Project Report

Data Science for Business (TECH)

Predicting House Sale Prices

By:

Rhea Chandok (rc5397)

Arya Goyal (ag9961)

Aditya Suresh (as17339)

Sanam Palsule (sp7940)

Content

Business Understanding

Business Problem:

In the real estate industry, accurately predicting house sale prices is of paramount importance for various stakeholders, including homeowners, real estate agents, investors, and policymakers. However, the complex and dynamic nature of the housing market, coupled with diverse factors influencing property values, presents a significant challenge in achieving precise estimations. The absence of reliable predictive models hampers informed decision-making, leading to suboptimal pricing strategies, financial risks, and market inefficiencies.

This project aims to address the pressing need for robust predictive models that can effectively forecast house sale prices with a high degree of accuracy. By leveraging machine learning algorithms, statistical techniques, and comprehensive datasets encompassing property attributes, economic indicators, and local market trends, the project seeks to develop a predictive framework capable of generating reliable estimates of house sale prices. The proposed solution will empower stakeholders in the real estate industry with actionable insights, facilitating informed decision-making, optimizing sales strategies, mitigating financial risks, and enhancing market efficiency.

Motivation:

The significance of predicting house sale prices lies in its pivotal role in facilitating informed decision-making processes within the real estate market. For prospective buyers, knowing the expected sale price of a property enables them to make budgetary decisions and determine affordability. Similarly, sellers benefit from accurate price predictions by setting competitive prices that attract potential buyers while maximizing the return on their investments. Additionally, investors rely on precise sale price predictions to assess the profitability and feasibility of real estate ventures.

Stakeholders and their benefits:

1. **Prospective Homebuyers:** Prospective homebuyers are interested in knowing the predicted sale prices of properties to make informed decisions

regarding their purchase. Accurate predictions help them determine whether a property fits within their budget constraints and financial goals.

2. **Home Sellers:** Sellers aim to maximize their returns on investment while ensuring a timely sale of their properties. Accurate price predictions assist sellers in setting competitive listing prices, attracting potential buyers, and optimizing their selling strategies.
3. **Real Estate Agents:** Real estate agents act as intermediaries between buyers and sellers, facilitating property transactions. Predictive models for house sale prices enable agents to provide valuable insights to their clients, negotiate effectively, and streamline the selling process.
4. **Real Estate Investors:** Investors in the real estate market rely on price predictions to assess the profitability and risks associated with various investment opportunities. Accurate predictions assist investors in identifying undervalued properties, optimizing investment portfolios, and maximizing returns on investment.
5. **Financial Institutions:** Banks and financial institutions involved in mortgage lending require accurate assessments of property values to determine loan eligibility and terms. Predictive models for house sale prices aid financial institutions in managing lending risks and ensuring responsible lending practices.

Dataset and Data Understanding

Dataset:

The data utilized for predictive analytics in addressing the business problem of predicting house sale prices is sourced from the Ames, Iowa Assessor's Office and obtained from Kaggle. This dataset comprises information on individual residential properties sold in Ames, IA from 2006 to 2010. It encompasses a comprehensive range of variables related to various aspects of the properties, including physical characteristics, amenities, location, and sale conditions.

Data Collection:

The data collection process involved compilation and editing by Dr. Dean DeCock, who meticulously curated the dataset for research and analysis purposes. The original dataset contains 82 columns with 2930 observations (houses) and includes details such as dwelling types, square footage, construction dates, zoning classifications, lot attributes, neighborhood information, and sale prices.

Dataset Features:

The dataset has **2930 rows and 82 columns**.

The dataset comprises 20 continuous variables, detailing various area dimensions for each observation including lot size, dwelling square footage, and specific areas like basements and porches.

It also includes numerous categorical variables, ranging from 2 to 28 classes, identifying dwelling types, garages, materials, and environmental conditions. Nominal variables categorize types, while ordinal variables rate items within the property.

Numerical features offer quantitative attributes like lot size, square footage, bedrooms, bathrooms, year built, and sale price.

Categorical variables capture qualitative attributes such as dwelling type, zoning, neighborhood, foundation type, and exterior material.

Ordinal features represent ordered categorical attributes like overall condition and quality, providing insights into relative property quality.

Temporal data includes sale year, construction year, and renovation dates, adding a temporal dimension to the dataset for trend analysis.

The primary target variable is the sale price, essential for predictive modeling tasks to predict property prices based on provided features.

Possible Biases

While the dataset provides a rich source of information for predictive analytics, it may contain certain biases inherent in real estate data collection and reporting processes. Biases could arise from factors such as:

1. **Temporal Bias:** The dataset spans from 2006 to 2010, during which economic conditions and real estate market dynamics may have varied. Temporal biases could affect the generalizability of predictive models to current market conditions.
2. **Selection Bias:** The dataset may primarily include properties that were sold under specific sale conditions (e.g., 'family' or 'normal'), potentially excluding properties sold under different conditions and introducing selection biases.
3. **Reporting Bias:** Variations in reporting practices among property assessors or real estate agents could introduce inconsistencies or inaccuracies in the data, impacting the reliability of predictive models.

Data Cleaning and Preparation

NAN Values:

In our dataset, we encountered a considerable number of NaN values spanning multiple columns. While the presence of NaN typically signifies missing data, a thorough examination of our dataset revealed that many of these values were not true null entries but rather specific representations of meaningful information. For instance, the 'Alley' attribute contained the value "NA," which, upon closer inspection, signifies "No Alley Access" rather than a missing value. However, when loading the dataset into our dataframe, these representations were erroneously interpreted as NaN.

