UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR D´ENGINYERIA INFORMÀTICA

PROYECTO ACADÉMICO:

Visualización de la Información

TRABAJO DE LA ASIGNATURA DE “Visualización” (VIS)

Daniel Oliver Belando & Carlos Gallego Andreu

VALENCIA, ENERO DE 2022

Índice

Introducción

Esta memoria es para la asignatura Visualización, del 3º curso de la carrera de Ciencia de Datos.

En este documento se ha intentado realizar un análisis/visualización para responder a una pregunta, pero para conseguir la respuesta vamos a tener que seguir un proceso. Es por ello por lo que primero de todo se va a explicar en qué consisten las distintas fases de la visualización de datos.

Una vez tengamos claro qué se realiza en cada una de las fases, pasaremos a describir cómo hemos empleado nosotros los conceptos adquiridos en la asignatura para poder conseguir la respuesta a la pregunta formulada.

Objetivo: ¿Qué queremos responder?

Antes de llevar a cabo cualquier análisis o visualización, tenemos que tener algo claro: ¿Qué es lo que queremos saber? ¿A qué queremos responder?

Es por ello que no hay que avanzar sin antes tener una pregunta a responder.

El tema del que queríamos extraer información estaba claro, el baloncesto. Un deporte con gran relevancia social y deportiva en nuestro día a día. No solo eso, sino que también es uno de los deportes con mayor recorrido histórico.

Por eso estuvimos pensando cuál era la liga más importante de baloncesto hoy en día, y no fue difícil adivinar que la NBA (National Basketball Association) es de las ligas más importantes a nivel mundial.

Habiendo decidido el entorno del que queríamos aprender, llegamos a la conclusión de que hay factores que pueden afectar al rendimiento de un equipo, como alguna vez se han señalado los entrenadores, las relaciones personales entre los jugadores y más factores sociales o emocionales. Sin embargo, nunca nos hemos parado a pensar si podría afectar más que lo social, lo geográfico.

Quizá salir de la zona de confort, de tu estadio, puede afectar a la forma de encarar un partido. En definitiva, la pregunta que hemos tratado de contestar en este proyecto ha sido: **¿Cómo afecta la ventaja de campo en la NBA?**

Logo de la *NBA*

Las 7 fases de la Visualización de Datos

Adquisición: Obtención de los datos

Esta primera fase, como el nombre indica, trata sobre la obtención de los datos.

La adquisición de estos datos es fundamental para realizar cualquier tipo de análisis o visualización por muy sencilla que sea, ya que sin ellos no podrás obtener información.

Por ello, vamos a explicar un poco los datos que hemos utilizado:

Los datos han sido recogidos de la web Kaggle, un sitio web dónde se suben ficheros de datos y también se hacen competiciones o la gente comparte sus estudios.

En nuestro caso, encontramos unos datasets llamados NBA games data. Este estaba formado por diversos ficheros de datos que contenían información sobre las temporadas desde 2004 hasta 2021.

Sin embargo, nosotros únicamente hemos utilizado "**ranking.csv**", que contiene las posiciones de cada uno de los equipos junto con información sobre victorias/derrotas, incluso dividido entre casa/visitante, y partidos totales jugados. También usamos "**games.csv**", en el que está la cantidad de puntos, además dividido en casa o visitante, para cada día del año.

Formateado: Modificar datos para que tengan utilidad

Para realizar cualquier tipo de estudio, una vez tienes los datos, es necesario modificar las distintas variables a los formatos que sean de provecho. Quizá hay algún tipo de dato que puede ser más útil en intervalos o similar.

Pues esa capacidad de reconocer en qué formatos nos pueden ayudar más las variables no es fácil de conseguir. A pesar de ello, nosotros hemos logrado modificar los datos que no nos eran nada fáciles de utilizar para, con estos próximos cambios, sacar el máximo partido.

