**DATA ANALYSIS**

**PROJECT REPORT**

**Organizing Quantitative Data:**

**Dependent Variable:**

**Monthly income:** The amount of money an employee earns in a month. Here in the dataset monthly income is used as dependent variable.

**Independent Variables:**

1. **Job satisfaction:** This variable describes whether an employee is satisfied with his job/monthly income. Scale for job satisfaction is: 1 - Very Dissatisfied 2 – Dissatisfied 3 – Satisfied 4 - Very Satisfied

2. **Total working years:** This variable defines the total working years of an employee.

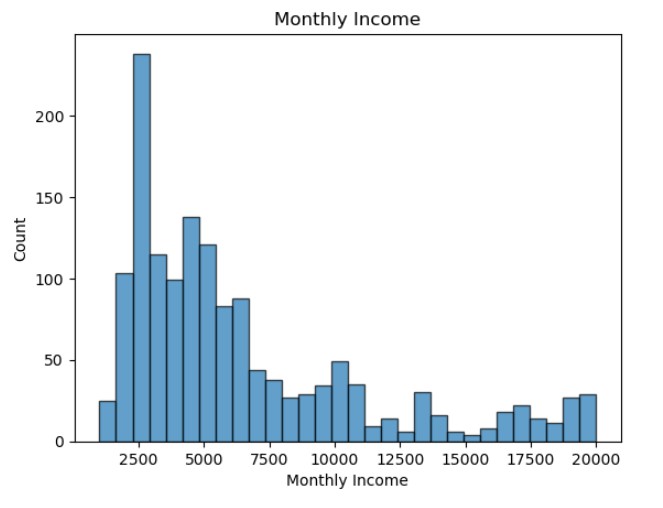
3. **Education:** This variable defines the education of the employee. Scale for Education is: 1 – School 2 – College 3 – Bachelors 4 – Masters 5 – Ph.D.

4. **Job level:** Job Level are grades or ranks of employees with different salaries and titles. Here the scale of Job Level is: 1 – Junior 2 - Mid level 3 – Senior 4 – Manager 5 - Director

5. **Years at company:** This variable defines the no of years an employee had worked in the company.

6. **Age:** This variable describes the age of employees from 18-60 years in the dataset.

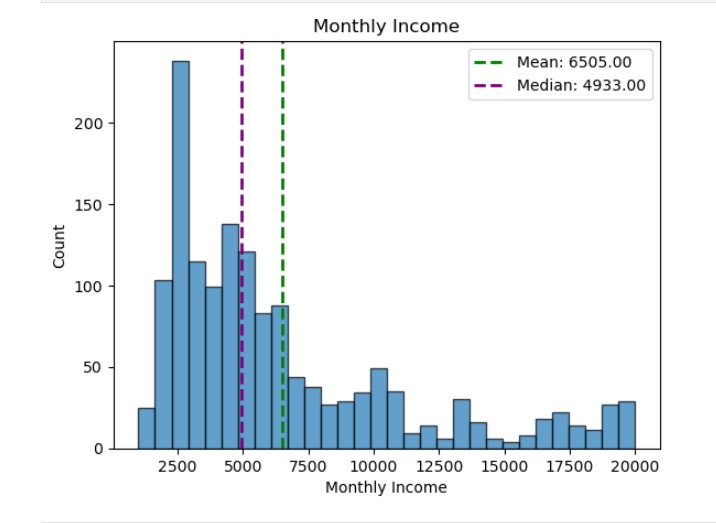
**Dependent Variable:** Monthly income



The shape of the graph is right skewed which tells us that mostly employees have lower income with a few having high incomes which is pulling the distribution to the right.

**Numerically Summarizing Data:**

|  |  |
| --- | --- |
| **Max** | 19999 |
| **Min** | 1009 |
| **Mean** | 6505 |
| **Median** | 4933 |
| **Mode** | 2342 |
| **Standard Deviation** | 4700.26 |
| **Variance** | 22092457.22 |
| **Range** | 18990 |
| **Skew** | 1.36 |

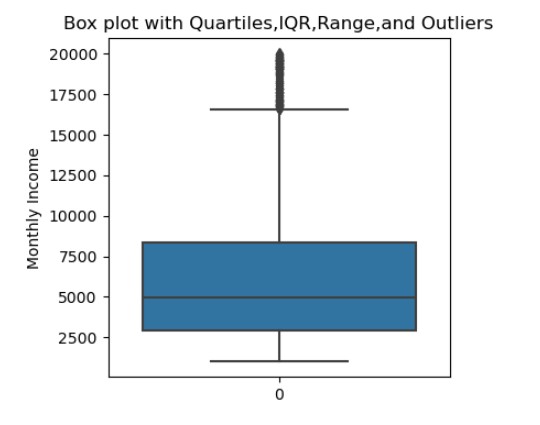


**Interpretation for Mean, Median, Mode, standard deviation, variance:**

The mean monthly income of an employee in the dataset is **6505,** while the mode income is **2342,** which is the most frequently occurring income in the dataset. Half of the employees in the dataset have income below 4933 and half have above 4933.The graph shows that the **mean is greater than the median**, indicating that the distribution is **right-skewed.** This suggests that most employees have **lower incomes**, while a few employees with significantly higher incomes are pulling the mean to the right. Here the standard deviation is 4700 which means that the data is 4700 units away from the mean. The higher the standard deviation and variance the wider the data is spread away from the mean.

**Quartiles and IQR:**

|  |  |
| --- | --- |
| **Quartiles** |  |
| **Q1** | 2922.5 |
| **Q2** | 4933 |
| **Q3** | 8383.75 |
|  |  |
| **Interquartile** |  |
| **IQR(Q3 - Q1)** | 5461.25 |



**Interpretation for Quartiles, IQR and Outliers:**

Q1 is the bottom of the box, middle line is Q2 also called the median and third is Q3. The IQR is the distance between Q3 and Q1 and the box in the box plot is represented as IQR. The line below the box is the minimum range and the line above the box is the maximum range. The data points outside the whiskers are called the outliers which are high values due to the high incomes of few employees.

**Interpretation for z-scores:**

Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is -1.08 standard deviations away and negative z score tells us that the monthly income is below the average monthly income.

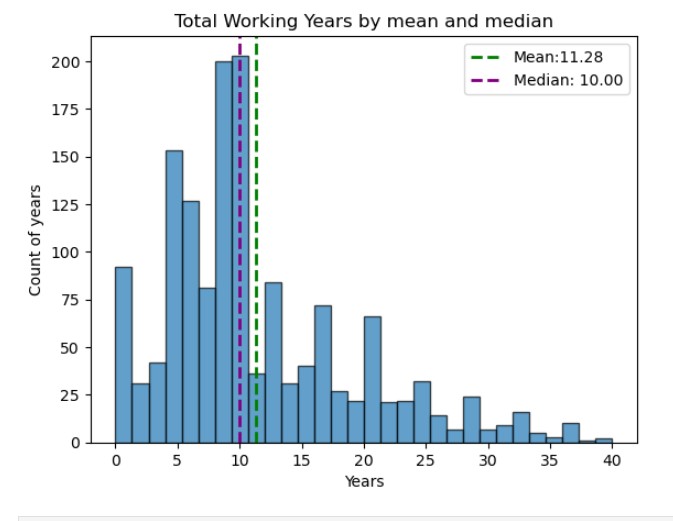
First 20 z-scores:

|  |  |
| --- | --- |
| **Monthly Income** | **Z-scores** |
| 1420 | -1.082217381 |
| 1200 | -1.129039109 |
| 1878 | -0.984743056 |
| 1051 | -1.160750189 |
| 1904 | -0.979209579 |
| 1611 | -1.041567608 |
| 1569 | -1.050506302 |
| 1514 | -1.062211734 |
| 1675 | -1.027946742 |
| 1483 | -1.068809341 |
| 2325 | -0.889609817 |
| 1102 | -1.149896061 |
| 2564 | -0.838744394 |
| 2121 | -0.933026329 |
| 2552 | -0.841298307 |
| 1859 | -0.988786751 |
| 2994 | -0.747229198 |
| 2926 | -0.761701369 |
| 2836 | -0.780855712 |
| 1009 | -1.169688883 |

