BFS Capstone Project

MID-SUBMISSION

Group Members:

1. Chayan Naskar
2. Avanish Kumar
3. Asa Singh

Business Understanding:

- CredX is a leading credit card provider that receives thousands of credit card applications every year.
- The CEO believes that the best strategy to mitigate credit risk is to 'acquire the right customers'

Problem Statement:

Help CredX identify the right customers using predictive models.
Using past data of the bank's applicants,

- Determine the factors affecting credit risk.
- Create strategies to mitigate the acquisition risk.
- Assess the financial benefit of your project.

Analysis Approach:

- Data Understanding and Cleaning
- Univariate Analysis
- Segmented Univariate and Bivariate Analysis
- Identification of important predictor variables
- Plotting of Categorical Variable, Continuous Variables and Correlation between the Variables.
- Build the Logistic Regression Model to predict the likelihood of default.
- Evaluation of the model in predicting the likeliness of default.
- Providing the Financial Benefit of the project.

Data Understanding:

There are two data sets in this project:

- Demographic/application data:

Information provided by the applicants at the time of credit card application like customer-level information on age, gender, income, marital status, no of dependents, education, profession, type of residence, no of months in current residence, no of months in current company, application id and performance tag.

- Credit bureau:

This file contains information provided by the credit bureau and contains variables such as 'number of times 30/60/90 DPD in last 3/6/12 months', 'outstanding balance', 'number of trades opened in last 6/12 months', 'number of PL trades opened in last 6/12 months', ''number of inquiries in last 6/12 months', Avgas CC Utilization in last 12 months, presence of open home loan, outstanding balance, total no of trades, presence of open auto loan, performance tag.

Data Quality Issue in Demographic Dataset:

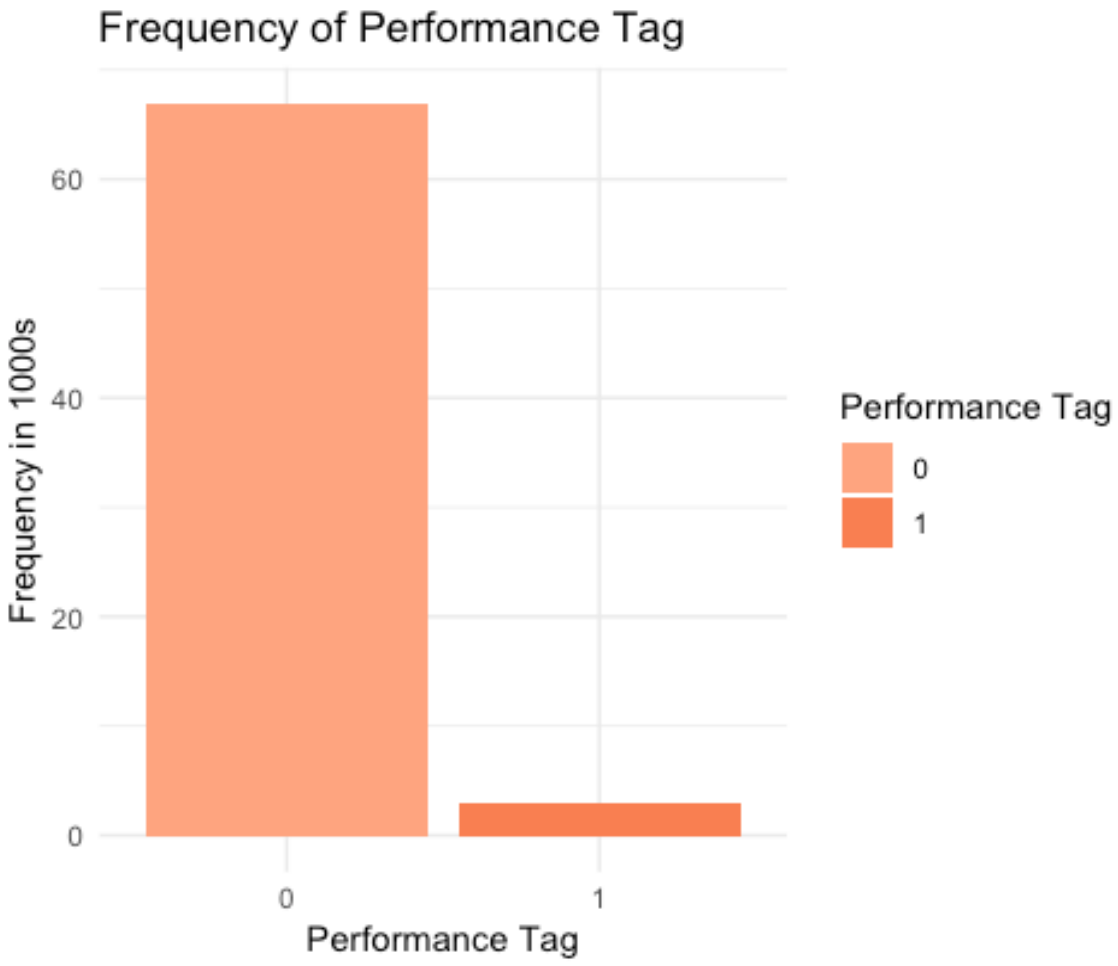
Column Name	Missing Data	Erroneous Data
Age	-	20 rows with value -3 and 0
Gender	2 NA's	-
Marital Status	6 NA's	-
No of Dependents	3 NA's	-
Income	-	81 rows have income less than 0
Education	119 rows are blank	-
Profession	14 rows are blank	-
Type of Residence	8 rows are blank	-
Performance Tag	1425 NA's	-

Data Quality Issue in Credit Dataset:

Column Name	Missing Data	Erroneous Data
Avgas CC Utilization in last 12 months	1058 NA's	-
No of trades opened in last 6 months	1 NA	-
Presence of open home loans	272 NA's	-

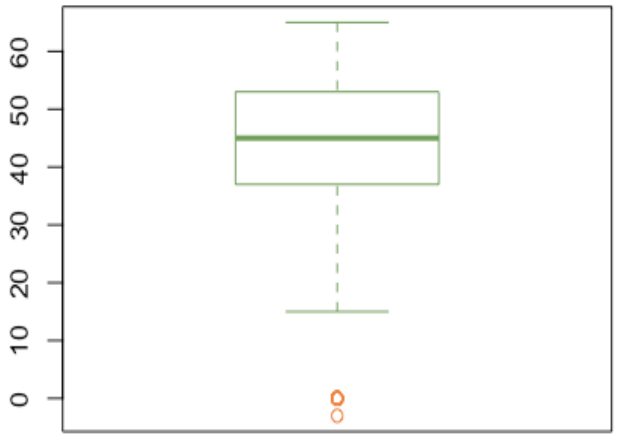
Exploratory Data Analysis:

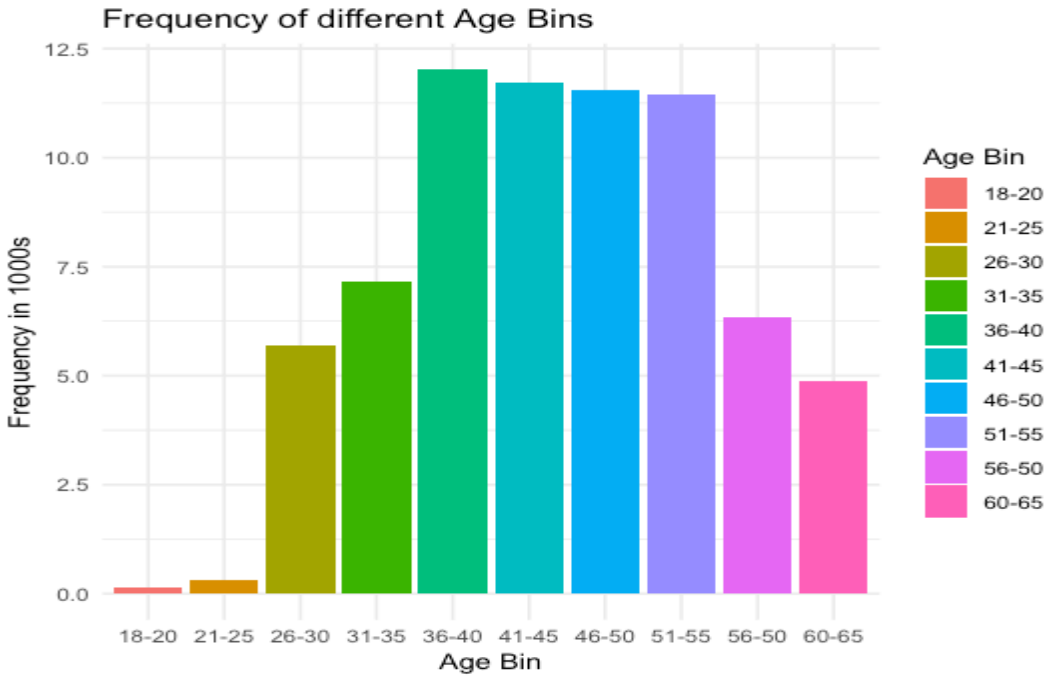
1. **Performance Tag –** The Performance Tag variable has 1425 rows having NA value. These rows are removed for EDA and would be used as dataset for prediction.

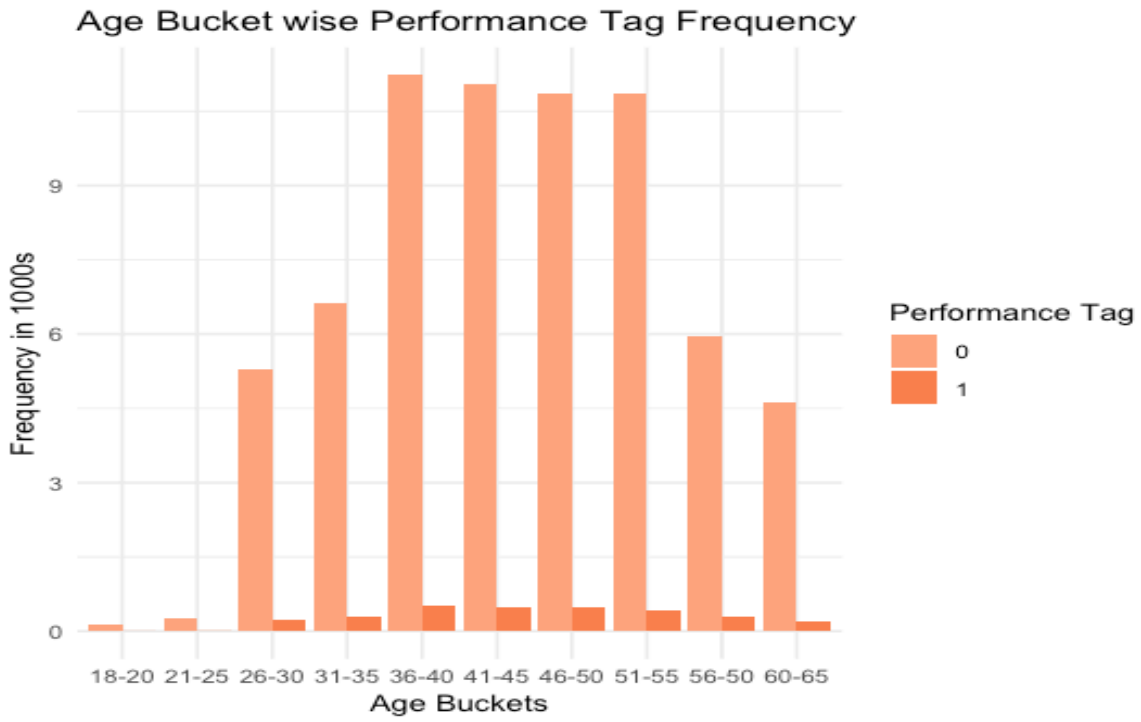


- The Default percentage is 95.78 %

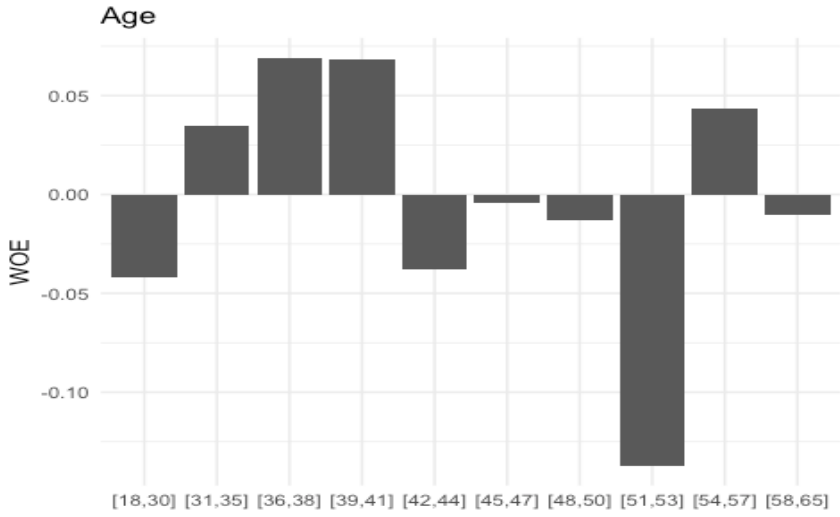
2. **AGE –** The Age variable has minimum age of -3 and some age of 0. Capping the minimum age at 18 since that is the minimum age to get Credit Card.



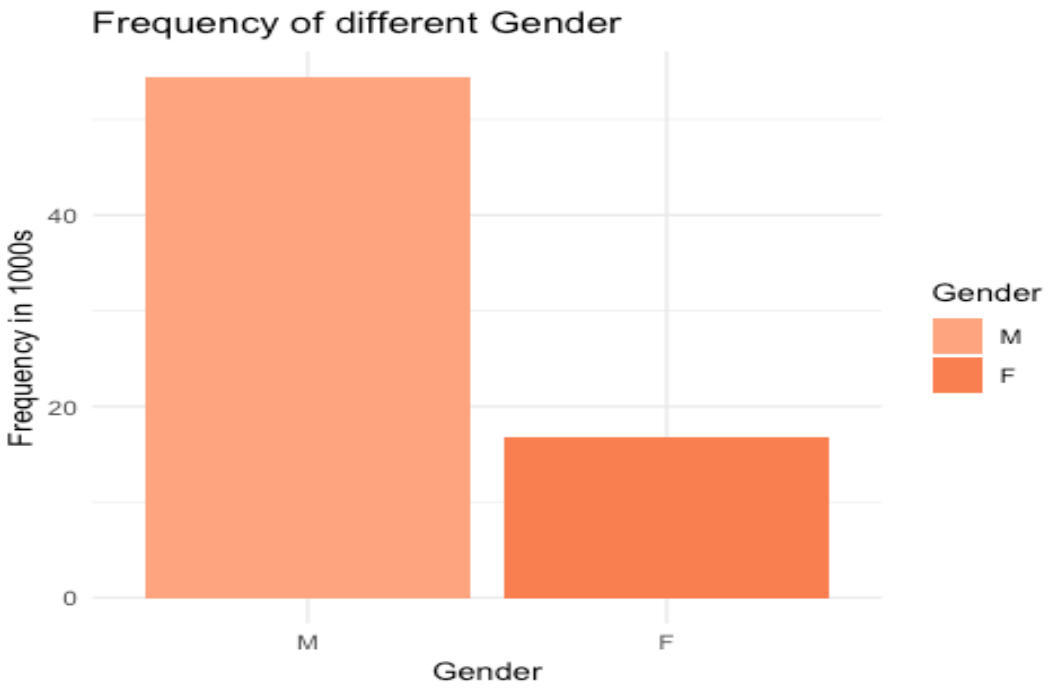




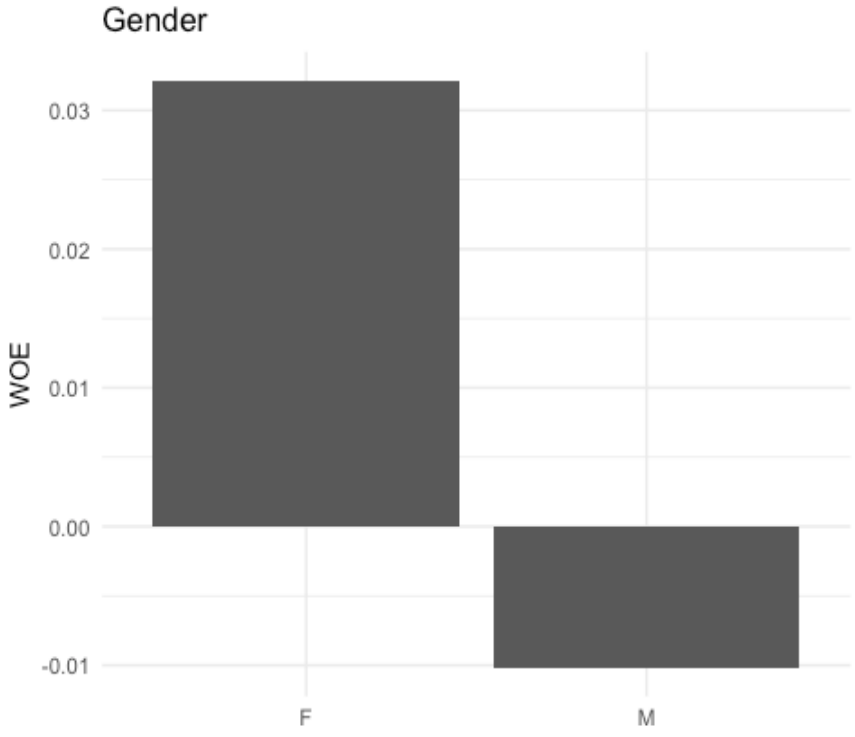
Age	N	Percent	WOE	IV
[18,30]	5946	0.08510821	-0.041635987	0.0001447595
[31,35]	6927	0.09914978	0.034486708	0.0002645613
[36,38]	6924	0.09910684	0.069027071	0.0007519893
[39,41]	7129	0.10204111	0.068252795	0.0012424782
[42,44]	7007	0.10029486	-0.037986486	0.0013847103
[45,47]	6830	0.09776136	-0.004003497	0.0013862744
[48,50]	6743	0.09651609	-0.012674135	0.0014016885
[51,53]	6841	0.09791881	-0.136950261	0.0031273031
[54,57]	7618	0.10904042	0.043497729	0.0033377714
[58,65]	7899	0.11306252	-0.010058241	0.0033491572



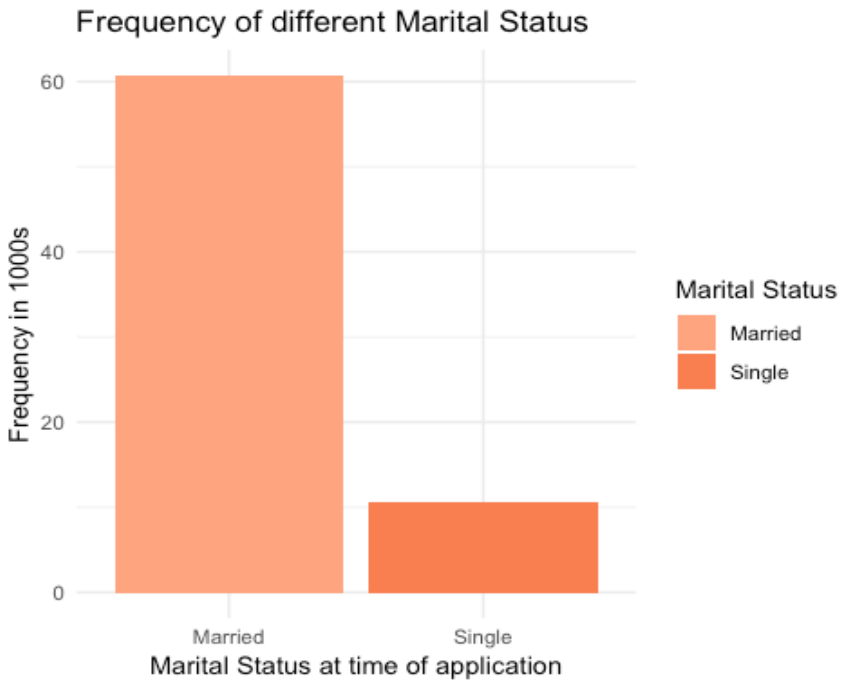
3. **Gender -** The Gender variable has 2 NA's which are replaced with Male since there is highest number of Male.

