Lending Club Case Study

Problem Statement:

- Given the past data of a lending company, apply the basics of data analysis and come up with meaningful conclusions about its defaulters

Approach:

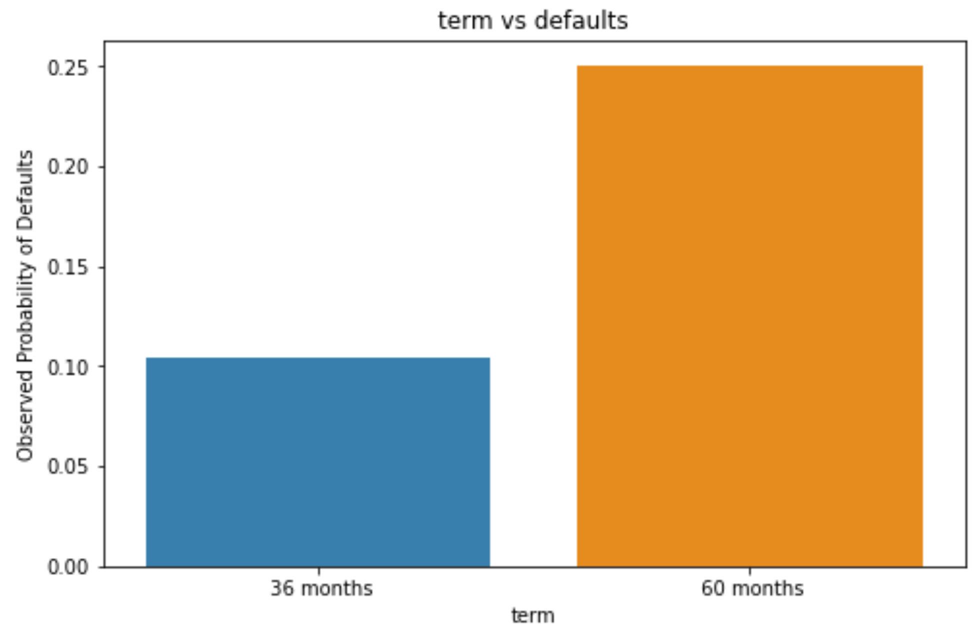
- We started by Understanding the problem domain and the variables in the data. We moved onto to cleaning up the data, removing outliers, if any and formatting them.
- For the data analysis itself, we began with univariate analysis and segmented univariate analysis to identify the driver variables. We then extended this with Bivariate analysis.

Note:

- If the presentation is viewed via the GitHub browser, please click on "More Pages" at the end of page 5, to view all the pages & conclusions

