**Escuela de Estudios de Postgrado**

**Facultad de Ingeniería**

**Universidad San Carlos de Guatemala**

**Maestría en Ingeniería para la Industria con**

**Especialización en Ciencias de la Computación.**

**Introduction Data Analytics**

**Ing. Ms. Marco Tulio Aldana Prillwitz**

**PROYECTO INICIAL**

**https://github.com/obduliogonzalez/proyecto\_inicial**

**Néstor Obdulio González López**

**Carné: 999013150**

**Introducción**

El objetivo de este proyecto es llevar a cabo un análisis exhaustivo de un conjunto de datos relacionado con el videojuego FIFA 2020. La finalidad es obtener insights significativos que informen decisiones estratégicas y ayuden a comprender patrones en el desempeño de jugadores y equipos. Para ello, se ha utilizado Python y PostgreSQL como herramientas principales para el análisis y manipulación de los datos.

**Discovery**

Para estructurar el proyecto, se utilizó el Canvas Data Product para definir el enfoque y las preguntas a resolver:

**Problema:** ¿Qué factores influyen en el desempeño de los jugadores y equipos en FIFA 2020?

**Hipótesis:** Los equipos con jugadores mejor calificados (overall rating) tienden a tener un mejor rendimiento en sus respectivas ligas.

**Solución esperada:** Identificar las características clave de los jugadores que contribuyen al éxito de los equipos.

**Datos utilizados:**

Players(15,16,17,18,19,20).csv: información detallada sobre jugadores (edad, calificación, potencial, valor, etc.).

teams\_and\_leagues.csv: información sobre equipos y ligas.

Métricas clave (KPI): Calificación general (overall), precio de transferencia (value\_eur), salarios (wage\_eur), y rendimiento por equipo (teams\_and\_leagues).

**Data Analysis (Análisis Exploratorio de Datos - EDA)**

Para el análisis inicial, se utilizó Python (Pandas, Numpy, Matplotlib y Seaborn). Aquí se presentan algunos de los resultados:

**Distribución de Calificación General (Overall Rating)**

plt.figure(figsize=(10, 6))

sns.histplot(df['overall'], bins=20, kde=True, color='blue')

plt.title(f'Distribución de la Calificación General de los Jugadores (Año: {year})')

plt.xlabel('Calificación General')

plt.ylabel('Frecuencia')

plt.grid(True)

plt.show()

La mayoría de los jugadores tienen una calificación general entre **70 y 80**, con algunos pocos superando el **90**.

**Relación entre Edad y Calificación General**

plt.figure(figsize=(12, 8))

sns.scatterplot(x='age', y='overall', hue='club', data=merged\_df)

plt.title(f'Relación entre Edad y Calificación General ({year})')

plt.xlabel('Edad')

plt.ylabel('Calificación General')

plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.grid(True)

plt.show()

Se observa que jugadores jóvenes tienden a tener un alto potencial (potential), aunque los jugadores con experiencia tienen calificaciones generales más estables.

**Data Cleaning (Limpieza de Datos)**

El proceso de limpieza de datos incluyó:

Eliminación de duplicados: Se eliminaron registros duplicados para evitar distorsiones en el análisis.

Tratamiento de valores nulos: Se imputaron valores faltantes en las columnas críticas como rating, value\_eur y wage\_eur.

Estándar de formatos: Se unificaron los formatos de fechas y nombres de equipos para facilitar la unión de datasets.

**Data Wrangling (Manipulación de Datos)**

Se realizó la combinación de los datasets players\_.csv y teams\_and\_leagues.csv utilizando la columna club y league\_name como clave:

merged\_df = pd.merge(players\_df, teams\_df, left\_on='club', right\_on='league\_name', how='inner')

Además, se crearon nuevas variables para el análisis, como la diferencia entre potential y overall, y el cálculo del rendimiento basado en wage\_eur y value\_eur.

