###### DATA SCIENCE PROJECT ON HOUSE PRICING ANALYSIS

Final project report submitted to

Graduate School at

The University of Arkansas at Little Rock

in fulfillment of

requirements for the graduate course of

202510 Spring: Data Science - Technologies (9U1)

in

May 2025

by

Deepak Singla

Instructors

Dr. Elizabeth Pierce

and

Dr. Serhan Dagtas

[INTRODUCTION 3](#_Toc197385572)

[Background and Context 4](#_Toc197385573)

[Current Knowledge 4](#_Toc197385574)

[Missing Knowledge 4](#_Toc197385575)

[Purpose 4](#_Toc197385576)

[LITERATURE REVIEW 5](#_Toc197385577)

[Data Selection, Research Design and Methods 6](#_Toc197385578)

[**Required Libraries** 7](#_Toc197385579)

[Load Data 7](#_Toc197385580)

[Initial Data Inspection: Let's look at the first few rows, data types, and summary statistics. This initial inspection helps identify the number of entries, columns, data types (numerical vs. categorical/object), presence of missing values (comparing non-null counts to total entries), and the basic statistical profile of numerical features (mean, median, min, max, standard deviation). 8](#_Toc197385581)

[To get the top 5 rows of the data, data frames head() is used as shown in figure below. 8](#_Toc197385582)

[We can see that there are some NaN values and there are some data time data also in some rows beyond 5 rows. This data needs to be cleaned before using it in EDA. 8](#_Toc197385583)

[Data Clean Up: After initial inspection, data shows that there are some inconsistencies in the data which might be problem while doing an exploratory data analysis on this sample data so it should be cleaned 10](#_Toc197385584)

[Analysis Work 12](#_Toc197385585)

[Exploratory Data Analysis (EDA) 12](#_Toc197385586)

[Target Variable Analysis 12](#_Toc197385587)

[ Distribution Plot: 12](#_Toc197385588)

[This plot typically shows the frequency of houses sold at different price points. For the Ames dataset, it reveals a right-skewed distribution, meaning most houses sell at lower to mid-range prices, with a long tail extending towards higher prices. 14](#_Toc197385589)

[ Log Transformation: Since many models perform better with normally distributed data, and the evaluation metric for this dataset often uses log-transformed prices, we analyze the log-transformed SalePrice. We use np.log1p which calculates log(1+x) to handle potential zero values gracefully. 14](#_Toc197385590)

[*print("\nLog-transformed SalePrice analysis skipped.")* 14](#_Toc197385591)

[This plot shows the distribution of the logarithm of the sale prices. It typically appears much more symmetrical, resembling a normal (bell-shaped) distribution, confirmed by a skewness value closer to 0. This transformation makes the target variable more suitable for certain modeling techniques. 15](#_Toc197385592)

[Numerical Value Analysis 16](#_Toc197385593)

[Discussion of Results 18](#_Toc197385594)

[Conclusions and Closing Thoughts 18](#_Toc197385595)

[Appendix 19](#_Toc197385596)

[References 20](#_Toc197385597)

# INTRODUCTION

This project will demonstrate some of the methods to analyze the housing price based on the different parameters in the Housing dataset. For the purpose of the demonstration Ames Housing Dataset has been taken from Kaggle.com.

House price analysis captivates my interest due to its practical real-world applications. Understanding the dynamics that drive property values, from macroeconomic trends to local market conditions, offers valuable insights into investment strategies, urban development, and even individual financial well-being. The ability to identify patterns, predict future movements, and ultimately inform sound decisions within such a significant sector of the economy makes this field highly relevant.

Understanding the drivers behind house prices is crucial for many parties including buyers, sellers, and real estate professionals. This project uses the Ames Housing dataset, data describing residential home sales in Ames, Iowa, between 2006 and 2010. It contains 79 explanatory variables detailing nearly every aspect of a home, making it ideal for in-depth analysis. Our goal is to explore these features and visualize their relationship with the Sales Price of the property

Background and Context: House price analysis is a well-established field, crucial for understanding economic trends, investment decisions, and housing market dynamics. Historical data on sales, prices, and property characteristics, coupled with economic indicators like interest rates, employment, and GDP, form the foundation of this analysis. Various methodologies, from simple statistical measures like median price to complex econometric models, are employed to interpret past trends and forecast future movements. The real estate market's sensitivity to economic cycles and government policies has long been recognized, leading to extensive research on these relationships.

Current Knowledge: A significant amount is already known about house price analysis. Sophisticated statistical and machine learning models are used to predict price movements, assess property values, and identify market bubbles. Geographic Information Systems (GIS) are increasingly integrated to analyze spatial patterns and the impact of location-specific factors. Real-time data and online platforms provide accessible information on market trends, sales volumes, and property listings. Indices like the S&P CoreLogic Case-Shiller Home Price Index and the National Association of Realtors' data offer comprehensive insights into price changes at national and regional levels. Furthermore, behavioral economics is being applied to understand the role of sentiment and expectations in housing market fluctuations.

Missing Knowledge: Despite the advancements, gaps remain. Predicting turning points in the market with high accuracy is still challenging due to unforeseen economic shocks and shifts in buyer behavior. The impact of emerging factors like climate change risks and the increasing prevalence of remote work on long-term property values needs further investigation. Granular, real-time data at the neighborhood level is not always consistently available or easily integrated across different sources. Understanding the precise mechanisms through which social media sentiment and information diffusion affect housing prices requires more research. Moreover, incorporating qualitative data, such as community perceptions and quality of life indicators, into quantitative models remains a complex task. A more nuanced understanding of the heterogeneity of buyer and seller behavior and its impact on market dynamics is also needed.

