

Understanding Loan Defaulter: An Exploratory Data Analysis (EDA)

Key Takeaways from Analysis

Categorical Variables

- Lower loan grades (D, E, F, G) have higher default rates.
- "OWN" home ownership status corresponds to fewer loan applicants.
- Debt consolidation and borrowers from California (CA) show higher loan application numbers.

Relationship between Categorical Variables and Loan Default

- Lower loan grades (D, E, F, G) are linked to increased default likelihood.
- Certain sub-grades (F4 and G3) have notably high default rates.
- "Verified" income verification status may indicate higher default risk.
- Small business loans and borrowers from Nebraska (NE) have elevated default probabilities.

Numerical Variables and Loan Default:

- Higher interest rates (>12.5%) correlate with increased default risk.
- More credit inquiries in the last 6 months are associated with higher default rates.
- Higher loan and principal payments postissuance reduce default likelihood.
- Loans with ~35-month terms tend to have better repayment rates.
- Lower late fees and public record bankruptcies indicate full repayment likelihood.

Numerical Variable Relationships with Loan Default:

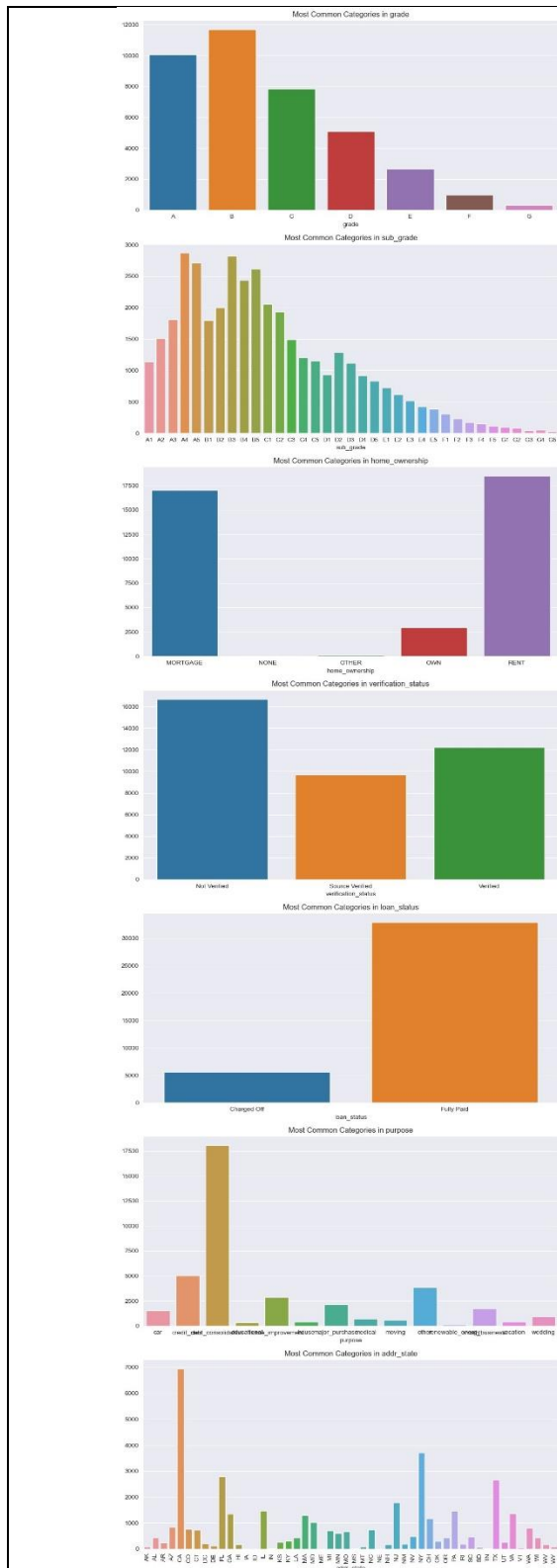
- Higher loan amounts (21k-23k and 29k-35k) exhibit higher default chances.
- Interest rates of 21%-23.5% have a 50% default probability.
- Employment lengths of 6.57 years and 9.5-10 years show higher default rates.
- Certain zip code areas (304xx 350xx and 850xx-999xx) have elevated default likelihood.
- Debt-to-income ratios between 11-26 raise default risk.
- More credit inquiries in the last 6 months are associated with higher default rates.
- Any number of derogatory public records (pub_rec) warrant thorough loan approval scrutiny.
- Higher revolving credit utilization (>16%) increases default probability.

