***Học viên:* Đoàn Văn Thái**

***Lớp:* Khoa học dữ liệu**

***Bài tập môn:* Đồ họa máy tính**

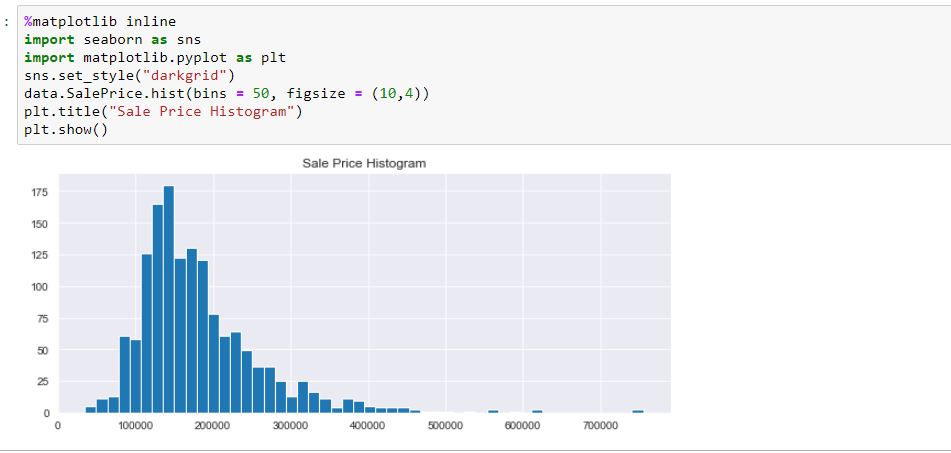
Đề bài: Thực hiện phân tích trên 1 bộ dữ liệu

Đường link đến bộ dữ liệu thực hiện phân tích: Bộ dữ liệu cho bài toán xác định giá nhà tại địa chỉ: <https://www.kaggle.com/alphaepsilon/housing-prices-dataset>

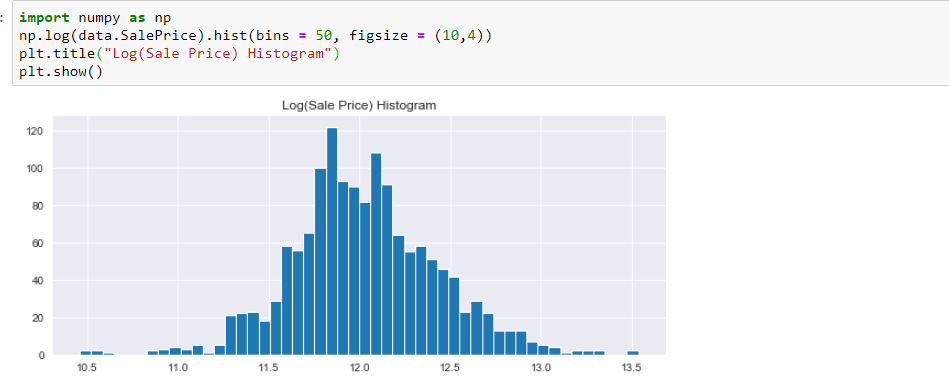
Trong bộ dữ liệu này, ta sẽ thực hiện phân tích dữ liệu trên các trường sau:

* SalePrice: Giá ngôi nhà
* YearBuild: Năm xây dựng ngôi nhà
* YrSold: Năm bán ngôi nhà
* MoSold: Tháng bán ngôi nhà
* HouseStyle: Kiểu của ngôi nhà
* FullBath: Số lượng phòng tắm đầy đủ
* GarageCar: Số lượng xe có thể đỗ trong gara ô tô
* LotArea: Diện tích lô đất
* LotFrontage: Diện tích phần sân trước ngôi nhà
* BsmtFinSF1: Diện tích tầng hầm loại 1
* BsmtFinSF2: Diện tích tầng hầm loại 2
* TotalBsmtSF: Tổng diện tích tầng hầm
* 1stFlrSF: Diện tích tầng 1
* 2ndFlrSF: Diện tích tầng 2

