Nicks EDA

Nicholas Mandarano and Patrick McHugh

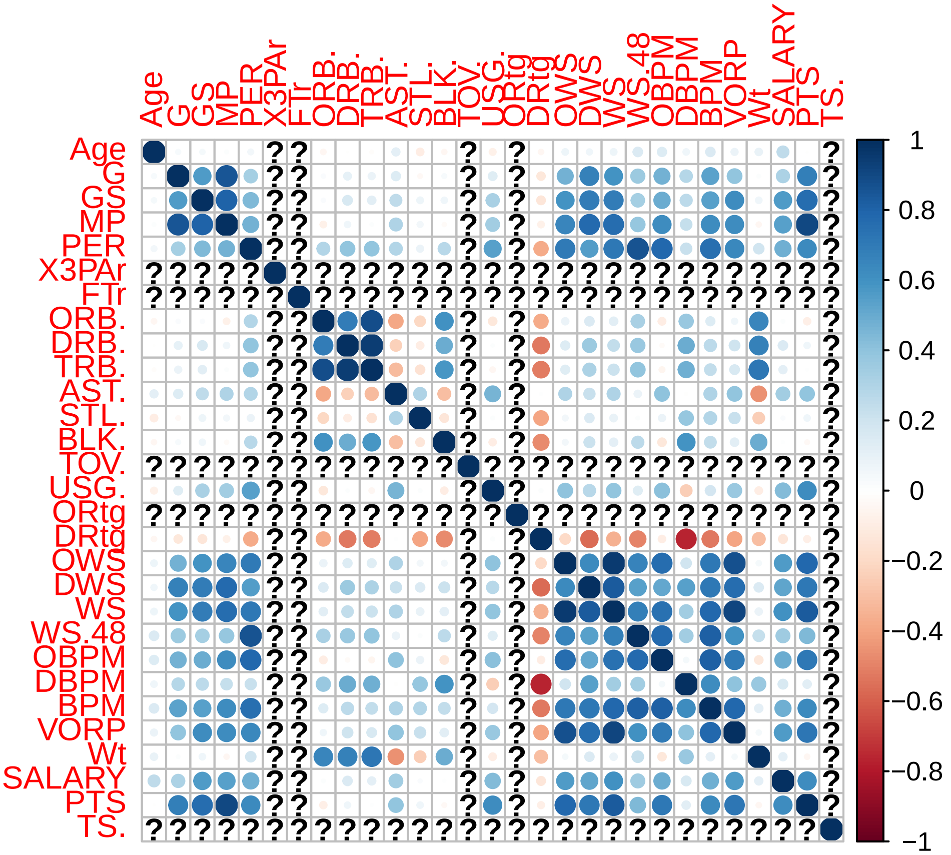
3/30/2022

Exploratory Data Analysis

Our data comes from four different datasets. We used three of Riguang Wen's datasets from figshare.com – `players cv`, `players salary`, and `players stat`. We also used a dataset called `NBA RS 2020-1950 Stats` uploaded to zenodo.org by Pablo Gomez and Sandra Giral. From these datasets, we considered the following variables.

Variable	Description	Type	Source
Player	Name of player	Character	`players stat`
Age	Age of player	Numeric	`players stat`
G	Games played	Numeric	`players stat`
GS	Games started	Numeric	`players stat`
MP	Minutes played	Numeric	`players stat`
PER	Player efficiency rating	Numeric	`players stat`
PTS	Points	Numeric	`NBA RS 2020-1950 Stats`
X3PAr	3PA/FGA	Numeric	`players stat`
FTr	FTA/FGA	Numeric	`players stat`
TS	True shooting percentage	Numeric	`NBA RS 2020-1950 Stats`
ORB	Offensive rebounds	Numeric	`players stat`
DRB	Defensive rebounds	Numeric	`players stat`
TRB	Total rebounds	Numeric	`players stat`
AST	Assists	Numeric	`players stat`
STL	Steals	Numeric	`players stat`
BLK	Blocks	Numeric	`players stat`
TOV	Turnovers	Numeric	`players stat`
USG	Usage percentage	Numeric	`players stat`
ORtg	Offensive rating	Numeric	`players stat`
DRtg	Defensive rating	Numeric	`players stat`
OWS	Offensive win shares	Numeric	`players stat`
DWS	Defensive win shares	Numeric	`players stat`
WS	Win shares	Numeric	`players stat`
WS.48	Win shares per 48 minutes	Numeric	`players stat`
OBPM	Offensive box +/-	Numeric	`players stat`
DBPM	Defensive box +/-	Numeric	`players stat`
BPM	Box +/-	Numeric	`players stat`
VORP	Value over replacement player	Numeric	`players stat`
Pos	Position	Factor	`players salary`
Ht	Height in inches	Numeric	`players salary`
Wt	Weight in pounds	Numeric	`players salary`
PwrSix	Power Six College?	Indicator	`players cv`
International	International Player?	Indicator	`players cv`
Salary	Salary in dollars	Numeric	`players salary`

Immediately we can recognize that some variables are functions of others and therefore do not need to be considered. Specifically, `BPM = OBPM + DBPM`, so there is no need to include `BPM` in our model. Similarly, `WS = OWS + DWS` and `TRB = ORB + DRB`, so we can exclude `WS` and `TRB` from consideration if we include `OWS`, `DWS`, ORB and DRB in our model. Some other multicollinearity issues will likely arise given the correlation matrix of the numerical variables under consideration below. Some examples of potential issues are the correlation between `WS` and `PER` as well as that of `MP` and `G`.



Also in the numeric variables are signs of non-normality. Of the 27 numeric variables considered after the exclusion of `BPM`, `WS` and `TRB`, 11 had medians that had 10% or more in difference of the mean, possibly indicating asymmetry. Of these, only the boxplots of `G` and `GS` did not signify outliers, though histograms of the data did show skewness. Histograms of the others (`FTr, ORB, AST, BLK, OWS, DWS, VORP, Salary`, and `PTS`) were all right-skewed.