Recommendations Based on EDA Analysis for Loan Defaulters

Risk-Based Pricing Model	Enhanced Credit Risk Assessment	Strengthened Verification Process	Targeted Intervention Strategies	Enhanced Credit Monitoring	Tighten loan approval criteria
<ul style="list-style-type: none">• Adjust interest rates based on identified risk factors like loan grade, sub-grade, employment length, debt-to-income ratio, and credit utilization to reflect borrower risk adequately.	<ul style="list-style-type: none">• Incorporate key driver variables such as loan grade, purpose (especially small business loans), employment length, zip code area, credit inquiries, and debt-to-income ratio into credit risk assessment models for improved accuracy	<ul style="list-style-type: none">• Improve income verification processes by implementing additional verification steps such as requesting supporting documents or conducting third-party verifications to address potential issues with "Verified" income status loans being defaulted	<ul style="list-style-type: none">• Implement personalized intervention strategies like financial counseling, debt consolidation, or alternative repayment plans for high-risk borrower segments identified through risk factors	<ul style="list-style-type: none">• Establish robust credit monitoring systems to track borrower credit behavior and identify early warning signs of default, including monitoring credit inquiries, credit utilization, and payment patterns	<ul style="list-style-type: none">• Implement stricter eligibility requirements for lower loan grades (D, E, F, G) and sub-grades (F4, G3)• Impose higher down payment or collateral requirements for small business loans• Consider declining loan applications from high-risk zip code areas (304xx-350xx and 850xx-999xx)

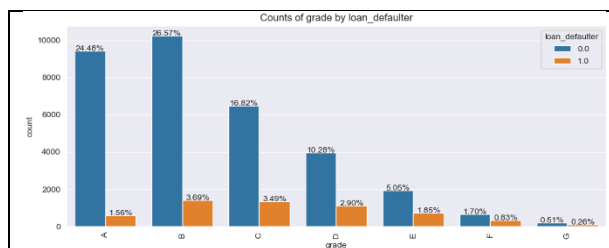
Detailed Analysis

1. Category Columns in BoxPlot



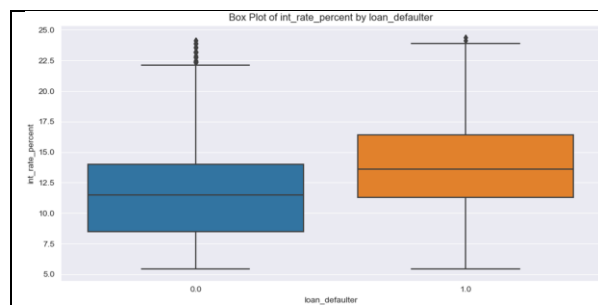
1. Very less Loan applicants in below:
 - a. Grade F and G
 - b. OWN house
2. A lot more Loan applicants are in below:
 - a. debt_consolidation
 - b. CA

2. CountPlot for Categorical Columns versus Loan Defalters

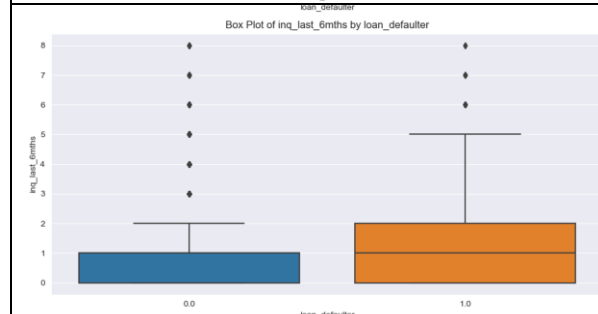


1. As the Grade towards D,E,F,G the percentage of having a defaulter is higher

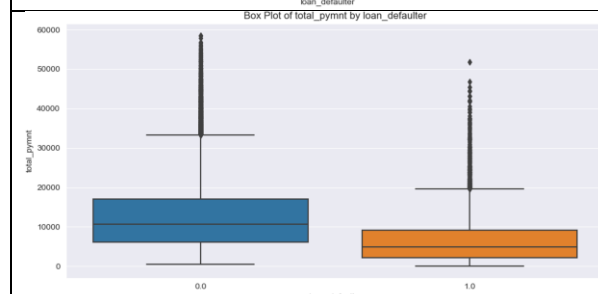
3. BoxPlot for Numerical columns versus Loan Defaulter



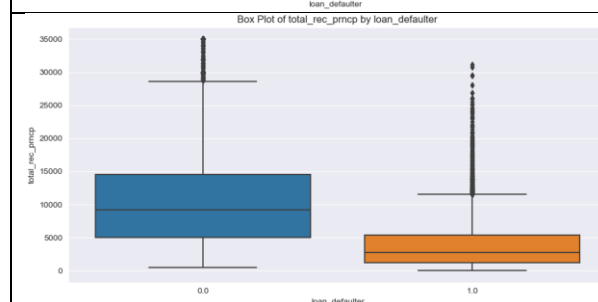
1. Higher Interest Rate has higher defaulter. esp., above 12.5%