Trước hết, ta sử dụng biểu đồ histogram đối với biến SalePrice để xác định phân phối của biến này:



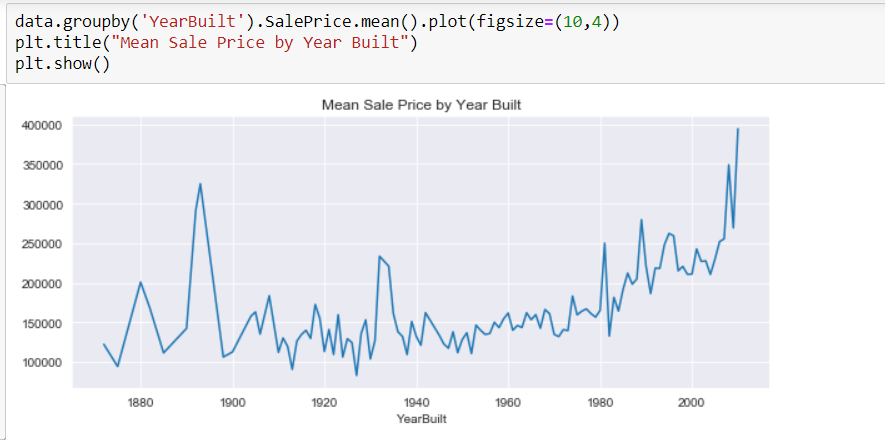
Từ biểu đồ trên, nhận thấy giá trị của biến SalePrice bị lệch phải. Để xử lý việc lệch phải này, ta có thể tiến hành lấy logarit của SalePrice. Ta có đồ thị tương ứng như sau:



Việc đưa giá trị của biến phụ thuộc (giá nhà) trong bài toán này về dạng phân phối chuẩn như trên sẽ giúp cho việc mô hình hóa phía sau có độ chính xác cao hơn.

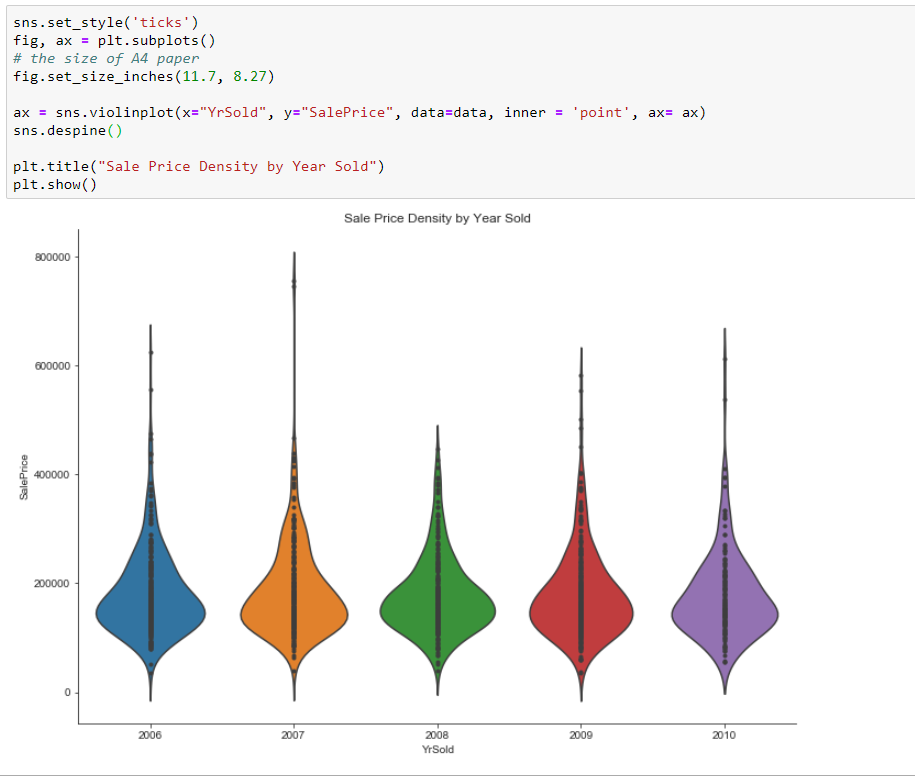
Tiếp theo, ta tiến hành phân tích sự phụ thuộc của SalePrice theo các giá trị của các biến rời rạc.

