analyze_feature_distributions

August 8, 2025

Connected to .venv (Python 3.11.9)

[ ]:
```
import sys

sys.path.append("..")
import pathlib
from IPython.display import display

from datasources.loaders import RegionLoader
from datasources.local import LocalDatasource
from base.dataset_loader import CategoricalLoader, CompanyDataFilter,␣
 ↪FinancialLoader, ScopeLoader
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
from base import OxariDataManager
from datasources.core import DefaultDataManager,␣
 ↪PreviousScopeFeaturesDataManager
from datasources.online import S3Datasource
from pathlib import Path

sns.set_palette('viridis')

PARENT_PATH = Path('..').absolute().resolve().as_posix()
PARENT_PATH
```

[ ]:
```
'C:/Users/User/Workspace/work_oxari/architectura'
```

[ ]:
```
dataset = PreviousScopeFeaturesDataManager(
    FinancialLoader(datasource=LocalDatasource(path=PARENT_PATH + "/model-data/
 ↪input/financials.csv")),
    ScopeLoader(datasource=LocalDatasource(path=PARENT_PATH + "/model-data/
 ↪input/scopes.csv")),
    CategoricalLoader(datasource=LocalDatasource(path=PARENT_PATH + "/
 ↪model-data/input/categoricals.csv")),
```

```
    RegionLoader(),
).set_filter(CompanyDataFilter(frac=1)).run()
DATA = dataset.get_data_by_name(OxariDataManager.ORIGINAL)
DATA
```

```
[I 2025-08-08 01:44:33,898] PreviousScopeFeaturesDataManager - INFO -
Remaining data points 0
[I 2025-08-08 01:44:33,900] FinancialLoader - INFO - Loading...
[I 2025-08-08 01:44:33,901] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-
data\input\financials.csv
[I 2025-08-08 01:44:42,631] FinancialLoader - INFO - Completed download
-- 8.729533672332764 seconds
[I 2025-08-08 01:44:42,632] PreviousScopeFeaturesDataManager - INFO -
Added loader_financialloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:42,633] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_0 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:42,634] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 110)
[I 2025-08-08 01:44:42,636] ScopeLoader - INFO - Loading...
[I 2025-08-08 01:44:42,638] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-data\input\scopes.csv
[I 2025-08-08 01:44:43,037] ScopeLoader - INFO - Completed download --
0.39867401123046875 seconds
[I 2025-08-08 01:44:43,039] CombinedLoader - INFO - Adding
(FinancialLoader + ScopeLoader)
[I 2025-08-08 01:44:43,040] CombinedLoader - INFO - Merging special
loader ScopeLoader to FinancialLoader
[I 2025-08-08 01:44:44,868] PreviousScopeFeaturesDataManager - INFO -
Added loader_scopeloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:44,869] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_1 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:44,870] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 122)
[I 2025-08-08 01:44:44,871] CategoricalLoader - INFO - Loading...
[I 2025-08-08 01:44:44,872] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-
data\input\categoricals.csv
[I 2025-08-08 01:44:47,358] CategoricalLoader - INFO - Completed
download -- 2.4865143299102783 seconds
[I 2025-08-08 01:44:47,359] CombinedLoader - INFO - Adding
(FinancialLoader-ScopeLoader + CategoricalLoader)
[I 2025-08-08 01:44:48,867] PreviousScopeFeaturesDataManager - INFO -
Added loader_categoricalloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:48,869] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_2 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:48,870] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 126)
```

```
[I 2025-08-08 01:44:48,871] RegionLoader - INFO - Loading…
[I 2025-08-08 01:44:48,874] OnlineCSVDatasource - INFO - Fetching data
from https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-
Codes/master/all/all.csv
[I 2025-08-08 01:44:49,231] RegionLoader - INFO - Completed download --
0.3571755886077881 seconds
[I 2025-08-08 01:44:49,232] CombinedLoader - INFO - Adding
(FinancialLoader-ScopeLoader-CategoricalLoader + RegionLoader)
[I 2025-08-08 01:44:49,233] CombinedLoader - INFO - Merging special
loader RegionLoader to FinancialLoader-ScopeLoader-CategoricalLoader
[I 2025-08-08 01:44:52,967] PreviousScopeFeaturesDataManager - INFO -
Added loader_regionloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:52,968] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_3 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:52,970] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 129)
[I 2025-08-08 01:44:52,972] PreviousScopeFeaturesDataManager - INFO -
Added merged to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:54,535] CompanyDataFilter - INFO - Filtered dataset
from 526241 to 526241 data points
[I 2025-08-08 01:44:54,541] PreviousScopeFeaturesDataManager - INFO -
Added reduced to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:44:54,544] PreviousScopeFeaturesDataManager - INFO -
Taking all previous year scopes
100%|      | 103752/103752 [03:12<00:00, 540.03it/s]
[I 2025-08-08 01:48:08,363] PreviousScopeFeaturesDataManager - INFO -
Added original to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:48:08,784] PreviousScopeFeaturesDataManager - INFO -
Data with original found retrieved: Dataset after transformation changes.
```

