ANALISIS DATA

Oleh: Dr Kamree & Mohd Fadil

Kandungan

- ➔ Pengenalan
- ➔ Jenis-Jenis Analisa Data
- ➔ Alat untuk Menganalisa Data
- ➔ Exploratory Data Analysis (EDA)
- ➔ Ringkasan Statistik
- ➔ Skewness & Kurtosis
- ➔ Jenis-jenis carta
- ➔ Pearson Correlation
- ➔ Null & Alt Hypothesis
- ➔ P-value
- ➔ Ordinary Least Square Regression
- ➔ Q&A

Pengenalan

Data Raya

Data raya adalah istilah yang menggambarkan jumlah data yang besar - baik terstruktur maupun tidak terstruktur - yang membanjiri perniagaan setiap hari. Tetapi bukan jumlah data yang penting. Organisasi melakukan data yang penting. Data raya dapat dianalisis untuk mendapatkan pandangan yang membawa kepada keputusan yang lebih baik dan pergerakan perniagaan yang strategik.

Analisis Data

Analisis data adalah proses mengumpulkan, memodelkan, dan menganalisis data untuk mengekstrak wawasan yang mendukung pembuatan keputusan. Terdapat beberapa kaedah dan teknik untuk melakukan analisis bergantung pada industri dan tujuan analisis.

Pembelajaran Mesin

Pembelajaran mesin adalah cabang kecerdasan buatan (AI) yang difokuskan pada pembangunan aplikasi yang belajar dari data dan meningkatkan ketepatannya dari masa ke masa tanpa diprogramkan untuk melakukannya.

Jenis-jenis Analisis Data

Bagaimana analisis data berfungsi?

Analisis data penyelidikan kualitatif terdiri daripada metodologi statistik yang menganalisis iteratif data dan lebih mudah jika dibandingkan berdasarkan nilai. Selain itu, para saintis dan penganalisis data turut mencari corak dalam pelbagai pemerhatian dalam data.

Descriptive

Fokus utama menggunakan teknik analisis data ini adalah untuk mengetahui alasan dan sebab di sebalik kenaikan atau kejatuhan perniagaan yang mahal pada masa lalu dan penyelesaiannya untuk masa depan. Ia kebanyakannya digunakan dalam bidang BI (Business Intelligence) dan data mining

Inferential

Teknik ini bertujuan untuk menguji teori mengenai tingkah laku dan sifat data apa pun bergantung pada beberapa sampel "subjek" atau "hasil" yang diambil dari pemerhatian. Ini mencerminkan matlamat model statistik namun ia masih bergantung terutamanya kepada populasi data dan juga skema pensampelan.

Predictive

Teknik analisis ini menambahkan rasa mengawal tuntutan masa depan dan mengakses risiko. Tambahan lagi, ia mengarahkan pelbagai hasil yang mungkin akan meningkatkan perniagaan utama, terutamanya berkaitan dengan "Apa yang harus dilakukan oleh prestasi perniagaan untuk mencapai tujuan tertentu?"

Alat untuk Menganalisis Data

Apakah fungsi penggunaan alat-alat analisis data?

Penggunaan alat adalah untuk meningkatkan keupayaan untuk melakukan analisis data pada data besar dengan cekap

Bahasa Pengaturcaraan

python™

Modul Pembinaan (AI)

pandas

NumPy

Modul Gambaran Data

seaborn

plotly

matplotlib

Power BI

+ableau

Exploratory Data Analysis (EDA)

Kegunaan EDA

Analisis Data Eksploratori merujuk kepada proses kritis untuk melakukan penyiasatan awal pada data sehingga dapat menemui corak, untuk melihat anomali, untuk menguji hipotesis dan untuk memeriksa andaian dengan bantuan statistik ringkasan dan gambaran grafik.

Anomali/Outliers

Outlier adalah nilai yang terletak sangat jauh dari hampir semua nilai lain.

Outlier juga dikenali sebagai nilai ekstrim.

Outliers dapat memberi kesan dramatik pada rata-rata, sisihan piawai, dan pada skala histogram sehingga hakikat sebarannya benar-benar dikaburkan.

