analyze_feature_missingness

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
import missingno as msno
import matplotlib.patches as mpatches
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 01:44:58,629] PreviousScopeFeaturesDataManager - INFO -
Remaining data points 0
[I 2025-08-08 01:44:58,631] FinancialLoader - INFO - Loading…
[I 2025-08-08 01:44:58,631] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-
data\input\financials.csv
[I 2025-08-08 01:45:07,678] FinancialLoader - INFO - Completed download
-- 9.04696249961853 seconds
[I 2025-08-08 01:45:07,679] PreviousScopeFeaturesDataManager - INFO -
Added loader_financialloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:07,680] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_0 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:07,681] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 110)
[I 2025-08-08 01:45:07,682] ScopeLoader - INFO - Loading…
[I 2025-08-08 01:45:07,683] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-data\input\scopes.csv
[I 2025-08-08 01:45:08,070] ScopeLoader - INFO - Completed download --
0.38664770126342773 seconds
[I 2025-08-08 01:45:08,071] CombinedLoader - INFO - Adding
(FinancialLoader + ScopeLoader)
[I 2025-08-08 01:45:08,072] CombinedLoader - INFO - Merging special
loader ScopeLoader to FinancialLoader
[I 2025-08-08 01:45:09,503] PreviousScopeFeaturesDataManager - INFO -
Added loader_scopeloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:09,505] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_1 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:09,506] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 122)
[I 2025-08-08 01:45:09,507] CategoricalLoader - INFO - Loading…
[I 2025-08-08 01:45:09,510] LocalDatasource - INFO - Fetching data from
C:\Users\User\Workspace\work_oxari\architectura\model-
data\input\categoricals.csv
[I 2025-08-08 01:45:12,260] CategoricalLoader - INFO - Completed
download -- 2.750828742980957 seconds
[I 2025-08-08 01:45:12,261] CombinedLoader - INFO - Adding
(FinancialLoader-ScopeLoader + CategoricalLoader)
[I 2025-08-08 01:45:13,848] PreviousScopeFeaturesDataManager - INFO -
Added loader_categoricalloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:13,849] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_2 to PreviousScopeFeaturesDataManager
```

```
[I 2025-08-08 01:45:13,850] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 126)
[I 2025-08-08 01:45:13,851] RegionLoader - INFO - Loading…
[I 2025-08-08 01:45:13,852] OnlineCSVDatasource - INFO - Fetching data
from https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-
Codes/master/all/all.csv
[I 2025-08-08 01:45:14,005] RegionLoader - INFO - Completed download --
0.1528468132019043 seconds
[I 2025-08-08 01:45:14,007] CombinedLoader - INFO - Adding
(FinancialLoader-ScopeLoader-CategoricalLoader + RegionLoader)
[I 2025-08-08 01:45:14,007] CombinedLoader - INFO - Merging special
loader RegionLoader to FinancialLoader-ScopeLoader-CategoricalLoader
[I 2025-08-08 01:45:15,678] PreviousScopeFeaturesDataManager - INFO -
Added loader_regionloader to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:15,679] PreviousScopeFeaturesDataManager - INFO -
Added merge_stage_3 to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:15,680] PreviousScopeFeaturesDataManager - INFO -
Remaining data points (526241, 129)
[I 2025-08-08 01:45:15,681] PreviousScopeFeaturesDataManager - INFO -
Added merged to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:16,725] CompanyDataFilter - INFO - Filtered dataset
from 526241 to 526241 data points
[I 2025-08-08 01:45:16,729] PreviousScopeFeaturesDataManager - INFO -
Added reduced to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:45:16,730] PreviousScopeFeaturesDataManager - INFO -
Taking all previous year scopes
100%|      | 103752/103752 [03:13<00:00, 537.54it/s]
[I 2025-08-08 01:48:31,368] PreviousScopeFeaturesDataManager - INFO -
Added original to PreviousScopeFeaturesDataManager
[I 2025-08-08 01:48:31,759] PreviousScopeFeaturesDataManager - INFO -
Data with original found retrieved: Dataset after transformation changes.
```

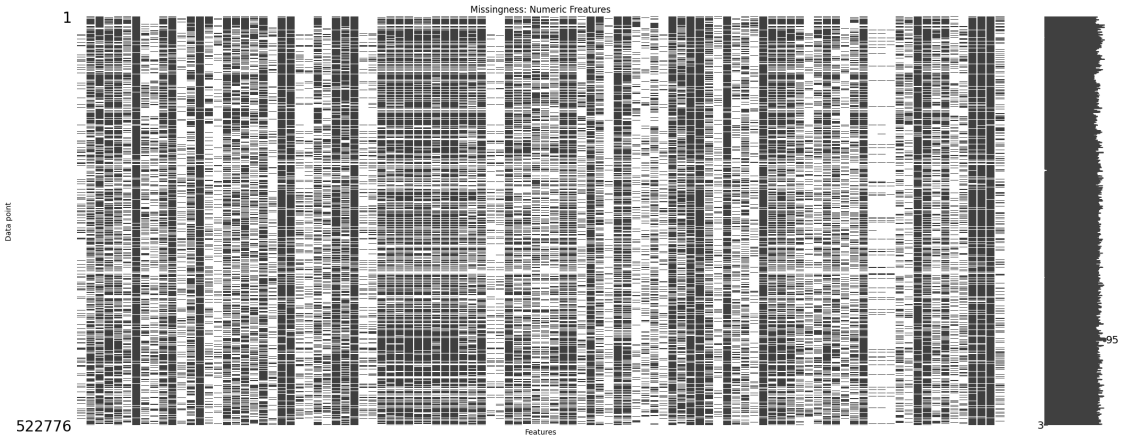
[ ]:

	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	…	

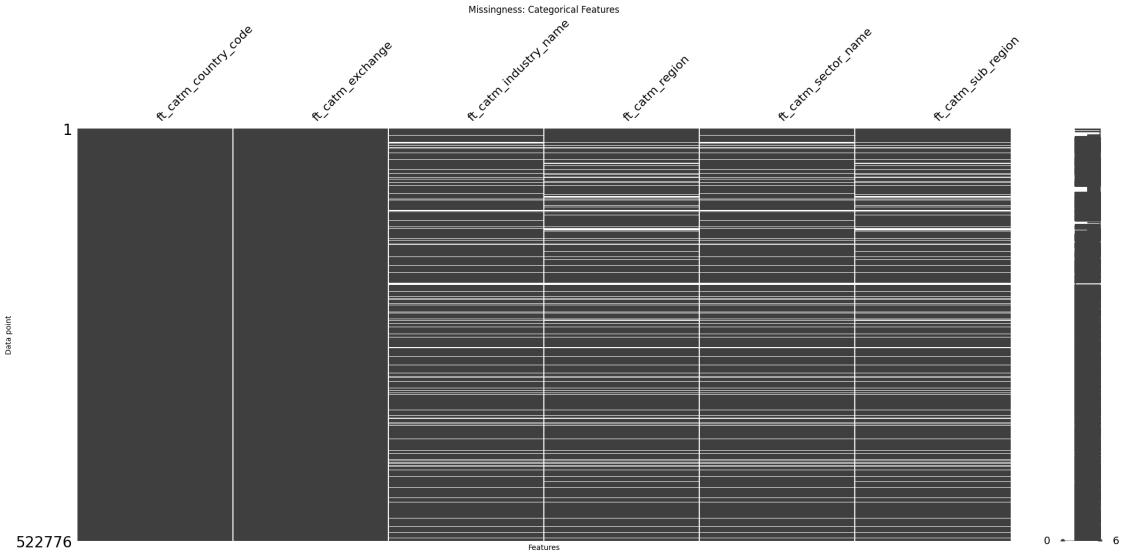
```
NaN
526239                  NaN              NaN                  NaN   …
NaN
526236                  NaN              NaN                  NaN   …
NaN
526237                  NaN              NaN                  NaN   …
NaN
526240                  NaN              NaN                  NaN   …
NaN

