D207 Exploratory Data Analysis, Performance Assessment

Alexa R. Fisher

Western Governors University

Degree: M.S. Data Analytics

Table of Contents

Part I. Research Question for Data Analysis 3
- A1. Research Question for Analysis 3
- A2. Benefit From Analysis 3
- A3. Data Identification 4

Part II. Data Analysis: Technique 5
- B1. Data Analysis Code 5
- B2. Output Results 6
- B3. Justification for Analysis Technique 7

Part III. Data Analysis: Statistics 9
- C. Identification of Distribution via Univariate Statistics 9
- C1. Visualization of Distribution via Univariate Statistics 10
- D. Identification of Distribution via Bivariate Statistics 11
- D1. Visualization of Distribution via Bivariate Statistics 11

Part IV. Data Analysis: Outcome 13
- E1. Discussion of Findings 13
- E2. Limitations 14
- E3. Recommendation of Action 14

Part V. Supporting Documents 15
- F. Panopto Video 15
- G. Third-Party Web Sources 15

H. References 16

Part I. Research Question for Data Analysis

A1. Research Question for Analysis

The research question selected for this thesis was, “Is there an association between the contract terms and the customers who left within the last month?”. This provided an analysis of the “Churn” data, which was represented by the customers who left and the terms of the contract they had with the telecommunication company. To solve this hypothesis, the “Churn” variable and the “Contract” variable were analyzed using the Chi-square Test of Independence also known as Pearson’s test (Chapman, n.d.). This Chi-square Test was used to see if there was any association between the two categorical variables.

A2. Benefit From Analysis

The telecommunication company stakeholders could gain insight from this analysis. The assessment could notate if there was any correlation between the contract term options and the clients who stopped using their product or service within the last month. This could provide guidance on if the contract terms were ideal for retaining clients. According to Forbes’ being more engaging with active clients and looking inward to develop plans to match client needs can help with retaining repeat customers (Expert Panel, Forbes Agency Council, 2019). Reviewing the correlation between the terms and retention could provide marketing strategies to clients that have specific terms, such as month-to-month, annually, and biennially. The contract terms could pinpoint what customers were looking for in connection with flexibility. This would gear marketing down the correct path of successful retention efforts.

A3. Data Identification

The telecommunication dataset was utilized in this research analysis. The dataset had 10,000 rows and 30 columns once the variables with high cardinality were dropped. High cardinalities were deemed as variables that had more than five unique values. The following qualitative or categorical variables were identified in this cleaned dataset: “Area”, “Marital”, “Gender”, “Churn”, “Techie”, “Contract”, “Port_modem”, “Tablet”, “InternetService”, “Phone”, “Multiple”, “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “StreamingTV”, “StreamingMovies”, “PaperlessBilling”, and “PaymentMethod”. The quantitative or numerical variables remaining in the cleaned dataset were “Population”, “Children”, “Age”, “Income”, “Outage_sec_perweek”, “Email”, “Contacts”, “Yearly_equip_failure”, “Tenure”, “MonthlyCharge”, and “Bandwidth_GB_Year”.

The dataset had a couple of variables relevant to the question “Is there an association between the contract terms and the customers who left within the last month?”. The variables that were identified as significant were the “Churn” data and the “Contract” data. The “Churn” data was whether or not a customer discontinued a service within the last month as a “yes” or “no” value. This variable had two levels of cardinality. The second variable would be the “Contract” data. This was a variable defining the contract terms the customer had. The options for this were month-to-month, one year, or two years. The cardinality for this variable was three levels. Both of these variables were nominal categorical variables. As these were identified as categorical, the statistical model utilized for analysis was the Chi-Square Test for Independence. This test provided insight into whether the two categorical variables are related to each other.

