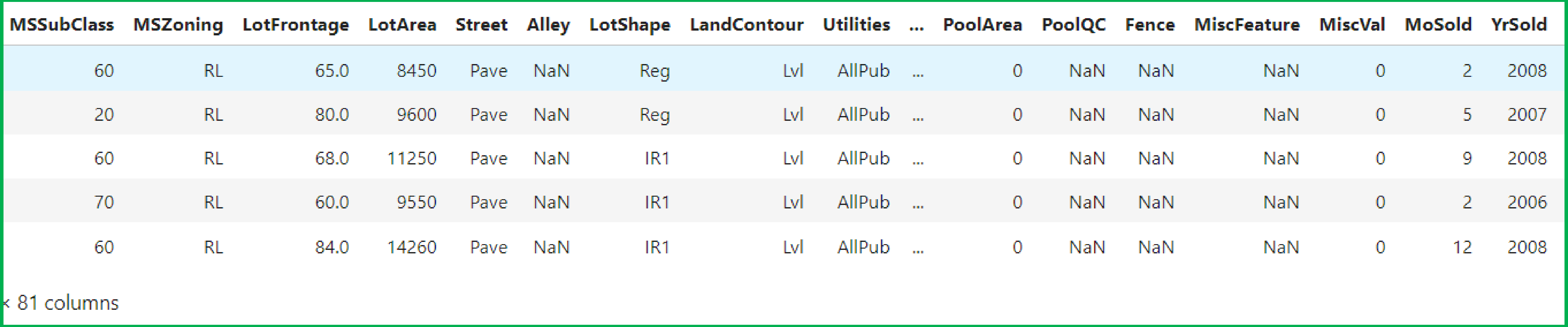
MACHINE LEARNING PROJECT

CLASSIFICATION AND REGRESSION

DATA DESCRIPTION

- The real estate dataset is typically used in housing market analysis. It includes detailed information about properties such as their physical attributes, conditions, neighbourhood details, and sales information. Here's a more detailed categorization:
- **Dataset Type** - Domain: Real Estate / Housing Market
- **Rows (Observations):** Each row represents a unique house.
- **Columns (Features):** Each column represents an attribute of the house or its sale.

- **Classification:** Predicting categorical variables such as zoning classification (MSZoning).
- **Regression:** Predicting continuous variables such as the sale price of houses (SalePrice).



MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold
60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008
20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007
60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008
70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006
60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008

× 81 columns

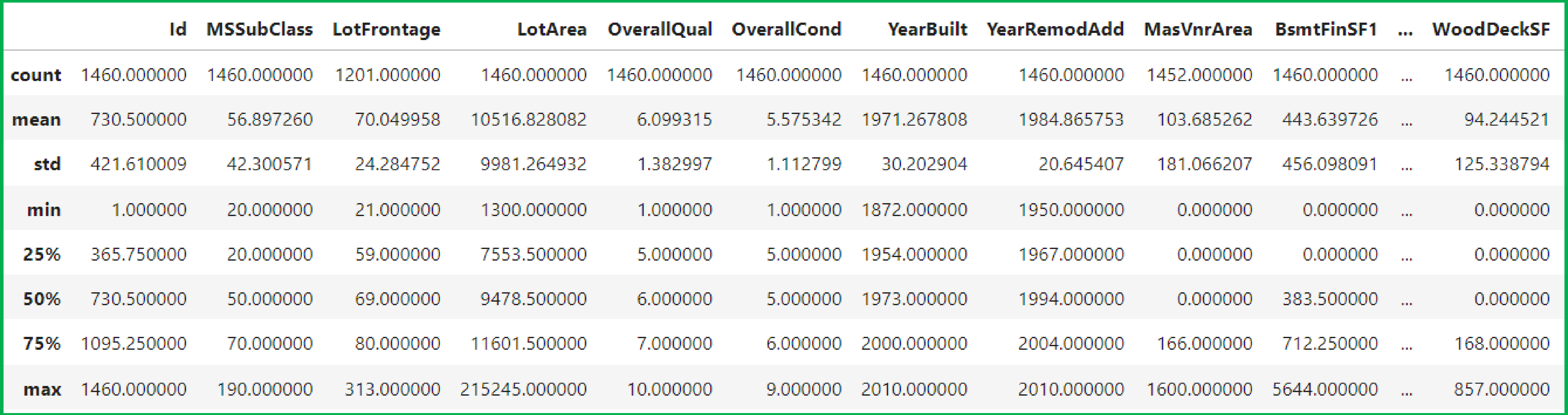
The dataset has 81 features, including the target variables. Here's a breakdown:

- ID: 1 feature (Id)
- Features: 79 features (from MSSubClass to MiscVal)
- Target Variables: 1 for classification (MSZoning), 1 for regression (SalePrice)

So, the total number of features (excluding target variables) is 79.

DATASET INSIGHTS

- The dataset consists of 1460 observations with 79 features, including numerical and categorical data points. Here's a detailed overview based on the provided statistics:



	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeckSF
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	...	1460.000000
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.865753	103.685262	443.639726	...	94.244521
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.112799	30.202904	20.645407	181.066207	456.098091	...	125.338794
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	0.000000	...	0.000000
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.000000	0.000000	0.000000	...	0.000000
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.000000	0.000000	383.500000	...	0.000000
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.000000	166.000000	712.250000	...	168.000000
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	...	857.000000

SNAPSHOT OF DATA

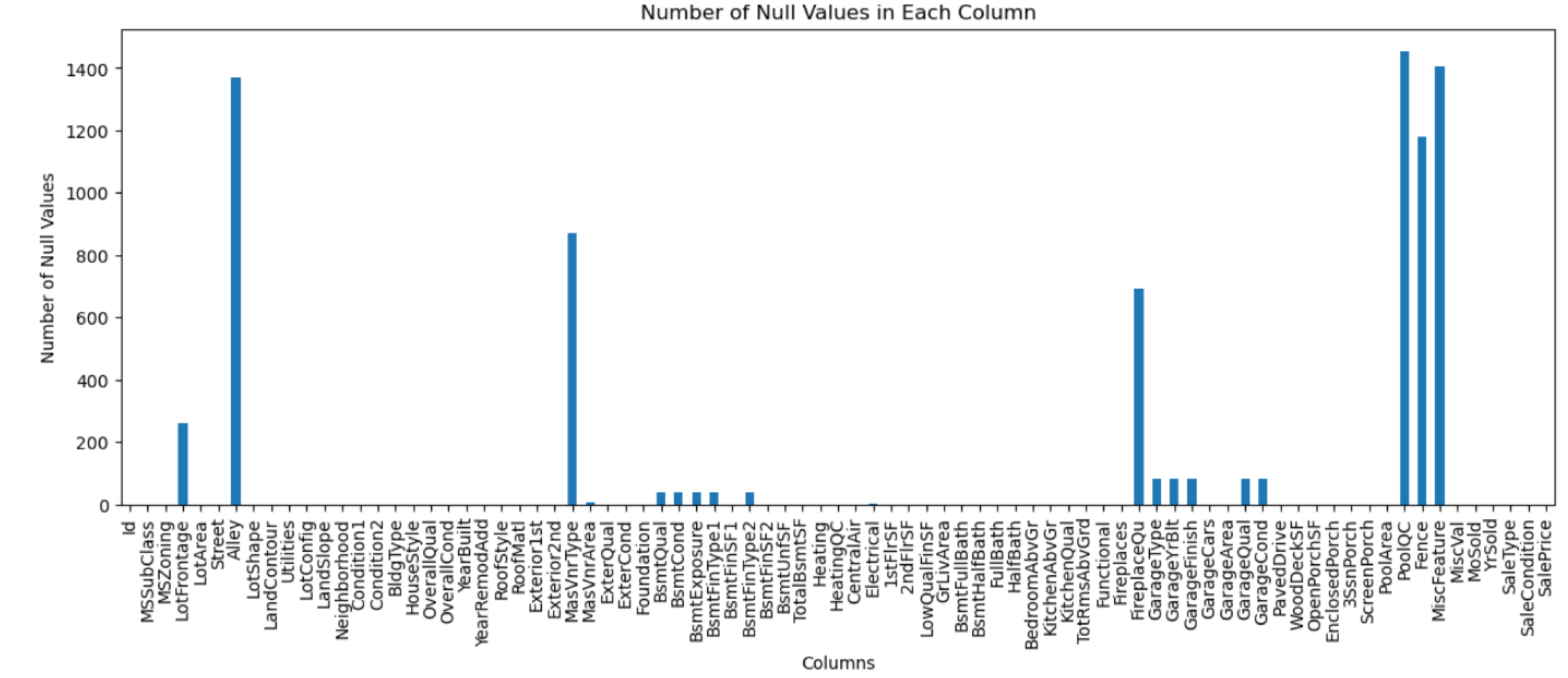
ANALYSING MISSING VALUES

- The dataset consists of **1460 observations**, but several features have missing values that need to be addressed for accurate analysis. The most significant missing values are in features like **Alley** (1369 missing), **PoolQC** (1453 missing), and **Fence** (1179 missing), indicating that these features are not present in the majority of the properties. Features related to garages, such as **GarageType**, **GarageYrBlt**, **GarageFinish**, **GarageQual**, and **GarageCond**, each have 81 missing values, suggesting that about 5.5% of the properties do not have a garage.

