A low-angle, close-up shot of several runners' legs and feet as they jog on a paved surface. The runners are wearing various colored athletic shoes, including orange, yellow, and blue. The background is blurred, showing more runners and spectators.

Predicción y prescripción de tiempos de maratón

(en base a datos de
entrenamiento)

INTRODUCCIÓN AL PROBLEMA

- Las maratones se caracterizan por ser carreras de fondo, es decir, larga distancia.
- En las carreras de velocidad, se utiliza la máxima capacidad física para agotarla en distancias de hasta 400mts y menos de 1 minuto.
- Pero en una maratón, se debe **dosificar el desgaste** que el cuerpo debe soportar durante decenas de kilómetros y horas de recorrido.



INTRODUCCIÓN AL PROBLEMA

- En tal sentido, se busca indagar en un **modelo predictivo –y prescriptivo–** que determine el tiempo de carrera de un corredor, a partir de sus parciales, rango de edad y género.
- Esto con el fin de determinar el ritmo de carrera en tiempo real, planificar entrenamientos y/o planificar las carreras en sí, para lograr el rendimiento adecuado a cada runner en cada etapa de su desarrollo.





OBJETIVO DEL PROYECTO

- En base al análisis de miles de registros de corredores totalmente heterogéneos, **generar un modelo predictivo, que determine el tiempo de carrera de un corredor a partir de sus parciales, rango de edad y género.**
- Dar al usuario objetivo, una herramienta que le permita controlar su ritmo de carrera en tiempo real, planificar entrenamientos y/o planificar sus carreras en sí.
- Además, del análisis exploratorio descriptivo, se puede derivar para futuros proyectos, la segmentación de clientes potenciales, para la oferta de productos y servicios conexos al running en el nivel amateur y profesional, según género y rango de edad y, de acuerdo a su locación geográfica.



USUARIO FINAL

- **Maratonistas** de toda categoría de rendimiento, rango de edad y sexo.
- **Entes comerciales** relacionados a prestación de servicios y venta de artículos relacionados al running.



BASES TEÓRICAS

DESDE LO MENTAL Y FISIOLÓGICO

- Muchos corredores suelen sentir el deseo apremiante de **correr demasiado lejos o rápido desde el primer momento**. Esto, puede provocar mucho dolor físico, agotamiento o, incluso, lesiones.
- Además, los corredores que se desgastan demasiado y muy pronto, se pueden sentir **derrotados** e incapaces de incluir la disciplina del running en su rutina.



BASES TEÓRICAS

EN CONSECUENCIA

- Una de las mejores cosas que se puede hacer es **tomar el trote con calma**, especialmente al comenzar. Incluso los deportistas experimentados hacen hincapié en la importancia de tomarse con calma algunos días.
- Si se empieza a un **ritmo más lento** se **desarrolla**, además de la **resistencia**, la **paciencia** y la **disciplina**. Al mismo tiempo, las **articulaciones**, los **tendones** y los **huesos** tendrán la posibilidad de **adaptarse mejor**.
- Además, **correr a intensidades más bajas fortalece el sistema de energía aeróbica**, es decir, el uso de carbohidratos y grasas para obtener energía cuando el oxígeno está presente, lo cual es fundamental para las disciplinas de fondo.

The background image shows a group of runners in silhouette against a bright, hazy sky at sunset or sunrise. In the foreground, a close-up of a runner's leg and foot is visible, wearing a dark running shoe with a white swoosh logo and orange accents on the sole. The overall mood is energetic and focused.

BASE DE DATOS

DATA ACQUISITION - INDICACIÓN DE LA FUENTE DEL DATASET


- El Dataset se extrajo de la plataforma de Kaggle. Enlace:
- <https://www.kaggle.com/datasets/rojour/boston-results>
- Es una matriz de 26.598 registros (+1 fila de encabezados) y 25 campos.

The background image shows a group of runners in silhouette against a warm, orange-hued sky at sunset or sunrise. In the foreground, a large, detailed leg and foot in a running shoe are visible, suggesting a first-person perspective of a runner. Other runners are seen in the background, some in motion and others more static, creating a sense of a race or marathon event.

