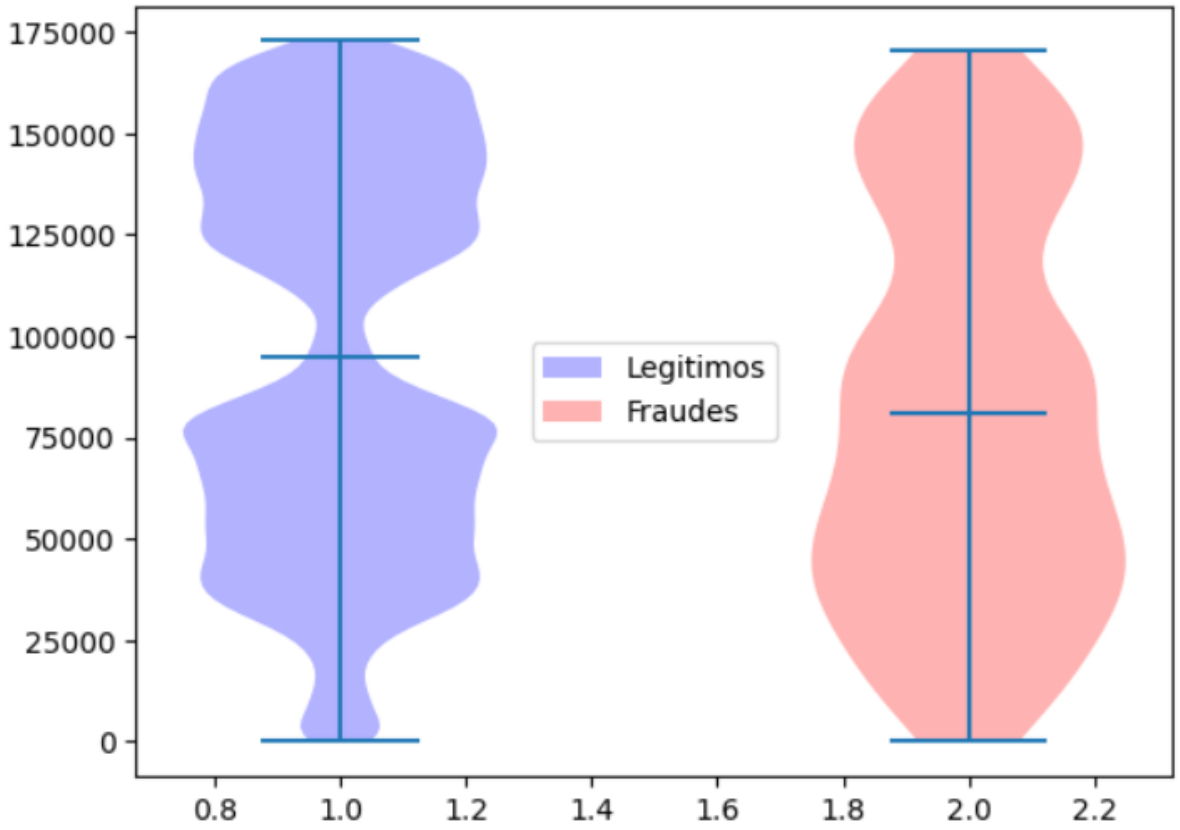
# Documentação sobre as análises

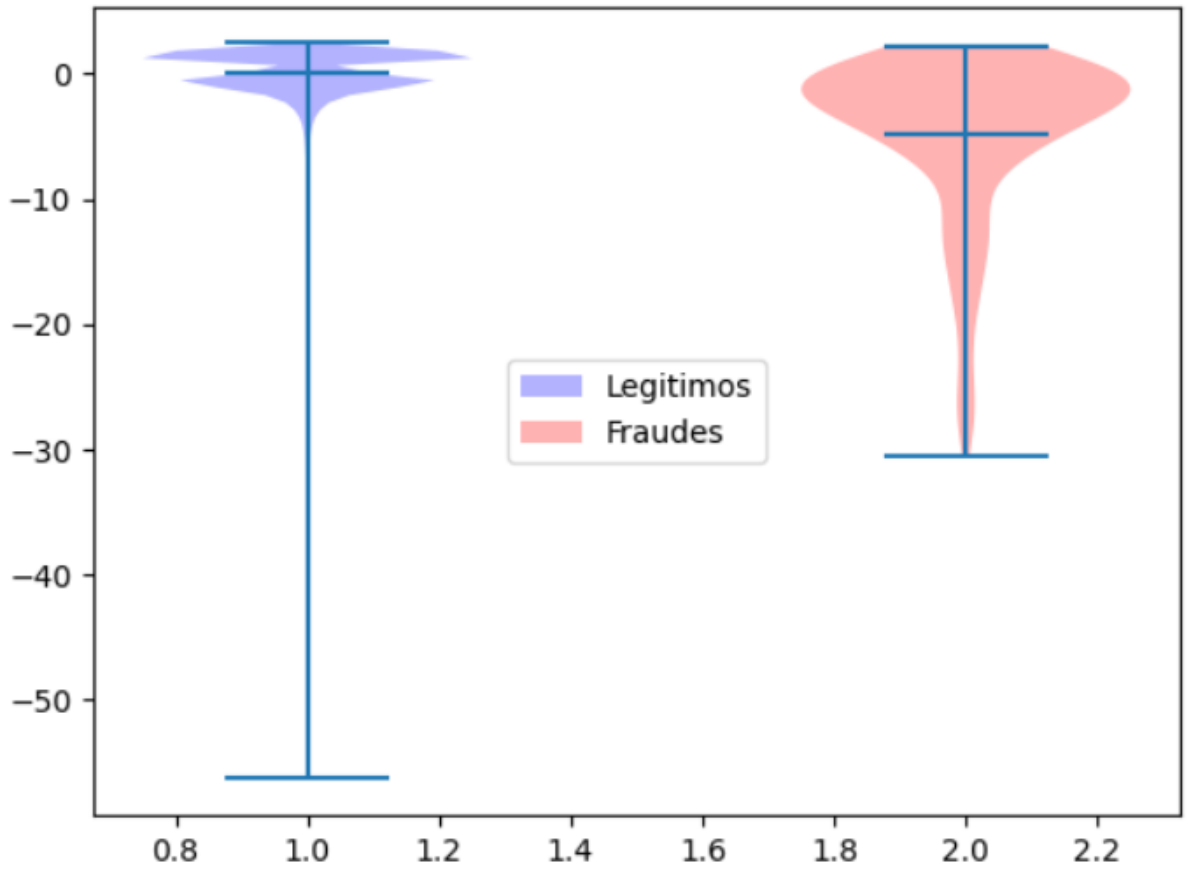
Analisando cada atributo e explicando o porquê cada um foi utilizado ou descartado, pegando as plotagens do colab:

- Time:



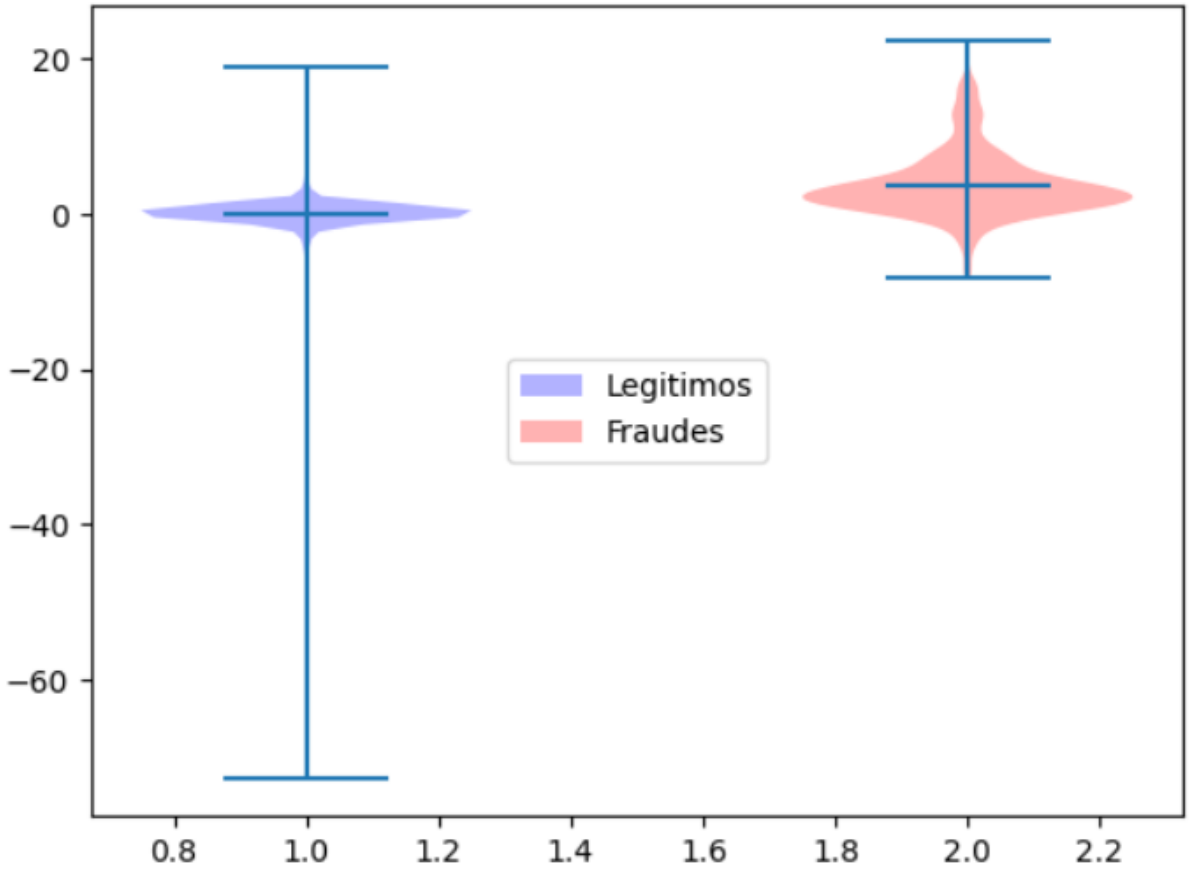

O atributo Time apresenta valores para as medianas muito próximos tanto dos legítimos quanto das fraudes. Por esse motivo, não é um bom atributo para diferenciação.

- V1:



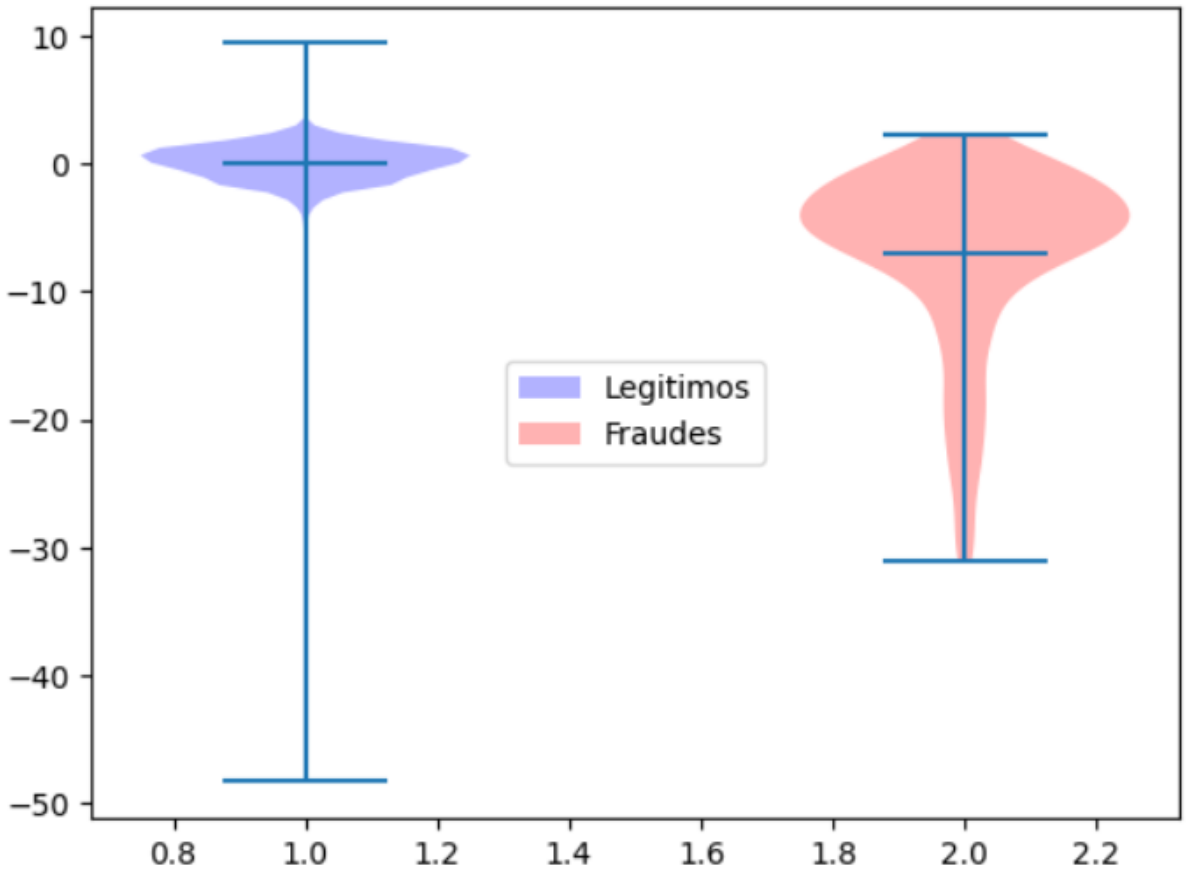

Assim como o anterior, não é bom para diferenciar legítimos de fraudes, além de possuir muitos outliers.

