Temasek POLYTECHNIC | School of Engineering

Data Visualisation and Analytics (ESE1008)

Project Report

Declaration of Originality

I am the originator of this work and I have appropriately acknowledged all other original sources used in this work.

I understand that Plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my own without proper acknowledgement.

I understand that Plagiarism is an academic offence
and if I am found to have committed or abetted the offence of plagiarism in relation to this submitted work, disciplinary action will be enforced.

Submitted By	**Student's Signature**
Riffaie Fathan Bin Rahim	

Class: PE13

AY2022/2023 OCT SEMESTER

Content Page

1. Pre-Project Plan

Goal Setting
I aim to complete my project by 20*/1/2023.* I shall take initiative to find out the information needed. I shall check the project rubric to ensure all items are done before submission.

My data set is TELCO-5

2. Preliminary questions

My preliminary questions that I will answer from my data set:

1. Is the tenure correlated to the monthly charge?

2. What is the composition of churn in this dataset?

3. What do boxplots of numerical features against variable “Churn” in the Telco-5 dataset suggest?

4. What is the mode of highest churn for monthly charges?

5. What is the most common Contract type? In addition on average, which contract type is cheaper monthly?

6. For customers with phone service, how many have multiple lines.

7. Which contract type are customers more likely to churn?

8. What are the range of monthly charges for each contract length?

3. Monitor

Task/Milestone	**By When**	**Actual Completed Date**	**Comment** (on-time/delay/early)
Download the data. Understand the rows and columns.	**3/12/2022**	**28/11/2022**	**Downloaded early so as to understand given dataset earlier**
Background research of delivery mode, function of eCommerce.	**3/12/2022**	**28/11/2022**	**Looked though Kaggle to gather more information about Telco Dataset**
Perform data cleaning.	**6/12/2022**	**6/12/2022**	**Submitted msteams data cleaning assignment and got feedback from my lecturer**
Perform data transformation.	**20/12/2022**	**18/12/2022**	**Earlier than intended to perform data analysis as soon as possible**
Exploratory Data Analysis	**22/12/2022**	**20/12/2022**	**Performed earlier than intended so as to submit before the deadline – less stress, more time**
Submit Report 1	**23 Dec 2022 (Due date)**	**22/12/2022**	**Submitted a day earlier**
Answer my preliminary questions	**28/12/2022**	**18/1/2023**	**Late, however managed to change and refine some prelimary questions**
Data modeling	**12/1/2023**	**12/1/2023**	**On time as I had did it in class with help**

			from my lecturer
Final report conclusion and reflection	**20/1/2023**	**20/1/2023**	**On time**
Create Dashboard	**20/1/2023**	**20/1/2023**	**On time**

4. Introduction

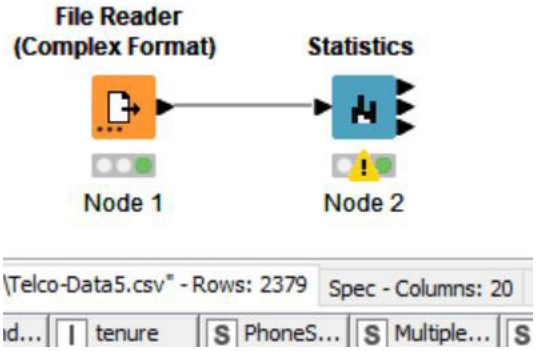
The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to customers in United State. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer.

Being able to retain customers is essential for the business of Telco companies to find out what is the exact cause of customer churning/leaving contracts, and also to stay competitive with other Telco companies.

My objective is to use the data to predict behaviour to retain customers from churning and analysing all relevant customer data and develop focused customer retention programs.

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents [1].

The statistic node is useful to get more information about the variables in the data.



There are 2379 Observations, and
20 Variables.

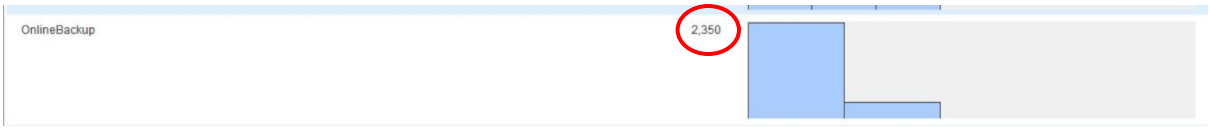
Name of Variable	Data type
customerID	Categorical
gender	Categorical, Male, Female
SeniorCitizen	Numerical, Discrete between 0 to 1
Partner	Categorical, No, Yes
Dependents	Categorical, No, Yes
tenure	Numerical, Discrete between 1 to 72
PhoneService	Categorical, No, Yes
MultipleLines	Categorical, No, Yes, No phone service
InternetService	Categorical, No, Fiber optic, DSL
OnlineSecurity	Categorical, No, Yes, No internet service
OnlineBackup	Categorical, No
TechSupport	Categorical, No, Yes, No internet service
StreamingTV	Categorical, No, Yes, No internet service
StreamingMovies	Categorical, No, Yes, No internet service
Contract	Categorical, Month-to-month, One year, Two year
PaperlessBilling	Categorical, No, Yes
PaymentMethod	Categorical, Electronic check, Credit card (automatic), Bank transfer (automatic)
MonthlyCharges	Numerical, Continuous between 18.55 to 118.75
TotalCharges	Numerical, Continuous between 19.25 to 8,684.8
Churn	Categorical, No, Yes

5. Data Cleaning

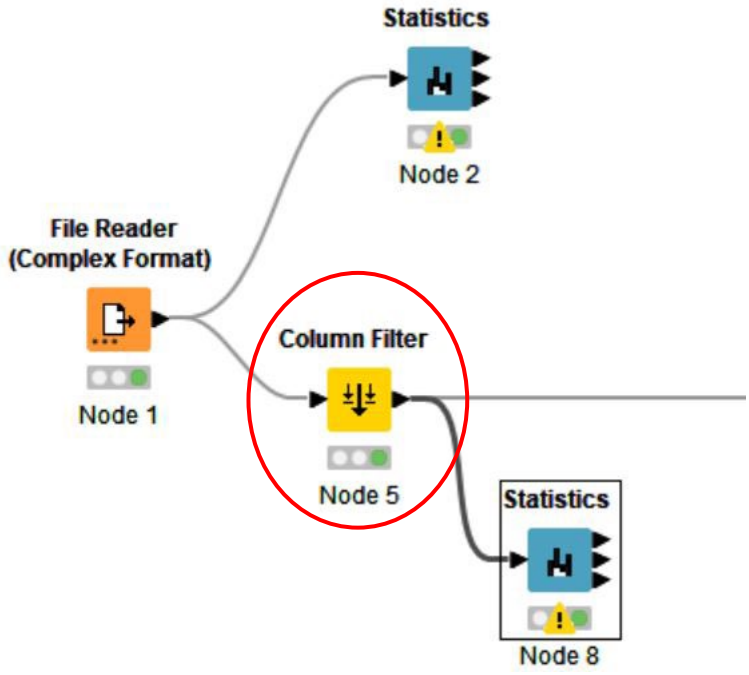
Row ID	custom...	gender	SeniorC...	Partner	Depend...	tenure	PhoneS...	Multiple...	Interne...	OnlineS...	OnlineB...	TechSu...	Streami...	Streami...	Contract	
Row0	5655-JSMZM	Male	1	No	No	49	Yes	Yes	Fiber optic	No	?	No	Yes	Yes	Month-to-month	Y
Row1	9861-PDSZP	Female	0	No	No	72	Yes	Yes	Fiber optic	No	?	Yes	Yes	Yes	Two year	Y
Row2	4505-EXZHB	Female	1	No	No	6	Yes	No	Fiber optic	No	?	No	No	No	Month-to-month	N
Row3	7225-CBZPL	Male	1	Yes	No	17	Yes	Yes	Fiber optic	No	?	No	Yes	Yes	Month-to-month	Y
Row4	6704-UTUKK	Male	0	Yes	No	29	No	No phone se...	DSL	No	?	No	Yes	Yes	Month-to-month	Y
Row5	4587-VVTOX	Female	0	Yes	No	6	Yes	Yes	Fiber optic	No	?	No	Yes	Yes	Month-to-month	Y
Row6	2019-HDCZY	Male	0	Yes	No	63	Yes	Yes	Fiber optic	No	?	No	Yes	Yes	Two year	N
Row7	4652-NNHNY	Male	0	Yes	No	16	Yes	Yes	Fiber optic	No	?	No	No	No	Month-to-month	Y
Row8	8788-DOXSU	Male	0	No	No	59	Yes	No	DSL	No	?	No	No	Yes	One year	Y
Row9	7404-JLKQG	Female	0	No	No	3	Yes	No	DSL	No	?	No	Yes	No	Month-to-month	N
Row10	8972-HJWNV	Female	1	Yes	No	7	Yes	Yes	Fiber optic	No	?	No	No	Yes	Month-to-month	Y
Row11	3274-NSDWE	Female	0	No	No	68	Yes	No	No	No internet ...	?	No internet ...	No internet ...	No internet ...	Two year	N
Row12	2632-IVXVF	Female	0	Yes	Yes	68	Yes	Yes	Fiber optic	Yes	?	Yes	Yes	Yes	Two year	N
Row13	3692-JHONH	Female	1	Yes	No	52	Yes	Yes	Fiber optic	No	?	No	Yes	Yes	One year	Y
Row14	8915-NNTRC	Male	0	Yes	Yes	72	Yes	Yes	Fiber optic	Yes	?	Yes	Yes	Yes	Two year	Y
Row15	7463-IFMQU	Female	0	Yes	No	72	Yes	No	No	No internet ...	?	No internet ...	No internet ...	No internet ...	Two year	N
Row16	2920-RNCEZ	Male	0	Yes	Yes	1	Yes	No	Fiber optic	No	?	No	No	No	Month-to-month	Y
Row17	2541-YGPKE	Male	0	Yes	Yes	42	Yes	No	DSL	Yes	?	Yes	No	Yes	One year	N
Row18	8515-OCTJS	Female	0	No	No	25	Yes	Yes	No	No internet ...	?	No internet ...	No internet ...	No internet ...	Two year	N
Row19	5382-TEMLV	Male	0	No	No	45	Yes	No	DSL	Yes	?	No	No	No	Month-to-month	Y
Row20	3441-CGZJH	Female	0	Yes	Yes	43	No	No phone se...	DSL	Yes	?	Yes	Yes	Yes	One year	Y
Row21	2860-RANUS	Female	1	No	No	20	Yes	Yes	Fiber optic	No	?	No	No	Yes	Month-to-month	Y
Row22	5261-QSHQM	Female	0	No	No	4	No	No phone se...	DSL	No	?	No	No	No	Month-to-month	Y
Row23	4778-IZARL	Male	0	Yes	No	63	Yes	No	Fiber optic	Yes	?	Yes	Yes	Yes	Two year	Y
Row24	7008-LZVOZ	Male	0	Yes	Yes	66	Yes	Yes	No	No internet ...	?	No internet ...	No internet ...	No internet ...	Two year	Y
Row25	8868-WOZGU	Male	0	No	No	28	Yes	Yes	Fiber optic	No	?	No	Yes	Yes	Month-to-month	Y
Row26	1200-TUZHR	Female	1	No	No	8	Yes	Yes	Fiber optic	No	?	No	No	Yes	Month-to-month	N
Row27	1334-FJSVR	Male	0	No	No	1	No	No phone se...	DSL	No	?	No	No	No	Month-to-month	Y
Row28	5884-FBCTL	Female	0	Yes	Yes	72	Yes	Yes	No	No internet ...	?	No internet ...	No internet ...	No internet ...	Two year	N
Row29	7130-CTCUS	Male	1	Yes	No	16	Yes	No	DSL	Yes	?	No	No	No	Month-to-month	Y
Row30	7625-XCQRH	Female	0	No	No	11	Yes	Yes	Fiber optic	No	?	No	No	No	Month-to-month	Y
Row31	6194-HBGQN	Male	0	No	No	51	Yes	Yes	DSL	Yes	?	Yes	Yes	No	One year	Y
Row32	7634-WSWDB	Female	0	No	Yes	8	No	No phone se...	DSL	Yes	?	No	No	Yes	Month-to-month	Y
Row33	6986-IXNDM	Male	0	No	No	14	Yes	No	Fiber optic	No	?	No	Yes	Yes	Month-to-month	Y
Row34	1731-TVIUK	Female	0	No	No	4	Yes	Yes	Fiber optic	No	?	Yes	No	Yes	Month-to-month	Y
Row35	9769-TSBZE	Female	0	No	Yes	70	Yes	Yes	DSL	Yes	?	Yes	No	No	Two year	N

Rows: 2379 | Spec - Columns: 20

Initial row of 2379 and column of 20

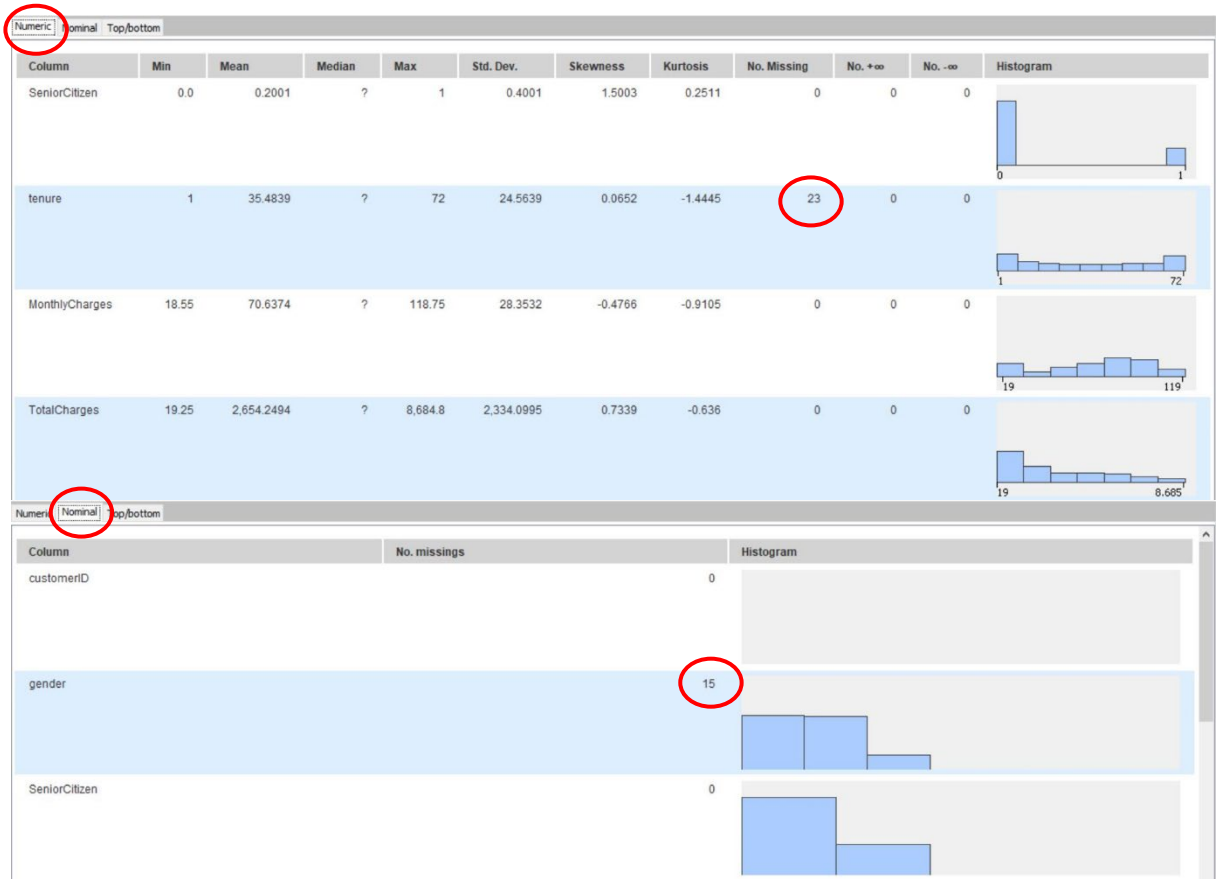


From the statistics page, column OnlineBackup is (2350/2379 x 100) = 98.781% missing, I had use the Column Filter node to entirely remove column OnlineBackup to clean the data.

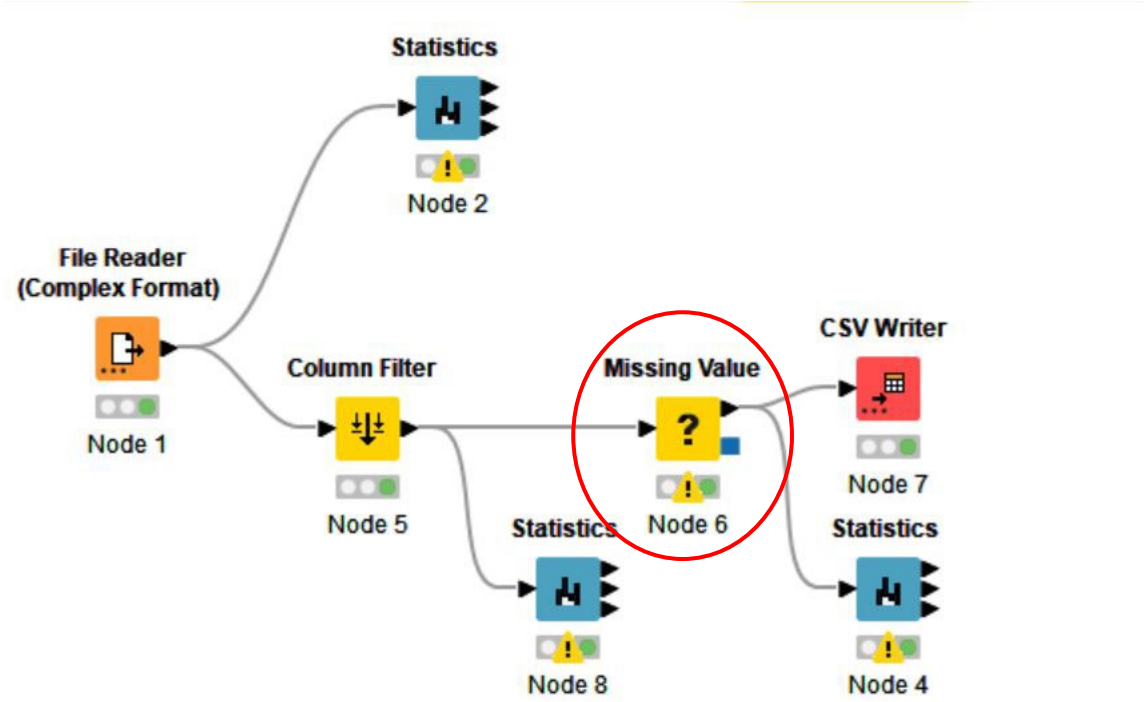


Rows: 2379 | Spec - Columns: 19

After removal, total column is now 19, however it is still not clean as there are still missing values of "tenure" and "gender" when viewing statistics view.



Since there are still missing values, using the Missing Value node will remove any rows of missing values to make the data more clean



After using the Missing Value node removing appropriate rows, data table would be cleaned as shown in image below

Numeric | Nominal | Top/bottom

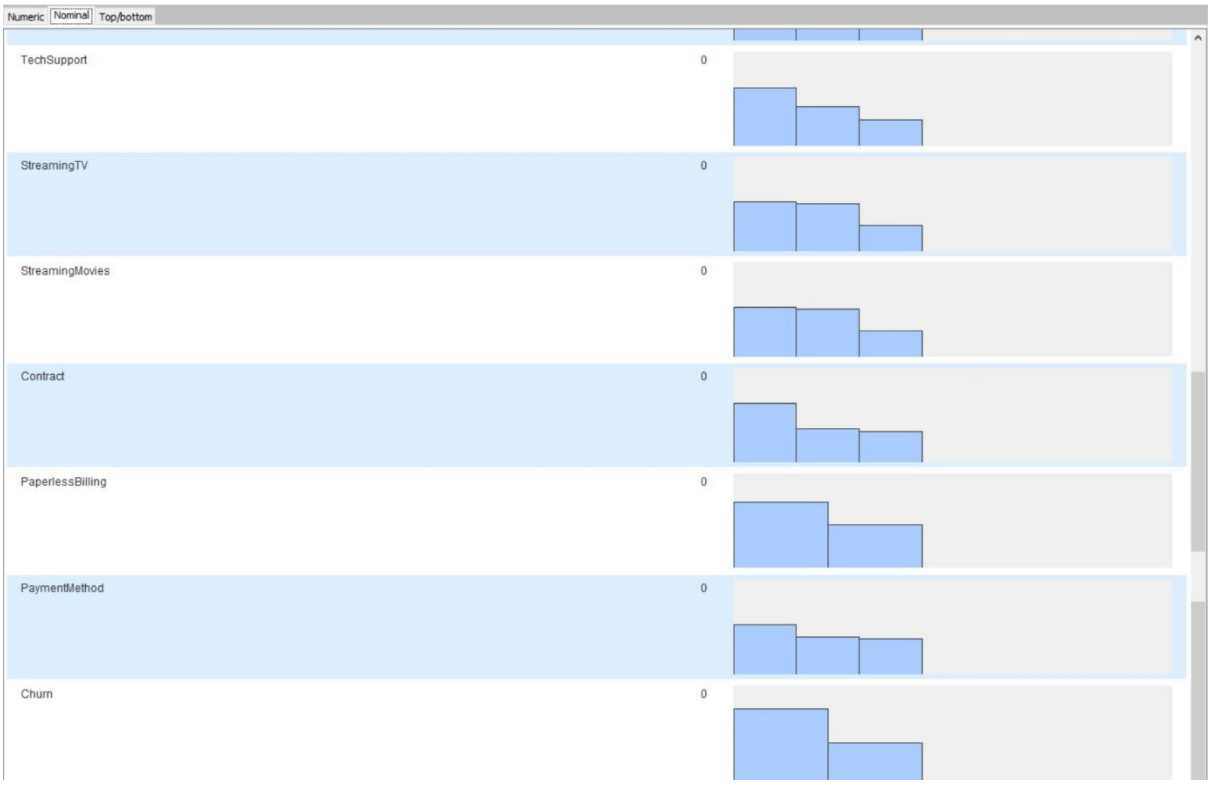
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
SeniorCitizen	0.0	0.2003	?	1	0.4003	1.4983	0.2451	0	0	0	
tenure	1	35.3622	?	72	24.5543	0.0728	-1.4431	0	0	0	
MonthlyCharges	18.55	70.216	?	118.75	28.3322	-0.463	-0.9288	0	0	0	
TotalCharges	19.25	2,612.1666	?	8,672.45	2,311.7513	0.7538	-0.5975	0	0	0	