Purpose: The purpose of this report is to explore the different ways of analyzing housing data which can further be used in training models which can predict the house pricing in future.

# LITERATURE REVIEW

It is important to demonstrate that you are familiar with other data science investigations on your topic. In this section, share how your project will fit into the other conversations about your topic by comparing and contrasting the main theories, methods, and debates. This is also an opportunity to examine the strengths and weaknesses of different approaches or to explain how you build on, challenge, or incorporate prior scholarship.

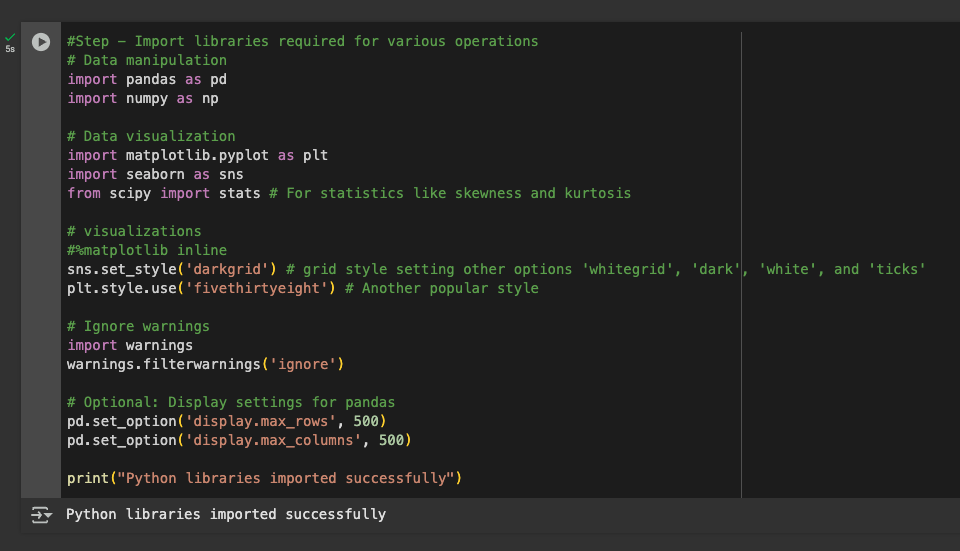
Useful resources: [Guidelines for Writing a Literature Review](https://www.scribbr.com/dissertation/literature-review/) and [Literature Review Steps](https://library.dsu.edu/c.php?g=22495&p=133184)

# Data Selection, Research Design and Methods

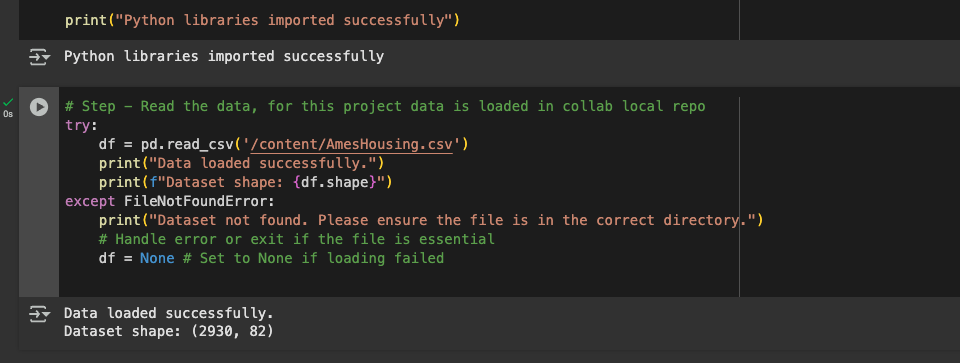
After going through numerous datasets available on [Kaggle.com/dataset](http://kaggle.com/datasets) and few other open-source datasets from [geeksforgeeks.com](http://geeksforgeeks.com/). I have chosen dataset from Ames Residential Project Dataset. The Ames Housing dataset has data describing residential home sales in Ames, Iowa, between 2006 and 2010. It contains 79 explanatory variables detailing nearly every aspect of a home.

The following steps will help in inspecting the dataset. Based on the quality of data, we can choose what operations would be required to use the dataset. The Kaggle dataset in its original form is not directly usable. While just glancing over the dataset in the file you can find some null values in some of the columns, data time values and some bad data like NaN values etc. These values will be a problem during the Exploratory Data Analysis phase. So, to clean up the data and transform it into usable for EDA following operations were performed.

