Proyecto No. 1
Desarrollo y consulta de Base de Datos
(Parejas)

Modalidad y fecha de entrega

- El proyecto se hará en parejas y debe de ser enviado antes de la fecha límite de entrega: lunes 26 de febrero a las 11:55 a.m.
- No se permitirá la entrega o envío de proyecto más allá de la fecha límite

Descripción general del proyecto

El proyecto contempla el uso de tecnologías de bases de datos para la creación y carga de modelos de datos, con el objetivo de utilizar lenguaje SQL para investigación, desarrollo y presentación de resultados sobre preguntas de negocio para apoyo de toma de decisiones.

El conjunto de datos a utilizar serán los archivos en formato CSV que pueden descargarse en Canvas. Este conjunto de datos incluye todos los juegos de futbol de las cinco principales ligas europeas entre las temporadas de 2,014 a 2,020 (es decir, 7 años) así como información de los jugadores y sus características.

El objetivo general del proyecto es investigar los datos presentados para responder a la siguiente pregunta: **Basado en el desempeño de los equipos y jugadores según este modelo, ¿a qué equipo le apostaría usted? (debe de dar fundamentos basados en los datos)**

Para responder fundamentadamente a esta pregunta usted deberá analizar y entender el modelo de datos presentado, definir métricas que puedan servir de base para justificar una decisión como esta y desarrollar los queries necesarios para calcular dichas métricas.

Etapa 1

En la etapa 1 se encargará de procesar los archivos CSV proporcionados y levantarlos en una base de datos PostgreSQL donde pueda ejecutar sus queries para análisis. Para esto deberá:

1. Descargar los archivos CSV y verificar qué información se incluye en cada uno

2. Crear la base de datos y las tablas en donde almacenará los datos

3. Desarrollar un script de Python (o lenguaje a su elección) que pueda leer la información de los archivos CSV, conectarse a su base de datos y alimentar la información en las tablas creadas

Etapa 2
En la etapa 2 usted ejecutará algunos queries que le permitan familiarizarse con el modelo de datos presentado.

Para esto deberá obtener lo siguiente:

Según estadísticas:

1. La cantidad de juegos jugados en cada temporada por cada equipo, de cada liga (tome en cuenta que cada equipo puede jugar como visitante o como anfitrión.

2. ¿Quién es el mejor equipo de todas las ligas y de todas las temporadas según las estadísticas de diferencia de goles?

   Hint: Obtenga la cantidad de goles a favor, goles en contra y la diferencia entre las dos anteriores, esto por cada temporada y por cada equipo de cada liga.

   Utilizando este mismo query, obtenga el ranking de los equipos por temporada y por liga, ordenados por ese ranking de manera descendente por diferencia (utilice la función `Rank () over patition`), para obtener el equipo ganador.

3. ¿Quiénes son los jugadores que han realizado mayor cantidad de goles a través de todas las temporadas? ¿Cuáles son los jugadores con mayor cantidad de pases izquierdos y pases derechos que han hecho goles? (Compare contra los resultados del inciso 2 y determine de manera textual si dichos jugadores pertenecen a los equipos del inciso anterior).

Según apuestas:

4. Realice un comparativo de las probabilidades de todas las casas de apuesta por temporada, liga y equipo, eliminando aquellos equipos que no tienen estadísticas en ninguna casa de apuesta.

Tome en cuenta de que en la tabla de GAMES se representan los datos de probabilidades de que se gane el local, que gane el extranjero o que empate, según diferentes casas de apuestas como Bet365 (B365), Bet&Win (BW), Interwetten (IW), Ladbrokes (LB), William Hill (WH), VC Bet (VC), etc. Por tanto, escoja el valor más alto de estas columnas.

Luego de obtener las probabilidades correctas, escoja la que mejor le convenga para determinar qué equipo tiene la mayor probabilidad de ganar en qué liga de que temporada.

Tome en cuenta que los valores que aparecen en las columnas (por ejemplo de B365H, B365D y B365A) no son en sí probabilidades porque no se encuentran en el rango entre 0 y 1. Por tanto, para obtener las probabilidades debe de realizar la división de 1 / b365h, por ejemplo. Este es un mecanismo que usan las casas de apuesta para poder confundir al jugador.

5. ¿Cuál es el mejor equipo de todas las ligas y de todas las temporadas según las apuestas?

Hint: Apóyese o complemente el query del inciso anterior para obtener este.

Otros:

6. ¿Quiénes son los jugadores de cada liga y cada temporada que tienen los mejores atributos – características de juego -pases, goles, etc.? ¿De acuerdo a este inciso, y comparándolo con el inciso 2 y 5 anteriores, alguno de los jugadores más valiosos se encuentra dentro del mejor equipo?

7. Obtenga el rendimiento de los equipos en promedio, comparando goles metidos contra la expectativa de goles, determinando qué equipo era quien tenía más expectativa de goles contra quien fue en realidad el que acertó más goles (goals vs expected goals, *xgoals*) en general, pero también es necesario que lo muestre si dichos equipos jugaron como locales o como extranjeros.

8. ¿Cuáles son las características/atributos de los equipos que han sido los líderes de sus ligas en las distintas temporadas? ¿Sus comportamientos son similares?

9. ¿Según la casa de apuesta Beat365 (tome la mejor probabilidad de las 3 medidas), cuales deberían de ser los equipos que tenían la mayor probabilidad de ganar en cada una de las temporadas (*seasons*)?

10. Obtenga el top 10 de estadísticas de los equipos más limpios en jugar (mejor faltas, menos tarjetas amarillas, menos tarjetas rojas) y también el top 10 de los equipos más sucios.

