**Documentación del proceso de limpieza de datos:**

**Creamos una carpeta dedicada a esta sección llamada “Proceso de datos sucios a datos limpios”**

1. Comenzamos descargando los archivos CSV directamente del Portal de Datos abiertos de la Ciudad de México. [*https://datos.cdmx.gob.mx/*](https://datos.cdmx.gob.mx/)*.* Ambos conjuntos de datos, Afluencia Diaria del Metro (Simple) y Afluencia Diaria del Metro (desglosada) se localizan en la siguiente liga: [*https://datos.cdmx.gob.mx/dataset/afluencia-diaria-del-metro-cdmx*](https://datos.cdmx.gob.mx/dataset/afluencia-diaria-del-metro-cdmx)*.* Estos conjuntos de datos en su forma “cruda” se localizan en la subcarpeta “stc\_afluencia\_archivos \_csv” de la primera carpeta “1. Afluencia Diaria Metro CDMX”.

A continuación, ambos CSV se transforman a una hoja de cálculo, de tal forma que los datos quedan transformados en formato de tabla. Se les asigna el nombre de “Metro\_desglosado” y “Metro\_simple” respectivamente.

* 1. *Comenzamos por analizar el conjunto de datos “Metro\_desglosado”.*

Ya que la tabla por defecto crea un desplegable para colocar filtros, nos apoyamos de éstos para buscar posibles errores en los registros.

Se verificó que la columna de “fecha” tuviera el formato de fecha corta y no se encontraron inconsistencias, para asegurar la consistencia, seleccionamos toda la columna correspondiente a fecha y seleccionamos el formato de fecha corta dd/mm/aaaa. No se detectaron valores en blanco o NA.

En la columna de mes, se detectaron NA como valores sugeridos para filtrar. Filtramos los datos que indicaban NA y se detectó que para el período de diciembre de 2023 para todas las líneas en lugar de indicar el mes de diciembre indicaba los valores NA. Corrigiendo el primer valor de NA a diciembre, se utilizó Auto llenado para sustituir los demás NA. Se borró luego el filtro de la columna Mes y ya no aparecieron los NA. El formato de la columna es formato General.

En la columna de “anio” se verificó que los años estuvieran registrados como 2021, 2022 y 2023. No se encontraron errores.

En la columna “línea” se verificó con ayuda del filtro que las estaciones abarcaran las líneas de la 1 a la 9, Línea 12, Línea A y Línea B. Teniendo un total de 12 líneas, que es consistente con las que se cuentan actualmente en toda la red del STC. El formato de la columna es formato general.

Para la columna de estación, se usó un método diferente, listado más adelante. \*

Para la columna de afluencia se utilizó el filtro y no se registraron valores menores a 0, todos los valores son numéricos y la columna es de formato general.

De igual forma, tenemos en cuenta que existen 195 estaciones en toda la red del STC, por cada estación se contabilizan 3 filas para la misma fecha haciendo distinción entre el tipo de ingreso. Esto se hizo durante 3 años (2021-2023). Es decir, un cálculo preliminar debería decir que existen 640575 entradas. (195\*3) \*3 \*(365) = 640575. Contando el número de filas con Excel, obtenemos exactamente el número calculado.

* 1. *Después se analizó el conjunto de datos “Metro\_simple”*

Se siguió el mismo procedimiento que el conjunto de datos “Metro\_desglosado” y de forma similar, se corrigieron los valores NA que se presentaban en todos los registros para todas las estaciones en diciembre de 2023. Fue la única corrección que se realizó a los datos.

Para el conjunto de datos simple, debemos tener en consideración que los registros se llevan a cabo a lo largo de 14 años (2010-2023). Tomando en cuenta que los años 2012, 2016 y 2020 fueron bisiestos y que para cada fecha y cada estación sólo se lleva un solo registro en lugar de tres. Haciendo un cálculo:

(195\*365)\*11 + (195\*366)\*3 = 997,035 registros. Verificando el número de filas con ayuda de Excel, efectivamente contamos con el número de registros calculados.

1. Volviendo al conjunto de datos “Metro\_desglosado” se presenta una particularidad. Como se mencionó anteriormente, por cada estación se contabilizan 3 filas para la misma fecha distinguiendo entre los tipos de ingreso. Para fines prácticos del análisis, se decidió colocar los tres tipos de observaciones en una fila, es decir pasar del formato largo al ancho. De tal manera que tuviéramos nuevas columnas, con boleto, prepago y gratuidad.

De esta forma, por cada tres observaciones por estación, se reduciría a una. Pero para lograr colocar los tres tipos de ingreso en una misma fila,

Se creó otro archivo llamado “Metro desglosado transpuesto”. En la carpeta “2. Conteo de afluencia transpuesta”

Se creó una copia del conjunto de datos en otra hoja.

Luego, en las celdas I1, J1 y K1 se colocaron las nuevas columnas, boleto, prepago y gratuidad.

En la celda I2 se escribió la siguiente fórmula:

=SI.ERROR(INDICE($G$2:$G$640575,(FILA()-4)\*3+COLUMNA()-2),"")

Arrastrándola a K2 y posteriormente usando el Auto llenado, se transpusieron los valores de las celdas hasta la fila 213526.