Xét sự phụ thuộc của SalePrice theo năm xây dựng ngôi nhà, ta có đồ thị biểu diễn trung bình giá trị của các ngôi nhà theo năm xây dựng như sau

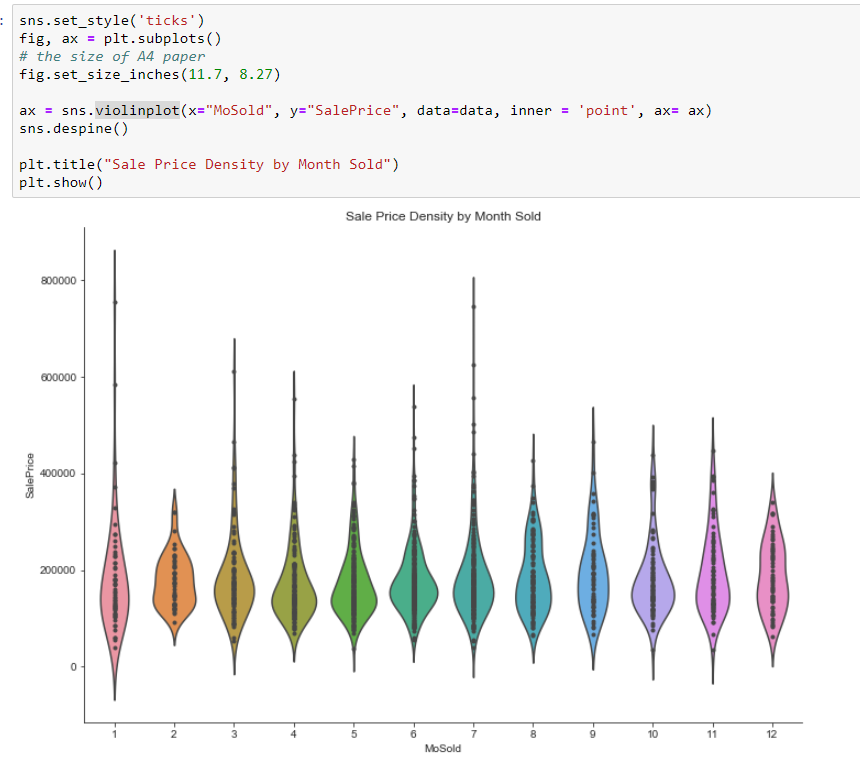


Từ biểu đồ trên, ta có thể thấy được, trung bình các ngôi nhà có năm xây dựng càng mới sẽ có giá cao hơn.

Xét theo năm bán ngôi nhà, ta có đồ thị Violin dưới đây biểu diễn phân phối của giá nhà theo năm bán ngôi nhà

