

Nama : Lisy Septyo Ningrum
Kompi : E
SC ID : DS5E07

ASSIGNMENT 1

SQL FOUNDATION

a) Analisis Geografis

Dengan menggunakan query sebagai berikut:

```
SELECT
  country,
  COUNT(*) AS jumlah_pesanan
FROM
  `bigquery-public-data.thelook_ecommerce.users`
GROUP BY
  country
ORDER BY
  jumlah_pesanan DESC;
```

Berdasarkan query tersebut didapatkan output sebagai berikut:

Row	country	jumlah_pesanan	Row	country	jumlah_pesanan
1	China	33697	12	Poland	220
2	United States	22574	13	Colombia	15
3	Brasil	14767	14	Austria	4
4	South Korea	5317	15	España	3
5	France	4753	16	Deutschland	1

Dari output yang diperoleh, dapat dilihat bahwa negara yang paling sering melakukan order adalah China, US, Brazil, Korea Selatan, dan Prancis. Sedangkan, negara yang paling jarang atau sedikit melakukan order adalah negara Polandia, Colombia, Austria, Spanyol, dan Jerman. Dapat disimpulkan bahwa kebanyakan negara yang jarang melakukan order adalah negara Maju. Hal ini kemungkinan mereka sudah mampu memproduksi barang untuk kebutuhannya sendiri. Mengingat kemajuan teknologi yang ada pada negaranya. Sehingga mereka jarang sekali terdeteksi melakukan transaksi atau order. Berbeda dengan negara yang sering melakukan order. Kebanyakan dari mereka adalah termasuk golongan negara berkembang. Hal ini kemungkinan mereka masih belum mampu memproduksi barang untuk memenuhi kebutuhannya sendiri. Sehingga mereka seringkali terdeteksi melakukan transaksi atau order.

Dengan demikian, kita dapat menyusun strategi dan melakukan promosi lebih besar atau memperluas pasar di wilayah negara berkembang. Karena sebagian besar dari mereka masih belum mumpuni untuk memproduksi barang dalam memenuhi kebutuhannya. Sehingga, ini menjadi peluang untuk memperoleh lebih banyak keuntungan.

b) Pelanggan Terloyal Berdasarkan Pembelian Total

Dengan menggunakan query sebagai berikut:

```
SELECT
  user_id,
  SUM(order_id) AS total_pembelian
FROM
  `bigquery-public-data.thelook_ecommerce.orders`
GROUP BY
  user_id
ORDER BY
  total_pembelian DESC
LIMIT
  10; -- Mengambil 10 pelanggan terbaik
```

Berdasarkan query tersebut didapatkan output sebagai berikut:

Row	user_id	total_pembelian
1	99984	497634
2	99971	497546
3	99966	497518
4	99959	497462
5	99939	497350
6	99918	497230
7	99883	497034
8	99873	496970
9	99850	496858
10	99831	496758

Dari output yang diperoleh, dapat dilihat 10 user_id mana saja yang paling banyak atau paling sering melakukan order. Mereka yang memiliki riwayat order terbanyak, dapat diberikan reward atau gift sebagai wujud apresiasi. Selain itu, ini juga merupakan trik untuk menarik perhatian pelanggan lainnya agar mereka berlomba-lomba untuk membeli barang dan mencetak riwayat order yang banyak. Sehingga, kita dapat memperoleh keuntungan yang lebih banyak lagi.

c) Analisis Events

Dengan menggunakan query sebagai berikut:

```
SELECT
  event_type,
  COUNT(*) AS total_event
FROM
  `bigquery-public-data.thelook_ecommerce.events`
GROUP BY
  event_type;
```

Berdasarkan query tersebut didapatkan output sebagai berikut:

Row	event_type	total_event
1	cart	595759
2	home	87593
3	cancel	124946
4	product	845724
5	purchase	181926
6	department	595850

