1. i)

Data Checks:

1. Check for missing values: Scan the dataset for any missing or blank values. If there are any, decide whether to remove those records or impute the missing values based on the nature of the data and the analysis requirements.

2. Check for outliers: Identify any extreme values that deviate significantly from the rest of the data. Outliers can skew the results of the analysis, so it's essential to investigate them further. Determine if they are genuine data points or errors that need to be corrected or removed.

3. Check for data consistency: Ensure that the data is consistent across all records. Look for any discrepancies or inconsistencies in the data format, units of measurement, or data types. Standardize the data if necessary to maintain consistency.

4. Check for data accuracy: Verify that the data values are within the expected range and align with the data description. If there are any values that seem unrealistic or implausible, investigate further to determine if they are errors that need to be corrected or removed.

5. Check for data relevance: Assess whether all the variables in the dataset are relevant to the analysis objectives. Remove any irrelevant variables that do not contribute to the analysis.

6. Check for data duplicates: Look for any duplicate records in the dataset. Duplicates can bias the analysis results, so it's important to identify and remove them.

7. Check for data completeness: Ensure that the dataset covers the required time period and includes all the necessary variables for the analysis. If there are any gaps or missing variables, consider whether additional data collection is needed.

Corrections and Deletions:

Based on the data checks performed, make the necessary corrections and deletions to ensure the data is fit for purpose. This may include:

1. Removing records with missing values, if appropriate.

2. Correcting or removing outliers, depending on their nature and impact on the analysis.

3. Standardizing data formats and units of measurement to ensure consistency.

4. Correcting any errors or inaccuracies identified in the data.

5. Removing irrelevant variables from the dataset.

6. Eliminating duplicate records.

7. Filling any gaps in the data or collecting additional data if necessary.

After making the corrections and deletions, document the changes made to the dataset and the rationale behind each decision. This documentation will help maintain transparency and reproducibility in the analysis.

Once the data checks, corrections, and deletions are complete, the dataset should be ready for further analysis and modeling as per the rest of the question.

**ii) Second Part**

(ii) Construct two separate charts to show the number of taxi pick-ups by day of the week and by the pick-up hour of the day.

To create these charts, you'll need to use the taxi pick-up data provided, which likely includes columns for the day of the week and the pick-up hour. In your Excel spreadsheet, follow these steps:

1. Count the number of taxi pick-ups for each day of the week:

- Create a new sheet named "Day of Week".

- In the first column, list the days of the week (e.g., Monday, Tuesday, etc.).

- Use the COUNTIF function to count the number of pick-ups for each day of the week. For example, if the day of the week is in column A of the original data sheet, use a formula like: =COUNTIF('Data Sheet'!A:A, "Monday") for Monday, and repeat for each day.

2. Create a chart for the number of taxi pick-ups by day of the week:

- Select the data range containing the days of the week and their corresponding pick-up counts.

- Go to the Insert tab and choose the appropriate chart type (e.g., Column chart or Bar chart).

- Customize the chart title, axis labels, and any other formatting as needed.

3. Count the number of taxi pick-ups for each hour of the day:

- Create a new sheet named "Hour of Day".

- In the first column, list the hours of the day (0 to 23).

- Use the COUNTIF function to count the number of pick-ups for each hour. For example, if the pick-up hour is in column B of the original data sheet, use a formula like: =COUNTIF('Data Sheet'!B:B, 0) for hour 0, and repeat for each hour.

4. Create a chart for the number of taxi pick-ups by pick-up hour:

- Select the data range containing the hours and their corresponding pick-up counts.

- Go to the Insert tab and choose the appropriate chart type (e.g., Column chart or Bar chart).

- Customize the chart title, axis labels, and any other formatting as needed.

By following these steps, you will create two separate charts in your Excel spreadsheet: one showing the number of taxi pick-ups by day of the week and another showing the number of taxi pick-ups by pick-up hour.

