Credit card payment default project inferential analysis

For statistical analysis in this credit card payment default project, I took the following steps to verify the findings from graphical exploratory data analysis and identify the correlations between variables.

- Correlation between credit limit and bill amount: Pearson R

Since credit limit and bill amount are both continuous variables, Pearson R could be used to detect the correlations between them. We assume customers who have a high credit limit will have a higher bill amount and indeed the Pearson R result shows a positive correlation between them.

- Correlation between credit limit and payment default: t-test

Ideally, we should have customers' income data, but since this data is not available, we can assume higher income customers have higher credit limits. Therefore, we will verify if there is a correlation between credit limit and default payment using a t-test.

- ❏ Null hypothesis: credit limit does not affect default likelihood.
- ❏ Alternative hypothesis: credit limit impact default likelihood.
- ❏ Set significance level α to 0.05

Test result: we get a p value as 0, therefore we need to reject the null hypothesis and accept the alternative hypothesis. Credit limit has an impact on payment default.

- Correlation between education and payment default: chi-squared test

Previous visualization indicates education impacts default likelihood. We will use a chi-squared test to verify this finding since these are both categorical variables.

- ❏ Null hypothesis: education does not affect default likelihood.
- ❏ Alternative hypothesis: education impacts default likelihood.
- ❏ Set significance level α to 0.05

Test result: the p value is close to 0, so we will reject the null hypothesis and accept the alternative hypothesis. Because education has a strong correlation with default probability, we should keep this variable in the machine learning model.

- Correlation between age and payment default: chi-squared test

Previous visualization indicates age impacts default likelihood. We will use a chi-squared test to verify this finding.

- ❏ Null hypothesis: age does not affect default likelihood.
- ❏ Alternative hypothesis: age impacts default likelihood.
- ❏ Set significance level α to 0.05

Test result: the p value is smaller than significance level α,we will reject the null hypothesis and accept the alternative hypothesis, which is age has impact on default probability.

- Correlation between sex and payment default: permutation test and t-test

In previous data visualization, it appears males tend to default more than females. Does sex have any correlations with default or was this observation due to chance event? Let's find out with a permutation test and a t-test on each group's default proportions and mean respectively.

- ❏ Null hypothesis: sex has no impact on default probability.
- ❏ Alternative hypothesis: sex has impact on default probability.
- ❏ Set significance level α to 0.05.

Test result: since the p value is 0, we should reject the null hypothsis and accept the alternative hypothsis, which is sex has impact on default likelihood. To be more concrete in this case, male customers tend to default more.

- Correlation between credit limit and sex: t-test

Since we know the credit limit is strongly correlated with default probability. Let's see if sex plays any role in credit limit and default likelyhood.

- ❏ Null hypothesis: sex has no impact on credit limit.
- ❏ Alternative hypothesis: sex has impact on credit limit.
- ❏ Set significance level α as0.05.

Test result: the p value is much lower than α , so we will reject null hypothesis and accept alternative hypothesis. We have verified that sex plays a role in credit limit and it is not due to chance.