**Independent Variables:**

**1.Total Working Years and Monthly Income:**

|  |  |
| --- | --- |
| **Max** | 40 |
| **Min** | 0 |
| **Mean** | 11 |
| **Median** | 10 |
| **Mode** | 10 |
| **Standard Deviation** | 7.77 |
| **Variance** | 60.38 |
| **Range** | 40 |
| **skewness** | 1.11 |



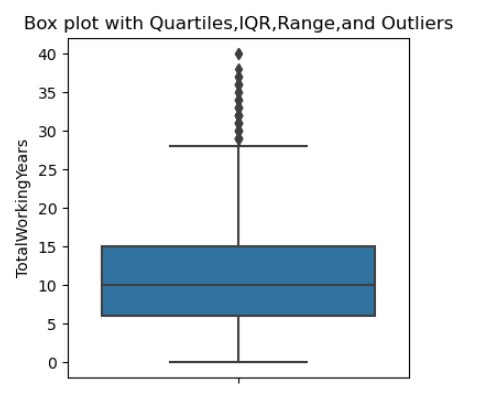
**Interpretation for Mean, Median, Mode, standard deviation, variance:**

The average number of years an employee has been with the company is **11 years,** while the most occurring number of years is **10 years**. Since the graph is **right-skewed**, it indicates that a small number of employees have significantly higher years of experience, pulling the mean towards the right. A right-skewed graph suggests that the majority of employees have **lower years of experience**, while only a few have **much higher years of experience.**

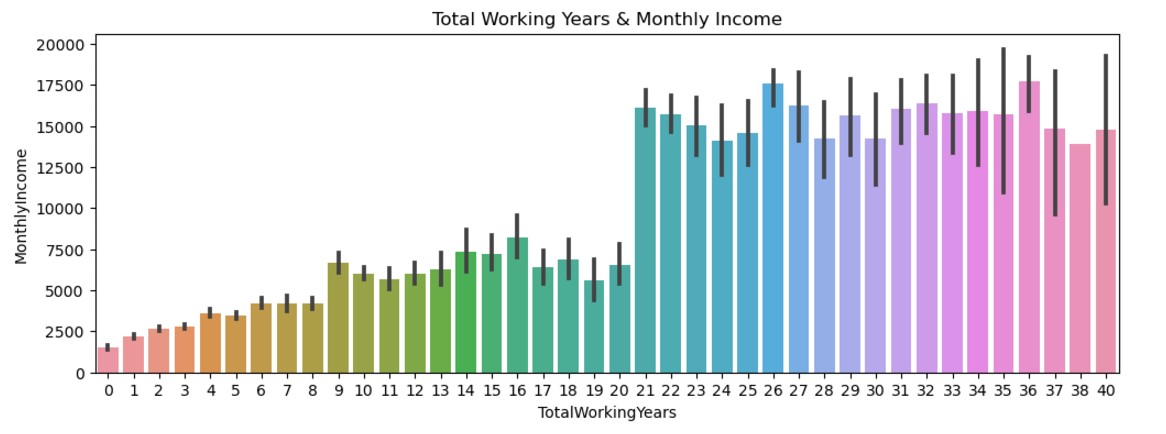
The standard deviation of the number of years an employee works is **7.8 years.** This means that the number of year’s employees work tends to **vary**, on average, by **7.8 years** from the mean of **11 years**. The variance is **60.38,** A higher variance suggests that data points are widely spread out around the mean.

**Quartiles and IQR:**

|  |  |
| --- | --- |
| **Quartiles** |  |
| **Q1** | 6 |
| **Q2** | 10 |
| **Q3** | 15 |
|  |  |
| **Interquartile** |  |
| **IQR(Q3 - Q1)** | 9 |



The first line this is the minimum line, then the bottom of the box is the Q1 the minimum range and the line in the middle of the box is Q2 also known as the median the top of the box is Q3 and the whole box is the IQR. The most top line is the maximum line. And the points outside the whiskers are outliers as can be shown in the graph.



This graph shows as the number of years increases the employee’s monthly income also increases due to his increase in experience.

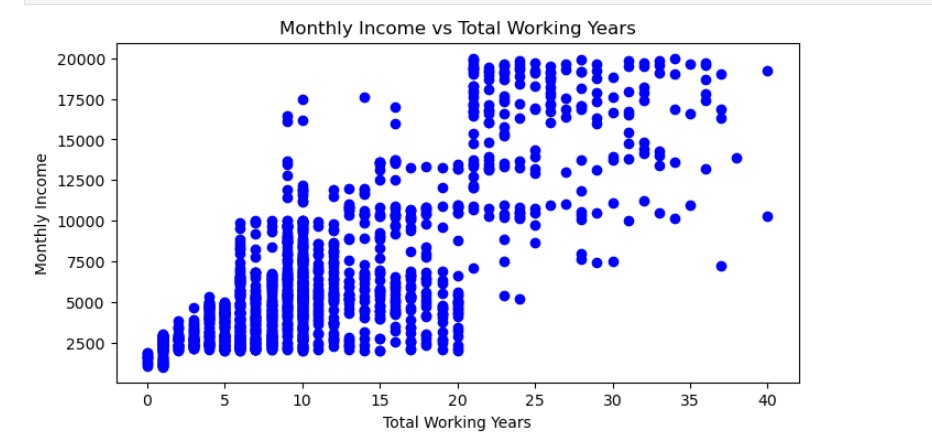
**Interpretation for z-scores:**

Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is -1.45 standard deviations away and negative z score tells us that the monthly income is below the average monthly income.

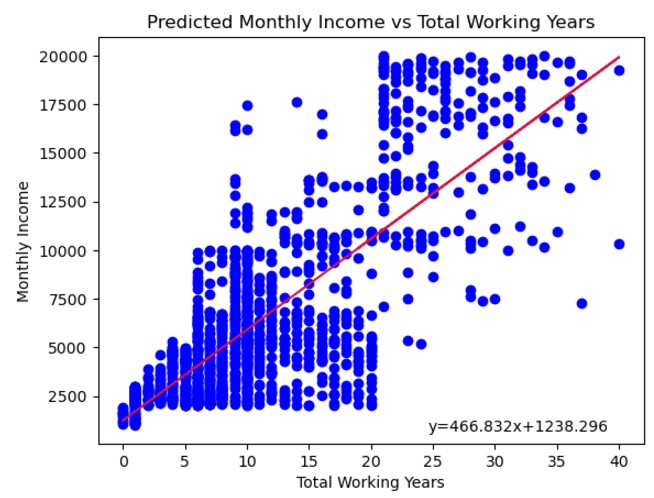
|  |  |
| --- | --- |
| **Monthly Income** | **Z-scores** |
| 25233 | -1.4522918 |
| 9724 | -1.4522918 |
| 8059 | -1.4522918 |
| 13493 | -1.4522918 |
| 13556 | -1.4522918 |
| 19305 | -1.4522918 |
| 18420 | -1.4522918 |
| 8018 | -1.4522918 |
| 26820 | -1.4522918 |
| 16102 | -1.3235626 |
| 20989 | -1.3235626 |
| 9241 | -1.3235626 |
| 18437 | -1.3235626 |
| 9947 | -1.3235626 |
| 7172 | -1.3235626 |
| 6148 | -1.3235626 |
| 21221 | -1.3235626 |
| 19783 | -1.3235626 |
| 11757 | -1.3235626 |
| 26999 | -1.3235626 |

**Describing the relation between two variables:**

|  |  |
| --- | --- |
| **Correlation coefficient of Monthly Income and Total Working Years:** | 0.77180685 |
| **Regression Line** | y = 1238.30 + 466.83x |
|  |  |
| **Coefficient of Determination** | 0.5956 |



The correlation coefficient of monthly income and total working years is 0.77180685 which shows a strong positive linear relationship between monthly income and total working years which means that as people's experience increases their monthly income also increases.



