LAPORAN
DATA ANALYTICS COMPETITION
FIND IT! 2023
Tim BukanRISTEK

Disusun Oleh :

Edward Salim

Fathi Qushoyyi Ahimsa

Muhammad Fakhri Robbani

Kota Depok

2023

1. Latar Belakang

Dalam era globalisasi yang terus berkembang, industri perhotelan menjadi salah satu sektor yang mengalami pertumbuhan pesat. Industri hotel merupakan pemain utama dalam perekonomian global dan Amerika Serikat dengan menyumbang sekitar 10% dari PDP dunia. Pada tahun 2022, industri hotel global bernilai lebih dari $4,548 triliun dan diperkirakan akan mengalami pertumbuhan tahunan sebesar 7% dari tahun 2021 hingga 2025. Di Amerika Serikat sendiri, industri hotel dan motel bernilai $177,6 miliar, dan industri pariwisata bernilai $545,11 miliar (Zippia, 2023). Diperkirakan terdapat 187.000 hotel yang ada di seluruh dunia, belum termasuk hotel kecil di bawah 10 kamar dan penginapan seperti RedDoorz, AirBnB, dan OYO.

Pertumbuhan jumlah wisatawan serta permintaan akan akomodasi yang nyaman dan berkualitas mendorong jaringan hotel untuk terus memperluas bisnis mereka. Dalam upaya mengoptimalkan investasi dan mengidentifikasi peluang baru, perusahaan perhotelan seperti Jaringan Hotel Kyozo perlu melakukan analisis harga yang efektif dan akurat. Dalam konteks ini, penggunaan model prediksi harga hotel menjadi sangat penting.

Jaringan Hotel Kyozo memiliki tujuan untuk mengembangkan hotel-hotel baru di berbagai lokasi. Namun, dalam pengembangan ini, mereka perlu menentukan harga kamar yang tepat agar tetap kompetitif dan sesuai dengan kebutuhan pasar. Oleh karena itu, penting untuk mengembangkan model prediksi harga yang dapat membantu perusahaan dalam pengambilan keputusan yang lebih baik. Model ini harus mampu memprediksi harga yang akurat berdasarkan karakteristik dan variabel yang relevan dari hotel-hotel yang sudah ada dalam jaringan mereka.

Penelitian ini memiliki beberapa justifikasi yang kuat untuk dilakukan. Pertama, dengan memiliki model prediksi harga yang handal, Jaringan Hotel Kyozo dapat mengoptimalkan pengembangan hotel baru mereka dengan lebih efisien. Dengan memperoleh perkiraan harga yang akurat, perusahaan dapat menghindari penentuan harga yang terlalu tinggi yang dapat mengurangi minat calon

pelanggan, atau penentuan harga yang terlalu rendah yang dapat mengurangi pendapatan perhotelan.

Kedua, melalui penelitian ini, Jaringan Hotel Kyozo dapat memanfaatkan data yang sudah mereka miliki dari ribuan hotel yang ada di jaringan mereka saat ini. Dengan menganalisis dataset ini, perusahaan dapat mengidentifikasi pola dan tren harga yang signifikan. Penggunaan model prediksi harga akan membantu mengungkapkan faktor-faktor yang berkontribusi terhadap perubahan harga, seperti lokasi geografis, fasilitas, dan faktor lainnya.

Dalam hal ini, penggunaan format avg/night pada harga menunjukkan bahwa model prediksi harga akan memberikan harga rata-rata per malam untuk kamar hotel. Format ini memberikan gambaran yang jelas dan mudah dipahami bagi calon pelanggan, serta memudahkan perusahaan dalam menghitung pendapatan yang diharapkan dari setiap hotel baru yang mereka kembangkan.

Melalui analisis yang komprehensif dan pemodelan yang cermat, diharapkan bahwa hasil penelitian ini akan memberikan kontribusi signifikan terhadap pengembangan hotel-hotel baru yang sukses dan berkelanjutan dalam jaringan Kyozo, serta memberikan wawasan berharga bagi industri perhotelan secara keseluruhan.

2. Tujuan dan Manfaat

Tujuan:

- **Mengembangkan model prediksi harga yang akurat**

  Tujuan utama dari analisis data ini adalah untuk mengembangkan model prediksi harga hotel yang dapat memberikan perkiraan harga yang akurat. Dengan menggunakan dataset yang tersedia dari ribuan hotel dalam jaringan Kyozo, tujuan ini mencakup identifikasi variabel dan karakteristik yang paling berpengaruh terhadap harga kamar hotel. Model prediksi yang handal akan membantu perusahaan dalam menetapkan harga yang tepat untuk hotel-hotel baru yang mereka kembangkan.

- **Meningkatkan pemahaman terhadap faktor-faktor yang mempengaruhi harga hotel**

  Analisis data akan membantu dalam mengidentifikasi faktor-faktor yang berkontribusi terhadap perubahan harga hotel. Hal ini mencakup faktor-faktor seperti lokasi geografis, fasilitas yang tersedia, dan faktor lain yang relevan. Dengan memahami faktor-faktor ini, perusahaan dapat mengoptimalkan strategi penetapan harga mereka untuk mencapai keuntungan maksimal.

Manfaat:

- **Pengambilan keputusan yang lebih baik**

  Dengan adanya model prediksi harga yang handal, perusahaan dapat membuat keputusan yang lebih baik dalam hal penetapan harga untuk hotel-hotel baru yang akan dikembangkan. Dengan menggunakan data yang akurat, perusahaan dapat menghindari penentuan harga yang terlalu tinggi yang dapat mengurangi minat calon pelanggan, atau penentuan harga yang terlalu rendah yang dapat mengurangi pendapatan perhotelan. Hasil analisis data akan memberikan wawasan yang berharga dalam mengoptimalkan strategi penetapan harga.

- **Pengembangan hotel yang efisien**

  Dengan memanfaatkan analisis data yang komprehensif, perusahaan dapat mengidentifikasi tren harga yang signifikan dan pola permintaan pasar. Hal ini akan membantu perusahaan dalam mengarahkan pengembangan hotel baru mereka ke lokasi yang tepat dan mengoptimalkan fasilitas yang ditawarkan. Dengan begitu, perusahaan dapat mengembangkan hotel-hotel baru dengan lebih efisien dan meningkatkan tingkat pengembalian investasi mereka.