- **FireplaceQu** is missing for 690 entries, indicating that nearly half of the houses lack a fireplace. Additionally, **LotFrontage** has 259 missing values, which might reflect data entry gaps or properties without defined frontage.

- Basement-related features also show a notable amount of missing data: **BsmtQual**, **BsmtCond**, **BsmtExposure**, **BsmtFinType1**, and **BsmtFinType2** have around 37 to 38 missing entries, suggesting that a small fraction of houses lack basements. **MasVnrType** and **MasVnrArea** have 872 and 8 missing values, respectively, which could indicate that many homes do not have masonry veneer. The Electrical feature has only one missing value, making it the least concern in terms of missing data.

- Addressing these missing values is crucial for building robust predictive models, whether through imputation or by considering them as categorical indicators of feature absence.

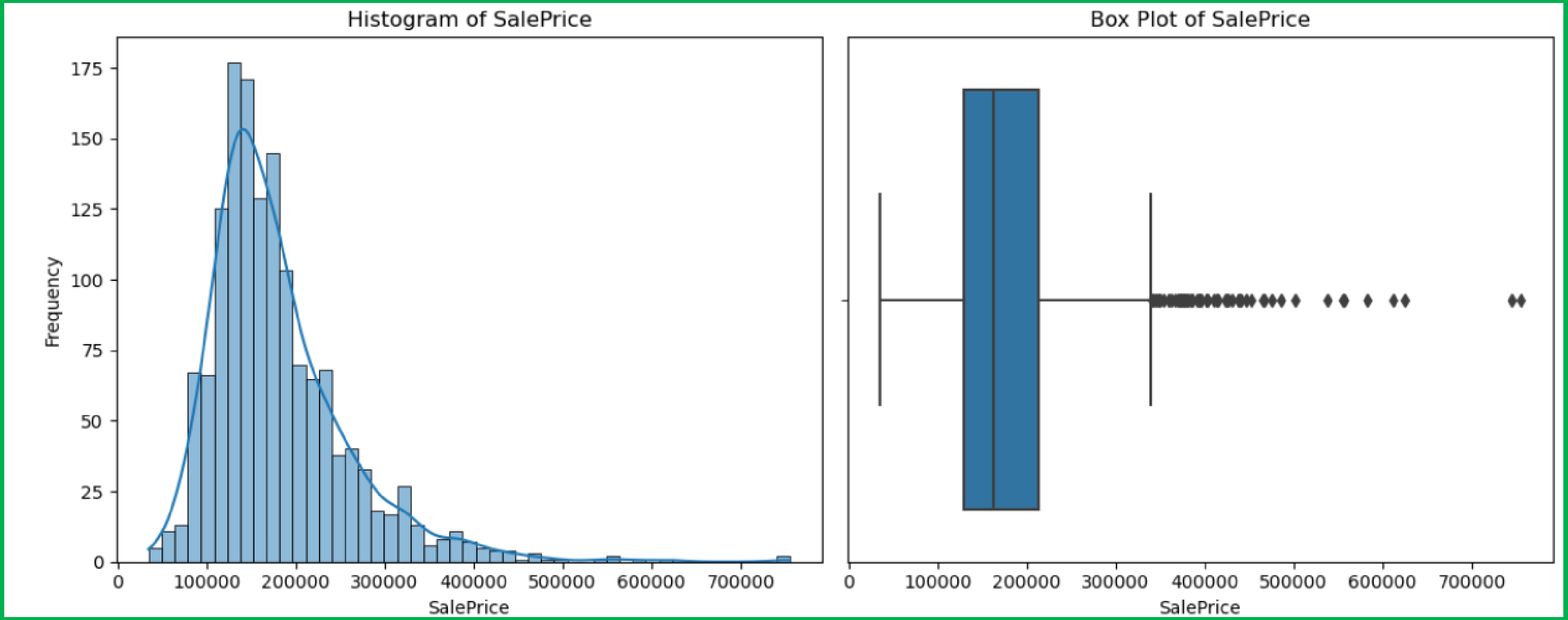


VISUALIZATION SHOWING MISSING VALUES USING BAR GRAPH

UNIVARIATE ANALYSIS

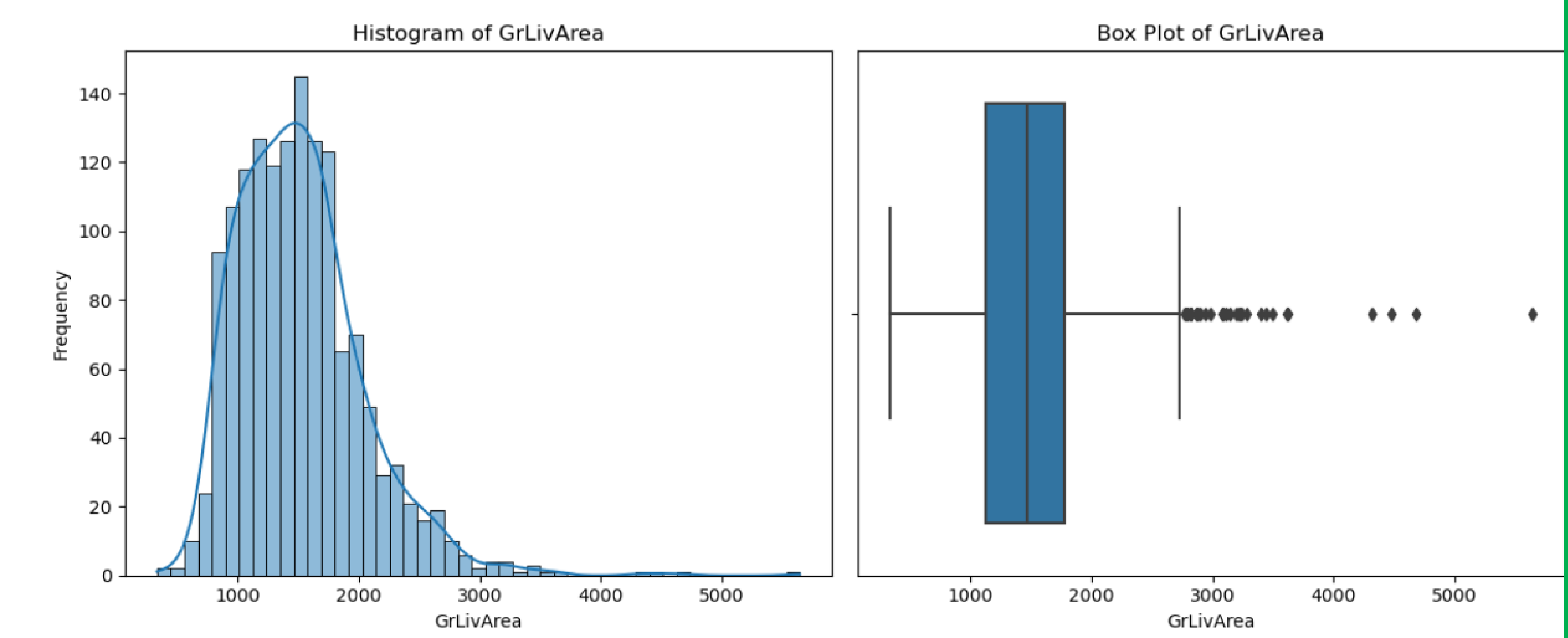
- **Univariate analysis is a statistical method used to describe and analyze data involving a single variable. In other words, it focuses on examining the distribution, summary statistics, and characteristics of a single variable at a time, without considering the relationships between variables.**
- **The goal of univariate analysis is to gain insights into the characteristics and patterns of individual variables within a dataset. It helps in understanding the central tendency, variability, and shape of the distribution of numerical variables, as well as the distribution of categories within categorical variables.**
- **Univariate analysis is often the first step in the exploratory data analysis process, providing a foundation for further analysis and hypothesis testing. It helps researchers or data analysts understand the nature of the data they are working with and identify any initial patterns or trends that may be present.**

