Bank Loan Case Study

Project Description:

As a data analyst at a finance company specializing in lending various types of loans to urban customers, the primary objective of this project is to analyse patterns in loan application data using Exploratory Data Analysis (EDA). The goal is to identify capable applicants and reduce the risk of loan defaults by understanding customer attributes and loan characteristics.

Approach:

1. Identify Missing Data: Identify the missing data in the dataset.
2. Identify Outliers: Identify outliers in the dataset using excel functions.
3. Analyse Data Imbalance: Determine if there is data imbalance in the dataset and calculate ratio of imbalance.
4. Perform Univariate, Segmented Univariate and Bivariate Analysis:

   Perform Univariate analysis to compare value distribution and bivariate analysis to explore the relationships between variables.
5. Identify Correlations for Different Scenarios: Segment dataset based on scenarios and identify the top correlations using excel functions.

Tech-Stack:

Microsoft Excel 2019: Utilized for data cleaning, analysis, visualizations and reports.

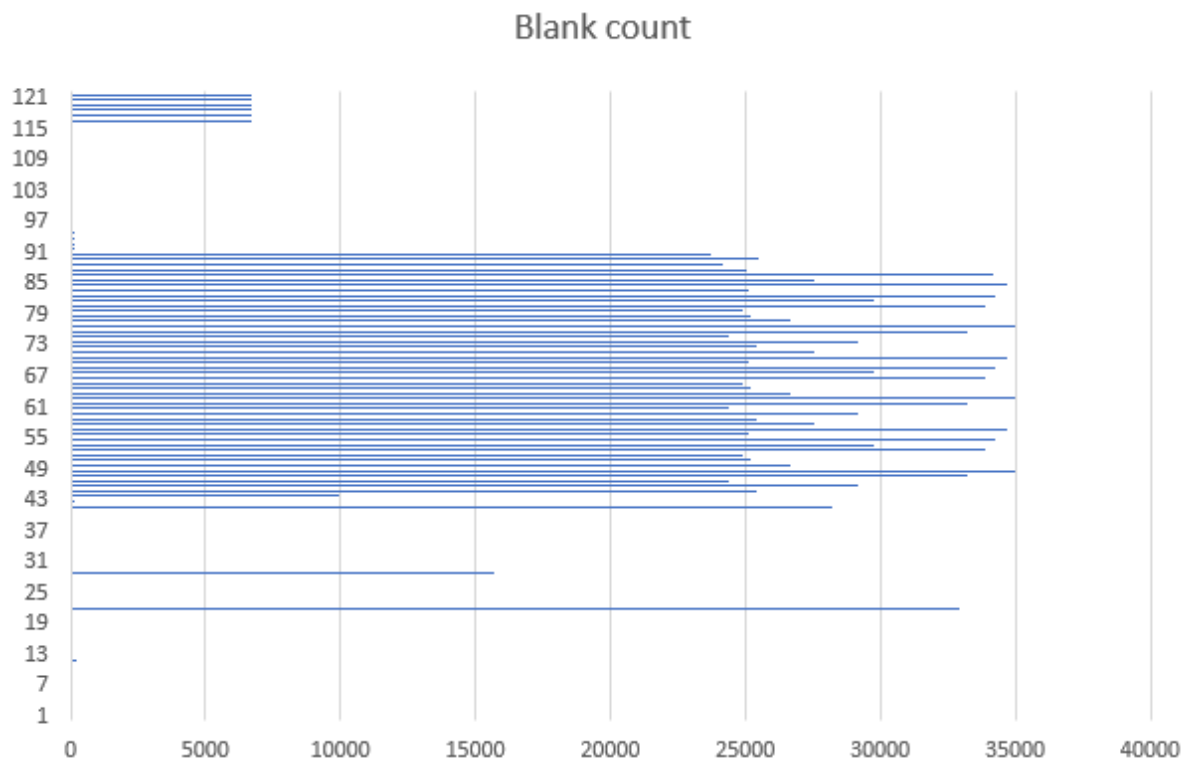
Insights:

1. Identify Missing Data and Handle Appropriately:
   - Missing data was identified across several columns.
   - Imputation using median values was more effective than average for skewed data.
   - Visualization showed that certain variables had a higher proportion of missing values, and those columns were dropped for further analysis.
   - COUNTBLANK() functions was used to identify the number of blanks and Median imputation for numerical data and Mode imputation for non-numerical data.

