ANALYSIS ON

LOAN APPLICANT

Data Visualisation Project Documentation

ABSTRACT:

Applying loan for immediate/essential financial needs is almost done by everyone. The only factor that varies is the reason and the amount required.

There are many factors that are involved in determining if the applicant is going to repay the loan or not.

Through this analysis process, I have tried to find if there are any correlations which has impact on the loan defaulters.

NAME : AKSHAYA KUMAR CHANDRASEKARAN

ID : 31021301

Lab No : 08

Tutor Name : Fahimeh Sadat Saleh

Intended Audience:

People with little knowledge on banking process , data analyst/scientist aspirants et.al.,

Contents

FACTORS INFLUENCING LOAN REPAYMENT:

1.INTRODUCTION:

Almost all of us apply for loan for some of the essential needs and requirements in an intention to repay to the loan within the stipulated date and time. However, few of them are unfortunately not able to repay the loan as planned and end up being a defaulter.

So what loan providers generally do is collect required basic information about the person to assess the application and then decide to proceed with the application or not.

Primary focus of this analysis is on some important factors like Annual Income, Employment length, House ownership, type of application, grade of the applicant, loan amount requested and other few factors.

2.DATA SOURCE:

The dataset used for the analysis is taken from the site Kaggle. One large data set is used for analysis.

1. **Dataset** → Tabular data consisting of 855696 records and 74 variables. (All kinds of variables such as **location, continuous, categorical** and **simple text with punctuations and numbers** are present in the dataset)
   Source Link : https://www.kaggle.com/sonujha090/xyzcorp-lendingdata

Dataset Description: Attributes mentioned below are **used in analysis** from the dataset. Other columns were removed as part of the data wrangling process as their contribution in the analysis were found insignificant.

Attribute Name	Description
annual_inc	The self-reported annual income provided by the borrower during registration.
Application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
loan_amnt	The listed amount of the loan applied for by the borrower.
Emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
Grade	XYZ assigned loan grade
Purpose	A category provided by the borrower for the loan request.
Title	The loan title provided by the borrower.
Emp_title	The job title supplied by the Borrower when applying for the loan. (Text Data)
Issue_d	Year the loan is issued.
addr_state	The state provided by the borrower in the loan application (Geo Location)
default_ind	If he is a defaulter or not. 0 is Non defaulter and 1 is defaulter.

Table 2.1 Major Attributes used for analysis

3.DESIGN:

The idea when creating the web page was to show the impact of attributes annual income, employment length and home ownership on defaulting a loan or not. Upon development of the web user interface, various new ideas were implemented for new findings.

3.1 Sheet One / IDEA Sheet:

Below are few of the ideas that has been considered after filtering out the repetitions in notion to answer the questions asked.

1. Choropleth Map : To find the total number of applicants from each state
2. Box Plot : To have a quick idea about the annual income attribute, its stats and distribution.
3. Sunburst graph : To find out in each year and in which proportion applicants have defaulted the loan.
4. Pie Chart : To show the percentage of purpose of loan
5. Sankey Diagram : To find the relationship between the top employment title, grade and default status.
6. Word Cloud : To signify the number of applicant's employee title.

3.2 Sheet Two:

		POSITIVE SIDE:	DRAW BACKS
LAYOUT	**Box Plot**	Distribution and stats of annual income can be easily understood.	If there are more number of outliers, then box plot will not be clearly visible.
	Bar Chart	Direct comparisons can be done easily between the count for the employee title.	Since the unique number of applicant's employee title is more than 1000, it is difficult to understand if plotted.
	Word Cloud	The most repeating category can be easily found out using word cloud.	The level of detail will be less if the count is more.
FOCUS/ZOOM	**Bar chart**	Count of the records can be easily obtained	Difficult to find the required count if the bar are too cluttered.
	Word Cloud	If we hover on the word, count of the word repeated will be known.	If the words are too small, then identifying the word is difficult.
OPERATIONS	**Radio Button**	Can analyse in detailed manner by choosing the required option. Choosing the option will change both the graphs dynamically.	The provided choices are limited.

Table 3.1 Sheet two contents of Five sheet design

3.3 Sheet Three:

		POSITIVE SIDE:	DRAW BACKS
LAYOUT	**Choropleth Map**	Can get the depth knowledge on the behaviour of the people state-wise.	If people are from different countries, then it is difficult to show in detail.
	Pie Chart	Most repeated category of purpose can be easily interpreted with the help of pie chart.	If there are more number of unique categories, then Pie chart will not be the best option.
	Tree Map	Can show categories and sub-categories as well.	It is same as that of pie chart and finding the count can be difficult as the size reduces.
FOCUS/ZOOM	**Pie Chart**	Hovering on a category, count will be displayed.	If there are too many unique values, then choosing the wanted category will be difficult.
OPERATIONS	**Drop Down**	Detailed information can be gathered.	If there are more number of unique values, multiple drop down option can some times be confusing.
	Radio Button	Can analyse in detailed manner by choosing the required option.	The provided choices are limited.

Table 3.2 Sheet Three contents of Five sheet design

3.4 Sheet Four:

		POSITIVE SIDE:	DRAW BACKS
LAYOUT	**Sankey Diagram**	Flow of the category can be easily identified and understood	Only top level understanding can be obtained if there are more number of unique categories.
	Sunburst Chart	Hierarchical visualization can be easily understood from sun burst and it also consumes less space.	The radial structure makes it more complicated to understand and difficult at times to read by human eye.
FOCUS/ZOOM	**Sankey Diagram**	Hovering on the line of flow will provide detailed information about count and categories.	If there are more levels, it can be difficult to understand the flow.
OPERATIONS	**Sankey Diagram**	Clicking on any one of the categories will generate another graph.	Difficult to implement.

Table 3.3 Sheet four contents of Five sheet design

3.5 Sheet Five/Design Sheet:

		POSITIVE SIDE:	DRAW BACKS
LAYOUT	**Sankey Diagram**	Flow of the category can be easily identified and understood	Only top level understanding can be obtained if there are more number of unique categories.
	Sunburst Chart	Hierarchical visualization can be easily understood from sun burst and it also consumes less space.	The radial structure makes it more complicated to understand and difficult at times to read by human eye.
	Choropleth Map	Can get the depth knowledge on the behaviour of the people state-wise.	If people are from different countries, then it is difficult to show in detail.
	Box Plot	Distribution and stats of annual income can be easily understood.	If there are more number of outliers, then box plot will not be clearly visible.
	Bar Chart	Direct comparisons can be done easily between the count for the employee title.	Since the unique number of applicant's employee title is more than 1000, it is difficult to understand if plotted.
	Word Cloud	The most repeating category can be easily found out using word cloud.	The level of detail will be less if the count is more.
FOCUS/ZOOM	**Sankey Diagram**	Hovering on the line of flow will provide detailed information about count and categories.	If there are more levels, it can be difficult to understand the flow.
	Bar chart	Count of the records can be easily obtained	Difficult to find the required count if the bar are too cluttered.
	Word Cloud	If we hover on the word, count of the word repeated will be known.	If the words are too small, then identifying the word is difficult.

OPERATIONS	**Sankey Diagram**	Clicking on any one of the categories will generate another graph.	Difficult to implement.
	Radio Button	Can analyse in detailed manner by choosing the required option.	The provided choices are limited.
	Drop Down	Detailed information can be gathered.	If there are more number of unique values, multiple drop down option can some times be confusing.

Table 3.4 Sheet five contents of Five sheet design

3.6 Changes from original design:

Five sheet design was a kick start for the data visualisation project. Upon developing the web user interface, there were few realisations and upgradations from the original ideas. Those changes were implemented in the UI for better understanding and to provide more information.

Below are the deviations from the initial design proposed and the actual design explained in the table 3.5 .

Category	Initial Design	Current Design	Reason
Layout	Planned to put the entire figures in one page	Introduced tabs	for aesthetic looks and good presentation and better understanding on the information provided.
Sheet 2	Three diagrams 1. Box plot 2. Bar chart 3. Word cloud	Bar chart and Box plot in one page and word cloud in a separate page.	Word cloud occupied more space. Visualisation was not clear. For better understanding on the information.
Choropleth Map	Initially planned to show choropleth in sheet 3	Displayed in the introduction page.	To have a quick background and understanding about the data set and spatial information, showed the map in the introduction page.
Pop-up window	Did not intend to implement.	Implemented in few pages	Quick gist about the image shown. Consumes less space in the UI and more user interaction also.
One additional page	Did not intend to plot.	Additional attributes were taken, and charts were drawn and implemented in a new page.	For better understanding and in-depth knowledge about those attributes and their impact for analysis.

