DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Social Network Analysis of Fitness Club Data

Karkala Pulkeri Ashwin Prabhu

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Social Network Analysis of Fitness Club Data

Social Network Analysis auf Fitnessclub Daten

Author:	Karkala Pulkeri Ashwin Prabhu
Supervisor:	PD. Dr. Georg Groh
Advisor:	MSc. Salla Niskanen
Submission Date:	February 3, 2020

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

München, February 3, 2020 Karkala Pulkeri Ashwin Prabhu

Acknowledgment

First and foremost, I am forever thankful to Professor Dr. Georg Groh for recommending me to MSc. Salla Niskanen and supervising my thesis at the Technical University of Munich at the Chair for Social Computing Research Group. I could not have imagined having a better Supervisor for always helping me with continuous guidance and support.

My special thanks to MSc. Salla Niskanen, without whom this thesis would not have been a reality. I am grateful for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me a lot during my research and writing of this thesis. She has been the continuous moral support for this whole thesis. With every meeting, she would always help me keep on the right path by giving valuable feedback and insights on the data and still being available anytime I had questions.

I am thankful to Paula María Mejía De Miguel who helped me understand the background to get started with the thesis and always making time out of her busy schedule to help me whenever I had some queries.

I am also thankful for my team lead, Mr. Steffen Scherle, and my colleagues for supporting me during my stressful times of thesis and making the office a fun place to work. I can't feel more fortunate for all the opportunities given to me.

Last but never the least, I am grateful to my parents, brother, and family for their unconditional love, support, and belief in perceiving my Masters. And Almighty Lord Venkataramana for everything. My special thanks to Ranjitha Prabhu for always being there for me and motivating me. Thanks to my friends Abhishek, Preethi, and Juanita for taking the time to proofread this report and help me improve it.

Abstract

Exploratory data analysis, together with machine learning, are powerful techniques to deal with large amounts of complex and unstructured data. Furthermore, knowledge of graph theory coupled with social network analysis facilitates us to apply that on these data and extract undiscovered aspects of it. All these techniques are extensively used in almost every field.

A fitness club is a place where a lot of people work out, interact, and even get to know each other better. It can be viewed as a place where people get into social interactions and develop a friendship over time. This thesis focuses on this aspect of user interactions and uses it as a base to build a social network. It creates a social network based on user interactions and then derives different features out of the system, which describes the social activeness of individual users.

Later we try to compare it with other user behaviors like the visiting frequency and contract duration to see if these get influenced by the social activity of the user. We use the exploratory data analysis to understand the relationship and patterns between these data. Also, various machine learning approaches are applied to build a predictive model based on the social activeness of the user.

Contents

1 Introduction

Fitness and health sectors are some of the leading industries, with the annual market growth worldwide 4-5%: IHRSA 2019. Despite all these factors, the fitness industry faces a high churn rate. So researches have been done to see what factors influence users to terminate their contract at the fitness club and how it can be prevented.

The social network has been an essential aspect of human behavior and could be a critical factor in enhancing user retention. Social engagement enhances the user experience to the extent that they start experiencing the sense of belonging to a community and having an environment that is not judgmental on the actions performed. These factors greatly influence the user experience of a facility provided to her/him and finally make an impact on the user's decision to continue or terminate the contract [21]. The fitness and health industry has been investing a lot of effort into user engagement and gathering information that would give them an understanding of aspects that they fall short of to spike the customer's turnout at the fitness club. However, the idea and research on the fact that a customer can influence the visiting behavior of other members of the club are very minimum.

With this idea in mind evaluating the aspect of the social network in a fitness club environment and calculating the effects of the same over user behavior is going to help us improve the consequences it has on user engagement and preservation. We suggest that this unexplored possibility of a social effect on fitness and health is going to boost their success story.

The results of this study contribute in two main ways, first of it being the fact that this study is unique. Customer retention is a hot topic in the fitness industry, and a lot of research has been done and ongoing to see how it can be increased. But viewing a fitness club as a hub for social engagements and making a decision based on the discovered social network features is very limited. Secondly, generalizing the impacts of distinct social networks on user behavior is not encouraged since the peer to peer effects are constrained within themselves and are unique [2]. Most of the prior studies performed on social network analysis were based on the data readily available and easy to obtain. The data used in this study is a real-time data provided by a service provider. The information is collected from different sensors and the CRM database and depicts the customer's usage behavior. It is one of the first studies that utilize data that is not extracted from the web.

1.1 Data and Methods

The data used in this study is real-time and is provided by a fitness chain based in Finland. The data is extracted from various sources like sensors, CRM systems. The information has login footprints of about 144964 unique customers collected from the year 2013 till 2018. The different features derived from the data can be categorized into three main groups, namely spatial, temporal, and demographic characteristics, and these characteristics provide a broad understanding of user behavior at the fitness club. Machine learning techniques have played a vital role in understanding the essence of the data by detecting the patterns and predicting future outputs. Machine learning techniques train models based on the provided data and predict results for new data based on the learning from the trained model.

Machine-learning techniques are the best tool to gain knowledge from high-resolution behavioral data with an exploratory point of view [22] [12]. To investigate the effects of social networks on user behavior, we make use of such tools and predict if they contribute to user behavior. We start by building a social network utilizing the login data of the fitness club. We make use of the graph theory to build a social network and then gain essential features like the degree centrality, betweenness centrality, and closeness centrality from it. Once these features that represent the social behavior of every user are calculated, we try to see the effects of these features on user behavior. We calculate the user's contract duration and visiting frequency to describe her/his usage behavior and then compare them with the social network features and see if they have any correlation between them. We also performed a categorical analysis to see if any specific section of the user exhibits any unique behavior. All of these are noted and shown in the results of this study. Further, the measured social network features were exploited using machine-learning techniques also to see if they influence user behavior and can be used to predict the same.

2 Technical background

Contents

2.1 Exploratory data analysis

Exploratory data analysis [18] [1] is a set of methods performed in the initial phase of data analytics to understand the behavior of data, explore unknown patters, point out differences, make predictions and check the validity of it with the help of various visual depictions.

2.1.1 Visual plots

It is always helpful to know the data well before starting any study on it. There are various ways to describe the information visually. Some of them are listed below.

Line Plot

A line graph [23] is a type of graphical representation used to depict data as a sequence of points and connect these points with lines. Usually, a line graph helps us to see the trend in the data over a period [20]. In the [Fig. 2.1], we can see a line graph in a negative direction. It has the count of visits on the Y-axis and day of the week on the X-axis, and we can see that the count of the event is more on a weekday and decreases over the weekend.

Box plot

A box plot [26][24][18] is a graphical way to represent the variation of the dataset along with different quartiles[9]. It also shows the distribution of values of a variable corresponding to the different values of another variable. In the [Fig. 2.2], we can see the distribution of contract duration for different age groups. The line in the mid depicts the median value. The box range is where most of the data points lie, and the lines extending from the boxes, which are called whisker, illustrate the outliers that don't fall in the box range.

