Laporan Projek 4
Dana Penelitian Startup

Kelompok 8:

- Bryan Christopher Wijaya 202000223
- James Patrick Oentoro 202000241
- Noel Christevent Mandak 202000436
- Stefannus Christian 202000338
- Tiffany Sondakh 191900199

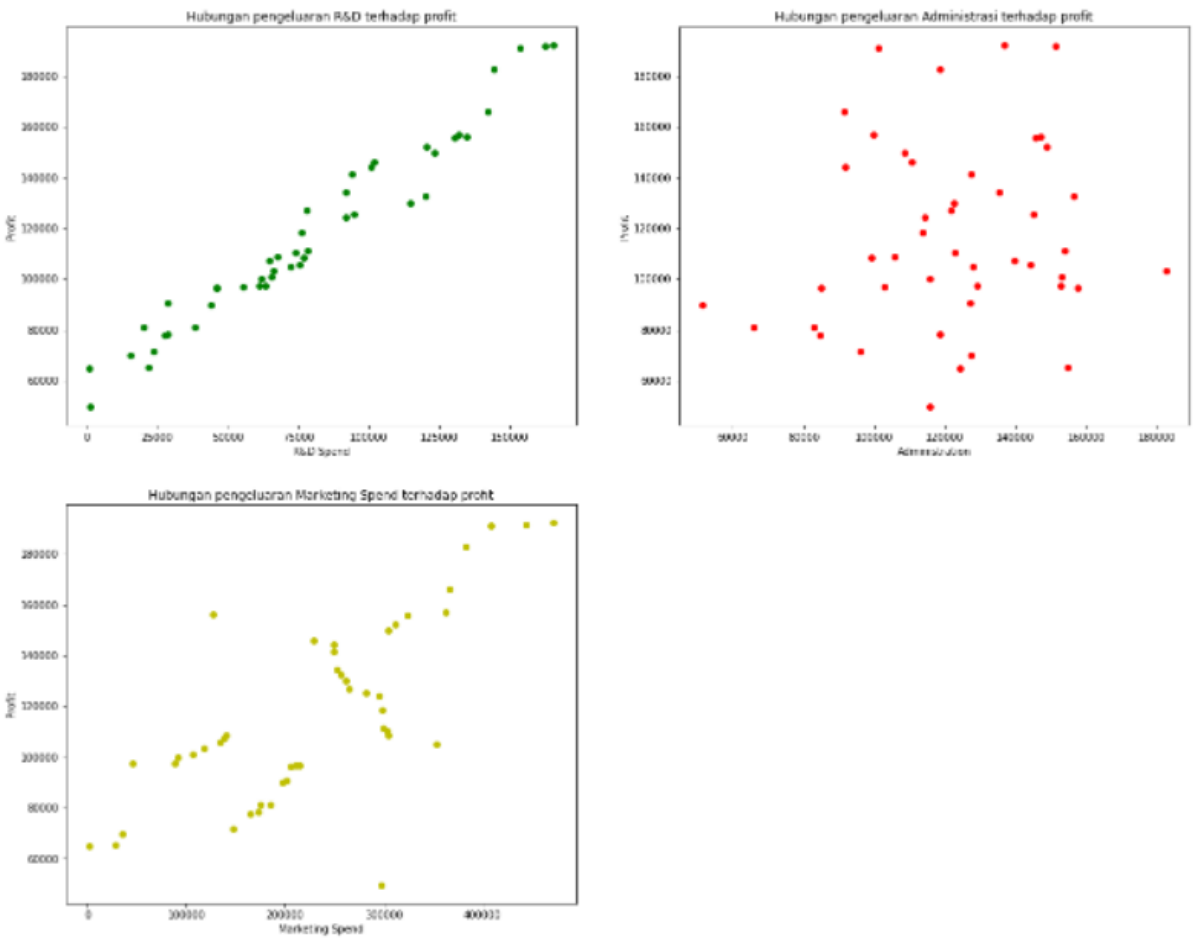
Link tugas pemrograman :
https://datalore.jetbrains.com/notebook/CPsQxEPW7yVpS665a3lC8b/VvRbW9E8rgEb0bH4IKZjgE/

Laporan Singkat:

