



Vehicle Loan & Amount of Money Sanctioned

Sai Vivek Bala
Sahil Agarwal
12/02/2019

MSBA-5324

Table of Contents

Abstract	3
Problem Identification	5
Methodology	6
Dataset description	6
Statistical Model	8
Logistic Regression	8
Multiple Regression	8
Major Findings	9
Descriptive Analysis	9
Relevant Findings	15
Statistical Analysis	16
Summary Statistics	16
Frequency Distribution	16
Chi-Square Test	17
Hypothesis Testings	18
Chi-Square Test	18
T-Test	21
Model	27
Logistic Regression	27
Multiple Regression	29
Marketing Implications and Conclusions	30
Marketing Implications	30
Conclusion	31
Appendix	32
Overall Efforts	33
SAS Codes	34
Python Codes	40

Abstract

Vehicle loan sanctioned has been widely used by several companies. Our main aim of the project is to understand the process of Loan sanctioned to the customer and how much amount of the Loan to be sanctioned to the customer. Our team has thought of collecting the dataset and retrieved a compatible dataset from the Kaggle. Our dataset comprises of all the variables which will give insight about the whole process. Once we got the dataset, our team noticed to clean the dataset with the help of SAS and Excel. After manipulating the dataset, we decided to use a particular method i.e. Logistic Regression and Multiple Regression which will help in giving the precise output. Loan Sanctioned is a categorical variable so we decided to use logistic regression to see which variables will help in deciding whether to give the loan or not. Once we will get the model of deciding the Loan Sanctioned then our aim is to decide the amount of money needs to give to the consumer. For the second step, we decided to use Multiple Regression which will give the model in which AMOUNT_DISBURSED will play as a dependent variable and the rest of the important variable will play as an independent variable.

Additionally, we decided to use Python (Pandas) to run statistical analysis which will help in understanding the significant difference in the proportion of the mean. For the statistical analysis, there will be two t-tests and three chi-square tests that we are going to run i.e. on the income, credit score and employment type.

Chi-Square Test -

The purpose of using the chi-square test is to see whether income greater than or equal to \$ 100,000 is significantly different than the income of less than \$ 100,000 of the customer.

Ho: No relationship between accepting Loan and Income (greater than 100,000)

Ha: Relationship between Loan and Income

Moreover, a credit score greater than 648 is significantly playing a role in deciding whether a person assigned a Loan or not.

Ho: No relationship between accepting Loan and Credit score (greater than 648).

Ha: Relationship between Loan and Credit score.

The use of statistical inference in Employment type (1=Self Employed) is to see whether the employment type is playing any role in deciding the Loan or not.

Ho: No relationship between Loan and Self Employed.

Ha: Relationship between Loan and Self Employed.

T-Tests

The use of the t-test is to check any significant difference between mean income with respect to Loan.

Ho: $\mu_y - \mu_n = 0$

Ha: $\mu_y - \mu_n \neq 0$

The purpose of using a t-test is to check whether any significant difference between mean credit score and loan.

Ho: $\mu_y - \mu_n = 0$

Ha: $\mu_y - \mu_n \neq 0$

Problem Identification

Loan companies usually found a problem in deciding whether a person needs to be assigned a Loan or not. Recently, one of my friends applied for a credit card to purchase a phone. The credit card company denied the application of the person by saying the credit score is low. We both wanted to know about the process of deciding the Loan sanctioned and the variables which robust in getting the Loan. Our main goal of choosing this problem is to explore the variables which decide whether a person needs to get a loan or not based on certain variables. Additionally, how much amount of credit limit or loan amount company should give to the consumer by putting one or multiple values in their model. Banks describes several different factors that need to be taken care of while evaluating Loan i.e. Loan to Value (LTV), Previous loan payments, etc. Once the bank sanctioned a loan, they try to take more time in investigating the profile of the borrower. Due to so much research, banks are able to eliminate the borrowers which may default their loan. But the problem is still not completely fixed. Our aim is to understand the whole process and find out various factors through which banks can efficiently take decisions with the help of the model.

Methodology

Dataset description

After searching on the internet and asking the experienced person in this field, we came to know about the variables required in this analysis. We have cleaned the dataset and removed some of the unnecessary variables i.e.

UID - Unique ID, e_t - Employment Type, p_c_s_D - Perform Credit Score Description, A_D - Amount Disbursed, B_I - Branch ID, S_I - Supplier ID, M_I - Manufacturer ID, CURRENT_PINCODE_ID, D_O_B - Date of Birth, D_D - Disbursal Date, EMPLOYEE_CODE_ID, MOBILENO_AVL_FLAG, STATE_ID, DRIVING_FLAG, PASSPORT_FLAG, PRI_O_A - Primary Overdue Amount, PRI_S_A - Primary Sanctioned Amount, PRI_D_A - Primary Disbursed Amount, SEC_O_A - Secondary Overdue Amount, SEC_S_A - Secondary Sanctioned Amount and SEC_D_A - Secondary Disbursed Amount.

Some of the variables which we created temporarily i.e. year_avg, month_avg, year_credit, and month_credit.

Some of the variables we created permanently in manipulating the dataset i.e. e_t_num - Employment Type in numbers (1 - Self-employed, 2 - Salaried, 3 - Others), Avg_acc_age_in_months, Credit_hist_len_in_months, Salaried, Self Employed, and Age_yrs.

Below Fig 1. is showing all the data types of the variables in the dataset.

There are 8-character variables and 37 numerical variables.

#	Variable	Type	Len	Format	Informat	Label
34	AVG_A_A	Char	10	\$10.	\$10.	AVG_A_A
2	A_D	Num	8	BEST.		A_D
41	Avg_acc_age_in_months	Num	8	BEST.		Avg_acc_age_in_months
4	B_I	Num	8	BEST.		B_I
35	CREDIT_H_L	Char	10	\$10.	\$10.	CREDIT_H_L
16	CREDIT_SCORE	Num	8	BEST.		CREDIT_SCORE
7	CURRENT_PINCODE_ID	Num	8	BEST.		CURRENT_PINCODE_ID
44	Credit_hist_len_in_months	Num	8	BEST.		Credit_hist_len_in_months
33	DEL_A_I_L_S_M	Num	8	BEST.		DEL_A_I_L_S_M
14	DRIVING_FLAG	Num	8	BEST.		DRIVING_FLAG
10	D_D	Num	8	DDMMYY10.		D_D
8	D_O_B	Num	8	DDMMYY10.		D_O_B
11	EMPLOYEE_CODE_ID	Num	8	BEST.		EMPLOYEE_CODE_ID
9	E_T	Char	13	\$13.	\$13.	E_T
3	INCOME	Num	8	BEST.		INCOME
37	LOAN	Num	8	BEST.		LOAN
12	MOBILENO_AVL_FLAG	Num	8	BEST.		MOBILENO_AVL_FLAG
6	M_I	Num	8	BEST.		M_I
43	Month_Credit	Char	4	\$4.	\$4.	Month_Credit
32	NEW_A_I_L_S_M	Num	8	BEST.		NEW_A_I_L_S_M
36	NO_O_I	Num	8	BEST.		NO_O_I
15	PASSPORT_FLAG	Num	8	BEST.		PASSPORT_FLAG
19	PRI_A_A	Num	8	BEST.		PRI_A_A
21	PRI_C_B	Num	8	BEST.		PRI_C_B
23	PRI_D_A	Num	8	BEST.		PRI_D_A
30	PRI_I_A	Num	8	BEST.		PRI_I_A
18	PRI_N_O_A	Num	8	BEST.		PRI_N_O_A
20	PRI_O_A	Num	8	BEST.		PRI_O_A
22	PRI_S_A	Num	8	BEST.		PRI_S_A
17	P_C_S_D	Char	53	\$53.	\$53.	P_C_S_D
25	SEC_A_A	Num	8	BEST.		SEC_A_A
27	SEC_C_B	Num	8	BEST.		SEC_C_B
29	SEC_D_A	Num	8	BEST.		SEC_D_A
31	SEC_I_A	Num	8	BEST.		SEC_I_A
24	SEC_N_O_A	Num	8	BEST.		SEC_N_O_A
26	SEC_O_A	Num	8	BEST.		SEC_O_A
28	SEC_S_A	Num	8	BEST.		SEC_S_A
13	STATE_ID	Num	8	BEST.		STATE_ID
5	S_I	Num	8	BEST.		S_I
1	UID	Num	8	BEST.		UID
42	Year_Credit	Char	2	\$2.	\$2.	Year_Credit
45	age_year_int	Num	8	BEST.		age_year_int
38	e_t_num	Num	8	BEST.		e_t_num
40	month_avg	Char	4	\$4.	\$4.	month_avg
39	year_avg	Char	2	\$2.	\$2.	year_avg

Fig 1. Format of all the variables

Statistical Model

The purpose of using the logistic regression is to predict whether the loan should be given to that person or not. As our dependent variable is categorical (1-Yes & 0-No) and independent variables are the mixture of categorical and quantitative variables.

