1. Preparation

- Packages :
  Pandas as pd
  Numpy as np
  Matplotlib.pyplot as plt
  Seaborn as sns
  sns.set()
- Data : ‘Telco Customer Churn.csv’

2. Initial EDA

Data Profilling :

- There are 7043 rows of data with 21 columns
- The target variable is `Churn` and possible dropped column is `customerID`
- Data has float, integer, and object values.
- Basically, there is no missing values/null in this data. But there is a blank values `...` on TotalCharges. Because of the blank values, the dtype of `['TotalCharges']` considered as an object. We need to change the blank value to get further analysis
- The value of the target `['Churn']` must be encoding in order to get an analysis correlation of the column with the others.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

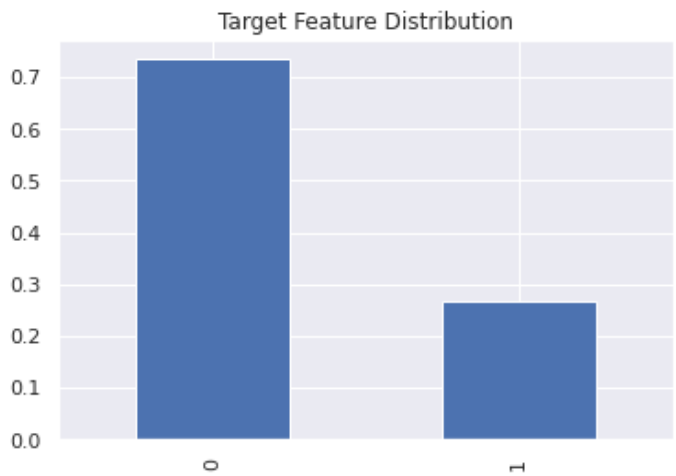
Action Plan :

In order to get full correlation between all columns, we need to encode all categorical with one-hot encode method. Or if the category had only two value, we could also use ordinal encoding. From the `.describe()` above we know that.

- `gender`, `Partner`, `Dependents`, `PhoneService`, `PaperlessBilling`, and the target `Churn` is two category data columns.
- `Tenure`, `MonthlyCharges`, and `TotalCharges` are continuous.
- The rest are three or more category data columns.
- Unique value happen on `['Contract']` and `['PaymentMethod']` , while other three category just had `['Yes' 'No' 'No phone service']` value.