1. Tulislah ringkasan singkat (maksimal 4 paragraf) untuk menceritakan proses statistika yang Anda lakukan, mulai dari mengekstrak informasi dari dataset hingga menghasilkan regresi.
2. Jawablah pertanyaan diskusi berikut dengan menggunakan justifikasi statistika:
   a. Bagaimanakah relasi antara masing-masing biaya pengeluaran dengan besarnya profit? Apakah linear, eksponensial, logaritmik, resiprokal, atau lainnya?
   b. Buatlah model regresi linear sederhana untuk hubungan antara profit dan dana R&D. Berikan evaluasi kecocokan modelnya.
   c. Berikan interval estimasi untuk prediksi profit suatu perusahaan startup yang mendedikasikan dana R&D sebesar 125.000 USD, dengan tingkat kepercayaan sebesar 95%.
   d. Selidikilah hasil regresi linear untuk relasi profit terhadap dana marketing. Berikan evaluasi kecocokan modelnya.
   e. Selidiki kembali model regresi linear untuk pasangan variabel yang sama, namun dipecah berdasarkan kategori state masing-masing (New York, Florida, California). Berikan evaluasi kecocokan model dan buatlah suatu kesimpulan yang relevan.
   f. Apa yang Anda dapat simpulkan mengenai kaitan antara dana administrasi dengan profit?

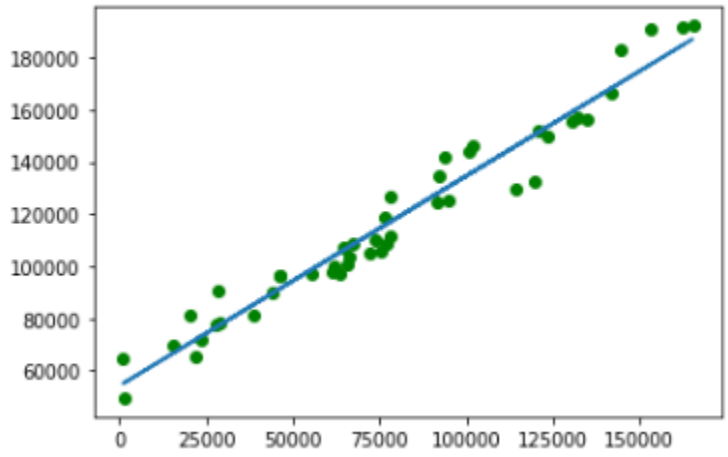
1.

Pertama kami mengimport data menggunakan dataframe pandas, kemudian kami membuat fungsi-fungsi yang nantinya akan digunakan untuk perhitungan statistika yaitu fungsi count_b() untuk menghitung b_0 dan b_1, make_model() untuk membuat model regresi $\hat{y} = b_0 + b_1 x$, plot_regression() untuk plotting data, dan regression_interval() untuk mencari interval model regresi dengan tingkat kepercayaan tertentu.

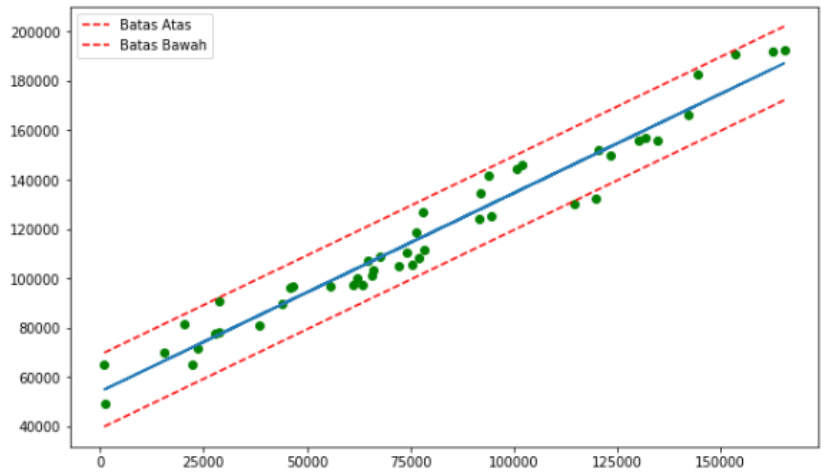


Melalui plot data menggunakan matplotlib, kami menyimpulkan bahwa relasi masing-masing pengeluaran terhadap Profit terlihat linear maka kami menggunakan regresi linear. Kami membuat beberapa visualisasi yaitu:

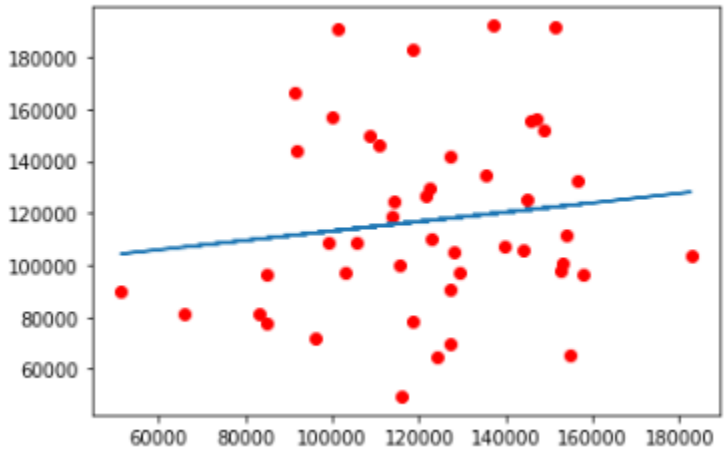
- Regresi linear data pengeluaran R&D Spend terhadap Profit.



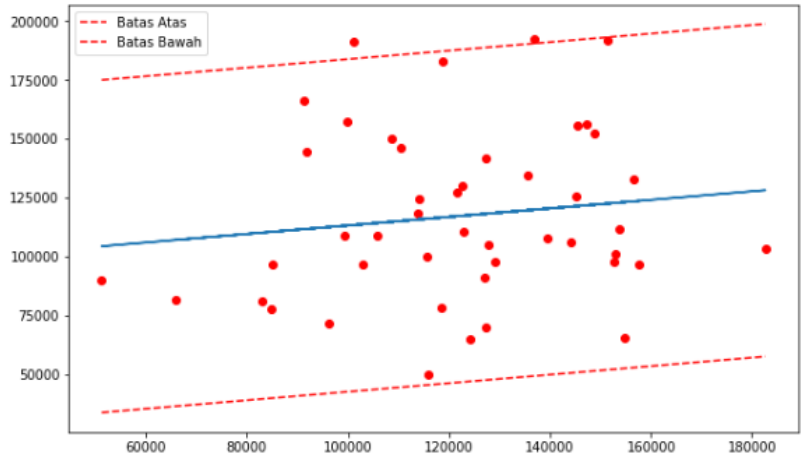
- Inferensi dengan tingkat kepercayaan 95% data pengeluaran R&D Spend terhadap Profit.



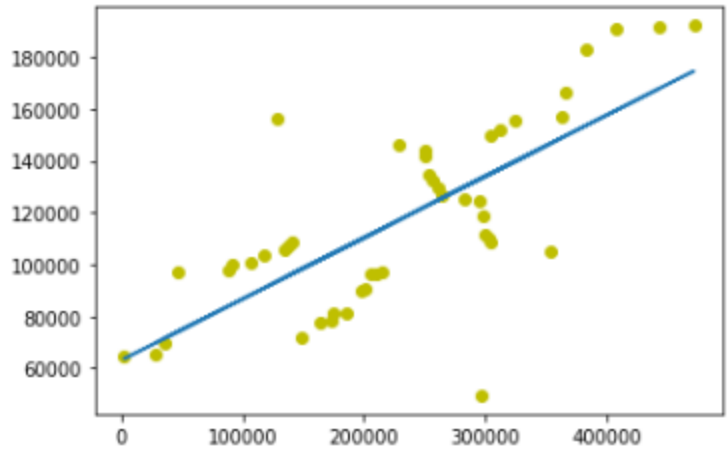
- Regresi linear data pengeluaran Administration terhadap Profit.