- V2:



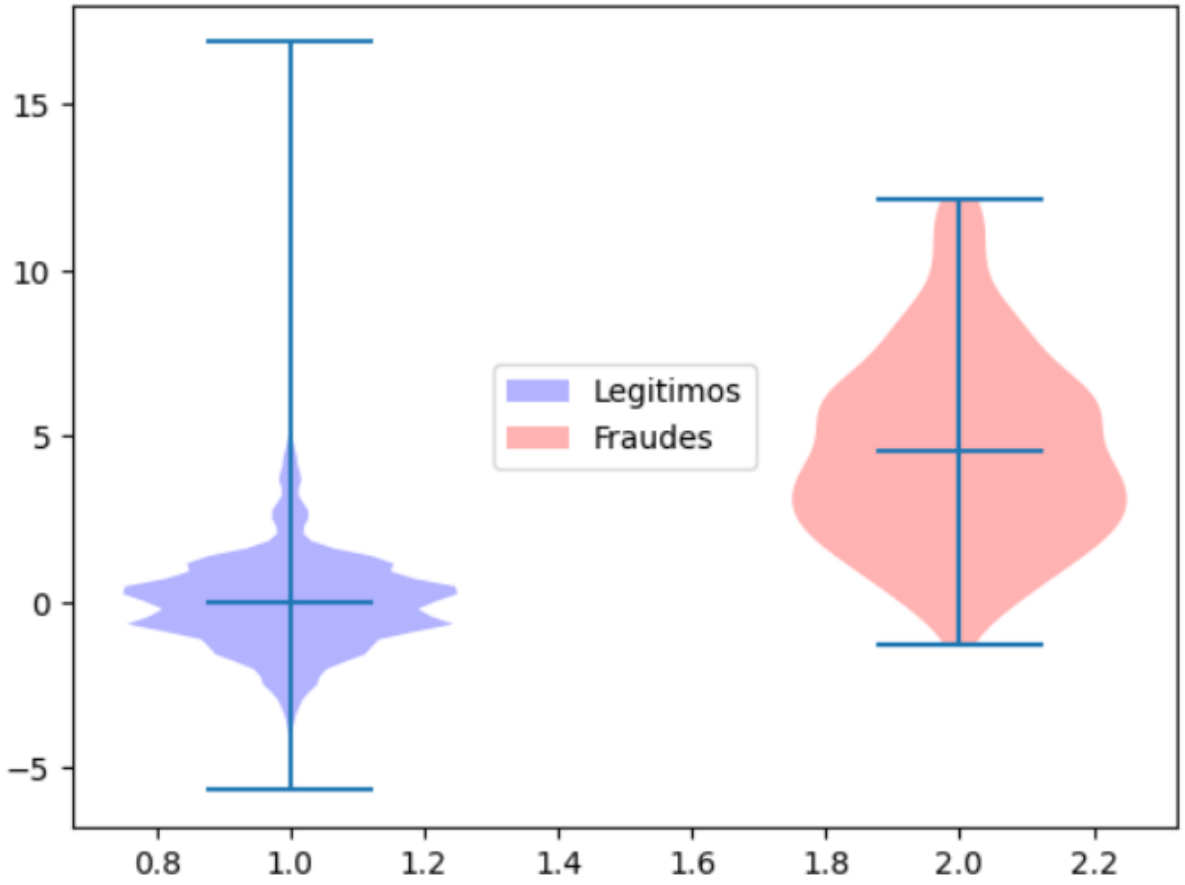

O mesmo ocorre com o atributo V2.

- V3:



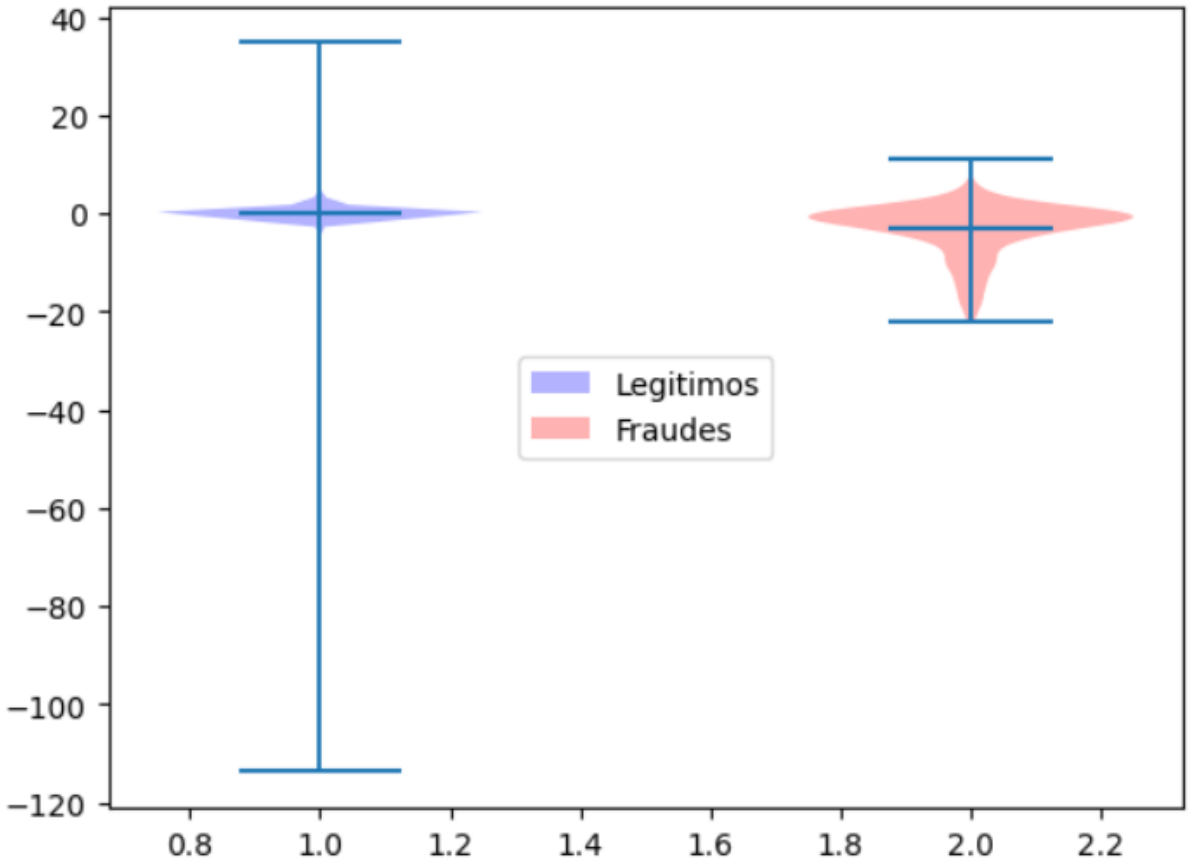

Além da distribuição ser diferente, a mediana também está um pouco distante. Por isso, utilizamos V3.

- V4:



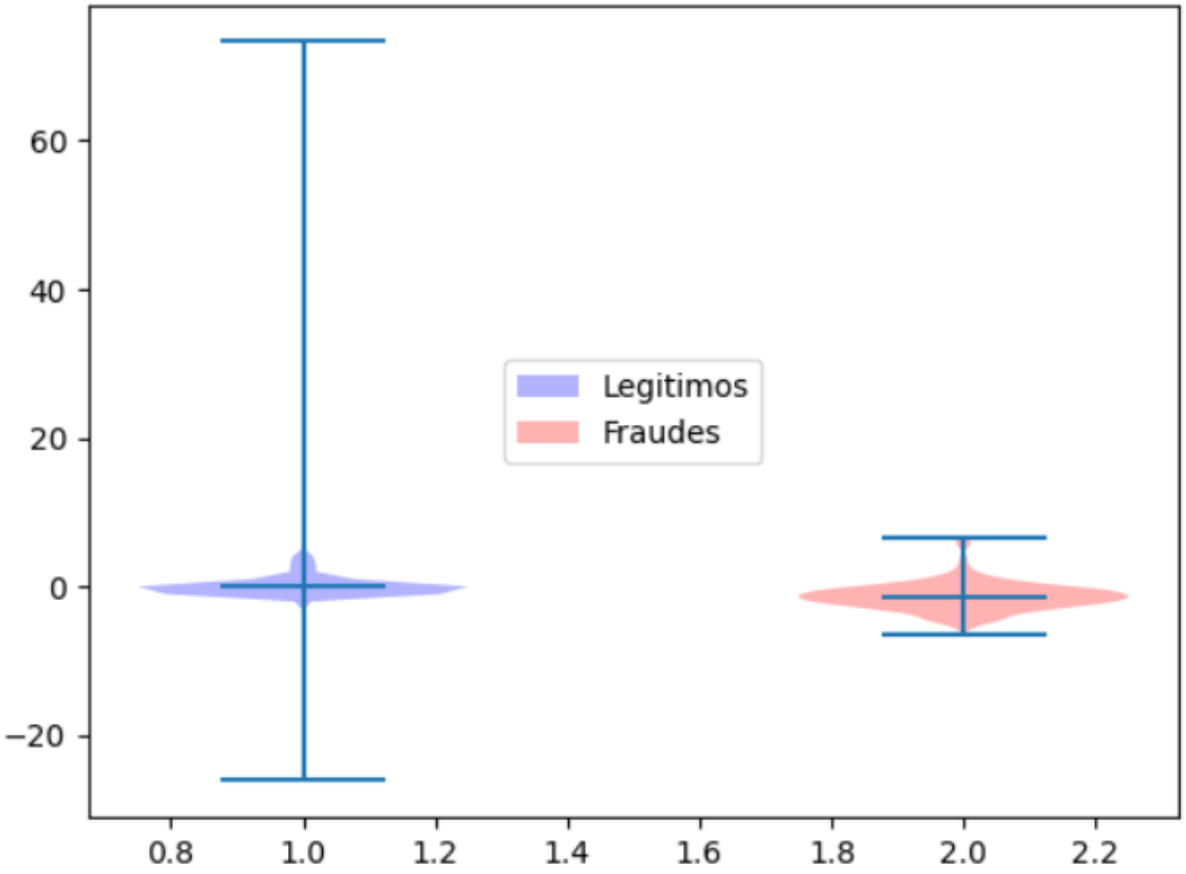

O mesmo que acontece com o atributo V3, ocorre no V4.

- V5:



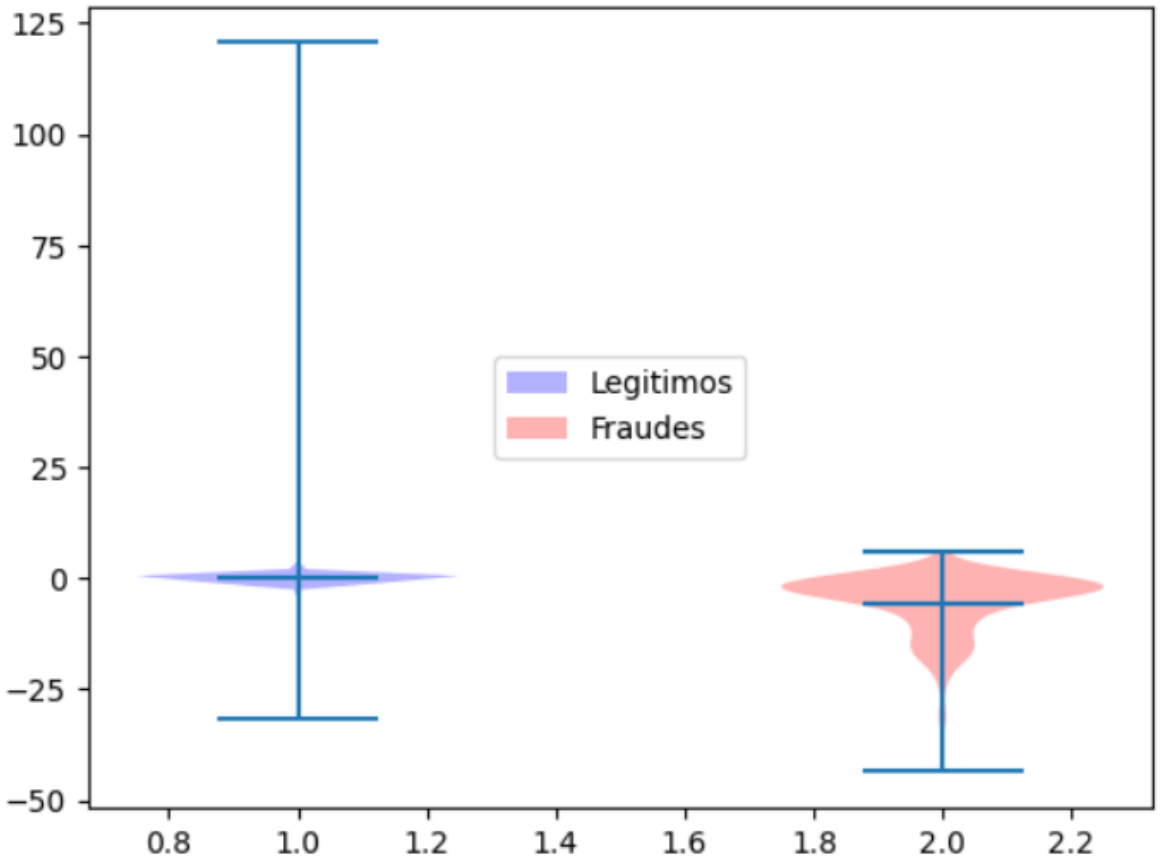

O atributo apresentado possui distribuições diferentes e medianas também, podendo utilizá-lo como diferenciador.

