**Symbiosis Centre of Management Studies**

**NOIDA**

****

**Assessment Component-2**

**(2020-21)**

**Predictive Analytics**

**Aman Tibrewal (18021021023)**

**Divye Kant (18021021089)**

**Vridhi Wadhvan (18021021354)**

**Submitted to:**

**Dr Kriti Priya Gupta**

**Associate Professor**

# Objective: To formulate an accurate predictive linear relation model for the Average Monthly Working Hours in the Employee Data File

## 

## Case A: Employee hours

Company ABC has very poor employee satisfaction and retention. They have recently conducted a series of exit interviews to understand what went wrong and how they could make an impact on employee retention. A large number of the leaving employees indicated that they would have stayed if they were compensated with overtime pay for their extra hours. While Company ABC may not have been tracking employee hours this year, they do have a sample of previous employee data from an in-depth employee quiz performed 2 years ago. The information available for the sample employees includes currently available information such as:

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Variable** | **Description** | **Data Type** |
| 1. | satisfaction\_level | Satisfaction level of employees | Continuous |
| 2. | last\_evaluation | Evaluation received in the last appraisal | Continuous |
| 3. | number\_project | Number of projects worked on | Discrete |
| 4. | average\_montly\_hours | Average hours an employee is likely to work | Continuous |
| 5. | time\_spend\_company | Tenure (in years) | Continuous |
| 6. | Work\_accident | Whether accident happened during work (Yes=1, No=0) | Nominal |
| 7. | left | Left the company or not (Yes=1, No=0) | Nominal |
| 8. | promotion\_last\_5years | Whether received promotion during last 5 years (Yes=1, No=0) | Nominal |
| 9. | dept | Department where employee is working | Nominal |
| 10. | salary | Salary level | Nominal |

The company wants to estimate the average hours an employee is likely to work so that it can estimate how much money it would have to pay out to its employees. Specifically, the company wants to build a predictive model for the average hours an employee is likely to work based on the other factors.

### Assumptions Testing:

#### Normality Testing

|  |  |  |
| --- | --- | --- |
| **Variable** | **Kurtosis** | **Skewness** |
| **satisfaction\_level** | 2.477300 | -0.539013 |
| **last\_evaluation** | 1.820919 | -0.031801 |
| **number\_project** | 2.604474 | 0.332810 |
| **average\_montly\_hours** | 1.916970 | 0.027523 |
| **time\_spend\_company** | 8.132913 | 1.815038 |
| **Log\_TimeSpent** | 3.185936 | 0.544127 |
| **Inv\_TimeSpent** | 2.191859 | 0.236493 |
| **Imp\_timeSpend** | 2.841937 | 0.613427 |

Table 1Descriptive Statistics for the variables of Employee Data

For a variable to be normally distributed, the skewness and kurtosis values have to be within the following threshold:

* Skewness: Between -2 and +2
* Kurtosis: Between -7 and +7 for R

Thus, As per the Descriptive Statistics reported for the variables in Table 1, all the variables expect

#### Outliers Test

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Table 2 Outliers Test for the Variables

As per Table 2 Outliers Test for the Variables, the boxplots for all the variables, it can be established that most of the data does not have many outliers, except Bay Views and Rental Returns %.