   The columns in red had missing value percentage above 40% and those columns were dropped.

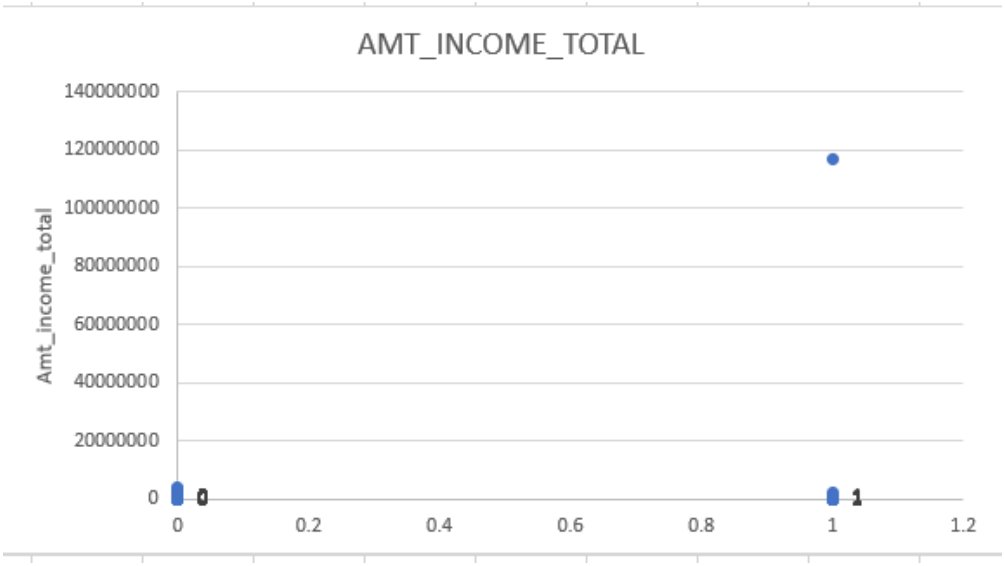
0	AMT_CREDIT
1	AMT_ANNUITY
38	AMT_GOODS_PRICE
192	NAME_TYPE_SUITE
0	NAME_INCOME_TYPE
0	NAME_EDUCATION_TYPE
0	NAME_FAMILY_STATUS
0	NAME_HOUSING_TYPE
0	REGION_POPULATION_RELATIVE
0	DAYS_BIRTH
0	DAYS_EMPLOYED
0	DAYS_REGISTRATION
0	DAYS_ID_PUBLISH
32950	OWN_CAR_AGE
0	FLAG_MOBIL
0	FLAG_EMP_PHONE
0	FLAG_WORK_PHONE
0	FLAG_CONT_MOBILE
0	FLAG_PHONE
0	FLAG_EMAIL

0	LIVE_CITY_NOT_WORK_CITY
0	ORGANIZATION_TYPE
28172	EXT_SOURCE_1
126	EXT_SOURCE_2
9944	EXT_SOURCE_3
25385	APARTMENTS_AVG
29199	BASEMENTAREA_AVG
24394	YEARS_BEGINEXPLUATATION_AVG
33239	YEARS_BUILD_AVG
34960	COMMONAREA_AVG
26651	ELEVATORS_AVG
25195	ENTRANCES_AVG
24875	FLOORSMAX_AVG
33894	FLOORSMIN_AVG
29721	LANDAREA_AVG
34226	LIVINGAPARTMENTS_AVG
25137	LIVINGAREA_AVG
34714	NONLIVINGAPARTMENTS_AVG
27572	NONLIVINGAREA_AVG
25385	APARTMENTS_MODE
29199	BASEMENTAREA_MODE
24394	YEARS_BEGINEXPLUATATION_MODE
33239	YEARS_BUILD_MODE
34960	COMMONAREA_MODE

