**HOUSE PRICE PREDICTION IN TAIWAN USING MACHINE LEARNING AND DATA ANALYTICS**

**Chapter 3: Methodology**

**3.1**

**Introduction**

A dissertation's research technique section is an essential part that offers an organized and methodical framework for doing research. It outlines the processes, strategies, and techniques used to collect, examine, and evaluate data, allowing the researcher to successfully address the research questions or hypotheses. This chapter walks the reader through the procedures and methods that form the basis of the study, acting as a blueprint for the investigation. One cannot stress the significance of a carefully considered study technique. It not only shows the audience how valid and rigorous the research is, but it also provides transparency, letting them know how the study was carried out and how the results were arrived at. This chapter acts as a link between the empirical data and conclusions that will be discussed later in the dissertation and the theoretical and conceptual framework that was previously established in the parts. The goal of this thorough study approach is to provide academics, professionals, and decision-makers with the information and resources needed to generate reliable predictions about Taiwanese real estate values. This dissertation aims to enhance the comprehension of Taiwan's housing market dynamics by using diverse data sources and analytical methods. This will facilitate the development of better-informed investment strategies and decision-making.

**3.2**

**Research Method**

There are many different research techniques that can be used in a dissertation, including action research, historical research, ethnography, case studies, mixed methods that combine both qualitative and quantitative techniques, and quantitative techniques like surveys and experiments. The research's questions, objectives, and data needs determine the method to choose. To best achieve their goals, researchers may choose to use one technique or a combination of them. The methods selected and their reasoning are explained in the dissertation's methodology section. Two essential methods for conducting research are qualitative and quantitative procedures. Investigating and comprehending non-numerical data, such as textual, visual, or observational information, is the focus of qualitative approaches. This method is appropriate for revealing deep insights and comprehending intricate social phenomena since it places an emphasis on context and depth. Quantitative methods, on the other hand, entail gathering and examining numerical data in order to do statistical analysis and find patterns or links. Generalizable conclusions can be drawn from objective, structured data obtained through quantitative research. Depending on the goals of the study, qualitative or quantitative approaches should be used. When a thorough grasp of a subject is required, as in the social sciences or market research, qualitative techniques are frequently employed. When numerical data is necessary for statistical analysis and extrapolating conclusions, as in scientific research and surveys, quantitative approaches are used. In order to capitalize on the advantages of both strategies and guarantee a thorough grasp of their research subject, researchers may also employ mixed methodologies. In this way, the study of Taiwan's house prediction is under qualitative techniques because it is a market research kind of analysis. The real estate market price according to the area and the units of the land data possesses a qualitative kind of nature. This study is followed by this study method.

**3.3**

**Research Philosophy**

Research philosophy in a dissertation refers to a collection of presumptions and ideas about the nature of knowledge and how it can be achieved. The entire strategy for the research process is determined by the research philosophy selected. Broadly speaking, there are three primary research philosophies that are employed.

**1.  Positivism**

This philosophy is based on the idea that techniques used in the natural sciences can also be applied to the study of society. The goal of positivism is to establish quantifiable, empirical, and objective knowledge. They place a strong emphasis on testing hypotheses and creating general rules or theories using quantitative data and statistical techniques.

**2.  Pragmatism**

Because pragmatism is a flexible philosophy, researchers are free to select the methodologies and procedures that best fit their goals and the study challenge. It recognizes the importance of both quantitative and qualitative data and methodologies, concentrating on the most effective means of addressing particular research problems.

**3. Interpretivism**

According to interpretivism, there are more complex relationships in the social world than just cause and effect. This school of thought focuses on understanding people's or groups' subjective meanings and perceptions. To investigate these individualized experiences, qualitative techniques like content analysis, observations, and interviews are frequently employed.

The prime selection of research philosophy is a crucial choice that has a big impact on the study design, data collection, analysis, and interpretation. Scholars frequently select a philosophy that is consistent with the nature of the issues they are trying to answer as well as the ontological and epistemological presuppositions they make. Certain dissertations may also include components from other ideologies, based on the objectives of the study and the intricacy of the research.

**3.4**

**Research Approach**

In a dissertation or research project, a research approach is a general plan or structure that is employed to carry out the investigation and address the research questions or goals. The entire research process, including technique selection, data collecting, analysis, and interpretation, is guided by the research methodology chosen (Mehrad *et al.* 2019). There are some popular research methodologies:

The deductive Approach method begins with a theory or hypothesis and aims to test it by gathering information and interpreting it in order to support or contradict the theory. The inductive approach to research starts with specific observations or data and builds general ideas or conclusions from these findings, in contrast to the deductive approach (Hall *et al.* 2021). The goal of abductive reasoning is to produce the best explanation for a collection of observations even in the absence of all relevant knowledge. It blends aspects of deductive and inductive reasoning. In situations where there is little to no prior research or material available, the exploratory research approach is employed to study the research problem. To obtain insights, it frequently entails gathering basic data and doing an open-ended study (Canlas *et al.* 2020). The goal of descriptive research is to present a thorough and accurate picture of a particular phenomenon or subject. It focuses on providing "what" answers and frequently uses surveys and observations. The goal of an explanatory study is to pinpoint the causes or reasons underlying a specific occurrence or phenomenon. It frequently makes use of experiments and statistical analysis and is centered on providing "why" answers. In order to test theories and spot trends, quantitative research collects and analyzes numerical data, frequently using surveys, experiments, and statistical methods. This is the main parameter of this analysis and this method is used here. Because this study depends on the survey data and the market price report. This thorough report helps to analyze the quantitative research approach in a very descriptive manner.

