SMU Data Science

Machine Learning 1 DS7331

Lab 1

Jeremy Otsap [jotsap@mail.smu.edu](mailto:jotsap@mail.smu.edu)

Shawn Jung [shawnj@mail.smu.edu](mailto:shawnj@mail.smu.edu)

Lance Dacy [ldacy@mail.smu.edu](mailto:ldacy@mail.smu.edu)

Amber Burnett [aburnett@mail.smu.edu](mailto:aburnett@mail.smu.edu)

Contents

[Business Understanding 3](#_Toc30438100)

[Data Understanding 3](#_Toc30438101)

[Data Dictionary 3](#_Toc30438102)

[Data Quality 4](#_Toc30438103)

[Appendix 1: Code 5](#_Toc30438104)

# Business Understanding

We are looking at customer data from a north American Telco provider. The purpose being to retain existing customers. In telecommunications, the estimated cost of new customer acquisition is approximately 5x higher than retaining an existing customer. Furthermore, only a third of customers switch carriers due to lower prices; more and more factors such as dissatisfaction with quality of service, advancing technology and media features, competitors having better cellular coverage, and poorly implemented loyalty programs are all contributing to customer attrition.

<https://www.entrepreneur.com/article/225415>

<https://www.demandjump.com/blog/customer-acquisition-cost-by-industry>

<https://www.performancemagazine.org/june-smartkpi-telecom-sub-acquisition/>

<https://www.invespcro.com/blog/customer-acquisition-retention/>

<https://www.annexcloud.com/blog/21-surprising-customer-retention-statistics-2018/>

# Data Understanding

This is the “Telco Churn Data Set” from Kaggle, url below:

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

It is a random sampling of approximately 7000 anonymized customers from an unnamed Telco provider. It contains 20 features that describe such characteristics as their demographics, their account information, and their subscribed services.

Data Dictionary  
We are provided a CSV set of 21 features and 7043 anonymized customers.

* **customerID**: Unique alpha-numeric string to anonymously represent an individual customer
* **gender**: Categorical String value to represent customer's gender (Male or Female)
* **SeniorCitizen**: Boolean int value to show whether the customer is a senior citizen or not (1, 0)
* **Partner**: Boolean string value showing whether the customer has a partner or not (Yes, No)
* **Dependents**: Boolean string value showing whether the customer has dependents or not (Yes, No)
* **tenure**: Numeric value showing number of months the customer has stayed with the company
* **PhoneService**: Boolean string value showing whether the customer has a phone service or not (Yes, No)
* **MultipleLines**: Categorical string value that shows if the customer has multiple lines or not (Yes, No, No phone service)
* **InternetService**: Categorical string value that shows the customer’s internet service provider (DSL, Fiber optic, No)
* **OnlineSecurity**: Categorical string value showing whether the customer has online security or not (Yes, No, No internet service)
* **OnlineBackup**: Categorical string showing whether the customer has online backup or not (Yes, No, No internet service)
* **DeviceProtection**: Categorical string showing whether the customer has device protection or not (Yes, No, No internet service)
* **TechSupport**: Categorical string showing whether the customer has tech support or not (Yes, No, No internet service)
* **StreamingTV**: Categorical string showing whether the customer has streaming TV or not (Yes, No, No internet service)
* **StreamingMovies**: Categorical string showing whether the customer has streaming movies or not (Yes, No, No internet service)
* **Contract**: Categorical string that represents the contract term (Month-to-month, One year, Two year)
* **PaperlessBilling**: Boolean string showing whether the customer has paperless billing or not (Yes, No)
* **PaymentMethod**: Categorical string that shows the customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* **MonthlyCharges**: Numeric value showing the amount charged to the customer each month
* **TotalCharges**: Numeric value showing the total amount charged to the customer
* **Churn**: Boolean string showing whether or not the customer 'churned' or terminated services (Yes or No)

Data Quality

For the most part this data is production ready. Two modifications were necessary.

First the **MonthlyCharges** has 11 missing values. [**Code Appendix 1A]** Because it is such a small percentage of the total 7000+ values it should not prove to be problematic. For the moment we are simply excluding these 11 observations from the initial data exploration and visualization exercises.

Second the **SeniorCitizen** variable is given as numeric, when for all practical purposes it is a categorical factor. I.e. a customer cannot be 0.77 of a senior citizen; they either are under 65 or they are not. For this we simply converted from an integer to a categorical factor. [**Code Appendix 1B**]

