




# Wk4 SQL Mini-Project

## Project Brief

	<p>select and explore datasets of your choice from various sources like Kaggle, APIs, or through web scraping (if you choose to use more than one, ensure they complement each other for a cohesive analysis.) → <b>we need at least 2 distinct sources</b></p> <p><b>TOPIC: Reforestation [!!!!]</b> Kaggle dataset in % <a href="https://www.kaggle.com/datasets/konradb/deforestation-dataset">https://www.kaggle.com/datasets/konradb/deforestation-dataset</a></p> <p>Food and Agriculture Org. of the United Nations data in hectares <a href="https://data.apps.fao.org/catalog/dataset/forest-area-1990-2020-1000-ha/resource/3cc5000d-184c-4ebe-8e54-b71910111f12">https://data.apps.fao.org/catalog/dataset/forest-area-1990-2020-1000-ha/resource/3cc5000d-184c-4ebe-8e54-b71910111f12</a></p> <p>WBG population dataset Csv file downloaded from World Data Group</p> <p>Tourism data: top spots <a href="https://worldpopulationreview.com/country-rankings/most-visited-countries">https://worldpopulationreview.com/country-rankings/most-visited-countries</a></p> <p>→ countries in same format in the 2 datasets :ISO 3166-1, alpha code 3 for country in 3 letters</p> <p><b>For HeatMap</b> <a href="https://gist.github.com/tadast/8827699">https://gist.github.com/tadast/8827699</a> tadast/countries_codes_and_coordinates.csv This link has a map with forests which would be interesting <a href="https://sdga2022.github.io/sdga2022/goal-15-life-on-land?lang=en">https://sdga2022.github.io/sdga2022/goal-15-life-on-land?lang=en</a></p>
	<p>formulate hypotheses related to the datasets selected <b>The goal is to craft a narrative using the data</b> → <b>we need at least 2 clear research questions OR problem statements</b></p>
	<p>sketch out an Entity-Relationship Diagram (ERD) highlighting significant data fields and relationships</p>

	→ <b>we need at least 3 tables</b> Make sure to apply suitable primary and foreign keys and delineate the relationships between tables
✓	then proceed to create and populate a functional database
✓	During the process, <ul style="list-style-type: none"> <li>• carry out data wrangling to clean your datasets and prepare them for analysis</li> <li>• use SQL to run queries, derive valuable insights, and summarise your findings</li> </ul>
✓	For the data analysis: → <b>we need at least 5 SQL queries using fundamental clauses</b>
✓	For the data visualisation: → <b>we need at least 2 graphs/plots using Matplotlib or Seaborn</b>
✓	<b>To conclude, visualise the derived insights using Python libraries and compile a comprehensive report that encapsulates your data story from exploration to insights</b>



## Deliverables

✓	GitHub Repo → sql-database <a href="https://github.com/TaniaZakowski/sql-database">https://github.com/TaniaZakowski/sql-database</a>
	A README.md contain information about the project, so that anyone that reads the README should be able to understand the project without having to look through all of the files
	Code and files that demonstrate the data pipeline coverage of acquisition, transformation, loading, analysis, and reporting.  This includes: <ul style="list-style-type: none"> <li>• Database: The exported <code>.sql</code> file should be included with the final schema ✓</li> <li>• Entity relationship diagram (ERD) ✓</li> </ul>

