Exploratory Data Analysis

From data visualization and exploration, I found that the gender, education and age affected the default_payment. The collected data shows that more female default than male card holders. The plot shows that the card holders with university level education have the maximum default, and the plot show some correlation between these variables. Age group with default plot shows that the younger the age group, higher the default. It shows an inverse relationship between age-group and default.

According to the data, I found that the female card holders default more than male card holders. This suggest that gender might have some effect on default. To check if this did not happen by chance, I performed a population proportion difference hypothesis T-test. The null hypothesis is that the population male default mean and population female default mean is equal. The alternative hypothesis is population male mean and population female default mean is not equal. We got a t-value of 6.85, and a very low p-value. The test suggests that the male mean default and female mean default is not equal, because the p-value calculated is less than the level of significance 0.05.

I also conducted a Chi-squared test to evaluate the relationship between gender and default payment, education-level and default payment, marital_status and default payment, and age-group and default payment. First, I evaluated the relationship between gender and default payment. The null hypothesis is that gender and default_payment are independent of each other or not related. The alternative hypothesis is that gender and default_payment are not independent or related. I used the level of significance alpha 5%. The p-value calculated from the chi-squared test was very low less than 0.001. Because the p-value is less than 0.05, we reject the null hypothesis and suggest the alternative hypothesis. Therefore the gender and default payment are related and dependent to each other.

Then, I performed the chi-squared test for education and default_payment. The p-value for this test was very low, less than 0.001. The result is significant at p<0.05, so we will reject the null hypothesis. We have enough evidence to support the alternative hypothesis that the education level and default_payment are related and dependent on each other.

Then, I performed the chi-squared test for marital_status and default_payment. The null hypothesis for this test is that marital_status and default_payment are not related. The alternative hypothesis for this test is that marital_status and default_payment are related. After the calculation, p-value we got is approximately 0.0000007, which is less than 0.001. The level of significance is 0.05. Because the p-value is less than 0.05, we reject the null hypothesis that the marital_status and default_payment are not related. We suggest the alternative hypothesis that the marital_status and default_payment are related.

Finally, I compared age-group and default_payment with chi-squared test. The null hypothesis is that age-group and default_payment are not related. The alternative hypothesis is that age-group and default_payment are related. The test suggested that the age-group and default_payment are dependent and related to each other. The p-value is approximately 0.0000003, which is less than the level of significance 0.05. The evidence suggest the alternative hypothesis and rejects the null hypothesis.