- **Keunggulan kompetitif**

  Melalui analisis data yang mendalam, perusahaan dapat mengungkapkan wawasan tentang tren pasar dan preferensi pelanggan. Hal ini akan memberikan perusahaan keunggulan kompetitif dalam menentukan harga yang kompetitif dan menarik bagi calon pelanggan. Dengan memanfaatkan informasi ini, Kyozo dapat memposisikan diri sebagai pemain utama di pasar perhotelan dengan menawarkan harga yang kompetitif dan layanan berkualitas.

- **Peningkatan pendapatan dan keuntungan**

  Dengan menggunakan model prediksi harga yang akurat, Kyozo dapat menyesuaikan harga kamar hotel sesuai dengan permintaan pasar dan faktor-faktor lain yang mempengaruhi harga. Hal ini akan membantu perusahaan dalam meningkatkan pendapatan dan keuntungan mereka dengan cara yang berkelanjutan. Dengan memaksimalkan penghasilan dari setiap kamar hotel, perusahaan dapat mencapai tujuan keuangan mereka dan melanjutkan pertumbuhan yang berhasil.

3. Metode Analisis Data

Dalam konteks analisis data dan pengembangan model prediksi harga hotel, kami menggunakan bahasa pemrograman Python dan Google Colab sebagai lingkungan kolaborasi antar tim. Terdapat beberapa modul Python dan perangkat lunak yang kami gunakan, antara lain:

- **NumPy:** modul Python untuk operasi numerik dan perhitungan array multidimensi.
- **Pandas:** modul Python untuk manipulasi, transformasi, dan analisis data yang efisien.
- **Matplotlib & Seaborn:** modul Python untuk visualisasi data.
- **Scikit-learn:** modul Python untuk pemrosesan data dan *machine learning*.
- **Scipy:** modul Python yang menyediakan berbagai *function* untuk keperluan ilmiah dan teknis.

- **XGBoost:** model *state-of-art* yang dapat digunakan untuk regresi.

Berikut adalah deskripsi *features* dari dataset yang berisi ribuan hotel yang dimiliki jaringan Hotel Kyozo:

- **`facilities`:** fasilitas yang disediakan oleh hotel.
- **`rating`:** rating yang diberikan oleh pengunjung hotel.
- **`location`:** lokasi kota tempat hotel berada.

Performa dari model prediksi harga hotel akan dievaluasi menggunakan metrik Mean Absolute Error (MAE) untuk variabel target (**`Price`**).

3.1. Impute Nilai NA

Nilai-nilai NA (*NaN values*) pada data dapat mempengaruhi kualitas dan akurasi analisis data. Beberapa metode analisis statistik memiliki persyaratan khusus terkait keberadaan *NaN values*. Misalnya, beberapa teknik seperti regresi atau analisis faktor membutuhkan dataset yang lengkap tanpa *NaN values*. Ketika kita meninggalkan nilai NA tanpa penanganan, hal ini dapat menyebabkan bias dalam hasil analisis dan mengurangi akurasi pada hasil yang diperoleh.

Sebelum melakukan analisa data, kita perlu mengetahui apakah terdapat nilai-nilai kosong atau *NaN values* pada data *training*. Berikut adalah persentase nilai NA pada data *training* yang diberikan.

facilities	rating	location
9.82%	20.78%	0.0%

Tabel 1: Persentase jumlah nilai NA pada data *training*

Pada tabel diatas kita mengetahui bahwa terdapat 9.82% nilai NA pada kolom **`facilities`** dan 20.78% pada kolom **`rating`**. Pada kolom **`facilities`**, kami melakukan *most frequent impute* dengan mengubah semua data menjadi *lowercase* terlebih dahulu untuk menghindari *sensitive case* saat menentukan *most frequent* pada data. Pada kolom **`rating`**, kami

impute dengan string kosong (“”) untuk mempermudah ekstraksi nilai dan mempertahankan pola *NaN values* pada kolom tersebut.

3.2. Penghapusan Baris Duplikat

Penghapusan nilai duplikat dapat memastikan bahwa analisis data dilakukan pada dataset yang bersih, akurat, dan konsisten. Ini akan menghasilkan hasil yang lebih andal, mengurangi distorsi, meningkatkan efisiensi komputasi, dan meminimalkan risiko kesalahan**.** Pada data *training* terdapat 175 nilai duplikat yang akan di *drop.* Berikut adalah perbedaan jumlah sebelum dan setelah penghapusan baris duplikat.

Sebelum *Drop*	**Setelah *Drop***
3066	2891

Tabel 2: Jumlah baris sebelum dan sesudah *drop duplicate*

3.3. Ekstraksi Fasilitas pada Hotel

Pada dataset ini terdapat beberapa *keyword* fasilitas yang tersedia pada suatu hotel. Diantaranya adalah `'restaurant'`, `'bar'`, `'swimmingpools'`, `'gym'`, `'internet'`, dan `'pool'`. Kami telah melakukan pemeriksaan terhadap dataset menggunakan algoritma pengecekan dan memastikan bahwa hanya terdapat fasilitas yang telah disebutkan sebelumnya pada kolom **`facilities`**. Pada kolom ini kami melakukan ekstraksi berdasarkan keyword diatas menjadi beberapa kolom *binary* yaitu **`has_restaurant`**, **`has_bar`**, **`has_swimmingpools`**, **`has_gym`**, **`has_internet`**, **dan `has_pool`** dengan keterangan 1 untuk fasilitas yang tersedia dan 0 untuk fasilitas yang tidak tersedia. Berikut adalah jumlah dan persentase hotel yang menyediakan fasilitas-fasilitas pada dataset.

