**CIDM/ECON 6308 Exam 1 -Part 2 (100 points)**

**Due 11:59 PM CDT July 26, 2023**

1. **Business Understanding: Describe Analytics Efforts (15 points)**

This case involves multiple domains such as telecom, customer, and churn analytics. Please answer the following two questions to help you better describe how analytics is applied in this case.

* 1. Analytics Domain and Orientation: Which of the following types of analytics is mainly involved in Mark’s task of summarizing some patterns and reporting possible relationship in the data? [3 points]
     + Descriptive Analytics
  2. Analytics Activities: Please indicate whether each of the following statements about analytics activities is true or false by typing T (i.e., True) or F (i.e., False). [12 points and 2 points for each]

 Sending a special retention deal to those customers who are predicted to churn based on the scoring score computed from the data mining model is an example of evidence generation activity. F

 Retrieving about 5,000 customers’ transaction data from MegaTelCo’s repository is an example of evidence selection activity. T

 Obtaining about 5,000 customers’ demographic information from MegaTelCo’s repository is an example of evidence acquisition activity. F

 Generating the probability that each customer is going to turn over from the data mining model is an example of evidence generation activity. T

 Creating a customized service package for high-value customers who are projected to turn over based on the data mining model is an example of evidence emission activity. T

 Producing business rules based on the data mining model and industry standard to determine which customers are our top priority to retain is an example of evidence assimilation activity. T

1. **Data Understanding & Acquisition (30 points)**

SQL Queries: Mark selects the data based on multiple conditions from the data warehouse system using SQL queries. Our class introduces SQL and you have also completed a SQL course on DataCamp, so please answer the following questions. You may refer back to SQL Tutorial or a SQL course on DataCamp.

* 1. Mark wants to select the records in the past three years, i.e., 2020, 2021, and 2022, using a numerical attribute, *year*, from a relational database. His team comes up with ten possible WHERE statements below. Please indicate whether each of them is true or false by typing T (i.e., True) or F (i.e., False) in the front. [10 points and 1 point for each]

|  |  |
| --- | --- |
|  | WHERE year >=2020 AND year <=2022 X |
|  | WHERE year >=2020 OR year <=2022 |
|  | WHERE year BETWEEN 2020 AND 2022 |
|  | WHERE year BETWEEN 2019 AND 2023 X |
|  | WHERE year IN 2020, 2021, 2022 |
|  | WHERE year IN (2020, 2021, 2022) X |
|  | WHERE year = 2020 OR year=2021 OR year=2022 X |
|  | WHERE year = 2020 AND year=2021 AND year=2022 |
|  | WHERE year > 2019 AND year <2023 X |
|  | WHERE year = 2020 OR 2021 OR 2022 X |

* 1. Mark wants to sort the obtained records by the variable *CustomerID* in ascending order. Which of the following statement can be used for this purpose? Choose all that apply [4 points]
     + ORDER BY CustomerID
     + ORDER BY CustomerID ASC X
     + ORDER BY CustomerID DESC
     + GROUP BY CustomerID
     + GROUP BY CustomerID ASC X
     + GROUP BY CustomerID DESC

* 1. Please indicate whether each of the following statements is true or false by typing T (i.e., True) or F (i.e., False) in the front (16 points: 2 points for each question).

