

# Data Science Applications

Assignment Semester 2 2025





## Preamble

The main purpose of the assignment from your perspective is to help you to:

- consider the business environment in which a problem is to be solved;
- apply data science techniques to solve a business problem; and
- communicate the outcomes of your analysis to business stakeholders.

These skills will also help you pass the end of semester assessment and perform well in the workplace.

The specific skills that are being developed and assessed in the assignment are the ability to<sup>1</sup>:

- evaluate how well data describes business activity;
- develop solutions to a range of classification problems using GLMs, tree-based models, ensembling and neural networks;
- evaluate solutions produced by classification models;
- perform k-means and hierarchical clustering;
- evaluate a clustering algorithm using internal, external, and manual validation;
- apply each step in the natural language processing pipeline to solve a variety of business problems;
- evaluate the outcomes of natural language processing models;
- implement strategies for gaining stakeholder support for data science projects;
- communicate relevant points in language appropriate to the audience, in a logical and coherent manner; and
- meet business standards for presentation of work, both modelling and written materials.

The assignment requires you to create a set of sensible assumptions and parameters. You need to demonstrate *how* you derived your assumptions or model parameters. It is important that you describe what you did as the marker(s) will want to understand if you are able to apply knowledge to the specific situation described in this assignment. We are also looking for you to demonstrate that you can deal with uncertainty in a reasonable way.

---

<sup>1</sup> The skills listed here are learning objectives from the subject's syllabus, apart from the last two skills on the list which are assessable in every subject. This assignment does not cover every component of the learning objectives listed above.



The assignment requires you to build and/or use a model. A key actuarial skill is to obtain a grasp of the qualitative nature of outputs from models and describe them. This assignment is designed to test your ability to explain your model(s) and their outputs to a non-technical audience.

There may not be a single correct answer to the question(s) posed. Ensure you have adequately demonstrated your steps, assumptions, reasoning, and checks to the marker(s) so that your answer can be considered in context.

## Marking Guide

A percentage mark of at least 60% is required to achieve at least a pass grade for this assignment, which is consistent across all the subjects' assignments. This is only an indicative pass mark and the final pass mark for the subject overall may be different.

This assignment represents 50% of the available marks for the Data Science Applications subject<sup>2</sup>. Your assignment mark will be combined with your exam mark to determine your overall result for the subject.

It is anticipated that you will spend at least 50 hours to complete the assignment. In past semesters, some students have spent significantly more time than this, particularly those students who aim for a grade of Above Pass Level or Significantly Above Pass Level.

A detailed rubric is provided with the assignment question and will be used by the markers to assess your performance. The rubric has been posted on the Assignments page of Canvas to guide you as to what is required to achieve full marks for each part of the assignment. You should check that the components of your answer cover the items in the rubric.

You should use a clear structure in your work, whether written, coded, or recorded, to make it easy for markers to find where you have responded to each of the rubric criteria.

---

<sup>2</sup> For students completing the subject as a microcredential Certificate path, the assignment represents 100% of the available marks for the microcredential.



## Submission

### Deadline

The deadline for submission is **12:00 pm (midday) Sydney time (AEST) on 3<sup>rd</sup> October 2025.**

Submit your assignment via the Assignments page in Canvas. If you experience technological issues when submitting your assignment, please send a copy of your assignment by email to [education@actuaries.asn.au](mailto:education@actuaries.asn.au).

Penalties apply for late submissions (see section on 'Penalties'). You should anticipate potential delays by preparing and submitting your work in advance of the deadline.

Should circumstances arise that mean you cannot submit your assignment on time, you should contact [education@actuaries.asn.au](mailto:education@actuaries.asn.au) in advance of the deadline and apply for special consideration.

### File format

The submitted documents must consist of an ipynb file, and two pdf files (one of which is a pdf version of the ipynb file). Files in other formats will not be marked. The naming convention for files is:

**DSA 2025 S2 Assignment member ID.(file extension as appropriate).**

Please note that if you resubmit an assessment, Canvas automatically adds a suffix to the file name (such as '-1' for the first resubmission). You do not have to make any adjustment for this.

