DATA ANALYTICS

Customer Churn in Telecom Industry

Lucie Stenger

October, 2024

Report structure

Use Case

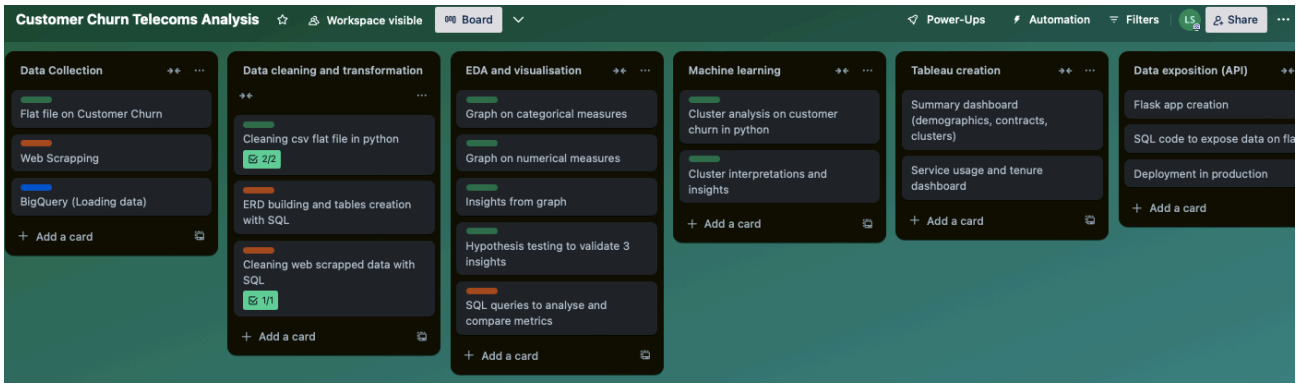
The telecommunications industry is facing significant challenges with customer retention. A recent survey, the *2022 State of Customer Churn in Telecom* revealed a dramatic 22% decrease in customer loyalty post-pandemic. Now, more than ever, customers prioritize experience over brand loyalty, and rising concerns about pricing further complicate the situation, with 58% of customers viewing telecom services as overpriced.

This report aims to analyse customer churn—the rate at which customers leave a telecom provider—and explore key factors driving this trend. Nearly 25% of customers indicate a strong likelihood of switching providers within the next few months or when their current contracts expire.

The average churn rate for telecom companies is alarmingly high, sitting between 30% and 35% compared to the 12-15% churn rate in broader utilities sectors across Western Europe. Reducing churn is critical to maintaining a stable customer base, and this report seeks to uncover how data analysis and machine learning can help telecom companies identify patterns and predict churn before it happens.

By leveraging machine learning models, telecom companies can better understand the causes of churn and develop targeted strategies to improve customer experience, optimize pricing, and ultimately reduce churn rates. This report will explore these possibilities and provide actionable insights for customer retention in the telecom sector using a synthetic data set of customer information from a telecom company, and web-scraped phone and internet plans information from a online North American carrier comparator.

Project Planning



Data sources and data collection

1. Web Scraping

For the SQL part, we are looking at the data scraped from www.whistleout.com, an online carrier comparator. It lists telecom plans, services, and pricing from multiple telecom companies, which are key to understanding churn dynamics.

2. Flat File

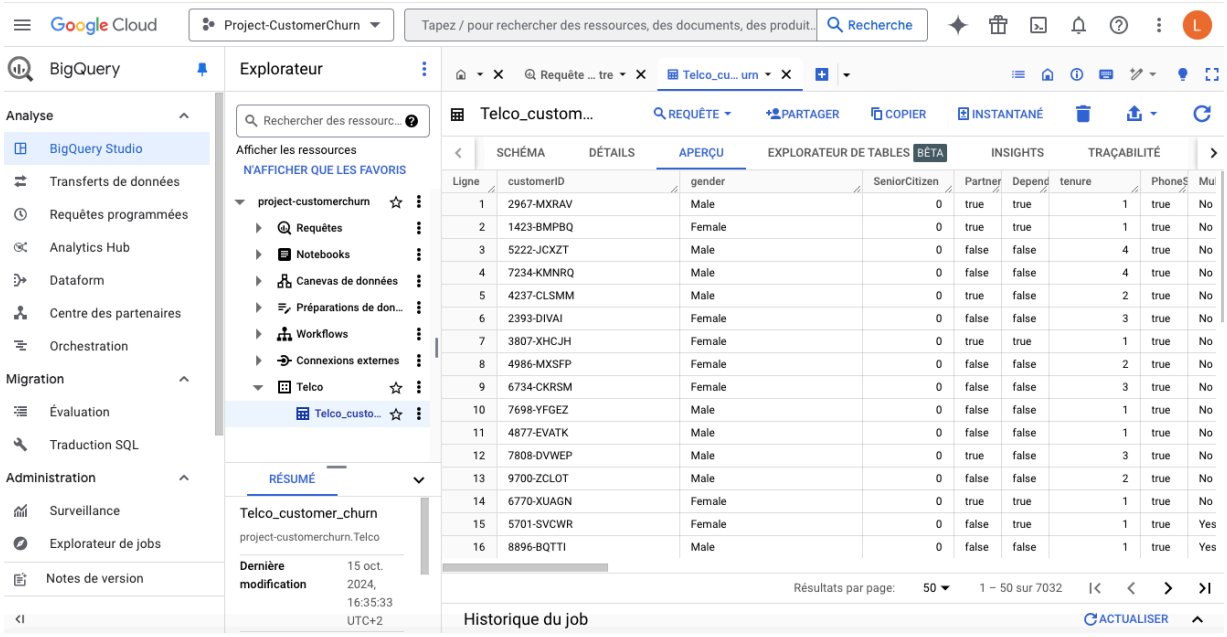
We used a .csv file from www.kaggle.com. This dataset contains customers' information from a fictive telecom company called Telco, including customer churn.

