Assignment

Use the "from the expert" (FTE) jupyter notebook as a starter for this assignment, and ask your instructor questions if you need help.

Use the `churn_data.csv` file to carry out a similar EDA and visualization process as what we did in the FTE. Create at least 2 EDA plots, and create a HTML file with an auto-EDA analysis using pandas-profiling or another auto-EDA Python package. Write a short analysis at the end of the assignment in markdown.

Data science process steps this week

We will carry out the first two parts of the CRISP-DM data science process this week:

1. Business understanding

This is customer churn data for a telecommunications company. Customers can have phone as well as other services. The company is looking to reduce customer churn, where customers stop using the company's services and cancel their account. The 'Churn' column has a binary target, yes or no, that denotes if a customer churned. We want to create a machine learning model to predict the Churn target using the other available data in the dataset. Ideally, we will deploy this model to integrate with the company's database, so that a churn risk column is created for each customer. This will enable customer service reps and others to devise and use strategies to reduce churn.

2. Data understanding

Carry out some EDA as we did in the FTE, such as using pandas-profiling. Create a histogram like we did in the FTE, where we plot a numeric column with the target as the 'hue'. Optional challenge: create other plots with the target as the hue, such as bar plots for the categorical columns.

In [1]:
```
!conda install -c conda-forge pandas-profiling  openpyxl -y
```

```
Retrieving notices: ...working... done
Channels:
 - conda-forge
 - defaults
Platform: win-64
Collecting package metadata (repodata.json): ...working... done
Solving environment: ...working... done

# All requested packages already installed.
```