### Coversheet

A coversheet for the assignment is provided on the Assignments page in Canvas. Complete and attach this coversheet as the front page of your pdf file that contains your answer to Question 4.

Ensure you are complying with the statements provided on the coversheet.



### Word limit

Some questions in the assignment have a specific word or page limit. Markers will not read any part of your answer that exceeds this limit. Keep your word count or page count within any limits that are specified. The word count includes any text within tables, text boxes or images consisting primarily of text. The word count does not include:

- contents table or index; and
- references to sources used.

Keep in mind one of the key principles taught in the Communication, Modelling and Professionalism subject: always write as clearly and succinctly as possible, while still including enough information that will be useful for your audience. With that in mind, consider whether each word, sentence or paragraph you include in your assignment adds to or detracts from the message you are trying to convey. Importantly, know that 'more' is usually not 'best'.

### Video

As part of this assignment, you are required to record a video presentation. Advice about how to record an effective video is provided in an Appendix. You should submit your video by following these steps:

- create a video recording using the naming convention 'DSA 2025 S2 Assignment member ID';
- use your video recording to create an 'unlisted' YouTube video (see instructions in the Appendix)<sup>3</sup>; and
- insert your YouTube video URL as a hyperlink in your assignment pdf file.

### Time limit

The video may have a specified time limit. Markers will not watch any part of your answer that exceeds this limit. Keep your presentation timing within any limits that are specified.

### Jupyter notebook

The Jupyter notebook should use the assignment notebook template provided. The notebook must be capable of running successfully in Google Colab as markers will use this platform to view and access the notebooks. Within the notebook you should:

---

<sup>3</sup> The Appendix also provides advice for students who do not have access to YouTube due to their location.



- explain each step taken in your analysis in a text cell above your code; and
- evaluate and comment on the output from each step in a text cell below the output.

Please note that, unless specified, there is no word limit for the comments in your notebook. However, markers will look more favourably on students who provide clear and succinct commentary, compared to those who provide no commentary or those who provide too much commentary, including those who repeat large sections of the subject materials in their comments. This latter approach makes it very difficult for a marker to assess your understanding of the step being taken or the output being produced.

## Plagiarism

By submitting your assignment, you are implicitly stating that the work is your own.

Remember that an important aspect of being a professional actuary is to always act with integrity. Committing plagiarism by copying another person's work or not properly referencing other sources used in your assignment is a breach of the Integrity principle under the Actuaries Institute's Code of Conduct.

Any suspected plagiarism will be referred to the Institute's Executive General Manager, Education for review. Depending on findings, a penalty may be applied, and/or a complaint regarding the member may be made to the Institute's Conduct Committee. Subject marks may not be released until the matter is resolved.

## Penalties

### Late submissions

Penalties will be applied to late submissions without prior approval.

If you submit an assessment after the due date and time (whether that is the original due date or any extended due date you have been granted), the following penalties apply:

- within 24 hours of due date and time: 20% x maximum mark available (i.e. deduct 4 marks if a 20 mark assignment, deduct 10 marks if a 50 mark assignment);
- more than 24 hours (1 day) late: 100% x maximum mark available (i.e. assessment score = 0).



### Incorrectly formatted submissions

There is no direct penalty if an assessment is submitted in a format with an incorrect file name or an incorrect format (e.g. submitted as a word document when a pdf document was required).

However, you may be required to resubmit your work with the correct file format, particularly relevant to modelling or coding assignments.

If a submission does not include the correct identifier (member ID) in the file name, then it may take time to identify you as the student and you may be asked to resubmit your work with an appropriate identifier.

If either situation arises then this will probably cause you to submit late and hence incur the late submission penalties outlined above. Students should therefore follow all assessment instructions provided.

### Feedback

Our approach to feedback is for students to receive general feedback and a sample assessment marked as 'Significantly above pass level'.

