DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

TP Final

24 de agosto de 2020 — Calidad de datos

Integrante	LU	Correo electrónico
Millassón, Matías	131/13	matiasmillasson@gmail.com
Giusto, Maximiliano	486/05	maxi.giusto@gmail.com
Venegas, David	783/18	venegasr.david@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
https://exactas.uba.ar

Índice

1. Introducción

La industria del fútbol mueve billones de dolares al año, y cada día es de mayor valor para el mercado su información. Ante esto, es de interés para el negocio asegurar la calidad de su data, por lo cual el presente trabajo se busca analizar y limpiar el dataset correspondiente a los partidos oficiales de fútbol internacional (es decir entre selecciones de diversas federaciones) desde 1872 hasta el presente, para realizar así una apreciación sobre el nivel de sus cualidades intrínsecas de calidad de datos, así como efectuar una análisis descriptivo y bivariado de sus variables, para finalmente determinar reglas de negocio y métricas correspondientes.

2. Resultados No Técnicos (Parte Gerencial)

2.1. Descripción Funcional

La base de datos permite la rápida búsqueda de partidos por campeonato, país donde se jugo, ciudad o equipos. Sin embargo, el no encontrar un resultado no garantiza que no exista ese juego sino solo que no esta registrado ya que el dataset no es 100 % exacto, sin embargo, tiene un margen de error aparentemente pequeño. La información más relevante que pudimos obtener es:

- Alrededor de la cuarta parte de los partidos son jugados en estadios neutrales
- Es mas frecuente que un equipo gane los partidos de local.
- En Europa es donde hay más partidos y a la vez donde hay más ciudades con la infraestructura suficiente para albergar un partido entre selecciones. En los países geográficamente chicos pero con buen desarrollo de la disciplina (por ejemplo Uruguay) se puede apreciar una ciudad que predomina por sobre el resto.

2.2. Informe sobre la Calidad

El estado de la calidad del dataset es bueno, con un alto nivel de completitud (no hay datos faltantes) y no tiene múltiples copias. Existes muy pocos partidos repetidos (menos del %0.01) y se poseen datos hasta de este año (2020). Por otra parte la precisión se podría mejorar, aunque esto representaría una inversión considerable de tiempo de desarrollo, ya que seria necesario registrar las fuentes de cada partido para llevar un mejor control de la veracidad de la información, sobre todo para juegos del siglo XIX.

3. Resultados Técnicos

El dataset incluye 41586 resultados de fútbol internacional desde 1872 hasta 2020, con partidos pertenecientes tanto a los mundiales de fútbol de la FIFA como a juegos amistosos. No incluye juegos olímpicos o partidos de selecciones juveniles. A continuación se describe cada columna de la base de datos:

- date: fecha del juego.
- home_team: equipo local.
- away_team: equipo visitante.
- home_score: puntaje del equipo local en tiempo completo + tiempo extra (no incluye penales).
- away_score: puntaje del equipo visitante en tiempo completo + tiempo extra (no incluye penales).

- tournament: el nombre del campeonato
- city: ciudad donde se jugo el partido.
- country: pais donde se jugó el partido.
- neutral: TRUE/FALSE indicando si el partido se jugó en un estadio neutral (no pertenecía a ninguno de los dos equipos). Notar que cuando este valor es verdadero home_team y away_team podrían ser intercambiados.

Es importante destacar sobre la metadata que dado que los nombres de país pueden variar con el paso de los años en el dataset se establecieron ciertas reglas para garantizar la coherencia histórica de éstos. Para los campos de *home_team* y *away_team* se utiliza el nombre del equipo correspondiente a su sucesor histórico para facilitar rastrear la historia de un equipo a lo largo del tiempo o hacer cálculos estadísticos. Ejemplo: en el caso de la unión soviética como equipo de fútbol se utiliza en nombre de Rusia.

Con respecto a los nombres de países, es utilizado el nombre oficial al momento del juego, aunque este no coincida con el nombre del equipo. Esto se puede deber a que la selección no represente a un país sino que represente a una nación. Un ejemplo es cuando Ghana jugo en Accra en 1950, aunque son diferentes el nombre de *home_team* y *country*, era un juego local para Ghana. Esto se indica en la columna neutral, que marca FALSE para estos partidos, implicando que no era un estadio neutral. Otro ejemplo es el caso del juego de Rusia-Finlandia en 1968, aunque el equipo local aparece Rusia por las razones mencionadas anteriormente, el valor de la variable *country* es la Unión Soviética.

3.1. Cualidades Intrínsecas de la Calidad

3.1.1. Precisión

Todos los campos del dataset tratan de representar el mundo real, dado la naturaleza histórica del mismo al registrar los partidos de fútbol. En el campo de *home_team* y *away_team* se puede encontrar 312 equipos diferentes. Esta cantidad es muchos mayor que las federaciones afiliadas actualmente a la FIFA pero recordemos que el dataset incluye partidos disputados por naciones que hoy en dia no existen pero en su momento tenían su propio equipo, como el caso de URSS, o bien partidos que no son parte del universo de la FIFA. Es importante resaltar que debido a lo extensa que es la historia del fútbol así como el grado de autenticidad de las fuentes de este dataset[1] es imposible alcanzar una total precisión histórica en esta base de datos, pero dada la cantidad de partidos que tiene podemos afirmar que es de gran utilidad como referencia de búsqueda. Ejemplos de estas imprecisiones son ausencias de juegos de la vida real como el partido Cuba-Panamá del 01/08/2016. Además, existen otros errores en algunos casos de la variable *neutral* tanto por razones burocráticas como ambigüedades técnicas. Por último los partidos con penales son mostrados como empates y no se posee información extra sobre el ganador oficial en la base de datos. Esta última decisión la vemos muy acertada ya que no tiene sentido mezclar los goles de una definición por penales con los goles realizados en el tiempo regular.