3. Database (store your data in relational database - MySQL)

Information collected both from www.whistleout.com and from www.kaggle.com were stored in a relational database created in MySQLWorkbench.

4. Big Data System

We loaded our table containing customer information into BigQuery to enable efficient handling of large datasets. This ensures scalability for our project, allowing us to perform fast queries and analysis on extensive data without performance limitations.



Data Dictionary

Web Scraping (www.whistleout.com)

Dataset Details

Dataset Name: Telecoms_deals

Description: The dataset contains four tables that provide detailed information on various telecom plans based in North America. The Cellphone_Plan_ID table includes information on individual cell phone plans, such as carrier, price, and best use case. The bundled_plans table details bundled mobile and internet plans with similar features. The internet_plans table focuses on carriers' internet offerings, listing the number and categories of internet plans available. Lastly, the tmobile_plans table outlines T-Mobile's subscription plans, detailing the number of lines, price per line, and total monthly costs.

Table Names:

- Cellphone_Plan_ID

Feature	Description
Cellphone_Plan_ID	A unique ID that identifies each cell-phone plan.
Carrier	Indicates carrier’s name.

Plan	Indicates plan's name.
Price	Indicates plan's price.
Best_For	Indicates advantages of particular plan.
Plan_Link	Website link to get more information on the plan.
Carrier_ID	A unique ID that identifies each carrier/provider.
Cleaned_Price	Price column stripped from extra words and symbols, making it possible to run aggregations and other calculation functions.

- bundled_plans

Feature	**Description**
Bundle_ID	A unique ID that identifies each bundle (mobile and internet) plan.
Carrier	Indicates carrier's name.
Bundle_Plan	Indicates plan's name.
Price	Indicates plan's price.
Best_For	Indicates advantages of particular plan.
Plan_Link	Website link to get more information on the plan.
Carrier_ID	A unique ID that identifies each carrier/provider.
Cleaned_Price	Price column stripped from extra words and symbols, making it possible to run aggregations and other calculation functions.

- internet_plans

Feature	**Description**
Carrier_ID	A unique ID that identifies each carrier/provider.
Carrier	Indicates carrier's name.
Plans_Available	Indicates the number of plans offered by carrier.
Internet_Categories	Indicates categories of internet services available, such as Fiber, Cable, Satellite, VDSL, DSL, Mobile Broadband.

- tmobile_plans

Feature	**Description**
Plan_ID	A unique ID that identifies each plan.
Plan	Indicates plan's name.
Line_Number	Indicates number of lines.
Price_Per_Line	Indicates price per line for a month of subscription.

Total_Monthly_Price	Indicates total price, all lines included, for a month of subscription
Carrier_ID	A unique ID that identifies each carrier/provider.
Cleaned_Price_Per_Line	Price per line column stripped from extra words and symbols, making it possible to run aggregations and other calculation functions.
Cleaned_Total_Monthly_Price	Total monthly price column stripped from extra words and symbols, making it possible to run aggregations and other calculation functions.

Kaggle TELCO churn file

Dataset Details

Dataset Name: Telco Customer Churn

Description: This dataset provides information about customers from a telecom company, including customer demographics, account information, services subscribed (e.g., internet, phone), and whether they churned. It includes categorical and numerical features.

Feature	**Description**
CustomerID	A unique ID that identifies each customer.
Gender	The customer's gender: Male, Female.
Senior Citizen	Indicates if the customer is 65 or older: 1 if Yes, 0 if No.
Partner	Indicates if the customer has a partner: Yes, No.
Dependents	Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
Tenure	Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.
Phone Service	Indicates if the customer subscribes to home phone service with the company: Yes, No.
Multiple Lines	Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No, No phone service.
Internet Service	Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic.

Online Security	Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No, No internet service.
Online Backup	Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No, No internet service.
Device Protection	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No, No internet service.
Tech Support	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No, No internet service.
Streaming TV	Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No, No internet service. The company does not charge an additional fee for this service.
Streaming Movies	Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No, No internet service. The company does not charge an additional fee for this service.
Contract	Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.
Paperless Billing	Indicates if the customer has chosen paperless billing: Yes, No.
Payment Method	Indicates how the customer pays their bill: Bank transfer (automatic), Credit card (automatic), Electronic check, Mailed check.
Monthly Charge	Indicates the customer's current total monthly charge for all their services from the company.
Total Charges	Indicates the customer's total charges, calculated to the end of the quarter specified above.
Cluster	Column added post-clustering analysis, using machine learning to group customers into 3 cluters (0,1,2).
Carrier_ID	Column added for ERD and for SQL analysis.

Data Cleaning

SQL cleaning on web-scraped data

In this data-cleaning process, I have focused on standardising the pricing information in the bundled_plans, cell_phone_plans, and tmobile_plans tables. Here's a summary of the steps taken:

1. Added Cleaned Price Columns

For each of the tables, I created new columns (Cleaned_Price, Cleaned_Price_Per_Line, and Cleaned_Total_Monthly_Price) to store the cleaned and standardised numeric values of the prices. These columns are of type DECIMAL(10, 2), which allows the storage of prices as decimal numbers with two decimal places.

2. Removed Non-Numeric Characters and Text

In each pricing column, there were non-numeric characters such as dollar signs ($), text descriptions like "per line," "per month," and extraneous words like "Starting at" or "line." These characters and words were removed using SQL functions like REPLACE(), TRIM(), and SUBSTRING_INDEX().

In cases where the price was embedded with ranges (e.g., "$88-176/month"), I cleaned out the additional numbers and symbols to extract a clear price that was consistant with the rest of the table's information.

The cleaned values were then inserted into the newly added columns.

3. Ensured Safe Updates

Since SQL Safe Updates mode restricts updates to tables without using keys, I disabled it temporarily (SET SQL_SAFE_UPDATES = 0) to allow the cleaning process to run. After performing the updates, I re-enabled Safe Updates to ensure data integrity moving forward (SET SQL_SAFE_UPDATES = 1).

