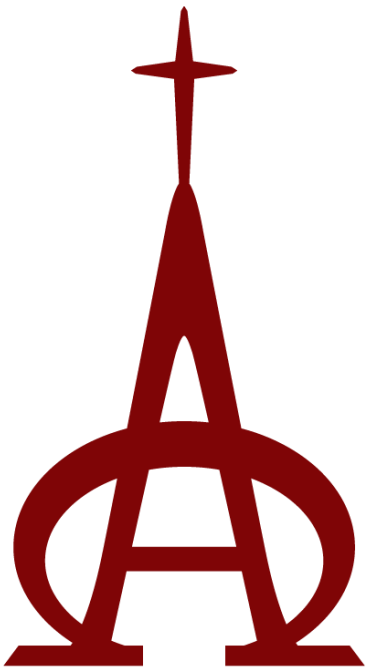
Laporan Proyek
Peluang dan Statistika
MATH1042



Stephen Theodorus (202000232)
Sean Wilbert N. D. (202000286)
Nathanael Hansen (202000424)
Samuel Kennard Sun (202000213)

CALVIN INSTITUTE OF TECHNOLOGY
2022

DAFTAR ISI

DAFTAR GAMBAR

1. Pendahuluan

1.1 Latar Belakang

Pada proyek 4 kali ini, kelompok kami akan melakukan analisa mengenai permasalahan profit dari perusahaan dan kaitannya dengan biaya R&D, Administrasi, dan Marketing. Melalui analisa permasalahan ini, kelompok kami berharap dapat memberikan konklusi terbaik untuk menjawab pertanyaan tentang seberapa penting peran divisi Penelitian dan Pengembangan (R&D) dalam sebuah perusahaan startup.

1.2 Persiapan Awal

Kelompok kami menggunakan Python sebagai peralatan statistika utama dan library Numpy, Pandas, serta Scipy untuk membantu melakukan proses statistika Uji Hipotesis pada proyek kali ini.

Data yang digunakan pada proyek kali ini diambil dari 50 perusahaan startup di Amerika Serikat. Data dari setiap perusahaan startup tersebut mencakup tiga jenis pengeluaran, yakni R&D, Administrasi, dan Marketing. Data-data dari 50 perusahaan startup tersimpan dalam sebuah file csb dengan nama 'Start-ups.csv". Kelompok kami kemudian menggunakan library pandas untuk mengekstrak data-data dari file csv tersebut.

	R&D Spend	Administration	Marketing Spend	State	Profit
0	73994.56	122782.75	303319.26	Florida	110352.25
1	165349.20	136897.80	471784.10	New York	192261.83
2	55493.95	103057.49	214634.81	Florida	96778.92
3	72107.60	127864.55	353183.81	New York	105008.31
4	119943.24	156547.42	256512.92	Florida	132602.65
5	91749.16	114175.79	294919.57	Florida	124266.90
6	131876.90	99814.71	362861.36	New York	156991.12
7	91992.39	135495.07	252664.93	California	134307.35
8	64664.71	139553.16	137962.62	California	107404.34
9	130298.13	145530.06	323876.68	Florida	155752.60
10	1000.23	124153.04	1903.93	New York	64926.08
11	27892.92	84710.77	164470.71	Florida	77798.83
12	38558.51	82982.09	174999.30	California	81005.76
13	44069.95	51283.14	197029.42	California	89949.14
14	15505.73	127382.30	35534.17	New York	69758.98
15	93863.75	127320.38	249839.44	Florida	141585.52
16	23640.93	96189.63	148001.11	California	71498.49
17	76253.86	113867.30	298664.47	California	118474.03
18	28663.76	127056.21	201126.82	Florida	90708.19
19	78389.47	153773.43	299737.29	New York	111313.02
20	144372.41	118671.85	383199.62	New York	182901.99

