Please use “sales.csv”, “variables.csv”, and “fiscal\_gregorian\_calendar.csv” to answer the following questions. We will refer “sales.csv” as the sales data and “variables.csv” as the causal variable data on going.

1. Dataset Information: Load and describe the three datasets: rows, columns, and column names for each.

See code

1. Within the sales data, modify column names. Change "BilledQty" to "unit\_sales," "DocCurrencyGrossAmt" to "dollar\_sales," and "BillingDate" to "ID\_Date."

Note: All columns other than “BilledQty”, “DocCurrencyGrossAmt” and “BillingDate” are hierarchy levels. We can uniquely define a product with all hierarchical columns. Let’s call all the hierarchical levels the “**lowest hierarchy level**”.

See code

1. Merge the "fiscal\_gregorian\_calendar" with the sales data using "ID\_Date" as the key.

Merge the causal variable data with sales data using columns other than “causal\_variable\_1” or “causal\_variable\_2”. Generate a consolidated table referred to as the "master table."

See code. I tried to fix the formatting of the dates (- vs /) and change the types to str so that they could be merged but the calendar data is still nan

1. Aggregate the "master table" to the lowest hierarchical/fiscal month level. Please use summation to aggregate unit sales and gross revenue, maximum to aggregate the holiday flags, and sales-weighted average to aggregate the two causal variables.
2. Examine and describe the sales trend in terms of units and dollars from the year 2014. Provide evidence either within the comments or within the code.

I would plot the sales over time in both units and dollars (and maybe combine them on a single plot using two y-axis) or I would use vertically stacked subplots. I would probably use lines with dots but bars could work too.

1. Compute the average, maximum, and minimum unit sales per year and per quarter/year using "FiscalQuarterNumber" for delineation.
2. Identify the five top-selling Business Orgs in both units and dollars. Verify if they align in both categories.

1. Implement data validation on the "master table." Identify and comment on at least five specific areas checked, offering takeaways and suggestions.

My master table has nan from the merge with the calendar table

I would look through the rest of the table for nan values

I would look through values that should be non-negative to make sure none are negative (unit counts, etc.)

I would plot distributions of key columns to see if there are outlier values that could be checked

Besides single variable distributions, bivariate distributions (or scatter plots) between pairs of columns would be good to look at as well

1. Ascertain the count of distinct "ID\_product" entries within the "master table." Recognize any pairs of similar products, articulate the definition of "similar products," and propose algorithms for identification.
2. Support management in comprehending the relationship between unit sales and the two causal variables.

I would make two scatter plots (variable1 vs sales and variable2 vs sales). If needed, I would also think about making a 3D scatter plot but those can be more difficult to digest.

1. As you may see, there are two causal variables and ten holidays affecting the unit sales. Suggest **at least three** methods for selecting the most influential variables for a predictive model.

Options 1: Calculate the correlation between each of the variables/holidays with the unit sales. Strong correlation would indicate that it is an influential variable

Option 2: Fit a separate linear model for each of the variables/holidays to predict unit sales and see which model performs best. Better performing model would indicate an influential variable

Option 3: Starting with the most influential variable/holiday from Option 2, fit models that contain that variable/holiday plus an additional one. Then incrementally increase the number of variables/holidays in the model, always using the best set and adding to it. Stop when the model stops getting better with more predictors.

1. Perform an A/B test comparing the efficacy of causal variables during the COVID period (starting from March 2020). Visualize and present the results.
2. Outline the methodology for constructing a predictive model and simulation to forecast unit sales. Include commentary on metrics and measurement methods, refraining from providing explicit code.

I would use cross validation with this dataset. So, design several models that use different functions to compute an estimate for sales given a set of inputs. Then for each model, fit it using 90% of the data and test it using the remaining 10%. Do this 10 times for each model, using a different 10% of the data as validation each time. Then for each model, average the test results and use the best model. Then fit that model on all of the available data to obtain the final predictive model.

Metrics to use to determine “best” could include smallest squared error or smallest maximum error (when comparing to the known sales numbers)