# D207 Data Exploration

**Research Question**

Today we will continue working with our cleaned churn data, as we move into further exploration of the dataset. Customer churn has a big impact on our organization, and it cost the company more money to acquire new customers than it does to retain current customers. In our exploration we are looking to address customer churn. Is there a significant difference in the average monthly charge and bandwidth usage per year between customers who churned and those who did not, and is there an association between customer churn and the type of contract or internet service they use? The goal of this question is to provide statistical validation that supports our analysis and help us further investigate what factors are driving our churn rate. Through our analysis we can provide stakeholders with valuable data related to customer behavior, business trends related to key performance indicators(KPI)’s and identify opportunity areas to improve our customer retention. Stakeholders gain insight that can be used to drive business decisions.

**Required Variables**

Our dataset is comprised of 10,000 rows and 50 columns. I’ll give a description of each variable, it’s data type and an example from the dataset to provide insight. The column names and datatypes represent the cleaned dataset. The variables from the dataset that we will be focusing on today are ‘Churn’, InternetService’, ‘Contract’, ‘MonthlyCharge’ and Bandwidth\_GB\_Year.

|  |  |  |
| --- | --- | --- |
| **Column** | **Datatype** | **Example** |
| UID | string | Unique IDs related to customer transactions, technical support, and  sign-ups: e885b299883d4f9fb18e39c75155d990, f2de8bef964785f41a2959829830fb8a |
| CaseOrder | string | Order of each case: 1,2,3 |
| Customer\_id | string | Customers Unique ID: D90850, K191035 |
| Interaction | string | Unique Interaction ID: fb76459f-c047-4a9d-8af9-e0f7d4ac2524 |
| City | string | Customer’s city: Del Mar, Yamhill, Needville |
| State | string | Customer’s state: GA, OK, TN |
| County | string | Customer’s county: Peach, Scott, Oklahoma |
| Zip | string | Customer’s zip: 31030, 37847,73109 |
| Lat | float | Customer’s location in Latitude: 32.57032, 35.43313 |
| Lng | float | Customer’s location in Longitude: -133.37571, -83.8904 |
| Population | integer | Customer’s city population: 23144, 2535 |
| Area | string | Type of area: Rural, Urban, Suburban |
| Timezone | string | Timezone customer lives in: America/Chicago, America/New\_York |
| Job | string | Customer’s Job: Applications developer, Broadcast presenter |
| Children | float | Number of children: 0, 7, 2 |
| Age | float | Customer’s age: 50, 27, 83 |
| Education | string | Customer’s education level: Regular High School Diploma, Master’s Degree |
| Employment | string | Customer’s employment status: Retired, Student, Full Time |
| Income | float | Customer’s income: 21704.77, 28561.99 |
| Marital | string | Customer’s marital status: Widowed, Married, Never Married |
| Gender | string | Customer’s gender: Female, Male |
| Churn | string | Churn status: Yes, No |
| Outage\_sec\_perweek | float | Outages in seconds per week: 44.72520233, 8.15350077 |
| Email | integer | Number of emails: 20, 16, 10 |
| Contacts | integer | Number of contacts: 2,3,0 |
| Yearly\_equip\_failure | integer | Yearly equipment failures: 0, 1, 2 |
| Techie | string | Technical savvy status: NA, No, Yes |
| Contract | string | Contract type: Month-to-month, One year |
| Port\_modem | string | Port modem status: Yes, No |
| Tablet | string | Tablet owner: Yes, No |
| InternetService | string | Internet service type: Fiber optic, DSL, None |
| Phone | string | Phone service status: Yes, No |
| Multiple | string | Multiple lines status: Yes, No |
| OnlineSecurity | string | Online security status: No, Yes |
| OnlineBackup | string | Online backup status: Yes, No |
| DeviceProtection | string | Device protection status: Yes, No |
| TechSupport | string | Tech support status: Yes, No |
| StreamingTv | string | Streaming tv status: Yes, No |
| StreamingMovies | string | Streaming movies status: Yes, No |
| PaperlessBilling | string | Paperless billing status: Yes, No |
| PaymentMethod | string | Payment method: Electronic Check, Mailed Check, Credit Card(automatic) |
| Tenure | integer | Tenure in months: 9.563362997, 6.732948946 |
| MonthlyCharge | float | Monthly service charge: 142.3770231, 116.095233 |
| Bandwidth\_GB\_Year | float | Bandwidth GB usage yearly: 1324.330108 |
| Timely response | string | Timely response customer satisfaction ratings: 1-7 |
| Timely fixes | string | Timely fixes customer satisfaction ratings: 1-7 |
| Timely replacements | string | Timely replacements customer satisfaction ratings: 1-8 |
| Reliability | string | Reliability customer satisfaction ratings: 1-7 |
| Options | string | Options customer satisfaction ratings: 1-7 |
| Respectful response | string | Respectful response customer satisfaction ratings: 1-8 |
| Courteous exchange | string | Courteous exchange customer satisfaction ratings: 1-7 |
| Evidence of active listening | string | Evidence of active listening customer satisfaction ratings: 1-8 |