Predictions for just first 20 monthly incomes:

|  |  |
| --- | --- |
| **X** | **Predicted Monthly Income by Total working years** |
| 1420 | 1238.295905 |
| 1200 | 1238.295905 |
| 1878 | 1238.295905 |
| 1051 | 1238.295905 |
| 1904 | 1238.295905 |
| 1611 | 1238.295905 |
| 1569 | 1238.295905 |
| 1514 | 1238.295905 |
| 1675 | 1238.295905 |
| 1483 | 1705.128334 |
| 2325 | 1705.128334 |
| 1102 | 1705.128334 |
| 2564 | 1705.128334 |
| 2121 | 1705.128334 |
| 2552 | 1705.128334 |
| 1859 | 1705.128334 |
| 2994 | 1705.128334 |
| 2926 | 1705.128334 |
| 2836 | 1705.128334 |
| 1009 | 1705.128334 |

**Interpretation for slope and y-intercept:**

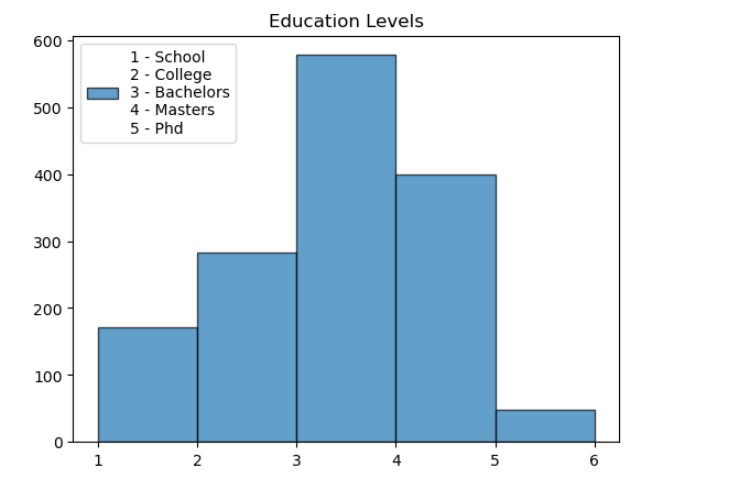
1238.30 is the intercept this is the income of the employee having zero experience. 466.83 is the slope which is positive which means that as the work experience of employee’s increases their monthly income increases by 466.83.

**Interpretation for Coefficient of Determination:**

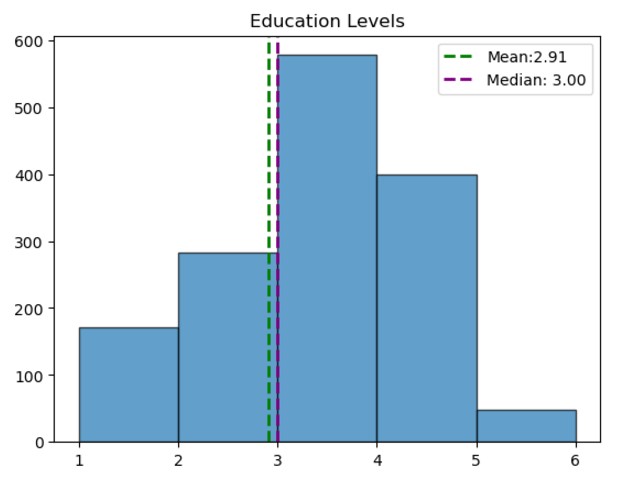
The Coefficient of Determination is 59.56% which tells us that 59.56% of variance in monthly income can be explained by total working years this tells us that total working years has a moderate relationship with monthly income as there are other factors also involved.

**2.Education and Monthly Income:**

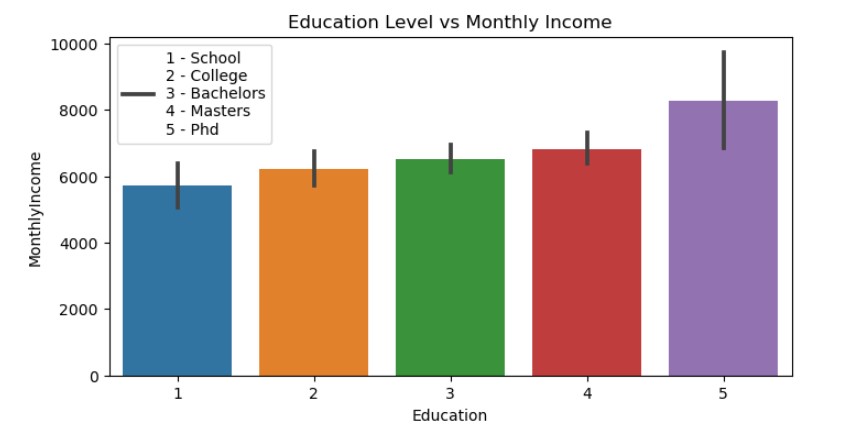
|  |  |
| --- | --- |
| **Max** | 5 |
| **Min** | 1 |
| **Mean** | 3 |
| **Median** | 3 |
| **Mode** | 3 |
| **Standard Deviation** | 1.023 |
| **Variance** | 1.04 |
| **Range** | 4 |
| **skewness** | -0.29 |



Average education for employees in the dataset is 3 which tells us that the average no of employees in the dataset are graduated with bachelor’s degree. While the most frequently repeated degree is 3 (bachelor’s). The graph is left skewed which shows that the it has higher values on the other side. The education level based on standard deviation of the employees is 1.023 which is close to the average 3 which means that data is not spread as far and is close to mean education level. Additionally, the variance is also close to the mean education level indicating that the data is not spread too much and is clustered near the mean.



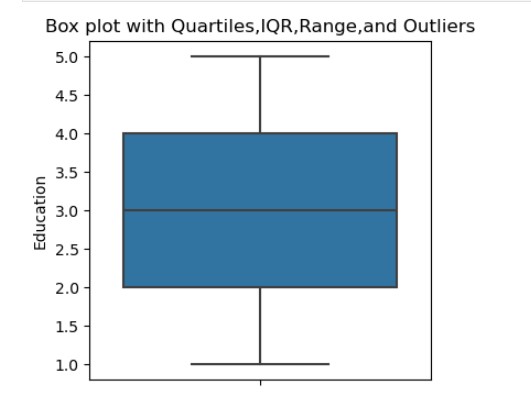
As median is greater than the mean supporting that the graph is skewed left. And the skewness of the graph is also negative which also indicates towards the graph to be left skewed.



The graph shows us the monthly income of the employees based on their education levels. This graph shows us that the employees having a PhD has higher incomes as compared to other employees with other degrees.

**Quartiles and IQR:**

|  |  |
| --- | --- |
| **Quartiles** |  |
| Q1 | 2 |
| Q2 | 3 |
| Q3 | 4 |
|  |  |
| **Interquartile** |  |
| IQR(Q3 - Q1) | 2 |



No outliers’ in the education column as there are no points outside the whiskers as shown in the graph. The center line of the box showing the median or Q2.

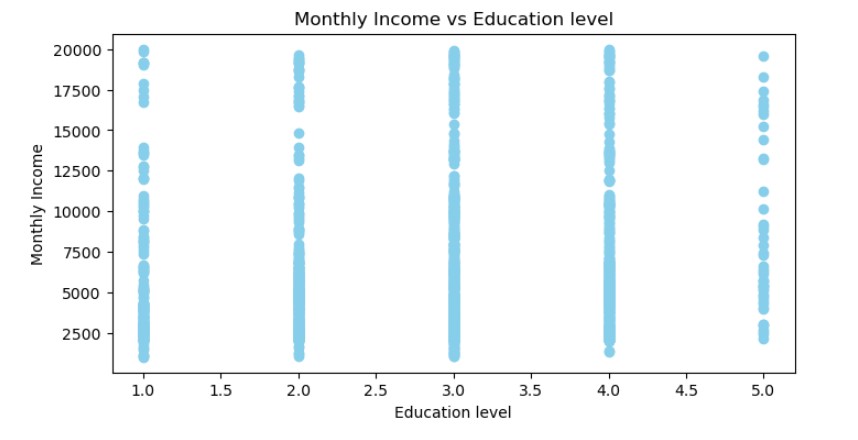
**Interpretation for z-scores:**

Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is 0.08 standard deviations away and positive z score tells us that the monthly income is above the average monthly income. And for the income of 1051 it has a negative z score which tells us that the income is below average.