Logistic Regression

$Y = \text{Loan (1 = Yes)}$

$X = \text{credit_score, income, e_t_num, Avg_acc_age_in_months, Credit_hist_len_in_months, Age_year_int, pri_c_b, no_o_i, del_a_i_l_s_m and new_a_i_l_s_m.}$

Where,

Del_a_i_l_s_m - Delinquent Account in the Last Six Months

New_a_i_l_s_m - New Account in the Last Six Months

No_o_i - Number of enquiries

E_t_num - Employment type

Once we decide whether the person should get the Loan then our next target to predict how much amount it should be given to the consumer. For that, our dependent variable is a continuous variable and the rest of the variables are the mixture of categorical and continuous variables.

Multiple Regression

$Y = A_D$

$X = \text{INCOME, PRI_C_B, SEC_C_B, Avg_acc_age_in_months, age_year_int, Salaried \& SelfEmployed.}$

Where,

A_D - Amount Disbursed

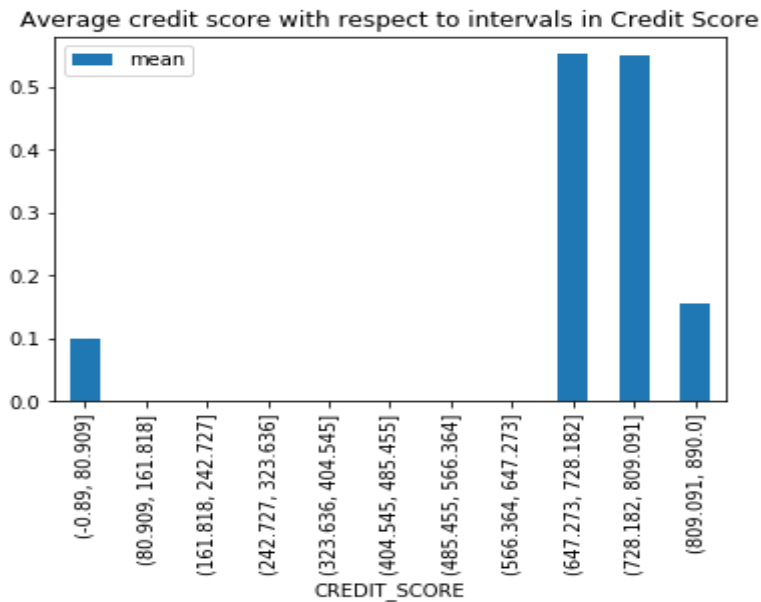
PRI_C_B - Primary Current Balance

SEC_C_B - Secondary Current Balance

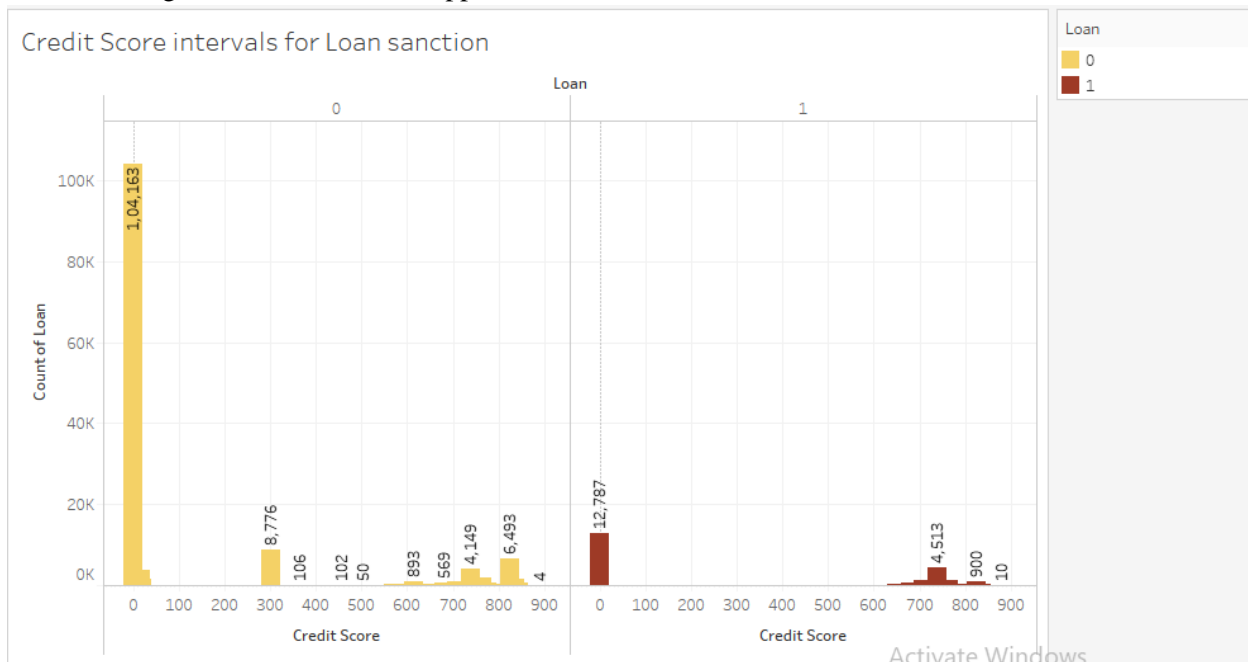
Major Findings

Descriptive Analysis

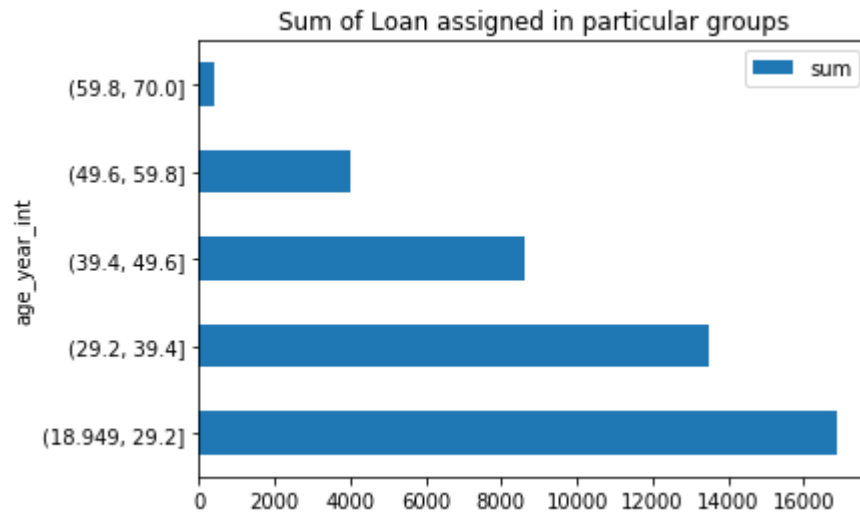
I. The minimum credit score that is required for loan approval is above 648, which can be noticed from the results below.



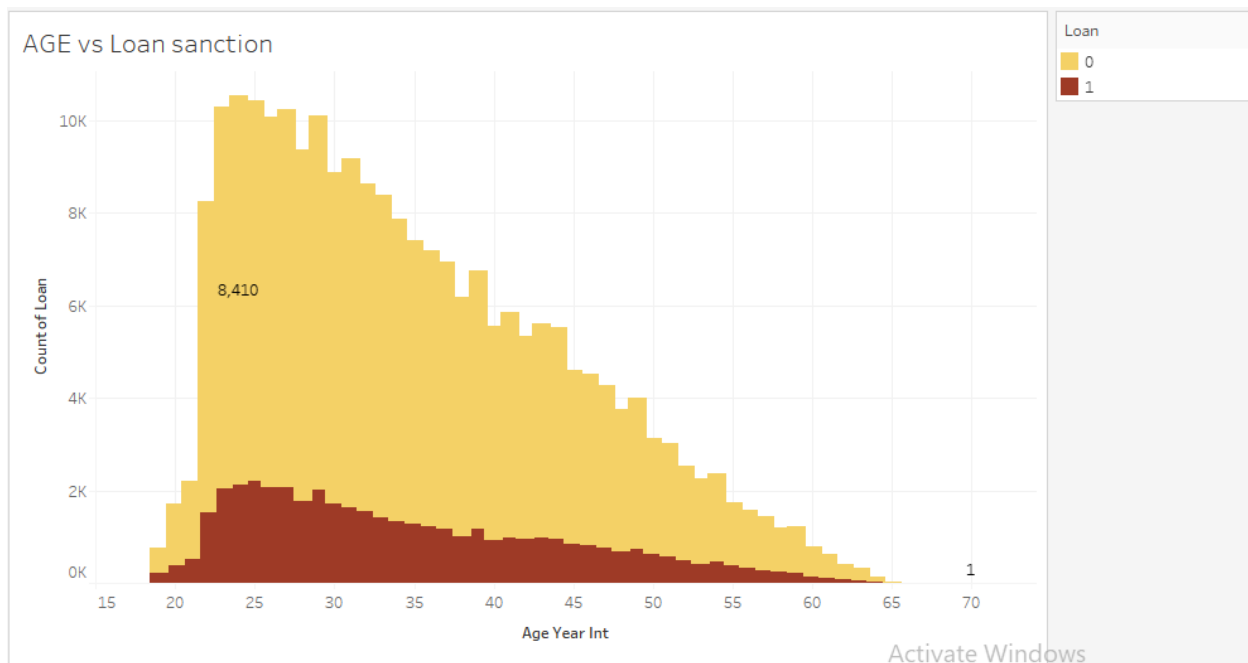
II. This gives the overall loans applied in several credit score intervals.