3.1.2. Completitud

Analizando la presencia o ausencia de características en la tabla, se evidenció que existe un alto nivel de información completa, con un 0 % de datos faltantes para las 41586 filas del dataset. A nivel lógico el modelo de datos es conciso y minimal, sin embargo se podría ampliar para proveer mayor información sobre los juegos a riesgo de un incremento en los datos faltantes.

[1] http://www.rsssf.com/, https://www.fifa.com , hhtps://www.wikipedia.com, sitios de las federaciones o confederaciones, etc

3.1.3. Consistencia

Evaluar la coherencia de los datos representados en múltiples copias en este caso de estudio no es posible ya que el dataset es único.

3.1.4. Unicidad

Cada combinación de fecha, equipo visitante, local y ciudad es única para casi todos los resultados excepto en 2 excepciones, en el caso del juego Tahiti vs. New Caledonia en 17/02/1974, donde su duplicado tiene invertido los goles. El otro caso es el partido de Guyana vs. Barbados el 22/10/1977. Se corroboró para ambos que no se jugó el mismo partido dos veces en un día.

3.1.5. Actualidad

El dataset esta actualizado hasta el año 2020 siendo la fecha de su último registro 01/02/2020 para Estados Unidos contra Costa Rica.

3.2. Análisis Descriptivo

A continuación se listan las principales variables estudiadas con sus respectivas métricas como media, moda, suma y desviación en en caso de las numéricas y frecuencia en el caso de las categóricas. Para la variable binaria solamente vamos a analizar la frecuencia.

3.2.1. Goles del local

La variable *home_score* tiene una media de 1,74, con una desviación estándar de 1,75, mediana de 1 y moda de 1. Además posee un valor mínimo de 0 goles, máximo de 31 y un total de 72599. Estos números son razonables ya que es raro que un equipo anote más de tres goles en un partido.

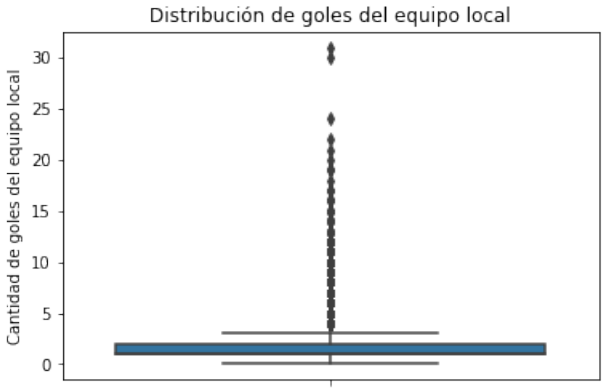

Figura 1: Distribución de la variable *home_score*

3.2.2. Goles del visitante

La variable *away_score*, tiene una media de 1,187587, con una desviación estándar de 1,405323 y una mediana de 1,0 y moda de 0. Por ultimo tiene un mínimo de 0 goles, un máximo de 21 y una suma total de 49387. Tiene sentido que la distribución sea similar a los goles del local. Aunque es levemente menor ya que es más difícil desempeñarse bien en un estadio desconocido.

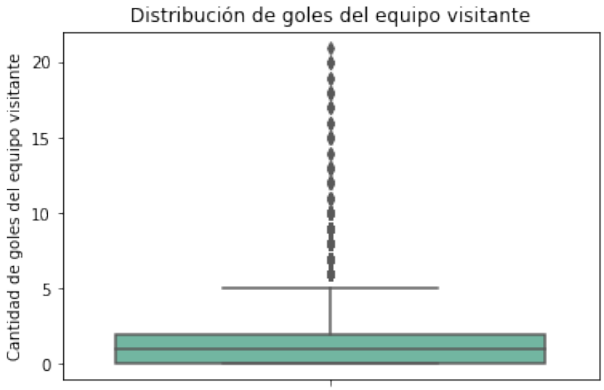

Figura 2: Distribución de la variable *away_score*

3.2.3. Fecha del partido

En la variable *date* se pueden encontrar 15196 días de juego para los 41586 partidos del dataset, con la notaria superposición de juegos en un mismo día, especialmente durante campeonatos o fechas FIFA (fechas en la que la FIFA les impone a las federaciones que jueguen partidos amistosos o por torneos). A su vez en la figura 3 se puede apreciar una tendencia creciente a través de los años a una mayor cantidad de partidos por día, debido al aumento en popularidad del deporte a nivel mundial y la afiliación de federaciones a las confederaciones. La fecha con la mayor cantidad de eventos es el 29-02-2012 con 66 partidos, que al igual que el resto de los valores más grandes resultó ser una fecha FIFA.

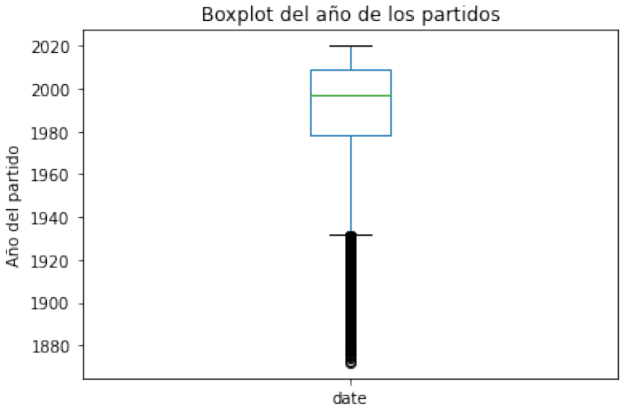

Figura 3: Distribución del año de los partidos

3.2.4. Estadio neutral

La variable *neutral* es binaria. Hay un 24,7 % de juegos neutrales como se observa en la figura 4 por lo que la mayoría de los partidos no se jugaron en estadio neutral. Este gran dominio de los partidos que no son en estadios neutrales es un resultado sensato ya que hasta hace algunos años viajar por el mundo era caro, por lo que se intentaba minimizar el trasporte a la hora de organizar amistosos. Además hay varios torneos en los que el reglamento exige que se juegue un partido en un estadio de cada equipo.

