D207 Exploratory Data Analysis Performance Task

Sean Simmons

WDU Data Analytics

MSDA D207

January 2023

Please Note- All code is included in the submission as "D207_PA_Code.ipynb"

A. Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:

1. Provide one question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).

2. Explain how stakeholders in the organization could benefit from an analysis of the data.

3. Identify *all* of the data in your data set that are relevant to answering your question in part A1.

1. "Does the survey response for Timely Fixes have an effect on Churn?" is my research question. I will use a chi-square test to determine if the variables have an association.
2. Stakeholders in the organization could benefit from the analysis of the data in two ways. Firstly, they can have a better understanding of the factors that make customers more likely to leave the business, better informing their marketing and retention. Secondly, this question can help inform all stakeholders of how the organization operates and whether more analysis needs to be done in how to increase the customer's satisfaction/survey response. Alternatively, this analysis informs the stakeholders that more data gathering needs to be done to better understand this relationship.

3. The data in the dataset relevant to answering my question involves the following variables from the churn dataset, please see the output below the list for data type and examples:
    - customer id - unique customer identification used as a key
    - Churn- Whether a customer will leave the company, yes/no
    - MonthlyCharge - Average monthly charge to the customer
    - Bandwidth_GB_Year - Internet usage of the customer, per year
    - Item 2 - customer response to the importance of timely fixes
        i. Renamed to item2_fixes
    - Item 4 - customer response to the importance of reliability
        i. Renamed to item4_reliability

Example of Variables:

```
 <class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Lat               10000 non-null  float64
 1   Lng               10000 non-null  float64
 2   Income            10000 non-null  float64
 3   Outage_sec_perweek  10000 non-null  float64
 4   Tenure            10000 non-null  float64
 5   MonthlyCharge     10000 non-null  float64
 6   Bandwidth_GB_Year   10000 non-null  float64
dtypes: float64(7)
memory usage: 547.0 KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 17 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   CaseOrder         10000 non-null  int64
 1   Zip               10000 non-null  int64
 2   Population        10000 non-null  int64
 3   Children          10000 non-null  int64
 4   Age               10000 non-null  int64
 5   Email             10000 non-null  int64
 6   Contacts          10000 non-null  int64
```

```
 7   Yearly_equip_failure  10000 non-null  int64
 8   item1_responses       10000 non-null  int64
 9   item2_fixes           10000 non-null  int64
 10  item3_replacements    10000 non-null  int64
 11  item4_reliability     10000 non-null  int64
 12  item5_options         10000 non-null  int64
 13  item6_respectfulness  10000 non-null  int64
 14  item7_courteous       10000 non-null  int64
 15  item8_listening       10000 non-null  int64
 16  TechSupport_numeric   10000 non-null  int64
dtypes: int64(17)
memory usage: 1.3 MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 27 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Customer_id      10000 non-null  object
 1   Interaction      10000 non-null  object
 2   UID              10000 non-null  object
 3   City             10000 non-null  object
 4   State            10000 non-null  object
 5   County           10000 non-null  object
 6   Area             10000 non-null  object
 7   TimeZone         10000 non-null  object
 8   Job              10000 non-null  object
 9   Marital          10000 non-null  object
 10  Gender           10000 non-null  object
 11  Churn            10000 non-null  object
 12  Techie           10000 non-null  object
 13  Contract         10000 non-null  object
 14  Port_modem       10000 non-null  object
 15  Tablet           10000 non-null  object
 16  InternetService  10000 non-null  object
 17  Phone            10000 non-null  object
 18  Multiple         10000 non-null  object
 19  OnlineSecurity   10000 non-null  object
 20  OnlineBackup     10000 non-null  object
 21  DeviceProtection 10000 non-null  object
 22  TechSupport      10000 non-null  object
 23  StreamingTV      10000 non-null  object
 24  StreamingMovies  10000 non-null  object
 25  PaperlessBilling 10000 non-null  object
 26  PaymentMethod    10000 non-null  object
dtypes: object(27)
memory usage: 2.1+ MB
None
```

B. Describe the data analysis by doing the following:

1. Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:

- **chi-square**
- **t-test**
- **ANOVA**

2. Provide the output and the results of *any* calculations from the analysis you performed.

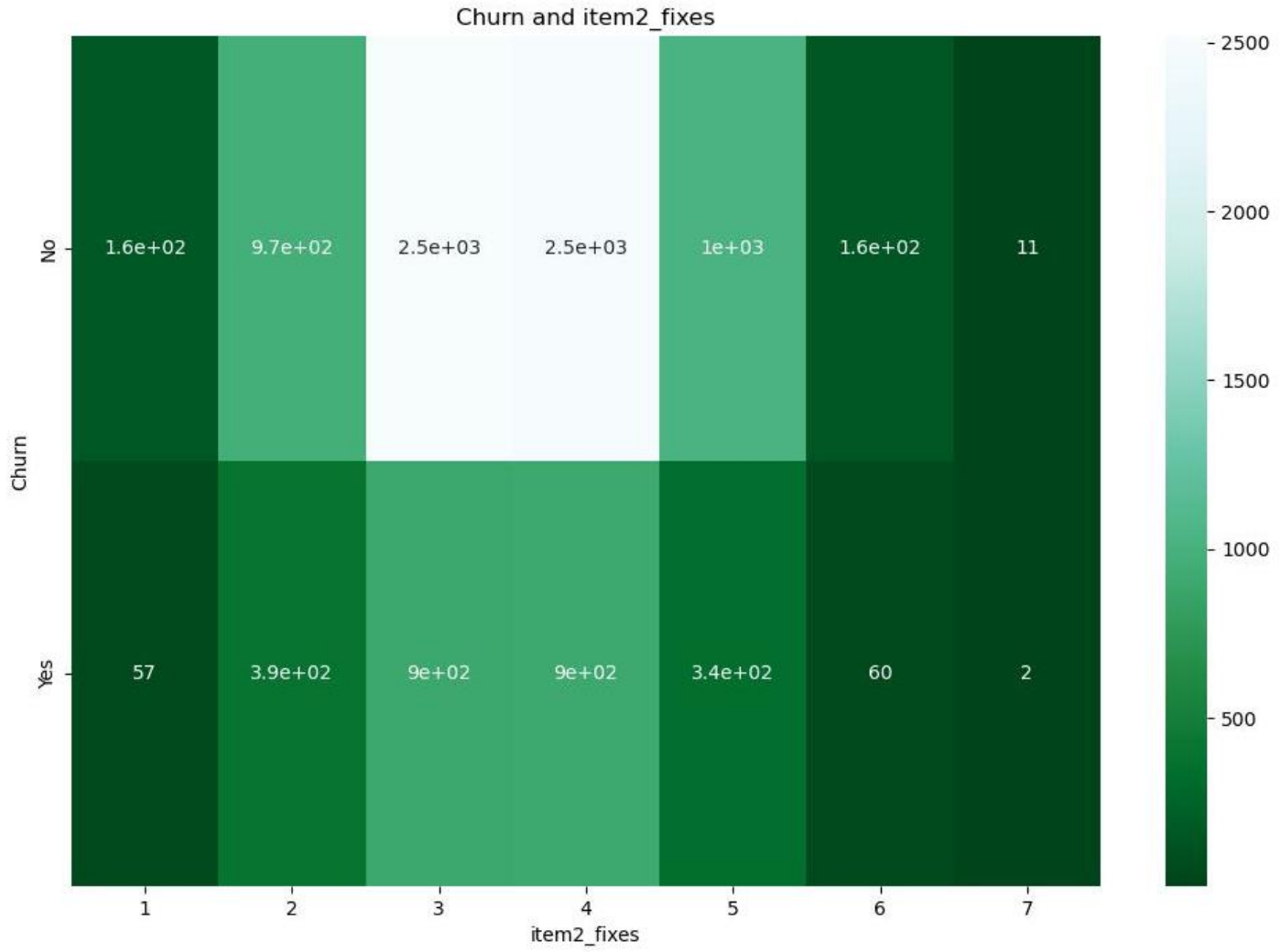
3. Justify why you chose this analysis technique.

