

Kennesaw State University
IS 8935 Business Intelligence: Traditional & Big Data Analytics
Dr. Reza Vaezi
Assignment 6
February 26, 2019

Linear Regression Analysis of Salary Requirements

By Karis Kim

Executive Summary

The goal of this report is to offer the management salary predictions, wage offer list and a projected budget in preparation for the acquisition of a company engaged in a similar line of business. The analysis predicts that the total weekly salary will be \$594,716 with the average employee's weekly salary being \$975. A wage offer list is provided based on average of prediction wage by Education and IQ. However, the model was underfit for the available data set, so this report recommends obtaining additional data with attributes like employees' job performance assessment, number of projects or duties assigned, and position rank in the company to yield a better fitting linear regression model.

CONTENTS

| | |
|--|----|
| Business Understanding | 3 |
| Assumptions | 3 |
| Data Understanding..... | 3 |
| Attribute Information | 3 |
| Data Assumptions..... | 5 |
| Data Preparation | 5 |
| Data Ranges for Attributes in Scoring set | 5 |
| Designate Target, ID Attribute | 6 |
| Modeling..... | 7 |
| Linear Regression..... | 7 |
| Apply Model on the Training set | 8 |
| Results of Linear Regression on Training Set | 8 |
| Performance (Regression) | 9 |
| Results of Performance of Linear Regression | 10 |
| Apply Model on the Scoring Set | 10 |
| Results of Linear Regression on Scoring Set | 11 |
| Evaluation of Findings | 13 |
| Business Recommendations | 15 |
| References | 17 |

BUSINESS UNDERSTANDING

Company A, for which this analysis is prepared, is acquiring Company B that is engaged in a similar line of business. This report aims to use current salary information to predict salary requirements and offers for the Company B employees who will be joining Company A, along with a total weekly salary budget that will help Company A management plan and prepare for the acquisition.

ASSUMPTIONS

This analysis and business recommendations (particularly with regard to total budget) are premised on the assumption that all employees in Company B will be joining Company A upon acquisition, and that the data set is a comprehensive and representative record of the employees aforementioned.

DATA UNDERSTANDING

The Wages data set contains 11 attributes, where each example/row represents a unique employee. The dataset represents weekly hours worked per employee along with their weekly wages, in addition to several additional information/attributes on each employee as listed below in the attribute information. All data set values are in the data type integer.

ATTRIBUTE INFORMATION

| | Attribute | Description |
|----|------------|--|
| 1 | EmployeeID | Employee ID (unique integer) |
| 2 | Wage | Weekly wages in dollars |
| 3 | Hours | Hours worked per week |
| 4 | IQ | Employee's IQ |
| 5 | Educ | Employee's education in terms of number of years |
| 6 | Exper | Number of years of job related experience |
| 7 | Tenure | Number of years being with the current employer |
| 8 | Age | Age of employee |
| 9 | Married | Marital status (0 = no, 1 = yes) |
| 10 | Urban | Residence status (0 = no, 1 = yes) |
| 11 | Sibs | Number of siblings |

The data set is divided into Wages_Training data set and Wages_Scoring data set for linear regression analysis. The Wages_Training data set contains 320 examples and 11 attributes, while the Wages_Scoring data set contains 615 examples and all attributes except the Wage attribute, which will be the label or predictor attribute.

Figure 1. Descriptive statistics of Training data set

Views: Design **Results** Turbo Prep More

Find data, operators, etc. All Studio

Result History: **ExampleSet (/Local Repository/Wages_Training)**

| Name | Type | Missing | Min | Max | Average |
|------------|---------|---------|------|------|----------|
| EmployeeID | Integer | 0 | 1001 | 1320 | 1160.500 |
| Wage | Integer | 0 | 233 | 2771 | 971.837 |
| Hours | Integer | 0 | 20 | 80 | 43.878 |
| IQ | Integer | 0 | 50 | 134 | 100.912 |
| Educ | Integer | 0 | 9 | 18 | 13.447 |
| Exper | Integer | 0 | 1 | 23 | 11.700 |
| Tenure | Integer | 0 | 0 | 21 | 7.072 |
| Age | Integer | 0 | 28 | 38 | 32.947 |
| Married | Integer | 0 | 0 | 1 | 0.884 |
| Urban | Integer | 0 | 0 | 1 | 0.738 |
| Sibs | Integer | 0 | 0 | 13 | 2.878 |

Showing attributes 1 - 11

Examples: 320 Special Attributes: 0 Regular Attributes: 11

Figure 2. Descriptive statistics of Scoring data set

Views: Design **Results** Turbo Prep More

Find data, operators, etc. All Studio

Result History: **ExampleSet (/Local Repository/Wages_Scoring)** ExampleSet (/Local Repository/Wages_Training)

| Name | Type | Missing | Min | Max | Average |
|------------|---------|---------|-----|-----|---------|
| EmployeeID | Integer | 0 | 101 | 715 | 408 |
| Hours | Integer | 0 | 24 | 80 | 43.956 |
| IQ | Integer | 0 | 55 | 145 | 101.475 |
| Educ | Integer | 0 | 9 | 18 | 13.480 |
| Exper | Integer | 0 | 1 | 23 | 11.493 |
| Tenure | Integer | 0 | 0 | 22 | 7.319 |
| Age | Integer | 0 | 28 | 38 | 33.150 |
| Married | Integer | 0 | 0 | 1 | 0.898 |
| Urban | Integer | 0 | 0 | 1 | 0.707 |
| Sibs | Integer | 0 | 0 | 14 | 2.974 |

Showing attributes 1 - 10

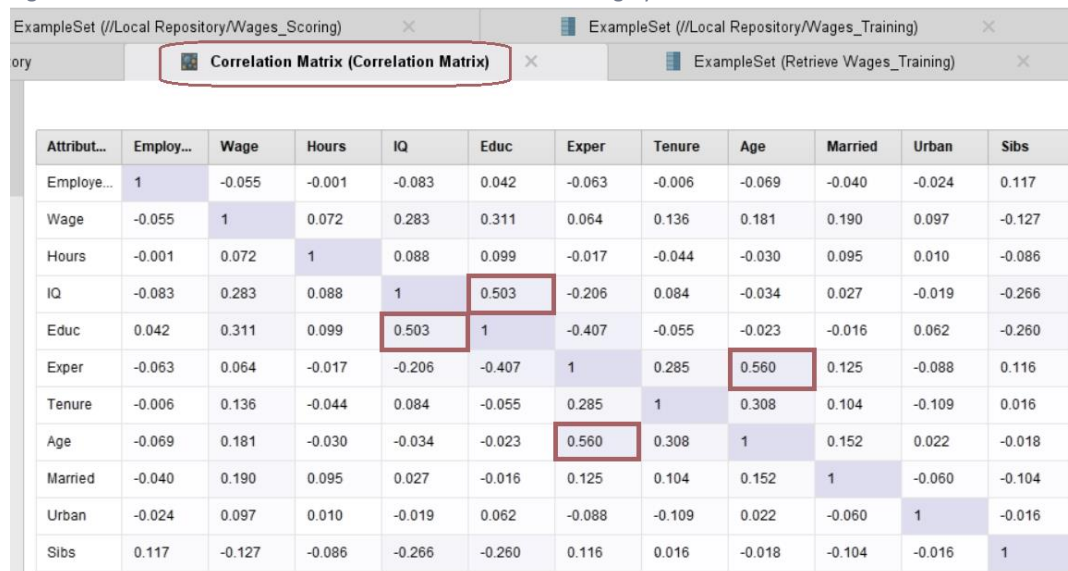
Examples: 615 Special Attributes: 0 Regular Attributes: 10

DATA ASSUMPTIONS

This report commences on the assumption that there will be a linear relationship between the dependent/target variable (Wages) and other independent variables/attributes, and that the variables follow a normal distribution.

Another assumption is that independent variables are not highly correlated. To test the validity of this assumption, a Correlation Matrix operator was launched in RapidMiner process. Figure 3 below shows that the top correlation is between Experience and Age, but at 0.56, it merely showed some correlation and verified that there are no independent attributes that are highly correlated. The next top correlation is between Education and IQ, but here also 0.503 barely showed “some” correlation.

