**Advance Statistics - Final Assignment**

**Question 1 (20 marks)**

A salary survey was conducted to explore the monthly salary of a sample of employees from different education background, who are working in public and private organizations.

Use the **Salary Data** Sheet provided to perform the following tasks and comment on the results:

1. Identify the type of each parameter (Qualitative & Quantitative).
2. Create summary table for Job Level.
3. Draw bar graph for Education.
4. Draw pie chart for Sector.
5. Draw a histogram for salary.
6. Calculate descriptive statistics of Salary and test its normality.
7. Draw box blot for Age and determine the existence of outliers.
8. Draw a Pareto chart for total salary of each Discipline & present your conclusion about the vital few.

**Answer:**

1. In the **Salary Data** provided, an analysis was conducted to discern the nature of each parameter, categorizing them into qualitative and quantitative types. This classification aids in understanding the fundamental characteristics of the dataset and informs subsequent analytical approaches.

**Qualitative Parameters:**

Qualitative parameters, also referred to as categorical data, encompass attributes that represent distinct categories or labels rather than numerical values. In the dataset, the following parameters were identified as qualitative:

['Job Level', 'Sector', 'Education', 'Discipline', 'Salary']

**Quantitative Parameters:**

Quantitative parameters entail measurable quantities, typically expressed as numerical values. These parameters facilitate quantitative analysis and statistical computations. Within the dataset, the following parameters were recognized as quantitative:

['Name', 'Age', 'Customer Satisfaction']

1. To create a summary table for the Job Level in your dataset, we use the groupby function in pandas to group the data by the Job Level and then calculate summary statistics or counts. The code by Python:

A close-up of a computer code

Description automatically generated

After we import our data and the necessary library we see that we have **83 junior ,**

**295 manager and 305 senior .**

1. **To draw a bar graph for the 'Education' we can use Plotly Express it`s strong library in python to draw graphs:**

**A screenshot of a computer

Description automatically generated**

1. **To draw pie chart for sector:**

**We use px.pie() to create a pie chart using Plotly Express.**

**Then, names='Sector' specifies the 'Sector' column as the data for the pie chart.**

**We sets the title of the pie chart**

**title='Pie Chart of Sector' sets the title of the pie chart.**

**Finaly, fig.show() displays the pie chart.**

**The final pie chart for Sector:**

**A blue and red circle with a red circle

Description automatically generated**

1. Draw a histogram for salary.

we can use Polly Express again to draw a histogram for the 'Salary' column in our dataset. Here's how we do it:

**First,** we need to **convert the salary data from strings to integers by removing the commas and converting them to numeric format.**

**Second,** create a DataFrame from the numeric salary data.

**Finally, we create a histogram using Plotly Express with the 'Salary' column as the data for the histogram, using the numeric values as bins, and count the frequency of occurrence of each salary value.**

**A graph with blue lines and text

Description automatically generated with medium confidenceThe Final histogram:**

1. To Calculate descriptive statistics of Salary and test normality we use **Minitab**

**First**, To Calculate Descriptive Statistics:

We Open Minitab and import our dataset.

Go to "Stat" > "Basic Statistics" > "Display Descriptive Statistics".

In the "Variables" box, select the 'Salary' column.

**And there is the final answer by the steps from Minitab:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Second, Test Normality:**

**After obtaining the descriptive statistics we test the normality assumption.**

**Go to "Stat" > "Basic Statistics" > "Normality Test".**

**In the "Variable" box, select the 'Salary' column.**

**And there is the final answer by the steps from Minitab:**

**A screenshot of a computer

Description automatically generated**

****

**We can also use python to answer the above question:**

we use the describe() function to obtain summary statistics and perform a normality test using statistical methods

**A screenshot of a computer

Description automatically generated**

Since the p-value is extremely small (close to zero), much smaller than any conventional significance level (such as 0.05), it provides strong evidence against the null hypothesis. Therefore, you would reject the null hypothesis and conclude that the data significantly deviates from a normal distribution.

**In summary, based on the normality test result with the given p-value, you would conclude that the 'Salary' data is not normally distributed.**

1. **To draw a boxplot for the 'Age' column in our dataset and determine the existence of outliers, we use seaborn, a Python data visualization library. Here's how you can do it:**

**A screen shot of a diagram

Description automatically generated**

**To identify the outliers, we use the interquartile range (IQR) method. Here's how we do it:**

**A screenshot of a computer

Description automatically generated**

**We conclude that there are no outliers in the “Age” factor.**

8)Draw a Pareto chart for total salary of each Discipline & present your conclusion about the vital few.

