8138 - SARANATHAN COLLEGE OF ENGINEERING

Department of CSE

NAAN MUDHALVAN

DAC Phase-3 Project Submission

Team-6 Customer Churn Prediction

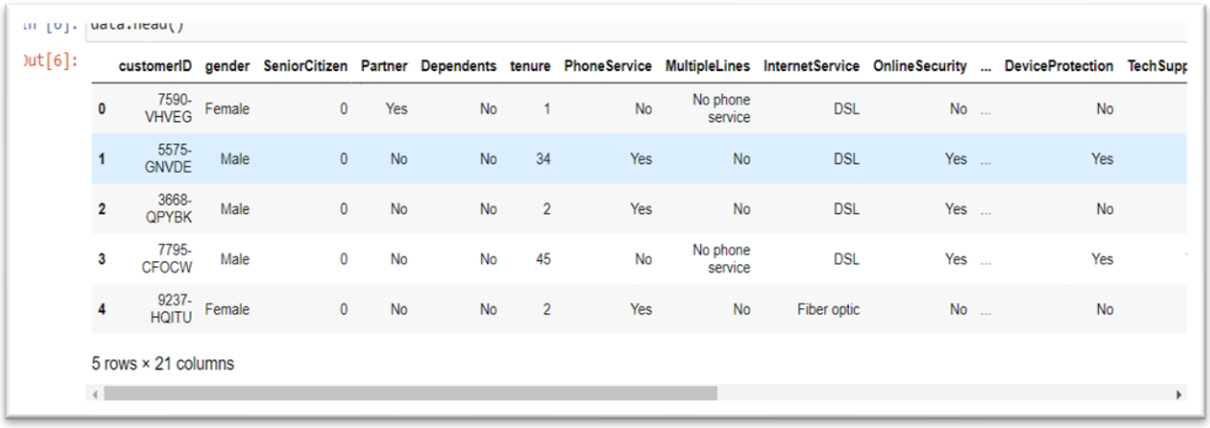
1.Loading the Dataset:

- The necessary libraries or packages have been installed and imported for working with the data.
- Initially, the Telco Customer Churn Dataset has been loaded as a csv file.

```
In [1]: import numpy as np
In [2]: import pandas as pd
In [3]: import matplotlib.pyplot as plt
In [4]: import seaborn as sns
In [5]: data=pd.read_csv("F:\\Telco-Customer-Churn.csv")
```

2. Inspecting the Data:

- After loading the required the dataset, data needs to be inspected to get an understanding.
- The following functions such as ‘head()’, ‘info()’, ‘describe()’ are used to check the first few rows of the dataset, datatypes and basic statistics.



Out[6]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupp
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

5 rows × 21 columns

3. Data Cleansing:

In this step, data is being cleaned by checking for missing values, dealing with outliers and converting data types.

```
In [9]: print(data.isnull().sum())

gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
dtype: int64
```

```
In [14]: cols=data.columns
         cols

Out[14]: Index(['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure',
                'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity',
                'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV',
                'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod',
                'MonthlyCharges', 'TotalCharges'],
               dtype='object')

In [15]: cat_cols=data.select_dtypes(exclude=['int','float']).columns
         cat_cols

Out[15]: Index(['Partner', 'Dependents', 'PhoneService', 'MultipleLines',
                'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection'
                'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
                'PaperlessBilling', 'PaymentMethod', 'TotalCharges'],
               dtype='object')
```

4. Data Preprocessing:

- In this step, data has been preprocessed by checking for categorical values and encoding them.

- This could involve feature engineering, scaling, or splitting the data into training and testing sets for machine learning.

```
In [7]: data.drop(['customerID','Churn'],axis=1,inplace=True)
```

```
In [12]: from sklearn.preprocessing import LabelEncoder

In [13]: le=LabelEncoder()
         data['gender']=le.fit_transform(data['gender'])
         data.head()
```

Out[13]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSuppo
0	0	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	N
1	1	0	No	No	34	Yes	No	DSL	Yes	No	Yes	N
2	1	0	No	No	2	Yes	No	DSL	Yes	Yes	No	N
3	1	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Y
4	0	0	No	No	2	Yes	No	Fiber optic	No	No	No	N

```
In [18]: enc_data=list(cat_cols)
         enc_data=enc_data[:-1]
         enc_data

Out[18]: ['Partner',
          'Dependents',
          'PhoneService',
          'MultipleLines',
          'InternetService',
          'OnlineSecurity',
          'OnlineBackup',
          'DeviceProtection',
          'TechSupport',
          'StreamingTV',
          'StreamingMovies',
          'Contract',
          'PaperlessBilling',
          'PaymentMethod']

In [19]: data[enc_data]=data[enc_data].apply(lambda col:le.fit_transform(col))
         data[enc_data].head()
```

Out[19]:

	Partner	Dependents	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMov
0	1	0	0	1	0	0	2	0	0	0	
1	0	0	1	0	0	2	0	2	0	0	
2	0	0	1	0	0	2	2	0	0	0	
3	0	0	0	1	0	2	0	2	2	0	
4	0	0	1	0	1	0	0	0	0	0	

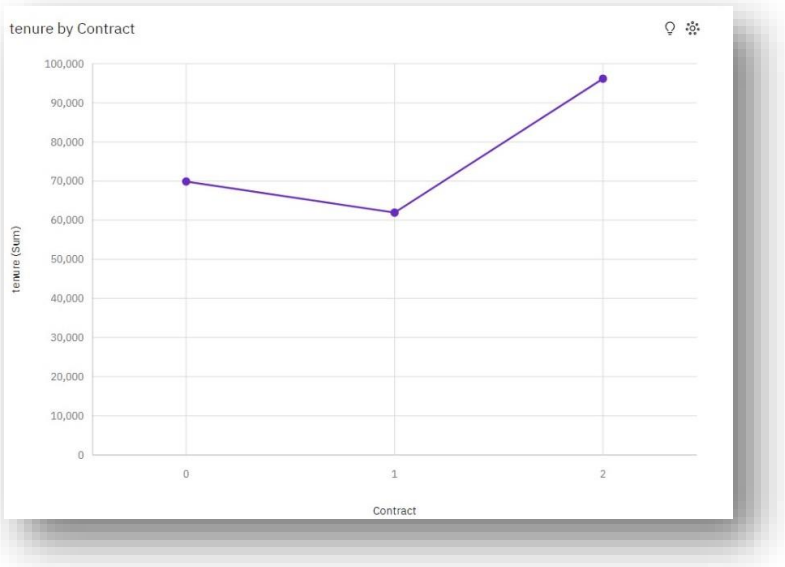
5. Data Visualization:

In this step, the preprocessed data has been visualized using IBM Cognos using various plots. Several insights are gained and data has been analyzed using various plots. The following plots have been used to visualize the data:

1.Packed Bubble

The above plot has visualized the gender feature from the dataset. In this plot 0 denotes Female and 1 represents Male.