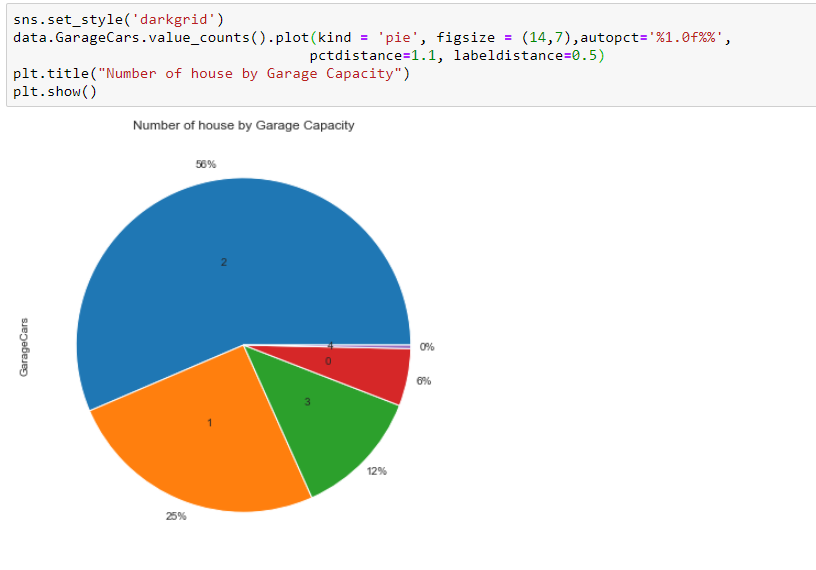


Từ đồ thị biểu diễn phân phối trên, ta thấy được giá của những ngôi nhà được bán vào năm 2007 biến động mạnh hơn giá nhà của những năm khác (biên độ dao động là dài nhất). Trong khi đó, giá nhà năm 2008 ít biến động nhất, còn những năm còn lại (2006, 2009, 2010) có giá nhà gần như cùng 1 phân phối.



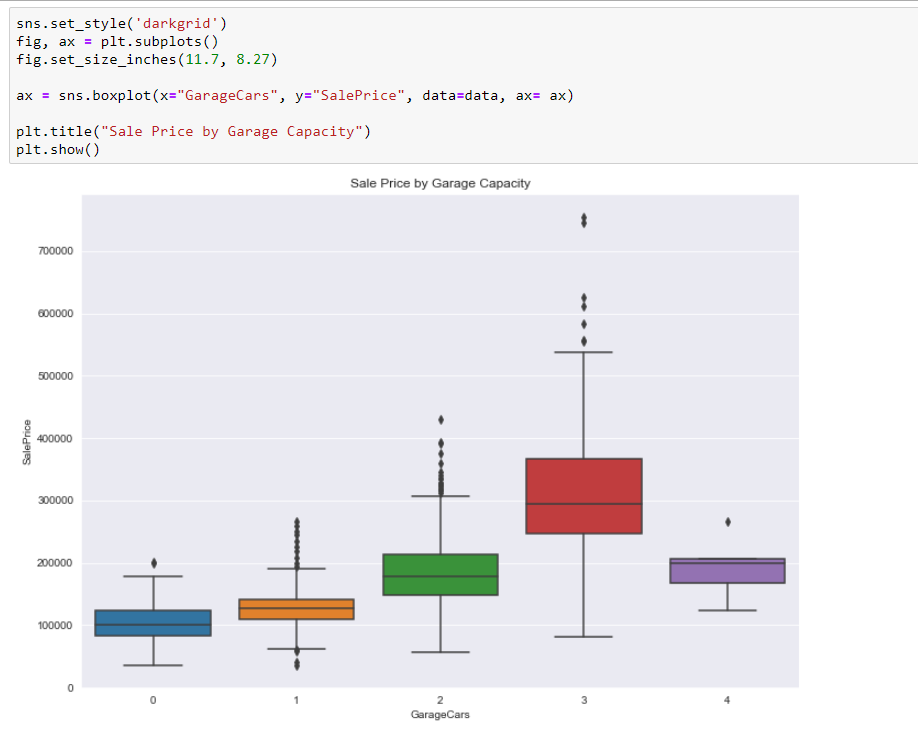
Xét tiếp về tháng bán ngôi nhà, ta có biểu đồ biểu diễn phân phối giá nhà như hình trên. Từ biểu đồ, ta thấy giá nhà bán vào tháng 1 có mức biến động cao nhất còn những tháng 4,5,6,7 có giá trị ổn định hơn

