**Project Instructions**

In this project, you will use population statistics from the US Census Bureau to determine where the greatest income exists around the country and whether there is a correlation between sales and income. We don’t know the incomes of our customers, but we should be able to predict it by looking at their purchase history and locations and comparing that against the census data. Additionally, we want to analyze our inventory, specifically customer ratings and return rate and see if there’s a correlation between the two.

**Draw conclusions**

Draw conclusions from your analysis and use visuals to answer the following questions:

Analysis Questions:

1. What is the correlation (R2 value) between sales and income?
2. What is the correlation (R2 value) between customer ratings and product return rate?
3. What are the linear regression formulas to predict customer income from customer sales?
4. Which customer do you predict has the highest income?
5. Which product will be advertised the most?

**Present your analysis**

You’ll need to present your analysis as a 1 page written summary and visual report in Power BI. Use the following visuals to present your data:

* Income Distribution: Histogram
* Household Income by Location: Map
* Correlations: Scatter Plot with Trendline and Card with R^2 value

Use other visuals as-needed to further present the results of your analysis.

**Histogram**

Although you will need to determine the parameters of the histogram(s) for your project, think about what a histogram is intended to do. In the example below, you can see that the histogram gives you a clear understanding of the distribution of values and helps you visualize the shape of the data. We can see here that it's a right-skewed distribution with a peak of around $40k and an average that is probably a bit to the right of the peak.

Your histogram(s) should also convey important statistics about the data you analyze, so you'll probably want to have at minimum 4 columns but preferably more.

Additionally, although there is a histogram visual you can download from the Power BI marketplace (if you have access to that resource) it's recommended that you use a column chart instead so that you can use a DAX formula to specify the ranges and number of columns (more on this in the detailed instructions).

**Detailed Instructions**

1. Import the census data, customer list, purchase list, and state list into Power BI. You’ll use the “Get Data” button to start the import process. Make sure you select each of the available tables within each Excel file. There should be 7 total tables:
   * Avg Income by State
   * Customer List
   * Incomes by State
   * Industries
   * Product Inventory
   * Purchase List
   * State List
2. Set up the data in Power Query so that all the columns are correctly formatted and structured. Some of the tables will require steps in Power Query to be correctly set up.
3. Set up your table relationships. All tables you import should be tied into the table relationships (if possible). Avoid many-to-many relationships as they’re not needed and will cause issues with cross-filtering.
4. Review the table relationships and confirm if they are properly set up. Most of the tables should follow a star schema and the overall set up should look similar to the sample image:

5.Create a map to visualize the distribution of average household income across the country.

6.Create a new ‘Regression Table’ from the ‘Customer List’ to summarize the average customer sales by state and set up the following calculated columns for your regression formula:

* y = average sales by state (read the note below)
* y2 = Y^2
* x = average income by state (read the note below)
* x2 = X^2
* xy = X\*Y

NOTE: You'll need to use this data to calculate predicted customer incomes, and this can be set up in 1 of 2 ways. Either (1) set up income as x and use x = b-y/-m to predict customer income, or (2) set up sales as x and use y = mx +b to predict customer income. The results will be very similar.

7.Create a scatterplot with a trendline to analyze the relationship between average incomes and average sales.

8.Create the remainder of the regression formula variables as calculated measures: b (y-intercept), m (slope), n, Sum of X, Sum of Y, Sum of XY, Sum of X2, Sum of Y2

9.Create a calculated column in your customer table for your predicted customer incomes.

10.Use DAX formulas to categorize the predicted incomes into ranges and to determine the best product fit for each customer. Present this data with a histogram and other visuals.

11.Perform additional analysis using 1-2 other variables of your choosing. These variables can be part of the provided data set or data that you find from additional sources. Think about variables that may help you better understand the customer demographics and market conditions and that will allow creating a more rigorous analysis.

12.Write a brief 1 page summary of your findings, citing the visuals and tables you created as evidence. The summary should address the 5 *analysis questions* described on the **Instructions Overview** page and provide an overview of your resulting marketing strategy.