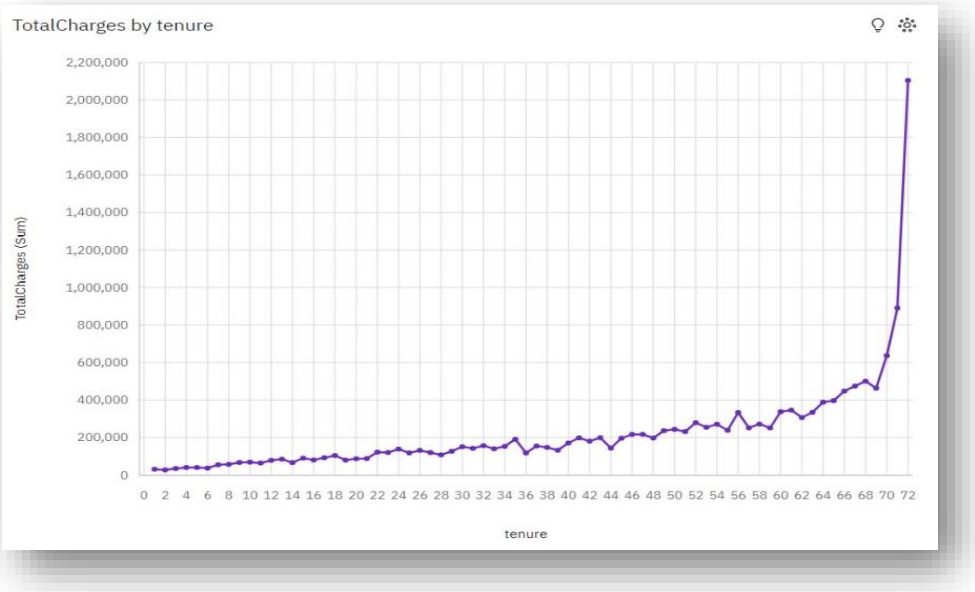
2.Line Chart



The above figure depicts plot with contract against tenure.

Tenure is unusually high when contract is two years.

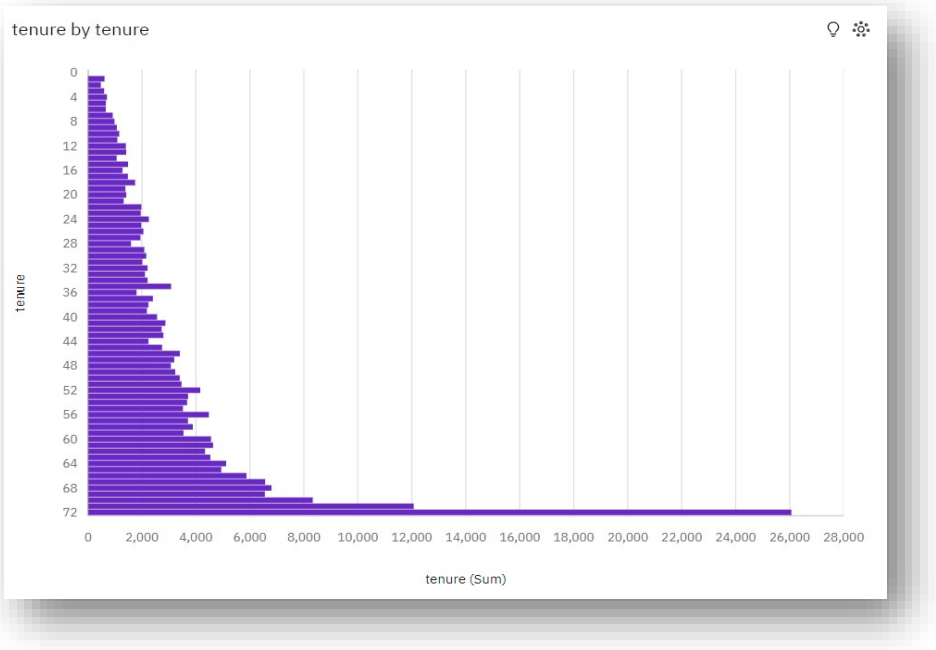
3. Line graph



The above figure shows a line plot of tenure against total charges.

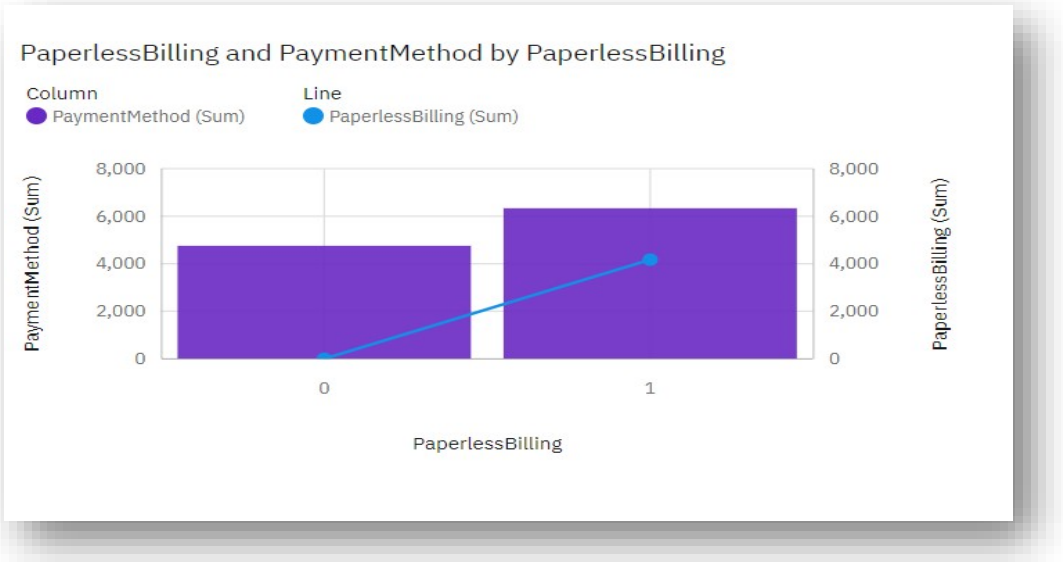
Total Charges are high when tenure is 72.

4. Bar graph



The above plot is simple bar graph representing the tenure feature from the dataset.
Across all values of tenure, the sum of the tenure is over 2500.

