What Banks Need to Know to Avoid Loan Defaults

Region vs Default Status

🟢 Dividing the US into 4 regions: North, South, North-East and Central, the probability was calculated to determine the region the loan applicant belongs to given that the applicant had defaulted the loan but it was found that the probability for the applicant belonging to North or South Regions is almost equal, approximately 46%. However, the probability that the person belongs from the North-East or Central region is about 7.5%.

🟢 The mean Credit Score is equal for both genders - Male and Female.

```
    Wilcoxon rank sum test with continuity correction

data:  Credit_Score[Gender == "Male"] and Credit_Score[Gender == "Female"]
W = 273401872, p-value = 0.1169
alternative hypothesis: true location shift is not equal to 0
```

Using the wilcoxon rank sum test, the hypothesis that the mean Credit Score for Male/Female is equal was accepted. [p-value> 0.05, accept above null hypothesis.]

🟢 The working class age group, ages 25-54, have average interest rates higher than 3.5%.

Using the Z-test method for hypothesis testing, the result was that there is 86.6% chance that ages 25-54 pay interest over 3.5% on average.

🟢 The co-applicant credit type has affects the possibility of loan default by the applicant.

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(loanDefault$co.applicant_credit_type, loanDefault$Status)
X-squared = 3092.4, df = 1, p-value < 0.00000000000000022
```

Using Chi-squared test on the data, p-value of almost 0 was obtained. This means that we can reject the hypothesis that co-applicant credit type does not have any effect on the loan default status.

🟢 The average rate of interest differs with respect to different age ranges and also differs with respect to different loan types.

The age groups like <25, 25-34, 35-44 and 45-54 have significantly different average rate of interest. The mid ages see lesser difference whereas we can say comparing with the ages above 65, there is no significant different in the average rate of interest.

95% family-wise confidence level

Differences in mean levels of age

95% family-wise confidence level

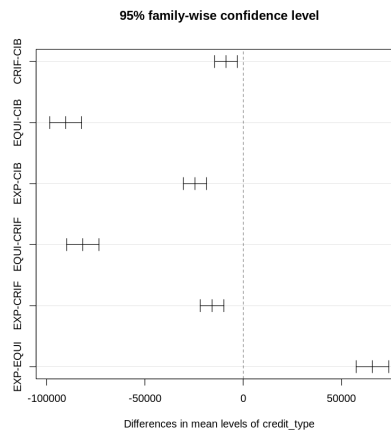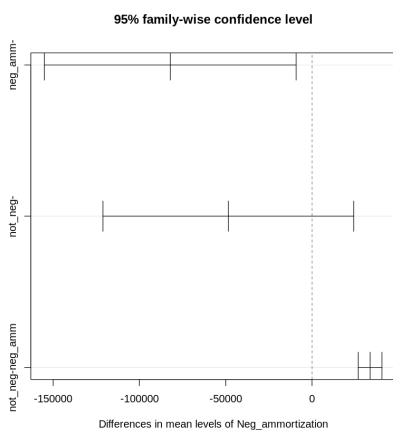Differences in mean levels of loan_type

The differences in mean levels of loan type for rate of interest are significantly different.

It was also found that age and loan type are significantly associated with each other.

Two-way ANOVA was used to analyse the group variances and comparison among the group means was done using Tukey HSD method.

● The property value varies with amortization status-ammortized or negatively ammortized. The average property value differs with credit type.

95% family-wise confidence level

Differences in mean levels of Neg_ammortization

95% family-wise confidence level

Differences in mean levels of credit_type

● Loan Default Prediction using Logistic Regression with 86% accuracy.

The accuracy obtained with the logistic regression model is 86.5% which is much greater than the baseline model accuracy, which is 75%.
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. Observing the above ROC curve, we can see that curve is far from 45 degree line, which suggests high performance of the model.

● The main influential factors that can be used to explain the variance of the data with respect to the applicant is inflow and outflow.
Inflow here contains factors like income and debt-to-income ratio which tells about the credit worthiness. Whereas, Outflow contains the loan amount(to pay) and property value(to purchase) , in other words, liability.