Capstone Project:
Advanced Data analytics
CRISTIAN TORRES BARON

Project Overview:

This capstone analyzes Lending Club loans to support an investment decision. An investment firms plans to invest $10M and hired me to evaluate **loan quality and risk** before selecting loans.

Dataset: (downloaded from Lending Club loan data). Scope is limited to **year 2015**, **36-month loans**, and **credit card debt purpose**. The dataset contains **7,151 loans** and **16 variables** (borrower, loan, and credit attributes).

Column	Description
id	A unique LC assigned ID for the loan listing.
member_id	A unique LC assigned ID for the borrower member.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
purpose	A category provided by the borrower for the loan request.
loan_status	Current status of the loan
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
acc_open_past_24mths	Number of trades opened in past 24 months.
total_pymnt	Payments received to date for total amount funded

Specific problem to solve:
Identify which loans are most suitable for investment by predicting default risk and estimating expected losses/returns, so the firm can build a loan portfolio that **maximizes expected return while controlling risk**.

What this project delivers:

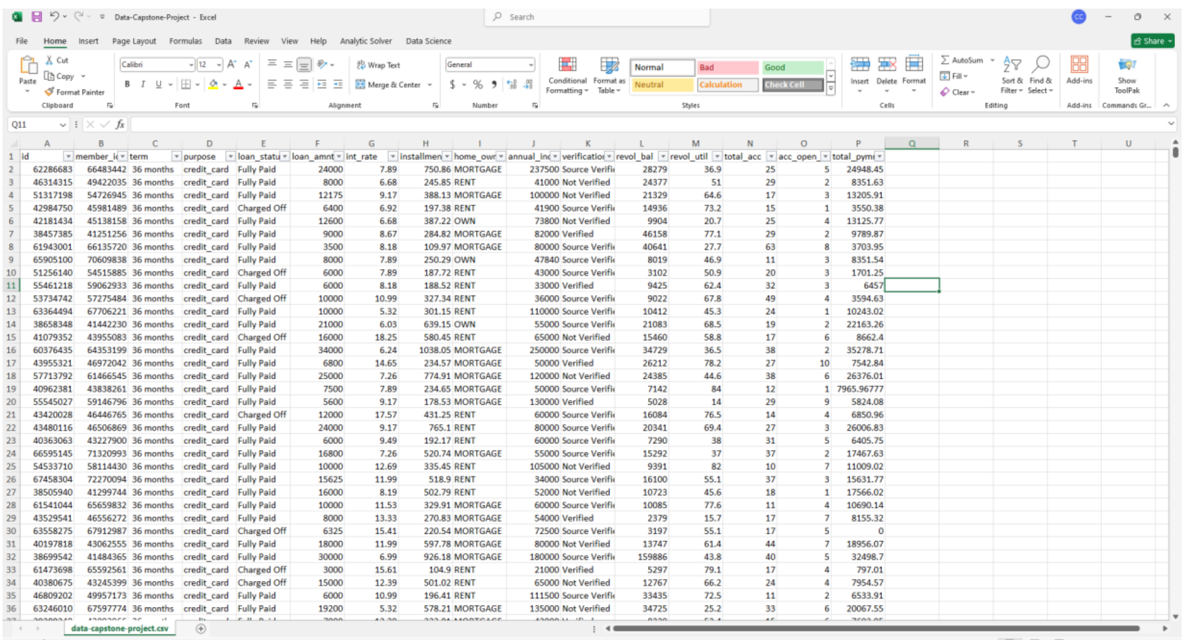
- Exploratory analysis to understand borrower/loan patterns and risk drivers
- Data cleaning + feature engineering based on the provided variables/data dictionary
- A predictive model (risk scoring) to estimate probability of default / loan quality
- A portfolio selection recommendation for allocating the **$10M** toward lower-risk, higher-expected-value loans\

Step 1 – Data preparation & Exploratory Analysis

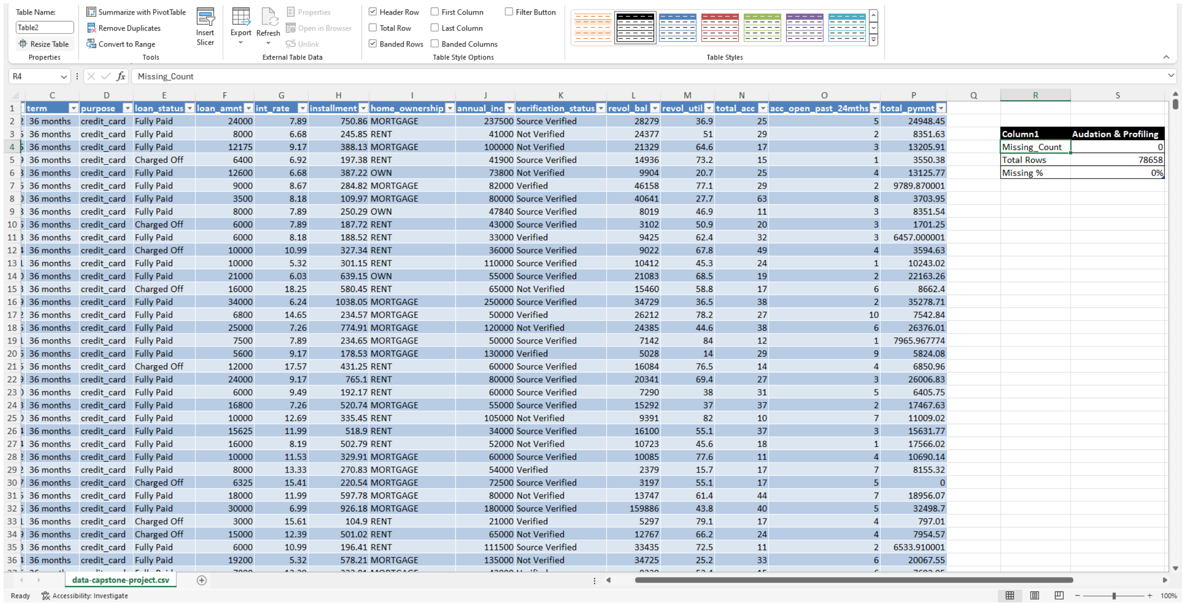
This step covers the first phase of the capstone: prepare the Lending Club 2015 dataset and run initial exploratory analysis to confirm the data is reliable for modeling default risk and building a $10M investment portfolio. Because the dataset comes from a public source and is already structured, the focus is on validation, cleaning, and understanding risk patterns.

CHANGELOG DATASET INGEST, VALIDATION, CLEANING:

Ingest the RAW data



Profiling Dataset if contain Null Values



	term	purpose	loan_status	loan_amnt	int_rate	installment	home_ownership	annual_inc	verification_status	revol_bal	revol_util	total_acc	acc_open_past_24mths	total_pymnt
2	36 months	credit_card	Fully Paid	24000	7.89	750.86	MORTGAGE	237500	Source Verified	28279	36.9	25	5	24948.45
3	36 months	credit_card	Fully Paid	8000	6.68	245.85	RENT	41000	Not Verified	24377	51	29	2	8351.63
4	36 months	credit_card	Fully Paid	12175	9.17	388.13	MORTGAGE	100000	Not Verified	21329	64.6	17	3	13205.91
5	36 months	credit_card	Charged Off	6400	6.92	197.38	RENT	41900	Source Verified	14936	73.2	15	1	3550.38
6	36 months	credit_card	Fully Paid	12600	6.68	387.22	OWN	73800	Not Verified	9904	20.7	25	4	13125.77
7	36 months	credit_card	Fully Paid	9000	8.67	284.82	MORTGAGE	82000	Verified	46158	77.1	29	2	9789.870001
8	36 months	credit_card	Fully Paid	3500	8.18	109.97	MORTGAGE	80000	Source Verified	40641	27.7	63	8	3703.95
9	36 months	credit_card	Fully Paid	8000	7.89	250.29	OWN	47840	Source Verified	8019	46.9	11	3	8351.54
10	36 months	credit_card	Charged Off	6000	7.89	187.72	RENT	43000	Source Verified	3102	50.9	20	3	1701.25
11	36 months	credit_card	Fully Paid	6000	8.18	188.52	RENT	33000	Verified	9425	62.4	32	3	6457.000001
12	36 months	credit_card	Charged Off	10000	10.99	327.34	RENT	36000	Source Verified	9022	67.8	49	4	3594.63
13	36 months	credit_card	Fully Paid	10000	5.32	301.15	RENT	110000	Source Verified	10412	45.3	24	1	10243.02
14	36 months	credit_card	Fully Paid	21000	6.03	639.15	OWN	55000	Source Verified	21083	68.5	19	2	22163.26
15	36 months	credit_card	Charged Off	16000	18.25	580.45	RENT	65000	Not Verified	15460	58.8	17	6	8662.4
16	36 months	credit_card	Fully Paid	34000	6.24	1038.05	MORTGAGE	250000	Source Verified	34729	36.5	38	2	35278.71
17	36 months	credit_card	Fully Paid	6800	14.65	234.57	MORTGAGE	50000	Verified	26212	78.2	27	10	7542.84
18	36 months	credit_card	Fully Paid	25000	7.26	774.91	MORTGAGE	120000	Not Verified	24385	44.6	38	6	26376.01
19	36 months	credit_card	Fully Paid	7500	7.89	234.65	MORTGAGE	50000	Source Verified	7142	84	12	1	7965.967774
20	36 months	credit_card	Fully Paid	5600	9.17	178.53	MORTGAGE	130000	Verified	5028	14	29	9	5824.08
21	36 months	credit_card	Charged Off	12000	17.57	431.25	RENT	60000	Source Verified	16084	76.5	14	4	6850.96
22	36 months	credit_card	Fully Paid	24000	9.17	765.1	RENT	80000	Source Verified	20341	69.4	27	3	26006.83
23	36 months	credit_card	Fully Paid	6000	9.49	192.17	RENT	60000	Source Verified	7290	38	31	5	6405.75
24	36 months	credit_card	Fully Paid	16800	7.26	520.74	MORTGAGE	55000	Source Verified	15292	37	37	2	17467.63
25	36 months	credit_card	Fully Paid	10000	12.69	335.45	RENT	105000	Not Verified	9391	82	10	7	11009.02
26	36 months	credit_card	Fully Paid	15625	11.99	518.9	RENT	34000	Source Verified	16100	55.1	37	3	15631.77
27	36 months	credit_card	Fully Paid	16000	8.19	502.79	RENT	52000	Not Verified	10723	45.6	18	1	17566.02
28	36 months	credit_card	Fully Paid	10000	11.53	329.91	MORTGAGE	60000	Source Verified	10085	77.6	11	4	10690.14
29	36 months	credit_card	Fully Paid	8000	13.33	270.83	MORTGAGE	54000	Verified	2379	15.7	17	7	8155.32
30	36 months	credit_card	Charged Off	6325	15.41	220.54	MORTGAGE	72500	Source Verified	3197	55.1	17	5	0
31	36 months	credit_card	Fully Paid	18000	11.99	597.78	MORTGAGE	80000	Not Verified	13747	61.4	44	7	18956.07
32	36 months	credit_card	Fully Paid	30000	6.99	926.18	MORTGAGE	180000	Source Verified	159886	43.8	40	5	32498.7
33	36 months	credit_card	Charged Off	3000	15.61	104.9	RENT	21000	Verified	5297	79.1	17	4	797.01
34	36 months	credit_card	Charged Off	15000	12.39	501.02	RENT	65000	Not Verified	12767	66.2	24	4	7954.57
35	36 months	credit_card	Fully Paid	6000	10.99	196.41	RENT	111500	Source Verified	33435	72.5	11	2	6533.910001
36	36 months	credit_card	Fully Paid	19200	5.32	578.21	MORTGAGE	135000	Not Verified	34725	25.2	33	6	20067.55

Column1	Audation & Profiling
Missing_Count	0
Total Rows	78658
Missing %	0%

Here use a simple function to handle if the data set contain blanks and the total % of this using and convert the root data in a functional table

Missing Count

- =COUNTBLANK(Table1)

Total Rows

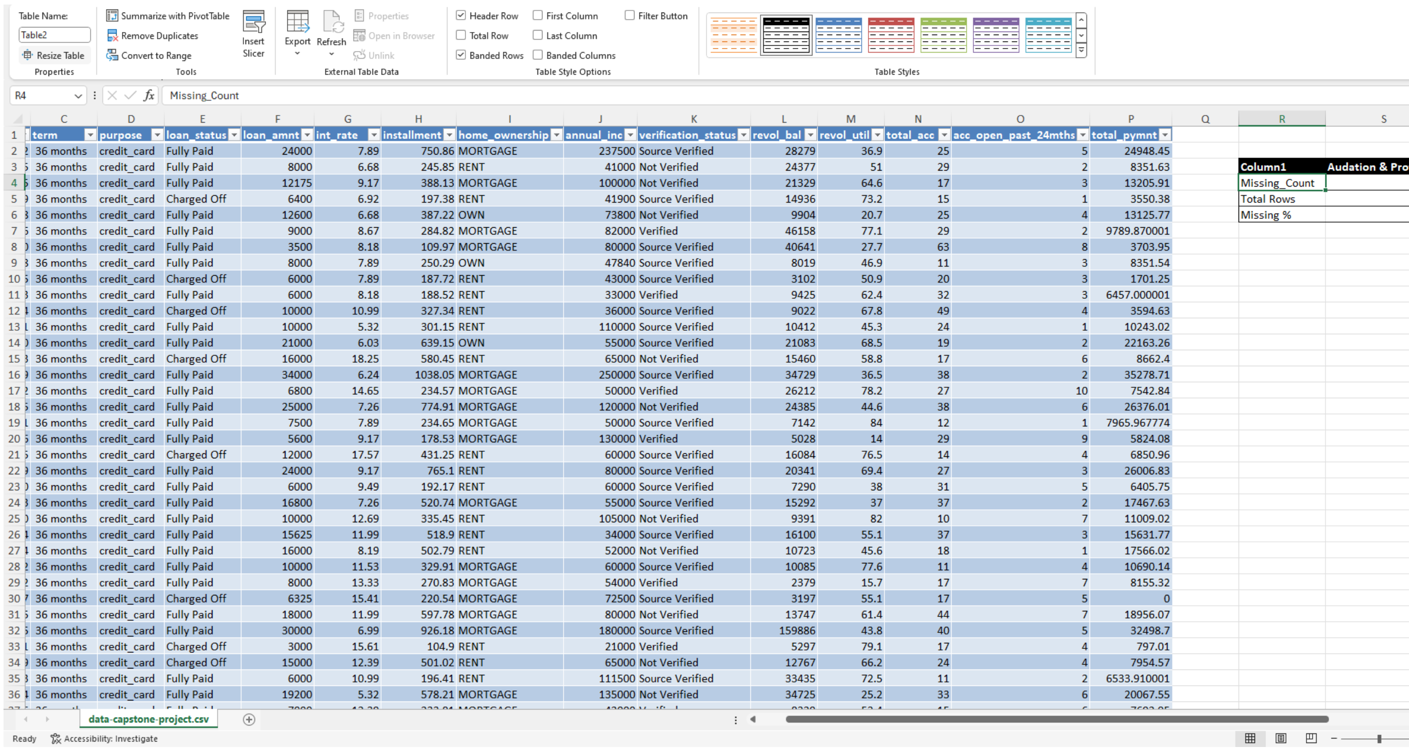
- =ROWS(Table1)