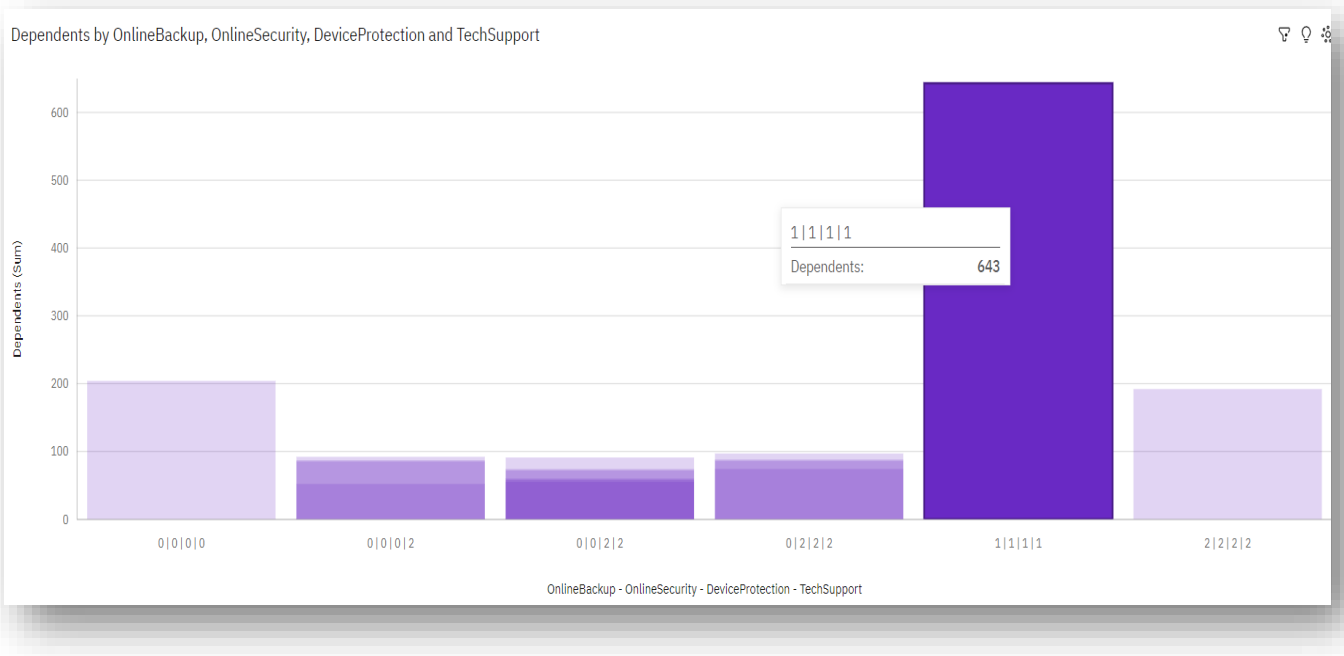
5. Line and Column chart



The above plot is a line and column chart plotted using the features Payment method and Paperless Billing.

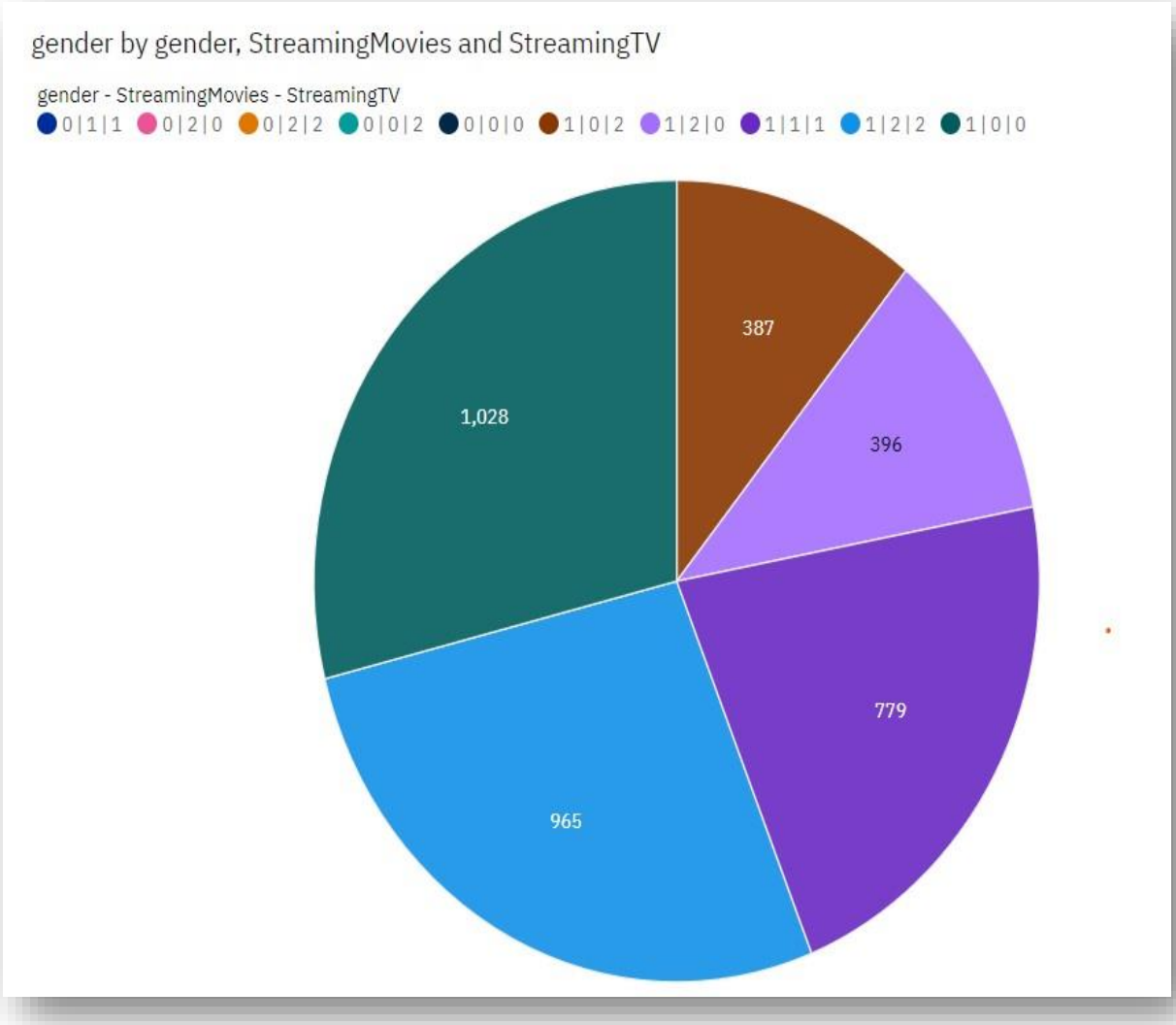
Paperless Billing when at 1 has the highest values of both payment methods and multiple lines.

6. Stacked Column



In the above plot 0 represents ‘No’, 1 represents ‘Yes’ and 2 represents ‘No Internet service’.
Tech Support when at 0 has the highest values of both Dependents and Multiple Lines.