Gambar 1. 20 Data Pertama Dari Dataset Start-ups.csv

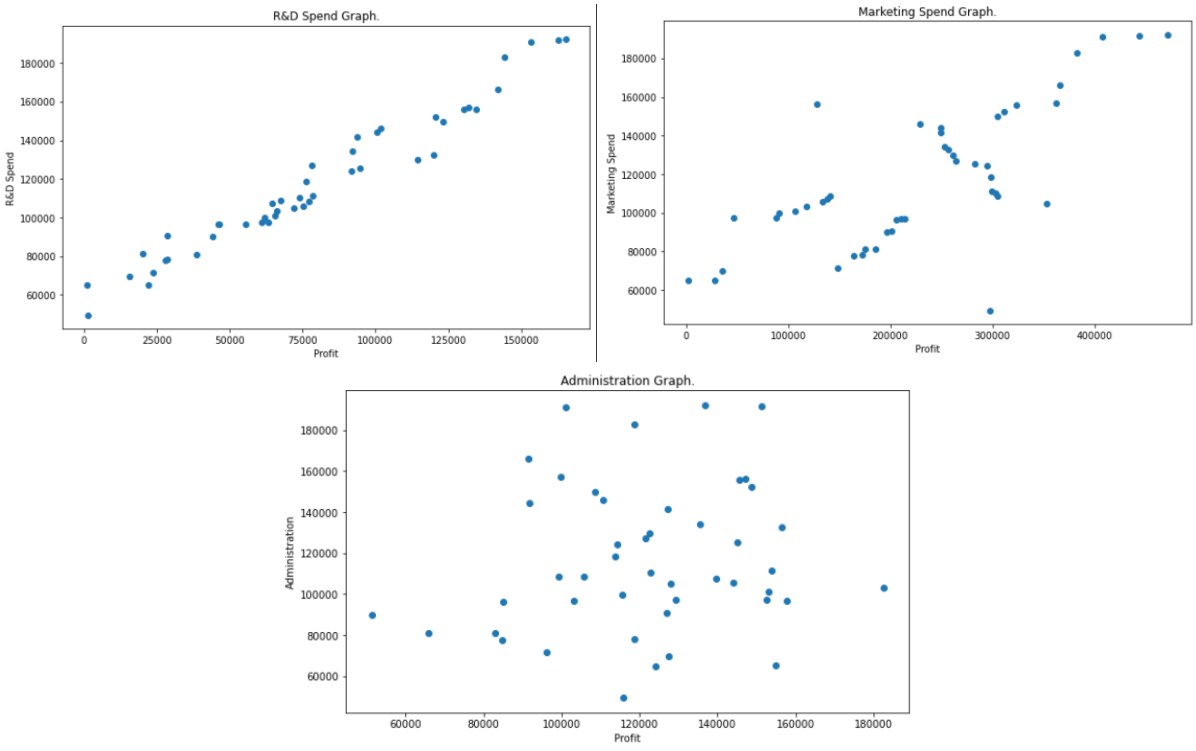
2. Rumusan Masalah

2.1 Masalah 1

Sebagai langkah pertama, kelompok kami akan mulai membuat suatu grafik yang menunjukkan hubungan antara masing-masing biaya pengeluaran dengan profit yang didapatkan. Kami menggunakan scatter plot untuk menggambarkan hubungan ini agar dapat dengan mudah dikaitkan dengan regresi linear nantinya.

```
#1. Membuat Grafik masing-masing biaya pengeluaran terhadap besarnya profit

fig, ax = plt.subplots(3,figsize=(10, 20))

ax[0].scatter(data['R&D Spend'], data['Profit'])
ax[0].set_title("R&D Spend Graph.")
ax[0].set_ylabel("R&D Spend")
ax[0].set_xlabel("Profit")

ax[1].scatter(data['Marketing Spend'], data['Profit'])
ax[1].set_title("Marketing Spend Graph.")
ax[1].set_ylabel("Marketing Spend")
ax[1].set_xlabel("Profit")

ax[2].scatter(data['Administration'], data['Profit'])
ax[2].set_title("Administration Graph.")
ax[2].set_ylabel("Administration")
ax[2].set_xlabel("Profit")
plt.show()
✓ 0.7s
```

Gambar 2. Kode Membuat Grafik Hubungan Pengeluaran dengan Profit

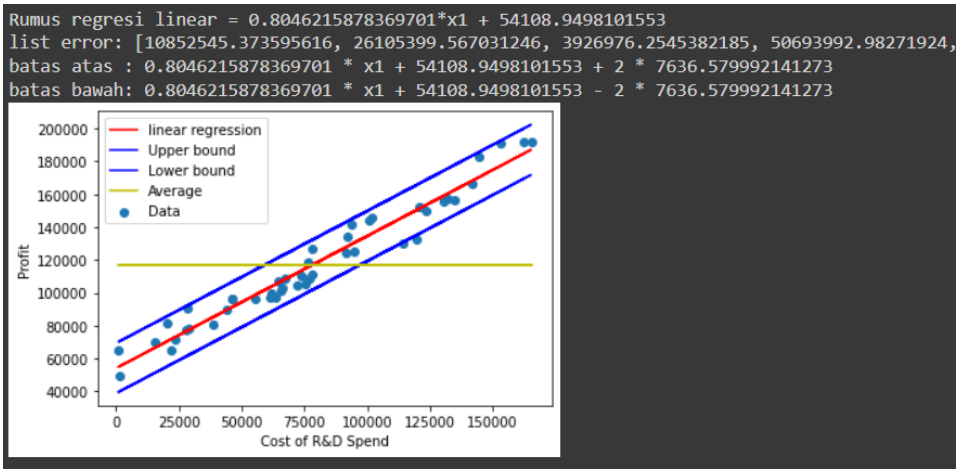


Gambar 3. Grafik Hubungan Masing-Masing Pengeluaran dengan Profit

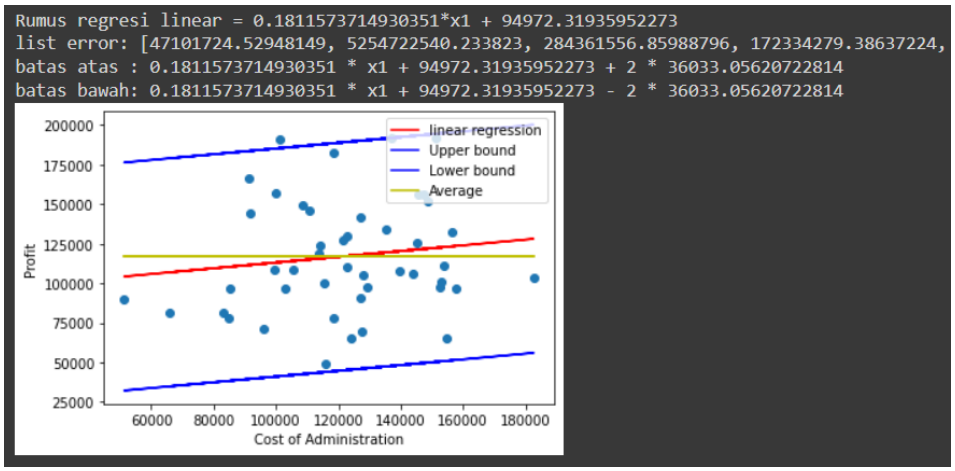
Dapat dilihat bahwa dari ketiga grafik scatter diatas, hanya R&D saja yang memiliki relasi linear yang kelihatan sangat jelas dengan profit. Kelompok kami juga memilih marketing sebagai sebuah fitur yang memiliki relasi linear dengan profit walaupun mungkin ada beberapa data outlier yang menyebabkan visualisasi datanya terkesan tidak linear.