Table 3.5 Changes in the design from the actual design

4.IMPLEMENATION:

The data set from the source contained more than 850,000 records and 74 columns. There were data anomalies. So data set was processed as part of data wrangling and data cleaning.

Data set used for the visualisation project consists of around 18 columns and all the records. Data anomalies were removed, and the data set is cleaned for the implementation of webpage using R-Shiny library.

	NAME	REASON
SOFTWARE	R and R-Studio	Choice of R was made over D3 since there are more number of packages for creating attractive and user friendly web interface.
LIBRARIES	Shiny	Web page implementation
	NetworkD3	For creating Sankey network
	Wordcloud2	Generate Word cloud
	Ggplot2	Create graphs
	Choropleth	Create choropleth maps
	SunburstR	To create sun burst graph
	Plyr , dplyr , margrittr	Data wrangling and cleaning
	RColorBrewer	For different color palettes

Table 4.1 Table showing software and packages used to build the UI

5.USER GUIDE:

5.1 Page One:

The web page starts with a welcome page which contains a quick introduction about the problem and it's summary and who the intended audience are as shown in the figure 5.1.

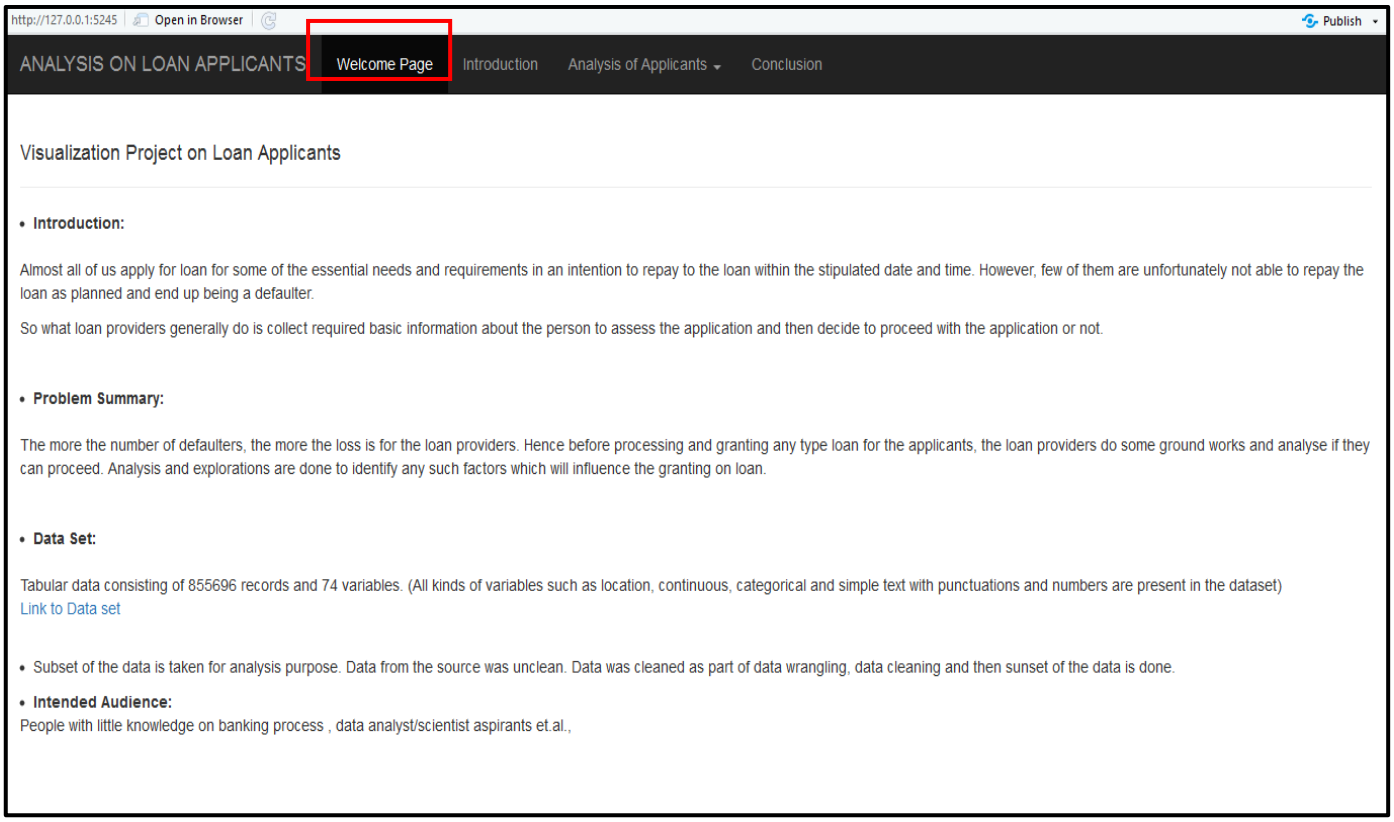


Figure 5.1 Image of the welcome page

5.2 Page Two:

Page Two is an introduction page representing the location of loan applicants. A radio button is provided for the user to toggle to know the count of defaulters and non-defaulters state wise. Legend is provided for the user to compare the gradient colour for comparison as shown in the figure 5.2.

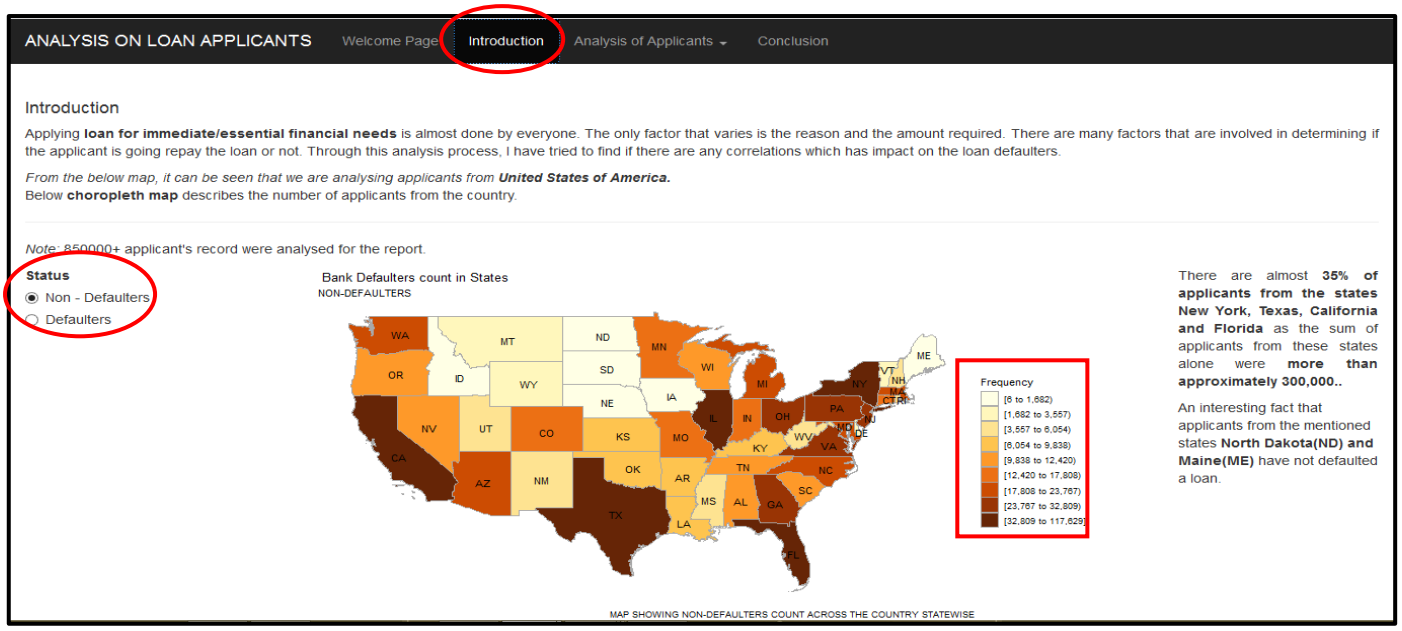


Figure 5.2 Image of the Introduction page

5.3 Page Three:

In page three, there are five tabs provided in drop down menu.