Statistik Ringkasan

Ringkasan statistik meringkaskan dan memberikan maklumat mengenai data sampel anda. Ia memberitahu anda tentang nilai-nilai dalam set data anda. Ini termasuk di mana rata-rata terletak dan sama ada data anda condong.

Hipotesis

Ujian hipotesis menilai dua pernyataan mengenai populasi. Kenyataan tersebut saling eksklusif. Ujian menyimpulkan pernyataan mana yang paling tepat menggambarkan data sampel. Ujian hipotesis membantu kita menentukan kepentingan statistik suatu penemuan.

Ringkasan Statistik

Kegunaan Statistik

Ringkasan statistik meringkaskan dan memberikan maklumat mengenai data sampel anda. Ia memberitahu anda tentang nilai-nilai dalam set data anda. Ini termasuk di mana rata-rata terletak dan sama ada data anda condong.

Measures of Location

Ukuran lokasi memberitahu anda di mana data anda berpusat, atau di mana arah aliran berada.

Jenis-jenis ukuran lokasi:

1. Mean
2. Median
3. Trimmed Mean

Measures of Spread

Langkah penyebaran memberitahu anda bagaimana penyebaran atau variasi set data anda. Ini boleh menjadi maklumat penting. Sebagai contoh, skor ujian yang berada dalam julat 60-90 mungkin dijangkakan sementara skor dalam julat 20-70 mungkin menunjukkan masalah.

Contoh adalah Range, IQR, Variance, Skewness dan Kurtosis

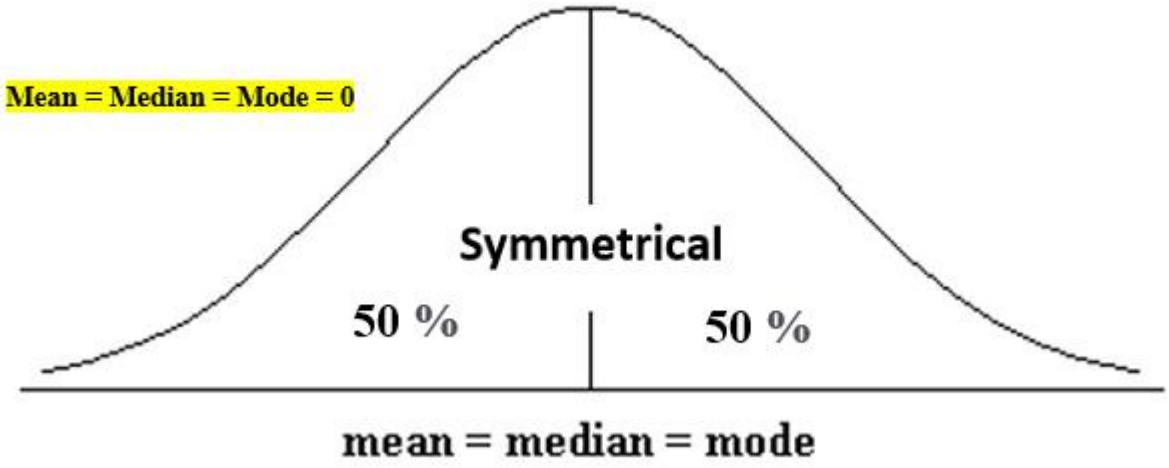
Charts

Carta adalah langkah bagi menggunakan pelbagai jenis carta untuk mencari corak dan informasi di dalam data.

Skewness & Kurtosis

Skewness

Dalam statistik, skewness adalah tahap asimetri yang diperhatikan dalam taburan kebarangkalian yang menyimpang dari taburan normal simetri (loceng lengkung) dalam kumpulan data yang diberikan.

