Logo, company name

Description automatically generated

**Skills test for Data Scientist**

Thank you for making time to undertake this skills test as a selection step for the position of Scientist: Data Science.

**General Instructions**

1. Complete the tasks described below using your own computer and submit the results of the test as a compressed folder via email (to k.bassier@sanbi.org.za and copying [s.khatieb@sanbi.org.za](mailto:s.khatieb@sanbi.org.za) and [j.currie@sanbi.org.za](mailto:j.currie@sanbi.org.za)). Attach the compressed folder to the email or via a shared link to an online storage platform. Time-stamps of files will be examined to assess adherence to the skills test time-limit.
2. Ensure that the answers you submit are clearly identifiable as yours. The folder containing your results should be labelled with your NAME\_SURNAME, and place all skills test items into that folder or appropriate subfolders, using relevant file and folder names. Even if you provide a functional script to answer the question, it is important to save the relevant outputs to files that are included in your submission.
3. This test has three questions, each containing more than one task and all need to be answered to the best of your ability. Be sure to read the questions carefully.
4. You have 90 minutes (1.5 hours) to complete all tasks. Score allocations for the questions are indicative as to how much they contribute to the overall test score and can be used as a guideline for how much time you should spend on each question.
5. You will be assessed predominantly on your explanation of the process, with a priority on the important steps, the logical and efficient sequence of these steps, and listing important considerations or pitfalls that need to be navigated. Providing the solution (even if correct) with minimal or no explanation of how you achieved it, will score very low. Demonstrating your understanding by providing a clear and logical (but concise) explanation of the steps that are needed, will score highly, even if you do not manage to provide the correct solution in the short time available. Include in your answer the software tools or packages and functions you would employ. A scripted workflow (programming code) will score extra points, but it is imperative that you document each step with comments that help the assessor understand your code, even if they are not familiar with the coding language.
6. The tasks should be completed entirely on your own without assistance from anyone else If we determine you have used AI to assist in writing the application, you will be disqualified.

**TASK 1 (18 marks) – Calculate the Protection Level for each of the ecosystem types**

**Background**

Ecosystem protection level is an indicator that tracks how well represented an ecosystem type is in the protected area network. It has been used as headline indicator in national reporting in South Africa since 2005 (Reyers et al. 2007). It is a relatively simple indicator, computed by combining the map of ecosystem types and the map of protected areas. The Protection Level categories of Ecosystem types are assigned based on the proportion of the biodiversity target achieved within one or more protected areas (Government of South Africa, 2008).

Ecosystems are classified into the categories Not protected, Poorly Protected, Moderately Protected and Well Protected according to the percentage of their biodiversity target that is protected, according to the following thresholds:

A green and grey bar with black text

Description automatically generated

**Instruction:**

Task 1.1: Using the Ecosystem Type (EcosystemTypes.shp) and Protected Area layers (ProtectedAreas.shp), calculate the protection level category for each ecosystem type.

The protected area layer includes areas that have been degazetted. (These are areas that have previously been declared as a protected area but no longer contribute to the conservation estate.) Degazetted Protected Areas are flagged in the NBA\_PA\_Not column.)

The name of each ecosystem type is displayed in the Name\_18 field. The biodiversity target (expressed as a percentage) for each ecosystem type is displayed in the CNSRV\_TRGT field.

Please clearly document your workflow. You may either save your workflow as a GIS model, or otherwise document each step taken in a script or a text file. Save your outputs of.csv file.

Task 1.2: Using your results from 1.1., what type of summary statistics would you use and how would you present them to policymakers, to encourage informed decisions? Include your reasoning. **(5 marks)**

**TASK 2 (18 marks) – Adding environmental data to survey locations**

**Background**

In the investigation of biodiversity pattern, we frequently need to combine environmental measures with biodiversity samples in statistical modelling frameworks. Sometimes the environmental variables of interest were not measured at the same time and locations as the samples were taken, but they can be retrospectively added from other spatial data layers.

**Instruction:**

Within the ‘Task2’ folder, there are two files, namely 1) a spreadsheet that contains the locations (unprojected latitude and longitude coordinates), dates and names of many genera that were encountered in demersal trawling surveys; and 2) a spatial layer of gridded average temperature of ‘bottom waters’ near the seafloor. Note that ‘NA’ indicates missing data.

Task 2.1: Explain/demonstrate the steps of how you would select the 10 most commonly encountered genera from the first file and add the average bottom water temperature values from the second file to the selected survey records. The expected outcome would be a spreadsheet (.csv).

Task 2.2: Describe/demonstrate how you would turn the resultant data into plots that effectively illustrate the average temperature profile for each genus. If you produce plots, save them as image files or PDF document(s) with appropriate names.

Task 2.3 (for bonus points, if you have completed 2.1 and 2.2 and have time remaining): If you had daily temperature hindcasts, instead of the single average temperature file, explain how you would perform Task 2.1, but adding the daily temperature records that match the date and location of the catch survey records (instead of the average temperature).

**TASK 3 (10 marks) Organised workflows and tracking/documenting versions**

**Background**

In the transition to improved workflows, data scientists frequently have to deal with legacy projects and workflows that may have grown ‘organically’ with little planning of their structure or consideration of methods to track the evolution of inputs and outputs. This question is about developing logical structure and efficiency in the curation of a large set of spatial files. There is no single correct solution, but we would like to see description of a realistic, efficient and well-organised solution that will aid future users and managers of these files or projects.

Your supervisor hands you a hard drive from a former staff member, who was in charge of multiple GIS projects, but was not organised and did not apply data science principles or tools. The hard drive contains many (about 100) spatial (both raster and vector) files, in a variety of file formats. The thematic contents of the spatial files range geological, climate, elevation and biodiversity data. Most of these files will not be further developed/updated, however it is important to document them for the sake of better understanding and transparency of previous projects and their outputs. A handful of the files, say 10 of them, will be actively developed or improved upon in future.

**Instruction:**

The following two tasks require a brief description of the system/software/file types used and the fields that you would include. You may demonstrate by setting up some example files, but we do not expect you to implement a solution. Concise description thereof will suffice for maximum points.

Task 3.1: Describe a simple inventory system that you could develop so that users of these files could ‘at a glance’ examine and/or update pertinent information about the spatial files, Include examples of the generic properties or metadata information that you think are most pertinent and useful to include in such an inventory.

Task 3.2: Describe what system you would recommend (or put in place) so that the few spatial files that will be developed/improved on in future have: a) a well documented history of their changes, b) are accessible in a centralised location by multiple collaborators that work on them, and c) are adequately and regularly backed up to ensure they will not be lost due to a hardware or software failure.