Loan Risk Analytics | Uncovering Patterns and Reducing Exposure

Advanced Data Analytics Capstone

Kavin (Nik) Supatravanij

Overview

01 | Problem Statement

02 | Data Processing

03 | Data Analysis

04 | Key Insights

05 | Limitations, Assumptions and Next Steps

Problem Statement

"I'm a Branch Manager at DMS Bank, a mid-sized financial institution, and we've been struggling with **increasing loan defaults.**

We have a lot of **data on our borrowers**—things like age, income, employment length, and loan details—but we **still can't easily identify who's most likely to default.**

If I can analyze this data and spot key patterns, I'll be able to **better predict which borrowers are high-risk**, help reduce defaults, and improve our lending decisions."

- *Charlene Yip, Bank Manager at DMS Bank*

Problem Statement | Data Processing | Data Analysis | Key Insights | Limitations and Next Steps

Dataset: Credit Risk Dataset, 2019
Source: [Credit Risk Dataset | Kaggle]
Details: 32,581 rows, 12 fields

This dataset contains information on:

A) Borrower demographics (Age, Income, etc.)

B) Loan attributes (intent, interest rate, defaults, etc.)

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate
22	59000	RENT	123	PERSONAL	D	35000	16.0
21	9600	OWN	5	EDUCATION	B	1000	11.1
25	9600	MORTGAGE	1	MEDICAL	C	5500	12.8
23	65500	RENT	4	MEDICAL	C	35000	15.2
24	54400	RENT	8	MEDICAL	C	35000	14.2
21	9900	OWN	2	VENTURE	A	2500	7.1
26	77100	RENT	8	EDUCATION	B	35000	12.4
24	78956	RENT	5	MEDICAL	B	35000	11.1
24	83000	RENT	8	PERSONAL	A	35000	8
21	10000	OWN	6	VENTURE	D	1600	14.7
22	85000	RENT	6	VENTURE	B	35000	10.3
21	10000	OWN	2	HOMEIMPROVEMENT	A	4500	8.6
23	95000	RENT	2	VENTURE	A	35000	7
26	108160	RENT	4	EDUCATION	E	35000	18.3
23	115000	RENT	2	EDUCATION	A	35000	7
23	500000	MORTGAGE	7	DEBTCONSOLIDATION	B	30000	10.6
23	120000	RENT	0	EDUCATION	A	35000	7
23	92111	RENT	7	MEDICAL	F	35000	20.2
23	113000	RENT	8	DEBTCONSOLIDATION	D	35000	18.2
24	10800	MORTGAGE	8	EDUCATION	B	1750	10.9
25	162500	RENT	2	VENTURE	A	35000	7.4
25	137000	RENT	9	PERSONAL	E	34800	16.7
22	65000	RENT	4	EDUCATION	D	34000	17.5
24	10980	OWN	0	PERSONAL	A	1500	7.2
22	80000	RENT	3	PERSONAL	D	33950	14.5
24	67746	RENT	8	HOMEIMPROVEMENT	C	33000	12.6
21	11000	MORTGAGE	3	VENTURE	E	4575	17.7
23	11000	OWN	0	PERSONAL	A	1400	9.3
24	65000	RENT	6	HOMEIMPROVEMENT	B	32500	9.9
21	11389	OTHER	5	EDUCATION	C	4000	12.8
21	11520	OWN	5	MEDICAL	B	2000	11.1
25	120000	RENT	2	VENTURE	A	32000	6.6
26	95000	RENT	7	HOMEIMPROVEMENT	C	31050	14.1
25	306000	RENT	2	DEBTCONSOLIDATION	C	24250	13.8
26	300000	MORTGAGE	10	MEDICAL	C	7800	13.4
21	12000	OWN	5	EDUCATION	A	2500	7.5
22	48000	RENT	1	EDUCATION	E	30000	18.3
24	64000	RENT	8	DEBTCONSOLIDATION	D	30000	14.5
25	75000	RENT	4	HOMEIMPROVEMENT	D	30000	16.8
23	71500	RENT	3	DEBTCONSOLIDATION	D	30000	
26	62050	RENT	6	MEDICAL	E	30000	17.9
24	12000	OWN	4	VENTURE	B	2500	12.6
26	300000	MORTGAGE	10	VENTURE	A	20000	7.8
23	300000	OWN	1	EDUCATION	F	24250	19.4
26	300000	OWN	9	HOMEIMPROVEMENT	B	10000	10.3
26	300000	MORTGAGE	0	EDUCATION	D	25000	15.3
25	300000	MORTGAGE	9	HOMEIMPROVEMENT	E	18000	16.4
26	80690	RENT	8	PERSONAL	A	30000	7.4
22	66300	RENT	4	MEDICAL	B	30000	12.6
26	89028	RENT	0	DEBTCONSOLIDATION	A	30000	6.6

Dataset Preparation

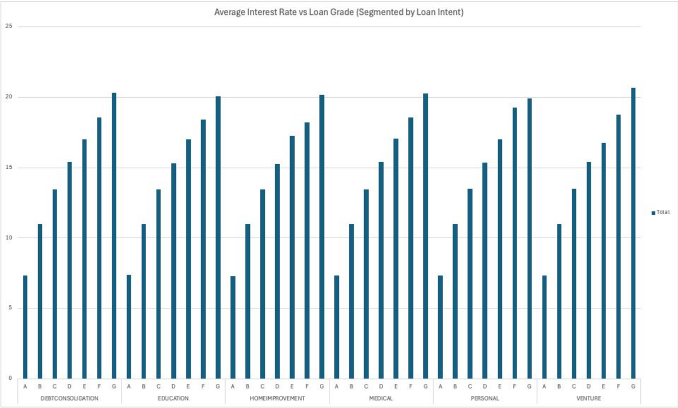
Excel:

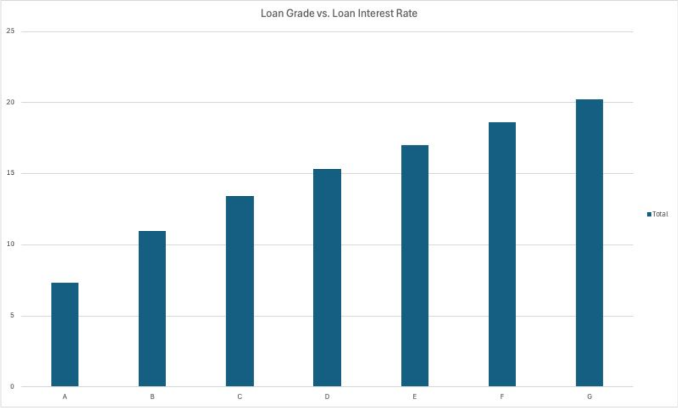
1. **314 duplicates:** since there is no unique ID in the raw data, it is uncertain whether these represent distinct data points and so these rows were kept

2. **Columns renamed** to improve readability

3. **Handle logical errors:**
   a. Remove rows where Years Employed > Person Age (2 rows deleted)
   b. Remove rows where Credit History Length > Age (No rows found)

4. **Missing data:**
   a. ~10% of rows have NULL Loan Interest rates: these were imputed with mean values of interest rate, segmented by loan grade. A PivotTable analysis showed that loan intent had no significant impact on average interest rates (see charts on the right), so segmentation was only done by loan grade. The missing values were filled using the IF/ISBLANK/INDEX/MATCH functions.
   b. ~3% of rows have NULL Years Employed: these were dropped as dataset is large and 3% would not represent a significant impact on data.





Dataset Preparation

Tableau:

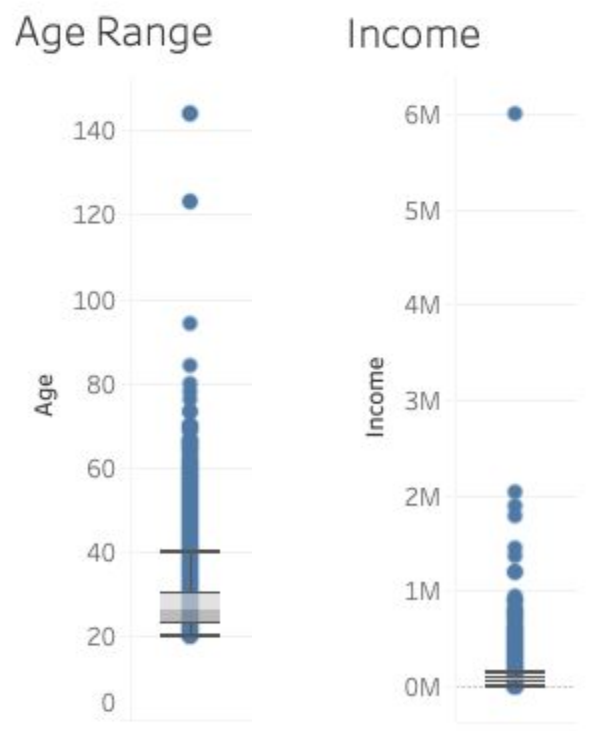
1. **Outliers:**
    a. Age: using a box plot, we can see 5 rows where customer age > 100. Considering the oldest person in the world was 122 years old (List of the verified oldest people - Wikipedia), it is highly likely these are errors and would skew the data, so the decision was made to delete these rows from the dataset on Excel
    b. There was an outlier for Income (Income = 6,000,000), however this was for one of the age outlier customers above, which was deleted along with the record for that outlier above. It is likely this particular record was an error.

Excel:

2. Added ID columns as no index in cleaned dataset

→ Final workable data: 31,679 rows (of 32,581 rows, 97.2% retained)

Age	Count of ID
123	2
144	3
Grand Total	**5**

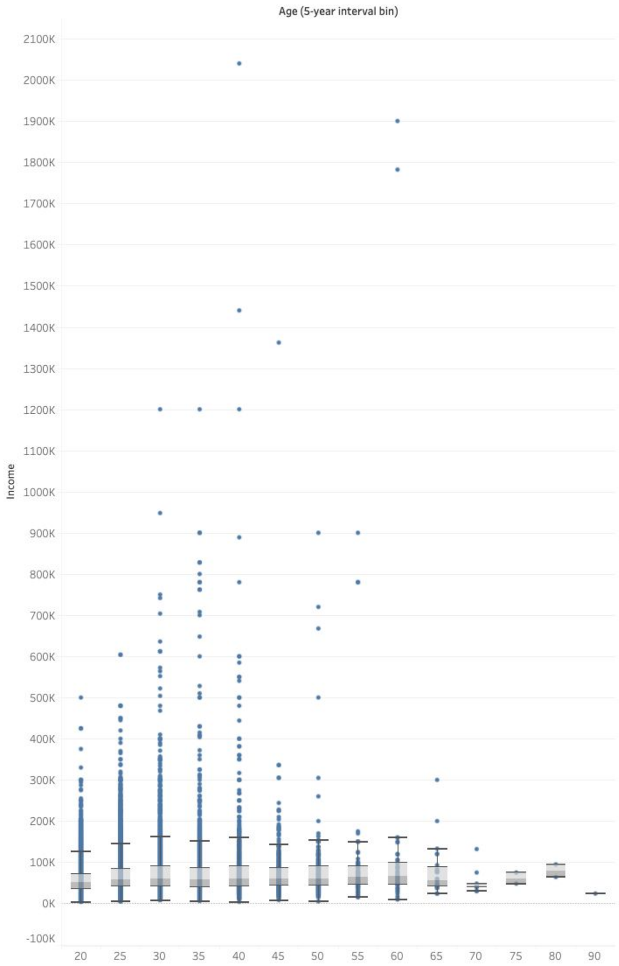


Understanding the customers

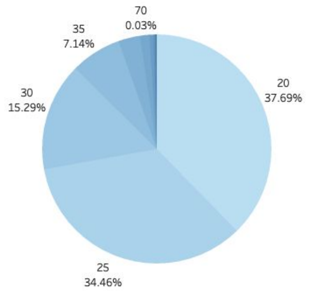
Demographic Breakdown by Age

- Majority of customers are young, between **20-30 years old** (these account for almost 75% of the total dataset)
- **Median income is ~$60,000** across all age ranges
- Loan intent among younger customers is split fairly evenly amongst all categories (with **Education loans being the highest amongst 20-40 year olds**). This shifts gradually toward **Medical and Personal loans as they get older**, with customers 80+ exclusively taking loans for only these 2 categories