Fasilitas	Hotel yang Tersedia	
	Jumlah	Persentase
Restaurant	2493	86.23%
Bar	2375	82.15%
Gym	660	22.82%
Swimming Pool	507	17.53%
Pool	505	17.46%
Internet	440	15.21%

Tabel 3: Jumlah hotel yang menyediakan fasilitas tertentu

3.4. Ekstraksi *Rating* pada Hotel

Pada dataset ini, kolom **`rating`** memiliki tiga nilai yang dapat diekstrak yaitu *rating* hotel dalam skala 1 - 10, jumlah review, dan level berdasarkan *rating* hotel. Pada kolom ini kami melakukan ekstraksi menjadi tiga kolom yaitu **`num_rating`, `num_reviews`, dan `level_rating`.** Kemudian jika nilai kolom **`rating`** berupa string kosong (tidak ada), maka kami menetapkan nilai 0 pada **`num_rating`** dan **`num_reviews`**, dan `'No Level'` pada **`level_rating.`**

3.5. Penambahan Fitur Baru

Dalam upaya meningkatkan performa model prediksi harga hotel, kami melakukan penggabungan fitur yang tepat dan relevan untuk memberikan informasi yang lebih kaya kepada model. Tujuan dari penggabungan ini adalah untuk meningkatkan pengaruh fitur-fitur yang mungkin memiliki kontribusi rendah terhadap variabel target, sehingga model dapat mempelajari pola yang lebih kompleks. Salah satu fitur baru yang kami buat adalah **`facilities_count`** yang menggabungkan semua *feature* fasilitas yang ada.

Selain itu, kami juga membuat fitur **`rating_ratio`** dan **`rating_review_combined`** yang didasarkan pada dua operasi fitur yang serupa namun dengan operator yang berbeda. Fitur **`rating_ratio`** dapat memberikan gambaran tentang tingkat kepuasan pengunjung terhadap hotel dan fitur **`rating_review_combined`** dapat mengekspresikan tingkat

popularitas hotel yang didasarkan pada interaksi antara rating dan ulasan dari pengunjung. Fitur yang kami buat memiliki formula sebagai berikut:

$$facilities\_count = \Sigma\, facilities\_cols$$

Dimana **facilities_cols** terdiri atas fitur *binary* **has_restaurant, has_bar, has_swimmingpools, has_gym, has_internet,** dan **has_pool.**

$$rating\_ratio = \frac{rating\ hotel}{jumlah\ review}$$

$$rating\_review\_combined = rating\ hotel \times jumlah\ review$$

3.6. *Encoding* Kolom Kategorikal

Mayoritas model pada *machine learning* hanya bisa melakukan prediksi dengan kolom numerik. Oleh karena itu kolom kategorik yang ada pada dataset perlu ditransformasikan menjadi kolom numerik menggunakan *One-Hot Encoding* dan *Ordinal Encoding.*

Kolom kategorikal yang akan kita *encoding* yaitu **level_rating** dan **location.** Kolom **level_rating** memiliki perbedaan yang konstan (*Symmetric difference)* antara level yang satu dengan yang lainnya (tingkatan) sehingga dapat ditransformasi menggunakan *Ordinal Encoding* (`"Excellent": 4, "Very Good": 3, "Good": 2, "Fair": 1, "No Level": 0).` Kemudian kolom **location** dapat ditransformasi menggunakan *One-Hot Encoding* karena tidak memiliki perbedaan yang konstan antara satu dengan yang lain (tidak ada tingkatan).

3.7. Penanganan *Outliers* pada Data Training

Dalam analisis regresi linear, terutama pada model seperti LinearRegression, LassoCV, dan RidgeCV, *outliers* dapat memiliki pengaruh yang signifikan terhadap hasil prediksi. Keberadaan *outliers* dapat mempengaruhi koefisien regresi dan dapat menghasilkan model yang tidak stabil atau tidak akurat. Oleh karena itu, kita perlu melakukan

pembuangan nilai yang diluar rentang nilai batas bawah (*Lower bound)* dan batas atas (*Upper bound)* yang ditentukan. Penghitungan rentang nilai batas adalah sebagai berikut:

$$Lower\ bound\ =\ Q_1\ -\ 1.5\ \times IQR$$

$$Upper\ bound\ =\ Q_3\ +\ 1.5\ \times IQR$$

Dimana IQR (*interquartile range*) diperoleh dari pengurangan Q3 dengan Q1. Pada submisi terakhir kami yang menjadi skor akhir di *leaderboard,* kami menggunakan *winsorize* dalam *handling outliers*, namun karena terdapat potensi bias yang tinggi dan memungkinkan terjadinya *overfit*, kami akhirnya menggantinya dengan melakukan *remove outliers*. Index data yang mengandung nilai di luar rentang batas nilai akan kami *drop* dan sisanya akan kami gunakan untuk *modeling*.

3.8. *Feature Scaling* dengan *Normalization*

Normalization atau normalisasi adalah proses yang digunakan untuk mengubah skala variabel-variabel dalam dataset sehingga memiliki rentang nilai yang seragam. Dalam konteks regresi linear, normalisasi memiliki peran penting dalam memastikan bahwa variabel-variabel memiliki bobot yang seimbang dan meminimalkan efek dominasi variabel yang memiliki rentang nilai besar. Model regresi linear seperti LinearRegression, LassoCV, dan RidgeCV dapat mendapatkan manfaat dari normalisasi ini.

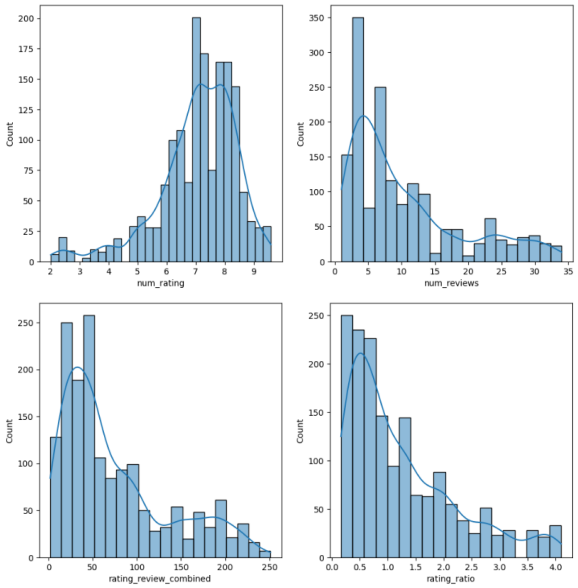

Figure 1: distribusi fitur numerik sebelum dilakukan transformasi

Dapat dilihat bahwa fitur **`num_rating`** memiliki distribusi yang condong ke kiri (*left-skewed distribution*), sementara fitur **`num_reviews`**, **`rating_ratio`**, dan **`rating_review_combined`** memiliki distribusi yang condong ke kanan (*right-skewed distribution*). Untuk menangani situasi ini, perlu dilakukan transformasi pada fitur-fitur tersebut agar dapat mendekati distribusi normal.