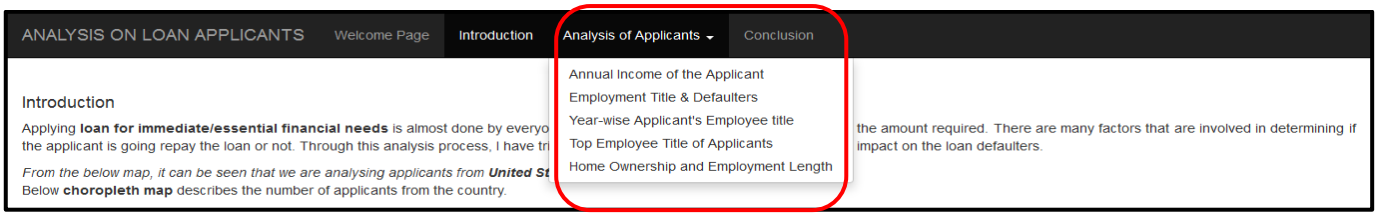


Figure 5.3 Image showing drop down menu provided in page three

5.3.1 Tab 1 (Annual Income of the Applicant):

In tab one, there are two main plots.

1. Box plot
2. Bar chart of top 15 Applicant's employee title.

User has been provided with the option of choosing the subset of data with the help of radio button provided as shown in the below figure 5.4. Choosing the radio button will change both the plots based on the input chosen.

Figure 5.4 Image of the Tab 1 in page 3-analysis on annual income and corresponding applicant's employee title

Also, there is an action button provided for the user as shown in the above figure to get a depth knowledge about the annual income. Once the user clicks on the dialogue box, based on the chosen box plot, the content will vary as shown in the below figures 5.5 and 5.6. Please click on dismiss upon completing the reading.

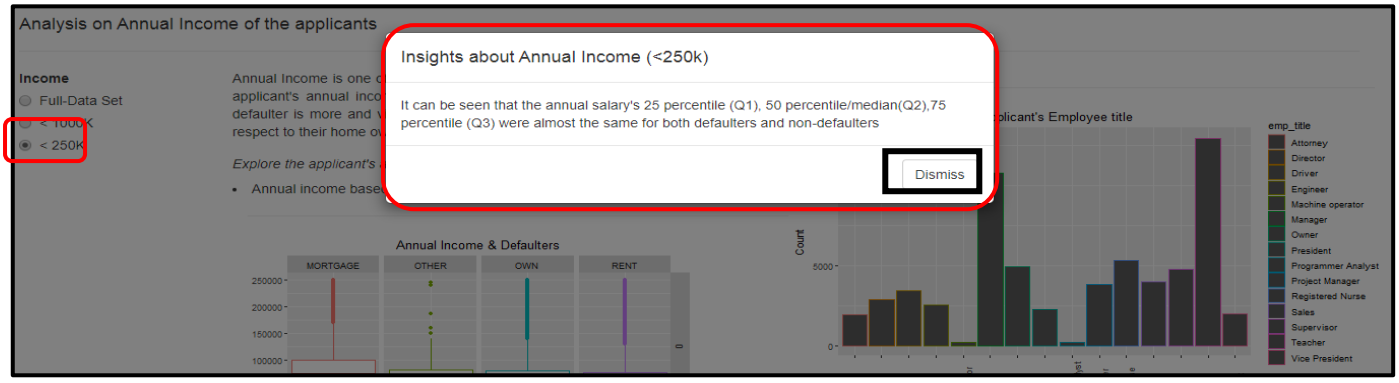


Figure 5.5 Image of the popup window message when clicked

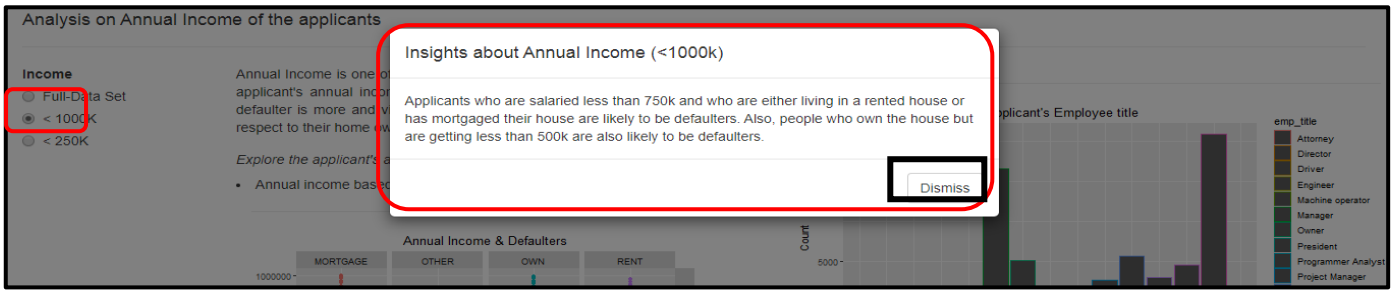


Figure 5.6 Image of the popup window message when clicked

5.3.2 Tab 2(Employment title & Defaulters):

In tab two, there are two main plots.

1. Donut chart
2. Sankey diagram for top 12 applicant's employee title.

User has been provided with the option of choosing the subset of data with the help of radio button provided as shown in the below figure 5.7. Choosing the radio button will change both the plots based on the input chosen. If you hover over in the greyed are of Sankey diagram, the source, destination and the count will be displayed.

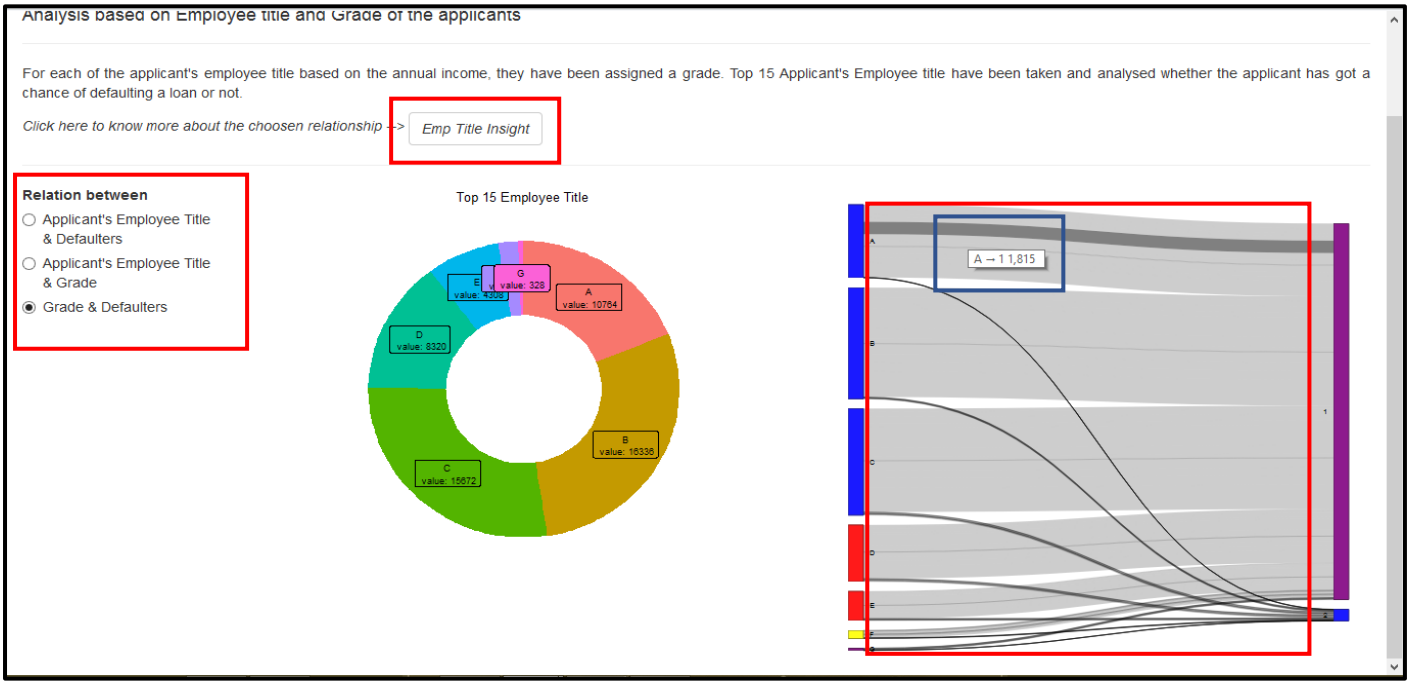


Figure 5.7 Image of the Tab 2 in page 3-analysis on applicant's employee title and defaulters

Also, there is an action button provided for the user as shown in the above figure to get a depth knowledge about the annual income. Once the user clicks on the dialogue box, based on the chosen box plot, the content will vary as shown in the below figures 5.8 and 5.9. Please click on dismiss upon completing the reading.

