Grading Rubric Data Scientist

July 2024 additions / changes are highlighted in yellow

<table>
<tr><th>Category</th><th>Competency</th><th>Sufficient</th><th>Insufficient</th></tr>
<tr><td>Data Validation</td><td>Assess data quality and perform validation tasks</td><td>Has validated all variables and where necessary has performed cleaning tasks to result in analysis-ready data.</td><td>Has not conducted all the required checks and/or has not cleaned the data. May have removed data rather than performed cleaning tasks.</td></tr>
<tr><td colspan="4">There should be either a written description of the data validation performed or a mix of code and written description.

For writing only: there must be a description for every column (8 or 9 columns) in the data.

For code and text: the **text should be the primary grading method**. Code can be taken into account where text has not been included for a column. The code must show that they have checked the distinct values (for categorical) or range (for continuous) for the columns. **They must have highlighted in text the problems that were encountered**. Only checking the data type is not sufficient.

All columns must be checked/described.

They must have resolved the issues that were included in the data as described in the project for associate or with their own justification for professional.

Code is not required for any role or level.

Information article provided to candidates</td></tr>
</table>

Data Visualization	Create data visualizations in R or Python to demonstrate the characteristics of data and represent the relationships between features	Has created at least two different visualizations of single variables (e.g. histogram, bar chart, single boxplot) Has created at least one visualization including two or more variables (e.g. scatterplot, filled barchart, multiple boxplots) Has used visualizations that support the findings being presented Data visualizations displayed in the presentation are clear and readable onscreen (e.g. components such as axis labels are large enough to be read onscreen, points and lines can be distinguished, key is included if needed, any chartjunk / clutter does not inhibit accurately reading the visualization)	Has used the same visualization throughout. Has not included graphics to represent single variables and relationships. Has not used visualizations that support the findings being presented. Visualizations are unreadable: too small, missing information (e.g. keys, axis labels), or unclear in other ways
There must be **two visualizations of single variables** (any graphic that includes only one variable i.e. only a variable defined on one axis). There must be one visualization that includes multiple variables (i.e. both an x and y axis variable). This plot can include any number of variables. There must be **at least two different types of visualization** i.e. they cannot include 3 bar charts, but two bar charts and a			

scatter plot is ok. The questions given will usually direct to a categorical and continuous plot as well as a relationship plot but this is not a requirement. The graphics included **should support their analysis description**. For instance, if they state that the data shows that sales increase over time, there should be a plot that shows sales increasing over time. The graphics can appear anywhere in the report. **Code is not required.** The readability / clarity criterion is for graphics presented; it is very reasonable to create graphics during analysis that are not meant to be seen and can therefore be too small or incomplete We **don't** expect candidates to demonstrate - understanding of accessibility features e.g. appropriate choice of color - refinements of layout e.g. order of bins - logical color choice e.g. warmer temperatures as red, colder as blue Information article provided to candidates.			
Model Fitting	Implement standard modeling approaches for supervised or unsupervised learning problems	Correctly identified the type of problem (regression, classification or clustering) Has selected and fitted a model for that problem to be used as a baseline. Has selected and fitted a comparison model for the problem that they were provided.	Has incorrectly identified the type of problem. Has not fitted a baseline model or has used a model for the wrong type of problem. Has not fitted a comparison model or has used a model for the wrong type of problem.

We expect to see **evidence that they have identified the type of problem** they will be working on. This will most likely be as a statement of the type of machine learning problem. The type of machine learning problem should be stated as one of regression, classification, or clustering. If they have not stated the problem type but they have fitted the right type of model, we will consider that evidence that they have identified the right type of problem. **Code needs to be included in Python or R**. The code should show that they have fitted the two models. They must **fit at least two models**. Fitting two models of the same type with different variables is permitted, e.g. linear regression with one variable and linear regression with multiple variables. Models fitted for the purpose of parameter tuning are not permitted. Grading should not reflect any other aspects of the model fitting such as approaches to splitting data, hyperparameter tuning etc. Information article provided to candidates.			
Model Evaluation	Use suitable methods to assess the performance of a model	Compared the performance of the two models/approaches using any method appropriate to the type of problem. Has described what the model comparison shows about the selected approaches.	Has selected a method not suitable for the type of problem. Has not described what the results show about the selected approaches.
We expect to **see code in Python or R** that demonstrates they have applied a comparison method to the model output. This could be a single metric or a graphic. It must be appropriate to the type of problem they have implemented.			

They **must include a description of what their method tells them**. A short sentence that one model performs better than the other is sufficient. This criteria should be graded in relation to the model the candidate has used. If they identify a regression problem as a classification problem, and they use metrics related to the classification problem, they should be graded as pass in model evaluation. There is no requirement to link this evaluation back to the problem.			
Data Communication	Employ multiple tactics (written and verbal) to communicate to business leaders	For each analysis step, has provided a written explanation of their findings and/or reasoning for selecting approaches Delivers a presentation addressing the business goals, outcomes and recommendations Delivers a presentation with a recognisable narrative (i.e. a logical progression between points) that can be followed without significant effort, and is supported by the findings of the data analysis	Has not provided a written summary for each step Has not delivered a presentation Delivers a presentation with no apparent narrative or no connection with the findings of the data analysis
They need to have **completed a report** with written text summaries throughout. In the report, steps in data cleaning are clearly understandable and have been presented through a short piece of text at minimum (e.g. 3 sentences)			

<table>
<tr><td colspan="4">In the report, findings are presented in short piece of text at minimum (e.g. a few sentences)
Data Scientist: The report must include all code for model development and evaluation
They must also have **delivered a presentation**. There is no requirement for slides or other visual aids. The presentation must include the business goal, their approach to the problem, what they found and their recommendations.</td></tr>
<tr><td>Business Focus</td><td>Make recommendations for analytic approaches based on business goal</td><td>Has described at least one of the business goals of the project
Has explained how their work has addressed the business problem
Has provided at least one business recommendation for future action to be taken by the company based on the outcome of the work done.</td><td>Has not identified any business goals
Has not explained how their work has addressed the business problem
Has not provided any recommendations for future actions for the company</td></tr>
<tr><td colspan="4">We expect to see/hear a **description of the business goal, how they addressed the problem** and at least one recommendation for next steps.
This criteria can be **graded from the report or presentation** (note all these things need to be included in the presentation to pass the communication criteria).
Repeating the information exactly as we provide is permitted. Repeating the customer questions we provide is permitted.
The recommendation can relate to either something the customer does or the individuals doing the work.
The recommendation should be an explicit statement on the best course of action; the reader should not need to infer this from the results of the analysis</td></tr>
</table>

Business Metrics	Judge performance of analytic results against relevant business criteria	Has defined a KPI to compare model performance to business criteria in the problem Has compared the performance of the two models/approaches using the defined KPI	Has not identified a KPI to compare the model performance to the business problem Has not compared the performance of the two approaches using the defined KPI
They should **state a way to show how their model performs** against the business goal. This could be a metric or graphic. They should **show how both of their models perform** using this metric. This should be graded independently of the model fitting and model evaluation criteria. The report must specify why the KPI was chosen			