2. Higher the enquiries on the credit status of the applicant, higher the chances of default

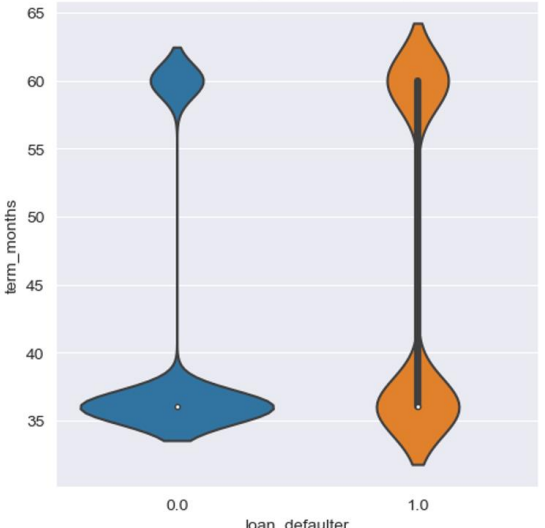
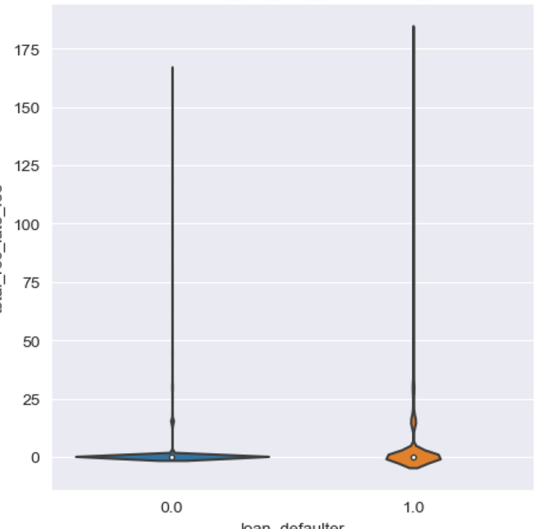
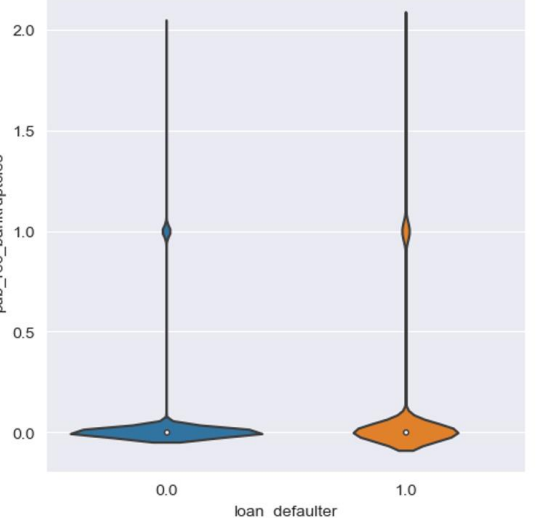