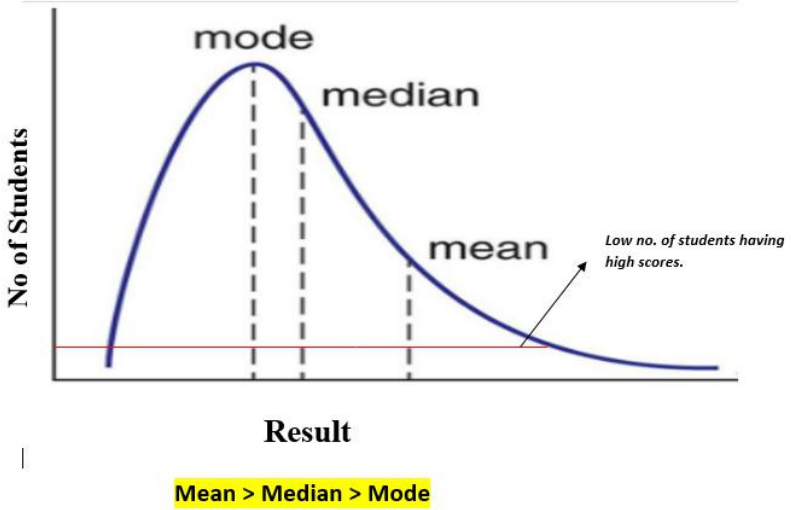


When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations.

Skewness & Kurtosis

Positif atau condong ke kanan

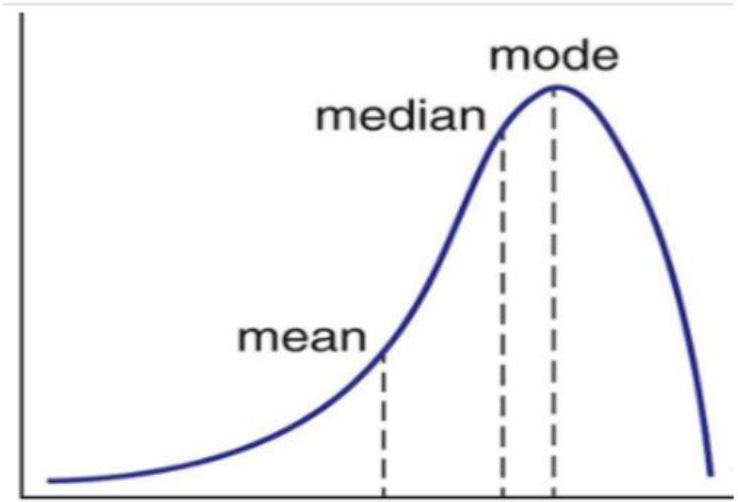
Dalam statistik, sebaran miring positif adalah sejenis pengedaran di mana, tidak seperti data yang disebarkan secara simetri di mana semua ukuran kecenderungan pusat (rata-rata, median, dan mod) sama antara satu sama lain, dengan data miring positif, ukurannya tersebar, yang bermaksud Positif Skewed Distribution adalah sejenis taburan di mana min, median, dan mod pengedaran adalah positif dan bukannya negatif atau sifar.



Skewness & Kurtosis

Negatif atau condong ke kiri

Sebaran condong negatif adalah pembalikan lurus dari taburan condong positif. Dalam statistik, pengedaran condong negatif merujuk kepada model pengedaran di mana lebih banyak nilai adalah di sebelah kanan grafik, dan ekor taburan menyebar di sebelah kiri.



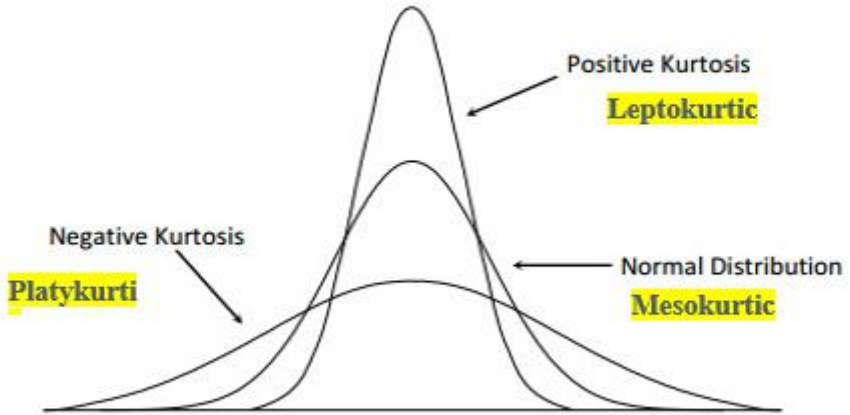
Mode > Median > mean

Skewness & Kurtosis

Kurtosis

Kurtosis merujuk kepada tahap kehadiran data `extreme` dalam pengedaran.

