Actividad 1: limpieza de datos

**Objetivo general**

Con esta actividad vas a llevar a cabo una de las tareas más importantes y comunes para un científico de datos: la limpieza de datos. Revisarás cada variable, comprobarás la existencia de errores, identificarás datos ausentes, realizarás conversión de formatos si aplica y finalmente, una vez preparado el dataset, lo llevarás a otra forma de persistencia de datos.

**Descripción**

Los datos para esta actividad están en el fichero **data\_act\_01.cvs**. Sobre dicho fichero llevarás a cabo la limpieza de datos. Existen al menos (6) seis errores que deben ser identificados y corregidos, con el fin de limpiar y homogeneizar el dataset final. Para cumplir con esta tarea se pide:

1. **Describe los problemas**: para cada uno de los elementos susceptibles de ser limpiados, indica:

* **Descripción del problema**, indicando claramente:
* Qué o cuál es problema.
* Cómo se ha detectado.
* Por qué crees que se trata de un problema.
* Un ejemplo completo extraído del propio dataset.
* **Forma de solucionarlo**. Ten en cuenta que un mismo problema puede tener diferentes soluciones, por ello debes escoger la solución(es) que consideres más adecuada en cada caso. Es importante también justificar la elección de dicha solución. Es decir, explicar por qué el dataset final, tras aplicar la solución, es el más idóneo para ser tratado o es mejor que el dataset original.

**Importante**: cada problema detectado puede afectar a una o varias filas/columnas de tus datos. No debes centrarte en buscar problemas puntuales sino en identificar problemas tipo a los que habrá que darles una solución. Por ejemplo, si encuentro un error del tipo “X”, aplicaré la solución “Y”.

1. **Propón un formato JSON** equivalente al CSV proporcionado. Cada documento JSON debe contener toda la información contenida en la fila correspondiente del documento CSV. Aprovecha esta conversión de formato (CSV a JSON) para eliminar los campos redundantes, los problemas del punto 1, etc. Describe todos estos cambios o transformaciones para que el profesor conozca el proceso utilizado.
2. **Describe una breve** [**metodología**](https://es.wikipedia.org/wiki/Metodolog%C3%ADa) (o aproximación de un método), clara y replicable que, basándose en tu proceso de limpieza, pueda ser aplicada sobre un nuevo dataset similar al que has tratado. El fin es poder afirmar que, usando tu metodología, el nuevo catálogo de datos estará limpio y listo para ser analizado. La [metodología](https://es.wikipedia.org/wiki/Metodolog%C3%ADa) debe ser un listado de pasos fáciles de seguir, no ambiguos. Se valorará la claridad de la metodología y su fácil aplicación.
3. **Propón al menos dos mejoras en el conjunto de datos**. Después de conocer el conjunto de datos utilizado, qué dos (o más) mejoras propondrías para garantizar la calidad de los datos. Ten en cuenta que estas mejoras pueden afectar al procedimiento de captura de datos o a la estructura de la base de datos, por ello explica tanto la mejora como el procedimiento para llevarla a cabo.

**Entrega**

* Entrega un informe con los elementos habituales (portada, índice, tu nombre, fecha, numeración de páginas, etc.).
* En él da respuesta a los cuatro apartados planteados anteriormente. Cada apartado debe ir resuelto en una sección diferente.
* El fichero JSON debe acompañar al informe propuesto y no debe contener ningún error de formato (*validar previamente el formato JSON del fichero*).

**Extensión**

El informe entregado no debe exceder las 6 páginas, sin contar portada e índices. Fuente utilizada Georgia 11 e interlineado 1,5. Utiliza la fuente Consolas para mostrar fragmentos de código.

**Entrega en el Aula virtual**

Al finalizar la actividad, **harás la entrega del informe final** a través del apartado «Envío de actividades» del aula virtual. El informe y el fichero JSON a entregar, deben ir nombrados como se muestra a continuación:

* APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_1.pdf
* APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_1.json
* Alternativa en ZIP: APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_1.zip

No uses tildes, espacios, apóstrofes u otro carácter que pudiera resultar conflictivo en el nombre del fichero.

**Recomendaciones finales**

* No utilices imágenes para mostrar textos o fragmentos de código. Utiliza directamente el texto o el código en cuestión.
* Puedes utilizar capturas de pantalla para mostrar los pasos ejecutados con tus herramientas de limpieza.
* Esta actividad no se trata de ver lo bueno que eres usando Python, R, OpenRefine, Excel o cualquier otra herramienta. Se espera que realices un buen trabajo de limpieza, independientemente de qué uses para ello.
* No hay ningún problema en usar más de una herramienta de apoyo siempre que conozcas las fortalezas y debilidades de cada una de ellas.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Trabajo: Limpieza de dataset | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | **Apartado 1**:  Para cada uno de los elementos limpiados:  -La descripción del problema es correcta.  -La solución aportada es correcta y es replicable. | 4.0 | 40% |
| Criterio 2 | **Apartado 2**:  -El formato propuesto JSON es correcto.  -El JSON propuesto contiene la misma información que el CSV de origen (en caso contrario se argumentan claramente los cambios).  - Se respeta el formato de todos los campos (en caso contrario se argumentan claramente los cambios).  - Se han eliminado los campos redundantes. | 2.5 | 25% |
| Criterio 3 | **Apartado 3**:  - La metodología es adecuada para la limpieza del dataset  - La metodología está explicada a través de pasos fáciles de seguir | 2.0 | 20% |
| Criterio 4 | **Apartado 4**:  - Las mejoras propuestas son pertinentes.  - Las mejoras propuestas están bien explicadas.  - La implantación de las mejoras propuestas está bien explicada. | 1.5 | 15% |
|  |  | **10.0** | **100 %** |