資料分析與學習基石
Homework 1 – First visit in Kaggle data
資訊系 F14076083 魏湧致

- **基本資料集描述, 資料特性描述**

NBA Players stats since 1950

這份資料集蒐集了1950年以來每位NBA球員的各種資料，包括身高、畢業大學、得分、失誤等等，主要有三份資料，分別為

Players : 有Player、height、weight、college、born、birth_city、birth_state這七個columns

player_data : 有name、year_start 、year_end、position、height、weight、birth_date、college這八個columns

以上兩份資料列出了所有球員的基本資料，包含了身高、體重以及鋒線位置等。

- 從數據的分布可以看出身高與體重大致上呈現正相關，也能看出身高是要進入NBA很重要的條件，大部分的球員身高都介於190~210cm，但也有特例，如史上最矮的球員Muggsy Bogues身高就只有160cm。
- 身高與鋒線位置的分布也有相關，由高到低分別適合擔任中鋒、前鋒、後衛。
- 也能從球員的畢業學校看出哪些大學是美國NCAA強隊，如培養出最多NBA球員的肯塔基大學、擁有許多名人堂球員的UCLA、Michael Jordan就讀的北卡羅來納大學等等。

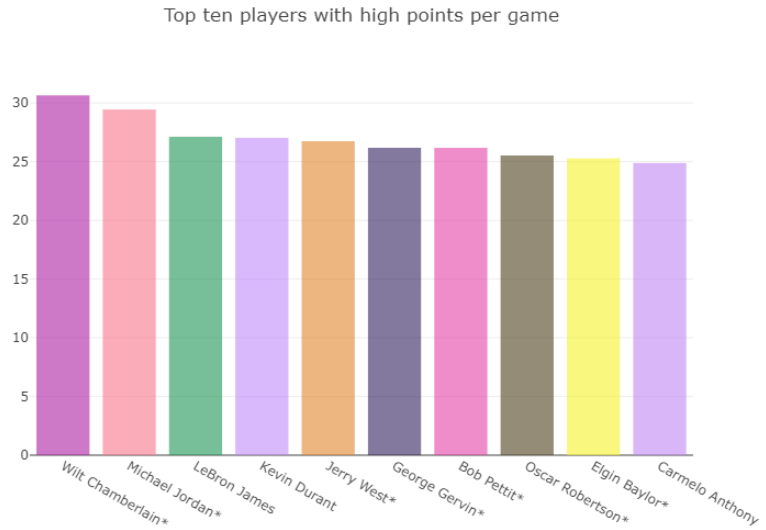
Seasons_Stats : 有Year、Player、Pos、Age、Tm、G、GS、MP、PER、TS%、3PAr、FTr、ORB%、DRB%、TRB%、AST%、STL%、BLK%、TOV%、USG%、OWS、DWS、WS、WS/48、OBPM、DBPM、BPM、VORP、FG、FGA、FG%、3P、3PA、3P%、2P、2PA、2P%、eFG%、FT、FTA、FT%、ORB、DRB、TRB、AST、STL、BLK、TOV、PF、PTS這些columns

這份資料列出了每個球員生涯內每年的各種數據

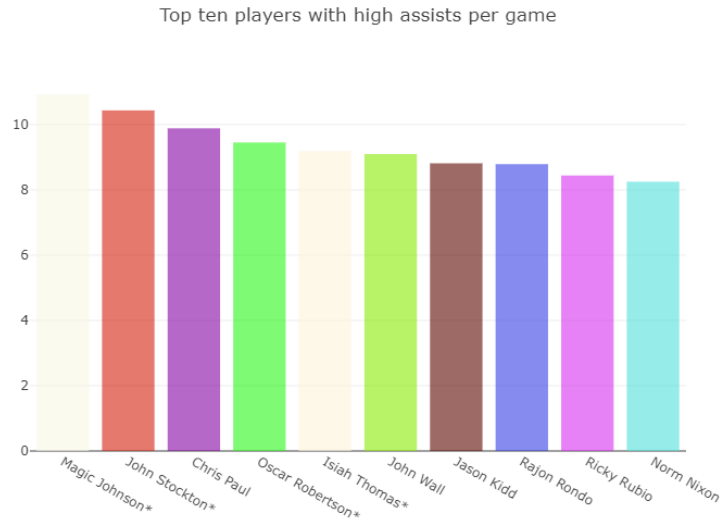
- 資料內有些較進階的數據(如PER、OBPM、DBPM、VORP等)在較早的年代就沒有紀錄。
- 因為三分線是在1979年才設立，所以關於三分球的數據在1980年後才開始有記錄。