Untuk fitur **`num_rating`** yang condong ke kiri, dapat dilakukan transformasi menggunakan metode Box-Cox. Sementara itu, untuk fitur **`num_reviews`**, **`rating_ratio`**, dan **`rating_review_combined`** yang condong ke kanan, dapat dilakukan transformasi menggunakan metode akar kuadrat (*square root transformation*).

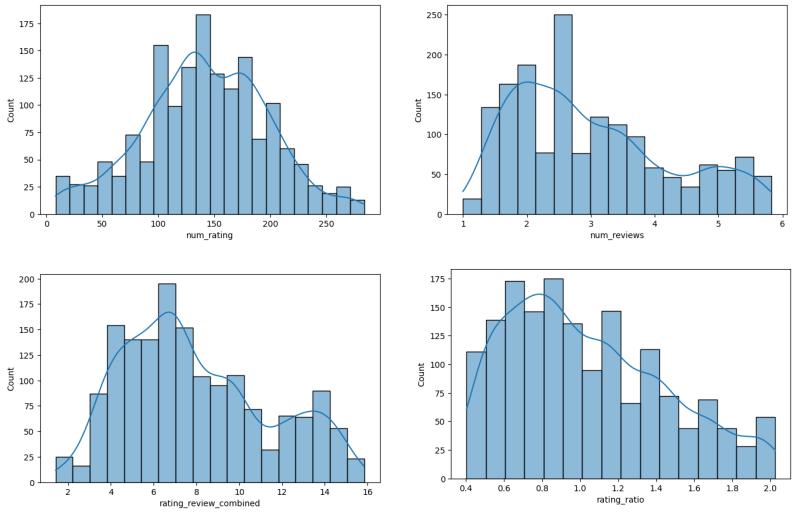

Figure 2: distribusi fitur numerik sesudah dilakukan transformasi

Setelah dilakukan transformasi, terlihat bahwa distribusi fitur **`num_rating`** telah berhasil diubah menjadi simetris dan mendekati distribusi normal. Sementara itu, fitur **`num_reviews`**, **`rating_ratio`**, dan **`rating_review_combined`** juga menunjukkan perbaikan signifikan dan mendekati simetri.

Dengan mendekati distribusi normal, fitur-fitur tersebut menjadi lebih sesuai dengan asumsi yang diperlukan dalam regresi linear dan algoritma machine learning lainnya. Ini akan membantu meningkatkan kualitas prediksi dan memastikan interpretasi yang lebih akurat dari hasil yang dihasilkan oleh model.

3.9. Pemodelan dan Evaluasi Kinerja

Dalam pemodelan prediksi harga hotel, kami telah memilih fitur-fitur berikut sebagai variabel independen (X): **`has_swimmingpools`**, **`has_pool`**, **`num_rating`**, **`num_reviews`**, **`rating_review_combined`**, **`rating_ratio`**, dan **`facilities_count`**. Keputusan ini didasarkan pada hasil analisis EDA yang telah kami lakukan, di mana fitur-fitur ini menunjukkan hubungan yang erat dengan variabel target (**`Price`**).

Untuk mengevaluasi kinerja model, kami membandingkan skor *Mean Absolute Error* (MAE) dari beberapa model yang telah kami pilih. Model-model tersebut termasuk LinearRegression, LassoCV, RidgeCV, BayesianRidge, ElasticNet, dan XGBRegressor. Penggunaan metode KFold dengan 10 split dilakukan untuk memastikan akurasi skor yang dihasilkan. KFold membagi data *train* menjadi 10 bagian yang sama ukurannya. Dalam setiap iterasi, satu bagian digunakan sebagai data uji, sedangkan sembilan bagian lainnya digunakan sebagai data latih. Proses ini dilakukan sebanyak 10 kali dengan bagian yang berbeda sebagai data uji setiap kali iterasi. Skor evaluasi dihitung pada setiap iterasi, dan skor akhir diambil sebagai rata-rata dari semua iterasi. Berikut ini hasil skor yang didapatkan setiap model:

Model	MAE Score
LinearRegression	6937.04
LassoCV	7013.03
RidgeCV	6943.48
BayesianRidge	6939.73
ElasticNet	7362.28
XGBRegressor	825.07

Tabel 4: Model yang digunakan serta MAE dengan KFold

Dalam menentukan model terbaik, biasanya kita menginginkan nilai MAE yang semakin kecil karena itu menunjukkan bahwa model memiliki tingkat kesalahan yang lebih rendah dalam memprediksi harga hotel. Namun, penting untuk dicatat bahwa hanya mengandalkan nilai MAE terendah saja tidak selalu menjamin bahwa model tersebut adalah yang terbaik.

Dalam kasus ini, meskipun XGBRegressor memiliki nilai MAE yang paling kecil, hal ini belum tentu menunjukkan bahwa model tersebut adalah yang terbaik. Terlalu kecilnya nilai MAE dapat menunjukkan kemungkinan adanya *overfitting*, di mana model terlalu spesifik terhadap

data *train* dan tidak mampu melakukan generalisasi yang baik pada data baru.

3.10. Pengecekan *Overfit* dan *Underfit*

Kami akan menggunakan *learning curve* untuk mengecek apakah model sebelumnya *overfit* atau *underfit*. Learning curve adalah alat yang digunakan untuk menganalisis kinerja model machine learning seiring dengan peningkatan jumlah data *train*. Grafik learning curve menunjukkan perubahan skor evaluasi model pada data *train* dan data *test* seiring dengan penambahan jumlah sampel data *train*.

Learning curve pada model yang baik menunjukkan *training loss* dan *validation loss* yang saling mendekati satu sama lain, dengan *validation loss* yang lebih tinggi daripada *training loss*. Pada awalnya, terjadi penurunan bertahap dalam *training loss* dan *validation loss*, dan setelah mencapai titik tertentu, tingkat kehilangan tersebut hampir tidak berubah (mendatar).

