­

**Exploratory Data Analysis: Churn Data**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. William Sewell

October 26, 2021

**Exploratory Data Analysis: Churn Data**

In this paper, I will use a data set containing cleaned customer data from a fictional telecommunications company. The primary purpose is to perform statistical analysis on the cleaned data set to determine which factor (or factors) are the greatest indicator of customer turnover.

# Part I: Research Question

What factor (or factors) are common among customers who have cancelled their service? By identifying patterns in customers who have terminated their service, organization stakeholders will be able to make changes in their company in order to 1) correct any flaws within the service model that may lead a customer to leave, and/or 2) identify trends in customers who have left and use that to target customers with similar profiles who may be at risk of leaving.

Relevant data to answer this question come from the Principal Component Analysis (PCA) I performed in the previous course, D206 Data Cleaning. I grouped the columns into each Principal Component (PC) as follows:

* PC1: **Customer Service** (Response, Fix, Replacement, Respectful, Courteous, Listening)
* PC2: **Location 1** (Zip, Lng)
* PC3: **Tenure/Bandwidth** (Tenure, Bandwidth\_GB\_Year)
* PC4: **Services** (Reliability, Options)
* PC5: **Location 2** (Lat, Population)
* PC6: **Family/Support** (Children, Age, Contacts)
* PC7: **Emails/Equipment** (Email, Yearly\_equip\_failure)
* PC8: **Finances** (Income, MonthlyCharge)

These values will be compared against the Churn Data (Yes/No) column in the data frame.

# Part II: Data Analysis

## Statistical Tests

I chose to perform both Wilcoxon Tests and ANOVA Tests based on data type; for qualitative data (such as those from the survey) I used ANOVA and for quantitative data (such as Tenure and Bandwidth\_GB\_Year) I used Wilcoxon Rank Sum as these tended to not follow normal distributions.

I began by creating a second data frame that contained numeric values for Yes (1) and No (0) in the Churn column so that I could perform the ANOVA operations. I addressed each group of variables according to the PCs outlined earlier.

I will provide an overview of my process and findings here. Please see the included *R* file: *churn\_data\_analysis.R* for the complete code.

## Principal Component 1

For PC1, I first visualized each variable based on the Yes/No value of Churn. Based on the visualizations in Figure 1 (above), there did not appear to be a strong indication that any of the variables in PC1 influence Churn. I then performed ANOVA tests for each variable to confirm this conclusion using p-values. The p-values found were not below the 0.05 standard threshold to reject the null hypothesis for each variable. (See Figure 1.1 and Figure 1.2 below).

**Figure 1.1**

*Histograms of PC1 Variables (Grouped by Churn)*

Chart, box and whisker chart

Description automatically generated

**Figure 1.2**

*Code Snippet: ANOVA Test & Summary of PC1 Variables*

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

## Principal Component 2

I also visualized the data for Zip and Lng (Longitude) based on Churn. These boxplots implied that Zip and Lng have no effect on Churn. I also performed an ANOVA test on Zip (discrete) and Wilcoxon Test on Lng (continuous) to confirm what the boxplots showed. The p-values for both are not below the 0.05 standard threshold. (See Figure 2.1 and Figure 2.2 below).

**Figure 2.1**

*Histograms of PC2 Variables (Grouped by Churn)*

Chart, box and whisker chart

Description automatically generated

**Figure 2.2**

*Code Snippet: ANOVA & Wilcoxon Tests of PC2 Variables*

Graphical user interface, text, application, email

Description automatically generated

## Principal Component 3

The boxplots for both Tenure and Bandwidth\_GB\_Year showed a statistically significant difference in distributions between customers who discontinued service and those that have stayed. After performing Wilcoxon Tests on both variables, the p-value of each was well below the 0.05 threshold. Both Tenure and Bandwith\_GB\_Year have a significant effect on whether a customer chooses to discontinue their service. (See Figure 3.1 and Figure 3.2 below).

**Figure 3.1**

*Histograms of PC3 Variables (Grouped by Churn)*

Chart, box and whisker chart

Description automatically generated

**Figure 3.2**

*Code Snippet: Wilcoxon Tests of PC3 Variables*

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

## Principal Component 4

The boxplots for both Reliability and Options did not appear to indicate a significant difference between groups. The ANOVA test showed p-values that were well above the 0.05 threshold confirming what the plots showed. (See Figure 4.1 and Figure 4.2 below).

**Figure 4.1**

*Histograms of PC4 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 4.2**

*Code Snippet: ANOVA Tests of PC4 Variables*

*Graphical user interface, text, application

Description automatically generated*

## Principal Component 5

The boxplots for both Lat (Latitude) and Population showed no statistically significant difference in distributions between customers who discontinued service and those that have stayed. The Wilcoxon Test on Latitude (continuous) and ANOVA test on Population (discrete) confirmed this with p-values well above the 0.05 standard threshold. (See Figure 5.1 and Figure 5.2 below).