• **Notebook方法介紹與比較**

1. **NBA Top players + Deep Learning**
   - Players position and age distribution
     直接由資料集統計取得，球員平均分配在各個鋒線位置，年紀大約落在20到33歲。
   - Points per game
     將球員得分除以出賽場數得到，Wilt Chamberlain的30.07是PPG最高的球員，而第10名的是Carmelo Anthony的24.88，在平常的比賽常常看到許多球員都能夠拿到20分以上，但要在整個生涯中都維持高得分是很困難的事，除了隨著年齡運動表現會下降外，也要避免受嚴重的傷。因此從場均得分這項數據不僅可以看出一位球員優異的得分能力，同時也能知道球員在生涯中很少受到復原後難以恢復水準的大傷。



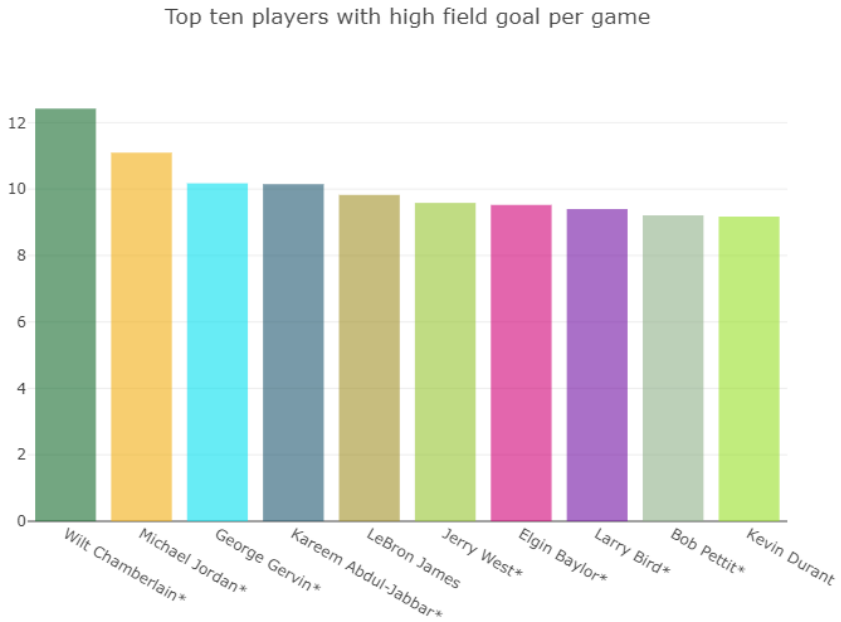
   - Assists per game