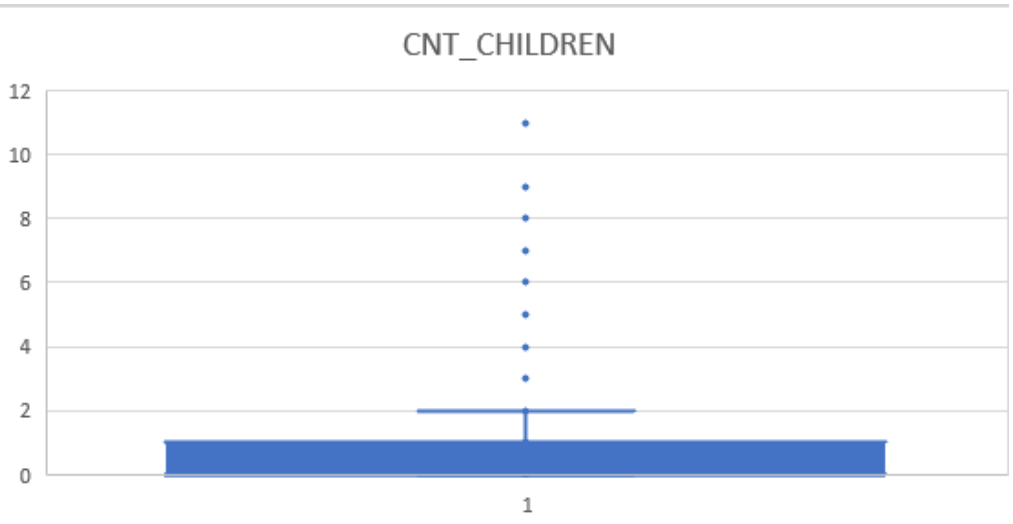


2. Identify Outliers in the Dataset:
    - Outliers were primarily found in Amt_income, Count_children and AMT_Credit.
    - Box plots revealed significant outliers that could distort analysis if not handled appropriately.
    - Calculated Inter Quartile (IQR), using Q3-Q1 formula where Q: Quartile

Quartile 1	112500
Qurtile 3	202500
Inter Quartile	90000
Upper limit	337500
lower limit	-22500