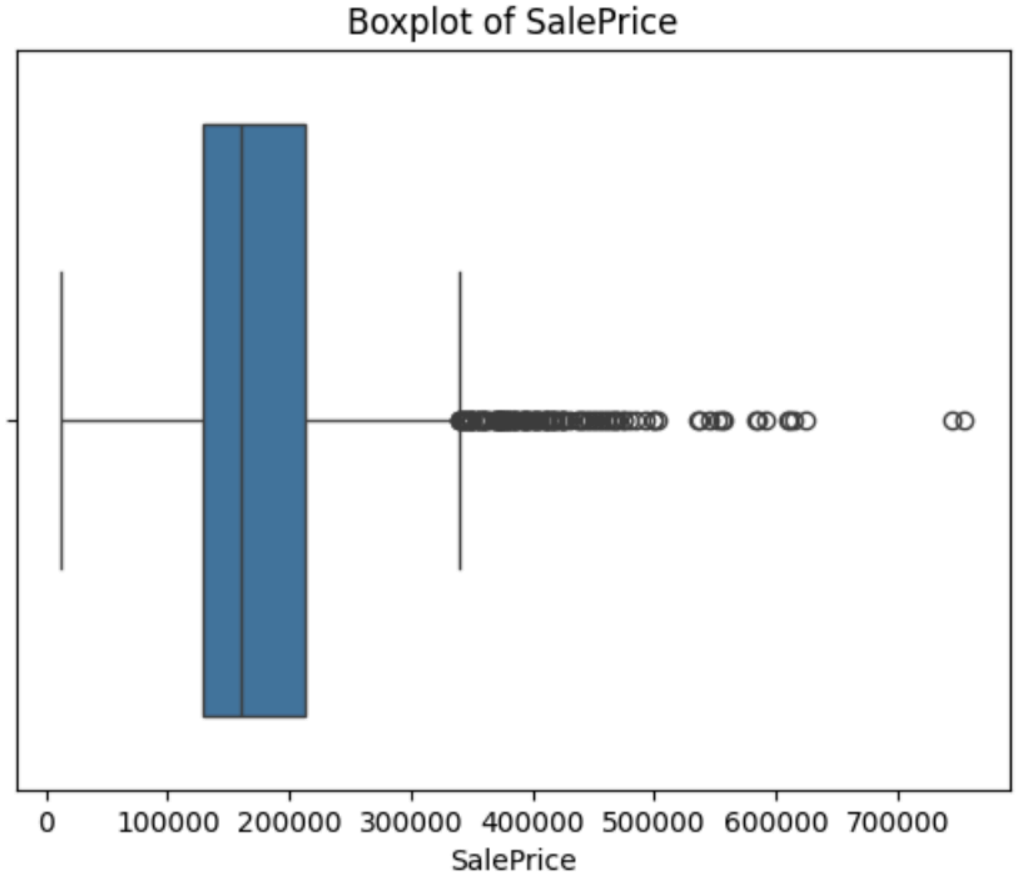
This situation prompted us to undertake a comprehensive understanding of the dataset's features to accurately interpret and handle these values during the cleaning and preparation stages. Simply dropping NaN values in certain columns would risk losing crucial information and adversely impact the accuracy of our predictions, as illustrated by the 'Alley' attribute example. Therefore, it was imperative to develop a nuanced approach to handle NaN values that preserved the integrity of our dataset while ensuring optimal performance in our predictive model. This involved identifying and correctly interpreting representations such as "NA" to accurately capture the underlying meaning of the data and avoid unintended data loss or distortion. Through meticulous data preprocessing techniques, we aimed to prepare a clean and reliable dataset that would serve as a solid foundation for our subsequent modeling efforts, ultimately leading to more robust and accurate predictions of house sale prices.

Categorical Variables:

Furthermore, our dataset comprised numerous columns containing categorical variables. To facilitate the incorporation of these variables into our predictive model, we employed label encoding techniques. This process involved transforming categorical variables into numerical representations, thereby enabling their integration into the model's computation framework. By utilizing label encoding, we effectively translated categorical attributes into a format conducive to statistical analysis and machine learning algorithms, thus enhancing the model's capacity to derive insights and make predictions based on these variables. This approach ensured that our model could effectively leverage the valuable information encapsulated within categorical attributes, contributing to the overall robustness and accuracy of our predictive framework.

Outliers:

Detecting outliers is crucial for assessing the accuracy of our model. A single outlier within a specific neighborhood has the potential to significantly influence the average sale price, thereby introducing errors in predicting the sale prices of similar properties. To identify outliers, we initially utilized a boxplot of the sale price, which provided insight into the range of the majority of prices while highlighting any potential outliers. Our analysis revealed approximately 30 records with unusually high sale prices. While this finding was noted, it did not raise significant concerns regarding the overall integrity of our dataset.



Feature Engineering:

We conducted data manipulation to derive additional meaningful attributes from existing ones in our dataset.

1. First, we calculated the age of each house by subtracting the year built from the current year, resulting in the creation of the "age" column. Similarly, we determined the age since remodeling by subtracting the year remodeled from the current year, generating the "age remodeled" column. Additionally, we created the "YearSinceSold" column by calculating the difference between

the year sold and the current year, as well as the "GarageAge" column by subtracting the garage year built from the current year.