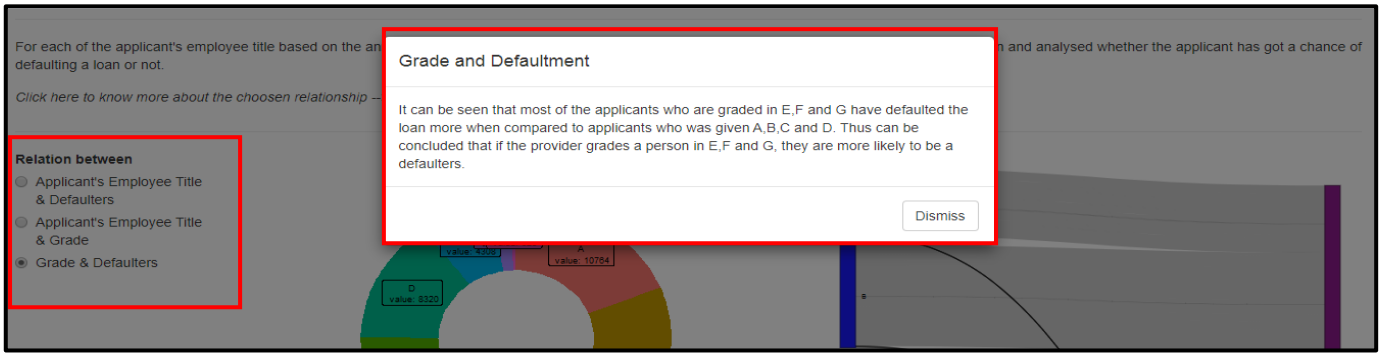


Figure 5.8 Image of the popup window message when clicked

When the user chooses a different option, both the Sankey diagram and the donut chart changing as shown in the below diagram. A pop up window when clicked, will provide a detailed information about the Sankey diagram.

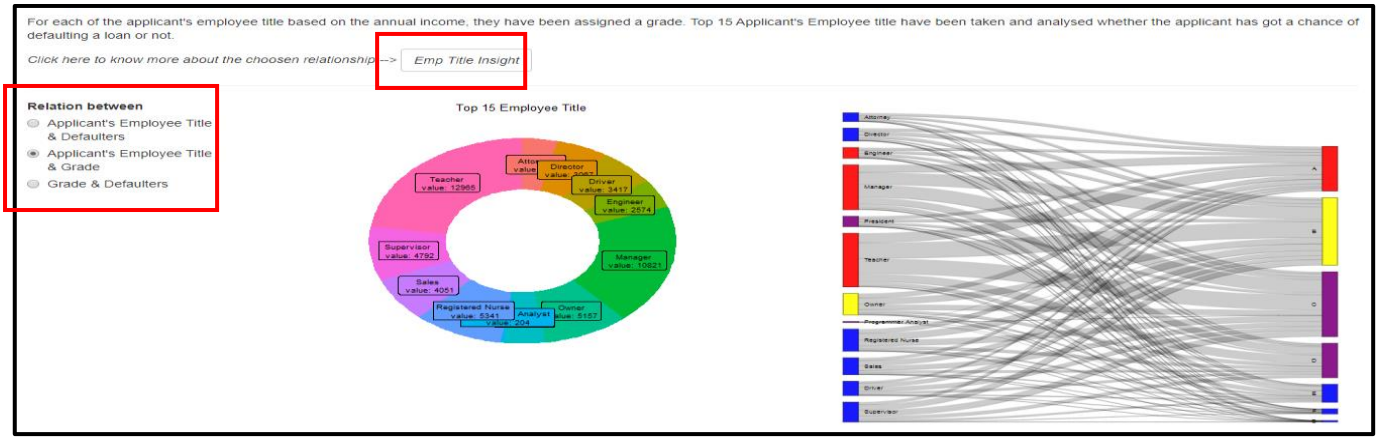


Figure 5.9 Image of dynamic changes in donut and Sankey diagram based on the input given

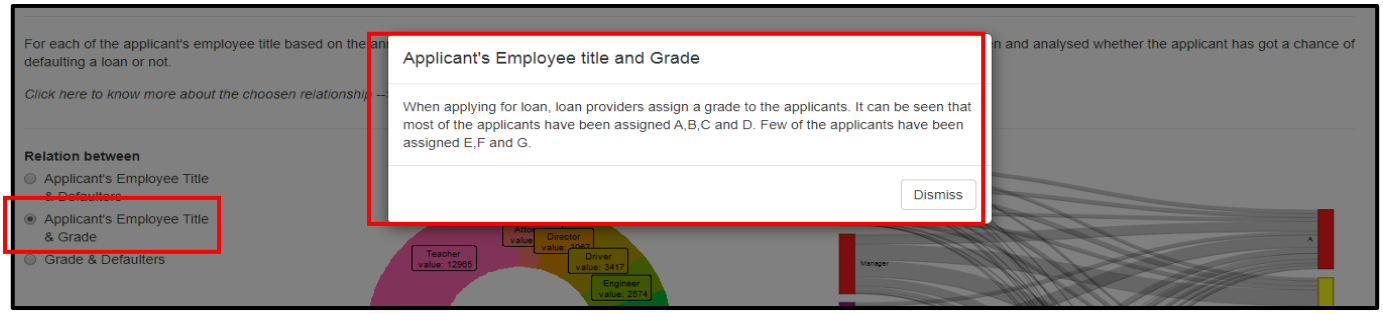


Figure 5.10 Image of dynamic changes in pop up message when clicked.

5.3.3 Tab 3(Year-wise Purpose of Loan):

In tab two, there are two main plots.

1. Pie Chart
2. Sun-burst chart

User has been provided with the option of choosing the subset of data with the help of radio button provided as shown in the below figure. Choosing the radio button will change the pie chart alone.

An action button is provided to describe about the pie chart. The display message in the action button will change dynamically based on the option chosen by the user in the radio button provided.

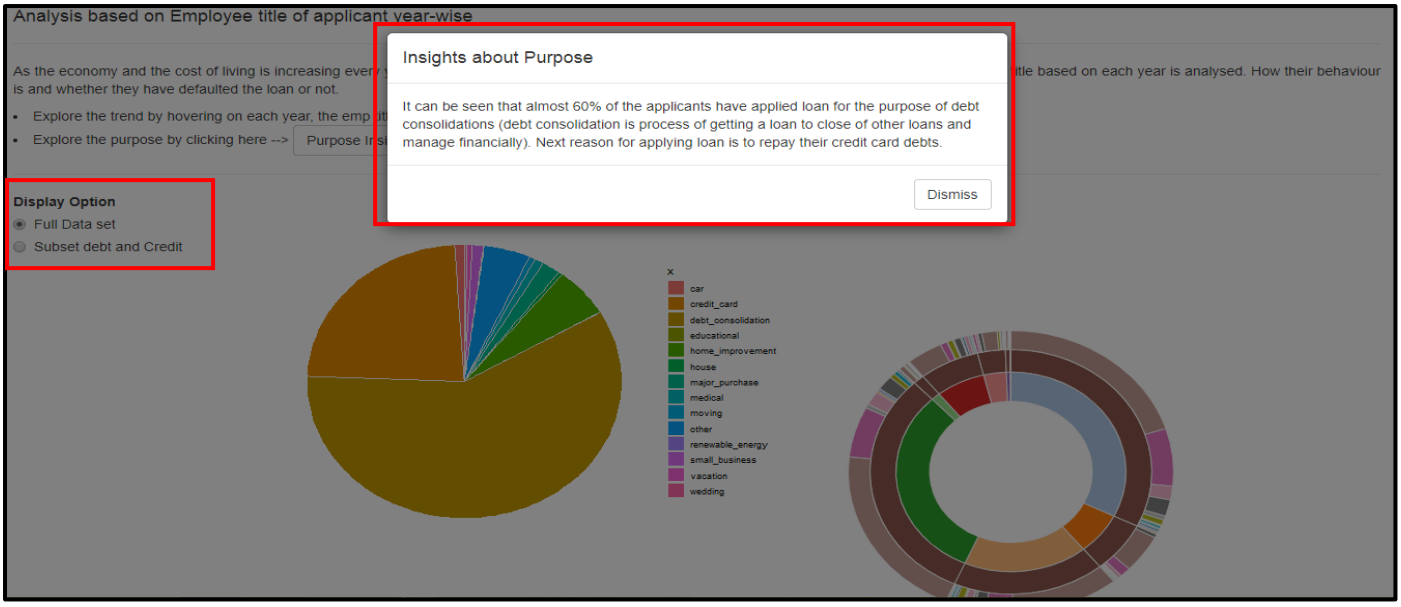


Figure 5.11 Image of the Tab 3 in page 3-analysis based on each year and the purpose of loan.

In sunburst chart, if you hover over on the radial vector, the percentage of that category in the data set will be displayed as shown in the below figures 5.12 and 5.13.