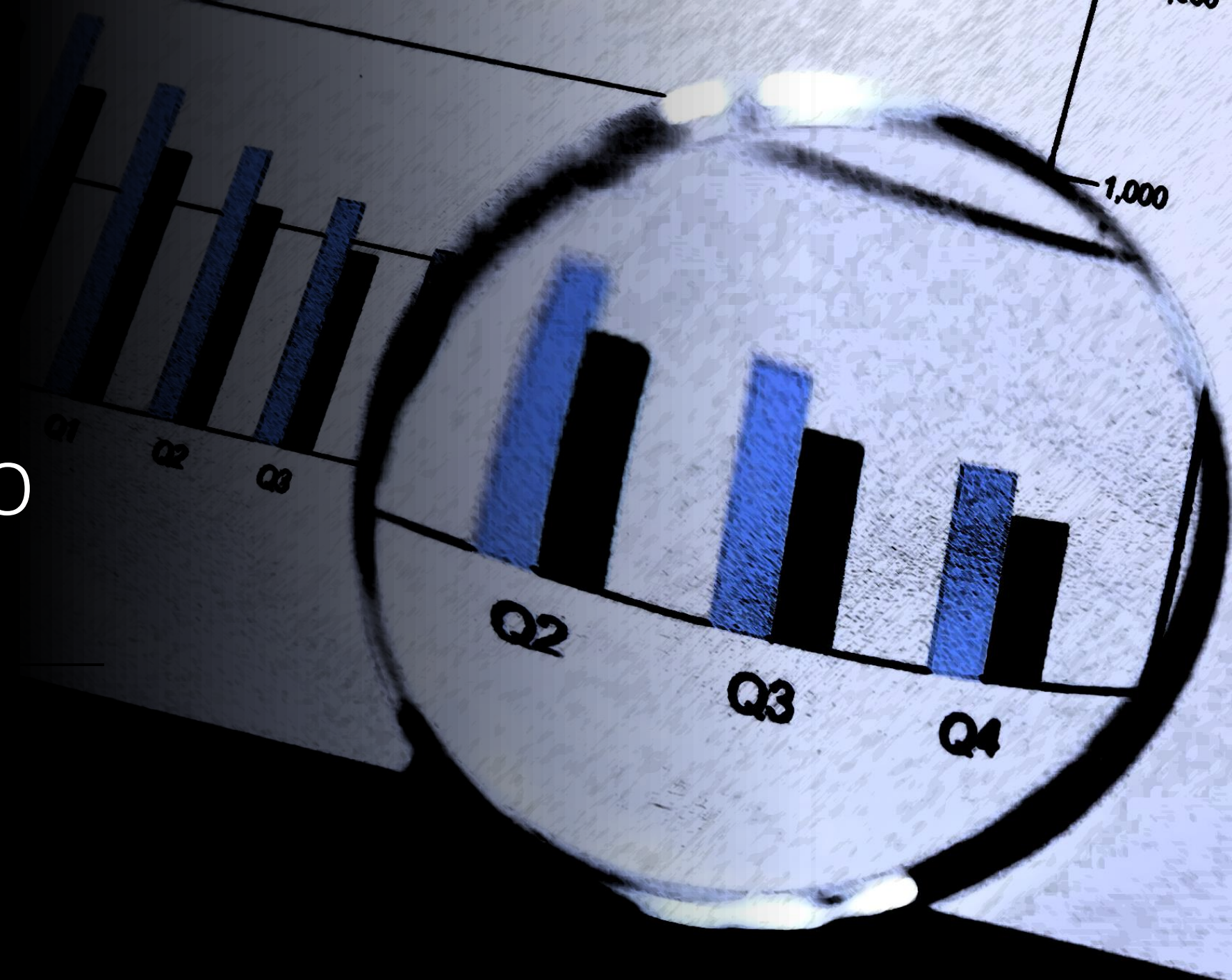
BASE DE DATOS

DATA ACQUISITION - CRITERIOS DE SELECCIÓN

- Su finalidad es la predicción de tiempos de maratón a partir de tiempos parciales de carrera, como así también un análisis descriptivo de la interacción de sus variables.
- La selección se perfecciona en tanto la base de datos, cuenta con una serie de **variables categóricas** que **inciden biológicamente en el desempeño físico** y, **cuantitativas**, que **reflejan la variación en la performance** a lo largo del recorrido.
- Además, estas últimas variables permiten generar nuevos campos a partir del tratamiento aritmético.
- Asimismo, se eligió este Dataset porque presenta un tamaño de cómodo manejo para el hosting web (4MB) como así también para el procesamiento de Google Colaboratory.