SALEPRICE



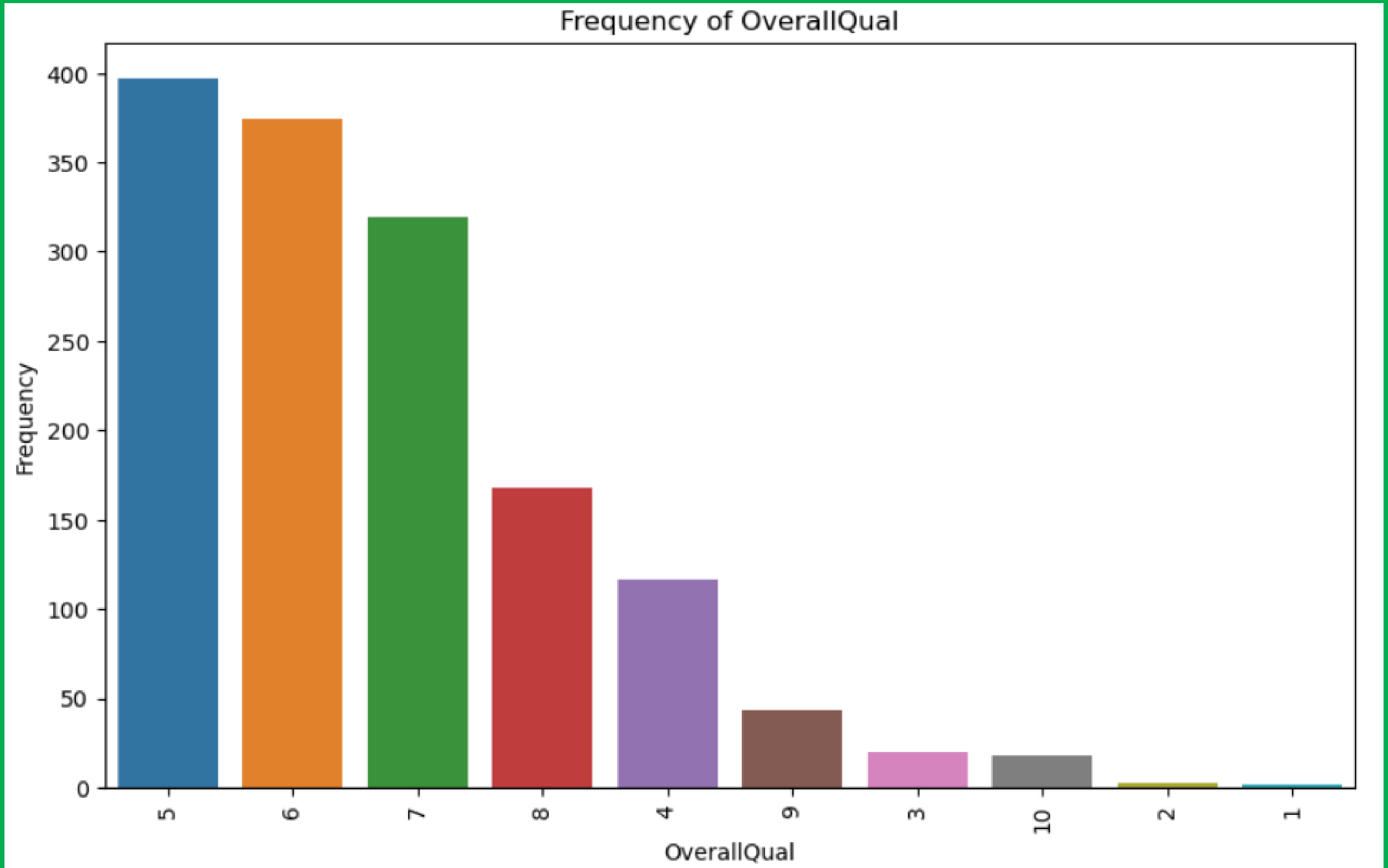
- **Histogram:** The histogram displays the frequency of sales prices across the range. The x-axis represents the sale price, and the y-axis represents the frequency. It appears there are more houses sold in the lower and middle price ranges compared to the higher price ranges.
- **Box Plot:** The box plot displays the median and quartile sales prices. The box in the center of the plot represents the middle quartiles (IQR), with the line in the middle representing the median sales price. The whiskers extend to the lowest and highest data points within 1.5 times the IQR from the median. There are outliers beyond the whiskers.

GRLIVAREA



- **Histogram:** The histogram displays the frequency of above-ground living areas across the range. The x-axis represents the above-ground living area, and the y-axis represents the frequency. It appears there are more houses with an above-ground living area in the 1500 to 2500 square foot range than any other range.
- **Box Plot:** The box plot displays the median and quartile above-ground living areas. The box in the center of the plot represents the middle quartiles (IQR), with the line in the middle representing the median above-ground living area. The whiskers extend to the lowest and highest data points within 1.5 times the IQR from the median. There are outliers beyond the whiskers.

OVERALLQUAL



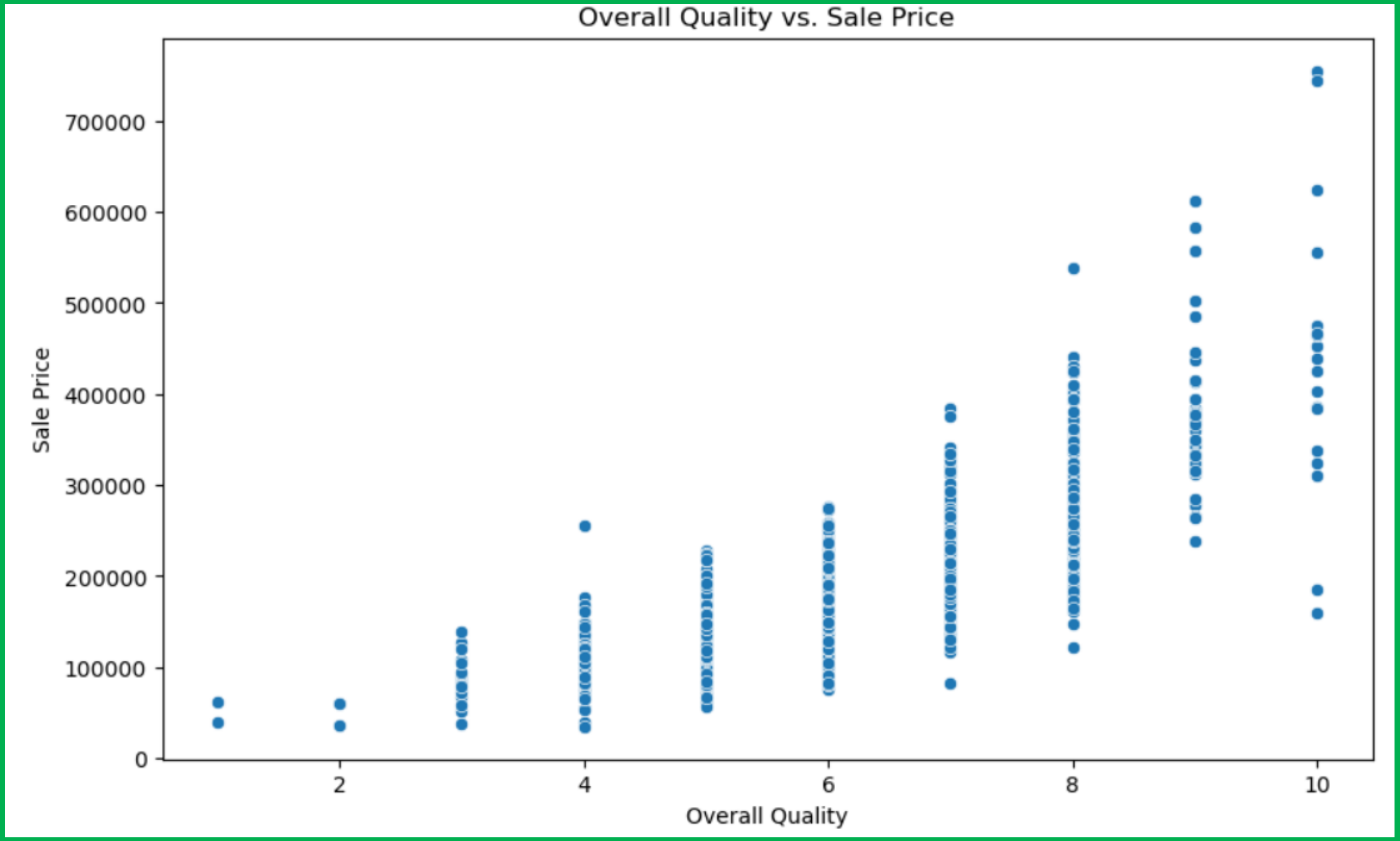
- There are five categories of OverallQual. It appears "10" is the most frequent category, followed by "9" and "8".
- It looks like higher OverallQual scores are more frequent than lower scores. This suggests that most of the data points represent houses with a higher overall quality.

BIVARIATE ANALYSIS

Bivariate analysis involves examining the relationship between two variables.