Figure 3. Correlation Matrix to check that attributes are not highly correlated



| Attribut... | Employe... | Wage | Hours | IQ | Educ | Exper | Tenure | Age | Married | Urban | Sibs |
|-------------|------------|--------|--------|--------|--------|--------|--------|--------|---------|--------|--------|
| Employe... | 1 | -0.055 | -0.001 | -0.083 | 0.042 | -0.063 | -0.006 | -0.069 | -0.040 | -0.024 | 0.117 |
| Wage | -0.055 | 1 | 0.072 | 0.283 | 0.311 | 0.064 | 0.136 | 0.181 | 0.190 | 0.097 | -0.127 |
| Hours | -0.001 | 0.072 | 1 | 0.088 | 0.099 | -0.017 | -0.044 | -0.030 | 0.095 | 0.010 | -0.086 |
| IQ | -0.083 | 0.283 | 0.088 | 1 | 0.503 | -0.206 | 0.084 | -0.034 | 0.027 | -0.019 | -0.266 |
| Educ | 0.042 | 0.311 | 0.099 | 0.503 | 1 | -0.407 | -0.055 | -0.023 | -0.016 | 0.062 | -0.260 |
| Exper | -0.063 | 0.064 | -0.017 | -0.206 | -0.407 | 1 | 0.285 | 0.560 | 0.125 | -0.088 | 0.116 |
| Tenure | -0.006 | 0.136 | -0.044 | 0.084 | -0.055 | 0.285 | 1 | 0.308 | 0.104 | -0.109 | 0.016 |
| Age | -0.069 | 0.181 | -0.030 | -0.034 | -0.023 | 0.560 | 0.308 | 1 | 0.152 | 0.022 | -0.018 |
| Married | -0.040 | 0.190 | 0.095 | 0.027 | -0.016 | 0.125 | 0.104 | 0.152 | 1 | -0.060 | -0.104 |
| Urban | -0.024 | 0.097 | 0.010 | -0.019 | 0.062 | -0.088 | -0.109 | 0.022 | -0.060 | 1 | -0.016 |
| Sibs | 0.117 | -0.127 | -0.086 | -0.266 | -0.260 | 0.116 | 0.016 | -0.018 | -0.104 | -0.016 | 1 |

Note that of all the attributes, Wage has the highest correlation coefficients with Education and then IQ, but 0.311 and 0.283 respectively does not indicate the presence of a correlation.

DATA PREPARATION

Data Type Transformation: Linear regression models use numeric data types, and since all data types in the data set were integer (see figure 1 and 2), no data type transformation was needed.

Data Preparation of Missing Values: No missing values were found in the data set.

DATA RANGES FOR ATTRIBUTES IN SCORING SET

All data ranges for attributes in the scoring set must be within the range of those in the corresponding training set. Comparison of the scoring and training data sets showed that for attribute IQ, the training data set only had a maximum of 134, while the scoring data set had values up to 145. Similarly, the attribute Tenure in training data set has a maximum of 21, but the scoring data set has 22; the attribute Sibs in training has a maximum of 13, but the scoring has 14.

Data Preparation for Out of Range Attributes: In RapidMiner, Filter Examples operator was used to remove those out of range observations from the data set. In Filter Examples Parameters, click Add Filters, and in the Create Filters window, enter the maximum values of IQ, Tenure and Sibs in from the training data set, so that all examples less than or equal to those maximum values may be passed through to the output (see figure 4).

Figure 4. Filter Examples operator to remove observations out of range from scoring set

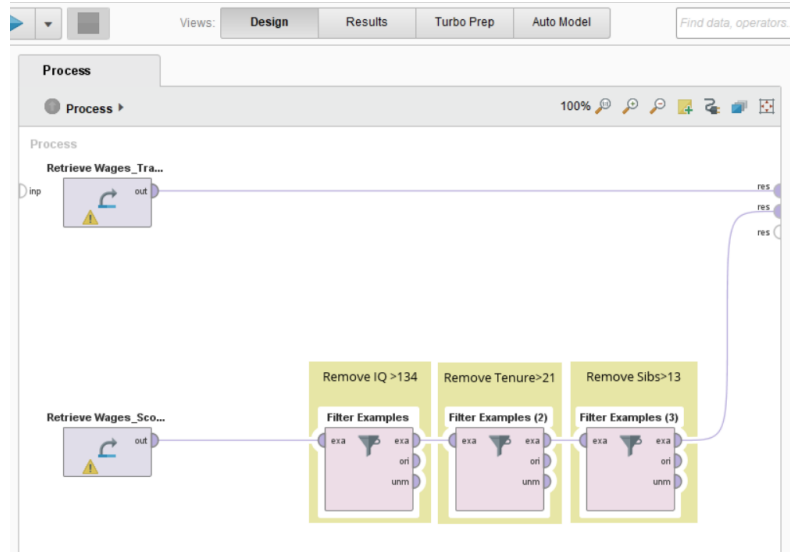


Figure 5. Results showing 610 examples in scoring set after removing out of range observations

Result History: ExampleSet (Filter Examples (3)) | ExampleSet (Set Role) | ExampleSet (Retrieve Wages_Scoring)

Open in: Turbo Prep | Auto Model

Filter (610 / 610 examples): all

| Row No. | EmployeeID | Hours | IQ ↓ | Educ | Exper | Tenure | Age | Married |
|---------|------------|-------|------|------|-------|--------|-----|---------|
| 48 | 149 | 40 | 134 | 18 | 10 | 10 | 37 | 1 |
| 68 | 169 | 40 | 134 | 13 | 9 | 9 | 30 | 0 |
| 137 | 238 | 40 | 132 | 17 | 8 | 9 | 32 | 1 |
| 169 | 270 | 43 | 132 | 18 | 8 | 13 | 38 | 1 |
| 210 | 312 | 40 | 132 | 17 | 8 | 1 | 33 | 1 |
| 151 | 252 | 48 | 131 | 12 | 7 | 1 | 28 | 0 |
| 230 | 332 | 40 | 131 | 18 | 9 | 10 | 33 | 1 |
| 313 | 416 | 40 | 131 | 16 | 10 | 2 | 33 | 1 |
| 453 | 556 | 40 | 131 | 16 | 9 | 3 | 30 | 0 |
| 527 | 632 | 50 | 130 | 14 | 15 | 1 | 33 | 1 |
| 9 | 109 | 75 | 129 | 18 | 8 | 12 | 38 | 1 |
| 252 | 354 | 40 | 129 | 18 | 4 | 5 | 36 | 1 |
| 323 | 426 | 40 | 129 | 14 | 9 | 3 | 32 | 0 |

ExampleSet (610 examples) 0 special attributes, 10 regular attributes

DESIGNATE TARGET, ID ATTRIBUTE

In RapidMiner, using Set Role operator, Wage attribute will be designated as the target attribute or the label in the training data set. In the scoring data set, the Wage attribute is omitted, so there is no need to designate Wage as the label. However, in both the training and scoring sets, EmployeeID attribute needs to be designated as ID, so that EmployeeID is not used in the model. A second Set Role operator is added to the scoring set for this function (see figure 6).

Figure 6. Set Role operators to designate Wage as label, EmployeeID as id

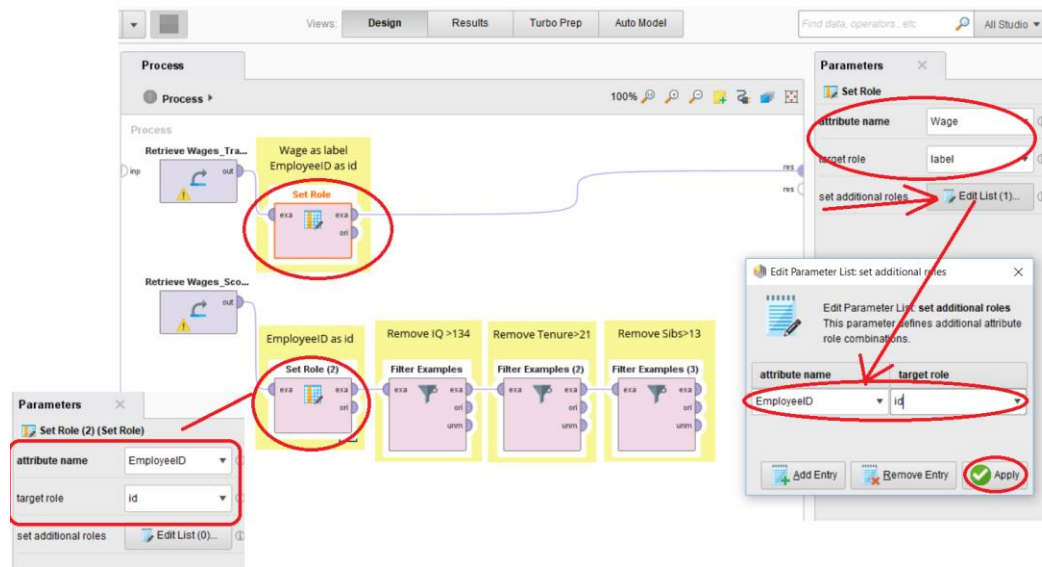


Figure 7. Results showing ID attribute in blue column, label attribute in green column

