Term Project

Arjun Varma

Bellevue University

DSC530-T301 Data Exploration and Analysis (2221-1)

Professor Shankar Parajulee

November 19, 2021

Contents

Cord-Cutting - Customer Churn in the Cable Industry 3

Did the Pandemic influence customer decisions with regards to, total cable disconnection or downgrading services? 3

Introduction: 3

The Dataset – Details 4

Methodology Used for Analysis 5

EDA & Data Wrangling 5

Description of Dataset Columns: 6

Preliminary EDA Data Analysis: 10

Correlation Analysis - Heatmap 12

Variables for Analysis 14

Explanation of Variables 14

Statistical Data Analysis 17

Histograms 17

Checking and Handling Outliers in Data – Box Plots 30

Summarizing Data Analysis – Bar Charts 35

Validating Correlations & Causation 35

Validating Correlation between Service Downgrade and Unresolved issues/Outages 36

Validating Correlation between Service Downgrade and Contracts 39

Validating correlation between service downgrade and TTS Tickets (Outages) 41

Validating correlation between service downgrade and customer reason 43

CDF – Cumulative Distribution Function 45

Scatter Plot – Analysis for Correlation 47

Logistic Regression 49

Summary and Conclusion 50

Constraints and Limitations Faced 50

Cord-Cutting - Customer Churn in the Cable Industry

Did the Pandemic influence customer decisions with regards to, total cable disconnection or downgrading services?

Introduction:

Cord cutting has been a common phenomenon which has gained speed over the past few years. Customers have been disconnecting Cable services due to multiple reasons; one of the primary reasons has been cost of the Cable service itself Vs. the streaming options which are cheap and available and being provided by big name companies.

However, since the beginning of the Pandemic and specifically early 2020, when the country went into lock down, something strange has happened. People were forced to work from home and have ended up not disconnecting their complete service package and have maintained Internet service which is crucial for communication and interactions for their jobs. Customers also want to be entertained so more and more people have been moving to the streaming option while they are stuck at home. In the end, research by surveys conducted has shown that customers are evaluating the Cable services they have against ‘value for money’, this is a perceived valuation for the cost of the services they are paying for.

This research paper attempts to pinpoint what variables are affecting customer decisions, regarding whether, to only cut their Video cable services or disconnect from Cable altogether. An open survey conducted by a third party indicated that customers did not perceive the value of the Cable products, directly proportional to the costs. This research will indicate what variables impacted this opinion of the customers. The information derived could be used to design a prediction-model for customer churn or downgrades however a final Model build has not been included as it’s out of the scope of this research paper.

The dataset which has been used is from a real company in existence *(Source: Public Limited Company, Private Data).*

The Dataset – Details

The dataset being used has 4000 records with 25 attributes. The data has been extracted for the city of Boston, MA and has data on both active and inactive customers from August 2021.

Third party surveys and actual data has already indicated an increase in cord-cutting; *In broadcast television, cord-cutting refers to the pattern of viewers, referred to as cord-cutters, cancelling their subscriptions to multichannel television services available over cable or satellite, dropping pay television channels or reducing the number of hours of subscription TV viewed in response to competition from rival media available over the Internet. This content is either free or significantly cheaper than the same content provided via cable.* However, the same customers who are cutting the cord are not disconnecting their Internet services.

Based on this information, the dataset being used will try to prove the following:

Hypothesis: Are customer are more likely to downgrade cable services, rather than, disconnecting cable altogether and do certain variables influence that decision?

Assumptions:

Customers need the Internet service to keep in touch with their offices and also for streaming and entertainment purposes. This would mean customers may disconnect video services but not Internet.

Methodology Used for Analysis

This research will involve:

1. EDA

2. Data Wrangling

3. Statistical Analysis

The tools being used are primarily Python 3 running via Jupyter Notebook in Anaconda.

EDA & Data Wrangling

Python Libraries used:

- matplotlib
- seaborn
- pandas
- pandas_profiling
- dtale
- numpy
- %matplotlib inline

Code used:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import pandas_profiling
import dtale
import numpy as np
%matplotlib inline
```

Description of Dataset Columns:

Column NAME	DESCRIPTION
CUSTOMER_TYPE	Type of customer: Residential or Business
Gender	Gender/sex of subscriber: M/F
HOUSE_ID	Unique company generated identifier for a customer account address-location
MARITAL_STATUS	Married or Single: SGL/MRD
INCOME_CODE	Income divided into range
Education	Highest education received, SCLG=College not finished, HSCL=High School, COLG=4-year college, GRAD=Graduate School, NHSD=High School not completed
CHILDREN_NUM_HH	Number of children at home
Age	Age
Truckrolls	Number of times a truck was sent to the customer's house for a trouble call
Unresolved_calls	Open tickets and trouble calls, not yet resolved
TTS_TOTAL_TICKETS	Number of outage tickets
TENURE_BY_ACTIVE_MO	Time since subscriber has been a customer (in months)
VIDEO_DISCONNECT	Flag field 1=Disconnected, 0= Not disconnected
CUSTOMER_DISCONNECT	Flag field 1=Disconnected, 0= Not disconnected
CUSTOMER_DISCONNECT_DT	Date when customer disconnected all services
CUST_DISCON_REASON_NAME	Reason provided for total disconnect
CONTRACT_FLAG	1=Contract exists, 0=No contract
CONTRACT_START_DATE	Date when contract started
CONTRACT_TERM_PERIOD	In months, 12- or 24-month contract
CALC_MRC_CURR_MONTH	Current monthly recurring charges
CALC_MRC_MONTH_01	Previous month's monthly recurring charges
NUMBER_OF_PRODUCTS	count of products/services on customer account
PREV_NUMBER_OF_PRODUCTS	count of products/services on customer account from previous month
Srvc_Dwngrd	1=service was cancelled for specific products, 0=service was not cancelled
PRODUCT_MIX	Type of services/products customer has. CDV=Digital Voice(Phone), HSD=High Speed Internet, XH=Home Security
PREV_PRODUCT_MIX	Type of services/products customer had in previous month
PAYMENT_RECURRING	Automated payment setup by customer every month

1. **print("The shape of the dataset is : ", df.shape)**

   The shape of the dataset is: (4000, 26)

2. **print(df.head())**

```
  CUSTOMER_TYPE Gender MARITAL_STATUS INCOME_CODE Education  CHILDREN_NUM_HH  \
0   RESIDENTIAL      M            SGL    175-200K      GRAD                0
1   RESIDENTIAL      F            SGL        <15K      HSCL                0
2   RESIDENTIAL      F            SGL      35-50K      SCLG                0
3   RESIDENTIAL      M            MRD        <15K      HSCL                0
4   RESIDENTIAL      M            MRD    150-175K      NHSD                0

   Age  Truckrolls Unresolved_calls  TTS_TOTAL_TICKETS  TENURE_BY_ACTIVE_MO  \
0   38           0                5                  5                   71
1   28           0                1                  0                   42
2   66           0                1                  0                  127
3   47           0                3                  0                   47
4   52           0                1                  0                  127

   VIDEO_DISCONNECT  CUSTOMER_DISCONNECT CUSTOMER_DISCONNECT_DT  \
0                 1                    0                    NaN
1                 1                    0                    NaN
2                 1                    0                    NaN
3                 1                    0                    NaN
4                 1                    0                    NaN

  CUST_DISCON_REASON_NAME  CONTRACT_FLAG CONTRACT_START_DATE  \
0          Too expensive              1            8/6/2021
1          Too expensive              1           8/18/2021
2          Too expensive              1            8/7/2021
3          Too expensive              0                   ?
4          Too expensive              0                   ?

   CONTRACT_TERM_PERIOD CALC_MRC_CURR_MONTH CALC_MRC_MONTH_01  \
0                    12               65.95            220.84
1                    12               79.95            133.69
2                    12               94.95            159.43
3                     0              119.95            215.24
4                     0              100.95            180.37

   NUMBER_OF_PRODUCTS  PREV_NUMBER_OF_PRODUCTS  Srvc_Dwngrd PRODUCT_MIX  \
0                   1                        3            1    HSD ONLY
1                   1                        3            1    HSD ONLY
2                   1                        3            1    HSD ONLY
3                   1                        3            1    HSD ONLY
4                   1                        2            1    HSD ONLY

  PREV_PRODUCT_MIX  PAYMENT_RECURRING
0    VIDEO/HSD/CDV                  0
1    VIDEO/HSD/CDV                  0
2    VIDEO/HSD/CDV                  0
```

```
3    VIDEO/HSD/CDV                  0
4        VIDEO/HSD                  1
```

3. df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3999 entries, 0 to 3998
Data columns (total 27 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   CUSTOMER_TYPE           3999 non-null   object 
 1   HOUSE_ID                3999 non-null   float64
 2   Gender                  3999 non-null   int64  
 3   MARITAL_STATUS          3999 non-null   int64  
 4   INCOME_CODE             3999 non-null   object 
 5   Education               3999 non-null   object 
 6   CHILDREN_NUM_HH         3999 non-null   int64  
 7   Age                     3999 non-null   int64  
 8   Truckrolls              3999 non-null   int64  
 9   Unresolved_calls        3999 non-null   int64  
 10  TTS_TOTAL_TICKETS       3999 non-null   int64  
 11  TENURE_BY_ACTIVE_MO     3999 non-null   int64  
 12  VIDEO_DISCONNECT        3999 non-null   int64  
 13  CUSTOMER_DISCONNECT     3999 non-null   int64  
 14  CUSTOMER_DISCONNECT_DT  2019 non-null   object 
 15  CUST_DISCON_REASON_NAME  3999 non-null   int64  
 16  CONTRACT_FLAG           3999 non-null   int64  
 17  CONTRACT_START_DATE     3999 non-null   object 
 18  CONTRACT_TERM_PERIOD    3999 non-null   int64  
 19  CALC_MRC_CURR_MONTH     3999 non-null   float64
 20  CALC_MRC_MONTH_01       3999 non-null   float64
 21  NUMBER_OF_PRODUCTS      3999 non-null   int64  
 22  PREV_NUMBER_OF_PRODUCTS  3999 non-null   int64  
 23  Srvc_Dwngrd             3999 non-null   int64  
 24  PRODUCT_MIX             3999 non-null   object 
 25  PREV_PRODUCT_MIX        3999 non-null   object 
 26  PAYMENT_RECURRING       3999 non-null   int64  
dtypes: float64(3), int64(17), object(7)
```

4. df.nunique()

```
CUSTOMER_TYPE                2
Gender                       3
MARITAL_STATUS               2
INCOME_CODE                 12
Education                    5
CHILDREN_NUM_HH              7
Age                         81
Truckrolls                   4
Unresolved_calls            13
TTS_TOTAL_TICKETS           13
TENURE_BY_ACTIVE_MO        127
VIDEO_DISCONNECT             2
CUSTOMER_DISCONNECT          2
```

```
CUSTOMER_DISCONNECT_DT       19
CUST_DISCON_REASON_NAME       4
CONTRACT_FLAG                 2
CONTRACT_START_DATE         387
CONTRACT_TERM_PERIOD          3
CALC_MRC_CURR_MONTH        1529
CALC_MRC_MONTH_01          2285
NUMBER_OF_PRODUCTS            4
PREV_NUMBER_OF_PRODUCTS       4
Srvc_Dwngrd                   2
PRODUCT_MIX                  12
PREV_PRODUCT_MIX             11
PAYMENT_RECURRING             3
dtype: int64
```

5. **df.isnull().sum()**

```
CUSTOMER_TYPE                 0
Gender                        0
MARITAL_STATUS                0
INCOME_CODE                   0
Education                     0
CHILDREN_NUM_HH               0
Age                           0
Truckrolls                    0
Unresolved_calls              0
TTS_TOTAL_TICKETS             0
TENURE_BY_ACTIVE_MO           0
VIDEO_DISCONNECT              0
CUSTOMER_DISCONNECT           0
CUSTOMER_DISCONNECT_DT     1980
CUST_DISCON_REASON_NAME    2005
CONTRACT_FLAG                 0
CONTRACT_START_DATE           0
CONTRACT_TERM_PERIOD          0
CALC_MRC_CURR_MONTH           0
CALC_MRC_MONTH_01             0
NUMBER_OF_PRODUCTS            0
PREV_NUMBER_OF_PRODUCTS       0
Srvc_Dwngrd                   0
PRODUCT_MIX                   0
PREV_PRODUCT_MIX              0
PAYMENT_RECURRING             0
dtype: int64
```

Preliminary EDA Data Analysis:

CUSTOMER_DISCONNECT_DT = 1980 NULL Values
CUST_DISCON_REASON_NAME = 2005 NULL Values

- These can be disregarded as the values will only be populated if the customer disconnected services.

- One row where CALC_MRC_CURR_MONTH = $30122 was removed as this was a major outlier bringing the count of records in dataset from 4000 to 3999.

CALC_MRC_CURR_MONTH and CALC_MRC_MONTH_01 show up as Dtype = OBJECT

- We will convert them to FLOAT/Numeric by using the following syntax:

-
  ```
  df['CALC_MRC_CURR_MONTH'] = df['CALC_MRC_CURR_MONTH'].replace(r'\s+', np.nan, regex=True)
  df['CALC_MRC_CURR_MONTH'] = pd.to_numeric(df['CALC_MRC_CURR_MONTH'])
  ```

-
  ```
  df['CALC_MRC_MONTH_01'] = df['CALC_MRC_MONTH_01'].replace(r'\s+', np.nan, regex=True)
  df['CALC_MRC_MONTH_01'] = pd.to_numeric(df['CALC_MRC_MONTH_01'])
  ```

CUST_DISCON_REASON_NAME has values which are Strings

- In order for this field to be counted towards correlation, we will replace the values to numeric Using the following syntax

  ```
  df.replace({'CUST_DISCON_REASON_NAME':{'Too expensive':14, 'Too many outages':15, 'NA':0, 'Unresolved issues':17, 'Video service not required': 18}})
  ```

  Too expensive = 14
  Too many outages = 15
  NA = 0
  Unresolved issues = 17
  Video service not required = 18

- As the values have been converted to Numeric, we will change the Dtype of CUST_DISCON_REASON_NAME to FLOAT

  ```
  df['CUST_DISCON_REASON_NAME'] =
  df['CUST_DISCON_REASON_NAME'].replace(r'\s+', np.nan, regex=True)
  df['CUST_DISCON_REASON_NAME'] =
  pd.to_numeric(df['CUST_DISCON_REASON_NAME'])
  ```

- We will implement the same changes for fields: MARITAL STATUS & GENDER

  ```
  df.replace({'MARITAL_STATUS':{'MRD':22, 'SGL':25}})
  df['MARITAL_STATUS'] = df['MARITAL_STATUS'].replace(r'\s+', np.nan, regex=True)
  df['MARITAL_STATUS'] = pd.to_numeric(df['MARITAL_STATUS'])
  ```

  ```
  df.replace({'Gender':{'M':99, 'F':100}})
  df['Gender'] = df['Gender'].replace(r'\s+', np.nan, regex=True)
  df['Gender'] = pd.to_numeric(df['Gender'])
  ```

