Analisa Jumlah Tamu Hotel Bintang di Setiap Provinsi dengan Multiple Linear Regression

Grup 4 : Pariwisata

Team Leader : Dennis Jonathan (23101910027)

Member: Steffi Victoria Yahya (23101910037), Gavrila Blenda (23101910031), Sherina Jakfar (23101910036), Irwin Budianto (23101910051)

S1 Business Mathematics - Universitas Prasetiya Mulya

Abstrak :

Tujuan dari penelitian ini adalah untuk menganalisis jumlah tamu hotel bintang (y) dengan menggunakan variabel independen (nilai x) yaitu mean lama inap tamu (per malam), persen tingkat huni kamar, jumlah tempat tidur, jumlah kamar tidur, jumlah akomodasi, kepadatan penduduk (jiwa/km2) dan jumlah perjalanan wisatawan nusantara. Dari hasil penelitian ini, peneliti menemukan bahwa persamaan untuk mengestimasi jumlah tamu total dengan persamaan:

$\hat{y} = 1521994.3430552406 + 1620436.3509497817x_1 + 373703.7081082469x_2$

dengan nilai *adjusted R-squared* sebesar 0.969.

Kata kunci: *Variance Inflation Factor, Ordinary Least Square, Regresi Linear Ganda, Pariwisata*

1. Pendahuluan

Pariwisata di Indonesia mulai dikenal sejak jaman kolonial pada tahun 1910-1920. Pada tahun tersebut dikeluarkan keputusan Gubernur Jenderal Belanda bernama VTV *(Vereneiging Touristen Verker)*. VTV ini merupakan suatu badan atau *official tourist bureau* yang memiliki kedudukan sebagai lembaga pariwisata dan juga bertindak sebagai *tour operator* atau *travel agent*. Pada tahun 1913, organisasi turis VTV menerbitkan *guide book* atau buku panduan bagi turis yang ingin berwisata ke Indonesia. Di dalam *guidebook* tersebut terdapat beberapa tempat yang direkomendasikan untuk menjadi tempat wisata. Seperti Banten, Jawa Barat, Jawa Timur, Jawa Tengah, Bali, Lombok, Sumatera Utara, Sumatera Barat, Sumatera Selatan, dan Tanah Toraja di Sulawesi. Sementara itu,

pada tahun 1923, juga diterbitkan sebuah surat kabar mingguan yang berisi panduan pelayanan akomodasi selama berwisata ke Indonesia.

Sejak diterbitkan surat kabar pada tahun 1923, pemerintah kolonial mulai menganggap serius permintaan masyarakat Eropa untuk berwisata ke Indonesia. Maka dari itu, pemerintah kolonial berusaha memberikan pelayanan yang terbaik kepada wisatawan asing yang sedang melakukan perjalanan wisata. Pemerintah pun akhirnya mendirikan *travel agent* di Batavia pada tahun 1926 bernama *Linsone Linderman (Lis Lind)* yang berpusat di negeri Belanda. Sejak itu, Bali mulai dikenal oleh wisatawan asing. Juga tempat-tempat di luar pulau jawa seperti Air Terjun Bantimurung pun menjadi tujuan wisata asing. Sekarang, pariwisata di Indonesia merupakan sektor ekonomi penting di Indonesia. Pada tahun 2009, pariwisata menempati urutan ketiga dalam hal penerimaan devisa setelah komoditas minyak dan gas bumi serta minyak kelapa sawit. Pada tahun 2020, jumlah wisatawan asing ada sebesar 4.052.923 dan pada tahun 2019 ada sebesar 16.108.600.

Berdasarkan data dari Badan Pusat Statistik, sebelas provinsi yang paling sering dikunjungi oleh para turis adalah Bali sekitar lebih dari 3,7 juta disusul, DKI Jakarta, Daerah Istimewa Yogyakarta, Jawa Timur, Jawa Barat, Sumatra Utara, Lampung, Sulawesi Selatan, Sumatra Selatan, Banten dan Sumatera Barat. Sekitar 59% turis berkunjung ke Indonesia untuk tujuan liburan, sementara 38% untuk tujuan bisnis.

2. Kajian Teori

2.1 Multiple Linear Regression Model

Multiple linear regression atau analisis regresi linier ganda adalah analisis statistik untuk mengetahui pengaruh beberapa variabel bebas (independen) terhadap variabel yang terkait (dependent).