Numeric | Nominal | Top/bottom

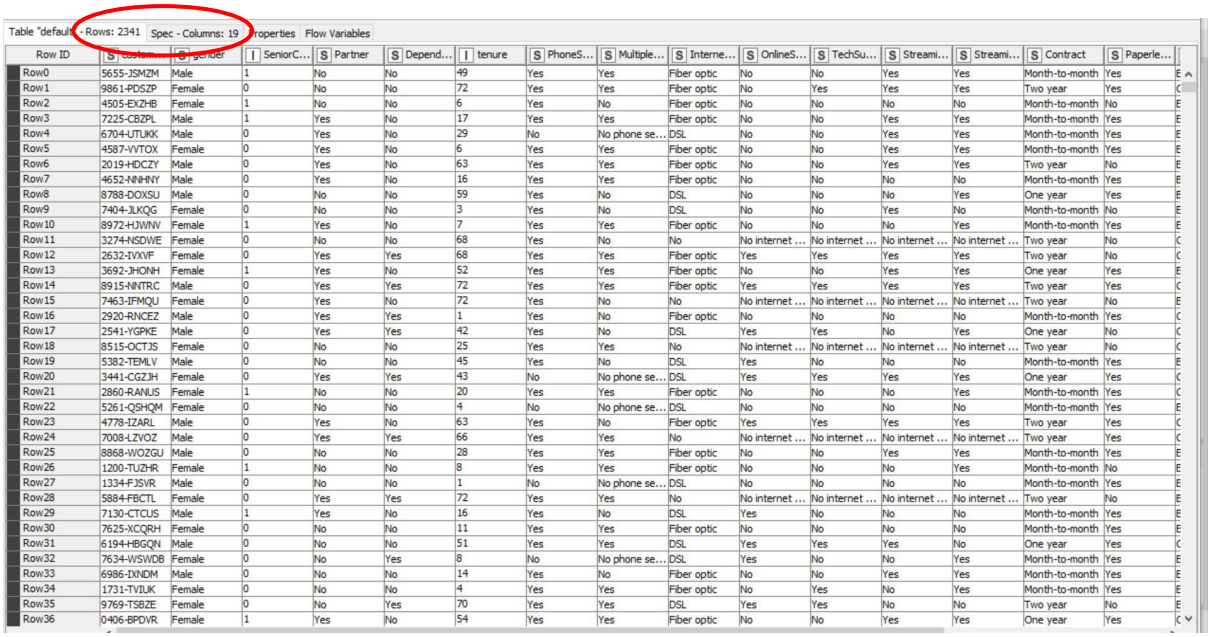
Column	No. missings	Histogram
customerID	0	
gender	0	
SeniorCitizen	0	
Partner	0	
Dependents	0	

Numeric | Nominal | Top/bottom

tenure	0	
PhoneService	0	
MultipleLines	0	
InternetService	0	
OnlineSecurity	0	

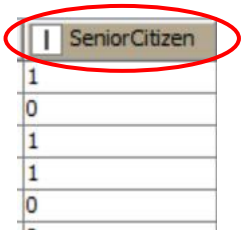


*Dataset after cleaning would have row of **2341** and **19** columns.*

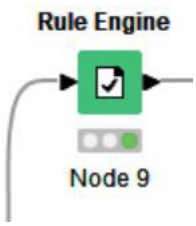


Row ID	customerID	gender	SeniorC...	Partner	Depend...	tenure	PhoneS...	Multiple...	Interne...	OnlineS...	TechSu...	Streami...	Streami...	Contract	Paperle...	E
Row0	5655-JSMZM	Male	1	No	No	49	Yes	Yes	Fiber optic	No	No	Yes	Yes	Month-to-month	Yes	E
Row1	9861-PDSZP	Female	0	No	No	72	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Two year	Yes	C
Row2	4505-EXZHB	Female	1	No	No	6	Yes	No	Fiber optic	No	No	No	No	Month-to-month	No	E
Row3	7225-CBZPL	Male	1	Yes	No	17	Yes	Yes	Fiber optic	No	No	Yes	Yes	Month-to-month	Yes	E
Row4	6704-UTUKK	Male	0	Yes	No	29	No	No phone se...	DSL	No	No	Yes	Yes	Month-to-month	Yes	E
Row5	4587-VVTOX	Female	0	Yes	No	6	Yes	Yes	Fiber optic	No	No	Yes	Yes	Month-to-month	Yes	E
Row6	2019-HDCZY	Male	0	Yes	No	63	Yes	Yes	Fiber optic	No	No	Yes	Yes	Two year	No	E
Row7	4652-NNHNY	Male	0	Yes	No	16	Yes	Yes	Fiber optic	No	No	No	No	Month-to-month	Yes	E
Row8	8788-DOXSU	Male	0	No	No	59	Yes	No	DSL	No	No	No	Yes	One year	Yes	E
Row9	7404-JLKQG	Female	0	No	No	3	Yes	No	DSL	No	No	Yes	No	Month-to-month	No	E
Row10	8972-HJWNV	Female	1	Yes	No	7	Yes	Yes	Fiber optic	No	No	No	Yes	Month-to-month	Yes	E
Row11	3274-NSDWE	Female	0	No	No	68	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	Two year	No	C
Row12	2632-IVXVF	Female	0	Yes	Yes	68	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Two year	No	C
Row13	3692-JHONH	Female	1	Yes	No	52	Yes	Yes	Fiber optic	No	No	Yes	Yes	One year	Yes	E
Row14	8915-NNTRC	Male	0	Yes	Yes	72	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Two year	Yes	C
Row15	7463-IFMQU	Female	0	Yes	No	72	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	Two year	No	E
Row16	2920-RNCEZ	Male	0	Yes	Yes	1	Yes	No	Fiber optic	No	No	No	No	Month-to-month	Yes	C
Row17	2541-YGPKE	Male	0	Yes	Yes	42	Yes	No	DSL	Yes	Yes	No	Yes	One year	No	C
Row18	8515-OCTJS	Female	0	No	No	25	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	Two year	No	C
Row19	5382-TEMLV	Male	0	No	No	45	Yes	No	DSL	Yes	No	No	No	Month-to-month	Yes	E
Row20	3441-CGZJH	Female	0	Yes	Yes	43	No	No phone se...	DSL	Yes	Yes	Yes	Yes	One year	Yes	C
Row21	2860-RANUS	Female	1	No	No	20	Yes	Yes	Fiber optic	No	No	No	Yes	Month-to-month	Yes	C
Row22	5261-QSHQM	Female	0	No	No	4	No	No phone se...	DSL	No	No	No	No	Month-to-month	Yes	E
Row23	4778-IZARL	Male	0	Yes	No	63	Yes	No	Fiber optic	Yes	Yes	Yes	Yes	Two year	Yes	C
Row24	7008-LZVOZ	Male	0	Yes	Yes	66	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	Two year	Yes	C
Row25	8868-WOZGU	Male	0	No	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Month-to-month	Yes	E
Row26	1200-TUZHR	Female	1	No	No	8	Yes	Yes	Fiber optic	No	No	No	Yes	Month-to-month	No	E
Row27	1334-FJSVR	Male	0	No	No	1	No	No phone se...	DSL	No	No	No	No	Month-to-month	Yes	E
Row28	5884-FBCTL	Female	0	Yes	Yes	72	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	Two year	No	E
Row29	7130-CTCUS	Male	1	Yes	No	16	Yes	No	DSL	Yes	No	No	No	Month-to-month	Yes	E
Row30	7625-XCQRH	Female	0	No	No	11	Yes	Yes	Fiber optic	No	No	No	No	Month-to-month	Yes	E
Row31	6194-HBGQN	Male	0	No	No	51	Yes	Yes	DSL	Yes	Yes	Yes	No	One year	Yes	C
Row32	7634-WSWDB	Female	0	No	Yes	8	No	No phone se...	DSL	Yes	No	No	Yes	Month-to-month	Yes	E
Row33	6986-IXNDM	Male	0	No	No	14	Yes	No	Fiber optic	No	No	Yes	Yes	Month-to-month	Yes	E
Row34	1731-TVIUK	Female	0	No	No	4	Yes	Yes	Fiber optic	No	Yes	No	Yes	Month-to-month	Yes	E
Row35	9769-TSBZE	Female	0	No	Yes	70	Yes	Yes	DSL	Yes	Yes	No	No	Two year	No	E
Row36	0406-BPDVR	Female	1	Yes	No	54	Yes	Yes	Fiber optic	No	No	Yes	Yes	One year	Yes	C

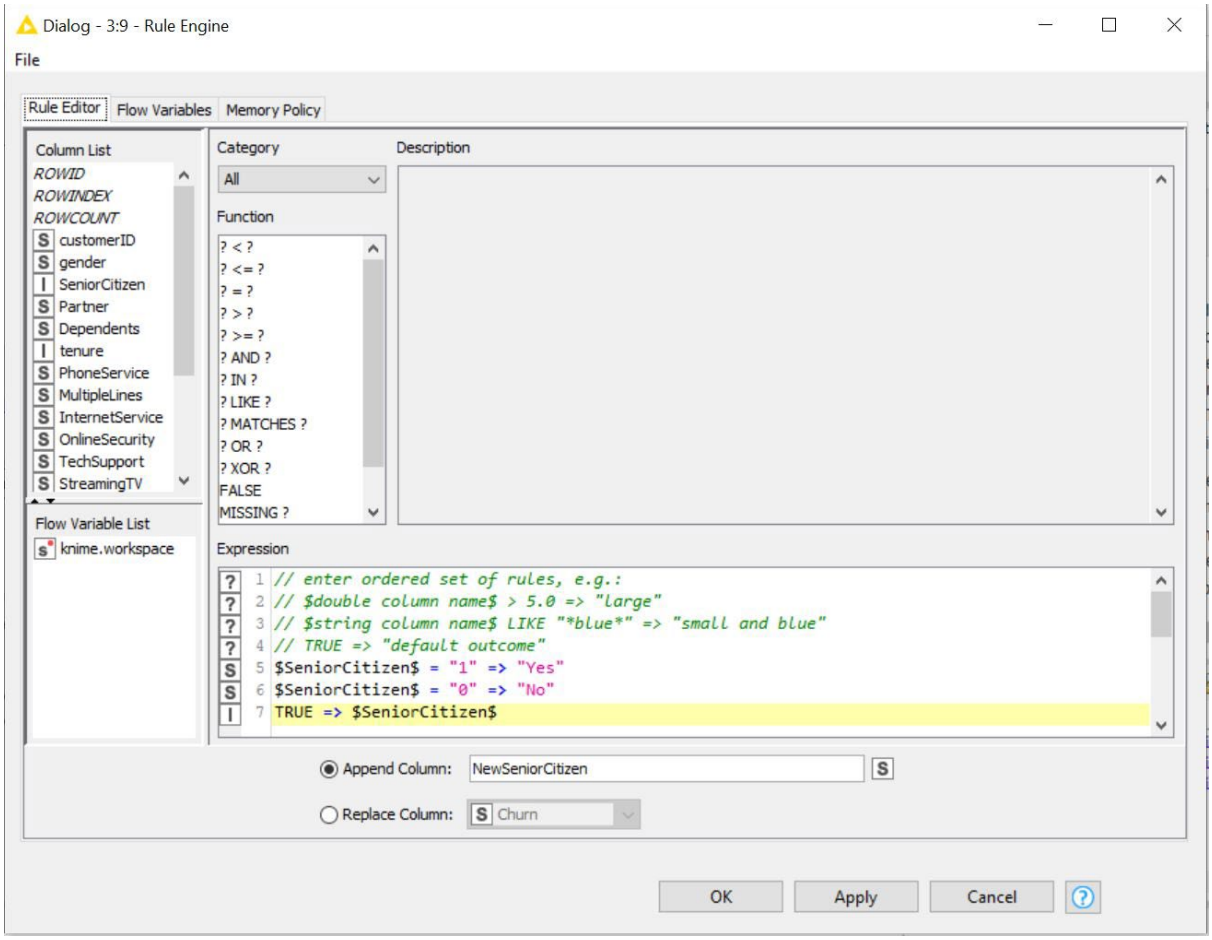
***Data Transformation** might be needed but not necessary. It is to make the dataset much more consistent across variables, as variable SeniorCitizen in the dataset is an integer.*



SeniorCitizen
1
0
1
1
0

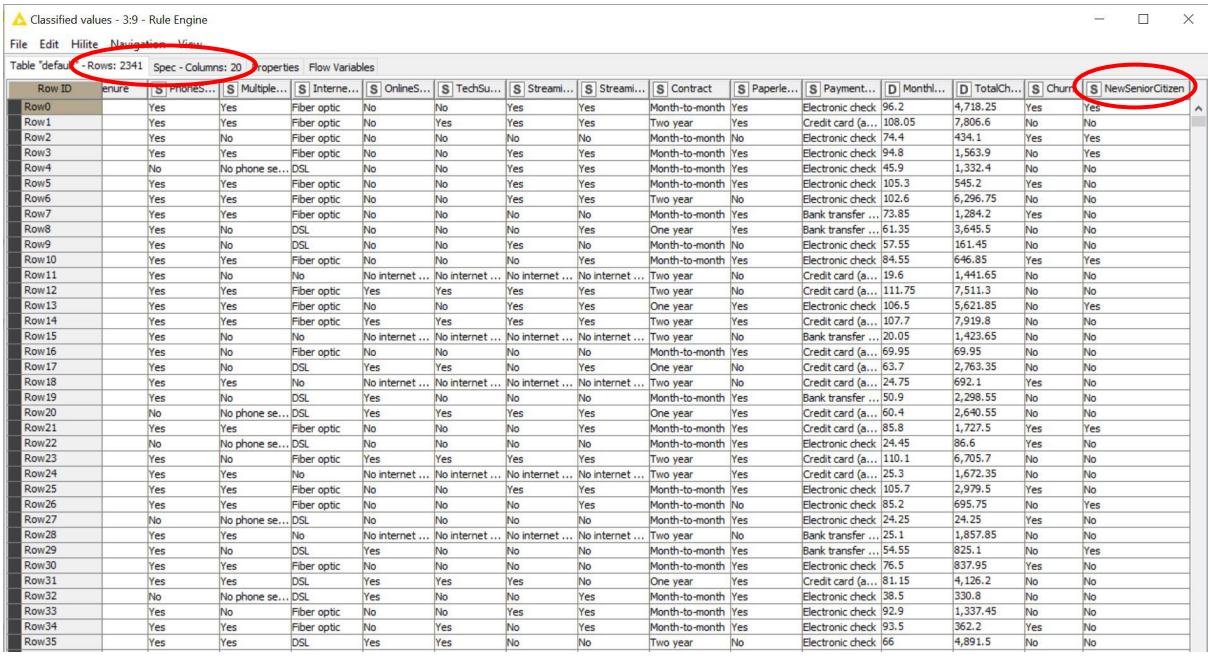


By using the Rule Engine Node, we can start the data transformation.



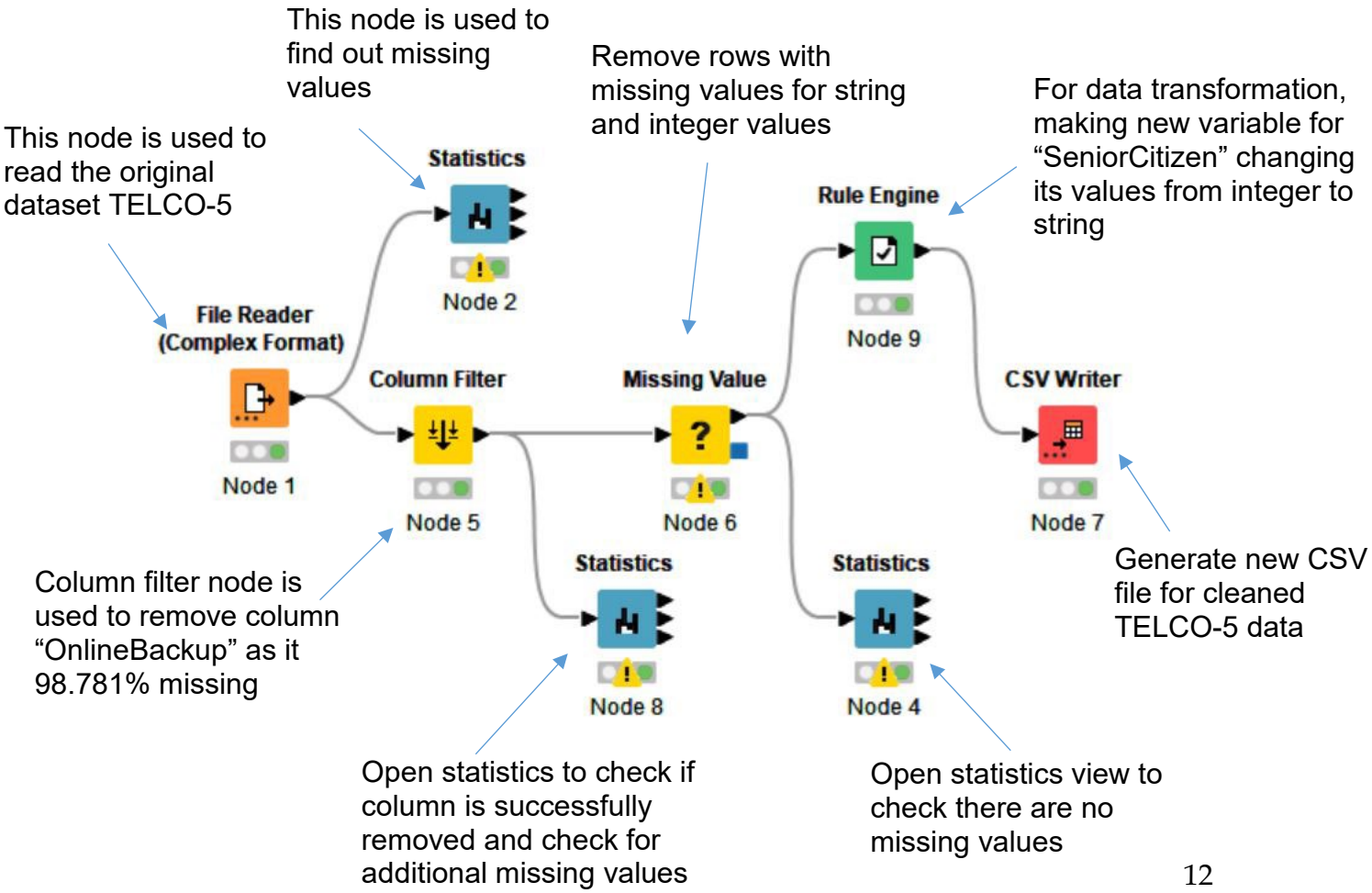
With data transformation, through the use of the Rule Engine Node we can make SeniorCitizen a string by making “1” => “Yes” and “0” => “No”.

After executing the Rule Engine Node, the classified values are as shown with a new column added called, “NewSeniorCitizen”.



*Dataset after cleaning and transformation should now have **2341 Rows** and **20 Columns**, keeping the old "SeniorCitizen" variable as reference for "NewSeniorCitizen".*

The Workflow:



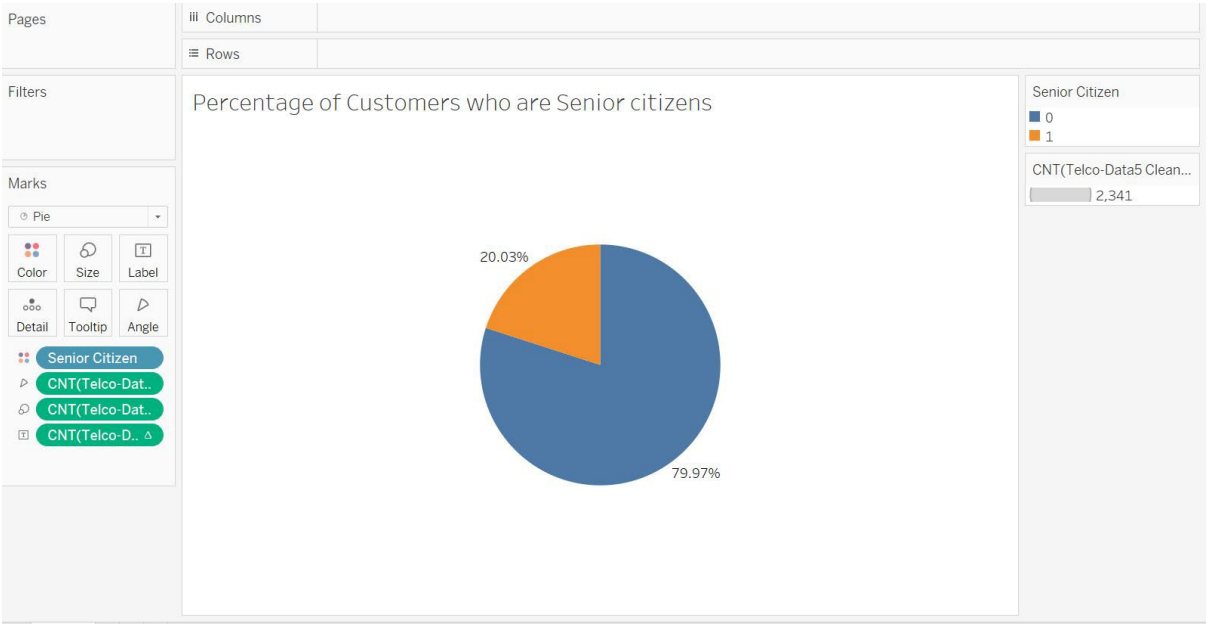
6. Exploratory Data Analysis

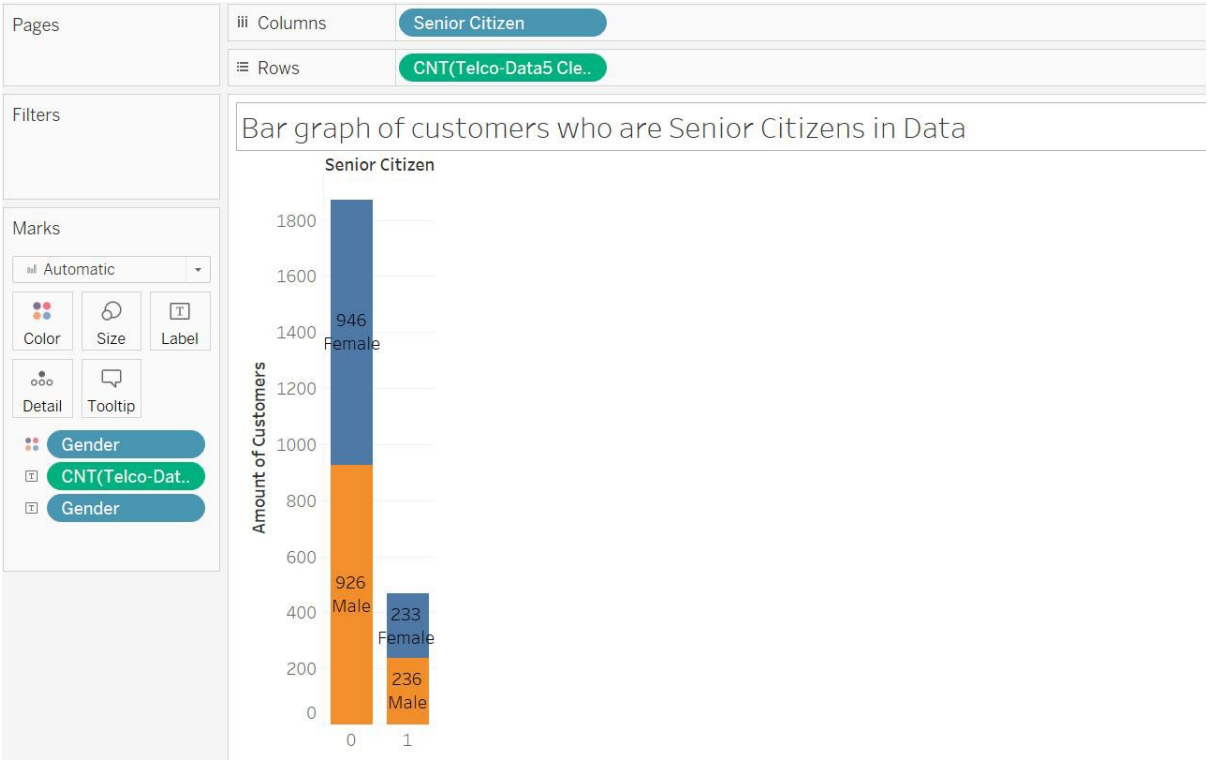
Variable	Min	Max	Mean	Standard deviation	Median
SeniorCitizen	0.0	1	0.2003	0.4003	0.0
tenure	1	72	35.3622	24.5543	35
MonthlyCharges	18.55	118.75	70.216	28.332	75.75
TotalCharges	19.25	8,672.45	2,612.1666	2,311.7513	1,848.8

- SeniorCitizen

For variable SeniorCitizen, “0” denotes No, whereas “1” denotes Yes. Only 469 out of 2341 customers are senior citizens, which makes 20.03% of customers who are senior citizens. The gender of customers are almost equal for both senior citizens and non-senior citizens.