Filtering

The **cell total_pymnt** column contained invalid values represented as "-" or zero. Since a loan cannot be charged off without any payment history, these values were treated as missing data. Rows with missing total payment values were removed to maintain data consistency and analytical validity.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	member_id	term	purpose	loan_status	loan_amnt	int_ratef	installment	home_ownership	annual_inc	verification_status	revol_bal	revol_util	total_acc	acc_open_past_24mths	total_pymnt	
30	63558275	67912987	36 months	credit_card	Charged Off	$ 6,325	15.41%	$ 220.54	MORTGAGE	$ 72,500.00	Source Verified	$ 3,197	55.1%	17	5	$ -	
2794	41051750	43927579	36 months	credit_card	Charged Off	$ 12,000	7.89%	$ 375.43	RENT	$137,000.00	Source Verified	$ 92,698	65.0%	30	8	$ -	
4784	66955316	71756031	36 months	credit_card	Charged Off	$ 12,000	7.89%	$ 375.43	RENT	$250,000.00	Verified	$ 117,760	56.1%	26	5	$ -	
5639	68393736	73283477	36 months	credit_card	Charged Off	$ 12,000	8.49%	$ 378.76	RENT	$ 60,000.00	Not Verified	$ 15,548	81.8%	22	3	$ -	
6292	53654320	57185045	36 months	credit_card	Charged Off	$ 5,000	12.29%	$ 166.77	RENT	$ 23,753.40	Verified	$ 7,981	38.2%	42	18	$ -	
6753	51506406	54926122	36 months	credit_card	Charged Off	$ 8,000	13.99%	$ 273.39	RENT	$ 20,000.00	Verified	$ 1,389	6.1%	21	13	$ -	
7153																	
7154																	

Data types & Standardization



term	purpose	loan_status	loan_amnt	int_rate	installment	home_ownership	annual_inc	verification_status	revol_bal	revol_util	total_acc	acc_open_past_24mths	total_pymnt
36 months	credit_card	Fully Paid	24000	7.89	750.86	MORTGAGE	237500	Source Verified	28279	36.9	25	5	24948.45
36 months	credit_card	Fully Paid	8000	6.68	245.85	RENT	41000	Not Verified	24377	51	29	2	8351.63
36 months	credit_card	Fully Paid	12175	9.17	388.13	MORTGAGE	100000	Not Verified	21329	64.6	17	3	13205.91
36 months	credit_card	Charged Off	6400	6.92	197.38	RENT	41900	Source Verified	14936	73.2	15	1	3550.38
36 months	credit_card	Fully Paid	12600	6.68	387.22	OWN	73800	Not Verified	9904	20.7	25	4	13125.77
36 months	credit_card	Fully Paid	9000	8.67	284.82	MORTGAGE	82000	Verified	46158	77.1	29	2	9789.870001
36 months	credit_card	Fully Paid	3500	8.18	109.97	MORTGAGE	80000	Source Verified	40641	27.7	63	8	3703.95
36 months	credit_card	Fully Paid	8000	7.89	250.29	OWN	47840	Source Verified	8019	46.9	11	3	8351.54
36 months	credit_card	Charged Off	6000	7.89	187.72	RENT	43000	Source Verified	3102	50.9	20	3	1701.25
36 months	credit_card	Fully Paid	6000	8.18	188.52	RENT	33000	Verified	9425	62.4	32	3	6457.000001
36 months	credit_card	Charged Off	10000	10.99	327.34	RENT	36000	Source Verified	9022	67.8	49	4	3594.63
36 months	credit_card	Fully Paid	10000	5.32	301.15	RENT	110000	Source Verified	10412	45.3	24	1	10243.02
36 months	credit_card	Fully Paid	21000	6.03	639.15	OWN	55000	Source Verified	21083	68.5	19	2	22163.26
36 months	credit_card	Charged Off	16000	18.25	580.45	RENT	65000	Not Verified	15460	58.8	17	6	8662.4
36 months	credit_card	Fully Paid	34000	6.24	1038.05	MORTGAGE	250000	Source Verified	34729	36.5	38	2	35278.71
36 months	credit_card	Fully Paid	6800	14.65	234.57	MORTGAGE	50000	Verified	26212	78.2	27	10	7542.84
36 months	credit_card	Fully Paid	25000	7.26	774.91	MORTGAGE	120000	Not Verified	24385	44.6	38	6	26376.01
36 months	credit_card	Fully Paid	7500	7.89	234.65	MORTGAGE	50000	Source Verified	7142	84	12	1	7965.967774
36 months	credit_card	Fully Paid	5600	9.17	178.53	MORTGAGE	130000	Verified	5028	14	29	9	5824.08
36 months	credit_card	Charged Off	12000	17.57	431.25	RENT	60000	Source Verified	16084	76.5	14	4	6850.96
36 months	credit_card	Fully Paid	24000	9.17	765.1	RENT	80000	Source Verified	20341	69.4	27	3	26006.83
36 months	credit_card	Fully Paid	6000	9.49	192.17	RENT	60000	Source Verified	7290	38	31	5	6405.75
36 months	credit_card	Fully Paid	16800	7.26	520.74	MORTGAGE	55000	Source Verified	15292	37	37	2	17467.63
36 months	credit_card	Fully Paid	10000	12.69	335.45	RENT	105000	Not Verified	9391	82	10	7	11009.02
36 months	credit_card	Fully Paid	15625	11.99	518.9	RENT	34000	Source Verified	16100	55.1	37	3	15631.77
36 months	credit_card	Fully Paid	16000	8.19	502.79	RENT	52000	Not Verified	10723	45.6	18	1	17566.02
36 months	credit_card	Fully Paid	10000	11.53	329.91	MORTGAGE	60000	Source Verified	10085	77.6	11	4	10690.14
36 months	credit_card	Fully Paid	8000	13.33	270.83	MORTGAGE	54000	Verified	2379	15.7	17	7	8155.32
36 months	credit_card	Charged Off	6325	15.41	220.54	MORTGAGE	72500	Source Verified	3197	55.1	17	5	0
36 months	credit_card	Fully Paid	18000	11.99	597.78	MORTGAGE	80000	Not Verified	13747	61.4	44	7	18956.07
36 months	credit_card	Fully Paid	30000	6.99	926.18	MORTGAGE	180000	Source Verified	159886	43.8	40	5	32498.7
36 months	credit_card	Charged Off	3000	15.61	104.9	RENT	21000	Verified	5297	79.1	17	4	797.01
36 months	credit_card	Charged Off	15000	12.39	501.02	RENT	65000	Not Verified	12767	66.2	24	4	7954.57
36 months	credit_card	Fully Paid	6000	10.99	196.41	RENT	111500	Source Verified	33435	72.5	11	2	6533.910001
36 months	credit_card	Fully Paid	19200	5.32	578.21	MORTGAGE	135000	Not Verified	34725	25.2	33	6	20067.55

Column1	Audation & Pro
Missing_Count	
Total Rows	
Missing %	

Data type fixes (convert to true numeric)

Currency fields converted to numeric

- Columns: **loan_amnt, installment, annual_inc, revol_bal, total_pymnt**
- Before: risk of being stored as text due to $ and separators.
- After: true numeric values; currency format applied only for display.

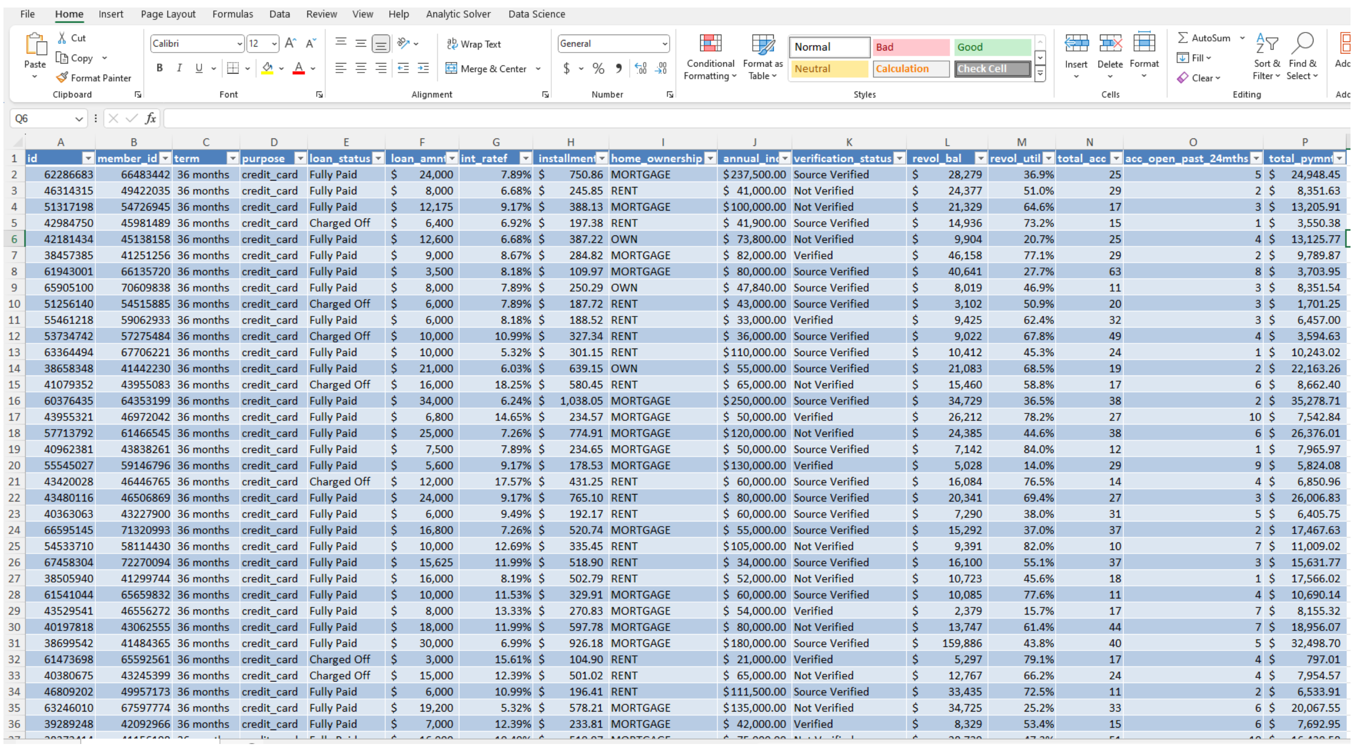
Percent fields corrected (scale + format)

- Columns: int_rate, revol_util
- Before: percent stored as text or scaled incorrectly (e.g., 789% instead of 7.89%).

- After: correct numeric percent values (e.g., 7.89%), not multiplied by 100.