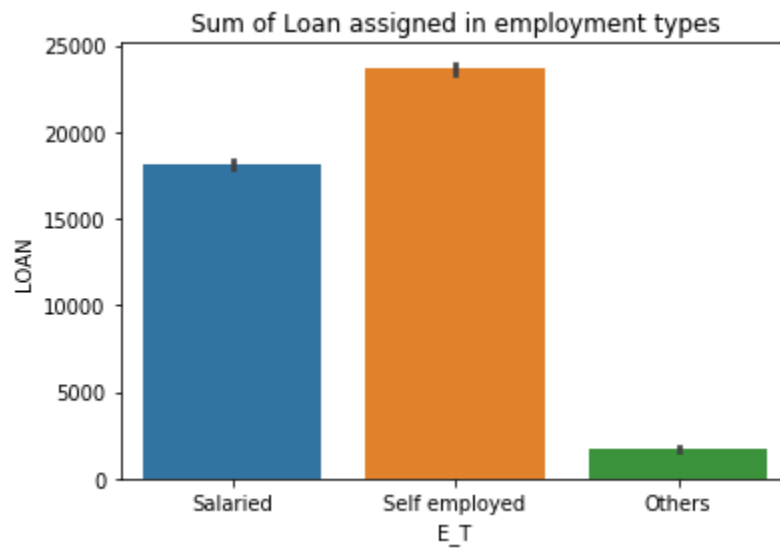
III. The chart below gives the no. of people who got a loan approved in the age groups shown below. The age group of 19-30 are the people who got the highest no. of loans approved.



IV.



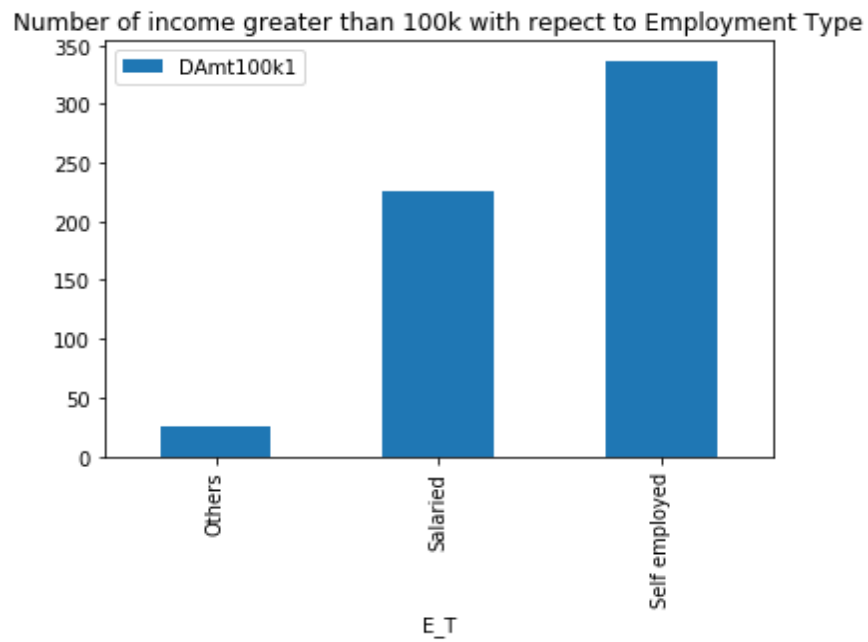
V. Self-employment type people are given the highest no.of loans.



VI. This is the comparison for Loan approvals and rejections of each employment type.



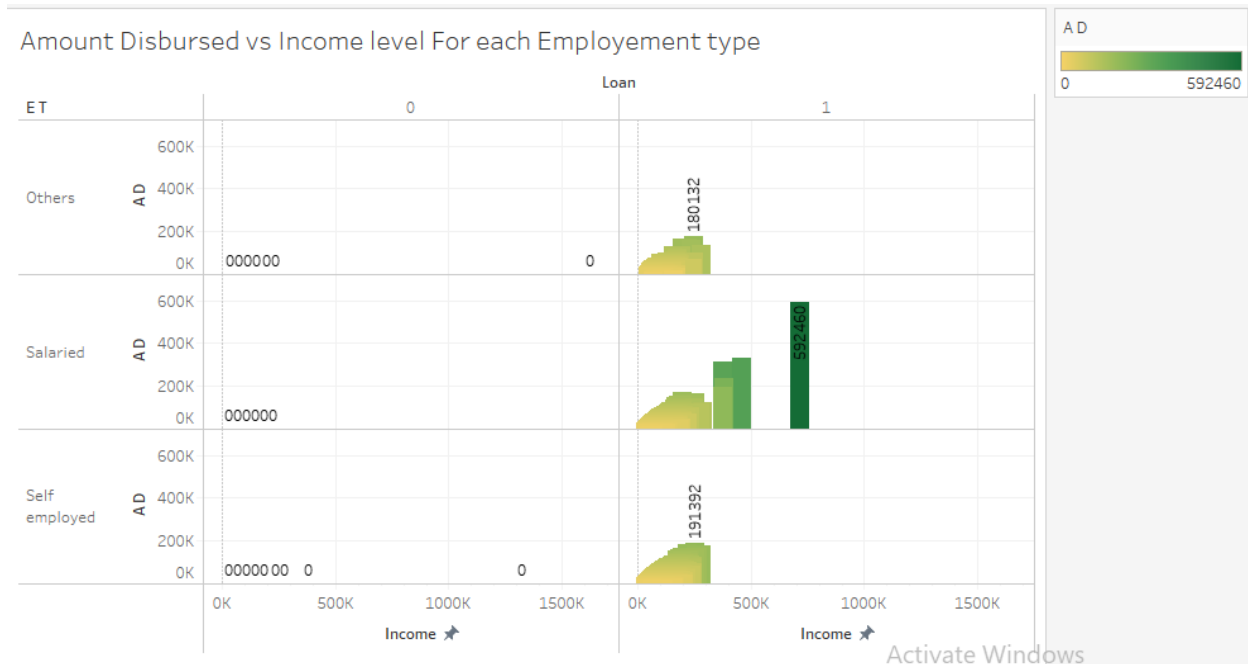
VII. Self-employment type people are more expected to get loans over \$100,000 than other types.



