Housing Data Comprehensive Analysis

In this project, the famous Ames housing sales dataset, which has numerous features like Ground Living Area, Lot Size, Lot Shape, Basement Quality, Pool Quality, and Sales Price, was thoroughly analysed. Due to high number of features many interesting insights about the data were discovered. The time period of sales spans from January 2006 to July 2010 (fig 1a). The month-to-month (MoM) sales chart shows a pattern in the sales count after certain time periods. The annually decomposition of MoM chart confirms this fact; it shows strong annual seasonality (fig 1b). The monthly sales trend in this dataset roughly follows the national trend of March to July having highest sales and November, January and February having lowest [1].

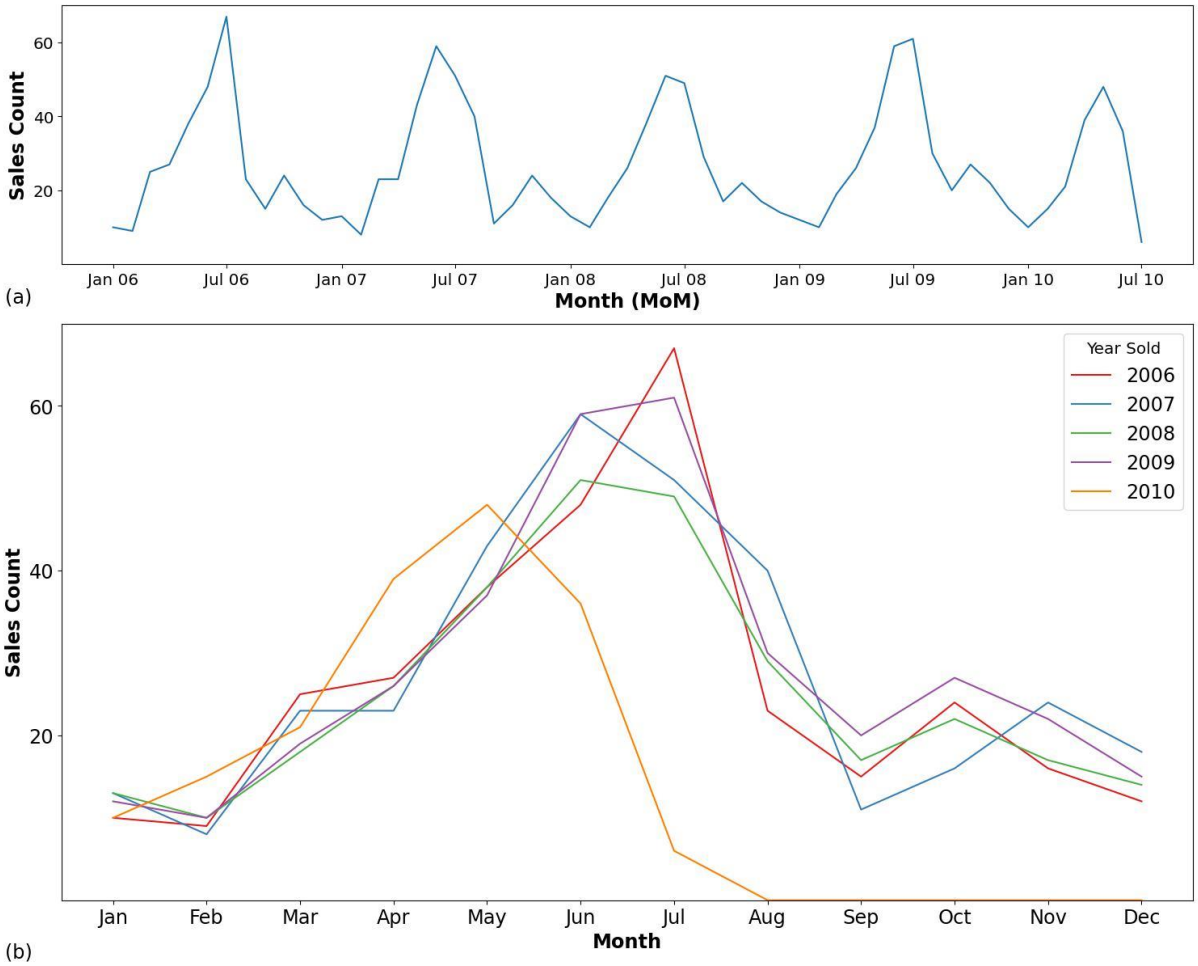

Figure 1: a. Month-on-Month time series of sales count b. Seasonal decomposition of Month-on-Month sales count

Most of the sales were witnessed by very few neighbourhoods, mainly North Ames, College Creek, Old Town and Edwards followed by Somerset, Gilbert and Northridge Heights (fig 2). Ames, with population of 66,427 according to 2020 census, is more of a college town than a metropolitan area with very few industries/ institutions other than Iowa State University, Ames National Lab, USDA research centres [2][3]. These facts were also reflected in types of houses featured in the dataset. Out of 1460 houses, 1220 were single family detached and 157 were townhouse (fig 3a). Very few houses had more than 2 levels (fig 3b). More than half of the houses were one story and 3 out of every 4