IDs preserved

- Columns: id, member_id
- Before: risk of scientific notation or losing precision.
- After: protected formatting (text or full numeric without scientific notation).



id	member_id	term	purpose	loan_status	loan_amnt	int_ratef	installment	home_ownership	annual_inc	verification_status	revol_bal	revol_util	total_acc	acc_open_past_24mths	total_pymnt
62286683	66483442	36 months	credit_card	Fully Paid	$ 24,000	7.89%	$ 750.86	MORTGAGE	$237,500.00	Source Verified	$ 28,279	36.9%	25	5	$ 24,948.45
46314315	49422035	36 months	credit_card	Fully Paid	$ 8,000	6.68%	$ 245.85	RENT	$ 41,000.00	Not Verified	$ 24,377	51.0%	29	2	$ 8,351.63
51317198	54726945	36 months	credit_card	Fully Paid	$ 12,175	9.17%	$ 388.13	MORTGAGE	$100,000.00	Not Verified	$ 21,329	64.6%	17	3	$ 13,205.91
42984750	45981489	36 months	credit_card	Charged Off	$ 6,400	6.92%	$ 197.38	RENT	$ 41,900.00	Source Verified	$ 14,936	73.2%	15	1	$ 3,550.38
42181434	45138158	36 months	credit_card	Fully Paid	$ 12,600	6.68%	$ 387.22	OWN	$ 73,800.00	Not Verified	$ 9,904	20.7%	25	4	$ 13,125.77
38457385	41251256	36 months	credit_card	Fully Paid	$ 9,000	8.67%	$ 284.82	MORTGAGE	$ 82,000.00	Verified	$ 46,158	77.1%	29	2	$ 9,789.87
61943001	66135720	36 months	credit_card	Fully Paid	$ 3,500	8.18%	$ 109.97	MORTGAGE	$ 80,000.00	Source Verified	$ 40,641	27.7%	63	8	$ 3,703.95
65905100	70609838	36 months	credit_card	Fully Paid	$ 8,000	7.89%	$ 250.29	OWN	$ 47,840.00	Source Verified	$ 8,019	46.9%	11	3	$ 8,351.54
51256140	54515885	36 months	credit_card	Charged Off	$ 6,000	7.89%	$ 187.72	RENT	$ 43,000.00	Source Verified	$ 3,102	50.9%	20	3	$ 1,701.25
55461218	59062933	36 months	credit_card	Fully Paid	$ 6,000	8.18%	$ 188.52	RENT	$ 33,000.00	Verified	$ 9,425	62.4%	32	3	$ 6,457.00
53734742	57275484	36 months	credit_card	Charged Off	$ 10,000	10.99%	$ 327.34	RENT	$ 36,000.00	Source Verified	$ 9,022	67.8%	49	4	$ 3,594.63
63364494	67706221	36 months	credit_card	Fully Paid	$ 10,000	5.32%	$ 301.15	RENT	$110,000.00	Source Verified	$ 10,412	45.3%	24	1	$ 10,243.02
38658348	41442230	36 months	credit_card	Fully Paid	$ 21,000	6.03%	$ 639.15	OWN	$ 55,000.00	Source Verified	$ 21,083	68.5%	19	2	$ 22,163.26
41079352	43955083	36 months	credit_card	Charged Off	$ 16,000	18.25%	$ 580.45	RENT	$ 65,000.00	Not Verified	$ 15,460	58.8%	17	6	$ 8,662.40
60376435	64353199	36 months	credit_card	Fully Paid	$ 34,000	6.24%	$ 1,038.05	MORTGAGE	$250,000.00	Source Verified	$ 34,729	36.5%	38	2	$ 35,278.71
43955321	46972042	36 months	credit_card	Fully Paid	$ 6,800	14.65%	$ 234.57	MORTGAGE	$ 50,000.00	Verified	$ 26,212	78.2%	27	10	$ 7,542.84
57713792	61466545	36 months	credit_card	Fully Paid	$ 25,000	7.26%	$ 774.91	MORTGAGE	$120,000.00	Not Verified	$ 24,385	44.6%	38	6	$ 26,376.01
40962381	43838261	36 months	credit_card	Fully Paid	$ 7,500	7.89%	$ 234.65	MORTGAGE	$ 50,000.00	Source Verified	$ 7,142	84.0%	12	1	$ 7,965.97
55545027	59146796	36 months	credit_card	Fully Paid	$ 5,600	9.17%	$ 178.53	MORTGAGE	$130,000.00	Verified	$ 5,028	14.0%	29	9	$ 5,824.08
43420028	46446765	36 months	credit_card	Charged Off	$ 12,000	17.57%	$ 431.25	RENT	$ 60,000.00	Source Verified	$ 16,084	76.5%	14	4	$ 6,850.96
43480116	46506869	36 months	credit_card	Fully Paid	$ 24,000	9.17%	$ 765.10	RENT	$ 80,000.00	Source Verified	$ 20,341	69.4%	27	3	$ 26,006.83
40363063	43227900	36 months	credit_card	Fully Paid	$ 6,000	9.49%	$ 192.17	RENT	$ 60,000.00	Source Verified	$ 7,290	38.0%	31	5	$ 6,405.75
66595145	71320993	36 months	credit_card	Fully Paid	$ 16,800	7.26%	$ 520.74	MORTGAGE	$ 55,000.00	Source Verified	$ 15,292	37.0%	37	2	$ 17,467.63
54533710	58114430	36 months	credit_card	Fully Paid	$ 10,000	12.69%	$ 335.45	RENT	$105,000.00	Not Verified	$ 9,391	82.0%	10	7	$ 11,009.02
67458304	72270094	36 months	credit_card	Fully Paid	$ 15,625	11.99%	$ 518.90	RENT	$ 34,000.00	Source Verified	$ 16,100	55.1%	37	3	$ 15,631.77
38505940	41299744	36 months	credit_card	Fully Paid	$ 16,000	8.19%	$ 502.79	RENT	$ 52,000.00	Not Verified	$ 10,723	45.6%	18	1	$ 17,566.02
61541044	65659832	36 months	credit_card	Fully Paid	$ 10,000	11.53%	$ 329.91	MORTGAGE	$ 60,000.00	Source Verified	$ 10,085	77.6%	11	4	$ 10,690.14
43529541	46556272	36 months	credit_card	Fully Paid	$ 8,000	13.33%	$ 270.83	MORTGAGE	$ 54,000.00	Verified	$ 2,379	15.7%	17	7	$ 8,155.32
40197818	43062555	36 months	credit_card	Fully Paid	$ 18,000	11.99%	$ 597.78	MORTGAGE	$ 80,000.00	Not Verified	$ 13,747	61.4%	44	7	$ 18,956.07
38699542	41484365	36 months	credit_card	Fully Paid	$ 30,000	6.99%	$ 926.18	MORTGAGE	$180,000.00	Source Verified	$ 159,886	43.8%	40	5	$ 32,498.70
61473698	65592561	36 months	credit_card	Charged Off	$ 3,000	15.61%	$ 104.90	RENT	$ 21,000.00	Verified	$ 5,297	79.1%	17	4	$ 797.01
40380675	43245399	36 months	credit_card	Charged Off	$ 15,000	12.39%	$ 501.02	RENT	$ 65,000.00	Not Verified	$ 12,767	66.2%	24	4	$ 7,954.57
46809202	49957173	36 months	credit_card	Fully Paid	$ 6,000	10.99%	$ 196.41	RENT	$111,500.00	Source Verified	$ 33,435	72.5%	11	2	$ 6,533.91
63246010	67597774	36 months	credit_card	Fully Paid	$ 19,200	5.32%	$ 578.21	MORTGAGE	$135,000.00	Not Verified	$ 34,725	25.2%	33	6	$ 20,067.55
39289248	42092966	36 months	credit_card	Fully Paid	$ 7,000	12.39%	$ 233.81	MORTGAGE	$ 42,000.00	Verified	$ 8,329	53.4%	15	6	$ 7,692.95

Preparing for Analytical y EDA

Default label created

Using Logical function can transform the categorical values for the EDA (Data modeling) “=IF(E2="Fully Paid",0,1)”

- Before: loan_status only as text.
- After: default_flag:
  - Charged Off = 1
  - Fully Paid = 0

Also for home_ownership (=IFS(J2="MORTGAGE",1,J2="RENT",0,J2="OWN",2))

- MORTAGAGE = 1
- RENT = 0
- OWN = 2

K2 =IFS(J2="MORTGAGE",1,J2="RENT",0,J2="OWN",2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	member_id	term	purpose	loan_status	Label_Loan	loan_amnt	int_ratef	installment	home_ownership	ownership_label	annual_inc	verification_	revol_bal	revol_util	total_acc	acc_open_pa
2	62286683	66483442	36 months	credit_card	Fully Paid	0	$ 24,000	7.89%	$ 750.86	MORTGAGE	1	$ 237,500.00	Source Verified	$ 28,279	36.9%	25	5
3	46314315	49422035	36 months	credit_card	Fully Paid	0	$ 8,000	6.68%	$ 245.85	RENT	0	41,000.00	Not Verified	$ 24,377	51.0%	29	2
4	51317198	54726945	36 months	credit_card	Fully Paid	0	$ 12,175	9.17%	$ 388.13	MORTGAGE	1	$ 100,000.00	Not Verified	$ 21,329	64.6%	17	3
5	42984750	45981489	36 months	credit_card	Charged Off	1	$ 6,400	6.92%	$ 197.38	RENT	0	$ 41,900.00	Source Verified	$ 14,936	73.2%	15	1
6	42181434	45138158	36 months	credit_card	Fully Paid	0	$ 12,600	6.68%	$ 387.22	OWN	2	$ 73,800.00	Not Verified	$ 9,904	20.7%	25	4
7	38457385	41251256	36 months	credit_card	Fully Paid	0	$ 9,000	8.67%	$ 284.82	MORTGAGE	1	$ 82,000.00	Verified	$ 46,158	77.1%	29	2
8	61943001	66135720	36 months	credit_card	Fully Paid	0	$ 3,500	8.18%	$ 109.97	MORTGAGE	1	$ 80,000.00	Source Verified	$ 40,641	27.7%	63	8
9	65905100	70609838	36 months	credit_card	Fully Paid	0	$ 8,000	7.89%	$ 250.29	OWN	2	$ 47,840.00	Source Verified	$ 8,019	46.9%	11	3
10	51256140	54515885	36 months	credit_card	Charged Off	1	$ 6,000	7.89%	$ 187.72	RENT	0	$ 43,000.00	Source Verified	$ 3,102	50.9%	20	3
11	55461218	59062933	36 months	credit_card	Fully Paid	0	$ 6,000	8.18%	$ 188.52	RENT	0	$ 33,000.00	Verified	$ 9,425	62.4%	32	3
12	53734742	57275484	36 months	credit_card	Charged Off	1	$ 10,000	10.99%	$ 327.34	RENT	0	$ 36,000.00	Source Verified	$ 9,022	67.8%	49	4
13	63364494	67706221	36 months	credit_card	Fully Paid	0	$ 10,000	5.32%	$ 301.15	RENT	0	$ 110,000.00	Source Verified	$ 10,412	45.3%	24	1
14	38658348	41442230	36 months	credit_card	Fully Paid	0	$ 21,000	6.03%	$ 639.15	OWN	2	$ 55,000.00	Source Verified	$ 21,083	68.5%	19	2
15	41079352	43955083	36 months	credit_card	Charged Off	1	$ 16,000	18.25%	$ 580.45	RENT	0	$ 65,000.00	Not Verified	$ 15,460	58.8%	17	6
16	60376435	64353199	36 months	credit_card	Fully Paid	0	$ 34,000	6.24%	$ 1,038.05	MORTGAGE	1	$ 250,000.00	Source Verified	$ 34,729	36.5%	38	2
17	43955321	46972042	36 months	credit_card	Fully Paid	0	$ 6,800	14.65%	$ 234.57	MORTGAGE	1	$ 50,000.00	Verified	$ 26,212	78.2%	27	10
18	57713792	61466545	36 months	credit_card	Fully Paid	0	$ 25,000	7.26%	$ 774.91	MORTGAGE	1	$ 120,000.00	Not Verified	$ 24,385	44.6%	38	6
19	40962381	43838261	36 months	credit_card	Fully Paid	0	$ 7,500	7.89%	$ 234.65	MORTGAGE	1	$ 50,000.00	Source Verified	$ 7,142	84.0%	12	1
20	55545027	59146796	36 months	credit_card	Fully Paid	0	$ 5,600	9.17%	$ 178.53	MORTGAGE	1	$ 130,000.00	Verified	$ 5,028	14.0%	29	9
21	43420028	46446765	36 months	credit_card	Charged Off	1	$ 12,000	17.57%	$ 431.25	RENT	0	$ 60,000.00	Source Verified	$ 16,084	76.5%	14	4
22	43480116	46506869	36 months	credit_card	Fully Paid	0	$ 24,000	9.17%	$ 765.10	RENT	0	$ 80,000.00	Source Verified	$ 20,341	69.4%	27	3
23	40363063	43227900	36 months	credit_card	Fully Paid	0	$ 6,000	9.49%	$ 192.17	RENT	0	$ 60,000.00	Source Verified	$ 7,290	38.0%	31	5
24	66595145	71320993	36 months	credit_card	Fully Paid	0	$ 16,800	7.26%	$ 520.74	MORTGAGE	1	$ 55,000.00	Source Verified	$ 15,292	37.0%	37	2
25	54533710	58114430	36 months	credit_card	Fully Paid	0	$ 10,000	12.69%	$ 335.45	RENT	0	$ 105,000.00	Not Verified	$ 9,391	82.0%	10	7
26	67458304	72270094	36 months	credit_card	Fully Paid	0	$ 15,625	11.99%	$ 518.90	RENT	0	$ 34,000.00	Source Verified	$ 16,100	55.1%	37	3
27	38505940	41299744	36 months	credit_card	Fully Paid	0	$ 16,000	8.19%	$ 502.79	RENT	0	$ 52,000.00	Not Verified	$ 10,723	45.6%	18	1
28	61541044	65659832	36 months	credit_card	Fully Paid	0	$ 10,000	11.53%	$ 329.91	MORTGAGE	1	$ 60,000.00	Source Verified	$ 10,085	77.6%	11	4
29	43529541	46556272	36 months	credit_card	Fully Paid	0	$ 8,000	13.33%	$ 270.83	MORTGAGE	1	$ 54,000.00	Verified	$ 2,379	15.7%	17	7
30	40197818	43062555	36 months	credit_card	Fully Paid	0	$ 18,000	11.99%	$ 597.78	MORTGAGE	1	$ 80,000.00	Not Verified	$ 13,747	61.4%	44	7
31	38699542	41484365	36 months	credit_card	Fully Paid	0	$ 30,000	6.99%	$ 926.18	MORTGAGE	1	$ 180,000.00	Source Verified	$ 159,886	43.8%	40	5
32	61473698	65592561	36 months	credit_card	Charged Off	1	$ 3,000	15.61%	$ 104.90	RENT	0	$ 21,000.00	Verified	$ 5,297	79.1%	17	4
33	40380675	43245399	36 months	credit_card	Charged Off	1	$ 15,000	12.39%	$ 501.02	RENT	0	$ 65,000.00	Not Verified	$ 12,767	66.2%	24	4
34	46809202	49957173	36 months	credit_card	Fully Paid	0	$ 6,000	10.99%	$ 196.41	RENT	0	$ 111,500.00	Source Verified	$ 33,435	72.5%	11	2
35	63246010	67597774	36 months	credit_card	Fully Paid	0	$ 19,200	5.32%	$ 578.21	MORTGAGE	1	$ 135,000.00	Not Verified	$ 34,725	25.2%	33	6
36	39289248	42092966	36 months	credit_card	Fully Paid	0	$ 7,000	12.39%	$ 233.81	MORTGAGE	1	$ 42,000.00	Verified	$ 8,329	53.4%	15	6

