Đăng Ký Đề Tài Đồ án Lập trình Python cho Máy học

Lớp: CS116.P21

I. Thông tin nhóm:

STT	Họ và tên	MSSV
1	Nguyễn Thị Lý	22520837
2	Hà Ngũ Long Nguyên	22520965
3	Nguyễn Thu Phương	22521167

II. Tên đồ án và Dataset:

- **Competition:** Housing Prices Competition for Kaggle Learn Users
- **Giới thiệu về cuộc thi:** Cuộc thi dự đoán giá nhà tại Ames, Iowa, là một thử thách thú vị dành cho các nhà khoa học dữ liệu và những người yêu thích phân tích dữ liệu. Với 79 biến giải thích, bộ dữ liệu này cung cấp cái nhìn sâu sắc về nhiều yếu tố ảnh hưởng đến giá trị bất động sản, từ kích thước phòng ngủ đến các đặc điểm kiến trúc độc đáo. Đây là cuộc thi diễn ra vô thời hạn với bảng xếp hạng liên tục
- **Metric:** Các bài nộp được đánh giá dựa trên Root-Mean-Squared-Error (RMSE) giữa logarit của giá trị dự đoán và logarit của giá bán quan sát được.

III. Số liệu ban đầu về quy mô dữ liệu:

- Train.csv - Tập huấn luyện.
- Test.csv - Tập kiểm tra

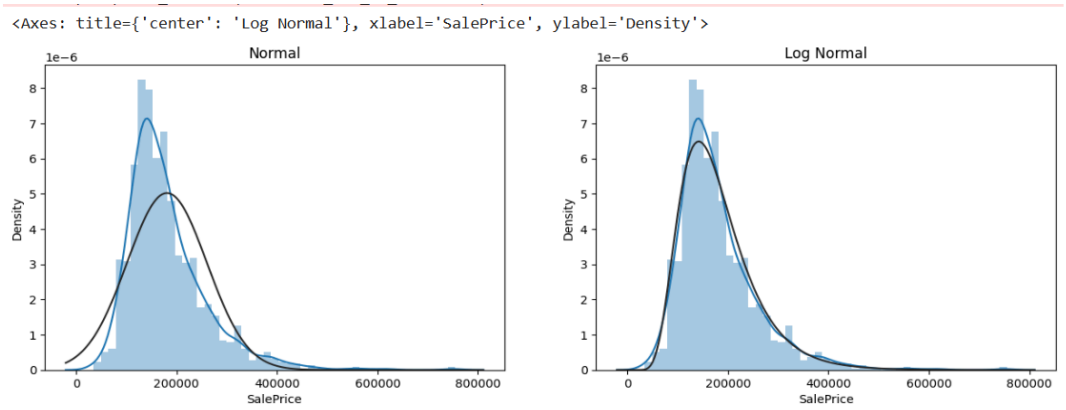
IV. Phân tích dữ liệu với EDA

- **EDA (Exploratory Data Analysis – Phân tích Khám phá Dữ liệu)** là một bước quan trọng trước khi làm bất kỳ một bài toán ML với dữ liệu dạng bảng nào.
- Các bước EDA bao gồm:
  + Data Distribution.
  + Phân tích đơn biến.
  + Phân tích đa biến: feature - feature và feature - label.
  + Phân tích Outlier.

