Data yang digunakan pada Tugas 3 ini bersumber dari Kagle – Black Friday Dataset. Kemudian, dilakukan *preprocessing data* untuk menghasilkan dan membandingkan 4 model regresi linier yang berbeda, meliputi:

1. Model 1 : Data yang telah diatasi *missing value*
2. Model 2 : Data yang telah diatasi *missing value* dan *outlier*
3. Model 3 : Data yang telah diatasi *missing value* dan ditransformasi
4. Model 4 : Data yang telah diatas *missing value*, *outlier*, dan ditransformasi

Variabel respon dan prediktor yang digunakan pada tugas ini, yaitu :

y = Purchase (jumlah pendapatan dari pembelian produk dalam satuan dollar)

x1 = Produk kategori 1

x2 = Produk kategori 2

Sehingga, dari ke-4 model tersebut menghasilkan model regresi linier sebagai berikut.

1. Model 1

**Regression Analysis: Purchase versus Product\_Category; Product\_Category**

The regression equation is

Purchase = 12717 - 528 Product\_Category\_1 - 57,2 Product\_Category\_2

Predictor Coef SE Coef T P

Constant 12717,0 382,2 33,27 0,000

Product\_Category\_1 -527,68 45,09 -11,70 0,000

Product\_Category\_2 -57,23 38,06 -1,50 0,133

S = 4726,25 R-Sq = 15,0% R-Sq(adj) = 14,8%

1. Model 2

**Regression Analysis: Purchase versus Product\_Category; Product\_Category**

The regression equation is

Purchase = 12915 - 738 Product\_Category\_1 + 7,9 Product\_Category\_2

Predictor Coef SE Coef T P

Constant 12914,7 361,9 35,68 0,000

Product\_Category\_1 -738,46 47,08 -15,68 0,000

Product\_Category\_2 7,89 36,51 0,22 0,829

S = 4468,90 R-Sq = 22,8% R-Sq(adj) = 22,6%

1. Model 3

**Regression Analysis: Purchase versus Product\_Category; Product\_Category**

The regression equation is

Purchase = 1,00 - 0,00767 Product\_Category\_1 - 0,00739 Product\_Category\_2

Predictor Coef SE Coef T P

Constant 1,00001 0,00000 1535679,56 0,000

Product\_Category\_1 -0,0076674 0,0007354 -10,43 0,000

Product\_Category\_2 -0,0073916 0,0006414 -11,53 0,000

S = 0,0000152623 R-Sq = 80,3% R-Sq(adj) = 80,3%

1. Model 4

**Regression Analysis: Purchase versus Product\_Category; Product\_Category**

The regression equation is

Purchase = 1,00 - 0,00826 Product\_Category\_1 - 0,00693 Product\_Category\_2

Predictor Coef SE Coef T P

Constant 1,00001 0,00000 1497635,81 0,000

Product\_Category\_1 -0,0082634 0,0007745 -10,67 0,000

Product\_Category\_2 -0,0069345 0,0006716 -10,33 0,000

S = 0,0000152986 R-Sq = 80,5% R-Sq(adj) = 80,4%

Perbandingan kebaikan dari 4 model regresi :

|  |  |  |
| --- | --- | --- |
| Model ke- | S | R-Sq |
| 1 | 4726,25 | 15,0% |
| 2 | 4468,90 | 22,8% |
| 3 | 0,0000152623 | 80,3% |
| 4 | 0,0000152986 | 80,5% |

Bedasarkan tabel diatas diketahui bahwa model ke-4 memiliki nilai R-Square tertinggi yaitu sebesar 80,5% yang berarti bahwa data yang telah diatasi *missing value*, *outlier*, dan ditransformasi memiliki model regresi linier yang paling baik. Sehingga, dapat disimpulkan bahwa produk kategori 1 dan produk kategori 2 berpengaruh terhadap jumlah pendapatan sebesar 80,5%. Oleh karena itu, *preprocessing data* perlu dilakukan terlebih dahulu untuk memastikan data yang akan diolah tidak terdapat *missing value*, *outlier*, dll. Karena semakin berkualitas sebuah data akan menghasilkan model yang semakin baik juga.