MGMT8520

Data Analytics and Financial Management II

Assignment 2 – Outlier Detection and Employee Absenteeism

## **PARTICULARS**

* There are two things to determine in this assignment. One is which employees are outliers in absenteeism within our company and two, how much does absenteeism cost our company? In other words, how does this affect the profitability of our company (in other words how much have we paid them to be non-contributing employees in their absence)? Are there any department, position, demographic patterns you can see? You will use employee data to help you measure the costs of absenteeism to Headgames Inc.
* Two apps will help you with the assignment: Excel and Rapidminer. There is also a Word memo to senior management of recommendations based on your analysis.
* You may do this assignment alone, or in teams of two (2) (1+1, 3-1, 4-2, 13-11…)
* There are eight main assignment questions to answer with data in Excel (including some calculations you will have to perform to answer these questions. Create as many worksheet tabs as necessary to answer them.

## **Background**

Once a year the human resources department manually calculates employee absenteeism for department managers across all divisions of Headgames Inc. It is a tedious process. It requires a manual gathering of data, manual calculations in Excel, and manual analysis of the results. On average it takes two weeks, and managers always have questions about the process and suggestions on how to better calculate it “because their department does things a little differently.” But you are feeling that nothing changes, and no one in the company really has a handle on the cost of absenteeism.

You have decided two things. One, you are standardizing the calculation so no more special departmental calculations, only a single company wide measure. Traditionally, outliers are anyone who has been absent more that 1.5 times the interquartile range (Quartile 3 – Quartile 1, added and subtracted from the quartile). But there are other ways to calculate outliers or anomalies, that will give you more options. Two, you want to create your solution as a data model so that next quarter when you want to measure results again, all you have to do it import in the new data and the process will calculate the results in minutes rather than weeks. You decide that Rapidminer will be the tool to help you create the data model to find the outliers, and then Excel will help you calculate the cost of absenteeism and chart the results where necessary.

Starting with the Employee Database in Excel, you import two other files that allow you to build a more comprehensive dataset.

## **CALCULATIONS**

1. Excel: Drag in the Absences worksheet into the Employee Database workbook. Use a VLOOKUP to copy over the unplanned absences by employee into column P.
2. Calculate the daily absentee wage (see note 1) by employee in column Q in the Employee Data worksheet. (Annual salary/No of working days in a year)
3. Calculate the yearly cost of absenteeism by employee in column R in the Employee Data worksheet. (days absent\*daily salary)
4. Bring in the absence code by employee into Column S from the Employee Absenteeism Causes workbook using a VLOOKUP.
5. Use this Employees Database for Outliers workbook to calculate the outliers in Rapidminer.
6. RAPIDMINER:
   1. In the Process pane, drag the Excel Employee Database for Outliers sheet you just built in Excel into the Process pane.
   2. Rapidminer will ask you which worksheet (Employee Data) to use, which columns (A-S), and the rest of questions are defaults. On the Where to Store the Data? Choose your Local DATA Repository as the location.
   3. Go back to Design mode, and drag the just saved file into the Rapidminer process workspace, where you can detect outliers by using the following process operators:
      1. Select Attributes: You only want two attributes, ID and Unplanned Absence. Rapidminer will detect outliers weighted across all columns, but we only want to determine outliers from the one column: unplanned absence.
      2. Set Role: Make the ID attribute an ID role.
      3. Search Detect Outliers (Univariate): Use the single attribute (unplanned absences) and quartiles as the method. Defaults for the remaining parameters.
      4. Search Generate Outlier Flag: Select Outlier\_score as the score column, and defaults for the remaining parameters.
      5. Write the output from the outlier process to a XLSX file using the Write Excel Operator. Name and save the file to a location where you can access the file (you will need this later).
   4. EXCEL: This new Rapidminer created file can now be dragged into the Employee Database workbook. Use the VLOOKUP function in column T to bring the Outlier Flag value for each employee into the Employee Data worksheet. Use ID as the lookup code.
   5. In column U, use VLOOKUP to bring in the position names.
   6. RAPIDMINER: You want to check correlations on the new columns you have created in Excel. In Excel, you can create a correlation matrix across multiple columns, but they must contain numeric data. This means you will have to use the nominal to numerical operator which will convert columns like Education Level, Sex, Location to dummy codes. Once the correlation matrix is created in Rapidminer, copy and paste the results from Rapidminer into Excel and open the results in a new sheet. Here is how the process should look:
7. Retrieve Employees: Drag your Employee database/Employee data worksheet into Rapidminer like you did in Q6 above. This time though you have more columns, A-U.
8. Watch the assignment 2 walkthrough in the class video to see how to create a correlation matrix in RapidMiner.
9. Correlation Matrix: Check both boxes: Normalize weights, squared correlation.
10. Copy the new correlation report into a new tab in the Employees database workbook.
    1. In the Employee Absenteeism Causes worksheet create the Bradford Factor (in column H) based on the following formula (BF=TS2 \* DA) where TS is the total spells in column G, and DA is the days absent in column F. Using VLOOKUP, bring this value into the EMPLOYEE DATA worksheet.
    2. Do a VLOOKUP in column W to bring in the department from the JOB CODES worksheet.

## **QUESTIONS (80 marks)**

Once you have created the worksheet (with the Rapidminer outliers and correlations sheets copied in), you have some questions to answer about the data you see. Answer these in a textbox in the worksheet to which they refer.

1. How does the outlier list in Rapidminer compare to the outliers that you could manually calculate in Excel? Explain. (5 marks)
2. What is the total cost of absenteeism to our company? What is the ratio of absentee cost to annual salaries? (5 marks)
3. How does the absentee/annual salaries ratio compare to the U.S. plant? Based on the data you have what accounts for the difference? Explain. (10 marks)
4. Is there a strong correlation between any of the absentee attributes found using RapidMiner? Explain. (5 marks)
5. Create pivot tables in a single worksheet to help you answer these questions. Which absentee code has the most lost wages cost associated with it? What % of lost wages? Which sex is more prone to injury and by what ratio? Which are the top two (2) positions in cost of absenteeism? Speculate as to why these two positions, in other words, what type of absenteeism is most prevalent in these positions? Write your brief answers next to each pivot table in a text box. (25 marks)
6. In Column X, calculate the Employee Absence Rate (Unplanned absences/Work days). Can you see a pattern between position, Pay Type, and Employee Absence Rate? Speculate about what you see. (5 marks)
7. Create a 1-page memo in Word describing 3 key points you have found in your analysis around the impact of absenteeism on our company, and how you plan to use the Bradford Factor score to help our company measure, understand, and reduce the cost of absenteeism. This may include text, numbers, charts, screen shots, whatever data you feel will help you recommend an approach to absenteeism for the company. (25 marks)

## **NOTES**

1. The company is using 251 standard working days in Canada for 2021. The total working hours for 2021 per employee is 2000. In the US, The working days was 249.
2. The Bradford factor measures an employee’s irregularity of attendance by combining measures of absence frequency and duration and is designed to highlight employees who take many short-term absences from work.

## **SUBMISSION**

The submission for this assignment is one (1) Excel workbook, with all the required worksheets attached and answers included in text boxes and one (1) Word document (question 7).