Credit Risk Analysis

Business Understanding: CredX needs an Acquisition Analytical Model. Need to:

a. Determine the factors affecting credit risk,
b. Create strategies to mitigate the acquisition risk,
c. Assess the financial benefit of model.

Data Understanding and Data Preparation:

a. Collect data: There are two csv data sources – Demographic and Credit Bureau Data. Both have Applicant ID as primary key. Master file can be made by joining these data sets on Applicant ID
b. Describe datasets: The data dictionay is provided with the sources. Following are data fields

Demographic	Credit Bureau
Application ID	Application ID
Age	No of times 90 DPD or worse in last 6 months
Gender	No of times 60 DPD or worse in last 6 months
Marital Status (at the time of application)	No of times 30 DPD or worse in last 6 months
No of dependents	No of times 90 DPD or worse in last 12 months
Income	No of times 60 DPD or worse in last 12 months
Education	No of times 30 DPD or worse in last 12 months
Profession	Avgas CC Utilization in last 12 months
Type of residence	No of trades opened in last 6 months
No of months in current residence	No of trades opened in last 12 months
No of months in current company	No of PL trades opened in last 6 months
Performance Tag	No of PL trades opened in last 12 months
	No of Inquiries in last 6 months (excluding home & auto loans)
	No of Inquiries in last 12 months (excluding home & auto loans)
	Presence of open home loan
	Outstanding Balance
	Total No of Trades
	Presence of open auto loan
	Performance Tag

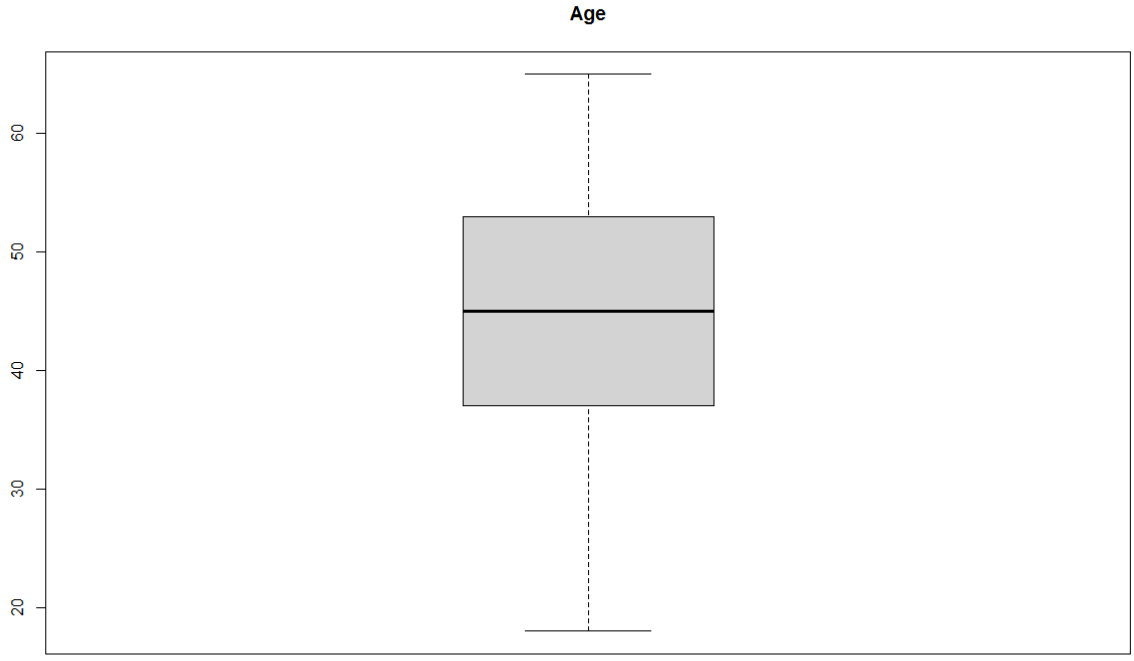
c. Check Data Quality:
    a. All Application.ID in demographic file are present in credit_bureau
    b. Both the data sets had 3 duplicate rows which needs to be deleted they they are not producing a unique row due to non-availability of timestamp. Hence, completely removed the Applicant IDs having duplicate rows as they will provide false observation. (less than 10% of data)
    c. Both files have a Synonym field “Performance Tag” hence, keeping only one field and removing the redundant column.
    d. The data sets have many NA values. Blanks too are coreced as NAs while importing.

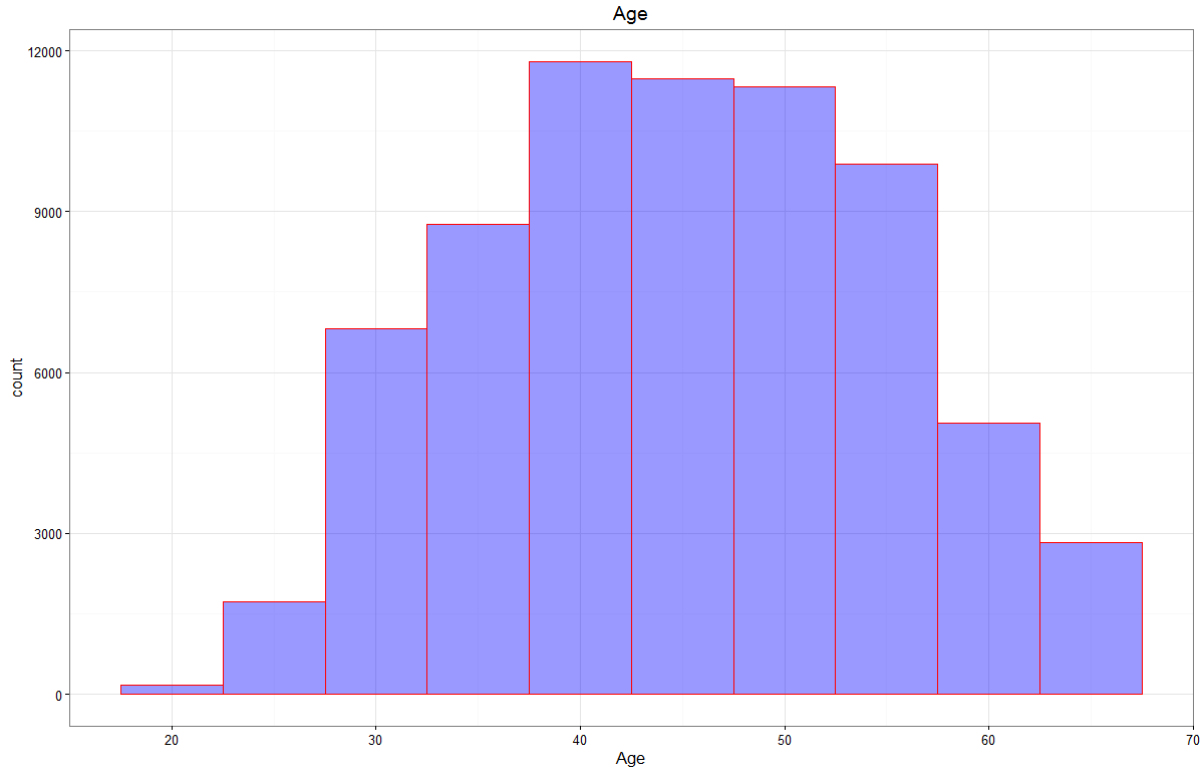
e. The data seems to have outliers also by looking at summary. These will be observed and taken care in the next process.
f. Performance Tag NA implies that no other bank provided card to them so data is unavailable. Making a separate data frame for the rows where Performance Tag is NA. Will check final model to get good accuracy on this data set of 1425 rows

Exploratory Data Analysis

1.Age : Age variable seems to have outliers. "-3" "0" "15" "16" "17" have to be capped at 18 as minimum age criteria for holding Credit card is 18 years

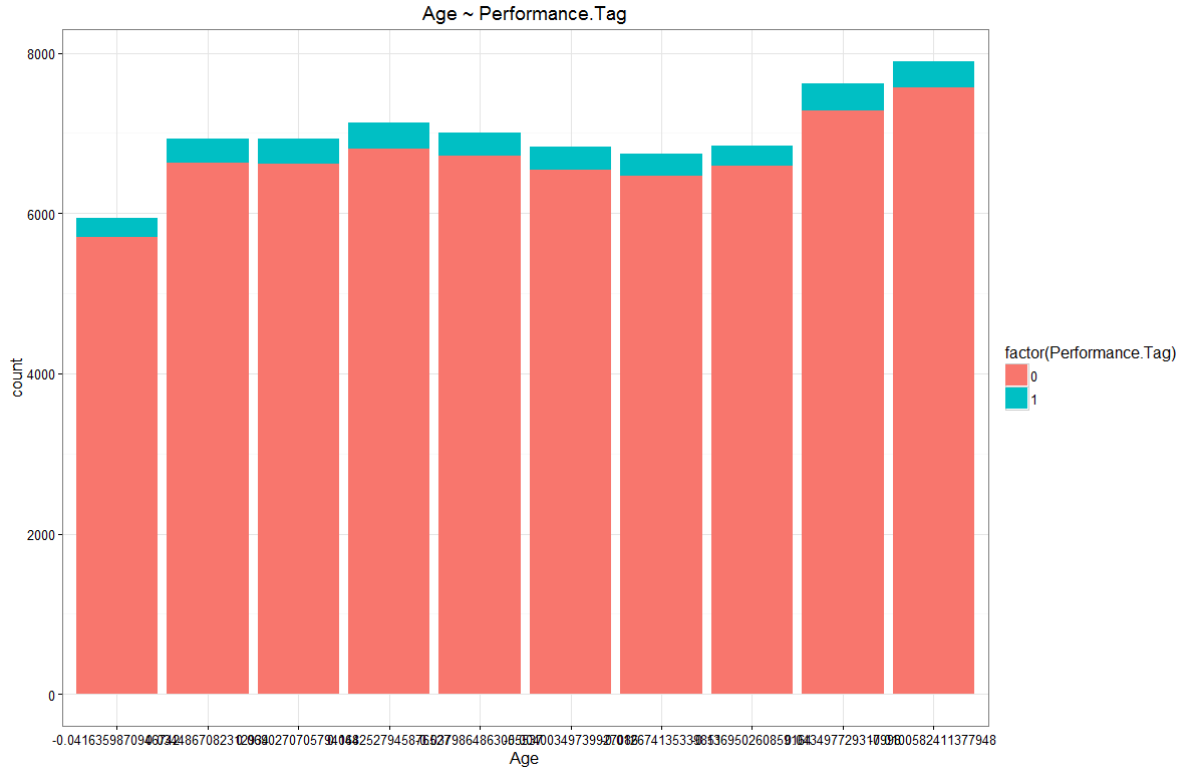
Post flooring Age to 18: Outliers are removed. Normal distribution obtained.



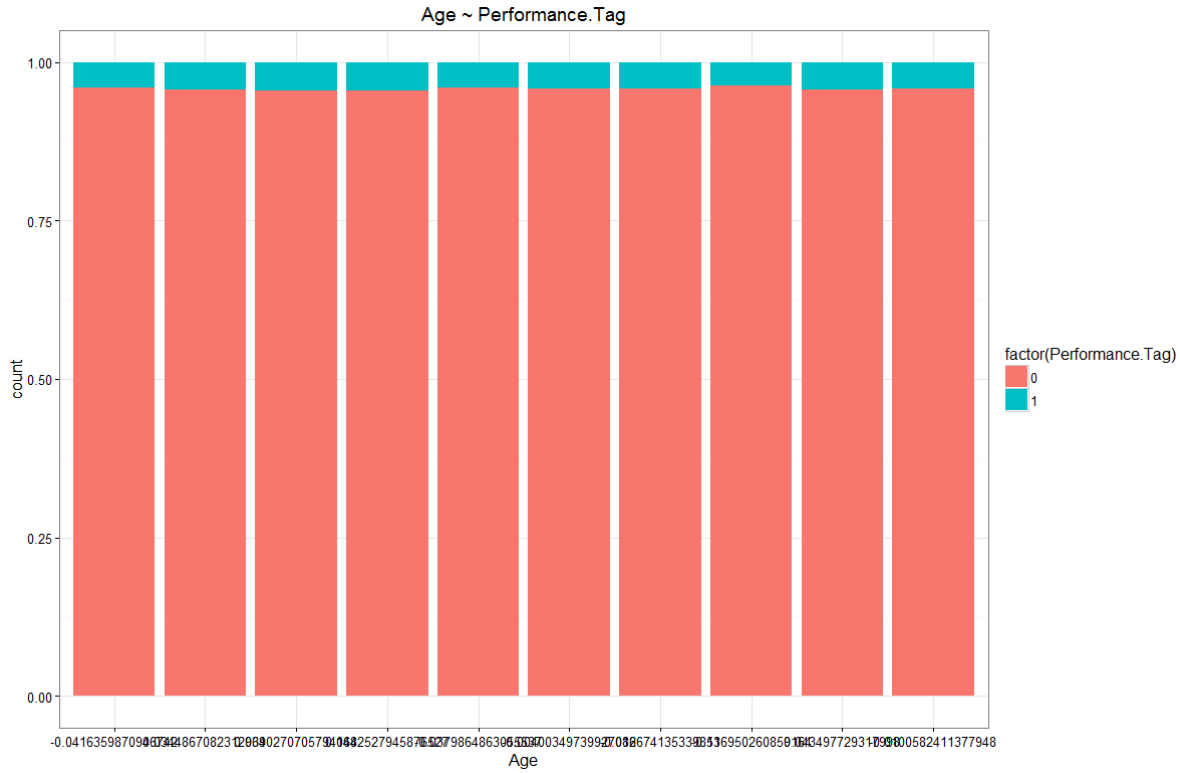


After binning age and changing the categories with their WOE Values:

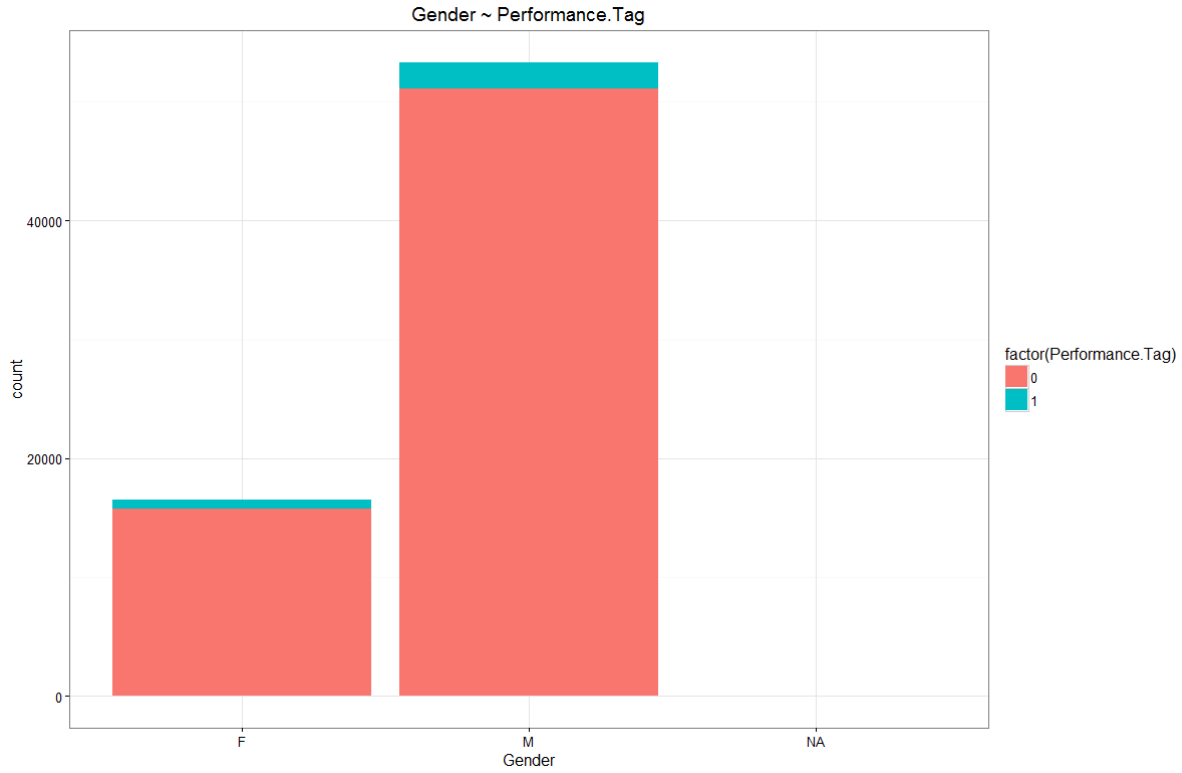
	Age	N	Percent	WOE	IV
1	(17,30]	5946	0.08510821	-0.041635987	0.0001447595
2	(30,35]	6927	0.09914978	0.034486708	0.0002645613
3	(35,38]	6924	0.09910684	0.069027071	0.0007519893
4	(38,41]	7129	0.10204111	0.068252795	0.0012424782
5	(41,44]	7007	0.10029486	-0.037986486	0.0013847103
6	(44,47]	6830	0.09776136	-0.004003497	0.0013862744
7	(47,50]	6743	0.09651609	-0.012674135	0.0014016885
8	(50,53]	6841	0.09791881	-0.136950261	0.0031273031
9	(53,57]	7618	0.10904042	0.043497729	0.0033377714
10	(57,65]	7899	0.11306252	-0.010058241	0.0033491572



By plotting an absolute bar chart, we can see that all have nearly equal distribution for default