Part II. Data Analysis: Technique

B1. Data Analysis Code

Please see below for a copy of my code that analyzed the dataset via the Chi-Square test of Independence.

```
#Chi-square test for independence. If the contract terms are associated with the churn status.
#H0: The contract terms have no impact on churn, which is if a customer discontinued service within the last month. The two variables are independent.
#Ha: The contract terms have an impact on churn, which is if a customer discontinued service within the last month. The two variables are dependent.

data['Contract'].value_counts()
data['Churn'].value_counts()
observed_all= pd.crosstab(data['Churn'], data['Contract'], margins=True)
print(observed_all)
value= np.array([observed_all.iloc[0][0:3].values,
         observed_all.iloc[1][0:3].values])
chistat, p, dof, expected = chi2_contingency(value)
print("Chi-Square statistic is: " + str(chistat))
print("Degree of Freedom is: " + str(dof))
print("Expected is: " + str(expected))
#chi2 critical value
prob = 0.95
critical_val = chi2.ppf(q=0.95, df=2)
```

```
print('probability: %.3f, critical: %.3f, chi-squared stat: %.3f' % (prob, critical_val, chistat))
if abs(chistat)>= critical_val:
   print("Reject the Null hypothesis(H0). The two variables are dependent.")
else:
   print("Fail to reject the Null hypothesis(H0). The two variables are independent.")
#alpha = 0.05, critical value at 0.05 and dof = 2 is 5.991
alpha = 1 - prob
print('significance: %.3f, p: %.3f' % (alpha, p))
#alpha = 0.05, critical value at 0.05 and dof = 2 is 5.991
if p <=  alpha:
   print("Reject the Null hypothesis(H0). The two variables are dependent.")
else:
   print("Fail to reject the Null hypothesis(H0). The two variables are independent.")
expected, observed, stats = pingouin.chi2_independence(data=data, x = 'Contract', y= 'Churn')
print(stats[stats['test'] == 'pearson'])
```

B2. Output Results

Please see below for a copy of the results from my code that analyzed the dataset via the Chi-Square test of Independence.

```
Month-to-month    5456
Two Year          2442
One year          2102
Name: Contract, dtype: int64

No     7350
```

```
Yes    2650
Name: Churn, dtype: int64

Contract  Month-to-month  One year  Two Year    All
Churn
No                  3422      1795      2133   7350
Yes                 2034       307       309   2650
All                 5456      2102      2442  10000

Chi-Square statistic is: 718.5915805949758
Degree of Freedom is: 2
Expected is: [[4010.16 1544.97 1794.87]
 [1445.84  557.03  647.13]]

probability: 0.950, critical: 5.991, chi-squared stat: 718.592

Reject the Null hypothesis(H0). The variables are dependent.

significance: 0.050, p: 0.000

Reject the Null hypothesis(H0). The variables are dependent.

test   lambda        chi2  dof           pval    cramer  power
pearson     1.0  718.591581  2.0  9.116348e-157  0.268066    1.0
```

B3. Justification for Analysis Technique

The analysis technique selected was based on the distinctive characteristics of the dataset, telecommunication. The telecommunication dataset had insightful information regarding the customers' personal information, services chosen, payment options, and service failures as well as their overall satisfaction with any interactions with their service provider. After reviewing the dataset, the two categorical variables selected were the "Churn" data and the "Contract" data to provide further analysis and insight into the research question posed.

In the Datacamp resource, the video noted that the best technique to analyze two nominal categorical variables is the Chi-Squared Test for Independence (Chapman, n.d.). This test was used to determine if there was a significant association between two categorical variables (Bobbitt, 2020). Before utilizing this test, the null hypothesis needed to be determined. The null hypothesis in this case was the contract terms and churn data have no association between them. This indicates that the variables were independent. The alternative hypothesis was the specified variables were dependent on each other.

