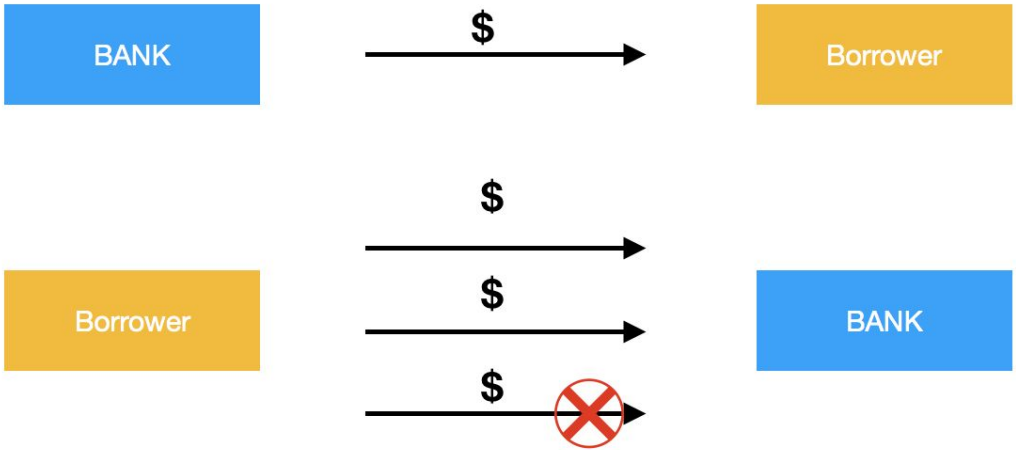
Credit Risk Prediction

Belle Pandya . Kate Weber

Introduction

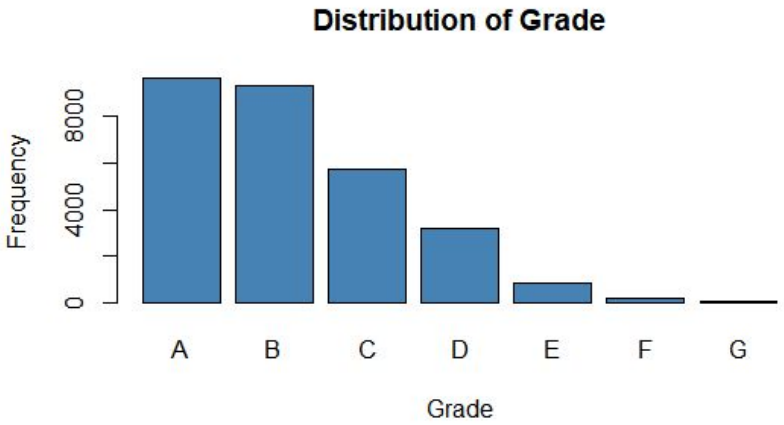
Banks determine whether to lend money to a customer based on many factors, including the customer's age, home ownership status, annual income, and credit rating. These factors are intended to be used to predict whether the customer will pay back the loan or defect, in which case, the bank would not lend money to them.

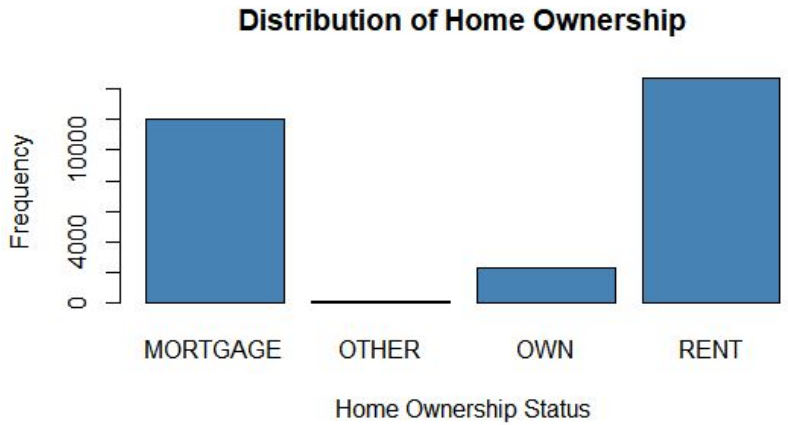


Dataset Description

Size (n)	~30,000 rows
Number of Predictors	8
Responding Variable	1, Loan_status (Binary)
Continuous Variables	5, (age, emp_length, loan_amount, int_rate,annual_income)
Categorical Variables	2, grade (A to G), home_ownership(own, mortgage, rent, others)