**iii) Third Part**

Understood. I'll focus on answering part (iii) of the question, which asks to construct a chart that shows the relationship between distance and journey duration. Since this is a modeling question, I'll provide guidance on how you can create this chart in Excel.

To create a chart showing the relationship between distance and journey duration in Excel, follow these steps:

1. Assuming your data is in two columns (e.g., Column A for distance and Column B for journey duration), select the data range including the headers.

2. Go to the "Insert" tab in the Excel ribbon and click on the "Scatter" chart type. Choose the first option, which is a simple scatter plot without lines.

3. Excel will automatically create a scatter plot with distance on the x-axis and journey duration on the y-axis.

4. To enhance the chart, you can add chart elements like a title, axis labels, and a trendline:

- Click on the chart to select it.

- Go to the "Chart Design" tab in the Excel ribbon.

- Click on the "Add Chart Element" button and select the desired elements from the dropdown menu.

- For example, you can add a title by selecting "Chart Title" and entering a meaningful title for your chart.

- You can also add a trendline by selecting "Trendline" and choosing the appropriate type (e.g., linear, exponential, etc.).

5. Customize the appearance of the chart, if desired, by changing colors, fonts, or other formatting options from the "Format" tab in the ribbon when the chart is selected.

By following these steps, you can create a scatter plot in Excel that visually represents the relationship between distance and journey duration based on the provided data.

Remember to properly label your chart, including a title and axis labels, to ensure clarity and readability for anyone viewing the chart.

**iv) Fourth Part**

(iv) Calculate the expected fare amount using the fare formula (1). [1]

To calculate the expected fare amount using the fare formula (1), you need to use the following formula in your spreadsheet model:

Expected Fare = $2.50 + $0.50 × distance + $0.50 × duration

Where:

- $2.50 is the initial charge

- $0.50 is the per-mile charge

- distance is the distance of the trip in miles

- $0.50 is the per-minute charge

- duration is the duration of the trip in minutes

To implement this in your spreadsheet, create a new column and enter the formula using the cell references for the distance and duration columns. For example, if distance is in column A and duration is in column B, and you want to calculate the expected fare in column C, you would enter the following formula in cell C2 and drag it down:

=2.50 + 0.50 \* A2 + 0.50 \* B2

This will calculate the expected fare for each trip based on the given fare formula.

**v) Fifth part**

To show that the fare formula provided by the manager is a good fit to the 'fare\_amount' data at a 95% confidence level, you can perform a hypothesis test.

Hypothesis:

H0 (null hypothesis): The fare formula is a good fit for the data.

H1 (alternative hypothesis): The fare formula is not a good fit for the data.

Assuming the 'fare\_amount' data follows a normal distribution, you can use a one-sample t-test.

Formula for one-sample t-test:

t = (x̄ - μ) / (s / √n)

Where:

x̄ = sample mean of the 'fare\_amount' data

μ = expected mean (calculated using the fare formula)

s = sample standard deviation of the 'fare\_amount' data

n = number of data points

Steps:

1. Calculate the expected fare amount for each data point using the given fare formula.

2. Calculate the sample mean (x̄) and sample standard deviation (s) of the 'fare\_amount' data.

3. Calculate the t-statistic using the formula above.

4. Determine the critical t-value for a 95% confidence level and the appropriate degrees of freedom (n-1).

5. Compare the calculated t-statistic with the critical t-value. If the absolute value of the calculated t-statistic is less than the critical t-value, fail to reject the null hypothesis (H0). Otherwise, reject the null hypothesis.

If you fail to reject the null hypothesis, it suggests that the fare formula is a good fit for the 'fare\_amount' data at a 95% confidence level.

**vi) Sixth part**

To analyze the impact of the fare increase calculated in part (vi) on the total number of taxi journeys in July 2021 that begin or end in neighborhood A:

1. Assume the price elasticity of demand for taxi journeys is -0.8 (a common estimate for taxi services). This means that for every 1% increase in fare, the demand for taxi journeys decreases by 0.8%.