Step 2 Exploratory analysis

Pivot table to prepare analysis Pre-Modeling

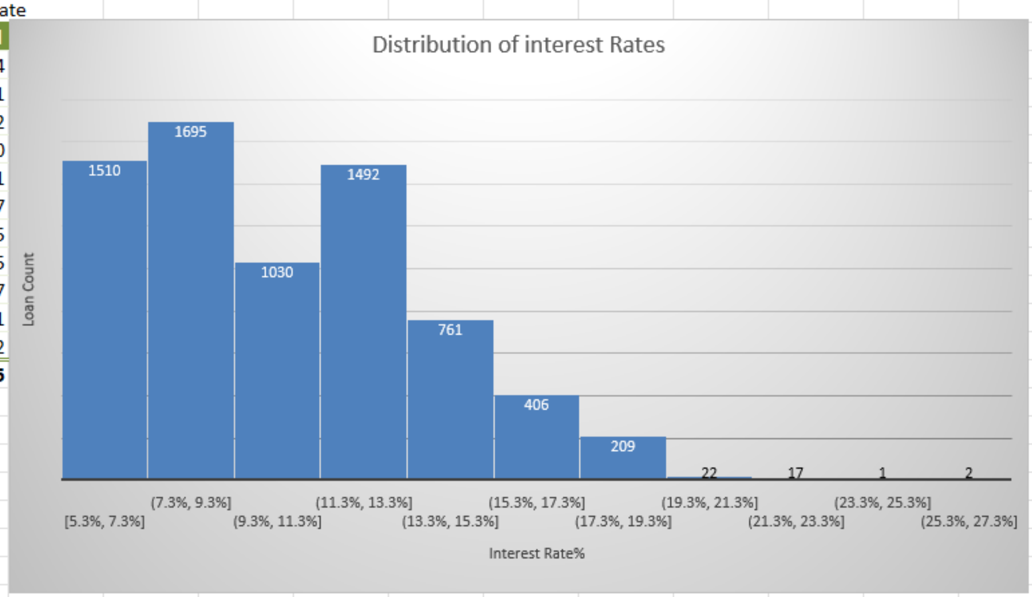
Objective

Use pivot tables to identify patterns related to loan default risk and validate that cleaned variables behave as expected before modeling. And based in the core business question **“Which loans should we select to maximize return while minimizing default risk?”**

First analyze the distribution of interest rate

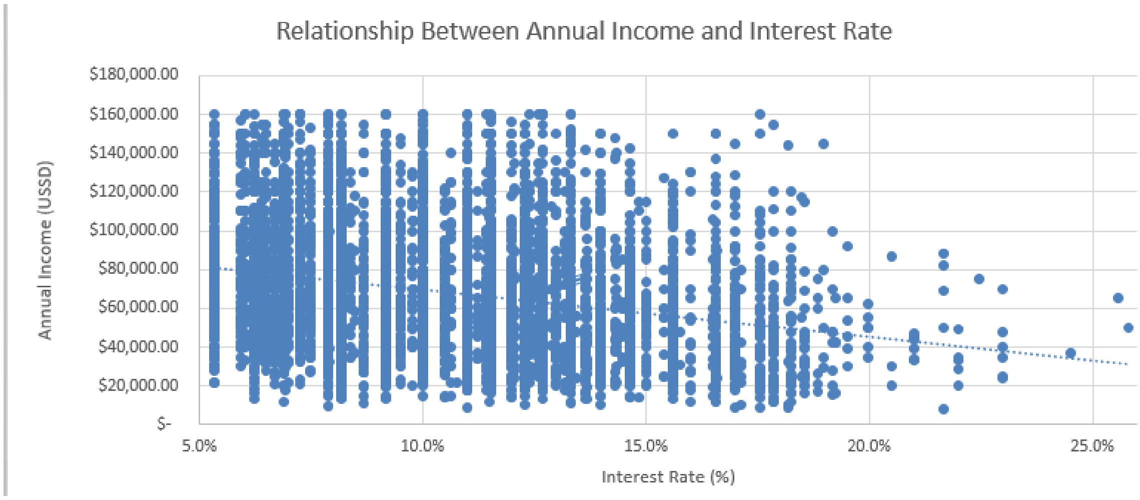
Distribution of interset rate

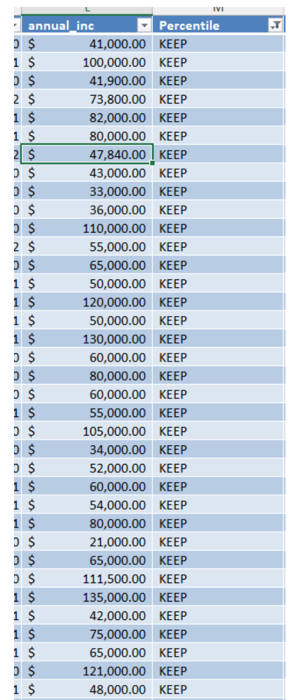
Int_Ratef	Count of id
0.05-0.07	1314
0.07-0.09	1441
0.09-0.11	1462
0.11-0.13	1480
0.13-0.15	791
0.15-0.17	397
0.17-0.19	205
0.19-0.21	35
0.21-0.23	17
0.23-0.25	1
0.25-0.27	2
Grand Total	**7145**



Interest rates were grouped into 2-percentage-point bins ranging from 5% to 27% using Excel Pivot Table grouping. Rates are stored as decimal values and displayed as percentages for interpretability. The distribution shows a strong concentration between 7% and 13%, indicating that most loans fall within moderate interest rate ranges.

Relationship between variables





annual_inc	Percentile
$ 41,000.00	KEEP
$ 100,000.00	KEEP
$ 41,900.00	KEEP
$ 73,800.00	KEEP
$ 82,000.00	KEEP
$ 80,000.00	KEEP
$ 47,840.00	KEEP
$ 43,000.00	KEEP
$ 33,000.00	KEEP
$ 36,000.00	KEEP
$ 110,000.00	KEEP
$ 55,000.00	KEEP
$ 65,000.00	KEEP
$ 50,000.00	KEEP
$ 120,000.00	KEEP
$ 50,000.00	KEEP
$ 130,000.00	KEEP
$ 60,000.00	KEEP
$ 80,000.00	KEEP
$ 60,000.00	KEEP
$ 55,000.00	KEEP
$ 105,000.00	KEEP
$ 34,000.00	KEEP
$ 52,000.00	KEEP
$ 60,000.00	KEEP
$ 54,000.00	KEEP
$ 80,000.00	KEEP
$ 21,000.00	KEEP
$ 65,000.00	KEEP
$ 111,500.00	KEEP
$ 135,000.00	KEEP
$ 42,000.00	KEEP
$ 75,000.00	KEEP
$ 65,000.00	KEEP
$ 121,000.00	KEEP
$ 48,000.00	KEEP
$ 65,000.00	KEEP

The variable (annual_inc) exhibited a right-skewed distribution with a small number of extreme high-income values (outliers)

Rather using an arbitrary cutoof, incomes values above the **95th percentile** (approximately at$160,000) **were removed only for visualization purposes**, preserving the natural income range for the majority of observations using logical functions such (=PERCENTIL.INC) and binning the interest rates.

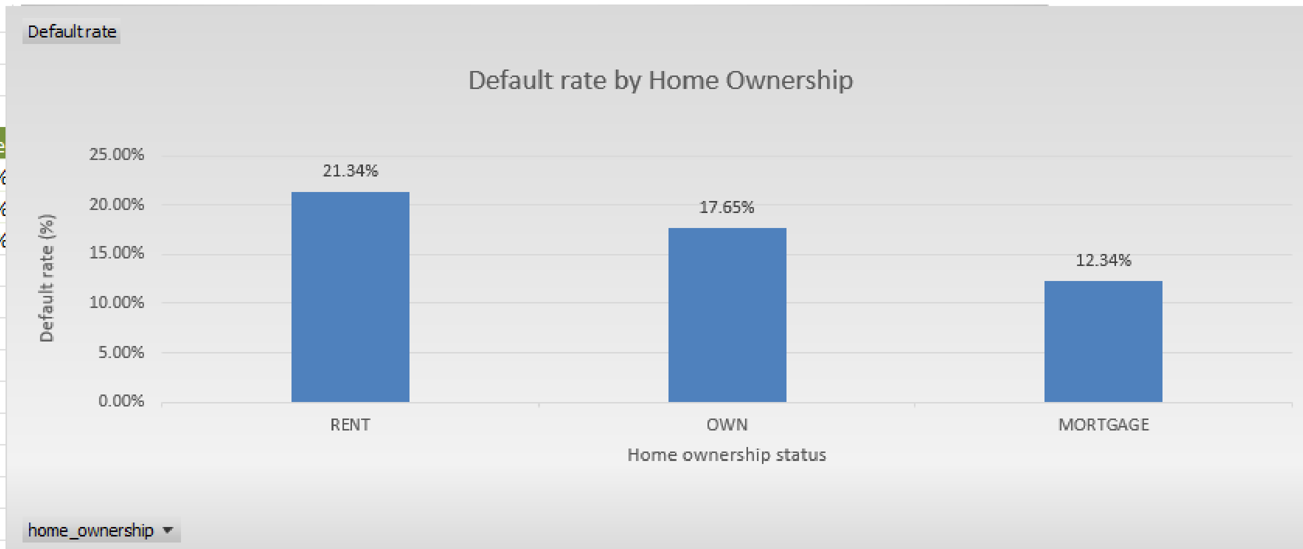
After that A **scatter plot** was used to visualize the relationship between **annual_inc** and **int_rate** after removing income values above the 95th percentile.

The visualization shows a **negative relationship**: borrowers with higher income generally receive lower interest rates, while lower-income borrowers are charged higher rates. Although the relationship is not perfectly linear, the trend aligns with expected credit-risk pricing behavior.

Default rate by Home Ownership

Compare **default behavior** across **home ownership** categories to assess borrower stability and investment risk

Row Labels	Default rate
RENT	21.34%
OWN	17.65%
MORTGAGE	12.34%



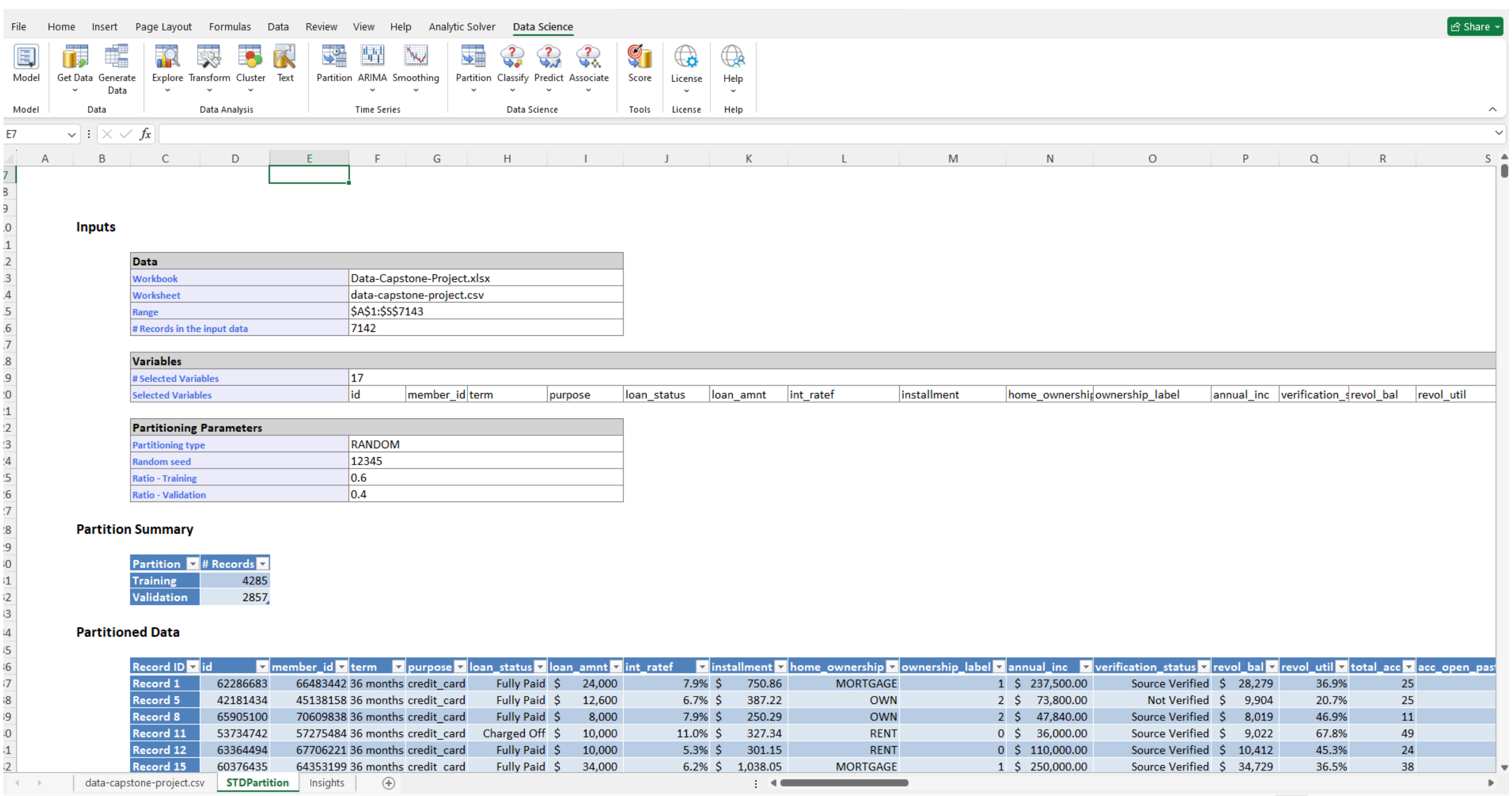
A PivotTable and bar chart were used to compare default rates across home ownership categories Renters exhibit higher default rates that borrowers with mortgages or owned homes, suggesting that home ownership is associated with lower credit risk and greater financial stability

Step 3 Performing predictive analytics

First using Analytic Solver (standart program to EDA)

Split dataset into Training model:Validation to support model

The dataset is split to **separate model learning from model evaluation**. This prevents information leakage and allows an objective assessment of how well the model generalizes to unseen data, which is critical in credit-risk classification (good vs. bad loans) also based in the Core of the business problem.