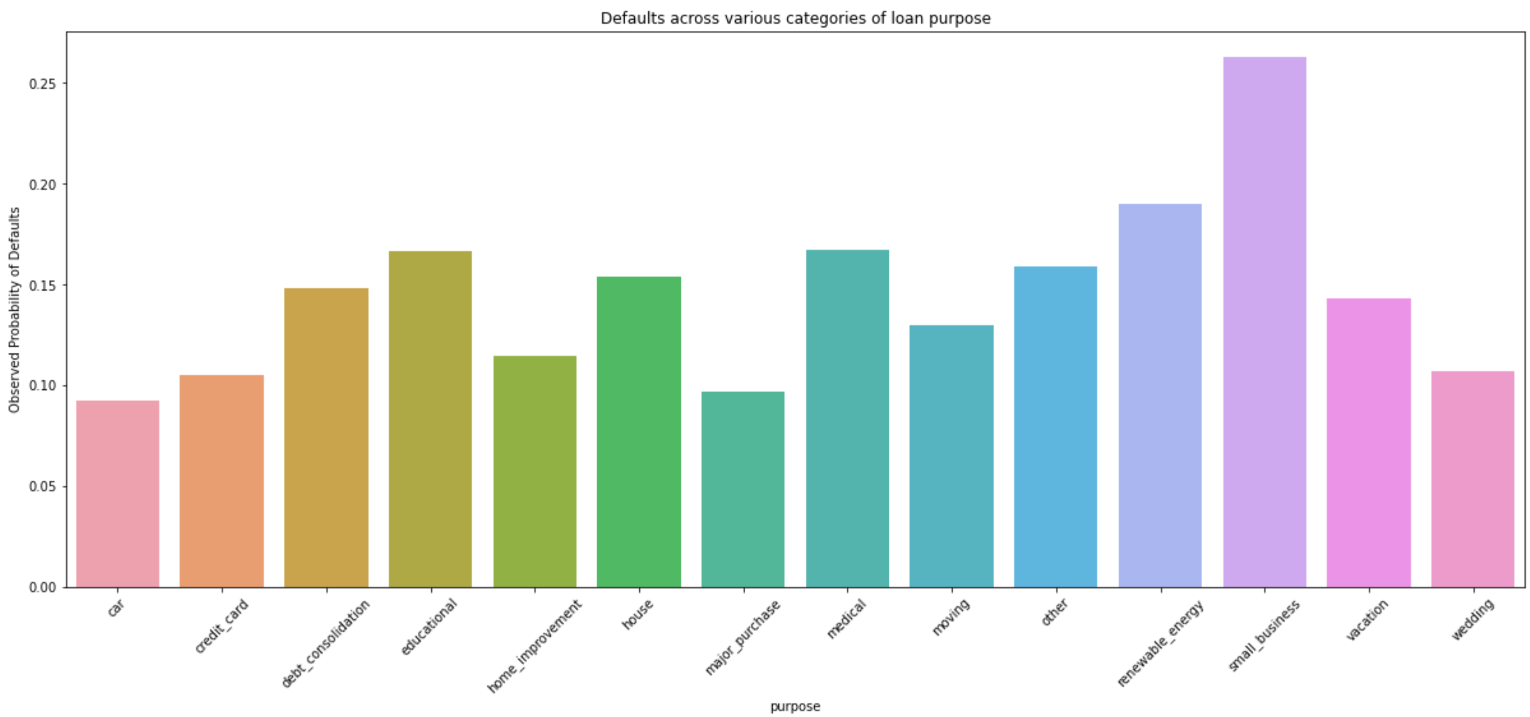
Conclusion 1

- “**term**” is a strong driver variable of default
- **At least 1 out of 4 loans** (25%) in “60 months” end up defaulting
- This helps the company to manage their portfolio by **not allocating a high number of loans** in the high risk **“60 month”** category