2.2 Masalah 2

Kelompok kami kemudian mencari regresi linear untuk setiap jenis pengeluaran terhadap nilai profit. Kami mengelompokkan terlebih dahulu data-data yang ada ke dalam kelompok-kelompok pengeluaran, seperti R&D, Administrasi, dan Marketing.

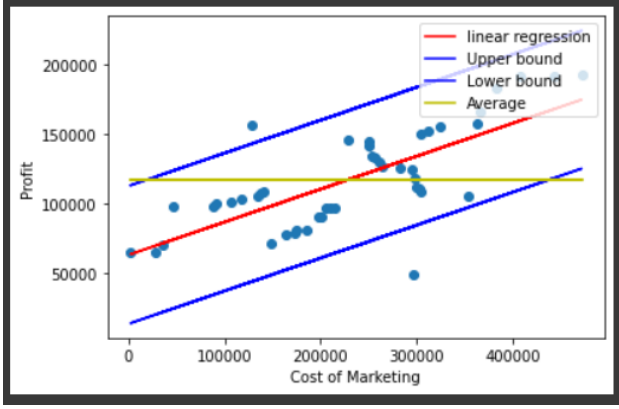


Gambar 4. Grafik Relasi Pengeluaran R&D dengan Profit



Gambar 5. Grafik Relasi Pengeluaran Administrasi dengan Profit

```
Rumus regresi linear = 0.2365138910721844*x1 + 63045.78840138968
list error: [596959605.8802425, 310906761.72493047, 290054364.0306718, 1728094462.
batas atas : 0.2365138910721844 * x1 + 63045.78840138968 + 2 * 24765.742002988387
batas bawah: 0.2365138910721844 * x1 + 63045.78840138968 - 2 * 24765.742002988387
```



Gambar 6. Grafik Relasi Pengeluaran Marketing dengan Profit

Berdasarkan grafik-grafik regresi linear yang telah ditampilkan, kelompok kami dapat menyimpulkan bahwa hanya R&D spend dan Marketing spend saja yang memiliki kaitan linear dengan profit perusahaan. Sedangkan untuk pengeluaran Administrasi tidak memiliki kaitan linear dengan profit dari perusahaan. Hal ini dapat dilihat dari R2 score dari ketiga model ini.

```
percentage of performance for R&D vs profit is 95.77663994296597
percentage of performance for Administration vs profit is 50.46033713482172
percentage of performance for marketing vs profit is 68.31670699584154
```

Gambar 7. Persentase R&D, Administration, dan Marketing terhadap Profit

Dapat dilihat bahwa data R&D cocok dimasukan kedalam regresi linear untuk melakukan prediksi nilai profit karena memiliki R2 score sebesar 95.7%. Walaupun tidak memiliki nilai R2 score yang setinggi R&D, marketing juga memiliki kecocokan yang lumayan bagus yaitu 68%. Akan tetapi Administrasi tidak memiliki hubungan yang tidak terlalu linear dengan profit karena R2 scorenya yang tergolong rendah yaitu 50.4%.

2.3 Masalah 3

```
result = m*125000 + b
upper_bound = m*125000 + b + 2*s
lower_bound = m*125000 + b - 2*s
print("result =",result)
print(upper_bound, lower_bound)
```

```
result = 154686.64828977658
169959.80827405912 139413.48830549404
```

Gambar 8. Rumus untuk Prediksi Profit

Dari hasil regresi linear terhadap R&D diatas, kami mendapatkan nilai prediksi jika perusahaan menghabiskan uang sebesar 125.000 USD untuk R&D, dengan tingkat kepercayaan 95% kita dapat mendapatkan profit sebesar 154686.64828977658 USD dengan batas atas yaitu 169959.80827405912 USD dan batas bawah yaitu 139413.48830549404 USD.

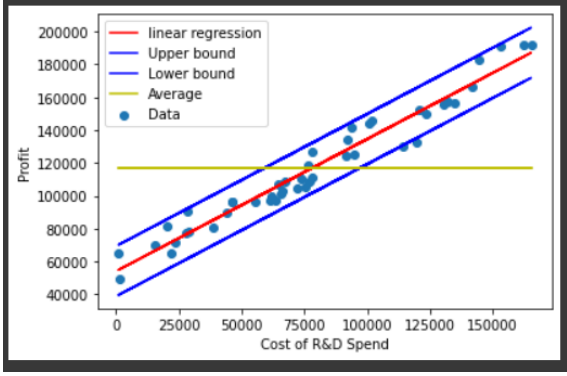
2.4 Masalah 4: Pertanyaan Diskusi

2.4.1 Relasi Masing-Masing Biaya Pengeluaran dengan Besarnya Profit

Jika kita melihat dari grafik persebaran data yang sudah ditampilkan di atas beserta regresi linearnya, maka kita dapat melihat pengeluaran R&D dan Marketing memiliki relasi linear dengan besarnya profit. Sedangkan Administrasi tidak memiliki relasi yang relevan dengan profit.