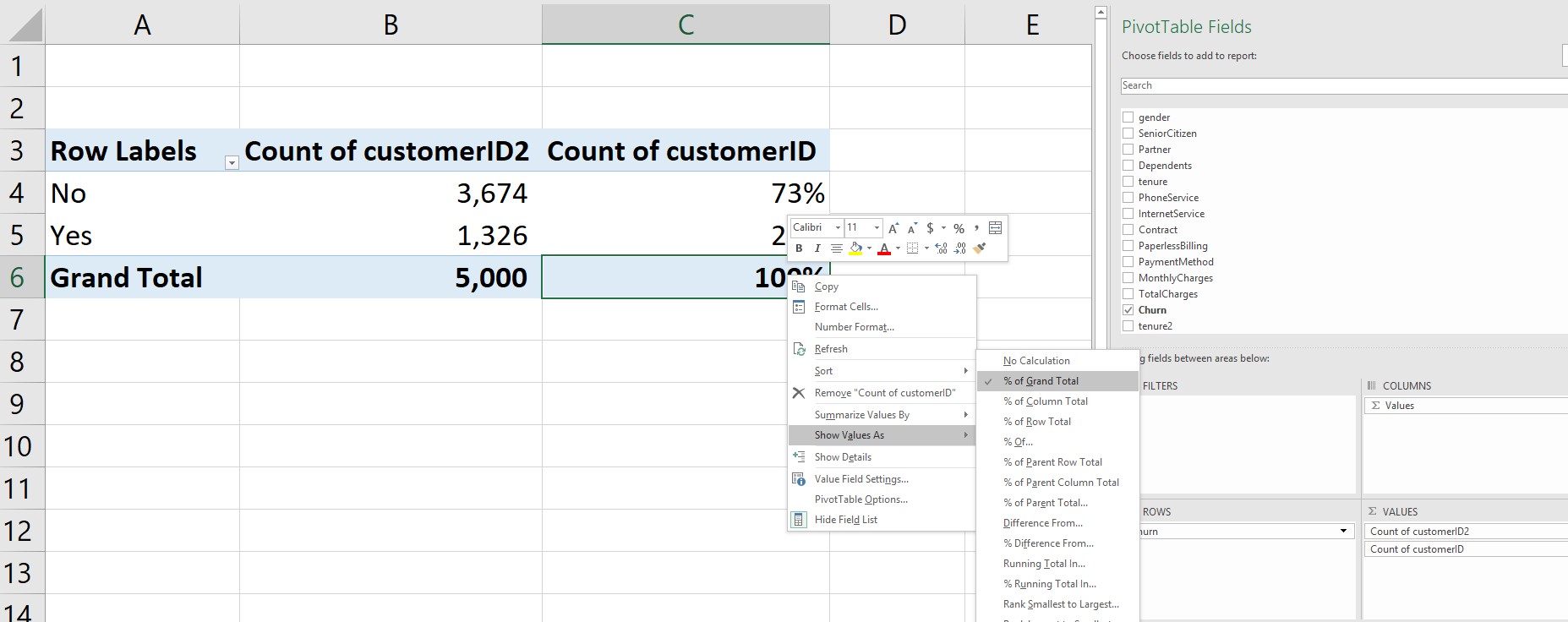
|  |  |
| --- | --- |
|  | The variable *MonthlyCharges* is a continuous numerical variable. T |
|  | The variable *SeniorCitizen* is a discrete numerical variable. T |
|  | The variable *PaymentMethod* is an ordinal variable. T |
|  | The variable *Tenure* is a discrete numerical variable. F |
|  | As a type of incidental data, profile data such as *gender* in this case is relatively static. T |
|  | In this case, the data Mark is using is qualitative. F |
|  | In this case, Mark uses internal data that is collected by ongoing business activities. T |
|  | In order to convert the categorical variable, *PaymentMethod*, to a numerical attribute, Mark should create four dummy variables. **T** |

1. **Data Understanding: Descriptive Analysis of Attributes (20 points)** 
   1. Describe categorical attributes (10 points: 1 point for each cell)

A summary table[[1]](#footnote-1) tallies the values as frequencies and/or percentages for each category of a variable. A summary table helps us see the differences among the categories in a separate column. Please complete the following summary table for all the categorical attributes (a few examples are already provided). You just need to type the percentage (rounded to the second decimal place such as 74.76), and the frequency is not required.

|  |  |  |  |
| --- | --- | --- | --- |
| Attributes | Category or value | Frequency | Percentage (round to the second decimal place such as 12.34%) |
| Churn | No | 3,705 | 74.10% |
| Yes | 1,295 | 25.90% |
| Gender | Female | 2,459 | 49.18% |
| Male | 2,541 | 50.82% |
| SeniorCitizen | No (0) | - | 84.14% |
| Yes (1) | - | 15.86% |
| Partner | No | 2595 | 51.90% |
| Yes | 2405 | 48.10% |
| Dependents | No | 3492 | 69.84% |
| Yes | 1508 | 30.16% |
| PhoneService | No | 479 | 9.58% |
| Yes | 4521 | 90.42% |
| InternetService | DSL | - | 34.06% |
| Fiber optic | - | 43.88% |
| No | - | 22.06% |
| Contract | Month-to-month | - | - |
| One year | - | - |
| Two year | - | - |
| PaperlessBilling | No | - | 40.88% |
| Yes | - | 59.12% |
| PaymentMethod | Bank Transfer | 1101 | 22.02% |
| Credit Card | 1086 | 21.72% |
| Electronic Check | 1686 | 33.72% |
| Mailed Check | 1127 | 22.54% |

Note: A quick way to get both frequency and percentage is to use PivotTable for each categorical attribute (you may refer to the instruction below or in LA1).