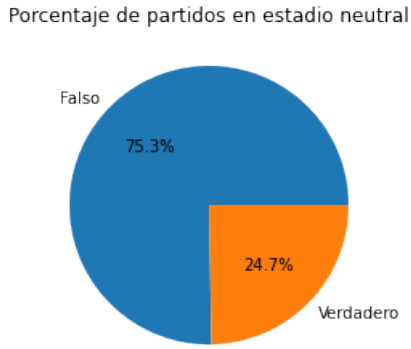

Figura 4: Frecuencia de la variable *neutral* en el dataset

3.2.5. Torneo

La variable *tournaments* toma 78 valores posibles siendo el mas frecuente Friendly con 17029 juegos, luego está la clasificación al mundial de la FIFA con 7236 y en tercer lugar la clasificatoria de la Eurocopa. A continuación se muestran los principales 20 torneos del dataset en la tabla 1. Las clasificaciones a las copas de las confederaciones o a la copa mundial de la FIFA tienen como objetivo seleccionar a las mejores federaciones por lo que todos los equipos deben poder participar y para elegir pocos equipos se necesitan jugar muchos partidos. Al mismo tiempo cualquier selección puede organizar amistosos contra otras.

torneo	frequencia
Friendly	17029
FIFA World Cup qualification	7236
UEFA Euro qualification	2582
African Cup of Nations qualification	1672
FIFA World Cup	900
Copa América	813
AFC Asian Cup qualification	724
African Cup of Nations	690
CECAFA Cup	620
CFU Caribbean Cup qualification	606
British Championship	505
Merdeka Tournament	503
Gulf Cup	380
AFC Asian Cup	370
Island Games	350
Gold Cup	327
AFF Championship	293
COSAFA Cup	292
UEFA Euro	286
Nordic Championship	283

Cuadro 1: Tabla de frecuencia de los 20 torneos con mas partidos del dataset

3.2.6. País

En el caso de la variable *countries* existen 177 países donde tuvo lugar un partido del dataset, siendo el mas común Estados Unidos con 1179 debido a que además de los partidos de su seleccionado albergó muchos partidos amistosos de México. Le sigue Francia con 806, e Inglaterra con 696, en el caso de Asia el mas popular resulto ser Malasia con 562 juegos. Malasia fue sede de varios torneos Merdeka, por eso tiene muchos partidos. La razón por la que Catar está entre los mejores 10 (a pesar de que su primer partido fue en 1974) porque los empresarios del petróleo se están encargando de llevar muchos partidos amistosos. Tailandia hace más de 50 años que organiza la King's Cup haciendo que tenga muchos partido llevandolo al décimo lugar. Francia, Inglaterra, Suecia, Alemania, Brasil y España son países muy importantes en el mundo del fútbol.

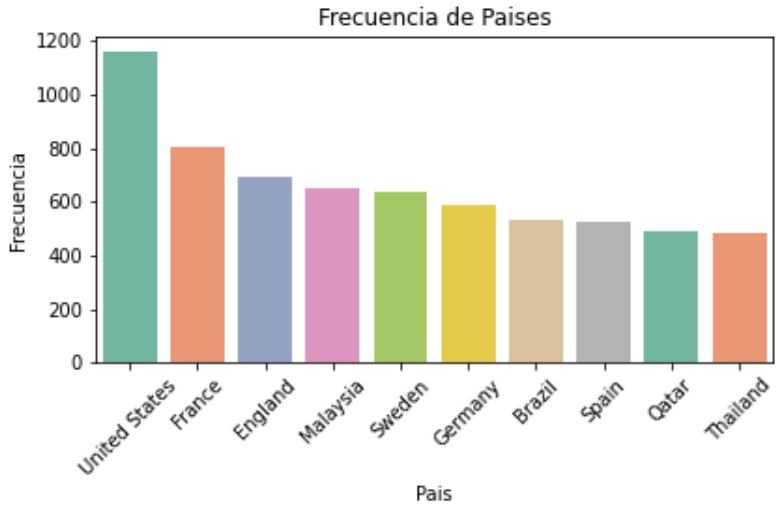

Figura 5: Frecuencia de los principales 10 países

3.2.7. Ciudad

Para la variable *city* se puede ver en el mapa 6 que la mayor parte de las ciudades están en Europa. América Central pareciera tener muchas ciudades pero la realidad es que hay muchos países muy pequeños. Kuala Lumpur queda como la ciudad con la mayor cantidad de partidos de las 187 ciudades en el dataset, llega a esta posición un total de 589, seguida de Doha con 459 y Bangkok con 417. Yendo a Europa nos encontramos con la capital de la cuna del fútbol, Londres (395) y Budapest (386). En América se destaca Montevideo con 350 partidos. Uruguay es un país con poca extensión y una gran historia futbolística por lo que la mayor parte de sus partidos se jugaron en el mejor estadio de su capital.