To create a Pareto chart for the total salary of each discipline, We do the following steps using Python:

1.Calculate the total salary for each discipline.

2.Sort the disciplines based on their total salary.

3.Create a Pareto chart showing the cumulative percentage of total salary.

4.Analyze the vital few disciplines based on their contribution to the total salary.

**As the following:**

**A screenshot of a computer program

Description automatically generated**

**A graph with a red line

Description automatically generatedAnd there is the output:**

**And Based on the Output , we can draw the following conclusions:**

**Based on these findings, it's evident that Engineering, IT, and Business are the vital few disciplines, collectively contributing to a significant portion of the total salary. These disciplines play a crucial role in shaping the overall salary distribution within the dataset. Therefore, focusing on these disciplines may be essential for strategic decision-making or resource allocation within the organization.**

**Question 2 (20 marks)**

1. You are working in TV set factory. The manufactured TV has a normal distribution life with *m* = 3,500 working hours and *s* = 200 hours.
2. What is the probability that a TV will work less than 3,350 hours?

We will answer the following question using Minitab:

**"Calc" > "Probability Distributions" > "Normal".**

**And we will input data as the following:**

A screenshot of a computer

Description automatically generated

**The output:**

**A screenshot of a computer function

Description automatically generated**

**The probability obtained from the Z-table is approximately 0.2266**

**Finally, probability as a percentage. Therefore, the probability that a TV will work less than 3350 hours is approximately 22.66%**

1. What is the probability that a TV will work more than 3,750 hours?

As the previous question we will do the same steps

A screenshot of a computer

Description automatically generated**input data as the following:**

A white background with black text

Description automatically generated

But in this question, we don’t want to calculate the cumulative probability of less than 3750

So we know that the sum of probability = 1 , So to find the probability that a TV will work more than 3,750 hours, we subtract this probability from 1.

**It will be = 1-0.894350 = 0.10565**

**Finally, probability as a percentage. Therefore, the probability that a TV will work more than 3750 hours is approximately 10.565%**

1. What is the probability that a TV will work between 3,350 & 3,750 hours?

To find the probability of a TV working between 3,350 and 3,750 hours, **we can subtract the probability of the TV working less than 3,750 hours from the probability of the TV working less than 3,350 hours**. Since we calculated these values previously, the calculation would be as follows:

**0.894350−0.2266=0.66775**

**Therefore, the probability that a TV will work between 3,350 and 3,750 hours is approximately** **66.775%."**

1. What is TV life that you are confident 95% it will keep working?

To find the TV life such that you are confident 95% it will keep working, we need to find the value of the TV's working hours at the 95th percentile of the normal distribution. This value represents the point below which 95% of the TV working hours fall.

And we will use Minitab to find it

A screenshot of a computer

Description automatically generatedWe will change the cumulative probability to the inverse cumulative probability and input value will be 1 – 0.95 = 0.05 as the following:

And the final result is

A white background with black text

Description automatically generated

Therefore, with a confidence level of 95%, we can expect the majority of TVs to continue working for at **least 3171.03 hours.**

1. You are working in a bank. You have collected enough data to determine the average time needed to serve one customer and found that it follows a normal distribution with *m* = 4.78 minutes and *s* = 1.32 minutes.
2. What is the probability that you will serve 10 customers every hour?

Mean (μ) = 4.78 minutes

Standard deviation (σ) = 1.32 minutes

**Let's calculate the total time to serve 10 customers:**

Since 10 customers are to serve in every hour , 60 minutes, each customer will be served in an average of **6 minutes**

To find this probability, we can use the cumulative distribution function (CDF) of the normal distribution, and we will use Minitab:

A screenshot of a computer

Description automatically generated

And the output is:

A black text on a white background

Description automatically generated

**Therefore, the probability that you will serve 10 customers every hour is nearly** **88.2320%.**

1. What is the probability that you will serve more than 15 customers every hour?

We will use Minitab But, in this question, we don’t want to calculate the cumulative probability of less than 15 customers every hour .

So we know that the sum of probability = 1 , So to find the probability that you will serve more than 15 customers every hour, we subtract this probability from 1.