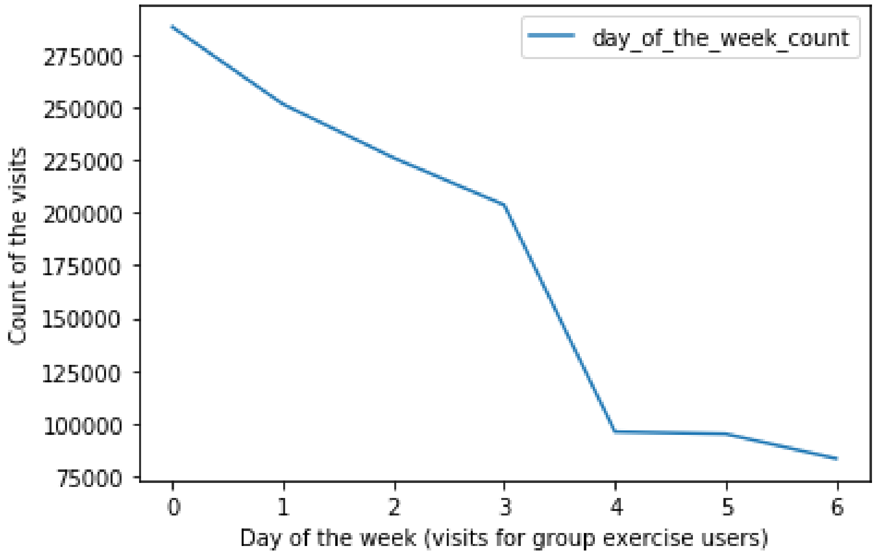


Figure 2.1: Line plot

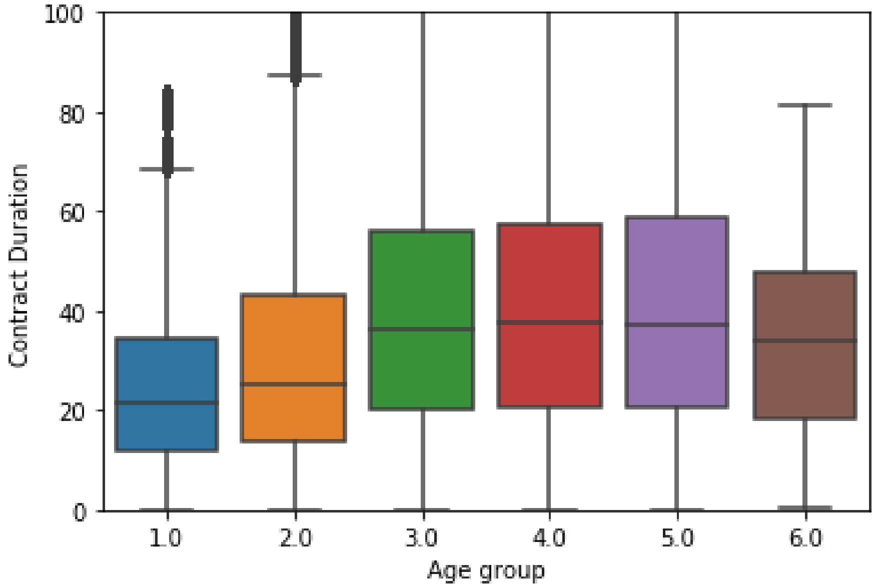


Figure 2.2: Box plot

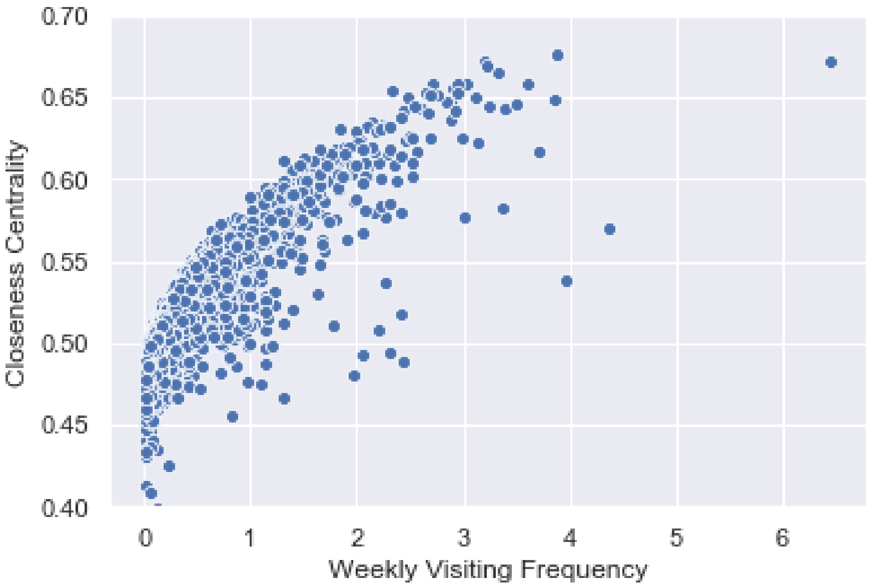


Figure 2.3: Scatter plot

Scatter plot

A scatter plot [8][11] is a graphical representation in the Cartesian coordinates system to show the distribution of data between 2 variables of a data set. Each dot in the graph depicts a data point of the data set. In [Fig. 2.3] we can see the correlation between the two variables.

Histogram plot

Histogram plot is a graphical representation to illustrate the frequency distribution of a feature in a data set. The data is split into different classes called bins. Each bin would have a fixed or varying width. The height of each bin depicts the frequency density. [Fig. 2.4] is an example of a histogram plot of the age group.

2.2 Social Network Analysis

Social network analysis[17][25] is a method of analyzing social features using graphs and network analysis. The social network consists of nodes and edges. The node in the

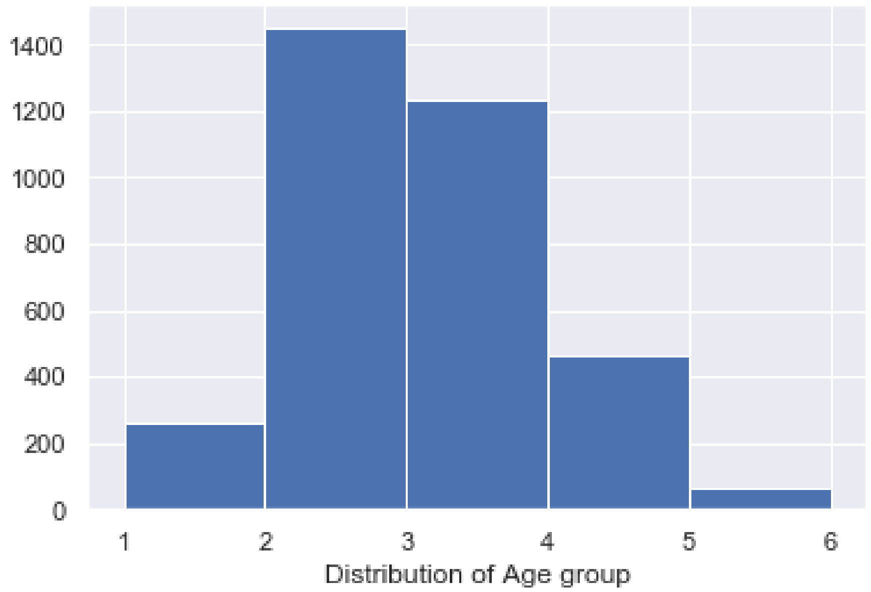


Figure 2.4: Histogram plot

network stands for the users, and edges are the social bonds between the users. Each edge will have a weight, higher the weight stronger is the relation between the user.

Various aspects can describe the social activeness of a node in a network. Centrality value justifies the importance of the node in the system. Higher is the centrality value of a node; higher is his contribution to the social network. There are several centrality values in a social network; some of the commonly used centrality values are Degree centrality, closeness centrality, and betweenness centrality.

2.2.1 Degree centrality

Degree centrality[7] is essential and the most simple centrality value to calculate from a social network. This centrality value signifies the number of connections a node has. In terms of network, number of edges a node has gives its degree centrality value.

The degree centrality [27] of a vertex a, in a graph $G := (V, E)$ is given as:

$$C_d(a) = deg(a) \tag{2.1}$$

2.2.2 Betweenness centrality

Betweenness centrality [6] of a node in a network computes the number of times it appears on the shortest path between two other nodes in the network. In terms of a social network, this value justifies the contribution of a user to help two different users in the system communicate with each other.

The betweenness centrality [27] of a node a, in a graph $G := (V, E)$ is given as:

$$C_b(a) = \sum_{x \neq a \neq y \in V} \frac{\sigma_{xy}(a)}{\sigma_{xy}} \tag{2.2}$$

σ_{xy} is count of shortest paths from node x to node y and $\sigma_{xy}(a)$ is the number of those paths that pass through a.

2.2.3 Closeness centrality

Closeness centrality [7] of a node in a network justifies how close a node is to all the other nodes in the network. The closer a user is to others in the network higher is the user's closeness centrality value. The closeness centrality of a node in a graph is the average length of the shortest path the node has with other nodes in the graph. So from this intuition, we can say as the node moves to the center of the network, higher it's closeness centrality gets.

The closeness centrality [27] of a node a, in a graph $G := (V, E)$ is given as:

$$C_c(a) = \frac{1}{\sum_b d(b, a)} \tag{2.3}$$

where $d(a, b)$ is the distance between vertices a and b

2.3 Interquartile range

The interquartile range [Fig. 2.5] of a distribution quantifies how disperse the data distribution is. IQR is least sensitive to outliers. We can calculate it by taking the difference between the first and the third quartile of the distribution. The IQR contains 50% of the data. From the [Fig. 2.6], we can see that a higher IQR value implies the data is well dispersed, and too much fluctuation from the median value.A smaller IQR value means the data is more stable.

$$IQR = Q_3 - Q_1 \tag{2.4}$$

Where Q_3, Q_1 are 1st and 3rd quartile respectively

2.4 Machine Learning

Machine learning[4] is a process of training a system to perform specific tasks without explicit instructions. The process involves pattern detection in the data and then using that as a base to learn and improve the results. This training process does not include much human intervention, as Artificial intelligence carries out everything. Machine learning algorithms, once trained with the data, generate mathematical models that can then predict outcomes of new data, or make better decisions with a new set of data.

2.4.1 Types of learning approach:

Supervised learning

Supervised learning [19] is a type of machine learning technique where both the input and desired output is known. Usually, a dataset is cleft into two sets, namely training and test set. The data that we use to train the model is called training data. We build a mathematical model using the training data based on the input and output values. Once we have the trained model, we use it to predict the results using test data. The results predicted using the test data set is matched against the actual value, and using this, we can estimate the precision of the model to predict the output.

