Section 1 – A Data Survey

Take some time to take a broad overview of the Ames housing data set. Read over the data documentation. What data do we have, and what is it supposed to represent?

The Ames Housing dataset presents a comprehensive collection of attributes related to residential properties in Ames, Iowa, aimed at estimating the sale price of homes.

Overview of the Data:

- **Basic Data Dimensions:** The dataset comprises 2,930 observations (or houses) and 82 variables (or features).
- **Target Variable:** `SalePrice` represents the price at which the property was sold, and it's the response variable that we aim to predict or explain.
- **Predictor Variables:** These can be broadly categorized into the following groups:

1. **Identification and Sale Info:**
   - `SID`: A simple serial number.
   - `PID`: Parcel identification number.
   - `SaleType`: Method of sale (e.g., warranty deed, foreclosure).
   - `SaleCondition`: Condition of the sale (e.g., normal, partial).
   - `MoSold`: Month the property was sold.
   - `YrSold`: Year the property was sold.
2. **Property Characteristics:**
   - MS Zoning (Nominal): General zoning classification of the sale
   - Lot Area (Continuous): Lot size in square feet
   - Lot Shape (Ordinal): General shape of the property
   - Land Contour (Nominal): Flatness of the property
   - Utilities (Ordinal): Type of utilities available
   - Lot Config (Nominal): Lot configuration
   - Land Slope (Ordinal): Slope of the property
   - House Style (Nominal): Style of dwelling
   - Year Built (Discrete): Original construction date
   - Roof Style (Nominal): Type of roof
   - Roof Matl (Nominal): Roof material
   - Exterior 1 (Nominal): Exterior covering on the house
   - Exterior 2 (Nominal): Exterior covering on the house (if more than one material)
   - Total Bsmt SF (Continuous): Total square feet of basement area
   - Total Bsmt SF (Continuous): Total square feet of basement area
   - 1st Flr SF (Continuous): Firstfloor square feet
   - 2nd Flr SF (Continuous): Secondfloor square feet
   - Gr Liv Area (Continuous): Above grade living area square feet
   - Yr Sold (Discrete): Year Sold
   - Sale Condition (Nominal): Condition of sale

3. **Sale Property Rating:**
   This considers all the different ratings given to the different parts of the property.
   - Overall Qual (Ordinal): Rates the overall material and finish of the house
   - Overall Cond (Ordinal): Rates the overall condition of the house
4. **Nearby Facilities:**
   This considers facilities available from the property.
   - Lot Frontage (Continuous): Linear feet of street connected to the property
   - Street (Nominal): Type of road access to the property
   - Alley (Nominal): Type of alley access to the property
   - Neighborhood (Nominal): Physical locations within Ames city limits
   - Condition 1 (Nominal): Proximity to various conditions
   - Condition 2 (Nominal): Proximity to various conditions (if more than one is present)

Potential Analysis:

1. Given the broad range of features, we can conduct various analyses:
2. Descriptive Statistics: To understand distributions, central tendencies, and variability.
3. Correlation and Association Analysis: To determine which variables are most related to `SalePrice`.
4. Predictive Modelling: To build models predicting `SalePrice` based on other attributes.
5. Segmentation and Clustering: To identify clusters of similar houses.

The Ames Housing dataset is a rich and multifaceted collection that can offer valuable insights into the housing market of Ames, Iowa. It captures both the physical attributes of properties and the surrounding environmental and locational factors, offering a holistic perspective on what drives property values.

It's essential to conduct thorough exploratory data analysis (EDA) before model building to understand data distributions, handle outliers, address missing values, and potentially create new features or transform existing ones to improve model performance.

In the linear regression component of this course we want to build linear regression models to predict the value of a property (or home). Do we have the right data to properly address our problem? Are there observations in the data that should be excluded?

Do we have the right data to properly address our problem?

Continuous Variables: Variables like `LotArea`, `GrLivArea`, and `GarageArea` directly influence a property's value. For instance, a home's value typically scales with its size. The year it was built or remodeled (`YearBuilt`, `YearRemodel`) provides information on the age and potentially the style and condition of the property.

Discrete Variables: These give an enumerated perspective on the property's features. For example, the overall quality (`OverallQual`) and condition (`OverallCond`) can significantly influence buyer decisions. Likewise, the number of bedrooms (`BedroomAbvGr`), bathrooms

(`FullBath`, `HalfBath`), and fireplaces (`Fireplaces`) are crucial elements homebuyers consider.

Are there observations in the data that should be excluded?

Missing Values: We have some columns with missing values, like `LotFrontage` (490 missing) and `GarageYrBlt` (159 missing).

Here's what can be done:

`LotFrontage`: Missing values can be imputed using various techniques such as median imputation or modelbased imputation. However, if missing values correspond to properties without frontage (e.g., homes without direct street access), it would be meaningful to assign a value of 0.

`GarageYrBlt`: Missing values here could imply properties without garages. This can be crossverified with `GarageArea` or `GarageCars`. If they also indicate no garage, we can assign a special value or impute based on another relevant variable like `YearBuilt`. Alternatively, we could drop the records with missing values.

Variables like `MasVnrType`, `MasVnrArea`, `BsmtQual`, and others with fewer missing values can be dropped for lack of domain expertise.

NonResidential Properties: If the dataset contains entries for nonresidential properties, it might skew our predictive model since the sale price determinants for commercial or industrial properties differ vastly from residential ones. Filtering out nonresidential sales or any outlier properties that don't represent typical homes would be advisable.

Redundancy: Some variables might offer redundant information. It's essential to check correlations among predictors to ensure multicollinearity doesn't affect the regression model. For instance, having both `GarageCars` (how many cars can fit) and `GarageArea` (size of the garage) might be redundant as they would be highly correlated.

In conclusion, while the dataset provides a robust foundation for building a linear regression model to predict property values, careful preprocessing is crucial. Addressing missing values, eliminating potential outliers or nonrelevant observations, and handling redundant information will ensure a more accurate and generalizable model.

What kinds of problems can we properly address given the data that we have? In particular, if we were to build a regression model with the variable SalePrice as the response variable (Y), what types of properties would we be valuing? Do we need to be careful about what we are doing here?

Given the data at hand, there are various problems and questions that we can address:

1. Predictive Modelling:

Price Prediction: The most straightforward use of the data is to build a predictive model for the sale price of a property. This would be a regression problem, with `SalePrice` as the response variable.

2. Property Characterization:

   Type of Property: With variables like `BldgType`, `HouseStyle`, and `MSZoning`, you can analyze different kinds of residential properties: singlefamily homes, townhouses, etc.

   Age and Condition: `YearBuilt` and `OverallCond` can provide insights into the age and overall condition of a property.

   Location and Surroundings: Variables such as `Neighborhood`, `Condition1`, and `Condition2` provide information about the property's location and its immediate surroundings.

3. Feature Importance:

   Determine which features have the most significant influence on the `SalePrice`. Is it the square footage, the neighborhood, the year it was built, or some other factor?

4. Time Series Analysis:

   Given that we have `YrSold`, we could look into how property prices have changed over time.

5. Analysis of Other Influencing Factors:

Examine how factors like proximity to various conditions, type of road access, or the type of utilities available affect the property's price.

What types of properties would we be valuing?

We would primarily be valuing residential properties in Ames, Iowa, across different neighbourhoods, conditions, and features. This includes various types of dwellings from singlefamily homes to townhouses, characterized by their size, age, style, quality, and more.

Do we need to be careful about what we are doing here?

Absolutely, for several reasons:

1. Handling Missing Data: As previously discussed, there are several variables with missing values. Imputing these incorrectly or neglecting them can lead to skewed results.
2. Outliers: Extreme values or outliers can disproportionately influence a regression model. Proper detection and handling of these (either through transformation, imputation, or removal) is crucial.

3. Multicollinearity: Some of the features might be highly correlated with each other, causing multicollinearity, which can destabilize regression models. Detecting and addressing multicollinearity (possibly through feature selection, combination, or regularization techniques) is essential.
4. Model Assumptions: Linear regression has several assumptions, including linearity, homoscedasticity, and normality of errors. Violations of these assumptions can lead to unreliable predictions. Regular diagnostics should be performed to ensure these assumptions are met or, if not, that appropriate adjustments or alternative models are considered.
5. External Validity: While the dataset provides a comprehensive look into properties in Ames, Iowa, the model might not generalize well to properties outside of this region or during significantly different time periods.
6. Influence of Categorical Variables: There are many categorical variables in the dataset. Proper encoding (e.g., onehot encoding) and understanding their impact are vital.

In summary, while the dataset is ripe for regression modelling, due diligence in preprocessing, modelling, and validating results is essential to draw meaningful and accurate conclusions.

Section 2 – Define the Sample Population

When building statistical models we have to define the population of interest, and then sample from THAT population. Frequently we will not actively perform the sampling function. Instead, the data will be made available and we will have to sample from it retrospectively, i.e. we will need to carve out the population of interest. In our case the objective of our application is to be able to provide estimates of home values for 'typical' homes in Ames, Iowa. We may not be able to define what 'typical' is, but we can use the data to find out what is atypical. Any values which are not atypical are then considered to be typical.