Since 15 customers or more are to serve in every hour , 60 minutes, each customer will be served in an average of 4 minutes. The probability that you will serve more than 15 customers every hour by Minitab:

A black text on a white background

Description automatically generated

**P (X>4) = 1-0.27729 = 0.72271**

Hence, the probability that you will serve more than 15 customers every hour is **0.72271 or 72.271%**

1. What is the probability that you will serve between 10 & 15 customers every hour?

To find the probability of serve between 10 & 15 customers every hour, **we can subtract the probability of less than 15 customers every hour from the probability of serve 10 customers every hour**. Since we calculated these values previously, the calculation would be as follows:

**0.882320- 0.27729 =0.60503**

**Therefore, the probability that** serve between 10 & 15 customers every hour **approximately** **60.503%."**

1. What is the number of customers you will be 95% confident that you will serve every hour?

To find the number of customers you will be 95% confident that you will serve every hour

we will use Minitab to find it

We will change the cumulative probability to the inverse cumulative probability and input value **will be 1 – 0.95 = 0.05 as the following:**

A white background with black text

Description automatically generated

Hence, the number of customers at which 95% confident is achieved that you will serve every hour is **nearly 3 customers.**

1. You are thinking about signing a contract, as a supplier for one of the biggest global exporting company. The draft contract obligates you to deliver 20 tons of orange every week. The delivery process of orange during this season follows a normal distribution with *m* = 22.5 tons every week and *s* = 3.2 tons.
2. What is the probability that you will achieve the contract terms?

To find the probability of achieving the contract terms, which is delivering at least 20 tons of oranges every week, we need to calculate the probability of the delivery process being greater than or equal to 20 tons.

**Mean (μ) = 22.5 tons**

**Standard deviation (σ) = 3.2 tons**

This represents the probability of delivering at least 20 tons of oranges every week.

In Minitab, you can use the Probability Distribution Calculator for the normal distribution to find this probability we will input data as the following:

A screenshot of a computer

Description automatically generated

And the output is :

A white background with black text

Description automatically generated

Based on the calculation, the probability of achieving the contract terms, which is delivering at least 20 tons of oranges every week, is approximately = 1-**0.217328 = 0.782672 or 78.2672%**

1. What is the orange quantity that you will be 95% confident that you will deliver every week?

To find the orange quantity that you will be 95% confident that you will deliver every week we will use Minitab to find it

We will change the cumulative probability to the inverse cumulative probability and input value **will be 1 – 0.95 = 0.05 as the following:**

**A white background with black text

Description automatically generated**

**Hence, the orange quantity at which 95% confident is achieved that you will deliver every week is 17.2365, nearly 17 tons oranges.**

**Question 3 (15 marks)**

A supplier was requested to deliver order within 25 to 35 days after receiving the Purchase Order.

Use **Delivery Time Data** to perform the following:

1. Calculate the capability indices and provide your comments on the results.

We will use Minitab to calculate the capability indices so we will import the **Delivery Time Data**

And we will divide the data we have into **13 groups**, each group containing **5 values as wrote in the Delivery Time Data**

A screenshot of a spreadsheet

Description automatically generated**The data in Minitab will be like that :**

After that we going to “State” 🡪 “Capability Analysis” 🡪 “Normal”

As the following:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**The output will be:**



**Based on the provided data and the estimated capability indices in Minitab, we analyze the results and provide it:**

**1)Overall Capability:**

* **Pp:** 0.51
* **PPL:** 0.59
* **PPU:** 0.43
* **Ppk:** 0.43
* **Cp:** 0.50
* **CPL:** 0.58
* **CPU:** 0.42
* **Cpk:** 0.42

2) **Process Data:**

* **Sample Mean:** 30.7539
* **Sample N:** 64
* **StDev(Overall):** 3.27027
* **StDev(Within):** 3.33123

3) **Performance:**

* **PPM < LSL:** 15625.00
* **PPM > USL:** 109375.00
* **PPM Total:** 125000.00
* **Expected Within:** 42059.40

**Conclusion:**

* The calculated capability indices indicate that the supplier's process has moderate capability to meet the specified delivery time requirements.
* The Ppk value of 0.43 suggests that the process is slightly off-center with respect to the specification limits, indicating the need for improvement.
* While the Ppk is slightly lower than desirable, it still indicates that the process is capable of meeting the customer requirements to a reasonable extent.
* The process data shows that the sample mean is close to the target, but there is variation within the process, as indicated by the standard deviations.