1. The test I've chosen to run is a chi-square independence test using Python.
2. Outputs and results from analysis: The P value is 0.5093789499498207, which is not less than 0.05, so I fail to reject the null hypothesis. We cannot say there is a strong dependent connection between Churn and item2_fixes.

```
item2_fixes    1    2     3     4     5    6   7
Churn
No           160  973  2519  2507  1025  155  11
Yes           57  387   896   905   343   60   2
item2_fixes         1         2         3         4         5         6  \
Churn
No           0.021769  0.132381  0.342721  0.341088  0.139456  0.021088
Yes          0.021509  0.146038  0.338113  0.341509  0.129434  0.022642

item2_fixes         7
Churn
No           0.001497
Yes          0.000755
```

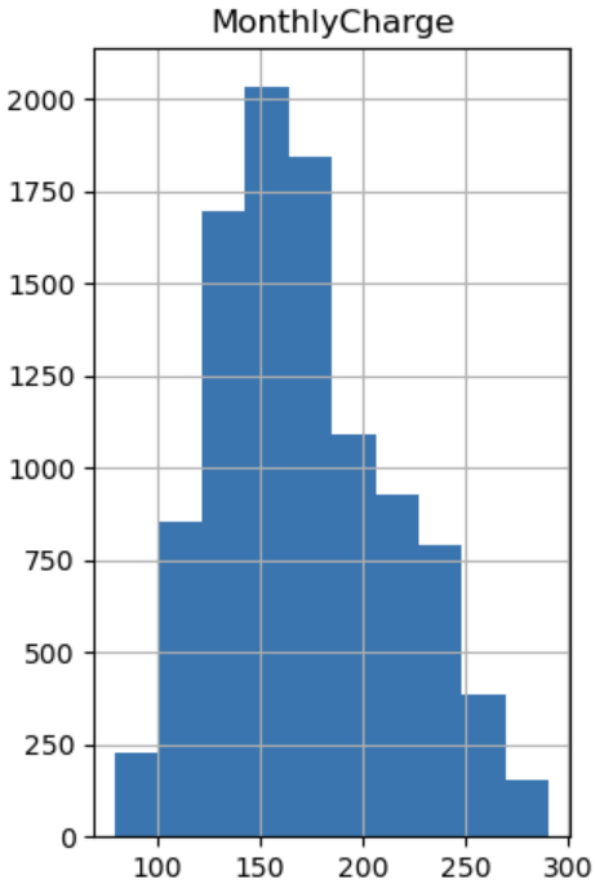
```
dof=6
probability=0.950, critical=12.592, stat=5.272
Fail to Reject Null Hypothesis
Fail to Reject Null Hypothesis
0.5093789499498207
```

