Laporan Final Project Data Science: Rekomendasi Pembelian Rumah

Business Understanding

1. Latar Belakang

   Rumah adalah kebutuhan utama manusia untuk hidup. Dalam pembelian rumah, terdapat proses pemilihan rumah yang melibatkan faktor-faktor tertentu untuk menentukan apakah pilihan rumah tersebut tepat dan menguntungkan. Dalam proses pembelian tersebut, pembeli rumah biasanya akan bingung bagaimana memilih dan menilai rumah yang tepat dari banyaknya pilihan rumah yang ada karena batasnya kemampuan informasi. Oleh karena itu, diperlukannya sistem untuk merekomendasikan pembelian rumah dan membantu pembeli membuat keputusan agar bisa membeli rumah yang tepat dan menguntungkan.

2. Tujuan Bisnis

   Menyediakan rekomendasi rumah yang optimal bagi calon pembeli, berdasarkan atribut yang relevan, seperti harga, lokasi, fasilitas, dan preferensi lainnya.

3. Tujuan Data Science
   - Mengidentifikasi atribut-atribut rumah yang berpengaruh kepada harga rumah dan visualisasi hubungan antar atribut rumah.
   - Memprediksi rumah yang tepat dan menguntungkan berdasarkan atribut yang ditentukan.

4. Permasalahan Bisnis

   Tidak adanya alat bantu yang dapat memberikan rekomendasi berbasis data.

5. Kebutuhan Data Science
   - Data: Informasi atribut rumah.
   - Analisis: Pengaruh antar atribut rumah.
   - Model: Sistem rekomendasi untuk pembelian rumah.

Implementation

```
import pandas as pd

df = pd.read_csv('data_porto_2.csv', index_col='date', parse_dates=True)
df.head(10)
```

date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	street	city	statezip	country
2014-05-02	313000.0	3.0	1.50	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Densmore Ave N	Shoreline	WA 98133	USA
2014-05-02	2384000.0	5.0	2.50	3650	9050	2.0	0	4	5	3370	280	1921	0	709 W Blaine St	Seattle	WA 98119	USA
2014-05-02	342000.0	3.0	2.00	1930	11947	1.0	0	0	4	1930	0	1966	0	26206-26214 143rd Ave SE	Kent	WA 98042	USA
2014-05-02	420000.0	3.0	2.25	2000	8030	1.0	0	0	4	1000	1000	1963	0	857 170th Pl NE	Bellevue	WA 98008	USA
2014-05-02	550000.0	4.0	2.50	1940	10500	1.0	0	0	4	1140	800	1976	1992	9105 170th Ave NE	Redmond	WA 98052	USA

Data asli dari dataset rumah

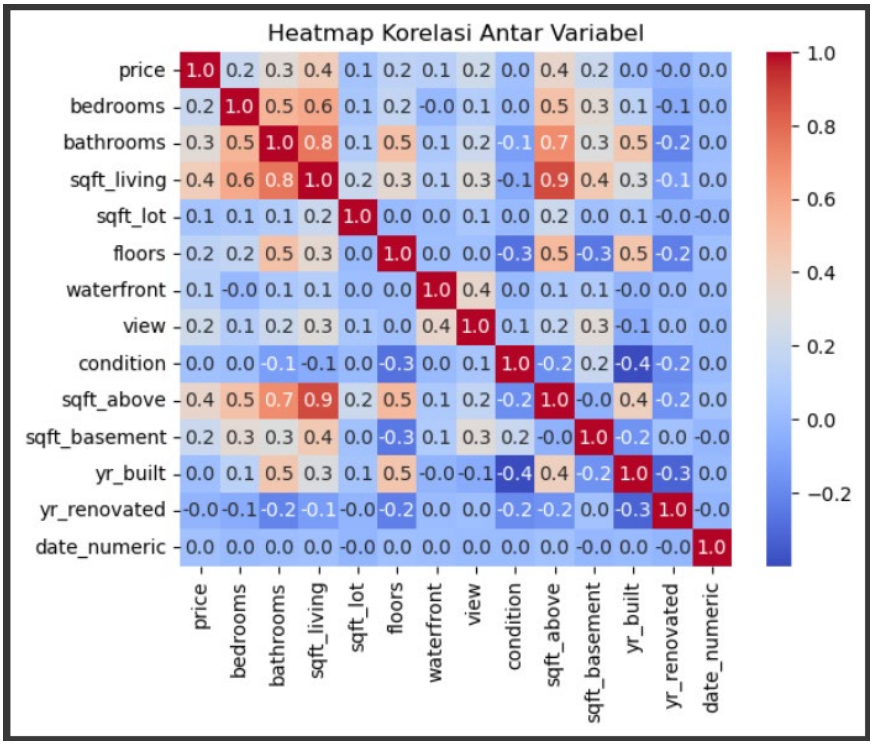
```
df.describe()
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated
count	4.600000e+03	4600.000000	4600.000000	4600.000000	4.600000e+03	4600.000000	4600.000000	4600.000000	4600.000000	4600.000000	4600.000000	4600.000000	4600.000000
mean	5.519630e+05	3.400870	2.160815	2139.346957	1.485252e+04	1.512065	0.007174	0.240652	3.451739	1827.265435	312.081522	1970.786304	808.608261
std	5.638347e+05	0.908848	0.783781	963.206916	3.588444e+04	0.538288	0.084404	0.778405	0.677230	862.168977	464.137228	29.731848	979.414536
min	0.000000e+00	0.000000	0.000000	370.000000	6.380000e+02	1.000000	0.000000	0.000000	1.000000	370.000000	0.000000	1900.000000	0.000000
25%	3.228750e+05	3.000000	1.750000	1460.000000	5.000750e+03	1.000000	0.000000	0.000000	3.000000	1190.000000	0.000000	1951.000000	0.000000
50%	4.609435e+05	3.000000	2.250000	1980.000000	7.683000e+03	1.500000	0.000000	0.000000	3.000000	1590.000000	0.000000	1976.000000	0.000000
75%	6.549625e+05	4.000000	2.500000	2620.000000	1.100125e+04	2.000000	0.000000	0.000000	4.000000	2300.000000	610.000000	1997.000000	1999.000000
max	2.659000e+07	9.000000	8.000000	13540.000000	1.074218e+06	3.500000	1.000000	4.000000	5.000000	9410.000000	4820.000000	2014.000000	2014.000000

Banyak data yang masih belum rapi untuk dilakukan analisis. Oleh karena itu diperlukan pembersihan data, seperti mencari data yang kosong, tidak konsisten, *outlier*, dan sebagainya, dengan mengupdate data terbaru seperti berikut.

