**HW – Plots for Analyses**

Submission: Completed Excel workbook via Canvas.

Part 1 – Linear Regression

Overview

* This problem extends the ideas from the notes ‘Basic Charts in Excel’, where we learned how to create summary tables, and from ‘Lab – More Charts in Excel’, where we created scatterplots.
* The first tab of the accompanying workbook contains a dataset of public transformation information related to the metro for a sample of hypothetical cities; a subset is shown below:

A screenshot of a data

Description automatically generated

* + The response variable *Y* is ‘Number of weekly riders’ and the other numeric variables are the explanatory *X* variables.

Assignment

* The goal is to analyze the number of weekly riders by creating a summary statistics table for each state and determining the best predictor via simple linear regression.

1. Complete the summary statistics table for each City and Overall.
   * Find the correct formulas to use and set them up so that they can be autofilled down easily.
2. Now we want to determine which *X* variable has the most impact on *Y*, the number of weekly riders. Repeat the following steps for each of the three numeric *X* variables.

* Create a scatterplot with the correct selection of the axes / variables.
* Add informative chart title and axes labels so it is clear which variables are being plotted.
* To add the line of best fit (regression line): Chart Design -> Add Chart Element -> Trendline -> Linear.
* To add the regression information: Right click on the trend line -> Format Trendline -> check the boxes for ‘Display Equation on chart’ and ‘Display R-squared value on chart’.
* Format the points to be a neutral color and the line a brighter color so that the regression line has more emphasis. Change the font color of the regression info to visually ‘link’ it to the regression line.
  + After finishing all plots, add a text box to your workbook and include a quick write-up discussing which *X* variable is the best predictor of *Y* the number of weekly riders. How do you know? Describe the relationship (either visually or from the regression equation) between *Y* and the best *X*.

What will be looked at for grading

* Completed summary statistics table with correct formulas, well-formatted plots with trendline information, sufficient write-up.

Part 2 – Plots of Aggregated Data

Overview

* This problem extends Part 1 of this Lab, where we continued introducing more types of Charts that can be made in Excel.
* The second tab of the accompanying workbook contains two tables: a dataset of daily sales and a table for monthly sales from the previous year. Subsets of both are shown below:

A table with numbers and letters

Description automatically generated

Assignment

* The goal is to analyze the sales data from several perspectives using visuals based on different levels of aggregation.
* After each analysis, add a text box with a quick write-up answering the directed questions.

1. The first analysis will compare monthly sales total for 2022 to that of 2021 via a Line Chart.
2. First, the daily data needs to be aggregated by month. Here’s how:
   * In the daily sales table, you need to create a new column representing the month for each sale date. Create a new column after Date named ‘Month #’ and determine the correct function to use to extract the month.
   * Now in the monthly sales table, add a column for 2022 and use the correct function to calculate the total sales based on the respective month.
   * To have more informative axis labels in the plot, create a new column after ‘Month #’ named ‘Month’. To convert the month # to the abbreviation (e.g. 1 = Jan), use the following formula template:

= TEXT(‘Month #’ \* 29, “mmm”)

This formula just converts the month number to a day number (\* 29) that would be in the desired month and changes the format to the month abbreviation.

1. Now, you can plot two lines, one for 2021 and one for 2022, at the same time by highlighting the Month, 2021 and 2022 -> Insert Line Chart.
   * Be sure to give the chart an informative title. Try a few different legend spots to find the one that is easiest for the reader to figure out which line is which.
   * In the write up: Compare monthly sales from the two years.
2. The second analysis will compare quarterly sales for via Comparative Boxplots.
3. First, the daily data needs to be setup. Here’s how:
   * Boxplots take raw data and summarize it for you. So all that needs to be done is adding a column for the group, in this case quarter of the sales date.
   * Add a column after ‘Month #’ called ‘Quarter’. There is no direct function to convert a date into quarter. So, write an algebra formula involving the month number and one of the various ROUND() functions to end up with the correct quarter (1, 2, 3 or 4). Ensure your formula works for all months.
4. Now, you can insert the Comparative Boxplot for Sales by Quarter.
   * Notice the X-axis labels; they aren’t very informative. This is something that can be fixed directly in data so that better labels are automatically pulled in.
   * Search how to combine text in Excel with a formula and apply this to the Quarter calculation. You want the values to be of the form Q1, Q2,….
   * One additional feature of Comparative Boxplots is connecting the mean markers for each boxplot. To do this: Right click any boxplot -> Format Data Series -> check box for ‘Show mean line’.
   * Be sure to give the chart an informative title.
   * In the write up: Discuss what type of information is shown by the final Comparative Boxplot. Is it more or less informative that the Line Chart in the first analysis? Why? Then compare the quarterly.

What will be looked at for grading

* Correct calculation of Quarter, correct Comparative Boxplot, sufficient write-up.