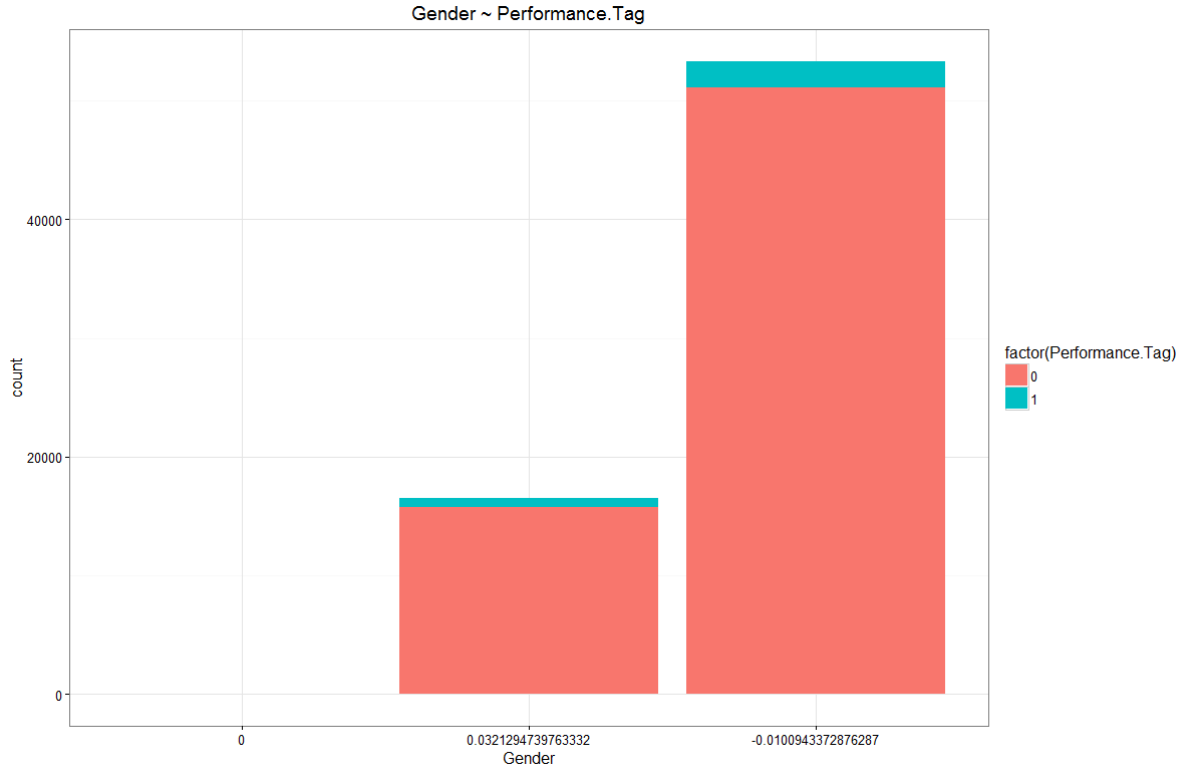


2. Gender: It has 2 NAs

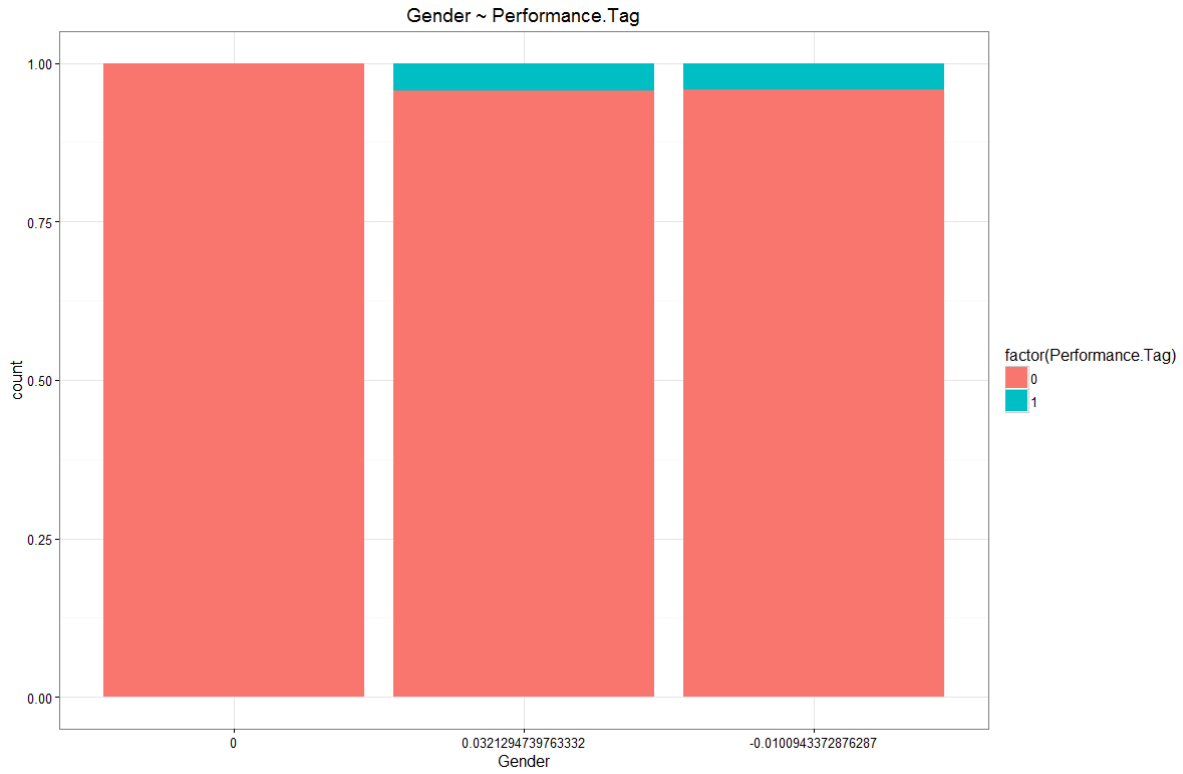


After changing the categories with their WOE Values:

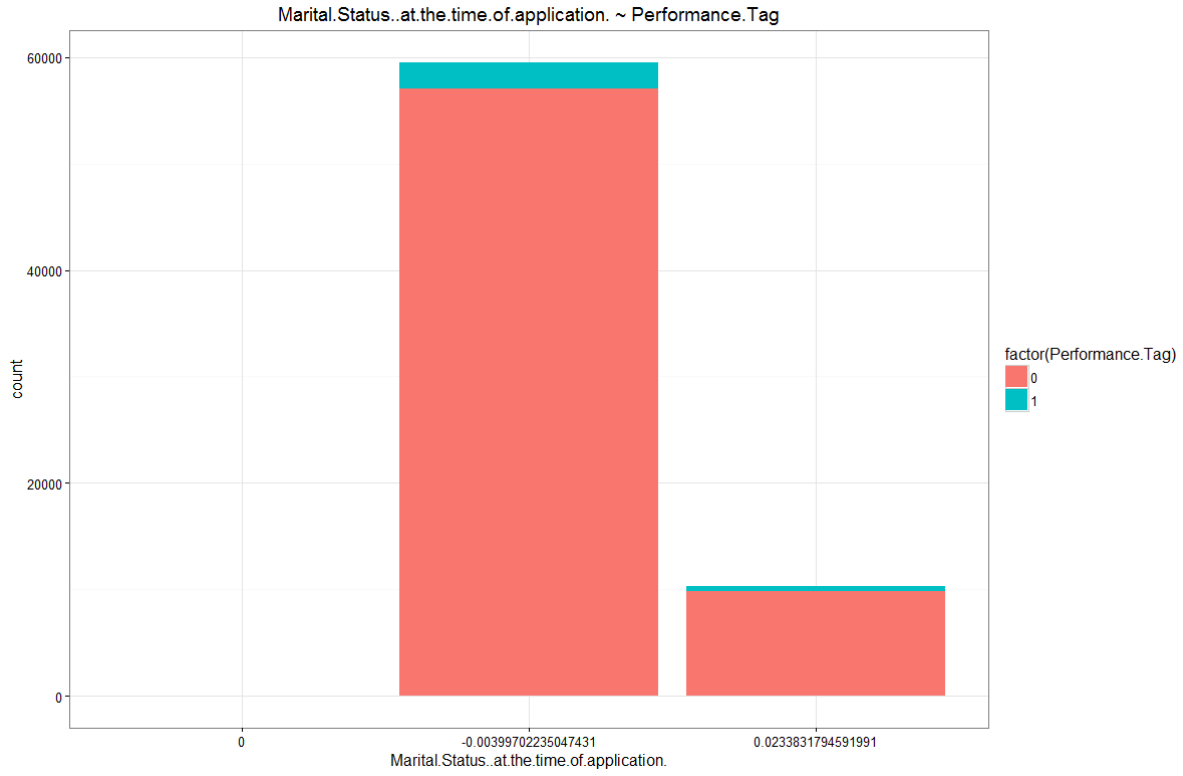
	Gender	N	Percent	WOE	IV
1		2	0.00002862705	0.00000000	0.0000000000
2	F	16506	0.23625901752	0.03212947	0.0002475104
3	M	53356	0.76371235543	-0.01009434	0.0003249707



By plotting an absolute bar chart, we can see that both Males and Females have equal distribution for default.

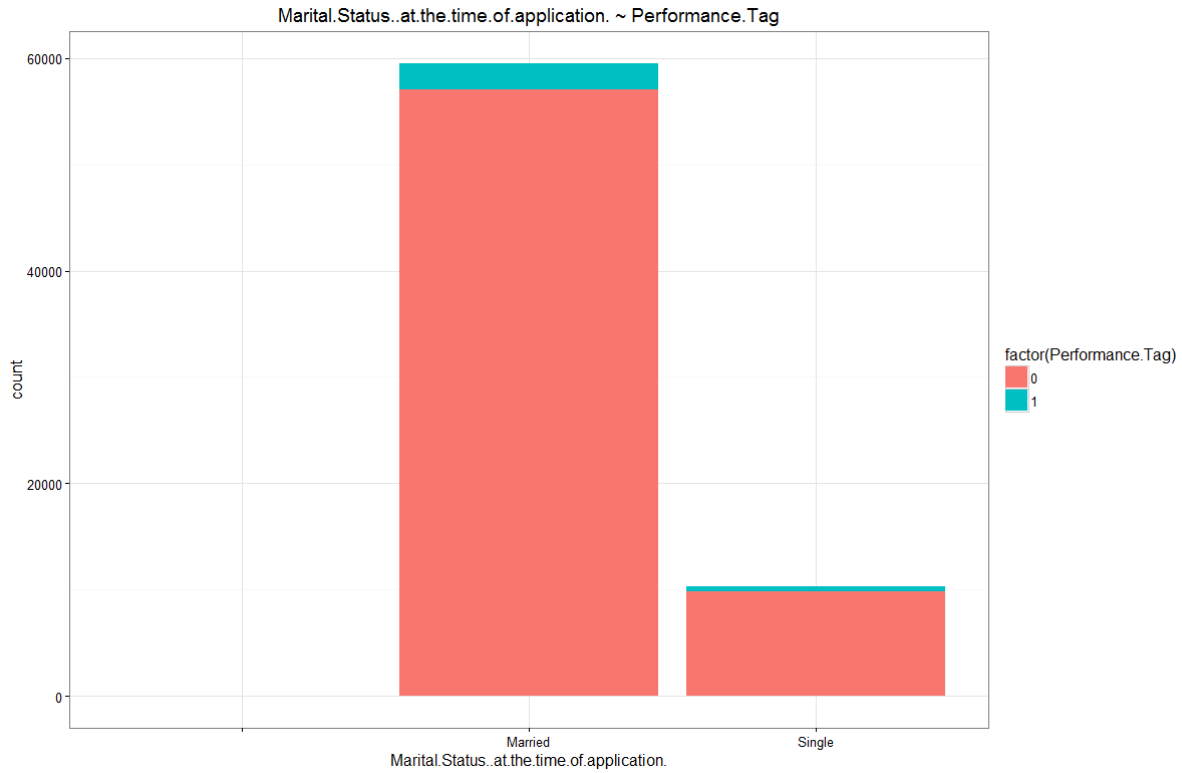


3. Marital.Status..at.the.time.of.application. : 6 NAs

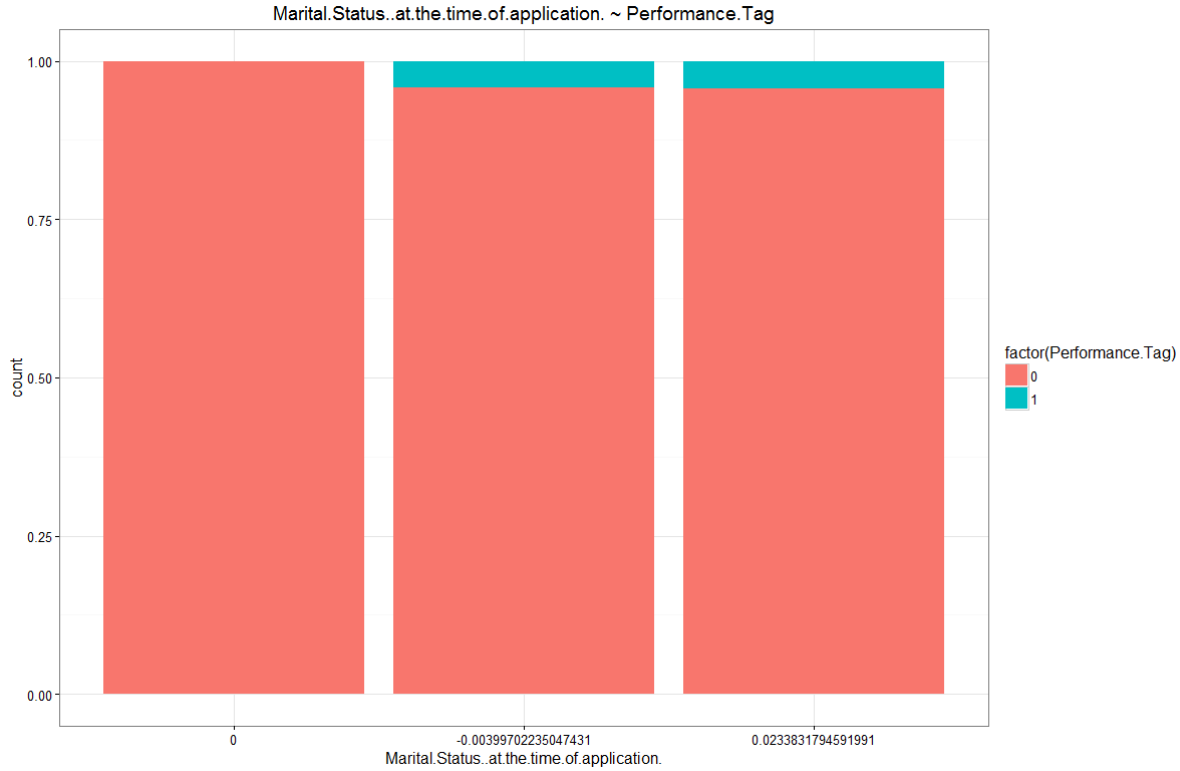


After changing the categories with their WOE Values:

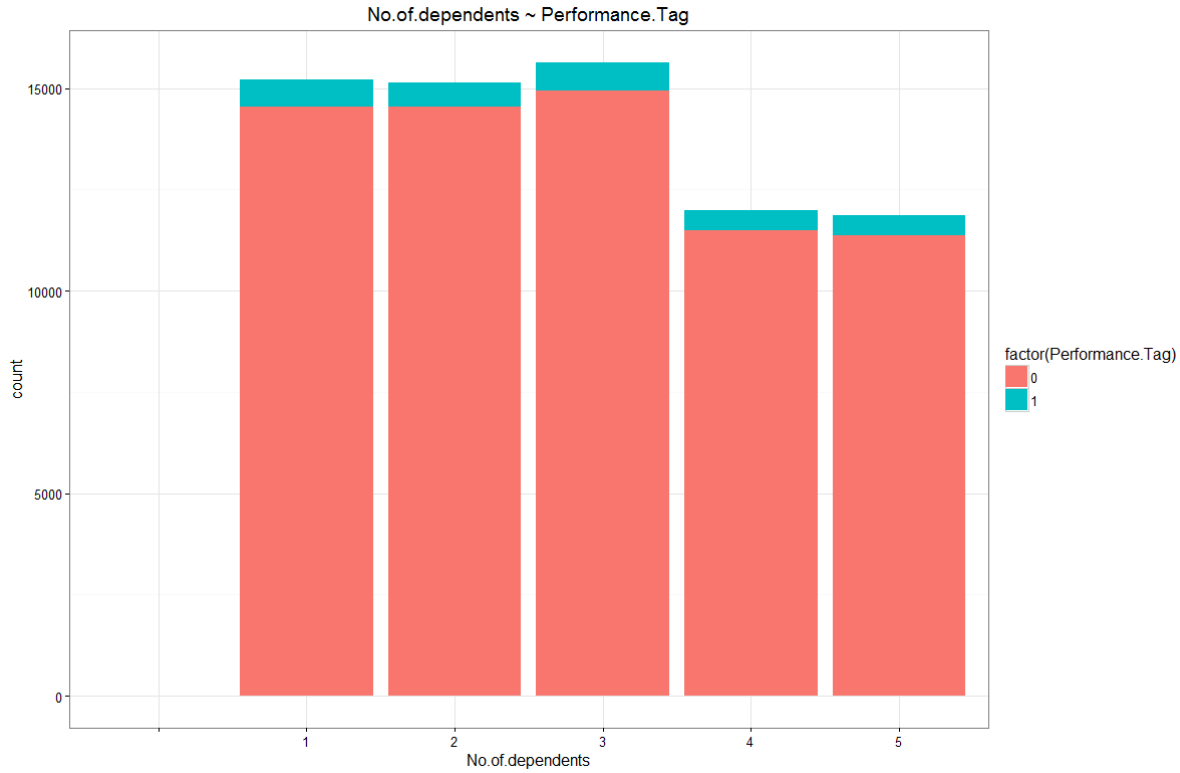
	Marital.Status..at.the.time.of.application.	N	Percent	WOE	IV
1		6	0.00008588114	0.000000000	0.00000000000
2	Married	59542	0.85225581129	-0.003997022	0.00001359091
3	Single	10316	0.14765830757	0.023383179	0.00009519639



By plotting an absolute bar chart, we can see that both Married and Singles have equal distribution for default



