Week 2 Code

```
from pathlib import Path
import pandas as pd

cwd = Path.cwd().resolve()
repo = next(p for p in [cwd, *cwd.parents] if (p / ".git").exists())

dev = pd.read_csv(repo / "data" / "raw" / "DEV _ March Madness.csv")
spell = pd.read_csv(repo / "data" / "raw" / "MTeamSpellings.csv")
games = pd.read_csv(repo / "data" / "raw" /
"MarchMadnessGameStats2003-2024.csv")

(repo / "data" / "processed").mkdir(parents=True, exist_ok=True)
(repo / "data" / "interim").mkdir(parents=True, exist_ok=True)

# TeamID -> team name
spell["nameKey"] =
spell["TeamNameSpelling"].astype(str).str.strip().str.lower()
dev["nameKey"] = dev["Mapped ESPN Team
Name"].astype(str).str.strip().str.lower()
dev = dev.merge(spell[["TeamID", "nameKey"]].drop_duplicates(),
on="nameKey", how="left")

# team and which season
teamSeason = (
    dev.dropna(subset=["TeamID"])
       .drop_duplicates(["Season", "TeamID"])
       .copy()
)
if "nameKey" in teamSeason.columns:
    teamSeason.drop(columns=["nameKey"], inplace=True)

# tournament outcomes
wins = (games.groupby(["Season", "WTeamID"]).size()
        .rename("tourneyWins").reset_index().rename(columns={"WTeamID"
: "TeamID"}))
losses = (games.groupby(["Season", "LTeamID"]).size()
          .rename("tourneyLosses").reset_index().rename(columns={"LTea
mID": "TeamID"}))

t = wins.merge(losses, on=["Season", "TeamID"], how="outer").fillna(0)
t["tourneyWins"] = t["tourneyWins"].astype(int)
t["tourneyLosses"] = t["tourneyLosses"].astype(int)
t["tourneyGames"] = t["tourneyWins"] + t["tourneyLosses"]

def winsToRound(w):
    if w <= 0: return "R64"
    if w == 1: return "R32"
    if w == 2: return "S16"
```

```
    if w == 3: return "E8"
    if w == 4: return "F4"
    if w == 5: return "Final"
    return "Champion"

t["roundReached"] = t["tourneyWins"].map(winsToRound)

teamSeason = teamSeason.merge(t, on=["Season", "TeamID"], how="left")
teamSeason["tourneyWins"] =
teamSeason["tourneyWins"].fillna(0).astype(int)
teamSeason["tourneyLosses"] =
teamSeason["tourneyLosses"].fillna(0).astype(int)
teamSeason["tourneyGames"] =
teamSeason["tourneyGames"].fillna(0).astype(int)
teamSeason["roundReached"] =
teamSeason["roundReached"].fillna("NoTourney")

# only march madness teams allowed
mm = (teamSeason.loc[teamSeason["Post-Season Tournament"].eq("March
Madness"), ["Season", "TeamID"]]
      .drop_duplicates())

g = games[[c for c in ["Season", "DayNum", "WTeamID", "LTeamID",
"WScore", "LScore"] if c in games.columns]].copy()
g = g.merge(mm.rename(columns={"TeamID": "WTeamID"}), on=["Season",
"WTeamID"], how="inner")
g = g.merge(mm.rename(columns={"TeamID": "LTeamID"}), on=["Season",
"LTeamID"], how="inner")

g["team1Id"] = g[["WTeamID", "LTeamID"]].min(axis=1)
g["team2Id"] = g[["WTeamID", "LTeamID"]].max(axis=1)
g["y"] = (g["team1Id"] == g["WTeamID"]).astype(int)
g["winnerTeamId"] = g["WTeamID"]

# convert id to names
names = (teamSeason[["Season", "TeamID", "Mapped ESPN Team Name"]]
         .drop_duplicates(["Season", "TeamID"])
         .rename(columns={"Mapped ESPN Team Name": "teamName"}))
g = g.merge(names.rename(columns={"TeamID": "team1Id", "teamName":
"team1Name"}),
            on=["Season", "team1Id"], how="left")
g = g.merge(names.rename(columns={"TeamID": "team2Id", "teamName":
"team2Name"}),
            on=["Season", "team2Id"], how="left")

# features
t1 = teamSeason.rename(columns={"TeamID": "team1Id"}).copy()
t1 = t1.rename(columns={c: f"team1_{c}" for c in t1.columns if c not
in ["Season", "team1Id"]})
```

```
t2 = teamSeason.rename(columns={"TeamID": "team2Id"}).copy()
t2 = t2.rename(columns={c: f"team2_{c}" for c in t2.columns if c not
in ["Season", "team2Id"]})

tourneyGames = g.merge(t1, on=["Season", "team1Id"],
how="left").merge(
    t2, on=["Season", "team2Id"], how="left"
)