  MARITAL_STATUS
  MRD = 22
  SGL = 25

  GENDER
  M = 99
  F = 100

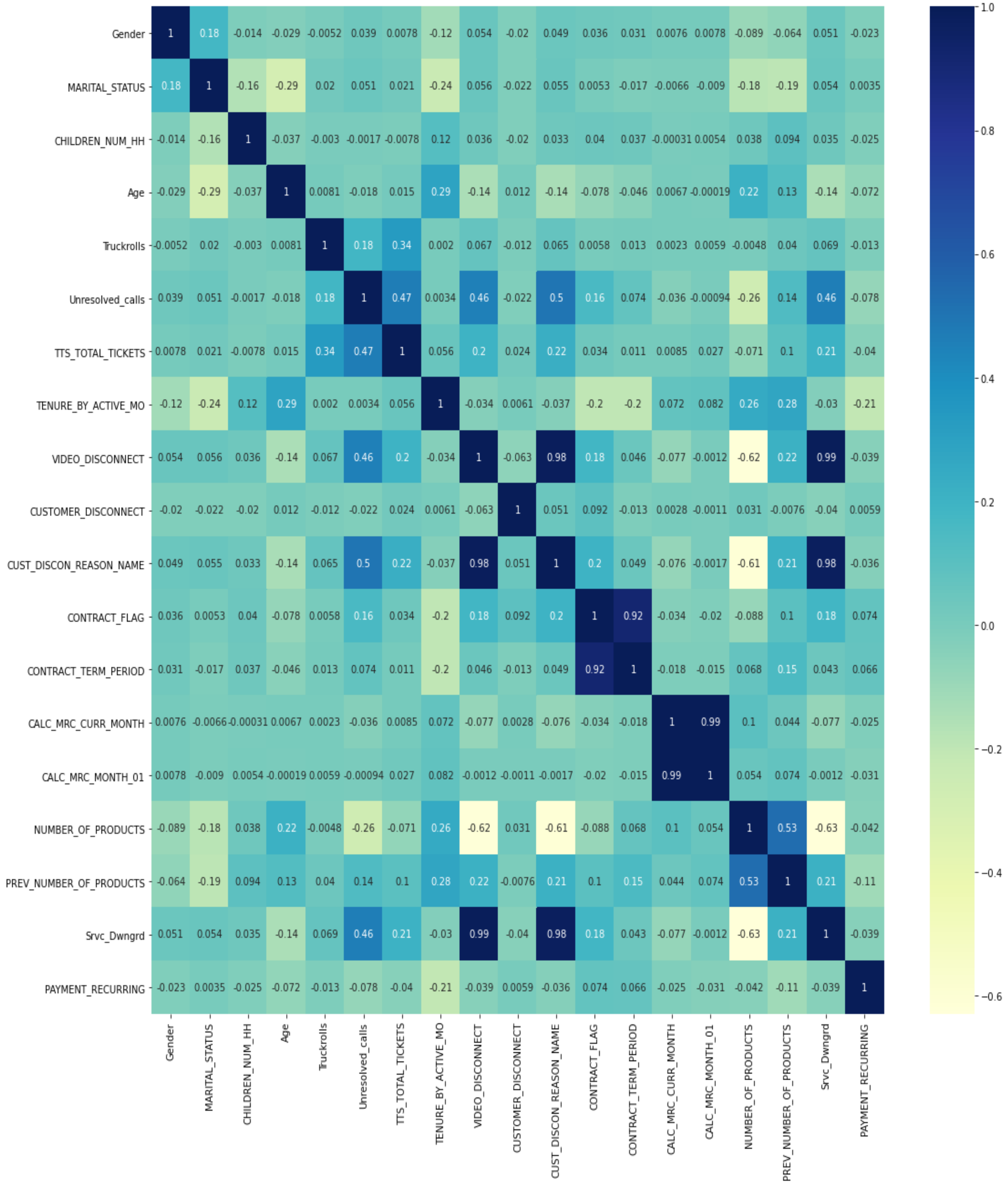
- At this point, the data has been cleaned and transformed for analysis.

- We will load the Seaborn Libraries for Correlation Chart analysis

  - Code:

  -
    ```
    plt.figure(figsize=(18,18))
    sns.heatmap(df.corr(), cmap="YlGnBu", annot=True)
    ```

Correlation Analysis - Heatmap



Correlation Co-Variance for Dependent Variable

- Code:
  show_correlations(df, show_chart=False)["Srvc_Dwngrd"].sort_values(ascending=False)

```
Srvc_Dwngrd                1.000000 - Positive Correlation
VIDEO_DISCONNECT           0.988565 - Positive Correlation
CUST_DISCON_REASON_NAME    0.975659 - Positive Correlation
Unresolved_calls           0.459341 - Positive Correlation
PREV_NUMBER_OF_PRODUCTS    0.213765 - Positive Correlation
TTS_TOTAL_TICKETS          0.209495 - Positive Correlation
CONTRACT_FLAG              0.181953 - Positive Correlation
CUST_DISCON_REASON_NAME    0.981133 - Positive Correlation
Truckrolls                 0.069499 - Positive Correlation
MARITAL_STATUS             0.053684 - Positive Correlation
Gender                     0.051223 - Positive Correlation
CONTRACT_TERM_PERIOD       0.042799 - Positive Correlation
CHILDREN_NUM_HH            0.035463 - Positive Correlation
CALC_MRC_MONTH_01         -0.001213 - Negative Correlation
TENURE_BY_ACTIVE_MO       -0.030407 - Negative Correlation
PAYMENT_RECURRING         -0.038698 - Negative Correlation
CUSTOMER_DISCONNECT       -0.039671 - Negative Correlation
CALC_MRC_CURR_MONTH       -0.076889 - Negative Correlation
Age                       -0.139682 - Negative Correlation
NUMBER_OF_PRODUCTS        -0.629395 - Negative Correlation
Name: Srvc_Dwngrd, dtype: float64
```

Variables for Analysis

Based on the Correlation chart, the following would be the variables chosen for this research:

Dependent Variable to prove Hypothesis:

SRVC_DWNGRD

Independent Variables:

TRUCKROLLS

UNRESOLVED_CALLS

TTS_TOTAL_TICKETS

CONTRACT_FLAG

VIDEO_DISCONNECT

CUST_DISCON_REASON_NAME

The research will attempt to prove whether customers are specifically disconnecting Video services Vs. cancelling, the complete Cable account subscription.

Lastly, if the customers do end up cancelling certain services, what factors influences their decision.

Explanation of Variables

SRVC_DWNGRD – This is our dependent variable and stands for Service Downgrade. This is important because it's a 'flag' field with values 1/0. A value of '1' means the customer downgraded their service and a value of '0' means there was no change or downgrade to their service from the previous month. This will be closely aligned with the field: Video_Disconnect as it would show whether the Service Downgrade was specifically for Video services only. The dependent variable has positive correlation with several fields; however, we are picking 6 dependent variables only. Customer demographic variables are not being chosen as the data does not indicate a specific demographic category in terms of

Age, Income level or Education which is influencing service downgrade decisions. Data in this column is from the current reporting month.

TRUCKROLLS – This is an independent variable and signifies the count of the number of times, a truck had to be sent with a technician to the customer's house. This is significant as the only time a truck is sent out is because a problem with the services is intense enough to warrant a physical check. This is usually associated with an outage of services. The higher number of truck rolls to a customer's house reduces customer satisfaction and indicates an instability in service delivery. This also works towards a customer's feeling about 'perceived value of the service' they pay for. A higher number of truck rolls, results in a lower perceived value of the service, in the customer's mind. Data in this column is from the previous reporting month.

UNRESOLVED_CALLS – This is an independent variable and is the count of trouble calls that are open or unresolved. This again reduces the perceived value of the service; the customer is paying for. A higher number of unresolved calls indicates bad customer service and quality of delivery services. Data in this column is from the previous reporting month.

TTS_TOTAL_TICKETS – This is an independent variable and is the count of tickets opened for specifically service outages. This counts the number of tickets opened and which were closed/resolved. This again reduces the perceived value of the service; the customer is paying for. A higher number of outages indicates low quality of delivery as there is interruption of services. Data in this column is from the previous reporting month.

CONTRACT_FLAG – This is an independent variable and is a flag field with values 1/0. A value of '1' indicates the customer is under contract with the cable company to retain their services for a specific amount of time and a value of '0' indicates that there is no contract and the customer can cancel services any time. A contract is either 12 or 24 months and once a customer signs a contract they get a promotional rate for their bundled services, which is a lot cheaper cost than, getting all those services separately. If a customer ends a contract pre-maturely there is a penalty ranging from $150-$299. This is a key independent variable which will show if customers avoid paying the penalty by not ending their contract but simply downgrading their service, which would technically still mean they are active customers.

VIDEO_DISCONNECT – This is an independent variable flag field. This has values of 1/0. A value of ‘1’ indicates that a customer discontinued, specifically their Video services only. A value of ‘0’ indicates there was no change to their Video services. This will be used to track if customers specifically cut down their Video service and still retained other services, like Mobile, Internet, etc. Data in this column is from the current reporting month.

CUST_DISCON_REASON_NAME – This is a comment field, the customer specifies a reason as to why they downgraded or totally disconnected their services. This is an independent variable and is key to finding out the reason for a disconnect or downgrade. This field has been converted into numeric values in order for analysis purposes. The mapping of the string values to numeric can be found here.

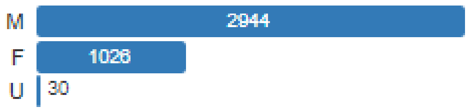
Statistical Data Analysis

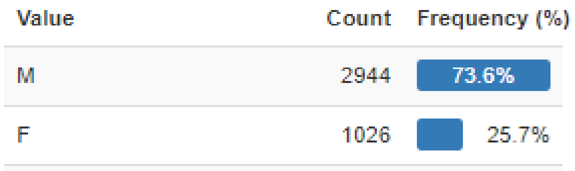
- Library used: Pandas Profiling, matplotlib
- Code: import pandas_profiling
- Code: import matplotlib.pyplot as plt
- Code: import seaborn as sns

Histograms

Demographic Data

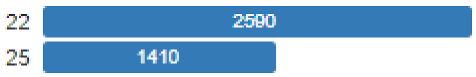
Gender

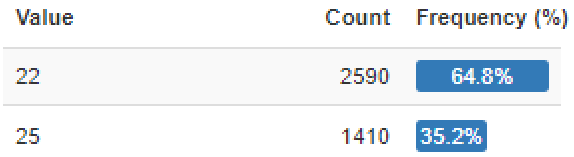




Value	Count	Frequency (%)
M	2944	73.6%
F	1026	25.7%

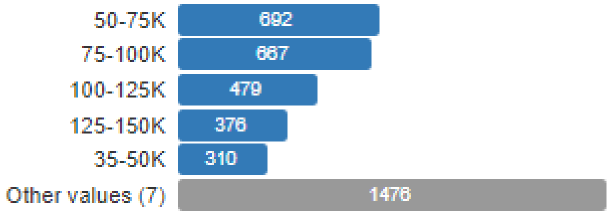
Marital Status

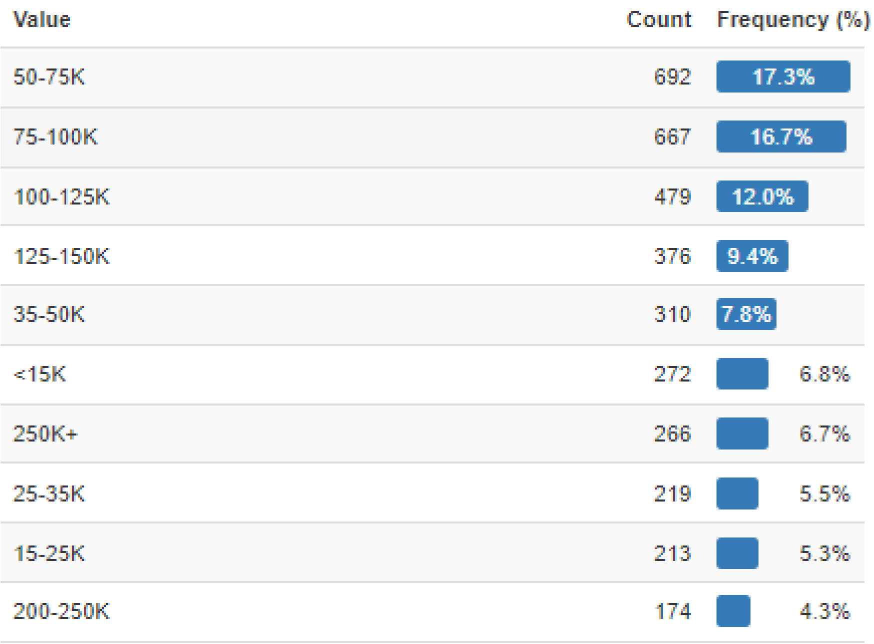




Value	Count	Frequency (%)
22	2590	64.8%
25	1410	35.2%

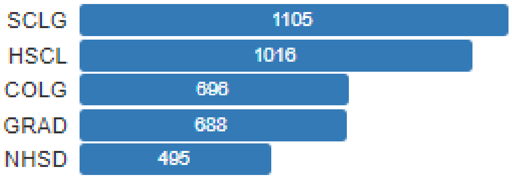
Income

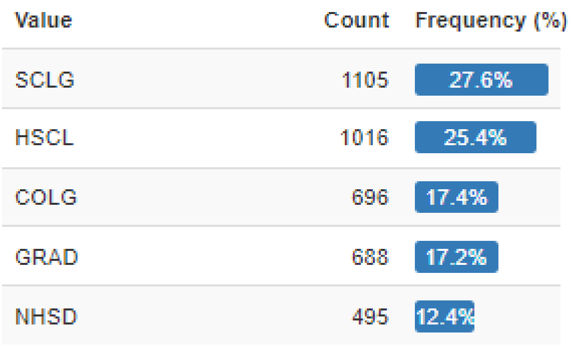




Value	Count	Frequency (%)
50-75K	692	17.3%
75-100K	667	16.7%
100-125K	479	12.0%
125-150K	376	9.4%
35-50K	310	7.8%
<15K	272	6.8%
250K+	266	6.7%
25-35K	219	5.5%
15-25K	213	5.3%
200-250K	174	4.3%

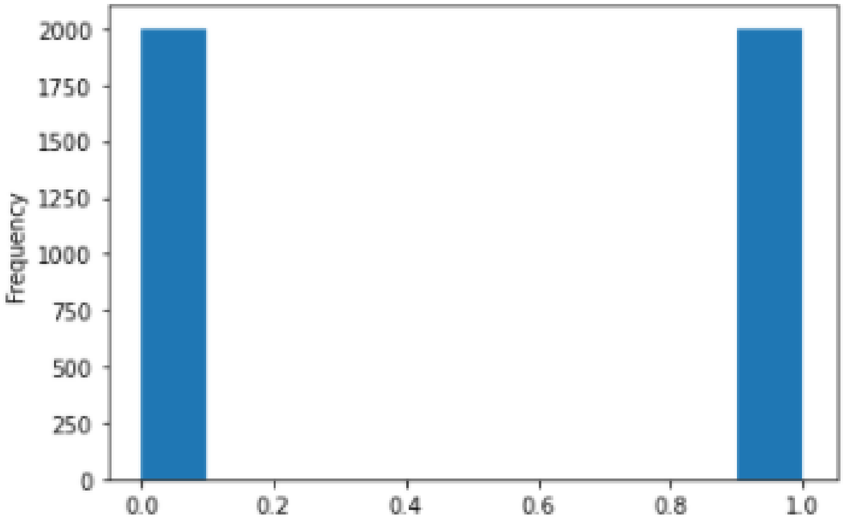
Education

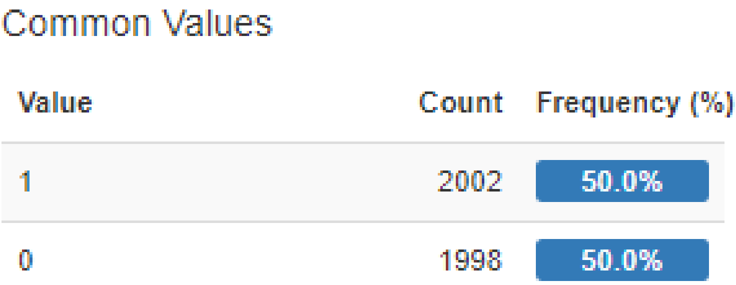




Value	Count	Frequency (%)
SCLG	1105	27.6%
HSCL	1016	25.4%
COLG	696	17.4%
GRAD	688	17.2%
NHSD	495	12.4%

Variable - SRVC_DWNGRD





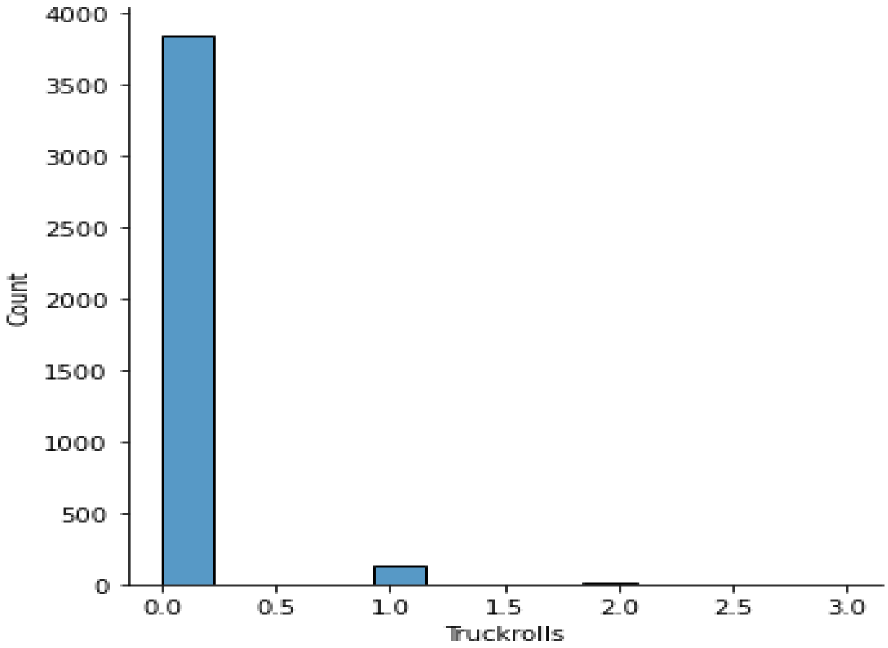
Common Values

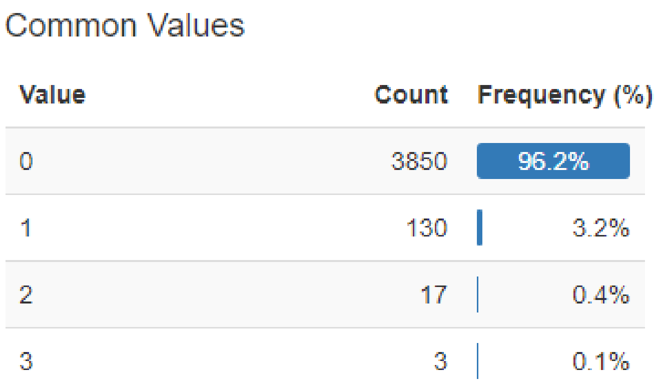
Value	Count	Frequency (%)
1	2002	50.0%
0	1998	50.0%