3. After loan issuance, higher the loan payment is done, lower are the chances of being default

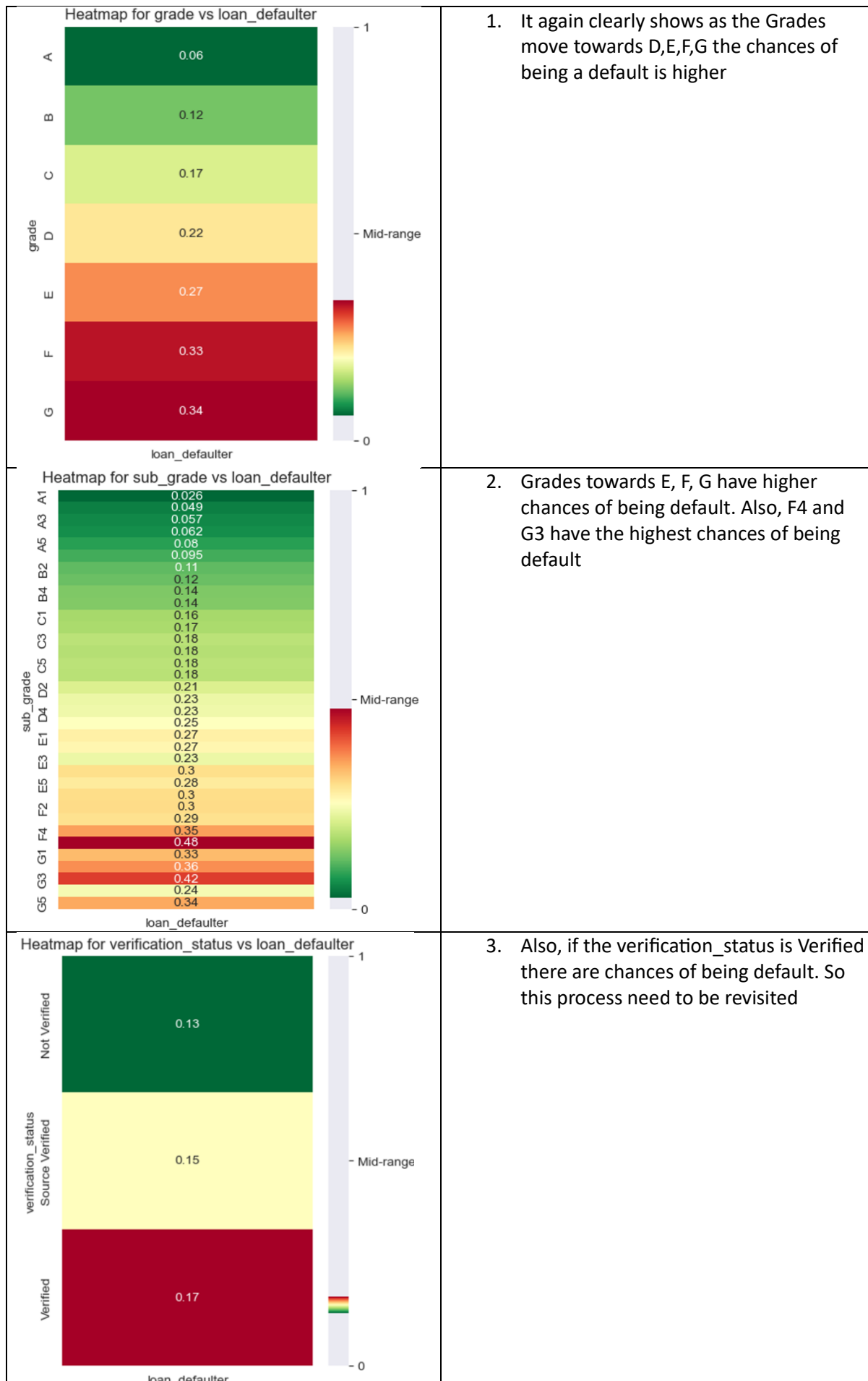


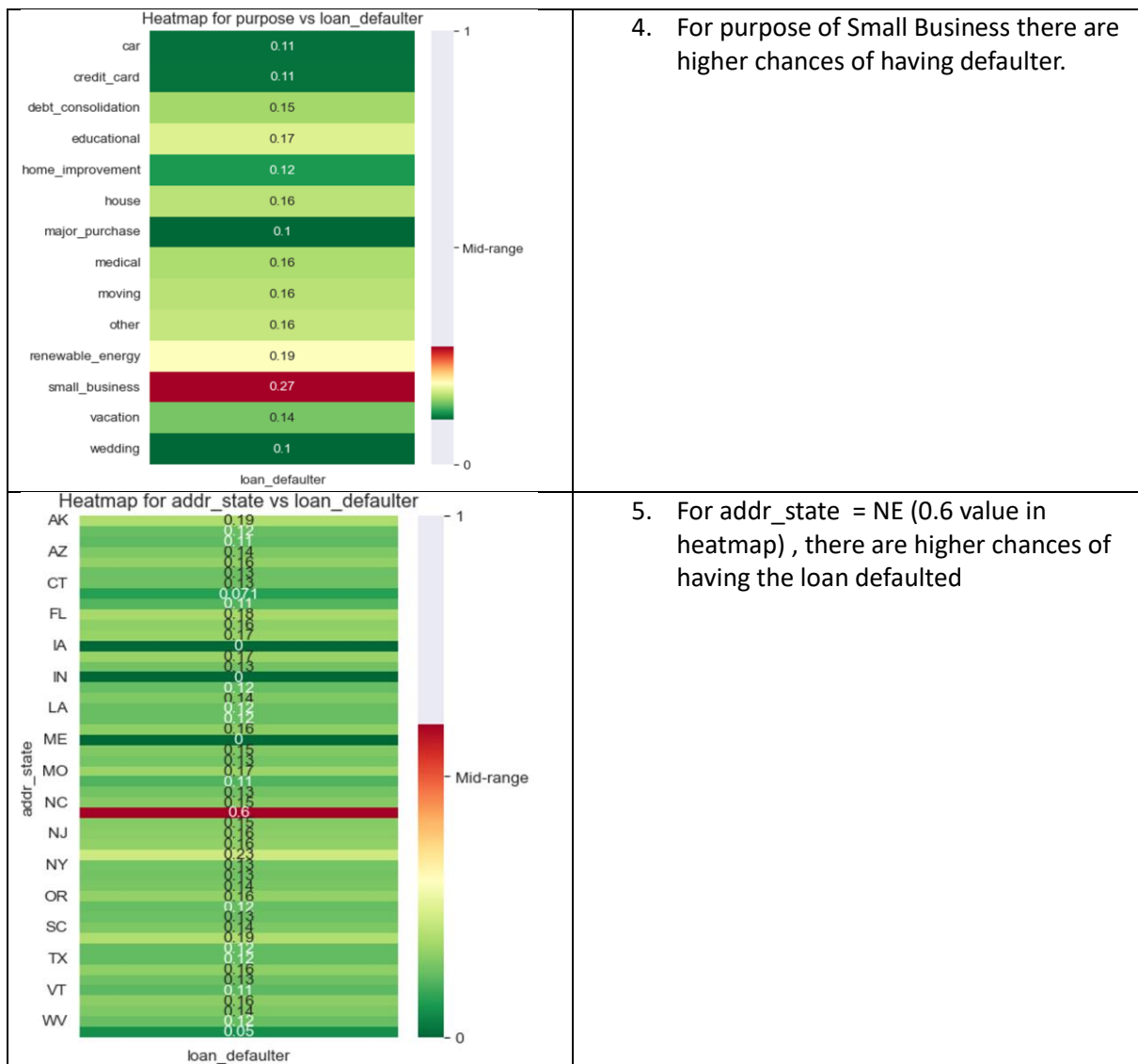
4. After loan issuance, higher the principal payment is done, lower are the chances of being default

