

KAUNAS UNIVERSITY OF TECHNOLOGY
INFORMATICS FACULTY

INTRODUCTION TO ARTIFICIAL INTELLIGENCE
(P176B101)

LABORATORY WORK №1
Dataset Processing and Analysis

Report author: **Ronan Bonnet**
Lecturer: **Dr Germanas Budnikas**

February 27, 2023

Table of contents

Introduction	1
1 - Purpose	1
2 - Objectives	1
I - Selection of dataset	2
1 - Task	2
2 - Dataset selected	2
3 - Dataset attributes	2
II - Dataset quality analysis	4
1 - Continuous type attributes	4
2 - Categorical type attributes	4
III - Graphical representations and data analysis	5
1 - Histograms columns	5
2 - Histograms and analysis	5
IV - Identifying and solving data quality problems	14
1 - Missing values	14
2 - Cardinality and outliers	14
V - Visualization of relationships between attributes	15
1 - Relationships between attributes of continuous type	15
Scatter plot examples	15
SPLOM diagram	16
2 - Relationships between attributes of categorical type	17
Bar plot examples	17
3 - Relationships between attributes of continuous and categorical type	18
Box plot	18
Histograms	19
VI - Covariance and correlation	21
1 - Covariance	21
2 - Correlation	21
3 - Correlation matrix	22
VII - Data normalization	23
VIII - Categorical attributes conversion	24
Conclusion	25
Dataset quality	25
Data processing	25
Attributes relationships	25

Introduction

1 - Purpose

The purpose of this laboratory work is to make data processing and analysis on a chosen dataset.

2 - Objectives

1. Select (create) a dataset
2. Perform quality analysis of the dataset
3. Draw histograms of attributes
4. Identify data quality problems
5. Identify relationships between attributes using visualization techniques
6. Calculate covariance and correlation values
7. Perform data normalization
8. Convert categorical variables into continuous variables

I - Selection of dataset

1 - Task

Select (create) a dataset to perform this and other laboratory works. Your choice must be approved by the tutor.

Data set requirements:

- Numeric (integer and real) and categorical values must exist.
- For a dataset, the number of records (rows) m must be at least 500, i.e., $m \geq 500$ and the number of attributes n must be at least 8 (columns) $n \geq 8$. If there are fewer attributes in the selected dataset, you have to add derivatives (created)

2 - Dataset selected

The dataset chosen is a set of customer information in IT. The aim of this dataset is to predict customer churn.

Link : [IT Customers Churn \(Goal : Imbalanced Dataset\).](https://www.kaggle.com/datasets/soheiltehranipour/it-customer-churn)

<https://www.kaggle.com/datasets/soheiltehranipour/it-customer-churn>

3 - Dataset attributes

The dataset has 3 different groups of information:

- Services that customer has signed up for
- Customer account information
- Customer demographic information

It also has a column saying if the customer has left the company within the last month.

You can find all the attributes, their type and description in the following table:

Attribute	Type	Description
gender	Category	Gender of the customer (male or female)
SeniorCitizen	Category	Senior citizen or not (1 or 0)
Partner	Category	Customer has a partner or not (Yes or No)
Dependents	Category	Customer has dependents or not (Yes or No)
tenure	Number (int)	Average time since customer has initiated contracts (in years)
PhoneService	Category	Customer has a phone service or not (Yes or No)
MultipleLines	Category	Customer has multiples lines
InternetService	Category	Type of customer's Internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Category	Customer has online security
OnlineBackup	Category	Customer has online backup
DeviceProtection	Category	Customer has device protection
TechSupport	Category	Customer has tech support
StreamingTV	Category	Customer subscribed to Streaming TV
StreamingMovies	Category	Customer subscribe to Streaming Movies
Contract	Category	Customer contract type (One year, Month-to-month, two year, ...)
PaperlessBilling	Category	Customer billing paperless or not (Yes or No)
PaymentMethod	Category	Customer payment method
MonthlyCharges	Number (float)	Customer monthly charges
TotalCharges	Number (float)	Customer total charges
Churn	Category	Customer left within the last month (Yes or No)

II - Dataset quality analysis

1 - Continuous type attributes

	Attribute	Total Values	Percentage of Missing Values	Cardinality	Min	Max	1st Quartile	3rd Quartile	Average	Median	Standard Deviation
0	tenure	7043	0	73	0	72	9	55	32.3711	29	24.5595
1	MonthlyCharges	7043	0	1585	18.25	118.75	35.5	89.85	64.7617	70.35	30.09
2	TotalCharges	7032	0.156183	6530	18.8	8684.8	401.45	3794.74	2283.3	1397.47	2266.77

We can see that the attribute TotalCharges has a small amount of missing values.
Minimum and maximum values of these attributes are coherent.

2 - Categorical type attributes

	Attribute	Total Values	Percentage of Missing Values	Cardinality	Mode	Frequency of Mode	Percentage of Mode	2nd Mode	Frequency of 2nd Mode	Percentage of 2nd Mode
0	gender	7043	0	2	Male	3555	50.4756	Female	3488	49.5244
1	SeniorCitizen	7043	0	2	0	5901	83.7853	1	1142	16.2147
2	Partner	7043	0	2	No	3641	51.6967	Yes	3402	48.3033
3	Dependents	7043	0	2	No	4933	70.0412	Yes	2110	29.9588
4	PhoneService	7043	0	2	Yes	6361	90.3166	No	682	9.68337
5	MultipleLines	7043	0	3	No	3390	48.1329	Yes	2971	42.1837
6	InternetService	7043	0	3	Fiber optic	3096	43.9585	DSL	2421	34.3746
7	OnlineSecurity	7043	0	3	No	3498	49.6663	Yes	2019	28.6668
8	OnlineBackup	7043	0	3	No	3088	43.845	Yes	2429	34.4881
9	DeviceProtection	7043	0	3	No	3095	43.9443	Yes	2422	34.3888
10	TechSupport	7043	0	3	No	3473	49.3114	Yes	2044	29.0217
11	StreamingTV	7043	0	3	No	2810	39.8978	Yes	2707	38.4353
12	StreamingMovies	7043	0	3	No	2785	39.5428	Yes	2732	38.7903
13	Contract	7043	0	3	Month-to-month	3875	55.0192	Two year	1695	24.0664
14	PaperlessBilling	7043	0	2	Yes	4171	59.2219	No	2872	40.7781
15	PaymentMethod	7043	0	4	Electronic check	2365	33.5794	Mailed check	1612	22.888
16	Churn	7043	0	2	No	5174	73.463	Yes	1869	26.537

No categorical attribute has missing values.