```
count    4000.000000
mean        0.500500
std         0.500062
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
mode         1.000000
Name: Srvc_Dwngrd, dtype: float64
```

There is a 50-50% split for this variable in the dataset of 4000 records. There are 2002 records which indicate customers who downgraded service and 1998 records where customers did not make any change to service. There are no outliers.

Variable - TRUCKROLLS

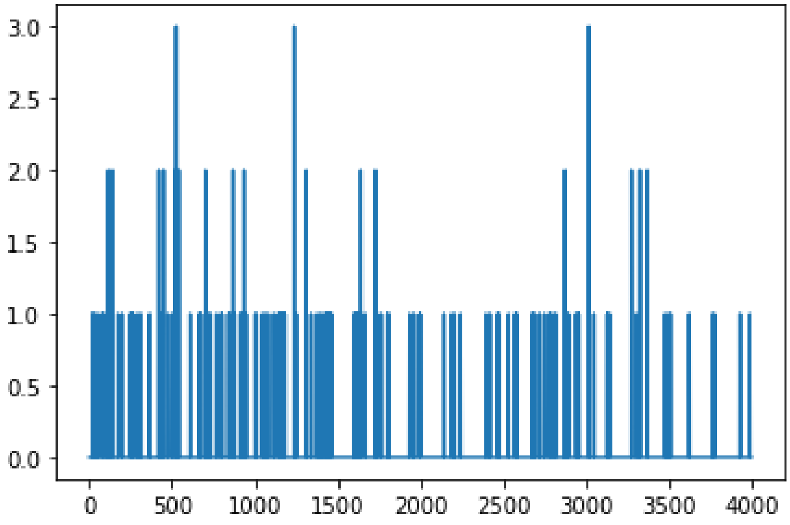




Common Values

Value	Count	Frequency (%)
0	3850	96.2%
1	130	3.2%
2	17	0.4%
3	3	0.1%

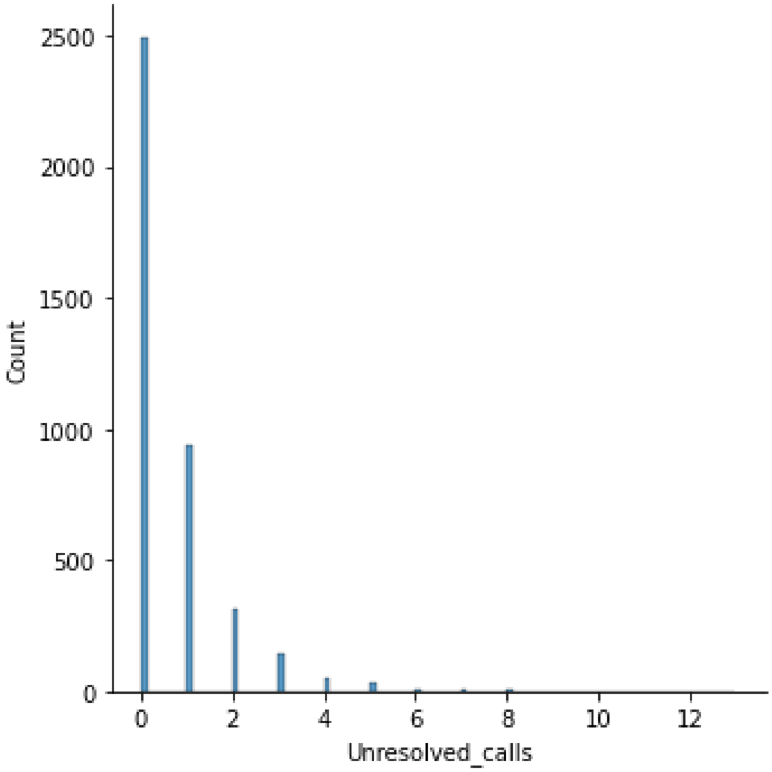
Including Line Plot to show better distribution:

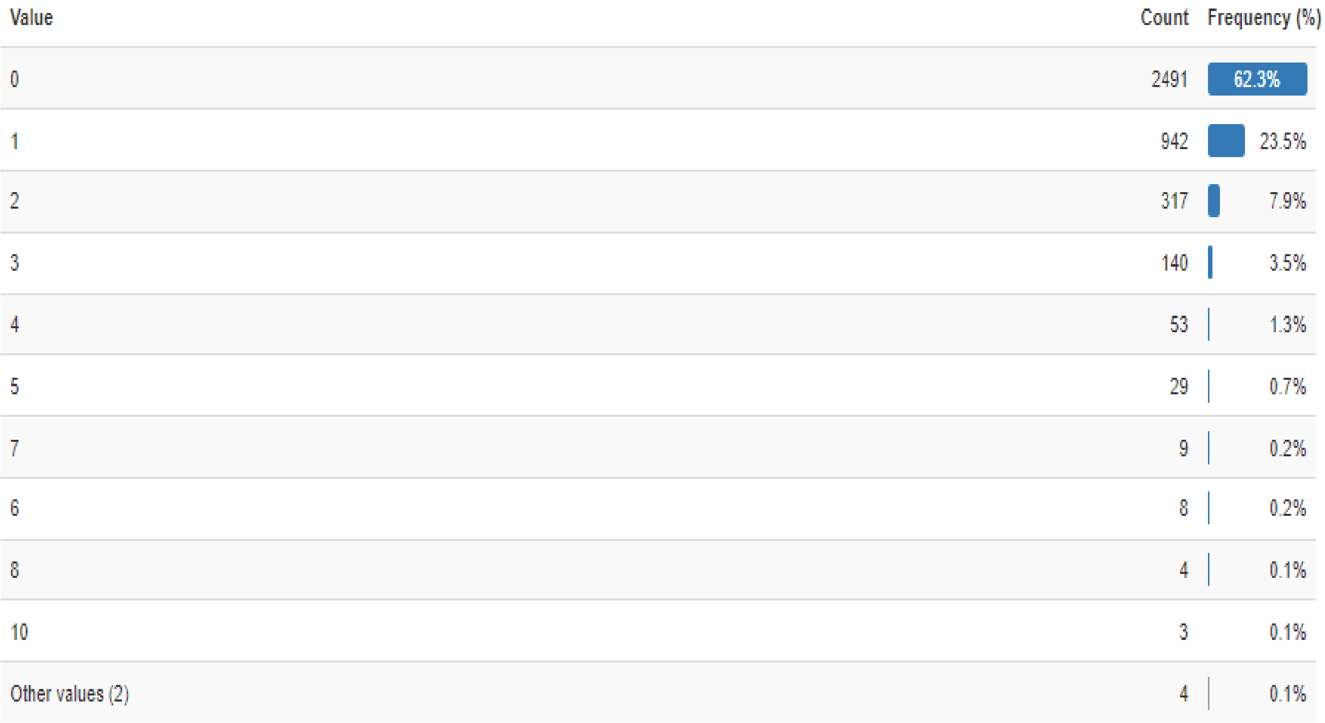


```
count    4000.000000
mean        0.043250
std         0.233223
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         3.000000
mode         0.000000
Name: Truckrolls, dtype: float64
```

Data shows that a majority of customers did not experience a truck roll. The truck rolls ranged from 1 to 3 times.

Variable - UNRESOLVED_CALLS



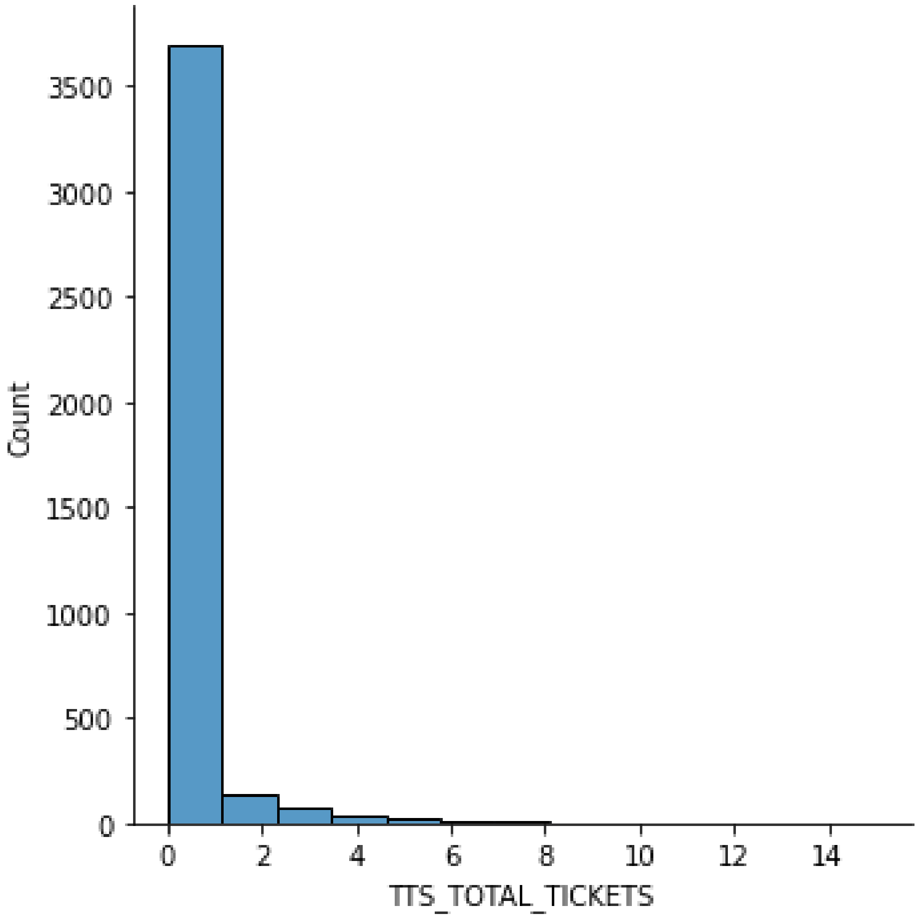


Value	Count	Frequency (%)
0	2491	62.3%
1	942	23.5%
2	317	7.9%
3	140	3.5%
4	53	1.3%
5	29	0.7%
7	9	0.2%
6	8	0.2%
8	4	0.1%
10	3	0.1%
Other values (2)	4	0.1%

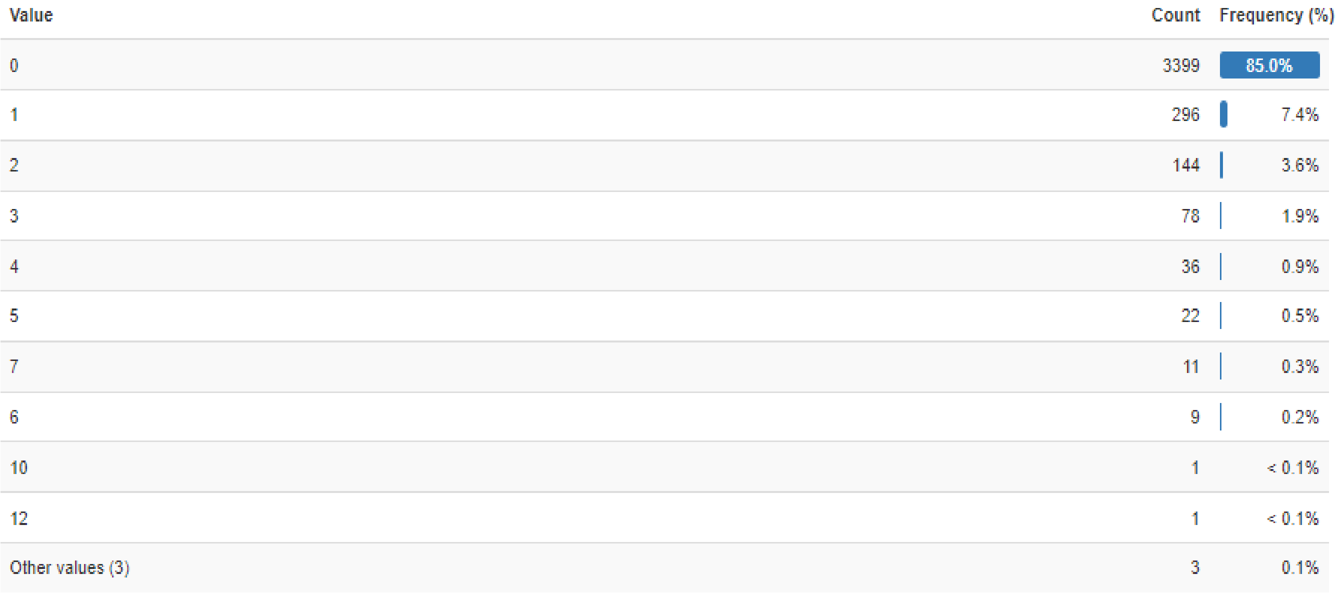
```
count    4000.000000
mean        0.641500
std         1.128626
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max        13.000000
mode        0.000000
Name: Unresolved_calls, dtype: float64
```

We can see in the data that there is a high number of customers who have unresolved calls; around 38% have unresolved issues and 62% do not have any.

Variable - TTS_TOTAL_TICKETS



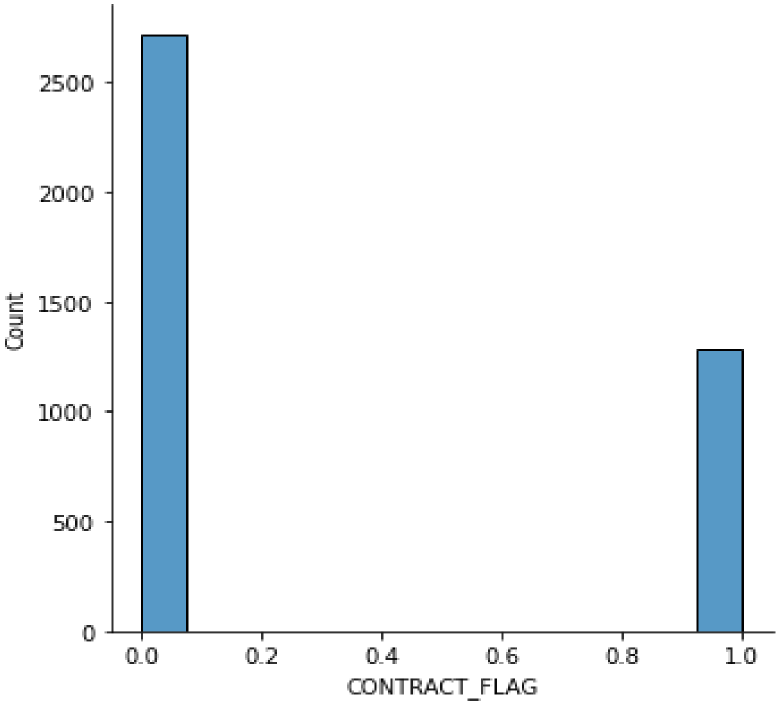
```
count    4000.000000
mean        0.315000
std         0.979548
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max        15.000000
mode        0.000000
Name: TTS_TOTAL_TICKETS, dtype: float64
```



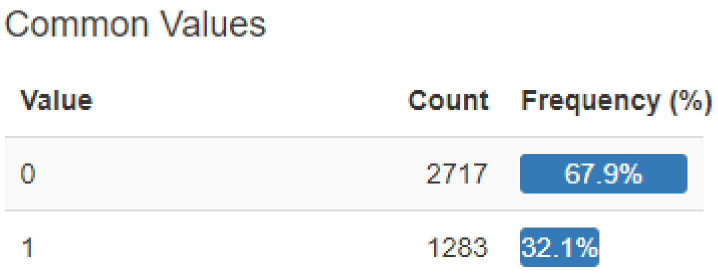
Value	Count	Frequency (%)
0	3399	85.0%
1	296	7.4%
2	144	3.6%
3	78	1.9%
4	36	0.9%
5	22	0.5%
7	11	0.3%
6	9	0.2%
10	1	< 0.1%
12	1	< 0.1%
Other values (3)	3	0.1%

Data shows a majority of customers not facing an outage and no TTS Tickets being opened.

Variable - CONTRACT_FLAG