4. Violin Plot for Numerical Columns versus Loan Defaulter:

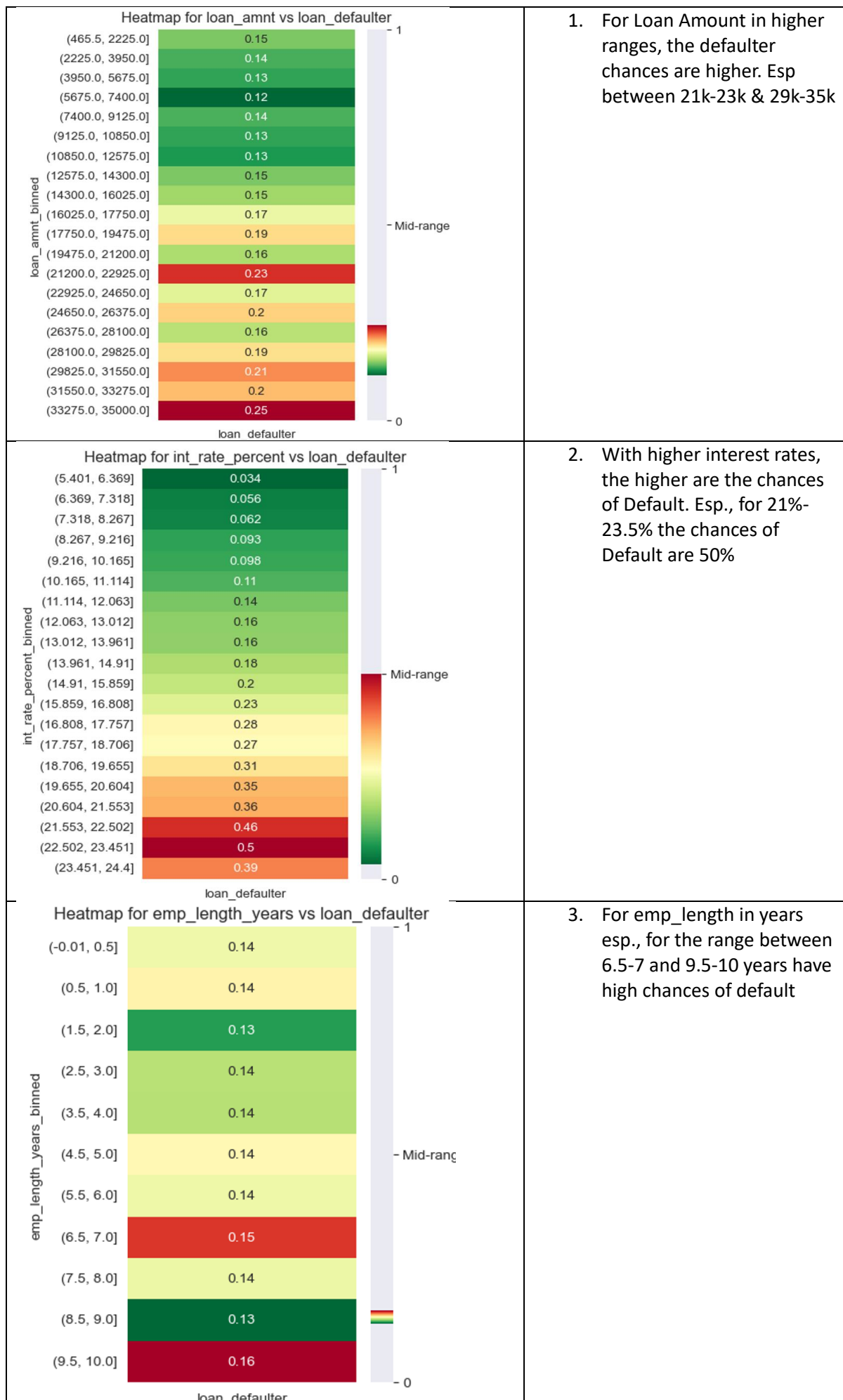
<p>Violin Plot of term_months by loan_defaulter</p> 	<ol style="list-style-type: none">1. There are more applicants at term_months around 35 months to be not Defaulted and have Fully Paid
<p>Violin Plot of total_rec_late_fee by loan_defaulter</p> 	<ol style="list-style-type: none">2. Lower the total_rec_late_fee higher the chances are to Fully Pay the loan. So having lesser recovery late fees would help in having the loans paid fully
<p>Violin Plot of pub_rec_bankruptcies by loan_defaulter</p> 	<ol style="list-style-type: none">3. Lower the pub_rec_bankruptcies higher the chances are to Fully Pay the loan. So the time pub_rec_bankruptcies are more its an indication of being a defaulter

