Penerapan Model Regresi Linear OLS dan Bayesian untuk Memprediksi Harga Rumah di California

Vika Valencia Susanto
2440062123
Teknik Informatika dan Statistika
vika.susanto@binus.ac.id

Nicholaus
2440102120
Teknik Informatika dan Statistika
nicholaus@binus.ac.id

Vieren Cristian
2440102202
Teknik Informatika dan Statistika
vieren.cristian@binus.ac.id

Abstrak

Dalam beberapa tahun belakangan ini, harga rumah di berbagai wilayah terus meningkat dikarenakan adanya peningkatan ketertarikan individu dalam membeli rumah di suatu daerah. Tempat tinggal yang baik ini mampu menciptakan kehidupan yang sehat, kondusif, serta ideal yang dapat mendukung dalam meningkatnya kualitas hidup. Oleh karena itu, mengetahui prediksi harga rumah merupakan sebuah keuntungan dan menjadi suatu hal yang penting. Dalam *paper* ini, *Dataset* yang digunakan adalah *California Housing Prices Dataset* untuk memprediksi harga rumah. Penelitian ini membandingkan dua metode prediksi, yaitu model regresi linear ganda OLS(*Ordinary Least Square)* dengan model regresi linear bayesian.

KEYWORDS: *California housing prices* - Regresi Linear Ganda OLS - Regresi Linear Ganda Bayesian.

Pendahuluan

Pada saat ini, banyak orang yang tertarik dengan transaksi jual beli rumah di berbagai wilayah sebagai tempat tinggal ataupun investasi. Dengan mempunyai tempat tinggal yang layak di suatu wilayah yang baik merupakan impian dari setiap individu untuk mendapatkan kehidupan yang sehat, kondusif, serta ideal. Akan tetapi, harga rumah yang cukup tinggi membuat orang bertanya-tanya apakah harga yang ditawarkan sesuai dengan lokasinya ataupun apakah harga di kawasan tersebut akan naik atau turun kedepannya.

Sama halnya seperti harga rumah yang ada di negara Amerika Serikat, salah satunya wilayah California yang merupakan negara bagian di Amerika Serikat. Tercatat hingga tahun 2022 harga rumah di California terbilang mahal dan menjadi sebagai negara bagian kedua yang mempunyai harga untuk membeli rumah termahal setelah New York. Kenaikan harga di California ini terus berlangsung dari tahun 2020.

Dari data ekonomi yang ada di negara Amerika dikatakan bahwa sedang terjadi kenaikan suku bunga dimana hal ini bisa saja berdampak pada harga rumah di wilayah California. Harga yang tidak stabil, tidak pasti, dan tidak bisa diprediksi membuat para pembeli dan investor membutuhkan sebuah sistem untuk memprediksi harga rumah.

Prediksi harga dari sebuah rumah selalu menjadi tantangan tersendiri bagi *data scientist* dan *engineer*. Pada *paper* ini dilakukan pengujian menggunakan beberapa model

untuk melakukan prediksi harga rumah di California. Berdasarkan deskripsi tersebut, dengan menggunakan *dataset* perumahan di California, beserta model linear regresi berganda, dan metode lainnya, diharapkan hasil dari penelitian ini bisa membantu mencari model terbaik untuk memprediksi harga rumah berdasarkan beberapa aspek perumahan lainnya.

Penelitian Sebelumnya

Prediksi harga rumah di daerah California sebelumnya sudah dilakukan oleh banyak penelitian dan berbagai metode. Salah satunya, terdapat *paper* yang berjudul "Prediction of California House Price Based On Multiple Linear Regression" yang ditulis oleh Zixu Wu. Dalam *paper* ini, model prediksi yang dilakukan dalam memprediksi harga rumah di California adalah menggunakan regresi linear berganda. Model prediksi ini bertujuan untuk memprediksi "annual average sales prices of local houses" dengan faktor yang paling berpengaruh pada harga rumahnya adalah "number of rooms", "income distribution", "teacher-student ratio".

Adapun penelitian lainnya yang berjudul "The Time-Series Properties of House Prices: A Case Study of the Southern California Market" yang ditulis oleh Rangan Gupta dan Stephen M. Miller. Dalam penelitian ini, hubungan antar kota berdasarkan MSA-nya (*Metropolitan Statistical Area)* dengan menggunakan model VAR (*Vector Autoregressive)* dan VEC (*Vector Error-correction),* Bayesian, dengan menggunakan *prior* yang beragam. Dalam *paper* ini, hubungan *time-series* sangat dipertimbangkan dalam memprediksi harga rumah. Ini terbukti dengan adanya perbedaan *time-series* dalam sebuah model dapat memberikan hasil yang lebih baik dalam memprediksi harga rumah dalam MSA yang berbeda.

Metode Penelitian

A. Regresi Linear Berganda

Regresi linear berganda adalah lanjutan dari regresi linear yang merupakan pendekatan statistik untuk memprediksi hasil variabel berdasarkan dua atau lebih variabel. Variabel dependennya adalah variabel yang akan diprediksi, dan faktor yang digunakan untuk memprediksi nilai dari variabel dependen disebut dengan variabel independen. Analisa menggunakan teknik ini untuk melihat variasi model dan kontribusi dari setiap variabel independen kepada total varians.

Model dari regresi linear berganda:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Dimana,

- Y = variable dependen
- X = variabel independen
- β = koefisien regresi
- ε = eror

B. *Ordinary Least Squares*

Regresi *Ordinary Least Squares* adalah metode untuk menghitung koefisien dari persamaan regresi linear yang menjelaskan tentang hubungan antara satu atau lebih variabel independen dan variabel dependen. Tujuan dari metode ini adalah untuk mengurangi perbedaan nilai sum of square antara nilai prediksi dan aktual.
Formula untuk menghitung koefisien vektor melalui OLS:

$$\beta = (X'X)^{-1}X'Y$$

Dimana,

- X = matriks berisi nilai dari variabel prediktor, di mana untuk nilai intercept atau X_0 diisi dengan nilai 1
- X' = transpose dari matriks X
- Y = vektor dari variabel dependen

C. Metode Bayes

Metode ini digunakan untuk melakukan estimasi parameter regresi linear berganda dengan mempertimbangkan informasi awal (prior). Distribusi prior akan dikombinasikan dengan fungsi likelihood untuk menghasilkan distribusi posterior yang akan menjadi dasar inferensi.

Rumus umum metode bayes adalah:

$$P(\mu|y) = \frac{P(y|\mu)P(\mu)}{P(y)}$$

Dimana:

- $P(\mu|y)$ = distribusi posterior μ
- $P(y|\mu)$ $P(\mu)$ pada umumnya tidak diketahui, biasanya hanya distribusi prior dan fungsi likelihoodnya yang dinyatakan.

D. Uji signifikansi setiap parameter

- *T-test*

Merupakan sebuah tes statistik untuk membandingkan rata-rata dari dua grup. Biasanya digunakan untuk tes hipotesis untuk melihat apakah treatment berpengaruh secara signifikan pada populasi. Berikut merupakan hipotesis dari *T-test*:

Hipotesis:
H_0: $\mu_i = 0$, variabel independen i tidak berpengaruh signifikan terhadap variabel dependen
H_1: $\mu_i \neq 0$, variabel independen i berpengaruh signifikan terhadap variabel dependen

Hasil dari *t-test* pada R akan menghasilkan dua nilai, yaitu t-value dan p-value. *T-value* merupakan cara untuk mengukur perbedaan rata-rata populasi dan nilainya biasa dibandingkan dengan nilai tabel. Sedangkan *P-value* merupakan probabilitas untuk memperoleh nilai t dengan nilai absolut.

Kriteria tolak H0 adalah jika nilai P-value < alpha (0.05) atau T-value > t-table

- *Credible Interval*

  Credible interval merupakan interval di mana nilai parameter yang tidak teramati jatuh dengan probabilitas tertentu. Berbeda dengan *confidence interval* yang memprediksi koefisien berada dalam suatu interval, *credible interval* akan memprediksi bahwa sebuah koefisien berada di atas nilai CI_low dan di bawah CI_high. Sebuah variabel independen bisa dibilang berpengaruh secara signifikan terhadap variabel dependen jika dalam *credible interval*-nya tidak mencakup nilai 0.

E. Backward Elimination Method

Metode *backward elimination* merupakan salah satu metode yang digunakan untuk feature selection, di mana fitur/variabel independen yang tidak mempengaruhi variabel dependen secara signifikan akan dieliminasi. Idenya adalah dengan menghilangkan variabel yang paling tidak berpengaruh satu per satu dengan melihat nilai *P-value* dari setiap variabel. Jadi, satu per satu variabel dengan nilai *P-value* tertinggi akan dibuang dan dilakukan pengecekan kembali terhadap nilai *P-value* dari model terbaru sampai mendapatkan model yang signifikan.

F. *Mean Squared Error* dan RMean Squared Error

Mean Squared Error merupakan fungsi yang menghitung seberapa dekat nilai regresi dengan sekumpulan data points dengan menghitung perbedaan kuadrat rata-rata antara nilai estimasi dengan nilai sebenarnya. Mayoritas nilai dari MSE akan selalu positif dan merupakan ukuran kualitas dari sebuah model regresi. Karena diturunkan dari Euclidean distance yang dikuadratkan, nilai MSE akan semakin berkurang jika errornya mendekati 0, yang artinya semakin kecil nilai MSE, maka model dianggap baik dalam memprediksi. Sedangkan, untuk RMSE sendiri merupakan akar dari nilai MSE.

G. Mean Absolute Percentage Error

Merupakan ukuran kesalahan relatif dan menyatakan persentase kesalahan hasil pendugaan. MAPE digunakan ketika ukuran variabel ramalan penting dalam mengevaluasi ketepatan pendugaan.