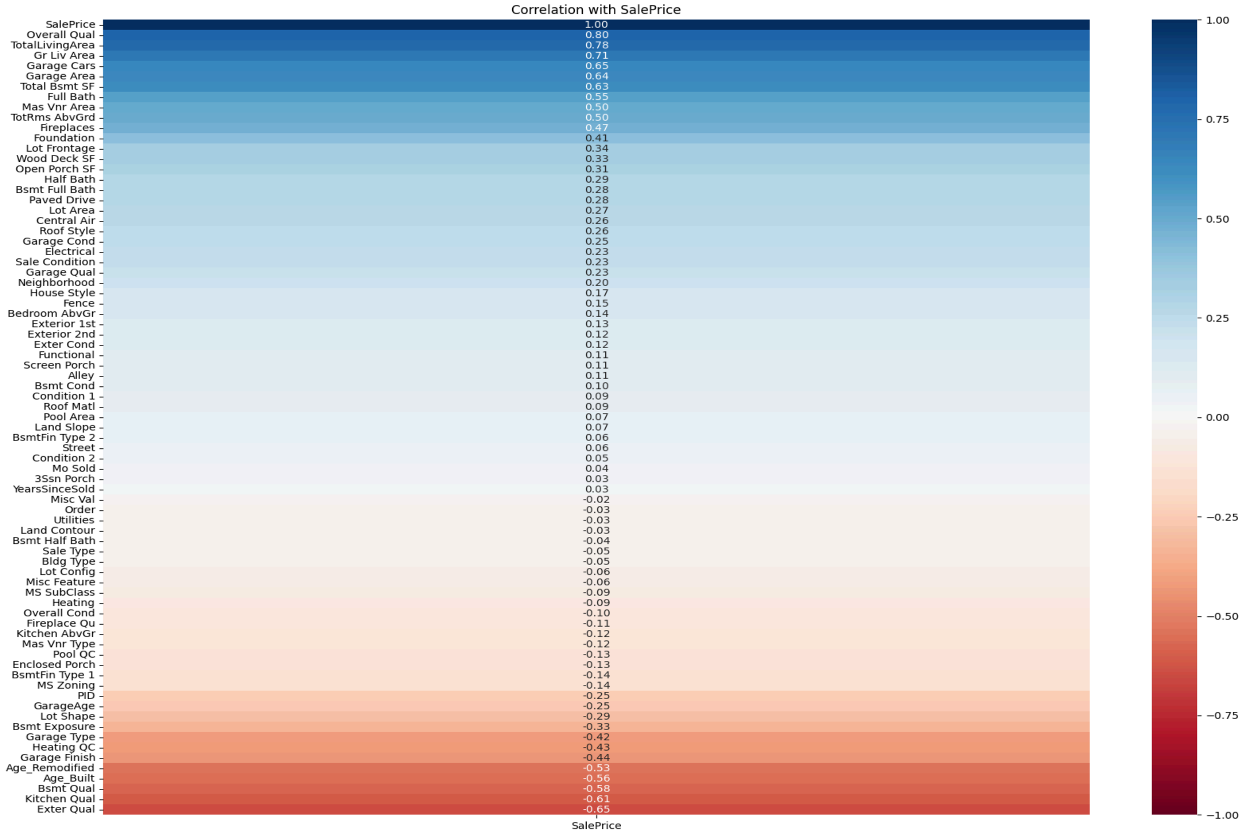
2. To streamline attribute interpretation, we renamed the "Gr Liv Area" column to "TotalLivingArea". Upon closer examination of the dataset, we identified columns that were combinations of other existing attributes. For example, "TotalLivingArea" was derived from the sum of the "1st floor SF", "2nd floor SF", and "Low Quality Fin SF" columns. Similarly, "Total basement SF" was calculated as the sum of the "BsmtFin SF 1" and "Bsmt Fin SF 2" columns. Recognizing redundancy, we decided to drop these redundant columns as they were unnecessary duplicates of attributes that had already been manipulated or derived.

Exploratory Data Analysis

In our exploration of the data, we employed various visualization techniques, including heatmaps, scatterplots, and histograms, to discern relationships between features and the target variable, Sale Price.

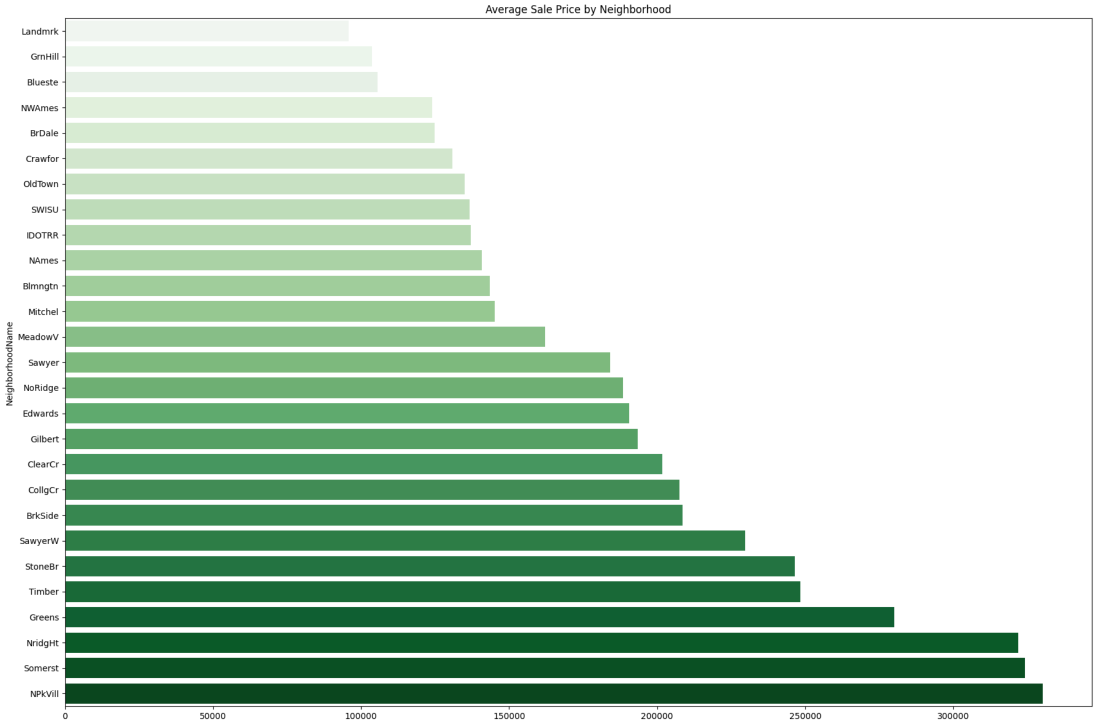
Correlation with Sale Price

First, we constructed a correlation heatmap to elucidate the degree of association between each feature and the Sale Price. This visualization unveiled attributes with strong correlations to the Sale Price, as well as those exhibiting minimal or negligible correlations. An intriguing revelation from this analysis was the negative correlation observed between the attribute "Exterior Quality" and the Sale Price, contrary to conventional expectations. Despite the common belief that exterior appearance significantly influences property value, our data suggested otherwise.