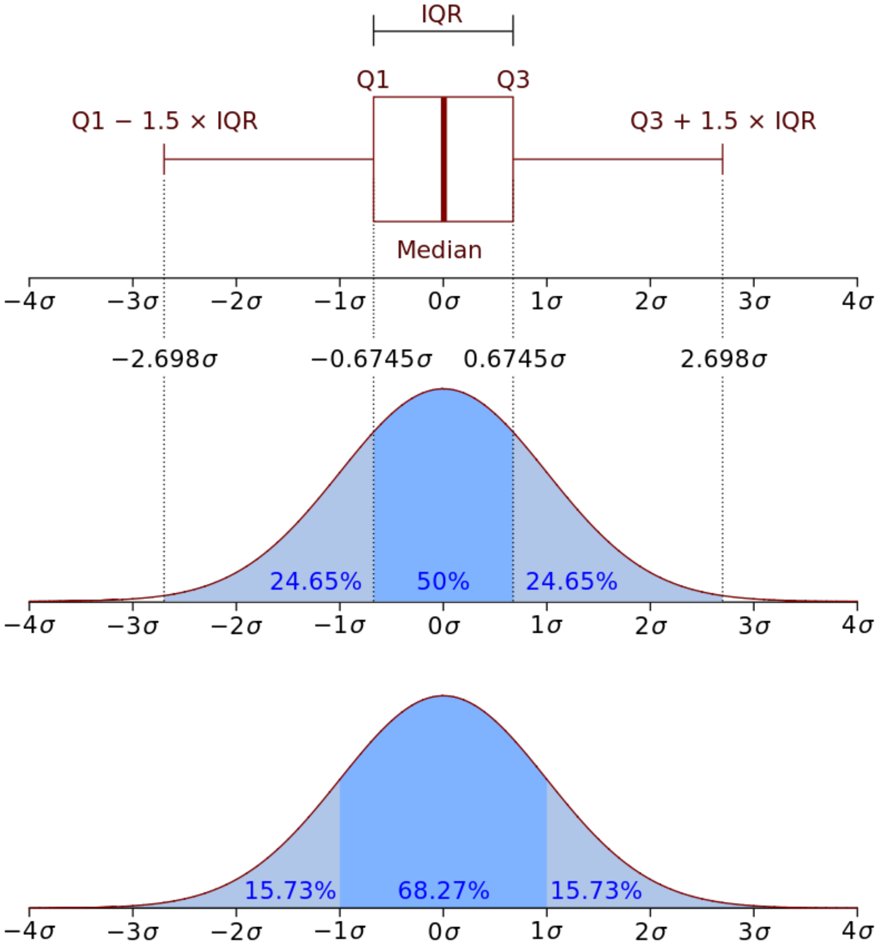
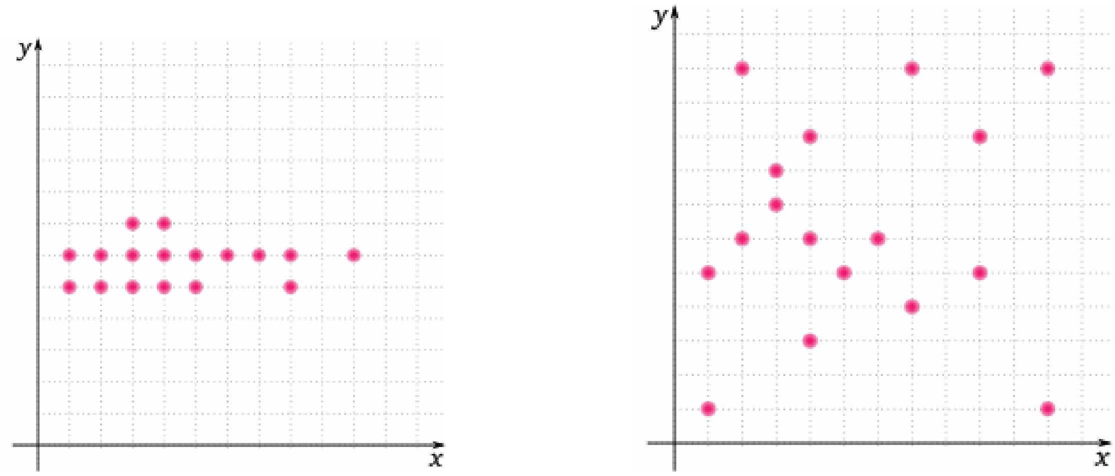


Figure 2.5: Interquartile Range (IQR) [28]



(a) The data points are close together; the Interquartile range is lower [14]

(b) The data points are widely scattered; the Interquartile range is high [13]

Figure 2.6: The data points distribution of higher and lower Interquartile range

Unsupervised learning

Unsupervised learning is an approach to train the system where the output is unknown. The algorithm aims to find commonalities and to realize hidden or unknown patterns in the data. It also uses a clustering approach to cluster data with similar behaviors.

Regression learning

Regression analysis is a method where we try to find a relationship between the input variable and the output variable by fitting a line to the data. The purest form of regression is linear regression, where a range is provided on the data to meet the condition of ordinary least squares. There are different varieties of regression like logistic regression, polynomial regression, kernel regression, each of which depends on the type of data available.

2.5 Dimensionality Reduction

Dimensionality Reduction is one of the critical steps to be followed in the machine learning process. Since the models are trained based on the data, the features used to

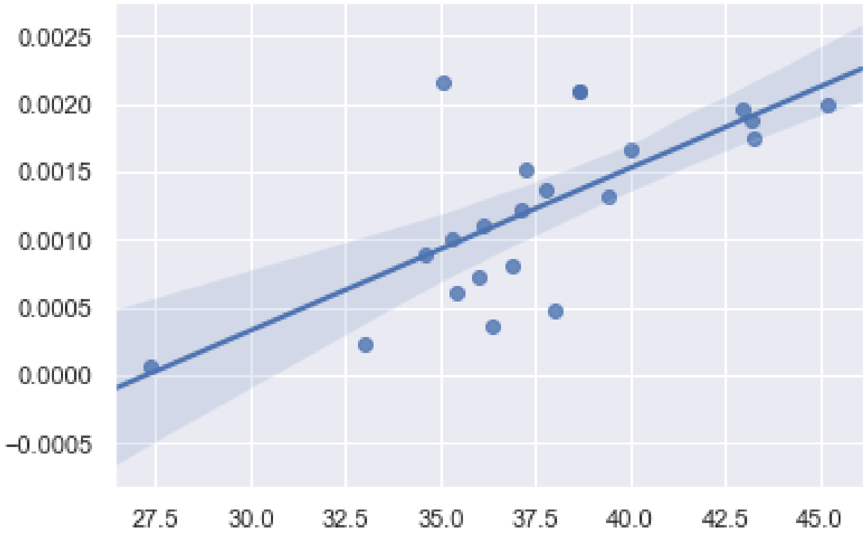


Figure 2.7: Linear Regression on a dataset

train must be accurate and unique. The data captured by the live system possesses a lot of redundant information in it. As a result, if we train a model with this data, the model would have a lot of instability, and prediction accuracy goes low. So, it's essential to make sure the feature set used to train the model be unique. There are a lot of techniques to reduce the dimensionality of the data and remove redundant features from it.

2.5.1 Backward elimination

Backward elimination is a process of removing the features from the dataset that does not make much of a contribution to the prediction process. In this method, we start by considering all the variables and calculate the variable of least significance. Removing this variable from the analysis should not have a substantial loss on the results of the statistical model. This removal process continues until all the features of the dataset are essential for building the mathematical regression model. This whole process is explained in the Algorithm 1

Algorithm 1 Backward elimination stepwise regression [5]

```
procedure BackwardElimination(set of all possible features)
    Set a significance level (SL) for a feature to stay in the model (e.g. SL =0.05)
    Fit the full model with all available features
    Consider the feature with highest P-value
    while P-value of the selected feature > SL do
        Reject the feature
        Fit the model without this rejected feature
        Select the feature with highest P-value
    return Set of unique and beneficial predictive features
```

3 Related work

Contents

This chapter outlines some recent and past research in the field of social network analysis, the importance and impact of it in various areas of day to day life. We discuss the related work accurately to show the critical role social engagement plays in different fields of life, and improving the same would help a service provider provide better service to users. We found this research fascinating, and it inspired us to apply a similar study on the social network aspect of fitness club data.

3.1 Social network analysis on fitness club data

Social network analysis has been a trending research topic across various fields of study. It has been seen that users get influenced by the social environment around them and make decisions based on that. Also, the effects of social engagement on user behavior of fitness club users were studied [15]. The study was made on different types of workouts at the fitness club, and the results for gym exercise were found to be more exciting. These results were then used to compare with the contract duration and monthly visiting frequency of the users, which showed a positive correlation.

3.2 Effects of the social relationship on terminating a service

The human behavior of adapting to something new is always influenced by the people close to them. It could be anything from ordinary to extraordinary things. If someone close to you likes it, it is a human tendency that you try it out too and then gradually get adapted to it. But the effects of the social relationship on churn rate is something that is not much looked upon. There are little researches that are focused on this aspect, and [16] is one of them. Studies so far show that there are a lot of factors that determine the churn rate. The way a customer uses his membership, his expectations from the service provided, and the satisfaction with the service contribute to a customer's churn. But significant evidence suggests that churn may also be by one's friend/colleague. We can see from the activities like hitting the gym, quit smoking, change of job, all depends on the social environment.

Inspired by such findings, a couple of researchers performed extensive research on the cellular data to realize that social relationships do play a role in a customer's decision to terminate a contract. For this experiment, 1 million customers' data like their calls, SMS were taken into consideration, and a social network based on their interactions was built.

Various experiments were carried out, a few of which are listed below.

- The research showed that social aspects profoundly influenced defection. With every neighboring customer defection, the focal customer was exposed to a 79.7% risk of defection.
- It was also found out that as the strength of social relationships between the users increases, the chance of focal customer defection also increases.
- From the study, it was found that a 1% increase in the strength of the tie between 2 customers, increases the chance of churn by 2.2% if the first person quits.
- The research was done on neighbors with similar attributes. The results showed that the risk of losing the focal customer is high when both the customers share the same characteristics.
- The effect of the neighbor's defection on the focal customer's defection percentage decreases over time, as shown in [Fig. 3.1].
- Loyal customers are less prone to defecting neighbors. The new customers get affected more quickly compared to customers from a long time as shown in the [Fig. 3.2]
- Highly connected users get affected easily but have a lesser influence on the churn rate of other users. The study shows that a user with ten neighboring bonds is 3.3% more prone to get defected when comparison with a user with just seven neighboring relationships.

There is another study on the churn model by considering mobile provider data [3]. In this study, the snowball sampling technique is used to get users with distinct user behavior and churn rate. The results obtained from this study is also supporting the results from the previous research and concludes that the social behavior of users does influence dropping out.