4.1. Data Distribution:

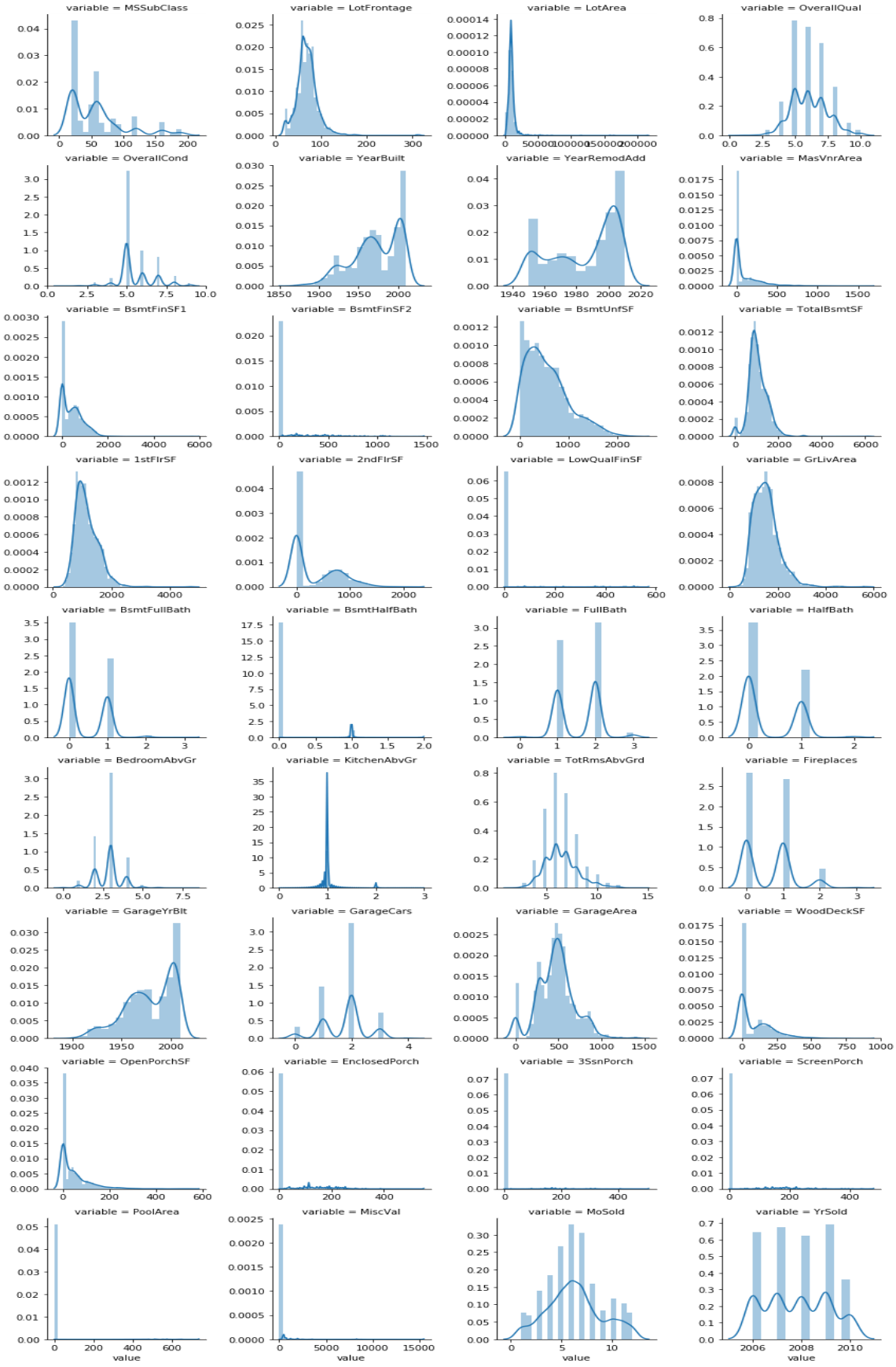
Bộ dữ liệu mà cuộc thi đang sử dụng bao gồm 1460 mẫu trong tập training và 1459 mẫu trong tập testing. Mỗi mẫu trong tập train có 79 attributes, trong đó 43 attributes là định tính (qualitative) và 36 attributes là định lượng (quantitative). Ngoài ra còn có thêm 2 trường là ***Id*** và ***SalePrice*** (giá bán nhà cần dự đoán).

