



Detailed Assignment Feedback

Question 1 – Domain knowledge

In this question, students had to research the key drivers of property prices and explain how well the data given to students captures these key drivers.

Students were generally able to identify and apply practical strategies for acquiring knowledge about the real estate industry. Most students could not only identify key drivers of property prices but also link this research to how well the dataset captured these drivers.

Stronger answers to this question:

- used clear structure and formatting, such as setting out, for each driver, a clearly labelled section explaining the driver and a separate section discussing how well the driver is captured in the dataset;
- provided evidence of having conducted thorough research using multiple domain knowledge acquisition skills (e.g. searching online and talking to a real estate agent);
- included visualisations and examples (such as data summaries, graphs, or screenshots from websites) to illustrate their arguments; and
- used convincing arguments about how each driver influences property prices.

Weaker answers to this question:

- described key drivers without explaining *how* they influenced prices;
- identified drivers that were too broad (e.g. 'supply and demand') without providing sufficient explanation of these drivers;
- identified drivers that were similar or overlapped with one another (e.g. 'population growth' and 'economic growth');
- listed references but did not link these to any sentence or paragraph; and
- lacked structure in their writing (e.g. long paragraphs without clear points).



Question 2 – Deployment decisions

In this question students had to discuss deployment decisions needed for planning the roll-out of the app.

Most students performed strongly in this question, with good discussion of the key deployment decisions that would be needed. The most common of these related to the form of deployment, model architecture, and testing of the app.

Stronger answers to this question:

- had five clear deployment considerations that were discussed well; and
- used suitable language for the audience (your friend), making the answers clear to read and mark.

Weaker answers to this question:

- outlined many considerations without discussing any in detail (e.g. the points were only listed, or deployment-related questions were posed without discussion of considerations that should be made when trying to answer these questions); and
- lacked a clear structure in their response, such as by grouping different considerations together into the one section and/or using section headings that did not match what was covered in each section.



Question 3 – Exploratory data analysis

In this question students had to describe the property dataset using exploratory data analysis and undertake cleaning steps to prepare the non-text data for analysis.

Most students examined more than three features and identified more than three issues in the dataset.

Stronger answers to this question:

- used different data exploration techniques depending on the features they were examining;
- justified the approach taken to explore each feature based on research undertaken in Question 1;
- linked their exploration of features to the problem and modelling context;
- commented on the usefulness of each feature for the analysis;
- created nicely designed visuals to present their findings;
- provided justification for their findings based on their industry research; and
- communicated using language suitable for the target audience;

Weaker answers to this question:

- repeated the same exploratory code and visualisation for each feature summarised;
- provided generic comments about each feature (e.g. number of duplicates and missing data);
- did not offer much insight into the relevance of their findings for the modelling context; and
- had incomplete code or errors in their code.



Question 4 – Vectorising features

This question required students to calculate vectorised features to numerically represent each property's ad heading and ad body.

Most students seemed to have a good grasp of the coding required to implement the key natural language processing steps. However, there was a varying degree to which students explained their understanding of each step and, more importantly, the interpretation of the results achieved.

Stronger answers to this question:

- structured their notebook with well labelled sections;
- commented on why each step was being performed, in a way that was tailored to this assignment context (e.g. by showing and interpreting a small passage of raw text from the data);
- evaluated code output using a range of reasonableness checks (e.g. spot checks and checks on dimensions of inputs v outputs);
- evaluated the output of vectorisation by checking, for example, TF-IDF weights assigned to words and seeing if these made sense after looking at the original text; and
- interpreted code output, in a way that was tailored to this assignment context.

Weaker answers to this question:

- lacked structure either through lack of labelling of sections or through using labels within code cells, meaning the labels don't show in the side menu bar for the notebook;
- only provided very brief outlines of each step or copied comments from last semester's sample assignment without adaption to this semester's context, leaving markers unsure about whether they understood what they were doing in their coding;
- only implemented spot checking of outputs; and
- either did not comment on outputs of checks or only briefly commented on them, making it difficult for markers to ascertain that the student understood the step they had taken.



Data Analytics Applications

Semester 2 2022 Assignment Feedback

Several students appeared to have copied comments from last semester's sample assignment. In some cases, these comments did not match the actual output from their code (e.g. their comments referenced books and book reviews, or their comments referenced, say, nine cleaning steps, when only two had been undertaken). Markers tended to mark these answers below pass level, as it was unclear whether the student understood what they were doing.



Question 5 – Classification

This question required students to build a classification model to predict property sale prices. They were required to take the following steps in building their models:

- create a response variable;
- create an additional feature based on externally sourced data;
- build a classification model; and
- evaluate the model.

5a – Response variable

A key consideration for this question was which bands to use to group property prices into.

Stronger answers to this question:

- provided a technical and context specific justification for their choice of bands; and
- considered several options in helping to determine a suitable option to use and then provided a strong justification for the final band selection.

Weaker answers to this question:

- did not select rounded bands, or selected bands that were too narrow, demonstrating a lack of understanding of the context for this question (e.g. unrounded bands could be confusing for users of the app);
- considered several options for banding but did not make it clear which banding option was selected and/or why;
- provided no justification for their band selection or only provided a technical justification, such as class balance, without reference to the problem context; and
- did not perform at least two types of checks on their output or did not interpret the output of their checks.

5b – Additional feature

There was a large variety of external data sources used in this question. Markers only allocated marks if the outputs were in a format that would allow them to be used for subsequent modelling. For example, if a student had imported an external data source but did not join the data source to the property data, no marks were awarded.



