analyze_feature_set_initial_discovery

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 13:56:28,001] PreviousScopeFeaturesDataManager - INFO -
Remaining data points 0
[I 2025-08-08 13:56:28,004] FinancialLoader - INFO - Loading…
[I 2025-08-08 13:56:28,005] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-
data\input\financials.csv
[I 2025-08-08 13:56:38,010] FinancialLoader - INFO - Completed download
-- 10.005178689956665 seconds
[I 2025-08-08 13:56:38,013] PreviousScopeFeaturesDataManager - INFO -
Added loader_financialloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:38,014] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_0 to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:38,015] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 110)
[I 2025-08-08 13:56:38,016] ScopeLoader - INFO - Loading…
[I 2025-08-08 13:56:38,019] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-data\input\scopes.csv
[I 2025-08-08 13:56:38,377] ScopeLoader - INFO - Completed download --
0.359450101852417 seconds
[I 2025-08-08 13:56:38,378] CombinedLoader - INFO - Adding
(FinancialLoader + ScopeLoader)
[I 2025-08-08 13:56:38,379] CombinedLoader - INFO - Merging special
loader ScopeLoader to FinancialLoader
[I 2025-08-08 13:56:39,715] PreviousScopeFeaturesDataManager - INFO -
Added loader_scopeloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:39,716] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_1 to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:39,718] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 122)
[I 2025-08-08 13:56:39,718] CategoricalLoader - INFO - Loading…
[I 2025-08-08 13:56:39,719] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-
data\input\categoricals.csv
[I 2025-08-08 13:56:42,924] CategoricalLoader - INFO - Completed
download -- 3.2042951583862305 seconds
[I 2025-08-08 13:56:42,925] CombinedLoader - INFO - Adding
(FinancialLoader-ScopeLoader + CategoricalLoader)
[I 2025-08-08 13:56:44,382] PreviousScopeFeaturesDataManager - INFO -
Added loader_categoricalloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:44,382] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_2 to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:44,383] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 126)
```

```
[I 2025-08-08 13:56:44,383] RegionLoader - INFO - Loading…
[I 2025-08-08 13:56:44,385] OnlineCSVDatasource - INFO - Fetching data
from https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-
Codes/master/all/all.csv
[I 2025-08-08 13:56:44,637] RegionLoader - INFO - Completed download --
0.2519216537475586 seconds
[I 2025-08-08 13:56:44,638] CombinedLoader - INFO - Adding
(FinancialLoader-ScopeLoader-CategoricalLoader + RegionLoader)
[I 2025-08-08 13:56:44,639] CombinedLoader - INFO - Merging special
loader RegionLoader to FinancialLoader-ScopeLoader-CategoricalLoader
[I 2025-08-08 13:56:46,313] PreviousScopeFeaturesDataManager - INFO -
Added loader_regionloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:46,314] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_3 to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:46,315] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 129)
[I 2025-08-08 13:56:46,316] PreviousScopeFeaturesDataManager - INFO -
Added merged to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:47,450] CompanyDataFilter - INFO - Filtered dataset
from 526241 to 526241 data points
[I 2025-08-08 13:56:47,454] PreviousScopeFeaturesDataManager - INFO -
Added reduced to PreviousScopeFeaturesDataManager
[I 2025-08-08 13:56:47,455] PreviousScopeFeaturesDataManager - INFO -
Taking all previous year scopes
100%|      | 103752/103752 [03:39<00:00, 472.79it/s]
[I 2025-08-08 14:00:28,702] PreviousScopeFeaturesDataManager - INFO -
Added original to PreviousScopeFeaturesDataManager
[I 2025-08-08 14:00:29,130] 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_num", axis=1)
categorical_features = df_scopes.filter(regex="^ft_cat", axis=1)
```

[ ]:
```
thresh = 0.5
from sklearn.impute import KNNImputer, SimpleImputer
```

[ ]:
```
correlations_original = numerical_features.corr()
correlations_original
```

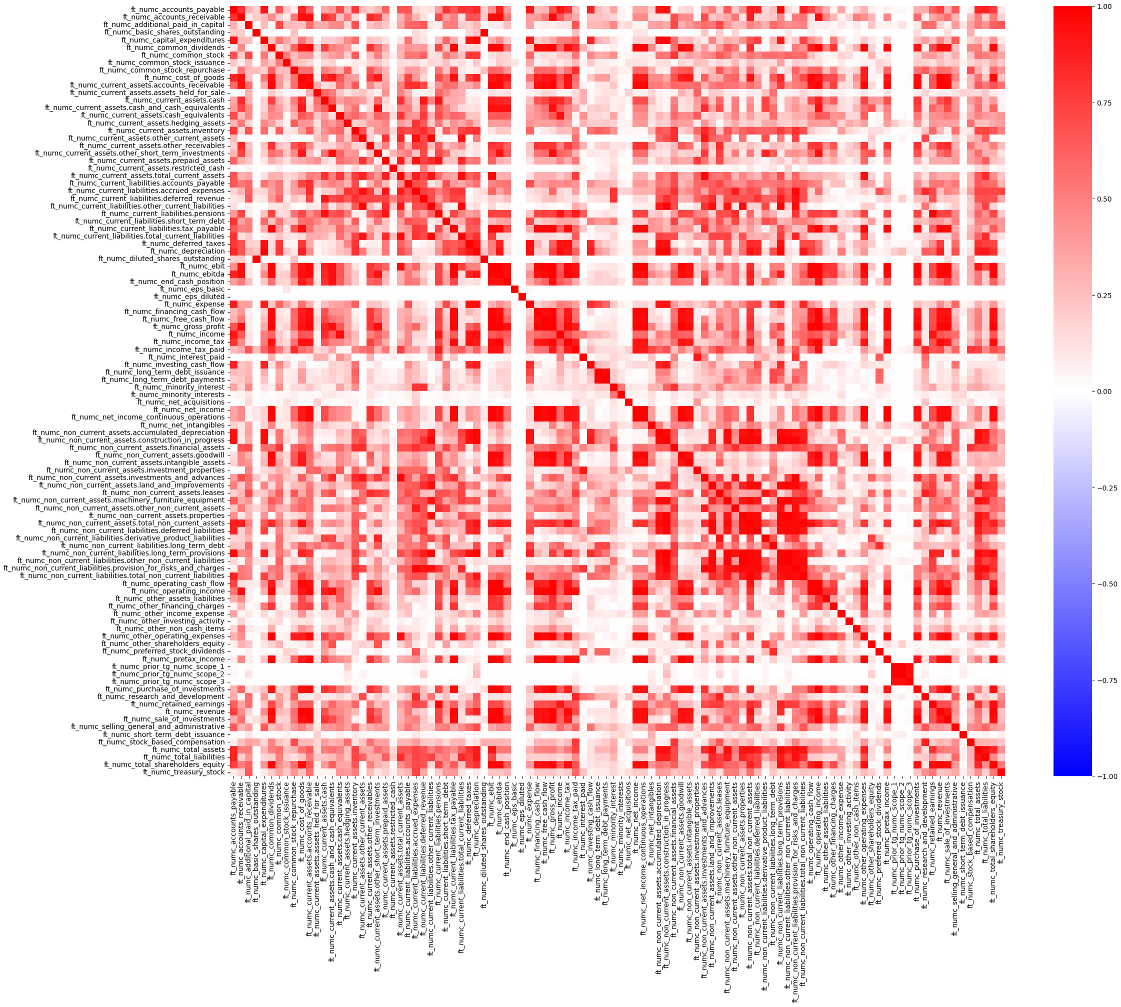
[ ]:
```
                      ft_numc_accounts_payable  ft_numc_accounts_receivable  \
ft_numc_accounts_…             1.000000                         -0.790710
```

```
ft_numc_accounts_…            -0.790710                  1.000000
ft_numc_additiona…             0.298424                 -0.089450
ft_numc_basic_sha…             0.184551                 -0.199001
ft_numc_capital_e…            -0.924032                  0.259657
…                                   …                        …
ft_numc_stock_bas…             0.061907                 -0.251039
ft_numc_total_assets             0.858600                 -0.509079
ft_numc_total_lia…             0.869143                 -0.189408
ft_numc_total_sha…             0.751994                 -0.523233
ft_numc_treasury_…             0.083284                 -0.185425

                      ft_numc_additional_paid_in_capital  …
ft_numc_total_liabilities  \
ft_numc_accounts_…             0.298424                 …
0.869143
ft_numc_accounts_…            -0.089450                 …
-0.189408
ft_numc_additiona…             1.000000                 …
0.415579
ft_numc_basic_sha…             0.004594                 …
0.001642
ft_numc_capital_e…            -0.263882                 …
-0.168674
…                                   …                 …
…
ft_numc_stock_bas…             0.352415                 …
0.301035
ft_numc_total_assets             0.523025                 …
0.930842
ft_numc_total_lia…             0.415579                 …
1.000000
ft_numc_total_sha…             0.616168                 …
0.465421
ft_numc_treasury_…             0.653782                 …
0.189129

                      ft_numc_total_shareholders_equity  ft_numc_treasury_stock
ft_numc_accounts_…             0.751994                           0.083284
ft_numc_accounts_…            -0.523233                          -0.185425
ft_numc_additiona…             0.616168                           0.653782
ft_numc_basic_sha…             0.005404                           0.002166
ft_numc_capital_e…            -0.415115                          -0.083936
…                                   …                                 …
ft_numc_stock_bas…             0.288338                           0.229405
ft_numc_total_assets             0.866850                           0.437628
ft_numc_total_lia…             0.465421                           0.189129
ft_numc_total_sha…             1.000000                           0.525521
```

```
ft_numc_treasury_…              0.525521                          1.000000

[102 rows x 102 columns]
```

[ ]:
```
plt.figure(figsize=(25, 20))
sns.heatmap(correlations_original.abs(), vmin=-1, vmax=1, cmap='bwr')
```

[ ]:
```
<Axes: >
```

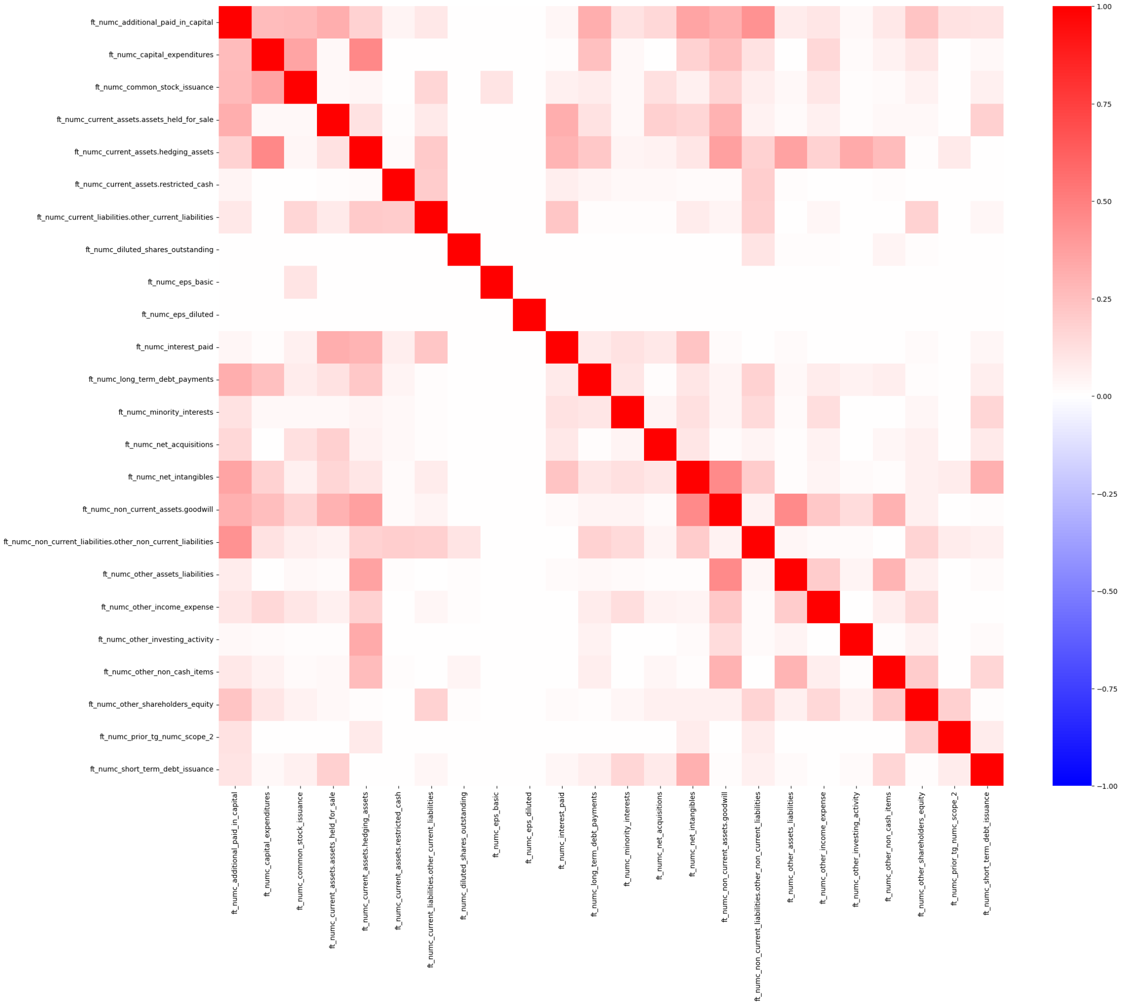


[ ]:
```
correlations = correlations_original.copy()
flag = True
while flag:
    highest_corrs = list(np.sum(correlations.abs() > thresh).sort_values().
 ↪items())[-1]
    if highest_corrs[1] < 2:
        break
    # print(f"Going to remove {highest_corrs}")
    correlations = correlations.drop(highest_corrs[0], axis=1).
  ↪drop(highest_corrs[0], axis=0)
    # display(correlations)
print('Iterative elimination\n')
print(f"features_iterative_corr_elimination = {list(correlations.columns)}")
```

```
Iterative elimination

features_iterative_corr_elimination = ['ft_numc_additional_paid_in_capital',
'ft_numc_capital_expenditures', 'ft_numc_common_stock_issuance',
'ft_numc_current_assets.assets_held_for_sale',
'ft_numc_current_assets.hedging_assets',
'ft_numc_current_assets.restricted_cash',
'ft_numc_current_liabilities.other_current_liabilities',
'ft_numc_diluted_shares_outstanding', 'ft_numc_eps_basic',
'ft_numc_eps_diluted', 'ft_numc_interest_paid',
'ft_numc_long_term_debt_payments', 'ft_numc_minority_interests',
'ft_numc_net_acquisitions', 'ft_numc_net_intangibles',
'ft_numc_non_current_assets.goodwill',
'ft_numc_non_current_liabilities.other_non_current_liabilities',
'ft_numc_other_assets_liabilities', 'ft_numc_other_income_expense',
'ft_numc_other_investing_activity', 'ft_numc_other_non_cash_items',
'ft_numc_other_shareholders_equity', 'ft_numc_prior_tg_numc_scope_2',
'ft_numc_short_term_debt_issuance']
```

```
[ ]: plt.figure(figsize=(25, 20))
     sns.heatmap(correlations.abs(), vmin=-1, vmax=1, cmap='bwr')
```

```
[ ]: <Axes: >
```

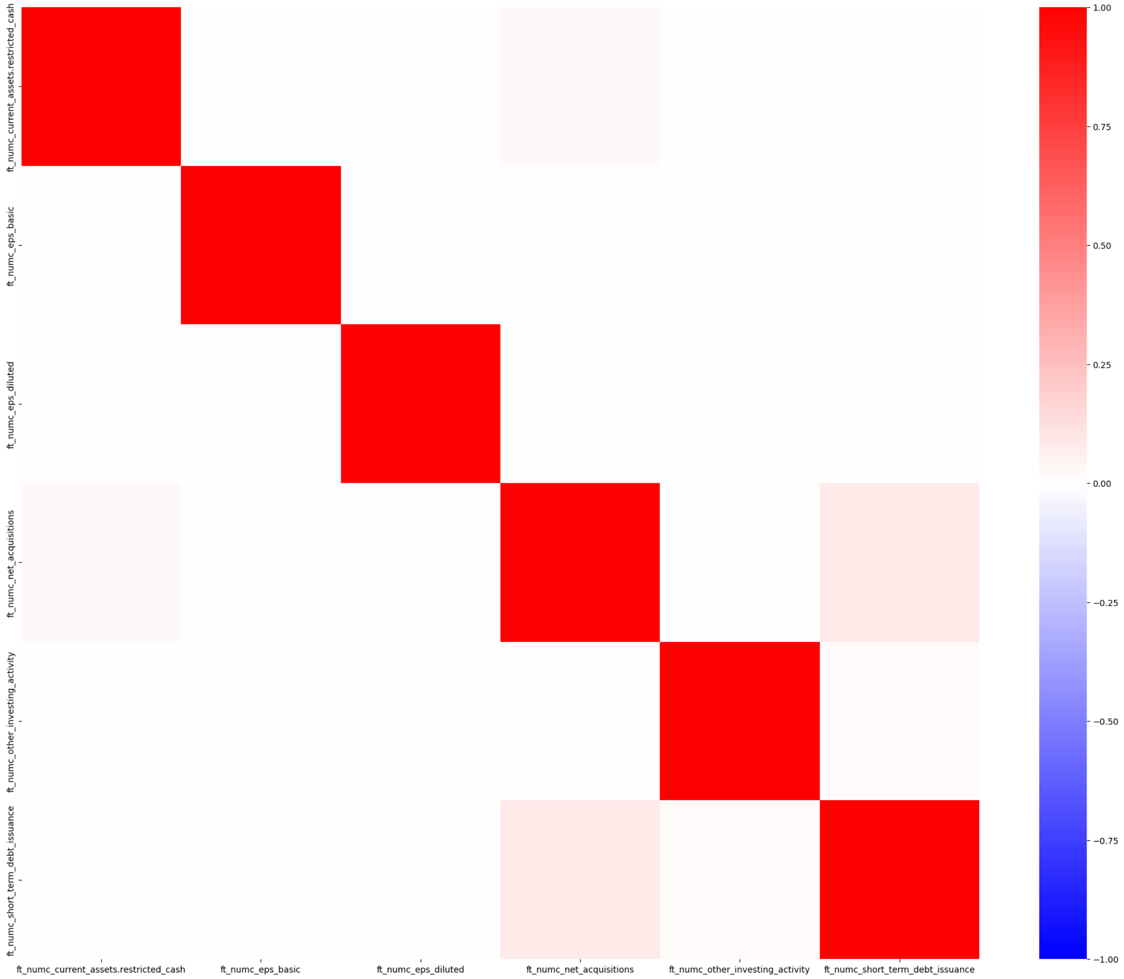


```
[ ]: correlations_strict = correlations_original.copy()
     l_highest_corrs = list(np.sum(correlations_strict.abs() > thresh).
      ↪sort_values(ascending=0).items())
     reversed(l_highest_corrs)
     for key, val in l_highest_corrs:
         if val > 1:
             # print(f"Going to remove {(key, val)}")
             correlations_strict = correlations_strict.drop(key, axis=1).drop(key,␣
      ↪axis=0)

     print('Strict elimination\n')
     print(f"features_strict_corr_elimination = {list(correlations_strict.columns)}")
```

```
Strict elimination

features_strict_corr_elimination = ['ft_numc_current_assets.restricted_cash',
'ft_numc_eps_basic', 'ft_numc_eps_diluted', 'ft_numc_net_acquisitions',
'ft_numc_other_investing_activity', 'ft_numc_short_term_debt_issuance']
```

```
[ ]: plt.figure(figsize=(25, 20))
     sns.heatmap(correlations_strict.abs(), vmin=-1, vmax=1, cmap='bwr')
```

```
[ ]: <Axes: >
```



```
[ ]: numerical_features = pd.DataFrame(SimpleImputer(strategy='median').
      ↪fit_transform(numerical_features), columns=numerical_features.columns,␣
      ↪index=numerical_features.index)
     numerical_features
```

```
[ ]:        ft_numc_accounts_payable  ft_numc_accounts_receivable
     ft_numc_additional_paid_in_capital  \
     0                      517000.0                    -702711.0
     1.107616e+07
```

```
1                  517000.0                 -702711.0
8.989978e+05
2                  517000.0                 -702711.0
-3.667970e+06
3                  517000.0                 -702711.0
-3.786688e+06
4                  517000.0                 -702711.0
1.107616e+07
…                     …                       …
…
526238             517000.0                 -702711.0
4.338612e+07
526239             517000.0                 -702711.0
3.504908e+07
526236             517000.0                 -702711.0
3.738345e+07
526237             517000.0                 -702711.0
4.255846e+07
526240             517000.0                 -702711.0
5.076744e+07

        …  ft_numc_total_liabilities  ft_numc_total_shareholders_equity
ft_numc_treasury_stock
0       …         1.492148e+08              7.779242e+07
13106995.0
1       …         7.654873e+08              4.072673e+08
13106995.0
2       …         1.118626e+09              4.894674e+08
13106995.0
3       …         1.860606e+09              6.209250e+08
13106995.0
4       …         1.492148e+08              7.779242e+07
13106995.0
…     …                   …                        …
…
526238  …         3.070824e+10              8.729744e+09
13106995.0
526239  …         1.737979e+08              1.221721e+10
13106995.0
526236  …         1.737979e+08              1.485952e+10
13106995.0
526237  …         1.737979e+08              1.652636e+10
13106995.0
526240  …         1.737979e+08              1.767752e+10
13106995.0

[522776 rows x 102 columns]
```

[ ]:

```
# https://www.projectpro.io/recipes/
 ↪drop-out-highly-correlated-features-in-python

from sklearn.feature_selection import SelectKBest

from sklearn.feature_selection import f_regression

# define number of features to keep

k = 10

# perform feature selection
y = DATA['tg_numc_scope_1']
selector = SelectKBest(f_regression, k=k).fit(numerical_features[~y.isna()],␣
 ↪y[~y.isna()])
X_new = selector.transform(numerical_features)
# get feature names of selected features

selected_features = numerical_features.columns[selector.get_support()]

# print selected features
print('SelectKBest elimination\n')
print(f"features_select_k_best = {list(selected_features)}")
```

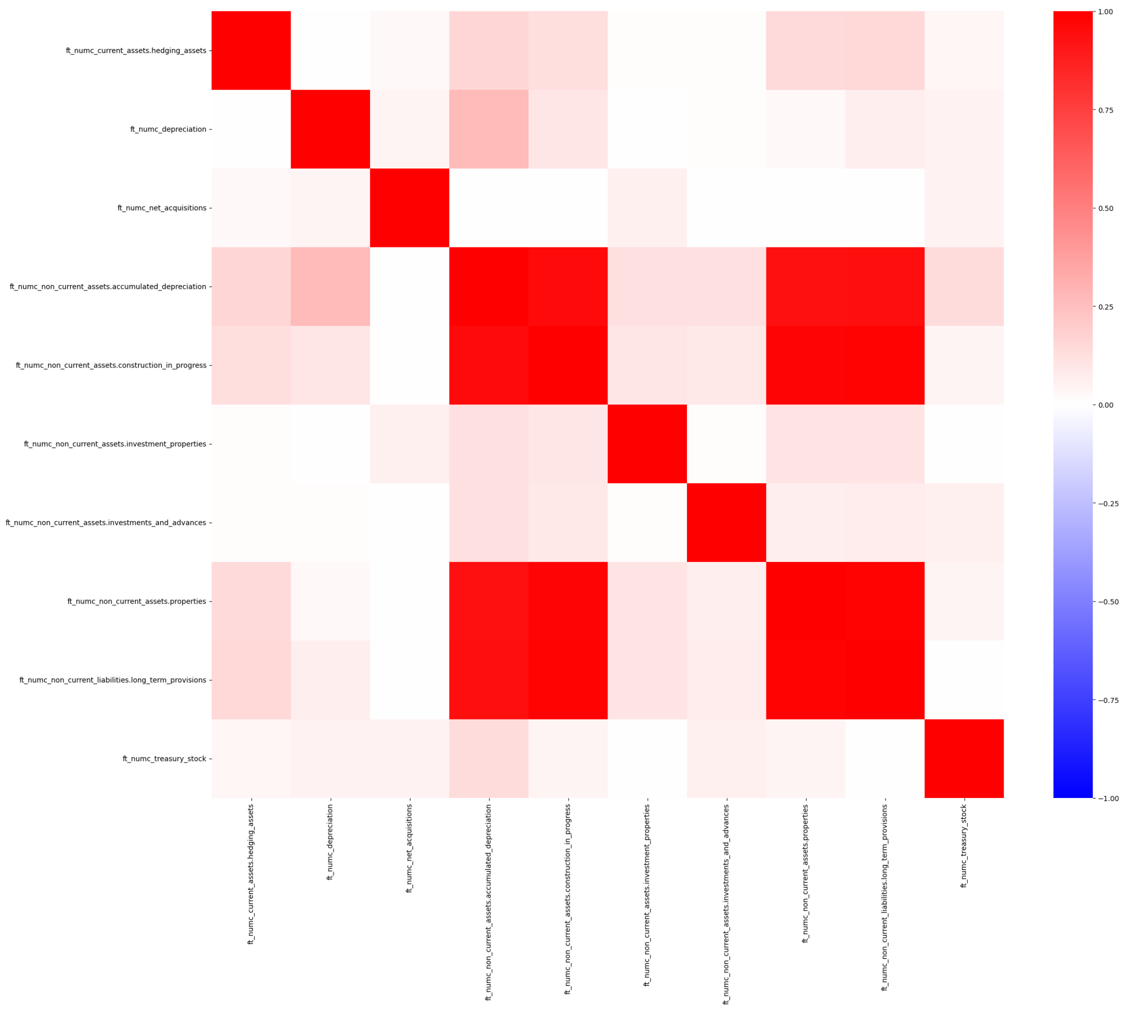
```
SelectKBest elimination

features_select_k_best = ['ft_numc_current_assets.hedging_assets',
'ft_numc_depreciation', 'ft_numc_net_acquisitions',
'ft_numc_non_current_assets.accumulated_depreciation',
'ft_numc_non_current_assets.construction_in_progress',
'ft_numc_non_current_assets.investment_properties',
'ft_numc_non_current_assets.investments_and_advances',
'ft_numc_non_current_assets.properties',
'ft_numc_non_current_liabilities.long_term_provisions',
'ft_numc_treasury_stock']
```

[ ]:

```
plt.figure(figsize=(25, 20))
sns.heatmap(numerical_features[selected_features].corr().abs(), vmin=-1,␣
 ↪vmax=1, cmap='bwr')
```

[ ]:

```
<Axes: >
```



[ ]:

```python
from statsmodels.stats.outliers_influence import variance_inflation_factor
from tqdm import tqdm
# calculate VIF for each feature

vif = pd.DataFrame()

vif["VIF Factor"] = [variance_inflation_factor(numerical_features, i) for i in tqdm(range(numerical_features.shape[1]))]

vif["features"] = numerical_features.columns

# print VIF values
```

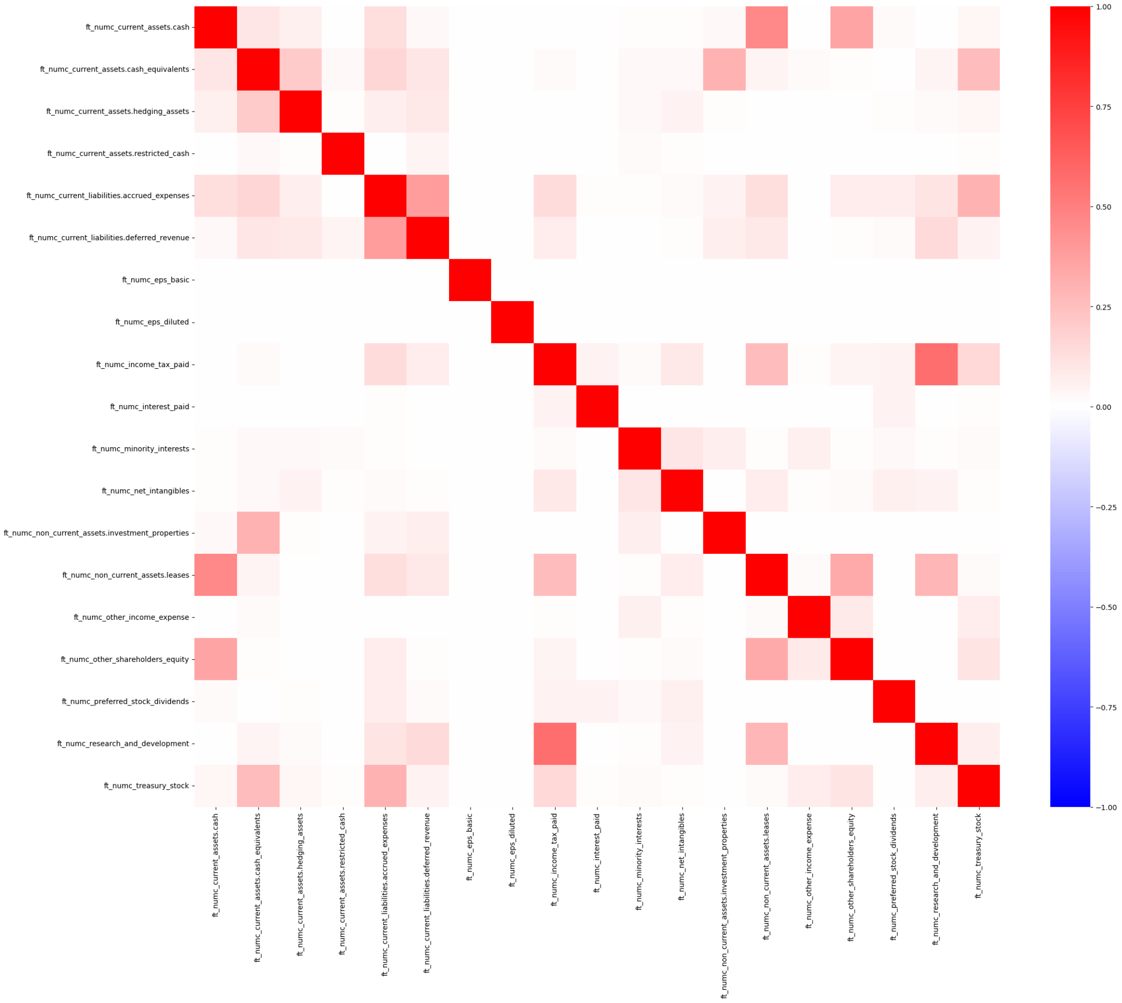
```
100%|          | 102/102 [29:08<00:00, 17.14s/it]
```

```
[ ]: print('VIF elimination\n')
     print(f"features_VIF_under_5 = {vif[vif['VIF Factor'] < 5].features.tolist()}")
```

```
VIF elimination

features_VIF_under_5 = ['ft_numc_current_assets.cash',
'ft_numc_current_assets.cash_equivalents',
'ft_numc_current_assets.hedging_assets',
'ft_numc_current_assets.restricted_cash',
'ft_numc_current_liabilities.accrued_expenses',
'ft_numc_current_liabilities.deferred_revenue', 'ft_numc_eps_basic',
'ft_numc_eps_diluted', 'ft_numc_income_tax_paid', 'ft_numc_interest_paid',
'ft_numc_minority_interests', 'ft_numc_net_intangibles',
'ft_numc_non_current_assets.investment_properties',
'ft_numc_non_current_assets.leases', 'ft_numc_other_income_expense',
'ft_numc_other_shareholders_equity', 'ft_numc_preferred_stock_dividends',
'ft_numc_research_and_development', 'ft_numc_treasury_stock']
```

```
[ ]: plt.figure(figsize=(25, 20))
     sns.heatmap(numerical_features[vif[vif["VIF Factor"] < 5].features.tolist()].
      ↪corr().abs(), vmin=-1, vmax=1, cmap='bwr')
```

```
[ ]: <Axes: >
```



```
[ ]: print('VIF elimination\n')
     print(f"features_VIF_under_10 = {vif[vif['VIF Factor'] < 10].features.
      ↪tolist()}")
```

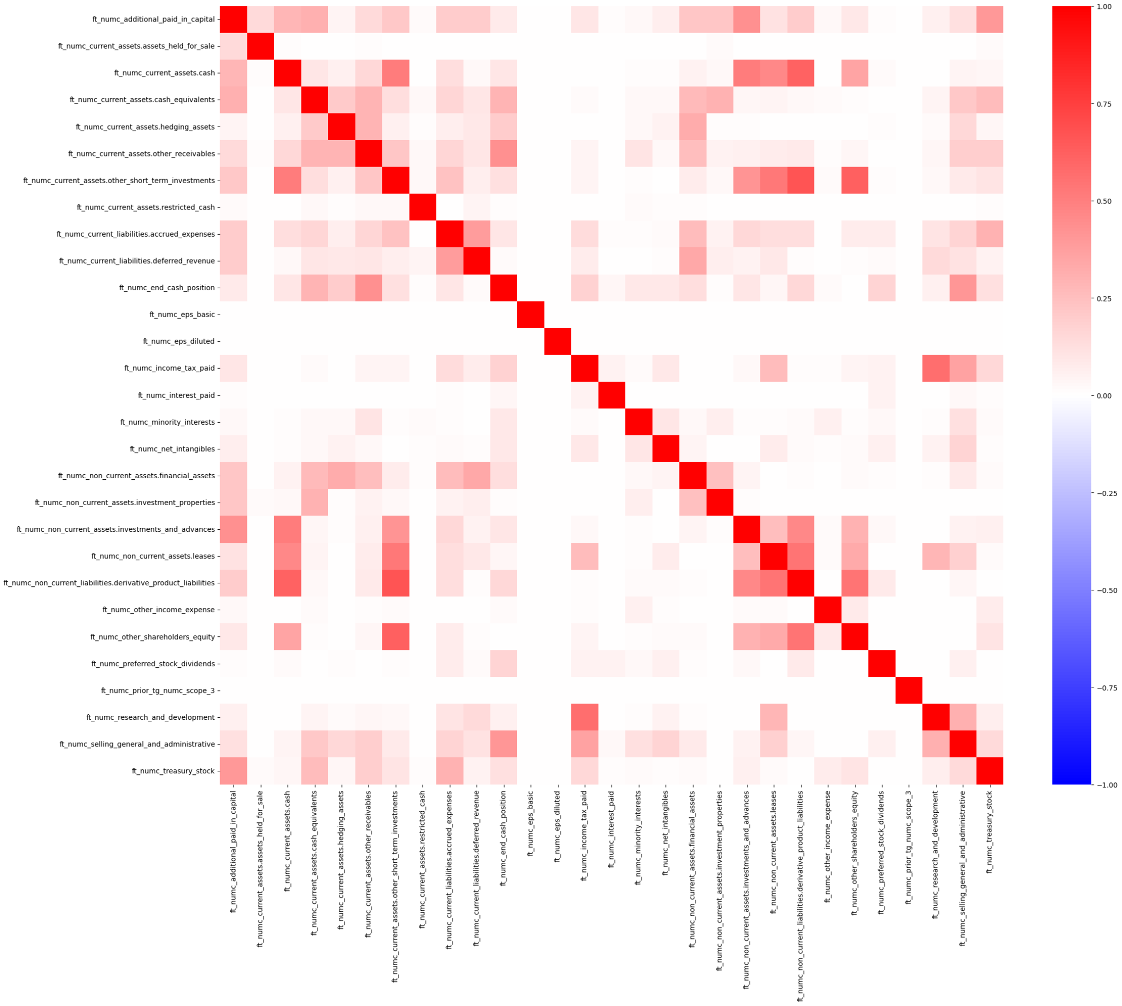
```
VIF elimination

features_VIF_under_10 = ['ft_numc_additional_paid_in_capital',
'ft_numc_current_assets.assets_held_for_sale', 'ft_numc_current_assets.cash',
'ft_numc_current_assets.cash_equivalents',
'ft_numc_current_assets.hedging_assets',
'ft_numc_current_assets.other_receivables',
'ft_numc_current_assets.other_short_term_investments',
'ft_numc_current_assets.restricted_cash',
'ft_numc_current_liabilities.accrued_expenses',
'ft_numc_current_liabilities.deferred_revenue', 'ft_numc_end_cash_position',
'ft_numc_eps_basic', 'ft_numc_eps_diluted', 'ft_numc_income_tax_paid',
'ft_numc_interest_paid', 'ft_numc_minority_interests',
'ft_numc_net_intangibles', 'ft_numc_non_current_assets.financial_assets',
'ft_numc_non_current_assets.investment_properties',
'ft_numc_non_current_assets.investments_and_advances',
'ft_numc_non_current_assets.leases',
'ft_numc_non_current_liabilities.derivative_product_liabilities',
'ft_numc_other_income_expense', 'ft_numc_other_shareholders_equity',
'ft_numc_preferred_stock_dividends', 'ft_numc_prior_tg_numc_scope_3',
'ft_numc_research_and_development',
'ft_numc_selling_general_and_administrative', 'ft_numc_treasury_stock']
```

[ ]:

```
plt.figure(figsize=(25, 20))
sns.heatmap(numerical_features[vif[vif["VIF Factor"] < 10].features.tolist()].
 ↪corr().abs(), vmin=-1, vmax=1, cmap='bwr')
```

[ ]:

```
<Axes: >
```



[ ]:

```
print('VIF elimination\n')
print(f"features_VIF_under_10 = {vif[vif['VIF Factor'] < 15].features.
 ↪tolist()}")
```

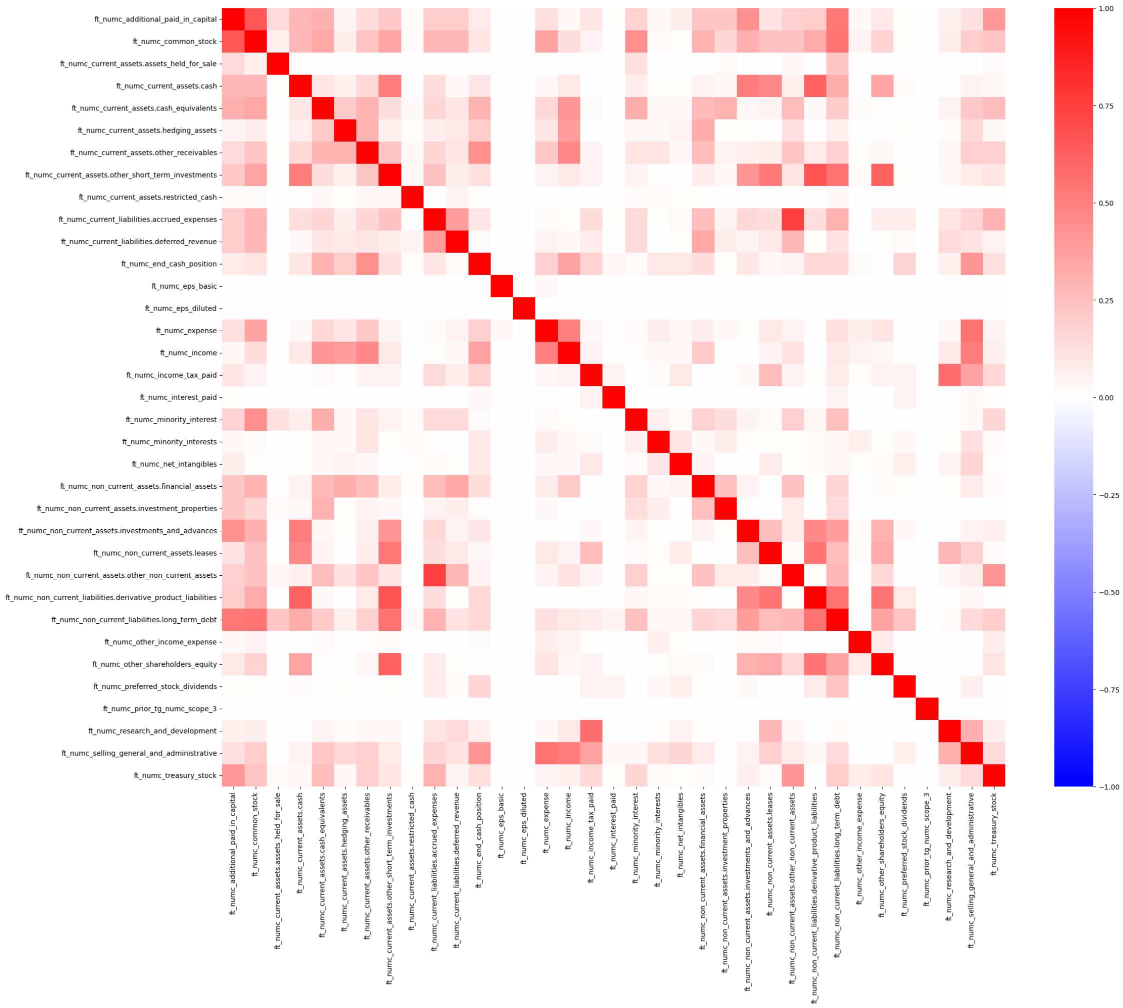
```
VIF elimination

features_VIF_under_10 = ['ft_numc_additional_paid_in_capital',
'ft_numc_common_stock', 'ft_numc_current_assets.assets_held_for_sale',
'ft_numc_current_assets.cash', 'ft_numc_current_assets.cash_equivalents',
'ft_numc_current_assets.hedging_assets',
'ft_numc_current_assets.other_receivables',
'ft_numc_current_assets.other_short_term_investments',
'ft_numc_current_assets.restricted_cash',
'ft_numc_current_liabilities.accrued_expenses',
'ft_numc_current_liabilities.deferred_revenue', 'ft_numc_end_cash_position',
'ft_numc_eps_basic', 'ft_numc_eps_diluted', 'ft_numc_expense', 'ft_numc_income',
'ft_numc_income_tax_paid', 'ft_numc_interest_paid', 'ft_numc_minority_interest',
'ft_numc_minority_interests', 'ft_numc_net_intangibles',
'ft_numc_non_current_assets.financial_assets',
'ft_numc_non_current_assets.investment_properties',
'ft_numc_non_current_assets.investments_and_advances',
'ft_numc_non_current_assets.leases',
'ft_numc_non_current_assets.other_non_current_assets',
'ft_numc_non_current_liabilities.derivative_product_liabilities',
'ft_numc_non_current_liabilities.long_term_debt',
'ft_numc_other_income_expense', 'ft_numc_other_shareholders_equity',
'ft_numc_preferred_stock_dividends', 'ft_numc_prior_tg_numc_scope_3',
'ft_numc_research_and_development',
'ft_numc_selling_general_and_administrative', 'ft_numc_treasury_stock']
```

[ ]:

```
plt.figure(figsize=(25, 20))
sns.heatmap(numerical_features[vif[vif["VIF Factor"] < 15].features.tolist()].
 ↪corr().abs(), vmin=-1, vmax=1, cmap='bwr')
```

[ ]:

```
<Axes: >
```

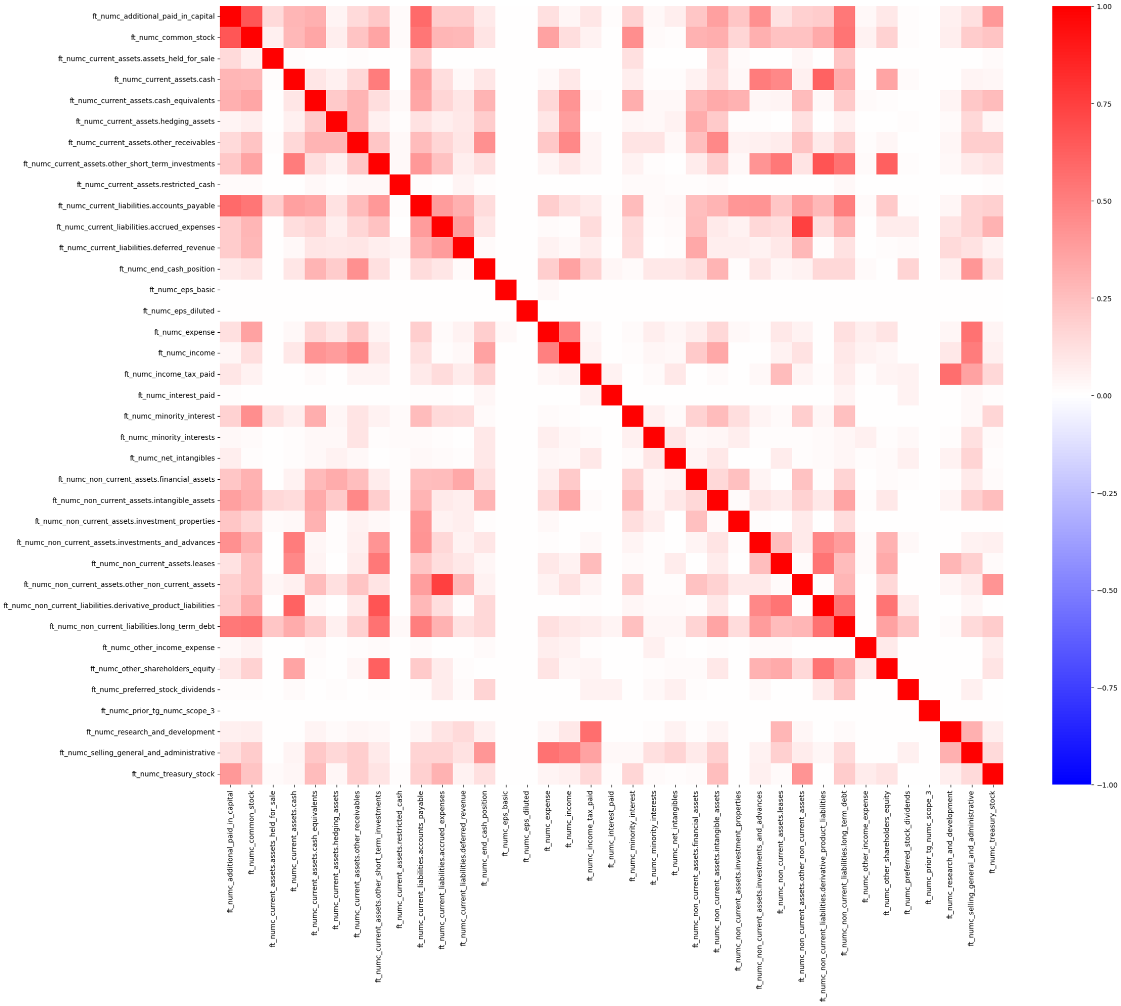


[ ]:
```
print('VIF elimination\n')
print(f"features_VIF_under_10 = {vif[vif['VIF Factor'] < 20].features.
  ↪tolist()}")
```

```
VIF elimination

features_VIF_under_10 = ['ft_numc_additional_paid_in_capital',
'ft_numc_common_stock', 'ft_numc_current_assets.assets_held_for_sale',
'ft_numc_current_assets.cash', 'ft_numc_current_assets.cash_equivalents',
'ft_numc_current_assets.hedging_assets',
'ft_numc_current_assets.other_receivables',
'ft_numc_current_assets.other_short_term_investments',
'ft_numc_current_assets.restricted_cash',
'ft_numc_current_liabilities.accounts_payable',
'ft_numc_current_liabilities.accrued_expenses',
'ft_numc_current_liabilities.deferred_revenue', 'ft_numc_end_cash_position',
'ft_numc_eps_basic', 'ft_numc_eps_diluted', 'ft_numc_expense', 'ft_numc_income',
'ft_numc_income_tax_paid', 'ft_numc_interest_paid', 'ft_numc_minority_interest',
'ft_numc_minority_interests', 'ft_numc_net_intangibles',
'ft_numc_non_current_assets.financial_assets',
'ft_numc_non_current_assets.intangible_assets',
'ft_numc_non_current_assets.investment_properties',
'ft_numc_non_current_assets.investments_and_advances',
'ft_numc_non_current_assets.leases',
'ft_numc_non_current_assets.other_non_current_assets',
'ft_numc_non_current_liabilities.derivative_product_liabilities',
'ft_numc_non_current_liabilities.long_term_debt',
'ft_numc_other_income_expense', 'ft_numc_other_shareholders_equity',
'ft_numc_preferred_stock_dividends', 'ft_numc_prior_tg_numc_scope_3',
'ft_numc_research_and_development',
'ft_numc_selling_general_and_administrative', 'ft_numc_treasury_stock']
```

```
[ ]: plt.figure(figsize=(25, 20))
     sns.heatmap(numerical_features[vif[vif["VIF Factor"] < 20].features.tolist()].
      ↪corr().abs(), vmin=-1, vmax=1, cmap='bwr')
```

```
[ ]: <Axes: >
```



[ ]:

```
print('VIF elimination\n')
print(f"features_VIF_under_10 = {vif[vif['VIF Factor'] < 25].features.
  ↪tolist()}")
```

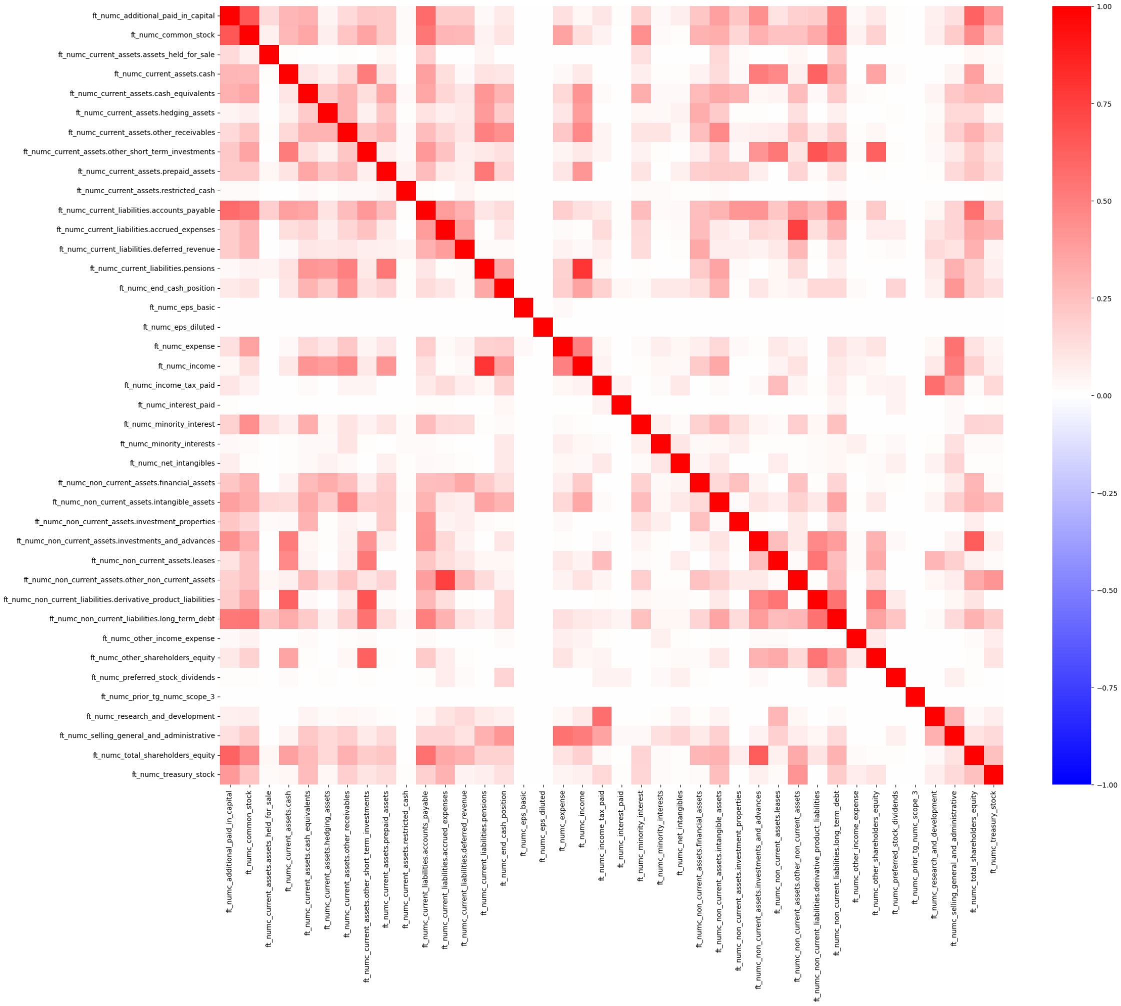
```
VIF elimination

features_VIF_under_10 = ['ft_numc_additional_paid_in_capital',
'ft_numc_common_stock', 'ft_numc_current_assets.assets_held_for_sale',
'ft_numc_current_assets.cash', 'ft_numc_current_assets.cash_equivalents',
'ft_numc_current_assets.hedging_assets',
'ft_numc_current_assets.other_receivables',
'ft_numc_current_assets.other_short_term_investments',
'ft_numc_current_assets.prepaid_assets',
'ft_numc_current_assets.restricted_cash',
'ft_numc_current_liabilities.accounts_payable',
'ft_numc_current_liabilities.accrued_expenses',
'ft_numc_current_liabilities.deferred_revenue',
'ft_numc_current_liabilities.pensions', 'ft_numc_end_cash_position',
'ft_numc_eps_basic', 'ft_numc_eps_diluted', 'ft_numc_expense', 'ft_numc_income',
'ft_numc_income_tax_paid', 'ft_numc_interest_paid', 'ft_numc_minority_interest',
'ft_numc_minority_interests', 'ft_numc_net_intangibles',
'ft_numc_non_current_assets.financial_assets',
'ft_numc_non_current_assets.intangible_assets',
'ft_numc_non_current_assets.investment_properties',
'ft_numc_non_current_assets.investments_and_advances',
'ft_numc_non_current_assets.leases',
'ft_numc_non_current_assets.other_non_current_assets',
'ft_numc_non_current_liabilities.derivative_product_liabilities',
'ft_numc_non_current_liabilities.long_term_debt',
'ft_numc_other_income_expense', 'ft_numc_other_shareholders_equity',
'ft_numc_preferred_stock_dividends', 'ft_numc_prior_tg_numc_scope_3',
'ft_numc_research_and_development',
'ft_numc_selling_general_and_administrative',
'ft_numc_total_shareholders_equity', 'ft_numc_treasury_stock']
```

[ ]:

```
plt.figure(figsize=(25, 20))
sns.heatmap(numerical_features[vif[vif["VIF Factor"] < 25].features.tolist()].
 ↪corr().abs(), vmin=-1, vmax=1, cmap='bwr')
```

[ ]:

```
<Axes: >
```



```
[ ]: import json
     import io
     json.dump(vif[vif["VIF Factor"] < 5].features.tolist()+categorical_features.
      ↪columns.tolist(), io.open(PARENT_PATH+'/res/vif_05.json', 'w'), indent=2)
     json.dump(vif[vif["VIF Factor"] < 10].features.tolist()+categorical_features.
      ↪columns.tolist(), io.open(PARENT_PATH+'/res/vif_10.json', 'w'), indent=2)
     json.dump(vif[vif["VIF Factor"] < 15].features.tolist()+categorical_features.
      ↪columns.tolist(), io.open(PARENT_PATH+'/res/vif_15.json', 'w'), indent=2)
     json.dump(vif[vif["VIF Factor"] < 20].features.tolist()+categorical_features.
      ↪columns.tolist(), io.open(PARENT_PATH+'/res/vif_20.json', 'w'), indent=2)
     json.dump(vif[vif["VIF Factor"] < 25].features.tolist()+categorical_features.
      ↪columns.tolist(), io.open(PARENT_PATH+'/res/vif_25.json', 'w'), indent=2)
```

```
[ ]: print('Corr Matrix\n')
```

```
Corr Matrix
```

[ ]:

```
corr_matrix = numerical_features[vif.features.tolist()].corr().abs()
corr_matrix
```

[ ]:

```
                            ft_numc_accounts_payable  ft_numc_accounts_receivable  \
ft_numc_accounts_…                  1.000000                     0.076831
ft_numc_accounts_…                  0.076831                     1.000000
ft_numc_additiona…                  0.057491                     0.019829
ft_numc_basic_sha…                  0.004922                     0.007751
ft_numc_capital_e…                  0.790640                     0.169406
…                                        …                            …
ft_numc_stock_bas…                  0.056203                     0.107678
ft_numc_total_assets                  0.050635                     0.085425
ft_numc_total_lia…                  0.059782                     0.052785
ft_numc_total_sha…                  0.040261                     0.141104
ft_numc_treasury_…                  0.033019                     0.040536

                            ft_numc_additional_paid_in_capital  …
ft_numc_total_liabilities  \
ft_numc_accounts_…                  0.057491                  …
0.059782
ft_numc_accounts_…                  0.019829                  …
0.052785
ft_numc_additiona…                  1.000000                  …
0.409724
ft_numc_basic_sha…                  0.001626                  …
0.000959
ft_numc_capital_e…                  0.082047                  …
0.100357
…                                        …                  …
…
ft_numc_stock_bas…                  0.052266                  …
0.038864
ft_numc_total_assets                  0.392316                  …
0.891294
ft_numc_total_lia…                  0.409724                  …
1.000000
ft_numc_total_sha…                  0.614992                  …
0.454531
ft_numc_treasury_…                  0.398818                  …
0.130252