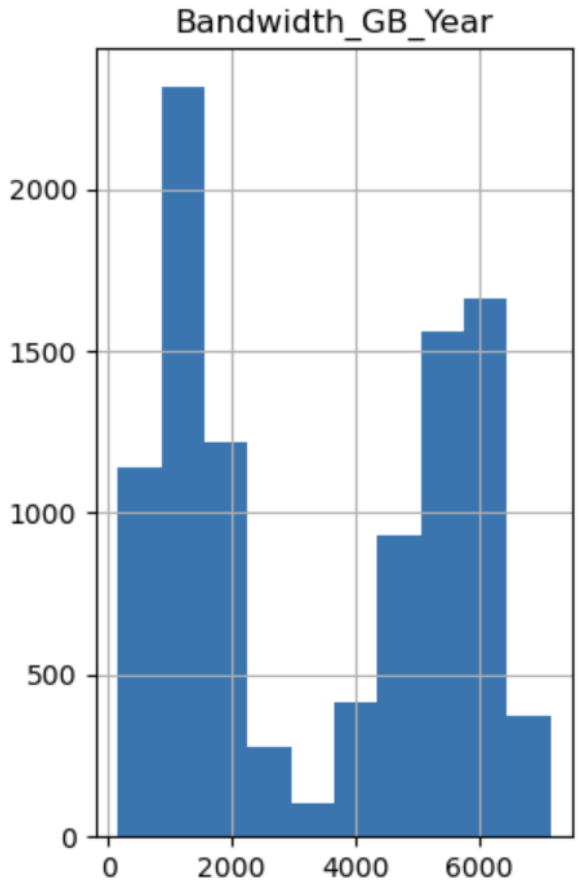


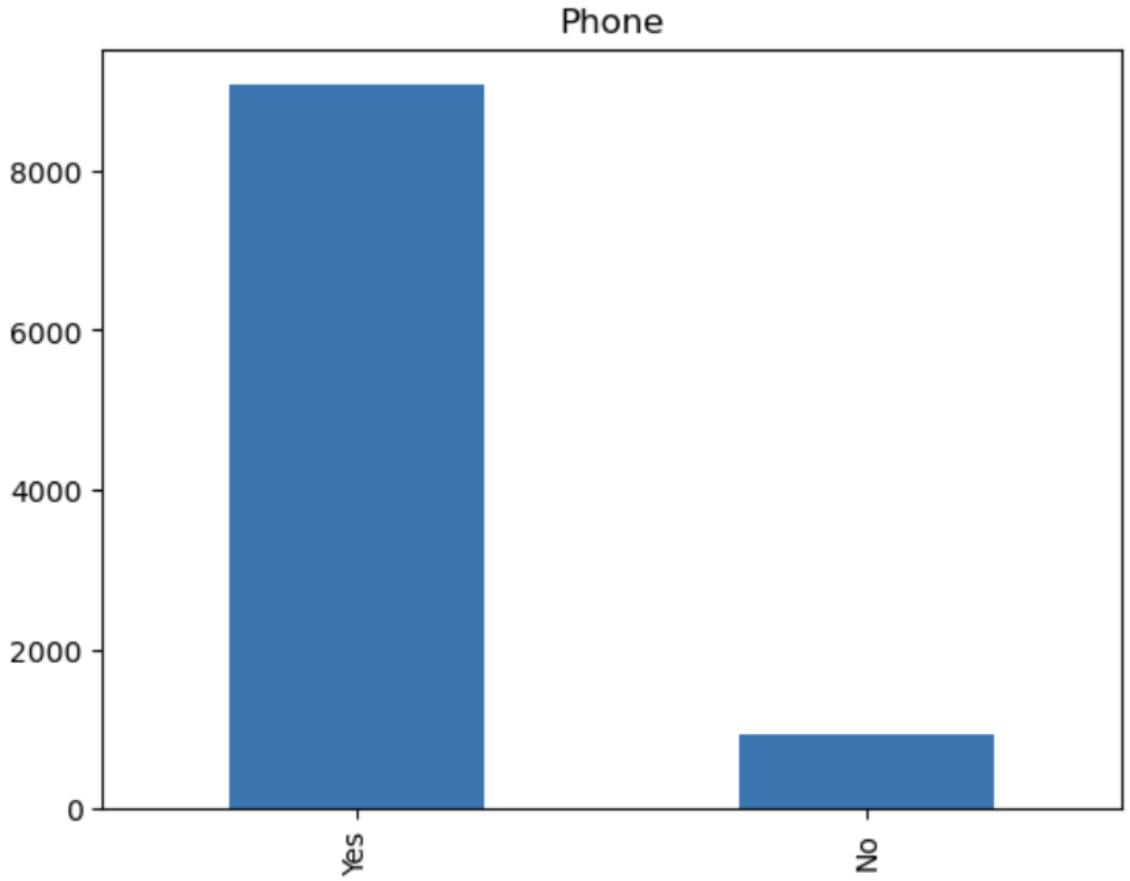
3. I chose this analysis technique because chi-square independence tests excel in showing whether two categorical variables are dependent or independent of each other. We are trying to see a statistically significant connection between Churn and item2_fixes. The most direct way of seeing this would be a chi-square independence test (McDonald, 2017).

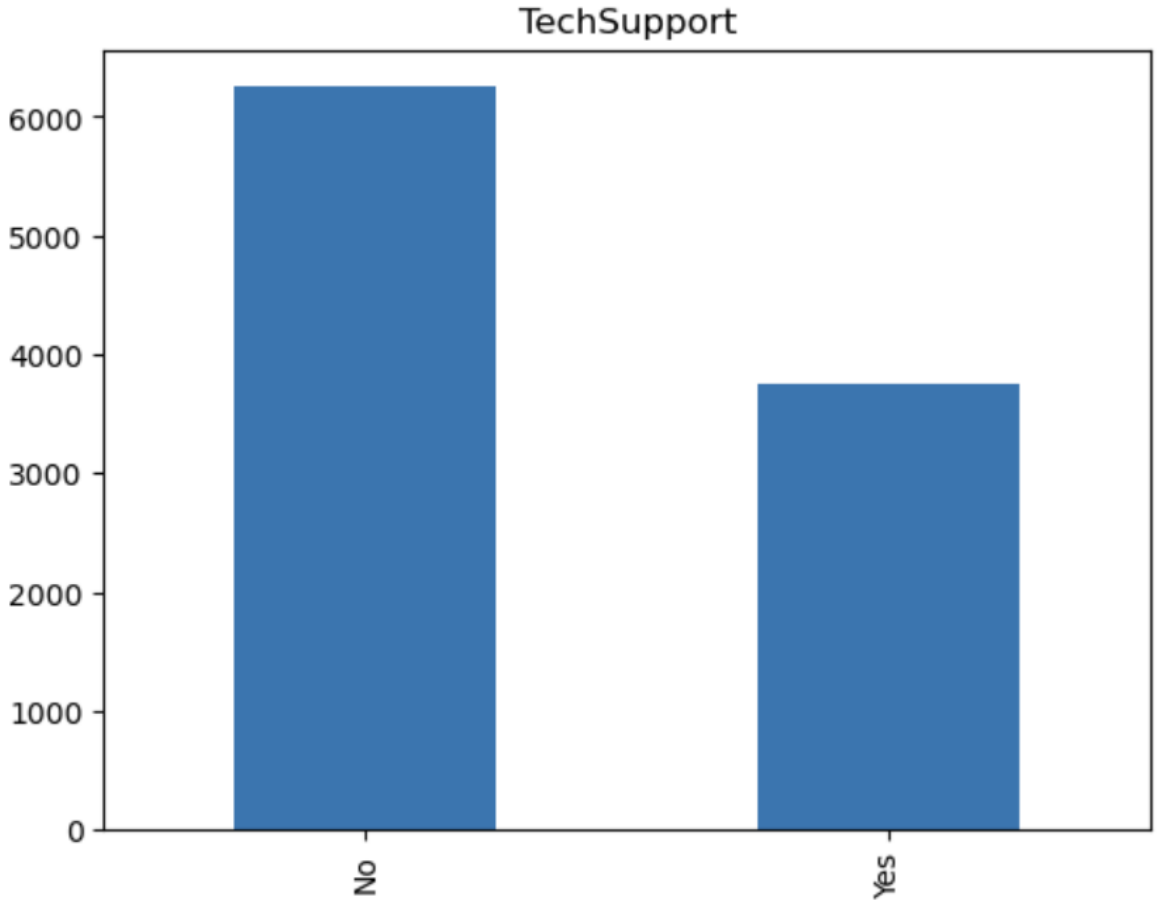
C. Identify the distribution of two continuous variables and two categorical variables using univariate statistics from your cleaned and prepared data.