       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]
```

[ ]:
```
def visualize_matrix(df, title="Missingness Matrix"):
    fig = plt.figure()
    ax = msno.matrix(df)
    ax.set_title(title)
    ax.set_ylabel('Data point')
    ax.set_xlabel('Features')
    fig.tight_layout()
    plt.show()

visualize_matrix(DATA.filter(regex="^ft_num", axis=1), 'Missingness: Numeric␣
 ↪Freatures')
visualize_matrix(DATA.filter(regex="^ft_cat", axis=1), 'Missingness:␣
 ↪Categorical Features')
visualize_matrix(DATA.filter(regex="^tg_", axis=1), 'Missingness: Targets')
```

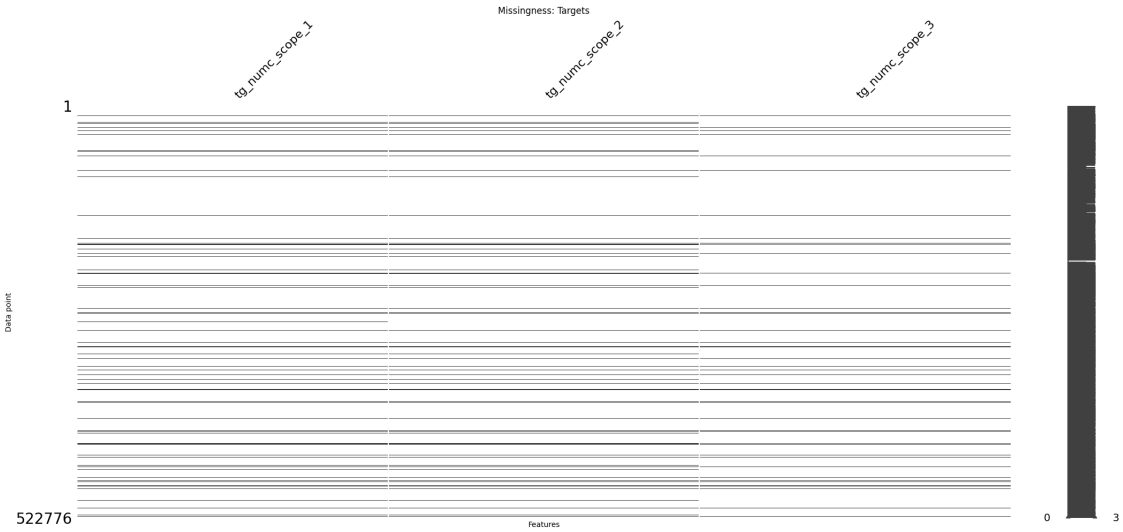
```
<Figure size 640x480 with 0 Axes>
```



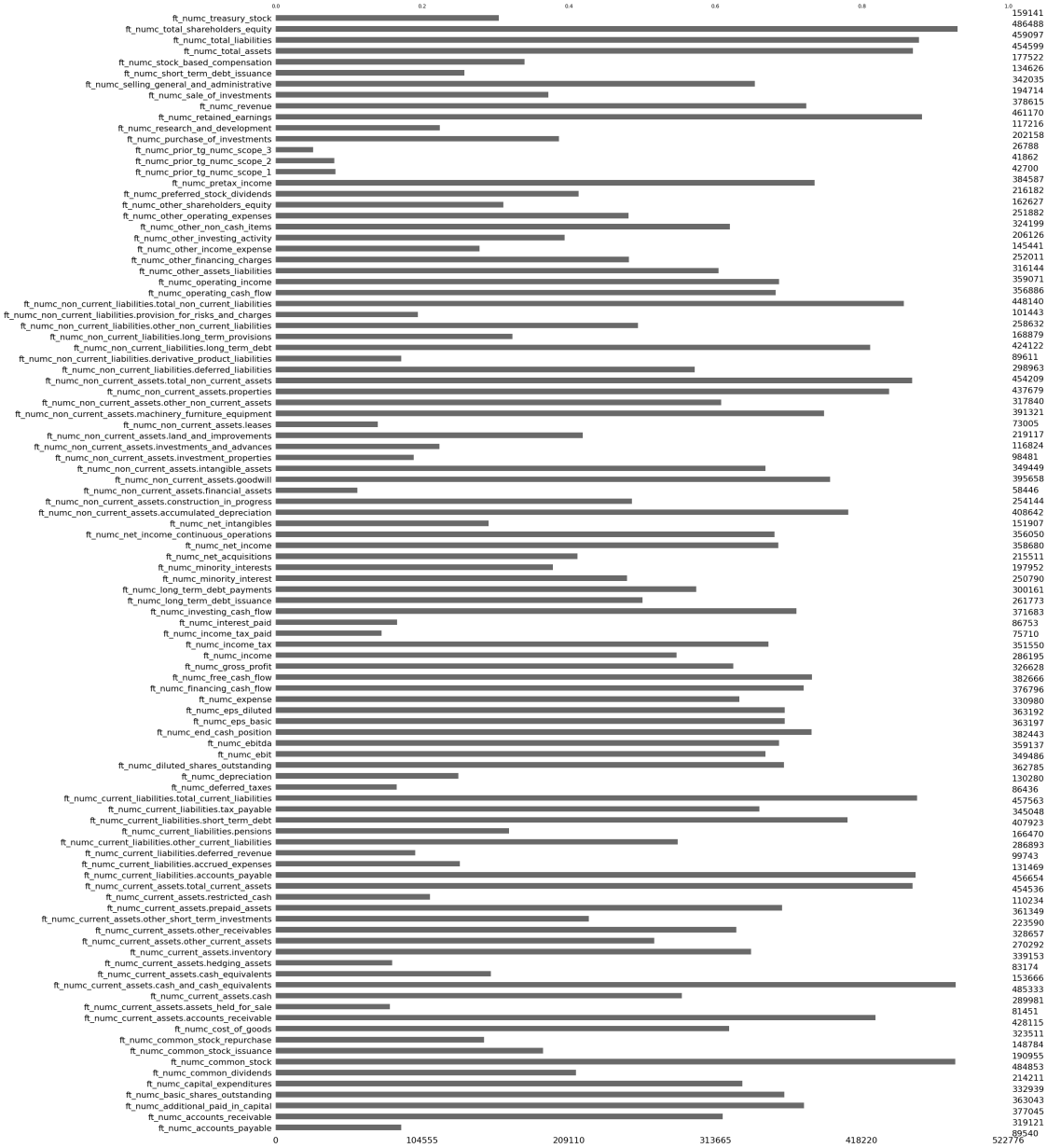
<Figure size 640x480 with 0 Axes>



<Figure size 640x480 with 0 Axes>



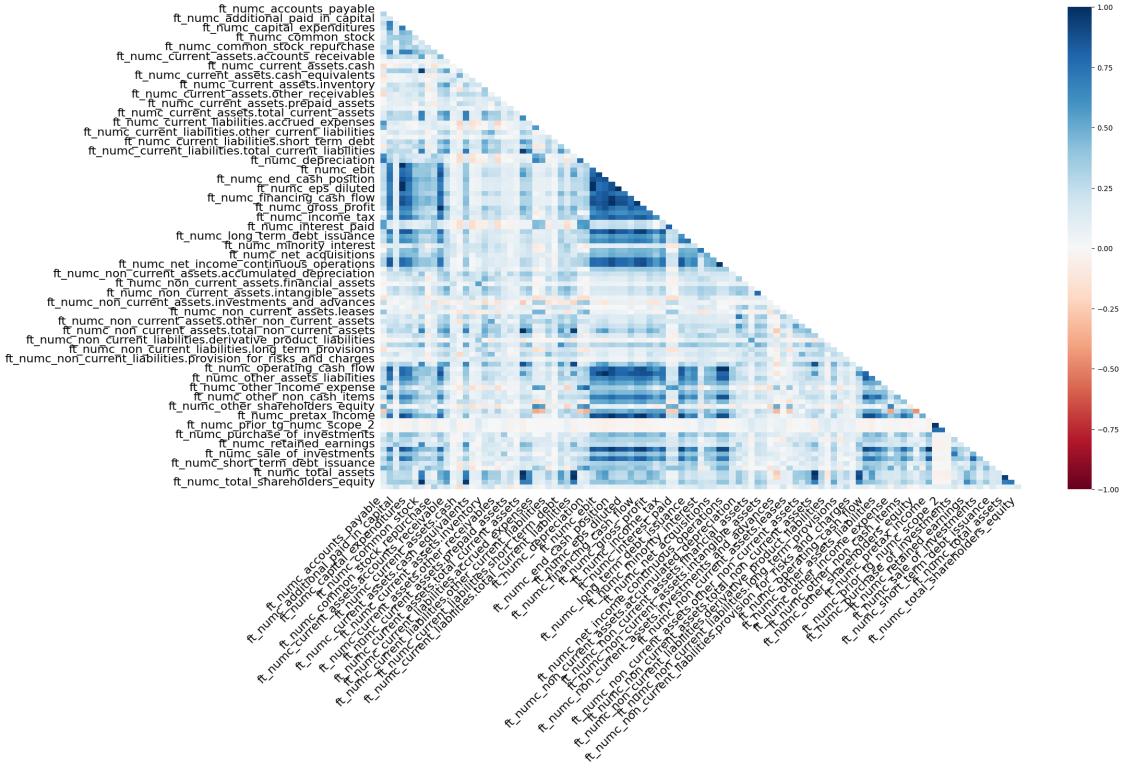
[ ]:
```
df:pd.DataFrame = DATA.filter(regex="^ft_num", axis=1)
msno.bar(df)
plt.show()
```



[ ]:
```
# NOTE: If we have ppe we might not have the others based on this image
msno.heatmap(df)
plt.show()
```

```
c:\Users\User\Workspace\work_oxari\architectura\.venv\Lib\site-
packages\seaborn\matrix.py:260: FutureWarning: Format strings passed to
MaskedConstant are ignored, but in future may error or produce different
behavior
  annotation = ("{:" + self.fmt + "}").format(val)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
File c:
 ↪\Users\User\Workspace\work_oxari\architectura\notebooks\analyze_feature_missi gness.
 ↪py:3
      1 # %%
      2 # NOTE: If we have ppe we might not have the others based on this image
----> 3 msno.heatmap(df)
      4 plt.show()

File c:\Users\User\Workspace\work_oxari\architectura\.
 ↪venv\Lib\site-packages\missingno\missingno.py:398, in heatmap(df, filter, n,␣
 ↪p, sort, figsize, fontsize, labels, label_rotation, cmap, vmin, vmax, cbar, a:)
    395 ax0.patch.set_visible(False)
    397 for text in ax0.texts:
--> 398     t = float(text.get_text())
    399     if 0.95 <= t < 1:
    400         text.set_text('<1')

ValueError: could not convert string to float: '--'
```



[ ]:
```
msno.dendrogram(df)
plt.show()
```

```
Cell was canceled due to an error in a previous cell.
```

[ ]:
```
(~df.isna()).sum().sort_values(ascending=False)[:10]
```

```
Cell was canceled due to an error in a previous cell.
```