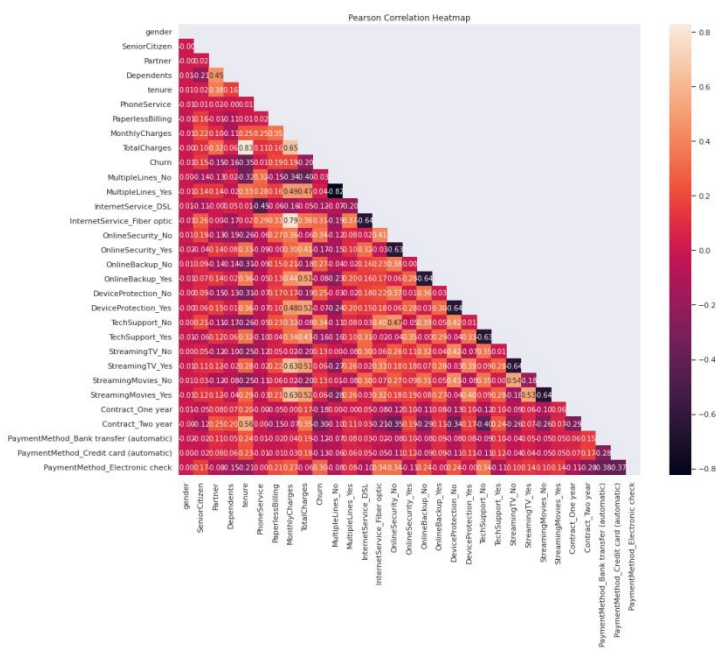
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 31 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   gender                                   7043 non-null   int64
 1   SeniorCitizen                            7043 non-null   int64
 2   Partner                                  7043 non-null   int64
 3   Dependents                               7043 non-null   int64
 4   tenure                                   7043 non-null   int64
 5   PhoneService                             7043 non-null   int64
 6   PaperlessBilling                         7043 non-null   int64
 7   MonthlyCharges                           7043 non-null   float64
 8   TotalCharges                             7043 non-null   float64
 9   Churn                                    7043 non-null   int64
 10  MultipleLines_No                         7043 non-null   int64
 11  MultipleLines_Yes                        7043 non-null   int64
 12  InternetService_DSL                      7043 non-null   int64
 13  InternetService_Fiber optic              7043 non-null   int64
 14  OnlineSecurity_No                        7043 non-null   int64
 15  OnlineSecurity_Yes                       7043 non-null   int64
 16  OnlineBackup_No                          7043 non-null   int64
 17  OnlineBackup_Yes                         7043 non-null   int64
 18  DeviceProtection_No                      7043 non-null   int64
 19  DeviceProtection_Yes                     7043 non-null   int64
 20  TechSupport_No                           7043 non-null   int64
 21  TechSupport_Yes                          7043 non-null   int64
 22  StreamingTV_No                           7043 non-null   int64
 23  StreamingTV_Yes                          7043 non-null   int64
 24  StreamingMovies_No                       7043 non-null   int64
 25  StreamingMovies_Yes                      7043 non-null   int64
 26  Contract_One year                        7043 non-null   int64
 27  Contract_Two year                        7043 non-null   int64
 28  PaymentMethod_Bank transfer (automatic)  7043 non-null   int64
 29  PaymentMethod_Credit card (automatic)    7043 non-null   int64
 30  PaymentMethod_Electronic check           7043 non-null   int64
dtypes: float64(2), int64(29)
memory usage: 1.7 MB
```

3. Target Distribution



73% : 27% distribution is still considered as balanced distribution

4. Correlation Analysis



- ‘TotalCharges` and `tenure`, also `InternetService_Fiber Optic` and `MonthlyCharges` are good candidates of predictors
- However, they have a very high correlation (0.83 and 0.79) which means we have to choose which one of them will be used as predictor

5. Total Charges Analysis

If we take a look back for a while, we must remember that `we replace few rows blank TotalCharge value with 0` . Now let's take a look into original data numerical feature.

	tenure	MonthlyCharges	TotalCharges
0	1	29.85	29.85
1	34	56.95	1889.5
2	2	53.85	108.15
3	45	42.30	1840.75
4	2	70.70	151.65

We could see a red line here, that `the TotalCharges value really close to value tenure * MonthlyCharges`. So, how about we take a look into specific data that had blank TotalCharges value.

	tenure	MonthlyCharges	TotalCharges
488	0	52.55	
753	0	20.25	
936	0	80.85	
1082	0	25.75	
1340	0	56.05	
3331	0	19.85	
3826	0	25.35	
4380	0	20.00	
5218	0	19.70	
6670	0	73.35	
6754	0	61.90	

Interesting! So the customer with blank TotalCharges have the tenure value 0. It means the customer with blank TotalCharges most likely `cancel their subscription to the company before their contract periods end`. It result their TotalCharges are undefined.

6. Family Analysis