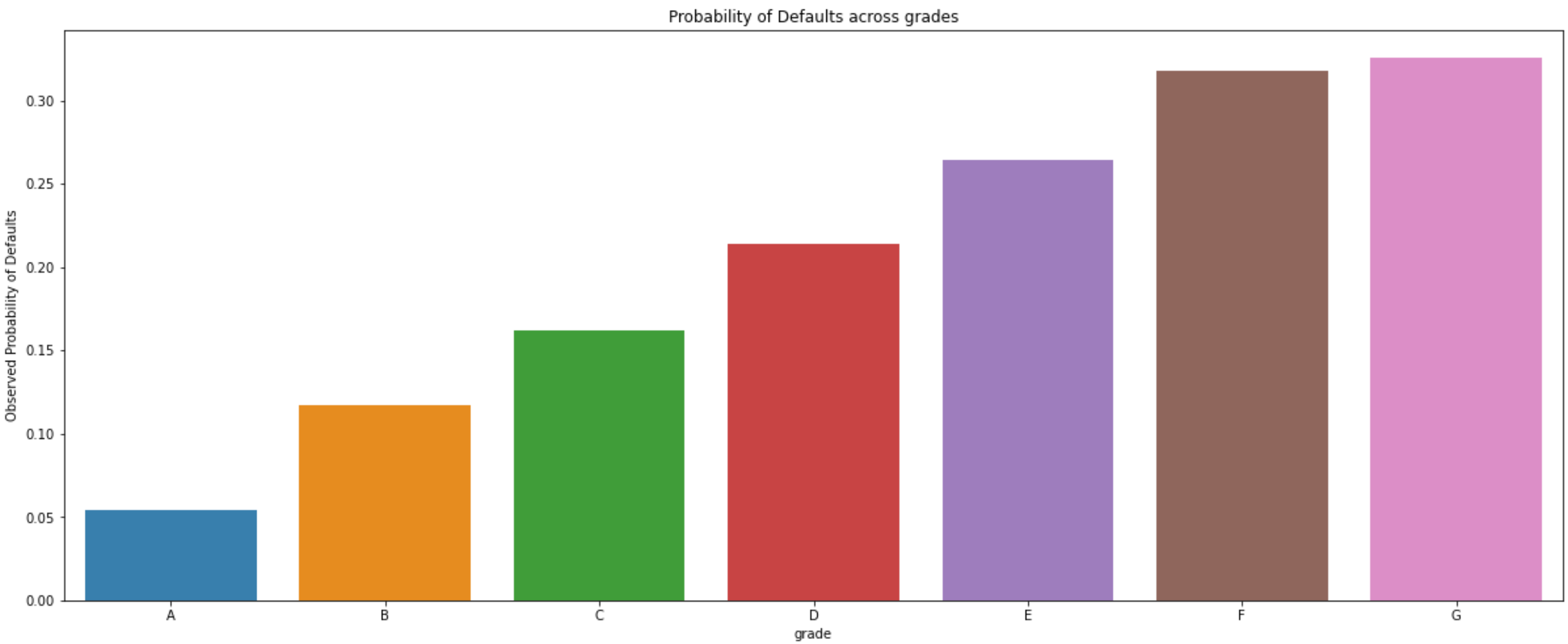
Conclusion 2

- "**purpose**" is a strong driver variable of default
- Ex: at least **1 out of 4 loans** (>25%) in "**small business**" end up defaulting
- This helps the company to manage their portfolio by **not allocating a high number of loans in the high risk "purpose" categories** like **"small business", "renewable energy".**