#### Linear Relationship and Multicollinearity Test

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Price** | **Rooms** | **Lot Size** | **Age** | **Area** | **To Train** | **To Bus** | **To Shops** | **Street** |
| Price | 1.000 | .505\*\* | .411\*\* | -.363\*\* | .568\*\* | 0.003 | -0.024 | -0.084 | .723\*\* |
| Rooms | .505\*\* | 1.000 | 0.085 | -.504\*\* | .769\*\* | 0.124 | -0.064 | -0.026 | 0.118 |
| Lot Size | .411\*\* | 0.085 | 1.000 | -0.154 | 0.139 | -0.027 | 0.110 | 0.136 | .371\*\* |
| Age | -.363\*\* | -.504\*\* | -0.154 | 1.000 | -.653\*\* | -0.015 | 0.135 | 0.067 | 0.106 |
| Area | .568\*\* | .769\*\* | 0.139 | -.653\*\* | 1.000 | 0.090 | -.204\* | -0.151 | 0.064 |
| To Train | 0.003 | 0.124 | -0.027 | -0.015 | 0.090 | 1.000 | -0.002 | .306\*\* | -0.009 |
| To Bus | -0.024 | -0.064 | 0.110 | 0.135 | -.204\* | -0.002 | 1.000 | .406\*\* | 0.139 |
| To Shops | -0.084 | -0.026 | 0.136 | 0.067 | -0.151 | .306\*\* | .406\*\* | 1.000 | 0.051 |
| Street | .723\*\* | 0.118 | .371\*\* | 0.106 | 0.064 | -0.009 | 0.139 | 0.051 | 1.000 |
| Storeys | .565\*\* | 0.142 | 0.113 | -0.089 | 0.110 | -0.065 | -0.069 | -0.060 | .323\*\* |
| Bedrooms | .540\*\* | .916\*\* | 0.104 | -.600\*\* | .846\*\* | 0.128 | -0.087 | -0.038 | 0.106 |
| Bathrooms | .331\*\* | .635\*\* | 0.082 | -.341\*\* | .511\*\* | -0.020 | -0.038 | -0.093 | 0.043 |
| Bay Views | .676\*\* | 0.027 | .195\* | 0.070 | -0.044 | 0.069 | 0.140 | 0.009 | .647\*\* |
| Weekly Rent $ | .666\*\* | .566\*\* | .217\* | -.465\*\* | .594\*\* | 0.120 | -0.026 | -0.059 | .408\*\* |
| Rental Return % | -.404\*\* | 0.040 | -.236\*\* | -0.090 | -0.043 | .206\* | 0.050 | 0.121 | -.348\*\* |
| **Variables** | **Price** | **Storeys** | **Bedrooms** | **Bathrooms** | **Bay Views** | **Weekly Rent $** | **Rental Return %** |
| Price | 1.000 | .565\*\* | .540\*\* | .331\*\* | .676\*\* | .666\*\* | -.404\*\* |
| Rooms | .505\*\* | 0.142 | .916\*\* | .635\*\* | 0.027 | .566\*\* | 0.040 |
| Lot Size | .411\*\* | 0.113 | 0.104 | 0.082 | .195\* | .217\* | -.236\*\* |
| Age | -.363\*\* | -0.089 | -.600\*\* | -.341\*\* | 0.070 | -.465\*\* | -0.090 |
| Area | .568\*\* | 0.110 | .846\*\* | .511\*\* | -0.044 | .594\*\* | -0.043 |
| To Train | 0.003 | -0.065 | 0.128 | -0.020 | 0.069 | 0.120 | .206\* |
| To Bus | -0.024 | -0.069 | -0.087 | -0.038 | 0.140 | -0.026 | 0.050 |
| To Shops | -0.084 | -0.060 | -0.038 | -0.093 | 0.009 | -0.059 | 0.121 |
| Street | .723\*\* | .323\*\* | 0.106 | 0.043 | .647\*\* | .408\*\* | -.348\*\* |
| Storeys | .565\*\* | 1.000 | 0.163 | 0.120 | .434\*\* | .315\*\* | -.299\*\* |
| Bedrooms | .540\*\* | 0.163 | 1.000 | .676\*\* | 0.001 | .600\*\* | 0.042 |
| Bathrooms | .331\*\* | 0.120 | .676\*\* | 1.000 | -0.028 | .325\*\* | -0.009 |
| Bay Views | .676\*\* | .434\*\* | 0.001 | -0.028 | 1.000 | .234\* | -.401\*\* |
| Weekly Rent $ | .666\*\* | .315\*\* | .600\*\* | .325\*\* | .234\* | 1.000 | .337\*\* |
| Rental Return % | -.404\*\* | -.299\*\* | 0.042 | -0.009 | -.401\*\* | .337\*\* | 1.000 |

\*\*. Correlation is significant at the 0.01 level (2-tailed).

\*. Correlation is significant at the 0.05 level (2-tailed).

Table 3 Correlation b/w the variables

As per Table 3, It can be established that all the data is significantly correlated with Price except the variables (Marked with Yellow in the table): To Train, To Bus and, to shops; This is because the p-value for the 3 variables was greater than 0.05 (i.e., at 95% confidence interval). Thus, it can be established that the Price Variable has a linear relationship with the 13 other discrete and continuous variables.

Additionally, based upon Table 3, it can also be established that there is a multicollinearity between the 2 variables (Marked with Red in the table), i.e., Rooms and Bedrooms, as there is a significant correlation between the 2 variables and the r-value is greater than 0.9.

