Exercise 3.6

1. **Check for and clean dirty data:** Find out if the film table and the customer table contain any dirty data, specifically non-uniform or duplicate data, or missing values.
   1. **Non-uniform**
      1. I ran several queries on different columns to check for non-uniform data, but there wasn’t anything noticeably out of place for either the film or the customer tables. If there was, I would use **UPDATE, SET** and **WHERE** to fix the necessary records.

A white rectangular object with a corner

Description automatically generated A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A white rectangular object with a white background

Description automatically generatedA white rectangular object with a black border

Description automatically generatedA white rectangular object with a white background

Description automatically generated

* 1. **Duplicate data**
     1. There were no duplicates found, but if there were I would have either deleted the duplicates, or created a “view” table that selected only unique values ensuring that only one of the duplicates was selected to view.

A screenshot of a computer program

Description automatically generated A screenshot of a computer program

Description automatically generated

* 1. **Missing values**
     1. There were no missing values found for either table, however if there were, I would try to fill the missing values depending on how necessary the data was. For example, if one of the customer’s last name information was missing, I would likely ignore it, but if something important like the customer id, or the film title or rating was missing, I would do my best to find that information and fill it in.

A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

1. **Summarize your data:** Use SQL to calculate descriptive statistics for both the film table and the customer table. For numerical columns, this means finding the minimum, maximum, and average values. For non-numerical columns, calculate the mode value.
   1. Film Table Numerical:

A screenshot of a computer program

Description automatically generated

* 1. Film Table non-numerical

A screenshot of a computer

Description automatically generated

* 1. Customer Numerical
     1. there are no columns in the customer table that would make sense calculated as numerical columns.
  2. Customer non-numerical
     1. A screenshot of a computer

        Description automatically generatedthe store\_id and activebool are the only two columns that really made sense to find the mode of:

1. **Reflect on your work: Back in Achievement 1 you learned about data profiling in Excel. Based on your previous experience, which tool (Excel or SQL) do you think is more effective for data profiling, and why? Consider their respective functions, ease of use, and speed. Write a short paragraph in the running document that you have started.**
   1. While Excel is generally more user-friendly and great for small datasets and quick visualizations, it struggles with large amounts of data, making it slower and more prone to errors with large datasets. For data profiling, SQL is generally more effective than Excel. SQL is specifically designed for managing and analyzing large datasets quickly and efficiently. It allows you to write powerful queries to filter, aggregate, and analyze data directly within the database, which ensures speed and accuracy.