Demo: Income Distribution by Age (bins)



Demo: Age Distribution (bins)



Demo: Loan Intent by Age

	Loan Intent					
Age (20-year i..	DEBTCONSOLID..	EDUCATION	HOMEIMPROVE..	MEDICAL	PERSONAL	VENTURE
20	16.10%	20.11%	10.96%	18.52%	16.67%	17.64%
40	14.17%	15.52%	13.62%	19.79%	20.77%	16.13%
60	11.84%	9.21%	5.26%	27.63%	42.11%	3.95%
80				66.67%	33.33%	

Understanding the customers

Demographic Breakdown by Home Ownership Status

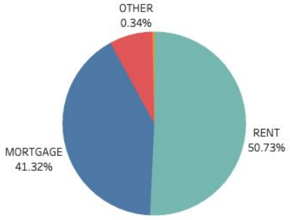
- **Most customers either rent (~50% of total) or mortgage (~41% of total)** their properties, with only a small percentage with other/full ownership status

- This **ratio remains relatively unchanged** as customers get older

- Customers with **higher Loan Grades (A) tend to own mortgages rather than rent** their properties, while this trend reverses as the loan grade decreases.

Home Own.. RENT MORTGAGE OWN OTHER

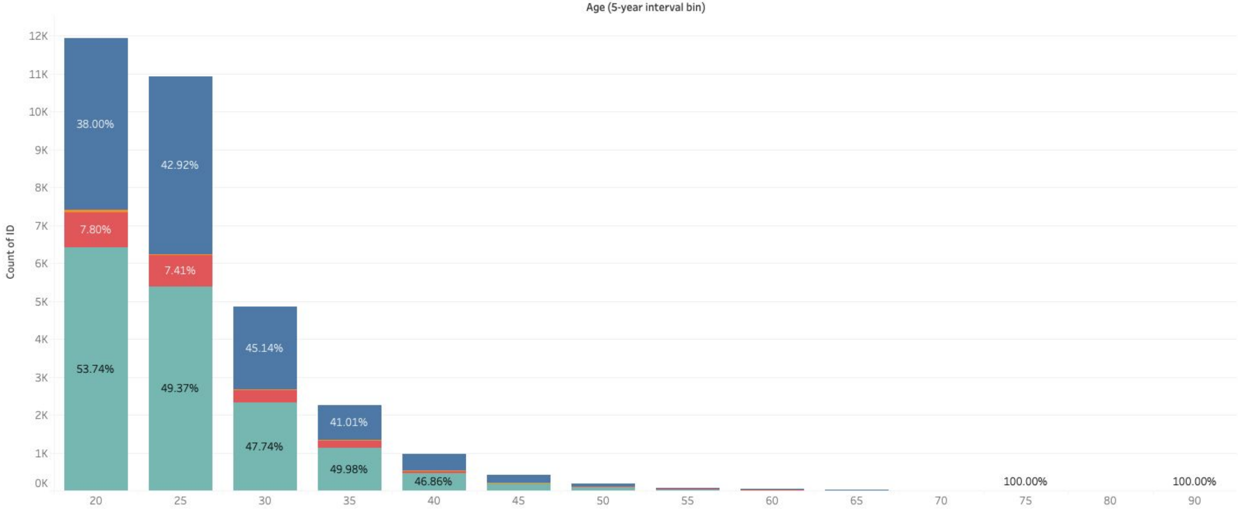
Demo: Home Ownership Status Total



Demo: Home Ownership by Loan Grade

Home Own..	A	B	C	D	E	F	G
MORTGAGE	50.68%	38.87%	35.13%	33.53%	35.50%	40.68%	48.44%
OTHER	0.25%	0.33%	0.27%	0.56%	0.84%	0.85%	
OWN	8.02%	7.46%	7.52%	7.62%	5.67%	5.51%	7.81%
RENT	41.05%	53.33%	57.08%	58.28%	57.98%	52.97%	43.75%

Demo: Home Ownership Status by Age (Stacked Bar)

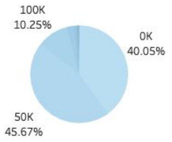


Understanding the customers

Demographic Breakdown by Income

- Majority of **customers (~85%) make less than $100k** in income
- Cross-tabulating Income vs. Age and Income vs. Home Ownership, we see that the **Loan Amount does not vary significantly by whether the customer rents/owns a mortgage or how old they are**
- **Income seems to play the biggest factor in determining the loan amount**, more than Home Ownership Status or even Age

Demo: Income Breakdown Total



Demo: Income vs. Home Ownership Determining Loan Amount

Income ($50k-interval bin)

Home Own..	0K	50K	100K	150K	200K	250K	300K	350K	400K	450K	500K	550K	600K	650K	700K	750K	800K	850K	900K	1200K	1350K	1400K	1750K	1900K	2000K
MORTGAGE	7,433	10,503	13,287	15,428	16,755	19,219	18,630	21,211	21,000	16,469	20,000	15,333	20,000	10,000	17,481	10,000	7,800		15,750	11,000	6,600	6,400		1,500	
OTHER	8,742	12,261	10,722	12,125	15,000								25,000												
OWN	6,894	10,485	14,187	15,483	20,684	20,000	19,583	32,000		24,000			16,000			10,000				10,000					
RENT	7,000	10,505	13,260	15,336	15,482	19,721	14,556	9,200	18,667		6,600	10,000	9,800		12,000	7,500		20,000	19,600				12,025		8,450