ANÁLISIS EXPLORATORIO DE DATOS



```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26598 entries, 0 to 26597
Data columns (total 25 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Final position        26598 non-null  int64
1   Bib                   26598 non-null  object
2   Name                  26598 non-null  object
3   Age                   26598 non-null  int64
4   M/F                   26598 non-null  object
5   City                  26598 non-null  object
6   State                 24047 non-null  object
7   Country               26598 non-null  object
8   Citizen               1064 non-null   object
9   _1                    67 non-null     object
10  5K                     26446 non-null  float64
11  10K                    26567 non-null  float64
12  15K                    26580 non-null  float64
13  20K                    26569 non-null  float64
14  Half                   26570 non-null  float64
15  25K                    26567 non-null  float64
16  30K                    26559 non-null  float64
17  35K                    26547 non-null  float64
18  40K                    26542 non-null  float64
19  Pace                   26598 non-null  int64
20  Proj Time              26598 non-null  object
21  Official Time          26598 non-null  int64
22  Overall                26598 non-null  int64
23  Gender                 26598 non-null  int64
24  Division               26598 non-null  int64
dtypes: float64(9), int64(7), object(9)
memory usage: 5.1+ MB

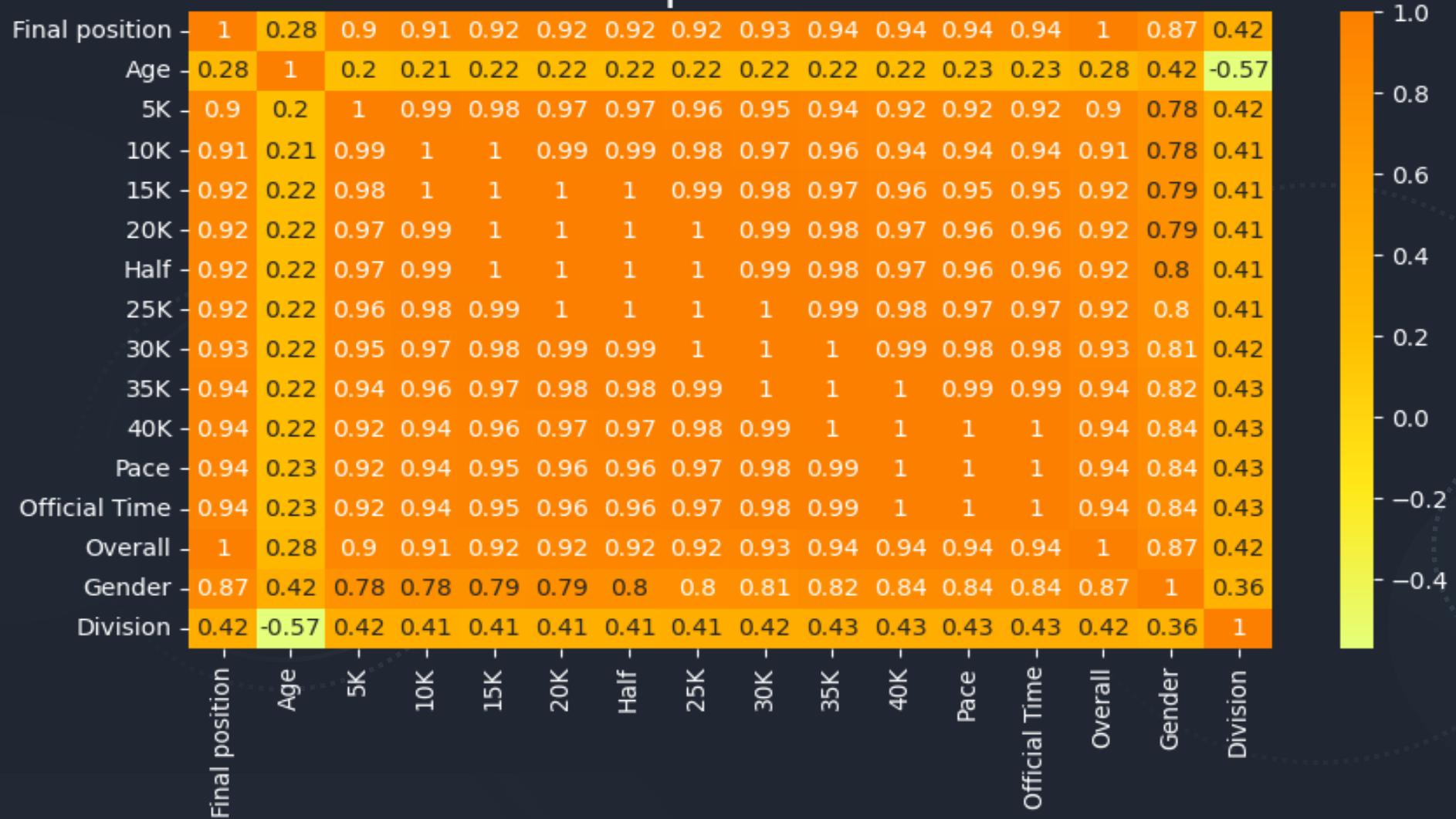
```

Tipificación de campos

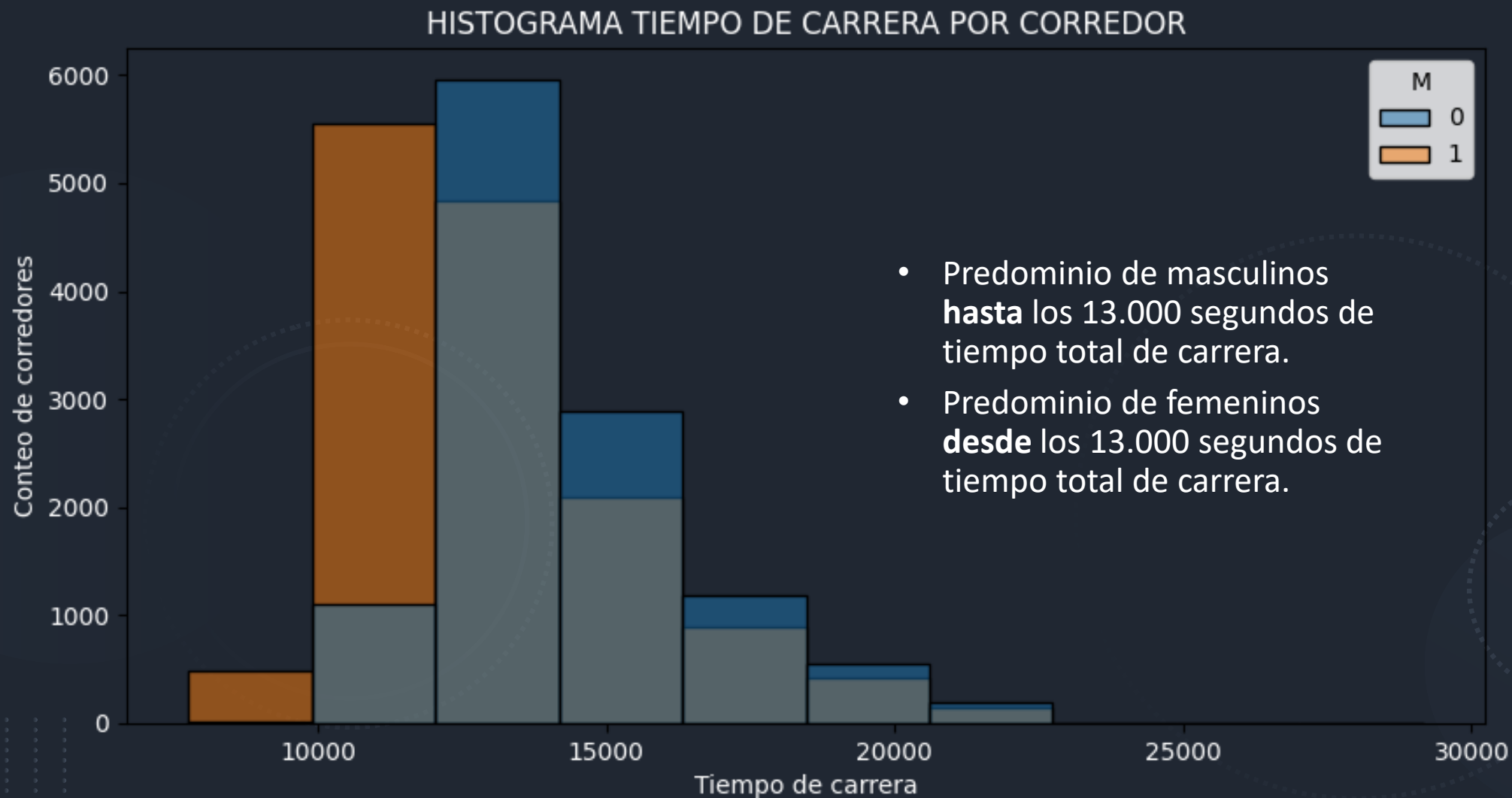
- Índice y nombre de columnas.
- Conteo de valores no nulos.
- Tipos de variables: cualitativas y cuantitativas.

