Soal Ujian - **Data Analytics & Visualization**

Purwadhika
Startup and Coding School

Soal 1 - `PyMySQL` - Sakila Database (40 poin)

Database `sakila` merupakan *sample dummy database* yang menyimpan informasi seputar toko rental DVD. Apabila belum ada database `sakila` di laptop Anda, ikuti panduan di laman ini.

Soal : Buatlah file *jupyter* (*.ipynb*) dan buat *connection* Python dengan server MySQL Anda, lalu gunakan database *sakila*. `Tuliskan langkah-langkah/query MySQL dan tampilkan hasilnya berupa DataFrame.` Anda **dilarang** membuat database baru, merubah struktur table, membuat view atau segala bentuk tindakan yang mengubah struktur database.

1. Tampilkan daftar **10 film komedi dengan durasi tersingkat**. Urutkan data berdasarkan film dengan durasi terpendek. Kolom yang diwajibkan tampil adalah **title**, **category** dan **length**. Output yang diharapkan:

```
+------------------+----------+--------+
| title            | category | length |
+------------------+----------+--------+
| DOWNHILL ENOUGH  | Comedy   |     47 |
| HEAVEN FREEDOM   | Comedy   |     48 |
| PARADISE SABRINA | Comedy   |     48 |
| HURRICANE AFFAIR | Comedy   |     49 |
| LION UNCUT       | Comedy   |     50 |
| ZORRO ARK        | Comedy   |     50 |
| CLOSER BANG      | Comedy   |     58 |
| AIRPLANE SIERRA  | Comedy   |     62 |
| LONELY ELEPHANT  | Comedy   |     67 |
| DOOM DANCING     | Comedy   |     68 |
+------------------+----------+--------+
```

2. Tampilkan daftar lengkap **kategori film beserta jumlah film tiap kategori & rata-rata harga sewa DVD film tiap kategori**. Urutkan data dari kategori dengan jumlah film terbanyak. Kolom yang diwajibkan ada minimal adalah **kategori**, **jumlah film** dan **rata-rata harga sewa**. Output yang diharapkan:

```
+-------------+-------------+---------------+
| kategori    | jumlahMovie | rataHargaSewa |
+-------------+-------------+---------------+
| Foreign     |          73 |      3.099589 |
| Sports      |          73 |      3.099589 |
| Family      |          69 |      2.758116 |
```

```
| Documentary |          68 |      2.666471 |
| Animation   |          66 |      2.808182 |
| Action      |          64 |      2.646250 |
| New         |          63 |      3.116984 |
| Drama       |          61 |      2.990000 |
| Games       |          61 |      3.252295 |
| Sci-Fi      |          61 |      3.219508 |
| Children    |          60 |      2.890000 |
| Comedy      |          58 |      3.162414 |
| Classics    |          57 |      2.744386 |
| Horror      |          56 |      3.025714 |
| Travel      |          56 |      3.275714 |
| Music       |          51 |      2.950784 |
+-------------+-------------+---------------+
```

3. Tampilkan daftar **10 aktor/aktris yang paling banyak membintangi film**. Kolom yang ditampilkan minimal: **id aktor**, **nama depan**, **nama belakang** dan **jumlah film yang dibintangi** kemudian urutkan dari aktor/aktris yang membintangi film terbanyak. Output yang diharapkan:

```
+----------+------------+-------------+--------------+
| actor_id | first_name | last_name   | jumlah_Movie |
+----------+------------+-------------+--------------+
|      107 | GINA       | DEGENERES   |           42 |
|      102 | WALTER     | TORN        |           41 |
|      198 | MARY       | KEITEL      |           40 |
|      181 | MATTHEW    | CARREY      |           39 |
|       23 | SANDRA     | KILMER      |           37 |
|       81 | SCARLETT   | DAMON       |           36 |
|      158 | VIVIEN     | BASINGER    |           35 |
|      144 | ANGELA     | WITHERSPOON |           35 |
|      106 | GROUCHO    | DUNST       |           35 |
|       60 | HENRY      | BERRY       |           35 |
+----------+------------+-------------+--------------+
```