1. Provide your recommendation to improve the supplier performance.

**from my point of view to improve supplier performance in a statistical way, we can focus on utilizing statistical tools and methodologies to identify areas for improvement and optimize the delivery process, like:**

1. **Statistical Process Control (SPC):** Implement SPC techniques such as control charts to monitor delivery times over time. By analyzing control charts, the supplier can identify trends, shifts, or anomalies in the process that may indicate underlying issues affecting performance. This proactive approach allows for early detection of problems and timely intervention.
2. **Regression Analysis:** Perform regression analysis to identify factors influencing delivery times. By analyzing historical data, regression models can help identify correlations between various factors such as order volume, transportation mode, or supplier location, and delivery performance. This insight can inform strategic decisions to optimize delivery processes and mitigate delays.
3. **Design of Experiments (DOE):** Utilize DOE techniques to systematically evaluate the impact of different process variables on delivery performance. By conducting controlled experiments, the supplier can identify the most effective combinations of factors that lead to improved delivery times. This empirical approach enables data-driven decision-making and optimization of delivery processes.

**There are many statistical methods other than that, but I think they are the most important to improve the supplier performance.**

**Question 4 (10 marks)**

Cycle Time was measured and found to be a non-normal distribution.

Use **Cycle Time** **Data** to perform the following:

1. Transform Cycle Time data using Box-Cox transformation.

At first, we import our data in Minitab.

A screenshot of a computer

Description automatically generatedAfter that we Check for non-normality using Minitab by the normality test (Anderson-Darling)

**A screenshot of a test

Description automatically generated**

**And this is our output from Minitab:**



Since the p-value is extremely small (close to zero), much smaller than any conventional significance level (such as 0.05), it provides strong evidence against the null hypothesis. Therefore, you would reject the null hypothesis and conclude that the data significantly deviates from a normal distribution.

**In summary, based on the normality test result with the given p-value, we see that our data is found to be a non-normal distribution.**

**Finally, we Apply Box-Cox transformation using Minitab as the following:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**



And we store the normalized data at C2:

A screenshot of a table

Description automatically generated

And We assess the normality of the transformed data using the same normality test employed in step 1 :



**Based on the obtained p-value of 0.444 from the normality test, we conclude that the data follows a normal distribution**

And from the results before : (95.0% Confidence Interval):

Estimate: -0.3

Lower Confidence Limit (CL): -1.12

Upper Confidence Limit (CL): -0.5

The estimated value for the parameter is -0.3. With a 95.0% confidence level, the lower limit of the confidence interval is -1.12, and the upper limit is -0.5.

Given the precision of the estimates, the rounded value for the parameter is -0.3.

**These results suggest that the parameter falls within the specified confidence interval, indicating a certain level of confidence in the estimated value.**

1. Transform Cycle Time data using Johnson transformation.

We already Check for non-normality in the previous point.

Wo we **Apply** Johnson transformation **using Minitab as the following:**

**A screenshot of a computer

Description automatically generated**

**We input our data and use p-value 0.05:**

**A screenshot of a computer

Description automatically generated**

**And there is our output:**

We save the new transformed data at C3 :

A screenshot of a table

Description automatically generated

Based on the results obtained from the Johnson transformation, the Anderson-Darling (AD) statistic is 37 with a corresponding **p-value of 0.217.**

This indicates that the transformed data does not significantly deviate from a normal distribution at a significance level of 0.05.

Furthermore, **the p-value for the best fit transformation is 0.830,** suggesting that the chosen transformation adequately approximates a normal distribution.

**The Z-score for the best fit is 0.74**, indicating that the transformed data closely resembles a normal distribution.

**Question 5 (20 marks)**

Car manufacturing plant is studying the stability of its processes.

Use **Car Factory Data** to perform the following:

1. Provide your comments on the stability of daily production.

At first we import our **Car Factory Data in Minitab:**

A screenshot of a computer

Description automatically generated

Step 1: Analysis of Daily Production

Calculate the average number of cars produced daily:

A close-up of a number

Description automatically generated

We see that the average number of cars produced daily is **957.2 cars.**

Step 2: Calculate the standard deviation of the number of cars produced daily and it`s **247.8.**

A screenshot of a computer

Description automatically generatedStep 3: Plot a control chart for daily production:



Based on the descriptive statistics provided for the number of cars produced daily and the time series plot:

Average Daily Production: **The average number of cars produced daily is approximately 957.2,** with a **standard error of 36.5**. This indicates the central tendency of the production process.