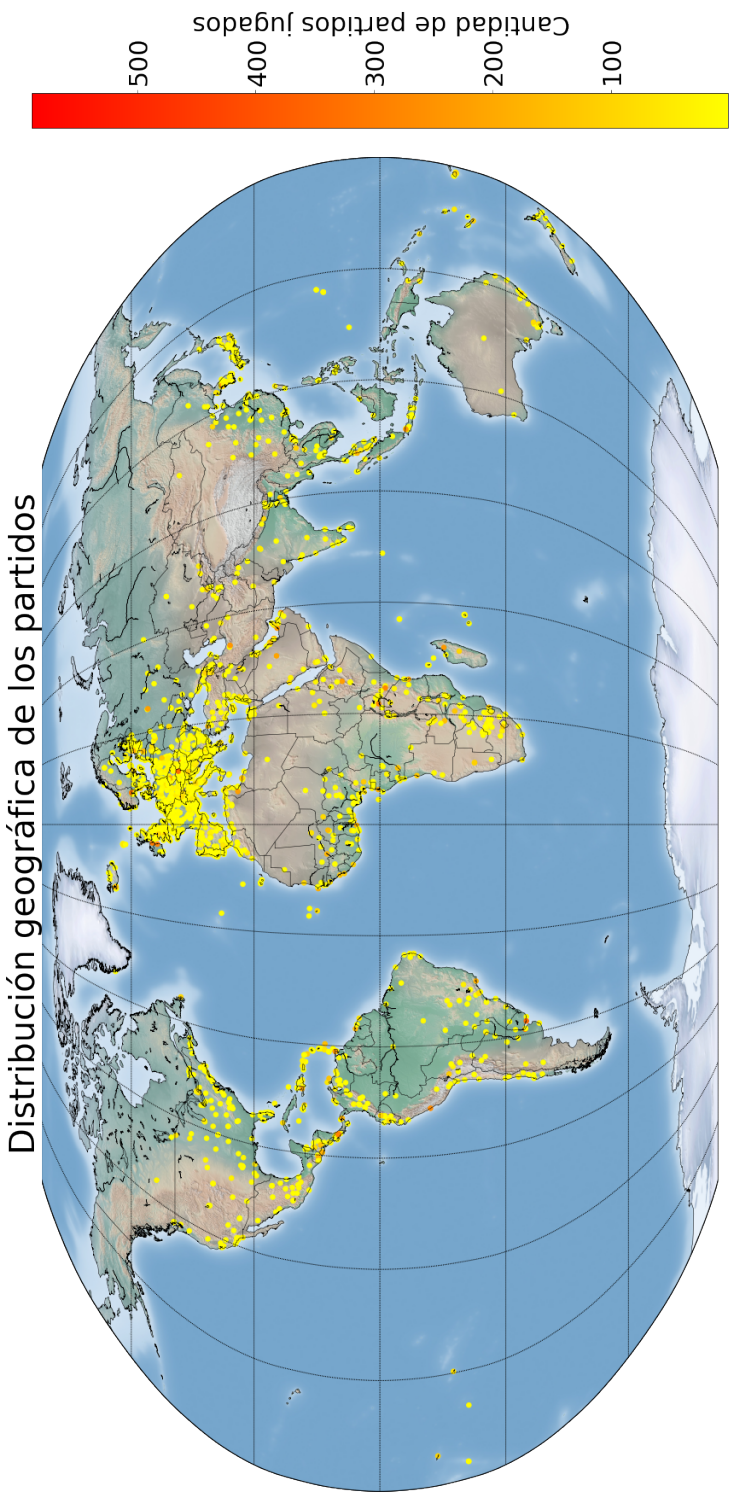

Figura 6: Cantidad de partidos por Ciudad

3.2.8. Equipo local

Para la variable *home_team* en la figura 7 podemos ver a cuatro potencias como lo son Brasil (568), Argentina (548), Alemania (506) e Inglaterra (498). Además Mexico fue local en 513 ocasiones.

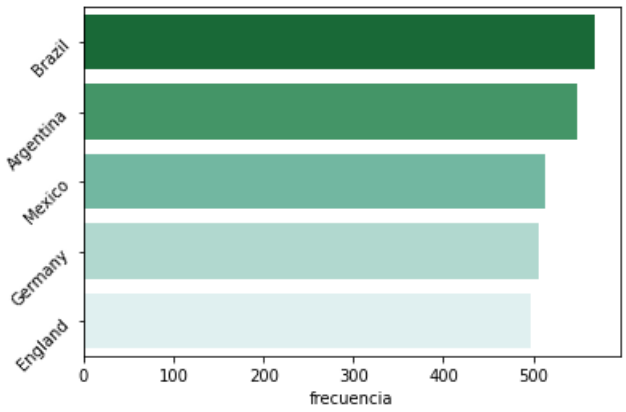

Figura 7: Frecuencia de los principales 5 equipos locales

3.2.9. Equipo visitante

En esta sección analizaremos la variable *away_team*. En la figura 8 podemos ver que Uruguay es la selección con más partidos jugados de visitante. Luego nos encontramos con tres selecciones europeas de las cuales solo Inglaterra es una potencia. Hungría fue una potencia hasta la década de 1970. Suecia y Paraguay jugaron muchos torneos en con sede fija pero no los organizaron.

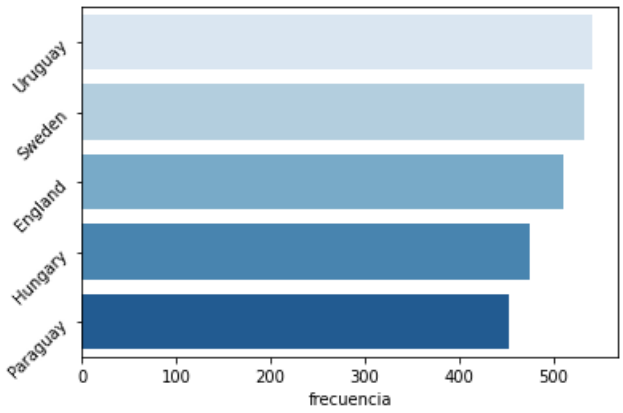

Figura 8: Frecuencia de los principales 5 equipos visitantes

3.3. Análisis bivariado

En esta sección vamos a hacer cinco análisis bivariados. Vamos a analizar dos variables numéricas (goles locales y visitantes, goles visitantes y fecha), dos variables categóricas (país y neutral, torneo y neutral) y una variable numérica con una variable categórica (goles locales y equipo local) para tratar de inferir información sobre el dataset y obtener información para el negocio.

3.3.1. Goles locales y goles visitantes

En esta sección se analizan dos variables numéricas del dataset, que son de gran relevancia en el negocio pues la cantidad de goles determinan el ganador de un partido, y por ende cualquier tipo de información al respecto es relevante para el negocio.