2.4.2 Model Regresi Linear Sederharna antara Profit dengan R&D Spend

```
Rumus regresi linear = 0.8046215878369701*x1 + 54108.9498101553
list error: [10852545.373595616, 26105399.567031246, 3926976.2545382185, 50693992.98271924,
batas atas : 0.8046215878369701 * x1 + 54108.9498101553 + 2 * 7636.579992141273
batas bawah: 0.8046215878369701 * x1 + 54108.9498101553 - 2 * 7636.579992141273
```

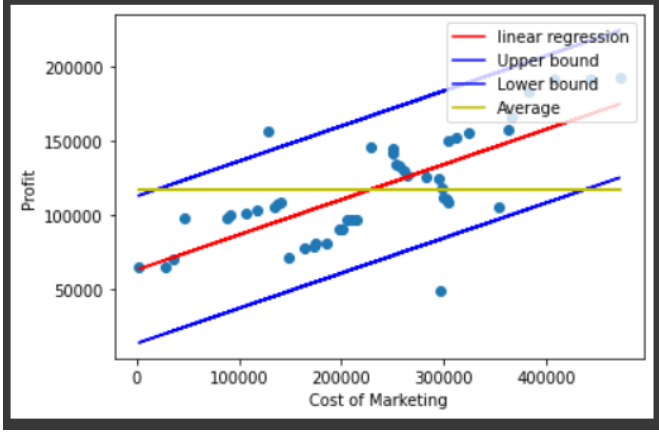


Gambar 9. Grafik Relasi Pengeluaran R&D dengan Profit

Hanya melalui gambar grafik diatas ini, kita dapat melihat bahwa model regresi linear sangat cocok untuk menggambarkan hubungan antara profit dengan R&D spend. Oleh karena itu, model ini sangat cocok dalam meprediksi nilai profit dari sebuah perusahaan start up.

2.4.3 Model Regresi Linear Sederharna antara Profit dengan Marketing Spend

```
Rumus regresi linear = 0.2365138910721844*x1 + 63045.78840138968
list error: [596959605.8802425, 310906761.72493047, 290054364.0306718, 1728094462.
batas atas : 0.2365138910721844 * x1 + 63045.78840138968 + 2 * 24765.742002988387
batas bawah: 0.2365138910721844 * x1 + 63045.78840138968 - 2 * 24765.742002988387
```



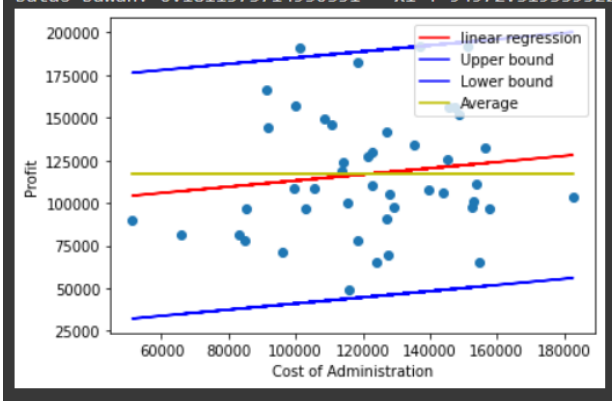
Gambar 10. Grafik Relasi Pengeluaran Marketing dengan Profit

Dengan melihat grafik regresi linear di atas, kita dapat mengevaluasi bahwa model regresi linear cukup dapat menggambarkan meskipun tidak secocok dengan data R&D. Dari visualisasi diatas, kita dapat melihat bahwa ada 2 data yang melebihi batas atas dan batas bawah. Kemungkinan besar, ini merupakan data outlier. Jika kita melakukan data

preprocessing dan menghilangkan beberapa data outlier, mungkin kita akan mendapatkan visualisasi yang lebih menggambarkan relasi linear antara data marketing spend dan profit.

2.4.4 Model Regresi Linear Sederharna antara Profit dengan Dana Administrasi

```
Rumus regresi linear = 0.1811573714930351*x1 + 94972.31935952273
list error: [47101724.52948149, 5254722540.233823, 284361556.85988796, 172334279.38637224,
batas atas : 0.1811573714930351 * x1 + 94972.31935952273 + 2 * 36033.05620722814
batas bawah: 0.1811573714930351 * x1 + 94972.31935952273 - 2 * 36033.05620722814
```



Gambar 11. Grafik Relasi Pengeluaran Administrasi dengan Profit

Kita dapat melihat grafik sebaran data dari dana pengeluaran administrasi kurang cocok dengan model regresi linear. Selain itu, kelompok kami menemukan bahwa nilai percentage of performance dari model regresi linearnya pun hanya sebesar 50.46%. Nilai ini tentu saja dapat menyimpulkan bahwa Administrasi tidak memiliki relasi linear yang bagus terhadap peningkatan profit dari perusahaan.

2.4.5 R^2 Score

```
percentage of performance for R&D vs profit is 95.77663994296597
percentage of performance for Administration vs profit is 50.46033713482172
percentage of performance for marketing vs profit is 68.31670699584154
```

Gambar 12. Persentase R&D, Administration, dan Marketing terhadap Profit

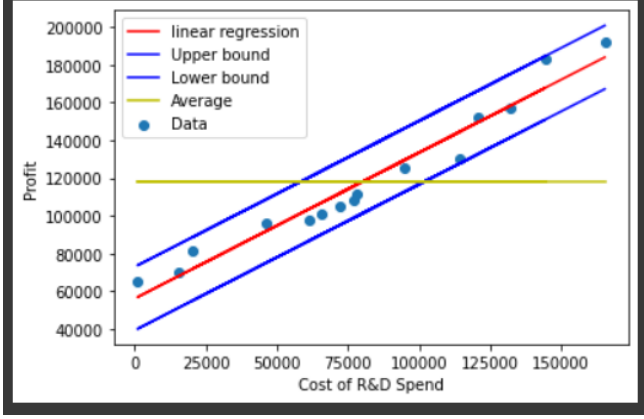
Dapat dilihat bahwa data R&D cocok dimasukan kedalam regresi linear untuk melakukan prediksi nilai profit karena memiliki R2 score sebesar 95.7%. Walaupun tidak memiliki nilai R2 score yang setinggi R&D, marketing juga memiliki kecocokan yang lumayan bagus yaitu 68%. Akan tetapi Administrasi tidak memiliki hubungan yang tidak terlalu linear dengan profit karena R2 scorenya yang tergolong rendah yaitu 50.4%.