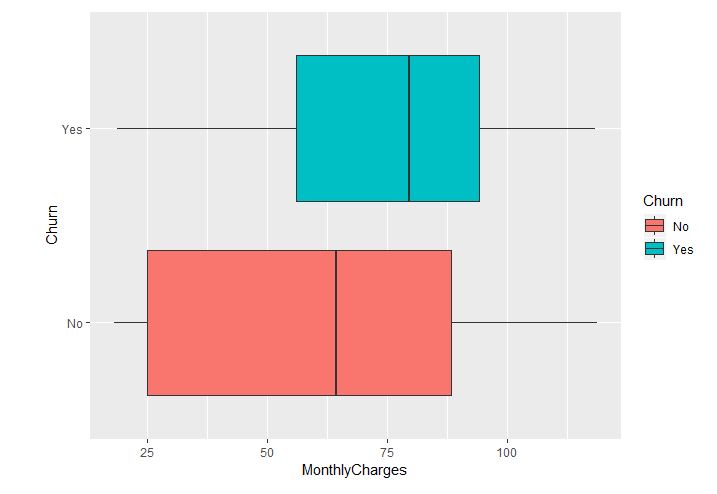
# Exploratory Analysis & Visualization

**AMBER & LANCE**: I’m embedding the images here w/ very minimal comments as I’m still uncertain where to focus

NUMERIC

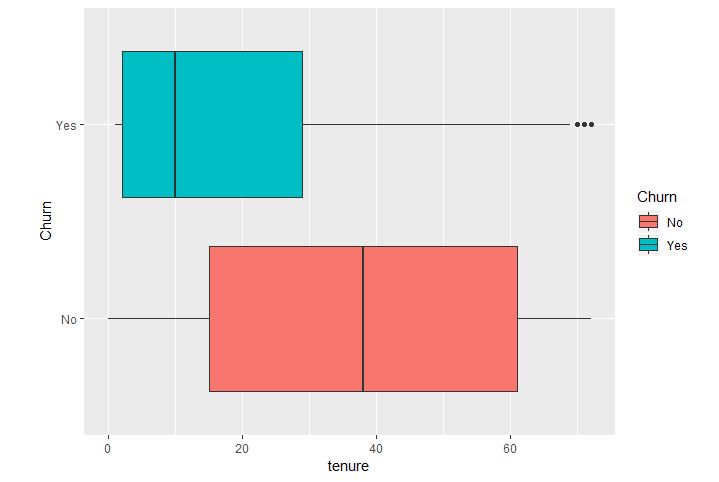
BoxPlot of Monthly Charges for each Churn Category

shows a significant difference in the distribution. **If we validated this w/ a 2-sample t-test it will likely show the 2 distributions differ significantly**. This is an observation so we can NOT make a reference to causality, BUT because its random we CAN make generalizations. Thus there is some correlation between Monthly Charges and Churn, which the Logistic Regression and / or LDA will illuminate



BoxPlot of Tenure for each Churn Category

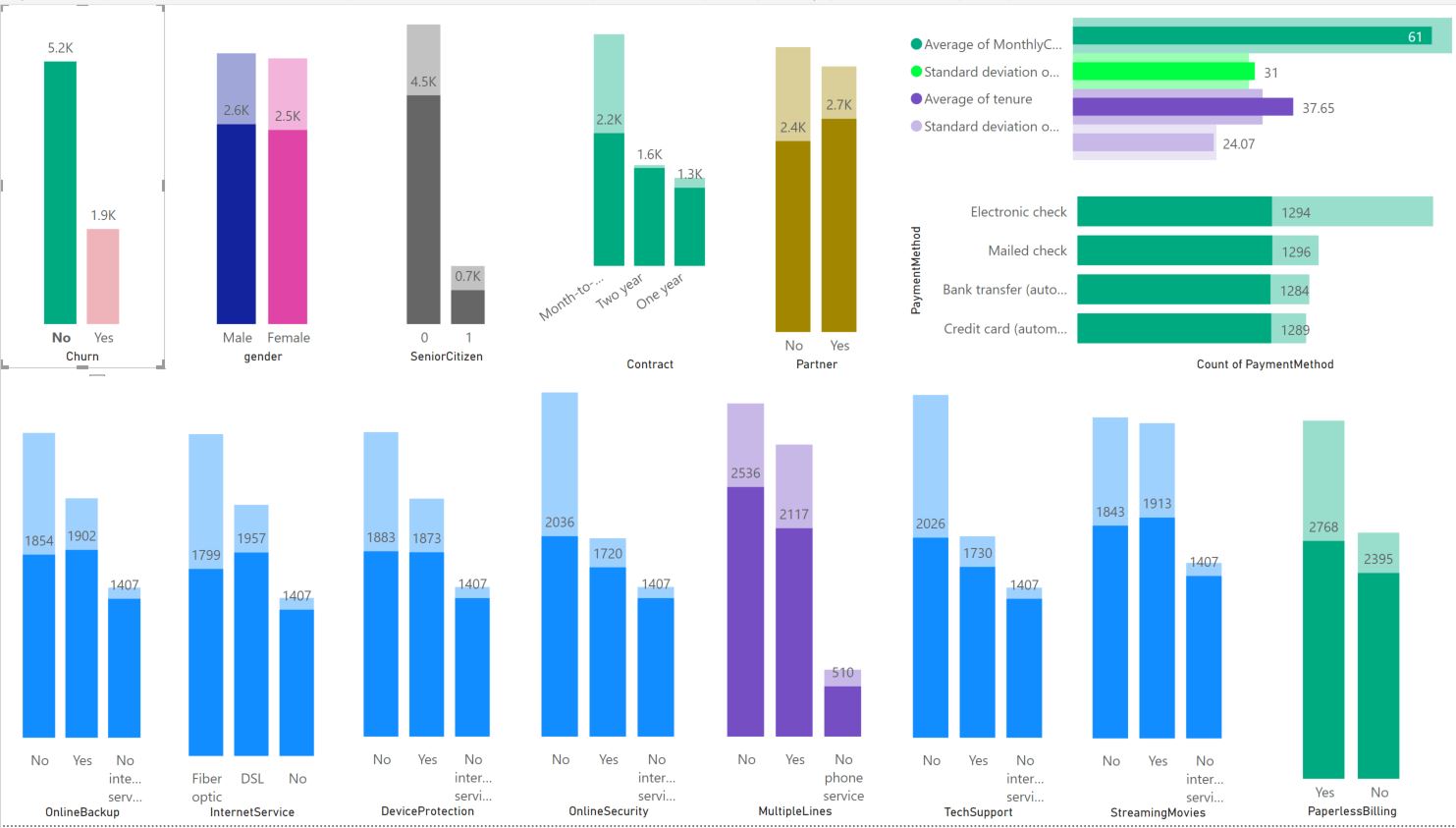
As per above, a **Welch’s 2 sample t-test** would likely show significant deviations in the two different distributions. This is fairly intuitive; customers who churn usually do so within the first 1 – 2 years of the service contract