In [8]:
```
import warnings
warnings.filterwarnings("ignore")
```

In [4]:
```
!pip install ydata-profiling
```

```
Collecting ydata-profiling
  Downloading ydata_profiling-4.8.3-py2.py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: scipy<1.14,>=1.4.1 in c:\users\geflo\anaconda3\lib\si
te-packages (from ydata-profiling) (1.11.4)
Requirement already satisfied: pandas!=1.4.0,<3,>1.1 in c:\users\geflo\anaconda3\lib
\site-packages (from ydata-profiling) (2.2.2)
Requirement already satisfied: matplotlib<3.9,>=3.2 in c:\users\geflo\anaconda3\lib
\site-packages (from ydata-profiling) (3.8.4)
Collecting pydantic>=2 (from ydata-profiling)
  Downloading pydantic-2.7.1-py3-none-any.whl.metadata (107 kB)
     ---------------------------------------- 0.0/107.3 kB ? eta -:--:--
     ----------- -------------------------- 30.7/107.3 kB 1.4 MB/s eta 0:00:01
     ---------------------------------------- 107.3/107.3 kB 2.1 MB/s eta 0:00:00
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in c:\users\geflo\anaconda3\lib\si
te-packages (from ydata-profiling) (6.0.1)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in c:\users\geflo\anaconda3\lib\s
ite-packages (from ydata-profiling) (3.1.3)
Collecting visions<0.7.7,>=0.7.5 (from visions[type_image_path]<0.7.7,>=0.7.5->ydata
-profiling)
  Downloading visions-0.7.6-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: numpy<2,>=1.16.0 in c:\users\geflo\anaconda3\lib\site
-packages (from ydata-profiling) (1.26.4)
Requirement already satisfied: htmlmin==0.1.12 in c:\users\geflo\anaconda3\lib\site-
packages (from ydata-profiling) (0.1.12)
Requirement already satisfied: phik<0.13,>=0.11.1 in c:\users\geflo\anaconda3\lib\si
te-packages (from ydata-profiling) (0.12.3)
Requirement already satisfied: requests<3,>=2.24.0 in c:\users\geflo\anaconda3\lib\s
ite-packages (from ydata-profiling) (2.31.0)
Requirement already satisfied: tqdm<5,>=4.48.2 in c:\users\geflo\anaconda3\lib\site-
packages (from ydata-profiling) (4.65.0)
Requirement already satisfied: seaborn<0.14,>=0.10.1 in c:\users\geflo\anaconda3\lib
\site-packages (from ydata-profiling) (0.12.2)
Requirement already satisfied: multimethod<2,>=1.4 in c:\users\geflo\anaconda3\lib\s
ite-packages (from ydata-profiling) (1.4)
Requirement already satisfied: statsmodels<1,>=0.13.2 in c:\users\geflo\anaconda3\li
b\site-packages (from ydata-profiling) (0.14.0)
Collecting typeguard<5,>=3 (from ydata-profiling)
  Downloading typeguard-4.2.1-py3-none-any.whl.metadata (3.7 kB)
Requirement already satisfied: imagehash==4.3.1 in c:\users\geflo\anaconda3\lib\site
-packages (from ydata-profiling) (4.3.1)
Collecting wordcloud>=1.9.1 (from ydata-profiling)
  Downloading wordcloud-1.9.3-cp311-cp311-win_amd64.whl.metadata (3.5 kB)
Collecting dacite>=1.8 (from ydata-profiling)
  Downloading dacite-1.8.1-py3-none-any.whl.metadata (15 kB)
Requirement already satisfied: numba<1,>=0.56.0 in c:\users\geflo\anaconda3\lib\site
-packages (from ydata-profiling) (0.59.0)
Requirement already satisfied: pillow in c:\users\geflo\anaconda3\lib\site-packages
(from imagehash==4.3.1->ydata-profiling) (10.2.0)
Requirement already satisfied: PyWavelets in c:\users\geflo\anaconda3\lib\site-packa
ges (from imagehash==4.3.1->ydata-profiling) (1.5.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\geflo\anaconda3\lib\site-
packages (from jinja2<3.2,>=2.11.1->ydata-profiling) (2.1.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\geflo\anaconda3\lib\site
-packages (from matplotlib<3.9,>=3.2->ydata-profiling) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\geflo\anaconda3\lib\site-pac
kages (from matplotlib<3.9,>=3.2->ydata-profiling) (0.11.0)
```

```
Requirement already satisfied: fonttools>=4.22.0 in c:\users\geflo\anaconda3\lib\sit
e-packages (from matplotlib<3.9,>=3.2->ydata-profiling) (4.25.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\geflo\anaconda3\lib\sit
e-packages (from matplotlib<3.9,>=3.2->ydata-profiling) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\geflo\anaconda3\lib\site-
packages (from matplotlib<3.9,>=3.2->ydata-profiling) (23.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\geflo\anaconda3\lib\site
-packages (from matplotlib<3.9,>=3.2->ydata-profiling) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\geflo\anaconda3\lib
\site-packages (from matplotlib<3.9,>=3.2->ydata-profiling) (2.8.2)
Requirement already satisfied: llvmlite<0.43,>=0.42.0dev0 in c:\users\geflo\anaconda
3\lib\site-packages (from numba<1,>=0.56.0->ydata-profiling) (0.42.0)
Requirement already satisfied: pytz>=2020.1 in c:\users\geflo\anaconda3\lib\site-pac
kages (from pandas!=1.4.0,<3,>1.1->ydata-profiling) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\geflo\anaconda3\lib\site-p
ackages (from pandas!=1.4.0,<3,>1.1->ydata-profiling) (2023.3)
Requirement already satisfied: joblib>=0.14.1 in c:\users\geflo\anaconda3\lib\site-p
ackages (from phik<0.13,>=0.11.1->ydata-profiling) (1.1.0)
Collecting annotated-types>=0.4.0 (from pydantic>=2->ydata-profiling)
  Downloading annotated_types-0.6.0-py3-none-any.whl.metadata (12 kB)
Collecting pydantic-core==2.18.2 (from pydantic>=2->ydata-profiling)
  Downloading pydantic_core-2.18.2-cp311-none-win_amd64.whl.metadata (6.7 kB)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\geflo\anaconda3
\lib\site-packages (from pydantic>=2->ydata-profiling) (4.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\geflo\anaconda3
\lib\site-packages (from requests<3,>=2.24.0->ydata-profiling) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\geflo\anaconda3\lib\site-pac
kages (from requests<3,>=2.24.0->ydata-profiling) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\geflo\anaconda3\lib\si
te-packages (from requests<3,>=2.24.0->ydata-profiling) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\geflo\anaconda3\lib\si
te-packages (from requests<3,>=2.24.0->ydata-profiling) (2024.2.2)
Requirement already satisfied: patsy>=0.5.2 in c:\users\geflo\anaconda3\lib\site-pac
kages (from statsmodels<1,>=0.13.2->ydata-profiling) (0.5.3)
Requirement already satisfied: colorama in c:\users\geflo\anaconda3\lib\site-package
s (from tqdm<5,>=4.48.2->ydata-profiling) (0.4.6)
Collecting typing-extensions>=4.6.1 (from pydantic>=2->ydata-profiling)
  Downloading typing_extensions-4.11.0-py3-none-any.whl.metadata (3.0 kB)
Requirement already satisfied: attrs>=19.3.0 in c:\users\geflo\anaconda3\lib\site-pa
ckages (from visions<0.7.7,>=0.7.5->visions[type_image_path]<0.7.7,>=0.7.5->ydata-pr
ofiling) (23.1.0)
Requirement already satisfied: networkx>=2.4 in c:\users\geflo\anaconda3\lib\site-pa
ckages (from visions<0.7.7,>=0.7.5->visions[type_image_path]<0.7.7,>=0.7.5->ydata-pr
ofiling) (3.1)
Requirement already satisfied: six in c:\users\geflo\anaconda3\lib\site-packages (fr
om patsy>=0.5.2->statsmodels<1,>=0.13.2->ydata-profiling) (1.16.0)
Downloading ydata_profiling-4.8.3-py2.py3-none-any.whl (359 kB)
   ---------------------------------------- 0.0/359.5 kB ? eta -:--:--
   ---------------------------- ----------- 256.0/359.5 kB 7.9 MB/s eta 0:00:01
   ---------------------------------------- 359.5/359.5 kB 5.5 MB/s eta 0:00:00
Downloading dacite-1.8.1-py3-none-any.whl (14 kB)
Downloading pydantic-2.7.1-py3-none-any.whl (409 kB)
   ---------------------------------------- 0.0/409.3 kB ? eta -:--:--
   -------------------------------- ------ 337.9/409.3 kB 10.6 MB/s eta 0:00:01
   ---------------------------------------- 409.3/409.3 kB 8.5 MB/s eta 0:00:00
Downloading pydantic_core-2.18.2-cp311-none-win_amd64.whl (1.9 MB)
```

```
   ---------------------------------------- 0.0/1.9 MB ? eta -:--:--
   ----- ---------------------------------- 0.3/1.9 MB 16.6 MB/s eta 0:00:01
   ---------- ----------------------------- 0.6/1.9 MB 7.4 MB/s eta 0:00:01
   ------------------- -------------------- 1.0/1.9 MB 7.9 MB/s eta 0:00:01
   ---------------------------- ---------- 1.4/1.9 MB 8.1 MB/s eta 0:00:01
   ----------------------------------- --- 1.8/1.9 MB 7.9 MB/s eta 0:00:01
   ---------------------------------------- 1.9/1.9 MB 7.2 MB/s eta 0:00:00
Downloading typeguard-4.2.1-py3-none-any.whl (34 kB)
Downloading visions-0.7.6-py3-none-any.whl (104 kB)
   ---------------------------------------- 0.0/104.8 kB ? eta -:--:--
   ---------------------------------------- 104.8/104.8 kB ? eta 0:00:00
Downloading wordcloud-1.9.3-cp311-cp311-win_amd64.whl (300 kB)
   ---------------------------------------- 0.0/300.2 kB ? eta -:--:--
   ---------------------------------------- 300.2/300.2 kB 19.3 MB/s eta 0:00:00
Downloading annotated_types-0.6.0-py3-none-any.whl (12 kB)
Downloading typing_extensions-4.11.0-py3-none-any.whl (34 kB)
Installing collected packages: typing-extensions, dacite, annotated-types, typeguar
d, pydantic-core, wordcloud, visions, pydantic, ydata-profiling
  Attempting uninstall: typing-extensions
    Found existing installation: typing_extensions 4.9.0
    Uninstalling typing_extensions-4.9.0:
      Successfully uninstalled typing_extensions-4.9.0
  Attempting uninstall: visions
    Found existing installation: visions 0.7.4
    Uninstalling visions-0.7.4:
      Successfully uninstalled visions-0.7.4
  Attempting uninstall: pydantic
    Found existing installation: pydantic 1.10.12
    Uninstalling pydantic-1.10.12:
      Successfully uninstalled pydantic-1.10.12
Successfully installed annotated-types-0.6.0 dacite-1.8.1 pydantic-2.7.1 pydantic-co
re-2.18.2 typeguard-4.2.1 typing-extensions-4.11.0 visions-0.7.6 wordcloud-1.9.3 yda
ta-profiling-4.8.3
ERROR: pip's dependency resolver does not currently take into account all the packag
es that are installed. This behaviour is the source of the following dependency conf
licts.
pandas-profiling 3.2.0 requires visions[type_image_path]==0.7.4, but you have vision
s 0.7.6 which is incompatible.
```