4. No.of.dependents: 3 NAs

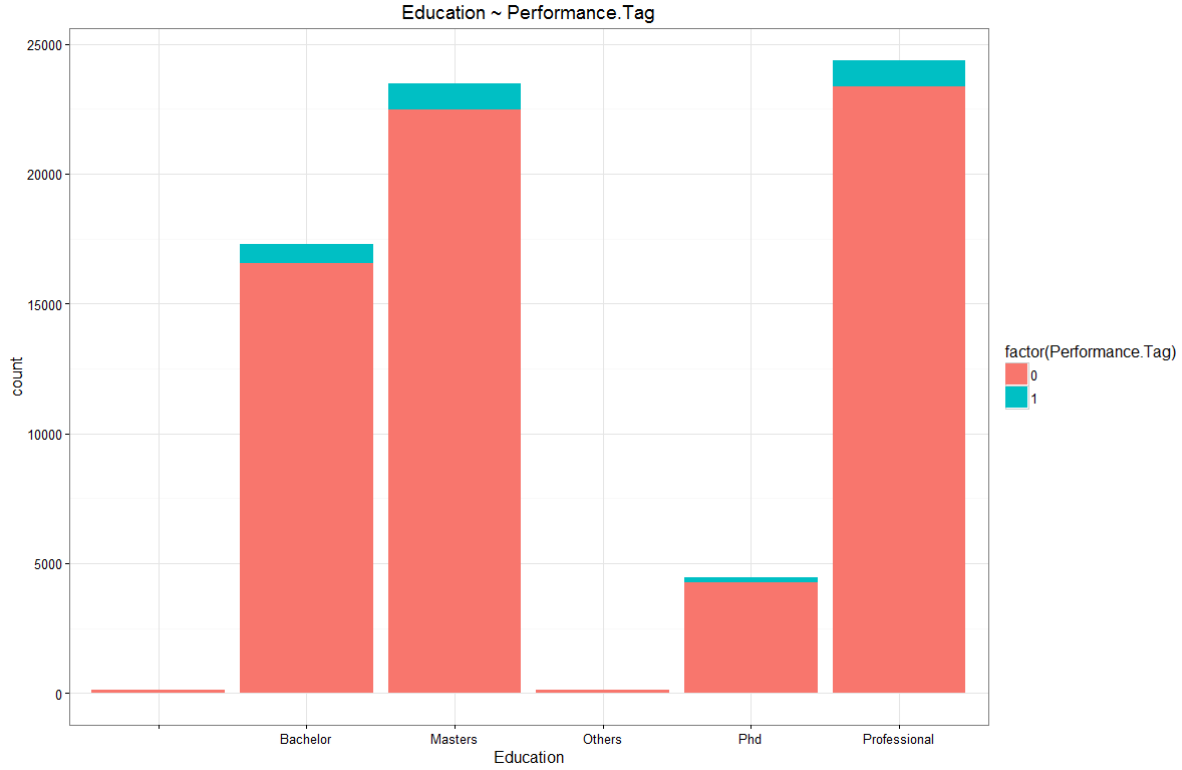


After changing the categories with their WOE Values:

	No.of.dependents	N	Percent	WOE	IV
1		3	0.00004294057	0.000000000	0.0000000000
2	1	15218	0.21782319936	0.040040389	0.0003556941
3	2	15127	0.21652066873	-0.085197683	0.0018674600
4	3	15644	0.22392076033	0.054177500	0.0025412603
5	4	11997	0.17171934043	-0.025162291	0.0026487390
6	5	11875	0.16997309058	0.004346039	0.0026519559

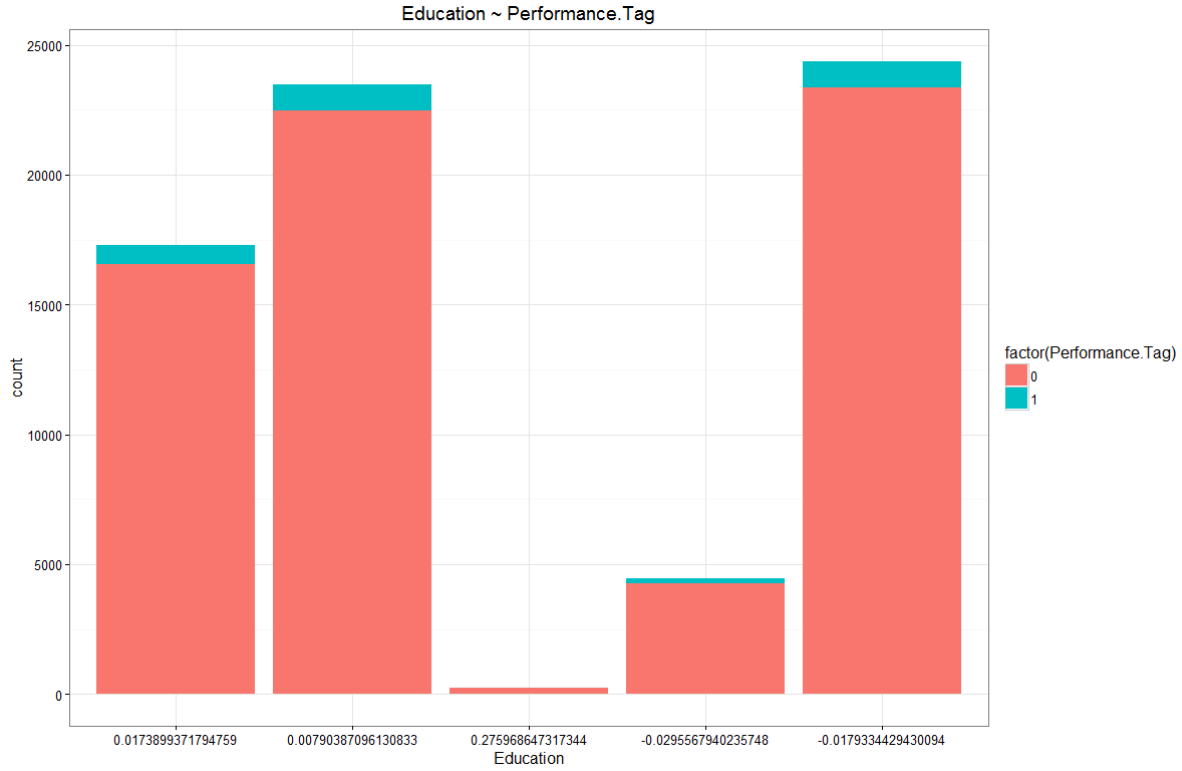
By plotting an absolute bar chart, we can see that all categories have equal distribution for default

5. Education: Has 118 NAs

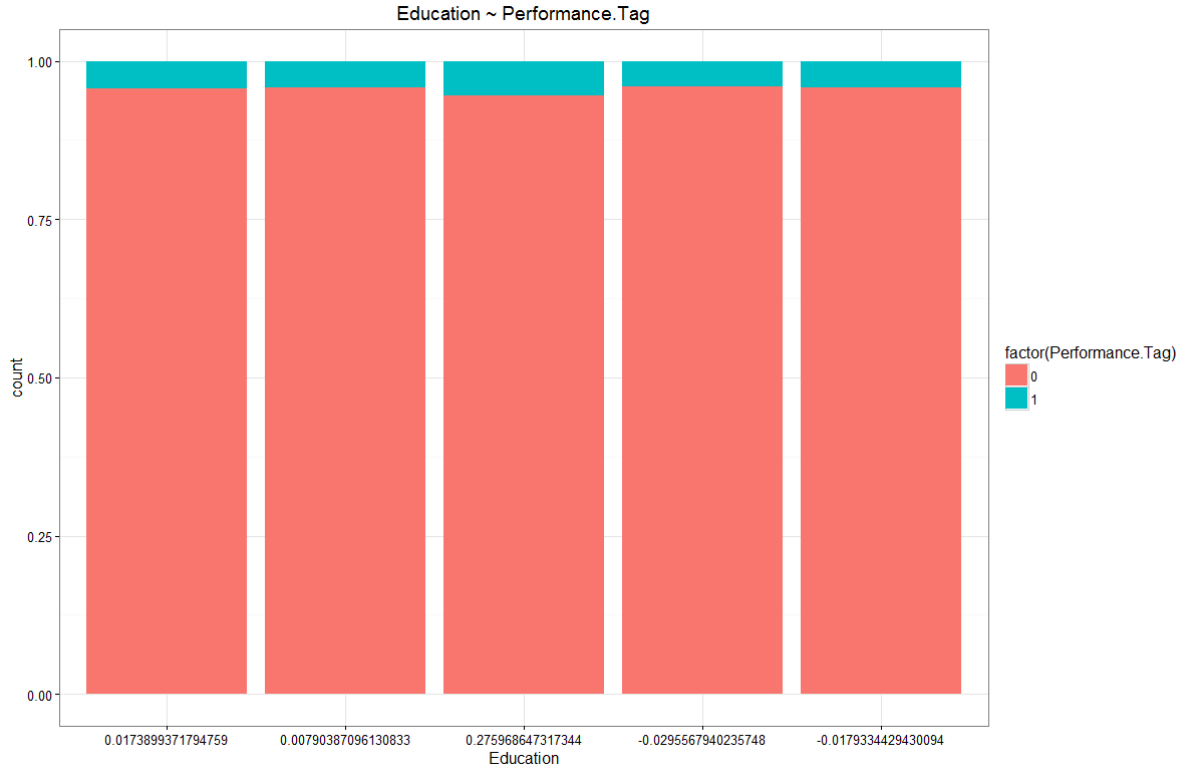


Merging Blanks to “Others” and after changing the categories with their WOE Values:

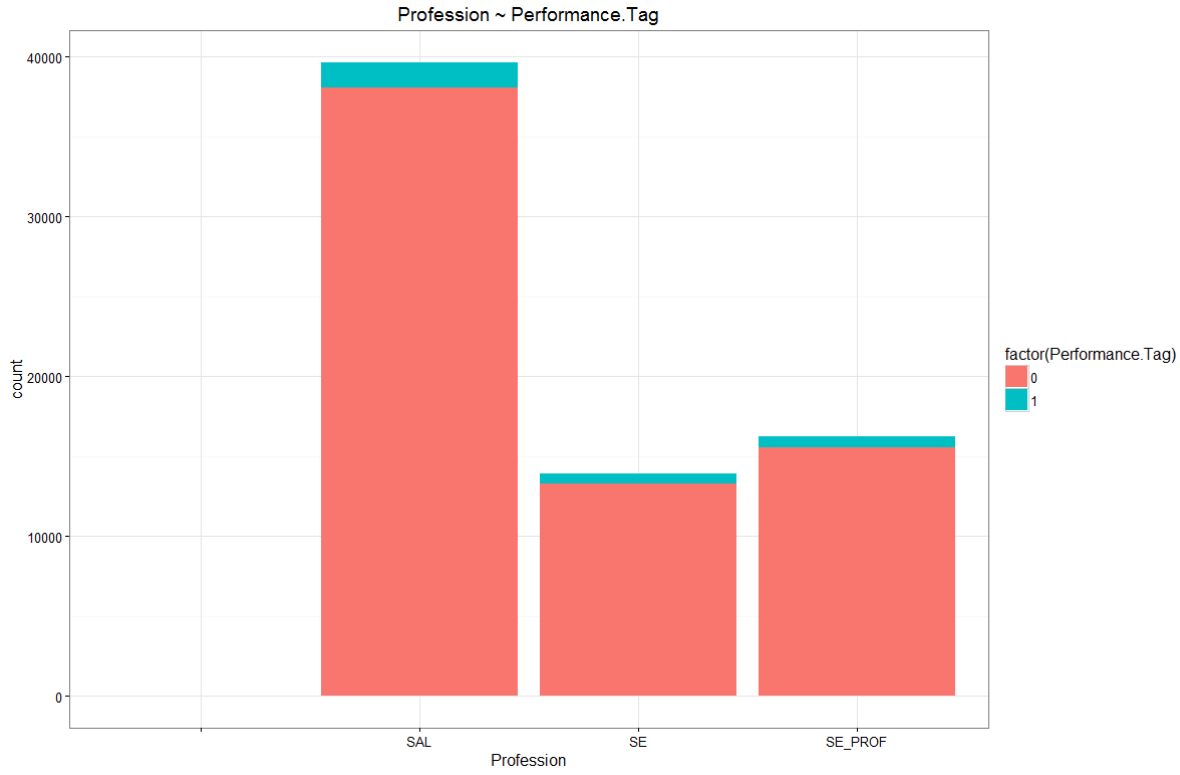
	Education	N	Percent	WOE	IV
1	Bachelor	17300	0.247623955	0.017389937	0.00007548299
2	Masters	23481	0.336095843	0.007903871	0.00009655543
3	Others	237	0.003392305	0.275968647	0.00039014062
4	Phd	4463	0.063881255	-0.029556794	0.00044519851
5	Professional	24383	0.349006641	-0.017933443	0.00055652497



By plotting an absolute bar chart, we can see that both all have equal distribution for default

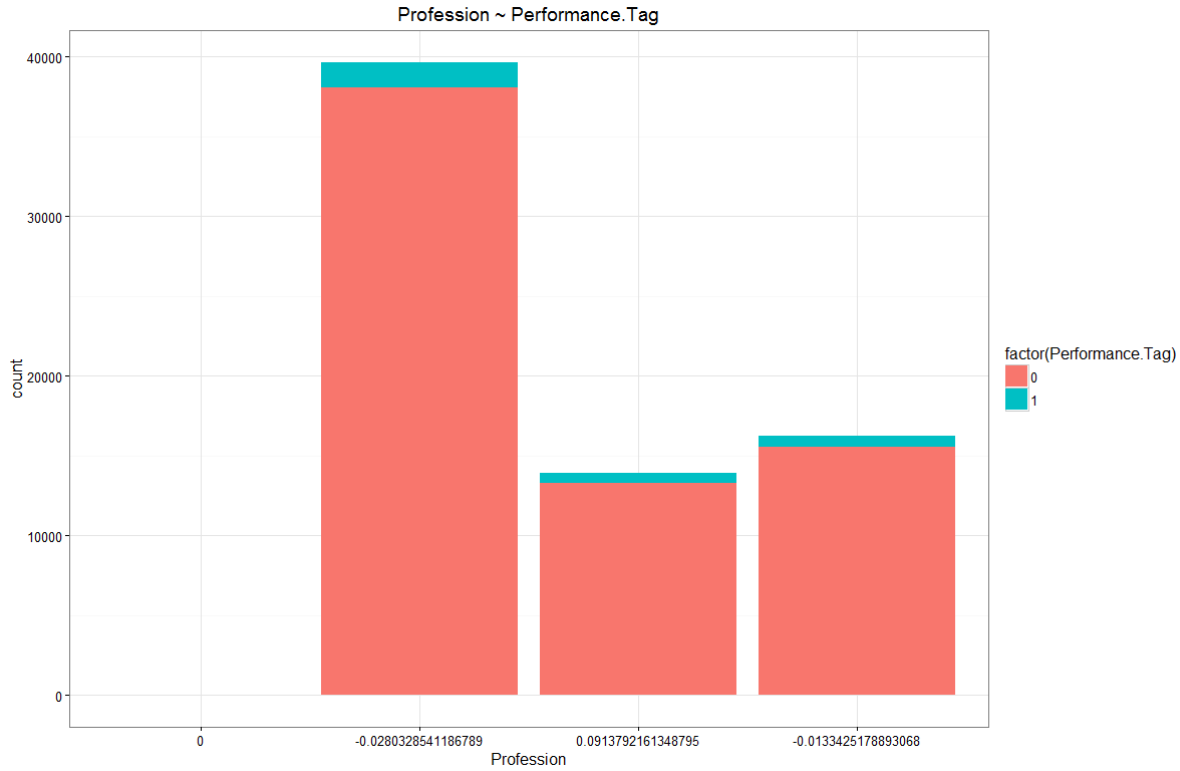


6. Profession: 13 NAs

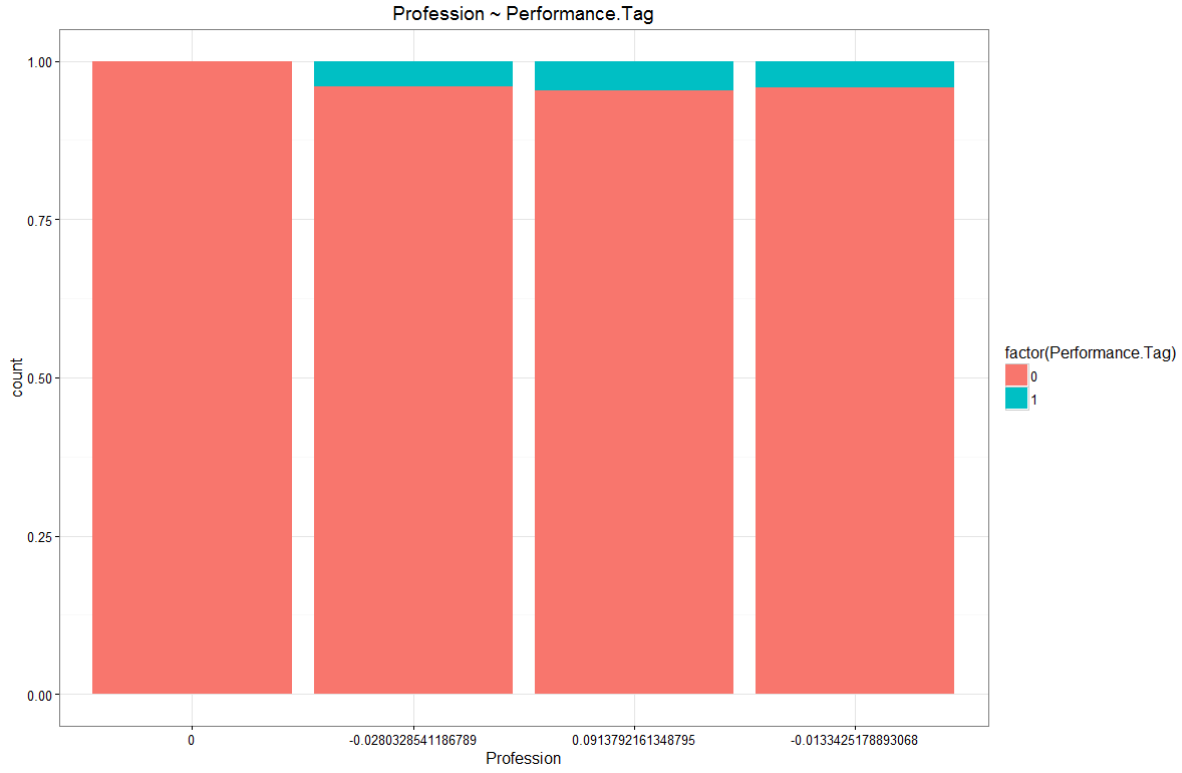


After changing the categories with their WOE Values:

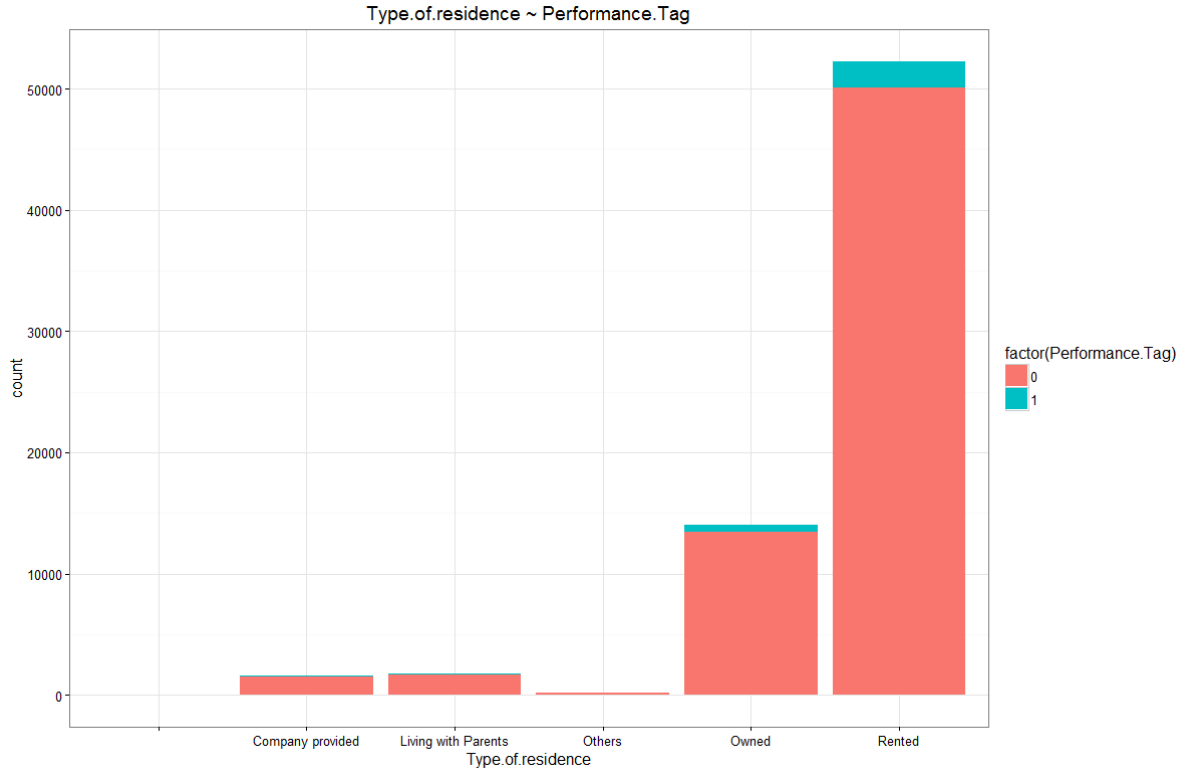
	Profession	N	Percent	WOE	IV
1		13	0.0001860758	0.00000000	0.0000000000
2	SAL	39670	0.5678174739	-0.02803285	0.0004405316
3	SE	13925	0.1993158136	0.09137922	0.0021762573
4	SE_PROF	16256	0.2326806367	-0.01334252	0.0022174277



By plotting an absolute bar chart, we can see that all have equal distribution for default

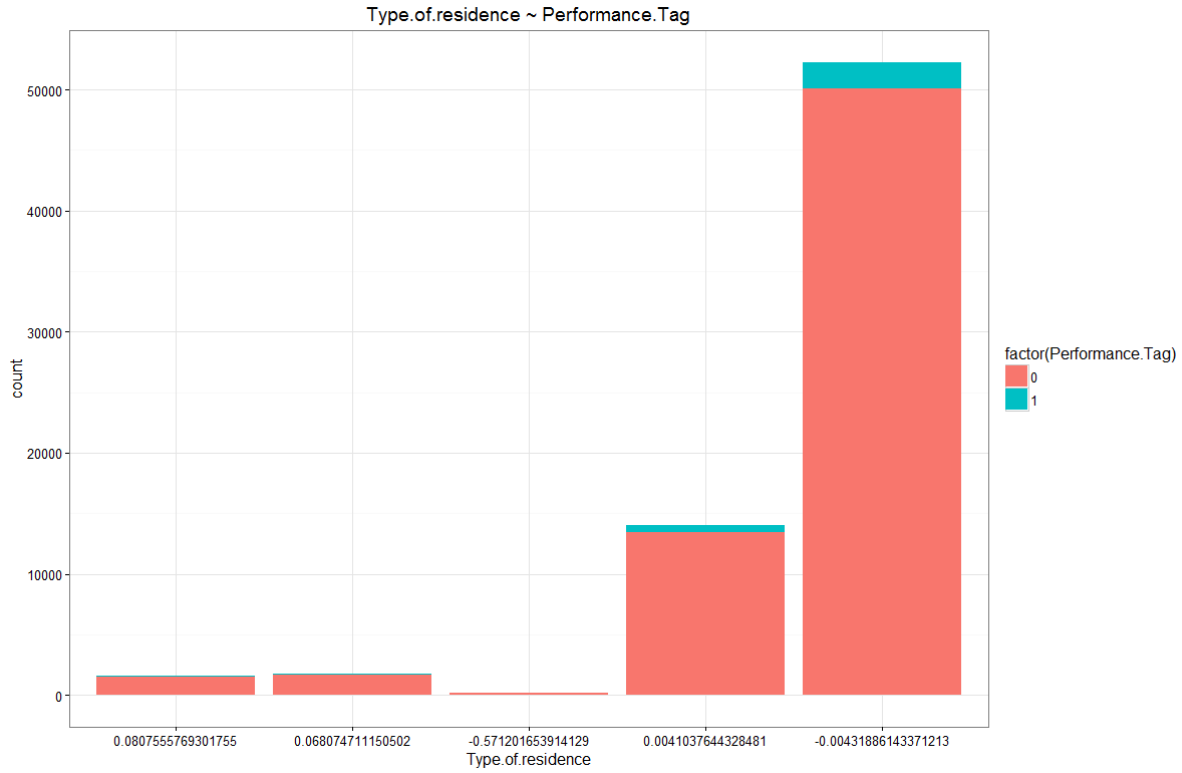


7. Type.of.residence: 8 NAs

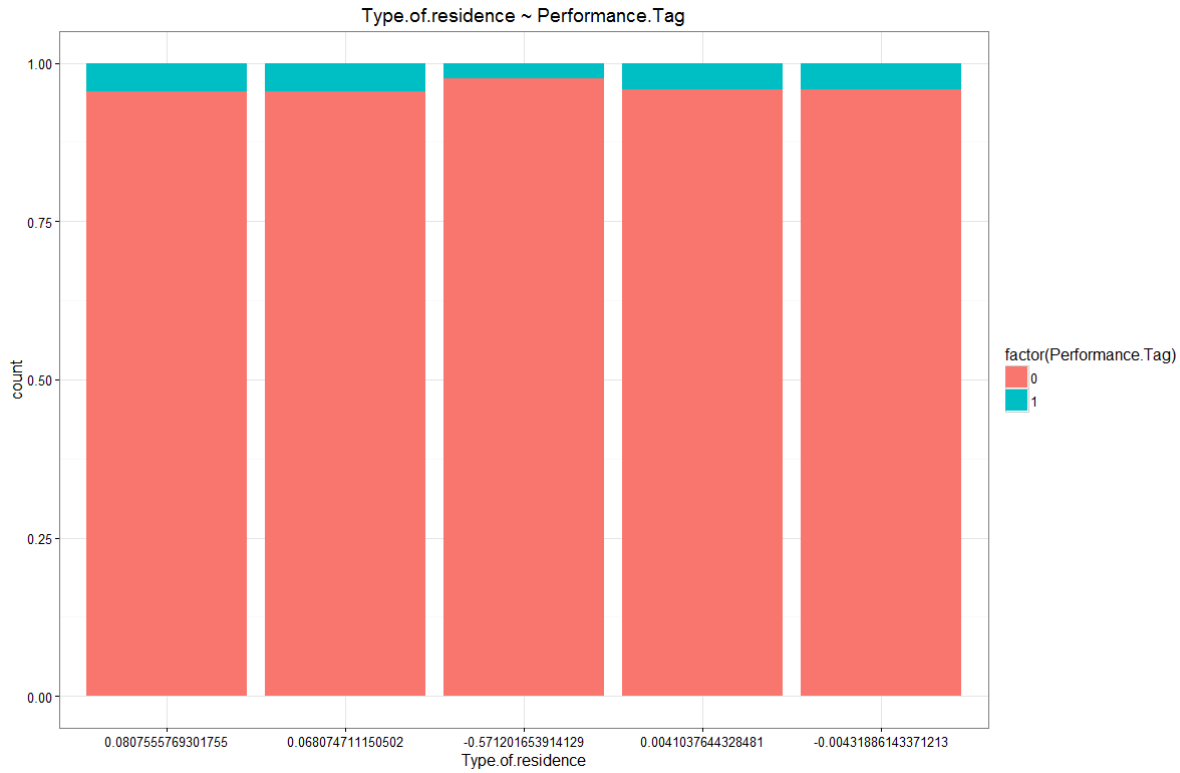


Merging Blanks to “Others” and after changing the categories with their WOE Values:

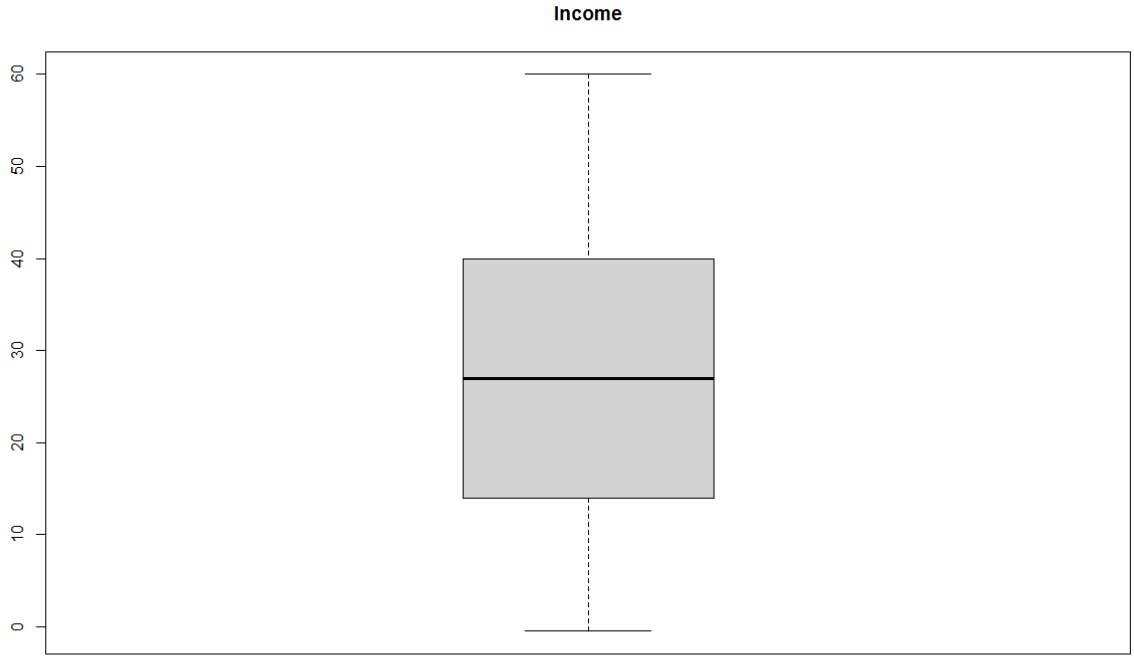
	Type.of.residence	N	Percent	WOE	IV
1	Company provided	1602	0.022930265	0.080755577	0.0001551922
2	Living with Parents	1777	0.025435131	0.068074711	0.0002768061
3	Others	206	0.002948586	-0.571201654	0.0010234125
4	Owned	14003	0.200432268	0.004103764	0.0010267944
5	Rented	52276	0.748253750	-0.004318861	0.0010407236

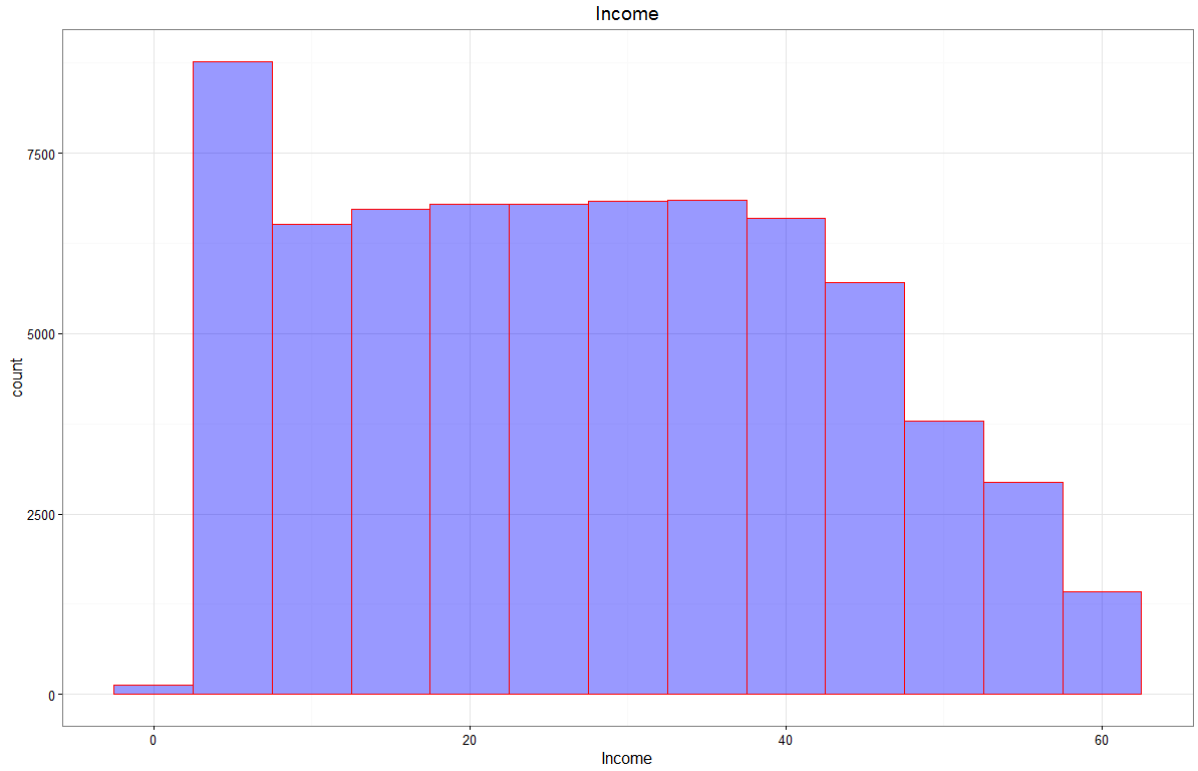


By plotting an absolute bar chart, we can see that all have nearly equal distribution for default