Kurtosis adalah ukuran statistik, sama ada datanya berat atau ringan di ekor dalam taburan normal

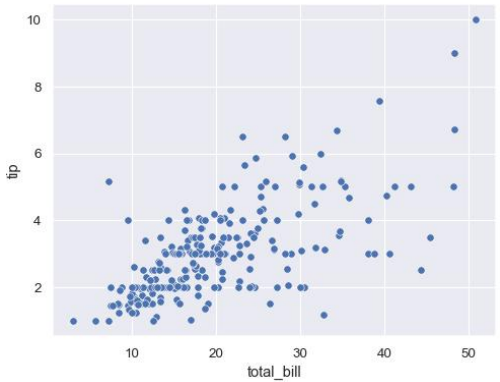


Jenis-jenis Carta

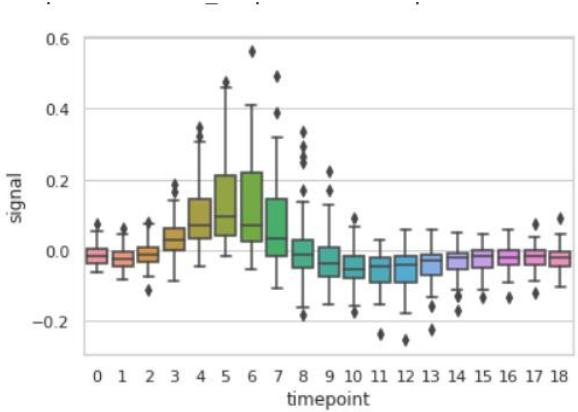
Kegunaan Carta

Carta dapat memberikan gambaran visual kepada penganalisa tentang corak dan informasi didalam data yang dikumpul. Carta merupakan aspek yang penting dalam fasa EDA, kerana carta membantu menganalisa untuk membuat `descriptive analytics`.

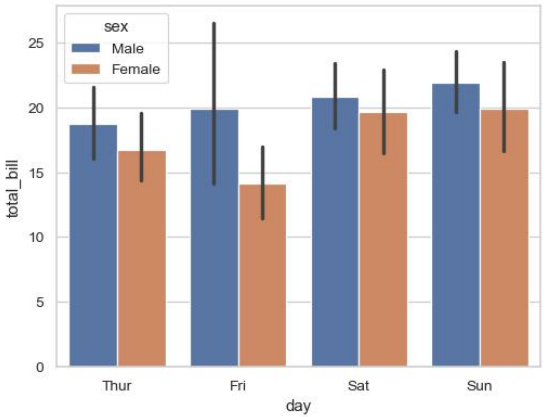
Scatter Plot



Box Plot



Bar Plot

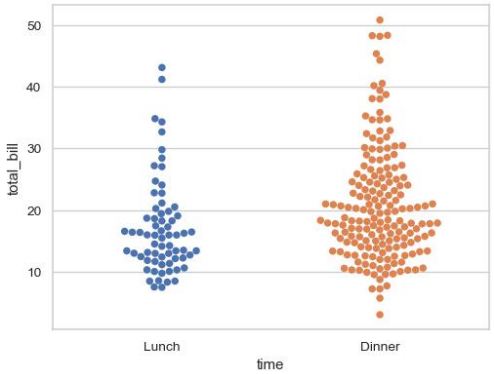


Jenis-jenis Carta (Cont.)

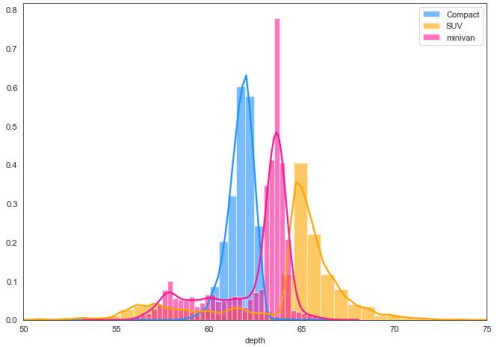
Kegunaan Carta

Carta dapat memberikan gambaran visual kepada penganalisa tentang corak dan informasi didalam data yang dikumpul. Carta merupakan aspek yang penting dalam fasa EDA, kerana carta membantu penganalisa untuk membuat `descriptive analytics`.