Dari output yang diperoleh, didapatkan bahwa ada 6 jenis event_type. Diantara keenam event_type tersebut event_type yang paling banyak jumlahnya adalah product. Kemudian department, lalu disusul dengan purchase. Selain itu, kita juga bisa melihat jumlah cancel yang cukup banyak. Ini menandakan perlu adanya evaluasi untuk mengurangi jumlah cancel. Agar pembelian semakin meningkat dan jumlah cancel semakin berkurang. Beberapa hal yang dapat di evaluasi, seperti lama pemrosesan barang, lama pengiriman, dan lain sebagainya.

d) Analisis Waktu Transaksi

Dengan menggunakan query sebagai berikut:

```
SELECT
  EXTRACT(YEAR FROM delivered_at) AS tahun,
  EXTRACT(MONTH FROM delivered_at) AS bulan,
  COUNT(*) AS jumlah_transaksi
FROM
  `bigquery-public-data.thelook_ecommerce.order_items`
GROUP BY
  tahun, bulan
ORDER BY
  tahun, bulan;
```

Berdasarkan query tersebut didapatkan output sebagai berikut:

Row	tahun	bulan	jumlah_transaksi	Row	tahun	bulan	jumlah_transaksi
1	nuli	nuli	117007	44	2022	7	1574
2	2019	1	6	45	2022	8	1594
3	2019	2	17	46	2022	9	1684
4	2019	3	56	47	2022	10	1885
5	2019	4	72	48	2022	11	1881
6	2019	5	95	49	2022	12	2159
7	2019	6	122	50	2023	1	2132

Dari output yang diperoleh, dapat dilihat bahwa dari awal tahun 2019 hingga awal tahun 2023, jumlah transaksi atau order semakin banyak. Ini menunjukkan sebuah kemajuan pesat dalam bisnis dan mencari keuntungan. Dengan menjaga kepercayaan dari pelanggan, dan juga senantiasa melayani pesanan dengan sepenuh hati diharapkan jumlah orderan yang masuk semakin banyak kedepannya.

e) Analisis Kategori Produk Terpopuler

Dengan menggunakan query sebagai berikut:

```
SELECT
  category,
  COUNT(*) AS total_penjualan
FROM
  `bigquery-public-data.thelook_ecommerce.products`
GROUP BY
  category
ORDER BY
  total_penjualan DESC;
```

Berdasarkan query tersebut didapatkan output sebagai berikut:

Row	category	total_penjualan
1	Intimates	2363
2	Jeans	1999
3	Tops & Tees	1868
4	Fashion Hoodies & Sweatshirts	1866
5	Swim	1798
6	Sleep & Lounge	1771
7	Shorts	1765
8	Sweaters	1737

Dari output yang diperoleh, dapat dilihat bahwa produk yang paling laris terjual 5 diantaranya adalah intimates, jeans, tops & tees, fashion hoodies & sweatshirts, dan juga swim. Dari penjualan produk terlaris tersebut, kita dapat senantiasa menjaga kualitas dan kuantitas barang tersebut. Agar pembeli senantiasa puas dengan produk yang ada. Selain itu, untuk produk yang penjualannya masih rendah, perlu adanya evaluasi penyebab rendahnya minat pembelian produk. Agar produk yang minat belinya rendah ini juga akan ikut naik penjualannya. Sehingga dapat meningkatkan keuntungan pula.

PYTHON FOUNDATION

Link Google Colab :

<https://colab.research.google.com/drive/1QL0QfZpSljCmO85SMlvT5ZUwQnx4SPfb?usp=sharing>

STATISTICS

Peluang (Probability) adalah nilai peluang dari suatu peristiwa yang akan terjadi pada variabel acak dengan nilai 0 sampai 1. Variabel acak adalah variabel yang berisikan data numerik dari suatu peristiwa setelah dilakukannya beberapa percobaan/survei. Ada beberapa macam propability, diantaranya :

a. Teorema Bayes

Dengan menggunakan teorema Bayes, dapat dicari peluang teradinya A, dengan syarat B telah terjadi. Disini, B adalah bukti, dan A adalah hipotesis. Persamaan bayes :

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Keterangan :

$P(A|B)$ = Peluang A jika B diketahui benar

$P(A|B)$ = Peluang B jika A benar

$P(A), P(B)$ = Probabilitas independen A dan B

b. Binomial

Ada beberapa syarat kejadian binomial, yaitu :

- Percobaan dilakukan sebanyak n kali.
- Setiap percobaan memiliki 2 kemungkinan hasil.
- Kemungkinan hasil dari setiap percobaan adalah sama.
- Hasil dari satu percobaan tidak mempengaruhi hasil percobaan yang lain (saling independen).