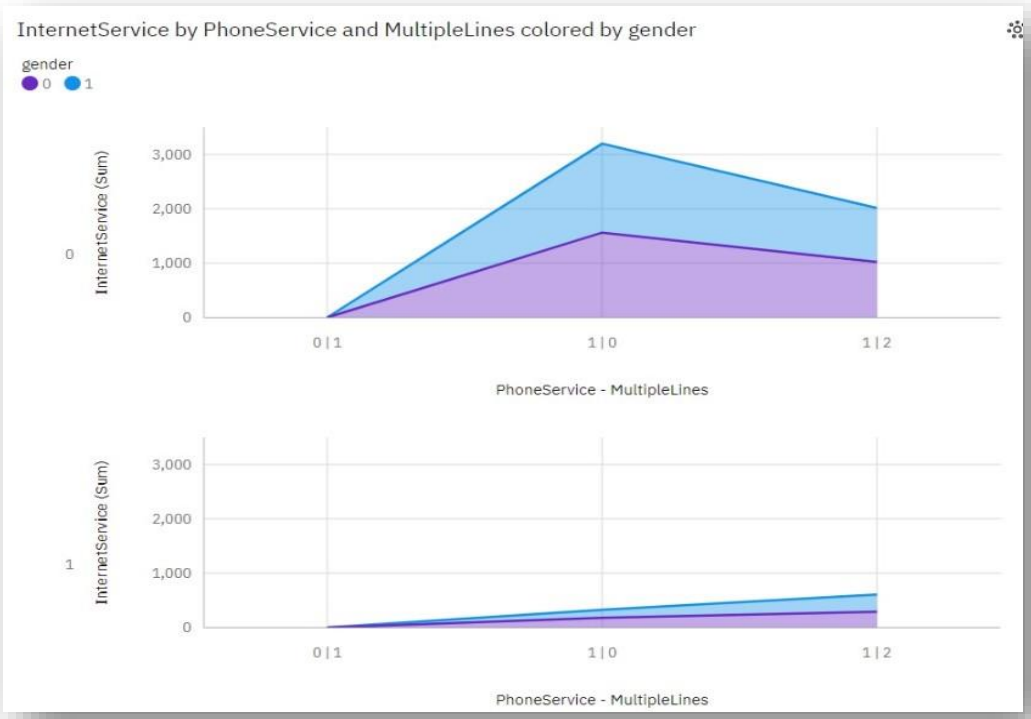
7. Pie chart



The above pie chart visualizes the features gender, Streaming Movies and Streaming TV. For gender, represents Female and 1 represents Male.

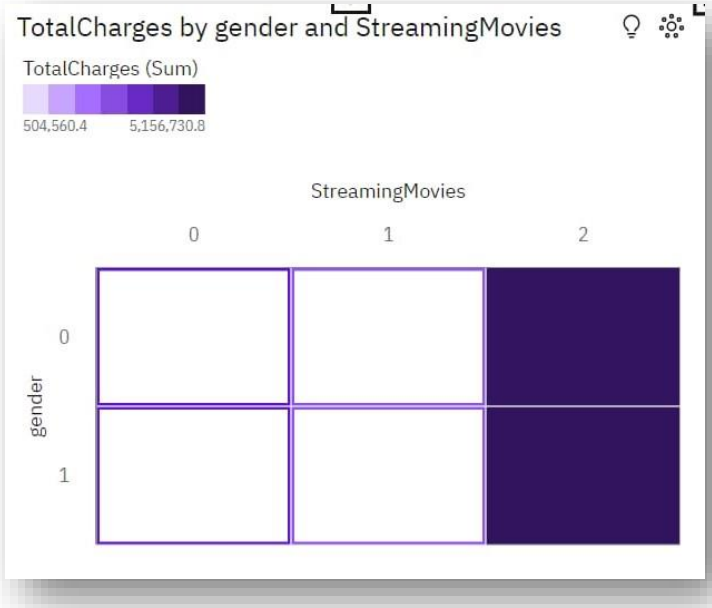
When the gender is at 0 both Streaming TV and Streaming Movies have the same values i.e., yes values.

8. Area



Gender 1 has the highest values of Internet Service and Total Charges.
Multiple lines when at 2 has the highest Total Charges but is ranked #2 in Internet service.
Multiple lines when at 0 has the highest total Internet Service but is ranked #2 in Total Charges.

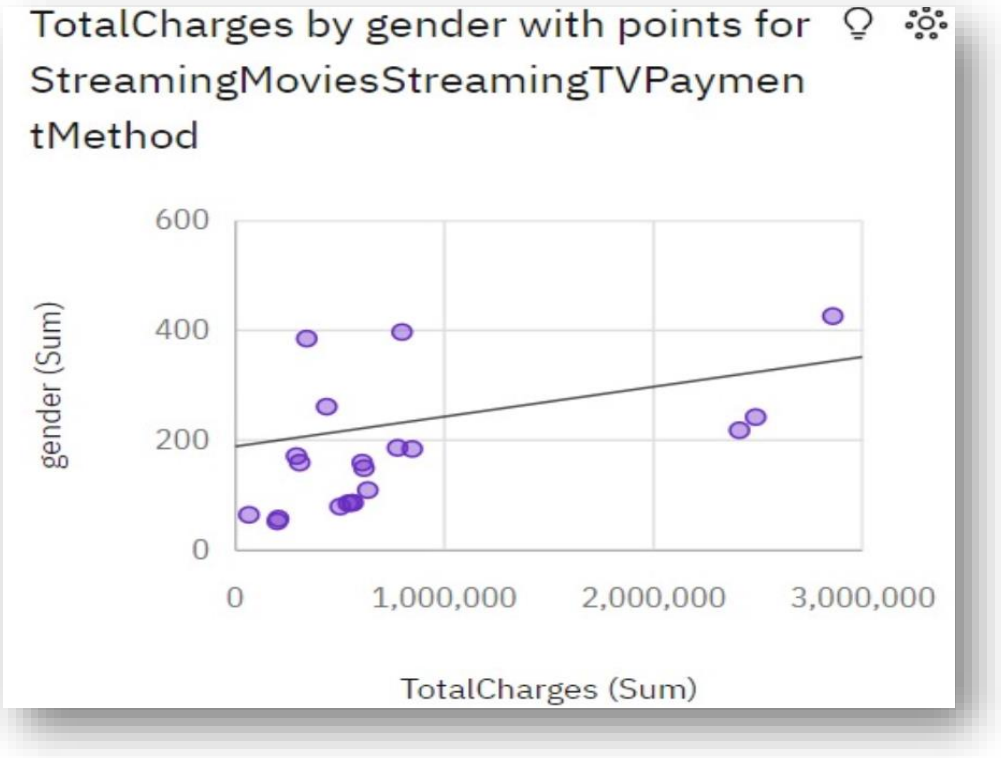
9. Heat Map



Gender 1 has the highest sum of Total Charges due to Streaming Movies is at 2.

