# Scikit-learn Experiment

Generated by: User

Generated on: 2021-10-08 16:28

[Scikit-learn Experiment 1](#_Toc38398026)

[Experiment Overview 1](#_Toc38398027)

[Data Overview 1](#_Toc32075692)

[Feature Importance 1](#_Toc38398030)

[Final Model 1](#_Toc38398031)

[Alternative Models 1](#_Toc38398032)

[Partial Dependence Plots 2](#_Toc38398033)

[Appendix 2](#_Toc38398035)

## Experiment Overview

Scikit-learn built a GradientBoostingRegressor to predict AGE given 27 original features from the input dataset. This is a regression experiment.

### Performance

|  |  |
| --- | --- |
| **Dataset** | **rmse** |
| Validation Data not Provided | NA |
| Test Data | 7.82 |

### System Specifications

|  |  |
| --- | --- |
| **CPU cores** | **Memory size** |
| 8 | 17.18 GB |

### Versions

|  |  |
| --- | --- |
| **Package** | **Version Number** |
| scikit-learn | 0.24.2 |

## Data Overview

This section provides information on the datasets used for the experiment.

|  |  |  |
| --- | --- | --- |
| **data** | **rows** | **cols** |
| train | 19,199 | 28 |
| test | 4,800 | 27 |

### Training Data

The training data consists of only numeric columns.

The summary of the columns is shown below:

#### Numeric Columns

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **name** | **min** | **mean** | **max** | **std** |
| LIMIT\_BAL | 1e+04 | 1.622e+05 | 1e+06 | 1.288e+05 |
| PAY\_0 | -2 | 0.03094 | 8 | 1.123 |
| PAY\_2 | -2 | -0.09662 | 8 | 1.203 |
| PAY\_3 | -2 | -0.1261 | 8 | 1.213 |
| PAY\_4 | -2 | -0.1907 | 8 | 1.167 |
| PAY\_5 | -2 | -0.2253 | 8 | 1.144 |
| PAY\_6 | -2 | -0.2536 | 8 | 1.17 |
| BILL\_AMT1 | -1.656e+05 | 5.008e+04 | 9.645e+05 | 7.138e+04 |
| BILL\_AMT2 | -3.335e+04 | 4.821e+04 | 9.839e+05 | 6.924e+04 |
| BILL\_AMT3 | -1.573e+05 | 4.574e+04 | 1.664e+06 | 6.706e+04 |
| BILL\_AMT4 | -1.7e+05 | 4.135e+04 | 8.916e+05 | 6.14e+04 |
| BILL\_AMT5 | -3.759e+04 | 3.952e+04 | 9.272e+05 | 5.918e+04 |
| BILL\_AMT6 | -3.396e+05 | 3.816e+04 | 9.617e+05 | 5.861e+04 |
| PAY\_AMT1 | 0 | 5516 | 5.05e+05 | 1.537e+04 |
| PAY\_AMT2 | 0 | 5693 | 1.684e+06 | 2.146e+04 |
| PAY\_AMT3 | 0 | 4667 | 8.96e+05 | 1.565e+04 |
| PAY\_AMT4 | 0 | 4711 | 4.97e+05 | 1.528e+04 |
| PAY\_AMT5 | 0 | 4733 | 4.18e+05 | 1.548e+04 |
| PAY\_AMT6 | 0 | 5073 | 5.287e+05 | 1.739e+04 |
| SEX.male | 0 | 0.402 | 1 | 0.4903 |
| EDUCATION.highschool | 0 | 0.1599 | 1 | 0.3665 |
| EDUCATION.other | 0 | 0.01182 | 1 | 0.1081 |
| EDUCATION.university | 0 | 0.4713 | 1 | 0.4992 |
| MARRIAGE.married | 0 | 0.4455 | 1 | 0.497 |
| MARRIAGE.other | 0 | 0.001615 | 1 | 0.04015 |
| MARRIAGE.single | 0 | 0.5409 | 1 | 0.4983 |
| DEFAULT\_PAYMENT\_NEXT\_MONTH | 0 | 0.2298 | 1 | 0.4207 |
| AGE | 21 | 35.26 | 79 | 9.317 |

#### Shifts Detected

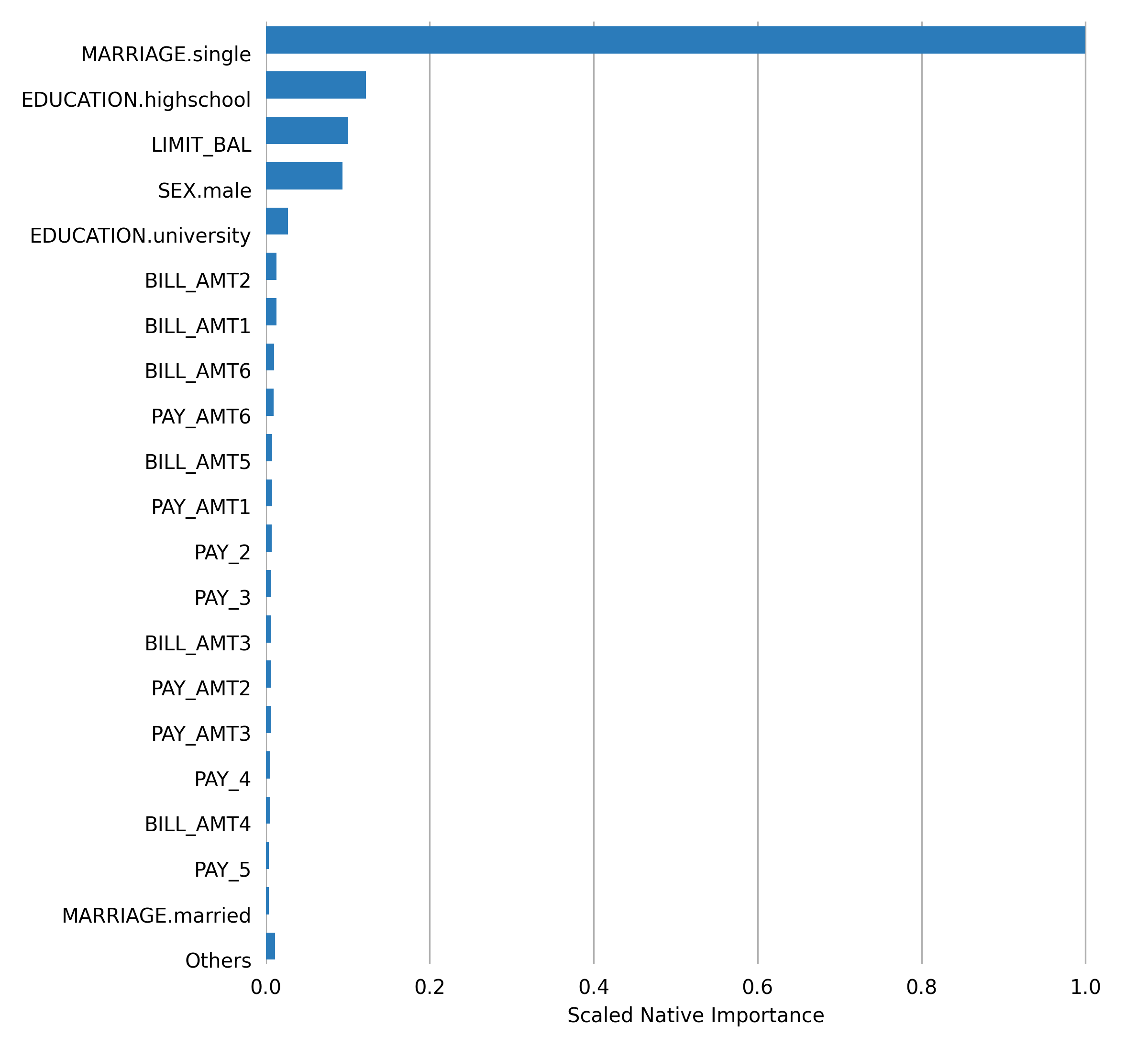
Scikit-learn can perform shift detection between the training, validation, and testing datasets. It does this by training a binomial model to predict which dataset a record belongs to. For example, it may find that it is able to separate the training and testing data with an AUC of 0.8 using only the column: C1 as the predictor. This indicates that there is some sort of drift in the distribution of C1 between the training and testing data.

For this experiment, Scikit-learn checked the train and test data for any shift in distributions but found none. This indicates that all the predictors/columns in the train and test data are from the same distribution.

## Feature Importance

The feature importance table shows the model-specific variable importance (Native Importance) and Scaled Native Importance (where values are scaled between 0 and 1).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Feature** | **Native Importance** | **Scaled Native Importance** |
| 0 | MARRIAGE.single | 0.6826 | 1.0 |
| 1 | EDUCATION.highschool | 0.0834 | 0.1222 |
| 2 | LIMIT\_BAL | 0.0681 | 0.0998 |
| 3 | SEX.male | 0.0637 | 0.0933 |
| 4 | EDUCATION.university | 0.0185 | 0.0272 |
| 5 | BILL\_AMT2 | 0.009 | 0.0131 |
| 6 | BILL\_AMT1 | 0.0088 | 0.0128 |
| 7 | BILL\_AMT6 | 0.0068 | 0.0099 |
| 8 | PAY\_AMT6 | 0.0065 | 0.0096 |
| 9 | BILL\_AMT5 | 0.0055 | 0.008 |
| 10 | PAY\_AMT1 | 0.0053 | 0.0077 |
| 11 | PAY\_2 | 0.005 | 0.0074 |
| 12 | PAY\_3 | 0.0046 | 0.0068 |
| 13 | BILL\_AMT3 | 0.0045 | 0.0066 |
| 14 | PAY\_AMT2 | 0.0043 | 0.0063 |
| 15 | PAY\_AMT3 | 0.0039 | 0.0058 |
| 16 | PAY\_4 | 0.0037 | 0.0054 |
| 17 | BILL\_AMT4 | 0.0037 | 0.0054 |
| 18 | PAY\_5 | 0.0023 | 0.0033 |
| 19 | MARRIAGE.married | 0.002 | 0.003 |
| 20 | Others | 0.0078 | 0.0115 |



## Final Model

|  |  |
| --- | --- |
| **Model Attribute** | **Value** |
| Type | GradientBoostingRegressor |
| n\_estimators | 100 |
| learning\_rate | 0.1 |
| loss | ls |
| criterion | friedman\_mse |
| min\_samples\_split | 2 |
| min\_samples\_leaf | 1 |
| min\_weight\_fraction\_leaf | 0.0 |
| subsample | 1.0 |
| max\_features | None |
| max\_depth | 3 |
| min\_impurity\_decrease | 0.0 |
| min\_impurity\_split | None |
| ccp\_alpha | 0.0 |
| init | None |
| random\_state | None |
| alpha | 0.9 |
| verbose | 0 |
| max\_leaf\_nodes | None |
| warm\_start | False |
| validation\_fraction | 0.1 |
| n\_iter\_no\_change | None |
| tol | 0.0001 |

**Performance of Final Model**

|  |  |  |
| --- | --- | --- |
| **scorer** | **training** | **test** |
| RMSE | 7.6413 | 7.8196 |
| MSE | 58.3902 | 61.1456 |
| MAE | 5.9058 | 6.007 |
| RMSLE | 0.201 | 0.2035 |

**Population Stability Index (PSI)**

Population Stability Index is a statistic used to describe a variable’s distribution shift. It can measure the shift between the training dataset’s model score distribution and any other given dataset (i.e. validation or test dataset).

A PSI value lower than 0.10 indicates a small shift in the model predictions, a value between 0.10 and 0.25 indicates a moderate shift, and a value greater than 0.25 indicates a strong shift. Strong shift values can indicate that the model trained on the training dataset might not be suitable for the provided validation or test datasets.