Xét về khu vực gara ô tô, ta có biểu đồ dưới đây thể hiện Số lượng các ngôi nhà theo từng mức thể tích của gara ô tô



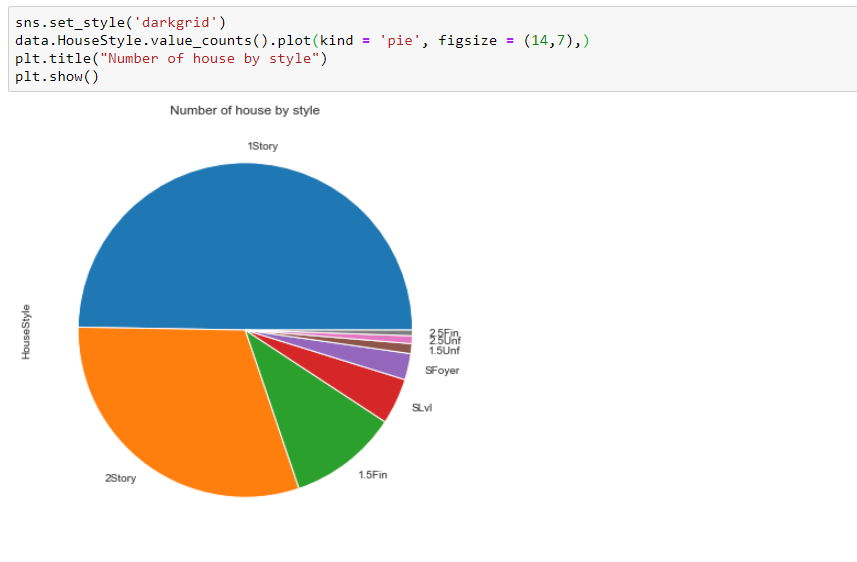
Từ biểu đồ, ta thấy phần lớn các ngôi nhà được xét có gara để được 2 xe có tỷ lệ chiếm 56%, tiếp theo là nhà có gara chỉ để được 1 xe chiếm 25%, các trường hợp còn lại chỉ chiếm 20%

Tiếp tục, ta xét biến động của giá nhà dựa theo dung lượng của gara ô tô, ta có biểu đồ hộp dưới đây ứng với từng loại dung lượng của gara ô tô

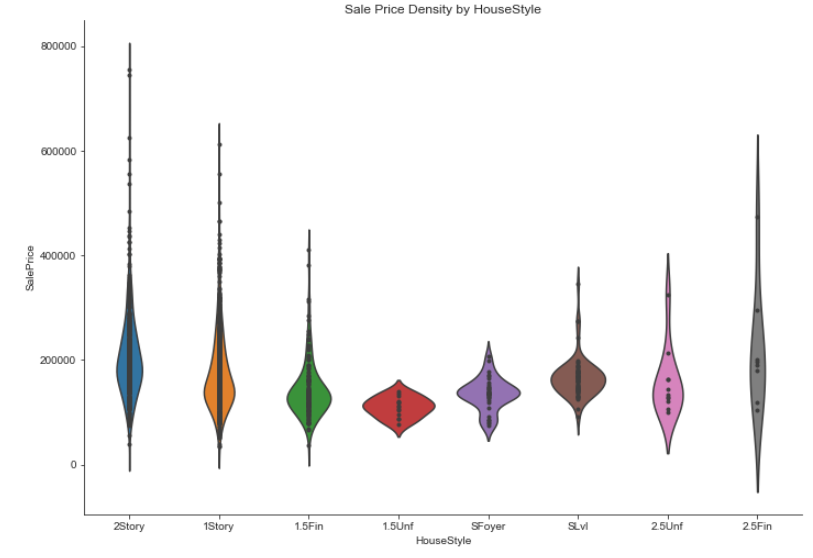


Từ biểu đồ này, ta thấy, trừ trường hợp nhà có gara để được 4 ô tô, giá nhà tăng khi gara ô tô có thể để được nhiều xe hơn. Nhà có gara để vừa 3 ô tô có giá trung bình là cao nhất.