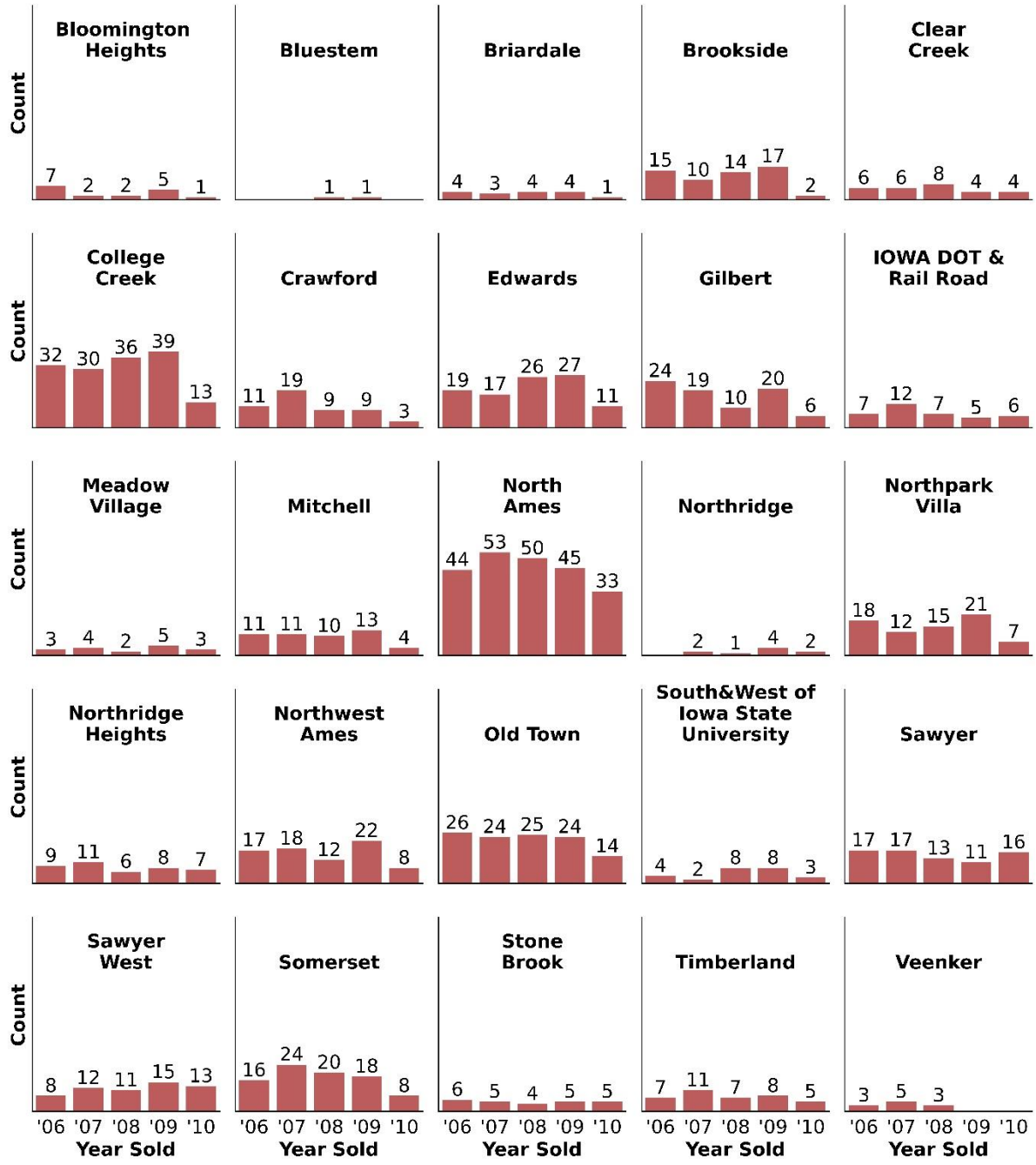

Figure 2 : Histogram of annual sales decomposed by neighbourhood

houses were either one story or two story. More than 75 percent of the houses were in low-density residential zones and only around 15 percent of them were in medium or high-density residential zones (fig 3c). Only 10 properties were in commercial zone.

For most of the neighbourhoods the annual sales count was below 20 throughout the five-year period. Almost 90 percent of the properties had near flat land contour (fig 4a). The topographical map of Ames shows some hilly region with slight depression at central and southern parts of city [4]. This can be attributed to around 10 percent of the houses having either banked, hillside or depressed contour. More than 96 percent of the property lots had either regular shape or were slightly irregular (fig 4b). Less than a percent of lots were completely irregular (fig 4c). More than 70 percent property lots had inside locality, around 18 percent had corner locality and less than 4 percent had frontage exposure to 2 or 3 sides. Almost all the property had access to paved street (fig 4d).

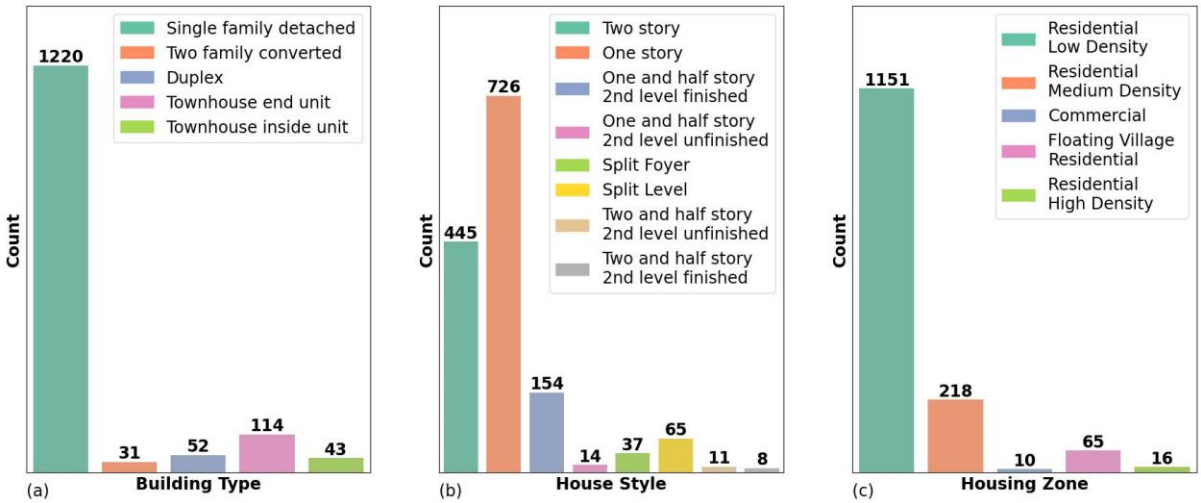

Figure 3 : Count plot of sales across a. building types b. House Style c. Housing Zone