```
count    4000.000000
mean        0.320750
std         0.466823
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         1.000000
mode         0.000000
Name: CONTRACT_FLAG, dtype: float64
```

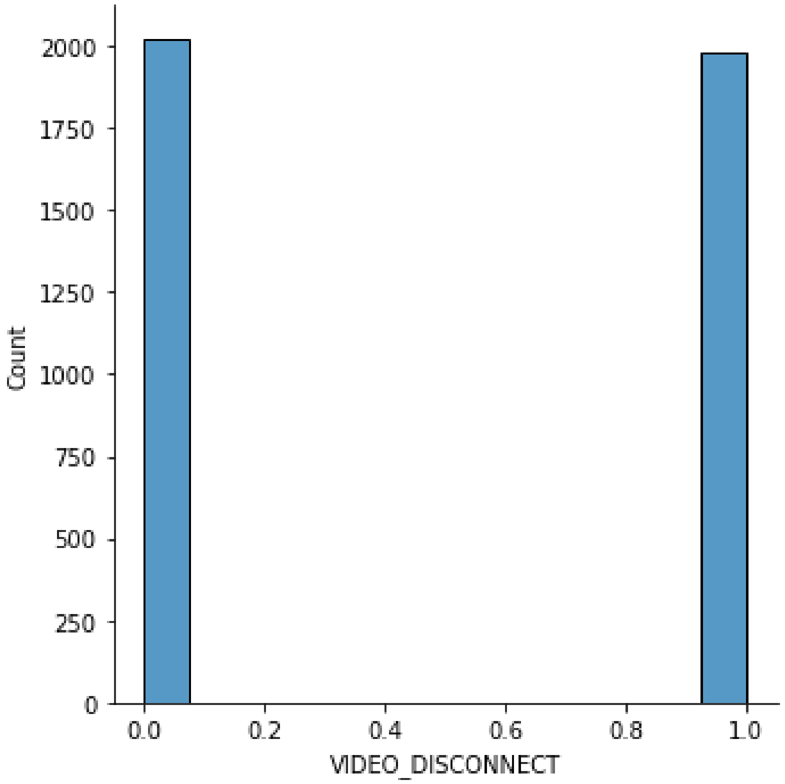


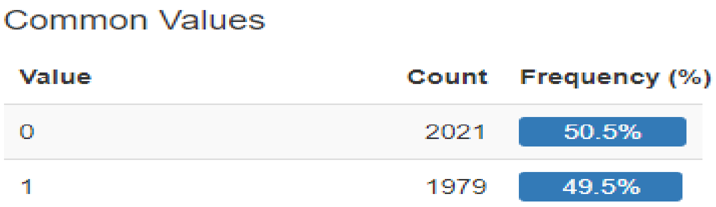
Common Values

Value	Count	Frequency (%)
0	2717	67.9%
1	1283	32.1%

Data shows a number of customers who are bound by a 12- or 24-month contract; 32.1% of customers are under contract to retain services

Variable - VIDEO_DISCONNECT





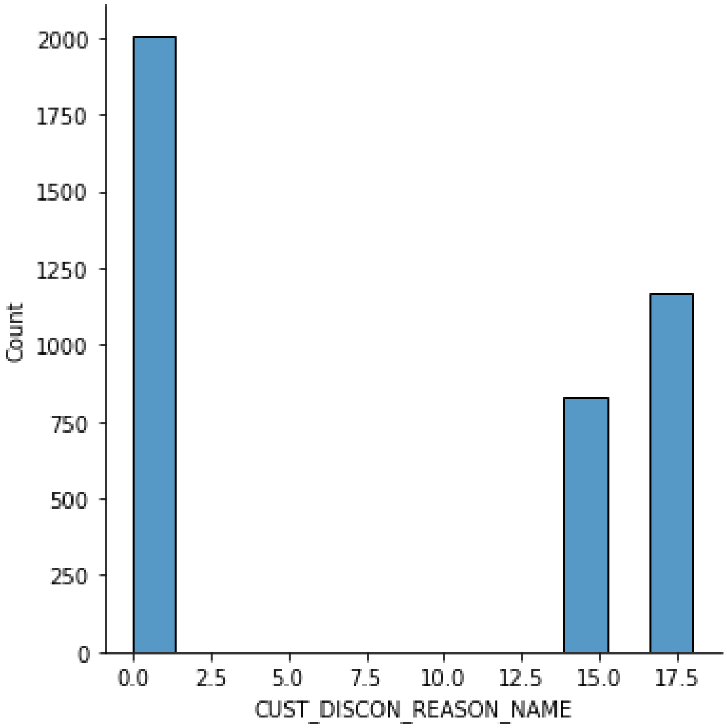
Common Values

Value	Count	Frequency (%)
0	2021	50.5%
1	1979	49.5%

```
count    4000.000000
mean        0.494750
std         0.500035
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         1.000000
mode        0.000000
Name: VIDEO_DISCONNECT, dtype: float64
```

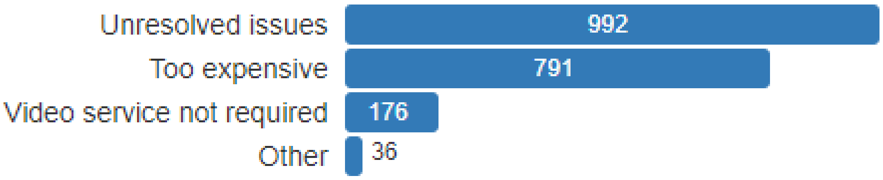
Data shows an almost even split between customers who did not downgrade services Vs. customers who downgraded. The data also indicates that the 49.5% customers downgrading services, specifically chose to remove Video services

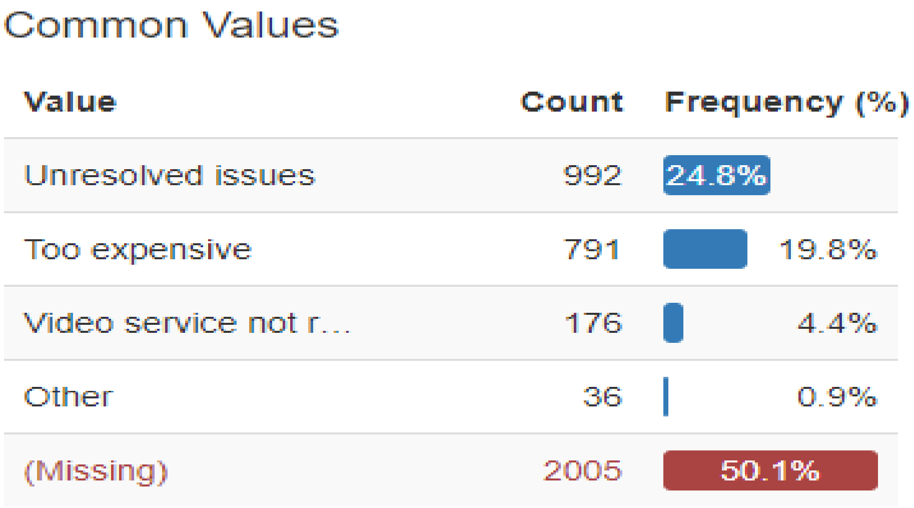
Variable - CUST_DISCON_REASON_NAME



Refer to mapping of reason names:

Too expensive = 14
Too many outages = 15
NA = 0
Unresolved issues = 17
Video service not required = 18





Common Values

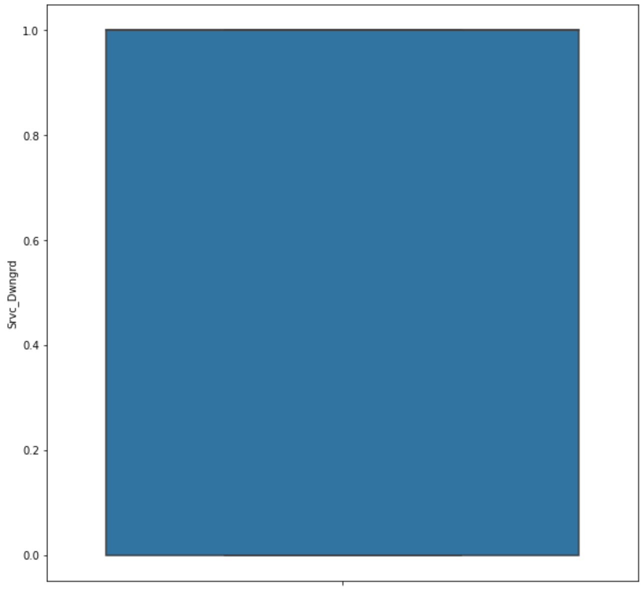
Value	Count	Frequency (%)
Unresolved issues	992	24.8%
Too expensive	791	19.8%
Video service not r...	176	4.4%
Other	36	0.9%
(Missing)	2005	50.1%