**Data Exploration Analysis**

I want to further analyze the relationship between our selected variables and how they may relate to customer churn. Using Python for my analysis I used .**describe().T**, **.skew()**, and **.kurtosis()** to get the univariate statistics for my **MonthlyCharge** and **Bandwidth\_GB\_Year** numerical variables. I also created histograms and boxplots to visual variable distribution. The monthly charges are distributed with a mean of $172.62. The interquartile range (Q3 - Q1) is $60.76, indicating that the middle 50% of the data is spread over this range. The distribution appears to have a reasonable spread, with the highest charge being $290.16. The bandwidth usage per year shows a wide range with a mean of 3392.34 GB.

The large standard deviation suggests high variability in data usage among customers. The interquartile range is 4349.68 GB, reflecting a significant spread in data usage. The maximum value of 7158.98 GB indicates that some customers have very high data consumption. The majority of customers have a month-to-month contract, which might indicate a preference for flexible service agreements. The distribution shows that fewer customers opt for longer-term contracts, with the two-year contract being more popular than the one-year contract. Fiber optic service is the most commonly used type of internet service among customers, followed by DSL. A significant portion of the customer base does not have internet service, which might include customers using the service for purposes other than internet connectivity. I then moved into bivariate statistical analysis of the two variables by creating a correlation matrix between the two variables and **Churn** along with bar plots to visualize how **MonthlyCharge** and **Bandwidth\_GB\_Year** affect **Churn.** The mean annual bandwidth usage for churned customers (1785.01 GB) is significantly lower than for non-churned customers (3971.86 GB). The distributions show that non-churned customers generally have higher bandwidth usage, with a clear separation between the two groups. The mean monthly charge for churned customers ($199.30) is significantly higher than for non-churned customers ($163.01). The distributions show that churned customers generally have higher monthly charges, with a noticeable separation between the two.

I then moved on to identifying the distribution of my two categorical variables, **Contract** and **InternetService**, using univariate and bivariate statistics. I got the frequency count and proportions using **.value\_counts()** and **.value\_counts(normalize=True)**. I then used bar plots to visualize the data for each variable. I created a contingency table for **Contract**, **InternetService**, and **Churn** and used a heatmap to visualize correlation and bar plot to compare each variable with **Churn** and visualize the data for bivariate distribution. Customers with month-to-month contracts have a significantly higher count of churn compared to those with one-year and two-year contracts. This suggests that shorter-term contracts might be associated with a higher likelihood of churn. Customers using Fiber Optic service have a higher count of churn compared to those using DSL or no internet service. This suggests that the type of internet service might influence the likelihood of churn, with Fiber Optic users being more likely to churn. I provided images below for our distribution findings.

