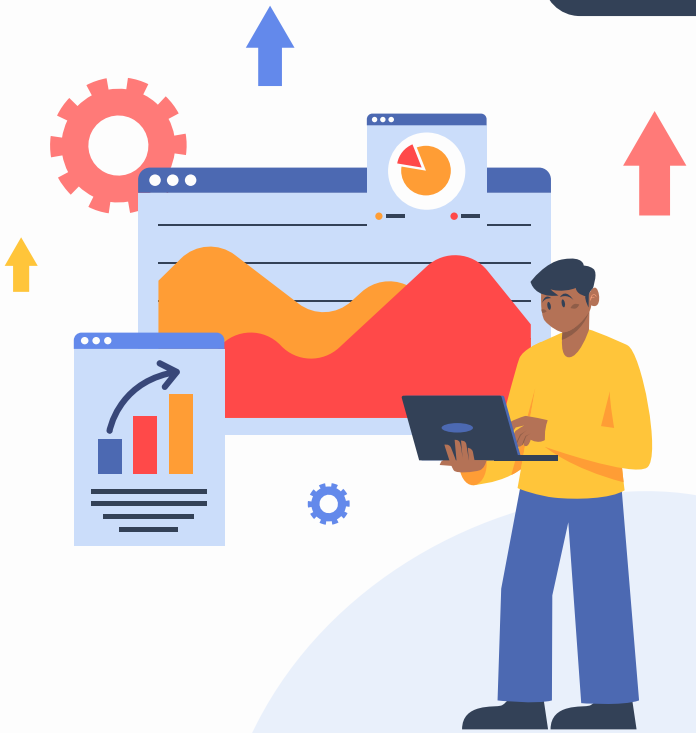
Exploratory Data Analysis (EDA) for House Prices

Business Intelligence



Understanding Housing Market Trends (2006–2010)

Author: Hijir Della Wirasti
Batch: Dibimbing Batch 13
Field: Business Intelligence

https://github.com/hijirdella/House-Price-Analysis-EDA-and-Correlation-Insights
linkedin.com/in/hijirdella/

Table of contents

1. Key Objectives:

A. Perform data profiling and visualize key trends.
B. Preprocess the dataset to ensure data quality.
C. Conduct exploratory data analysis (EDA) to uncover insights.
D. Use advanced statistical methods for deeper analysis:
    - Correlation Analysis
    - Chi-Square Test
    - Linear Regression

2. Dataset Overview

Source: House Prices - Advanced Regression Techniques

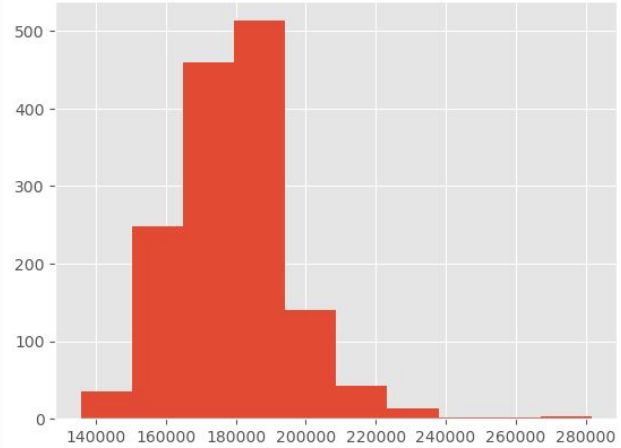
Key Information:

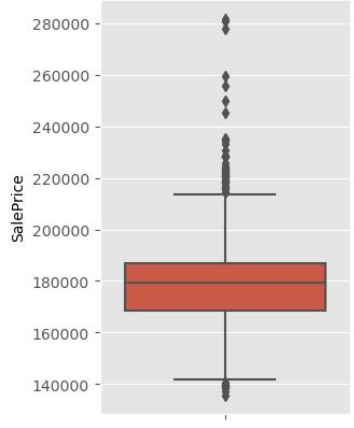
- Rows: 1460
- Columns: 81 (Numerical: 38, Categorical: 43)
- Target Variable: SalePrice

3. Data Profiling Distribusi Harga Rumah

Reviewed central tendency measures (mean, median, mode).