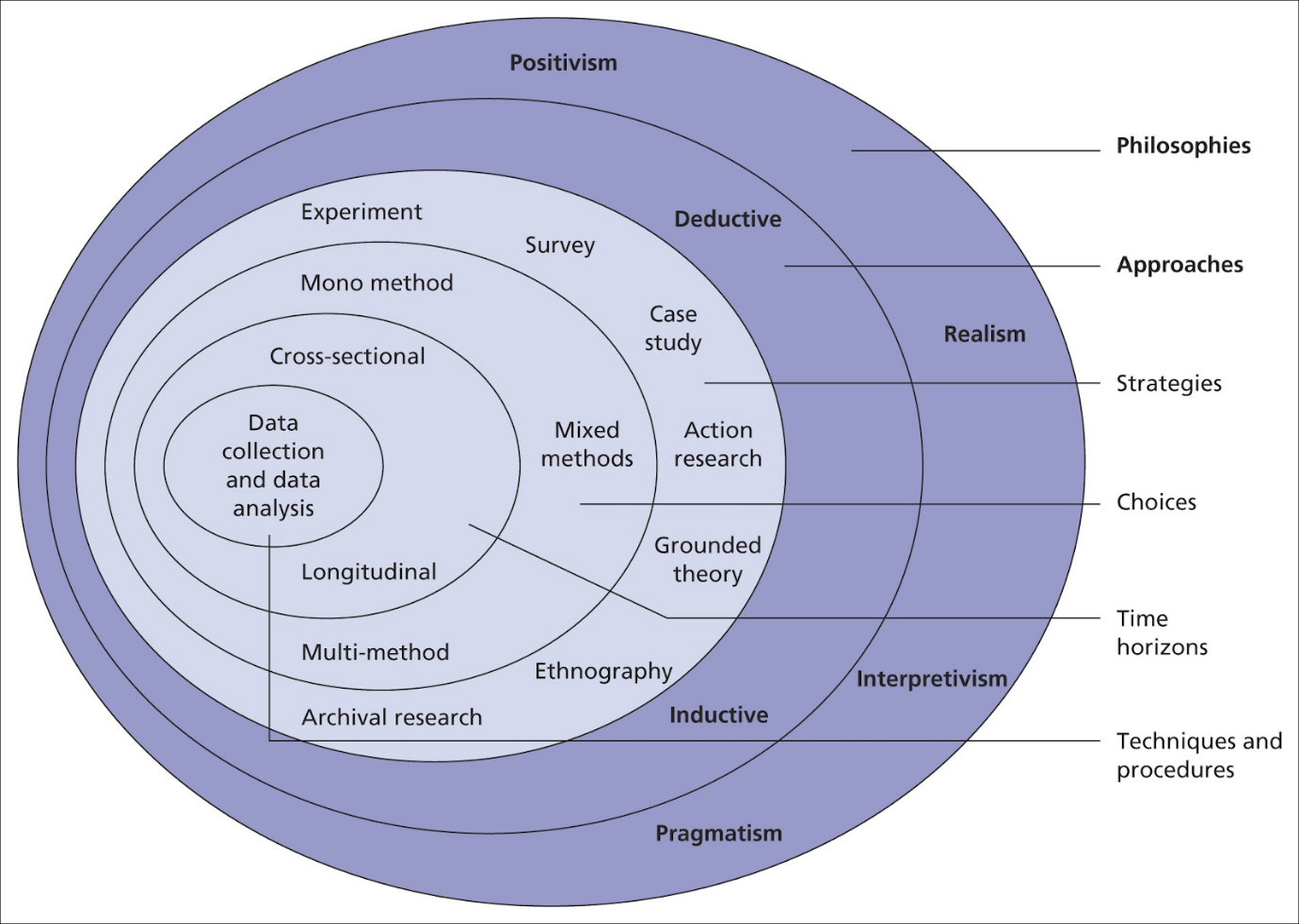
**3.5**

**Research Outline**

A dissertation research outline functions as an all-encompassing road map that directs the complete research effort in an organized and rational way. The dissertation begins with a Title Page that includes the title, author name, institutional affiliation, and date. The goals, procedures, major discoveries, and conclusions of the study are succinctly outlined in the abstract. For ease of navigation, a list of the chapters, sections, and subsections is included in the Table of Contents, along with page numbers. Readers can discover any visual aids used in the dissertation with the use of the List of Figures and Tables. Acknowledgments are optional and provide the author a chance to thank others (Gläser *et al.* 2022). The research problem, justification for the study, aims, and significance of the study are all introduced in the Introduction section, which also sets the scene. A review of the literature lays out the theoretical underpinnings of the study and points out any shortcomings. The research concept, methodology, data collecting and analysis strategies, sample tactics, and ethical considerations are all described in detail. The acquired data is visually presented in Data Presentation, and the findings are analyzed and discussed in Data Analysis and Results (Oghazi *et al.* 2022). The discussion part provides an interpretation of the data, draws links to the research objectives, and places the findings in the perspective of previous studies. Key findings are summed up, research problems are addressed, and contributions to the field are highlighted in the Conclusion (Buschle *et al.* 2022). The Appendices include further information if needed, while the References identify all sources cited in the dissertation. With the flexibility to be customized in accordance with particular needs and guidelines from the academic institution, this descriptive outline offers a thorough foundation for organizing a dissertation. This study depends on Taiwan’s house prediction which is based on the secondary dataset based on another’s servery record. This analysis format follows the traditional dissertation rules and its authenticity according to the structure.

**3.6**

**Research Onion**

****

**Figure 1: Research Onion**

(Source: https://15writers.com)

**3.7**

**Research Strategy**

The general plan or methodology that directs the course of a research endeavor, including a dissertation, is known as a research strategy. It describes the approaches and strategies that must be applied in order to successfully address the goals or research questions (Jiang  *et al.* 2020). This is a succinct description of a research strategy: A research strategy is a project's road map that specifies how the study will be conducted. It includes choices about research approach (e.g., deductive, inductive, exploratory), research philosophy (e.g., positivism, interpretivism, or pragmatism), data collection methods (e.g., surveys, interviews, experiments, etc.), data analysis techniques (qualitative or quantitative), and the study's overall design. Selecting a research strategy is essential since it affects the reliability and validity of the findings (Nickel  *et al.* 2020). By deciding on the project essentials this analysis maintains the primary analysis method on the secondary data collection on Taiwan’s house prediction, In academic and research settings, having a clear and well-defined research strategy is crucial for defining research objectives and questions (Beres  *et al.* 2020). It increases productivity by simplifying the research process, improves the caliber and coherence of research, synchronizes operations with overall objectives, takes ethical considerations into account, and permits flexibility when new information becomes available (Oliva  *et al.* 2019). In addition, it helps with decision-making, resource management, and communication, which all contribute to the effectiveness, rigor, and impact of research in answering difficult problems and expanding knowledge.

**3.8**

**Data Collection**

Primary and secondary data gathering are the two basic categories into which data collection techniques in dissertations can be generally divided. Directly obtaining original data through techniques including surveys, interviews, observations, and experiments is known as primary data collection. In order to answer certain research questions, researchers plan, carry out, and modify data collection tools as necessary (Lobe  *et al.* 2020). On the other hand, secondary data collecting entails utilizing pre-existing data sources, like government reports, government publications, archives, and publicly accessible information. To get new ideas and findings, researchers examine and combine this existing data. The study's objectives, resources, and design all influence the decision between primary and secondary data collecting. Frequently, a mix of primary and secondary data collection techniques (Archibald  *et al.* 2019). Depending on how the information is gathered and applied in a research setting, statistics on house prices in Taiwan may be primary or secondary (Zhang *et al.* 2023).  Primary data is that which is obtained by researchers directly from survey respondents, interview subjects, field workers, or on-site investigators. Primary data collection is the process of gathering original information for a particular study goal. Secondary data is what happens when researchers use pre-existing sources—like government records, real estate databases, or already published research—to get information on property prices (Roh *et al.* 2019). Though they did not actually acquire the data, researchers examine and use this pre-existing data to get ideas and findings. In this instance, the data is collected by others and used here which makes the data collected secondary.

**3.9**

**Data Analysis**

The use of secondary data analysis is crucial in augmenting the precision and profundity of predictions about property values in Taiwan. Secondary data can include a wide range of information, such as past property prices, economic factors, demographics, and housing market patterns. It is typically sourced from multiple existing databases (Hariri *et al.* 2019). Researchers and analysts can learn more about trends, correlations, and variables influencing property values by examining this pre-existing data. Predictive models that use these insights can yield more accurate projections of future housing costs. Furthermore, secondary data analysis offers an important background for comprehending the dynamics of Taiwan's real estate market and enables the identification of long-term market trends. Predictive modeling uses both historical and present data to help real estate agents make well-informed decisions. When a dataset is described as being based on secondary data collection, it indicates that information from pre-existing sources was gathered and compiled, as opposed to fresh surveys or experiments being conducted to acquire original data (Amarasinghe *et al.* 2020). Secondary data gathering in the context of Taiwan's home price prediction may involve publicly available data that might offer insightful information about the housing market, such as real estate databases, government records, historical sales records, or other data sources (Lester *et al.* 2020). Because secondary data makes use of information that is already available and may include a large range of historical and current data points, it can be an effective tool for analysis and prediction.

**3.10**

**Research Ethics**

Within the framework of a dissertation, research ethics constitute a vital element that sets forth the moral obligations and guidelines that researchers must follow when carrying out their investigations. These guidelines cover a wide range of important factors, such as participant informed consent, privacy and confidentiality protection, truthfulness and integrity in the reporting of results, a responsible research design that reduces harm and assures justice, and adherence to pertinent laws and regulations. Moreover, beneficence is emphasized by research ethics, which seeks to optimize benefits and reduce potential risks to study participants and society at large (Cascio *et al.* 2021). Important components of ethical research techniques include responsibility, cultural sensitivity, transparency, and data handling, Respecting these ethical guidelines is not only morally required but also essential to the credibility of research, protecting participants' rights and welfare, and maintaining the validity and integrity of the dissertation's conclusions (Drolet *et al.* 2023). To preserve the legitimacy and confidence that underpin the research process, researchers should strictly adhere to these ethical norms.