Demo: Income vs. Age Determining Loan Amount

Income ($50k-interval bin)

Age (5-year ..	0K	50K	100K	150K	200K	250K	300K	350K	400K	450K	500K	550K	600K	650K	700K	750K	800K	850K	900K	1200K	1350K	1400K	1750K	1900K	2000K
20	6,930	10,468	13,226	15,653	15,368	17,521	18,188	20,000	10,000		30,000														
25	7,318	10,522	13,405	15,633	17,122	19,352	16,210	18,400	19,333	19,600			25,000												
30	7,136	10,476	13,082	14,724	18,276	20,780	21,850	16,867	23,250	8,875	2,750	10,000	9,800		12,000	30,000			2,000	12,000					
35	6,977	10,736	13,434	15,035	14,928	18,400	18,767	13,333	23,583		11,075		13,000		25,000	8,750	7,800		15,500	10,000					
40	7,298	10,135	14,888	16,216	17,195	17,783	15,350	26,000	19,500	20,000	30,000	15,333	25,000			1,000		20,000		10,000		6,400			8,450
45	7,335	10,946	11,315	15,213	14,700		12,000														6,600				
50	6,759	10,765	13,211	15,029	7,000	14,000	5,000				20,000			10,000	7,925				30,000						
55	5,933	8,668	14,280	11,660												4,500			19,600						
60	7,977	13,017	16,100	21,333																			12,025	1,500	
65	9,677	20,731	16,667		3,000		28,000																		
70	6,114	7,800	14,000																						
75	3,000	15,000																							
80		8,400																							
90	6,500																								

Problem Statement | Data Processing | Data Analysis | Key Insights | Limitations and Next Steps

Understanding the customers

Demographic Breakdown by Income

- Interestingly, loan interest rates do not seem to be affected by income level
- Those with **higher incomes do not necessarily enjoy a lower interest rate**, and vice versa
- The **purpose of the loan also does not appear to play a factor** in the loan interest rate
- **Overall, loan grade is the biggest factor in determining loan interest rate**, all other factors being equal

Demo: Income vs Loan Grade Determining Loan Interest Rate

	Income ($50k-interval bin)																								
Loan Grade	0K	50K	100K	150K	200K	250K	300K	350K	400K	450K	500K	550K	600K	650K	700K	750K	800K	850K	900K	1200K	1350K	1400K	1750K	1900K	2000K
A	7.39	7.31	7.35	7.33	7.58	7.47	7.76	7.44	7.20	6.25	7.43	7.22			8.94	8.13		7.35		7.13	7.74	7.40		7.35	
B	11.00	11.00	11.02	11.00	11.14	10.84	10.89	10.93	10.32	11.11	10.79	10.54	9.94		11.22	11.42	11.12		11.23						
C	13.44	13.46	13.44	13.48	13.52	13.81	13.90	13.21	13.24		13.06	13.45	12.73	13.98	12.53				13.43				14.27		12.29
D	15.33	15.35	15.42	15.41	15.22	15.12	15.71	15.88	15.31	15.20	15.67		14.74		15.33										
E	16.97	16.93	17.21	17.20	18.03	18.18	17.19	17.82	16.45	14.38			17.19												
F	18.88	18.57	18.19	18.24	18.86	17.26	18.50			18.43															
G	20.14	20.10	20.42	21.09		22.48																			

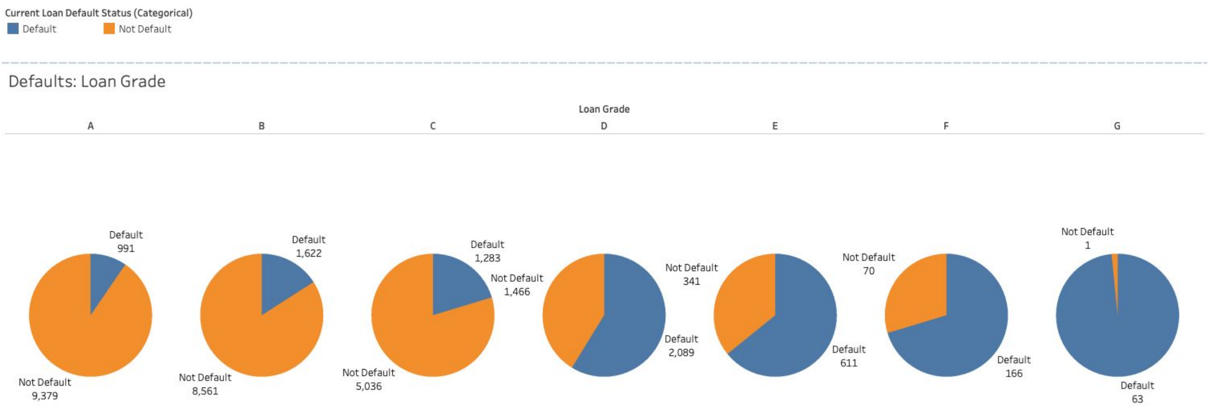
Demo: Income vs Loan Intent Determining Loan Interest Rate

