Performance Assessment

D207 – Exploratory Data Analysis

Jessica Hosey

MSDA, College of Information Technology

Western Governors University

April 14th, 2024

A1) Provide one question that is relevant to your chosen data set.

What customers are most at risk for churn, and which variables/features drive this churn?

A2) Explain how stakeholders in the organization could benefit from an analysis of the data.

Stakeholders can use this information to change or modify how these specific variables/features may stop customers from churning on the services, whether this information would lead to new company promotions, marketing strategies, or overhaul of customer service. This information should provide the organization's stakeholders with valuable knowledge about their high-risk customers, with some confidence about why they churn and what might stop them from churning.

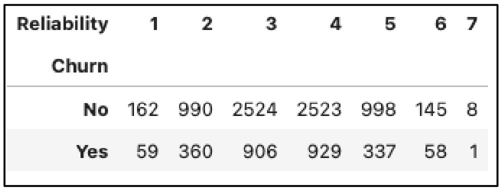
A3) Identify all data in your data set relevant to answering your question in part A1.

The following data will be used in this exploratory analysis: Churn, Tenure, Monthly Charge, Bandwidth, and Survey Responses (timely response, timely fixes, timely replacements, reliability, options, respectful response, courteous exchange, and evidence of active listening).

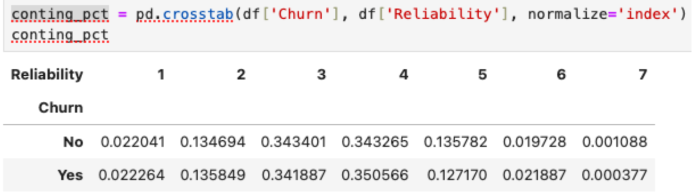
Churn (categorical), Tenure (continuous numerical), Monthly Charge (continuous numerical), and Bandwidth (continuous numerical) will allow insight into what services customers are using and at what cost to them. The Survey Responses (discrete numerical) is a group of data gathered on customers' feelings on specific services provided by the company. These responses range from 1 to 8 and could help show whether the company needs to change/change specific service supports to slow/cut customer churn rate.

B1) Python Code Link

B2) Provide the output and the results of any calculations from your analysis.

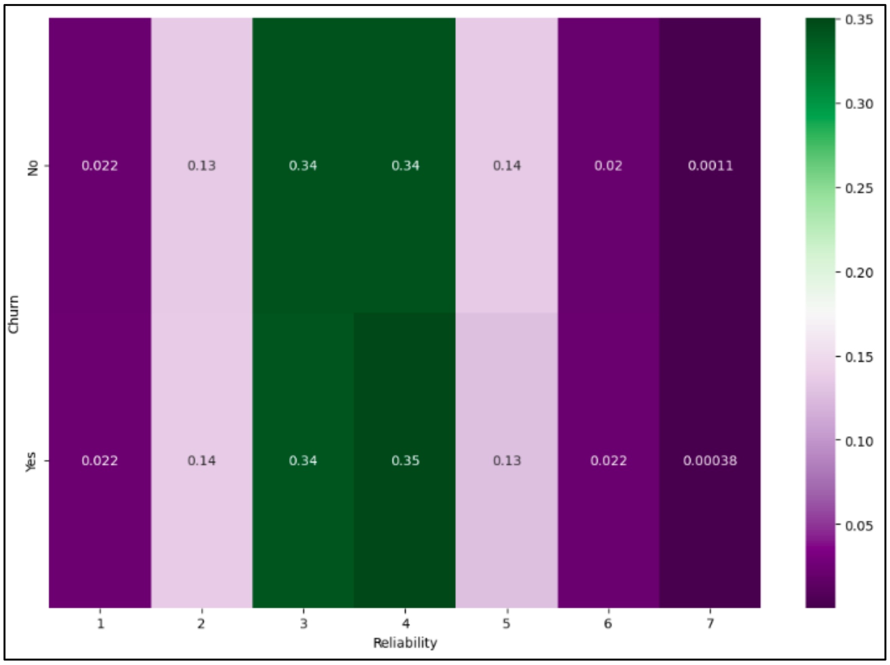


Reliability	1	2	3	4	5	6	7
Churn							
No	162	990	2524	2523	998	145	8
Yes	59	360	906	929	337	58	1

