Briana Churchill

D207

Instructor Gagner

14 July 2023

Performance Assessment: Exploratory Data Analysis (OEM2)

A. Describe a real-world organizational situation

1. Research Question

The research question I'd like to ask given the dataset is "Are customers that rate themselves technically inclined more or less likely to churn?" Answering this question involves analyzing two non-related columns within the dataset. The analysis completed will be evaluated in relation to the null hypothesis; which is H0: Technological inclination (techie) and churn are independent. The alternative hypothesis is H1: Technological inclination (techie) and churn are not independent.

2. Benefit to stakeholders

Answering this question could allow the business to determine if there is a relationship between customers churning for reasons related to their self-reported technological capabilities. This review of information may inform the business of potential changes they could make to customer packages, marketing, FAQ sections, etc. to better maintain current customers regardless of their technological capabilities. Considering that the cost to obtain customers is far greater than maintaining them, stakeholders could benefit from avoiding the excessive costs to acquire new customers. (D207 data dictionary)

3. Data relevant to answering the research question

This dataset consists of 50 columns, which all contain separate variables relating to 10,000 individual customers. The variables applicable to the research question described above are:

- Churn
- Techie

B. Describe the data analysis

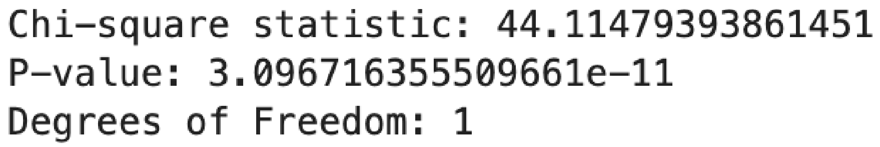
Both of the variables applicable to answering the question are considered nominal data. Testing for independence of these nominal variables utilizing the chi-square test is most applicable. Using Python, I reviewed the clean data CSV provided. After reviewing the dataset, there were necessary changes noted for it to be considered completely clean. I started by renaming certain columns, and data types. I also corrected outliers in “Zip” and cleaned up the time zones again as well. These changes were necessary to allow proper utilization of the data and correct punctuation. Once the data was clean, I was to perform a chi-square test. Creating a contingency table and performing the chi-square test to review the proportions of the two columns was the first step. Once the test was performed the chi-square statistic, p-value and degrees of freedom were available to analyze and come to a conclusion of dependency between the variables.

1. Write code to run the analysis chosen

The executable script file is attached to this submission, for review of functions used.

2. Provide the output for the chi-square test analysis

As seen below, I have provided a screengrab of the results given from executing the chi-square test. The output values are the chi-square statistic, p-value and the degree of freedom.



```
Chi-square statistic: 44.11479393861451
P-value: 3.096716355509661e-11
Degrees of Freedom: 1
```

3. Justify why you chose this analysis technique

Reviewing the data within the two variables and determining whether or not the two variables have a relationship means that a chi-square test is most appropriate. “[A] chi-square test is used to determine if there is evidence that the two variables are not independent” (2023, Sewell) Whether or not the results of the test confirm the independence of each variable from the other, would allow insight to answer the question.

C. Use univariate statistics to identify the distribution of two continuous variables and two categorical variables

Identifying the distribution of two continuous variables using univariate statistics was done with columns “Outages_sec_per_week” and “Tenure.” A histogram was used for both of these variables in this step. For the identification of the distribution of two categorical variables using univariate statistics, columns “Area” and “Children” were used. A pie chart was used for “Area” and a bar graph was used for “Children.”

For outages in seconds per week, the histogram demonstrates a unimodal distribution with the peak being between 10 and 11 seconds per week. This graph also indicates that outages range between 0 and 20 seconds per week, with 10-11 seconds being the most frequent duration.