The data shows that the majority of customers are non-senior citizens below 65 years of age.



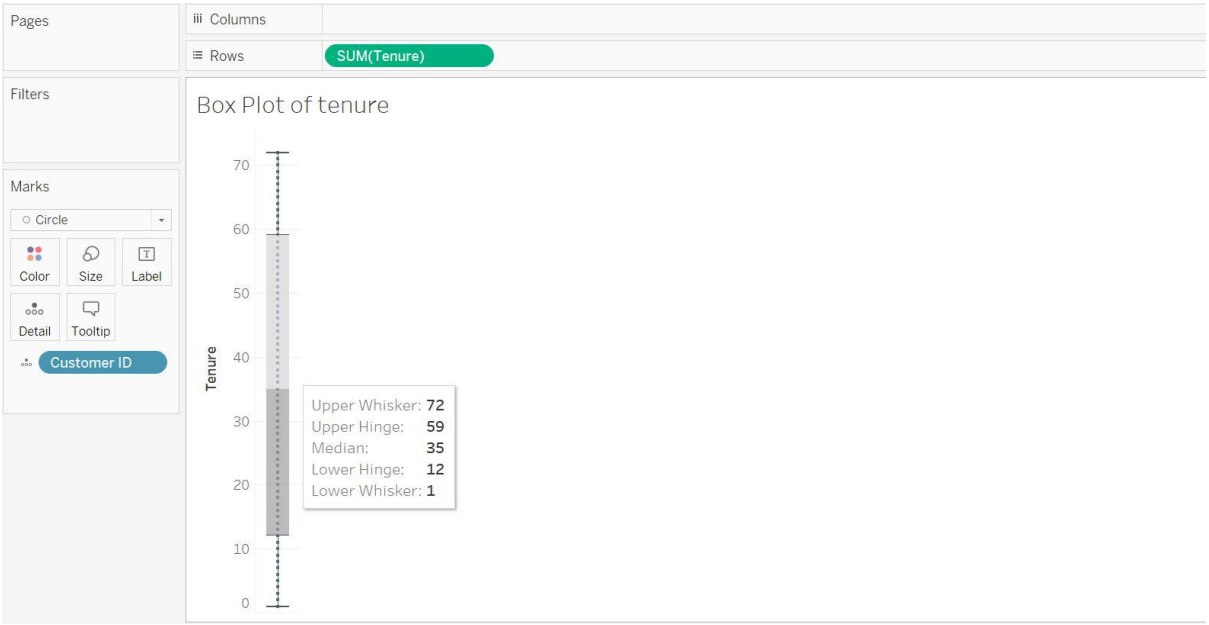


For Senior citizens, the majority of them are independent with only 27 of senior citizens being dependent. For senior citizens who are independent, 223 of them have partners while 219 of them do not. For senior citizens who are dependent, all of them have partners.

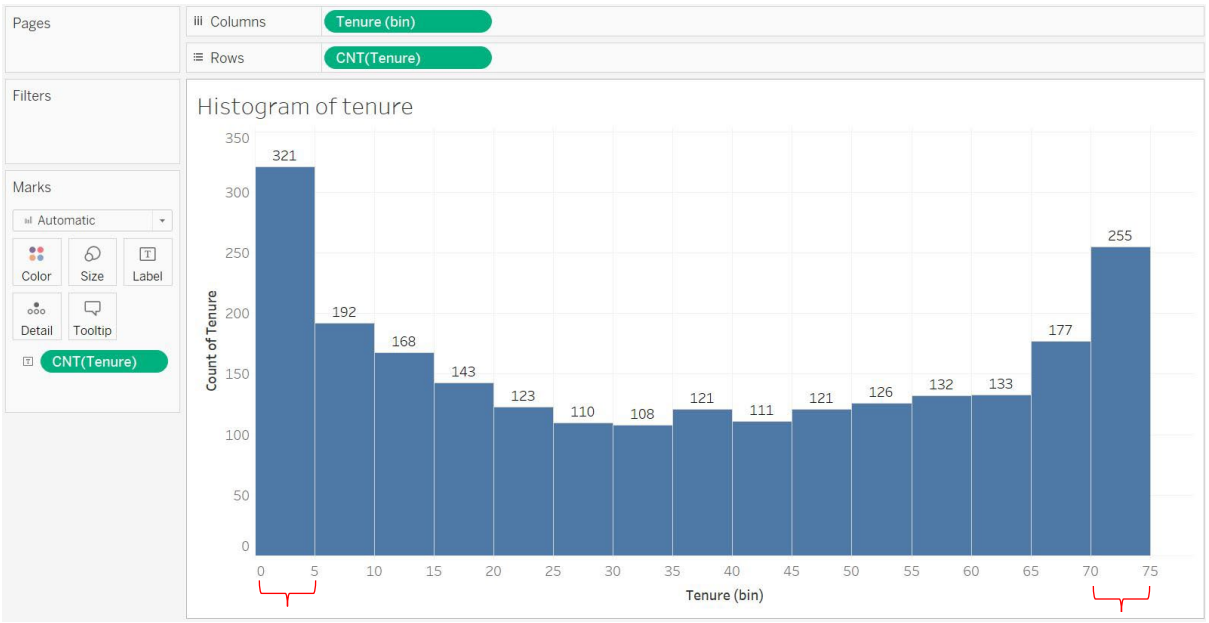
- **tenure**

The tenure ranges from 1 to 72. The Maximum of tenure is 72, while minimum is 1. The range of tenure is 71. *(72-1 = 71).*
The Upper Quartile of tenure is 59, while the Lower Quartile is 12. The Inter-quartile range of tenure is 47. *(59 - 12 = 47).*

The tenure has a median value of 35. 50% of tenure is less than 35 and higher than 35. It has a roughly **uniform** spread.

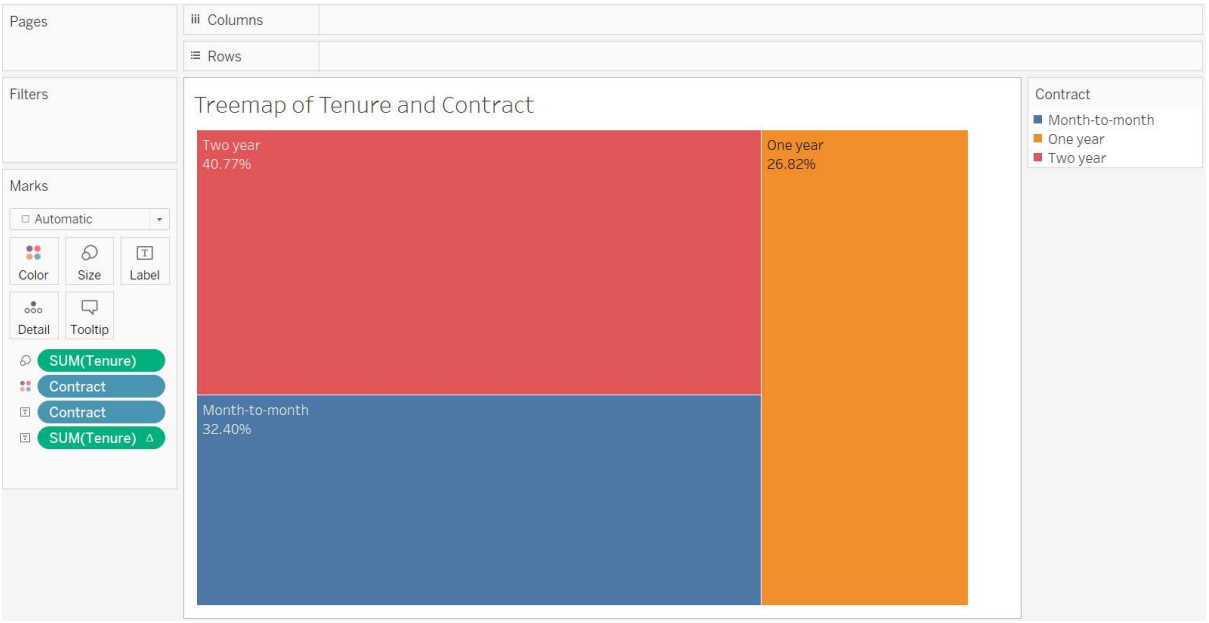


The histogram of tenure shows that the mode is between 0 to 5 as it has the highest occurrence rate at a count of 321 *(13.71% count of tenure).* Tenure > 5 and < 70 are at a lower number. The lowest tenure occurrence is between 30 to 35 at 108 *(4.61% count of tenure).*



Insight: This could mean that many are just starting their contracts or likely churns their contract after 5 months of tenure.

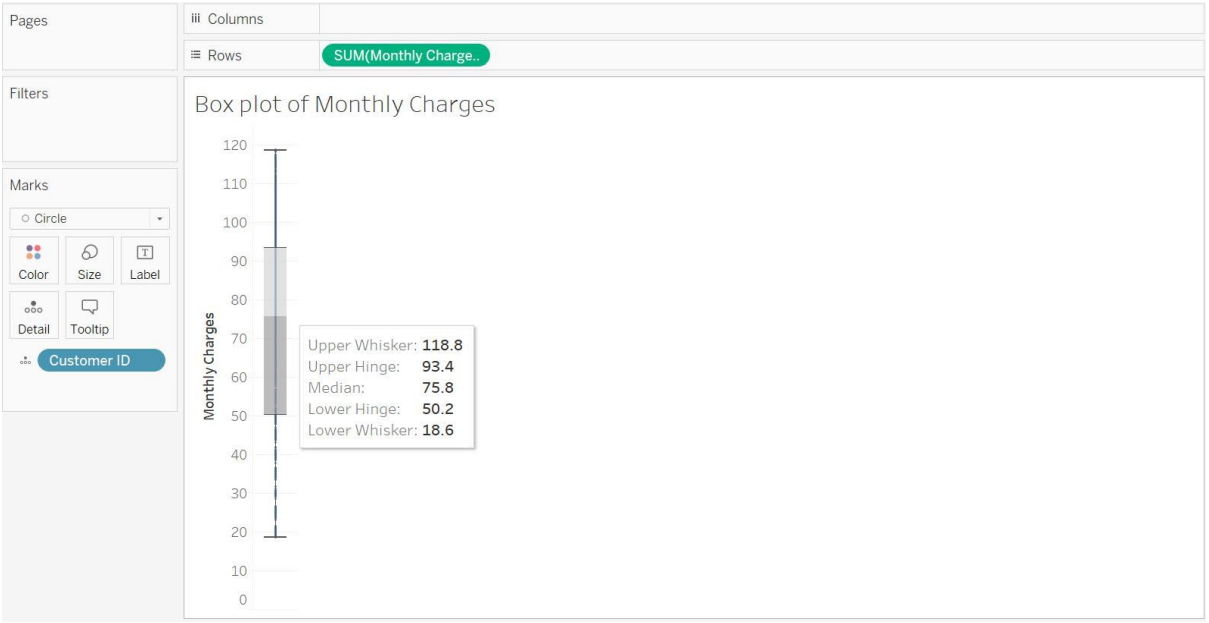
In addition, the treemap below shows that for variable tenure, customers are more likely to stay within a two-year contract (at 40.77%) and less likely to stay within a one-year contract (at 26.82%). This could be because it is easier to break a contract with a month-to-month or one-year contract.



- **MonthlyCharges**

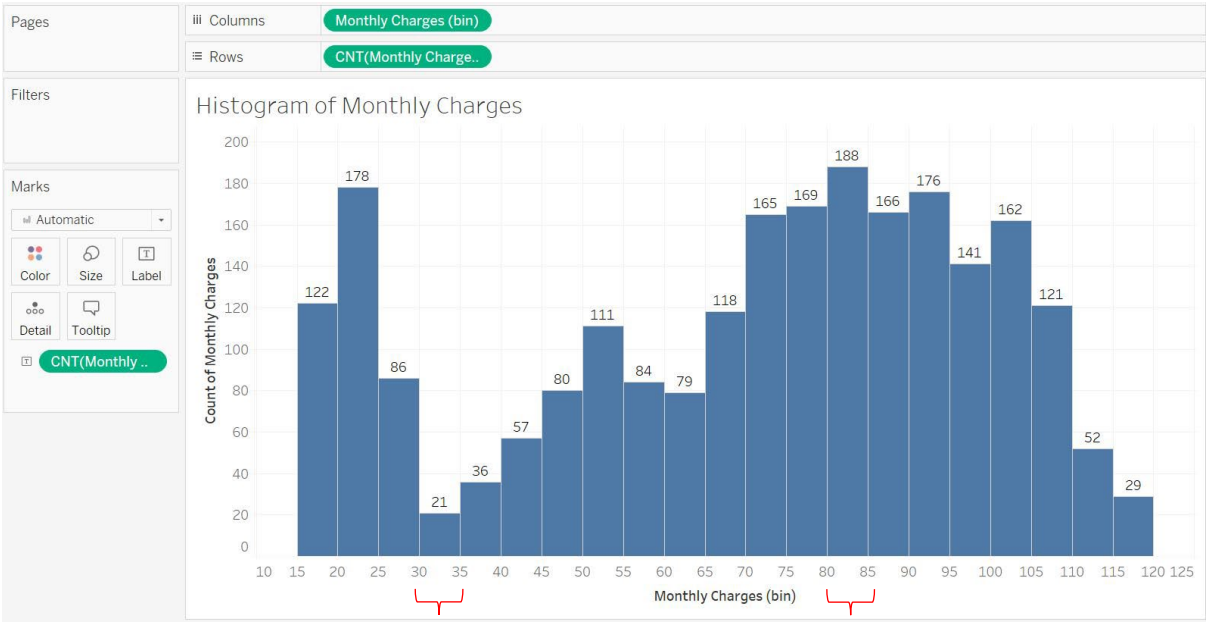
The monthly charges range from 18.6 to 118.8. The Maximum being 118.8, and Minimum of 18.6, with the range being 100.2. *(118.8 – 18.6 = 100.2).*
The Upper Quartile range of monthly charges is 93.4, while the Lower Quartile is 50.2. The interquartile range of monthly charges is 43.2. *(93.4 – 50.2 = 43.2).*

The monthly charges have a median value of 75.8. 50% of the monthly charges are less than 75.8, however, the spread is **stronger** in the lower quartile.



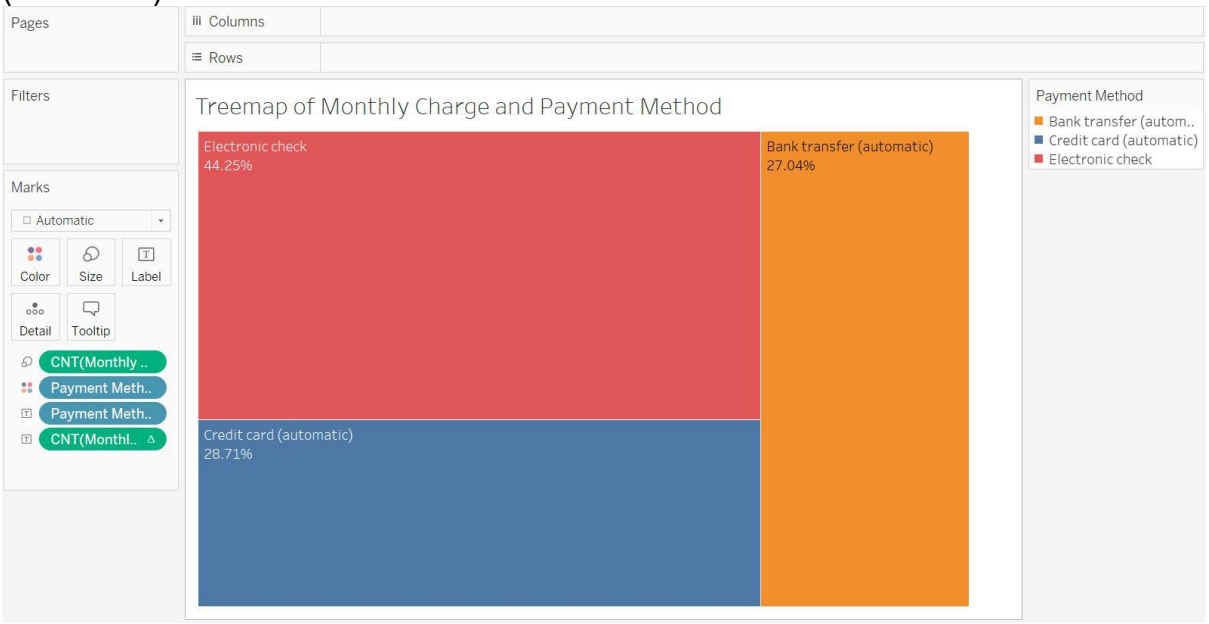
In the Histograms shown below, the mode of monthly charges is between 80 to 85 as it has the highest occurrence rate at a count of 188 *(8.03% count of*

monthly charges). Lowest occurrence being 30 to 35 at a count of 21 *(0.9% count of monthly charges).*



Insight: From the Histogram, on average many customers prefer paying between $65 to $105 a month on a Telco Company. However, for a more affordable option, many customers prefer spending $15 to $25 a month on a Telco Company.

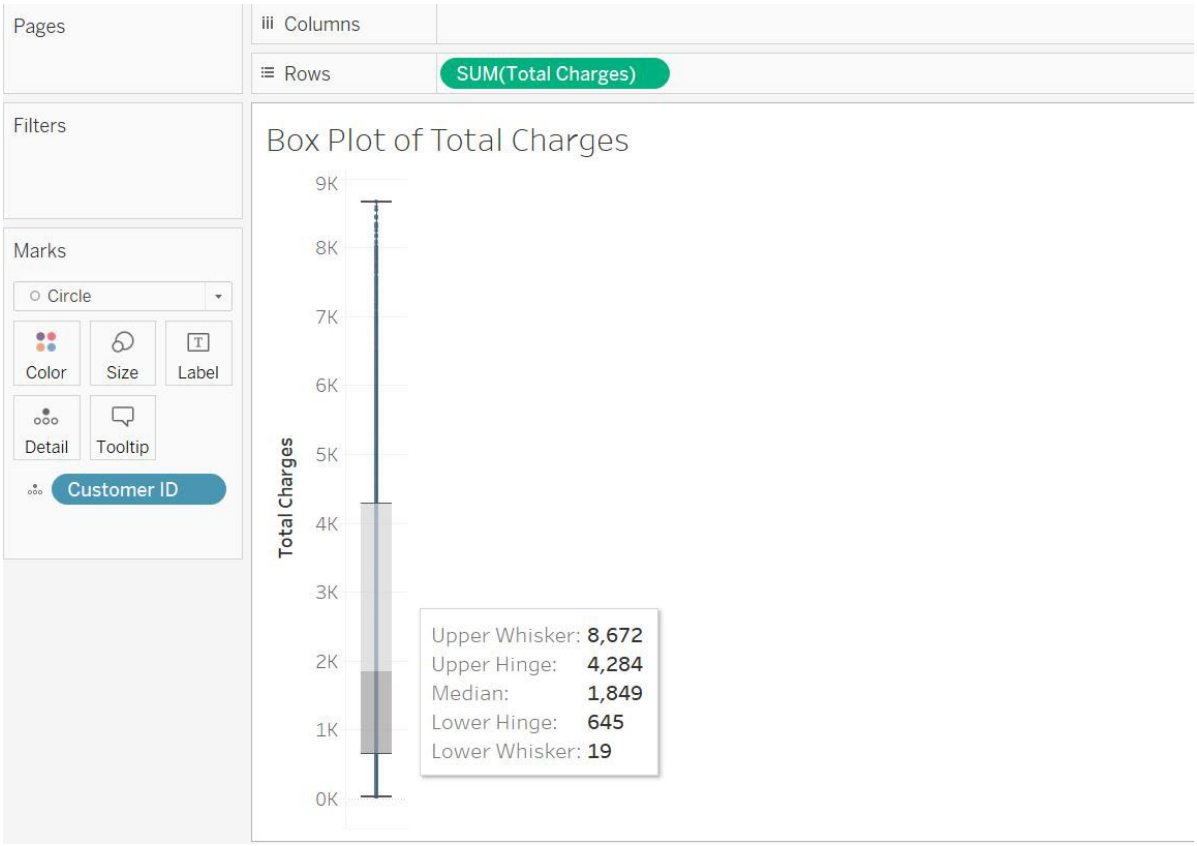
In addition, from the treemap below, the most popular payment method is by Electronic check while the least popular is by Bank Transfer (automatic). 44.25% of the monthly charges are paid by an Electronic Check, 28.71% paid by Credit card (automatic) and lastly, 27.04% of monthly charges is paid by Bank transfer (automatic).