In [9]:
```
import pandas as pd
#from pandas_profiling import ProfileReport
from ydata_profiling import ProfileReport
import matplotlib.pyplot as plt

%matplotlib inline
```

In [10]:
```
# we can give an index number or name for our index column, or leave it blank
df = pd.read_csv('churn_data.csv', index_col='customerID')
df
```

Out[10]:

customerID	tenure	PhoneService	Contract	PaymentMethod	MonthlyCharges	TotalCharg
7590-VHVEG	1	No	Month-to-month	Electronic check	29.85	29.
5575-GNVDE	34	Yes	One year	Mailed check	56.95	1889.
3668-QPYBK	2	Yes	Month-to-month	Mailed check	53.85	108.
7795-CFOCW	45	No	One year	Bank transfer (automatic)	42.30	1840.
9237-HQITU	2	Yes	Month-to-month	Electronic check	70.70	151.
...	...	...	...	...	...	
6840-RESVB	24	Yes	One year	Mailed check	84.80	1990.
2234-XADUH	72	Yes	One year	Credit card (automatic)	103.20	7362.
4801-JZAZL	11	No	Month-to-month	Electronic check	29.60	346.
8361-LTMKD	4	Yes	Month-to-month	Mailed check	74.40	306.
3186-AJIEK	66	Yes	Two year	Bank transfer (automatic)	105.65	6844.

7043 rows × 7 columns

In [11]:
```
df.head()
```

Out[11]:

customerID	tenure	PhoneService	Contract	PaymentMethod	MonthlyCharges	TotalCharg
7590-VHVEG	1	No	Month-to-month	Electronic check	29.85	29.
5575-GNVDE	34	Yes	One year	Mailed check	56.95	1889.
3668-QPYBK	2	Yes	Month-to-month	Mailed check	53.85	108.
7795-CFOCW	45	No	One year	Bank transfer (automatic)	42.30	1840.
9237-HQITU	2	Yes	Month-to-month	Electronic check	70.70	151.