Swarm Plot



Histogram Plot



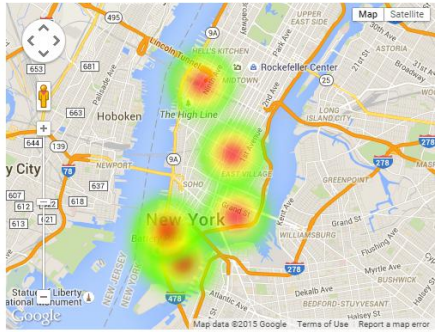
Line Plot

Jenis-jenis Carta (Cont.)

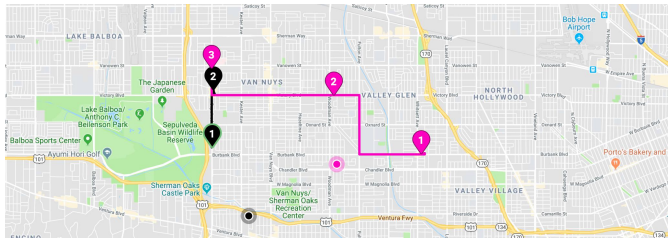
Kegunaan Carta

Carta dapat memberikan gambaran visual kepada penganalisa tentang corak dan informasi didalam data yang dikumpul. Carta merupakan aspek yang penting dalam fasa EDA, kerana carta membantu penganalisa untuk membuat `descriptive analytics`.