YrSold	count	mean	std	min	25%	50%	75%	max
2006	314.0	182549.458599	79426.838855	35311.0	131375.0	163995.0	218782.5	625000.0
2007	329.0	186063.151976	85768.171410	39300.0	129900.0	167000.0	219500.0	755000.0
2008	304.0	177360.838816	69735.610685	40000.0	131250.0	164000.0	207000.0	446261.0
2009	338.0	179432.103550	80879.237311	34900.0	125250.0	162000.0	212750.0	582933.0
2010	175.0	177393.674286	80451.280085	55000.0	128100.0	155000.0	213250.0	611657.0

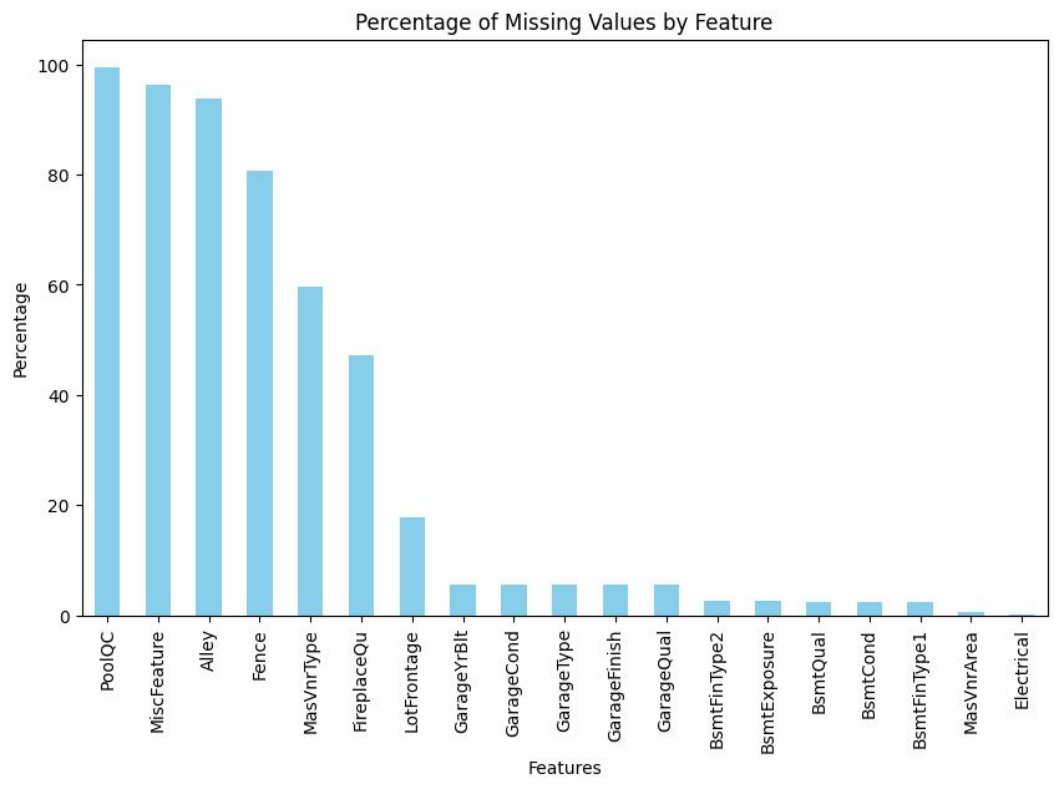




Insight:

- Stabilitas Harga: Rata-rata harga rumah stabil dari 2006 hingga 2010, dengan sedikit kenaikan di 2007.
- Distribusi Skewed: Mayoritas rumah dijual di kisaran $160,000–$200,000, sementara sedikit outliers mewakili properti premium.
- Penyebaran Tinggi: Variasi harga menunjukkan segmen pasar yang beragam dari menengah hingga kelas atas.

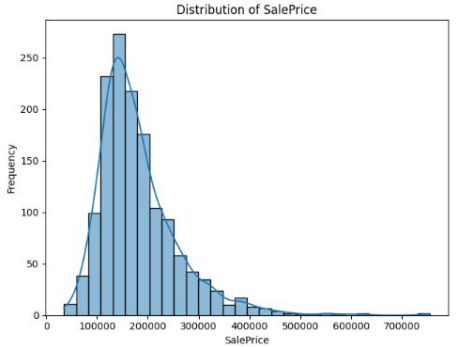
4. Data Preprocessing



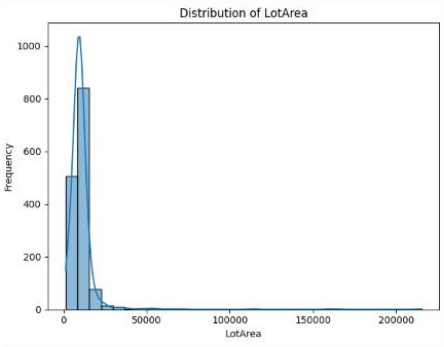
Steps Taken:

1. **Handling Missing Values:**
   - Columns with <5% missing: Imputed with median or mode.
   - Columns with 5–20% missing: Imputed with median or mode.
   - Columns with >20% missing: Dropped (e.g., `PoolQC`, `Alley`).
2. **Removing Duplicates:**
   - Duplicates identified and removed.
3. **Outliers Detection:**
   - Significant outliers found in `LotArea`, `GrLivArea`, and `SalePrice`.

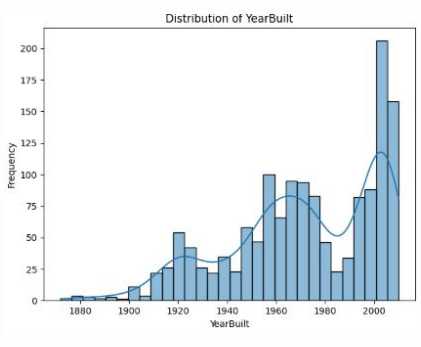
5. EDA – Univariate Analysis (1)



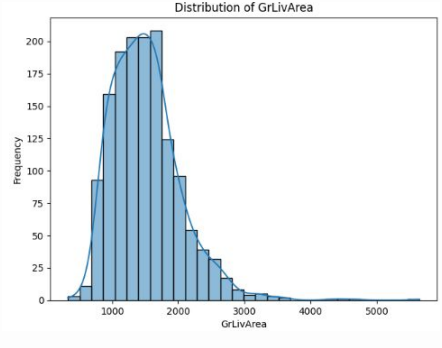
Distribusi `SalePrice` terlihat **right-skewed**, artinya sebagian besar rumah memiliki harga di bawah rata-rata, sementara sedikit rumah memiliki harga yang sangat tinggi. Ini menunjukkan adanya **outliers** atau properti premium.



Distribusinya juga sangat **right-skewed**, menunjukkan mayoritas properti memiliki lahan kecil hingga sedang, sementara hanya sedikit yang memiliki ukuran lahan besar. Properti dengan ukuran lahan besar mungkin menjadi outlier yang memengaruhi rata-rata harga.



`Mayoritas properti dibangun antara tahun 1970–2010, dengan peningkatan signifikan pada tahun 2000-an. Ini bisa menunjukkan bahwa properti yang lebih baru memiliki pengaruh signifikan pada harga properti.`

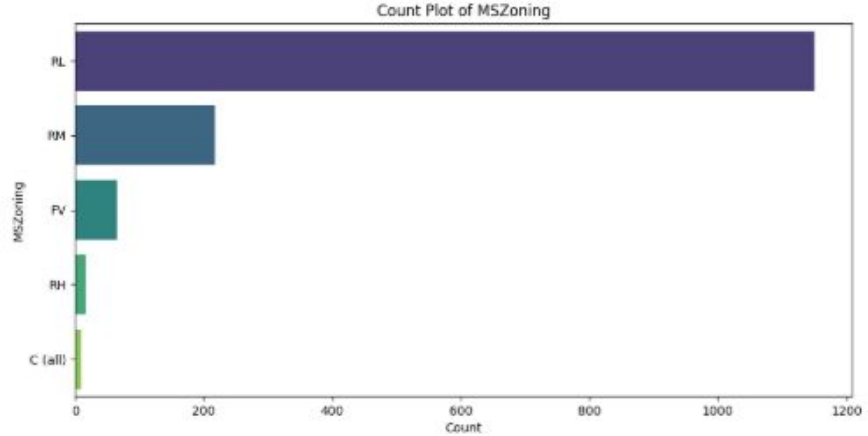


Distribusi terlihat mirip dengan `SalePrice`, **right-skewed**, menunjukkan bahwa properti dengan luas area yang lebih besar cenderung memiliki harga lebih tinggi. Properti yang jauh lebih besar dari rata-rata perlu diperiksa sebagai potensi outlier.

Properti yang dibangun lebih baru (YearBuilt) lebih menarik bagi pembeli dan mungkin memiliki harga lebih tinggi. Fokus pada properti baru dapat memberikan keuntungan di pasar.

