**1)**

**film-table:**

**checking for duplicates:**

SELECT title,

release\_year,

language\_id,

rental\_duration,

COUNT(\*)

FROM film

GROUP BY title,

release\_year,

language\_id,

rental\_duration

HAVING COUNT(\*) >1;

⇒ No duplicates were found

⇒ if there are only a few duplicates I could conduct research by myself if two movies under the same name have been published in this year etc…

**checking for incorrect/Missing data:**

SELECT language\_id,

FROM film

GROUP BY language\_id;

⇒ I repeated this for the columns rating, release\_year, rental\_duration

⇒ Now I am curious why there are only movies in English, published in 2006?? Did I do something wrong when importing this data?

⇒ No missing or incorrect Data was found

**customer-table:**

**checking for duplicates:**

SELECT first\_name,

last\_name,

email,

address\_id,

COUNT(\*)

FROM customer

GROUP BY first\_name,

last\_name,

email,

address\_id

HAVING COUNT(\*) >1;

⇒ No duplicates were found

⇒ I would ask if people can create a second account in another store. If not, I could delete the duplicated data.

**checking for incorrect/Missing data:**

SELECT store\_id

FROM customer

GROUP BY store\_id;

⇒ there are

⇒ Since we only have two stores in our database, I would ignore the dirty data.

**2)**

**film-table:**

| **column name** | **AVG** | **MIN** | **MAX** |
| --- | --- | --- | --- |
| rental\_duration | 4.985 | 3 | 7 |
| rental\_rate | 2.98 | 0.99 | 4.99 |
| length(minutes) | around 115 | 46 | 185 |
| replacement\_cost(Dollar) | 19.98 | 9.99 | 29.99 |

SELECT AVG(**column**),

MIN(**column**),

MAX(**column**)

FROM film;

| **column name** | **Mode** |
| --- | --- |
| release\_year | 2006 |
| language\_id | 1 ⇒ english |
| rating | PG-13 |
| special\_features | {Trailers,Commentaries,"Behind the Scenes"} |

SELECT mode() WITHIN GROUP (ORDER BY **column**)

AS modal\_value

FROM film;

**customer-table:**

⇒ There are no numerical columns in this table where a summary would make sense.

| **column name** | **Mode** |
| --- | --- |
| store\_id | 1 |
| first\_name | Jamie |
| last\_name | Abney |
| create date | 2006-02-14 |

SELECT mode() WITHIN GROUP (ORDER BY **column**)

AS modal\_value

FROM column;

**3)**

Finding out where the data has to be cleaned appears to be simpler using SQL. Excel, on the other hand, makes it easy to delete duplicates. Conducting statistical summaries does not appear to be easier or more difficult in any tool. In my opinion, it is simply a matter of personal preference. The working time is a significant advantage for SQL. When working with data in Excel, I've had an occasional lag. When using SQL, this is not the case.