Code.Hub

Univators: Skilling Future Digital Innovators

Project Assignment

Oceanographic data analysis, visualization, and reporting

May 2023

Code.Hub

Abstract

In this project oceanographic data from CalCOFI database, publicly available in open formats over the internet, are used to perform Business Intelligence techniques and produce a dashboard offering added value to the professional and researcher.

The data are in raw form and should be explored to obtain insight and comprehension of the available measurements, their value, and the appropriate summaries and groupings that can be calculated. Finally, reports and technical dashboards will be created to visualize the data, highlight results, demonstrate possible correlations, and provide predictions.

Contents

1 Background

CalCOFI, California Cooperative Oceanic Fisheries Investigations, conducts quarterly cruises off southern & central California to collect a suite of environmental and marine ecosystem data. The database contains oceanographic data measured from seawater samples collected at CalCOFI stations.

Prior to 1993, CalCOFI collected seawater samples using Niskin, Nansen, and "Wally" (in-house design by Walt Bryant & George Anderson) bottles with reversing thermometers. In 1990, CalCOFI began using a CTD-Rosette (pictured here underwater), and starting Aug 1993 (9308NH) CTD-Rosette sampling became the primary method of seawater collection.

Two CSV data files are provided in the Kaggle repository. The Cast table contains the metadata. This table includes the date, time, latitude, longitude, weather, etc. for each CTD cast ever completed on a CalCOFI cruise. Each row is a unique cast, numbered sequentially/indexed by the "Cst_Cnt" column.

The Bottle table contains oceanographic data. This table includes oceanographic measurements for each bottle/sampling depth ever completed on a CalCOFI cruise. There are additional data codes and precision columns describing the quality of each oceanographic measurement. Each row is a unique bottle/sampling depth, numbered sequentially/indexed by the "Btl_Cnt" column. This table is much larger than the Cast table because each cast can have up to 24 bottles/sampling depths. The Bottle table also has a "Cst_Cnt" column, which can be used as a unifier between the two tables.
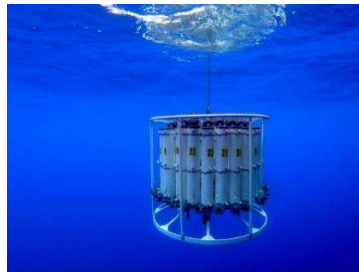
CTD-Rosette samplers, Measuring device.

2 Aim and objectives

The aim of the project is to use the provided hydrographic survey data and apply the Business Intelligence methodologies using visual tools and minimum programming code.

Specifically, the project objectives are:

- load data from public repositories to analyze locally,
- get an insight into the data,
- perform data sanitization techniques,

Code.Hub

- create basic diagrams,
- group and summarize data,
- perform correlations,
- present predictions
- implement user-friendly, informative, and intuitive dashboards.

3   Deliverables

To organize the project, it has been divided into four milestones. The pipeline of the work is proposed to follow the below analysis.

Milestone A

a. Perform Exploratory Data Analysis and get an insight into the data.
b. Identify appropriate data summarizing and grouping.
c. Propose possible correlations.
d. Examine the time series to provide trends and predictions.
e. Design a suitable Star Schema based on the data given from the available repository to identify the dimensions and measurements for further analysis.

It is expected to deliver:

A PowerPoint document with the tables, diagrams, and bullets with the main points and assumptions.

Milestone B

a. Load the data to Power BI.
b. Perform transformations to sanitize data.
c. Implement the data model that has been designed.
d. Produce basic visualization of the data.

It is expected to deliver:

A Power BI file with the data, the model, and the needed transformations.

Milestone C

a. Create reports of the provided time series data in Power BI.
b. Design summarized reports with filters, user interaction, and adaptive behavior.
c. Provide correlations, KPIs, and estimations that apply. (nice-to-have)

It is expected to deliver:

The previous *PBIX* file with the requested reports.

Milestone D

   a. Prepare a brief presentation summarizing your work, for use during the final presentation of your project.
   b. Also, revise any assumptions and decisions you made as well as their justifications, and provide points of future improvement or expansion.

You are expected to deliver:

The final version of the PowerPoint file of the presentation.

4   Group formation

The participants form groups of six, organize their work, prepare all the deliverables, and present their work on the scheduled date. This is a simulation of a real-life case study with project management and delivery in a modern consulting company.

5   Submission details

You will need to submit the work that you have done for the project, namely the presentation deck and the Power BI files. Submission will involve uploading the file in Google Drive. Specific links will be provided to each team.

6   Achievement of Learning Outcomes

Demonstrating the two project's criteria below serve as evidence of successfully achieving the program's learning objectives:

   ● Understanding of the business problem and ability to provide solutions (correct schemas, useful measures or calculated columns, useful, organized, and readable reports).
   ● Technical accuracy and performance (correct use of tools, absence of logical errors, use of best practices for performance and stability).

Code.Hub

7 Important dates

- Saturday May 20: Hackathon - boosting project development
- Friday May 26: Submission of project files in Google Drive
- Saturday May 27: Final presentations

8 Links

https://en.wikipedia.org/wiki/Rosette_sampler

https://calcofi.org/about/

https://calcofi.org/data/oceanographic-data/bottle-database/

https://www.kaggle.com/datasets/sohier/calcofi

https://learn.microsoft.com/en-us/training/powerplatform/power-bi