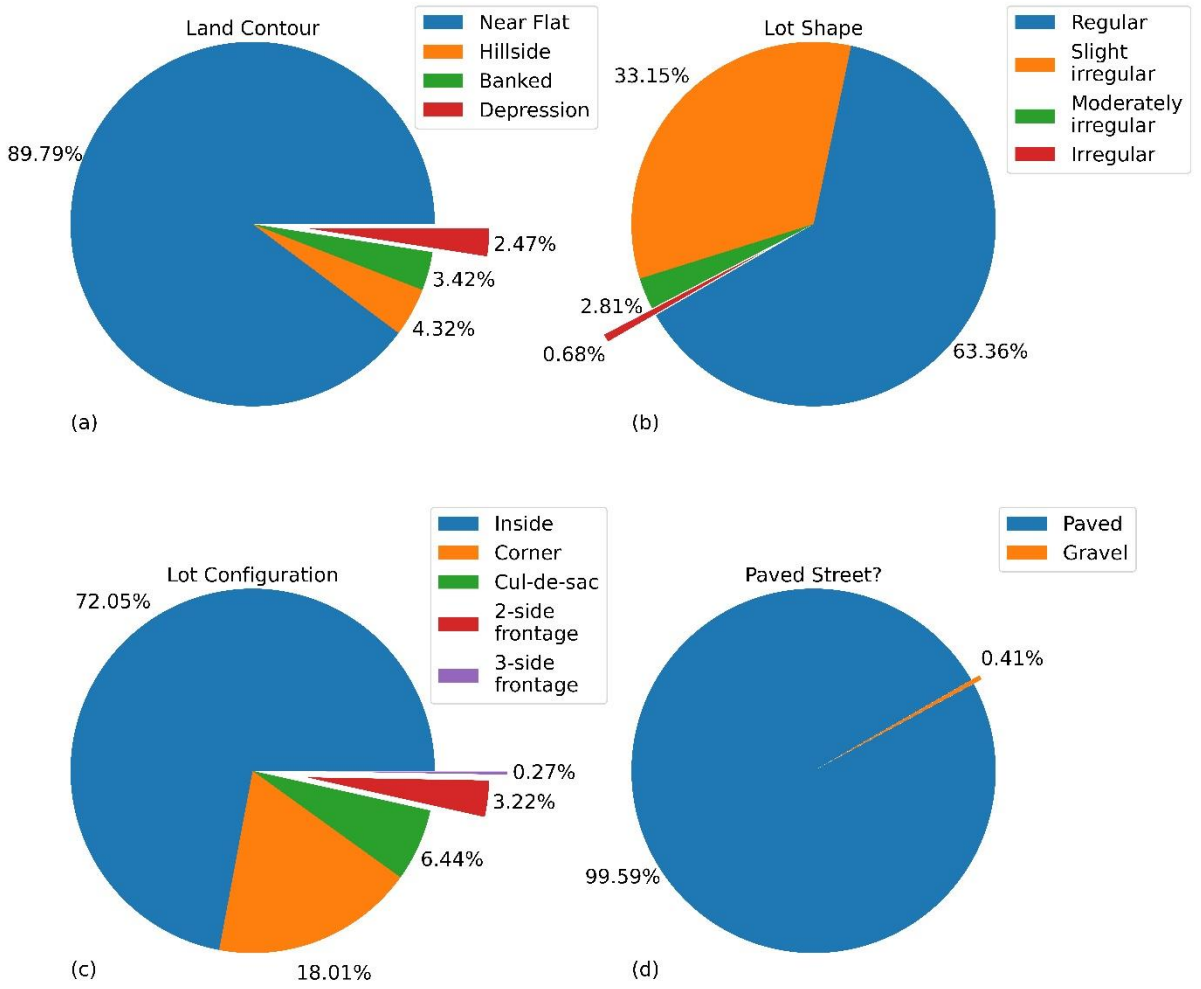

Figure 4: Pie chart of property features a. Land Contour b. Lot Shape c. Lot Configuration and d. Street Pavement

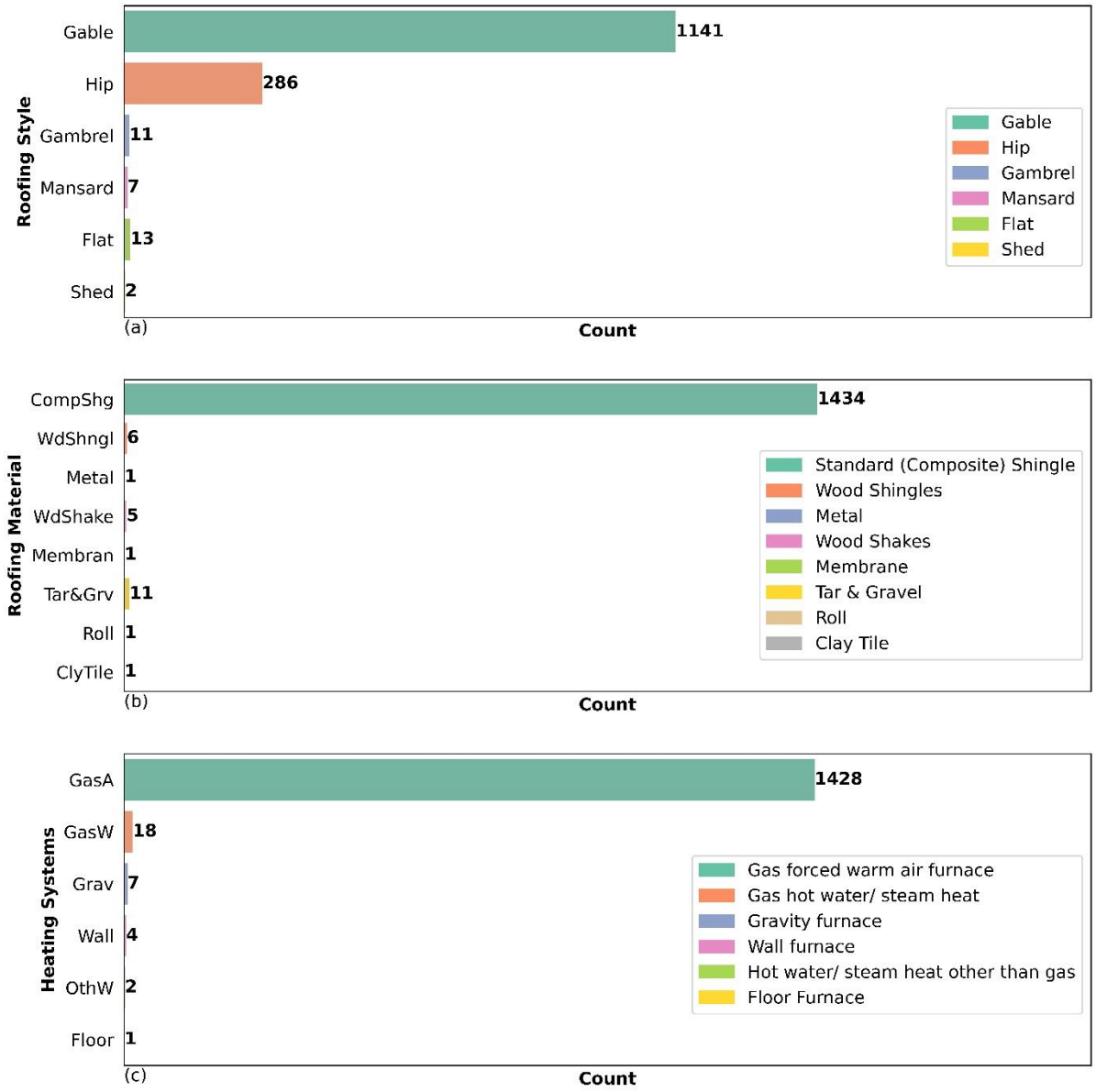

Figure 5: Count plot representing distribution of a. Roofing Style b. Roofing Material c. Heating System

It snows from Nov to Mar in Ames [3] and also, the tornado index is 337.09 [5]. This justifies the disproportionately high number of properties with Gable and Hip roofing style and shingled roofs (fig 5a, 5b) [6]. Furthermore, each and every house had heating system since this is an essential amenity in cold regions. Almost 98 percent of the houses were equipped with Gas forced warm air furnace.