8. Income: Mean income is 14 (units). No outlier.

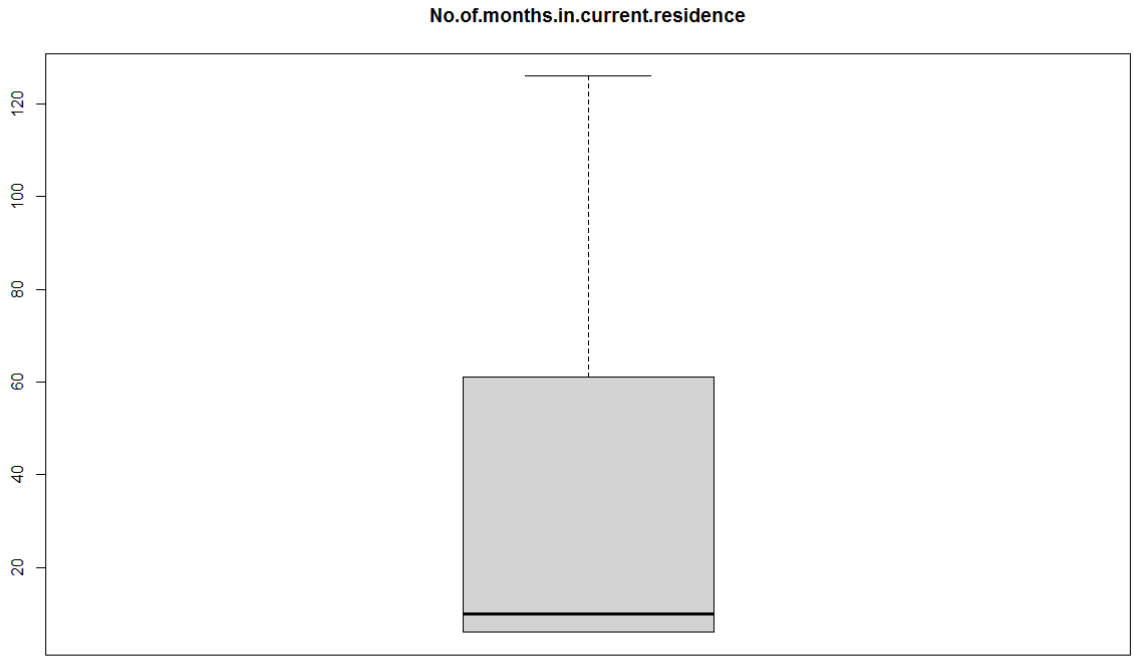


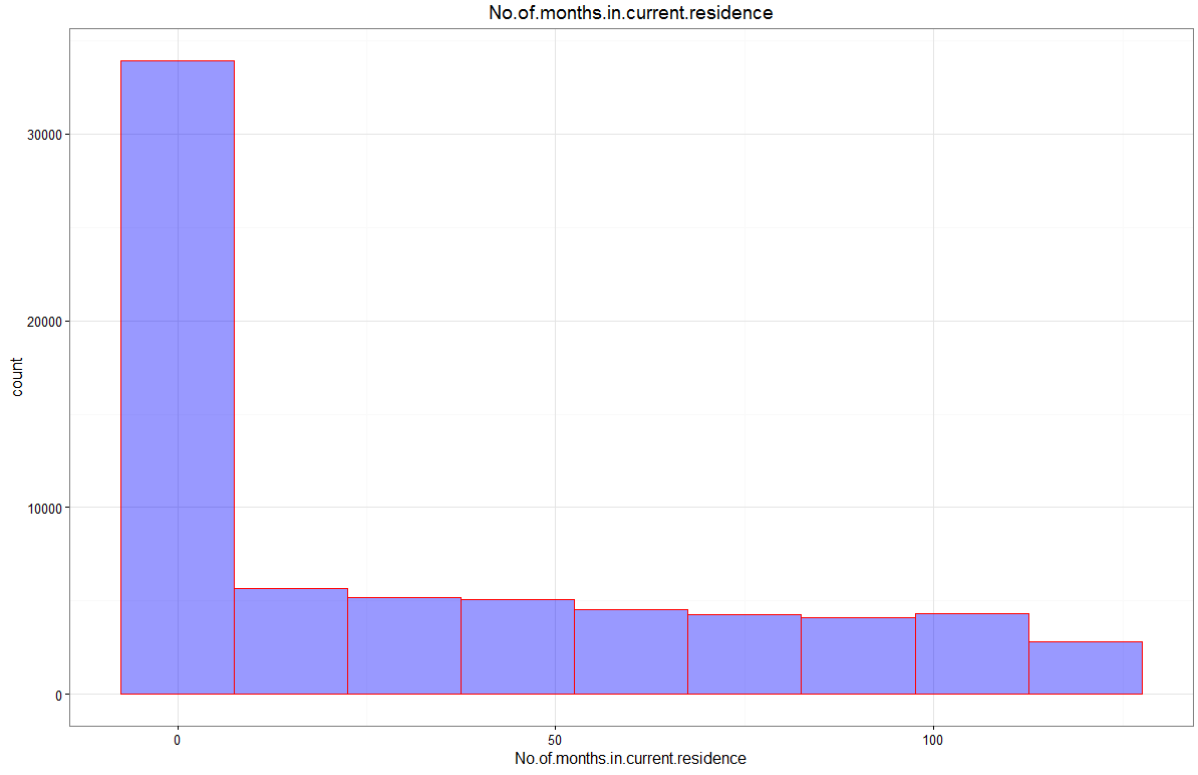


Income groups will have to binned as per IV

	Income	N	Percent	WOE	IV
1	[0,5]	6329	0.09059029	0.30259148	0.009544033
2	[6,10]	6510	0.09318104	0.27570608	0.017592008
3	[11,16]	7923	0.11340605	0.06604411	0.018101897
4	[17,21]	6803	0.09737490	0.08075769	0.018760966
5	[22,26]	6827	0.09771842	0.02517224	0.018823603
6	[27,31]	6817	0.09757529	0.07860384	0.019448649
7	[32,36]	6829	0.09774705	-0.15584790	0.021660495
8	[37,41]	6723	0.09622982	-0.26372600	0.027601688
9	[42,48]	7784	0.11141647	-0.17690835	0.030819626
10	[49,60]	7319	0.10476068	-0.36083049	0.042417800

9. No.of.months.in.current.residence : Median is 10 months

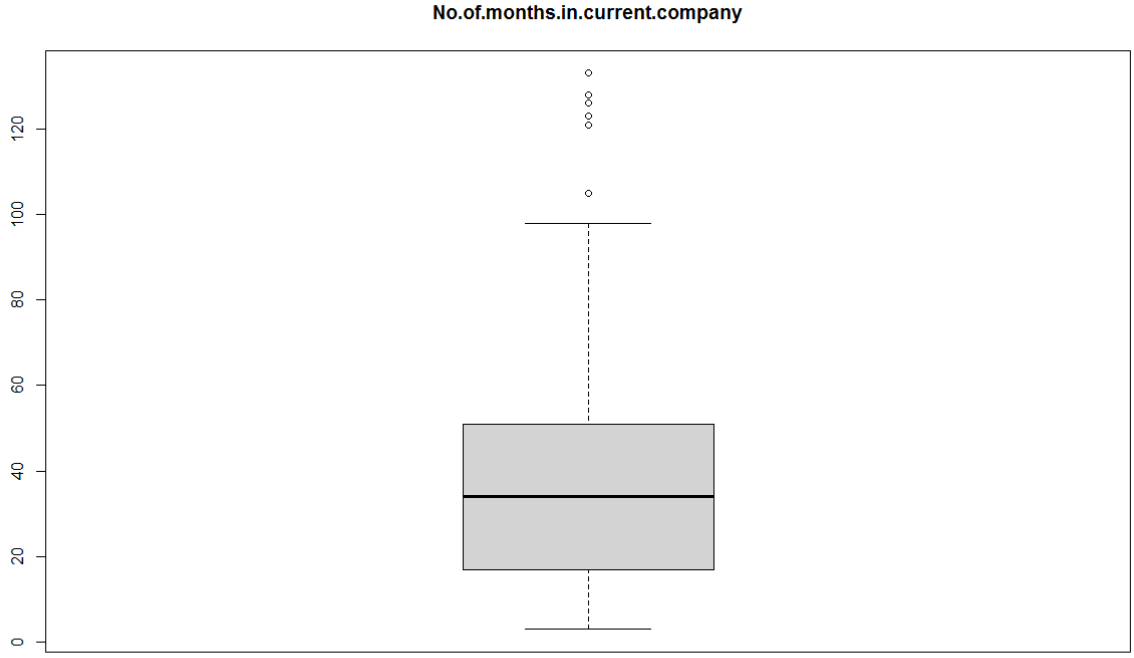


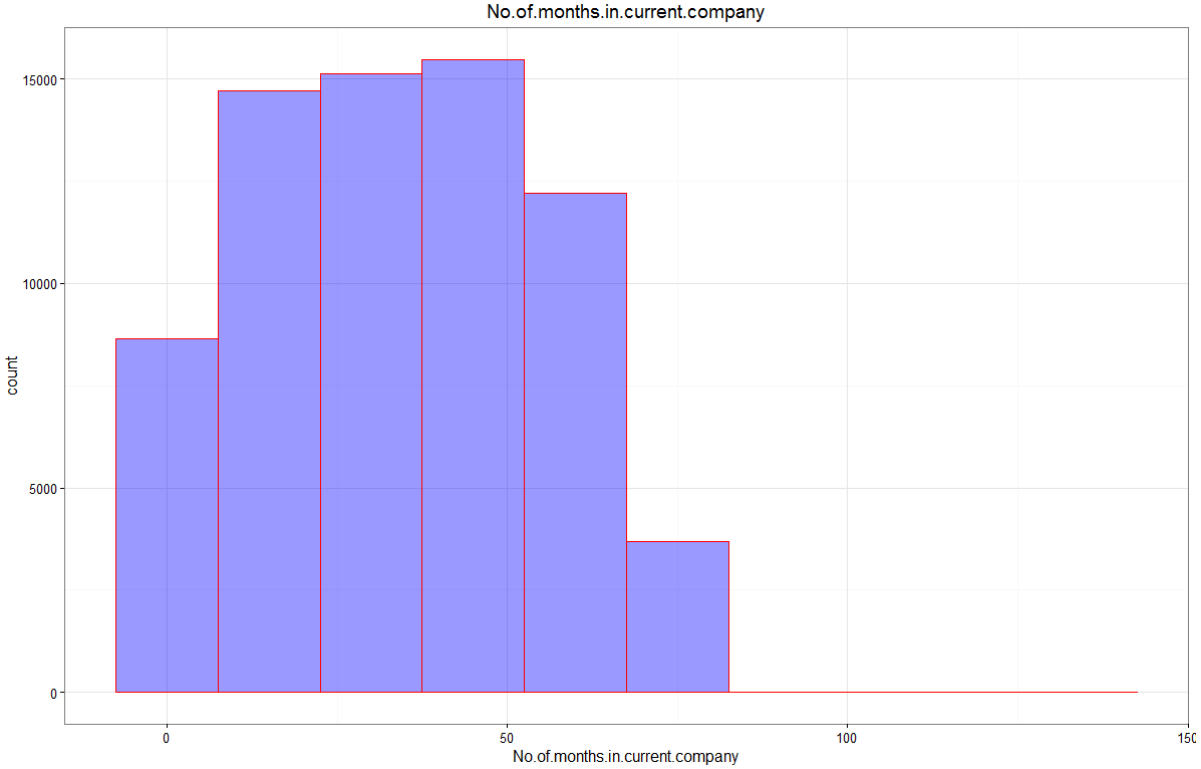


This field has to be binned as per IV:

	No.of.months.in.current.residence	N	Percent	WOE	IV
1	[6,9]	34693	0.49657907	-0.27220657	0.03253901
2	[10,28]	6922	0.09907821	0.49867827	0.06363545
3	[29,49]	7210	0.10320050	0.30113949	0.07439660
4	[50,72]	6988	0.10002290	0.13397271	0.07630615
5	[73,97]	6931	0.09920703	0.13943606	0.07836294
6	[98,126]	7120	0.10191229	-0.07681208	0.07894353

10. No.of.months.in.current.residence: Median is 34 months. It seems to have some outliers which will be taken care of when we map WOE weights to the bin

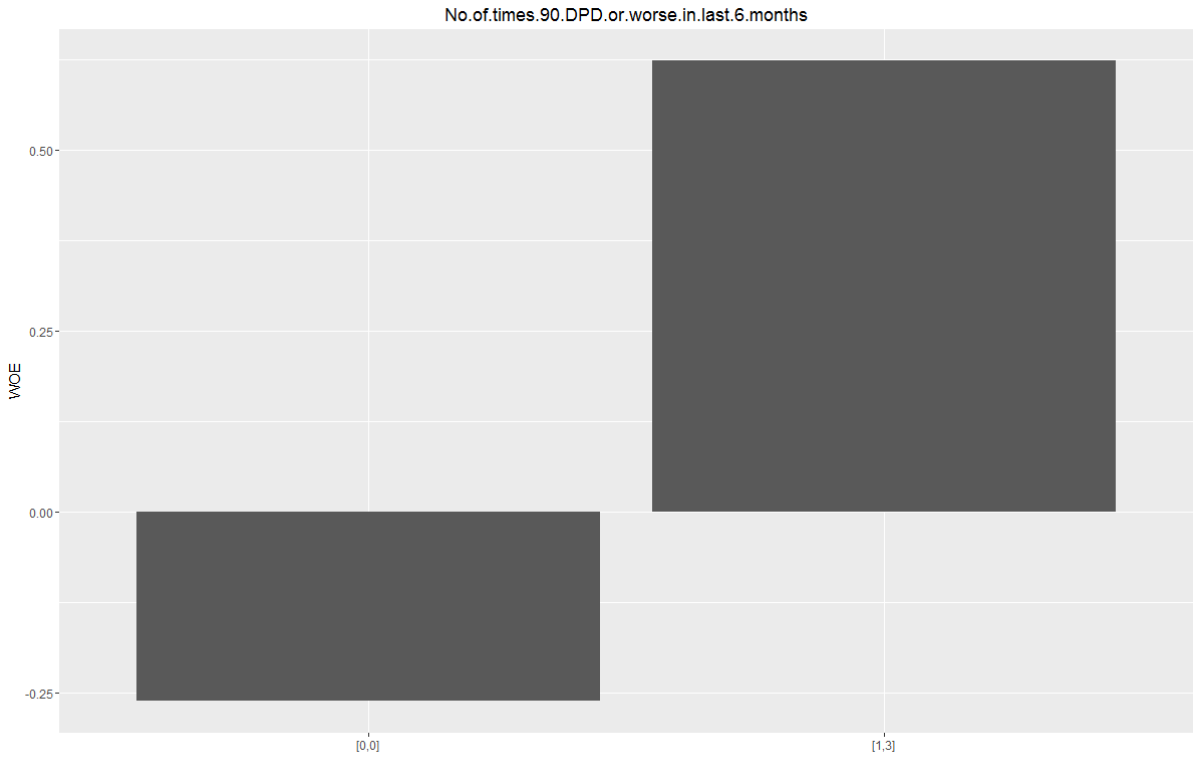

No.of.months.in.current.company
0
20
40
60
80
100
120


No.of.months.in.current.company
count
0
5000
10000
15000
0
50
100
No.of.months.in.current.company

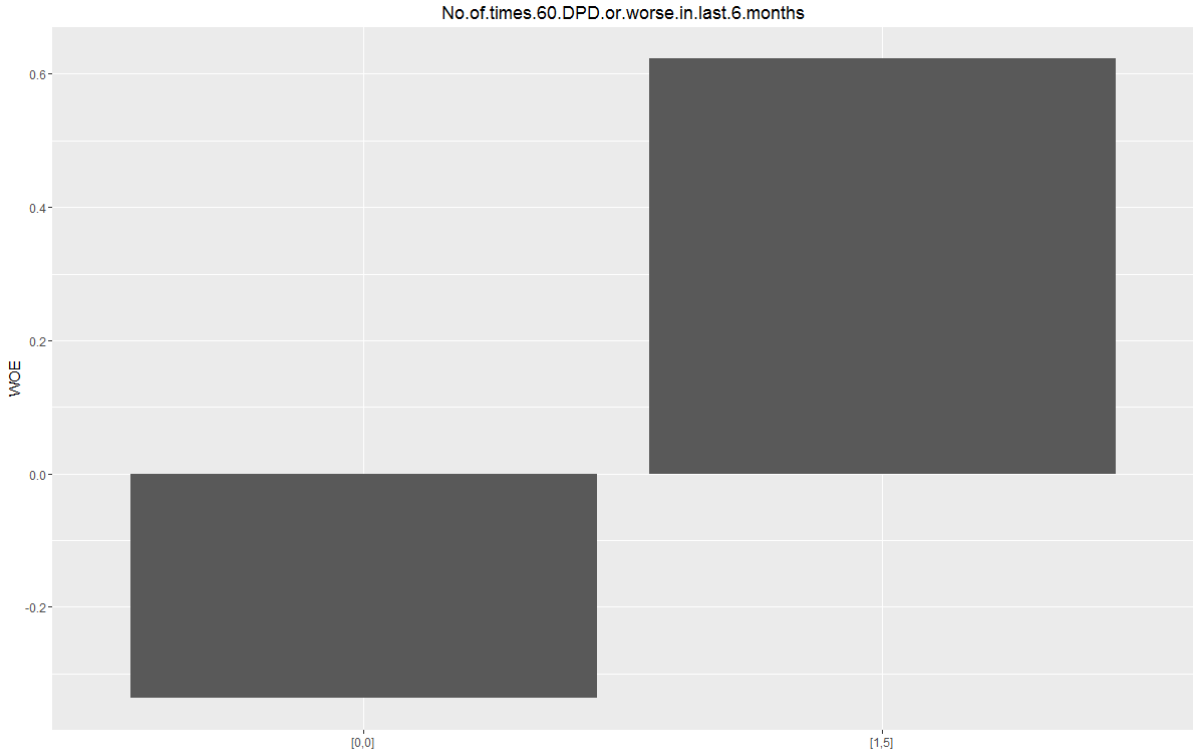
	No.of.months.in.current.company	N	Percent	WOE	IV
1	[3,5]	6689	0.09574316	0.09847101	0.0009713844
2	[6,12]	6797	0.09728902	0.17559050	0.0042241321
3	[13,19]	6933	0.09923566	0.20626208	0.0088680976
4	[20,26]	6919	0.09903527	0.03915191	0.0090226567
5	[27,33]	7104	0.10168327	-0.08572088	0.0097411937
6	[34,40]	7182	0.10279973	0.03074914	0.0098397718
7	[41,47]	7217	0.10330070	-0.17619333	0.0128001924
8	[48,53]	6169	0.08830013	-0.21796666	0.0166009719
9	[54,61]	7822	0.11196038	-0.21618018	0.0213452997
10	[62,133]	7032	0.10065270	0.06284108	0.0217544128

Plotting WOE Analysis for below numeric variables:

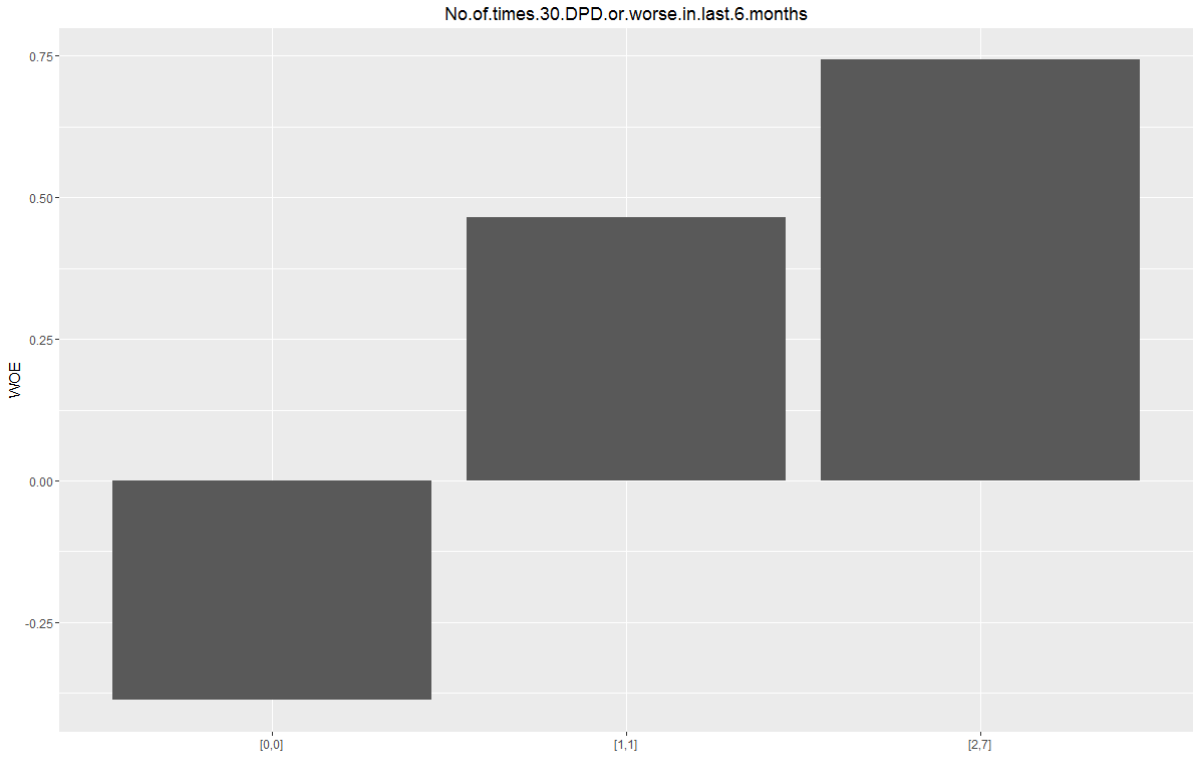
11. No of times 90 DPD or worse in last 6 months