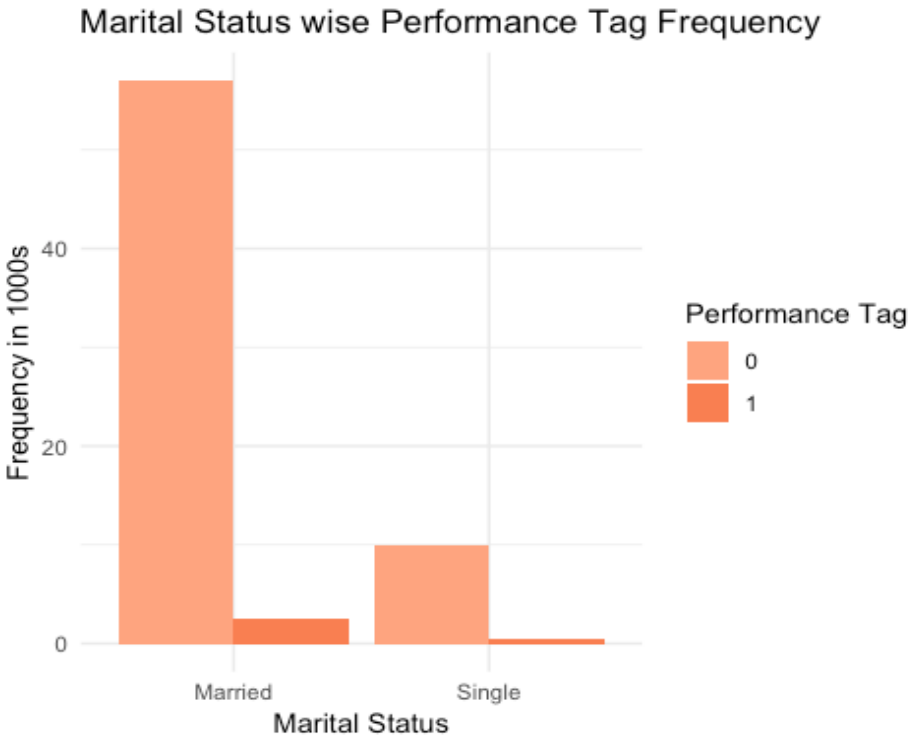


Gender	N	Percent	WOE	IV
F	16506	0.236259	0.03212947	0.0002475104
M	53358	0.763741	-0.01013345	0.0003255737

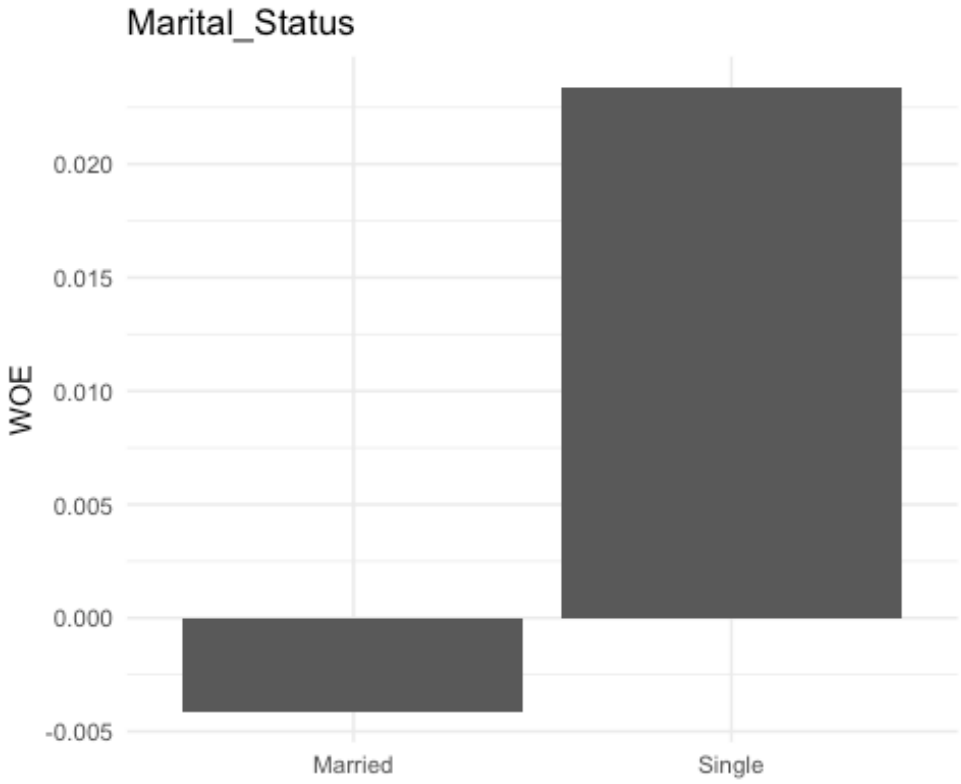


4. **Marital Status –** The Marital Status variable has 6 NA's which are replaced with Married.

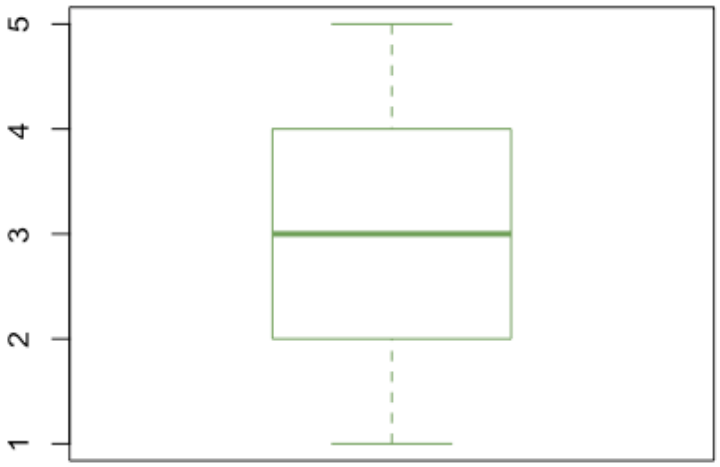


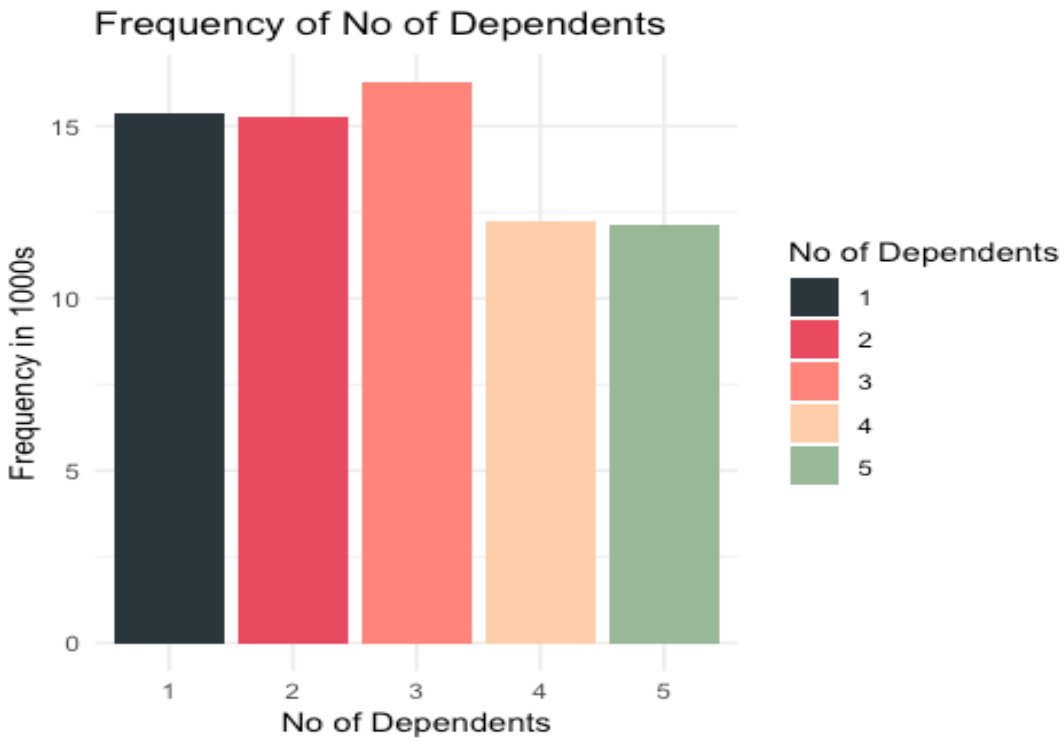


Marital_Status	N	Percent	WOE	IV
Married	59548	0.8523417	-0.004102206	1.431638e-05
Single	10316	0.1476583	0.023383179	9.592186e-05

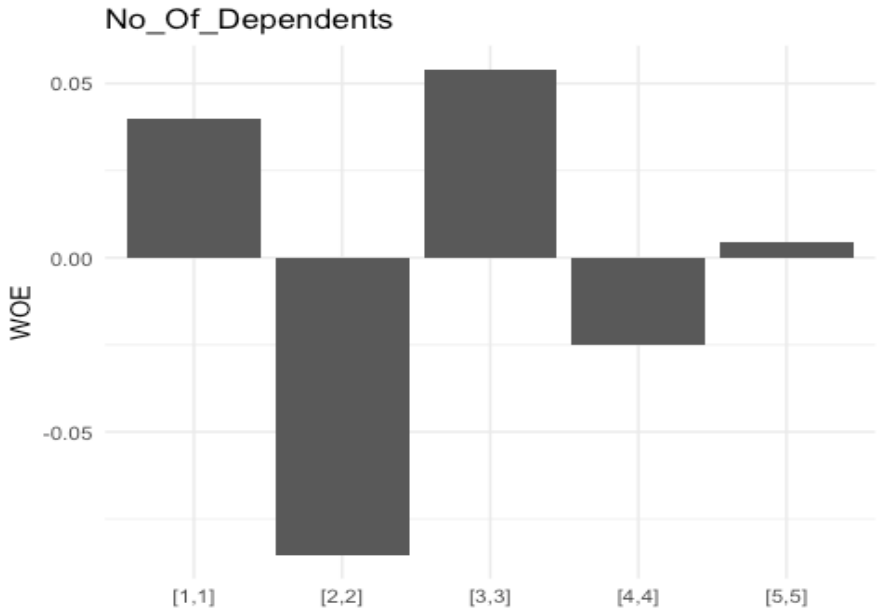


5. **No of Dependents –** The No of Dependents variable has 3 NA's which are replaced with 3.

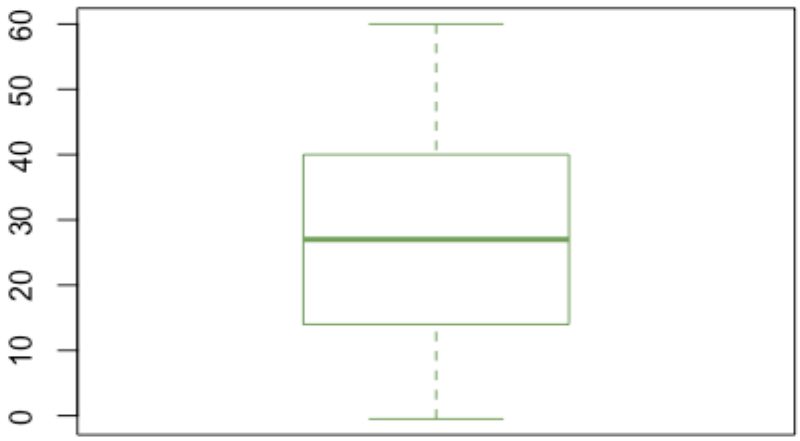


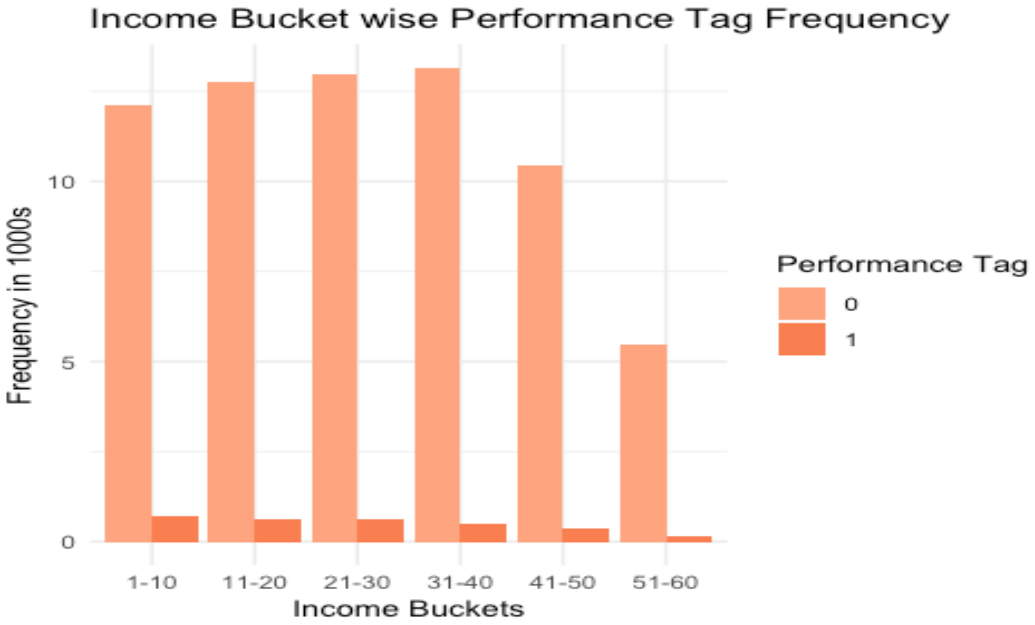


No_Of_Dependents	N	Percent	WOE	IV
1	15218	0.2178232	0.040040389	0.0003556941
2	15127	0.2165207	-0.085197683	0.0018674600
3	15647	0.2239637	0.053976838	0.0025363448
4	11997	0.1717193	-0.025162291	0.0026438235
5	11875	0.1699731	0.004346039	0.0026470404