In [12]:
```
df.tail()
```

Out[12]:

customerID	tenure	PhoneService	Contract	PaymentMethod	MonthlyCharges	TotalCharg
6840-RESVB	24	Yes	One year	Mailed check	84.80	1990.
2234-XADUH	72	Yes	One year	Credit card (automatic)	103.20	7362.
4801-JZAZL	11	No	Month-to-month	Electronic check	29.60	346.
8361-LTMKD	4	Yes	Month-to-month	Mailed check	74.40	306.
3186-AJIEK	66	Yes	Two year	Bank transfer (automatic)	105.65	6844.

In [13]:
```
# use the argument minimal=True to speed this up, although you won't get all the pl
report = ProfileReport(df)
report.to_file('churn_eda.html')
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

In [14]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 7043 entries, 7590-VHVEG to 3186-AJIEK
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   tenure          7043 non-null   int64  
 1   PhoneService    7043 non-null   object 
 2   Contract        7043 non-null   object 
 3   PaymentMethod   7043 non-null   object 
 4   MonthlyCharges  7043 non-null   float64
 5   TotalCharges    7032 non-null   float64
 6   Churn           7043 non-null   object 
dtypes: float64(2), int64(1), object(4)
memory usage: 698.2+ KB
```

In [15]:
```
df.describe(include='all')
```

Out[15]:

	tenure	PhoneService	Contract	PaymentMethod	MonthlyCharges	TotalCharg
count	7043.000000	7043	7043	7043	7043.000000	7032.000(
unique	NaN	2	3	4	NaN	N
top	NaN	Yes	Month-to-month	Electronic check	NaN	N
freq	NaN	6361	3875	2365	NaN	N
mean	32.371149	NaN	NaN	NaN	64.761692	2283.300
std	24.559481	NaN	NaN	NaN	30.090047	2266.771
min	0.000000	NaN	NaN	NaN	18.250000	18.800(
25%	9.000000	NaN	NaN	NaN	35.500000	401.450(
50%	29.000000	NaN	NaN	NaN	70.350000	1397.475(
75%	55.000000	NaN	NaN	NaN	89.850000	3794.737
max	72.000000	NaN	NaN	NaN	118.750000	8684.800(

In [16]:
```
df.tenure.median()
```

Out[16]:
```
29.0
```

In [17]:
```
col = df.columns
col
```

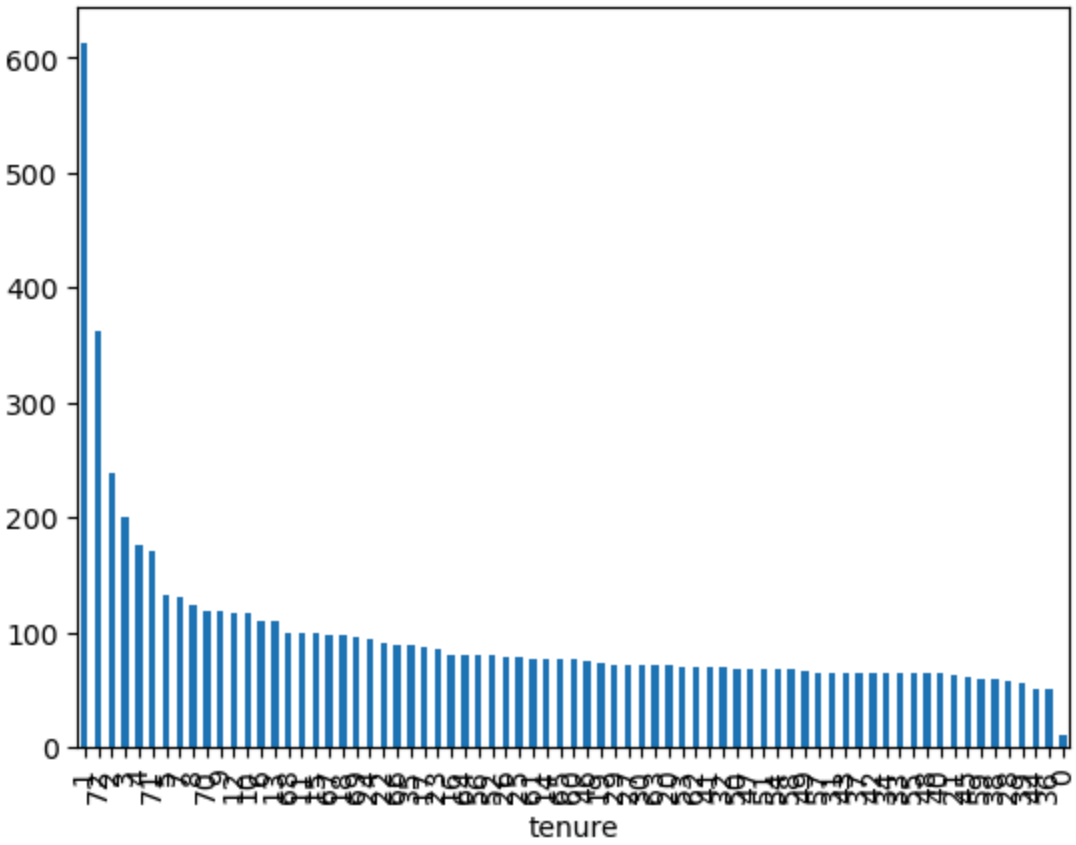
Out[17]:
```
Index(['tenure', 'PhoneService', 'Contract', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'],
      dtype='object')
```