Para ello, hemos utilizado el software R con su IDE Rstudio. En este hemos realizado cada uno de los cambios deseados:

También modificamos los valores de alguna variable porque tenían la forma “integer - integer” y no podíamos utilizarlo cómodamente, así que decidimos separar esos dos integers:

```
g82 = ranking[grepl('-07-20$', ranking$STANDINGSDATE),]
g82$HOME = gsub("-", "", substring(g82$HOME_RECORD, 1, last = 2))
g82$AWAY = gsub("-", "", substring(g82$ROAD_RECORD, 1, last = 2))
g82$STANDINGSDATE = substring(g82$STANDINGSDATE, 1, last = 4)
```

Lo mismo hacemos con las fechas que tenemos en entre los datos, tenemos que indicar el formato que tienen para poder emplearlas en posibles gráficos, como eje de coordenadas, o en plots de R como un filtro temporal:

```
year = format(as.Date(ranking$STANDINGSDATE[1], format="%Y-%m-%d"),"%Y")
date = as.double(year)+c
```

Por último creamos nuevas variables (% victorias y derrotas en casa, y % de victorias y derrotas fuera (VC, DC y VF DF respectivamente), ya que podrían ser útiles para el estudio, y las podríamos aprovechar. También hicimos un merge de 2 los dos conjuntos de datos, para los que tuvimos que formatear los nombres de las variables comunes.

```
g82$VC = g82$HOME / (g82$AWAY + g82$HOME) * 100
g82$VF = g82$AWAY / (g82$AWAY + g82$HOME) * 100
g82$DC = (41 - g82$HOME) / (41 - g82$AWAY + 41 - g82$HOME) *100
g82$DF = (41 - g82$AWAY) / (41 - g82$AWAY + 41 - g82$HOME) *100
names(games)[names(games) == "HOME_TEAM_ID"] <- "TEAM_ID"
names(games)[names(games) == "SEASON"] <- "STANDINGSDATE"
final = merge(g82,games, by = c("TEAM_ID", "STANDINGSDATE"))
write.csv(final,"final.csv", row.names = FALSE)
```

Este CSV final (“**final.csv**”) es el que utilizaremos más adelante en Tableau

Filtrado: Suprimir variables no relevantes.

No solo hay que prepararse las variables de forma que extraigamos el máximo partido sino que también hay que hacer un cribado de los datos o variables que no vamos a utilizar para excluirlas del estudio.

Esta fase nos sirve para reducir la información a procesar, ya que conocemos qué variables no se va utilizar para responder a la pregunta formulada.

En nuestro caso, tuvimos dos procesos distintos de análisis. Por una parte, utilizamos RStudio para crear una herramienta que más adelante se explicará, sin embargo, cabe destacar que excluimos las variables innecesarias con algunas de las instrucciones mostradas en el apartado anterior. En ningún caso las eliminamos del todo, creamos subsets del principal para preservar dicha información por si en algún momento utilizamos tales variables.

Por otro lado, utilizamos el software Tableau en el que se pueden decidir qué variables incorporar a las visualizaciones sin mayor problema. Es por ello que tampoco eliminamos ninguna variable, ya que no habríamos sacado provecho de todo el proceso de filtrado.

Por último, también hemos eliminado alguno de los equipos que aparecía en el conjunto de datos ya que habían errores, por ejemplo, aparecía LA Clippers 2 veces, u Oklahoma City 2 veces.

Minado: Datos en un contexto matemático.

Los datos por sí mismos no expresan nada, son valores con información aparentemente irrelevante, pero esto cambia cuando tratamos de extraer la información gracias al uso de las matemáticas:

Dicha información se puede obtener a partir de modelos estadísticos muy elaborados o, por el contrario, una forma más sencilla pero no simple es relacionando las propias variables del Dataset.

Por nuestra parte, tratamos de extraer la información relacionando las distintas variables, y nuestras estrategias fueron las siguientes:

Primero de todo, aclarar que no todas las gráficas existentes están en el dashboard mostrado en el vídeo, sin embargo, sí las representaciones visuales más útiles en función de lo que consigue explicar.

Por tanto, sobre todo en Tableau creamos nuevas variables o incorporamos medidas sobre las ya existentes. Por ejemplo, en la mayoría utilizamos las medias para crear un valor único y no utilizar intervalos de valores.

