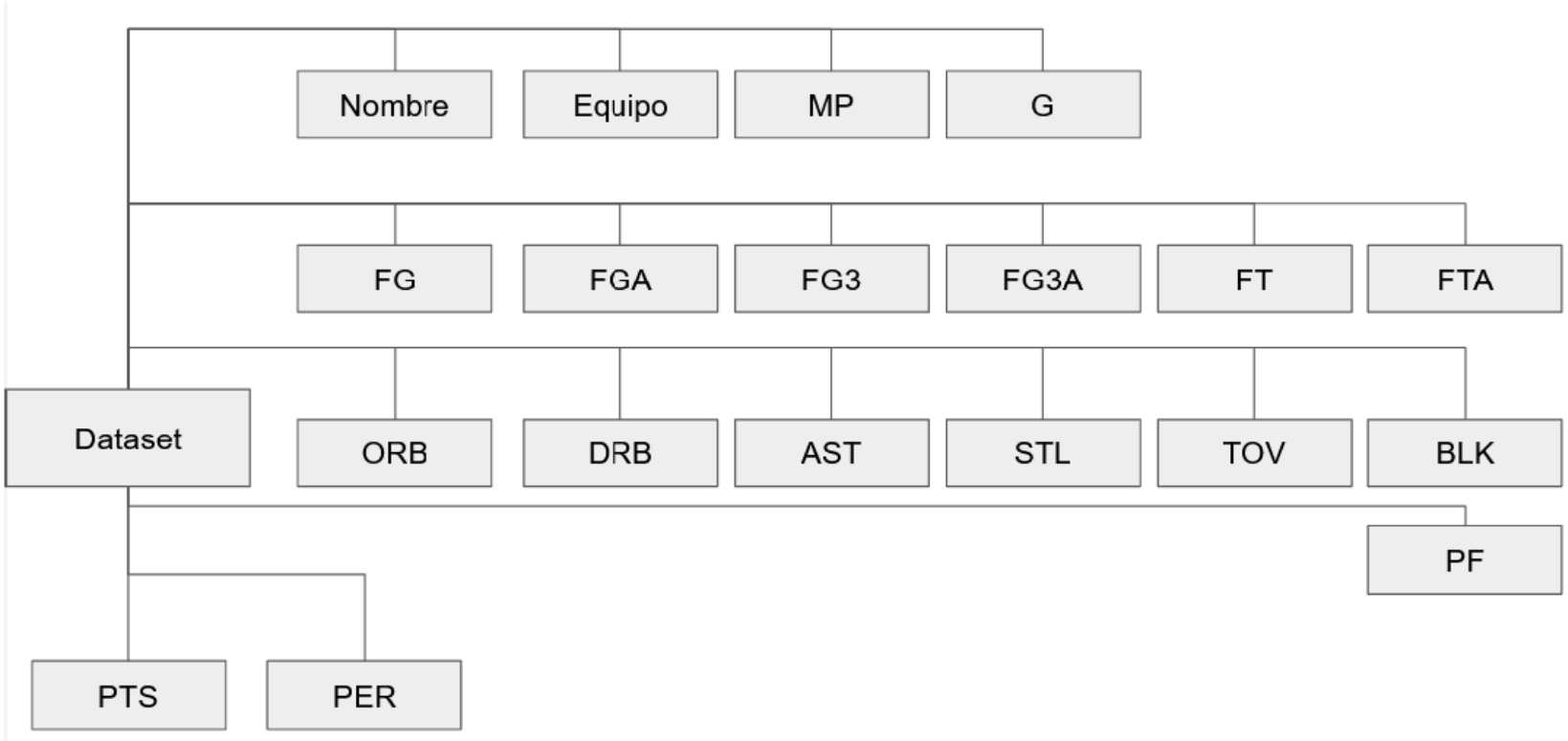
Práctica 2: Limpieza y análisis de datos

Tipología y ciclo de vida de los datos

David Lucas y Fco Javier Cantero

1. Descripción del dataset

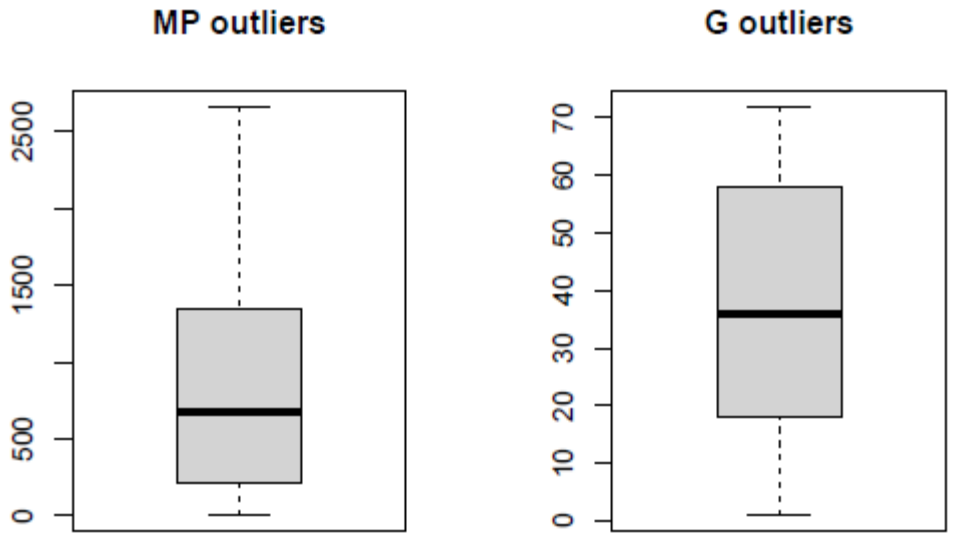


2. Integración y selección de los datos de interés a analizar

- **Conjunto de datos adicional - Nba 2020-2021 Season Player Stats:** https://www.kaggle.com/datasets/umutalpaydn/nba-20202021-season-player-stats?resource=download
- Adición al conjunto de datos principal de las variables de Edad y Posición de cada jugador.
- Sustitución de las variables FG3, FG3A por FG3%, FT, FTA por FT% y creación FG2%.
- Eliminación de FG y FGA.

3. Limpieza de los datos

- Comprobación y eliminación de elementos vacíos.
- Comprobación y eliminación de elementos NANs.
- Eliminación de jugadores que estén por debajo del primer percentil en las variables MP y G.



4. Análisis de los datos

- Comprobación de normalidad y homogeneidad de la varianza
- Análisis sobre el conjunto de datos:
  - Correlación lineal
  - Regresión lineal
  - Regresión logística
  - Contraste de hipótesis

5. Resolución del problema (i)

Correlación lineal

```
##          FG2p        FG3p       FTp      ORB       DRB       AST       STL
## PER 0.4073303 -0.01475301 0.1507436 0.473949 0.5751889 0.500181 0.3701477
##           BLK       TOV        PF       PTS PER
## PER 0.4007462 0.5740526 0.3344931 0.6709724   1
```

Regresión lineal

- Creación de un modelo capaz de predecir el PER al 56%.
- Se pueden predecir más de la mitad de los datos.
- La variable Age no mejora el modelo.

5. Resolución del problema (ii)

Regresión logística

- Se crea un modelo capaz de detectar si un jugador es grande o pequeño con un 83% de acierto.
- Se pueden calificar a los jugadores por tamaño según sus estadísticas

Contraste de hipótesis

- Se comprueba que los jugadores que juegan en la posición de Center tienen un mayor PER que en el resto de posiciones