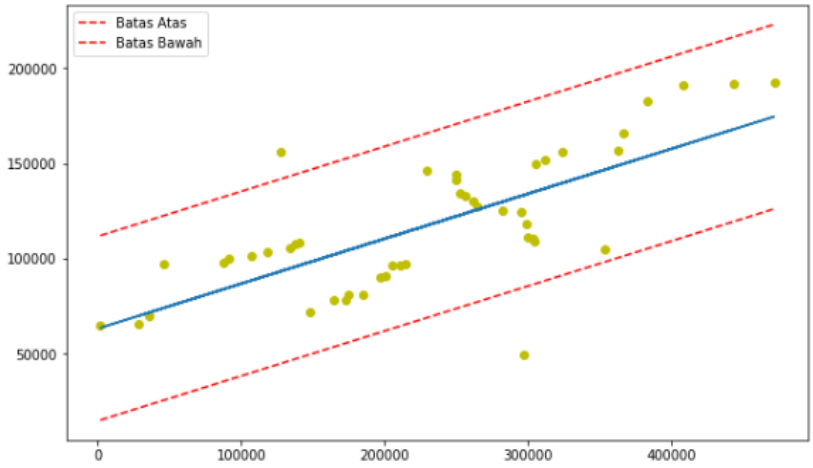
- Inferensi dengan tingkat kepercayaan 95% data pengeluaran Administration terhadap Profit.



- Regresi linear data pengeluaran Marketing Spend terhadap Profit.



- Inferensi dengan tingkat kepercayaan 95% data pengeluaran Administration terhadap Profit.



Kemudian kami mencari nilai koefisien korelasi dan mendapatkan hasil sebagai berikut :

```
Koefisien korelasi dari RnD Spend dan Profit adalah: 0.9777034670669674
Koefisien korelasi dari Administration dan Profit adalah: 0.13507591841115776
Koefisien korelasi dari Marketing dan Profit adalah: 0.732276732172417
```

Karena koefisien korelasi dari Administration dan provit hanya 0.135, maka kami menyimpulkan bahwa tidak ada kaitan secara langsung antara dana administrasi dengan profit.

Kemudian kami membuat fungsi find_SSE(), find_SSR(), find_SST(), find_r_squared() untuk mencari nilai R^2, dan isApproximatelyEqual() untuk mengecek apakah hasil perhitungan sudah benar. Akhirnya kami mendapatkan hasil sebagai berikut,

- Evaluasi Kecocokan Model R&D-Profit : $R^2 = 95,59\%$, yang artinya performa keberhasilan regresi kami sangat baik.
- Evaluasi Kecocokan Model Marketing-Profit : $R^2 = 53,62\%$, yang artinya performa keberhasilan regresi kami cukup baik.

- Sedangkan Evaluasi Kecocokan Model Marketing-Profit kami menyimpulkan bahwa data tidak berhubungan.

Maka menjawab cerita yang telah diberikan, kami menyimpulkan bahwa dengan tingkat kepercayaan 95% semakin banyak dana R&D yang diberikan sebuah perusahaan semakin banyak keuntungan yang akan diperoleh.

2.

a. Relasi pengeluaran R&D terhadap Profit : linear
   Relasi pengeluaran Administrasi terhadap Profit : tidak mempengaruhi / berhubungan
   Relasi pengeluaran Marketing Spend terhadap Profit : linear

b. Regresi Linear hubungan Profit dan dana R&D :
   $\hat{y} = 54108.94981015532 + 0.8046215878369701\,x$
   Evaluasi kecocokan :

```
Kecocokan Model Hubungan R&D - Profit

Sum of Square Error (SSE) = 2565963574.9603853
Sum of Square Regression (SSE) = 55624521277.59975
Sum of Square Total (SST) = 58190484852.560135

SSE + SSR = SST
2565963574.9603853 + 55624521277.59975 = 58190484852.560135

R² = 95.59%
```

Semakin tinggi nilai dari R^2 maka semakin baik model prediksi. Dengan kata lain, semakin dekat nilai R^2 ke 100% semakin baik model. Model Koefisien determinasi yang dihasilkan oleh grafik marketing-profit state Florida adalah 95.59% maka dari itu dapat dikatakan bahwa model yang dihasilkan sangat baik.

c. Interval Estimasi untuk prediksi profit suatu perusahaan startup yang mendedikasikan dana R & D sebesar 125.000 USD dengan tingkat kepercayaan 95%.

