Exploratory Data Analysis

Data at a glance - initial observations

The data in Microsoft Excel Spreadsheet consists of 21 data fields (columns) and 7043 tuples (records).

A few initial observations in respect of data are given below:

(1) Data field 'tenure' appears to be in months, as for individual records the values in field 'TotalCharges' are close to the corresponding product of 'MonthlyCharges' multiplied by 'tenure'. The differences in 'TotalCharges' and corresponding product may be due to billing plan changes.

(2) There are many new customers as Mode is 1 for 'tenure'.

(3) There are many customers with low monthly charges of around 20 (close to Minimum of 18.25) as Mode is 20.05 for 'MonthlyCharges'.

(4) The data had 11 records that had data field 'TotalCharges' as "blank", for records with 'customerID' as mentioned below:

No.	customerID
1	4472-LVYGI
2	3115-CZMZD
3	5709-LVOEQ
4	4367-NUYAO
5	1371-DWPAZ
6	7644-OMVMY
7	3213-VVOLG
8	2520-SGTTA
9	2923-ARZLG
10	4075-WKNIU
11	2775-SEFEE

These 11 records represent 0.16% of total records (0.17% of records for males and 0.14% of records for females). The count of these 11 records is not very significant considering the overall size of data.

All these 11 records having data field 'TotalCharges' as "blank" were for customers with 'tenure' as '0'. There was no other 'tenure' (i.e. apart from '0') for which records had 'TotalCharges' as "blank".

It appears that the data is for customers on post-paid billing plans for which customers make the payments after the bill is generated by the telecom company at the end of each billing period; and 'TotalCharges' for recently onboarded customers (having 'tenure' as '0') were yet to be populated as '0' (being calculated as 'tenure' of '0' multiplied by the corresponding 'MonthlyCharges').

It may also be the case that the telecom company's systems that store and/or generate the data keep 'TotalCharges' as blank if either the 'tenure' or the calculated value for 'TotalCharges' is '0'.

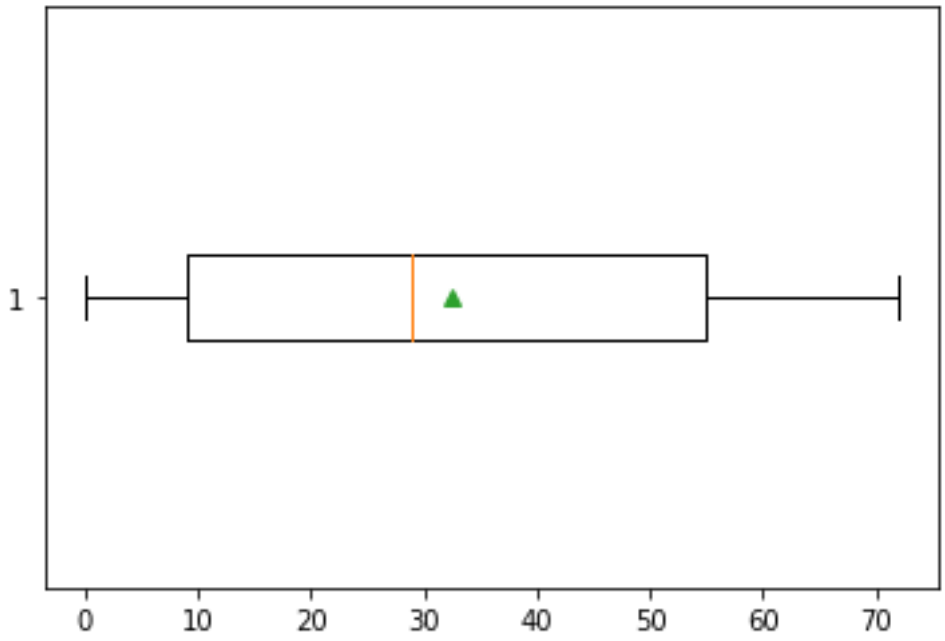
'TotalCharges' for these 11 records have been replaced by 0 for completeness of data, and to enable calculations and graphs using complete data.

(5) A summary of the data fields (after the correction for 11 records that had data field 'TotalCharges' as "blank") is included below:

No.	Data Field	Data Example (First Row)	Data Type	Comments
1	customerID	7590-VHVEG	Text	Alphanumeric, with punctuation (hyphen "-")
2	gender	Female	Text	Binary (Male, Female)
3	SeniorCitizen	0	Numeric	Binary (0, 1)
4	Partner	Yes	Text	Binary (Yes, No)
5	Dependents	No	Text	Binary (Yes, No)
6	tenure	1	Numeric	Minimum = 0, Maximum = 72, Average = 32.37, Median = 29, Mode = 1. Tenure appears to be in months, as values in field ‘TotalCharges’ are approximately ‘MonthlyCharges’ multiplied by ‘tenure’. There are many new customers as Mode is 1.
7	PhoneService	No	Text	Binary (Yes, No)
8	MultipleLines	No phone service	Text	Yes, No, No phone service
9	InternetService	DSL	Text	DSL, Fiber optic, No
10	OnlineSecurity	No	Text	Yes, No, No internet service
11	OnlineBackup	Yes	Text	Yes, No, No internet service
12	DeviceProtection	No	Text	Yes, No, No internet service
13	TechSupport	No	Text	Yes, No, No internet service
14	StreamingTV	No	Text	Yes, No, No internet service
15	StreamingMovies	No	Text	Yes, No, No internet service
16	Contract	Month-to-month	Text	Month-to-month, One year, Two year
17	PaperlessBilling	Yes	Text	Binary (Yes, No)
18	PaymentMethod	Electronic check	Text	Bank transfer (automatic), Credit card (automatic), Mailed check, Electronic check
19	MonthlyCharges	29.85	Numeric	Minimum = 18.25, Maximum = 118.75, Average = 64.76, Median = 70.35, Mode = 20.05. There are many customers with low monthly charges of around 20 (close to the Minimum of 18.25).
20	TotalCharges	29.85	Numeric	Values in field ‘TotalCharges’ are approximately ‘MonthlyCharges’ multiplied by ‘tenure’ for individual records. Minimum = 0 for customers with ‘tenure’ as ‘0’, and 18.80 for customers with ‘tenure’ not as ‘0’ (for a customer with ‘tenure’ as ‘1’). Maximum = 8,684.80 (for a customer with ‘tenure’ as ‘72’), Average = 2,279.73, Median = 1,394.55, Mode = 20.20 (for 11 customers with ‘tenure’ as ‘1’).
21	Churn	No	Text	Binary (Yes, No)

The boxplots (including mean) were created for data fields ‘tenure’, ‘MonthlyCharges’ and ‘TotalCharges’.