**3.11**

**Research Limitation**

The term "research limitations" describes the elements or restrictions that could affect a study's validity, generalizability, and scope. To provide readers with a clear idea of the limits and potential flaws in the research, it is crucial to recognize and discuss these constraints in a dissertation. Sample size restrictions, restrictions on data collecting, financial and resource limitations, and possible participant bias are examples of common study limitations. The study's scope may also be limited by ethical considerations, such as prohibitions on intrusive research techniques. Factors that may create limits include time constraints, presumptions in the study model, and the researcher's personal prejudices. Furthermore, the research may be impacted by outside variables that are beyond the researcher's control, such as changes in the economy or unanticipated events. Transparency demands that these limits be acknowledged and discussed, as doing so enables readers to evaluate the study's validity and usefulness in light of these restrictions. When designing and carrying out their studies, researchers should take techniques to mitigate limitations into account. Data Collection, Time, and resource constraints, among other issues, may restrict the extent of information that may be obtained. Data has a lot of missing values and the columns are not so useful according to the dominating result.

**3.12**

**Conclusion**

In summary, this dissertation's methodology chapter on Taiwan's housing price prediction is essential since it establishes the framework for an in-depth and methodical study strategy. This chapter functions as a research roadmap by thoroughly detailing the methods, approaches, and procedures. It highlights how crucial a well-thought-out research strategy is to the study's validity and transparency, providing details on the study's methodology and findings acquisition. Additionally, it does a good job of bridging the gap between the empirical facts and the conclusions that follow by tying them into well-established theoretical and conceptual frameworks. The analysis has been done despite the data being less transparent. The goal of the methodology chapter is to provide professionals, scholars, and decision-makers with the knowledge and resources needed to forecast Taiwanese real estate values with confidence. This is made possible by combining a variety of data sources and analytical techniques that, taken as a whole, improve our comprehension of the dynamic character of Taiwan's housing market. As a result, our study technique aids in the creation of more informed investment plans and decision-making procedures, thereby advancing our understanding of the variables affecting Taiwanese housing prices and their future trends. The dissertation aims to contribute significantly to the field of real estate analysis and prediction within this painstakingly created research framework.

**Chapter 4: Findings and Analysis**

**4.1 Introduction**

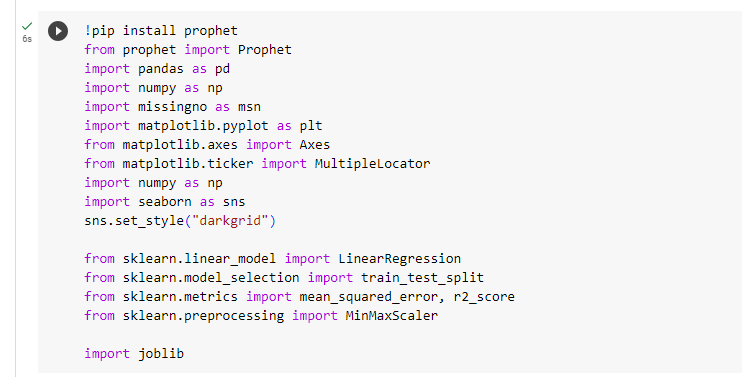
Findings and analysis are the most important part of this kind of analysis-based study. The examination of the selected data sets and the collection of the desired results based on the analysis. Here the dataset is based on the Taiwan’s house pricing, area, and districts. The prime aim of this analytical study is to predict the house price depending on the dataset based on the housing details by using the machine learning modules and then comparing them with the several components is the key to visualizing the results and the prediction rates. To use the data first they have to prepare for the algorithm-worthy format. The methods used here are data cleaning and data preparation to develop this kind of data as an algorithm. This project will help to predict the price in Taiwan and the analysis that has taken place depends on the linear regression analysis. The results of this analysis will help to understand the area per unit price and the total price depending on the total unit area. Overall this study will give the reader a comprehensive understanding of the challenges and complexities of price prediction along with the potential benefits of the implications of a desired and clear forecast.

**4.2 Analysis**

This analysis is based on Taiwan's housing data and the aim will be the prediction of those exact prices which is divided into several individual expenses in the dataset. This analysis will help to understand Taiwan’s housing price situation. For this housing price prediction, there are certain steps,

**Data collection:**

The first step is to gather the historical data of the housing components. That data should contain the housing price prediction components like main area, floors, per unit price, and so on. Collecting the data and choosing the data according to the merits is very important.

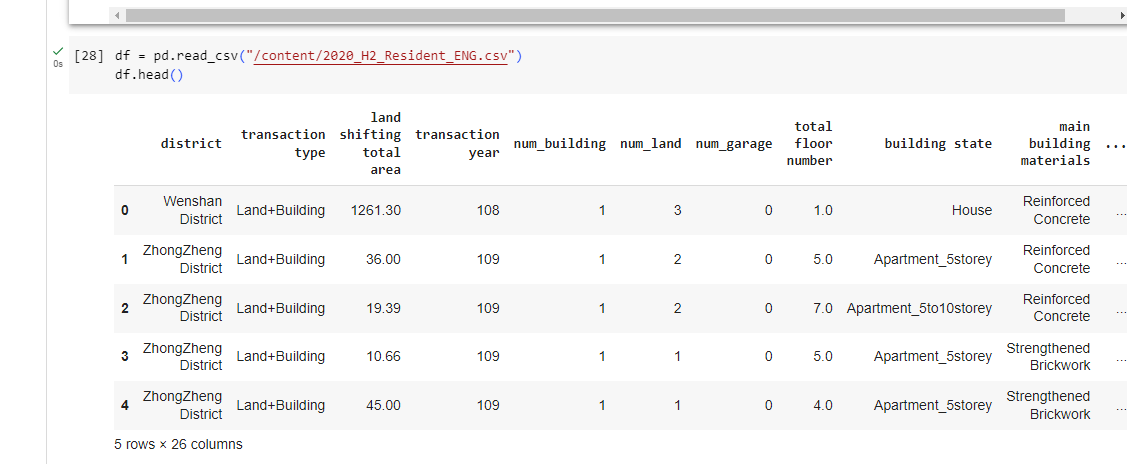
****

**Figure 1: Installing and importing libraries**

(Source: Self-created on Google Colab)

This part of the code has imported the essential libraries and installed the models by using the Python packages.

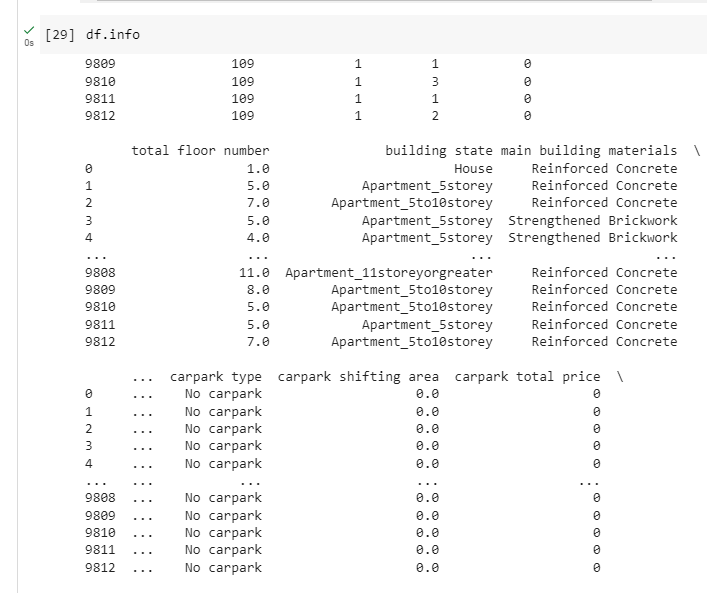
**Pre-processing of data:**

****

**Figure 2: Importing dataset**

(Source: Self-created on Google Colab)