3

. For the percentage, please choose to

show values as % of Grand Total.

1

. Drag Churn to ROWS

2

. Drag CustomerID to

∑

VALUES twice: one

for frequency, the other one for Percentage

4

. Once you complete the first attribute, replace

Churn using next attribute

* 1. Describe Numerical Attributes

Please obtain the minimum, maximum, mean, and sample standard deviation for the three numerical attributes, Tenure, MonthlyCharges, and TotalCharges and then complete the following blanks (2 points for each, 10 points in total). Round your answers to the second decimal place such as 11.11 except the ones already provided.

* + - The customers’ tenure ranges from 0 month (i.e., new customer) to 72 months and their average tenure is 32.34 months, with a sample standard deviation of 24.46 months.

* + - The customers’ monthly charge ranges from 18.40 dollars to 118.65 dollars and their average monthly charge is 64.56 dollars, with a sample standard deviation of 30.19 dollars.

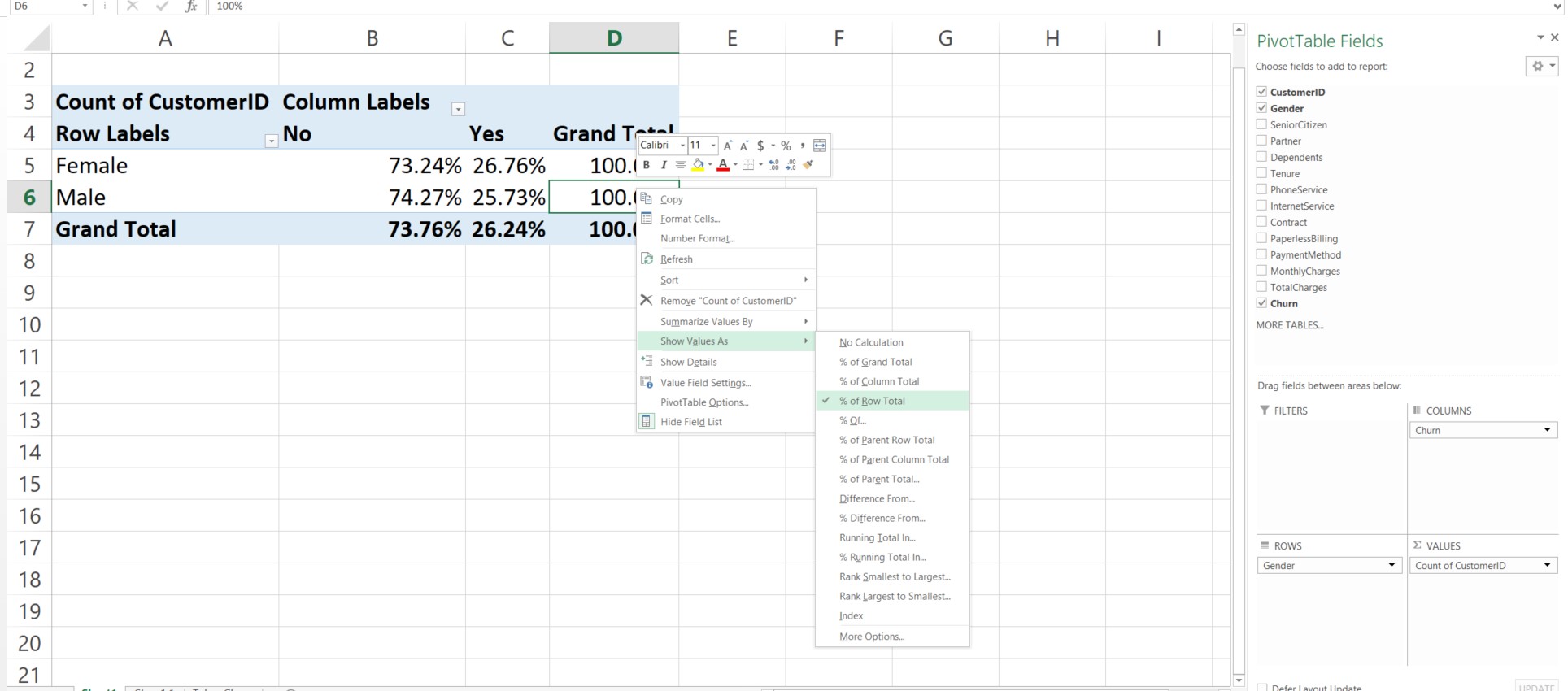
* + - The customers’ total charge ranges from 0.00 dollars to 8,684.80 dollars and their average total charge is 2,279.97 dollars, with a sample standard deviation of 2,270.69 dollars.