The screenshot shows the Results view with a data table. The **ExampleSet (Set Role)** window is active. The table has 18 rows and 12 columns. The **EmployeeID** column is highlighted in blue, and the **Wage** column is highlighted in green. Red arrows point to these columns with the text: "EmployeeID in blue indicates ID attribute" and "Wage in green indicates label attribute".

| Row No. | EmployeeID | Wage | Hours | IQ | Educ | Exper | Tenure | Age | Married | Urban | Sibs |
|---------|------------|------|-------|-----|------|-------|--------|-----|---------|-------|------|
| 1 | 1001 | 722 | 55 | 113 | 16 | 8 | 1 | 38 | 0 | 1 | 1 |
| 2 | 1002 | 1000 | 40 | 112 | 14 | 15 | 4 | 35 | 1 | 1 | 4 |
| 3 | 1003 | 425 | 30 | 106 | 17 | 8 | 0 | 36 | 0 | 0 | 0 |
| 4 | 1004 | 978 | 45 | 89 | 12 | 17 | 2 | 38 | 1 | 1 | 0 |
| 5 | 1005 | 600 | 40 | 87 | 10 | 20 | 7 | 36 | 1 | 0 | 0 |
| 6 | 1006 | 1444 | 40 | 120 | 16 | 12 | 15 | 37 | 1 | 1 | 1 |
| 7 | 1007 | 622 | 45 | 119 | 16 | 12 | 1 | 33 | 1 | 1 | 0 |
| 8 | 1008 | 756 | 40 | 78 | 14 | 9 | 2 | 31 | 1 | 1 | 2 |
| 9 | 1009 | 1250 | 40 | 102 | 13 | 13 | 11 | 30 | 1 | 1 | 5 |
| 10 | 1010 | 1444 | 40 | 109 | 15 | 11 | 1 | 29 | 1 | 1 | 2 |
| 11 | 1011 | 850 | 40 | 72 | 11 | 19 | 10 | 38 | 1 | 1 | 2 |
| 12 | 1012 | 1212 | 40 | 98 | 12 | 11 | 12 | 28 | 0 | 1 | 4 |
| 13 | 1013 | 1250 | 40 | 95 | 12 | 12 | 2 | 32 | 1 | 1 | 2 |
| 14 | 1014 | 2137 | 45 | 102 | 13 | 8 | 2 | 30 | 1 | 0 | 0 |
| 15 | 1015 | 578 | 40 | 109 | 13 | 7 | 4 | 30 | 0 | 1 | 1 |
| 16 | 1016 | 1000 | 60 | 101 | 12 | 11 | 6 | 30 | 1 | 1 | 3 |
| 17 | 1017 | 800 | 40 | 88 | 12 | 16 | 7 | 38 | 1 | 1 | 3 |
| 18 | 1018 | 750 | 70 | 85 | 10 | 13 | 0 | 35 | 1 | 1 | 3 |

ExampleSet (320 examples: 2 special attributes, 9 regular attributes)

MODELING

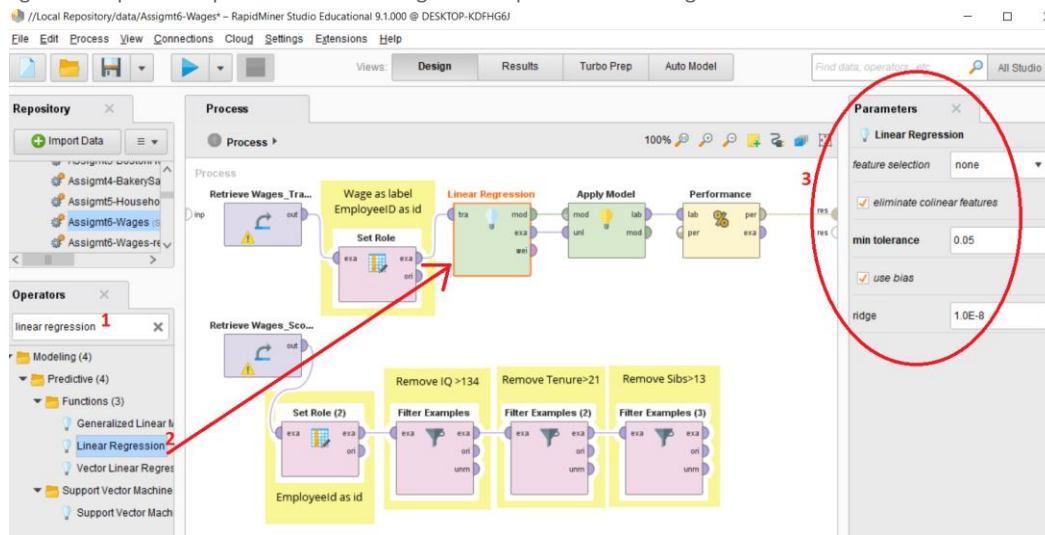
Following data preparation, the linear regression modeling process will involve running the Linear Regression operator on the training set, then running the Apply Model operator, and the Performance (Regression) operator to measure the performance of the linear regression model. Then the model can be deployed on the scoring set.

LINEAR REGRESSION

First, the Linear Regression operator is added to the training set. In the Parameters, feature selection is set to none and all defaults settings are maintained (see figure 8). Setting feature selection to none will prevent RapidMiner from removing least significant factors from the model. Checking the eliminate colinear features will

enable RapidMiner to remove factors that are linearly correlated from the modeling process. Checking the use bias will enable RapidMiner to build a model with an intercept.

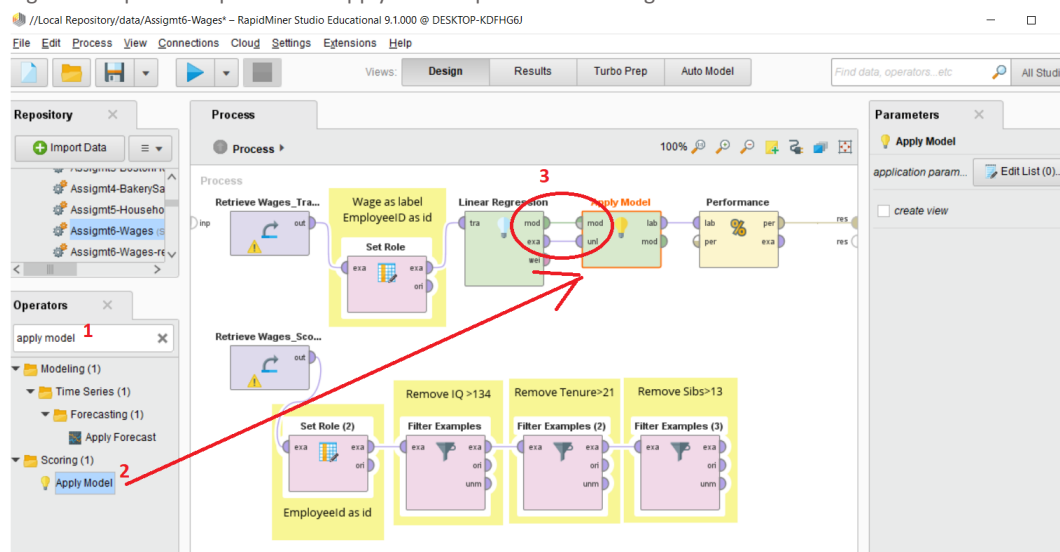
Figure 8. RapidMiner process for Linear Regression operator to training set



APPLY MODEL ON THE TRAINING SET

Next, the Apply Model operator is added after the Linear Regression operator to apply the model to the training set.

