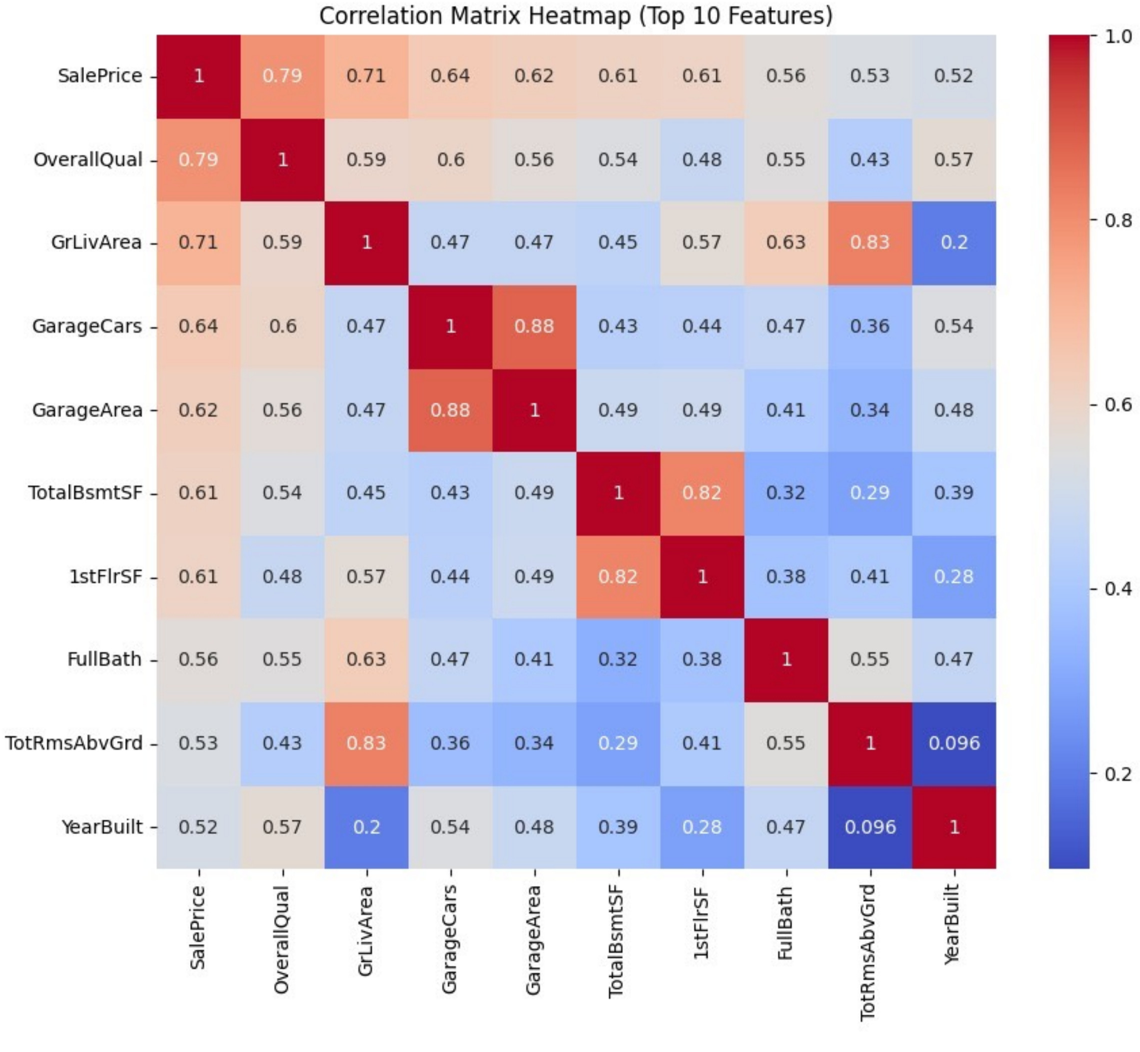
Data Science For Business - Status Update