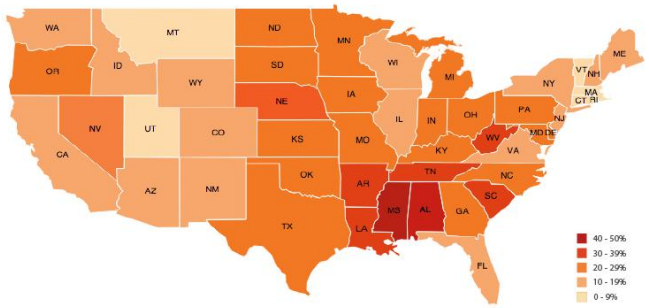
Heat Map Plot



Marker Plot



Choropleth Plot

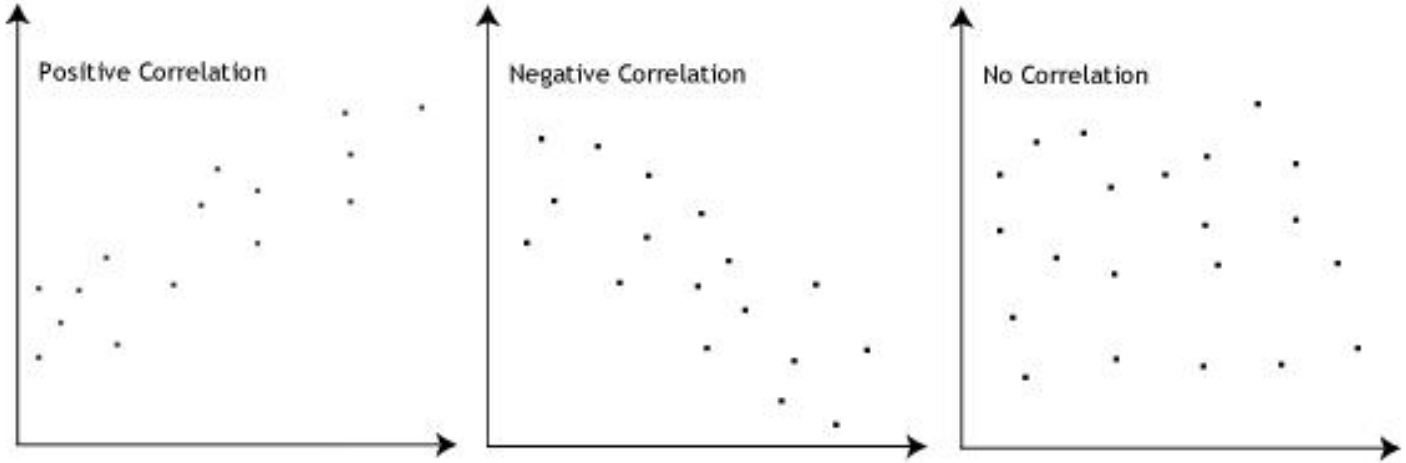


Pearson Correlation

Kegunaan Pearson Correlation

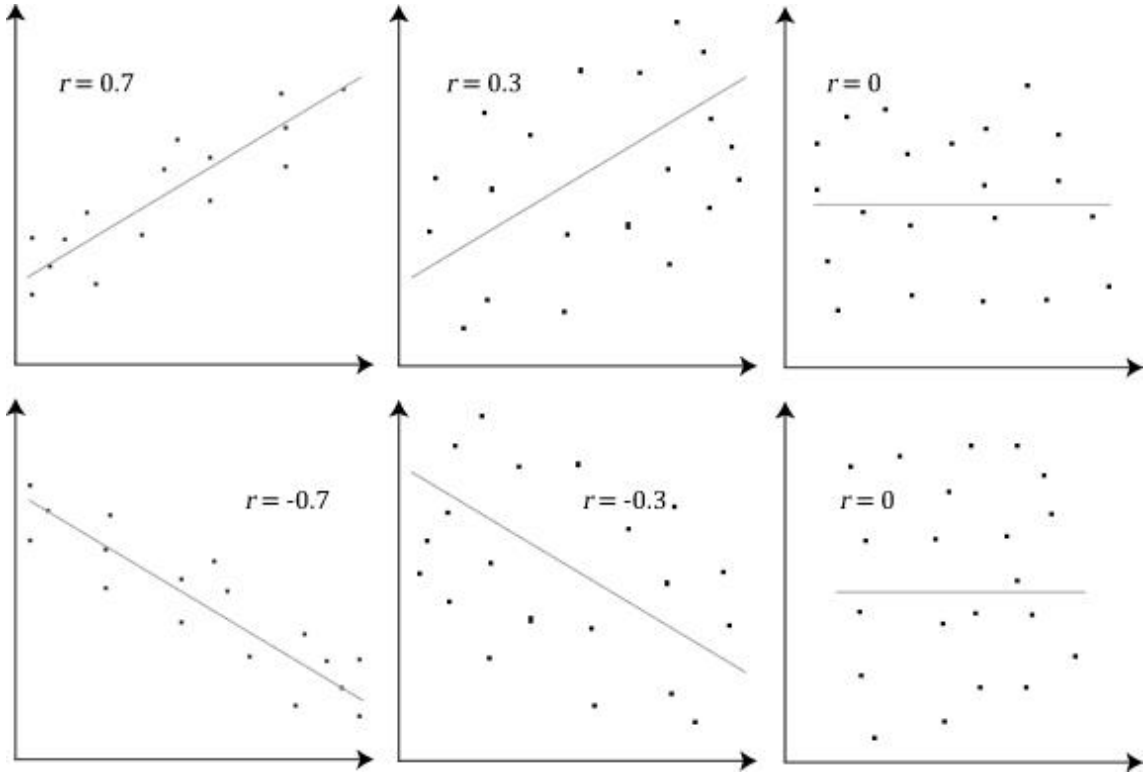
Pearson Correlation adalah ukuran kekuatan hubungan linear antara dua pemboleh ubah.

Pekali korelasi Pearson dapat mengambil julat nilai dari +1 hingga -1. Nilai 0 menunjukkan bahawa tidak ada perkaitan antara dua pemboleh ubah tersebut. Nilai lebih besar daripada 0 menunjukkan hubungan positif, iaitu, apabila nilai satu pembolehubah meningkat, begitu juga nilai pemboleh ubah yang lain. Nilai kurang dari 0 menunjukkan hubungan negatif; iaitu, apabila nilai satu pembolehubah meningkat, nilai pembolehubah lain menurun.



