IME 672 : Data Mining and Knowledge Discovery
Assignment 1
Devansh Kumar Sahu 190271

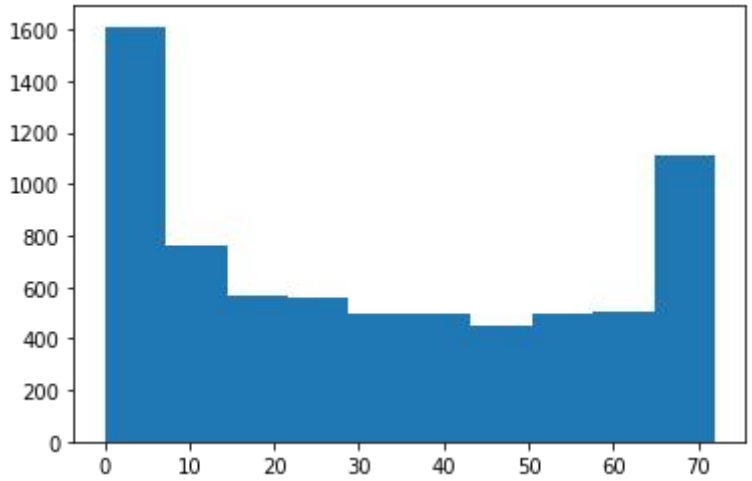
We have a dataset on whether customers of a telecom service provider will churn or not. From what we have been taught on the chapter 2 of this course, some interpretations are made to know more about the dataset.

Attributes

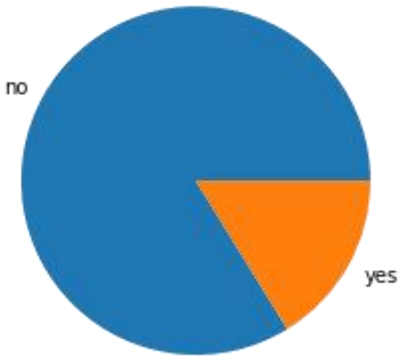
This dataset has 7043 cutomer IDs (objects) and 21 attributes (dimentions). Attributes include, customerID, gender, Senior Citizen, Partner, Dependents, tenure, Phone Service, Multiple Lines ,Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges, Churn.

Out of these attributes tenure , monthly charges and total charges are continuous attributes and rest are discrete attributes, either binary or categorical.

Refer to the following histogram and pie charts to know more.

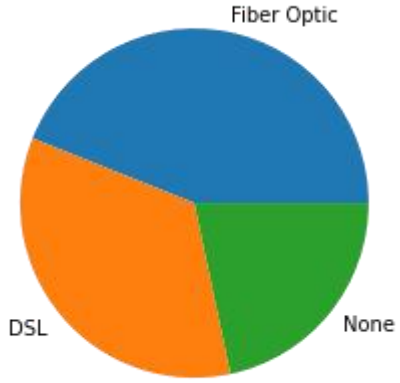


1. Distribution of Tenure

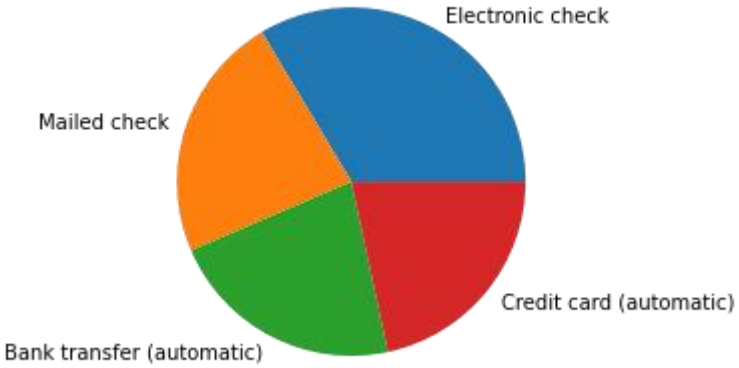


2.Senior Citizen

3.Contract Duration

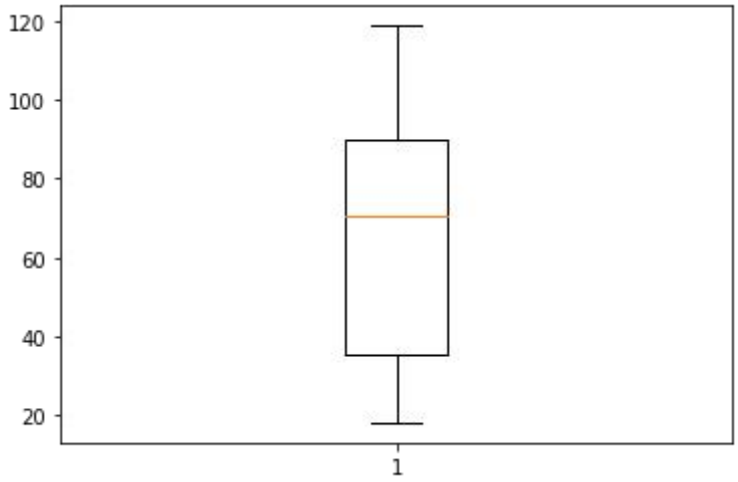


4. Internet Service



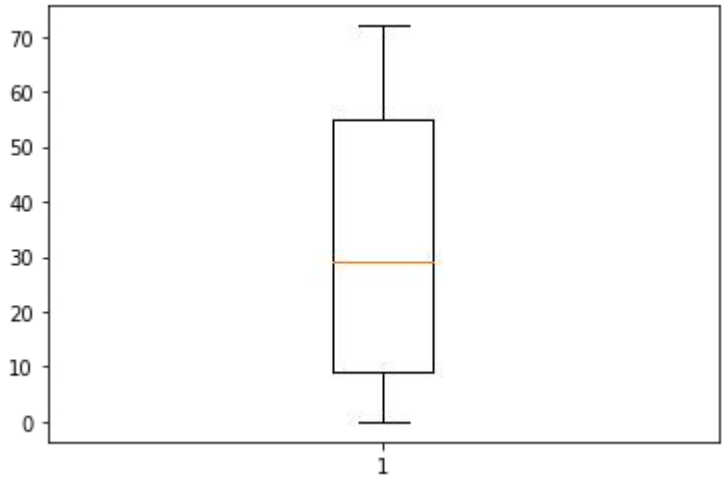
5.Payment Method

Boxplot analysis of monthly charges reveals that average monthly chage is 64.76 units, with first quartile of 35.5 units and third quartile of 89.85 units. Minimum and maximum monthly charges are 18.25 units and 118.75 units respectively. There is no outlier observed.



5. Boxplot For Monthly Charges

Boxplot analysis of tenure reveals that average tenure is 32.37 units, with first quartile of 9 units and third quartile of 55 units. Minimum and maximum tenure are 0 units and 72 units respectively. There is no outlier observed.



6. Boxplot For Tenure

Similarity between first 5 users is also calculated using Jaccard distance formula. Assymetrics attributes like, senior citizen, partner, phone service, online security/ backup etc. are used for finding out the similarity.

	0	1	2	3	4
0	0.000000	1.000000	0.666667	1.000000	0.8
1	1.000000	0.000000	0.666667	0.500000	0.8
2	0.666667	0.666667	0.000000	0.857143	0.4
3	1.000000	0.500000	0.857143	0.000000	1.0
4	0.800000	0.800000	0.400000	1.000000	0.0

8.Similarity Table for First Five Users