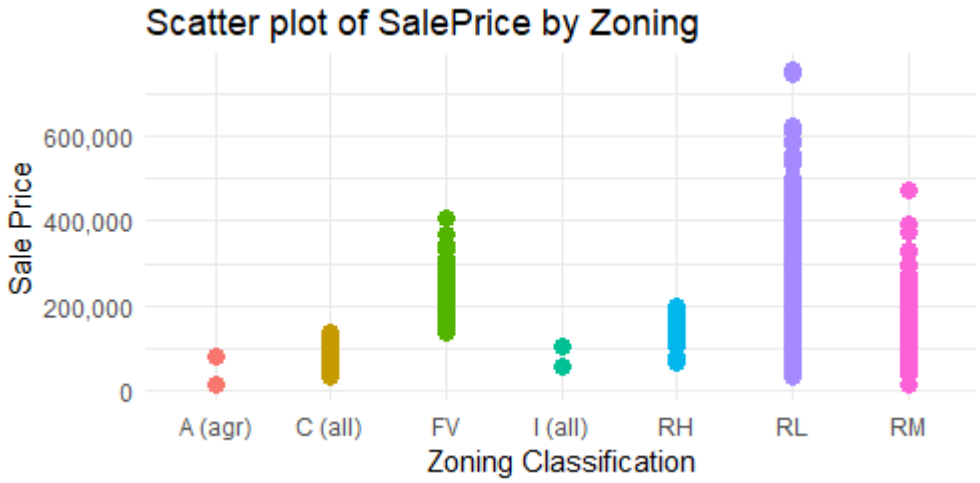
Define the appropriate sample population for your statistical problem. Hint: We are building regression models for the response variable SalePrice. Are all properties the same? Would we want to include an apartment building in the same sample as a singlefamily residence? Would we want to include a warehouse or a shopping center in the same sample as a singlefamily residence? Would we want to include condominiums in the same sample as a singlefamily residence?

When pinpointing our target population, specifically the 'typical' homes in Ames, Iowa, it's imperative to delineate what sets apart an 'atypical' home. This distinction helps us to sieve out the unconventional properties and center our attention on the mainstream or 'typical' homes. Our strategy to ascertain this is rooted in our observations from the boxplots for SalePrice and LivingArea, supplemented by the scatterplots of SalePrice against variables like Zoning, SaleCondition, BldgType, and Functional. Here's our proposed roadmap to refine our dataset for this purpose.

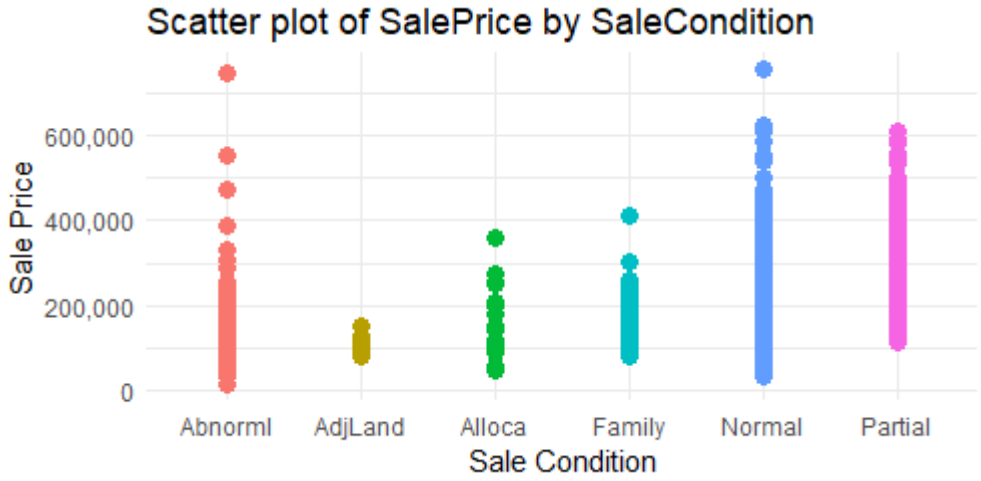
Define your sample using ‘drop conditions’. Create a waterfall for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population.

Defining 'Typical' Homes in Ames, Iowa:

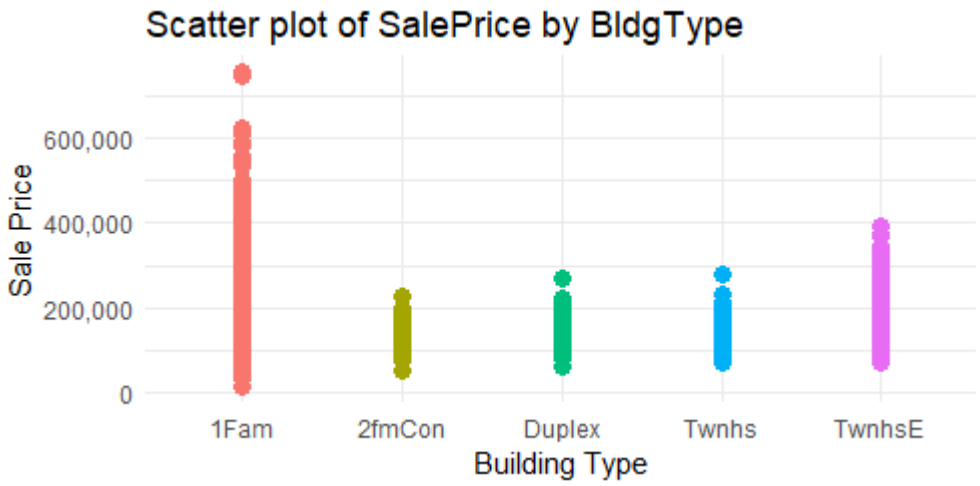
1. Residential Properties: We will retain only those properties that have a `Zoning` value of 'RH', 'RL', or 'RM'. This ensures that we are looking at residential properties only, filtering out commercial and other nonresidential zones.



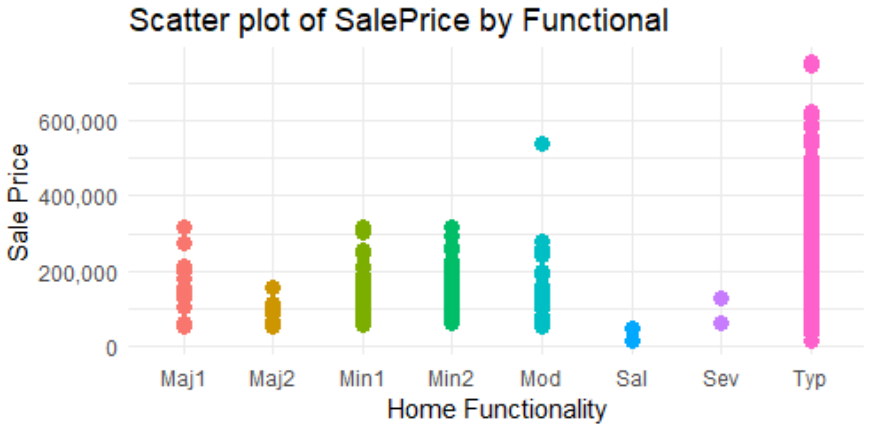
2. Normal Sale Condition: We will filter out properties where the `SaleCondition` is 'Normal'. This ensures that the dataset represents typical sales and excludes properties that may have been sold between family members or under other unusual circumstances.



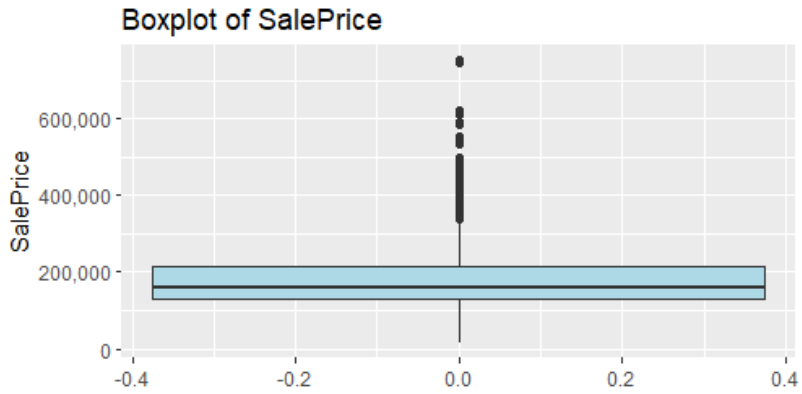
3. Single Family Houses: By selecting homes with `BldgType` as '1Fam', we ensure our dataset represents singlefamily homes. This excludes properties like townhouses, duplexes, and other building types that might not represent the typical homebuyer's preference in Ames.



4. Typical Home Functionality: By retaining only homes with a 'Typical' value for the `Functional` variable, we're excluding homes with any kind of damage or any other functionality issues. This is critical because homes with significant damages or functionality problems may not be representative of a 'typical' home.