- V6:



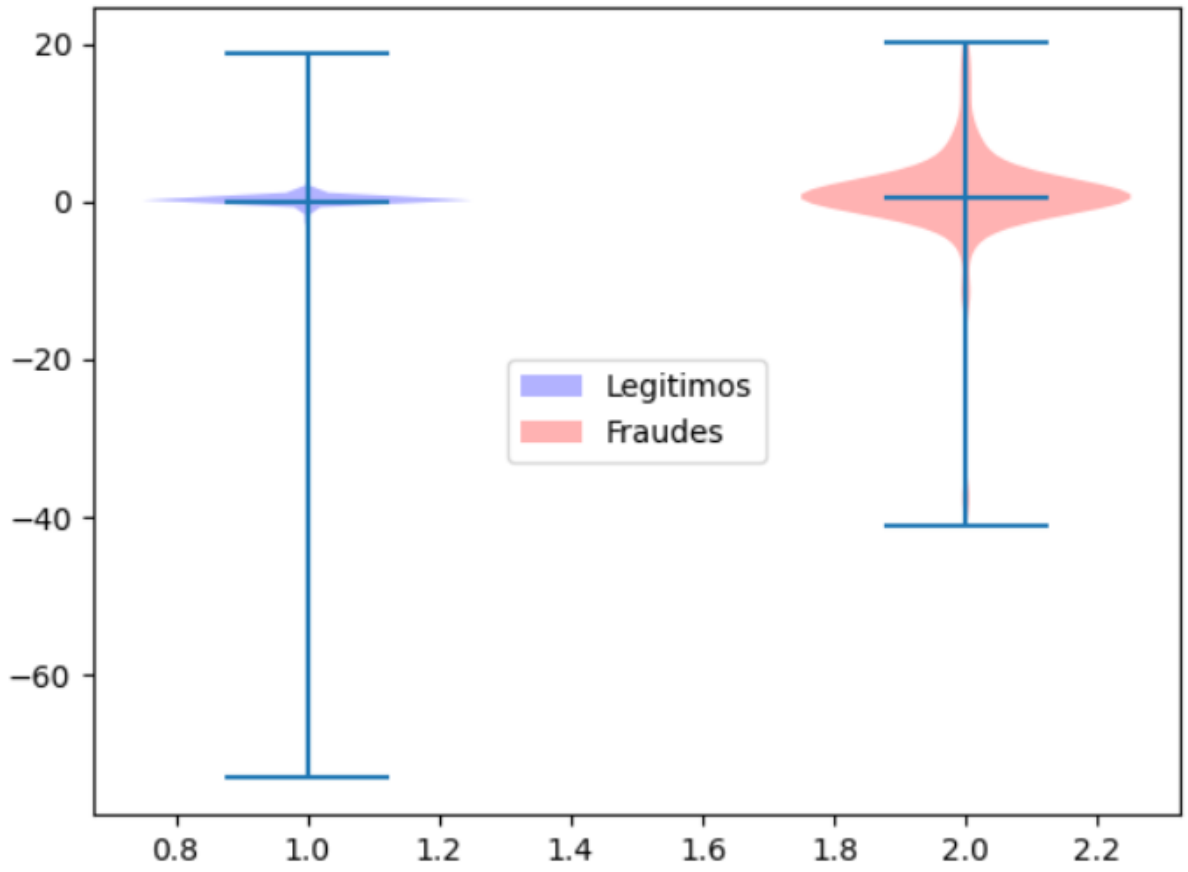

Muito próximo a distribuição e a mediana, por isso não utiliza-se.

- V7:



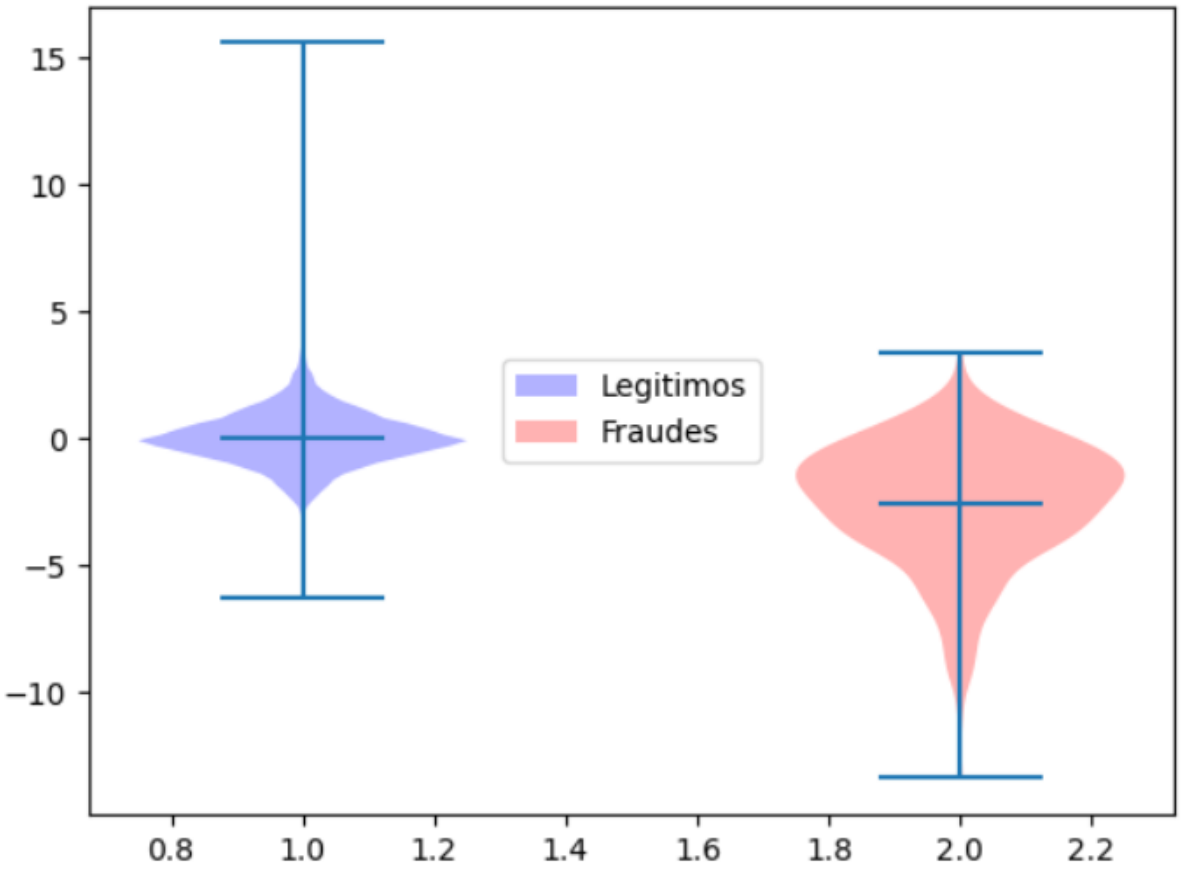

Distribuição diferente e mediana diferente. Por esse motivo, utilizamos V7.

- V8:



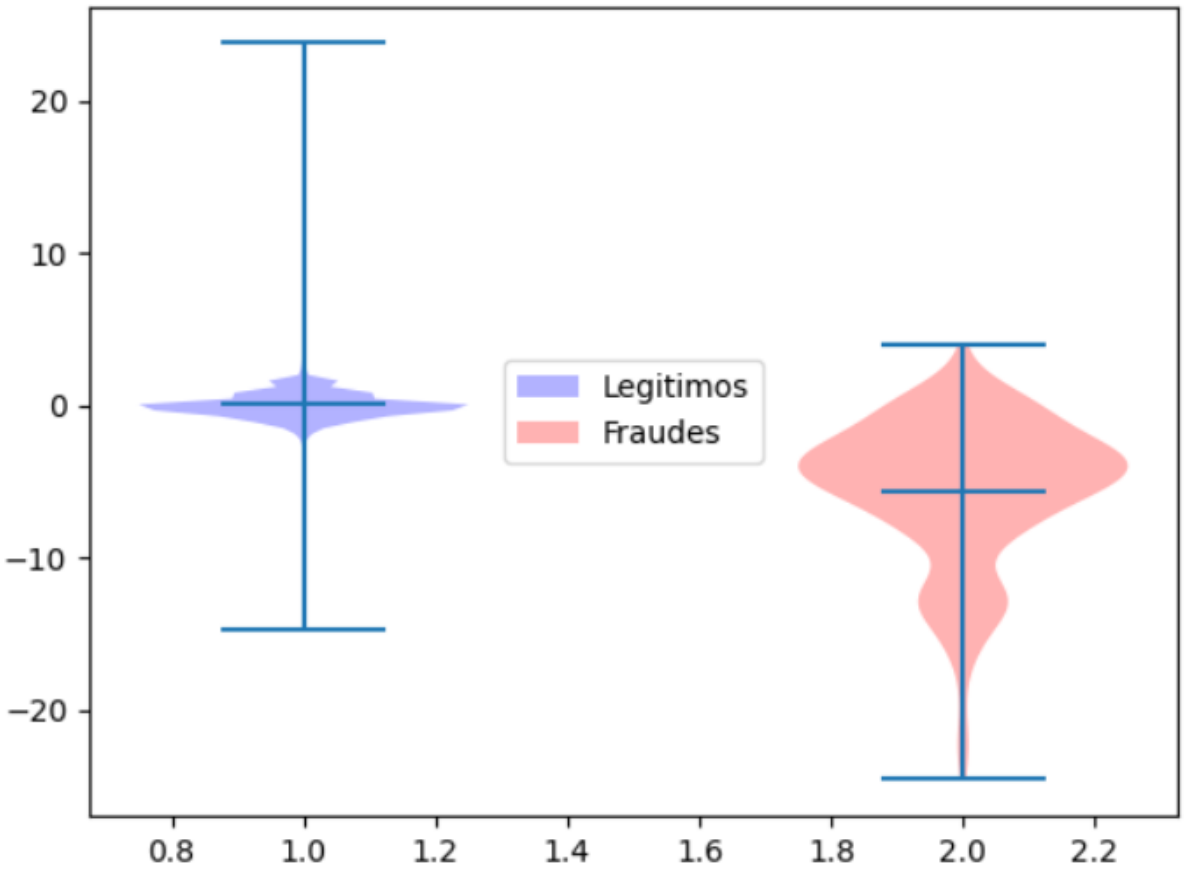

Distribuição semelhante e mediana também. Não podemos usar para diferenciar.

- V9:



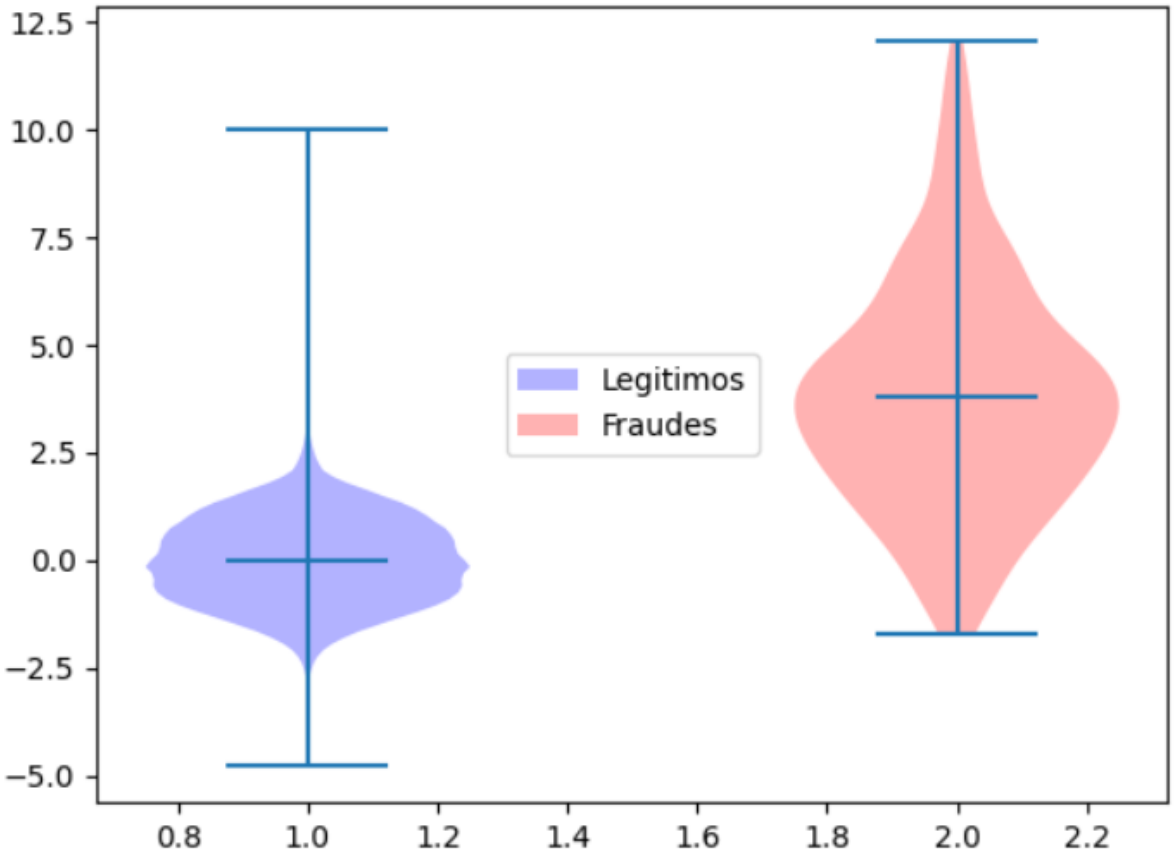

As distribuições dos dados não são parecidas e, ademais, a mediana possui valores diferentes. Usamos V9.