將球員助攻數量除以出賽場數取得，前10落在8到10次之間，後衛因為是主要控球的球員，所以在這項數據的表現通常較為出色，助功前10名的球員幾乎都是控球後衛

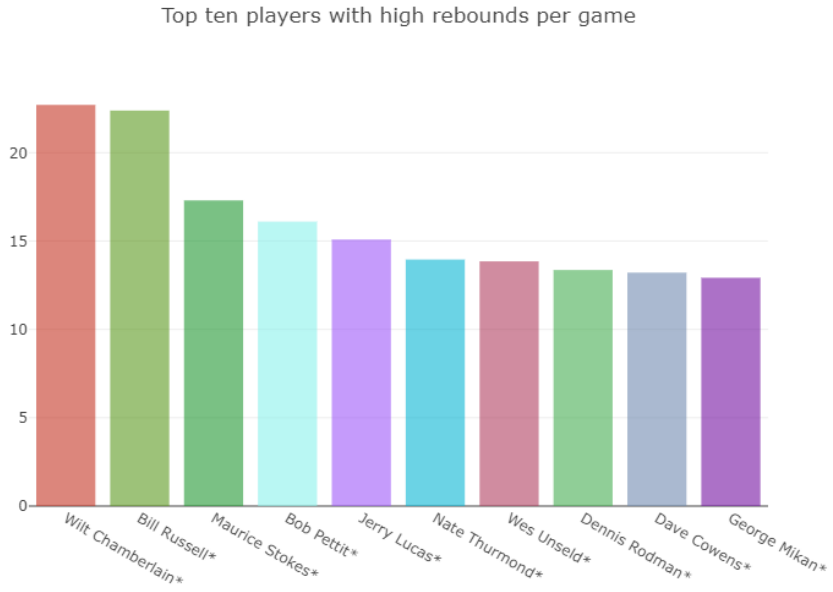
- Field goal per game

將球員FG除以出賽場數取得，這項數據紀錄球員除了三分與罰球的場均得分次數，有幾位球員可能因為三分球的關係，雖然場均FG很高但場均得分並沒有在前10名。



- Rebounds per game

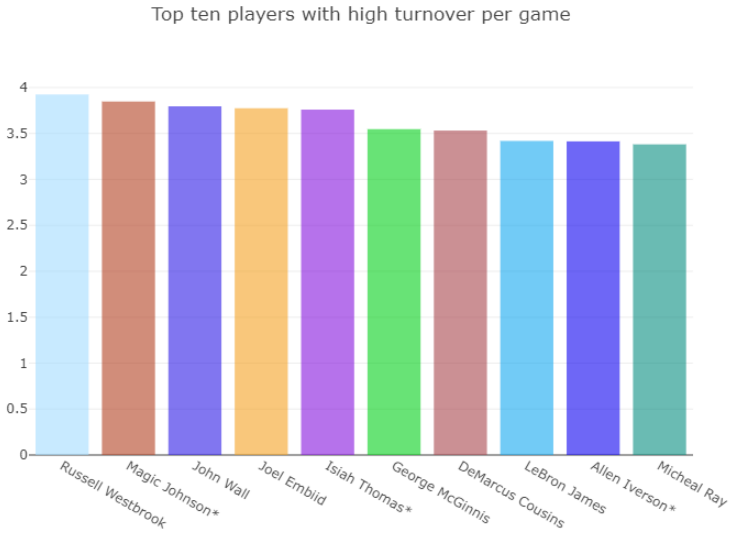
將球員籃板數量除以出賽場數取得，這項數據前10名幾乎都為中鋒或大前鋒。



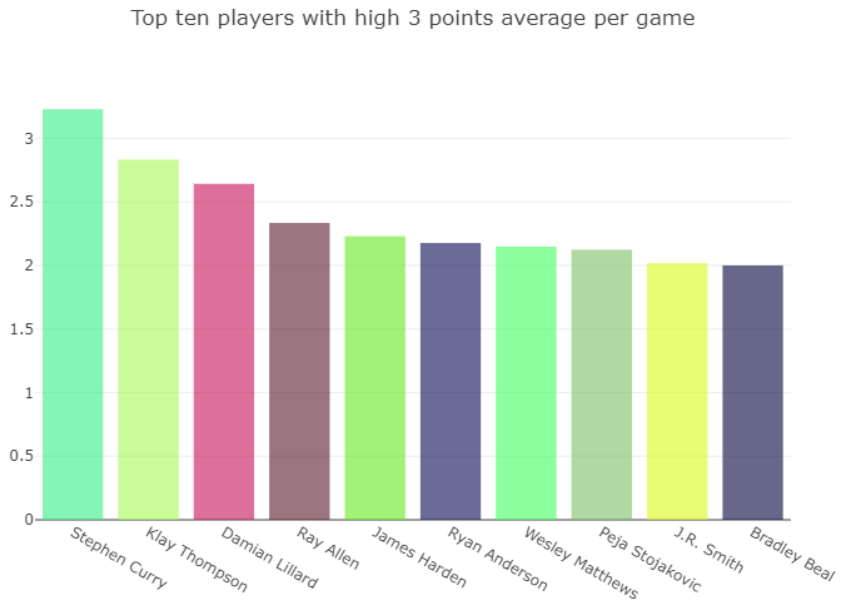
- Turnover per game

將球員失誤數量除以出賽場數取得，這10位幾乎都是明星球

員，還有一些拿過MVP，因此失誤高的原因應該是這些球員主要是隊上的持球與決策者，失誤率才會較高。