Numerical Data

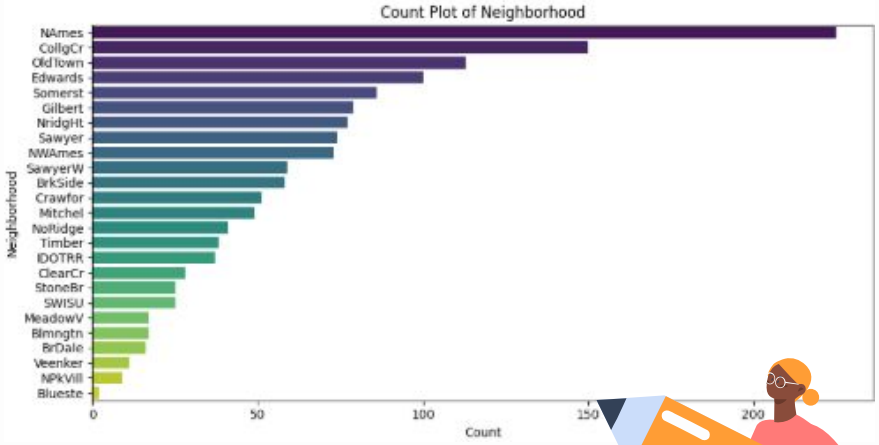
5. EDA – Univariate Analysis (2)



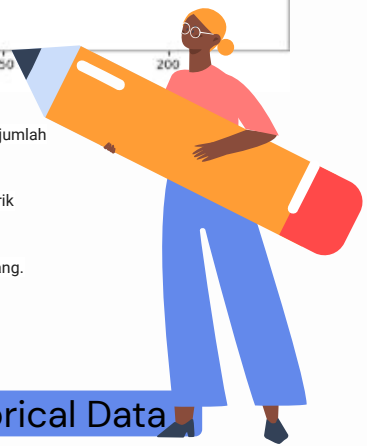
- **Observasi**: Sebagian besar properti berada di zona **RL (Residential Low Density)**, diikuti oleh zona **RM (Residential Medium Density)**.

Insight Bisnis:

- Fokus pemasaran properti di zona RL karena dominasi jumlah properti.
- Promosikan properti di zona lain (seperti Commercial atau Floating Village) untuk menarik pembeli dengan kebutuhan spesifik.



- **Observasi**: Beberapa lingkungan seperti **NridgHt** dan **CollgCr** memiliki jumlah properti yang dominan.
- **Insight Bisnis**:
  - Properti di lingkungan dengan volume tinggi dapat menarik karena popularitasnya.
  - Properti di lingkungan dengan jumlah lebih kecil dapat dipromosikan sebagai area yang lebih eksklusif atau tenang.