**Data Transformation (Transformación de Datos)**

Para preparar los datos para un análisis más avanzado, se aplicaron técnicas de normalización y escalado utilizando Scikit-learn:

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

merged\_df['scaled\_overall'] = scaler.fit\_transform(merged\_df[['overall']])

merged\_df['scaled\_value'] = scaler.fit\_transform(merged\_df[['value\_eur']])

**Hallazgos y Conclusiones**

* El análisis mostró que la calificación general (overall) y el potencial (potential) son indicadores clave del rendimiento de los jugadores.
* Los equipos con mayores salarios (wage\_eur) tienden a tener jugadores con calificaciones más altas.
* Existen ligas (por ejemplo, la Premier League y La Liga) que concentran a los jugadores mejor calificados.

**Recomendaciones**

* Los equipos podrían optimizar su presupuesto enfocándose en jugadores jóvenes con alto potential.
* Analizar la relación entre el costo de transferencia (value\_eur) y el salario (wage\_eur) podría ayudar a los clubes a hacer contrataciones más eficientes.

GRAFICAS:

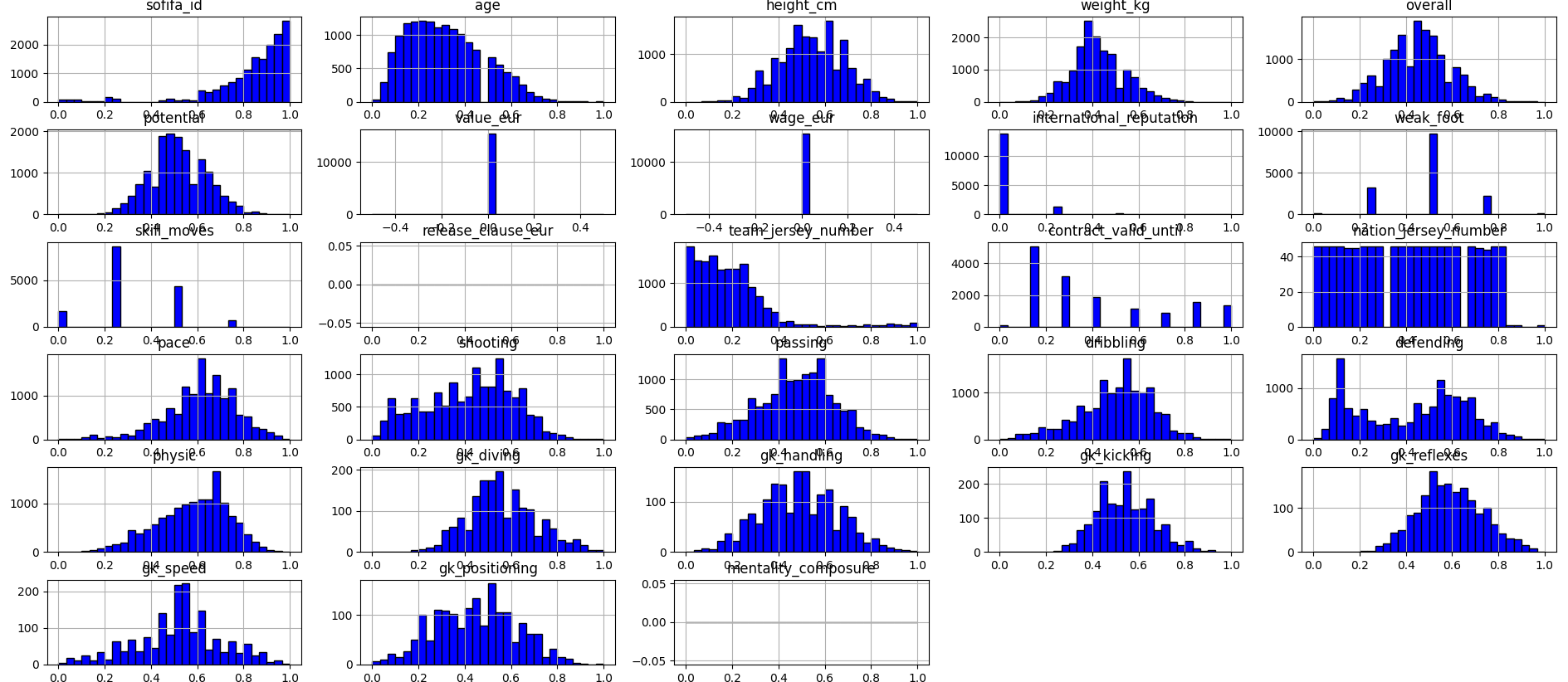


Ilustración 1 PLANTILLA 2015