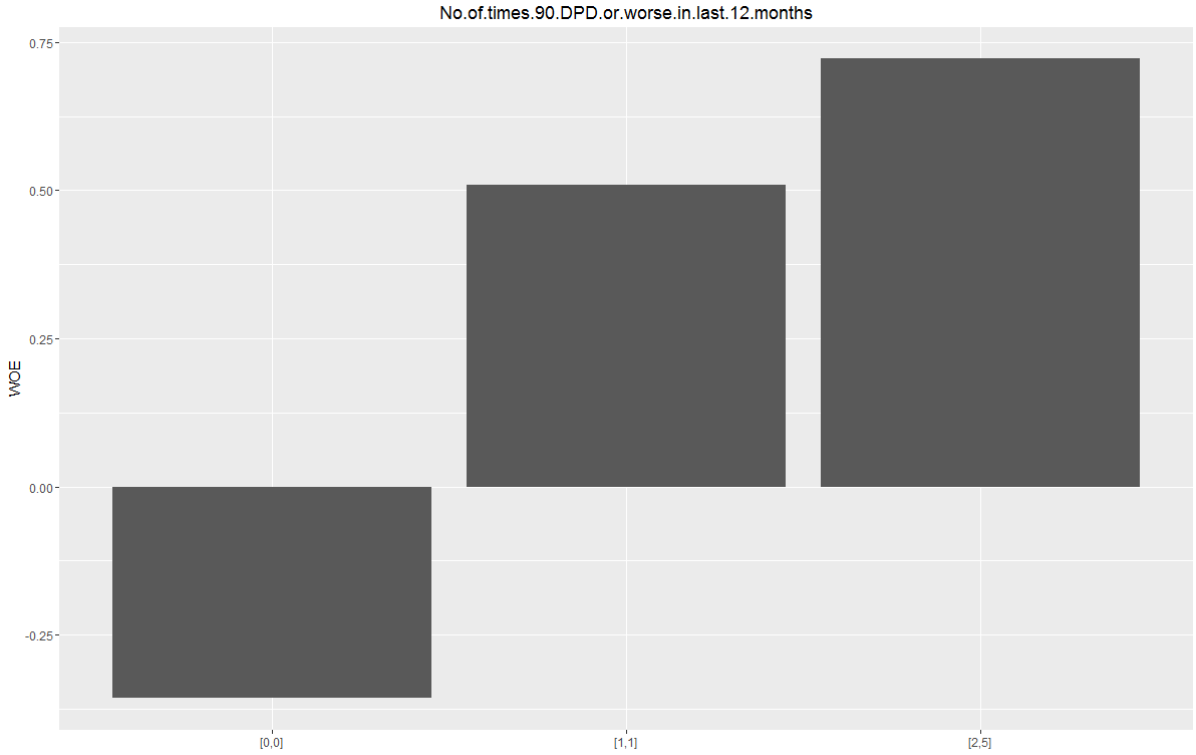
12. No of times 60 DPD or worse in last 6 months



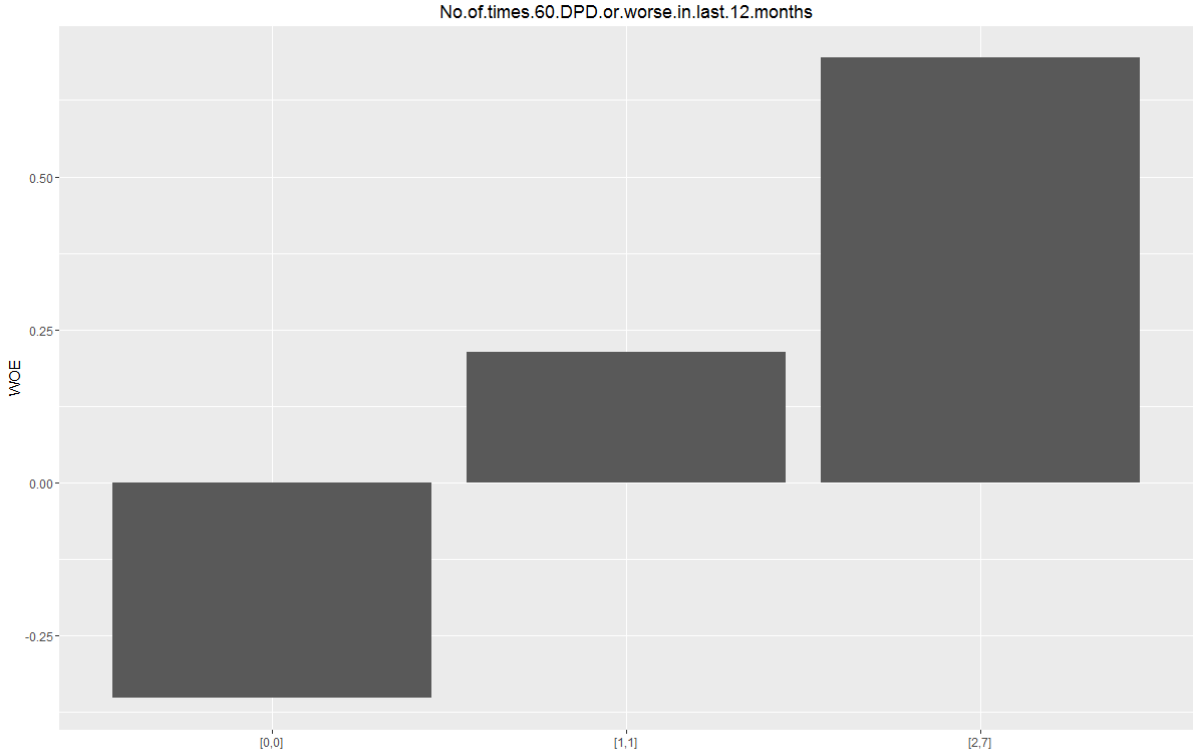
13. No of times 30 DPD or worse in last 6 months



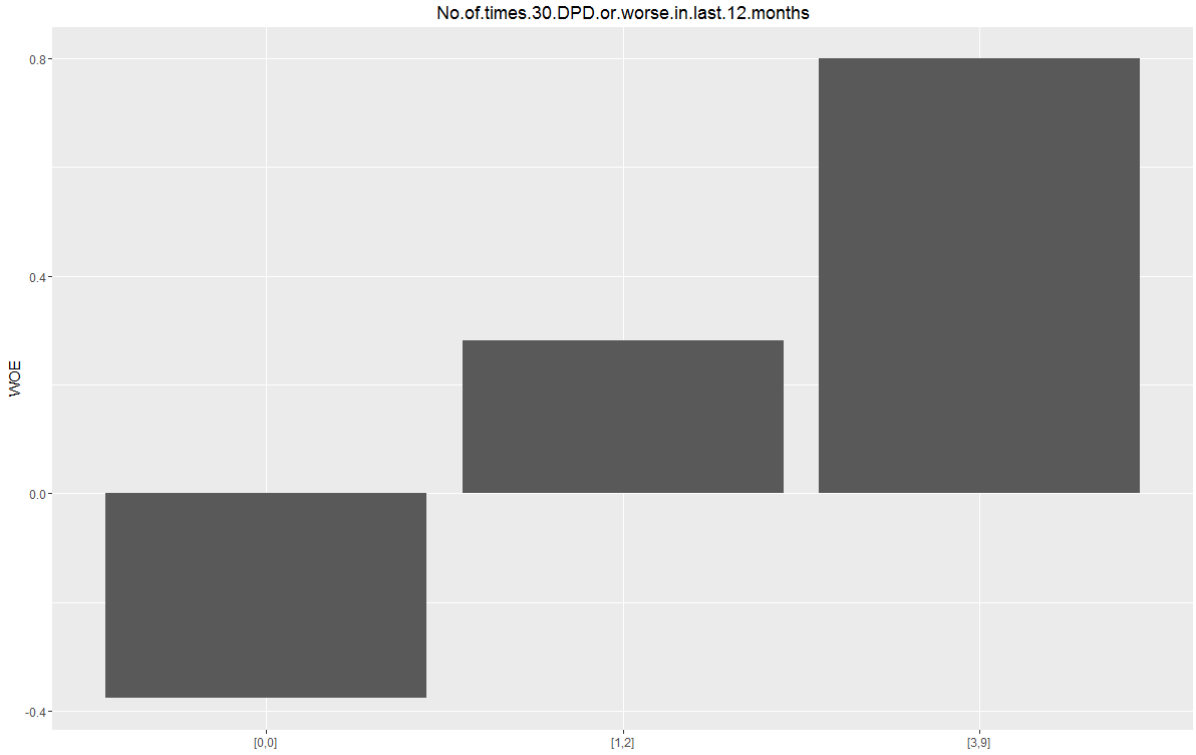
14. No of times 90 DPD or worse in last 12 months



15. No of times 60 DPD or worse in last 12 months

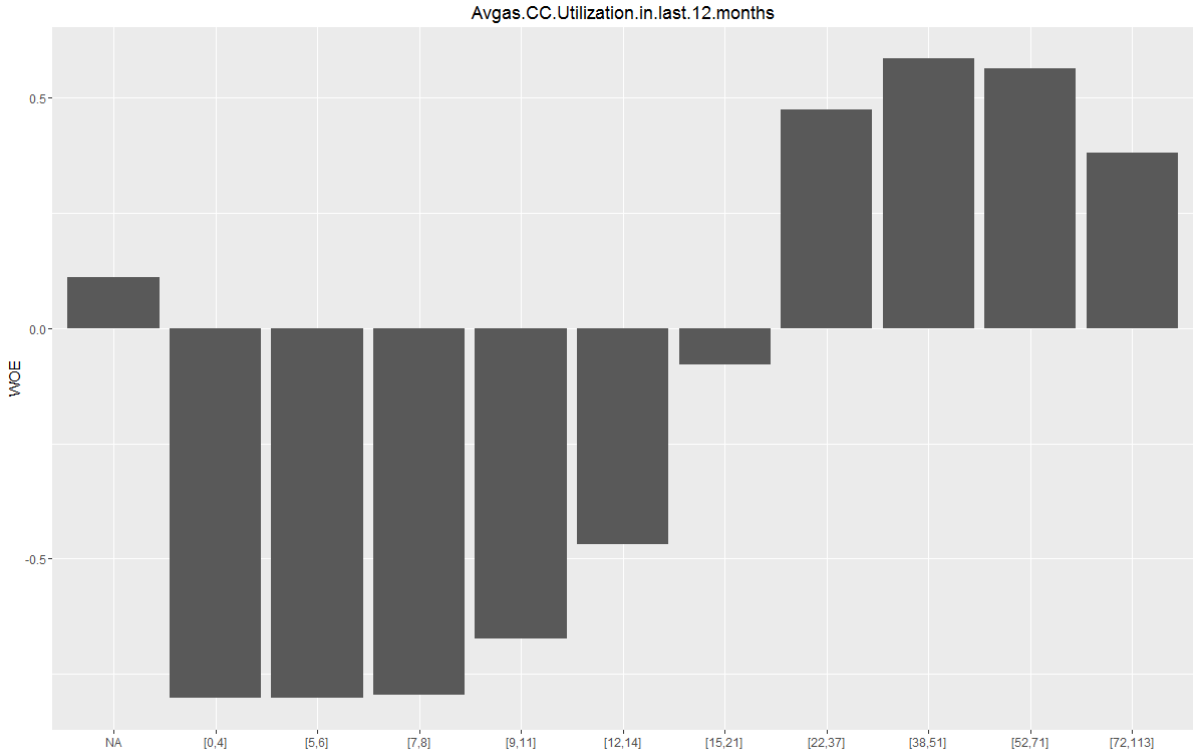


16. No of times 30 DPD or worse in last 12 months

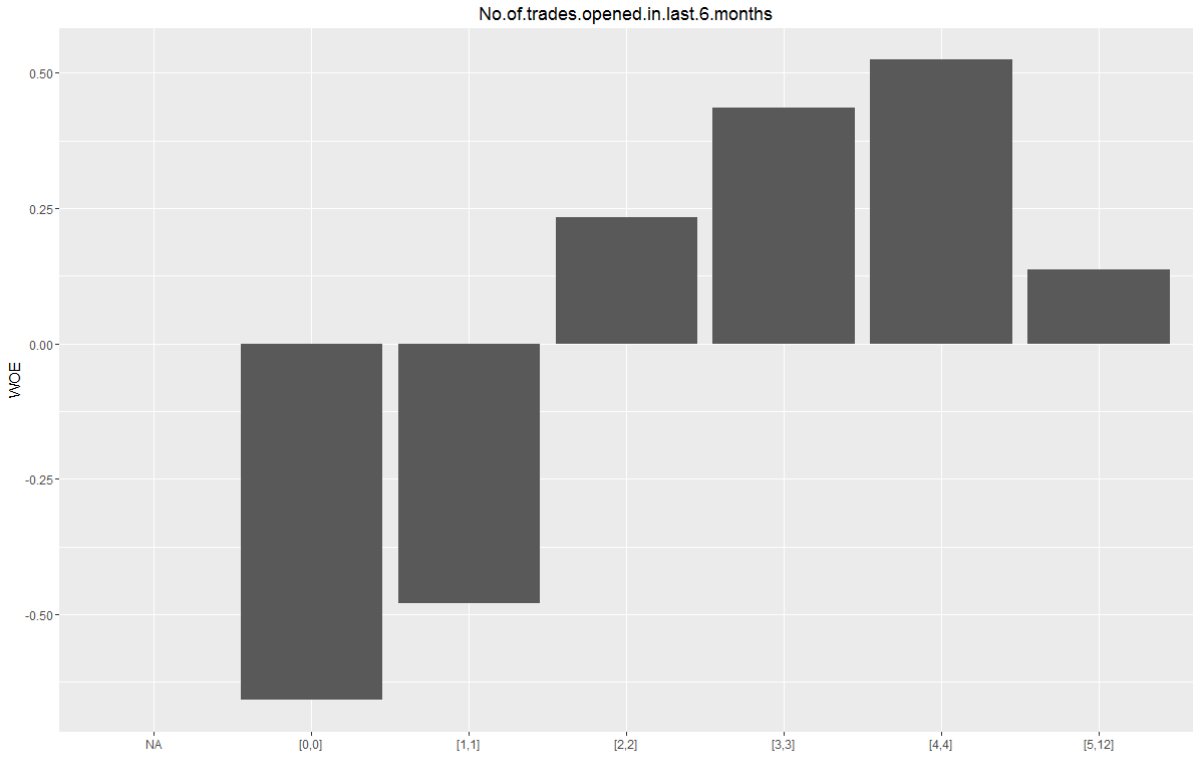


17. Avgas CC Utilization in last 12 months: 1058 NAs. This is the most important field holding the most Information content

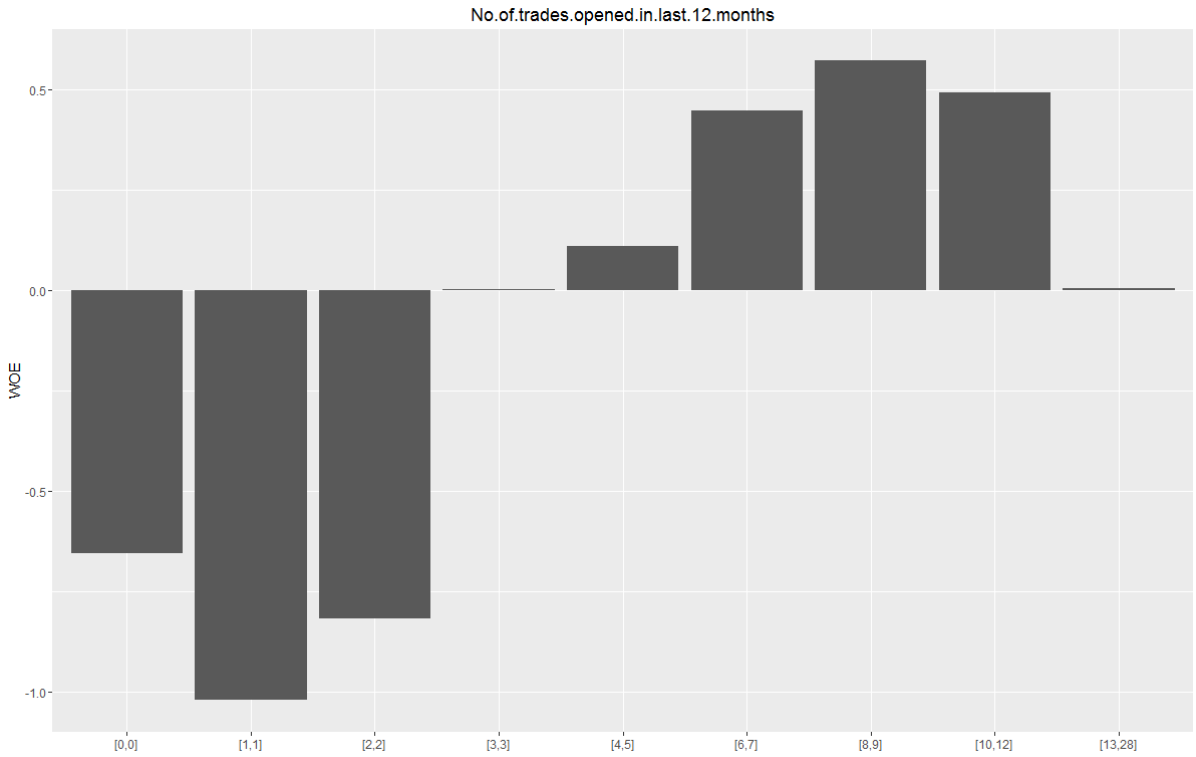
Variable	IV
Avgas.CC.Utilization.in.last.12.months	0.30993640495
No.of.trades.opened.in.last.12.months	0.29795710779
No.of.PL.trades.opened.in.last.12.months	0.29589547357
No.of.Inquiries.in.last.12.months..excluding.home...aut...	0.29542430724
Outstanding.Balance	0.24626922759
No.of.times.30.DPD.or.worse.in.last.6.months	0.24156273923
Total.No.of.Trades	0.23660492197
No.of.PL.trades.opened.in.last.6.months	0.21970498073
No.of.times.90.DPD.or.worse.in.last.12.months	0.21387483771
No.of.times.60.DPD.or.worse.in.last.6.months	0.20583387648
No.of.Inquiries.in.last.6.months..excluding.home...auto...	0.20518701285
No.of.times.30.DPD.or.worse.in.last.12.months	0.19825485836
No.of.trades.opened.in.last.6.months	0.18600887061
No.of.times.60.DPD.or.worse.in.last.12.months	0.18549887262
No.of.times.90.DPD.or.worse.in.last.6.months	0.16011692406
No.of.months.in.current.residence	0.07894352677
Income	0.04241780038
No.of.months.in.current.company	0.02175441278
Presence.of.open.home.loan	0.01762652922
Age	0.00334915723
No.of.dependents	0.00265195588
Profession	0.00221742766
Presence.of.open.auto.loan	0.00165481980
Type.of.residence	0.00104072364
Education	0.00055652497
Gender	0.00032497070
Marital.Status..at.the.time.of.application.	0.00009519639