**Summary PSI table**

|  |  |
| --- | --- |
| **Dataset** | **PSI** |
| Test | 0.0293 |

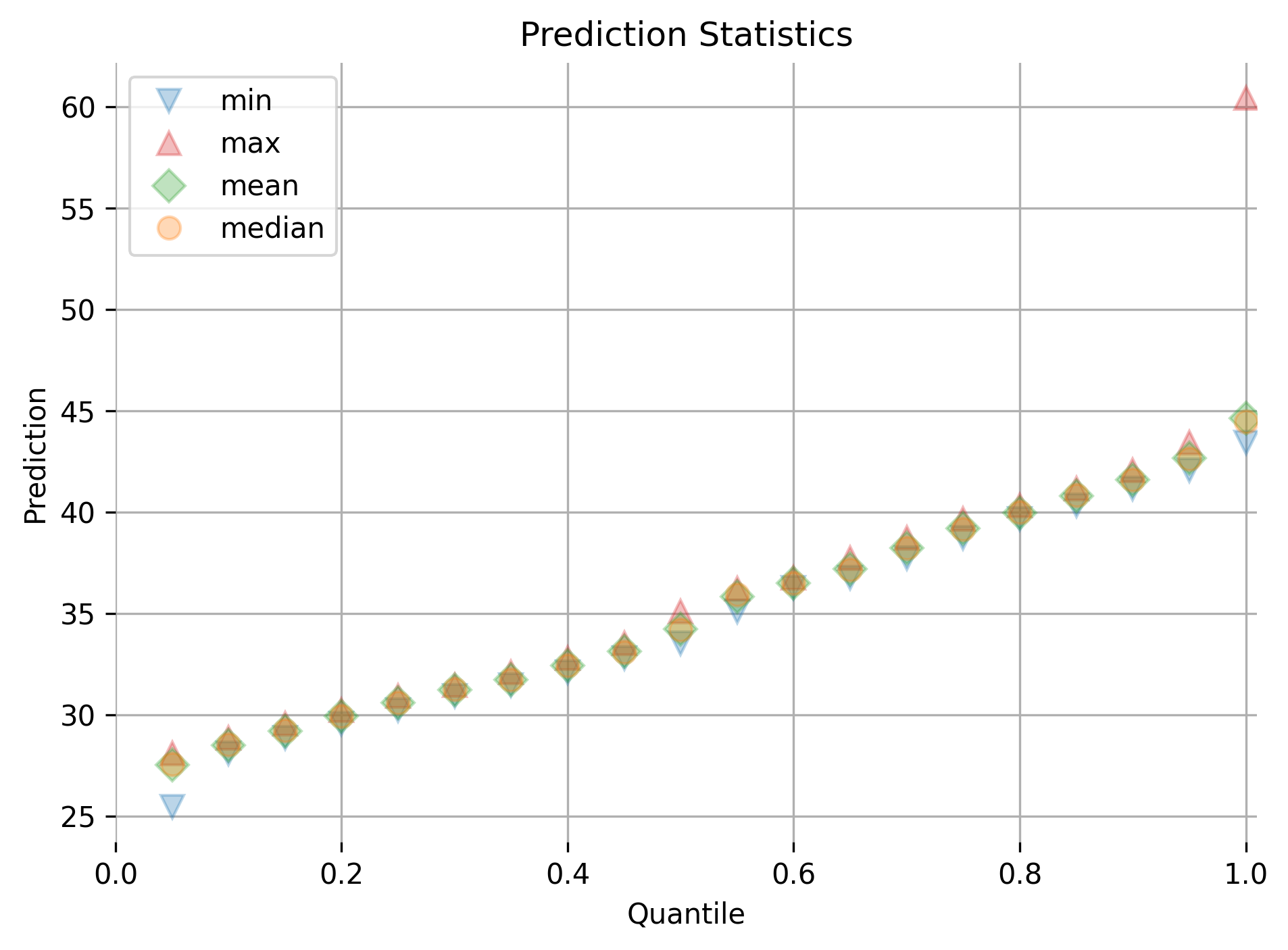
Details on the PSI calculations can be found in the Appendix.

**Prediction statistics**

The following tables and plots show the min, max, mean, and median quantile prediction values for each dataset split. Note: values are rounded to the fourth decimal place. For example, .000025 and .000010 would both appear as 0.0.

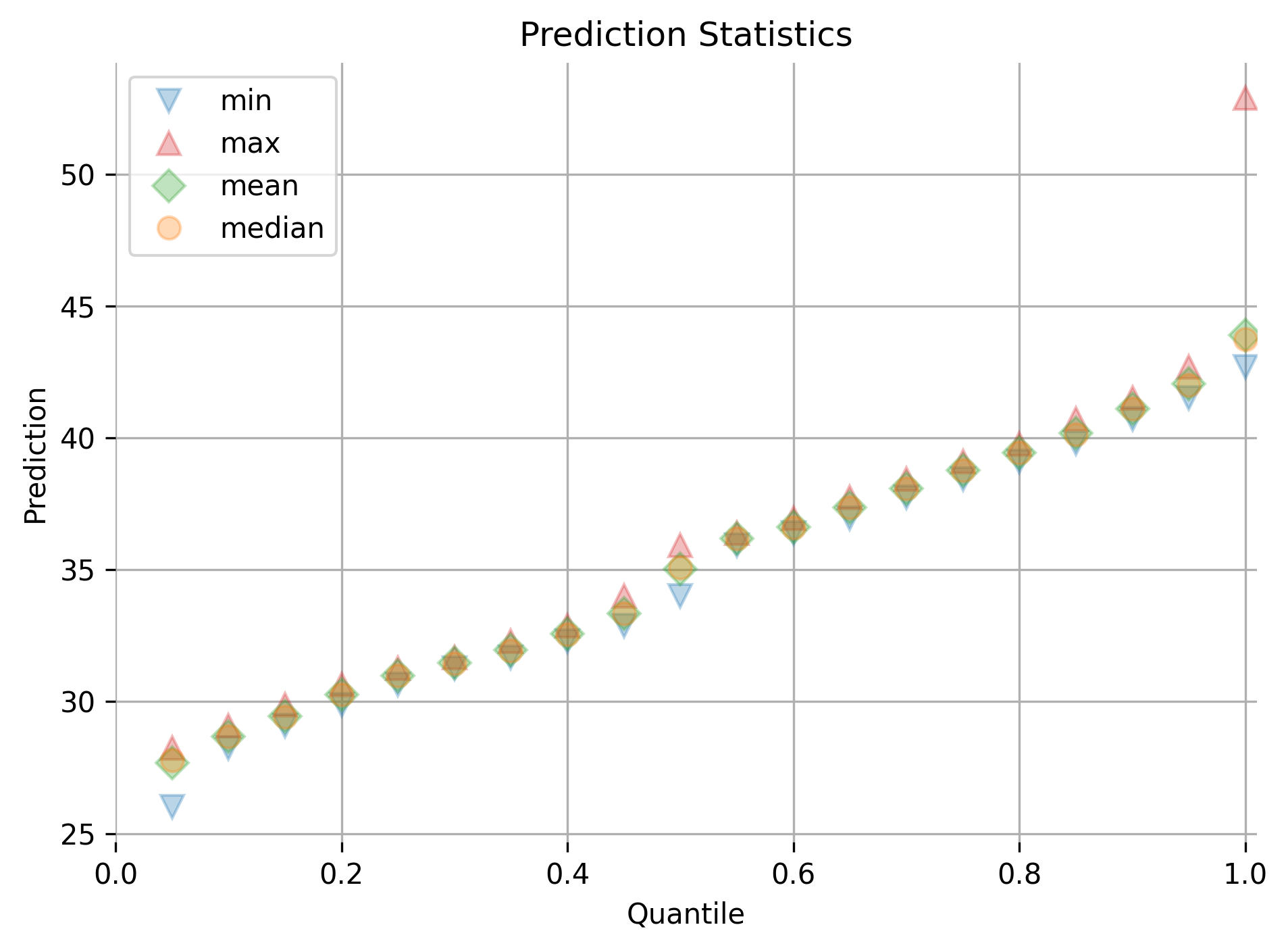
**Train**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **quantile** | **min** | **max** | **mean** | **median** |
| 0.05 | 25.4645 | 28.0871 | 27.521 | 27.5428 |
| 0.1 | 28.0895 | 28.8521 | 28.4828 | 28.5239 |
| 0.15 | 28.8534 | 29.5616 | 29.2058 | 29.2172 |
| 0.2 | 29.5622 | 30.2372 | 29.9133 | 29.9255 |
| 0.25 | 30.2382 | 30.9219 | 30.584 | 30.5998 |
| 0.3 | 30.9221 | 31.4632 | 31.2225 | 31.2456 |
| 0.35 | 31.4641 | 32.0995 | 31.7376 | 31.7154 |
| 0.4 | 32.0996 | 32.7807 | 32.4264 | 32.4184 |
| 0.45 | 32.7827 | 33.5275 | 33.1155 | 33.0928 |
| 0.5 | 33.5284 | 35.1017 | 34.2191 | 34.1914 |
| 0.55 | 35.1026 | 36.2447 | 35.8287 | 35.9371 |
| 0.6 | 36.2476 | 36.7748 | 36.4916 | 36.4989 |
| 0.65 | 36.7756 | 37.7317 | 37.2127 | 37.1775 |
| 0.7 | 37.7333 | 38.738 | 38.229 | 38.2261 |
| 0.75 | 38.7381 | 39.658 | 39.1918 | 39.1784 |
| 0.8 | 39.6588 | 40.3288 | 39.9749 | 39.9869 |
| 0.85 | 40.3307 | 41.1824 | 40.809 | 40.8478 |
| 0.9 | 41.1827 | 42.0538 | 41.593 | 41.6061 |
| 0.95 | 42.0579 | 43.4212 | 42.6564 | 42.6373 |
| 1.0 | 43.4214 | 60.4341 | 44.6323 | 44.4596 |



**Test**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **quantile** | **min** | **max** | **mean** | **median** |
| 0.05 | 26.0187 | 28.2306 | 27.6776 | 27.7678 |
| 0.1 | 28.2313 | 29.0797 | 28.6805 | 28.6817 |
| 0.15 | 29.0813 | 29.881 | 29.447 | 29.4304 |
| 0.2 | 29.893 | 30.6405 | 30.2633 | 30.2589 |
| 0.25 | 30.6438 | 31.2735 | 30.9868 | 30.9793 |
| 0.3 | 31.2771 | 31.6664 | 31.4623 | 31.4487 |
| 0.35 | 31.6673 | 32.292 | 31.9603 | 31.9397 |
| 0.4 | 32.2954 | 32.8767 | 32.5806 | 32.5531 |
| 0.45 | 32.8787 | 33.9974 | 33.3479 | 33.3291 |
| 0.5 | 34.0039 | 35.916 | 35.0394 | 35.0872 |
| 0.55 | 35.9198 | 36.3907 | 36.1738 | 36.1795 |
| 0.6 | 36.3967 | 36.9454 | 36.6195 | 36.5847 |
| 0.65 | 36.9547 | 37.7545 | 37.3707 | 37.37 |
| 0.7 | 37.7557 | 38.446 | 38.091 | 38.0832 |
| 0.75 | 38.4493 | 39.1099 | 38.7655 | 38.7718 |
| 0.8 | 39.1107 | 39.7744 | 39.4416 | 39.4261 |
| 0.85 | 39.7845 | 40.7073 | 40.1818 | 40.1244 |
| 0.9 | 40.7203 | 41.5063 | 41.1139 | 41.1047 |
| 0.95 | 41.5071 | 42.6889 | 42.0402 | 42.0058 |
| 1.0 | 42.6904 | 52.8778 | 43.8898 | 43.7504 |



## Alternative Models

Alternative models were not provided by the user.