More than half of the houses had fireplaces (fig 6a). Almost 95 percent of them had a garage and more than 97 percent had a basement (fig 6b, 6c). The overall weather and temperature profile of Ames is more characteristic of a tundra than a tropical region. This has translated in less than half a percent of homes having a pool (fig 6d). All the properties except one were equipped with all public utilities (fig 6e). Less than 4 percent of the properties had additional amenities like shed, 2nd garage and tennis court (fig 6f).

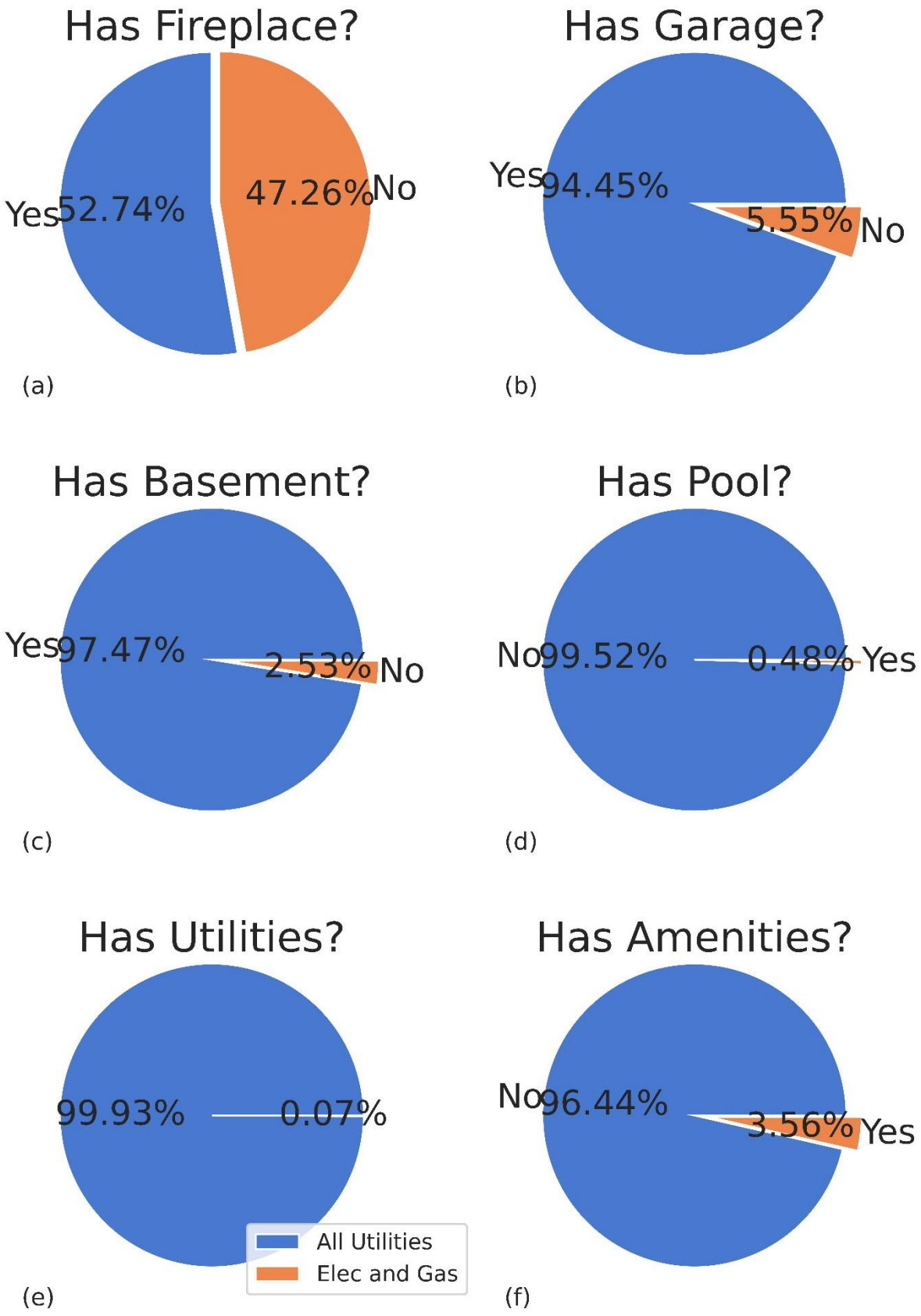

Figure 6: Pie chart representing present/absence of a. Fireplace b. Garage c. Basement d. Pool e. Utilities f. Amenities