Rumus dari MAPE:

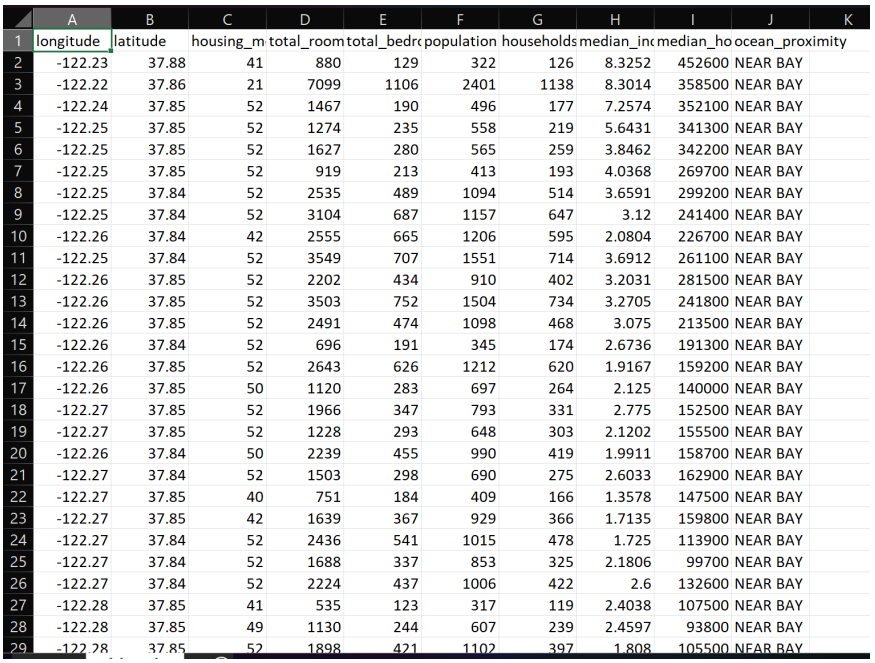
$$MAPE = \sum_{t=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right|$$

Dimana,

y = nilai hasil aktual

$\hat{y}$ = nilai hasil pendugaan

Dataset (California_Housing)



	A	B	C	D	E	F	G	H	I	J	K
1	longitude	latitude	housing_m	total_room	total_bedr	population	households	median_inc	median_ho	ocean_proximity	
2	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY	
3	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY	
4	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY	
5	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY	
6	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY	
7	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY	
8	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY	
9	-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY	
10	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY	
11	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY	
12	-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY	
13	-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY	
14	-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY	
15	-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY	
16	-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY	
17	-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY	
18	-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY	
19	-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY	
20	-122.26	37.84	50	2239	455	990	419	1.9911	158700	NEAR BAY	
21	-122.27	37.84	52	1503	298	690	275	2.6033	162900	NEAR BAY	
22	-122.27	37.85	40	751	184	409	166	1.3578	147500	NEAR BAY	
23	-122.27	37.85	42	1639	367	929	366	1.7135	159800	NEAR BAY	
24	-122.27	37.84	52	2436	541	1015	478	1.725	113900	NEAR BAY	
25	-122.27	37.84	52	1688	337	853	325	2.1806	99700	NEAR BAY	
26	-122.27	37.84	52	2224	437	1006	422	2.6	132600	NEAR BAY	
27	-122.28	37.85	41	535	123	317	119	2.4038	107500	NEAR BAY	
28	-122.28	37.85	49	1130	244	607	239	2.4597	93800	NEAR BAY	
29	-122.28	37.85	52	1898	421	1102	397	1.808	105500	NEAR BAY	

Gambar 1: Data teratas dari dataset

Data yang digunakan pada paper ini berkaitan dengan rumah yang ditemukan di distrik California tertentu dan beberapa statistik ringkasan mengenai mereka berdasarkan data sensus di California.

Dataset ini akan digunakan untuk memprediksi *housing price* di distrik California dengan membuat model regresi.

Dataset ini berisi 20640 data dengan total 10 variabel:

No	Variabel	Penjelasan variabel
1.	longitude	(bujur) Ukuran seberapa jauh rumah ke arah barat; semakin tinggi nilainya maka bisa dibilang rumah terletak semakin jauh ke barat
2.	latitude	(lintang) Ukuran seberapa jauh rumah ke arah timur; semakin tinggi nilainya maka bisa dibilang rumah terletak semakin jauh ke timur
3.	housingMedianAge	Median usia rumah dalam satu blok; angka yang lebih rendah adalah bangunan yang lebih baru
4.	totalRooms	Jumlah total kamar dalam satu blok

5.	totalBedrooms	Jumlah total kamar tidur dalam satu blok
6.	population	Jumlah total orang yang tinggal dalam satu blok
7.	households	Jumlah total rumah tangga, sekelompok orang yang tinggal dalam satu unit rumah, untuk satu blok
8.	medianIncome	Pendapatan rata-rata untuk rumah tangga dalam satu blok rumah (diukur dalam puluhan ribu Dolar AS)
9.	medianHouseValue	Median nilai/harga rumah untuk rumah tangga dalam satu blok (diukur dalam Dolar AS)
10.	oceanProximity	Lokasi jarak rumah dengan laut

Ringkasan statistik dari dataset dijabarkan dalam gambar di bawah ini,

```
> summary(housing)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms     population
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0   Min.   :    3
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 296.0   1st Qu.:  787
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0   Median : 1166
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9   Mean   : 1425
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0   3rd Qu.: 1725
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0   Max.   :35682
                                                                    NA's   :207
   households     median_income     median_house_value ocean_proximity
 Min.   :   1.0   Min.   : 0.4999   Min.   : 14999     Length:20640
 1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600     Class :character
 Median : 409.0   Median : 3.5348   Median :179700     Mode  :character
 Mean   : 499.5   Mean   : 3.8707   Mean   :206856
 3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725
 Max.   :6082.0   Max.   :15.0001   Max.   :500001
```

Gambar 2: Ringkasan statistik dari dataset.

Hasil dan Pembahasan

Komputasi dilakukan dengan menggunakan *software* R. Untuk model regresi linear berganda, variabel independen atau variabel yang akan digunakan untuk memprediksi adalah variabel “housingMedianAge”, “totalRooms”, “totalBedrooms”, “population”, “households”, “medianIncome”, “oceanProximity”. Sedangkan untuk variabel respon atau variabel yang akan diprediksinya adalah “medianHouseValue”. Dalam penelitian ini, dari 20640 observasi, diambil sampel sebanyak 500 observasi secara acak untuk mengolah model. Pada 500 observasi yang sudah diambil didapatkan 3 *missing value* yang dibuang sehingga untuk membuat model, hanya digunakan 497 sampel.

Variabel “longitude” dan “latitude” tidak bisa langsung digunakan karena tidak ada hubungan linear yang nyata. Agar bisa digunakan, variabel ini akan di convert menjadi sebuah variabel baru yaitu “county”, kota di California. Karena variasi dari data county ini terlalu banyak, hanya 5 data terbanyak yang terpilih dan sisanya diberikan kategori “others”.

Data kemudian akan diproses dengan menghilangkan missing value, dan mengubah nilai dari data kategorik menjadi faktor dengan label numerik:

- Ocean Proximity

Nama kategori	Faktor
<1H OCEAN	1
INLAND	2
NEAR BAY	3
NEAR OCEAN	4

- County

Nama kategori	Faktor
Los Angeles	1
Orange	2
San Diego	3
Alameda	4
San Benardino	5
Others	6

Terdapat sebanyak 4 model regresi yang akan dibuat, yaitu model regresi OLS serta bayesian tanpa menggunakan variabel “county”, dan model regresi OLS serta model bayesian dengan menggunakan variabel longitude dan latitude yang sudah dikonversi menjadi variabel “county”. Uji signifikansi akan digunakan pada setiap model untuk melihat(memeriksa) dan mengeliminasi variabel independen yang tidak signifikan.

Berikut adalah nilai p-value dan credible interval tiap variabel untuk tiap model.

Variabel	Tanpa County		Dengan County	
	OLS	Bayes	OLS	Bayes
housing_median_age	5.08e-5	690.53134 - 1636.096984	0.00166	433.56571 - 1426.819239
total_rooms	0.636991	-11.43033 - 6.364199	0.77730	-10.99380 - 7.384582
total_bedrooms	0.975752	-62.38696 - 59.855762	0.84086	-65.17475 - 55.023618
population	0.000183	-47.46991 - (-18.246358)	0.00022	-47.45576 - (-18.671912)
households	0.002066	57.79552 -	0.00180	56.11488 -

		190.918566		190.297701
median_income	< 2e-16	37386.17610 - 44412.315017	<2e-16	37082.15330 - 44030.697240
ocean_proximity2	6.79e-16	-78380.56271 - (-52459.906679)	1.46e-13	-86898.24904 - (-55704.214624)
ocean_proximity3	0.643590	-11533.43989 - 21563.383115	0.81960	-16800.91701 - 22901.684379
ocean_proximity4	0.000613	18523.70422 - 49626.867254	0.00113	18653.45682 - 54391.836547
county2	-	-	0.03741	-48907.98465 - (-5288.479419)
county3	-	-	0.00500	-72231.79195 - (-19597.538245)
county4	-	-	0.66541	-38203.14771 - 22260.635087
county5	-	-	0.62375	-38506.68226 - 20660.221188
county6	-	-	0.67267	-19728.98199 - 11834.671766

Variabel yang tidak memiliki pengaruh yang signifikan ditunjukkan dengan nilai p-value yang diwarnai merah. Sebelum melakukan metode backward elimination, setiap variabel kategorik yang faktornya tidak berpengaruh signifikan akan digabung menjadi satu kategori, seperti contohnya faktor "ocean_proximity1" dan "ocean_proximity3" pada variabel "ocean_proximity" serta faktor "county1", "county4", "county5", dan county "6" pada variabel "county", sehingga ketika diuji signifikansinya kembali dihasilkan:

Variabel	Tanpa County		Dengan County	
	OLS	Bayes	OLS	Bayes
housing_median_age	2.57e-05	719.26293 - 1650.511507	0.000870	479.05129 - 1444.667095
total_rooms	0.672185	-10.69278 - 5.994393	0.669681	-11.11485 - 6.360366
total_bedrooms	0.938375	-62.52828 - 55.275273	0.901521	-61.74676 - 56.874309

population	0.000103	-48.16117 - (-20.060287)	0.000183	-46.18240 - (-17.936466)
households	0.001550	62.54078 - 194.193909	0.001643	57.19351 - 188.633153
median_income	< 2e-16	37327.24965 - 44056.507583	< 2e-16	37168.04529 - 44023.337295
ocean_proximity1&3	< 2e-16	54237.55127 - 78806.224737	< 2e-16	60728.55456 - 85556.466023
ocean _proximity4	< 2e-16	82586.95996 - 115980.745298	< 2e-16	90536.11415 - 124875.101248
county3	-	-	0.334959	47305.18144 - 14016.698883
county1,4,5,6	-	-	0.038414	6247.38833 - 46349.100542

Setelah dilakukan kalkulasi ulang, untuk variabel kategorik “county” masih memiliki faktor yang tidak signifikan, sehingga faktor “county2” dan “county3” akan digabungkan, sehingga hasil dari uji signifikansi variabel menghasilkan nilai p-value dan credible interval:

Variabel	Tanpa County		Dengan County	
	OLS	Bayes	OLS	Bayes
housing_median _age	2.57e-05	719.26293 - 1650.511507	0.000916	483.65189 - 1428.130089
total_rooms	0.672185	-10.69278 - 5.994393	0.582850	-11.45799 - 5.665622
total_bedrooms	0.938375	-62.52828 - 55.275273	0.956144	-59.45503 - 56.279157
population	0.000103	-48.16117 - (-20.060287)	0.000216	-46.94662 - (-17.600586)
households	0.001550	62.54078 - 194.193909	0.001748	60.36244 - 190.899565
median_income	< 2e-16	37327.24965 - 44056.507583	< 2e-16	37342.35131 - 44239.072406
ocean_proximity1&3	< 2e-16	54237.55127 - 78806.224737	< 2e-16	60465.14344 - 85514.454079

ocean _proximity4	< 2e-16	82586.95996 - 115980.745298	< 2e-16	88931.69514 - 121986.738009
county2&3	-	-	0.001097	-48690.42876 - (-16173.919535)

Dengan menggunakan backward elimination method untuk mengevaluasi variabel independen yang tidak berpengaruh secara signifikan terhadap variabel dependen, variabel pertama yang dieliminasi adalah variabel “total_bedrooms” karena memiliki nilai P-value paling tinggi. Setelah backward elimination method yang pertama, dilakukan kembali uji signifikansi:

Variabel	Tanpa County		Dengan County	
	OLS	Bayes	OLS	Bayes
housing_median _age	2.35e-05	703.365798 - 1642.328285	0.00087	468.819693 - 1438.357606
total_rooms	0.48883	-8.227105 - 3.543235	0.39493	-8.965355 - 3.113507
population	1.57e-05	45.959502 - (-21.237328)	3.76e-05	-44.482803 - (-19.072793)
households	4.72e-06	81.088128 - 168.180658	5.39e-06	78.937013 - 168.776428
median_income	< 2e-16	37811.111501 - 43757.219302	< 2e-16	37993.651583 - 43743.888400
ocean_proximit y1&3	< 2e-16	54419.660571 - 78471.909976	< 2e-16	60744.522686 - 85422.464362
ocean _proximity4	< 2e-16	82699.892674 - 116328.666253	< 2e-16	88355.236287 - 122246.333653
county2&3	-	-	0.00108	-49149.209497- (-16407.639845)

Karena masih ada variabel independen yang tidak berpengaruh secara signifikan terhadap variabel dependen yaitu “total_rooms”, maka dilakukan eliminasi kedua untuk variabel “total_rooms”. Sehingga hasil uji signifikansi untuk eliminasi kedua menjadi:

Variabel	Tanpa County		Dengan County	
	OLS	Bayes	OLS	Bayes
housing_median_age	1.41e-05	765.73872 - 1679.57027	0.00053	515.30437 - 1456.37264
population	6.18e-06	-47.04821- (-22.27013)	1.35e-05	-45.10364 - (-20.57769)
households	1.46e-07	79.66084 - 148.79683	3.49e-07	75.06816 - 143.17612
median_income	< 2e-16	37665.72668 - 42700.86454	< 2e-16	37592.71266 - 42710.37327
ocean_proximity1&3	< 2e-16	56241.16999 - 79317.12169	< 2e-16	63174.17470 - 86742.74248
ocean_proximity4	< 2e-16	83863.05371 - 117057.60755	< 2e-16	89819.36025 - 123261.29835
county2&3	-	-	0.00122	-48411.77304 - (-15601.85196)

Karena hasil dari uji signifikansi parameter sudah menunjukkan bahwa semua variabel independen berpengaruh secara signifikan terhadap variabel dependen, maka keempat model bisa dianggap merupakan model yang baik. Keempat model yang didapatkan dijabarkan sebagai berikut:

Model OLS tanpa county

```
model_ols3$coefficients
     (Intercept) housing_median_age         population         households      median_income   ocean_proximity4 ocean_proximity1&3
     -42764.4322          1209.2900           -34.5838           114.1889         40168.9858         100714.9037          68065.6983
```

Gambar 3: Hasil dari model OLS tanpa county

$$\text{Median\_house\_value} = -42764.4322 + 1209.29\text{housing\_median\_age} - 34.5838\text{population} + 114.1889\text{households} + 40168.9858\text{median\_income} + 100714.9037\text{ocean\_proximity4} + 68065.6983\text{ocean\_proximity1\&3}$$

Dengan nilai

- MSE: 4430303930
- RMSE: 66560.53
- MAPE: 0.2776473

Model bayes tanpa county

```
model_bayes3$coefficients
     (Intercept) housing_median_age         population         households      median_income   ocean_proximity4 ocean_proximity1&3
     -42648.47493          1217.72788           -34.51848           114.01065         40161.82995         100422.44893          68089.75909
```

Gambar 4: Hasil dari model bayes tanpa county

Median_house_value = -42648.47493 +1217.72788housing_median_age - 34.51848population + 114.01065households + 40161.82995median_income + 100422.44893ocean_proximity4 + 68089.75909ocean_proximity1&3

Dengan nilai:

- MSE: 4430416557
- RMSE: 66561.37
- MAPE: 0.2784858

Model OLS dengan county

```
> model_ols_county3_2$coefficients
      (Intercept) housing_median_age         population         households      median_income   ocean_proximity4 ocean_proximity1&3
      -37175.57846          982.78365          -33.02444          109.73343        40144.00614       106849.07296        74933.17126
        county2&3
      -32187.25373
```

Gambar 5: Hasil dari model OLS dengan county

Median_house_value =-37175.57846 + 982.78365housing_median_age - 33.02444population + 109.73343households + 40144.00614median_income + 106849.07296ocean_proximity4 + 74933..17126ocean_proximity1&3 -32187.25373county2&3

Dengan nilai:

- MSE: 4336482857
- RMSE: 65851.98
- MAPE: 0.2748224

Model bayes dengan county

```
> model_bayes_county3_2$coefficients
      (Intercept) housing_median_age         population         households      median_income   ocean_proximity4 ocean_proximity1&3
      -37493.61583          990.18837          -32.47699          109.01365        40145.83766       106956.74972        74794.20321
        county2&3
      -31957.14438
```

Gambar 6: Hasil dari model bayes dengan county

Median_house_value = -37493.61583 + 990.18837housing_median_age - 32.47699population + 109.01365households + 40145.83766median_income +106956.74972ocean_proximity4 + 74794.20321ocean_proximity1&3 - 31957.14438county2&3

Dengan nilai:

- MSE: 4336678512
- RMSE: 65853.46
- MAPE: 0.2753984

Kesimpulan

Keempat model tersebut setelah dilakukan uji signifikansi dihasilkan model dengan menggunakan variabel “housing_median_age”, “population”, “households”, “median_indome”, “ocean_proximity”, dan “county” (untuk dua model menggunakan county) dan akan dibandingkan berdasarkan nilai MSE, RMSE, dan MAPE yang didapatkan.

	Tanpa County		Dengan County	
	OLS	Bayes	OLS	Bayes
MSE	4430303930	4430416557	4336482857	4336678512
RMSE	66560.53	66561.37	65851.98	65853.46
MAPE	0.2776473	0.2784858	0.2748224	0.2753984

Semakin kecil nilai MSE, RMSE, dan MAPE, menandakan bahwa model lebih baik dalam melakukan prediksi. Berdasarkan nilai *output* diatas, model regresi tanpa variabel “county” serta model dengan variabel “county” paling baik dengan menggunakan penduga parameter OLS. Selain itu, model yang paling baik untuk melakukan prediksi harga rumah berdasarkan beberapa aspek perumahan adalah model OLS dengan menggunakan variabel “county”.

Model:

Median_house_value =-37175.57846 + 982.78365housing_median_age - 33.02444population + 109.73343households + 40144.00614median_income + 106849.07296ocean_proximity4 + 74933..17126ocean_proximity1&3 -32187.25373county2&3

Dengan demikian, paper ini bisa digunakan sebagai referensi untuk memprediksi harga rumah, juga diantisipasi sebagai titik awal penelitian yang terkait di masa mendatang, dan juga bisa digunakan sebagai acuan rencana marketing bagi broker dan pemain di bidang *real estate* atau perumahan.