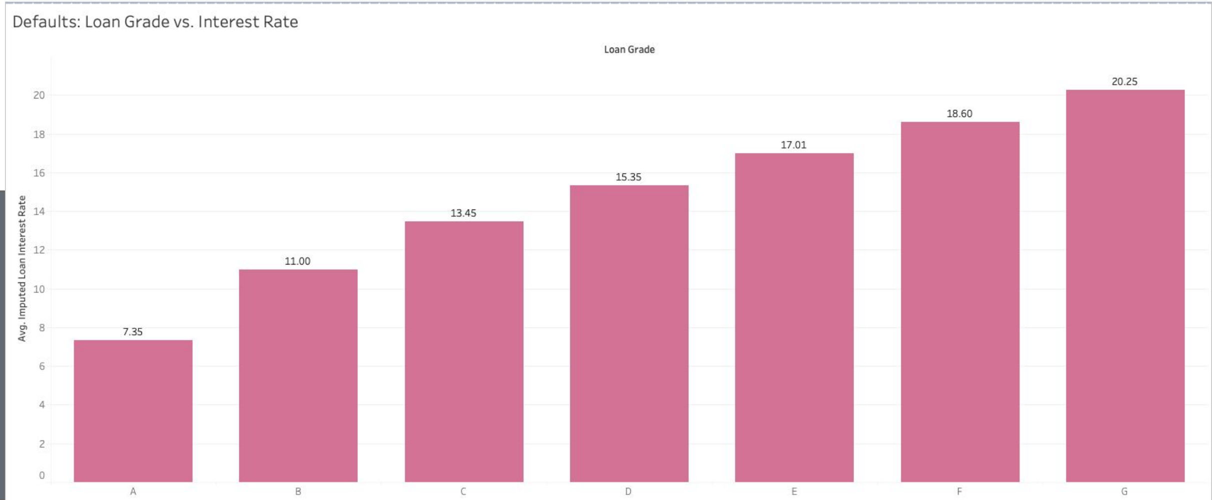
	Income ($50k-interval bin)																								
Loan Intent	0K	50K	100K	150K	200K	250K	300K	350K	400K	450K	500K	550K	600K	650K	700K	750K	800K	850K	900K	1200K	1350K	1400K	1750K	1900K	2000K
DEBTCONS..	11.349	10.789	10.968	10.813	11.322	12.514	11.327	11.984	12.730	10.013	12.038		12.105		12.180	9.910	11.360		13.335			7.400			
EDUCATION	11.055	10.914	10.998	11.001	11.369	11.502	13.991	9.990	11.433	14.875			14.740					7.347					14.270		
HOMEIMPR..	11.625	11.042	11.071	11.041	12.788	13.139	14.514	13.018	10.424		7.430		12.495			10.560	11.002								
MEDICAL	11.122	11.108	10.370	10.599	12.003	12.235	12.073	12.645	8.516	11.925	15.990	10.276	12.730		12.703	8.940	11.002			6.944					
PERSONAL	11.229	10.802	10.955	11.221	11.346	10.782	13.142	10.868	7.825	14.610	11.925	7.290	11.737	13.980	8.940	6.910			11.287					7.347	
VENTURE	11.112	10.782	11.018	11.189	11.922	11.964	12.037	13.278	11.855		13.979	11.120				12.180				7.510	7.740				12.290

Understanding loan defaults

Breakdown by Loan Grade

- Loan grade, loan defaults, and loan interest rates are all naturally correlated with each other
- If a person is awarded a high loan grade (A or B for example), it typically means they are less likely to default, and are given a lower interest rate as they are less risky to lend to



