Lista de verificación para el análisis de datos en Python

Esta lista proporciona una visión resumida de los pasos a seguir durante el proceso de análisis de datos en Python. Puede usarse como guía durante el proceso de análisis, como rúbrica para evaluar proyectos de análisis de datos o como una forma de evaluar la calidad de un análisis de datos reportado.El análisis de datos en Python es un proceso de varias etapas que implica la recopilación, limpieza, exploración, modelado y evaluación de datos. El objetivo del análisis de datos es obtener información valiosa de los datos para tomar decisiones informadas.

El proceso de análisis de datos en Python generalmente sigue los siguientes pasos:

1. **Recopilación de datos**: Esta etapa implica la recopilación de datos de diversas fuentes, como encuestas, bases de datos, archivos de texto y APIs. Los datos pueden estar estructurados, semiestructurados o no estructurados.
2. **Limpieza de datos**: Esta etapa implica la limpieza de los datos para eliminar valores faltantes, duplicados y erróneos. También implica la normalización de los datos para garantizar la coherencia y la consistencia.
3. **Exploración de datos**: Esta etapa implica la exploración de los datos para comprender su distribución, correlaciones y patrones. Esto se puede hacer mediante técnicas de visualización de datos como histogramas, diagramas de dispersión y gráficos de líneas.
4. **Modelado de datos**: Esta etapa implica la creación de modelos estadísticos o de aprendizaje automático para predecir o clasificar los datos. Los modelos se pueden entrenar utilizando técnicas de aprendizaje supervisado o no supervisado.
5. **Evaluación del modelo**: Esta etapa implica la evaluación del rendimiento del modelo utilizando métricas como la precisión, la sensibilidad y la especificidad. El modelo también se puede evaluar utilizando técnicas de validación cruzada para garantizar que no esté sobreajustado.
6. **Interpretación de los resultados**: Esta etapa implica la interpretación de los resultados del modelo para obtener información valiosa. Esto se puede hacer mediante técnicas de análisis estadístico o de visualización de datos.
7. **Comunicación de los resultados**: Esta etapa implica la comunicación de los resultados del análisis a las partes interesadas. Esto se puede hacer mediante informes, presentaciones o visualizaciones interactivas.

El proceso de análisis de datos en Python es un proceso iterativo que puede repetirse varias veces para mejorar la calidad del análisis. Es importante documentar cada paso del proceso para garantizar la transparencia y la reproducibilidad del análisis.

1 Responder a la pregunta

1. ¿Especificó el tipo de pregunta analítica de datos (por ejemplo, exploración, asociación, causalidad) antes de tocar los datos?
2. ¿Definió la métrica de éxito antes de comenzar?
3. ¿Entendió el contexto de la pregunta y su aplicación científica o empresarial?
4. ¿Registró el diseño experimental?
5. ¿Consideró si la pregunta podría responderse con los datos disponibles?

2 Verificar los datos

1. ¿Graficó resúmenes univariados y multivariados de los datos?
2. ¿Verificó la existencia de valores atípicos?
3. ¿Identificó el código de datos faltantes?

3 Organizar los datos

1. ¿Cada variable es una columna?
2. ¿Cada observación es una fila?
3. ¿Aparecen diferentes tipos de datos en cada tabla?
4. ¿Registró la receta para pasar de datos crudos a datos organizados?
5. ¿Creó un libro de códigos?
6. ¿Registró todos los parámetros, unidades y funciones aplicadas a los datos?

4 Análisis exploratorio

1. ¿Identificó los valores faltantes?
2. ¿Realizó gráficos univariados (histogramas, gráficos de densidad, diagramas de caja)?
3. ¿Consideró las correlaciones entre variables (gráficos de dispersión)?
4. ¿Verificó las unidades de todos los puntos de datos para asegurarse de que estén en el rango correcto?
5. ¿Intentó identificar errores o codificación incorrecta de variables?
6. ¿Consideró graficar en una escala logarítmica?
7. ¿Sería más informativo un gráfico de dispersión?

5 Inferencia

1. ¿Identificó a qué gran población está tratando de describir?
2. ¿Identificó claramente las cantidades de interés en su modelo?
3. ¿Consideró posibles factores de confusión?
4. ¿Identificó y modeló posibles fuentes de correlación, como mediciones a lo largo del tiempo o del espacio?
5. ¿Calculó una medida de incertidumbre para cada estimación en la escala científica?

6 Predicción

1. ¿Identificó de antemano su medida de error?
2. ¿Dividió sus datos inmediatamente en entrenamiento y validación?
3. ¿Usó validación cruzada, remuestreo o bootstrap solo en los datos de entrenamiento?
4. ¿Creó características usando solo los datos de entrenamiento?
5. ¿Estimó parámetros solo en los datos de entrenamiento?
6. ¿Fijó todas las características, parámetros y modelos antes de aplicarlos a los datos de validación?
7. ¿Aplicó solo un modelo final a los datos de validación y reportó la tasa de error?

7 Causalidad

1. ¿Identificó si su estudio fue aleatorio?
2. ¿Identificó posibles razones por las que la causalidad podría no ser apropiada, como factores de confusión, datos faltantes, abandonos no ignorables o experimentos no cegados?
3. Si no, ¿evitó usar lenguaje que implicara causa y efecto?

8 Análisis escritos

1. ¿Describió la pregunta de interés?
2. ¿Describió el conjunto de datos, el diseño experimental y la pregunta que está respondiendo?
3. ¿Especificó el tipo de pregunta analítica de datos que está respondiendo?
4. ¿Especificó con notación clara el modelo exacto que está ajustando?
5. ¿Explicó en la escala de interés qué significa cada estimación y medida de incertidumbre?
6. ¿Reportó una medida de incertidumbre para cada estimación en la escala científica?

9 Figuras

1. ¿Cada figura comunica una pieza importante de información o aborda una pregunta de interés?
2. ¿Todas sus figuras incluyen etiquetas de ejes en lenguaje simple?
3. ¿El tamaño de la fuente es lo suficientemente grande para leer?
4. ¿Cada figura tiene un título detallado que explica todos los ejes, leyendas y tendencias en la figura?

10 Presentaciones

1. ¿Comenzó con una declaración breve y comprensible para todos sobre su problema?
2. ¿Explicó los datos, la tecnología de medición y el diseño experimental antes de explicar su modelo?
3. ¿Explicó las características que usará para modelar los datos antes de explicar el modelo?

4. ¿Se aseguró de que todas las leyendas y ejes fueran legibles desde el fondo de la sala?

11 Reproducibilidad

1. ¿Evitó hacer cálculos manualmente?
2. ¿Creó un script que reproduzca todos sus análisis?
3. ¿Guardó las versiones crudas y procesadas de sus datos?
4. ¿Registró todas las versiones del software que usó para procesar los datos?
5. ¿Intentó que alguien más ejecutara su código de análisis para confirmar que obtuvieran las mismas respuestas?

12 Paquetes de Python

1. ¿Hizo que el nombre de su paquete sea fácilmente encontrable en Google?
2. ¿Escribió pruebas unitarias para sus funciones?
3. ¿Escribió archivos de ayuda para todas las funciones?
4. ¿Escribió una guía o tutorial?
5. ¿Intentó reducir las dependencias a paquetes mantenidos activamente?
6. ¿Eliminó todos los errores y advertencias al ejecutar las pruebas?