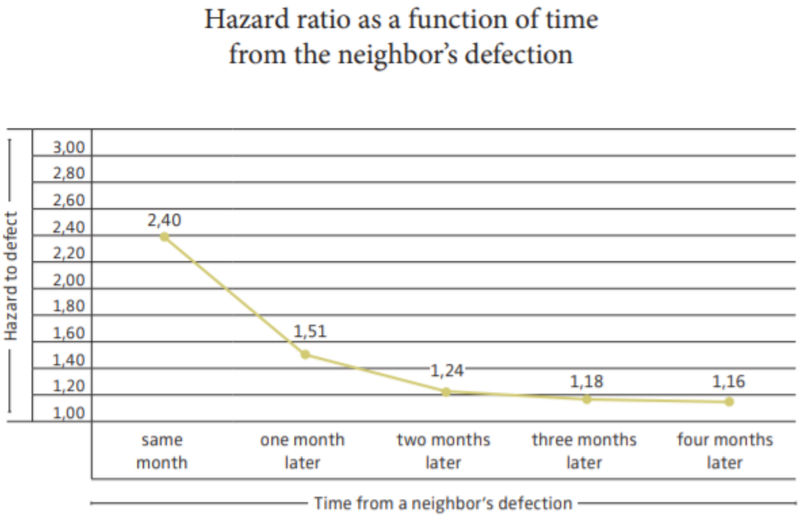


Figure 3.1: Hazard to defect of a focal user decreases over time [16]

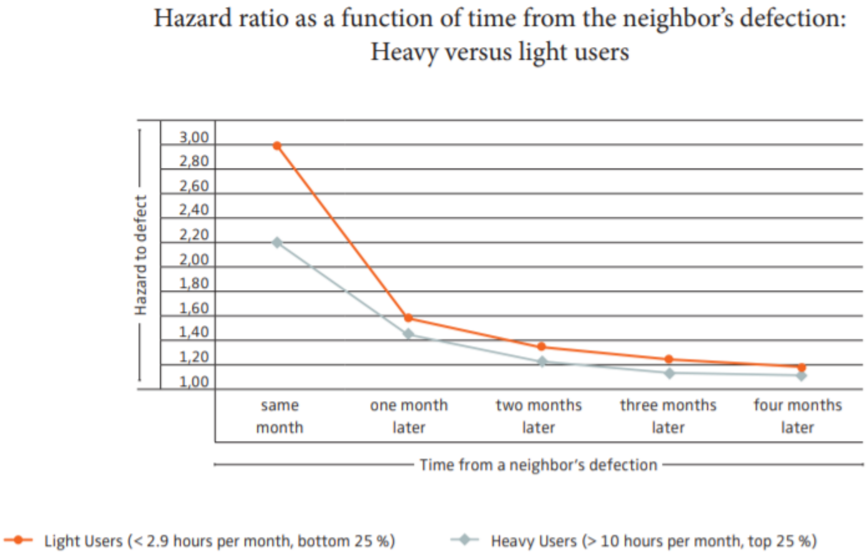


Figure 3.2: Hazard to defect of loyal users is less than new users [16]

4 Methodology

Contents

The main aim of this thesis is to understand the effect of social activeness of fitness club members over their usage behavior of fitness club contract. We build a social network using the login footprints of the fitness club users and calculate various attributes of the social network. Later machine learning techniques were used to see if these social network features influence user behavior.

4.1 Data pre-processing

The data used in this analysis was a real-time data. Hence it was required to pre-process the data before moving forward with any analysis.

Since the aim of the thesis was to find the social activity of the users based on their visiting behavior, the users with no or very fewer visits were filtered out. Also, users with a contract duration of less than a month were filtered out. Then we calculate contract duration in months for every user. Since there were few wrong data in the system with users having their start of the contract in the year 1800s and way ahead in the future, i.e., 2100, such data were omitted out to avoid outliers and miscalculations.

The structure of the data before the pre-processing step is shown in [Table 4.1] and [Table 4.2] depicts the resulting structure after applying pre-processing.

Total number of customers	144964
Number of login footprints	9092758
Number of fitness clubs	43
Sources	ERP, Sensor Data for Logins,Company CRM
Variables used in the study	Contract duration, login entries, age, gender and location.

Table 4.1: Shape of the data before the pre-processing

4.2 Exploratory data analysis

The first and essential step towards data mining is to understand the data. Exploratory data analysis (EDA) helps us interpret the main characteristics of data and how it varies numerically as well as visually.

Total number of customers	141918
Number of login footprints	8904910
Number of fitness clubs	43
Sources	ERP, Sensor Data for Logins,Company CRM
Variables used in the study	Contract duration, login entries, age, gender and location.

Table 4.2: Shape of the data after the pre-processing

It is essential to understand if there are any specific clusters or groups that exhibit any behavior in the fitness club usage pattern. The data obtained after pre-processing gives us an overall picture without distinguishing between different subgroups present in it. So, we had to classify the data based on various features before we start analyzing it.

We can classify the data into two groups based on the type of activity performed, the gym data, and the group exercise data. Further, we can again segregate these groups based on the features of the users. Since we know the gender and age group of each user, we used this information to create more refined subgroups.

Once we are done classifying user data based on various attributes, our next focus was to explore the behavior of each subgroup.

We listed out various features that describe a user's behavior of using the fitness club facility. The main features are their visiting frequency, the hour of the day they visit the gym, and their contract duration.

4.2.1 Visiting frequency of the user

Since the data that was received was in raw form and didn't have many features in it, we had to process it and deduce some features out of it. Visiting frequency is how many times a user visits the fitness club over a given interval of time. It could be on a weekly, monthly, or yearly basis.

We started off calculating per day visits of users using the login data to see how it varies over time. Then we extended it to weekly, monthly, and yearly to see the trend. We then plotted the variation in the overall visits of all users to the different time frames for each subgroup we defined initially. To get a better understanding, we calculated the minimum, the maximum, mean, and standard deviation of visiting frequency.

4.2.2 Contract duration

The contract duration of users is the number of days a user retained his fitness club membership. We calculate that by subtracting the contract start date from the contract end date.

We calculated contract duration for every user and then did a distribution plot for every subgroup to see the variation between each other. Also, we figured the minimum, maximum, mean, and standard deviation for each of the subgroups.

4.2.3 Visiting hours of the day

Visiting hour indicates the time of the day a user uses the facility. For every hour from 0 to 23, we calculate the number of user login entries and then do a distribution plot. We also compute the minimum, maximum, mean, and standard deviation for every subgroup. From this, we get an idea of the rush hours at the fitness club.

4.3 Social network analysis

The primary motivation to apply social network analysis at a fitness club is because of the understanding that people, when workout at the fitness club, tend to meet and communicate with familiar faces on a daily or weekly basis. This interaction would develop social bonding between them. We assume that these social interactions would influence the various user behaviors like visiting frequency and contract duration.

Since the users workout and interact with each other at the fitness club, it does not make sense to consider the overall data of 141918 users while building the social network. Also, it is not practical since its a lot of data and would need massive computing capacity. It is better if we choose one or two locations of the fitness club and perform social network analysis separately for each of the locations. We perform various exploratory data analysis to see which location would be ideal for conducting social network analysis. We saw the distribution of users based on their age at every location and selected a location that had an almost equal number of users across all age groups. This would ensure that the study didn't do favoritism and had a significant amount of users in each age group. Once the location was decided, we started with the procedure to build a social network. We use the python library Networkx[10] to build the social netowork.

We take into consideration the fact that every time two or more users use the facility at a current time, some amount of social interactions takes place between them. Using this concept, we build a network graph where users using the facility represent the nodes, and edges between these nodes represent a social interaction between themselves. More the number of co-logins between 2 users, stronger is their social bonding hence the weight of the edge connecting both the nodes. One best scenario to explain this is when users who know each other well would usually plan to visit the fitness club together.

4.3.1 Regularity factor

We calculate a regularity value for every user based on their visiting frequency. We can classify users into two types: Regular and Irregular users. A regular user is the one who maintains a consistent visiting frequency. Irregular users in contradiction are the once who don't follow a fixed routine and have an inconsistent visiting frequency.

To calculate the regularity factor, we compute the list of gaps between visits for all the users and apply the Interquartile range (IQR) [2.3] to it. A high IQR value means that the gap between visits is inconsistent, hence implying an irregular user [Fig. 4.2]. If the list contains almost consistent values, it can be considered that the user visits the fitness club regularly [Fig. 4.1].

4.3.2 Weight of edges

Calculating the weight of an edge is a crucial part of building a network. For this, we calculate the time difference between users' visiting time to the fitness club, and based on this value, we define a primary weight W to the edge between the users. The value is high when the time difference is low and decreases with an increase in time. We use an exponential decay function to calculate this value.

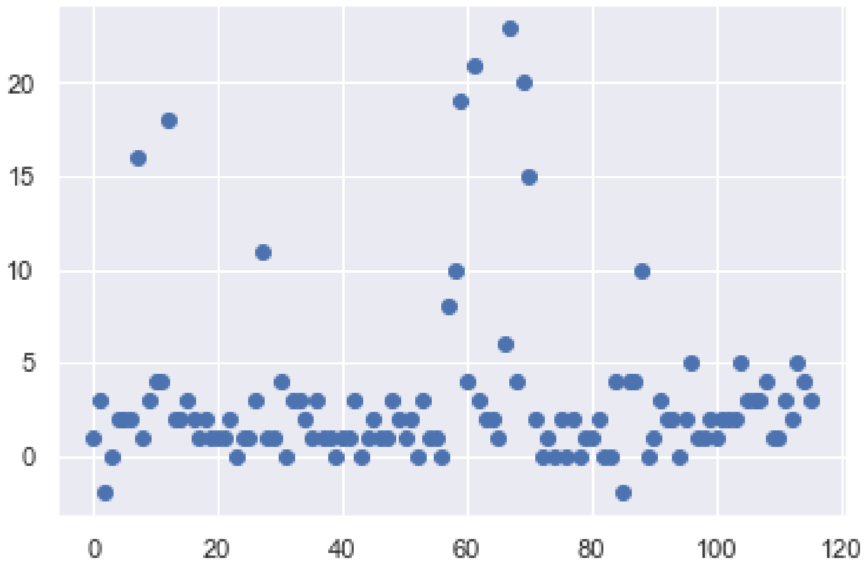


Figure 4.1: Gap between visits of a user with lower IQR. X-axis carries the number of visits by the user, Y-axis has the gap value between every visit.