Trước khi đi vào phân tích attributes của bộ dữ liệu, ta cần phân tích biến ***SalePrice***. Đây chính là giá trị mà mô hình cần dự đoán.*SalePrice* sẽ được mô tả qua histogram:



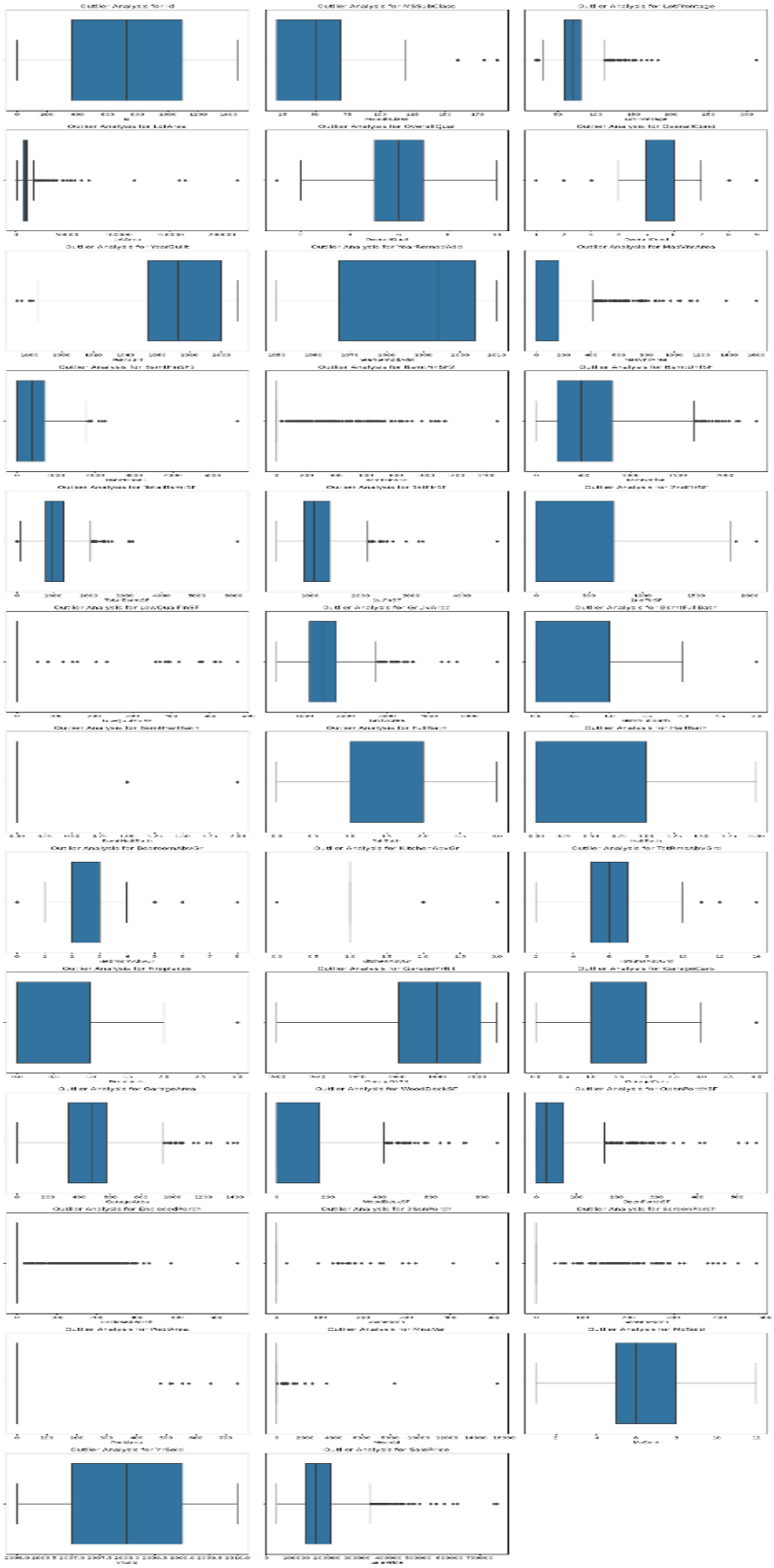
Có thể thấy rằng, phân bố của biến ***SalePrice*** không phải là phân bố chuẩn (normal) mà giống với phân bố log-normal. Vì vậy, trước khi đưa vào thuật toán Machine Learning, biến *SalePrice* nên được biến đổi bằng log transformation.