En la tabla de contingencia del cuadro 2 utilizando el índice de Pearson se aprecia que no hay una correlación directa entre ambas variables, lo cual es lo esperado, pues de haberla se debería revisar los reglamentos oficiales de los torneos con respecto a la localidad de los partidos y la cantidad de los mismo para lograr mayor equidad. Por otra parte, como se menciono anteriormente en el presente trabajo en el diagrama de dispersión se puede apreciar que hay una mayor tendencia a anotar goles por el equipo local, encontrando outliers de hasta 30-0 goles, a diferencia de 21-0 para el visitante, lo cual es lógico ya que es mas difícil desempeñarse igual en un estadio desconocido en comparación al del entrenamiento.

	home_score	away_score
home_score	1.000000	-0.136095
away_score	-0.136095	1.000000

Cuadro 2: Tabla de contingencia de goles locales y visitantes

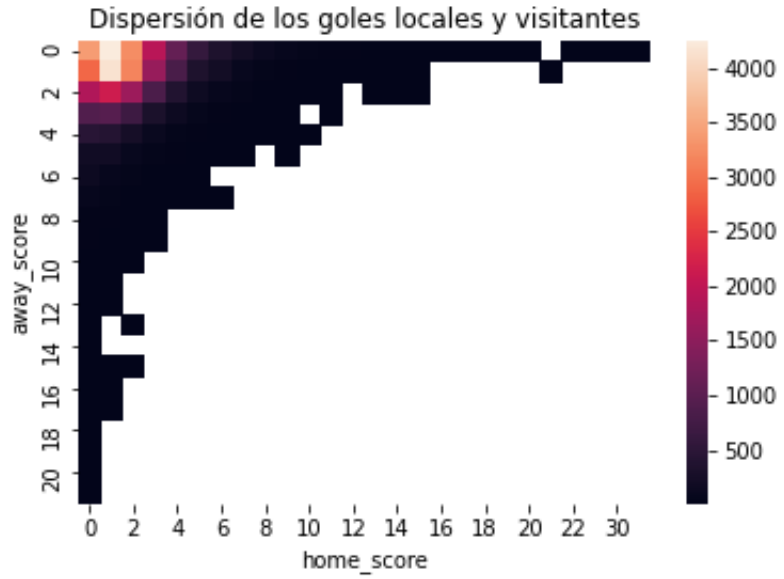

Figura 9: Heatmap de Dispersión de los goles

3.3.2. País y estadio neutral

En este sección vamos a analizar cómo se distribuyen los partidos neutrales en cada país. En el gráfico 10 podemos observar que los únicos países que tienen una gran cantidad de partidos neutrales son Estados Unidos y Catar (muchos amistosos entre selecciones de renombre), Malasia (organiza la King's Cup), Francia (organizó varias copas mundiales y copas de Europa). En general los partidos neutrales no dominan en ningún país.