[ ]:

	ft_catm_country_code	ft_catm_exchange	ft_catm_industry_name	… tg_numc_scope_1 \
0	PRT	XBER	Utilities - Rene…	… NaN
1	PRT	XBER	Utilities - Rene…	… NaN
2	PRT	XBER	Utilities - Rene…	… NaN
3	PRT	XBER	Utilities - Rene…	… NaN
4	PRT	XDUS	Utilities - Rene…	… NaN
…	…	…	…	… …
526238	NaN	NaN	NaN	… NaN
526239	NaN	NaN	NaN	…

```
NaN
526236                  NaN              NaN                  NaN   …
NaN
526237                  NaN              NaN                  NaN   …
NaN
526240                  NaN              NaN                  NaN   …
NaN

       tg_numc_scope_2 tg_numc_scope_3
0                  NaN             NaN
1                  NaN             NaN
2                  NaN             NaN
3                  NaN             NaN
4                  NaN             NaN
…                  …               …
526238             NaN             NaN
526239             NaN             NaN
526236             NaN             NaN
526237             NaN             NaN
526240             NaN             NaN

[522776 rows x 129 columns]
```

[ ]:
```
df_scopes = DATA
df_scopes["grp_scope_1"] = None
df_scopes["log_scope_1"] = None
df_scopes.loc[df_scopes["tg_numc_scope_1"].isna(), ["grp_scope_1"]] = "Not␣
 ↪reported"
df_scopes.loc[df_scopes["tg_numc_scope_1"] == 0, ["grp_scope_1"]] = "Zero␣
 ↪Emissions"
df_scopes.loc[df_scopes["tg_numc_scope_1"] < 0, ["grp_scope_1"]] = "Impossible"
df_scopes.loc[df_scopes["tg_numc_scope_1"].between(0, 1, inclusive='right'),␣
 ↪["grp_scope_1"]] = "Weird"
df_scopes.loc[df_scopes["tg_numc_scope_1"] > 1, ["grp_scope_1"]] = "Emittor"
df_scopes["log_scope_1"] = np.log(df_scopes["tg_numc_scope_1"])
indices = df_scopes["tg_numc_scope_1"] > 0
df_scopes
```

```
c:\Users\User\Workspace\work_oxari\architectura\.venv\Lib\site-
packages\pandas\core\arraylike.py:402: RuntimeWarning: divide by zero
encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

[ ]:
```
       ft_catm_country_code ft_catm_exchange ft_catm_industry_name  …
tg_numc_scope_3  \
0                       PRT             XBER  Utilities - Rene…   …
NaN
```

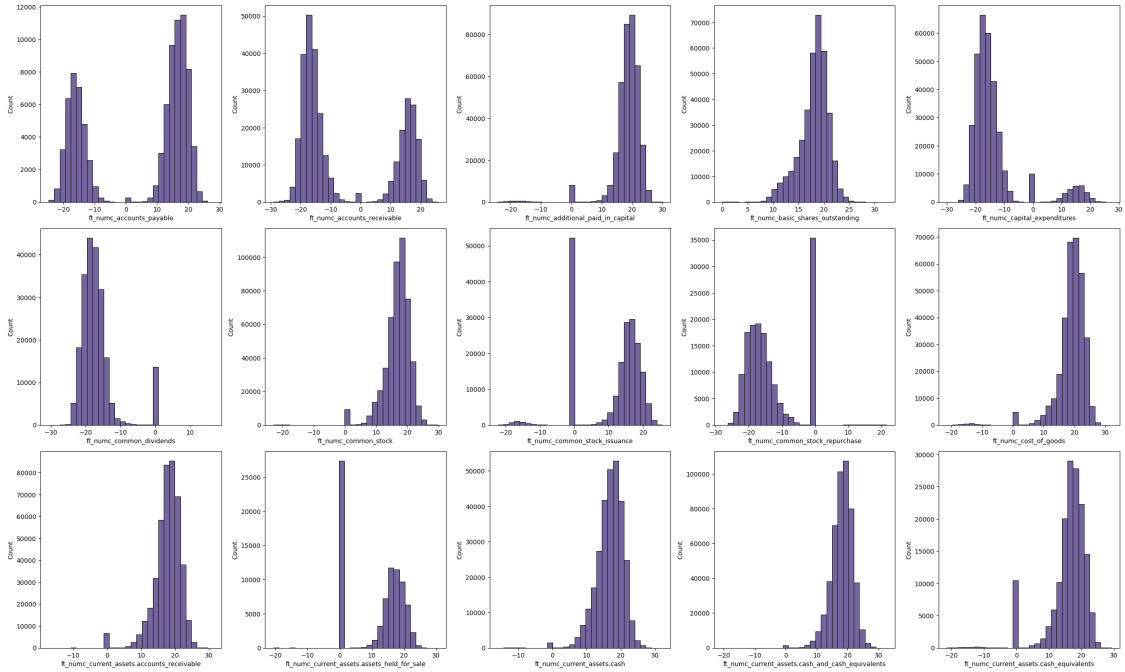
```
1                       PRT             XBER  Utilities - Rene…   …
NaN
2                       PRT             XBER  Utilities - Rene…   …
NaN
3                       PRT             XBER  Utilities - Rene…   …
NaN
4                       PRT             XDUS  Utilities - Rene…   …
NaN
…                       …               …                    …   …
…
526238                  NaN              NaN                 NaN   …
NaN
526239                  NaN              NaN                 NaN   …
NaN
526236                  NaN              NaN                 NaN   …
NaN
526237                  NaN              NaN                 NaN   …
NaN
526240                  NaN              NaN                 NaN   …
NaN