**Required Libraries** This code has been written in Python and executed in Collab.google.com. To perform the operations on data, we need to import some python libraries as below



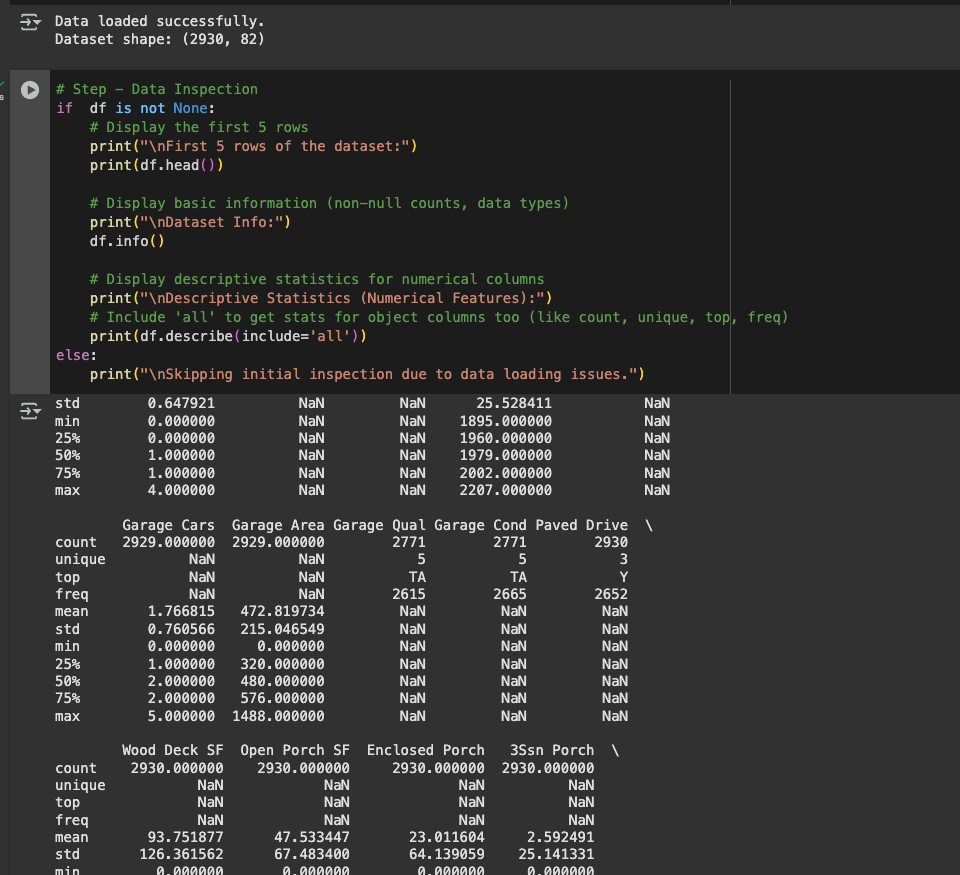
Load Data: After successfully importing Python libraries, we need to load the data and get a basic understanding of its structure. To load the sample dataset (AmesHousing.csv) pandas(pd) pd.read(filepath) method is used which return a result of type data frames as below



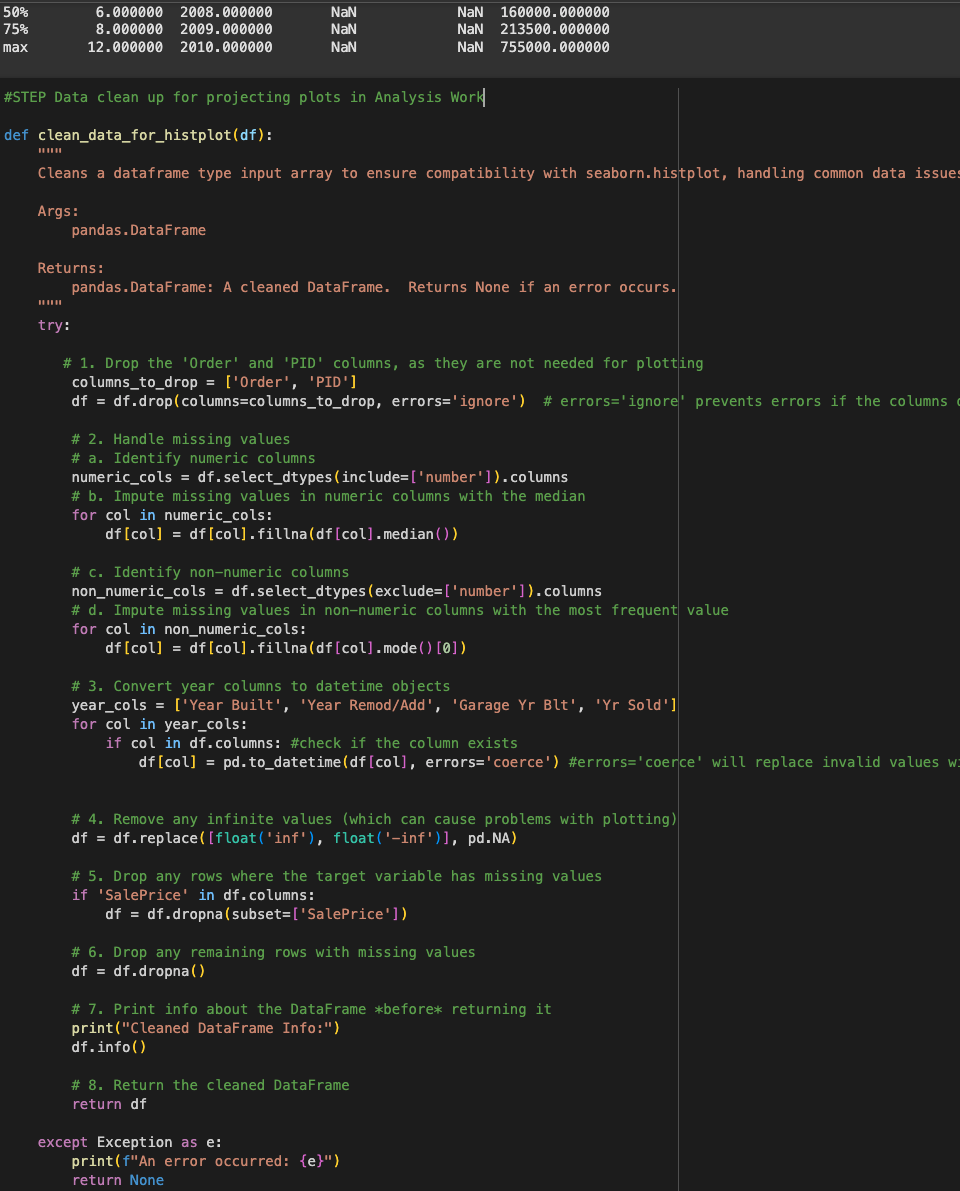
# Initial Data Inspection: Let's look at the first few rows, data types, and summary statistics. This initial inspection helps identify the number of entries, columns, data types (numerical vs. categorical/object), presence of missing values (comparing non-null counts to total entries), and the basic statistical profile of numerical features (mean, median, min, max, standard deviation).