- V10:



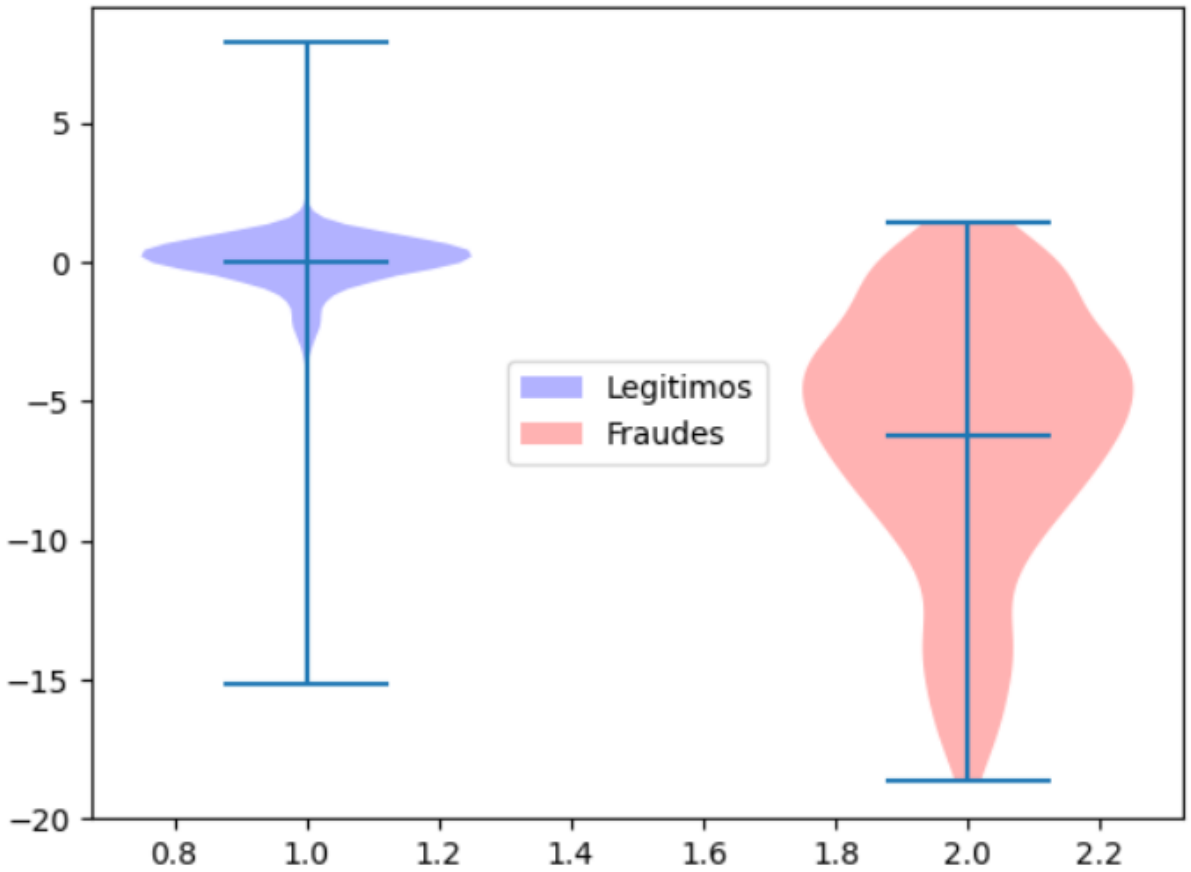

Semelhante ao atributo anterior.

- V11:



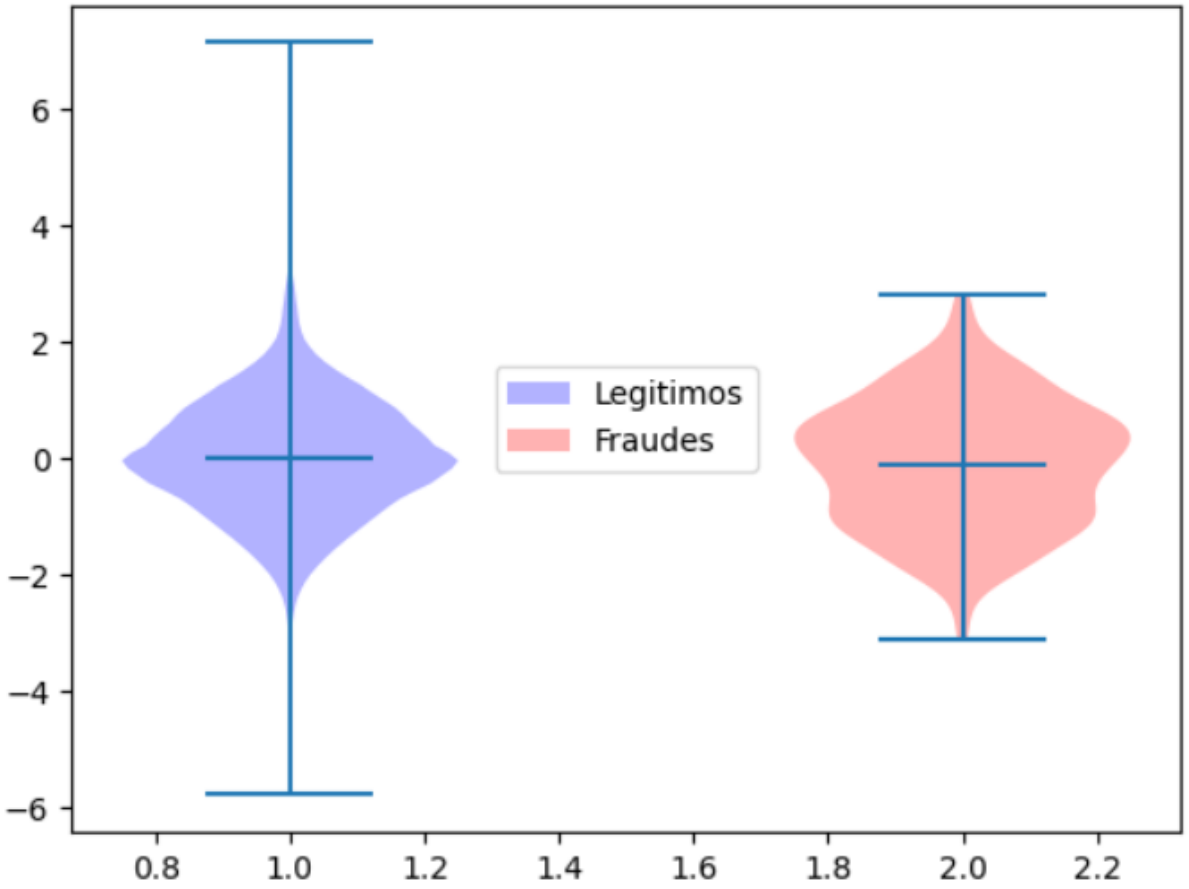

Bem discrepante e evidente a diferença de distribuição e de valor das medianas.

- V12:



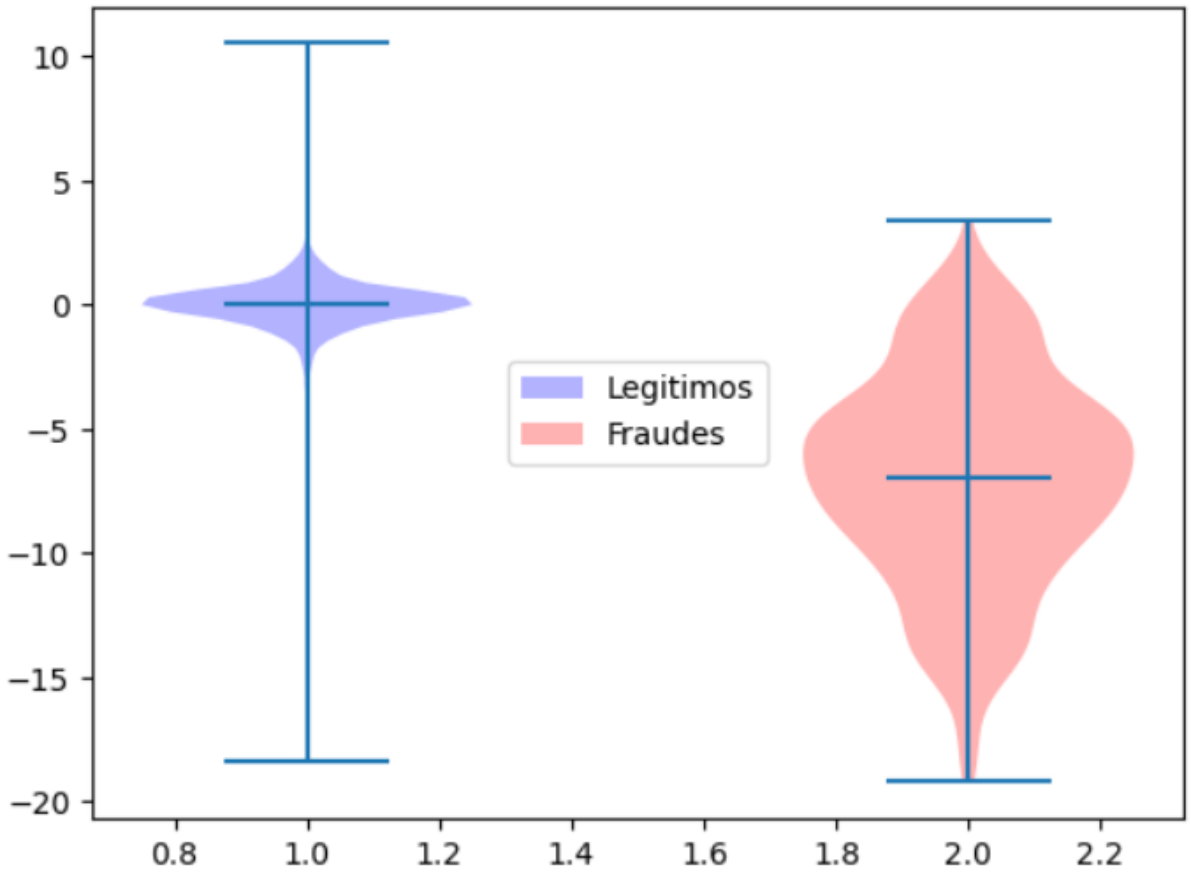

Semelhante ao atributo anterior e, desse modo, utilizamos.

- V13:



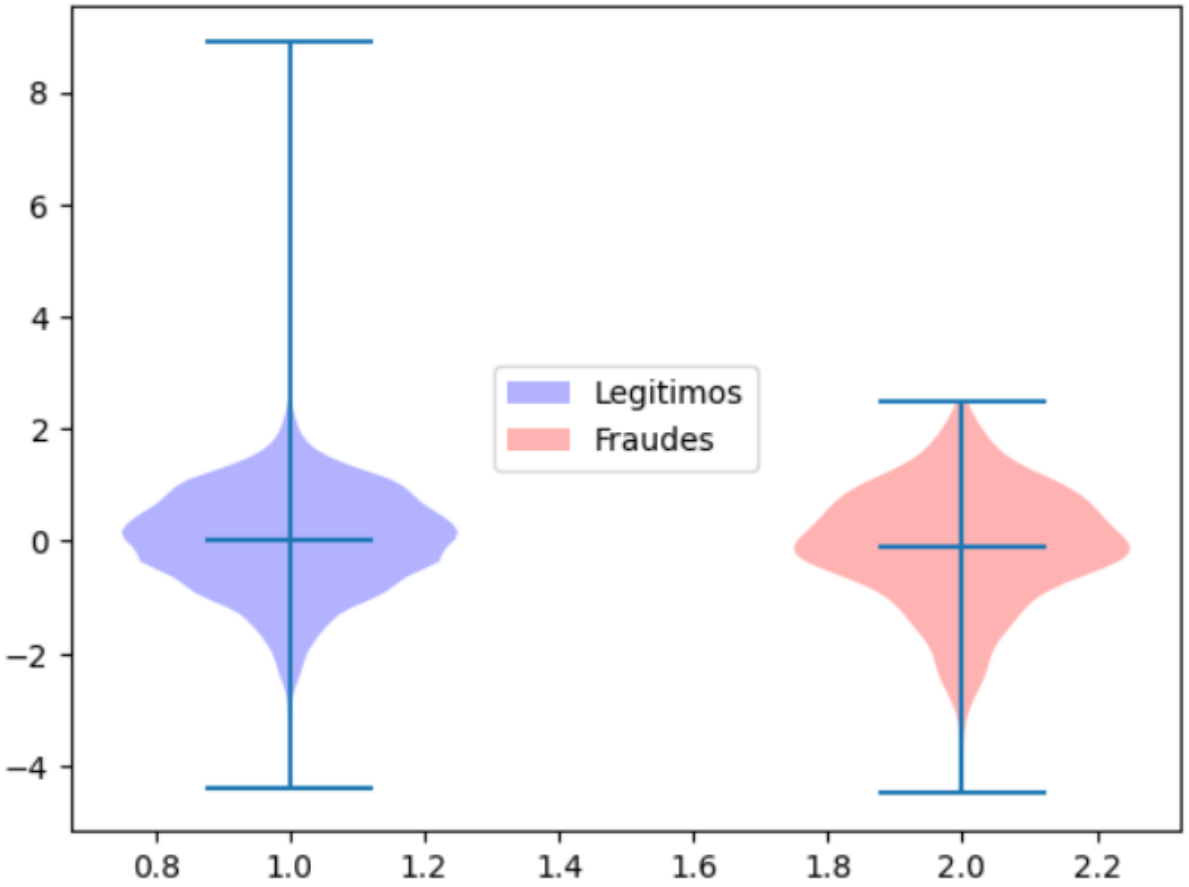

Distribuição parecida e mediana também. Dessa forma, não vamos usar.