                            ft_numc_total_shareholders_equity  ft_numc_treasury_stock
ft_numc_accounts_…                  0.040261                            0.033019
ft_numc_accounts_…                  0.141104                            0.040536
ft_numc_additiona…                  0.614992                            0.398818
```

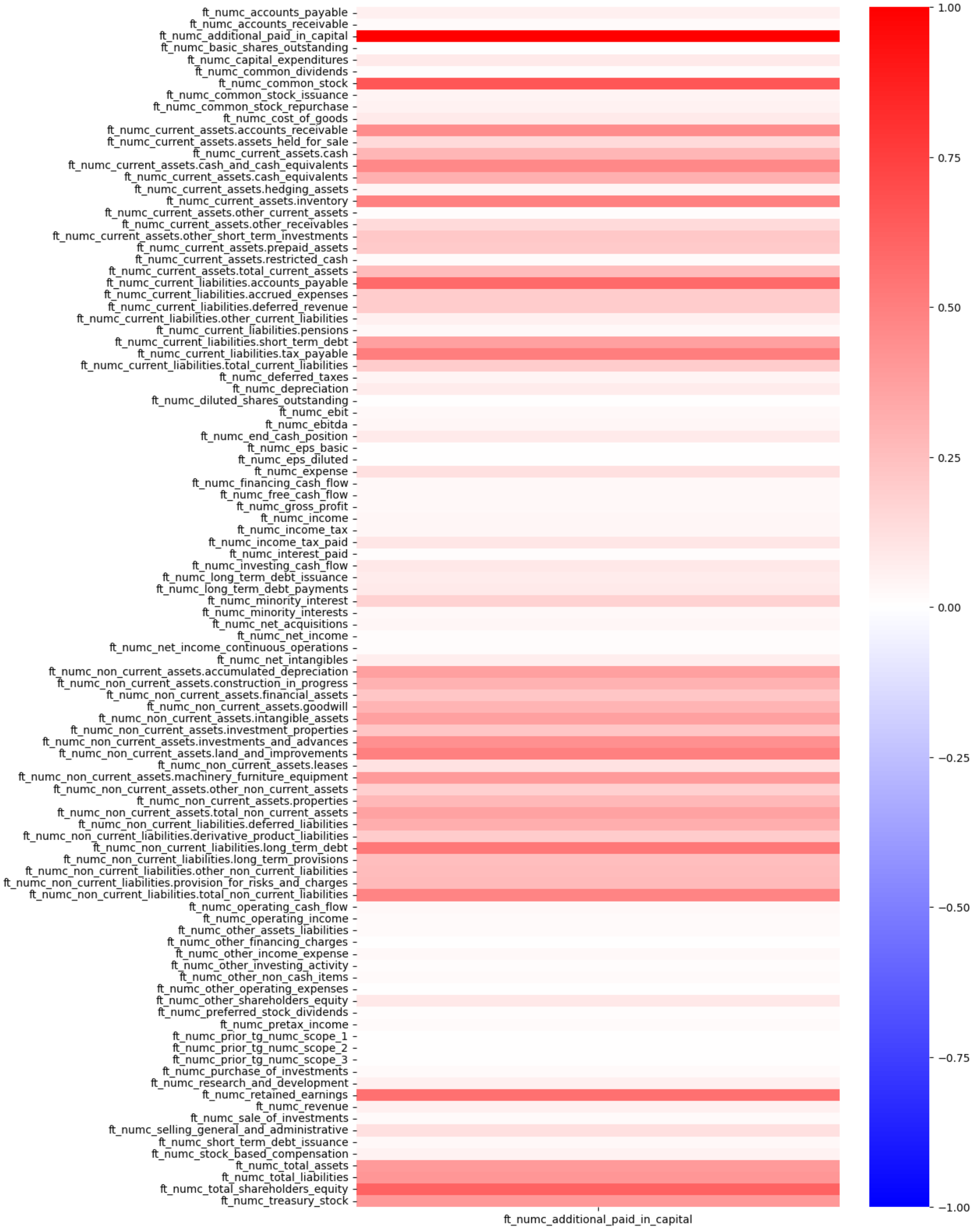
```
ft_numc_basic_sha…         0.001479                          0.000885
ft_numc_capital_e…         0.113998                          0.035130
…                               …                                 …
ft_numc_stock_bas…         0.075882                          0.050916
ft_numc_total_assets         0.529527                          0.210104
ft_numc_total_lia…         0.454531                          0.130252
ft_numc_total_sha…         1.000000                          0.243658
ft_numc_treasury_…         0.243658                          1.000000

[102 rows x 102 columns]
```

[ ]:

```
plt.figure(figsize=(10, 20))
sns.heatmap(corr_matrix[["ft_numc_additional_paid_in_capital"]], vmin=-1,␣
 ↪vmax=1, cmap='bwr')



# # %%
# from sklearn.feature_selection import RFECV
# from xgboost import XGBRegressor
# from sklearn.svm import SVR
# from sklearn.neighbors import KNeighborsRegressor
# from sklearn.ensemble import RandomForestRegressor

# estimator = RandomForestRegressor()
# selector = RFECV(estimator, step=0.1, cv=10, verbose=True)
# selector = selector.fit(numerical_features[~y.isna()], y[~y.isna()])

# # %%

# plt.figure(figsize=(25, 20))
# sns.heatmap(numerical_features.iloc[:, selector.support_].corr().abs(),␣
 ↪vmin=-1, vmax=1, cmap='bwr')

# # %%
```

[ ]:

```
<Axes: >
```



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
ft_numc_current_assets.cash_and_cash_equivalents
ft_numc_current_assets.cash_equivalents
ft_numc_current_assets.hedging_assets
ft_numc_current_assets.inventory
ft_numc_current_assets.other_current_assets
ft_numc_current_assets.other_receivables
ft_numc_current_assets.other_short_term_investments
ft_numc_current_assets.prepaid_assets
ft_numc_current_assets.restricted_cash
ft_numc_current_assets.total_current_assets
ft_numc_current_liabilities.accounts_payable
ft_numc_current_liabilities.accrued_expenses
ft_numc_current_liabilities.deferred_revenue
ft_numc_current_liabilities.other_current_liabilities
ft_numc_current_liabilities.pensions
ft_numc_current_liabilities.short_term_debt
ft_numc_current_liabilities.tax_payable
ft_numc_current_liabilities.total_current_liabilities
ft_numc_deferred_taxes
ft_numc_depreciation
ft_numc_diluted_shares_outstanding
ft_numc_ebit
ft_numc_ebitda
ft_numc_end_cash_position
ft_numc_eps_basic
ft_numc_eps_diluted
ft_numc_expense
ft_numc_financing_cash_flow
ft_numc_free_cash_flow
ft_numc_gross_profit
ft_numc_income
ft_numc_income_tax
ft_numc_income_tax_paid
ft_numc_interest_paid
ft_numc_investing_cash_flow
ft_numc_long_term_debt_issuance
ft_numc_long_term_debt_payments
ft_numc_minority_interest
ft_numc_minority_interests
ft_numc_net_acquisitions
ft_numc_net_income
ft_numc_net_income_continuous_operations
ft_numc_net_intangibles
ft_numc_non_current_assets.accumulated_depreciation
ft_numc_non_current_assets.construction_in_progress
ft_numc_non_current_assets.financial_assets
ft_numc_non_current_assets.goodwill
ft_numc_non_current_assets.intangible_assets
ft_numc_non_current_assets.investment_properties
ft_numc_non_current_assets.investments_and_advances
ft_numc_non_current_assets.land_and_improvements
ft_numc_non_current_assets.leases
ft_numc_non_current_assets.machinery_furniture_equipment
ft_numc_non_current_assets.other_non_current_assets
ft_numc_non_current_assets.properties
ft_numc_non_current_assets.total_non_current_assets
ft_numc_non_current_liabilities.deferred_liabilities
ft_numc_non_current_liabilities.derivative_product_liabilities
ft_numc_non_current_liabilities.long_term_debt
ft_numc_non_current_liabilities.long_term_provisions
ft_numc_non_current_liabilities.other_non_current_liabilities
ft_numc_non_current_liabilities.provision_for_risks_and_charges
ft_numc_non_current_liabilities.total_non_current_liabilities
ft_numc_operating_cash_flow
ft_numc_operating_income
ft_numc_other_assets_liabilities
ft_numc_other_financing_charges
ft_numc_other_income_expense
ft_numc_other_investing_activity
ft_numc_other_non_cash_items
ft_numc_other_operating_expenses
ft_numc_other_shareholders_equity
ft_numc_preferred_stock_dividends
ft_numc_pretax_income
ft_numc_prior_tg_numc_scope_1
ft_numc_prior_tg_numc_scope_2
ft_numc_prior_tg_numc_scope_3
ft_numc_purchase_of_investments
ft_numc_research_and_development
ft_numc_retained_earnings
ft_numc_revenue
ft_numc_sale_of_investments
ft_numc_selling_general_and_administrative
ft_numc_short_term_debt_issuance
ft_numc_stock_based_compensation
ft_numc_total_assets
ft_numc_total_liabilities
ft_numc_total_shareholders_equity
ft_numc_treasury_stock
ft_numc_additional_paid_in_capital
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00