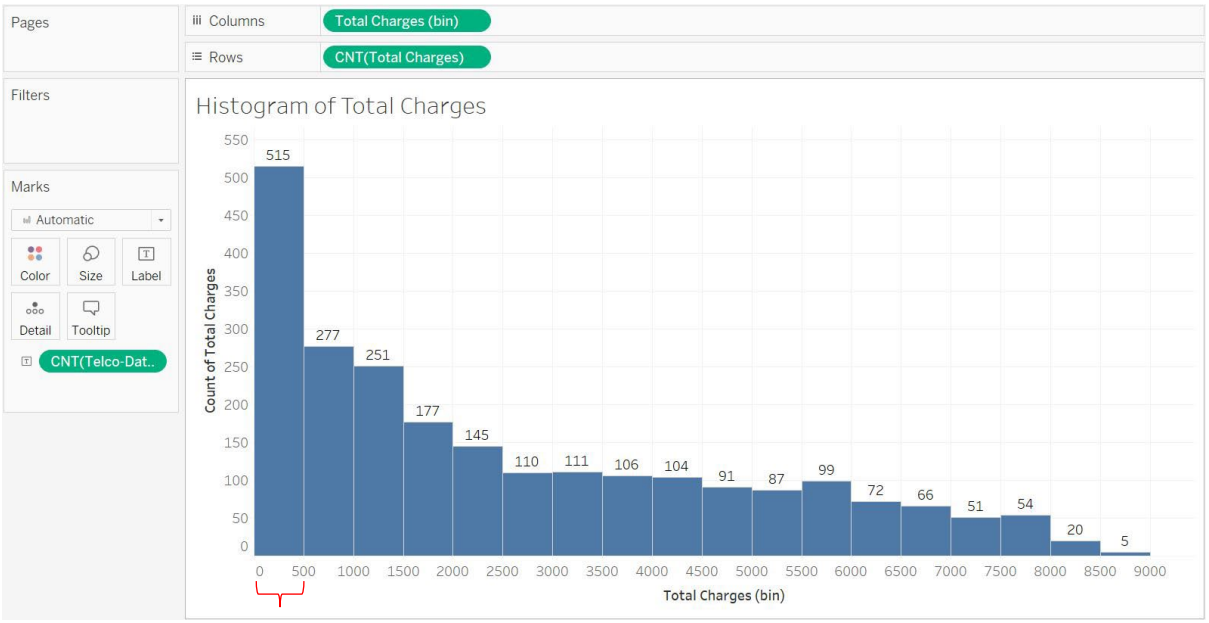
- **TotalCharges**

The total charges ranges from 19 to 8,672. The Maximum being 8,672, and Minimum of 19, with the range being 8,653. *(8,672 – 19 = 8,653)*.
The Upper Quartile range of total charges is 4,284, while the Lower Quartile is 645. The interquartile range of total charges is 3,639. *(4,284 – 645 = 3,639)*.

The total charges have a median value of 1,849. 50% of the total charges are less than 1,849, however, the spread is **stronger** in the Upper Quartile.



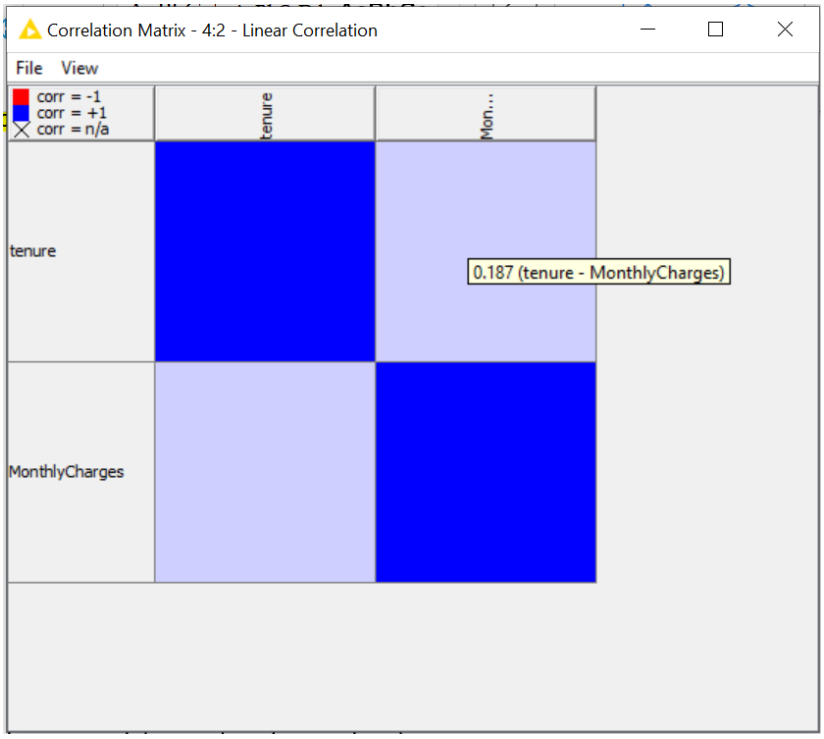
The histogram below shows that the mode of total charges is between 0 to 500 at a count of 515 (22% count of total charges). The lowest occurrence being 8500 to 9000 at a count of 5 (0.21% count of total charges).



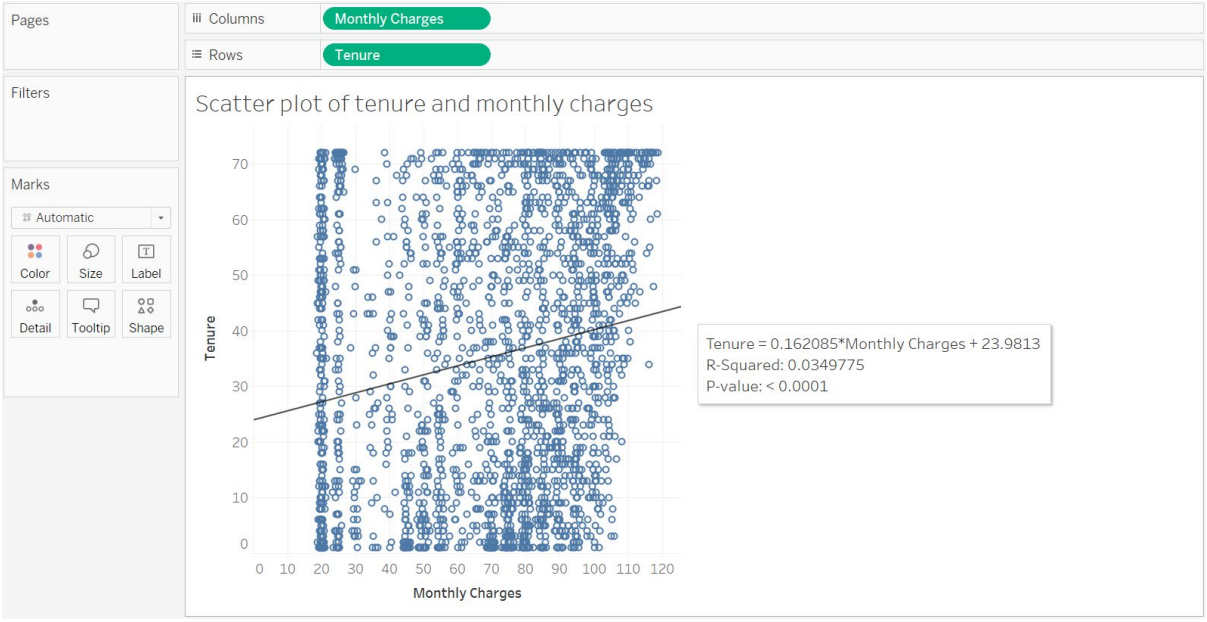
Insight: From the histogram, it can be seen that the graph of total charges is **skewed right.** This could mean that customers are less likely to pay a high amount of money over time.

7. Answers to preliminary questions

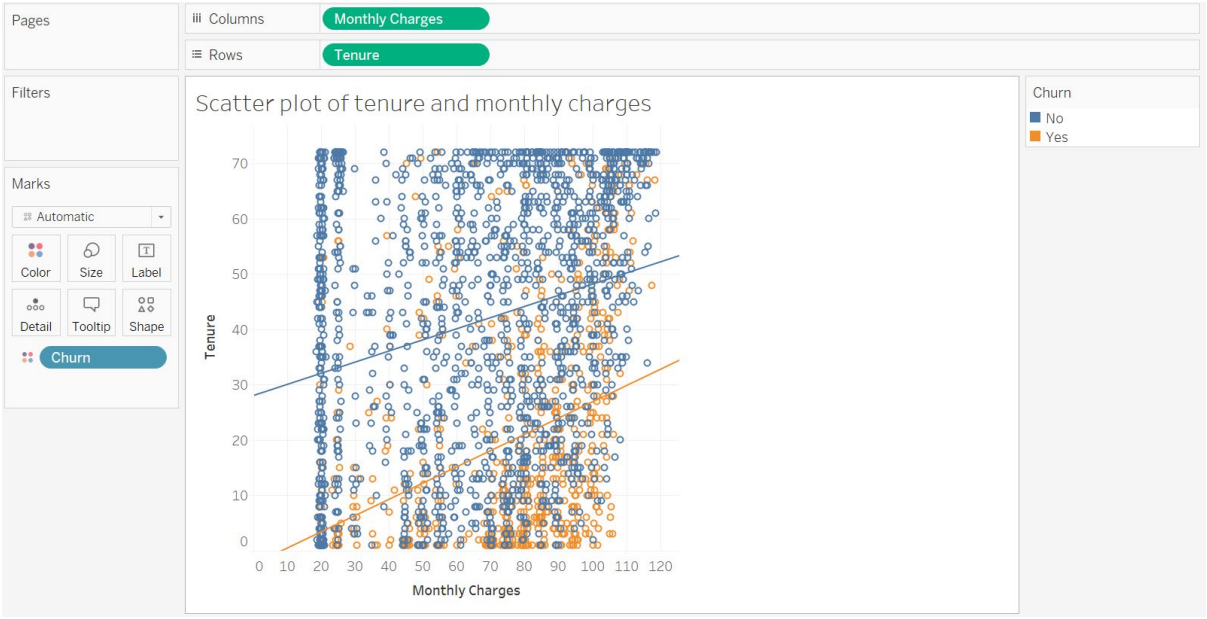
1. Is the tenure linearly correlated to the monthly charge?



From the linear correlation matrix above, the r-value is shown to be 0.187 which is more than 0, hence there is a weak positive linear correlation between tenure and MonthlyCharges. Therefore, when tenure increases, MonthlyCharges also increases.

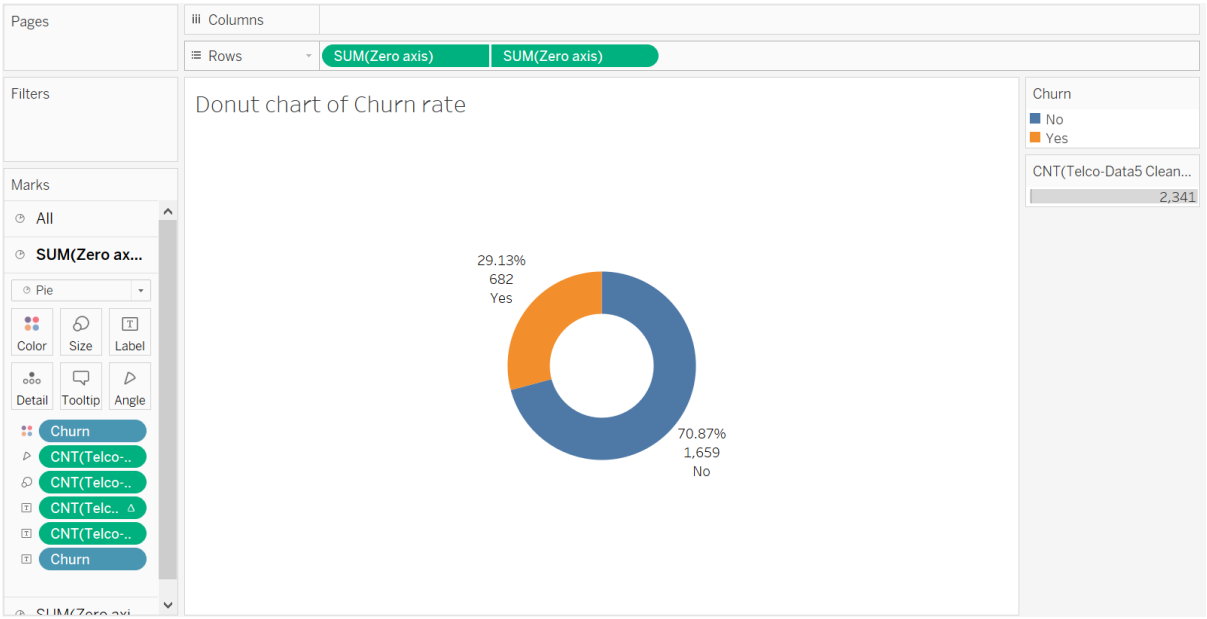


For further proof of linear correlation, the scatter plot above shows the trend line which denotes a weak positive linear correlation with r value of sqrt(0.0349775) = 0.187.



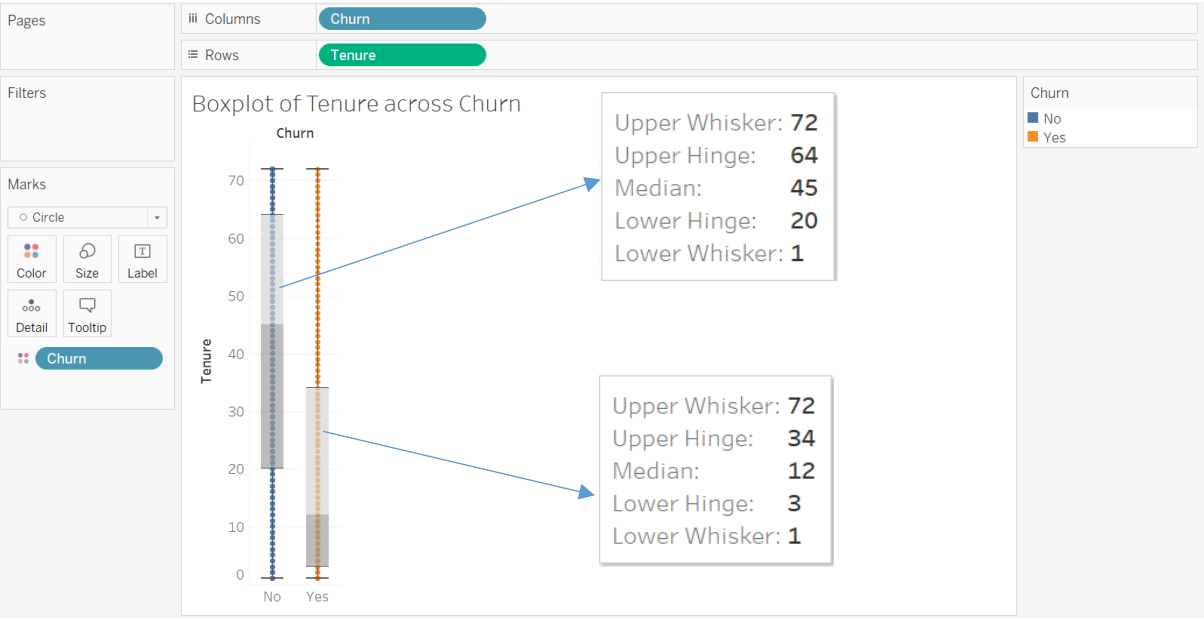
In addition, from the scatter plot above, we can see that churn is directly proportionate to the tenure.

2. What is the composition of churn in this dataset?

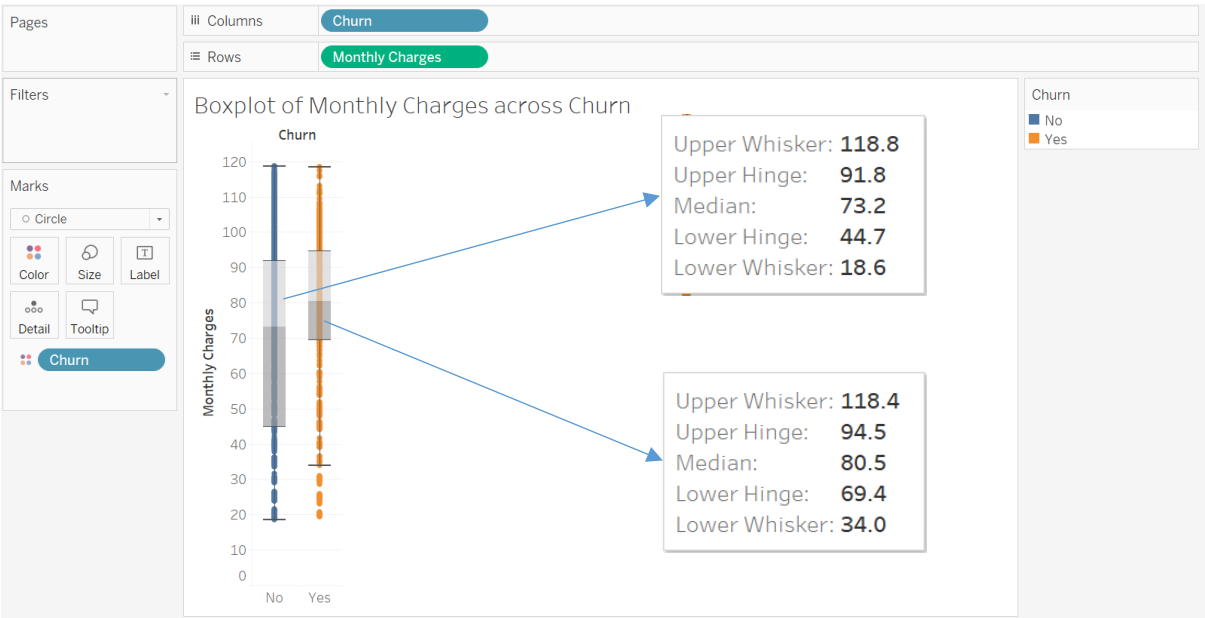


From the donut chart above, we can see the composition of churn in this Telco-5 dataset. (682 of 2341) 29.13% of the customers chose to churn, while (1,659 of 2341) 70.87% of the customers chose to stay.

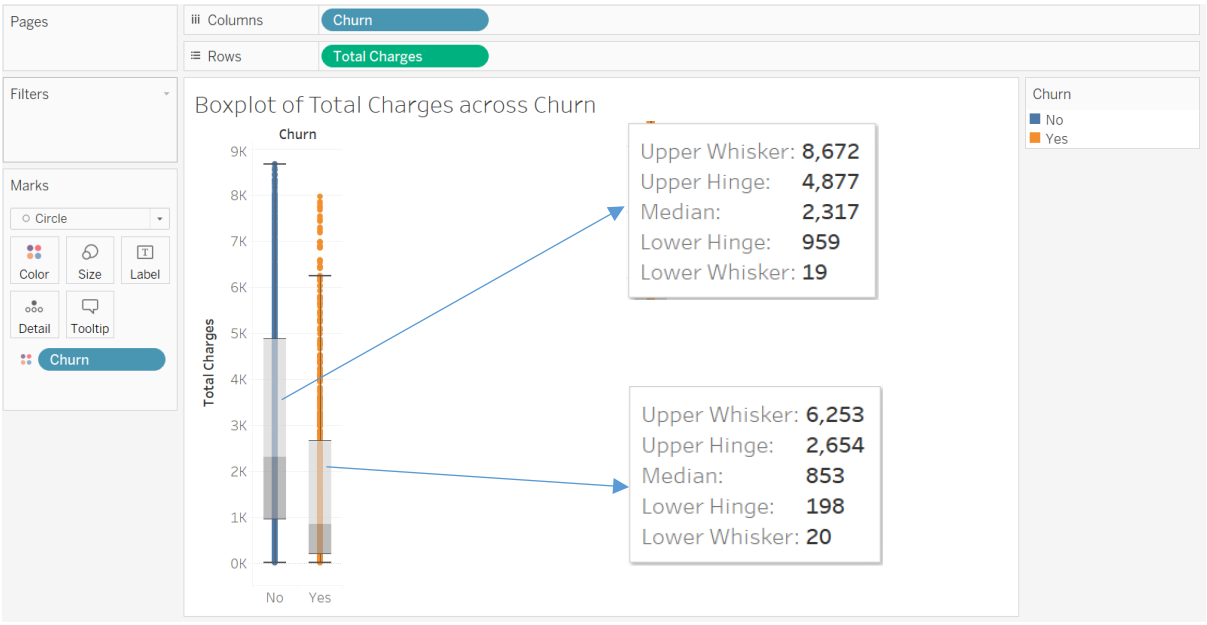
3. What do boxplots of numerical features against variable “Churn” in the Telco-5 dataset suggest?



For the boxplots of Tenure across Churn above, it suggests that customers with a longer tenure are less likely to churn.