4. Dari soal sebelumnya diketahui `Gina Degeneres` merupakan aktris yang paling banyak membintangi film, dengan total **42** judul film. Kategori film apakah yang paling banyak dibintanginya? Untuk mengetahuinya, tampilkan daftar **kategori film beserta jumlah film yang pernah dibintangi oleh** `Gina Degeneres`. Kolom yang diwajibkan ada yaitu **kategori film** dan **jumlah film yang dibintangi**. Output yang diharapkan:

```
+-------------+--------------+
| category    | jumlah_Movie |
+-------------+--------------+
| Documentary |            3 |
| Foreign     |            2 |
| Music       |            4 |
| New         |            1 |
| Sci-Fi      |            7 |
```

```
| Action      |           3 |
| Drama       |           2 |
| Animation   |           4 |
| Horror      |           1 |
| Family      |           4 |
| Comedy      |           3 |
| Children    |           2 |
| Classics    |           2 |
| Sports      |           2 |
| Games       |           1 |
| Travel      |           1 |
+-------------+-------------+
```

5. Dari soal sebelumnya diketahui `Gina Degeneres` paling banyak membintangi film bergenre science-fiction, dengan total **7** judul film. Tampilkan daftar **judul film sci-fi yang pernah dibintangi oleh** `Gina Degeneres`. Kolom yang diwajibkan ada yaitu **judul film** dan **kategorinya**. Output yang diharapkan:

```
+--------------------+----------+
| title              | category |
+--------------------+----------+
| CHARIOTS CONSPIRACY | Sci-Fi   |
| COLDBLOODED DARLING | Sci-Fi   |
| FRISCO FORREST      | Sci-Fi   |
| GOODFELLAS SALUTE   | Sci-Fi   |
| LICENSE WEEKEND     | Sci-Fi   |
| OPEN AFRICAN        | Sci-Fi   |
| SPIRITED CASUALTIES | Sci-Fi   |
+--------------------+----------+
```

Soal 2 - EDA (`Exploratory Data Analysis`) Real Estate (40 poin)

Anda adalah seorang Data Scientist di sebuah perusahaan real estate di Melbourne. Anda diberikan dataset tentang harga rumah berbagai tipe di Melbourne beserta variabel-variabel terkait rumah seperti jumlah kamar, luas tanah, metode penjualan, dan lainnya.

Data ini adalah cuplikan data yang dibuat oleh Tony Pino. Data harga rumah ini merupakan hasil web scrapping dari data yang terpublikasikan untuk publik melalui Domain.com.au.

Informasi mengenai keterangan variabel (*features*) dan apabila Anda kesulitan download dataset pada repo ini, Anda bisa akses Kaggle

Soal: Buatlah sebuah file *notebook* (**EDA_HouseMarket.ipynb**) dan ikuti panduan *Exploratory Data Analysis* berikut:

1. Untuk *feature/column* `Landsize` & `Price`, hitung dan beri penjelasan (*insight*) terkait:
    - Titik Pusat Data (*Central Tendency*)
    - Persebaran Data (*Dispersion*)

    - Distribusi Data (*Distribution*)

2. Carilah data *outlier* dan beri penjelasan (*insight*) pada *feature/column* `Distance` ! Tambahkan plot untuk menampilkan adanya data *outlier*!

3. Tampilkan distribusi (sebaran data) dalam bentuk plot yang tepat, untuk *feature/column* `YearBuilt`. Jelaskan *insight* apa yang Anda dapatkan di setiap *plot* tersebut!

4. Tampilkan perkembangan rata-rata harga (*price*) rumah berdasarkan tahun pembangunan rumah (*Year*) menggunakan plot yang tepat! Jelaskan (*insight*) perkembangan rata-rata harga rumah dari plot yang Anda buat!