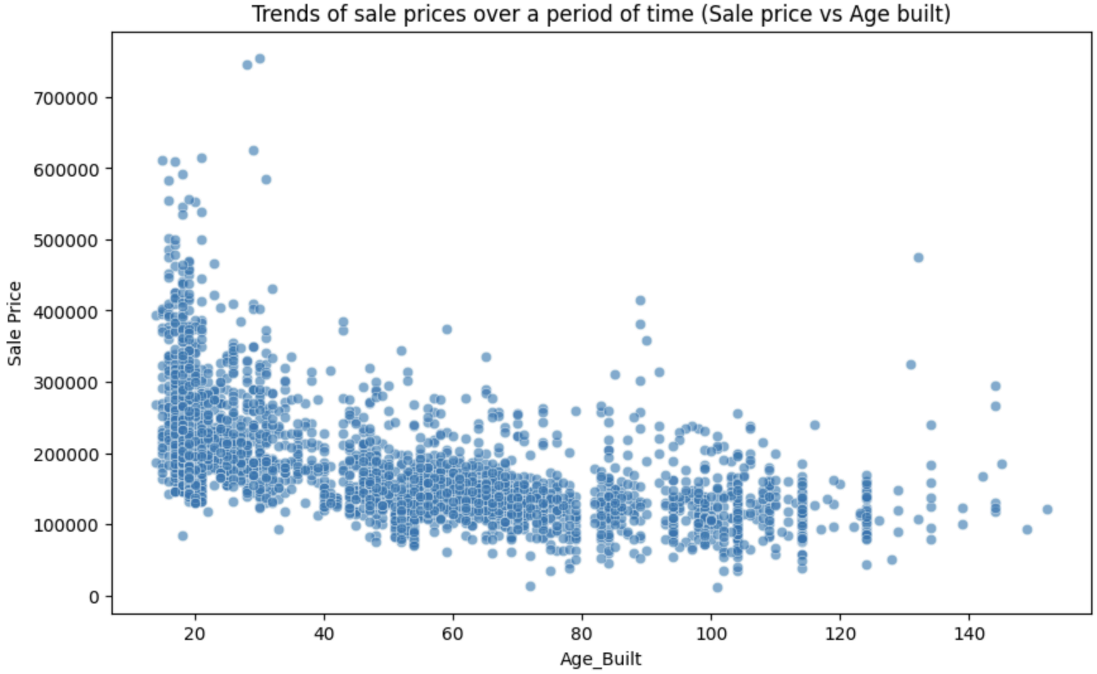


Visualisations and Interesting Insights:

Additionally, we generated a histogram of sale prices segmented by neighborhood to gauge the quality of housing stock within each locality. This visualization offered potential insights for investors seeking to identify neighborhoods with properties of higher value, presenting opportunities for strategic investment.

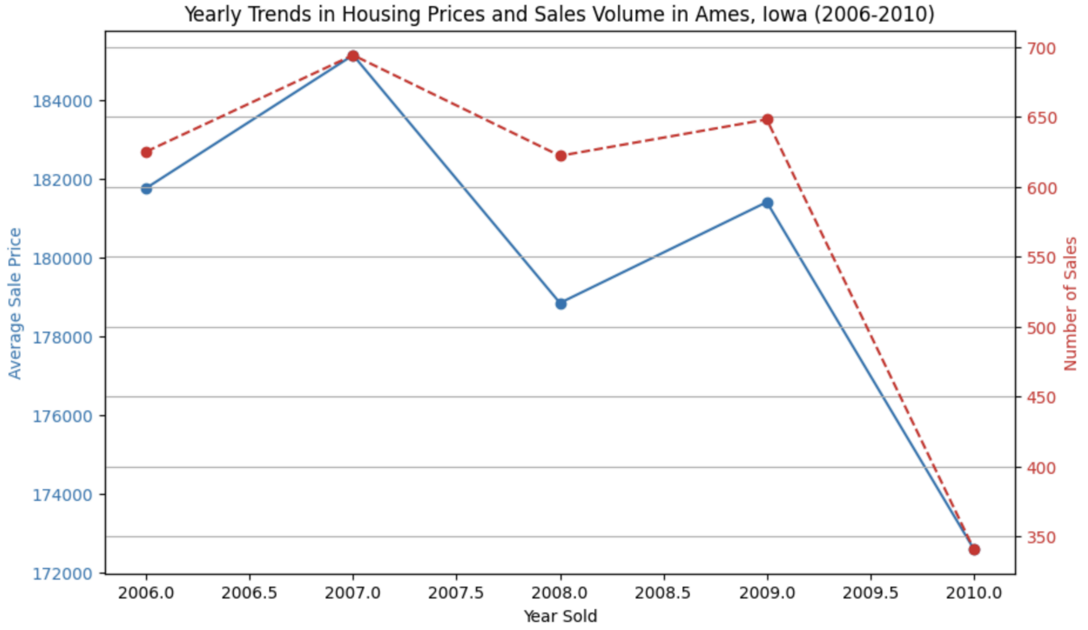


We plotted a scatterplot to examine the trend of sale prices over time, specifically juxtaposing the sale price against the age of the house. Notably, we observed a positive correlation between newer properties (aged 10-35 years) and higher average sale prices, aligning with the expectation that newer homes command premium prices. However, beyond the 40-year threshold, age exhibited minimal impact on sale prices, with most properties within the 40-120-year age range displaying similar average sale prices.



We then constructed a line plot illustrating yearly trends in sale prices from 2006 to 2010. Notably, a significant downturn in sale prices was observed in 2008, coinciding with the global financial crisis. This finding underscored the susceptibility of the real estate market to broader economic fluctuations, providing valuable insights into market dynamics and long-term price trends.

These visualizations served to unravel intricate relationships within the data, offering valuable insights for stakeholders and informing strategic decision-making in the real estate domain.



Analysis of SalePrice in 2008 considering the financial crisis:

We wanted to assess potential bias in the house sales price especially in the context of the 2008 financial crisis. Hence, we employed Kruskal-Wallis-H test to evaluate the null hypothesis that there are no statistically significant differences in the distribution of sale prices across the years 2006 to 2010.

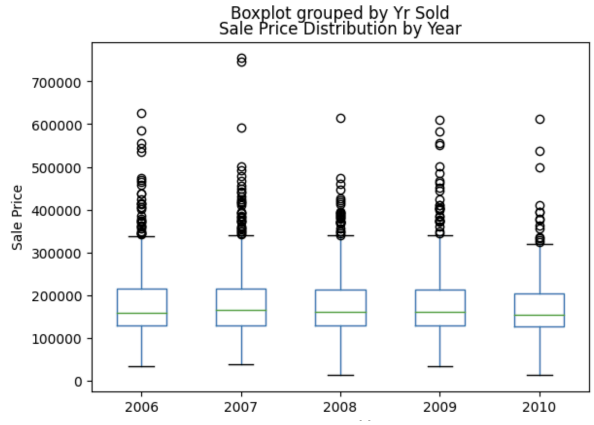
Result :

- Statistic : 4.654
- P-value : 0.325

Since the P-value is greater than the significance level (alpha = 0.05), we failed to reject the null hypothesis.

The analysis indicates no significant differences in the sale price distributions from 2006 to 2010. Despite the economic downturn in 2008, the data does not show a statistically significant impact on the housing prices in the Ames dataset for the years analyzed.

