Name: Rishabh Arora **Roll No.: 190698**

IME 672A: Data Mining & Knowledge Discovery

Assignment 2: Telecom Churn: Data Preprocessing

We have provided the data of a Telecom Company and we have to analyze on which factors the customers are churning their service. So, first we Cleanse the Data Set and after that we Analyze the data on various attributes and find out in which case, they most likely to churn from the service.

1. Data Cleansing:

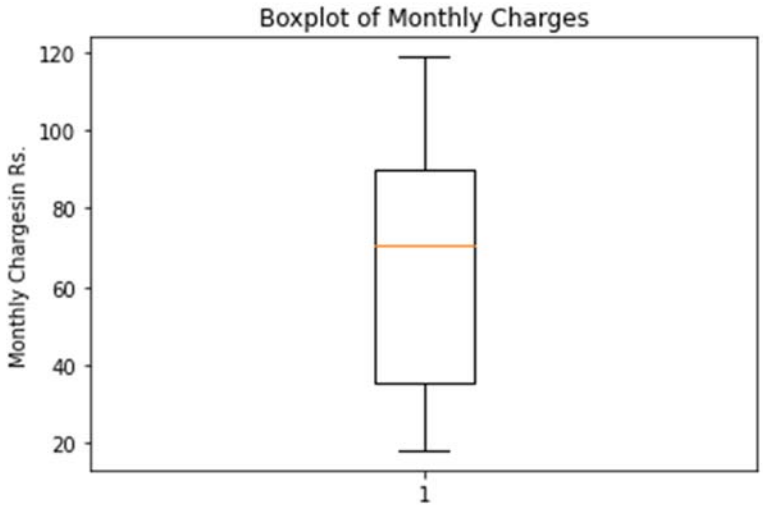
Missing Values:

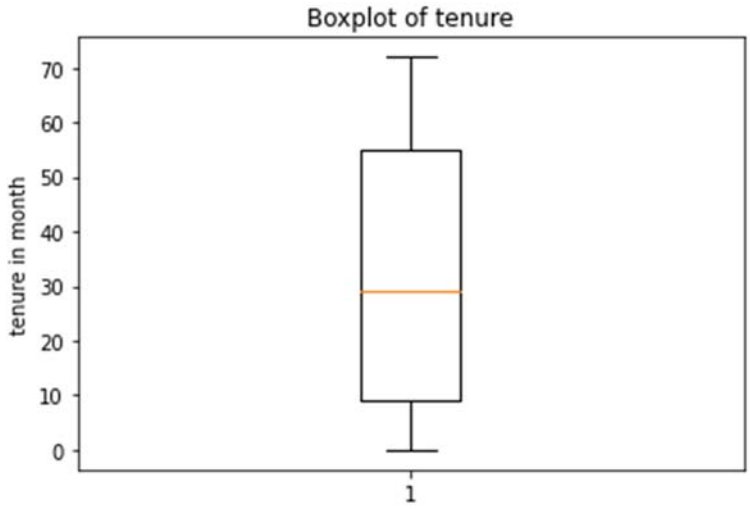
In the dataset we have provided the data of 7,043 customers with 21 attributes. Out of which 3 are numeric and 18 are categorical. After data cleansing, we convert 'TotalCharges' to numeric and got 11 null entries. After that we find out that the null values, we got have the customers with tenure of 0 month. So 'TotalCharges' of null entries is equal to 'MonthlyCharges' for these customers.

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2279.798992
std	0.368612	24.559481	30.090047	2266.730170
min	0.000000	0.000000	18.250000	18.800000
25%	0.000000	9.000000	35.500000	398.550000
50%	0.000000	29.000000	70.350000	1394.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

Noise/Outliers:

Change the "No phone service" and "No internet service" with "No" in different attributes.





There is no Outlier in the attributes ‘tenure’, ‘MonthlyCharges’, ‘TotalCharges’ with IQR method.