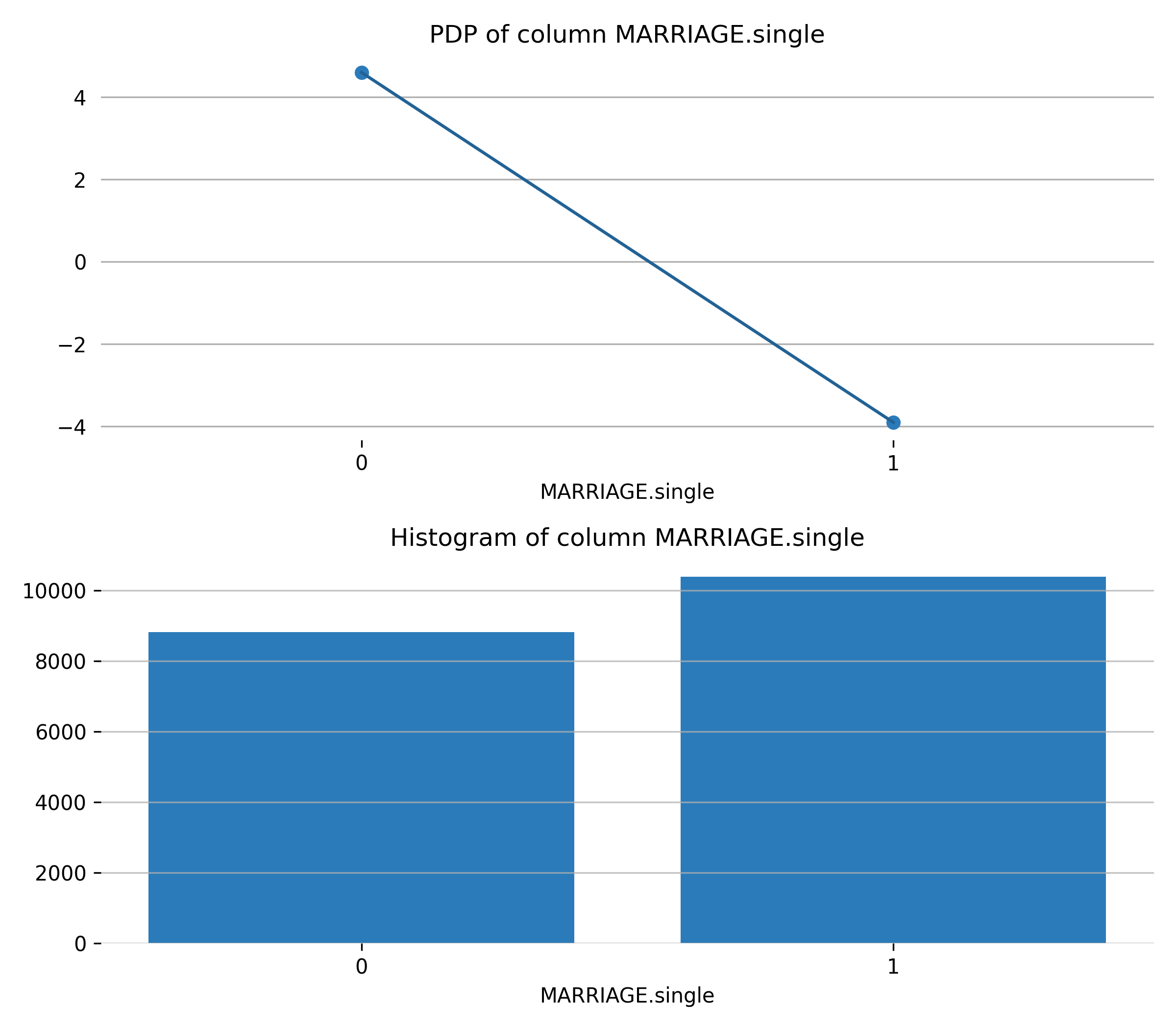
## Partial Dependence Plots

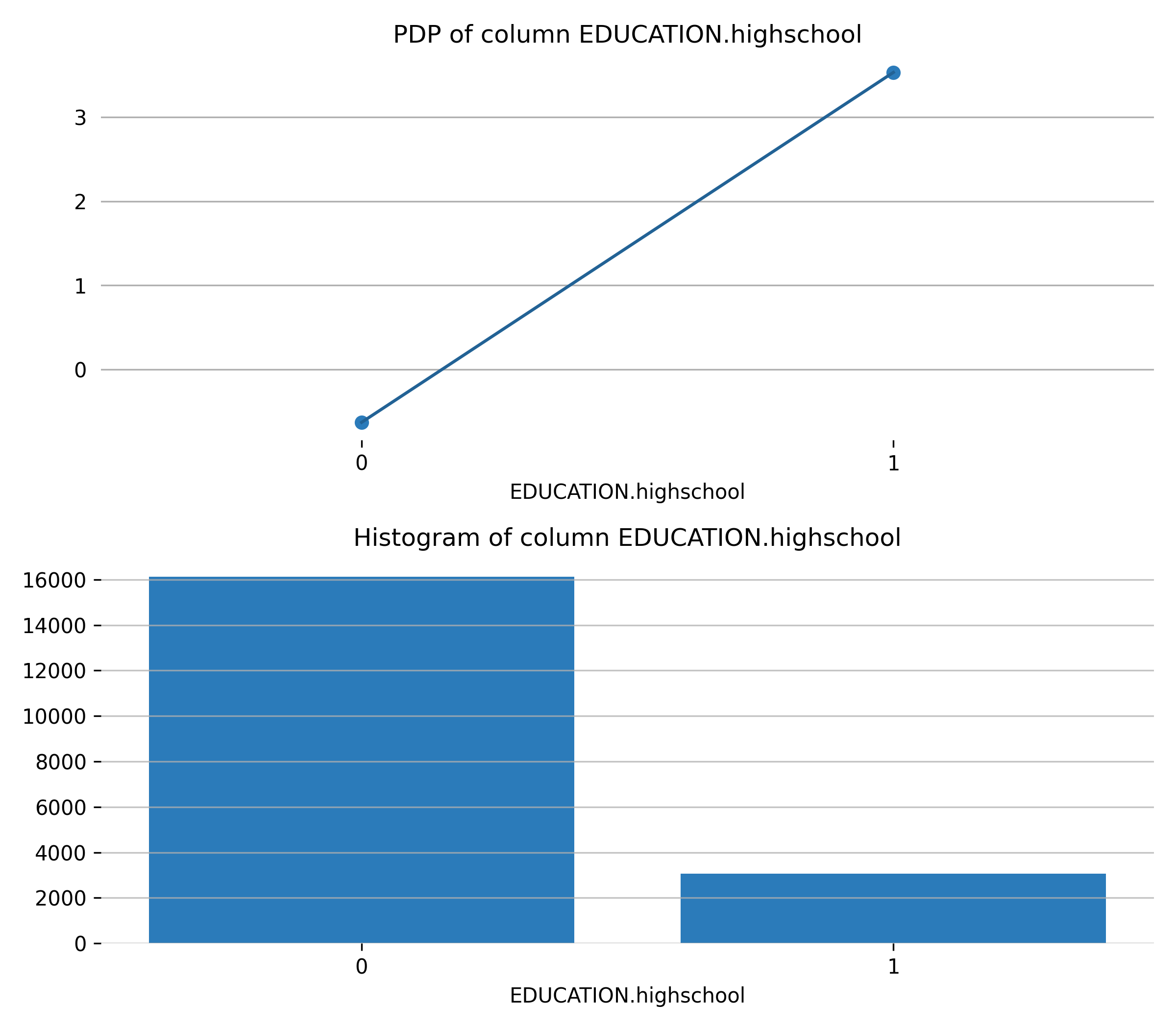
Partial dependence plots show the partial dependence as a function of specific values for a feature subset. The plots show how machine-learned response functions change based on the values of an input feature of interest, while taking nonlinearity into consideration and averaging out the effects of all other input features. Partial dependence plots enable increased transparency in a model and enable the ability to validate and debug a model by comparing a feature’s average predictions across its domain to known standards and reasonable expectations.

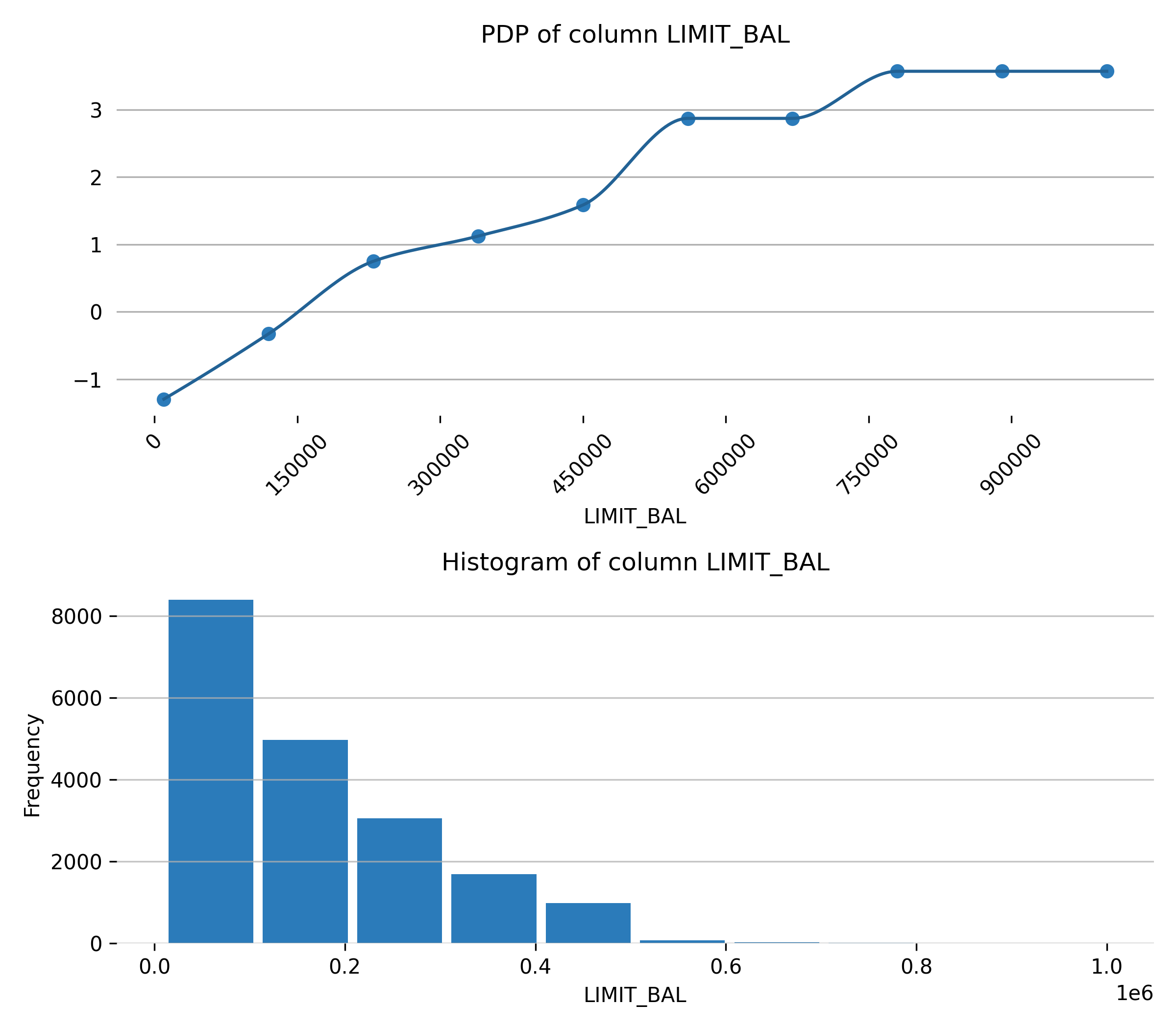
The partial dependence plots are shown for the top 20 original variables. The top 20 original variables are chosen based on their model specific variable importance.