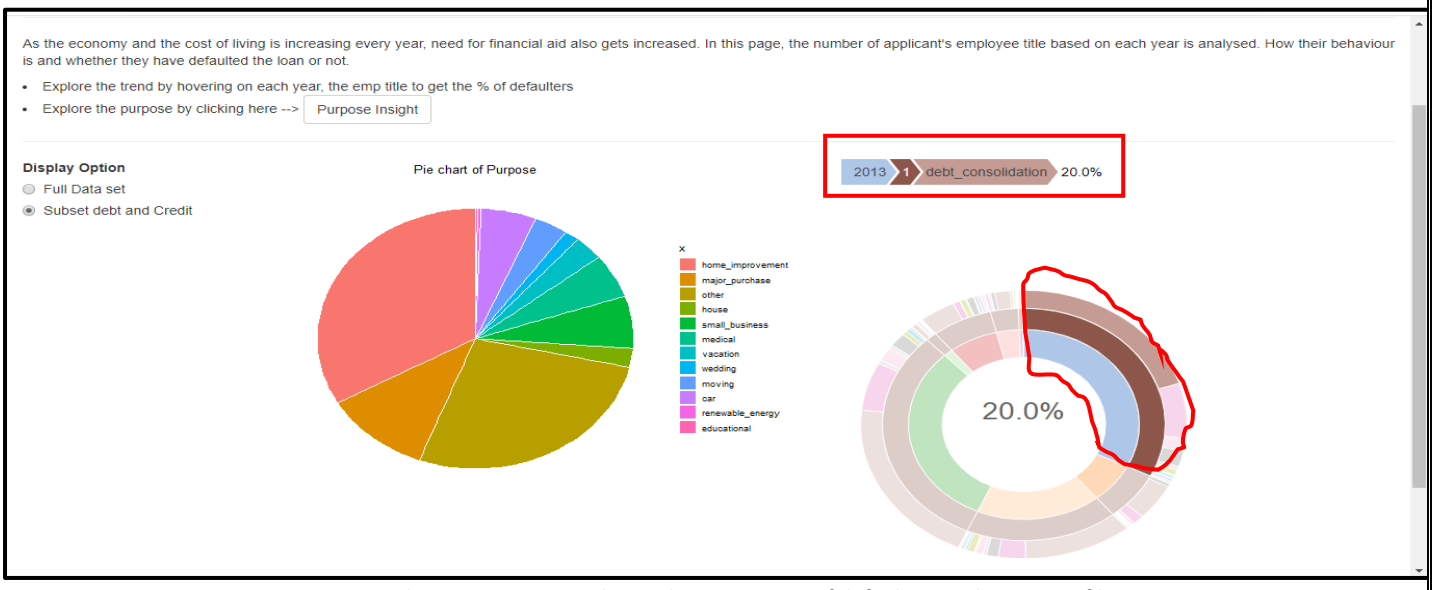


Figure 5.12 Sunburst interaction to know the percentage of defaulters and purpose of loan.

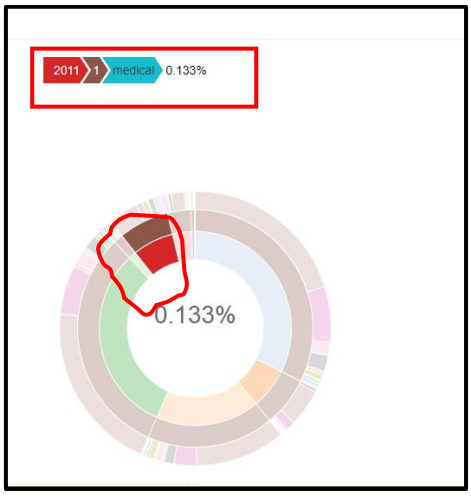


Figure 5.13 Sunburst interaction to know the percentage of defaulters and purpose of loan.

5.3.4 Tab 4(Top Employee title of Applicants):

In tab four, the main plot is word cloud.

In this tab, user has been provided with two different options.

1. Drop down → To choose the category for which the word cloud should be generated
2. Slider Option → To set the minimum frequency the word should be repeated in the data set for display as shown in the below figure 5.14.

Also, when the user hover over the word in word cloud, the count of word repeated in the word cloud will be displayed as shown in the below figure 5.15.

Based on the category chosen, the insight description also changes dynamically.

Figure 5.14 Image of the Tab 4 in page 3-analysis based Applicant's Employee title and purpose of loan.

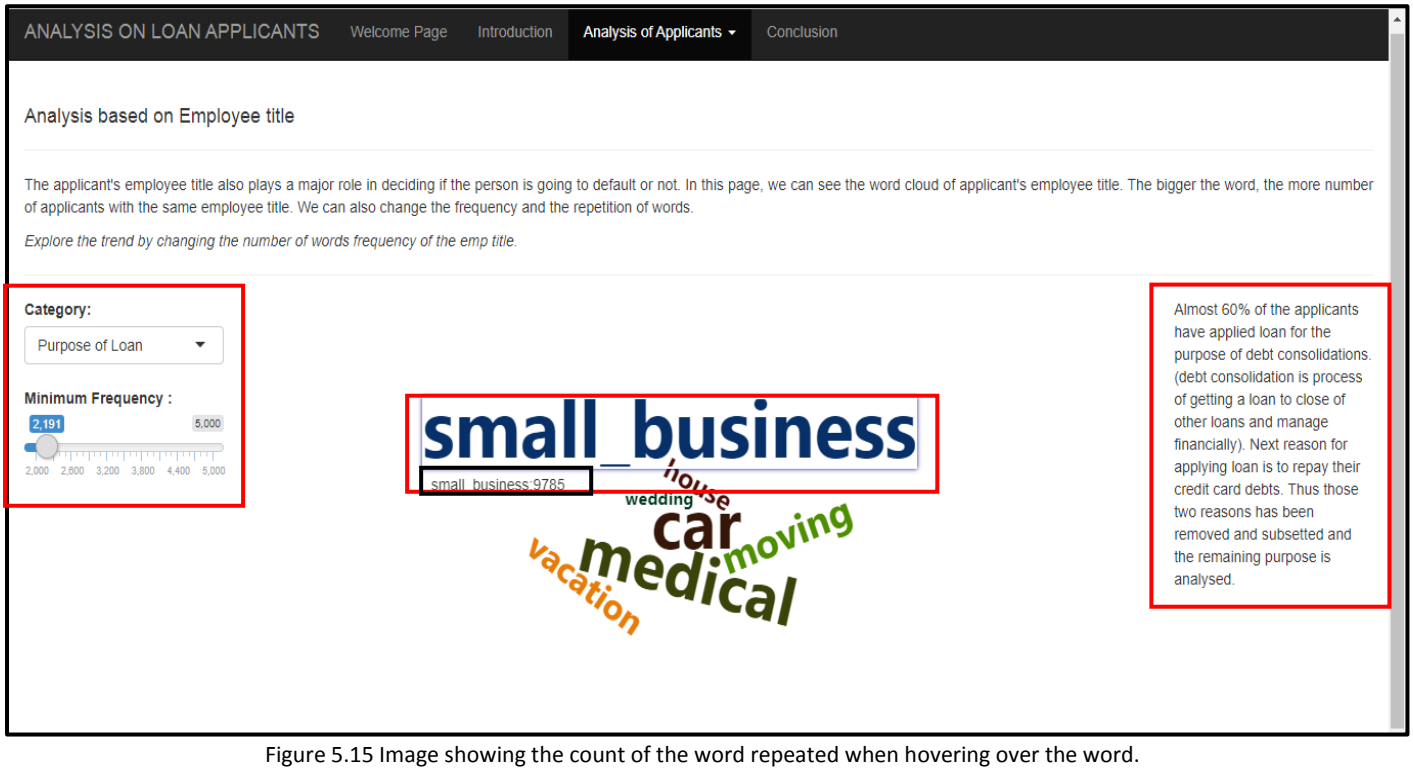


Figure 5.15 Image showing the count of the word repeated when hovering over the word.

5.3.5 Tab 5(Home Ownership and employee length):

In tab five, there are two main plots.

1. Scatterplot
2. Bar-chart

User has been provided with the option of choosing the subset of data with the help of radio button provided as shown in the below figure. Choosing the radio button will change both the plots based on the option chosen.

An action button is provided to describe about the scatter plot. The display message in the action button will change dynamically based on the option chosen by the user in the radio button provided.