Etapa 3

A continuación, debe plantear sus propias preguntas que le permitan justificar la decisión que tomará acerca de en qué equipo invertirá. Todas sus conclusiones deben estar basadas en el resultado de consultas SQL. Por ejemplo (sugerencias):

- Podría plantearse apostar en el equipo que sea más consistente en la cantidad de partidos que gana por temporada
- Podría plantearse apostar en el equipo que haya mejorado en las últimas tres temporadas
- Podría plantearse invertir en el equipo tienen características que tienen mas goles, menos faltas, más pases, etc
- Etc.

Requerimientos mínimos a completar:

- Se debe presentar el resultado de al menos 18 queries en todo el proyecto
- Deben presentarse al menos tres queries *diferentes* con agrupaciones (GROUP BY)
- Deben presentarse al menos tres queries *diferentes* con JOINs entre dos o más tablas
- Debe presentarse al menos una consulta que haga uso de subqueries

Especificación de tecnología:

- Sistema gestor de base de datos: PostgreSQL
- Interfaz de interacción con base de datos: a discreción

Temas a reforzar:

- Lenguaje SQL: DDL / DML
- PostgreSQL
- Consultas SQL hacia lógica de negocio

Documentos a entregar:

- Archivo comprimido con:
    a. Diagrama Entidad / Relación de la base de datos construida
    b. Script desarrollado para procesar archivos CSV y alimentar base de datos
    c. Documento PDF con las preguntas, queries y resultados obtenidos; que incluya la respuesta a la pregunta de negocio planteada y su justificación

Evaluación:

1. Diseño y construcción de base de datos: 10 puntos
2. Diseño y construcción de script para procesar archivos CSV y alimentar base de datos: 10 puntos
3. Diseño de queries iniciales e interacción con base de datos: 25 puntos
4. Diseño de preguntas preguntas propias y queries para responderlas: 40 puntos
5. Análisis de resultados y presentación de solución a pregunta de negocio: 15 puntos

Total: 100 puntos

Puntos extras:

- Creatividad para presentación de análisis
- Creatividad para presentación de resultados de queries de forma gráfica

Glosario:

Datos de probabilidades de apuestas

B365H = Bet365 home win odds
B365D = Bet365 draw odds
B365A = Bet365 away win odds
BSH = Blue Square home win odds
BSD = Blue Square draw odds
BSA = Blue Square away win odds
BWH = Bet&Win home win odds
BWD = Bet&Win draw odds
BWA = Bet&Win away win odds
GBH = Gamebookers home win odds
GBD = Gamebookers draw odds
GBA = Gamebookers away win odds
IWH = Interwetten home win odds
IWD = Interwetten draw odds
IWA = Interwetten away win odds
LBH = Ladbrokes home win odds
LBD = Ladbrokes draw odds
LBA = Ladbrokes away win odds
PSH = Pinnacle Sports home win odds
PSD = Pinnacle Sports draw odds
PSA = Pinnacle Sports away win odds
SOH = Sporting Odds home win odds
SOD = Sporting Odds draw odds
SOA = Sporting Odds away win odds
SBH = Sportingbet home win odds
SBD = Sportingbet draw odds
SBA = Sportingbet away win odds
SJH = Stan James home win odds
SJD = Stan James draw odds
SJA = Stan James away win odds
SYH = Stanleybet home win odds
SYD = Stanleybet draw odds
SYA = Stanleybet away win odds
VCH = VC Bet home win odds
VCD = VC Bet draw odds
VCA = VC Bet away win odds
WHH = William Hill home win odds
WHD = William Hill draw odds
WHA = William Hill away win odds

Indicadores

XGOALS OR XG: Los goles esperados le asignan un valor a cada disparo entre 0.00 y 1.00, para reflejar la probabilidad de que ese golpeo termine en gol. Un disparo que mide 0.01 xG sugiere que podría ser un gol en cada 100 oportunidades que se repita. En otras palabras,

hay solo un 1% de posibilidades de que termine en gol y, por ende, no es una muy buena opción. Un disparo que mida 0.99 xG, debería terminar en acierto en 99 de 100 veces por un jugador. Así, es una opción de gol que rara vez puede ser fallada.

Si bien la métrica de los goles esperados se utiliza para calcular valores de disparo individuales, suele ser más útil cuando se utiliza en períodos prolongados, como una temporada completa. También es útil para equipos enteros, en vez de jugadores individuales.

XGOALCHAIN: Esta métrica se define por la **cantidad de Goles Esperados en los que ha participado un jugador; si el jugador ha participado en una jugada que acaba en tiro, el xG de la jugada se sumará a esta métrica.** De esta forma, damos valor a jugadores que con frecuencia forman parte de secuencias de pase que acaban en tiro.

XGOALSBUILDUP: se basa en conocer qué jugadores forman parte de cadenas de pase que acaban en tiro pero sin tener en cuenta los dos últimos eslabones -tiro y último pase-, midiendo con ello a centrocampistas, centrales o laterales que tienen una influencia importante en las posesiones de su equipo.

XASSISTS: El modelo de Expected Assists (xA) de Stats Perform mide la probabilidad de que un pase se convierta en una asistencia de gol. El modelo premia a los jugadores que dan un pase en zonas de peligro, independientemente de si el receptor realiza o no un disparo. La xA se mide en una escala entre cero y uno, donde cero representa un pase que nunca resultará en asistencia y uno representa un pase del que se esperaría que el receptor anotara siempre.