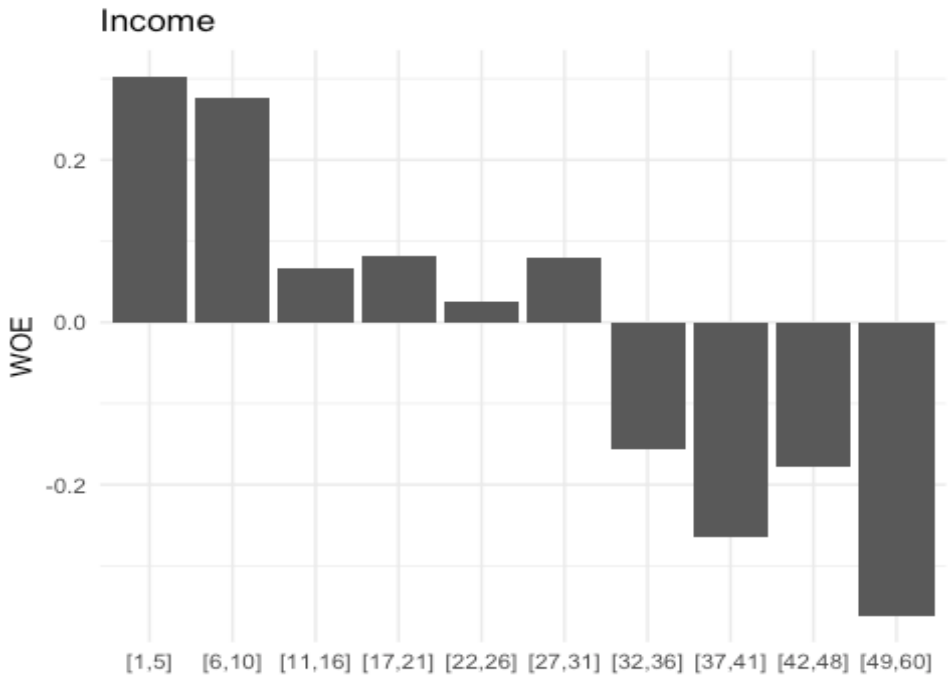


6. **Income -**

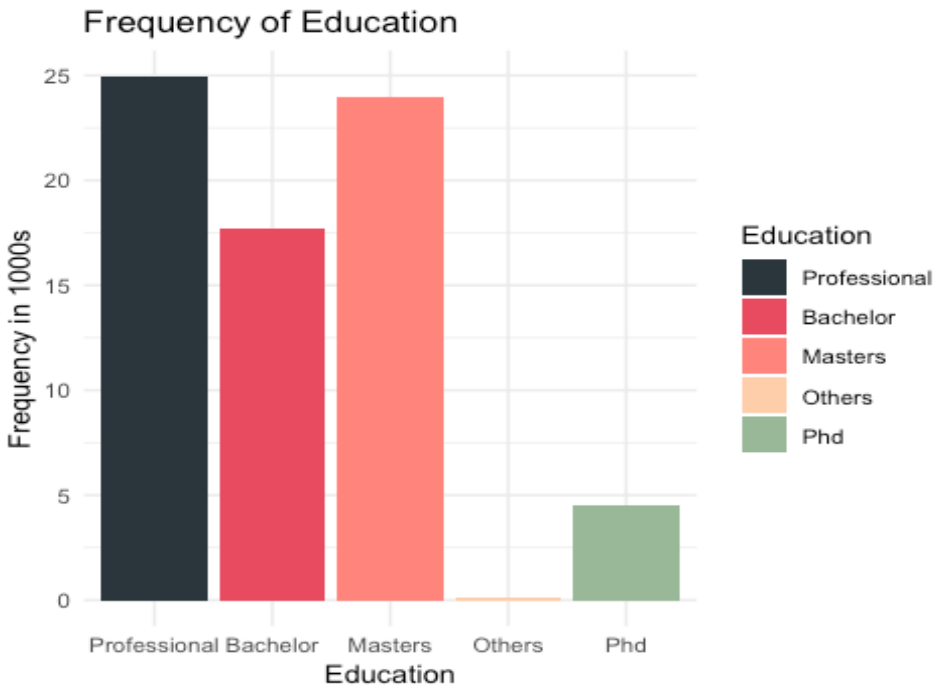

0
10
20
30
40
50
60

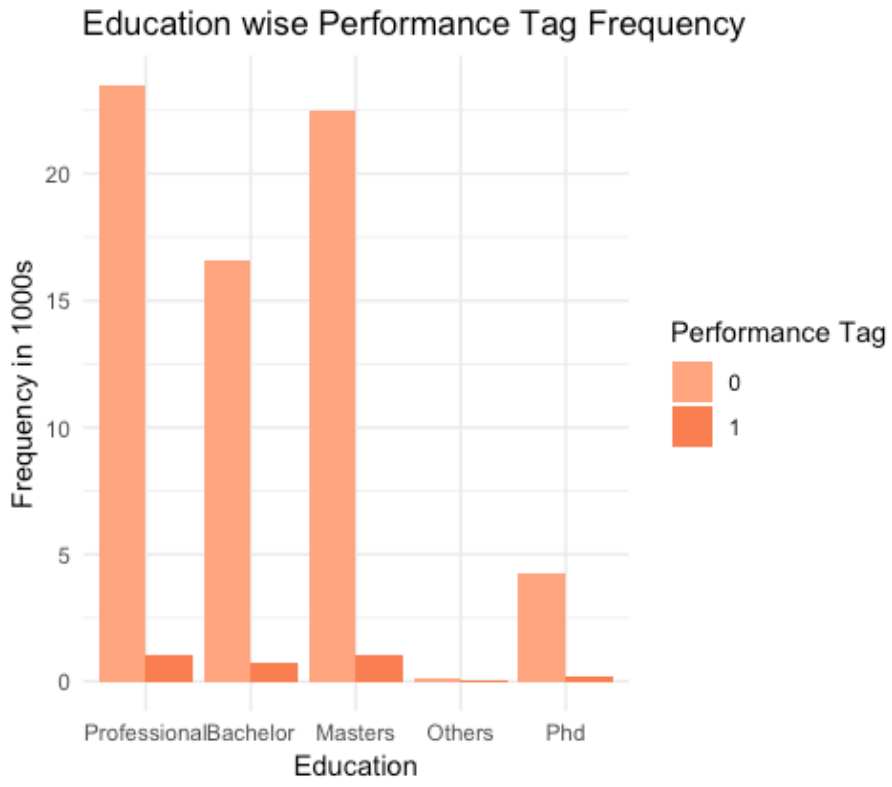

Income Bucket wise Performance Tag Frequency
Frequency in 1000s
0
5
10
1-10
11-20
21-30
31-40
41-50
51-60
Income Buckets
Performance Tag
0
1

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

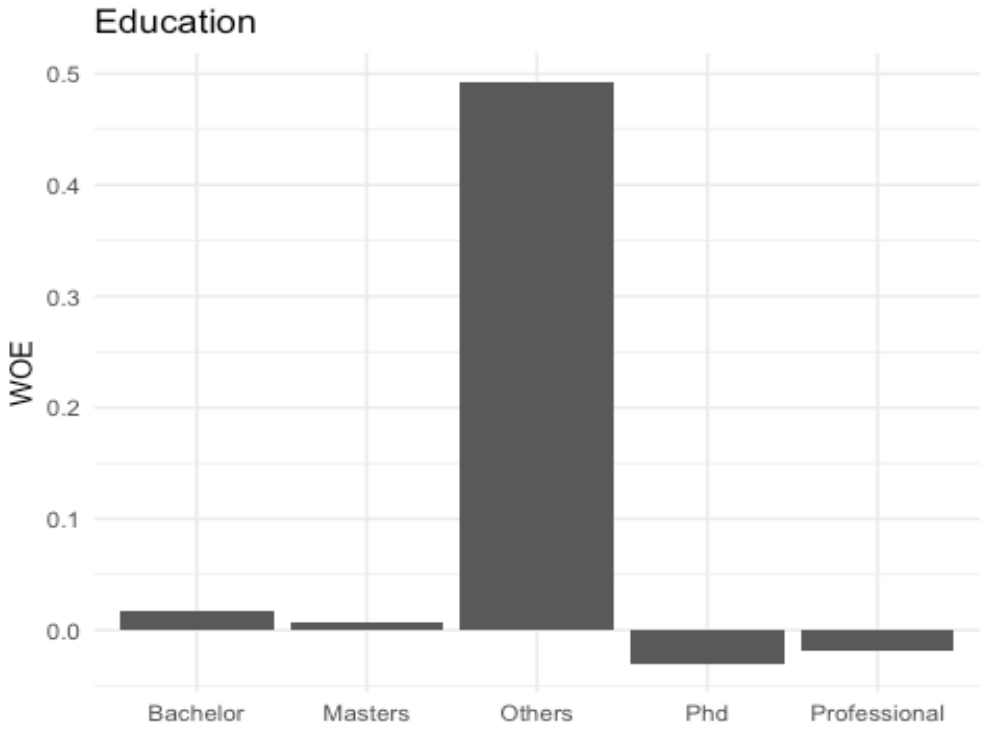


7. **Education –** The blank rows are replaced with “Professional”.

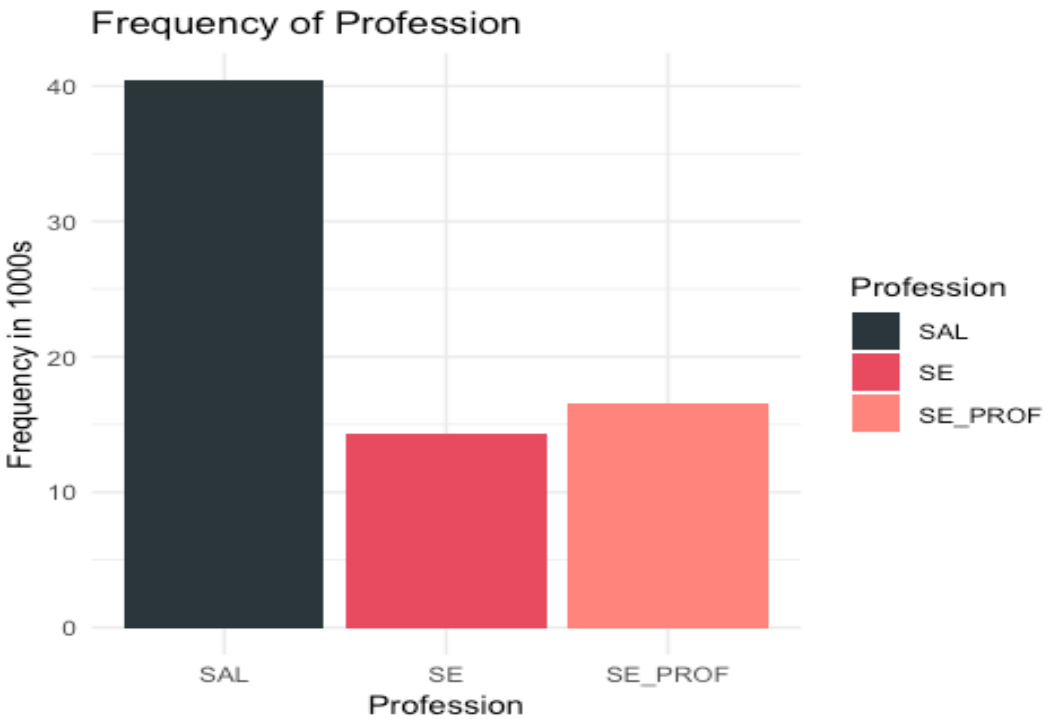


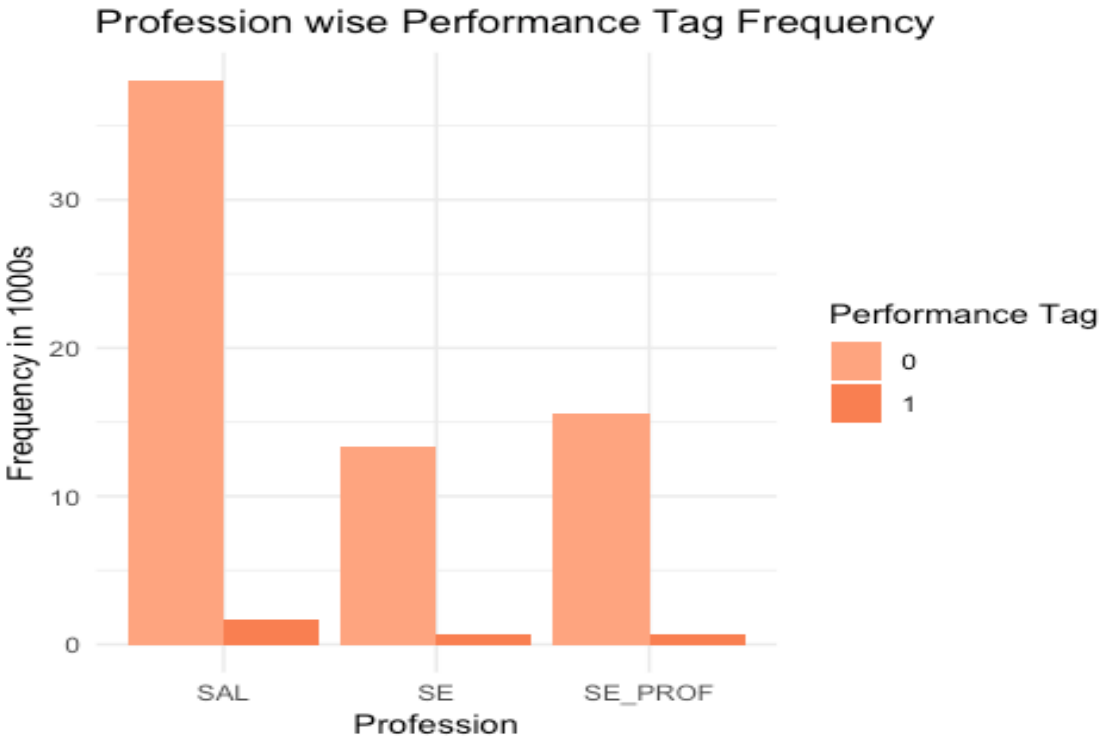


Education	N	Percent	WOE	IV
Bachelor	17300	0.247623955	0.017389937	7.548299e-05
Masters	23481	0.336095843	0.007903871	9.655543e-05
Others	119	0.001703309	0.492576682	6.166444e-04
Phd	4463	0.063881255	-0.029556794	6.717023e-04
Professional	24501	0.350695637	-0.017823229	7.822023e-04

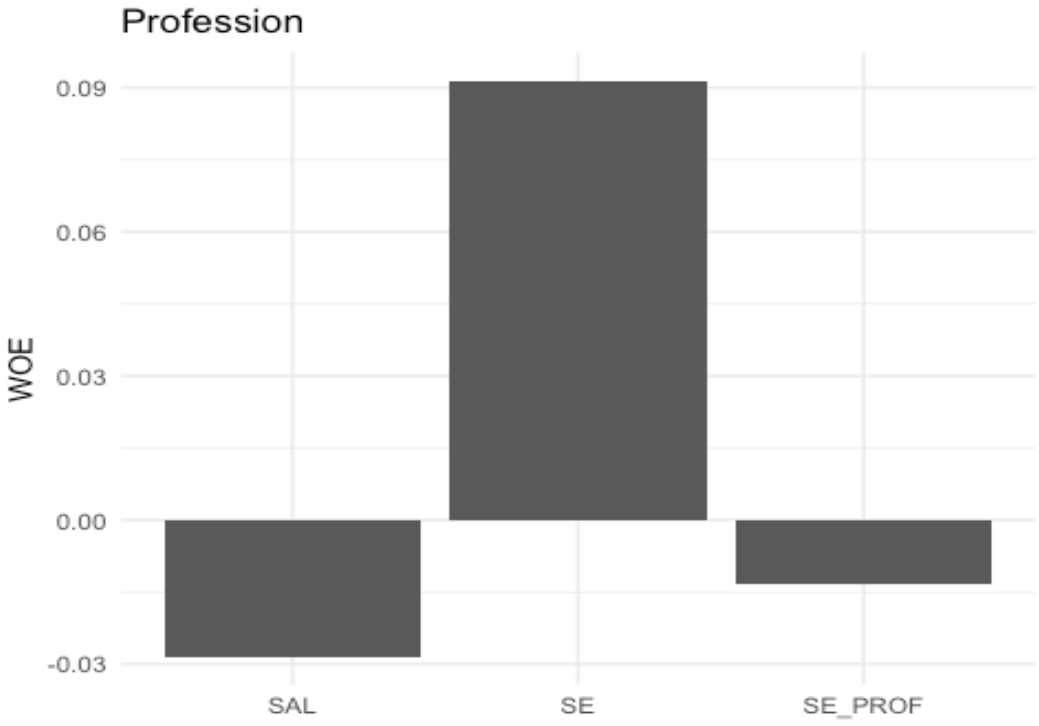


8. **Profession –** The blank rows are replaced with “SAL”.