2. Correlation:

Chi-Square test:

For Chi Square test if p <=0.05 then the attributes are Dependent or reject null hypothesis(H0)
Chi Square test results for different attributes w.r.t Churn:

```
Chi-Square Test for gender
The Churns are independent on gender : Independent (fail to reject H0)
```

Chi-Square Test for SeniorCitizen
The Churns are dependent on SeniorCitizen : Dependent (reject H0)

Chi-Square Test for Partner
The Churns are dependent on Partner : Dependent (reject H0)

Chi-Square Test for Dependents
The Churns are dependent on Dependents : Dependent (reject H0)

Chi-Square Test for PhoneService
The Churns are independent on PhoneService : Independent (fail to reject H0)

Chi-Square Test for MultipleLines
The Churns are dependent on MultipleLines : Dependent (reject H0)

Chi-Square Test for InternetService
The Churns are dependent on InternetService : Dependent (reject H0)

Chi-Square Test for OnlineSecurity
The Churns are dependent on OnlineSecurity : Dependent (reject H0)

Chi-Square Test for OnlineBackup
The Churns are dependent on OnlineBackup : Dependent (reject H0)

Chi-Square Test for DeviceProtection
The Churns are dependent on DeviceProtection : Dependent (reject H0)

Chi-Square Test for TechSupport
The Churns are dependent on TechSupport : Dependent (reject H0)

Chi-Square Test for StreamingTV
The Churns are dependent on StreamingTV : Dependent (reject H0)

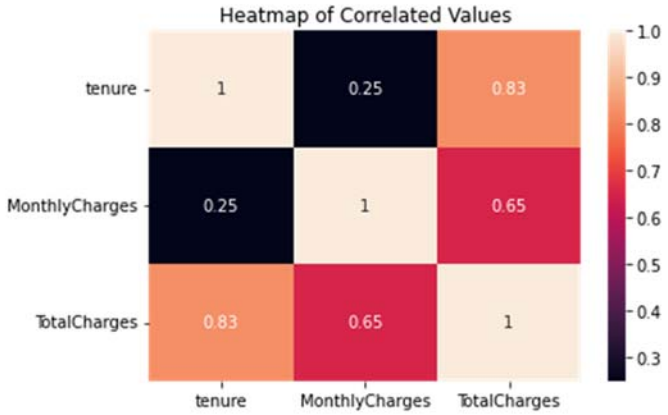
Chi-Square Test for StreamingMovies
The Churns are dependent on StreamingMovies : Dependent (reject H0)