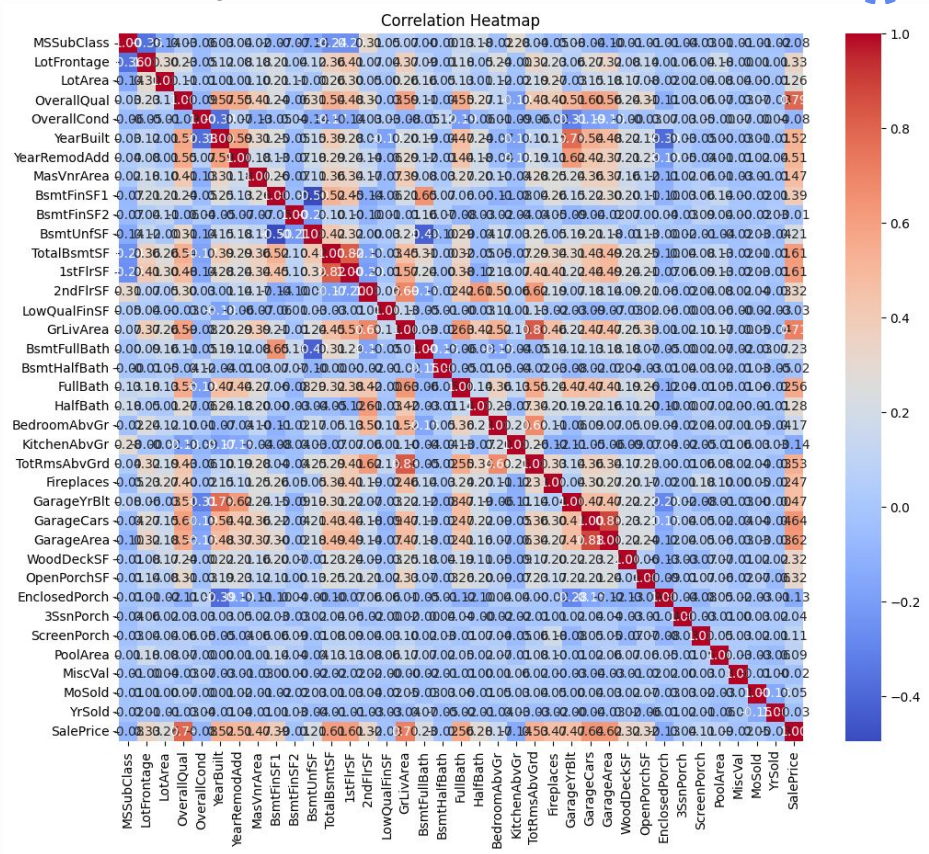


Categorical Data

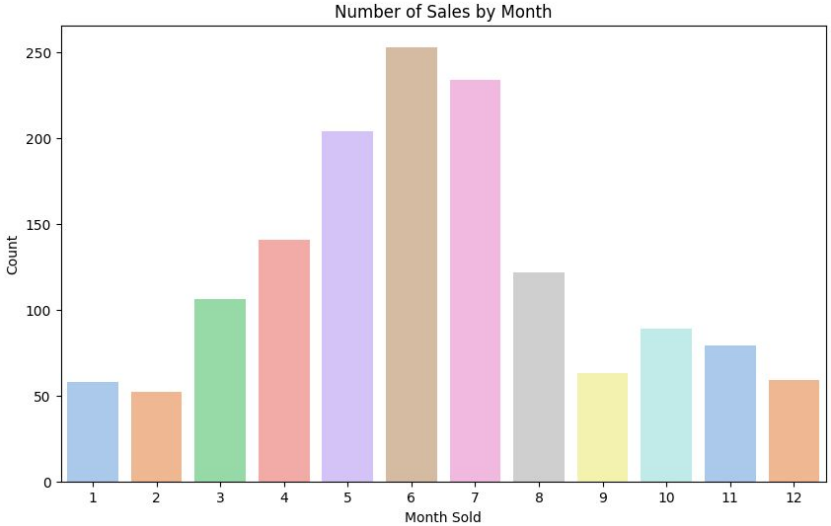
6. EDA – Multivariate Analysis

Insight Utama dari Heatmap Korelasi

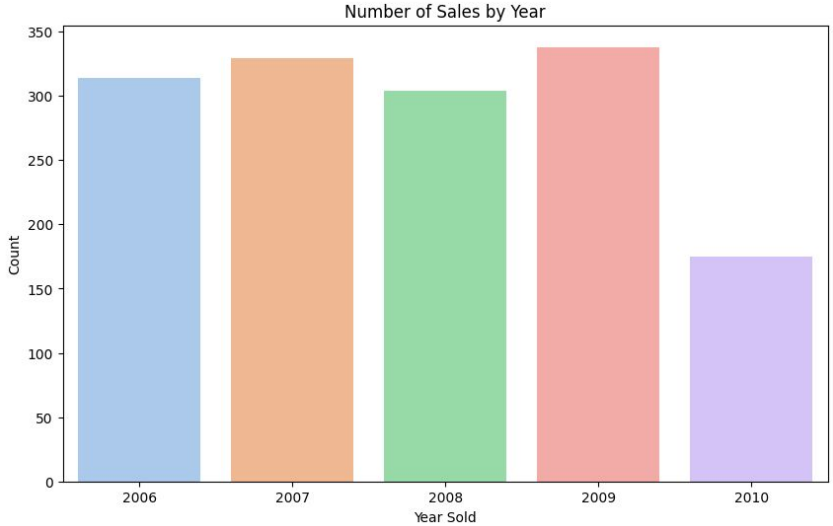
1. **Korelasi Kuat dengan SalePrice**:
   - **OverallQual, GrLivArea, GarageArea, TotalBsmtSF**: Kualitas bangunan, luas area layak huni, garasi, dan basement yang besar berkontribusi pada harga properti yang lebih tinggi.
2. **Korelasi Lemah**:
   - **YrSold, MiscVal**: Tidak memiliki pengaruh signifikan terhadap harga properti.
3. **Korelasi Antar Fitur**:
   - Hubungan kuat seperti **GarageArea & GarageCars**. Pilih salah satu untuk menghindari redundansi dalam model.
4. **Insight Bisnis**:
   - **Properti Premium**: Fokus pada properti berkualitas tinggi dan luas.
   - **Properti Budget**: Targetkan properti kecil dengan kualitas rendah untuk pembeli dengan anggaran terbatas.
   - Gunakan fitur kunci untuk menyusun strategi harga yang lebih akurat.



