Project Plan Computational Statistics

Leila Maillefer, Fabia Schreyer, Loan Strübi

April 2025

Project Plan: Impact of House Style on Sale Price

Objective

This project aims to evaluate the influence of architectural style (`HouseStyle`) and building type (`BldgType`) on the sale price (`SalePrice`) of houses in the Ames dataset. We seek to determine whether certain combinations of structural characteristics are systematically more valued on the housing market, while controlling for other influential factors such as living area and overall quality. The project includes the application of classical statistical inference, 2^k factorial design, regression and ANOVA, and time series analysis.

1. Classical Statistical Inference

- **Descriptive Statistics**: We will compute measures such as the mean, median, variance, and standard deviation of `SalePrice` across different `HouseStyle` and `BldgType` categories. This gives a preliminary understanding of the impact of structural features on housing prices.
- **Data Visualization**: Boxplots and histograms will be used to visually compare the distribution of `SalePrice` by style/type. This will help identify potential outliers, skewness, and group-specific characteristics.
- **Hypothesis Testing**: For example, we can test whether the mean price of 2-story houses is significantly different from that of 1-story houses using a two-sample t-test. We may also use ANOVA to compare means across more than two categories.
- **Confidence Intervals**: We will compute 95% confidence intervals for the mean `SalePrice` within each `HouseStyle` group to assess variability and statistical reliability.

2. 2^k Factorial Design and ANOVA

- **Factor Selection**: We will select four binary explanatory variables:
  - `HouseStyle`: 2Story (1) vs 1Story (0)

  - `BldgType`: 1Fam (1) vs Twnhs (0)
  - `OverallQual`: High quality (>7) vs others
  - `CentralAir`: Yes (1) vs No (0)

- **Design Matrix**: A full 2^4 factorial design will be constructed using these variables. Each row represents a hypothetical combination of the four binary factors.
- **ANOVA Analysis**: Using the design matrix, we will estimate main effects and two-way interactions using ANOVA. This will help identify which factors (or combinations) most affect the house price.
- **Model Evaluation**: The statistical significance of each factor will be tested, and effect sizes will be interpreted to draw practical conclusions.

3. Regression and ANOVA Analysis

- **Model Building**: We will construct multiple linear regression models to predict `SalePrice` using the following predictors:
  - `GrLivArea`, `OverallQual`, `HouseStyle`, `BldgType`
- **Interaction Terms**: To capture conditional effects, we will include interactions such as `GrLivArea * HouseStyle`. These terms can reveal whether the effect of living area differs depending on the house style.
- **Model Diagnostics**: Residual analysis, multicollinearity checks (via VIF), and performance metrics (e.g., R^2, RMSE) will be used to assess the models.
- **ANOVA between Models**: Nested models will be compared using ANOVA to assess whether the inclusion of new variables or interaction terms significantly improves the model.

4. Time Series Analysis

- **Data Aggregation**: We will group the sales by month and year using `MoSold` and `YrSold`, then compute the average `SalePrice` for each time point.
- **Subgroup Analysis by Style**: Separate time series will be constructed for two main house styles (e.g., 1Story and 2Story) to compare their price evolution over time.
- **SARIMA Modeling**: A SARIMA model will be fitted to at least one of the time series to model the trend and seasonal components. The model order will be selected based on AIC/BIC criteria and residual diagnostics.
- **Interpretation**: We will discuss market trends, seasonality (e.g., peak selling periods), and compare long-term price dynamics across house styles.

Expected Outcomes

This project will provide a comprehensive statistical analysis of how architectural and structural characteristics influence house prices. We expect to identify the most impactful style and type combinations, evaluate model performance using real-world data, and interpret trends over time. The project will showcase the practical application of classical inference, factorial design, regression modeling, and time series analysis in a unified real estate context.