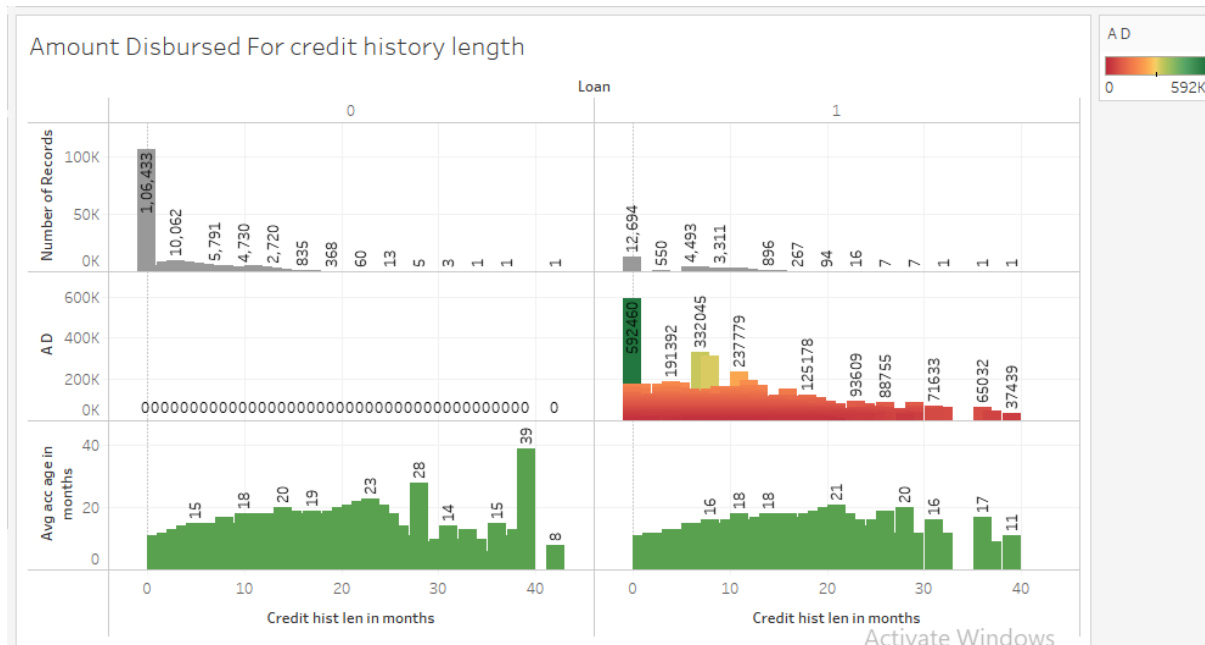
VIII. The chart below shows the amount of disbursed based on their income level. That shows a person who is getting a loan about \$590,000 is having an income of about \$750,000



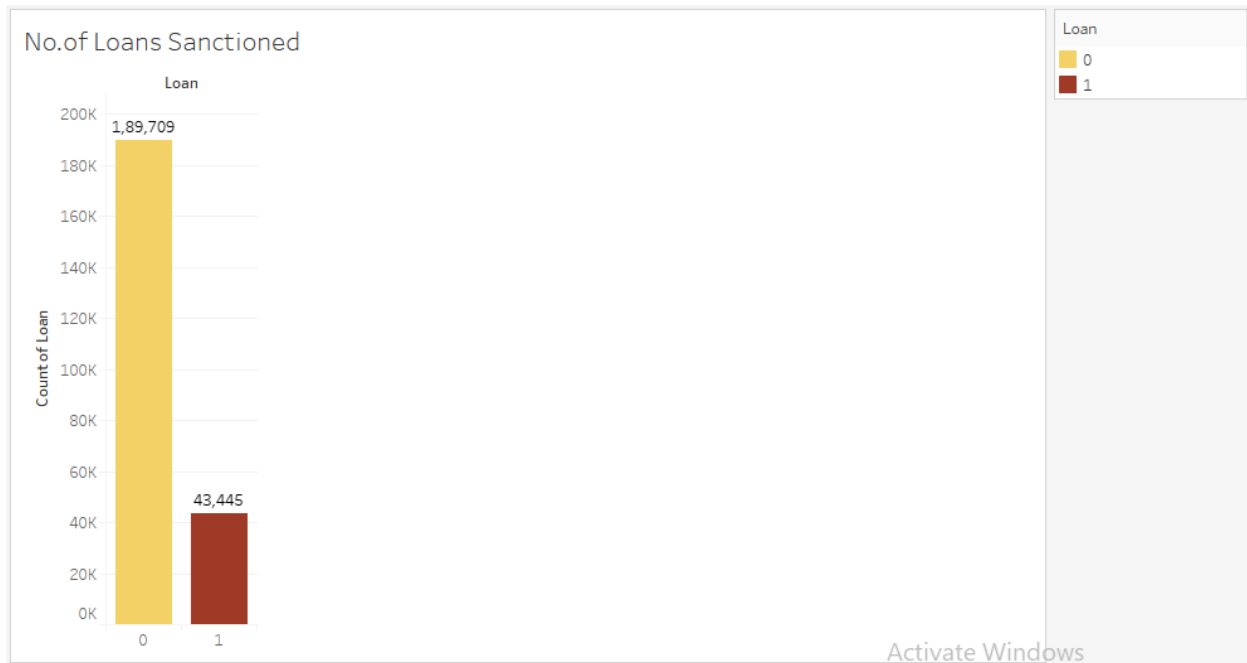
IX. The chart below shows the amount disbursed for each type of Employment type with respect to their income levels. It is observed that the person who got the highest loan amount falls under the Salaried type.



X. The chart below shows the amount of disbursed based on their credit history. It is observed that the person who got the highest loan has zero credit history. This would be because that person might be new to the country.



XI. This is the total no.of people who got loan approval and rejection. (1=approved, 0=rejected)



Relevant Findings

The below table shows finding the average income of the person who applied for the loan by categorizing the employment type. From the table, we can interpret that the average income of the employment type (1=Self Employed) who got the loan is \$ 92515 and did not get the loan is \$ 72953. Moreover, employment type (2=Salaried) who got the loan is \$ 86056 and did not get the loan is \$ 71752. We can take a look at the other values which might be helpful like minimum and maximum income is self-employed, salaried and others.

Table of Employment Type in which Loan has been sanctioned or not w.r.t Income

Analysis Variable : INCOME INCOME							
e_t_num	LOAN	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	0	103981	103981	72953	14397	37129	1328954
	1	23654	23654	92515	28998	38098	281164
2	0	79773	79773	71752	12142	37000	259000
	1	18085	18085	86056	27370	37000	715188
3	0	5955	5955	75941	24088	38881	1628992
	1	1708	1708	108522	26847	45514	282800

Fig 2. Statistics of different employment types assigned loans based on income.

This table shows the average credit score in which consumers can get the loan. It can be easily noticed that the value for not getting the Loan varies from 0 to 890 and getting the loan is vary from 0 to 879. We can say, the credit score is not only the variable which will affect on Sanctioning the Loan. Furthermore, people who got the loan are 18.6% as compared to the people who did not get the loan which is 81.4%.

To check average credit_score required to sanctioned Loan

Analysis Variable : CREDIT_SCORE CREDIT_SCORE						
LOAN	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	189709	189709	238	317	0	890
1	43445	43445	516	335	0	879

Fig 3. Investigating the minimum credit score to consider customer's eligibility

Statistical Analysis

Summary Statistics

Summary Statistics of the necessary variables						
The MEANS Procedure						
Variable	Label	N	Mean	Std Dev	Minimum	Maximum
INCOME	INCOME	233154	75885.1	18944.8	37000.0	1628992.0
CREDIT_SCORE	CREDIT_SCORE	233154	289.5	338.4	0.0	890.0
PRI_N_O_A	PRI_N_O_A	233154	2.4	5.2	0.0	453.0
PRI_A_A	PRI_A_A	233154	1.0	1.9	0.0	144.0
PRI_C_B	PRI_C_B	233154	165900.1	942273.6	-8678298.0	98524920.0
SEC_N_O_A	SEC_N_O_A	233154	0.1	0.6	0.0	52.0
SEC_A_A	SEC_A_A	233154	0.0	0.3	0.0	36.0
SEC_C_B	SEC_C_B	233154	5427.8	170237.0	-574647.0	36032852.0
PRI_I_A	PRI_I_A	233154	13105.5	151367.9	0.0	25642808.0
SEC_I_A	SEC_I_A	233154	323.3	15553.7	0.0	4170901.0
NEW_A_I_L_S_M	NEW_A_I_L_S_M	233154	0.4	1.0	0.0	35.0
DEL_A_I_L_S_M	DEL_A_I_L_S_M	233154	0.1	0.4	0.0	20.0
NO_O_I	NO_O_I	233154	0.2	0.7	0.0	36.0
Avg_acc_age_in_months	Avg_acc_age_in_months	233154	3.2	4.0	0.0	39.0
Credit_hist_len_in_months	Credit_hist_len_in_months	233154	3.6	4.7	0.0	42.0
age_year_int	age_year_int	233154	35.0	9.8	19.0	70.0

Fig 4. Summary Statistics to the problem variables

Frequency Distribution

Frequency distribution of Loan and Employment Type				
The FREQ Procedure				
Frequency Percent	Table of LOAN by e_t_num			
	e_t_num(e_t_num)			
LOAN(LOAN)	1	2	3	Total
0	103981 44.60	79773 34.21	5955 2.55	189709 81.37
1	23654 10.15	18085 7.76	1706 0.73	43445 18.63
Total	127635 54.74	97858 41.97	7661 3.29	233154 100.00

Fig 5. Frequency distribution of Loan and Employment Type

Chi-Square Test

Ho: No relationship between Loan and Employment Type

Ha: Relationship between Loan and Employment Type

Assumed,

Significance level < 0.05

Title: Loan Sanctioned of Employment type

Summary Statistics:

Frequency distribution of Loan and Employment Type

The FREQ Procedure

Frequency Percent	Table of LOAN by e_t_num				
	LOAN(LOAN)	e_t_num(e_t_num)			
		1	2	3	Total
0	103981	79773	5955	189709	
	44.80	34.21	2.55	81.37	
1	23854	18085	1708	43445	
	10.15	7.78	0.73	18.63	
Total	127835	97858	7661	233154	
	54.74	41.97	3.29	100.00	

Statistics for Table of LOAN by e_t_num

Statistic	DF	Value	Prob
Chi-Square	2	89.1324	<.0001
Likelihood Ratio Chi-Square	2	88.1026	<.0001
Mantel-Haenszel Chi-Square	1	14.8877	0.0001
Phi Coefficient		0.0172	
Contingency Coefficient		0.0172	
Cramer's V		0.0172	

Sample Size = 233154

Method: Chi-Square, 1-tail

Conclusion: Relationship between Loan and Employment Type ($p < .0001$)

Hypothesis Testing's

After analyzing the dataset, we would like to run multiple tests to check the significant difference in our proportion of the mean. Firstly, we believe that there is a significant difference between the people who got the loan based on salary equal to or more than 100,000. So,

Assuming that the credit score is normally distributed.

Chi-Square Test

Ho: No relationship between accepting Loan and Income (greater than 100,000)

Ha: Relationship between Loan and Income.

Title: Loan sanctioned with income greater than 100k.

Summary Statistics:

The FREQ Procedure

Frequency Percent	Table of LOAN by income_int			
	LOAN	income_int		Total
		0	1	
	0	185478 79.55	4231 1.81	189709 81.37
	1	25627 10.99	17818 7.64	43445 18.63
	Total	211105 90.54	22049 9.46	233154 100.00

Decision: Reject Ho

Conclusion: Significant difference between proportions for income groups (1=Greater than 100k) with respect to accepting loans. ($p < 0.0001$)

Our next hypothesis is to check the minimum credit score value which we believe is 648. So, we would like to run the chi-square test to verify our hypothesis.

Assuming that the credit score is normally distributed.

Ho: No relationship between accepting Loan and Credit score (greater than 648).

Ha: Relationship between Loan and Credit score.

Title: Loan sanctioned with a credit score greater than 648.

Summary Statistics:

The FREQ Procedure

Frequency Percent	Table of LOAN by credit_score_int			
		credit_score_int		
	LOAN	0	1	Total
	0	154679 66.34	35030 15.02	189709 81.37
	1	12787 5.48	30658 13.15	43445 18.63
	Total	167466 71.83	65688 28.17	233154 100.00

Decision: Reject Ho

Conclusion: Significant difference between proportions for credit score groups with respect to accepting loans. ($p < 0.0001$)

A final hypothesis is to check whether employment type (1=Self Employed) affecting the decision of giving the Loan or not. For that, we would like to run a hypothesis testing which will conclude our result.

Assuming that the credit score is normally distributed.

Ho: No relationship between the loan and self-employed (1 = Yes).

Ha: Relationship between loan and self-employed.

Assumed,

Significance level < 0.05

Title: Loan Sanctioned of employment type (1 = Self Employed)

Summary Statistics:

Summary statistics of Loan and Self Employed
The FREQ Procedure

Frequency Percent	Table of Self_Employed by LOAN			
	Self_Employed	LOAN		
		0	1	Total
	0	85728 36.77	19791 8.49	105519 45.26
	1	103981 44.60	23654 10.15	127635 54.74
	Total	189709 81.37	43445 18.63	233154 100.00

Statistics for Table of Self_Employed by LOAN

Statistic	DF	Value	Prob
Chi-Square	1	1.9002	0.1681
Likelihood Ratio Chi-Square	1	1.8997	0.1681
Continuity Adj. Chi-Square	1	1.8855	0.1697
Mantel-Haenszel Chi-Square	1	1.9002	0.1681
Phi Coefficient		-0.0029	
Contingency Coefficient		0.0029	
Cramer's V		-0.0029	

Decision: Fail to reject Ho.

Conclusion: No relationship between Loan and Self Employed. ($p=0.1681$)

T-Test

The use of the t-test is to check any significant difference between mean income with respect to Loan.

Ho: No significant difference between the mean income and loan.

Ha: Significant difference between the mean income and loan.

Summary Statistics:

The MEANS Procedure				
Loan_char=No				
Analysis Variable : INCOME				
N	Mean	Std Dev	Minimum	Maximum
185478	71358.02	10045.86	37000.00	99999.00

Loan_char=Yes				
Analysis Variable : INCOME				
N	Mean	Std Dev	Minimum	Maximum
17818	118709.08	22509.07	100000.00	715186.00

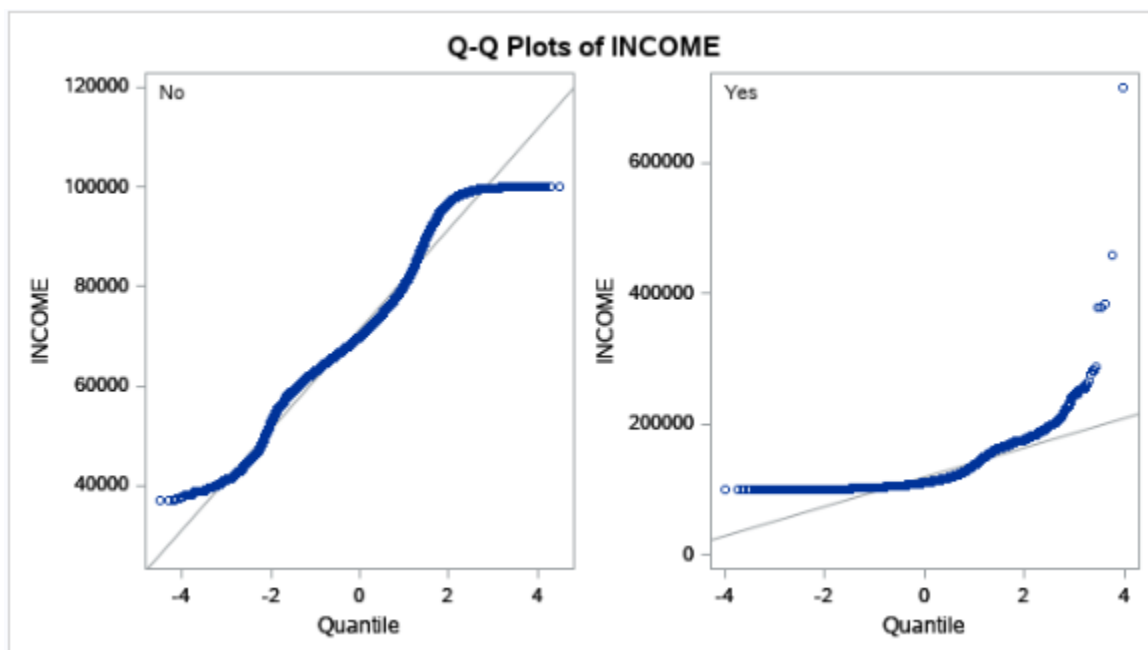
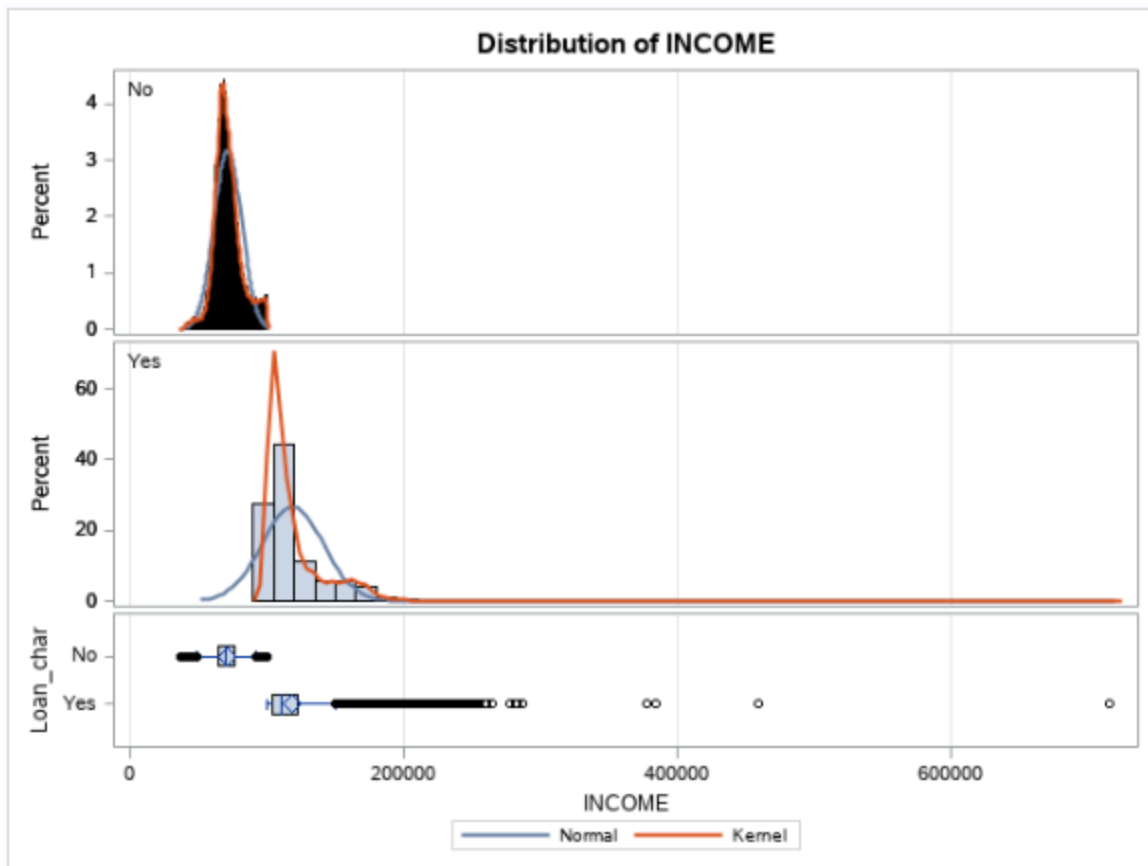
Results:

The TTEST Procedure							
Variable: INCOME							
Loan_char	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
No		185478	71358.0	10045.9	23.3260	37000.0	99999.0
Yes		17818	118709	22509.1	168.6	100000	715186
Diff (1-2)	Pooled		-47351.0	11682.4	91.6266		
Diff (1-2)	Satterthwaite		-47351.0		170.2		

Loan_char	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
No		71358.0	71312.3 71403.7	10045.9	10013.6 10078.3
Yes		118709	118379 119040	22509.1	22277.8 22745.2
Diff (1-2)	Pooled	-47351.0	-47530.6 -47171.4	11682.4	11646.6 11718.4
Diff (1-2)	Satterthwaite	-47351.0	-47684.7 -47017.4		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	203294	-516.78	<.0001
Satterthwaite	Unequal	18505	-278.15	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	17817	185477	5.02	<.0001



F-test (always 2-tail)

$$H_0: \sigma^2_y / \sigma^2_n = 1$$

$$H_a: \sigma^2_y / \sigma^2_n \neq 1$$

Decision: Reject H_0

Conclusion: Variances are not equal.

T-test

$$H_0: \mu_y - \mu_n = 0$$

$$H_a: \mu_y - \mu_n \neq 0$$

Decision: Reject H_0

Conclusion: Significant difference between Loan groups Yes and No with respect to mean income.
($p < 0.0001$)

T-Test

:

The purpose of using a t-test is to check whether any significant difference between mean credit score and loan.

Ho: No significant difference between the mean credit score and loan.

Ha: Significant difference between the mean credit score and loan.

Summary Statistics:

The MEANS Procedure				
Loan_char=No				
Analysis Variable : CREDIT_SCORE				
N	Mean	Std Dev	Minimum	Maximum
88	259.6511628	310.1263411	0	836.0000000

Loan_char=Yes				
Analysis Variable : CREDIT_SCORE				
N	Mean	Std Dev	Minimum	Maximum
14	529.8571429	350.4491310	0	825.0000000

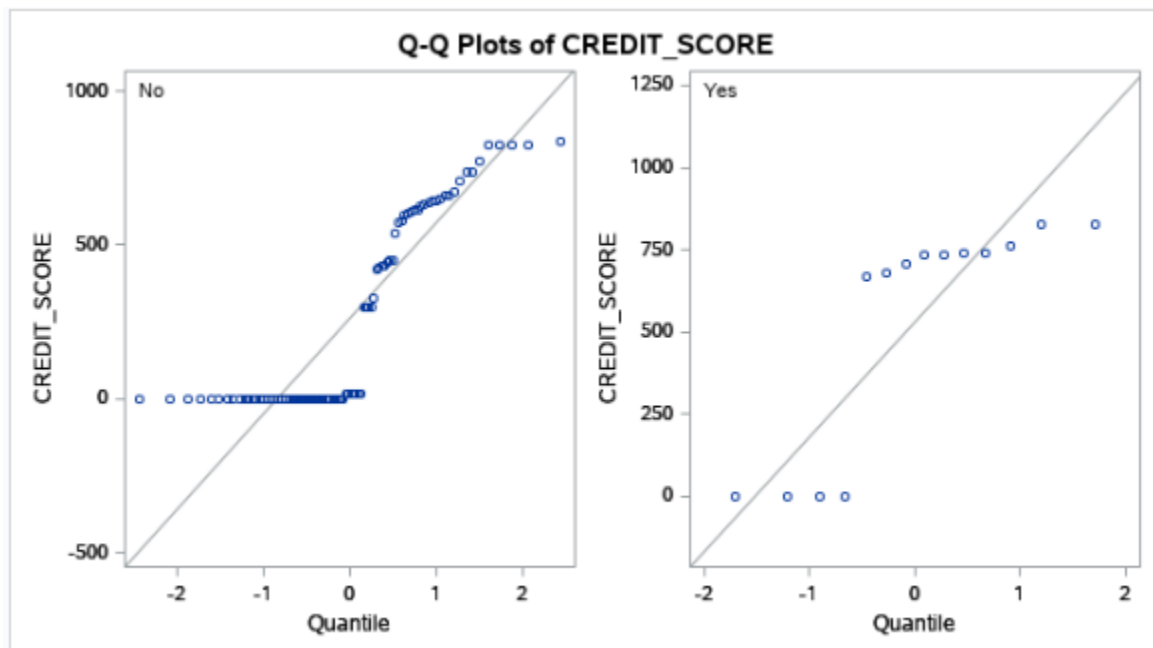
Results:

The TTEST Procedure							
Variable: CREDIT_SCORE							
Loan_char	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
No		88	259.7	310.1	33.4418	0	836.0
Yes		14	529.9	350.4	93.6615	0	825.0
Diff (1-2)	Pooled		-270.2	315.8	91.0039		
Diff (1-2)	Satterthwaite		-270.2		99.4526		

Loan_char	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
No		259.7	193.2 326.1	310.1	289.7 364.9
Yes		529.9	327.5 732.2	350.4	254.1 564.6
Diff (1-2)	Pooled	-270.2	-450.8 -89.6118	315.8	277.1 367.1
Diff (1-2)	Satterthwaite	-270.2	-480.5 -59.8787		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	98	-2.97	0.0038
Satterthwaite	Unequal	16.485	-2.72	0.0149

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	13	85	1.28	0.4858



F-test (always 2-tail)

$$H_0: \sigma^2_y / \sigma^2_n = 1$$

$$H_a: \sigma^2_y / \sigma^2_n \neq 1$$

Decision: Fail to reject H_0

Conclusion: Variances are equal.

T-test

$$H_0: \mu_y - \mu_n = 0$$

$$H_a: \mu_y - \mu_n \neq 0$$

Decision: Reject H_0

Conclusion: Significant difference between Loan groups Yes and No with respect to mean credit score.

($p=0.0038$)

Model

Logistic Regression

Title: Predicting Loan sanctioned for the vehicle loan.

Summary Statistics:

Variables	N (%)	Mean	Std. Dev.	Range
Credit_score	233154 (100)	289.46	338.37	0 - 890
NEW_A_I_L_S_M	233154 (100)	0.38	0.96	0 - 35
DEL_A_I_L_S_M	233154 (100)	0.10	0.38	0 - 20
Avg_acc_age_in_monhs	233154 (100)	3.17	4.04	0 - 39
Credit_hist_len_in_months	233154 (100)	3.65	4.81	0 - 42
Age_year_int	233154 (100)	35.01	9.81	19 - 70

Model:

Log odds (Loan = Yes) = - 8.31 + 0.000074 * Income + 0.003 * Credit Score + 0.164 * New_a_i_l_s_m + (-4.728) * Del_A_i_l_s_m + (-0.146) * Avg_acc_age_in_months + 0.205 * Credit_hist_len_in_months + (-0.015) * age_year_int

Odds Ratio (95 % CI) with interpretation:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
INCOME	1.000	1.000	1.000
CREDIT_SCORE	1.003	1.003	1.003
NEW_A_I_L_S_M	1.179	1.164	1.194
DEL_A_I_L_S_M	0.009	0.008	0.010
Avg_acc_age_in_month	0.864	0.860	0.869
Credit_hist_len_in_m	1.227	1.222	1.232
age_year_int	0.985	0.983	0.986

Interpretation:

Income - For every one dollar (\$) increase in income the likelihood of Loan = Yes increases by 1.000.