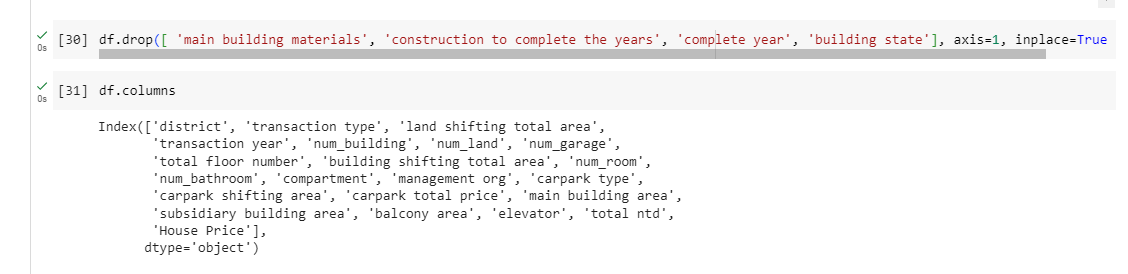
After choosing the data according to the desire importing the data in the code is a very crucial task. Then display the data for future convenience (Tang *et al.* 2020). This data has 26 columns.

****

**Figure 3: Data information**

(Source: Self-created on Google Colab)

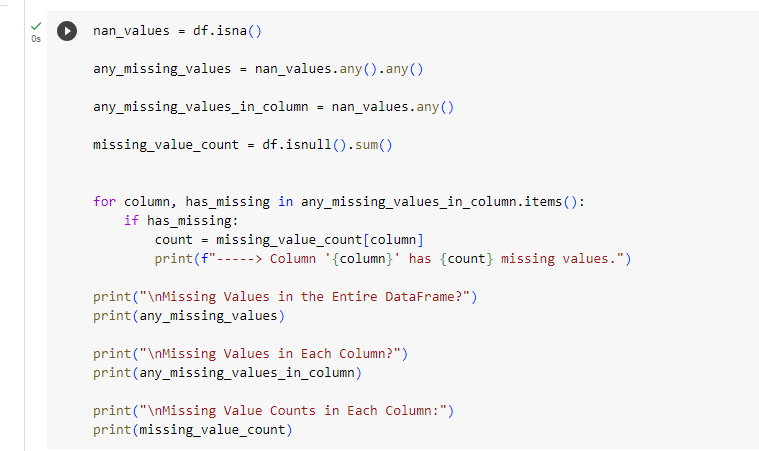
This data set has 9812 rows means 9812 data variants. This information chart gives detailed and deep information on the data (Nesca *et al.* 2022). Cleaning the data by handling missing values and outliers and by machine learning techniques convert the categorical data into numerical values.

****

**Figure 4: Drop the columns from the data set**

(Source: Self-created on Google Colab)

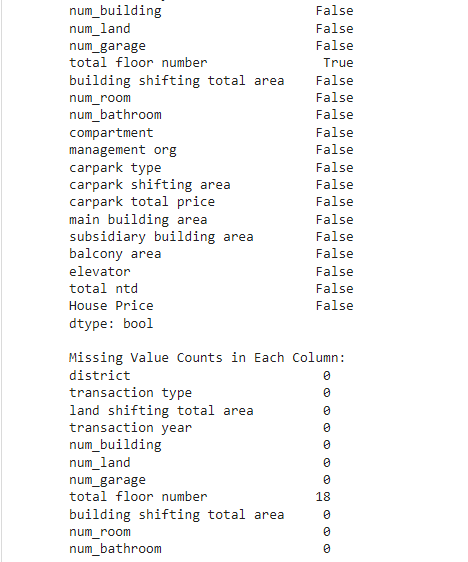
Here the unnecessary columns will drop and create a fresh dataset. These data columns have no usage in this study that is they have to drop. The dataset should only contain the necessary consecutive data column (Burdack *et al.* 2020). For this dataset building state, state of the building, and materials of the building are not necessary.

****

**Figure 5: Checking the missing values from the entire columns**

(Source: Self-created on Google Colab)

After the columns drop check the missing values from the entire data existing columns (Kuhlmann, 2021). That is very important to check the missing values before creating the models and implementing them.

****

**Figure 6: Finding the missing value in the total floor number column**

(Source: Self-created on Google Colab)

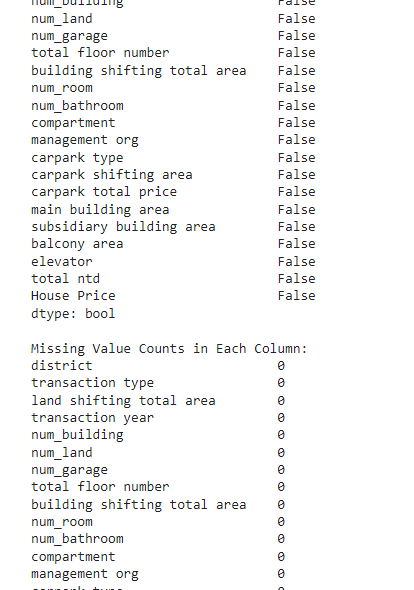
This report of the outcome has a clear message that the total floor number has the missing values and the number of the missing value is 18 in different columns (Emmanuel *et al.* 2021). This anomaly will have the potential to give the wrong result of the analysis because of these missing sets.

****

**Figure 7: Drop all missing values and replace them from columns**

(Source: Self-created on Google Colab)

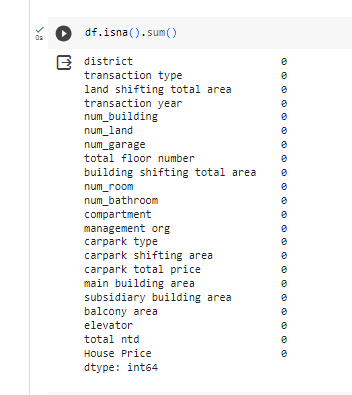
Implement the Python feature to drop the NA values in the first place. With this feature, all tables will have equally distributed numbers.

****

**Figure 8: Checking the missing value again**

(Source: Self-created on Google Colab)

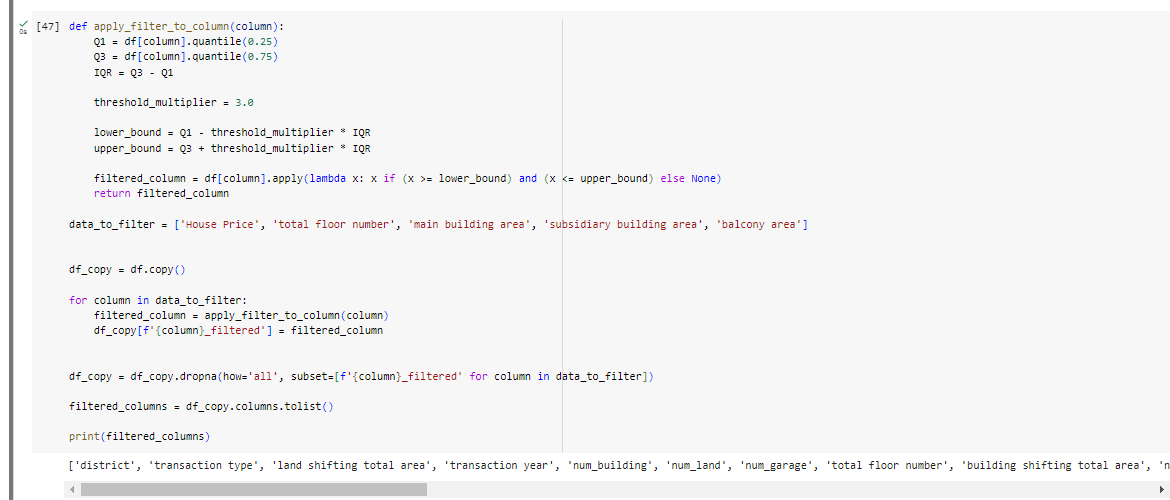
Checking after the dropping of NA values. This display shows that now all columns are well distributed and have no NA values (Park *et al.* 2021). Now this dataset is worthy of implementation in any machine-learning model. This data cleaning and pre-processing process is very straightforward but essential.

