Polars Cheat Sheet

Open in Colab

General

Install

```
pip install polars
```

Import

```
import polars as pl
```

データフレームの作成/読み込み

データフレームの作成

nrs	names	random	groups
1	"foo"	0.3	"A"
2	"ham"	0.7	"A"
3	"spam"	0.1	"B"
null	"egg"	0.9	"C"
5	null	0.6	"B"

```
df = pl.DataFrame(
  {
    "nrs": [1, 2, 3, None, 5],
    "names": ["foo", "ham", "spam", "egg", None],
    "random": [0.3, 0.7, 0.1, 0.9, 0.6],
    "groups": ["A", "A", "B", "C", "B"],
  }
)
```

Read CSV

```
df = pl.read_csv("https://j.mp/iriscsv",
                 has_header=True)
```

Read parquet

```
df = pl.read_parquet("path.parquet",
                     columns=["select",
                     "columns"])
```

Expressions

Polarsのエクスプレッションは順番に 実行することができ、コードの可読性が向上します。

```
df \
  .filter(pl.col("nrs") < 4) \
  .group_by("groups") \
  .agg(pl \
  .all() \
  .sum()
)
```

観測値の部分集合 - 行

フィルター: 論理条件を満たす行を抽出します。

```
df.filter(pl.col("random") > 0.5)
df.filter(
  (pl.col("groups") == "B")
  & (pl.col("random") > 0.5)
)
```

サンプル

```
# Randomly select fraction of rows.
df.sample(frac=0.5)

# Randomly select n rows.
df.sample(n=2)
```

先頭、末尾の行を選択する

```
# Select first n rows
df.head(n=2)

# Select last n rows.
df.tail(n=2)
```

観測値の部分集合 - 列

特定の名前を持つ複数の列を選択する

```
df.select(["nrs", "names"])
```

正規表現に一致する名前の列を選択する

```
df.select(pl.col("^n.*$"))
```

部分集合 - 行と列

2~4行を選択

```
df[2:4, :]
```

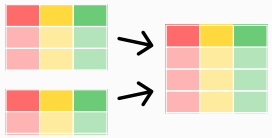
位置1と3の列を選択する（最初の列は0）

```
df[:, [1, 3]]
```

論理条件を満たす行を選択し、特定の列のみを抽出する

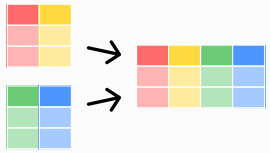
```
df[df["random"] > 0.5, ["names", "groups"]]
```

データの整形 - レイアウト変更、並べ替え、名前変更



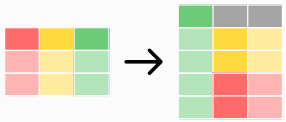
データフレームの行を追加する

```
pl.concat([df, df2])
```



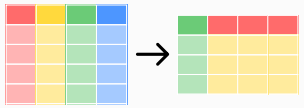
データフレームの列を追加する

```
pl.concat([df, df3], how="horizontal")
```



列を行に集約する

```
df.melt(
  id_vars="nrs",
  value_vars=["names", "groups"]
)
```



行を列に展開する

```
df.pivot(values="nrs", index="groups",
         columns="names")
```

列の値で行を並べ替える

```
# low to high
df.sort("random")

# high to low
df.sort("random", reverse=True)
```

データフレームの列名を変更する

```
df.rename({"nrs": "idx"})
```

データフレームから列を削除する

```
df.drop(["names", "random"])
```

データを要約する

変数ごとのユニークな値に基づいて行数を数える

```
df["groups"].value_counts()
```

データフレームの行数

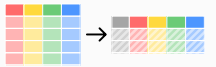
```
len(df)
# or
df.height
```

データフレームの行数と列数のタプル

```
df.shape
```

列内のユニークな値の数

```
df["groups"].n_unique()
```



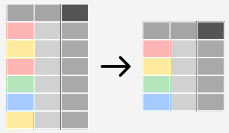
各列の基本的な記述統計量

```
df.describe()
```

集計関数

```
df.select(
    [
      # Sum values
      pl.sum("random").alias("sum"),

      # Minimum value
      pl.min("random").alias("min"),

      # Maximum value
      pl.max("random").alias("max"),
      # or
      pl.col("random").max().alias("other_max"),

      # Standard deviation
      pl.std("random").alias("std dev"),

      # Variance
      pl.var("random").alias("variance"),

      # Median
      pl.median("random").alias("median"),

      # Mean
      pl.mean("random").alias("mean"),

      # Quantile
      pl.quantile("random", 0.75) \
        .alias("quantile_0.75"),
      # or
      pl.col("random").quantile(0.75) \
        .alias("other_quantile_0.75"),

      # First value
      pl.first("random").alias("first"),
    ]
)
```

データのグループ化



列名「col」で値ごとにグループ化し、GroupByオブジェクトを返す

```
df.group_by("groups")
```

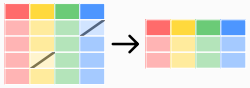
上記のすべての集計関数は、グループにも適用できます。

```
df.group_by(by="groups").agg(
  [
      # Sum values
      pl.sum("random").alias("sum"),

      # Minimum value
      pl.min("random").alias("min"),

      # Maximum value
      pl.max("random").alias("max"),
      # or
      pl.col("random").max().alias("other_max"),

      # Standard deviation
      pl.std("random").alias("std_dev"),

      # Variance
      pl.var("random").alias("variance"),

      # Median
      pl.median("random").alias("median"),

      # Mean
      pl.mean("random").alias("mean"),
      # Quantile
      pl.quantile("random", 0.75) \
        .alias("quantile_0.75"),
      # or
      pl.col("random").quantile(0.75) \
        .alias("other_quantile_0.75"),
      # First value
      pl.first("random").alias("first"),
  ]
)
```

追加のGroupBy関数

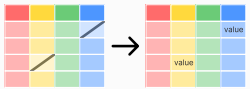
```
df.group_by(by="groups").agg(
  [
    # Count the number of values in each group
    pl.count("random").alias("size"),

    # Sample one element in each group
    pl.col("names").apply(
      lambda group_df: group_df.sample(1)
    ),
  ]
)
```

欠損データの処理



任意の列にnull値が含まれる行を削除する

```
df.drop_nulls()
```



指定された値でnull値を置換する

```
df.fill_null(42)
```

前の行の値を使用してnull値を置換する

```
df.fill_null(strategy="forward")
```

他の補完方法には "backward", "min", "max", "mean", "zero" "one" があります。

浮動小数点のNaN値を指定された値に置換する

```
df.fill_nan(42)
```

新しい列を作成する

データフレームに新しい列を追加する

```
df.with_column(
  (pl.col("random") * pl.col("nrs")) \
    .alias("product")
)
```

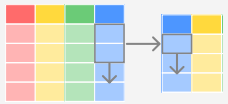
データフレームに複数の新しい列を追加する

```
df.with_columns(
    [
        (pl.col("random") * pl.col("nrs")) \
          .alias("product"),
        pl.col("names").str.lengths() \
          .alias("names_lengths"),
    ]
)
```

インデックス0に行数をカウントする列を追加する

```
df.with_row_count()
```

ローリング関数



利用可能なローリング関数は以下の通りです

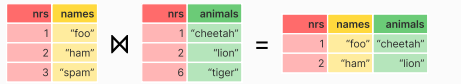
```
df.select(
    [
        # Rolling maximum value
        pl.col("random") \
          .rolling_max(window_size=2) \
          .alias("rolling_max"),

        # Rolling mean value
        pl.col("random") \
          .rolling_mean(window_size=2) \
          .alias("rolling_mean"),

        # Rolling median value
        pl.col("random") \
          .rolling_median(
              window_size=2, min_periods=2) \
          .alias("rolling_median"),

        # Rolling minimum value
        pl.col("random") \
          .rolling_min(window_size=2) \
          .alias("rolling_min"),

        # Rolling standard deviation
        pl.col("random") \
          .rolling_std(window_size=2) \
          .alias("rolling_std"),

        # Rolling sum values
        pl.col("random") \
          .rolling_sum(window_size=2) \
          .alias("rolling_sum"),

        # Rolling variance
        pl.col("random") \
          .rolling_var(window_size=2) \
          .alias("rolling_var"),

        # Rolling quantile
        pl.col("random") \
          .rolling_quantile(
            quantile=0.75, window_size=2,
            min_periods=2
          ) \
          .alias("rolling_quantile"),

        # Rolling skew
        pl.col("random") \
          .rolling_skew(window_size=2) \
          .alias("rolling_skew"),

        # Rolling custom function
        pl.col("random") \
          .rolling_apply(
            function=np.nanstd, window_size=2) \
          .alias("rolling_apply")])
```

ウィンドウ関数

ウィンドウ関数を使用すると、 複数の列で同時にグループ化することができます。

```
df.select(
    [
        "names",
        "groups",
        pl.col("random").sum().over("names") \
          .alias("sum_by_names"),
        pl.col("random").sum().over("groups") \
          .alias("sum_by_groups"),
    ]
)
```

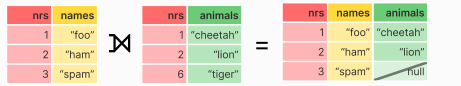
データセットの結合



内部結合

他のセットに一致する行のみを保持します。

```
df.join(df4, on="nrs")
# or
df.join(df4, on="nrs", how="inner")
```



左結合

"left" セット (データフレーム) の各行を保持します。

```
df.join(df4, on="nrs", how="left")
```

外部結合

一致する行が存在しない場合でも、各行を保持します。

```
df.join(df4, on="nrs", how="outer")
```

アンチ結合

df には、df4 と一致しない行がすべて含まれています。

```
df.join(df4, on="nrs", how="anti")
```