Leith McIndewar
Machine Learning
Assignment 2
4/12/16

Columns and datatypes

SeriousDlqin2yrs	int64
RevolvingUtilizationOfUnsecuredLines	float64
age	int64
NumberOfTime30-59DaysPastDueNotWorse	int64
DebtRatio	float64
MonthlyIncome	float64
NumberOfOpenCreditLinesAndLoans	int64
NumberOfTimes90DaysLate	int64
NumberRealEstateLoansOrLines	int64
NumberOfTime60-89DaysPastDueNotWorse	int64
NumberOfDependents	float64

Descriptive Statistics

	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome
Count	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000	1.202690e+05
Mean	0.066840	6.048438	52.295207	0.421033	353.005076	6.670221e+03
Std	0.249746	249.755371	14.771866	4.192781	2037.818523	1.438467e+04
Min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00
25%	0.000000	0.029867	41.000000	0.000000	0.175074	3.400000e+03
Median	0.000000	0.154181	52.000000	0.000000	0.366508	5.400000e+03
75%	0.000000	0.559046	63.000000	0.000000	0.868254	8.249000e+03
Max	1.000000	50708.000000	109.000000	98.000000	329664.000000	3.008750e+06

	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
Count	150000.000000	150000.000000	150000.000000	150000.000000	146076.000000
Mean	8.452760	0.265973	1.018240	0.240387	0.757222
Std	5.145951	4.169304	1.129771	4.155179	1.115086
Min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	5.000000	0.000000	0.000000	0.000000	0.000000
Median	8.000000	0.000000	1.000000	0.000000	0.000000
75%	11.000000	0.000000	2.000000	0.000000	1.000000
Max	58.000000	98.000000	54.000000	98.000000	20.000000

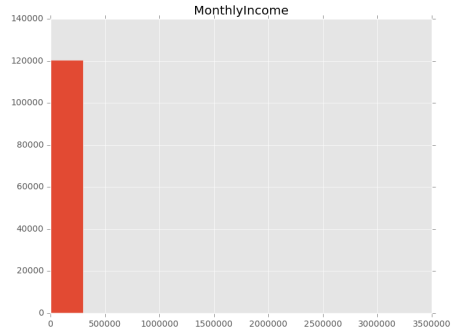
Modes

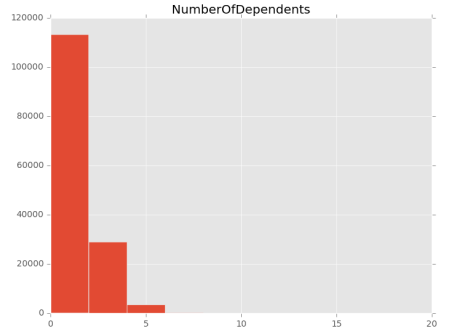
SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents	
0	0	0.0	49	0	0.0	5000.0	6	0	0	0	0.0

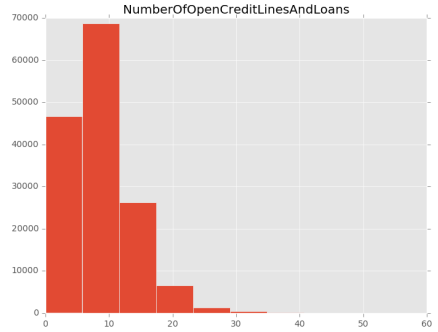
Missing Values

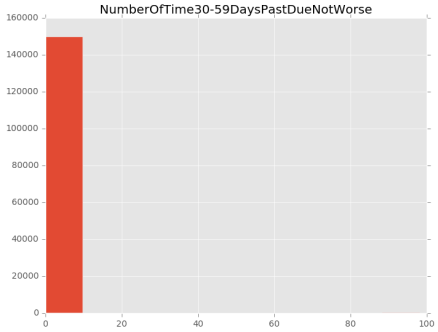
SeriousDlqin2yrs	0
RevolvingUtilizationOfUnsecuredLines	0
age	0
NumberOfTime30-59DaysPastDueNotWorse	0
DebtRatio	0
MonthlyIncome	29731
NumberOfOpenCreditLinesAndLoans	0
NumberOfTimes90DaysLate	0
NumberRealEstateLoansOrLines	0
NumberOfTime60-89DaysPastDueNotWorse	0
NumberOfDependents	3924

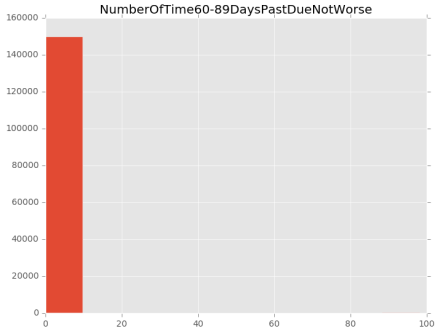
Data Histograms
Many of the histograms are not good illustrations of the data because of outliers that skew the scale of the x-axis. I would continue to refine these to get better representations.