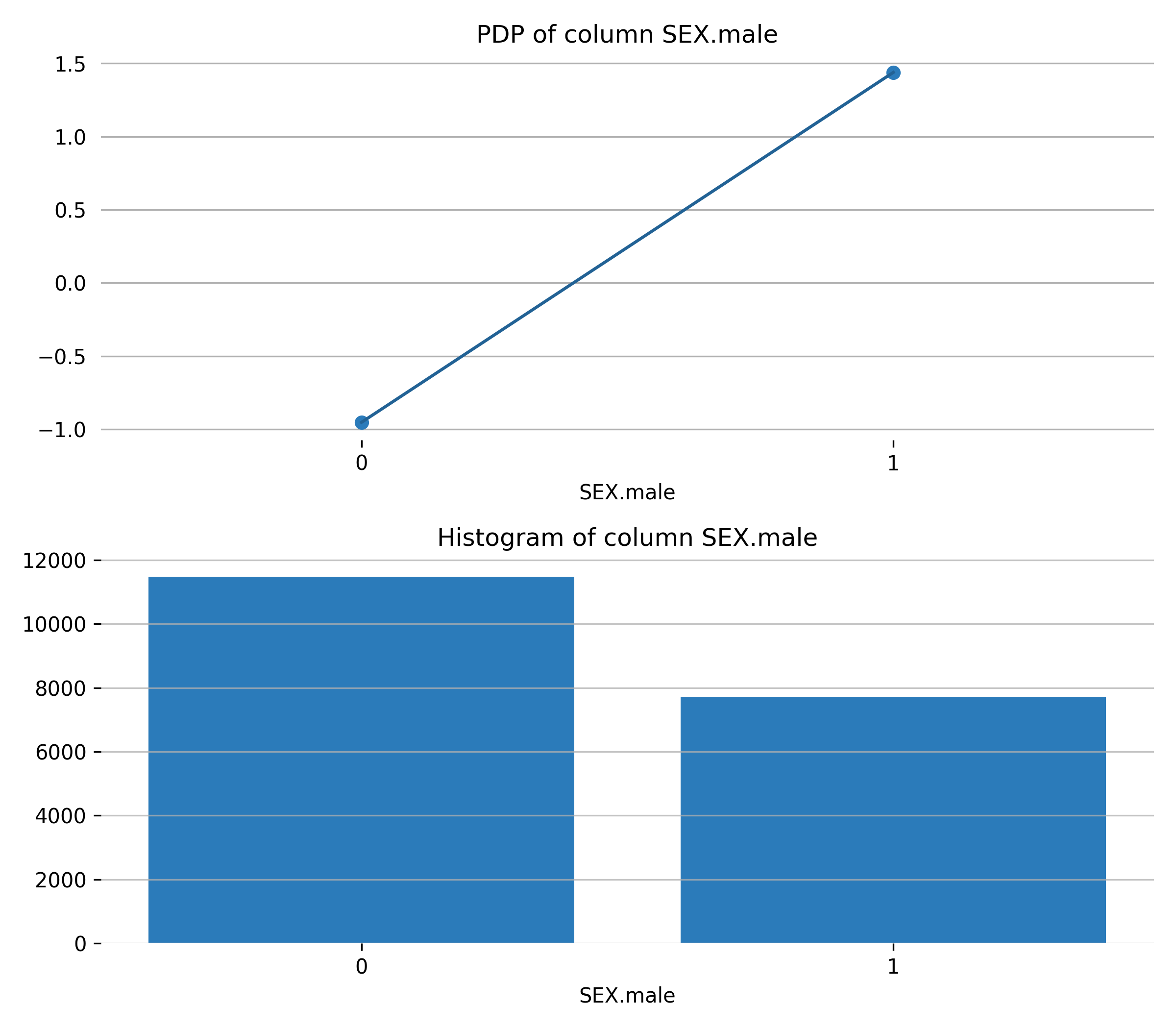
**Plot Details**

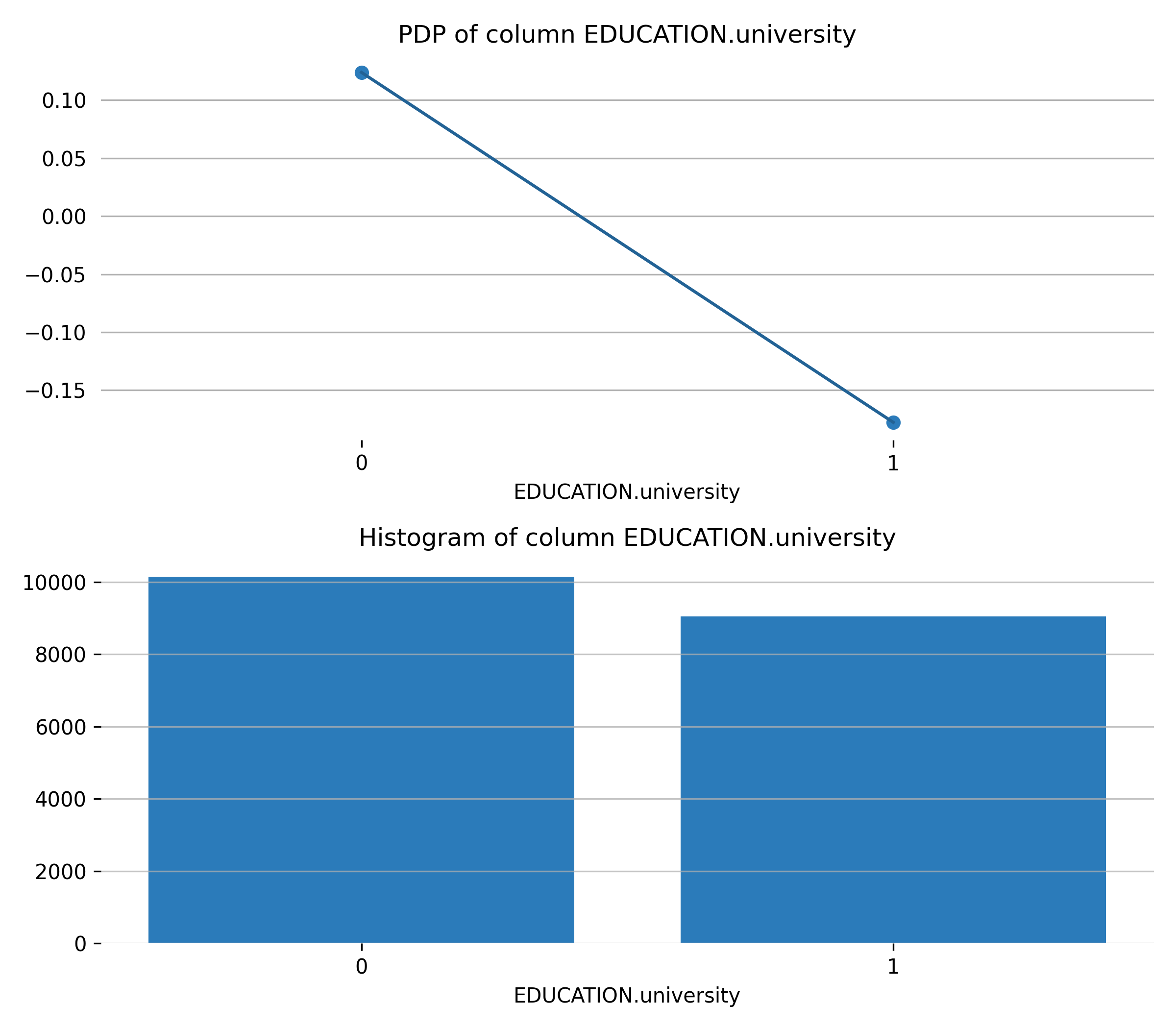
In the Scikit-learn PDP, the y-axis represents the mean response, and a shaded region (for numeric features) or shaded bar (for categorical features) represents ± 1 standard deviation.