Referensi

- Fernandez-Duran, L., Llorca, A., Ruiz, N.,Valero, S., Botti, V. (2011). The impact of location on housing prices: applying the Artificial Neural Network Model as an analytical tool. *ERSA conference papers*.
- Simplilearn. What Is Backward Elimination Technique In Machine Learning?. [Online]; 2022 [diakses 2023 Januari 19]. Tersedia dari:

https://www.simplilearn.com/what-is-backward-elimination-technique-in-machine-learning-article

- Rangan Gupta, Stephen M. Miller. The Time-Series Properties of House Prices: A Case Study of the Southern California Market. [Online]; 2015 [diakses 2023 Januari 23]. Tersedia dari: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1352768

- Zixu Wu, Prediction of California House Price Based on Multiple Linear Regression. Academic Journal of Engineering and Technology Science. Academic Journal of Engineering and Technology Science. ISSN 2616-5767 Vol.3, Issue 7: 11-15, DOI: 10.25236/AJETS.2020.030702

Code

Mengimport dataset ke dalam R

```
> getwd()
[1] "C:/Users/62821/Documents/Binus Semester 5/Bayesian Data Analysis/Bayesian"
> housing <- read.csv(file = "cal_housing.csv")
> head(housing)
  longitude latitude housing_median_age total_rooms total_bedrooms population households median_income median_house_value
1   -122.23    37.88                41         880            129        322        126        8.3252             452600
2   -122.22    37.86                21        7099           1106       2401       1138        8.3014             358500
3   -122.24    37.85                52        1467            190        496        177        7.2574             352100
4   -122.25    37.85                52        1274            235        558        219        5.6431             341300
5   -122.25    37.85                52        1627            280        565        259        3.8462             342200
6   -122.25    37.85                52         919            213        413        193        4.0368             269700
  ocean_proximity
1        NEAR BAY
2        NEAR BAY
3        NEAR BAY
4        NEAR BAY
5        NEAR BAY
6        NEAR BAY
```

Mengambil 500 observasi secara acak dan menetapkan seed agar sampel tidak berubah setiap kali dijalankan

```
> set.seed(42)
> dataset <- housing[sample(nrow(housing), 500), ]
> head(dataset)
      longitude latitude housing_median_age total_rooms total_bedrooms population households median_income median_house_value
18882   -122.25    38.10                52         248             86        173         69        2.3000              109400
19341   -122.87    38.61                23        2676            521       1456        500        3.7361              173700
5906    -118.42    34.29                34        1489            326       1389        313        3.4821              160300
17138   -122.16    37.46                32        2663            661       1403        733        4.2667              410200
13244   -117.65    34.14                 9        3877            490       1815        490        8.4839              406700
10712   -117.70    33.60                25        1321            295        396        278        3.1131               77100
      ocean_proximity
18882        NEAR BAY
19341       <1H OCEAN
5906        <1H OCEAN
17138        NEAR BAY
13244          INLAND
10712       <1H OCEAN
> summary(dataset)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms     population       households
 Min.   :-124.2   Min.   :32.57   Min.   : 2.00      Min.   :   24   Min.   :   6.0   Min.   :   23.0   Min.   :   5.0
 1st Qu.:-122.0   1st Qu.:33.97   1st Qu.:18.00      1st Qu.: 1442   1st Qu.: 285.0   1st Qu.:  771.8   1st Qu.: 271.0
 Median :-119.0   Median :34.66   Median :28.00      Median : 2068   Median : 434.0   Median : 1148.0   Median : 409.0
 Mean   :-119.7   Mean   :35.79   Mean   :27.32      Mean   : 2643   Mean   : 535.1   Mean   : 1401.2   Mean   : 493.3
 3rd Qu.:-118.1   3rd Qu.:37.77   3rd Qu.:35.00      3rd Qu.: 3100   3rd Qu.: 643.0   3rd Qu.: 1757.0   3rd Qu.: 606.2
 Max.   :-114.6   Max.   :41.54   Max.   :52.00      Max.   :20908   Max.   :4183.0   Max.   :10988.0   Max.   :3510.0
                                                                    NA's   :3
 median_income     median_house_value   ocean_proximity
 Min.   : 0.4999   Min.   : 25000     <1H OCEAN :212
 1st Qu.: 2.4435   1st Qu.:115600     INLAND    :163
 Median : 3.5444   Median :174000     ISLAND    :  0
 Mean   : 3.8857   Mean   :204416     NEAR BAY  : 63
 3rd Qu.: 4.7067   3rd Qu.:264450     NEAR OCEAN: 62
 Max.   :15.0001   Max.   :500001
```

Menghitung missing-value dan melakukan drop kepada observasi yang memiliki missing value.

```
> sum(is.na(dataset))
[1] 3
> dataset <- na.omit(dataset)
```

Mengubah variabel ocean proximity menjadi faktor dengan label numerik

```
> #mengubah ocean proximity menjadi sebuah faktor
> dataset$ocean_proximity = factor(dataset$ocean_proximity,levels =
+                                    c('<1H OCEAN', 'INLAND',
+                                      'NEAR BAY', 'NEAR OCEAN'),
+                                  labels = c(1, 2, 3, 4))
> summary(dataset)
   longitude          latitude     housing_median_age  total_rooms    total_bedrooms     population      households
 Min.   :-124.2   Min.   :32.57   Min.   : 2.00      Min.   :   24   Min.   :   6.0   Min.   :   23   Min.   :   5.0
 1st Qu.:-122.0   1st Qu.:33.97   1st Qu.:18.00      1st Qu.: 1435   1st Qu.: 285.0   1st Qu.:  768   1st Qu.: 268.0
 Median :-119.0   Median :34.69   Median :28.00      Median : 2066   Median : 434.0   Median : 1142   Median : 407.0
 Mean   :-119.7   Mean   :35.80   Mean   :27.28      Mean   : 2641   Mean   : 535.1   Mean   : 1397   Mean   : 492.4
 3rd Qu.:-118.0   3rd Qu.:37.78   3rd Qu.:35.00      3rd Qu.: 3065   3rd Qu.: 643.0   3rd Qu.: 1744   3rd Qu.: 606.0
 Max.   :-114.6   Max.   :41.54   Max.   :52.00      Max.   :20908   Max.   :4183.0   Max.   :10988   Max.   :3510.0
 median_income     median_house_value ocean_proximity
 Min.   : 0.4999   Min.   : 25000     1:209
 1st Qu.: 2.4464   1st Qu.:114100     2:163
 Median : 3.5500   Median :173900     3: 63
 Mean   : 3.8837   Mean   :203942     4: 62
 3rd Qu.: 4.7062   3rd Qu.:264200
 Max.   :15.0001   Max.   :500001
> datas <- dataset
```

Note: setiap model yang dibuat diuji signifikansi parameternya

Membuat model OLS dan bayesian

```
> model_ols <- lm(median_house_value~housing_median_age +
+                   total_rooms+ total_bedrooms + population + households +
+                   median_income + ocean_proximity, data = datas)
> summary(model_ols)

Call:
lm(formula = median_house_value ~ housing_median_age + total_rooms +
    total_bedrooms + population + households + median_income +
    ocean_proximity, data = datas)

Residuals:
    Min      1Q  Median      3Q     Max
-198008  -40534   -8897   27222  380113

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       22367.290  15385.447   1.454 0.146647
housing_median_age   1160.508    283.838   4.089 5.08e-05 ***
total_rooms            -2.505      5.306  -0.472 0.636991
total_bedrooms         -1.097     36.089  -0.030 0.975752
population            -33.191      8.803  -3.771 0.000183 ***
households            125.708     40.588   3.097 0.002066 **
median_income       40759.629   2093.261  19.472  < 2e-16 ***
ocean_proximity2   -65648.272   7855.793  -8.357 6.79e-16 ***
ocean_proximity3     4633.852  10008.880   0.463 0.643590
ocean_proximity4    34061.588   9878.136   3.448 0.000613 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67190 on 487 degrees of freedom
Multiple R-squared:  0.6808,    Adjusted R-squared:  0.6749
F-statistic: 115.4 on 9 and 487 DF,  p-value: < 2.2e-16


> library(rstanarm)
Loading required package: Rcpp
rstanarm (Version 2.18.2, packaged: 2018-11-08 22:19:38 UTC)
- Do not expect the default priors to remain the same in future rstanarm versions.
Thus, R scripts should specify priors explicitly, even if they are just the defaults.
- For execution on a local, multicore CPU with excess RAM we recommend calling
options(mc.cores = parallel::detectCores())
- Plotting theme set to bayesplot::theme_default().
> model_bayes <- stan_glm(median_house_value~housing_median_age+
+                           total_rooms+ total_bedrooms + population + households +
+                           median_income + ocean_proximity, data = datas, prior =
+                           normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 7.877 seconds (Warm-up)
Chain 1:                1.195 seconds (Sampling)
Chain 1:                9.072 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
```

```
> summary(model_bayes)

Model Info:

 function:     stan_glm
 family:       gaussian [identity]
 formula:      median_house_value ~ housing_median_age + total_rooms + total_bedrooms +
           population + households + median_income + ocean_proximity
 algorithm:    sampling
 priors:       see help('prior_summary')
 sample:       4000 (posterior sample size)
 observations: 497
 predictors:   10

Estimates:
                  mean     sd       2.5%     25%     50%     75%     97.5%
(Intercept)      21767.2  15483.1 -8393.2  11092.3  21925.6  32292.8  51961.6
housing_median_age  1166.6   286.5   603.1    974.2   1160.5   1363.8   1720.3
total_rooms         -2.6     5.4   -13.1     -6.4     -2.6      1.0      8.0
total_bedrooms      -0.3    36.9   -73.1    -25.0     -0.2     24.7     70.2
population         -33.0     8.9   -50.2    -38.9    -33.1    -26.8    -16.0
households         124.9    40.6    45.3     97.7    124.7    152.3    204.3
median_income    40823.3  2128.2  36612.6  39369.9  40795.7  42272.6  45023.4
ocean_proximity2 -65388.7  7845.9 -80745.4 -70612.3 -65316.4 -60220.6 -49809.3
ocean_proximity3   4842.7 10186.2 -14702.6  -2163.1   5043.8  11745.2  24343.9
ocean_proximity4  34261.4  9508.1  15153.9  27922.0  34261.2  40600.3  52938.2
sigma             67314.8  2117.5  63239.8  65859.8  67288.9  68721.2  71591.0
mean_PPD         203851.5  4331.9 195610.7 200942.5 203786.3 206833.1 212389.2
log-posterior      -6254.5     2.4  -6260.0  -6255.9  -6254.2  -6252.7  -6250.9

Diagnostics:
                  mcse  Rhat  n_eff
(Intercept)      307.3   1.0 2538
housing_median_age   4.7   1.0 3769
total_rooms          0.1   1.0 1698
total_bedrooms       1.0   1.0 1486
population           0.2   1.0 1897
households           1.0   1.0 1636
median_income       47.7   1.0 1994
ocean_proximity2   147.7   1.0 2823
ocean_proximity3   186.4   1.0 2986
ocean_proximity4   174.6   1.0 2965
sigma               34.4   1.0 3788
mean_PPD            70.6   1.0 3768
log-posterior        0.1   1.0 1760
```

```
> #variabel independen bisa dibilang berpengaruh secara signifikan terhadap variabel
> #dependen jika dalam posterior/credible intervalnya tidak mencakup nilai 0
> posterior_interval(model_bayes)
                              5%           95%
(Intercept)          -4204.11874  47060.850297
housing_median_age     690.53134   1636.096984
total_rooms            -11.43033      6.364199
total_bedrooms         -62.38696     59.855762
population             -47.46991    -18.246358
households              57.79552    190.918566
median_income        37386.17610  44412.315017
ocean_proximity2   -78380.56271 -52459.906679
ocean_proximity3   -11533.43989  21563.383115
ocean_proximity4    18523.70422  49626.867254
sigma               63904.62272  70927.334367
```