En cuanto a las nuevas variables, creamos dos:

Una en la que se almacena la proporción de victorias en casa respecto a las victorias fuera (en porcentaje). Por ejemplo, si un equipo gana 6 partidos en casa pero solo 3 fuera, esta variable tendría como valor: 30% que significa que ha ganado en casa más veces (en dicho valor porcentual) que fuera de casa. (Variable VC - VF)

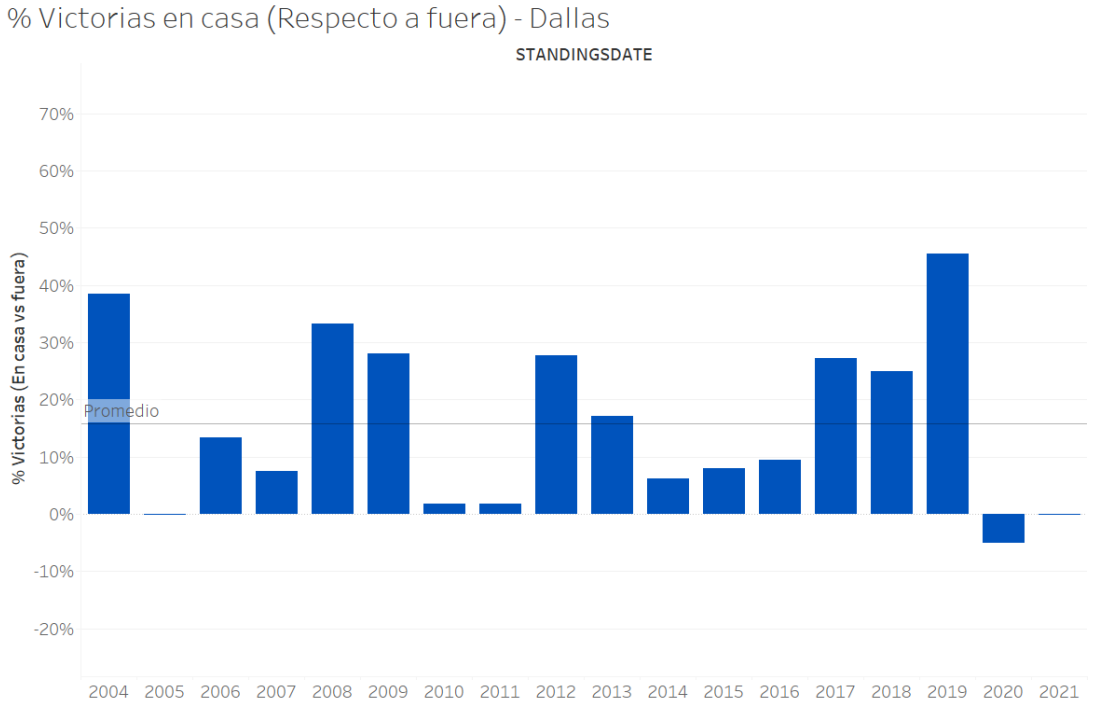
El otro método matemático que añadimos fue una regresión lineal sobre los puntos conseguidos las distintas temporadas, comparando los conseguidos en casa y los conseguidos como visitantes. Esta regresión nos sirve para extraer si se ganaron más puntos fuera o en casa y en qué porcentaje se anotó más en casa respecto a visitante (o viceversa). Más adelante mostraremos cómo la aprovechamos en el gráfico de dispersión

Representación : Modelo visual básico.

En este apartado tuvimos muy presente las clases de Visualización, ya que en esta fase se tiene que proceder a elegir el tipo de gráfico o elemento visual que se quiere utilizar.