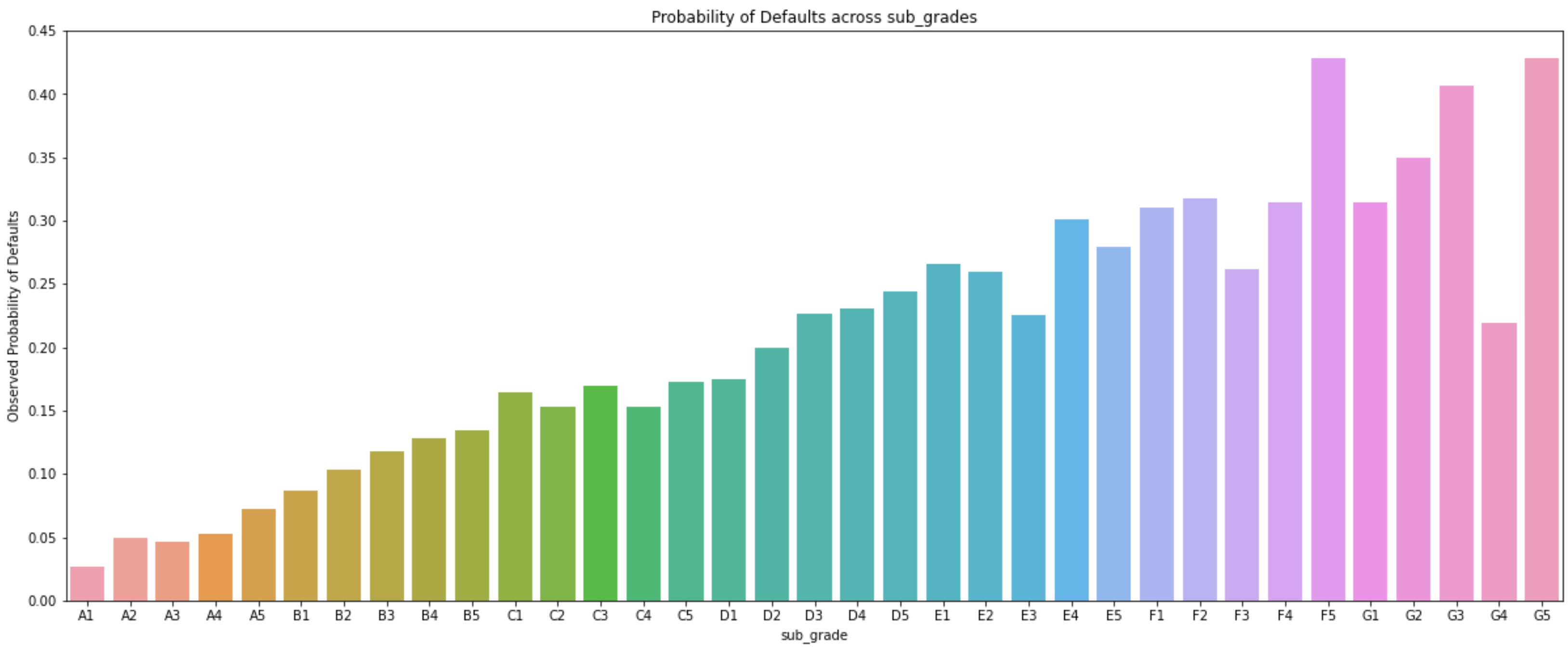
Conclusion 3

- "**grade**" is a strong driver variable of default
- Ex: **more than 30% of loans in grade G & F** end up defaulting
- This helps the company to manage their portfolio by **not allocating a high number of loans in the high risk "grade" categories** like **"G", "F"**