Other noted information needed to perform the analysis were the following: the degree of freedom, standard significance level, and the chi-square distribution table (Chapman, n.d.). These were needed along with the hypothesis to provide understanding. The degree of freedom was the number of rows minus one multiplied by the number of columns minus one (Chapman, n.d.). In this case, there were two rows and three columns, leaving a degree of freedom of two. The standard significance level or alpha utilized was 0.05 or 5% (Chapman, n.d.). In conjunction with the significance level, we used the degrees of freedom and the chi-square distribution table to find our critical value. The chi-square distribution table values could be calculated in Python using the SciPy.stats library chi2 package. This package used the significance level and the degree of freedom to calculate the critical value. The SciPy.stats.chi2.ppf() function passed the q value, which is the significance level, and the df value, which is the degree of freedom to return the critical value. (Bobbitt, 2020). The critical value for these specific variables was 5.991 based off of this calculation and confirmed with the chi-square distribution table. This means that the chi-square statistic needs to be greater than 5.991 for us to reject the null hypothesis.

The second Python library utilized was Pingouin to provide similar data and visualizations to the chi-squared test in SciPy (Bobbitt, 2020). The chi2 package in Pingouin provided the chi-squared test summaries of various chi-squared test theories.

Pearson’s test is the only one utilized as it is known as the Chi-Squared test for Independence (Lewinson, 2020). Both of these packages could be used interchangeably to provide insight into the hypothesis.

Part III. Data Analysis: Statistics

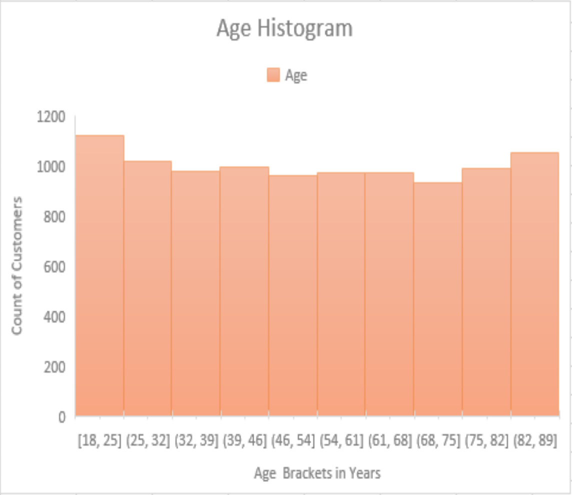
C. Identification of Distribution via Univariate Statistics

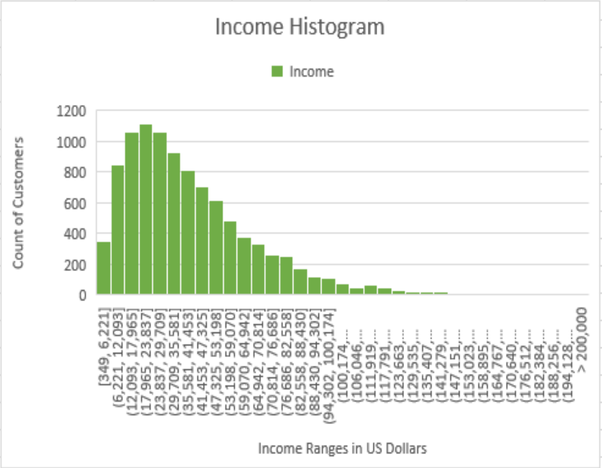
In the telecommunication dataset, there were various continuous and categorical variables. A continuous variable is data that can take any value within an interval (Bruce, Bruce, & Gedeck, 2020). The two continuous variables selected to identify their univariate distributions were “Age” and “Income”. The “Age” variable had a uniform distribution type. A uniform distribution type is described as a distribution that has an equal chance of occurrence (Matsui, n.d.). This distribution is usually visualized in a rectangle shape on a histogram due to the variable having the same mean and median. This would be a symmetric distribution. The “Age” variable has an equal occurrence across numerous clients. In contrast to the previous variable, the “Income” variable had a right-skewed distribution type. A right-skewed distribution is said to be positively skewed and asymmetrical. This implies that the mean is usually to the right of the median.