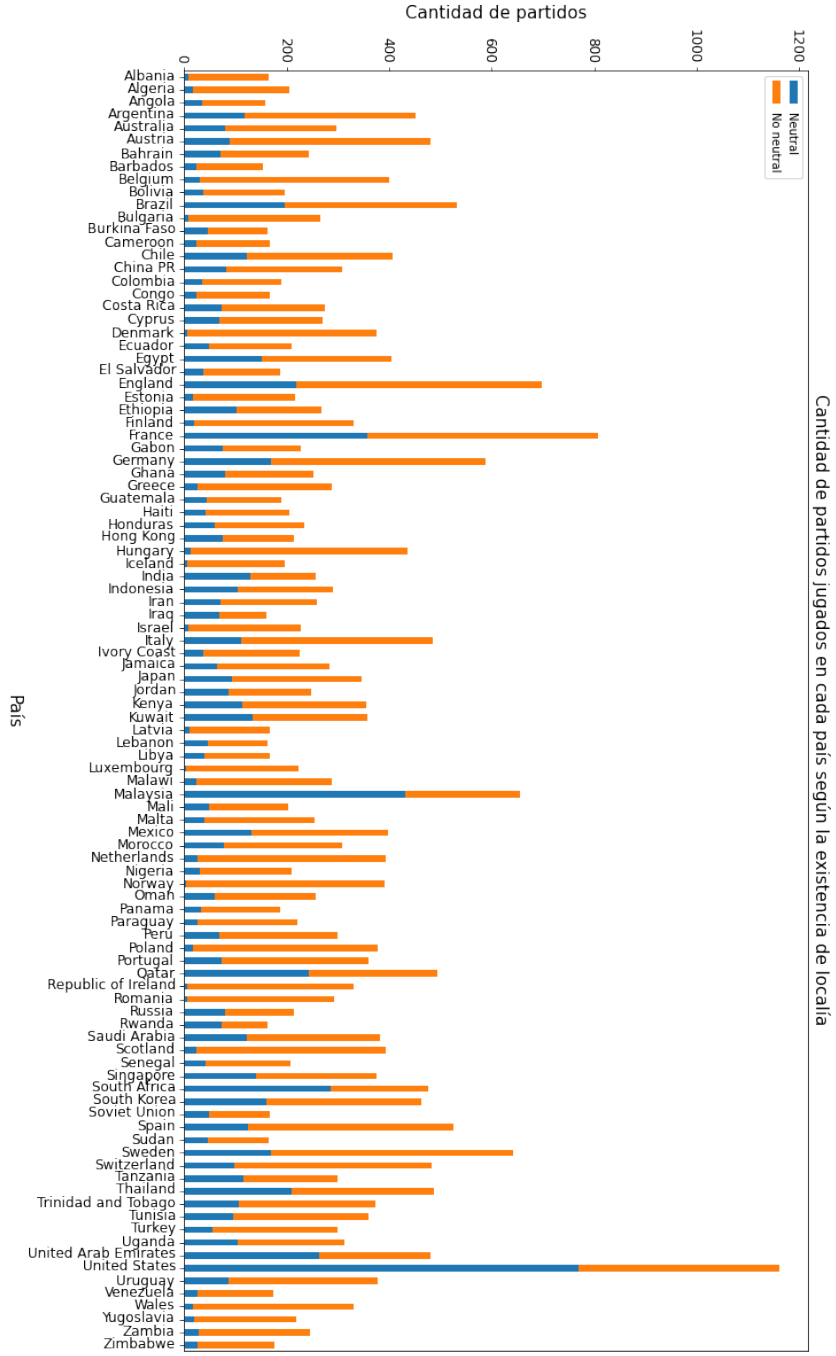

Figura 10: Partidos por país

3.3.3. Goles locales y equipo local

En este caso vamos a analizar la variable *home_score* y *home_team*. En el gráfico 11 podemos encontrar la suma de los goles convertidos en condición de local. Dado que son muchas selecciones solamente mostramos las que hicieron más de 200 goles. Inglaterra, Alemania, Hungría, Países Bajos y Suecia son las selecciones más goleadoras de local. Brasil, Italia y Francia están un escalón más abajo. A excepción de Suecia todas son selecciones muy destacadas a lo largo de la historia. Además el fútbol es muy popular en las naciones correspondientes a esas selecciones.

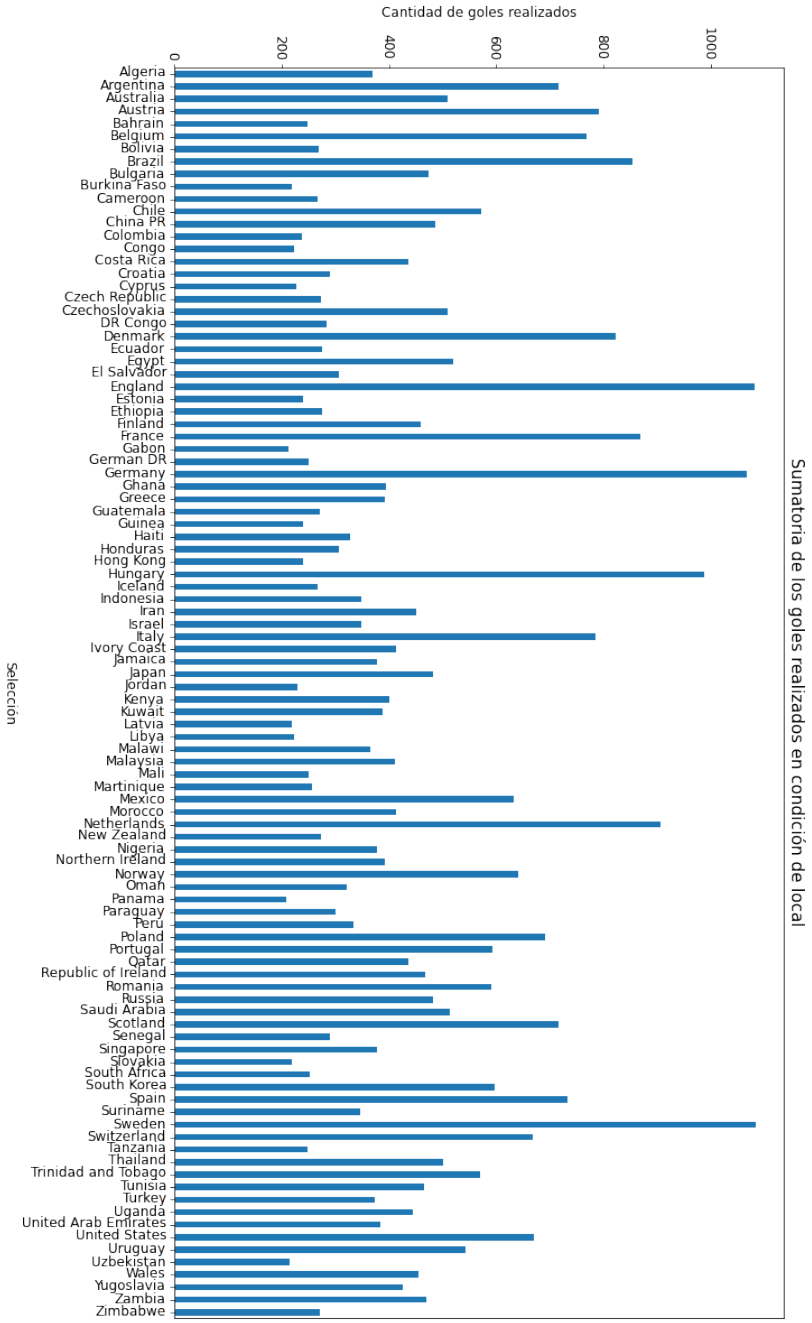

Figura 11: Goles de local de las selecciones

3.3.4. Torneo y estadio neutral

La motivación de este análisis es poder ver si los torneos que suelen disputarse en una sola sede tienen una gran proporción de partidos en estadio neutral. Dado que hay cuatro competencias con una gran cantidad de partidos (los amistosos y las clasificaciones a la copa mundial, a la copa de Europa y a la copa de África) decidimos separar estos casos para poder apreciar mejor las barras. En el gráfico 12 podemos observar que la Copa mundial tiene muchos partidos neutrales, al igual que la Copa africana de las naciones y la Copa América. En contrapartida la mayor parte de los partidos correspondientes al Campeonato nórdico, Campeonato británico, a la clasificación a la Copa del caribe y a la Copa asiática poseen un equipo local y otro visitante.