6. EDA



Penjualan Bulanan: Penjualan memuncak pada **Juni-Juli** dan menurun pada **November-Februari**, mencerminkan tren musiman. **Strategi**: Fokus pemasaran di musim panas dan promosi diskon di musim dingin.

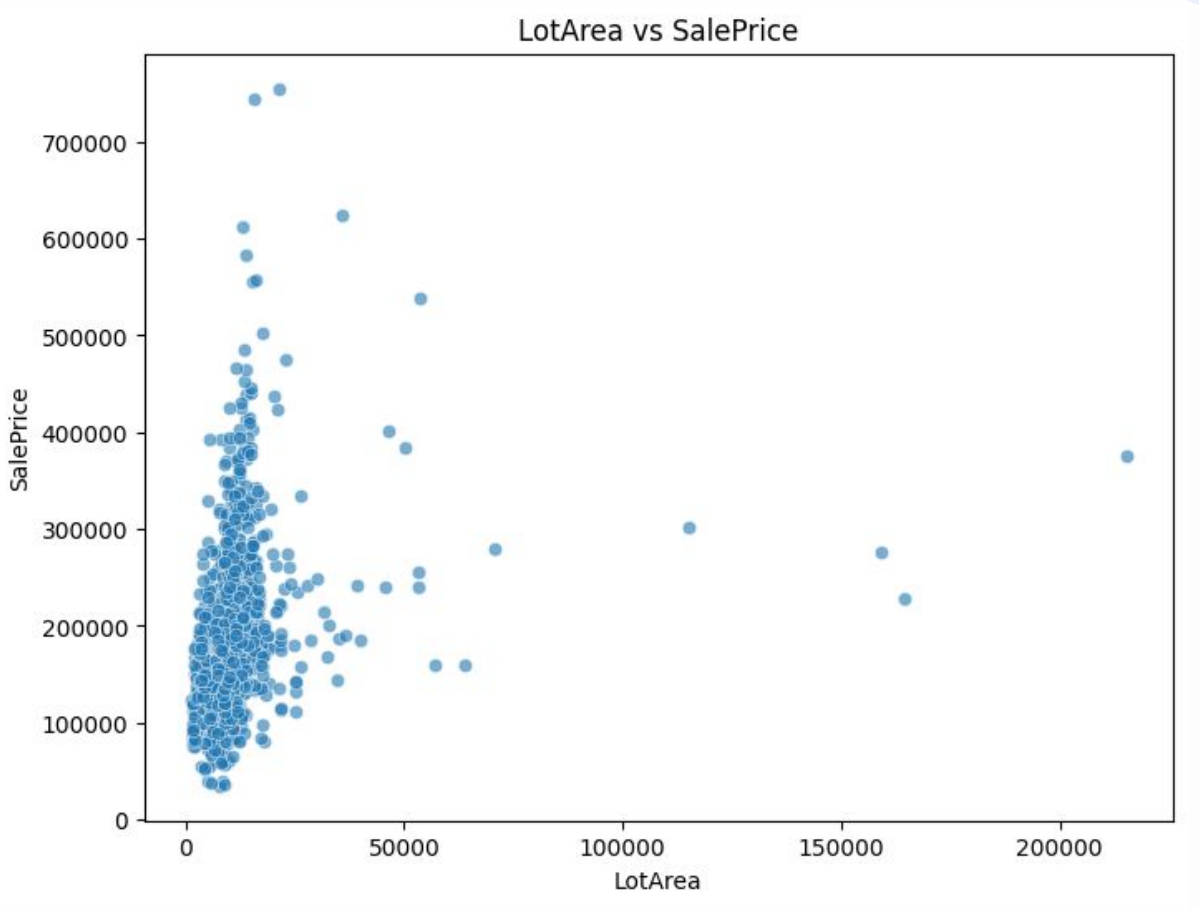


Penjualan Tahunan: Stabil dari **2006-2009**, menurun tajam di **2010**. **Strategi**: Analisis lebih lanjut untuk memahami penurunan dan optimalkan penjualan pada tahun rendah.