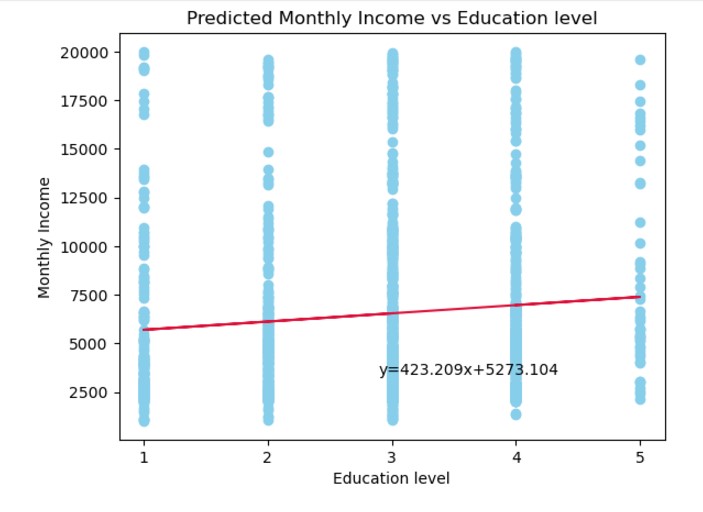
|  |  |
| --- | --- |
| **Monthly Income** | **z-scores** |
| 1420 | 0.087145591 |
| 1200 | 0.087145591 |
| 1878 | 0.087145591 |
| 1051 | -0.889941342 |
| 1904 | -1.867028276 |
| 1611 | 0.087145591 |
| 1569 | -0.889941342 |
| 1514 | 0.087145591 |
| 1675 | -1.867028276 |
| 1483 | -1.867028276 |
| 2325 | -1.867028276 |
| 1102 | 0.087145591 |
| 2564 | -0.889941342 |
| 2121 | 0.087145591 |
| 2552 | -0.889941342 |
| 1859 | 0.087145591 |
| 2994 | 0.087145591 |
| 2926 | 0.087145591 |
| 2836 | 0.087145591 |
| 1009 | -1.867028276 |

**Describing the relation between two variables:**

|  |  |
| --- | --- |
| **Correlation coefficient of monthly income and education level** | 0.09218208 |
| **Regression Line** | y = 5273.10 + 423.21x |
|  |  |
| **Coefficient of determination** | 0.00849754 |



Monthly income and education level has a week linear relationship between each other. They both are connected with each other but still their connection isn't that strong.



Predictions for just first 20 monthly incomes:

|  |  |
| --- | --- |
| **X** | **Predicted Monthly Income** |
| 1420 | 6542.731484 |
| 1200 | 6542.731484 |
| 1878 | 6542.731484 |
| 1051 | 6119.522421 |
| 1904 | 5696.313359 |
| 1611 | 6542.731484 |
| 1569 | 6119.522421 |
| 1514 | 6542.731484 |
| 1675 | 5696.313359 |
| 1483 | 5696.313359 |
| 2325 | 5696.313359 |
| 1102 | 6542.731484 |
| 2564 | 6119.522421 |
| 2121 | 6542.731484 |
| 2552 | 6119.522421 |
| 1859 | 6542.731484 |
| 2994 | 6542.731484 |
| 2926 | 6542.731484 |
| 2836 | 6542.731484 |
| 1009 | 5696.313359 |

**Interpretation for slope and y-intercept:**

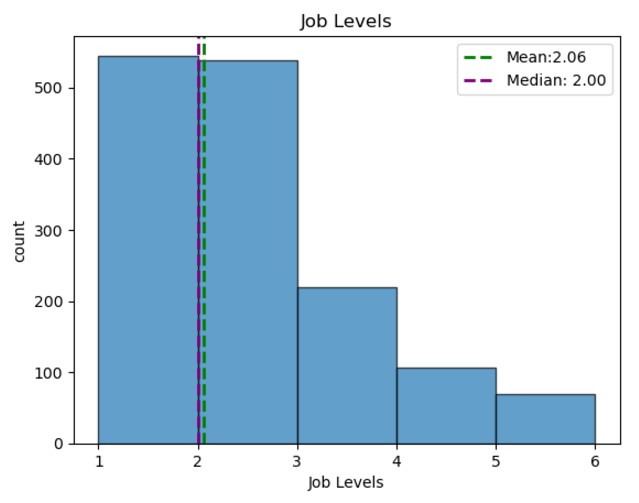
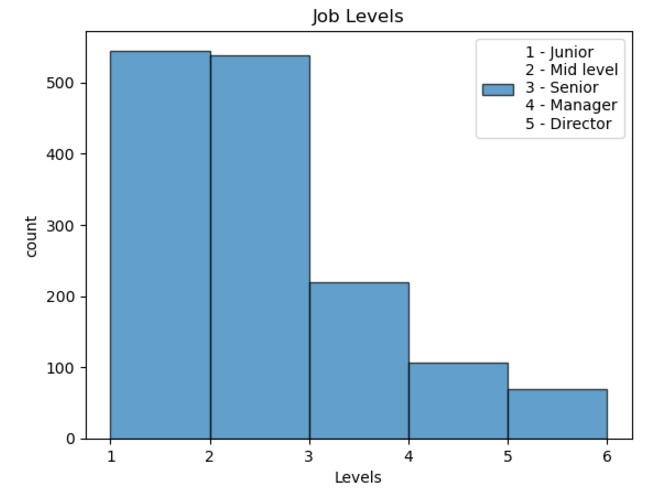
5273.10 is the intercept this is the income of the employee having zero education level.423.21 is he slope which is positive which means that as the education level of the employee increases the income increases by 423.21.

**Interpretation for Coefficient of Determination:**

The Coefficient of Determination is 8.5% which tells us that only 8.5% of the variance in the monthly income can be explained by education it is due as there are other factors also involved.

**3.Job Level and Monthly income:**

|  |  |
| --- | --- |
| **Max** | 5 |
| **Min** | 1 |
| **Mean** | 2.06 |
| **Median** | 2 |
| **Mode** | 1 |
| **Standard Deviation** | 1.105 |
| **Variance** | 1.222 |
| **Range** | 4 |
| **skew** | 1.02 |



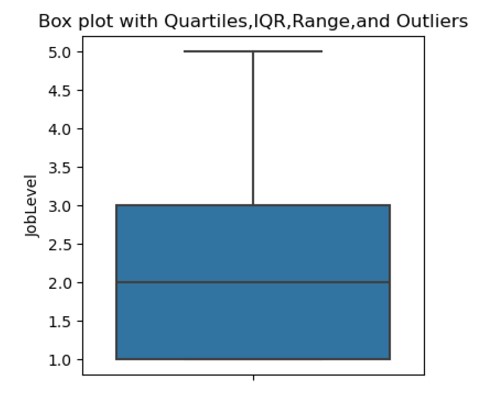
The average level of employee is in the dataset is mid-level employees. And the most frequently occurring position is junior employees. The graph is right skewed as mean is greater than the median and also it is supported by the skewness which is positive indicating the graph to be right skewed. As larger portion employees job levels are lower and a fewer of them have higher positions the standard deviation and variance also shows this that the most of the data is close to the mean.



This graph indicates that employees with a higher posts or higher job levels has higher monthly income as compared to employees with job levels as lower job levels. The most common job level(junior) has a very low monthly income.

**Quartiles and IQR:**

|  |  |
| --- | --- |
| **Quartiles** |  |
| **Q1** | 1 |
| **Q2** | 2 |
| **Q3** | 3 |
|  |  |
| **Interquartile** |  |
| **IQR(Q3 - Q1)** | 2 |



The box plot helps us to show the minimum and maximum called the range. In this box plot there is no bottom whiskers to show minimum value as Q1 is close to minimum. This graph also shows that there are no outliers in this graph as there are no points outside the whiskers.