Variability: The standard deviation of **approximately 247.8** suggests a considerable variability in daily production around the mean. **This variability may indicate fluctuations or inconsistencies in the production process.**

Range of Production: The range of daily production spans from a **minimum of 540 cars to a maximum of 1380 cars**. **This wide range further emphasizes the variability in production levels.**

Trends or Fluctuations: To assess stability, it's essential to examine trends or fluctuations in daily production over time. This can be achieved by visualizing the data using time series plots or control charts. Any consistent trends, cycles, or irregular fluctuations should be identified and analyzed to understand their impact on production stability.

**Overall, while the average daily production provides an indication of the expected output, the significant variability and range suggest that the production process may not be entirely stable. Further investigation into the causes of variability and any observed trends or fluctuations is warranted to improve production stability and efficiency.**

1. Provide your comments on the stability of defective cars.

At first we need to Combine cars inspected and Defective Cars Data using Minitab and we take the same rate for the inspected cars so we will use the formula Defective Car Rate = Number of fixed cars) / (Number of cars inspected) and will store the values at column C6:

A screenshot of a calculator

Description automatically generated

A screenshot of a computer

Description automatically generated

A group of numbers and symbols

Description automatically generatedDescriptive Statistics for Defective Car Rate:

Histogram for Defective Car Rate:



**Based on the descriptive statistics provided for the Defective Car Rate and the histogram:**

Average Defective Car Rate: The mean defective car rate is **approximately 0.1501**, indicating that, on average, **around 15.01%** of the cars produced are defective.

Variability in Defective Car Rate: **The standard deviation of 0.0928** suggests a moderate variability in the defective car rate around the mean. **This indicates that the rate of defective cars may vary noticeably from the average value.**

Common Value: **The mode values, particularly 0.0641026 and 0.0789474**, suggest that these are the most frequent defective car rates observed in the data. **This implies that there are specific rates at which defects occur more frequently than others.**

Consistency in Defective Car Rate: While the median value of **0.1202** suggests a relatively consistent central tendency, the mode values indicate that there are several peaks in the distribution, indicating potential variability in the defective car rates.

**Overall Stability**: The descriptive statistics, including the mean, median, standard deviation, and mode values, **suggest a moderate level of stability in the defective car rates**. However, the presence of multiple mode values and the range between the minimum and maximum values indicate some variability in the data.

Based on these observations, further investigation may be warranted to understand the factors contributing to the observed variability in defective car rates and to ensure that production processes are adequately controlled to minimize defects.

1. Provide your comments on the stability of car defects.

First, we add new column to our data by Calculate **Average Parts Fixed per Car** on Minitab and we save it on C7, Use the formula: "Number of fixed parts" / "Number of fixed cars":

A screenshot of a calculator

Description automatically generated

**A screenshot of a computer

Description automatically generated**

Descriptive Statistics for **Average Parts Fixed per Car**:

A number of cars with numbers and letters

Description automatically generated with medium confidence

Histogram for **Average Parts Fixed per Car**:

****

**As we see the data have outliers so we will draw a box plot to see the outliers clearly**

**box plot for Average Parts Fixed per Car:**

****

**Outliers:**

* **The maximum value of 22 indicates the presence of outliers in the data. Additionally, the mode values of 1.5, 1.66667, and 4 suggest that these values are more frequently observed, but there might be some irregularities in the data distribution.**
* **The special outlier with a value of 22 warrants further investigation, as it significantly deviates from the typical range of values and may indicate an anomaly in the production process.**

1. **Average Parts Fixed per Car:**
   * The mean average parts fixed per car is approximately **2.309**, indicating that, on average, around **2.309** parts are fixed per car during the production process.
   * The standard deviation of **3.070** suggests variability in the number of parts fixed per car, with values ranging from **1 to 22**.
   * The quartile values **(Q1 = 1.356, Median = 1.598, Q3 = 2.229**) provide insights into the distribution of the average parts fixed per car.
2. **Outliers**:
   * **The special outlier with a value of 22 warrants further investigation,** as it significantly deviates from the typical range of values and may indicate an anomaly in the production process.
3. **Graphical Analysis**:
   * The presence of outliers and the special outlier can be visually confirmed through graphical analysis, **such as boxplots or histogram plots**, which may highlight extreme values or patterns in the data distribution.
4. **Root Cause Analysis**:
   * Conduct a root cause analysis to identify the factors contributing to the observed variability and outliers in the average parts fixed per car. This analysis can help uncover any underlying issues in the production process that may need to be addressed.
5. **Process Improvement**:
   * Based on the analysis, identify opportunities for process improvement to reduce variability and address outliers in the average parts fixed per car. This may involve implementing quality control measures, enhancing training programs, or optimizing equipment maintenance procedures.
6. **Recommendations**:
   * Provide recommendations for optimizing production processes and minimizing variability in the average parts fixed per car. These recommendations should be based on insights gained from the analysis and aimed at improving overall production quality and efficiency.