Tương tự, chúng ta kiểm tra phân bố của tất cả các biến định lượng (quantitative) trong bộ dữ liệu.



Có thể thấy, có một vài biến cần phải biến đổi log transformation như: ***TotalBsmtSF***, ***KitchenAbvGr***, ***LotFrontage***, ***LotArea***, ...

4.2. Phân tích Outlier:



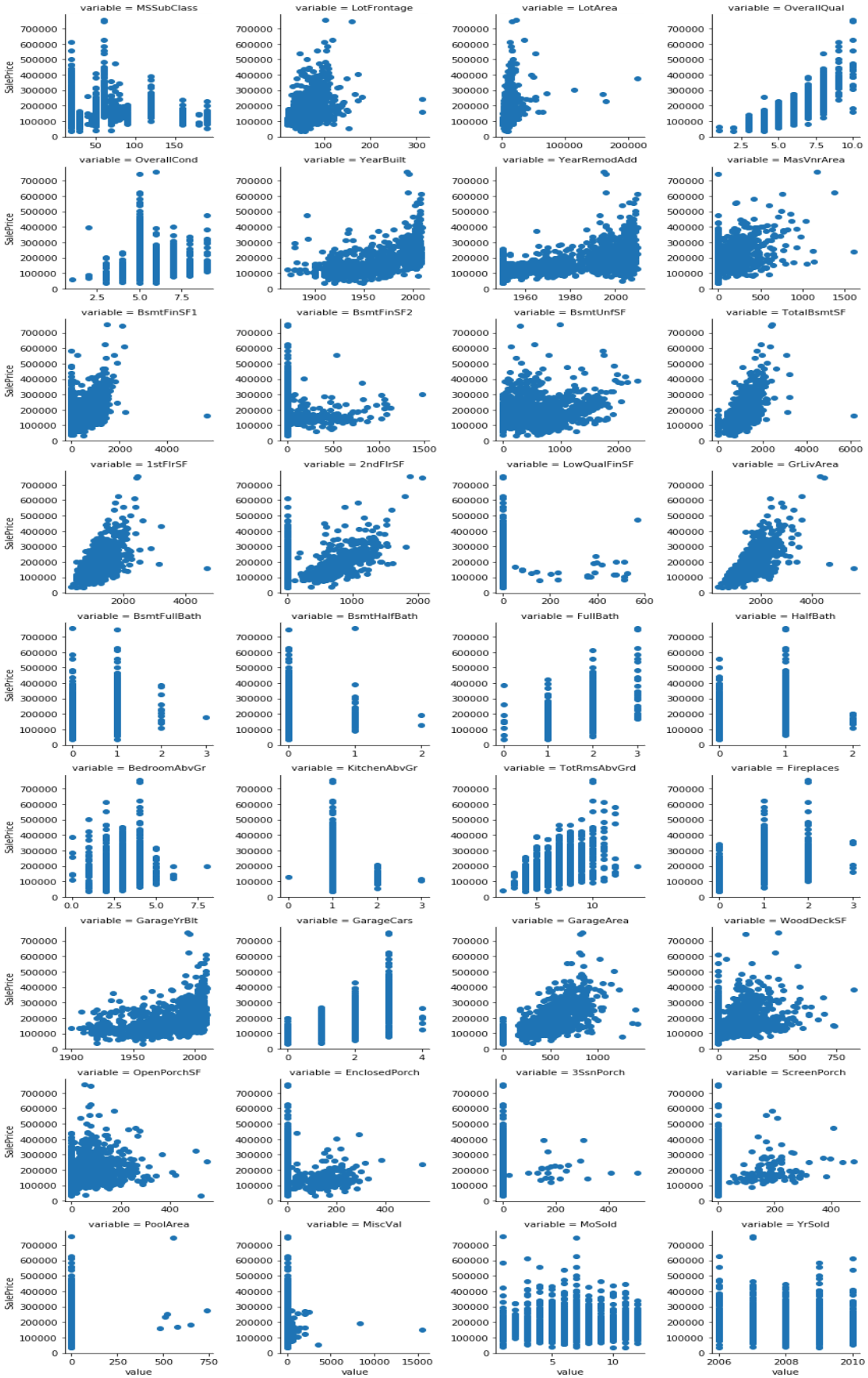
Các biểu đồ boxplot cho thấy nhiều biến có điểm ngoại lai rõ rệt. Ví dụ:

- **"GrLivArea"** và **"TotalBsmtSF"**: Có nhiều điểm ngoại lai, có thể đại diện cho các căn nhà rất lớn so với phần còn lại, cần điều tra thêm.

- **"SalePrice":** Có một số nhà với giá rất cao, kéo median lên, cho thấy sự chênh lệch lớn trong giá trị bất động sản.

4.3. Phân tích đơn biến, đa biến:

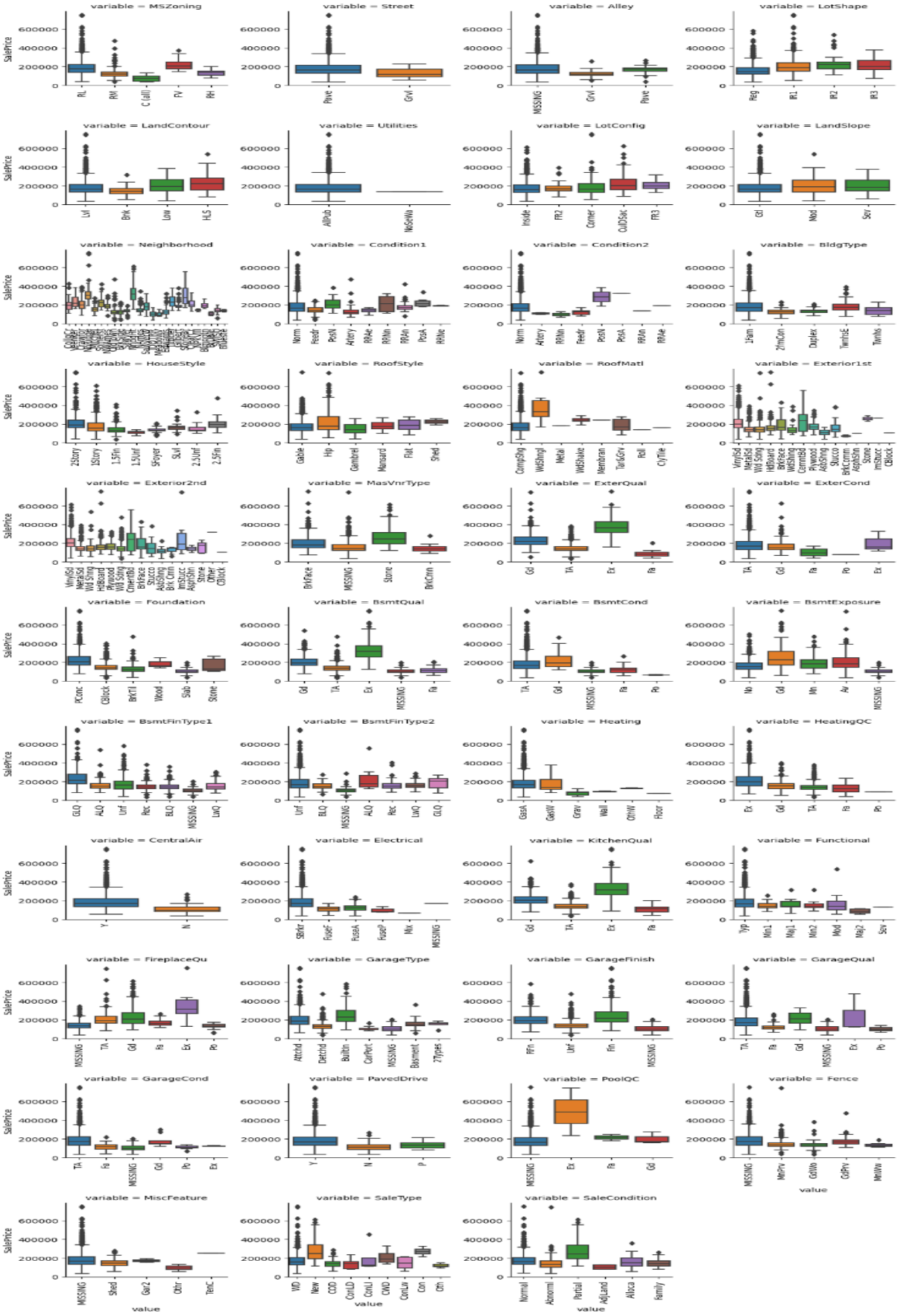
Tiếp theo, chúng ta quan sát quan hệ giữa giá nhà với các giá trị định lượng thông qua biểu đồ scatter plot.



