T Level Technical Qualification in Digital Business Services

Occupational specialism assessment (OSA)

Data Technician

Task 2 - Distinction

Guide standard exemplification materials

v1.1: Specimen assessment materials
September 2021
603/6902/4

Internal reference: DBS-GSEM-02

T Level Technical Qualification in Digital Business Services
Occupational specialism assessment

Guide standard exemplification materials

Data Technician

Task 2

Contents

Introduction

The material within this document relates to the Data Technician occupational specialism sample assessment. These exemplification materials are designed to give providers and students an indication of what would be expected for the lowest level of attainment required to achieve a pass or distinction grade.

The examiner commentary is provided to detail the judgements examiners will undertake when examining the student work. This is not intended to replace the information within the qualification specification and providers must refer to this for the content.

In task 2, part A, the student must use the provided datasets and join the data into one single dataset, cleaning the data where required and remembering the client's business objectives. Students must also produce a written decision-making log to keep track of their progress and record any decisions made throughout this task. In part B, students must examine the database and add a section to their decision-making log from part A. This section must include, identify, describe and explain several related points.

After each live assessment series, authentic student evidence will be published with examiner commentary across the range of achievement.

Task 2

Time limit and marks available

Maximum time allowed = 10 hours (you can use this time how you want during each session, but task 2 must be completed within this time limit).

(52 marks)

Instructions for students

The client intends to open 2 new stores. The client wants one shop to focus on their high-end range for wealthier customers and another to focus on their budget range for less wealthy customers.

The client has decided to locate their high-end store in the KT postcode area and their budget store in the BS postcode area. They are currently undecided which postcode sector to open their respective stores in and want to use a combination of their in-house data and publicly available data to inform their decision.

Tony Slater has provided you with some internal ecommerce sales data and external data sets.

Part A

The client wishes to open their high-end store in a postcode sector where the average house price is over one million, and their budget store in a postcode sector where the average house price is under £250,000.

Tony has asked you to join the external data into one single clean dataset. Make sure the single dataset has appropriate variables which reflect the client's business objectives, as it will eventually help to create a dashboard for the client.

Once cleaned and validated, you must calculate the average house price per postcode sector from the prices dataset. You should exclude postcode sectors without a significant number of sales. The final dataset should also include any calculations which may help you design a dashboard.

Tony would like you to keep a log of your progress and any decisions you make.

This log must include:

- which variables you consider relevant to the business objectives and why
- errors you have found in the datasets
- ways you have validated the data
- which columns you feel are appropriate to the business objectives and why
- the primary keys for each dataset
- data you have removed and why
- the minimum number of sales you considered significant and why
- any calculations and aggregations you have applied to the data
- how you reformatted the data to be joined to the clients' internal data

Include any code or formulas you used to automate the above tasks.

Part B

For this part of the task, the internal data received from the client has been exported from their relational MySQL database. They plan to upload the single dataset you created in part A to their infrastructure. The database will include the following tables:

- CLIENT_PRODUCT_LIST
- CLIENT_DATA_FINANCE
- CLIENT_DATA_PERSONAL
- CLIENT_DATA_SALES
- your new demographics dataset

In addition to part A, Tony has asked you to write a separate additional section in your log, which must include the following:

- describe the normalisation form of this new database, giving a clear explanation of your reasons
- identify the primary, alternate, and foreign keys in each table – write a sentence for each key describing why you have identified it as such
- explain how you reformatted the data to be joined to the external data
- explain how you manipulated date of birth to a format appropriate to the context
- provide a data validation template for each column in your new table which includes **data types** and **constraints**
- explain how you removed any variables from the internal datasets that is not applicable for your analysis

Include any code or formulas you used to automate the above tasks.

Resources

You will have access to the following resources, plus the original brief:

- task 2 data sets (provided by NCFE)
    - Ages_sctr
    - Client_data_finance
    - Client_data_personal
    - Client_data_sales
    - Client_product_list
    - Number_of_bedrooms
    - Number_of_rooms
    - Prices_housetype_key
    - Prices_part_1
    - Prices_part_2
    - Prices_part_3

- software applications to clean and blend data (Microsoft or Google)
- word processing software (Microsoft or Google)

Note: you will not have access to the internet during this task.