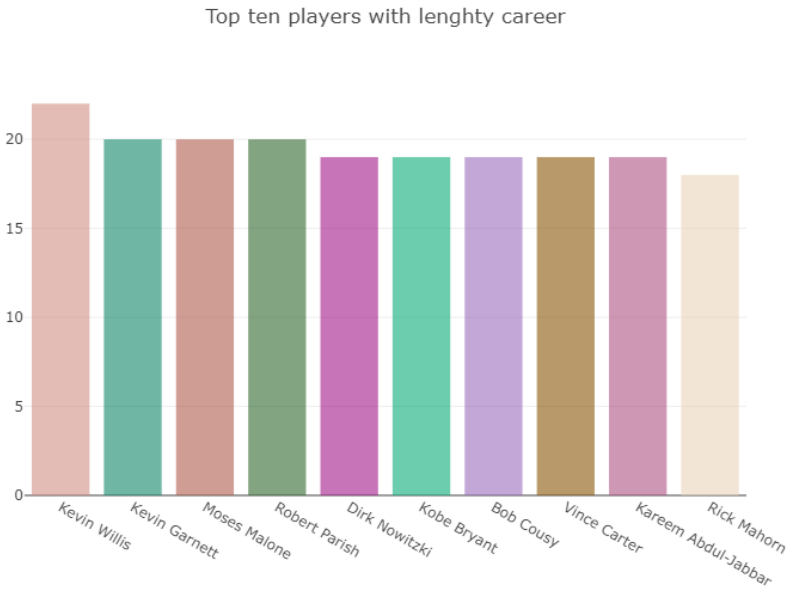


- 2 points per game
  將球員2分球數量除以出賽場數取得。
- 3 points per game
  將球員3分球數量除以出賽場數取得，可以看到都是近期的球員居多，應該是近年來NBA各隊的打法都偏向於外線出手居多。



- Dribbles per game
  這項數據Notebook的作者將DRB(防守籃板)誤以為運球次數，故可刪掉不計。
- Players career length

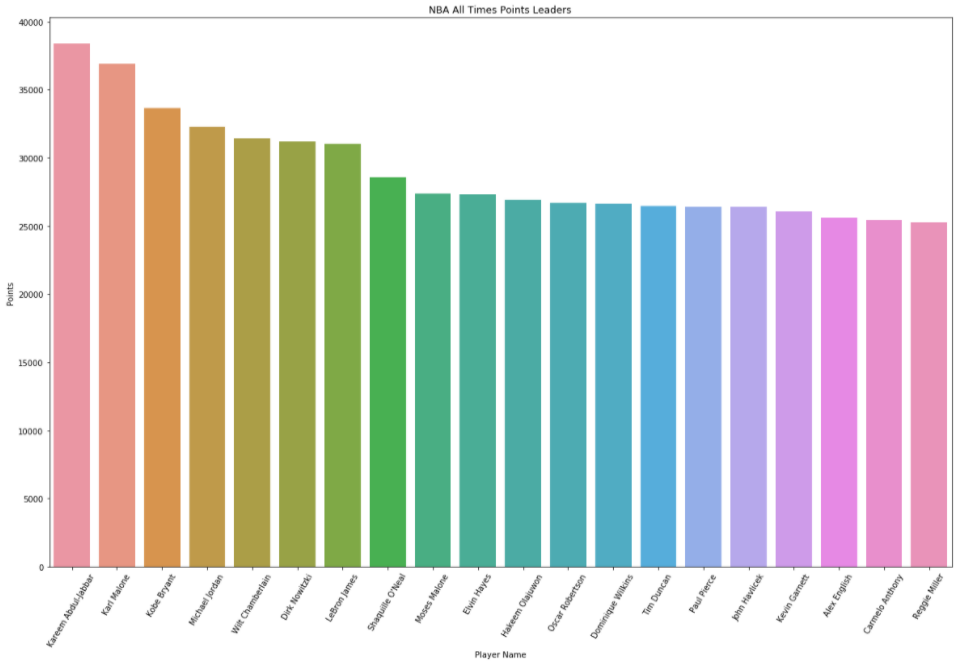
計算球員出賽球季取得，可以看到生涯長度到達18年就已經是前10了，這項數據也與球員年紀分布(大多20歲進入聯盟，幾乎沒人40歲以上)相符合。



- Neural network for predicting Points per game of the players
  作者利用各項數據進行deep learning來預測球員的得分。