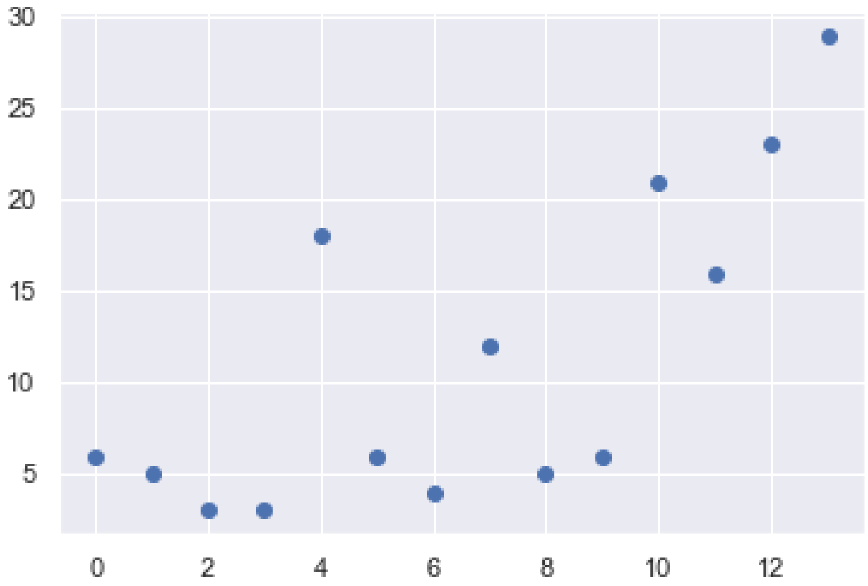


Figure 4.2: Gap between visits of a user with higher IQR. X-axis carries the number of visits by the user, Y-axis has the gap value between every visit.

$$W = 1 - \exp\left(-\frac{10}{|x|}\right) \tag{4.1}$$

Once we calculate the primary weight W using the formula 4.1, we introduce a regularity coefficient calculated using the formula 4.2 which is the product of IQR values of both the users. As we have seen in 2.4 that irregular users have high IQR value, if both the users are irregular, the regularity coefficient would have a higher value. Later we introduce this regularity coefficient into the weight function, as shown in 4.3. We use this term with an understanding that regular users always follow the same routine to work out, and there could be a coincidence of two regular users working out at the same time of the day but don't know each other or haven't spoken to each other. But when two irregular users have multiple gym entries with similar visiting times, this implies that there is a high possibility they know each other and plan to work out together at the gym. So after applying the following formula, we have a higher weightage for the social bond between two irregular users than the bond between two regular users.

$$\delta = IQR_{USER\_1} * IQR_{USER\_2} \tag{4.2}$$

$$W_R = W * (1 + \delta * W) \tag{4.3}$$

Where W_R is the weight after considering the regularity of users, and δ is the regularity factor.

4.3.3 Calculate Social Network Features

Once we calculate the weight for every edge between users, we move forward to compute the graph that represents the social behavior of the users. After the graph is ready, we calculate the essential features of it. We compute the page rank and the centrality values like degree, betweenness, and closeness [2.2].

Page Rank

Page rank of a node depicts its importance in the network. This value takes into consideration the degree of the node and its connection to other vital nodes in the network. The higher the value of page rank, the more significant is the node in the network. In the context of the fitness club, a higher page rank implies higher connections of the users with other co-members at the club. The page rank value of a node is directly proportional to the social activeness of the user.

Degree Centrality

Degree centrality is the most straightforward centrality value to calculate and understand. It is the number of connections a node has in the network. In the context of the fitness club, a user's degree centrality defines the number of social connections she/he has with other customers. It doesn't have to be a close friend but could also be a friend of a friend. The higher the value, the more people she/he knows.

Betweenness Centrality

Betweenness centrality of a node justifies its presence on the shortest path between 2 other nodes in the group. Higher value implies the existence in more paths. The nodes with higher betweenness value are the ones who play a significant role in information exchange in the network as they influence the shortest path. Betweenness centrality value of a user in a fitness club justifies her/his role as a bond maker between two other users who don't know each other. Higher the value signifies more roles she/he has played in bonding strangers at the fitness club.

Closeness Centrality

Closeness Centrality value of a node is the average distance of itself from other nodes in the network. Lower the average distance value higher its closeness centrality value. A node with higher closeness centrality value contributes more to the information transfer in the network. In terms of the fitness club, the closeness centrality of a user depicts how close she/he is to other users. More the number of users she/he is close to, higher is her/his closeness centrality value.

Once we calculate all the main features depicting the social network, we move forward with the analysis of the results and deciding on further steps over it. The foremost goal is to see the effect of social networks on the usage of the fitness club. So, once we compute the main features of the social network, we would like to build regression and predictive models that would predict the behavior of the user based on his social engagement at the fitness club.

We start by plotting the contract duration of the users to the centrality values calculated above. We could see a positive trend in the contract duration with the increase in the centrality values. From this, we can interpret that social activities influence the contract duration of the users positively. To understand the results better, we used various features of our results and classified each section of the data and the corresponding results. We split the data based on the type of exercise a user performed at the fitness club. Since we had data for gym and group exercise, we analyzed the output separately for each of these groups. We also had the gender of every user. So, we could classify users based on this feature and create two more groups and saw the trend in the variation of contract for every group. We also exploited the age feature in the data. We segregated users into five age groups based on their age. Later plotted the above graphs for every age group to see if there is any unusual pattern in the user behavior based on the age group of the user.

Once we plot the graph, we wanted to see the distribution of the contract duration of the users. We started by plotting an overall distribution of contract duration. Later, to dig deep into the data, we used various available features like gender, age, and type of exercise and plotted the distribution graph. We also wanted to see how the distribution varies across various centrality values. For every centrality type, we split the users based on the different slab of values and plotted the distribution of the contract duration for every slab. All these results were cumulative and gave us an overview of the complete system. We wanted to see the trend in the results, so we calculated the rolling mean of contract duration for every centrality type. We define the size of the rolling window and apply it to the centrality value. We take into consideration all the points that lie in the current window and then calculate the average of all these points. Once completed with all the points in the current window, we slide it to the next and continue the same process. We start from the lowest point until we cover the point with the highest value. Once completed, we plot the average value of contract duration to the average centrality value for every window.

We later wanted to analyze more on user behavior, so we compared the monthly visiting frequency of users to the social network features. We followed all the processes carried out earlier for the contract duration of the users and applied it to the monthly visiting frequency.

4.4 Overall visiting pattern

It was necessary to see how a user's visiting behavior changes over time from the day he joins the fitness club. The overall visiting pattern of the data calculates the frequency of fitness club usage and variation over time. To start, we calculate the visiting frequency of the first month for every user and then take the average of it. Similarly, we calculate the visiting frequency of every month till the last month. This list of all the average values starting from the first till the last month is the overall visiting pattern of the fitness club. The overall visiting pattern appears to be decreasing over time, meaning the users tend to reduce their visiting frequency over time. So, we wanted to see if this is a general case or if this varies based on the type of users and users with different centrality values. We took into consideration the different features of the users like their gender, age, and the exercise they do. We split the user set based on these features and then calculated the overall visiting pattern. We also used the centrality value and calculated the overall visiting pattern for different slabs. We could see that as the centrality value increases, the overall visiting pattern stays constant, implying the users get more motivated to visit the fitness club as they get more socially involved with the co-users at the club.

4.5 Churned percentage of friends

Next, we need to calculate the churned percentage of friends for a customer who has quit her/his contract. It is the percentage of friends who have also discontinued their contracts. We wanted to calculate the extent to which a person by quitting her/his fitness club contract influences her/his friends to cancel their contract. This value would help us understand the fact if socially active users control the fitness club membership of his friends at the facility or not.

To calculate the same, we start by selecting all the users who have already quit their membership. Then we analyze their contribution in the social network, which is already computed earlier by computing various centrality and page rank values. Once we retrieve these values, we get a list of close friends for every churned user, and from that, we calculate the percentage of users in the list who have also terminated their contract. By this, for every churned user, we get the percentage of friends who have also discontinued their contract.

Once we have calculated all the required details, we start analyzing these values. We start by plotting a graph of churn percentage of friends to various centrality values computed and try to see if it shows any specific trends. We applied the rolling mean over these plots and saw a positive trend in the value. As the centrality value of the user increases, the churn percent of friends also increases. Looking at the results, we could conclude that a user who is more socially powerful influence his friends' contact duration. We separated the users based on their centrality values and divided them into four groups from lower centrality to high centrality value. We then did a distribution plot for every user group. The results showed that the standard deviation of the churn percent reduces as the centrality value increases. Also, the mean percentage of churned friends increases with an increase in the centrality value. These observations add to the assumption that as a person gets more socially influential, he can dictate his friends' contract duration.

4.6 Machine Learning

The primary motivation to apply machine learning techniques is to build a predictive model using the available features. Since we have features that define various aspects of the user like his age, gender, visiting frequency, social networking features, we wanted to make use of these data and try to predict a pattern and see if they have an impact on the users' behavior.

Machine learning has various types of machine learning algorithms, all of which fall into three main categories, namely.

- **Supervisor learning:** is training a model based on labeled data. The data provided during the training phase guides the model. Once we train the model, it is fed with test data to check the accuracy of the model
- **Unsupervised learning:** , on the other hand, doesn't have labeled data. The model tries to figure out various patterns and structures on its own and learns the data. It tries to classify data into clusters based on similarities.
- **Reinforcement learning:** is a type of training process where the model interacts with the environment itself to maximize fruitful outcomes/results.

4.6.1 Regression

Regression is a type of supervised learning. Since we had different features that describe the user data efficiently, we wanted to see with the help of these features if we could predict the contract duration of the user. We could see a high correlation between centrality values and the visiting frequency of the users. So, we fitted a regression line to each of the centrality values and calculated the coefficient values for the same. The closeness and betweenness centrality values converge with a second-order polynomial regression, but degree centrality goes up to 6th order polynomial.