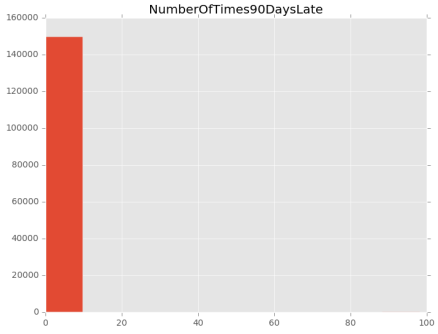


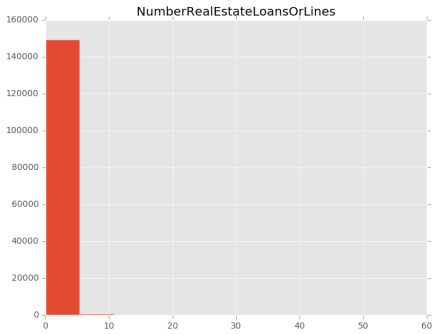



NumberOfOpenCreditLinesAndLoans
70000
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
60


NumberOfTime30-59DaysPastDueNotWorse
160000
140000
120000
100000
80000
60000
40000
20000
0
0
20
40
60
80
100


NumberOfTime60-89DaysPastDueNotWorse
160000
140000
120000
100000
80000
60000
40000
20000
0
0
20
40
60
80
100


NumberOfTimes90DaysLate
160000
140000
120000
100000
80000
60000
40000
20000
0
0
20
40
60
80
100



Cleaning data
To prepare the data for generating the model, I filled in missing information for MonthlyIncome and NumberOfDependents. I used mean imputation for missing MonthlyIncome data. To fill in the missing number of dependents, I made the (big) assumption that those that did not fill in number of dependents were more likely to not have dependents. Though it was a big assumption, it seems reasonable that those who did not fill in dependents would be more likely not to have any. I would try other methods of imputation for both, especially for dependents if I were to continue using this dataset.
I chose to categorize and create dummy variables DebtRatio and Age. For debt ration, there were several large outliers that may have affected the model if it were left as is. For example, 75% of the 150,000 samples had a debt ratio at or below 0.86, yet there was at least one at 329,664. I created dummy variables for a ranges of debt ratio divided by 0.25 up to 1. I dropped the original debt ratio variable as well. I ran a similar process for age. If I were to continue refining this model, it would be worth checking if it was a the right decision to drop the original variables and if the bins sizes added to the model.

Model 1
This model uses the raw values for age and number od dependents.

	Columns	Model Coefficients
0	RevolvingUtilizationOfUnsecuredLines	-0.000050
1	age	-0.042671
2	NumberOfTime30-59DaysPastDueNotWorse	0.503462
3	DebtRatio	-0.000017
4	MonthlyIncome	-0.000049
5	NumberOfOpenCreditLinesAndLoans	-0.018155
6	NumberOfTimes90DaysLate	0.356532
7	NumberRealEstateLoansOrLines	0.084392
8	NumberOfTime60-89DaysPastDueNotWorse	-0.831169
9	NumberOfDependents	0.098085

Model 2
This model includes the binned dummy variables for age and number of dependents.

	Columns	Model Coefficients
0	RevolvingUtilizationOfUnsecuredLines	-0.000043
1	NumberOfTime30-59DaysPastDueNotWorse	0.481873
2	MonthlyIncome	-0.000030
3	NumberOfOpenCreditLinesAndLoans	-0.012714
4	NumberOfTimes90DaysLate	0.514969
5	NumberRealEstateLoansOrLines	0.054729
6	NumberOfTime60-89DaysPastDueNotWorse	-0.966976
7	NumberOfDependents	0.072990
8	2-.4	-0.296910
9	.4-.6	-0.051176
10	.6-.8	0.264402
11	.8-1	0.213745
12	1-10	0.237350
13	10+	-0.117229
14	20-30	0.333624
15	30-40	0.215526
16	40-50	-0.063297
17	50-60	-0.352283
18	60-70	-0.881538
19	70-80	-1.015456
20	80+	-0.636311