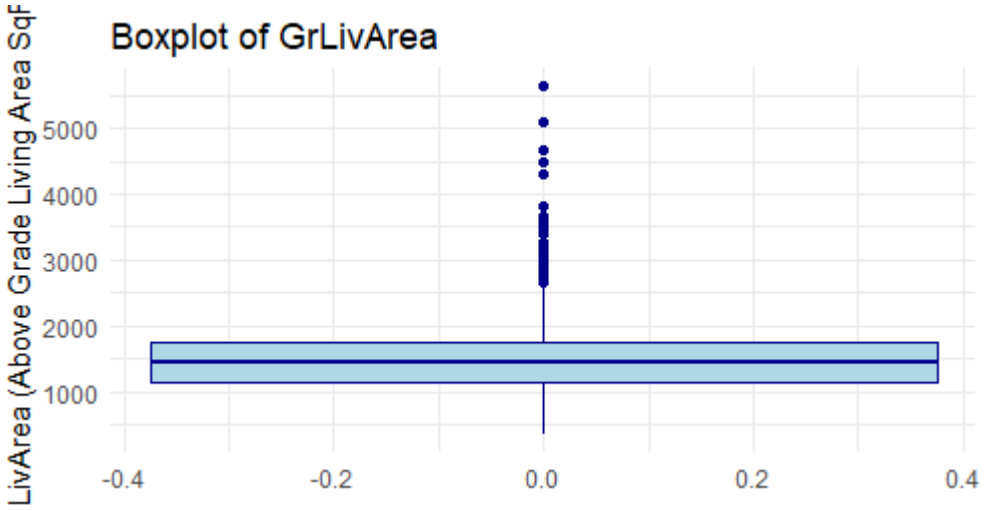


5. Price Constraint: From the provided insight, properties priced above $ 339500 (Q3+1.5*IQR) are considered outliers. By filtering out these properties, we're ensuring that we're modelling prices that are representative of the majority of homes in Ames.

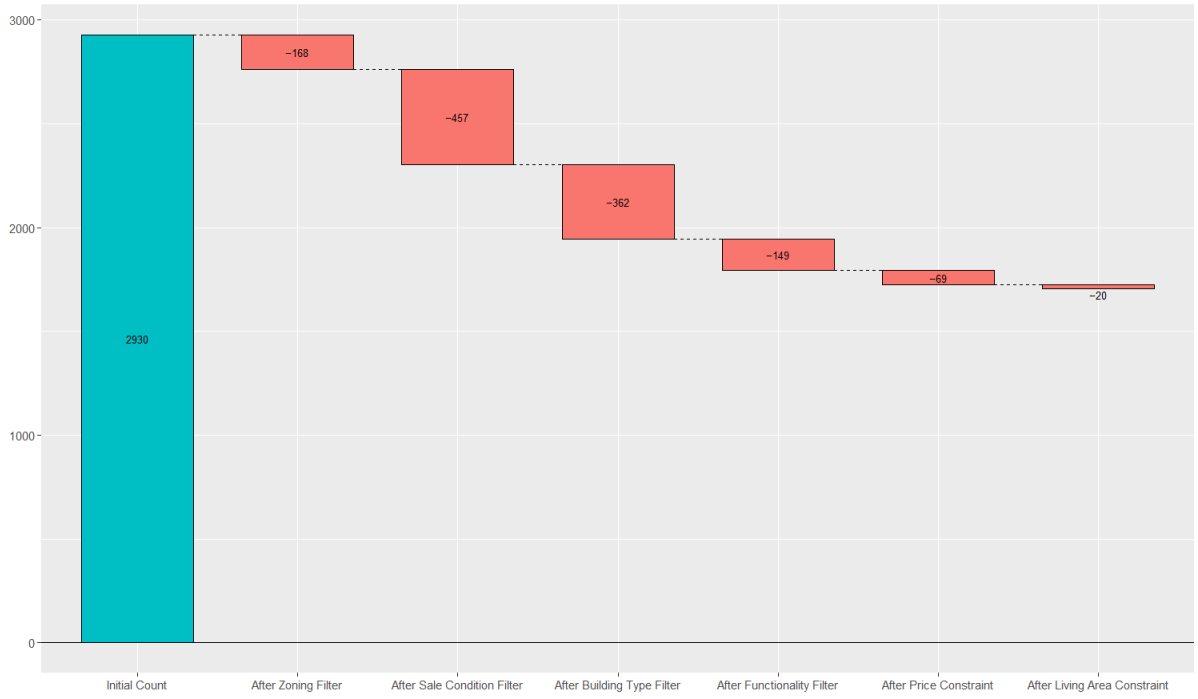


6. Living Area Constraint: By retaining properties with a GrLivArea of 2667(Q3+1.5*IQR) sq. ft. or less, we're further refining our dataset to exclude atypically large homes. Our

observation from the box showed that properties above this threshold were unusual compared to the rest.



Define your sample using 'drop conditions'. Create a waterfall for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population.



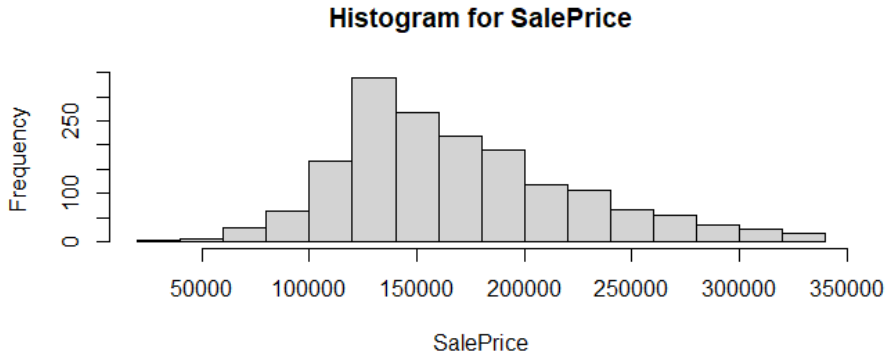
The sample population (filtered dataframe) has 1705 rows and 82 columns.

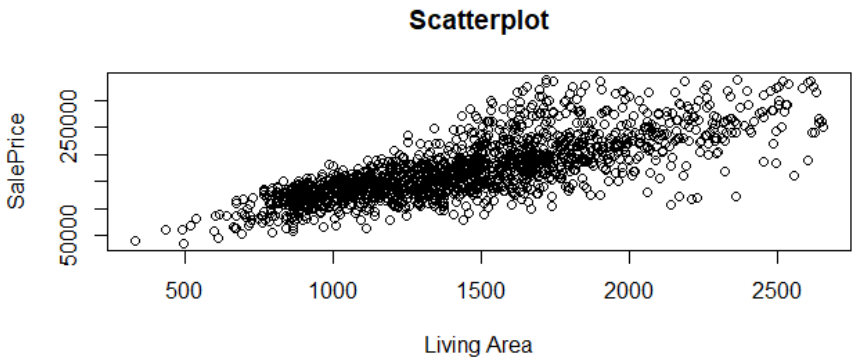
Section 3 – A Data Quality Check

In practice your data will not be 'clean'. You will need to examine your data for errors and outliers. Errors will not always show as outliers, and outliers are not necessarily errors.
If you have a data dictionary that states the set of proper values for each field, then you will want to check your data against the data dictionary.

If you do not have a data dictionary, then you will need to reason and explore your way to a proper data set.

1. Missing Values: There are 338 missing values in the `LotFrontage` column.
    a. The `MasVnrType` and `MasVnrArea` columns each have 9 missing values.
    b. The `BsmtExposure` column has 2 missing values.
    c. The `BsmtFinType2` and `Electrical` columns each have 1 missing value.
    d. The `GarageYrBlt` column has 61 missing values.
2. Outlier Detection: Outliers concerning `SalePrice` have been previously removed.
3. Duplicate Rows: There are no duplicate rows in the dataframe.
4. Consistency Check: YearBuilt < YearRemodel for all records.
5. Other Anomalies: SalePrice > 0 for all records.
6. Visual Exploration:





Consider the use of R functions table(), summary(), quantile(), mean(), sd(). Also consider the use of lapply() to vectorize R computations across an R data frame when forming your data quality check. Pick twenty variables that you want to consider and run a data quality check on these twenty variables.

Here is the output summary in words for the selected variables:

Continuous/Discrete Variables:

1. LotArea:
   Minimum: 2887
   1st Quartile: 8050
   Median: 9600
   Mean: 10343
   3rd Quartile: 11455
   Maximum: 159000
2. YearBuilt:
   Minimum: 1872
   1st Quartile: 1950
   Median: 1968
   Mean: 1967
   3rd Quartile: 1995
   Maximum: 2010
3. TotalBsmtSF:
   Minimum: 0.0
   1st Quartile: 803.8
   Median: 953.0
   Mean: 1009.1
   3rd Quartile: 1180.2
   Maximum: 3206.0
4. GrLivArea:
   Minimum: 334
   1st Quartile: 1080
   Median: 1405
   Mean: 1426
   3rd Quartile: 1694
   Maximum: 2654
5. PoolArea:
   Minimum: 0.0000
   1st Quartile: 0.0000
   Median: 0.0000
   Mean: 0.7183
   3rd Quartile: 0.0000
   Maximum: 648.0000
6. YrSold:
   Minimum: 2006
   1st Quartile: 2007
   Median: 2008
   Mean: 2008
   3rd Quartile: 2009
   Maximum: 2010

7. YearRemodel:
   Minimum: 1950
   1st Quartile: 1962
   Median: 1990
   Mean: 1982
   3rd Quartile: 2002
   Maximum: 2010
8. LotFrontage:
   Minimum: 30.00
   1st Quartile: 60.00
   Median: 70.00
   Mean: 71.36
   3rd Quartile: 80.00
   Maximum: NA's (There are 338 missing values)

Nominal/Ordinal Variables:

1. Fence:
   GdPrv: 82
   GdWo: 79
   MnPrv: 238
   MnWw: 10
   NA: 1295
2. Utilities:
   AllPub: 1704
3. HouseStyle:
   1.5Fin: 204
   1.5Unf: 17
   1Story: 853
   2.5Fin: 2
   2.5Unf: 14
   2Story: 482
   SFoyer: 39
   SLvl: 93
4. RoofStyle:
   Flat: 9
   Gable: 1374
   Gambrel: 17
   Hip: 296
   Mansard: 5
   Shed: 3
5. Street:
   Grvl: 2
   Pave: 1702
6. Alley:
   Grvl: 77

NA: 1613
Pave: 14

7. Neighborhood: (Counts for multiple neighborhoods)
   Blmngtn: 1
   BrkSide: 86
   ClearCr: 30
   CollgCr: 210
   Crawfor: 67
   Edwards: 105
   Gilbert: 126
   IDOTRR: 46
   Mitchel: 78
   NAmes: 325
   NoRidge: 41
   NridgHt: 37
   NWAmes: 103
   OldTown: 156
   Sawyer: 102
   SawyerW: 82
   Somerst: 21
   StoneBr: 4
   SWISU: 30
   Timber: 41
   Veenker: 13
8. Condition1:
   Artery: 60
   Feedr: 98
   Norm: 1447
   PosA: 14
   PosN: 29
   RRAe: 18
   RRAn: 29
   RRNe: 4
   RRNn: 5
9. Condition2:
   Artery: 1
   Feedr: 10
   Norm: 1691
   RRNn: 2
10. OverallQual:
    1: 1
    2: 5
    3: 18
    4: 111

5: 548
6: 481
7: 370
8: 148
9: 21
10: 1

11. OverallCond:
5: 834
6: 367
7: 293
8: 122
9: 30

12. SaleCondition:
Normal: 1704
13. GarageType:
2Types: 7
Attchd: 1027
Basment: 15
BuiltIn: 100
CarPort: 3
Detchd: 491
NA: 61

Section 4 – An Initial Exploratory Data Analysis

Pick ten variables from the twenty variables from your data quality check to explore in your initial exploratory data analysis. Perform an initial exploratory data analysis. How do we perform an exploratory data analysis for continuous versus discrete (or categorical) data? Consider the use of scatterplots, scatterplot smoothers such as LOESS, and boxplots to produce relevant graphics when appropriate.

Selecting ten variables from the dataset to perform an initial exploratory data analysis (EDA). We will choose a mix of continuous and discrete (categorical) variables for analysis. Here are the ten variables we'll explore:

1. LotArea (Continuous)
2. YearBuilt (Continuous)
3. TotalBsmtSF (Continuous)
4. GrLivArea (Continuous)
5. OverallQual (Categorical)

6. OverallCond (Categorical)
7. SaleCondition (Categorical)
8. GarageType (Categorical)
9. Fence (Categorical)
10. HouseStyle (Categorical)

Now, let's perform an initial exploratory data analysis for these variables. We'll use appropriate graphical methods for continuous and discrete (categorical) data.

For Continuous Variables (LotArea, YearBuilt, TotalBsmtSF, GrLivArea):

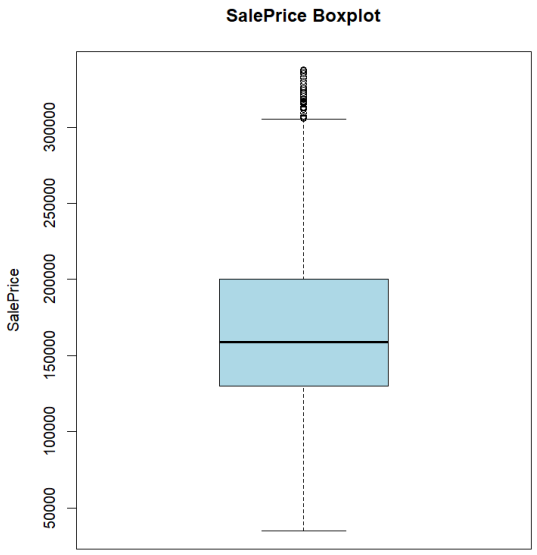
1. Histograms: We can create histograms to visualize the distribution of these continuous variables.

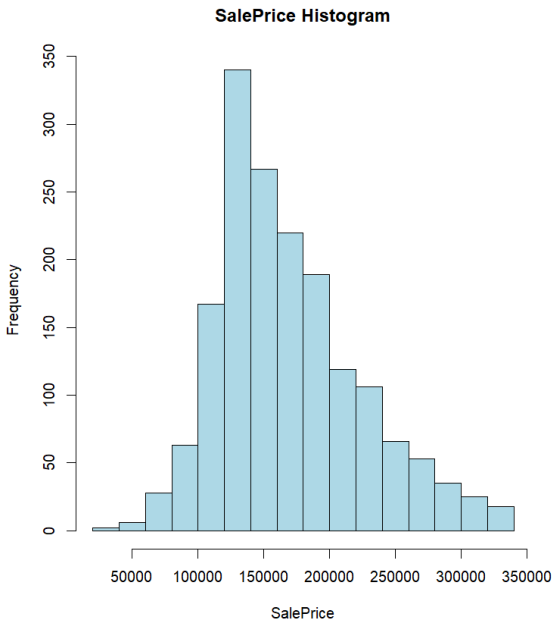
2. Boxplots: Boxplots can help us identify outliers and understand the spread of data.

3. Scatterplots: For pairwise comparisons, we can use scatterplots to explore relationships between these continuous variables.

EDA of Continuous Variables:

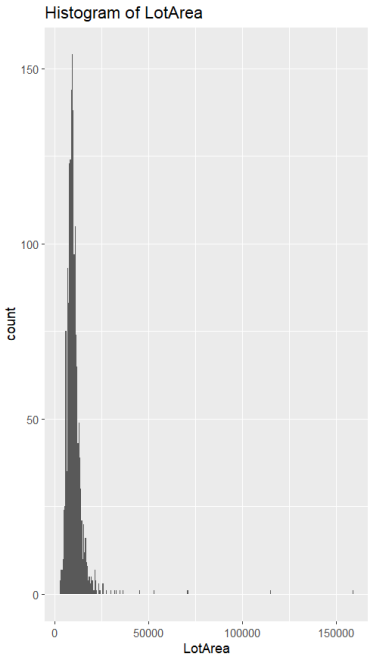
1. SalePrice

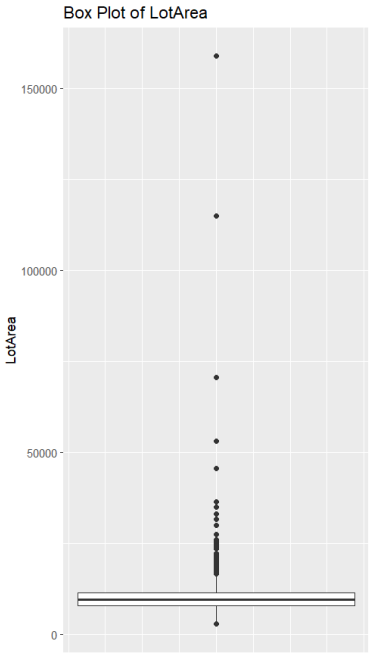


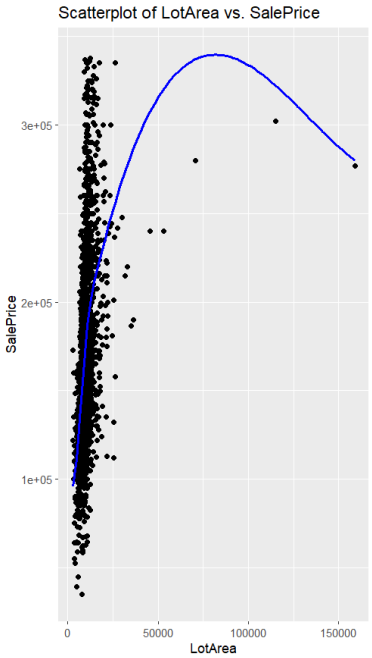



```
Variable: SalePrice
Lower Bound: 25000
Upper Bound: 305000
Number of Outliers Below: 0
Number of Outliers Above: 40
Statistical Summary:
Min.: 35000 1st Qu.: 130000 Median: 159000 Mean: 169992.064553991 3rd Qu.:
200000 Max.: 337500
```