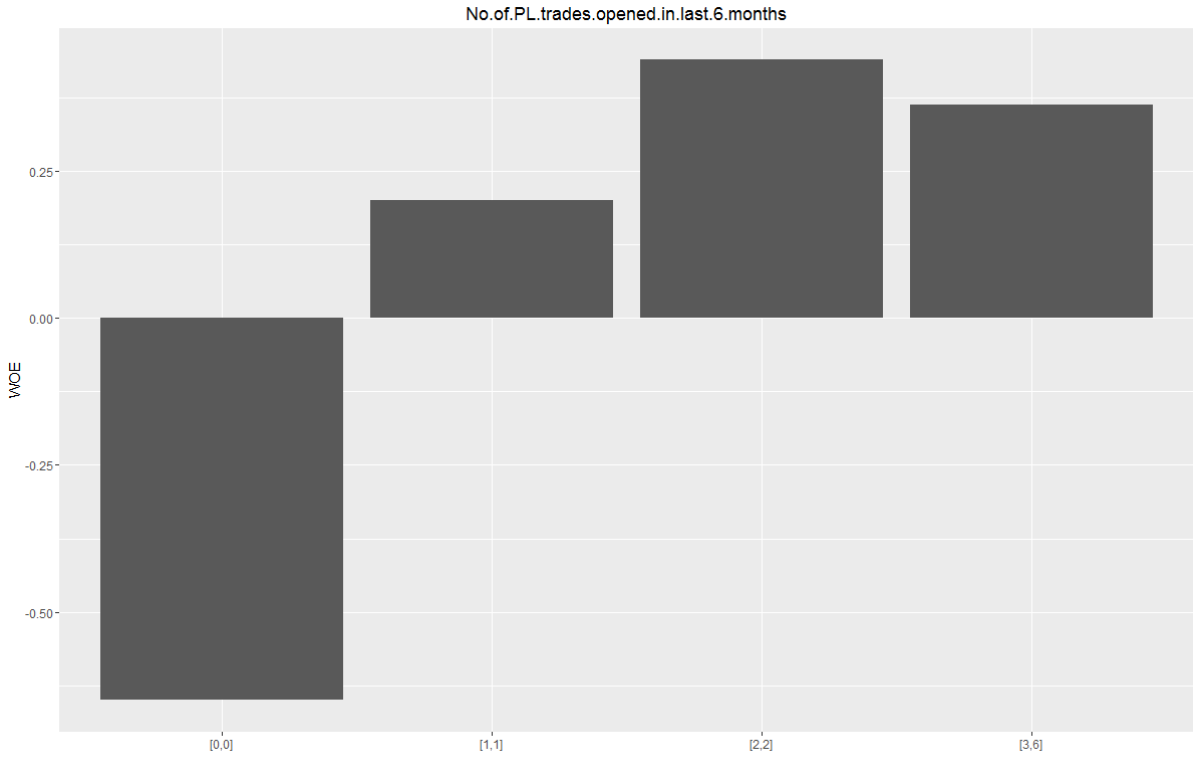
18. No of trades opened in last 6 months: 1 NA.



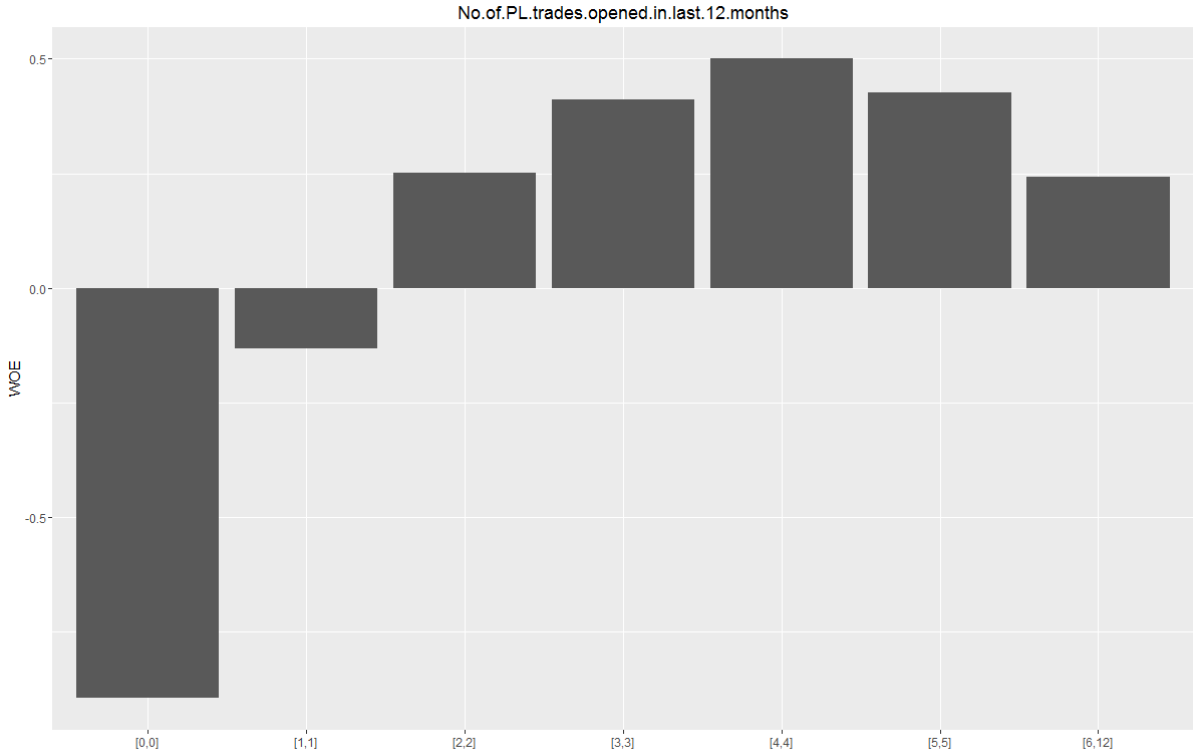
19. No of trades opened in last 12 months:



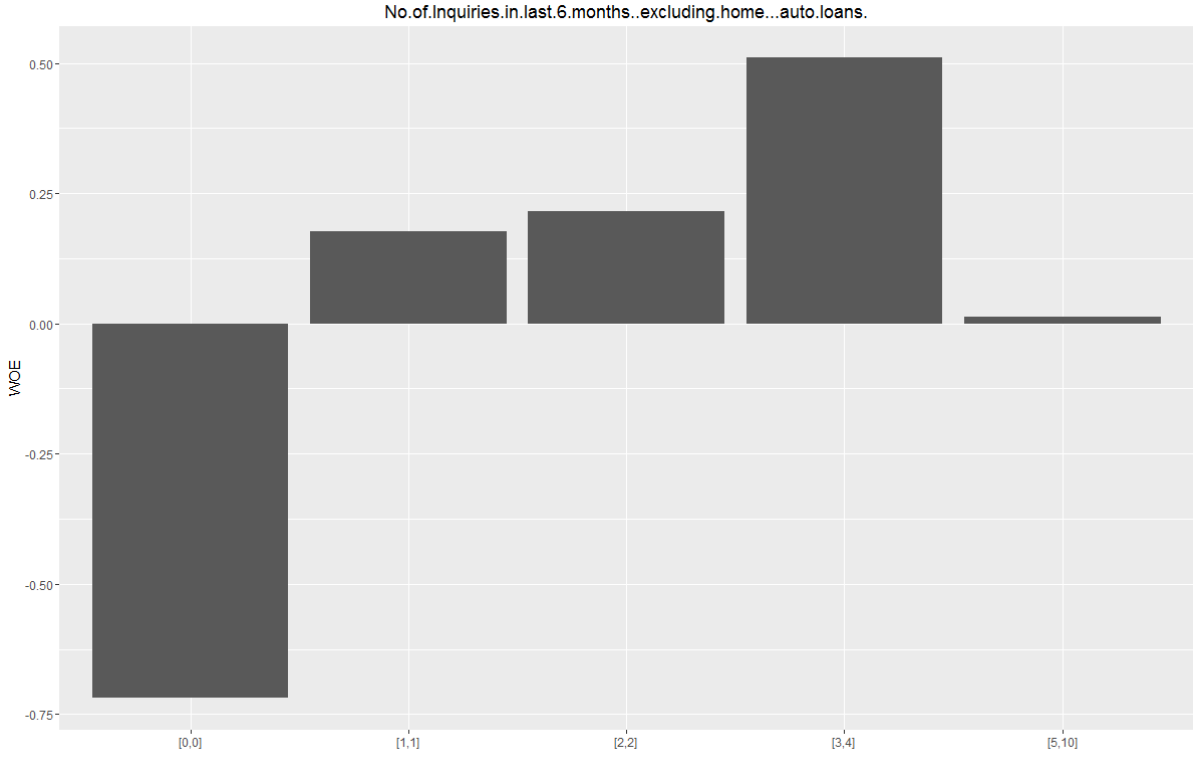
20. No of PL trades opened in last 6 months:



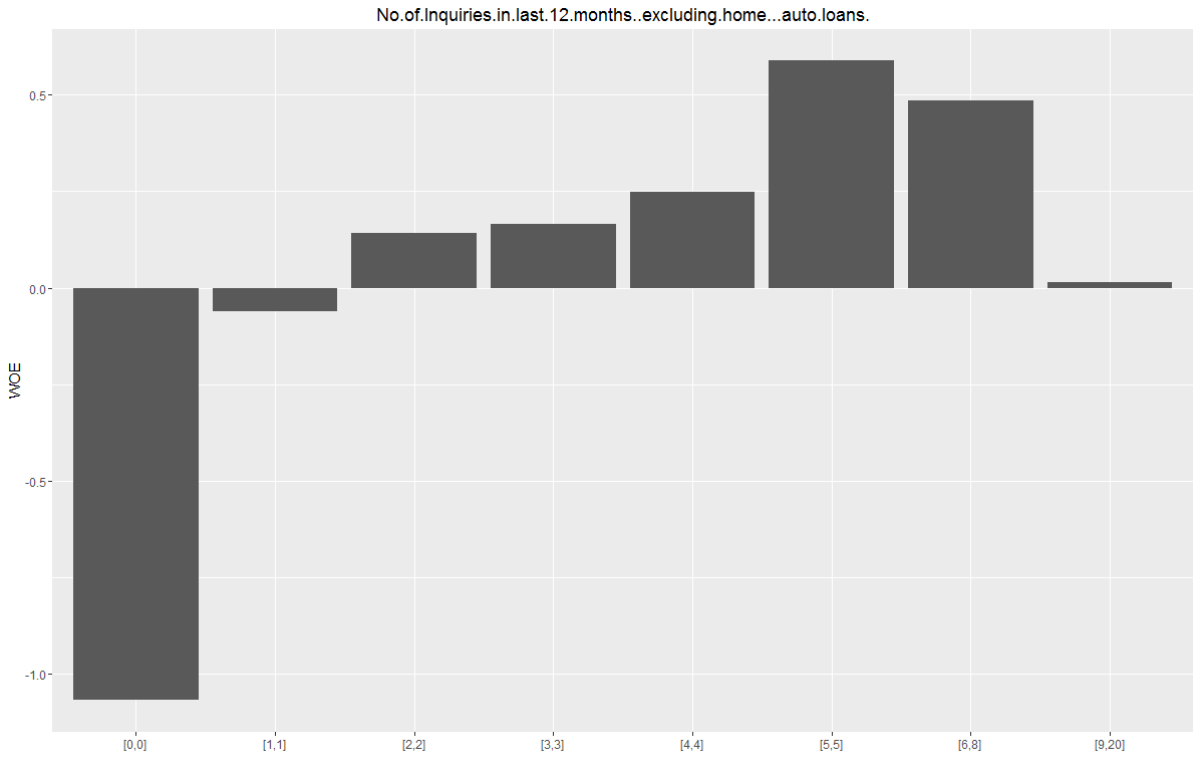
21. No of PL trades opened in last 12 months



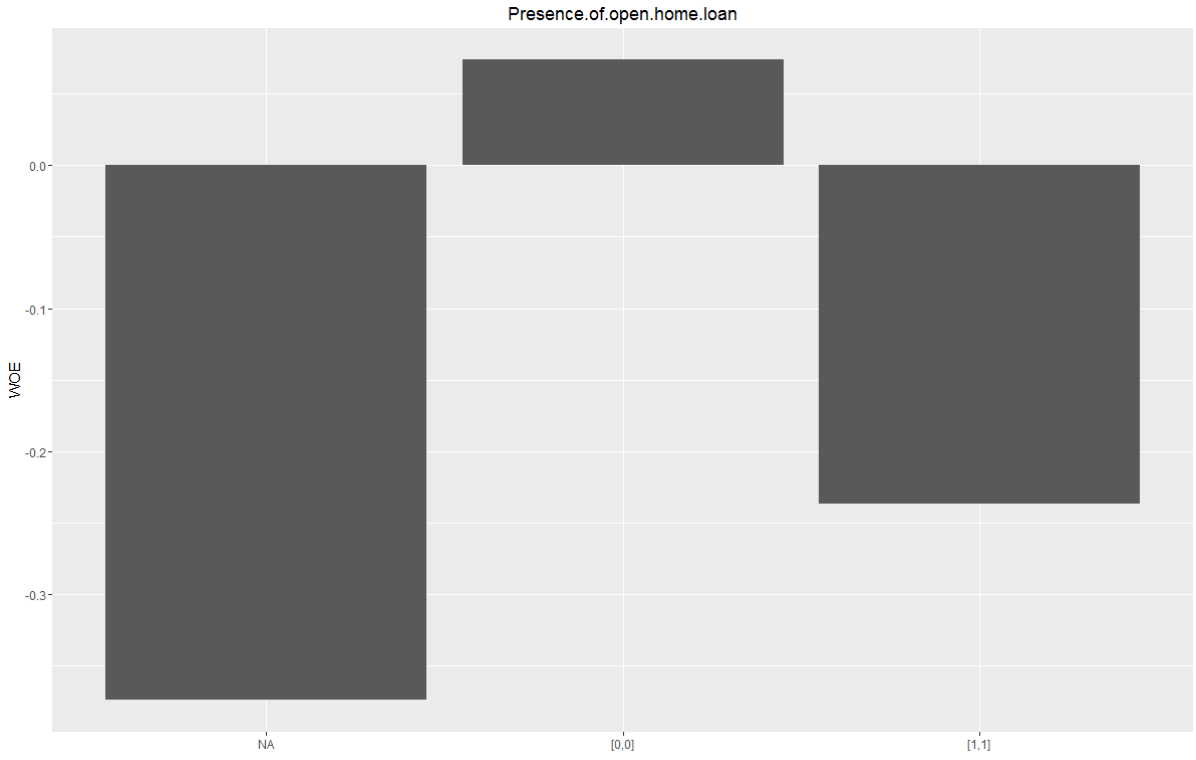
22. No of Inquiries in last 6 months (excluding home & auto loans):