Berikut adalah persamaan binomial :

$$P(X = x) = \frac{n!}{(n-x)! \cdot x!} p^x \cdot q^{n-x}$$

Keterangan :

$P(X = x)$ = Peluang dari nilai variabel random

p = peluang sukses

q = peluang gagal

n = jumlah percobaan

c. Poisson

Merupakan probabilitas yang bertujuan untuk menghitung peluang terjadinya suatu peristiwa langka dalam suatu periode waktu/daerah tertentu.

Hipotesis adalah sebuah pernyataan sementara/praduga peneliti terhadap masalah yang dihadapi.

Ada 2 jenis hipotesis, yaitu :

1) Hipotesis Nol (H_0)

Hipotesis yang diasumsikan sebagai "status quo" atau tidak adanya efek atau perbedaan. Hipotesis ini berisi pernyataan yang akan diuji atau dibantah. Selalu ternotasi dengan tanda [=], [\leq], atau [\geq].

2) Hipotesis Alternatif (H_a) atau (H_1)

Hipotesis yang menyatakan adanya efek, perbedaan, atau hubungan yang ingin diuji. Tidak pernah bernotasi [=], [\leq], atau [\geq].

Kesalahan jenis I dan II

Peluang	H_0 benar	H_0 salah
Menolak H_0	Keputusan salah α (Kesalahan Jenis I)	Keputusan Benar
Menerima H_0	Keputusan Benar	Keputusan Salah β (Kesalahan Jenis II)

• Kesalahan Jenis I

Penolakan hipotesis nol yang benar.

• Kesalahan Jenis II

Penerimaan hipotesis nol yang salah.

Distribusi diskrit dan distribusi kontinu adalah dua konsep dasar dalam statistik yang menggambarkan cara data atau variabel acak terdistribusi. Berikut adalah perbedaan antara keduanya:

1) Distribusi Diskrit

- Distribusi diskrit digunakan untuk mewakili variabel diskrit atau kategorikal, yang hanya dapat mengambil sejumlah nilai tertentu yang terpisah. Contoh variabel diskrit termasuk jumlah koin yang dilempar (1, 2, 3, ...), jumlah siswa dalam kelas (tidak dapat 1,5 siswa), atau jumlah kejadian tertentu dalam satu periode waktu.
- Nilai-nilai dalam distribusi diskrit terpisah satu sama lain, dan ada celah antara nilai-nilai tersebut. Probabilitasnya dapat dihitung untuk setiap nilai individu. Distribusi ini sering direpresentasikan dengan tabel atau grafik batang (bar chart).
- Distribusi diskrit merujuk pada distribusi probabilitas dari variabel acak yang mengambil nilai-nilai diskrit atau terbatas pada sejumlah nilai yang terpisah.
- Nilai-nilai dalam distribusi diskrit adalah diskret atau terbatas pada angka-angka tertentu. Contoh umum adalah distribusi Poisson (digunakan untuk menghitung peristiwa yang jarang terjadi dalam interval waktu tertentu) dan distribusi binomial (digunakan untuk menghitung jumlah keberhasilan dalam serangkaian percobaan Bernoulli yang independen).

2) Distribusi Kontinu

- Distribusi diskrit digunakan untuk mewakili variabel diskrit atau kategorikal, yang hanya dapat mengambil sejumlah nilai tertentu yang terpisah. Contoh variabel diskrit termasuk jumlah koin yang dilempar (1, 2, 3, ...), jumlah siswa dalam kelas (tidak dapat 1,5 siswa), atau jumlah kejadian tertentu dalam satu periode waktu.
- Nilai-nilai dalam distribusi diskrit terpisah satu sama lain, dan ada celah antara nilai-nilai tersebut. Probabilitasnya dapat dihitung untuk setiap nilai individu. Distribusi ini sering direpresentasikan dengan tabel atau grafik batang (bar chart).
- Distribusi kontinu merujuk pada distribusi probabilitas dari variabel acak yang dapat mengambil nilai-nilai di seluruh rentang nyata, yang berarti ada tak terbatas jumlah nilai yang mungkin.
- Nilai dalam distribusi kontinu adalah kontinu dan membentuk kurva yang halus. Contoh yang paling terkenal adalah distribusi normal (distribusi Gauss) yang sering digunakan dalam statistik.