Team G1 - Anna Dominic, Siri Desiraju, Manisha Goyal, Karthikeyan Shanmugam

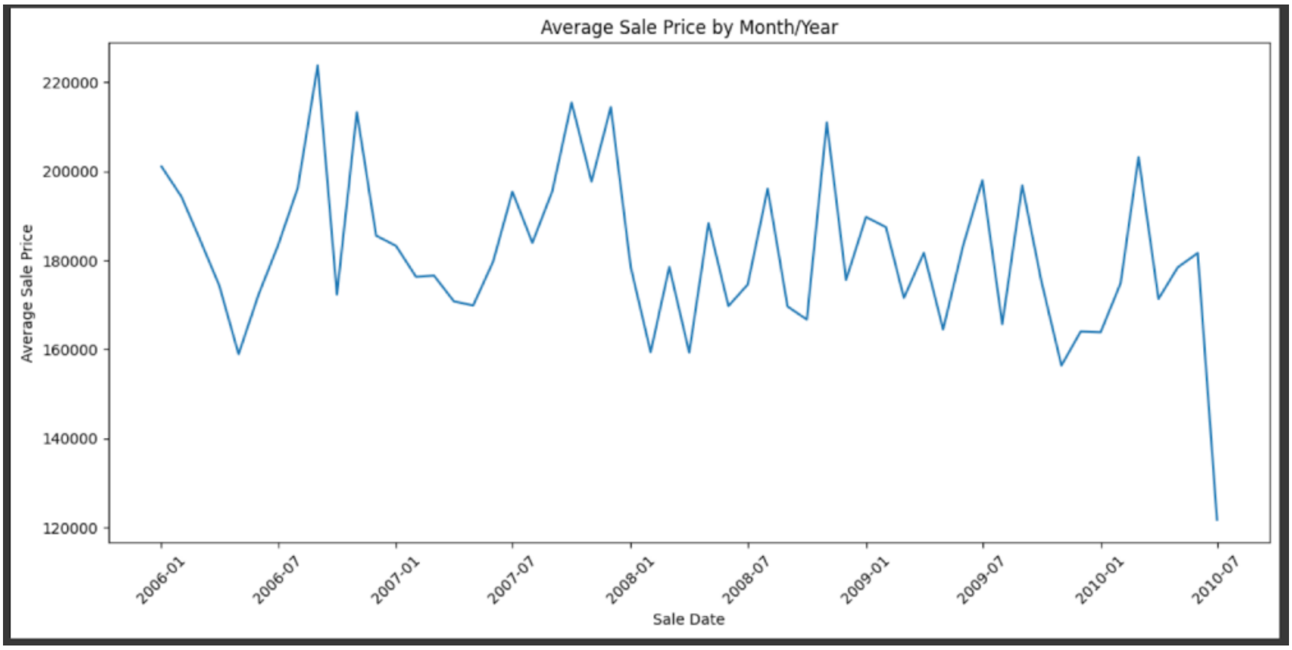
House Value Prediction for Strategic Real Estate Decision-Making

Our project has advanced through exploratory data analysis, where we utilized various visualization techniques to explore the distribution and relationships between different features in the dataset. Through a combination of histograms for frequency distribution, boxplots for comparing distributions across different categories, and scatter plots for examining correlations, we gained insights into the underlying patterns and trends within the data.

Through correlation analysis, we discovered that the most strongly correlated variables with SalePrice encompassed a diverse range of property characteristics. Notably, features such as Overall Quality, Ground Floor Living Area, and Garage Area intuitively align with common expectations of what drives property values. However, the inclusion of less obvious factors like the number of cars the garage fits, total basement space, and first-floor space shed light on additional dimensions influencing sale prices. Furthermore, variables such as the presence of full baths, total rooms above ground, and the year built underscored the importance of both functional amenities and historical aspects in determining property values. This comprehensive set of correlated features offers valuable insights into the multifaceted nature of the real estate market and highlights the complexity of factors influencing property pricing dynamics.