2.5 Masalah 4

2.5.1 NEW YORK

2.5.1.1 Regresi Linear Sederharna antara Profit dengan R&D Spend

```
Rumus regresi linear = 0.7728067740825835*x1 + 56105.661582674235
list error: [70110253.3057149, 46547779.68531459, 1060700.699593543, 64761192.337611
batas atas : 0.7728067740825835 * x1 + 56105.661582674235 + 2 * 8399.944316929159
batas bawah: 0.7728067740825835 * x1 + 56105.661582674235 - 2 * 8399.944316929159
```

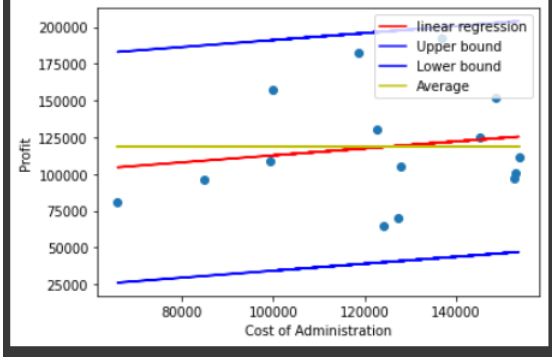


Gambar 13. Grafik State New York Relasi Pengeluaran R&D dengan Profit

Hanya melalui gambar grafik diatas ini, kita dapat melihat bahwa model regresi linear sangat cocok untuk menggambarkan hubungan antara profit dengan R&D spend. Oleh karena itu, model ini sangat cocok dalam meprediksi nilai profit dari sebuah perusahaan start up.

2.5.1.2 Regresi Linear Sederharna antara Profit dengan Administration

```
Rumus regresi linear = 0.23770541120170532*x1 + 88869.59268362902
list error: [5019848539.3649645, 203215805.32992032, 1970918744.8021898, 2857475738.1798077, 2439379477
batas atas : 0.23770541120170532 * x1 + 88869.59268362902 + 2 * 39246.56836272714
batas bawah: 0.23770541120170532 * x1 + 88869.59268362902 - 2 * 39246.56836272714
```

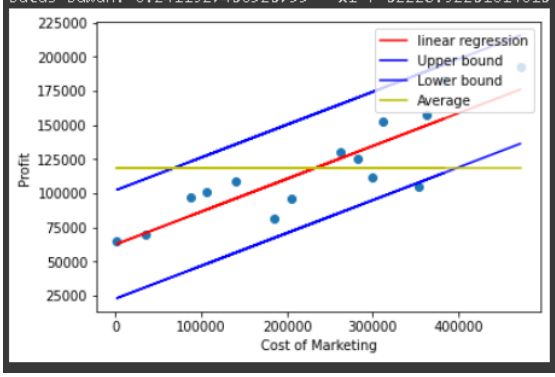


Gambar 14. Grafik State New York Relasi Pengeluaran Administrasi dengan Profit

Kita dapat melihat grafik sebaran data dari dana pengeluaran administrasi kurang cocok dengan model regresi linear. Dari visualisasi diatas, kita dapat melihat bahwa perbedaan batas atas dan batas bawah dengan nilai regresi dari regresi ini sangat besar. Kita juga dapat melihat bahwa banyak titik-titik data yang memiliki jarak yang sangat jauh dengan garis regresi sehingga model ini dianggap tidak akan terlalu akurat untuk dijadikan model prediksi.

2.5.1.3 Regresi Linear Sederharna antara Profit dengan Marketing Spend

```
Rumus regresi linear = 0.2411927460526755*x1 + 62228.92261014016
list error: [263802718.84702116, 1798267615.1642513, 52456261.993961535, 5008390.146284768, 1082695.711
batas atas : 0.2411927460526755 * x1 + 62228.92261014016 + 2 * 19872.85063640718
batas bawah: 0.2411927460526755 * x1 + 62228.92261014016 - 2 * 19872.85063640718
```



Gambar 15. Grafik State New York Relasi Pengeluaran Marketing dengan Profit

Dengan melihat grafik regresi linear di atas, kita dapat mengevaluasi bahwa model regresi linear cukup dapat menggambarkan meskipun tidak secocok dengan data R&D. Perbedaan batas atas dan batas bawah dengan nilai regresi dari regresi ini lebih kecil dibandingkan dengan batas atas dan batas bawah dari model regresi dengan data marketing yang belum dibagi berdasarkan state. Ini mendandakan bahwa, pada state New York, marketing spend mempunyai relasi linear dengan profit dari perusahaan start up.

2.5.1.4 R^2 Score

```
percentage of performance for R&D  vs profit is 95.73389519677275
percentage of performance for Administration vs profit is 50.68979033770897
percentage of performance for marketing vs profit is 80.03706224436996
```

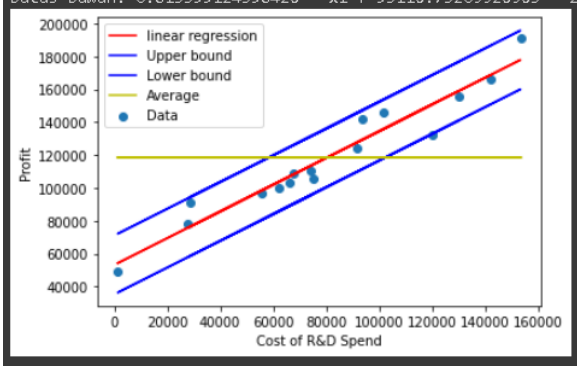
Gambar 16. R^2 Score New York

Dapat dilihat bahwa data R&D cocok dimasukan kedalam regresi linear untuk melakukan prediksi nilai profit karena memiliki R2 score sebesar 95.7%. Walaupun tidak memiliki nilai R2 score yang setinggi R&D, marketing juga memiliki kecocokan yang bagus yaitu 80%. Ini merupakan loncatan yang begitu tinggi jika dibandingkan dengan R2 score dari model regresi dengan data marketing yang belum dibagi berdasarkan state. Akan tetapi Administrasi tetap tidak menunjukan hubungan yang linear dengan profit karena R2 scorenya yang tergolong rendah yaitu 50.6%.

