The scenario I chose for this task was the telecommunications industry scenario. As stated in the data dictionary, it costs 10 times more to acquire new customers than maintaining current customers. With that being said, stakeholders wanted to understand what customers are at a high risk of churn. My question for this analysis is, “What factors related to customers are related to churn. The goal of this analysis is to segment the data, so I can understand what factors are closely related to churn.

For this scenario, I will use hierarchical clustering. I decided to use this clustering technique as hierarchical clustering does not require me to set the number of clusters and instead provides a dendrogram that can help me with making the best decision for a more accurate model. An assumption of this clustering technique is that it will be a higher computational cost, than k-means. This computational cost will be worth it, as it will create a stronger model (Data & Analytics, 2023).

There were different packages and libraries I used. As usual I used pandas and numpy as they provide many basic functions used in data science. I chose to use matplotlib.pylot and seaborn for visualizations. I utilized sklearn.preprocessing for standardization, sklearn.cluster for my hierarchical agglomerative clustering (*Plot Hierarchical Clustering Dendrogram*, n.d.), and sklearn.manifold for t-SNE. T-sne will preserve the structure of the data, but be helpful in visualizing the pattern of the data (GeeksforGeeks, 2025). Lastly, to further add to my hierarchical analysis I used scipy.cluster.hierarchy to create my dendrogram, linkage, and the fcluster command (Joern, 2015).

After creating my dataframe, it was time for data preprocessing. I noticed right away there were null values in my internet service column. With maintaining 7871 rows of data and there appearing to be no relationship in the missing data to other datapoints, I decided it was best to drop these null rows. This way there was no noise created trying to impute this data, with an accurate, significant data to analyze. After dropping redundant or irrelevant variables, I was left with binary, ordinal, nominal, and continuous variables. My ordinal variables were Item1-Item8. My continuous variables were population, children, age, income, outage seconds per week, number of emails, contacts, yearly equip failures, tenure, monthly charge, and bandwidth GB per year. My binary variables were churn, techie, port modem, tablet, phone, multiple, online security, online backup, device protection, tech support, streaming tv, streaming movies, and paperless billing. Lastly, my nominal variables were gender, marital, job, area, contract, internet service, and payment method. There were a few steps I took to prepare the data. This included removing redundant and irrelevant data, handling null values by dropping them, encoding my variables accordingly, and standardizing my continuous variables.

After creating my dendrogram, I was aware of how optimal clusters were. As screenshotted below, we can see 2 clusters were optimal. The method I chose was utilizing the dendrogram and cutting off after a huge jump in distance. This could be seen at about 120, which was used for the cutoff (265 Data Science, 2019). This also aligned with what I was seeing previously when plotting the data via t-SNE.

A diagram of a city

AI-generated content may be incorrect.

A yellow and purple dots

AI-generated content may be incorrect.

As seen above we could see the two separate components easily with their colors separating them. The results of these clusters showed that one cluster had about 7.31% churn, while the other one had 44.12% churn. We could tell then which segment was more attached to churn. Using that cluster, I created a correlation matrix to further understand what is related to churn. These results can be seen below. Monthly charges, if the customer streamed movies or tv, and whether they had multiple lines were all connected to churn. What we can assume is that the higher the monthly charges (that would be typically related to streaming movies, streaming tv, and multiple lines), the more likely the customer would churn. I suggest looking into ways to reduce the monthly charges for customers, so we can maintain our highest users. I feel it is also important to clarify that while we understand what is connected to churn the most, we could further analyze this data by seeing what factors are most likely to cause churn by using a supervised model like a random forest. The limitation to this model is that we can see what in this segment is tied to churn, but not precisely what is causing it. I suggest further analysis to specifically understand what causes churn.

A screenshot of a computer program

AI-generated content may be incorrect.

References:

365 Data Science. (2019, January 15). *Flat and hierarchical clustering | the Dendrogram explained* [Video]. YouTube. <https://www.youtube.com/watch?v=ijUMKMC4f9I>

Data & Analytics. (2023, November 7). *K-means Clustering vs. Hierarchical Clustering: Clash of Clusters* [Video]. YouTube. <https://www.youtube.com/watch?v=Lj6f_Jd0ZXA>

GeeksforGeeks. (2025b, February 10). *ML | TDistributed Stochastic Neighbor Embedding (TSNE) Algorithm*. GeeksforGeeks. <https://www.geeksforgeeks.org/ml-t-distributed-stochastic-neighbor-embedding-t-sne-algorithm/>

Joern. (2015, August 26). *SciPY Hierarchical Clustering and Dendrogram tutorial | Jörn’s blog*. <https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/>

*Plot Hierarchical Clustering Dendrogram*. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html>

Sachinsoni. (2024, February 11). *Mastering t-SNE(t-distributed stochastic neighbor embedding).* Medium. <https://medium.com/@sachinsoni600517/mastering-t-sne-t-distributed-stochastic-neighbor-embedding-0e365ee898ea>