You should review the general feedback that is provided to all students as well as the sample assessment. After reviewing the general feedback, you should use the rubric to grade the sample assessment and your submission. This will help you to compare the assessments and identify areas where your submission could have been improved.

Our belief is that this active approach to studying will provide you with a deeper understanding of where you need to improve. This is the best way for you to learn about your areas of strength and weakness. We do not provide students with individual feedback on their assessments.

At the end of the semester, you will receive:

- a letter to indicate whether you have passed or failed the subject;
- if you have failed the subject, a breakdown of your grade for each assessment;
- general feedback to all students about assessment performance; and
- sample assessment(s) that were graded as 'Significantly above pass level'.



## Assignment Context

You are a data science consulting actuary engaged by Betahelp, a large healthcare provider operating across the United States of America. At a recent executive strategy retreat, Betahelp's management team decided to develop a healthcare program that proactively targets patients who are likely to receive an acute diagnosis over the next 12 months. They have asked for your advice on the following three points:

- Betahelp's pathologists provide medical advice via a commentary section in the lab result reports. When providing this advice, are the pathologists adding information above what is otherwise captured in the structured data?  
*Your work in questions 1, 2, and 3 will be useful for addressing this point.*
- Is it possible to predict which of Betahelp's current patients will receive an acute diagnosis over the coming 12 months?  
*Your work in Question 2 will be useful for addressing this point.*
- Is it possible to split the causes of acute diagnoses between broad environmental factors (e.g. socio-economic status) versus patient-specific factors (e.g. a history of acute disease or poor lifestyle choices)?  
*Your work in Question 2 will be useful for addressing this point.*

To help you complete this task, Betahelp has provided you with a file ('DSA 2025 S2 assignment data.xlsx' or 'the assignment dataset') containing a sample of data on patient statuses and events from 2020 to 2024. The data contains a random sample of patients, filtered to only include those who had a pathology report during that period. The data dictionary for this dataset is set out in Table 1 and Table 2 which can be found in the appendix of this assignment.



# Assignment Questions (Total 100 Marks)

Answer questions 1, 2, and 3 in your Jupyter notebook using the assignment template provided. When you are ready to submit, print your ipynb file to a pdf (using file print in your browser) and submit BOTH files (both the ipynb AND the pdf version of that file). Answer Question 4 in your pdf document.

Different markers will review different questions in this assignment, so your answer to each question and part of that question should be self-contained. No marks will be awarded for answers to a question that are only contained in your answers to other questions.

## 1. Explore and examine the pathology comments

Your work with Betahelf will start by providing them with a summary and analysis of the pathology commentary data they provided. *At this stage, you are performing descriptive analytics, and therefore, you will not partition the data until you work on predictive models in Question 2.*

Answer Question 1 in your Jupyter notebook.

- a. Calculate vectorised features on the commentary section of the pathology report text within the 'written\_report' column of the 'Pathology' table. Use both embeddings and TF-IDF vectorisation. For embeddings, use the [xlreator/biosyn-bioBERT-snomed](#) sentence transformer available at Hugging Face. **(5 marks)**
- b. Apply a clustering algorithm using both a subset of the embedding features and a subset of the TF-IDF features calculated in Question 1a to provide insights into the nature of the pathology report commentaries. **(5 marks)**

*You should use either feature selection or feature extraction methods to significantly reduce the dimensionality of the embedding and TF-IDF features before using these features in the clustering algorithm. This will enhance the effectiveness of your clustering outputs.*

*You should justify your choice of features, clustering algorithm, distance measure, and number of clusters, comparing your choices to alternatives. Use internal validation and the business context to support your justifications. Do not perform manual validation at this point.*



- c. Apply discriminator modelling that predicts your clusters (from Q1b) using only the full set of TF-IDF vectorisations, to understand the top 10 keywords that distinguish each cluster from another. **(5 marks)**

*You should use the Scikit-learn random forest algorithm and variable importance. Use the default hyperparameters for a random forest algorithm. You are not required to assess the performance of the random forest model.*