5. Categorical Columns – Heatmap versus Loan Defaulter

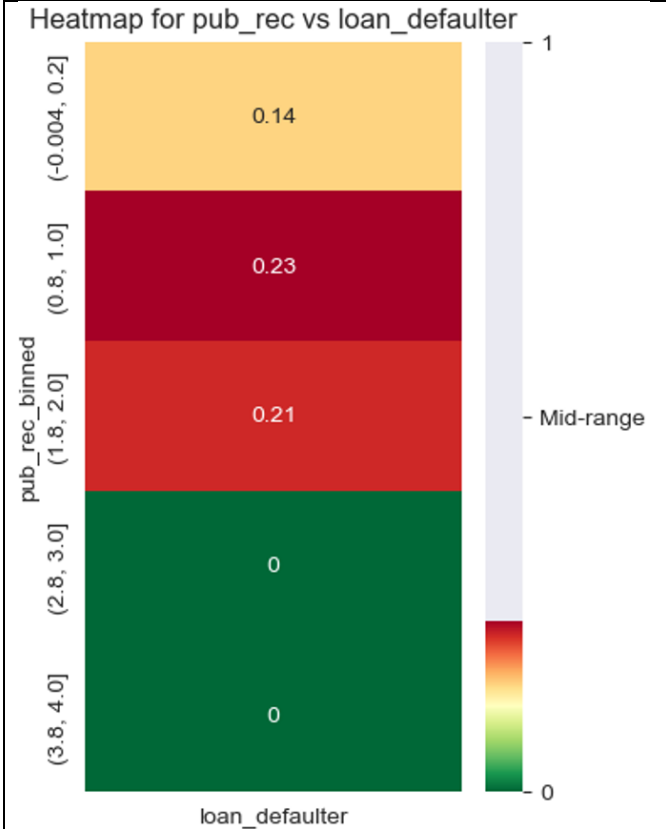




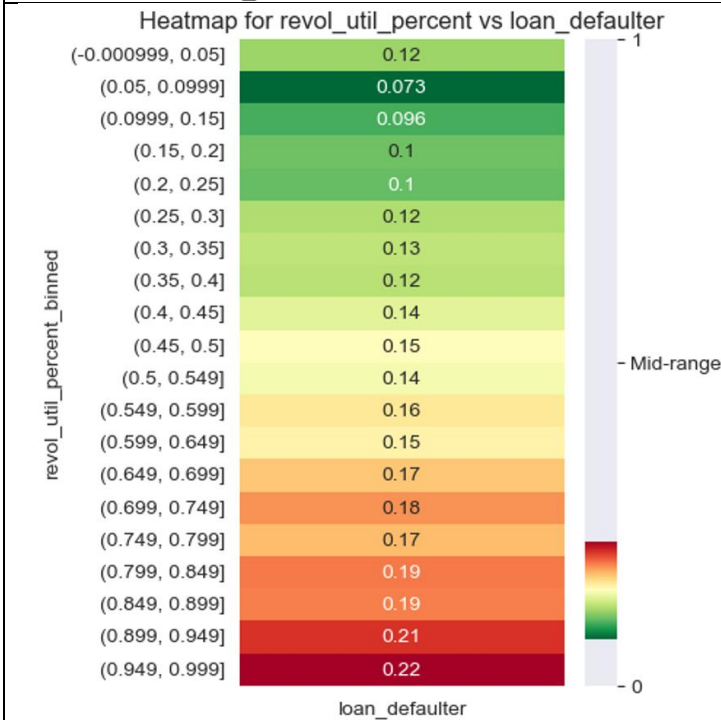
6. Numerical-Columns-Heatmap versus Loan Defaulter