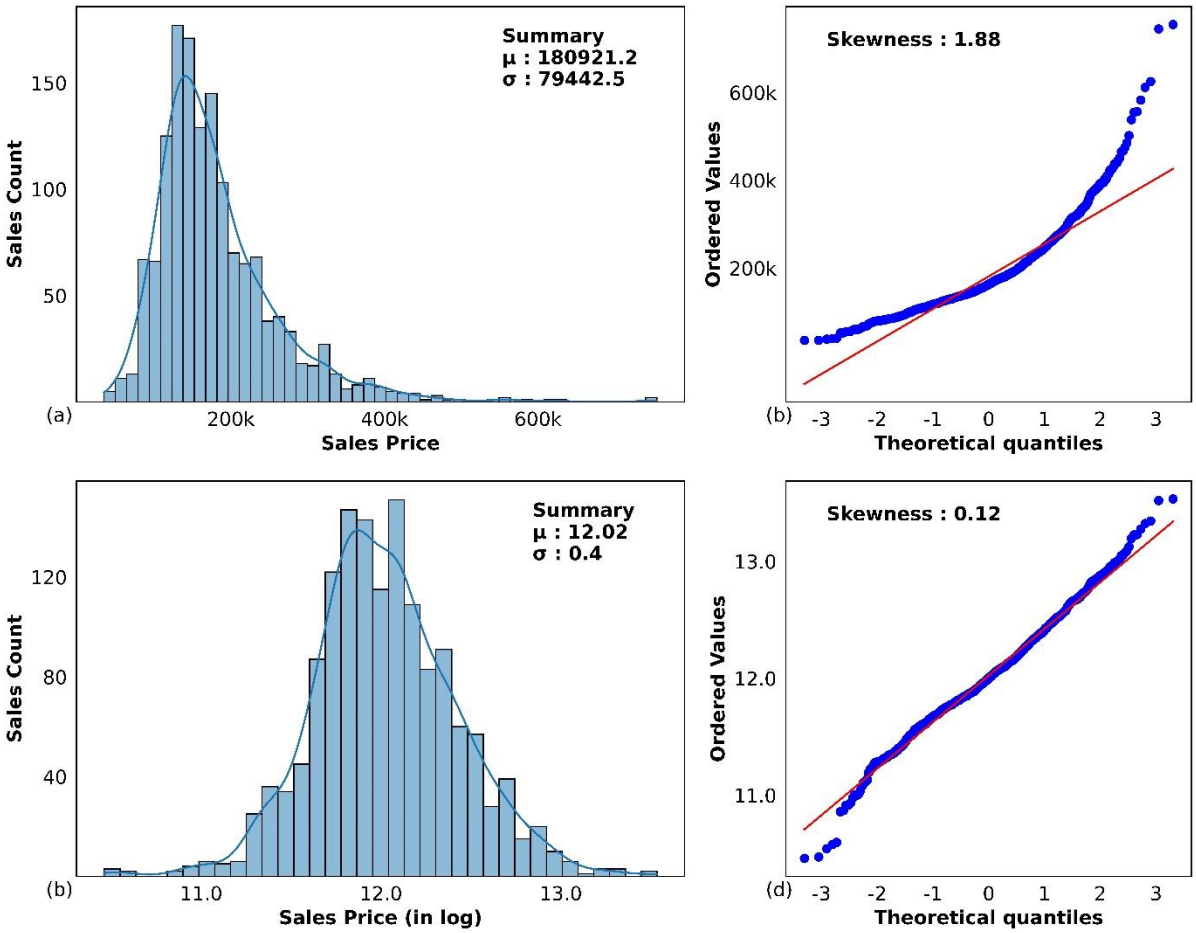

Figure 7: a. Distribution of Sales Price b. Q - Q plot of distribution of Sales Price c. Distribution of Sales Price on log scale d. Q - Q plot of distribution of Sales Price on log scale

Average selling price of the dataset was 180921.2 US Dollars with a standard deviation of 79442.5 US Dollars (fig 7a). The distribution of sales price feature was close to normal distribution and was skewed to right. The skewness was reflected in the Quantile – Quantile plot where empirical values vs theoretical quantile graph was concaved upward; it was flatter than ideal line before median value and it became steeper after that (fig 7b). The skewness of this distribution was 1.88. The most common technique to remove skewness is to use log transform of the given feature values and observe the distribution of log values. To avoid the occurrence of inf or highly negative log values, log1p transform was used ($p \rightarrow \log(1+p)$). The distribution of the log values was much closer to the normal distribution than the regular scale values (fig 7c). The QQ plot also reflect the near normal distribution with right skewness reduced to 0.12 (fig 7d). More than a percent of houses had selling price 3 standard deviations above mean which also contributed to the right skewness.

If a numerical variable spans over a semi-infinite range, from 0 to ∞, it is more possible to have right skewness than left skewness. The analysis of other important numerical features like ground living area, lot size, basement area added evidence to this hypothesis. The ground living area, basement area had right skewness of 1.37 and 1.52 respectively (fig 8a, 8b, 8c, 8d). Since some of the houses did not had basement, the frequency count at 0 was high. Few houses had lot area greater than 40000 sq ft. These outliers spiked the right skewness of the lot size feature to 12.2 (fig 8d, 8f).

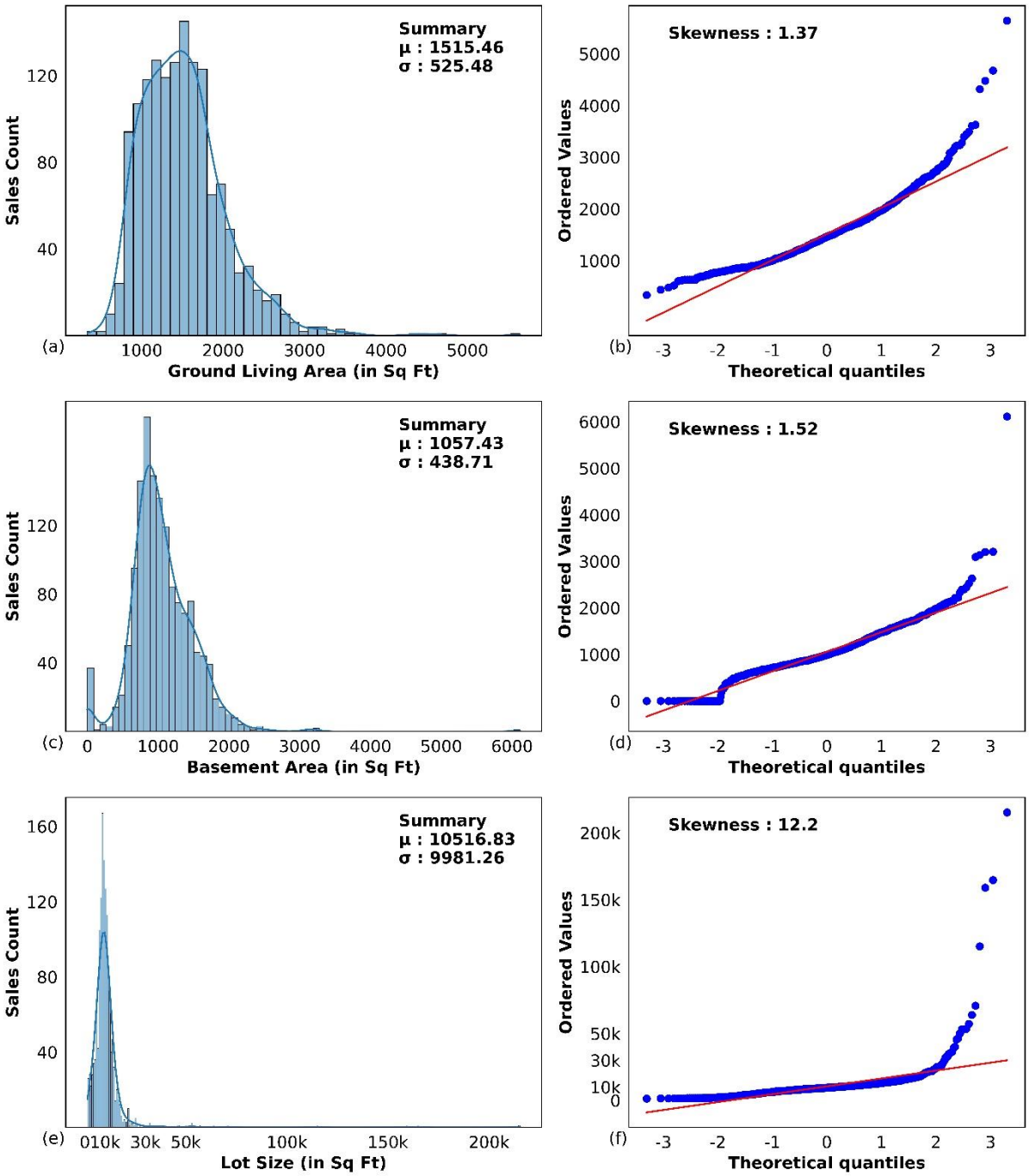

Figure 8 : a. Distribution of Ground Living Area b. Q - Q plot of distribution of Ground Living Area c. Distribution of Basement Area d. Q - Q plot of distribution of Basement Area e. Distribution of Lot Size f. Q -Q plot of distribution of Lot Size

