Case Study

General Information

Congratulations on making it to the final stage of the Data Scientist Professional certification!

So far we have assessed a range of the skills required of a data scientist. In this case study we will be assessing your skills in model development and communication. We want to see that you can take a business problem, select appropriate techniques and implement them, and then give a relevant summary of what you have found to the defined audience.

You will be joining a 4 hour session. The session will be led by a member of the DataCamp certification team. There will be up to 8 certification candidates on the call.

At the start we will recap the details of how the session will work. Once we have covered all the information:

- We will put you into a Zoom breakout room on your own. You will be able to talk to the DataCamp examiner through the chat but not to other participants.
- You will need to share your screen so the DataCamp examiner can see your work throughout the session and you will have to leave your webcam on.
- The DataCamp examiner will provide a DataCamp Workspace environment. This workspace will contain the data that is described in the brief in this document. You will be able to use either R or Python.
- You can then start working and perform any analysis that you want.
- You can use any resources that you want to but the work must be done by you.
- You will also need to prepare to present for up to 15 minutes. Your presentation can be in any format that you want to use. Check the project brief for the target audience.
- The analysis and preparation should be done in 90 minutes. After 90 minutes, we will inform you of your position in the order of presentations and approximately how long you will have until then.
- If you are ready to present earlier, you should inform the examiner as directed. They will get to you as soon as possible.
- The examiner may give you feedback and the chance to resolve any issues and present again. This will depend on the number of people in the session and the time available.
- After you have given your presentation, you will be free to leave.
- We will contact you within 3-5 business days to confirm the outcome.

You should already have an invite to the session which includes a Zoom link. If not, let us know as soon as possible. Your final presentation will be recorded for quality assurance. We may also review your code.

Project - Used Car Sales

You have been hired as a data scientist at a used car dealership in the UK. The sales team are having problems with pricing used cars that arrive at the dealership. They would like your help. To start with, they would like you to work with the Toyota specialist to test your idea(s). They have collected some data from other retailers on the price that a range of Toyota cars were listed at. It is known that cars that are more than £1500 above the estimated price will not sell. The sales team wants to know whether you can make predictions within this range.

You will need to present your findings to the Head of Sales, who has no technical data science background.

Data

You will be provided with a data set that has the following columns:

Column Name	**Details**
model	Character, the model of the car, 18 possible values
year	Numeric, year of manufacture from 1998 to 2020
price	Numeric, listed value of the car in GBP
transmission	Character, one of "Manual", "Automatic", "Semi-Auto" or "Other"
mileage	Numeric, listed milage of the car at time of sale
fuelType	Character, one of "Petrol", "Hybrid", "Diesel" or "Other"
tax	Numeric, road tax in GBP. Calculated based on CO2 emissions or a fixed price depending on the age of the car.
mpg	Numeric, miles per gallon as reported by manufacturer
engineSize	Numeric, listed engine size, one of 16 possible values

Grading

The case study will test your skills in modelling, and communicating your findings. You are not expected to fit a perfect model. You will need to show that you have chosen appropriate methods for the problem you are solving. You will have to show that you can communicate at the right level for the target audience.

In order to pass you have to reach "Satisfactory" in all of the criteria in the table below.

	Excellent	**Good**	**Satisfactory**	**Needs Improvement**
Analysis - shown in script/notebook				
Exploratory Data Analysis (EDA)	The EDA included has been refined to reflect the most relevant elements to the problem and elements that may impact further analysis or decisions.	Some refinement of EDA has been considered, with the most relevant elements being reported. Does not always demonstrate a clear connection to the problem.	Large number of graphics and tables created in the EDA, it is not always clear as to the connection to the problem under consideration	There is no evidence that EDA has been conducted
Visualizations/Tables	Graphics and tables reflect good practices (e.g., use of color, when to use a table instead of a plot), have been clearly labelled, and add value.	Consideration has been given as to when to use each graphic or table. Some good practices have been followed (e.g., visualizations are clearly labelled and titled).	Graphics do not follow good practices but are appropriate to the data type. Tables are extensive and not well formated to make them easy to read.	Graphics/tables have been included that are not appropriate to the data or analysis.
Analysis notes	Notes connect any results/graphics/tables to the problem being considered and the approach being taken. It is clear from the notes how each element of the analysis relates	Notes are clear and go beyond restating what is seen in the results, with a clear connection to the analysis and problem.	Notes on results/graphics/tables are basic and simply restate what is shown.	Results, tables, and graphics are created with no additional notes.

	to the overall problem.			
Model fitting	The choice of model has a clear connection to the problem and the author has shown a clear understanding of modelling good practices, selecting a model that is also suitable for the size of the data as well as the type of problem.	Modelling good practices have been demonstrated, the choice of model is suitable for the problem.	A suitable model has been selected and fitted.	Models selected are not appropriate for the data or problem statement.
Evaluation	The model evaluation is clear, appropriate, and connected to the problem.	The model(s) have been evaluated using appropriate techniques and a clear interpretation of the evaluation has been given.	Model evaluation techniques appropriate to the model have been demonstrated.	The model fit has not been evaluated in any way.
Results - shown in either code or presentation				
Outcome	The outcome clearly relates back to the original problem and defined users, making clear how the analysis helps to solve the defined problem and the impact it has.	The outcomes are stated and some attempt is made to connect back to the original problem statement and how the problem has been solved with these results.	Results restate findings earlier in the report (e.g., stating which is the best model, which features are most relevant).	There is no final summary of the results obtained.
Future work	Future actions are not limited to the analysis conducted but also consider the	Future actions consider the analysis and data collection but may not	Future improvements to the analysis conducted are provided.	No consideration has been given to future actions or work.

	end user, how they may be able to make use of the work conducted and any actions they may be able to take to improve the work.	connect to the end user or the original problem statement.		
Communication - shown in presentation				
Motivation	The work is clearly motivated and demonstrates why the work adds value to a specific person/group. The end user is the focus rather than the analysis.	The work is clearly motivated and shows a connection to practical application. The end user may not have been fully considered.	A basic motivation has been stated that shows some connection to the analysis being used in a practical application.	The presentation provides no motivation for why the analysis has been conducted or why the audience may be interested in the outcomes.
Timing	Within the expected timing, appropriate weighting provided to each aspect of the presentation	May have exceeded timing by up to 3 minutes. In general a well balanced presentation with good weighting to each aspect	May have been between 3 and 5 minutes over time. Some aspects are given more focus than necessary	Exceeds timing by more than 5 minutes
Audience	Generally chooses not to use technical terminology, but where absolutely necessary provides simple to understand explanations, often using analogies that can easily followed	Uses minimal technical terminology and as much as possible provides clear explanations to terms used	May include technical details (such as model information, metrics etc) but makes an attempt to explain relevant information	Frequently uses technical terms with little to no non-technical explanation
Organization	Presentation is structured to	The presentation has a logical flow	The presentation includes all of the	Presentation lacks

	clearly tell a story and provides the right information at the right time to progress that story	that typically follows the analysis conducted	relevant information but may occasionally seem out of order or introduce information too soon/too late	organization or structure to convey the appropriate message clearly