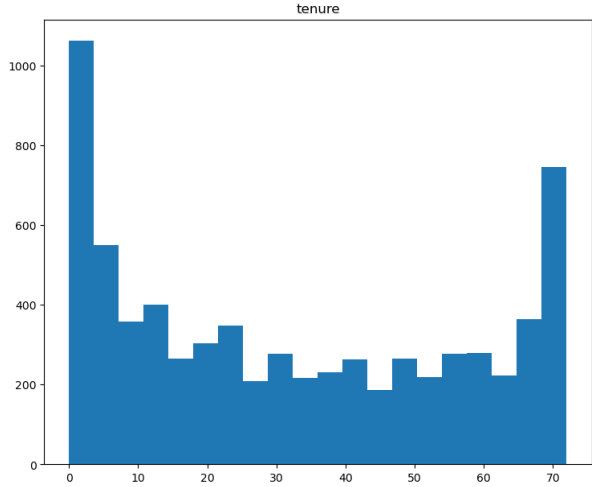
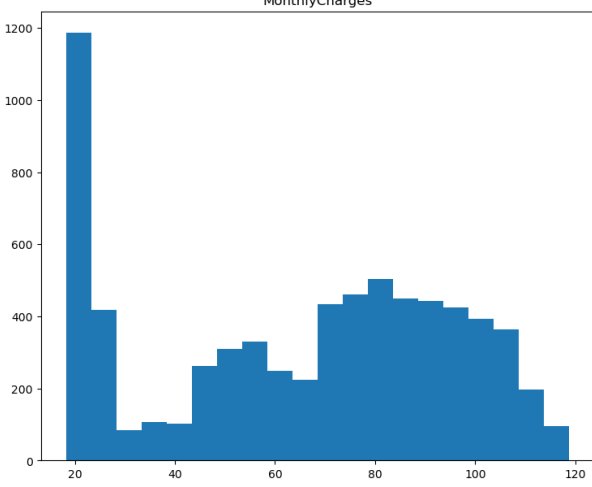
III - Graphical representations and data analysis

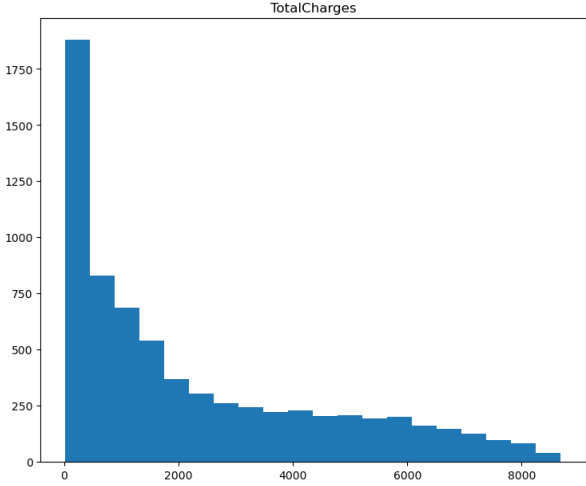
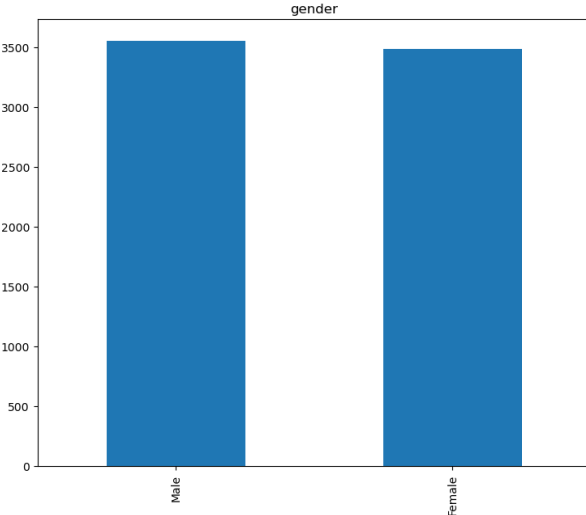
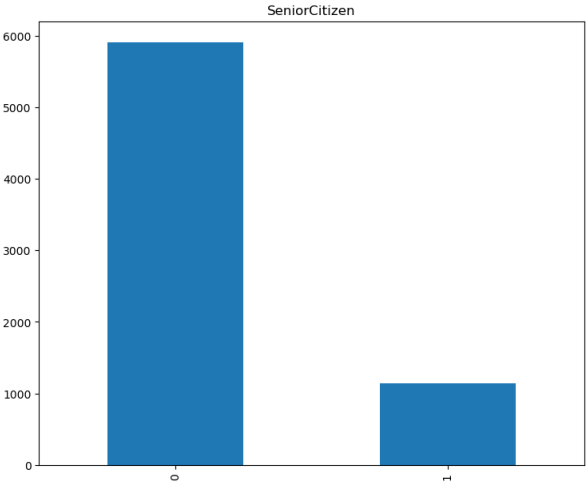
1 - Histograms columns

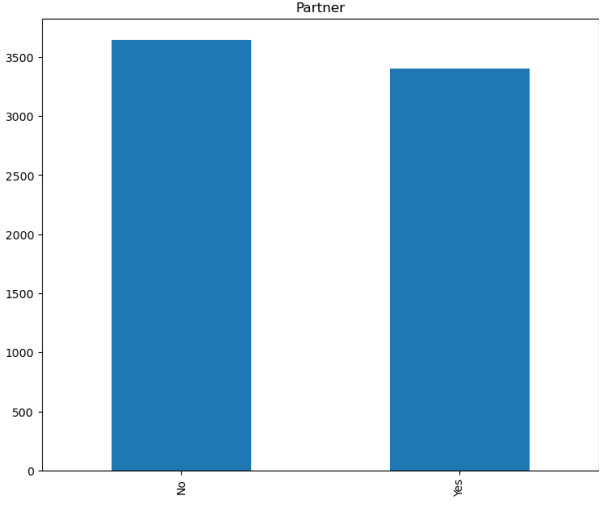
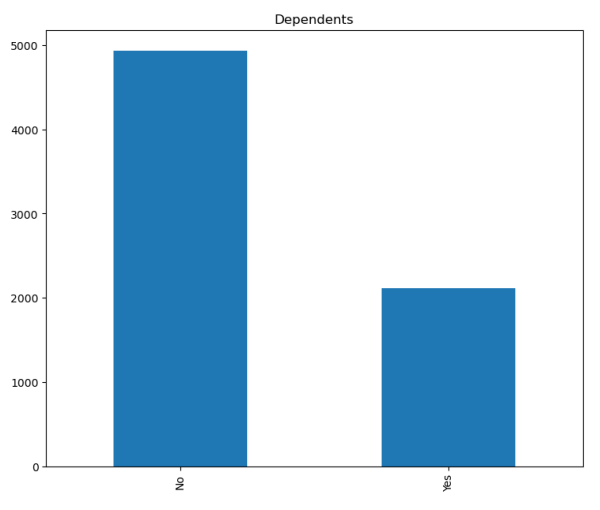
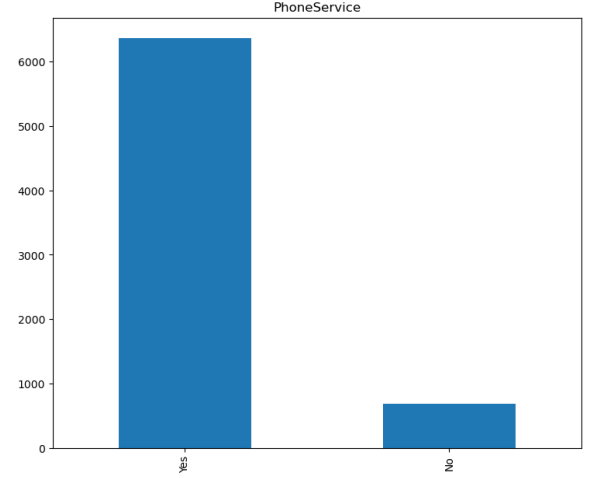
Recommended number of histogram columns is defined by a formula:

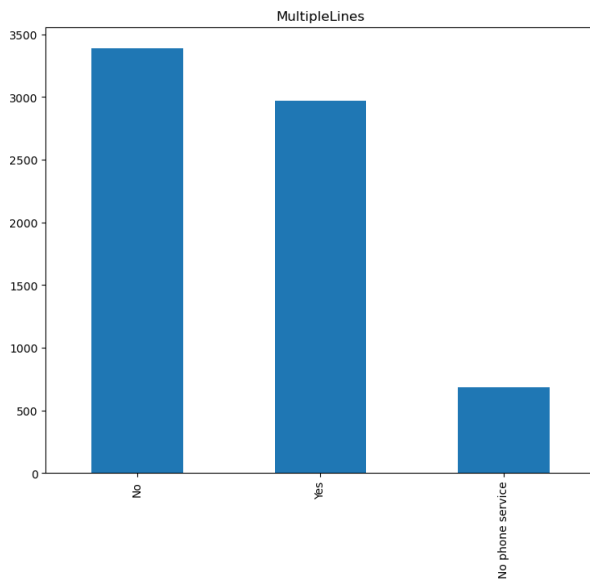
$1 + 3.22 \times \log_e(n)$, where n is sample size

2 - Histograms and analysis

Histograms	Descriptions
	Parabolic distribution
	Normal distribution with a peak on the very left

 <p>A histogram titled 'TotalCharges' showing the frequency of total charges. The x-axis represents total charges from 0 to 8000, and the y-axis represents frequency from 0 to 1750. The distribution is highly positively skewed, with a very high frequency (over 1750) for charges near 0, which rapidly decreases as charges increase.</p>	<p>Positively skewed</p>
 <p>A bar chart titled 'gender' showing the frequency of individuals by gender. The x-axis has two categories: 'Male' and 'Female'. The y-axis represents frequency from 0 to 3500. Both categories have a frequency of approximately 3500, indicating an even distribution.</p>	<p>Two categories Evenly distributed</p>
 <p>A bar chart titled 'SeniorCitizen' showing the frequency of individuals by senior status. The x-axis has two categories: '0' (non-senior) and '1' (senior). The y-axis represents frequency from 0 to 6000. The '0' category has a frequency of approximately 5800, while the '1' category has a much lower frequency of approximately 1100, indicating an uneven distribution.</p>	<p>Two categories Unevenly distributed. There are less seniors than non-seniors.</p>

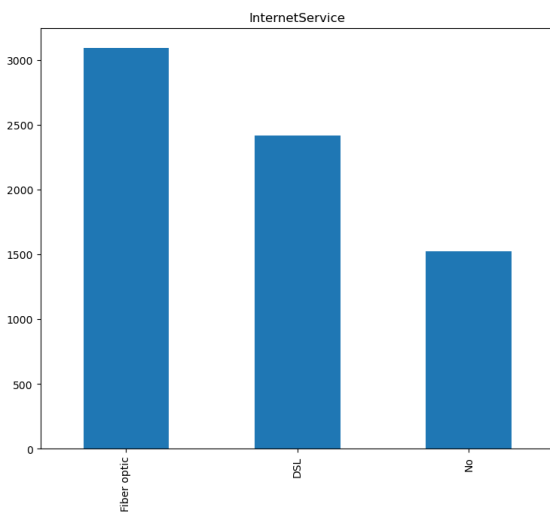
 <p>Partner</p> <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>No</td> <td>~3600</td> </tr> <tr> <td>Yes</td> <td>~3400</td> </tr> </tbody> </table>	Category	Count	No	~3600	Yes	~3400	<p>Two categories</p> <p>Evenly distributed</p>
Category	Count						
No	~3600						
Yes	~3400						
 <p>Dependents</p> <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>No</td> <td>~4900</td> </tr> <tr> <td>Yes</td> <td>~2100</td> </tr> </tbody> </table>	Category	Count	No	~4900	Yes	~2100	<p>Two categories</p> <p>Unevenly distributed</p> <p>There are less dependents people</p>
Category	Count						
No	~4900						
Yes	~2100						
 <p>PhoneService</p> <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>Yes</td> <td>~6300</td> </tr> <tr> <td>No</td> <td>~700</td> </tr> </tbody> </table>	Category	Count	Yes	~6300	No	~700	<p>Two categories</p> <p>Unevenly distributed</p> <p>There are more people who have contracted phone service.</p>
Category	Count						
Yes	~6300						
No	~700						



Three categories

This attribute depends on the PhoneService attribute. That is why there are two possible "No" options.

For people having a phone service, a little bit less have multiple lines than those who have a single line

