Consumer loans in India

Team 2 - America

Cooper Atkins, Ricardo Diaz, Varun Shah, Alton Kessely

We analyzed consumer loans in India and the characteristics of those who did and didn’t default

Questions we posed

1. Do customers who default on loans have statistically lower **incomes** than those who don't default?
2. Does **homeownership** correlate with lower rates of default?
3. Does being **married** decrease the likelihood of default?
4. Does an additional **year of homeownership** reduce the likelihood of default?
5. Does **job experience** or **age** show a larger impact on someone defaulting on their loan?
6. Are customers who default on loans **younger** than those who do not?

Our dataset originally contained 13 variables and 252K observations

The dataset contains customer data including incomes (in rupees), job experience, home ownership, and other characteristics along with a Risk Flag signaling default vs. non-default loan status.

```
252,000 observations with 13 variables
```

```
 [1] "Id"               "Income"           "Age"              "Experience"       "Married.Single"
 [6] "House_Ownership"  "Car_Ownership"    "Profession"       "CITY"             "STATE"
[11] "CURRENT_JOB_YRS"  "CURRENT_HOUSE_YRS" "Risk_Flag"
```

2

Exploratory Data Analysis

In preparation, we dropped 4 unnecessary variables and converted 3 to factors

Changed the variables, Married.Single, Home Ownership & Car Ownership to factor (categorical) type

Dropped the variables: Id, City, State & Profession

Checked for Null values and Outliers

Means -> **Income**: ₹5 Million, **Experience**: 10 Years, **Age**: 50, **Current Home in Years**: 12 Years

Summary Statistics for Loan Default Prediction

	Income	Age	Experience	Married.Single	House_Ownership	Car_Ownership	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
Min	Min. : 10310	Min. :21.00	Min. : 0.00	Single :226272	Renting:231898	No :176000	Min. : 0.000	Min. :10	Non-Defaulted:221004
Q1	1st Qu.:2503015	1st Qu.:35.00	1st Qu.: 5.00	Married: 25728	Owning : 12918	Yes: 76000	1st Qu.: 3.000	1st Qu.:11	Defaulted : 30996
Median	Median :5000694	Median :50.00	Median :10.00	NA	Neither: 7184	NA	Median : 6.000	Median :12	NA
Mean	Mean :4997117	Mean :49.95	Mean :10.08	NA	NA	NA	Mean : 6.334	Mean :12	NA
Q3	3rd Qu.:7477502	3rd Qu.:65.00	3rd Qu.:15.00	NA	NA	NA	3rd Qu.: 9.000	3rd Qu.:13	NA
Max	Max. :9999938	Max. :79.00	Max. :20.00	NA	NA	NA	Max. :14.000	Max. :14	NA

Examining the two sub-populations, we see minor differences between the two groups

Not-Defaulted

Majority rent their home, own a car, and are not married

Means:

Age: 50 Years

Income: ₹5 Million

Experience: 10 Years

Current Job: 6.5 Years

Current Home: 12 Years

Defaulted

Majority rent their home, own a car, and are not married

Means:

Age: 49 Years

Income: ₹4.9 Million

Experience: 9.5 Years

Current Job: 6 Years

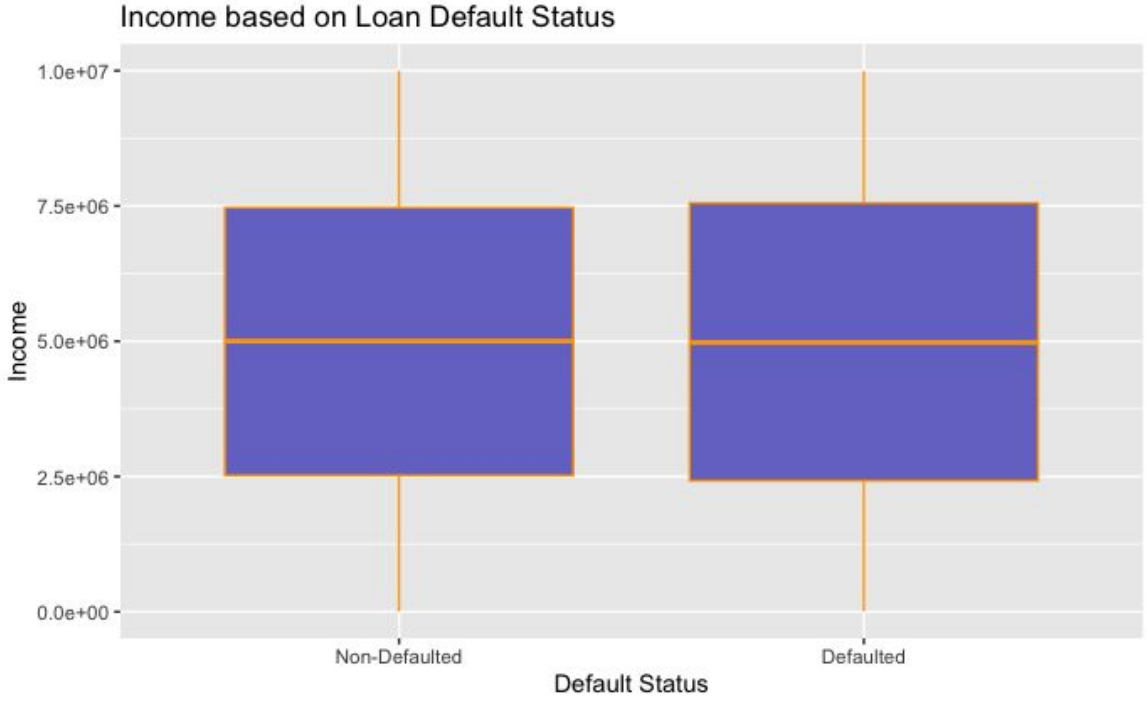
Current Home: 12 Years

3

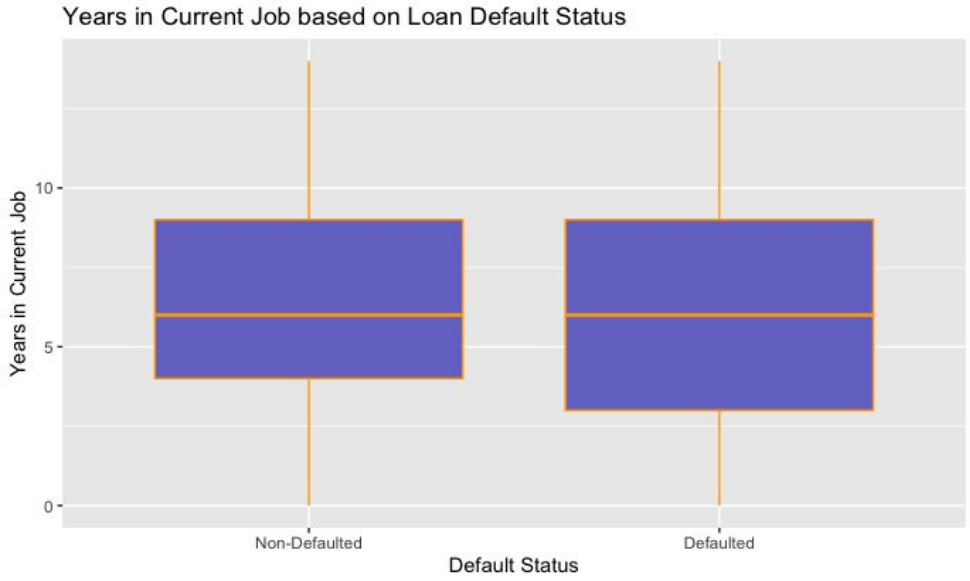
Variables Analysis

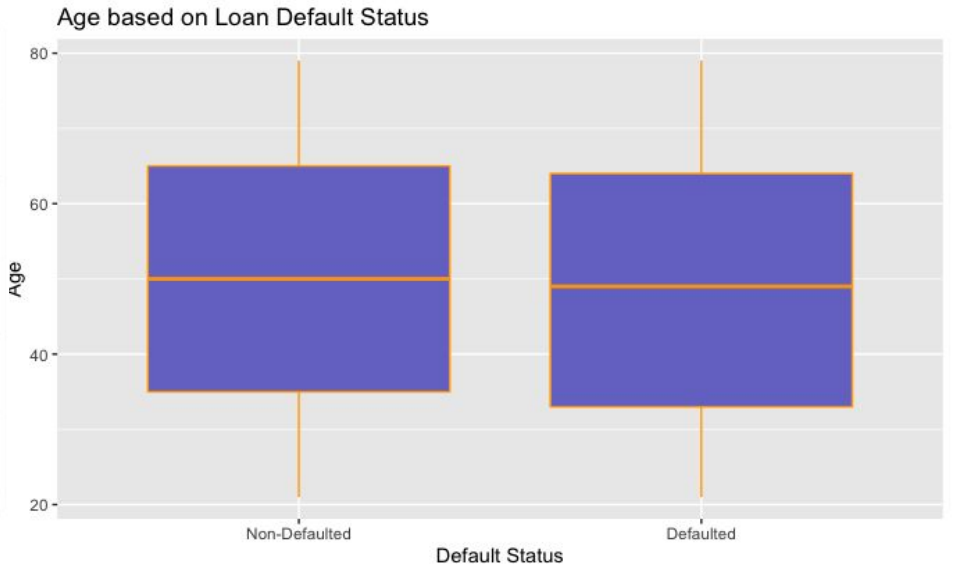
Defaulted vs. Non-Defaulted

Income Level doesn't appear to differ between defaulted and non-defaulted consumers



