UNIVERSIDAD DEL VALLE DE GUATEMALA

CC3074 - Minería de datos

Sección 20

Ing. Luis R. Furlán

Laboratorio 2

Dáriel Villatoro, 20776

Cristian Aguirre, 20231

GUATEMALA, 3 de febrero de 2023

Exploración de datos

¿Cuál es el tamaño de los datos?
En total hay 17 columnas y 2463 filas en los datos recibidos

¿Cuáles son las columnas de los datos?
Las columnas encontradas de los datos fueron:

- attendance
- away_team
- away_team_errors
- away_team_hits
- away_team_runs
- boxscore_url
- date
- field_type
- game_duration
- game_type
- home_team
- home_team_errors
- home_team_hits
- home_team_runs
- other_info_string
- start_time
- venue

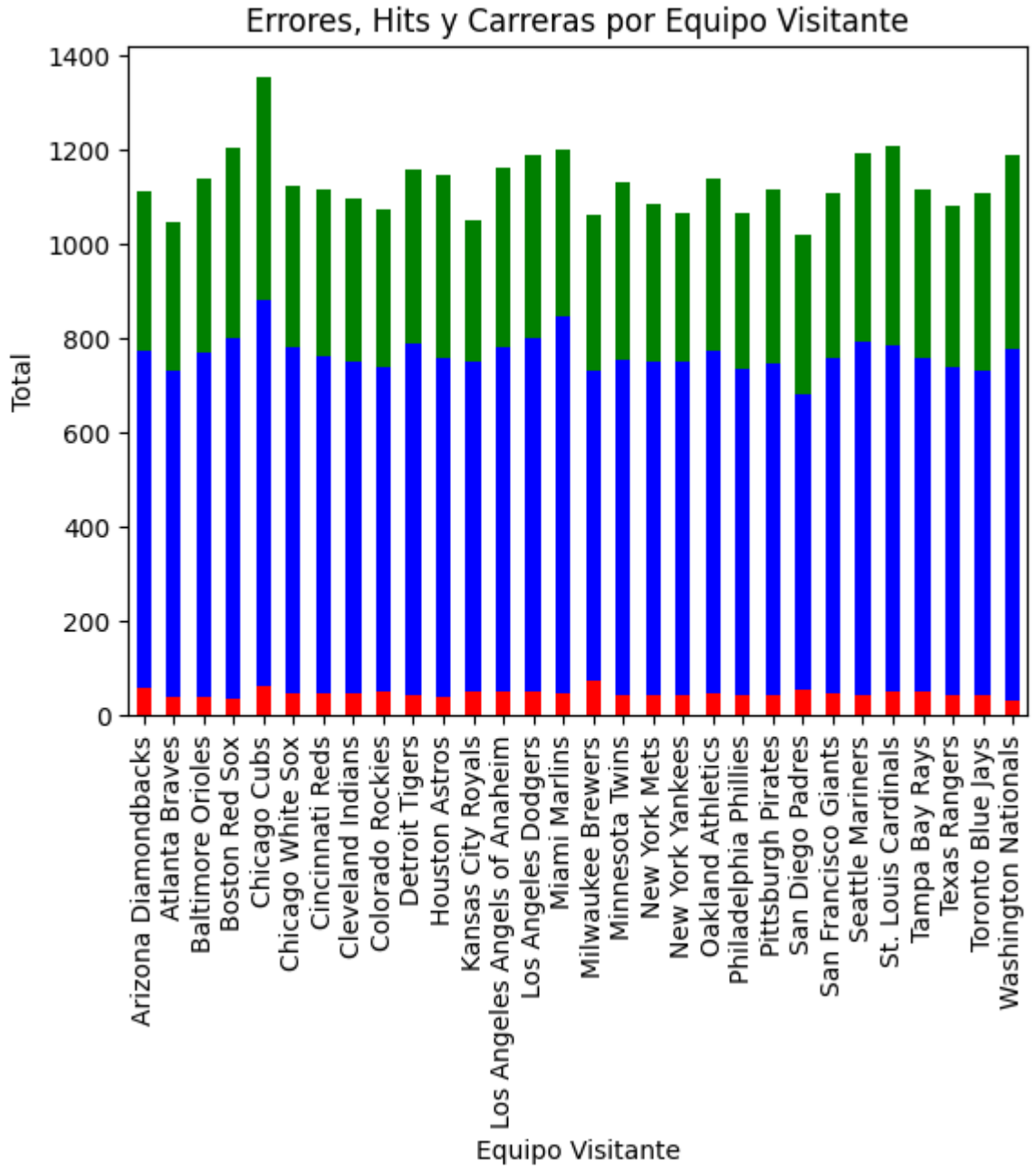
¿Cuáles son los tipos de las variables?
Según la información dada por Pandas, se determino la siguiente distribución de los tipos de los datos:

- Datos categoricos:
    - attendance
    - away_team
    - boxscore_url
    - date
    - game_duration
    - game_type
    - home_team
    - other_info_string
    - start_time
    - venue
- Datos cuantitativos continuos:
    - field_type
- Datos cuantitativos discretos:
    - away_team_errors

- away_team_hits
- away_team_runs
- home_team_errors
- home_team_hits
- home_team_runs

Sin embargo, como se determinará más adelante en la limpieza de los datos, la variable attendance es en realidad una variable cuantitativa discreta.

Gráficas exploratorias



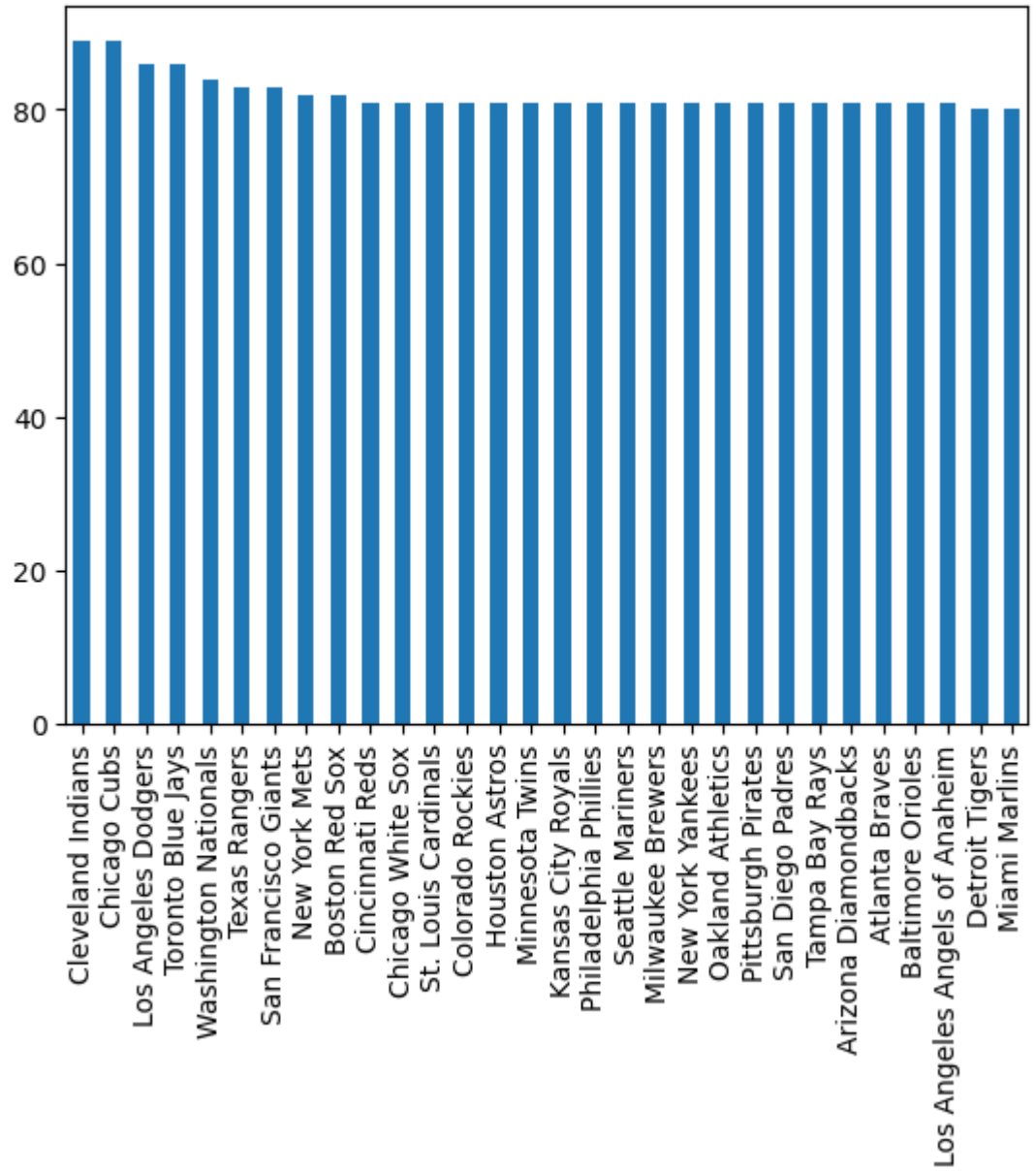


Tabla de frecuencia de los tipos de juegos

Night Game, on grass	1560
Day Game, on grass	733
Night Game, on turf	104
Day Game, on turf	63