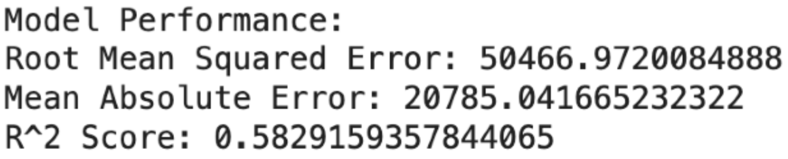
Another interesting finding from our analysis was through a time-series analysis of how house prices changed over time. The plot for this is given as follows:



The plot reflects the interplay between seasonality and economic factors in the housing market from 2006 to 2010. Peaks in mid-2006 and mid-2008 suggest strong spring/summer sales, possibly enhanced by economic stimuli, while dips in early 2007 and 2008 align with seasonal slowdowns and the onset of the global financial crisis, respectively, indicating a contraction in buyer activity and a decrease in average sale prices.

Subsequent hypothesis testing (ANOVA tests, OLS regression, Tukey's HSD), gave insight into influential features. For example, these analyses revealed significant market valuation differences based on the property's type and style. Specific types like single-family homes and architectural styles such as one-story and two-story houses are more sought after, reflecting their higher prices in the market. This insight can be valuable for sellers, buyers, and real estate professionals when making decisions about investments or renovations.

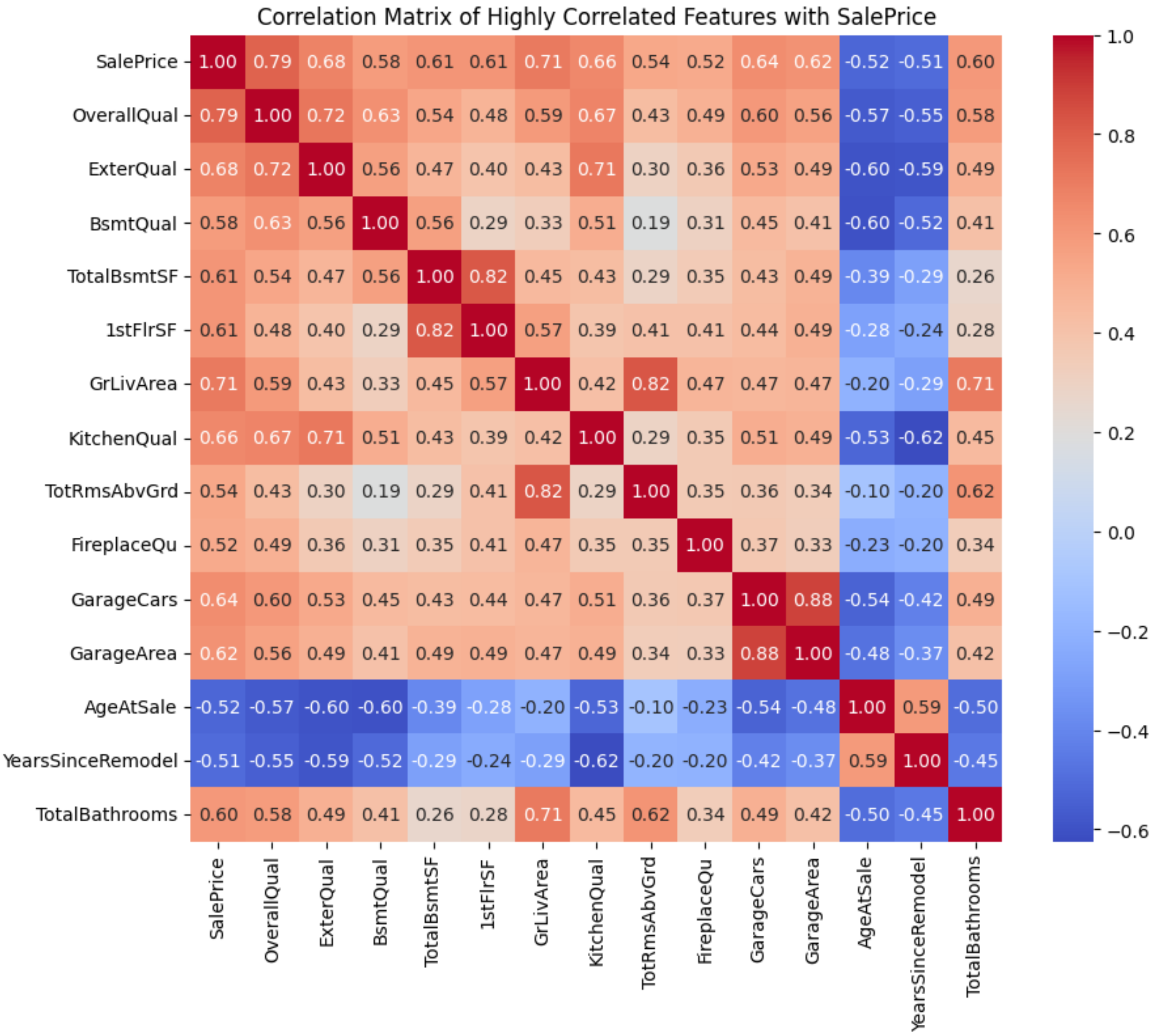
The data cleaning phase included addressing missing values, encoding categorical variables, and feature engineering. After data cleaning, we had a total of 205 features, and a baseline linear regression model was developed using all of them.



```
Model Performance:
Root Mean Squared Error: 50466.9720084888
Mean Absolute Error: 20785.041665232322
R^2 Score: 0.5829159357844065
```