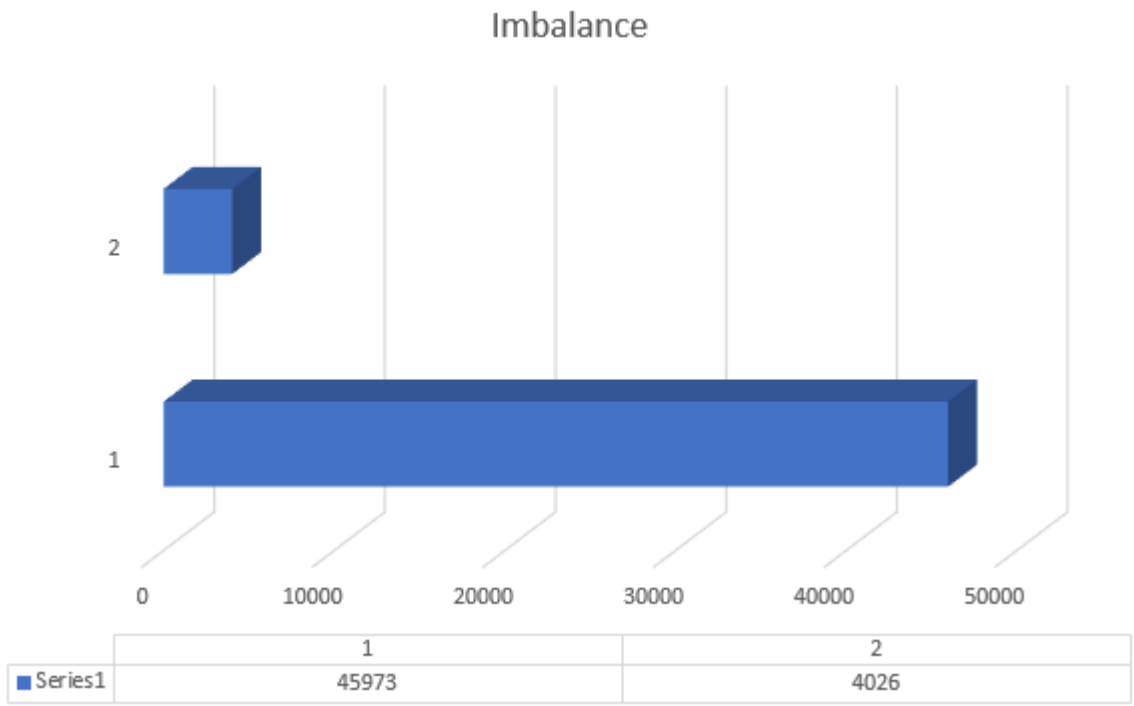
Qurtile 1	0
Quartile 3	1
IQR	1
Upper limit	2.5
Lower limit	-1.5



In AMT_INCOME_TOTAL the outlier is approximate 11,700,000 in the dataset this value is a lot more than any other value and hence appears unreal.

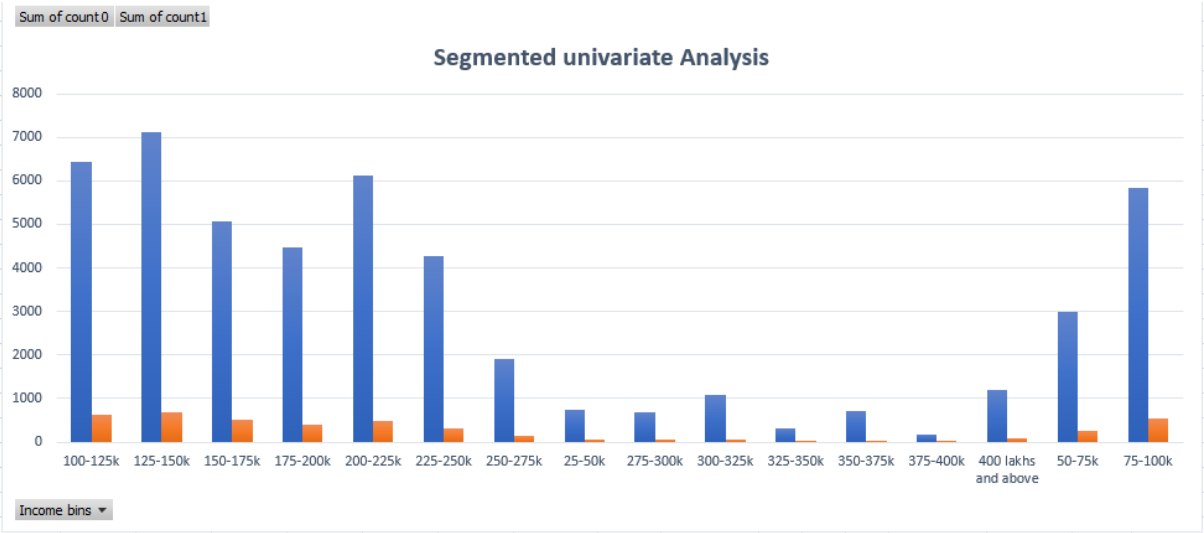
3. Analyse Data Imbalance:
    - There is significant imbalance in the target variable, with fewer instances of loan defaulters.
    - Visualizations highlighted that there are 45973 loan repairs and 4026 are loan defaulters or did not pay loan on time.