Evidence required for submission to NCFE

- single joined data set
- decision log of processes and steps taken as described in the instructions for both parts A and B

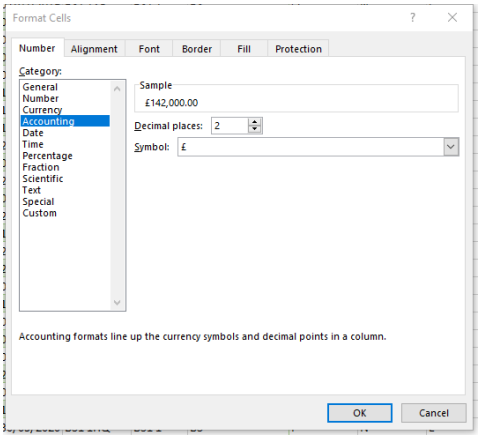
Student evidence

Please also see the following files for student evidence for task 2:

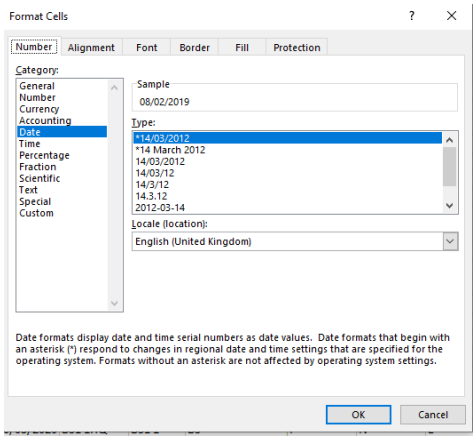
- task 2 evidence

Decision making log

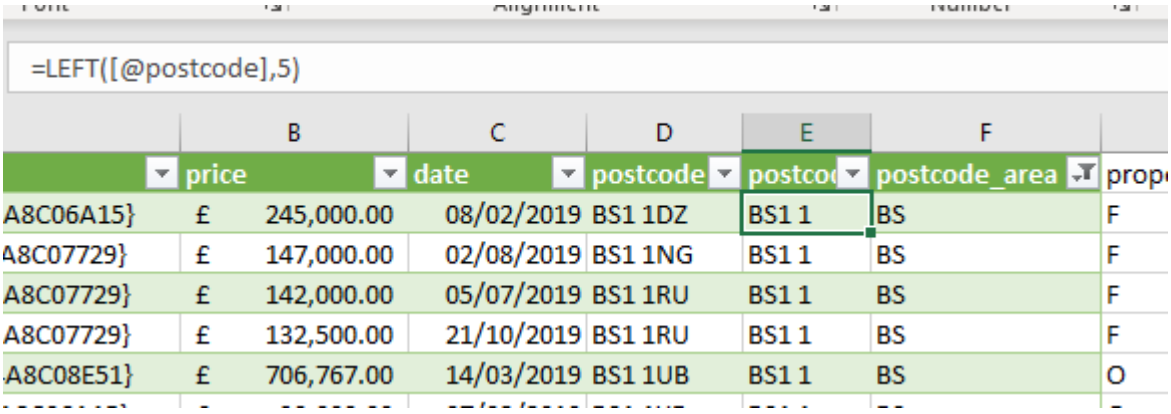
1. Preparation work:
    - each file opened and issues identified
    - plan prepared of edits and mergers
    - software choices were considered including R, SPSS and Python – Excel was chosen on the basis that it is available on the computers of the user and their client and likely to be able to produce accessible material without creating issues, such as additional training needs or installation of software that might then disrupt workflows
    - review of the market and customer base for the business to ensure that the dataset can be prepared to reflect the key demographics and behaviours that the business seeks to target, including:
        - house prices over one million and under £250,000
        - customers between 31 and 55, and under 30
        - sizes of houses to identify larger purchasers
        - postcodes with numbers of new purchases to spot potential for larger sales
2. Cleaning and merging of data:
    - data validation results were established
    - postcode sector columns were confirmed as being a string, with the number of characters used to differentiate postcode sector and area
    - the average price column was formatted as a number and decimal points removed
    - prices_part_1, 2 and 3 were processed using Power Query Editor
    - each column was checked for blanks and those rows removed
    - price column was formatted for currency



