ASSIGNMENT 01 // TAXI EMISSIONS DATA

This project uses public data sets for (1) NYC taxi records from 2024; and (2) emissions data for each taxi type. You will work with both YELLOW and GREEN trip records for all 12 months of 2024, which are available as Parquet files from the NYC Taxi Comission. Your assignment is to insert those records into consolidated tables of a local database, clean and transform it, and then correlate values for each ride based on known emissions data. This will allow you to calculate several required outputs for an initial analysis of the data. Points are earned for each element fulfilled correctly.

QUESTIONS ARE DESIGNED TO EVALUATE STAGES OF DATA PROCESSING AND SKILL LEVELS FROM LOW TO HIGH. THESE MIMIC THE REQUIREMENTS A DATA ENGINEER IS PRESENTED IN INDUSTRY OR RESEARCH PROJECTS.

Student Name: ______________________________

	YES	NO	TOTAL
GENERAL REQUIREMENTS			11 points
PROJECT SUBMISSION: ALL REQUIRED FILES (CODE, OUTPUT IMAGE, ETC.) ARE SUBMITTED BY THE DEADLINE.	□	□	2
CODE EXECUTION: THE PROVIDED CODE RUNS WITHOUT MAJOR ERRORS (E.G., SYNTAX ERRORS, OR UNHANDLED EXCEPTIONS THAT CRASH THE PROGRAM).	□	□	2
FORKED SUBMISSION: THE SUBMITTED REPOSITORY IS A FORK OF THE SOURCE.	□	□	1
COMMENTS: SUBMISSION INCLUDES INLINE COMMENTS FOR EACH CLASS/FUNCTION	□	□	1
FILE ORGANIZATION: PROJECT FILES ARE REASONABLY ORGANIZED (E.G., SEPARATE FOLDERS FOR DATA, SCRIPTS).	□	□	1
LANGUAGE: PROJECT IS WRITTEN IN PYTHON, MARKDOWN, AND SQL. BASH SCRIPTS ARE DISALLOWED.	□	□	1
ERROR HANDLING: PYTHON INCLUDES PROPER ERROR HANDLING THROUGHOUT.	□	□	1
LOGGING: PYTHON INCLUDES LOGGING TO LOG FILEs IN THE PROJECT DIRECTORY. ONE LOG SHOULD EXIST FOR EACH OF THE STAGES BELOW.	□	□	1
NO CRUFT: YOUR REPOSITORY SHOULD NOT CONTAIN ANY (a) PARQUET FILES; (b) LOG FILES; OR (c) LOCAL DATABASE FILES.	□	□	1
DATA LOADING			6 points
DATA LOADING: SUCCESSFULLY LOADS DATA FROM BOTH YELLOW AND GREEN TRIP FILES FOR 2024 INTO DUCKDB TABLE(S). SUCCESSFULLY LOADS EMISSIONS TABLE INTO SEPARATE DUCKDB TABLE.	□	□	2
BASIC DATA SUMMARIZATION: PERFORMS BASIC DESCRIPTIVE STATISTICS OR SIMPLE AGGREGATIONS ON THE INGESTED DATA. OUTPUT TO SCREEN AND TO LOG.	□	□	2
PROGRAMMATIC LOADING: MULTIPLE FILE SOURCES ARE LOADED PROGRAMMATICALLY, INSTEAD OF STATICALLY.	□	□	2
DATA CLEANING			7 points
TABLES UPDATED TO REMOVE DUPLICATE TRIPS	□	□	1
TABLES UPDATED TO REMOVE TRIPS WITH 0 PASSENGERS	□	□	1
TABLES UPDATED TO REMOVE TRIPS 0 MILES IN LENGTH	□	□	1
TABLES UPDATED TO REMOVE TRIPS GREATER THAN 100 MILES IN LENGTH	□	□	1
TABLES UPDATED TO REMOVE TRIPS GREATER THAN 24 HOURS IN LENGTH	□	□	1
INCLUDE TESTS TO VERIFY ABOVE CONDITIONS NO LONGER EXIST. OUTPUT TO SCREEN AND LOG.	□	□	2

DATA TRANSFORMATION			13 points
INSERT CALCULATED COLUMN WITIH CO2 PER TRIP IN KILOGRAMS.	□	□	2
INSERT CALCULATED COLUMN WITH AVG MPH PER TRIP	□	□	1
INSERT CALCULATED COLUMN TRIP HOUR	□	□	1
INSERT CALCULATED COLUMN TRIP DAY OF WEEK	□	□	1
INSERT CALCULATED COLUMN WEEK NUMBER	□	□	1
INSERT CALCULATED COLUMN MONTH	□	□	1
DBT: PERFORM ALL TRANSFORMATIONS ABOVE USING DBT INSTEAD OF SCRIPT	□	□	6

DATA ANALYSIS			8 pts
OUTPUTS: CODE RENDERS SUMMARY ANALYSIS RESULTS AS DESCRIBED IN README. OUTPUTS RESULTS TO BOTH SCREEN AND LOG FILE. **SEE README FILE FOR ITEMIZED LIST OF REQUIRED OUTPUTS (6).**	□	□	6
PLOT: CODE RENDERS TREND PLOT AS DESCRIBED IN README. MUST BE IN JPG/GIF/PNG FORMAT, ADDED AND COMMITTED TO REPOSITORY.	□	□	2

EXTENDED PROCESS			5 pts
BROADEN THE SCOPE OF THIS ENTIRE PROJECT TO ENCOMPASS THE YEARS 2015 THROUGH 2024 (10 YEARS). MODIFY LOAD SCRIPTING, CLEANING, TRANSFORMATIONS APPROPRIATELY	□	□	3
BROADEN THE SCOPE OF YOUR ANALYSIS TO PROVIDE HIGHEST/LOWEST FIGURES ACROSS THE ENTIRE TIMESPAN, NOT JUST 2024. PLOT SHOULD REPRESENT ALL 10 YEARS.	□	□	2