Persamaan multiple linear regression:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Keterangan:

$\beta_0 = intercept$

$\beta_n = slope$

X_n = variabel bebas (independent)

$\hat{Y}$ =variabel yang akan diprediksi

ε = *error*

Pada model regresi linier ganda ini, tidak semua faktor/variabel independen bisa dimasukkan sebagai pembuatan model. Jika menambahkan variabel yang tidak memiliki relasi atau tidak relevan, model yang dihasilkan tidak akurat untuk memprediksi variabel yang ingin diprediksi.

2.2 *Ordinary Least Square* (OLS)

OLS adalah salah satu metode untuk analisis regresi linear. Metode ini bertujuan untuk mengetahui pengaruh variabel independen (variabel bebas) terhadap variabel dependent (variabel tidak bebas). Kriteria OLS adalah *"Line of Best Fit"* atau dengan kata lain jumlah kuadrat dari deviasi antara titik-titik observasi dengan garis regresi adalah minimum.

2.3 *Variance Inflation Factor* (VIF)

Uji multikolinearitas atau VIF adalah metode uji asumsi klasik (normalitas dan heteroskedastisitas) dalam analisis regresi linear berganda. Tujuan dari VIF ini adalah untuk menguji apakah model regresi ditemukan adanya korelasi/ hubungan kuat antara variabel independen. Untuk model regresi yang baik sebaiknya tidak ada korelasi yang kuat antar variabel independent agar tidak terjadi multikolinearitas. Jika nilai VIF yang ditampilkan pada suatu variabel lebih dari 5, maka variabel tersebut memiliki multikolinearitas.

2.4 *Root Mean Square Error* (RMSE)

RMSE adalah metode yang mengukur tingkat akurasi hasil perkiraan suatu model. Nilai RMSE yang rendah menunjukkan bahwa variasi nilai yang dihasilkan oleh suatu model prakiraan mendekati variasi nilai observasinya. Semakin kecil nilai RMSE, semakin dekat nilai yang diprediksi dan diamati.

Rumus RMSE :

$$RMSE = \left(\frac{\Sigma(y_i - \hat{y}_i)}{n}\right)^{1/2}$$

RMSE = nilai root mean square error

y = nilai hasil observasi

ŷ = nilai hasil prediksi

i = urutan data

n = jumlah data

3. Metodologi

3.1 Pengumpulan Data

Dari semua kemungkinan data yang ada di internet, data yang digunakan untuk penelitian ini berkisar pada jumlah wisatawan domestik dan mancanegara yang menginap pada hotel bintang. Data yang kami ambil merupakan agregat dari beberapa data yang ada di situs Badan Pusat Statistik dengan menggunakan fitur tabel dinamis. Kemudian kami juga menambahkan satu variabel yaitu kepadatan penduduk untuk melihat apakah kepadatan penduduk berhubungan dengan jumlah tamu yang datang ke suatu provinsi. Data tersebut diambil dari dua tahun, yaitu 2016 dan 2019 dan memiliki kolom yang terdiri dari 9 variabel yang termasuk:

1	Rata-rata Lama Menginap Tamu pada Hotel Bintang (Malam)	68 non-null
2	Tingkat Penghunian Kamar pada Hotel Bintang Menurut Provinsi (Persen)	68 non-null
3	Jumlah Tempat Tidur Hotel Bintang	68 non-null
4	Jumlah Kamar Tidur Hotel Bintang	68 non-null
5	Jumlah Akomodasi Hotel Bintang	68 non-null
6	Kepadatan Penduduk menurut Provinsi (jiwa/km2)	68 non-null
7	Jumlah Perjalanan Wisatawan Nusantara (Orang)	68 non-null
8	Jumlah Tamu Indonesia pada Hotel Bintang (Ribu Orang)	68 non-null
9	Jumlah Tamu Asing pada Hotel Bintang (Ribu Orang)	68 non-null

Tabel 1: Daftar Kolom-Kolom di Data

Dapat dilihat dari tabel di atas bahwa semua kolom yang digunakan untuk penelitian ini mengandung non-null values. Data tersebut memiliki populasi 68 namun kami menggunakan random sampling untuk mengambil sampel 30 baris dari data menggunakan *random state* dengan nilai 0.

3.2 Analisa Data

Variabel dependen (nilai y) yang digunakan dalam penelitian ini adalah jumlah tamu yang menginap di hotel bintang. Untuk mendapatkan jumlah total tamu yang menginap, kami menambahkan jumlah tamu Indonesia dengan jumlah tamu asing (baris 8 dan 9 dari *tabel 1*). Sedangkan variabel independen (nilai x) yang digunakan untuk penelitian ini adalah mean lama inap tamu (per malam), persen tingkat huni kamar, jumlah tempat tidur, jumlah kamar tidur, jumlah akomodasi, kepadatan penduduk (jiwa/km2) dan jumlah perjalanan wisatawan nusantara.