- V14:



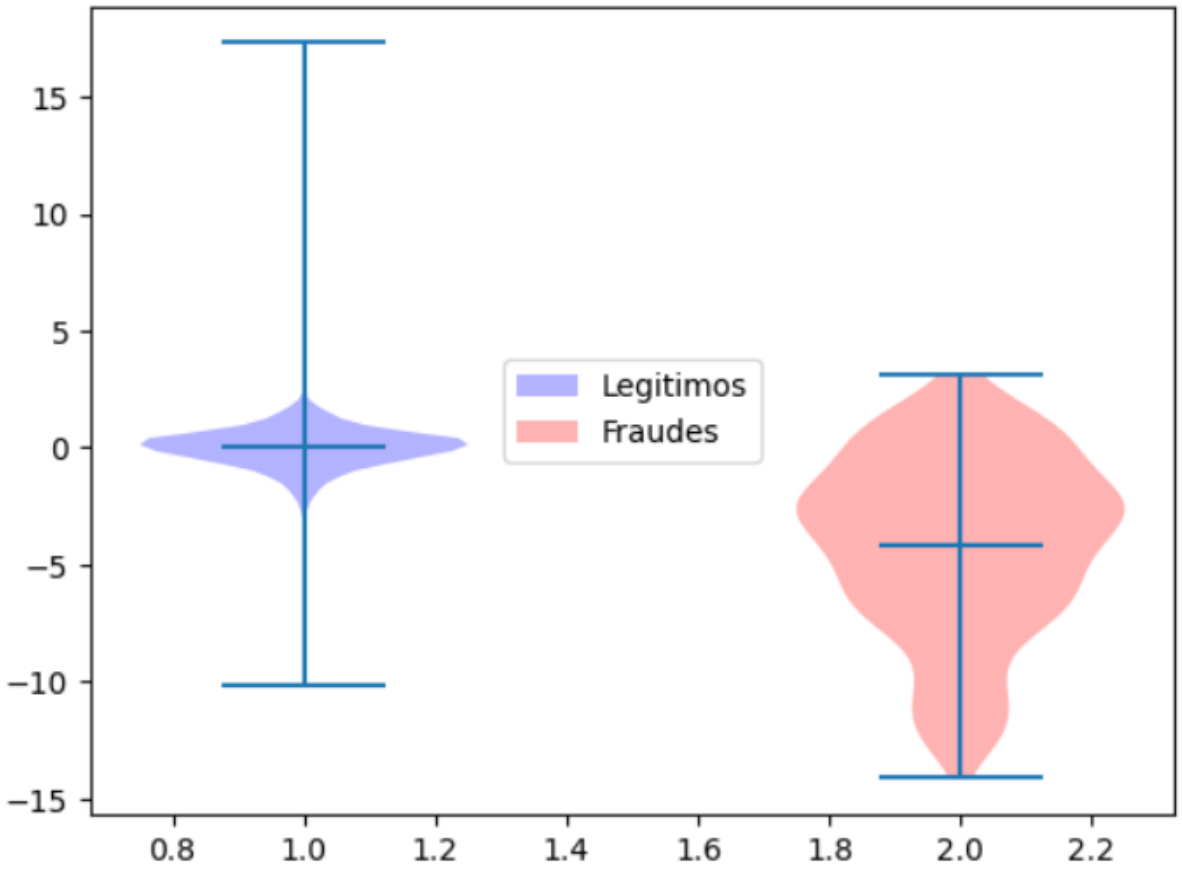

Parecido com os atributos anteriores aceitos para a diferenciação entre dados legítimos e fraudulentos.

- V15:



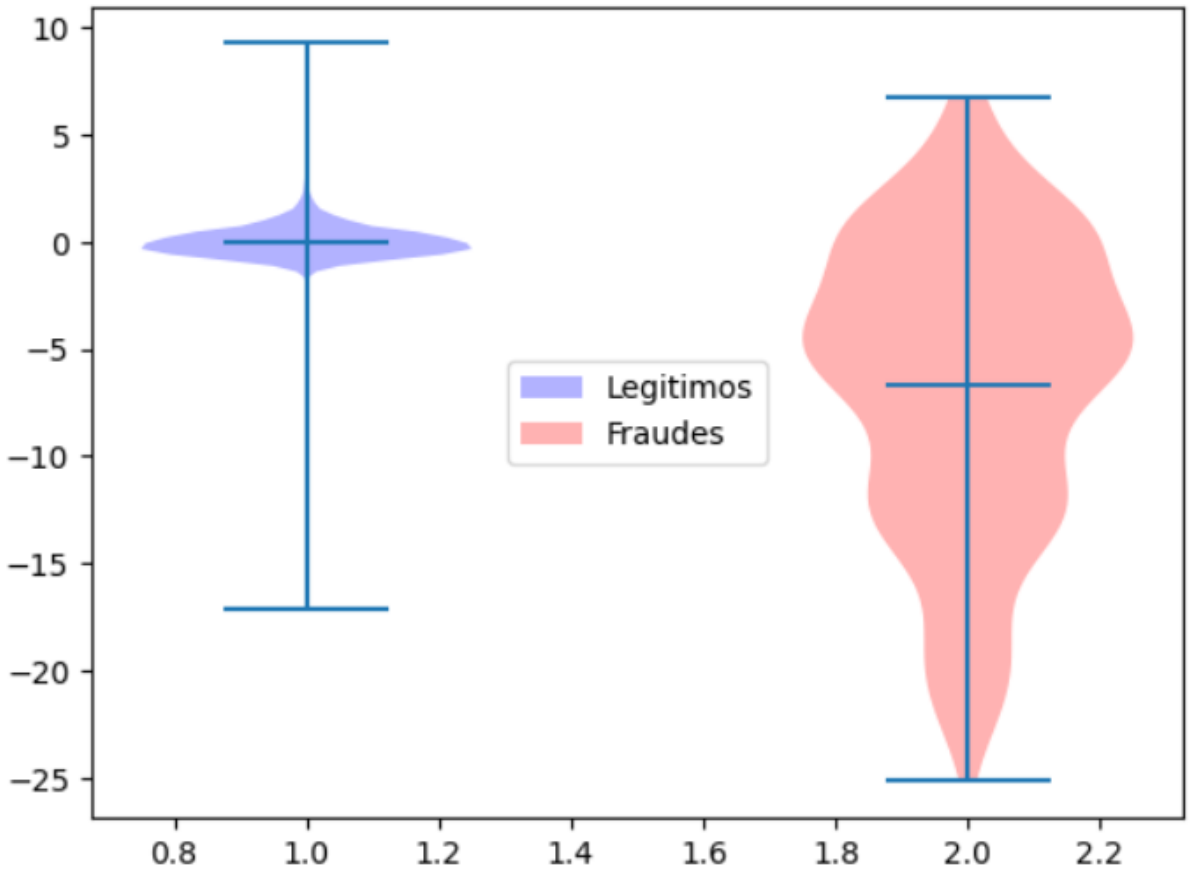

Não usado pelo fato de serem muito parecidos, exceto pelos outliers.

- V16:



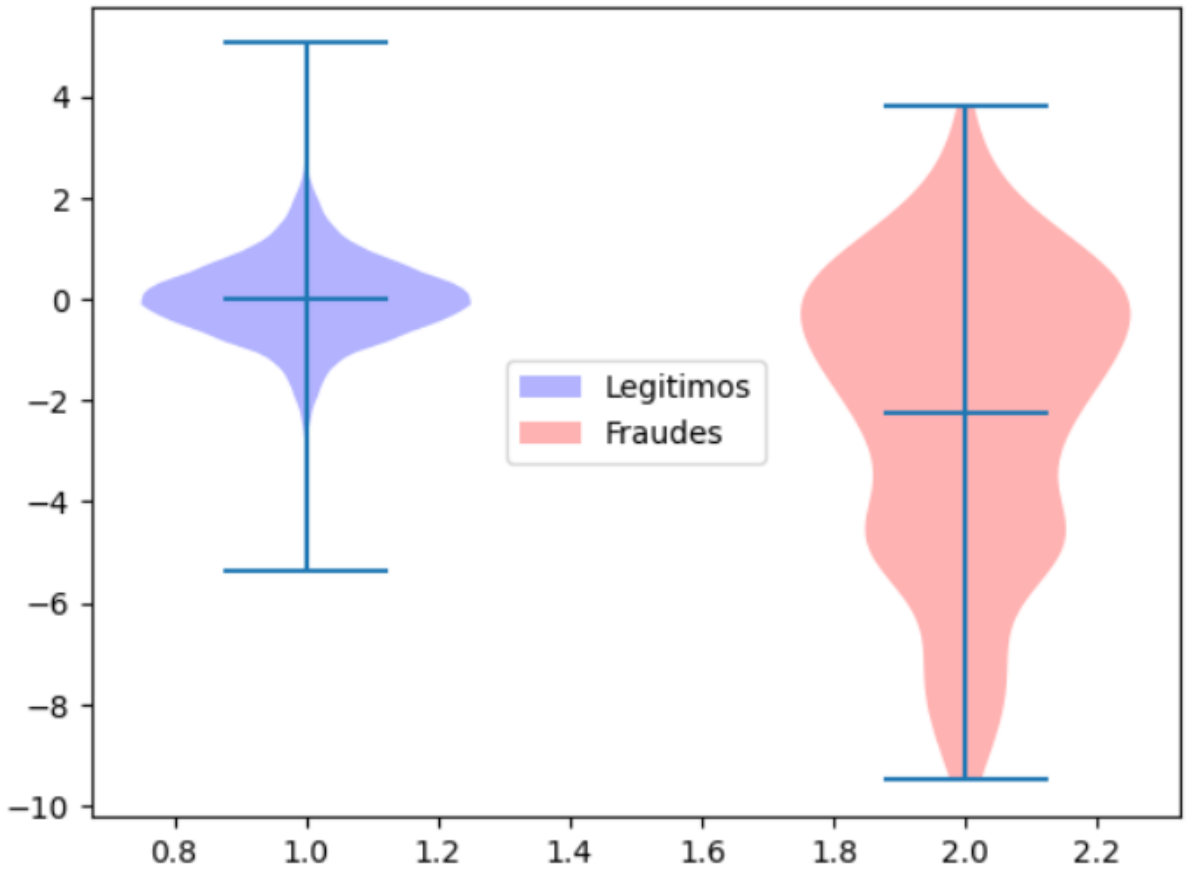

Distribuição bem diferente dos dados e valor discrepante das medianas. Assim, pode ser usado.

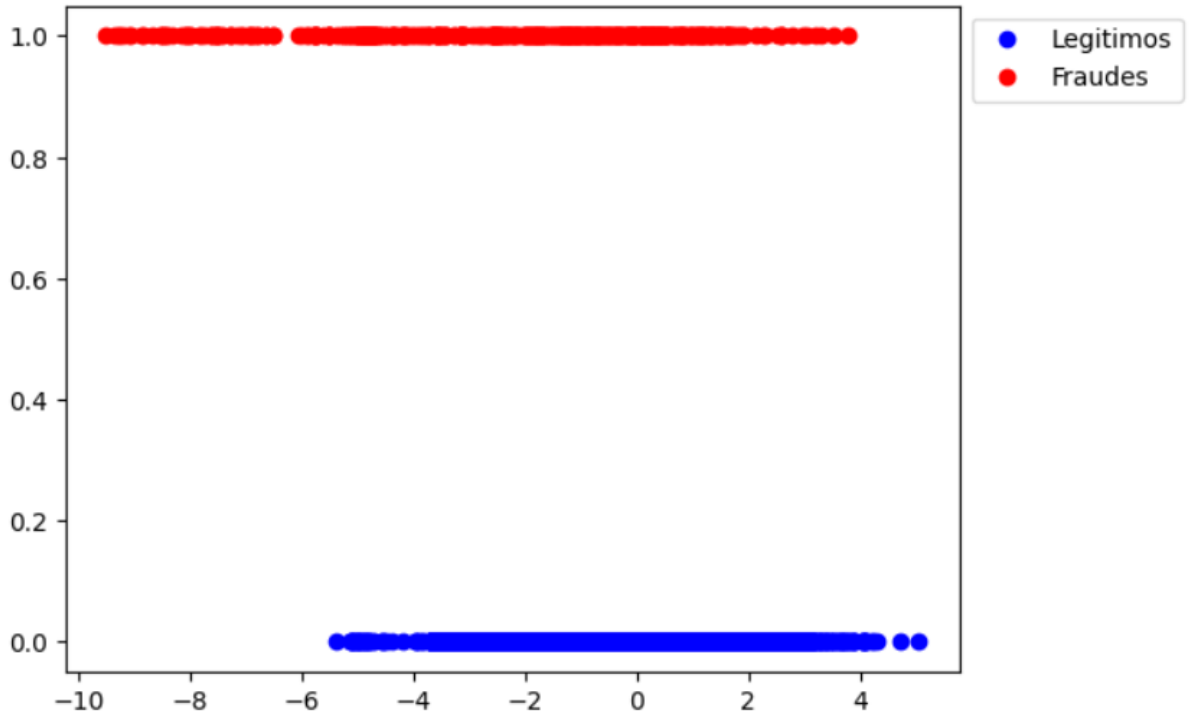
- V17:




Distribuição bem diferente entre os legítimos e fraudulentos, sem contar a mediana que também é.