Categorical variables





Adding Predictors

Decomposed categorical variables into dummy variables:

Grade

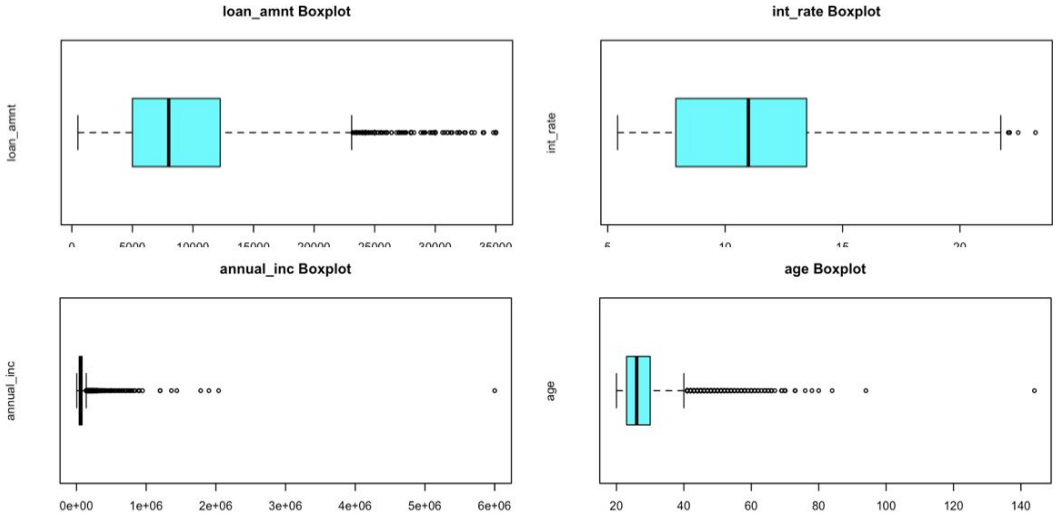
- grade.A
- grade.B
- grade.C
- grade.D
- grade.E
- grade.F
- grade.G

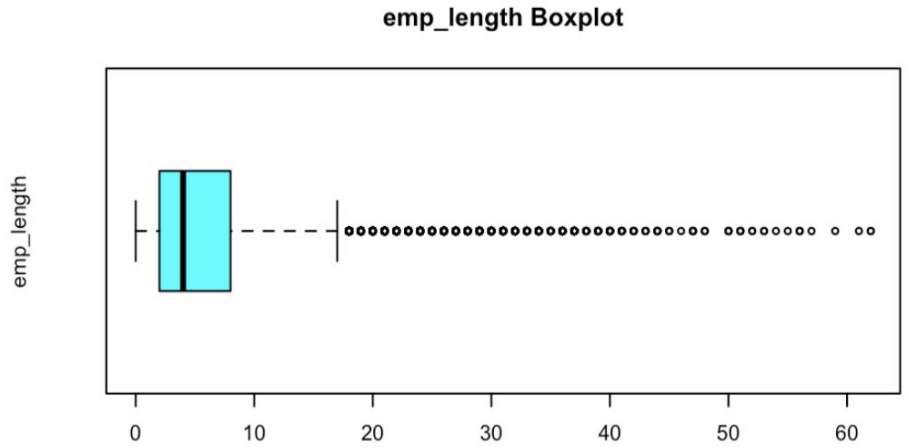
Home Ownership

- home_ownership.RENT
- home_ownership.MORTGAGE
- home_ownership.OWN
- home_ownership.OTHER

1. Outliers

Box Plots

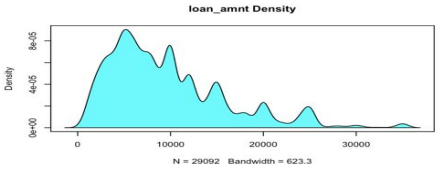

loan_amnt Boxplot
loan_amnt
int_rate Boxplot
int_rate
annual_inc Boxplot
annual_inc
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
age Boxplot
age
20
40
60
80
100
120
140