Hint: You can either use the formulas mentioned in LA1 (Step 1.4) or the analysis ToolPak in MS Excel. For loading the analysis ToolPak, please refer to [this link;](https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4) please use it to perform data analysis, please check [this link](https://support.microsoft.com/en-us/office/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6) (Data Data Analysis  Descriptive Statistics) to find all summary statistics above.

1. **Data Exploration: Computing Churn Rate (35 points)**

A contingency table[[2]](#footnote-2) cross-tabulates or tallies jointly the value of two or more categorical variables, allowing us to study patterns that may exist between the variables. Tallies can be shown as a frequency, a percentage of the overall total (grand total), a percentage of the row total, or a percentage of the column total. A multi-dimensional contingency table tallies the response of three or more categorical attributes. In Excel, we can construct PivotTable for this purpose. All these efforts are called data discovery, which is defined as the methods enabling us to perform preliminary analysis by manipulating interactive summarizations. Using the contingency table, we can compute churn rate for different groups of customers segmented by each variable and then determine whether or not such variable may be a significant predictor of customer churn before running any data mining models (Note: here, we just explore whether or not such a variable matters, rather than statistically test its significance).

For example, by constructing the following PivotTable, we find that the churn rate of female customers is 26.60%, slightly higher than the churn rate of male customers (25.23%). Because the difference of the two churn rates is quite small, we can conclude that gender may not be a good differentiator for customer churn. In contrast, using the similar way, we find that the churn rate of customers using paperless billing is 32.71% while the churn rate of customers who do not use paperless billing is 16.052%, indicating that the variable paperless billing may be a good differentiator for customer churn.



* 1. Please compute the churn rate of each group of customers and round them the second decimal place, as illustrated in the two examples below (1.5 point for each, 27 points in total).

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Customer group | Churn rate | Compare churn rates to determine whether it is a good differentiator |
| Gender | Female | 26.60% | Female customers are slightly more likely to churn than males, indicating gender is not a |
|  |  |
|  | Male | 25.23% | good differentiator. |
| SeniorCitizen | No (0) | 22.94% | good differentiator. |
| Yes (1) | 41.61% |
| Partner | No | 32.45% |  |
| Yes | 18.84% |
| Dependents | No | 30.73% | good differentiator. |
| Yes | 14.72% |
| PhoneService | No | 23.17% |  |
| Yes | 26.19% |
| InternetService | DSL | 18.26% | good differentiator. |
| Fiber optic | 40.79% |
| No | 8.07% |
| Contract | Month-to-month | 41.54% | good differentiator. |
| One year | 10.62% |
| Two year | 3.04% |
| PaperlessBilling | No | 16.05% | good differentiator. |
|  |  |
|  | Yes | 32.71% | good differentiator. |
| PaymentMethod | Bank Transfer | 15.99% | good differentiator. |
| Credit Card | 14.46% |
| Electronic Check | 44.54% |
| Mailed Check | 18.72% |

* 1. In data mining, one of our primary goals is to identify informative attributes that can effectively reduce uncertainty in prediction, providing valuable insights (for further details on informative attributes, refer to DSB Chapter 3). Here, we will use a simple way to identify informative attributes, which we name as differentiators.

During data exploration, it becomes imperative to pinpoint these differentiators from the numerous variables at hand. To achieve this, we adopt a straightforward rule: when a variable divides the customer base into multiple groups, we consider it a good differentiator if, within these groups, there is at least one group of customers with a churn rate at least twice as high as another group's churn rate. To better understand this concept, consider the two examples provided in the last column of the table above. These examples serve to illustrate whether a variable qualifies as a good differentiator or not. By applying this criterion, we can effectively identify the most influential attributes that significantly impact customer churn rates, aiding us in making more informed predictions and decisions. Using the same rule, which of the seven other categorical variables can be considered as a good differentiator? Choose all that apply. (8 points)

----------------------------------------------------------------------------------------------------------------------------------

This is the end of Exam 1 Part 2

1. You are supposed to learn this in your previous statistics course or the leveling course, CIDM 6300 before taking this course. Here is a quick review to help you refresh your memory and get ready for the second module of this course: Technical Application. [↑](#footnote-ref-1)
2. You are supposed to learn this in your previous statistics course or the leveling course, CIDM 6300 before taking this course. Here is a quick review to help you refresh your memory and get ready for more advanced analytics techniques. [↑](#footnote-ref-2)