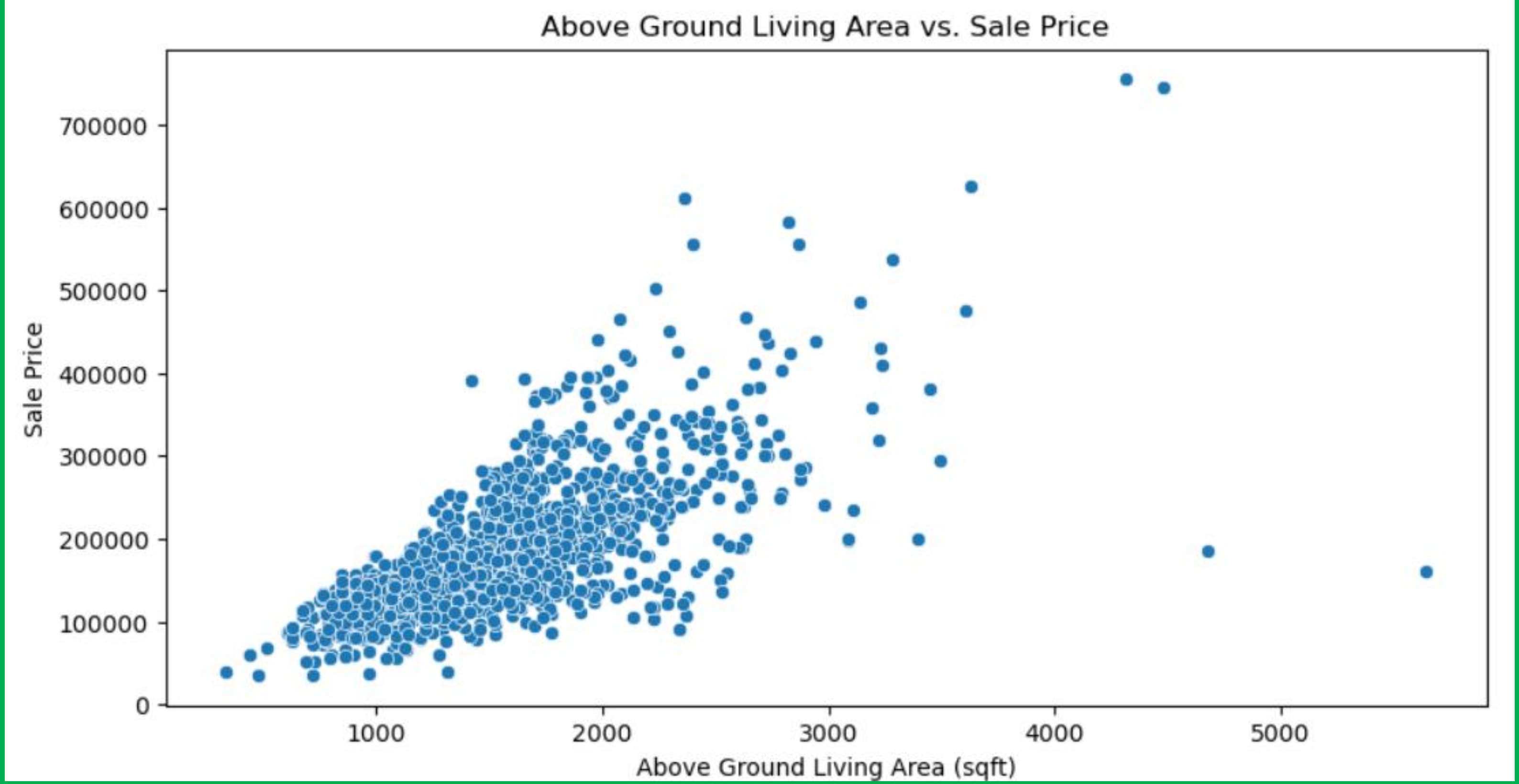
- **OverallQual vs. SalePrice:** We have used scatter plot with OverallQual on the x-axis and SalePrice on the y-axis to see how the overall quality of the house relates to its sale price.
- **GrLivArea vs. SalePrice:** Another scatter plot with GrLivArea (above ground living area square feet) on the x-axis and SalePrice on the y-axis can reveal the correlation between the size of the living area and the sale price.
- **Neighborhood vs. SalePrice:** We created a box plot with Neighborhood on the x-axis and SalePrice on the y-axis to compare the distribution of sale prices across different neighborhoods.

OVERALLQUAL VS. SALEPRICE



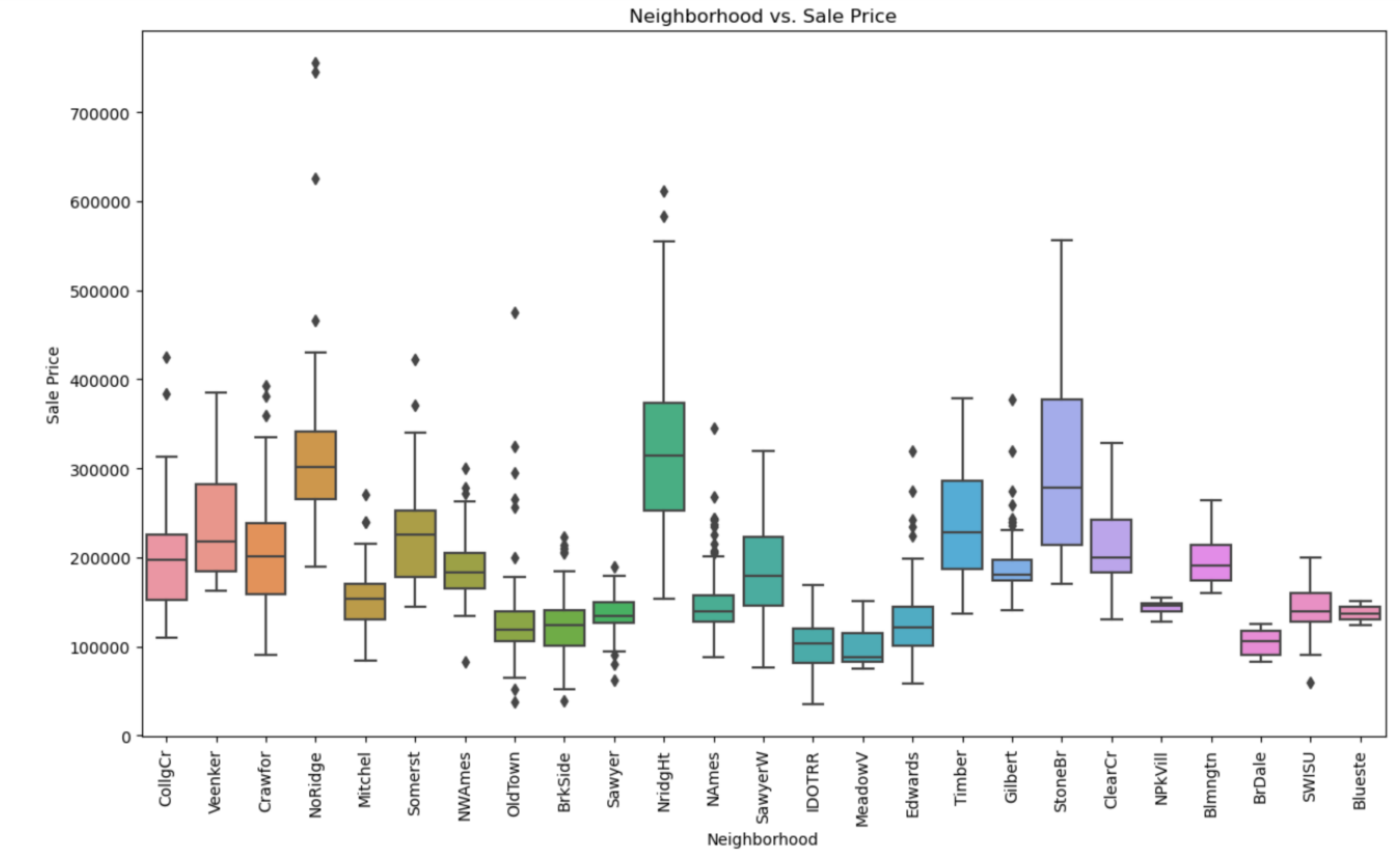
- There is a positive correlation between overall quality and sale price. This means that as the overall quality of a house increases, the sale price also tends to increase.
- The data points are spread out, which means there is a variation in sale price for houses of similar quality. This could be due to many factors, such as location, size, and amenities.

GRLIVAREA VS. SALEPRICE



- There is a positive correlation between above ground living areas and sale prices. This means that as the above ground living area of a house increases, the sale price also tends to increase.
- The data points are spread out, which means there is a variation in the sale price for houses of similar above ground living areas. This could be due to several factors, such as location, number of bedrooms, and amenities.