```
count    4000.000000
mean        7.911500
std         8.008508
min         0.000000
25%         0.000000
50%         0.000000
75%        17.000000
max        18.000000
mode        0.000000
Name: CUST_DISCON_REASON_NAME, dtype: float64
```

Data shows customer reasons provided when they disconnect or downgrade services. 50% records do not have reasons allocated as those customers did not disconnect or downgrade. There are outliers present for 36 records only as they are classified as ‘Other’ however the quality of Outliers is very small so we will not make any changes to handle these as they account for less than 1% of the whole dataset.

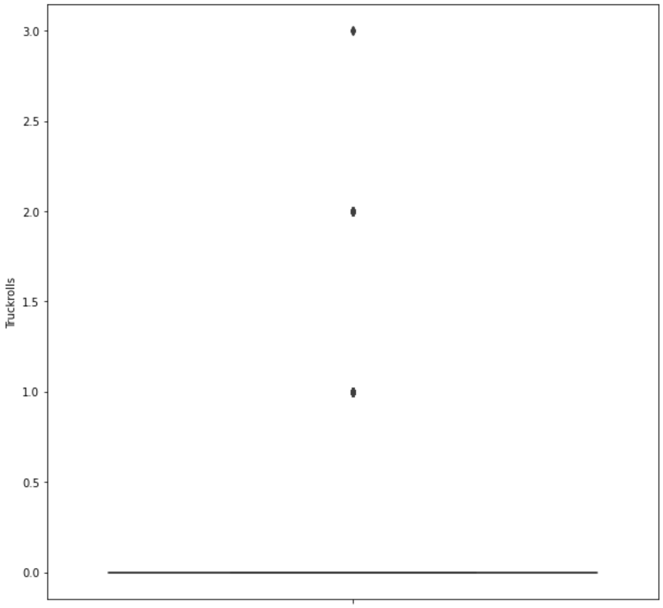
Checking and Handling Outliers in Data – Box Plots

Variable - SRVC_DWNGRD



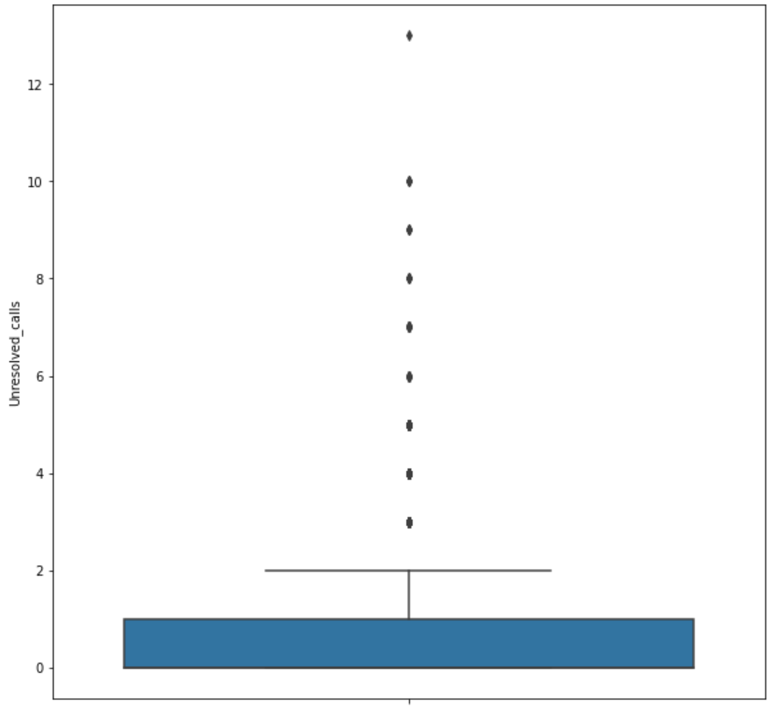
No Outliers present. Values are 1 or 0

Variable – TRUCKROLLS



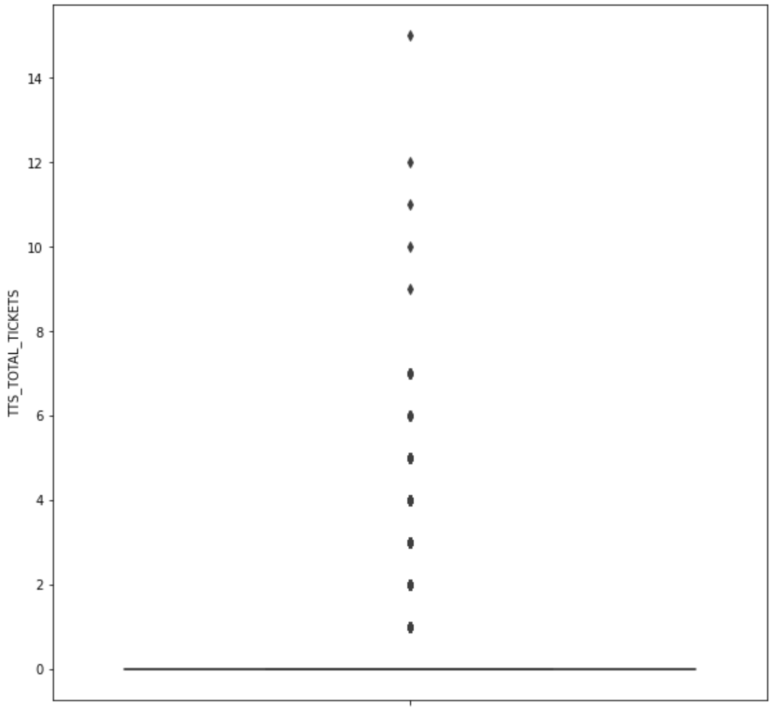
Data shows majority of records having zero or no Truck rolls. There are outliers with Truck Rolls of 1, 2 and 3 however these will be included in the analysis. It is possible to have 'repeat' trouble calls and truck rolls to the same customer account if the issue has not been fixed the first time so this data is valid.

Variable - UNRESOLVED_CALLS



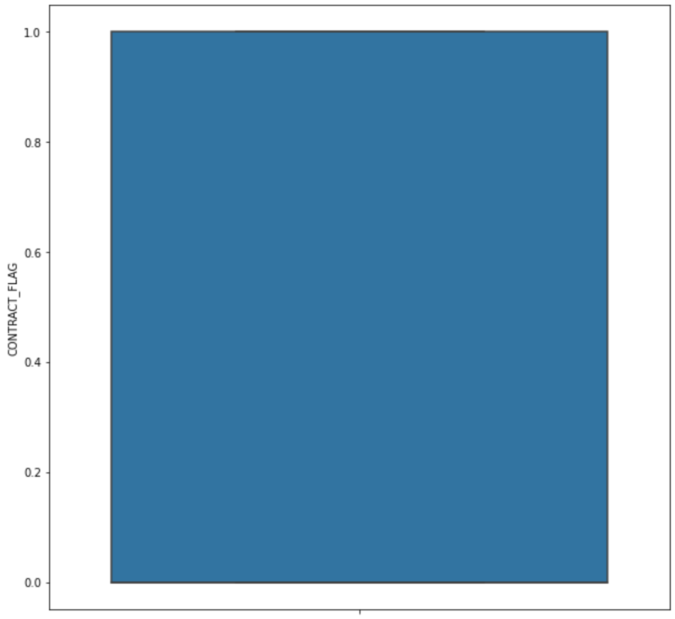
Data shows outliers from a 25% threshold of 2 unresolved calls and increases to 3 and all the way to 15. Having so many unresolved calls is not common but does happen and we will not make any changes to the data but will include these in the analysis. It is important to know if the number of unresolved calls plays a part in the customer's decision to disconnect or downgrade services. In this case the number of records related to the outliers is high so we need to include them in that analysis.

Variable – TTS_TOTAL_TICKETS



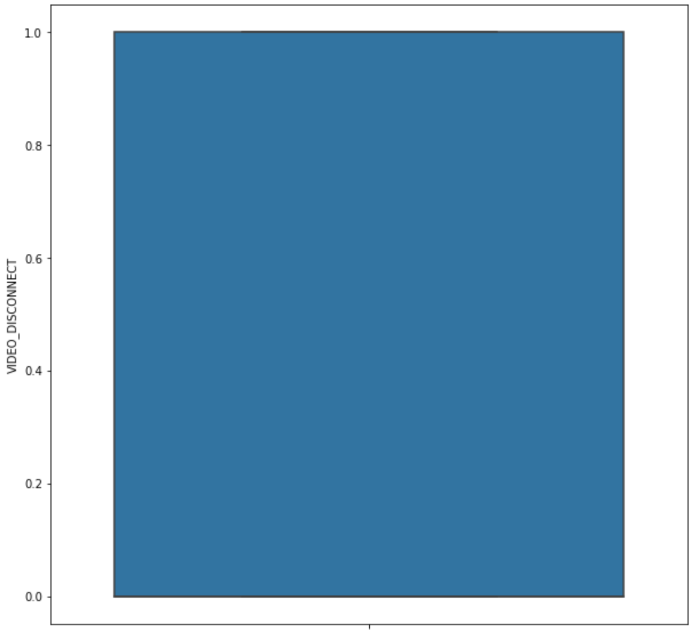
Data shows majority of customers have zero Outage tickets opened. There are outliers from 1 ticket opened to 16 tickets. This is not common but can happen and we will be including these in the analysis. It is important to know if the number of Total TTS/Outage tickets played a part in the customer's decision to disconnect or downgrade services.

Variable - CONTRACT_FLAG



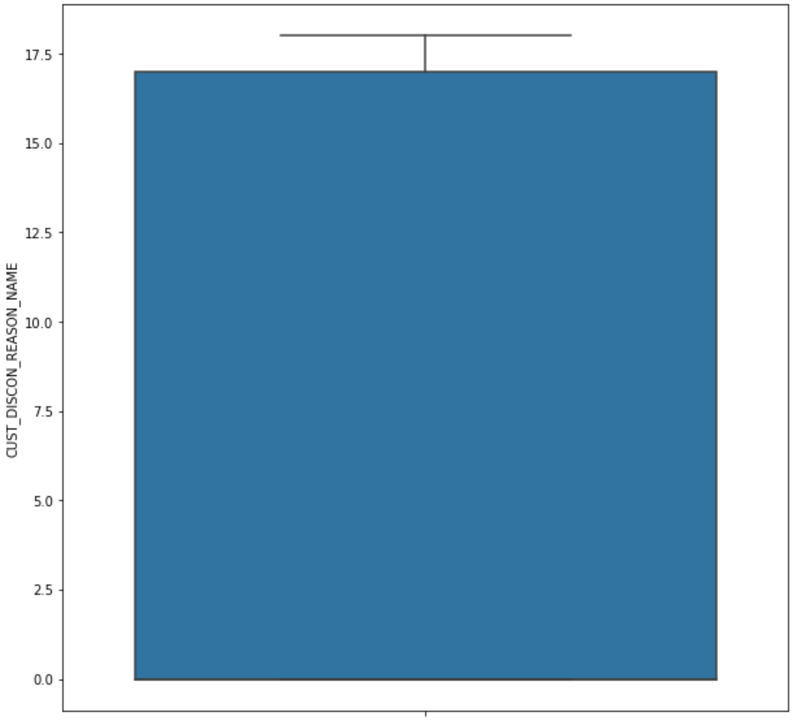
No Outliers present. Values are 1 or 0

Variable - VIDEO_DISCONNECT



No Outliers present. Values are 1 or 0

Variable - CUST_DISCON_REASON_NAME



No Outliers present.

Customer Disconnect Reason mapping to Numeric values:

Too expensive = 14
Too many outages = 15
NA = 0
Unresolved issues = 17
Video service not required = 18

Summarizing Data Analysis – Bar Charts

EDA and Data Wrangling has been completed and we have accounted for Outliers. Based on what we have seen so far, we need to confirm if there is causation present in the correlation indicated.

Revisiting the Hypothesis:

Are, customer more likely to downgrade cable services, rather than, disconnecting cable altogether and do certain variables influence that decision?

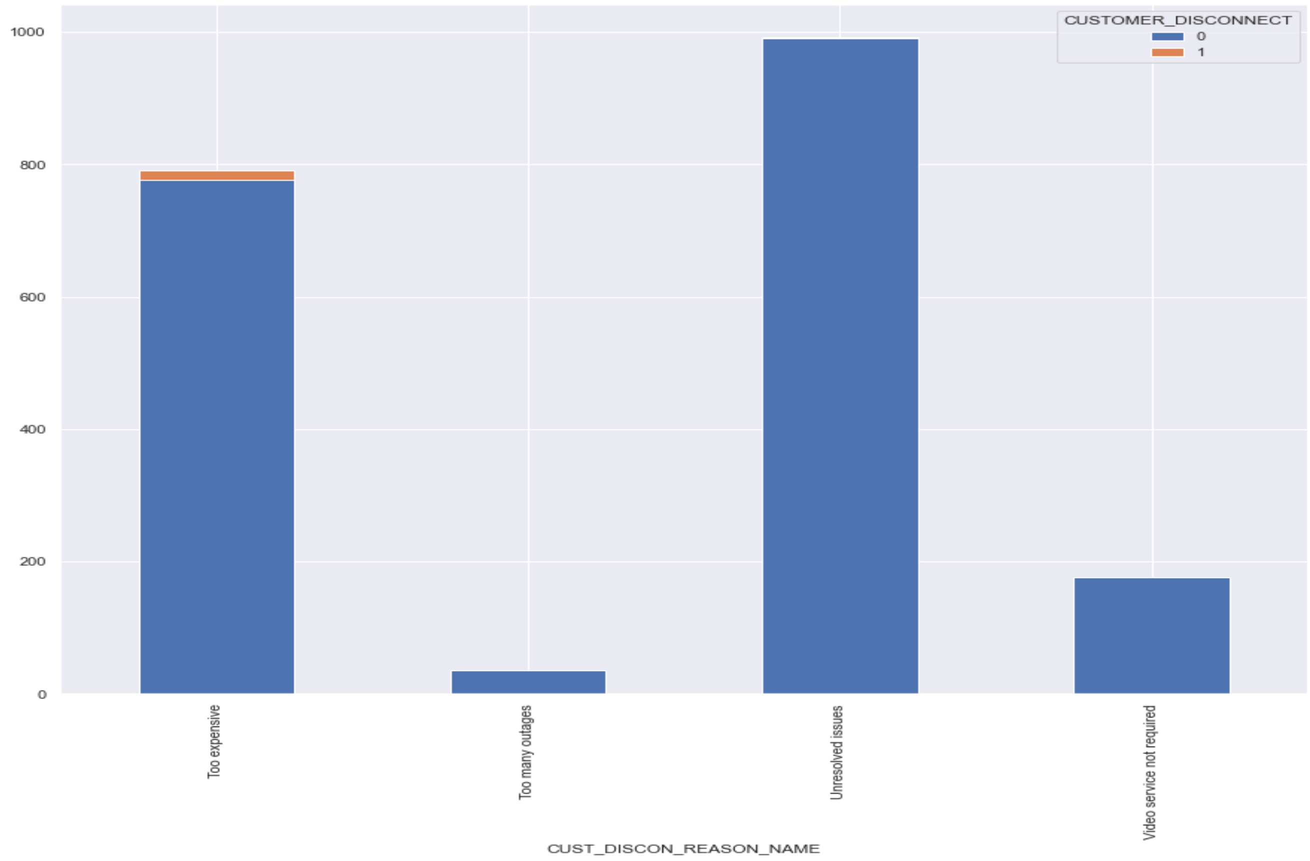
Validating Correlations & Causation

- We started with a dataset of 4000 records
- After data cleaning we have a final dataset of 3999 records

For the Month of August, 2021

Total Service disconnects = **16**

Active Customers remaining = **3983**

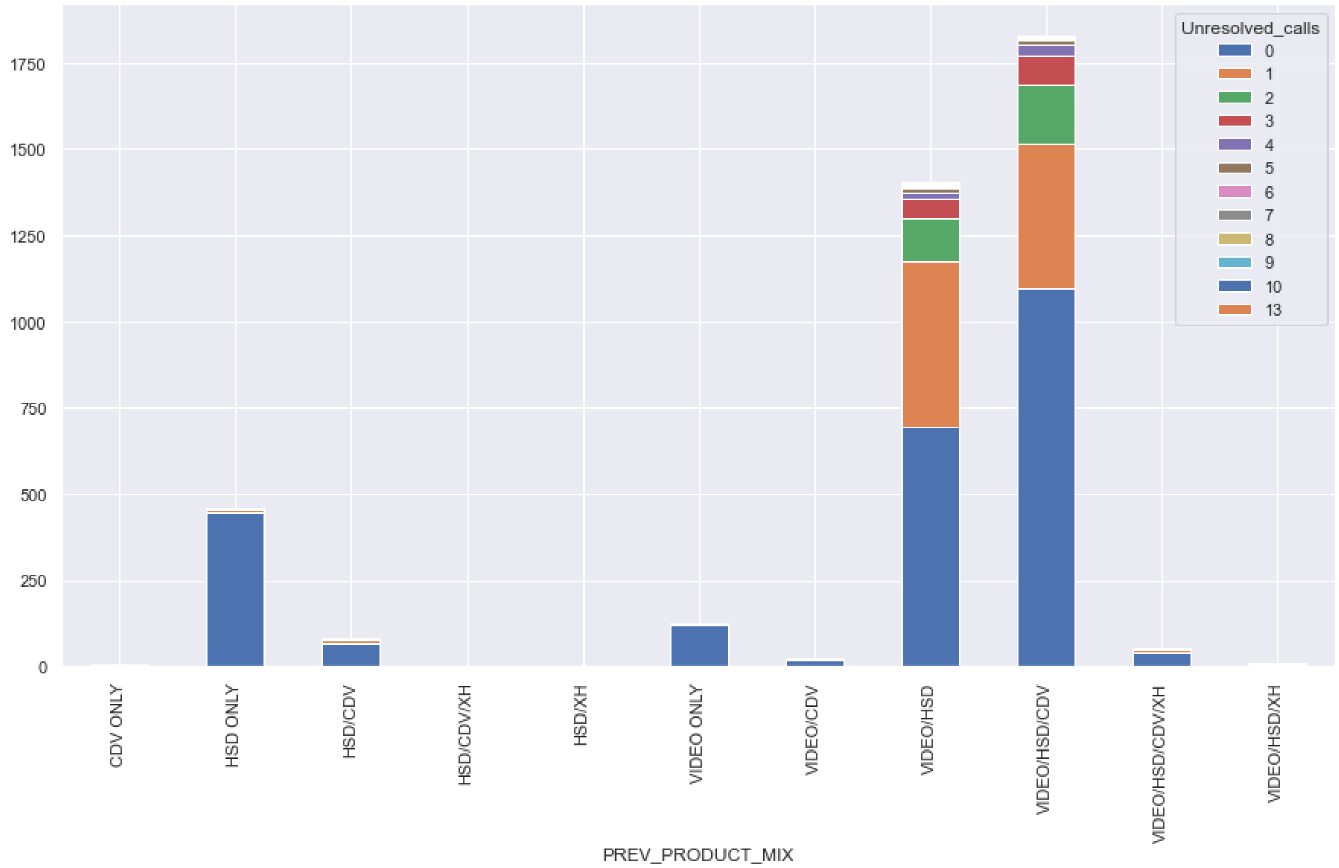