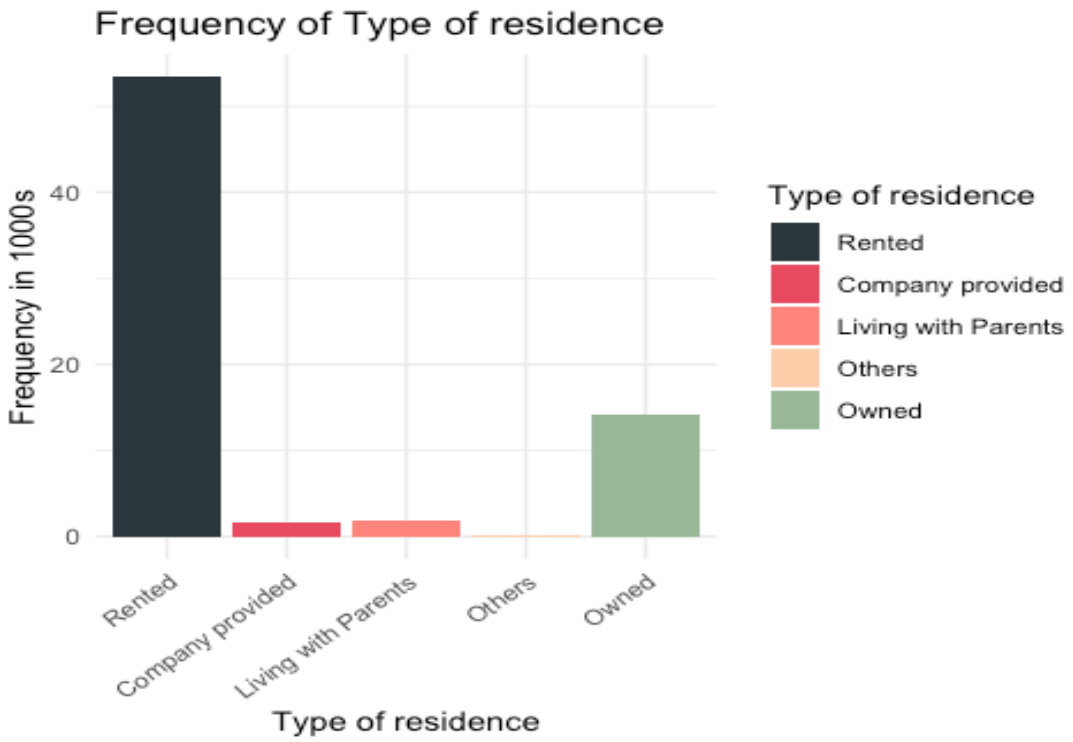


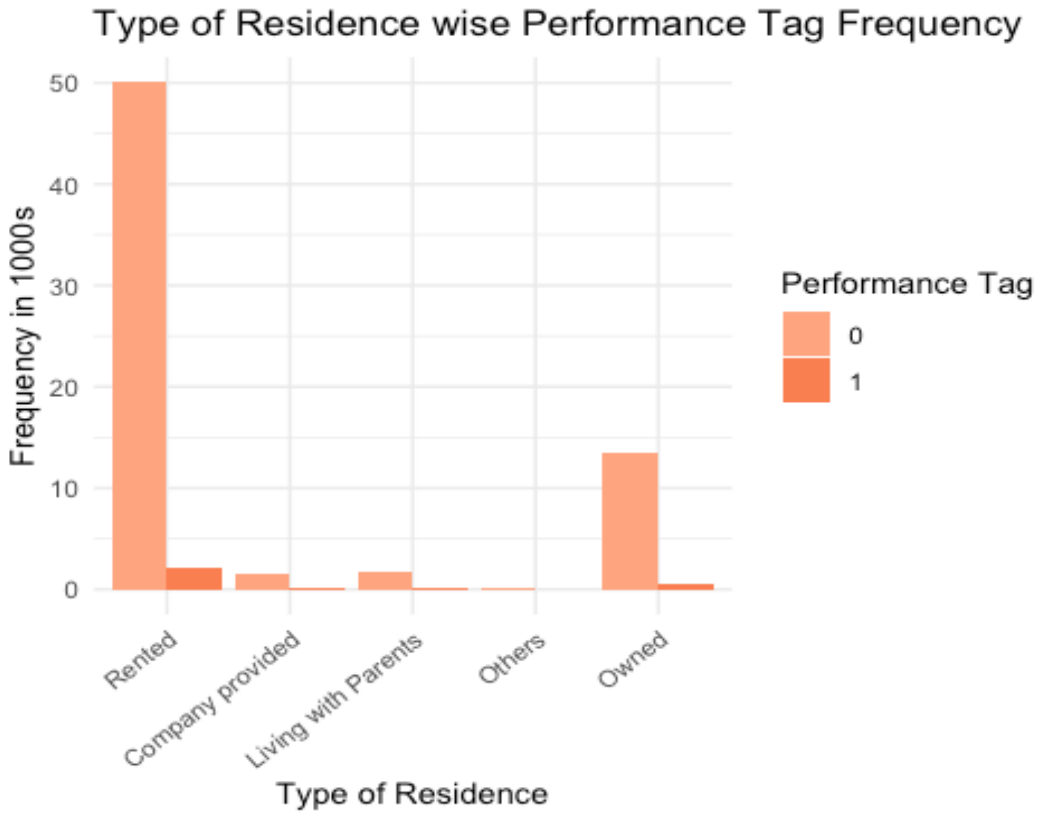


Profession	N	Percent	WOE	IV
SAL	39683	0.5680035	-0.02837453	0.0004514133
SE	13925	0.1993158	0.09137922	0.0021871391
SE_PROF	16256	0.2326806	-0.01334252	0.0022283094

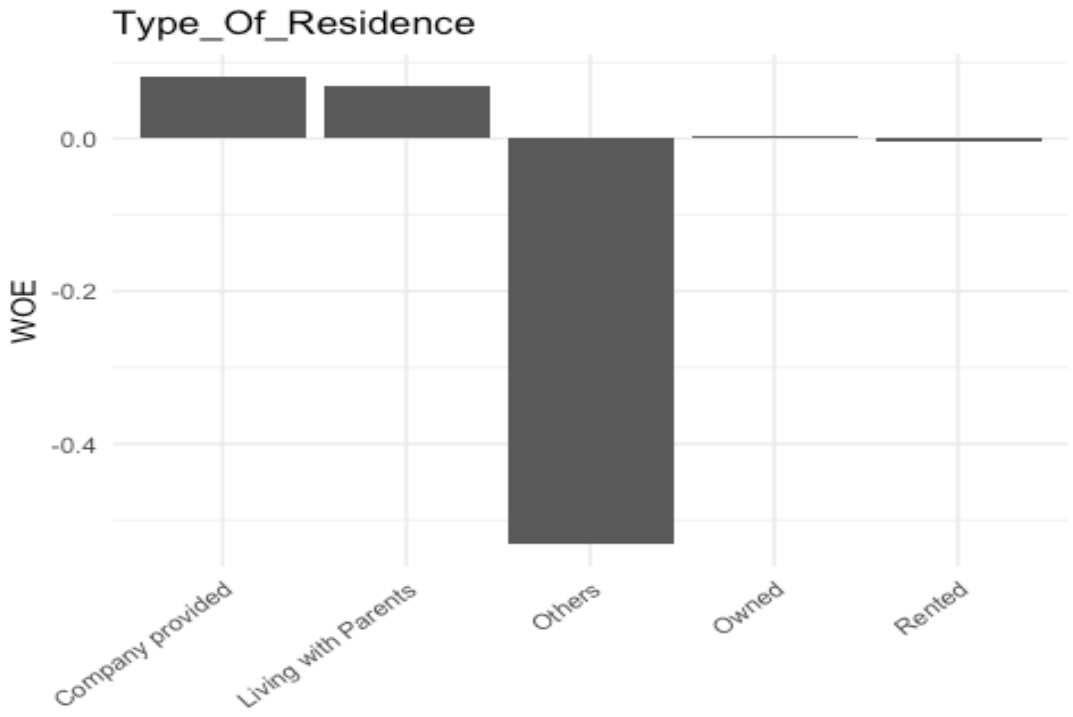


9. **Type of Residence –** The blank rows are replaced with “Rented”.

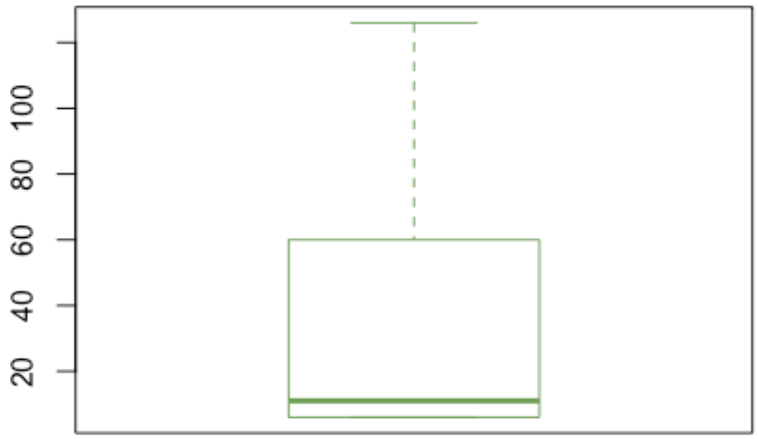


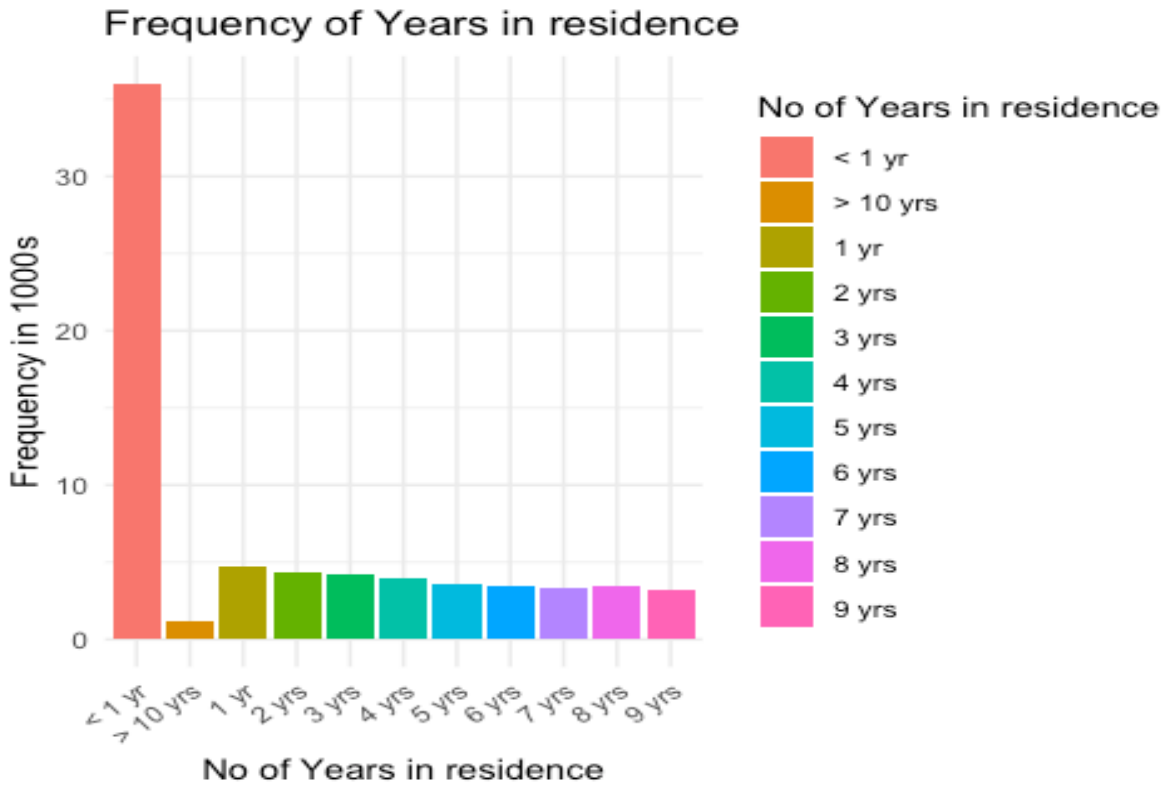


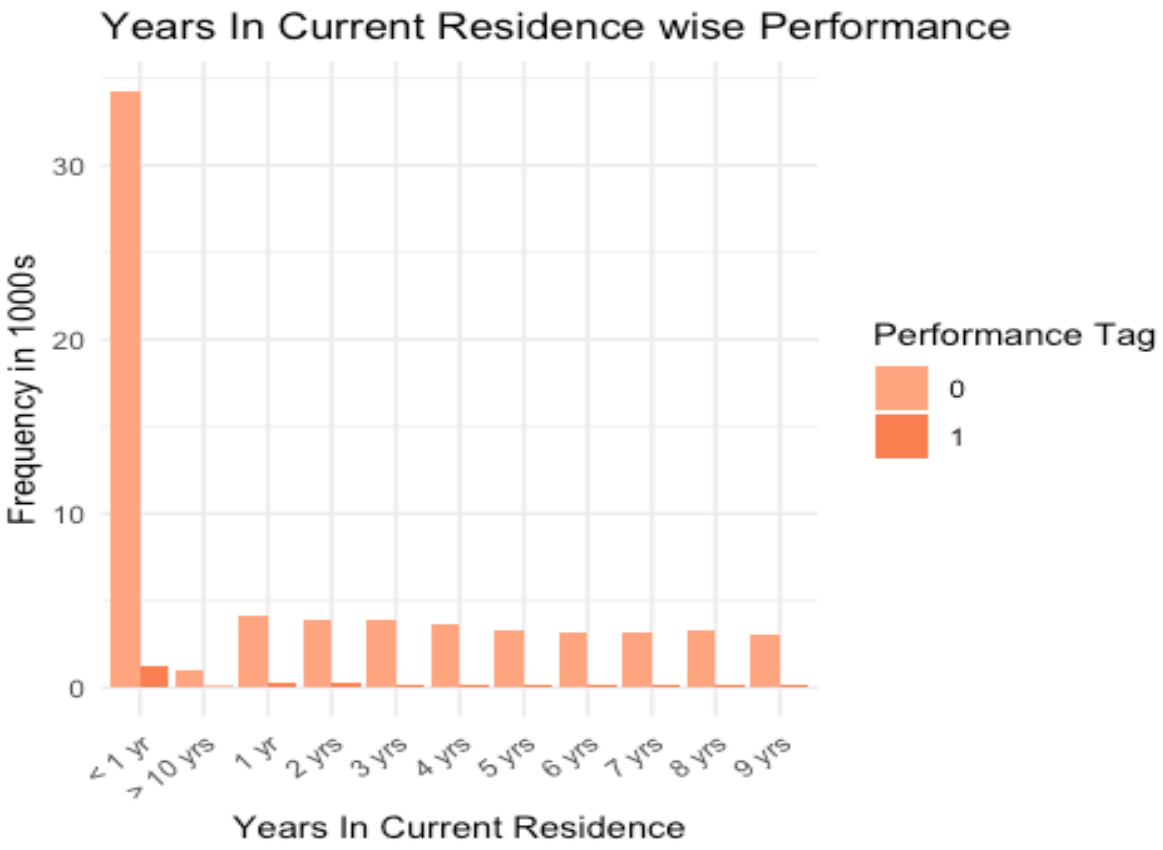
Type_Of_Residence	N	Percent	WOE	IV
Company provided	1602	0.022930265	0.080755577	0.0001551922
Living with Parents	1777	0.025435131	0.068074711	0.0002768061
Others	198	0.002834078	-0.530586935	0.0009068936
Owned	14003	0.200432268	0.004103764	0.0009102754
Rented	52284	0.748368258	-0.004478593	0.0009252553