Output of the Analysis:

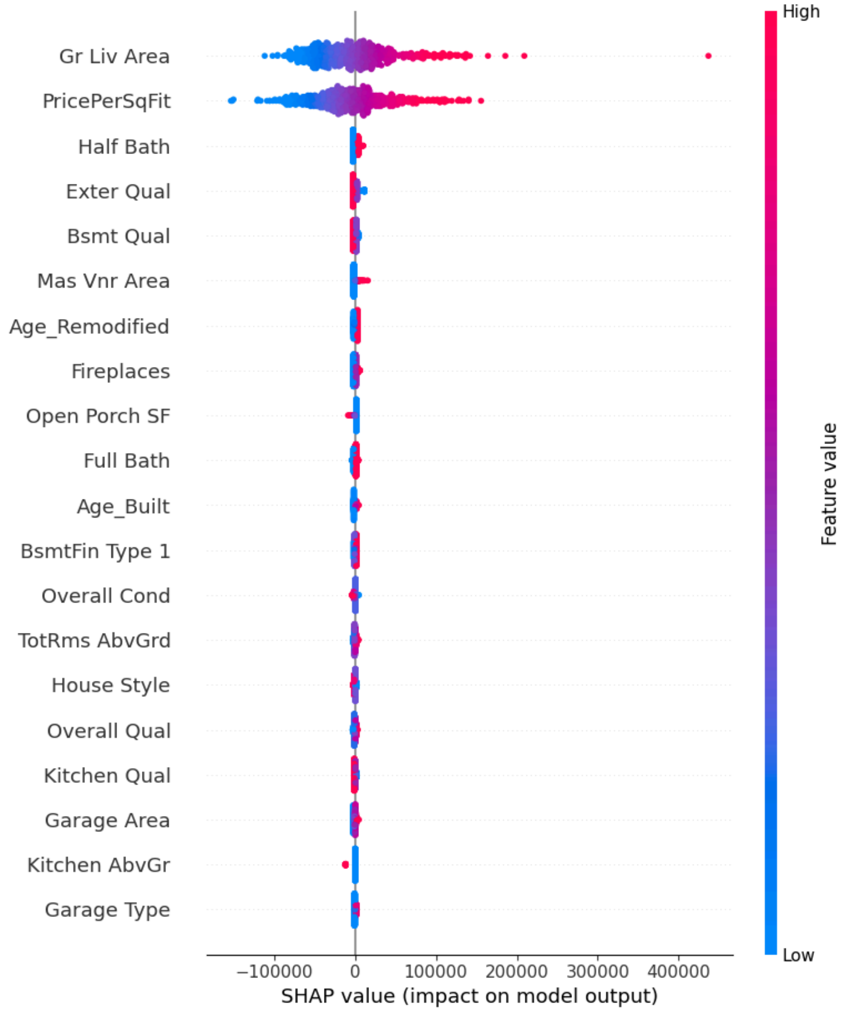


No significant differences between the sale prices across different years.

In our analysis, we employed various techniques to assess the influence of different features on the target variable, Sale Price.

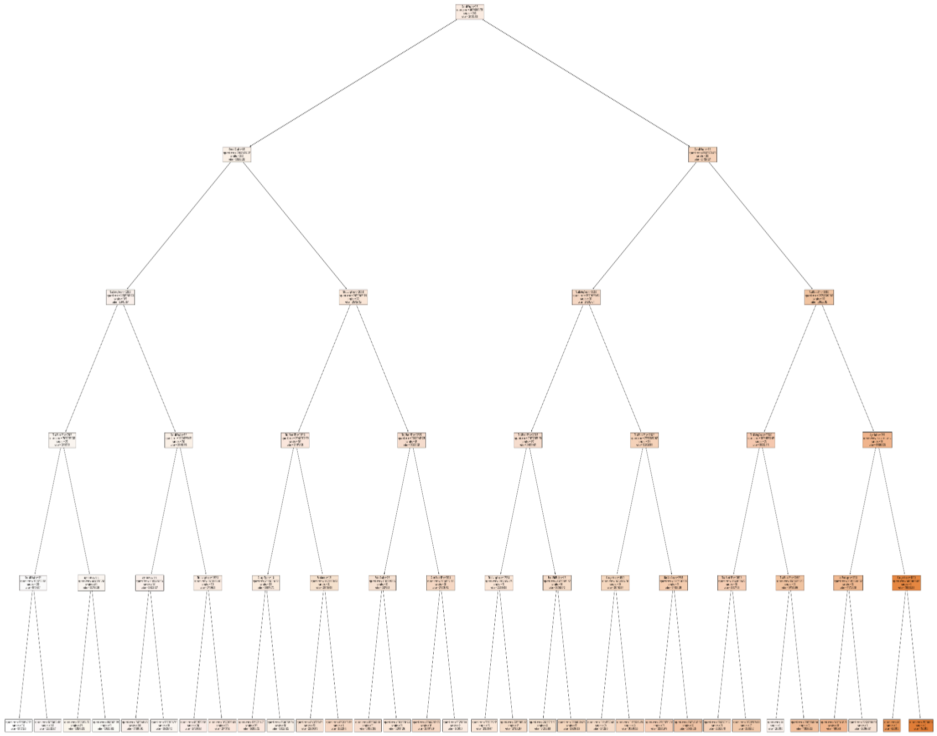
Shapley plots:

Firstly, we utilized Shapley plots to gain insights into which features had the most significant impact on the Sale Price of the house. Our analysis revealed that attributes such as overall quality and TotalLivingArea played a prominent role in influencing Sale Price.