Once we did a line fitting, we wanted to train a model to see how well it predicts a user's contract duration based on the features available. Since we had a lot of available features in the data, it was essential to filter out the ones that are not so useful or redundant. We used the backward elimination method to remove all unwanted features. Once we achieve this, we started training the regression model. We trained three models with different algorithms, namely Linear regression, decision tree regression, and random forest regression. From all these algorithms, the results obtained from the decision tree regression was having high predictive accuracy.

5 Results and Observations

Contents

The outcome of this experiment was to see how social activeness of a user influences her/his fitness club usage. In this study, we try to evaluate various users' behavior like her/his weekly or monthly usage frequency, the usual hour of the day she/he does the training, her/his contract duration. And then see how these vary with social activity. We build a social network and calculate the various social network features and use these values to realize the change in user behavior. We also try to understand if a highly social user by quitting his contract also influences his close friends to terminate their agreement at the fitness club.

5.1 Exploratory data analysis

In this experiment, we tried to get an insight into the data. We saw the data at various angles to identify underlying patterns and tried to understand the reason behind it.

5.1.1 Hourly visiting frequency

We grouped both the group exercise and the gym data based on the hour of the day and plotted the overall graph just to see how the usage varies over different hours of the day. Looking at the results [Fig 5.1], we were convinced that users prefer to go to the gym in the evening, usually around 18:00. Also, we could see the second-best preferred time to work out is around 10:00.

We later granulated the data by splitting it into two groups based on the exercise type, namely gym exercise, and group exercise. To our surprise, there was a significant difference in the visiting patterns, as shown in [Fig 5.2a] and [Fig 5.2b]. There is an upward trend in the visiting frequency for gym exercise, but the group exercise is mostly done from 16:00 to 19:00.

We also checked the visiting frequency based on age groups. From this analysis, we can see that young users prefer to work out in the evening, whereas senior users prefer the morning hours to work out. From figure 5.3, we can see that as the age factor increases, the preferred workout time shifts to morning.

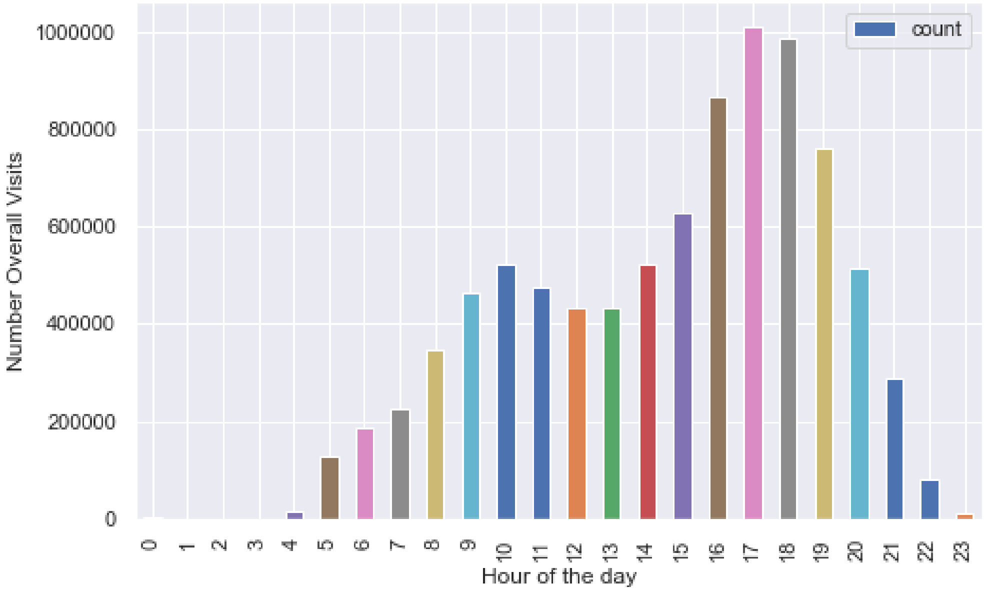
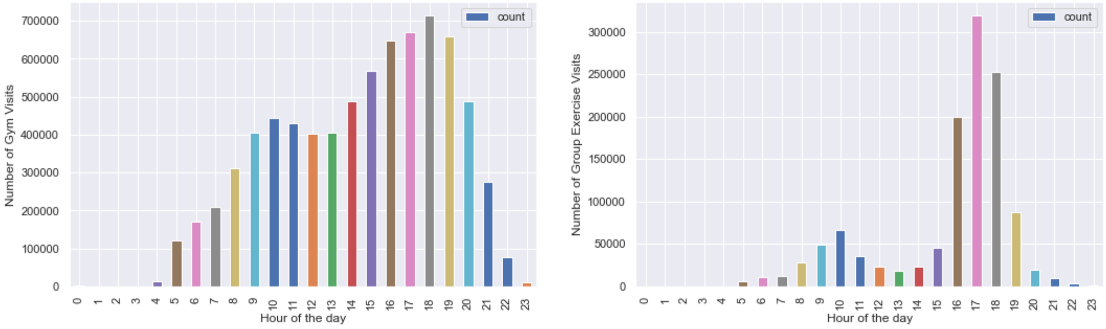


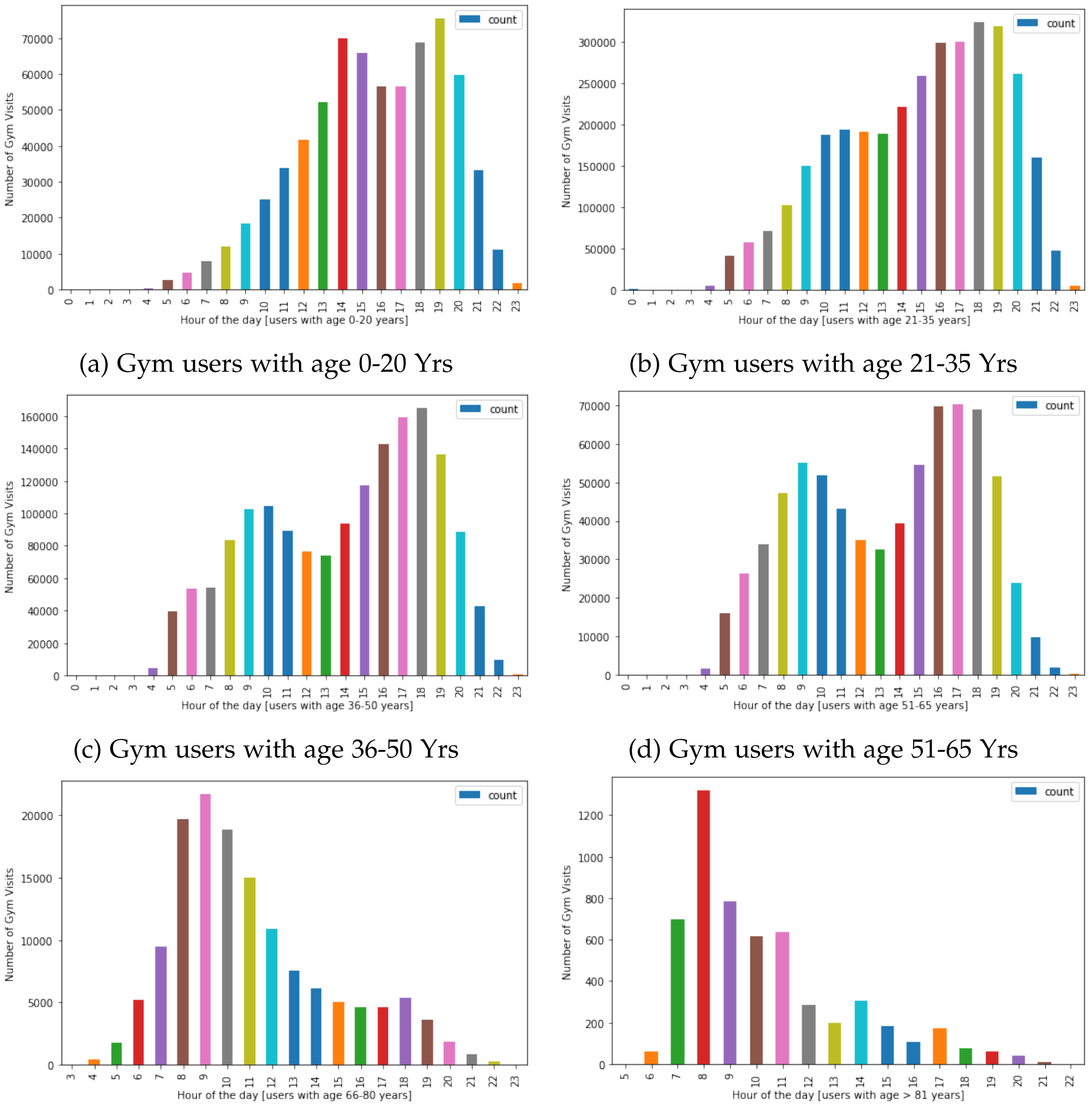
Figure 5.1: Overall hourly visiting frequency of users



(a) Gym users

(b) Group Exercise users

Figure 5.2: Hourly visiting frequency of gym and group Exercise users



(a) Gym users with age 0-20 Yrs

(b) Gym users with age 21-35 Yrs

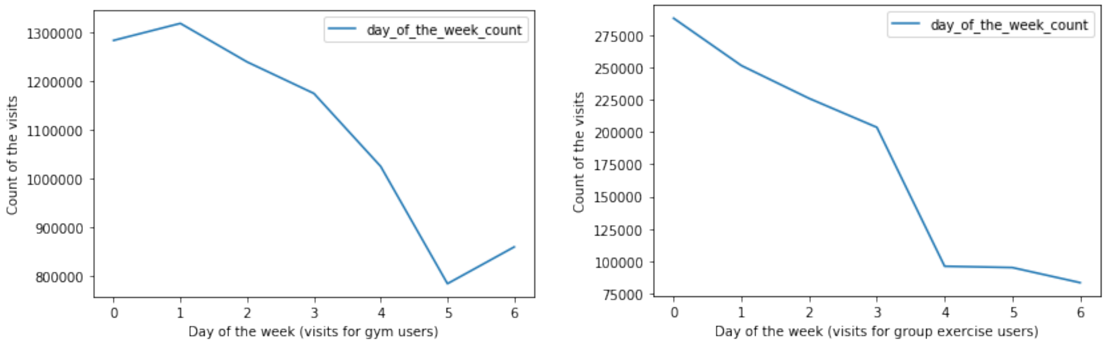
(c) Gym users with age 36-50 Yrs

(d) Gym users with age 51-65 Yrs

(e) Gym users with age 66-80 Yrs

(f) Gym users with age >80 Yrs

Figure 5.3: Hourly visiting frequency of gym users with different age group



(a) Gym users

(b) Group Exercise users

Figure 5.4: Visiting frequency for the day of the week - gym and group Exercise users

5.1.2 Weekly visiting frequency

After having analyzed the hourly frequency, we wanted to see how the data varies on a weekly basis and get to know if people like to work out on weekdays or weekends or both. So, we first calculated the weekly visiting frequency for gym and group exercise separately. The results [Fig. 5.4a] and [Fig. 5.4b] show that people prefer to workout on a weekday rather than the weekend. Here in the graph, on the X-axis, days of the week 0-6 stand for the days of the week Monday-Sunday.