- date column was split using space as a delimiter and the time column was deleted as it is not needed

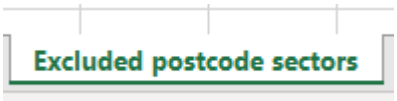


- postcode sector was populated with data using the formula =left([@postcode,x]) where x was the number of characters needed to unconcatenate the postcode column
- prices_part_1 was appended to the end of prices_part_2 and then the columns were filtered so that only the BS and KT postcodes were left



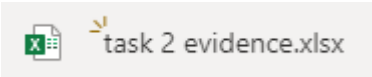
	B	C	D	E	F	
	price	date	postcode	postco	postcode_area	prop
A8C06A15}	£ 245,000.00	08/02/2019	BS1 1DZ	BS1 1	BS	F
A8C07729}	£ 147,000.00	02/08/2019	BS1 1NG	BS1 1	BS	F
A8C07729}	£ 142,000.00	05/07/2019	BS1 1RU	BS1 1	BS	F
A8C07729}	£ 132,500.00	21/10/2019	BS1 1RU	BS1 1	BS	F
A8C08E51}	£ 706,767.00	14/03/2019	BS1 1UB	BS1 1	BS	O

- in order to exclude areas with few sales, a tab was added called excluded postcodes and the 10 areas with the least sales were identified so they could be filtered from consideration



postcode_area		postcode_area	
BS		KT	
Row Labels	Count of price	Row Labels	Count of price
BS1 1	32	KT1 1	28
BS1 2	24	KT1 2	135
BS1 3	33	KT1 4	56
BS1 4	46	KT2 6	150
BS1 8	1	KT2 7	143
BS26	83	KT3 4	147
BS28	63	KT3 5	127
BS29	49	KT8 0	79
BS4 5	91	KT8 1	94
BS8 1	83	KT8 9	113
Grand Total	505	Grand Total	1072

- all data was saved as an XLSX file to allow access to features within this software



- data was loaded as data model due to file size constraints on XLS workbooks
- features in the software were used to check for duplicated values and these were removed
- number_of_bedrooms, number_of_rooms and ages_sctr were merged – the latter file was used as the basis for this blending
- duplicated and empty rows were removed using the same process as the previous dataset – this allows only complete records to be processed as part of the dashboard, ensuring greater validity in the dataset
- new columns were created in the dataset, combining age groups into demographic groups more in line with the goals of the business
- a new column was added for the number of houses with 5 or more bedrooms and was added with the sum of all houses over 5 rooms – this will remove any outliers from the dataset and make the data better fit the client needs, which will also allow:
    - age groups not targeted by the analysis to be omitted from the dashboard

  - more detailed statistics via the dashboards if needed, but will also facilitate an easier summary of key data points
- data was included in line with the requirements of the business objectives – this was done by:
  - removing postcodes with an average house price of over £250,000 but less than one million
  - removing irrelevant columns such as total urban which did not contain any data
- data validation was completed using the built-in tools in Excel and Power Query, ensuring that the dataset was accurately entered, columns contained the correct values and the data was ready for further processing – end user needs were considered as part of this process leading to renaming of columns to make content more intuitive to parse
- data on new house buys has been included and mapped against postcodes to help identify the main areas where larger sales can be made
- under 30 and over 55 age groups have been broken up more clearly in the dataset to help the business better identify key customer groups
- data was filtered to include only the KT and BS postcodes as all other data is not relevant to the clients' needs

postcode_area: BS		postcode_area: KT	
Row Labels	Count of price	Row Labels	Count of price
BS1 1	32	KT1 1	28
BS1 2	24	KT1 2	135
BS1 3	33	KT1 4	56
BS1 4	46	KT2 6	150
BS1 8	1	KT2 7	143
BS26	83	KT3 4	147
BS28	63	KT3 5	127
BS29	49	KT8 0	79
BS4 5	91	KT8 1	94
BS8 1	83	KT8 9	113
Grand Total	505	Grand Total	1072

- the combination of data was based on a series of steps:
  - prices_part_1 has not column names, this data was combined with prices_part_3 which does have the column names
  - prices_part_2 can then be merged into this file using the import data function in excel which gets around the fact that this is a different filetype
  - the postcode sector data was then modified to align in format and presentation with the internal company data provided