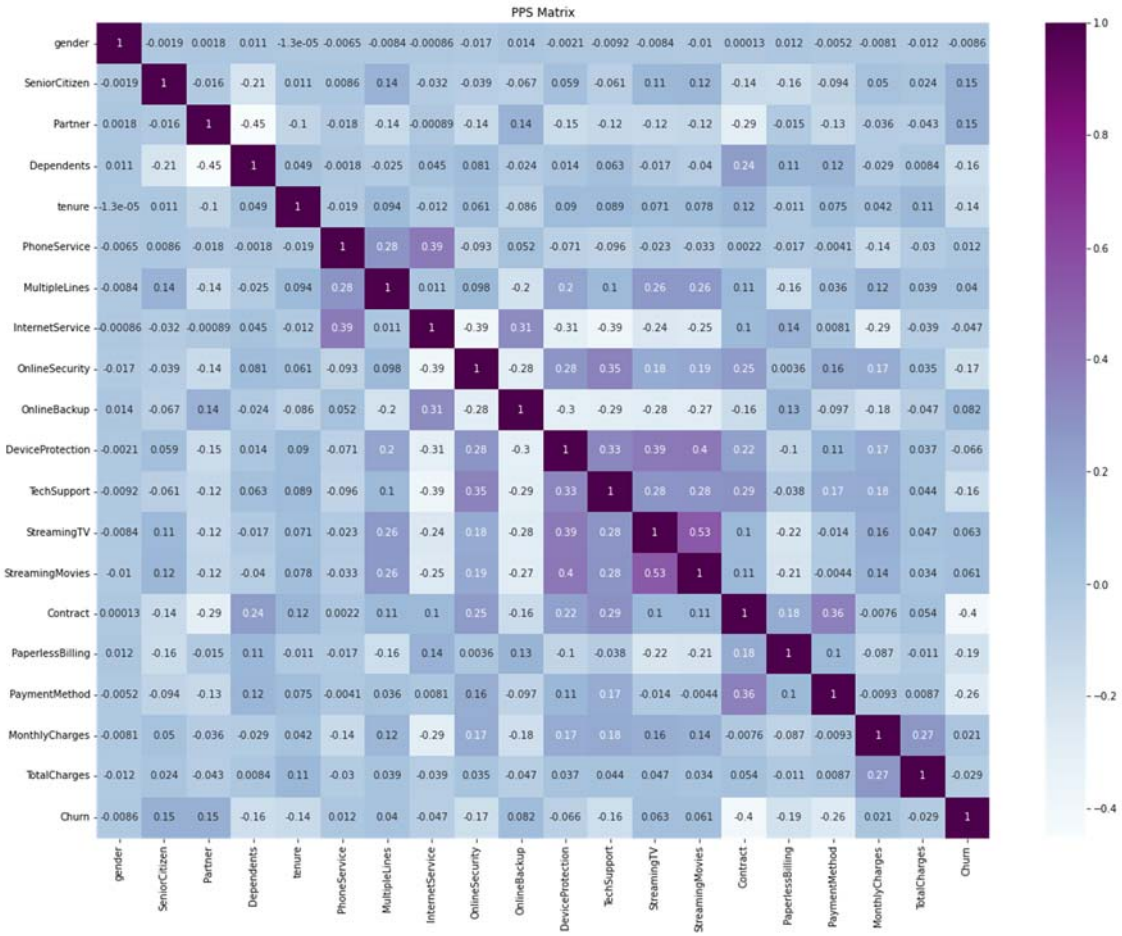
Chi-Square Test for Contract
The Churns are dependent on Contract : Dependent (reject H0)

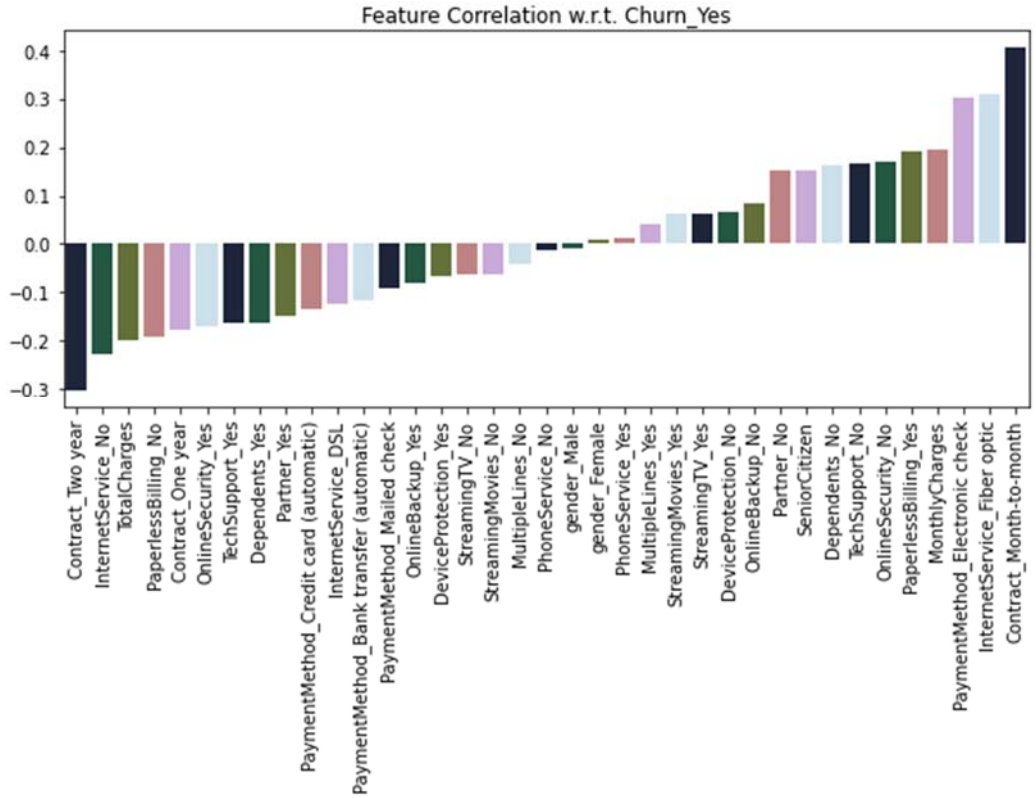
Chi-Square Test for PaperlessBilling
The Churns are dependent on PaperlessBilling : Dependent (reject H0)