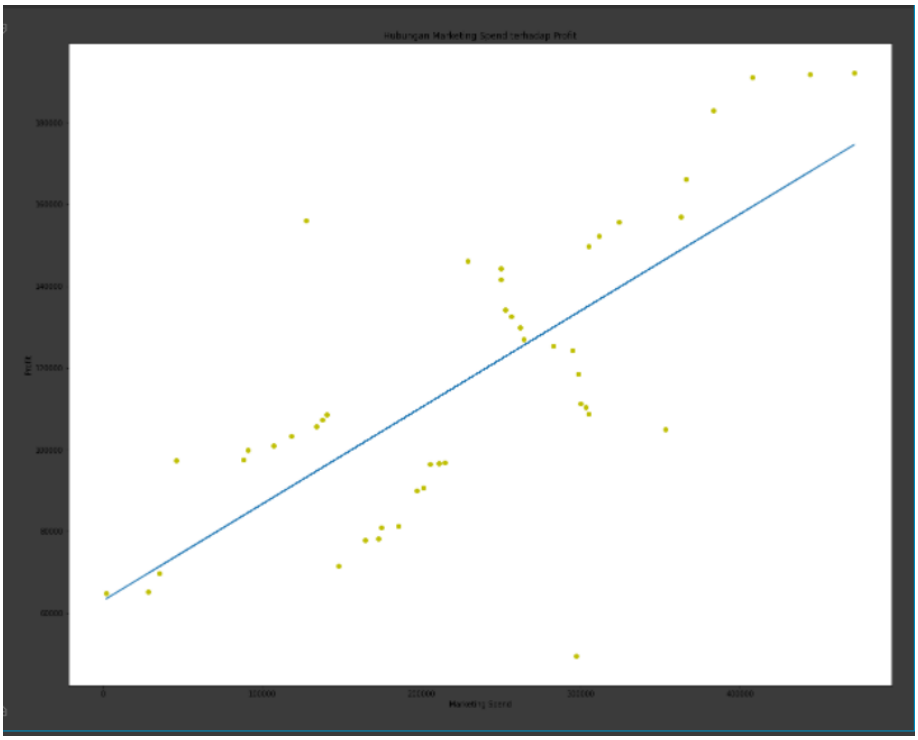
```
0.1s
```

```python
def predict_interval(x, y, tingkat_kepercayaan, x_test):
    n = y.count()
    model = make_model(x, y)
    y_hat = model(x)
    s = (((y - y_hat) ** 2).sum() / (n - 2)) ** 0.5
    alpha = (1 - tingkat_kepercayaan) / 2
    range_ = s * stats.norm.ppf(1 - alpha)
    y_pred = model(x_test)
    batas_atas = y_pred + range_
    batas_bawah = y_pred - range_
    return batas_bawah, batas_atas

predict_interval(x_rd, y, 0.95, 125000)
```

```
(139719.22654012052, 169654.07003943264)
```

Dengan tingkat kepercayaan 95%, maka besarnya profit yang akan dihasilkan oleh perusahaan tersebut jika perusahaan tersebut mengeluarkan 125 ribu USD untuk R&D berada dalam interval 139719 - 169654 dollar.