2.5.2 FLORIDA

2.5.2.1 Regresi Linear Sederharna antara Profit dengan R&D Spend

```
Rumus regresi linear = 0.813355124358426*x1 + 53116.73209926963
list error: [8692689.001241786, 2172972.881582443, 326544090.0951028, 12072021.977712749, 11174203.7280
batas atas : 0.813355124358426 * x1 + 53116.73209926963 + 2 * 8942.431622472492
batas bawah: 0.813355124358426 * x1 + 53116.73209926963 - 2 * 8942.431622472492
```

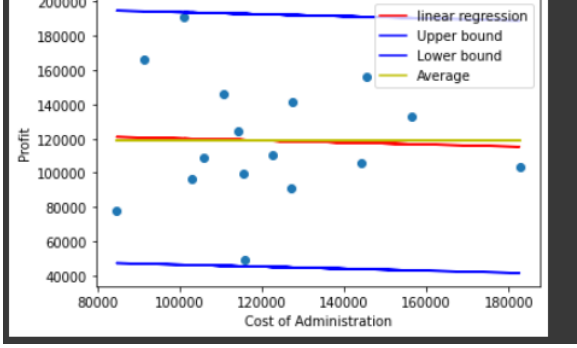


Gambar 17. Grafik State Florida Relasi Pengeluaran R&D dengan Profit

Hanya melalui gambar grafik diatas ini, kita dapat melihat bahwa model regresi linear sangat cocok untuk menggambarkan hubungan antara profit dengan R&D spend. Oleh karena itu, model ini sangat cocok dalam meprediksi nilai profit dari sebuah perusahaan start up.

2.5.2.2 Regresi Linear Sederharna antara Profit dengan Administration

```
Rumus regresi linear = -0.058704985460230535*x1 + 125922.46573224581
list error: [69927328.39868604, 533312387.8551693, 251866367.58157802, 25473444.112920925, 1472523531.7
batas atas : -0.058704985460230535 * x1 + 125922.46573224581 + 2 * 36823.00971204951
batas bawah: -0.058704985460230535 * x1 + 125922.46573224581 - 2 * 36823.00971204951
```

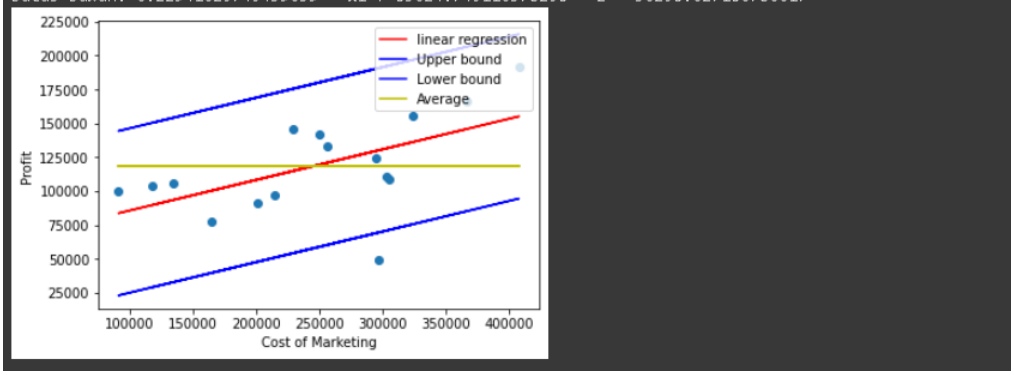


Gambar 18. Grafik State Florida Relasi Pengeluaran Administrasi dengan Profit

Kita dapat melihat grafik sebaran data dari dana pengeluaran administrasi kurang cocok dengan model regresi linear. Dari visualisasi diatas, kita dapat melihat bahwa perbedaan batas atas dan batas bawah dengan nilai regresi dari regresi ini sangat besar. Kita juga dapat melihat bahwa banyak titik-titik data yang memiliki jarak yang sangat jauh dengan garis regresi sehingga model ini dianggap tidak akan terlalu akurat untuk dijadikan model prediksi.

2.5.2.3 Regresi Linear Sederharna antara Profit dengan Marketing Spen

```
Rumus regresi linear = 0.22541029749435035*x1 + 63024.749110378296
list error: [442840834.2113979, 213941098.41710737, 138232863.63310018, 27413152.629936595, 388985372.5
batas atas : 0.22541029749435035 * x1 + 63024.749110378296 + 2 * 30296.027160780017
batas bawah: 0.22541029749435035 * x1 + 63024.749110378296 - 2 * 30296.027160780017
```



Gambar 19. Grafik State Florida Relasi Pengeluaran Marketing dengan Profit

Dengan melihat grafik regresi linear di atas, kita dapat mengevaluasi bahwa model regresi linear cukup dapat menggambarkan meskipun tidak secocok dengan data R&D. Dari visualisasi diatas, kita dapat melihat bahwa ada sebuah data yang melebihi batas bawah. Kemungkinan besar, ini merupakan data outlier. Jika kita melakukan data preprocessing dan menghilangkan data outlier tersebut, mungkin kita akan mendapatkan visualisasi yang lebih menggambarkan relasi linear antara data marketing spend dan profit dengan batas atas dan batas bawah yang lebih sempit daripada yang divisualisasikan.