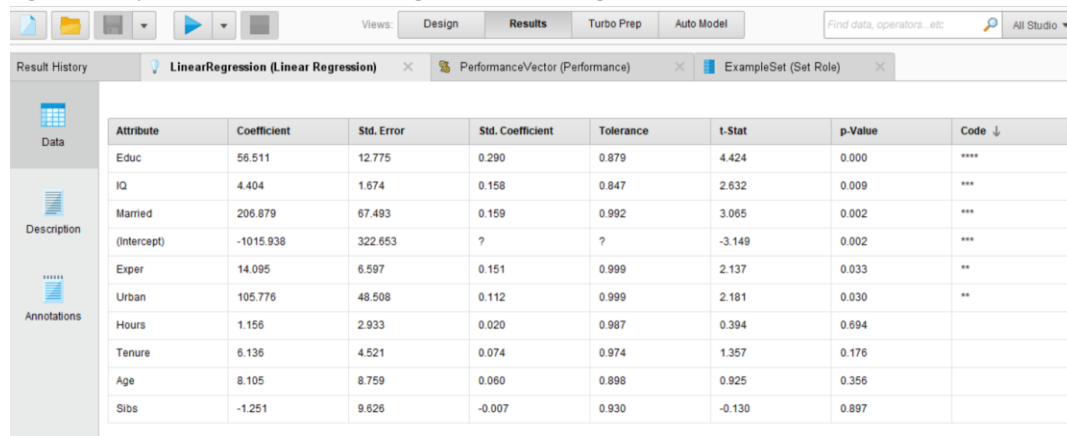
Figure 9. RapidMiner process for Apply Model operator to training set



RESULTS OF LINEAR REGRESSION ON TRAINING SET

At this point, the process can be run to generate the following results in Figure 10. In the Results view, in the Linear Regression tab, doubling clicking the Code column will sort the attributes according to decreasing levels of significance. Attributes with 4 stars is highly significant, while any star below 2 should be disregarded. Figure 10 shows that Education is the most significant, followed by IQ and Married. The Description of Linear Regression shows a list of the coefficients of the linear regression function for each attribute (see figure 11).

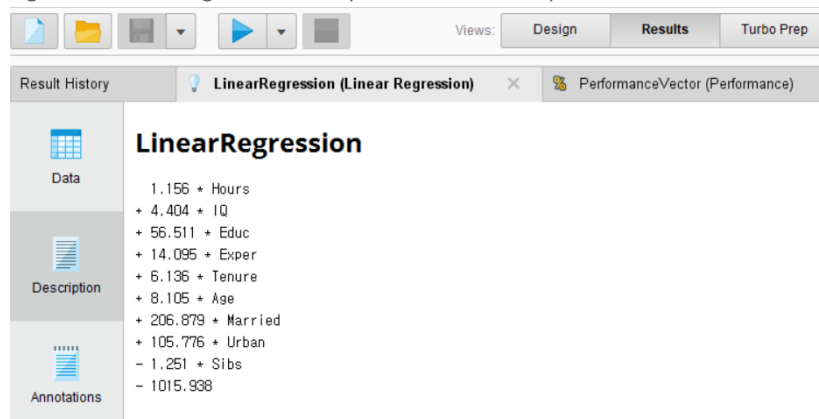
Figure 10. RapidMiner Result of Linear Regression on training set



The screenshot shows the 'Results' view in RapidMiner for a Linear Regression model. The left sidebar has 'Data', 'Description', and 'Annotations' tabs. The main area displays a table with the following data:

| Attribute | Coefficient | Std. Error | Std. Coefficient | Tolerance | t-Stat | p-Value | Code ↓ |
|-------------|-------------|------------|------------------|-----------|--------|---------|--------|
| Educ | 56.511 | 12.775 | 0.290 | 0.879 | 4.424 | 0.000 | **** |
| IQ | 4.404 | 1.674 | 0.158 | 0.847 | 2.632 | 0.009 | *** |
| Married | 206.879 | 67.493 | 0.159 | 0.992 | 3.065 | 0.002 | *** |
| (Intercept) | -1015.938 | 322.653 | ? | ? | -3.149 | 0.002 | *** |
| Exper | 14.095 | 6.597 | 0.151 | 0.999 | 2.137 | 0.033 | ** |
| Urban | 105.776 | 48.508 | 0.112 | 0.999 | 2.181 | 0.030 | ** |
| Hours | 1.156 | 2.933 | 0.020 | 0.987 | 0.394 | 0.694 | |
| Tenure | 6.136 | 4.521 | 0.074 | 0.974 | 1.357 | 0.176 | |
| Age | 8.105 | 8.759 | 0.060 | 0.898 | 0.925 | 0.356 | |
| Sibs | -1.251 | 9.626 | -0.007 | 0.930 | -0.130 | 0.897 | |

Figure 11. Linear Regression Description of coefficients per attribute and the intercept



The screenshot shows the 'Description' view in RapidMiner for a Linear Regression model. The left sidebar has 'Data', 'Description', and 'Annotations' tabs. The main area displays the following text:

LinearRegression

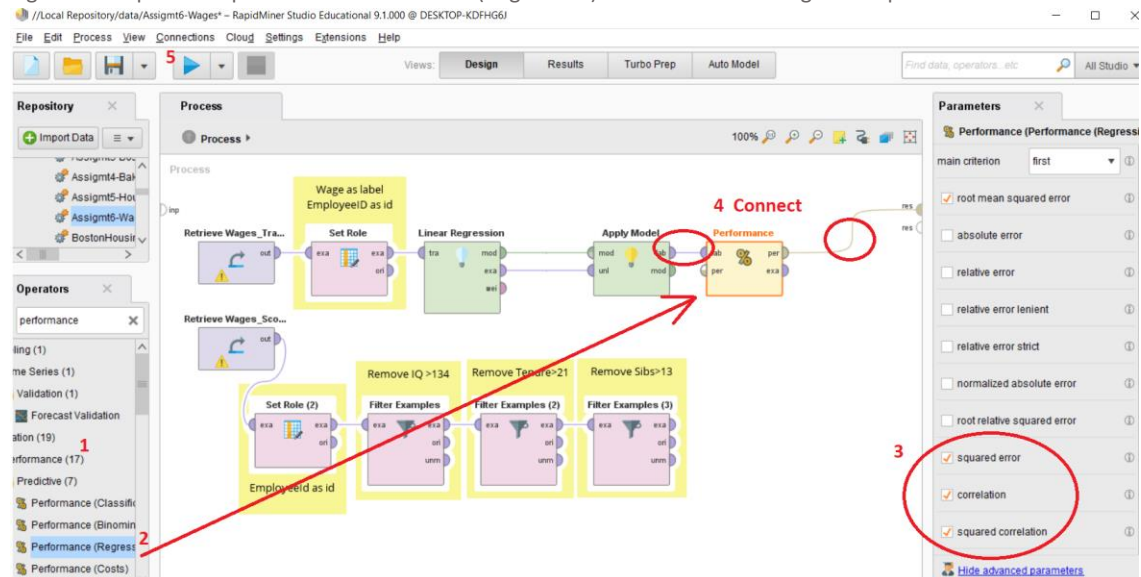
- 1.156 * Hours
- + 4.404 * IQ
- + 56.511 * Educ
- + 14.095 * Exper
- + 6.136 * Tenure
- + 8.105 * Age
- + 206.879 * Married
- + 105.776 * Urban
- 1.251 * Sibs
- 1015.938

PERFORMANCE (REGRESSION)

Before applying the Linear Regression model to the scoring set, the Performance (Regression) operator is added to measure how good the model fits, with the squared correlation (or R^2). Squared correlation values can be from 0.0 to 1.0, and better models will be closer to 1.0. Our result view shows that the squared correlation is 0.209 (see figure 13). Low values below 0.2 generally mean that attributes in model do not explain the prediction outcome satisfactorily [1]. However, social and behavioral science models typically do accept low values [1].

In the Performance (Regression) operator's Parameters, *squared error*, *correlation*, and *squared correlation* are selected before input and output are connected appropriately, and the process is run (see figure 12).

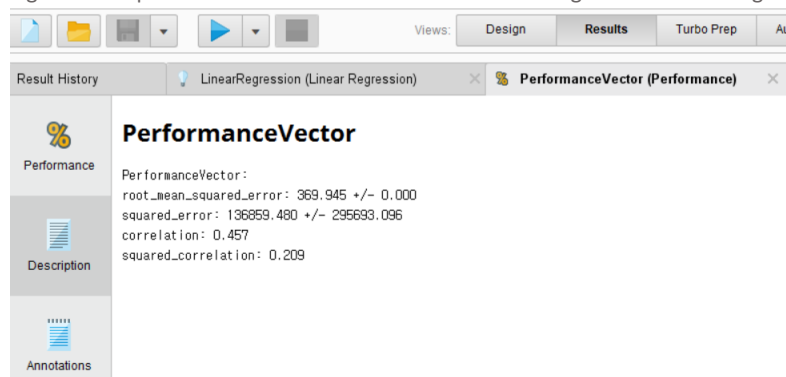
Figure 12. RapidMiner process for Performance (Regression) to measure linear regression performance



RESULTS OF PERFORMANCE OF LINEAR REGRESSION

The Performance Vector Description lists the measurements in one view (figure 13).