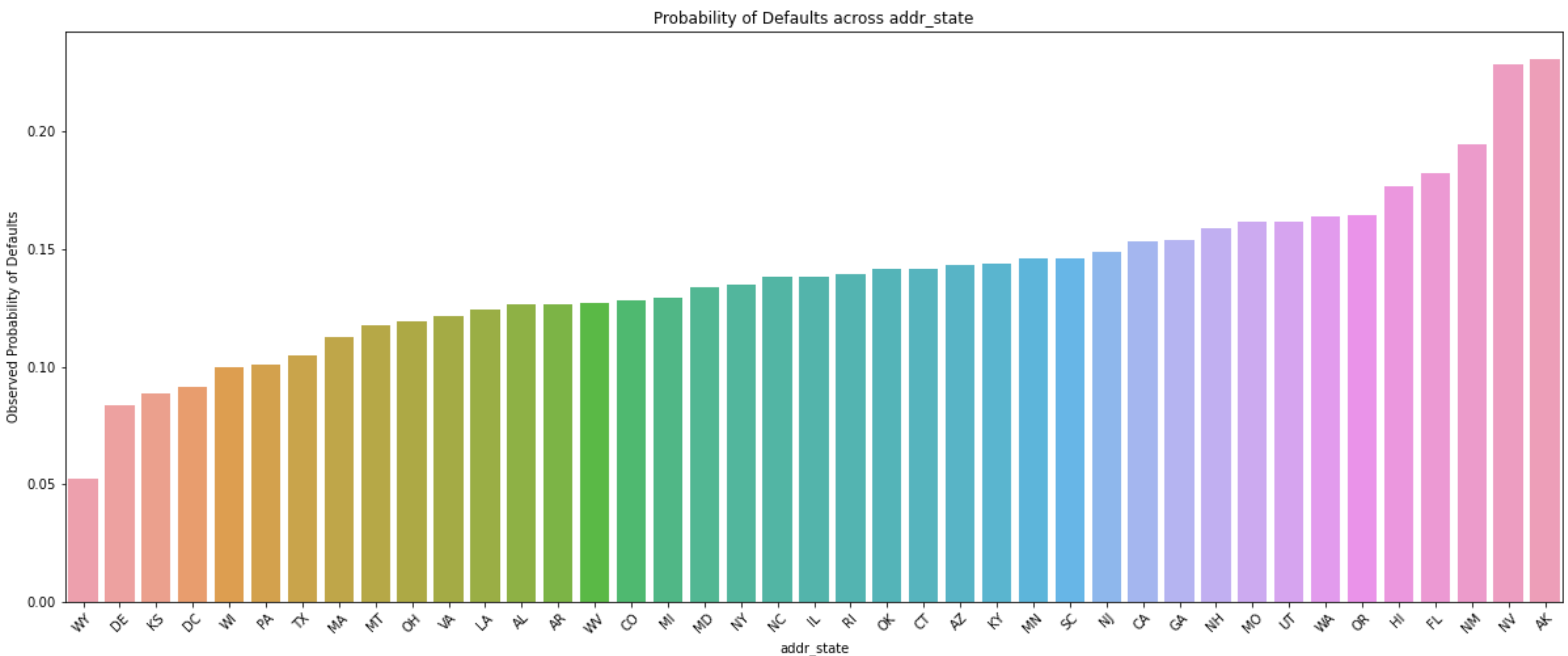
Conclusion 4

- "**sub_grade**" is a strong variable of default
- Ex: **more than 40% of loans** in "F5" & "G5" end up **defaulting**
- This helps the company to manage their portfolio by **not allocating a high number of loans in such high risk "sub_grade" categories**



Conclusion 5

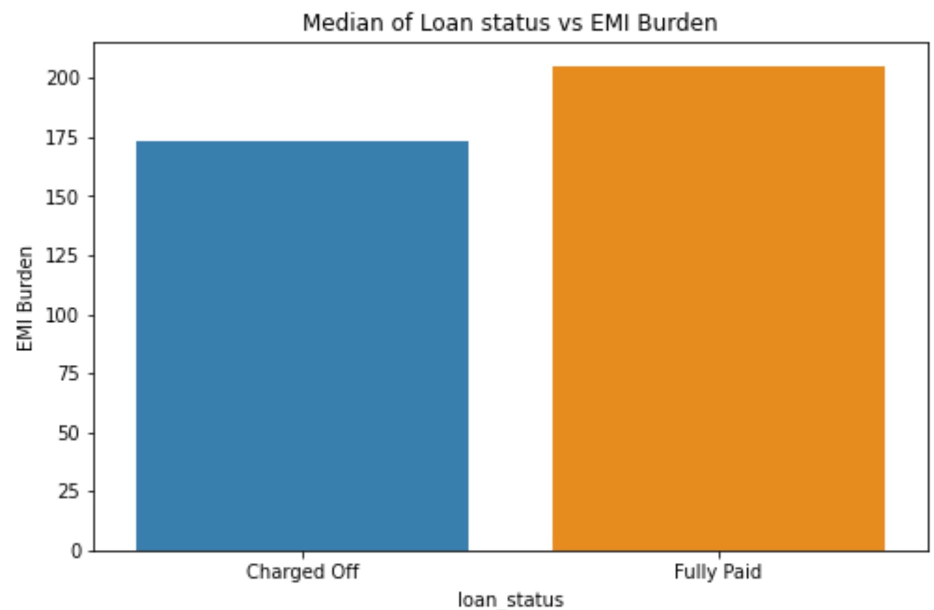
- “**addr_state**” is a strong variable of default
- Ex: **more than 20% of loans** in “NV” & “AK” end up **defaulting**
- This helps the company to manage their portfolio by **not allocating a high number of loans in such high risk “addr_state” categories**