Inputs

Data	
Workbook	Data-Capstone-Project.xlsx
Worksheet	data-capstone-project.csv
Range	$A$1:$S$7143
# Records in the input data	7142

Variables														
# Selected Variables	17													
Selected Variables	id	member_id	term	purpose	loan_status	loan_amnt	int_ratef	installment	home_ownershi	ownership_label	annual_inc	verification_s	revol_bal	revol_util

Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.6
Ratio - Validation	0.4

Partition Summary

Partition	# Records
Training	4285
Validation	2857

Partitioned Data

Record ID	id	member_id	term	purpose	loan_status	loan_amnt	int_ratef	installment	home_ownership	ownership_label	annual_inc	verification_status	revol_bal	revol_util	total_acc	acc_open_pas
Record 1	62286683	66483442	36 months	credit_card	Fully Paid	$ 24,000	7.9%	$ 750.86	MORTGAGE	1	$ 237,500.00	Source Verified	$ 28,279	36.9%	25	
Record 5	42181434	45138158	36 months	credit_card	Fully Paid	$ 12,600	6.7%	$ 387.22	OWN	2	$ 73,800.00	Not Verified	$ 9,904	20.7%	25	
Record 8	65905100	70609838	36 months	credit_card	Fully Paid	$ 8,000	7.9%	$ 250.29	OWN	2	$ 47,840.00	Source Verified	$ 8,019	46.9%	11	
Record 11	53734742	57275484	36 months	credit_card	Charged Off	$ 10,000	11.0%	$ 327.34	RENT	0	$ 36,000.00	Source Verified	$ 9,022	67.8%	49	
Record 12	63364494	67706221	36 months	credit_card	Fully Paid	$ 10,000	5.3%	$ 301.15	RENT	0	$ 110,000.00	Source Verified	$ 10,412	45.3%	24	
Record 15	60376435	64353199	36 months	credit_card	Fully Paid	$ 34,000	6.2%	$ 1,038.05	MORTGAGE	1	$ 250,000.00	Source Verified	$ 34,729	36.5%	38	

Categorical variables

In this step was convert the categorical variables into categorical predictors creating dummys to able define a valid and unbiased set of predictors for the classifcations task into the loan_status

tal_pymnt	loan_status_Charged Off	loan_status_Fully Paid
24,948.45	0	1
13,125.77	0	1
8,351.54	0	1
3,594.63	1	0
10,243.02	0	1
35,278.71	0	1
7,542.84	0	1
7,965.97	0	1
6,850.96	1	0
26,006.83	0	1
6,405.75	0	1
17,467.63	0	1
11,009.02	0	1
15,631.77	0	1
17,566.02	0	1
8,155.32	0	1
18,956.07	0	1
7,954.57	1	0
6,533.91	0	1
7,692.95	0	1
21,508.24	0	1
27,561.73	0	1
28,052.89	0	1
6,315.16	0	1
20,752.22	0	1
6,129.10	0	1
21,032.83	0	1
23,881.36	0	1
9,479.09	0	1
6,276.65	0	1
2,600.42	0	1
1,051.86	1	0
1,643.84	1	0
4,016.34	1	0

Logistic Regression Model

This is with the objective to classify loans into good and bad loans using logistic regression model,

Predictions and probabilities were generated **within the same dataset** using the partition column.

True Positives (TP): Bad loans correctly predicted as bad.

True Negatives (TN): Good loans correctly predicted as good.

TP and TN were computed using validation rows only.

Model performance was assessed using the **ROC curve and AUC** reported by Analytical solver program where use the loan status created in the partition data (Model training) such dependent variable the **loan_status** (Core problem) with the explanatory variables excel total_pymnt

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	Charged Off	Fully Paid
Charged Off	24	437
Fully Paid	17	2379

Error Report			
Class	# Cases	# Errors	% Error
Charged Off	461	437	94.79393
Fully Paid	2396	17	0.709516
Overall	2857	454	15.89079

Metrics	
Metric	Value
Accuracy (#correct)	2403
Accuracy (%correct)	84.10920546
Specificity	0.052060738
Sensitivity (Recall)	0.992904841
Precision	0.844815341
F1 score	0.912893323
Success Class	Fully Paid
Success Probability	0.5

Here able to see the success probability in the portafolio for investment based in the good or bad loans where use

Now will be use the Model Metrics (Validation Perfomance)

To this convert the logistic regression outputs (probabilities) into measurable model quality on Validation data

With the Confusion Matrix Counts based in different cutoffs for the necessary validation

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	Charged Off	Fully Paid
Charged Off	55	406
Fully Paid	73	2323

Error Report			
Class	# Cases	# Errors	% Error
Charged Off	461	406	88.06941432
Fully Paid	2396	73	3.046744574
Overall	2857	479	16.76583829

Metrics	
Metric	Value
Accuracy (#correct)	2378
Accuracy (%correct)	83.23416171
Specificity	0.119305857
Sensitivity (Recall)	0.969532554
Precision	0.851227556
F1 score	0.906536585
Success Class	Fully Paid
Success Probability	0.6

Data Science: Logistic Regression - Prediction of Validation Data

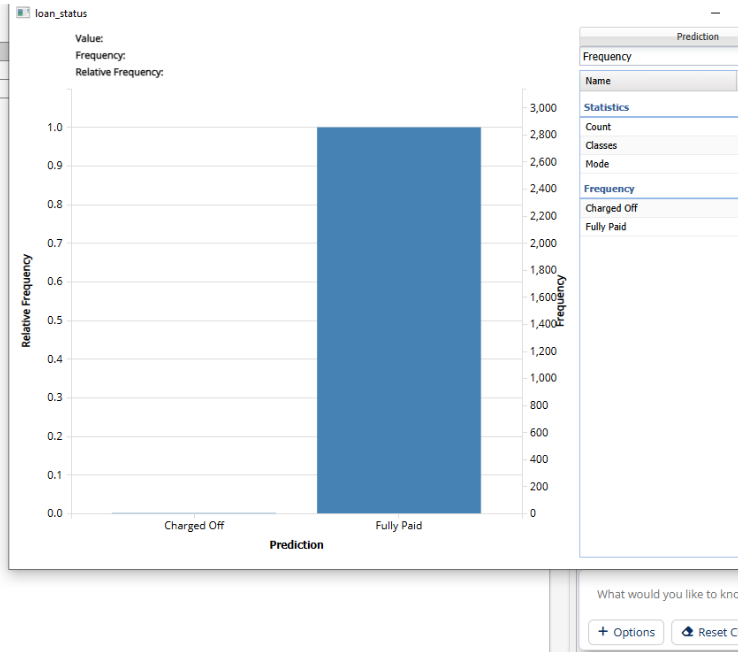
Output Navigator		
Inputs	Regression Summary	Predictor Screening
Training: Classification Summar	Training: Classification Details	Validation: Charts

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	Charged Off	Fully Paid
Charged Off	2	459
Fully Paid	1	2395

Error Report			
Class	# Cases	# Errors	% Error
Charged Off	461	459	99.56616052
Fully Paid	2396	1	0.041736227
Overall	2857	460	16.10080504

Metrics	
Metric	Value
Accuracy (#correct)	2397
Accuracy (%correct)	83.89919496
Specificity	0.004338395
Sensitivity (Recall)	0.999582638
Precision	0.83917309
F1 score	0.912380952
Success Class	Fully Paid
Success Probability	0.3



Validation: Classification Details

Data Science: Logistic Regression - Prediction of Validation Data

Output Navigator		
Inputs	Regression Summary	Predictor Screening
Training: Classification Summar	Validation: Classification Summary	Validation: Classification Details

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	Charged Off	Fully Paid
Charged Off	6	455
Fully Paid	6	2390

Error Report			
Class	# Cases	# Errors	% Error
Charged Off	461	455	98.69848156
Fully Paid	2396	6	0.250417362
Overall	2857	461	16.13580679

Metrics	
Metric	Value
Accuracy (#correct)	2396
Accuracy (%correct)	83.86419321
Specificity	0.013015184
Sensitivity (Recall)	0.997495826
Precision	0.840070299
F1 score	0.912039687
Success Class	Fully Paid
Success Probability	0.4

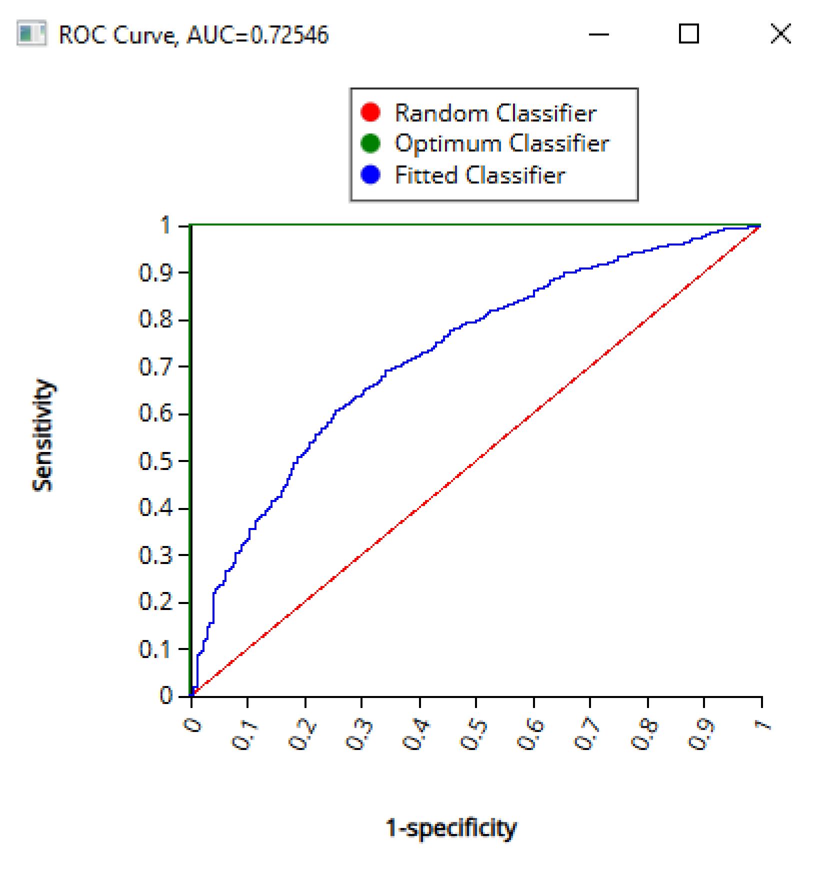
Here was be resumed the cutoffs for the validation and to asses the risk and meet more accuracy to evaluate the logistic regression model beyond a single arbitratry threshold, multiple probability cutoffs were tested using the validation dataset

Cutoff	**Default Detected (TP)**	**Defaults Missed (FN)**	**False Positives (FP)**	**Key Behavior**
0.3	Very low (2)	Extremely high	Minimal	Extremely permissive; almost all loans classified as Fully Paid
0.4	Very low (6)	Very high	Very low	Slight improvement, but defaults largely undetected
0.5	Moderate (24)	High	Low	Balanced baseline; still conservative toward Fully Paid
0.6	Higher (55)	Lower	Higher	Improved default detection with increased false positives

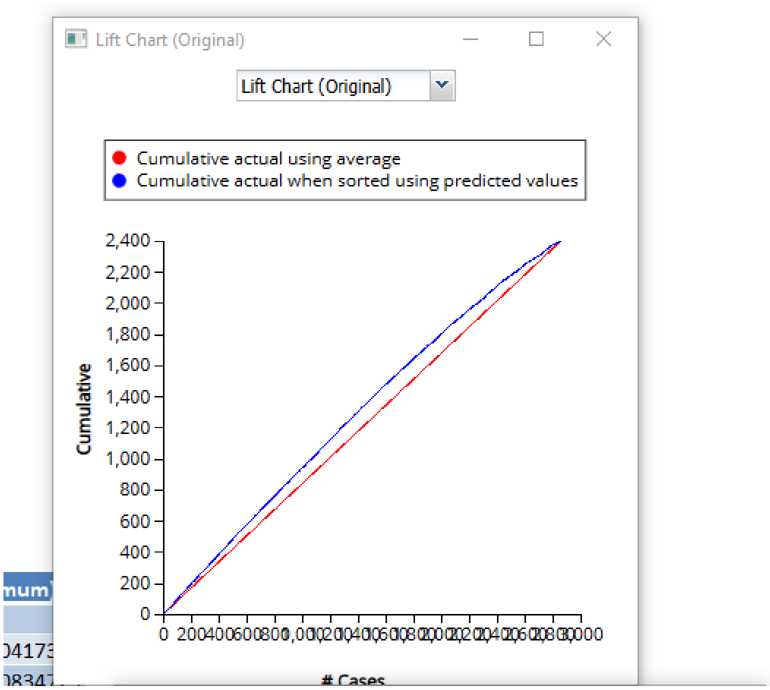
This model consistently identifying Fully Paid loans, as reflected by high recall for the classes

Step 4 Model Perfomance

The model was evaluated on unseen validation data using ROC analysis. The area AUC (Area Under the Curve) of 0.725 indicates good discriminatory power between defaulted and non-defaulted loans



Also the lift decile analysis show that loans ranked by predicted probability outperform random selection, especially in the top deciles, validating the use of the model for loan prioritization



And decision to take 0.6 such decision tool based in the probability cutoff were analyzed to understand classification trade-odds. However, cutoffs selection was driven by business risk tolerance rather than statistical optimization alone

cutoff 0.6

Validation: Classification Summary

Confusion Matrix		
Actual\	Chargec	Fully Pa
Charged C	55	406
Fully Paid	73	2323

cutoff 0.3

Validation: Classification Summary

Confusion Matrix		
Actual\	Chargec	Fully Pa
Charged C	2	459
Fully Paid	1	2395

cutoff 0.5

Validation: Classification Summary

Confusion Matrix		
Actual\	Chargec	Fully Pa
Charged C	24	437
Fully Paid	17	2379

Step 5 Decision Framework (Business Answer)

Turn this model into probabilities into a concrete investment decision in a portfolio where wich loan are selected and how much capital is allocated and what risk and return is expected where just now with the validation dataset is able to expected a conservative investor, the priority is to reduce defaults = 0.6 or if want maximize acceptance with controlled risk = 0.5.