Dalam kasus *overfitting*, terlihat bahwa terdapat perbedaan yang signifikan antara *training loss* dan *validation loss*. *Validation loss* terus menurun secara bertahap tanpa mendatar, sementara *training loss* sangat rendah dan cenderung stabil. Berikut ini grafiknya:

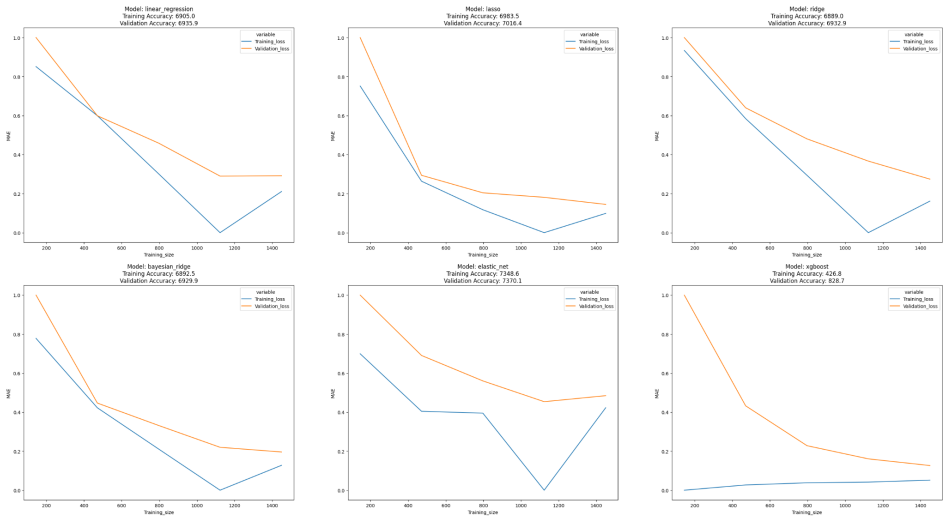

Figure 3: *Learning Curve* untuk melihat performa belajar model

Dari grafik yang disajikan, terlihat bahwa model-model seperti LinearRegression, RidgeCV, BayesianRidge, ElasticNet, dan XGBRegressor mengalami *overfitting*. Hal ini terlihat dari penurunan *validation loss* yang signifikan dan terus-menerus tanpa mencapai tingkat kestabilan. Sebagai akibatnya, model-model tersebut akan memiliki kinerja yang buruk dalam melakukan generalisasi pada data yang belum pernah dilihat sebelumnya.

Namun, berdasarkan analisis tersebut, kami menyimpulkan bahwa model LassoCV (baris: 1, kolom: 2) merupakan pilihan terbaik untuk prediksi harga hotel. Meskipun nilai skor MAE bukan yang terendah, *learning curve* pada model ini mencapai titik tengah antara *overfitting* dan *underfitting*. Hal ini menunjukkan bahwa model LassoCV mampu menyesuaikan dengan baik terhadap data *train* yang diberikan dan memiliki kemampuan yang lebih baik dalam melakukan generalisasi pada data yang belum pernah dilihat sebelumnya. Selain itu, perlu dicatat bahwa dalam pemilihan model LassoCV, kita tidak perlu khawatir tentang adanya masalah *multicollinearity* antara fitur-fitur yang digunakan. LassoCV memiliki kemampuan bawaan untuk mengatasi masalah ini.

3.11. *Hyperparameter Tuning* dan Submisi

Tujuan dari melakukan *hyperparameter tuning* adalah untuk meningkatkan akurasi model LassoCV. Namun, setelah proses *tuning* dilakukan, kami mengalami penurunan performa model. Dalam rangka mengatasi hal tersebut, kami memutuskan untuk kembali menggunakan *default* parameter. Dalam hasil pengujian model melalui file submisi, kami mendapatkan skor sebesar 10990.70 pada *private leaderboard* dan 11528.50 pada *public leaderboard* di platform Kaggle. Meskipun terdapat sedikit perbedaan dengan skor yang kami capai pada tahap penyisihan (*private*: 10461.63, *public*: 10842.52), kami yakin bahwa kami telah berhasil meminimalisir potensi *overfitting* dan *bias* pada model terbaru kami. Setelah melakukan investigasi lebih lanjut, kami menemukan bahwa model yang kami gunakan pada tahap penyisihan masih memiliki *bias*

yang tinggi. Oleh karena itu, kami berusaha semaksimal mungkin untuk menghasilkan model yang dapat memberikan prediksi harga hotel yang akurat dan dapat diandalkan.

4. Analisis

4.1. Distribusi Data Fitur Kategorik

Mengetahui distribusi suatu data sangat penting karena dapat memberikan wawasan yang berharga mengenai karakteristik data tersebut. Distribusi data mencerminkan bagaimana nilai-nilai data tersebar dan seberapa sering nilai-nilai tersebut muncul. Terdapat data dengan fitur kategorik, yaitu **`location`** dan **`level_rating`**, serta data dengan fitur *binary* yaitu kolom-kolom fasilitas**.** Berikut adalah distribusi data (barplot) pada kolom-kolom tersebut sebelum dilakukan penghapusan *outlier*:

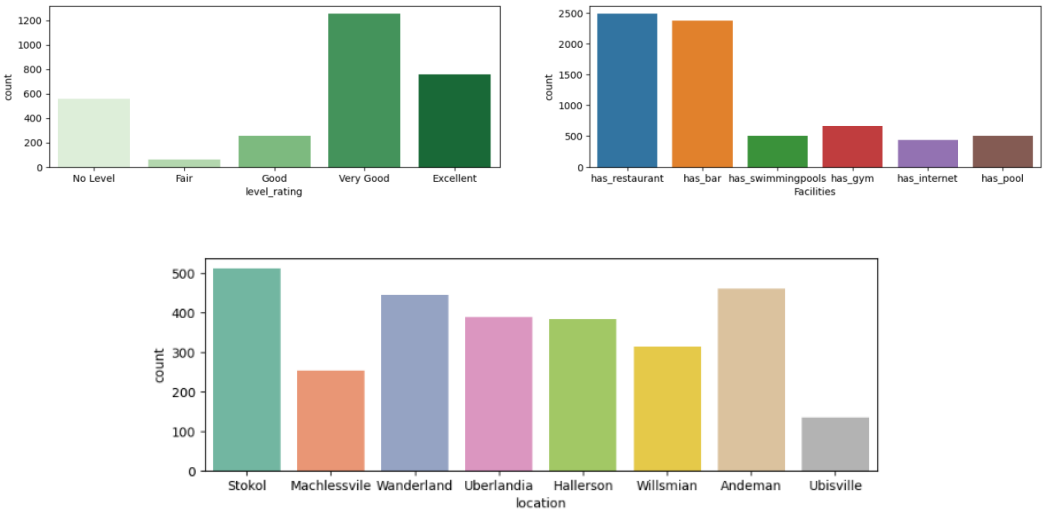

Figure 4: distribusi fitur kategorik dengan *outlier*

Dari sini kami menyimpulkan bahwa sebagian besar hotel pada data *training* berasal dari "`Stokol`", dan sebagian besar hotel yang ada pada data *training* memiliki level rating "`Very Good`". Selain itu, sebagian besar hotel pada data *training* juga memiliki "`restaurant`" atau "`bar`".