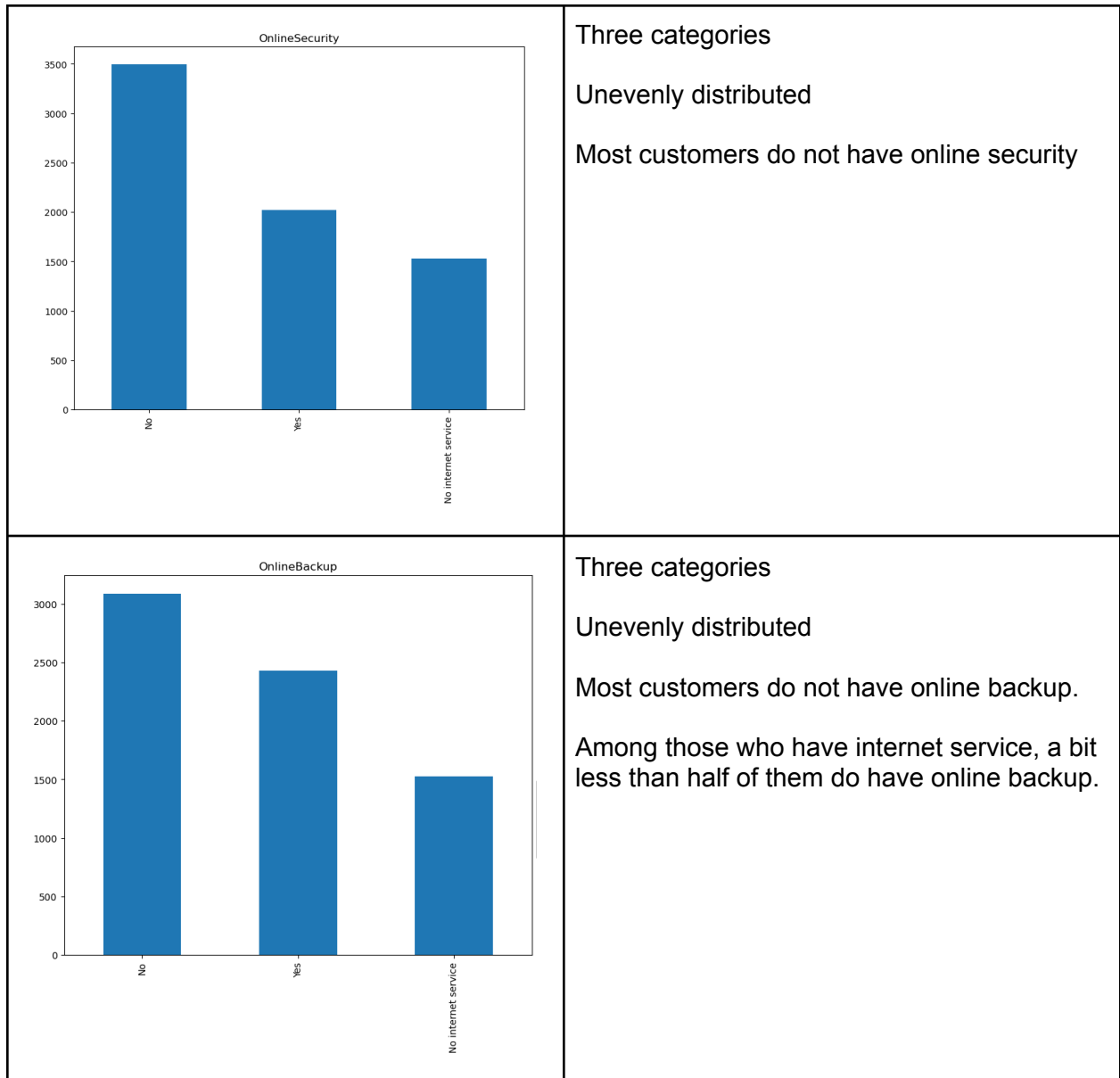


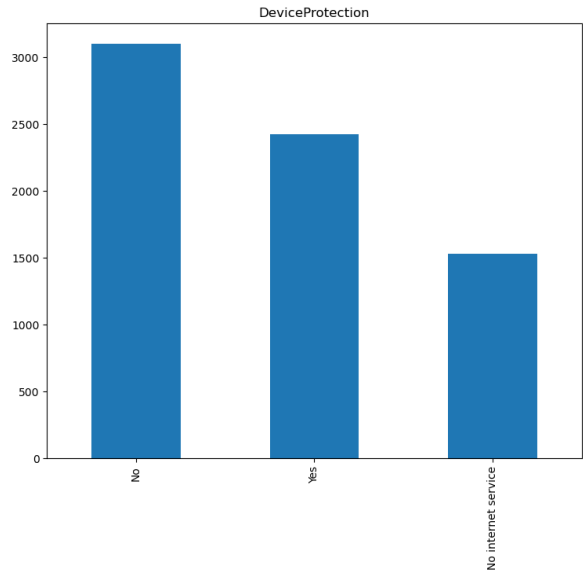
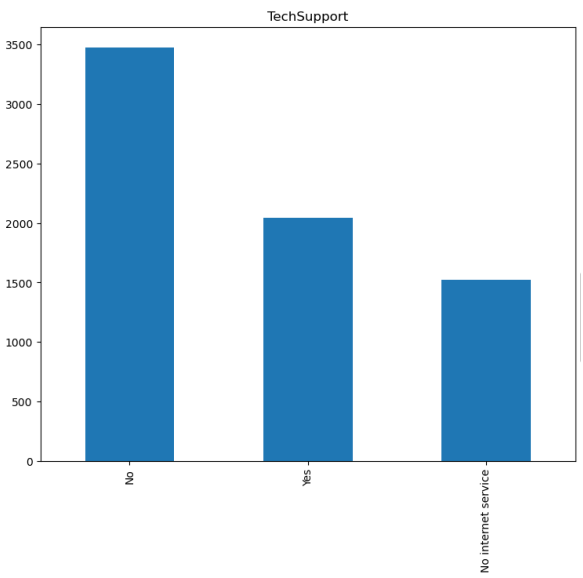
Three categories

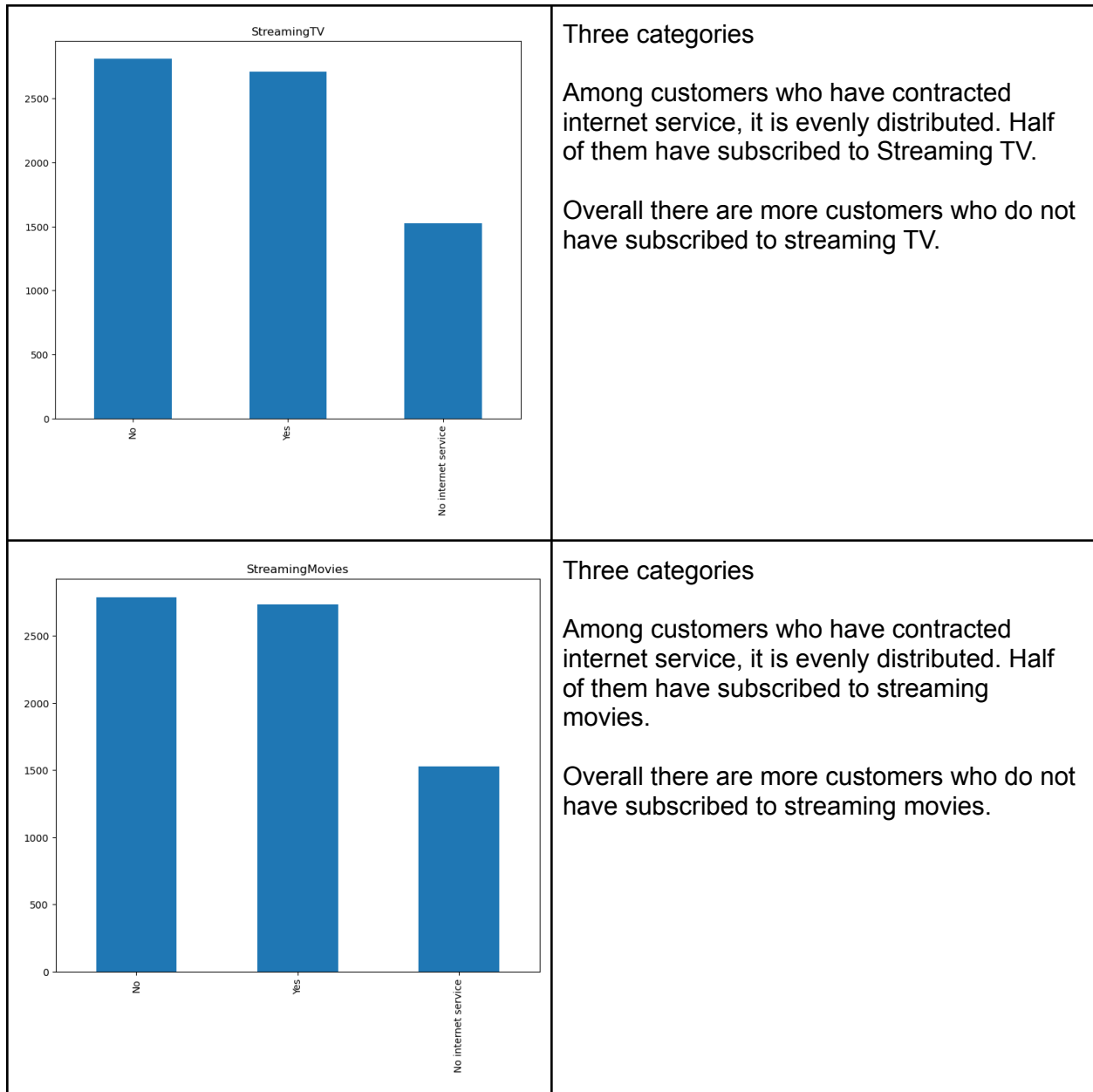
Two types of Internet services: Fiber optic or DSL.