Credit_Score - For every one unit increase in credit score the likelihood of Loan = Yes increases by 1.003.

New_A_I_L_S_M - For every one unit increase in new accounts in the last six months the likelihood of Loan = Yes increases by 1.179.

DEL_A_I_L_S_M - For every one unit increase in delinquent account in the last six months the likelihood of Loan = Yes decreases by 0.009.

Avg_acc_age_in_months - For every one unit increase in average account age in months the likelihood of Loan = Yes decreases by 0.864.

Credit_hist_len_in_months - For every one unit increase in credit history length in months the likelihood of Loan = Yes increases by 1.227.

Age_year_int - For every one unit increase in age the likelihood of Loan = Yes decreases by 0.985.

Multiple Regression

Title: Predicting disbursed amount for the sanctioned loan.

Summary Statistics:

Variables	N (%)	Mean	Std. Dev	Range
Income	43445 (100)	90376	28601	37000 - 715186
PRI_C_B	43445 (100)	259652	860654	-99572 - 32027420
SEC_C_B	43445 (100)	7352	257964	-117138 - 36032852
Avg_acc_age_in_months	43445 (100)	4.75	7.09	0 - 27
Age_year_int	43445 (100)	34.69	10.01	19 - 65
Salaried (1 = Yes)	18085 (41.63)			
Self Employed (1 = Yes)	23654 (54.45)			

Model:

Estimated Amount Disbursed = $18349.25 + 0.48 * \text{Income}$

Fit of the model:

Adjusted R-square = 0.6481

That means, 64.81% of the variability in the amount disbursed is explained by income.

Marketing Implications and Conclusions

Marketing Implications

According to the research, our team finds out major variables that need to be considered in sanctioning the loan and amount. The marketing research evaluates credit score as the top priority in assigning loans. Credit scores evaluate the ability of the people to repay their loan amount. Moreover, less credit score leads towards high-interest rates or vice versa.¹ The income which is used to generate new parameters such as the ratio of debt to income is giving the idea to the lender either you are paying your previous loans constantly or not.² If you have high income with a high debt then the bank will consider as a negative attitude towards repayment of the loan and it will decrease your chances of getting the loan. The LTV (Loan to Value), meaning the amount requested by the borrower divided by the value of the vehicle. Bank compare based on higher the ratio value more the risk that the borrower will use extra money to repay other loans.³ Additionally, this gives an extra hint to check other variables like how many loan borrowers already taken from other places.

¹ "9 Factors to Consider Before Taking Out a Car Loan - Tweak" 6 Sep. 2019, <https://tweakyourbiz.com/finance/financial-planning/car-loan-factors>. Accessed 23 Nov. 2019.

² "Factors that affect your auto loan | RoadLoans." 21 Aug. 2014, <https://roadloans.com/blog/factors-that-affect-your-auto-loan>. Accessed 23 Nov. 2019.

³ "12 Credit Union Factors That Affect Your Auto Loan Payments." 27 Mar. 2018, <https://www.weokie.org/blog/-12-credit-union-factors-that-affect-your-auto-loan-payments->. Accessed 23 Nov. 2019.

Conclusion

The existing variables in our datasets and with the internet search, our team can conclude that the bank needs to consider several factors in predicting whether the borrower will repay the loan. Or, how much amount sanctioned to the borrower which will give high chances to repay the loan.

From our model, we can conclude that the borrower's with the employment type (1 = Self Employed, 2 = Salaried or 3 = Others), credit score, primary current balance, new account in the last six months, delinquent account in the last six months, number of enquiries, average account age, credit history length and age consider to be the important independent variable to predict lending a Loan or not. In our model, the new account in the last six months defines as the new loans taken by the borrower's in the last six months and delinquent account in the last six months defines as the loan defaulted by the borrower's in the last six months.

Model:

$$\text{Log odds (Loan = Yes)} = -8.31 + 0.000074 * \text{Income} + 0.003 * \text{Credit Score} + 0.164 * \text{New_a_i_l_s_m} + (-4.728) * \text{Del_A_i_l_s_m} + (-0.146) * \text{Avg_acc_age_in_months} + 0.205 * \text{Credit_hist_len_in_months} + (-0.015) * \text{age_year_int}$$

For sanctioning the loan amount, we got the linear regression model. In the Linear Regression model, we got income as the best predictor after rejecting several independent variables. Our model is explaining 64.81% of the variability in the amount disbursed by the income variable. As the sum of the squared error of the residual is bigger 4.38×10^{12} .

Model:

$$\text{Estimated Amount Disbursed} = 18349.25 + 0.48 * \text{Income}$$

Accuracy Predictor:

Accuracy of logistic regression classifier on test set is 0.81.

Appendix

Loan can be rejected once the disbursed amount has been given to the customer.

Once the loan amount sanctioned to the customer there is a possibility to cancel the loan. There are several factors which impact on cancelling the loan decision.

Seller background: This is an important factor which banks considered in investigating during the vehicle loan sanction. If the seller has a criminal history, then bank will hold the sanctioned amount given to the customers. There might be a possibility that the vehicle is not under the seller's name. So, banks consider the history of seller's as an important aspect in evaluating the loan decision.



Income: The income submitted by the customer during the vehicle loan has been analyzed. Banks needs to figure out whether the income is a part of full-time work or an extra income i.e. rental, interest etc. Moreover, banks need to verify their income because it is an important factor in deciding the sanctioned amount. More the income bank is likely to sanction more amount to the customer. So, bank needs to investigate this variable thoroughly.

Delinquent: This is the history of the past loan payments of the customer who applied for other loan. This is also one of the crucial factors in rejecting the application. If the customer has missed a payment in the first 6 months, then it is more likely that the customer will miss this loan payment also. So, banks look at their default history in previous loans and decide whether to give the loan or not.

Overall Efforts

Our team consisting of two members and area of interest is to know about the evaluation of Loan sanctioned. The main reason for choosing the dataset is to analyze and understand the procedure of eliminating the unnecessary variable and keep necessary variable in the Loan sanction problem. Initially, we were concentrating on getting the variables which will be a good predictor for Loan Sanctioned. When we gave our first presentation, professor (Dr. Kanghyun Yoon) suggested to look more questions from the dataset i.e how much amount to be sanctioned to the consumer. For Loan predictors, we searched on the internet for predicting loan which variable plays an important role. After searching on the internet, we came to know a few of the independent variables which we found in our dataset. Using those variables, we predicted Loan Sanctioned. After discussing with a team member, we decided to use only Loan (1=Yes) for predicting disbursed amount. The reason for choosing Loan (1=Yes) only because the amount will be disbursed once the person will be eligible for Loan. For disbursed amount, we had taken important variables after searching on the internet and some of them from our own guess as an independent variable. We presented our model with the help of Multiple Regression.

SAS Codes

 ---- shows the comments.
 ---- shows the codes.

***** Data Manipulation using SAS *****

/* Reading the original dataset which we downloaded from the Internet */

```
proc import datafile='/folders/myfolders/main_file_formatted.csv'  
dbms=csv out=project replace;  
run;
```

/* Manipulating the dataset by using different procedure in the SAS */

```
data project1 (drop= UID e_t p_c_s_D A_D B_I S_I M_I CURRENT_PINCODE_ID D_O_B D_D  
EMPLOYEE_CODE_ID      MOBILENO_AVL_FLAG      STATE_ID      DRIVING_FLAG  
      PASSPORT_FLAG PRI_O_A PRI_S_A PRI_D_A SEC_O_A SEC_S_A SEC_D_A  
year_avg month_avg year_credit month_credit);  
set project;
```

/* e_t has missing values filled by Others */

```
if e_t = ' ' then e_t = 'Others';
```

/* Converting employment_type to numeric */

```
if e_t = "Salaried" then e_t_num = 2;  
else if e_t = "Self employed" then e_t_num = 1;  
else e_t_num = 3;
```

/* Converting Avg Acc. Age to proper format */

```
year_avg = scan(avg_a_a,1,' ');  
month_avg = scan(avg_a_a,2,' ');  
length year_avg month_avg $10.;  
year_avg = tranwrd(year_avg,'yrs','');  
month_avg = tranwrd(month_avg,'mon','');  
month_avg = tranwrd(month_avg,'mo','');  
Avg_acc_age_in_months = sum(year_avg,month_avg);
```