****

**Figure 9: Checking the dataset using another method**

(Source: Self-created on Google Colab)

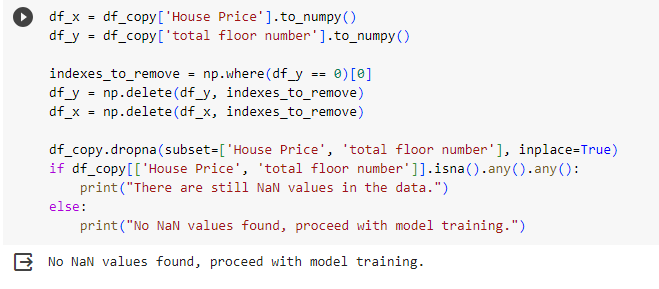
Showing the dataset, a second time to crosscheck by using the different features.

****

**Figure 10: Applying the filter on the desired columns**

(Source: Self-created on Google Colab)

This part of the code has to add the filters in desired columns and store the name of the columns in a data variable. Then use the pre-defined filtering function to filter the whole data columns using the quantile and the threshold multiplier to create the upper and lower bound.

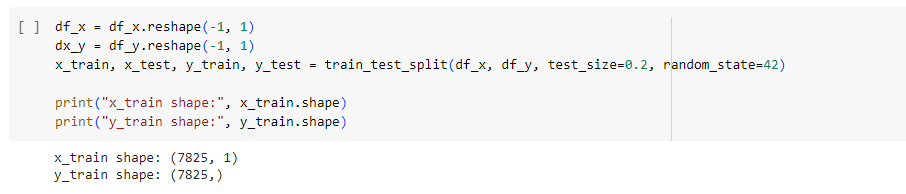


**Figure 11: Cheating NAN values preceding the dataset for the model**

(Source: Self-created on Google Colab)

Creating a function where the desired data columns will search and check whether there is any NAN value that exists or not. This check is very necessary for the model training after this testing the model that is selected for this data set can be used for better results than the predictions (Cui *et al.* 2020). After the check, the result is clearly visible that the desired columns have no NAN values so they are ready for further model inputs.

**Splitting the dataset:**



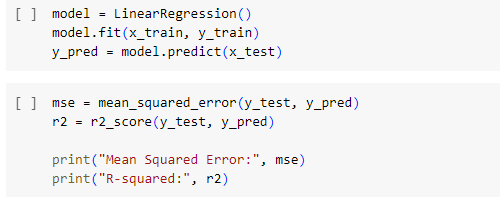
**Figure 12: Splitting the train and test dataset**

(Source: Self-created on Google Colab)

In this section the dataset has split into two parts first is the train and the second is for the test. By training the dataset using the data that is separate for the training the model will get more efficient. The split is fairly equal in this case the 7825 data from the 1 row has been allocated for the training and the 7825 is used to test the model (Nguyen *et al.* 2021). The result will decide how better they predict according to the data that is provided here. This is a very important step in machine learning-based algorithmic analysis.

**Model selection:**

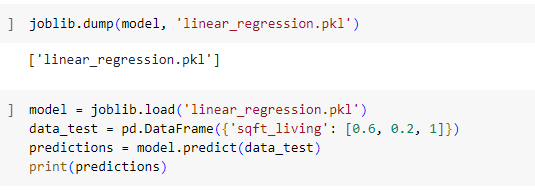
Now this part of the analysis has to choose a definite model for the prediction and here the chosen model is linear regression. Linear regression model may not be the prime and optimum choice for the situations where the relationship in between the existing variables is significantly high as non-linear. The simple data and the interpretable results along with the linear assumptions and the quick and inexpensive responses the linear regression is the best choice. linear regression is good for the baseline modelling and the diagnostic tool.



**Figure 13: Applying the Linear regression**

(Source: Self-created on Google Colab)

Here the linear regression is implemented on the training data model and as the required analysis data the test data is used. Along with the model setup, the consecutive mean squared and the r-squared data have been printed.



**Figure 14: Dumping the tested file in a. pkl file**

(Source: Self-created on Google Colab)

After the analysis the dump file will be allocated in the .pkl file named liner\_  
regression. Then print the accuracy result of the trained dataset.

**4.3 Result and Discussion**

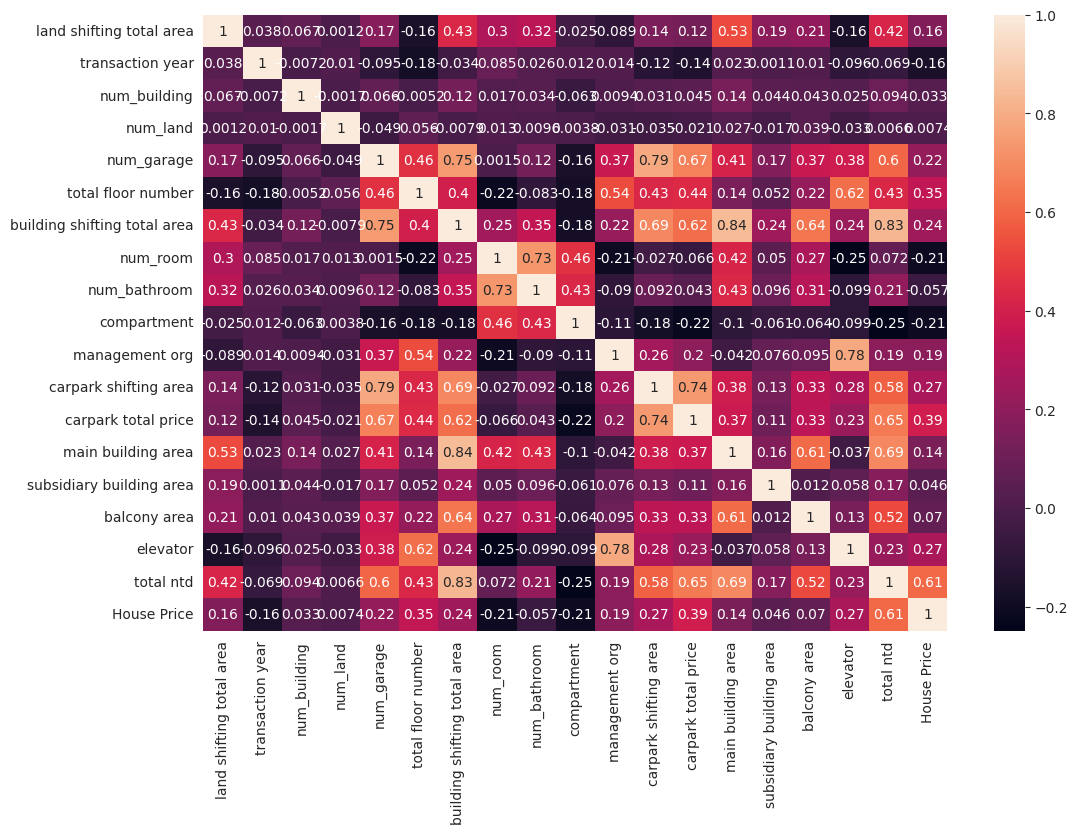
In this section, the main focus is on the outcome of the analysis, insights, and the visual representation of the previously declared methods. The model which is used here for the analysis is linear regression. Gaining significant insights into Taiwan's real estate market and helping potential buyers, investors, and regulators make wise decisions depends on the examination of housing data. This data analysis aims to identify trends, patterns, and key variables that affect Taiwanese housing demand and pricing. The main conclusions and outcomes of this analysis are briefly summarized in the part that follows.