Sebelum menggunakan data untuk membuat model, pertama-tama kita harus membersihkan data untuk memastikan model yang lebih akurat. Hal pertama yang kami lakukan adalah menghapus semua pencilan yang ada di data. Kami menganggap semua nilai yang diatas 3 kali deviasi standar ke distribusi normal sebagai pencilan.

Setelah itu, kami juga menerapkan penskalaan pada data kami, penskalaan adalah metode yang digunakan untuk menormalkan rentang variabel independen. Ini digunakan agar fitur dengan rentang nilai yang lebih tinggi tidak mendominasi model dan juga mengurangi sensitivitas model terhadap perubahan input.

4. Hasil dan Diskusi

4.1 Statistik Deskriptif

Statistik deskriptif merupakan proses analisis statistik yang berfokus pada pengumpulan, penyajian, dan klasifikasi data untuk memberikan gambaran informasi dari keseluruhan data. Metode ini dibagi menjadi 2 bagian, yaitu menurut ukuran pemusatan data(mean, median, dan modus) dan ukuran sebaran data(range, kuartil, varians, standar deviasi, dan lain-lain).

	mean lama inap tamu(malam)	persen tingkat huni kamar	jumlah tempat tidur	jumlah kamar tidur	jumlah akomodasi	kepadatan penduduk (jiwa/km2)	jumlah perjalanan wisatawan nusantara	jumlah tamu total
count	68.000000	68.000000	68.000000	68.000000	68.000000	68.000000	6.800000e+01	6.800000e+01
mean	1.739706	51.290000	13041.014706	8775.823529	86.808824	730.750000	8.047991e+06	2.353951e+06
std	0.294071	7.108257	19393.436979	13686.189647	112.729932	2653.019797	1.228440e+07	3.166043e+06
min	1.260000	36.070000	492.000000	313.000000	3.000000	9.000000	4.915310e+05	6.952000e+04
25%	1.552500	46.595000	1937.000000	1294.500000	18.750000	48.500000	1.907504e+06	3.148575e+05
50%	1.680000	50.580000	5386.000000	3618.500000	42.000000	101.000000	3.206628e+06	9.136050e+05
75%	1.855000	56.177500	15981.500000	9625.000000	93.250000	251.250000	7.764839e+06	2.978440e+06
max	2.910000	71.120000	97099.000000	70146.000000	507.000000	15900.000000	5.208172e+07	1.335323e+07

Tabel 2: Tabel Deskriptif Statistik

4.2 Hasil Modelling

Dari 7 variabel dependen ini, kami menghapus semua variabel yang memiliki nilai VIF tertinggi satu per satu, hingga variabel yang tersisa semuanya memiliki nilai VIF di bawah 5. Berikut adalah nilai VIF dari masing-masing kolom sebelum ada perubahan:

	Var	VIF
0	mean lama inap tamu(malam)	1.206423
1	persen tingkat huni kamar	1.112550
2	jumlah tempat tidur	50.222239
3	jumlah kamar tidur	76.975326
4	jumlah akomodasi	60.256694
5	kepadatan penduduk (jiwa/km2)	3.288676
6	jumlah perjalanan wisatawan nusantara	8.570167

Tabel 3: Variance Inflation Factor dari Kolom - Kolom

Selama proses ini, kami juga memastikan untuk hanya mengambil variabel yang memiliki nilai probabilitas nilai absolut *t* (*p-value t*) dibawah 0,05 karena nilai *p-value t* yang di atas tingkat signifikansi berarti variabel tersebut tidak signifikan terhadap model kami. Berikut adalah tabel yang dengan nilai-nilai parameter untuk model pertama (model sebelum pengurangan kolom).

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.499e+06	7.31e+04	20.494	0.000	1.35e+06	1.65e+06
mean lama inap tamu(malam)	-1.957e+04	6.11e+04	-0.320	0.752	-1.46e+05	1.07e+05
persen tingkat huni kamar	1.284e+05	6.42e+04	2.002	0.058	-4627.539	2.61e+05
jumlah tempat tidur	2.636e+06	3.86e+05	6.823	0.000	1.83e+06	3.44e+06
jumlah kamar tidur	-1.059e+06	5.02e+05	-2.109	0.047	-2.1e+06	-1.78e+04
jumlah akomodasi	-3.156e+05	4.18e+05	-0.755	0.459	-1.18e+06	5.52e+05
kepadatan penduduk (jiwa/km2)	1.624e+05	1.09e+05	1.495	0.149	-6.28e+04	3.88e+05
jumlah perjalanan wisatawan nusantara	6.042e+05	1.64e+05	3.690	0.001	2.65e+05	9.44e+05

Tabel 4: Tabel Hasil Analisis Persamaan Regresi dan P-value

Setelah menghilangkan kolom - kolom yang berpotensi multikolinear dan juga tidak signifikan terhadap model yang digunakan, kami mendapatkan sisa 2 variabel, yaitu jumlah tempat tidur dan jumlah perjalanan wisatawan nusantara.