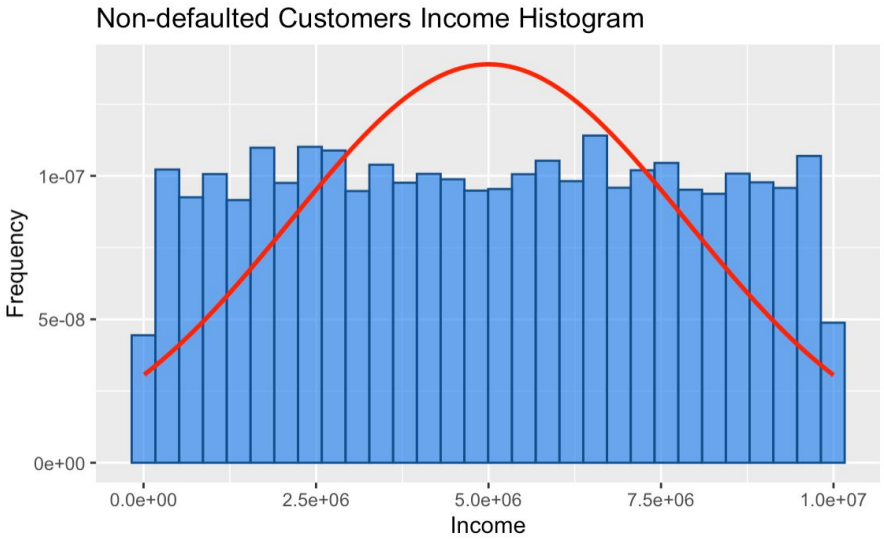
Years in Job and Age appear lower for those who have defaulted on loans