In [18]:
```
df.MonthlyCharges
```

Out[18]:
```
customerID
7590-VHVEG     29.85
5575-GNVDE     56.95
3668-QPYBK     53.85
7795-CFOCW     42.30
9237-HQITU     70.70
               ...
6840-RESVB     84.80
2234-XADUH    103.20
4801-JZAZL     29.60
8361-LTMKD     74.40
3186-AJIEK    105.65
Name: MonthlyCharges, Length: 7043, dtype: float64
```

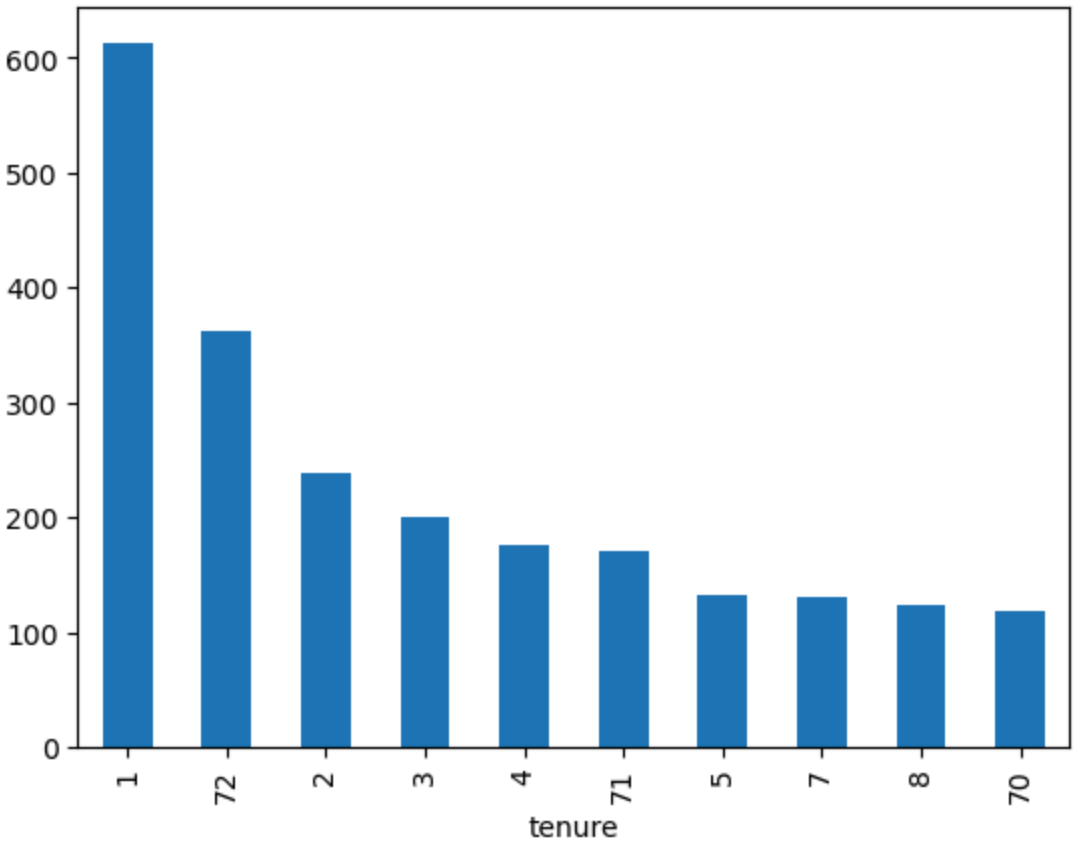
In [19]:
```
df['tenure'].value_counts().plot.bar()
```