Menggabungkan faktor dari variabel “ocean_proximity” yaitu ocean_proximity 1 dan 3.

```
> datas$ocean_proximity = as.character(datas$ocean_proximity)
> datas$ocean_proximity = as.array(datas$ocean_proximity)
>
> summary(datas)
   longitude          latitude     housing_median_age  total_rooms     total_bedrooms
 Min.   :-124.2   Min.   :32.57   Min.   : 2.00       Min.   :   24   Min.   :   6.0
 1st Qu.:-122.0   1st Qu.:33.97   1st Qu.:18.00       1st Qu.: 1435   1st Qu.: 285.0
 Median :-119.0   Median :34.69   Median :28.00       Median : 2066   Median : 434.0
 Mean   :-119.7   Mean   :35.80   Mean   :27.28       Mean   : 2641   Mean   : 535.1
 3rd Qu.:-118.0   3rd Qu.:37.78   3rd Qu.:35.00       3rd Qu.: 3065   3rd Qu.: 643.0
 Max.   :-114.6   Max.   :41.54   Max.   :52.00       Max.   :20908   Max.   :4183.0
   population      households     median_income     median_house_value
 Min.   :   23   Min.   :   5.0   Min.   : 0.4999   Min.   : 25000
 1st Qu.:  768   1st Qu.: 268.0   1st Qu.: 2.4464   1st Qu.:114100
 Median : 1142   Median : 407.0   Median : 3.5500   Median :173900
 Mean   : 1397   Mean   : 492.4   Mean   : 3.8837   Mean   :203942
 3rd Qu.: 1744   3rd Qu.: 606.0   3rd Qu.: 4.7062   3rd Qu.:264200
 Max.   :10988   Max.   :3510.0   Max.   :15.0001   Max.   :500001
 ocean_proximity
 Length:497
 Class :array
 Mode  :character
```

```
> View(datas)
> c2 <- c('2','4')
> datas$ocean_proximity[!(datas$ocean_proximity %in% c2)] <- 'others'
> datas$ocean_proximity = factor(datas$ocean_proximity,levels =
+                                  c('2', '4', 'others'),
+                                labels = c(2, 4, '1&3'))
`
```

Membuat model baru

Model OLS

```
> summary(model_ols1)

Call:
lm(formula = median_house_value ~ housing_median_age + total_rooms +
    total_bedrooms + population + households + median_income +
    ocean_proximity, data = datas)

Residuals:
    Min      1Q  Median      3Q     Max
-197836  -39999   -8691   26902  383565

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -43745.574  12892.555  -3.393 0.000747 ***
housing_median_age   1184.677    278.772   4.250 2.57e-05 ***
total_rooms            -2.231      5.268  -0.423 0.672185
total_bedrooms         -2.775     35.877  -0.077 0.938375
population            -33.907      8.659  -3.916 0.000103 ***
households            128.068     40.234   3.183 0.001550 **
median_income       40716.528   2089.506  19.486  < 2e-16 ***
ocean_proximity4    99702.668  10230.331   9.746  < 2e-16 ***
ocean_proximity1&3  66693.280   7518.480   8.871  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67140 on 488 degrees of freedom
Multiple R-squared:  0.6806,    Adjusted R-squared:  0.6754
F-statistic:   130 on 8 and 488 DF,  p-value: < 2.2e-16
```

Melakukan backward elimination untuk model OLS biasa

```
> model_ols2 <- lm(median_house_value~.-total_bedrooms-longitude
+                  -latitude, data = datas)
> summary(model_ols2)

Call:
lm(formula = median_house_value ~ . - total_bedrooms - longitude -
    latitude, data = datas)

Residuals:
    Min      1Q  Median      3Q     Max
-197927  -40014   -8711   26888  383481

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -44045.482  12283.249  -3.586  0.00037 ***
housing_median_age   1186.206    277.788   4.270 2.35e-05 ***
total_rooms            -2.525      3.645  -0.693  0.48883
population            -33.602      7.703  -4.362 1.57e-05 ***
households            125.774     27.173   4.629 4.72e-06 ***
median_income       40797.837   1804.017  22.615  < 2e-16 ***
ocean_proximity4    99658.974  10204.337   9.766  < 2e-16 ***
ocean_proximity1&3  66588.335   7387.536   9.014  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67070 on 489 degrees of freedom
Multiple R-squared:  0.6806,    Adjusted R-squared:  0.6761
F-statistic: 148.9 on 7 and 489 DF,  p-value: < 2.2e-16
```

```
> model_ols3 <- lm(median_house_value~.-total_bedrooms-longitude
+                  -latitude-total_rooms, data = datas)
> summary(model_ols3)

Call:
lm(formula = median_house_value ~ . - total_bedrooms - longitude -
    latitude - total_rooms, data = datas)

Residuals:
    Min      1Q  Median      3Q     Max
-197515  -41766   -9035   27071  382941

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -42764.432  12136.779  -3.524 0.000466 ***
housing_median_age   1209.290    275.635   4.387 1.41e-05 ***
population            -34.584      7.568  -4.570 6.18e-06 ***
households            114.189     21.405   5.335 1.46e-07 ***
median_income       40168.986   1558.114  25.781  < 2e-16 ***
ocean_proximity4   100714.904  10084.476   9.987  < 2e-16 ***
ocean_proximity1&3  68065.698   7069.212   9.628  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67030 on 490 degrees of freedom
Multiple R-squared:  0.6803,    Adjusted R-squared:  0.6764
F-statistic: 173.8 on 6 and 490 DF,  p-value: < 2.2e-16
```

Model Bayes

```
> model_bayes1 <- stan_glm(median_house_value~housing_median_age+
+                          total_rooms+ total_bedrooms + population + households +
+                          median_income + ocean_proximity, data = datas, prior =
+                          normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 7.473 seconds (Warm-up)
Chain 1:                1.454 seconds (Sampling)
Chain 1:                8.927 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
```

```
> summary(model_bayes1)

Model Info:

 function:     stan_glm
 family:       gaussian [identity]
 formula:      median_house_value ~ housing_median_age + total_rooms + total_bedro
               population + households + median_income + ocean_proximity
 algorithm:    sampling
 priors:       see help('prior_summary')
 sample:       4000 (posterior sample size)
 observations: 497
 predictors:   9

Estimates:
                      mean     sd       2.5%     25%      50%      75%      97.5%
(Intercept)        -43276.5  12693.9 -68259.4 -52096.3 -43232.9 -34578.5 -18888.0
housing_median_age   1180.7    280.5    635.1    985.4   1178.7   1374.6   1723.4
total_rooms            -2.2      5.2    -12.5     -5.7     -2.3      1.5      7.6
total_bedrooms         -2.8     35.4    -75.3    -25.8     -1.7     21.1     64.8
population            -34.0      8.5    -51.1    -39.7    -34.2    -28.2    -16.9
households            128.1     39.6     49.5     101.5    128.1    154.7    206.5
median_income       40707.5   2051.9  36775.4  39315.2  40688.1  42110.1  44635.2
ocean_proximity4    99182.9  10179.8  79629.8  92140.5  99112.3 106119.7 118928.1
ocean_proximity1&3  66501.2   7398.9  52089.3  61524.2  66411.6  71493.5  81061.8
sigma               67247.4   2124.3  63212.7  65803.3  67160.7  68637.0  71582.8
mean_PPD           203911.0   4297.6 195596.4 200978.7 203929.3 206894.0 212268.4
log-posterior       -6253.2      2.2  -6258.4  -6254.4  -6252.8  -6251.5  -6249.8

Diagnostics:
                    mcse  Rhat  n_eff
(Intercept)        222.7   1.0 3248
housing_median_age   4.7   1.0 3625
total_rooms          0.1   1.0 1899
total_bedrooms       0.9   1.0 1576
population           0.2   1.0 2059
households           0.9   1.0 1750
median_income       44.4   1.0 2133
ocean_proximity4   192.0   1.0 2811
ocean_proximity1&3 154.1   1.0 2305
sigma               35.7   1.0 3536
mean_PPD            68.9   1.0 3891
log-posterior        0.1   1.0 1902
```

```
> posterior_interval(model_bayes1)
                                    5%             95%
(Intercept)         -64113.45161 -22746.923612
housing_median_age     719.26293   1650.511507
total_rooms            -10.69278      5.994393
total_bedrooms         -62.52828     55.275273
population             -48.16117    -20.060287
households              62.54078    194.193909
median_income        37327.24965  44056.507583
ocean_proximity4     82586.95996 115980.745298
ocean_proximity1&3   54237.55127  78806.224737
sigma                63848.35418  70826.775545
```