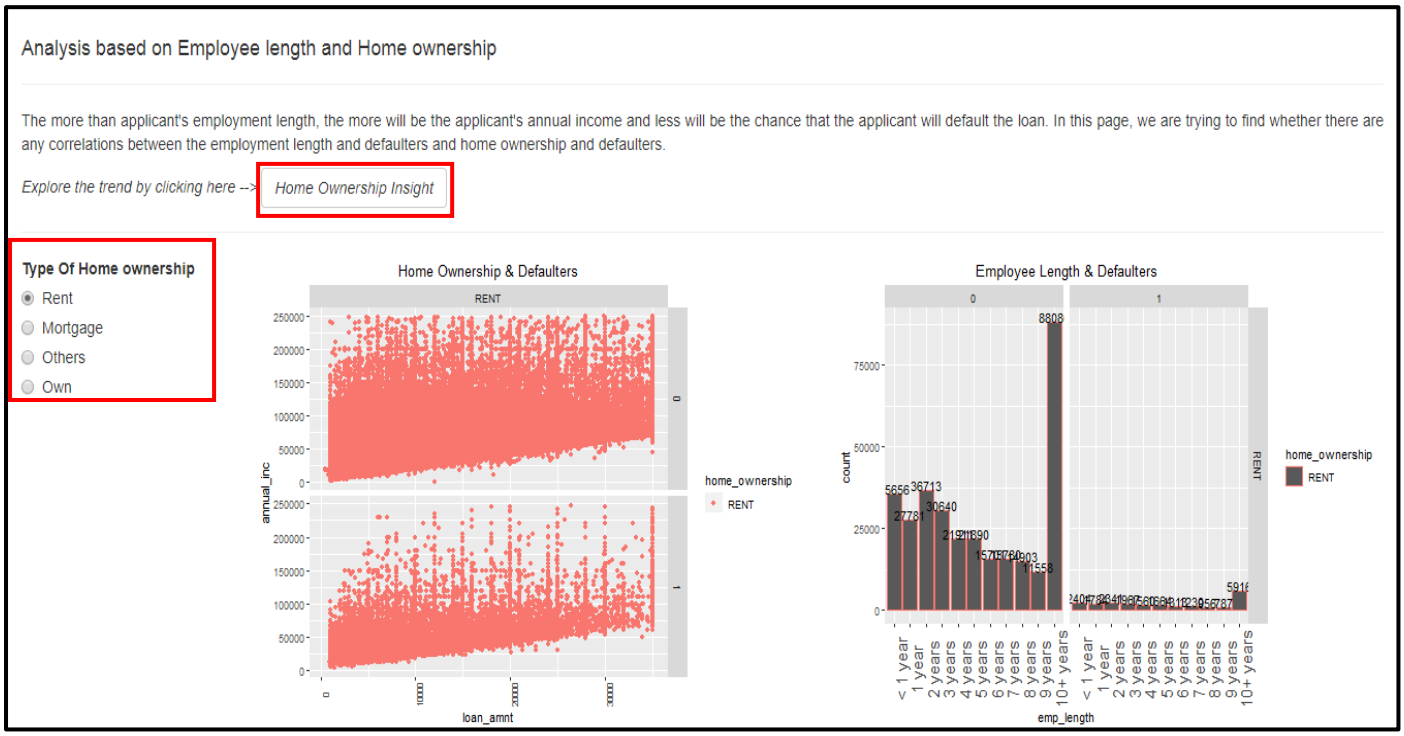


Figure 5.16 Image of the Tab 5 in page 3-Analysis based on the applicant's home ownership.

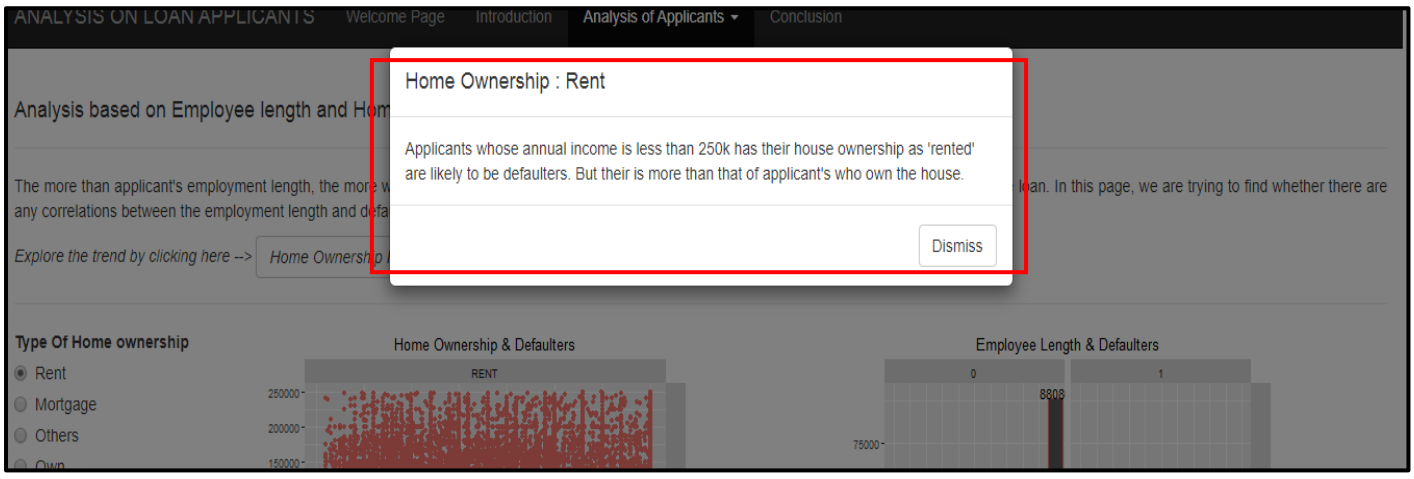


Figure 5.17 Image of dynamic changes in pop up message when clicked.

5.4 Conclusion:

The Web interface comes to an end with the conclusion tab. This tab consists about the summary analysis from the visualizations in the UI. Also the reference links that helped to develop the web user interactions has also been attached in this page.

User can click the hyper link and go to the page for quick visit as shown in the below figure.

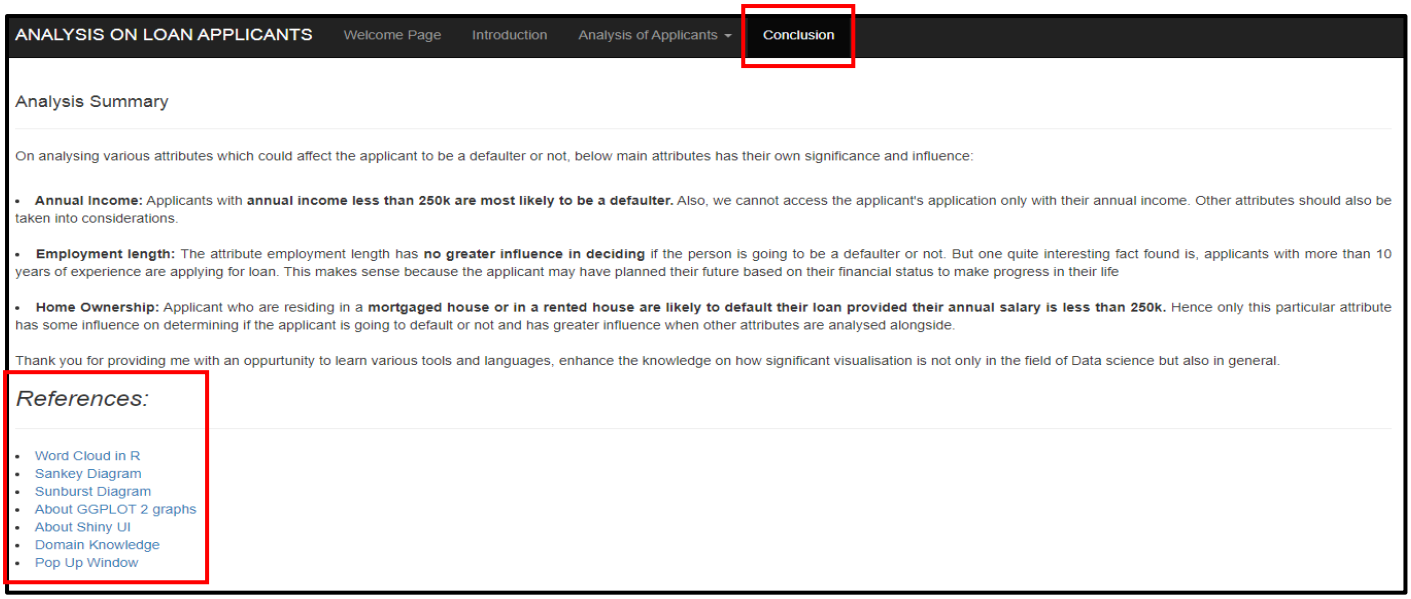


Figure 5.18 Image of Conclusion page

6.CONCLUSION:

As an outcome of this visualisation project, I found out the correlations between the various factors affecting the loan status. This project also helped me gain depth knowledge on data visualisation, different types of visualisation, data processing, data wrangling and data manipulations.