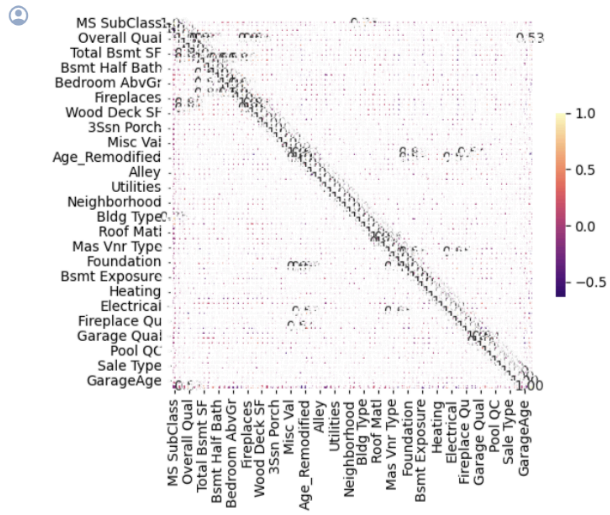
Decision Tree:

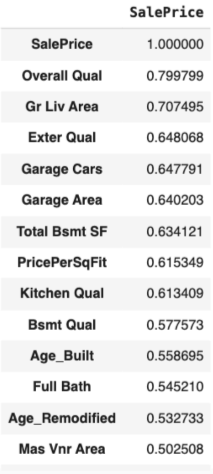
Furthermore, we visualized a decision tree to identify which attributes were utilized as optimal splits to reduce the mean square error. This visualization provided a clear understanding of the importance of individual attributes in predicting Sale Price. In the decision tree analysis, attributes such as Overall Quality, TotalLivingArea, Total Bsmt SF, GarageAge, and Kitchen Quality emerged as significant factors for splitting the data and having the highest reduction in mean square error.



Permutation importance:

Lastly, we employed permutation importance to quantify the effect of each attribute on Sale Price. This technique involved randomly shuffling the values of each feature and observing the resulting impact on the target variable. By aggregating the results, we obtained an importance ranking for each attribute, allowing us to identify features with minimal influence on Sale Price. Attributes such as Pool Area, Mo Sold, Utilities etc. had a permutation score of 0 and this allowed us to confirm their minimal impact on the Sale Price.





	SalePrice
SalePrice	1.000000
Overall Qual	0.799799
Gr Liv Area	0.707495
Exter Qual	0.648068
Garage Cars	0.647791
Garage Area	0.640203
Total Bsmt SF	0.634121
PricePerSqFit	0.615349
Kitchen Qual	0.613409
Bsmt Qual	0.577573
Age_Built	0.558695
Full Bath	0.545210
Age_Remodified	0.532733
Mas Vnr Area	0.502508

Through the combination of these methods, we identified common attributes that exhibited little to no effect on the target variable. Consequently, we determined that the following attributes could be safely dropped from our analysis, streamlining our feature selection process, and enhancing the efficiency of our predictive model:

'BsmtFin Type 2', 'Roof Matl', 'Pool Area', 'Land Slope', 'Street', 'Condition 2', 'Mo Sold', '3Ssn Porch', 'YearsSinceSold', 'Misc Val', 'Utilities', 'Land Contour', 'Sale Type', 'Bldg Type', 'Lot Config', 'Misc Feature', 'MS SubClass', 'Order', 'PID', 'Neighborhood'

Modeling

Baseline Model:

Following the exploratory data analysis (EDA) phase, we identified 57 features deemed suitable for predicting the target variable, Sale Price. To establish a benchmark for subsequent regression analysis, we implemented a Mean Baseline model. In this approach, we predicted the Sale Price of each house to be equal to the mean of the sale prices observed in the dataset.

Through our analysis, we calculated the mean sale price of the entire dataset to be $180,796.04.

This Mean Baseline model served as the initial reference point against which the performance of upcoming models was evaluated and compared. **Implementation of this model resulted in a Root Mean Square Error of $79,202.66 and a Mean Absolute Error of $58,286.04.**

<u>Models Used:</u>

Linear Regression:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to fit a straight line that best represents the data points and predicts the outcome variable.

We first used a linear regression model on the data and tried to predict the sale price of the houses. We also used five-fold cross validation to try and utilize every bit of data available.

Pros: It's computationally efficient and easier to interpret.

Cons: Sensitive to outliers, limited to linear relationships.

Ridge Regression:

Ridge Regression, a regularization technique, is commonly applied to the Ames Housing dataset to mitigate multicollinearity and overfitting. By adding a penalty term to the least squares objective function, Ridge Regression helps stabilize the model's coefficients, enhancing predictive accuracy for housing price estimation.

We used cross-validation on the dataset using a list of folds ranging from 1-10 and alpha ranging from 10^{-15} to 100 as parameters.

Pros: Reduces overfitting due to regularization term, handles multicollinearity

Cons: Limited to linear relationships, computationally complex

Lasso Regression:

Lasso Regression, a variant of linear regression, is often employed in modeling the Ames Housing dataset to mitigate multicollinearity and perform feature selection by penalizing the absolute size of the regression coefficients. By imposing a penalty term based on the L1 norm of the coefficients, Lasso Regression encourages sparsity in the model, effectively selecting a subset of the most influential features while shrinking others to zero.

We used mean squared error as the scoring criterion, and we also performed hyperparameter tuning on the parameter alpha, where we set the range of alpha to be from 0.001 to 100.

Pros: Reduces overfitting due to regularization term, performs automatic feature selection by shrinking the coefficients of less important features to zero.

Cons: Sensitive to small changes in the data, biased estimates for large coefficients

Decision Tree Regressor:

The Decision Tree Regressor is a machine learning algorithm commonly applied to the Ames Housing dataset for predicting house prices based on various features. By recursively partitioning the data into subsets based on feature values, decision trees provide interpretable models capable of capturing non-linear relationships between predictors and the target variable.

We used mean squared error as the scoring criterion and performed hyperparameter tuning on the parameters:

i) max_depth: Determines the maximum depth of the tree. It ranges from 5 to 100.
ii) min_samples_split: Determines the minimum number of samples required to split a particular node. It ranges from 10 to 201.
iii) min_samples_leaf: Determines the minimum number of samples required for a particular node to be a leaf node. It ranges from 10 to 101.

Pros: Handles non-linearity, Handles irrelevant features

Cons: Prone to overfitting, prone to high variance

Random Forest Regressor:

Random Forest Regressor is a powerful machine learning algorithm commonly applied to the Ames Housing dataset for predicting house prices based on a multitude of features. By leveraging an ensemble of decision trees, it captures complex relationships between the input variables and the target variable, offering robust and accurate predictions.

We used gridsearch to perform hyperparameter tuning on the parameters max_depth ranging from 10 to 201 and n_estimators. Mean squared error was used as the scoring criterion.

Pros: Resistant to overfitting, highly accurate predictions

Cons: complexity, computation cost, sensitive to noise.

Evaluation Metrics

1. Root Mean Square Error:

Root mean square error (RMSE) is a commonly used metric for evaluating the accuracy of a predictive model. It measures the average magnitude of the differences between predicted values and actual values in a dataset, considering both the magnitude and direction of errors.