Boxplot for ‘tenure’

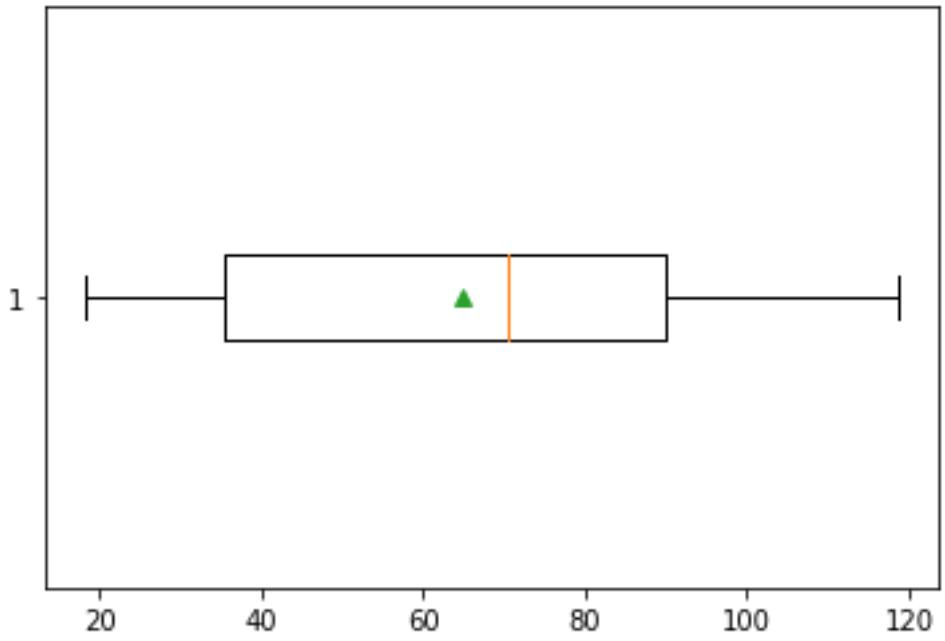


The boxplot above shows that ‘tenure’ has a minimum of 0, a maximum of 72, and hence a Range of 72.

The mean for ‘tenure’ is 32.37 and the median is 29.

For ‘tenure’, the Quartile 1 (Q1) value is 9, the Quartile 3 (Q3) value is 55, and the Interquartile range (IQR) is 46. The distribution of ‘tenure’ is sightly right-skewed.

Boxplot for ‘MonthlyCharges’

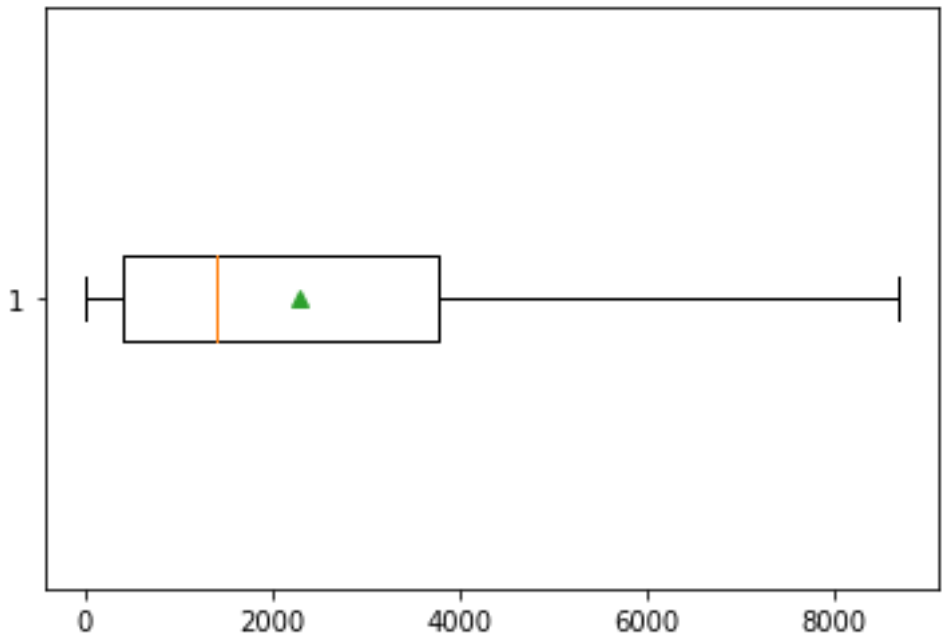


The boxplot above shows that ‘MonthlyCharges’ have a minimum of 18.25, a maximum of 118.75, and a Range of 100.50.

The mean for ‘MonthlyCharges’ is 64.76 and the median is 70.35.

For ‘MonthlyCharges’, the Quartile 1 (Q1) value is 35.50, the Quartile 3 (Q3) value is 89.85, and the Interquartile range (IQR) is 54.35. The distribution of ‘MonthlyCharges’ is sightly left-skewed.

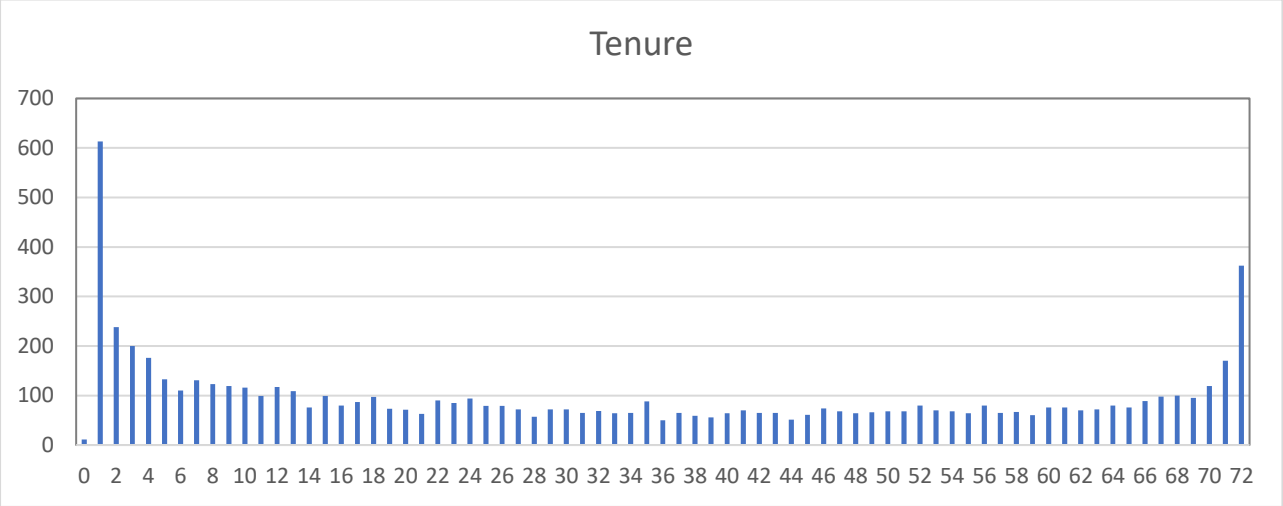
Boxplot for ‘TotalCharges’