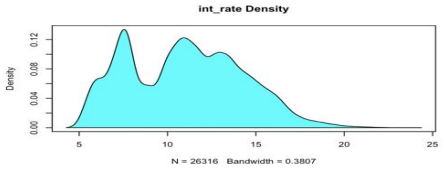

emp_length Boxplot
emp_length
0
10
20
30
40
50
60

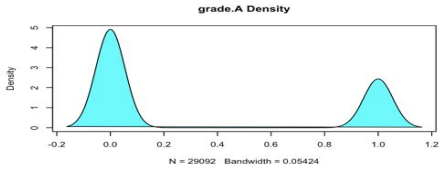
2. Skewness

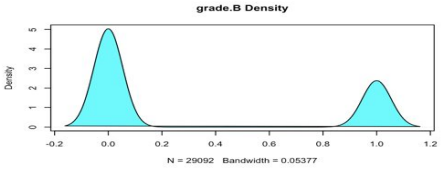
Density plots

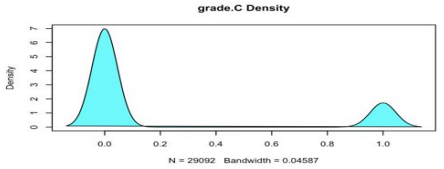
- Highly Skewed : 11
- Moderately Skewed : 2
- Approx Normal : 3