- V18:

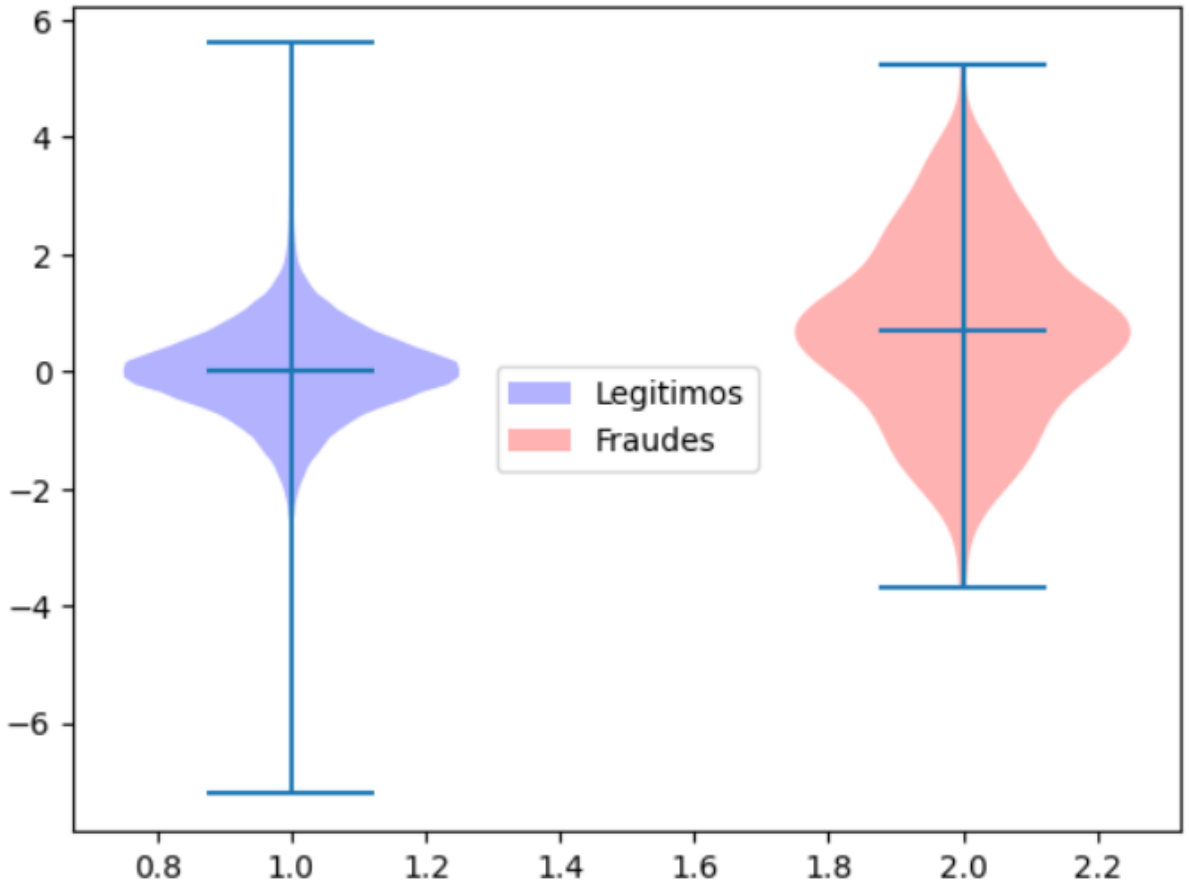



Ponto importante: os dados estão muito esparramados tanto para cima quanto para baixo, e a grande quantidade de dados fraudulentos está concentrada próximo à mediana

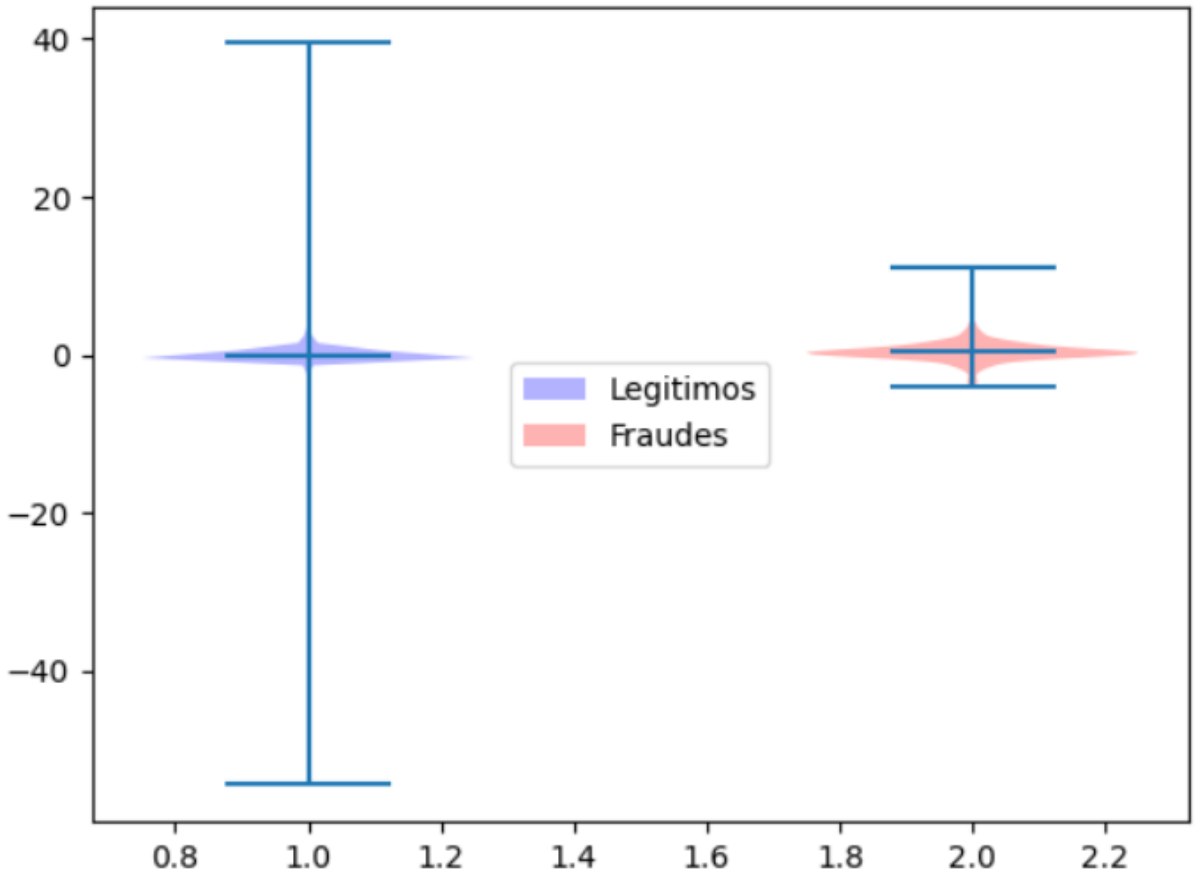
das legítimas, o que causa confusão na hora de decidir se é fraude ou não. Desse modo, não usamos.

- V19:



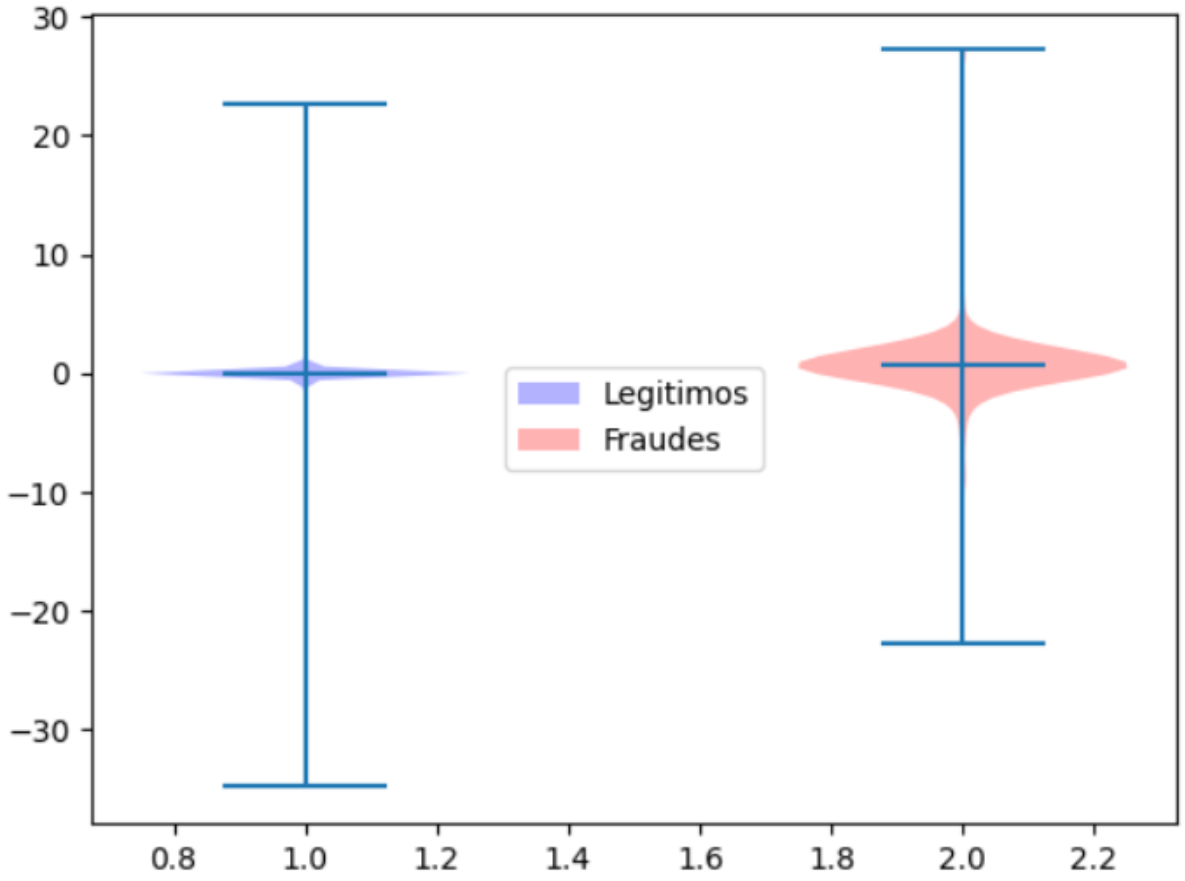

Outro atributo importante para diferenciar, já que as distribuições são diferentes mesmo que a mediana seja relativamente próxima.