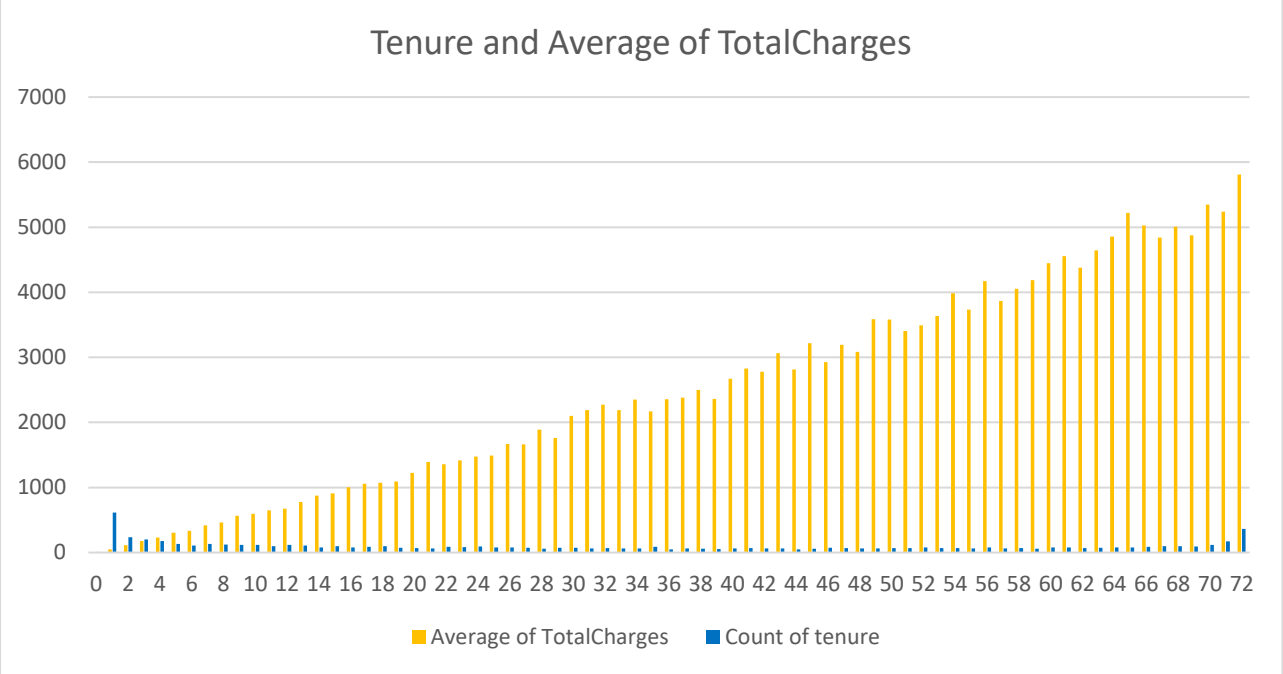


The boxplot above shows that ‘TotalCharges’ have a minimum of ‘0’ for customers with ‘tenure’ as ‘0’. The minimum for customers with ‘tenure’ not as ‘0’ is 18.80 (for a customer with ‘tenure’ as ‘1’). ‘TotalCharges’ have a maximum of 8,684.80 (for a customer with ‘tenure’ as ‘72’). Hence, the Range for ‘TotalCharges’ is 8684.80 on an overall basis for ‘tenure’ ranging from 0 to 72.
The mean for ‘TotalCharges’ is 2,279.73 and the median is 1,394.55.

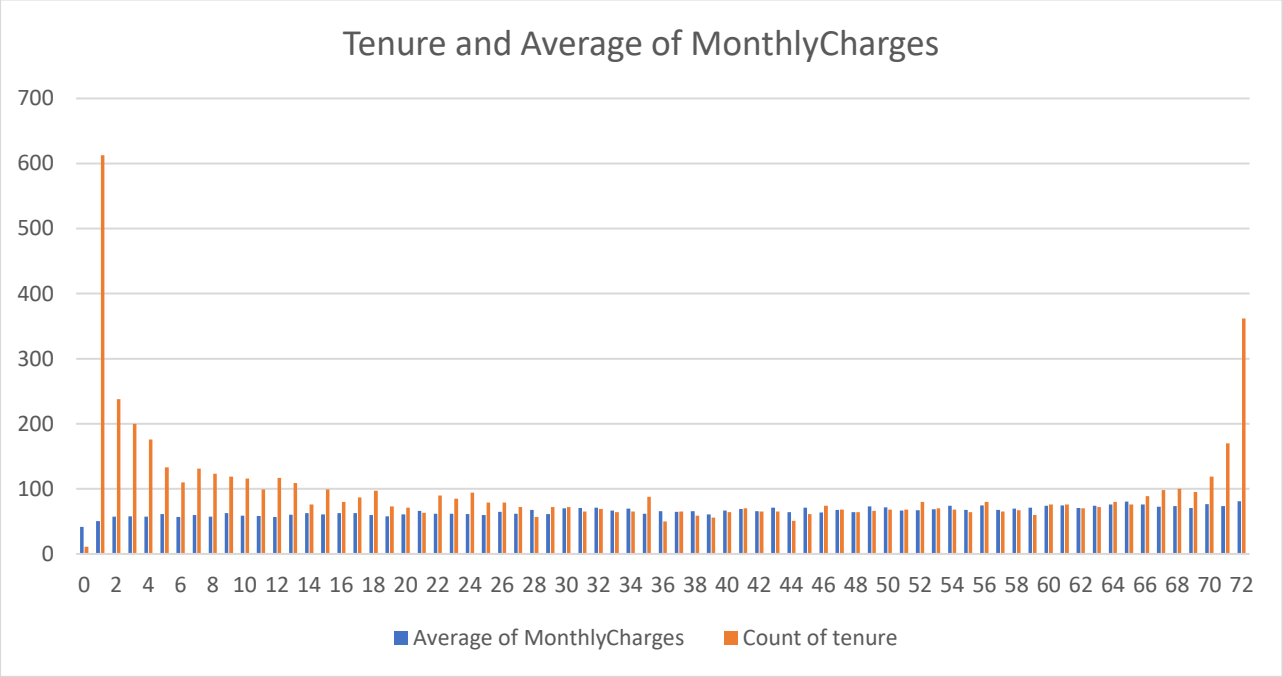
For ‘TotalCharges’, the Quartile 1 (Q1) value is 398.55, the Quartile 3 (Q3) value is 3,786.60, and the Interquartile range (IQR) is 3,388.05.

The distribution of ‘TotalCharges’ is significantly right-skewed, which may be attributed to higher cumulation of charges over time for ‘TotalCharges’ as ‘tenure’ increases, as indicated by the charts below created in Microsoft Excel.