Correlación entre variables

Heatmap del Dataset

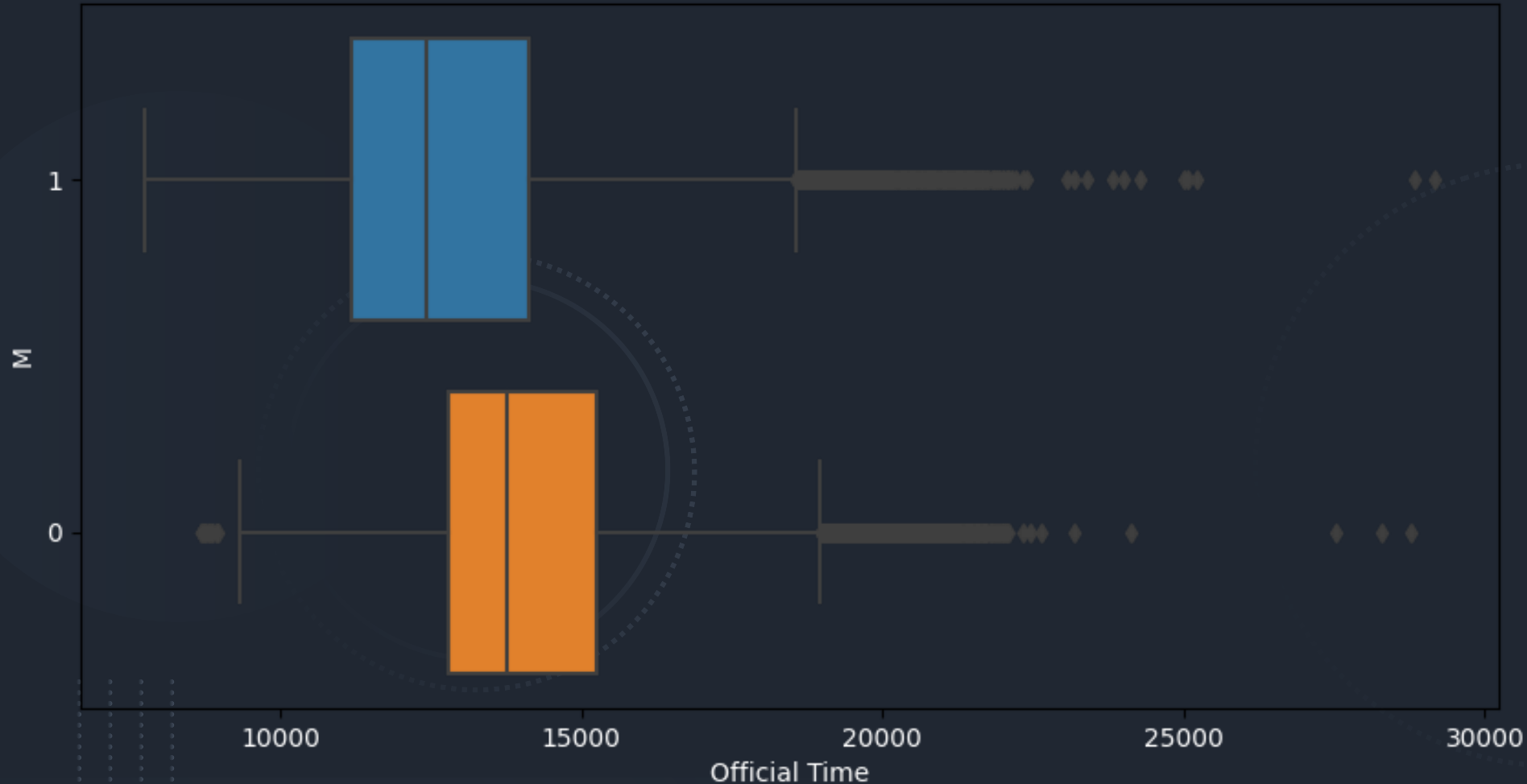


Corredores por rango de rendimiento



M = 1 masculino, M = 0 femenino

Outliers por tiempo de carrera

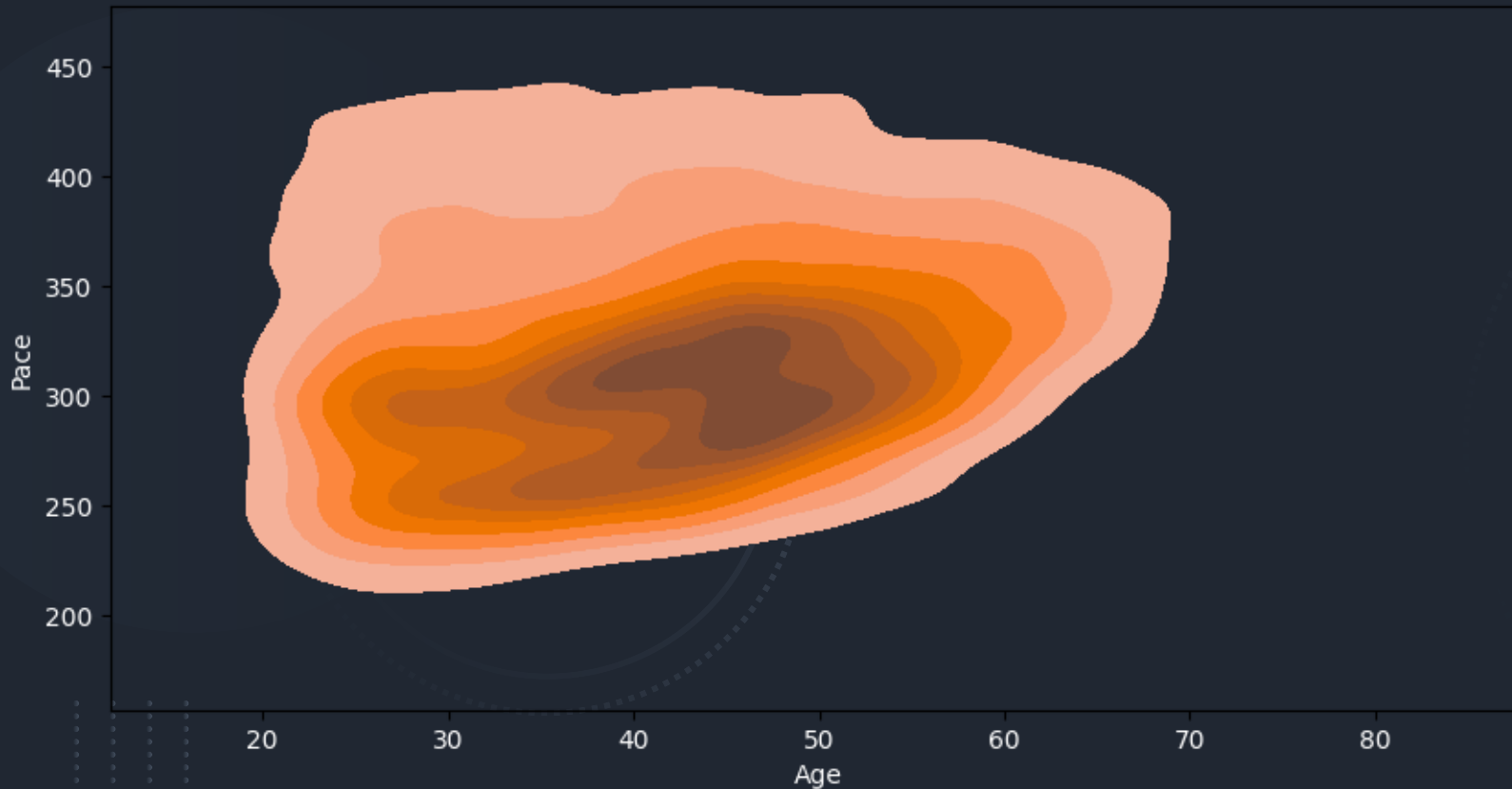


M = 1 masculino
M = 0 femenino

- Outliers superiores en sujetos de **bajo rendimiento, lesionados** o con **extenuación**.
- Outliers inferiores femeninos, mayor **brecha** en **corredoras profesionales**.

Concentración de corredores

(POR EDAD Y PASO PROMEDIO)