Part B

Normalisation of the internal data has taken place following the formal steps:

- reorganising sheets so the primary key is always the leftmost column in each sheet

	A	B	C	D	E	F	G	H	I	J	K
	customer_id	first_name	last_name	gender	date_of_birth	postcode	Postco	Purcha	Product		
2	bcoleg9	Blakelee	Cole	Female	17/12/1979	KT11 3GT	KT11 3	2342	MyHome Curtain Alarm		
3	bcrambht	Bendix	Cramb	Male	26/04/1998	BS31 1QQ	BS31 1	2342	MyHome Curtain Alarm		
4	bdunne91	Bennett	Dunne	Male	30/01/2000	KT12 4PJ	KT12 4	2342	MyHome Curtain Alarm		
5	bedscerl9	Bay	Edscer	Male	30/04/1997	BS34 8VW	BS34 8	2342	MyHome Curtain Alarm		
6	bgaitskellj	Birgitta	Gaitskell	Female	08/10/1999	BS30 5CW	BS30 5	2342	MyHome Curtain Alarm		
7	bgillsonia	Babara	Gillson	Female	20/08/1979	BS34 5DP	BS34 5	2342	MyHome Curtain Alarm		
8	bgouthier9j	Barnie	Gouthier	Male	18/01/1989	BS34 8CW	BS34 8	2342	MyHome Curtain Alarm		
9	bisakovitchbr	Burton	Isakovitch	Male	22/06/1970	BS30 5BX	BS30 5	2342	MyHome Curtain Alarm		
0	blambird8f	Barron	Lambird	Male	29/03/2003	KT16 9JB	KT16 9	2342	MyHome Curtain Alarm		
1	bmackey31	Breena	MacKey	Female	27/08/1997	KT15 1YM	KT15 1	2342	MyHome Curtain Alarm		
2	brenadf6	Baillie	Renad	Male	18/12/1981	BS32 4IV	BS32 4	2342	MyHome Curtain Alarm		

	A	B	C	D	E	F	G	H	I	J
	id	price	date	postcode	postco	postcode_area	property_ty	new_build	estate_type	
94	{85866A64-6C7D-143F-E053-6B04A8C06A15}	£ 245,000.00	08/02/2019	BS1 1DZ	BS1 1	BS	F	N	L	
95	{965B6D91-06E0-95E4-E053-6C04A8C07729}	£ 147,000.00	02/08/2019	BS1 1NG	BS1 1	BS	F	N	L	
96	{965B6D91-07D5-95E4-E053-6C04A8C07729}	£ 142,000.00	05/07/2019	BS1 1RU	BS1 1	BS	F	N	L	
97	{965B6D91-07D6-95E4-E053-6C04A8C07729}	£ 132,500.00	21/10/2019	BS1 1RU	BS1 1	BS	F	N	L	
98	{8CAC1318-F4BC-0253-E053-6B04A8C08E51}	£ 706,767.00	14/03/2019	BS1 1UB	BS1 1	BS	O	N	F	
99	{85866A65-852F-143F-E053-6B04A8C06A15}	£ 30,000.00	07/02/2019	BS1 1XR	BS1 1	BS	O	N	L	

- removing any duplicated values in each table so that content is optimised
- establishing the primary and foreign keys for each table
- formatting of data to ensure that it is appropriate for the task including standardising the formatting of data within columns and assigning the right data type to each column such as currency and date
- postcode sectors were added to the client data table by using the formula =LEFT([@postcode],6) which allowed the postcode to be split to give the relevant data. This can be seen in the following screenshot:

30	blambird8f	Barron	Lambird	Male	29/03/2003	KT16 9JB	KT16 9	2342	MyHome Curtain Alarm

Data not relevant to the task was not included in the final analysis such as the credit card information of customers as this was not relevant.