Melakukan backward elimination untuk model bayes biasa

```
> model_bayes2 <- stan_glm(median_house_value~.-total_bedrooms-longitude
+                          -latitude, data = datas, prior =
+                            normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 10.783 seconds (Warm-up)
Chain 1:                0.815 seconds (Sampling)
Chain 1:                11.598 seconds (Total)
Chain 1:
```

```
> posterior_interval(model_bayes2)
                              5%           95%
(Intercept)        -64072.103573 -22966.096639
housing_median_age    703.365798   1642.328285
total_rooms            -8.227105      3.543235
population            -45.959502    -21.237328
households             81.088128    168.180658
median_income       37811.111501  43757.219302
ocean_proximity4    82699.892674 116328.666253
ocean_proximity1&3  54419.660571  78471.909976
sigma               63718.278470  70843.513921
```

```
> model_bayes3 <- stan_glm(median_house_value~.-total_bedrooms-longitude
+                          -latitude-total_rooms, data = datas, prior =
+                            normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 10.238 seconds (Warm-up)
Chain 1:                0.595 seconds (Sampling)
Chain 1:                10.833 seconds (Total)
Chain 1:
```

```
> posterior_interval(model_bayes3)
                              5%           95%
(Intercept)        -63431.22752 -23362.84039
housing_median_age    765.73872   1679.57027
population            -47.04821    -22.27013
households             79.66084    148.79683
median_income       37665.72668  42700.86454
ocean_proximity4    83863.05371 117057.60755
ocean_proximity1&3  56241.16999  79317.12169
sigma               63791.73609  70805.26152
```

Mengkonversi variabel longitude dan latitude menjadi county

```
> library(maps)
> startm <- Sys.time()
> county<-map.where(database="county",
+                   dataset$longitude, dataset$latitude)
> endm <- Sys.time()
> county1 <- as.array(county)
> county2 <- as.data.frame(county)
> dataset$county <- county1
> summary(dataset)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms
 Min.   :-124.2   Min.   :32.57   Min.   : 2.00      Min.   :   24   Min.   :   6.0
 1st Qu.:-122.0   1st Qu.:33.97   1st Qu.:18.00      1st Qu.: 1435   1st Qu.: 285.0
 Median :-119.0   Median :34.69   Median :28.00      Median : 2066   Median : 434.0
 Mean   :-119.7   Mean   :35.80   Mean   :27.28      Mean   : 2641   Mean   : 535.1
 3rd Qu.:-118.0   3rd Qu.:37.78   3rd Qu.:35.00      3rd Qu.: 3065   3rd Qu.: 643.0
 Max.   :-114.6   Max.   :41.54   Max.   :52.00      Max.   :20908   Max.   :4183.0
   population      households     median_income     median_house_value ocean_proximity
 Min.   :   23   Min.   :   5.0   Min.   : 0.4999   Min.   : 25000     1:209
 1st Qu.:  768   1st Qu.: 268.0   1st Qu.: 2.4464   1st Qu.:114100     2:163
 Median : 1142   Median : 407.0   Median : 3.5500   Median :173900     3: 63
 Mean   : 1397   Mean   : 492.4   Mean   : 3.8837   Mean   :203942     4: 62
 3rd Qu.: 1744   3rd Qu.: 606.0   3rd Qu.: 4.7062   3rd Qu.:264200
 Max.   :10988   Max.   :3510.0   Max.   :15.0001   Max.   :500001
    county
 Length:497
 Class :array
 Mode  :character
```

Faktor county yang diambil hanya 5 faktor paling banyak dan sisanya dikategorikan menjadi others, lalu dijadikan faktor dengan label numerik.

```
> summary(county2)
                       county
 california,los angeles   :122
 california,orange        : 36
 california,san diego     : 23
 california,alameda       : 22
 california,san bernardino: 19
 (Other)                  :235
 NA's                     : 40
> c <- c('california,los angeles','california,orange', 'california,san diego',
+        'california,alameda', 'california,san bernardino')
> dataset$county[!(dataset$county %in% c)] <- 'others'
>
> dataset$county = factor(dataset$county,levels =
+                           c('california,los angeles','california,orange', 'california,san
 diego',
+                             'california,alameda', 'california,san bernardino', 'others'),
+                         labels = c(1, 2, 3, 4, 5, 6))
> summary(dataset)
```

Membuat model OLS dan Bayesian dengan menggunakan tambahan variabel county

```
> model_ols_county <- lm(median_house_value~housing_median_age+ county+
+                    total_rooms+ total_bedrooms + population + households +
+                    median_income + ocean_proximity, data = dataset)
> summary(model_ols_county)

Call:
lm(formula = median_house_value ~ housing_median_age + county +
    total_rooms + total_bedrooms + population + households +
    median_income + ocean_proximity, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-206124  -38330   -9166   28960  386581

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         38206.823  17139.096   2.229  0.02626 *
housing_median_age    934.751    295.455   3.164  0.00166 **
county2            -27395.158  13126.051  -2.087  0.03741 *
county3            -46387.813  16449.436  -2.820  0.00500 **
county4             -8059.451  18625.023  -0.433  0.66541
county5             -9127.673  18595.425  -0.491  0.62375
county6             -4052.445   9586.054  -0.423  0.67267
total_rooms            -1.547      5.465  -0.283  0.77730
total_bedrooms         -7.351     36.587  -0.201  0.84086
population            -33.155      8.906  -3.723  0.00022 ***
households            127.289     40.560   3.138  0.00180 **
median_income       40529.803   2101.893  19.283  < 2e-16 ***
ocean_proximity2   -71058.122   9338.734  -7.609 1.46e-13 ***
ocean_proximity3     2841.954  12454.541   0.228  0.81960
ocean_proximity4    36294.870  11080.885   3.275  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66730 on 482 degrees of freedom
Multiple R-squared:  0.6884,    Adjusted R-squared:  0.6793
F-statistic: 76.06 on 14 and 482 DF,  p-value: < 2.2e-16
```

```
> model_bayes_county <- stan_glm(median_house_value~housing_median_age+ county+
+                            total_rooms+ total_bedrooms + population + households +
+                            median_income + ocean_proximity, data = dataset,
+                           prior = normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 9.177 seconds (Warm-up)
Chain 1:                1.478 seconds (Sampling)
Chain 1:                10.655 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
```

```
> summary(model_bayes_county)

Model Info:

 function:     stan_glm
 family:       gaussian [identity]
 formula:      median_house_value ~ housing_median_age + county + total_rooms +
           total_bedrooms + population + households + median_income +
           ocean_proximity
 algorithm:    sampling
 priors:       see help('prior_summary')
 sample:       4000 (posterior sample size)
 observations: 497
 predictors:   15

Estimates:
                    mean     sd       2.5%     25%      50%      75%      97.5%
(Intercept)         37939.8  17417.2   3939.4  25982.2  37743.9  49852.3  72562.2
housing_median_age    936.9    301.0    343.7    735.9    936.1   1141.8   1524.1
county2           -27294.0  13108.7 -52842.5 -35743.0 -27251.9 -18767.1  -1417.2
county3           -45888.0  16235.9 -77509.2 -56622.7 -45682.5 -35205.0 -13987.2
county4            -7949.0  18410.1 -44095.6 -20812.8  -7921.5   4552.9  27615.6
county5            -9053.9  18008.5 -43076.4 -21240.7  -9331.5   2885.7  27063.2
county6            -4004.1   9512.8 -22817.2 -10235.4  -3992.0   2170.5  14879.8
total_rooms           -1.6      5.5    -12.7     -5.3     -1.5      2.1      9.0
total_bedrooms        -5.9     36.9    -75.7    -31.1     -7.2     19.0     67.7
population           -32.9      8.8    -49.5    -38.8    -32.9    -27.0    -15.5
households           125.4     40.5     44.0     98.6    125.9    153.3    201.9
median_income      40571.8   2122.3  36506.0  39133.2  40584.1  41990.0  44708.6
ocean_proximity2  -71018.6   9465.4 -89913.8 -77404.3 -70928.3 -64713.9 -52716.2
ocean_proximity3    2888.9  12229.5 -20838.2  -5532.8   2873.0  11195.7  26301.6
ocean_proximity4   36140.2  10869.7  15051.3  28966.1  36049.5  43162.6  57570.8
sigma              66862.0   2134.1  62885.9  65438.3  66805.5  68281.7  71207.1
mean_PPD          203933.4   4282.9 195701.0 201090.5 203948.0 206742.5 212579.8
log-posterior       -6255.5      2.9  -6262.1  -6257.2  -6255.1  -6253.5  -6251.0

Diagnostics:
                  mcse  Rhat  n_eff
(Intercept)       375.1   1.0 2156
housing_median_age   5.2   1.0 3376
county2            231.6   1.0 3204
county3            316.1   1.0 2639
county4            363.2   1.0 2569
county5            346.6   1.0 2700
county6            230.4   1.0 1704
total_rooms          0.1   1.0 1764
total_bedrooms       0.9   1.0 1513
population           0.2   1.0 1792
households           1.0   1.0 1702
median_income       47.2   1.0 2024
```

```
> posterior_interval(model_bayes_county)
                                5%            95%
(Intercept)           9872.87716  66615.418667
housing_median_age     433.56571   1426.819239
county2             -48907.98465  -5288.479419
county3             -72231.79195 -19597.538245
county4             -38203.14771  22260.635087
county5             -38506.68226  20660.221188
county6             -19728.98199  11834.671766
total_rooms            -10.99380      7.384582
total_bedrooms         -65.17475     55.023618
population             -47.45576    -18.671912
households              56.11488    190.297701
median_income        37082.15330  44030.697240
ocean_proximity2   -86898.24904 -55704.214624
ocean_proximity3   -16800.91701  22901.684379
ocean_proximity4    18653.45682  54391.836547
sigma               63410.06766  70443.682846
```