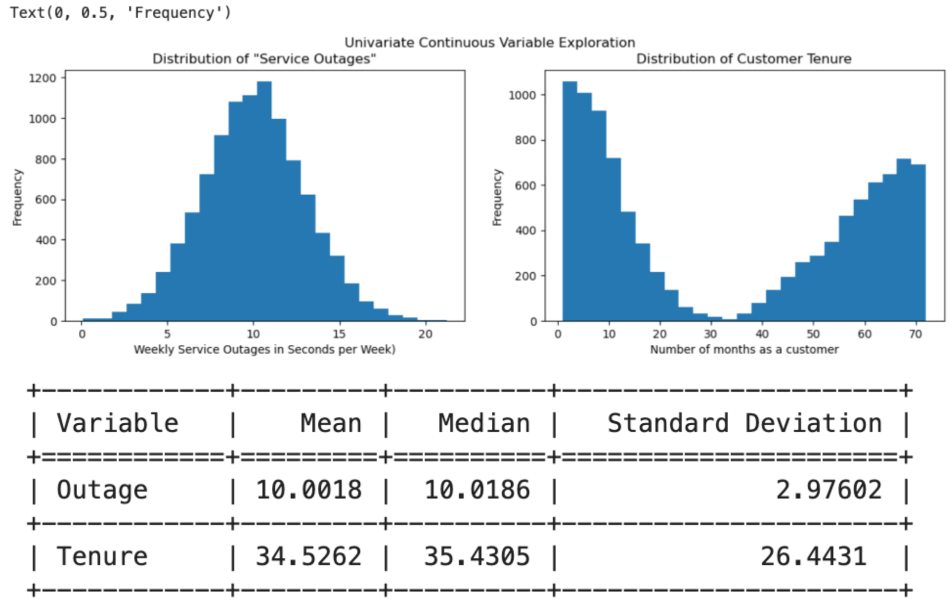
For the tenure variable, an inverted bell curve is seen. There is a higher frequency of newer customers compared to customers with longer tenure. A significant decrease in frequency can be seen starting at around 20 months, with the lowest point of frequency being at ~ 35 months. Frequency picks back up after 40 months increasing up to around 67-68 months with a very slight drop before 70 months.

The percentage of areas lived in by customers (ie: rural, suburban, urban) is almost evenly distributed. There is a 33.3% distribution for each of customers in urban and rural areas. For suburban, the distribution is 33.5%.

Distribution of children per customer as seen in the box plot indicates a skewness to the left with a decreasing trend in the central tendency of the data. The highest frequency is that of customers with zero children, with customers having one child being close in frequency. Beyond one child, the number of customers with 2+ children decreased. There is the slightest increase at 8 children per customer, but the decreasing trend continues with more children.

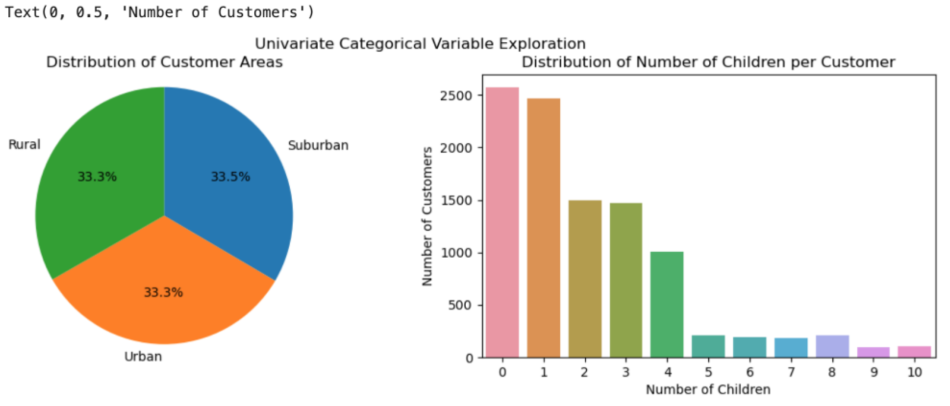
1. **Represent your findings in Part C, visually as part of your submission.**

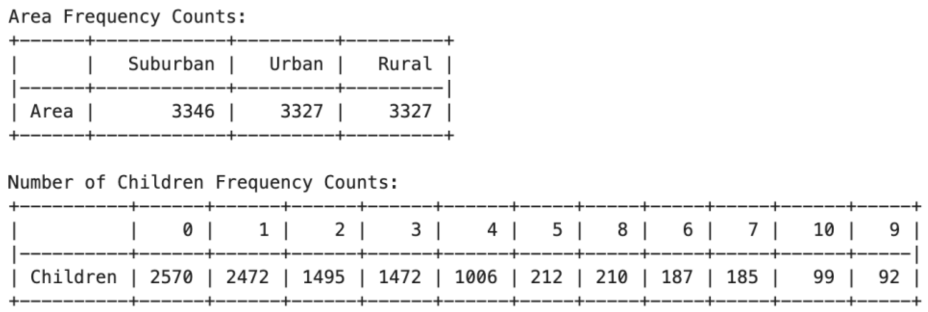
The applicable graphs used to determine the distributions of the variables described above, and their respective tables, can be seen in the images below.


