BIA 3713: Introduction to Business Intelligence

Fall 2021

ASSIGNMENT 2 POINTERS

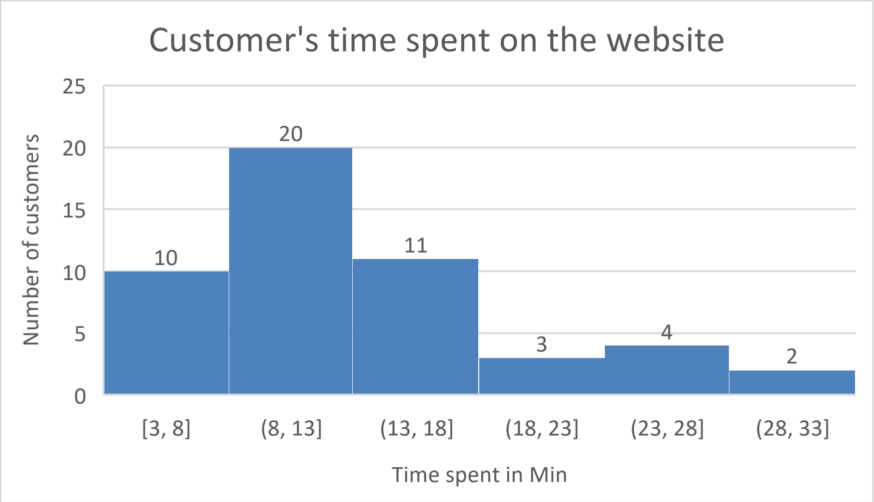
Mini Case

MyHealthcare manufactures and sells blood pressure measurement and control products. Last year the company began selling its products online. Online sales have exceeded the company's expectations, and management is now considering strategies to increase sales even further. To learn more about the online shoppers, a sample of 50 transactions (refer to Data file - ShoppersData.xlsx on Canvas under Assignment 2) were selected from the previous month's sales. Data showing the day of the week each transaction was made, the type of browser used, the time spent on the website, the number of website pages viewed, and the amount spent by each of the 50 shoppers. MyHealthcare would like to understand the shoppers buying pattern in general. Specifically, the company uses the sample data to determine if online shoppers who spend more time and view more pages also spend more money during their visit to the website. The company would also like to investigate the effect that day of the week and the types of browsers have on sales.

Managerial Report

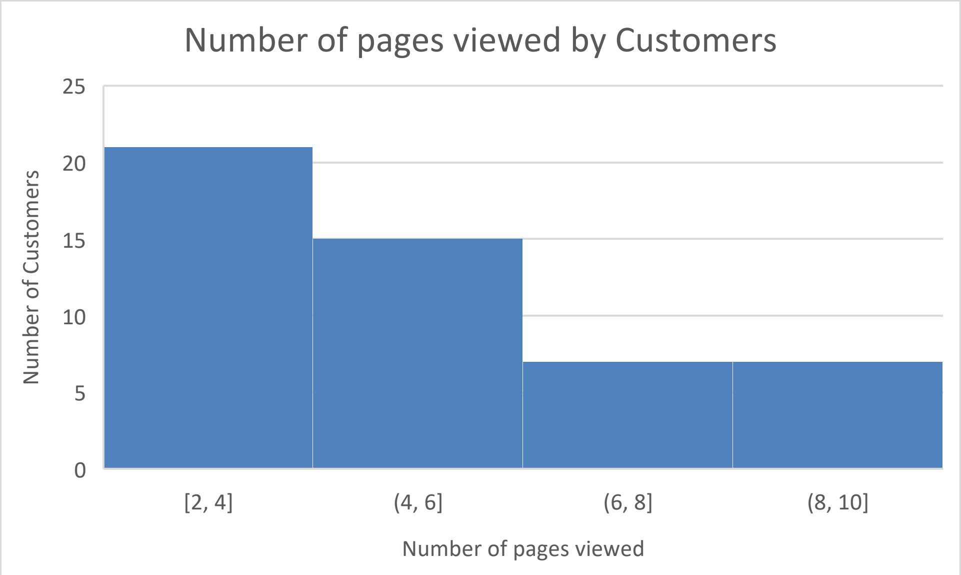
Use the methods of data analytics to learn about the shoppers who visit the MyHealthcare site. Include the following in your report.

a. Appropriate graphical and numerical summaries for the length of time the shopper spends on the website, the number of pages viewed, and the amount spent per transaction. Interpret and discuss what you learn about MyHealthcare's online shoppers from these summaries. (5 points)