To calculate RMSE, you:

1. Compute the squared differences between predicted and actual values for each observation.

2. Calculate the mean of these squared differences.

3. Take the square root of the mean to obtain the RMSE.

RMSE provides a single numerical value that represents the typical deviation of predicted values from actual values. A lower RMSE indicates better model performance, with values closer to zero indicating higher accuracy. RMSE is widely used in regression analysis and machine learning to compare the performance of different models and assess their predictive capabilities.

2. Mean Absolute Error:

Relying solely on RMSE gives rise to higher errors since RMSE is sensitive to outliers in a dataset. To overcome this, we use Mean absolute error (MAE) which is a metric used to measure the average magnitude of errors between predicted and actual values in a dataset. It is calculated by taking the average of the absolute differences between predicted and actual values for each observation. Unlike root mean square error (RMSE), which squares the errors before averaging, MAE does not consider the direction of errors and provides a straightforward measure of prediction accuracy. A lower MAE indicates better model performance, as it signifies smaller discrepancies between predicted and actual values. MAE is commonly used in regression analysis and machine learning to assess the accuracy of predictive models.

Below is the overview of the results achieved using the different models:

Model	**Root Mean Square Error**	**Mean Absolute Error**	**Comments**
Baseline model	79202.66	58286.04	
Linear Regression Model	36680.14	19415.03	
Ridge Regression	36447.63	19468.99	
Lasso Regression	36680.12	19415.06	These values are noticeably almost identical to the linear regression model. This is because of the fact that lasso regression works by eliminating various columns that have a very low impact on the sale price, but we had already done the process of elimination in our exploratory data analysis phase.
Decision Tree Regressor	35849.98	23200.52	
Random Forest Regressor	28724.68	18283.59	This is a huge improvement from all our models, particularly because random forest is an ensemble model, which means that it fits multiple decision trees to make its predictions.

Outcome:

Through meticulous model training, testing and hyperparameter tuning, we were able to come to a conclusion that the random forest regressor was our best-performing model, giving an error of about 10% of the mean.

Therefore, after performing data cleaning and feature engineering, we were able to achieve a model better than the baseline model.

Costs and Benefits of Model:

Costs:

1. **Financial Loss:** If the model consistently overestimates or underestimates house sale prices, it could lead to financial losses for buyers, sellers, or real estate investors. Overestimating prices may deter potential buyers, while underestimating prices could result in selling properties below their market value.
2. **Market Distortion:** Inaccurate predictions may contribute to market distortion by influencing pricing trends. If the model consistently overvalued properties, it could contribute to inflated market prices, leading to potential instability or housing bubbles.
3. **Customer Dissatisfaction:** Incorrect predictions may lead to dissatisfaction among clients, especially if they rely on the model's estimates for decision-making purposes. This could harm the reputation of real estate agents, brokerage firms, or other stakeholders involved in the transaction.
4. **Data Collection:** Expenses related to acquiring, cleaning, and preprocessing the housing dataset.
5. **Model Development and Maintenance:** Model development and maintenance encompass various costs involved in creating and sustaining the predictive model. This includes expenses related to hiring data scientists or analysts to develop and refine the model, as well as investing in computing resources, software tools, and technology infrastructure necessary for both development and deployment phases.

Benefits:

1. **Improved Decision-Making:** A well-performing predictive model provides stakeholders with valuable insights and information for making informed decisions about pricing, investment, risk management, and strategic planning in the real estate market.
2. **Efficiency and Automation:** Predictive models automate the process of analyzing and interpreting data, enabling faster and more efficient

decision-making compared to traditional methods. This can lead to cost savings and increased productivity for businesses and organizations.

3. **Competitive Advantage:** Organizations that leverage predictive models effectively gain a competitive edge by staying ahead of market trends, understanding customer behavior, and making timely, data-driven decisions.
4. **Customer Satisfaction:** By providing accurate predictions and personalized recommendations, predictive models enhance customer satisfaction, loyalty, and trust, leading to improved customer relationships and retention.
5. **Financial Returns:** Ultimately, the successful implementation of a predictive model can lead to tangible financial returns, including increased revenues, reduced costs, improved profitability, and enhanced shareholder value.

Return on Investment (ROI) Calculation:

Say we've approached a reputable real estate company with our solution, highlighting its capabilities in attracting customers and providing accurate price predictions. After negotiations, the real estate company agrees to purchase our model for **$200,000.**

Costs:

1. **Model Development and Deployment:** We've invested $80,000 in developing and deploying the model. This includes expenses related to hiring skilled data scientists, acquiring necessary software and computing resources, and setting up the infrastructure for model deployment.

Benefits:

1. **Increased Sales Revenue:** By leveraging our predictive model, the real estate company experiences a significant boost in sales revenue. The accurate price predictions attract more potential buyers, leading to increased property transactions and higher sales volumes.

2. **Cost Savings:** Additionally, the model helps the company optimize its pricing strategies and identify lucrative investment opportunities. This leads to cost savings through efficient resource allocation and reduced marketing expenditures.

Interpretation:

ROI = (($200,000 - $80,000) / $80,000) * 100 = 150%

This ROI calculation indicates that for every dollar invested in developing and deploying the predictive model, the real estate company gains $1.50 in return. The 150% ROI demonstrates the substantial value our model brings to the company, making it a highly lucrative investment.

By presenting such detailed insights into the financial implications of implementing our predictive model, we empower stakeholders to make informed decisions and confidently proceed with the investment.

To sustain revenue generation beyond the initial sale of our predictive model, we can:

1. Offer subscription or licensing options.

2. Provide customization and consulting services.

3. Charge for maintenance and support.

4. Develop additional features or modules.

5. Invest in continuous innovation.

These strategies ensure ongoing value delivery to clients while expanding our market reach and maintaining competitiveness.