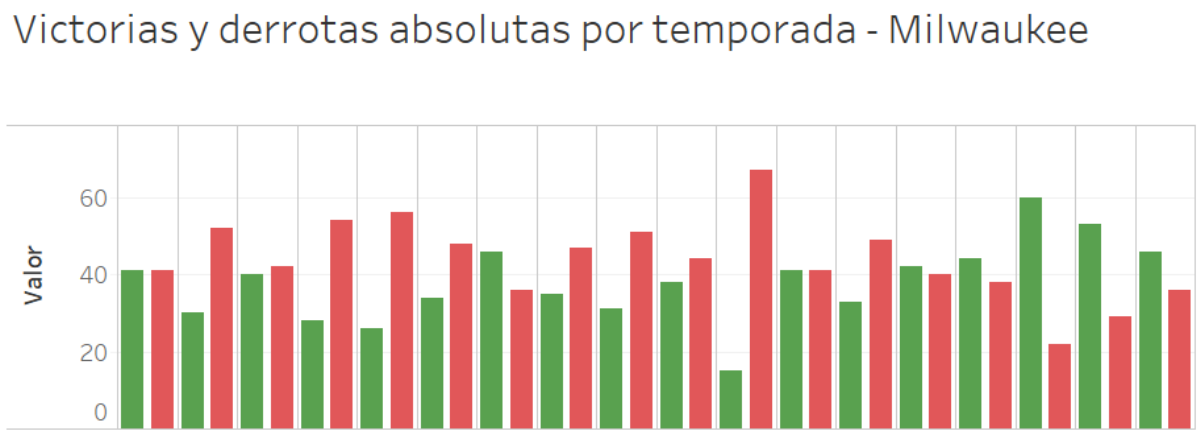
Por tanto, nosotros teníamos claro que queríamos intentar mostrar la información de la forma más clara y certera posible. Es por ello que evitamos gráficos que tuviesen como dimensión principal áreas o similar y nos quedamos con Scatter plots y diagramas de barras que siempre nos ayudan a comparar de manera más sencilla los distintos valores del gráfico.

Primero tenemos un gráfico de barras simples, con el % de victorias en casa con respecto a las de fuera (0% indicaría mismas victorias en casa que fuera), un porcentaje positivo más victorias en casa y viceversa



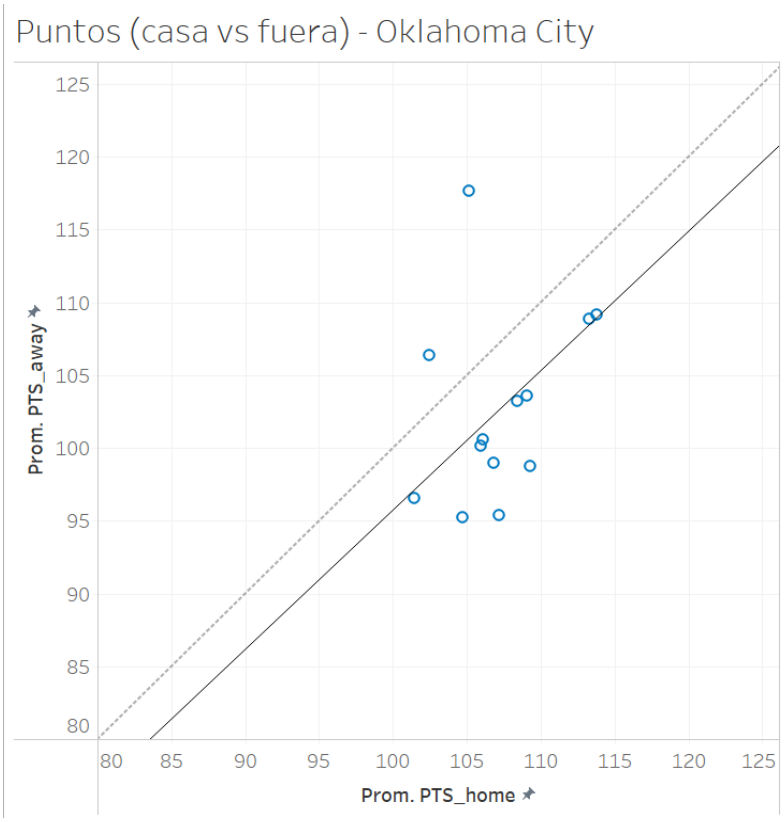
También hemos añadido una línea promedio para ver el valor medio de cada equipo

Para compaginar este gráfico (que es el más útil para entender el efecto de el estadio en cada equipo), añadimos otro gráfico de barras con las victorias y derrotas absolutas de cada equipo, cada temporada, así por ejemplo, una temporada con un alto porcentaje de diferencia en las victorias en casa, y además con muchas victorias será más significativo que una temporada con casi no victorias, ya que la sensibilidad del porcentaje es más alta.