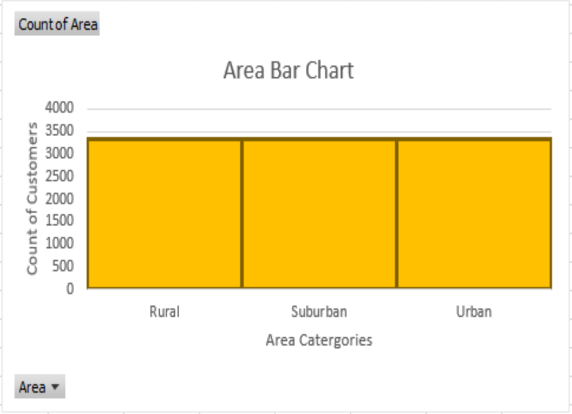
The two categorical variables selected were “Area” and “InternetService”. A categorical variable is data that can only take a specific set of values showing in a set of categories (Bruce, Bruce, & Gedeck, 2020). The “Area” variable had a uniform distribution type. The “InternetService” variable had a positive, right-skewed distribution type. Positive skewed distribution types indicated that the mode, mean, and median were all positive with more of the data occurring on one side of the scale with a long tail of fewer instances on the right side (Matsui, n.d.).

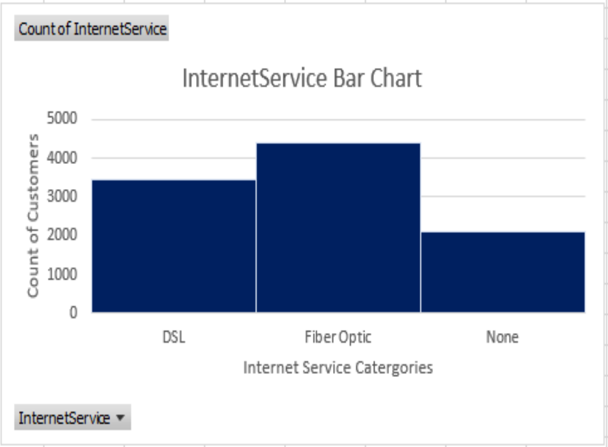
C1. Visualization of Distribution via Univariate Statistics

Please see below for my visual univariate analysis findings for the two continuous variables and two categorical variables. The two continuous variables of "Age" and "Income" were charted using Excel to show a histogram. The histogram provided an excellent view of the specific distributions found by each of these variables. The two categorical variables of " Area" and "Internet Service" were also charted via Excel utilizing the bar chart representation.









D. Identification of Distribution via Bivariate Statistics

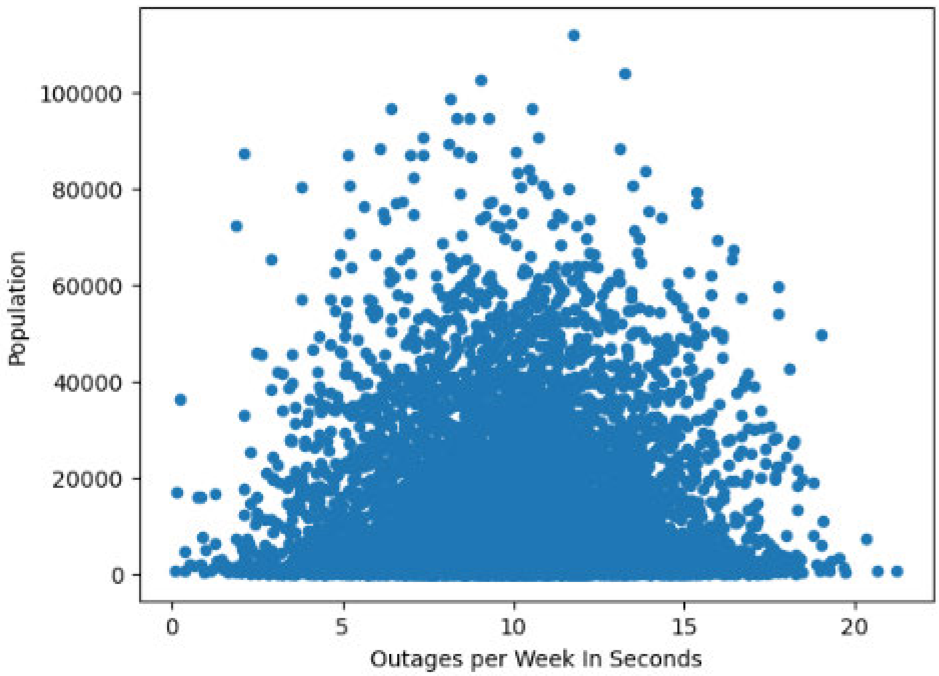
There were various variables within the given dataset that could identify the distribution by bivariate statistics. The following two continuous variables selected were "Population" and "Outage_sec_perweek". This was shown via utilizing a scatterplot. The scatterplot noted the outage seconds per week against the population. It could be visualized that this data resulted in a normal distribution type. The normal distribution type is data that correlates with the mean of the dataset. The data was more frequent near the mean with less frequency the farther it got from the mean (Matsui, n.d.). It could be represented as a symmetrical bell-like shape.