In this case the investor choose the option 0.6 to control default risk Loans predicted probability of default below this threshold were considered eligible for investment computed as the product of probability of default and loan amount

Where need use the Expected Loss + PD * Loan_amnt

To this need create a unique value ID in both dataset (Dataset original (in this case was created a copy from the original for security asses) loan_amnt) and the (Validation 0.6)

In the original dataset with simple (=ROW) can add a Primary KEY ID

1	KEYID	id	member_i	term	purpose	loan_status	Label_Loan	loan_amnt	int_ratef	installment	home_ownership	ownership_label	annual_inc
3	1	46314315	49422035	36 months	credit_card	Fully Paid	0	$ 8,000	6.7%	$ 245.85	RENT	0	$ 4
4	2	51317198	54726945	36 months	credit_card	Fully Paid	0	$ 12,175	9.2%	$ 388.13	MORTGAGE	1	$ 10
5	3	42984750	45981489	36 months	credit_card	Charged Off	1	$ 6,400	6.9%	$ 197.38	RENT	0	$ 4
6	4	42181434	45138158	36 months	credit_card	Fully Paid	0	$ 12,600	6.7%	$ 387.22	OWN	2	$ 7
7	5	38457385	41251256	36 months	credit_card	Fully Paid	0	$ 9,000	8.7%	$ 284.82	MORTGAGE	1	$ 8
8	6	61943001	66135720	36 months	credit_card	Fully Paid	0	$ 3,500	8.2%	$ 109.97	MORTGAGE	1	$ 8
9	7	65905100	70609838	36 months	credit_card	Fully Paid	0	$ 8,000	7.9%	$ 250.29	OWN	2	$ 4
10	8	51256140	54515885	36 months	credit_card	Charged Off	1	$ 6,000	7.9%	$ 187.72	RENT	0	$ 4
11	9	55461218	59062933	36 months	credit_card	Fully Paid	0	$ 6,000	8.2%	$ 188.52	RENT	0	$ 3
12	10	53734742	57275484	36 months	credit_card	Charged Off	1	$ 10,000	11.0%	$ 327.34	RENT	0	$ 3
13	11	63364494	67706221	36 months	credit_card	Fully Paid	0	$ 10,000	5.3%	$ 301.15	RENT	0	$ 11
14	12	38658348	41442230	36 months	credit_card	Fully Paid	0	$ 21,000	6.0%	$ 639.15	OWN	2	$ 5
15	13	41079352	43955083	36 months	credit_card	Charged Off	1	$ 16,000	18.3%	$ 580.45	RENT	0	$ 6
17	15	43955321	46972042	36 months	credit_card	Fully Paid	0	$ 6,800	14.7%	$ 234.57	MORTGAGE	1	$ 5
18	16	57713792	61466545	36 months	credit_card	Fully Paid	0	$ 25,000	7.3%	$ 774.91	MORTGAGE	1	$ 12
19	17	40962381	43838261	36 months	credit_card	Fully Paid	0	$ 7,500	7.9%	$ 234.65	MORTGAGE	1	$ 5
20	18	55545027	59146796	36 months	credit_card	Fully Paid	0	$ 5,600	9.2%	$ 178.53	MORTGAGE	1	$ 13
21	19	43420028	46446765	36 months	credit_card	Charged Off	1	$ 12,000	17.6%	$ 431.25	RENT	0	$ 6
22	20	43480116	46506869	36 months	credit_card	Fully Paid	0	$ 24,000	9.2%	$ 765.10	RENT	0	$ 8
23	21	40363063	43227900	36 months	credit_card	Fully Paid	0	$ 6,000	9.5%	$ 192.17	RENT	0	$ 6
24	22	66595145	71320993	36 months	credit_card	Fully Paid	0	$ 16,800	7.3%	$ 520.74	MORTGAGE	1	$ 5
25	23	54533710	58114430	36 months	credit_card	Fully Paid	0	$ 10,000	12.7%	$ 335.45	RENT	0	$ 10
26	24	67458304	72270094	36 months	credit_card	Fully Paid	0	$ 15,625	12.0%	$ 518.90	RENT	0	$ 3
27	25	38505940	41299744	36 months	credit_card	Fully Paid	0	$ 16,000	8.2%	$ 502.79	RENT	0	$ 5
28	26	61541044	65659832	36 months	credit_card	Fully Paid	0	$ 10,000	11.5%	$ 329.91	MORTGAGE	1	$ 6
29	27	43529541	46556272	36 months	credit_card	Fully Paid	0	$ 8,000	13.3%	$ 270.83	MORTGAGE	1	$ 5
30	28	40197818	43062555	36 months	credit_card	Fully Paid	0	$ 18,000	12.0%	$ 597.78	MORTGAGE	1	$ 8
32	30	61473698	65592561	36 months	credit_card	Charged Off	1	$ 3,000	15.6%	$ 104.90	RENT	0	$ 2
33	31	40380675	43245399	36 months	credit_card	Charged Off	1	$ 15,000	12.4%	$ 501.02	RENT	0	$ 6
34	32	46809202	49957173	36 months	credit_card	Fully Paid	0	$ 6,000	11.0%	$ 196.41	RENT	0	$ 11
35	33	63246010	67597774	36 months	credit_card	Fully Paid	0	$ 19,200	5.3%	$ 578.21	MORTGAGE	1	$ 13
36	34	39289248	42092966	36 months	credit_card	Fully Paid	0	$ 7,000	12.4%	$ 233.81	MORTGAGE	1	$ 4
37	35	38372414	41156198	36 months	credit_card	Fully Paid	0	$ 16,000	10.5%	$ 519.97	MORTGAGE	1	$ 7
38	36	68434571	73324336	36 months	credit_card	Fully Paid	0	$ 20,000	12.0%	$ 664.20	MORTGAGE	1	$ 6
40	38	58110712	61914445	36 months	credit_card	Fully Paid	0	$ 25,000	11.5%	$ 824.76	RENT	0	$ 12
41	39	56632962	60314707	36 months	credit_card	Fully Paid	0	$ 10,000	13.3%	$ 338.54	MORTGAGE	1	$ 4

data-loan | data-capstone-project.csv | Validation0.6 | Reports | Matrix | STDPartition | Validation ...

On the validation set use the formula (=RIGHT) to arrange the RowKey where each number is in the original dataset

Validation: Classification Details

KEYID	Record ID	loan_status	Prediction: loan_sta	PostProb: Fully Paid	PostPro
2689	Record 2689	Fully Paid	Fully Paid	0.870499015	0.129501
5353	Record 5353	Fully Paid	Fully Paid	0.921826852	0.078173
1637	Record 1637	Fully Paid	Fully Paid	0.961300438	0.0387
3555	Record 3555	Fully Paid	Fully Paid	0.780806853	0.219193
5151	Record 5151	Charged Off	Fully Paid	0.653485363	0.346515
4246	Record 4246	Fully Paid	Fully Paid	0.863877669	0.136122
5467	Record 5467	Charged Off	Charged Off	0.557777969	0.442222
6372	Record 6372	Charged Off	Fully Paid	0.733939059	0.266061
6980	Record 6980	Fully Paid	Fully Paid	0.848458508	0.151541
3816	Record 3816	Fully Paid	Fully Paid	0.951646477	0.048354
6929	Record 6929	Charged Off	Fully Paid	0.921627799	0.078372
3403	Record 3403	Fully Paid	Fully Paid	0.906068458	0.093932
3326	Record 3326	Fully Paid	Fully Paid	0.888188166	0.111812
2578	Record 2578	Fully Paid	Fully Paid	0.901627846	0.098372
1823	Record 1823	Fully Paid	Fully Paid	0.924439816	0.07556
5806	Record 5806	Fully Paid	Fully Paid	0.92754895	0.072451
2937	Record 2937	Charged Off	Fully Paid	0.613615351	0.386385
7087	Record 7087	Fully Paid	Fully Paid	0.852899869	0.1471
4389	Record 4389	Fully Paid	Fully Paid	0.641825381	0.358175
2798	Record 2798	Fully Paid	Fully Paid	0.988659594	0.01134
6197	Record 6197	Fully Paid	Fully Paid	0.697131746	0.302868
694	Record 694	Fully Paid	Fully Paid	0.691775948	0.308224
6146	Record 6146	Charged Off	Fully Paid	0.828811638	0.171188
2793	Record 2793	Fully Paid	Fully Paid	0.650581601	0.349418
4861	Record 4861	Fully Paid	Fully Paid	0.882669786	0.11733
1789	Record 1789	Charged Off	Fully Paid	0.839489671	0.16051
6625	Record 6625	Charged Off	Charged Off	0.485081958	0.514918
6538	Record 6538	Fully Paid	Fully Paid	0.682105987	0.317894
2584	Record 2584	Fully Paid	Fully Paid	0.783038732	0.216961
2474	Record 2474	Charged Off	Charged Off	0.546165363	0.453835
4642	Record 4642	Fully Paid	Fully Paid	0.868481207	0.131519

Now use JOIN with the **VLOOKUP or XLOOKUP** to arrange the specific data to the dataset to run the model in this case I used **VLOOKUP**

=VLOOKUP(A2,Validation__Classification_Details_14[#All],6,FALSE)

U102 =VLOOKUP(A102,Validation__Classification_Details_14[#All],6,FALSE)

	bel	annual_inc	Percentile	verificatio	revol_bal	revol_util	total_acc	acc_open_past	total_pymnt	PD-Default
102	2	$ 84,000.00	REMOVE	Source Verifi	$ 44,411	57.7%	27	4	$ 7,416.05	0.092863807
103	2	$ 65,000.00	REMOVE	Verified	$ 54,589	89.5%	16	1	$ 11,524.25	0.557226875
104	1	$ 40,000.00	REMOVE	Not Verified	$ 7,257	36.3%	21	6	$ 10,230.34	0.087386855
112	1	$ 100,000.00	REMOVE	Source Verifi	$ 46,925	81.5%	21	5	$ 11,058.23	0.205212471
117	0	$ 67,000.00	REMOVE	Verified	$ 8,569	19.6%	45	18	$ 6,768.33	0.321611392
122	1	$ 85,000.00	REMOVE	Source Verifi	$ 8,722	46.4%	24	3	$ 16,339.69	0.085436712
125	2	$ 47,000.00	REMOVE	Verified	$ 10,681	55.1%	29	11	$ 9,942.83	0.273601922
127	0	$ 45,000.00	REMOVE	Not Verified	$ 24,538	47.4%	17	3	$ 7,247.59	0.1064399
128	1	$ 30,000.00	REMOVE	Source Verifi	$ 9,953	38.7%	39	7	$ 11,314.45	0.182722523
132	0	$ 60,000.00	REMOVE	Source Verifi	$ 5,627	25.8%	56	4	$ 8,039.73	0.143779015
136	2	$ 95,000.00	REMOVE	Source Verifi	$ 39,420	63.5%	43	2	$ 7,802.82	0.095019587
138	1	$ 53,000.00	REMOVE	Not Verified	$ 17,707	28.2%	37	12	$ 14,027.66	0.123602948
139	1	$ 77,000.00	REMOVE	Source Verifi	$ 30,869	71.1%	30	4	$ 15,934.28	0.086659582
140	1	$ 95,000.00	REMOVE	Not Verified	$ 4,638	53.3%	38	5	$ 10,510.09	0.076282982
141	1	$ 129,000.00	REMOVE	Verified	$ 73,793	35.0%	67	8	$ 28,872.20	0.044842465
145	1	$ 90,000.00	REMOVE	Source Verifi	$ 23,268	71.6%	22	3	$ 35,037.49	0.163096976
148	0	$ 30,000.00	REMOVE	Not Verified	$ 12,112	44.0%	11	6	$ 10,504.87	0.283976986
150	1	$ 84,000.00	REMOVE	Source Verifi	$ 15,236	39.8%	43	10	$ 15,622.74	0.065915349
152	1	$ 50,000.00	REMOVE	Source Verifi	$ 8,458	34.1%	22	7	$ 15,661.94	0.088752508
153	1	$ 169,900.00	REMOVE	Not Verified	$ 11,989	54.2%	57	9	$ 13,635.76	0.04319477
156	2	$ 47,000.00	REMOVE	Not Verified	$ 19,273	47.8%	32	3	$ 12,615.36	0.103612278
161	2	$ 110,500.00	REMOVE	Source Verifi	$ 26,535	32.4%	38	11	$ 20,510.37	0.086446188
163	2	$ 42,804.00	REMOVE	Source Verifi	$ 16,330	39.2%	30	9	$ 7,288.52	0.197904077
166	2	$ 22,000.00	REMOVE	Verified	$ 18,015	79.0%	10	1	$ 11,255.18	0.482632198
171	1	$ 140,000.00	REMOVE	Source Verifi	$ 35,967	61.7%	38	5	$ 30,139.59	0.053154758
173	0	$ 28,000.00	REMOVE	Not Verified	$ 10,254	39.6%	14	5	$ 8,003.56	0.217589104
174	2	$ 82,000.00	REMOVE	Not Verified	$ 16,820	19.4%	30	10	$ 20,474.16	0.135855067
176	1	$ 120,000.00	REMOVE	Source Verifi	$ 34,876	48.4%	39	10	$ 26,804.89	0.073102527
177	2	$ 95,000.00	REMOVE	Source Verifi	$ 11,302	38.8%	27	3	$ 22,281.69	0.063433968
178	1	$ 85,000.00	REMOVE	Not Verified	$ 6,102	55.0%	21	5	$ 6,641.92	0.118408794
179	0	$ 40,000.00	REMOVE	Source Verifi	$ 14,557	54.5%	36	2	$ 15,012.76	0.157328487
181	1	$ 65,000.00	REMOVE	Source Verifi	$ 15,724	80.6%	23	7	$ 9,499.65	0.145881855
184	2	$ 43,000.00	REMOVE	Not Verified	$ 4,095	59.3%	18	1	$ 11,081.40	0.20820942
192	1	$ 40,000.00	REMOVE	Source Verifi	$ 5,122	41.0%	37	9	$ 7,784.86	0.158892838
195	0	$ 69,000.00	REMOVE	Source Verifi	$ 9,414	42.6%	20	3	$ 11,274.56	0.09913113
198	0	$ 47,200.00	REMOVE	Not Verified	$ 4,350	48.4%	23	6	$ 3,288.40	0.139380505
199	1	$ 126,000.00	REMOVE	Source Verifi	$ 57,311	61.7%	24	3	$ 21,150.62	0.051757546