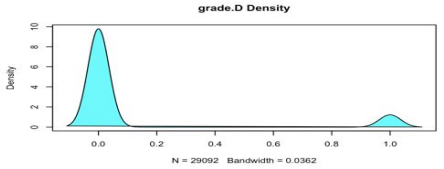


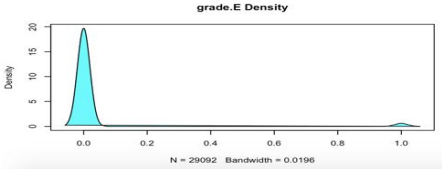


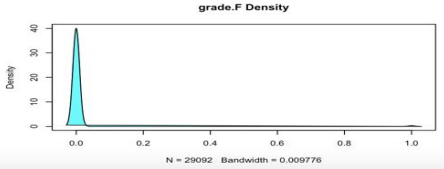


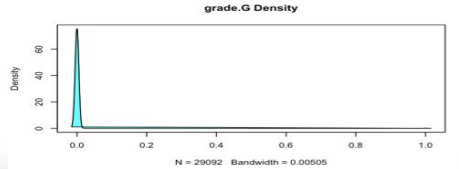


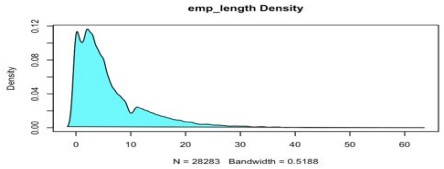


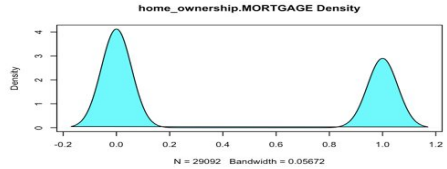


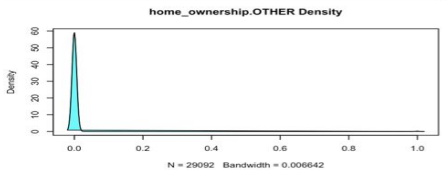


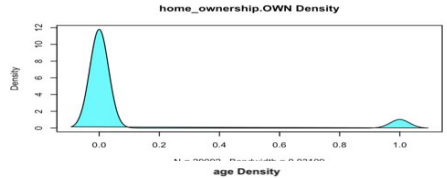


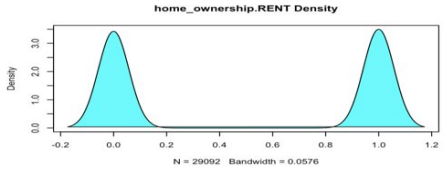


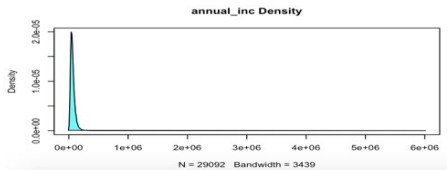


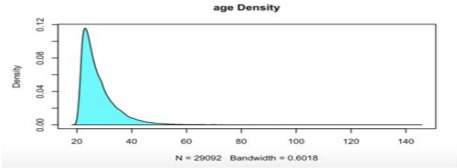












Skewness Summary

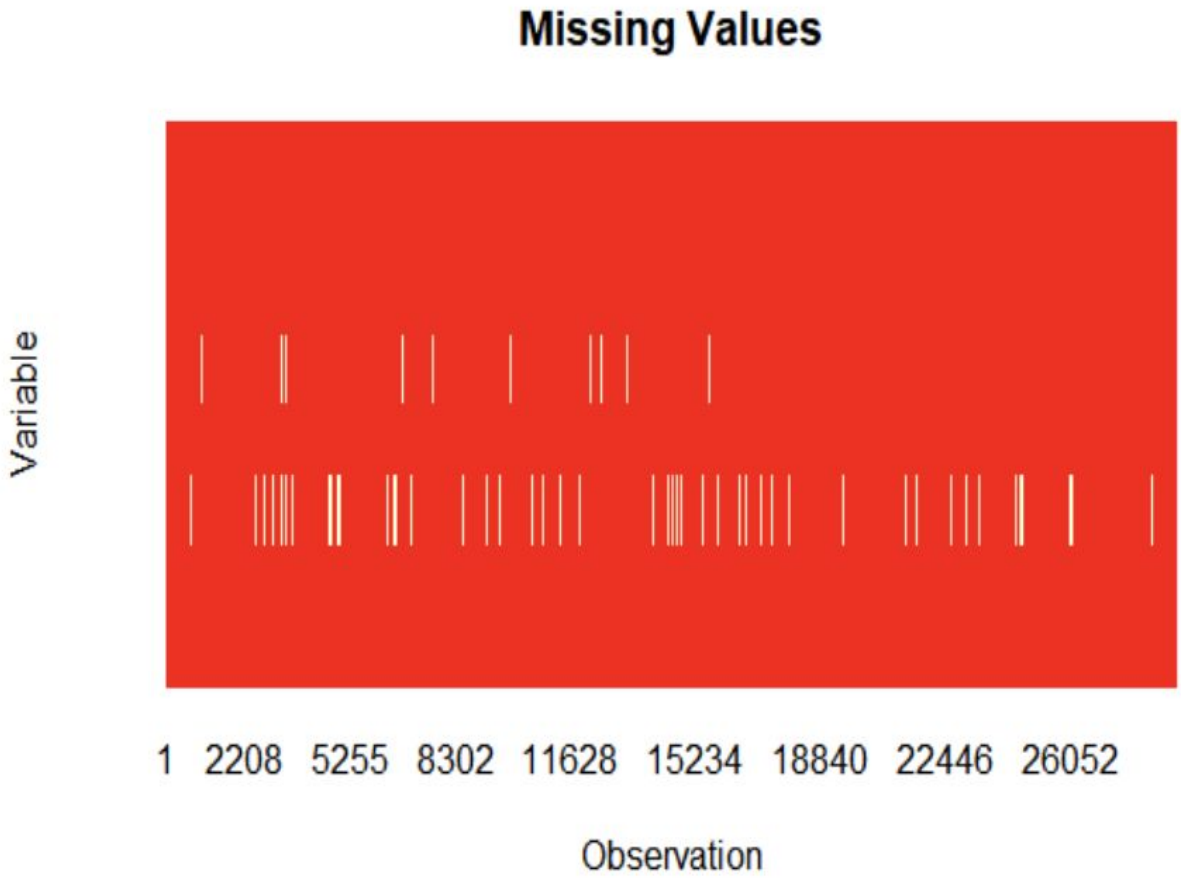
Variable	Skewness Coefficient	Skewed
Loan Amount	1.19104171	Highly Skewed
Interest Rate	0.20972676	Approx. Normal
Grade A	0.71501461	Moderately Skewed
Grade B	0.76839487	Moderately Skewed
Grade C	1.51895722	Highly Skewed
Grade D	2.47554614	Highly Skewed
Grade E	5.52663850	Highly Skewed
Grade F	11.61336188	Highly Skewed
Grade G	22.72550700	Highly Skewed
Empl. Length	2.10235517	Highly Skewed
Home Own. Mortgage	0.35524409	Approx. Normal
Home Own. Other	17.23050357	Highly Skewed
Home Own. Own	3.11898740	Highly Skewed
Home Own. Rent	-0.02007422	Approx. Normal
Annual Income	33.75117342	Highly Skewed
Age	2.15185778	Highly Skewed