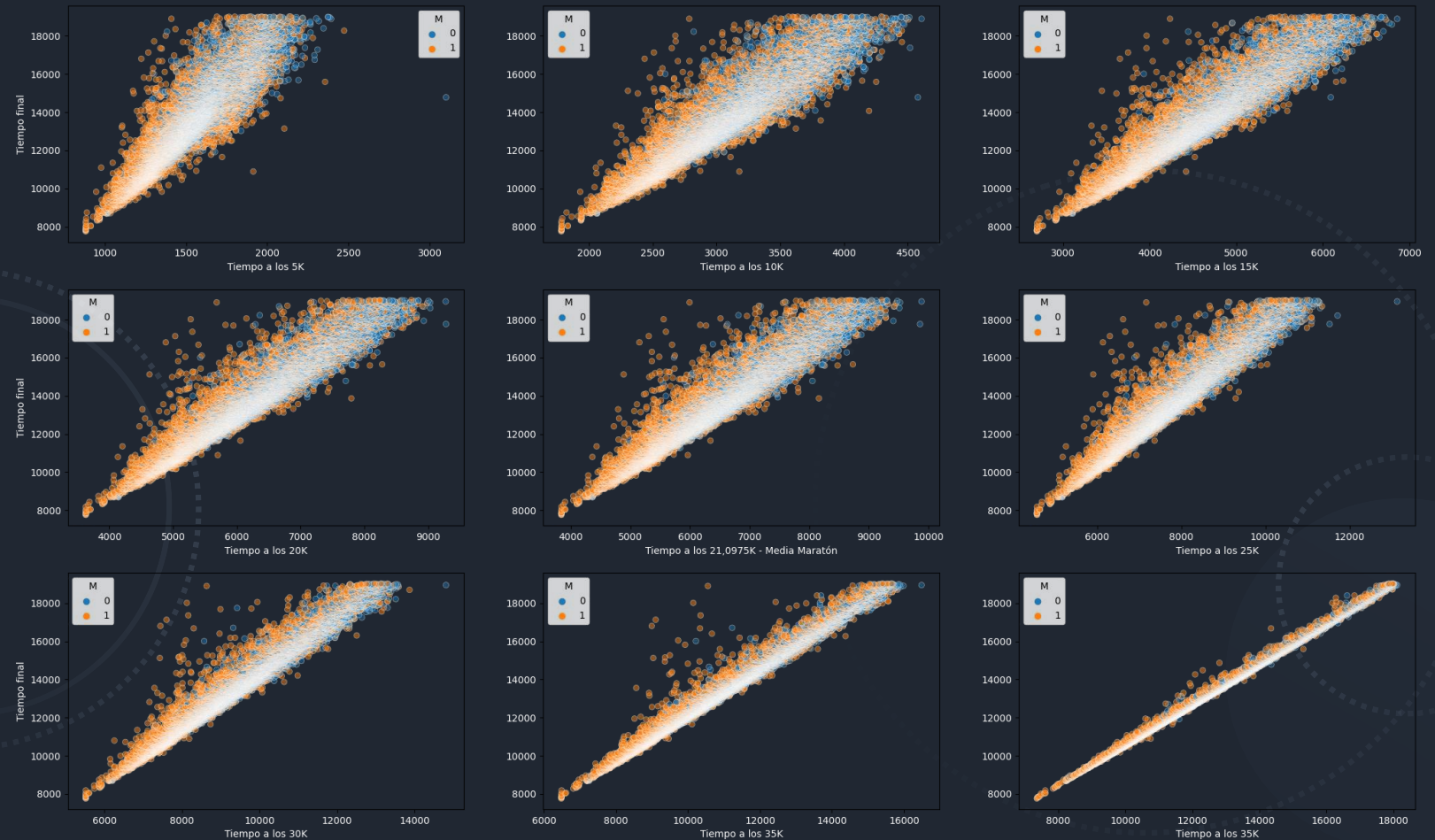


MÁS OSCURO, MÁS CORREDORES

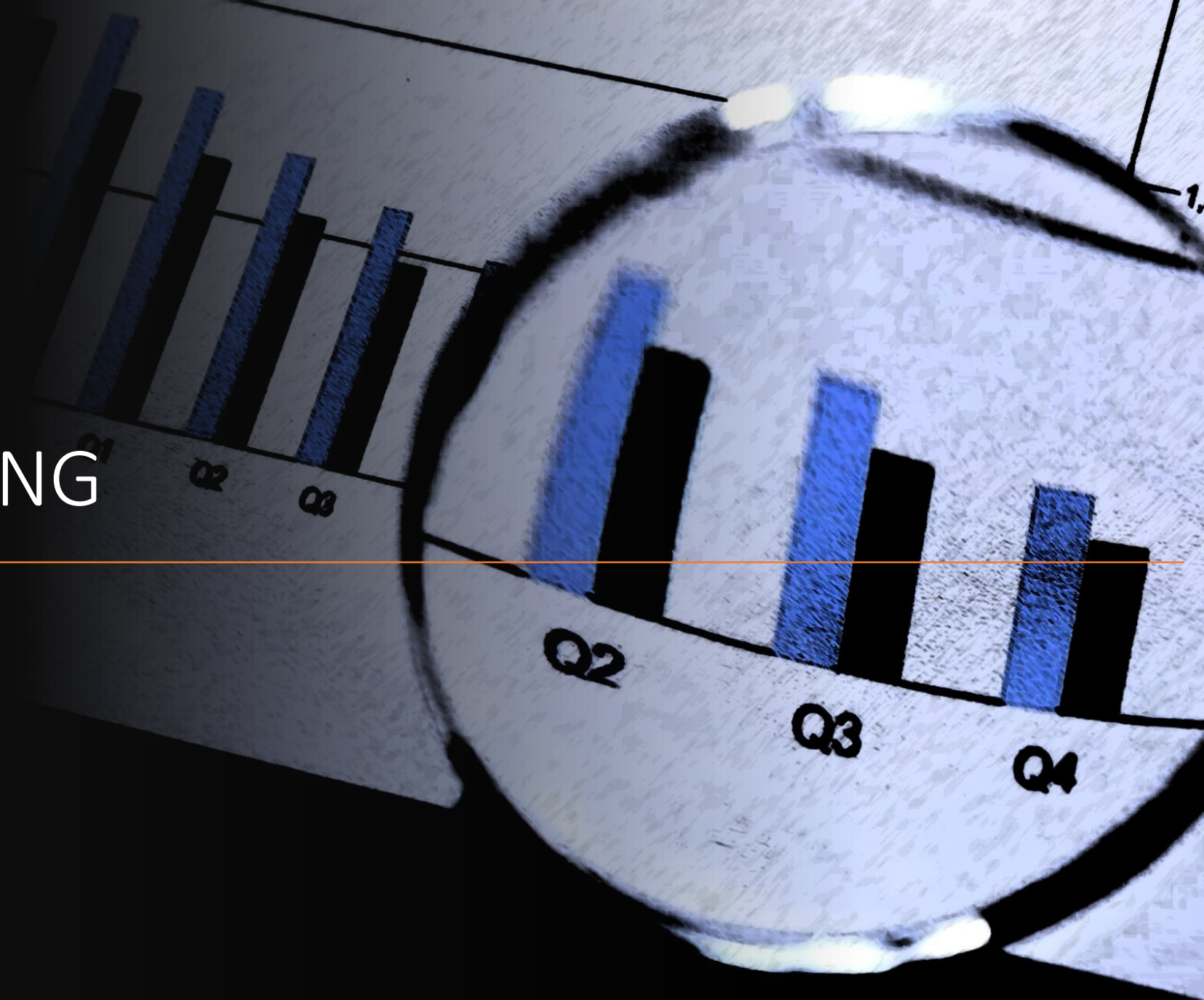
- **Mayor concentración** entre los 40 y 50 años y;
- Entre los 275 y 325 segundos por kilómetro.
- El paso de carrera mínimo **mejora** de los 18 a los 25 años;
- Luego, esos ritmos se vuelven **más holgados** a medida que aumenta la edad.