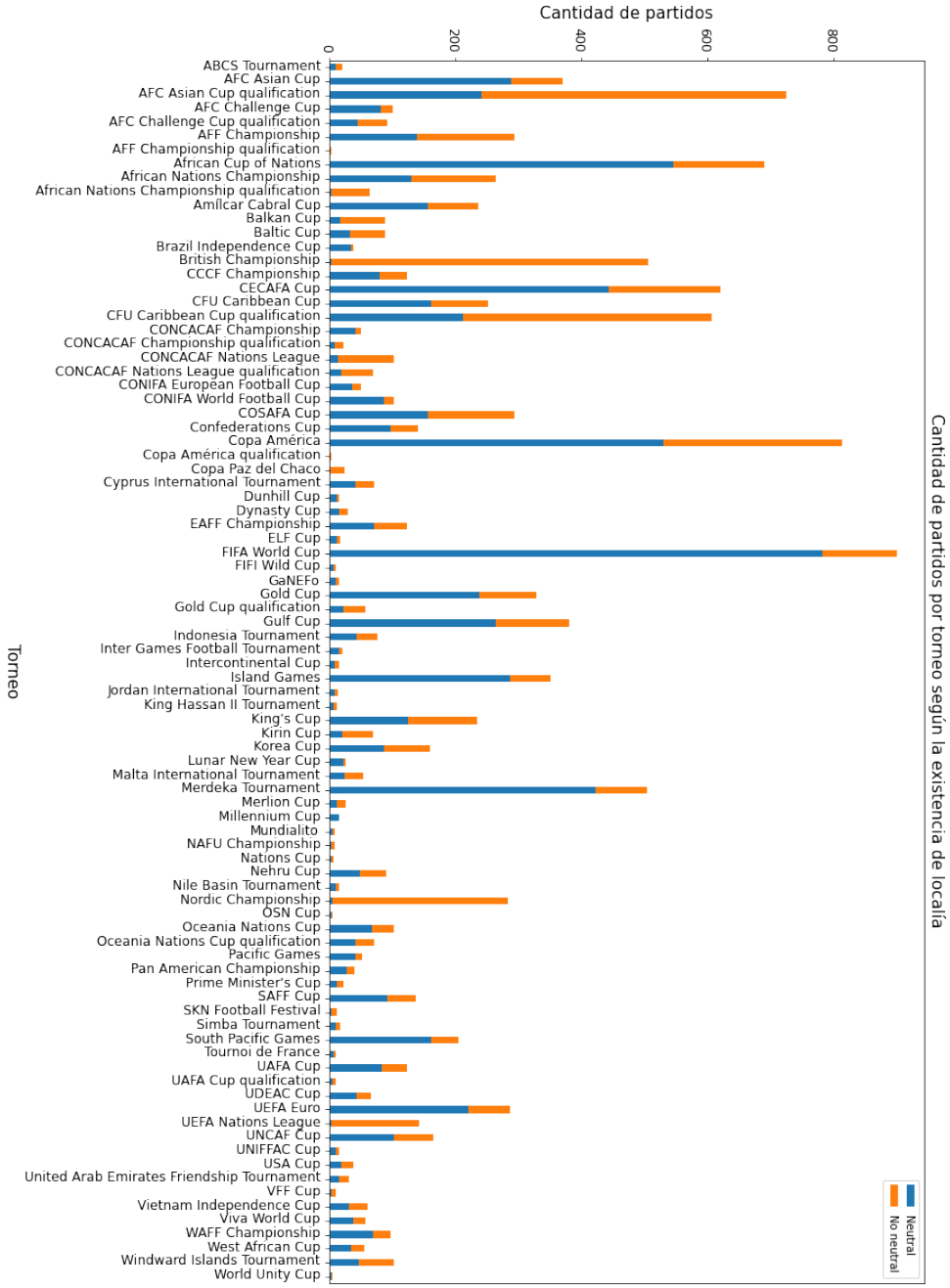

Figura 12: Cantidad de partidos neutrales y no neutrales de todos los otros torneos

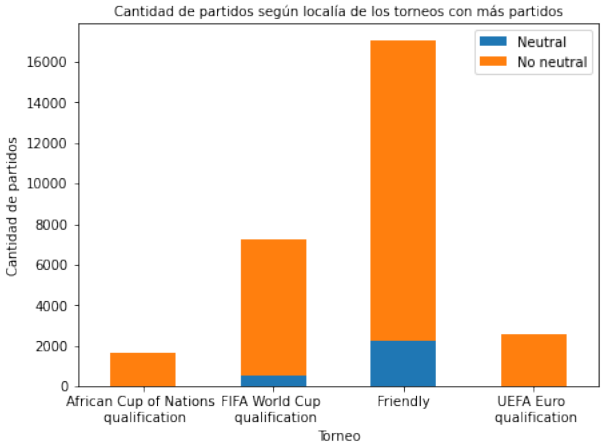

Figura 13: Cantidad de partidos neutrales y no neutrales de los torneos con más partidos

En la figura 13 tenemos a los torneos con más partidos. A diferencia de los otros torneos los partidos no suelen jugarse en un estadio neutral. Esto condice con los reglamentos de los torneos presentados. Con respecto a los amistosos hace relativamente poco tiempo que se empezaron a organizar partidos amistosos entre selecciones reconocidas en lugares donde el fútbol no es popular con el objetivo de promover la actividad. Si bien hay varios torneos que se juegan en una sola sede, abundan los torneos que se juegan en varios lugares. Además los amistosos tienen la mayor parte de los partidos y predominan los partidos con local y visitante.

3.3.5. Goles visitantes y fecha

Pasemos a analizar en conjunto las variables *home_score* y *date*. En la figura 14 podemos observar la suma de todos los goles de los equipos visitantes en cada día. Este gráfico es muy similar a la cantidad de partidos que hubo en cada fecha. Es sensato que hayan habido más goles si se disputaron más partidos.