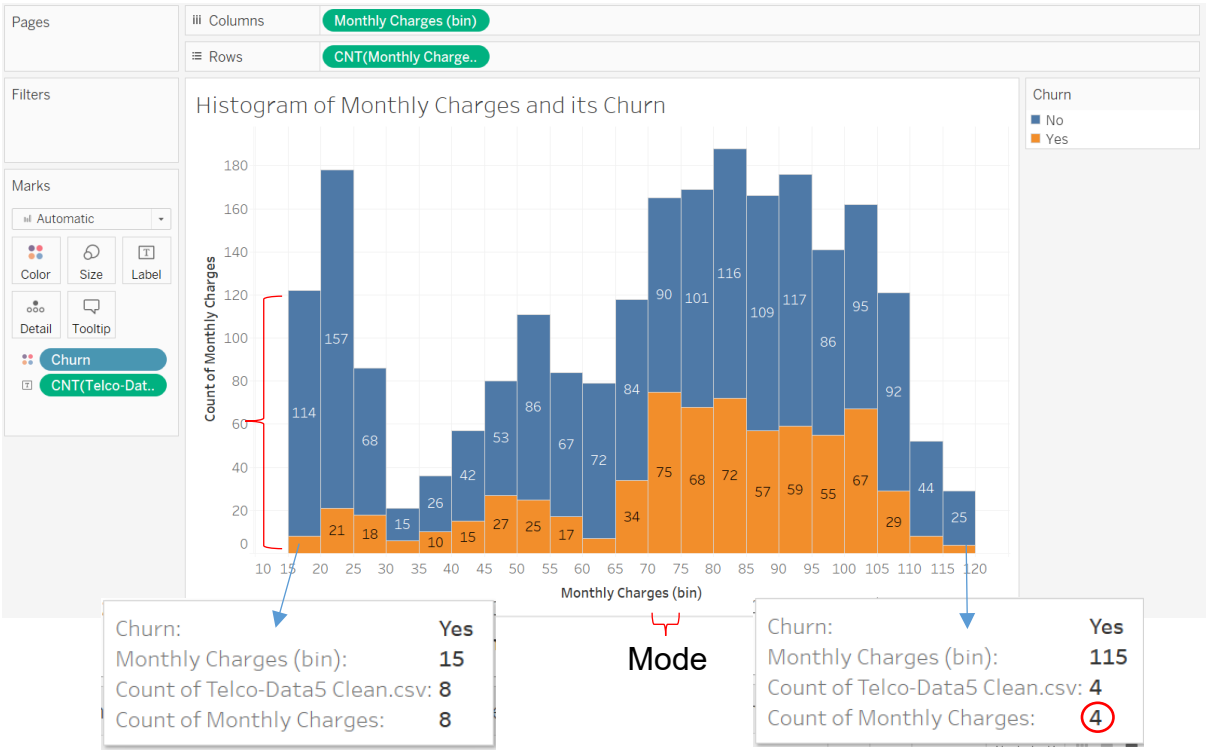


For the boxplots of Monthly Charges across Churn above, it suggests that customers with a lower monthly charge are less likely to churn. However, “lower monthly charge” is vague and data needs to be transformed to find out affordability, hence requires further analysis and insights in part 8.



For boxplots of Total Charges across Churn above, there are outlying data for customers who churn. Total Charges are likely correlated to Monthly Charges and Tenure.

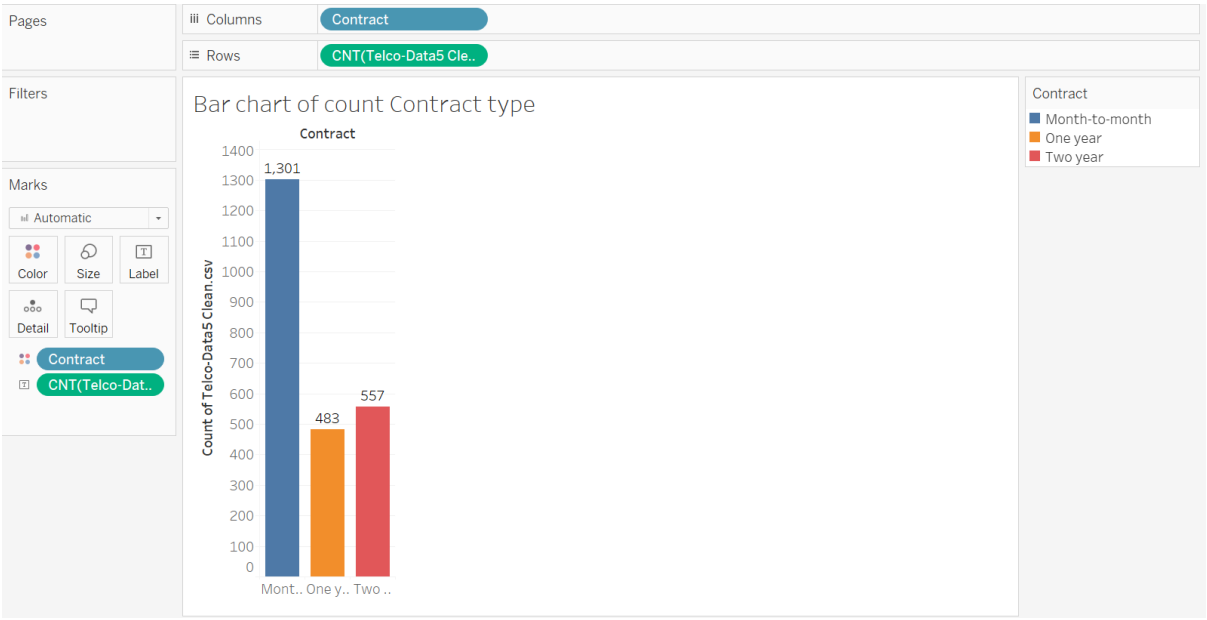
2. What is the mode of highest churn for monthly charges



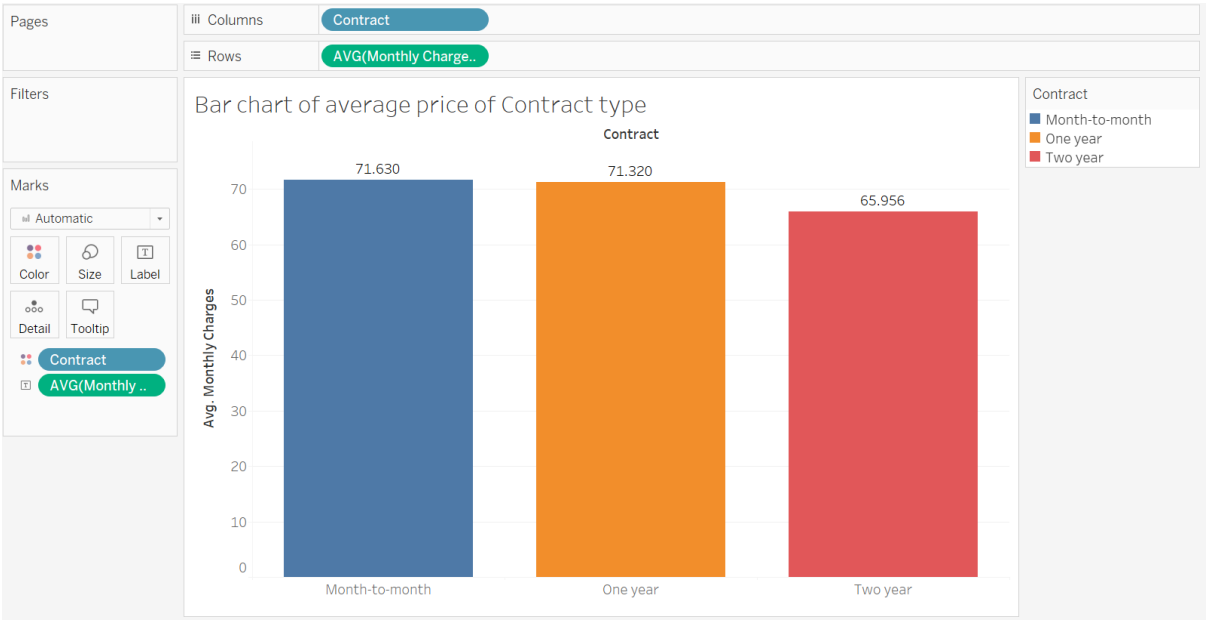
From the histogram above, the color orange denotes “Yes” for Churn, while the color blue denotes “No” for Churn. The **mode for the highest churn for monthly charges is 70 to 75**, 75 chose to churn while 90 did not. While the least amount of churn for monthly charges is between 115 to 120, where 4 chose to churn while 25 did not.

However, for churn RATE (percentage), the least amount of churn rate is between a monthly charge of 15 to 20, at a 6.56% churn rate (8/8+114 x 100). While the highest churn rate is still between monthly charges of 70 to 75, at a 45.45% churn rate (75/75+90 x 100).

3. What is the most common Contract length? In addition, on average, which contract length is cheaper monthly?



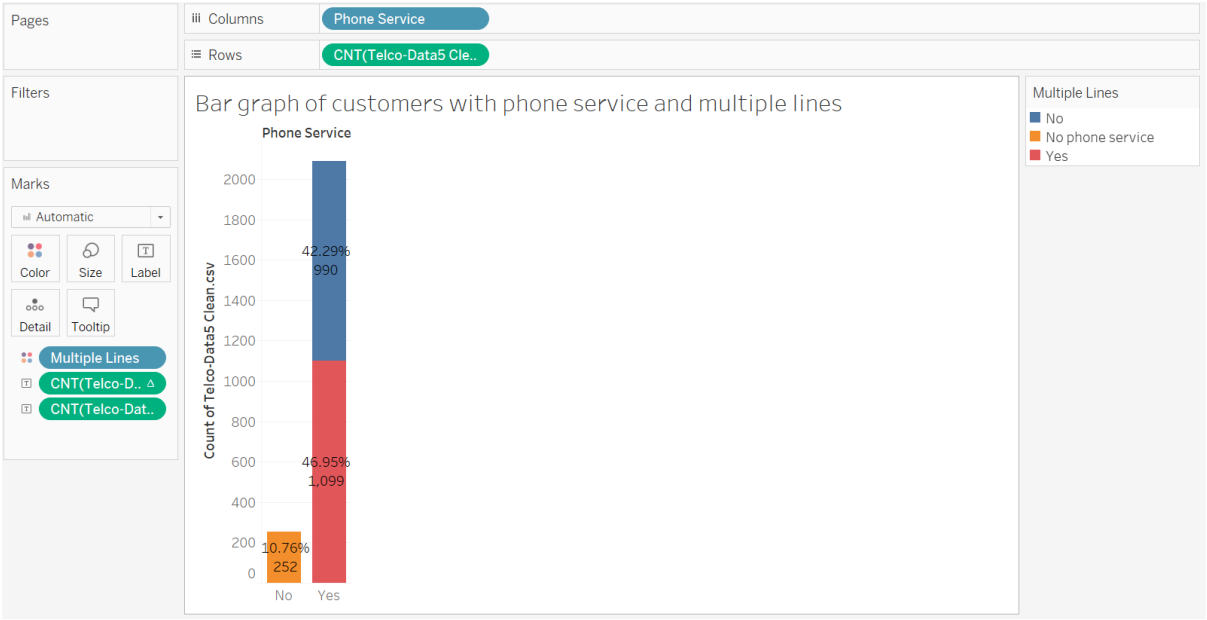
As seen from the bar chart above, the **most common Contract type is "Month-to-month" at 1,301 counts** (55.57% count of Contract), while the lowest Contract type is "One year" at 483 counts (20.63% count of Contract).



However, on average, the **Two year contract is cheaper monthly at 65.956** while the Month-to-month is more expensive on average at 71.630.

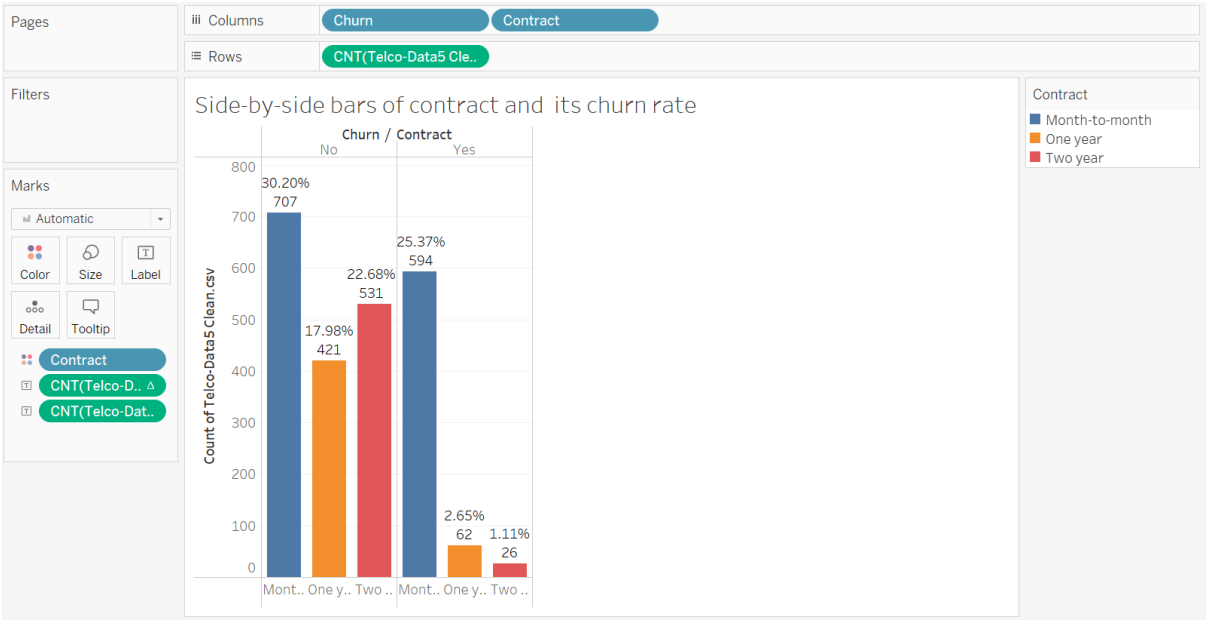
Insight: Most customers prefer opting for a Month-to-month contract as opposed to the other contract types, despite the higher price. As stated in a previous graph, this could be because of the terms of the contract and that a Month-to-month contract is easier to be broken than a One year or Two year contract.

4. For customers with phone service, how many have multiple lines.



From the bar graph above, 2,089 customers are with phone service while 252 do not. Moreover, of the 2,089 customers with phone service, 1,099 have multiple lines.

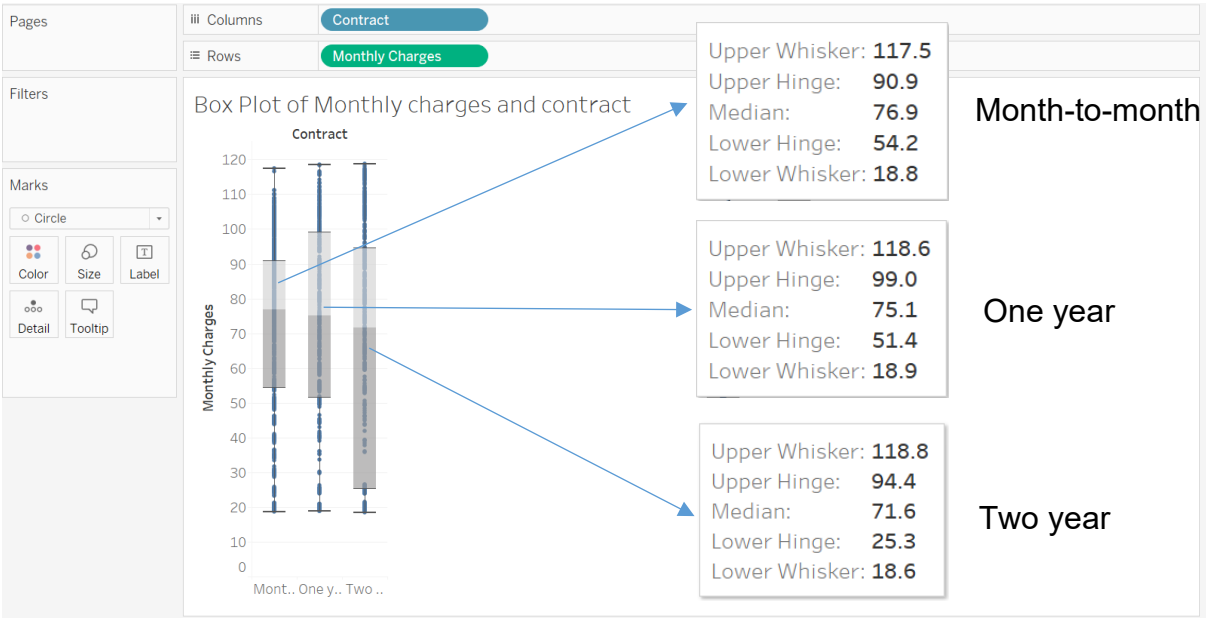
5. Which contract type are customers more likely to churn?



The graph above shows that the **Month-to-month contract is the most likely to be churned at 25.37%** whereas the least likely to be churned is the Two year contract at 1.11%.

Insight: The graph shows evidence that a Month-to-month contract is the most easily broken, at a pretty big percentage of 25.37%. It is likely customers chose a Month-to-month contract to try it out for the first month to see if they like it as it is easier to break the contract after the first month of trial.

6. What are the range of monthly charges for each contract length?



From the box plot above we can see that,

Month-to-month: The monthly charges for Month-to-month Contract range from **18.8 to 117.5** while the range is **98.7** *(117.5 – 18.8 = 98.7)*. The median monthly charge is **76.9**. The interquartile range of monthly charges is **36.7** *(90.9 – 54.2 = 36.7).*

One year: The monthly charges for One year Contract range from **18.9 to 118.6** while the range is **99.7** *(118.6 – 18.9 = 99.7).* The median monthly charge is **75.1**. The interquartile range of monthly charges is **47.6** *(99.0 – 51.4 = 47.6).*

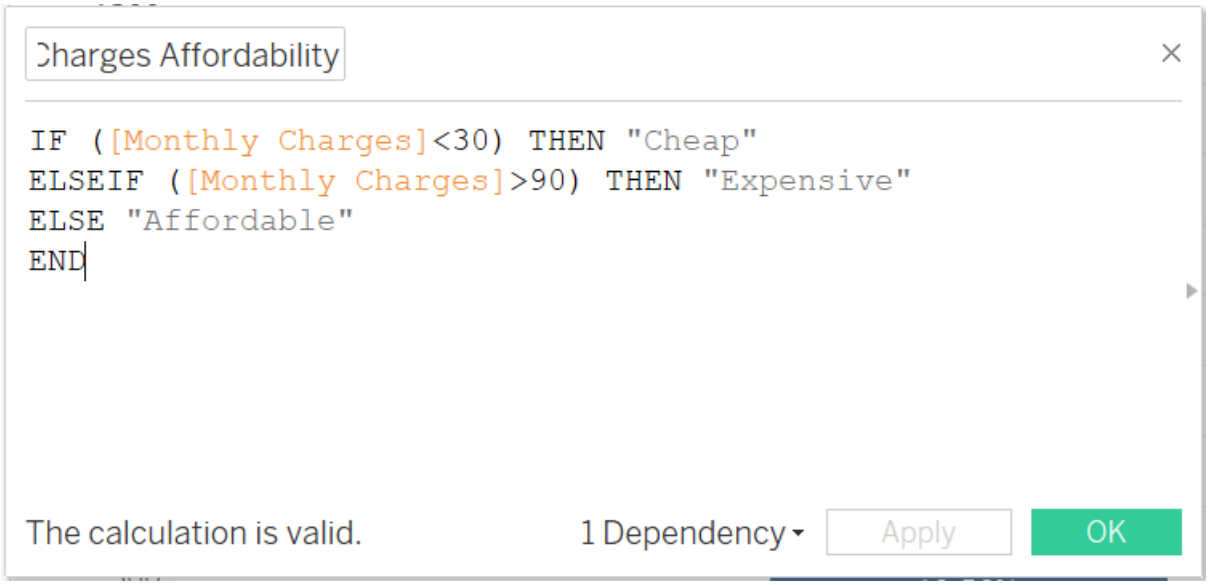
Two year: The monthly charge for Two year Contract range from **18.6 to 118.8** while the range is **100.2** *(118.8 – 18.6 = 100.2).* The median monthly charge is **71.6**. The interquartile range of monthly charges is **69.1** *(94.4 – 25.3 = 69.1).*

Insight: The spread of monthly charges is much lower in the Month-to-month contract due to its lowest range of 98.7 and lowest interquartile range of 36.7, while the Two year contract has a much higher spread due to its highest range of 100.2 and highest interquartile range of 69.1.

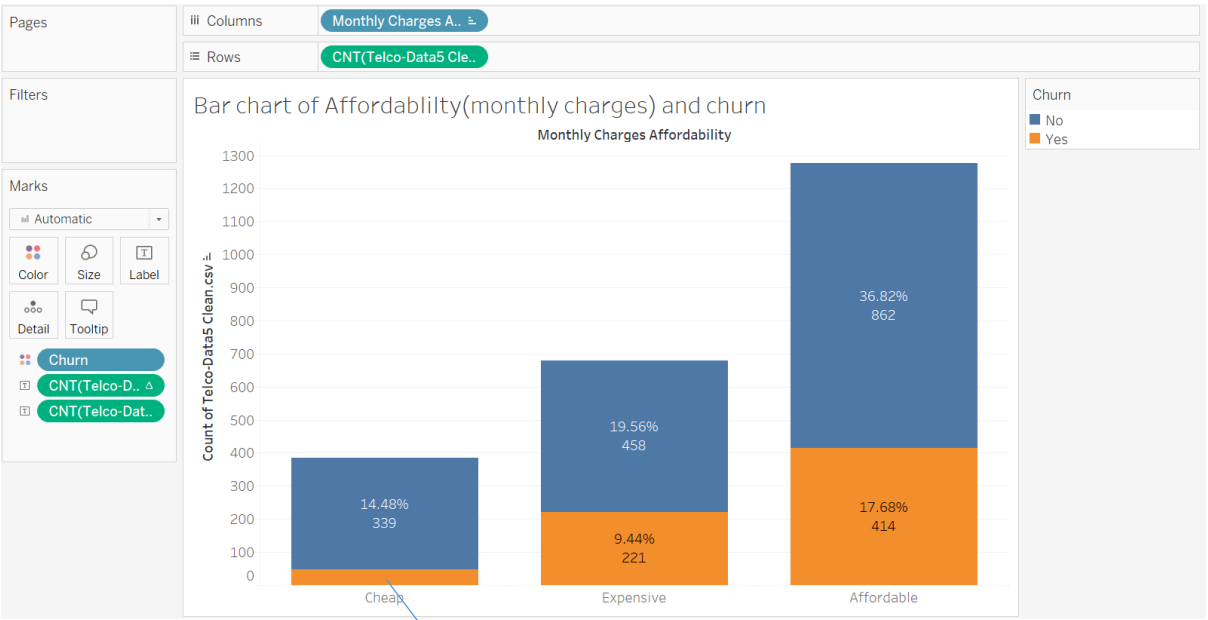
8. Further Insights Questions (At least 3) and answers

Q1. Are customers more likely to stay on contract with a cheaper monthly charge? What percentage of customers is within the range?