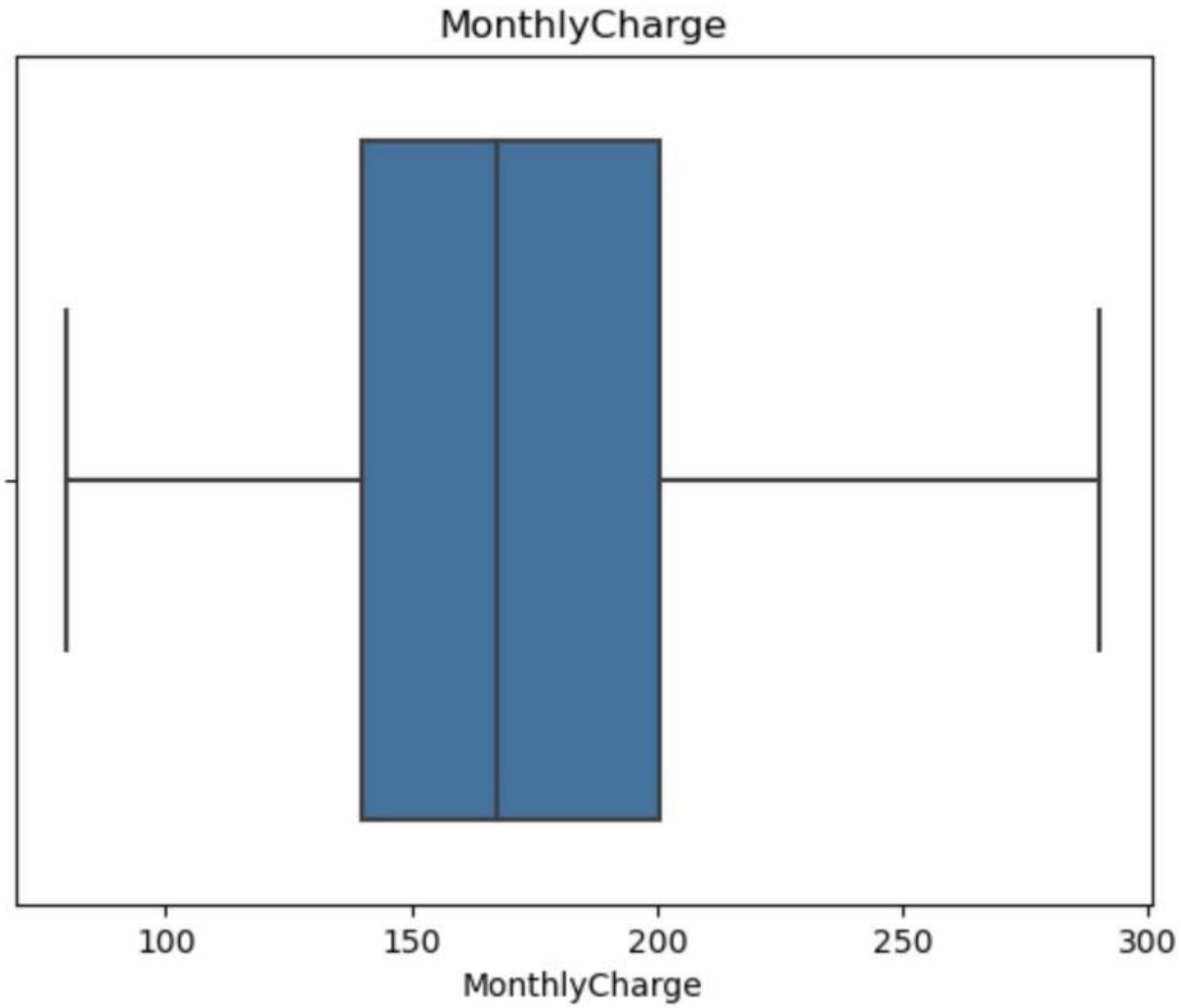
1. **Represent your findings in Part C, visually as part of your submission.**
    - Two continuous variables: MonthlyCharge, Bandwidth_GB_Year
    - Two categorical variables: Phone, TechSupport
    - Univariate statistics technique: I will use histograms and boxplots to analyze my continuous and categorical variables one at a time (Zach 2021).
    - Output using Jupyter Notebook:

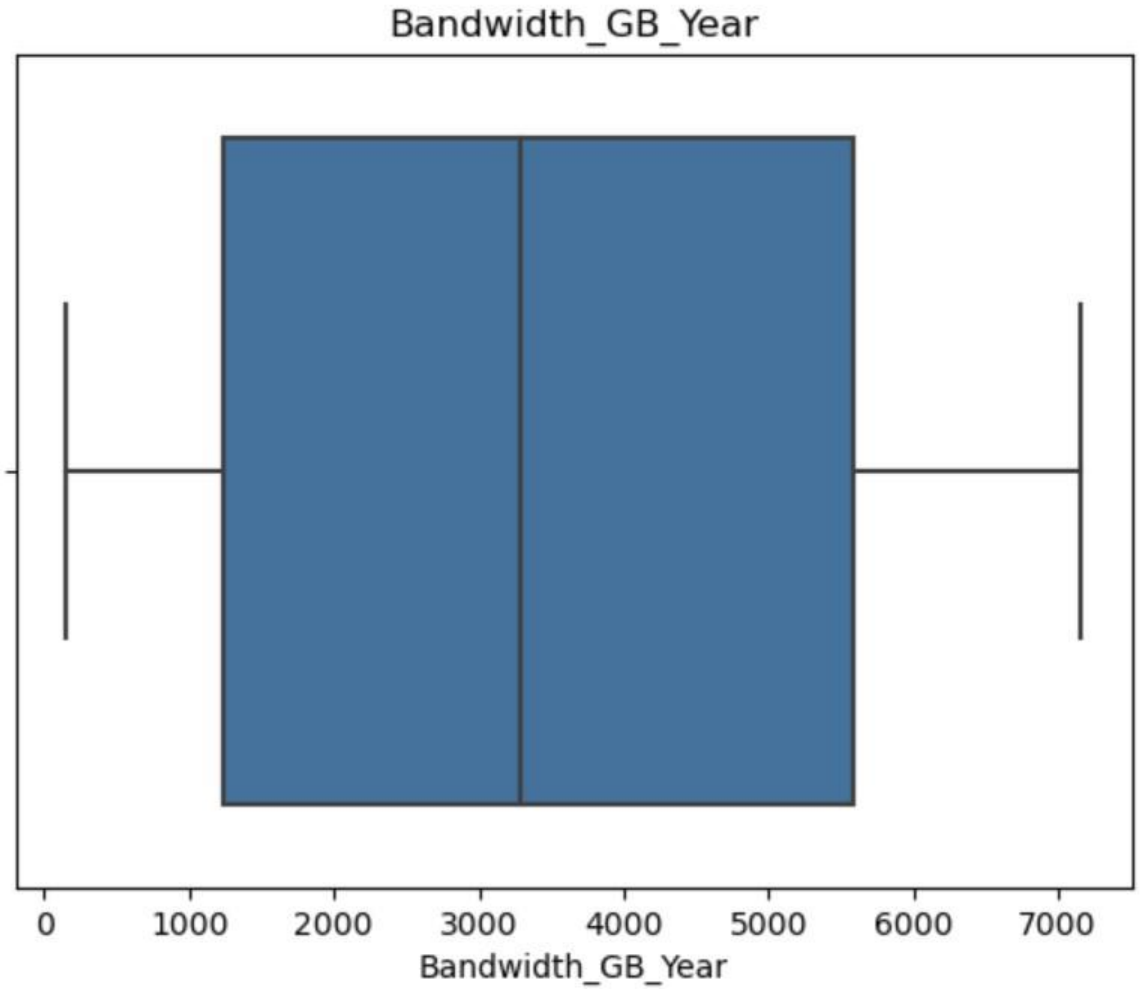

MonthlyCharge
2000
1750
1500
1250
1000
750
500
250
0
100
150
200
250
300

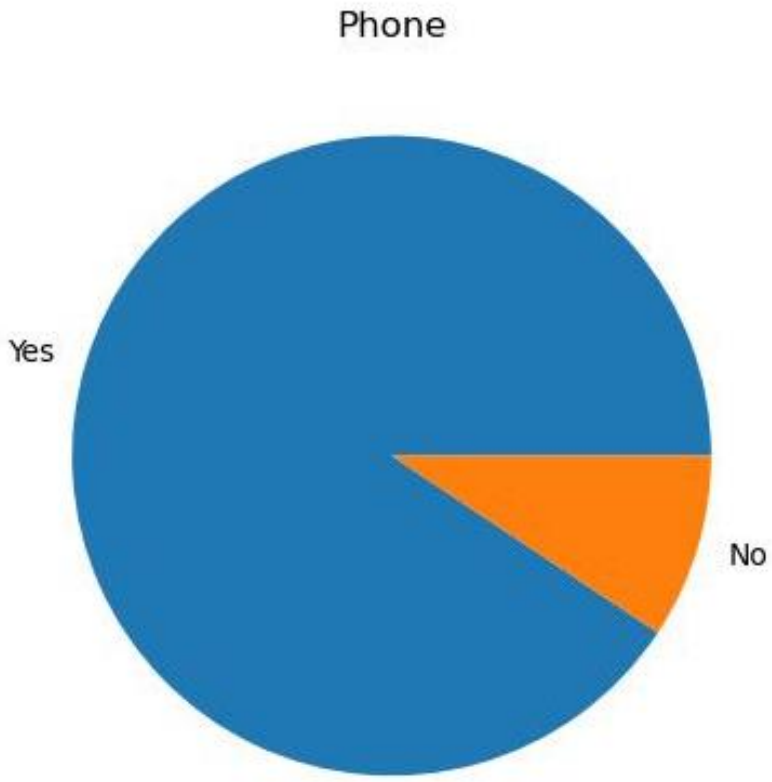

Bandwidth_GB_Year
2000
1500
1000
500
0
0
2000
4000
6000

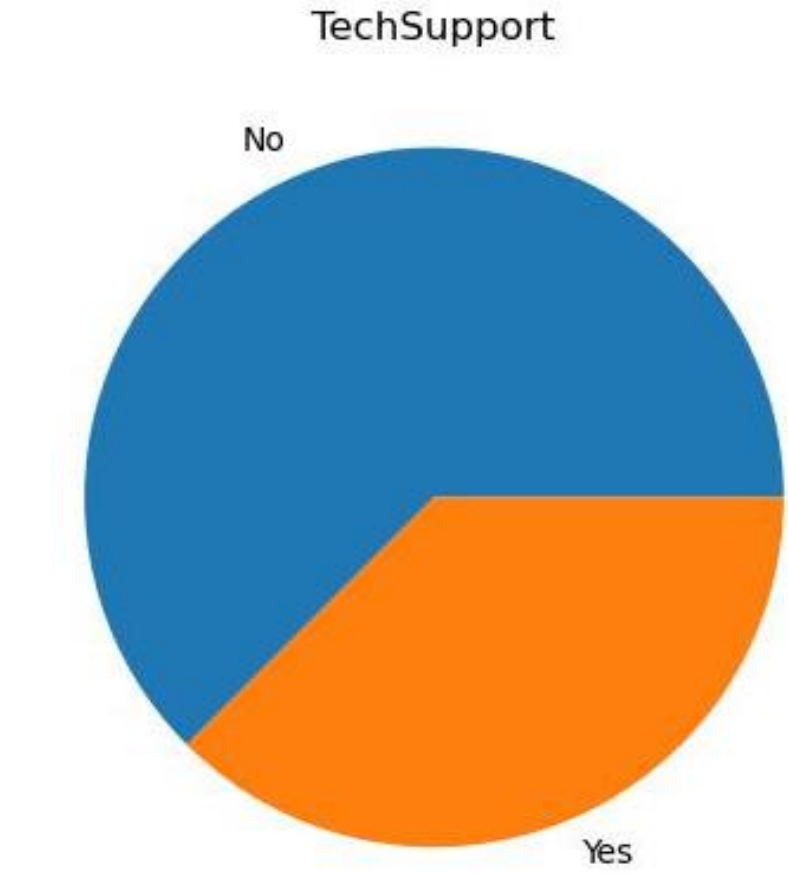