Menggabungkan faktor county 1, 4, 5, 6 dan ocean_proximity 1 dan 3 serta membuat model baru.

```
> View(dataset)
> c2 <- c('2','4')
> dataset$ocean_proximity[!(dataset$ocean_proximity %in% c2)] <- 'others'
> dataset$ocean_proximity = factor(dataset$ocean_proximity,levels =
+                                    c('2', '4', 'others'),
+                                  labels = c(2, 4, '1&3'))
>
> dataset$county = as.character(dataset$county)
> dataset$county = as.array(dataset$county)
> summary(dataset)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms     population      households
 Min.   :-124.2   Min.   :32.57   Min.   : 2.00      Min.   :   24   Min.   :   6.0   Min.   :   23   Min.   :   5.0
 1st Qu.:-122.0   1st Qu.:33.97   1st Qu.:18.00      1st Qu.: 1435   1st Qu.: 285.0   1st Qu.:  768   1st Qu.: 268.0
 Median :-119.0   Median :34.69   Median :28.00      Median : 2066   Median : 434.0   Median : 1142   Median : 407.0
 Mean   :-119.7   Mean   :35.80   Mean   :27.28      Mean   : 2641   Mean   : 535.1   Mean   : 1397   Mean   : 492.4
 3rd Qu.:-118.0   3rd Qu.:37.78   3rd Qu.:35.00      3rd Qu.: 3065   3rd Qu.: 643.0   3rd Qu.: 1744   3rd Qu.: 606.0
 Max.   :-114.6   Max.   :41.54   Max.   :52.00      Max.   :20908   Max.   :4183.0   Max.   :10988   Max.   :3510.0
 median_income     median_house_value ocean_proximity    county
 Min.   : 0.4999   Min.   : 25000     2  :163         Length:497
 1st Qu.: 2.4464   1st Qu.:114100     4  : 62         Class :array
 Median : 3.5500   Median :173900     1&3:272         Mode  :character
 Mean   : 3.8837   Mean   :203942
 3rd Qu.: 4.7062   3rd Qu.:264200
 Max.   :15.0001   Max.   :500001
```

```
> View(dataset)
> c3 <- c('2', '3')
> dataset$county[!(dataset$county %in% c3)] <- 'others'
> dataset$county = factor(dataset$county,levels =
+                           c('2', '3', 'others'),
+                         labels = c(2, 3, '1,4,5,6'))
>
> model_ols_county2 <- lm(median_house_value~housing_median_age+ county+
+                         total_rooms+ total_bedrooms + population + households +
+                         median_income + ocean_proximity, data = dataset)
> summary(model_ols_county2)

Call:
lm(formula = median_house_value ~ housing_median_age + county +
    total_rooms + total_bedrooms + population + households +
    median_income + ocean_proximity, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-206373  -38406   -8689   27934  382286

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -63589.336  17073.690  -3.724 0.000219 ***
housing_median_age    955.195    285.110   3.350 0.000870 ***
county3           -17881.411  18527.456  -0.965 0.334959
county1,4,5,6      25547.484  12305.765   2.076 0.038414 *
total_rooms           -2.245      5.260  -0.427 0.669681
total_bedrooms        -4.410     35.617  -0.124 0.901521
population           -32.441      8.604  -3.771 0.000183 ***
households           126.182     39.856   3.166 0.001643 **
median_income      40622.545   2083.112  19.501  < 2e-16 ***
ocean_proximity4  107837.084  10534.273  10.237  < 2e-16 ***
ocean_proximity1&3  73200.047   7711.591   9.492  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66480 on 486 degrees of freedom
Multiple R-squared:  0.6882,    Adjusted R-squared:  0.6817
F-statistic: 107.2 on 10 and 486 DF,  p-value: < 2.2e-16
```

```
> model_bayes_county2 <- stan_glm(median_house_value~housing_median_age+ county+
+                                   total_rooms+ total_bedrooms + population + households +
+                                   median_income + ocean_proximity, data = dataset,
+                                 prior = normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 7.206 seconds (Warm-up)
Chain 1:                1.368 seconds (Sampling)
Chain 1:                8.574 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
```

```
> summary(model_bayes_county2)

Model Info:

 function:     stan_glm
 family:       gaussian [identity]
 formula:      median_house_value ~ housing_median_age + county + total_rooms +
          total_bedrooms + population + households + median_income +
          ocean_proximity
 algorithm:    sampling
 priors:       see help('prior_summary')
 sample:       4000 (posterior sample size)
 observations: 497
 predictors:   11

Estimates:
                      mean     sd       2.5%     25%      50%      75%      97.5%
(Intercept)        -63713.7  16798.3 -96902.1 -75296.5 -63742.3 -52141.9 -31459.6
housing_median_age    957.8    291.9    395.0    753.5    956.8   1154.2   1537.0
county3           -17148.7  18440.3 -52286.0 -29496.8 -17183.2  -4656.2  19576.1
county1,4,5,6      25822.9  12100.0   2969.9  17503.6  25769.7  33724.5  50928.6
total_rooms           -2.2      5.3    -12.8     -5.9     -2.1      1.5      7.7
total_bedrooms        -3.6     35.9    -71.8    -28.4     -3.7     20.7     66.7
population           -32.1      8.6    -48.8    -37.7    -32.2    -26.5    -14.5
households           124.3     39.9     44.6     98.0    124.5    150.7    201.3
median_income      40607.0   2085.0  36617.8  39173.5  40580.5  42027.5  44697.3
ocean_proximity4  107639.9  10492.2  87411.1 100816.0 107667.9 114522.1 128273.3
ocean_proximity1&3 73120.0   7704.7  58244.0  67846.9  72943.8  78338.1  88732.8
sigma              66625.1   2154.3  62608.4  65165.1  66533.6  68025.7  70969.5
mean_PPD          204077.4   4251.1 195783.4 201192.7 204060.8 207020.8 212494.2
log-posterior      -6250.2      2.5  -6255.8  -6251.6  -6249.8  -6248.3  -6246.3

Diagnostics:
                  mcse  Rhat  n_eff
(Intercept)       305.5   1.0 3024
housing_median_age  4.9   1.0 3497
county3           357.8   1.0 2656
county1,4,5,6     243.9   1.0 2460
total_rooms         0.1   1.0 1725
total_bedrooms      0.9   1.0 1554
population          0.2   1.0 1995
households          0.9   1.0 1767
median_income      46.4   1.0 2017
ocean_proximity4  192.6   1.0 2969
ocean_proximity1&3 159.1   1.0 2346
sigma              34.5   1.0 3892
mean_PPD           67.5   1.0 3964
log-posterior       0.1   1.0 1511
```

```
> posterior_interval(model_bayes_county2)
                                  5%           95%
(Intercept)        -90915.81514 -36053.553521
housing_median_age    479.05129   1444.667095
county3            -47305.18144  14016.698883
county1,4,5,6        6247.38833  46349.100542
total_rooms           -11.11485      6.360366
total_bedrooms        -61.74676     56.874309
population            -46.18240    -17.936466
households             57.19351    188.633153
median_income       37168.04529  44023.337295
ocean_proximity4    90536.11415 124875.101248
ocean_proximity1&3  60728.55456  85556.466023
sigma               63230.09240  70346.976220
```

Menggabungkan faktor county 2 dan 3 dan membentuk model baru.

```
> dataset$county = as.character(dataset$county)
> dataset$county = as.array(dataset$county)
> summary(dataset)
   longitude          latitude     housing_median_age  total_rooms    total_bedrooms     population      households
 Min.   :-124.2   Min.   :32.57   Min.   : 2.00      Min.   :   24   Min.   :   6.0   Min.   :   23   Min.   :   5.0
 1st Qu.:-122.0   1st Qu.:33.97   1st Qu.:18.00      1st Qu.: 1435   1st Qu.: 285.0   1st Qu.:  768   1st Qu.: 268.0
 Median :-119.0   Median :34.69   Median :28.00      Median : 2066   Median : 434.0   Median : 1142   Median : 407.0
 Mean   :-119.7   Mean   :35.80   Mean   :27.28      Mean   : 2641   Mean   : 535.1   Mean   : 1397   Mean   : 492.4
 3rd Qu.:-118.0   3rd Qu.:37.78   3rd Qu.:35.00      3rd Qu.: 3065   3rd Qu.: 643.0   3rd Qu.: 1744   3rd Qu.: 606.0
 Max.   :-114.6   Max.   :41.54   Max.   :52.00      Max.   :20908   Max.   :4183.0   Max.   :10988   Max.   :3510.0
 median_income     median_house_value ocean_proximity    county
 Min.   : 0.4999   Min.   : 25000     2  :163         Length:497
 1st Qu.: 2.4464   1st Qu.:114100     4  : 62         Class :array
 Median : 3.5500   Median :173900     1&3:272         Mode  :character
 Mean   : 3.8837   Mean   :203942
 3rd Qu.: 4.7062   3rd Qu.:264200
 Max.   :15.0001   Max.   :500001
> View(dataset)
> c4 <- c('1,4,5,6')
> dataset$county[!(dataset$county %in% c4)] <- 'others'
> dataset$county = factor(dataset$county,levels =
+                           c('1,4,5,6', 'others'),
+                         labels = c('1,4,5,6', '2&3'))
`
```

Model OLS

```
> model_ols_county3 <- lm(median_house_value~housing_median_age+ county+
+                           total_rooms+ total_bedrooms + population + households +
+                           median_income + ocean_proximity, data = dataset)
> summary(model_ols_county3)

Call:
lm(formula = median_house_value ~ housing_median_age + county +
    total_rooms + total_bedrooms + population + households +
    median_income + ocean_proximity, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-204850  -39869   -9546   27256  382000

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -38457.784  12866.378  -2.989 0.002940 **
housing_median_age    950.853    285.054   3.336 0.000916 ***
county2&3           -32575.056   9919.584  -3.284 0.001097 **
total_rooms            -2.869      5.220  -0.550 0.582850
total_bedrooms         -1.955     35.524  -0.055 0.956144
population            -32.024      8.592  -3.727 0.000216 ***
households            125.410     39.845   3.147 0.001748 **
median_income       40852.600   2069.284  19.742  < 2e-16 ***
ocean_proximity4   105667.262  10290.847  10.268  < 2e-16 ***
ocean_proximity1&3  73289.881   7710.486   9.505  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66480 on 487 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.6818
F-statistic: 119.1 on 9 and 487 DF,  p-value: < 2.2e-16
```