5.1.3 Monthly visiting frequency

The weekly visiting frequency kind of gave us an idea that users tend to workout on a weekday. We wanted to see if the visiting frequency remains the same every month, or if it varies based on seasons. We can see from [Fig. 5.5] that there is genuinely a dip in the curve from May to August, which is the peak summer. Looking at the results and after consulting the domain experts, we can conclude that the users would indulge more in outdoor activities like hiking, cycling and might go on vacation, which leads to lesser attendance at the fitness club. Also, because of the holiday season, we can see low attendance in December.

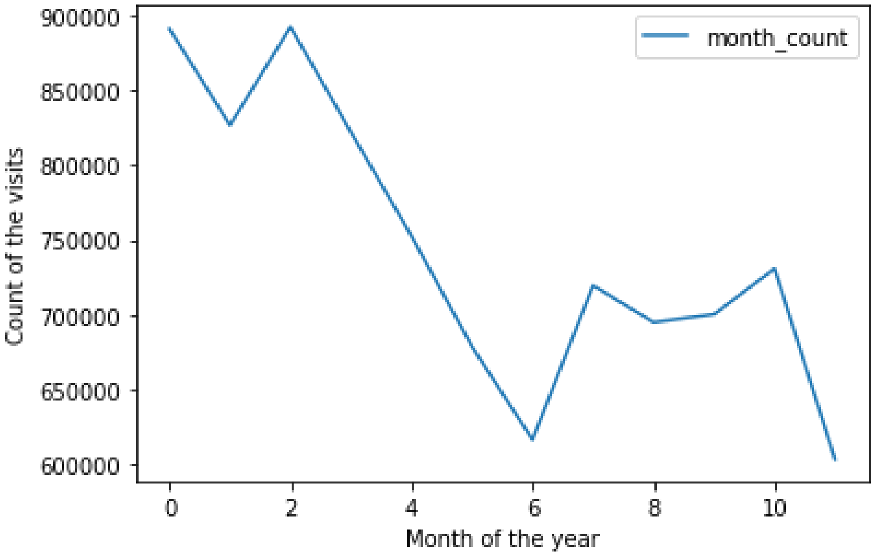


Figure 5.5: Monthly Visiting frequency

5.1.4 Contract duration

The contract duration of a user is the time duration during which the contract of a user is active. We calculate the contract duration of a user by subtracting the start date from the date of contract termination.

We try to visualize the variation of contract duration based on gender and age. Looking at the results [Fig. 5.6], we can say gender doesn't matter as both male and female users have almost the same mean and standard deviation.

We also plot box-plots for different age groups. We split the users into six age groups 1 - 6 based on their age. The age group 1 represents all the users between age 0-20 years, age group 2 holds all the users from age 21-35 years, and so on till age group 6, which represents users of age greater than 80 years. Looking at the results, we can infer that the contract duration of mid-aged users is high compared to the younger and senior users.

5.1.5 Location wise user distribution

There were 141918 users, and the login data of all these users would be a huge chunk of data. So we had to sample the data to reduce the workload on the system that

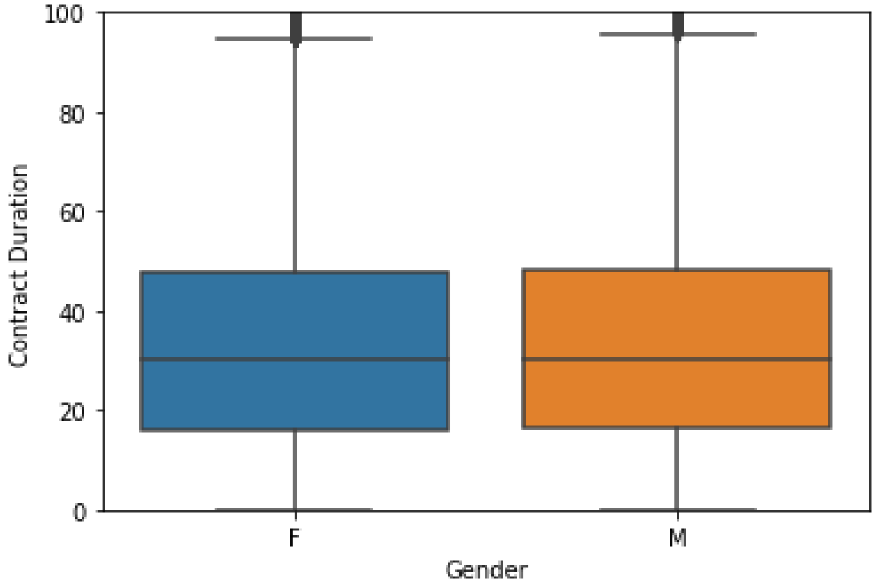


Figure 5.6: Contract duration box-plot based on gender

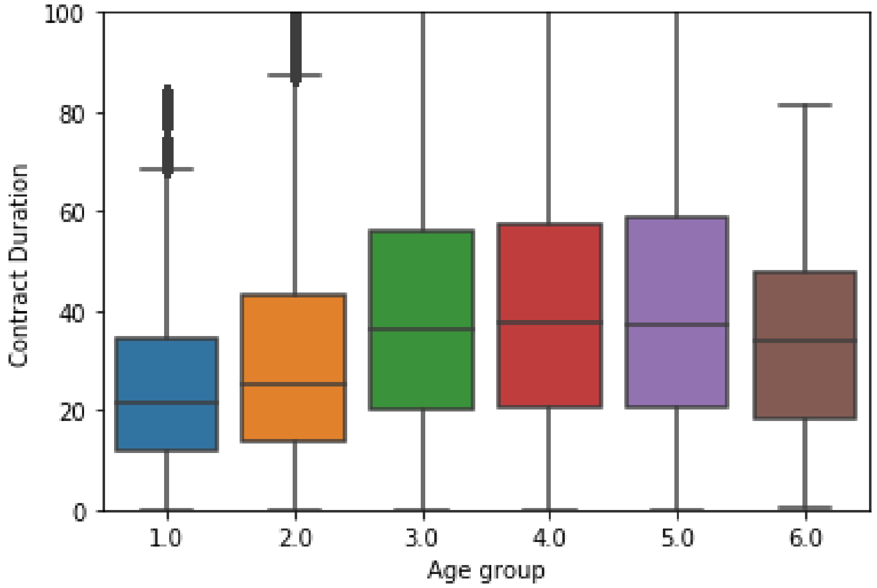


Figure 5.7: Contract duration box-plot based on age

calculates the social network. We did a location-based distribution to see how users of different age groups are distributed across different locations and chose a couple of locations that had almost equal counts of users across different age groups.

We wanted to see how the data is distributed based on their age group across different locations. So we plotted graphs for every location with user count on Y-axis and their age on X-axis. Each location showed a variation in the distribution of users. Some locations were dominated by youngsters [Fig. 5.6]. Some locations had a majority of mid-aged users [Fig. 5.6].

But we wanted to select two locations that had good user count and are of diverse age groups. We can see in the [Fig. 5.10] that there are users from all the age groups and count of users are almost the same. We use this location to carry out social network analysis.

We also plotted a box-plot to see the variation in contract duration. [Fig. 5.11] shows the mean and standard deviation of contract duration across different locations. Many factors would contribute to this variation in the contract duration. It could be the year in which the gym started, the locality of the gym (i.e., if it's a residential location and far away from the city center, the crowd would be less), the retention rate of users and many more.

5.2 Social Network Analysis

We performed social network analysis over the data and calculated various social network features from the built network. Once we have the computed social network features for all the users of the fitness club, we start by plotting these values against each other and see how these values correlate with each other.

5.2.1 Betweenness Centrality

The betweenness centrality value of a user defines his role as a bond creator between two other users at the fitness club. Higher its value better is his role in introducing two strangers at the club.