2. Lot Area

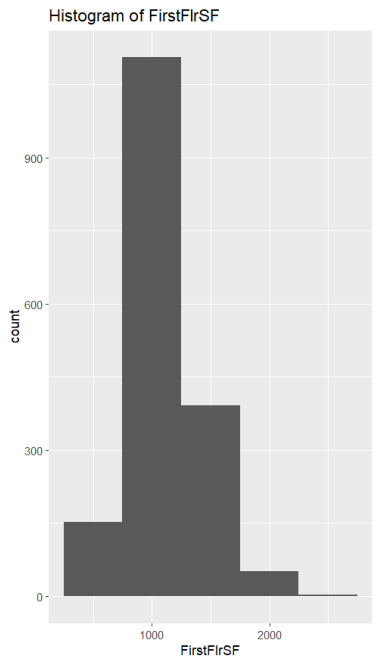


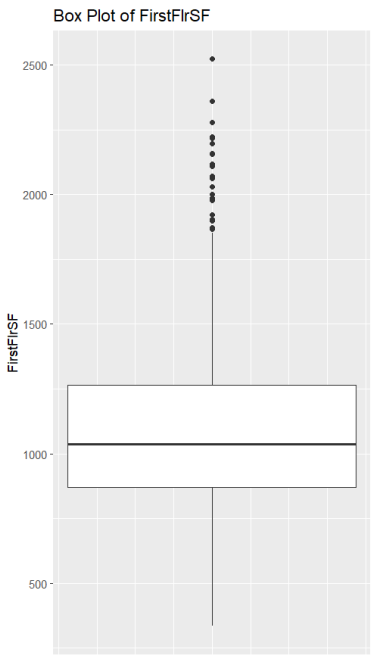


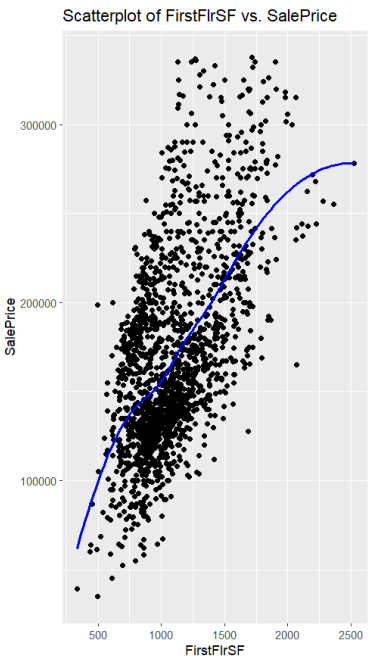



```
Variable: LotArea
Lower Bound: 2942.25
Upper Bound: 16562.25
Number of Outliers Below: 1
Number of Outliers Above: 82
Statistical Summary:
Min.: 2887 1st Qu.: 8049.75 Median: 9600 Mean: 10343.4460093897 3rd Qu.:
11454.75 Max.: 159000
```

3. FirstFlrSF

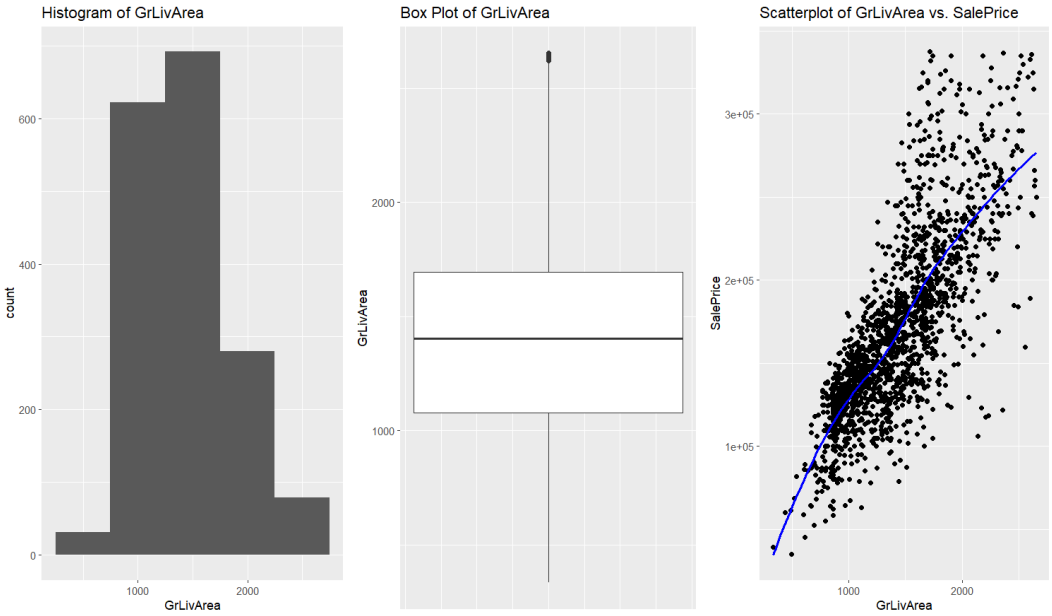






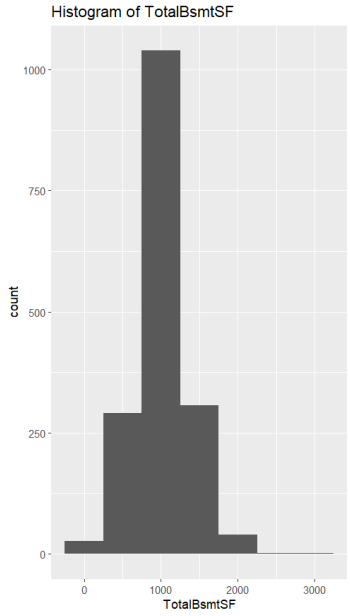
```
Variable: FirstFlrSF
Lower Bound: 274.625
Upper Bound: 1859.625
Number of Outliers Below: 0
Number of Outliers Above: 27
Statistical Summary:
Min.: 334 1st Qu.: 869 Median: 1037 Mean: 1095.99178403756 3rd Qu.:
1265.25 Max.: 2524
```

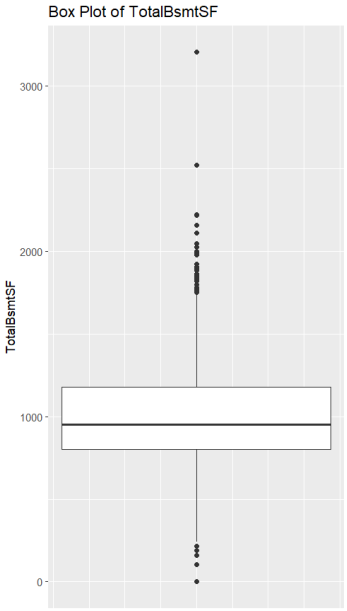
4. GrLivArea

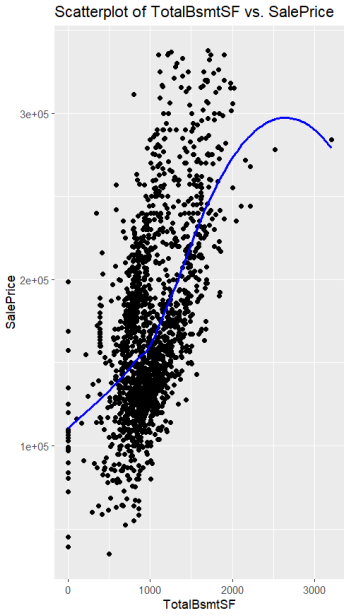



```
Variable: GrLivArea
Lower Bound: 158.375
Upper Bound: 2615.375
Number of Outliers Below: 0
Number of Outliers Above: 7
Statistical Summary:
Min.: 334 1st Qu.: 1079.75 Median: 1405 Mean: 1426.15727699531 3rd Qu.:
1694 Max.: 2654
```

5. TotalBsmtSF








```
Variable: TotalBsmtSF
Lower Bound: 239
Upper Bound: 1745
Number of Outliers Below: 25
Number of Outliers Above: 42
Statistical Summary:
Min.: 0 1st Qu.: 803.75 Median: 953 Mean: 1009.09507042254 3rd Qu.:
1180.25 Max.: 3206
```

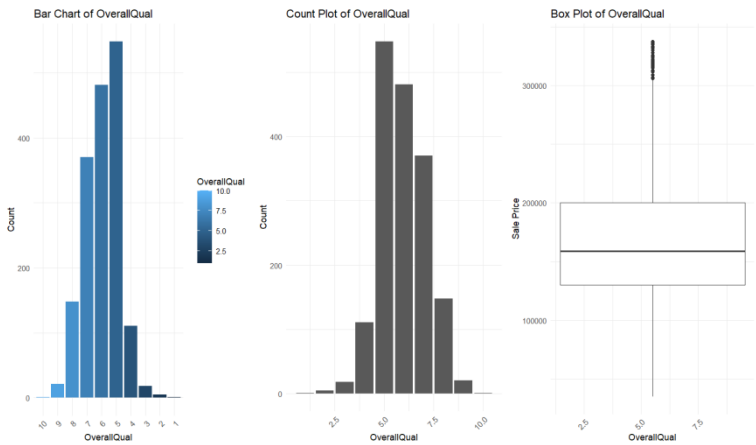
For Categorical Variables (OverallQual, OverallCond, SaleCondition, GarageType, Fence, HouseStyle):

1. Bar Charts: We can create bar charts to visualize the distribution of categories within each categorical variable.

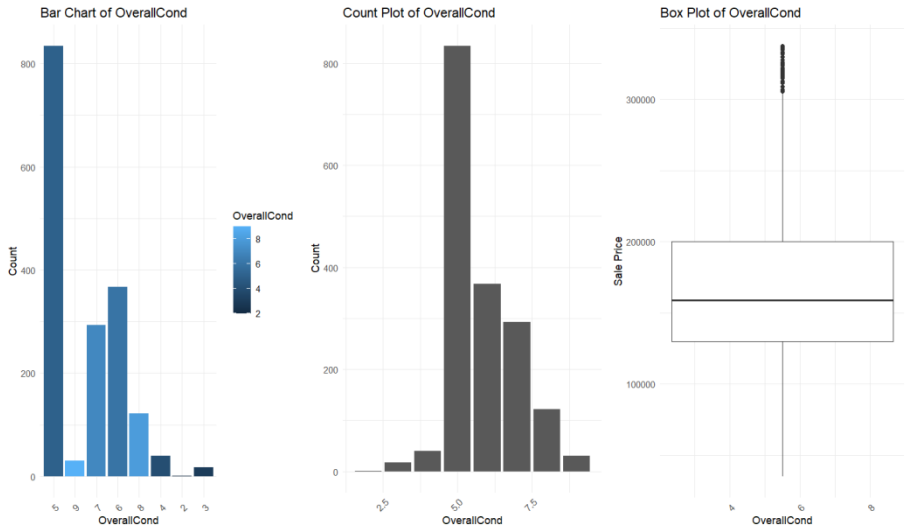
2. Count Plots: Count plots can provide insights into the frequency of different categories.