Out[19]:
```
<Axes: xlabel='tenure'>
```



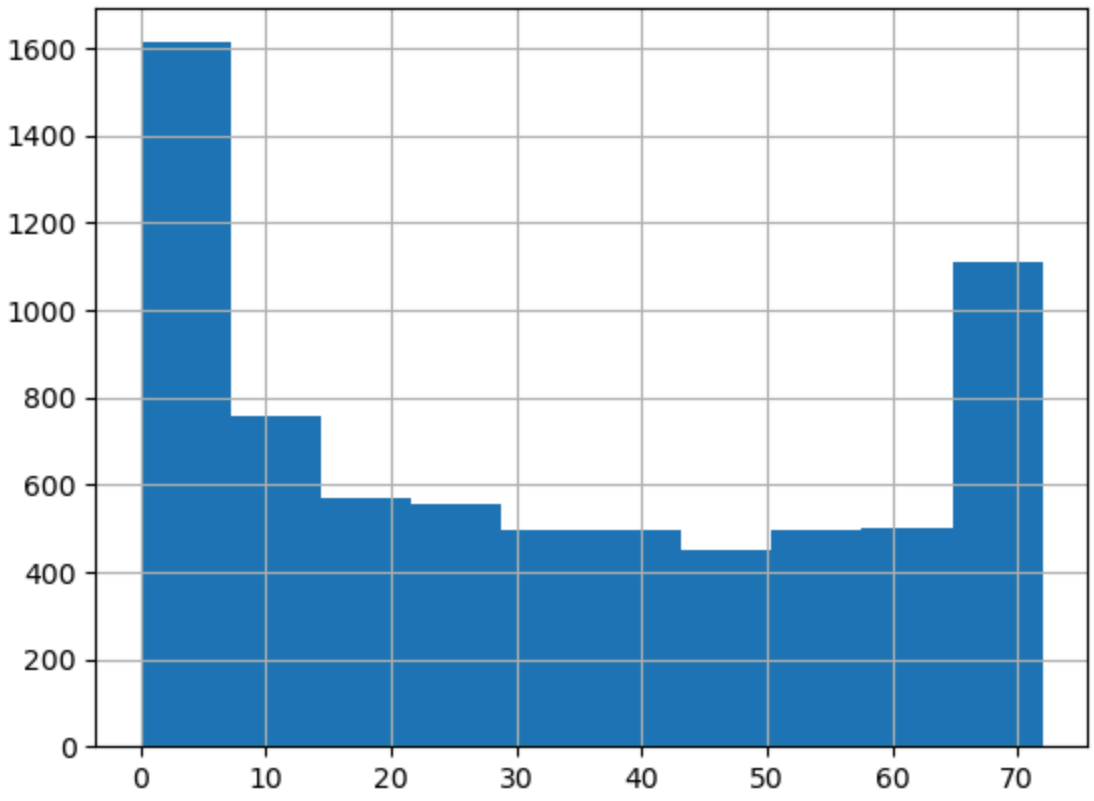
In [20]:
```
df['tenure'].value_counts()[:10].plot.bar()
```