To define what's cheap and what's expensive, we will be creating a calculated field in tableau.



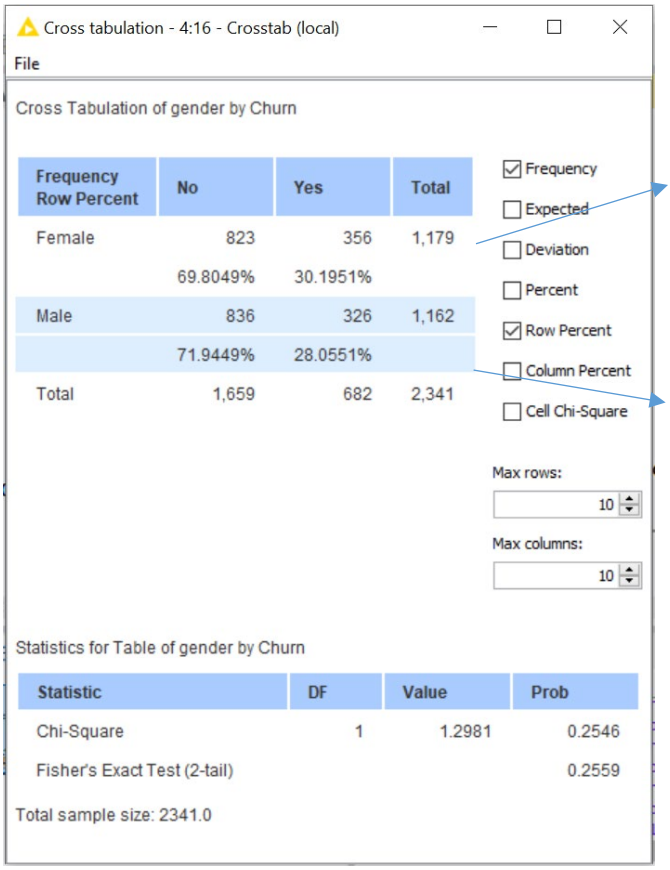
With the calculated field above, monthly charges below 30 would be categorized as "Cheap", monthly charges above 90 would be categorized as "Expensive", while those in between are categorized as "Affordable".



From the bar chart above, we can see that **customers are more likely to stay on contract with a cheaper (below 30) monthly charge at only 2.01% churn rate**. While customers are more likely to churn a contract with an “Affordable” (above 30 but less than 90) monthly price range at 17.68% churn rate.

> Q2. Using a contingency table, determine which variables correlate / affect the customers’ churn rate.

Since variables to correlate with churn are mostly categorical, “Crosstab (local)” node is used.



Gender

In the contingency table, for the 1,179 female customers, **69.8% chose not to churn**, while **30.2% chose to churn.**

For the 1,162 male customers, **71.94% chose not to churn**, while the remaining **28.06% chose to churn.**

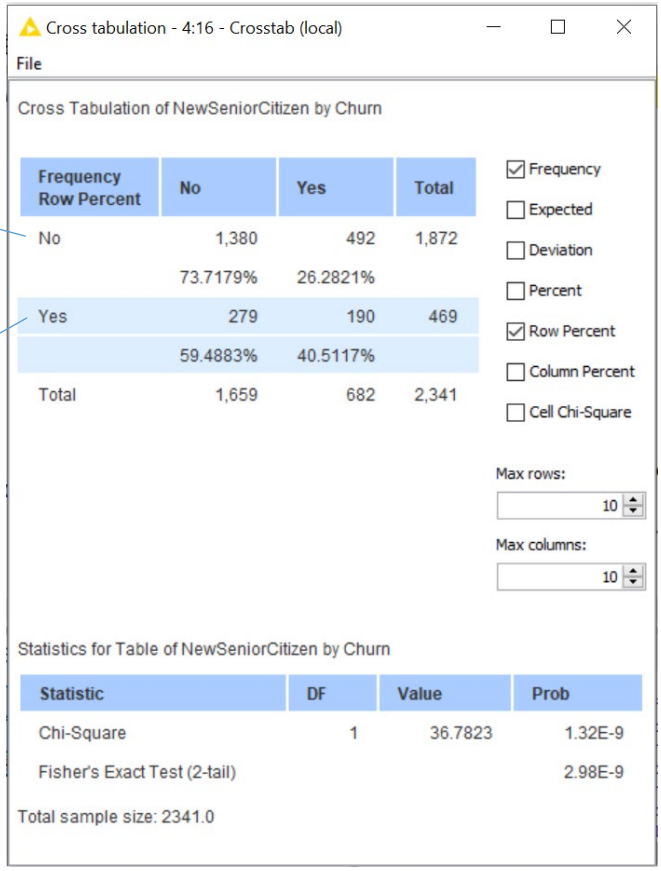
As both male and female customers have a roughly equal chance of churning, gender is not a factor that affects the churn rate.

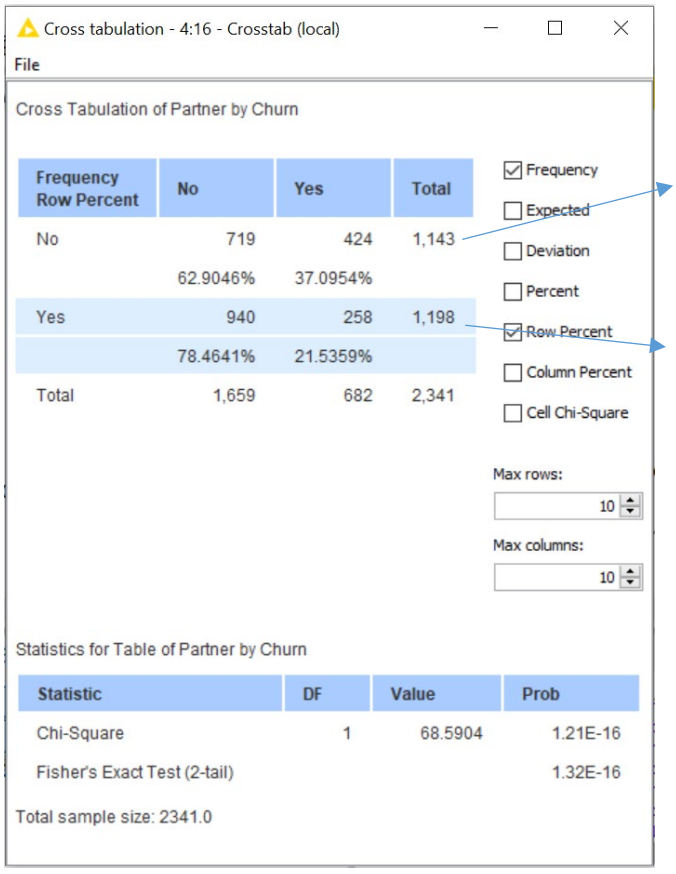
NewSeniorCitizen

For the 1,872 customers who are not senior citizens, **73.72% chose not to churn**, while **26.28% chose to churn.**

As for the 469 customers who are senior citizens, **59.49% chose not to churn**, while **40.51% chose to churn.**

Therefore, customers who are senior citizens are more likely to churn (by ~14%) than those who are not senior citizens.





Partner

For the 1,143 customers with no partners, **62.9% chose not to churn**, while **37.1% chose to churn.**

As for the 1,198 customers with partners, **78.46% chose not to churn**, while the remaining **21.54% chose to churn.**

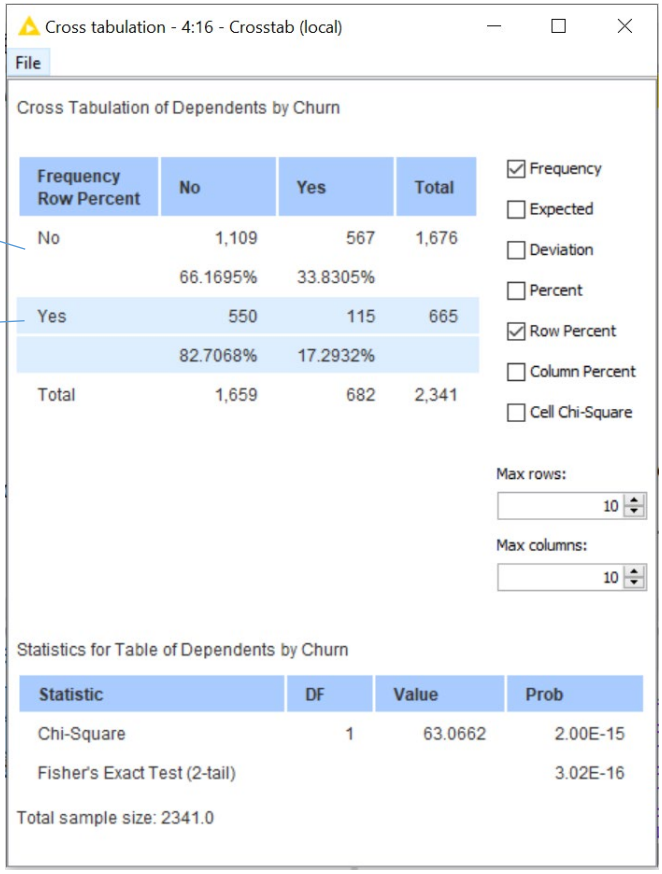
Therefore, customers with no partners are more likely to churn (by ~16%) than those with partners.

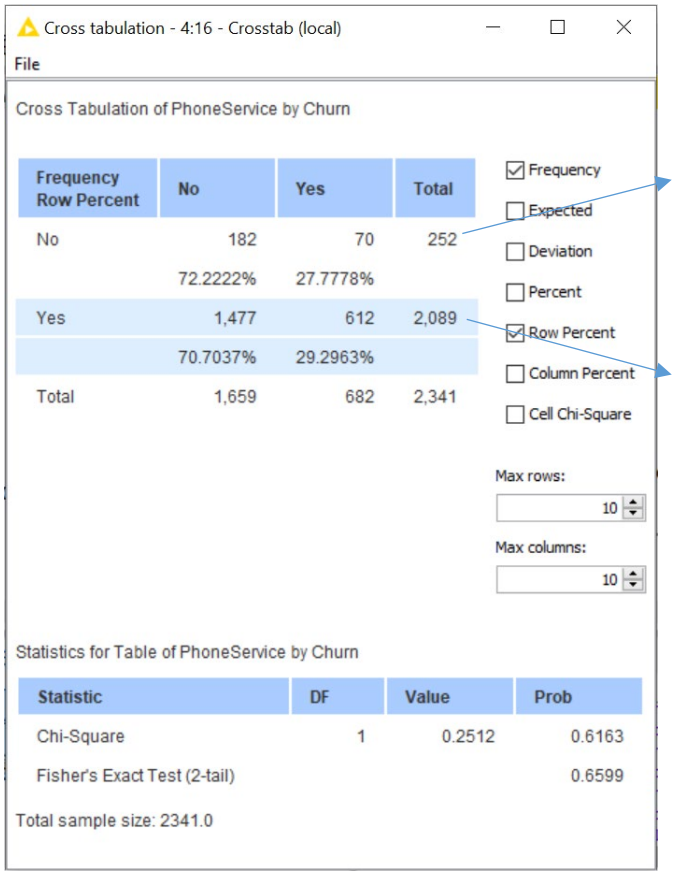
Dependents

For the 1,676 customers with no dependents, **66.17% chose not to churn**, while **33.83% chose to churn.**

As for the 665 customers with dependents, **82.71% chose not to churn**, while the remaining **17.29% chose to churn.**

Hence, customers with no dependents are more likely to churn (by ~17%) than those with dependents.





PhoneService

For the 252 customers with no phone service, **72.22% chose not to churn**, while the remaining **27.78% chose to churn.**

As for the 2,089 customers with phone service, **70.7% chose not to churn**, while the remaining **29.3% chose to churn**.

As both with and without phone service churn rate is similar, phone service is not a factor affecting churn rate.

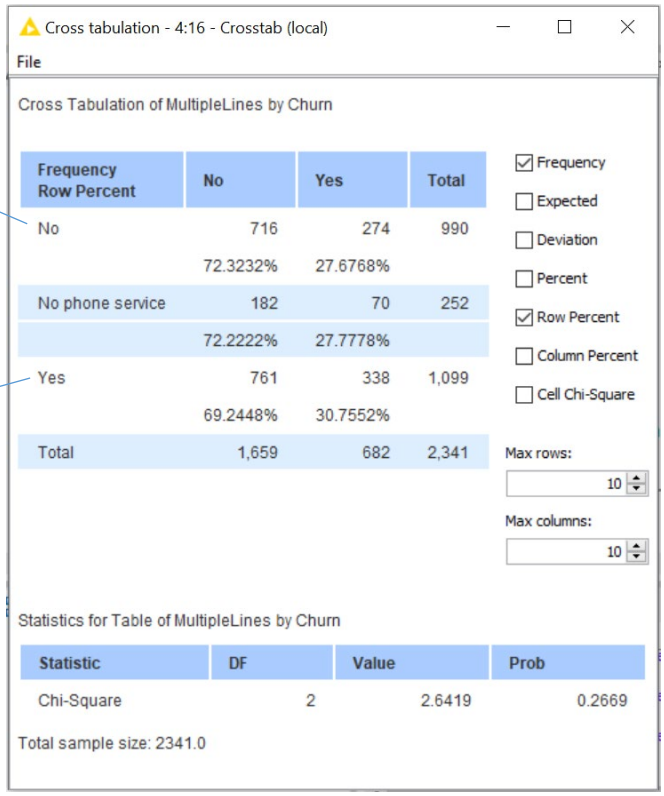
MultipleLines

For the 990 customers with no multiple lines, **72.32% chose not to churn**, while the remaining **27.68% chose to churn.**

As for the 252 customers with no phone service, **72.22% chose not to churn**, while the remaining **27.78% chose to churn.**

Lastly, for the 1,099 customers with multiple lines, **69.24% chose not to churn**, while the remaining **30.76% chose to churn.**

As the churn rate is similar across variable MultipleLines, it is not a factor affecting the churn rate.


Cross tabulation - 4:16 - Crosstab (local)

File

Cross Tabulation of MultipleLines by Churn

Frequency Row Percent	No	Yes	Total
No	716	274	990
	72.3232%	27.6768%	
No phone service	182	70	252
	72.2222%	27.7778%	
Yes	761	338	1,099
	69.2448%	30.7552%	
Total	1,659	682	2,341

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

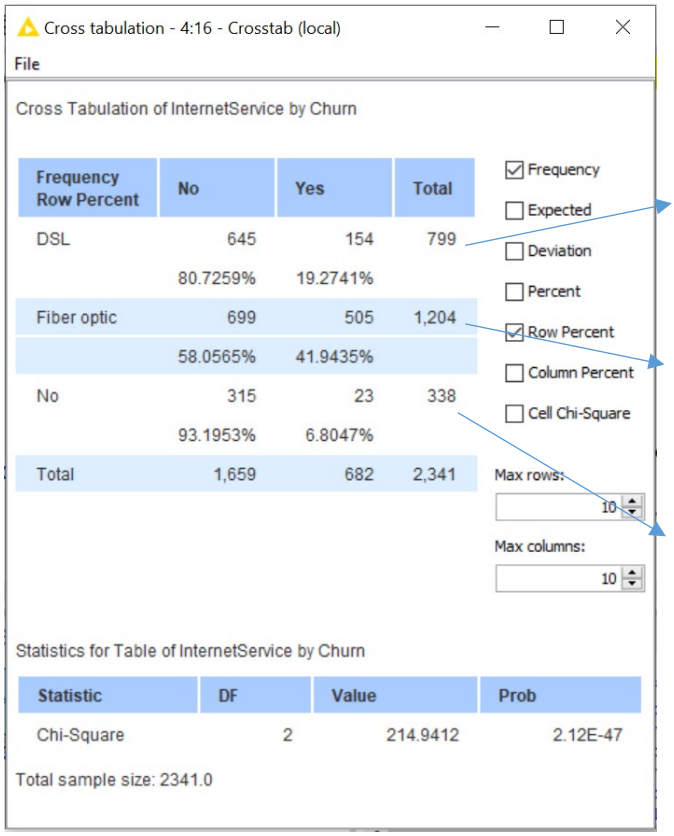
Max rows: 10

Max columns: 10

Statistics for Table of MultipleLines by Churn

Statistic	DF	Value	Prob
Chi-Square	2	2.6419	0.2669

Total sample size: 2341.0


Cross tabulation - 4:16 - Crosstab (local)

File

Cross Tabulation of InternetService by Churn

Frequency Row Percent	No	Yes	Total
DSL	645	154	799
	80.7259%	19.2741%	
Fiber optic	699	505	1,204
	58.0565%	41.9435%	
No	315	23	338
	93.1953%	6.8047%	
Total	1,659	682	2,341

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of InternetService by Churn

Statistic	DF	Value	Prob
Chi-Square	2	214.9412	2.12E-47

Total sample size: 2341.0

InternetService

For the 799 customers with a DSL internet service, **80.73% chose not to churn**, while **19.27% chose to churn.**

As for the 1,204 customers with a Fiber optic internet service, **58.1% chose not to churn**, while **41.9% chose to churn.**

Lastly, for the 338 customers with no internet service, **93.2% chose not to churn**, while **6.8% chose to churn.**

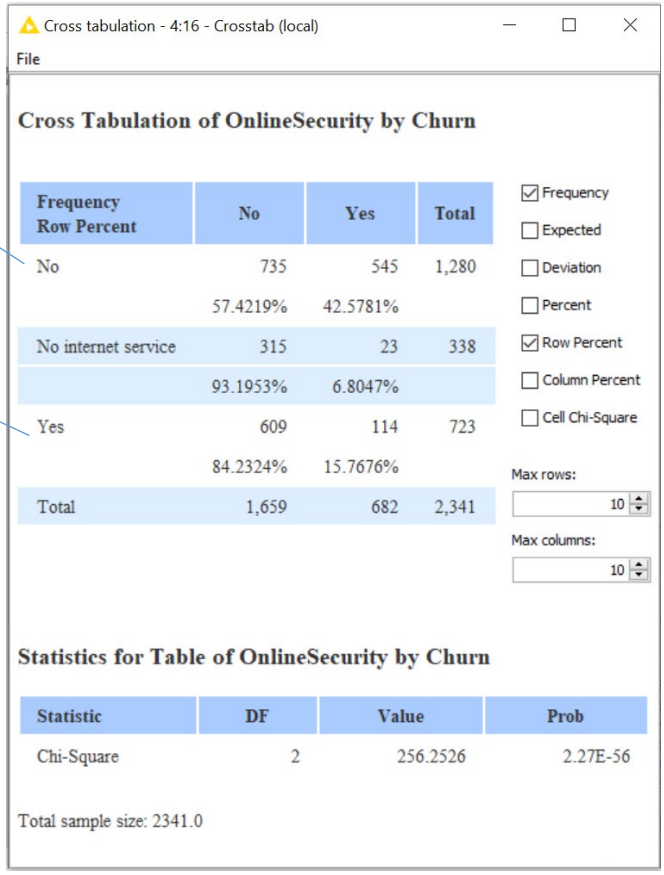
Therefore, customers with a Fiber optic internet service are likely to churn (by ~16%) compared to those with DSL or no internet service.

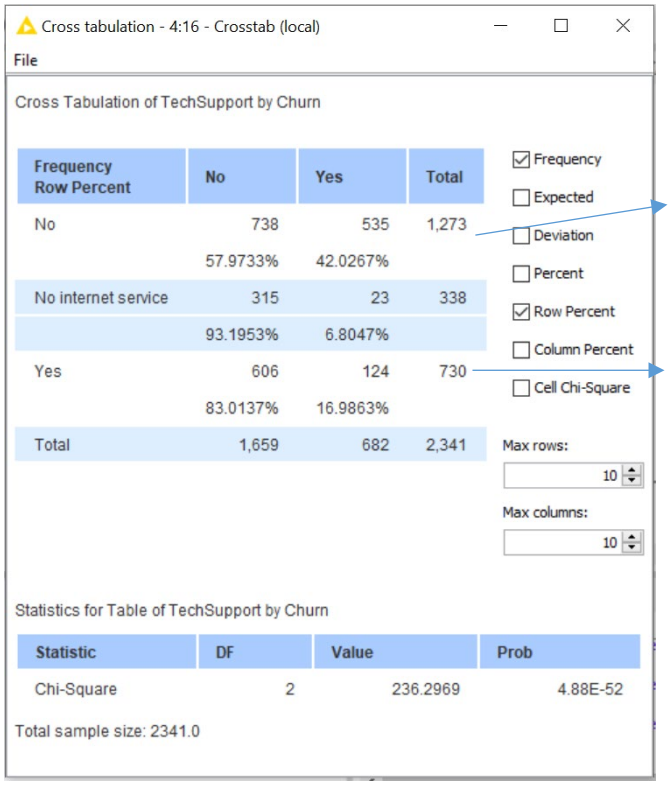
OnlineSecurity

For the 1,280 customers with no online security, **57.42% chose not to churn**, while **42.58% chose to churn**.

As for the 723 customers with online security, a majority of **84.23% chose not to churn**, while only **15.77% chose to churn.**

Hence, customers with no online security are more likely to churn (by ~27%) than those with online security.