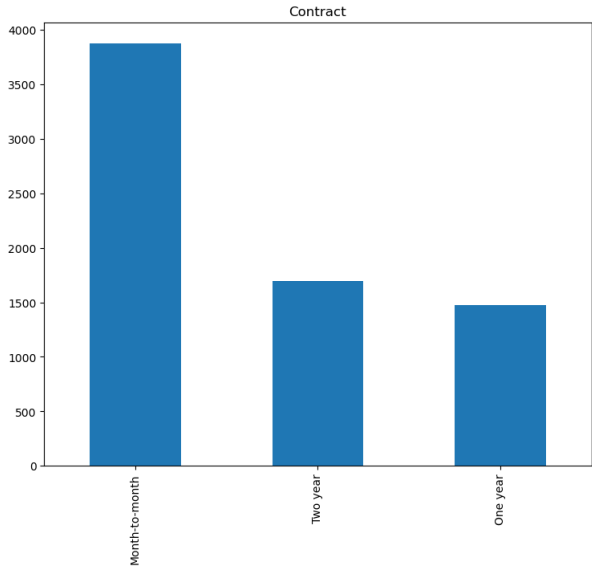
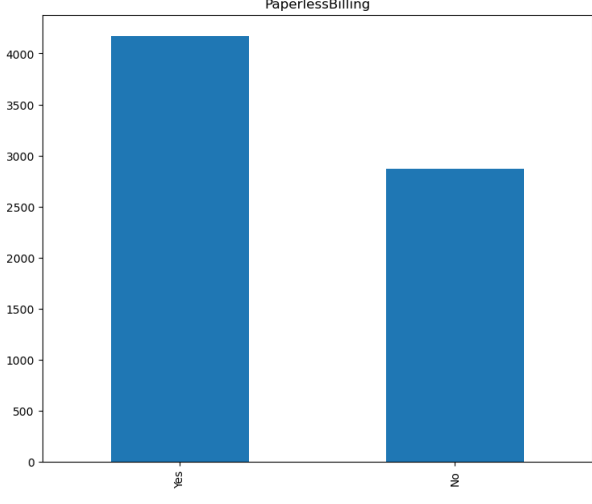
A small number of customers do not have Internet service.

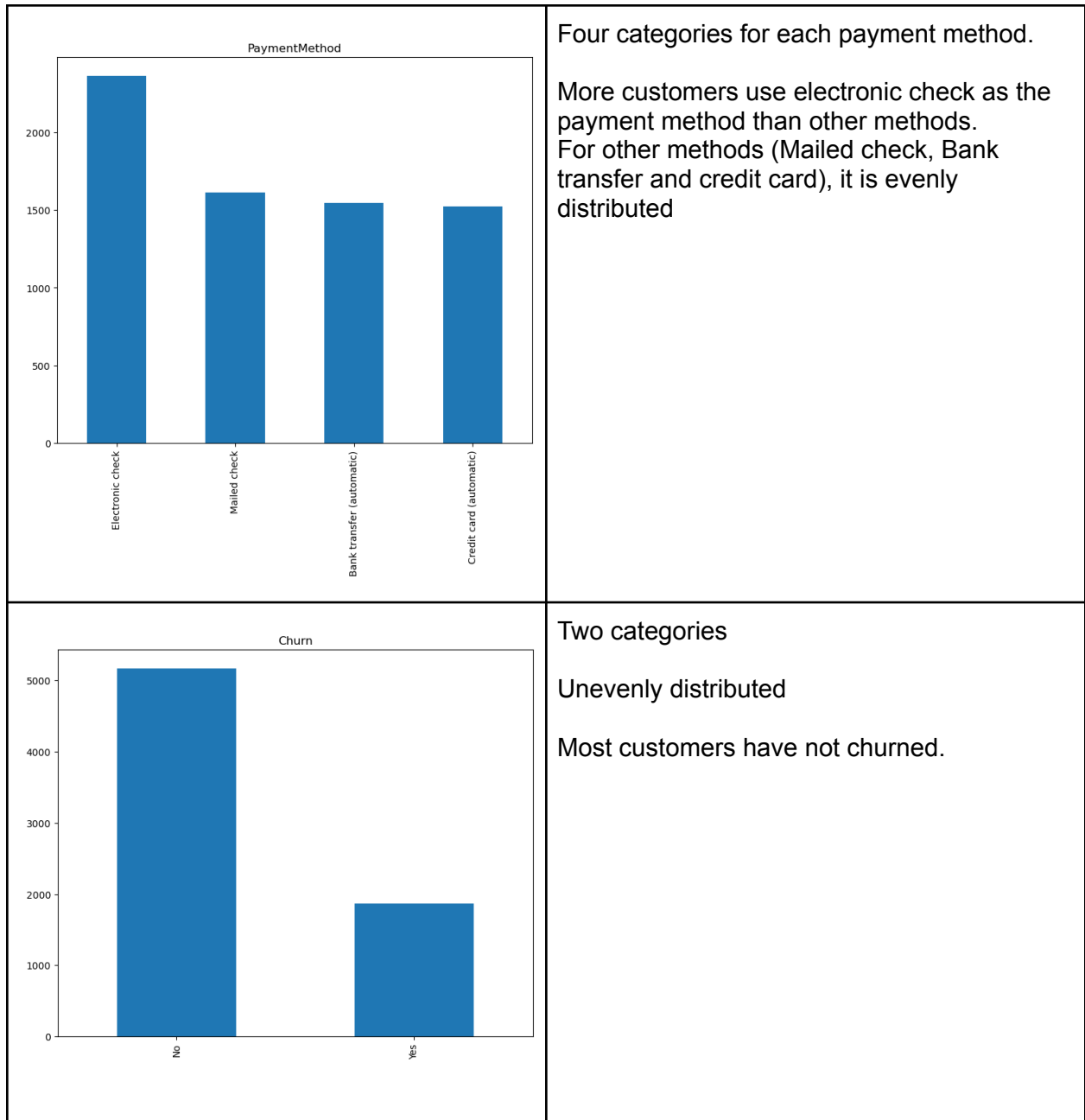
Among those who have, there are a bit more customers who have fiber optic than those who have DSL.