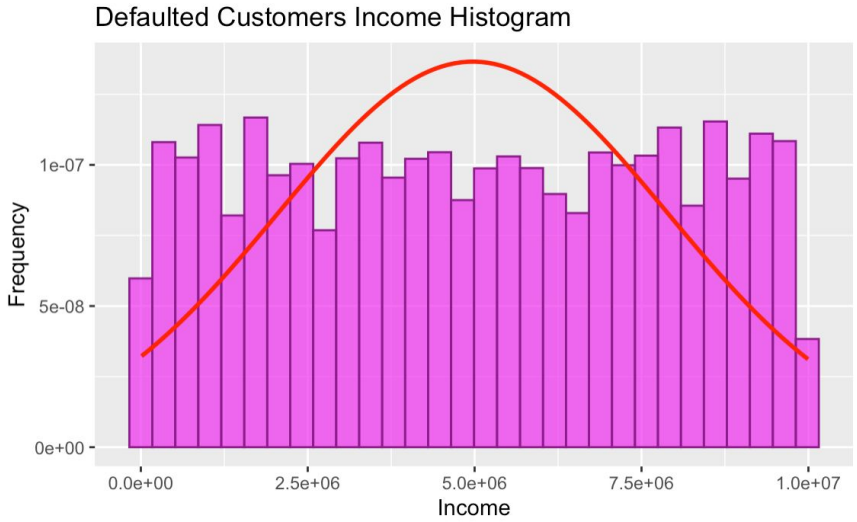


Income appears relatively uniformly distributed for both sample populations

Non-Defaulted

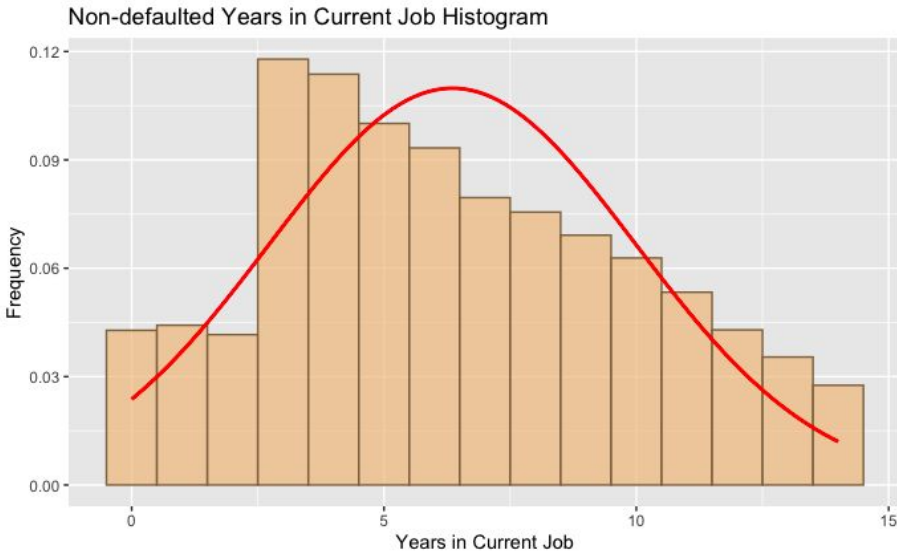


Defaulted

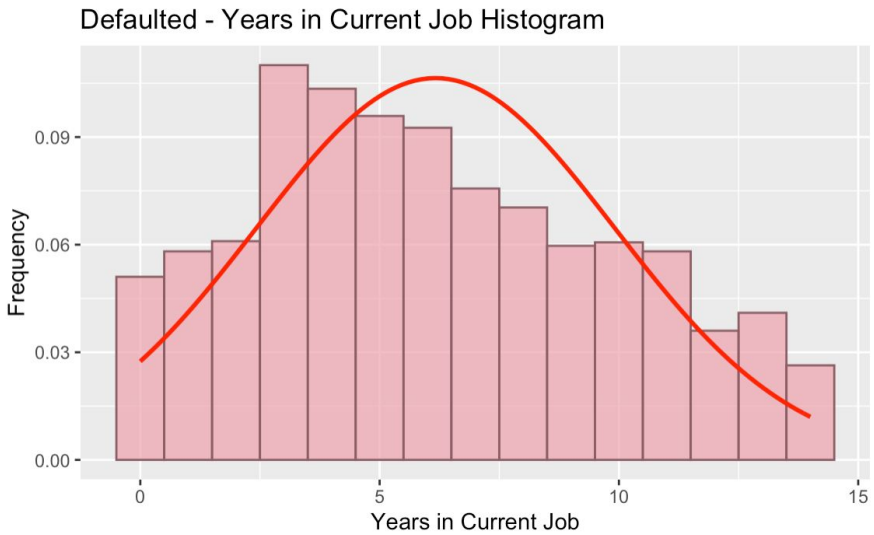


Years in Current Job follows an exponential distribution with a uniform left tail for both samples