As a result of this cleaning process, we now have clearly defined numerical values for the price-related columns across all relevant tables. This allows for more accurate data analysis, comparison, and aggregation of the prices in each of the plans.

Python cleaning on Kaggle flat file

Telco customer churn flat file available on Kaggle.com was fortunately clean and relatively ready for EDA and deeper data analysis.

1. Converted and removed invalid values

The data cleaning process simply consisted in converting the 'TotalCharges' column to a numeric format, handling any non-numeric or missing values by removing rows with invalid data.

2. Categorized numerical features

I also categorized key numerical features—'tenure', 'MonthlyCharges', and 'TotalCharges'—into defined ranges to group customers based on their time with the service, monthly billing, and total charges. This allowed for an easier visualisation of all categorical measures using bar charts.

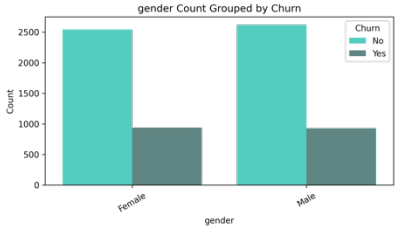
Exploratory Data Analysis (EDA)

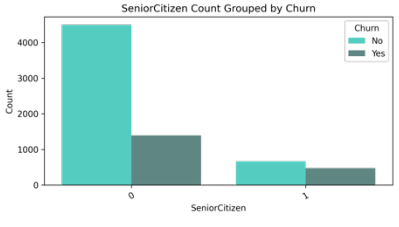
First interpretations

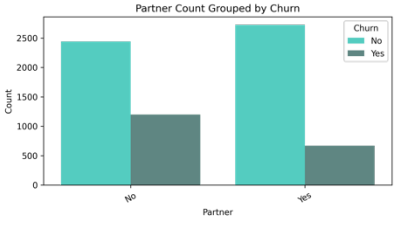
Clients are more likely to churn if:

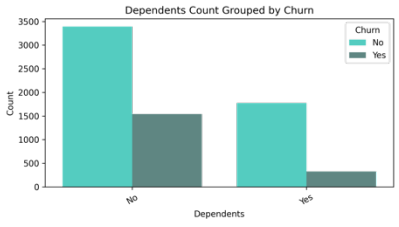
- They have no dependent and no partner
- They are a senior citizen (older than 65y)
- They are recent customers (less than 1 year)
- They have multiple lines
- They have fiber optic Internet service
- They subscribe to Internet but do not subscribe to Online security, Online backup, Device protection, or Tech support services
- Their contract is month-to-month
- They pay via electronic check