Phone
8000
6000
4000
2000
0
Yes
No


TechSupport
6000
5000
4000
3000
2000
1000
0
No
Yes


MonthlyCharge
100
150
200
250
300
MonthlyCharge

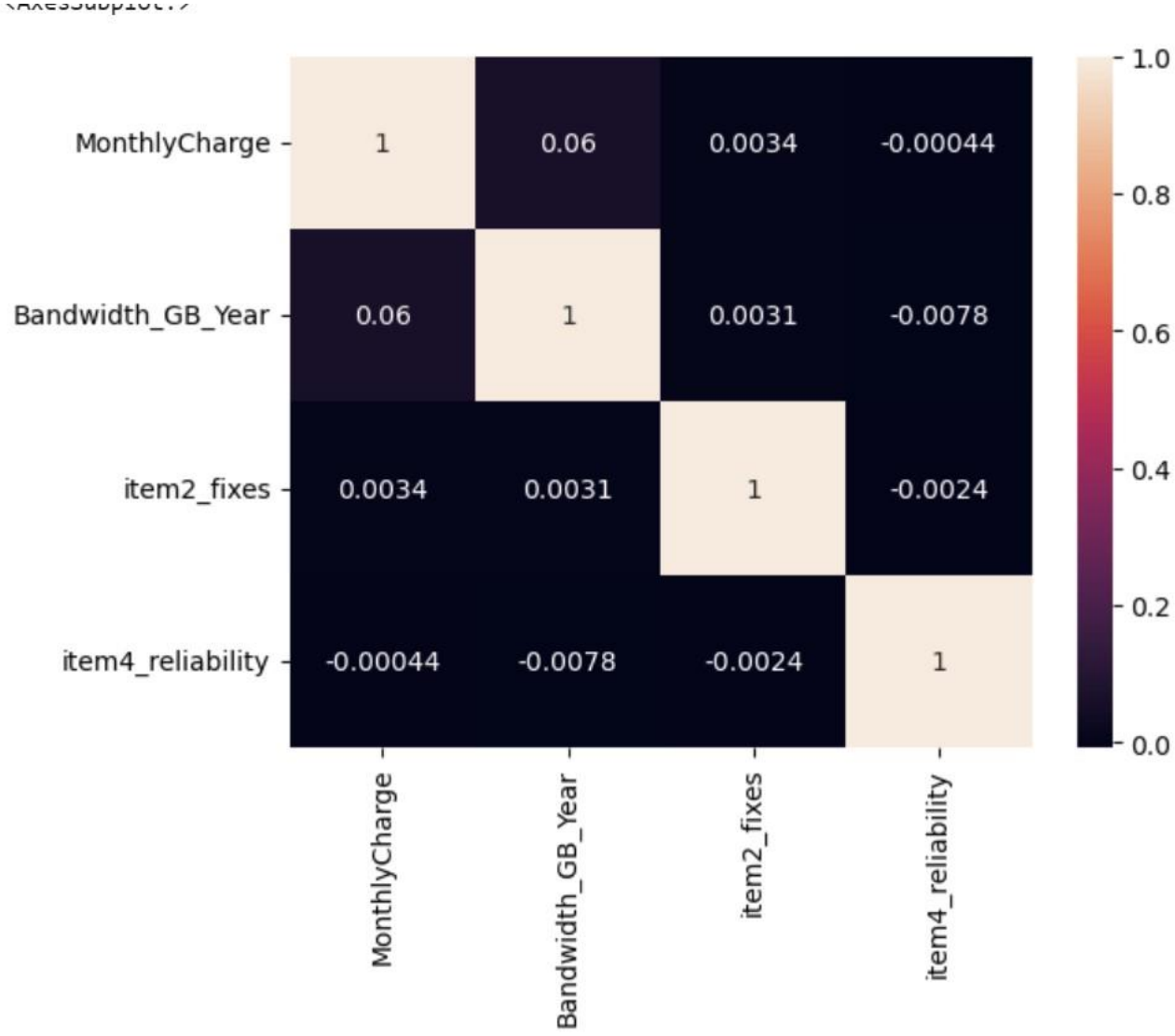

Bandwidth_GB_Year
0
1000
2000
3000
4000
5000
6000