The two categorical variables selected to identify for bivariate analysis were "Area" and "InternetService". The bivariate statistical model utilized with these two categorical variables was a combination of a cross-tabulation table and a bar chart for visualization. The Pandas' crosstab function was used to place the two categorical variables into a contingency table (Matsui, n.d.). Once the table was populated the data was visualized as a bar chart. The bar chart results showed that in each specific area there was a right-skewed distribution type. The results read from the contingency table noted a 44% selection rate for Fiber Optic regardless of area type noted.

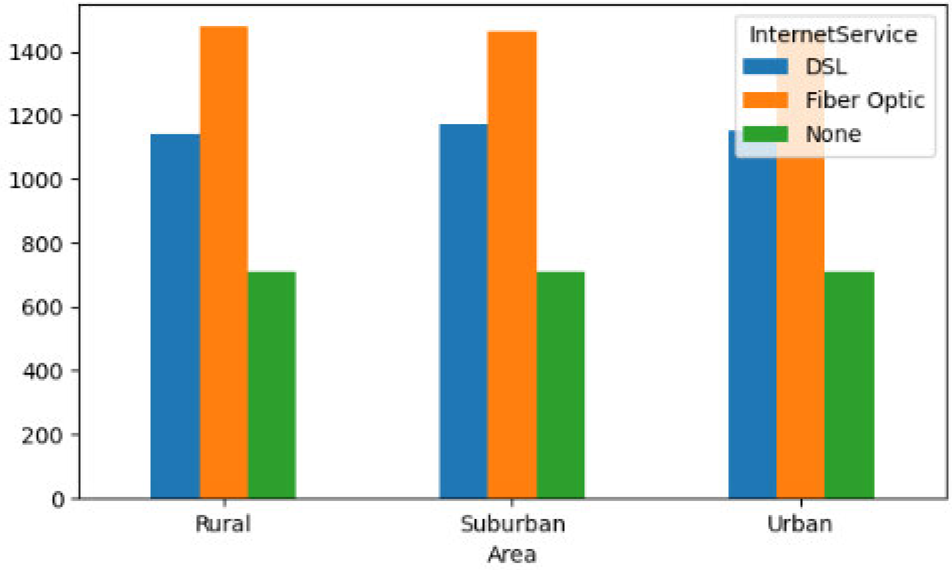
D1. Visualization of Distribution via Bivariate Statistics

Please see below for my visual bivariate analysis findings for the two continuous variables and two categorical variables. The two continuous variables of "Population" and "Outage_sec_perweek" were charted as a scatterplot. The scatterplot provided a view of the two variables' normal distribution type. The two categorical variables of

“Area” and “Internet Service” were analyzed via a cross-tabulation contingency table and the results were further plotted as a bar chart for visualization.



```
InternetService     DSL  Fiber Optic  None
                    Area
Rural              1142         1477   708
Suburban           1170         1465   711
Urban              1151         1466   710
```



Part IV. Data Analysis: Outcome

E1. Discussion of Findings

The Chi-Test for Independence was utilized to determine if there was a significant association between two categorical variables in the given setting. The hypothesis for the test was the “Churn” and “Contract” variables were not associated and found to be independent. The alternate hypothesis was these variables were associated and dependent on each other. As these variables were categorical, the contingency table was created and the chi-square statistic and degree of freedom were discovered. The chi-square statistic for the two variables of “Churn” and “Contract” was 718.59. The degree of freedom was 2. The observed and expected values were noted in the following charts.