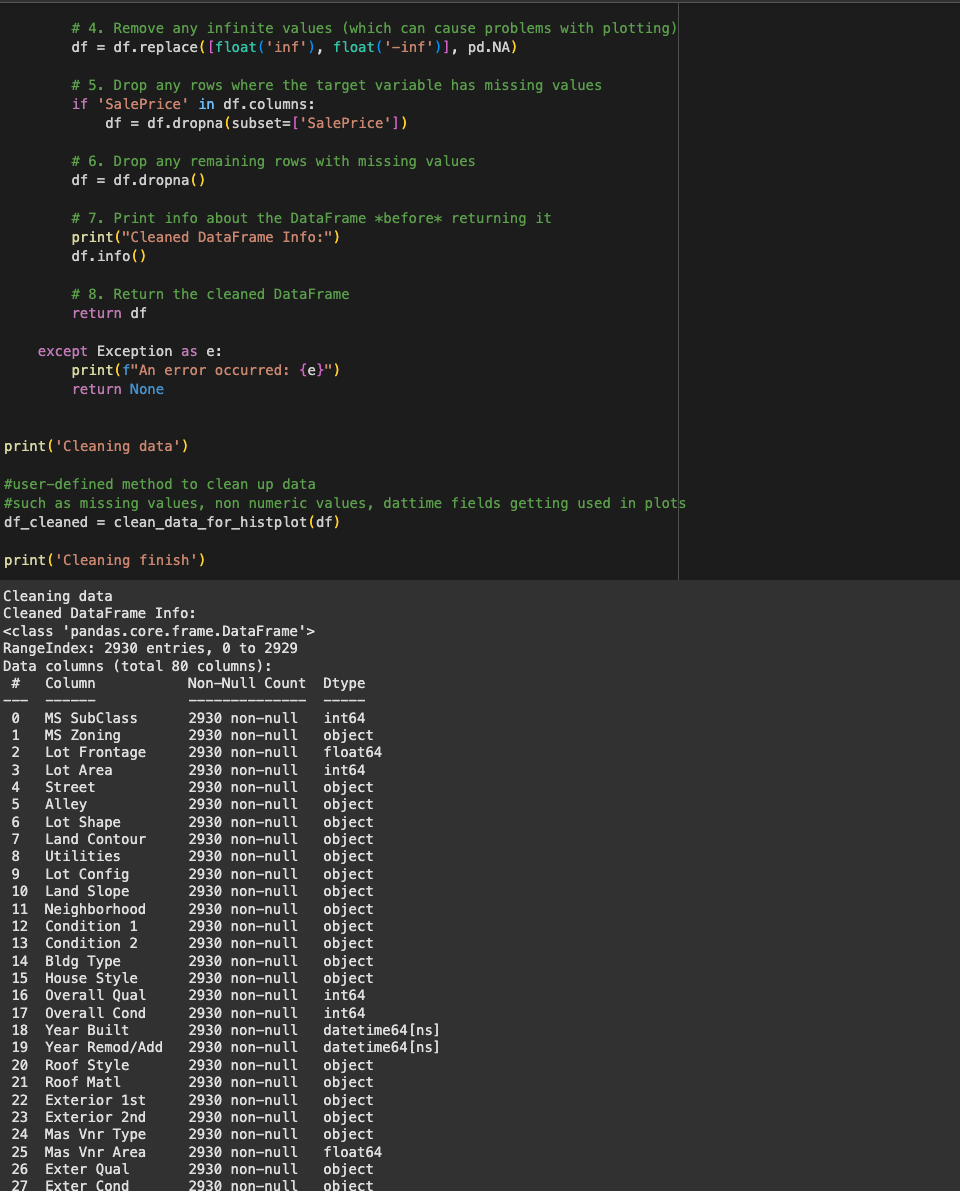
# To get the top 5 rows of the data, data frames head() is used as shown in figure below.

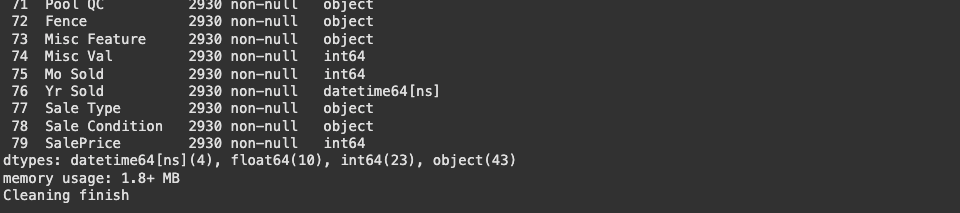
# We can see that there are some NaN values and there are some data time data also in some rows beyond 5 rows. This data needs to be cleaned before using it in EDA.