*You should provide insights on what the results of this analysis imply about the cluster analysis's validity and individual clusters' interpretation.*

- d. Examine the clustering outputs using manual validation. **(10 marks)**
- e. Summarise, in 500 words or less, the results of your cluster analysis, and what it indicates about the information content of pathologists' commentaries. **(5 marks)**

*Your answer should be communicated using language suitable for sharing with the management team at Betahelp.*



## 2. Predict acute diagnoses

Your next task is to build a neural network model that predicts acute diagnoses. Answer Question 2 in your Jupyter notebook

- a. Clean the assignment dataset in preparation for using it to build the classifier. **(5 marks)**

*Do not clean the 'written\_report' data column, as that was done earlier in this assignment. However, you may need to revise any TF-IDF vectorisation of the column should you choose to use it in this question.*

*Note that you must split the data into training, validation, and test sets within this question. You should apply your judgement in deciding when to perform that split.*

*Consider privacy issues when cleaning the data.*

- b. Propose a unit of analysis to use in your model, stating the entity and timestamp(s) that identify each unique row in your training, validation, and testing data. **(5 marks)**

*As part of your answer, you should propose whether to use regularly spaced timestamps, event-driven timestamps, or a combination of both.*

*You should create a table called `unit_of_analysis_df` that contains the specific entity and timestamp values that you will use to train and validate your model.*

- c. Construct a response variable for your classification model. **(5 marks)**

*You must explain your choices in the design of the response variable you constructed in language suitable for a data scientist.*

*Add a column to `unit_of_analysis` called `AcuteDiagnosis`, and fill it with the response variable values you calculate for each row.*

- d. Suggest four metrics you will use to evaluate the success of your classifier. **(5 marks)**



- e. Construct your neural network classifier. **(20 marks)**

*You should use feature engineering techniques to create a set of features that use data from every table in the dataset and are inspired by what you have learned about the business context.*

*You must demonstrate an ability to fine-tune model architecture, hyperparameters, feature selection, regularisation, and optimisation algorithms. You are not required to tune the algorithm beyond these requirements.*

- f. Interpret, in 500 words or less, the performance of your chosen neural network classifier using the metrics suggested in Question 2d and a comparison to benchmarks. **(5 marks)**

*Your answer should be communicated using language suitable for sharing with the management team at Betahelf.*

- g. Interpret, in 500 words or less, the behaviours of your chosen neural network classifier using feature importance, partial dependence (on the top 5 most important features), SHAP explanations (on three validation examples), and fairness metrics. **(5 marks)**

*Your answer should be communicated using language suitable for sharing with the management team at Betahelf.*



## 3. Artificial Intelligence Pathologists

Betahelf's CEO is excited about the opportunities for AI to provide better healthcare advice and cost savings. She has asked you for a preliminary investigation of the practicality of using AI to automate pathologists' commentaries and advice.

Answer Question 3 in your Jupyter notebook.

- a. Construct a Python prototype that generates an LLM prompt for five randomly sampled lab results, and applies each prompt to the Mistral-7B-Instruct:free model hosted by OpenRouter.

**(5 marks)**

*Note that your Python-generated prompt must include not just the patient's key attributes and lab result data (the same details in the current pathology reports) but must also include the structure of a well-written LLM prompt, including sufficient context for the LLM to provide expert advice.*

*Your prompt should focus on extracting insights about patient risk factors and outcomes based on the lab results. You are not expected to replicate medical expertise, but rather to demonstrate effective prompt engineering that helps the LLM provide structured, analytical commentary similar to the existing human-written pathology reports.*

- b. Critique, in 500 words or less, the healthcare advice provided by the LLM.