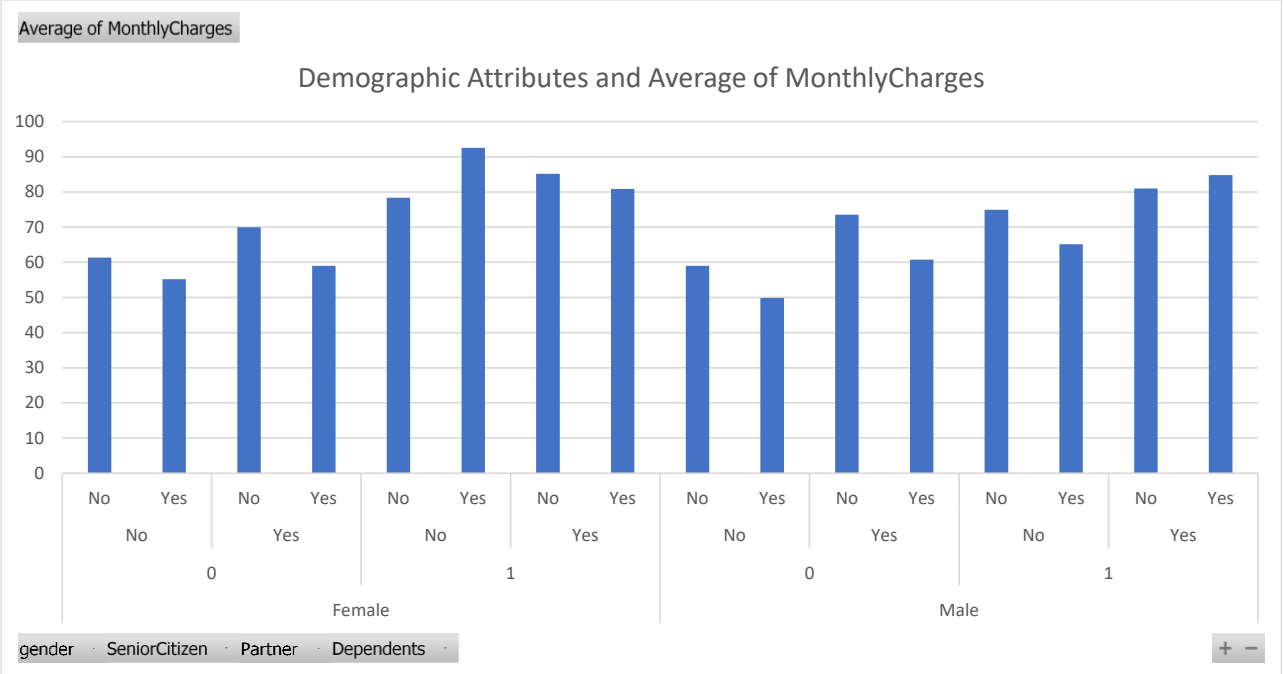




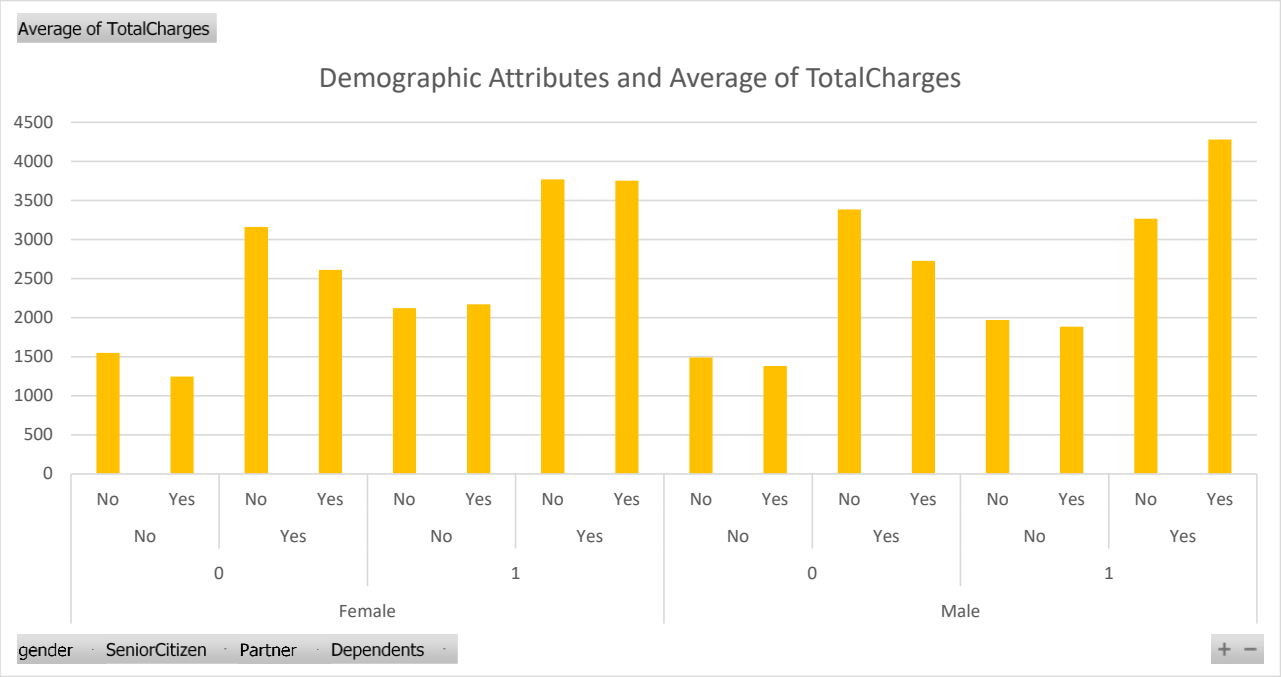
As compared to 'TotalCharges', the chart for 'MonthlyCharges' shows that the customers are spending more per month at the lower and the higher ends of 'tenure', while the spend decreases for tenures towards the middle of the range for 'tenure'.



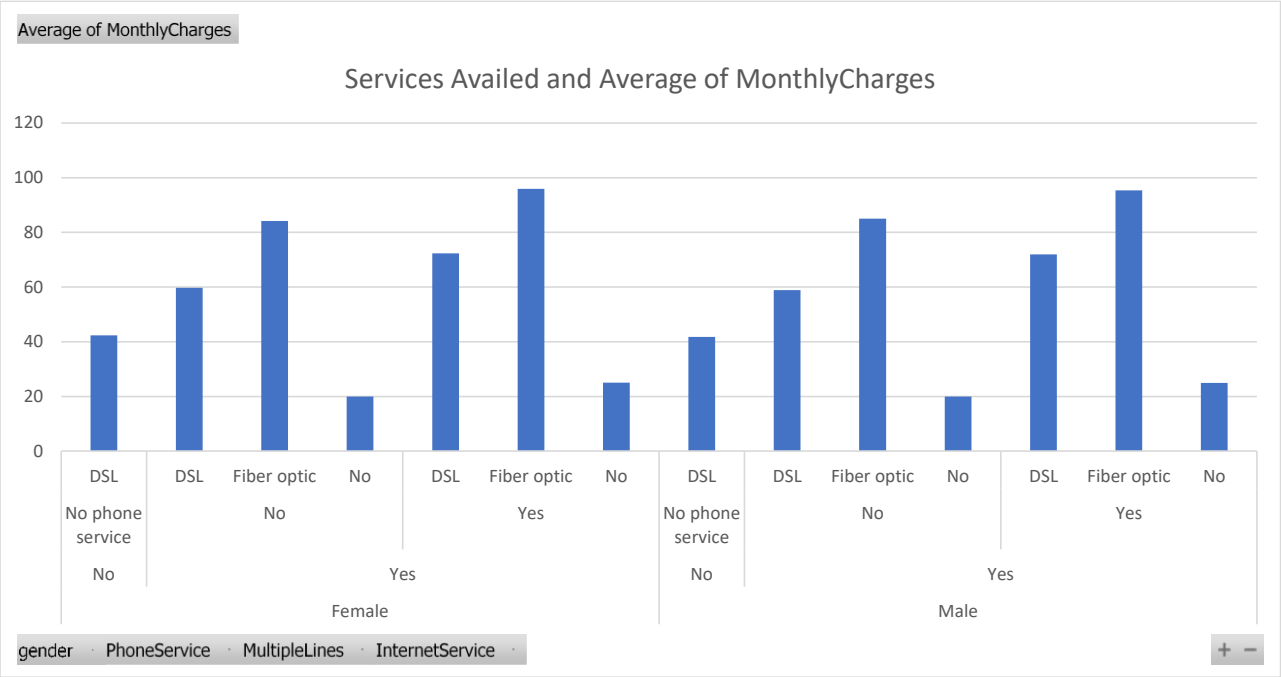
The effect of demographic attributes 'gender', 'SeniorCitizen', 'Partner', and 'Dependents' on 'MonthlyCharges' is shown in the chart below. The chart is split by the demographics in the sequence of firstly 'gender', then 'SeniorCitizen', then 'Partner', and finally 'Dependents'.