Non-Defaulted

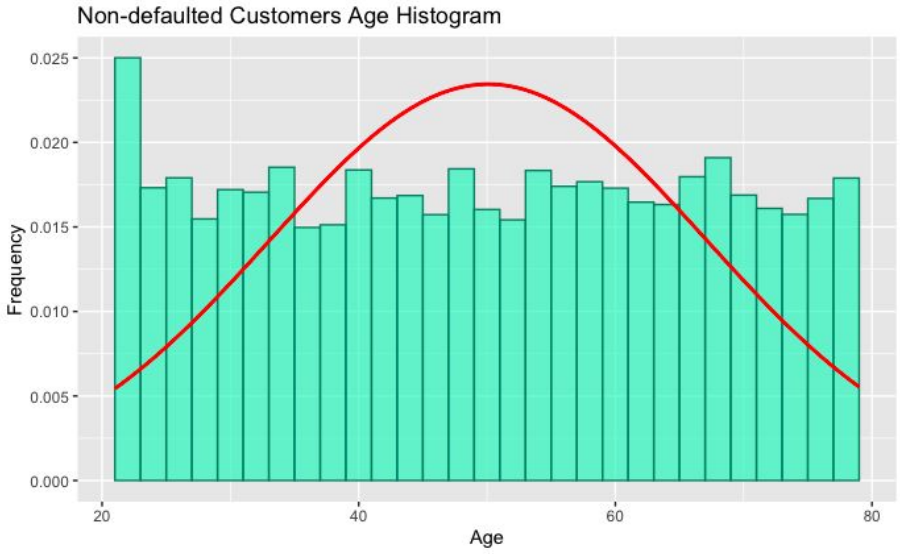


Defaulted

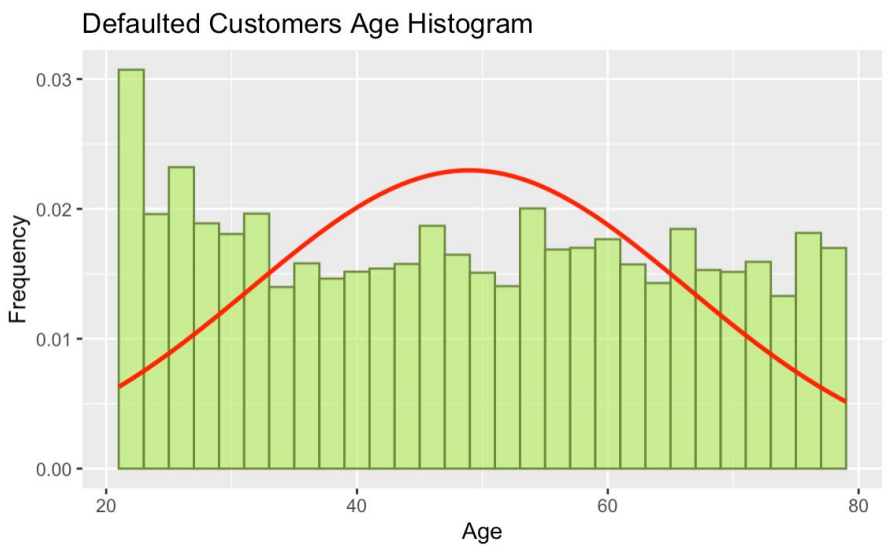


Age appears uniformly distributed for Defaulted and Non-defaulted customer samples

Non-Defaulted



Defaulted



4

Hypothesis Testing

Let’s start with the Chi-Squared Test

Marital status is not independent of default status

Question:

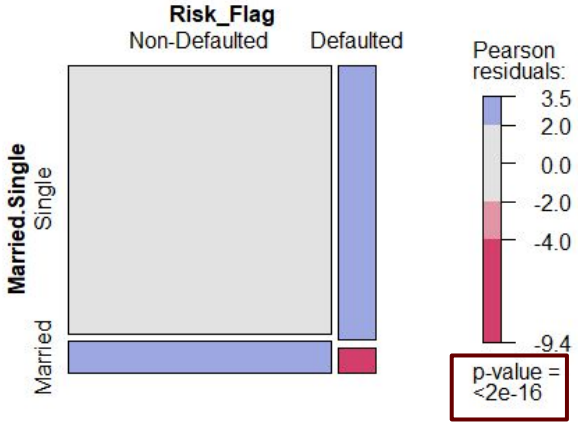
Does marital status have an effect on default status?

H_o : Marital Status and Default Status are independent

H_A : Marital Status and Default Status are dependent

Contigency table for Risk Flag vs Marital Status

	Non-Defaulted	Defaulted
Single	197912	28360
Married	23092	2636



α = 0.05

Average income is not statistically different for Defaulted and Non-Defaulted Customers

Question: Do customers who default on loans have statistically lower incomes than those who do not?

H_0 : Income Default ≥ Income Non-Defaulted

H_A : Income Default < Income Non-Defaulted

$\alpha = 0.05$

T-Test	
Degree of Freedom	39868
P-value	0.06

Defaulted customers are statistically younger than non-defaulted customers

Question: Are customers who default on loans younger than those who do not?

H_0 : Age Default ≥ Age Non-Defaulted

H_A : Age Default < Age Non-Defaulted

$\alpha = 0.05$

T-Test	
Degree of Freedom	39800
P-value	< 0.00002

Defaulted customers have statistically less time in their current house than non-defaulted customers

Question: Does an additional year of homeownership reduce the likelihood of default?

H_0 : Years in Current House Default ≥ Years in Current House Non Defaulted

H_A : Years in Current House Default < Years in Current House Non Defaulted

$\alpha = 0.05$

T-Test	
Degree of Freedom	40170
P-value	0.014

Through our EDA, we were able to answer the majority of our questions posed

Questions

1. Do customers who default on loans have statistically lower **incomes** than those who don't default?
2. Does **homeownership** correlate with lower rates of default?
3. Does being **married** decrease the likelihood of default?
4. Does an additional **year of homeownership** reduce the likelihood of default?
5. Does **job experience** or **age** show a larger impact on someone defaulting on their loan?
6. Are customers who default on loans **younger** than those who do not?

Answers

1. Not statistically significant
2. Statistically significant difference
3. Statistically significant difference
4. Defaulted is statistically significantly lower
5. Statistically significant difference
6. Defaulted is statistically significant lower

Questions

References

Data sourced from Kaggle:
https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior?select=Training+Data.csv