Correlación entre las variables numéricas y las categóricas

	away_team_errors	away_team_hits	away_team_runs	\
away_team_errors	1.000000	0.033057	0.042442	
away_team_hits	0.033057	1.000000	0.780760	
away_team_runs	0.042442	0.780760	1.000000	
field_type	NaN	NaN	NaN	
home_team_errors	0.024280	0.190945	0.280002	
home_team_hits	0.153876	0.114150	0.091607	
home_team_runs	0.218470	0.052616	0.038996	

	field_type	home_team_errors	home_team_hits	home_team_runs
away_team_errors	NaN	0.024280	0.153876	0.218470
away_team_hits	NaN	0.190945	0.114150	0.052616
away_team_runs	NaN	0.280002	0.091607	0.038996
field_type	NaN	NaN	NaN	NaN
home_team_errors	NaN	1.000000	-0.007762	-0.008251
home_team_hits	NaN	-0.007762	1.000000	0.769776
home_team_runs	NaN	-0.008251	0.769776	1.000000

Limpieza de los datos

Al obtener la información de nuestros datos notamos que había que hacer dos limpiezas notables. Primero, notamos que la columna ‘game_type’ tenía valores nulos, decidimos tratarlos reemplazando los valores nulos por un valor string ‘undefined’ para no perder esos datos. Además, la columna ‘field_type’ no contenía ningún valor, por lo que se eliminó de los datos.

Luego se corrigió la columna ‘attendance’, que como se mencionó anteriormente, era identificada por Pandas como una variable categórica, por lo que se tuvo que modificar para convertirla a un dato numérico. También se encontraron valores no válidos en la columna ‘attendace’ al hacer la corrección. Ya que los datos faltantes representaban una fracción muy pequeña del conjunto de datos, se optó por eliminarlos.

Se modificó la columna ‘date’ para que se usará un formato soportado por Pandas y a partir de ello se añadió una nueva columna con el día de la semana que correspondía a cada fecha. Esto debido a que se consideró que ese dato sería importante para el desarrollo del modelo

Finalmente, se modificó la columna ‘start_time’ a un formato soportado por pandas para que ayudara en la categorización.

Modelo de regresión lineal

Una vez realizamos la limpieza de datos correspondiente, analizamos las variables que consideramos útiles y debido a que identificamos la necesidad de usar más de una variable, decidimos utilizar un modelo de regresión lineal múltiple. Para comenzar analizamos el problema; se nos solicita calcular la predicción de asistencia de partidos de béisbol, es decir la cantidad de personas que compraron sus tickets, para ello consideramos las variables, away_team, home_team, game_type, venue y day, pues en nuestro análisis e investigación concluimos que estos son factores importantes que influyen en la decisión de los fanáticos al momento de comprar una entrada para un partido.

El primer paso es descartar las variables que no vamos a utilizar en nuestro modelo y crear un dataframe que contenga las variables que mencionamos anteriormente. Una vez completada esta tarea, aplicamos una codificación sobre las variables categóricas en la cual convertimos dichas variables categóricas a numéricas para que puedan ser utilizadas por nuestro modelo y proceder a realizar la separación de datos en dos conjuntos separados “X” y “y”.

El siguiente paso en nuestro modelo consiste en la división del conjunto de datos en un conjunto para entrenamiento de datos y otro para la prueba de datos, para luego poder entrenar el modelo de regresión lineal multiple. Posteriormente realizamos nuestra predicción sobre los datos obtenidos en el paso anterior y podemos concatenar con los datos originales encontrados en nuestro conjunto de prueba y, una vez tenemos estos dos conjuntos concatenados (y_prueba, y_pred) podemos graficar y observar la regresión lineal de mejor manera.

Al usar la funcion R^2_score de la libreria scikit-learn podemos obtener que el valor de R^2 es de 0.150 y la ecuación obtenida es:
y = 37205.179 + (-21.65)x1 + (-1485.97)x2 + (150.32)x3 + (-269.08)x4 + (-853.98)x5