# Video Streaming Churn Project | Preliminary Data Summary (Milestone 2)

Executive Summary Report

Prepared by: PhD Aleksandar Osmanli

## OVERVIEW

I'm currently developing a data analytics project aimed at increasing overall growth by preventing monthly user subscription churn on the video streaming service and continue their subscriptions for another month. For the purposes of this project, churn quantifies the number of users who have canceled the subscription on the video streaming service.

**This report offers a preliminary data summary, information on the project status and key insights of Milestone 2, which impact the future development of the overall project.**

## PROJECT STATUS

**Milestone 2 - Compile Summary Information**

🎯 **Target Goal:** Inspect user data to learn important relationships between variables.

🎯 **Methods:**

- Built a dataframe
  - Each row represents a single observation, and each column represents a single variable
- Collected preliminary statistics
- Analyzed user behavior

🎯 **Impact:** important relationships were determined between variables that will guide further analysis of user data.

## NEXT STEPS

➔ **It is recommended to analyze the impact of some variables on the user churn**. It's obvious that variables 'AccountAge', 'AverageViewingDuration', 'ViewingHoursPerWeek', 'ContentDownloadsPerMonth' have negative impact on user churn, while 'MonthlyCharges'. 'TotalCharges' and 'SupportTicketsPerMonth' have positive impact on user churn

➔ **The immediate next step is to conduct thorough EDA and feature engineering** to reveal the correlation between features and user churn.

## KEY INSIGHTS

- I was provided with two datasets:

1. Train dataset contains 21 variables and 70% of the overall sample, and will reveal whether or not the subscription was continued into the next month (the "ground truth").
2. Test dataset contains 20 variables with the exact same information about the remaining sample of 30%, but does not disclose the "ground truth" for each subscription. I should predict the outcome.

- The types of variables include 11 objects (10 categorical variables), 5 floats, and 5 integers; two categorical variables contains spaces between values, so I concatenated them to avoid future problems with analysis.
- There were no missing values in both train and test sets.
- There were only 741 outliers out of 104.480 values in the 'TotalCharges' (less than 1%), so I decided to leave them.