data-capstone-project.csv | Reports | Matrix | data-loan | Validation.06 | STDPartition | Validation0.3 ...

Ready 2818 of 7142 records found Accessibility: Investigate

Next step is find the better portafolio for Business answer

With the next formula

Expected Loss = PD * Loan_amnt

acc_open_past	total_pymnt	PD-Default	Expected_Loss	Elegible
3	$ 6,933.65	0.079627279	$ 477.76	Yes
4	$ 11,593.05	0.156883919	$ 1,600.22	Yes
7	$ 22,398.21	0.130148901	$ 2,733.13	Yes
9	$ 8,866.60	0.058522067	$ 453.55	Yes
2	$ 1,669.03	0.138626131	$ 207.94	Yes
12	$ 2,187.25	0.181273879	$ 362.55	Yes
7	$ 2,211.77	0.12156219	$ 267.44	Yes
1	$ 1,073.43	0.481166732	$ 481.17	Yes
4	$ 6,166.16	0.050057071	$ 300.34	Yes
7	$ 8,242.11	0.055269768	$ 382.74	Yes
7	$ 3,145.20	0.136533989	$ 382.30	Yes
9	$ 2,108.49	0.055666371	$ 111.33	Yes
9	$ 511.43	0.376936018	$ 565.40	Yes
5	$ 1,861.80	0.026951711	$ 48.51	Yes
6	$ 1,069.51	0.147747161	$ 443.24	Yes
3	$ 37,608.02	0.044396378	$ 1,553.87	Yes
7	$ 2,058.50	0.266405531	$ 479.53	Yes
5	$ 2,238.09	0.179712762	$ 345.95	Yes
8	$ 1,522.37	0.085942606	$ 128.91	Yes
11	$ 1,060.33	0.19855358	$ 198.55	Yes
3	$ 1,008.56	0.089511186	$ 89.51	Yes
7	$ 404.78	0.023876869	$ 47.75	Yes
6	$ 595.24	0.070000537	$ 210.00	Yes
5	$ 5,782.20	0.094159993	$ 470.80	Yes
2	$ 8,927.61	0.224867847	$ 1,911.38	Yes
3	$ 10,415.74	0.163074391	$ 1,598.13	Yes
1	$ 3,717.79	0.088849791	$ 310.97	Yes
2	$ 5,507.21	0.130483733	$ 652.42	Yes
7	$ 4,748.16	0.181896649	$ 818.53	Yes
5	$ 4,120.57	0.274748558	$ 1,098.99	Yes
3	$ 4,056.51	0.063836657	$ 255.35	Yes
7	$ 5,222.25	0.125117488	$ 625.59	Yes
0	$ 5,496.73	0.534897109	$ 2,674.49	Yes
5	$ 7,433.12	0.100617966	$ 654.02	Yes
4	$ 962.60	0.122324954	$ 244.65	Yes
9	$ 3,018.39	0.031307514	$ 93.92	Yes
8	$ 5,016.56	0.16747764	$ 837.39	Yes

ta-loan | Validation.06 | STDPartition | Validation0.3 | ...

Additionally add filter to see only the Elegible Loans for the investor

Step 6 Consolidation Portfolio for Investor

Now need build the portafolio for the investor at only elegible loans

Using Sum and logical functions

1	verificatio	revol_bal	revol_util	total_acc	acc_open_past	total_pymnt	PD-Default	Expected_Loss	Elegible	Cum_Investment	Selected
4	Source Verifi	$ 10,908	74.2%	19	6	$ 15,357.69	0.079627279	$ 1,194.41	Yes	$ 15,000	1
5	Not Verified	$ 9,272	74.2%	27	4	$ 10,172.41	0.156883919	$ 1,490.40	Yes	$ 24,500	1
7	Not Verified	$ 10,561	76.5%	7	0	$ 9,707.42	0.130148901	$ 1,145.31	Yes	$ 45,900	1
8	Verified	$ 9,213	8.4%	31	11	$ 2,215.66	0.058522067	$ 117.04	Yes	$ 47,900	1
.0	Not Verified	$ 22,193	83.4%	10	4	$ 9,346.68	0.138626131	$ 1,247.64	Yes	$ 64,900	1
.1	Not Verified	$ 10,032	46.7%	69	7	$ 13,230.59	0.181273879	$ 2,175.29	Yes	$ 76,900	1
.4	Source Verifi	$ 91,165	66.6%	34	4	$ 1,902.73	0.12156219	$ 203.62	Yes	$ 98,575	1
.5	Source Verifi	$ 7,108	64.6%	7	3	$ 5,512.52	0.481166732	$ 2,405.83	Yes	$ 103,575	1
.8	Not Verified	$ 889	2.8%	28	1	$ 2,150.34	0.050057071	$ 100.11	Yes	$ 146,375	1
20	Verified	$ 3,749	50.0%	12	3	$ 1,008.56	0.055269768	$ 55.27	Yes	$ 154,875	1
28	Not Verified	$ 14,804	68.2%	23	3	$ 2,022.32	0.136533989	$ 273.07	Yes	$ 257,300	1
31	Source Verifi	$ 6,674	42.8%	27	5	$ 5,471.19	0.055666371	$ 278.33	Yes	$ 288,300	1
32	Not Verified	$ 13,726	48.8%	26	7	$ 2,058.50	0.376936018	$ 678.48	Yes	$ 290,100	1
35	Verified	$ 1,213	10.5%	45	11	$ 1,060.33	0.026951711	$ 26.95	Yes	$ 312,100	1
37	Verified	$ 17,797	73.8%	10	3	$ 4,461.43	0.147747161	$ 590.99	Yes	$ 323,100	1
39	Not Verified	$ 6,091	88.3%	38	7	$ 2,327.39	0.044396378	$ 93.23	Yes	$ 345,200	1
41	Verified	$ 7,547	64.5%	24	12	$ 3,296.73	0.266405531	$ 799.22	Yes	$ 373,200	1
44	Verified	$ 5,748	58.1%	17	12	$ 2,187.25	0.179712762	$ 359.43	Yes	$ 404,800	1
45	Source Verifi	$ 7,693	90.5%	18	4	$ 9,409.34	0.085942606	$ 730.51	Yes	$ 413,300	1
47	Source Verifi	$ 3,895	34.2%	19	5	$ 1,861.80	0.19855358	$ 357.40	Yes	$ 434,100	1
52	Source Verifi	$ 5,285	43.7%	14	3	$ 965.80	0.089511186	$ 322.24	Yes	$ 494,700	1
53	Verified	$ 7,593	17.5%	62	13	$ 6,439.11	0.023876869	$ 143.26	Yes	$ 500,700	1
59	Verified	$ 5,028	14.0%	29	9	$ 5,824.08	0.070000537	$ 392.00	Yes	$ 531,325	1
54	Source Verifi	$ 4,477	32.4%	9	2	$ 1,669.03	0.094159993	$ 141.24	Yes	$ 598,700	1
55	Verified	$ 10,681	55.1%	29	11	$ 9,942.83	0.224867847	$ 2,023.81	Yes	$ 607,700	1
58	Source Verifi	$ 19,528	32.9%	30	13	$ 10,662.40	0.163074391	$ 1,630.74	Yes	$ 627,700	1
70	Source Verifi	$ 5,786	24.3%	21	7	$ 5,222.25	0.088849791	$ 444.25	Yes	$ 667,700	1
73	Source Verifi	$ 80,028	77.6%	34	9	$ 342.38	0.130483733	$ 521.93	Yes	$ 680,900	1
75	Not Verified	$ 36,761	67.5%	36	3	$ 1,214.62	0.181896649	$ 545.69	Yes	$ 695,900	1
76	Not Verified	$ 6,910	50.4%	26	9	$ 511.43	0.274748558	$ 412.12	Yes	$ 697,400	1
30	Not Verified	$ 1,274	7.9%	11	5	$ 3,236.31	0.063836657	$ 201.09	Yes	$ 718,875	1
36	Not Verified	$ 6,819	17.0%	47	7	$ 2,226.10	0.125117488	$ 275.26	Yes	$ 783,075	1
37	Not Verified	$ 7,914	60.0%	19	4	$ 1,105.12	0.534897109	$ 1,069.79	Yes	$ 785,075	1
38	Not Verified	$ 4,737	48.3%	13	3	$ 2,898.01	0.100617966	$ 261.61	Yes	$ 787,675	1
91	Not Verified	$ 3,830	7.5%	28	6	$ 2,556.32	0.122324954	$ 305.81	Yes	$ 819,175	1
94	Not Verified	$ 6,384	7.2%	21	5	$ 6,333.73	0.031307514	$ 187.85	Yes	$ 858,800	1
95	Source Verifi	$ 38,941	58.2%	27	9	$ 1,582.73	0.16747764	$ 251.22	Yes	$ 860,300	1

Once elegible loans were identified and ranked by expected loss, a cumulative investment approach was applied to construct final portfolio under a fixed capital constraint for each elegible loan where 1 is loans included in portfolio and 0 excluded

Where this model is based default risk estimates and expected monetary loss per loan and fixed investment budget of 10M the resulting is the risk elegibility rule

Where have this Portfolio performance Summary

Due to discrete loan sizes, the final portfolio slightly exceeds the target capital constraint, can fixed with a 2 filters but for this capstone the investor take the portfolio

By combining predictive modeling, probability cutoffs, and capital constraints, the analysis moves from model performance to a concrete investment recommendation. The resulting portfolio demonstrates how data-driven decision-making can be applied to lending scenarios to balance risk and return in monetary terms.

FROM RISK TO RETURN

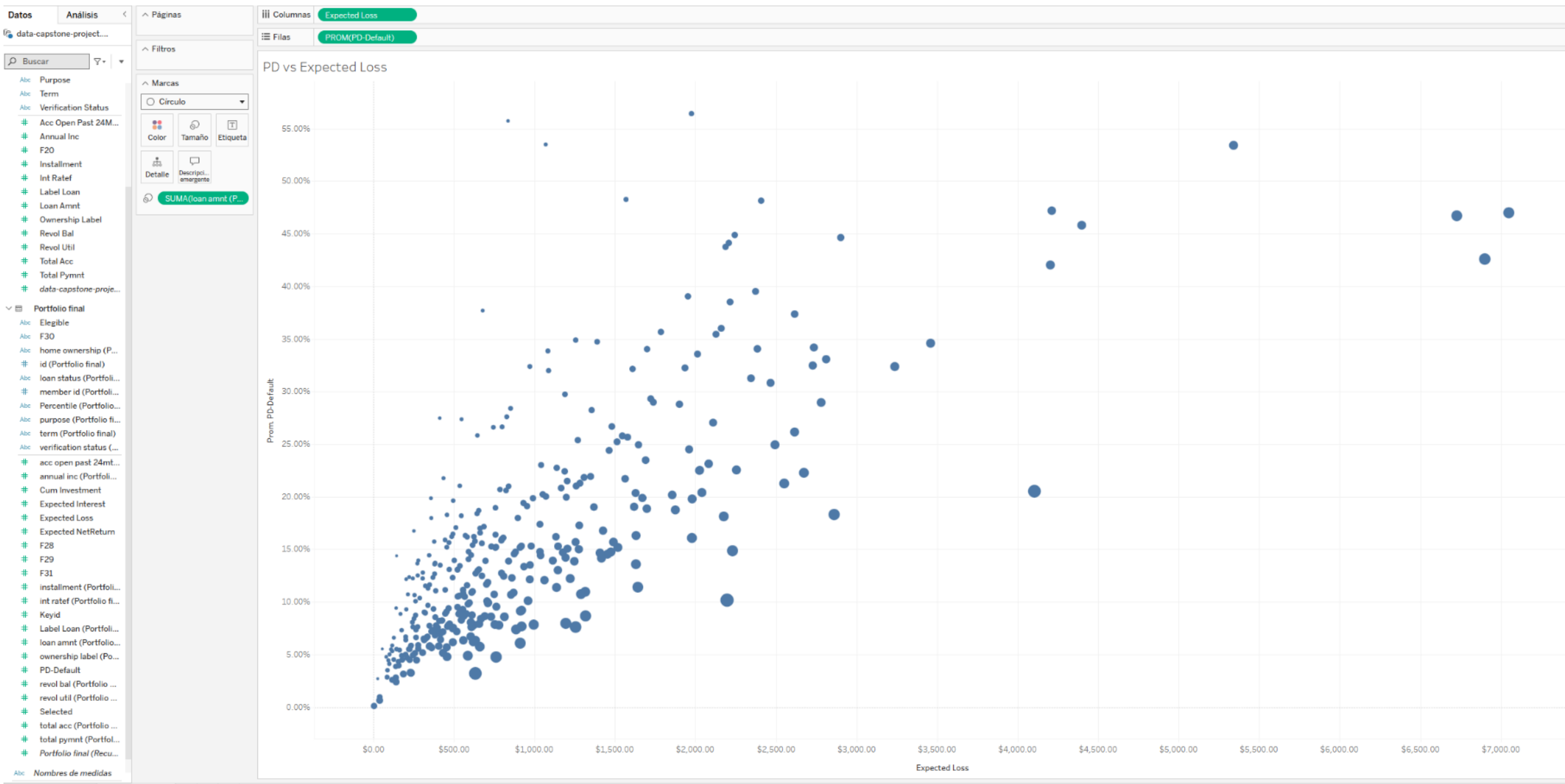
While probability of default and expected loss quantify downside risk, investment decision requires an estimate of expected return. To translate model outputs into an economic an economic decision framework a proxy was constructed loan level.

