Google Data Analytics Capstone Project

Erik Hub

Scenario

I'm a junior data analyst working for a business intelligence consultant. I've been at the job for six months, and my boss feels I'm ready for more responsibility. He has asked me to lead a project for a brand new client — this will involve everything from defining the business task all the way through presenting my data-driven recommendations. I will choose the topic, ask the right questions, identify a fresh dataset and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

Company and Objectives

The company itself will be a forest management company, called Growing-Gnomes and are seeking to accomplish the following

- Identify the most suitable tree species for afforestation based on the study locations and the target location and climate.
- Analyze the growth rates and survival rates of different tree species in various forest plots to determine which species are more likely to succeed in afforestation projects.
- Identify any trends or patterns in the forestry data that could inform the development of new afforestation strategies or techniques.
- Create visualizations and presentations to communicate your findings to stakeholders, such as government agencies, environmental organizations, or forestry companies.

Ask

1. What topic are you exploring?

   The topic will be forestry and the dataset will be a joint of several different datasets related to forest health and density in europe.

2. What is the problem you are trying to solve?

   Utilize the available datasets to gain knowledge and insights for an upcoming afforestation project, especially focusing in

   - Species selection
   - Growth and survival rates
   - Location selection
   - Planting techniques
   - Monitoring and maintenance

3. What metrics will you use to measure your data to achieve your objective?

   - Tree Species
   - Diameter at Breast Height (DBH)
   - Location

4. Who are the stakeholders?

   Both the BI Consultant and Growing-Gnomes, government and environmental organizations involved in the upcoming afforestation project.

5. Who is your audience?

   Anyone interested in the results of the project, both stakeholders and the general public.

6. How can your insights help your client make decisions?

   Providing valuable information and insights, such as identifying the best suitable species planting techniques and understanding survival rates under different conditions.

7. Where will you obtain that data?

   - Figshare
   - EU Open Data

Prepare

1. Where is your data located?

   Online, both in Figshare and the EU Open Data portal.

2. How is the data organized?

   Data is spread through 5 datasets, which are:

   - EUForestgenus: Tree Genus by Country
   - EuForestspecies: Tree Species by Country, Diameter at breast height, Forest Floor, Basal Area and Estimated Elevation
   - IcelandTestPlot: Test plot for afforestation project
   - Parcels: More detailed info about the TestPlot
   - Romanian Surface Sample: Control subject and example of growth

3. Are there issues with bias or credibility in this data? Does your data ROCCC?

   The data is reliable, free of any bias, and has been collected by government agencies.

4. How are you addressing licensing, privacy, security, and accessibility?

   The data is publicly available, has no personal information of any kind, with the following licenses:
   CC0 1.0 Universal (CC0 1.0) Public Domain Dedication
   Attribution 4.0 International (CC BY 4.0)

5.

6. How did you verify the data's integrity?

   Data Exploration, Validation and Profiling in Jupyter.

7. How does it help you answer your question?

   Datasets provide information and insights.

8. Are there any problems with the data?

   Missing data in the EUForestspecies dataset and Data not explained in the IcelandTestPlot dataset.

Process

1. What tools are you choosing and why?

   Jupyter and Pandas from Python, more flexibility and fastest response and cell execution order.

2. Have you ensured your data's integrity?

   Yes, Data Exploration, Validation and Profiling

3. What steps have you taken to ensure that your data is clean?

   Addressed in notebook

4. How can you verify that your data is clean and ready to analyze?

   Addressed in notebook

5. Have you documented your cleaning process so you can review and share those results?

   Proof in the Jupyter Notebook of the repository

Analyze

1. How should you organize your data to perform analysis on it?

   Data Cleaned and Organized in the the Jupyter Notebook of the repository, only species that can survive subarctic weather will be used in the study.

2. Has your data been properly formatted?

   Yes, cleaned and formatted, .CSV files and Jupyter notebooks will be used.

3. What surprises did you discover in the data?

   The very few amount of tree species that can survive subarctic weather, and the relation between survivability and location (either inside an already populated area or an expansion range)

4. What trends or relationships did you find in the data?

   Subplot density, distribution of DBH (diameter at breast height) and survival species composition, and most successful subplot.

5. How will these insights help answer your business questions?

Share

1. Were you able to answer the business question?

   Yes, the insights gained from the analysis related to growth and survivability in subarctic climate will be useful at the new afforestation project.

2. What story does your data tell?

   The defining aspects of success and failure of new tree plantations in subarctic environments.

3. How do your findings relate to your original question?

   They answer the original question.

4. Who is your audience? What is the best way to communicate with them?

   As stated before, the audience are both the BI Consultant and Growing-Gnomes, government and environmental organizations involved in the upcoming afforestation project.
   The best way to communicate with them will be through an informative document (attached in repository), the graphs in the analysis notebook and a Tableau presentation.

5. Can data visualization help you share your findings?

   Yes, and will be the main way to communicate findings.

6. Is your presentation accessible to your audience?

   Yes, will be published in both GitHub repository and Tableau.

Act

1. What is your final conclusion based on your analysis?

   Final conclusion will be on the Growing-Gnomes Report

2. How could your team and business apply your insights?

   The insights need to be communicated to Growing-Gnomes and the other entities.

3. What next steps would you or your stakeholders take based on your findings?

   Start obtaining the correct species of trees for the afforestation project and preparing the planting areas according to the findings.

4. Is there additional data you could use to expand on your findings?

   Yes, additional data is always welcome, any kind of data related to soil and climate conditions, plus a constant follow-up of the project will be needed.