# Data Clean Up: After initial inspection, data shows that there are some inconsistencies in the data which might be problem while doing an exploratory data analysis on this sample data so it should be cleaned







# Analysis Work

## Exploratory Data Analysis (EDA)

EDA is crucial for understanding the data's characteristics, identifying patterns, and visualizing relationships. Sale Price is the target variable in my research. To analyse the effect of all the explanatory variables in the dataset on the Sale price of the house. I am going to use two different ways

Target Variable Analysis **(**Sale Price**):** To do the EDA of the data understanding the distribution of our target variable is key.  To understand the distribution, I am going to use two different kind of plots

* Distribution Plot: To draw a histogram combined with a Kernel Density Estimate (KDE) to visualize the distribution.

# Distribution Plot

*if df is not None:*

*print(df\_cleaned.info())*

*plt.figure(figsize=(10,6))*

*sns.histplot(data=df\_cleaned, x='SalePrice')*

*#plt.show()*

*plt.title('Distribution of SalePrice')*

*plt.xlabel('Sale Price ($)')*

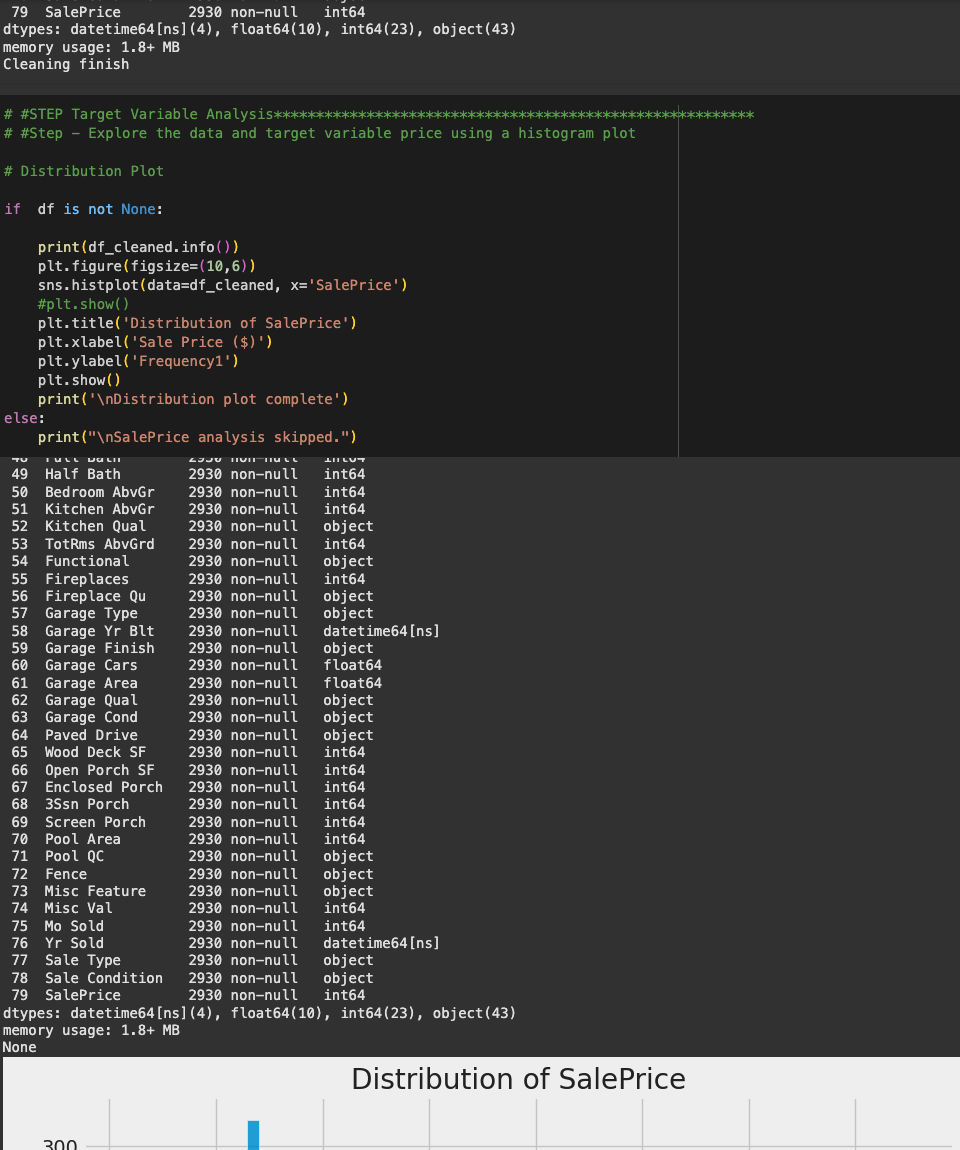
*plt.ylabel('Frequency1')*

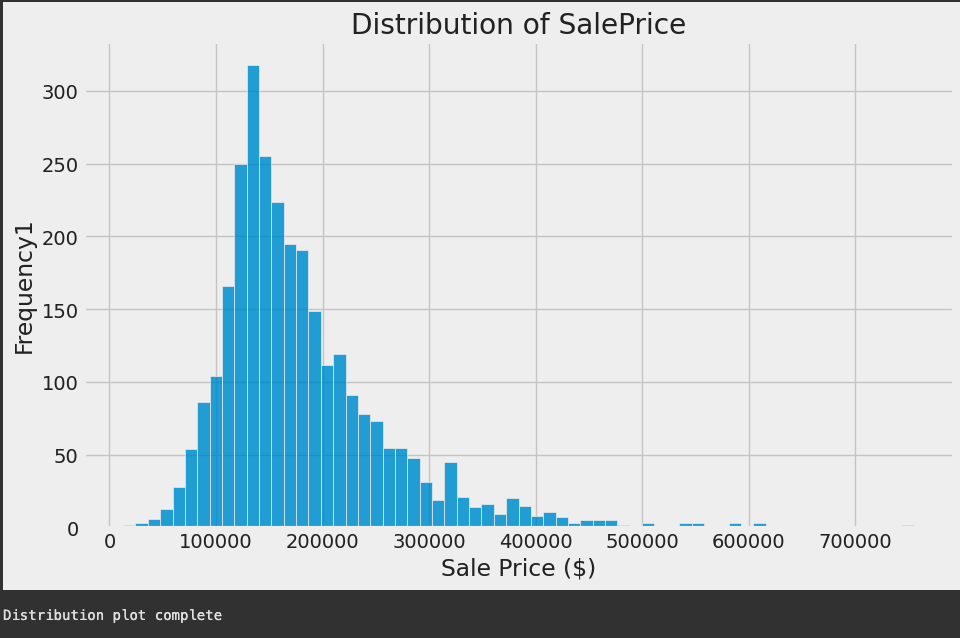
*plt.show()*

*print('\nDistribution plot complete')*

*else:*

*print("\nSalePrice analysis skipped.")*





# This plot typically shows the frequency of houses sold at different price points. For the Ames dataset, it reveals a right-skewed distribution, meaning most houses sell at lower to mid-range prices, with a long tail extending towards higher prices.

# Log Transformation: Since many models perform better with normally distributed data, and the evaluation metric for this dataset often uses log-transformed prices, we analyze the log-transformed SalePrice. We use np.log1p which calculates log(1+x) to handle potential zero values gracefully.

