# Análisis multivariado de la performance de universidades en el basketball profesional

Datos del NBA draft 1989 - 2020

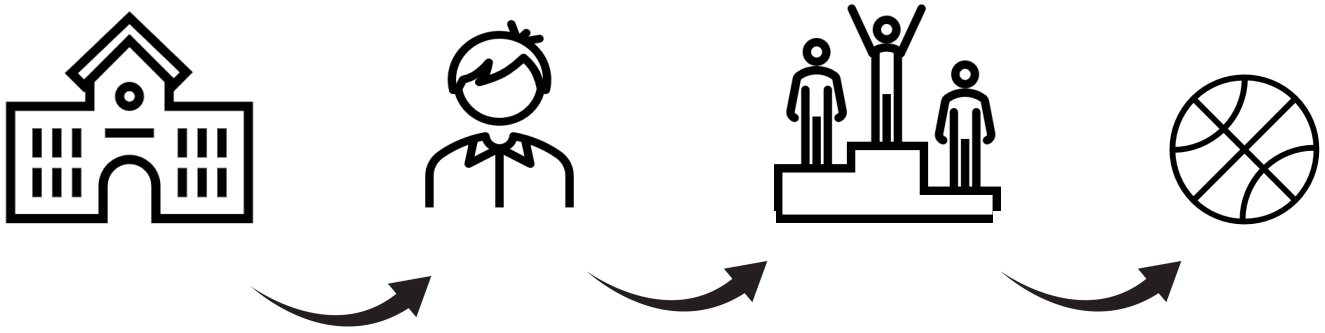
Francisco Fernandez Pedraza

# Introducción:

Con el dataset  que disponemos del draft de la NBA analizaremos desde el año 1989 hasta el 2020 la evolución de las universidades estadounidenses en este deporte y las valoraremos de acuerdo a la performance de sus jugadores en el d  aft de la NBA.

# Objetivo:

Nuestro proyecto va dirigido a estudiantes universitarios y preuniversitarios que quieren dedicarse profesionalmente al basketball. Pretendemos mediante técnicas descriptivas y de aprendizaje de máquinas identificar las mejores opciones de universidades basándonos en datos históricos del NBA draft y sus mejores jugadores



# Preguntas primarias:

1.¿Cuántos jugadores rank 1 lograron cumplir las expectativas dentro de la NBA en sus años de carrera?

2.¿Cuál es el top 10 de universidades que han llevado jugadores a la NBA?

3.¿Qué universidades tienen los jugadores con más años activos en la NBA?

4.¿Cuáles universidades son las mejores si el jugador se quiere desarrollar en anotaciones, asistencias o rebotes?

**Preguntas Secundarias**

- ¿Las universidades con más jugadores en los top5 de cada año son las que más jugadores llevan a los drafts?
- ¿ El mejor top 5 está compuesto por jugadores de la última década, declarando que actualmente los jugadores tienen más nivel.?
- ¿ En la última década de la NBA ha bajado la cantidad de anotaciones, podríamos asumir que se hace énfasis en un juego más defensivo?

# Objetivo:

Nuestro objetivo va dirigido a estudiantes universitarios y preuniversitarios que quieren dedicarse profesionalmente al basketball. Pretendemos mediante técnicas descriptivas y de aprendizaje autónomo identificar las mejores opciones de universidades basándonos en datos históricos de la NBA Draft y sus mejores jugadores

# Descripción de los datos:

- **'id': Numero de identificacion**
- **'year': Año de su ingreso a la NBA**
- **'rank': lugar que fue escogido**
- **'team': Equipo que lo escogio**
- **'player': Nombre del jugador**
- **'college': Universidad de la que se graduo**
- **'years_active': Años en la NBA**
- **'games': Juegos totales**
- **'minutes_played': minutos totales jugados en su carrera**
- **'points': Anotaciones totales**
- **'total_rebounds': Rebotes totales**
- **'assists': Asistencias totales**
- **'field_goal_percentage':Porcentaje de tiro**
- **'3_point_percentage': Porcentaje de tiros de 3 PTS**
- **'free_throw_percentage': Porcentaje de tiros libres**
- **'average_minutes_played': Promedio de minutos jugados**
- **'points_per_game': Promedio de puntos por partido**
- **'average_total_rebounds': Promedio de rebotes totales**
- **'average_assists': Promedio de asistencias**
- **'win_shares': Victorias**
- **'win_shares_per_48_minutes': Victorias antes de los 48 min**

- **'box_plus_minus': Cuantos minutos extra jugaron**
- **'value_over_replacement': Valor en el tiempo**

El dataset contiene 24 variables y 1922 registros donde se reflejan las estadísticas de cada jugador drafteado en la NBA desde 1989 -2020, que nos dara informacion de qué College han logrado una mayor tasa de éxito en crear profesionales de basketball y comparar si de estas universidades han salidos los mejores jugadores de la NBA, a partir de esta información poder prever cuál sería la mejor universidad para que un estudiante vaya a formarse si su objetivo es la NBA.