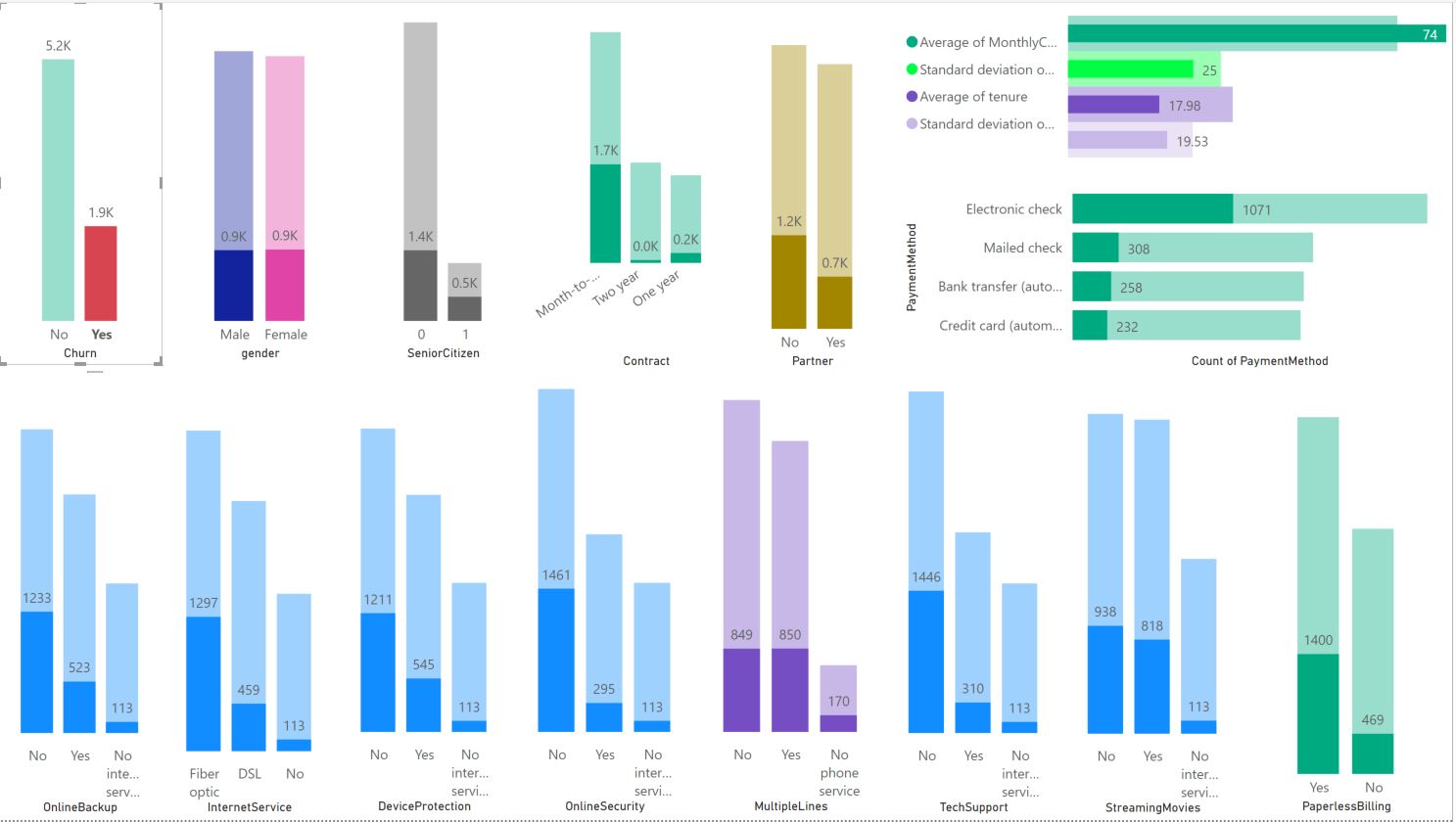


CATEGORICAL

COMPARING CHURN: The Churn “Yes” group

* Has disproportionate number of eCheck for payment method
* Shorter Average Tenure
* Higher Average Monthly Cost
* Lower ratio of those calling Tech Support
* Lower ratio of advanced data services [i.e. device protection, security, online backup]