- V20:



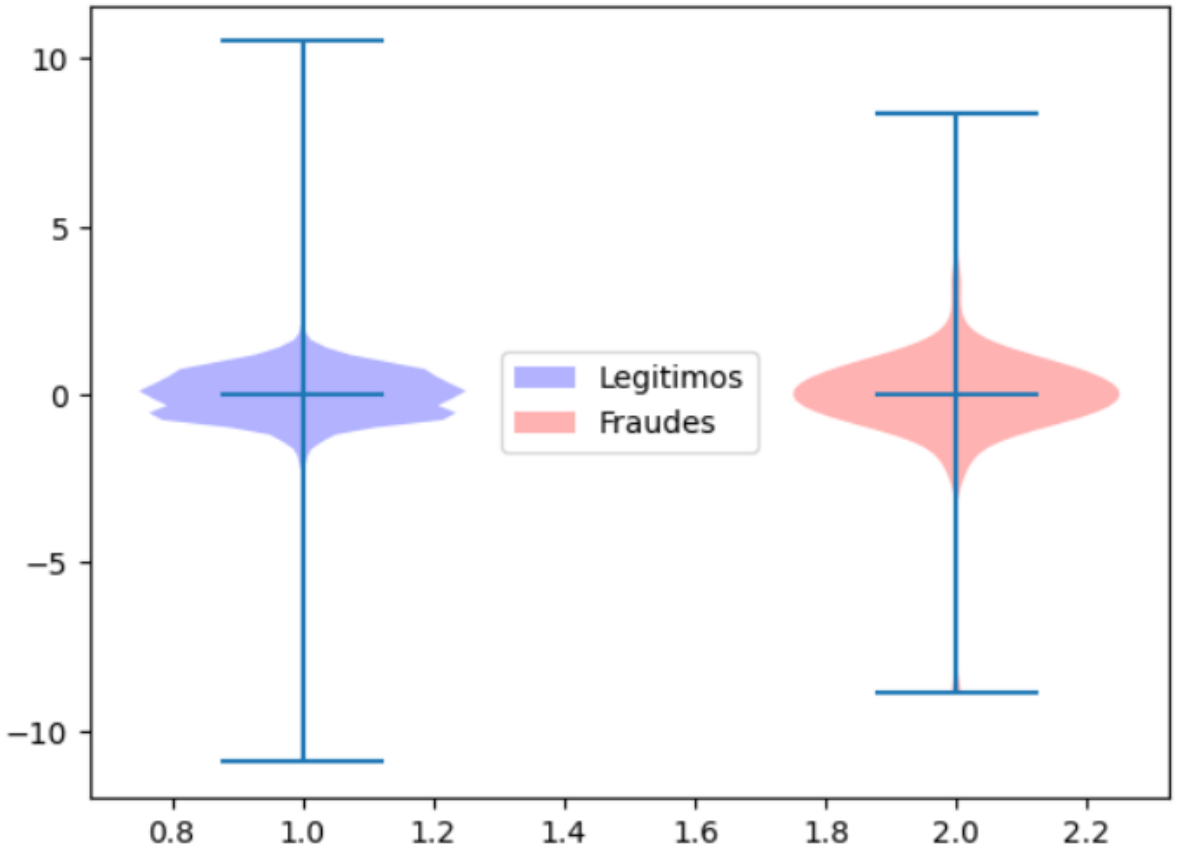

Distribuição parecida, muitos outliers nos dados legítimos e mediana parecida. Por esses motivos, não é um atributo utilizado.

- V21:



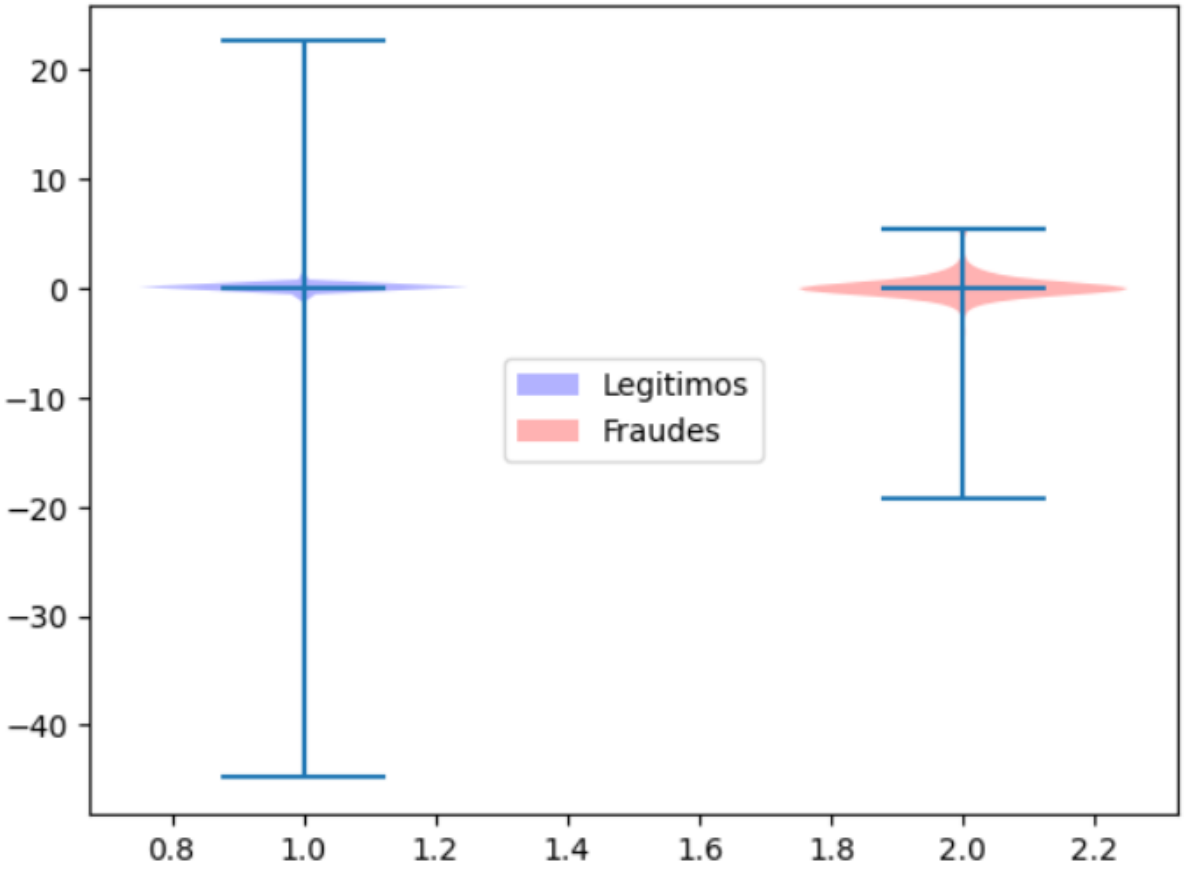

Semelhante ao atributo anterior.

- V22:



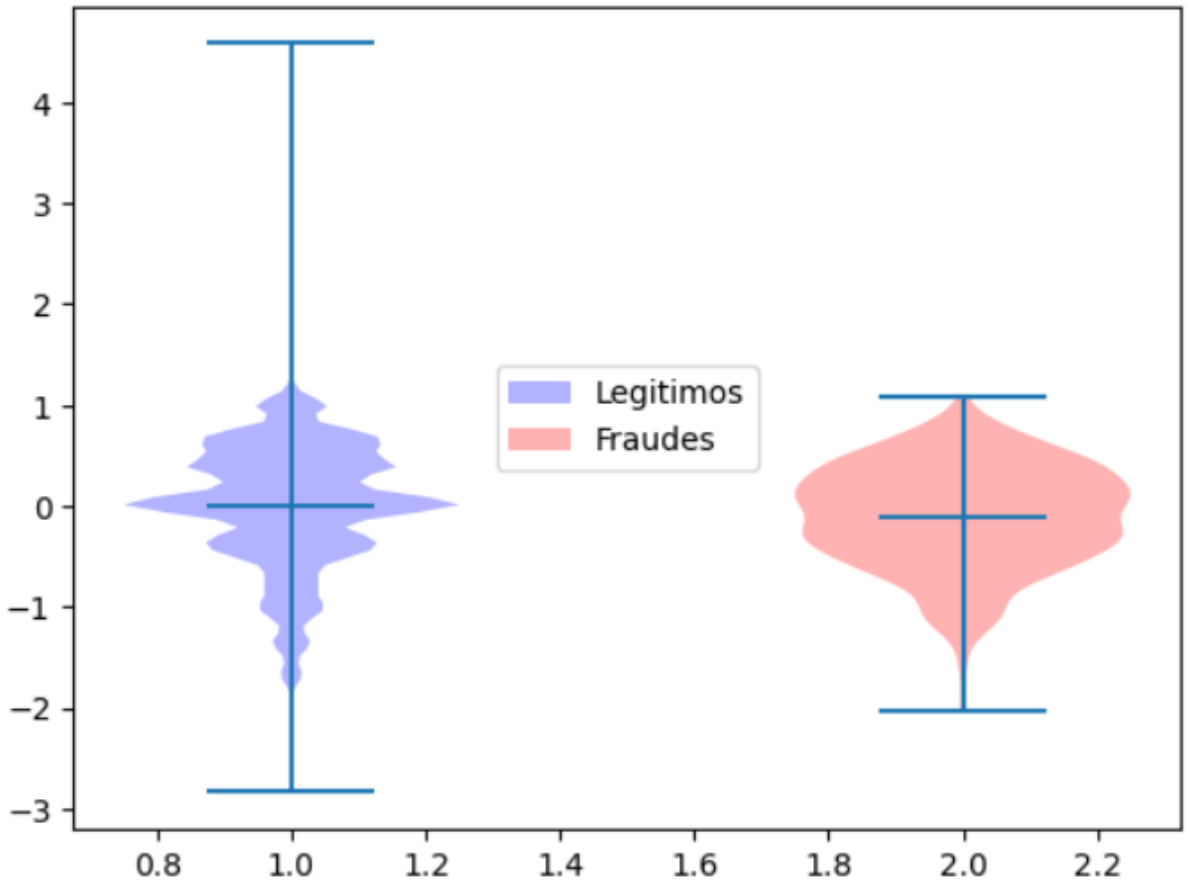

Idêntico aos 2 últimos atributos.

- V23:



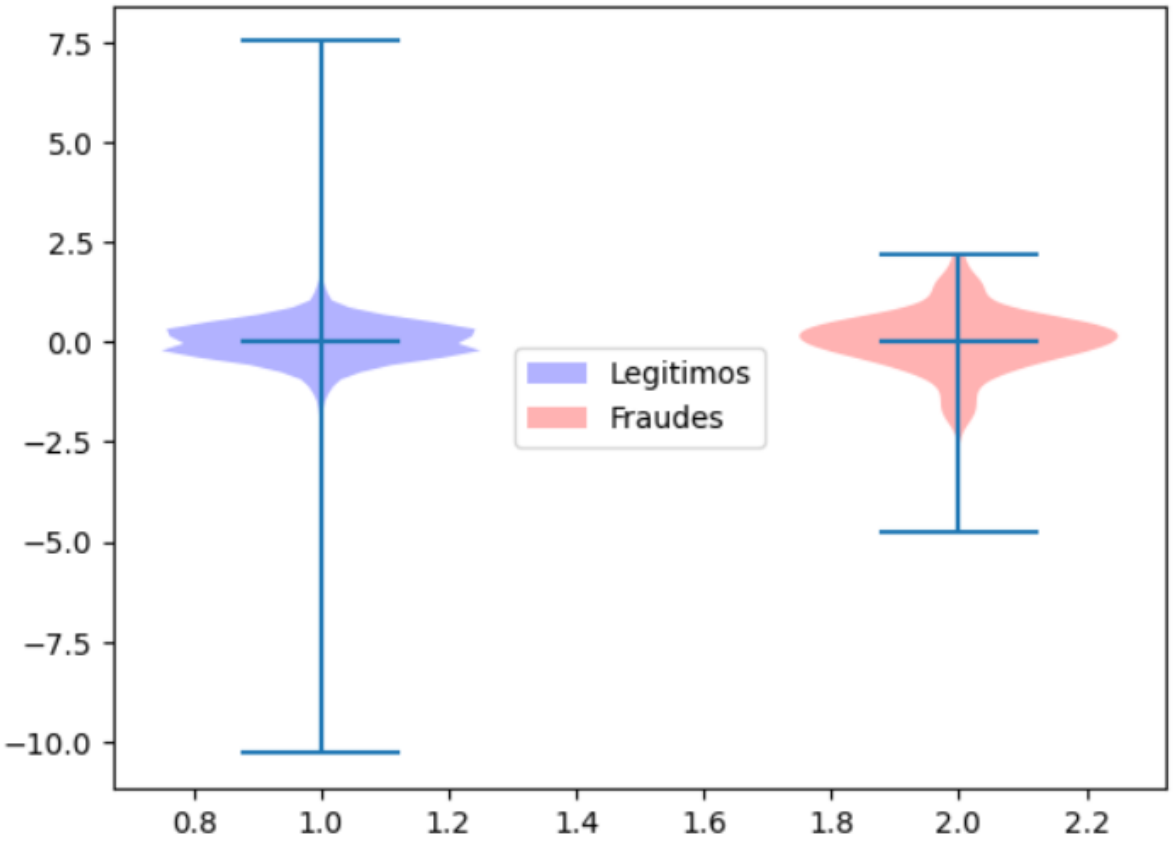

Semelhante aos últimos.