#### Testing the Linear Relation between the Categorical Variables and the Price Variable

**Suburb**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test of Homogeneity of Variances** | | | | | | | | | | |
|  | | | Levene Statistic | | | df1 | df2 | | Sig. | |
| Price ($'000) | Based on Mean | | 2.101 | | | 2 | 117 | | .127 | |
| **ANOVA** | | | | | | | | | | |
| Price ($'000) | | | | | | | | | | |
|  | | Sum of Squares | | df | Mean Square | | | F | | Sig. |
| Between Groups | | 1850893.907 | | 2 | 925446.954 | | | 10.106 | | .000 |
| Within Groups | | 10714355.418 | | 117 | 91575.687 | | |  | |  |
| Total | | 12565249.325 | | 119 |  | | |  | |  |

Table 4 ANOVA Result for Price and Suburb

Levene’s test (Levene statistic=2.101, p=0.127>0.05) indicates that the variances of prices in the three groups of Suburb are homogeneous (no significant difference in variance). Hence, Welch ANOVA will not be used, simple ANOVA will be used. As per the ANOVA Results (F= 10.106 and p=0.000<0.05), Since the p-value is less than 0.05 then we can conclude that there is a linear relation between price and suburb as there is significant impact of the suburb values on the price.

**Condition**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test of Homogeneity of Variances** | | | | | | | | | | |
|  | | | Levene Statistic | | | df1 | df2 | | Sig. | |
| Price ($'000) | Based on Mean | | 1.144 | | | 3 | 116 | | .334 | |
| **ANOVA** | | | | | | | | | | |
| Price ($'000) | | | | | | | | | | |
|  | | Sum of Squares | | df | Mean Square | | | F | | Sig. |
| Between Groups | | 1964770.203 | | 3 | 654923.401 | | | 7.167 | | .000 |
| Within Groups | | 10600479.122 | | 116 | 91383.441 | | |  | |  |
| Total | | 12565249.325 | | 119 |  | | |  | |  |

Levene’s test (Levene statistic=1.144, p=0.334>0.05) indicates that the variances of prices in the 4 groups of Condition are homogeneous (no significant difference in variance). Hence, Welch ANOVA will not be used, simple ANOVA will be used. As per the ANOVA Results (F= 7.167 and p=0.000<0.05), Since the p-value is less than 0.05 then we can conclude that there is a linear relation between price and suburb as there is significant impact of the Condition values on the price.

**Rental Status**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test of Homogeneity of Variances** | | | | | | | | | | |
|  | | | Levene Statistic | | | df1 | df2 | | Sig. | |
| Price ($'000) | Based on Mean | | .540 | | | 2 | 117 | | .584 | |
| **ANOVA** | | | | | | | | | | |
| Price ($'000) | | | | | | | | | | |
|  | | Sum of Squares | | df | Mean Square | | | F | | Sig. |
| Between Groups | | 150850.624 | | 2 | 75425.312 | | | .711 | | .493 |
| Within Groups | | 12414398.701 | | 117 | 106105.972 | | |  | |  |
| Total | | 12565249.325 | | 119 |  | | |  | |  |

Levene’s test (Levene statistic=0.545, p=0.584>0.05) indicates that the variances of prices in the 3 groups of Rental Status are homogeneous (no significant difference in variance). Hence, Welch ANOVA will not be used, simple ANOVA will be used. As per the ANOVA Results (F= 0.711 and p=0.493>0.05), Since the p-value is greater than 0.05 then we can conclude that there is no linear relation between price and suburb as there is no significant impact of the Rental Status values on the price.

**Style**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Independent Samples Test** | | | | | | | | | |
|  | Price ($'000) | | | | | | | | |
| Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| F | Sig. | t | df | Sig. (2tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Equal variances assumed | 5.925 | .016 | -4.255 | 118 | .000 | -236.545 | 55.598 | -346.644 | -126.445 |
| Equal variances not assumed |  |  | -4.173 | 100.601 | .000 | -236.545 | 56.682 | -348.992 | -124.097 |

Levene’s test (Levene statistic=5.925, p=0.016<0.05) indicates that the variances of prices in the 2 groups of Style are not homogeneous (significant difference in variance). Hence, it is established that there are no equal variances. As per the Independent Samples T-Test Results (T= -4.173 and p=0.000<0.05), Since the p-value is less than 0.05 then we can conclude that there is a linear relation between price and suburb as there is a significant impact of the Style values on the price.