Chi-Square Test for PaymentMethod
The Churns are dependent on PaymentMethod : Dependent (reject H0)

Chi-Square Test for SeniorCitizen
The Churns are dependent on SeniorCitizen : Dependent (reject H0)



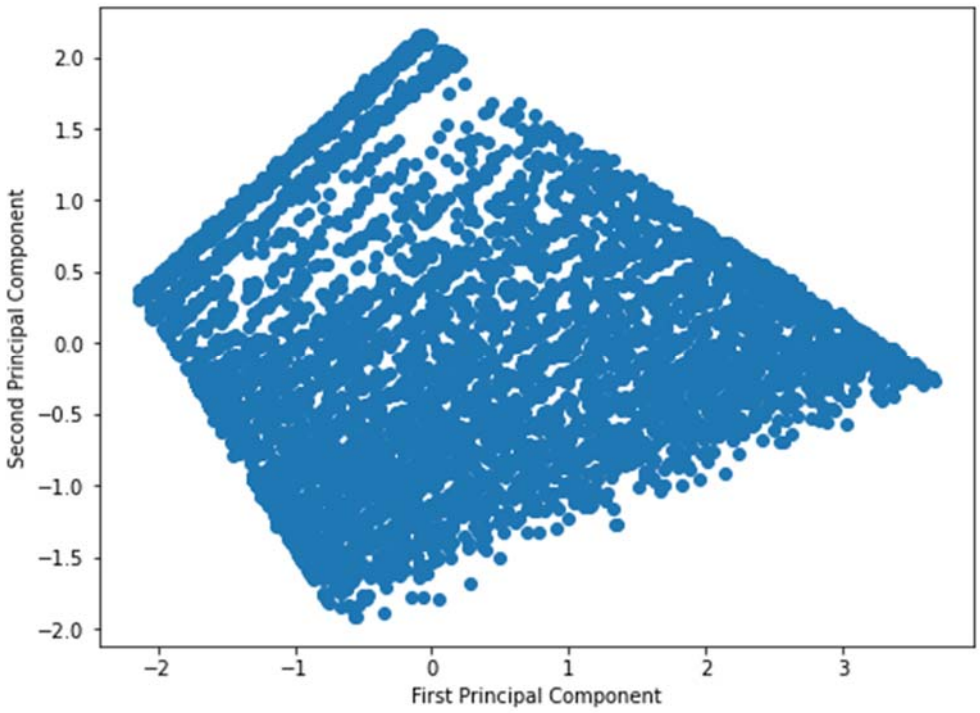

PPS Matrix
gender 1 -0.0019 0.0018 0.011 -1.3e-05 -0.0065 -0.0084 -0.00086 -0.017 0.014 -0.0021 -0.0092 -0.0084 -0.01 0.00013 0.012 -0.0052 -0.0081 -0.012 -0.0086
SeniorCitizen -0.0019 1 -0.016 -0.21 0.011 0.0086 0.14 -0.032 -0.039 -0.067 0.059 -0.061 0.11 0.12 -0.14 -0.16 -0.094 0.05 0.024 0.15
Partner 0.0018 -0.016 1 -0.45 -0.1 -0.018 -0.14 -0.00089 -0.14 0.14 -0.15 -0.12 -0.12 -0.12 -0.29 -0.015 -0.13 -0.036 -0.043 0.15
Dependents 0.011 -0.21 -0.45 1 0.049 -0.0018 -0.025 0.045 0.081 -0.024 0.014 0.063 -0.017 -0.04 0.24 0.11 0.12 -0.029 0.0084 -0.16
tenure -1.3e-05 0.011 -0.1 0.049 1 -0.019 0.094 -0.012 0.061 -0.086 0.09 0.089 0.071 0.078 0.12 -0.011 0.075 0.042 0.11 -0.14
PhoneService -0.0065 0.0086 -0.018 -0.0018 -0.019 1 0.28 0.39 -0.093 0.052 -0.071 -0.096 -0.023 -0.033 0.0022 -0.017 -0.0041 -0.14 -0.03 0.012
MultipleLines -0.0084 0.14 -0.14 -0.025 0.094 0.28 1 0.011 0.098 -0.2 0.2 0.1 0.26 0.26 0.11 -0.16 0.036 0.12 0.039 0.04
InternetService -0.00086 -0.032 -0.00089 0.045 -0.012 0.39 0.011 1 -0.39 0.31 -0.31 -0.39 -0.24 -0.25 0.1 0.14 0.0081 -0.29 -0.039 -0.047
OnlineSecurity -0.017 -0.039 -0.14 0.081 0.061 -0.093 0.098 -0.39 1 -0.28 0.28 0.35 0.18 0.19 0.25 0.0036 0.16 0.17 0.035 -0.17
OnlineBackup 0.014 -0.067 0.14 -0.024 -0.086 0.052 -0.2 0.31 -0.28 1 -0.3 -0.29 -0.28 -0.27 -0.16 0.13 -0.097 -0.18 -0.047 0.082