Univariate Continuous Variable Exploration
Distribution of "Service Outages"
Frequency
0
200
400
600
800
1000
1200
0
5
10
15
20
Weekly Service Outages in Seconds per Week)
Distribution of Customer Tenure
Frequency
0
200
400
600
800
1000
0
10
20
30
40
50
60
70
Number of months as a customer

```
Text(0, 0.5, 'Frequency')
```

```
+------------+----------+-----------+----------------------+
| Variable   |     Mean |    Median |   Standard Deviation |
+============+==========+===========+======================+
| Outage     | 10.0018 |  10.0186 |              2.97602 |
+------------+----------+-----------+----------------------+
| Tenure     | 34.5262 |  35.4305 |             26.4431  |
+------------+----------+-----------+----------------------+
```


```
Text(0, 0.5, 'Number of Customers')
```


```
Area Frequency Counts:
+------+------------+---------+---------+
|      |   Suburban |   Urban |   Rural |
|------+------------+---------+---------|
| Area |       3346 |    3327 |    3327 |
+------+------------+---------+---------+

Number of Children Frequency Counts:
+----------+------+------+------+------+------+-----+-----+-----+-----+------+-----+
|          |    0 |    1 |    2 |    3 |    4 |   5 |   8 |   6 |   7 |   10 |   9 |
|----------+------+------+------+------+------+-----+-----+-----+-----+------+-----|
| Children | 2570 | 2472 | 1495 | 1472 | 1006 | 212 | 210 | 187 | 185 |   99 |  92 |
+----------+------+------+------+------+------+-----+-----+-----+-----+------+-----+
```

D. Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data.

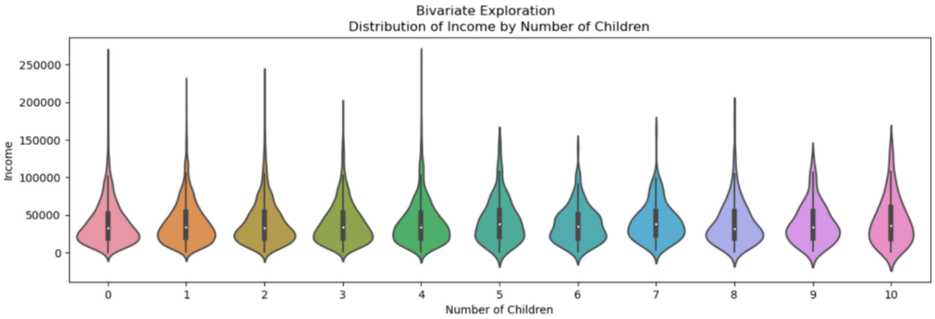
Identifying the distribution of one continuous variable and one categorical variable using bivariate statistics was done with columns “Income” (continuous) and “Children.” (categorical) As well as with columns “Tenure” (continuous) and “Area” (categorical). A violin plot was used for the first analysis, and a scatterplot for the second.

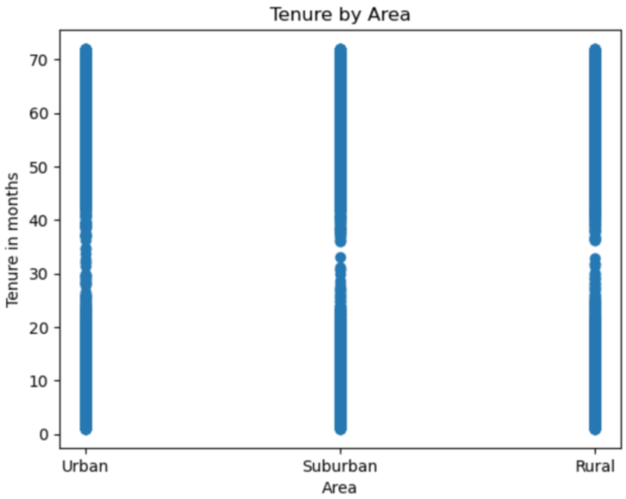
The distribution of income by children within the violin plot shows a similar median income value regardless of the number of children a customer has. The shape of each violin is also very similar to one another. This shape indicates the density of the data points in each applicable violin, which shows that there are many similar income values for each number of children.

Viewing the scatter plot to identify the distribution of tenure by area shows that there are similar patterns of tenure by area. With urban areas having more customers having 30 to 35 months of tenure in comparison to rural and suburban areas.