<p>Heatmap for zip_code_area vs loan_defaulter</p> <p>zip_code_area_binned</p> <p>loan_defaulter</p> <p>Mid-range</p> <table border="1"> <thead> <tr> <th>zip_code_area_binned</th> <th>loan_defaulter</th> </tr> </thead> <tbody> <tr><td>(6.008, 56.6]</td><td>0.13</td></tr> <tr><td>(56.6, 106.2]</td><td>0.14</td></tr> <tr><td>(106.2, 155.8]</td><td>0.14</td></tr> <tr><td>(155.8, 205.4]</td><td>0.11</td></tr> <tr><td>(205.4, 255.0]</td><td>0.15</td></tr> <tr><td>(255.0, 304.6]</td><td>0.15</td></tr> <tr><td>(304.6, 354.2]</td><td>0.18</td></tr> <tr><td>(354.2, 403.8]</td><td>0.13</td></tr> <tr><td>(403.8, 453.4]</td><td>0.14</td></tr> <tr><td>(453.4, 503.0]</td><td>0.14</td></tr> <tr><td>(503.0, 552.6]</td><td>0.14</td></tr> <tr><td>(552.6, 602.2]</td><td>0.14</td></tr> <tr><td>(602.2, 651.8]</td><td>0.14</td></tr> <tr><td>(651.8, 701.4]</td><td>0.13</td></tr> <tr><td>(701.4, 751.0]</td><td>0.12</td></tr> <tr><td>(751.0, 800.6]</td><td>0.13</td></tr> <tr><td>(800.6, 850.2]</td><td>0.12</td></tr> <tr><td>(850.2, 899.8]</td><td>0.18</td></tr> <tr><td>(899.8, 949.4]</td><td>0.16</td></tr> <tr><td>(949.4, 999.0]</td><td>0.16</td></tr> </tbody> </table>	zip_code_area_binned	loan_defaulter	(6.008, 56.6]	0.13	(56.6, 106.2]	0.14	(106.2, 155.8]	0.14	(155.8, 205.4]	0.11	(205.4, 255.0]	0.15	(255.0, 304.6]	0.15	(304.6, 354.2]	0.18	(354.2, 403.8]	0.13	(403.8, 453.4]	0.14	(453.4, 503.0]	0.14	(503.0, 552.6]	0.14	(552.6, 602.2]	0.14	(602.2, 651.8]	0.14	(651.8, 701.4]	0.13	(701.4, 751.0]	0.12	(751.0, 800.6]	0.13	(800.6, 850.2]	0.12	(850.2, 899.8]	0.18	(899.8, 949.4]	0.16	(949.4, 999.0]	0.16	<p>4. For the zip_code area 304xx-350xx and 850xx-999xx , the chances of default are higher</p>
zip_code_area_binned	loan_defaulter																																										
(6.008, 56.6]	0.13																																										
(56.6, 106.2]	0.14																																										
(106.2, 155.8]	0.14																																										
(155.8, 205.4]	0.11																																										
(205.4, 255.0]	0.15																																										
(255.0, 304.6]	0.15																																										
(304.6, 354.2]	0.18																																										
(354.2, 403.8]	0.13																																										
(403.8, 453.4]	0.14																																										
(453.4, 503.0]	0.14																																										
(503.0, 552.6]	0.14																																										
(552.6, 602.2]	0.14																																										
(602.2, 651.8]	0.14																																										
(651.8, 701.4]	0.13																																										
(701.4, 751.0]	0.12																																										
(751.0, 800.6]	0.13																																										
(800.6, 850.2]	0.12																																										
(850.2, 899.8]	0.18																																										
(899.8, 949.4]	0.16																																										
(949.4, 999.0]	0.16																																										
<p>Heatmap for dti vs loan_defaulter</p> <p>dti_binned</p> <p>loan_defaulter</p> <p>Mid-range</p> <table border="1"> <thead> <tr> <th>dti_binned</th> <th>loan_defaulter</th> </tr> </thead> <tbody> <tr><td>(-0.03, 1.499]</td><td>0.12</td></tr> <tr><td>(1.499, 2.999]</td><td>0.12</td></tr> <tr><td>(2.999, 4.498]</td><td>0.13</td></tr> <tr><td>(4.498, 5.998]</td><td>0.11</td></tr> <tr><td>(5.998, 7.497]</td><td>0.13</td></tr> <tr><td>(7.497, 8.997]</td><td>0.13</td></tr> <tr><td>(8.997, 10.496]</td><td>0.13</td></tr> <tr><td>(10.496, 11.996]</td><td>0.14</td></tr> <tr><td>(11.996, 13.495]</td><td>0.15</td></tr> <tr><td>(13.495, 14.995]</td><td>0.15</td></tr> <tr><td>(14.995, 16.494]</td><td>0.15</td></tr> <tr><td>(16.494, 17.994]</td><td>0.16</td></tr> <tr><td>(17.994, 19.493]</td><td>0.17</td></tr> <tr><td>(19.493, 20.993]</td><td>0.17</td></tr> <tr><td>(20.993, 22.492]</td><td>0.17</td></tr> <tr><td>(22.492, 23.992]</td><td>0.17</td></tr> <tr><td>(23.992, 25.492]</td><td>0.16</td></tr> <tr><td>(25.492, 