Out[20]:
```
<Axes: xlabel='tenure'>
```



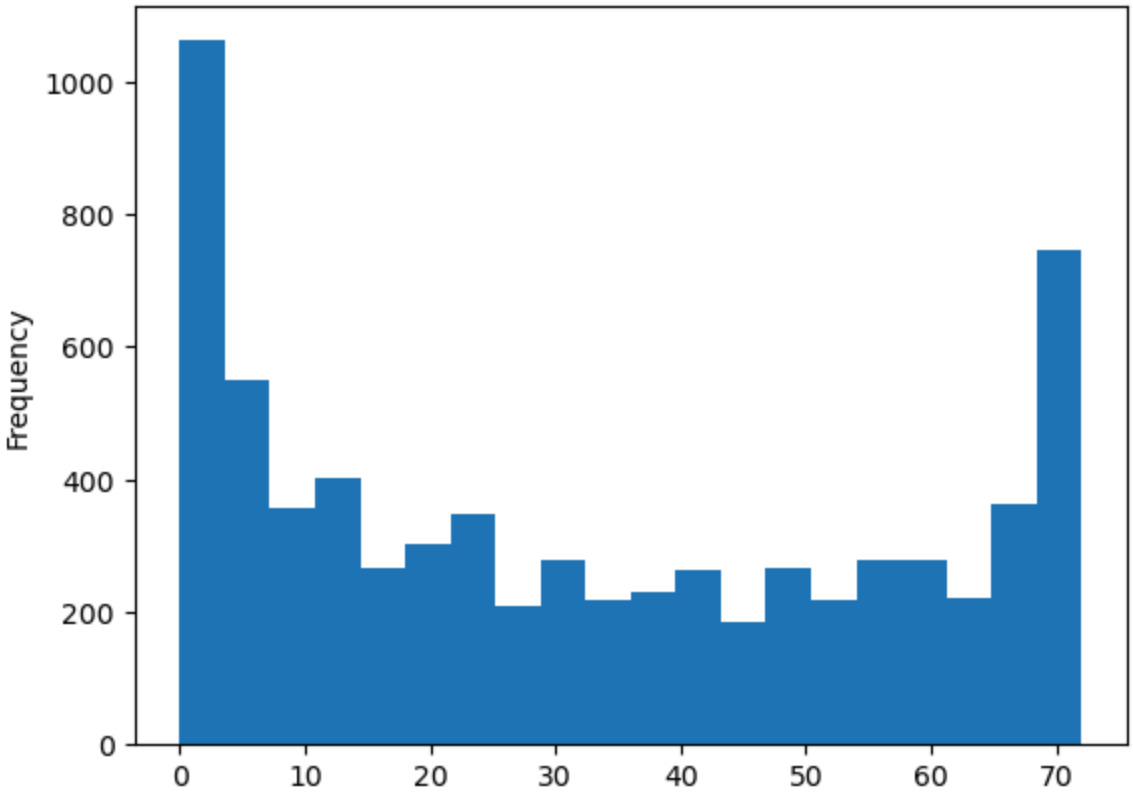
In [21]:
```
df['tenure'].hist()
```

Out[21]: <Axes: >



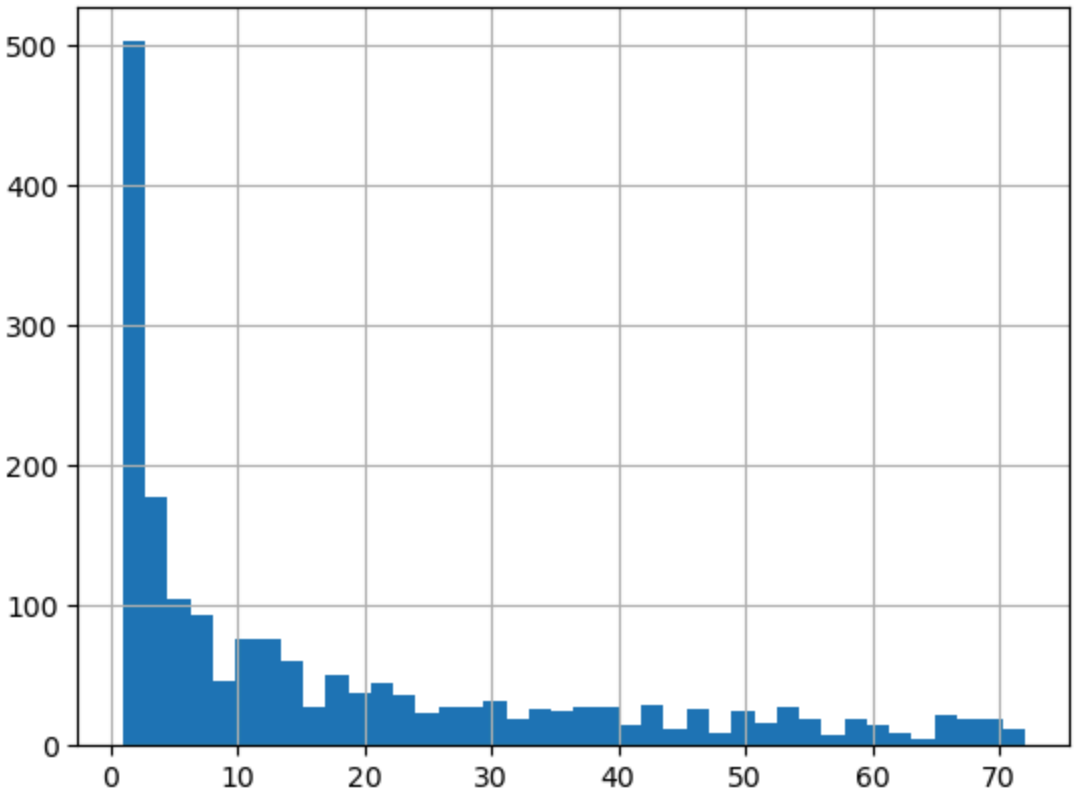
In [23]:
```
df['tenure'].plot.hist(bins=20)
# this has Yes and No Churn data so there are two spikes
```

Out[23]: <Axes: ylabel='Frequency'>

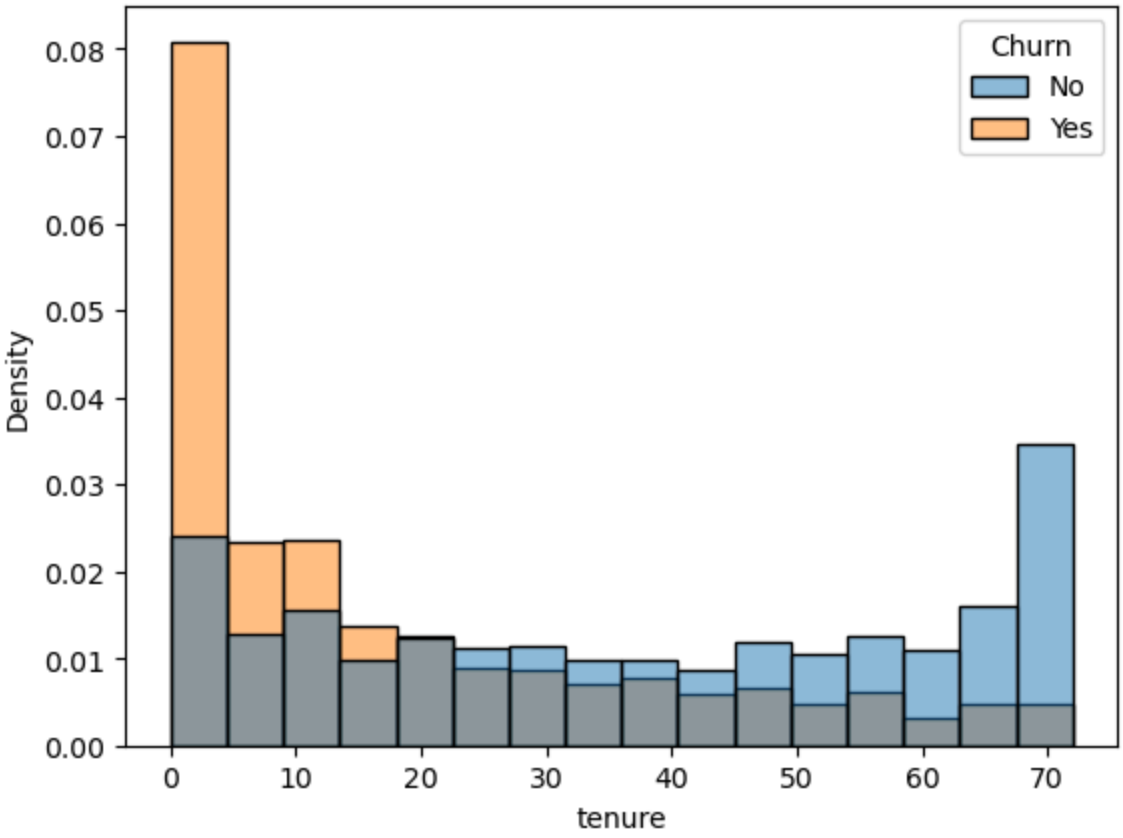


In [24]:
```
df[df.Churn=='Yes'].tenure.hist(bins=40)
# found this on stack overlow, filtered on Churn=Yes to get a clearer view of churn
```