Contoh Distribusi:

1) Contoh Distribusi Diskrit, yaitu Distribusi Poisson

Distribusi Poisson digunakan ketika kita ingin menghitung berapa kali peristiwa yang jarang terjadi akan terjadi dalam interval waktu atau ruang tertentu. Contoh, jika kita ingin menghitung berapa kali telepon berdering di kantor selama satu jam.

2) Contoh Distribusi Kontinu, Distribusi Normal (Gaussian)

Distribusi normal adalah distribusi kontinu yang membentuk kurva lonceng. Ini digunakan untuk menggambarkan banyak fenomena dalam kehidupan sehari-hari, seperti tinggi badan manusia, berat badan, atau hasil tes IQ. Ketika data mengikuti distribusi normal, kita dapat menggunakan properti distribusi ini untuk melakukan analisis statistik lebih lanjut.

Uji-t

Uji-t adalah metode statistik yang digunakan untuk membandingkan dua kelompok atau lebih dan menguji apakah perbedaan antara kelompok-kelompok tersebut adalah signifikan secara statistik (Walpole, 2012). Ini adalah alat statistik yang sering digunakan untuk menguji apakah rata-rata dari

dua kelompok berbeda secara signifikan atau tidak. Uji-t memperhitungkan variasi dalam data dan mengukur seberapa besar perbedaan antara rata-rata kelompok dibandingkan dengan variasi dalam kelompok tersebut.

Pemahaman sederhana tentang uji-t adalah bahwa ini membantu dalam menentukan apakah perbedaan yang diamati antara dua kelompok adalah hasil kebetulan atau apakah perbedaan tersebut nyata. Uji-t menghasilkan nilai statistik yang disebut t-statistik dan nilai probabilitas (p-value) yang menunjukkan seberapa signifikan perbedaan tersebut. Semakin kecil p-value, semakin signifikan perbedaan tersebut.

Misalnya, menggunakan uji-t untuk membandingkan rata-rata nilai tes antara dua kelompok siswa dan menentukan apakah perbedaan dalam rata-rata tersebut adalah signifikan secara statistik atau hanya hasil kebetulan.

Berikut adalah langkah-langkah melakukan uji-t (Walpole, 2012):

1. Hipotesis Nol (H_0) dan Hipotesis Alternatif (H_1)
Menyusun hipotesis nol yang menyatakan bahwa tidak ada perbedaan yang signifikan antara dua kelompok yang dibandingkan. Hipotesis alternatif menyatakan sebaliknya, yaitu ada perbedaan yang signifikan.
2. Perhitungan Statistik
Menghitung t-statistik berdasarkan data yang dimiliki. T-statistik ini mengukur seberapa besar perbedaan antara dua kelompok relatif terhadap variabilitas dalam kelompok-kelompok tersebut.
3. Uji Signifikansi
Membandingkan nilai t-statistik dengan nilai kritis dari tabel distribusi t. Jika t-statistik lebih besar dari nilai kritis, maka dapat menolak hipotesis nol dan menyatakan bahwa perbedaan antara dua kelompok adalah signifikan secara statistik.
4. Interpretasi Hasil
Jika menolak hipotesis nol, dapat menyimpulkan bahwa perbedaan antara dua kelompok adalah signifikan secara statistik. Jika tidak, tidak memiliki cukup bukti statistik untuk menyatakan adanya perbedaan yang signifikan.

DAFTAR PUSTAKA

Walpole, R. e. (2012). *Probability & Statistics for Engineers & Scientists* (9th ed.). Boston: Pearson Education.