*# Log Transformation*

*print('Log Transformation')*

*if df\_cleaned is not None:*

*# Apply log1p transformation*

*train\_df = np.log1p(df\_cleaned['SalePrice'])*

*plt.figure(figsize=(10, 6))*

*sns.histplot(train\_df, kde=True)*

*plt.title('Distribution of Log-Transformed SalePrice')*

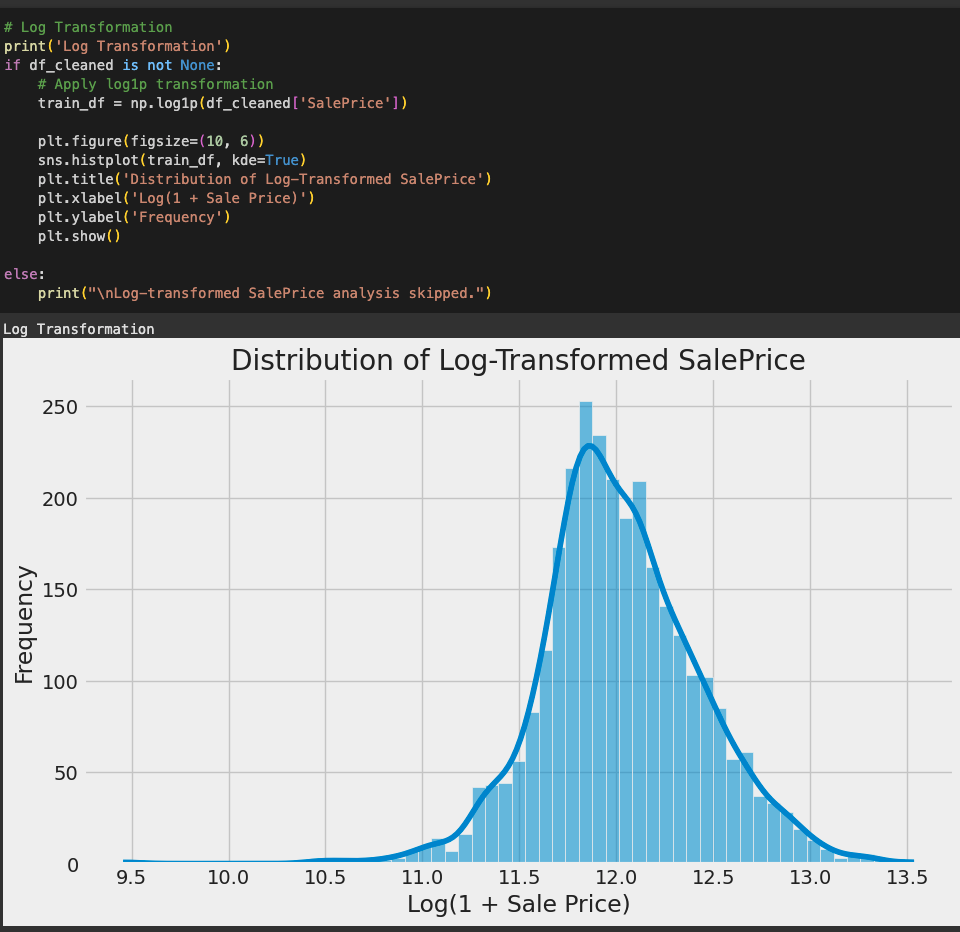
*plt.xlabel('Log(1 + Sale Price)')*

*plt.ylabel('Frequency')*

*plt.show()*

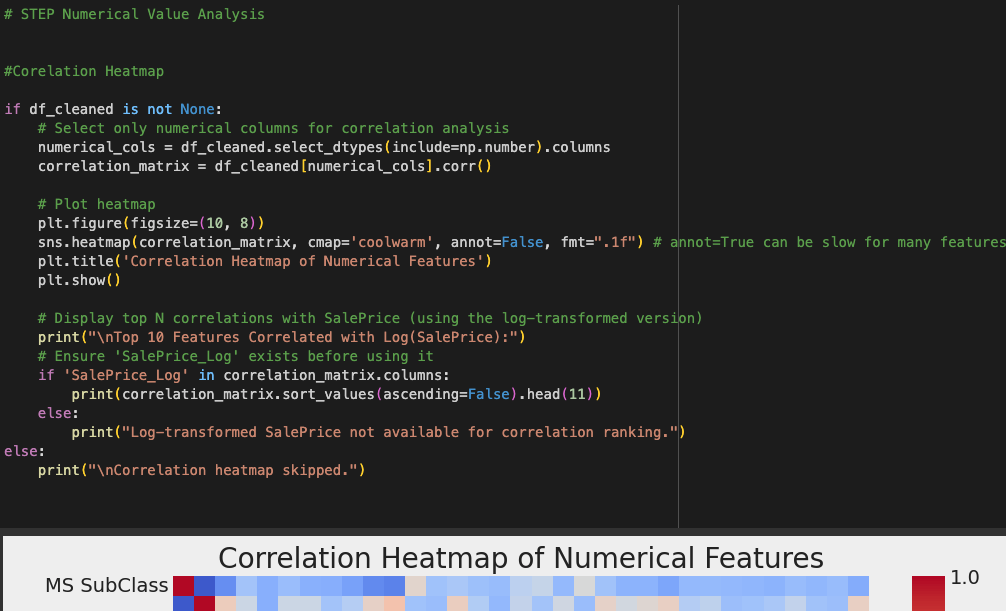
*else:*

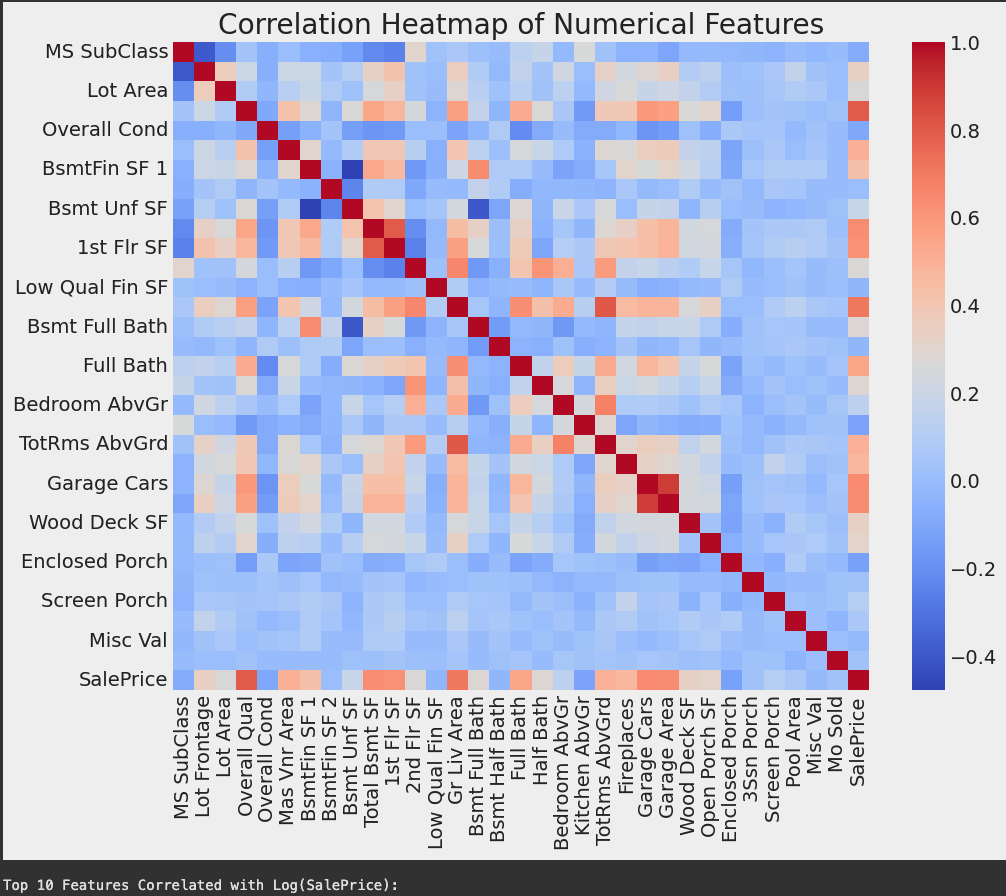
# *print("\nLog-transformed SalePrice analysis skipped.")*