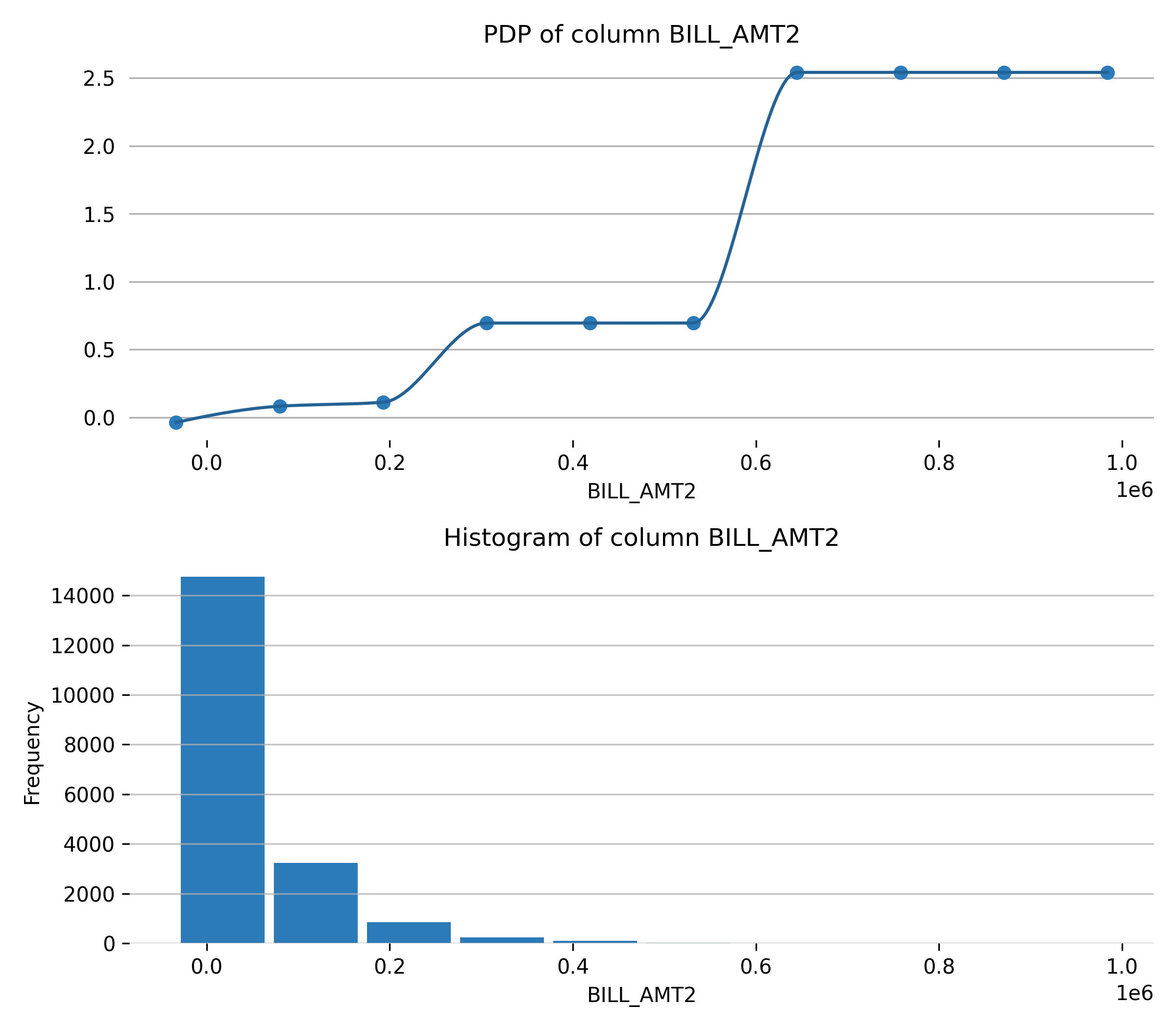
Feature **MARRIAGE.single**

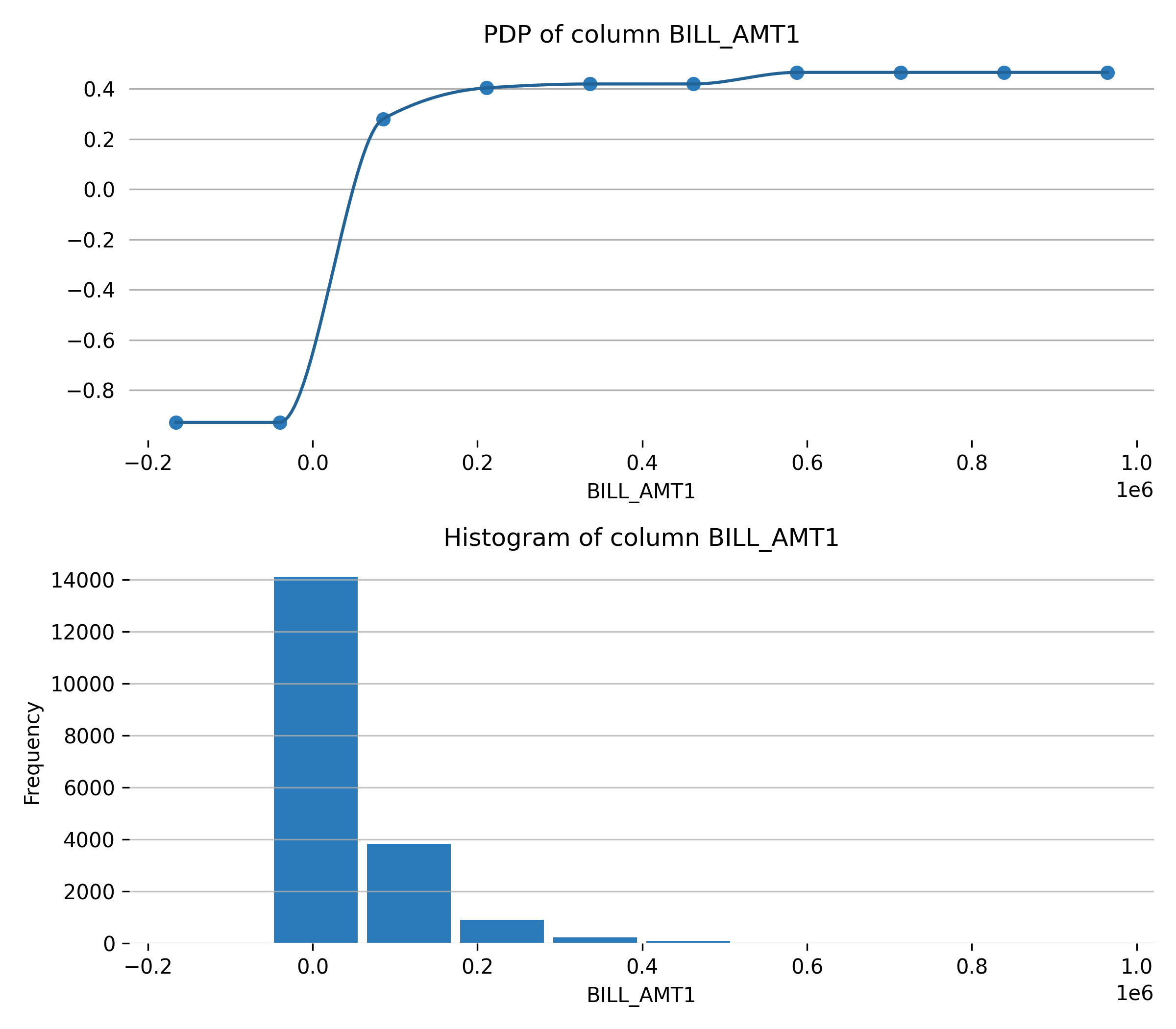
Feature **EDUCATION.highschool**

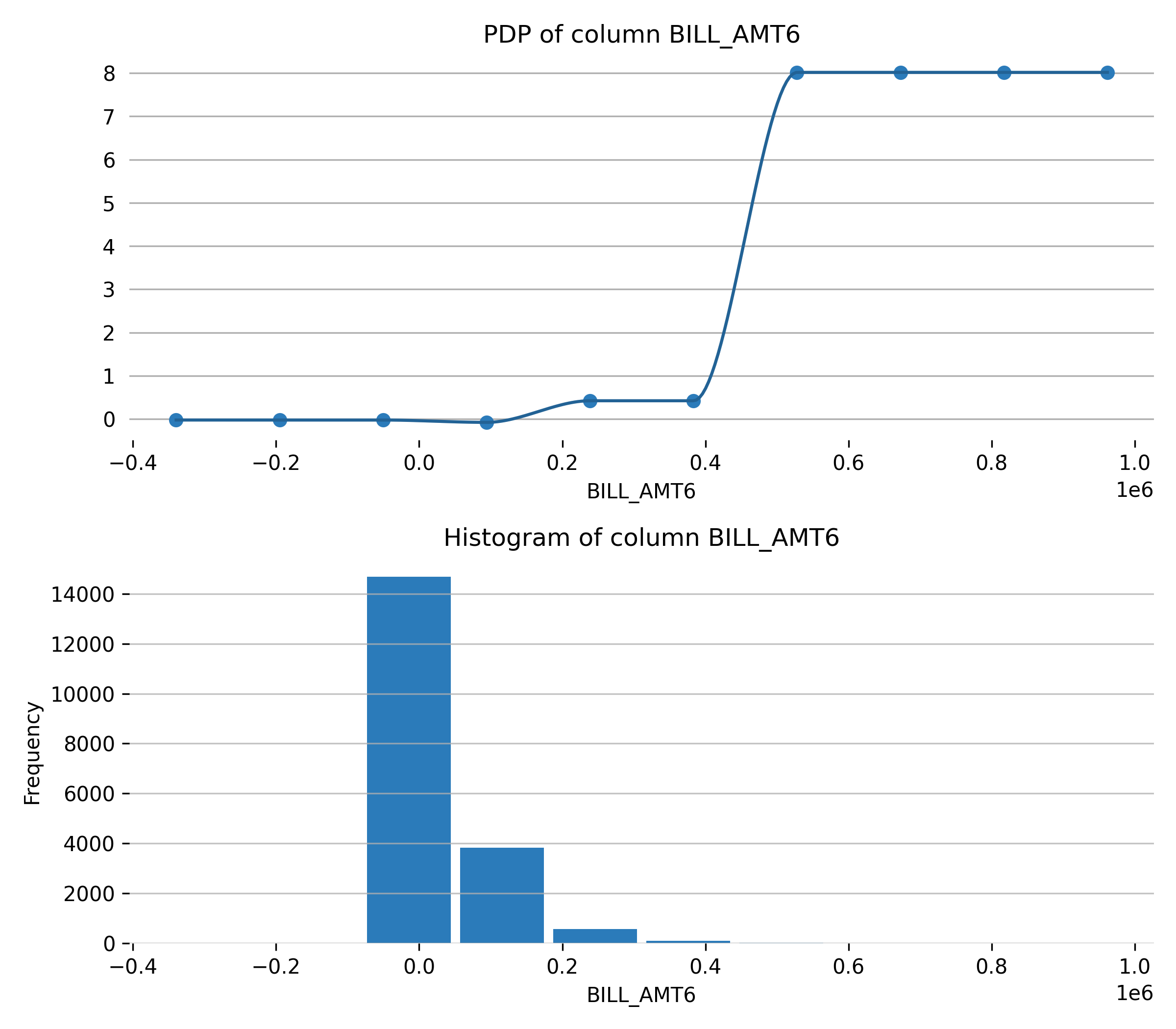
Feature **LIMIT\_BAL**

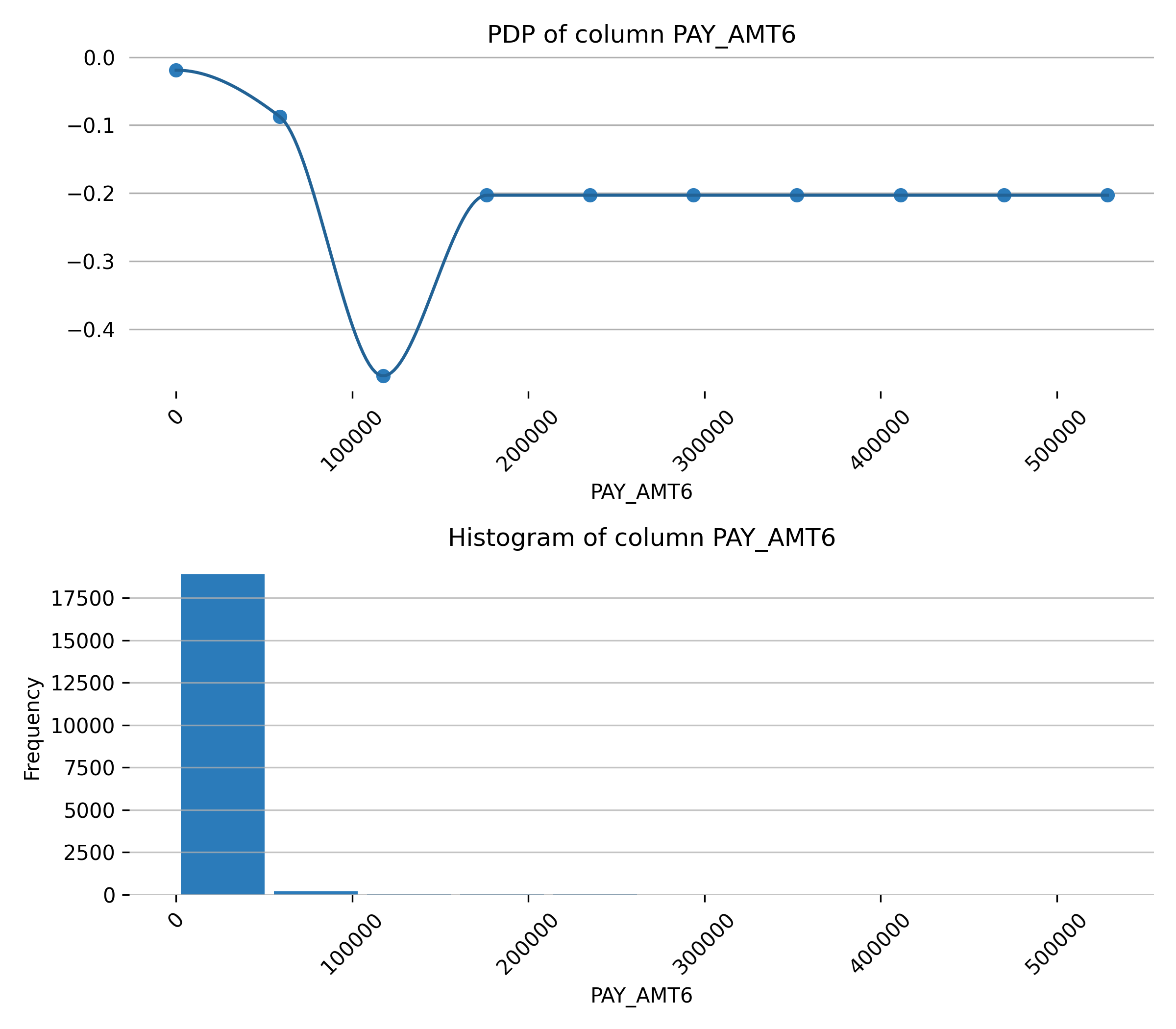
Feature **SEX.male**

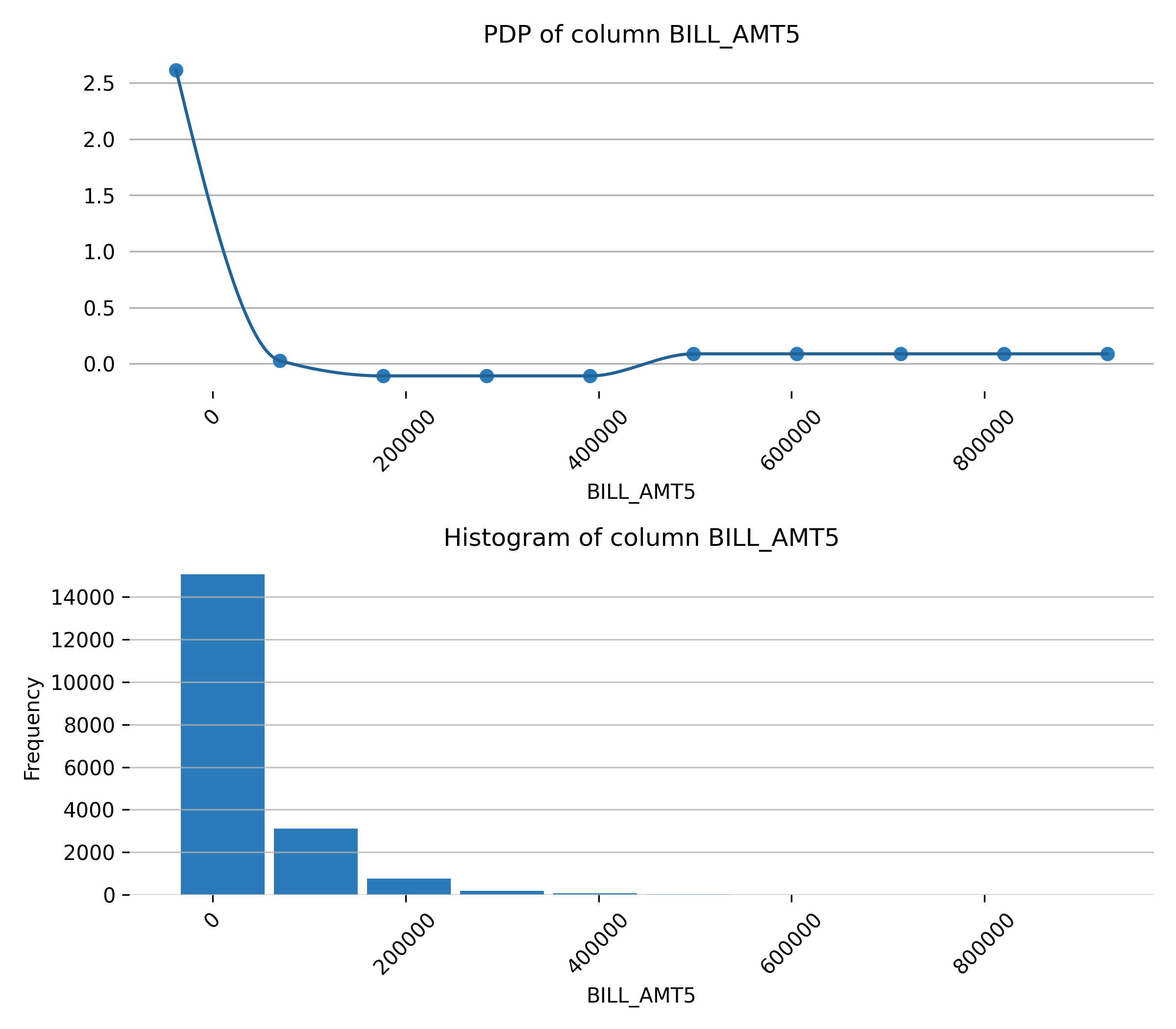
Feature **EDUCATION.university**

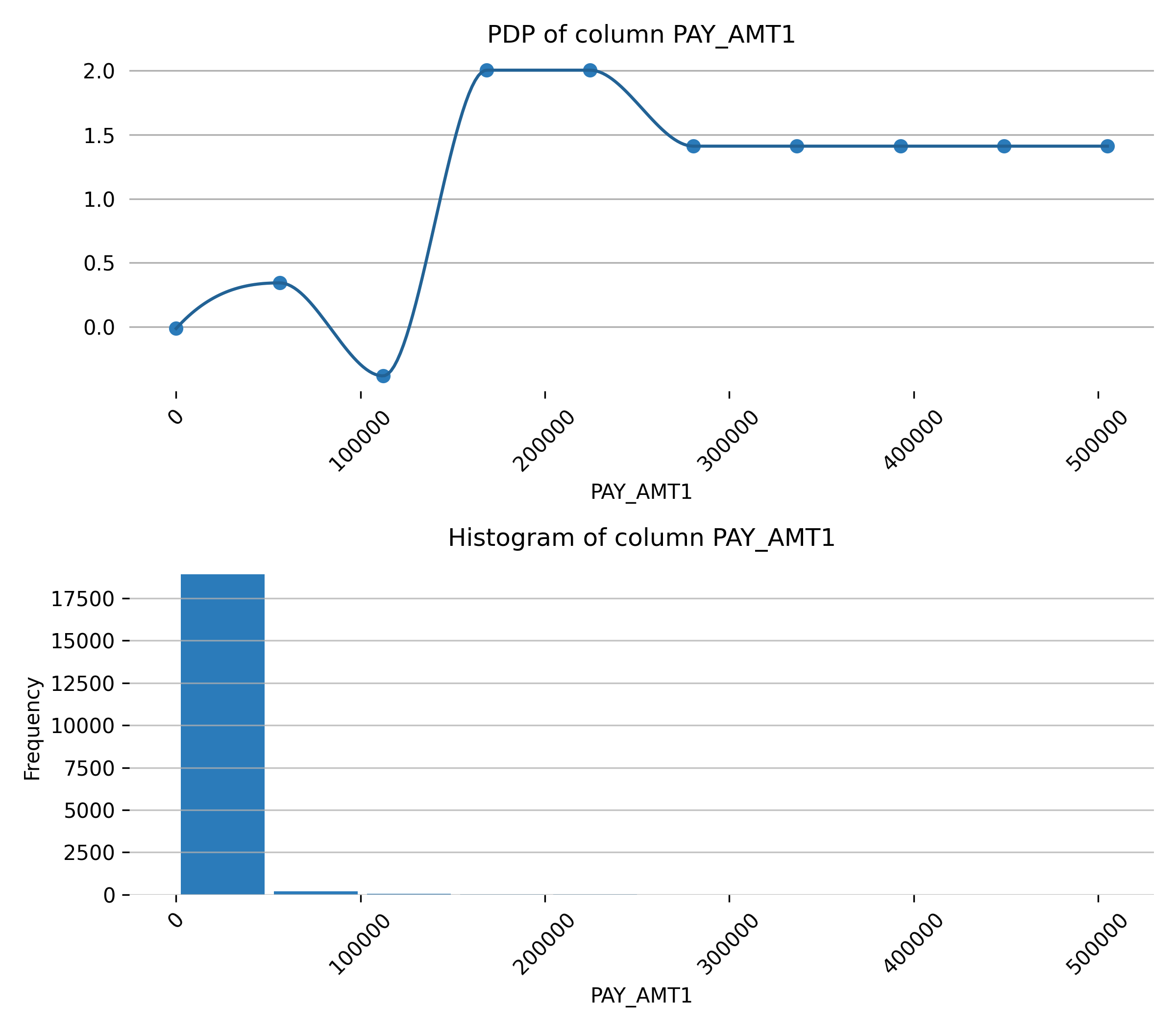
Feature **BILL\_AMT2**

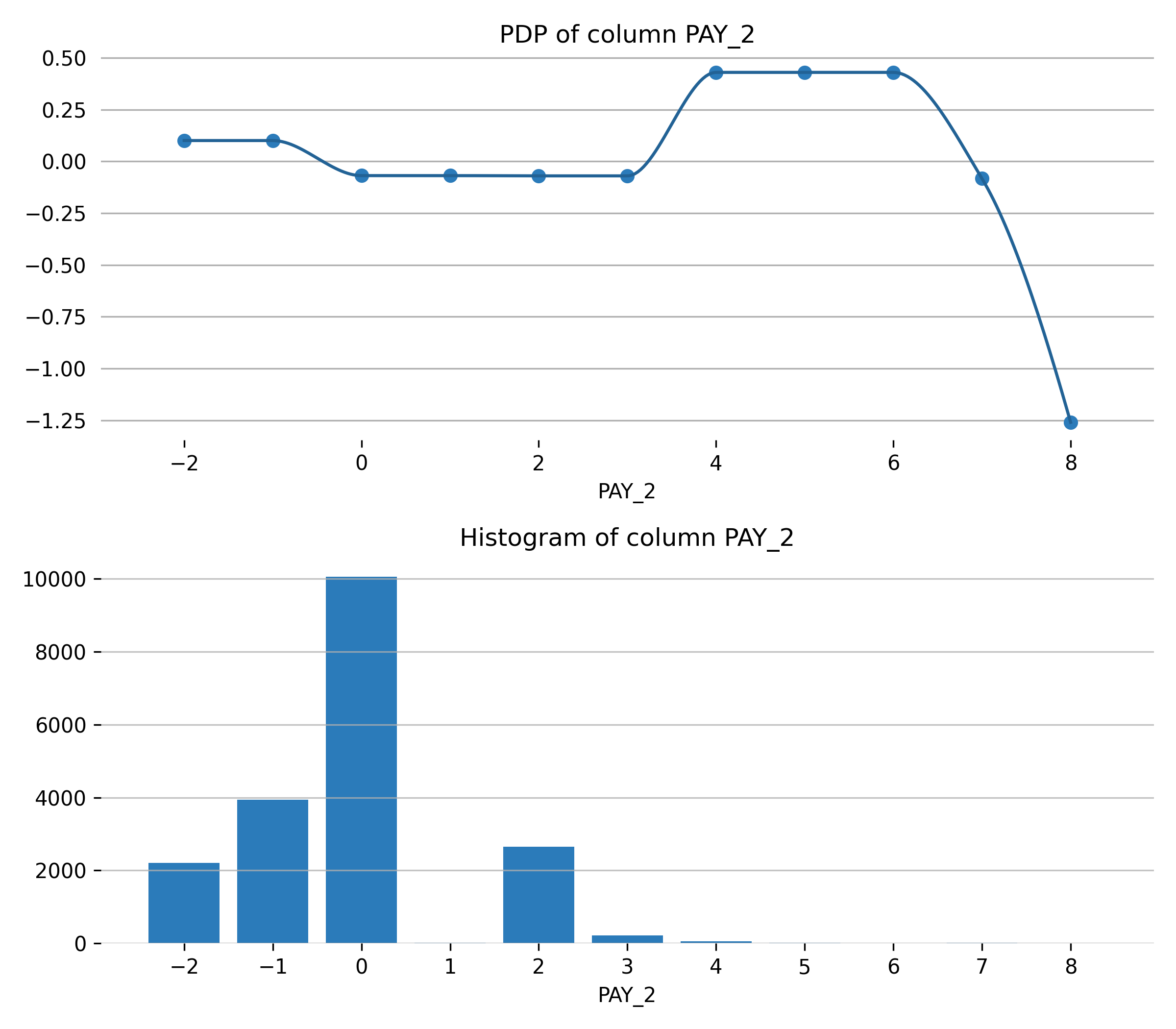
Feature **BILL\_AMT1**

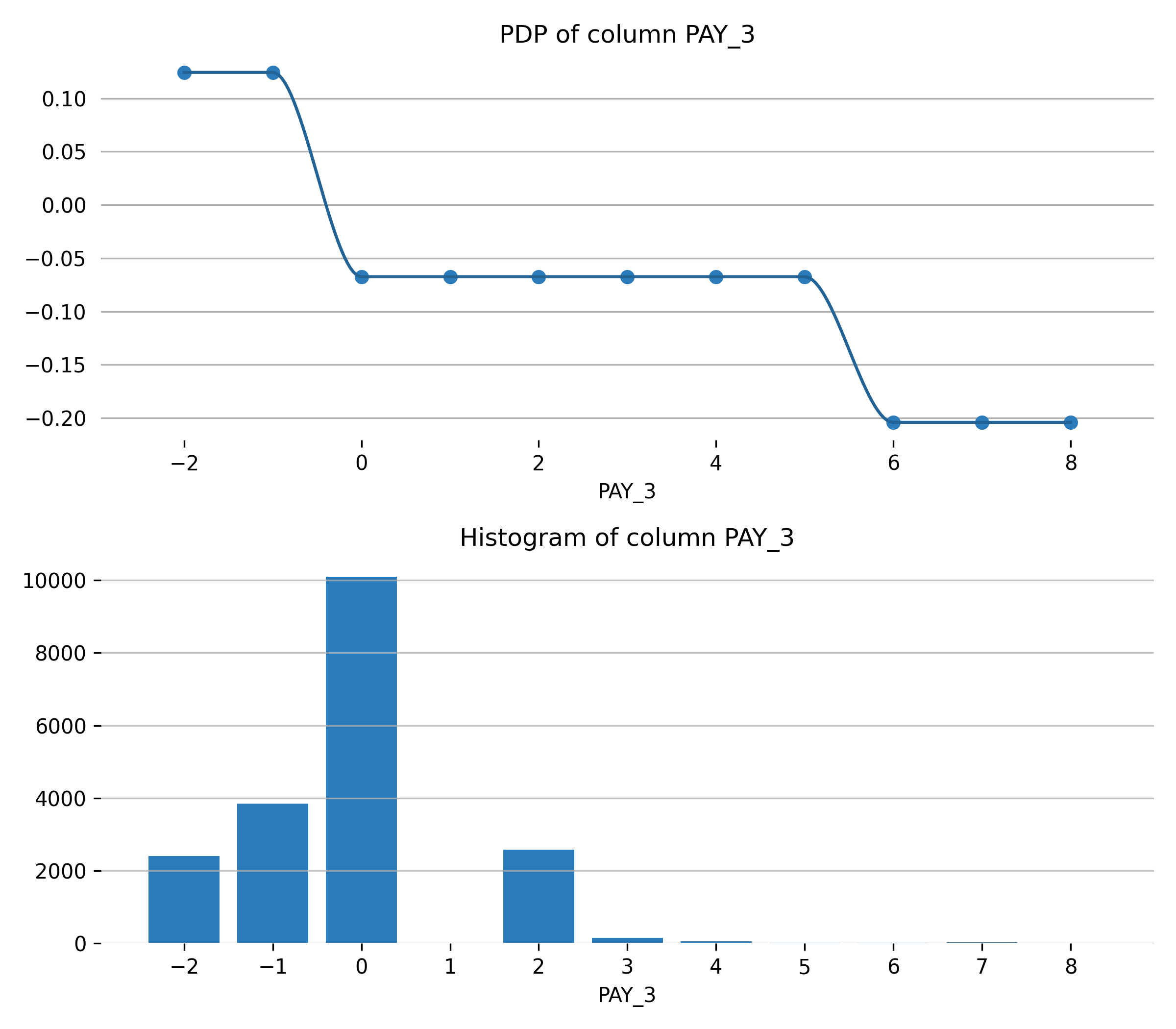
Feature **BILL\_AMT6**

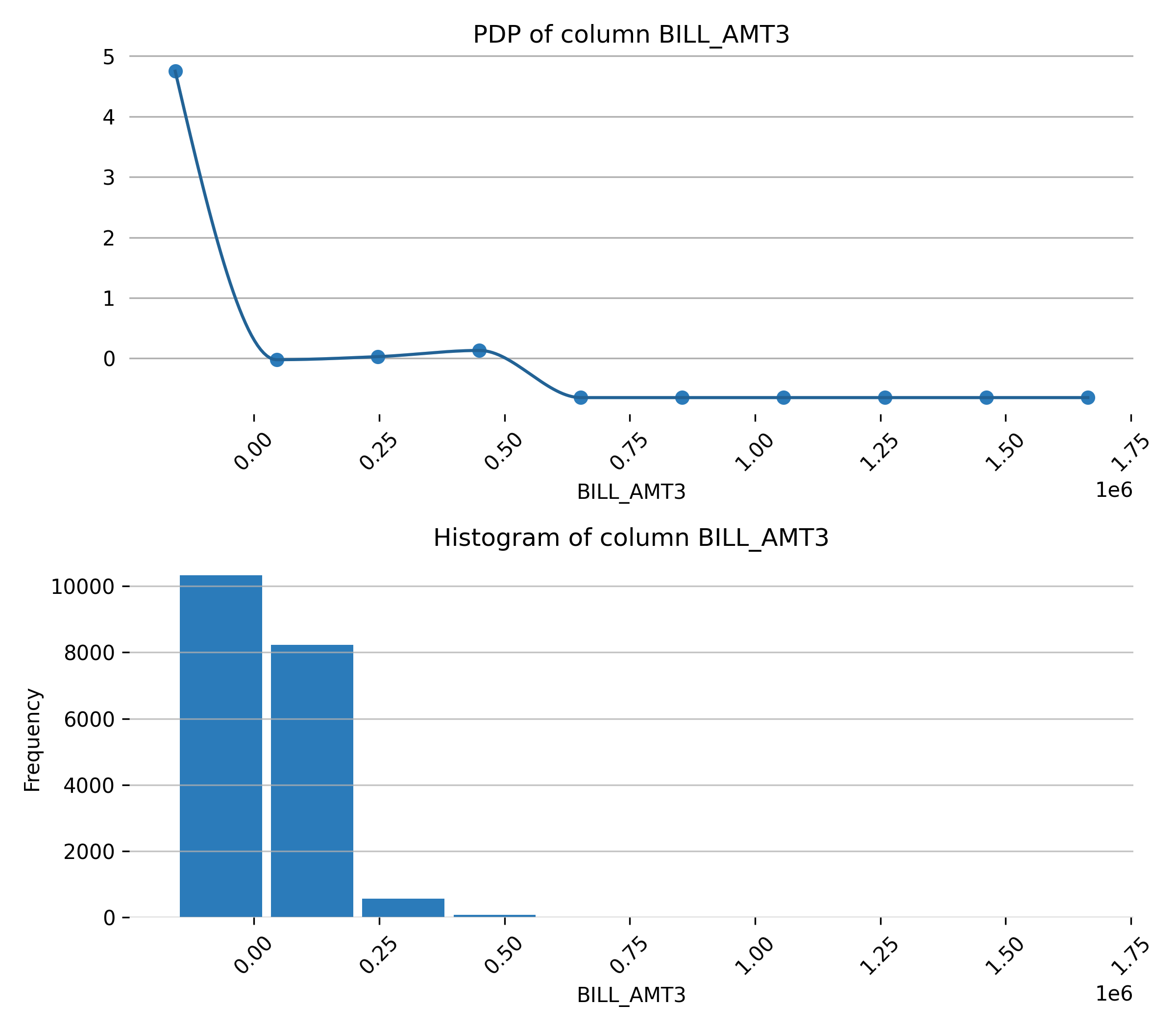
Feature **PAY\_AMT6**

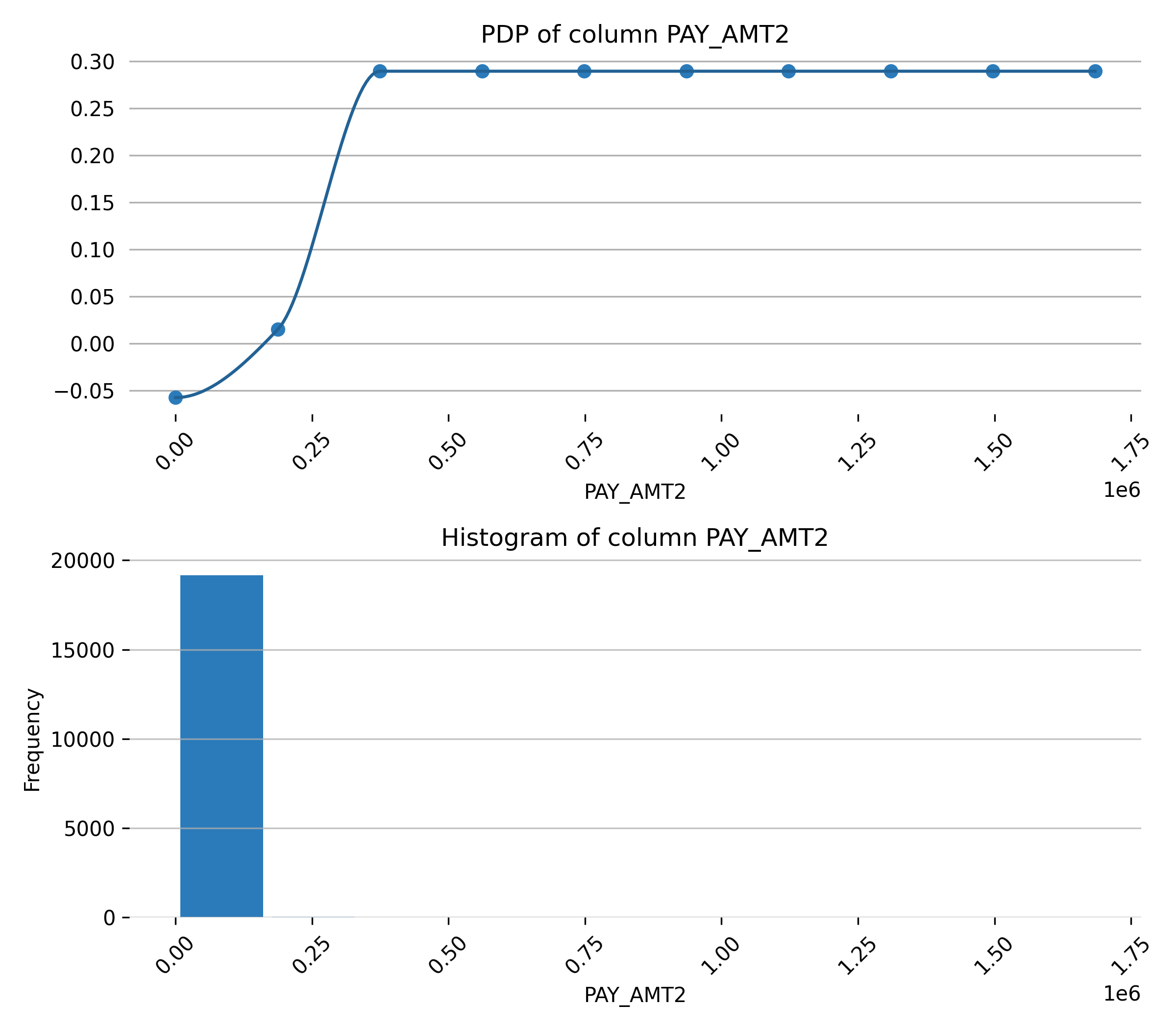
Feature **BILL\_AMT5**

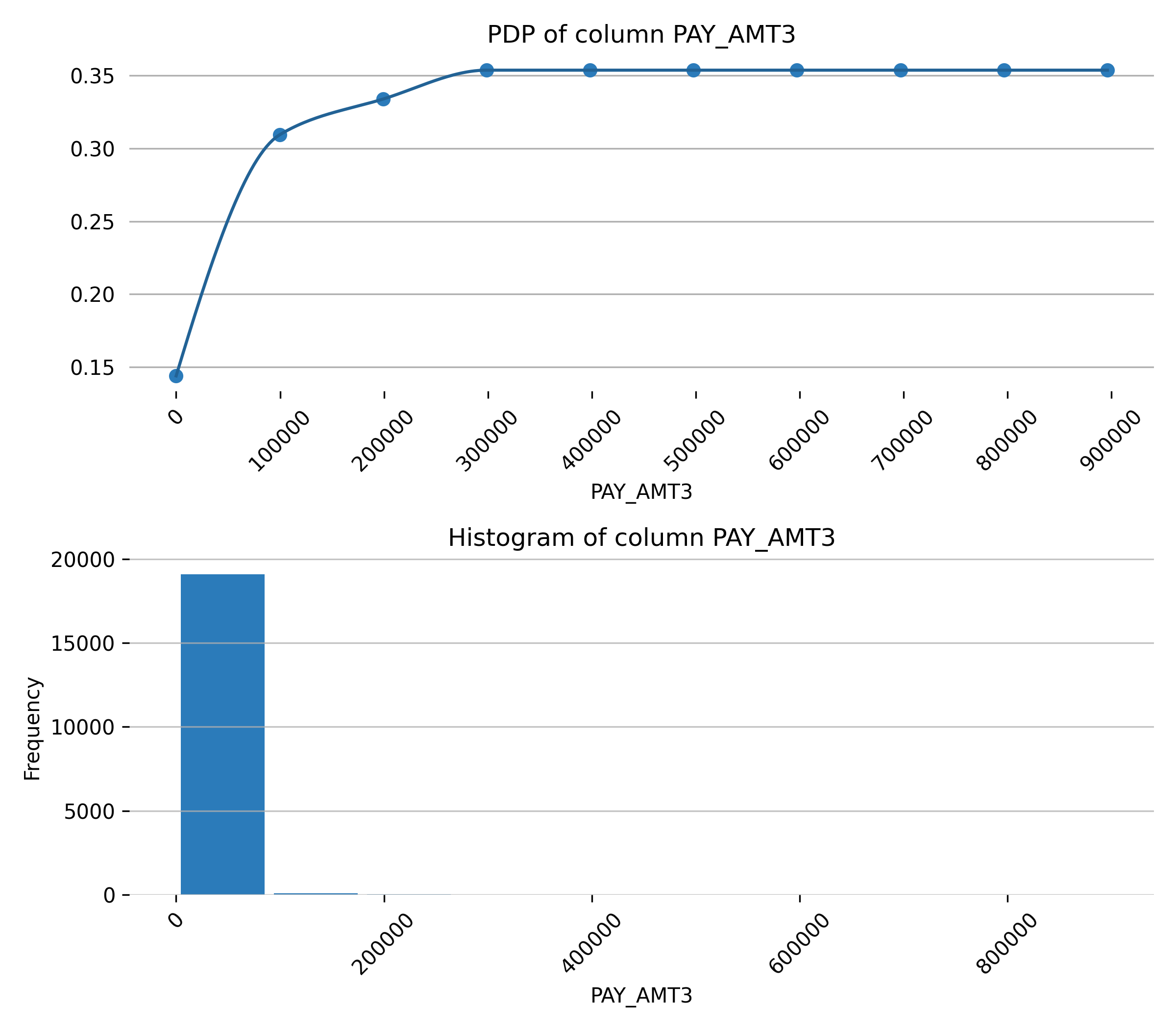
Feature **PAY\_AMT1**

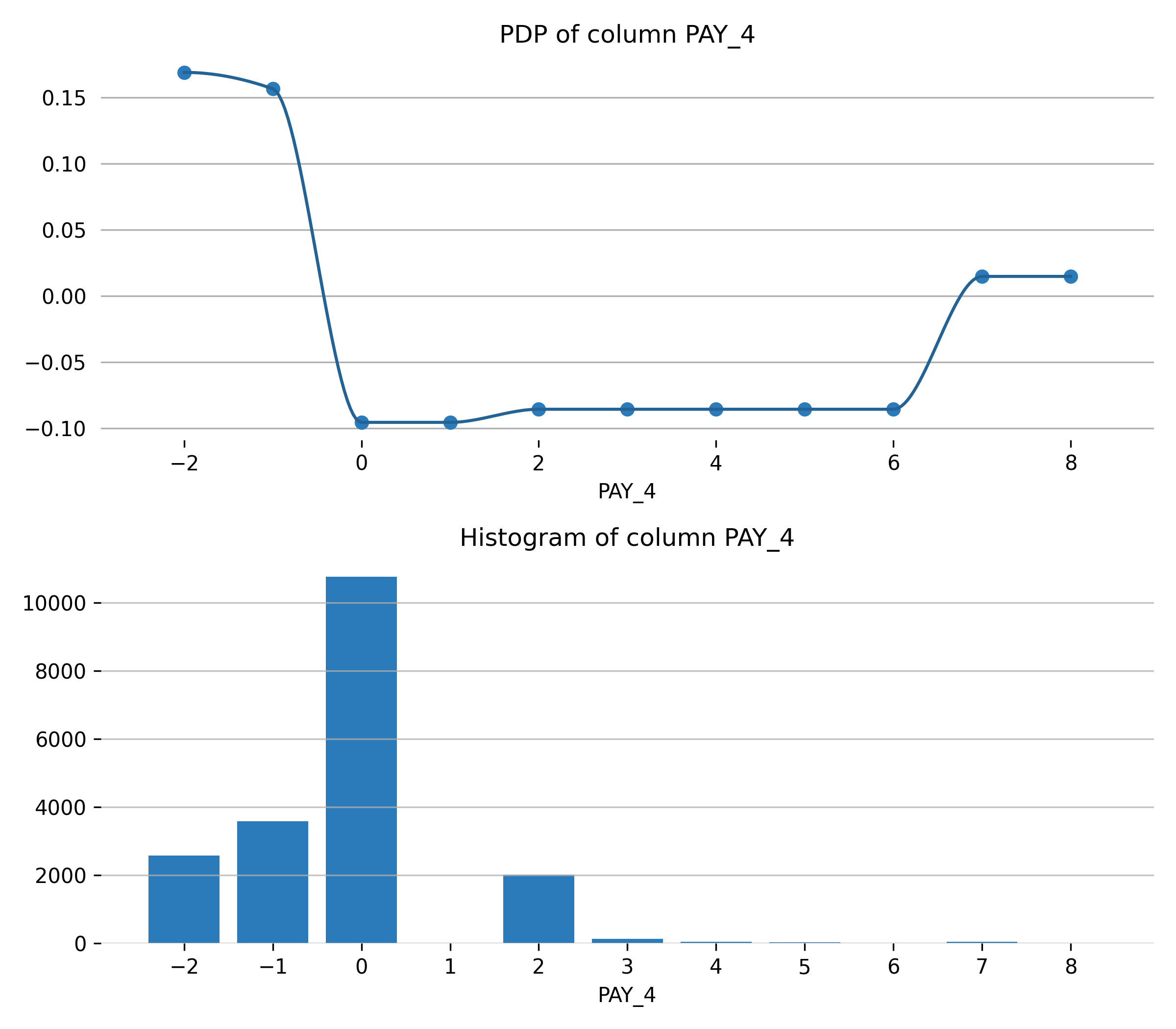
Feature **PAY\_2**

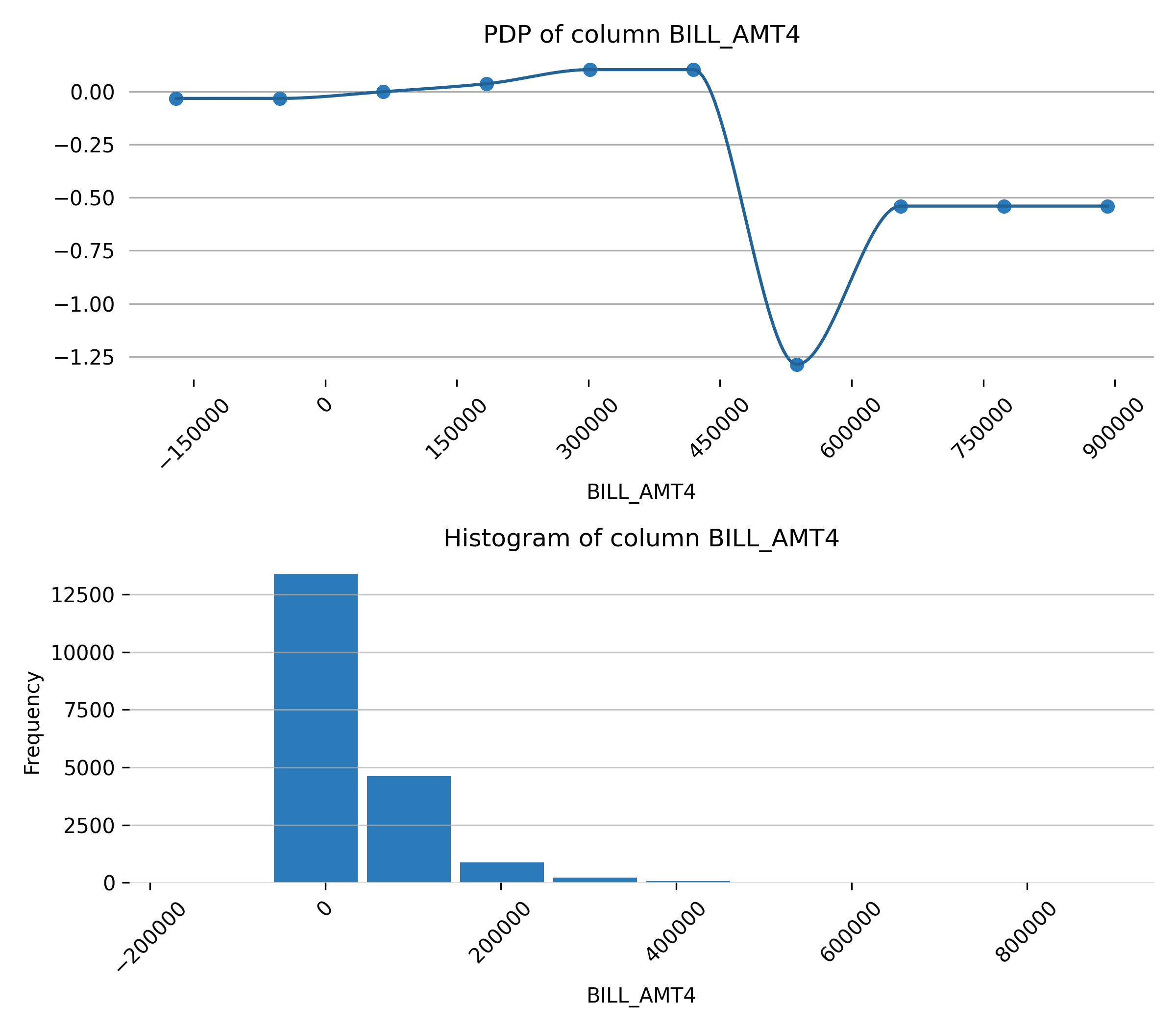
Feature **PAY\_3**

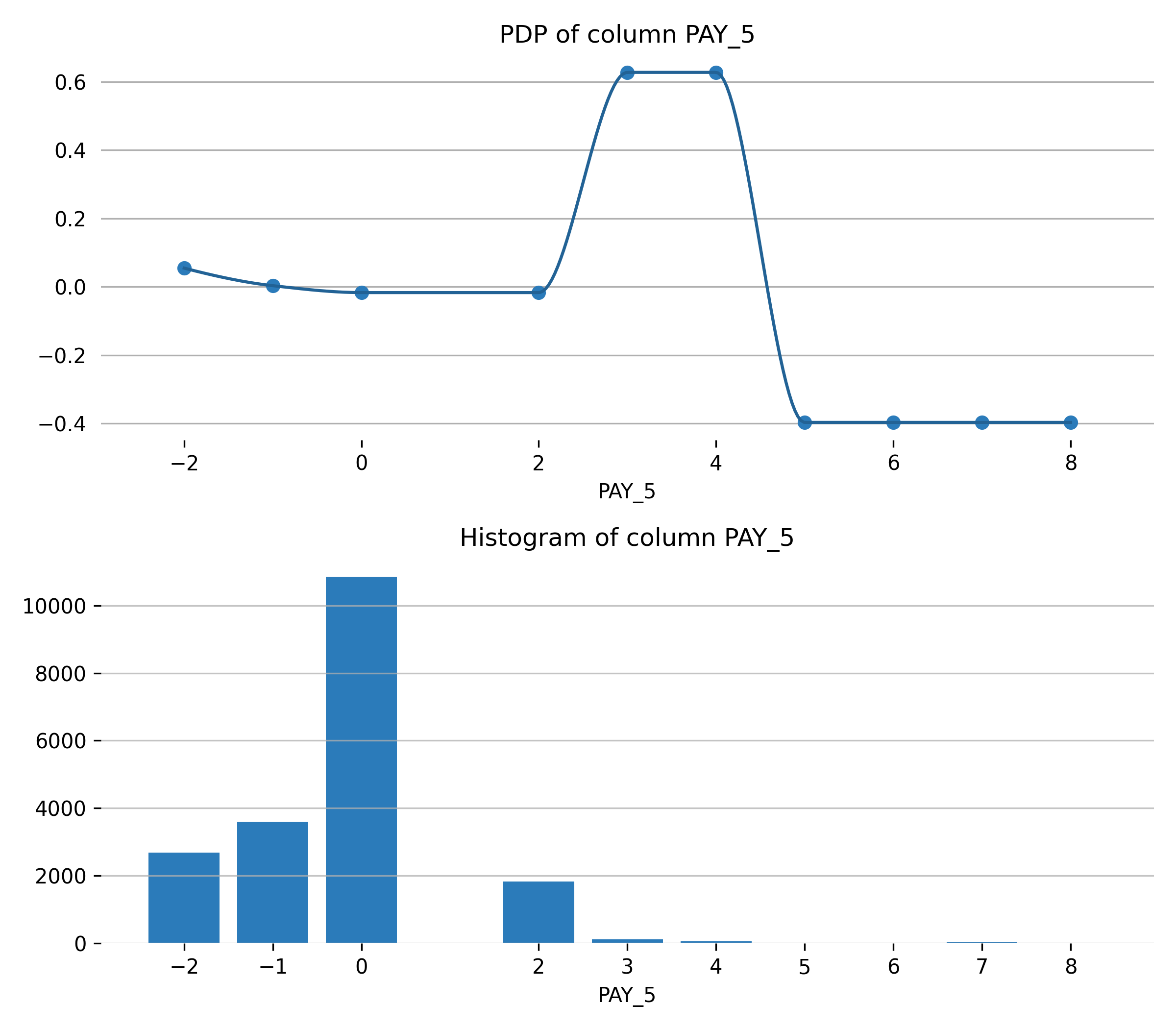
Feature **BILL\_AMT3**