 <table border="1"> <caption>DeviceProtection Data</caption> <thead> <tr> <th>Category</th> <th>Count (approx.)</th> </tr> </thead> <tbody> <tr> <td>No</td> <td>3100</td> </tr> <tr> <td>Yes</td> <td>2400</td> </tr> <tr> <td>No internet service</td> <td>1500</td> </tr> </tbody> </table>	Category	Count (approx.)	No	3100	Yes	2400	No internet service	1500	<p>Three categories</p> <p>Unevenly distributed</p> <p>Less than half of customers who contracted internet service have device protection.</p> <p>The histogram is almost the same as the histogram of OnlineBackup attribute</p>
Category	Count (approx.)								
No	3100								
Yes	2400								
No internet service	1500								
 <table border="1"> <caption>TechSupport Data</caption> <thead> <tr> <th>Category</th> <th>Count (approx.)</th> </tr> </thead> <tbody> <tr> <td>No</td> <td>3400</td> </tr> <tr> <td>Yes</td> <td>2000</td> </tr> <tr> <td>No internet service</td> <td>1500</td> </tr> </tbody> </table>	Category	Count (approx.)	No	3400	Yes	2000	No internet service	1500	<p>Three categories</p> <p>Unevenly distributed</p> <p>Most customers do not have tech support</p>
Category	Count (approx.)								
No	3400								
Yes	2000								
No internet service	1500								



 <table border="1"> <caption>Contract Data</caption> <thead> <tr> <th>Contract Type</th> <th>Count (Approximate)</th> </tr> </thead> <tbody> <tr> <td>Month-to-month</td> <td>3900</td> </tr> <tr> <td>Two year</td> <td>1700</td> </tr> <tr> <td>One year</td> <td>1500</td> </tr> </tbody> </table>	Contract Type	Count (Approximate)	Month-to-month	3900	Two year	1700	One year	1500	<p>Three categories for each type of contract</p> <p>Unevenly distributed</p> <p>Half of customers have a month-to-month contract. The other ones evenly have either a two-year or a one-year contract.</p>
Contract Type	Count (Approximate)								
Month-to-month	3900								
Two year	1700								
One year	1500								
 <table border="1"> <caption>PaperlessBilling Data</caption> <thead> <tr> <th>Paperless Billing</th> <th>Count (Approximate)</th> </tr> </thead> <tbody> <tr> <td>Yes</td> <td>4200</td> </tr> <tr> <td>No</td> <td>2900</td> </tr> </tbody> </table>	Paperless Billing	Count (Approximate)	Yes	4200	No	2900	<p>Two categories</p> <p>Unevenly distributed.</p> <p>There are more customers who have paperless billing.</p>		
Paperless Billing	Count (Approximate)								
Yes	4200								
No	2900								



IV - Identifying and solving data quality problems

1 - Missing values

The attribute TotalCharges misses 11 values.

	gender	SeniorCitizen	tenure	PhoneService	MultipleLines	InternetService	MonthlyCharges	Contract	PaymentMethod	TotalCharges	Churn
488	Female	0	0	No	No phone service	DSL	52.55	Two year	Bank transfer (automatic)	NaN	No
753	Male	0	0	Yes	No	No	20.25	Two year	Mailed check	NaN	No
936	Female	0	0	Yes	No	DSL	80.85	Two year	Mailed check	NaN	No
1082	Male	0	0	Yes	Yes	No	25.75	Two year	Mailed check	NaN	No
1340	Female	0	0	No	No phone service	DSL	56.05	Two year	Credit card (automatic)	NaN	No
3331	Male	0	0	Yes	No	No	19.85	Two year	Mailed check	NaN	No
3826	Male	0	0	Yes	Yes	No	25.35	Two year	Mailed check	NaN	No
4380	Female	0	0	Yes	No	No	20.00	Two year	Mailed check	NaN	No
5218	Male	0	0	Yes	No	No	19.70	One year	Mailed check	NaN	No
6670	Female	0	0	Yes	Yes	DSL	73.35	Two year	Mailed check	NaN	No
6754	Male	0	0	Yes	Yes	DSL	61.90	Two year	Bank transfer (automatic)	NaN	No

I have selected some important attributes to show the rows in which TotalCharges misses values.

We can see that these customers all have 0 tenure. We can imagine that these customers are new customers. So I set their TotalCharges to their MonthlyCharges values.

2 - Cardinality and outliers

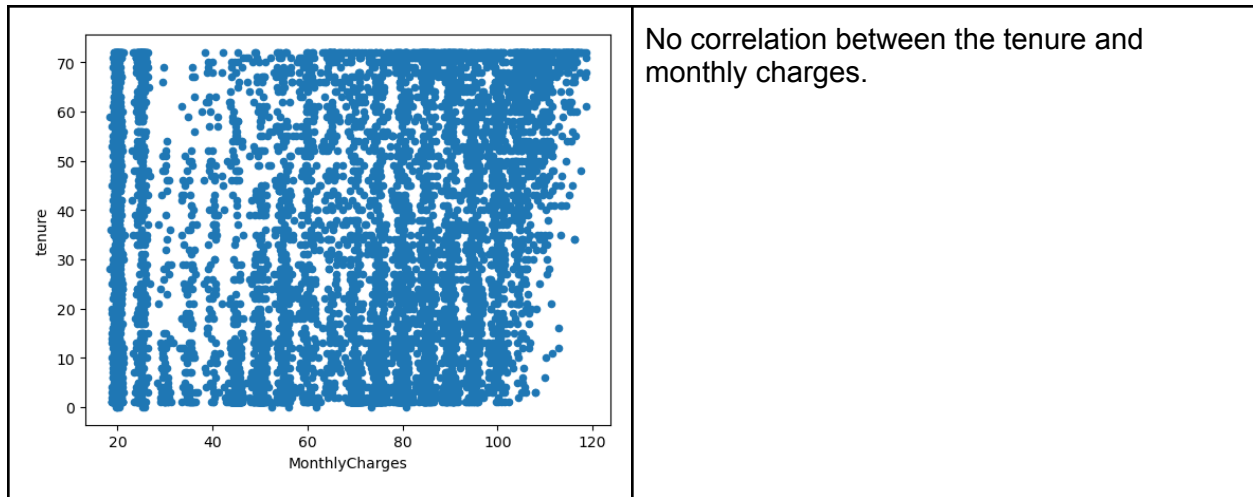
There are no cardinality problems nor outliers.