Xét về kiểu của nhà, ta có biểu đồ thể hiện tỷ lệ nhà ứng với từng kiểu như trong biểu đồ dưới đây:

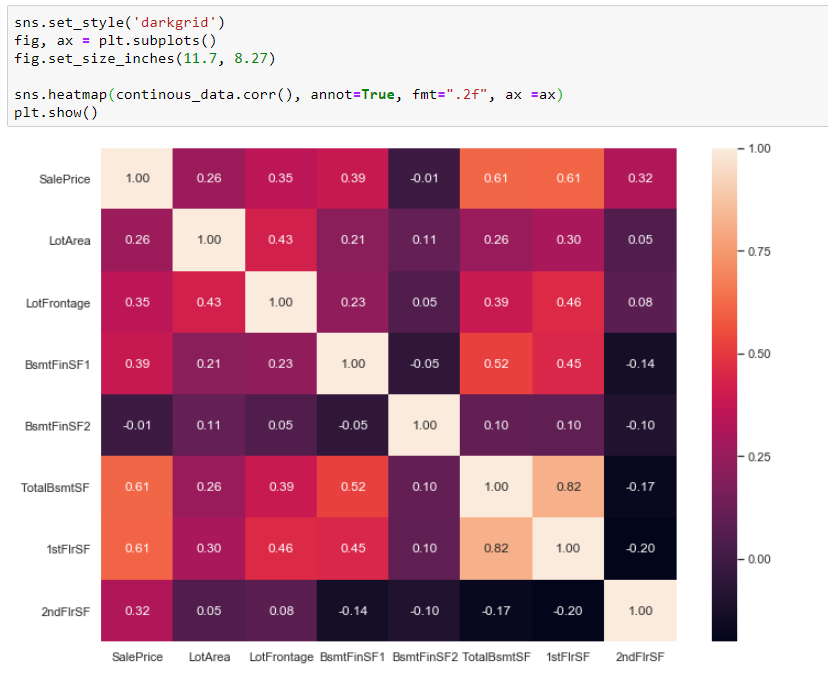


Từ biểu đồ này, ta thấy phần lớn nhà được xét là nhà có kiểu thiết kế 1 hoặc 2 tầng (chiếm khoảng 75%), tổng tỉ lệ của các kiểu thiết kế còn lại khoảng 25%