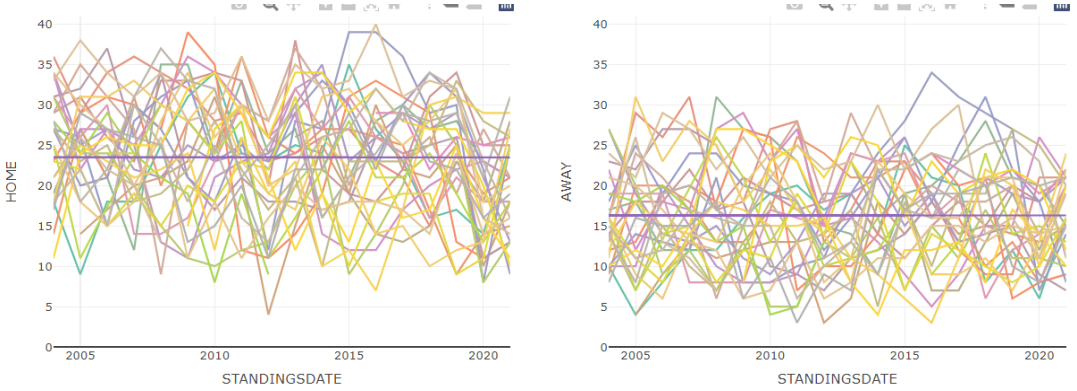
El eje y de temporadas está suprimido para evitar duplicación de información (ya está en el 1º gráfico)

Por último tenemos un gráfico de dispersión comparando los puntos en casa y fuera. Incluyendo esa línea de tendencia para entender mejor si un equipo mete más puntos en su estadio o fuera.



La línea de tendencia de los puntos de las temporadas pasará por encima o por debajo de una línea de referencia de la forma y = x, de 45º de inclinación, de forma que sabremos de un solo vistazo si ese equipo promedia más puntos en casa o fuera. Aquí Oklahoma City promedia más puntos en su casa.

Por último, también realizamos un mini - dashboard en plotly con 2 gráficos que más tarde explicaremos, pero que utilizan gráficos de lineas para comparar los equipos. Decidimos utilizar líneas en lugar de barras por la gran cantidad de equipos que hay.



Los gráficos representan las victorias en casa y fuera absolutas de cada equipo en cada temporada. Hay también añadida una línea promedio con un solo valor, el promedio de las victorias de todos los equipos en todas las temporadas, tanto fuera como en casa.

Refinado: Mejorando la representación.

Aunque se ha explicado mejor y se puede visualizar en el apartado anterior, a algunos gráficos simples les hemos añadido pequeños detalles que pueden ayudar a contextualizar mejor la información. Como puede ser:

- La línea promedio del primer gráfico: Por encima o por debajo de 0 indicará que el equipo gana más en casa o fuera respectivamente, de media.

- La eliminación del eje y en el segundo gráfico, ya que en el dashboard de Tableau coincide con el del primer gráfico.

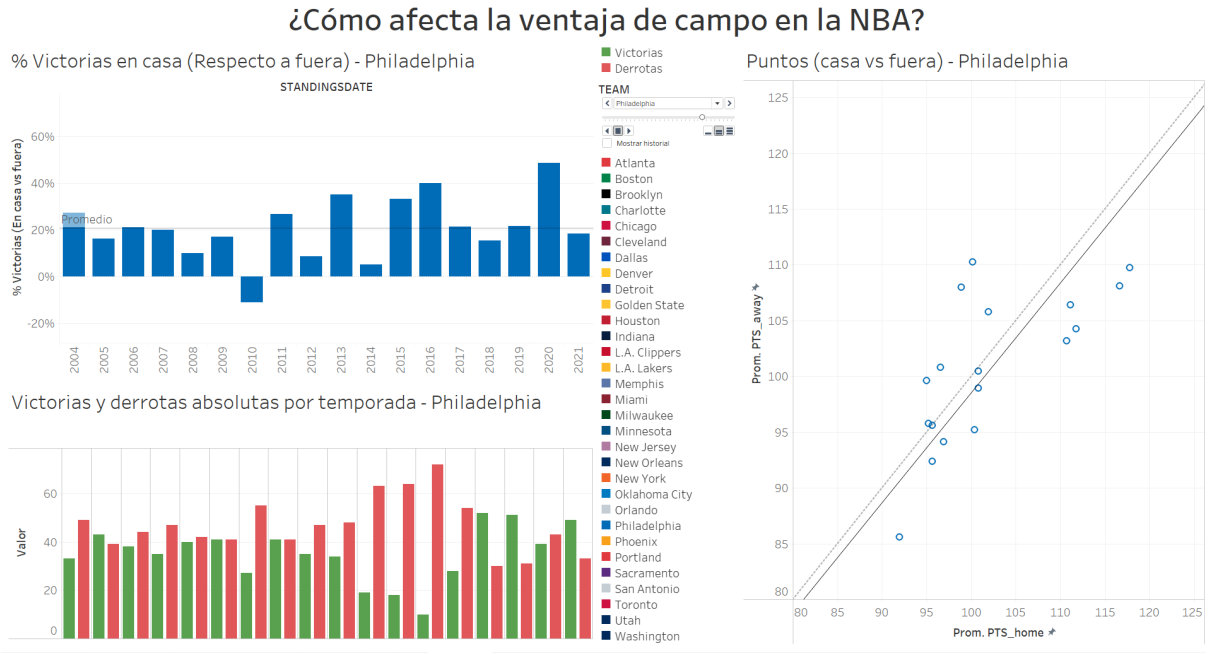
- La línea de referencia y la línea de tendencia del gráfico de dispersión mejoran exponencialemente la representación y la conclusión que queremos sacar de él

