Análisis Exploratorio de los Datos Originales

Raquel García Marañón

2024-04-23

1 Análisis exploratorio de los datos originales

Se lleva a cabo un análisis exploratorio de los datos (EDA), que incluye varias técnicas y herramientas para examinar y resumir sus características principales. Durante el EDA, se realiza un análisis de las estadísticas descriptivas y un análisis visual de las distribuciones. Esto permite entender la naturaleza global de los datos originales para tomar decisiones informadas en el procesamiento de los mismos, lo cual es fundamental durante la fase de modelado y para mejorar la precisión y la robustez del modelo final.

Se comienza con un análisis descriptivo. El conjunto de datos consta de 122.712 registros distribuidos en 22 columnas. Las variables presentes son cruciales para el estudio de las tormentas solares y sus efectos en la Tierra.

Las columnas del conjunto de datos contienen información sobre los parámetros descritos en la sección **??**, información del viento solar y el IMF. La mayoría de las variables son de tipo `float64`, con una excepción de tipo `datetime64[ns]`. En la Tabla **??** se presentan las medias y desviaciones estándar de las variables más relevantes.

Variable	ID_IMF	Bmag	dev_Bmag	Bx	By_gse	Bz_gse
Media	55,12	5,67	0,35	0,05	-0,05	-0,004
Desviación estandar	8,09	3,14	0,45	3,39	3,84	2,84

Variable	By_gsm	Bz_gsm	dev_Bx	Dst	P_density	dev_P_density
Media	-0,05	0,02	1,01	-12,39	5,65	0,62
Desviación estandar	3,74	2,97	0,81	20,35	4,66	0,97

Variable	AP	dev_AP	E_field	plasma_T	dev_plasma_T	plasma_V
Media	0,035	0,005	-0,01	96664,83	14899,26	431,72
Desviación estandar	0,021	0,006	1,39	99703,65	30230,36	106,77

Table 1: Estadísticas descriptivas de las variables seleccionadas

A partir del análisis descriptivo de los datos, se pueden hacer varias observaciones importantes. Las variables como `Bmag`, `Bx`, `By_gse`, `Bz_gse`, `P_density` y otras muestran un amplio rango, indicando una variabilidad significativa en las mediciones. Por ejemplo, la magnitud del campo magnético (`Bmag`) varía desde 0,4 hasta 62, mientras que las componentes del campo magnético en diferentes sistemas de coordenadas (`Bx`, `By_gse`, `Bz_gse`) presentan rangos amplios de valores mínimos y máximos.

La variable `Dst`, que mide la actividad geomagnética, varía considerablemente desde -422 hasta 77. Esto refleja diferentes niveles de actividad geomagnética, que son críticos para entender el impacto de las tormentas solares en la magnetosfera terrestre. Esta variabilidad en la actividad geomagnética es fundamental para la interpretación de los fenómenos asociados con el clima espacial.

Es importante señalar que varias columnas presentan datos faltantes, lo cual debe ser considerado en el análisis. Por ejemplo, el `ID_plasma` tiene 122.586 registros válidos, mientras que `P_density` y `dev_P_density` tienen 119.308 registros válidos. La amplitud de la perturbación (`AP`) y sus desviaciones (`dev_AP`) tienen datos disponibles en 108.014 registros, indicando que no todas las mediciones están completas y que el número de datos faltantes en esta variable es considerable.

La temperatura del plasma (`plasma_T`) muestra una media considerablemente alta de $96664,83$ con una desviación estándar de $99703,65$, lo cual sugiere una amplia variabilidad en las condiciones del plasma solar. Las desviaciones de la temperatura del plasma (`dev_plasma_T`) también presentan un rango amplio, desde 0 hasta 3532095, indicando fluctuaciones significativas en la temperatura.

Además, la velocidad del plasma (`plasma_V`) tiene una media de $431,72$ y una desviación estándar de $106,77$, con valores que varían desde 233 hasta 1.189. Este amplio rango de valores destaca la variabilidad en la dinámica del plasma, esencial para comprender los procesos de transporte y energía en el espacio interplanetario.

La variabilidad en las mediciones y la presencia de datos faltantes son aspectos clave que deben ser considerados en el preprocesado de los datos. Este análisis descriptivo inicial sienta las bases del análisis exploratorio de los datos y será de ayuda para las estadísticas descriptivas subsecuentes, proporcionando una comprensión más profunda de la naturaleza y características del conjunto de datos.