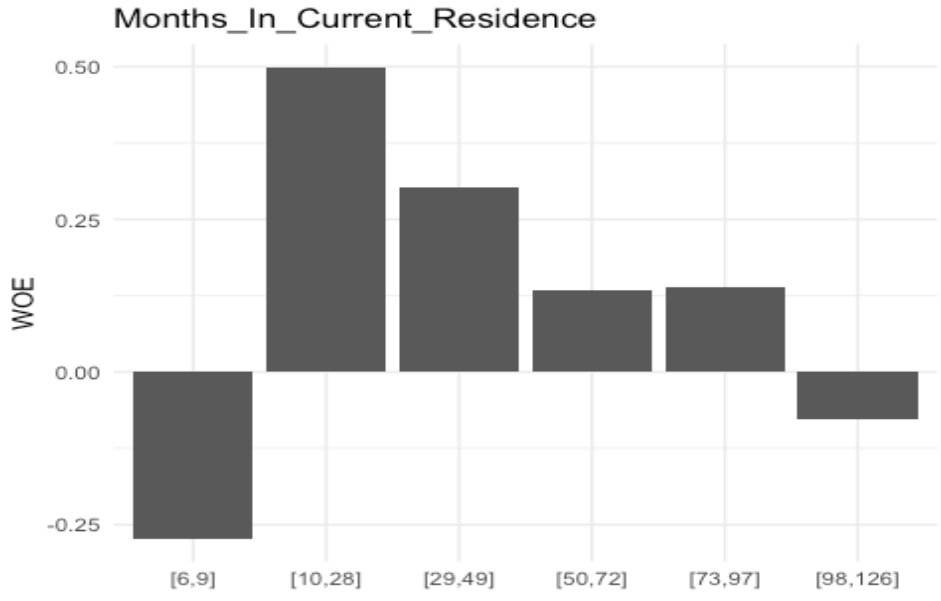
10. No of Months in Current Residence –



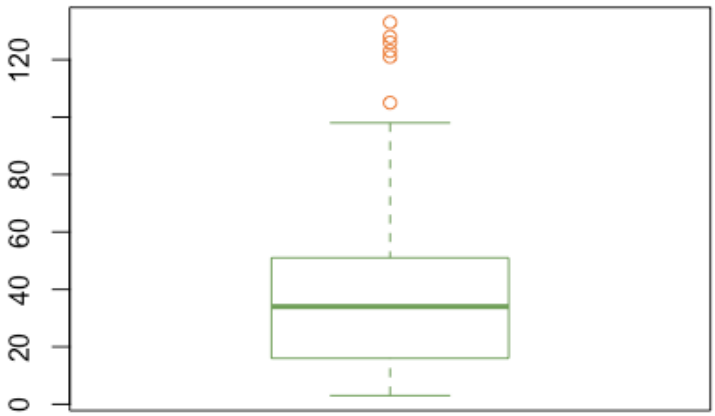


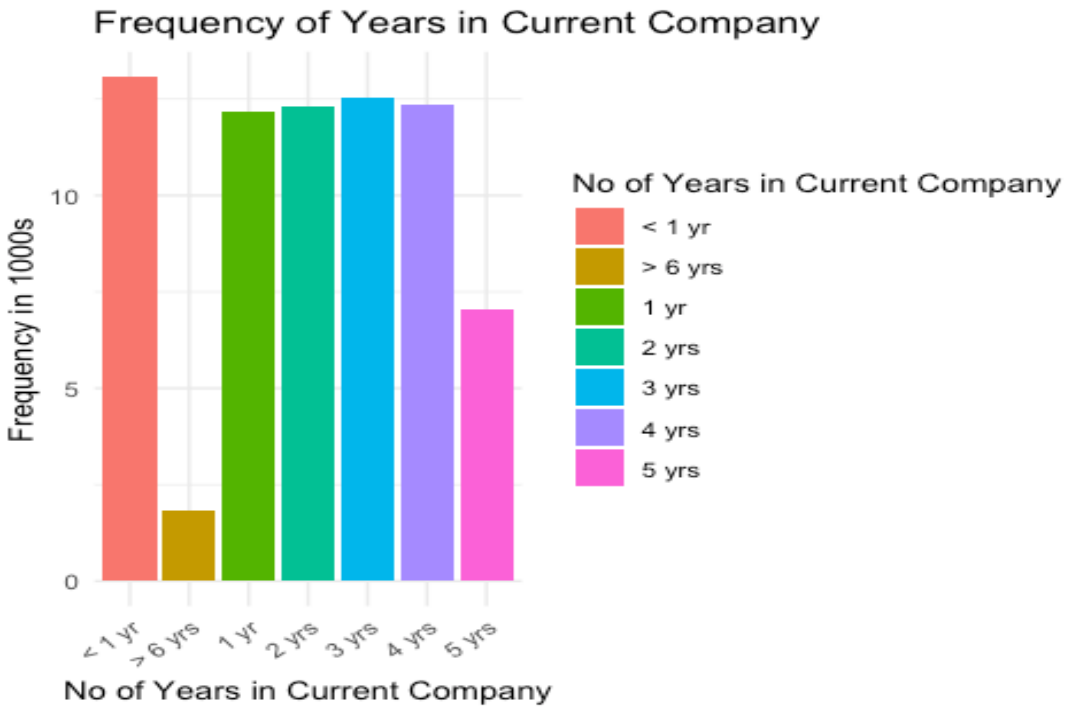


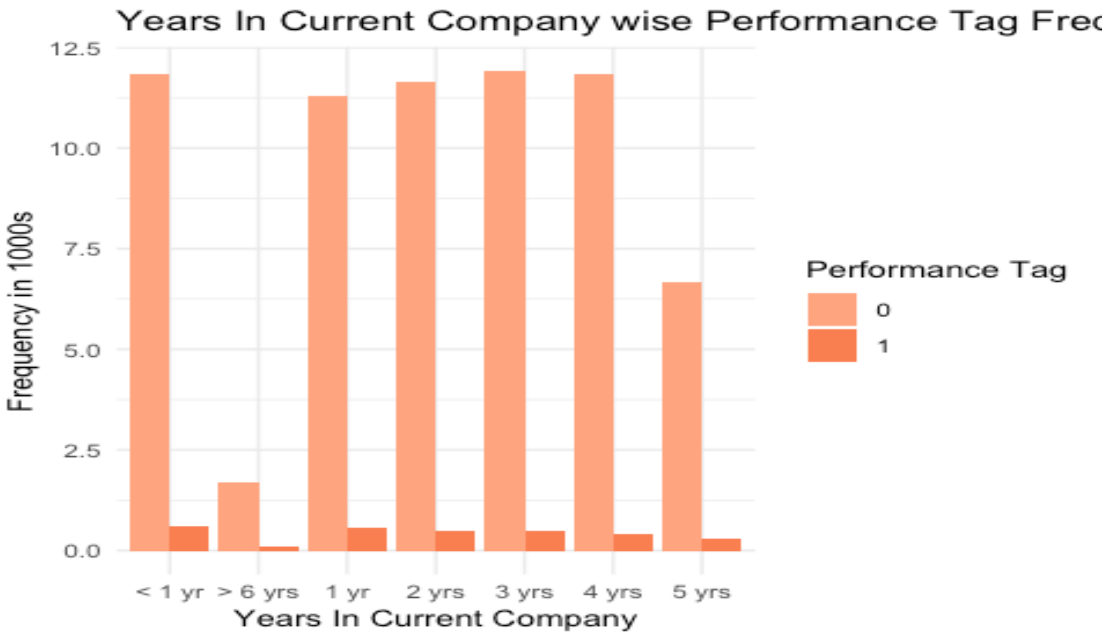
Months_In_Current_Residence	N	Percent	WOE	IV
[6,9]	34693	0.49657907	-0.27220657	0.03253901
[10,28]	6922	0.09907821	0.49867827	0.06363545
[29,49]	7210	0.10320050	0.30113949	0.07439660
[50,72]	6988	0.10002290	0.13397271	0.07630615
[73,97]	6931	0.09920703	0.13943606	0.07836294
[98,126]	7120	0.10191229	-0.07681208	0.07894353



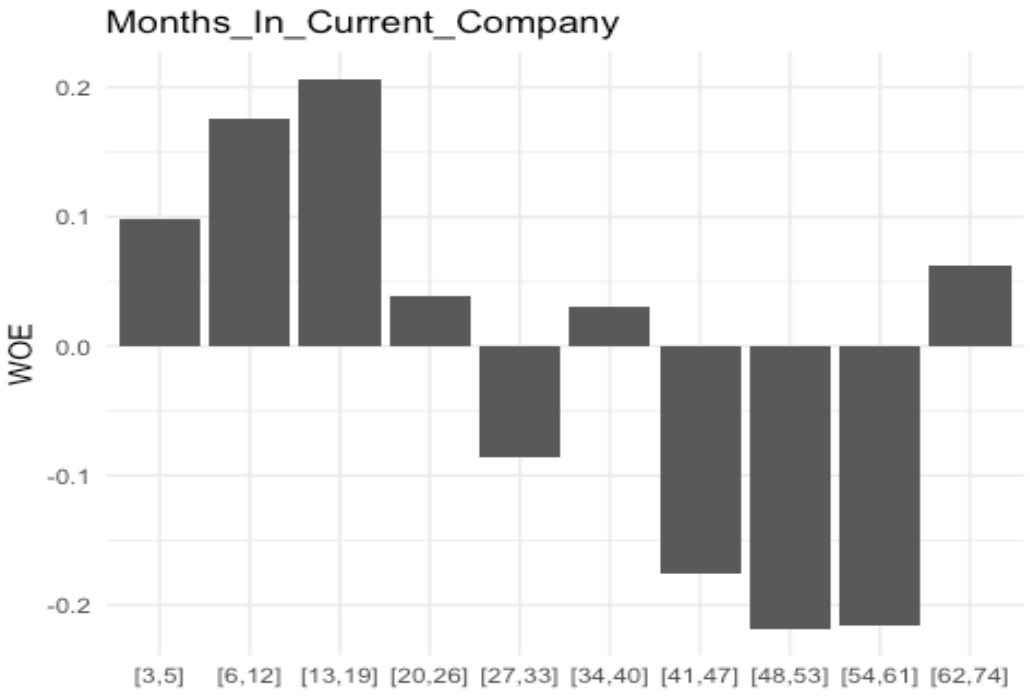
11. **No of Months in Current Company –** The outliers are treated by capping the max value at 74.





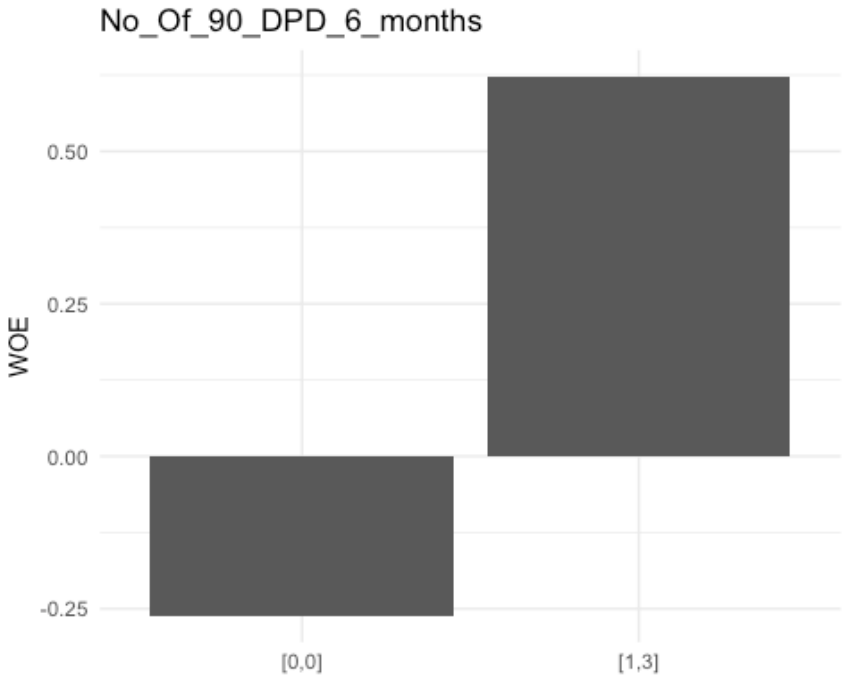


Months_In_Current_Company	N	Percent	WOE	IV
[3,5]	6689	0.09574316	0.09847101	0.0009713844
[6,12]	6797	0.09728902	0.17559050	0.0042241321
[13,19]	6933	0.09923566	0.20626208	0.0088680976
[20,26]	6919	0.09903527	0.03915191	0.0090226567
[27,33]	7104	0.10168327	-0.08572088	0.0097411937
[34,40]	7182	0.10279973	0.03074914	0.0098397718
[41,47]	7217	0.10330070	-0.17619333	0.0128001924
[48,53]	6169	0.08830013	-0.21796666	0.0166009719
[54,61]	7822	0.11196038	-0.21618018	0.0213452997
[62,74]	7032	0.10065270	0.06284108	0.0217544128



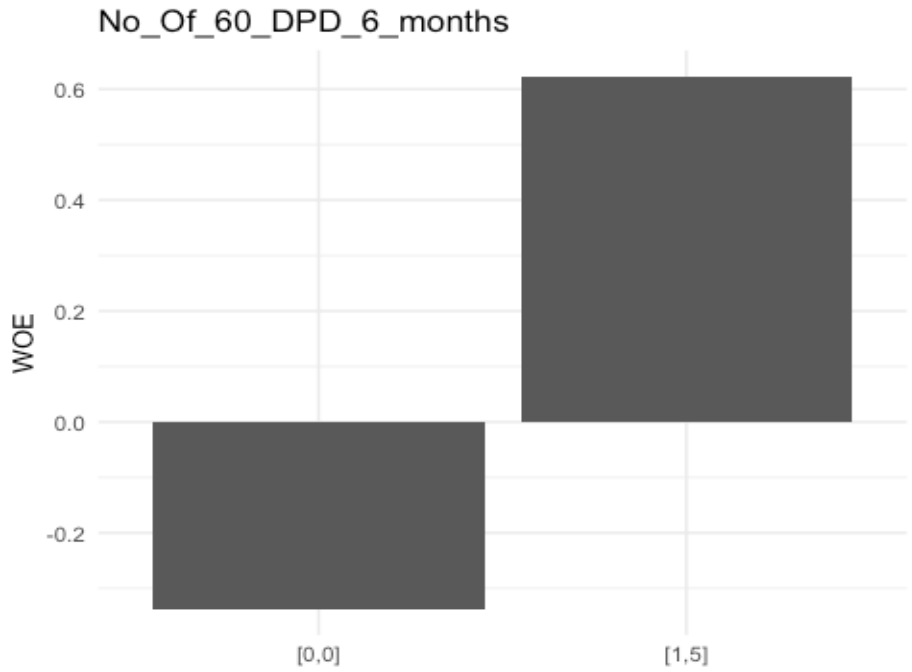
12. No of 90 DPD or more in last 6 months –

No_Of_90_DPD_6_months	N	Percent	WOE	IV
[0,0]	54662	0.7824058	-0.2606851	0.04726187
[1,3]	15202	0.2175942	0.6224814	0.16011692



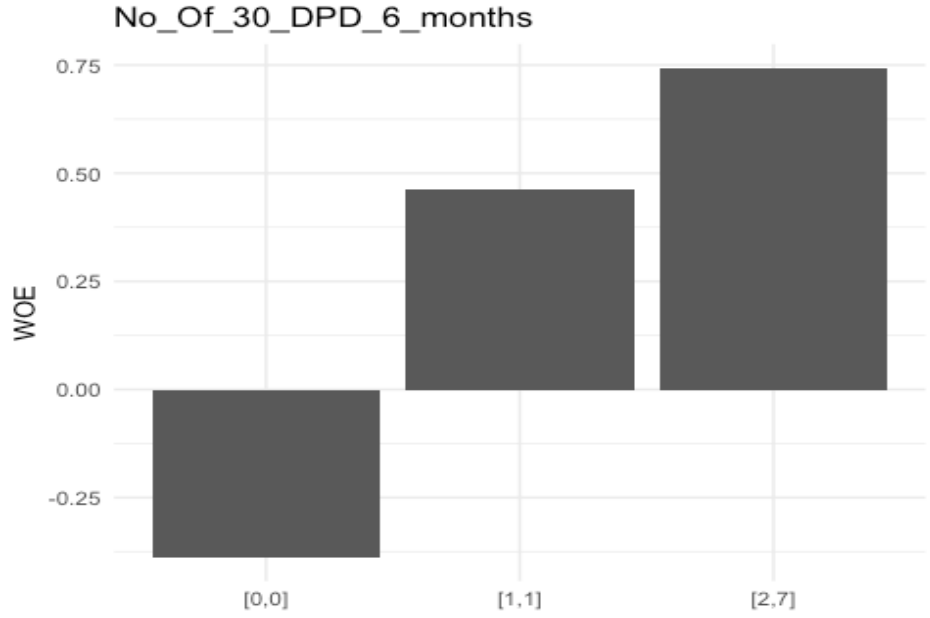
13. No of 60 DPD or more in last 6 months –

No_Of_60_DPD_6_months	N	Percent	WOE	IV
[0,0]	51868	0.7424138	-0.3363715	0.07220252
[1,5]	17996	0.2575862	0.6225513	0.20583388

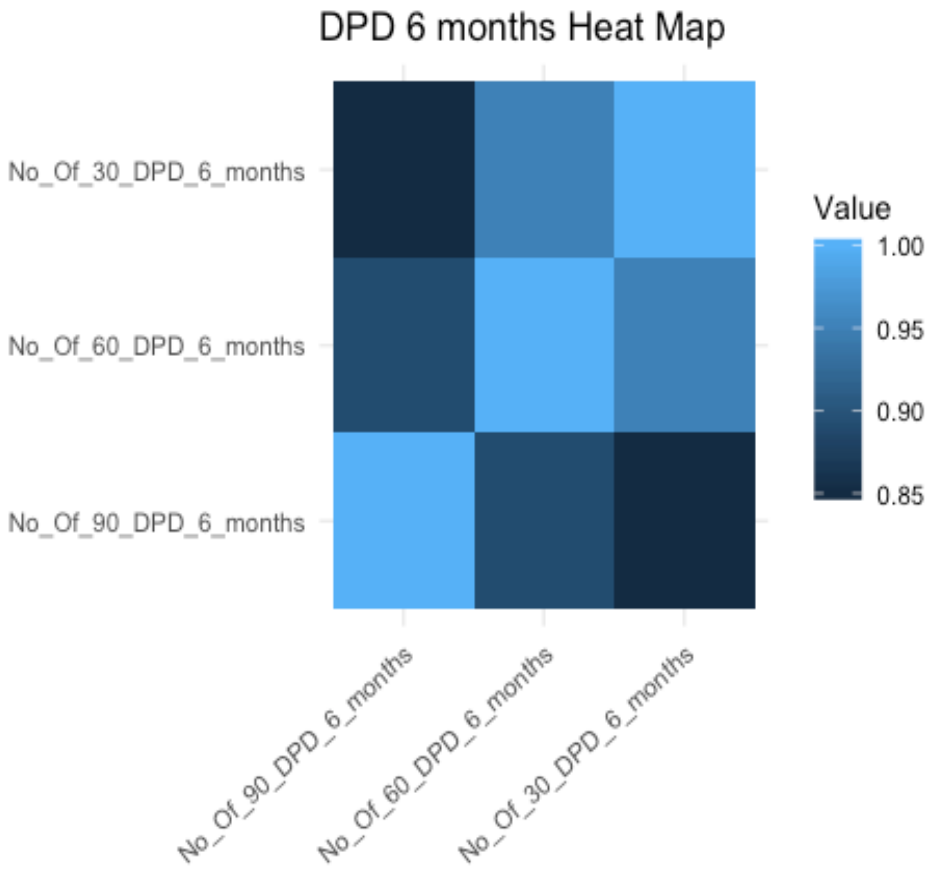


14. No of 30 DPD or more in last 6 months –

No_Of_30_DPD_6_months	N	Percent	WOE	IV
[0,0]	50096	0.7170503	-0.3867956	0.09018646
[1,1]	9500	0.1359785	0.4642738	0.12658101
[2,7]	10268	0.1469713	0.7429064	0.24156274

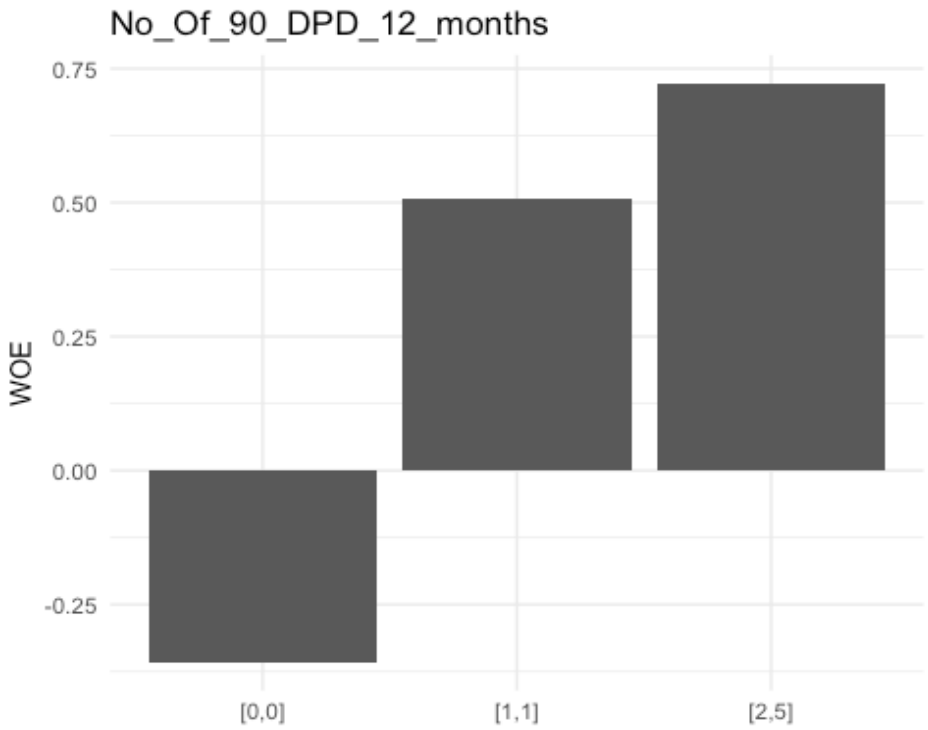


DPD for 6 months correlation – High



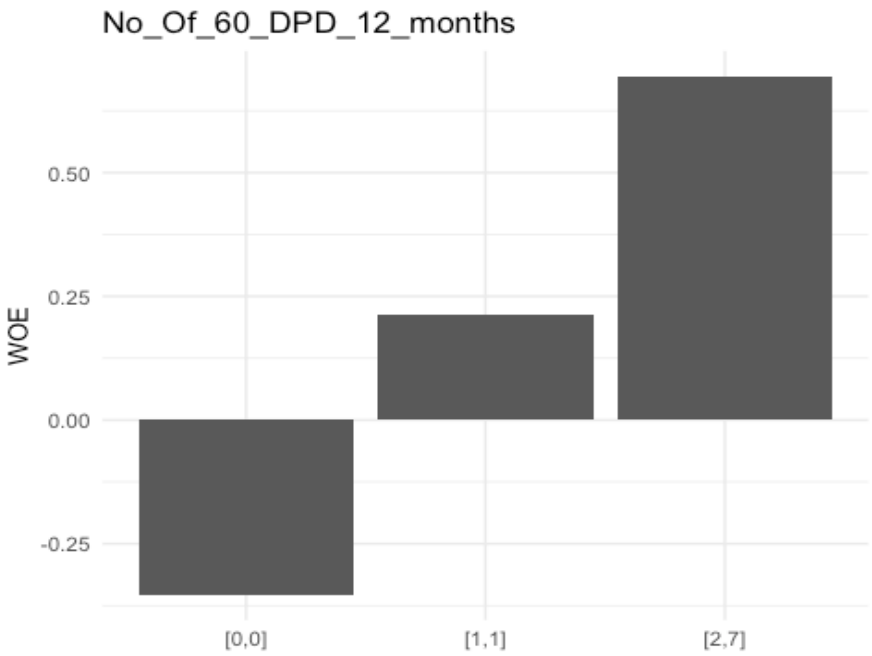
15. No of 90 DPD or more in last 12 months –