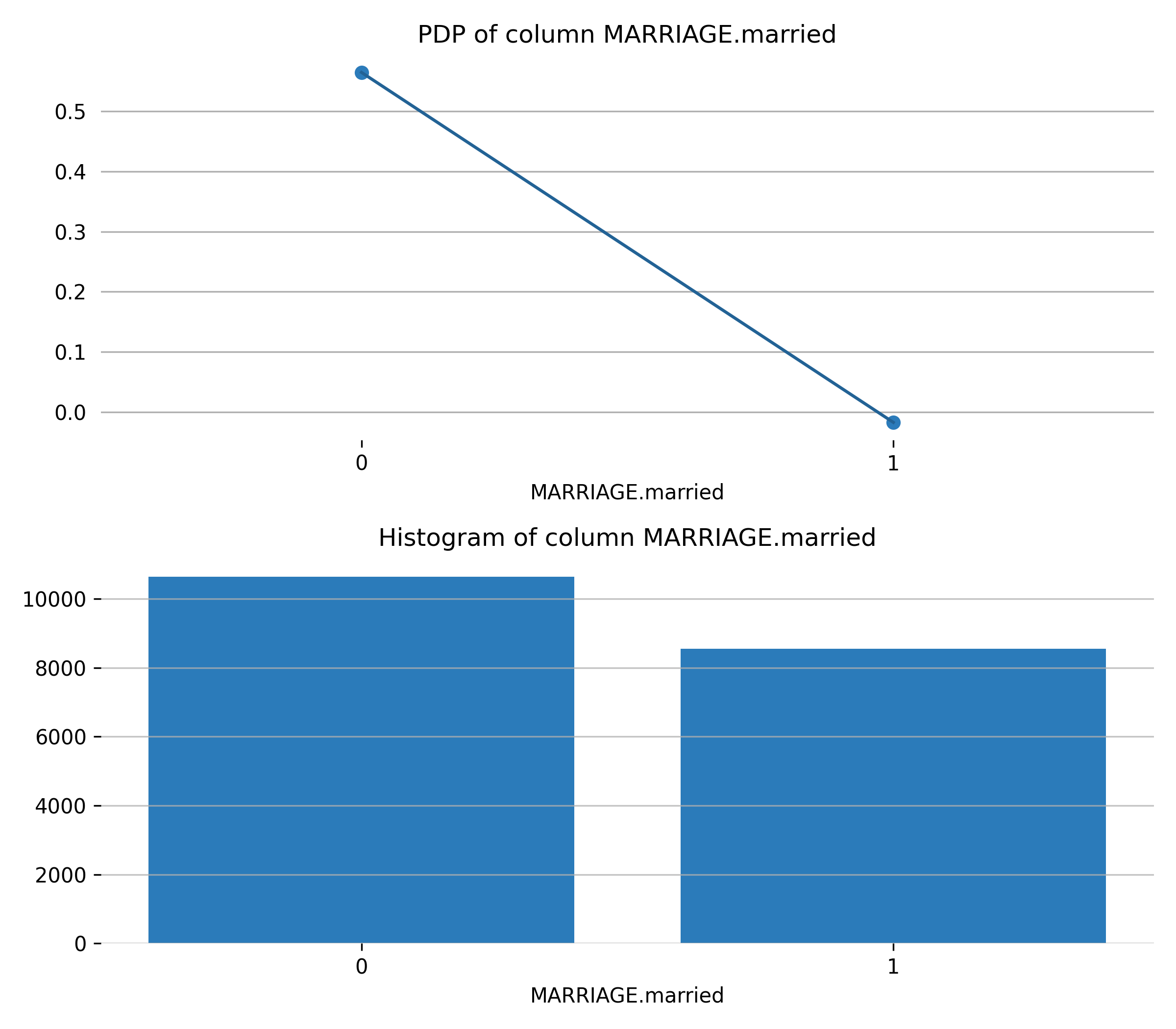
Feature **PAY\_AMT2**

Feature **PAY\_AMT3**

Feature **PAY\_4**

Feature **BILL\_AMT4**

Feature **PAY\_5**

Feature **MARRIAGE.married**

## Appendix

### Final Model Details

**Population Stability Index (PSI) Final Model Details**

Population Stability Index is a statistic used to describe a variable’s distribution shift. It can measure the shift between the training dataset’s model score distribution and any other given dataset (i.e. validation or test dataset).

A PSI value lower than 0.10 indicates a small shift in the model predictions, a value between 0.10 and 0.25 indicates a moderate shift, and a value greater than 0.25 indicates a strong shift. Strong shift values can indicate that the model trained on the training dataset might not be suitable for the provided validation or test datasets.

