Nama : Erda Putriana
ID DS : A23
Mentor : Tantut Wahyu

SQL Foundation

1. Top Produk yang Paling Banyak Dibeli

```
SELECT name, COUNT(*) as jumlah_pembelian
FROM `bigquery-public-data.thelook_ecommerce.products`
GROUP BY name
ORDER BY jumlah_pembelian DESC
LIMIT 5
```

Insight :
informasi tentang produk-produk yang paling sering dibeli oleh pelanggan. Informasi yang didapatkan membantu mengidentifikasi produk-produk yang paling diminati oleh pelanggan.

2. Negara dengen jumlah order paling sedikit

```
SELECT country, COUNT(DISTINCT u.id)order_count FROM `bigquery-public-
data.thelook_ecommerce.users` u INNER JOIN `bigquery-public-
data.thelook_ecommerce.orders` o ON u.id=o.user_id GROUP BY country ORDER BY
order_count ASC LIMIT 1
```

Insight :
negara dengan jumlah pesanan paling sedikit dalam data e-commerce, membantu mengidentifikasi peluang bisnis yang potensial di negara tersebut.

3. Tingkat Pembatalan Pesanan

```
SELECT
  o.status
FROM
  `bigquery-public-data.thelook_ecommerce.orders` o
 WHERE o.status = "Cancelled"
```

Insight :
Menghitung tingkat pembatalan pesanan dapat memberikan gambaran mengenai kepuasan pelanggan dan potensi masalah operasional yang perlu ditangani.

4. Kategori penjualan teratas

```
SELECT p.category AS Category_product,
       COUNT(p.category) AS TotalOrders
FROM `bigquery-public-data.thelook_ecommerce.products` p
       INNER JOIN bigquery-public-data.thelook_ecommerce.order_items` o
ON p.id = o.product_id
       GROUP BY p.category
       ORDER BY TotalOrders DESC
LIMIT 10
```

Insight :
kategori penjualan teratas berdasarkan jumlah pesanan, memberikan wawasan tentang preferensi pelanggan terhadap jenis produk tertentu.

5. Jumlah order oleh male

```
SELECT
COUNT(DISTINCT o.user_id)pembelian_male FROM `bigquery-public-
data.thelook_ecommerce.orders`o WHERE o.gender = 'M'
```

Insight :
jumlah pembelian yang dilakukan oleh pelanggan pria, memberikan pemahaman tentang tingkat partisipasi pembelian pria dalam platform tersebut.

6. Traffic Source

```
SELECT
  u.traffic_source
FROM
  `bigquery-public-data.thelook_ecommerce.users` u
 WHERE u.traffic_source = 'Facebook'
```

Insight :
Mengetahui pelanggan (pengguna) dalam menemukan toko e-commerce tertentu 'Facebook'.

7. Wilayah pemesanan yang paling banyak

```
SELECT
e.state,
e.city
FROM `bigquery-public-data.thelook_ecommerce.events`  AS e
```

Insight :
Dengan mengetahui wilayah asal pesanan, dapat menentukan justifikasi kebutuhan pemasaran dan/atau dukungan operasional terkait penawaran dan permintaan di wilayah tersebut.

8. Pola Pembelian Menurut Waktu (Bulan/Tahun)

```
SELECT EXTRACT(MONTH FROM created_at) as bulan, EXTRACT(YEAR FROM created_at) as
tahun, COUNT(*) as jumlah_pembelian
FROM `bigquery-public-data.thelook_ecommerce.orders`
GROUP BY bulan, tahun
ORDER BY tahun, bulan
```

Insight :
melihat pola pembelian seiring berjalannya waktu, dengan menghitung jumlah pembelian per bulan dan tahun.

9. Produk yang Paling Banyak Terjual dalam Setiap Kategori

```
SELECT category, name, SUM(num_of_item) as jumlah_terjual
FROM `bigquery-public-data.thelook_ecommerce.products` p
JOIN `bigquery-public-data.thelook_ecommerce.orders` o
ON p.id = o.order_id
GROUP BY category, name
ORDER BY category, jumlah_terjual DESC;
```

Insight :
mengetahui produk-produk terlaris dalam setiap kategori.

10. Rata-rata Jumlah Produk dalam Setiap Pesanan

```
SELECT AVG(num_of_item) as rata_rata_produk_per_pesanan
FROM `bigquery-public-data.thelook_ecommerce.orders`;
```

Insight :
Ini akan memberikan gambaran tentang seberapa suksesnya bisnis dengan menghitung total pendapatan dari semua penjualan.