No_Of_90_DPD_12_months	N	Percent	WOE	IV
[0,0]	50490	0.7226898	-0.3566371	0.07830539
[1,1]	11663	0.1669386	0.5087786	0.13310592
[2,5]	7711	0.1103716	0.7220790	0.21387484



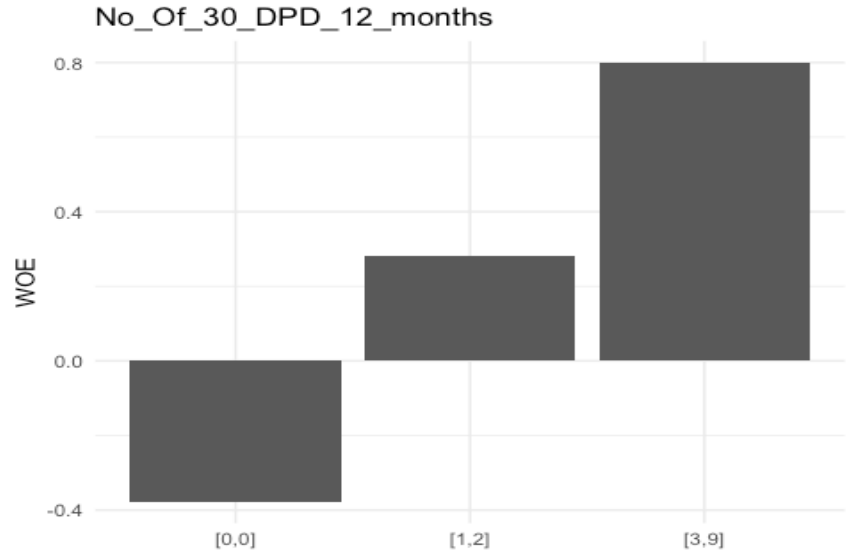
16. No of 60 DPD or more in last 12 months –

No_Of_60_DPD_12_months	N	Percent	WOE	IV
[0,0]	45866	0.6565041	-0.3519210	0.06940917
[1,1]	12816	0.1834421	0.2141090	0.07869326
[2,7]	11182	0.1600538	0.6941383	0.18549887

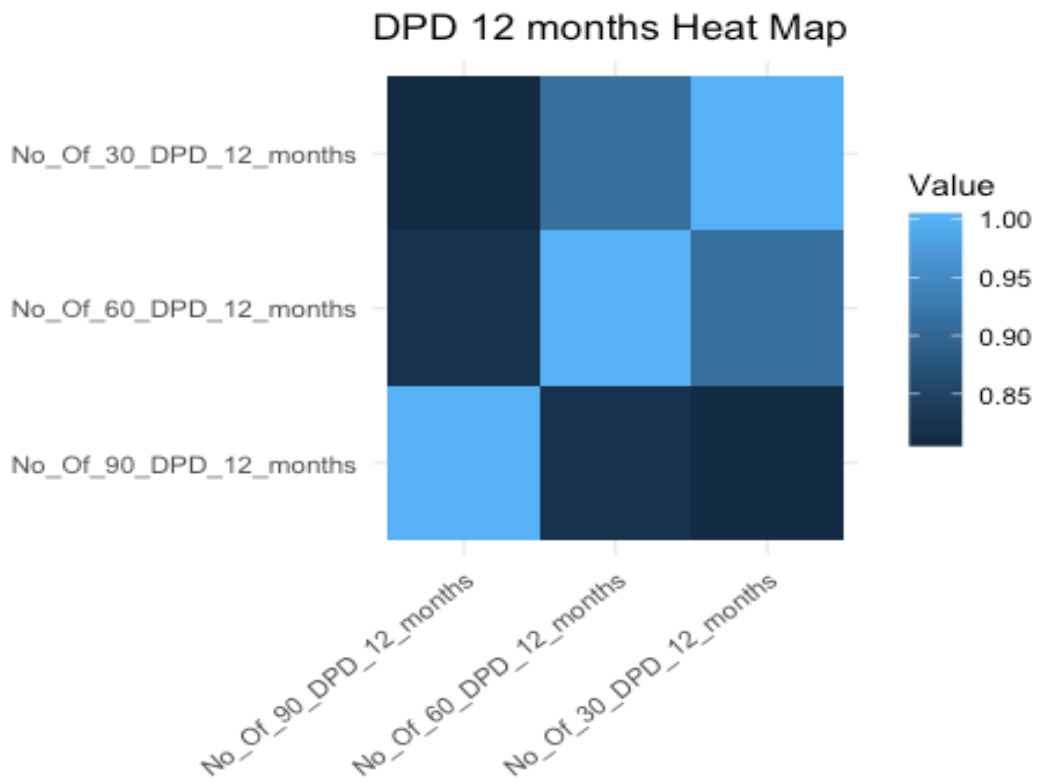


17. No of 30 DPD or more in last 12 months –

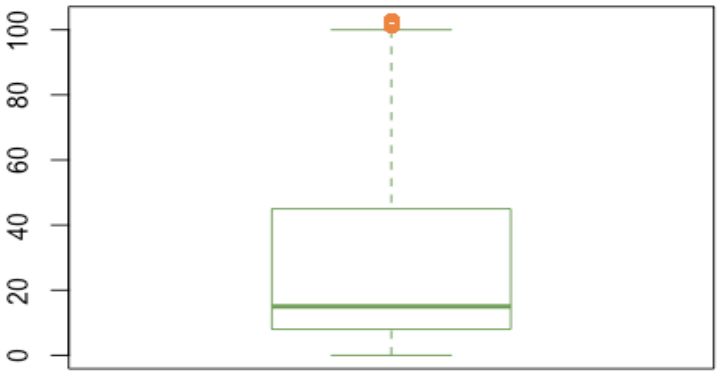
No_Of_30_DPD_12_months	N	Percent	WOE	IV
[0,0]	44855	0.6420331	-0.3763949	0.07681694
[1,2]	17590	0.2517749	0.2805077	0.09937723
[3,9]	7419	0.1061920	0.7995966	0.19825486



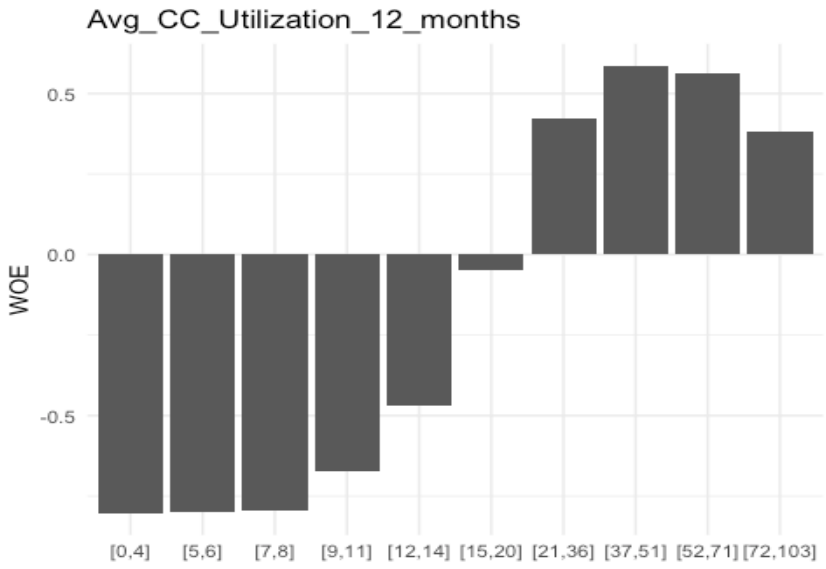
DPD for 12 months correlation – High



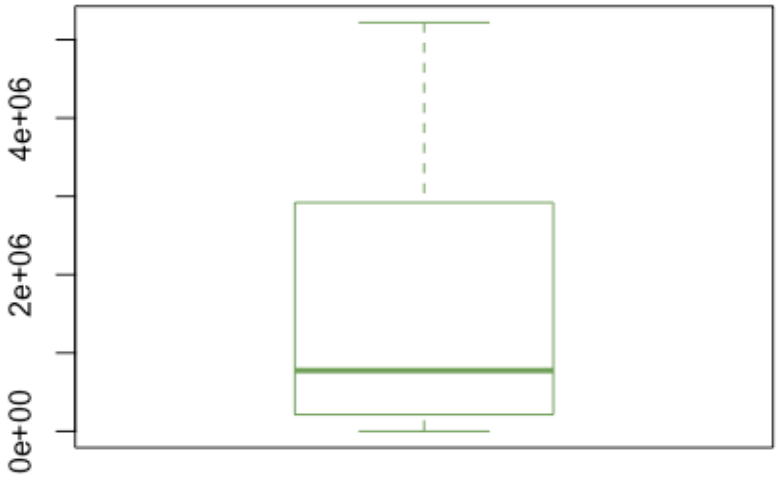
18. Average Credit Card Utilization –



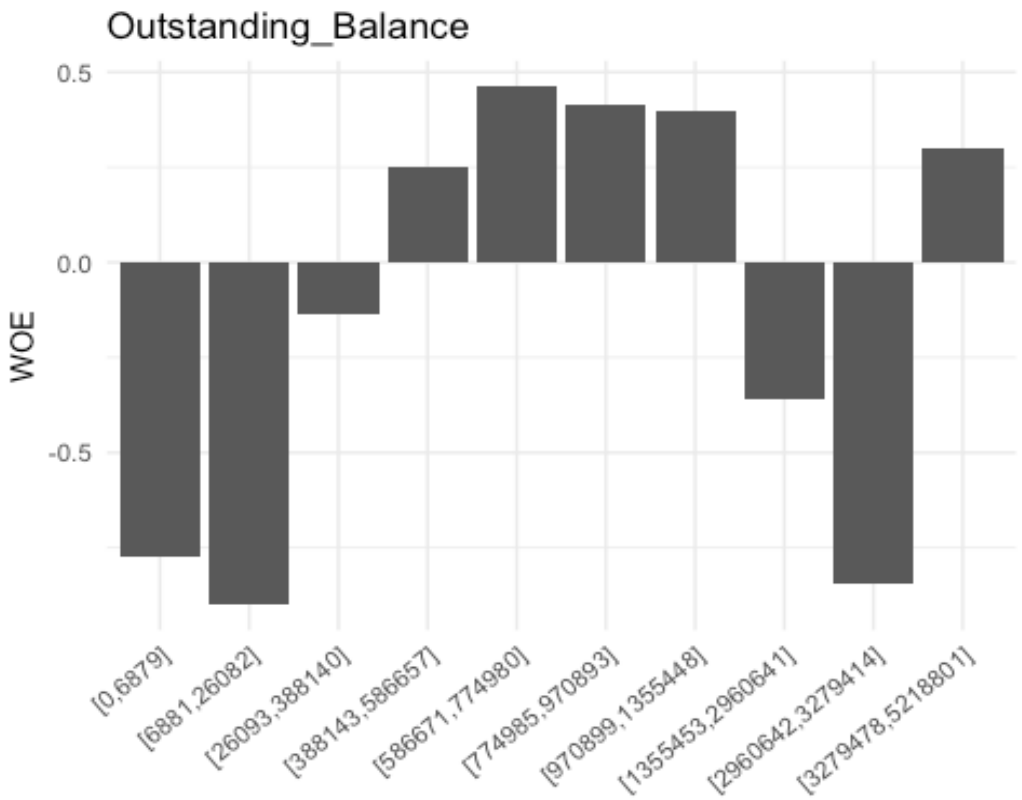
Avg_CC_Utilization_12_months	N	Percent	WOE	IV
[0,4]	5524	0.07906790	-0.8018033	0.03579484
[5,6]	5471	0.07830929	-0.8015472	0.07122737
[7,8]	6869	0.09831959	-0.7945679	0.11506932
[9,11]	9596	0.13735257	-0.6723445	0.16122586
[12,14]	6595	0.09439769	-0.4680467	0.17800476
[15,20]	7197	0.10301443	-0.0477444	0.17823452
[21,36]	7372	0.10551929	0.4244121	0.20139476
[37,51]	7175	0.10269953	0.5857448	0.24774164
[52,71]	7016	0.10042368	0.5638397	0.28930046
[72,103]	7049	0.10089603	0.3812966	0.30681520


Avg_CC_Utilization_12_months
WOE
0.5
0.0
-0.5
[0,4]
[5,6]
[7,8]
[9,11]
[12,14]
[15,20]
[21,36]
[37,51]
[52,71]
[72,103]

19. Outstanding Balance –

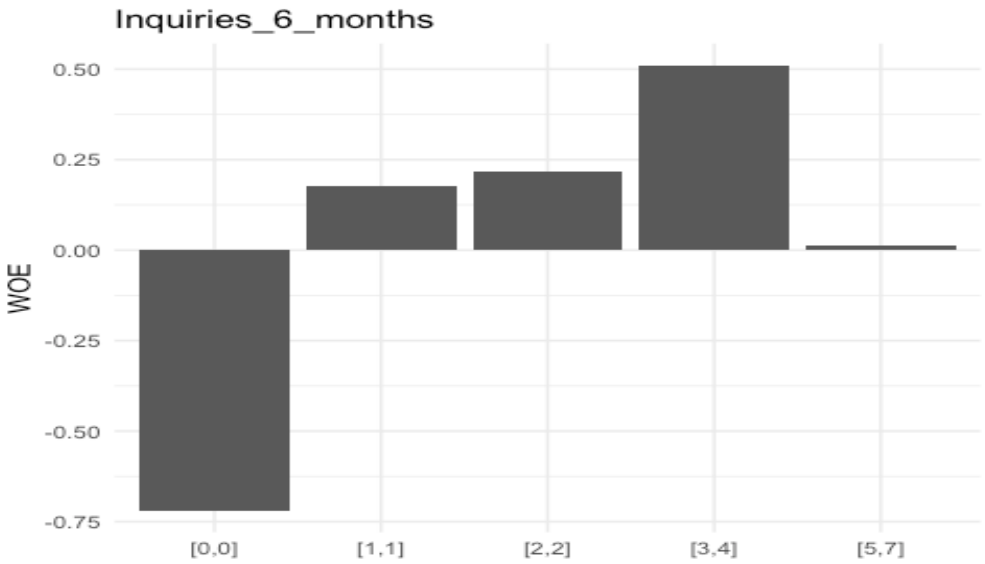


Outstanding_Balance	N	Percent	WOE	IV
[0,6879]	6985	0.09997996	-0.7742805	0.04269323
[6881,26082]	6987	0.10000859	-0.8991959	0.09739804
[26093,388140]	6986	0.09999427	-0.1343912	0.09909695
[388143,586657]	6986	0.09999427	0.2531012	0.10629870
[586671,774980]	6914	0.09896370	0.4626534	0.13258154
[774985,970893]	7059	0.10103916	0.4141617	0.15359820
[970899,1355448]	6987	0.10000859	0.3956813	0.17242067
[1355453,2960641]	6986	0.09999427	-0.3612339	0.18351395
[2960642,3279414]	6987	0.10000859	-0.8428255	0.23269920
[3279478,5218801]	6987	0.10000859	0.2971542	0.24283436



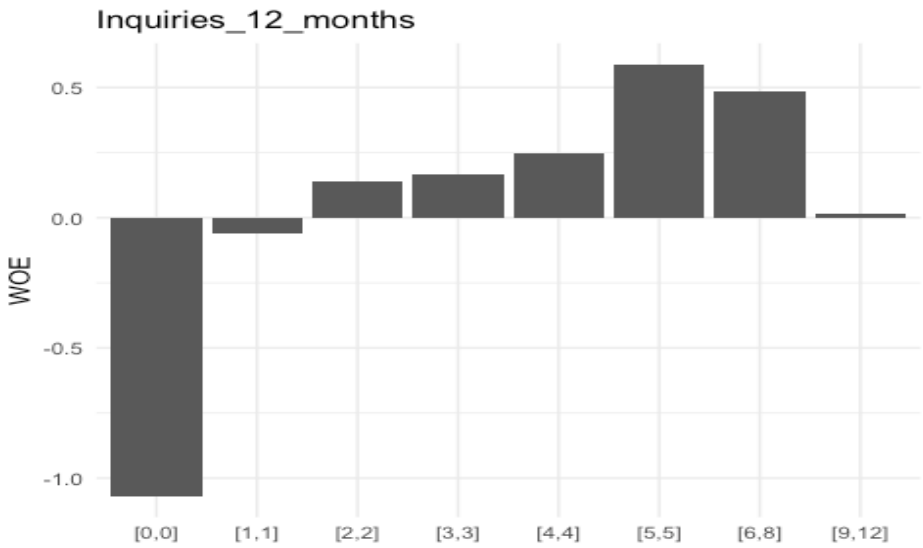
20. Inquiries in last 6 months excluding home and auto loans –