Figure 13. RapidMiner Result of Performance of Linear Regression on training set



APPLY MODEL ON THE SCORING SET

Now the model will be applied to the scoring set, by connecting the Apply Model operator's unlabeled input port with the examples from the scoring data set stream (see figure 14). The process of applying the Linear Regression model on the scoring set will yield a data table of the example set with the prediction(Wage) column containing the predicted wages per employee for the 610 examples in the scoring set (see figure 16).

In order to also provide the business actionable intelligence from the prediction that can help management prepare for the acquisition, an Aggregate operator is added to the process to find the sum of the predicted wages and the average of predicted wages per week (see figure 15).

Figure 14. Connect Apply Model operator to examples from the scoring set

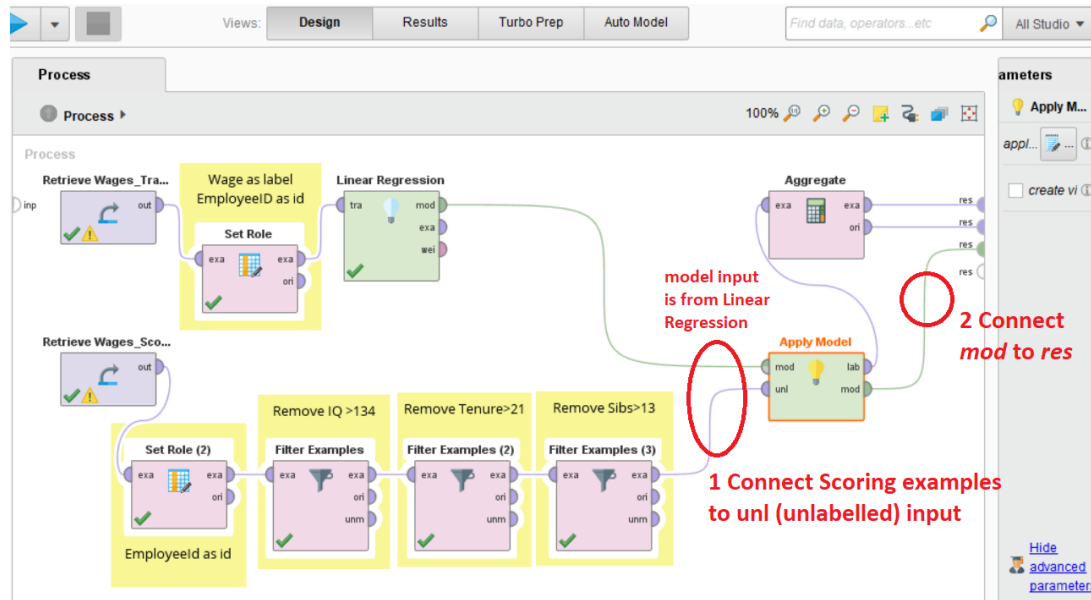
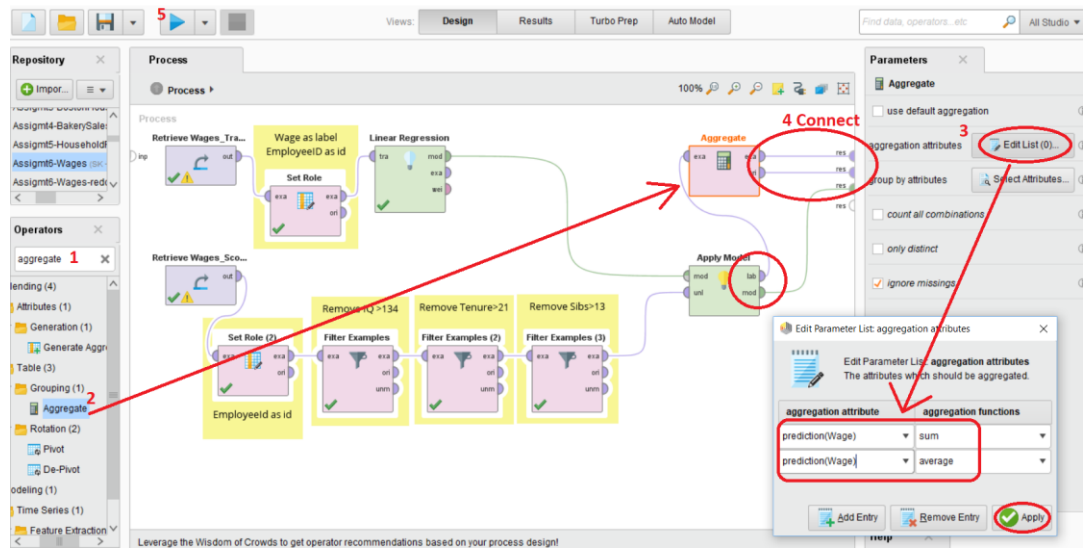


Figure 15. Add Aggregate operator to show sum and average of prediction wages



RESULTS OF LINEAR REGRESSION ON SCORING SET

The following Figure 16 shows the results of the linear regression model applied on the scoring set, with the prediction wages highlighted in the green column. The highest predicted weekly wage is \$1450.02 and the lowest predicted weekly wage is \$418.64. The descriptive statistics view of the scoring set with the prediction wages shows that the label variable follows a nice bell curve, but has a high standard deviation of 185.483 (see figure 17). The total sum of weekly predicted wages is \$594,716.11 and the average weekly predicted wages is \$974.94 (see figure 18).

Figure 16. Results of applying Linear Regression model on scoring set yielding prediction(Wages)

| Row No. | EmployeeID | prediction... | Hours | IQ | Educ | Exper | Tenure | Age | Married | Urban | Sibs |
|---------|------------|---------------|-------|-----|------|-------|--------|-----|---------|-------|------|
| 48 | 149 | 1450.020 | 40 | 134 | 18 | 10 | 10 | 37 | 1 | 1 | 2 |
| 520 | 625 | 1447.684 | 50 | 119 | 18 | 11 | 16 | 37 | 1 | 1 | 1 |
| 395 | 498 | 1445.127 | 55 | 125 | 18 | 12 | 12 | 34 | 1 | 1 | 1 |
| 360 | 463 | 1435.153 | 45 | 118 | 18 | 12 | 12 | 38 | 1 | 1 | 1 |
| 415 | 518 | 1418.826 | 45 | 121 | 18 | 14 | 3 | 38 | 1 | 1 | 3 |
| 116 | 217 | 1400.509 | 40 | 127 | 16 | 16 | 12 | 37 | 1 | 1 | 4 |
| 230 | 332 | 1384.039 | 40 | 131 | 18 | 9 | 10 | 33 | 1 | 1 | 7 |
| 220 | 322 | 1382.554 | 50 | 121 | 18 | 11 | 7 | 35 | 1 | 1 | 3 |
| 320 | 423 | 1366.396 | 40 | 114 | 18 | 8 | 14 | 38 | 1 | 1 | 2 |
| 9 | 109 | 1354.882 | 75 | 129 | 18 | 8 | 12 | 38 | 1 | 0 | 2 |
| 486 | 589 | 1354.308 | 55 | 127 | 18 | 10 | 3 | 32 | 1 | 1 | 1 |
| 528 | 633 | 1353.697 | 50 | 112 | 18 | 9 | 10 | 38 | 1 | 1 | 6 |
| 472 | 575 | 1341.727 | 47 | 122 | 18 | 6 | 13 | 34 | 1 | 1 | 3 |
| 169 | 270 | 1338.479 | 43 | 132 | 18 | 8 | 13 | 38 | 1 | 0 | 1 |
| 432 | 535 | 1337.145 | 55 | 123 | 17 | 10 | 7 | 36 | 1 | 1 | 1 |
| 281 | 383 | 1334.147 | 45 | 121 | 18 | 7 | 13 | 32 | 1 | 1 | 2 |
| 46 | 147 | 1315.704 | 40 | 120 | 16 | 16 | 3 | 37 | 1 | 1 | 3 |
| 43 | 143 | 1311.606 | 40 | 115 | 18 | 11 | 1 | 36 | 1 | 1 | 0 |