**(5 marks)**

*Your answer should be communicated using language suitable for sharing with the management team at Betahelf.*



### 4. Video Presentation

You have been invited to give a presentation of your analysis and recommendations at Betahelf's next monthly executive management team meeting. Your agenda is to answer the following three points:

- Betahelf's pathologists provide medical advice via a commentary section in the lab result reports. When providing this advice, are the pathologists adding information above what is otherwise captured in the structured data?
- Is it possible to predict which of Betahelf's current patients will receive an acute diagnosis over the coming 12 months?
- Is it possible to split the causes of acute diagnoses between broad environmental factors (e.g. socio-economic status) versus patient-specific factors (e.g. a history of acute disease or poor lifestyle choices)?

Prepare a 5-minute video to answer these three questions.

**(10 marks)**

*Your presentation should be communicated using language suitable for sharing with the management team at Betahelf.*

*You should use model performance metrics, feature importances, cluster analysis, descriptive analytics, SHAP prediction explanations, and your critique of the AI healthcare advice to support your answers.*

Answer Question 4 in a pdf document with 'Q4' added to the filename as follows:

**DSA 2025 S2 Assignment Q4 member ID.pdf.**

The assignment coversheet should appear as the first page in this pdf file.



## Assignment Appendix: Data Dictionary

Table 1: Data dictionary - tables

Table	Rows	Columns	Table type
Patient	3,626	16	Slowly Changing Dimension
Diagnosis	26,213	7	Event
Visit	89,154	12	Event
LabResult	13,489	4	Event
LabObservation	134,269	8	Event
Pathology	13,489	2	Event
Smoking	2,091	6	Slowly Changing Dimension
ICD9	17,553	4	Dimension
Specialty	65	3	Dimension
StateDetails	56	14	Slowly Changing Dimension
Prescription	109,147	12	Event



Table 2: Data dictionary – columns

Sheet	Column Name	Column Type	Missing Value Count	Unique Value Count	Rows With Missing Values	All Values Unique	Description	Special Column	Entity
Patient	RowID	character	0	3,626	0%	TRUE	Unique ID for each patient version record.	Primary Key	
Patient	ValidFrom	DateTime	0	562	0%	FALSE	Start date/time from which the record is valid.		
Patient	ValidTo	DateTime	0	466	0%	FALSE	End date/time until which the record is valid.		
Patient	PatientGuid	character	0	3,118	0%	FALSE	Globally unique patient identifier.		Patient
Patient	Gender	character	0	2	0%	FALSE	Patient's gender (e.g., Male, Female).		
Patient	DateOfBirth	DateTime	0	1,431	0%	FALSE	Patient's date of birth.		
Patient	StateCode	character	0	51	0%	FALSE	State code of residence.		StateDetails
Patient	BloodType	character	0	8	0%	FALSE	Patient's blood type (e.g., A+, O-).		
Patient	Title	character	0	4	0%	FALSE	Patient's title (e.g., Mr, Ms, Dr).		
Patient	GivenName	character	0	1,036	0%	FALSE	Patient's first name.		
Patient	Surname	character	0	1,988	0%	FALSE	Patient's last name.		
Patient	StreetAddress	character	0	3,566	0%	FALSE	Street address of residence.		
Patient	City	character	0	1,412	0%	FALSE	City of residence.		
Patient	ZipCode	numeric	0	2,089	0%	FALSE	Postal code.		
Patient	Latitude	numeric	0	3,565	0%	FALSE	Latitude coordinate of address.		
Patient	Longitude	numeric	0	3,565	0%	FALSE	Longitude coordinate of address.		
Diagnosis	DiagnosisGuid	character	0	26,213	0%	TRUE	Unique diagnosis record identifier.	Primary Key	Diagnosis
Diagnosis	PatientGuid	character	0	3,055	0%	FALSE	Identifier linking to the patient record.		Patient
Diagnosis	Timestamp	DateTime	0	26,207	0%	FALSE	Date and time diagnosis was recorded.		
Diagnosis	tz_offset	character	0	6	0%	FALSE	Time zone offset for the timestamp.		
Diagnosis	ICD9Code	character	0	1,882	0%	FALSE	ICD-9 code representing the diagnosis.		ICD9
Diagnosis	DiagnosisDescription	character	142	1,750	1%	FALSE	Text description of the diagnosis.		
Diagnosis	Acute	character	0	2	0%	FALSE	Flag indicating if the diagnosis is acute (T/F).		
Visit	VisitGuid	character	0	89,154	0%	TRUE	Unique identifier for the visit record.	Primary Key	Visit
Visit	PatientGuid	character	0	3,118	0%	FALSE	Identifier linking to the patient record.		Patient
Visit	Timestamp	DateTime	0	89,087	0%	FALSE	Date and time of the visit.		
Visit	tz_offset	character	0	6	0%	FALSE	Time zone offset for the visit.		
Visit	Height	numeric	57,467	300	64%	FALSE	Patient's height (in centimetres/inches).		
Visit	Weight	numeric	47,091	1,484	53%	FALSE	Patient's weight (in kilograms/pounds).		
Visit	BMI	numeric	57,993	6,778	65%	FALSE	Body Mass Index calculated for the patient.		
Visit	SystolicBP	numeric	40,732	134	46%	FALSE	Systolic blood pressure reading.		
Visit	DiastolicBP	numeric	40,581	96	46%	FALSE	Diastolic blood pressure reading.		
Visit	RespiratoryRate	numeric	62,991	31	71%	FALSE	Patient's respiratory rate.		
Visit	Temperature	numeric	69,158	157	78%	FALSE	Body temperature recorded during visit.		