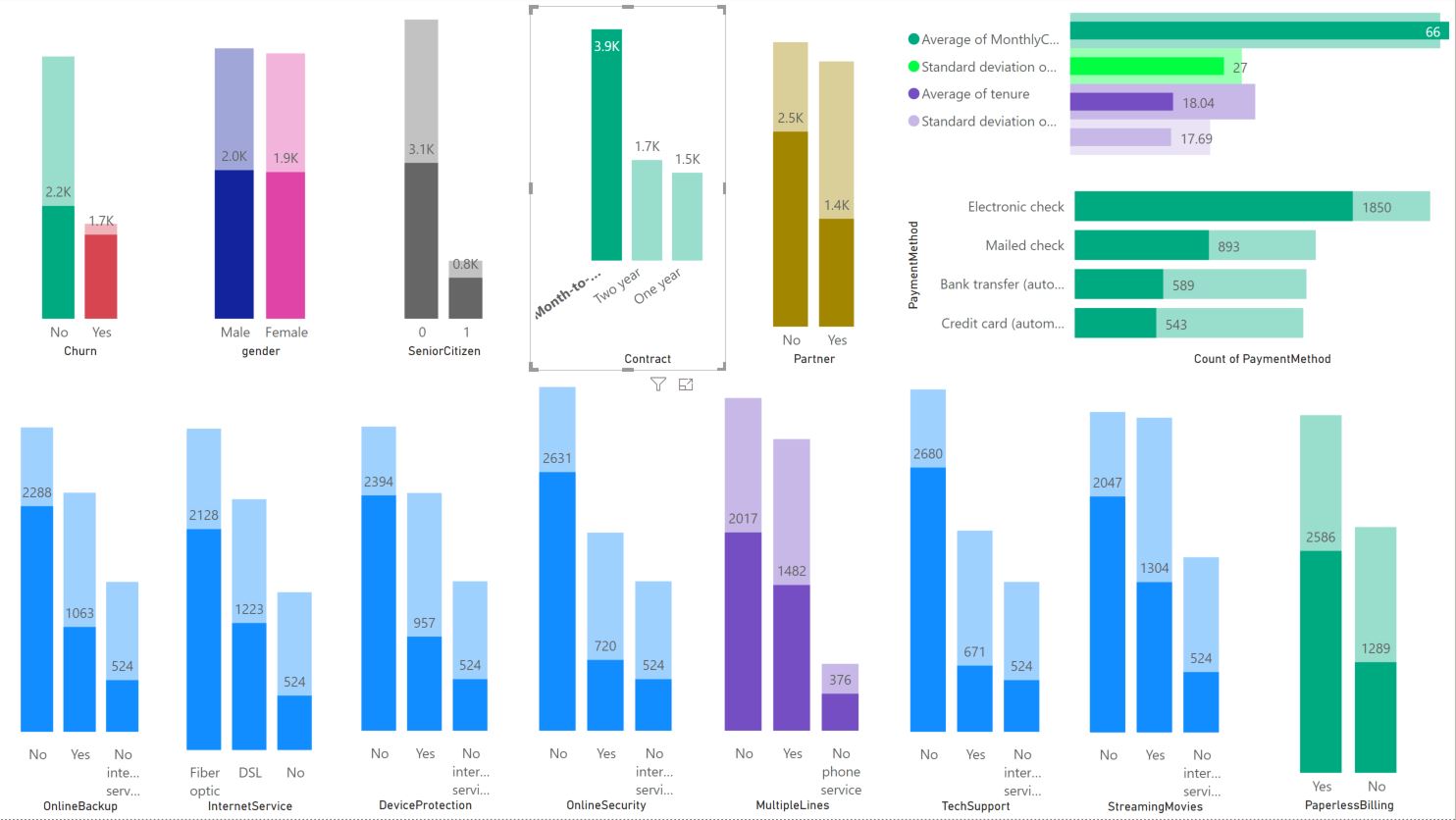




CONTRACT: MONTHLY

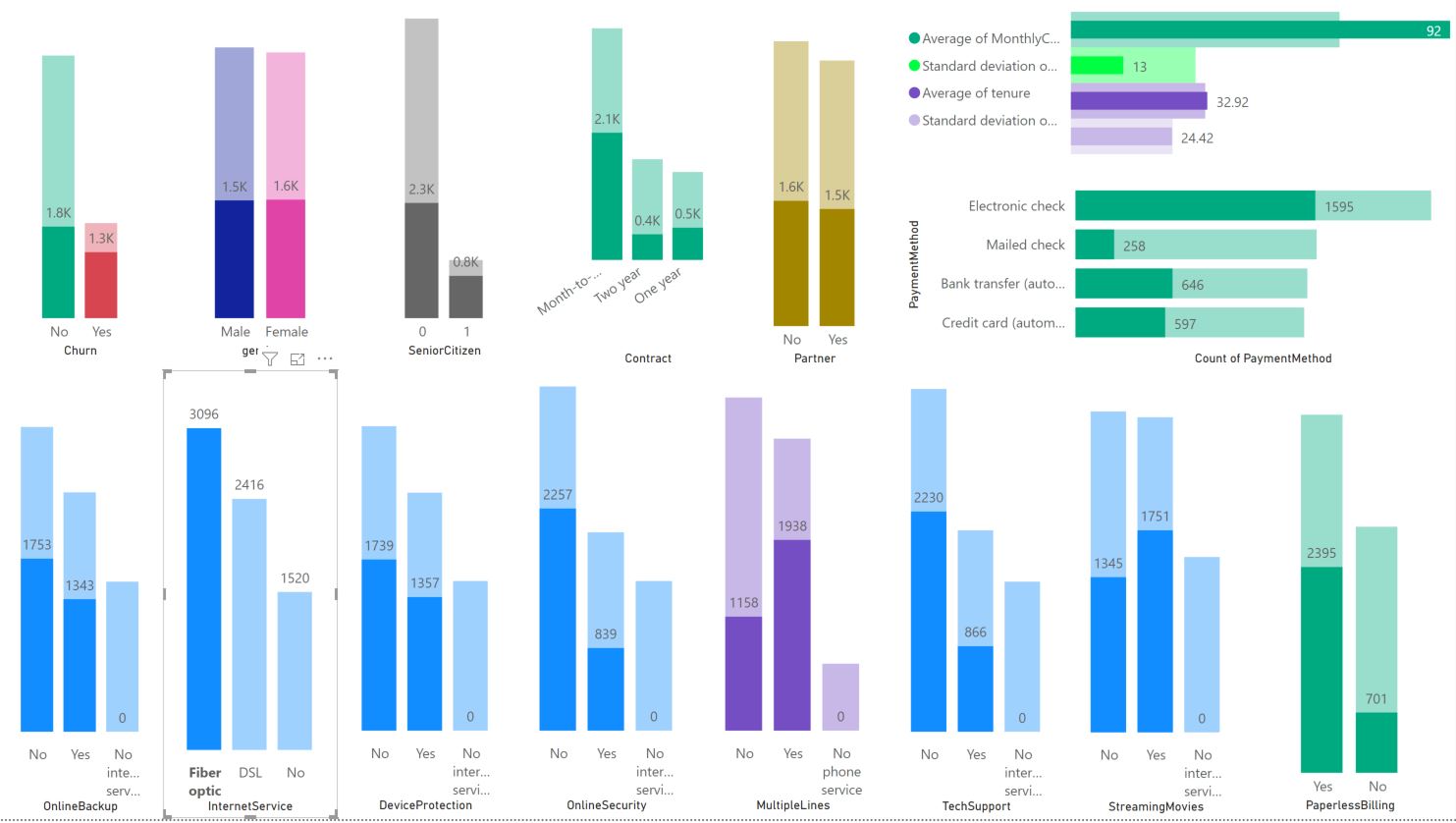
Notice the RATIO of Churn Yes vs No is disproportionate, having almost a 1 to 1. Compared to 1 year and 2 year that more closely align to the total Yes vs No behavior.