Inquiries_6_months	N	Percent	WOE	IV
[0,0]	25068	0.3588114	-0.71823458	0.1349649
[1,1]	13175	0.1885807	0.17702727	0.1413777
[2,2]	12830	0.1836425	0.21613413	0.1508576
[3,4]	11505	0.1646771	0.50984899	0.2051710
[5,7]	7286	0.1042883	0.01237065	0.2051870



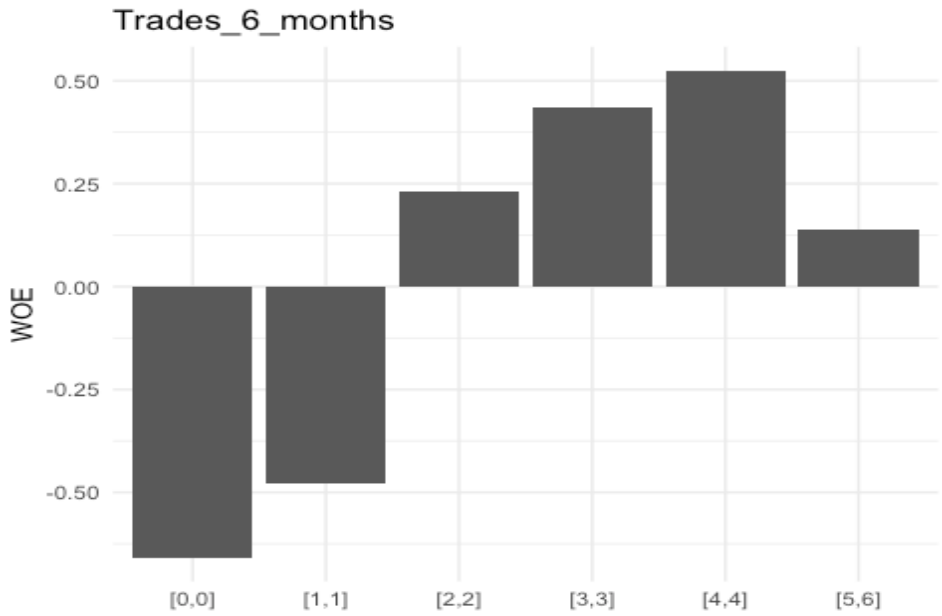
21. Inquiries in last 12 months excluding home and auto loans –

Inquiries_12_months	N	Percent	WOE	IV
[0,0]	20580	0.29457231	-1.06753214	0.2122079
[1,1]	3899	0.05580843	-0.06181938	0.2124153
[2,2]	7906	0.11316272	0.14223276	0.2148596
[3,3]	8978	0.12850681	0.16430448	0.2186019
[4,4]	7113	0.10181209	0.24806051	0.2256288
[5,5]	4926	0.07050842	0.58835484	0.2577723
[6,8]	8951	0.12812035	0.48408671	0.2954041
[9,12]	7511	0.10750887	0.01366001	0.2954243



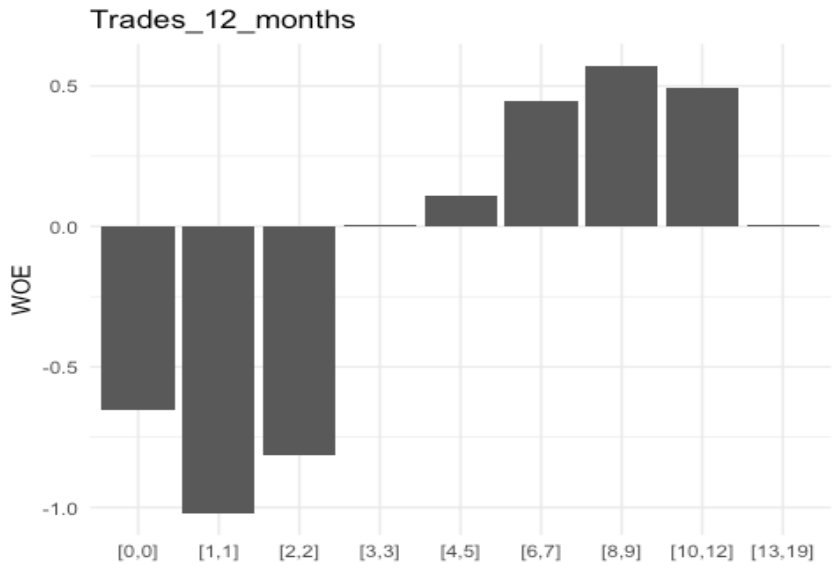
22. No of Trades opened in last 6 months –

Trades_6_months	N	Percent	WOE	IV
[0,0]	12193	0.17452479	-0.6575894	0.05645327
[1,1]	20120	0.28798809	-0.4795091	0.10991361
[2,2]	12116	0.17342265	0.2328162	0.12038229
[3,3]	9402	0.13457575	0.4350791	0.15158030
[4,4]	6297	0.09013226	0.5242321	0.18322344
[5,6]	9736	0.13935646	0.1368108	0.18600147

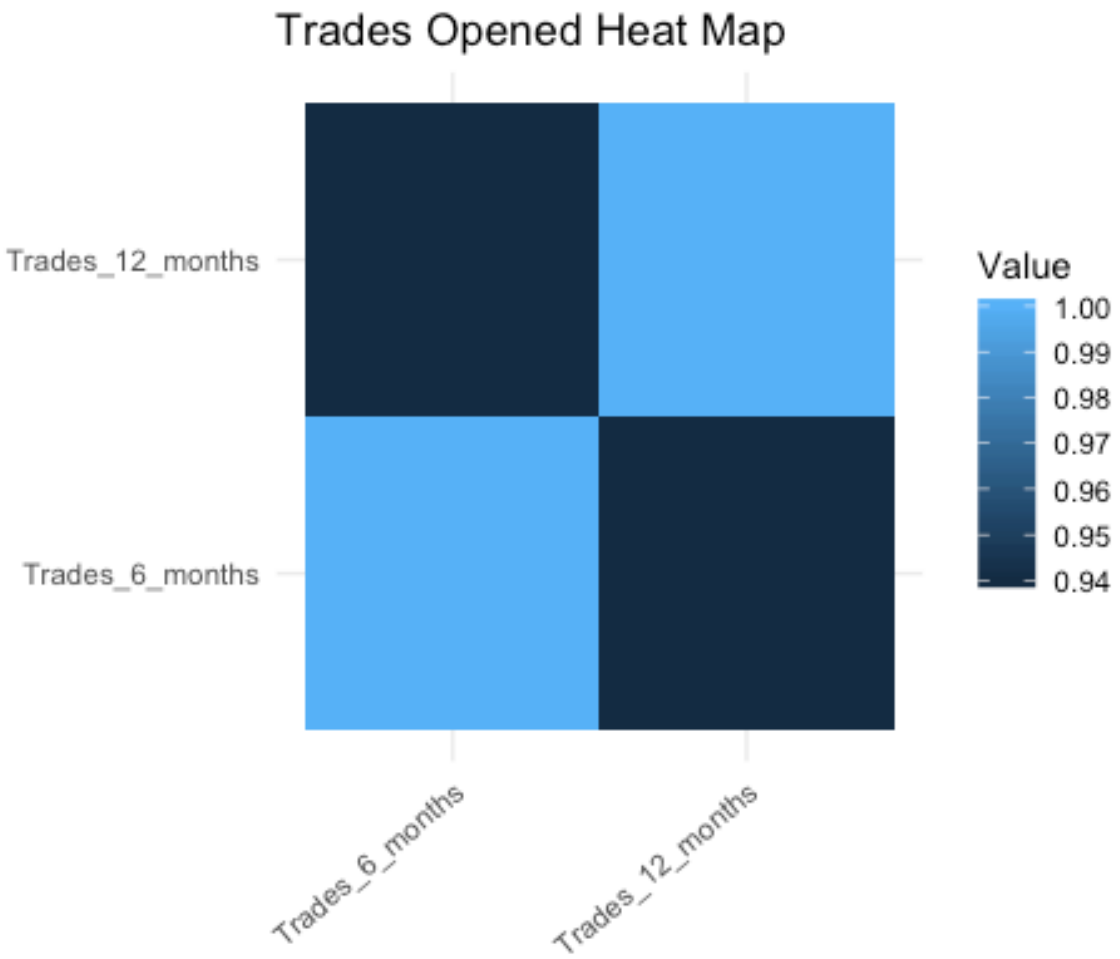


23. No of Trades opened in last 12 months –

Trades_12_months	N	Percent	WOE	IV
[0,0]	4955	0.07092351	-0.653300562	0.0226844
[1,1]	11377	0.16284496	-1.019130876	0.1316867
[2,2]	9322	0.13343067	-0.816404313	0.1939347
[3,3]	4678	0.06695866	0.003554047	0.1939355
[4,5]	9397	0.13450418	0.109249440	0.1956237
[6,7]	8296	0.11874499	0.448065618	0.2249993
[8,9]	7175	0.10269953	0.571295242	0.2687861
[10,12]	6699	0.09588629	0.491736195	0.2979526
[13,19]	7965	0.11400721	0.006261375	0.2979571

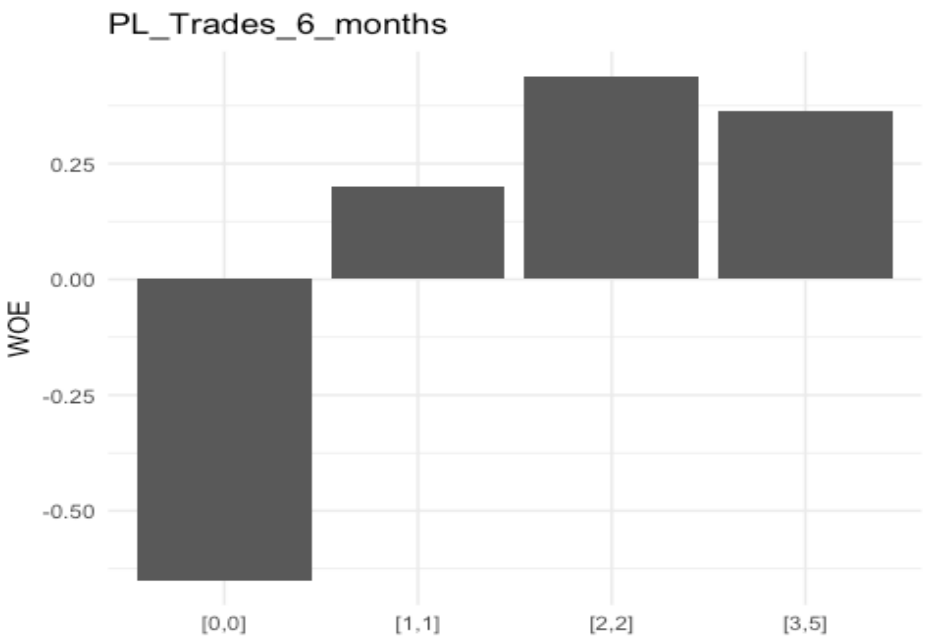


Trades Opened Correlation – High



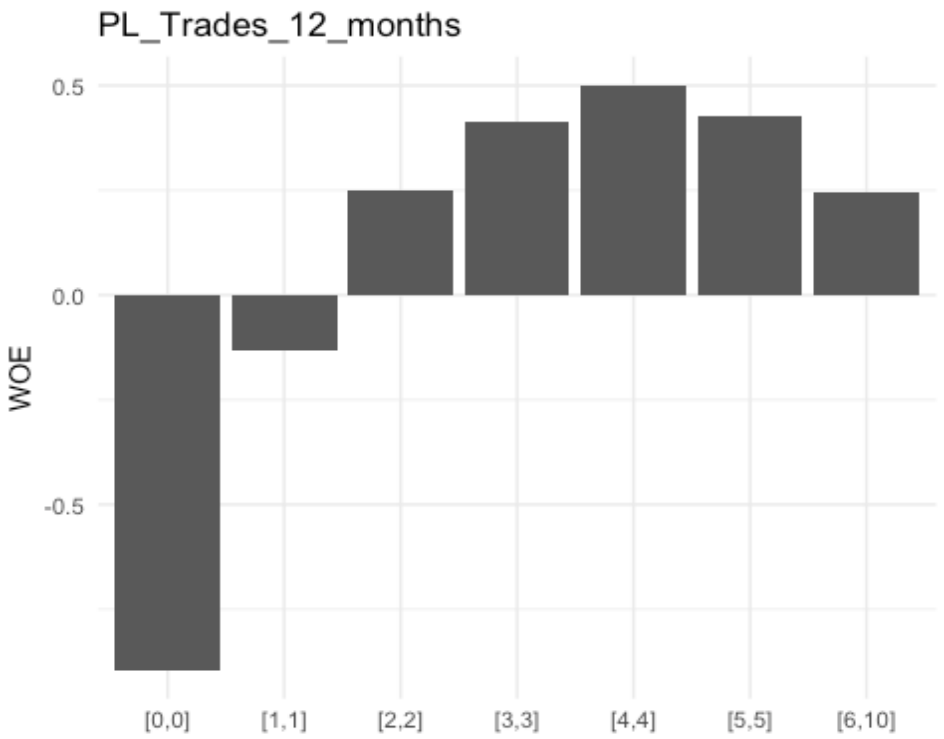
24. No of PL Trades opened in last 6 months –

PL_Trades_6_months	N	Percent	WOE	IV
[0,0]	31078	0.4448357	-0.6491908	0.1407380
[1,1]	13545	0.1938767	0.1993948	0.1491898
[2,2]	12565	0.1798494	0.4383908	0.1915868
[3,5]	12676	0.1814382	0.3619170	0.2197050

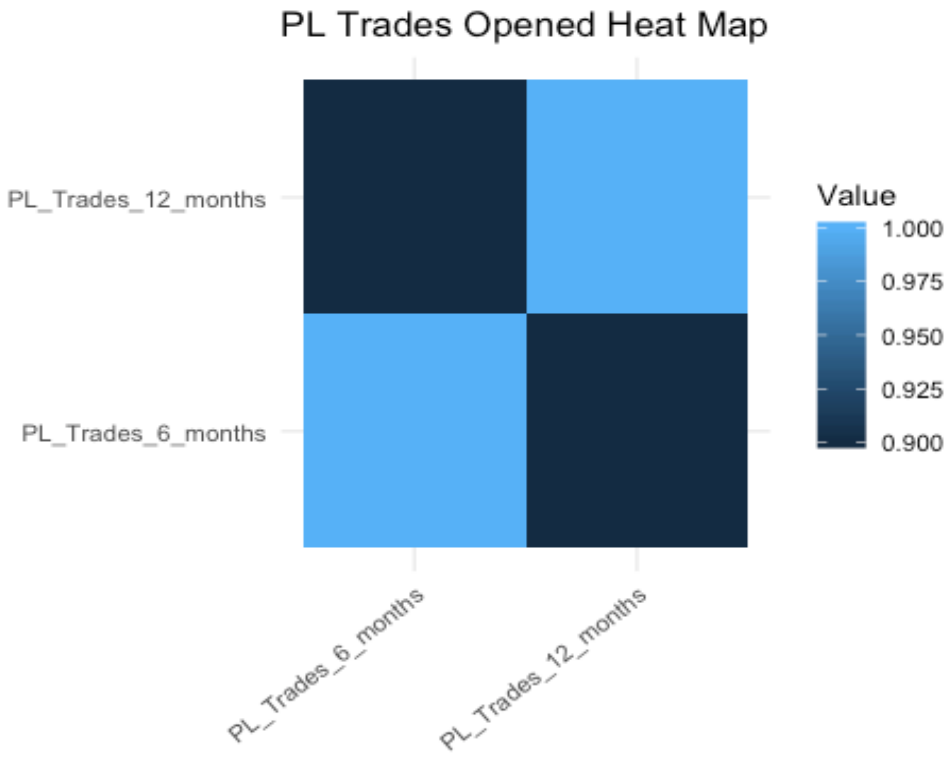


25. No of PL Trades opened in last 12 months –

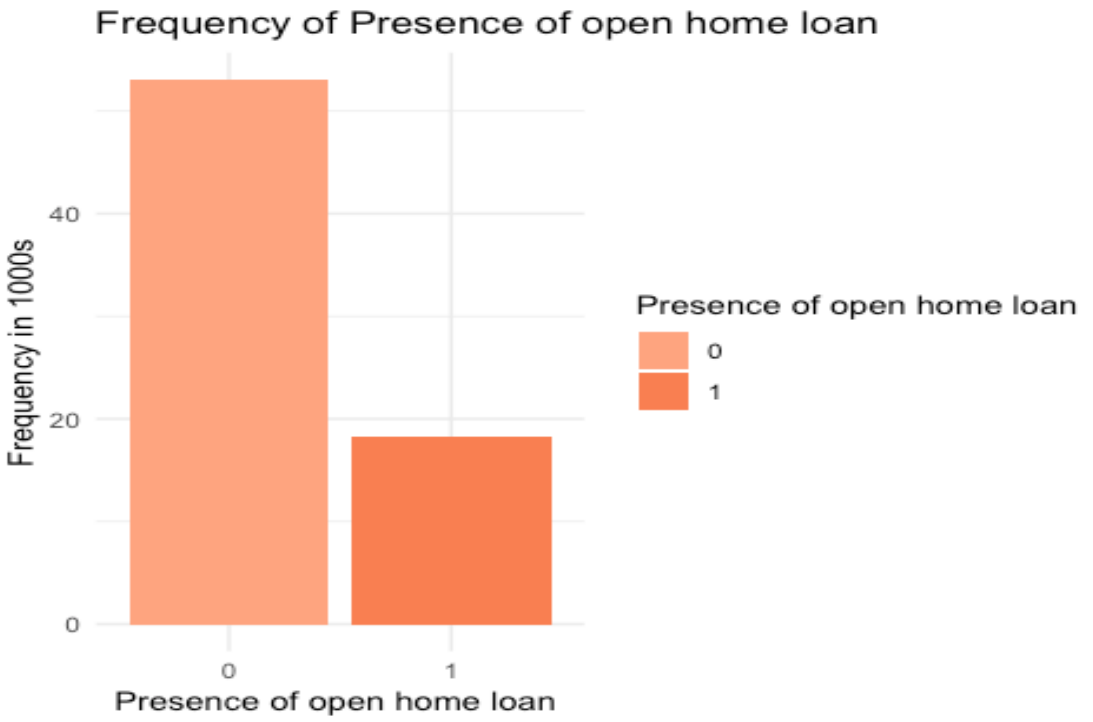
PL_Trades_12_months	N	Percent	WOE	IV
[0,0]	25823	0.36961812	-0.8938162	0.2002092
[1,1]	6640	0.09504180	-0.1309052	0.2017437
[2,2]	6830	0.09776136	0.2512951	0.2086786
[3,3]	8130	0.11636895	0.4122511	0.2326395
[4,4]	7902	0.11310546	0.5001662	0.2683759
[5,5]	6189	0.08858640	0.4261046	0.2879906
[6,10]	8350	0.11951792	0.2431127	0.2958955