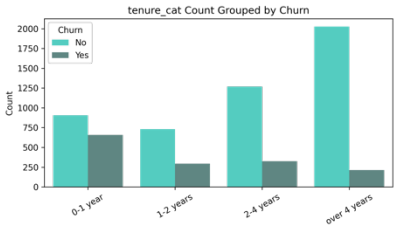
Visualisations Using Python

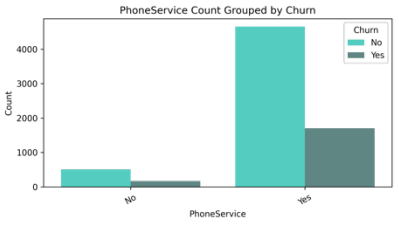


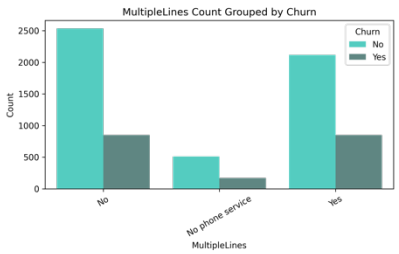


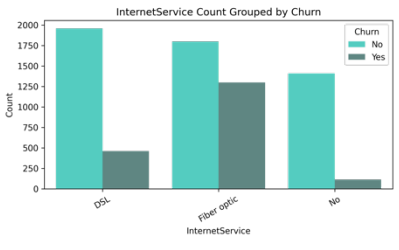


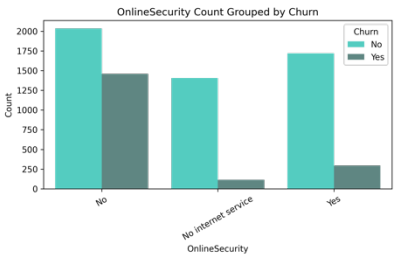


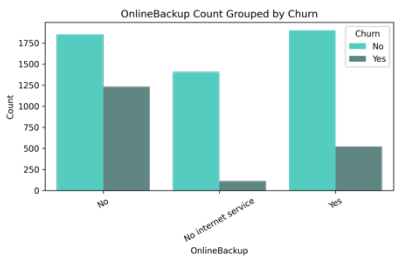


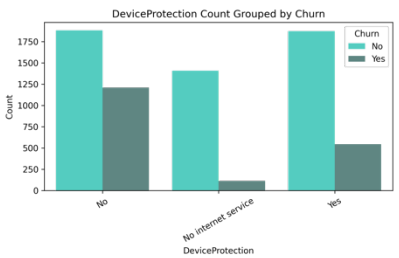


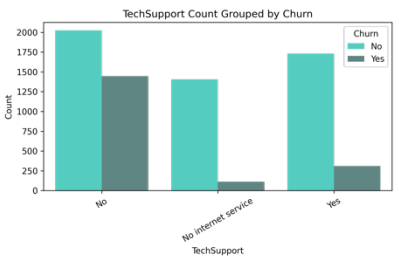


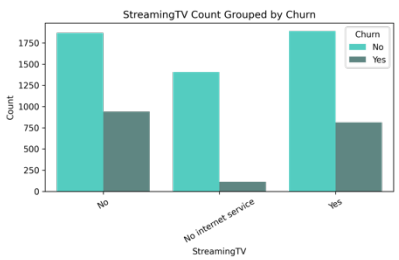


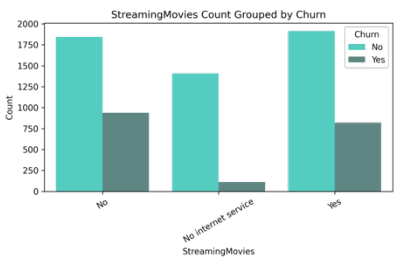


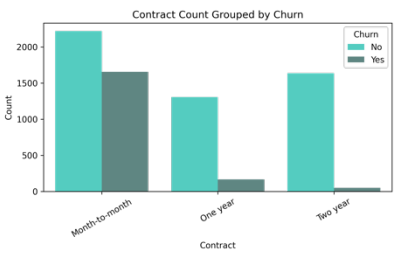


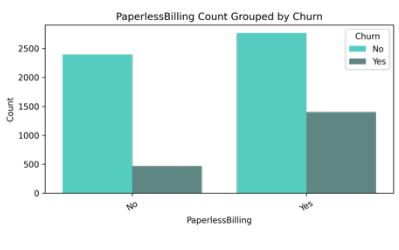


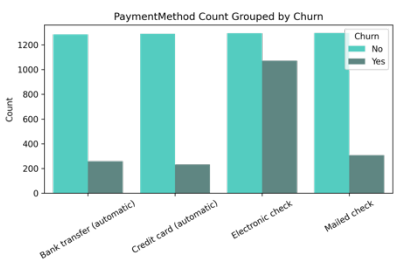


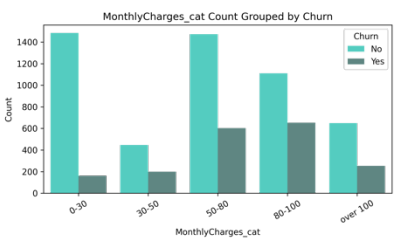


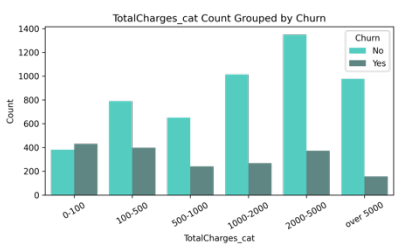


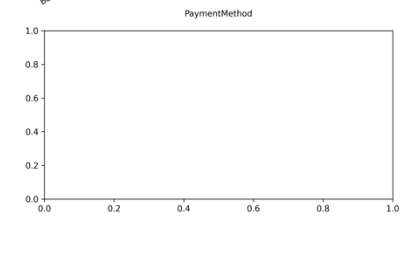


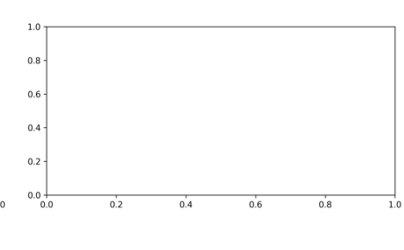


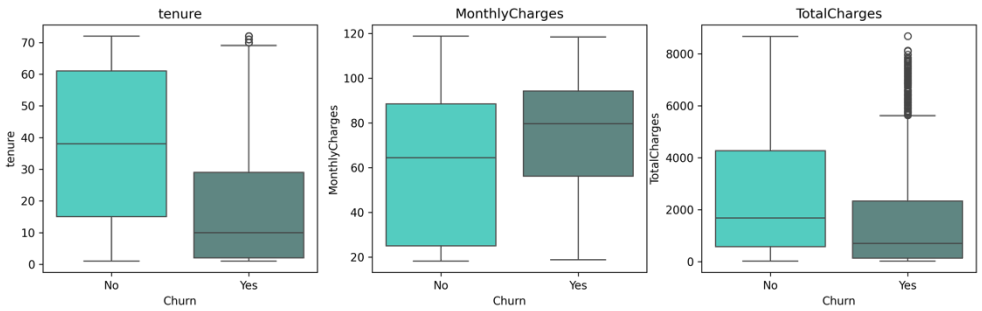












Hypothesis Testing

1. Customers who have been using Telco's services for over one year are less likely to churn.

Null Hypothesis (H0): Customers who have been using Telco's services for over one year are equally likely to churn as those who have been using it for less than or equal to one year.

Alternative Hypothesis(H1): Customers who have been using Telco's services for over 1 year are less likely to churn.

Chi-Square Statistic: 708.78

P-Value: 0

Using the chi-square test of independence, we reject the null hypothesis.

2. Customers with month-to- month contracts are more likely to churn.

Null Hypothesis (H0): Customers with month-to-month contracts are equally likely to churn as those with other types of contracts.

Alternative Hypothesis (H1): Customers with month-to-month contracts are more likely to churn.

Chi-Square Statistic: 1153.97

P-Value: 0

Using the chi-square test of independence, we reject the null hypothesis.

3. Senior citizens are more likely to churn than non-senior citizens.

Null Hypothesis (H0): Senior citizens are equally likely to churn as non- senior citizens.