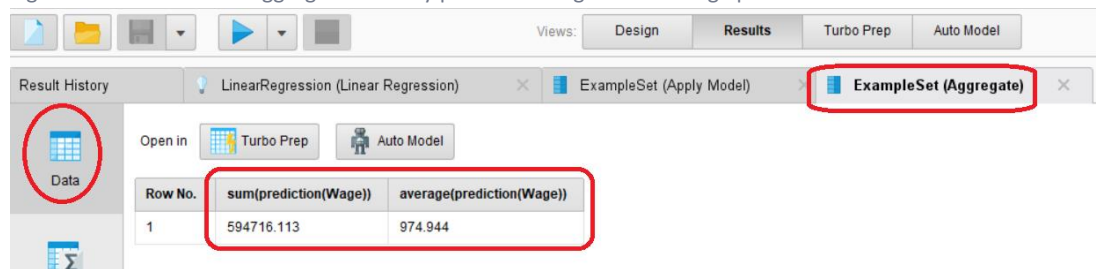
ExampleSet (610 examples, 2 special attributes, 9 regular attributes)

Figure 17. Results showing descriptive statistics on scoring set with prediction(Wages)

| Name | Type | Missing | Statistics |
|--------------------|---------|---------|---|
| ✓ Id EmployeeID | Integer | 0 | Min 101, Max 715, Average 407.982 |
| ✓ prediction(Wage) | Integer | 0 | Min 418.643, Max 1450.020, Average 974.944, Deviation 185.483 |
| ✓ Hours | Integer | 0 | Min 24, Max 80, Average 43.967 |
| ✓ IQ | Integer | 0 | Min 55, Max 134, Average 101.398 |
| ✓ Educ | Integer | 0 | Min 9, Max 18, Average 13.485 |
| ✓ Exper | Integer | 0 | Min 1, Max 23, Average 11.467 |
| ✓ Tenure | Integer | 0 | Min 0, Max 21, Average 7.274 |
| ✓ Age | Integer | 0 | Min 28, Max 38, Average 33.136 |
| ✓ Married | Integer | 0 | Min 0, Max 1, Average 0.897 |
| ✓ Urban | Integer | 0 | Min 0, Max 1, Average 0.707 |
| ✓ Sibs | Integer | 0 | Min 0, Max 13, Average 2.939 |

Showing attributes 1 - 11 Examples: 610 Special Attributes: 2 Regular Attributes: 9

Figure 18. Results of the aggregated weekly prediction wages and average per week



EVALUATION OF FINDINGS

The Linear Regression results data table (see figure 10) showed that Education and IQ were ranked at the top as the most significant factors, with a Code rating of 4 stars and 3 stars respectively. The following figure 18 shows a plot of Education and the target variable, prediction(Wage). Figure 19 shows a plot of IQ and the target variable, prediction(Wage).

Figure 18. Scatter plot of the top ranked variable *Education* and label *Prediction(Wage)*

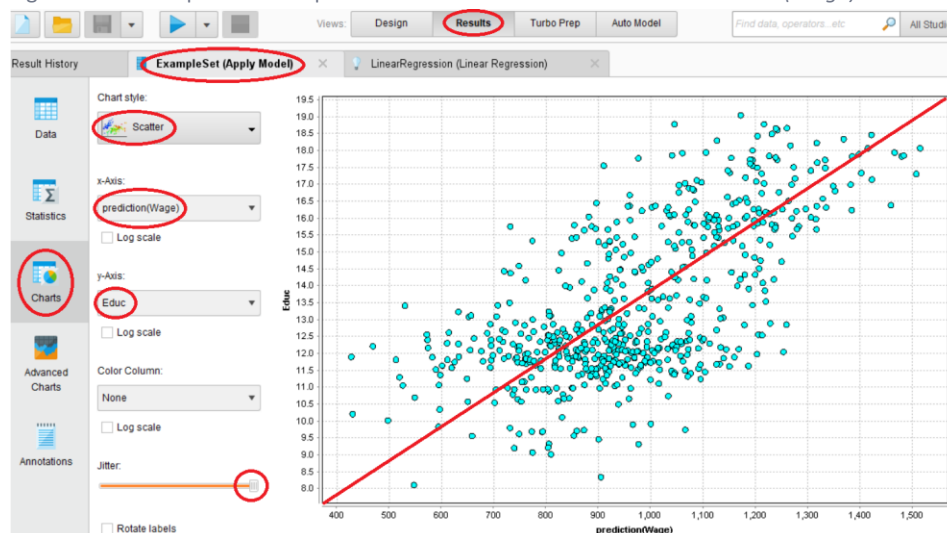
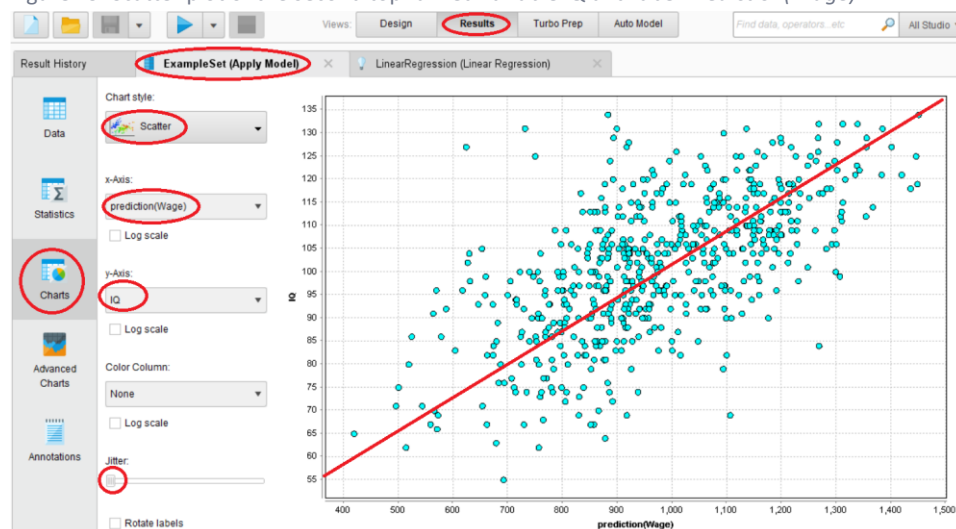


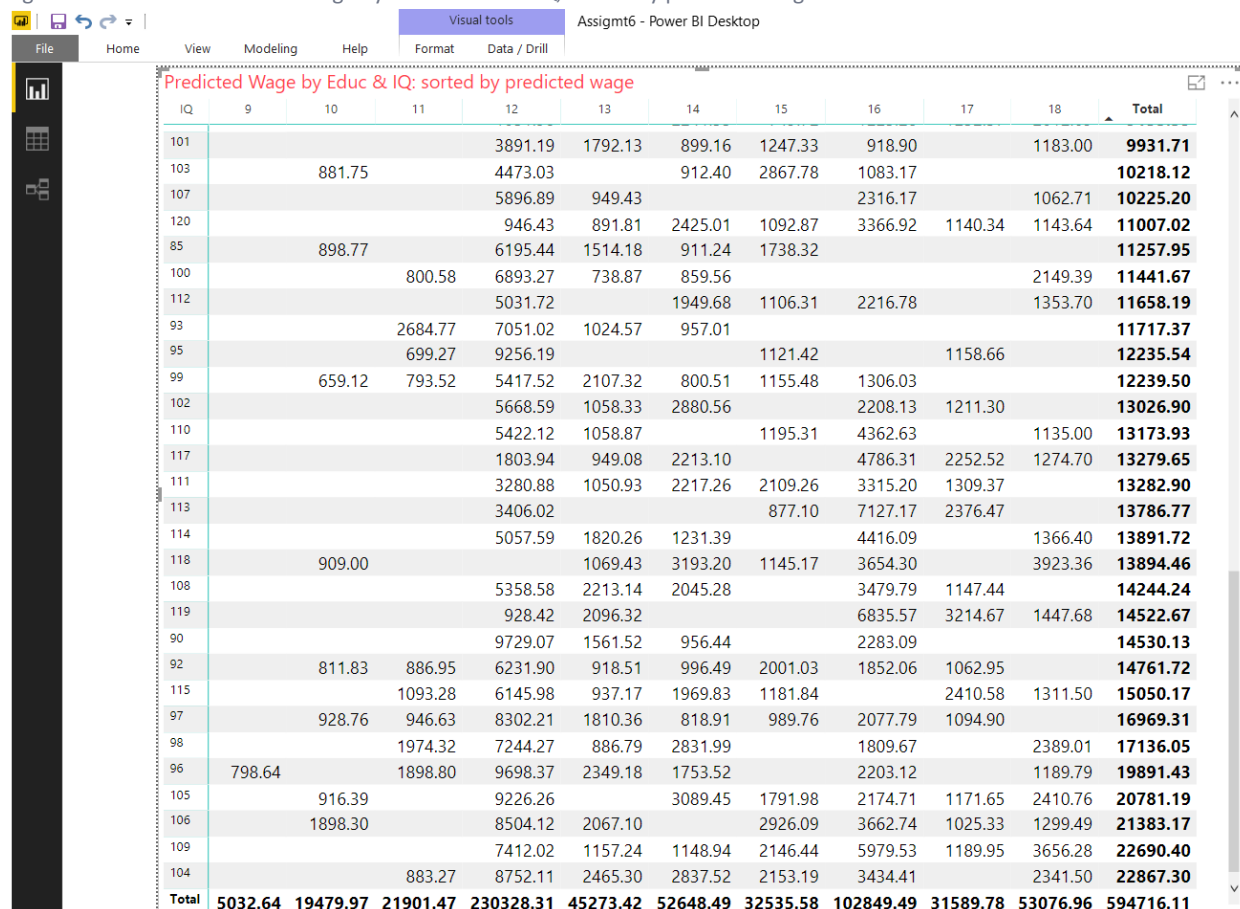
Figure 19. Scatter plot of the second top ranked variable *IQ* and label *Prediction(Wage)*