The PSI and calculation table is provided for each dataset below. The corresponding table columns are defined as follows:

* *Quantile: the bin to which the ordered predicted probabilities belong.*
* *Upper Bound: the upper bound of the corresponding bin.*
* *Test Count: the total number of Test records within the corresponding bin.*
* *Test Fraction (Tst): Test Count divided by the total number of Test records.*
* *Train Count: the total number of Train records within the corresponding bin.*
* *Train Fraction (Trn): Train Count divided by the total number of Train records.*
* *Tst - Trn: the difference between the Test Fraction and the Train Fraction.*
* *ln(Tst / Trn): the natural logarithm of the Test Fraction divided by the Train Fraction.*
* *PSI: the Population Stability Index for each bin - the dataset PSI is the total sum of these PSI values.*

**Test**

The Population Stability Index is 0.02935.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Quantile** | **Upper Bound** | **Test Count** | **Test Fraction (Tst)** | **Train Count** | **Train Fraction (Trn)** | **Tst - Trn** | **ln(Tst / Trn)** | **PSI** |
| 0.1 | 28.8531 | 401 | 0.0835 | 1,920 | 0.1 | -0.0165 | -0.1799 | 0.003 |
| 0.2 | 30.2378 | 426 | 0.0888 | 1,920 | 0.1 | -0.0113 | -0.1194 | 0.0013 |
| 0.3 | 31.4641 | 502 | 0.1046 | 1,916 | 0.0998 | 0.0048 | 0.0468 | 0.0002 |
| 0.4 | 32.7827 | 537 | 0.1119 | 1,923 | 0.1002 | 0.0117 | 0.1106 | 0.0013 |
| 0.5 | 35.1026 | 418 | 0.0871 | 1,920 | 0.1 | -0.0129 | -0.1384 | 0.0018 |
| 0.6 | 36.7755 | 547 | 0.114 | 1,920 | 0.1 | 0.014 | 0.1306 | 0.0018 |
| 0.7 | 38.7381 | 634 | 0.1321 | 1,920 | 0.1 | 0.0321 | 0.2782 | 0.0089 |
| 0.8 | 40.3296 | 546 | 0.1138 | 1,920 | 0.1 | 0.0137 | 0.1288 | 0.0018 |
| 0.9 | 42.0546 | 438 | 0.0912 | 1,920 | 0.1 | -0.0088 | -0.0916 | 0.0008 |
| 1.0 | inf | 351 | 0.0731 | 1,920 | 0.1 | -0.0269 | -0.3131 | 0.0084 |