3. Missing Values

Missing Values

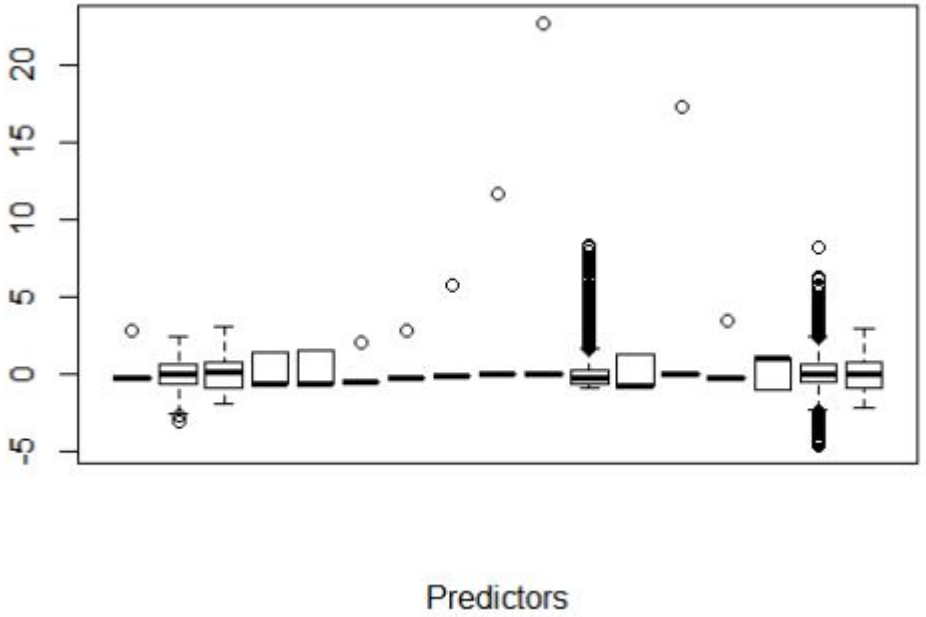
Variables with missing values-

1. Interest rate (int_rate)
2. Employment length (emp_length)



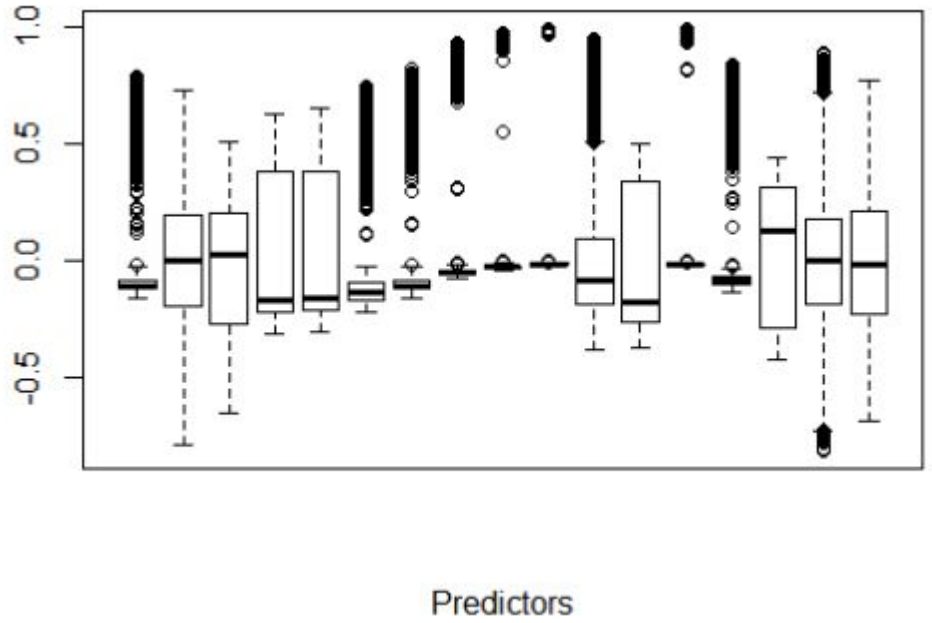
Transformations

- Center
- Scale
- Box-Cox
- KNN Imputation



Transformations

- Spatial Sign



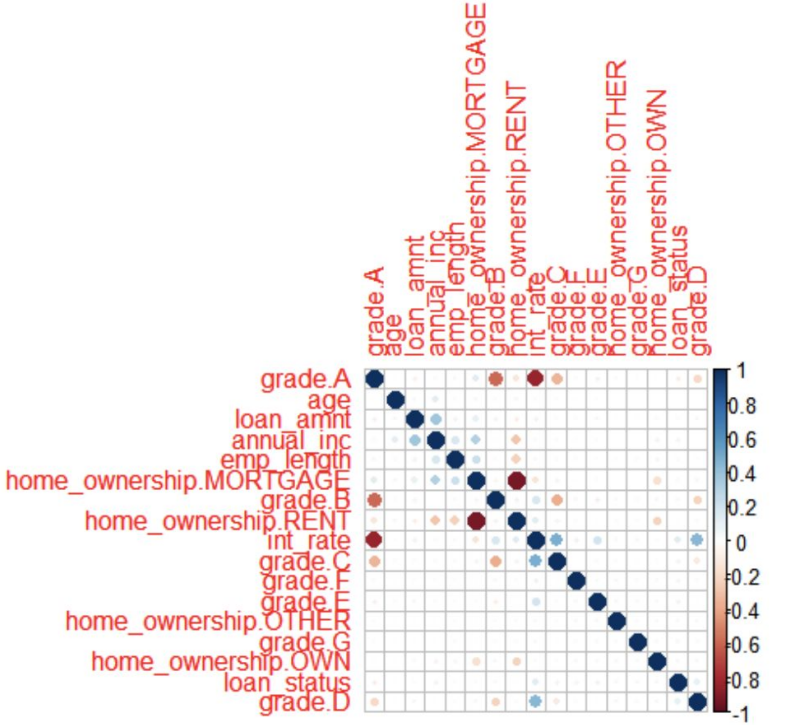
Removing predictors

Near Zero Variance

We did not find any predictors that had near-zero variance, which would have been an indication that the predictor should be removed.

```
> nearZeroVar(loan.pp.ss)
integer(0)
>
```

Correlation Plot



Principal Component Analysis

- Cutoff Value = 90%
- Requires 12 PCs to reach threshold
- Removing 5 vars - not worth complexity

PC	Cum. % Variance
1	15.27427