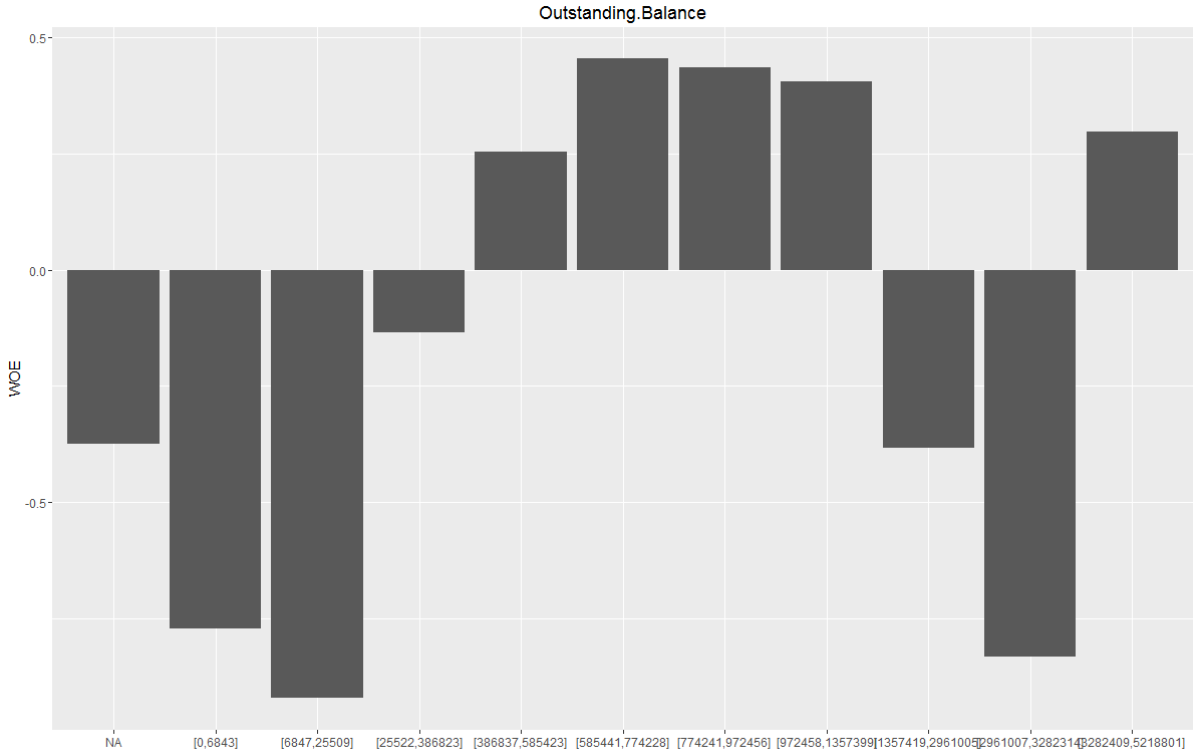
23. No of Inquiries in last 12 months (excluding home & auto loans):



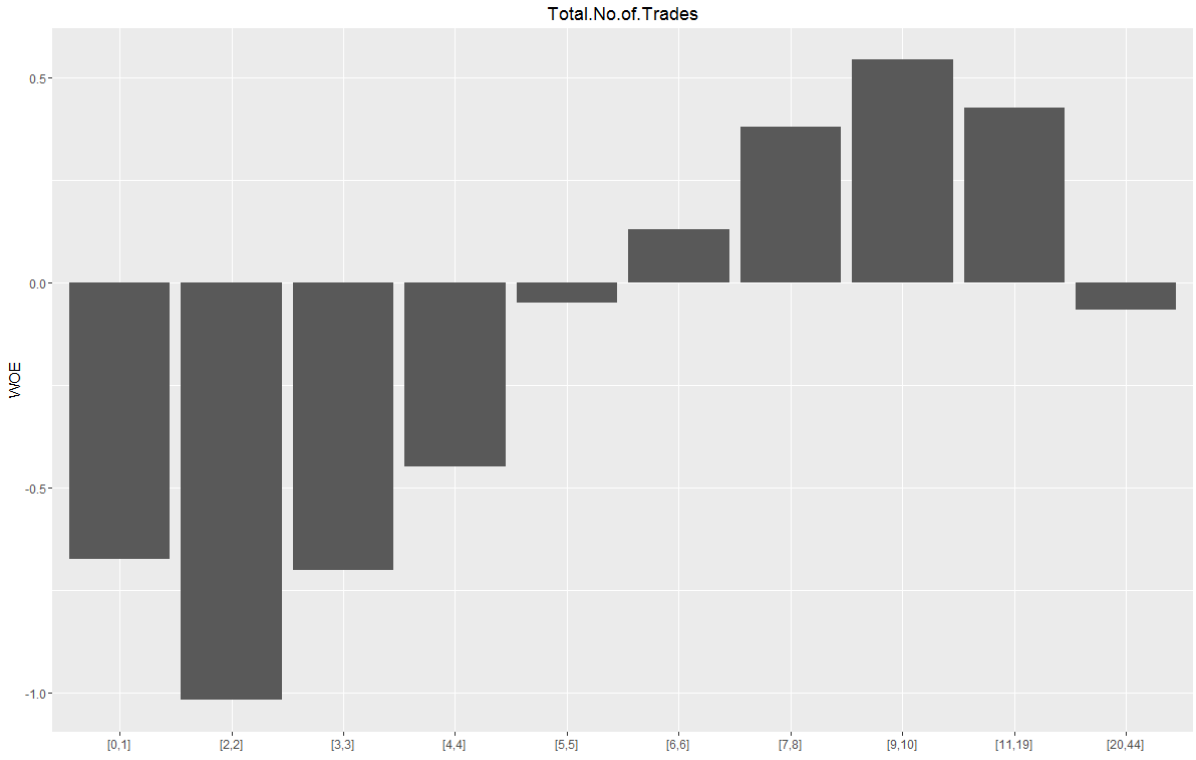
24. Presence of open home loan: 272 NAs



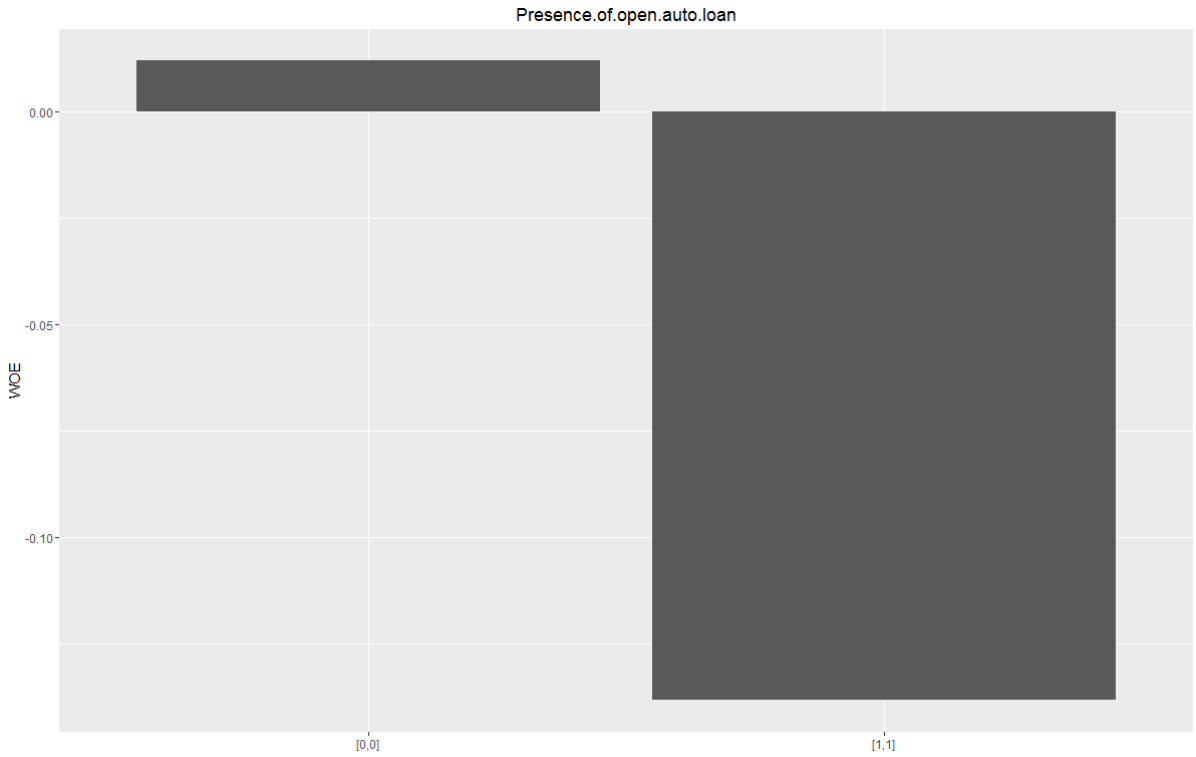
25. Outstanding Balance: 272 NAs



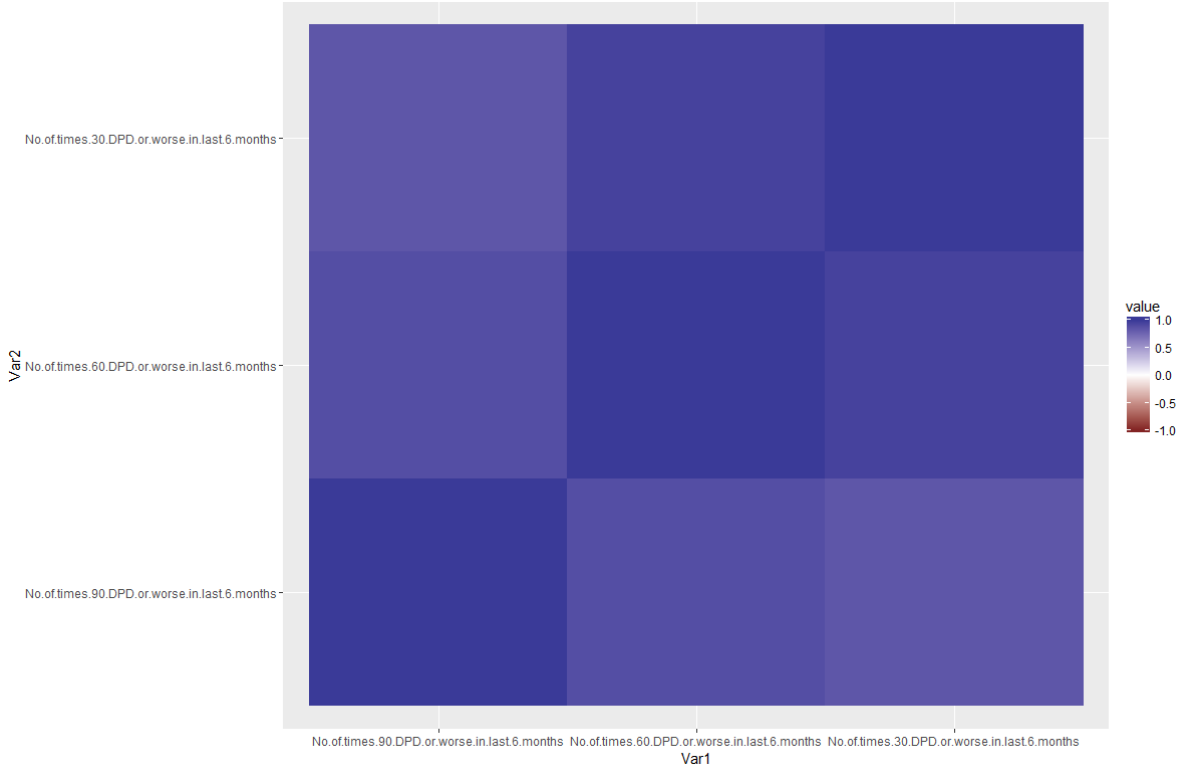
26. Total No of Trades



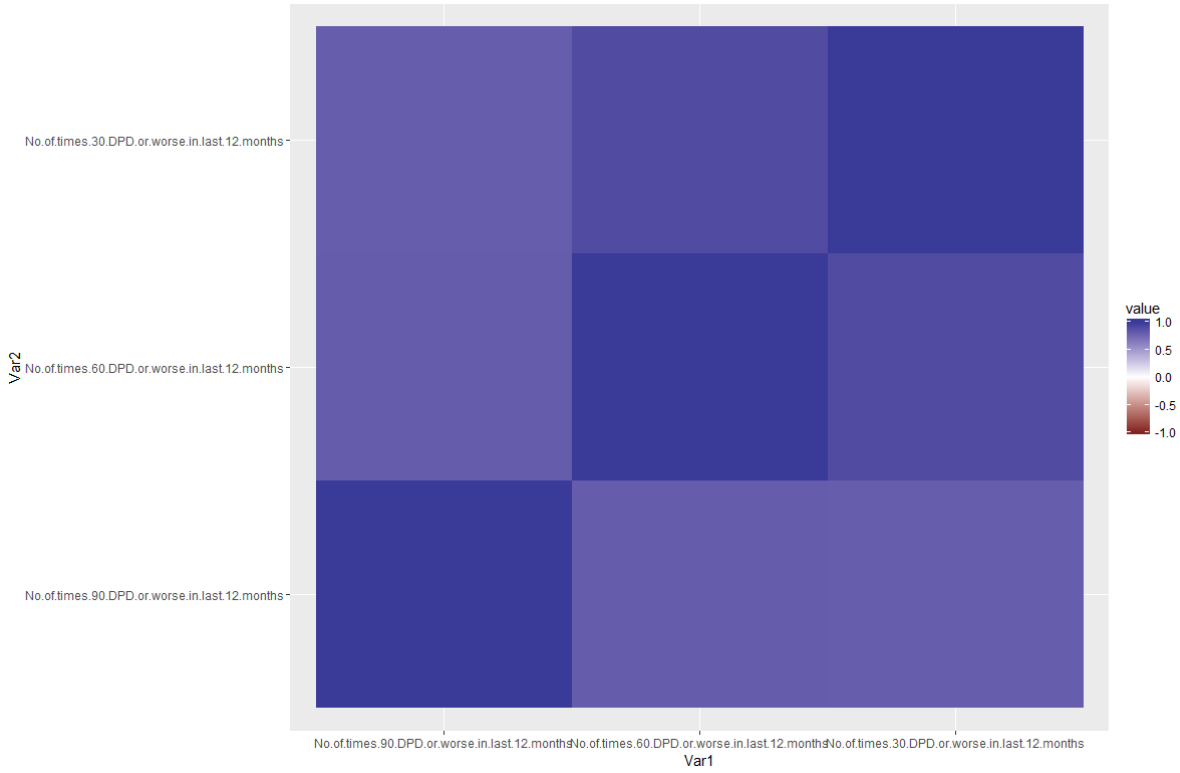
27. Presence of open auto loan:


Presence.of.open.auto.loan
0.00
-0.05
-0.10
WOE
[0,0]
[1,1]

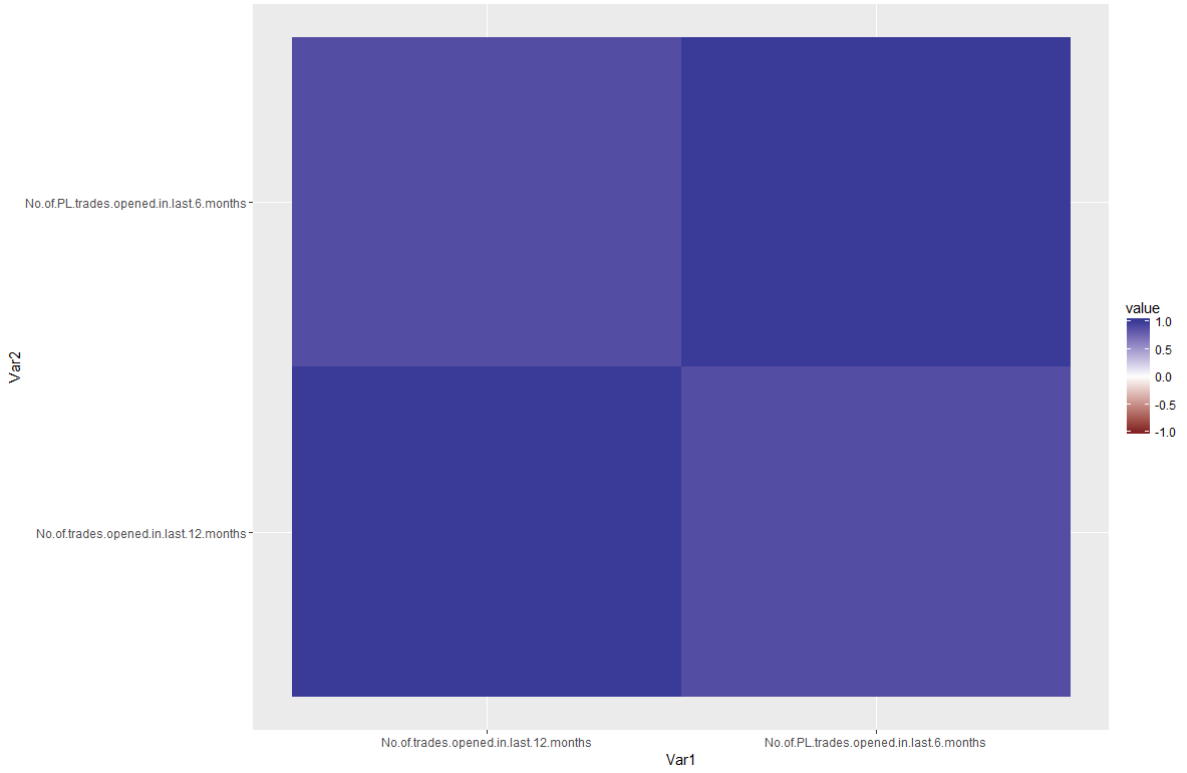
Correlation between No of times 90/60/30 DPD or worse in last 6 months: High



Correlation between No of times 90/60/30 DPD or worse in last 12 months: High



Correlation between No of PL trades opened in last 6/12 months: High



Insights from EDA and further approach

1. We observed that the Information value of the Demographic variables is less than Credit Bureau Variables. Just for checking we ran logistic regression on the original data set after cleaning and found out that none of the Demographic variable is significant.
2. We came to conclusion that imputing of NAs with Mean/Mode/Median is not required in this case and found out that WOE values can be used in that place. Fo rthis purpose, We will mostly use Fine Classing, and only in the case where we can replace NAs with “Other” category, we will use coarse classing technique.
3. Further, scaling or outlier tretment will not be required as we will be running the model on the separate data frame containing the WOE values for all variables {categorical, numeric}
4. Application ID will not be part of model, as it represents a series/row number.
5. Further we will make two data frames, one for model building with Demograpic variables and other with both Demographic and Credit Bureau variable.
6. We will split the data into train and test in 70:30 ratio
7. With the final equation of Logistic Regression model, we should be able to detremine the factors affecting the risk involved in acquiring the customers.
8. Based on those factors, we can try creating startegies to mitigate the risk. We can even see if require more data or need to derive more data variables out of the existing ones.
9. For evaluating our Logistic Regression model, we will try to see metrics like C-statistic, KS-statistic, Accuracy, Sensitivity, Specificity, AUC. Will try to find the optimum threshold value as per the confusion matrix.
10. Creation of Application Scorecard and accessing the financial benefit of our project will be done after our model satifies all the business constraints.