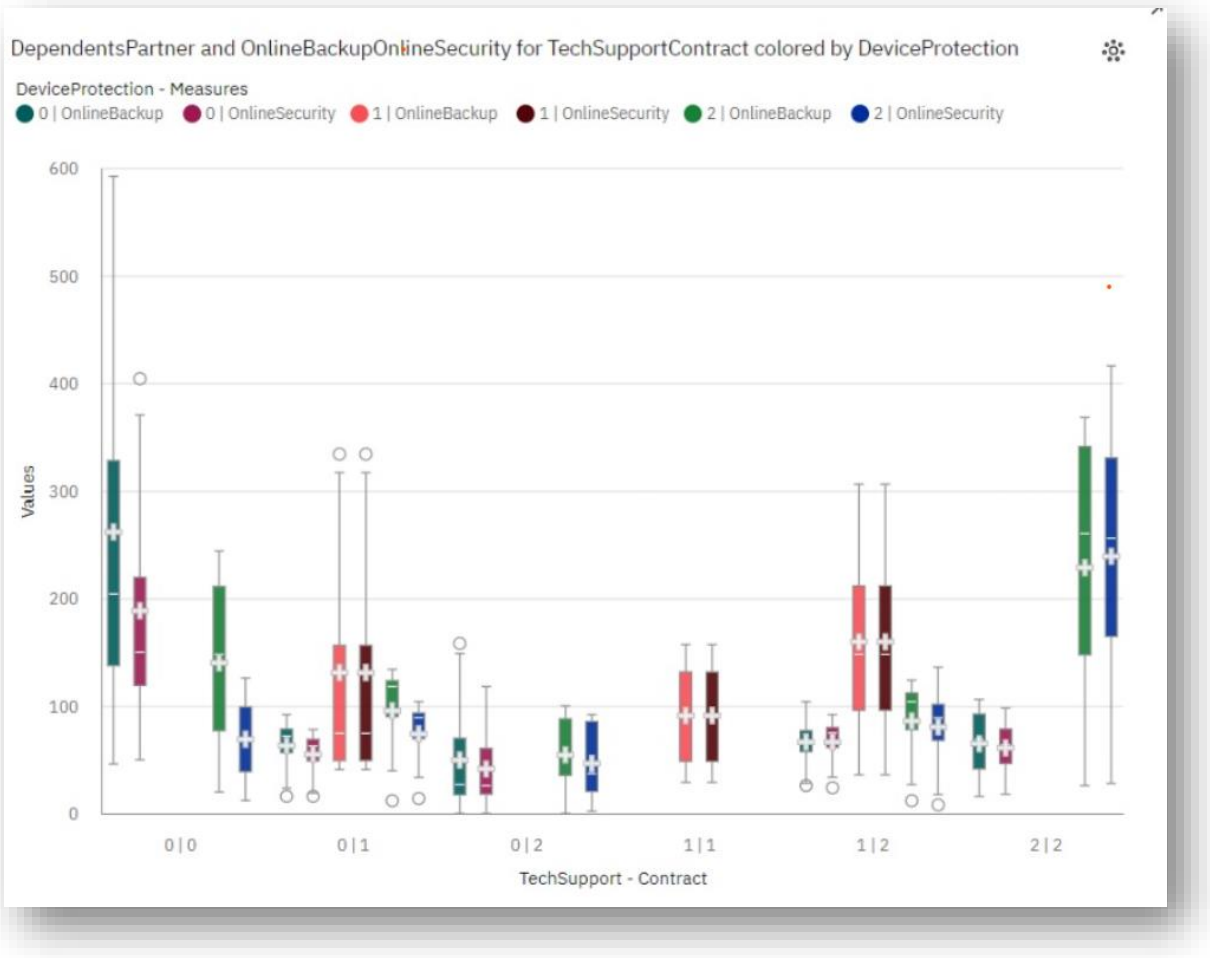
Total Charges is unusually high when Streaming Movies is at 2.

10. Scatter Plot



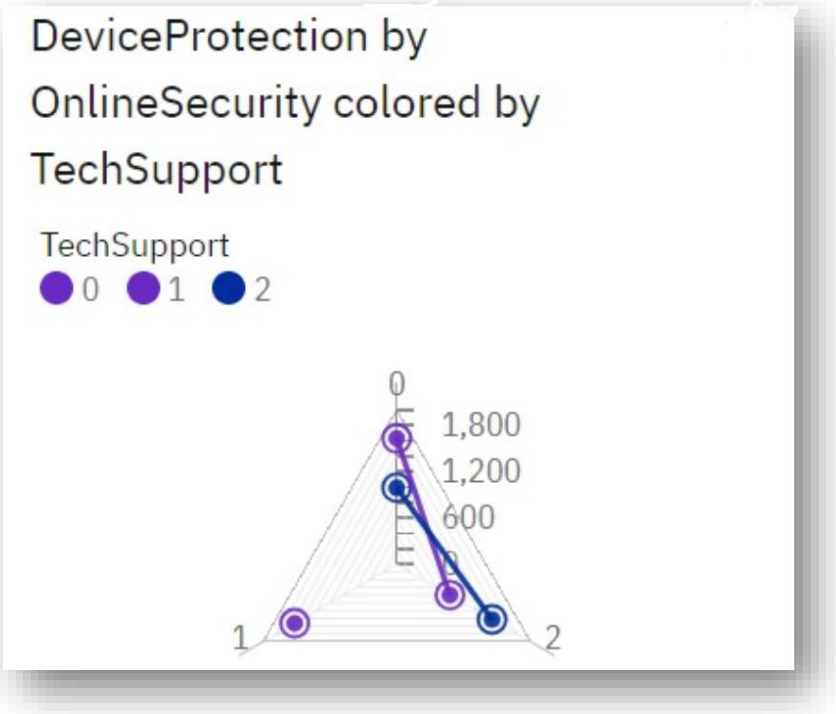
Streaming Movies when at 2 and Streaming TV when at 2 have the highest values of both Total Charges and Multiple Lines.
Total Charges weakly affects gender, the relationship is positively linear.