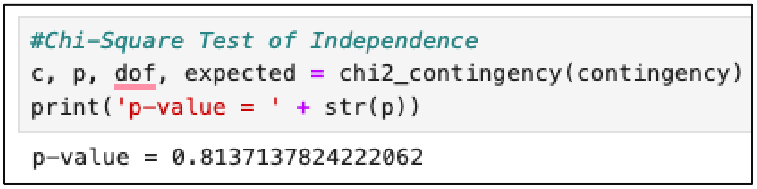


```
conting_pct = pd.crosstab(df['Churn'], df['Reliability'], normalize='index')
conting_pct
```

Reliability	1	2	3	4	5	6	7
Churn							
No	0.022041	0.134694	0.343401	0.343265	0.135782	0.019728	0.001088
Yes	0.022264	0.135849	0.341887	0.350566	0.127170	0.021887	0.000377



Using a Chi-squared Test, here are the outputs of the analysis code:



```
#Chi-Square Test of Independence
c, p, dof, expected = chi2_contingency(contingency)
print('p-value = ' + str(p))
```

```
p-value = 0.8137137824222062
```

B3) Justify why you chose this analysis technique.

Churn is the dependent variable being examined in the analysis of telecom companies. Churn is a categorical variable with responses including yes or no answers from customers telling whether they chose to continue using their services or stop using this company's services. This is why the chi-square test was selected for analysis.

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

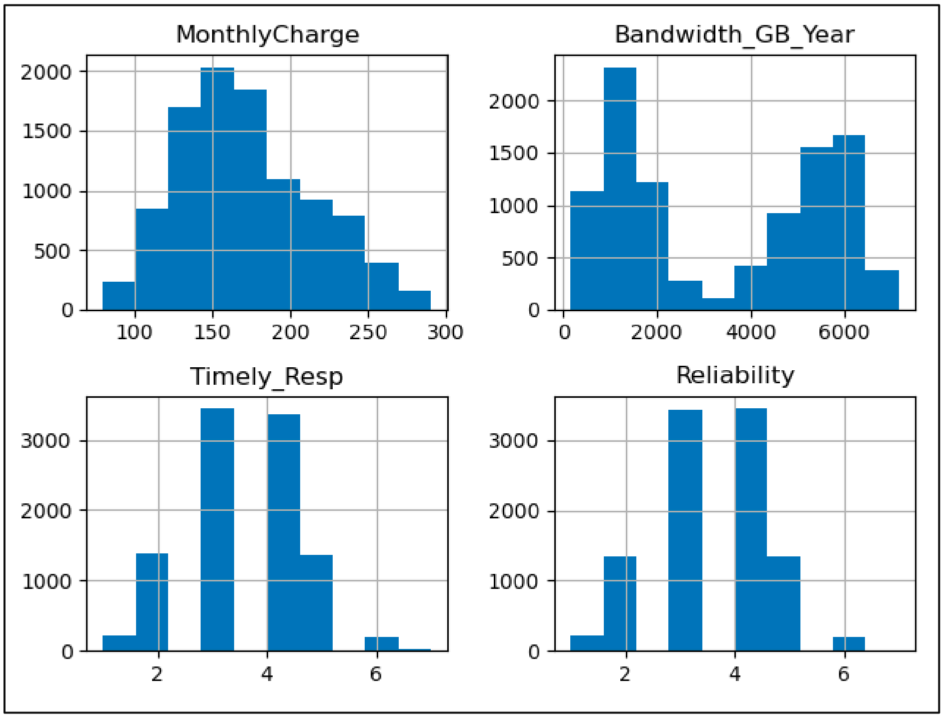
Two Continuous Variables:

- MonthlyCharge
- Bandwidth_GB_Year

Two Categorical Variables:

- Item 1 – Which was renamed to 'Timely_Resp'
- Item 4 – Which was renamed to 'Reliability'