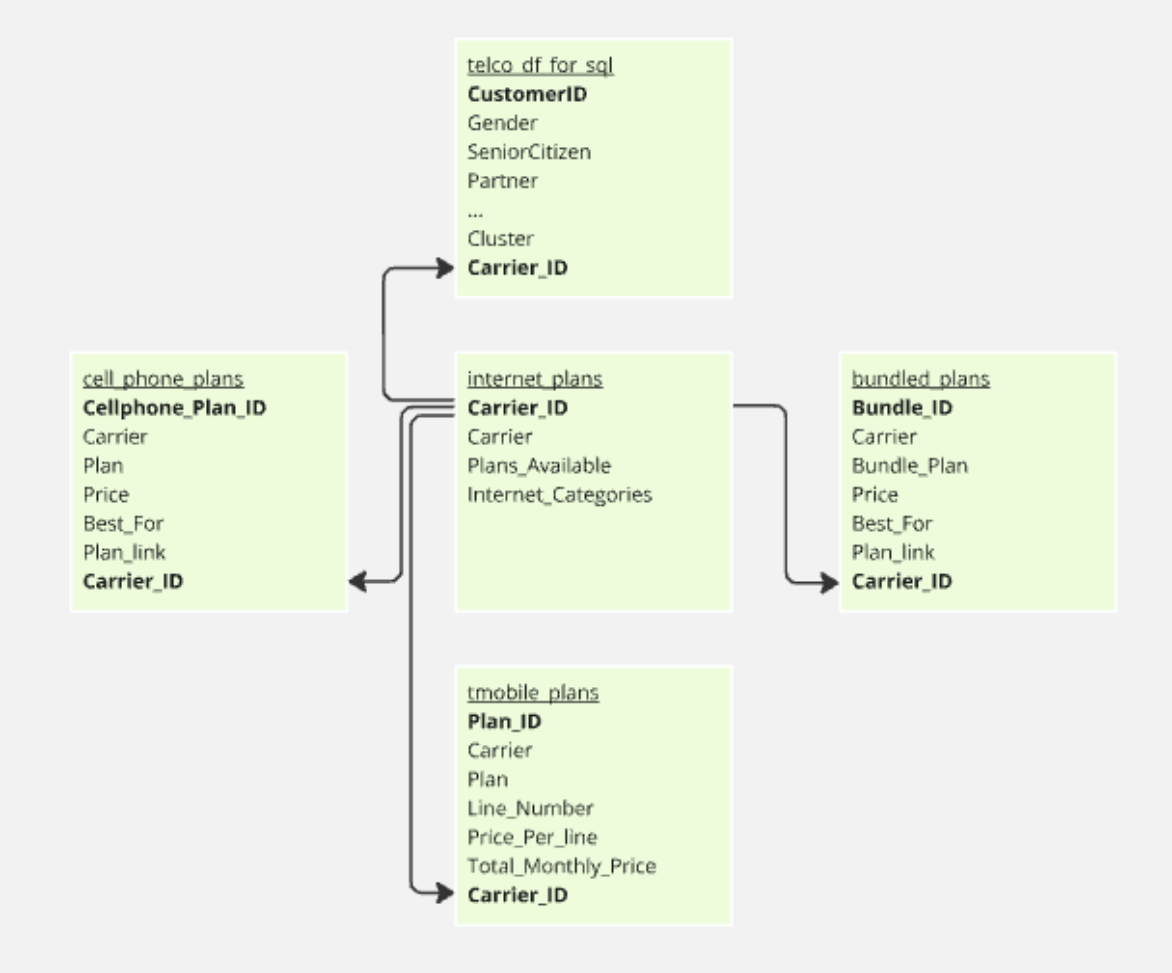
Alternative Hypothesis (H1): Senior citizens are more likely to churn than non-senior citizen

Z-**Statistic: 12.663**

P-Value: 0

Using the proportions z-test, we reject the null hypothesis.

Entity-Relationship Diagram (ERD)



Data Insights using SQL

```sql
-- Churn rate for customers with bundle (phone service + internet service) option
SELECT
    (SUM(CASE WHEN Churn = 'Yes' THEN 1 ELSE 0 END) * 100.0 / COUNT(*)) AS Churn_Rate
FROM
    telco_df_for_sql
WHERE
    PhoneService = 'Yes'
    AND InternetService != 'No';
-- Churn rate for customers with only internet service option
SELECT
    (SUM(CASE WHEN Churn = 'Yes' THEN 1 ELSE 0 END) * 100.0 / COUNT(*)) AS Churn_Rate
FROM
    telco_df_for_sql
WHERE
    PhoneService = 'No'
    AND InternetService != 'No';
-- Churn rate for customers with only phone service option
SELECT
    (SUM(CASE WHEN Churn = 'Yes' THEN 1 ELSE 0 END) * 100.0 / COUNT(*)) AS Churn_Rate
FROM
    telco_df_for_sql
WHERE
    PhoneService = 'Yes'
    AND InternetService = 'No';
```

In our synthetic telco customer data, the churn rate for customers with bundle services is higher than for phone service alone or for internet service alone.

This might be because customers will spend more with specific provider, and therefore save more when switching to a cheaper provider.

```sql
SELECT AVG(bp.Cleaned_Price) AS Bundle_Average_Price, AVG(cp.Cleaned_Price) AS Cellphone_Average_Price
FROM bundled_plans AS bp, cell_phone_plans AS cp;
```

```sql
SELECT
    AVG(CASE WHEN PhoneService = 'Yes' AND InternetService = 'No' THEN MonthlyCharges
            ELSE NULL END) AS Average_Price_phone,
    AVG(CASE WHEN PhoneService = 'No' AND InternetService != 'No' THEN MonthlyCharges
            ELSE NULL END) AS Average_Price_internet,
    AVG(CASE WHEN PhoneService = 'Yes' AND InternetService != 'No' THEN MonthlyCharges
            ELSE NULL END) AS Average_Price_bundle
FROM
    telco_df_for_sql;
```

The average bundle price in our telco company exceeds the combined average price of standalone phone and internet services. However, in our web-scraped data from real telecommunications companies, the price difference between bundle services and phone service alone is much smaller, indicating that these companies offer significant savings on bundles. This suggests that our telco company either does not offer bundle discounts or provides only minor reductions, which may contribute to the higher churn rate among customers using both internet and phone services, especially when competitors provide more attractive bundle deals.

```sql
-- Churn rate for customers with multiple phone lines
SELECT
    (SUM(CASE WHEN Churn = 'Yes' THEN 1 ELSE 0 END) * 100.0 / COUNT(*)) AS Churn_Rate
FROM
    telco_df_for_sql
WHERE
    MultipleLines = 'Yes';
-- Churn rate for customers with only one phone line
SELECT
    (SUM(CASE WHEN Churn = 'Yes' THEN 1 ELSE 0 END) * 100.0 / COUNT(*)) AS Churn_Rate
FROM
    telco_df_for_sql
WHERE
    MultipleLines = 'No';
```

In our synthetic telco customer data, the churn rate for customers subscribing to multiple phone lines is slightly higher than for customers subscribing to only one phone line.