- V24:



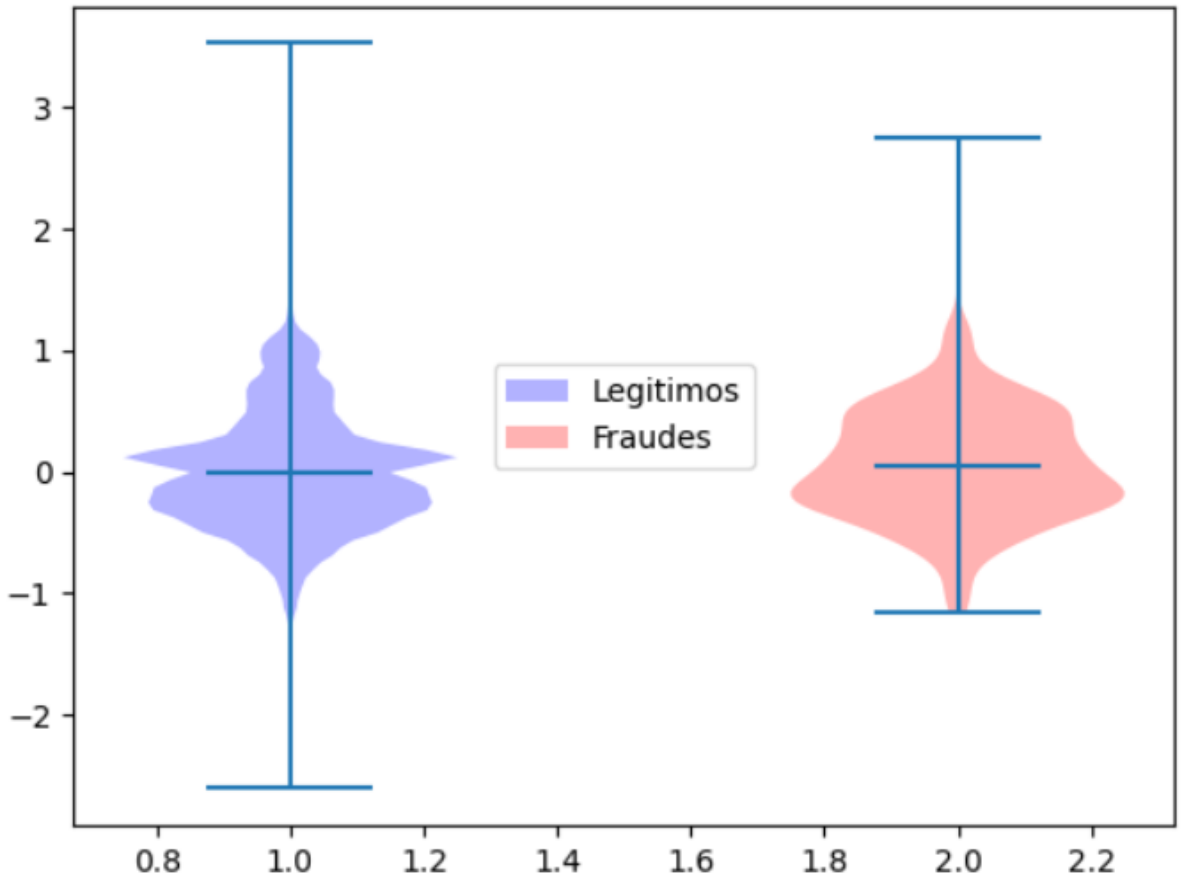

Motivo igual aos últimos atributos.

- V25:



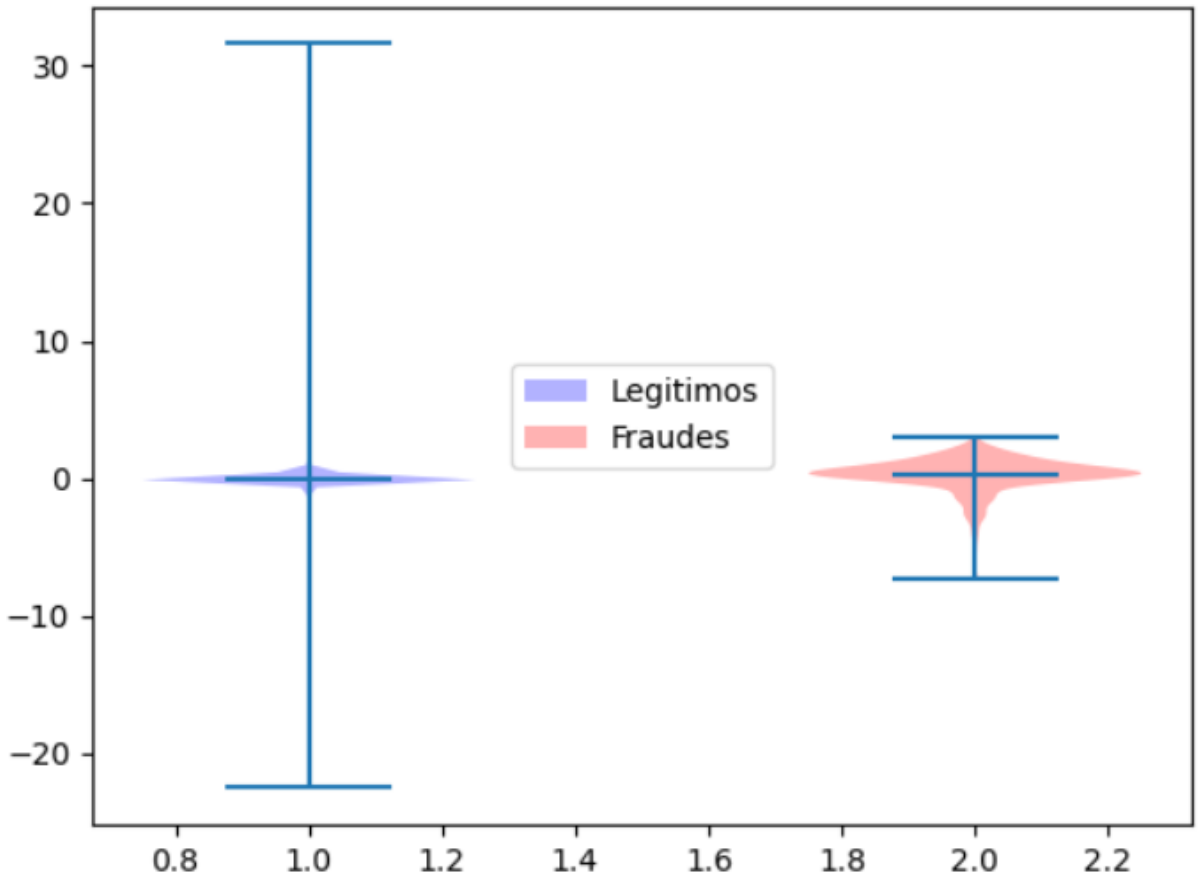

Também não utilizado pelas mesmas razões.

- V26:



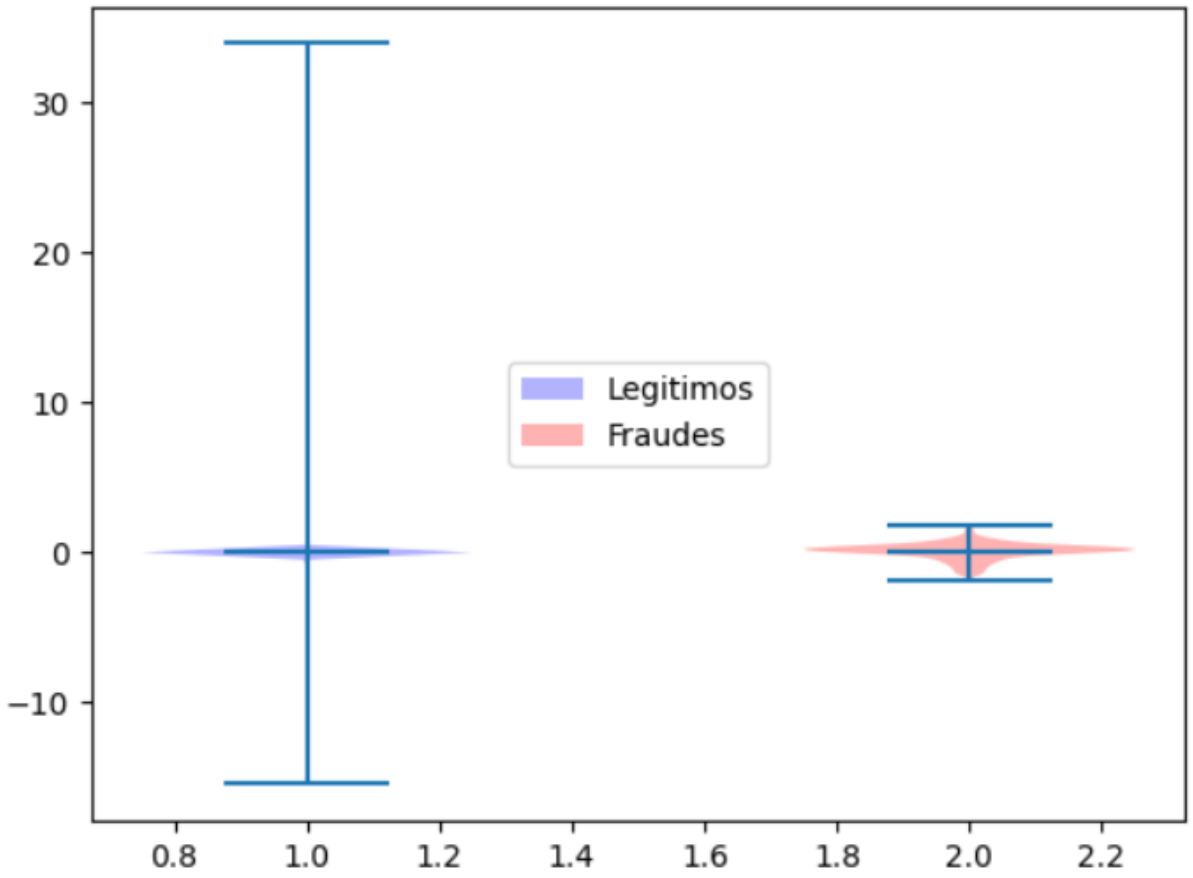

Mesma razão.

- V27:



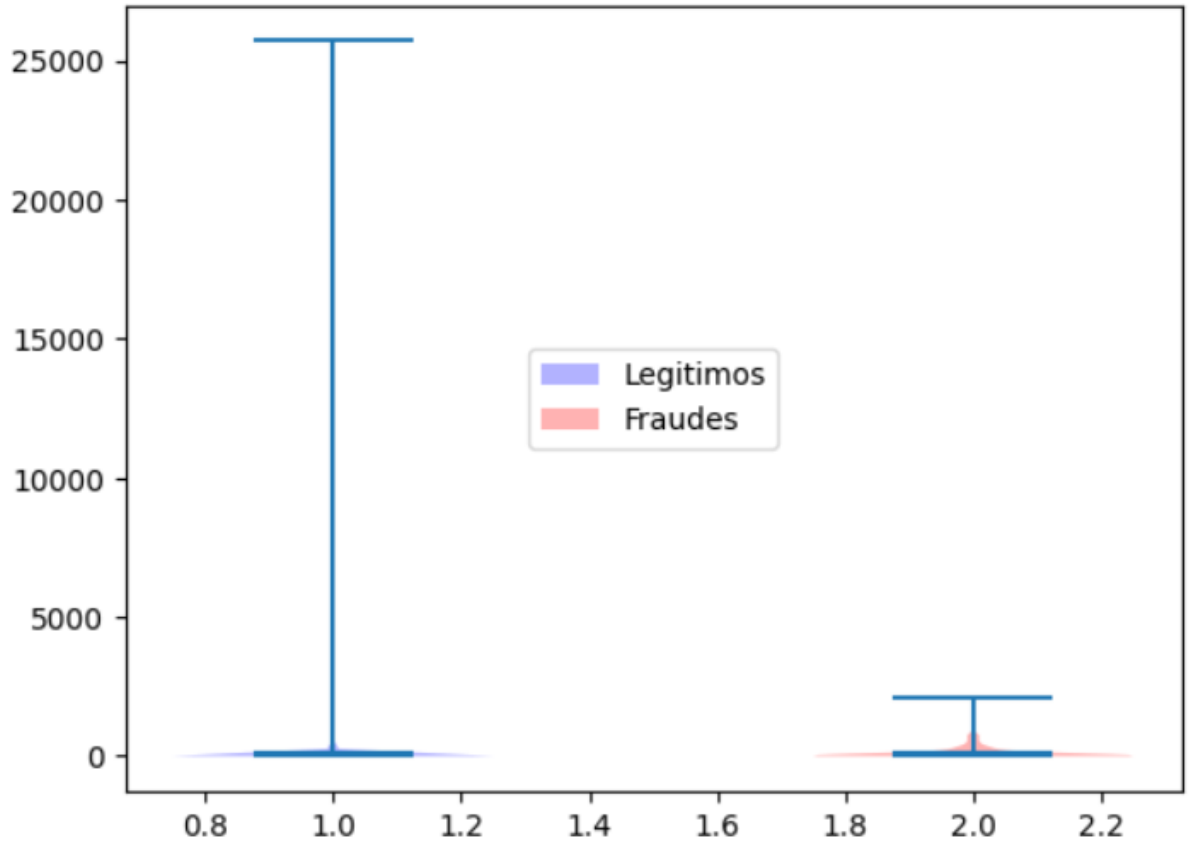

Distribuição próxima da mediana e esta muito parecida entre os dados legítimos e fraudulentos caracterizam um péssimo atributo para diferenciá-los.

- V28:




Mesmo motivo do atributo V27.

- Amount:




Bom atributo, mas para uma plotagem nem tanto, por causa dos valores serem muito altos.

Diante disso tudo, os atributos relevantes são: V3, V4, V5, V7, V9, V10, V11, V12, V14, V16, V17, V19 e Amount. Outro ponto importante a ser mostrado é o que foi pesquisado.

### Correlation Matrices

Correlation matrices are the essence of understanding our data. We want to know if there are features that influence heavily in whether a specific transaction is a fraud. However, it is important that we use the correct dataframe (subsample) in order for us to see which features have a high positive or negative correlation with regards to fraud transactions.

### Summary and Explanation:

- **Negative Correlations:** V17, V14, V12 and V10 are negatively correlated. Notice how the lower these values are, the more likely the end result will be a fraud transaction.
- **Positive Correlations:** V2, V4, V11, and V19 are positively correlated. Notice how the higher these values are, the more likely the end result will be a fraud transaction.
- **BoxPlots:** We will use boxplots to have a better understanding of the distribution of these features in fradulent and non fradulent transactions.

**Note:** We have to make sure we use the subsample in our correlation matrix or else our correlation matrix will be affected by the high imbalance between our classes. This occurs due to the high class imbalance in the original dataframe.