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

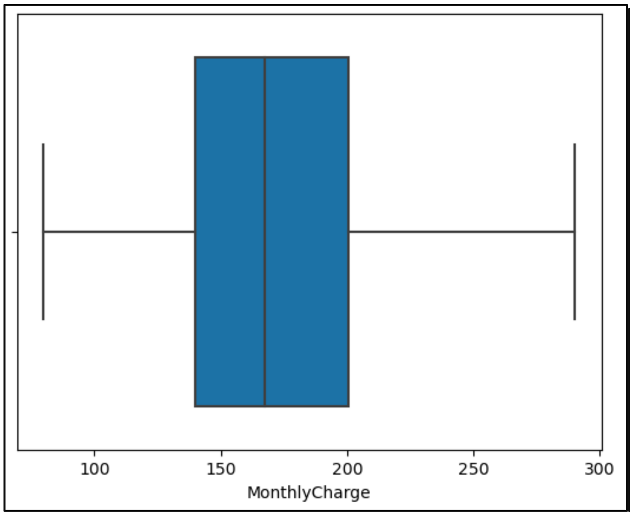


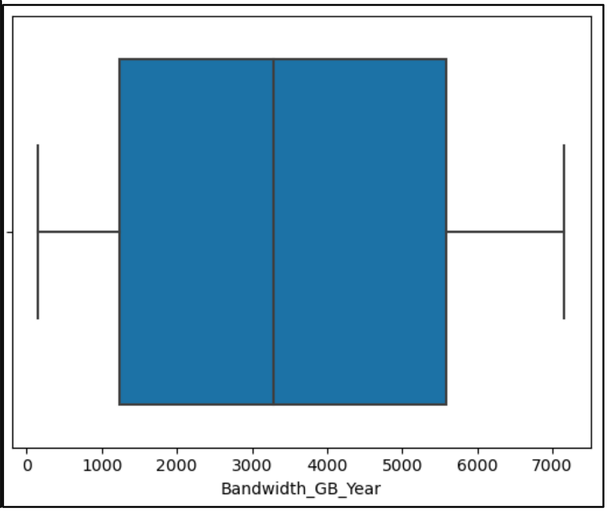
Two Continuous Variables:

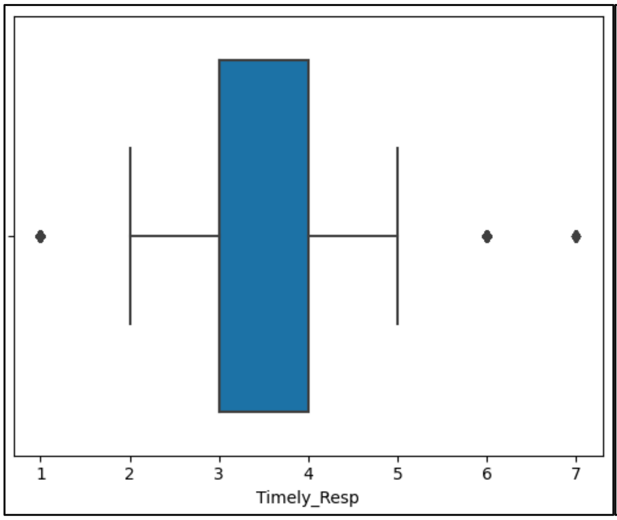
- MonthlyCharge
- Bandwidth_GB_Year

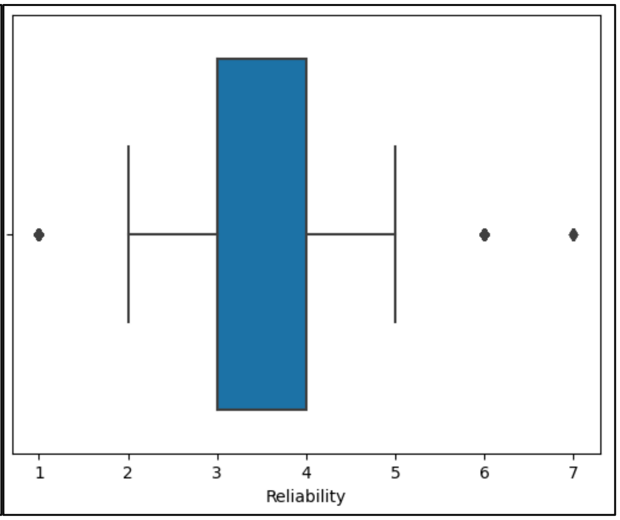
Two Categorical Variables:

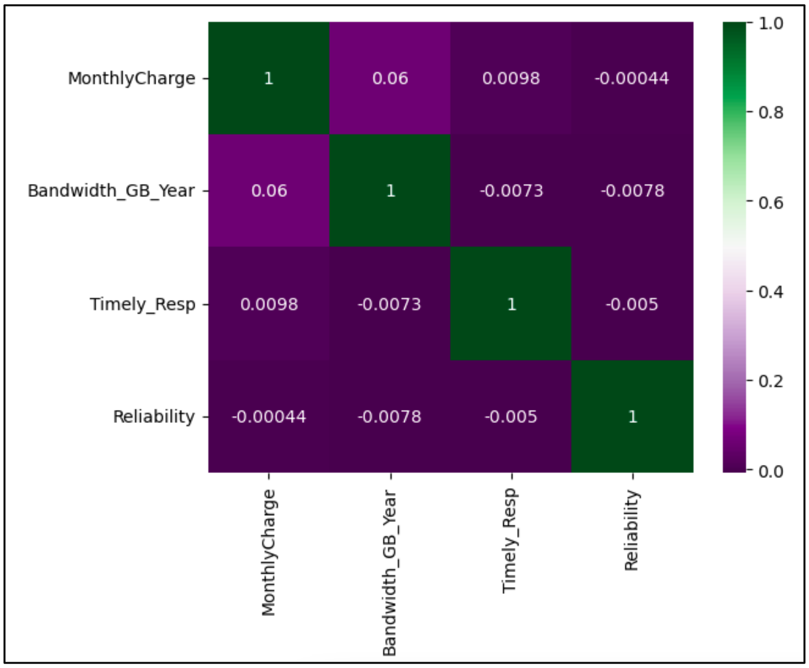
- Churn
- Item4 – Renamed to "Reliability"