TechSupport

For the 1,273 customers with no tech support, **57.97% chose not to churn**, while **42.03% chose to churn.**

As for the 730 customers with tech support, **83.01% chose not to churn**, while **16.99% chose to churn.**

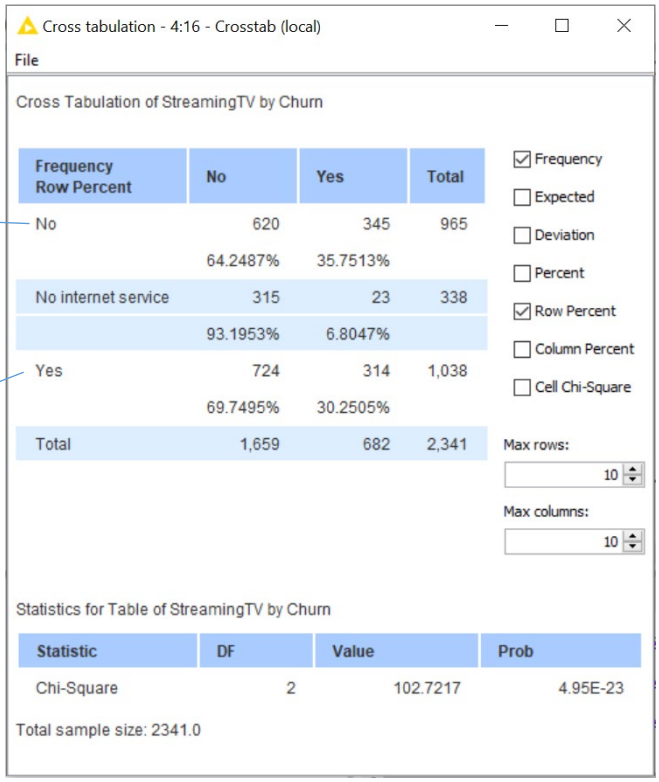
Therefore, customers with no tech support are more likely to churn (by ~25%) than those with tech support.

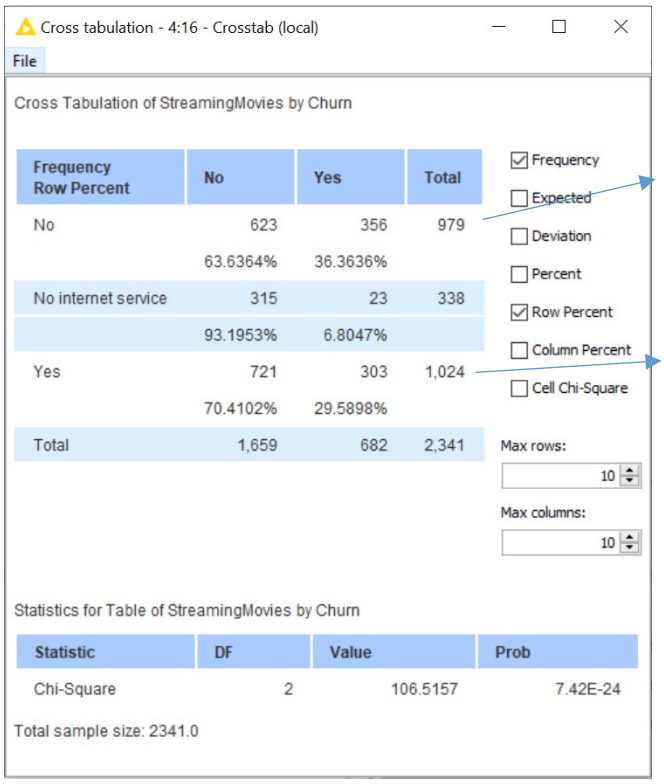
StreamingTV

For the 965 customers with no Streaming TV, **64.25% chose not to churn**, while **35.75% chose to churn.**

As for the 1,038 customers with Streaming TV, **69.75% chose not to churn**, while **30.35 chose to churn.**

Hence, customers with no Streaming TV are more likely to churn (by ~5%) than those with Streaming TV.





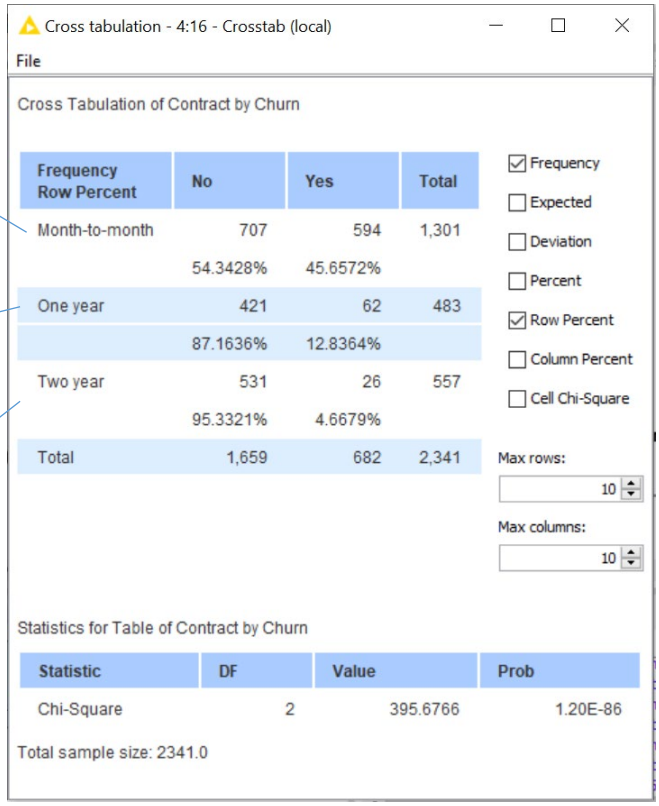
StreamingMovies

For the 979 customers with no Streaming Movies, **63.64% chose not to churn**, while **36.36% chose to churn.**

As for the 1,024 customers with Streaming Movies, **70.41% chose not to churn**, while **29.59% chose to churn.**

Therefore, customers with no Streaming Movies are more likely to churn (by ~7%) than those with Streaming Movies.

Contract

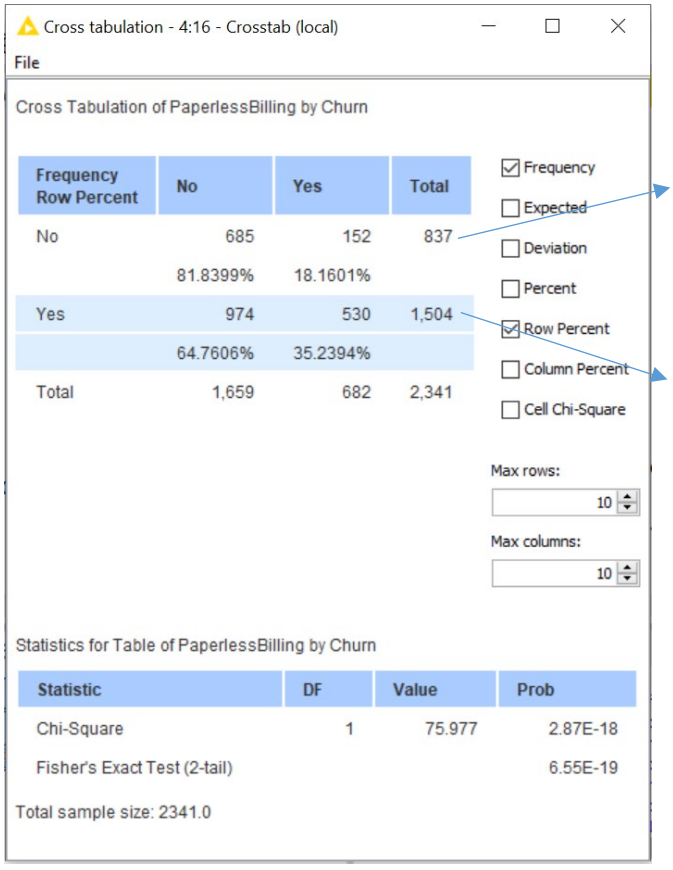


For the 1,301 customers with a Month-to-month contract**, 54.34% chose not to churn**, while **45.66% chose to churn.**

As for the 483 customers with a One year contract, **87.16% chose not to churn**, while **12.84% chose to churn.**

Lastly, for the 557 customers with a Two year contract, **95.33% chose not to churn**, while **4.67% chose to churn.**

Hence, customers with a Month-to-month contract are more likely to churn (by ~28%) than those with a One year or Two year contract.



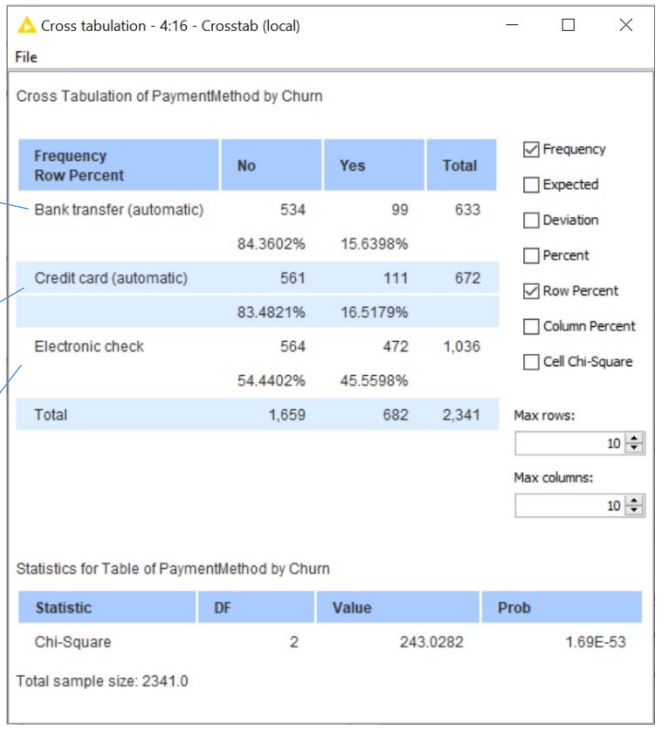
PaperlessBilling

For the 837 customers with no paperless billing, **81.84% chose not to churn**, while **18.16% chose to churn.**

As for the 1,504 customers with paperless billings, **64.76% chose not to churn**, while **35.24% chose to churn.**

Therefore, customers with paperless billings are more likely to churn (by ~17%) than those with no paperless billings.

PaymentMethod



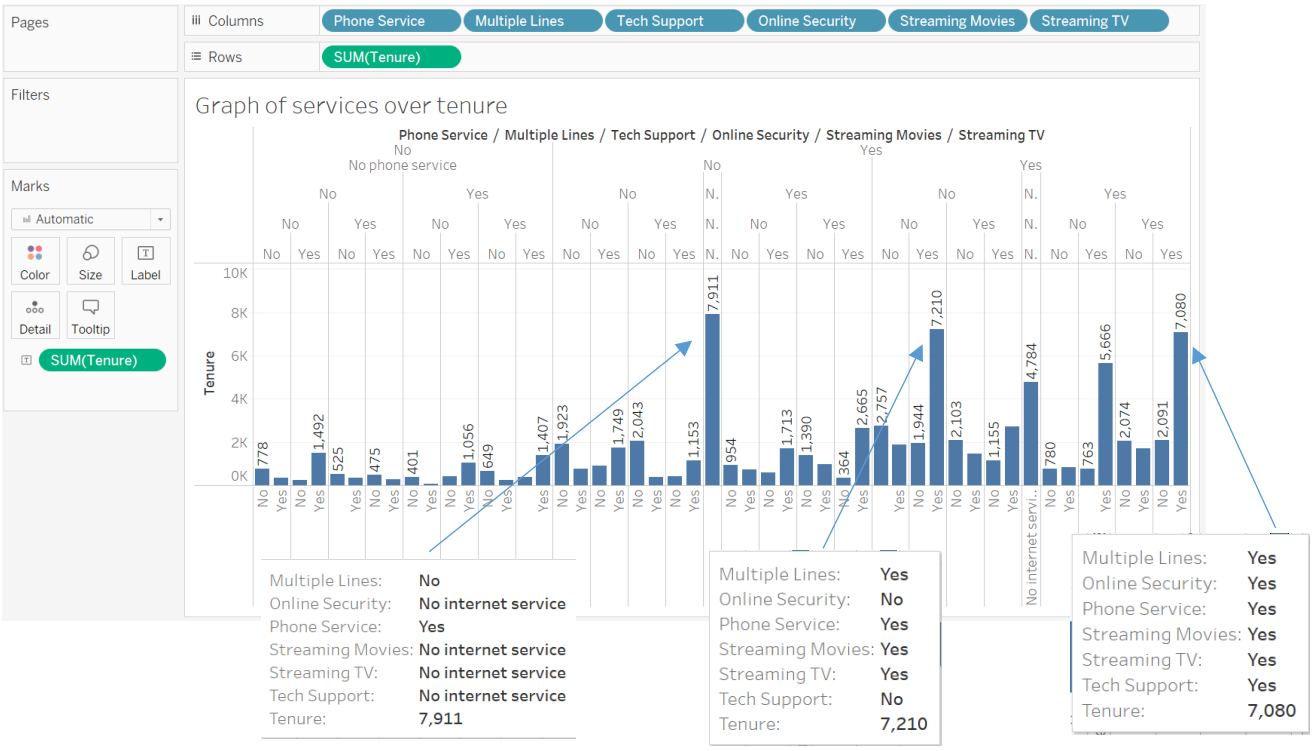
For the 633 customers paying by Bank transfer (automatic), **84.36% chose not to churn**, while **15.64% chose to churn.**

As for the 672 customers paying by Credit card (automatic), **83.48% chose not to churn**, while **16.52% chose to churn.**

Lastly, for the 1,036 customers paying by Electronic check, 54.44% chose not to churn, while 45.56% chose to churn.

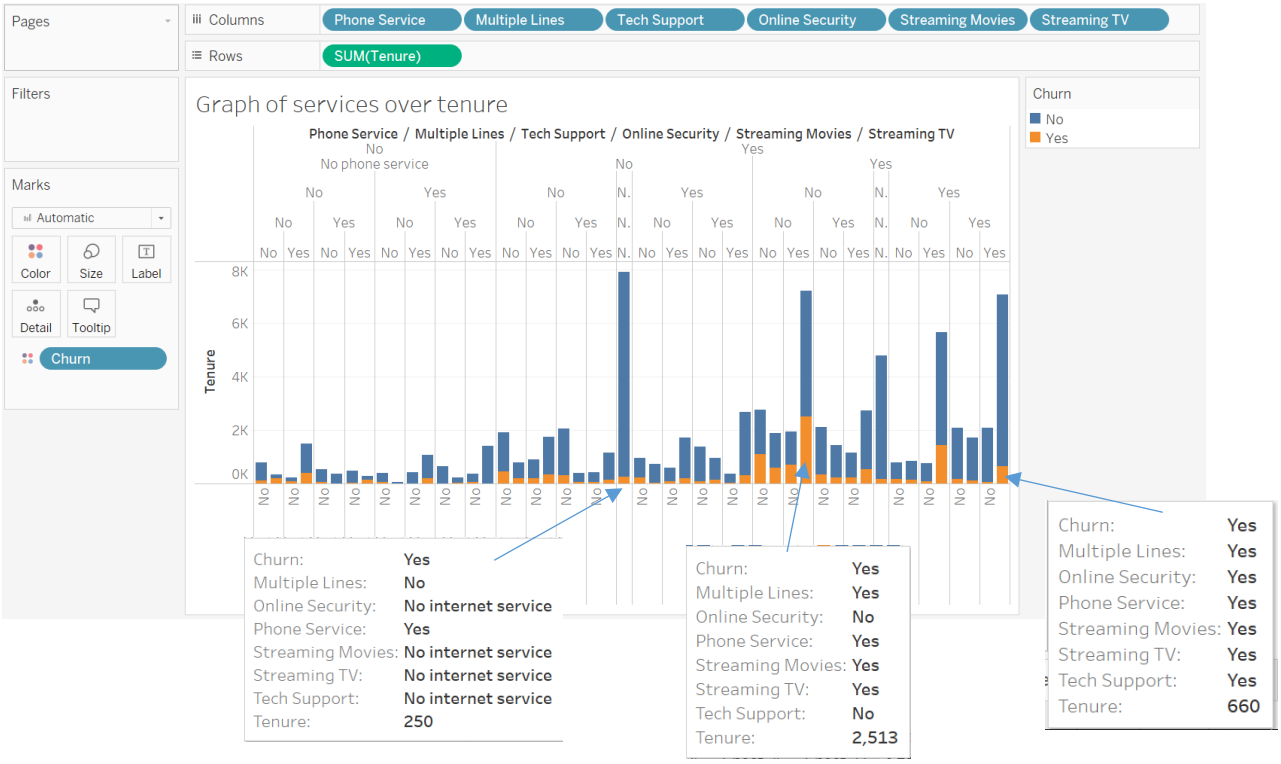
Hence, customers paying by Electronic check are more likely to churn (by ~30%) than those paying by Bank transfer (automatic) OR Credit card (automatic).

Q3. Which combination of variables/subscriptions (eg. StreamingTV, StreamingMovie) are responsible for the highest tenure?



As seen from the graph above, the three peaks show the combination of variables that makes up the highest count of tenure (number of months the customer stayed with a

company). The highest peak shows that **Phone Service** alone makes up for the highest count of tenure at 7,911. Whereas the second peak shows that variables (**Phone Service, Multiple Lines, Streaming Movies and Streaming TV**) are variables that caused its peak count of tenure at 7,210. While the third peak shows ALL variables present which caused its peak count of tenure at 7,080. In all three peaks, the variable **Phone Service** is present.

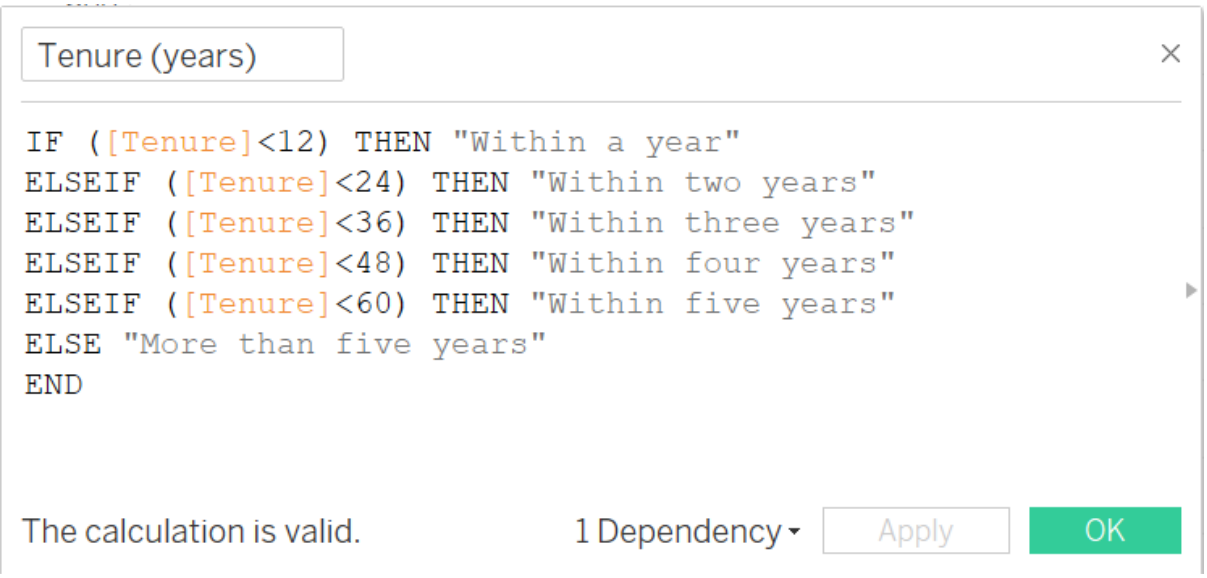


For the highest peak count of tenure at 7,911 with the variable Phone Service only, there is a **3.16% churn rate** *(250/7,911 x 100)*. Whereas for the second peak count of tenure at 7,210 with variables (Phone Service, Multiple Lines, Streaming Movies and Streaming TV), there is a **34.85% churn rate** *(2,513/7,210 x 100).* While the third peak count of tenure at 7,080 with all variables present, there is a **9.32% churn rate** (660/7,080 x 100).

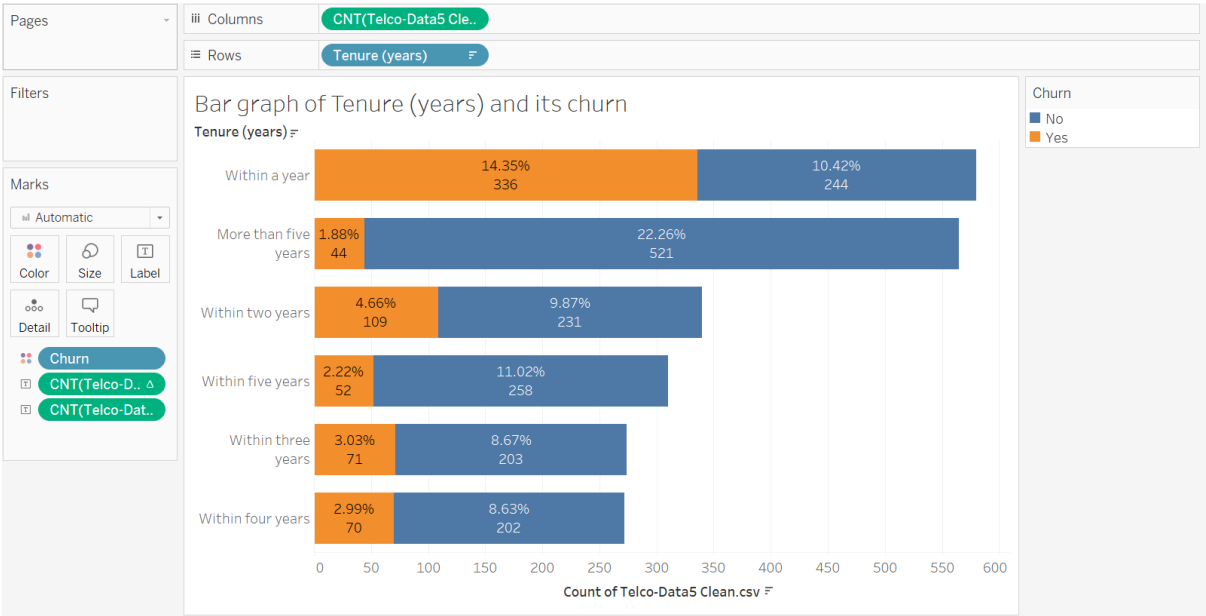
Insight: This could mean that for the second peak count of tenure with a **34.85%** churn rate, customers may churn due to the lack of Tech Support and Online Security as compared to having all variables including Tech Support and Online Security in the third highest peak of tenure with only a **9.32%** churn rate. In addition, Phone Service is the most popular variable and responsible for the highest tenure.