2.5.2.4 R^2 Score

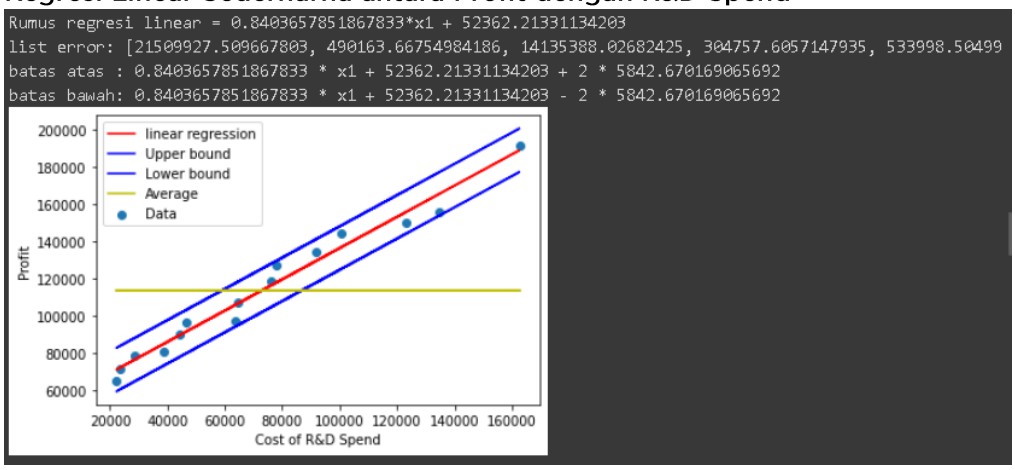
```
percentage of performance for R&D vs profit is 94.44005122465843
percentage of performance for Administration vs profit is 50.04364057934407
percentage of performance for marketing vs profit is 59.675402554577396
```

Gambar 20. R^2 Score Florida

Dapat dilihat bahwa data R&D cocok dimasukan kedalam regresi linear untuk melakukan prediksi nilai profit karena memiliki R2 score sebesar 94.4%. Dapat kita lihat bahwa nilai R2 score dari marketing memiliki penurunan yang lumayan jauh dibandingkan dengan R2 score dari model regresi dengan data marketing yang belum dibagi berdasarkan state yaitu hampir 9% dari 68% menjadi 59.6%. Hal ini menunjukan bahwa dalam state Florida, marketing spend tidak memiliki relasi linear yang bagus dengan profit. Akan tetapi Administrasi tetap tidak menunjukan hubungan yang linear dengan profit karena R2 scorenya yang tergolong rendah yaitu 50%.

2.5.3 CALIFORNIA

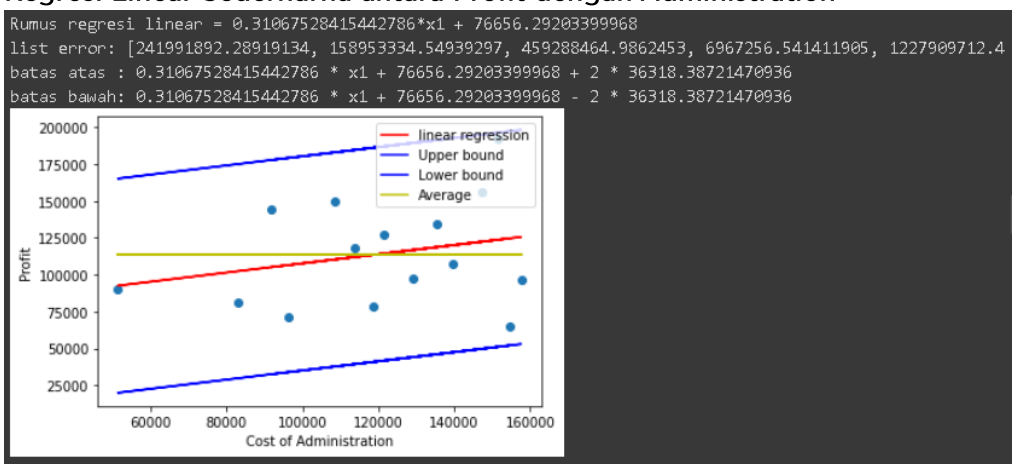
2.5.3.1 Regresi Linear Sederharna antara Profit dengan R&D Spend



Gambar 21. Grafik State California Relasi Pengeluaran R&D dengan Profit

Hanya melalui gambar grafik diatas ini, kita dapat melihat bahwa model regresi linear sangat cocok untuk menggambarkan hubungan antara profit dengan R&D spend. Oleh karena itu, model ini sangat cocok dalam meprediksi nilai profit dari sebuah perusahaan start up.