	Var	p_val	VIF
0	jumlah tempat tidur	7.118215e-13	4.854553
1	jumlah perjalanan wisatawan nusantara	9.590953e-03	4.651378

Tabel 5: Tabel P-value dan VIF Setelah Melakukan Pengurangan Kolom

Menggunakan data terbaru, kami mendapatkan rangkuman dari *Ordinary Least Squares* sebagai berikut:

OLS Regression Results

Dep. Variable:	jumlah tamu total	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.969
Method:	Least Squares	F-statistic:	456.0
Date:	Fri, 30 Apr 2021	Prob (F-statistic):	1.56e-21
Time:	16:49:13	Log-Likelihood:	-430.80
No. Observations:	30	AIC:	867.6
Df Residuals:	27	BIC:	871.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.522e+06	8.08e+04	18.837	0.000	1.36e+06	1.69e+06
jumlah tempat tidur	1.62e+06	1.28e+05	12.668	0.000	1.36e+06	1.88e+06
jumlah perjalanan wisatawan nusantara	3.737e+05	1.34e+05	2.788	0.010	9.87e+04	6.49e+05

Omnibus:	1.550	Durbin-Watson:	1.686
Prob(Omnibus):	0.461	Jarque-Bera (JB):	0.557
Skew:	-0.219	Prob(JB):	0.757
Kurtosis:	3.504	Cond. No.	3.76

Tabel 6: Tabel analisis Ordinary Least Squares setelah Eliminasi Variabel Independen

Penjelasan OLS Regression:

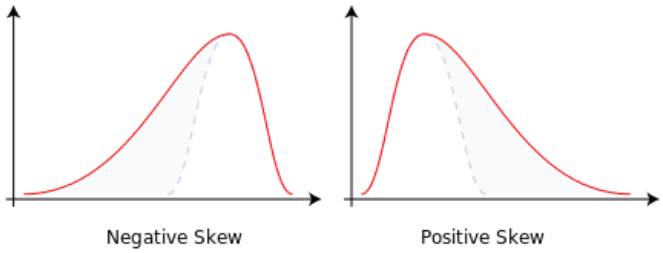
1. *Adj. R-squared* :

   Penggunaan *R-squared* tidak terlalu akurat dikarenakan setiap pertambahan variabel pastinya *R-squared* akan bertambah. Jika kita menambahkan variabel dengan sembarangan dan nilai R-squared meningkat, kita tidak tahu apakah variabel bebas itu berhubungan dengan variabel terikat atau tidak, maka dari itu kami menggunakan nilai *Adj. R-squared*. Pada data yang kami analisis, nilai *Adj. R-squared* adalah 0.969. Hal ini menjelaskan bahwa *Adj. R-squared* pada data kami baik dan variabel independen yang digunakan dapat menjelaskan variabel dependen.

2. *P-value* :

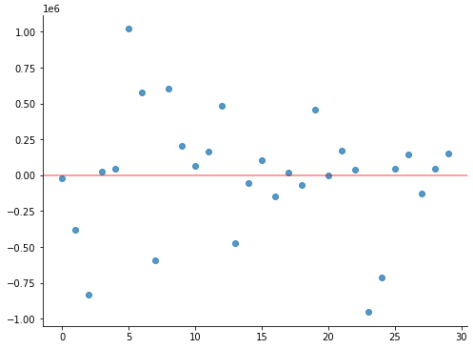
   Nilai *p-value* adalah nilai kesalahan yang didapat peneliti dari hasil perhitungan statistik. Nilai *p-value* biasanya digunakan sebagai indikator untuk menolak atau menerima hipotesis. Jika nilai $p<0.05$ maka hipotesis nol ditolak, sedangkan nilai $p>0.05$ maka hipotesis nol diterima. Berdasarkan hasil OLS diatas, nilai *p-value* untuk variabel independen jumlah tempat tidur dan jumlah perjalanan wisatawan nusantara sudah significant dan dapat diterima. Karena nilai *p-value* ini sudah di bawah 0,05.