The bar graph shows that the 16 customers that disconnected their complete cable service provided a reason which indicated Cost was the main factor for them discontinuing their services. These customers were also under a contract however the contract term remaining was ‘0’ months, meaning they do not incur a penalty for cancelling services.

Validating Correlation between Service Downgrade and Unresolved issues/Outages

Out of remaining Active Customers = **3983**

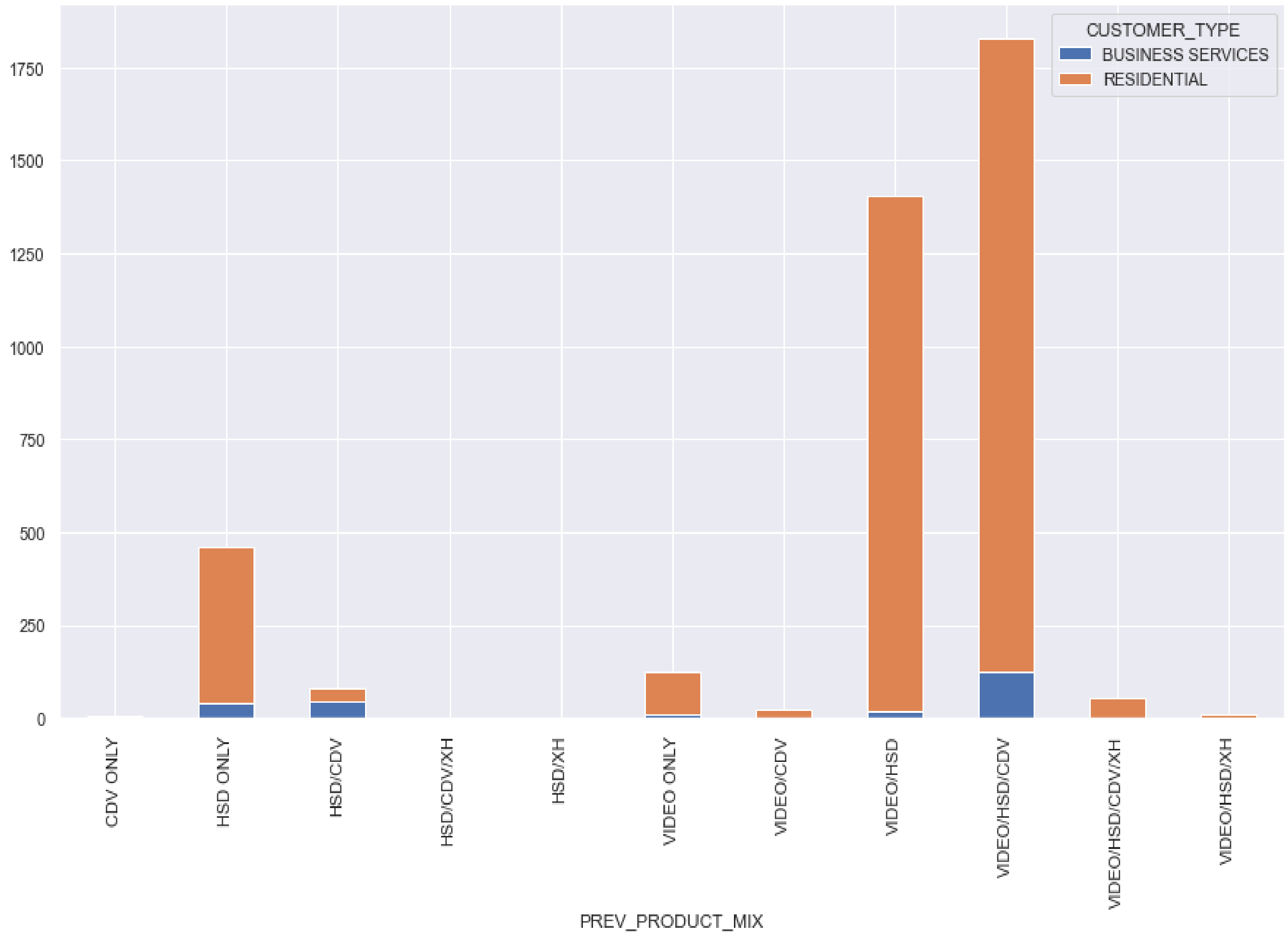
Number of customers who chose to downgrade services by disconnecting Video = **1979**



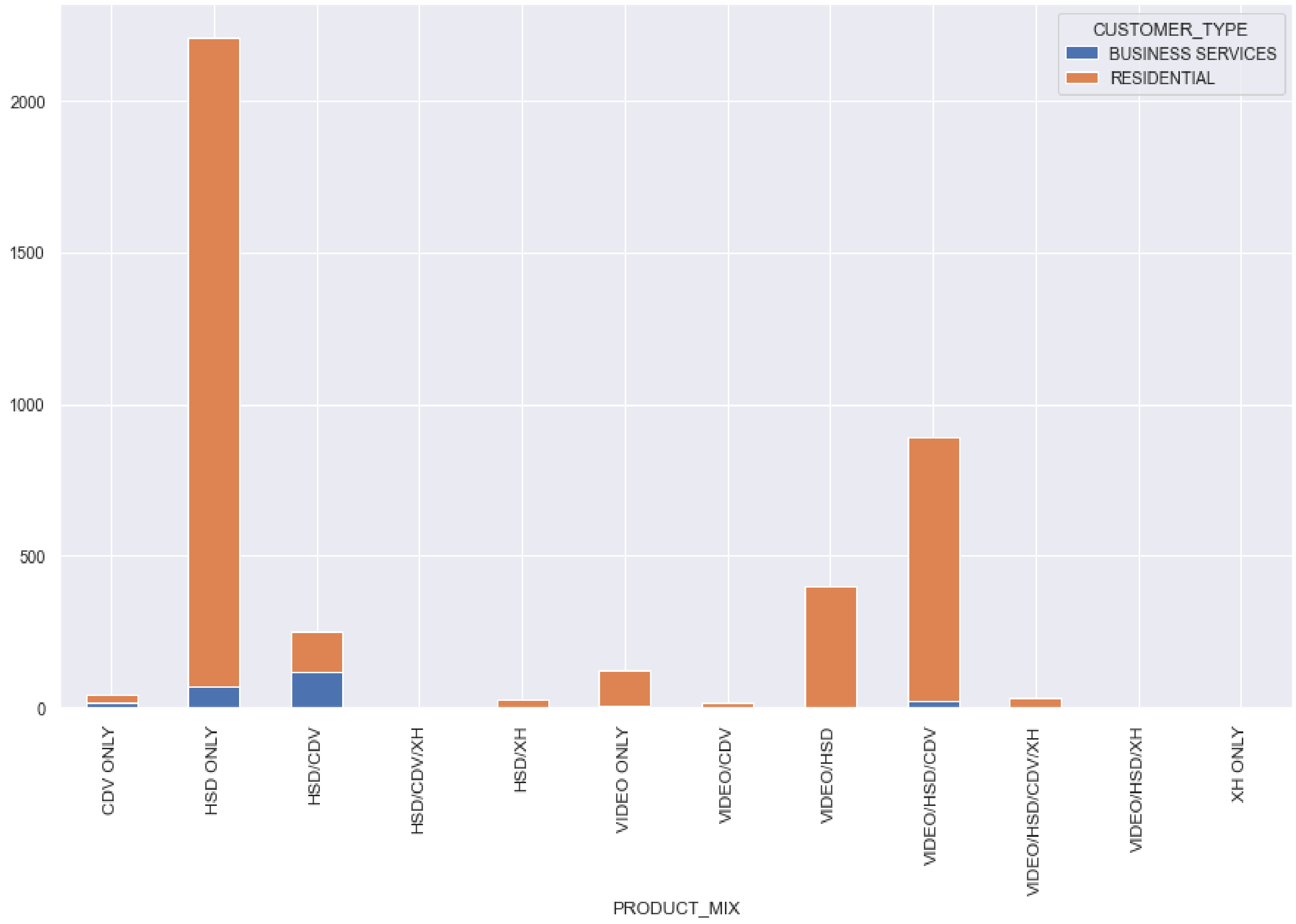
There are a substantial number of tickets/open issues which can be seen in the bar graph which Indicate an instability and bad quality of service delivery to the customers. We can see the maximum number of tickets opened and unresolved specifically with product bundles which include

VIDEO Service

The following Bar Chart shows the Product bundle mix by customer in the previous month of July 2021:

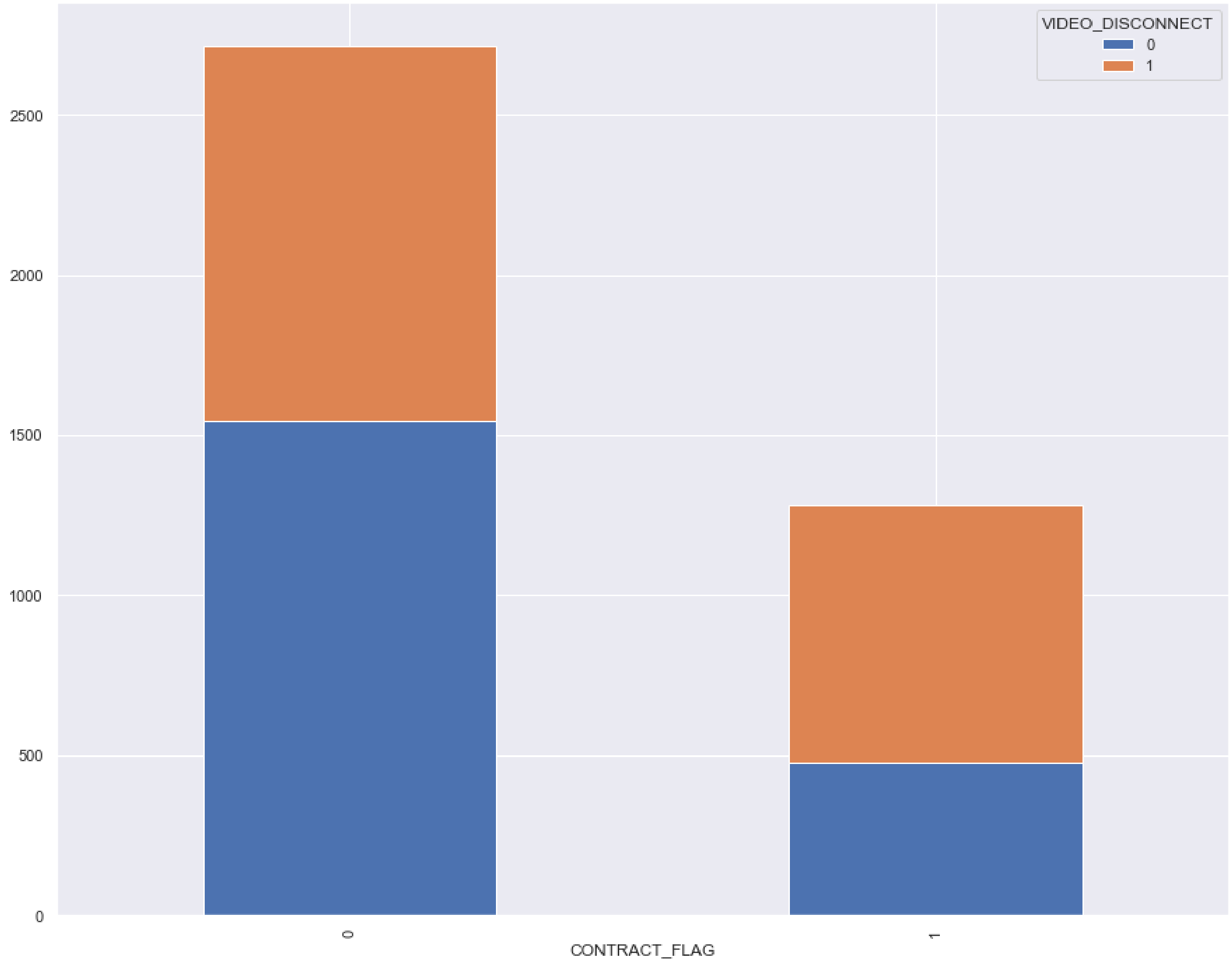


The following Bar Chart shows the product bundle mix after customer downgrade, for August 2021:



We can see that a large number of customers moved away from the Video/HSD, Video/HSD/CDV and Video/HSD/CDV/XH AND Video/HSD/XH bundles to HSD Only. This indicates that the customers chose to remain active subscribers but downgraded to a single product reducing costs and reducing probability of trouble-issues with multiple products.

Validating Correlation between Service Downgrade and Contracts



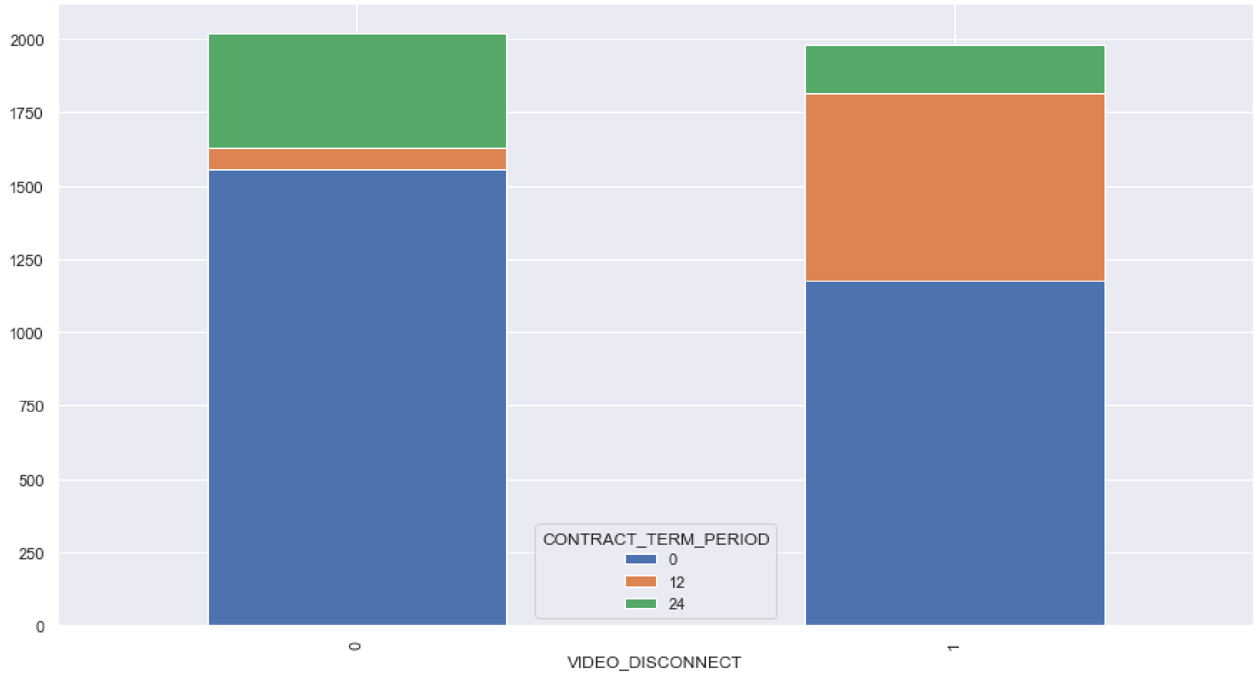
The data shows there are 1175 customers who downgraded their services by disconnecting Video only and they were not bound by a contract.

There were 806 customers who downgraded their services by disconnecting Video only and they were bound by a contract.

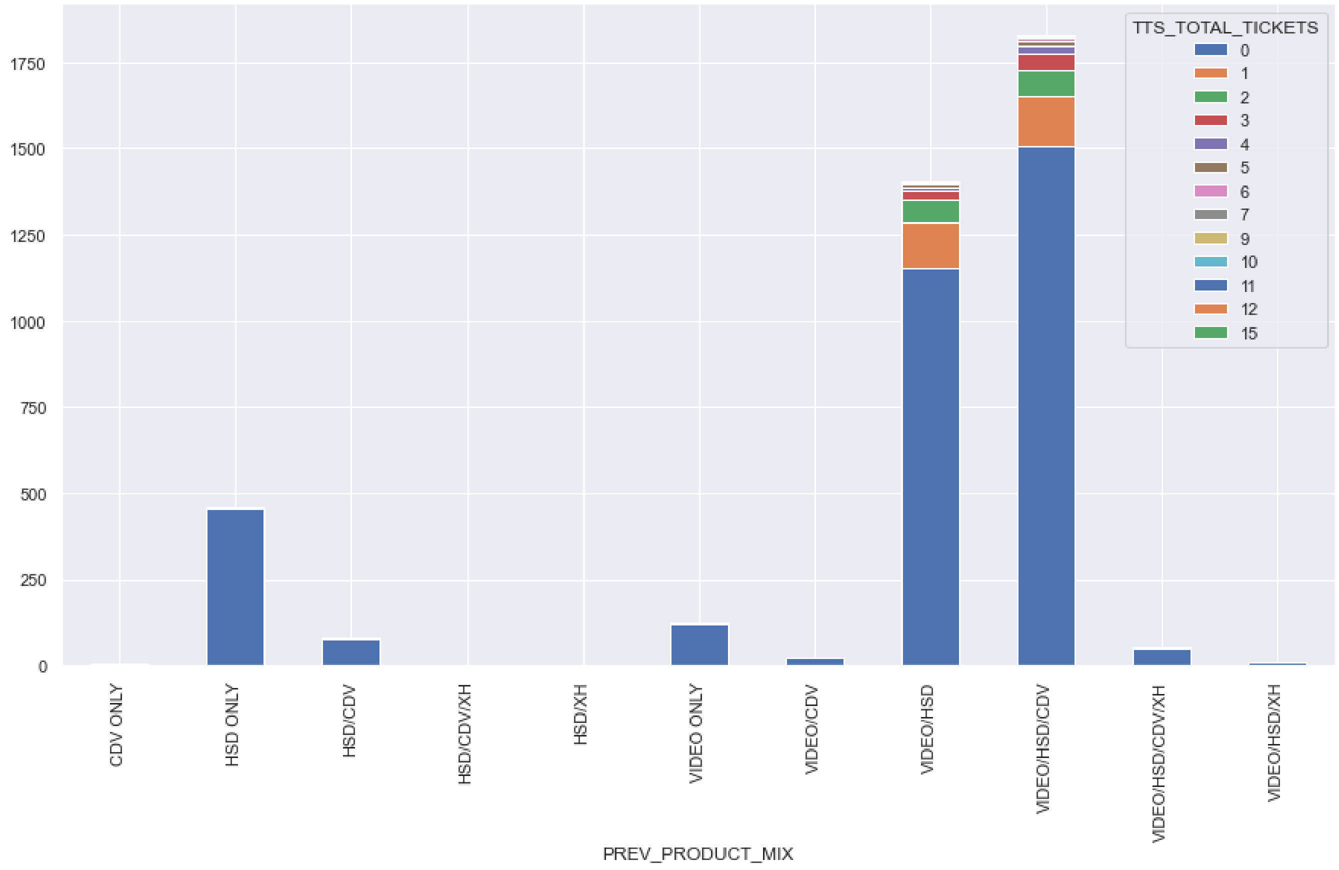
So, we can see that the 1175 customers who downgraded had no obligation to keep cable services and could have totally cancelled the same however they decided to retain their other products by simply removing video.

The 806 customers who downgraded and bound by a contract were also bound to a 12- or 24-month period contract. This does indicate that customers would rather finish the Contract term than pay penalty fees as they tend to downgrade service rather than outright cancel.

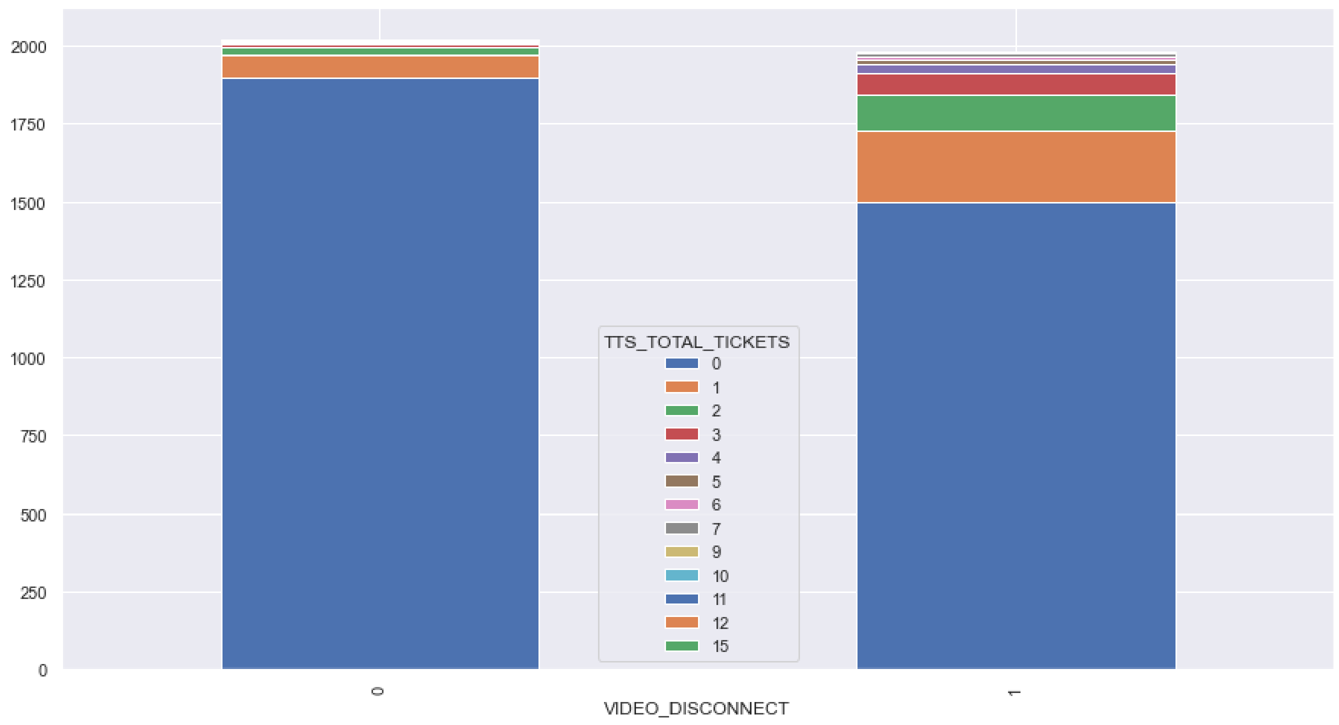
The fact that customers would tend to finish their contact terms is also indicated by the contract term period shown in months:



Validating correlation between service downgrade and TTS Tickets (Outages)

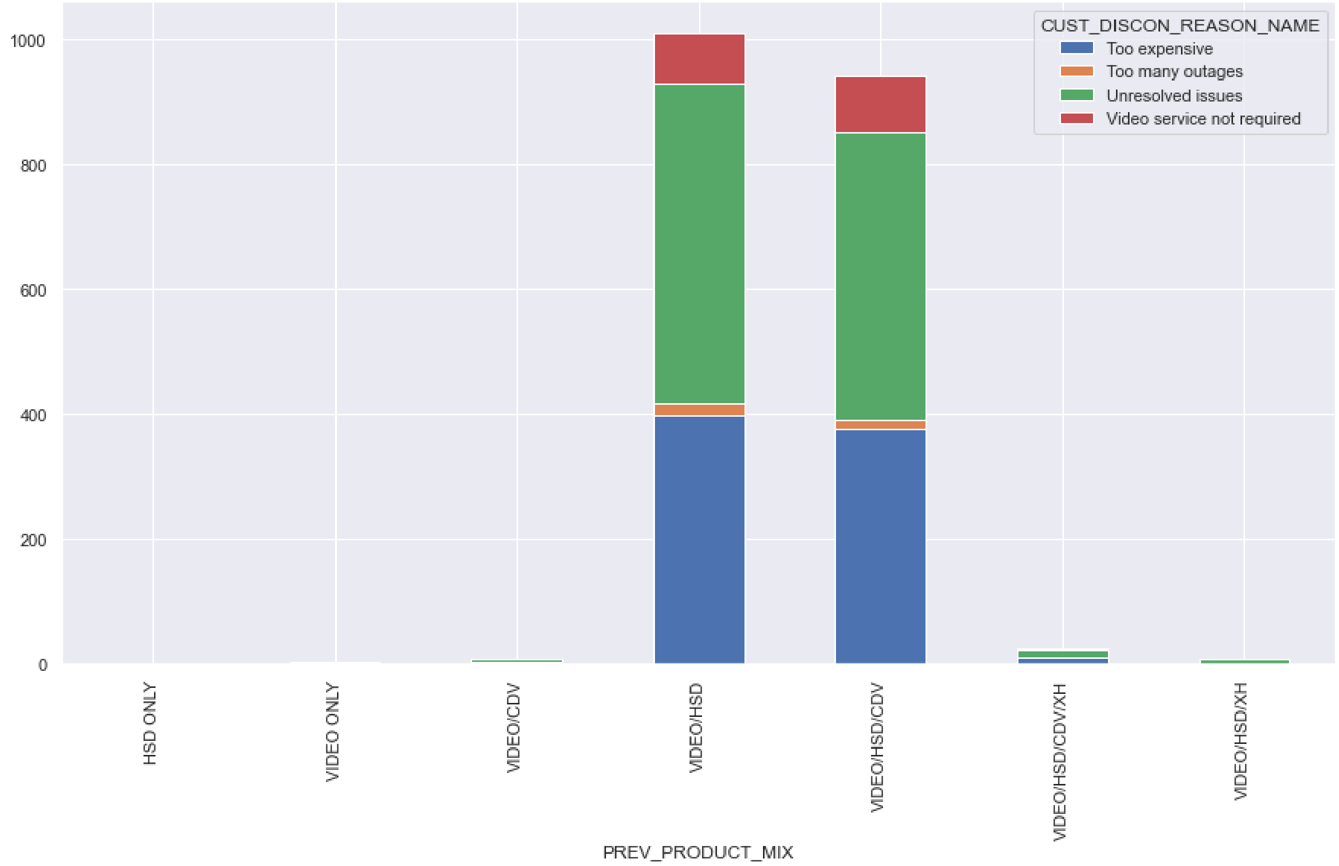


There were close to 600 Outage tickets which were opened and the Bar Graph shows the maximum number of outages took place with product bundles which had the Video service.

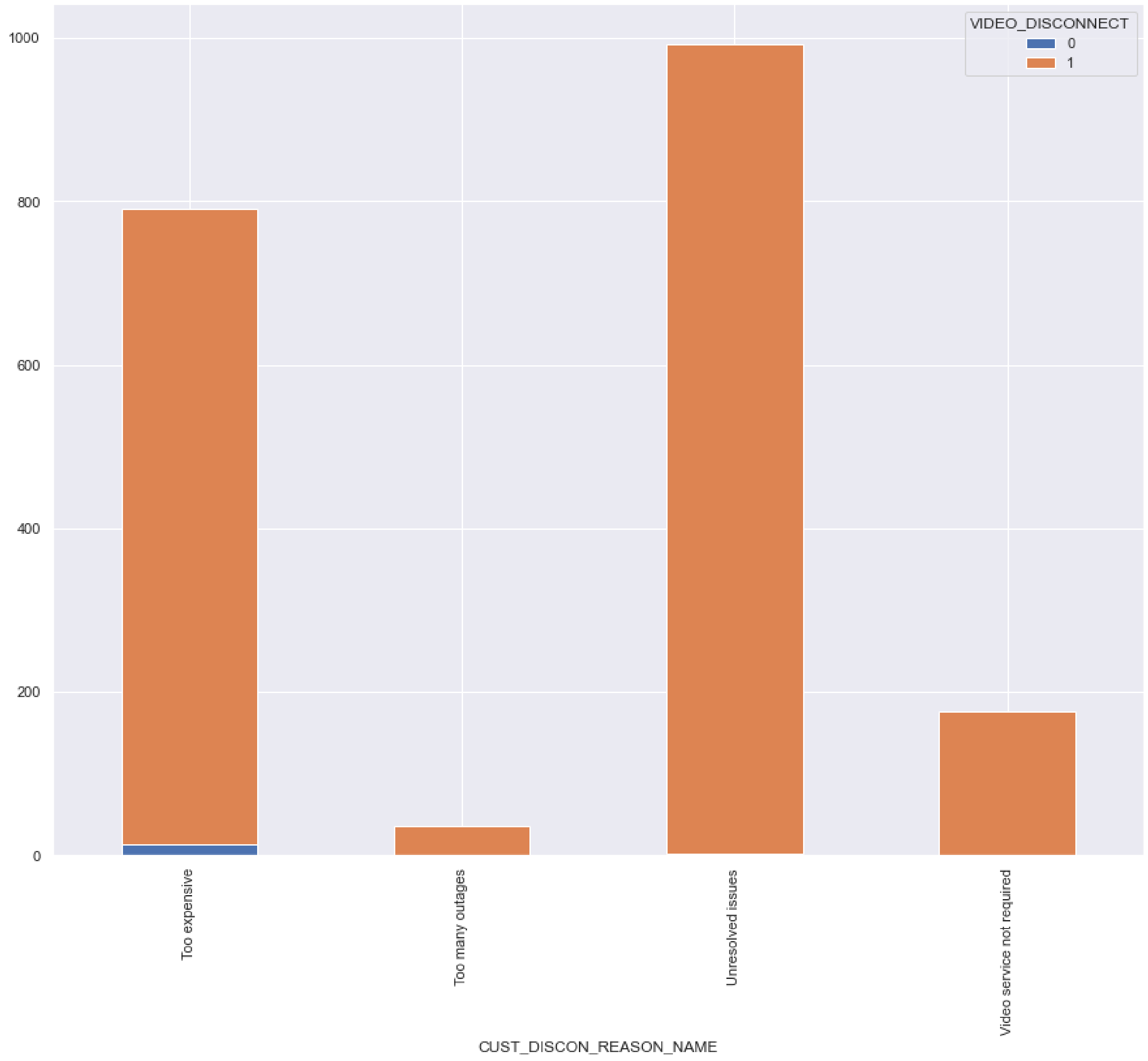


The Bar Chart above shows the customers who disconnected specifically Video services, indicted on X-Axis with ‘1’. There were 480 customers who downgraded services and faced outages.

Validating correlation between service downgrade and customer reason



The Bar Chart above shows the reason for customer disconnects based on product bundles. It is evident that the bundles which have the Video service included include the highest number of disconnection reasons. ‘Too expensive’ and ‘Unresolved Issues’ are almost evenly split in volume.