Python Foundation

[Assignment 1 DS.ipynb - Colaboratory (google.com)]

Statistic

Probabilitas (peluang)

1+1=2 jika di ilmu matematika itu pasti, jika di statistic berbicara tentang sebagai seberapa yakin kita terhadap jawaban kita dan peluang terhadap jawaban kita. Ketika kita berbicara tentang probabilitas dai tidak akan keluar dari bentuknya (contoh : coin).

Basic probabilitas

Chance events

- Adalah menghitung kemungkinan.
- Isi angka nya antara 0 dan 1, semakin kecil probabilitasnya semakin kecil peluangnya.
- Konsep statistic yang paling dipakai dalam data science.
- True probabilitas 50/50 kemungkinan output
- Semakin banyak percobaan, semakin mendekati true probabilitas

Expectation

- Ketika melakukan percobaan yang banyak, maka angka dadu mendekati ekspetasi

Variance

- Ketika melakukan percobaan pada kartu yang banyak, maka kartu yang dikeluarkan semakin mendekati varian

Probalititas recap

- Percobaan = menghasilkan event
- Hasil = peristiwa percobaan
- Event = hasil dari satu percobaan
- Peluang = nilai peluang (antara 0 dan 1)
- Variabel acak = data numerik dari event

Conditional probabilitas

- Probabilitas A harus terjadi dulu dari probabilitas B
- Ada actual dan expected
- Bayes theorem

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

A, B = events
$P(A|B)$ = probability of A given B is true
$P(B|A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B

Probabilitas diasease positif harus bersyarat

$$P(\text{Disease}|+) = \frac{P(+|\text{Disease})P(\text{Disease})}{P(+)}$$

P(diasease) adalah probabilitas prior

Distribusi

- Adalah penyebaran data
- Memiliki titik pemusatan, panjangnya dipengaruhi oleh variansinya
- Bisa digambarkan oleh histogram
- Random variabel = di generate dari empirical distribusi
- Long tail distribusi

- + ada dua tail tergantung distribusi apa
- Perbedaan head
- + high impact, popular, few in number, mainstream
- Long tail
- +low impact, niche, many in number, obsuce

Distriibusi diskrit = dapat dihitung atau berkelanjutan
- Probabilitas mass function
- Camulative function
- Ketika dijumlahkan harus hasilnya 1
- A Bernouli random variable takes the value 1 with probability of p and value 0 with probability of 1- p. (dilakuan sekali saja)
- Binomial = probabilitas barnouli secara berulang

Distribusi kontinu = tipe data numerik (interfall dan kontinal)
- Ketika miu =0 dan sigma = 1 itu disebut distribusi normal standar
- Ketika datanya semakin lancip homogennya, maka semakin besar
- Ketika datanya semakin landai / homogenya, maka semakin kecil
- Satu sigma secara normal 34.13 x 2 (distribusi normal sandar)

Hipotesis (menduga)
Hypothesis testing = prosedur statistik yang digunakan untuk menguji hipotesis atau asumsi tentang populasi berdasadrkan sampel data
1. Null hypothesis = tidak adanya efek atau perbedan. Hipotesis nol berisi pernyataan yang akan diuji atau dibantah
2. Alternative hypothesis = menyatakan adanya efek, perbedaan, atau hubungan yang ingin diuji.

Kesalahan hipotesis
Type I & II errors
- Type I = membuat klaim tentang adanya efek atau perbedaan, pafahal tidak ada efek perbedaan yang signifikan dalam populasi.
- Type II = tidak menerima klaim

Confidence interval = rentang nilai yang digunakan untuk mengukur seberapa yakin terhdap parameter statistik tertentu
P value = ukuran statistic yang digunakan untuk mengukur tingkat bukti terhdap hipotesis nol dalam pengujian hipotesis
T-test = jenis uji statistik yang digunakan untuk membandingkan rata-rata dua kelompok atau sampel untuk melihat apakah ada perbedaan
- Independen sample t-test : digunakan untuk membandingkan rata-rata dua kelompok yang berbeda secara independen.

Contoh : ingin mengetahui apakah rata-rata skor ujian siswa laki-laki berbeda secara signifikan dari rata-rata skor siswa perempuan.
- Paired sample t-test : digunakan ketika memiliki data yang diambil dari dua waktu atau dua kondisi yang berbeda pada kelompok yang sama.

Contohnya : ingin mengetahui apakah pembelajaran track data science memiliki efek yang signifikan pada mahasiswa sebelum dan sesudah ikut startup campus
- One-sample t-test : digunakan untuk membandingkan rata-rata sample dengan nilai tertenru yang diketahui

Contoh : ingin tahu apakah rata-rata waktu pengerjaan suatu tugas oleh sekelompok pekerja sana dengan waktu yang diharapkan.