NEIGHBORHOOD VS. SALEPRICE

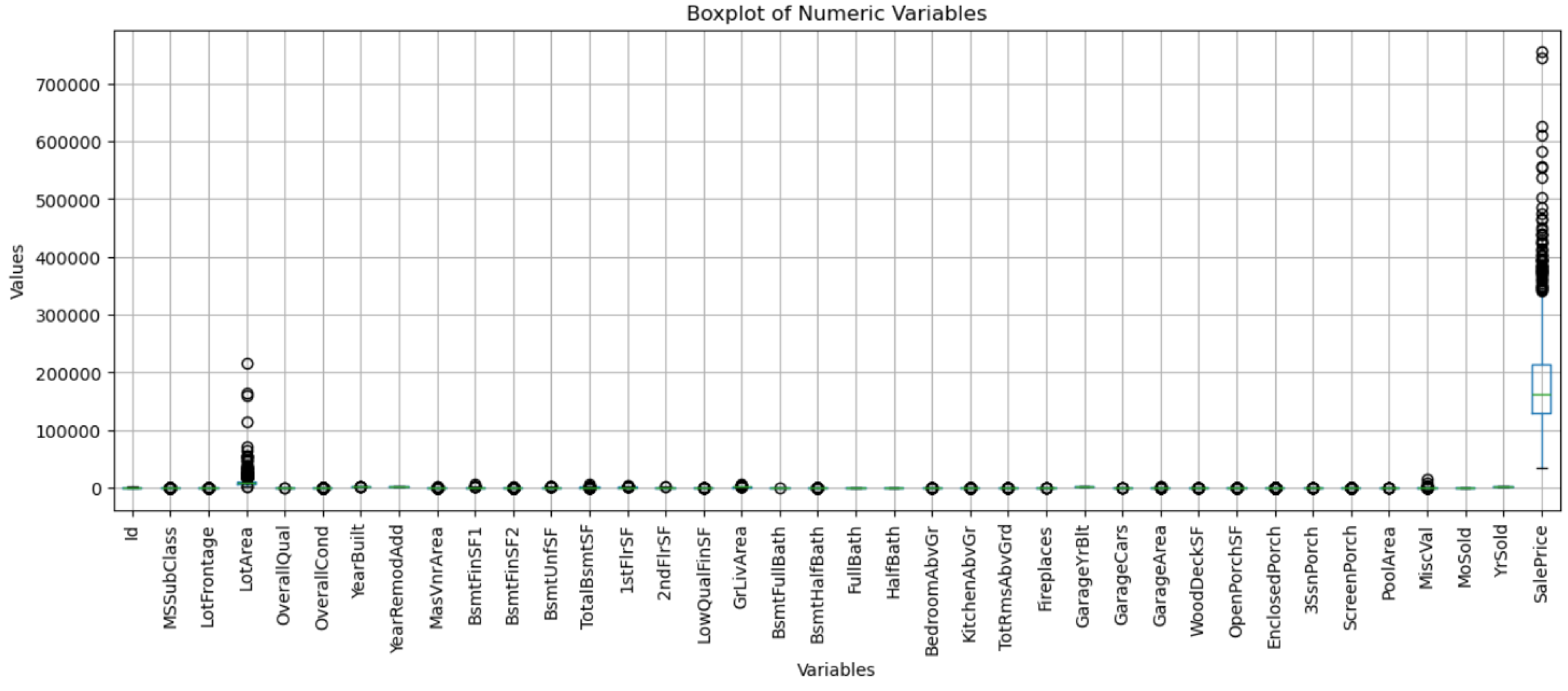


- There is a large variation in the sale price of houses between different neighborhoods. This suggests that neighborhood is a major factor affecting the sale price of a house.
- Some neighborhoods, such as College Creek and Veenker, have a much higher average sale price than other neighborhoods, such as Edwards and Meadowview.
- This could be due to several factors, such as the quality of the schools, the crime rate, the available amenities, or the desirability of the location. It is important to note that the graph only shows the average sale price in each neighborhood. There will be some variation in sale prices within each neighborhood.

DETECTING OUTLIERS IN THE DATASET USING THE TUKEY METHOD

- During the pre-processing of the house price dataset, a crucial step involved detecting outliers using the Tukey method, which is based on the Interquartile Range (IQR). First, we defined a function to identify outliers by computing the first (Q1) and third (Q3) quartiles for each numeric variable. The IQR, representing the middle 50% of the data, was calculated as the difference between Q3 and Q1. We then determined the lower and upper bounds as \(Q1 - 1.5 \times IQR \) and \(Q3 + 1.5 \times IQR \) respectively, and flagged data points falling outside these bounds as outliers.

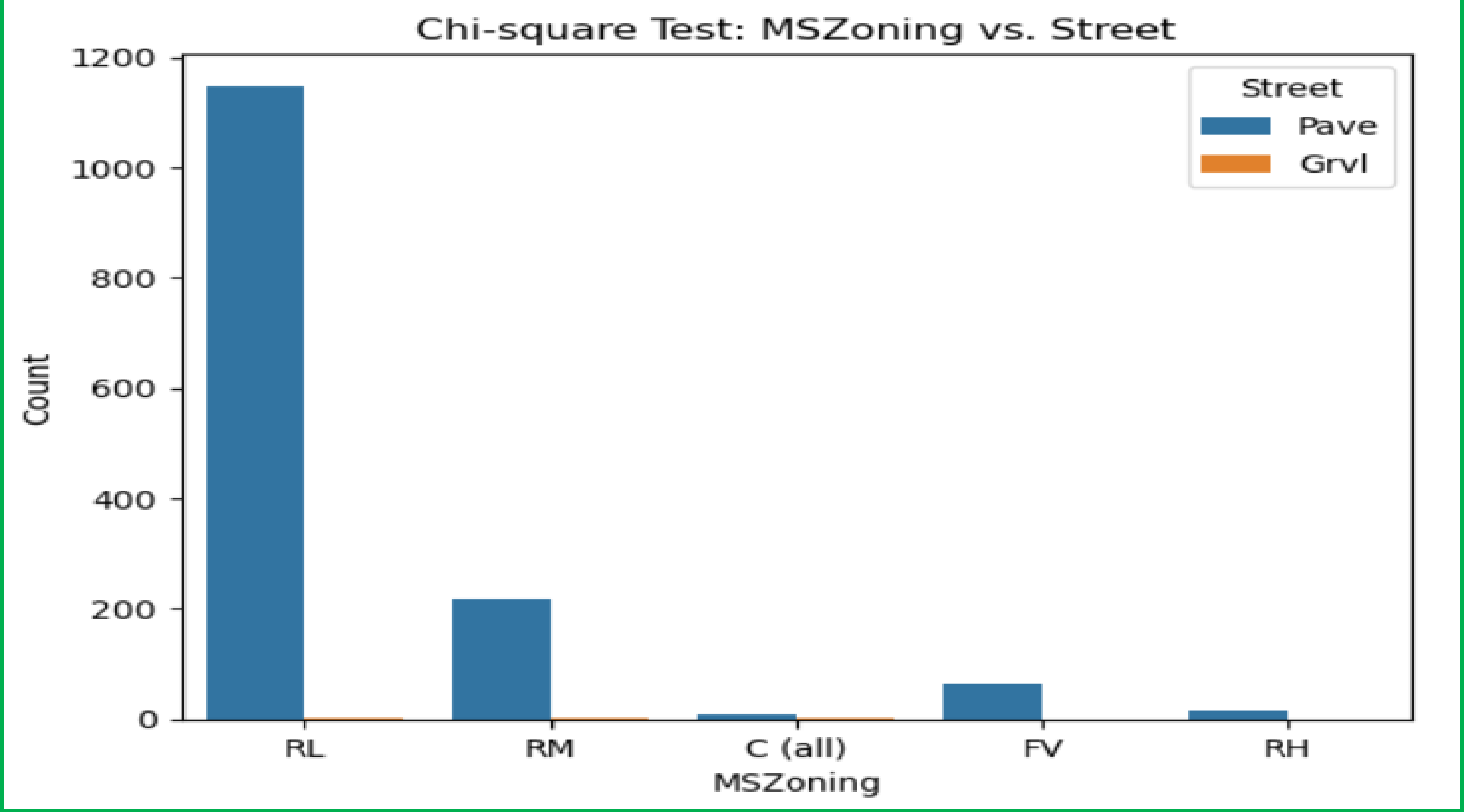
- Applying this function to each numeric variable in the dataset revealed several variables with significant outliers. For instance, variables like `LotFrontage`, `LotArea`, `BsmtFinSF1`, `BsmtFinSF2`, `GrLivArea`, `TotalBsmtSF`, and `SalePrice` exhibited extreme values. `LotArea` showed notably large lot sizes deviating from the typical range, while `SalePrice` had high-value homes considerably above the average. Identifying these outliers allows for informed decisions on handling them, such as capping extreme values, transforming variables, or removing outlier rows. This step ensures the dataset is cleansed of anomalies that could skew analysis and modeling, leading to more accurate and reliable results.


Boxplot of Numeric Variables
Values
700000
600000
500000
400000
300000
200000
100000
0
Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
Variables

CHI-SQUARE TEST FOR INDEPENDENCE

In this section, we will perform a Chi-square test for independence to examine the relationship between two categorical variables in the house price dataset: 'MSZoning' and 'Street'. This test will help us determine if there is a significant association between these variables.