7000
Bandwidth_GB_Year

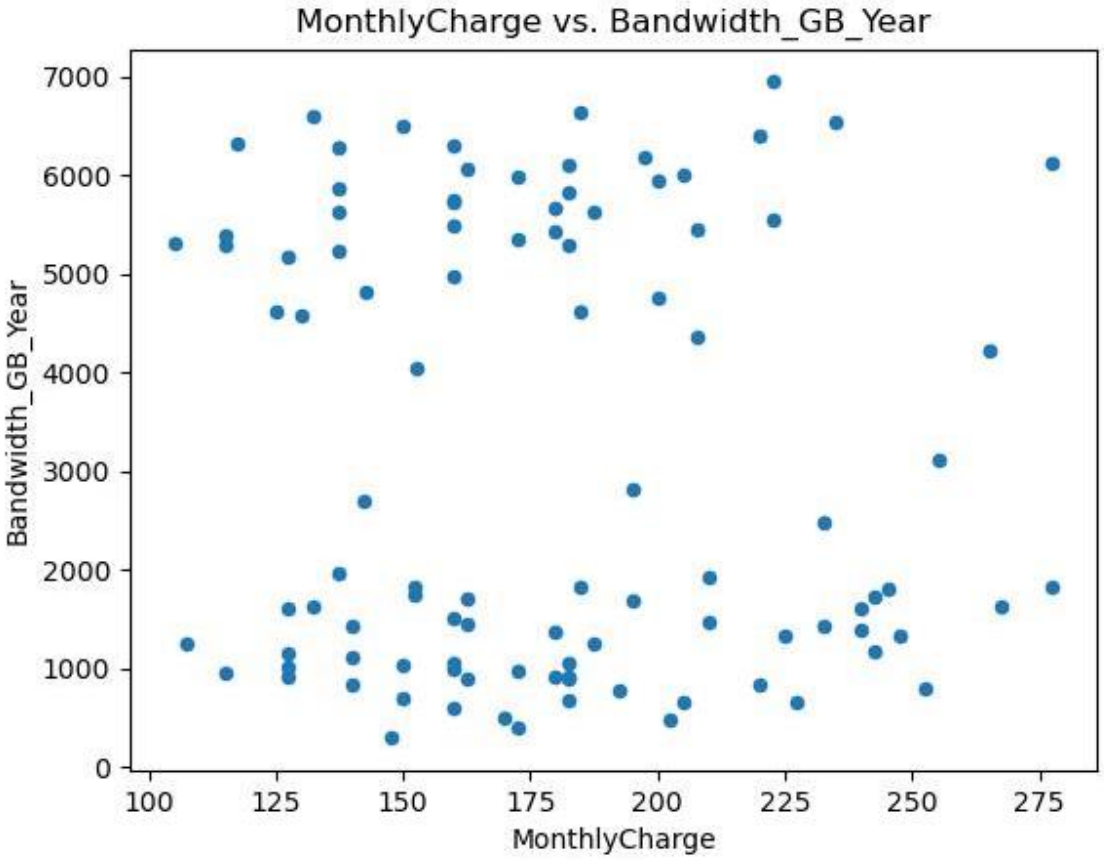

Phone
Yes
No

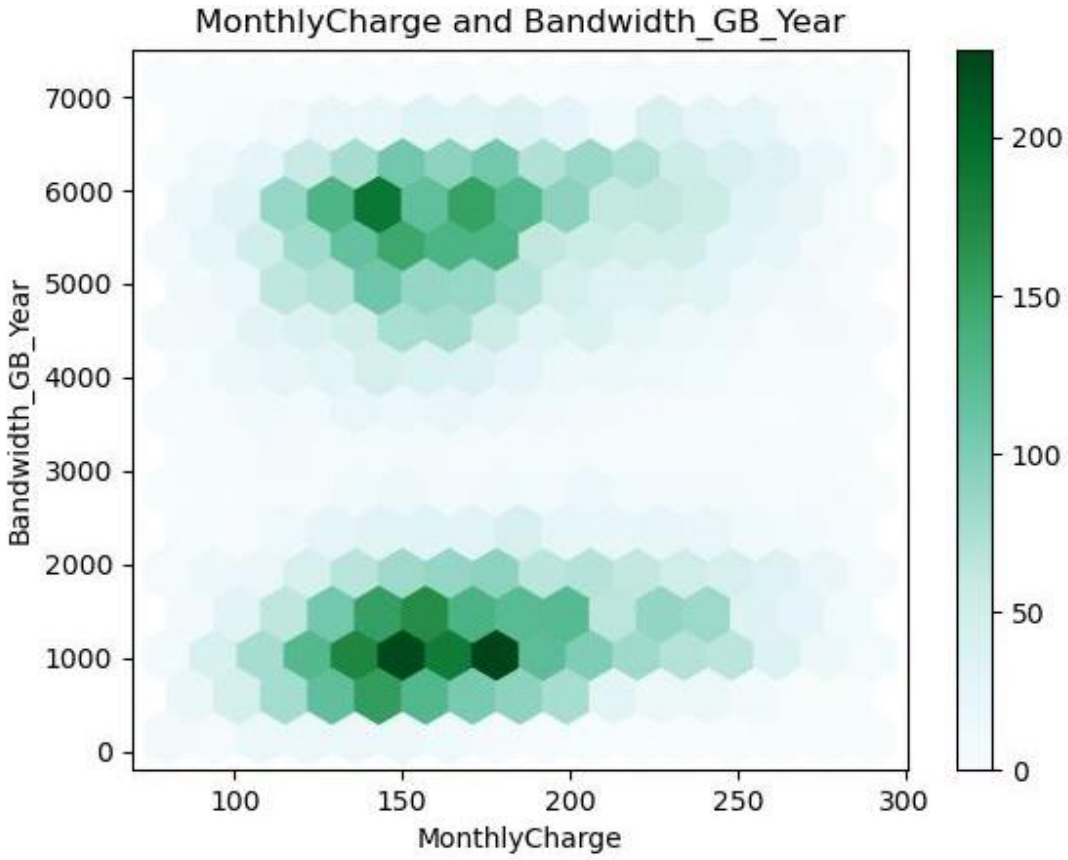

TechSupport
No
Yes

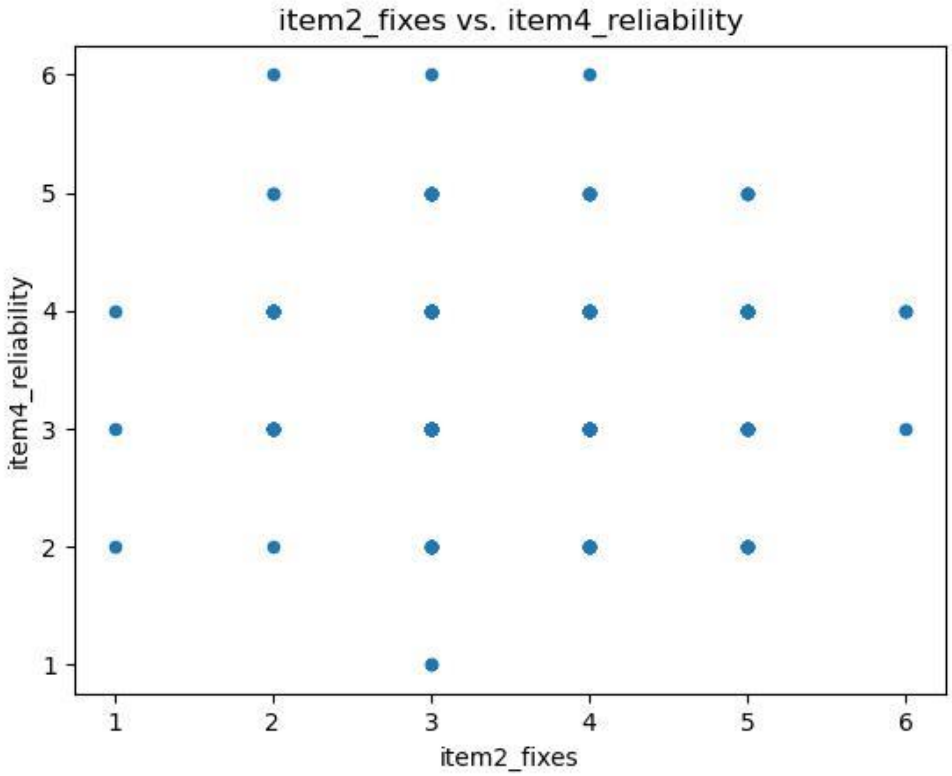
D. Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data.