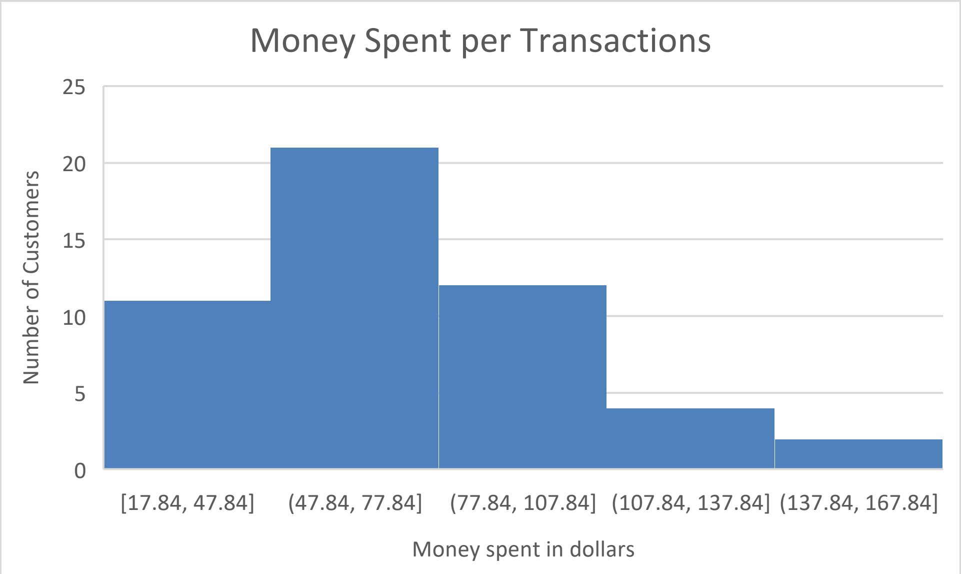
On an average, customers typically spend approximately 13.14 minutes on the MyHealthcare website. However, the time spent varies from just over 3.0 minutes to around 32.9 minutes in length. The distribution of time spent on MyHealthcare's website is approximately bell-shaped and positively skewed with a mean of 13.14 and a standard deviation of 6.52. Positive skewness means the data is skewed to the right with a long tail on the right side of the peak. The fact that the mean (13.14 minutes) is greater than the median (11.65 minutes) implies that the outliers are on the right side. In other words, there are a few customers who spend significantly more time on the website than the majority of customers do.

Note: The next step for you is to use the above facts and suggest potential action items.



On an average, customers typically view approximately 5.32 pages on the MyHealthcare website. However, the number of pages varies from 2 to 10 pages. The distribution of time spent on MyHealthcare's website is positively skewed with a mean of 5.32 and a standard deviation of 2.34. Positive skewness means the data is skewed to the right with a long tail on the right side of the peak. The fact that the mean (5.32 pages) is greater than the median (5 pages) implies that the outliers are on the right side. In other words, there are a few customers who view significantly more pages on the website than the majority of customers do.

Note: The next step for you is to use the above facts and suggest potential action items.



On an average, customers typically spend approximately $69.44 on the MyHealthcare website. However, the amount spent varies from approximately $17.84 to $167.10. The distribution of the amount spent on MyHealthcare's website is positively skewed with a mean of $69.44 and a standard deviation of $32.24. Positive skewness means the data is skewed to the right with a long tail on the right side of the peak. The fact that the mean ($69.44) is greater than the median ($65.22) implies that the outliers are on the right side. In other words, there are a few customers who spend significantly more money on the website than the majority of customers do.

Note: The next step for you is to use the above facts and suggest potential action items.

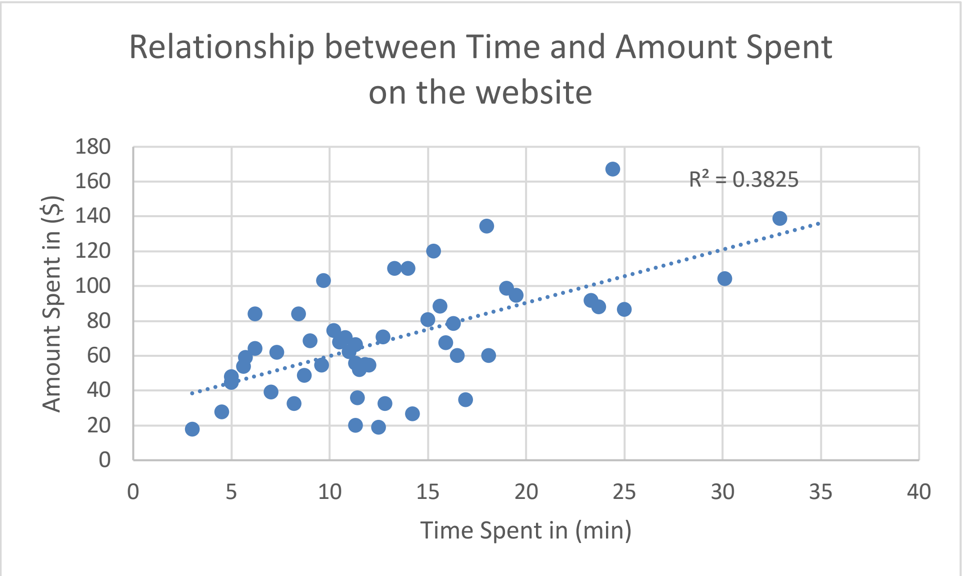
b. Report the frequency of transactions, the total dollars spent, and the mean amount spent for each day of week. What observations can you make about MyHealthcare's business based on the day of the week? Discuss and interpret your results. (5 points)