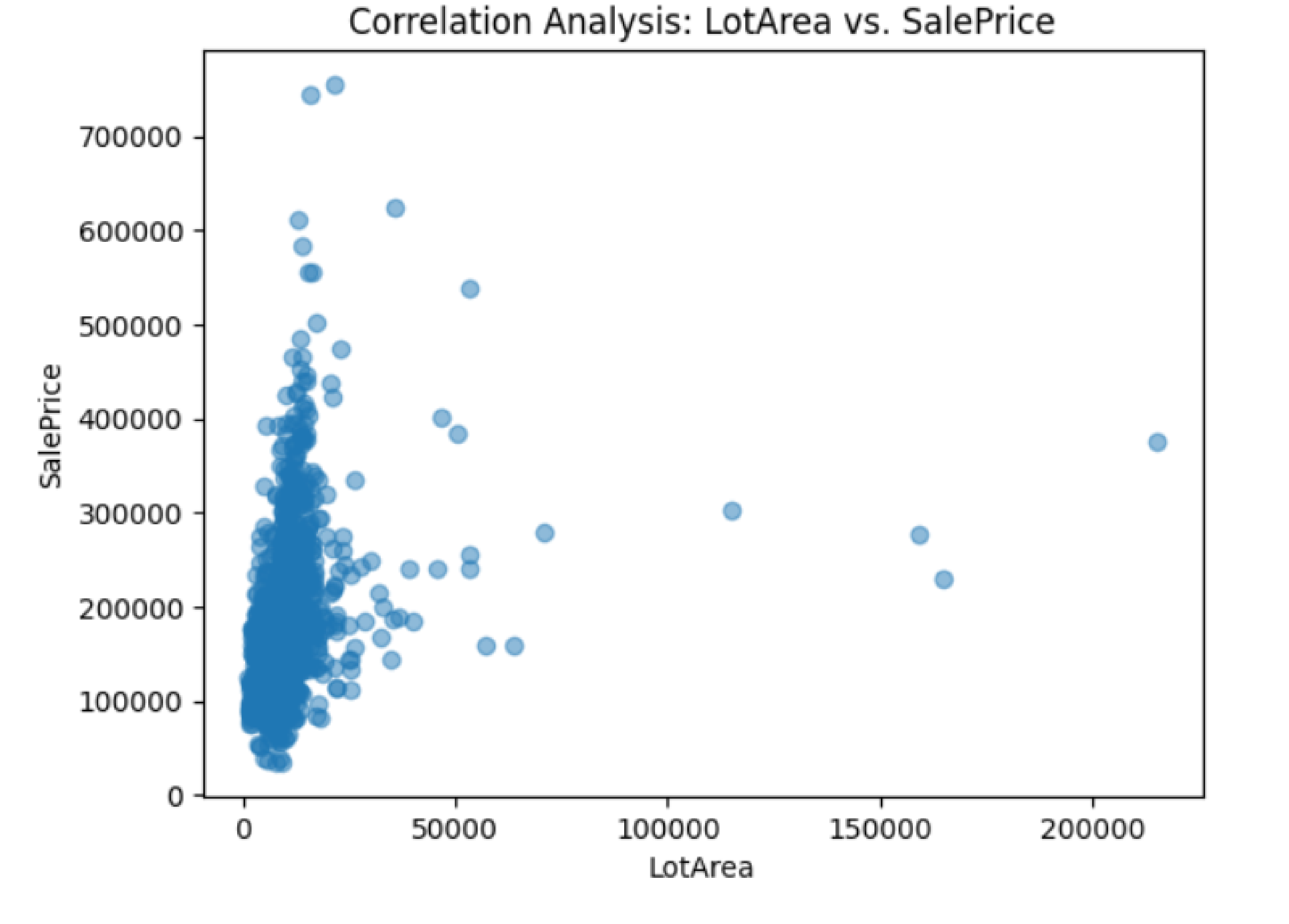
The chi-square test results between the variables 'MSZoning' and 'Street' reveal a statistically significant association between them. With a chi-square statistic of 94.74 and an extremely low p-value of approximately 1.29e-19, we reject the null hypothesis of independence, indicating that there is a significant relationship between the zoning classification of properties and the type of street they are located on. This suggests that the choice of street type (e.g., paved or gravel) is not random and may be influenced by the zoning regulations or other factors. Understanding this association can be valuable in urban planning, real estate development, and infrastructure management, as it provides insights into the interconnectedness of zoning policies and street infrastructure within a geographic area. Further analysis can explore the specific nature and implications of this relationship, potentially informing policy decisions and urban development strategies.

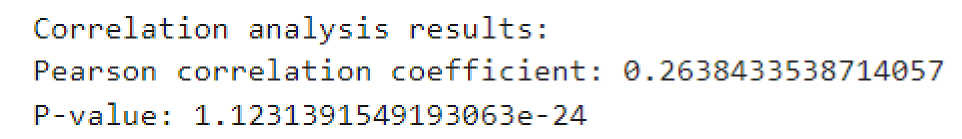


CHI-SQUARE TEST FOR INDEPENDENCE

CORRELATION ANALYSIS

- The correlation analysis between '**LotArea**' and '**SalePrice**' yields valuable insights into their relationship within the dataset. With a Pearson correlation coefficient of approximately 0.264, we observe a moderately positive correlation between the two variables.
- This indicates that as the lot area increases, there tends to be a corresponding increase in the sale price of properties. The very low p-value (approximately 1.12e-24) suggests that this correlation is statistically significant, providing strong evidence against the null hypothesis of no correlation.
- Understanding this relationship can be pivotal in real estate valuation and investment decisions, as it highlights the importance of lot size in determining property prices. Additionally, it underscores the need for property developers, investors, and urban planners to consider lot areas as a critical factor in assessing property values and market dynamics. Further exploration of this correlation can potentially uncover nuanced insights into regional housing markets and buyer preferences, facilitating informed decision-making and strategic planning in the real estate sector.

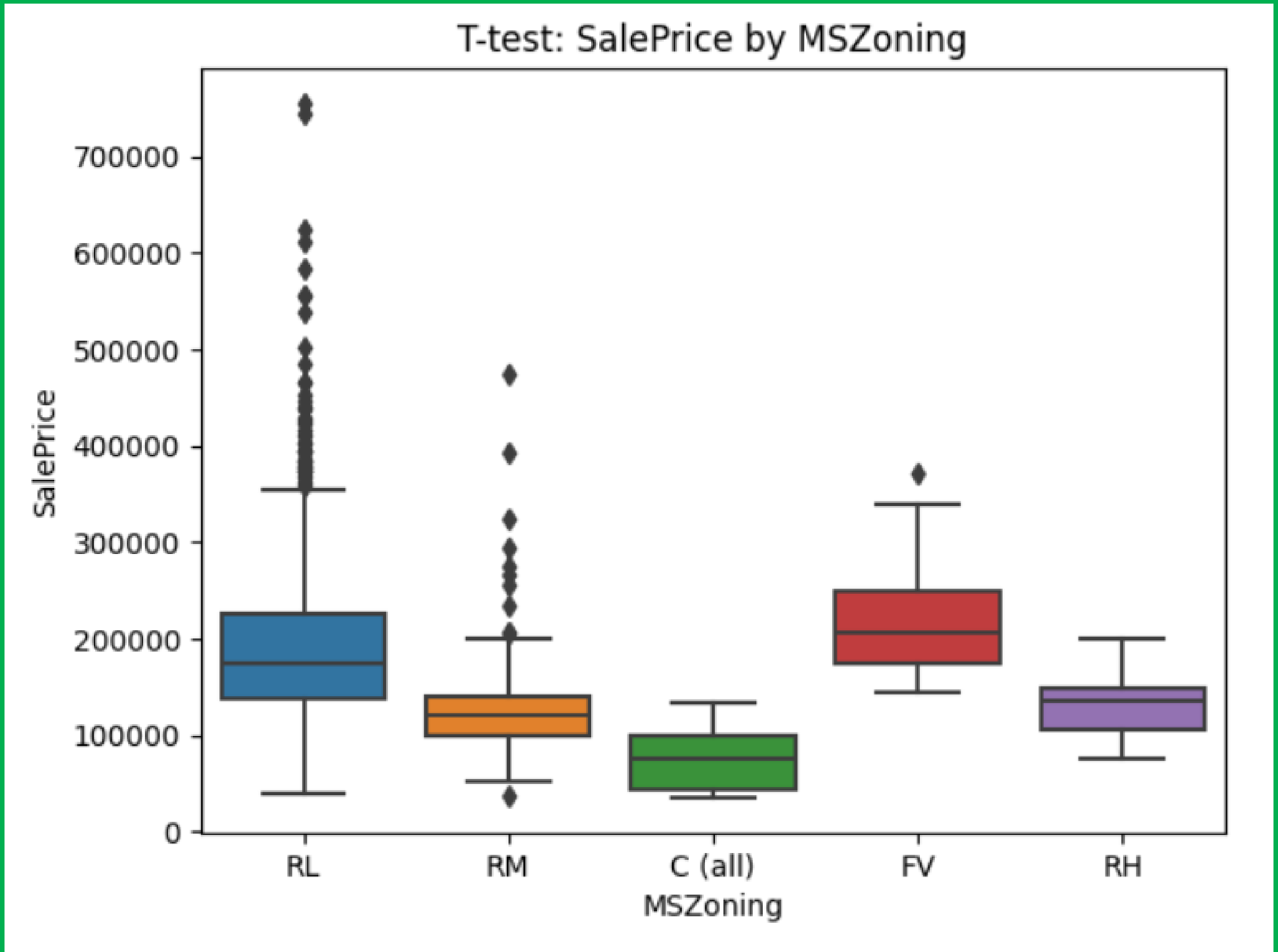




```
Correlation analysis results:
Pearson correlation coefficient: 0.2638433538714057
P-value: 1.123139154919306e-24
```

T-TEST FOR INDEPENDENCE

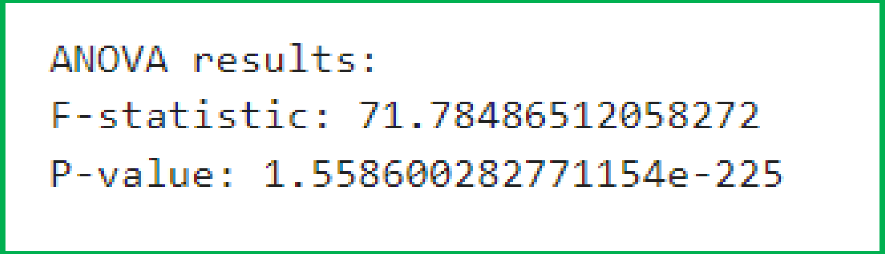
- The T-test results further support these observations, indicating a statistically significant difference in mean sale prices between the zoning classifications.
- With a high T-statistic of approximately 11.44 and an extremely low p-value of approximately 5.31e-29, we reject the null hypothesis of equal means, highlighting significant variations in property prices across different zoning categories.
- This suggests that zoning classification plays a crucial role in determining property values, with certain zoning categories associated with higher or lower sale prices. Such insights can inform various stakeholders, including real estate developers, investors, and policymakers, in understanding market dynamics and making informed decisions regarding property development, investment strategies, and urban planning initiatives.
- Understanding the relationship between zoning classifications and property prices can also aid in identifying potential investment opportunities, optimizing land use, and shaping sustainable urban development strategies that align with economic and social objectives.