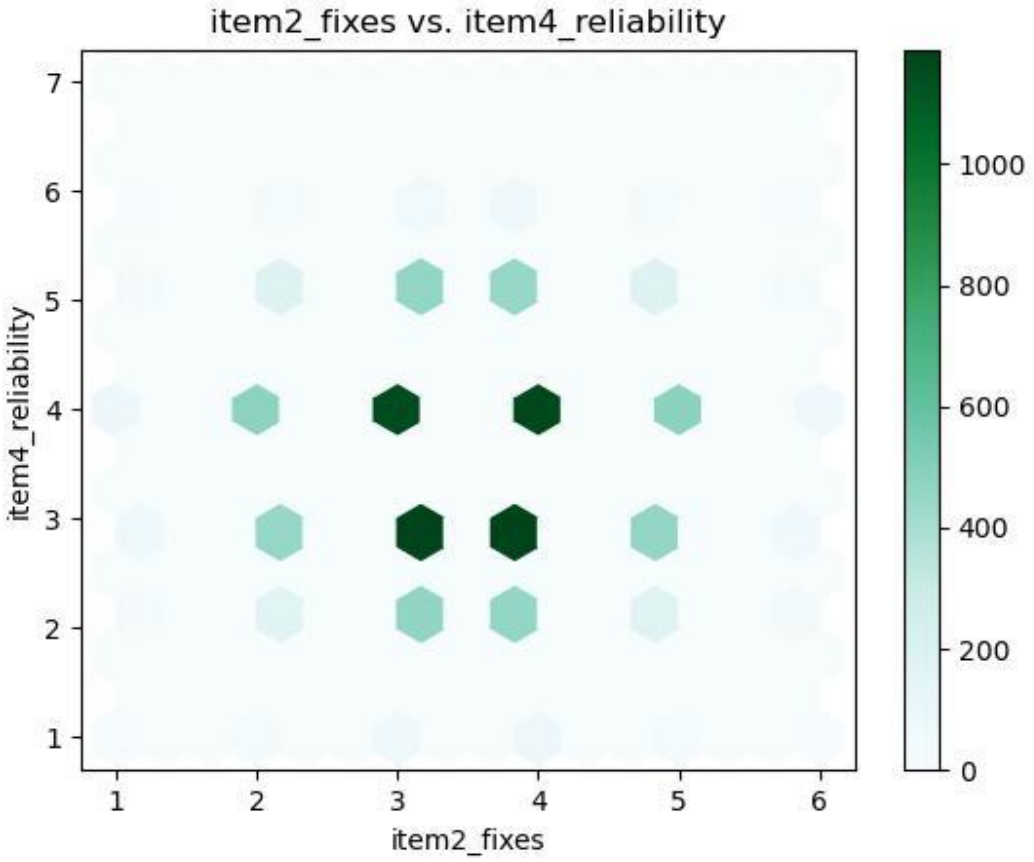
1. **Represent your findings in Part D, visually as part of your submission.**
    - Two continuous variables: MonthlyCharge, Bandwidth_GB_Year
    - Two categorical variables: item2_fixes, item4_reliability'
    - Bivariate statistics: I will use scatter plots and heat maps to analyze two variables at a time (Zach 2021).
    - Output:




MonthlyCharge vs. Bandwidth_GB_Year
Bandwidth_GB_Year
MonthlyCharge
0
1000
2000
3000
4000
5000
6000
7000
100
125
150
175
200
225
250
275


MonthlyCharge and Bandwidth_GB_Year
Bandwidth_GB_Year
MonthlyCharge
0
1000
2000
3000
4000
5000
6000
7000
100
150
200
250
300
0
50
100
150
200


item2_fixes vs. item4_reliability
item4_reliability
item2_fixes
1
2
3
4
5
6


item2_fixes vs. item4_reliability
item4_reliability
item2_fixes
1
2
3
4
5
6
7
0
200
400
600
800
1000

E. Summarize the implications of your data analysis by doing the following:

1. **Discuss the results of the hypothesis test.**
2. **Discuss the limitations of your data analysis.**
3. **Recommend a course of action based on your results.**

1. The chi- test uses a p-value to reject a null hypothesis.The P value is 0.5093789499498207, which is not less than 0.05, so I fail to reject the null hypothesis. We cannot say there is a strong dependent connection between Churn and item2_fixes.
2. The limitations of my data analysis is that this test indicates a probability of association between two variables. The extent of which and exact nature are not described and should be further investigated.
3. The course of action I recommend is to gather more customer data and continue analyzing for further factors that could be associated with churning. I suggest the organization should take a look at their procedures for dealing with how timely they fix customers issues and how they record customer data to determine if they want to have an association with churning in their business model and how they can make that connection to help with retention. A larger data set could also provide a more clear analysis. Further statistical analysis with other methods can be performed between item2_fixes and churn to further strengthen our knowledge about the connection between the two.

F. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.

Link for Panopto Video:

This link is also provided in the submission for the PA.

G. Reference the web sources used to acquire segments of third-party code to support the analysis.

The univariate and bivariate code was informed by the following:

Zach. (2021, November 22). *How to Perform Univariate Analysis in Python (With Examples)*. Statology. https://www.statology.org/univariate-analysis-in-python/

Zach. (2021, November 22). *How to Perform Bivariate Analysis in Python (With Examples)*. Statology. How to Perform Bivariate Analysis in Python (With Examples) - Statology

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

The choice of chi square test was informed by the following:

McDonald, J.(2017, June 27) *2.3: Chi-Square Test of Goodness-of-Fit*. Statistics LibreTexts. 2.3: Chi-Square Test of Goodness-of-Fit - Statistics LibreTexts

I. Demonstrate professional communication in the content and presentation of your submission.

This part of the rubric cannot be summarized and instead shows throughout the document.