3. Boxplots: We can use these plots to explore relationships between categorical variables and continuous variables if applicable.

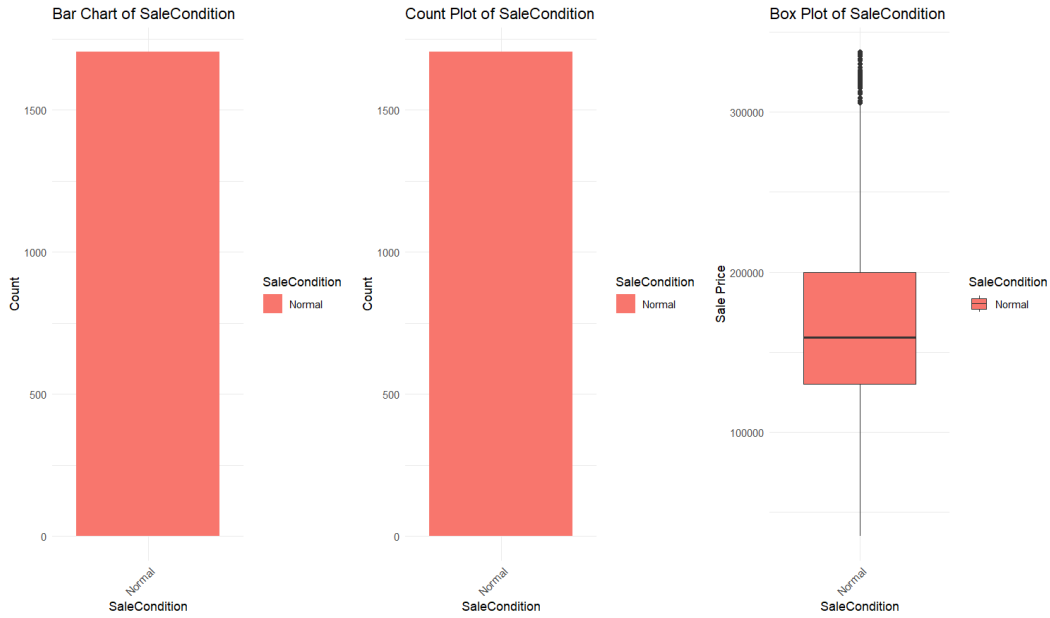
1. OverallQual



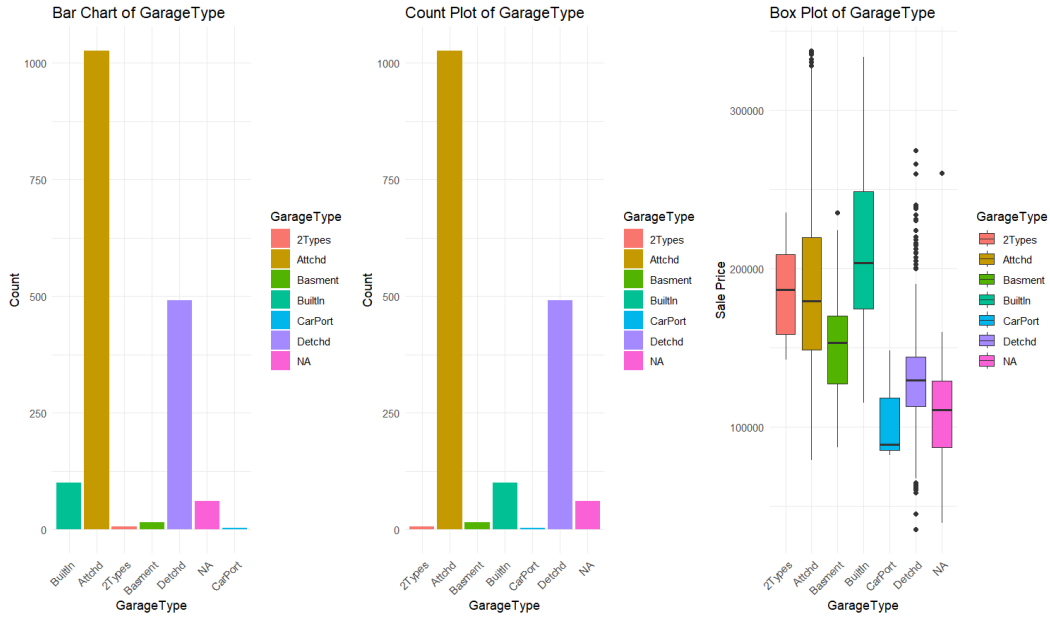
2. OverallCond



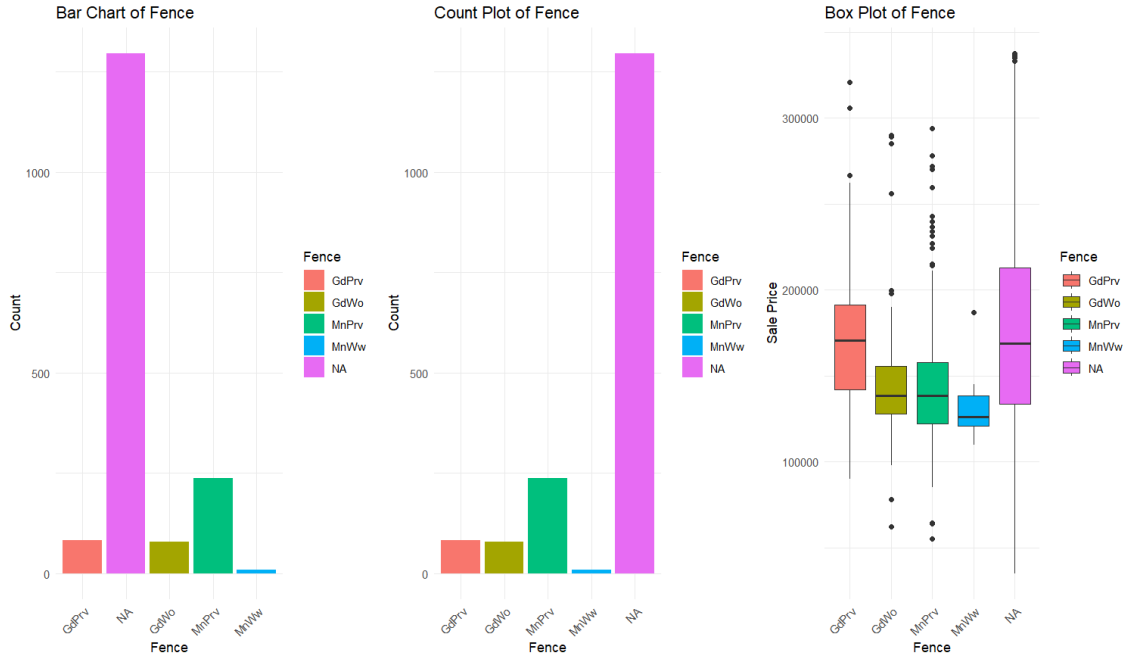
3. SaleCondition



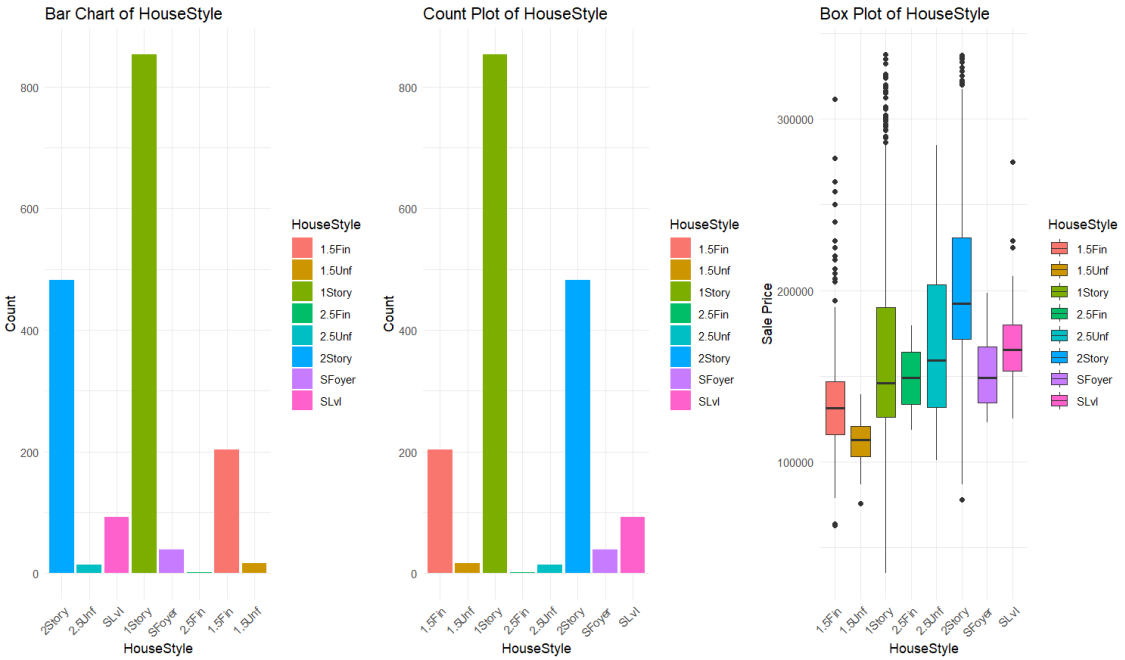
4. GarageType



5. Fence



6. HouseStyle



Section 5 – An Initial Exploratory Data Analysis for Modelling

What is the response variable in this problem? In addition to the raw response variable should we consider a transformation of the response variable? Consider SalePrice and log(SalePrice).

