**1. Without considering correlation among features.**

1. Quality of Data.

Correlation matrix

The synthetic data tries to maintain the structure of the original data, but the strength of some correlations is different.

Some business logic correlations, like those involving **gross income**, **Unit price**, and **Sales**, seem to degrade slightly in the synthetic version.

Some correlations between categorical features (like **Branch** and **City**) remain weak in both datasets, but that's to be expected due to the nature of those variables.

KL Divergence for Continuous Features

For columns like **"Unit price"**, **"Quantity",** and **"Rating",** the KL divergence is very low, meaning the distributions of these columns in the synthetic data are very close to the original data.

Columns like **"Sales",** **"gross income",** and **"Tax 5%"** have higher KL divergence, meaning the synthetic data deviates from the original data distribution in these columns. This suggests that the generation process could be fine-tuned for these columns to better match the original distribution.

**"cogs"** has the highest KL divergence, indicating a significant deviation between the distributions of the synthetic and original data for this column.

Most continuous features have acceptable KL divergence scores, but the higher KL values for "Sales,"

"cogs," and "gross income" suggest that the distribution of these columns in the synthetic data differs significantly from the original data

Statistical tests

Most of statistical measurements are close to the original data with close means and standard deviations, except for only 4 columns which are "**Tax 5%", "Sales", "Cogs"** and **"gross income"** , where the synthetic data tends to overestimate both the mean and standard deviation.

**Chi-Square Test for Categorical Features**

The **Chi-Square test** evaluates whether there is a significant difference between the observed frequencies (in the original data) and the expected frequencies (in the synthetic data) for categorical columns.

**Interpretation**:

* For columns like **"Branch"**, **"City"**, **"Customer type",** and **"Gender",** the p-values are high, meaning there is no significant difference between the categorical distributions of the original and synthetic data. This is a positive outcome, indicating that the distribution of categories in these columns is well-replicated.
* For **"Product line"** and **"Payment",** the p-values are a bit lower but still acceptable. The lower p-value for "Payment" indicates that there may be slight differences in the category distribution for this column.

**Conclusion**: The synthetic data performs well in replicating the categorical distributions of the original data. Most of the p-values are high, meaning the synthetic data distribution closely resembles the original distribution.

**Overall Conclusion:**

The synthetic data performs quite well in replicating the original data, especially in terms of:

* Categorical feature distribution (as confirmed by the Chi-Square test).
* Low KL divergence for most continuous features.

However, there are areas for improvement:

* For continuous features like **Sales**, **Cogs**, **Tax 5%**, and **Gross Income**, the generation process overestimates the data. Focusing on these columns and incorporating dependencies between them could reduce the KL divergence.
* **Business rules and constraints** between correlated features (like **Sales**, **Cogs**, and **Gross Income**) should be better modeled to improve the overall accuracy of the synthetic data.

2. Data Distribution

Numerical columns: according to the histogram of them, I assume they have these distributions which I used later to generate the synthetic data:

Unit price: Uniform

Quantity: Uniform

Tax 5%: gamma

Sales: gamma

Cogs: gamma

gross margin percentage: No distribution (fixed value)

gross income: gamma

Rating: Uniform

Categorical columns: according to the bar plot of them, I assume they have these distributions which I used later to generate the synthetic data:

Branch: Discrete Uniform

City: Discrete Uniform

Customer type: Binomial

Gender: Binomial

Product line: weighted categorical

Payment: weighted categorical

**2. With considering correlation among features.**

1. Quality of Data.

Correlation matrix

The copula-generated data successfully preserves many of the correlations observed in the original data, but some correlations are notably weaker or stronger:

* **Business logic correlations**: Key correlations involving **Unit price**, **Quantity**, and **Sales** are mostly preserved, but with slight degradation in the strength of correlations. For example, the correlation between **Quantity** and **Sales** is still present, though slightly weaker in the synthetic data.
* **Cogs (Cost of Goods Sold) and Gross Income**: These correlations, which are logically expected due to business rules, are also mostly captured but with minor variations in strength.
* **Categorical features**: Correlations between categorical features such as **Branch** and **City** remain weak in both datasets, which is expected given the nature of those variables.