Pearson Correlation

Cara Untuk Menentukan Kekuatan Pearson Correlation



Kekuatan	Positive	Negative
Lemah	.1 -> .3	-0.1 -> -0.3
Kuat	.3 -> .5	-0.3 -> -0.5
Sangat Kuat	.5 -> 1.0	-0.5 -> -1.0

Null & Alt Hypothesis

Null Hypothesis

Hipotesis nol selalunya merupakan tuntutan awal yang berdasarkan analisis sebelumnya atau pengetahuan khusus.

Alt Hypothesis

Hipotesis alternatif adalah yang mungkin anda percayai benar atau berharap dapat membuktikan yang benar.

P-value

Pengenalan kepada P-value

Nilai p digunakan dalam pengujian hipotesis untuk membantu anda menyokong atau menolak hipotesis nol. Nilai p adalah bukti terhadap hipotesis nol. Semakin kecil nilai p, semakin kuat bukti bahawa anda harus menolak hipotesis nol.

Pengenalan kepada Alpha Value

Tahap alpha dikendalikan oleh penyelidik dan berkaitan dengan tahap keyakinan.

P kecil (≤ 0,05), tolak hipotesis nol. Ini adalah bukti kuat bahawa hipotesis nol tidak sah.

P besar (> 0,05) bermaksud hipotesis alternatif lemah, jadi anda tidak menolak nol.

P-value vs Alpha	Kesimpulan
P > 0.10	tidak signifikan
p ≤ 0.10	signifikan sedikit
p ≤ 0.05	signifikan
p ≤ 0.01	sangat ketara

Ordinary Least Square Regression

Kegunaan OLS Regression

Ordinary Least Squares Regression (OLS) adalah kaedah analisis statistik yang menganggarkan hubungan antara satu atau lebih pembolehubah bebas dan pemboleh ubah bersandar; kaedah mengira hubungan dengan meminimumkan jumlah petak dalam perbezaan antara nilai yang diperhatikan dan diramalkan dari pembolehubah bersandar yang dikonfigurasikan sebagai garis lurus.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.426
Model:                            OLS   Adj. R-squared:                  0.426
Method:                 Least Squares   F-statistic:                 1.484e+04
Date:                Wed, 23 Jun 2021   Prob (F-statistic):               0.00
Time:                        07:16:56   Log-Likelihood:            -2.5598e+05
No. Observations:               20000   AIC:                         5.120e+05
Df Residuals:                   19998   BIC:                         5.120e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7745.1829   1749.485      4.427      0.000    4316.048    1.12e+04
median_income  1.601e+05   1314.561    121.823      0.000    1.58e+05    1.63e+05
==============================================================================
Omnibus:                     3494.030   Durbin-Watson:                   1.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6543.462
Skew:                           1.091   Prob(JB):                         0.00
Kurtosis:                       4.757   Cond. No.                         5.71
==============================================================================
```

Ordinary Least Square Regression

Pembolehubah yang ingin diramal

OLS Regression Results

Dep. Variable:	median_house_value	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.426
Method:	Least Squares	F-statistic:	1.484e+04
Date:	Wed, 23 Jun 2021	Prob (F-statistic):	0.00
Time:	07:16:56	Log-Likelihood:	-2.5598e+05
No. Observations:	20000	AIC:	5.120e+05
Df Residuals:	19998	BIC:	5.120e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7745.1829	1749.485	4.427	0.000	4316.048	1.12e+04
median_income	1.601e+05	1314.561	121.823	0.000	1.58e+05	1.63e+05

Omnibus:	3494.030	Durbin-Watson:	1.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6543.462
Skew:	1.091	Prob(JB):	0.00
Kurtosis:	4.757	Cond. No.	5.71

Coef memberitahu anda sama ada terdapat hubungan positif atau negatif antara setiap pembolehubah bebas dan pemboleh ubah bersandar.

R-squared adalah ukuran yang sesuai untuk model regresi linear. Statistik ini menunjukkan peratusan varians dalam pemboleh ubah bersandar yang dijelaskan oleh pemboleh ubah bebas secara kolektif.

Nilai p untuk pekali menunjukkan sama ada hubungan ini signifikan secara statistik.

Q&A

TERIMA KASIH