Pick three variables from the ten variables from your initial exploratory data analysis and explore their relationship with SalePrice and log(SalePrice).

Selection of Variables:

These three variables, "LotArea," "FirstFlrSF," and "GrLivArea," have been selected for analysis based on their significance in the real estate market and their potential impact on property sale prices. By examining the relationships between these variables and sale prices, we aim to gain insights into how property size and land area contribute to variations in pricing.

The choice of these variables aligns with common real estate considerations where prospective buyers and sellers often assess properties based on their lot size and living space. Analyzing these factors can provide valuable information for both property buyers and sellers, helping them make informed decisions in the real estate market.

The consideration of the logarithm of SalePrice, denoted as log(SalePrice), in real estate analysis and data exploration is based on several important statistical and practical reasons:

1. Normalization of Data: SalePrice, which represents the sale price of properties, is often characterized by a skewed or non-normal distribution. In many real estate datasets, there may be a wide range of sale prices, including a few very high-priced properties (outliers). Such a distribution can make it challenging to analyze and visualize the data effectively.

2. Reducing Skewness: Taking the logarithm of SalePrice helps in reducing the skewness of the data distribution. Skewness refers to the degree of asymmetry in the data distribution. A highly skewed distribution can affect the accuracy of statistical analyses and visualizations, as it may not meet the assumptions of normality.

3. Linear Relationships: Many statistical techniques, including regression analysis, assume that the response variable (SalePrice) and predictor variables (e.g., LotArea, FirstFlrSF, GrLivArea) have a linear relationship. Transforming SalePrice with a logarithm can make this relationship more linear, which can improve the performance of regression models and the interpretability of coefficients.

4. Equalizing Variance: In some cases, taking the log of SalePrice can help equalize the variance across different levels of the predictor variables. This is important for ensuring that the spread of predicted values is consistent, which is a key assumption in linear regression.

5. Interpretability: Using log(SalePrice) can make the interpretation of results more intuitive. For example, a coefficient in a regression model for log(SalePrice) can be interpreted as a percentage change in the original SalePrice for a one-unit change in the predictor variable.

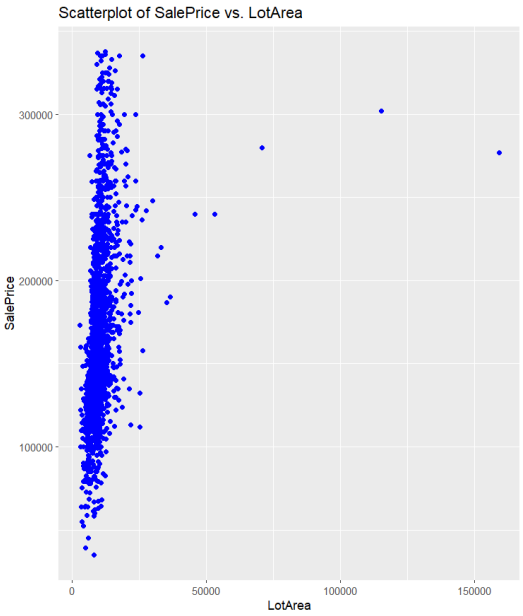
6. Visual Clarity: When creating scatterplots and other visualizations, the use of log(SalePrice) can often result in more visually interpretable and informative plots, especially when SalePrice spans a wide range.

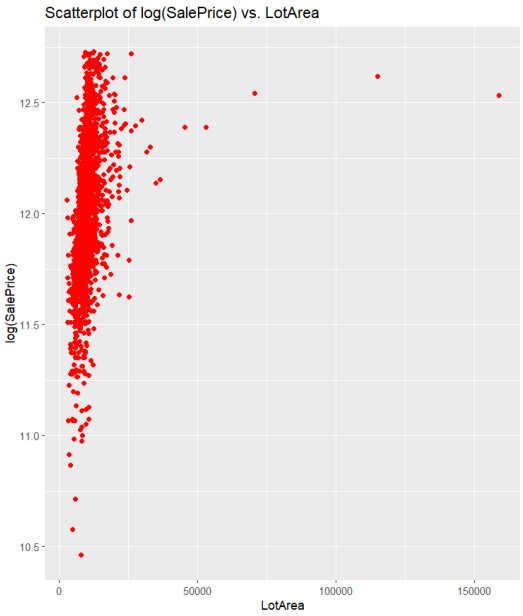
7. Residual Analysis: When performing regression analysis, examining the residuals (the differences between observed and predicted values) is a critical step. Using log-transformed SalePrice can result in more normally distributed residuals, which is beneficial for regression diagnostics.

In summary, the consideration of log(SalePrice) is a common practice in data analysis, particularly in situations where the original variable's distribution is skewed or non-normal. It can improve the accuracy and interpretability of statistical analyses and visualizations, making it a valuable tool in real estate data exploration and modelling.

Exploring the visible impacts of using scatterplots for "SalePrice" and "log(SalePrice)" in relation to the variables "LotArea," "FirstFlrSF," and "GrLivArea":

1. SalePrice vs. LotArea:

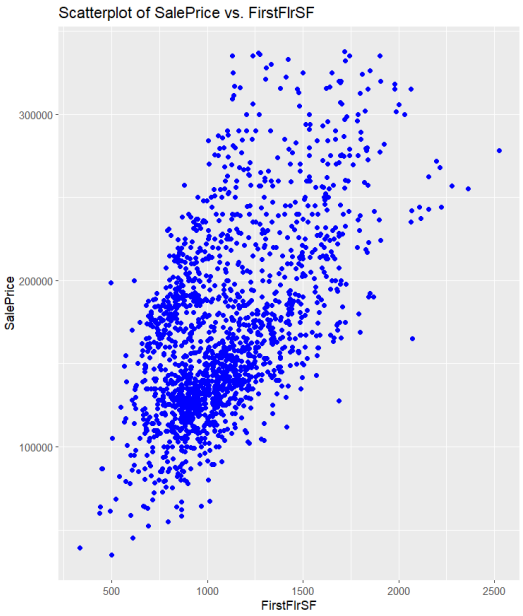




In the scatterplot of "SalePrice" against "LotArea," we observe a spread of data points. However, it might be challenging to discern a clear trend due to the wide range of sale prices across different lot sizes.

In contrast, when we take the logarithm of "SalePrice" and create a scatterplot against "LotArea," the data points tend to cluster more closely. This transformation reveals a clearer linear pattern, indicating that changes in lot area have a relatively consistent impact on the percentage change in sale price.

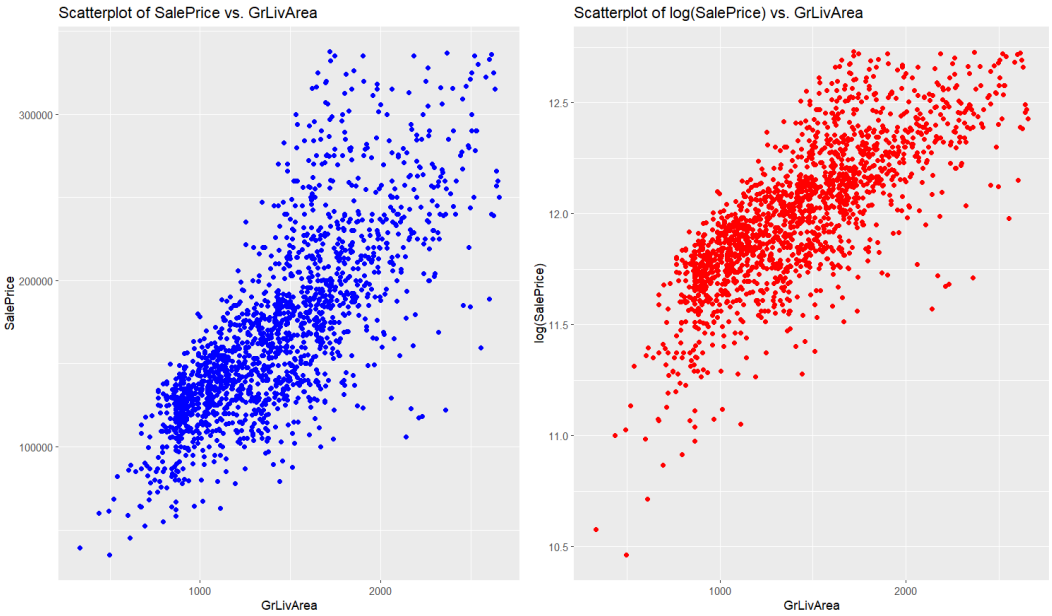
2. SalePrice vs. FirstFlrSF:



In the original scatterplot of "SalePrice" versus "FirstFlrSF," we see some variation in sale prices concerning firstfloor square footage, but the relationship may not be immediately apparent.