         grp_scope_1 log_scope_1
0       Not reported         NaN
1       Not reported         NaN
2       Not reported         NaN
3       Not reported         NaN
4       Not reported         NaN
…               …          …
526238  Not reported         NaN
526239  Not reported         NaN
526236  Not reported         NaN
526237  Not reported         NaN
526240  Not reported         NaN

[522776 rows x 131 columns]
```

[ ]:
```
numerical_features = df_scopes.filter(regex="^ft_numc", axis=1)
structure = pd.concat([numerical_features.skew(), numerical_features.
 ↪kurtosis()], axis=1)
structure.columns = ['skew', 'kurstosis']
structure
```

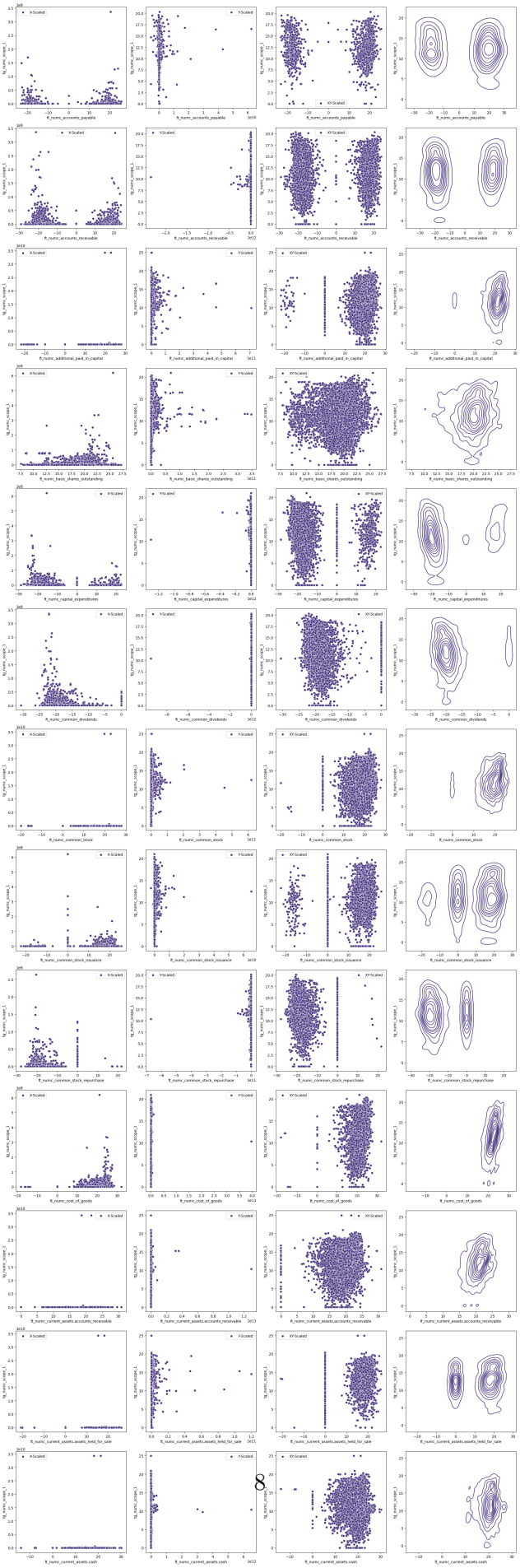
[ ]:
```
                            skew      kurstosis
ft_numc_accounts_…  179.676367   44387.550250
ft_numc_accounts_…  -76.119417   12329.606070
ft_numc_additiona…  122.782491   19847.173972
ft_numc_basic_sha…  424.577872  180674.384236
```