11. Box plot



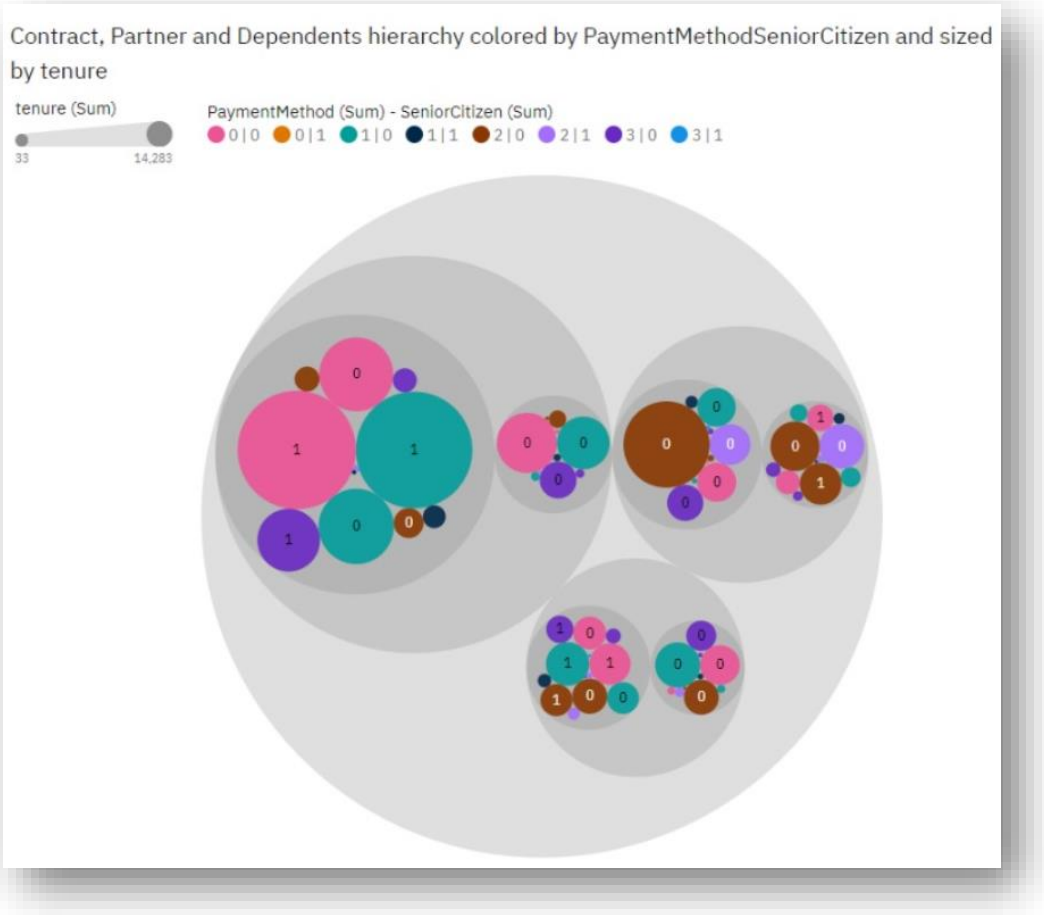
Contract 0 has the highest total Online Backup due to Dependents 0.
Dependents 0 and Tech Support 0 have the highest values of both Online Backup and Multiple Lines.
Partner 1 has the highest values of both Online Backup and Multiple Lines.

12. Radar



Online Security 0 has the highest values of both Device Protection and Multiple Lines.

13. Hierarchy Bubble



Dependents 0 and Partner 1 have the highest values of both Tenure and Multiple Lines.
Tenure is unusually high when the combination of Contract-Partner-Dependents and Payment Method-Senior Citizen is 0|1|0 and 2|0.
Payment Method 0 has the highest total Tenure due to Contract 2.

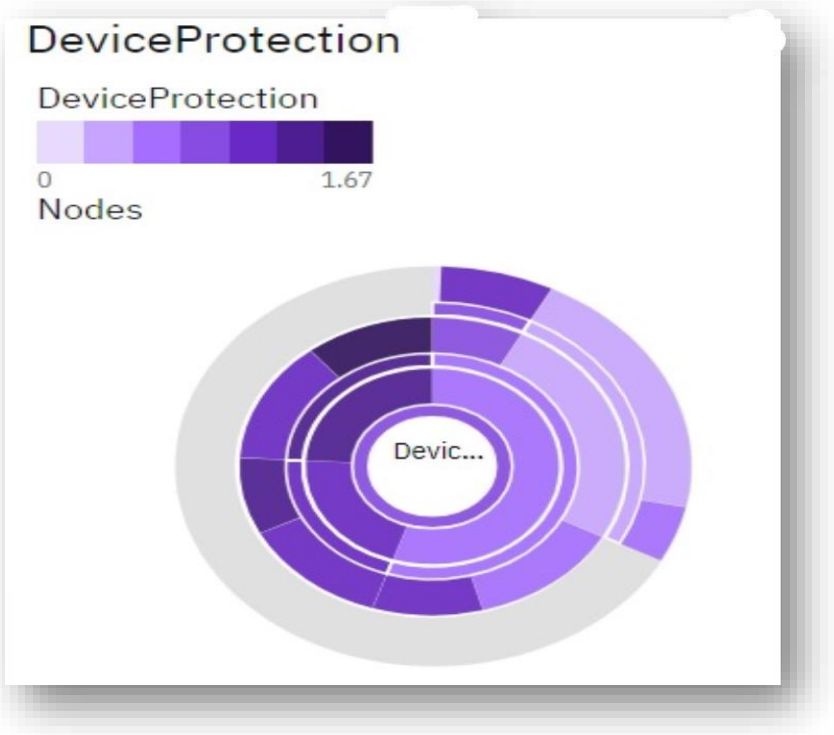
14. Table

TotalCharges, OnlineSecurity and InternetService

TotalCharges	OnlineSecurity	InternetService
16,056,168.7	5,564	6,148

The sum of Total Charges is 16,056,168.7
The sum of Online Security is 5,564.
The sum of Internet Service is 6,148.

15. Sunburst

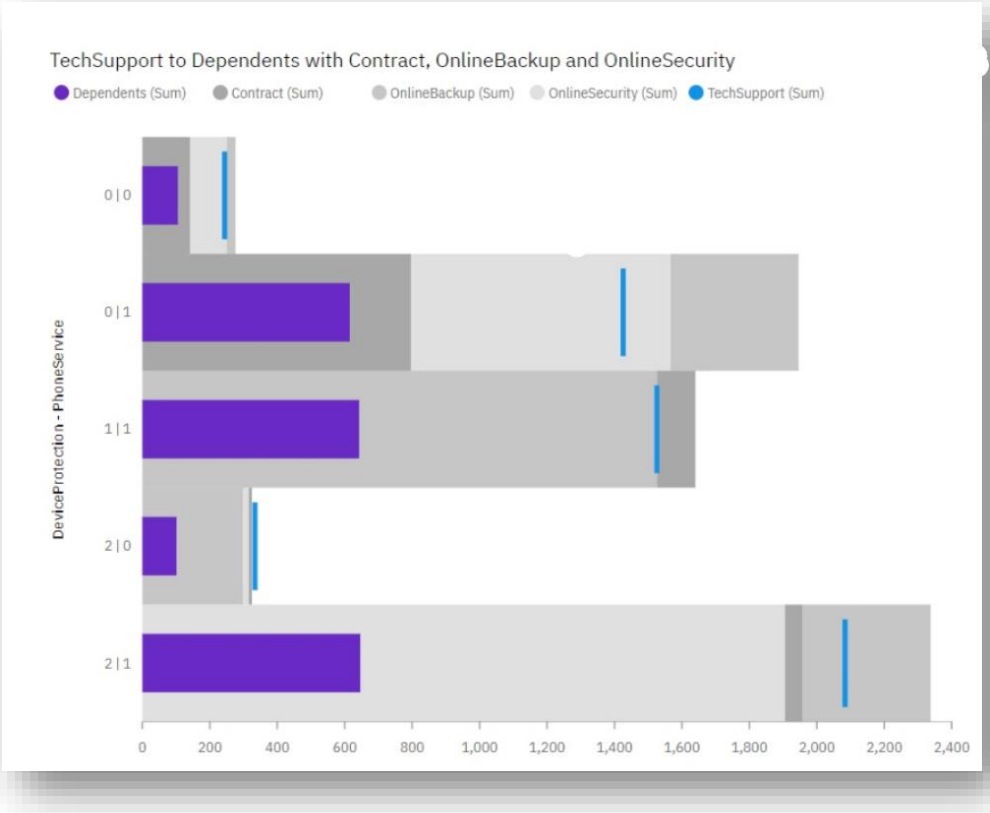


The plot depicts that Monthly Charges and Streaming Movies have a slight effect on Device Protection.