DeviceProtection -0.0021 0.059 -0.15 0.014 0.09 -0.071 0.2 -0.31 0.28 -0.3 1 0.33 0.39 0.4 0.22 -0.1 0.11 0.17 0.037 -0.066
TechSupport -0.0092 -0.061 -0.12 0.063 0.089 -0.096 0.1 -0.39 0.35 -0.29 0.33 1 0.28 0.28 0.29 -0.038 0.17 0.18 0.044 -0.16
StreamingTV -0.0084 0.11 -0.12 -0.017 0.071 -0.023 0.26 -0.24 0.18 -0.28 0.39 0.28 1 0.53 0.1 -0.22 -0.014 0.16 0.047 0.063
StreamingMovies -0.01 0.12 -0.12 -0.04 0.078 -0.033 0.26 -0.25 0.19 -0.27 0.4 0.28 0.53 1 0.11 -0.21 -0.0044 0.14 0.034 0.061
Contract 0.00013 -0.14 -0.29 0.24 0.12 0.0022 0.11 0.1 0.25 -0.16 0.22 0.29 0.1 0.11 1 0.18 0.36 -0.0076 0.054 -0.4
PaperlessBilling 0.012 -0.16 -0.015 0.11 -0.011 -0.017 -0.16 0.14 0.0036 0.13 -0.1 -0.038 -0.22 -0.21 0.18 1 0.1 -0.087 -0.011 -0.19
PaymentMethod -0.0052 -0.094 -0.13 0.12 0.075 -0.0041 0.036 0.0081 0.16 -0.097 0.11 0.17 -0.014 -0.0044 0.36 0.1 1 -0.0093 0.0087 -0.26
MonthlyCharges -0.0081 0.05 -0.036 -0.029 0.042 -0.14 0.12 -0.29 0.17 -0.18 0.17 0.18 0.16 0.14 -0.0076 -0.087 -0.0093 1 0.27 0.021
TotalCharges -0.012 0.024 -0.043 0.0084 0.11 -0.03 0.039 -0.039 0.035 -0.047 0.037 0.044 0.047 0.034 0.054 -0.011 0.0087 0.27 1 -0.029
Churn -0.0086 0.15 0.15 -0.16 -0.14 0.012 0.04 -0.047 -0.17 0.082 -0.066 -0.16 0.063 0.061 -0.4 -0.19 -0.26 0.021 -0.029 1
gender SeniorCitizen Partner Dependents tenure PhoneService MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies Contract PaperlessBilling PaymentMethod MonthlyCharges TotalCharges Churn
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