## Tipo de variables

```
id                          int64
year                        int64
rank                        int64
overall_pick                int64
team                       object
player                     object
college                    object
years_active              float64
games                     float64
minutes_played            float64
points                    float64
total_rebounds            float64
assists                   float64
field_goal_percentage     float64
3_point_percentage        float64
free_throw_percentage     float64
average_minutes_played    float64
points_per_game           float64
average_total_rebounds    float64
average_assists           float64
win_shares                float64
win_shares_per_48_minutes float64
box_plus_minus            float64
value_over_replacement    float64
```

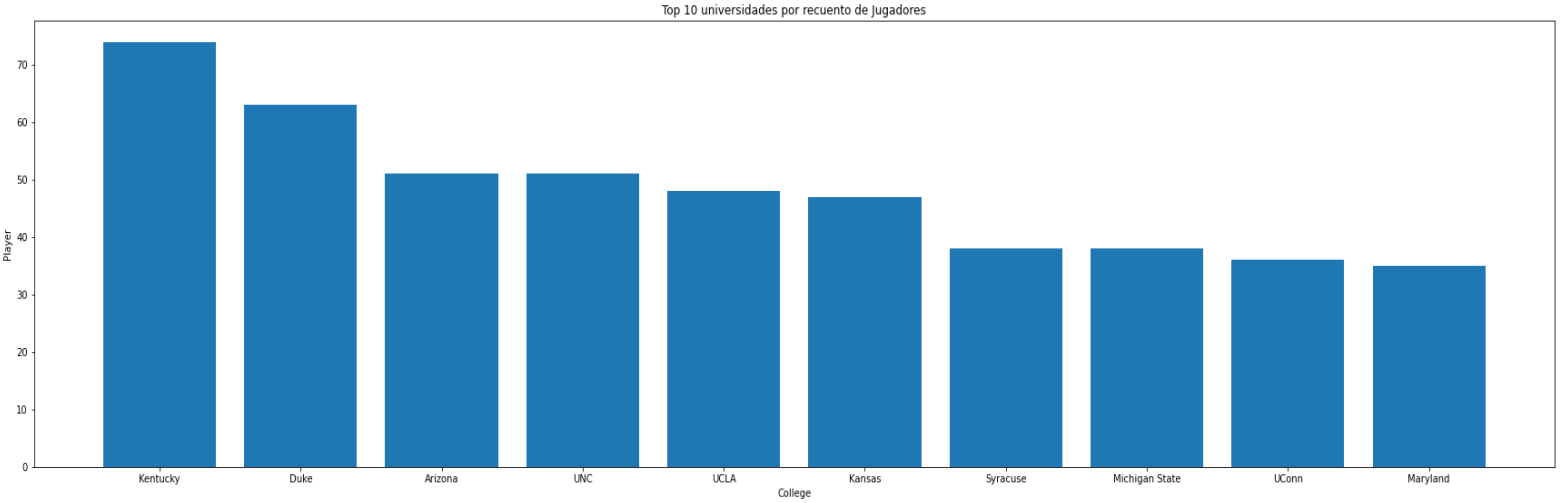
Decidí convertir la variable college y team a números para poder usarlas sin tener que dropearlas, buscando los valores únicos y cambiandolos por códigos que reemplacen el valor de

texto:

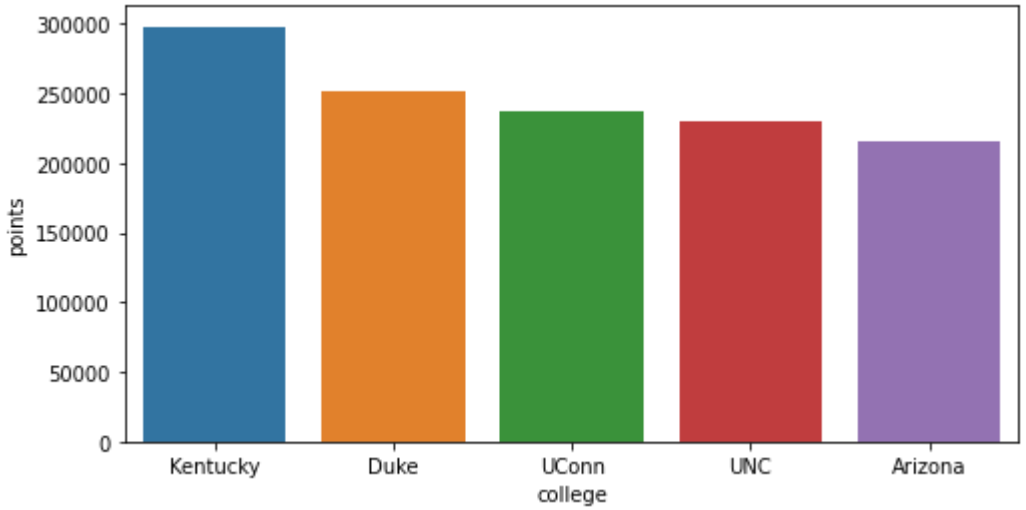
```
id                                int64
year                              int64
rank                              int64
overall_pick                      int64
team                              int64
player                           object
college                           int64
years_active                    float64
games                           float64
minutes_played                  float64
points                          float64
total_rebounds                  float64
assists                         float64
field_goal_percentage           float64
3_point_percentage              float64
free_throw_percentage           float64
average_minutes_played          float64
points_per_game                 float64
average_total_rebounds          float64
average_assists                 float64
win_shares                      float64
win_shares_per_48_minutes       float64
box_plus_minus                  float64
value_over_replacement          float64
```

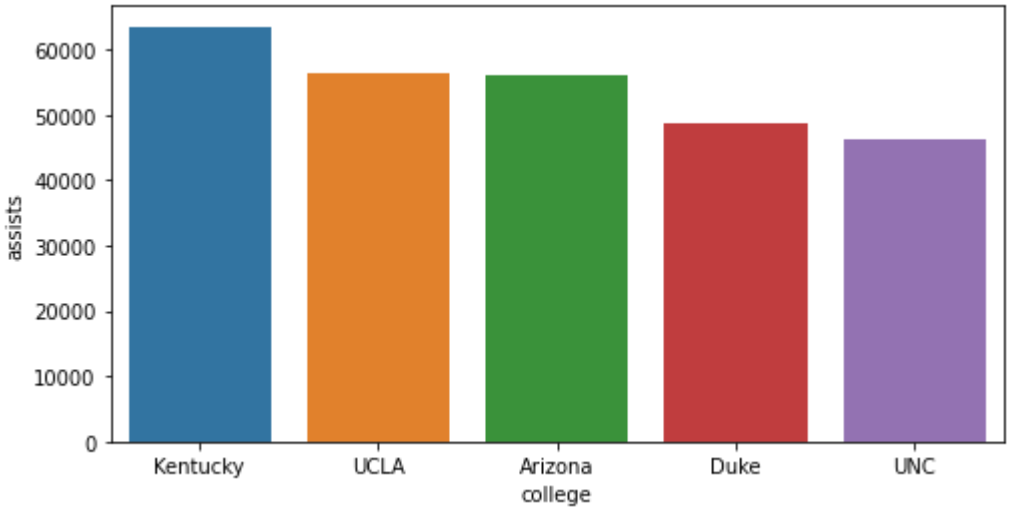
# EDA, Exploratory Data Analysis:

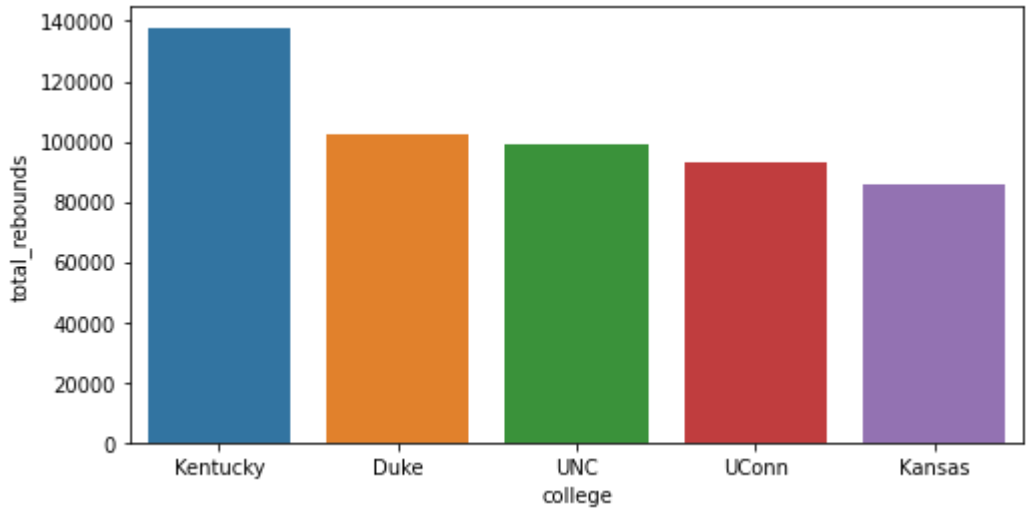
Durante la exploración de datos se busco las 10 universidades más exitosas respecto llevar estudiantes a la NBA:




Partiendo de esto quisimos tambien analizar si eran tambien las mismas con mejores estadísticas en: rebote, puntos y asistencias, datos importantísimos en la NBA:

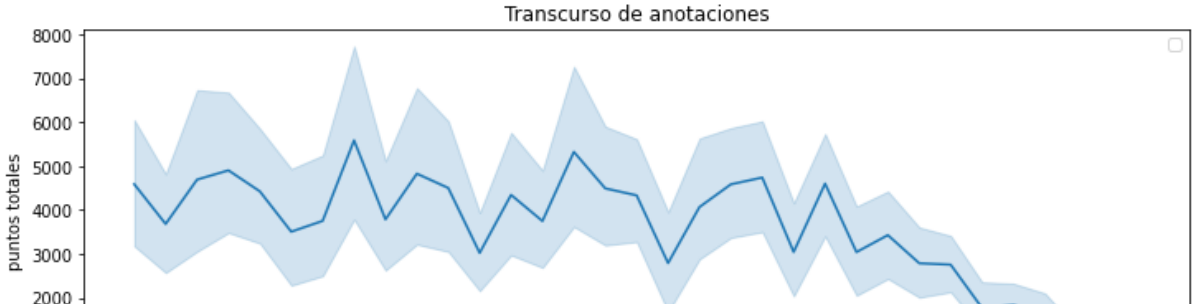




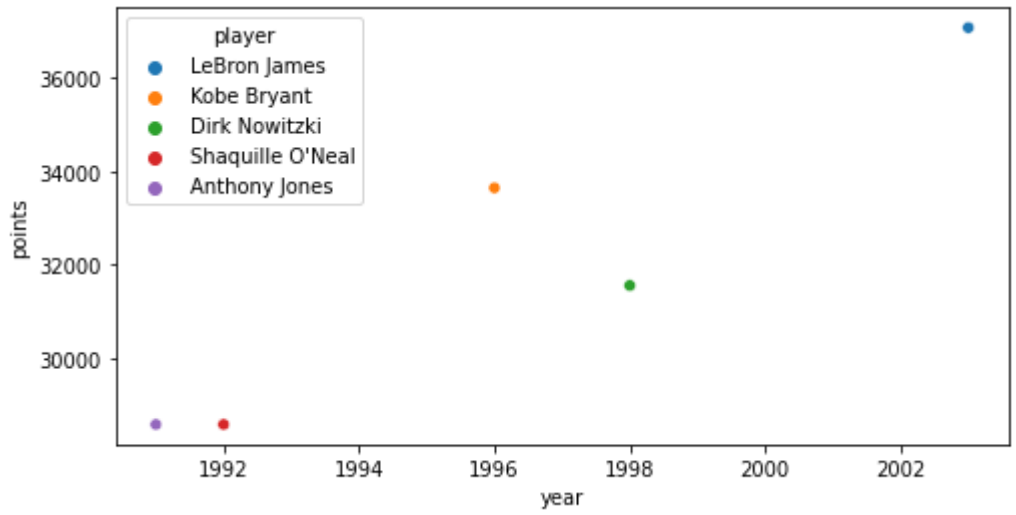




Analizando la cantidad de estudiantes que ha llevado Kentucky a la NBA podríamos excluirla por sesgar la información en diferencia total de jugadores, y de ahí analizar los otros 4 college que se pelean la posición en el top 5.

Por otra parte buscamos analizar la evolución de anotaciones de la NBA:

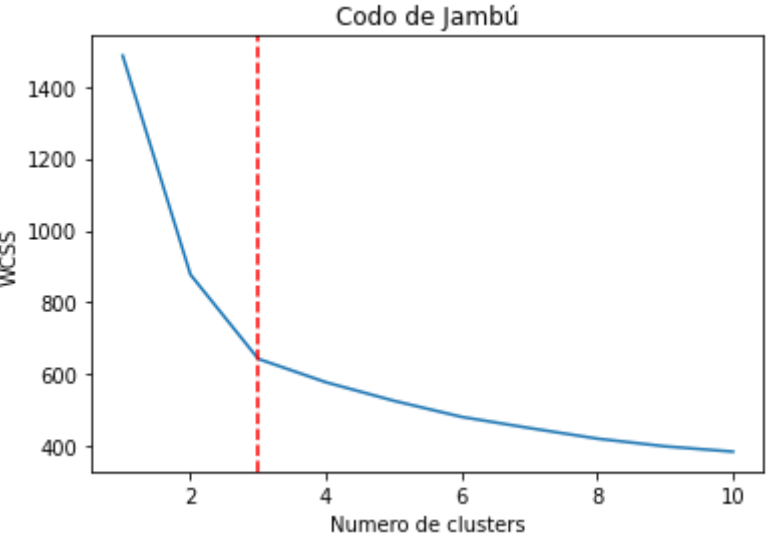


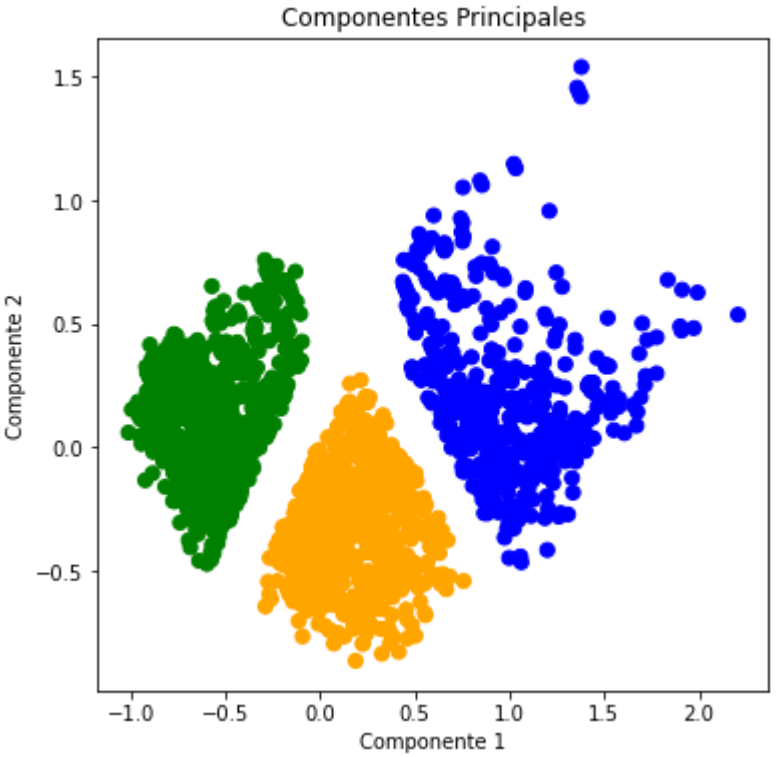
Donde aun si retiramos los últimos 5 años podemos ver que las anotaciones han decrecido que puede resultar en dos casos, un aumento en el estilo de juego defensivo haciendo más difícil  llevar un juego ofensivo o que han disminuido los jugadores ofensivos en la NBA. También se analizó cuales eran los jugadores más decisivos del deporte y que su año de draft refleja el aumento de anotaciones en la historia del juego.




# Algoritmo Elegido:

1. PCA
   Para el PCA, aplicamos el codo de jambu donde se nos reflejo la curvatura en 3 codo, luego para la visualización de esto es que fue que volvimos college y team en variables numéricas, eso nos permitió una división más clara de los 3 grupos de variables:




2. Arbol de decision

Se aplicaron 2 arboles de decisión el A y B. El A fue como estaba el dataset al igual que con el PCA pero el accuracy dio `% de aciertos sobre el set de evaluación: 0.043327556325823226`

Por ello en el B aplicamos el getdummies y nos enfocamos en un solo college en este caso “Duke”, dando mejores resultados `% de aciertos sobre el set de evaluación: 0.9688041594454073`

Validamos:

```
precision    recall  f1-score   support

           0       0.97      1.00      0.99       561
           1       0.00      0.00      0.00        16

    accuracy                           0.97       577
   macro avg       0.49      0.50      0.49       577
weighted avg       0.95      0.97      0.96       577
```

y finalmente con gridsearch optimizamos la busqueda de parametros:

```
Mejores Parametros {'colsample_bytree': 0.6, 'criterion': 'entropy',
'gamma': 0.5, 'max_depth': 5, 'min_child_weight': 1, 'subsample': 0.6}
Mejor CV score 0.9661686769005847
Accuracy del modelo = 0.97143
```

Se aplico tambien una regresión logística a la variable de College_101 (que representa al college Duke) para ver su relación con team , years y years_active.