6.1 Thoughts:

Having had such a large data set, the journey was challenging yet fruitful. Had I chosen a smaller data set, execution time and the UI page load time would have reduced to a considerable amount leading to speedy navigation and retrieval of images.

7.BIBLIOGRAPHY:

1. STHDA - Text mining and word cloud fundamentals in R : 5 simple steps you should know - http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know → (Word Cloud in R)

2. GGPLOT2 – CHEAT SHEET - https://rstudio.com/wp-content/uploads/2016/11/ggplot2-cheatsheet-2.1.pdf → (About GGPLOT 2 graphs)

3. WIKIPEDIA - Default (finance) - https://en.wikipedia.org/wiki/Default_(finance) → (Domain knowledge)

4. PEXELS - Finance Images - https://www.pexels.com/search/finance/ → (Cover Photo)

5. DISPLAYR – Sankey Diagram - https://www.displayr.com/sankey-diagrams-r/ → (Sankey Diagram)

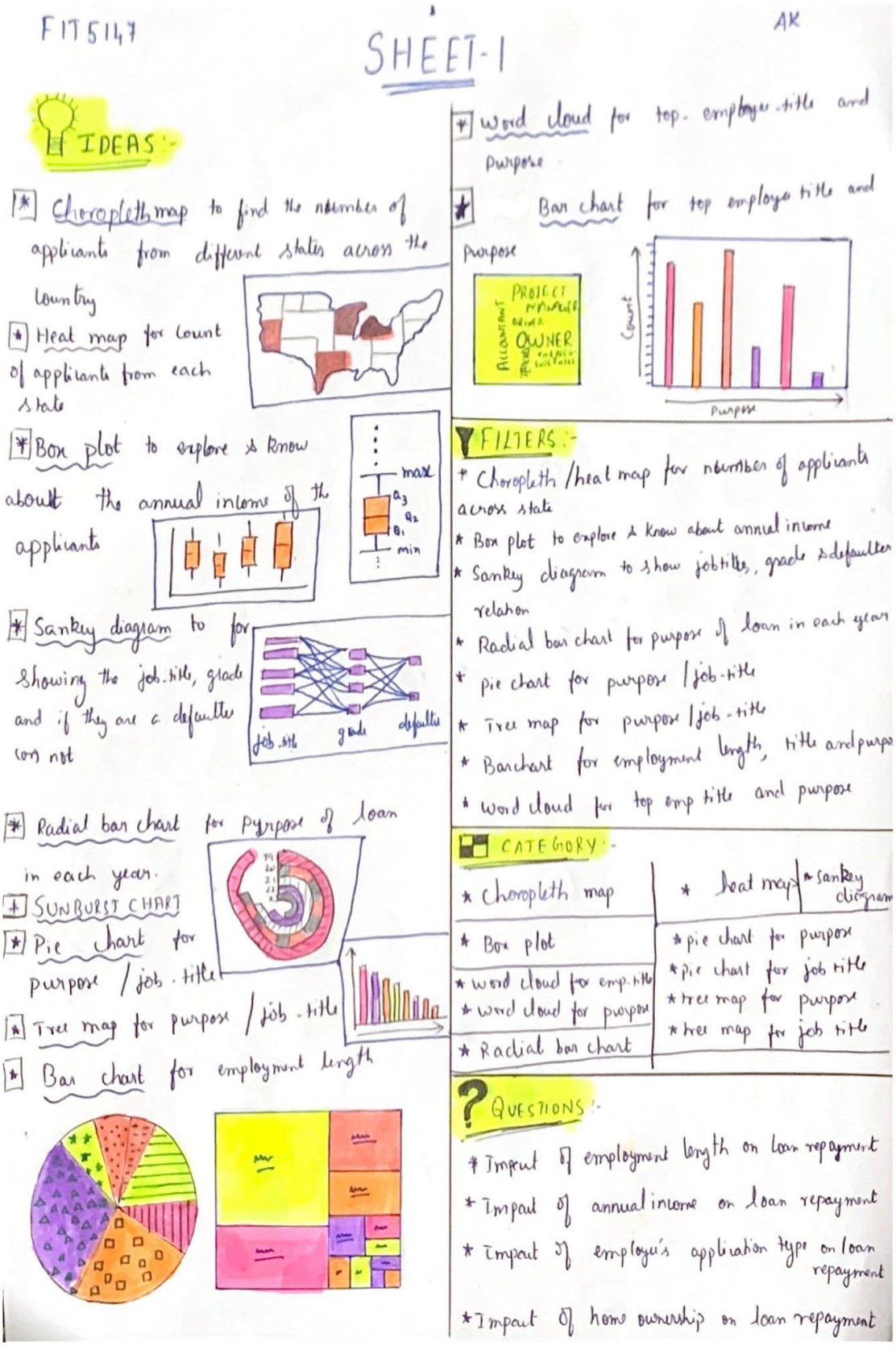
6. DATA-TO-VIZ - Sunburst Diagram- https://www.data-to-viz.com/graph/sunburst.html → (Sun Burst Graph)

7. SHINY– CHEAT SHEET - https://shiny.rstudio.com/images/shiny-cheatsheet.pdf → (About Shiny UI)

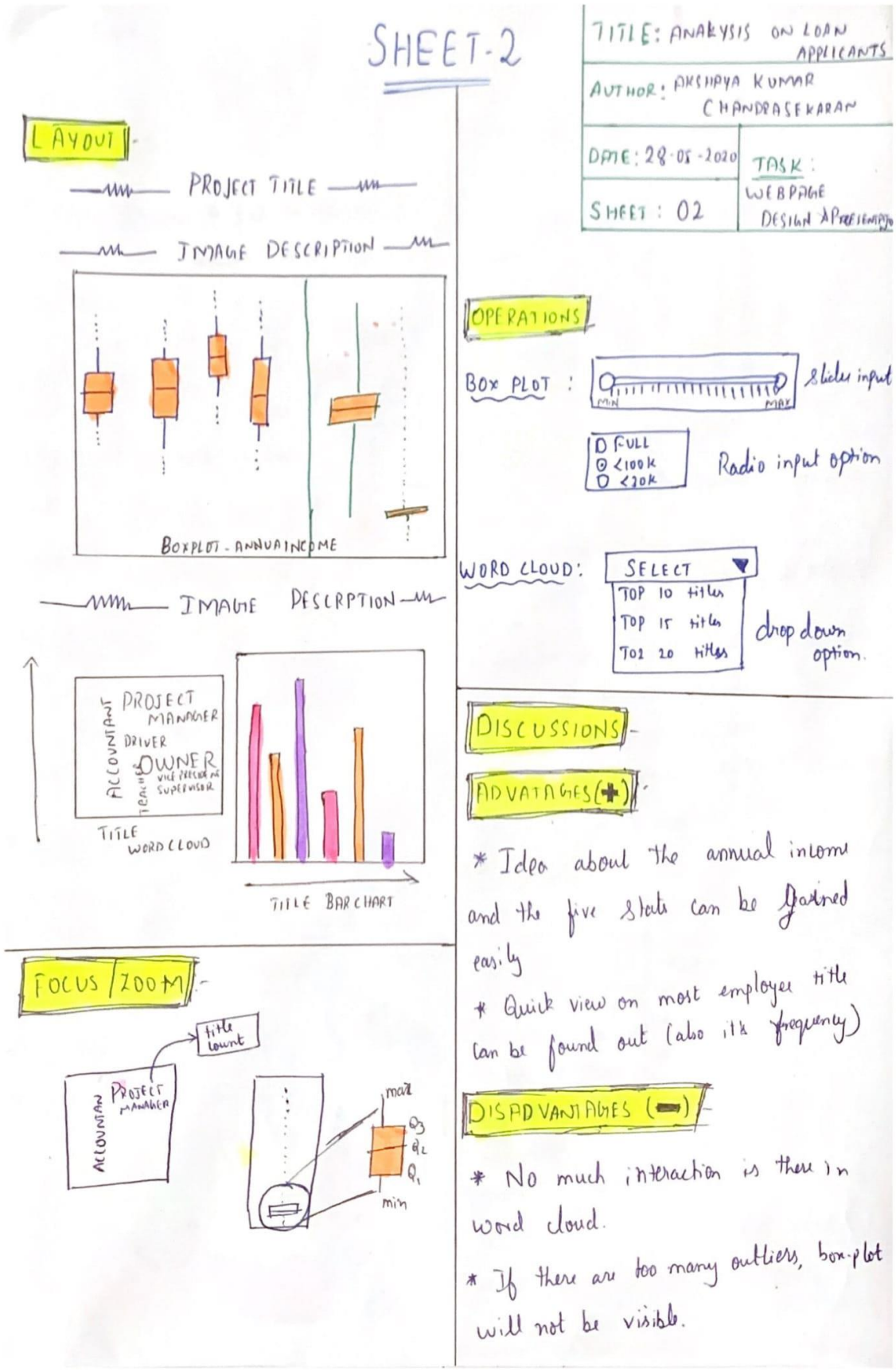
8. SHINY – Pop-Up Window - https://shiny.rstudio.com/reference/shiny/latest/modalDialog.html → (Pop Up Windows)