Contract vs Churn Contingency Observed Table				
	Month to Month	**One Year**	**Two Year**	**Total**
No	3422	1795	2133	7350
Yes	2034	307	309	2650
Total	5456	2102	2442	10000

Contract vs Churn Contingency Expected Table				
	Month to Month	**One Year**	**Two Year**	**Total**
No	4010.16	1544.97	1794.87	7350
Yes	1445.84	557.03	647.13	2650
Total	5456	2102	2442	10000

The charts visualized that 26.5% of customers in the dataset discontinued service within the last month. There was a variance on the observed results and the expected results. Each observed value was less than the expected values across each cardinality level. The bivariate analysis resulted in a right-skewed distribution type.

After calculating the critical value via the chi2.ppf function by passing the significance percentage and the degree of freedom, the chi-square statistic was compared to the critical value (Chapman, n.d.). This statistic was greater than the critical value for the specific degree of freedom. The results of the completed hypothesis test were noted to reject the hypothesis that the variables were independent. The next calculation completed was the alpha versus the p-value. The p-value in this case was determined to be 0. Since the p-value is less than the alpha then the results are to reject the null hypothesis per the chi-square test of independence (Bobbitt, 2020).

E2. Limitations

There were various limitations to using the chi-square test for independence. One of the limitations would be that the variables must have a frequency of greater than five. Both the “Churn” and “Contract” data met that qualification to proceed. Another limitation was interpretation of the results. Month to month showed a higher occurrence of churn compared to other contract terms, but that does not mean that the correlation between these variable implied that the contract terms were the cause of discontinuance of service. Another limitation would be that the results could be different if a sample size were used. The size of the dataset could provide differing results and distribution. The chi-square test for independence was not ideal for small datasets under twenty values or exceptionally large datasets.

E3. Recommendation of Action

The result of the research question deemed there was an association between the customers who left within the last month and the contract terms specified. The recommended course of action based on the results of the research analysis was to suggest the telecommunication company put efforts into its marketing strategies in relation to retention of month-to-month clients. This would aid the company into building as well as maintaining the customer base (Expert Panel, Forbes Agency Council, 2019). The marketing strategies could provide rewards or discounts to recurring customers that maintain long tenure with the company.

Part V. Supporting Documents

F. Panopto Video

Please see attached Panopto video link. This is a video providing an overview of the Python code used to discover anomalies, the data cleaning process, and the chi-square test of independence analysis.   The recording will demonstrate the code's warning and error-free functionality as well as provide an overview of the programming environment used, Jupyter Notebook.

Link Found here:

[redacted]

G. Third-Party Web Sources

There were no web sources used to acquire data nor segments of third-party code to support the application. All code and segments are original work.

H. References

Bobbitt, Z. (2020, April 27). *Chi-Square Test of Independence: Definition, Formula, and Example*. Retrieved from Statology Study: https://www.statology.org/chi-square-test-of-independence/

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python.* Sebastopol: O'Rielly.

Chapman, J. (n.d.). *Chi-square Test of Independence | Python*. Retrieved October 13, 2022, from Datacamp: https://campus.datacamp.com/courses/hypothesis-testing-in-python/proportion-tests-3?ex=7

Expert Panel, Forbes Agency Council. (2019, December 20). *How To Increase Client Retention: 10 Effective Strategies*. Retrieved from Forbes: https://www.forbes.com/sites/forbesagencycouncil/2019/12/20/how-to-increase-client-retention-10-effective-strategies/?sh=3f663798390b

Lewinson, E. (2020, June 4). *The new kid on the statistics-in-Python block: pingouin*. Retrieved from Towards Data Science: https://towardsdatascience.com/the-new-kid-on-the-statistics-in-python-block-pingouin-6b353a1db57c