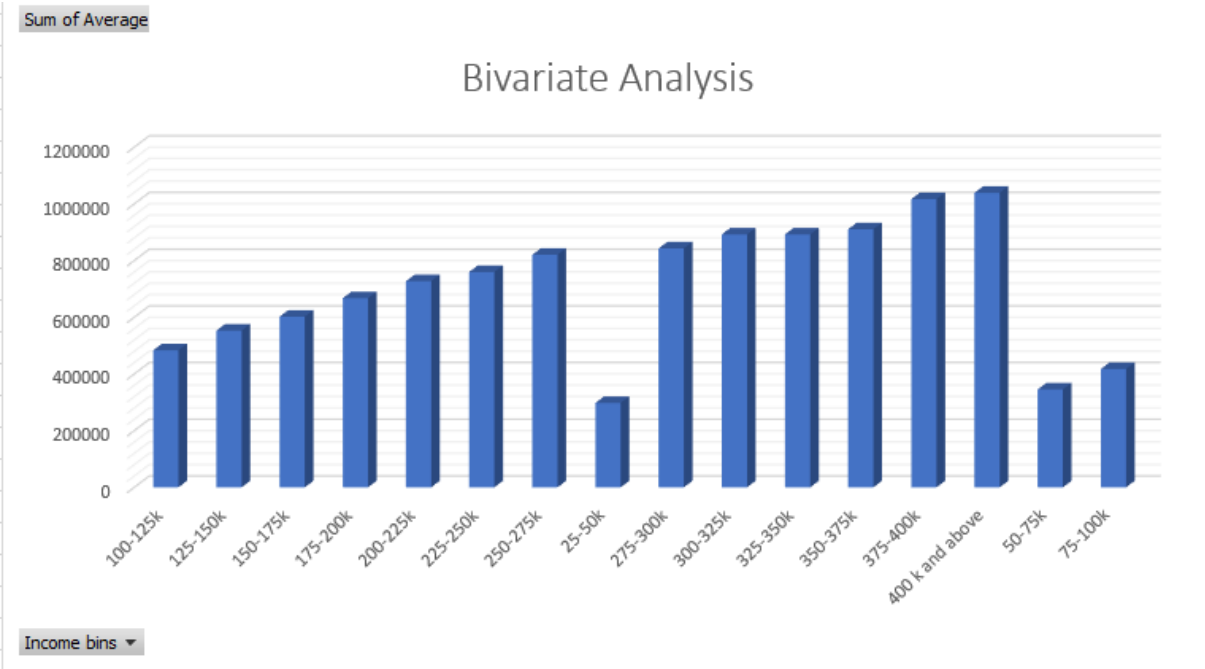
0 count	1 count	ratio of imbalance
45973	4026	11.41902633



4. Perform Univariate, Segmented Univariate and Bivariate Analysis:
    - Univariate analysis is where we consider or analyse only one variate at a time.
    - The most applicants were in range of 125k-150k i.e. 7804 applicants.
    - Bivariate analysis identified strong relationship between income bins and credit bins, deriving the no.of applicants that lie under the credit bins category.
    - With help of Univariate analysis, it showed that higher income bins had fewer defaults.
    - Pivot tables were used to visualize the data with bar graphs and column graphs.

Income bins	count 0	count1	Total count	Average
25-50k	741	63	804	297752.08
50-75k	2980	246	3226	345240.36
75-100k	5826	536	6362	417267.88
100-125k	6428	620	7048	483568.81
125-150k	7126	678	7804	553042.16
150-175k	5060	501	5561	602034.4
175-200k	4458	389	4847	667004.42
200-225k	6121	491	6612	727198.44
225-250k	4279	304	4583	759541.38
250-275k	1919	143	2062	820255.35
275-300k	681	45	726	842725.65
300-325k	1076	59	1135	892300.07
325-350k	322	24	346	892332.65
350-375k	723	34	757	910363.05
375-400k	186	14	200	1016814.4
400 lakhs and above	1205	98	1303	1038904.9





5. Identify Top Correlations for Different Scenarios:

Correlation matrices and Heatmaps were used to effectively visualize the relationship.