8.APPENDIX:

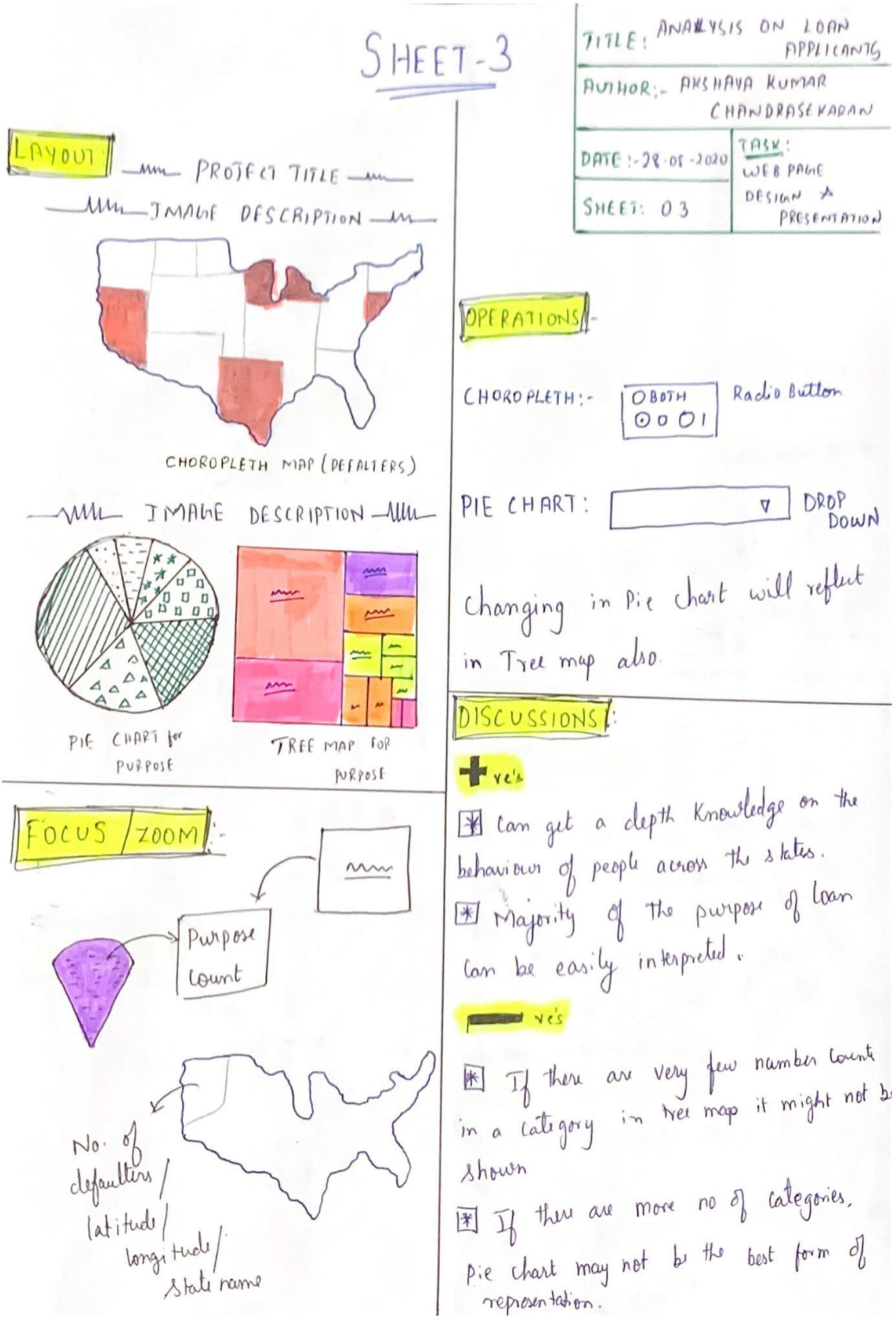
SHREET – ONE / IDEA SHEET



SHREET – TWO


SHEET-2
TITLE: ANALYSIS ON LOAN APPLICANTS
AUTHOR: AKSHAYA KUMAR CHANDRASEKARAN
DATE: 28-05-2020
TASK: WEBPAGE DESIGN PRESENTATION
SHEET: 02
LAYOUT
PROJECT TITLE
IMAGE DESCRIPTION
BOXPLOT - ANNUAINCOME
IMAGE DESCRIPTION
PROJECT MANAGER
DRIVER
OWNER
ACCOUNTANT
TITLE WORD CLOUD
TITLE BARCHART
OPERATIONS
BOX PLOT :
MIN
MAX
Slider input
FULL
<100K
<70K
Radio input option
WORD CLOUD:
SELECT
TOP 10 titles
TOP 15 titles
TOP 20 titles
drop down option.
DISCUSSIONS
ADVANTAGES (+)
* Idea about the annual income and the five stats can be gained easily
* Quick view on most employee title can be found out (also its frequency)
DISADVANTAGES (-)
* No much interaction is there in word cloud.
* If there are too many outliers, box-plot will not be visible.
FOCUS/ZOOM
title count
PROJECT MANAGER
ACCOUNTANT
max
Q3
Q2
Q1
min

SHREET – THREE



SHREET – FOUR

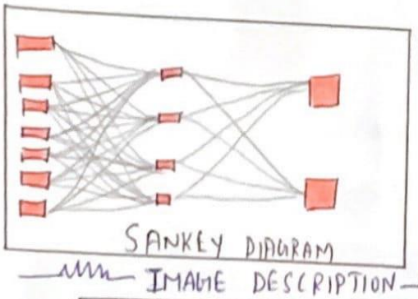
SHEET-4

TITLE: ANALYSIS ON LOAN APPLICANTS	
AUTHOR: AKSHAYA KUMAR CHANDRASEKARAN	
DATE: 28-05-2020	TASK: WEB PAGE DESIGN & PRESENTATION
SHEET: 04	

LAYOUT:

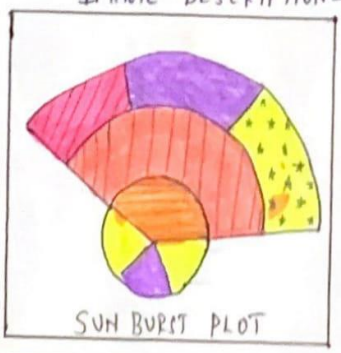
PROJECT TITLE

IMAGE DESCRIPTION



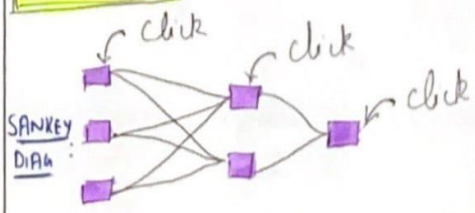
SANKEY DIAGRAM

IMAGE DESCRIPTION



SUN BURST PLOT

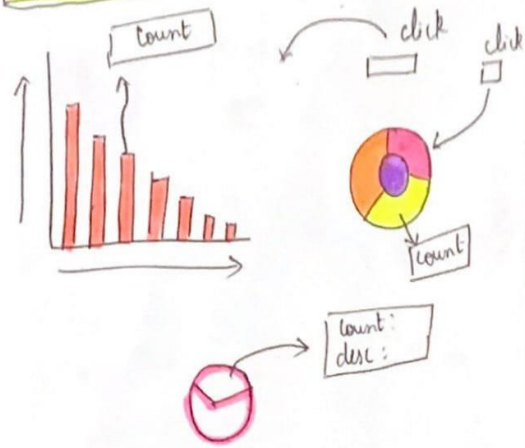
OPERATIONS:



Clicking on Sankey diagram will give us a new bar/donut chart.

SUNBURST: Choosing/clicking on a category further expands more data based on the category chosen

FOCUS / ZOOM:



DISCUSSIONS

-ve

* Sankey diagram cannot be drawn for the entire data set. Only for Top 10 (or Top 20 categories can be made compactly.
* If there is a lot of hierarchy, then Sunburst would not be the best way of representation. Becomes a complicated process.

+ve

* More user interactive and more info can be gained.

SHREET – FIVE/ DESIGN SHEET