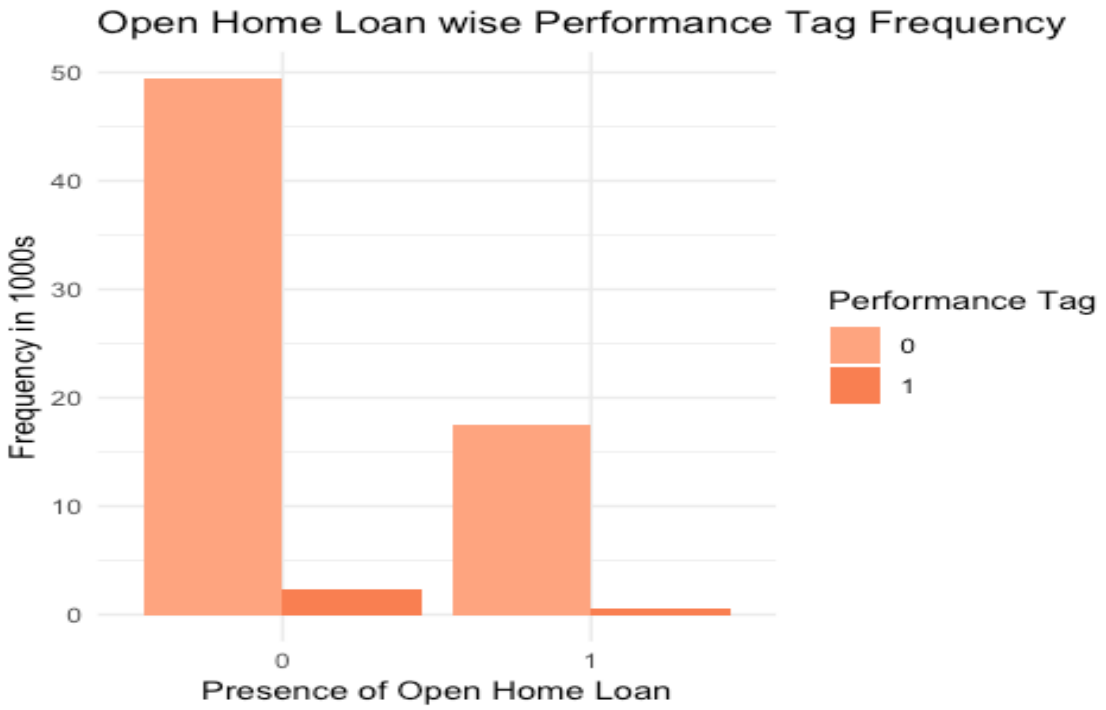


PL Trades Opened Correlation – High

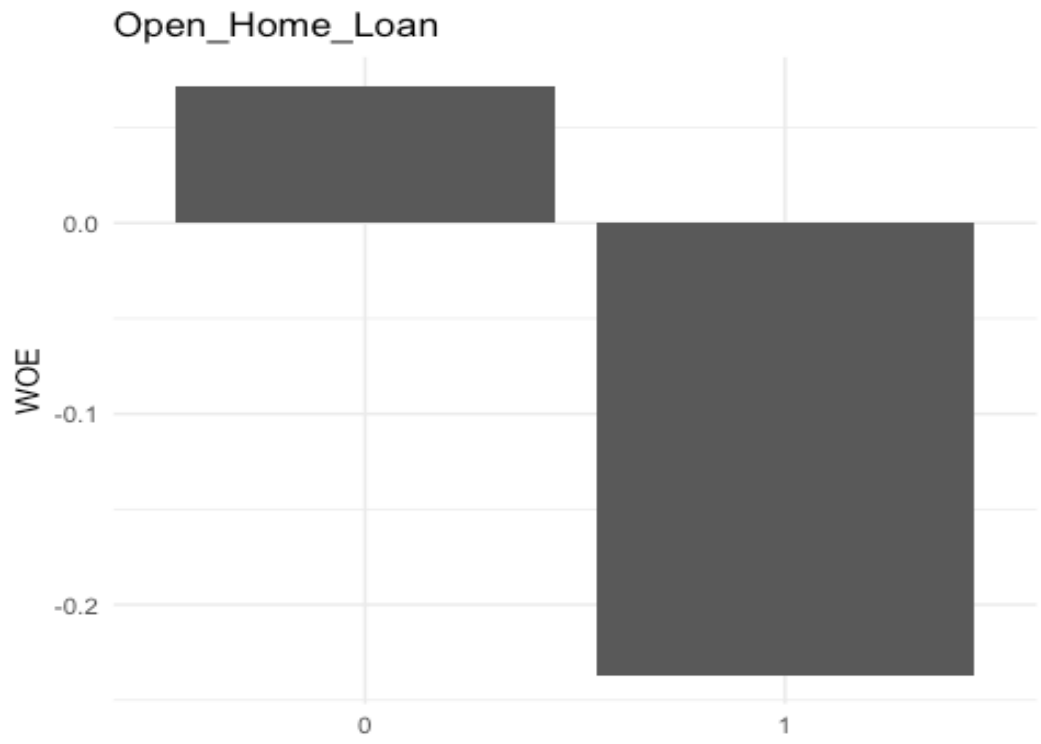


26. Open Home Loan – The Presence of Open Home Loan variable has 272 NA's which are replaced with median value of 0.

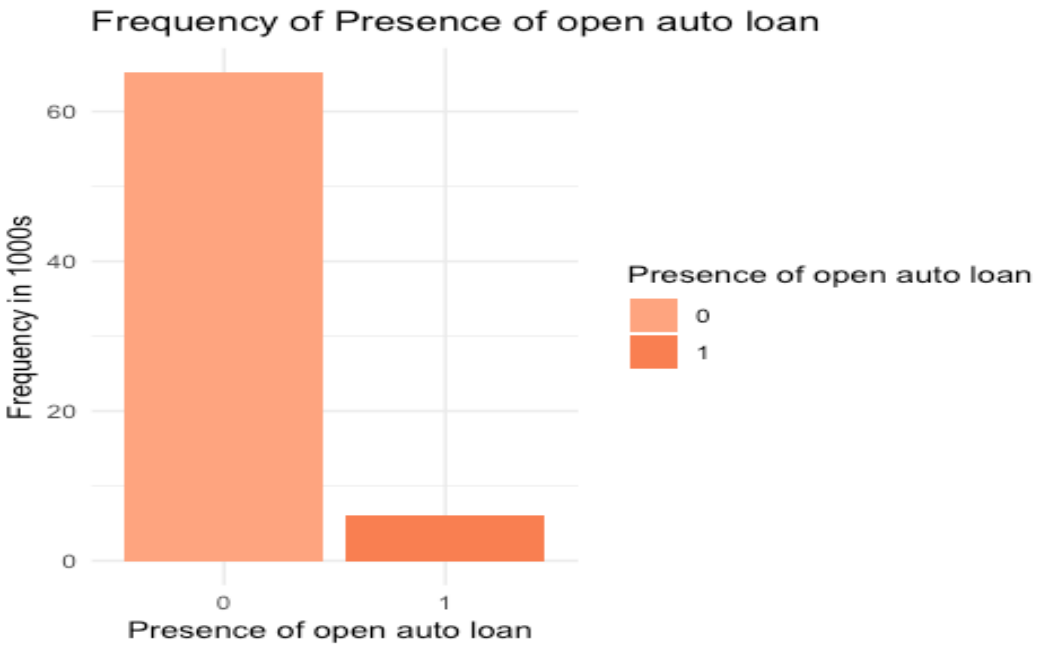


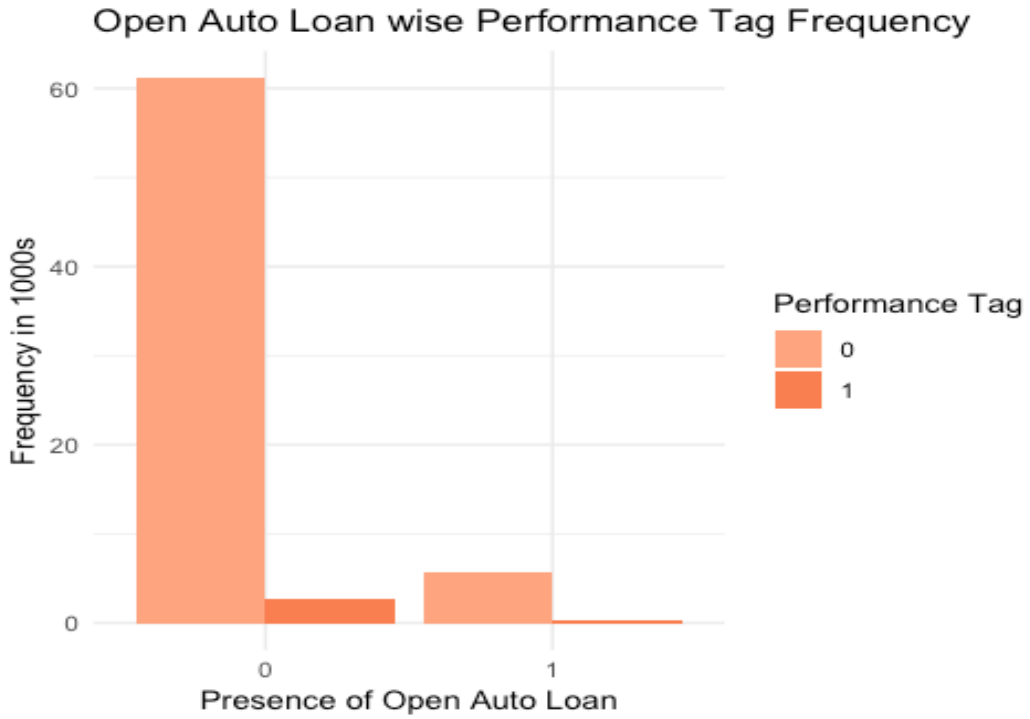


Open_Home_Loan	N	Percent	WOE	IV
0	51793	0.7413403	0.07179355	0.003949208
1	18071	0.2586597	-0.23670277	0.016969717

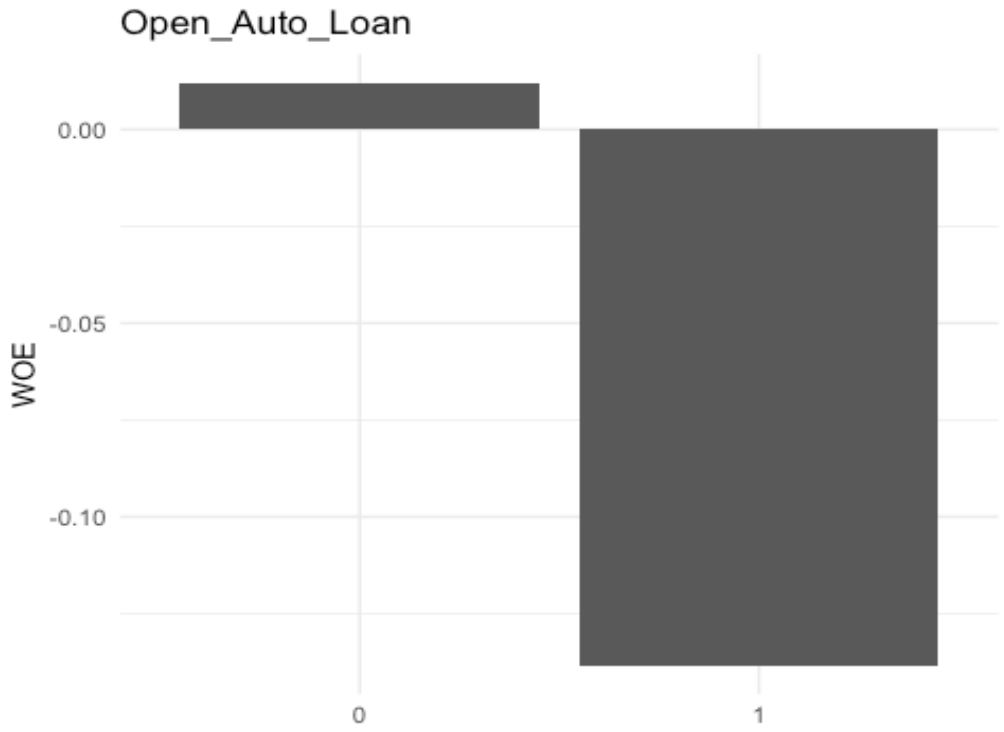


27. Open Auto Loan –

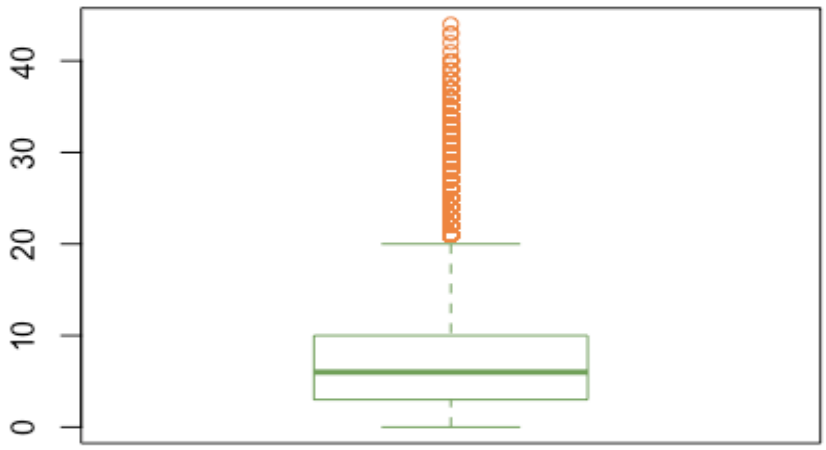




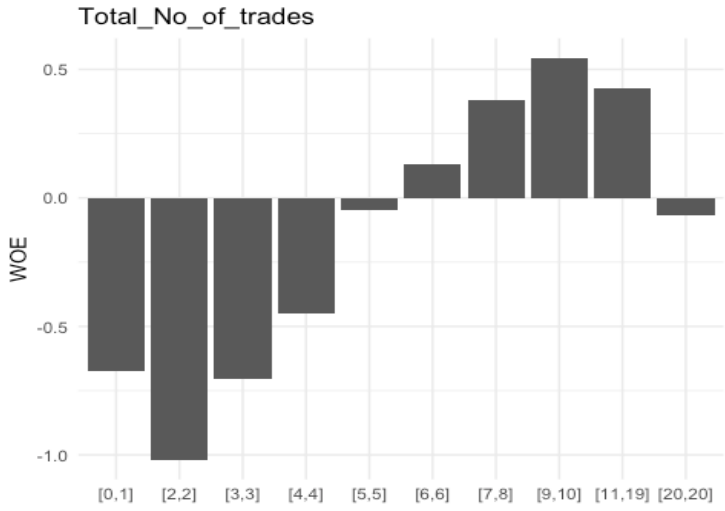
Open_Auto_Loan	N	Percent	WOE	IV
0	63935	0.91513512	0.01197252	0.000131898
1	5929	0.08486488	-0.13823723	0.001654820



28. Total No of Trades – The outliers are treated by capping the maximum value at 20.



Total_No_of_trades	N	Percent	WOE	IV
[0,1]	3914	0.05602313	-0.67308511	0.01886185
[2,2]	6765	0.09683099	-1.01762017	0.08352403
[3,3]	8614	0.12329669	-0.70195095	0.12812640
[4,4]	7490	0.10720829	-0.44789740	0.14573053
[5,5]	5714	0.08178747	-0.04884539	0.14592136
[6,6]	4966	0.07108096	0.12925644	0.14718176
[7,8]	9360	0.13397458	0.37943447	0.17019188
[9,10]	7133	0.10209836	0.54389543	0.20913701
[11,19]	8476	0.12132142	0.42713095	0.23614255
[20,20]	7432	0.10637811	-0.06694280	0.23660492



Insights:

1. Out of 71295 rows of both the Demographic and Credit datasets, 6 rows have duplicate Application ID. Since these Application IDs have different data, they cannot be merged with surety that each Application Id have exact data. So these rows are removed.

2. The Performance Tag variable has 1425 rows having NA value. These rows are removed for EDA and would be used as test dataset.

3. We observed that the Information Value for the Demographic Variables are less compared to the Credit Variables.

4. The NA's of the variables is either replaced by the median or the value with most frequency.

Next Step and Approach:

- The two types of model we need to build:
    1. Demographic Data Model: Model build only with the Demographic dataset.
    2. Credit and Demographic Data Model: Model build with the merged dataset of Credit and Demographic Dataset.

- The model will be evaluated using Accuracy, Sensitivity, Specificity, Gain & Lift and KS Statistics.

- The correct model will be selected with K-Fold cross validation.

- Build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

- Access and explain the potential financial benefits of the project.