**Kitchen**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Independent Samples Test** | | | | | | | | | |
|  | Price ($'000) | | | | | | | | |
| Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| F | Sig. | t | df | Sig. (2tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Equal variances assumed | .088 | .767 | .610 | 118 | .543 | 48.690 | 79.806 | -109.348 | 206.728 |
| Equal variances not assumed |  |  | .597 | 26.617 | .555 | 48.690 | 81.543 | -118.735 | 216.115 |

Levene’s test (Levene statistic=0.088, p=0.767>0.05) indicates that the variances of prices in the 2 groups of Kitchen are homogeneous (no significant difference in variance). Hence, it is established that there are equal variances. As per the Independent Samples T-Test Results (T= 0.610 and p=0.543>0.05), Since the p-value is greater than 0.05 then we can conclude that there is no linear relation between price and Kitchen as there is no significant impact of the Kitchen values on the price.

**Heating**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Independent Samples Test** | | | | | | | | | |
| Price ($'000) | | | | | | | | | |
|  | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Equal variances assumed | .393 | .532 | -2.729 | 118 | .007 | -161.551 | 59.188 | -278.759 | -44.344 |
| Equal variances not assumed |  |  | -2.679 | 92.096 | .009 | -161.551 | 60.308 | -281.327 | -41.775 |

Levene’s test (Levene statistic=0.393, p=0.532>0.05) indicates that the variances of prices in the 2 groups of Heating are homogeneous (no significant difference in variance). Hence, it is established that there are equal variances. As per the Independent Samples T-Test Results (T= -2.729 and p=0.007<0.05), Since the p-value is less than 0.05 then we can conclude that there is a linear relation between price and Heating as there is a significant impact of the Heating values on the price.

**Aircon**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Independent Samples Test** | | | | | | | | | |
|  | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Equal variances assumed | 2.365 | .127 | -1.340 | 118 | .183 | -82.202 | 61.351 | -203.694 | 39.290 |
| Equal variances not assumed |  |  | -1.430 | 107.742 | .156 | -82.202 | 57.474 | -196.129 | 31.725 |

Levene’s test (Levene statistic=2.365, p=0.127>0.05) indicates that the variances of prices in the 2 groups of Style are homogeneous (no significant difference in variance). Hence, it is established that there are equal variances. As per the Independent Samples T-Test Results (T= -1.340 and p=0.183>0.05), Since the p-value is greater than 0.05 then we can conclude that there is no linear relation between price and Aircon as there is no significant impact of the Aircon values on the price.

### Final Acceptable Variables Results as per the Assumptions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Normality Test** | **Outliers Test** | **Linear Relationship Test** | **Multicollinearity Test** |
| **Price ($'000)** | **DV** | **DV** | **DV** | **DV** |
| *Rooms* | Passed | Passed | Passed | Failed (with Bedrooms) |
| *Lot Size (sq. m)* | Passed | Passed | Passed | Passed |
| *Age* | Passed | Passed | Passed | Passed |
| *Area (sq. m)* | Passed | Passed | Passed | Passed |
| *To Train(km)* | Passed | Passed | Failed | Passed |
| *To Bus(km)* | Passed | Passed | Failed | Passed |
| *To Shops(km)* | Passed | Passed | Failed | Passed |
| *Street* | Passed | Passed | Passed | Passed |
| *Storeys* | Passed | Passed | Passed | Passed |
| *Style* | NA | Passed | Passed | Passed |
| *Bedrooms* | Passed | Passed | Passed | Failed (with Rooms) |
| *Bathrooms* | Passed | Passed | Passed | Passed |
| *Kitchen* | NA | Passed | Failed | Passed |
| *Heating* | NA | Passed | Passed | Passed |
| *Aircon* | NA | Passed | Failed | Passed |
| *Bay Views* | Passed | Passed | Passed | Passed |
| *Suburb* | NA | Passed | Passed | Passed |
| *Weekly Rent $* | Passed | Passed | Passed | Passed |
| *Rental Return %* | Passed | Passed | Passed | Passed |
| *Condition* | NA | Passed | Passed | Passed |
| *Rental Status* | NA | Passed | Failed | Passed |

Table 5 Acceptable Variables Results as per the Assumptions

### Transformed Variables

Since there were 2 variables with high skewness, i.e., Bay Views and Rental Return %, The 2 variables were transformed and the following are the Statistics of the Transformed Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | |
|  | Bay Views | Sqrt\_BayViews | Rental Return % | Sqrt\_RentReturn |
| Mean | .19 | .2489 | 3.715 | 1.9045 |
| Skewness | 1.457 | .936 | 1.024 | .422 |
| Kurtosis | .876 | -.805 | 2.367 | .767 |

Price= 254.68 + 0.081\*(LotSize)+ 0.331\*(Area)+ 0.2