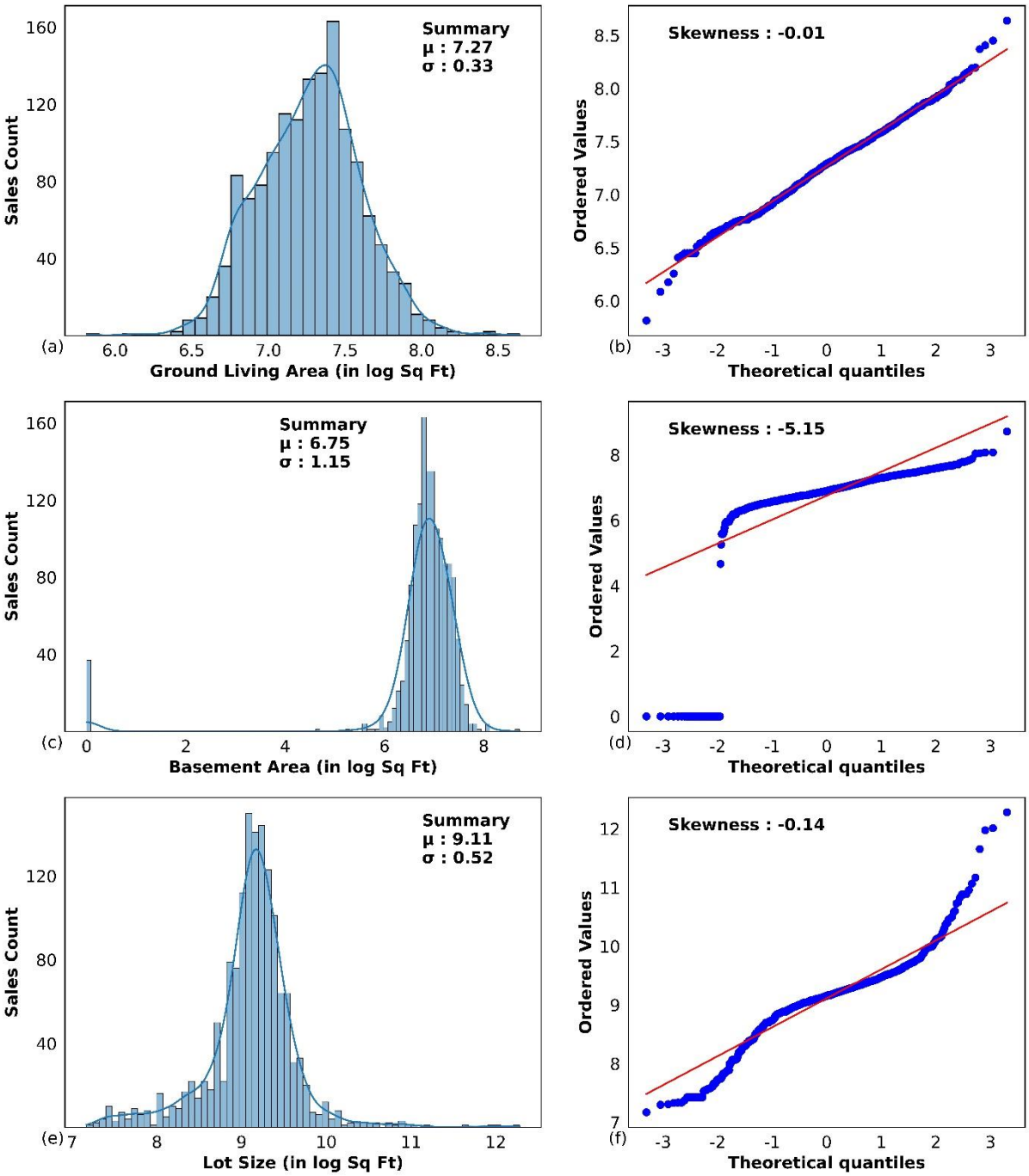

Figure 9 : a. Distribution of Ground Living Area on log scale b. Q - Q plot of distribution of Ground Living Area on log scale c. Distribution of Basement Area d. Q - Q plot of distribution of Basement Area on log scale e. Distribution of Lot Size on log scale f. Q -Q plot of distribution of Lot Size on log scale