```
[38] dfpartner
```

	Churn	
	count	mean
Partner		
0	3641	0.329580
1	3402	0.196649

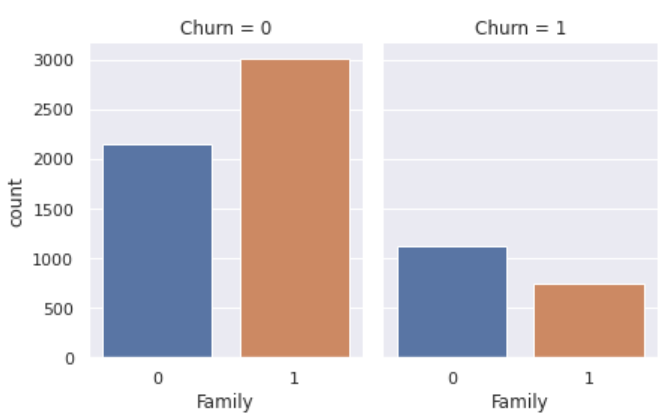
```
[39] dfdep
```

	Churn	
	count	mean
Dependents		
0	4933	0.312791
1	2110	0.154502

Take a look of data above. As we can see `Partner` and `Dependents` had similiarly distribution characteristic : `if customer had partner or dependents they had lower chance of churn`. Also, if we take a look at the columns description on kaggle, we know that :

- Partner = the customer that had partners, something like couple or married
- Dependents = the customer that had dependents

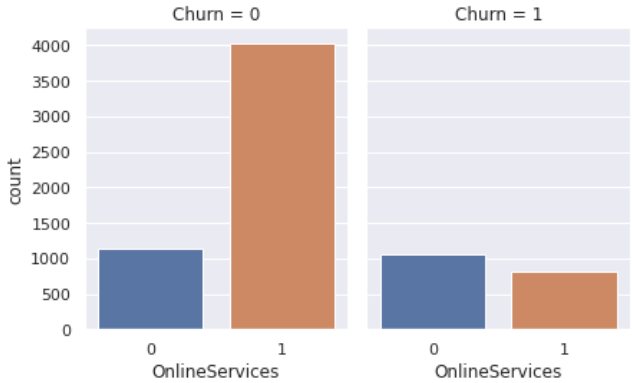
Now, how about we just merge this into one columns as Family columns?



From the data above, we know that customer with family `less likely to Churn`

7. Online Services Analysis

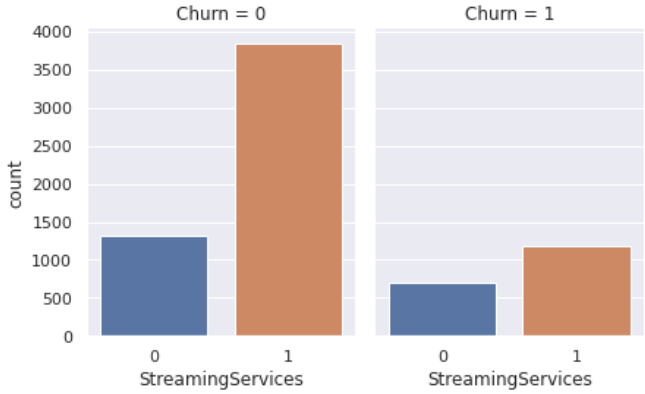
With the same method, how about we analyze Online Security and Online Backup columns?



From the data above, we know that customer with no online services (either Security or Backup) had `little higher chance to Churn`.

8. Streaming Services Analysis

One more column to merge & analyze. Streaming Service will consist of StreamingTV and StreamingMovies.



From the data above, we know that customer opted for StreamingServices (StreamingTV or StreamingMovies) had higher chance to Churn

9. Conclusion

From the analysis, i had few theories and recommendations :

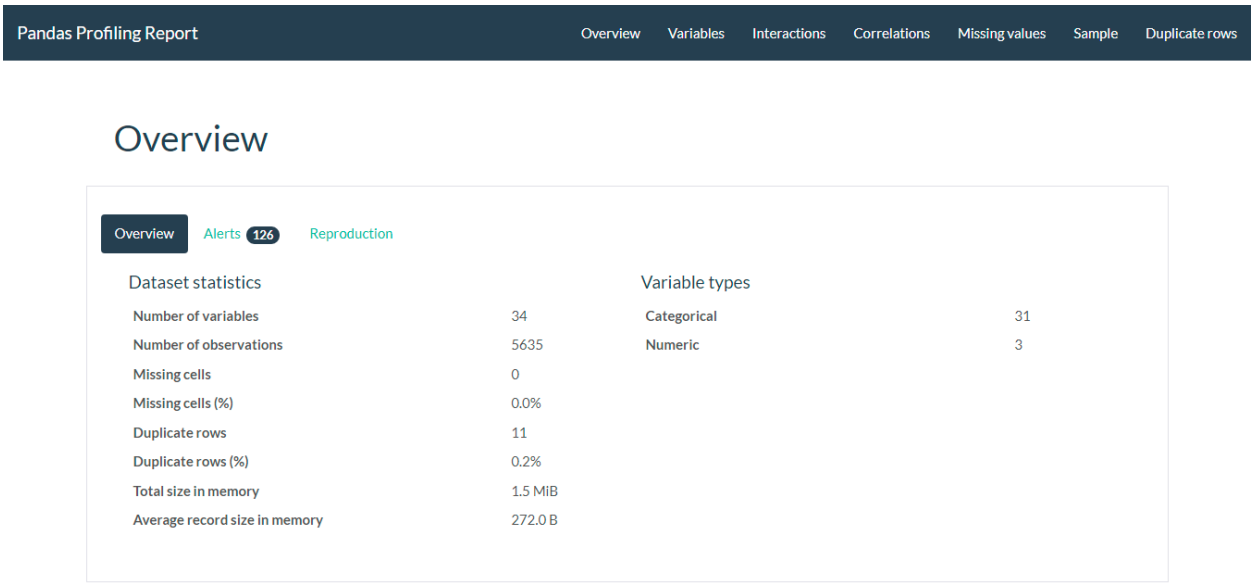
- The Telco company had good family contents because the family customers are slightly loyal to the company. But they need to engage the single customers, maybe company can give one-day vacation for loyal customer so they could enjoy quality time for theirself.
- The Telco company not yet seems securable, it cause the people didn't opted for OnlineService and more likely to Churn. The Telco company must added new feature to their OnlineService like online backup storage with encrypted message maybe help engage customer and decrese Churn rate.
- Lastly, the Telco company has little poor of streaming service. So the people with StreamingService more higher chance to Churn. To prevent that, the Telco company must improve their streaming service. How about premium movies for loyal customer with earlier release date than other company?

10. Google Colab Link

https://colab.research.google.com/drive/1K46b-Sc3N_PaEyXHWSjJqXDfWS5etY09?usp=sharing

Homework Day 16 – Andhyka Cakrabuana Adhitama

Data Profiler



Sweetviz

Homework Day 16 – Andhyka Cakrabuana Adhitama

Data Explorer