Deployment and Future Work

After data mining and developing a predictive model for house sale prices, the results can be deployed in various ways to support decision-making and enhance operations in the real estate industry. Here's how the results might be deployed:

i. **Integration into Real Estate Platforms:** The predictive model can be integrated into online real estate platforms, property listing websites, and mobile applications. Users can access the model to obtain accurate price estimates for properties, explore market trends, and make informed decisions about buying or selling homes.

ii. **Incorporation into Real Estate Agent Tools:** Real estate agents and brokers can incorporate the predictive model into their tools and software to assist them in pricing properties, conducting market analysis, and advising clients. The model's predictions can support agents in negotiating deals and maximizing returns for their clients.

iii. **Integration with Mortgage Lending Systems:** Financial institutions and mortgage lenders can integrate the predictive model into their loan origination systems and online banking platforms. The model's predictions can help lenders assess property valuations, evaluate mortgage applications, and manage risk in their lending portfolios.

iv. **Utilization in Urban Planning and Policy Making:** Government agencies and urban planners can utilize the predictive model to analyze housing market trends, monitor affordability issues, and inform policy-making decisions related to housing and urban development. The model's insights can support evidence-based planning and promote sustainable growth.

v. **Research and Academic Purposes:** Research institutions and academic organizations can use the predictive model for research purposes, conducting studies on housing market dynamics, demographic trends, and socio-economic impacts. The model's predictions can contribute to scholarly research and advance knowledge in the field of real estate economics.

After deployment, the predictive model should be monitored to ensure its continued accuracy and effectiveness. This can be achieved through:

i. **Regular Performance Evaluation:** The model's performance should be periodically evaluated using real-world data to assess its accuracy and reliability over time. This involves comparing predicted sale prices with actual sale prices and analyzing any discrepancies or errors.

ii. **Feedback Mechanisms:** Feedback from users, stakeholders, and domain experts should be collected to identify any issues, limitations, or areas for improvement with the model. This feedback can inform adjustments, updates, or refinements to the model to enhance its performance and usability.

iii. **Monitoring for Drift:** The model should be monitored for concept drift, which occurs when the underlying relationships between variables change over time. Regular monitoring and retraining of the model may be necessary to address concept drift and maintain its predictive accuracy.

In subsequent analyses, additional datasets that could be collected to further enhance understanding of house sale prices and related factors may include:

i. **Demographic Data:** Collect demographic information about the areas surrounding the properties, such as average income, population growth, age distribution, and migration patterns. This data can help understand demand dynamics in the housing market.

ii. **Market Conditions and Economic Indicators**: Data on local market conditions, economic indicators, interest rates, and demographic trends can help contextualize sale prices and identify broader market trends and influences.

iii. **Historical Sales Data:** Historical sales data spanning multiple years can be used to analyze trends over time, assess seasonality effects, and identify patterns in sale prices and market dynamics.

iv. **Customer Preferences and Behavior:** Surveys or customer feedback data can provide insights into buyer preferences, motivations, and decision-making processes, which can influence sale prices and market demand.

v. **Technological and Environmental Factors:** Data on the adoption of smart home technologies, energy efficiency ratings, and environmental risk factors (like flood risk) could be increasingly relevant in predicting house prices.

vi. **Feature Re-evaluation:** Regularly review and revise the set of features used in the model. As new data becomes available and as market dynamics change, the relevance of features may evolve.

By collecting and analyzing additional datasets, stakeholders can gain a more comprehensive understanding of house sale prices and develop more accurate and robust predictive models to support decision-making in the real estate industry.

Associated Risks

There are several important ethical considerations, privacy concerns, and risks associated with conducting an analysis of house sale prices and deploying predictive models in the real estate industry. Here are some key considerations and potential mitigation strategies:

1. Ethical Considerations:

i. Fairness and Bias: Predictive models may inadvertently perpetuate biases present in historical data, leading to unfair outcomes for certain demographic groups. It's essential to ensure fairness and equity in model development and deployment by carefully considering the selection of features, monitoring for bias, and implementing fairness-aware algorithms.

ii. Transparency: Stakeholders should have transparency into how predictive models are developed, the factors influencing predictions, and the potential implications of model outputs. Providing clear explanations and documentation can enhance trust and accountability.

iii. Accountability: Establishing accountability mechanisms and governance structures to oversee model development, deployment, and use can help mitigate ethical risks and ensure compliance with legal and regulatory requirements.

2. Privacy Concerns:

i. Data Protection: Real estate data often contains sensitive information about individuals, such as property addresses, transaction details, and personal identifiers. It's crucial to implement robust data protection measures, such as encryption, access controls, and anonymization techniques, to safeguard privacy and comply with data protection regulations.

ii. Informed Consent: When collecting and processing personal data, obtaining informed consent from individuals is essential to respect their privacy rights and ensure compliance with privacy laws and regulations. Clear communication about data usage and obtaining consent can help build trust with stakeholders.

3. Other Risks and Mitigation Strategies:

i. Model Accuracy: Predictive models may not always produce accurate or reliable predictions, leading to potential financial losses or missed opportunities. Conducting thorough testing, validation, and model evaluation procedures can help identify and mitigate risks associated with model accuracy.

ii. Data Quality and Bias: Poor data quality, incomplete datasets, or biased data can compromise the accuracy and fairness of predictive models. Employing data cleaning, preprocessing, and bias detection techniques can help mitigate these risks and improve the quality of model inputs.

iii. Model Interpretability: Black-box models may lack interpretability, making it challenging to understand how predictions are generated and assess their reliability. Employing interpretable models or techniques, such as model explanation methods or transparency frameworks, can enhance interpretability and facilitate stakeholder understanding.

iv. Regulatory Compliance: Compliance with legal and regulatory requirements, such as fair housing laws, anti-discrimination statutes, and data protection regulations, is essential when deploying predictive models in the real estate industry. Conducting legal reviews, risk assessments, and compliance checks can help ensure adherence to regulatory requirements and mitigate legal risks.

Overall, addressing ethical considerations, privacy concerns, and other risks associated with predictive modeling in the real estate industry requires a comprehensive approach that prioritizes fairness, transparency, privacy protection, and regulatory compliance. Collaboration among stakeholders, including data scientists, domain experts, legal professionals, and policymakers, is essential to effectively mitigate risks and promote the responsible use of predictive analytics in real estate.

Appendix

Contributions by each member :

Aditya Suresh: Data cleaning, Feature Correlation, and Model Evaluation

Rhea Chandok: Defining the business problem and its significance, Feature Engineering, and Model Evaluation

Sanam Palsule: Bias analysis, Data Visualization, and Modelling

Arya Goyal: Data Visualization, Feature importance, and Modelling

All members contributed to the presentation and report.

Link to the dataset:

https://www.kaggle.com/datasets/marcopale/housing

Link to the Google Colab file (code) :

https://colab.research.google.com/drive/16HFlFpaw4lMAC--aV51RG_U0P3zKk01i?usp=chrome_ntp