6. EDA

Insight:

1. **Mayoritas Properti**: Sebagian besar properti memiliki luas lahan di bawah 20,000 sq.ft. dengan harga yang bervariasi, menunjukkan pasar yang dominan untuk properti kecil hingga sedang.
2. **Outlier**: Properti dengan luas lahan di atas 50,000 sq.ft. memiliki harga yang lebih tinggi, tetapi tidak selalu proporsional, menunjukkan faktor lain seperti lokasi atau fasilitas memengaruhi harga.
3. **Strategi Bisnis**: Promosikan properti dengan luas lahan besar sebagai segmen premium, dan fokuskan strategi harga untuk properti kecil sesuai permintaan pasar utama.



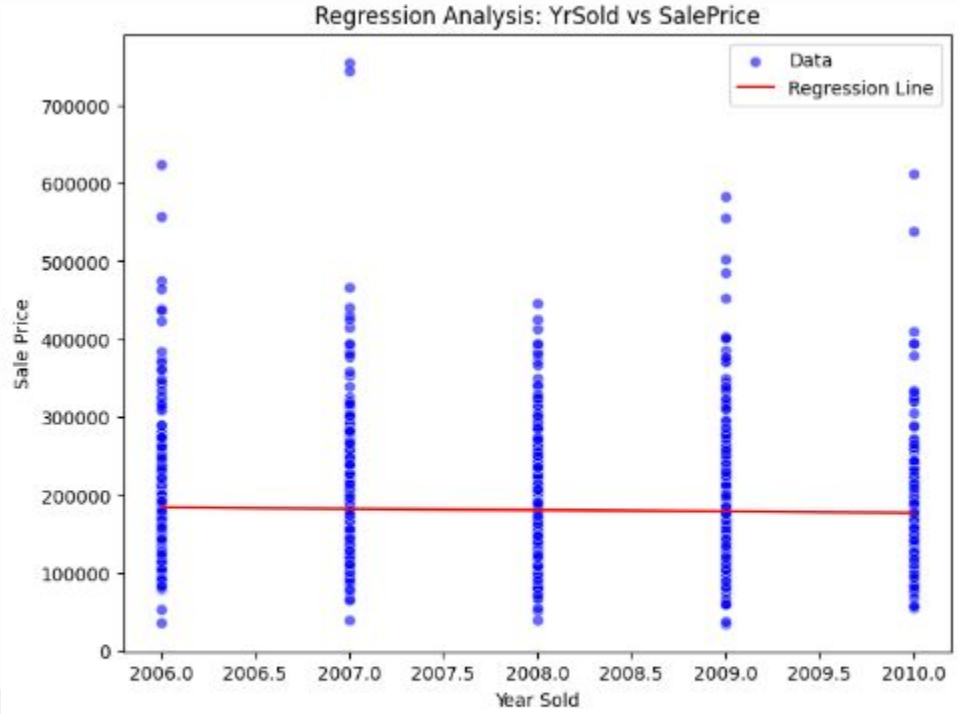
7. Correlation Analysis

Hipotesis

HO: Tidak ada perbedaan signifikan antara harga rumah pada tahun 2006 sampai 2010, yang menunjukkan bahwa kenaikan harga rumah antara tahun-tahun ini tidak signifikan.

H1: Ada perbedaan signifikan antara harga rumah pada tahun 2006 sampai 2010, yang menunjukkan bahwa kenaikan harga rumah antara tahun-tahun ini tidak signifikan

7. Correlation Analysis



Pengujian yang Digunakan	Regresi Linear
Tujuan	Menguji hubungan linear antara tahun penjualan (YrSold) dan harga rumah (SalePrice)
Statistik Uji	R-squared: 0.001, F-statistic: 1.221
P-Value	0.2690

```
Hipotesis:
H0: Tidak ada perbedaan signifikan antara harga rumah pada tahun 2006 sampai 2010.
H1: Ada perbedaan signifikan antara harga rumah pada tahun 2006 sampai 2010.

Kesimpulan: Gagal menolak Hipotesis Nol (H0).
Tidak ada perbedaan signifikan antara harga rumah pada tahun 2006 sampai 2010.
```