4.2. Distribusi Data Fitur Numerikal

Selain fitur kategorik, data *train* juga memiliki fitur numerik. Berikut visualisasi distribusi kolom numerik pada data *training* menggunakan histogram:

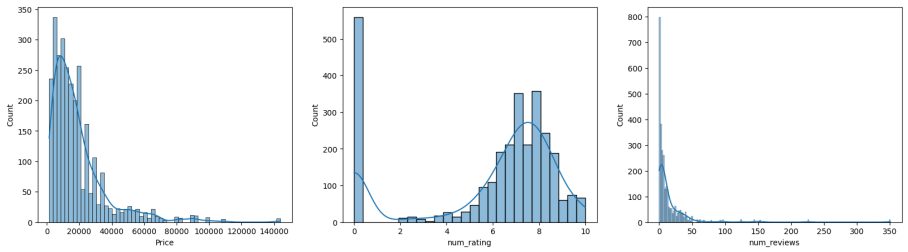

Figure 5: distribusi fitur numerik sebelum dilakukan transformasi

Berdasarkan histogram diatas, dapat dilihat bahwa kolom **`num_rating`** menunjukkan distribusi yang mendekati normal dengan puncak di sekitar nilai rata-rata. Namun, perlu diperhatikan bahwa terdapat dominasi nilai 0 (*zero-inflated*) yang cukup signifikan. Selanjutnya, kolom **`num_reviews`** dan **`Price`** menunjukkan *right-skewed distribution*. Hal ini dapat disebabkan oleh adanya data *outliers* pada kedua kolom tersebut. Hal ini kami teliti lebih lanjut melalui visualisasi boxplot sebagai berikut:

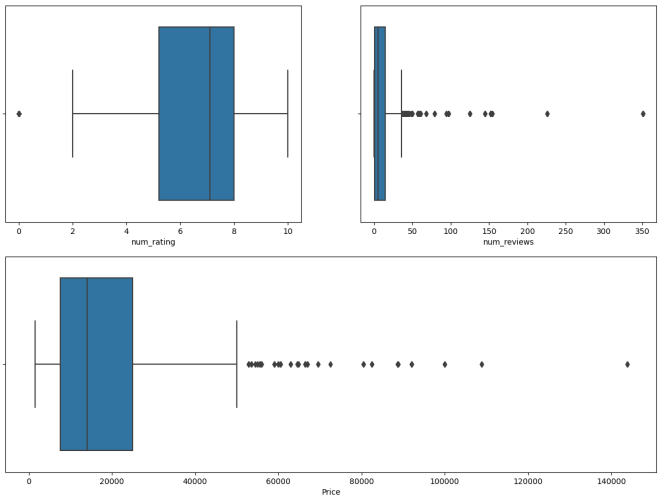

Figure 6: distribusi fitur numerik dengan boxplot

Boxplot diatas meyakinkan kita bahwa terdapat *outliers* yang cukup banyak pada kolom **`num_reviews`** dan **`Price.`** Nilai yang ekstrem tersebut dapat menyebabkan *bias* yang signifikan dalam estimasi parameter dan prediksi. Oleh karena itu, kami akan melakukan penghapusan *outliers* pada ketiga kolom tersebut berdasarkan formula yang kami jelaskan pada Subbab 3.7. Dengan menghapus *outlier*, kita dapat mengurangi pengaruh yang tidak wajar dan memperoleh hasil yang lebih akurat. Berikut tabel banyak *row* data *train* sebelum dan setelah penghapusan *outliers* beserta distribusi datanya*:*

Sebelum *Drop*	**Setelah *Drop***
2891	1612

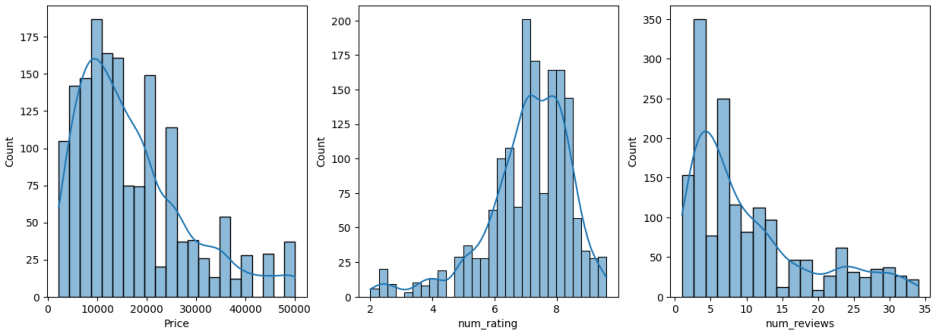

Figure 7: Jumlah row sebelum dan setelah *remove outlier* (atas) dan distribusi data sebelum dan sesudah *remove outlier* (bawah)

Selanjutnya, kami akan melakukan analisis menggunakan data training setelah dilakukan penghapusan *outlier*. Dengan *outliers* yang telah dihapus, kita dapat melihat pengaruh fitur-fitur lain terhadap harga hotel secara lebih akurat.

