

INDIVIDUAL ASSIGNMENT: DATA ANALYTICS AT VK OFFICE SUPPLY LTD

THE PROBLEM

VK Office Supply LTD sell printing-related consumables to companies. Their client base consists mostly of consulting firms, law firms and hedge funds. Located in Chicago, where they have dominated the local market the last few years, VK are considering launching new stores in major cities in the West Coast (San Francisco, Los Angeles) and in the East Coast (New York, Washington). Before making such bold decisions, Allison Jones, VK's CEO and founder, wants to understand better what drives revenue, by investigating some of their recent transactions.

THE TASK

As a close friend of Allison, you are confident that you can help her dissect insights that exist in her company's data. In particular, she would like to know how important each customer sector is, and if this information is relevant towards deciding their expansion decisions.

After a first meeting with the sales manager, you obtained data for some transactions in the Chicago area. This data is in file "VK template.xlsx", in which the revenue is reported in thousands of dollars.

You are tasked with analysing VK's data and communicating your findings in a managerial report. Make explicit any assumptions underlying your answers, interpret your results and justify your answers, conclusions and recommendations. Make sure your report is concise (about five pages) and that any voluminous tables or figures are placed in an Appendix. Also place in an Appendix material that is too technical and may distract the reader from the main focus of the report.

A note with solutions as well as completed spreadsheets will be posted later on Canvas.

Good luck!

1. A FIRST LOOK AT THE DATA (20%)

Consider the data shown in sheet Q1 of the file “VK template.xlsx”.

- Make a graph of revenue against the transaction date. Do you observe any trend or any periodical pattern?
- Analyse and interpret summary statistics for the given data.

2. TIME SERIES FORECASTING (20%)

Allison wants to examine how a simple moving average method performs when it comes to predicting future sales.

- Calculate two moving averages using 2 periods and 8 periods respectively, and compute their mean absolute percentage error (MAPE)¹. Which one seem to perform best in terms of prediction accuracy?
- Calculate a 9-period centered moving average (i.e., when calculating the average for each observation, include 4 observations before and 4 observations after the current observation). Graph this moving average together with the original sales graph. What does the centered moving average depict, and what does it reveal in this particular case?

3. MULTIPLE REGRESSION (20%)

Run a multiple regression using all the data provided (remember to avoid the dummy variable trap).

- What is the proportion of variability explained by the model?
- Which variables seem to exert a significant influence on sales?
- Is there any specific professional sector that appears to be associated with higher spending? If so, which one, and by how much?

4. MAKING PREDICTIONS (20%)

A law firm had made a purchase on 17 March 2021 and has 80 printers. However, the corresponding revenue has not been recorded in the data.

- Build a linear regression model by using the revenue as dependent variable and the transaction date, number of printers and the "Law" dummy as independent variables (you may use sheet Q4 for convenience). What is the proportion variability explained by this model? How does it compare to that of Q3?
- Consider the case of the law firm described above. What is your best estimate for the revenue of that transaction? Would it help us to make a more accurate prediction if we also knew the number of employees of that firm?

5. IMPROVING THE MODEL USING EXTRA DATA (20%)

Allion's partner has found some data on historical revenue values that have occurred in past dates. These are shown in sheet Q5.

- Carry out a linear multiple regression by using all the variables, including the new data. Has the model improved in terms of predictive accuracy?
- Which variables are significant? Can you explain any discrepancies between the significant variables found in this part compared to the significant variables found in Q3?