7. Correlation Analysis

Pengujian yang Digunakan	Uji Pearson
Tujuan	Menguji hubungan signifikan antara tahun penjualan dan harga rumah
Statistik Uji	Pearson Correlation: -0.0289
P-Value	0.2694
Kesimpulan	Gagal menolak Hipotesis Nol (H0)
Interpretasi	Tidak ada hubungan signifikan antara tahun penjualan dan harga rumah

Pengujian yang Digunakan	Uji Chi-Square
Tujuan	Menguji perbedaan signifikan dalam distribusi harga rumah
Statistik Uji	Chi2 Statistic: 8.6713
P-Value	0.7307
Kesimpulan	Gagal menolak Hipotesis Nol (H0)
Interpretasi	Tidak ada perbedaan signifikan dalam distribusi harga rumah dari tahun 2006-2010

Pengujian yang Digunakan	Uji T-Test
Tujuan	Menguji perbedaan signifikan dalam rata-rata harga rumah
Statistik Uji	T-Statistic: 0.4950
P-Value	0.5344
Kesimpulan	Gagal menolak Hipotesis Nol (H0)
Interpretasi	Tidak ada perbedaan signifikan dalam rata-rata harga rumah dari tahun 2006-2010

Pengujian yang Digunakan	Uji ANOVA
Tujuan	Menguji perbedaan signifikan dalam rata-rata harga rumah
Statistik Uji	F-Statistic: 0.6455
P-Value	0.6301
Kesimpulan	Gagal menolak Hipotesis Nol (H0)
Interpretasi	Tidak ada perbedaan signifikan dalam rata-rata harga rumah dari tahun 2006-2010

8. Key Insight

Price Drivers

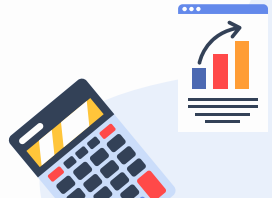
1. OverallQual and GrLivArea significantly influence house prices.
2. Premium neighborhoods like StoneBr and NridgHt command higher prices.

Market Stability

No significant year-over-year price variations observed (2006–2010).

Seasonal Trends

Sales peak during May–July; decline during November–February.



9. Conclusion

Temuan Utama:

- **Profiling Data**: Analisis statistik deskriptif mengungkapkan distribusi harga rumah berdasarkan ukuran pemusatan (mean, median, mode) dan ukuran penyebaran (standar deviasi, IQR), menunjukkan bahwa sebagian besar harga rumah terkonsentrasi di kisaran menengah.
- **Preprocessing Data**: Penanganan missing values dilakukan dengan teknik imputasi untuk kolom dengan <20% missing values, sementara kolom dengan >80% missing values dihapus karena tidak relevan.
- **EDA dan Korelasi**: Analisis menunjukkan korelasi kuat antara fitur seperti `OverallQual` (0.79), `GrLivArea` (0.71), dan `SalePrice`. Fitur ini memiliki dampak signifikan terhadap penentuan harga properti.
- **Pengujian Statistik**: Pengujian Pearson, Chi-Square, T-Test, ANOVA, dan Regresi Linear menunjukkan tidak ada perubahan signifikan pada harga rumah antara tahun 2006 hingga 2010 (contoh: Pearson Correlation = -0.0289, ANOVA F-Statistic = 0.6455, p-value = 0.63).

9. Conclusion

Aplikasi Bisnis:

- **Optimasi Harga Properti**: Fokuskan strategi harga pada fitur yang berdampak besar, seperti kualitas bangunan, luas area, dan fasilitas tambahan seperti garasi besar.
- **Strategi Pemasaran Musiman**: Maksimalkan penjualan pada bulan dengan permintaan tinggi (Mei-Juli) dengan kampanye yang terfokus.
- **Fokus pada Lingkungan Premium**: Prioritaskan pemasaran di lingkungan dengan harga tinggi, seperti `NridgHt` dan `StoneBr`, untuk menarik pembeli potensial dari segmen atas.
- **Efisiensi Sumber Daya**: Dengan tidak adanya perbedaan signifikan dalam harga berdasarkan tahun penjualan, alokasikan sumber daya pada pengembangan fitur utama yang relevan dengan kebutuhan pembeli.

Thanks!

Do you have any questions?

hijirdw@gmail.com

https://github.com/hijirdella/House-Price-Analysis-EDA-and-Correlation-Insights

https://www.linkedin.com/in/hijirdella/