Average house prices were added to the sheet using a Vlookup formula that pulled them from the average house price tab. The formula used was:

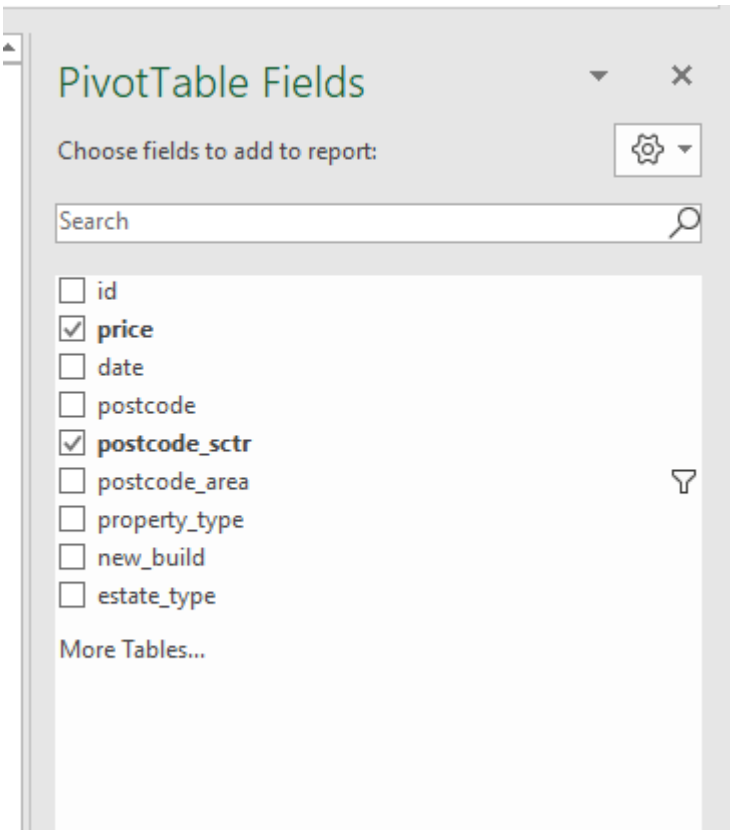
=VLOOKUP([@[Postcode Sector]],'Average Prices Per Sector'!$C$2:$D$70,2)

The formula used a name range for efficiency and in the definition of the cells in which to look up the data an absolute cell reference was added to ensure that the formula could be copied down the column without changing. The use of this formula can be seen in this screen shot:

'67	BS15 3	£ 270,004.89	8	60	256	390	2978	2841	4	454	1136	1027	162	58	8812	1766	1059

The formula is filled down the column and the results of the Vlookup can be seen on the spreadsheet.

The average house prices were calculated using a pivot table because while it would have been possible to analyse the data in that way using a formula such as =average() it was more efficient for the process to be undertaken in a pivot table and because this gave the end user more options for drilling down into the data when using this file later.

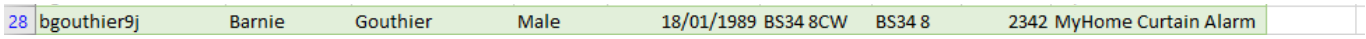


The excluded postcode sectors tab was also created in this way for similar reasons. By adding a slicer to each table, it allows users of the document more flexibility to drill down into the data set later.

Data on the purchases made by specific clients were added to the table using a Vlookup, using the following formula: =VLOOKUP([@[customer_id]],Client_data_sales[#All],2) and =VLOOKUP([@Purchases],Client_product_list[#All],2).

This allows the data to be pulled from the other tables using the primary key.

This can be seen in this screen shot:



28	bgouthier9j	Barnie	Gouthier	Male	18/01/1989	BS34 8CW	BS34 8	2342	MyHome Curtain Alarm

Key choices

There is no transient dependency in the dataset and the database is presented in 3NF.

Primary key – customer ID – this is the correct key for the prices data set and all other data sets. This is common to most tables and will allow the creation of suitable relationships between tables.

Foreign keys – product ID – this gives a unique identifier to the products in the product table.

Transaction ID which gives an identifiable value to each row in the sales table but there is already a primary key for this table.

Postcode sector can also be used as a key and this allows the alignment of the internal and public datasets.

These keys were used to create a blended dataset containing the client data showing the products that the customers bought in each postcode and postcode sector.