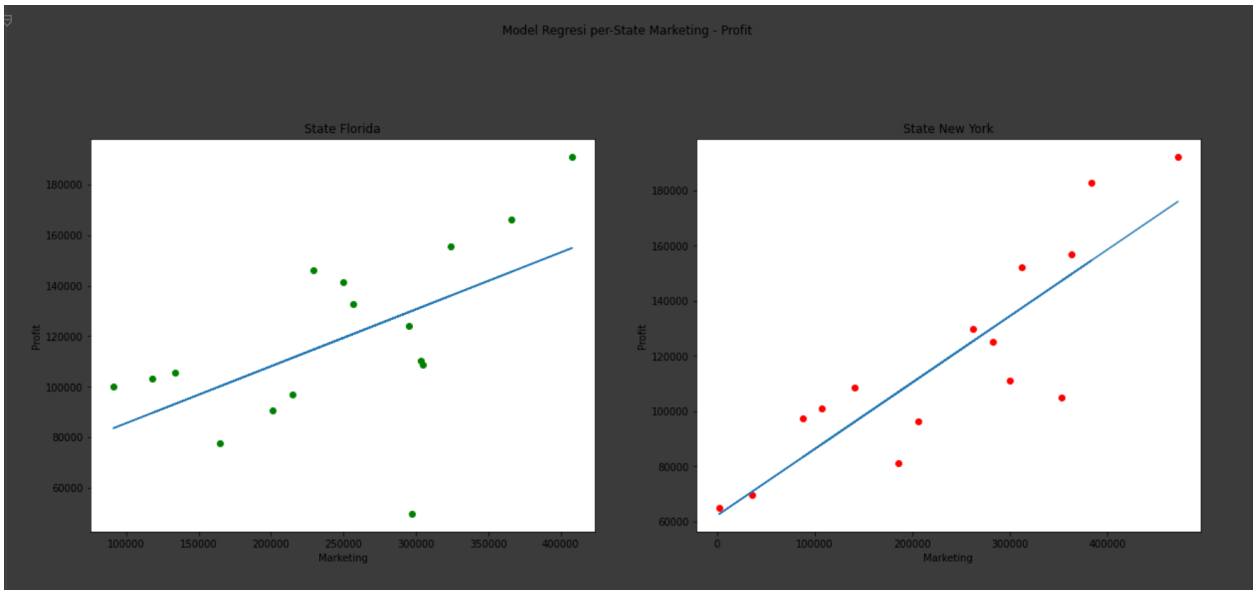
d. Hasil regresi linear untuk relasi profit terhadap dana marketing dan Evaluasi Modelnya

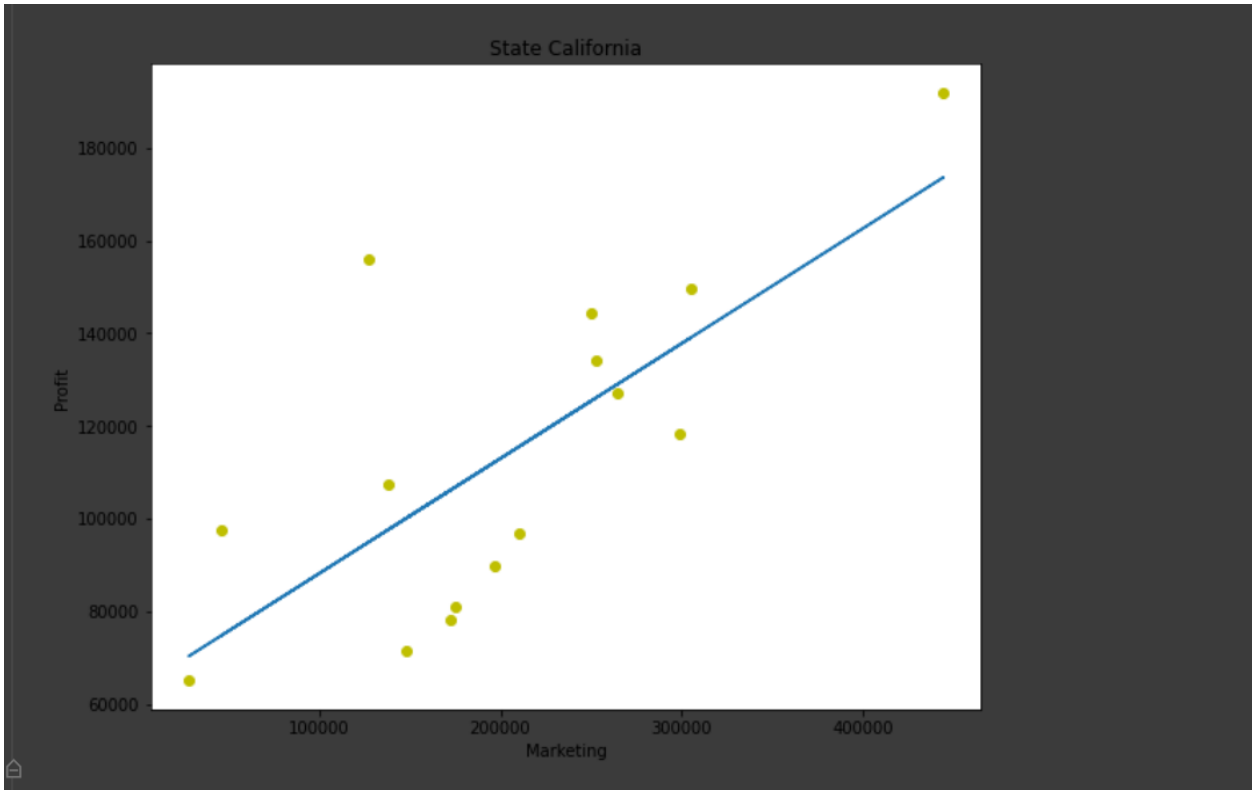


```
Kecocokan Model Hubungan Marketing - Profit

Sum of Square Error (SSE) = 26987046986.177666
Sum of Square Regression (SSE) = 31203437866.382477
Sum of Square Total (SST) = 58190484852.560135

SSE + SSR = SST
26987046986.177666 + 31203437866.382477 = 58190484852.560135

R² = 53.620000000000005%
```

