

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. Create a new "Answers 3.6" document and copy-paste your queries into it. Next to each query write 2 to 3 sentences explaining how you would clean the data (even if the data is not dirty).

a. *Film title contains the word Uptown in any position*

```
SELECT film_id, title, description
FROM film
WHERE title LIKE '%Uptown%'
```

b. *Film length is more than 120 minutes and rental rate is more than 2.99*

```
SELECT film_id, title, description
FROM film
WHERE length > 120 AND rental_rate > 2.99
```

c. *Rental duration is between 3 and 7 days (where 3 and 7 aren't inclusive)*

```
SELECT film_id, title, description
FROM film
WHERE rental_duration > 3 AND rental_duration < 7
ORDER BY rental_duration ASC
```

d. *Film replacement cost is less than 14.99*

```
SELECT film_id, title, description
FROM film
WHERE replacement_cost < 14.99
```

e. *Film rating is either PG or G*

```
SELECT film_id, title, description
FROM film
WHERE rating IN ('PG', 'G')
```

2. **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. Copy-paste your SQL queries and their outputs into your answers document.

- *Count of the movies*
- *Average rental rate*
- *Maximum rental duration and minimum rental duration*

```
SELECT rating,
COUNT(title),
AVG(rental_rate),
MIN(rental_duration),
MAX(rental_duration)
FROM film
WHERE rating IN ('PG', 'G')
GROUP BY rating;
```

3. **To make the output easier for your coworkers to understand, give your aggregate columns the following aliases: “count of movies,” “average movie rental rate,” “maximum rental duration”, and “minimum rental duration”. Run the query and transfer the result into your Excel file on a new sheet as well as the code you used to get there.**

```
SELECT rating,
COUNT(title) AS count_of_movies,
AVG(rental_rate) AS average_movie_rental_rate,
MIN(rental_duration) AS minimum_rental_duration,
MAX(rental_duration) AS maximum_rental_duration
FROM film
WHERE rating IN ('PG', 'G')
GROUP BY rating;
```

4. **The customer team would like to see the fields you calculated in step 3 grouped by rating. The totals in your results table should look the same as in step 3, but broken down by the rating column. Copy-paste your query and its output in your answers on a new sheet.**

```
SELECT rating,
COUNT(title) AS count_of_movies,
AVG(rental_rate) AS average_movie_rental_rate,
MIN(rental_duration) AS minimum_rental_duration,
MAX(rental_duration) AS maximum_rental_duration
FROM film
WHERE rating IN ('PG', 'G')
GROUP BY rating;
```