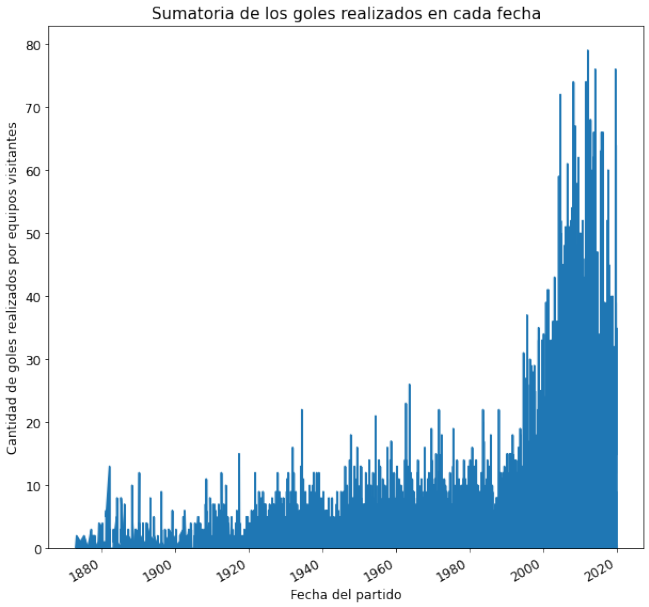

Figura 14: Serie temporal de los goles de visitante

Por esta razón en el gráfico 15 podemos encontrar el promedio de los goles visitantes por partidos según la fecha en la que se jugaron. Ignorando a los outliers podemos ver que hasta 1920 los visitantes no solían hacer más de 2 goles. En el período 1930-1970 encontramos el mejor momento para los equipos visitantes en cuanto a goles realizados.

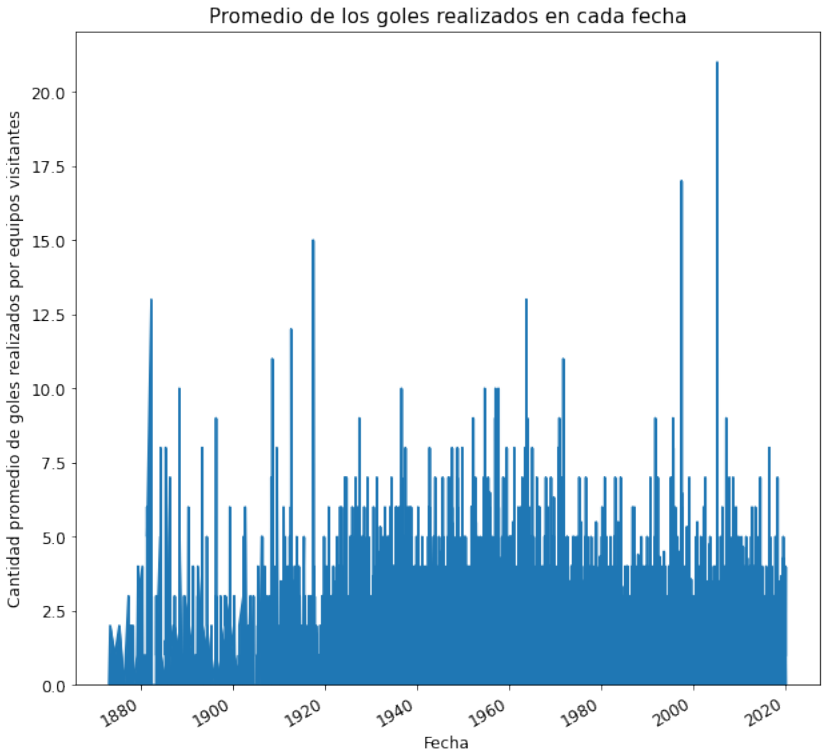

Figura 15: Serie temporal del promedio de los goles de visitante

3.4. Reglas de Negocio y Métricas de la Calidad

Desde un análisis de los datos determinamos las siguientes reglas de negocio para los campos de la tabla con sus correspondientes métricas o rangos válidos:

- Ninguno de los valores de las celdas puede ser nulo
- *date* debe ser una fecha válida entre 1872 y la actualidad
- *home_team* y *away_team* deben ser asociaciones de fútbol válidas
- *tournament* debe ser amistoso o un torneo internacional
- *home_score* y *away_score* deben ser mayores o iguales que cero
- *city* y *country* deben ser válidos para la fecha que se jugó el partido. Además la ciudad debe pertenecer al país en la fecha del partido.
- *neutral* sólo puede tomar los valores TRUE o FALSE

3.5. Mecanismo de Remediación Automática

Un mecanismo de remediación automática es difícil de establecer debido a que no se cuentan con otras tablas desde las cuales deducir partidos faltantes o resultados erróneos que hacen al negocio. Las principales correcciones se pueden hacer en el momento de la carga de los datos, como validaciones numéricas o de equipos o lugares.

Tanto las celdas que contienen valores de fecha, como los que corresponden a los goles hechos por cada bando pueden ser validados en el momento de la carga a través de restricciones. Los valores ya existentes para dichos campos cumplen con las restricciones.

Para los valores de las celdas correspondientes a la ciudad y el país donde se jugó el partido de esa entrada en la tabla se podrían llegar a validar con una fuente externa. Eventualmente habría que conseguir una tabla de referencia donde se tenga la validez del país por rango histórico ya que varios países cambian de nombre a través del tiempo.
Dicha validación puede hacerse al momento de la carga de los datos y para el caso de los ya existentes, pensando en una remediación automática, emitir un alerta cuando no haya coincidencia para ver el caso puntual.

El mismo tratamiento sería para ambos equipos y el Torneo en cuestión.

4. Anexo (código fuente)

Los diferentes algoritmos para el manejo de la base de datos, el análisis de los campos y los gráficos decidimos hacerlos en Python sobre un notebook de Colab tomando la data del sitio Kaggle[2].
Se deja más abajo el link a dicho notebook. Junto con la correspondencia a cada ejercicio y una breve explicación sobre el razonamiento sobre cómo realizamos los algoritmos para poder analizar los diferentes aspectos del trabajo.
https://colab.research.google.com/drive/1D9j8YtehJ6UeX1Guml99sd9nzobTxDIn?usp=sharing

[2] www.kaggle.com