Backward elimination untuk model OLS dengan variabel county

```
> model_ols_county3_1 <- lm(median_house_value~.-longitude-latitude-
+                           total_bedrooms, data = dataset)
> summary(model_ols_county3_1)

Call:
lm(formula = median_house_value ~ . - longitude - latitude -
    total_bedrooms, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-204916  -40025   -9600   27098  381941

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -38668.375  12271.348  -3.151  0.00173 **
housing_median_age    951.902    284.126   3.350  0.00087 ***
total_rooms            -3.076      3.612  -0.851  0.39493
population            -31.810      7.646  -4.160 3.76e-05 ***
households            123.795     26.912   4.600 5.39e-06 ***
median_income       40909.879   1786.515  22.899  < 2e-16 ***
ocean_proximity4   105637.193  10265.823  10.290  < 2e-16 ***
ocean_proximity1&3  73216.750   7587.298   9.650  < 2e-16 ***
county2&3          -32578.895   9909.201  -3.288  0.00108 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66410 on 488 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.6824
F-statistic: 134.2 on 8 and 488 DF,  p-value: < 2.2e-16
```

```
> model_ols_county3_2 <- lm(median_house_value~.-longitude-latitude-
+                           total_bedrooms-total_rooms, data = dataset)
> summary(model_ols_county3_2)

Call:
lm(formula = median_house_value ~ . - longitude - latitude -
    total_bedrooms - total_rooms, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-204330  -39709   -9536   26663  381303

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -37175.58   12142.04  -3.062  0.00232 **
housing_median_age    982.78     281.72   3.488  0.00053 ***
population            -33.02       7.51  -4.397 1.35e-05 ***
households            109.73      21.24   5.166 3.49e-07 ***
median_income       40144.01    1543.12  26.015  < 2e-16 ***
ocean_proximity4   106849.07   10163.82  10.513  < 2e-16 ***
ocean_proximity1&3  74933.17    7312.54  10.247  < 2e-16 ***
county2&3          -32187.25    9895.74  -3.253  0.00122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66390 on 489 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6826
F-statistic: 153.4 on 7 and 489 DF,  p-value: < 2.2e-16
```

Model bayes

```
> model_bayes_county3 <- stan_glm(median_house_value~housing_median_age+ county+
+                                   total_rooms+ total_bedrooms + population + households
+                                   median_income + ocean_proximity, data = dataset,
+                                 prior = normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 8.021 seconds (Warm-up)
Chain 1:                1.132 seconds (Sampling)
Chain 1:                9.153 seconds (Total)
```

```
> summary(model_bayes_county3)

Model Info:

 function:     stan_glm
 family:       gaussian [identity]
 formula:      median_house_value ~ housing_median_age + county + total_rooms +
           total_bedrooms + population + households + median_income +
           ocean_proximity
 algorithm:    sampling
 priors:       see help('prior_summary')
 sample:       4000 (posterior sample size)
 observations: 497
 predictors:   10

Estimates:
                    mean     sd       2.5%     25%      50%      75%      97.5%
(Intercept)        -38472.8  12830.5 -63467.8 -47138.1 -38312.9 -29785.0 -13805.8
housing_median_age    957.0    287.9    393.4    768.5    951.9   1148.0   1518.4
county2&3          -32416.9   9919.9 -51941.2 -38879.4 -32431.5 -25874.3 -12827.9
total_rooms            -2.8      5.2    -13.0     -6.3     -2.7      0.7      7.4
total_bedrooms         -1.6     35.5    -70.1    -26.2     -1.1     23.2     67.6
population            -32.1      8.8    -49.8    -37.9    -32.0    -26.3    -15.0
households            125.0     39.6     47.7     98.6    124.7    151.2    203.1
median_income       40860.2   2071.4  36816.1  39482.2  40893.3  42238.2  44914.9
ocean_proximity4   105383.4  10060.9  85503.0  98637.8 105255.6 112060.6 124610.1
ocean_proximity1&3  73038.6   7668.3  58226.0  67756.6  73016.5  78121.3  87969.4
sigma               66600.5   2040.2  62881.6  65142.4  66546.4  68010.0  70643.1
mean_PPD           203997.2   4325.2 195553.0 200999.6 203951.7 206942.9 212543.0
log-posterior       -6249.2      2.4  -6254.6  -6250.6  -6248.8  -6247.5  -6245.6

Diagnostics:
                   mcse  Rhat  n_eff
(Intercept)        253.6   1.0 2559
housing_median_age   5.2   1.0 3089
county2&3          184.1   1.0 2902
total_rooms          0.1   1.0 1458
total_bedrooms       1.0   1.0 1307
population           0.2   1.0 1729
households           1.0   1.0 1526
median_income        48.0   1.0 1861
ocean_proximity4   190.5   1.0 2790
ocean_proximity1&3 162.4   1.0 2231
sigma               32.4   1.0 3953
mean_PPD            72.3   1.0 3580
log-posterior        0.1   1.0 1653
```

```
> posterior_interval(model_bayes_county3)
                             5%           95%
(Intercept)        -59809.02299 -17604.330800
housing_median_age    483.65189   1428.130089
county2&3          -48690.42876 -16173.919535
total_rooms           -11.45799      5.665622
total_bedrooms        -59.45503     56.279157
population            -46.94662    -17.600586
households             60.36244    190.899565
median_income       37342.35131  44239.072406
ocean_proximity4    88931.69514 121986.738009
ocean_proximity1&3  60465.14344  85514.454079
sigma               63326.96343  69999.940647
```

Backward elimination untuk model bayes dengan variabel county.

```
> model_bayes_county3_1 <- stan_glm(median_house_value~.-longitude-latitude-
+                                   total_bedrooms, data = dataset,
+                                 prior = normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 15.48 seconds (Warm-up)
Chain 1:                0.808 seconds (Sampling)
Chain 1:                16.288 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
```

```
> posterior_interval(model_bayes_county3_1)
                             5%           95%
(Intercept)        -59065.750566 -17386.700446
housing_median_age    468.819693   1438.357606
total_rooms            -8.965355      3.113507
population            -44.482803    -19.072793
households             78.937013    168.776428
median_income       37993.651583  43743.888400
ocean_proximity4    88355.236287 122246.333653
ocean_proximity1&3  60744.522686  85422.464362
county2&3          -49149.209497 -16407.639845
sigma               63168.855614  70187.451587
```

```
> model_bayes_county3_2 <- stan_glm(median_house_value~.-longitude-latitude-
+                                   total_bedrooms-total_rooms, data = dataset,
+                                 prior = normal(), prior_intercept = normal())

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 12.245 seconds (Warm-up)
Chain 1:                0.474 seconds (Sampling)
Chain 1:                12.719 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
```

```
> posterior_interval(model_bayes_county3_2)
                             5%           95%
(Intercept)        -57510.12007 -17550.93268
housing_median_age    515.30437   1456.37264
population            -45.10364    -20.57769
households             75.06816    143.17612
median_income       37592.71266  42710.37327
ocean_proximity4    89819.36025 123261.29835
ocean_proximity1&3  63174.17470  86742.74248
county2&3          -48411.77304 -15601.85196
sigma               63175.87378  70161.04095
```

Menampilkan koefisien parameter yang sudah diestimasi oleh setiap model.

```
> model_ols3$coefficients
       (Intercept) housing_median_age          population          households       median_income   ocean_proximity4 ocean_proximity1&3
        -42764.4322           1209.2900            -34.5838            114.1889         40168.9858          100714.9037           68065.6983
> model_bayes3$coefficients
       (Intercept) housing_median_age          population          households       median_income   ocean_proximity4 ocean_proximity1&3
        -42648.47493          1217.72788            -34.51848           114.01065         40161.82995         100422.44893           68089.75909
> model_ols_county3_2$coefficients
       (Intercept) housing_median_age          population          households       median_income   ocean_proximity4 ocean_proximity1&3
        -37175.57846           982.78365            -33.02444           109.73343         40144.00614         106849.07296           74933.17126
          county2&3
        -32187.25373
> model_bayes_county3_2$coefficients
       (Intercept) housing_median_age          population          households       median_income   ocean_proximity4 ocean_proximity1&3
        -37493.61583           990.18837            -32.47699           109.01365         40145.83766         106956.74972           74794.20321
          county2&3
        -31957.14438
```

Mengkalkulasi RMSE, MSE, SMAPE, dan MAPE untuk keempat modelnya.

```
> library(Metrics)
> ypredols = predict(model_ols3, datas)
> ypredbayes = predict(model_bayes3, datas)
> rmse(datas$median_house_value, ypredols)
[1] 66560.53
> rmse(datas$median_house_value, ypredbayes)
[1] 66561.37
> mse(datas$median_house_value, ypredols)
[1] 4430303930
> mse(datas$median_house_value, ypredbayes)
[1] 4430416557
> smape(datas$median_house_value, ypredols)
[1] 0.2557983
> smape(datas$median_house_value, ypredbayes)
[1] 0.2560828
> mape(datas$median_house_value, ypredbayes)
[1] 0.2784858
> mape(datas$median_house_value, ypredols)
[1] 0.2776473
```

```
> ypredolscounty = predict(model_ols_county3_2, dataset)
> ypredbayescounty = predict(model_bayes_county3_2, dataset)
> rmse(dataset$median_house_value, ypredolscounty)
[1] 65851.98
> rmse(dataset$median_house_value, ypredbayescounty)
[1] 65853.46
> mse(dataset$median_house_value, ypredolscounty)
[1] 4336482857
> mse(dataset$median_house_value, ypredbayescounty)
[1] 4336678512
> smape(dataset$median_house_value, ypredolscounty)
[1] 0.2533173
> smape(dataset$median_house_value, ypredbayescounty)
[1] 0.2534443
> mape(dataset$median_house_value, ypredolscounty)
[1] 0.2748224
> mape(dataset$median_house_value, ypredbayescounty)
[1] 0.2753984
```