**Univariate Statistics/Visualizations**

A close-up of numbers

Description automatically generated

A graph of a diagram

Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence

A white paper with black text

Description automatically generated

A blue rectangular bars with text

Description automatically generated with medium confidence

A blue rectangular object with text

Description automatically generated

**Bivariate Statistics/Visualizations**

A graph of different colored bars

Description automatically generated with medium confidence

A graph of blue and orange bars

Description automatically generated

A close-up of numbers

Description automatically generated

A diagram of a heatmap

Description automatically generated

A graph with blue and orange squares

Description automatically generated

A graph of a number of blue and orange squares

Description automatically generated

A close-up of a table

Description automatically generatedA close-up of a table

Description automatically generated

A screenshot of a graph

Description automatically generated

A diagram of a heatmap

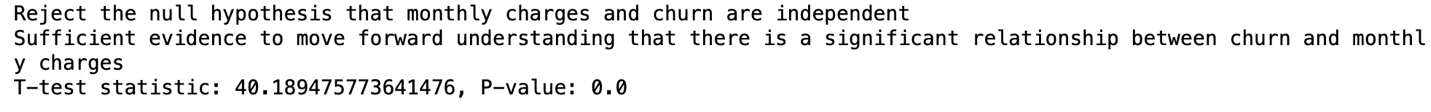
Description automatically generated

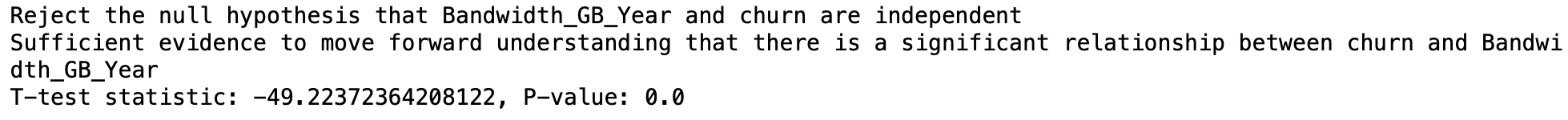
A close-up of numbers

Description automatically generated

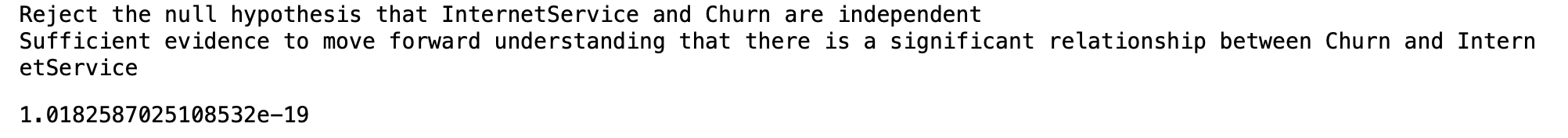
**Hypothesis Testing**

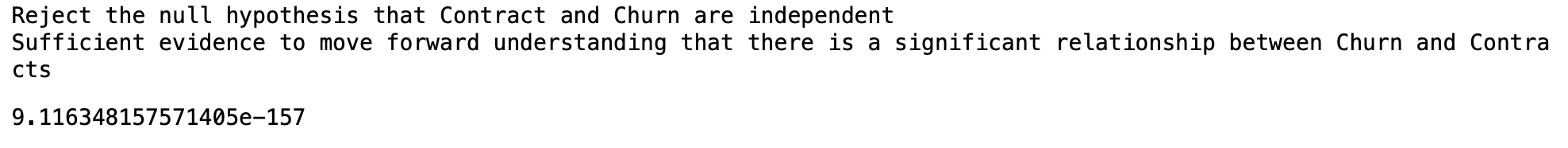
We have analyzed and visualized our data to this point and gained good insight, but we need to verify the statistical support of what we have found. We need to verify the significance of our chosen variables. This is where hypothesis testing comes in. We will use t-test for our numerical variables in this case because we are comparing their continues values between two independent groups, **Churn** and **Not Churn**. The Null Hypothesis for both variables is each is independent of the churn variable and the Alternate Hypothesis is there’s a relationship between the two variables and **Churn**.