16. Tree Map

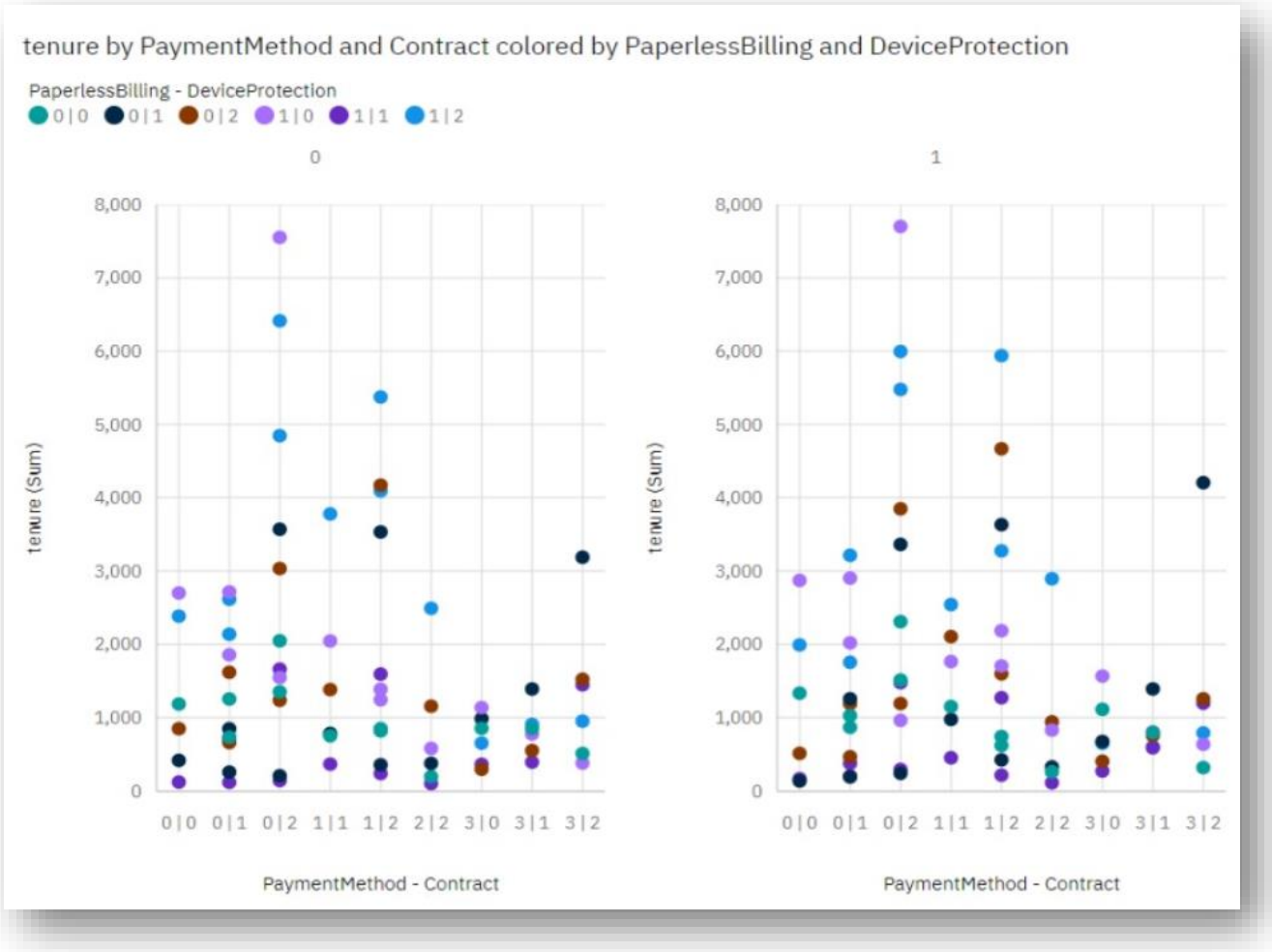
Internet Service 0 has the highest total Tech Support due to Online Backup 2.
Multiple Lines and Tech Support diverged the most when Internet Service 1.
Online Security 2 has the highest total Tech Support but is ranked #2 in total Multiple Lines.

17. Bullet



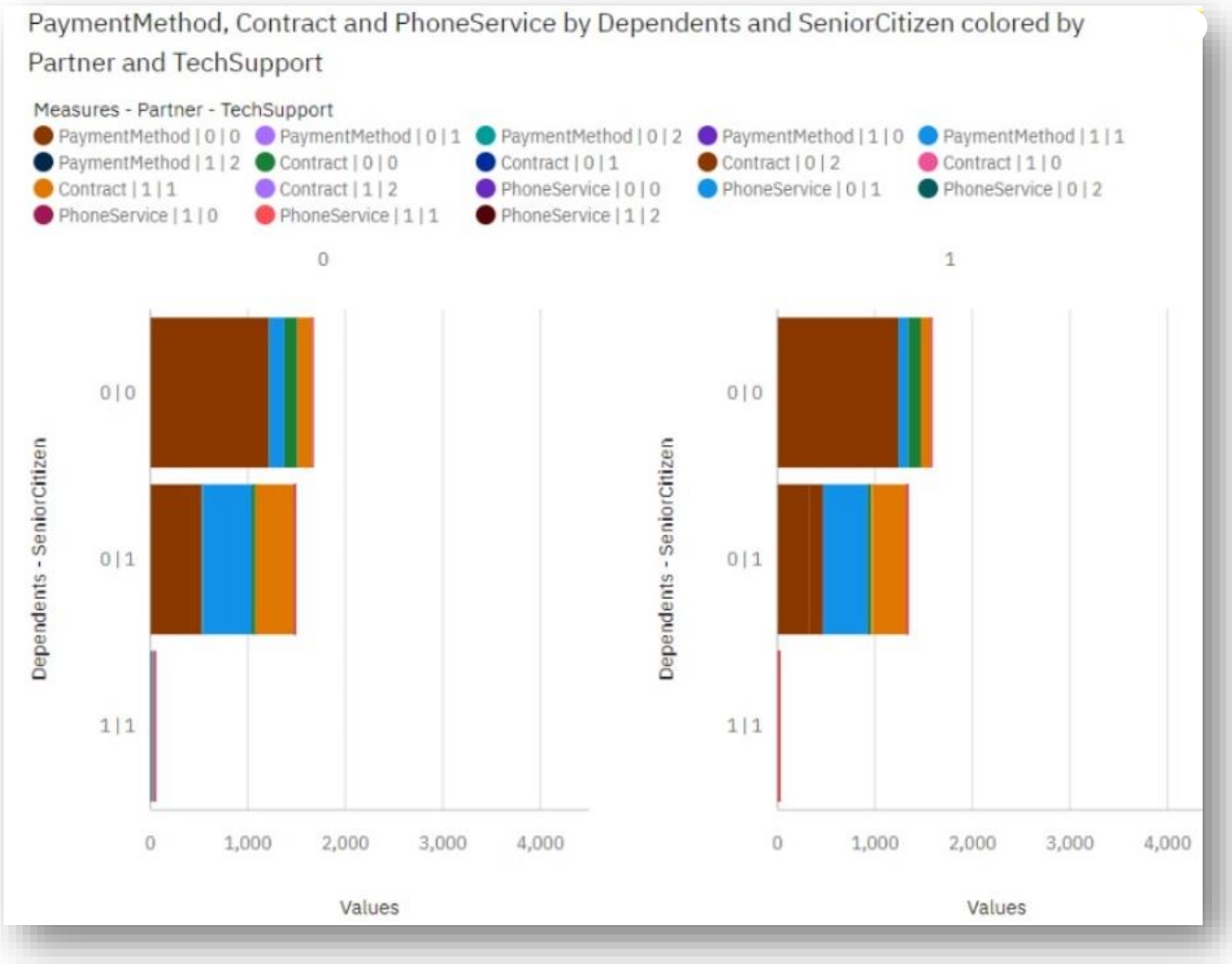
Device Protection 2 has the highest values of both Contract and Multiple Lines.
Multiple Lines and Contract diverged the most when Device Protection is 0.
Across all values of Device Protection-Phone Service, the sum of Dependents is over 2000.