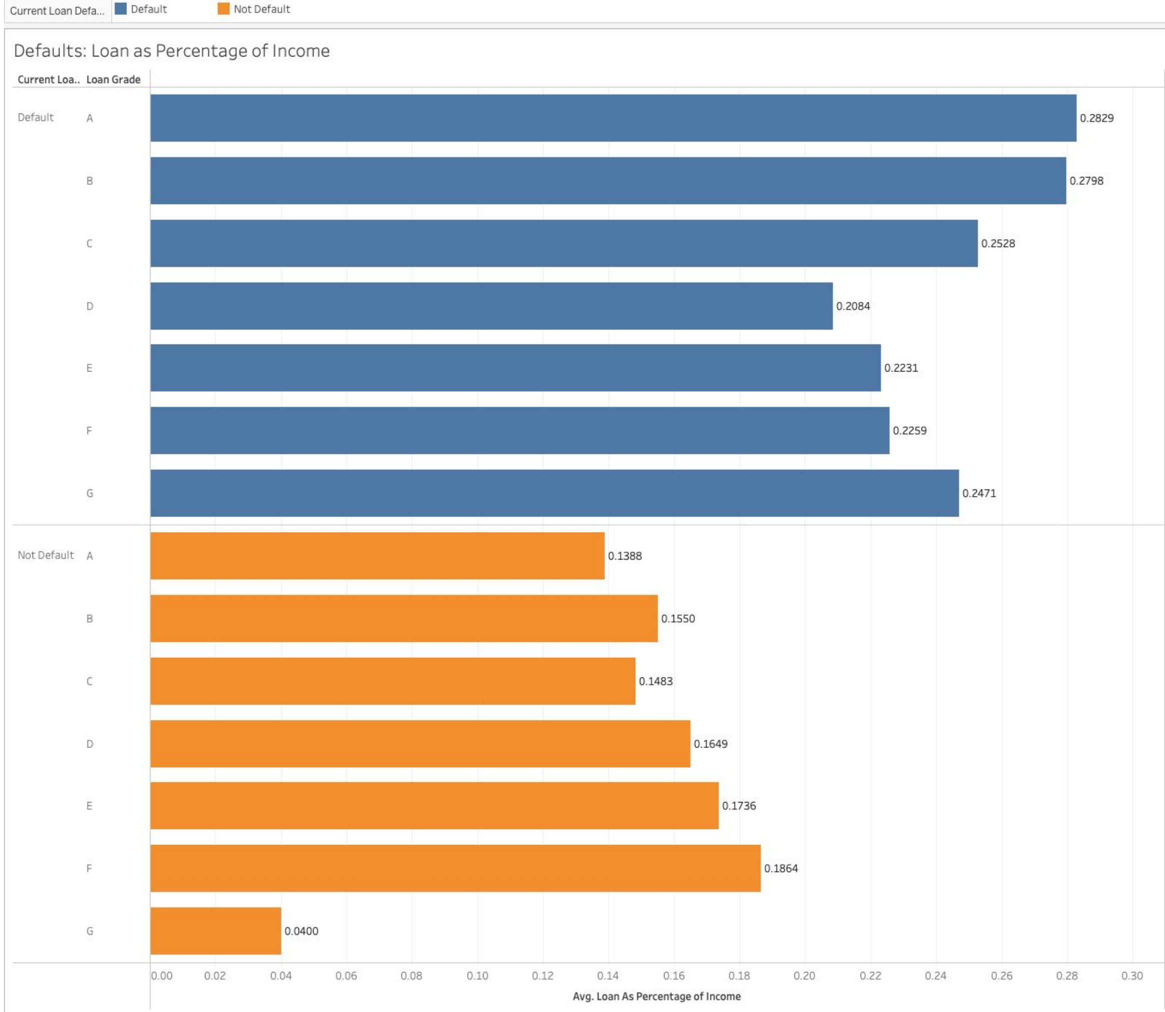


- Key insight: we can use **loan grade as a proxy for default risk** - so what behaviours/profiles determine what loan grade they get? **This will help us identify high-risk profiles in the future**.

Understanding loan defaults

Breakdown by Loan as Percentage of Income

- People who take a loan that is a **larger percentage of their income are more likely to default** on their loan
- Generally those that default on their loans take a **loan that is more than 20% of their income**
- This is true across all loan grades



Understanding loan defaults

Breakdown by Historical Defaults

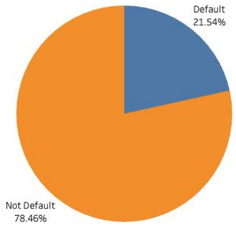
- **Most customers do not default (~80%)**, this is also true for historical defaults

- It is noted that for this reason, the **data is skewed towards non-default records**

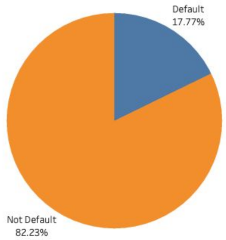
- We can also see that **Loan Grade A and B make up ~60% of the total**, although again this may be because the data is skewed towards good loans (those that do not, and are unlikely to default)