# drops columns with more than 30% of rows missing the data.
# https://stackoverflow.com/questions/60450808/remove-columns-with-
missing-values-above-a-threshold-pandas
keyCols = {
    "Season", "DayNum", "team1Id", "team2Id", "winnerTeamId", "y",
    "WTeamID", "LTeamID", "WScore", "LScore", "team1Name",
"team2Name",
    "team1_Post-Season Tournament", "team2_Post-Season Tournament"
}
dropCols = [c for c, frac in tourneyGames.isna().mean().items() if
frac > 0.30 and c not in keyCols]
tourneyGames = tourneyGames.drop(columns=dropCols)

teamSeason.to_csv(repo / "data" / "processed" / "teamSeason.csv",
index=False)
tourneyGames.to_csv(repo / "data" / "interim" / "tourneyGames.csv",
index=False)

print("tourneyGames data")
display(tourneyGames.head(5))
print("teamSeason data")
display(teamSeason.head(5))
```

```
tourneyGames
```

```
   Season  DayNum  WTeamID  LTeamID  WScore  LScore  team1Id  team2Id
y  \
0    2003     134     1421     1411      92      84     1411     1421
0
1    2003     136     1112     1436      80      51     1112     1436
1
2    2003     136     1113     1272      84      71     1113     1272
1
3    2003     136     1141     1166      79      73     1141     1166
1
4    2003     136     1143     1301      76      74     1143     1301
1

   winnerTeamId  ... team2_Post-Season Tournament Sorting Index  \
0          1421  ...                                          1
1          1112  ...                                          1
```

```
2          1113  ...                                                1
3          1141  ...                                                1
4          1143  ...                                                1

  team2_Vulnerable Top 2 Seed? team2_Tournament Winner?  \
0                           No                       No
1                           No                       No
2                           No                       No
3                           No                       No
4                           No                       No

   team2_Tournament Championship?  team2_Final Four?  \
0                              No                 No
1                              No                 No
2                              No                 No
3                              No                 No
4                              No                 No

   team2_Top 12 in AP Top 25 During Week 6?  team2_tourneyWins  \
0                                        No                  1
1                                        No                  0
2                                        No                  0
3                                        No                  0
4                                        No                  0

   team2_tourneyLosses  team2_tourneyGames  team2_roundReached
0                    1                   2                R32
1                    1                   1                R64
2                    1                   1                R64
3                    1                   1                R64
4                    1                   1                R64

[5 rows x 330 columns]

teamSeason

   Season Short Conference Name  Adjusted Temo  Adjusted Tempo Rank  \
0    2025                   ACC           66.0                  269
1    2015                   SEC           62.4                  274
2    2025                   B12           61.9                  359
3    2021                   WCC           73.8                    7
4    2024                    BE           64.6                  330

   Raw Tempo  Raw Tempo Rank  Adjusted Offensive Efficiency  \
0       66.6             256                          130.1
1       63.8             242                          121.3
2       62.4             360                          123.4
3       74.3              14                          126.4
4       66.0             305                          127.5
```

```
   Adjusted Offensive Efficiency Rank  Raw Offensive Efficiency  \
0                                   1                     125.0
1                                   6                     115.5
2                                  12                     116.4
3                                   1                     121.9
4                                   1                     123.0

   Raw Offensive Efficiency Rank  ...  Vulnerable Top 2 Seed?  \
0                              1  ...                      No
1                              9  ...                      No
2                             17  ...                      No
3                              1  ...                      No
4                              1  ...                      No

   Tournament Winner?  Tournament Championship?  Final Four?  \
0                  No                        No           No
1                  No                        No          Yes
2                  No                        No           No
3                  No                       Yes          Yes
4                 Yes                       Yes          Yes

   Top 12 in AP Top 25 During Week 6?  TeamID  tourneyWins
tourneyLosses  \
0                                 Yes  1181.0            0
0
1                                 Yes  1246.0            4
1
2                                  No  1222.0            0
0
3                                 Yes  1211.0            5
1
4                                 Yes  1163.0            6
0

   tourneyGames  roundReached
0             0     NoTourney
1             5            F4
2             0     NoTourney
3             6         Final
4             6      Champion

[5 rows x 170 columns]
```