When looking at T-Mobile family plans, a company offering savings on multiple lines, it appears clear that this strategy is a way to reduce churn rate.

It seems even more relevant when analysing the total cost for 2 lines versus 3 lines in various phone plans:

```sql
SELECT
    Plan,
    Line_Number,
    Cleaned_Total_Monthly_Price,
    Cleaned_Total_Monthly_Price - LAG(Cleaned_Total_Monthly_Price) OVER (PARTITION BY Plan ORDER BY Cleaned_Line_Number)
    AS Price_Difference_Per_New_Line
FROM
    tmobile_plans
ORDER BY
    Plan,
    Line_Number;
```

We can see that for both entry-level family plans (essentials and Go5G) there is no price difference in total spent between 2 lines and 3 lines.

We can interpret this as the following: customers churn more when adding a third line to their subscription, probably looking at competitors before making this choice.

If the first two lines in a family plan are typically subscribed by parents, the third line might be for a child, prompting a reassessment of their carrier's suitability in terms of price and features. By maintaining the same cost for 2 or 3 lines, carriers make it easier for families to stay with the same provider, increasing the likelihood of adding more lines in the future.

Exposing Stored Data via API (flask)

Portal content

Our flask web application includes two main routes: one for fetching carrier information based on a specific carrier_id and another for retrieving customer data within a specified cluster while considering GDPR compliance. The "/carrier/<int:carrier_id>" route queries the internet_plans table in a MySQL database, returning details of the requested carrier or a 404 error if not found. The "/Telco/cluster "route enables pagination and can filter data based on the GDPR flag, allowing either all customer details or a limited set of features not impacted by GDPR. The application connects to a MySQL database, handles SQL queries, and returns the results in JSON format. Proper error handling and resource management are implemented by closing the database connection after queries are executed. Finally, the application runs in debug mode for easier troubleshooting during development.

To run our app , we typed the below on the terminal:

- flask --app notebooks/flask_app run --port 8080 --debug

http://localhost:8080/carrier/3

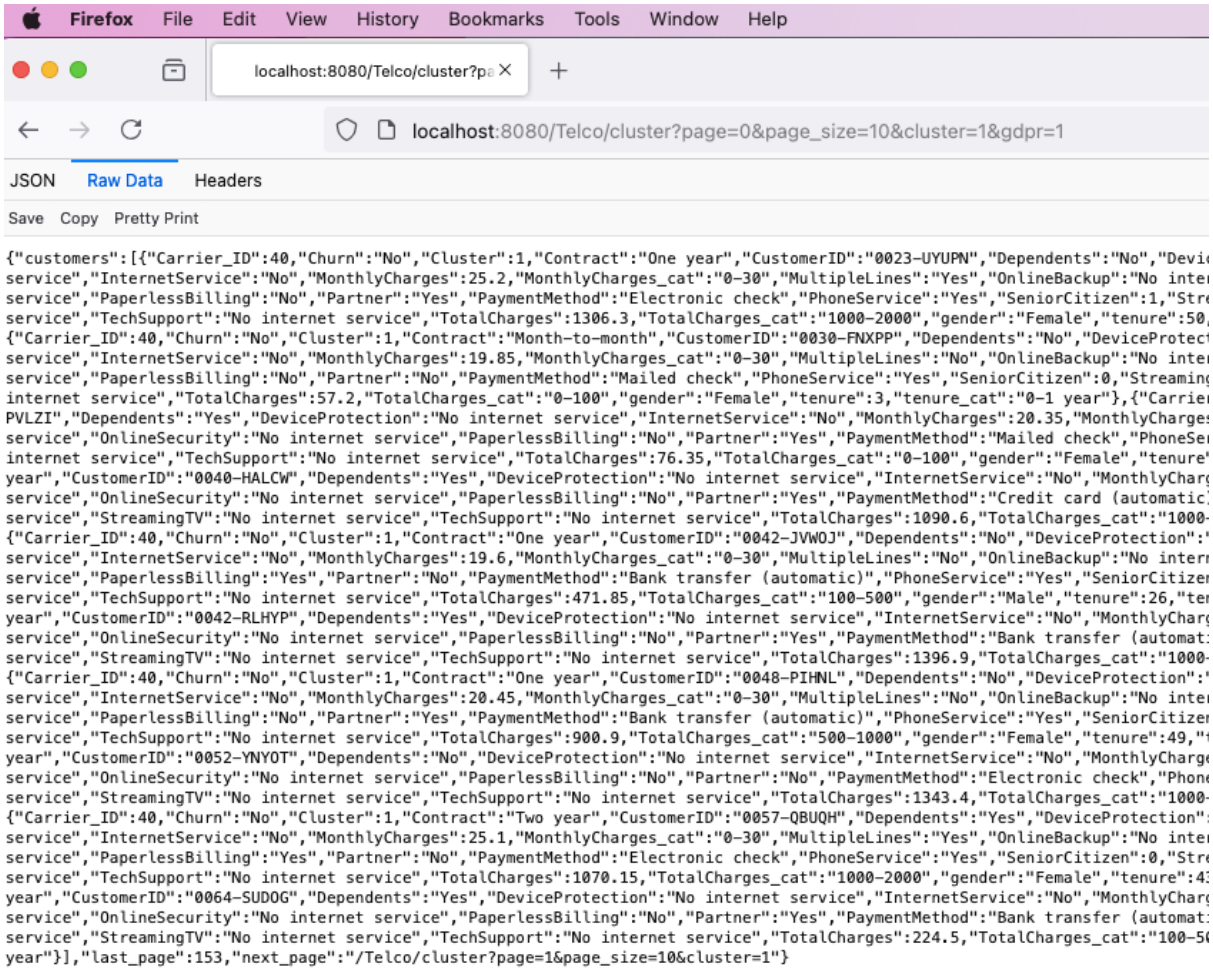

```
{
  "Carrier": "Spectrum",
  "Carrier_ID": 3,
  "Internet_Categories": "Cable",
  "Plans_Available": 15
}
```