Sheet	Column Name	Column Type	Missing Value Count	Unique Value Count	Rows With Missing Values	All Values Unique	Description	Special Column	Entity
Visit	PhysicianSpecialty	character	3	51	0%	FALSE	Specialty of physician seen during visit.		Specialty
LabResult	LabResultGuid	character	0	13,489	0%	TRUE	Unique identifier for the lab result.	Primary Key	LabResult
LabResult	PatientGuid	character	0	3,118	0%	FALSE	Identifier linking to the patient record.		Patient
LabResult	Timestamp	DateTime	0	13,488	0%	FALSE	Date and time when lab result was recorded.		
LabResult	tz_offset	character	0	6	0%	FALSE	Time zone offset for lab result timestamp.		
LabObservation	LabObservationGuid	character	0	134,269	0%	TRUE	Unique identifier for the lab observation.	Primary Key	LabObservation
LabObservation	LabResultGuid	character	0	13,447	0%	FALSE	Identifier linking to the corresponding lab result.		LabResult
LabObservation	HL7Text	character	1	106	0%	FALSE	Raw HL7 text message for the observation.		
LabObservation	ObservationValue	numeric	36,856	2,240	27%	FALSE	Numeric value of the observation.		
LabObservation	Units	character	37,036	45	28%	FALSE	Measurement units for observation (e.g. mg/dL).		
LabObservation	ReferenceRange	character	52,988	177	39%	FALSE	Reference range for the observation value.		
LabObservation	AbnormalFlags	character	126,002	7	94%	FALSE	Flag indicating whether the value is abnormal.		
LabObservation	IsAbnormalValue	logical	0	2	0%	FALSE	Boolean flag for abnormal observation (T/F).		
Pathology	LabResultGuid	character	0	13,489	0%	TRUE	Identifier linking to the lab result for pathology.	Primary Key	LabResult
Pathology	written_report	character	0	13,489	0%	TRUE	Text of the written pathology report.		
Smoking	PatientSmokingStatusGuid	character	0	2,091	0%	TRUE	Unique ID for smoking status version record.	Primary Key	
Smoking	PatientGuid	character	0	1,991	0%	FALSE	Identifier linking to the patient record.		Patient
Smoking	ValidFrom	DateTime	0	305	0%	FALSE	Effective start date/time of smoking status.		
Smoking	ValidTo	DateTime	0	101	0%	FALSE	Effective end date/time of smoking status.		
Smoking	Description	character	0	8	0%	FALSE	Description of smoking status (e.g. smoker, non-smoker).		
Smoking	NISTcode	numeric	0	7	0%	FALSE	Standard code representing smoking status (per NIST).		
ICD9	ICD9Code	character	0	17,553	0%	TRUE	ICD-9 code representing a diagnosis.	Primary Key	ICD9
ICD9	Group1	character	0	17	0%	FALSE	Primary grouping for the ICD-9 code.		
ICD9	Group2	character	0	156	0%	FALSE	Secondary grouping or subcategory.		
ICD9	Group3	character	0	1,225	0%	FALSE	Detailed tertiary grouping.		
Specialty	PhysicianSpecialty	character	1	65	2%	TRUE	Identifier or code for the physician's specialty.	Primary Key	Specialty
Specialty	Specialty	character	0	40	0%	FALSE	Name of the physician's specialty.		
Specialty	SpecialtyGroup	character	0	6	0%	FALSE	Broader category grouping the specialty.		
StateDetails	StateGuid	character	0	56	0%	TRUE	Unique identifier for the state record.	Primary Key	
StateDetails	StateCode	character	0	56	0%	TRUE	State abbreviation or code.		StateDetails
StateDetails	StateName	character	0	56	0%	TRUE	Full name of the state.		
StateDetails	CentroidLatitude	numeric	0	56	0%	TRUE	Latitude coordinate of the state's centroid.		
StateDetails	CentroidLongitude	numeric	0	56	0%	TRUE	Longitude coordinate of the state's centroid.		
StateDetails	Area	numeric	0	56	0%	TRUE	Total area of the state.		