Note also that most of them pay with eCheck



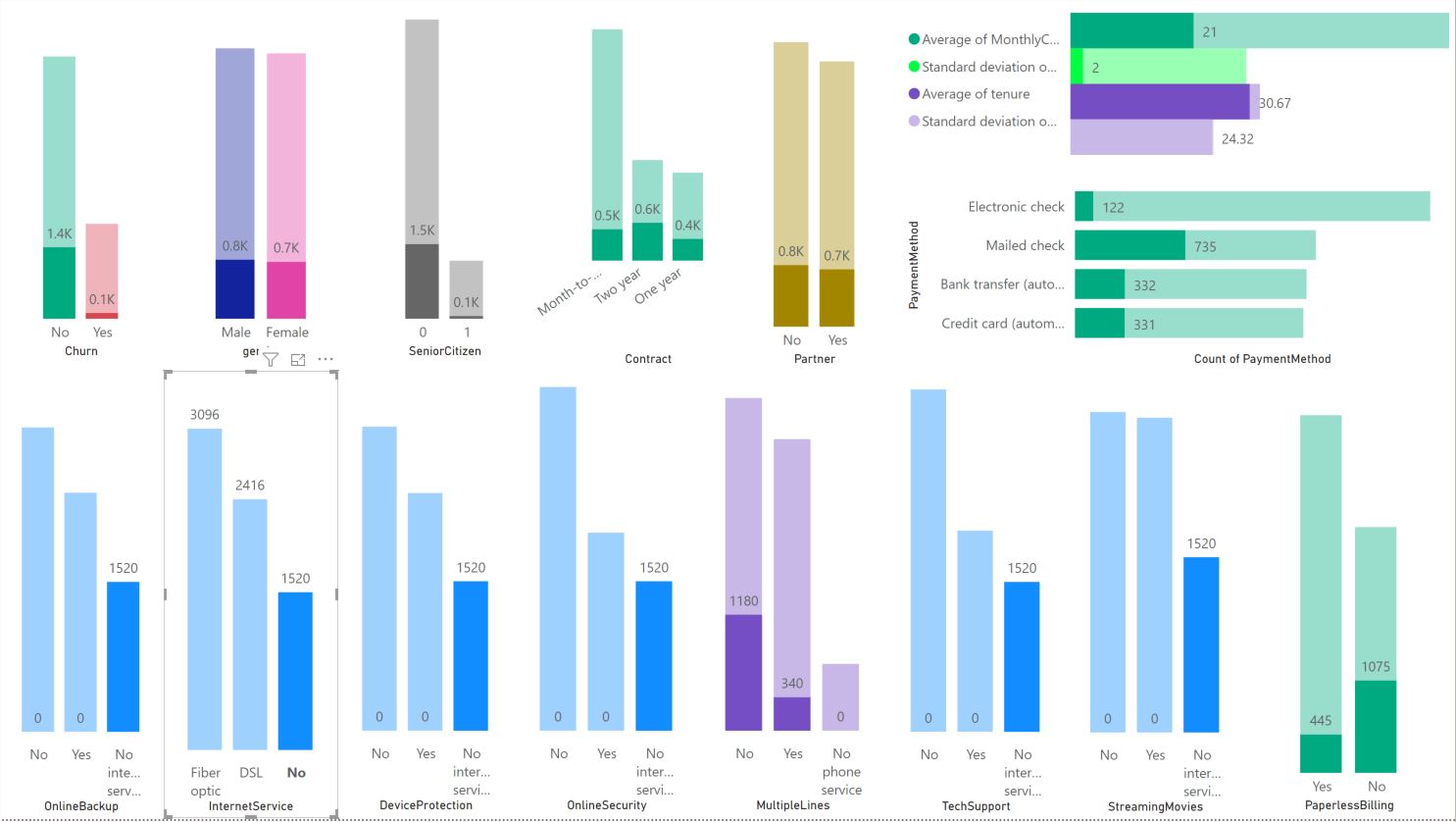
FIBER INTERNET SERVICE

* Almost a 1 to 1 Churn rate
* All of them have some kind of phone service
* Have much higher Avg Monthly Charges
* Prefer eChecks



NO INTERNET SERVICE

* Very few customers are in the Churn Yes category
* Notice also very few Seniors; this might imply that Seniors prefer “bundling”
* NONE of these users are in the “No Phone Service category