Conclusion 6

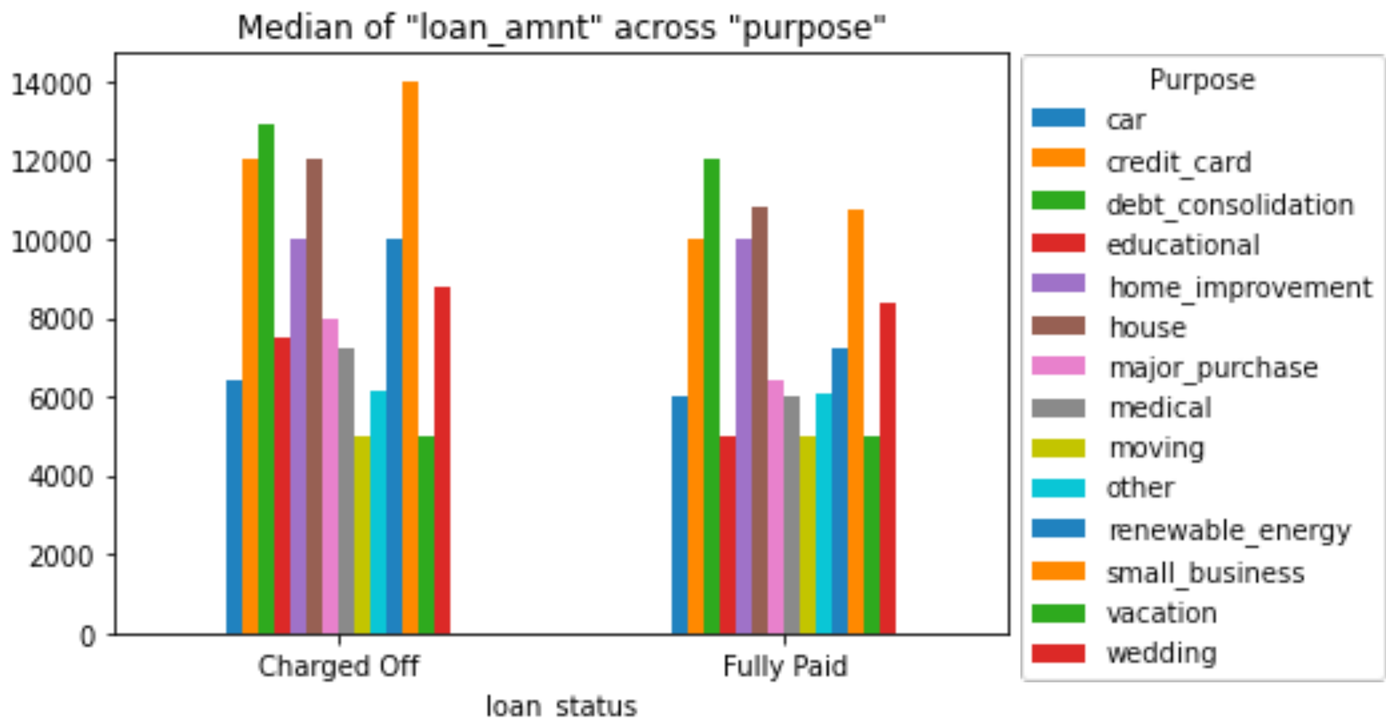
- **Business Driven Metric – EMI_BURDEN**
- Defined as the ratio of “annual_inc” to “installment”
- The medians of the calculated emi_burden are ”**173**” & “**205**” for “Charged Off & “Fully Paid” respectively.
- Fully paid loans **have much higher emi_burden** values than that of defaults
- This derived metric helps in predicting defaults at the time of approval

- **Note**: The metric must be used with caution. An increase/decrease in “annual_inc” during the term can greatly affect its importance



Conclusion 7

- Defaulters have significantly higher “**loan_amnt**” for “purpose” – “small_business”, “credit_card”, “renewable_energy” categories
- This helps the company to manage their portfolio by **not approving loans of higher “loan_amnt” among such categories**



Conclusion 8

- Among loans in grade “G”, “defaulters” have **significantly lesser “annual_inc”** compared to the ones who have “Fully Paid”
- This potentially means that the **company must consider higher values of “annual_inc”** for approval of loans among applicants of grade “G”