NHH

TERM PAPER
BAN436

Spring, 2025

Start: 10.01.2025 12:00

End: 24.01.2025 12:00

THE TERM PAPER SHOULD BE SUBMITTED IN WISEFLOW

You can find information on how to submit your paper here:
https://www.nhh.no/en/for-students/examinations/home-exams-and-assignments/

Your candidate number will be announced on StudentWeb. The candidate number should be noted on all pages (not your name or student number). In case of group examinations, the candidate numbers of all group members should be noted.

Collaboration between individuals or groups on submission preparation, as well as exchange of self-produced materials between individuals or groups is prohibited. The answer paper must consist of individual's or the group's own assessments and analysis. All communication during the home exam is considered cheating. All submitted assignments are processed in Ouriginal, a plagiarism control system used by NHH.

SUPPLEMENTARY REGULATIONS FOR EXAMINATIONS

You can find supplementary regulations under the headline “Regulations”

https://www.nhh.no/en/for-students/regulations/

Find more information under chapter 4.0 in the Supplementary provisions to the regulations for fulltime study programmes

Number of pages, including front page: 5

Number of attachments: 2 (AmesHousing.csv, data_description.txt)

BAN436 Term paper

About the exam:

- Deadline for submission is Friday January 24 at 12:00 on Wiseflow.
- You can submit individually or as a group (max. 3 students).
- Your submission must be a zipped folder that contains a Jupyter notebook and any other additional data files that are necessary to execute the code.
- Please write the candidate numbers of all group members in the beginning of the Jupyter notebook (do not write your names).
- You are allowed (and expected) to consult online sources for assistance. However, your code must be mainly written by you, and you are not allowed to submit a solution that is directly generated by generative AI (e.g., ChatGPT).

The Ames Housing Data competition

You are entering the Ames Housing Data competition on Kaggle! Kaggle is a platform for predictive modeling and analytics competitions in which participants compete to develop the best predictive models or solutions for given datasets and problems. The Ames Housing Data competition is a data science competition for beginners that challenges participants to develop predictive models for housing prices based on a comprehensive dataset related to the housing market in Ames, Iowa, USA.

The primary goal of the competition is to predict the final sale prices of homes based on various features and attributes provided in the dataset. `AmesHousing.csv` contains the sale price of 2,930 residential homes sold between 2006 and 2011. In addition to the sale price, the file also contains 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. A description of all the variables and their values is provided in `data_description.txt`.

In this assignment, you are asked to prepare for the competition by exploring the Ames Housing Data and predict the sale price using simple linear regression. The assignment consists of the following tasks:

Task 1: Import and explore the data

The data contains 79 explanatory variables, in which 36 are numerical and 43 are categorical. You should import the file and provide some simple exploration of the data, e.g., column data types, missing observations, descriptive statistics, correlations, value counts etc.

Task 2: Visualize the data

To better understand the data, you should create plots that visualize both the dependent variable (i.e., the sale price) and the potential explanatory variables. Choose any plots that you think are helpful to understand the explanatory variables and their relationship with the sale price of homes, e.g., line plots, histograms, scatter plots, heat maps etc.

Task 3: Data wrangling

Based on your exploration of the data in the previous tasks, you should wrangle the data into a format that is appropriate for the regression analysis in the next task. Operations that you can perform include (but are not limited to):

- Handling missing observations (e.g., drop, impute value)
- Remove outliers and/or wrong values
- Modify data types
- Recode variables (e.g., from numerical to categorical)

You should make sure to consult the data description provided in the text file when you modify the variables:

Note that several of the columns contain "fake" NaNs. For example, the data description states that a missing value in the variable that measures the quality of the basement ("BsmtCond") indicates that the house does not have a basement.

Note also that some numerical variables in the data are in fact categorical. For example, the variable "MSSubClass" uses numbers to represent the type of dwelling (e.g., 20 means "1-STORY 1946 & NEWER ALL STYLES").

It is up to you to decide how to clean the data and which explanatory variables to include in the final data set. However, you should try to include as many of the variables from the original data as possible, and the final dataset should include both numerical and categorical variables.

Task 4: Predict sale price

Using the tidy data created in the previous task, you should now find the simple linear regression model that performs best in predicting housing prices.

For each explanatory variable in your final data:

1. Estimate a simple linear regression model:

$$SalePrice_i = \alpha + \beta \times explanatory\_variable_i$$

2. Evaluate the predicative power of the model by calculating the Root-Mean-Squared-Error (RMSE) of the model.

Note that the RMSE is a common metric for evaluating model performance in predicative modelling and it is calculated according to the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

where:

- y_i is the actual sale price of house i
- $\hat{y}_i$ is the predicted sale price of house i
- N is the total number of homes in your data

The lower the RMSE, the better is the model at predicting the sale price of homes.

Which explanatory variable in your data has the lowest RMSE? Plot the in-sample predictions from the best model along the 45-degree line. According to the plot, does it seem like your simple model does a good job of predicting the sale prices in the data?

Task 5: Personal statement

Write a short statement (in a Markdown cell) in which you describe your use of online sources, e.g., Stackoverflow, official function documentation, generative AI etc. If you have copied code from online, you can use this statement to point out the code snippet that you have not written yourself along with a reference to the source. In addition, if you have used generative AI, e.g., ChatGPT, your statement should explain how you have used such tools to help with the assignment.

Assessment of term paper:

Submissions will be graded based on the point system in the rubric below. The criteria for assessment are based on the learning outcomes in the course. To receive a pass, a submission must achieve at least **6 out of 12** possible points.

Points	**Criteria**	**Learning outcome**
1	Jupyter notebook compiles without errors.	• write, modify and execute Python code in Jupyter notebook.

1	Code and analysis are explained and documented using Markdown cells.	• conduct reproducible research in Jupyter Notebook. • understand the importance of documentation when coding.
1	Use of functions from packages/modules not covered in lectures.	• use package documentation and online sources for help with coding.
1	Use of control structure (loops, if-else statements) to avoid unnecessary duplication of code.	• create functions and loops.
1	Use of self-defined functions to increase code reusability.	• create functions and loops.
1	Professional-looking graphs that summarize the data.	• visualize data.
1	Exploration and handling of missing observations.	• load, manipulate and save data.
1	Exploration and handling of column data types.	• load, manipulate and save data. • distinguish between the different data types and structures in Python.
1	Exploration and handling of outliers in the data.	• load, manipulate and save data.
1	Cleaning/transformation of variables to a format that is suitable for statistical analysis.	• load, manipulate and save data. • identify the appropriate format of data sets with regards to data analysis.
1	Estimation of linear regression models to predict the sale price of the homes in the data.	• perform simple data analysis.
1	Calculation of the RMSE to estimate the predicative power of regression models.	• perform simple data analysis.