18. Point



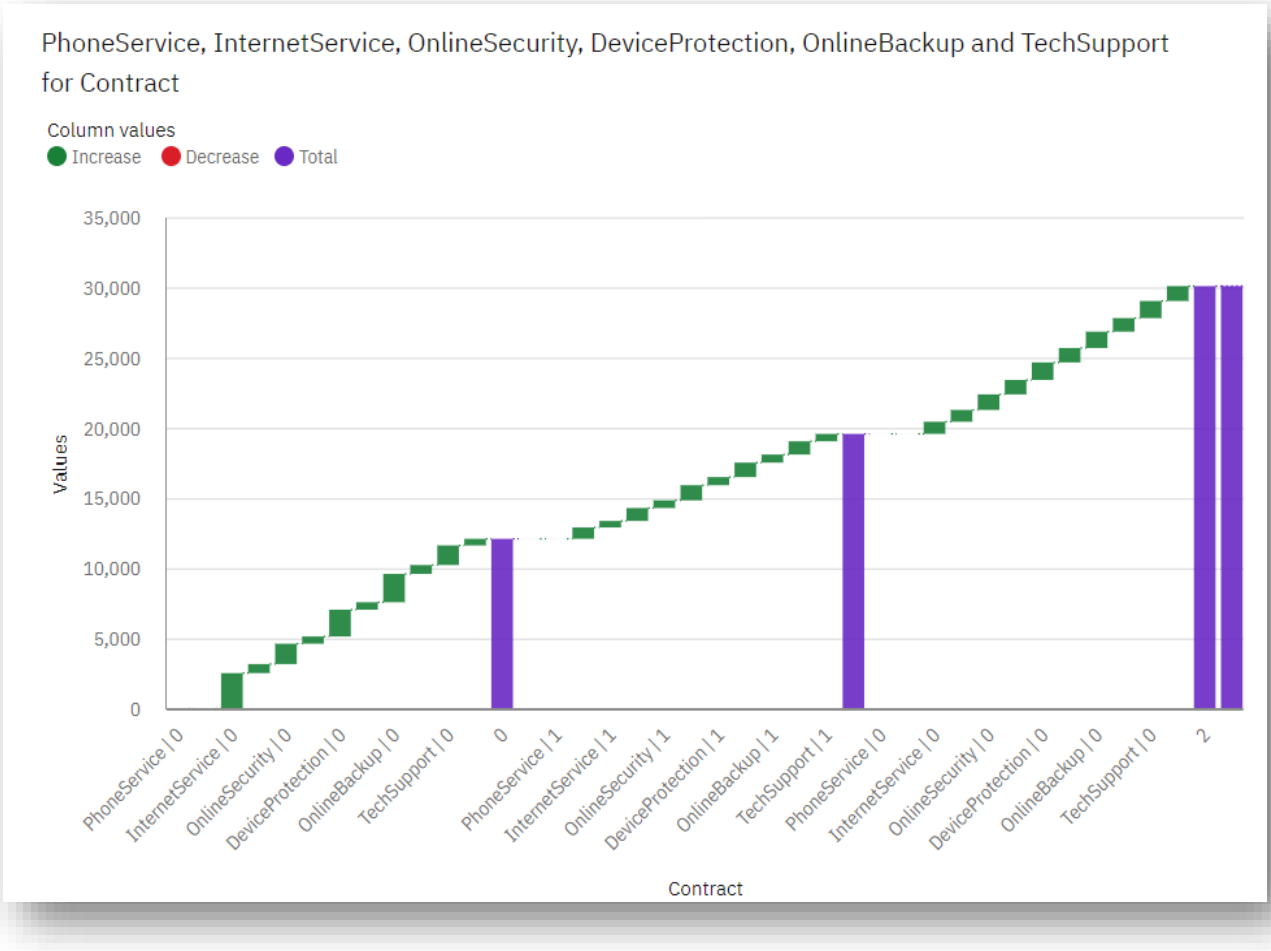
Gender 1 has the highest total Tenure due to Payment Method 0.
Device Protection 2 and Paperless Billing 1 have the highest values of both Tenure and Multiple Lines.
Payment Method 2 has the highest total Multiple Lines but is ranked #3 in total Tenure.
Payment Method 0 has the highest total Tenure but is ranked #2 in total Multiple Lines.

19. Stacked Bar



Gender 1 has the highest total Contract due to Dependents 0.
Dependents 0, Partner 1 and Senior Citizen 0 have the highest values of both Contract and Multiple Lines.
Tech Support 0 has the highest total Multiple Lines but is ranked #3 in total Contract.
Tech Support 2 has the highest total Contract but is ranked #2 in total Multiple Lines.
The overall number of results for Phone Service is over 7000.

20. Waterfall



Contract 0 has the highest total Device Protection due to Dependents 0.
0 is the most frequently occurring category of Contract with a count of 3875 items with Internet Service values (55 % of the total).
Multiple Lines and Device Protection diverged the most when Contract is 0, and when Multiple Lines was 902 higher than the Device Protection.
The total number of results for Device Protection, across all contracts, is over seven thousand.