4.3. Pengaruh Lokasi terhadap Harga Hotel

Setelah menghapus *outlier*, kami akan mengeksplorasi pengaruh fitur **`location`** terhadap harga hotel. Kami melakukan visualisasi menggunakan KDE plot dan Boxenplot untuk memahami hubungan antara lokasi dan harga hotel sebagai berikut:

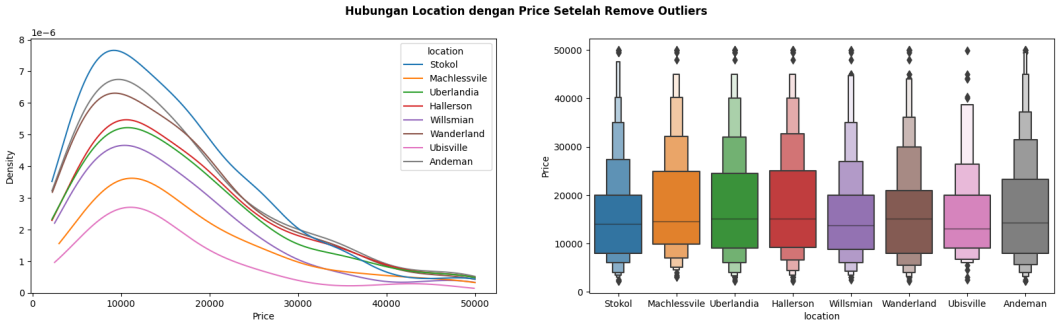

Figure 8: Hubungan antara lokasi terhadap harga hotel

Berdasarkan visualisasi diatas, tidak terlihat adanya pengaruh yang signifikan antara lokasi dan harga hotel setelah melakukan penghapusan *outlier*. Kemungkinan penyebabnya adalah variasi harga yang relatif kecil antara lokasi-lokasi yang berbeda.

4.4. Pengaruh Level Rating terhadap Harga Hotel

Selanjutnya, kami akan menganalisis pengaruh level rating terhadap harga hotel. Berikut ini adalah visualisasi menggunakan KDE Plot dan Boxenplot untuk melihat hubungan antara level rating dengan harga hotel:

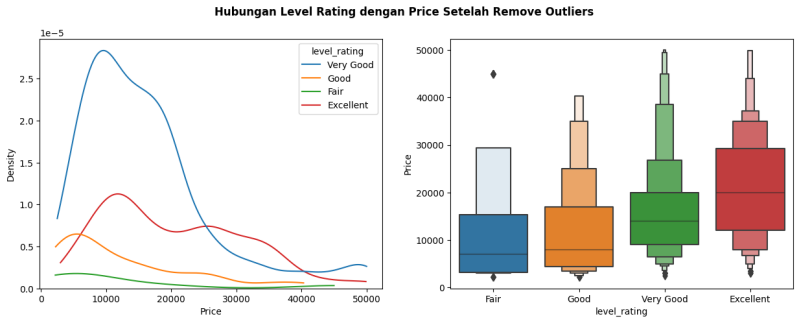

Figure 9: Hubungan level rating terhadap harga hotel

Grafik menunjukkan bahwa hotel dengan level rating yang tinggi cenderung memiliki harga yang tinggi pula. Dapat dilihat bahwa visualisasi KDE plot mendukung kesimpulan tersebut. Hotel dengan harga lebih dari 25000 cenderung didominasi oleh hotel dengan level rating "`Very Good`" dan "`Excellent`". Hotel dengan level rating yang tinggi umumnya memiliki reputasi yang baik di kalangan wisatawan. Karena popularitas yang tinggi, permintaan untuk menginap di hotel tersebut juga

dapat meningkat, yang kemudian dapat mempengaruhi harga hotel menjadi lebih tinggi.

4.5. Pengaruh Fasilitas terhadap Harga Hotel

Secara logis, kita dapat mengasumsikan bahwa fasilitas yang tersedia dalam sebuah hotel dapat mempengaruhi harga hotel tersebut. Untuk menganalisis hubungan antara fasilitas dan harga hotel, kami akan melihat grafik yang menunjukkan hubungan antara setiap fasilitas yang tersedia dengan harga hotel.

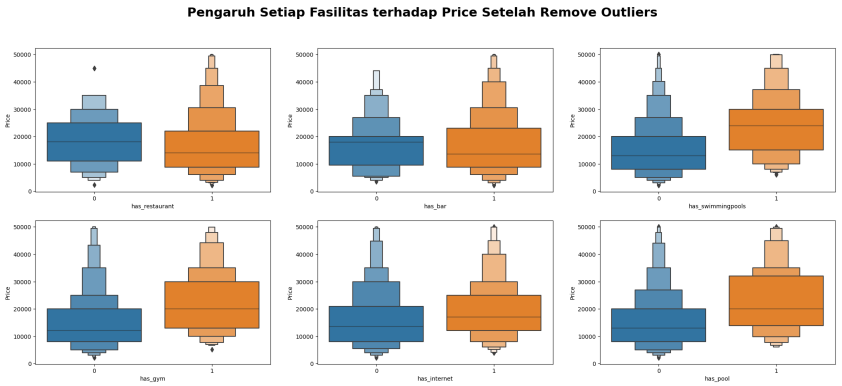

Figure 10: Hubungan setiap fitur dengan harga hotel

Berdasarkan *boxenplot* di atas, terlihat bahwa adanya fasilitas seperti "*Swimming Pools*", "*Gym*", dan "*Pool*" memiliki pengaruh yang cukup signifikan terhadap peningkatan harga hotel dibandingkan dengan fasilitas lainnya. Hotel-hotel yang menyediakan fasilitas-fasilitas tersebut cenderung menawarkan layanan yang lebih lengkap dan berkualitas, yang secara logis mempengaruhi peningkatan harga kamar.

Namun, perlu dicatat bahwa ada pengecualian untuk fasilitas "*Restaurant*". Meskipun restoran dianggap sebagai fasilitas yang diharapkan ada dalam hotel, ternyata hotel yang memiliki fasilitas restoran cenderung memiliki harga yang lebih rendah. Hal ini mungkin disebabkan oleh kompetisi di sekitar lokasi hotel yang menawarkan banyak pilihan tempat makan di luar.