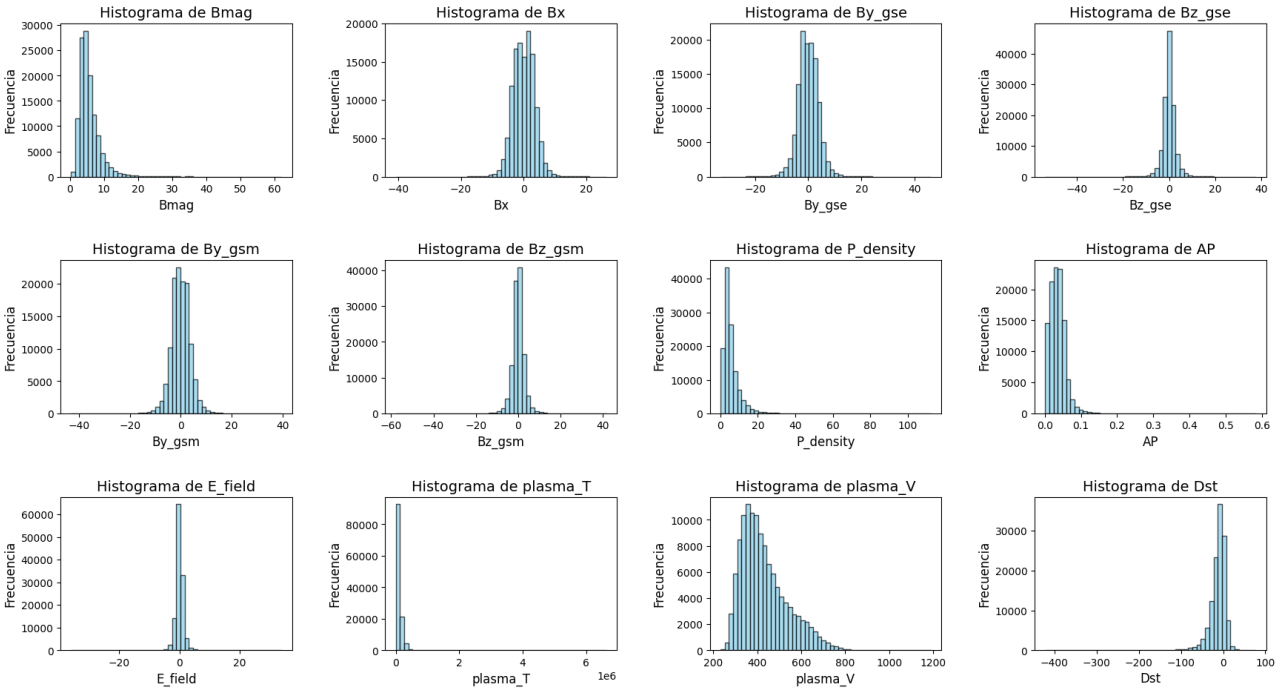

Figure 1: Distribución de los datos originales

A partir del análisis de los histogramas presentados, se pueden realizar varias observaciones importantes sobre la distribución de los datos. Las variables relacionadas con el campo magnético, como `Bmag`, `Bx`, `By_gse`, `Bz_gse`, `By_gsm` y `Bz_gsm`, presentan en su mayoría distribuciones simétricas centradas alrededor de 0, con `Bmag` mostrando una ligera sesgo a la derecha debido a algunos valores atípicos que alcanzan hasta 62. Estas distribuciones indican una variabilidad significativa pero concentrada alrededor de valores medios.

La densidad del plasma (`P_density`) y la amplitud de la perturbación (`AP`) presentan distribuciones sesgadas a la derecha, con la mayoría de los valores concentrados en rangos bajos y

unos pocos valores atípicos altos. Esto sugiere que aunque la mayoría de las mediciones están en rangos esperados, existen eventos ocasionales con valores significativamente mayores.

Por otro lado, el campo eléctrico (`E_field`) muestra una distribución simétrica centrada alrededor de 0, similar a las componentes del campo magnético, indicando una variabilidad que oscila igualmente en ambos sentidos. La temperatura del plasma (`plasma_T`) y su velocidad (`plasma_V`) presentan distribuciones sesgadas a la derecha. La temperatura muestra una amplia variabilidad con algunos valores extremadamente altos, mientras que la velocidad del plasma tiene la mayoría de los valores concentrados alrededor de 400 y algunos picos que alcanzan hasta 1200.

Finalmente, la actividad geomagnética (`Dst`) también muestra una distribución sesgada a la derecha, con la mayoría de los valores concentrados cerca de 0 y unos pocos valores negativos que alcanzan hasta -400. Esta distribución refleja diferentes niveles de actividad geomagnética, cruciales para entender el impacto de las tormentas solares en la magnetosfera terrestre.

La mayoría de las variables presentan distribuciones simétricas centradas alrededor de sus medias, con algunas excepciones que muestran sesgo a la derecha.