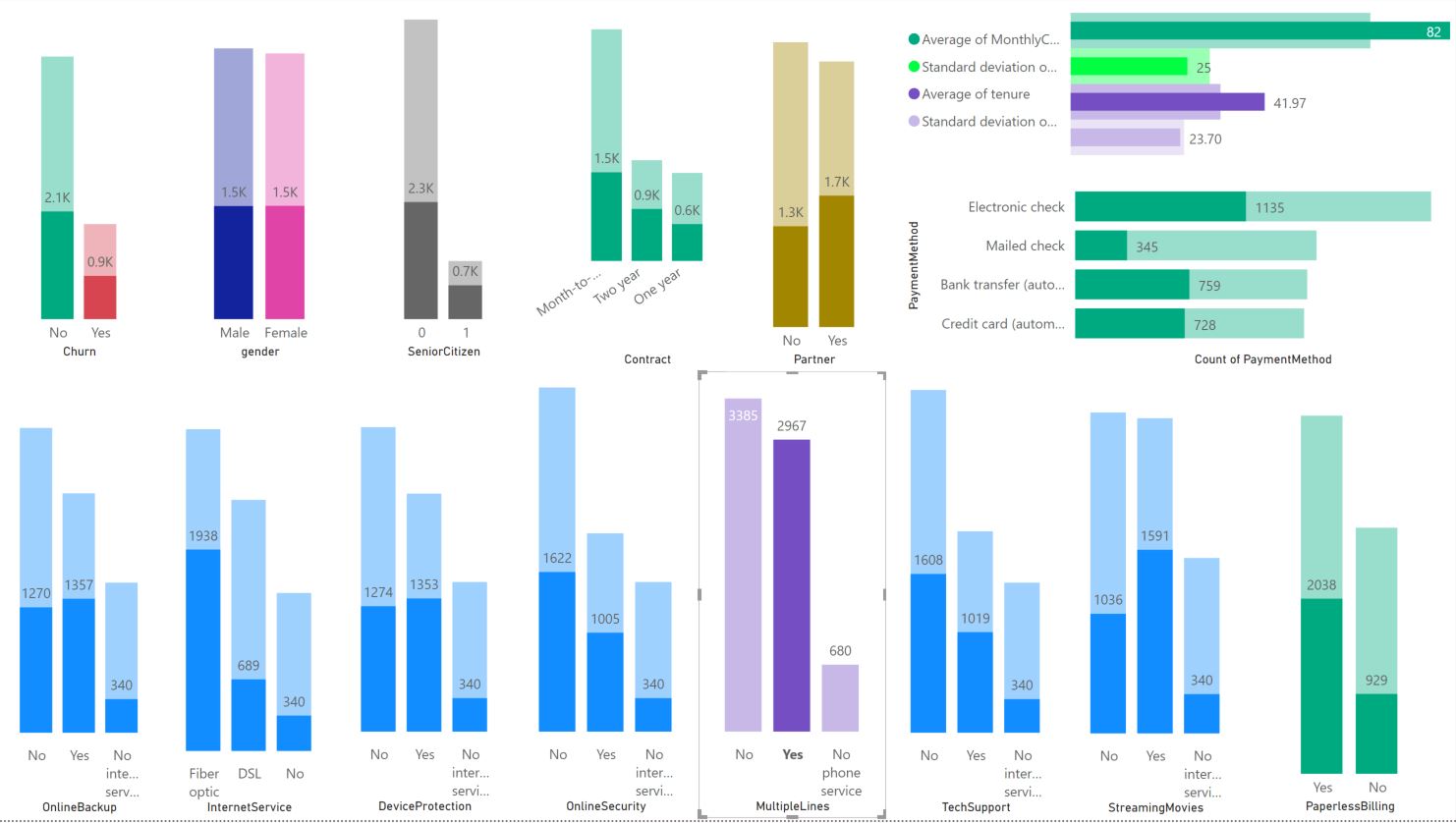
PHONE SERVICE: MULTI LINE

Of the 2967 customers who have multiple phone lines, 340 of them have no Internet. Possible opportunity to market services like Skype or WebEx; advanced VoIP services and video conferencing.

* They have higher Avg Monthly Spend and higher Avg Tenure