LAMPIRAN SQL FOUNDATION

Analisis Geografis

Query :

```
2023-09-05 11:31:23 SQL TUGAS 1 [RUN]

1 #Analisis Geografis
2 SELECT
3   country,
4   COUNT(*) AS jumlah_pesanan
5 FROM
6   `bigquery-public-data.thelook_ecommerce.users`
7 GROUP BY
8   country
9 ORDER BY
10  jumlah_pesanan DESC;
11
```

Output :

Query results

JOB INFORMATION		RESULTS	JSON	EXEC
Row	country	jumlah_pesanan		
1	China	33929		
2	United States	22305		
3	Brasil	14724		
4	South Korea	5216		
5	France	4814		
6	United Kingdom	4568		
7	Germany	4124		
8	Spain	4079		

Pelanggan Terloyal Berdasarkan Pembelian Total

Query:

```
2023-09-05 11:31:23 SQL TUGAS 1 [RUN]

12
13 #Pelanggan Terbaik Berdasarkan Pembelian Total
14 SELECT
15   user_id,
16   SUM(order_id) AS total_pembelian
17 FROM
18   `bigquery-public-data.thelook_ecommerce.orders`
19 GROUP BY
20   user_id
21 ORDER BY
22   total_pembelian DESC
23 LIMIT
24   10; -- Mengambil 10 pelanggan terbaik
25
```

Output:

Query results

JOB INFORMATION		RESULTS	JSON
Row	user_id	total_pembelian	
1	99983	499062	
2	99973	499006	
3	99938	498858	
4	99914	498718	
5	99911	498694	
6	99895	498598	
7	99891	498566	
8	99831	498274	
9	99790	498090	
10	99765	497970	

Analisis Events

Query:

```
24 2023-09-11 10:36:35 LS-1_ query 2 RUN
25 #Analisis Events
26 SELECT
27   event_type,
28   COUNT(*) AS total_event
29 FROM
30   `bigquery-public-data.thelook_ecommerce.events`
31 GROUP BY
32   event_type;
```

Output:

Query results

JOB INFORMATION		RESULTS	JSON	EXEC
Row	event_type	total_event		
1	cart	592288		
2	home	87211		
3	cancel	124659		
4	product	843008		
5	purchase	181004		
6	department	592994		

Analisis Waktu Transaksi

Query:

```
34 2023-09-05 11:31:23 SQL TUGAS 1 RUN
35 #Analisis Waktu Transaksi
36 SELECT
37   EXTRACT(YEAR FROM delivered_at) AS tahun,
38   EXTRACT(MONTH FROM delivered_at) AS bulan,
39   COUNT(*) AS jumlah_transaksi
40 FROM
41   `bigquery-public-data.thelook_ecommerce.order_items`
42 GROUP BY
43   tahun, bulan
44 ORDER BY
45   tahun, bulan;
```

Output:

Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION
Row	tahun	bulan	jumlah_transaksi	
1	nul	nul	117731	
2	2019	1	14	
3	2019	2	19	
4	2019	3	40	
5	2019	4	68	
6	2019	5	81	
7	2019	6	129	
8	2019	7	120	
9	2019	8	162	
10	2019	9	163	
11	2019	10	203	

Analisis Kategori Produk Terpopuler

Query:

```
46 2023-09-11 10:36:35 LS-1_ query 2 RUN
47 #Analisis kategori produk terpopuler
48 SELECT
49   category,
50   COUNT(*) AS total_penjualan
51 FROM
52   `bigquery-public-data.thelook_ecommerce.products`
53 GROUP BY
54   category
55 ORDER BY
56   total_penjualan DESC;
```

Output:

Query results

JOB INFORMATION		RESULTS	JSON	EXEC
Row	category	total_penjualan		
1	Intimates	2363		
2	Jeans	1999		
3	Tops & Tees	1868		
4	Fashion Hoodies & Sweatshirts	1866		
5	Swim	1798		
6	Sleep & Lounge	1771		
7	Shorts	1765		
8	Sweaters	1737		
9	Accessories	1559		
10	Active	1432		
11	Outerwear & Coats	1420		
12	Underwear	1088		