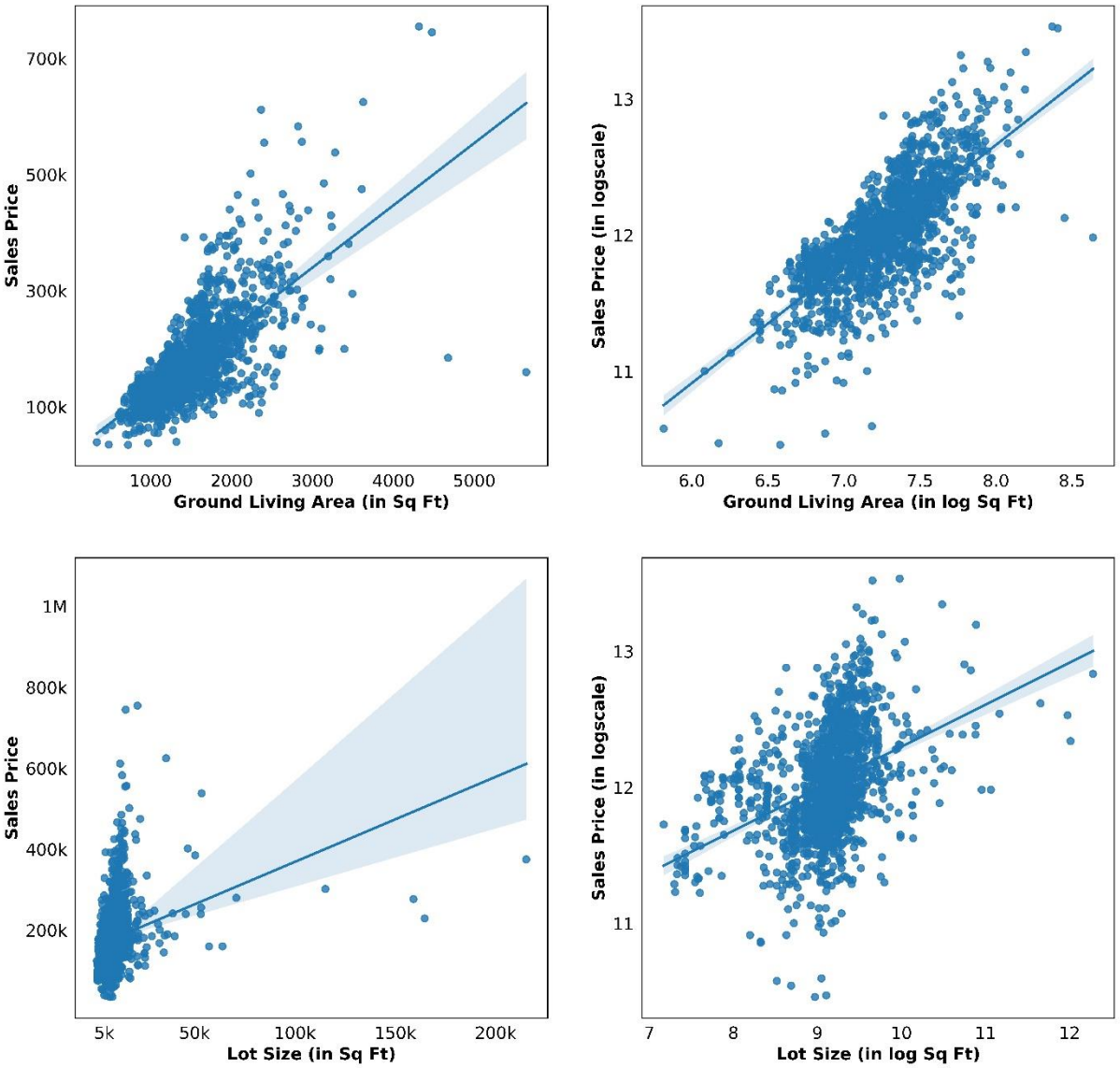

Figure 10: a. Regression plot between Sales Price and Ground Living Area b. Regression plot between log of Sale Price and log Ground Living Area c. Regression plot between Sales Price and Lot Size d. Regression plot between log of Sale Price and log Lot Size

Log1p transform made the ground living area distribution almost normal with small left skewness of 0.01 (fig 9a, 9b). However, the same transform brought about a huge shift from slight right skewness on normal scale to large left skewness on logscale (fig 9c, 9d). Careful inspection of the Q – Q plot of the basement area (normal scale) shows that the change of basement area from no-basement property to a property with basement is not smooth and gradual but rather abrupt and steep since no - basement property had zero basement area and most of the properties with basement had basement area similar to ground living area (which is quite obvious given ground being built on top of basement). Hence, in figure 8c, we see a tall bar at zero followed by really short bars eventually followed by taller ones. In log1p transform all the datapoints with zero basement area are clustered to zero and other data points form a near gaussian distribution centered around $\mathbf{e^{6.75}}$. This caused left skewness of 5.15 in the distribution. The lot size distribution had a slight left skewness of 0.14 after log1p transform (fig 9e, 9f). This is significant improvement.

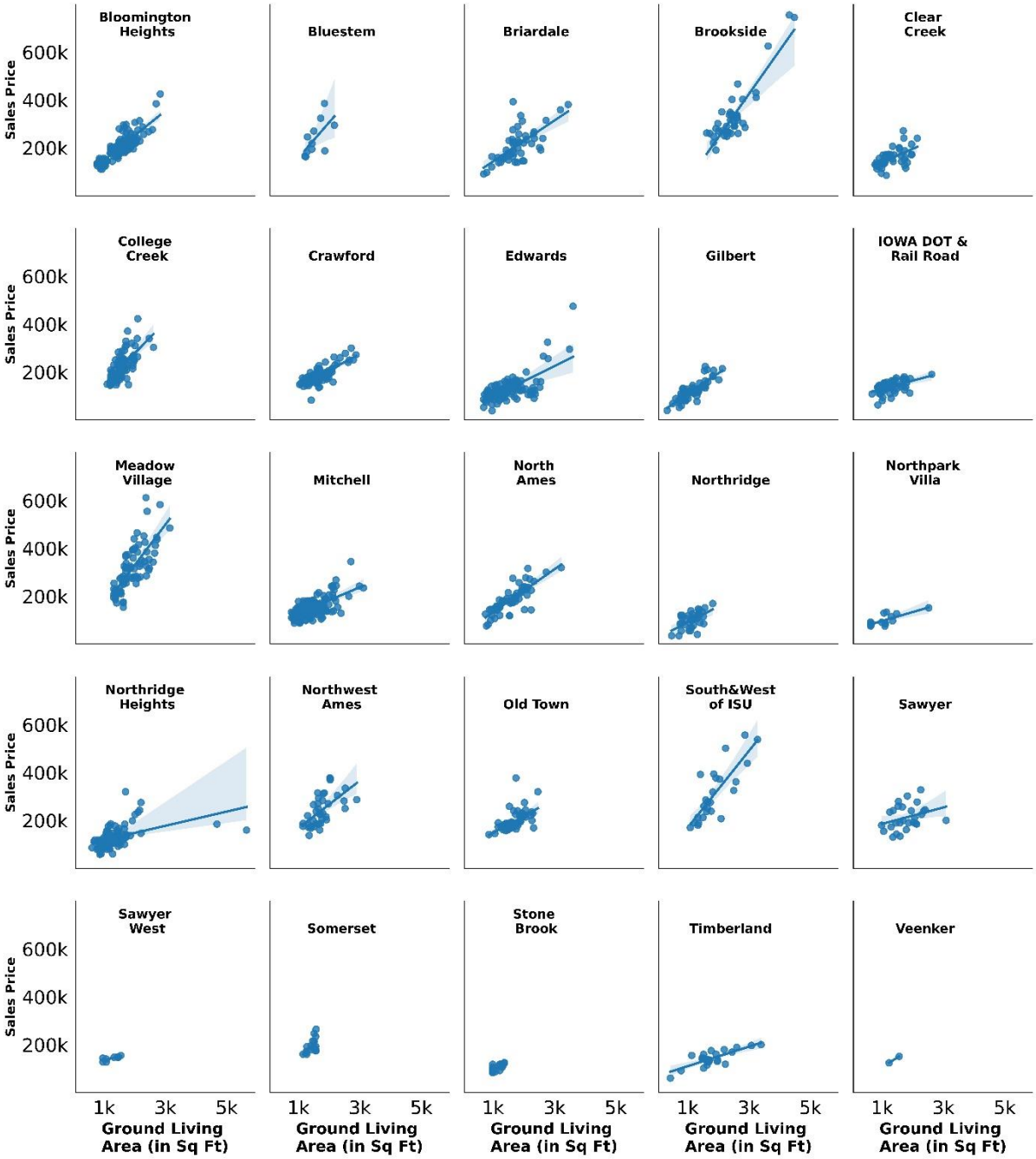

Figure 11 : Regression plot between Sales Price and Ground Living Area decomposed by Neighbourhoods