2. Calculate the percentage increase in fare:

Percentage increase = (New fare - Original fare) / Original fare \* 100

3. Calculate the percentage decrease in demand using the elasticity:

Percentage decrease in demand = Price elasticity \* Percentage increase in fare

4. Calculate the new number of taxi journeys:

New number of journeys = Original number of journeys \* (1 + Percentage decrease in demand / 100)

For example, if the original fare was $10, the new fare is $12, and there were originally 1000 journeys:

- Percentage increase in fare = ($12 - $10) / $10 \* 100 = 20%

- Percentage decrease in demand = -0.8 \* 20% = -16%

- New number of journeys = 1000 \* (1 + (-16%) / 100) = 840

So the total number of taxi journeys would decrease from 1000 to 840.

**vii) Seventh Part**

For part (vii), you need to analyze the impact of the fare increase calculated in part (vi) on the total number of taxi journeys in July 2021 that begin or end in neighborhood A.

To model this in Excel, you can follow these steps:

1. Calculate the percentage of taxi journeys in the sample data that begin or end in neighborhood A.

2. Assume that this percentage remains constant for July 2021.

3. Estimate the total number of taxi journeys in July 2021 based on the sample data.

4. Apply the percentage from step 1 to the total number of journeys in July 2021 to estimate the number of journeys that begin or end in neighborhood A.

5. Use the price elasticity of demand formula to calculate the change in the number of journeys due to the fare increase:

Percentage change in quantity demanded = Price elasticity of demand × Percentage change in price

Assuming a price elasticity of demand of -0.6 (a common value for taxi services), the formula would be:

Percentage change in journeys = -0.6 × Percentage change in fare

6. Apply this percentage change to the number of journeys that begin or end in neighborhood A to estimate the impact of the fare increase.

For example, if the fare increase is 10% and the number of journeys beginning or ending in neighborhood A is estimated to be 1,000, the impact would be:

Percentage change in journeys = -0.6 × 10% = -6%

Change in journeys = 1,000 × (-6%) = -60

So, the fare increase would lead to a decrease of 60 journeys in July 2021 that begin or end in neighborhood A.

**viii) Eight Part**

To calculate R, the additional adjustment needed to the fare increase factor, to reduce the number of taxi journeys that begin or end in neighbourhood A to 80% of those implied by the sample data, we need to follow these steps:

1. Determine the number of taxi journeys in the sample data that begin or end in neighbourhood A. Let's call this value N.

2. Calculate the expected fare after the proposed increase, as done in part (vi). Let's denote this fare as F.

3. Use the price elasticity of demand formula to determine the fare increase factor (1 + R) that would reduce the number of journeys to 80% of N.

The price elasticity of demand formula is:

Elasticity = (% Change in Quantity Demanded) / (% Change in Price)

We want the quantity demanded to be 80% of N, so the % Change in Quantity Demanded is -20%.

Let's assume the price elasticity of demand for taxi journeys is -0.8 (this means a 1% increase in price leads to a 0.8% decrease in quantity demanded).

Using the formula:

-0.8 = -20% / (% Change in Price)

Solving for % Change in Price:

% Change in Price = -20% / -0.8 = 25%

This means the fare should be increased by 25% to achieve an 80% reduction in quantity demanded.

4. Calculate R using the equation:

F \* (1 + R) = F \* (1 + 25%)

Solving for R:

R = 25% = 0.25

Therefore, the additional adjustment factor R should be 0.25 or 25% to reduce the number of taxi journeys that begin or end in neighbourhood A to 80% of those implied by the sample data.

In the spreadsheet, you can create a separate section for this calculation:

- Count the number of journeys in neighbourhood A (N)

- Use the fare (F) calculated in part (vi)

- Apply the formula: F \* (1 + 0.25) to get the adjusted fare

- Interpret the results and document your findings