**By addressing outliers and variability in the average parts fixed per car, the stability of car defects can be improved, leading to enhanced product quality and customer satisfaction. Further investigation into the special outlier with a value of 22 is recommended to understand its root cause and implement appropriate corrective actions.**

**Question 6 (15 marks)**

Sales manager wishing to predict the future sales.

Use **Sales Data** to perform the following:

1. Provide your comments on the sales trend.

At first we import our **Sales Data** **in Minitab:**

**A screenshot of a computer

Description automatically generated**

**To provide comments on the sales trend using Minitab, we can perform a simple linear regression analysis to identify any significant trend in the sales data over time and we can also use Trend Analysis, and we will have the same results.**

**A screenshot of a computer

Description automatically generatedFirst using Trend Analysis: we input our data:**

**A screenshot of a computer

Description automatically generated**

**And this is our output:**

**A white paper with black text

Description automatically generated**

**A screenshot of a computer

Description automatically generatedSecond using simple linear regression analysis: we input our data:**

A screenshot of a computer

Description automatically generated

**And this is our output and it’s the same as first way using Trend Analysis:**



A screenshot of a computer

Description automatically generated

**Based on the provided outputs from the trend analysis for sales, we can comment on the sales trend as follows:**

1. **Trend Equation:**
   * The trend equation obtained from the analysis indicates a polynomial trend:
   * **Yt = 20098 + 13808t + 41t^2**
   * This equation suggests a quadratic trend in the sales data, indicating that the sales pattern is not linear over time but rather exhibits a curvature.
2. **Accuracy Measures:**
   * The Mean Absolute Percentage Error (MAPE) is 8%, indicating that, on average, the model's predictions deviate by 8% from the actual sales values. Lower MAPE values suggest better model accuracy.
   * The Mean Absolute Deviation (MAD) is 8619, representing the average absolute difference between predicted and actual sales values.
   * The Mean Squared Deviation (MSD) is 96786492, which measures the average squared difference between predicted and actual sales values.
3. **Regression Analysis:**
   * The regression equation obtained from the polynomial regression analysis is:
   * **ales = 20098 + 13808 × Quarter + 41.3 × Quarter^2**
   * The R-squared value (R-Sq) is 98.3%, indicating that 98.3% of the variation in sales can be explained by the quadratic regression model. This suggests a strong fit of the model to the data.
   * The adjusted R-squared value (R-Sq(adj)) is 98.1%, which adjusts for the number of predictors in the model.
4. **Analysis of Variance:**
   * The analysis of variance (ANOVA) table shows that the regression model is highly significant, with an F-statistic of 444.21 and a p-value of 0.000. This indicates that the overall regression model is a good fit for the data.
   * Both the linear and quadratic terms in the regression model are individually significant, as indicated by their low p-values.
5. **Trend Analysis Plot:**
   * The trend analysis plot visually displays the fitted trend line along with the actual sales data points. This helps in visualizing the trend pattern and assessing how well the trend line fits the data.

**Overall, the trend analysis indicates a significant quadratic trend in the sales data, with a strong fit of the polynomial regression model. This suggests that sales exhibit a curvilinear pattern over time, rather than a simple linear trend. The model's accuracy measures also suggest relatively low prediction errors, further supporting the reliability of the trend analysis results.**

1. Provide your comments on the sales seasonality.

We will do **Seasonal Decomposition** to analyze sales seasonality using Minitab:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**And there is Decomposition Plots:**





**A table with numbers and numbers

Description automatically generated**

**Based on the Time Series Decomposition analysis results for sales data, here are the comments on the sales seasonality:**