4.6. Pengaruh *Rating* dan *Reviews* terhadap Harga Hotel

Selanjutnya, kami akan menganalisis pengaruh kolom numerik, yaitu **num_rating** dan **num_reviews**, terhadap harga hotel. Berikut ini Scatterplot yang menunjukkan hubungannya:

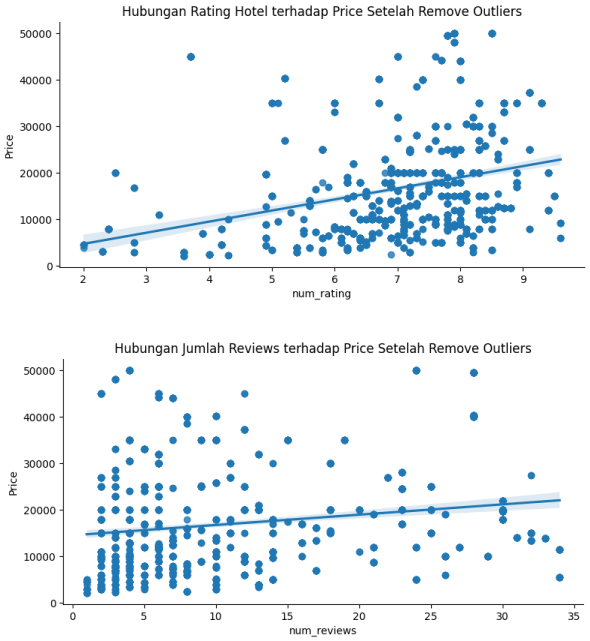

Figure 11: Hubungan rating dan reviews terhadap harga hotel

Grafik tersebut mengindikasikan bahwa kecenderungan linear pada **num_reviews** lemah, sedangkan **num_rating** memiliki kecenderungan linear yang moderat. Sehingga dapat disimpulkan bahwa semakin tinggi **num_rating**, maka **Price** juga akan cenderung lebih tinggi dengan tingkat kenaikan yang moderat. Dalam industri perhotelan, hotel dengan *rating* yang tinggi cenderung menargetkan pasar yang lebih eksklusif atau menghadirkan pengalaman premium kepada tamu. Oleh karena itu, hotel-hotel tersebut dapat menetapkan harga yang lebih tinggi untuk mencerminkan nilai tambah yang mereka berikan kepada pelanggan.

4.7. Pemilihan Fitur untuk *Modeling*

Fitur-fitur yang kami pilih untuk dimasukkan dalam model LassoCV adalah **has_swimmingpools**, **has_pool**, **num_rating**, **num_reviews**, **rating_review_combined**, **rating_ratio**, dan **facilities_count**

dengan alasan yang sudah disebutkan pada Subbab 3.9. Kami menghilangkan beberapa fitur lainnya, seperti **`has_bar`**, **`has_internet`**, **`has_restaurant`**, dan **`location`**, karena fitur-fitur tersebut memiliki korelasi yang rendah dengan variabel target dan cenderung tidak memberikan informasi yang signifikan dalam memprediksi harga hotel. Dalam hal ini, fitur-fitur tersebut hanya akan menambah *noise* yang dapat menurunkan akurasi model.

Kami juga memutuskan untuk tidak menggunakan fitur **`level_rating`**, karena fitur ini merupakan kategori yang sebenarnya didasarkan pada kolom **`num_rating`**. Model linear sulit untuk mempelajari fitur kategorikal secara efektif dibandingkan dengan fitur numerikal, sehingga penggunaan **`level_rating`** hanya akan menciptakan *noise* dalam model. Selain itu, kami memilih untuk tidak menggunakan fitur **`has_gym`** karena kami menemukan bahwa skor model LassoCV lebih akurat ketika fitur ini tidak digunakan. Meskipun tidak dapat dijelaskan secara eksplisit, kemungkinan terdapat pola yang lebih baik diidentifikasi oleh LassoCV tanpa melibatkan fitur **`has_gym`**.

5. Kesimpulan

Jaringan Hotel Kyozo memiliki tujuan untuk mengembangkan hotel-hotel baru di berbagai lokasi dengan harga kamar yang kompetitif dan sesuai dengan kebutuhan pasar. Dalam pengembangan ini, penting bagi mereka untuk memiliki model prediksi harga yang dapat membantu dalam pengambilan keputusan yang lebih baik. Dalam membangun model prediksi, telah dilakukan perbandingan antara dua model, yaitu XGBRegressor dan LassoCV. Meskipun model XGBRegressor menunjukkan tingkat kesalahan (MAE) yang lebih rendah, hasil dari learning curve menunjukkan bahwa model LassoCV merupakan pilihan terbaik untuk prediksi harga hotel.

Analisis data menunjukkan bahwa sebagian besar hotel pada data *training* berasal dari "Stokol" dan memiliki level *rating* "Very Good". Keberadaan fasilitas seperti restoran dan bar juga umum pada hotel-hotel tersebut. Penghapusan outlier pada

kolom **`num_rating`**, **`num_reviews`**, dan **`Price`** diperlukan untuk meningkatkan akurasi prediksi. Pengaruh lokasi terhadap harga hotel tidak terlihat signifikan, kemungkinan karena variasi harga yang relatif kecil antara lokasi yang berbeda. Namun, hotel dengan level *rating* tinggi cenderung memiliki harga yang tinggi pula, karena reputasi yang baik dan permintaan yang meningkat.

Keberadaan fasilitas kolam renang, jumlah *review*, *rating*, dan jumlah fasilitas suatu hotel, memiliki pengaruh yang signifikan terhadap prediksi harga hotel. Semakin tinggi *rating*, harga hotel cenderung meningkat dengan tingkat kenaikan yang moderat. Hotel dengan fasilitas yang lengkap cenderung memiliki harga yang lebih tinggi, mencerminkan nilai tambah yang mereka berikan kepada pelanggan.

Dengan mempertimbangkan faktor-faktor tersebut, Jaringan Hotel Kyozo dapat menggunakan model prediksi harga yang telah dikembangkan untuk mengambil keputusan yang lebih baik dalam penetapan harga kamar hotel baru mereka, memastikan kompetitivitas dan kepuasan pelanggan.

6. Daftar Pustaka

McCain, A. (2023). *25 Hotel Industry Statistics [2023]: Hotel Rate Trends And Market Data*. Zippia. Available at: https://www.zippia.com/advice/hotel-industry-statistics/

Kirenz, J. (2021). *Lasso Regression with Python | Jan Kirenz*. Jan Kirenz. Available at: https://www.kirenz.com/post/2019-08-12-python-lasso-regression-auto/

Frost, J. (2017). *Choosing the Correct Type of Regression Analysis*. Statistics By Jim. Available at: https://statisticsbyjim.com/regression/choosing-regression-analysis/

KSV Muralidhar (2021). *Learning Curve to identify Overfitting and Underfitting in Machine Learning*. Medium. Available at:

https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5#

Jordan Hollander (2022). *75+ Hospitality Statistics You Should Know (2023)*. Available at: https://hoteltechreport.com/news/hospitality-statistics

7. Lampiran

Dataset dan *source code* (notebook analisis) [Klik]