**Figure 5.1**

*Histograms of PC5 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 5.2**

*Code Snippet: Wilcoxon & ANOVA Tests of PC5 Variables*

*Graphical user interface, text, application

Description automatically generated*

## Principal Component 6

The boxplots for Children, Age, and Contacts did not show a statistically significant difference in distributions between customers who discontinued service and those that have stayed. The ANOVA tests performed on each (all discrete variables) confirmed what the histograms implied. (See Figure 6.1 and Figure 6.2 below).

**Figure 6.1**

*Histograms of PC6 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 6.2**

*Code Snippet: ANOVA Tests of PC6 Variables*

*Graphical user interface, text

Description automatically generated*

*Graphical user interface, text

Description automatically generated*

## Principal Component 7

The boxplots for both Email and Yearly\_Equip\_Failure did not show a difference between Churn groups, and this was confirmed with ANOVA tests on both variables. (See Figure 7.1 and Figure 7.2 below).

**Figure 7.1**

*Histograms of PC7 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 7.2**

*Code Snippet: ANOVA Tests of PC7 Variables*

*Graphical user interface, text, application

Description automatically generated*

*Graphical user interface, text, application

Description automatically generated*

## Principal Component 8

While the boxplot for Income did not show a difference in Churn groups, the MonthlyCharge Tenure did show a statistically significant difference. These conclusions were confirmed in the ANOVA and Wilcoxon tests for the two variables (respectively). (See Figure 8.1 and Figure 8.2 below).

**Figure 8.1**

*Histograms of PC8 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 8.2**

*Code Snippet: ANOVA & Wilcoxon Tests of PC8 Variables*

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

# Part III: Conclusions

## Analysis of Variables of Interest

The variables identified in the analysis above were Tenure, Bandwidth\_GB\_Year, and MonthlyCharge. To provide further context to how these values interact with on anther based on Churn, the following visualizations were performed. (See Figure 9, Figure 10, Figure 11, and Figure 12 below).

The histograms in Figure 9 revealed a relatively normal distribution for MonthlyCharge and similarly shaped bimodal distributions for Tenure and Bandwidth\_GB\_Year. This company apparently loses customers around the 24-month (two-year) mark; however, there are still many that have been with the company more than 60 months (five years). This leads me to wonder what happened around two years ago. Several options come to mind:

* Did the company offer a new-customer discount that expired after two years?
* Did a new company enter the market that offered discounted or new services that would draw customers away?
* Did something happen that caused customers to lose faith in the company (such as faulty service or a change in leadership to an unknown/unproven candidate)?

After analyzing the scatterplots, there appears to be a high concentration of customers who left the company with a high Monthly Charge and low Bandwidth usage (Figure 10) and high Monthly Charge and low Tenure (Figure 11). The logical conclusion would be that customers who are paying a high price for services they are not using and/or who have not been with the company for a long time are at high risk of leaving. The final scatterplot shows a strong positive relationship between Tenure and Bandwidth (Figure 12). The longer a customer is with this company, the greater the bandwidth they consume. This pattern is true for both former and current customers; however, it appears that customers who left tended to consume more bandwidth.

Based on the analysis of these three variables and the graphs, the greatest factor that is causing a customer to cancel their service is the cost of high bandwidth consumption. A review of the current market rate for this service is recommended and, should the rate for this company be notably higher than the market average, company leadership should consider reducing the price to match.

## Limitations on Analysis

No statistical test is perfect, and I restricted my analysis to only the numeric variables identified in my Principal Component Analysis from the previous course. Therefore, there could be other factors that were overlooked or confounders that were not identified. However, the conclusion based on Tenure, Bandwidth, and Monthly Charge seem strong and worth pursuing.

**Figure 9**

*Histograms of Tenure, Bandwidth\_GB\_Year, and MonthlyCharge*

Chart, bar chart, histogram

Description automatically generated

**Figure 10**

*Scatterplot of Bandwidth\_GB\_Year vs MonthlyCharge, Color-coded by Churn*

*Chart, scatter chart

Description automatically generated*

Note: The alpha value is for the opacity of each point.

**Figure 11**

*Scatterplot of Tenure vs MonthlyCharge, Color-coded by Churn*

*Chart, scatter chart

Description automatically generated*

Note: The alpha value is for the opacity of each point.

**Figure 12**

*Scatterplot of Bandwidth\_GB\_Year vs Tenure, Color-coded by Churn*

*Chart

Description automatically generated*

Note: The alpha value is for the opacity of each point.

# Part IV: Supporting Documents

Please find the code file, Panapto recording with my submission, and the references below.

**References**

Jaap & BenBarnes (2012, March 23). How to get summary statistics by Group. Stack Overflow. Retrieved October 28, 2021, from <https://stackoverflow.com/questions/9847054/how-to-get-summary-statistics-by-group>

Kassambara, A. (2017, September 1). GGPLOT2 - easy way to mix multiple graphs on the same page. STHDA. Retrieved October 28, 2021, from <http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/>

One-way ANOVA test in R. STHDA. (n.d.). Retrieved October 28, 2021, from <http://www.sthda.com/english/wiki/wiki.php?title=one-way-anova-test-in-r>

Unpaired two-samples Wilcoxon test in R. STHDA. (n.d.). Retrieved October 27, 2021, from <http://www.sthda.com/english/wiki/unpaired-two-samples-wilcoxon-test-in-r>