/* Converting Credit_history_length into proper format */

```
Year_Credit = scan(credit_h_l,1,' ');
```

```

Month_Credit = scan(credit_h_l,2,' ');
length year_credit month_credit $10.;
Year_Credit = tranwrd(Year_Credit,'yrs','');
Month_Credit = tranwrd(Month_Credit,'mon','');
Month_Credit = tranwrd(Month_Credit,'mo','');
Credit_hist_len_in_months = sum(Year_Credit, Month_Credit);

```

/* Generating Age column in the table */

```

Age_yrs = yrdif(d_o_b,today(),'ACT/ACT');
Age_yrs = int(age_yrs);
run;

```

***** Data Manipulation using Excel *****

/* Reading the same dataset but with the manipulation in the Excel */

```

proc import datafile='/folders/myfolders/project_file2.csv'
dbms=csv out=project2 replace;
run;

```

/* Creating binary variable of the Employment Type for further analysis */

```

data project_final (drop= UID e_t p_c_s_D B_I S_I M_I CURRENT_PINCODE_ID D_O_B D_D
EMPLOYEE_CODE_ID
MOBILENO_AVL_FLAG STATE_ID DRIVING_FLAG PASSPORT_FLAG
PRI_O_A PRI_S_A PRI_D_A
SEC_O_A SEC_S_A SEC_D_A Avg_A_A Credit_h_l year_avg
month_avg year_credit month_credit);
set project2;

```

```

if e_t_num = 1 then Self_Employed = 1;
else Self_Employed = 0;

```

```

if e_t_num = 2 then Salaried = 1;
else Salaried = 0;

```

```

if e_t_num = 3 then Others = 1;
else Others = 0;

```

```

if Loan = 1 then Loan_char = 'Yes';
else Loan_char = 'No';

```

```
run;
```

***** Summary Statistics *****

/* Statistics for Continuous variables */

```
proc means data=project_final maxdec=1;
var INCOME CREDIT_SCORE PRI_N_O_A PRI_A_A PRI_C_B SEC_N_O_A SEC_A_A SEC_C_B
PRI_I_A SEC_I_A NEW_A_I_L_S_M
DEL_A_I_L_S_M NO_O_I Avg_acc_age_in_months Credit_hist_len_in_months age_year_int;
title "Summary Statistics of the necessary variables";
run;
```

/* Statistics for Categorical Variable And Chi-Square test */

```
proc freq data=project_final;
tables LOAN*e_t_num/chisq norow nocol;
title "Frequency distribution of Loan and Employment Type";
run;
```

***** Relevant Findings *****

/* To see which employment type are getting loan, average income, min & max income. */

```
ods noproctitle;
proc means data=project_final maxdec=0;
class e_t_num loan;
var income;
title 'Table of Employment Type in which Loan has been sanctioned or not w.r.t Income';
run;
```

/* To check the average credit score required to pass the loan */

```
ods noproctitle;
proc means data=project_final maxdec=0;
class loan;
```

```
var credit_score;  
title 'To check average credit_score required to sanctioned Loan';  
run;
```

***** Hypothesis Testings *****

/* Randomly picking 100 observations from the dataset */

```
proc surveyselect data=project_final  
    method=srs n=100 out=SampleSRS;  
run;
```

/* Keeping income and loan column to measure the effect of income on loan */

```
data salary;  
set samplesrs;  
keep income loan_char;  
run;
```

```
proc sort data=salary; by loan_char;  
proc means data=salary; by loan_char; var income;  
proc ttest data=salary; class loan_char; var income;  
run;
```

/* Keeping credit score and loan to measure an effect on loan */

```
data credit_score;  
set samplesrs;  
keep credit_score loan_char;  
run;
```

```
proc sort data=credit_score; by loan_char;  
proc means data=credit_score; by loan_char; var credit_score;  
proc ttest data=credit_score; class loan_char; var credit_score;
```

```
run;
```

```
/* Keeping loan and self employed to measure a relationship on loan */
```

```
proc freq data = project_final;  
tables self_employed * loan/chisq norow nocol;  
title 'Summary statistics of Loan and Self Employed';  
run;
```

```
***** Multiple Regression *****
```

```
/* To retrieve all the rows with the Loan Sanctioned = 'Yes' or 1 */
```

```
data regression_dataset;  
set project_final;  
    if loan = 1;  
run;
```

```
/* Selecting 100 random observations from the regression dataset */
```

```
proc surveyselect data=regression_dataset  
    method=srs n=100 out=SampleSRS1;  
run;
```

```
/* Summary Statistics & correlation table */
```

```
proc means data=regression_dataset; var INCOME PRI_C_B SEC_C_B  
Avg_acc_age_in_months age_year_int Salaried Self_Employed;
```

```
proc freq data=regression_dataset; tables salaried*self_employed/chisq;  
proc corr data=regression_datase; var INCOME PRI_C_B SEC_C_B Avg_acc_age_in_months  
age_year_int Salaried Self_Employed;
```

/* Initializing the Stepwise process */

```
proc reg data = samplesrs1;  
model A_D = INCOME PRI_C_B SEC_C_B Avg_acc_age_in_months age_year_int Salaried  
Self_Employed /selection=stepwise;
```

```
proc reg data = samplesrs1;  
model A_D = INCOME Salaried age_year_int Avg_acc_age_in_months Self_Employed  
PRI_C_B/selection=adjrsq;
```

```
proc reg data = samplesrs1;  
model A_D = INCOME/R influence;  
    plot A_D*Income;  
    plot r.*npp.;  
Run;
```

Python Codes

```
[1]: # Importing all the necessary libraries.....
import pandas as pd
import numpy as np
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
pd.options.display.max_columns= None
```

0.0.1 Reading the main dataset file

```
[2]: ma=pd.read_csv('project_file2.csv')
```

```
[4]: ma.shape
```

```
[4]: (233154, 51)
```

0.0.2 Preparing bins for 'CREDIT_SCORE' and finding the minimum CREDIT_SCORE required to get a loan.

```
[8]: # Dividing the column into number of bins....
credit_cut=pd.cut(ma['CREDIT_SCORE'],11)
#Grouping to categorize credit score according to LOAN....
credit=ma.groupby(credit_cut)['LOAN'].agg(['mean'])
credit.reset_index()
# Plotting
credit.plot(kind='bar',title='Average credit score with respect to intervals in_
↳Credit Score')
```

0.0.3 Preparing bins for 'AGE' and finding the age group which is getting more loans sanctioned.

```
[10]: # Dividing the column into number of bins.
age_cut=pd.cut(ma['age_year_int'],5)
# Grouping and plotting the data which shows Loan according to different age_
↳groups.
ma['age_cut']=age_cut
ma['Loan1'] = np.squeeze(np.asarray(ma['LOAN']))
ma.groupby(age_cut)['LOAN'].agg(['sum']).plot(kind='barh',title='Sum of Loan_
↳assigned in particular groups')
```

```
[11]: # Plotting number of loan distributed in employment type.
sns.barplot(x='E_T',y='LOAN', data=ma, estimator=np.sum).set_title('Sum of Loan_
↳assigned in employment types')
```



```
[12]: # Grouping employment type & loan to get the count of Loan (1=Yes).
ax = ma.groupby(['E_T', 'LOAN']).agg({'LOAN': 'count'})
ax.rename(columns={"LOAN": 'LOAN_COUNT'}, inplace=True)
ax = ax.reset_index()
ax
# Plotting to visualise number of loan sanctioned to each employment type.
sns.barplot(x='E_T', y='LOAN_COUNT', data=ax, hue='LOAN').set_title('Number of
↳Loans assigned to each Employment Type')
```

0.0.4 Finding which employment type people are getting many loans over \$100K

```
[13]: # Dividing income into two different groups (greater than or equal to 100k &
↳less than 100k)
ma['Damt100k'] = pd.cut(ma['A_D'], [-1, 100000, ma['A_D'].max()], labels=[0, 1])
# Plotting the different group to show salary greater than 100k & their
↳repective employment type.
ma['Damt100k1'] = np.squeeze(np.asarray(ma['Damt100k']))
ma.groupby('E_T').agg({'Damt100k1': 'sum'})\
    .plot(kind='bar', title='Number of income greater than 100k with repect to
↳Employment Type')
```

Accuracy predictor code

```
import statsmodels.api as smf
logit_model = smf.Logit(y, X)
result = logit_model.fit()
result.summary()
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier, export_graphviz
```

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, test_size=0.3, random_state=0)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
```

```
y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))
```