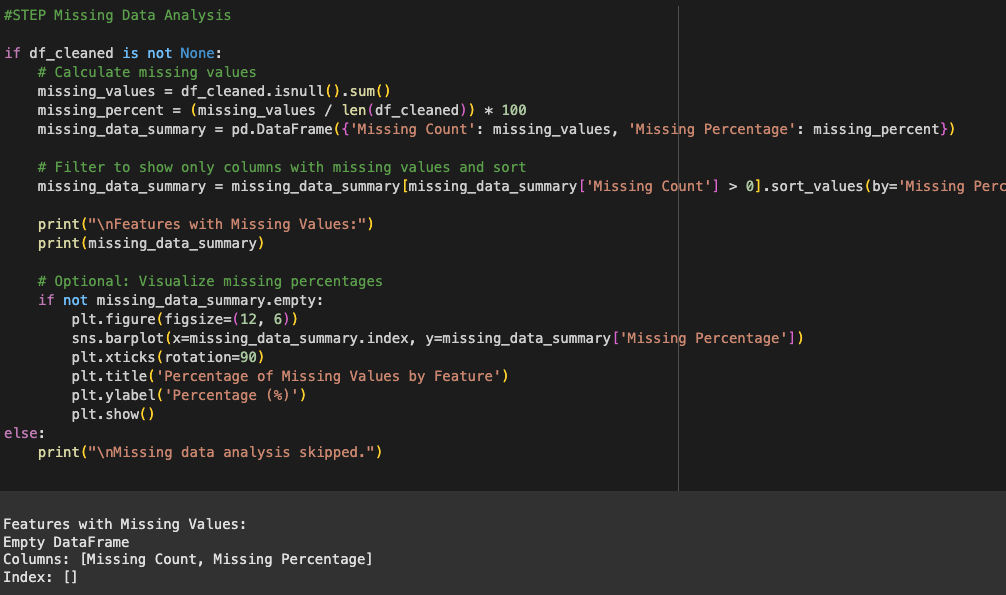


# This plot shows the distribution of the logarithm of the sale prices. It typically appears much more symmetrical, resembling a normal (bell-shaped) distribution, confirmed by a skewness value closer to 0. This transformation makes the target variable more suitable for certain modeling techniques.

Numerical Value Analysis **(**Sale Price**):** To do the EDA of the data understanding the distribution of our target variable is key.  To understand the distribution, I am going to use two different kind of plots







# Discussion of Results

~~Use this section to explain how your results help answer each of your project objectives/questions.~~

Based on the research, the results section would summarize the findings obtained from applying the chosen exploratory data analysis (EDA) methods:

* **Descriptive Statistics:** This part of the results section would present a clear and concise summary of the key statistical measures of the house price data. This would include things like the mean, median, standard deviation, range, and distribution characteristics (skewness and kurtosis) of house prices. It would also describe these measures for the variables that influence house prices, such as size, location, and age.
* **Data Visualization:** This section would describe the visual patterns and insights revealed through the use of charts and graphs. It would explain trends, correlations, and outliers in the data. For example, it might describe how house prices vary across different locations, how they relate to size, or how they change over time.
* **Correlation Analysis:** Here, the results of the correlation analysis would be presented. It would detail the strength and direction of the relationships between house prices and the factors that influence them. It would highlight which factors have the strongest correlations with price and whether those relationships are positive or negative.

# Conclusions and Closing Thoughts

This project demonstrates a structured approach to analyzing house price data using EDA and visualization. Key findings from analyzing the Ames dataset typically include:

* SalePrice is right-skewed but approaches normality after log transformation.
* Features related to **overall quality** (OverallQual) and **size** (GrLivArea, TotalBsmtSF, GarageArea, engineered TotalSF) are strongly correlated with price.
* **Location** (Neighborhood) significantly impacts price.
* **Age** (YearBuilt, YearRemodAdd) plays a role, with newer/recently remodeled homes generally fetching higher prices.
* The data contains significant **missing values**, but many represent the absence of a feature (like a pool or garage) and require careful, context-aware handling.

Thorough EDA, as outlined with code snippets and chart descriptions, is fundamental to understanding the data's nuances and informing subsequent preprocessing, feature engineering, and modeling steps in any data science project.

# Appendix

Include your Project Schedule to show a timeline of your project showing what you did at each stage of the project and how many hours spent.

Reference List - Include correct citations for every source you have used.

Any additional project documentation that you would like to include to showcase your work.

Additional Resources include [Structure of a Data Analysis Report.](http://www.stat.cmu.edu/~brian/701/notes/paper-structure.pdf) and [Data Analysis Write-Ups](https://jgscott.github.io/teaching/writeups/write_ups/)

Example of a Dissertation demonstrating these report features - <https://www.proquest.com/pqdtlocal1009983/docview/2902800999/1AEA2A01CF584BEAPQ/4?accountid=14482&sourcetype=Dissertations%20&%20Theses>

File or Link to your 3 Minute Thesis. You can use a one slide PowerPoint with audio, or you can record a selfie video. Remember to do your summary talk in 3 minutes or less. Longer talks will not be viewed. There are many good references on the web on how to do a good 3-minute summary (e.g., <https://pipettegazette.uthscsa.edu/2020/08/21/10-tips-for-3-minute-thesis-competition/>)

# References

Report format has been copied from

<https://www.proquest.com/pqdtlocal1009983/docview/2902800999/fulltextPDF/C967C462E7DF4DF2PQ/1?%20Theses&accountid=14482&sourcetype=Dissertations%20>

<https://www.kaggle.com/datasets>

<https://www.geeksforgeeks.org/>