Day of the Week	**Total amount spent per day**	**Average amount $ spent per day**	**Number of transactions per day**
Sun	235.98	47.20	5
Mon	814.92	90.55	9
Tue	423.85	60.55	7
Wed	413.68	68.95	6
Thu	271.83	54.37	5
Fri	933.17	84.83	11
Sat	378.74	54.11	7
Grand Total	**3,472.17**	**69.44**	**50**

Reviewing the buying patterns of the sample produced a few observations related to the day a sale was made. Friday (frequency 11) was the most common day for sales followed by Monday (frequency 9). The average spending on the MyHealthcare website was highest on Monday ($90.55) and Friday ($84.83), suggesting that customers spend more on each purchase on these two days than any other day. Sunday yields the least average revenue ($47.20). It is also tied with Thursday for least number of sales per day (frequency 5). Overall, the company can benefit from efforts to attract more customers and convert more sales on days like Sunday and Thursday with lower activity.

Note: The next step for you is to use the above facts and suggest potential action items.

d. Use appropriate graphical method and numerical summary statistics to explore the relationship between the time spent on the website and the dollar amount spent. Use the horizontal axis for the time spent on the website. Discuss and interpret your results. (3 points)



Exploring the relationship between the amount of time spent on the MyHealthcare website and the amount of money a shopper spends indicates a moderate average positive linear relationship. The correlation coefficient of 0.618 suggests these two variables are not very strongly related (in a positive linear fashion) to each other. Further research into what keeps shoppers on the website could be an important variable. It would also be worth including data for visitors who did not make a purchase and learning how much time they spent on the website as possible variables contributing toward an incomplete purchase.

Problem 2
Refer to the article: The Data Science Process and answer the following question.

What is involved in processing data for analysis, and why is this step important? Discuss your answer in about 100 words. (3 Points)

The steps involved in processing data for analysis involve the following: Checking for missing (null) values in the data set; checking for duplicate values in the data sets; checking for corrupted values or invalid entries in the data sets. Other steps involve checking for date such as time zone differences and data range. This step, which is basically checking for and correcting irregularities in the data, is important because without it a data scientist will not be able to get accurate insights of the data, which could lead to misleading assumptions or business decisions. It would help ensure that the analysis is an accurate representation of the data.

Problem 3
Refer to the article: Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says and answer the following questions:

Why is a data scientist, at times, described as a "data janitor?" Discuss your answer in about 150 words. (3.5 Points)

Data is not always in the cleanest or most organized formats. It requires a "cleanup" before it can be used for analysis. The data janitor description comes from the data pre-processing, or "data cleaning," part of a data scientist's job. The majority of data scientists spend 70-80% of their time pre-processing data that needs to be cleaned up and formatted in a way that allows for accurate analysis. Checking for proper data types, fixing masses of missing or corrupted values, taking into account geographical variables such as time zone differences or location data, furthers accuracy in resulting analysis. Data scientists must make sense of these large quantities and varieties of data, and that requires data wrangling and cleaning, The data scientist is like a janitor because they both clean up messes. Cleaning data and formatting it for use in analyses makes data scientists not unlike a data janitor.

Briefly describe why startup companies have focused on automating a solution to data preparation. Discuss your answer in about 150 words. (3.5 Points)

Data scientists regard cleaning and organizing data as the least enjoyable part of their work. Startups have focused on automating data wrangling processes to help data scientists because a lot of time is spent in cleaning and organizing data. The majority of data scientists consider data cleaning mundane and a necessary burden. Data scientists' salaries are considerable, reported around $104,000 annual salary in 2015, and their time can be more efficiently spent on analysis rather than data preparation if this step can be automated. Machine learning can possibly begin to replace the manual data pre-processing, and so automating these processes would lessen the work of preparing data and allow data scientist to focus more in data modeling or building algorithms and training sets, and it could shorten project times. Automation could also cut down on the potential for human error when preparing data for analysis.

Overall, the following criteria will be used to grade (credit/no-credit) the assignment:

- Convincing conclusions are drawn and demonstrate an understanding of investigation results and how to apply it.
- Writing style is understandable and organized in explaining investigation results and supporting conclusions.
- Data are explored and analyzed in many different ways.
- Calculations are detailed, accurate, and answers are correct.
- Graphs/pictures are labeled.
- Rigorous approach is used to solve a specific problem.

************************************** NOTE ************************************
Please feel free to consult your instructor by email or phone, if you have questions or need assistance!!