V - Visualization of relationships between attributes

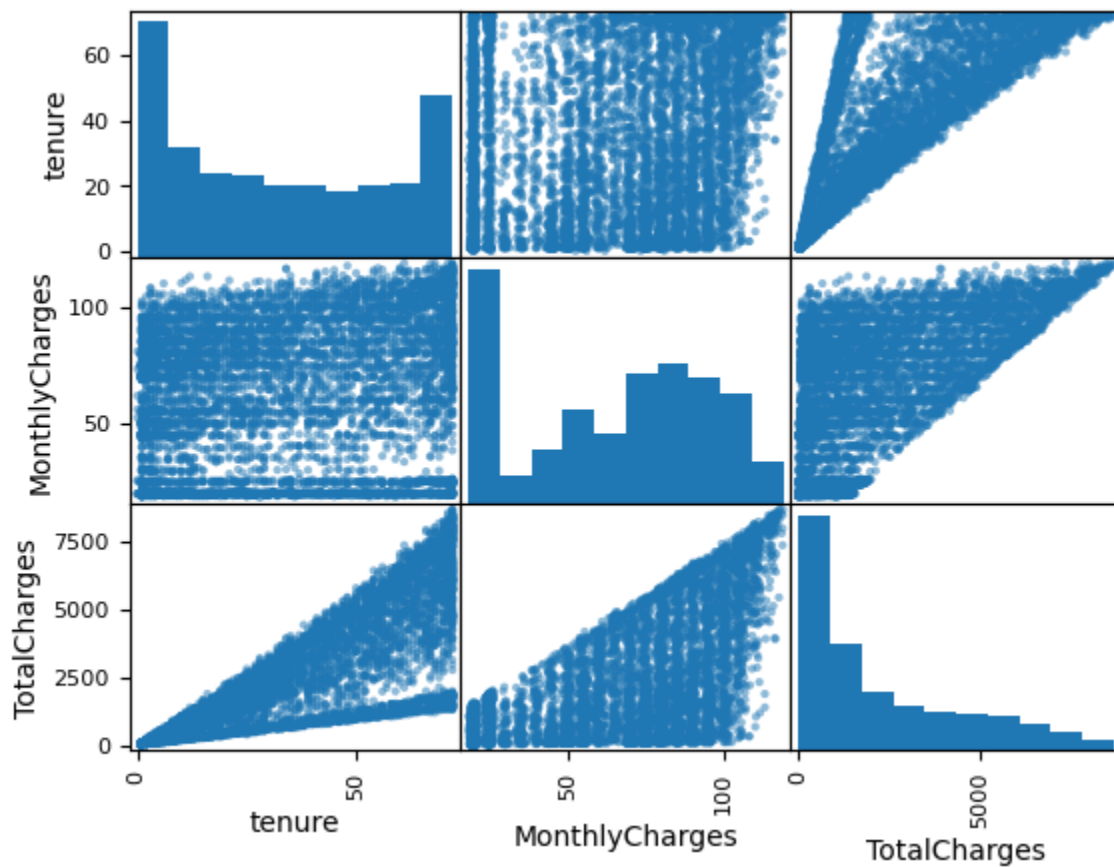
1 - Relationships between attributes of continuous type

Scatter plot examples





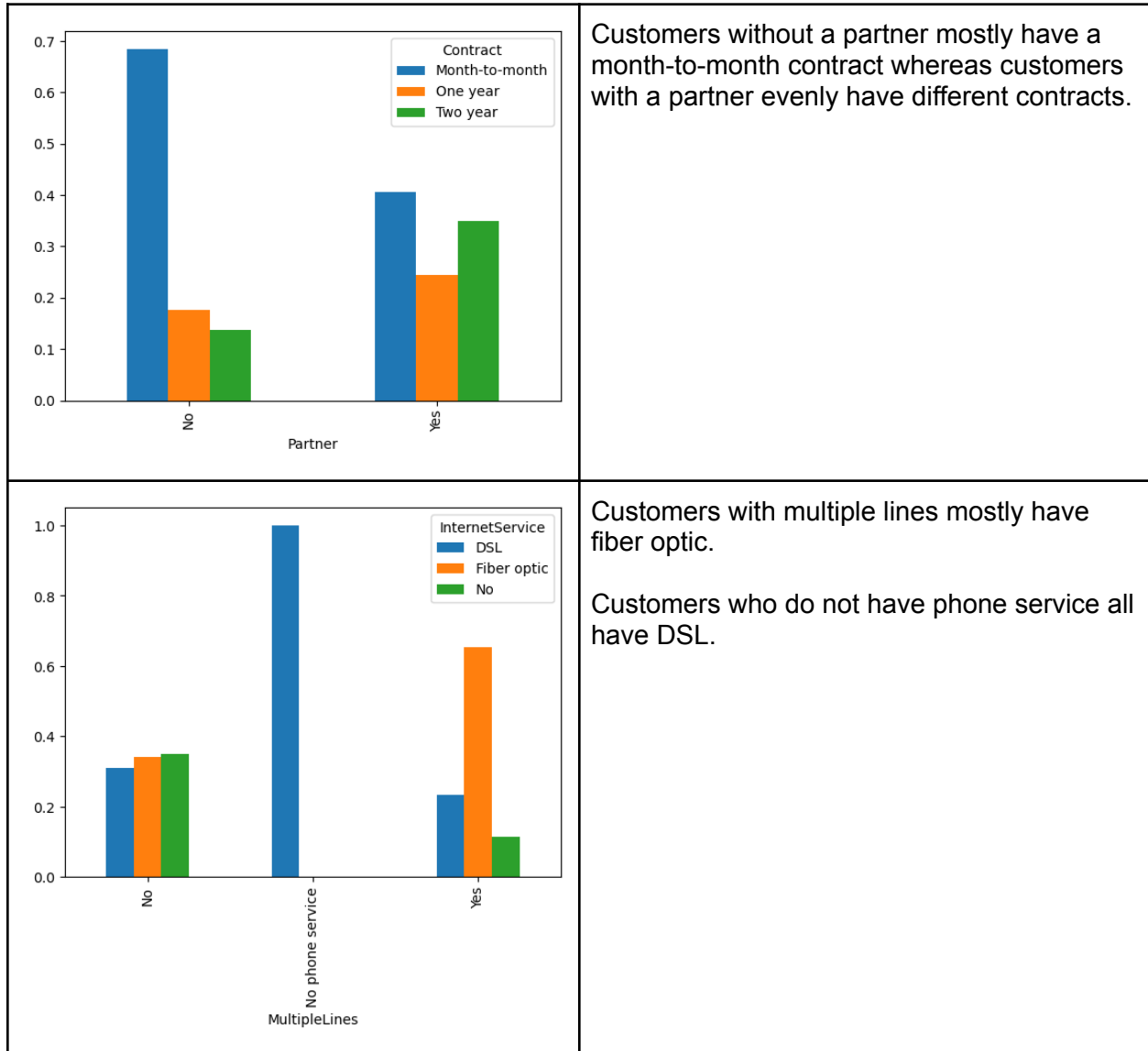
SPLOM diagram



As I have only 3 continuous attributes, I have already concluded on their correlations above.