	<ul style="list-style-type: none"> <li>• SQL Queries: A compilation of all SQL queries used during the project ✓</li> <li>• Python files. Remember, when using Python the code must be put in wrapped functions ✓</li> <li>• At least 1 Jupyter notebook containing the report in full with visualisations ✓</li> </ul>
✓	<p>URL of the presentation</p> <ul style="list-style-type: none"> <li>• Talking with Slides: <b>7 minutes</b></li> <li>• Demo: <b>3 minutes</b></li> <li>• Total: <b>10 minutes</b></li> </ul>
✓	<p>Presentation structure:</p> <ol style="list-style-type: none"> <li>1. <b>Title Slide</b> (1 slide): <ul style="list-style-type: none"> <li>• Project title.</li> <li>• Team members' names.</li> </ul> </li> <li>2. <b>Project Overview</b> (1 slides): <ul style="list-style-type: none"> <li>• Briefly describe the chosen dataset and its origin.</li> <li>• Highlight the business problem and the hypotheses that guided your approach.</li> </ul> </li> <li>3. <b>Data Acquisition, Enrichment, and Examination</b> (1 slide): <ul style="list-style-type: none"> <li>• Highlight the primary sources of data and any complementary datasets.</li> <li>• Discuss challenges faced during the data sourcing and integration.</li> <li>• Briefly describe how the supplemental data aligns with your primary dataset.</li> </ul> </li> <li>4. <b>Database Design &amp; Data Transformation</b> (1 slide): <ul style="list-style-type: none"> <li>• Provide a concise visualization or description of the Entity-Relational-Model.</li> <li>• Discuss the key challenges faced during data transformation, such as inconsistencies, and how they were addressed.</li> </ul> </li> <li>5. <b>SQL Insights &amp; Advanced Analysis</b> (1 slide):</li> </ol>

	<ul style="list-style-type: none"> <li>• Showcase one or two standout insights derived from advanced SQL queries.</li> <li>• Highlight any specific analyzes that were particularly challenging or revealing.</li> </ul> <p>6. <b>Visualization &amp; Key Insights</b> (2 slides):</p> <ul style="list-style-type: none"> <li>• Show some of the primary visualizations developed.</li> <li>• Discuss the major insights derived from the visualizations.</li> </ul> <p>7. <b>Conclusions &amp; Business Implications</b> (1-2 slides):</p> <ul style="list-style-type: none"> <li>• Reflect on the original business problem and whether your findings support or refute your initial hypotheses.</li> <li>• Discuss the potential business implications of your findings.</li> </ul> <p>8. <b>Major Obstacle</b> (1 slide):</p> <ul style="list-style-type: none"> <li>• Discuss the biggest obstacle or mistake you encountered during this project.</li> <li>• Share what you learned from it and how it influenced your project.</li> </ul> <p>9. <b>Closing Slide</b> (1 slide):</p> <ul style="list-style-type: none"> <li>• Reiterate the project title.</li> <li>• Team members' names.</li> <li>• A "Thank You" note or any final thoughts.</li> </ul> <p><b>Total:</b> Approximately 10 slides. &lt;&lt;&lt;&lt;&lt;&lt; HEAD</p>
--	--


## Weekly Schedule

### Monday

✓	<b>Data selection</b> → find datasets ether csv files or through APIs or web scraping
✓	<b>Business framing:</b> what business challenge are we addressing?

	<p>→ <b>Craft hypotheses</b></p> <p><b>Business frame: Eco-tourism</b>  We are tourism agency* looking to expand into ecotourism and therefore need to find suitable locations to start building eco-tour packages</p> <p><i>*business name needed at some point</i></p> <p><b>Hypothesis</b></p> <ul style="list-style-type: none"> <li>• We suspect countries with reforestation would be good for eco-tourists, so which countries worldwide are experiencing reforestation?</li> <li>• Which of the most-visited countries are relevant for landuse in term of forests? → Which of the reforested countries are among the most-visited countries?</li> <li>• <del>Countries having similar trends in forest evolution are geographically close.</del></li> <li>• <del>Countries with higher population growth tend to experience higher rates of deforestation.</del></li> </ul>
✓	Examine the data selected
✓	Project planning: Working doc & trello <a href="https://trello.com/b/zdoVTrRs/week4group4">https://trello.com/b/zdoVTrRs/week4group4</a>