Out[24]: <Axes: >



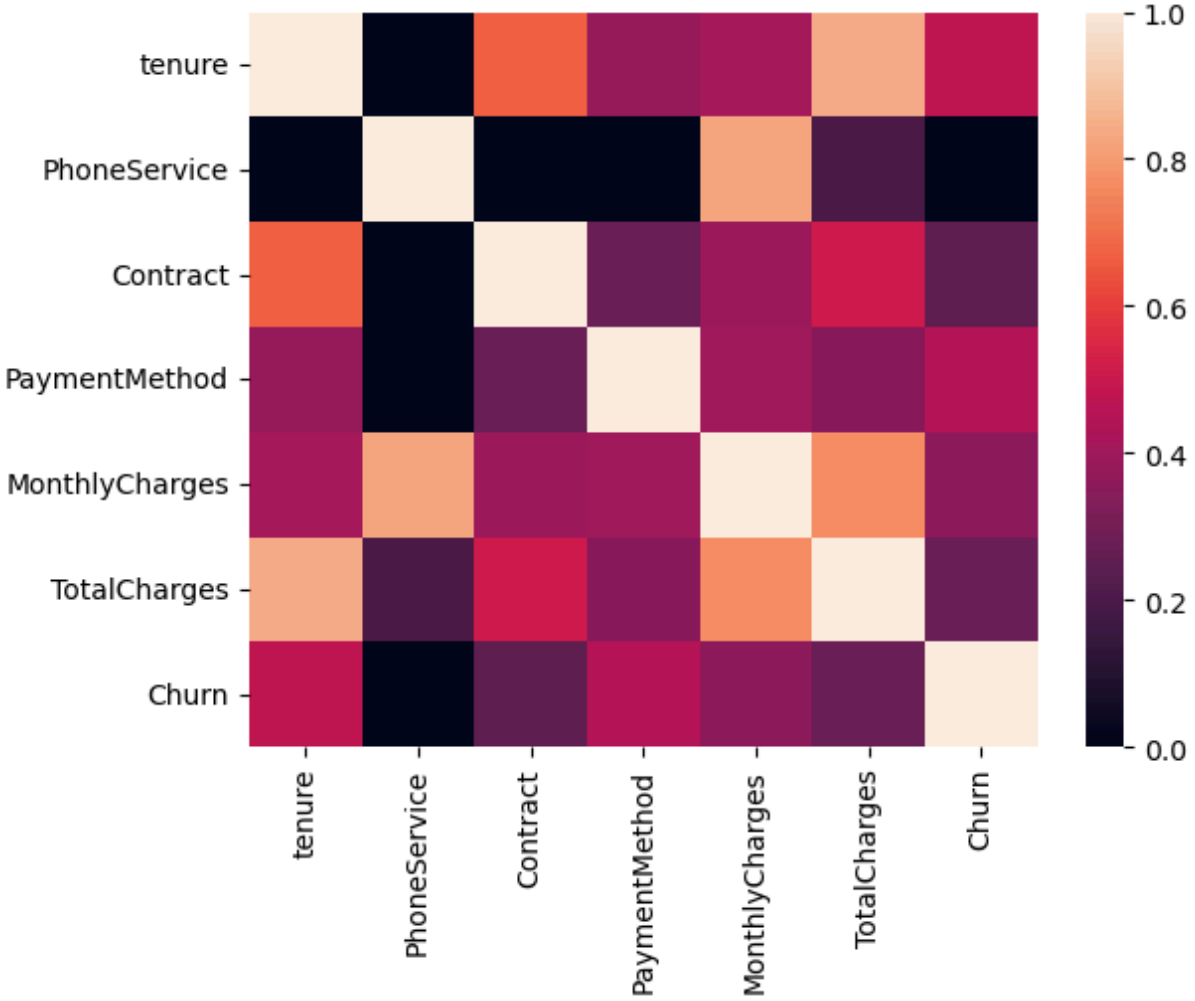
In [25]:
```
import phik
import seaborn as sns

_ = sns.histplot(data=df, x='tenure', hue='Churn', stat='density', common_norm=Fals
```



In [27]:
```
sns.heatmap(df.phik_matrix())
```

```
interval columns not set, guessing: ['tenure', 'MonthlyCharges', 'TotalCharges']
```

Out[27]: <Axes: >



Churn has a large peak around a tenure of 1; tenure ranges from 1 to 72. The churn column has many more customers without churn that with churn.