2 - Relationships between attributes of categorical type

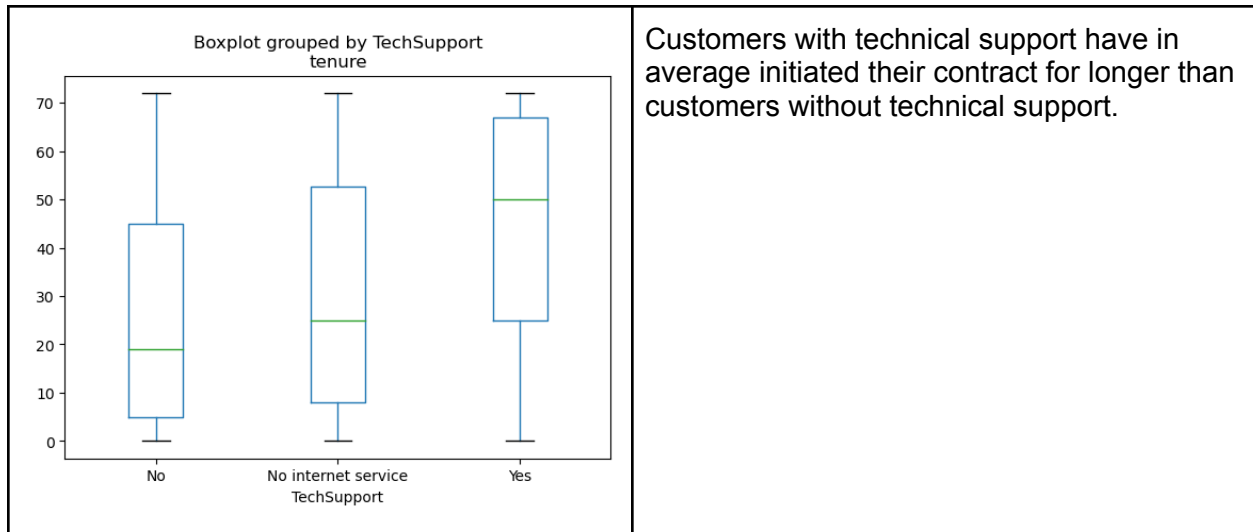
Bar plot examples



3 - Relationships between attributes of continuous and categorical type

Box plot





Histograms





VI - Covariance and correlation

1 - Covariance

	tenure	MonthlyCharges	TotalCharges
tenure	603.168108	183.196987	4.595074e+04
MonthlyCharges	183.196987	905.410934	4.440133e+04
TotalCharges	45950.743236	44401.333073	5.138252e+06

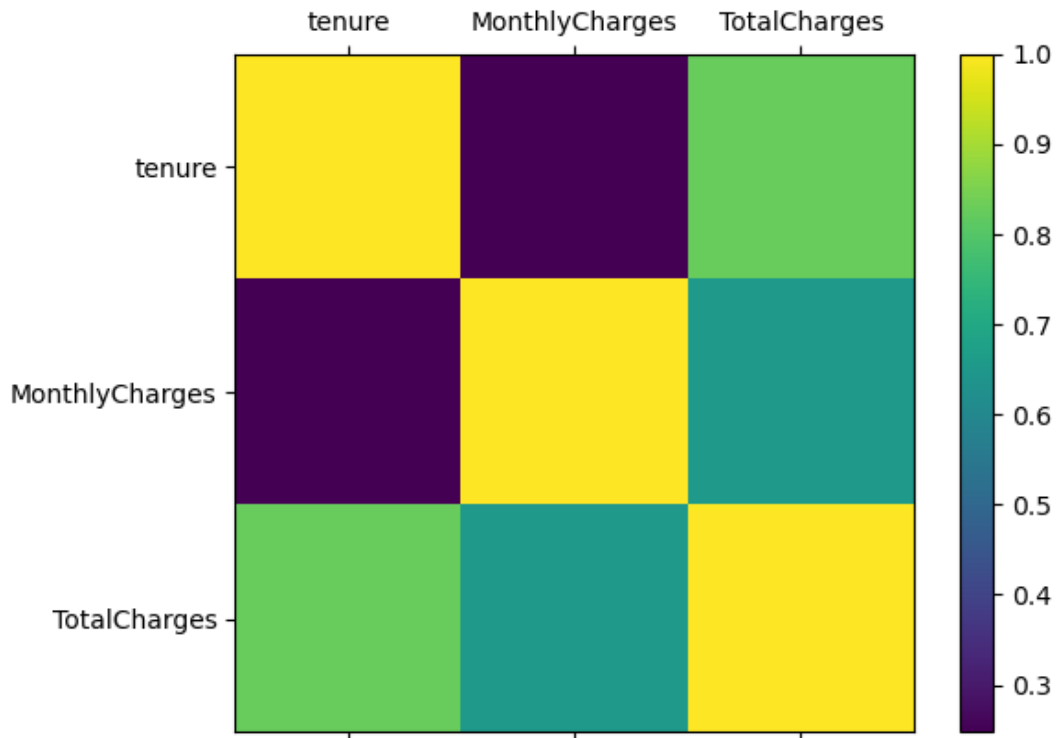
We can see that the covariance is low between MonthlyCharges and tenure. It is quite high between TotalCharges and other continuous attributes.

2 - Correlation

	tenure	MonthlyCharges	TotalCharges
tenure	1.00000	0.247900	0.825880
MonthlyCharges	0.24790	1.000000	0.651065
TotalCharges	0.82588	0.651065	1.000000

We find the same analysis as covariance analysis.

3 - Correlation matrix



Tenure and TotalCharges are highly correlated, the longest the customer has initiated his contract the most he has paid.

Monthly Charges and Total Charges are correlated.

Monthly Charges and tenure are poorly correlated.

VII - Data normalization

I have normalized all continuous values following a min-max scale to values between 0 and 1.

I have chosen the min-max scale because it uses a linear scale.

Here is a sample of the normalized data:

	tenure	MonthlyCharges	TotalCharges
6126	0.263889	0.772637	0.202989
3157	0.125000	0.265672	0.047565
6321	0.416667	0.013930	0.073511
1523	0.111111	0.251741	0.036534
1340	0.000000	0.376119	0.004298

VIII - Categorical attributes conversion

I wrote a function to map categorical values to numbers and applied it to the dataset. You can see a table showing some attributes:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
4910	1	0	0	0	0.416667	1	0	2	1	1	1
6779	0	0	0	1	0.402778	0	2	2	1	1	1
5868	1	0	0	0	0.013889	1	0	2	0	0	0
1220	0	0	0	0	0.236111	1	1	2	1	0	0
2560	0	0	0	0	0.638889	1	1	1	1	1	0

As the new values do not belong to the normalization boundaries. I normalized the categorical values to the same boundaries as continuous ones.

Conclusion

Dataset quality

The dataset was found to contain a few missing values. It has some peaks on sides for some attributes. However, no extreme values or other problems have been detected.

Data processing

As we had a few missing values, the solution chosen to solve missing values would not change the dataset a lot.

To solve the missing values problem, many options were possible (law of mean, law of median...). The option selected is the one that would be the most realistic.

Attributes relationships

We have analyzed many relationships. For continuous attributes, only one or two have a high correlation. For the categorical attributes, we can find a lot of relationships, especially between extra services such as online security, online backup, etc.

Overall, there are not a lot of attributes which are highly correlated.