- Por último, las líneas de referencia en los gráficos de líneas de plotly nos indican de un vistazo que se ganan más partidos en casa que fuera, y también podemos comparar ese valor con cada equipo, cada temporada.

Otro factor importante ha sido el color, en Tableau, hemos utilizado el color rojo y verde para representar derrotas y victorias, ya que es fácil asociar estos colores a algo negativo y positivo. Por otra parte, los colores de las barras y los puntos en los otros gráficos coinciden con el color de ese equipo, por ejemplo, los Chicago Bulls están representados en rojo, o Los Angeles Lakers en dorado. Nos puede ayudar en un momento dado, a saber de qué equipo estamos hablando. La elección de colores ha sido sacada del enlace: Colores de los equipos en la NBA

Otra cosa destacable es que nos hubiera gustado poder añadir la camiseta de cada equipo a su representación, en forma de jpg o png. No lo hemos conseguido, pero en el vídeo la hemos incluido de forma artificial.

Por último la elección de la posición de los gráficos también ha sido clave. Los dos primeros gráficos están situados uno encima del otro, ya que podríamos decir que van en forma de pack. En el centro hemos situado las leyendas y la paginación (interacción del usuario), de la que hablaremos en el siguiente punto. Y por último a la derecha del todo hemos situado el gráfico de dispersión ya que es un poco secundario, no tiene una relación como la que tienen los dos primeros gráficos



En plotly, se han situado los elementos de interacción en el lado izquierdo, y los gráficos en la parte derecha, el pequeño dashboard quedaría así:

Interacción: El usuario, dueño del Panel de Control.

Para poder representar toda la información de todos los equipos, hemos añadido paginación por la variable equipo en Tableau, el usuario puede elegir qué equipo quiere ver en cualquier momento, los gráficos están conectados de forma que los 3 se actualizan al cambiar el equipo. También es posible pulsar “play”, dejar que el Dashboard corra por su cuenta y observar como cambian las realidades en cada equipo.

En el caso de plotly el usuario tiene 3 formas de interactuar:

- Es posible indicar los equipos que quiere ver, ya que al haber 30 es difícil visualizar la información, a través de una caja de búsqueda (filter select).

- Se puede indicar el nº de victorias en casa, para poder comparar mejor entre equipos con pocas victorias o muchas victorias (filter slider)

- Podemos indicar las temporadas que queremos visualizar marcando las cajas correspondientes. ¡CUIDADO! Porque marcar un sólo año no representará nada. (filter checkbox)

COSAS QUE PODRÍAN MEJORAR: Hubiera sido muy interesante poder añadir una interacción que permitiera alternar variables de los ejes X e Y en Tableau. Ya que tenemos representados los puntos en casa con respecto a los de fuera, también la posibilidad de ver asistencias, rebotes o tiros libres de la misma forma y en el mismo panel de control. Podríamos tener representada prácticamente toda la información útil para el estudio de la pregunta inicial en un solo Dashboard.