After applying the log transformation to "SalePrice" and plotting it against "FirstFlrSF," the data points exhibit a more linear pattern. This suggests that variations in firstfloor square footage correspond to relatively consistent changes in the percentage of sale price.

3. SalePrice vs. GrLivArea:



When examining the scatterplot of "SalePrice" against "GrLivArea," we notice a spread of data points, indicating that larger living areas generally result in higher sale prices. However, this relationship may not be entirely clear from the plot alone.

Upon logtransforming "SalePrice" and creating a scatterplot with "GrLivArea," we observe a more linear and easily interpretable pattern. This suggests that changes in living area size have a consistent impact on the percentage change in sale price.

In summary, the log transformation of "SalePrice" enhances the visibility of linear trends and relationships between sale price and the variables "LotArea," "FirstFlrSF," and "GrLivArea." It helps make these impacts more visible and interpretable in the scatterplots, facilitating a better understanding of how these factors influence the percentage change in sale price.

Skewness:

The skewness values for `SalePrice` and `log(SalePrice)` are 0.7452331 and -0.1948676, respectively. Let's compare these skewness values and explain the difference in their impact on model creation:

1. SalePrice (Skewness = 0.7452331):

Positive Skewness: The original `SalePrice` data has a rightskewed distribution, indicating that there is a longer right tail with some properties having significantly higher sale prices than the majority.

Impact on Model Creation: Rightskewed data can lead to issues in regression models. Models assume a normal distribution, and when the target variable is rightskewed, it can violate the assumption. As a result, the model may be less accurate in predicting high sale prices because it tends to underestimate them.

2. log(SalePrice) (Skewness = 0.1948676):

Negative Skewness: The `log(SalePrice)` transformation results in a leftskewed distribution, where there is a longer left tail. This transformation is often used to make the data more symmetric.

Impact on Model Creation: Leftskewed data, or data that has been transformed to be less rightskewed, can be beneficial for model creation. By taking the logarithm of `SalePrice`, we've made the distribution closer to a normal distribution, which is a key assumption in many statistical models. This transformation can help stabilize variances and improve model performance, especially in linear regression models.

In summary, the difference in skewness has a significant impact on model creation. Transforming the target variable from `SalePrice` to `log(SalePrice)` can make the data more suitable for linear modeling techniques and improve the model's ability to make accurate predictions across a wider range of sale prices. It helps mitigate the influence of extreme high prices and brings the data distribution closer to the normal distribution assumption that many statistical models rely on. This transformation is a common practice in regression analysis when dealing with skewed target variables.

Regression Models:

We are performing a regression analysis to predict housing prices (`SalePrice`) using three predictor variables: `LotArea`, `FirstFlrSF` (First Floor Square Feet), and `GrLivArea` (Above Ground Living Area Square Feet). Additionally, we want to compare the models for predicting `SalePrice` and `log(SalePrice)` to understand the impact of taking the logarithm of the response variable.

Step 1: Data Splitting

To assess the performance of our models, we split the dataset into two parts: a training dataset (70% of the data) and a testing dataset (30% of the data). This allows us to train our models on one subset and evaluate their performance on another to check for generalization.

Step 2: Model Building

We build two linear regression models:

Model 1 (`lm_model_saleprice_train`): This model predicts `SalePrice` directly using the predictor variables `LotArea`, `FirstFlrSF`, and `GrLivArea`. It helps us understand how well we can predict the actual sale prices.

Model 2 (`lm_model_log_saleprice_train`): This model predicts `log(SalePrice)` using the same predictor variables. Taking the logarithm of `SalePrice` can help normalize its distribution and address issues like skewness.

Step 4: Model Evaluation

We make predictions using both models on the testing dataset and calculate the Root Mean Squared Error (RMSE) for each model. RMSE measures the accuracy of our predictions by quantifying the average error between predicted and actual values.

Step 5: Comparison

Finally, we compare the RMSE values of the two models. Lower RMSE indicates a betterperforming model. By comparing the RMSE for `SalePrice` and `log(SalePrice)`, we can assess the impact of the logarithmic transformation on model performance.

This analysis helps us determine whether transforming the response variable (`SalePrice`) by taking its logarithm improves the model's ability to predict housing prices using the given predictor variables.

The output you provided shows the Root Mean Squared Error (RMSE) for two different models when applied to the test data:

1. RMSE for SalePrice on Test Data: 31278.21

This RMSE value represents the error or the average difference between the predicted values of the "SalePrice" made by a model and the actual "SalePrice" values in the test dataset.

In this context, an RMSE of 31278.21 means that, on average, the predicted SalePrice values are off by approximately $31,278.21 from the actual SalePrice values in the test data. A lower RMSE would indicate better model performance.

2. RMSE for log(SalePrice) on Test Data: 0.1834198

This RMSE value represents the error or the average difference between the predicted values of the "log(SalePrice)" made by a different model and the actual "log(SalePrice)" values in the test dataset.

In this context, an RMSE of 0.1834198 means that, on average, the predicted log(SalePrice) values are off by approximately 0.1834198 units from the actual log(SalePrice) values in the test data.

Since we are working with logtransformed values, this low RMSE suggests that the model is performing very well in terms of predicting logtransformed SalePrice.

In summary, when comparing the two RMSE values:

The RMSE for predicting SalePrice directly is quite high, indicating that the model has relatively high prediction errors when working with the original SalePrice values.

The RMSE for predicting log(SalePrice) is significantly lower, suggesting that the model performs much better when predicting the logtransformed SalePrice values. This indicates that taking the logarithm of SalePrice has likely improved the model's performance, as the RMSE is much smaller in this case.

Section 6 – Conclusion

The exploratory data analysis (EDA) conducted on the dataset provides valuable insights that can inform the model building process. In this analysis, several potential difficulties and concerns for the model building process emerge, along with the need to consider transformations in predictor variables. Let's delve into these aspects in detail.

1. Potential Difficulties or Concerns for Model Building:
    a. Skewness in Target Variable: The EDA reveals that the target variable, SalePrice, has a rightskewed distribution with a skewness value of 0.7452331. This skewness can be problematic for linear regression models, as they often assume normally distributed errors. In the case of SalePrice, this skewness may lead to model bias and inefficiency.
    b. Outliers: Outliers can significantly impact the performance of predictive models. The analysis identifies outliers in several continuous variables such as LotArea, TotalBsmtSF, and GrLivArea. These outliers can influence regression coefficients and predictions, potentially leading to less accurate models.
    c. Missing Data: The LotFrontage variable has 338 missing values. Handling missing data appropriately is crucial, as it can affect model stability and predictive accuracy. Imputing missing values or considering alternative approaches is necessary.

    d. Variability in Categorical Variables: Nominal and ordinal categorical variables like Neighborhood, OverallQual, and OverallCond exhibit significant variability in their categories. Incorporating these variables directly into a model may result in high cardinality, making it challenging to interpret coefficients and increasing the risk of overfitting.
    e. NonNormality in Predictor Variables: Continuous variables like TotalBsmtSF and GrLivArea display rightskewed distributions. While transformation of the target variable (SalePrice) is discussed, these predictor variables may also benefit from transformations to improve model performance.
2. Need for Predictor Variable Transformations:
    a. LogTransformation of SalePrice: The EDA highlights the impact of applying a logarithmic transformation to SalePrice. The transformed variable, log(SalePrice), exhibits a skewness of 0.1948676, indicating a more symmetrical distribution. This transformation aligns SalePrice closer to a normal distribution, meeting one of the assumptions of linear regression models. Considering the improved distribution, models built using log(SalePrice) as the target variable are likely to provide more reliable and interpretable results compared to models using the original SalePrice.
    b. Transformation of Predictor Variables: Given the skewness in predictor variables like TotalBsmtSF and GrLivArea, it may be beneficial to explore transformations such as

logtransformations, squareroot transformations, or BoxCox transformations. These transformations can help make these variables more suitable for linear regression models. Additionally, addressing outliers in predictor variables through winsorization or transformation is essential to reduce their influence on model outcomes.

c. Feature Engineering for Categorical Variables: To handle the variability in categorical variables with many levels, feature engineering techniques like onehot encoding, target encoding, or grouping similar categories could be applied. This reduces the dimensionality of categorical variables and enhances model interpretability.
d. d. Missing Data Handling: For variables with missing data, strategies such as imputation with appropriate methods or considering their significance in the modelbuilding process need to be explored.

In conclusion, the EDA conducted on the dataset has illuminated potential challenges and opportunities for enhancing the model building process. Addressing issues related to skewed distributions, outliers, missing data, and high cardinality in categorical variables is essential for building robust and accurate predictive models. Moreover, the transformation of the target variable, SalePrice, to log(SalePrice) has been identified as a promising approach to improve model performance, potentially reducing the impact of skewed data on regression outcomes. These findings serve as a foundation for further data preprocessing and model development, ultimately leading to more reliable predictions in the housing price regression task.