2	27.51806
3	36.55126
4	43.77207
5	50.77822
6	56.99055
7	62.94289
8	68.68281
9	74.39220
10	80.05765

Data Splitting and Re-sampling

Data Splitting

- Stratified Random Sampling
- Split the data into train and test set with a ratio of 80:20

Re-Sampling

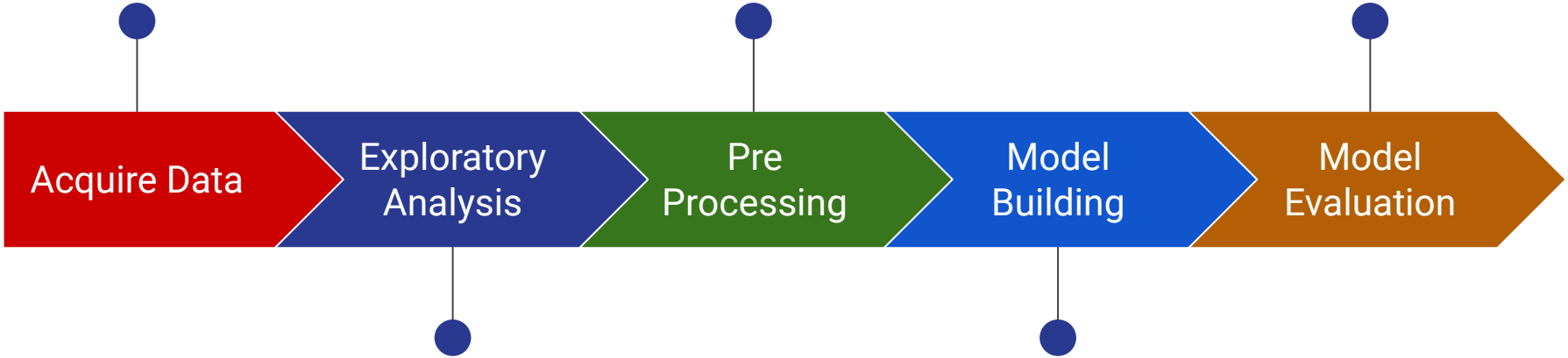
- K-fold cross validation, where k=10.

Summary

Acquired data from Kaggle and converted categorical variables into dummy variables

Performed center, scale,Box-cox, kNN imputation, spatial sign transformation.

TBD..



Plotted box plots to identify outliers, density plots for skewness and error plot to identify variables with missing values.

Splitting data into training and test using stratified random approach. Planning to perform K fold cross validation.