The scatter plots of Prediction(Wage) with Education and IQ does show a linear pattern, but the data points are loosely scattered around the linear regression line. In the previous section, the results of the Performance (Regression) vectors yielded a squared correlation of 0.209 (figure 13), which may be an indication that the attributes in this linear regression model do not explain the prediction outcomes very well, and we now see that the linear regression line underfits the data (figure 18, 19).

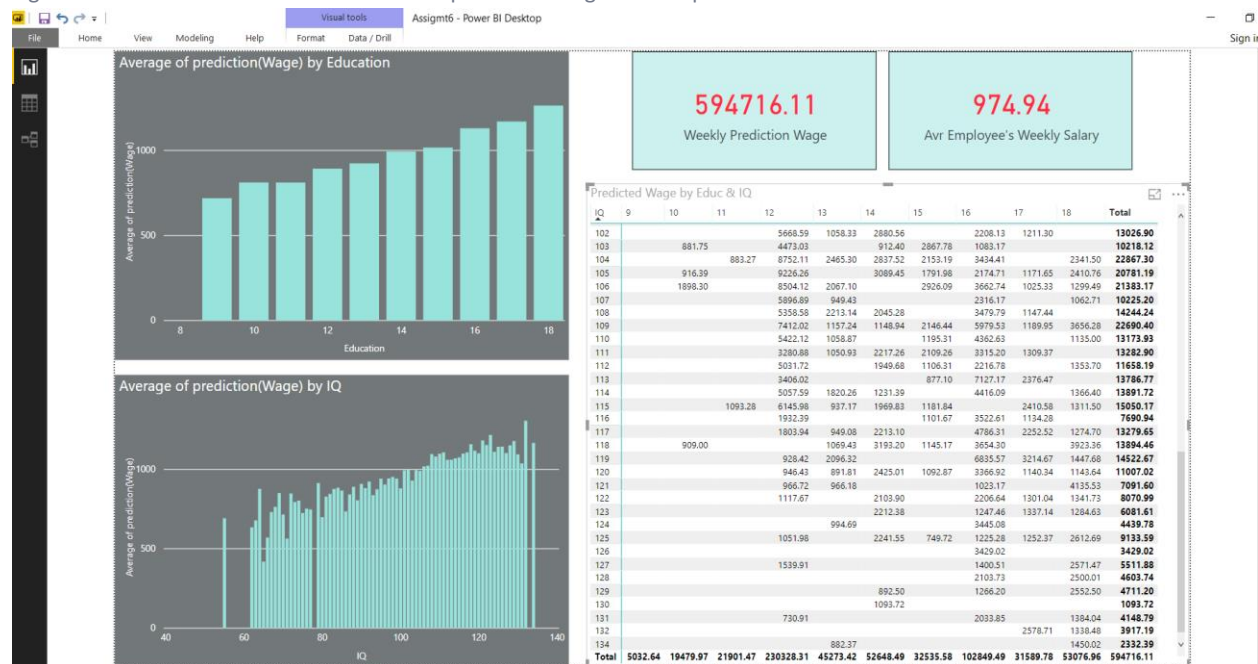
The following Power BI matrix shows Predicted(Wages) by Education and IQ, sorted by wages from low to high (figure 20). The matrix shows that high IQ does not necessarily yield high salary, and highest education does not entirely correspond to the highest salary. Just as the linear regression Performance and plot line revealed above, Education and IQ are not necessarily the best variables for wage prediction.

Figure 20. Power BI Predicted Wage by Education and IQ, sorted by predicted wage



| | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | Total |
|--------------|----------------|-----------------|-----------------|------------------|-----------------|-----------------|-----------------|------------------|-----------------|-----------------|------------------|
| 101 | | | | 3891.19 | 1792.13 | 899.16 | 1247.33 | 918.90 | | 1183.00 | 9931.71 |
| 103 | | 881.75 | | 4473.03 | | 912.40 | 2867.78 | 1083.17 | | | 10218.12 |
| 107 | | | | 5896.89 | 949.43 | | | 2316.17 | | 1062.71 | 10225.20 |
| 120 | | | | 946.43 | 891.81 | 2425.01 | 1092.87 | 3366.92 | 1140.34 | 1143.64 | 11007.02 |
| 85 | | 898.77 | | 6195.44 | 1514.18 | 911.24 | 1738.32 | | | | 11257.95 |
| 100 | | | 800.58 | 6893.27 | 738.87 | 859.56 | | | | 2149.39 | 11441.67 |
| 112 | | | | 5031.72 | | 1949.68 | 1106.31 | 2216.78 | | 1353.70 | 11658.19 |
| 93 | | | 2684.77 | 7051.02 | 1024.57 | 957.01 | | | | | 11717.37 |
| 95 | | | 699.27 | 9256.19 | | | 1121.42 | | 1158.66 | | 12235.54 |
| 99 | | 659.12 | 793.52 | 5417.52 | 2107.32 | 800.51 | 1155.48 | 1306.03 | | | 12239.50 |
| 102 | | | | 5668.59 | 1058.33 | 2880.56 | | 2208.13 | 1211.30 | | 13026.90 |
| 110 | | | | 5422.12 | 1058.87 | | 1195.31 | 4362.63 | | 1135.00 | 13173.93 |
| 117 | | | | 1803.94 | 949.08 | 2213.10 | | 4786.31 | 2252.52 | 1274.70 | 13279.65 |
| 111 | | | | 3280.88 | 1050.93 | 2217.26 | 2109.26 | 3315.20 | 1309.37 | | 13282.90 |
| 113 | | | | 3406.02 | | | 877.10 | 7127.17 | 2376.47 | | 13786.77 |
| 114 | | | | 5057.59 | 1820.26 | 1231.39 | | 4416.09 | | 1366.40 | 13891.72 |
| 118 | | 909.00 | | | 1069.43 | 3193.20 | 1145.17 | 3654.30 | | 3923.36 | 13894.46 |
| 108 | | | | 5358.58 | 2213.14 | 2045.28 | | 3479.79 | 1147.44 | | 14244.24 |
| 119 | | | | 928.42 | 2096.32 | | | 6835.57 | 3214.67 | 1447.68 | 14522.67 |
| 90 | | | | 9729.07 | 1561.52 | 956.44 | | 2283.09 | | | 14530.13 |
| 92 | | 811.83 | 886.95 | 6231.90 | 918.51 | 996.49 | 2001.03 | 1852.06 | 1062.95 | | 14761.72 |
| 115 | | | 1093.28 | 6145.98 | 937.17 | 1969.83 | 1181.84 | | 2410.58 | 1311.50 | 15050.17 |
| 97 | | 928.76 | 946.63 | 8302.21 | 1810.36 | 818.91 | 989.76 | 2077.79 | 1094.90 | | 16969.31 |
| 98 | | | 1974.32 | 7244.27 | 886.79 | 2831.99 | | 1809.67 | | 2389.01 | 17136.05 |
| 96 | 798.64 | | 1898.80 | 9698.37 | 2349.18 | 1753.52 | | 2203.12 | | 1189.79 | 19891.43 |
| 105 | | 916.39 | | 9226.26 | | 3089.45 | 1791.98 | 2174.71 | 1171.65 | 2410.76 | 20781.19 |
| 106 | | 1898.30 | | 8504.12 | 2067.10 | | 2926.09 | 3662.74 | 1025.33 | 1299.49 | 21383.17 |
| 109 | | | | 7412.02 | 1157.24 | 1148.94 | 2146.44 | 5979.53 | 1189.95 | 3656.28 | 22690.40 |
| 104 | | | 883.27 | 8752.11 | 2465.30 | 2837.52 | 2153.19 | 3434.41 | | 2341.50 | 22867.30 |
| Total | 5032.64 | 19479.97 | 21901.47 | 230328.31 | 45273.42 | 52648.49 | 32535.58 | 102849.49 | 31589.78 | 53076.96 | 594716.11 |

Figure 21. Power BI interactive dashboard of prediction wages with top two variables



BUSINES RECOMMENDATIONS

The findings of the report initially recommends that the company plan to budget for the total sum of weekly predicted wages of \$594,716.11 (\$30,925,232 per annum for employee wages) and the average weekly predicted wages of \$974.94 upon acquisition. A wage offer list is provided based on average of prediction wage by education (figure 22). A wage offer list based on average of prediction wage by IQ is also provided (figure 23).