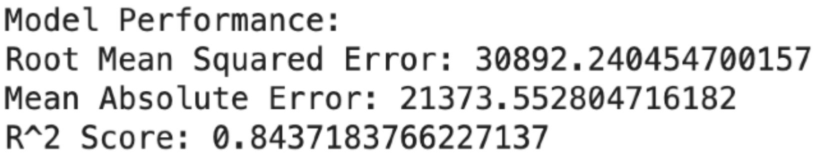
The R² score of approximately 0.5829 from the linear regression model indicates that about 58% of the variability in SalePrice can be explained by the model's features, signifying that simply using all the available features is not very useful for predicting house prices.

To address the performance of the model, we calculated the Pearson correlation coefficient between the features and filtered for the features that had an absolute correlation coefficient of more than 0.5 with SalePrice. We generated an updated matrix with cleaned data and the filtered features. This matrix was pivotal in isolating top features that correlate strongly with SalePrice.


Correlation Matrix of Highly Correlated Features with SalePrice

	SalePrice	OverallQual	ExterQual	BsmtQual	TotalBsmtSF	1stFlrSF	GrLivArea	KitchenQual	TotRmsAbvGrd	FireplaceQu	GarageCars	GarageArea	AgeAtSale	YearsSinceRemodel	TotalBathrooms
SalePrice	1.00	0.79	0.68	0.58	0.61	0.61	0.71	0.66	0.54	0.52	0.64	0.62	-0.52	-0.51	0.60
OverallQual	0.79	1.00	0.72	0.63	0.54	0.48	0.59	0.67	0.43	0.49	0.60	0.56	-0.57	-0.55	0.58
ExterQual	0.68	0.72	1.00	0.56	0.47	0.40	0.43	0.71	0.30	0.36	0.53	0.49	-0.60	-0.59	0.49
BsmtQual	0.58	0.63	0.56	1.00	0.56	0.29	0.33	0.51	0.19	0.31	0.45	0.41	-0.60	-0.52	0.41
TotalBsmtSF	0.61	0.54	0.47	0.56	1.00	0.82	0.45	0.43	0.29	0.35	0.43	0.49	-0.39	-0.29	0.26
1stFlrSF	0.61	0.48	0.40	0.29	0.82	1.00	0.57	0.39	0.41	0.41	0.44	0.49	-0.28	-0.24	0.28
GrLivArea	0.71	0.59	0.43	0.33	0.45	0.57	1.00	0.42	0.82	0.47	0.47	0.47	-0.20	-0.29	0.71
KitchenQual	0.66	0.67	0.71	0.51	0.43	0.39	0.42	1.00	0.29	0.35	0.51	0.49	-0.53	-0.62	0.45
TotRmsAbvGrd	0.54	0.43	0.30	0.19	0.29	0.41	0.82	0.29	1.00	0.35	0.36	0.34	-0.10	-0.20	0.62
FireplaceQu	0.52	0.49	0.36	0.31	0.35	0.41	0.47	0.35	0.35	1.00	0.37	0.33	-0.23	-0.20	0.34
GarageCars	0.64	0.60	0.53	0.45	0.43	0.44	0.47	0.51	0.36	0.37	1.00	0.88	-0.54	-0.42	0.49
GarageArea	0.62	0.56	0.49	0.41	0.49	0.49	0.47	0.49	0.34	0.33	0.88	1.00	-0.48	-0.37	0.42
AgeAtSale	-0.52	-0.57	-0.60	-0.60	-0.39	-0.28	-0.20	-0.53	-0.10	-0.23	-0.54	-0.48	1.00	0.59	-0.50
YearsSinceRemodel	-0.51	-0.55	-0.59	-0.52	-0.29	-0.24	-0.29	-0.62	-0.20	-0.20	-0.42	-0.37	0.59	1.00	-0.45
TotalBathrooms	0.60	0.58	0.49	0.41	0.26	0.28	0.71	0.45	0.62	0.34	0.49	0.42	-0.50	-0.45	1.00

A second linear regression model, utilizing these rigorously selected features, was developed and it serves as an initial benchmark for predictive evaluation.