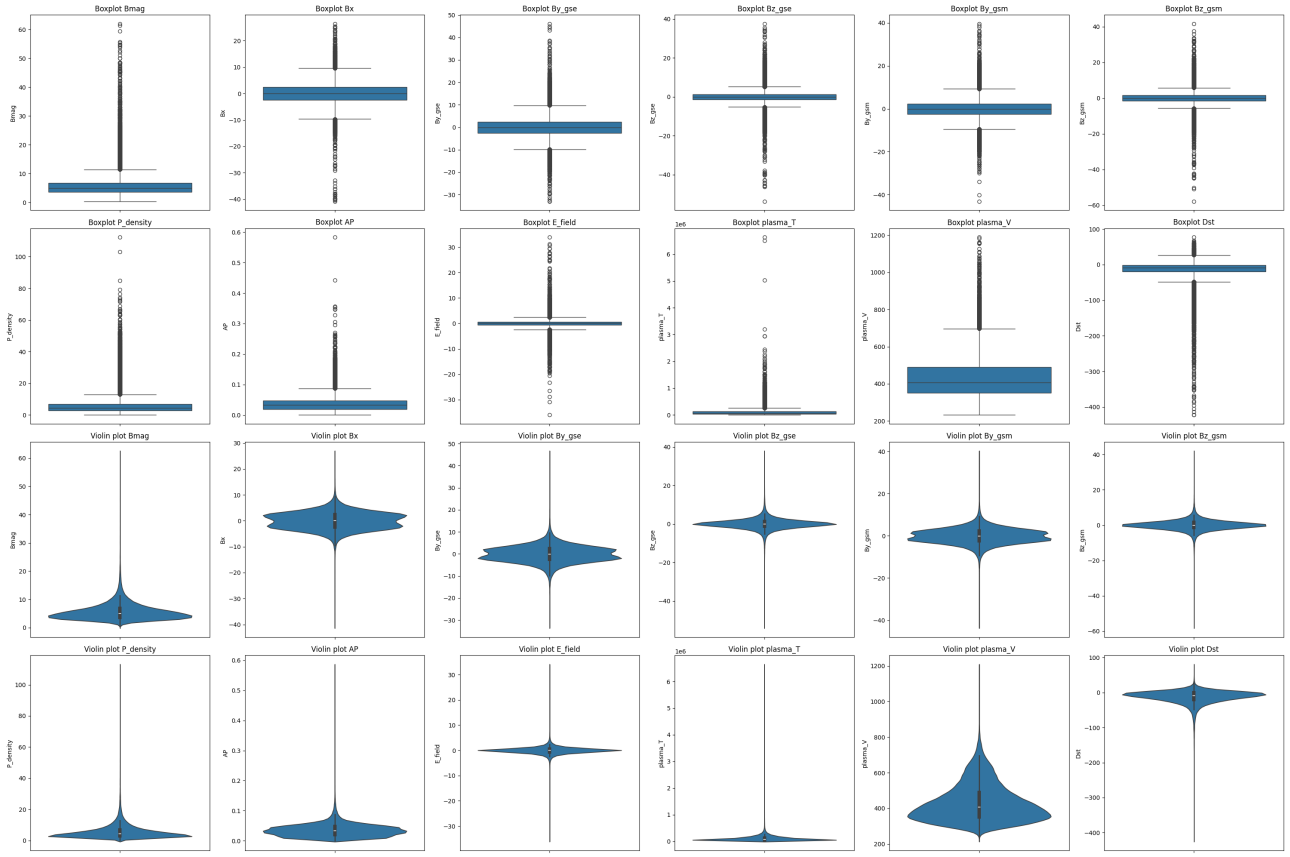

Figure 2: Diagrama de cajas y de violín de los datos originales.

En la Figura 2, se ven las comparaciones entre diferentes características del viento solar utilizando diagramas de caja (*boxplots*) y diagramas de violín (*violin plots*).

En términos generales, los diagramas de caja proporcionan una idea clara de la mediana, los cuartiles y los valores atípicos (*outliers*) de las diferentes características del viento solar, mientras que los diagramas de violín ofrecen una visión más detallada de la distribución de los datos, mostrando la densidad de los datos a diferentes valores.

En primer lugar, para la magnitud del campo magnético `Bmag`, el diagrama de caja muestra una mediana cercana a 5 con muchos valores atípicos hacia arriba. El diagrama de violín muestra una distribución que es bastante asimétrica con una cola larga hacia valores más altos,

lo que indica que aunque la mayoría de los datos están cerca de la mediana, hay algunos valores significativamente más altos.

Los componentes del campo magnético `Bx`, `By`, `Bz`, muestran distribuciones centradas alrededor de cero con valores atípicos en ambas direcciones, indicando variaciones en la intensidad del campo magnético entre los registros.

Para la densidad de protones en el viento solar (`P_density`), el diagrama de caja muestra una mediana baja con una amplia gama de valores atípicos hacia arriba. El diagrama de violín muestra una distribución muy sesgada hacia la derecha con una alta densidad de valores bajos y una cola larga hacia valores más altos, sugiriendo que aunque la mayoría de los valores son bajos, hay algunos eventos de alta densidad, como cabría esperar. El ratio alfa/protón `AP`, así como la temperatura y la velocidad del plasma, (`plasma_T` y `plasma_V`) muestran una distribución similar a la densidad, sugiriendo que la mayoría son valores bajos, pero existen eventos de alta intensidad.

El campo eléctrico (`E_field`) presenta una distribución con mediana baja y algunos valores atípicos altos. La forma del diagrama de violín sugiere que la mayoría de los datos están cerca de la mediana con una distribución que se extiende hacia valores más altos.

Finalmente, el índice de tormenta geomagnética (`Dst`) presenta una mediana cerca de cero (calma) con valores atípicos hacia valores negativos, indicando eventos de tormenta. El diagrama de violín muestra similarmente una distribución con una mayor densidad cerca de cero y una cola que se extiende hacia valores negativos, lo cual es esperable para un índice de tormenta.