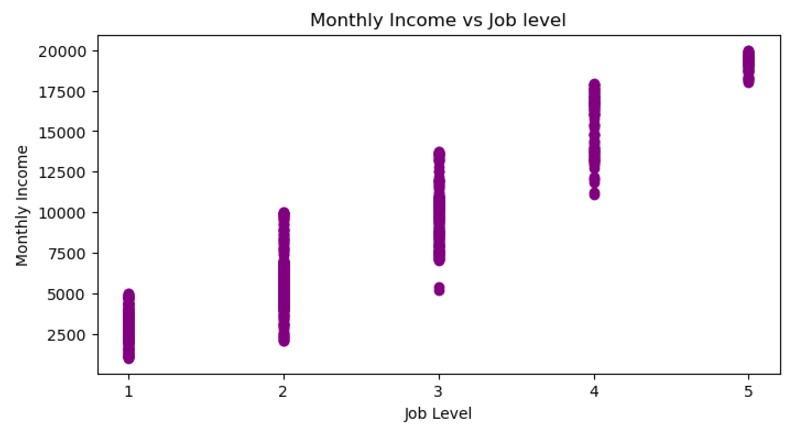
**Interpretation for z-scores:**

Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is -0.9635 standard deviations away and negative z score tells us that the monthly income is below the average monthly income. And for the income of 1051 it has a negative z score of -0.9635 which tells us that the income is below average.

|  |  |
| --- | --- |
| **Monthly Income** | **z-scores** |
| 1420 | -0.9635038 |
| 1200 | -0.9635038 |
| 1878 | -0.9635038 |
| 1051 | -0.9635038 |
| 1904 | -0.9635038 |
| 1611 | -0.9635038 |
| 1569 | -0.9635038 |
| 1514 | -0.9635038 |
| 1675 | -0.9635038 |
| 1483 | -0.9635038 |
| 2325 | -0.9635038 |
| 1102 | -0.9635038 |
| 2564 | -0.9635038 |
| 2121 | -0.9635038 |
| 2552 | -0.9635038 |
| 1859 | -0.9635038 |
| 2994 | -0.9635038 |
| 2926 | -0.9635038 |
| 2836 | -0.9635038 |
| 1009 | -0.9635038 |

**Describing the relation between two variables:**

|  |  |
| --- | --- |
| **Correlation coefficient of monthly income and education level** | 0.94983461 |
| **Regression Line** | y = 4038.15x + (-1833.24) |
|  |  |
| **Coefficient of determination** | 0.90218579 |



The value of coefficient of correlation 0.92 indicating a strong positive correlation between monthly income and job level also can be seen in the graph.



|  |  |
| --- | --- |
| **X** | **Predicted Monthly Income** |
| 1420 | 2204.90474 |
| 1200 | 2204.90474 |
| 1878 | 2204.90474 |
| 1051 | 2204.90474 |
| 1904 | 2204.90474 |
| 1611 | 2204.90474 |
| 1569 | 2204.90474 |
| 1514 | 2204.90474 |
| 1675 | 2204.90474 |
| 1483 | 2204.90474 |
| 2325 | 2204.90474 |
| 1102 | 2204.90474 |
| 2564 | 2204.90474 |
| 2121 | 2204.90474 |
| 2552 | 2204.90474 |
| 1859 | 2204.90474 |
| 2994 | 2204.90474 |
| 2926 | 2204.90474 |
| 2836 | 2204.90474 |
| 1009 | 2204.90474 |

First 20 predicted values using regression line:

**Interpretation for slope and y-intercept:**

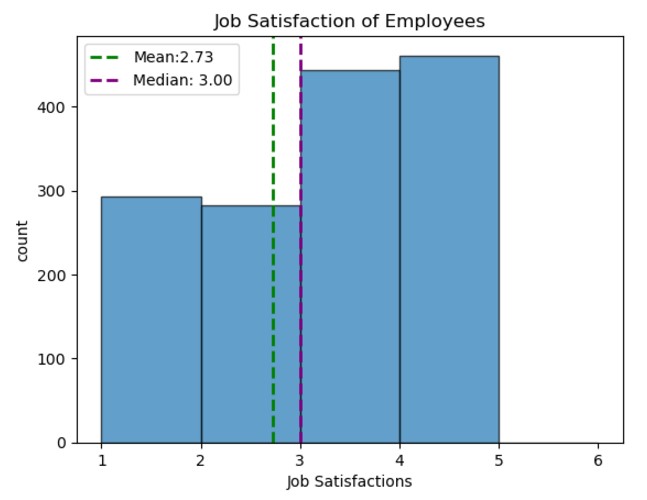
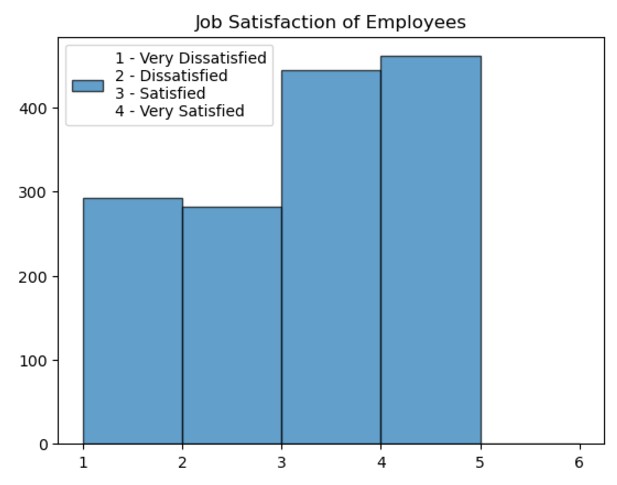
The intercept for job levels and monthly is negative 1833.24. 4038.15 is the slope which is positive which means that as the job level of the employee increases the income increases by 4038.15.

**Interpretation for Coefficient of Determination:**

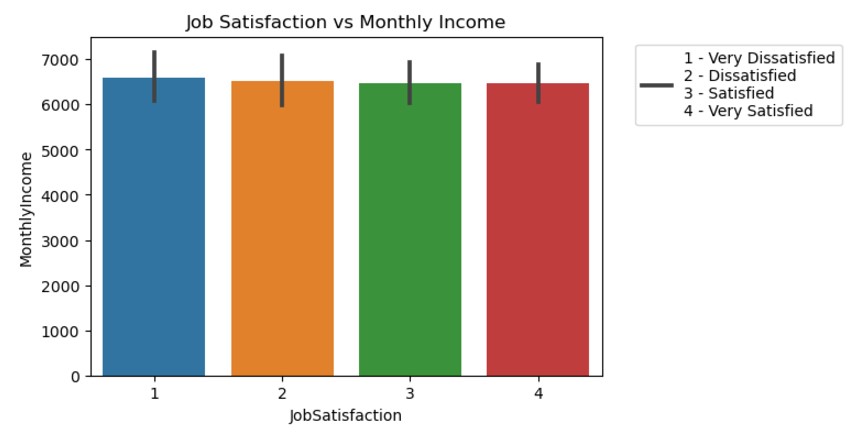
The Coefficient of Determination is 90.2% which tells us that only 90.2% of the variance in the monthly income can be explained by job levels. As it is easy to explain that higher job levels employees will have higher incomes and lower will have lower incomes.

**4. Job Satisfaction and Monthly income:**

|  |  |
| --- | --- |
| **Max** | 4 |
| **Min** | 1 |
| **Mean** | 2.73 |
| **Median** | 3 |
| **Mode** | 4 |
| **Standard Deviation** | 1.104 |
| **Variance** | 1.219 |
| **Range** | 3 |
| **skewness** | -0.325 |

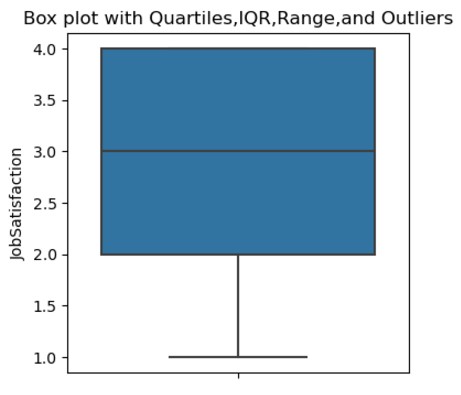


The average employees in the dataset are satisfied. And the most of the employees in the dataset are very much satisfied with their jobs not sure about income. This graph also shows us that median is greater than mean which shows that the graph is left skewed as there are less employees who are satisfied. The left skewness of the graph can also be seen by the value of skewness which is negative supporting the graph to be left skewed. As the standard deviations and variance is not too high it means that they are very close to mean and the most of the employee might seem to have same levels of satisfaction.