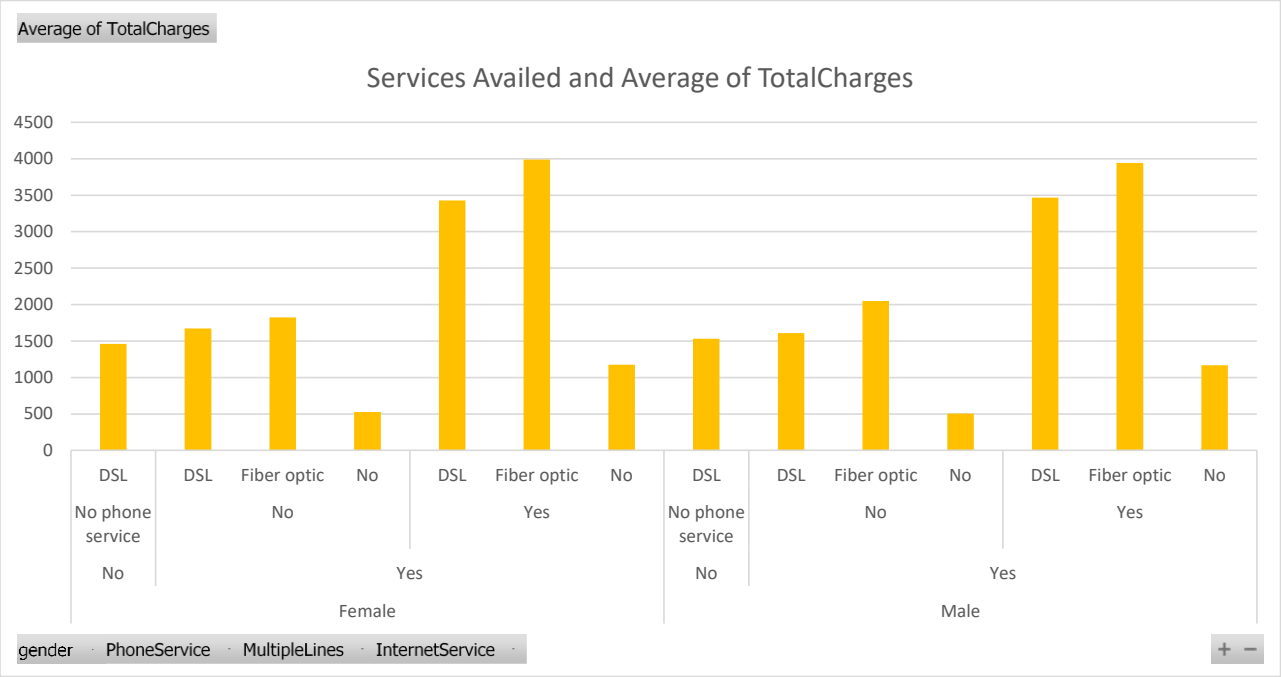
The effect of demographic attributes 'gender', 'SeniorCitizen', 'Partner', and 'Dependents' on 'TotalCharges' is shown in the chart below. The chart is split by the demographics in the sequence of firstly 'gender', then 'SeniorCitizen', then 'Partner', and finally 'Dependents'.



The effect of attributes related to services availed for 'PhoneService', 'MultipleLines' and 'InternetService', on 'MonthlyCharges' (by 'gender') is shown in the chart below. The chart is split by the services availed in the sequence of firstly 'gender', then 'PhoneService', then 'MultipleLines', and finally 'InternetService'.

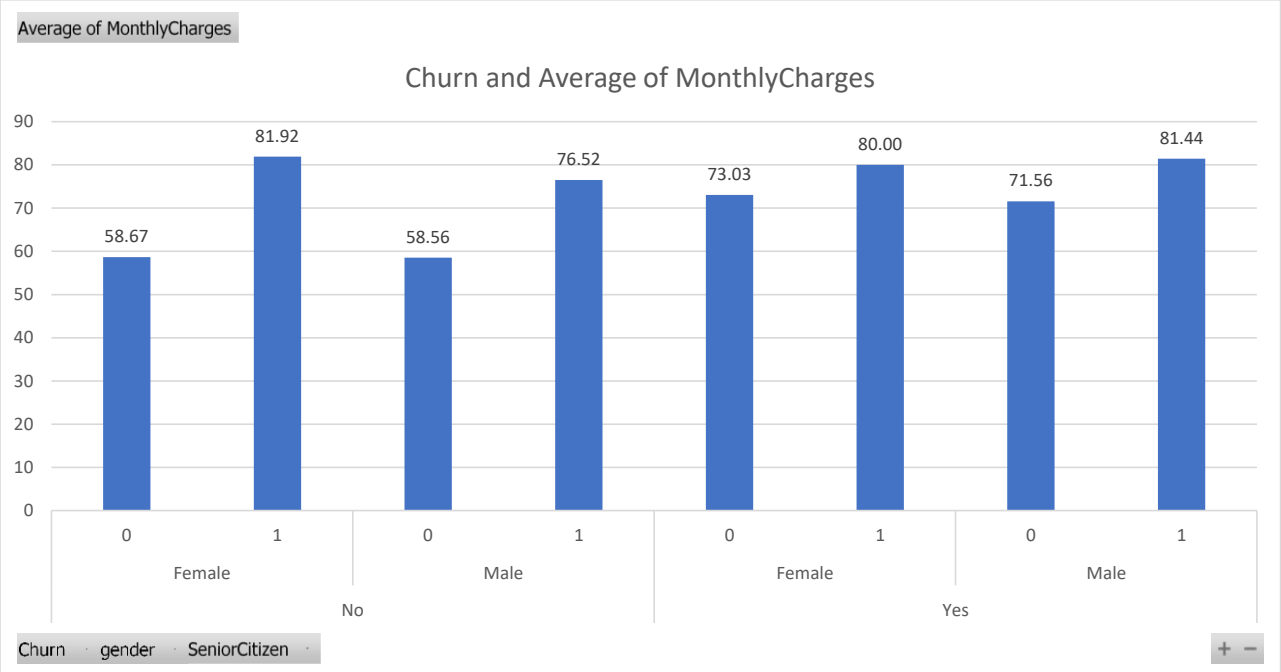


The effect of attributes related to services availed for 'PhoneService', 'MultipleLines' and 'InternetService', on 'TotalCharges' (by 'gender') is shown in the chart below. The chart is split by the services availed in the sequence of firstly 'gender', then 'PhoneService', then 'MultipleLines', and finally 'InternetService'.

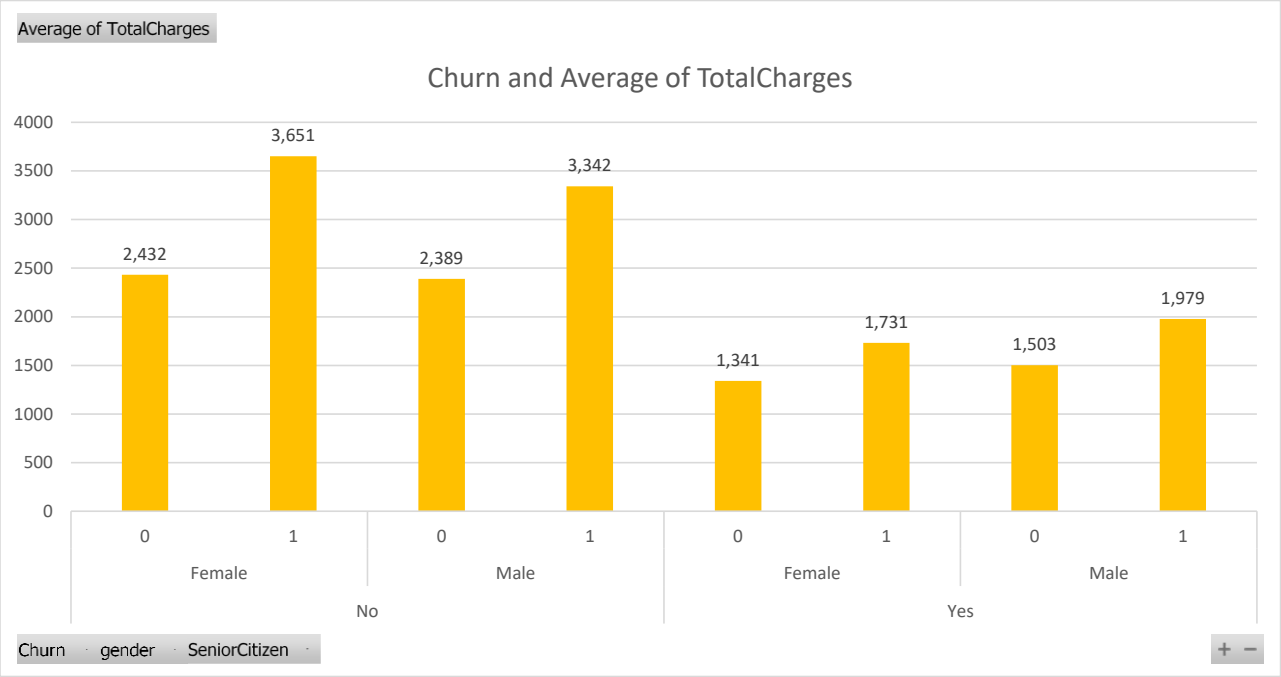


From the perspective of the 'Churn' of customers, we can see whether customers with higher or lower average 'MonthlyCharges' or 'TotalCharges' are likely to switch out from the telecom company, in the chart below.