```
ft_numc_capital_e… -461.952518  241096.514527
…                             …             …
ft_numc_stock_bas…   23.407000     893.557416
ft_numc_total_assets  266.609469   84006.487859
ft_numc_total_lia…  204.797473   49095.991778
ft_numc_total_sha…  272.278577   88622.148168
ft_numc_treasury_…  105.557676   16264.509630

[102 rows x 2 columns]
```

[ ]:
```python
# https://www.researchgate.net/publication/
 ↪233967063_A_bi-symmetric_log_transformation_for_wide-range_data
# https://support.cytobank.org/hc/en-us/articles/
 ↪206148057-About-the-Arcsinh-transform: The hyperbolic arcsine (arcsinh) is a␣
 ↪function used in Cytobank for transforming data. It serves a similar purpose␣
 ↪as transformation functions such as biexponential, logicle, hyperlog, etc.
# https://dillonhammill.github.io/CytoExploreR/articles/
 ↪CytoExploreR-Transformations.html
# https://opendatascience.com/transforming-skewed-data-for-machine-learning/
#   Skewness test with shapiro-wilk test

num_bins = 30
fig, axes = plt.subplots(3, 5, figsize=(25, 15))
faxes = axes.flatten()
for ax, feature in zip(faxes, numerical_features.columns):
    tmp_df = df_scopes
    # tmp_df = tmp_df.dropna(how="any", subset=[target, feature])

    # sns.scatterplot(tmp_df, x=feature, y=target, ax=ax)

    sns.histplot(x=np.arcsinh(tmp_df[feature]), ax=ax, bins=num_bins)
    # sns.histplot(x=tmp_df[feature], ax=ax, bins=num_bins)
fig.tight_layout()
plt.show()
```



[ ]:
```
num_bins = 14
fig, axes = plt.subplots(13, 4, figsize=(20, 60))
faxes = axes.flatten()

years = df_scopes["key_year"].unique()
target = "tg_numc_scope_1"
# target = "log_scope_1"
for axs, feature in zip(axes, numerical_features.columns):
    tmp_df = df_scopes[[target, feature]]
    tmp_df = tmp_df.dropna(how="any")

    ax1, ax2, ax3, ax4 = axs
    sns.scatterplot(x=np.arcsinh(tmp_df[feature]), y=tmp_df[target], ax=ax1,␣
 ↪label="X-Scaled")
    sns.scatterplot(x=tmp_df[feature], y=np.arcsinh(tmp_df[target]), ax=ax2,␣
 ↪label="Y-Scaled")
    sns.scatterplot(x=np.arcsinh(tmp_df[feature]), y=np.
 ↪arcsinh(tmp_df[target]), ax=ax3, label="XY-Scaled")
    sns.kdeplot(x=np.arcsinh(tmp_df[feature]), y=np.arcsinh(tmp_df[target]),␣
 ↪ax=ax4, label="XY-Scaled")

fig.tight_layout()
plt.show()
```


X-Scaled
Y-Scaled
XY-Scaled
ft_numc_accounts_payable
ft_numc_accounts_receivable
ft_numc_additional_paid_in_capital
ft_numc_basic_shares_outstanding
ft_numc_capital_expenditures
ft_numc_common_dividends
ft_numc_common_stock
ft_numc_common_stock_issuance
ft_numc_common_stock_repurchase
ft_numc_cost_of_goods
ft_numc_current_assets.accounts_receivable
ft_numc_current_assets.assets_held_for_sale
ft_numc_current_assets.cash