Furthermore, some of the features, believed to be impacting sales, were studied. The regression plot of Sales Price vs Ground Living Area reveals partial monotonic behavior where the houses with larger ground living area are more likely to be expensive than houses with smaller ground living area (fig 10a). The solid blue straight line represents the linear regression model that fits the data with light blue shaded region being the 95 percent confidence interval (CI). Since the spread of the data increases as we move in positive direction along both the variable axes, the confidence interval also gets wider simultaneously. When the same two features were plotted on log1p scale the extent of linearity increased considerably. Few datapoints had abnormally low selling price compared to datapoints with similar ground living area (fig 10b). Except for those points almost all the datapoints

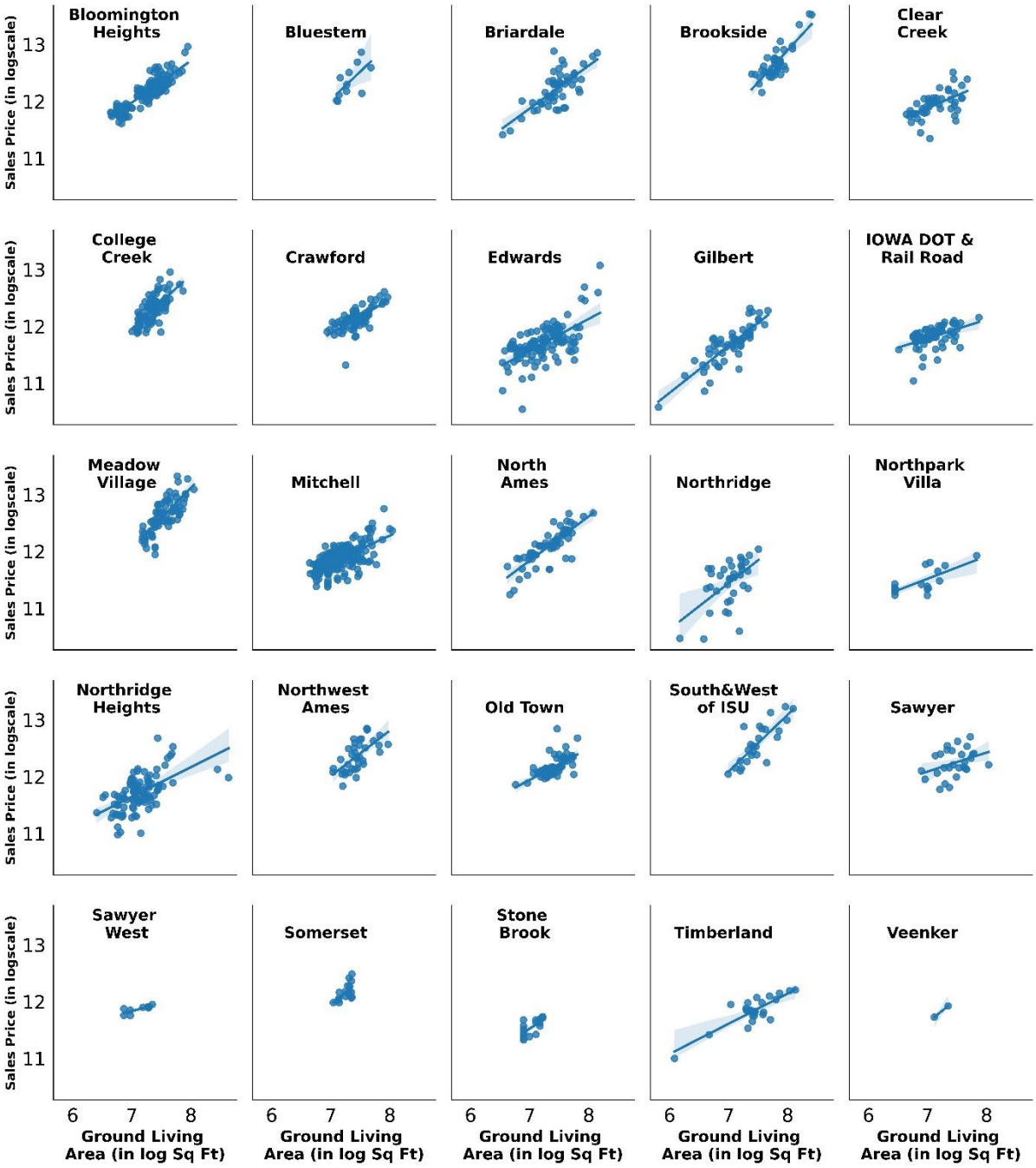

Figure 12 : Regression plot between log of Sales Price and log of Ground Living Area decomposed by Neighbourhoods

were clustered around the linear regression model. The CI of the regression model was also consistently narrow.

Another such feature, lot size was plotted against the sales price. Due to few irregularly shaped lots with exceptionally large size the linear regression model failed to accommodate the underlying data (fig 10c). Since these large lots had really low sales prices the CI of the regression model widened sharply in the positive direction of both axes. Log1p transform of the two did help in bringing more monotonous and linear relation between the variable but still the spread of datapoints was not consistent. Removing the datapoints with abnormal lot sizes and repeating the line fitting might reveal different results.

The regression plot between sales price and ground living area was further decomposed by neighborhoods (fig 11). In almost all the neighborhood datapoints were clustered around a small range of sales price and ground living area and rendered linear behavior. However, the extent of linearity and gradient of the regression line varied across the neighborhood. In Brookside, College Creek, Meadow Village, South & West of Iowa State University the Sales Price was more sensitive to the ground living area than in other neighborhoods. For all the neighborhoods the gradient of regression line was almost similar when fitted on log1p scale (fig 12). The CI of all the regression line improved for majority of the neighborhoods except the places where sales prices and ground living area were very low.

References:

1. https://www.attomdata.com/news/market-trends/home-sales-prices/attom-data-solutions-2021-best-days-to-sell-a-home-analysis/
2. https://www2.census.gov/programs-surveys/decennial/2020/data/01-Redistricting_File--PL_94-171/Iowa/
3. https://en.wikipedia.org/wiki/Ames,_Iowa
4. https://en-us.topographic-map.com/maps/cyk/Ames/
5. http://www.usa.com/ames-ia-natural-disasters-extremes.htm
6. https://www.nrel.gov/docs/fy13osti/56145.pdf