1. **Seasonal Indices:**
   * **The seasonal indices represent the relative strength of the seasonal component at different periods**.
   * The indices indicate how much above or below the trend line the seasonal component is during each period.
   * In the provided data:
     + **Period 1 has an index of 1015.6**, suggesting a strong positive seasonal effect.
     + **Period 2 has an index of 3515.6**, indicating an even stronger positive seasonal effect.
     + **Period 3 also shows a strong positive seasonal effect with an index of 8515.6.**
     + **Period 4 stands out with a negative index of -13046.9,** indicating a significant deviation from the trend, **possibly reflecting a seasonal trough or anomaly.**
2. **Decomposition Plot:**
   * The decomposition plot shows **the fitted values (FITS1) and residuals (RESI1)** after decomposing the sales data into trend, seasonal, and random components.
   * The fitted values represent the trend and seasonal components combined, while the residuals represent the random component or noise in the data.
   * Examining the plot can provide insights into how well the decomposition captures the underlying patterns in the sales data and any remaining variability not accounted for by the model.
3. **Accuracy Measures:**
   * Accuracy measures such as Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), and Mean Squared Deviation (MSD) provide an assessment of the model's performance in capturing the observed data.
   * **A low MAPE, MAD, and MSD indicate good accuracy and fit of the decomposition model to the sales data.**
4. **Overall Seasonality:**
   * Based on the seasonal indices and decomposition plot, it's evident that the sales data exhibit strong seasonal patterns.
   * These seasonal patterns are characterized by periodic fluctuations that repeat at regular intervals, influencing sales performance throughout the year.
   * Understanding these seasonal patterns is crucial for forecasting and planning purposes, as they can help anticipate fluctuations in sales volume and adjust strategies accordingly.

**By analyzing the seasonal indices, decomposition plot, and accuracy measures, we can conclude that the sales data exhibit significant seasonality, with distinct seasonal patterns affecting sales performance over time.**

1. Provide your comments on the obtained accuracy model.

Based on the accuracy measures provided in the Time Series Decomposition analysis results, here are the comments on the obtained accuracy model:

1. **Mean Absolute Percentage Error (MAPE):**
   * MAPE measures the average absolute percentage difference between the observed and predicted values.
   * A lower MAPE indicates better accuracy, with 8% being relatively low.
   * A MAPE of 8% suggests that, on average, the model's predictions deviate from the actual sales values by approximately 8%.
2. **Mean Absolute Deviation (MAD):**
   * MAD measures the average absolute difference between the observed and predicted values.
   * Similar to MAPE, a lower MAD indicates better accuracy.
   * The MAD of 7365 suggests that, on average, the model's predictions deviate from the actual sales values by approximately 7365 units.
3. **Mean Squared Deviation (MSD):**
   * MSD measures the average squared difference between the observed and predicted values.
   * It provides a measure of the variability of prediction errors.
   * The MSD value of 82610155 indicates the average squared deviation of approximately 82610155 units.
4. **Overall Assessment:**
   * **The accuracy model appears to perform reasonably well, as indicated by the relatively low MAPE and MAD values.**
   * **The model's ability to predict sales values with an average deviation of around 8% and 7365 units suggests that it captures a significant portion of the variability in the data.**
5. Predict the upcoming 3 quarter sales.

To predict the upcoming 3 quarter sales using the provided time series decomposition analysis for sales, we'll utilize the trend equation and seasonal indices. Here's the process:

**Trend Equation: Yt = (16509 + 14668 × t) and add the Seasonal Indices**

**Seasonal Indices:**

**Quarter 1: 1015.6**

**Quarter 2: 3515.6**

**Quarter 3:** **8515.6**

**Quarter 4: -13046.9**

Forecasting Process:

For each upcoming quarter, we calculate the trend component using the trend equation.

Adjust the trend component using the corresponding seasonal index to incorporate the seasonal effect.

Add any remaining residuals or random component from the decomposition.

Forecasted Sales:

Once the calculations are performed for each quarter, we will obtain the forecasted sales figures for **Quarters 19, 20, and 21.**

* Quarter 19: using the trend equation and adjusting for the Q3 effect.
* Quarter 20: using the trend equation and adjusting for the Q4 effect.
* Quarter 21: using the trend equation and adjusting for the Q1 effect.

Results:

**Quarter 19: Y19 = (16509 + 14668 × 19 ) + 8515.6 = 303716.6**

**Quarter 20: Y20 = (16509 + 14668 × 20) + -13046.9 = 296822.1**

**Quarter 21: Y21 = (16509 + 14668 × 21) +1015.6 = 325552.6**