Semakin tinggi nilai dari R^2 maka semakin baik model prediksi. Dengan kata lain, semakin dekat nilai R^2 ke 100% semakin baik model. Model Koefisien determinasi yang dihasilkan oleh grafik marketing-profit adalah 53,62% maka dari itu dapat dikatakan bahwa model yang dihasilkan kurang baik.

e. Hasil regresi linear untuk relasi profit terhadap dana marketing per state dan Evaluasi Model per state






```
Kecocokan Model Hubungan Marketing - Profit State Florida

Sum of Square Error (SSE) = 12849889664.174088
Sum of Square Regression (SSE) = 6166353201.718111
Sum of Square Total (SST) = 19016242865.892197

SSE + SSR = SST
12849889664.174088 + 6166353201.718111 = 19016242865.892197

R² = 32.43%
```

Semakin tinggi nilai dari R^2 maka semakin baik model prediksi. Dengan kata lain, semakin dekat nilai R^2 ke 100% semakin baik model. Model Koefisien determinasi yang dihasilkan oleh grafik marketing-profit state Florida adalah 32,43% maka dari itu dapat dikatakan bahwa model yang dihasilkan kurang baik.

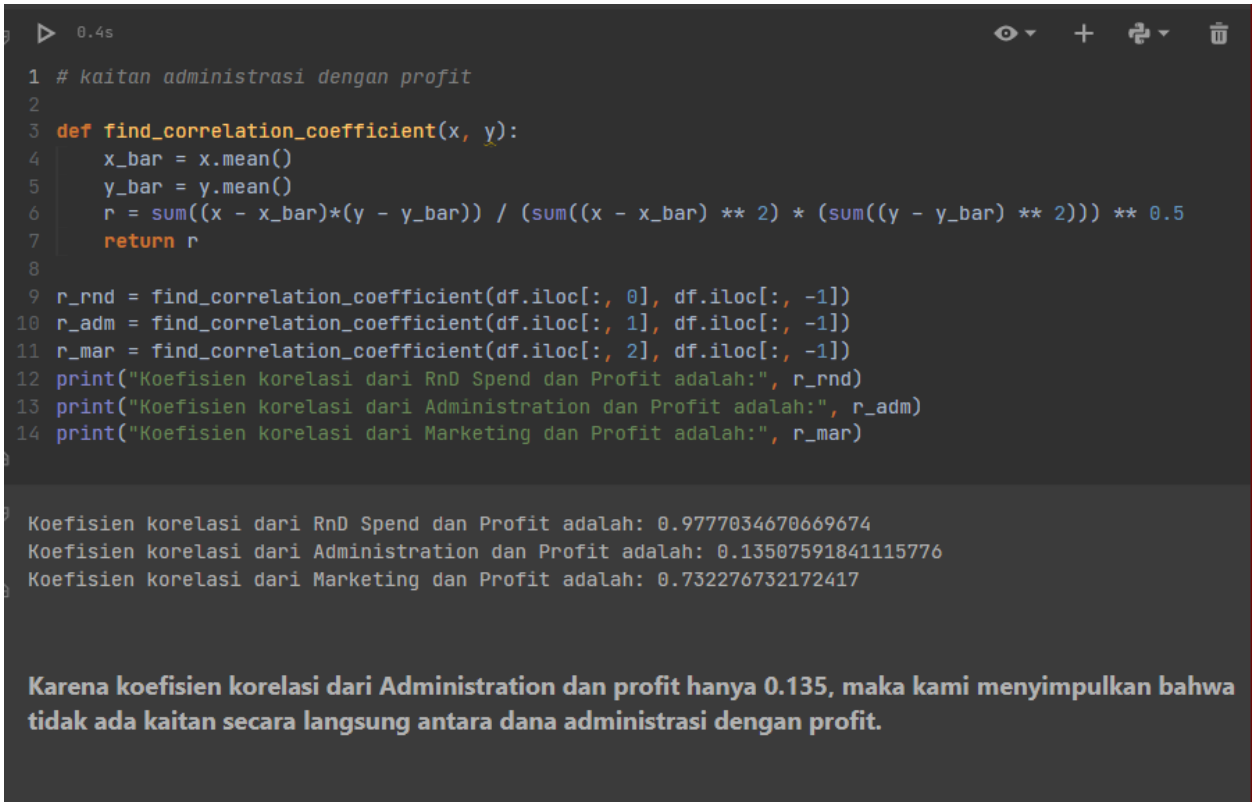
```
Kecocokan Model Hubungan Marketing - Profit State New York

Sum of Square Error (SSE) = 5134092501.420339
Sum of Square Regression (SSE) = 15449936068.655422
Sum of Square Total (SST) = 20584028570.075768

SSE + SSR = SST
5134092501.420339 + 15449936068.655422 = 20584028570.075768

R² = 75.06%
```

