Ans

1. Determine the number of active ads on the site per day (use generate\_series)

Use the generate series function and create the select query. Which is in generate\_series.sql

1. Determine the YoY growth of Beef Cattle ads in county Tipperary from 2016 to 2017 (use lag).

Attached in lag.sql

B. Car Data

Analyse the integrity of the car data.

The purpose is

1.Check the data along with data type.

2. Take the count, mean, and do the cleaning of the data then load into database.

|  |  |
| --- | --- |
| id |  |
| date |  |
| price | Use mode function to fill the banks |
| county |  |
| make |  |
| model | Filter the data and remove the unnecessary characters. |
| year | Use dropna as well filter the year. |
| colour | Done data profiling |
| mileage | Use mode function to fill the banks |
| mileageType | Done data profiling |
| engine\_cc | Use mode function to fill the banks |
| body\_type | Done data profiling |
| fuel\_type | Done data profiling |
| transmission | Done data profiling |
| doors | Use mode function to fill the banks |
| seller\_type | Done data profiling |

B. What data quality issues did you find?

Use Pandas Profiling to get the data quality. It shows the distinct, min, max , missing in percentage as well in numbers.

Pearson co-relation test to find the attributes as co related.

C. Write an ETL script using Python and PETL to standardize the mileage in km.

Use PETL package to filter the data and use preprocessing to standardize the data.