

Cleaning documentation:

Orders_data Tab:

- User_ID & Order_ID:
 - Several rows had scientific notation for the id columns - changed all of these to the general number format that the other rows displayed
- Purchase_TS:
 - One row had a value of /N for this column, which is nonsensical. Changed this to blank to align with some of the other blank values
- Marketing_CHANNEL and ACCOUNT_CREATION_METHOD:
 - Set blanks to “unknown” values to align with the unknown values.
- Product_Name:
 - 27in 4K Gaming Monitor values had typos - consolidated these
- Refund_TS:
 - There were several nonsensical values (e.g. 2025 Refund code - set these values to blank)
- Region:
 - Pulled in Regions as a separate column from the Country_Lookup Tab. There were several rows where lookups returned blanks. For these, these were because a corresponding country code was not provided. For these, I looked at the Currency code, searched which country the currency code corresponded to, and filled in the Region where that country resides.
- Duplicate Order IDs:
 - About 14% of the Order IDs were duplicates of the entire Order IDs. I analyzed these to see if these were true duplicates or multiple quantities of the same product tied to a single order ID. These were true duplicates (contained duplicate data across all columns of the dataset) - thus, they were removed.
- USD_Price \$0 transactions:
 - There were 158 0\$ transactions. I removed these from the dataset.

Country_Lookup Tab:

- US Region was set to “North America” - changed this to “NA” to align with a majority of the other values that had the same region
- There were regions that had a value of x for the country code, US. Changed this to “NA”
- Two Country codes had blanks - “BJ” and “BM” - looked up where these country codes refer to. They are Benin and Bermuda respectively. Also discovered where these countries reside in terms of regions (EMEA and NA) - filled these values in.