* BUT NOTICE the ratio for Churn [~ 2 to 1] is slightly higher vs the ENTIRE data set’s Churn ratio [~ 3 to 1]

CONCLUSION: maybe they spend their money more “productively” getting modern features like VoIP and Video Conferencing for the same money, and now multiple lines aren’t needed. Thus

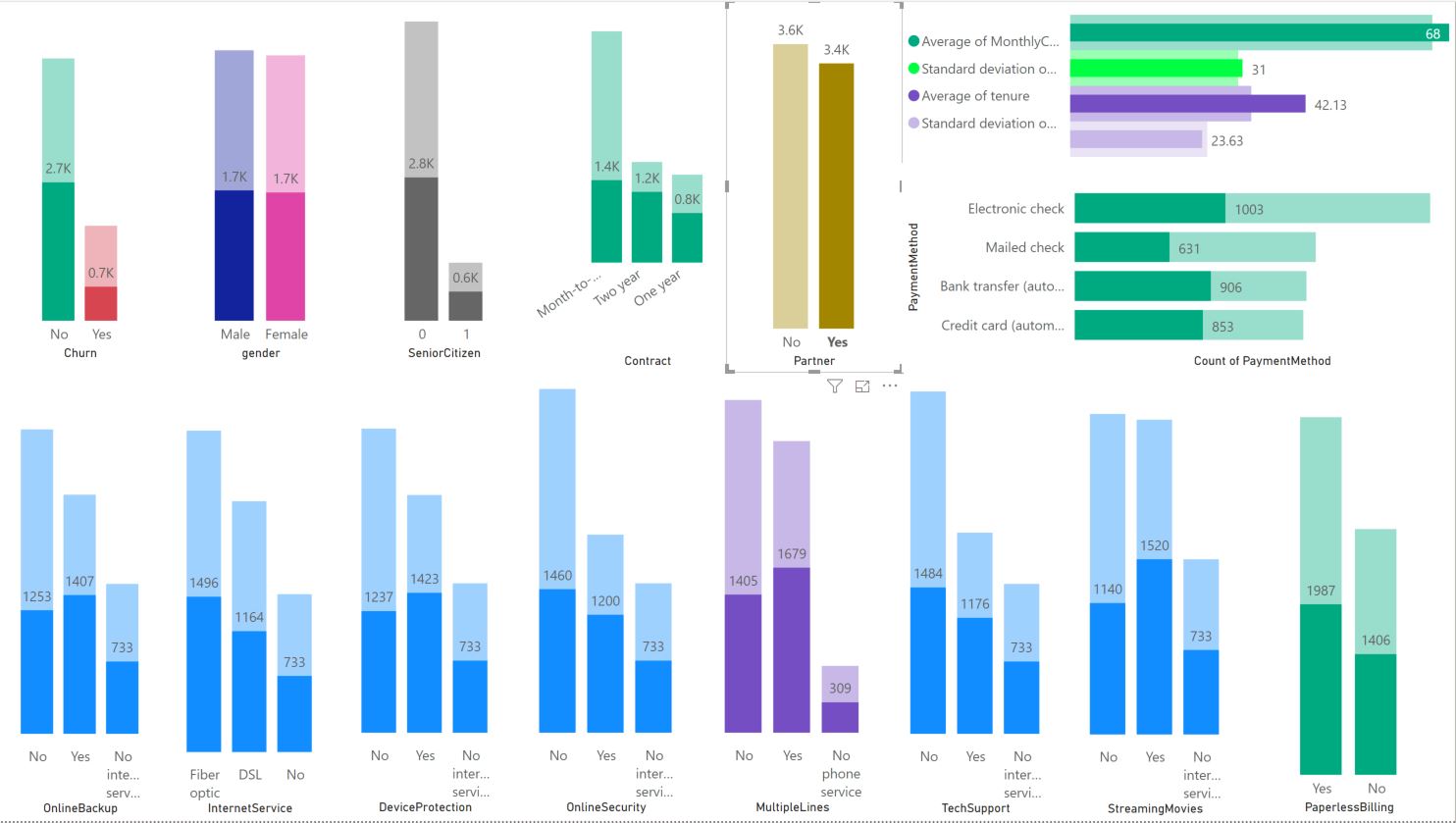
* Better loyalty
* Better service
* Migrate off LEGACY tech like analog phone lines