The pair of charts below are split in the sequence of firstly 'Churn', then 'gender', and finally 'SeniorCitizen'.

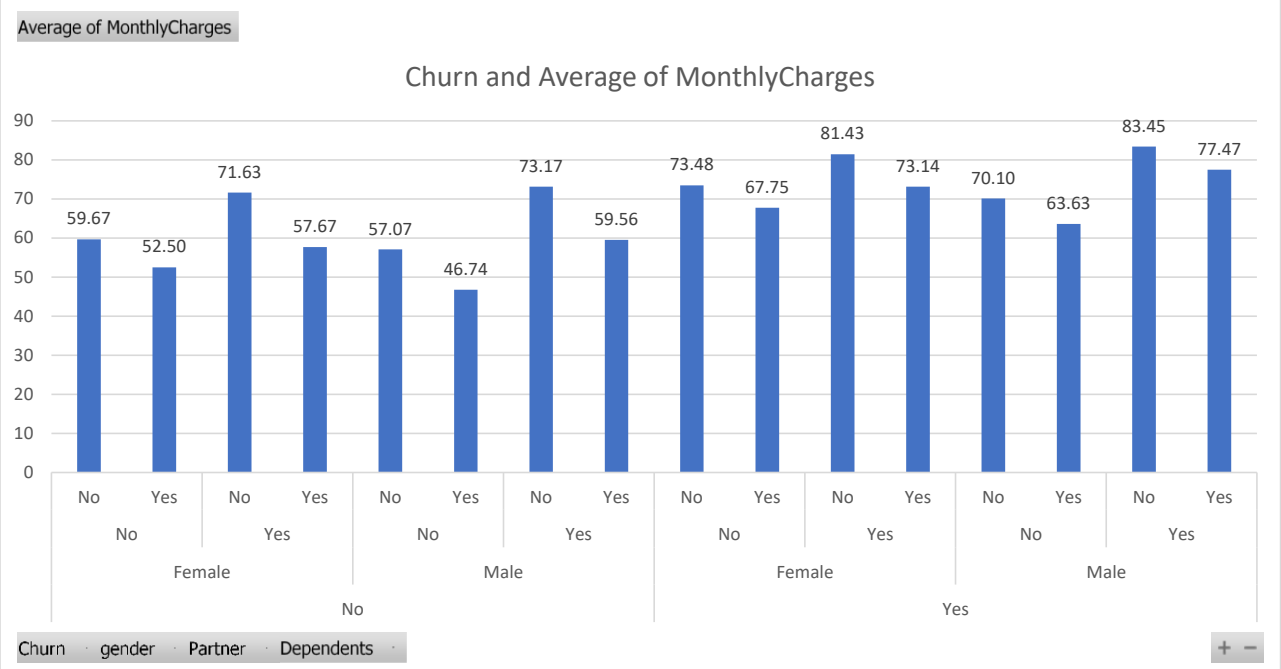


The chart above shows that the female senior citizens who are not churning have the highest average monthly charges of 81.92. In contrast, male non-senior citizens who are not churning have the lowest average monthly charges of 58.56.

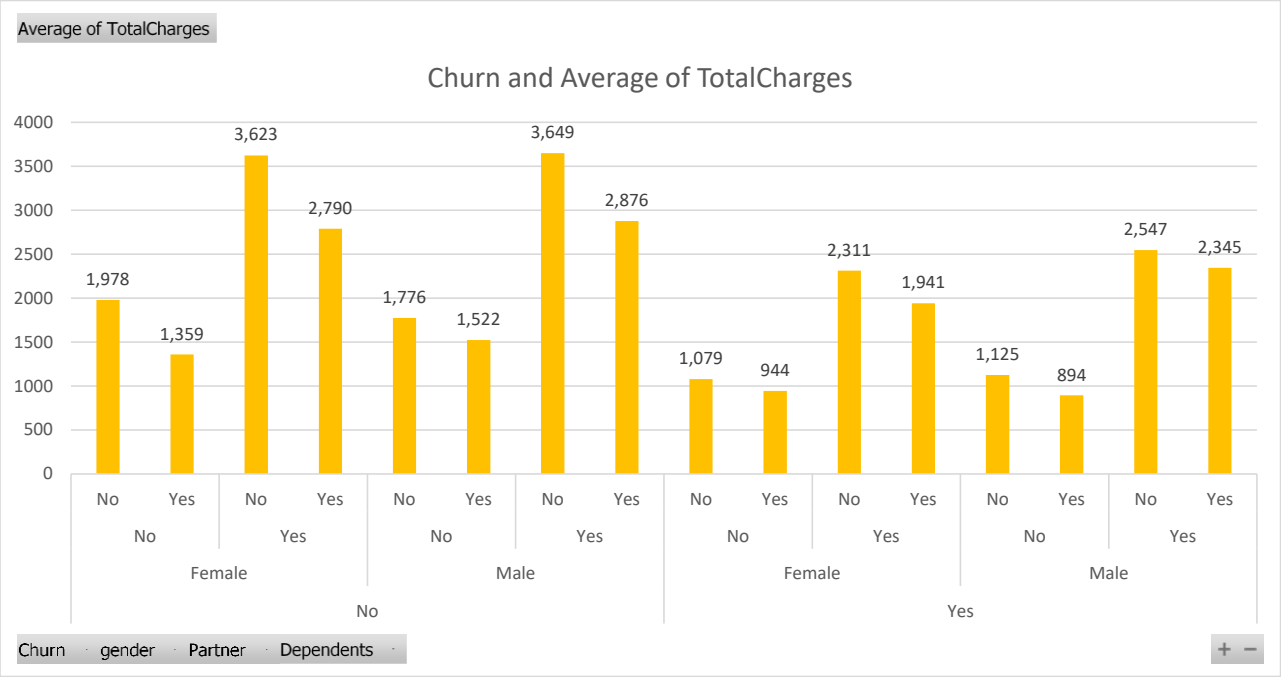


The chart above shows that the female senior citizens who are not churning have the highest average total charges of about 3,651. In contrast, female non-senior citizens who are churning have the lowest average total charges of about 1,341.

Another pair of charts below are split in the sequence of firstly 'Churn', then 'gender', then 'Partners', and finally 'Dependents'.