Q4. Among customers who chose to stay, what percentage stayed the longest?

As tenure is counted in months and is numerical, we can categorize them in years so as to simplify the data. This is where the calculated field can be used again.



Tenure less than 12 is categorized as “Within a year”, tenure less than 24 is categorized as “Within two years”, tenure less than 36 is categorized as “Within three years”, tenure less than 48 is categorized as “Within four years”, tenure less than 60 is categorized as “Within five years”, and lastly for more than 60, it would be categorized as “More than five years”.



From the bar graph above, we can see that count of customers during their first year are at its peak. However, during that same first year, the churn rate is also at its peak at 14.35%.

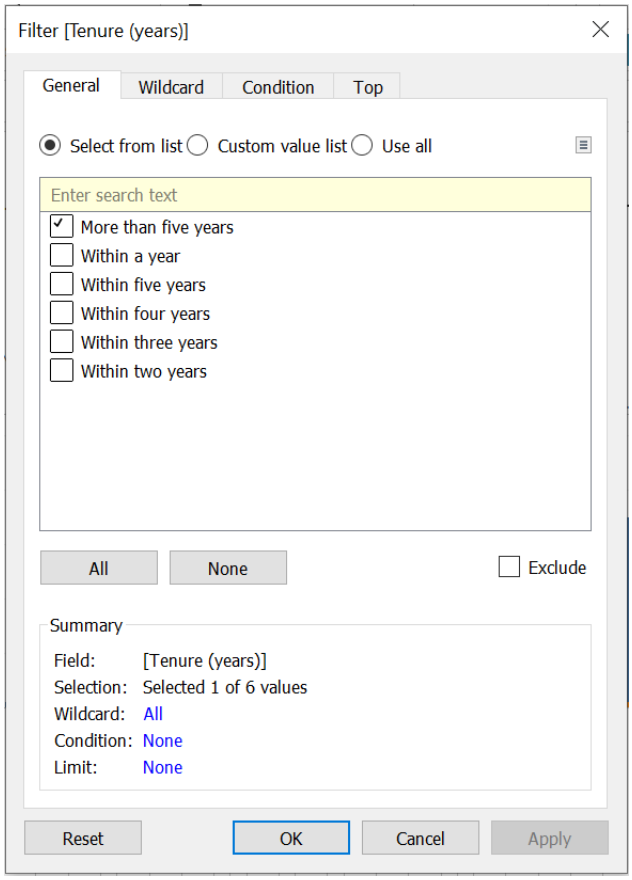
Conversely, the customers’ churn rate is at its lowest when tenure is at more than five years, while the churn rate is also at its lowest at 1.88%. With 22.26% of customers retained for more than five years.

Insight: Within the first year, customers are usually just trying out the services of a Telco company, while at more than five years, they are more accustomed to the Telco company, hence seeing a lesser churn rate.

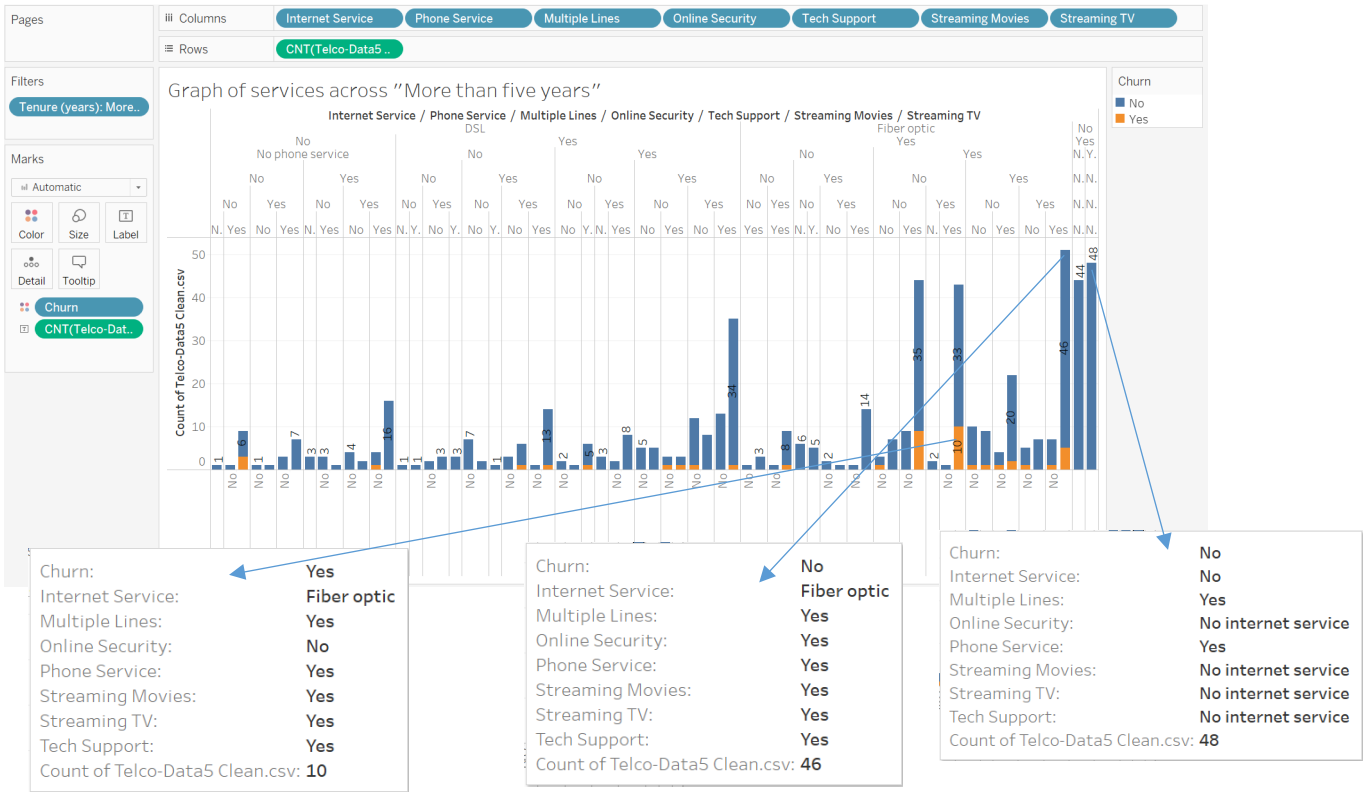
Q5. For those that stayed for “More than five years”, what are the variables that contribute to their long tenure?

As we have categorized tenure into years, we can try to find what variable exactly are making customer choose to stay within a company.

In order to focus on “More than five years” we can use the filter option in tableau.



After filtering to only include “More than five years”, we can proceed to add the appropriate services provided by Telco companies to determine its peak.



As we can see from the graph above, there are multiple peaks. The highest peak shows customers prefer a Fiber optic internet service with all services provided, with a churn rate of 9.8% as 5 out of the 46 customers chose to churn.

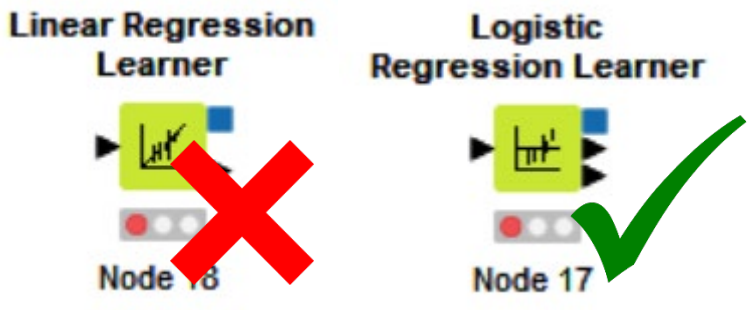
Whereas for the second peak, it shows no churn at all. The services provided are just Phone Service and Multiple Lines.

Lastly, the highest churn rate for “More than five years” comes from a combination of Fiber optic internet, Multiple Lines, no Online Security, Phone Service, Streaming Movies, Streaming TV, and Tech support, at a 23.26% churn rate as 10 out of the 33 customers chose to churn.

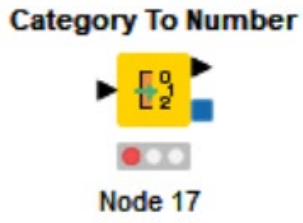
Insight: Customers may prefer more services provided by a Telco Company, hence the tenure of more than five years. However, some customers just prefer phone service with multiple lines with no additional services provided.

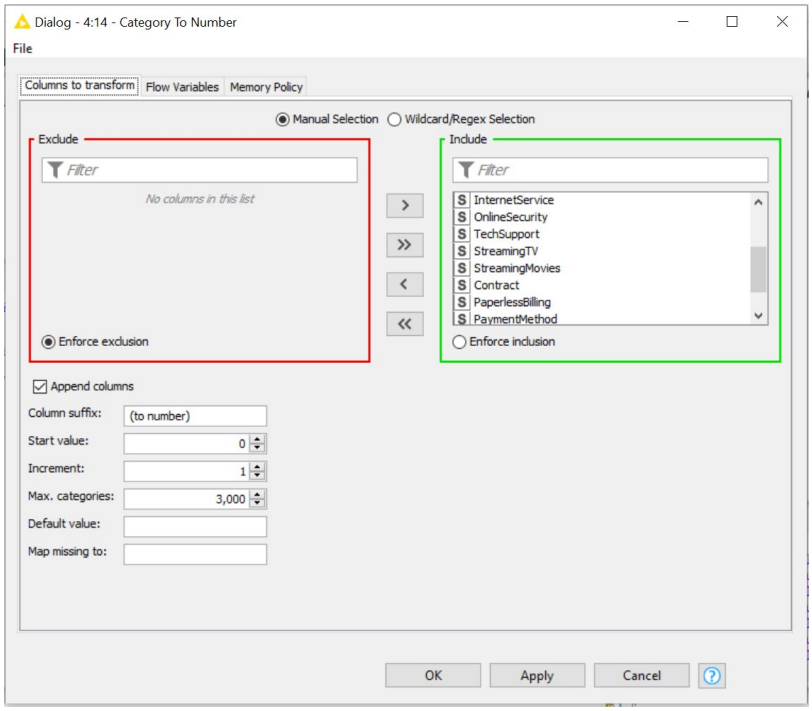
9. Data Modeling

The target for prediction is “Churn”, as to find out and predict customers’ stay on or leaving the Telco company. In addition, as the as the outcome of “Churn” is a categorical data type, linear regression learner cannot be used. Hence, logistics regression learner is used to generate the model instead.



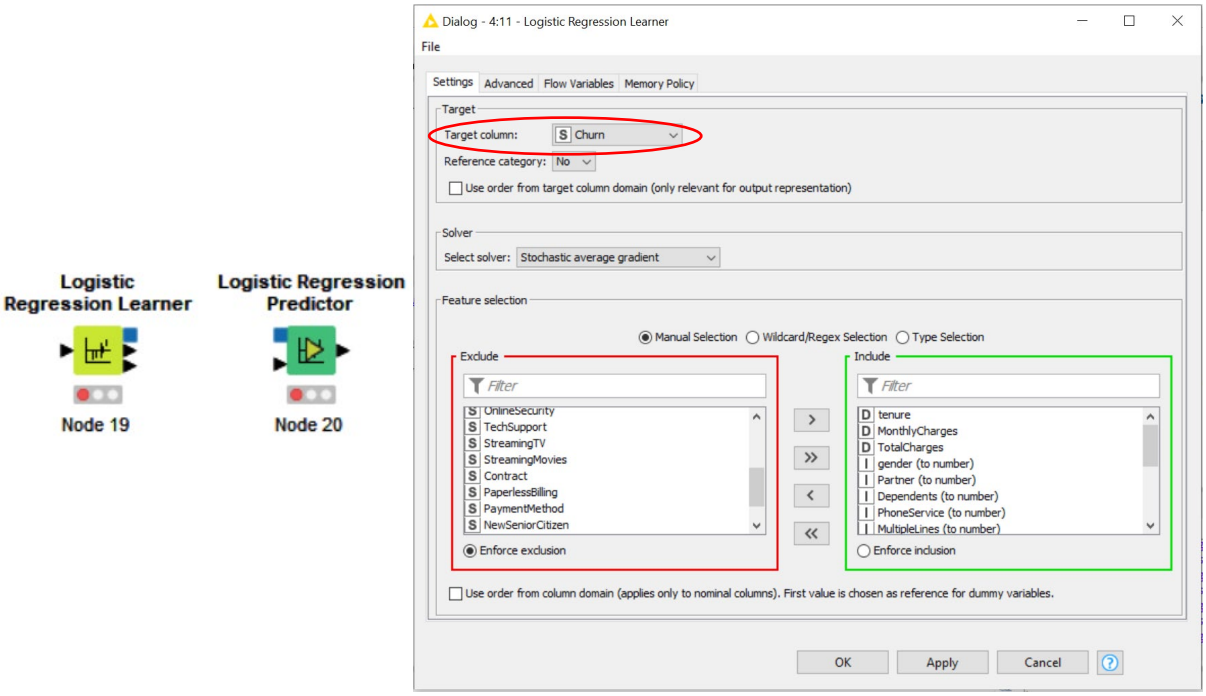
To use categorical data for training, they have to be converted to numerical data by using “Category To Number” node.



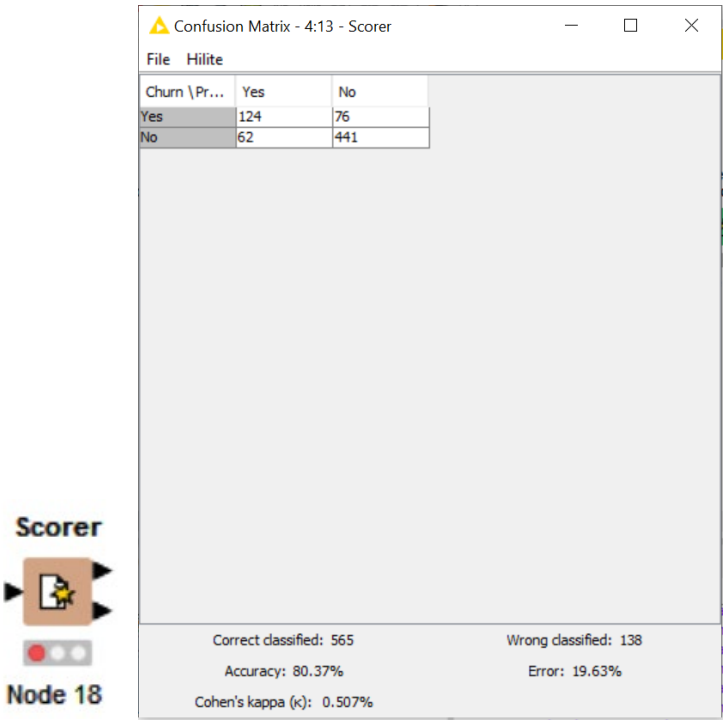


Once converted, apply Z transform by using the “Normalizer” node to normalize numerical data.

Once converted and normalized, then can we proceed to partitioning, splitting data 70% for training and 30% for testing.



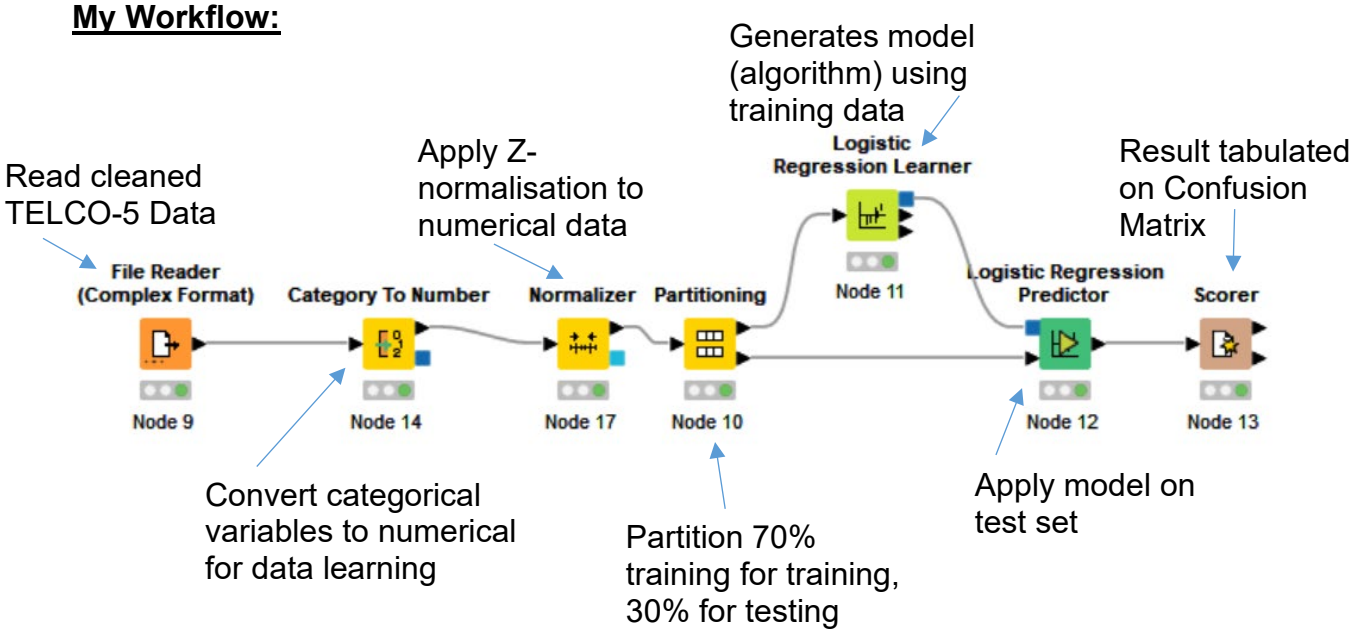
Using Logistics Regression Learner and Predictor, “Churn” will be the target column. In addition, we will include all relevant variables to generate a model and to test for variable “Churn” accuracy. A “Scorer” node will be used to display a confusion matrix and to check for prediction accuracy.



Correct classified: 565	Wrong classified: 138
Accuracy: 80.37%	Error: 19.63%
Cohen's kappa (κ): 0.507%	

In the confusion matrix above, with a sample size of 703, it shows 138 wrong predictions, while showing 565 correct predictions. With an accuracy of 80.37% and an error of 19.63%, the prediction of churn is mostly accurate.

My Workflow:



10. Conclusion

- Customers prefer paying between $65 to $105 a month on a Telco Company. For a more cheaper option, many customers prefer spending $15 to $25 a month on a Telco Company.
- There is a weak positive linear correlation between tenure and MonthlyCharges. When tenure increases, MonthlyCharges also increases.
- Churn is directly proportionate to the tenure.
- In the Telco-5 dataset, a majority of customers chose to stay while only 29.13% chose to churn.
- The least amount of churn rate is between a monthly charge of 15 to 20, at a 6.56% churn rate. While the highest churn rate is between monthly charges of 70 to 75, at a 45.45% churn rate.
- the most common Contract type is “Month-to-month, while the lowest Contract type is “One year”.
- However, on average, the Two year contract is cheaper monthly at 65.956 while the Month-to-month is more expensive on average at 71.630.
- Most customers prefer opting for a Month-to-month contract as opposed to the other contract types, despite the higher price.

- Month-to-month contract is the **most likely to be churned** at 25.37% whereas the least likely to be churned is the Two year contract at 1.11%.
- It is likely customers chose a Month-to-month contract to try it out for the first month to see if they like it as it is easier to break the contract after the first month of trial.
- Customers are **more likely to stay on contract** with a cheaper (below 30) monthly charge at only 2.01% churn rate. While customers are more likely to churn a contract with an “Affordable” (above 30 but less than 90) monthly price range at a 17.68% churn rate.
- As both male and female customers have a roughly equal chance of churning, gender is not a factor that affects the churn rate.
- Customers who are senior citizens are **more likely to churn** than those who are not senior citizens.
- Customers with no partners are **more likely to churn** than those with partners.
- Customers with no dependents are **more likely to churn** than those with dependents.
- As both with and without phone service churn rate is similar, Phone Service is not a factor affecting churn rate.
- As the churn rate is similar across variable MultipleLines, it is not a factor affecting the churn rate.
- Customers with a Fiber optic internet service are **more likely to churn** compared to those with DSL or no internet service.
- Customers with no online security are **more likely to churn** than those with online security.
- Customers with no tech support are **more likely to churn** than those with tech support.
- Customers with no Streaming TV are **more likely to churn** than those with Streaming TV.
- Customers with no Streaming Movies are **more likely to churn** than those with Streaming Movies.
- Customers with a Month-to-month contract are **more likely to churn** than those with a One year or Two year contract.
- customers with paperless billings are **more likely to churn** than those with no paperless billings.
- Customers paying by Electronic check are **more likely to churn** than those paying by Bank transfer (automatic) OR Credit card (automatic).
- **Phone Service** makes up for the highest count of tenure (retaining)
- Within the first year, customers are usually just trying out the services of a Telco company, while at more than five years, they are more accustomed to the Telco company, hence seeing a lesser churn rate.
- Customers are likely to stay with more services provided OR when they are subscribed to just phone service with multiple lines.
- Lack of **Tech Support** and **Online Security** are major factors in churn rate.

11. Reflection

Working on this project has been really fun for me as it has opened my eyes to Data Visualization and its many ways to visualize data. Exploring and playing around with both Knime and Tableau has been really interesting and has given me more proficiency in using both programs. With the help of my lecturer, I have learned a lot this semester, such as modeling data and transforming them to learn more and gather unforeseen insights. Although it was enjoyable making and visualizing data, there were challenges such as knowing whether the right variables (eg. Categorical, numerical) are used in the visualization of my dataset. However, I have managed to resolve them via trial and error and by understanding my data correctly. In addition, coming up with suitable questions was also a challenge as I did not know if my questions were of use to the objective, to resolve it, I had researched on Kaggle.com to gather more information on my dataset. If I had to do thing differently, I would definitely come up with more complex questions and answer, however due to time constraints I could not. I have learned a lot from Tableau such as making various different plots and graphs to analyze them to gather insights. In Knime, I had learnt a variety of different nodes with its different functions. With my proficiency in Knime and Tableau, I can definitely apply it to future job prospects to gather insights and visualize data.

12. References

Kaggle (2022). Retrieved from https://www.kaggle.com/datasets/blastchar/telco-customer-churn **[1]**