As seen in the charts that all the bars are close to each other based on monthly incomes. As job satisfaction increases, so does the average monthly income.

|  |  |
| --- | --- |
| **Quartiles** |  |
| **Q1** | 2 |
| **Q2** | 3 |
| **Q3** | 4 |
|  |  |
| **Interquartile** |  |
| **IQR(Q3 - Q1)** | 2 |



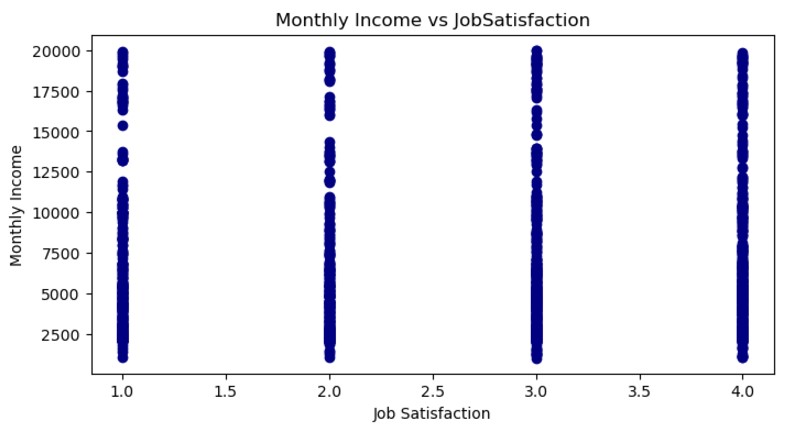
No outliers in the Job Satisfaction column as there are no points outside the whiskers as shown in the box plot. In this box plot there are no upper whiskers as Q3 is close to maximum value. The box represents the IQR with Q1, Q2(median) and Q3.

**Interpretation for z-scores:**

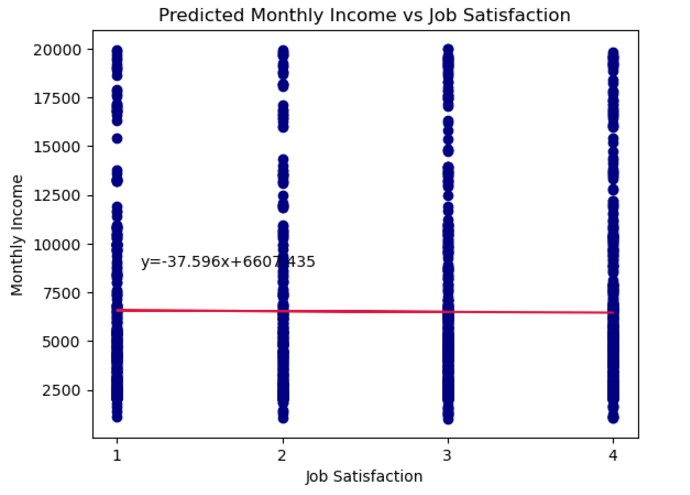
Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is 0.249 standard deviations away and positive z score tells us that the monthly income is above the average monthly income. And for the income of 1878 it has a negative z score of -0.6586 which tells us that the income is below average income.

|  |  |
| --- | --- |
| **Monthly Income** | **z-scores** |
| 1420 | 0.2491476 |
| 1200 | 0.2491476 |
| 1878 | -0.656844 |
| 1051 | 1.1551387 |
| 1904 | 0.2491476 |
| 1611 | 1.1551387 |
| 1569 | 1.1551387 |
| 1514 | 0.2491476 |
| 1675 | 0.2491476 |
| 1483 | -0.656844 |
| 2325 | -1.562835 |
| 1102 | 1.1551387 |
| 2564 | 1.1551387 |
| 2121 | -0.656844 |
| 2552 | -1.562835 |
| 1859 | -0.656844 |
| 2994 | 1.1551387 |
| 2926 | 1.1551387 |
| 2836 | -0.656844 |
| 1009 | 0.2491476 |

|  |  |
| --- | --- |
| **Correlation coefficient of monthly income and job Satisfaction** | -0.00883171 |
| **Regression Line** | y = 6607.44 + (-37.60x) |
|  |  |
| **coefficient of determination** | 7.80E-05 |



Monthly income and Job satisfaction has a week linear relationship between each other. They both are connected with each other but still their connection isn't that strong. There is no pattern or trend between monthly income and job satisfaction.



|  |  |
| --- | --- |
| **X** | **Predicted monthly income** |
| 1420 | 6494.646855 |
| 1200 | 6494.646855 |
| 1878 | 6532.243058 |
| 1051 | 6457.050651 |
| 1904 | 6494.646855 |
| 1611 | 6457.050651 |
| 1569 | 6457.050651 |
| 1514 | 6494.646855 |
| 1675 | 6494.646855 |
| 1483 | 6532.243058 |
| 2325 | 6569.839262 |
| 1102 | 6457.050651 |
| 2564 | 6457.050651 |
| 2121 | 6532.243058 |
| 2552 | 6569.839262 |
| 1859 | 6532.243058 |
| 2994 | 6457.050651 |
| 2926 | 6457.050651 |
| 2836 | 6532.243058 |
| 1009 | 6494.646855 |

**Interpretation for slope and y-intercept:**

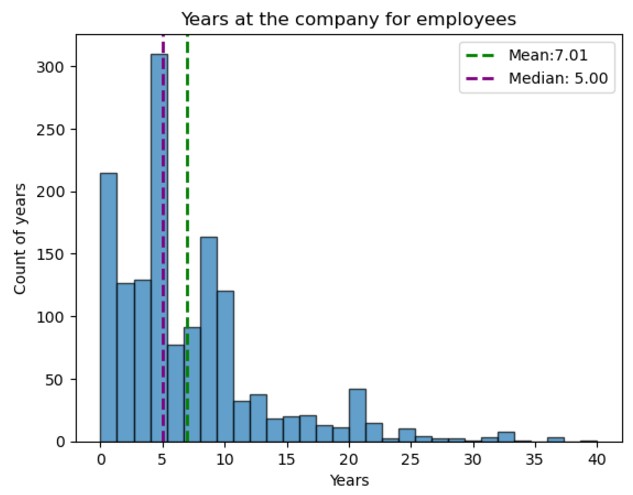
The intercept for job satisfaction and monthly income is 6607.435 it represents the predicted monthly income when job satisfaction is zero but it is not possible for job satisfaction to be zero. We have a negative slope of 37.596 and it shows that the relationship between monthly income and job satisfaction is very week as there are other factors also involved.

**Interpretation for Coefficient of Determination:**

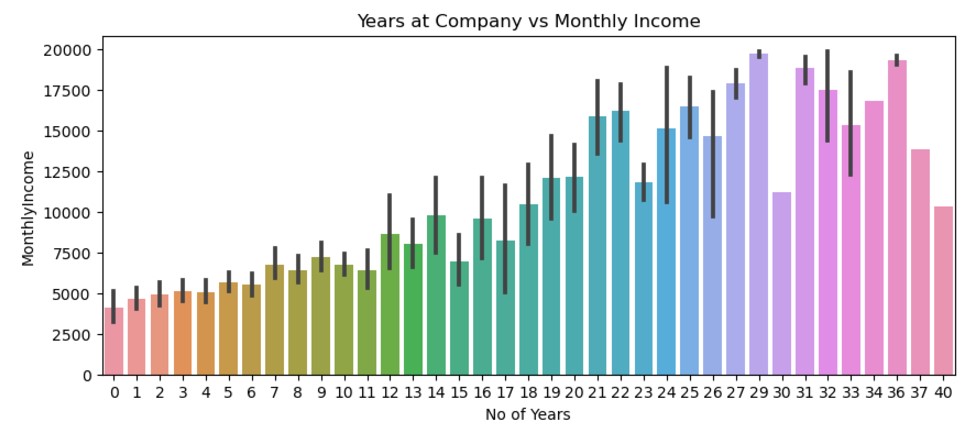
The coefficient of determination is 7.80E-05 which is a very small value. This indicates that only a small fraction of the variability in monthly income can be explained by job satisfaction.