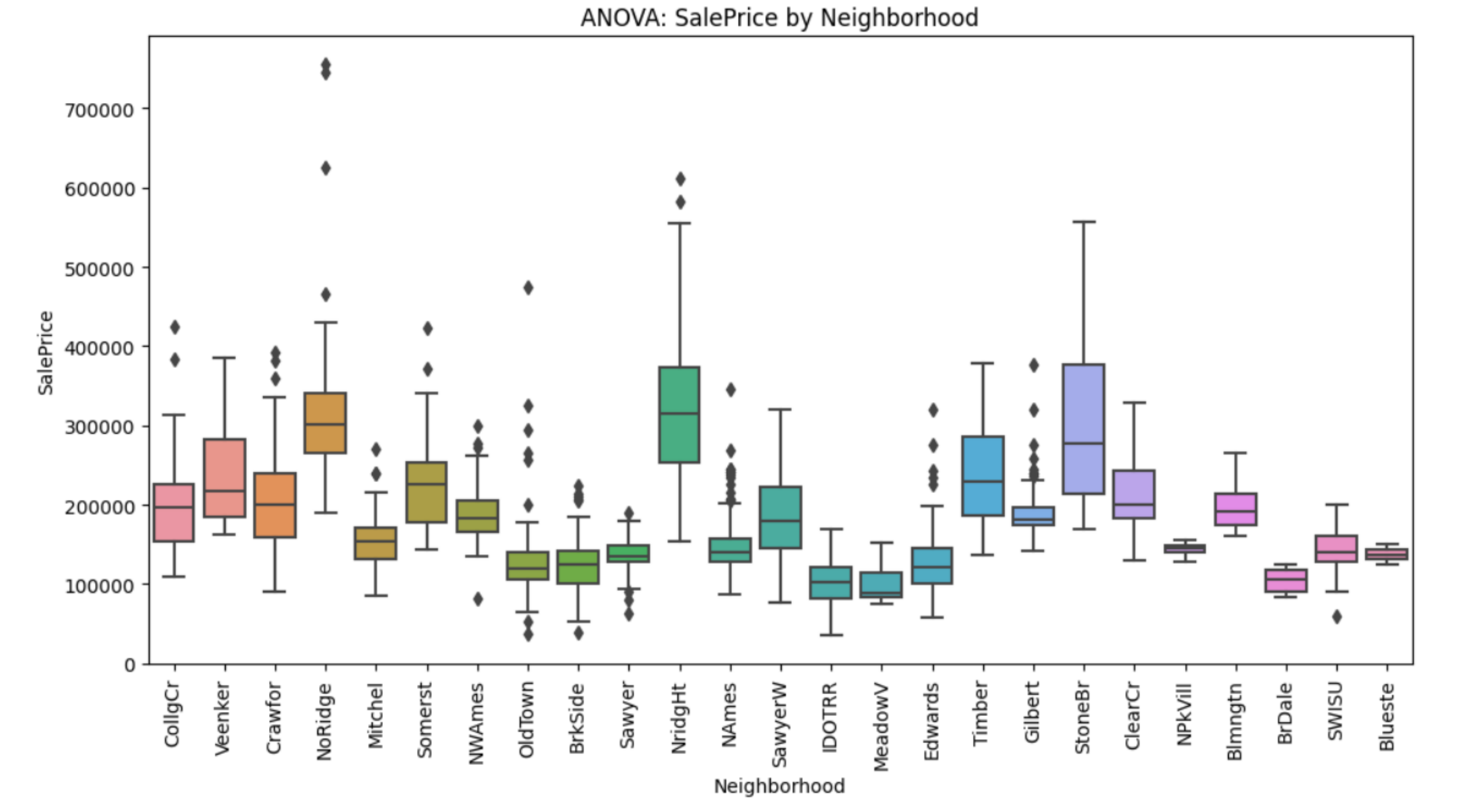
The boxplot visualization of 'SalePrice' across different zoning classifications ('MSZoning') provides visual insights into the distribution of property prices within each zoning category.

ANOVA TEST FOR INDEPENDENCE

- The analysis of variance (**ANOVA**) results for **'SalePrice**' across different neighborhoods unveils significant variations in property prices among the distinct neighborhood groups.
- With a substantial F-statistic of approximately 71.78 and an exceedingly low **p-value** of approximately 1.56e-225, we reject the null hypothesis of equal mean sale prices across neighborhoods. This underscores the influence of neighborhood location on property values, indicating that the choice of neighborhood significantly impacts sale prices.



```
ANOVA results:
F-statistic: 71.78486512058272
P-value: 1.558600282771154e-225
```



Such findings offer crucial insights for various stakeholders in the real estate industry, urban planners, and policymakers. For homebuyers and investors, understanding the disparities in property prices across neighbourhoods can guide decision-making processes, aiding in selecting neighbourhoods that align with budgetary constraints and investment goals. Similarly, for developers and urban planners, these insights inform land use planning, infrastructure development, and community revitalization efforts, facilitating the creation of vibrant and sustainable neighbourhoods.

CLASSIFICATION RESULT ANALYSIS

Support Vector Machine (SVM):

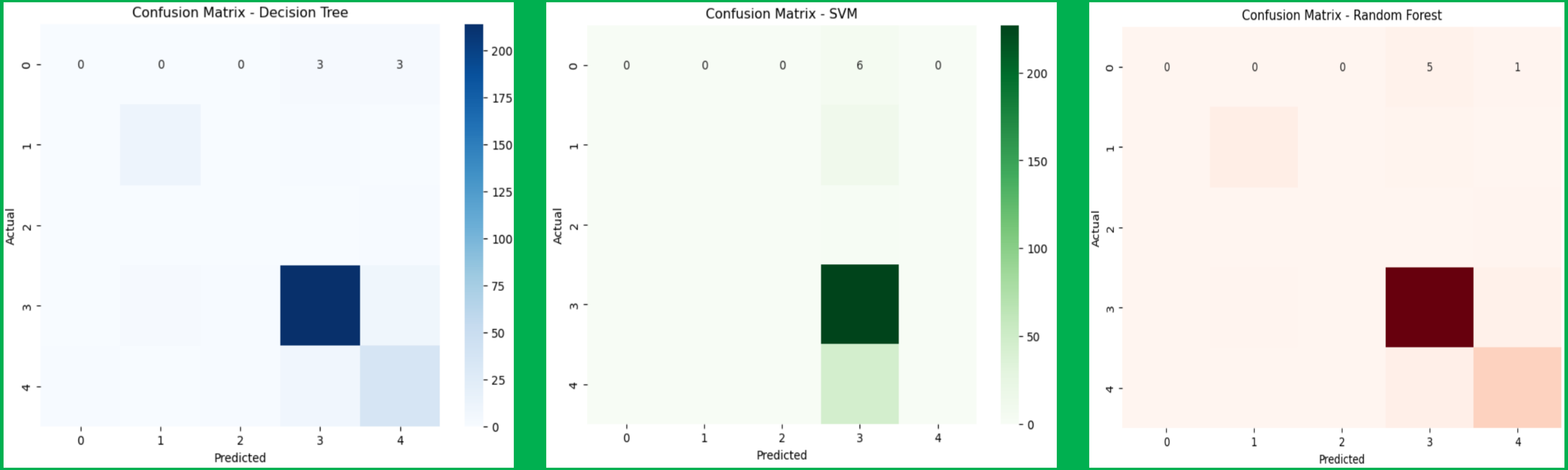
The SVM classifier achieved an accuracy of **77.74%**. Looking at the precision, recall, and F1-score metrics, it shows strong performance for class **1** with precision of **79%** and recall of **92%**, indicating it correctly identifies a high proportion of instances from this class. However, it struggled with classes **0** and **2**, both having a precision and recall of **0%**, implying it failed to correctly classify any instances for these classes. Overall, the weighted average F1-score is **0.89**, suggesting a balanced performance across classes.

Decision Tree Classifier:

The Decision Tree classifier slightly outperformed the SVM with an accuracy of **89.38%**. It shares similar precision and recall patterns as the SVM, with strong performance in class **1** (precision of **79%** and recall of **92%**) and deficiencies in classes **0** and **2** (both with precision and recall of **0%**). The Decision Tree model's overall weighted average F1-score is **0.89**, matching that of the SVM.