A continuación se muestra un ejemplo. Si tenemos estas 6 observaciones, transponiendo datos con la fórmula, deberíamos reducir las filas a sólo dos.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **fecha** | **mes** | **anio** | **linea** | **estacion** | **tipo\_pago** | **afluencia** |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | Boleto | 1848 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | Prepago | 790 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | Gratuidad | 325 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | Boleto | 1103 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | Prepago | 922 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | Gratuidad | 651 |

Aplicando la fórmula mencionada, obtendríamos el siguiente conjunto de datos:

|  |  |  |
| --- | --- | --- |
| Boleto | Prepago | Gratuidad |
| 1848 | 790 | 325 |
| 1103 | 922 | 651 |

Y de esta manera se reducen 640575 registros a 213525.

Pero como queremos que sólo exista un registro por fecha para cada estación, una vez transpuestos los datos, necesitamos eliminar las dos últimas columnas. A continuación, necesitamos conservar sólo una fila y eliminar las siguientes dos. Por ejemplo, queremos pasar de esto:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **fecha** | **mes** | **anio** | **linea** | **estacion** | **tipo\_pago** | **afluencia** |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | Boleto | 1848 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | Prepago | 790 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | Gratuidad | 325 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | Boleto | 1103 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | Prepago | 922 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | Gratuidad | 651 |

A esto:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **fecha** | **mes** | **anio** | **linea** | **estacion** |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas |

Aquí nos apoyamos de una macro para realizar esta acción. Por lo cual, del conjunto de datos ya limpio de “Metro\_desglosado”, aplicando filtro al año, dividimos el conjunto en los años 2021, 2022 y 2023. Y por cada año, lo dividimos en los respectivos meses. Teniendo así tres nuevos libros de trabajo, “2021\_metro”, “2022\_metro” y “2023\_metro”. Cada libro de trabajo posee 13 hojas, 12 correspondientes a los meses y una hoja correspondiente al año, donde se pegaron los resultados obtenidos por la macro para cada mes. Esta división fue necesaria dado el tiempo de procesamiento de Excel y de la computadora.

La macro utilizada fue la siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Por cada año, deberían existir 71,175 registros, pues ahora se registra una fila por cada estación por cada día del año: 195\*365 = 71,175. Se verificó este cálculo para cada hoja en cada libro de trabajo y coincidió perfectamente. Con lo cual no se perdieron observaciones al pasar de formato largo a formato ancho.

Los tres libros de trabajo se agregaron a la carpeta “3. Datos limpios”.

Finalmente, para lograr un nuevo conjunto de datos ya consolidado a formato ancho se realizó lo siguiente: Se creó un nuevo libro de trabajo titulado “Metro\_desglosado\_limpio”. Luego, de los tres libros de trabajo “2021\_metro”, “2022\_metro” y “2023\_metro” se copiaron y pegaron las tablas de cada año en este nuevo libro de trabajo. Registrando con ayuda de Excel un total de 213,525 registros. Finalmente, del libro de trabajo “Metro desglosado transpuesto” se copiaron los valores obtenidos de la matriz creada y se pegaron al lado. Coincidieron nuevamente 213,525 registros, con lo cual logramos obtener un nuevo conjunto de datos. Luciendo el nuevo conjunto de datos de la siguiente manera:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **fecha** | **mes** | **anio** | **linea** | **estacion** | **boleto** | **prepago** | **gratuidad** |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balbuena | 1848 | 790 | 325 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Balderas | 1103 | 922 | 651 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Boulevard Puerto Aéreo | 4890 | 2797 | 1053 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Candelaria | 0 | 1605 | 1127 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Chapultepec | 4480 | 2942 | 162 |
| 01/01/2021 | Enero | 2021 | Línea 1 | Cuauhtémoc | 857 | 998 | 654 |

En la misma carpeta se adjuntan los archivos “Metro\_desglosado\_limpio” y “Metro\_simple\_limpio”.

* \* Para la columna de estación se utilizó un método diferente. Ya que son en total 195 estaciones pero existen transbordos, es decir Pantitlán aparece como parte de la Línea A, 1, 5 y 9, Centro Médico como parte de la línea 3 y 9. Por lo cual, debemos contar los nombres únicos de las estaciones, y cuántas repeticiones ocurren en los nombres. De tal forma que ambos conteos deben sumar 195 estaciones en total.

Tomamos un conjunto de muestra de la hoja “metro\_desglosado”. Este conjunto consistió en tomar una muestra, filtrando el año, mes y día. Por ejemplo, se tomó el día 1 de abril de 2021 para hacer el conjunto de muestra. Se copió este conjunto en el libro de trabajo “estaciones\_unicas”.

Se creó una tabla dinámica. En el campo de filas se colocó la estación y en la suma de valores se colocó estación. De esta forma nos retorna un conteo del número de estaciones que aparecen en el día mencionado anteriormente.

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Se copió el resultado a un nuevo libro de trabajo llamado “estaciones\_unicas” y se copiaron los resultados. Y ahora sí contabilizamos los nombres únicos. Los resultados aparecen en el libro de trabajo, contabilizando así 163 nombres únicos, donde existen 135 nombres de estaciones que no tienen transbordos y 28 nombres de estaciones que cuentan con transbordos (2, 3 y 4).

Los resultados se guardaron en la carpeta “3. Datos limpios”