26.991]</td><td>0.15</td></tr> <tr><td>(26.991, 28.49]</td><td>0.14</td></tr> <tr><td>(28.49, 29.99]</td><td>0.13</td></tr> </tbody> </table>	dti_binned	loan_defaulter	(-0.03, 1.499]	0.12	(1.499, 2.999]	0.12	(2.999, 4.498]	0.13	(4.498, 5.998]	0.11	(5.998, 7.497]	0.13	(7.497, 8.997]	0.13	(8.997, 10.496]	0.13	(10.496, 11.996]	0.14	(11.996, 13.495]	0.15	(13.495, 14.995]	0.15	(14.995, 16.494]	0.15	(16.494, 17.994]	0.16	(17.994, 19.493]	0.17	(19.493, 20.993]	0.17	(20.993, 22.492]	0.17	(22.492, 23.992]	0.17	(23.992, 25.492]	0.16	(25.492, 26.991]	0.15	(26.991, 28.49]	0.14	(28.49, 29.99]	0.13	<p>5. For Debt-to-Income ratios of 11-26, there are chances of loan default.</p>
dti_binned	loan_defaulter																																										
(-0.03, 1.499]	0.12																																										
(1.499, 2.999]	0.12																																										
(2.999, 4.498]	0.13																																										
(4.498, 5.998]	0.11																																										
(5.998, 7.497]	0.13																																										
(7.497, 8.997]	0.13																																										
(8.997, 10.496]	0.13																																										
(10.496, 11.996]	0.14																																										
(11.996, 13.495]	0.15																																										
(13.495, 14.995]	0.15																																										
(14.995, 16.494]	0.15																																										
(16.494, 17.994]	0.16																																										
(17.994, 19.493]	0.17																																										
(19.493, 20.993]	0.17																																										
(20.993, 22.492]	0.17																																										
(22.492, 23.992]	0.17																																										
(23.992, 25.492]	0.16																																										
(25.492, 26.991]	0.15																																										
(26.991, 28.49]	0.14																																										
(28.49, 29.99]	0.13																																										
<p>Heatmap for inq_last_6mths vs loan_defaulter</p> <p>inq_last_6mths_binned</p> <p>loan_defaulter</p> <p>Mid-range</p> <table border="1"> <thead> <tr> <th>inq_last_6mths_binned</th> <th>loan_defaulter</th> </tr> </thead> <tbody> <tr><td>(-0.008, 0.4]</td><td>0.12</td></tr> <tr><td>(0.8, 1.2]</td><td>0.16</td></tr> <tr><td>(1.6, 2.0]</td><td>0.17</td></tr> <tr><td>(2.8, 3.2]</td><td>0.21</td></tr> <tr><td>(3.6, 4.0]</td><td>0.16</td></tr> <tr><td>(4.8, 5.2]</td><td>0.19</td></tr> <tr><td>(5.6, 6.0]</td><td>0.25</td></tr> <tr><td>(6.8, 7.2]</td><td>0.29</td></tr> <tr><td>(7.6, 8.0]</td><td>0.21</td></tr> </tbody> </table>	inq_last_6mths_binned	loan_defaulter	(-0.008, 0.4]	0.12	(0.8, 1.2]	0.16	(1.6, 2.0]	0.17	(2.8, 3.2]	0.21	(3.6, 4.0]	0.16	(4.8, 5.2]	0.19	(5.6, 6.0]	0.25	(6.8, 7.2]	0.29	(7.6, 8.0]	0.21	<p>6. For more inquiries in credit status of the applicant in last 6 months have higher chances of default</p>																						
inq_last_6mths_binned	loan_defaulter																																										
(-0.008, 0.4]	0.12																																										
(0.8, 1.2]	0.16																																										
(1.6, 2.0]	0.17																																										
(2.8, 3.2]	0.21																																										
(3.6, 4.0]	0.16																																										
(4.8, 5.2]	0.19																																										
(5.6, 6.0]	0.25																																										
(6.8, 7.2]	0.29																																										
(7.6, 8.0]	0.21																																										



7. With number of derogatory public records between 0-2 , there are higher chances of default. Even in case of higher than 2, it seem the value is 0. So for any number of pub_rec there needs to be proper investigation for loan approval



8. With higher percent of the credit used by the applicant among the total available credit to him from various sources, the higher chances of default. Esp., if the revol_util_percent is higher than 0.16 then more chance for default