Top Correlation for Re-payers:

CNT_FAM AND CNT_CHILDREN

AMT_ANNUITY AND AMT_CREDIT

DAYS_EMPLOYED AND DAYS_PUBLISHED

AMT_ANNUITY AND REGION_POPULATION

Correlation of diff in Payments								
CNT_children	1	0.009588558	0.00497156	0.026178823	-0.025555665	-0.2415396	0.0321158	0.880453
AMT_INCOME	0.009588558	1	0.069315897	0.083008508	0.029841469	-0.0315103	-0.003507	0.011226
AMT_CREDIT	0.00497156	0.069315897	1	0.769498914	0.095111221	-0.0677394	0.0122288	0.063997
AMT_ANNUITY	0.026178823	0.083008508	0.769498914	1	0.115111507	-0.1087096	-0.006716	0.077377
REGION_POP	-0.025555665	0.029841469	0.095111221	0.115111507	1	-0.0041583	0.0043451	-0.023037
DAYS_EMP	-0.241539565	-0.03151033	-0.06773941	-0.108709643	-0.004158337	1	0.2727667	-0.230761
DAYS_PUB	0.032115773	-0.003506646	0.012228765	-0.006716454	0.004345136	0.2727667	1	0.026074
CNT_FAM	0.880453292	0.011225511	0.063997155	0.077377071	-0.02303741	-0.2307607	0.0260743	1
	CNT_children	AMT_INCOME	AMT_CREDIT	AMT_ANNUITY	REGION_POP	DAYS_EMP	DAYS_PUB	CNT_FAM

Top Correlations for Payments Defaulters:

AMT_ANNUITY AND AMT_CREDIT

DAYS_EMPLOYED AND DAYS_PUBLISHED

REGION_POPULATION AND AMT_CREDIT

CNT_FAM AND CNT_CHILDREN

Correlation of Payment made								
CNT-CHILDREN	1	0.009588558	0.00497156	0.026178823	-0.025555665	-0.241539565	0.032115773	0.805140097
AMT_TOTAL	0.009588558	1	0.069315897	0.083008508	0.029841469	-0.03151033	-0.00350665	0.011178205
AMT_CREDIT	0.00497156	0.069315897	1	0.769498914	0.095111221	-0.06773941	0.012228765	0.061329491
AMT_ANNUITY	0.026178823	0.083008508	0.769498914	1	0.115111507	-0.108709643	-0.00671645	0.074021491
REG_POP	-0.025555665	0.029841469	0.095111221	0.115111507	1	-0.004158337	0.004345136	-0.0212927
DAYS_EMP	-0.241539565	-0.03151033	-0.06773941	-0.108709643	-0.004158337	1	0.272766672	-0.21097453
DAYS_PUB	0.032115773	-0.003506646	0.012228765	-0.006716454	0.004345136	0.272766672	1	0.022590141
CNT_FAM	0.805140097	0.011178205	0.061329491	0.074021491	-0.021292698	-0.210974532	0.022590141	1
	CNT-CHILDREN	AMT_TOTAL	AMT_CREDIT	AMT_ANNUITY	REG_POP	DAYS_EMP	DAYS_PUB	CNT_FAM

Result:

The project provided a comprehensive understanding of factors influencing loan defaults. Key insights included the importance of income and loan amount in predicting defaults, the impact of data imbalance on analysis, and the identification of outliers and missing data. These findings will help the company make informed decisions about loan approvals and risk management. The insights shows that most of the clients are loan re-payers.

Contact Details: manumoolimani7@gmail.com