```
Model Performance:
Root Mean Squared Error: 30892.240454700157
Mean Absolute Error: 21373.552804716182
R^2 Score: 0.8437183766227137
```

The major insight from the model developed is the strong predictive relationship between house prices and the selected features, particularly the overall quality and the living area of the houses. The R^2 score of approximately 0.8437 from the linear regression model indicates that about 84% of the variability in SalePrice can be explained by the model's features, signifying that the features chosen through the correlation matrix are highly informative for predicting house prices. This insight underscores the importance of feature selection in building effective predictive models and suggests that focusing on quality and space-related attributes could be key in accurately estimating house prices.

We will now move on to iterative enhancement, aiming to refine the model's efficacy through additional feature engineering and more sophisticated modeling techniques. Our analysis plan includes:

Expand Feature Engineering

- Investigate interaction terms between top correlated features to capture compound effects.
- Consider temporal features such as the month or season of sale.
- Explore dimensionality reduction techniques like Principal Component Analysis for enhanced feature representation.

Sophisticated Modeling Techniques

- Evaluate tree-based models like Random Forest and Gradient Boosting for capturing non-linear patterns.
- Experiment with Support Vector Machines for potential improvements in prediction accuracy.
- Use cross-validation to assess model stability, generalizability and hyperparameter tuning.
- Use grid search for optimal hyperparameter selection.
- Explore ensemble methods to combine predictions from multiple models

Further Analysis

- If resource and time constraints permit, we will also try to explore comparisons with housing markets in other cities and states, leveraging average income data to assess similarities and differences in housing trends.
- By combining these datasets and using features common across them, and leveraging the insights from our primary regression model to predict house prices in Ames, Iowa, we envision building a secondary, generalizable model that scales across different cities.