****

**Figure 15: Floor count graph visualization**

(Source: Self-created on Google Colab)

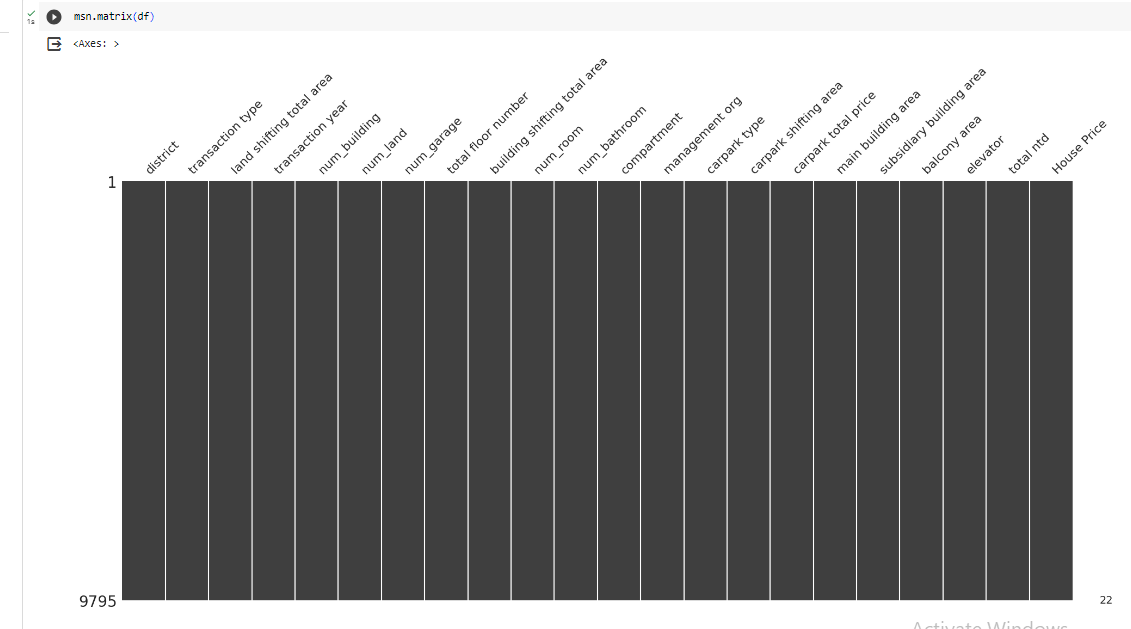
In a time of analysing complicated relationships within data, spotting trends, and generating predictions, graph visualization in data analysis is a potent tool. Here is the reason why it is necessary and how it can help with data prediction (Wang *et al.* 2020). High-dimensional and complicated data sets can frequently make it difficult to identify patterns or relationships. By demonstrating data points as nodes and connections as edges, graph visualization makes it easier to understand the underlying structure of complicated data structures. Here the filter is used for the count plot in the car shifting area. The graph shows the total number of floors and the cunt (Solarte *et al.* 2022). The car park shifting area has a very saturated result as the graph shows. The subplot and the count plot data will help to understand the data transparency in a very vivid manner.

****

**Figure 16: Correlation matrix**

(Source: Self-created on Google Colab)

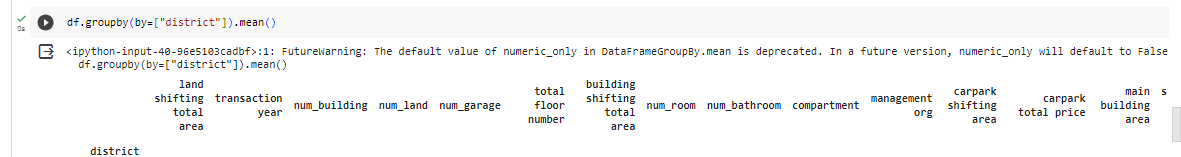
The coefficients of correlation between numerous variables are shown in a table called a correlation matrix. The correlation between the two variables is displayed in each cell of the table. It is a square matrix where the cells show the correlation coefficient between the variables in the respective row and column and the rows and columns represent variables. Correlation coefficients, which normally range from -1 (perfectly negative correlation) to 1 (perfectly positive correlation), with 0 denoting no correlation, assess the degree and direction of the linear link between two variables (Teng *et al.* 2020). This is the visualization of the correlation that is established in the code. This helped to understand the data more vividly and extensively. Analysts can better comprehend the relationships between variables by using a correlation matrix. Beginning of one variable rises, the other tends to rise as well, according to high positive correlations (near to 1), while high negative correlations (close to -1) imply that when one variable rises, the other tends to fall (Sattler *et al.* 2022). No linear relationship is shown by a correlation of 0. Here this matrix shows the strong relationship between the columns of the dataset.

****

**Figure 17: Msn Matrix visualization**

(Source: Self-created on Google Colab)

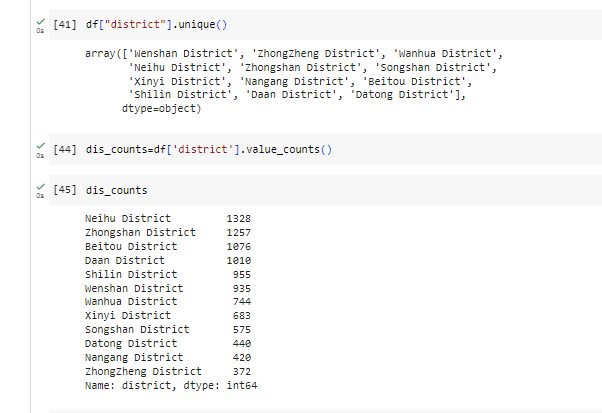
This is the visualization of the dataset using the msn feature.

****

**Figure 18: Separating the district column**

(Source: Self-created on Google Colab)

Here the District data is extracted separately from the data set. Taiwan is a big place and has many states and districts. According to the district, the price of the units per area is different which is why grouping districts according to the same features is important for the analysis.

****

**Figure 19: Districts and the D-type**

(Source: Self-created on Google Colab)

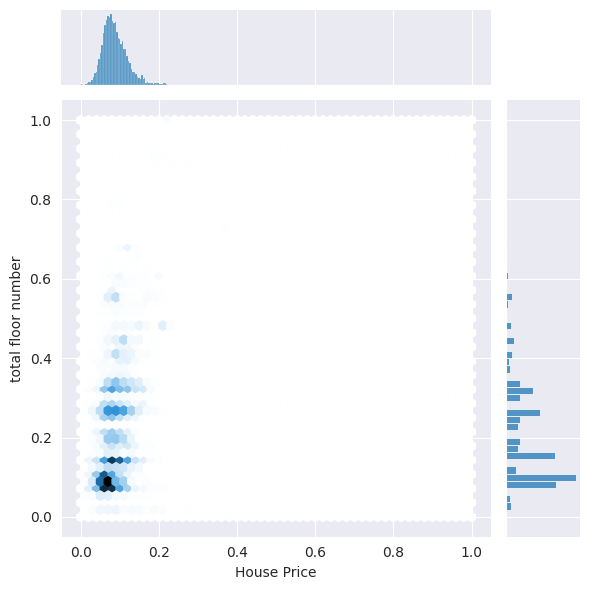
Gathering all districts' data into an array they try to display the values in each district through this approach the data analysis will be more efficient. By this visualization, the analysis and the prediction should be based on the prime aspects of the dataset.

