UNIVERSITY OF LONDON

MSc Data Science, Oct 2025 Session

Module: Statistics and Statistical Data Mining

Task Name: Data Engineering and Statistical Analysis using Python 3

Submission Deadline: Wed, 7 Jan 2026, 13:00

- Please Note: You are permitted to upload your Coursework in the final submission area as many times as you like before the deadline. You will receive a similarity/originality score which represents what the Turnitin system identifies as work similar to another source. The originality score can take over 24 hours to generate, especially at busy times e.g. submission deadline.

- If you upload the wrong version of your Coursework, you are able to upload the correct version of your Coursework via the same submission area. You simply need to click on the 'submit paper' button again and submit your new version before the deadline.

  In doing so, this will delete the previous version which you submitted and your new updated version will replace it. Therefore your Turnitin similarity score should not be affected. If there is a change in your Turnitin similarity score, it will be due to any changes you may have made to your Coursework.

- Please note, when the due date is reached, the version you have submitted last, will be considered as your final submission and it will be the version that is marked.

- **Once the due date has passed, it will not be possible for you to upload a different version of your assessment. Therefore, you must ensure you have submitted the correct version of your assessment which you wish to be marked, by the due date.**

You are asked to submit a Jupyter notebook that contains your solution (Weighted at 50% of final mark for the module). You will be given a Jupyter notebook that you can use as a skeleton/guide. Please make sure you use Python 3 and not Python 2. Python 2 code will not be marked and will be considered as a non-submission.

Coursework Description

In this task: you will implement several data engineering steps that are common in data science and machine learning. These steps involve several key topics in statistics. After that you will implement some statistical analysis.

As this is a postgraduate course, you are expected to do self-study or/and research whenever required. Use the data to answer the questions and perform the tasks below (Important: You must submit your answers and code in the Jupyter notebook. Also, for all questions, you must justify your choice/method, provide your Python code and your interpretation of the results).

Data description: the dataset you will use for this task contains data about house sale prices. The file 'data_description.txt' contains a detailed description of all the variables, what they represent, their values and so on. The target variable is 'SalePrice', which is the house's sale price in US dollars. The dataset and its description will be available to download when the coursework is released.

Here is a description of the steps you are asked to implement and their corresponding marks:
Import the required packages and load the data using Python's pandas package. This code is provided for you. You only need to change the path to the data on your machine.

1. Is the SalePrice column normally distributed or not? Back up your answer with a statistical test. Make sure you provide your interpretation of its results.

   [3 marks]
2. For the categorical variables in the data, draw a barchart that shows the *cardinality* of each variable. Make sure the variables are sorted based on the cardinality value (in ascending or descending order, it is up to you).

   [4 marks]
3. Drop all columns that have more than 30% of their data missing, after that drop all rows that contain any missing data.

[3 marks]

4. Is there a statistically significant difference in the mean sale price (SalePrice) between houses with and without central air conditioning (CentralAir)?

[6 marks]

5. Does the mean sale price differ significantly across different Neighborhoods?

[10 marks]

6. Is there a significant association between HouseStyle and the likelihood of having a GarageType?

[6 marks]

7. Does the relationship between OverallQual and SalePrice remain significant after controlling for GrLivArea (living area)?

[10 marks]

8. Are the distributions of LotArea significantly different between homes built before and after 1980?

[6 marks]

9. Is there evidence that higher-quality homes (OverallQual) tend to have newer YearBuilt values?

[6 marks]

10. Do finished basements (BsmtFinType1 = "GLQ") lead to significantly higher sale prices compared to unfinished basements (BsmtFinType1 = "Unf") after accounting for total basement area?

[10 marks]

11. Is there a statistically significant difference in the median sale price (SalePrice) among houses with different foundation types (Foundation)?

[6 marks]

12. To what extent can we build a robust predictive model for house prices in Ames that remains statistically valid across neighborhoods, housing styles, and economic conditions?
    Specifically:
    a- How do key predictors (e.g., GrLivArea, OverallQual, YearBuilt) interact with categorical variables like Neighborhood?

    [10 marks]

    b- Do the relationships (slopes) between predictors and SalePrice vary significantly across subgroups (i.e., is there interaction or moderation)?

    [10 marks]

    c- Can we statistically demonstrate model generalizability using cross-validation and nested F-tests between hierarchical models (e.g., base model vs. model with interactions)?

    [10 marks]

Please refer to Appendix C of the Programme Regulations for detailed Assessment Criteria.

Plagiarism:

This is cheating. Do not be tempted and certainly do not succumb to temptation. Plagiarised copies are invariably rooted out and severe penalties apply. All assignment submissions are electronically tested for plagiarism.