2. NBA Leaders and Records

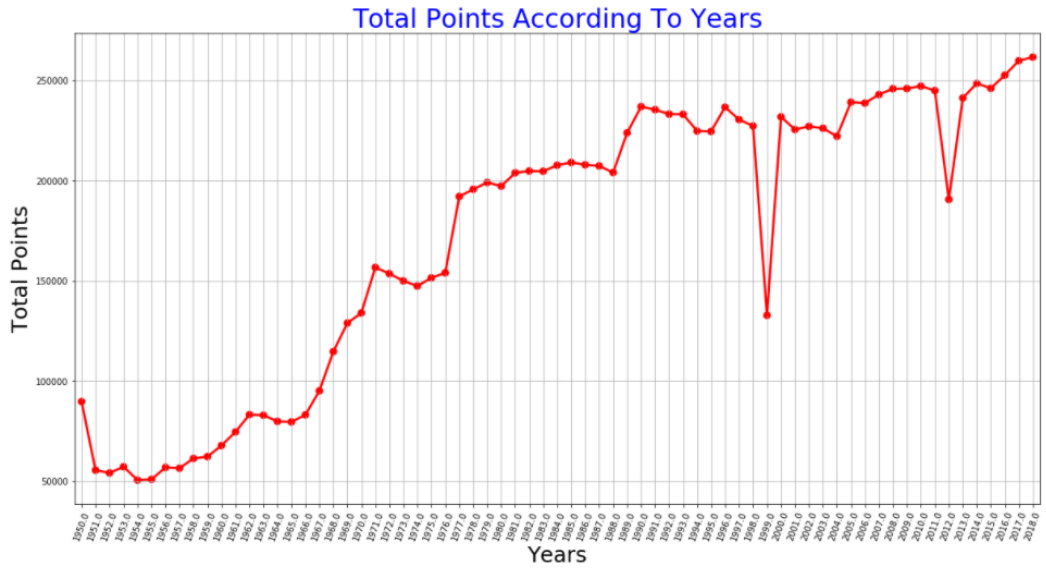
- All Times Points
  將球員生涯得分總計而得



- All Times Assists
  將球員生涯助攻總計而得
- All Times Rebounds

將球員生涯籃板總計而得

- All Times Blocks
  將球員生涯火鍋總計而得
- Total Points According To Years
  每一賽季的各個球員的得分加總，可以看到近年來NBA快節奏的打法讓得分增加許多，而其中1998-99及2011-12得分少是因為開季勞資問題導致場次減少。



- NBA Leaders for 3-Pt Field Goal Percentage
  將三分得手除以出手次數而得
- NBA Leaders for Field Goal Percentage
  將二分得手除以出手次數而得
- Total Points According To Conferences
  計算各隊各球員的得分總計而得

3. 比較

上述的兩個 Notebook 對資料處理的方式有滿大的差異，一個是計算場均數據一個是統計生涯總計數據，兩種方法的前 10 名球員並不會完全相同。場均數據能看出球員在生涯中能否維持一致的表現，但若是球員在一開始的幾個賽季表現很好後馬上退役就會造成偏差，而總計數據是球員整個生涯影響力的重要指標，但較難看出球員平均的表現，生涯總長較長的球員在總計數據也會表現較好。

• My insight

- 很多 Notebook 在一開始都會先對資料集做前處理，將較不

相干的資料去除，著重在有興趣且有用的資料。因此對資料去除雜訊的動作是很重要的，能夠排除無用資料的干擾，同時增加分析的精確性。

- 有些球員在各項數據上表現可能沒有非常出色，但卻是球隊贏球的關鍵因子，可能是這位球員在場上時能讓對方得分下降，也可能是能使自己隊伍的進攻效率提高，這些都是在數據面沒辦法顯現的，因此有時候簡單的數據統計沒辦法真正的展現結果，必須要對各個面向更深入的分析。
- 一般球迷在討論 GOAT(Greatest Of All Time) 時都會把 Michael Jordan 和 LeBron James 兩者拿來做比較，在上面兩個 Notebook 做出的統計數據中，不管是場均數據或生涯總計(如得分、助攻、籃板…)中，Wilt Chamberlain 或 Bill Russell 等傳奇球員都有更好的表現，但 MJ 和 LBJ 還是有較高的聲望，說明了單一面向的資料集(球員的場上個人數據)難以預測出全面的結果，因此我們在詮釋一個資料集時要注意是否有超出資料及所能解釋的面向。