http://localhost:8080/Telco/cluster?page=0&page_size=10&cluster=1&gdpr=0

```
{
  "customers": [
    {
      "Carrier_ID": 40,
      "Churn": "No",
      "Cluster": 1,
      "Contract": "One year",
      "DeviceProtection": "No internet service",
      "InternetService": "No",
      "MonthlyCharges": 25.2,
      "MonthlyCharges_cat": "0-30",
      "MultipleLines": "Yes",
      "OnlineBackup": "No internet service",
      "OnlineSecurity": "No internet service",
      "PaperlessBilling": "No",
      "PaymentMethod": "Electronic check",
      "TotalCharges": 1306.3,
      "TotalCharges_cat": "1000-2000",
      "tenure": 50,
      "tenure_cat": "over 4 years"
    },
    {
      "Carrier_ID": 40,
      "Churn": "No",
      "Cluster": 1,
      "Contract": "Month-to-month",
      "DeviceProtection": "No internet service",
      "InternetService": "No",
      "MonthlyCharges": 19.85,
      "MonthlyCharges_cat": "0-30",
      "MultipleLines": "No",
      "OnlineBackup": "No internet service",
      "OnlineSecurity": "No internet service",
      "PaperlessBilling": "No",
      "PaymentMethod": "Mailed check",
      "TotalCharges": 57.2,
      "TotalCharges_cat": "0-100",
      "tenure": 3,
      "tenure_cat": "0-1 year"
    },
```


To deploy our app in production, we typed the below on the terminal:

- flask --app notebooks/flask_app run --host 0.0.0.0 --port 8080

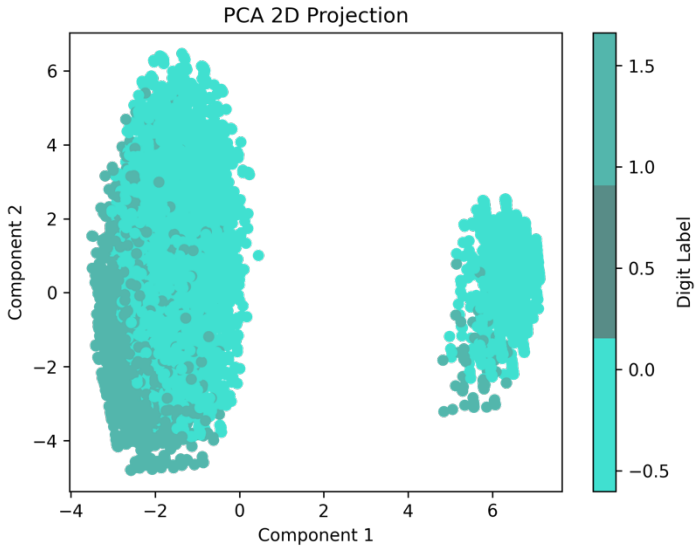


Machine Learning and Initial Findings

Clustering Algorithms

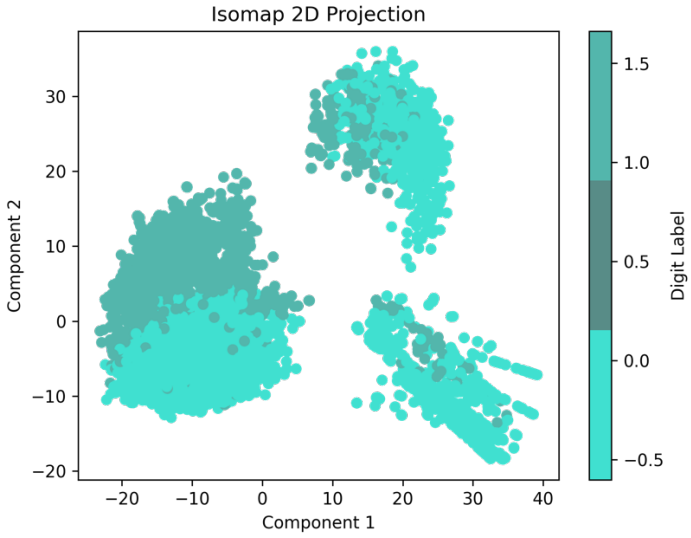
1. Transforming categorical features in numerical (dummification)
2. Standardisation of numerical data
3. Dimensionality reduction method
   - Testing relevancy of PCA method

PCA may not be optimal as the first two components explain only 0.38 of the variance, which is less than 80%.



- Testing relevancy of Isomap method

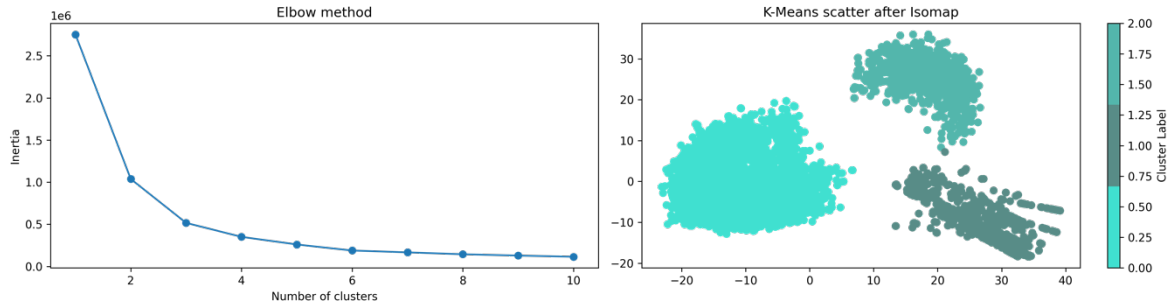
Isomap method seems to be more relevant than PCA for the 2D projection as we discovered a 3rd cluster.



