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) Charts**

* **Taxi Pick-ups by Day of the Week**
  1. **Data Aggregation:** Use a COUNTIFS formula or a Pivot Table to count journeys for each day of the week. You can extract the day of the week from pickup\_datetime using WEEKDAY function.
  2. **Chart Type:** Create a column or bar chart visualizing your aggregated data.
* **Taxi Pick-ups by Hour of the Day**
  1. **Data Aggregation:** Similar to the above, use COUNTIFS or a Pivot Table. Extract the hour from pickup\_datetime using the HOUR function.
  2. **Chart Type:** Create a column or bar chart to represent the hourly distribution.

**(iii) Relationship Between Distance and Duration**  
1. **Chart Type:** Create a scatter plot with trip\_distance on the x-axis and trip\_duration on the y-axis.

**(iv) Expected Fare Amount**  
1. **Formula Application:** In a new column, apply formula (1) to each journey, using your calculated trip\_distance and trip\_duration.

**(v) Hypothesis Test for Fare Formula**  
1. **Approach:** You can use a paired t-test to compare your calculated expected\_fare\_amount (from step iv) with the actual fare\_amount.  
2. **Null Hypothesis:** The null hypothesis would be that there's no significant difference between the average expected fare and the average actual fare.  
3. **Excel's T.TEST Function:** Use Excel's T.TEST function to perform the test and obtain the p-value. Remember to select the appropriate type of t-test (paired, in this case).  
4. **Interpretation:** If the p-value is less than your significance level (0.05), you would reject the null hypothesis, suggesting that the formula is not a good fit. If the p-value is greater than 0.05, you'd fail to reject the null hypothesis, indicating the formula might be a reasonable proxy.

**(vi) Expected Fare After Increase**  
1. **Apply Increase Factor:** Create a new column. For each journey, multiply the expected\_fare\_amount (calculated in step iv) by the increase factor given by formula (2).

**(vii) Impact of Fare Increase on Neighbourhood A**  
1. **Identify Neighbourhood A Journeys:** Use the Taxi\_location&demand worksheet and criteria to filter journeys that start or end in neighborhood A.  
2. **Demand Adjustment:** For the filtered journeys, use the "impact on demand" data to calculate the new number of journeys after the fare increase. Be sure to incorporate the continuous nature of the demand adjustment. You can use VLOOKUP function to retrieve the corresponding demand impact factor.  
3. **Analysis:** Compare the number of journeys before and after the fare increase in neighborhood A.

**(viii) Determining R**  
1. **Target Journeys:** Calculate the target number of journeys in neighborhood A (80% of the original number).  
2. **Goal Seek/Solver:** Use Excel's Goal Seek or Solver tool:  
\* **Set Cell:** A cell containing the total number of neighborhood A journeys *after* the fare increase (with the adjustable 'R' in the formula).  
\* **To Value:** Your target number of journeys.  
\* **By Changing Cell:** A cell containing a starting value for 'R'.  
3. **Result:** Goal Seek/Solver will find the value of 'R' that achieves the target reduction in neighborhood A journeys.

**Additional Notes:**

* **Audit Trail:** Document all your assumptions, data cleaning steps, formulas, and decisions in a separate worksheet (your audit trail). This is crucial for review and reproducibility.
* **Reasonableness Checks:** At each step, pause and ask yourself if the results make sense. Compare your outputs to the provided context.
* **Chart Clarity:** Label your charts clearly and provide concise titles that summarize the insights.

Remember, this is a framework. The specifics of your Excel formulas will depend on your data organization and modeling choices. Good luck with your CP2 exam!