Random Forest Classifier:

The Random Forest classifier achieved the highest accuracy of **91.78%** among the three models. It maintained similar precision and recall patterns as the SVM and Decision Tree for most classes, with standout performance on class **3** (precision and recall both at **95%**), indicating strong identification of instances from this class. Like the other models, it struggled with classes **0** and **2**, both showing **0%** precision and recall. The weighted average F1-score for Random Forest is **0.89**, aligning closely with SVM and Decision Tree.



VISUALIZATION DISPLAYING CONFUSION MATRICES FOR DIFFERENT CLASSIFICATION ALGORITHMS

REGRESSION RESULT ANALYSIS

Linear Regression:

- The Linear Regression model exhibits robust performance metrics across both training and testing datasets. It achieves an R-squared value of 0.79, indicating that the model explains 79% of the variance in the target variable, '**SalePrice**'. Despite this solid explanatory power, the Root Mean Squared Error (RMSE) values are notable: the training RMSE stands at 35,679.74, while the testing RMSE is slightly higher at 40,124.08. These RMSE values reflect the average magnitude of error in the predicted sale prices compared to the actual values. The higher RMSE in the testing set suggests that the model may encounter challenges in accurately predicting sale prices for unseen data points. Therefore, while the model captures a substantial portion of the variance in sale prices, efforts to reduce prediction errors further could enhance its reliability and applicability in real-world scenarios.

Random Forest Regressor:

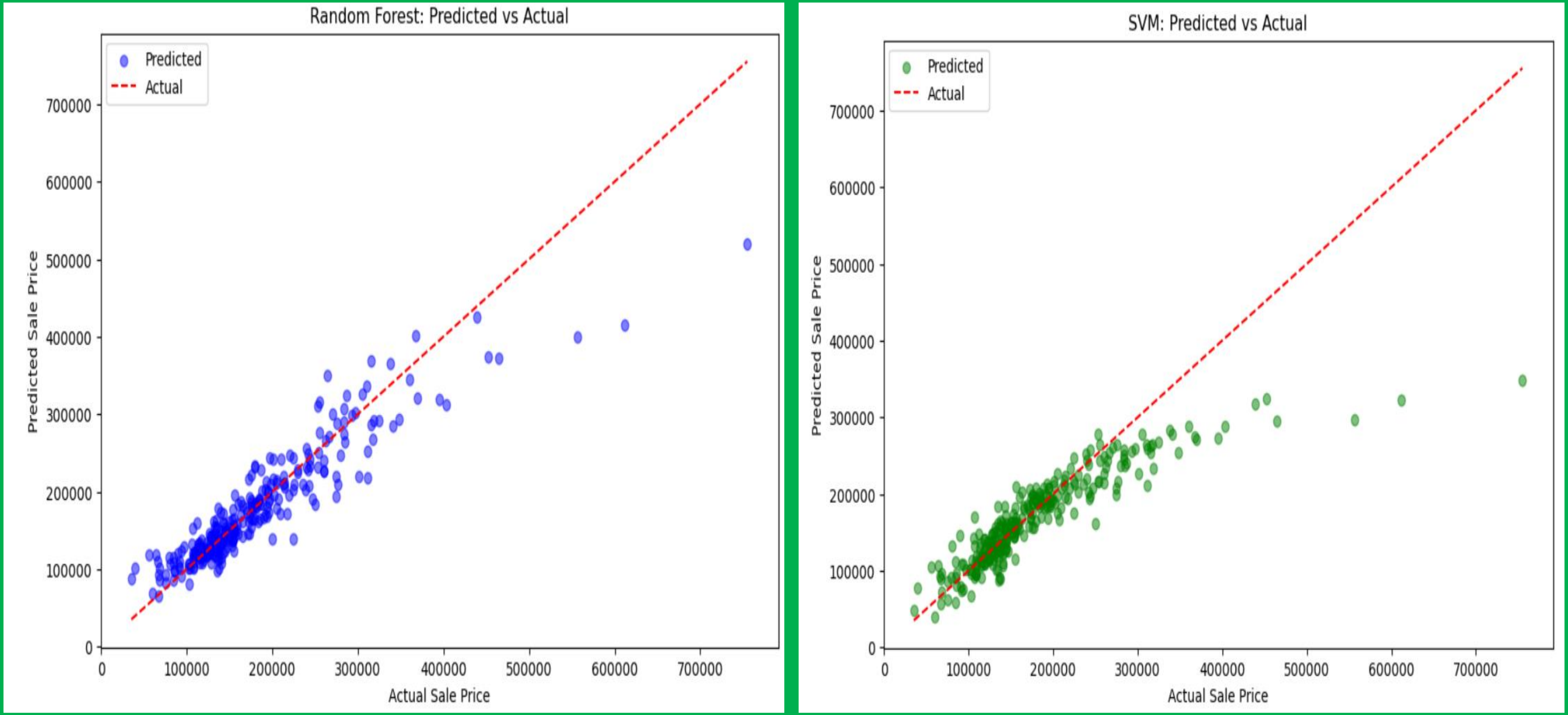
- The Random Forest Regressor demonstrates superior performance across both training and testing datasets, leveraging optimized hyperparameters. With a Best CV RMSE of 34,940.56, the model showcases strong predictive capability during cross-validation. In terms of predictive accuracy, the model achieves a Training RMSE of 16,692.76 and a Testing RMSE of 33,856.77. These RMSE values indicate the average magnitude of error in predicting '**SalePrice**', with the lower testing RMSE compared to Linear Regression highlighting the Random Forest's ability to generalize effectively to unseen data. This improvement is attributed to the model's capacity to capture intricate relationships between features and the target variable, utilizing ensemble learning to enhance predictive accuracy beyond the capabilities of simpler linear regression methods. Overall, the Random Forest Regressor's robust performance underscores its practical utility in real-world applications, particularly in predicting housing prices with greater accuracy and reliability.

Support Vector Machine (SVM):

- The SVM (Support Vector Machine) Regressor, utilizing a linear kernel and optimized hyperparameters {'C': 100, 'degree': 2, 'epsilon': 0.1, 'kernel': 'linear'}, demonstrates the highest RMSE values among the three models, with a Training RMSE of 39,992.64 and a Testing RMSE of 46,117.37. The Best CV RMSE is 41,053.37, indicating its performance during cross-validation. The R-squared values are 0.7318 for the training set and 0.7227 for the testing set, which is slightly lower than those observed with Linear Regression. These metrics suggest that the SVM model with a linear kernel provides a comparatively weaker fit, as evidenced by its higher RMSE and lower R-squared values, reflecting less accuracy in capturing the variance in the target variable **'SalePrice**'.

Conclusion Result Analysis:

- In summary, Random Forest outperforms both Linear Regression and SVM in terms of RMSE on the testing set, suggesting it provides the most accurate predictions among the models evaluated here. Linear Regression falls behind Random Forest in predictive accuracy, likely due to its assumption of linear relationships between features and the target variable, which may not hold in more complex datasets. SVM, while offering competitive R-squared values, exhibits higher RMSE values, indicating it struggles more with prediction errors compared to Random Forest.
- In practical terms, if the primary goal is accurate prediction with minimal error, Random Forest would be the preferred choice among the three models evaluated here. Its ability to handle non-linear relationships and interactions between features makes it well-suited for a wide range of predictive tasks. However, the choice of model should also consider other factors such as computational efficiency, interpretability, and scalability depending on the specific application context.
- Therefore, based on the provided metrics and analysis, **Random Forest** emerges as the best-performing model for this particular dataset, offering the lowest RMSE on the testing set and demonstrating robust predictive capabilities.



VISUALIZATION DISPLAYING PREDICTED VS ACTUAL FOR DIFFERENT REGRESSION ALGORITHMS

CONCLUSION

- The dataset comprises a comprehensive array of features related to residential properties, including characteristics such as lot size, building class, zoning classification, and various amenities. With a total of 81 features, each row represents a distinct property, allowing for detailed analysis and predictive modeling. The dataset's richness enables a multitude of analyses, ranging from classification tasks, such as predicting zoning classifications, to regression tasks, such as estimating sale prices. For classification, the target variable is 'MSZoning,' which categorizes properties into different zoning classifications, while for regression, the target variable is 'SalePrice,' representing the sale price of the properties. By leveraging the dataset's diverse features, stakeholders in the real estate sector can gain valuable insights into property valuation, market trends, and neighborhood dynamics, ultimately informing strategic decisions in investment, development, and urban planning endeavors.
- Through a series of statistical analyses and modeling techniques, our exploration of the dataset has unveiled valuable insights into various facets of real estate dynamics. Model performance assessments showcased the strengths and weaknesses of different modeling approaches, with linear regression excelling in predictive accuracy, while decision tree classification demonstrated superior classification abilities. Association analyses, including chi-square tests and correlation assessments, elucidated significant relationships between variables, such as zoning classifications and street types, as well as the correlation between lot area and property sale prices. Moreover, the impact of neighborhood location on property values emerged prominently, as ANOVA results highlighted considerable variations in sale prices across different neighborhoods. These insights carry significant implications for stakeholders ranging from investors and developers to urban planners and policymakers, informing decision-making processes in property investment, urban development, and community planning endeavors. By leveraging these insights, stakeholders can make informed decisions that align with market dynamics, optimize investment strategies, and contribute to the creation of vibrant, sustainable communities.