5. Di antara `Landsize`, `Distance`, & `Rooms`, manakah *feature/column* yang memiliki nilai korelasi tertinggi pada *feature/column* `Price`? Jelaskan metode korelasi apa yang Anda gunakan, serta apa *insight* untuk perusahaan setelah mengetahui nilai korelasi tersebut?

6. Temukan `RegionName` yang rata-rata harga (*Price*) propertinya paling tinggi, serta tampilkan RegionName dan rata-rata harga rumah dalam bentuk Barplot! Lalu, di region yang memiliki rata-rata harga properti tertinggi, carilah:

    - Lima **`CouncilArea`** yang memiliki rata-rata harga rumah tertinggi!
    - Tipe properti (**`Type`**) yang memiliki rata-rata harga rumah tertinggi!
    - Metode penjualan properti (**`Method`**) yang paling sering dilakukan!

7. Lakukan Uji Hipotesis untuk menguji apakah ada perbedaan harga properti antar Region Name!

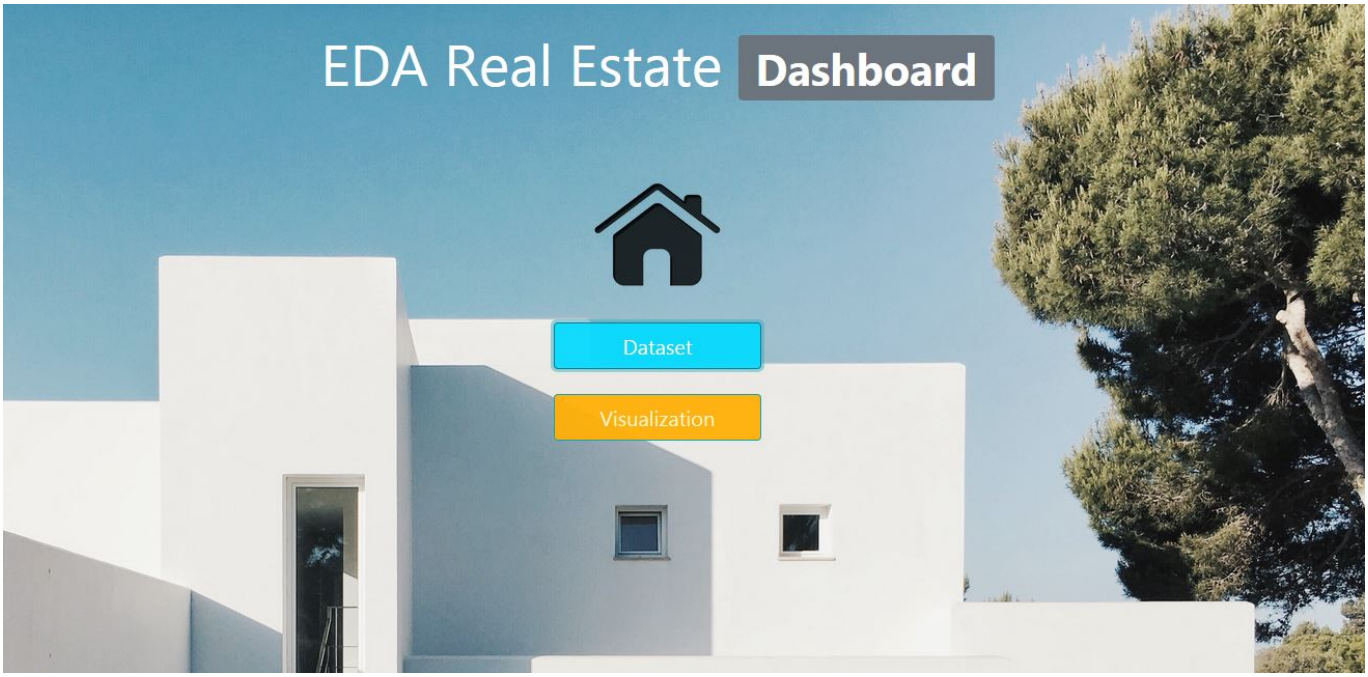
✅ Di setiap visualisasi, penjelasan (*insight*) yang Anda sampaikan sebaiknya yang berhubungan dengan keuntungan/kerugian perusahaan *real estate* atau manfaat ke konsumen.