Ilustración 2 PLANTILLA 2016

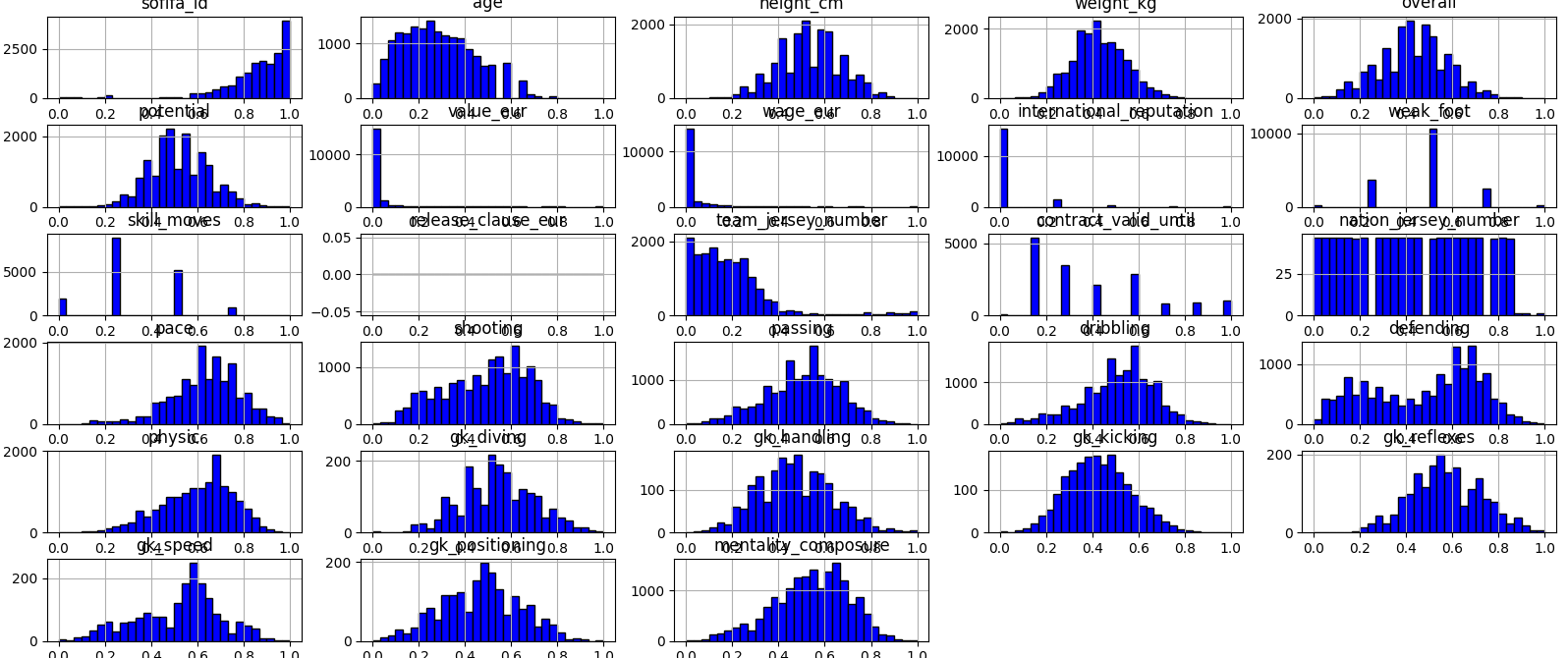


Ilustración 3 PLANTILLA 2017

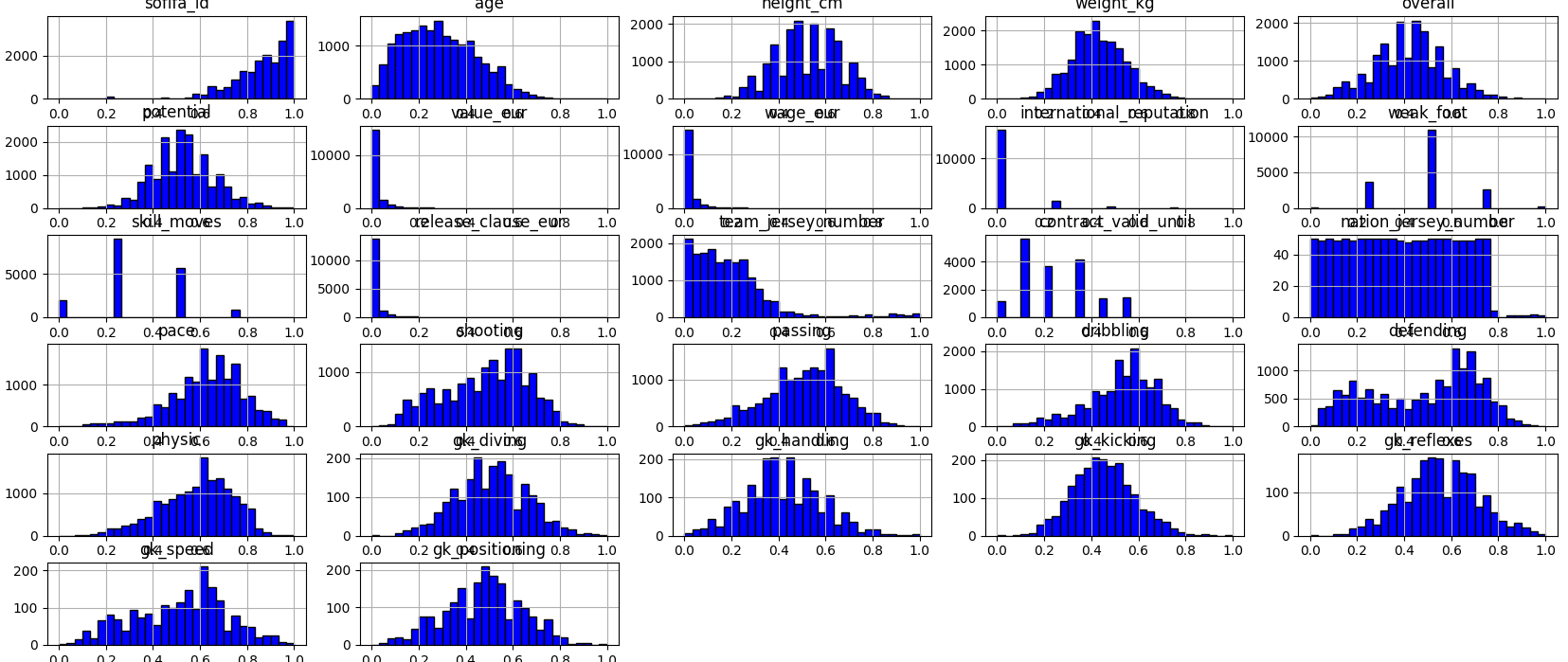


Ilustración 4 PLANTILLA 2018

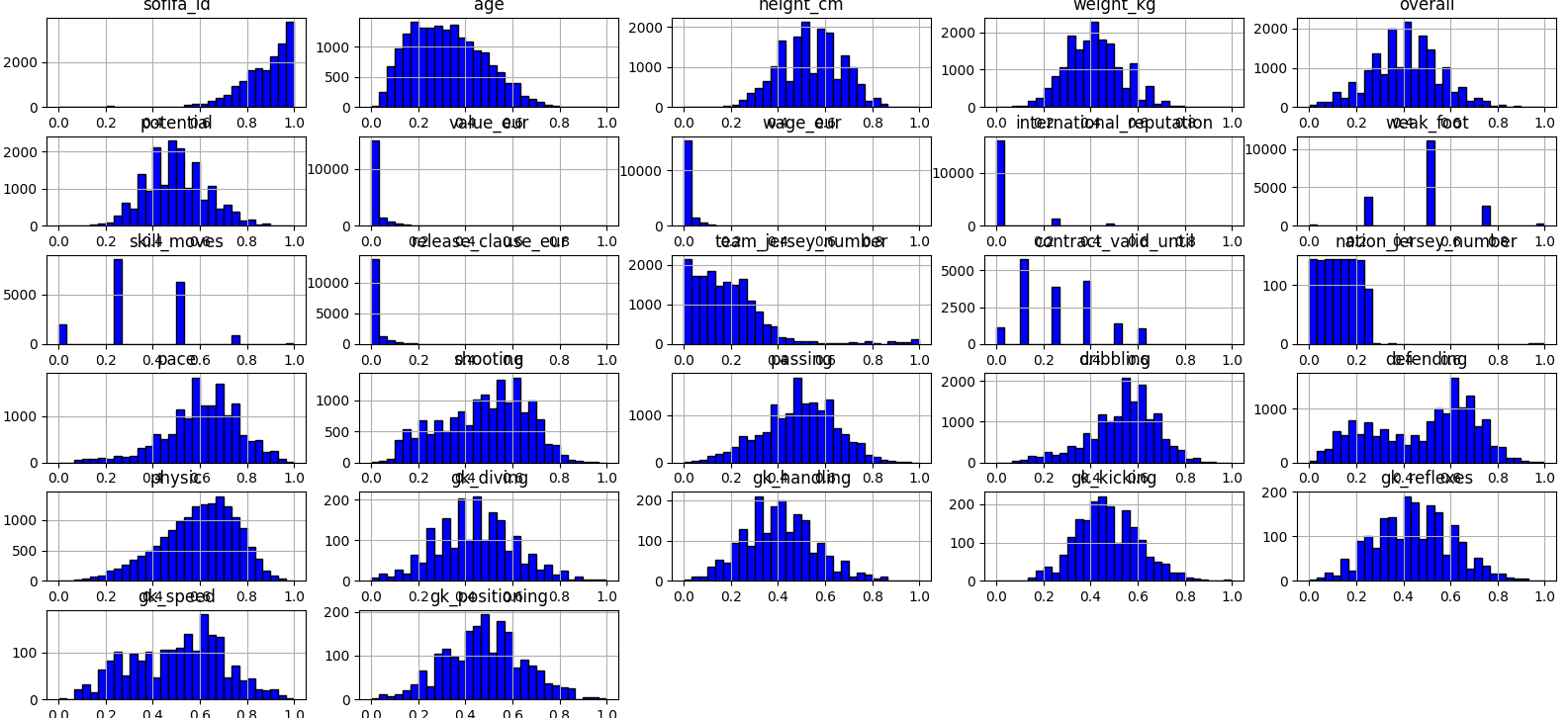


Ilustración 5 PLANTILLA 2019

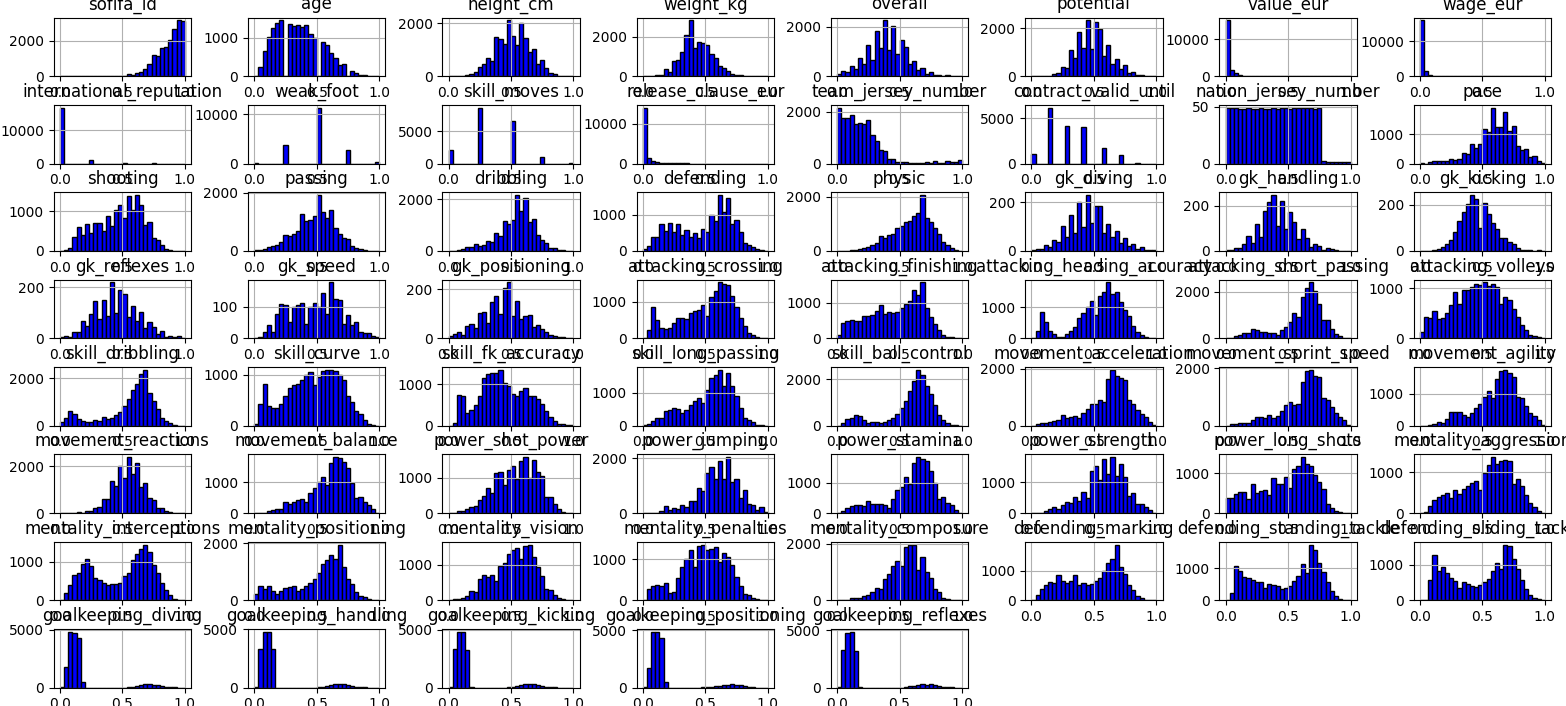