Qua phân tích các scatter plot, các biến như ***LotFrontage***, ***LotArea***, ***OverallQual***, ***TotalBsmtSF***, ... có quan hệ tuyến tính (linear) với ***SalePrice***. Một

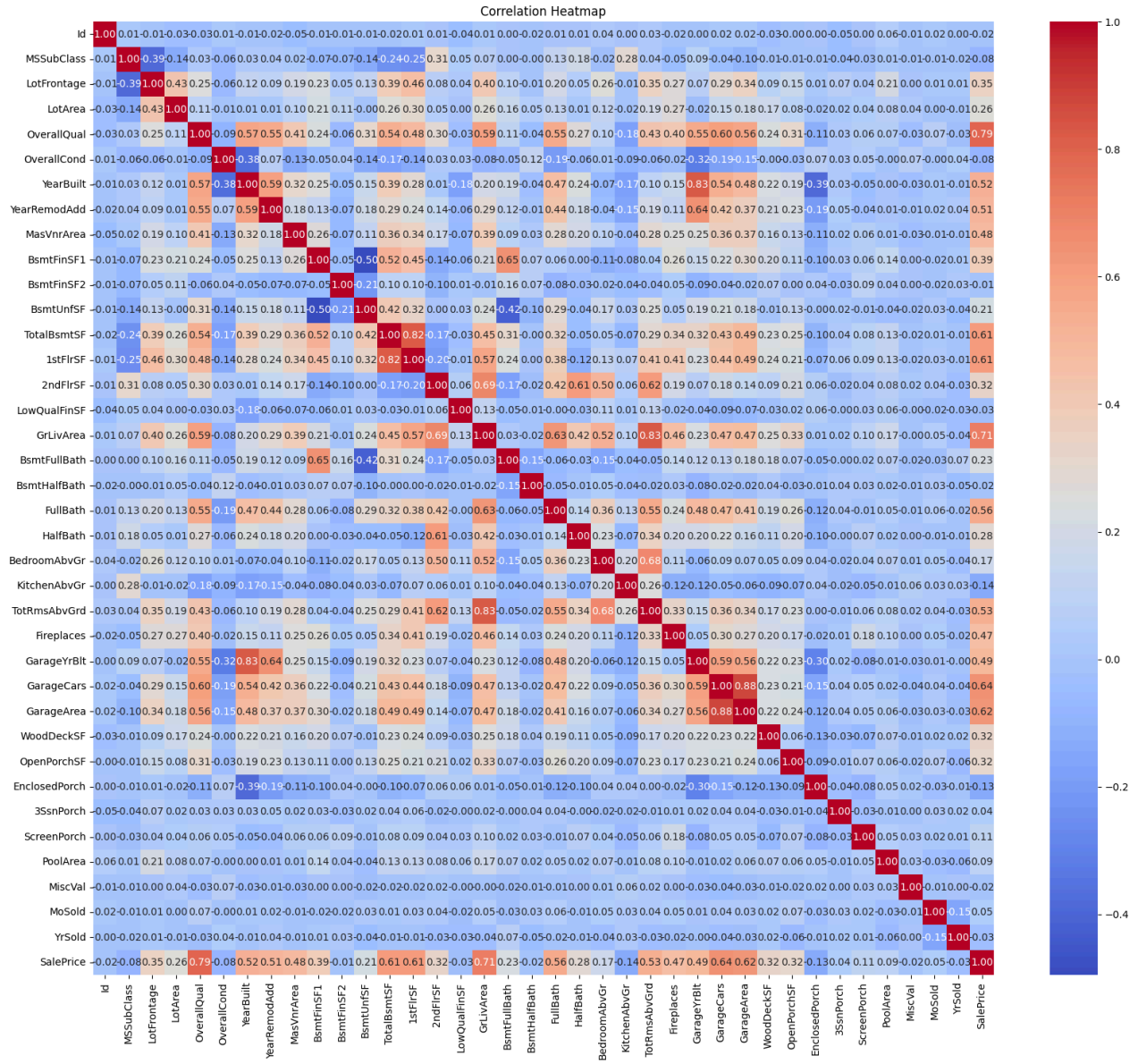
số biến khác như ***BsmtFullBath***, ***HalfBath***, ***Fireplaces***, ... lại có thể biểu diễn ***SalePrice*** thông qua phương trình bậc hai (quadratic equation).

Đối với các biến định tính, boxplot sẽ được sử dụng để phân tích phân bố của ***SalePrice*** đối với từng attribute:


variable = MSZoning
variable = Street
variable = Alley
variable = LotShape
variable = LandContour
variable = Utilities
variable = LotConfig
variable = LandSlope
variable = Neighborhood
variable = Condition1
variable = Condition2
variable = BldgType
variable = HouseStyle
variable = RoofStyle
variable = RoofMatl
variable = Exterior1st
variable = Exterior2nd
variable = MasVnrType
variable = ExterQual
variable = ExterCond
variable = Foundation
variable = BsmtQual
variable = BsmtCond
variable = BsmtExposure
variable = BsmtFinType1
variable = BsmtFinType2
variable = Heating
variable = HeatingQC
variable = CentralAir
variable = Electrical
variable = KitchenQual
variable = Functional
variable = FireplaceQu
variable = GarageType
variable = GarageFinish
variable = GarageQual
variable = GarageCond
variable = PavedDrive
variable = PoolQC
variable = Fence
variable = MiscFeature
variable = SaleType
variable = SaleCondition
SalePrice
value

Một số biến thể hiện sự đa dạng rõ rệt của ***SalePrice*** so với các biến còn lại, ví dụ như: ***Neighborhood***, ***SaleCondition***, …

Cuối cùng, chúng ta xem xét mối tương quan giữa các biến của dữ liệu.



Chúng ta thấy có một số biến có tương quan rõ rệt với nhau như họ các biến *Garage* hay họ các biến *Bsmt*. Do đó, có thể lược bỏ một số biến để làm giảm độ phức tạp của mô hình mà không làm ảnh hưởng quá nhiều đến hiệu suất.