Expected_Interest	Expected_NetReturn
$ 786.6	$ (193.15)
$ 266.0	$ (143.56)
$ 933.1	$ (1.66)
$ 1,919.5	$ 277.80
$ 981.6	$ 95.86
$ 266.0	$ (492.93)
$ 409.5	$ (489.05)
$ 549.1	$ (382.24)
$ 549.5	$ (1,234.07)
$ 568.1	$ (404.56)
$ 880.3	$ 126.09
$ 200.4	$ (138.70)
$ 383.0	$ 118.43
$ 1,229.0	$ (2,230.56)
$ 1,766.9	$ (428.92)
$ 694.3	$ (367.66)
$ 383.0	$ (1,179.92)
$ 839.4	$ 571.43
$ 1,269.0	$ (605.73)
$ 1,639.2	$ 9.20
$ 617.1	$ (658.48)
$ 288.3	$ (357.42)
$ 757.4	$ (649.39)
$ 278.1	$ (809.47)
$ 405.6	$ (2,496.05)
$ 1,729.5	$ (500.13)
$ 999.0	$ (4,341.61)
$ 479.4	$ (188.22)
$ 886.3	$ 46.83
$ 573.3	$ 167.38
$ 778.8	$ (1,894.14)
$ 237,479.52	$ (125,606.62)

The resulting expected net return is conservative by design, reflecting a risk-controlled investment strategy and the use of simplified return proxies rather tan full cash flow modeling

Now will be explain with Visualization for better understanding using Tableau

PD vs Expected Loss



scatter plot illustrates the relationship between the predicted probability of default (PD) and the expected loss for individual loans.

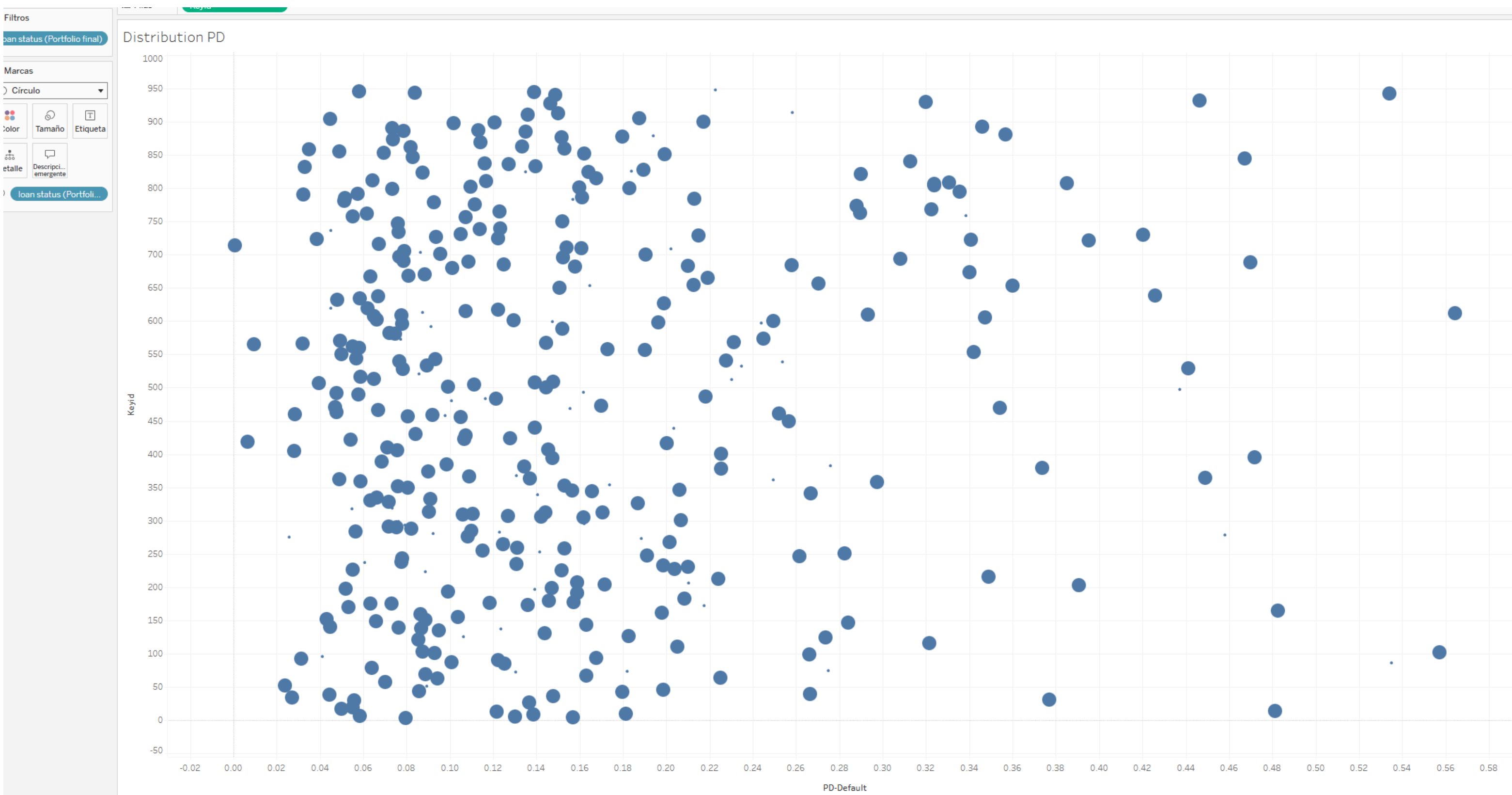
As expected, higher probabilities of default are associated with higher expected losses, reflecting the direct dependency between risk estimates and potential downside exposure.

The dispersion of points highlights heterogeneity in loan sizes, showing that loans with similar default probabilities can still carry very different loss magnitudes depending on the loan amount.

Distribution of Probability of Default

This scatter plot shows the distribution of predicted probabilities of default across the loan population.

The loans is concentrated at lower PD values, while a smaller number of loans exhibit significantly higher default risk.



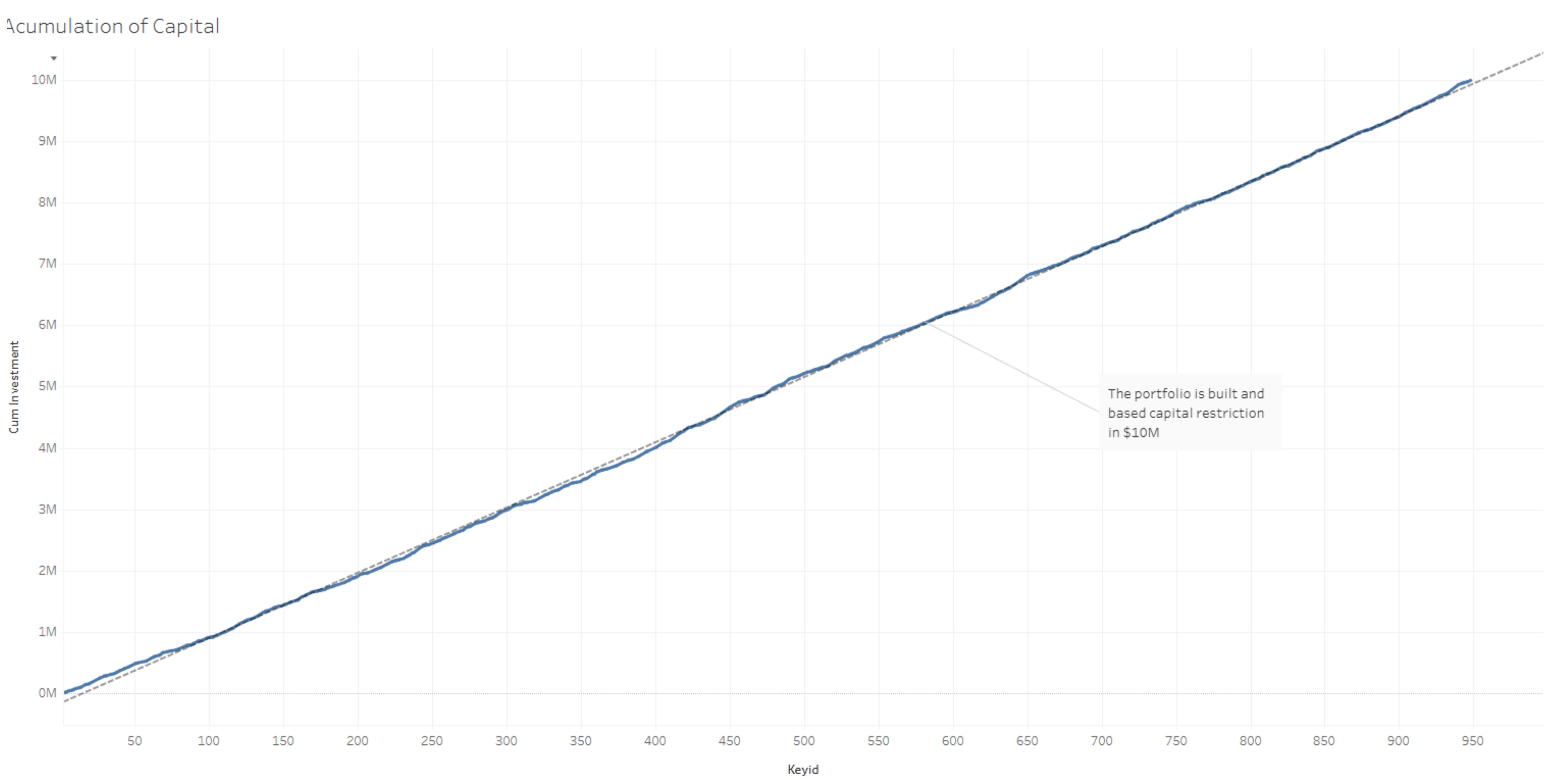
The PD have big concentration in Lower PD

Accumulation of Capital

The cumulative investment amount as loans are sequentially added according to the portfolio ranking criteria.

The upward trajectory reflects the mechanical accumulation of loan amounts when ordered by risk-adjusted priority.

While the cumulative curve approaches the $10M capital threshold, the final portfolio selection enforces a strict investment constraint, excluding loans that would breach the limit due to discrete loan sizes.



KPI BOX

A	B	C	D
Selected	1		
Row Labels	Count of KEYID	Average of PD-Default	Sum of Expected_Loss
credit_card	377	15.58%	$ 363,086.14
MORTGAGE	144	15.39%	$ 145,057.93
OWN	50	17.40%	$ 59,806.55
RENT	183	15.22%	$ 158,221.67
Grand Total	**377**	**15.58%**	**$ 363,086.14**

The final investment portfolio was evaluated using a concise set of key performance indicators that summarize capital deployment, risk exposure and expected economic impact, the EL was calculated a 100% Loss given default due to the absence of recovery or collateral data as a result overstate realized losses Although a total capital budget of 10M was available, only 2.4M was deployed. This reflects a **risk-controlled investment strategy**, where the capital was not forcibly allocated when loans exceeded acceptable risk thresholds or would breach the capital constraint due to discrete loan sizes, These KPIs provide a transparent and economically grounded summary of the final portfolio enabling decision-making

Finally, this project demonstrates how predictive analysis can be operationalized into a disciplined investment decision framework a logistic regression model was used to estimate the probability of default loans and was rigorously evaluated on validation data, achieving an AUC of 0.725, indicating solid discriminatory power. Under a fixed capital constraint of 10M, a cumulative investment approach was applied to an eligible loan ranked by expected loss. Due to conservative risk thresholds, discrete loan sizes and a strict enforcement of the capital constraint, the final deploy 2.4M this outcome reflect disciplined risk control rather than forced capital allocation, as resulting in portfolio controlled average default risk and transparent expected downside exposure, with expected losses intentionally overstated due to a 100% loss given-default assumption in the absence of recovery data.