Many students selected some form of postcode or location grouping, such as SA4 grouping, generic socio-economic indicators, or the ABS's Index of Relative Social Advantage and Disadvantage (IRSAD).

Stronger answers to this question:

- experimented with different options for an additional feature, using a one-way analysis of variance to decide on the most suitable additional feature; and
- justified their chosen additional feature with a link to the problem context and their research findings from Question 1.

Weaker answers to this question:

- did not show a consideration of the question context (e.g. the use of features such as altitude or electorate did not, on their own, demonstrate an appropriate level of judgment in selecting a feature suitable for predicting property prices);
- did not perform at least two reasonable checks on the output of their code; and
- did not justify their selection (e.g. they stated that income or location are good predictors of property prices without elaboration or supporting evidence).

5c – Classification model

Most students did not know how to manage missing values. For example, many students just ignored the highly predictive feature 'ad price' because it had missing values. Instead, they might have imputed a value for observations with a missing ad price (e.g. using the mean or median), along with a 'missing ad price' flag as an additional feature.

Many students also incorrectly believed that volatile training curves indicated over fitting.

Stronger answers to this question:

- linked each model iteration to the outcome of the previous model;
- included an additional feature from 5b that was strongly correlated with sales prices;
- used early stopping to avoid over training; and
- used the most important features, such as ad price and location, in their initial models, rather than using less important features such as number of car spaces.

Weaker answers to this question:



- had feature leakage in their model (e.g. including the days between sold date and first advertised date, when sold date is not known for properties that have not sold);
- did not direct their iterations towards model improvements, but rather randomly tried different model architectures, features and/or hyperparameter values;
- misinterpreted the outcomes of training and validation plots;
- optimised model architecture (e.g. the number of hidden neurons in a neural network) separately to the features selected (as you add more features, your model usually needs to be more complex); and
- appeared to follow a template (either from a case study or a past sample assignment), without showing much understanding of what they were doing.

5d – Model evaluation

Stronger answers to this question:

- made very clear links to the problem context in their evaluations, especially the desire to avoid under-predicting a property's price;
- used custom metrics, rather than just standard measures such as accuracy; and
- used ad price as a suitable benchmark, along with commentary about how to deal with properties that did not have an ad price listed.

Weaker answers to this question:

- misapplied the random model as a benchmark, assuming each class was weighted equally despite the classes being imbalanced; and
- inappropriately compared accuracy rates for external benchmarks that had a different number of banded outcomes (such as models discussed in research papers).



Question 6 – Executive summary

In this question students were required to present a five-minute executive summary of their findings to their 'friend' in the real estate industry.

6a – communication skills

Stronger answers to this question:

- presented their points clearly and concisely;
- used non-technical language to help explain their findings to their friend in real estate;
- appeared to be well prepared and rehearsed; and
- used presentation slides or inserted images to supplement their verbal communication (although some students scored full marks without the help of these visual aids).

Weaker answers to this question did not use language that was suitable for their friend in real estate (e.g. they used technical terms such as 'Gradient Boosting Machine' and 'confusion matrices').

6b – content

Most students explained or described the business idea, model performance, and their recommendation on whether to proceed further.

Almost all students noted at least one limitation of the provided dataset, for example, its lack of properties sold in states outside New South Wales and Victoria, and its short time period. However, the ability to discuss other limitations not related to the dataset, and appropriate steps to overcome these other limitations was a good differentiator between the strongest answers and those just around the pass level.

Stronger answers to this question were closely related to the business context and linked their model's limitations with implications for the business problem of building an app.



Question 7 – Proposed refinement to the app

In this question students were required to explain advantages and disadvantages of a refinement to the app that the friend in real estate had proposed.

Stronger answers to this question used a clear structure and explained their advantages and disadvantages clearly, without the use of jargon.

Weaker answers to this question:

- had messy structure in their answers, lacking clear headings and/or having paragraphs and sections that contained multiple ideas;
- provided overlapping advantages or disadvantages and therefore did not explain two distinct advantages and two distinct disadvantages;
- outlined generic disadvantages not related to the problem context; and
- did not provide enough supporting information to explain each of their points.



Sample assignment graded as 'Significantly above pass level'

A sample assignment is provided as an example of one that was graded as 'Significantly above pass level'. Students should use this example, along with the assignment rubric, to help them self-assess their own assignment attempts.

This assignment was marked as Significantly above pass level for the following reasons:

- **Question 1:** The domain research conducted was very thorough. References used were clearly listed against each finding, outcomes from the research were clearly explained, and strong conclusions on the adequacy of the dataset were made.
- **Question 2:** Five key deployment considerations were discussed using language that was easy to follow for someone without a data science background.
- **Question 3:** The exploratory data analysis conducted was very comprehensive and well explained. Appropriate methods were used for exploring each type of feature and these varied depending on the feature being explored.
- **Question 4:** The code had good structure and clear explanations were provided of what each NLP step was doing. Multiple reasonableness checks were performed for each step and the output of each step was interpreted well.
- **Question 5:** Appropriately rounded property price bands were selected and a strong justification for the choice of bands was provided. A suitable external data source was imported, again with a strong justification for the choice. In the modelling, a wide number of iterations were tried, and strong justification was provided for the final model choice. The model evaluation was thorough and tailored well to the context of predicting property prices that will fall within a buyer's budget.
- **Question 6:** Each point was presented clearly and concisely, using suitable language for the intended audience (your friend in real estate).
- **Question 7:** Two relevant advantages and disadvantages of the proposed model refinement were explained, with good links to the problem context.

Please note that this assignment is not 'perfect' and there were other ways to answer each of the questions and still achieve very high marks.