Ilustración 6 PLANTILLA 2020

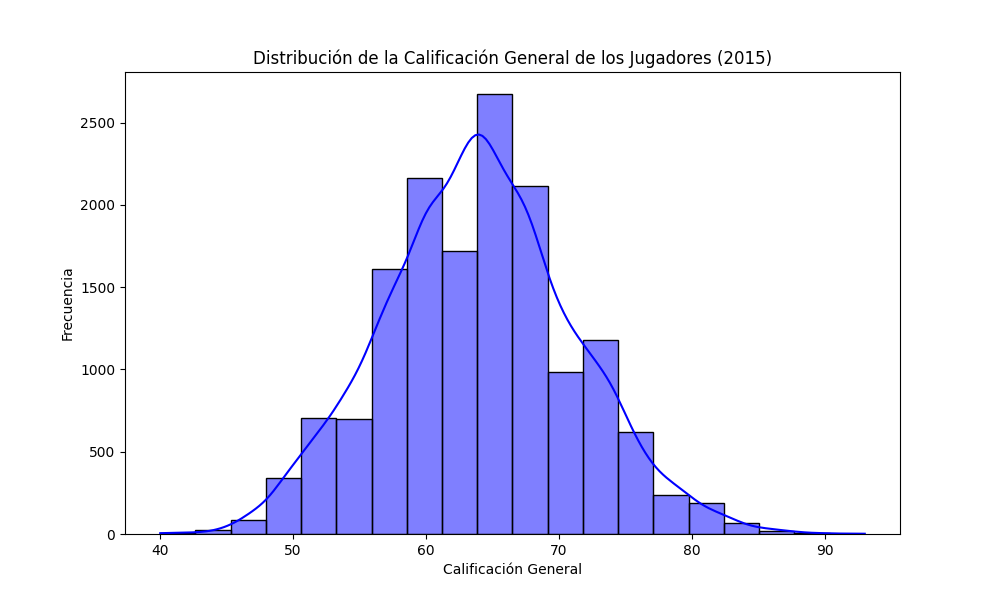


Ilustración 7 DISTRIBUCION GENERAL 2015

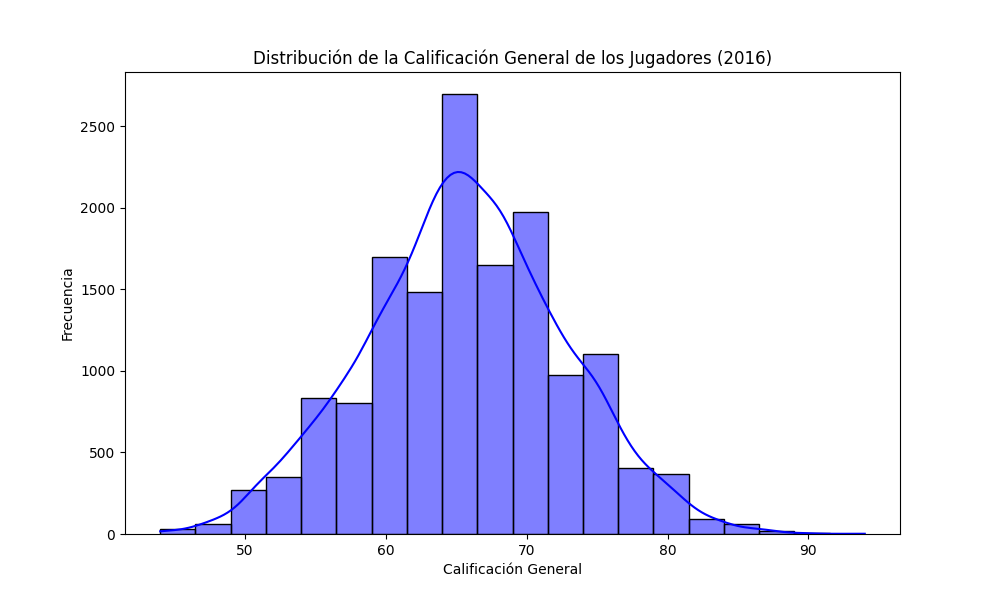


Ilustración 8 DISTRIBUCION GENERAL 2016

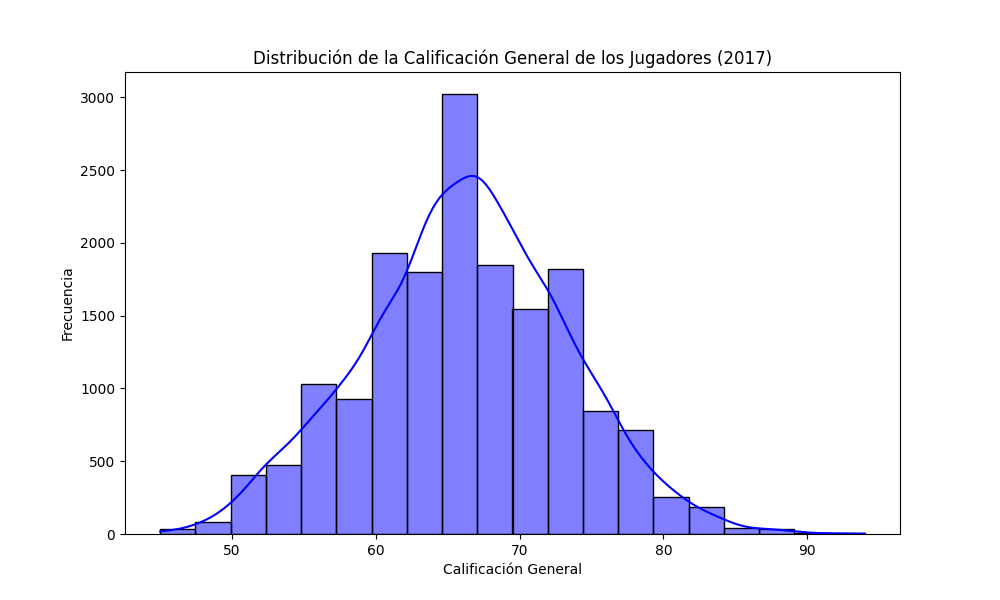


Ilustración 9 DISTRIBUCION GENERAL 2017