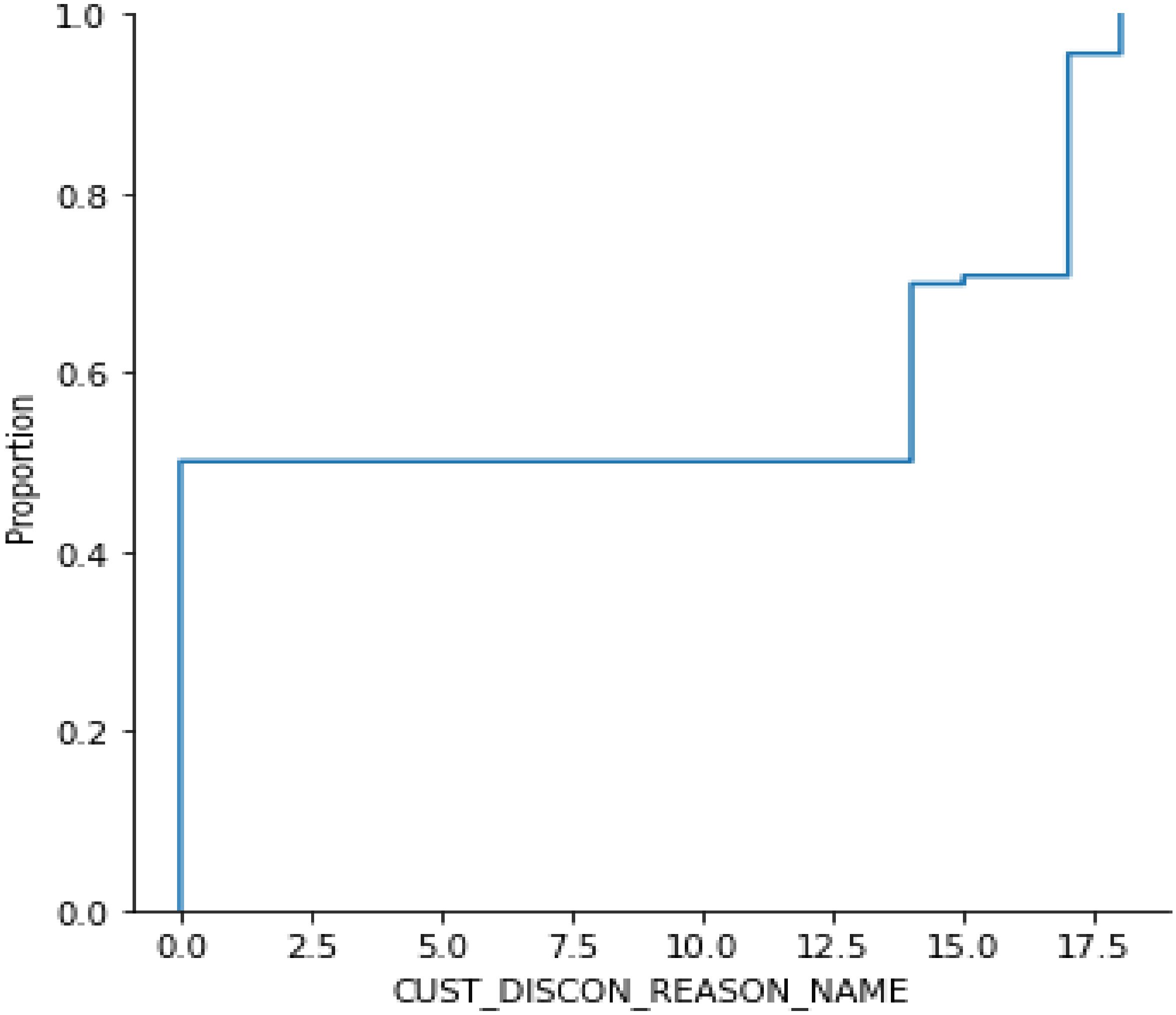


The data above shows customer who specifically disconnected Video services only. The reason again shows that the downgrade is related to costs and service issues being faced.

1979 Customers who downgraded services in August 2021 spent a total of $ 3,56287 in the previous month of July, 2021.

After downgrading services and removing Video, the same 1979 customers spent a total of $ 2,08350 in August 2021.

CDF – Cumulative Distribution Function



Customer disconnect reason has been chosen for the CDF. This variable was specifically chosen as it displays the specific customer reason which has been provided. Based on the data we have seen so far; the customer reasons point to prohibitive costs of service and quality of service delivery and outages. This variable provides direct link to independent variables which may be responsible for customer's disconnection/downgrade decision and those independent variables are: Monthly costs, Unresolved tickets, Outages and Contracts. This again leads to a customer's perceived value for a product or service they purchase. The customer reason string values were converted to numeric identifiers as indicated below:

Too expensive = 14
Too many outages = 15
NA = 0
Unresolved issues = 17
Video service not required = 18

Break-out by reason:

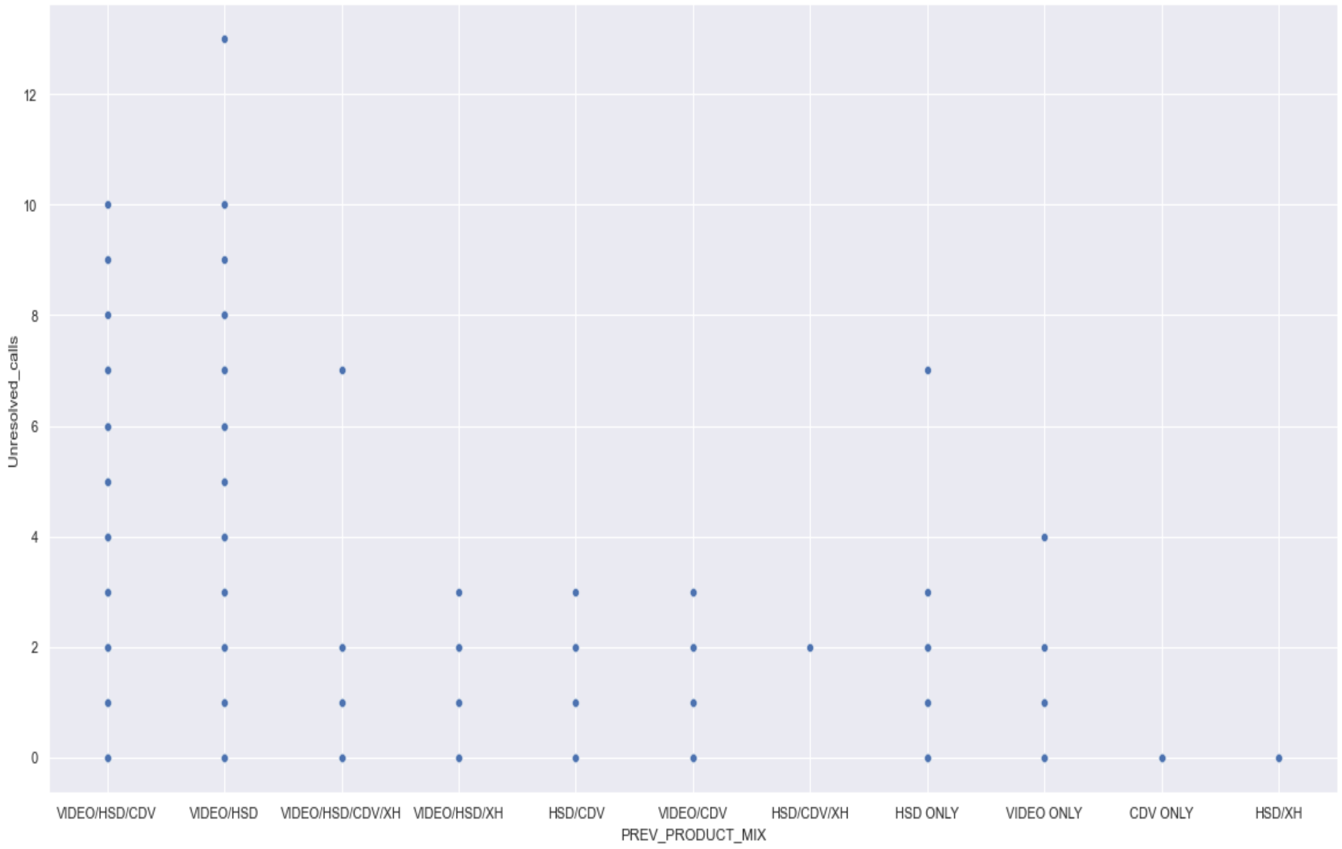
CUST_DISCON_REASON_NAME	Counts
Too expensive	798
Too many outages	38
Unresolved issues	992
Video service not required	176

Data indicates 50% of customers not having a disconnect reason; they are still active and have not downgraded or disconnected. Around 15% of customers stated Costs as being a factor for downgrades/disconnects and 20% stated Unresolved Issues and the remaining 25% state Video service not required or Outages.

The company can use this information to reduce costs of their services and also improve quality of delivery.

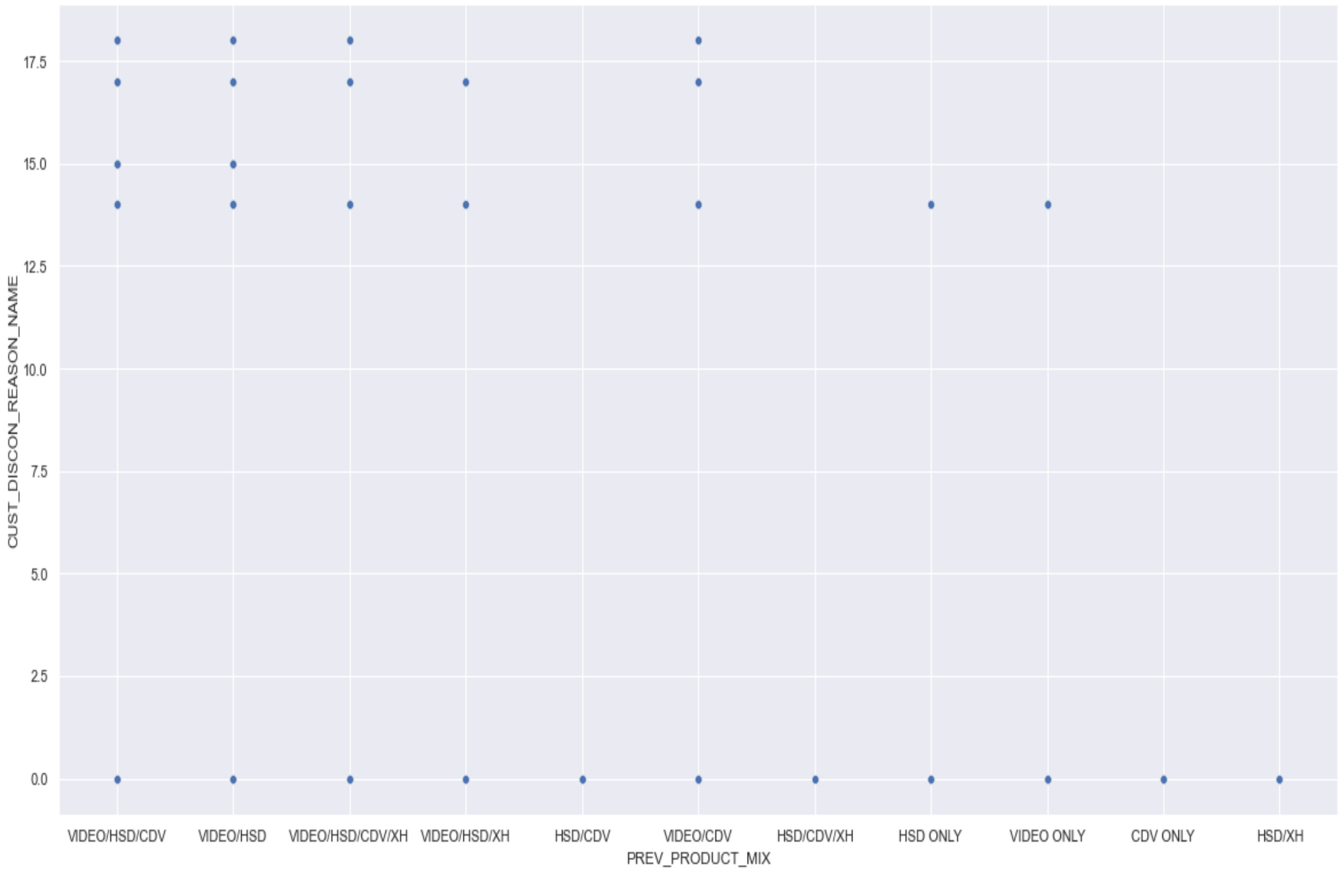
Scatter Plot – Analysis for Correlation

Previous Product Mix Vs. Unresolved Calls



The data shows bundles with Video services having a higher number of unresolved calls Vs. all other bundles or single products. This indicates instability of the Video service when offered as a bundled product. This data shows the product mix owned by the customer in the previous month and indicates a possible causation for downgrades.

Customer Disconnect Reason Vs. Previous Product Mix



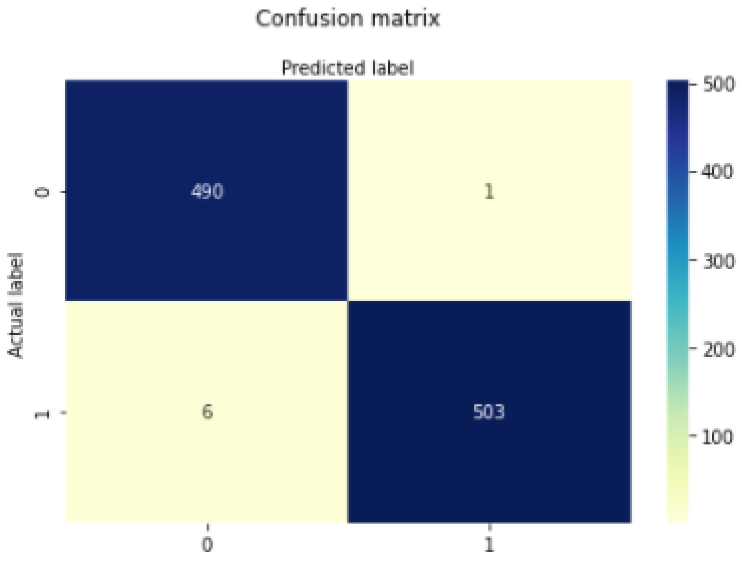
The data above further indicates one of the reasons for causation for downgrades or disconnects; customer dissatisfaction over cost of the service provided and the quality-of-service delivery. The cost shown is from the previous month, indicating a higher amount being paid by the customers before they downgraded. It can be clearly seen that for all bundled services, one of the reasons chosen is ‘Too expensive’, denoted by numeric indicator = 14

Customer Reason Name to Numeric mapping:

Too expensive = 14
Too many outages = 15
NA = 0
Unresolved issues = 17
Video service not required = 18

Logistic Regression

Running a Logistic Regression on the data with 75% data for model training and 25% for model testing, we get the following Confusion Matrix:



Target variable used: Service Downgrade (Srvc_Dwngrd)

The accuracy we are receiving with this model:

```
Accuracy: 0.893
Precision: 0.998015873015873
Recall: 0.9882121807465619
```

This again adds emphasis to the fact that the following independent variable we passed for the Regression analysis: ['Truckrolls', 'Unresolved_calls', 'TTS_TOTAL_TICKETS', 'CONTRACT_FLAG','VIDEO_DISCONNECT'], impact the customer decision to initiate a downgrade to their services, specifically removing the Video service

Summary and Conclusion

Constraints and Limitations Faced

1. The authenticity of customer Demographic data is questionable so any kind of modelling on the basis of demographics is not possible
2. The dataset used is a snapshot in time and I haven’t been able to get data for the following months to compare any model predictions
3. Data pertaining to any new customers who were on-boarded, is not available. This was required to see if new customers were joining, choosing only the Internet service.

- Total customer records analyzed/Population Dataset = 4000
- Number of Customers who disconnected all services = 16 = 0.4%
- Customers downgrading service by removing the VIDEO product = 2002 = 50.0%
  - Customers bound by a contract who downgraded service = 812 = 40.5%
  - Customers who downgraded and chose ‘Too Expensive’ as reason = 778 = 38.8%
  - Customers who downgraded and chose ‘Too Many Outages’ as reason = 36 = 1.7%
  - Customers who downgraded and chose ‘Unresolved Issues’ as reason = 992 = 49.5%
  - Customers who downgraded and chose ‘Video service not required’ as reason = 176 = 8.7%
- Customers who downgraded services (Cohort=2002)
  - Total charges/month after downgrade = $ 3,73,483.06
  - Total charges/month before downgrade = $ 5,21,797.33

The data provides strength to the theory of causation, as we can see based on the Correlation Heatmap and the Co-Variance data, that there is a high correlation which exists between customers who downgraded services to → customer disconnect reasons to → unresolved calls. The second correlation exists between customer disconnect reasons to → Monthly Recurring Charge from current Month

The company should increase the quality and stability of services, specially related to bundles products. The product bundles should also undergo a price reduction, specifically for the VIDEO product as that happens to be the primary service which is being cancelled.