2.5.3.2 Regresi Linear Sederharna antara Profit dengan Administration

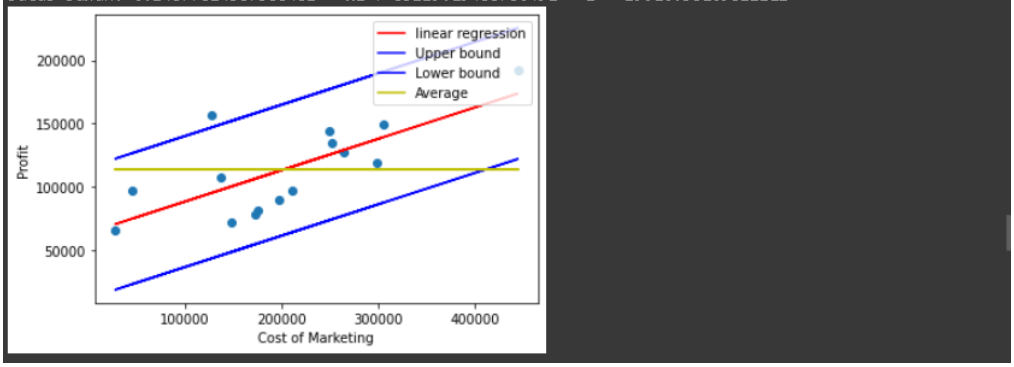


Gambar 22. Grafik State California Relasi Pengeluaran Administrasi dengan Profit

Kita dapat melihat grafik sebaran data dari dana pengeluaran administrasi kurang cocok dengan model regresi linear. Dari visualisasi diatas, kita dapat melihat bahwa perbedaan batas atas dan batas bawah dengan nilai regresi dari regresi ini sangat besar. Kita juga dapat melihat bahwa banyak titik-titik data yang memiliki jarak yang sangat jauh dengan garis regresi sehingga model ini dianggap tidak akan terlalu akurat untuk dijadikan model prediksi.

2.5.3.3 Regresi Linear Sederharna antara Profit dengan Marketing Spen

```
Rumus regresi linear = 0.24877824387366462*x1 + 63225.15486780451
list error: [67644991.95184578, 97162160.34805265, 663341319.4525522, 496962153.37395513, 814881030.138
batas atas : 0.24877824387366462 * x1 + 63225.15486780451 + 2 * 25910.36105822212
batas bawah: 0.24877824387366462 * x1 + 63225.15486780451 - 2 * 25910.36105822212
```



Gambar 23. Grafik State California Relasi Pengeluaran Marketing dengan Profit

Dengan melihat grafik regresi linear di atas, kita dapat mengevaluasi bahwa model regresi linear cukup dapat menggambarkan meskipun tidak secocok dengan data R&D. Dari visualisasi diatas, kita dapat melihat bahwa ada sebuah data yang melebihi batas atas. Kemungkinan besar, ini merupakan data outlier. Jika kita melakukan data preprocessing dan menghilangkan data outlier tersebut, mungkin kita akan mendapatkan visualisasi yang lebih menggambarkan relasi linear antara data marketing spend dan profit dengan batas atas dan batas bawah yang lebih sempit daripada yang divisualisasikan.

2.5.3.4 R^2 Score

```
percentage of performance for R&D vs profit is 97.64147021358623
percentage of performance for Administration vs profit is 51.72410069506661
percentage of performance for marketing vs profit is 67.79471815125756
```

Gambar 24. R^2 Score California

Dapat dilihat bahwa data R&D cocok dimasukan kedalam regresi linear untuk melakukan prediksi nilai profit karena memiliki R2 score sebesar 97.6% dan R2 score ini merupakan nilai yang paling tinggi dibandingkan dengan 2 state lainnya. Walaupun tidak memiliki nilai R2 score yang setinggi R&D, marketing juga memiliki kecocokan yang lumayan bagus yaitu 67.7%. Akan tetapi Administrasi tetap tidak memiliki hubungan yang tidak terlalu linear dengan profit karena R2 scorenya yang tergolong rendah yaitu 51.7%.

3. Kesimpulan

Secara keseluruhan, kita dapat dengan yakin menyimpulkan bahwa dari data-data perusahaan startup, hanya R&D dan Marketing spend yang memiliki relasi linear yang jelas dengan profit. Administration tidak memiliki relasi linear yang bagus dengan profit suatu perusahaan startup. Ini merupakan hal yang masuk akal di dunia nyata juga. Semakin bagus sebuah product yang tentunya dikembangkan melalui riset (R&D) dan semakin banyak orang yang mengetahui mengenai product yang dijual (marketing) tentu saja akan membuahkan profit yang setimpal juga. Dari pembagian data berdasarkan state, kita dapat melihat bahwa R&D memiliki relasi linear yang sangat bagus dengan profit di California dan Marketing memiliki relasi linear yang sangat bagus dengan profit di New York. Hal ini bukan merupakan suatu kejutan bagi kelompok kami. Walaupun memang R&D memiliki nilai R2 score yang baik di setiap state akan tetapi California memiliki nilai yang paling tinggi dan hal ini terjadi bukan tanpa alasan. California merupakan tempat yang populer dengan Silicon Valley yang merupakn tempat penghasil banyak sekali perusahaan teknologi yang sangat terkenal seperti Apple, Facebook dan lain-lain. Banyak sekali perusahaan start up dibidang teknologi di California, tentu saja teknologi harus dikembangkan agar dapat menjawab kebutuhan pasar. Jika perusahaan-perusahaan tersebut dapat mengikuti ombak pasar dengan R&D yang bagus maka tentu saja penjualan akan semakin meningkat (profit meningkat). Di New York, marketing spend memiliki relasi yang sangat bagus dengan profit. Hal ini merupakan hal yang lazim. Kita sendiri tahu bahwa New York sangat terkenal dengan Times Square, dimana banyak sekali Ads marketing yang bisa ditampilkan dalam layar besar sepanjang jalan tersebut. New York merupakan state yang padat penduduknya, melakukan marketing di sana tentu saja akan memberikan profit yang sangat tinggi karena dapat menjangkau banyak konsumen dibandingkan state-state lainnya.

Data-data ini memang masih tergolong sangat kecil, perlu lebih banyak lagi data agar dapat melihat secara lebih luas lagi. Seperti pada kasus data administration, pada data yang kita punya, data ini tidak membentuk apapun, melainkan hanya pure chaos. Akan tetapi hal ini mungkin dapat berubah jika kita mengambil lebih banyak data, misalnya 1000 data. Ada kemungkinan bahwa data administration yang kita miliki mempunyai banyak data outlier sehingga relasi linear dengan profit tidak begitu bagus.