Relación entre parciales y marca final de carrera

- **EJE X:** tiempo a los 5, 10, 15 y 20km, media maratón, 25, 30, 35 y 40km (Izquierda a derecha y arriba hacia abajo).
- **EJE Y:** Tiempo final de carrera.
- A **mayor distancia** de carrera, más se **estabiliza el ritmo** de trote, **respecto al tiempo final**, lo que se refleja en el **acotamiento de la dispersión**.



DATA WRANGLING



EXPLORACIÓN

- VERIFICACIÓN DE NULOS.
- Verificación de valores no válidos.
- Verificación de registros duplicados.

```
df.isnull().sum()
```

Final position	0	20K	29
Bib	0	Half	28
Name	0	25K	31
Age	0	30K	39
M/F	0	35K	51
City	0	40K	56
State	2551	Pace	0
Country	0	Proj Time	0
Citizen	25534	Official Time	0
_1	26531	Overall	0
5K	152	Gender	0
10K	31	Division	0
15K	18	dtype: int64	

SOBRE UN TOTAL DE 26.598 REGISTROS

EXPLORACIÓN

- Verificación de nulos.
- VERIFICACIÓN DE VALORES NO VÁLIDOS.
- Verificación de registros duplicados.

```
df['Proj Time'].unique()
```

```
array(['-'], dtype=object)
```

El campo “Proj Time” solo contiene el carácter “-” en todos los registros, en lugar de un valor cuantitativo.

SOBRE UN TOTAL DE 26.598 REGISTROS

EXPLORACIÓN

- Verificación de nulos.
- Verificación de valores no válidos.
- VERIFICACIÓN DE REGISTROS DUPLICADOS.

```
df['Name'].nunique()
```

26540

El campo “Name” contiene
58 valores duplicados
(26.598 – 26.540)

SOBRE UN TOTAL DE 26.598 REGISTROS

TRANSFORMACIONES

- Dropeo de filas para las variables de tiempo con nulos (por su incidencia no significativa).
- Dropeo de columnas totalmente o casi vacías.
- Dropeo de columnas que duplican información y con valores inválidos.
- Relleno de valores nulos para el campo "State" (utilidad futura en análisis espacial).

```
df.dropna()
```

```
df.fillna()
```

TRANSFORMACIONES

- Comprobación de que los nombres repetidos son diferentes personas.
- Transformación de la variable categórica 'M/F' en booleana mediante dummies.
- Creación de nuevos campos como relaciones de los ya existentes (para futuros análisis).

```
duplicados.groupby()[].nunique().count()
```

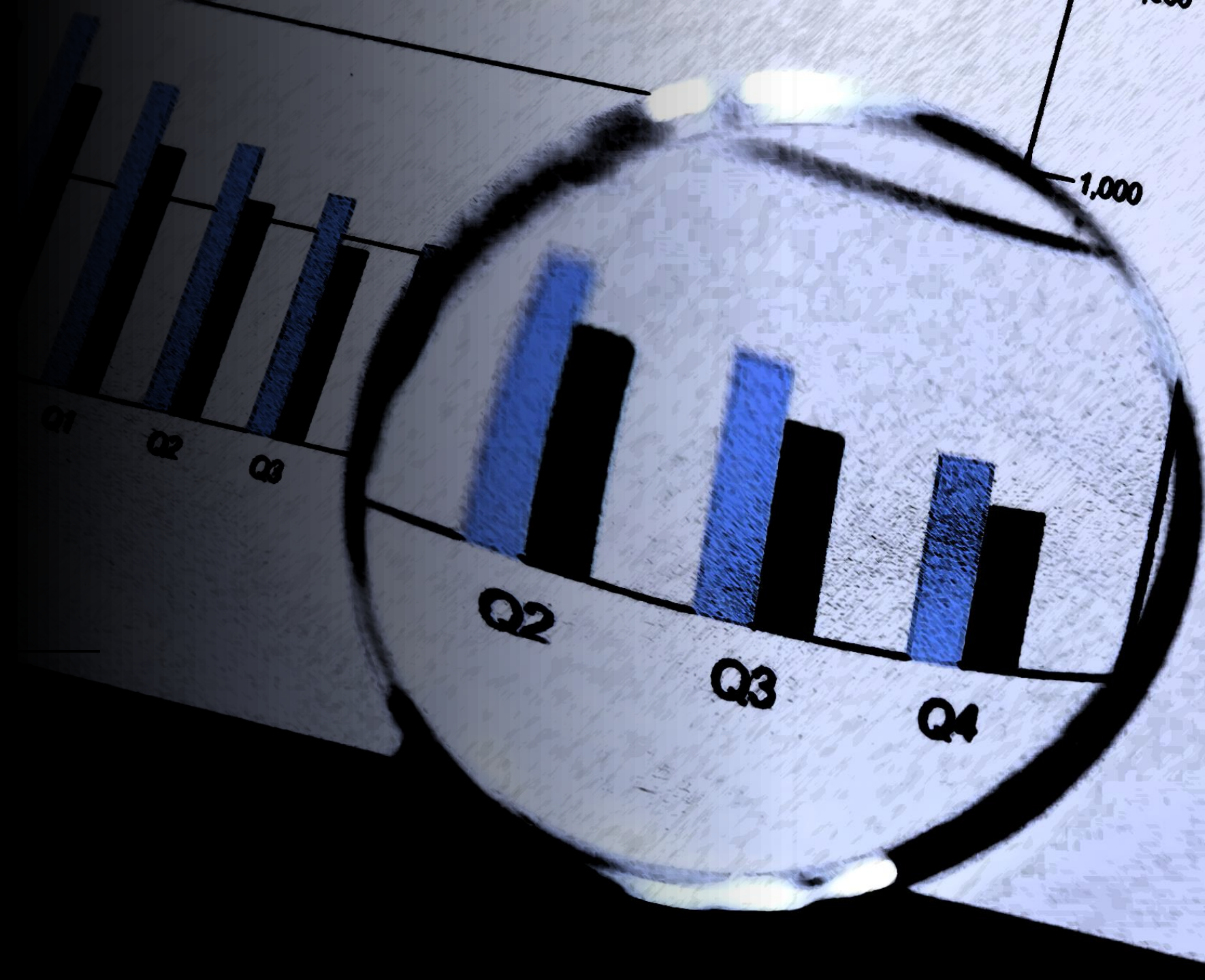
```
pd.get_dummies()
```

```
df[] = df[] - df[]
```

```
df[] = df[] / num
```



