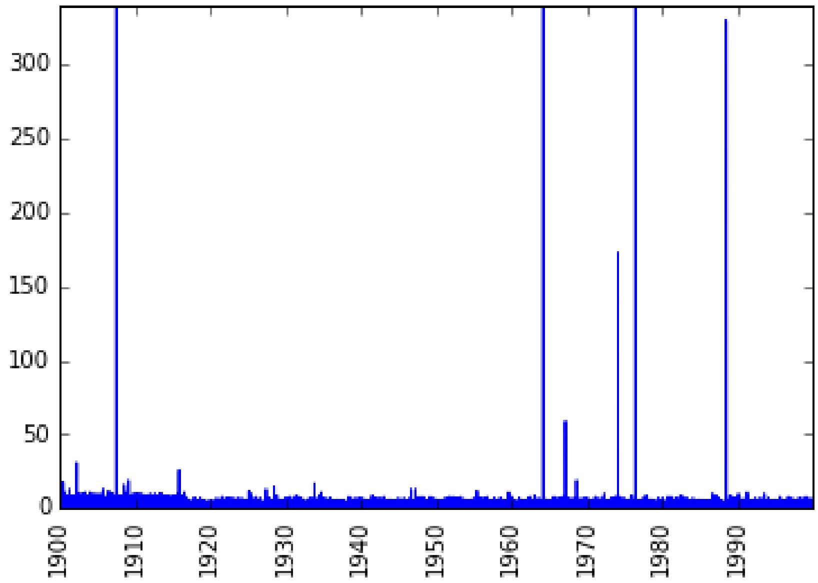

Figure 14: Distribution of dates of births of applicants

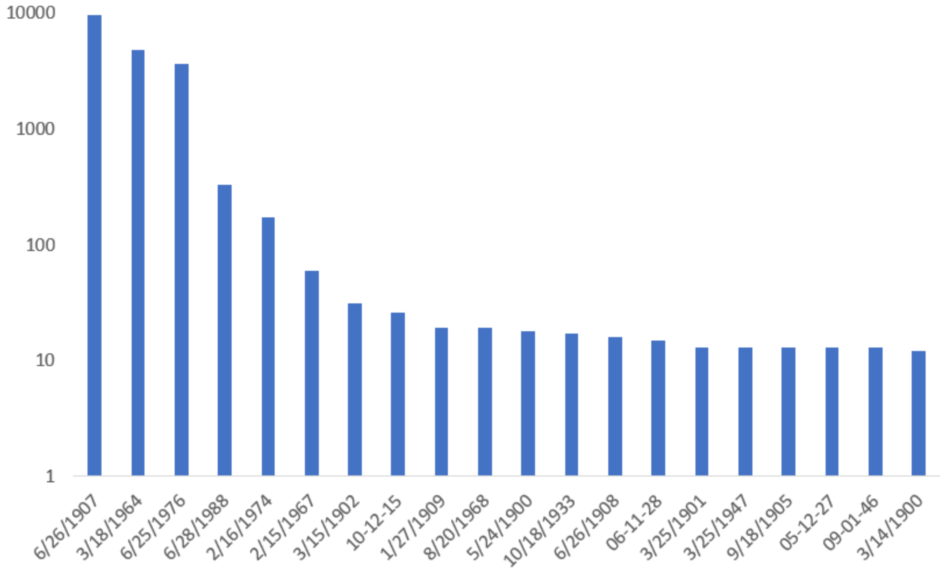

Figure 15: Log scaled frequency plot of 20 most recurring dates of births

10. FRAUD indicates whether a given application was determined to be a fraudulent application. A value of ‘1’ indicates fraud, whereas ‘0’ denotes a regular application. As can be seen from the pie chart below, 2 out of 10 applications in the given dataset have been marked as fraudulent.

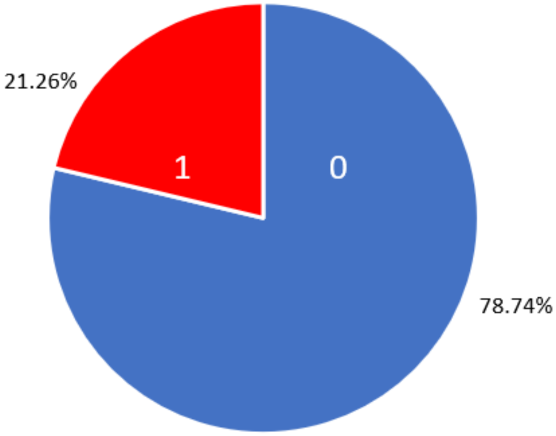

Figure 16: Pie chart showing the share of fraudulent applications in the dataset

Part III: Conclusions

This dataset is to be used to build a fraud detection model for a future project. Having tabulated and visualized all the variables, it is evident that there’s a small fraction of anomalies in the data. Going forward, our aim is to build a supervised learning model to detect fraud with reasonable accuracy.