Overall, while the copula-generated data maintains the structure of the original data, the strength of some correlations is not perfectly replicated.

### KL Divergence for Continuous Features

* **Low KL divergence for some features**: For columns like **Unit price** (0.0437), **Rating** (0.0709), and **gross margin percentage** (0.0), the KL divergence is low. This means that the copula-generated data’s distribution for these features closely matches the original data.
* **Higher KL divergence for key features**: Columns such as **Sales** (2.2417), **Tax 5%** (2.6861), **gross income** (2.6861), and **cogs** (2.2488) have higher KL divergence values. This suggests that the synthetic data’s distribution deviates significantly from the original data for these columns. These features are often business-driven and may require more fine-tuning to capture their distributions accurately.
* **Overall**: Most continuous features have acceptable KL divergence scores, but the higher KL values for **Sales**, **cogs**, **Tax 5%**, and **gross income** highlight areas where the copula model could be improved.

### Statistical Tests

* **Close mean and standard deviation**: Many of the statistical measurements (like means and standard deviations) are quite close between the copula-generated data and the original data. However, some features deviate:
  + **Tax 5%**, **Sales**, **Cogs**, and **gross income**: The copula-generated data tends to overestimate both the mean and standard deviation for these columns, which results in higher KL divergence and statistical differences.
  + **Unit price** and **Rating**: These columns exhibit close means and standard deviations between the original and synthetic data, indicating that the copula model performed well for these features.

### Chi-Square Test for Categorical Features

The Chi-Square test evaluates whether there is a significant difference between the observed frequencies in the original data and the synthetic data for categorical columns:

* **Good performance**: For columns like **Customer type**, **Gender**, and **Payment**, the p-values indicate no significant difference between the categorical distributions in the original and synthetic data. This is a positive outcome, meaning that the copula model effectively replicates the categorical feature distributions.
* **Lower p-values**: For columns like **Branch**, **City**, and **Product line**, the p-values are much lower, indicating that the copula-generated data does not perfectly replicate the categorical distribution in these columns. However, some deviations are expected when using statistical models like copulas for categorical data.

### Conclusion

The copula model is effective in preserving the general structure and relationships in the original dataset. Most categorical features are well-replicated, and several continuous features exhibit good KL divergence values. However, specific columns like **Sales**, **cogs**, **Tax 5%**, and **gross income** need further tuning to reduce the deviations between the original and synthetic data.

2. Data Distribution detected

The copula assumes that the dependence structure between variables follows a multivariate Gaussian distribution. This means it captures the correlation between different features based on a multivariate normal distribution.

3. Why did you choose that specific approach in the second case?

I used it because it was able to generate numerical and categorical data in the same time, it was flexible unlike some methods that are specialized for either numerical or categorical data, and also was able to closely capture the correlation and dependencies between features.

4. Specify which case is better to use and why?

### ****Column-wise Distribution Generation****:

* **Pros**:
  + Great for **precisely matching the distribution** of individual features.
  + **Lower KL divergence** for many continuous features like "Unit price" and "Quantity."
* **Cons**:
  + Does not capture **relationships between features**, leading to poor correlation preservation.
  + Business metrics like "Sales" and "Cogs" may deviate significantly due to the loss of dependency information.
* **Best for**: Univariate analysis and simple datasets with fewer interdependencies between features.

### ****Copula-based Generation****:

* **Pros**:
  + Excels at preserving **dependencies and correlations** between features.
  + Can handle both **categorical and numerical data** simultaneously.
  + More realistic for **business metrics** that rely on relationships between features, such as "Sales" and "Gross income."
* **Cons**:
  + Slightly higher **KL divergence** for some individual features.
  + More complex and computationally intensive than column-wise generation.
* **Best for**: Datasets with strong feature correlations and **multivariate relationships**, especially in business or finance settings.

### ****Conclusion****:

* If **feature correlations** are important and your data involves **complex relationships**, use **copulas**.
* If **individual feature distributions** are your primary concern, use **column-wise generation**.

In your case, **copulas** are likely the better choice due to their ability to capture **dependencies between features**.