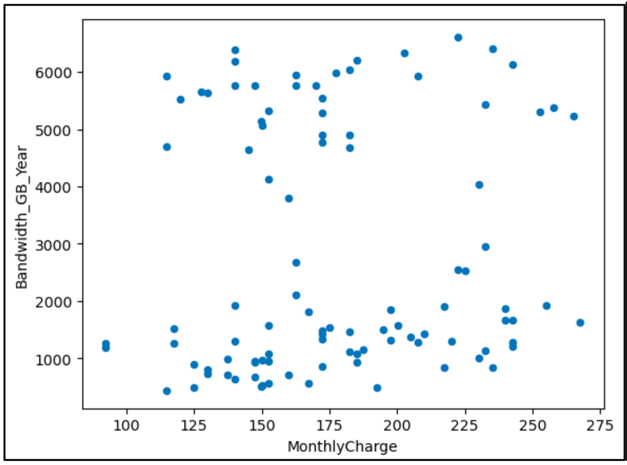


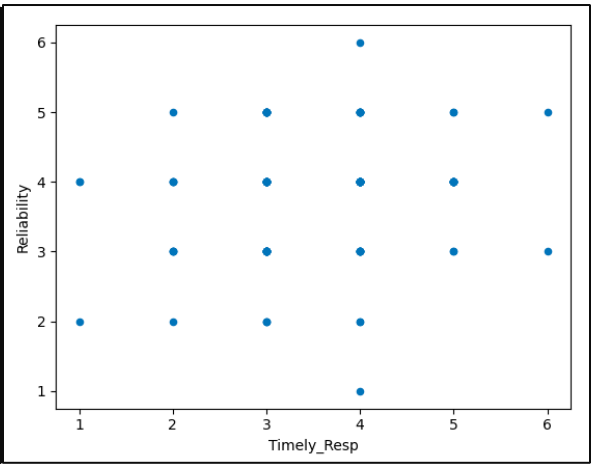


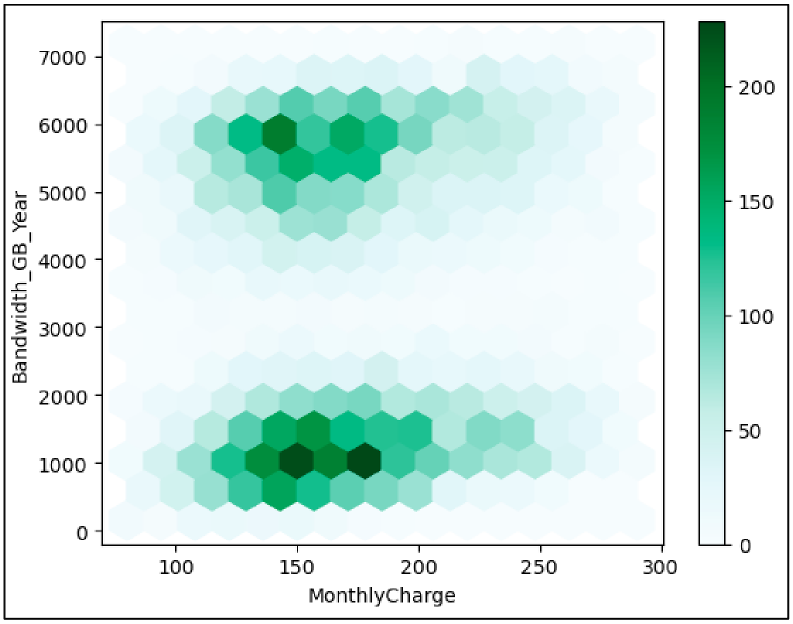

1
2
3
4
5
6
7
Timely_Resp


1
2
3
4
5
6
7
Reliability


MonthlyCharge
Bandwidth_GB_Year
Timely_Resp
Reliability
1
0.06
0.0098
-0.00044
0.06
1
-0.0073
-0.0078
0.0098
-0.0073
1
-0.005
-0.00044
-0.0078
-0.005
1
MonthlyCharge
Bandwidth_GB_Year
Timely_Resp
Reliability
1.0
0.8
0.6
0.4
0.2
0.0







E1) Discuss the results of the hypothesis test.

Unfortunately, rejecting the null hypothesis is impossible as the p-value exceeds the standard significance point of 0.05 with a p-value of 0.8137137824222062. Therefore, whether there is a statistical significance between Churn and Reliability (equipment/services) that needs further exploration is unknown.

E2) Discuss the limitations of your data analysis.

As the p-value is at 0.8137137824222062, there is a need to explore the data further (or retrieve more data) about this issue.

E3) Recommend a course of action based on your results.

Based on these results, more analysis needs to be done to see if more variables are involved in customer Churn. This analysis looked at a small percentage of the data given; for example, the other "item1-8" columns may have more influence over customer churn than what was analyzed above. The company should investigate providing customers who do churn a survey with those eight items they surveyed previously to see which of the eight items was a significant player in the customer's choice to end their services with the company.

F) Panopto Video Link

G) Code Sources

Kaggle. (2018, May 01). Bivariate plotting with pandas. Kaggle. https://www.kaggle.com/residentmario/bivariate-plotting-with-pandas

Wikipedia. (2021, May 31). Bivariate Analysis. https://en.wikipedia.org/wiki/Bivariate_analysis#:~:text=Bivariate%20analysis%20is%20one%20of,the%20empirical%20relationship%20between%20them.&text=Like%20univariate%20analysis%2C%20bivariate%20analysis%20can%20be%20descriptive%20or%20inferential.

H) Web Sources

NIH. (2020). National Library of Medicine. https://www.nlm.nih.gov/nichsr/stats_tutorial/section2/mod11_significance.html#:~:text=In%20statistical%20tests%2C%20statistical%20significance,set%20to%200.05%20(5%25).

P-Values. (2020). StatsDirect Limited. https://www.statsdirect.com/help/basics/p_values.htm