```
df_cleaned = df[
    (df['price'] <= 7000000) &
    (df['price'] > 0) &
    (df['sqft_living'] <= 12000) &
    (df['sqft_above'] <= 9000) &
    (df['bathrooms'] <= 7) &
    (df['sqft_lot'] <= 800000) &
    (df['sqft_basement'] <= 4000)
]
```

Setelah data *well prepared,* tahap selanjutnya yaitu Explanatory Data Analysis (EDA).



Dengan ini data dapat dipahami korelasi antara atributnya.

1. Tahun dibangun berbanding terbalik dengan kondisi. Semakin besar tahun dibangun, semakin buruk kondisinya.
2. Luas area di atas tanah ‘sqft_living’ berbanding lurus dengan beberapa atribut seperti price, bedrooms, bathrooms, serta sqft_above.

Berdasarkan EDA di atas, ditemukan bahwa ‘**sqft_living’** dan **‘price’** memiliki korelasi yang sangat tinggi sebesar 0.9, yang menunjukkan bahwa harga rumah dipengaruhi oleh luas ruangan yang bisa dihinggapi (sqft_living). Selain itu, sqft_living juga berpengaruh kepada luas bathroom dan bedroom dimana semakin besar sqft_living, maka luas ruangan lainnya akan besar juga. Oleh karena itu, kedua atribut ini dapat digunakan untuk menentukan kriteria rumah yang baik. Dengan ini kita bisa menghitung rasio harga per kaki persegi (price_per_sqft), untuk menilai apakah harga rumah yang ditawarkan wajar dibandingkan dengan luas area yang dihuni (sqft_living). Rasio ini membantu mengevaluasi efisiensi harga properti, memungkinkan pembeli untuk memutuskan membeli rumah yang tepat.

```
filtered_houses['price_per_sqft'] = filtered_houses['price'] / filtered_houses['sqft_living']
recommended_houses = filtered_houses.sort_values(by='price_per_sqft').head(10)
print(recommended_houses)
```

Selanjutnya, dengan memfilter rumah dengan tahun bangun minimal 1990 dan tahun renovasi rumah (jika dibangun sebelum tahun 1990), maka kita bisa mendapatkan rekomendasi rumah dengan cara berikut.

```
built_after_1990 = df_cleaned[df_cleaned['yr_built'] >= 1990]
renovated_after_2000 = df_cleaned[(df_cleaned['yr_built'] < 1990) & (df_cleaned['yr_renovated'] > 2000)]
filtered_houses = pd.concat([built_after_1990, renovated_after_2000])
filtered_houses['price_per_sqft'] = filtered_houses['price'] / filtered_houses['sqft_living']
recommended_houses = filtered_houses.sort_values(by=['price', 'price_per_sqft']).head(10)

print(recommended_houses)
```

Output:

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	\
83300.0	3.0	2.00	1490	7770	1.0	
90000.0	2.0	1.00	790	2640	1.0	
90000.0	2.0	1.00	580	7500	1.0	
100000.0	2.0	1.00	910	22000	1.0	
110000.0	2.0	1.00	800	15000	1.0	
129000.0	2.0	1.00	1150	30184	1.0	
132250.0	4.0	2.25	2192	12128	2.0	
132250.0	4.0	2.25	1830	8734	2.0	
132500.0	3.0	1.00	1080	10500	1.0	
134000.0	2.0	1.50	980	5000	2.0	

waterfront	view	condition	sqft_above	sqft_basement	yr_built	\
0	0	4	1490	0	1990	
0	0	3	790	0	1973	
0	0	3	580	0	1943	
0	0	3	910	0	1956	