Semakin tinggi nilai dari R^2 maka semakin baik model prediksi. Dengan kata lain, semakin dekat nilai R^2 ke 100% semakin baik model. Model Koefisien determinasi yang dihasilkan oleh grafik marketing-profit state Florida adalah 75.06% maka dari itu dapat dikatakan bahwa model yang dihasilkan cukup baik.

```
Kecocokan Model Hubungan Marketing - Profit State California

Sum of Square Error (SSE) = 8727508532.176632
Sum of Square Regression (SSE) = 9644601480.104303
Sum of Square Total (SST) = 18372110012.280933

SSE + SSR = SST
8727508532.176632 + 9644601480.104303 = 18372110012.280933

R² = 52.5%
```

Semakin tinggi nilai dari R^2 maka semakin baik model prediksi. Dengan kata lain, semakin dekat nilai R^2 ke 100% semakin baik model. Model Koefisien determinasi yang dihasilkan oleh grafik marketing-profit state Florida adalah 52.5% maka dari itu dapat dikatakan bahwa model yang dihasilkan kurang baik.

f. Kaitan antara dana administrasi dengan profit



```
0.4s

# kaitan administrasi dengan profit

def find_correlation_coefficient(x, y):
    x_bar = x.mean()
    y_bar = y.mean()
    r = sum((x - x_bar)*(y - y_bar)) / (sum((x - x_bar) ** 2) * (sum((y - y_bar) ** 2))) ** 0.5
    return r

r_rnd = find_correlation_coefficient(df.iloc[:, 0], df.iloc[:, -1])
r_adm = find_correlation_coefficient(df.iloc[:, 1], df.iloc[:, -1])
r_mar = find_correlation_coefficient(df.iloc[:, 2], df.iloc[:, -1])
print("Koefisien korelasi dari RnD Spend dan Profit adalah:", r_rnd)
print("Koefisien korelasi dari Administration dan Profit adalah:", r_adm)
print("Koefisien korelasi dari Marketing dan Profit adalah:", r_mar)
```

```
Koefisien korelasi dari RnD Spend dan Profit adalah: 0.9777034670669674
Koefisien korelasi dari Administration dan Profit adalah: 0.13507591841115776
Koefisien korelasi dari Marketing dan Profit adalah: 0.732276732172417
```

Karena koefisien korelasi dari Administration dan profit hanya 0.135, maka kami menyimpulkan bahwa tidak ada kaitan secara langsung antara dana administrasi dengan profit.

Karena hasil koefisien korelasi dari Administration hanya 0.135 (sangat rendah), maka menurut website https://accurate.id/akuntansi/koefisien-korelasi/, koefisien korelasi 0.135 artinya kedua variabel sangat rendah korelasinya.