2. Represent your findings in Part D, visually as part of your submission.

The applicable graphs used to determine the distributions of the variables described above, and their respective tables, can be seen in the images below.





```
Income Summary Statistics:
+--------+---------+---------+---------+--------+---------+---------+---------+---------+
|        |   count |    mean |     std |    min |     25% |     50% |     75% |     max |
|--------+---------+---------+---------+--------+---------+---------+---------+---------|
| Income |   10000 | 39806.9 | 28199.9 | 348.67 | 19224.7 | 33170.6 | 53246.2 |  258901 |
+--------+---------+---------+---------+--------+---------+---------+---------+---------+
Children Counts:
+----+----------------------+---------+
|    |   Number of Children |   Count |
|----+----------------------+---------|
|  0 |                    0 |    2570 |
|  1 |                    1 |    2472 |
|  2 |                    2 |    1495 |
|  3 |                    3 |    1472 |
|  4 |                    4 |    1006 |
|  5 |                    5 |     212 |
|  6 |                    8 |     210 |
|  7 |                    6 |     187 |
|  8 |                    7 |     185 |
|  9 |                   10 |      99 |
| 10 |                    9 |      92 |
+----+----------------------+---------+

Tenure Summary Statistics:
+--------+---------+---------+---------+---------+---------+---------+---------+---------+
|        |   count |    mean |     std |     min |     25% |     50% |     75% |     max |
|--------+---------+---------+---------+---------+---------+---------+---------+---------|
| Tenure |   10000 | 34.5262 | 26.4431 | 1.00026 | 7.91769 | 35.4305 | 61.4798 | 71.9993 |
+--------+---------+---------+---------+---------+---------+---------+---------+---------+
Area counts:
+----+----------+---------+
|    | Area     |   Count |
|----+----------+---------|
|  0 | Suburban |    3346 |
|  1 | Urban    |    3327 |
|  2 | Rural    |    3327 |
+----+----------+---------+
```

E. Summarize the implications of the data analysis

1. Discuss the results of the hypothesis test

The chi-square statistic is a measure of the difference between what frequencies are observed in the contingency table compared to what the expected frequencies would be. Because the chi-square statistic is a large value, there is an indication that there is greater deviation from the expected frequencies. This value suggests that there may potentially be a relationship between the variables.

The next value in the image below is the p-value. "The P stands for probability and measures how likely it is that any observed difference between groups is due to chance." (2014, ABU) The P-value is close to zero which leads to rejection of the null hypothesis. Rejecting the null hypothesis suggests that there is an associated between churn and techie, and that it is likely not due to chance.

Degrees of freedom for this analysis is a value of 1. This means that there is one independent variable needed to assess the relationship between the variables.

Based on these results, there seems to be a signification relationship between techie and churn. The null hypothesis was rejected, the variables are not independent of each other, and there is evidence that suggests the variables are related.

2. Discuss the limitations of your data analysis

Limiting the analyses to only assessing the relationship of two variables may have resulted in the null hypothesis being rejected incorrectly. Although in this analysis, there seems to be a relationship between level of technological inclination and churn, there may still be other variables that could explain the churn as well. In addition, this is a limited dataset of only 10,000 customers within a 6-year time period. A larger data set with more information to review in this analysis would allow a more detailed answer to the question at hand.

3. Recommend a course of action

Given the limited amount of data applicable to the hypothesis testing completed, I would recommend repeating the chi-square analyses, or potentially an ANOVA test, with more variables to assess other potential relationships regarding churn. Upon completion of another analyses, a new recommended course of action would be given referencing the updated results.

F. Provide a Panopto video

The panopto videos may be found in the corresponding assignments folder for D207.

G. Reference web sources

Sewell, W. (2023) Episode 6. Western Governors University.

<https://wgu.webex.com/recordingservice/sites/wgu/recording/8ebcbb20a09b1039bf7f0050568116ea/playback >

Sewell, W. (2023) Episode 5. Western Governors University.

<https://wgu.webex.com/recordingservice/sites/wgu/recording/57b9a2e55ba9103ab7f70050568fa8a9/playback

H. Acknowledge References

Ahmadu Bello University, Zaria, Nigeria. (2014). Understanding the p-value: "It's the difference between possible and probable." The American Statistician, 68(1), 9-13.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111019/#:~:text=The%20P%20value%20is%20defined,groups%20is%20due%20to%20chance>