Soal 3 - Dashboard for EDA Real Estate (20 poin)

Buatlah dashboard menggunakan Flask yang berisi visualisasi yang Anda buat di soal nomor 2! Beri penjelasan sekilas di setiap plot yang Anda tampilkan!

Kurang lebih tampilan Dashboard seperti berikut:

Contoh Tampilan Home

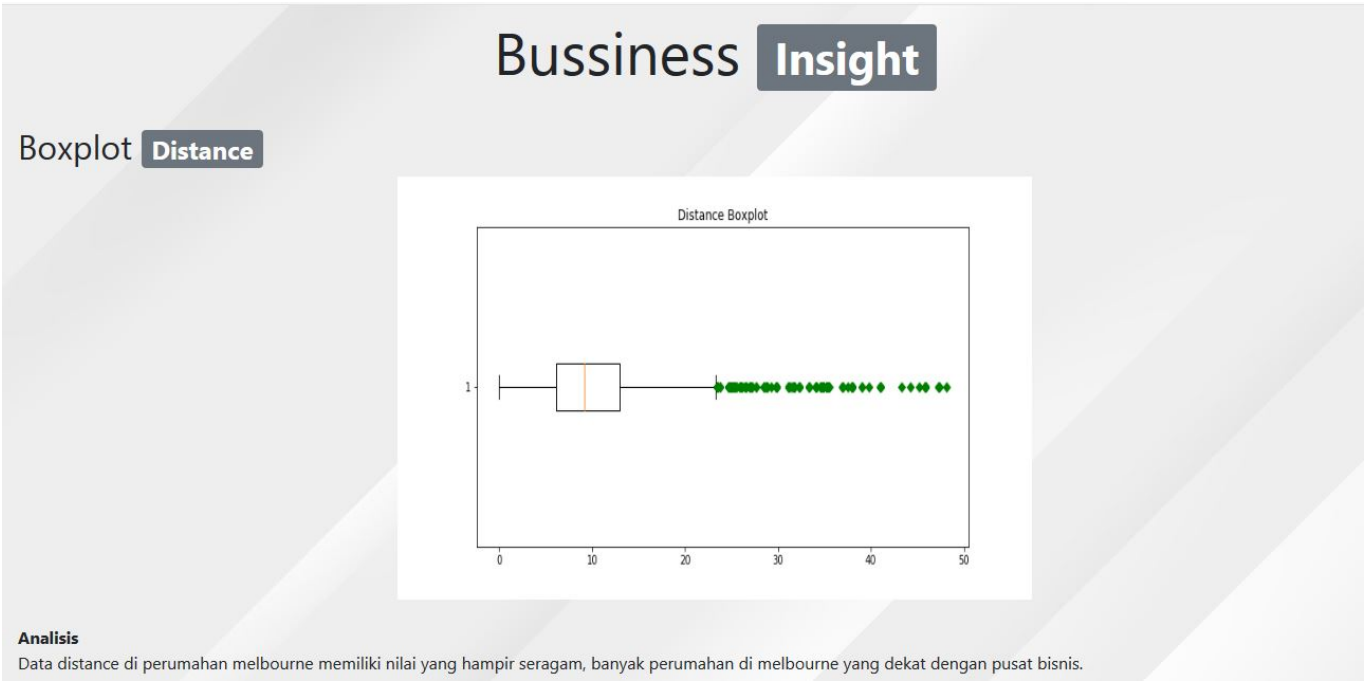


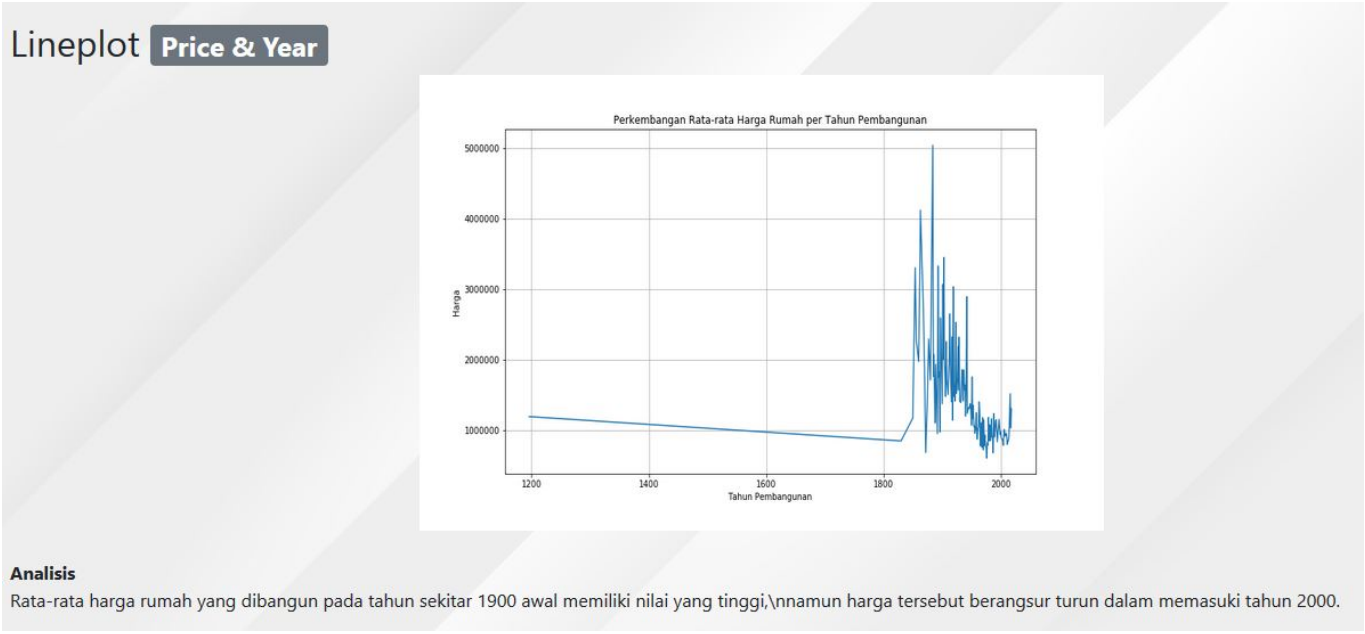
Contoh Tampilan Dataset

Melbourne House Price **Dataset**

Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt
Abbotsford	85 Turner St	2	h	1480000.0	S	Biggin	3/12/2016	2.5	3067.0	2.0	1.0	1.0	202.0		
Abbotsford	25 Bloomburg St	2	h	1035000.0	S	Biggin	4/02/2016	2.5	3067.0	2.0	1.0	0.0	156.0	79.0	1900.0
Abbotsford	5 Charles St	3	h	1465000.0	SP	Biggin	4/03/2017	2.5	3067.0	3.0	2.0	0.0	134.0	150.0	1900.0
Abbotsford	40 Federation La	3	h	850000.0	PI	Biggin	4/03/2017	2.5	3067.0	3.0	2.0	1.0	94.0		
Abbotsford	55a Park St	4	h	1600000.0	VB	Nelson	4/06/2016	2.5	3067.0	3.0	1.0	2.0	120.0	142.0	2014.0
Abbotsford	129 Charles St	2	h	941000.0	S	Jellis	7/05/2016	2.5	3067.0	2.0	1.0	0.0	181.0		

Contoh Tampilan Visualization





Tampilkan seluruh visualisasi yang Anda buat di soal nomor 2!

Good luck & Happy Coding