This table was created by following these steps:

- import data into powerquery
- use built in tools to remove duplicates and empty rows
- format cells using the following validation rules:
  - customer ID is text
  - first name is text
  - last name is text
  - gender is text
  - date of birth is date
  - postcode is string
  - purchase ID is number
  - product is text
- a postcode sector column was created using the formula:
  - =LEFT([@postcode],6)
  - this allows the data to be analysed alongside the public sector data on houses
- Vlookups were used to add the purchase ID and product descriptions to this table
- data showing as #NA in the lookup columns were filtered to remove them.

Examiner commentary

The student has blended the relevant datasets into a single workable document, documenting the steps taken to clean and validate the data. In doing so, the student has demonstrated a secure grasp of the knowledge and skills needed by a new employee within this sector.

The work contains minimal errors and omissions, and it has a unified structure. The selection of data to blend is good but could be improved upon. The identification of keys is good, although the explanation of why they were chosen is slightly confused, this is a minor error and still allows the work to be placed at the distinction boundary.

The information is structured in a way that is appropriate to the business needs and the data has been cleaned, removing the majority of the empty or duplicate values. The student's log has shown an understanding of the needs

of the business in the given scenario, demonstrating their ability to process and analyse a complex dataset with minimal supervision.

The student has shown a good understanding of data principles and has shown good manipulation skills with relatively few minor errors. Data formats are mostly correct and this allows the student to achieve a mark at the distinction boundary.

Grade descriptors

The performance outcomes form the basis of the overall grading descriptors for pass and distinction grades.

These grading descriptors have been developed to reflect the appropriate level of demand for students of other level 3 qualifications and the threshold competence requirements of the role, and have been validated with employers within the sector to describe achievement appropriate to the role.

Grade	Demonstration of attainment
Pass	The evidence is logical and displays the basic knowledge and skills expected of an employee in this sector in the context of the set brief.
	The student demonstrates theoretical knowledge of the sources, foundations, usage and quality of data that is used for analysis. They are able to carry out routine administrative and analytical tasks using simple datasets.
	The student demonstrates an understanding of data blending techniques and is able to carry out routine data blending tasks.
	The student is able to give a simple explanation of how and why data is analysed by a business. They are able to follow the data process in order to build and test a dataset.
	The student is able to demonstrate understanding of visualisation and communication techniques. They are able to provide evidence of communicating data which is relevant to stated business objectives.
	The student is able to state legal and professional principles that are relevant to the manipulation of data. They are able to carry out routine tasks using data in a way that complies with relevant laws and professional standards.
	The student is able to explain how appropriate sources of information can be selected and evaluated. They are able to search for relevant information and can assess the reliability of the knowledge that they generate.
Distinction	The evidence produced in response to the brief is precise and logical, displaying a secure grasp of the knowledge and skills that would be expected of a new recruit in the industry.
	The student demonstrates a thorough understanding of the sources, foundations, usage and quality of data that is used for analysis. They are able to carry out complex and non-routine administrative and analytical tasks with minimal supervision, using both simple and complex datasets.
	The student demonstrates a secure understanding of a range of data blending techniques and is able to carry out both routine and non-routine data blending tasks competently.
	The student is able to demonstrate a detailed understanding of the reasons why a range of businesses might analyse data. They are able to use their own initiative to follow the data process with minimal supervision in order to build and test a complex dataset in response to a specified business problem.
	The student is able to demonstrate a detailed understanding of a range of visualisation and communication techniques that might be appropriate to a range of organisational needs. They are able to work collaboratively to communicate and visualise data, showing links to business objectives in the materials that they produce.
	The student is able to explain the legal and professional principles that are relevant to a range of

	different data manipulation tasks. They are able to consistently carry out both routine and non-routine tasks in a way that complies with legal requirements and professional standards.
	The student is able to give a detailed explanation of how to select and evaluate a range of different sources of information for a specific task. They are able to search for data that is appropriate to a given task and can corroborate their findings using appropriate methods to evaluate the suitability of data and making appropriate recommendations for improvements in the collation of data for future tasks.

Document information

The T Level Technical Qualification is a qualification approved and managed by the Institute for Apprenticeships and Technical Education.

Owner: Head of Assessment Design

Change History Record

Version	Description of change	Approval	Date of Issue
v1.0	Published final version.		May 2021
v1.1	NCFE rebrand		September 2021