****

**Figure 20: Scatter plot of the selected columns**

(Source: Self-created on Google Colab)

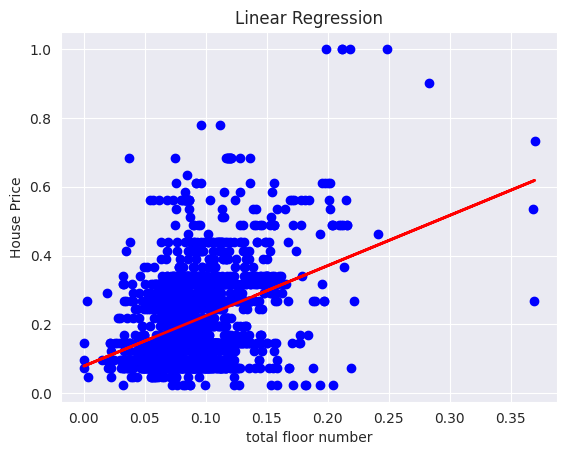
A scatter plot is a two-dimensional graphical depiction of data points, where each point on the plot denotes the values of two different variables. The x- and y-axes are used to plot the two variables, respectively, on the horizontal and vertical planes (Jack *et al.* 2019). A visual depiction of the relationship between the two variables is produced by placing each point on the plot according to its corresponding values on the two axes. To illustrate the correlations between two variables, scatter plots are indispensable (Wang *et al.* 2021). Users may rapidly identify patterns, trends, clusters, and any possible relationships between the variables by showing the data points. It takes this visual clue to comprehend the material fully. Predictive models can be severely impacted by outliers, or data points that differ significantly from the norm. Outliers are easily identified using scatter plots, which aids analysts in determining whether to include or omit them from the modelling process (Negishi *et al.* 2020). Understanding the association between the predictor variables and the target variable—the variable to be predicted—is essential to predictive analysis. When choosing features and creating models, it might be helpful to determine the direction and intensity of this link with the use of scatter plots. This essential tool in prediction analysis is the scatter plot (Nguyen *et al.* 2020). This offers a visual way to comprehend how variables relate to one another, spot trends, evaluate correlations, find outliers, and make well-informed choices regarding variable selection and model building. They are an essential part of the predictive modelling data exploration stage.

****

**Figure 21: joint plot of the House price and the total floor number**

(Source: Self-created on Google Colab)

A joint plot is a visual aid that shows the relationship between two variables in a dataset by combining various plot types. To show the distributions of each variable separately, it usually includes extra marginal histograms or kernel density plots in addition to a scatter plot of the two variables (Bandler *et al.* 2022). Joint plots are frequently made with Python data visualization libraries. The main reason why combined plots are necessary for prediction analysis is the relationship between two variables and the establishment of the correlation (Zhao *et al.* 2020). Those are the reasons the joint plot is taking place on the house price and the total number of the floor.

****

**Figure 22: Result of linear regression on the total floor number and the house price**

(Source: Self-created on Google Colab)

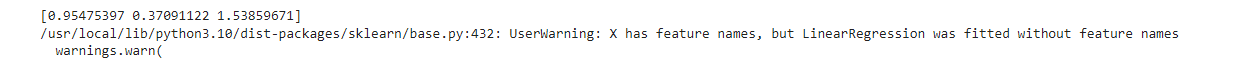
This is the final result of the linear regression as the graph is shown here. This dataset does not have clear data transparency that is the the plot is not as linear as it should be. This precise graph gives a rough idea about the dataset and the regressing outcome.

**https://lh3.googleusercontent.com/OZAw3dWvD66SIwFNdFqMmiLMiMLHy7hbzSbsE-20Co0fORwwqpOxQkv1KizUZDGoH4BbjnDtY8QbZNntTaGCiyY7DSC6TS9XzFNAVdCDgIy1UW--QZy3zKUkDifo2ataMbCHIH9LbSBDkJKisVZ3GUk**

**Figure 23: Mean and R-squared values**

(Source: Self-created on Google Colab)

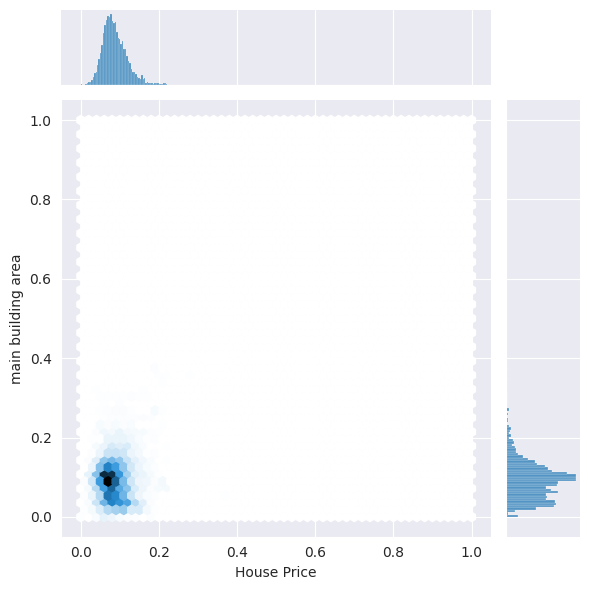
This is the mean squared error value and the r-squared value of this regression model as an outcome. Two crucial metrics in linear regression analysis are mean and R-squared values, and they are crucial in prediction analysis. Mean helps to evaluate the model and compare it and the r-squared helps to identify the model's stability and performance. According to the result, both has very low count here.



**Figure 24: Result of the regression**

(Source: Self-created on Google Colab)

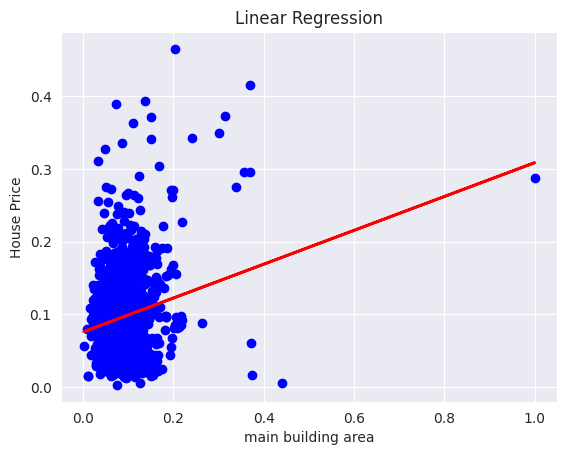
This result of the regression has a very satisfying result as the prediction. 0.954 is well but the graph does not go with the result.

****

**Figure 25: joint plot of the Main building area and the House price**

(Source: Self-created on Google Colab)

This is a second way to implement the linear regression upon separate two columns. This plot is on the main building area and the house price.

****

**Figure 26: Regression of the main building and the House price**

(Source: Self-created on Google Colab)

This is the graph plot of the second linear regression which is based on the house price and main building. The result is not so promising either.

https://lh6.googleusercontent.com/PTVE10zZfQlAsB4GqyDa0rJUZKnWPcpvl1nN68YNguUQZ5LOaVr7nTYthyLskdSoiePyt1tCOFHdbFJE-ZGwXUCK6EtZdxOECfReprF-vhHYsrA4qyYYw-zf2TBzBKeEshzCv91G_YHQF6zMpMptuOg

**Figure 27: results of this regression**

(Source: Self-created on Google Colab)

This indicates the results of the regression and that suggests the data set is not well distributed and structured.

**4.4 Conclusion**

This analytical study is very well established in the fact that the data on Taiwan real estate is not so promising. The model of the linear regression and the other visualization is igniting the same fact. This study has merit and a big scope for further future studies. Less encouraging than expected are the findings of the use of a linear regression model to explore Taiwan's real estate housing price prediction. Despite being a useful and popular method for simulating correlations between variables, linear regression has limits that must be understood in light of the intricacies of the real estate market.