3. Kurtosis :
   Kurtosis menandakan kelancipan dari distribusi suatu data. Semakin besar nilai kurtosis maka kurva akan semakin runcing. Nilai referensi kurtosis adalah 3. Jika nilai dari kurtosis lebih dari 3, maka kurva distribusi disebut leptokurtik. Jika nilai dari kurtosis kurang dari 3, maka kurva distribusi disebut platikurtik. Jika nilai dari kurtosis sama dengan 3 kurva berdistribusi normal atau mesokurtik. Dalam kasus ini, kurtosis yang dimiliki bernilai 3.504, sehingga distribusi residu dapat dibilang leptokurtik.
4. *Skewness*
   Skewness adalah ukuran ketidaksimetrisan dalam distribusi nilai. *Skewness* dapat bernilai positif, negatif, dan nol.



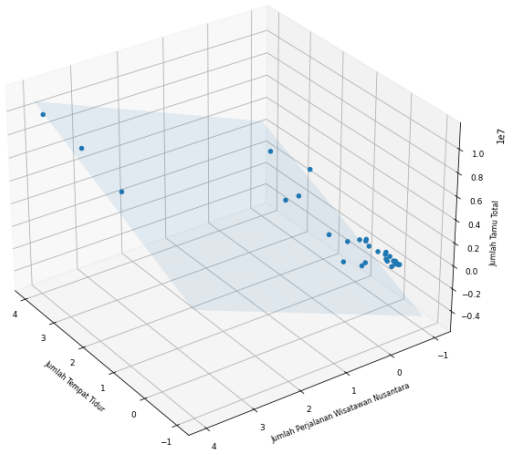
Gambar 1: Positive and Negative Skewness

Nilai skewness dalam data adalah -0.219. Nilai ini menunjukkan bahwa nilai skewness ini negatif.



Gambar 2: Grafik Residual

Kita juga dapat melihat melalui grafik residu di atas bahwa residu kita tidak memiliki pola tertentu, sehingga dapat dibilang bahwa model yang dihasilkan model baik untuk data yang digunakan.



Gambar 3: Grafik Bidang Regresi dengan Target

Karena ada model akhir hanya memiliki dua variabel independen, kita dapat membuat grafik dari bidang yang dibentuk dari persamaan regresi dengan titik - titik yang melambangkan target yang kita inginkan. Grafik tersebut dapat dilihat di atas.

5. Kesimpulan

Berdasarkan data yang kami miliki, kita dapat memodelkan jumlah total tamu di hotel berbintang dengan persamaan regresi linear ganda. Persamaan tersebut adalah:

$$\hat{y} = 1521994.3430552406 + 1620436.3509497817x_1 + 373703.7081082469x_2$$

Dimana $x_1$adalah jumlah kamar tidur di hotel berbintang dan $x_2$adalah jumlah perjalanan wisatawan nusantara. Model ini memiliki *Adjusted R-Squared* 0.969. Variabel independen yang dipilih sudah mengurangi kemungkinan multikolinear dan juga signifikan terhadap populasi.

Daftar Pustaka

Aindhae. (2019a, October 22). *Cara Menghitung Root Mean Squared Error (RMSE) dengan Excel*. Aindhae | My ID & Cuitan Sederhana. https://www.aindhae.com/2019/10/cara-menghitung-root-mean-squared-error.html

Muhammad, E. (2020, July 26). *Sejarah Pariwisata Indonesia Berawal dari Kebijakan Turis Kolonial*. Harapan Rakyat Online. https://www.harapanrakyat.com/2020/07/sejarah-pariwisata-indonesia/

Pezzullo, J. (2016, March 26). *The Symmetry and Shape of Data Distributions Often Seen in Biostatistics*. Dummies. https://www.dummies.com/education/science/biology/the-symmetry-and-shape-of-data-distributions-often-seen-in-biostatistics/

Raharjo, S. (2015, April 8). *Uji Multikolinearitas dengan Melihat Nilai Tolerance dan VIF*. KONSISTENSI. https://www.konsistensi.com/2013/07/uji-multikonieritas-dengan-melihat.html

Skillplus. (2019, December 3). *Multiple Linear Regression – Pendahuluan*. SkillPlus. https://skillplus.web.id/multiple-linear-regression-pendahuluan/

Statistik, K. (2021, July 1). *Koefisien Determinasi pada Regresi Linear*. Konsultan Statistik. https://www.konsultanstatistik.com/2011/07/koefisien-determinasi-pada-regresi.html