Sheet	Column Name	Column Type	Missing Value Count	Unique Value Count	Rows With Missing Values	All Values Unique	Description	Special Column	Entity
StateDetails	CensusRegion	character	0	12	0%	FALSE	Census-designated region of the state.		
StateDetails	HospitalCount	numeric	0	49	0%	FALSE	Number of hospitals in the state.		
StateDetails	HospitalBedCount	numeric	0	56	0%	TRUE	Total count of hospital beds.		
StateDetails	BelowPovertyLevel	numeric	0	53	0%	FALSE	Population (or percentage) below poverty level.		
StateDetails	Aged65Plus	numeric	0	39	0%	FALSE	Population (or percentage) aged 65 and above.		
StateDetails	TotalPopulation	numeric	0	56	0%	TRUE	Total population of the state.		
StateDetails	ValidFrom	DateTime	0	1	0%	FALSE	Start date/time for which the state data is valid.		
StateDetails	ValidTo	DateTime	0	1	0%	FALSE	End date/time for which the state data is valid.		
Prescription	PrescriptionGuid	character	0	109,147	0%	TRUE	Unique identifier for the prescription record.	Primary Key	Prescription
Prescription	PatientGuid	character	0	13,087	0%	FALSE	Identifier linking to the patient record.		Patient
Prescription	Timestamp	DateTime	0	108,962	0%	FALSE	Date and time when prescription was issued.		
Prescription	tz_offset	character	0	8	0%	FALSE	Time zone offset for prescription timestamp.		
Prescription	Quantity	numeric	0	232	0%	FALSE	Quantity of medication prescribed.		
Prescription	NumberOfRefills	numeric	0	19	0%	FALSE	Number of refills allowed.		
Prescription	RefillAsNeeded	logical	0	2	0%	FALSE	Indicates if refills are allowed as needed (T/F).		
Prescription	GenericAllowed	logical	0	2	0%	FALSE	Indicates if generic substitution permitted (T/F).		
Prescription	NdcCode	character	0	2,322	0%	FALSE	National Drug Code for the medication.		
Prescription	MedicationName	character	646	1,094	1%	FALSE	Name of the prescribed medication.		
Prescription	MedicationStrength	character	646	488	1%	FALSE	Strength or dosage of the medication.		
Prescription	Schedule	numeric	98,007	6	90%	FALSE	DEA controlled drugs schedule category.		

END OF ASSIGNMENT