Since the analysis revealed that the model was underfit for the available data set, this report recommends that the business obtain more data with other attributes that may potentially have a more significant relation to the label variable and yield a better linear regression result. Other attributes that may potentially be more helpful in making a linear regression based prediction that is neither underfit or overfit could be data/attributes on employees' job performance assessment, number of projects or duties assigned, position rank in the company, and so on.

Figure 22. Power BI wage offer list breakdown by education based on average of prediction wage

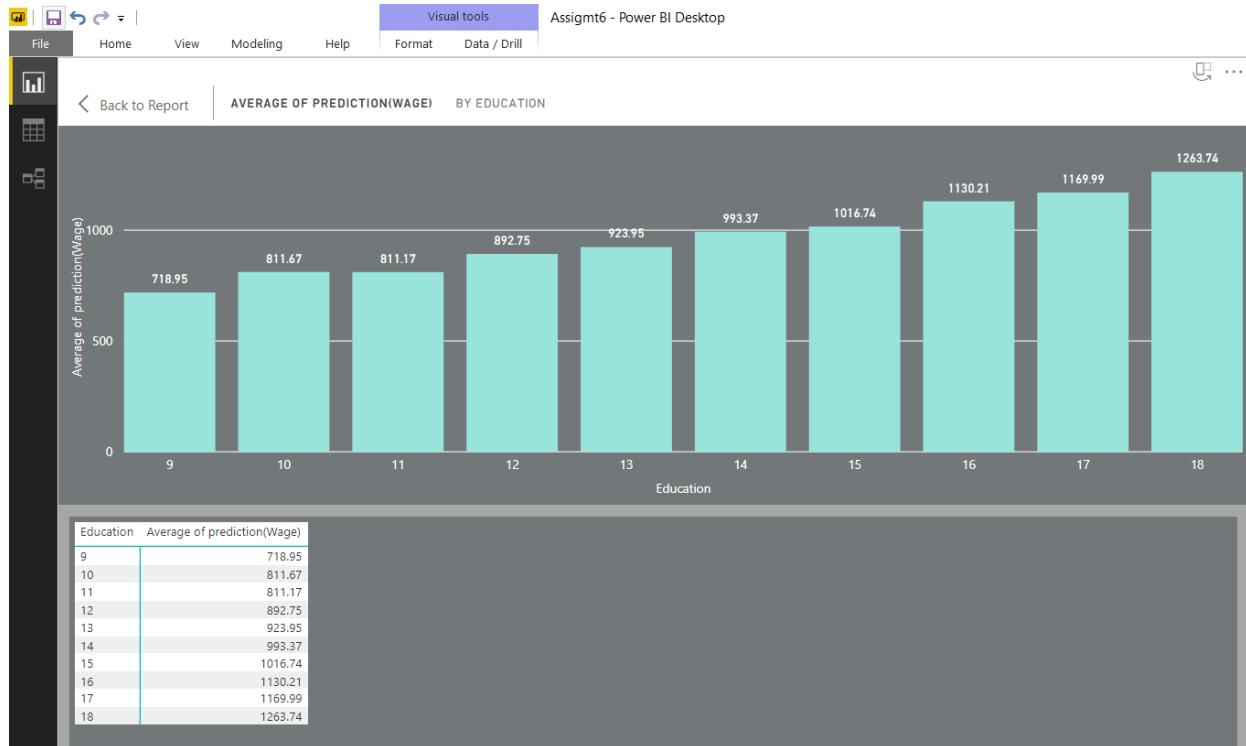
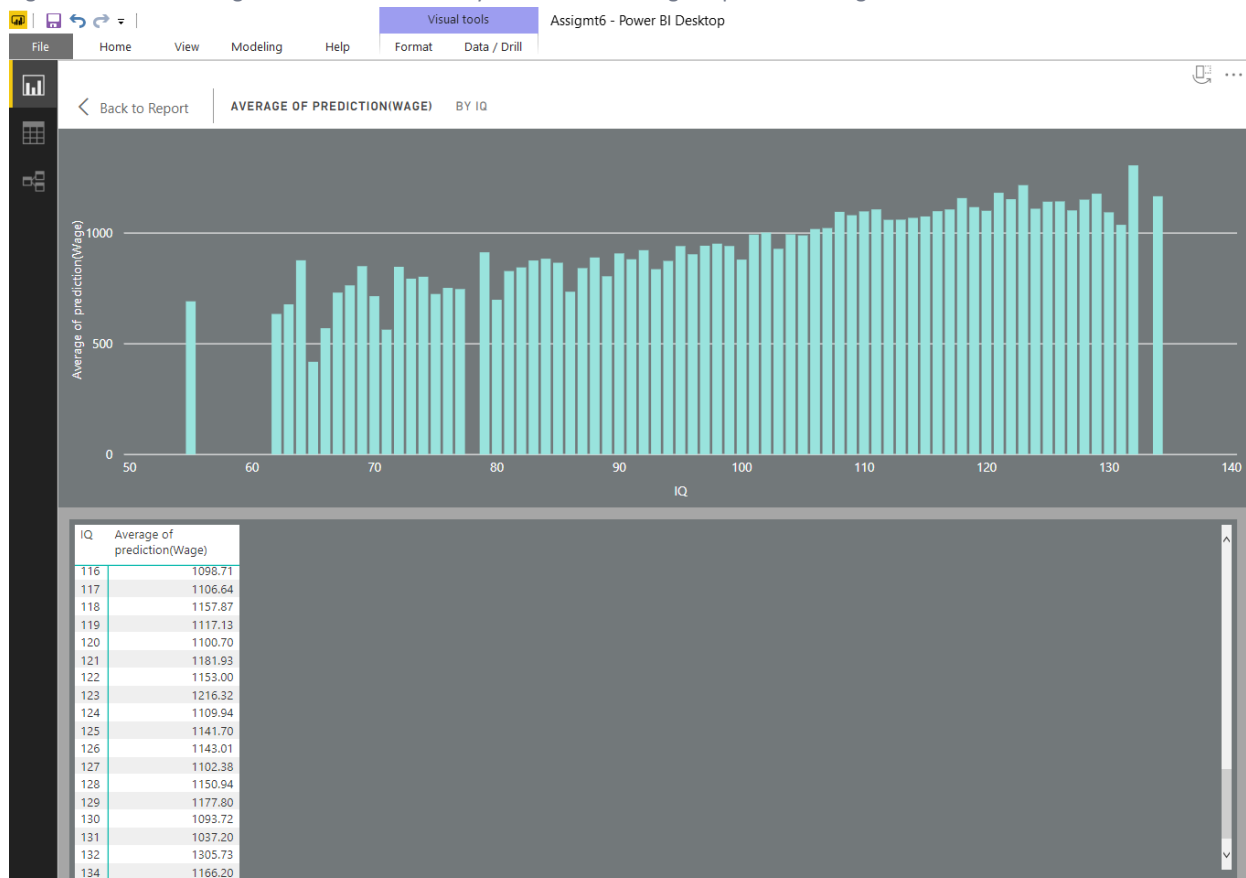


Figure 23. Power BI wage offer list breakdown by IQ based on average of prediction wage



REFERENCES

1. Vijay, K., & Bala, D. (2015). *Predictive Analytics and Data Mining*. Elsevier.