- Those that do not default historically tend not to default on current loans

Defaults: Total Defaults

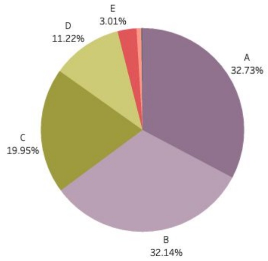


Defaults: Historical Defaults

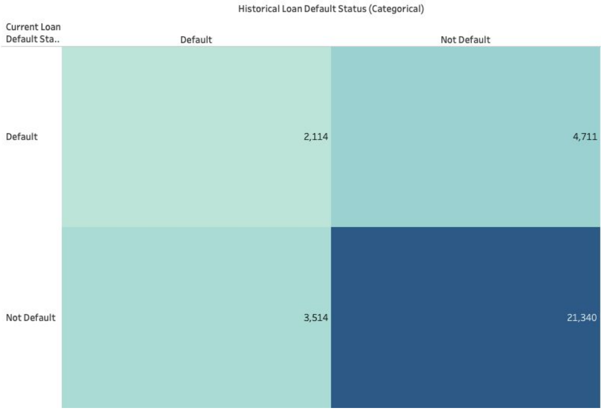


Loan Grade A B C D E F G

Defaults: Loan Grades



Defaults: Current vs. Historical Defaults

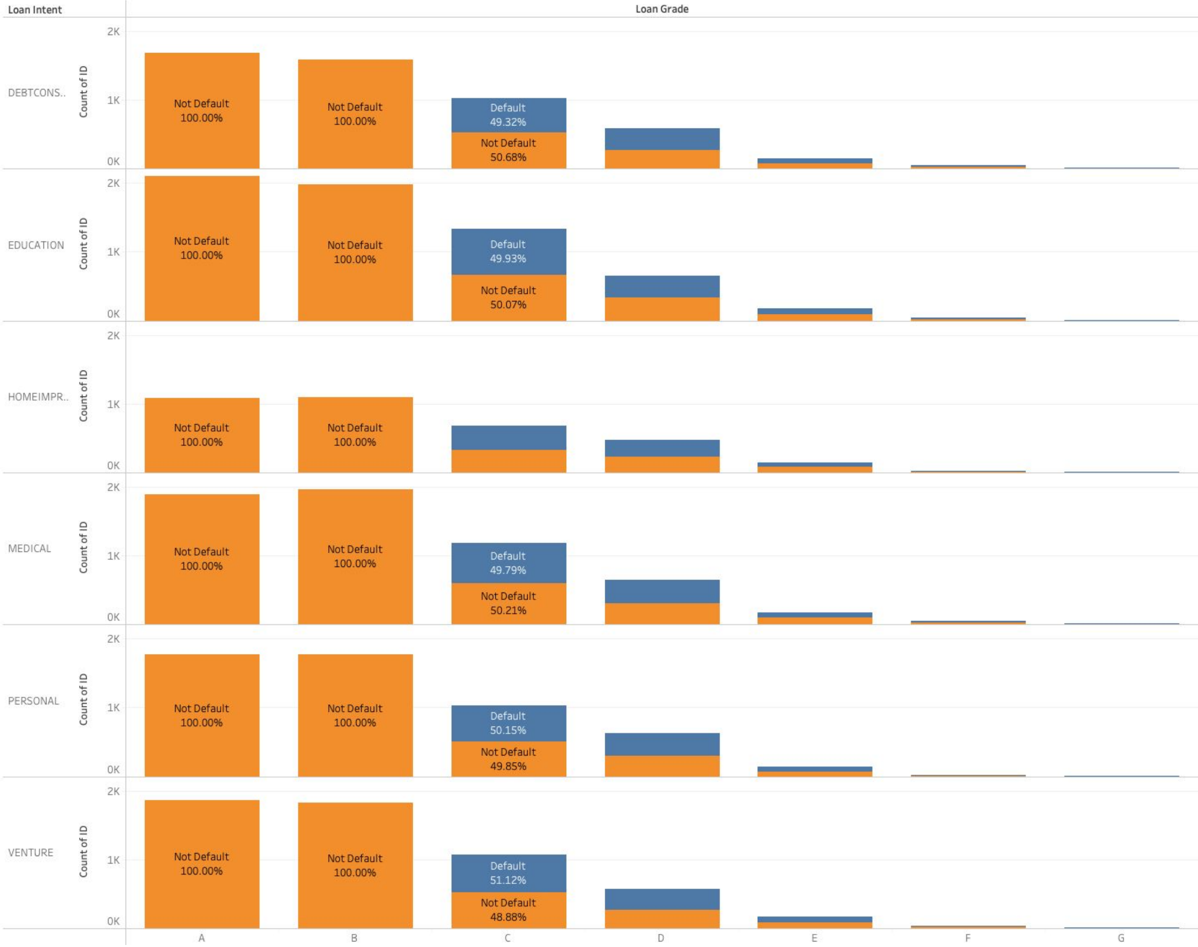


Understanding loan defaults

Breakdown by Historical Defaults

- While those with higher grade loans still default (although at lower rates than those with lower grades), **those who get a loan grade A or B have never defaulted**
- This means that historical default is a key indicator for identifying low-risk loans
- This is true across all types of loan intent as well

Defaults: Historical Status vs. Loan Grade



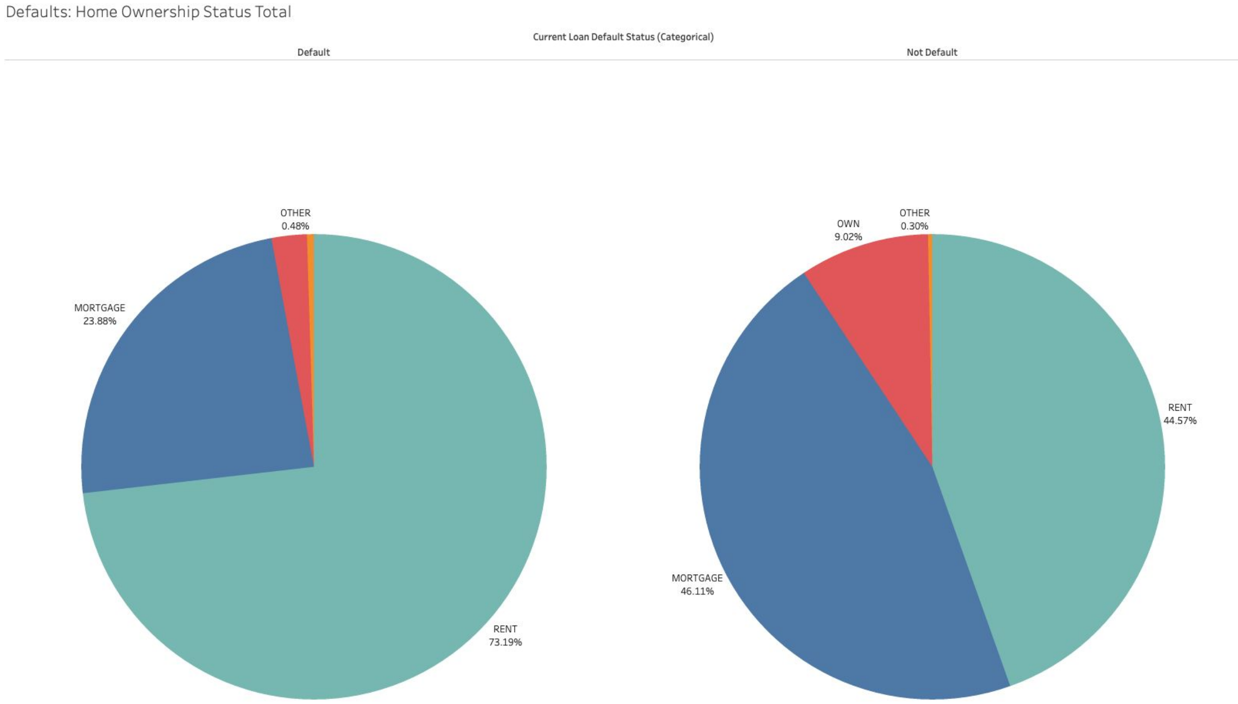
Understanding loan defaults

Breakdown by Home Ownership Status

- **Customers that have a mortgage appear less likely to default on their loan** than those that rent

- This may be because individuals with a mortgage have **already been approved for a housing loan, making them more reliable from a credit standpoint** and, therefore, less likely to default on additional loans they take on

- It is noted that correlation does not equate to causation, so **this observation would require further analysis in the future**