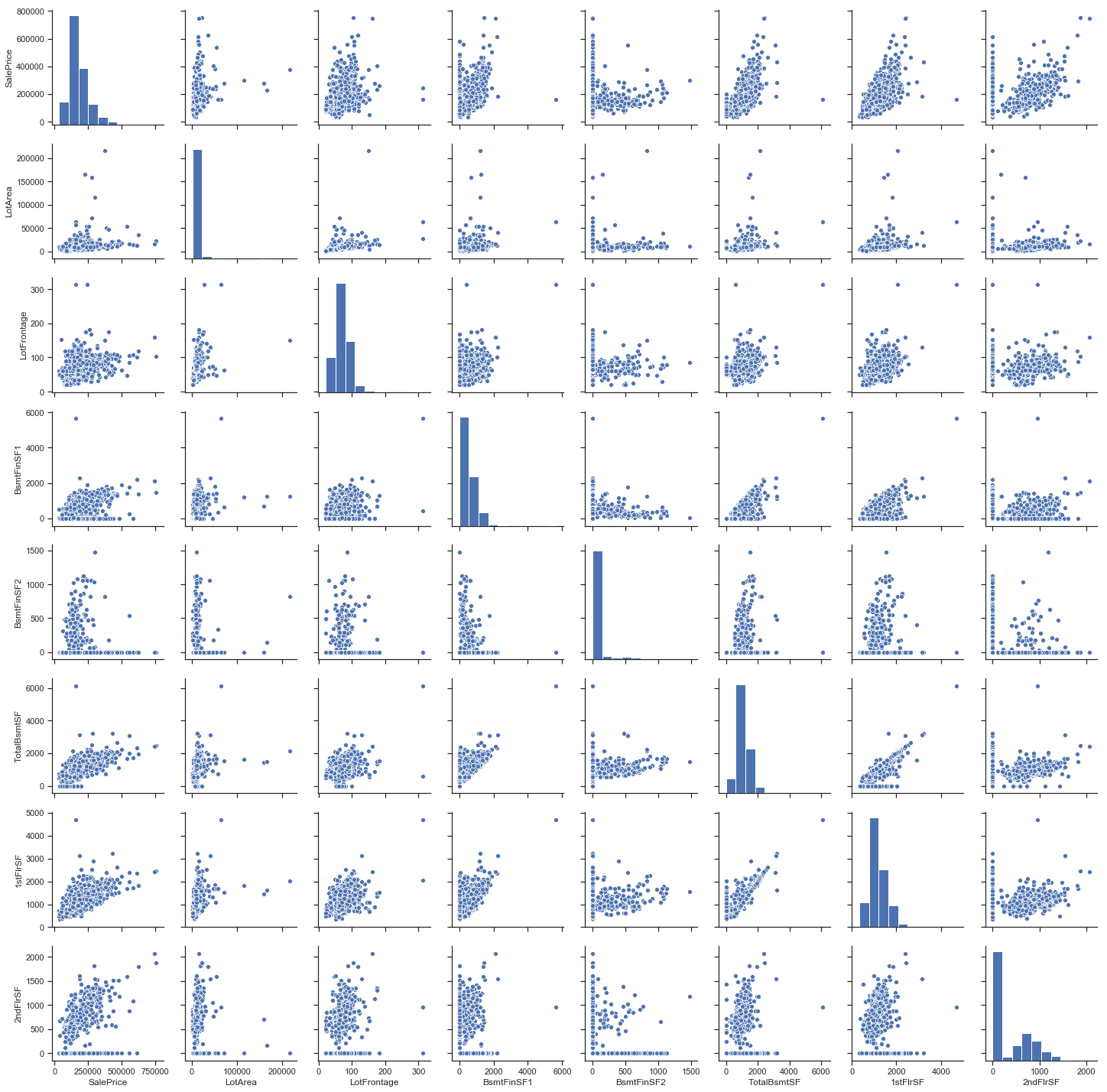
Từ biểu đồ trên, ta thấy các ngôi nhà có kiểu thiết kế 2 tầng có giá trị cao nhất, trong khi đó các ngôi nhà có kiểu thiết kế 1.5Unf có giá trị thấp nhất

Cuối cùng, ta thực hiện phân tích sự tương quan giữa các biến liên tục bằng 2 kiểu biểu đồ chính là Correlation matrix và scatter plot.



Từ correlation matrix này, ta thấy được SalePrice có tương quan mạnh nhất với các biến TotalBsmtSF, 1stFlrSF (hệ số tương quan 0.61) và tương quan rất yếu với biến, BsmtFinSF2 (hệ số tương quan 0.01)

Ta có thể xem thêm sự tương quan giữa các biến liên tục bằng biểu đồ scatter matrix dưới đây:



Từ biểu đồ này, ta có thể thấy rõ hơn sự tương quan giữa các biến và phân phối của các biến này.

Trên đây là phần phân tích dữ liệu dựa trên sử dụng các biểu đồ cơ bản, sử dụng ngôn ngữ lập trình Python. Qua phần phân tích này, ta có thể tìm được những ý tưởng để tạo ra những thuộc tính mới, biến đổi bộ dữ liệu, loại bỏ các biến không mang lại ý nghĩa để xây dựng các mô hình tốt hơn.