Feature Correlation w.r.t. Churn_Yes
0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3
Contract_Two year
InternetService_No
TotalCharges
PaperlessBilling_No
Contract_One year
OnlineSecurity_Yes
TechSupport_Yes
Dependents_Yes
Partner_Yes
PaymentMethod_Credit card (automatic)
InternetService_DSL
PaymentMethod_Bank transfer (automatic)
PaymentMethod_Mailed check
OnlineBackup_Yes
DeviceProtection_Yes
StreamingTV_No
StreamingMovies_No
MultipleLines_No
PhoneService_No
gender_Male
gender_Female
PhoneService_Yes
MultipleLines_Yes
StreamingMovies_Yes
StreamingTV_Yes
DeviceProtection_No
OnlineBackup_No
Partner_No
SeniorCitizen
Dependents_No
TechSupport_No
OnlineSecurity_No
PaperlessBilling_Yes
MonthlyCharges
PaymentMethod_Electronic check
InternetService_Fiber optic
Contract_Month-to-month

3. Data Reduction:

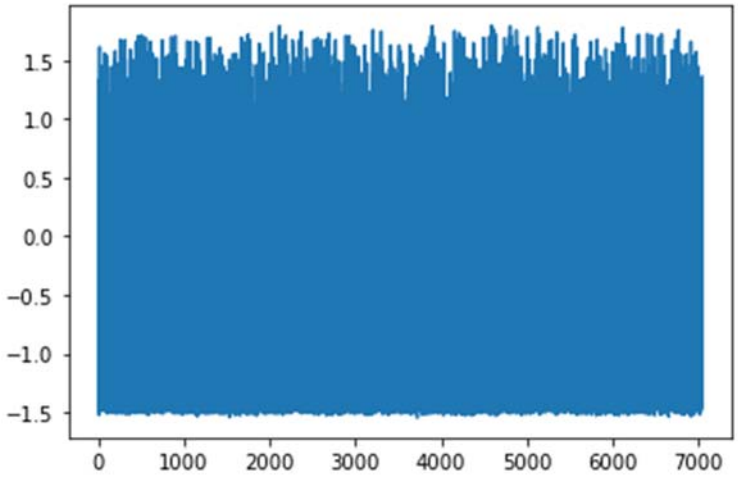
Data Reduction using Principal Component Analysis (PCA):

We reduce our data from 'tenure', 'MonthlyCharges' and 'TotalCharges' to PC1 and PC2.



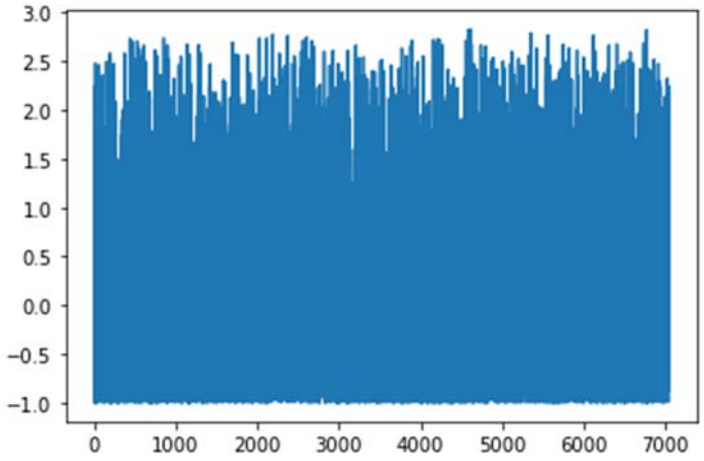
4. Data Transformation and Data Discretization:

Data Normalization using Z-score:



1. Z- Score for Monthly Charges:

2. Z- Score for Total Charges:



3. Z- Score for tenue: