Python ↔ R Data Wrangling Cheat Sheet (Improved Layout)

Setup Example Data

R

```
library(dplyr)
library(tidyr)

df <- tibble(
  name = c("A","B","C","C"),
  age  = c(25, 30, 22, 22),
  score= c(88, 92, 95, 95),
  grp  = c("g1","g1","g2","g2")
)
```

Python

```
import pandas as pd

df = pd.DataFrame({
  "name": ["A","B","C","C"],
  "age":  [25, 30, 22, 22],
  "score":[88, 92, 95, 95],
  "grp":  ["g1","g1","g2","g2"]
})
```

Select Columns

R

```
df %>% select(name, score)
```

Python

```
df[["name", "score"]]
```

Filter Rows

R

```
df %>% filter(age > 24, grp == "g1")
```

Python

```
df[(df["age"] > 24) & (df["grp"] == "g1")]
```

Mutate / Create Columns

R

```
df %>%
```

```
  mutate(
    passed = score > 90,
    score_z = (score - mean(score)) / sd(score)
  )
```

Python

```
df.assign(
  passed = df["score"] > 90,
  score_z = (df["score"] - df["score"].mean()) / df["score"].std()
)
```

Group and Summarize

R

```
df %>%
  group_by(grp) %>%
  summarise(
    n = n(),
    mean_score = mean(score),
    max_age = max(age),
    .groups = "drop"
  )
```

Python

```
df.groupby("grp").agg(
  n=("grp", "size"),
  mean_score=("score", "mean"),
  max_age=("age", "max")
).reset_index()
```

Join

R

```
lookup <- tibble(grp=c("g1","g2"), label=c("Group 1","Group 2"))
df %>% left_join(lookup, by="grp")
```

Python

```
lookup = pd.DataFrame({
  "grp":["g1","g2"],
  "label":["Group 1","Group 2"]
})

df.merge(lookup, on="grp", how="left")
```

Pivot Wider

R

```
df %>%
  group_by(name, grp) %>%
  summarise(mean_score=mean(score), .groups="drop") %>%
  pivot_wider(names_from = grp, values_from = mean_score)
```

Python

```
tmp = (df.groupby(["name","grp"], as_index=False)
         .agg(mean_score=("score","mean")))

tmp.pivot(index="name", columns="grp", values="mean_score").reset_index()
```