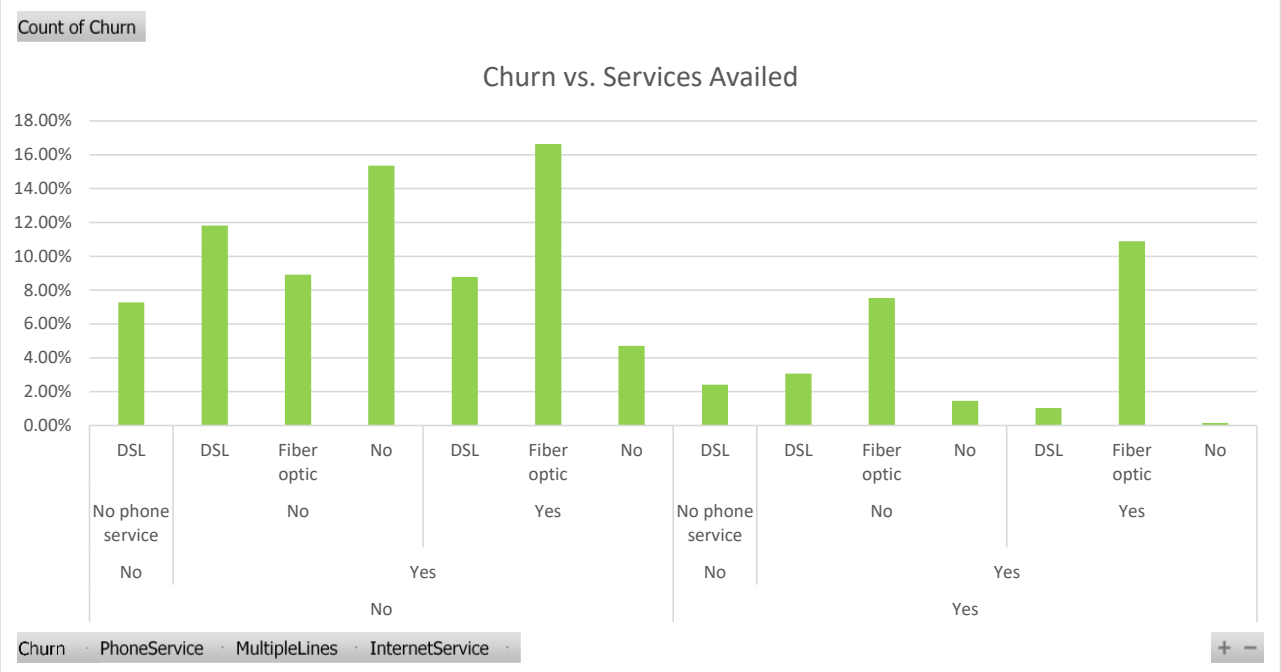
The chart above shows that males with partner but no dependents who are churning have the highest average monthly charges of 83.45. In contrast, males without partner but having dependents who are not churning have the lowest average monthly charges of 46.74.



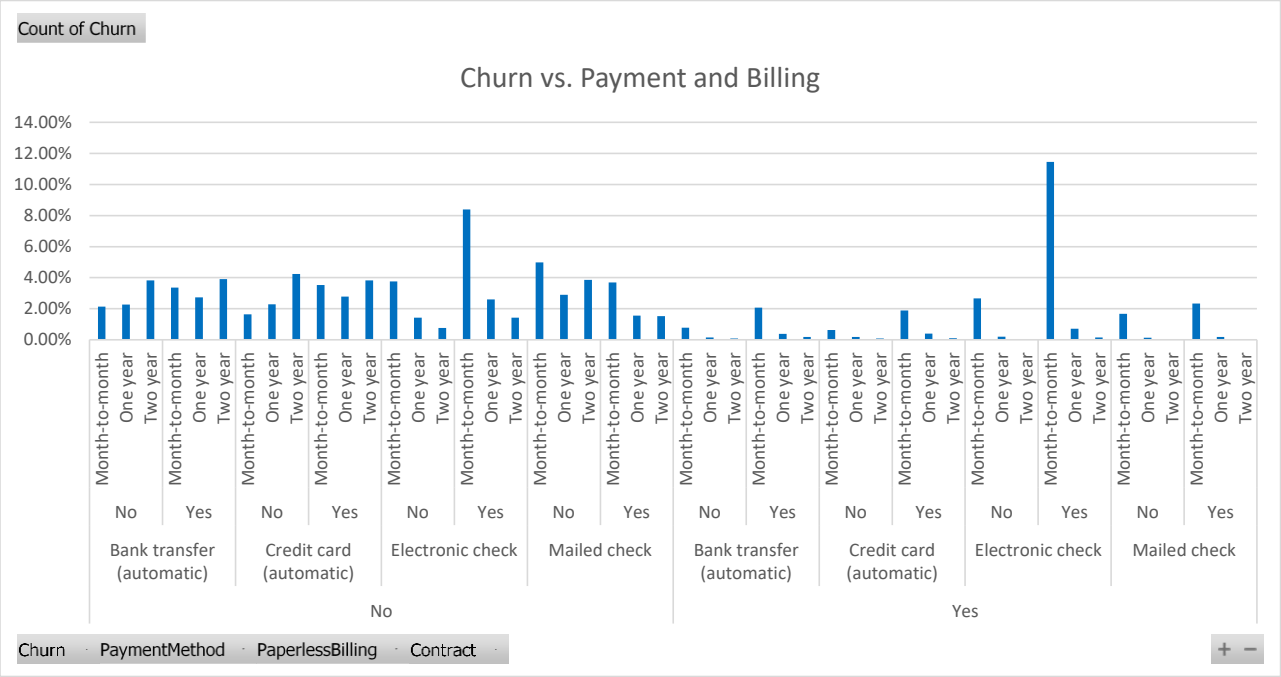
The chart above shows that males with partner but no dependents who are not churning have the highest average total charges of about 3,649. In contrast, males without partner but having dependents who are churning have the lowest average total charges of about 894.

In view of the above, the demographics have some contrasting effects of attributes on 'Churn', 'MonthlyCharges' and 'TotalCharges'.

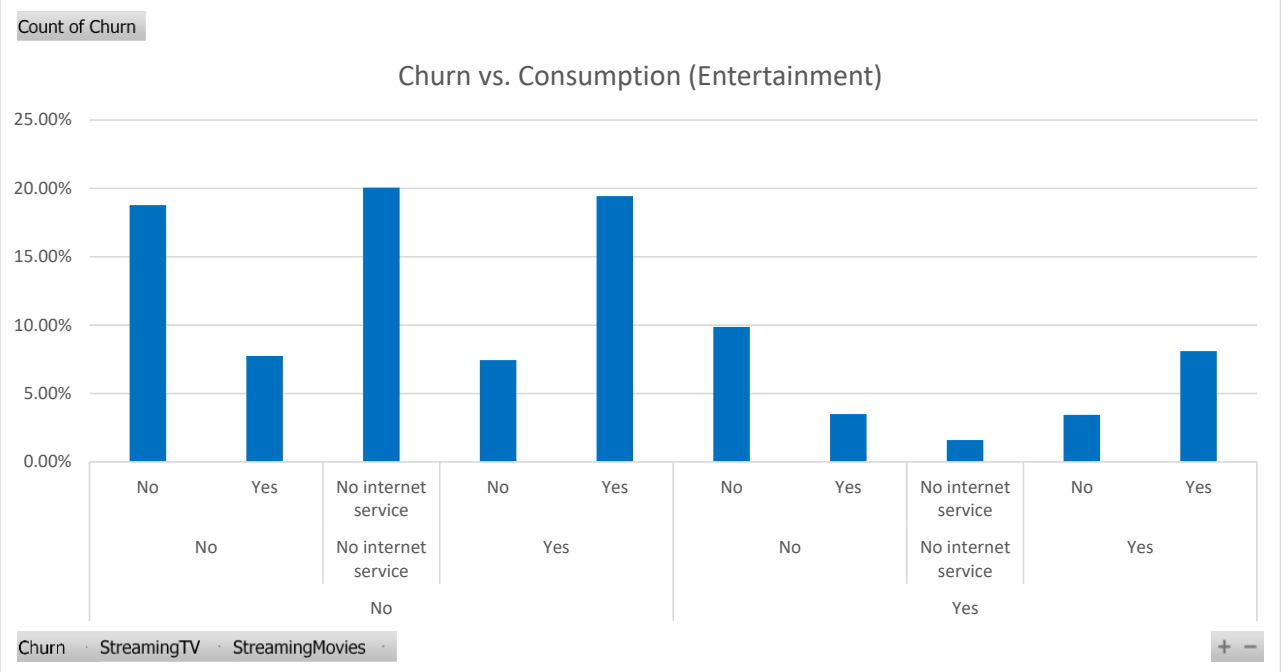
We can see the effect of attributes related to services availed on 'Churn'. The chart below is split in the sequence of firstly 'Churn', then 'PhoneService', then 'MultipleLines', and finally 'InternetService'.