PARTNER: YES

Have overall higher Tenure compared to those in the Partner NO category

* Avg Tenure of Partner No = 23.37
* Avg Tenure of Partner Yes = 42.13 [almost double]
* However the Avg Monthly Spend is almost the same

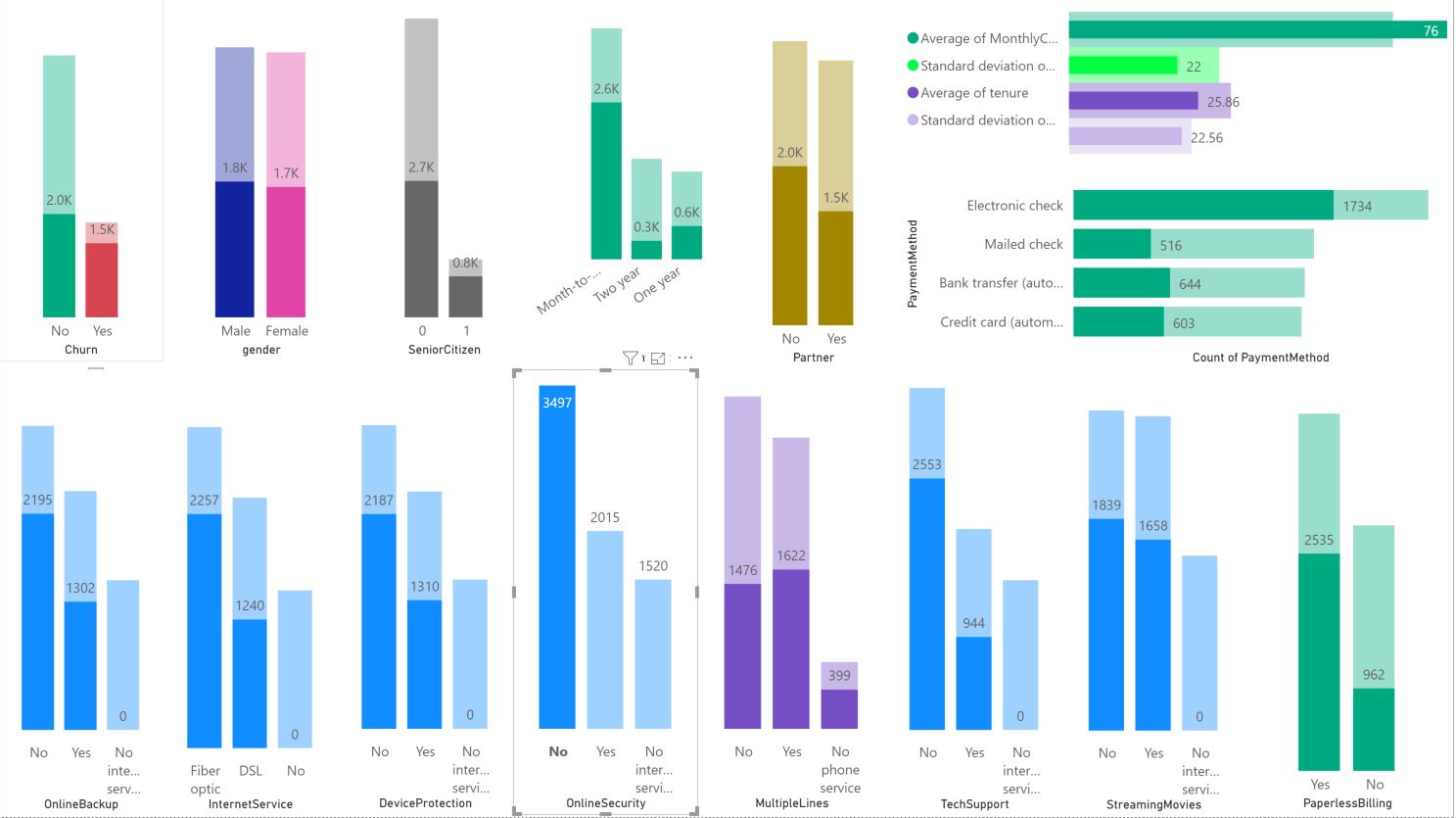


ONLINE SECURITY: NO

* Almost a 1 to 1 ratio of Churn
* Lower Avg Tenure
* Prefer eCheck
* Lower ratio of engaging Tech Support

NOTE: this behavior is the same for Online Backup, Security, Device Protection

NOTE: interestingly this behavior is also very similar for those who DO NOT leverage Tech Support



# Appendix 1: Code

Still unclear on how we reference here; lets discuss at our next meeting