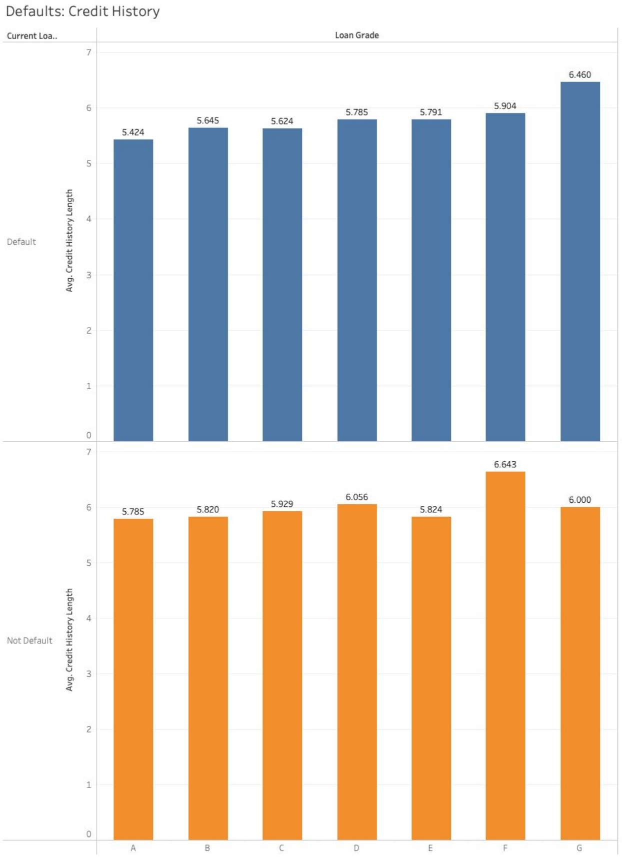
Understanding loan defaults

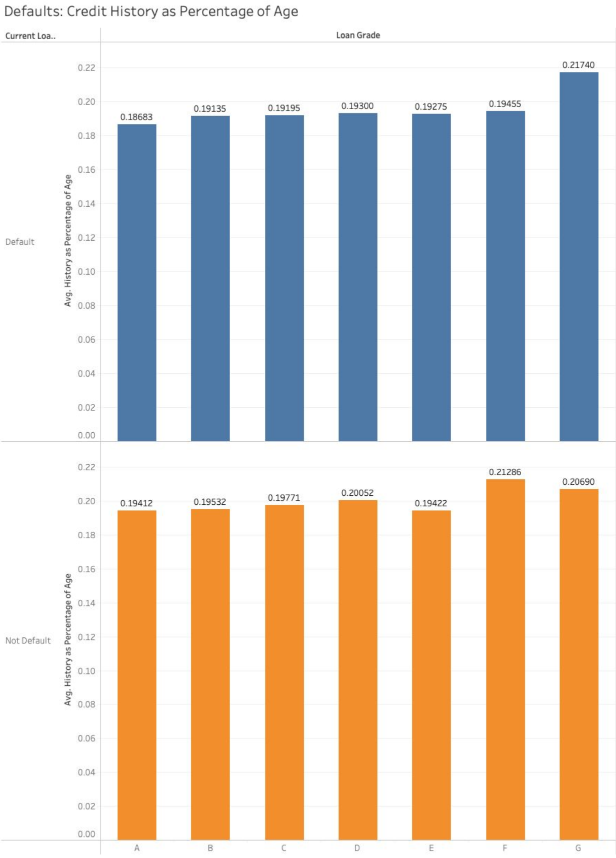
Breakdown by Credit History

- **Surprisingly, credit history length does not seem to be a predictor of default risk**, either absolute history length in years, or as percentage of customer age

- This may be due to the data within the dataset, however **across all loan grades the average credit history is ~5-6 years or ~20% of customer age**

- Both **default and non-default customers share these values**, signalling that credit history length does not vary between these two groups





Key Insights

What profiles are usually low-risk borrowers?

"Sarah is an example of a low-risk borrower: she is borrowing only around **15% of her total income, and has never defaulted on any of her loans before.**

She also has a mortgage, which she services every month. Although she makes an average salary, and has only been with the bank for ~2 years, she would be considered a good borrower and not likely to default.

The recommendation would be to give her a loan, and probably with a lower interest rate to keep her business with us"

Loan as % of income	Low % = ideally < 18% of total income
Historical default	Have not had any historical defaults
Home ownership status	Ideally have a mortgage, but rental or otherwise is OK
Income level	Not relevant
Credit history length	Not relevant
Age	Not relevant
Loan intent	Not relevant

Key Insights

What profiles are usually high-risk borrowers?

"Alvin is an example of a high-risk borrower: he is **borrowing quite a lot at 27% of his total income, even though he makes an above average salary.**

He is also a renter, and it is unclear whether he applied for a mortgage previously. Perhaps **most importantly is that he has had a history of defaults**: he has defaulted a few times on two other loans he has with the bank.

If we proceed to give him a loan, it is highly likely he will have a low grade loan of C or lower, with high interest rates because of his default risk."

Loan as % of income	High % = above 18% of total income
Historical default	May have had historical defaults
Home ownership status	Most likely a renter
Income level	Not relevant
Credit history length	Not relevant
Age	Not relevant
Loan intent	Not relevant

Key Recommendations

How can the loan process be improved?

1. **Collect more demographic information on customers**: currently in the dataset only Home Ownership Status appears to play a role in predicting low/high risk borrowers. Adding more data collection points could increase accuracy of this identification process.

2. **Understand more about historical defaults**: did the customer default more than once? What was the loan amount they defaulted on? This could help to segment the customers further.

3. **Target customers who have a mortgage**: those who have been approved for a mortgage are likely to have a good credit record already, and could be low-risk borrowers for other banking products.

4. **Offer micro-loans as a separate product**: since the median income is $60,000 amongst all ages, micro-loans of 5% (i.e. up to $3000) could be offered as a relatively low-risk product for all customers, regardless of whether they have a credit history with the bank.

Limitations, assumptions and next steps

Limitations

1. The dataset shows a strong bias towards younger borrowers, who tend to have shorter credit histories.
2. A majority of borrowers, across all age groups, fall within the low to middle-income range (below $150K USD annually).
3. Lack of comprehensive data on other key demographic factors (such as ethnicity, geographic location, gender, and occupation) as well as loan behaviours (repayment frequency, historical loan amounts, etc.)

Assumptions

4. Loan grade has not yet, or will not, account for loan default status. Borrowers who have defaulted on their loans have not experienced a downgrade in their loan grade.
5. Borrowers in the dataset are assumed to have consistent financial behavior, but fluctuations in income or unexpected expenses are not accounted for

Next Steps

6. Collect more data on key demographic factors and more granular loan performance data

Tableau Storyboard Visuals

[Loan Risk Analytics | Uncovering Patterns and Reducing Exposure]