Polars Cheat Sheet

Open in Colab

General

Install

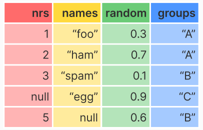
```
pip install polars
```

Import

```
import polars as pl
```

□□□□□□□□□□/□□□□

□□□□□□□□□□



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"])
```

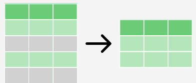
Expressions

Polars□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□□□□□
□□□□

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

Download at polars-cheat-sheet-ja

□□□□□□□□□ -□



□□□□□:□□□□□□□□□□□□□□□□

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

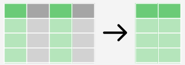
□□□□

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

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

□□□□□□□□□□□□

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

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

□□□□□□□□□ -□



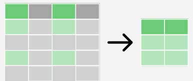
□□□□□□□□□□□□□□□□□

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

□□□□□□□□□□□□□□□□□□

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

□□□□ -□□□



2~4□□□□

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

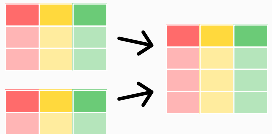
□□1□3□□□□□□□□□□□□□0□

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

□□□□□□□□□□□□□□□□□□□□□□□□□

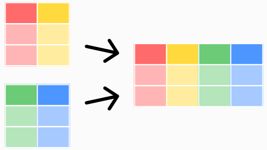
```
df[df["random"] > 0.5, ["names", "groups"]]
```

□□□□□□ -□□□□□□□□□□□□□□□□□



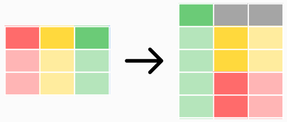
□□□□□□□□□□□□□□

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



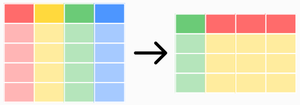
□□□□□□□□□□□□□□

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



□□□□□□□□

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



□□□□□□□□

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

□□□□□□□□□□□

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

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

□□□□□□□□□□□□□□□

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

□□□□□□□□□□□□□□□

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

□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□

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

□□□□□□□□□□

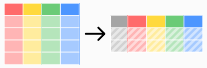
```
len(df)
# or
df.height
```

□□□□□□□□□□□□□□□□□

```
df.shape
```

□□□□□□□□□□□

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



□□□□□□□□□□□□

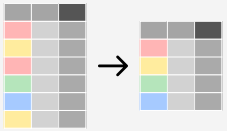
```
df.describe()
```

□□□□

```
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"),
    ]
)
```

□□□□□□□ (Link), inspired by Pandas Cheat Sheet

□□□□□□□□□□



□□□col□□□□□□□□□□□□□GroupBy□□□□□□□□

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

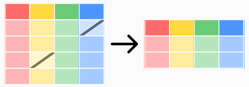
□□□□□□□□□□□□□□□□□□□□□□□□□

```
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"),
  ]
)
```

□□□GroupBy□□

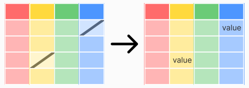
```
df.group_by(by="groups").agg(
  [
    # Count the number of values in each group
    pl.count("random").alias("size"),

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

□□□□□□□□□



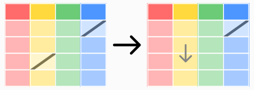
□□□□□null□□□□□□□□□□□

```
df.drop_nulls()
```



□□□□□□□null□□□□□□

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



□□□□□□□□□□□null□□□□□□

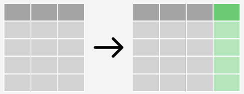
```
df.fill_null(strategy="forward")
```

□□□□□□□□ "backward", "min", "max", "mean", "zero" "one" □□□□□□

□□□□□□NaN□□□□□□□□□□□□

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

□□□□□□□□□□



□□□□□□□□□□□□□□□□

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

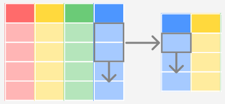
□□□□□□□□□□□□□□□□□□□

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

□□□□□□0□□□□□□□□□□□□□□□

```
df.with_row_count()
```

□□□□□□□□



□□□□□□□□□□□□□□□□□□□

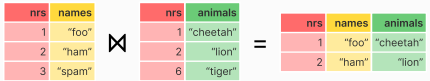
```
df.select(
    [
        # Rolling maximum value
        pl.col("random") \
          .rolling_max(window_size=2) \
          .alias("rolling_max"),

        # 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")])
```

□□□□□□□□

□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□□□□□□□

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

□□□□□□□□□



nrs	names
1	"foo"
2	"ham"
3	"spam"

⋈

nrs	animals
1	"cheetah"
2	"lion"
6	"tiger"

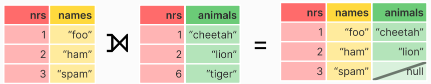
=

nrs	names	animals
1	"foo"	"cheetah"
2	"ham"	"lion"

□□□□

□□□□□□□□□□□□□□□□□□□□

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



nrs	names
1	"foo"
2	"ham"
3	"spam"

⟕

nrs	animals
1	"cheetah"
2	"lion"
6	"tiger"

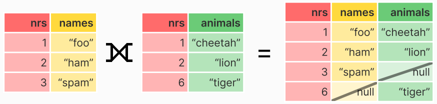
=

nrs	names	animals
1	"foo"	"cheetah"
2	"ham"	"lion"
3	"spam"	null

□□□

"left" □□□ (□□□□□□□□)□□□□□□□□□□□

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



nrs	names
1	"foo"
2	"ham"
3	"spam"

⟗

nrs	animals
1	"cheetah"
2	"lion"
6	"tiger"

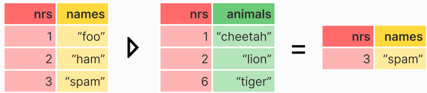
=

nrs	names	animals
1	"foo"	"cheetah"
2	"ham"	"lion"
3	"spam"	null
6	null	"tiger"

□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□

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



nrs	names
1	"foo"
2	"ham"
3	"spam"

▷

nrs	animals
1	"cheetah"
2	"lion"
6	"tiger"

=

nrs	names
3	"spam"

□□□□□

df □□□ df4 □□□□□□□□□□□□□□□□□□□□

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

Download at polars-cheat-sheet-ja

□□□□□□□ (Link), inspired by Pandas Cheat Sheet