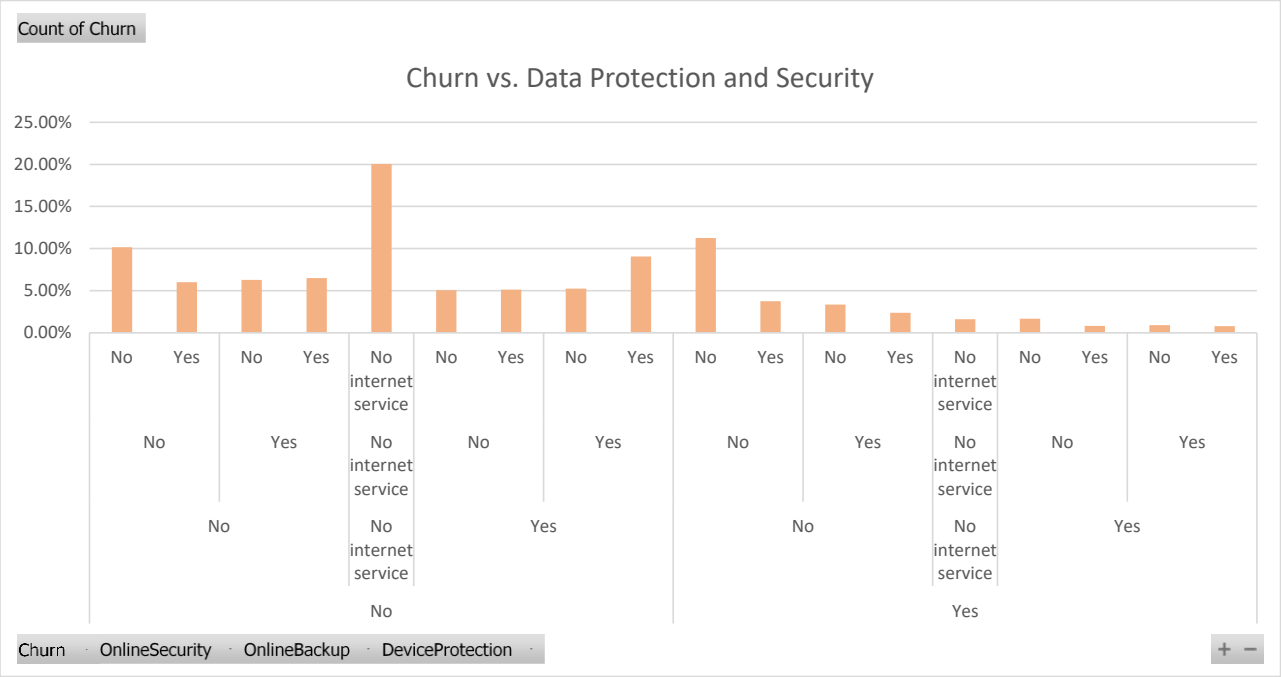
We can also look at the effect of payment and billing related attributes on 'Churn'. The chart below is split in the sequence of firstly 'Churn', then 'PaymentMethod', then 'PaperlessBilling', and finally 'Contract'.



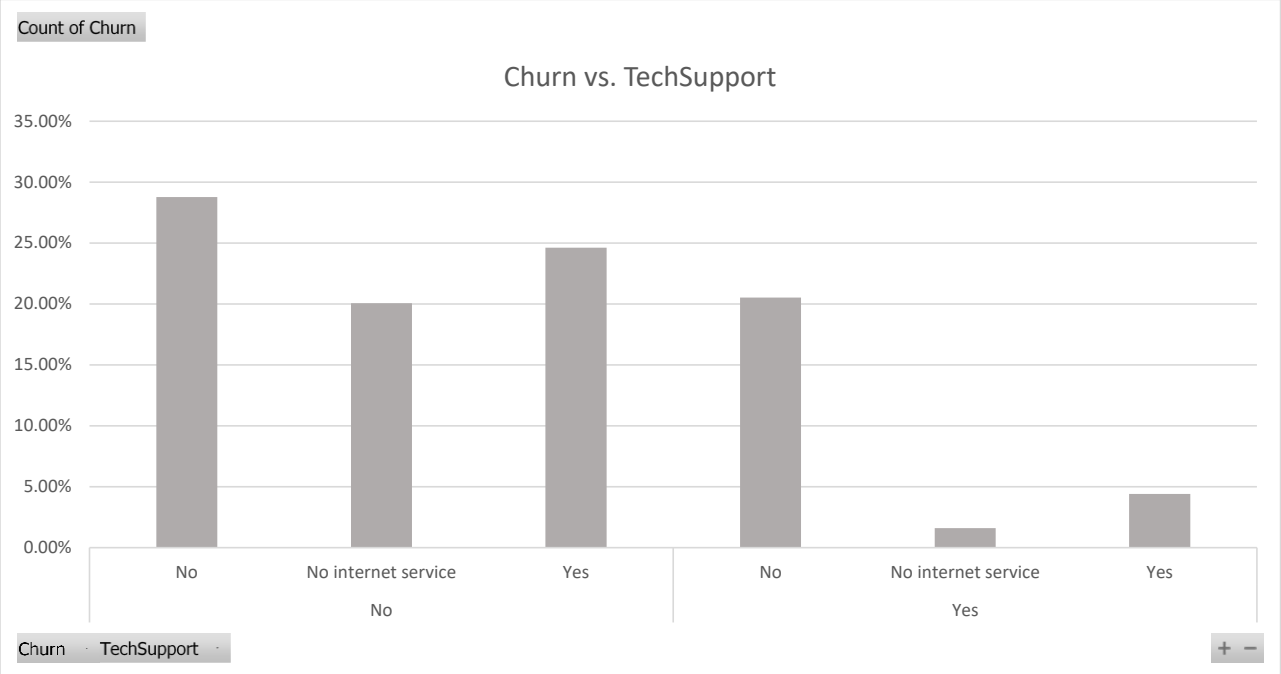
Further, we can see the effect of consumption (entertainment) patten related attributes on 'Churn'. The chart below is split in the sequence of firstly 'Churn', then 'StreamingTV', and finally 'StreamingMovies'.



In addition, we can also look at the effect of data protection and security related attributes on 'Churn'. The chart below is split in the sequence of firstly 'Churn', then 'OnlineSecurity', then 'OnlineBackup', and finally 'DeviceProtection'.



For the effect of technology support related attribute on 'Churn', the chart below shows 'Churn' against 'TechSupport'.



==========