|  |
| --- |
| **Linear Regression** |



|  |  |
| --- | --- |
| **Title: Machine Learning with Spark** | **Advisor: KiKi** |
| **Name: Seungsoo Lee** | **Date: 22-01-2020** |

**Index**

1. **Training Data**
2. **Calculation for LinearRegression**
3. **Data Visualization**
4. **Linear Regression Equation**
5. **Solve the Question**

**1. Training Data**

|  |  |
| --- | --- |
| year | salary |
| 1.1 | 39343 |
| 1.3 | 46205 |
| 1.5 | 37731 |
| 2 | 43525 |
| 2.2 | 39891 |
| 2.9 | 56642 |
| 3 | 60150 |
| 3.2 | 54445 |
| 3.2 | 64445 |
| 3.7 | 57189 |
| 3.9 | 63218 |
| 4 | 55794 |
| 4 | 56957 |
| 4.1 | 57081 |
| 4.5 | 61111 |
| 4.9 | 67938 |
| 5.1 | 66029 |
| 5.3 | 83088 |
| 5.9 | 81363 |
| 6 | 93940 |
| 6.8 | 91738 |
| 7.1 | 98273 |
| 7.9 | 101302 |
| 8.2 | 113812 |
| 8.7 | 109431 |
| 9 | 105582 |
| 9.5 | 116969 |
| 9.6 | 112635 |
| 10.3 | 122391 |
| 10.5 | 121872 |

**2. Calculation for Linear Regression**

|  |  |  |  |
| --- | --- | --- | --- |
| y-y' | (y-y')^2 | (x-x')^2 | (x-x')(y-y') |
| -36660 | 1343955600 | 17.75218 | 154460.8 |
| -29798 | 887920804 | 16.10684 | 119589.3 |
| -38272 | 1464745984 | 14.54151 | 145943.9 |
| -32478 | 1054820484 | 10.97818 | 107610.4 |
| -36112 | 1304076544 | 9.692844 | 112428.7 |
| -19361 | 374848321 | 5.824178 | 46724.55 |
| -15853 | 251317609 | 5.351511 | 36673.27 |
| -21558 | 464747364 | 4.466178 | 45559.24 |
| -11558 | 133587364 | 4.466178 | 24425.91 |
| -18814 | 353966596 | 2.602844 | 30353.25 |
| -12785 | 163456225 | 1.997511 | 18069.47 |
| -20209 | 408403681 | 1.724844 | 26541.15 |
| -19046 | 362750116 | 1.724844 | 25013.75 |
| -18922 | 358042084 | 1.472178 | 22958.69 |
| -14892 | 221771664 | 0.661511 | 12112.16 |
| -8065 | 65044225 | 0.170844 | 3333.533 |
| -9974 | 99480676 | 0.045511 | 2127.787 |
| 7085 | 50197225 | 0.000178 | -94.4667 |
| 5360 | 28729600 | 0.344178 | 3144.533 |
| 17937 | 321735969 | 0.471511 | 12316.74 |
| 15735 | 247590225 | 2.210178 | 23392.7 |
| 22270 | 495952900 | 3.192178 | 39789.07 |
| 25299 | 640039401 | 6.690844 | 65440.08 |
| 37809 | 1429520481 | 8.332844 | 109142 |
| 33428 | 1117431184 | 11.46951 | 113209.5 |
| 29579 | 874917241 | 13.59151 | 109047.9 |
| 40966 | 1678213156 | 17.52818 | 171511 |
| 36632 | 1341903424 | 18.37551 | 157029.2 |
| 46388 | 2151846544 | 24.86684 | 231321.5 |
| 45869 | 2103965161 | 26.90151 | 237907.2 |

**3. Data Visualization**

**4. Linear Regression Equation**

M = [∑](https://terms.naver.com/entry.nhn?docId=2073787&cid=47324&categoryId=47324) ( (x-x’)(y-y’) / (x-x’)^2 ) = 9449.962

C = y’ – Mx’

* 76003 – 9449.962 \* 5.31333 = 25792.2

**5. Solve the Question**

**Q1 Salary in 15 years?**

15 \* M + C = 15 \* 9449.962 + 25792.2 = 167,541.63‬

**Q2 Salary in 20 years?**

20 \* M + C = 20 \* 9449.962 + 25792.2 = 214,791.44