Verifying the statistical significance of our categorical values we will use CHI Square testing. The CHI Square is used to verify if two categorical variables are independent of each other or there is an association between them. The Null and Alternate Hypothesis will be the same and we are looking at the P-value as our indicator of relevance.





Our hypothesis testing for each of our four variables reveal there is a significant relationship with **Churn** so we must reject the null hypothesis that each variable is independent of **Churn**.

**Summary**

The p-value is 0.0, which is much less than the common significance level of 0.05, we reject the null hypothesis. This indicates that there is a statistically significant difference in the mean**MonthlyCharge** between churned and not churned customers. The large t-statistic (40.189) further supports this conclusion, showing a substantial difference between the groups.

The significant difference in **MonthlyCharge**suggests that the cost might be influencing customer churn. Higher or lower charges could be associated with higher churn rates. The actionable plan could be reviewing pricing strategies. If higher charges lead to higher churn, implementing customer loyalty programs, discounts, or optimizing pricing might help reduce churn rates.

**Bandwidth\_GB\_Year** also had a p-value of 0.0, leading us to reject the null hypothesis. This indicates that there is a statistically significant difference in the mean **Bandwidth\_GB\_Year**between churned and not churned customers. The t-statistic is -49.224, suggesting a substantial negative difference between the groups. This indicates that the amount of bandwidth used annually is related to churn behavior. This could be due to dissatisfaction with the amount of bandwidth provided or differences in customer usage patterns.

An actionable plan here would be to investigate customer usage patterns and satisfaction with bandwidth offerings. Consider offering tailored plans or additional bandwidth to high-risk customers to improve retention.

The CHI Squared hypothesis test performed on **Contract** and **InternetService** both reject the null hypothesis that **Contract** and **InternetService** are independent of **Churn.** Different types of internet services (DSL, Fiber Optic, None) have different churn rates and the type of contract (Month-to-month, One Year, Two Year) is significantly associated with churn behavior. We should investigate why certain internet services have higher churn rates and focus on improving the quality of the internet service with higher churn or offering incentives to customers using that service. Analysis reveals longer-term contracts influence lower churn rates compared to month-to-month contracts. We should consider promoting longer-term contracts with incentives or improving the conditions of month-to-month contracts to make them more attractive and reduce churn.

**Limitations of Analysis**

The limitations of my analysis include uncontrolled confounders, other variables could be influencing the relationship between our analyzed variables. We may need to analyze other variables to make sure are conclusions are correct. Customer interactions with other variables in the dataset may be influencing our relationships. My analysis primarily identifies associations, not causation. The finding of significant relationships doesn't imply that one variable causes the other. Further analysis or experiments are needed to establish causation.

**Recommended Course of Action**

I recommend optimizing pricing by implementing tiered pricing and discounts for high paying customers. Ensure the pricing matches the value through enhanced services and benefits. Increase customer engagement by encouraging higher bandwidth usage through targeted content and promotions. Then provide incentives for active usage and educate customers on how to maximize their services. We should also offer attractive incentives for customers to switch from month-to-month to longer-term contracts. By focusing on optimizing the monthlycharge  while simultaneously enhancing engagement and promoting longer-term contracts, we can create a comprehensive strategy to effectively reduce churn and improve customer retention.

**References**

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy-1.6.3/reference/generated/scipy.stats.ttest\_ind.html.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html#skew.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html#kurtosis.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html#scipy.stats.chisquare.

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA) accessed 09 June 2024, https://apps.cgp-oex.wgu.edu.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/numpy/numpy\_array\_iterating.asp.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/pandas/default.asp.