**5. Years at Company and Monthly income:**

|  |  |
| --- | --- |
| **Max** | 40 |
| **Min** | 0 |
| **Mean** | 7 |
| **Median** | 5 |
| **Mode** | 5 |
| **Standard Deviation** | 6 |
| **Variance** | 37.43 |
| **Range** | 40 |
| **skewness** | 1.76 |

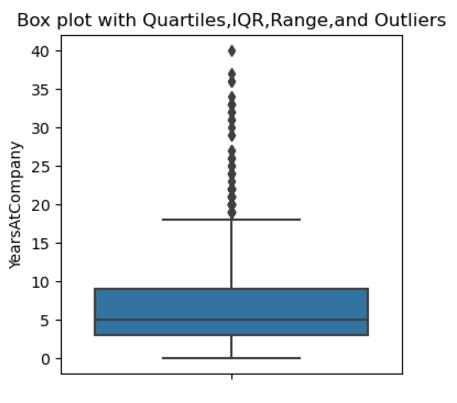


The average number of years at company is 7 years at the company for an employee. Half of the employees have been at the company for **5 years or less**, while the other half have been there for more than 5 years. And the most frequently occurring no of years is 5. In the graph it can be seen that the mean is greater than the median pulling the graph to the right. The skewness of the graph is also a positive indicating the graph to be right skewed. A value of 6 years suggests that the tenure of most employees deviates by about 6 years from the average of 7.01 years. This shows a moderate level of variability, indicating some employees stay much longer or shorter than the average tenure. A higher variance shows a wide spread in the year’s employees stay at the company.



This graph shows the no of years at the company and monthly income. It says that as the no of years of an employee working in the company increases it monthly income also increases.

|  |  |
| --- | --- |
| **Quartiles** |  |
| **Q1** | 3 |
| **Q2** | 5 |
| **Q3** | 9 |
|  |  |
| **Interquartile** |  |
| **IQR(Q3 - Q1)** | 6 |



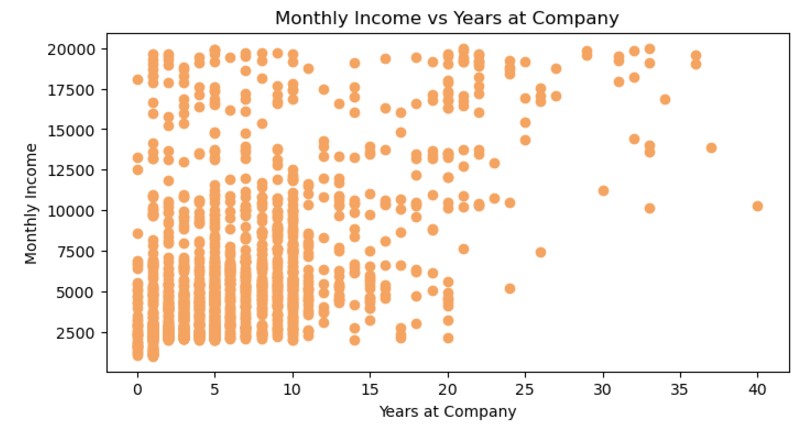
The box representing the IQR with Q1 at the bottom, Q2 at the middle also the median and the top of the box as Q3. Bottom of the whiskers is the minimum value and top of the whiskers called the maximum. Above the whiskers are called the outliers.

**Interpretation for z-scores:**

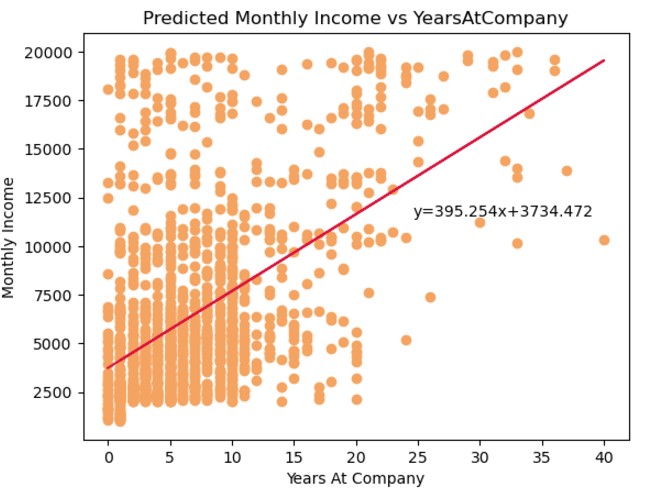
Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is -1.16 standard deviations away and negative z score tells us that the monthly income is below the average monthly income.

|  |  |
| --- | --- |
| **Monthly Income** | **z-scores** |
| 1420 | -1.14611 |
| 1200 | -1.14611 |
| 1878 | -1.14611 |
| 1051 | -1.14611 |
| 1904 | -1.14611 |
| 1611 | -1.14611 |
| 1569 | -1.14611 |
| 1514 | -1.14611 |
| 1675 | -1.14611 |
| 1483 | -0.9826 |
| 2325 | -1.14611 |
| 1102 | -0.9826 |
| 2564 | -0.9826 |
| 2121 | -0.9826 |
| 2552 | -0.9826 |
| 1859 | -0.9826 |
| 2994 | -0.9826 |
| 2926 | -0.9826 |
| 2836 | -0.9826 |
| 1009 | -0.9826 |

|  |  |
| --- | --- |
| **Correlation coefficient of monthly income and job Satisfaction** | 0.51446911 |
|  |  |
| **Regression Line** | y = 3734.47 + 395.25x |
|  |  |
| **Coefficient of Determination** | 0.26467847 |



The value of coefficient of correlation between years at company and monthly income has a strong positive correlation between each other. As the value of one variable increases the other also increases.



|  |  |
| --- | --- |
| **X** | **Predicted Monthly Income** |
| 1420 | 3734.47239 |
| 1200 | 3734.47239 |
| 1878 | 3734.47239 |
| 1051 | 3734.47239 |
| 1904 | 3734.47239 |
| 1611 | 3734.47239 |
| 1569 | 3734.47239 |
| 1514 | 3734.47239 |
| 1675 | 3734.47239 |
| 1483 | 4129.72589 |
| 2325 | 3734.47239 |
| 1102 | 4129.72589 |
| 2564 | 4129.72589 |
| 2121 | 4129.72589 |
| 2552 | 4129.72589 |
| 1859 | 4129.72589 |
| 2994 | 4129.72589 |
| 2926 | 4129.72589 |
| 2836 | 4129.72589 |
| 1009 | 4129.72589 |

**Interpretation for slope and y-intercept:**

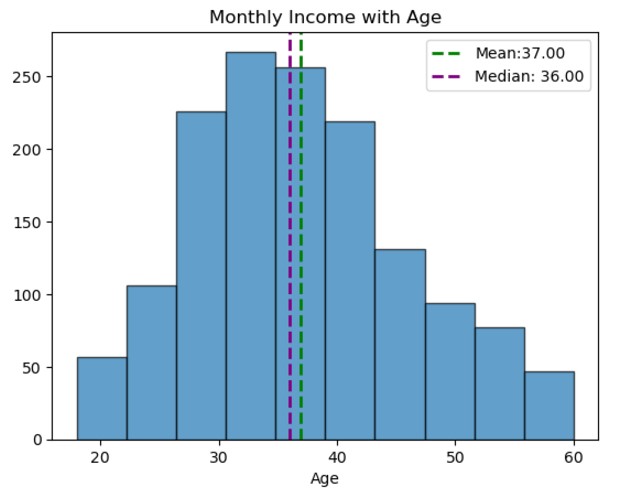
The intercept for Years at the company and monthly income is 3734.47 it represents the predicted monthly income when years at the company is zero but it is not possible years at the company to be zero and have this income. We have a positive slope of 395.25 and it shows that the relationship between monthly income and years at the company that as years increases the income increases.