DESARROLLO DEL MODELO



ENTENDIMIENTO DE VARIABLES

VARIABLES INDEPENDIENTES

- Surgen del análisis exploratorio, como explicativas, en menor o mayor medida, del comportamiento del rendimiento de los corredores y, en particular, del comportamiento de la variable objetivo.

VARIABLE DEPENDIENTE

- Determina una predicción a la vez que una prescripción del tiempo a lograr.





EVALUANDO MODELOS DE MACHINE LEARNING

REGRESIÓN

- Se ha seleccionado este tipo de modelo, en tanto los algoritmos de regresión en vez de predecir categorías, como lo hacen los algoritmos de clasificación, predicen valores numéricos.
- Es decir, la variable target en un problema de regresión es de tipo cuantitativa, y se busca predecir su valor, como el caso que nos atañe.

PARÁMETROS GENERALES

VARIABLES EXPLICATIVAS

- Tiempos en puntos de control (5k, 10k, 15k, 20k, Media Maratón, 25k, 30k, 35k, 40k).
- Sexo.
- Rango de edad.

VARIABLE OBJETIVO

- Tiempo final de maratón.
-

PARÁMETROS PARTICULARES

Se realizaron diferentes combinaciones de categorías, agrupándose distintos conjuntos de parciales de los 5 a los 30K.

Esto en tanto los parciales posteriores, amén de aumentar la correlación del modelo, no son útiles en la práctica, debido a su proximidad de distancia con la variable objetivo.

A modo de ejemplo:

- Sexo: masculino
- Rango de edad: 40 – 44
- Parciales: 5K, 10K, 15K, Half Marathon, 25K, 30K

PRUEBA DEL EJEMPLO

```
# Dataframe de entrenamiento y testeo
df_M_40_44_train=df.loc[(df['M'] == 1) & (df['Age'].between(40, 44, inclusive=True))]
-----
df_M_40_44_test=df3.loc[(df3['M'] == 1) & (df3['Age'].between(40, 44, inclusive=True))]
```

```
# Nuevo set de entrenamiento
X_train_k5 = df_M_40_44_train[['5K','10K','15K','Half','25K','30K']]
# Variable objetivo de set de entrenamiento
y=df_M_40_44_train[['Official Time']]
# Variable objetivo test
y_test=df_M_40_44_test[['Official Time']]
```

```
# Modelo
modelo_final=LR()
modelo_final.fit(X_train_k5,y)
y_pred_test_k5=modelo_final.predict(df_M_40_44_test[['5K','10K','15K','Half','25K','30K']])
```

```
# Métrica chequeo rendimiento
r2_score(y_pred_test_k5,y_test)
0.862716527193796
```

- ¡EL COEFICIENTE DE DETERMINACIÓN CONSEGUIDO ES SATISFACTORIO!



PREDICCIÓN FINAL

La segmentación de registros por grupos categóricos como el sexo y el rango de edad, en conjunto con la selección del rango limitado de parciales hasta los 30kms –inclusive-, ha dado como resultado la obtención de un coeficiente de determinación satisfactorio, con una explicación de variabilidad por **encima del 85%**, para lograr buenas predicciones.

Aún así, se pretende avanzar en la implementación de un modelo de regresión logística en tanto que, se observa en los corredores en los extremos de rendimiento opuestos, que podrían desarrollar una función de desempeño logarítmica en su recorrido, en menor y mayor grado.

Una vez realizadas las pruebas pertinentes, se confrontarán los resultados contra el modelo lineal y, se aplicará el nuevo modelo a aquellos subconjuntos donde las predicciones sean más acertadas.



CODERHOUSE – DATA SCIENCE

Diciembre 2023 - Buenos Aires

Oscar Leonardo Giménez

leonardo-gimenez@hotmail.com