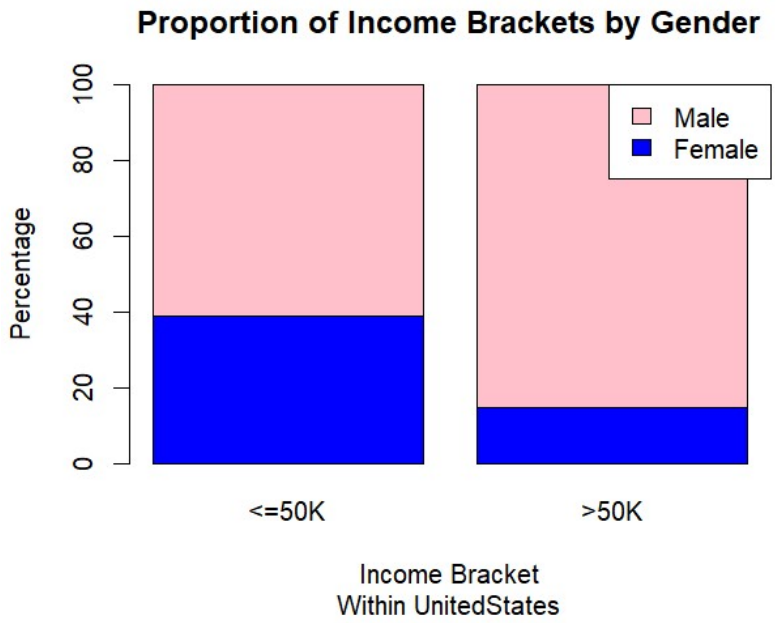
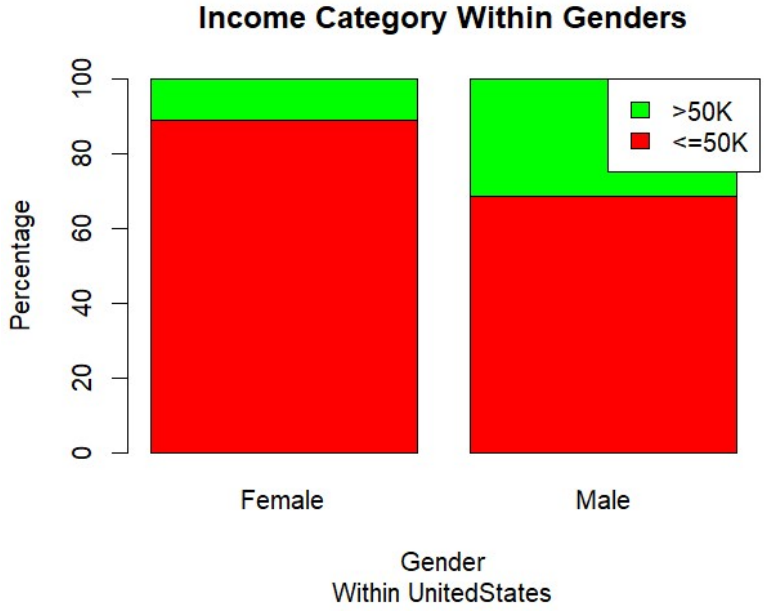
1) Visualization

1.1 Appropriate plot for the RQ output of an R script (NOT a screenshot) and required supplementary graph/table

USA: Proportion of Gender within Income Brackets



USA: Income category Distribution within Genders



Contingency table

United states: Sex vs Income Bracket Contingency table

	<=50K	**>50K**
Female	**8594**	**1068**
Male	**13371**	**6087**

- Sample size of the contingency table is 29,120, which is having 9,662 females in total with a proportion of 93.3% earning <=50K i.e., around 8,594 and remaining >50K. There are 19,458 males where nearly 68.75 earn <=50K i.e., around 13,371 and remaining >50K with respect to USA.
- For comparison of proportions analysis, we will use a chi-square test to check for the likelihood that there is a relationship between the two nominal variables (reason for absenteeism and age group in example). This non-parametric test makes no assumptions about the shape of the data (which is nominal, not interval) so we do not include a histogram for this test.

1.2 Additional information relating to understanding the data

The bar plot makes the difference easy to spot, males clearly have a larger green section, showing that more of them earn over 50K. The contingency table then backs this up with actual numbers, helping us confirm that what we see in the chart is also supported by the data needed for the chi-square test.

1.3 Useful information for the data understanding

The plot makes it clear that most females are concentrated in the <=50K category, while males appear much more often in the >50K group. Even though both genders mainly earn below 50K, males are still far more likely to be high earners. This pattern suggests that gender could be playing a meaningful role in income differences.

2 Analysis

2.1 Statistical test used to test the hypotheses and output

We used a Chi-Square Test because both of our variables—gender and income bracket—are categorical. Our aim is to see whether the proportion of high earners is different for males and females. The chi-square test checks whether the differences we see in the

data are larger than what we'd expect by chance. It doesn't require normally distributed data, which makes it ideal for large tables like those in the Adult Income dataset. The test gives us a chi-square value, degrees of freedom, and a p-value.

2.2 The null hypothesis is rejected /not rejected (select one) based on the p-value.

The chi-square test for the USA data returned a very small p-value ($p < 0.0001$), meaning it's extremely unlikely that the differences we see happened by chance. Because of this, we reject the null hypothesis. This shows a clear relationship between gender and income bracket. Males are much more likely to earn over 50K, while females are mostly in the ≤50K group. These results match what we observed in both the bar chart and the contingency table, supporting the idea that gender influences income levels in this dataset.