4. Clustering algorithm

- Testing relevancy of K-means algorithm

Testing visually with elbow method and scatter plot:



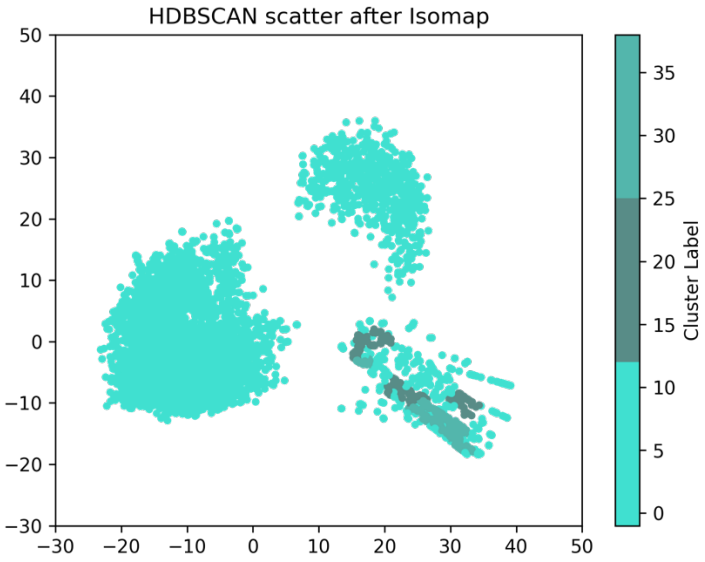
Testing with the average silhouette metric:

- For n_clusters = 2 The average silhouette_score is : 0.6287866644619318
- **For n_clusters = 3 The average silhouette_score is : 0.6834975292525928**
- For n_clusters = 4 The average silhouette_score is : 0.5229949160215177
- For n_clusters = 5 The average silhouette_score is : 0.509927850952412
- For n_clusters = 6 The average silhouette_score is : 0.48045968416837737

We proved visually and with the average silhouette metric that we could analyse our dataset using 3 clusters.

- Testing relevancy of HDBSCAN algorithm

Testing visually with a scatter plot:



HDBSCAN is not the optimal algorithm here as it did not identify the 3 clusters, which are however clearly defined on the graph. We can indeed notice that the hdbscan labels are not accurate.

Conclusions

Cluster Analysis

Interpretations

- Cluster 0 : customers who usually have no dependent, who are subscribing to phone, DSL, and fiber optic services. Most of them have a month-to-month contract, paperless billing, and pay with electronic check. Our highest paying customers are part of this population, median monthly charge is around USD 80. They have a 33% chance of churn.
- Cluster 1 : customers subscribing to phone services but no internet service, usually having only one line, paying less than USD30 monthly mostly by mailed check, and mostly under 65 years old. These customers are unlikely to churn (less than 10%).
- Cluster 2 : customers subscribing to DSL internet services but no phone service, paying between 0 and USD80 per month, with the majority of customers paying between USD30 and USD50. They have about 25% chance of churn.

General Analysis

Our analysis of both synthetic telco data and real-world data reveals key insights into customer churn drivers. Customers using bundled services (internet and phone) have a higher churn rate compared to those using standalone services, likely due to our telco company's higher bundle prices, unlike competitors that offer discounts. This pricing disparity encourages customers to switch to cheaper providers.

We also found that customers with multiple phone lines are slightly more prone to churn, and real-world data from T-Mobile suggests that competitive pricing for additional lines can help reduce churn. Other factors that increase churn risk include having no dependents or partner, being a senior citizen, being a new customer (under 1 year), and subscribing to fiber optic internet without additional services like tech support. Customers on month-to-month contracts, especially those paying via electronic check, are also more likely to churn. Addressing these risks with targeted offers and retention strategies could significantly reduce churn rates.

General Insights & Business actions

Churn prevention

- **Cluster 0**: Retain customers with specific offers on additional internet service when using Fiber Optic, especially Online Backup and Device Protection. Intensify these offers in the first year of subscription.
- **Cluster 2**: Run special offer on Tech Support.

Loyalty program

- **Cluster 1**: Reward customers to maintain engagement.

Upselling

- **All clusters**: Run incentive to subscribe to multiple telephone lines.
- **Cluster 1**: Offer service upgrades, such as internet services.

Note on GDPR

Our data scraped on whistleout.com does not contain personal information, however, data available on the kaggle.com flat file does. This data is fictive, so technically we are not running into any GDPR compliance issue. Nonetheless, in the case where this customer data would have been from a real telecommunications company, a few precautions can be taken:

- Inform our customers when they provide this information (here, probably in their contract) about data processing, use, and purpose as well as data retention.
- Ensure no personal data is being transferred outside EU or to third-party organisations without customers' consent.
- Make sure the API allowing data access to external parties is protected with authentication, and/or set up a specific argument to avoid sharing personal data by default (just like we did using flask).

Github Link

https://github.com/luciest06/Ironhack-Final_Project

https://public.tableau.com/app/profile/lucie.stenger/viz/Tableau_Telco_customers/SummaryDashboard?publish=yes

Sources

https://www.akkio.com/post/telecom-customer-churn#:~:text=Customer%20churn%20in%20the%20Telco%20industry,-The%20telecom%20industry&text=While%20the%20broader%20utilities%20market,%2C%20product%20failure%2C%20and%20price.

https://techsee.com/resources/reports/state-of-customer-churn-telecom-survey-report/

https://www.whistleout.com/

https://www.kaggle.com/datasets/blastchar/telco-customer-churn