**Interpretation for Coefficient of Determination:**

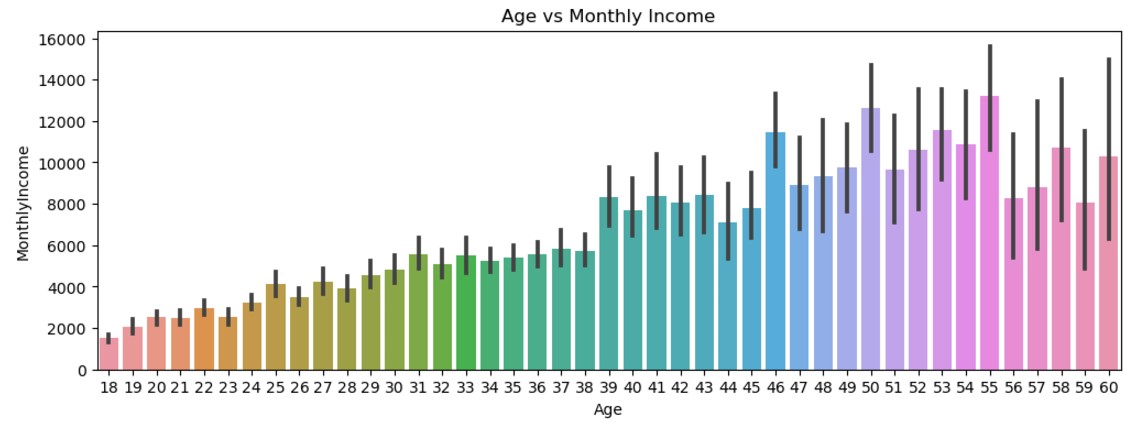
The Coefficient of Determination is 26.46% which tells us that only 26.46% of the variance in the monthly income can be explained by years at the company. As we can explain that as the number of the years increases employees will have higher incomes.

**6. Age and Monthly income:**

|  |  |
| --- | --- |
| **Max** | 60 |
| **Min** | 18 |
| **Mean** | 36.9 |
| **Median** | 36 |
| **Mode** | 34 |
| **Standard Deviation** | 9.12 |
| **Variance** | 83.33 |
| **Range** | 42 |
| **Skewness** | 0.414 |

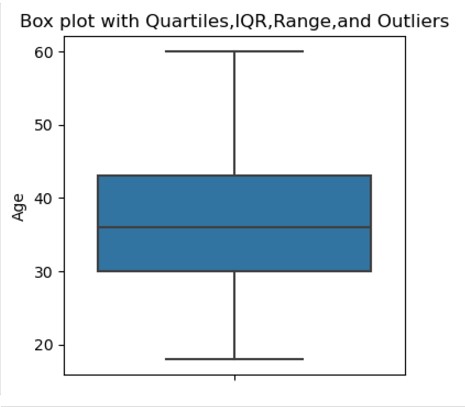


The mean age of employees in the dataset is 37. Half of the employees in the dataset are greater than age 36 **years and half are less than 36 years.** The most frequently occurring age of employee in the dataset is 34 years. By seeing the graph, we can see that the mean is greater than the median identified by green and purple lines which tells us that the graph is right skewed and also the skewness is also a positive value which also tells us that the graph is right skewed. The standard deviation of 9.12 suggests that the ages deviate on an average of 9.12 years and the variance of 83.33 years this value shows a moderate spread of data with some individuals being young and some being old.



This bar graph shows us that as the employees age increases their monthly income increases as their experience increases they are promoted to a new position so as age increases monthly income also increases.

|  |  |
| --- | --- |
| **Quartiles** |  |
| Q1 | 30 |
| Q2 | 36 |
| Q3 | 43 |
|  |  |
| **Interquartile** |  |
| IQR(Q3 - Q1) | 13 |



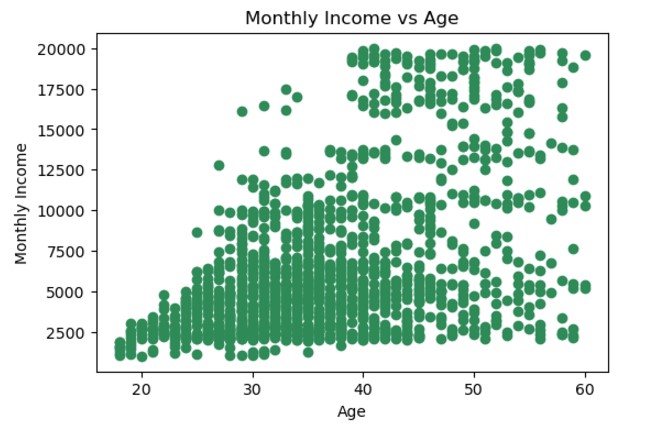
This box plot shows the Quartiles, IQR, Range and outliers. As we can see the bottom of the whiskers are the minimum age and the top of the whiskers are the maximum age. The box is the IQR with the bottom Q1 and middle line Q2 also called the median and top of the box is Q3. There are no outliers in this column of Ages.

**Interpretation for z-scores:**

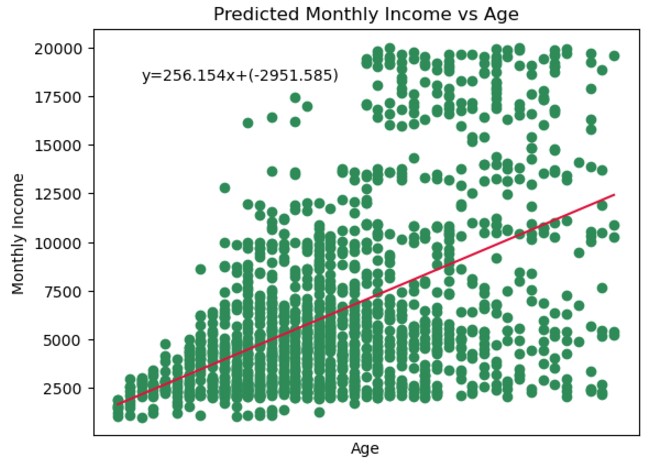
Z-scores help us to measure how many standard deviations the data point is away from the mean. Negative z score indicates that the data point is lower than average monthly income and a positive z score indicates that the data point is above the average monthly income. For example, if monthly income is 1420 then z score is -2.073 standard deviations away and negative z score tells us that the monthly income is below the average monthly income. And as we are going down the monthly income column z scores are changing from negative to positive.

|  |  |
| --- | --- |
| **Monthly Income** | **z-score** |
| 1420 | -2.073050277 |
| 1200 | -2.073050277 |
| 1878 | -2.073050277 |
| 1051 | -2.073050277 |
| 1904 | -2.073050277 |
| 1611 | -2.073050277 |
| 1569 | -2.073050277 |
| 1514 | -2.073050277 |
| 1675 | -1.963466935 |
| 1483 | -1.963466935 |
| 2325 | -1.963466935 |
| 1102 | -1.963466935 |
| 2564 | -1.963466935 |
| 2121 | -1.963466935 |
| 2552 | -1.963466935 |
| 1859 | -1.963466935 |
| 2994 | -1.963466935 |
| 2926 | -1.853883593 |
| 2836 | -1.853883593 |
| 1009 | -1.853883593 |

|  |  |
| --- | --- |
| **Correlation coefficient of monthly income and Age** | 0.4974 |
|  |  |
| **Regression Line** | y = -2951.58 + 256.15x |
|  |  |
| **coefficient of determination** | 0.2474 |



The correlation coefficient is a positive 0.4974 indicating a positive correlation between Age and monthly income indicating that as one increases the other also tends to increase.



First 20 predicted values:

|  |  |
| --- | --- |
| **X** | **Predicted Monthly By Age** |
| 1420 | 1659.18136 |
| 1200 | 1659.18136 |
| 1878 | 1659.18136 |
| 1051 | 1659.18136 |
| 1904 | 1659.18136 |
| 1611 | 1659.18136 |
| 1569 | 1659.18136 |
| 1514 | 1659.18136 |
| 1675 | 1915.335034 |
| 1483 | 1915.335034 |
| 2325 | 1915.335034 |
| 1102 | 1915.335034 |
| 2564 | 1915.335034 |
| 2121 | 1915.335034 |
| 2552 | 1915.335034 |
| 1859 | 1915.335034 |
| 2994 | 1915.335034 |
| 2926 | 2171.488709 |
| 2836 | 2171.488709 |
| 1009 | 2171.488709 |

**Interpretation for slope and y-intercept:**

The intercept for Years at the company and monthly income is a negative 2951.58. This represents the predicted monthly income when age is zero but it is not possible age to be zero. We have a positive slope of 256.15 and it shows that the relationship between monthly income and age that as age increases the income also tends to increase. Overall it suggests a positive relationship between age and monthly income.

**Interpretation for Coefficient of Determination:**

The Coefficient of Determination is 24.74% which tells us that only 24.74% of the variance in the monthly income can be explained by age of the employee. As we can explain that as age increases employees will have higher income as his experience & education also increases with his increasing age.