**Reference List**

**Journals**

Kuhlmann, D., 2021. Upzoning and single-family housing prices: A (very) early analysis of the Minneapolis 2040 plan. *Journal of the American planning association*, *87*(3), pp.383-395.

Burdack, J., Horst, F., Giesselbach, S., Hassan, I., Daffner, S. and Schöllhorn, W.I., 2020. Systematic comparison of the influence of different data preprocessing methods on the performance of gait classifications using machine learning. *Frontiers in bioengineering and biotechnology*, *8*, p.260.

Tang, S., Yuan, S. and Zhu, Y., 2020. Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery. *IEEE Access*, *8*, pp.149487-149496.

Nesca, M., Katz, A., Leung, C.K. and Lix, L.M., 2022. A scoping review of preprocessing methods for unstructured text data to assess data quality. *International Journal of Population Data Science*, *7*(1).

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O., 2021. A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), pp.1-37.

Park, D.J., Park, M.W., Lee, H., Kim, Y.J., Kim, Y. and Park, Y.H., 2021. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific reports*, *11*(1), p.7567.

Cui, Z., Ke, R., Pu, Z. and Wang, Y., 2020. Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values. *Transportation Research Part C: Emerging Technologies*, *118*, p.102674.

Nguyen, Q.H., Ly, H.B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I. and Pham, B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*, pp.1-15.

Wang, H., Rong, X., Yang, L., Feng, J., Xiao, J. and Tian, Y., 2020. Weakly supervised semantic segmentation in 3d graph-structured point clouds of wild scenes. *arXiv preprint arXiv:2004.12498*.

Solarte, B., Liu, Y.C., Wu, C.H., Tsai, Y.H. and Sun, M., 2022. 360-dfpe: Leveraging monocular 360-layouts for direct floor plan estimation. *IEEE Robotics and Automation Letters*, *7*(3), pp.6503-6510.

Teng, L., Wu, X., Günther, M. and Ehrhardt, M., 2020. A new methodology to create valid time-dependent correlation matrices via isospectral flows. *ESAIM: Mathematical Modelling and Numerical Analysis*, *54*(2), pp.361-371.

Sattler, P. and Pauly, M., 2022. Testing Hypotheses about Correlation Matrices in General MANOVA Designs. *arXiv preprint arXiv:2209.04380*.

Wang, F., Morita, K., DiNardo, C.D., Furudate, K., Tanaka, T., Yan, Y., Patel, K.P., MacBeth, K.J., Wu, B., Liu, G. and Frattini, M., 2021. Leukemia stemness and co-occurring mutations drive resistance to IDH inhibitors in acute myeloid leukemia. *Nature communications*, *12*(1), p.2607.

Nguyen, H.Q., Ly, H.B., Tran, V.Q., Nguyen, T.A., Le, T.T. and Pham, B.T., 2020. Optimization of artificial intelligence system by evolutionary algorithm for prediction of axial capacity of rectangular concrete filled steel tubes under compression. *Materials*, *13*(5), p.1205.

Jack Jr, C.R., Wiste, H.J., Botha, H., Weigand, S.D., Therneau, T.M., Knopman, D.S., Graff-Radford, J., Jones, D.T., Ferman, T.J., Boeve, B.F. and Kantarci, K., 2019. The bivariate distribution of amyloid-β and tau: relationship with established neurocognitive clinical syndromes. *Brain*, *142*(10), pp.3230-3242.

Bandler, R.C., Vitali, I., Delgado, R.N., Ho, M.C., Dvoretskova, E., Ibarra Molinas, J.S., Frazel, P.W., Mohammadkhani, M., Machold, R., Maedler, S. and Liddelow, S.A., 2022. Single-cell delineation of lineage and genetic identity in the mouse brain. *Nature*, *601*(7893), pp.404-409.

Negishi, T., Abe, S., Matsui, T., Liu, H., Kurosawa, M., Kirimoto, T. and Sun, G., 2020. Contactless vital signs measurement system using RGB-thermal image sensors and its clinical screening test on patients with seasonal influenza. *Sensors*, *20*(8), p.2171.

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S. and Wang, G., 2020. ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *Bmc Bioinformatics*, *21*(1), pp.1-14.

Mehrad, A. and Zangeneh, M.H.T., 2019. Comparison between qualitative and quantitative research approaches: Social sciences. *International Journal For Research In Educational Studies, Iran*, *5*(7), pp.1-7.

Hall, J., Gaved, M. and Sargent, J., 2021. Participatory research approaches in times of Covid-19: a narrative literature review. *International Journal of Qualitative Methods*, *20*, p.16094069211010087.

Canlas, I.P. and Karpudewan, M., 2020. Blending the principles of participatory action research approach and elements of grounded theory in a disaster risk reduction education case study. *International Journal of Qualitative Methods*, *19*, p.1609406920958964.

Gläser, J., Ash, M., Buenstorf, G., Hopf, D., Hubenschmid, L., Janßen, M., Laudel, G., Schimank, U., Stoll, M., Wilholt, T. and Zechlin, L., 2022. The independence of research—a review of disciplinary perspectives and outline of interdisciplinary prospects. *Minerva*, pp.1-34.

Oghazi, P., Parida, V., Wincent, J. and Mostaghel, R., 2022. Ecosystems transformation through disruptive innovation: A definition, framework and outline for future research. *Journal of Business Research*, *147*, pp.16-26.

Buschle, C., Reiter, H. and Bethmann, A., 2022. The qualitative pretest interview for questionnaire development: outline of programme and practice. *Quality & Quantity*, *56*(2), pp.823-842.

Jiang, Y., Geng, M. and Bai, L., 2020. Targeting biofilms therapy: Current research strategies and development hurdles. *Microorganisms*, *8*(8), p.1222.

Beres, B.L., Hatfield, J.L., Kirkegaard, J.A., Eigenbrode, S.D., Pan, W.L., Lollato, R.P., Hunt, J.R., Strydhorst, S., Porker, K., Lyon, D. and Ransom, J., 2020. Toward a better understanding of genotype× environment× management interactions—a global wheat initiative agronomic research strategy. *Frontiers in Plant Science*, *11*, p.828.

Oliva, R., 2019. Intervention as a research strategy. *Journal of Operations Management*, *65*(7), pp.710-724.

Nickel, J.C., 2020. Chronic prostatitis/chronic pelvic pain syndrome: it is time to change our management and research strategy. *BJU international*, *125*(4), pp.479-480.

Lobe, B., Morgan, D. and Hoffman, K.A., 2020. Qualitative data collection in an era of social distancing. *International journal of qualitative methods*, *19*, p.1609406920937875.

Archibald, M.M., Ambagtsheer, R.C., Casey, M.G. and Lawless, M., 2019. Using zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. *International journal of qualitative methods*, *18*, p.1609406919874596.

Zhang, C., Kong, J., Wu, D., Guan, Z., Ding, B. and Chen, F., 2023. Wearable Sensor: An Emerging Data Collection Tool for Plant Phenotyping. *Plant Phenomics*, *5*, p.0051.

Roh, Y., Heo, G. and Whang, S.E., 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), pp.1328-1347

Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), pp.1-16.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, *21*(1), pp.1-16.

Lester, J.N., Cho, Y. and Lochmiller, C.R., 2020. Learning to do qualitative data analysis: A starting point. *Human resource development review*, *19*(1), pp.94-106.

Drolet, M.J., Rose-Derouin, E., Leblanc, J.C., Ruest, M. and Williams-Jones, B., 2023. Ethical Issues in research: perceptions of researchers, research ethics board members and research ethics experts. *Journal of Academic Ethics*, *21*(2), pp.269-292.

Cascio, M.A., Weiss, J.A. and Racine, E., 2021. Making autism research inclusive by attending to intersectionality: A review of the research ethics literature. *Review Journal of Autism and Developmental Disorders*, *8*, pp.22-36.