## Tuesday

✓	<b>What data do we need from the datasets we sourced?</b> → identify what we need to keep vs drop → identify potential issues (ie. mismatching names) & flag clean-up tasks
✓	<b>Database Design &amp; Creation</b> → sketch an E-R model → <b>We need 3 tables [!!!]</b> → what are the primary keys, foreign keys, table relationships → what are the proper data types per column
✓	<b>Transform the Sourced Datasets</b> → do necessary data wrangling to clean the datasets (what we flagged earlier) → MAKE NEW DATASET/ FILES that we will use to feed our SQL tables as we have outlined in our E-R model
✓	<b>SQL Database Creation</b> → translate the E-R model into a functional database
	<b>Load the SQL Database</b>

→ import the sanitised data
-----------------------------

## SOURCED DATASET BREAKDOWN ↓↓↓

### Food and Agriculture Org. of the United Nations Dataset: Forest in hectares

<https://data.apps.fao.org/catalog/dataset/forest-area-1990-2020-1000-ha/resource/3cc5000d-184c-4ebe-8e54-b71910111f12>

- Q: What columns are we keeping?
  - Unnamed: 0 (or NaN)
  - Keep the columns for the years we are looking at
  - [1990, 2000, 2010, 2015, 2016, 2017, 2018, 2019, 2020]
- Q: What columns/rows will need cleaning?
  - Column name Unnamed: 0 (or NaN) needs to be renamed **Renamed 'Country'**
  - Some country names have '(Desk study)' as part of the name, that will need to be dropped from the name & *potentially* add a column identifying the desk study countries (I think desk study might indicate iffy data) **Column added (True or False values) before removing the string from column Country**
  - Drop the rows with non-country names (ie. '© FRA 2021' & '2021-01-05') **:26 rows but may need renaming so kept (for example 'United States of America') fait 2 datasets**
- Any other notes:
  - The first line of the dataset (header) can be dropped as it only shows the unit of measurement, which is 'Forest (1000 ha)' → the true header row starts at row 2 and contains 10 columns, of which the first is for country names and the rest are years
  - Country names ONLY, we are missing country codes
  - We have 239 rows

### Kaggle Dataset: Deforestation in %

<https://www.kaggle.com/datasets/konradb/deforestation-dataset>

- Q: What columns are we keeping?
  - All columns
- Q: What columns will need cleaning?
  - Forest\_2000
  - Forest\_2020
  - Trend (the change)
- Any other notes:
  - Empty cells in the "trend" column
  - Zero (0) values in some rows.
  - Column name for country code needs to be modified (From 'iso3c' to meaningful name like 'country\_code')
  - We need country names

- Specify the unit of the values.

### Tourism Dataset: Top Visited Countries

<https://worldpopulationreview.com/country-rankings/most-visited-countries>

- Q: What columns are we keeping?
  - Country Code, “years?”
  - DROP - Country Name, Indicator Name, Indicator Code, Unnamed: 68.
- Q: What columns will need cleaning?
  - A:
- Any other notes:

### Country Codes Dataset



[https://gist.github.com/tadast/8827699#file-countries\\_codes\\_and\\_coordinates-csv](https://gist.github.com/tadast/8827699#file-countries_codes_and_coordinates-csv)

- Q: What columns are we keeping?
  - A: all but Alpha-2 code and numeric code **DONE**
- Q: What columns will need cleaning?
  - A: possibly column country to fit other dataset (remove blank?) **NO NEED FOUND**
- Any other notes:

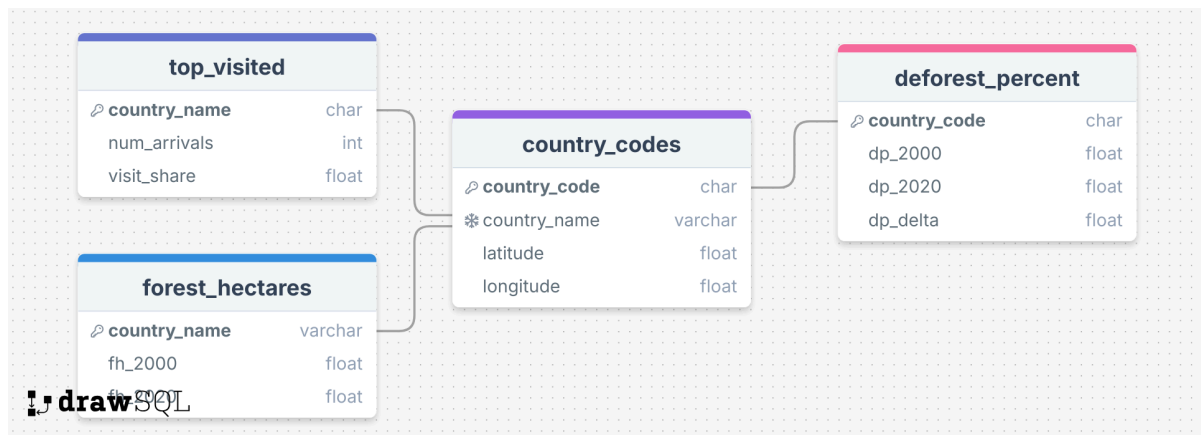
## Wednesday

*Image of the ER model can be found below the table*

	<b>Make final changes to the sourced datasets</b> → column names from the sourced datasets must match <u>100</u> to the column names in the SQL database → number of columns have to match <u>100</u> as well
✓	<b>SQL DB Creation: Team-Wide</b> → since we are not sharing a server, we each have to create a db in our SQL servers → Sasha has written the db+table creation code, but wanted to go over it with team so we can double-check & adjust together (eg. country code is 3 Char long correct?)   code will be shared w/ team & each will have to run it indiv.
✓	<b>Load the SQL Database</b> → import the sanitised data
👁️	<b>SQL Queries &amp; Analysis</b> → we need at least 5 SQL queries using fundamental clauses

	Use functionalities like <b>JOIN</b> , <b>GROUP BY</b> , <b>ORDER BY</b> , <b>CASE</b> and subqueries
	Summarise the data using <b>mean</b> , <b>max</b> , <b>min</b> , <b>std</b> and more

## E-R Model INFO



### Entity

- Central entity is country\_codes

### Attributes

- The supporting columns to the primary key




### Relationship

- We have a series of 1:1 relationships

### Quirks, in our case...

- we have a unique example where our primary keys are foreign keys
- In Country Codes tbl we had to assign a unique key in addition to the primary key
- our tables are not stacked for ease of analysis & given that these are not transactional tables (that will not be updated) → but you could stack the tables further by say country\_code, years, some\_quantity, which would turn our 1:1 relationships to 1:N (cause country\_codes/ name would no longer be unique in the non-central tables)

## Thursday

	<b>SQL Queries &amp; Analysis @ Sasha &amp; @ Joel</b> → Use functionalities like <b>JOIN</b> , <b>GROUP BY</b> , <b>ORDER BY</b> , <b>CASE</b> and subqueries & summarise the data using <b>mean</b> , <b>max</b> , <b>min</b> , <b>std</b> and more → Top 10 most visited countries in 2020  → Top 10 countries with the highest percent of forest 
---	--



	→ Top 10 countries with the most forest per total area ✓ → we need to put all SQL queries into ONE single SQL file ✓
✓	<b>Data Visualisations</b> → use Python libraries to create compelling visual representations of your findings → we need at least 2 graphs/plots using Matplotlib or Seaborn ✓ → Forest Map @ Tiago → Most-visited countries trend @ Tania
✓	<b>Report Compilation and Finalisation @ Tiago</b> → we need to put all python code into ONE single report → report should be clean & aesthetically pleasing (aka use markdown code blocks) & MUST BE DISTINCT from the code used for cleaning/processing the data (all that can be in separate notebooks)  Remember, when using Python the code must be put in wrapped functions  → How much more forest do these top 10 countries have over the avg?
✓	<b>Presentation @ Sasha &amp; @ Joel</b> <a href="#">LINK HERE</a> → Design a compelling presentation that encapsulates the essence of your project, challenges, insights, and outcomes
✓	<b>Prepare a README.md File @ Tania</b> → Anyone that reads the README should be able to understand the project without having to look through all of the files.

Friday

PRESENTATION DAY ✓