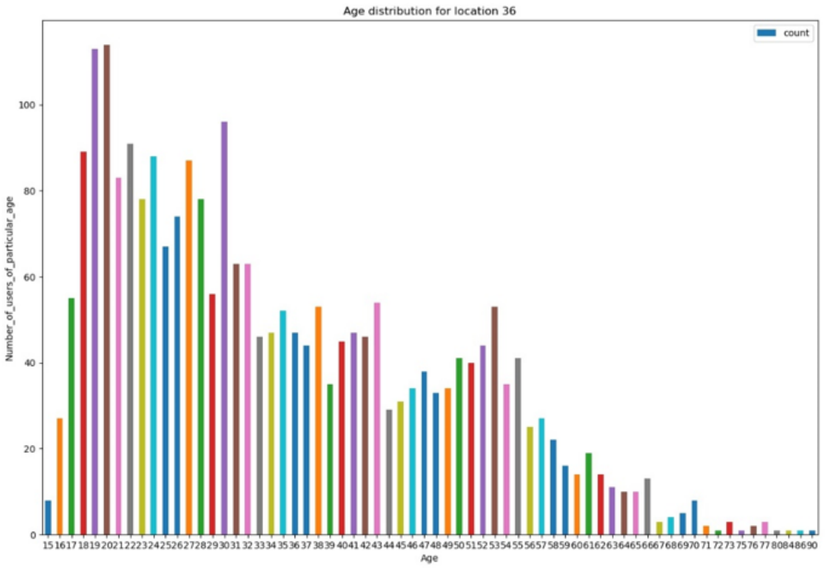


Figure 5.8: Location with majority of young users

Contract duration to betweenness centrality

We first see how the overall betweenness centrality of a user contributes to his contract duration in the [Fig. 5.12]. Every dot in the graph represents a user and where she/he stands in terms of his betweenness centrality and contract duration value.

We then plot the distribution separately for gym and group exercise to see if there is any significance in the distributions obtained. We can see that there is not much variation in the betweenness centrality value in the case of group exercise users [Fig. 5.14b]. Whereas gym users have quite some variation in the value, and it is increasing with the increase in contract duration [Fig. 5.14a]. This variation gives us an indication that more the role of a user in building bonds at the gym better is his contract duration.

We also wanted to analyze if other features like gender and age of the user influence the betweenness centrality value. So, we plot the graph based on gender, as shown in [Fig. 5.17]. We can see that male and female users almost have the same distribution of betweenness centrality value. From these results, we can conclude that gender doesn't play a significant role in building social networks at the fitness club.

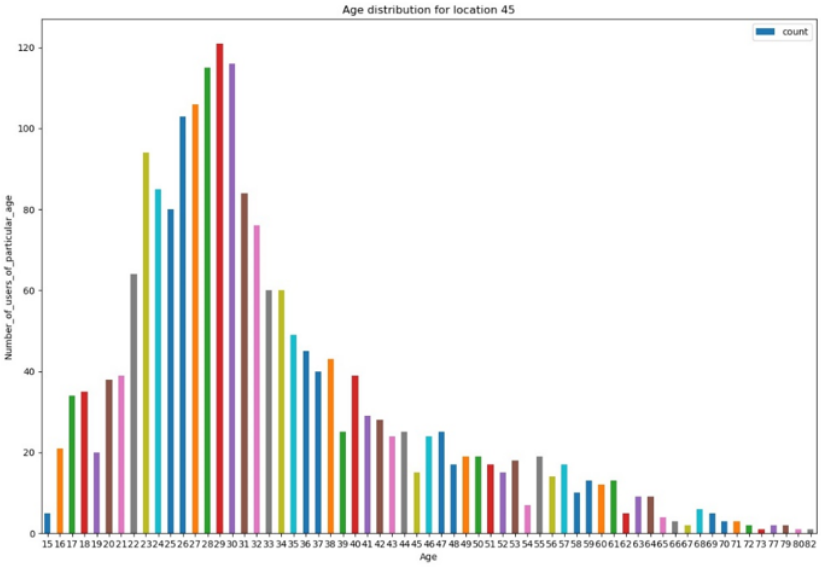


Figure 5.9: Location with majority of mid-aged users

When we do an age-based analysis, we can see that users with the age group from 20 to 40 years are active at forming social bonds [Fig. 5.18]. We can see that users of other age groups are not so keen on getting people close to each other. As the user's age increase, his betweenness centrality goes down, showing a lack of her/his interest in social involvement at the gym [Fig. 5.19].

Visiting frequency to betweenness centrality

The next user behavior that we wanted to monitor and see its variation based on social activeness is the users visiting frequency. It defines how often users workout.[Fig. 5.20]

Like earlier steps, we try to see the significance of betweenness centrality to weekly visiting frequency for gym and group exercise users. And looking at the results, we can conclude that group exercise users don't involve much in social bonding like the gym users.

We then move forward with the analysis and try to see the significance of the age group in this relation. Similar to contract duration, correlation is high for users with age

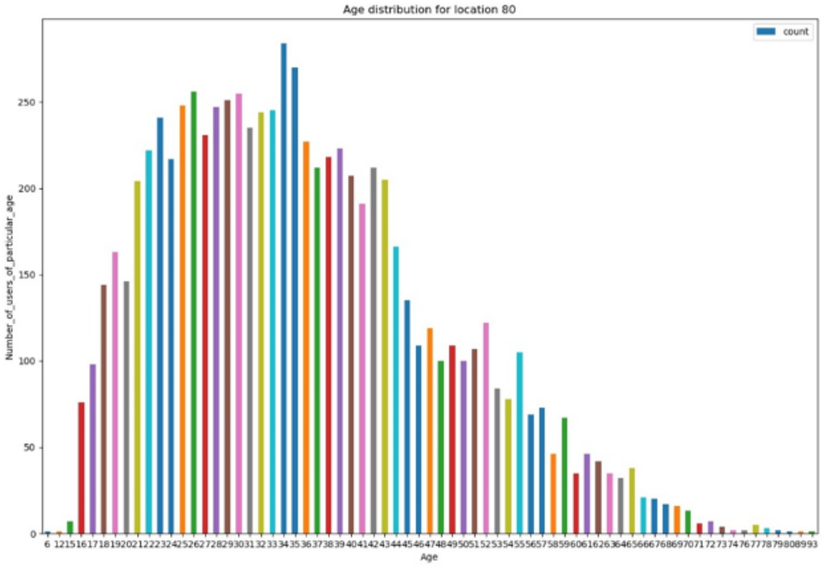


Figure 5.10: Location with uniformly distributed users of different age group

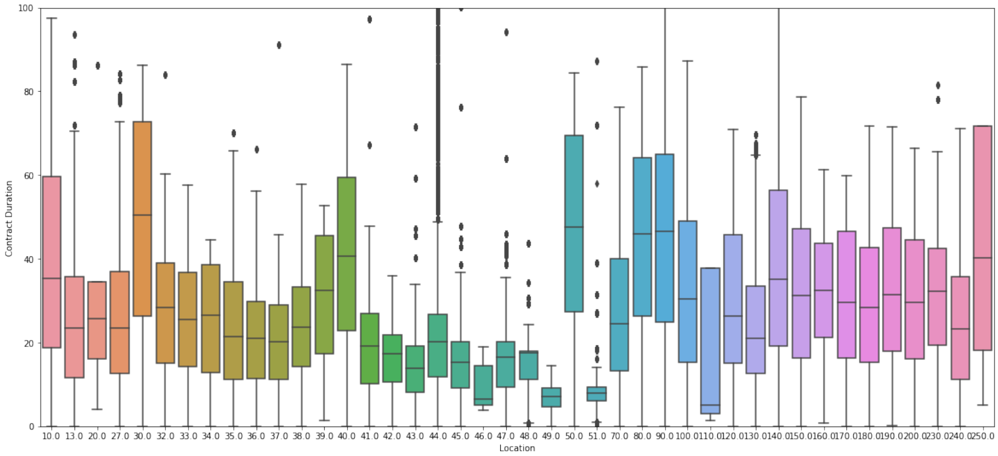


Figure 5.11: Location wise box-plot for contract duration

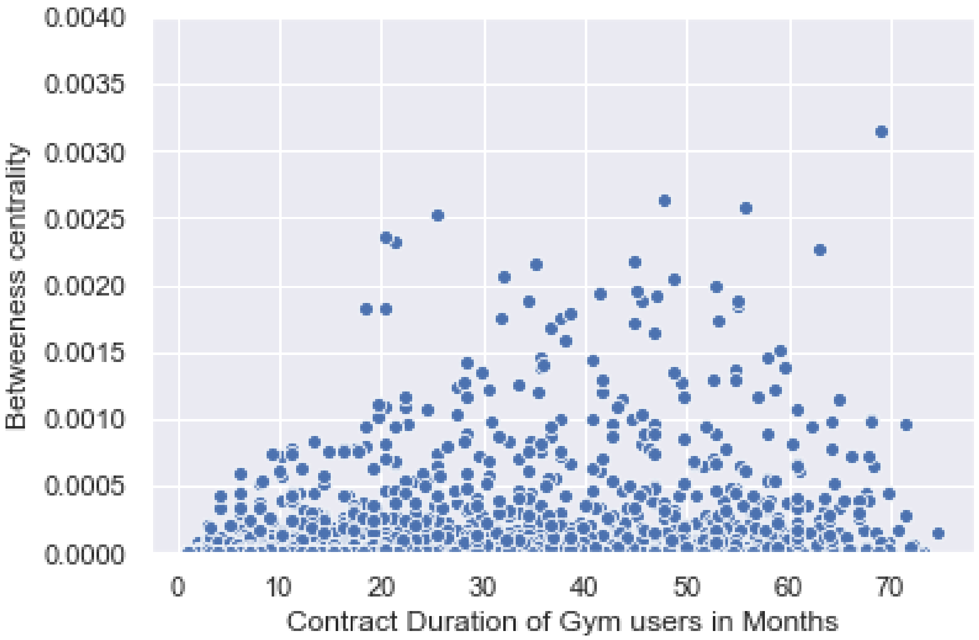
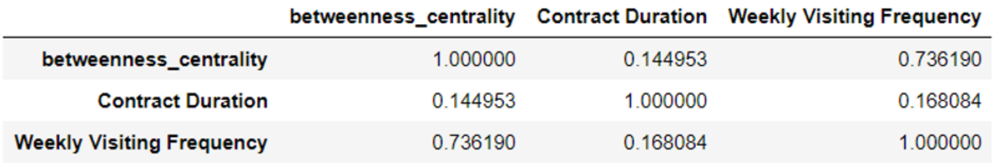
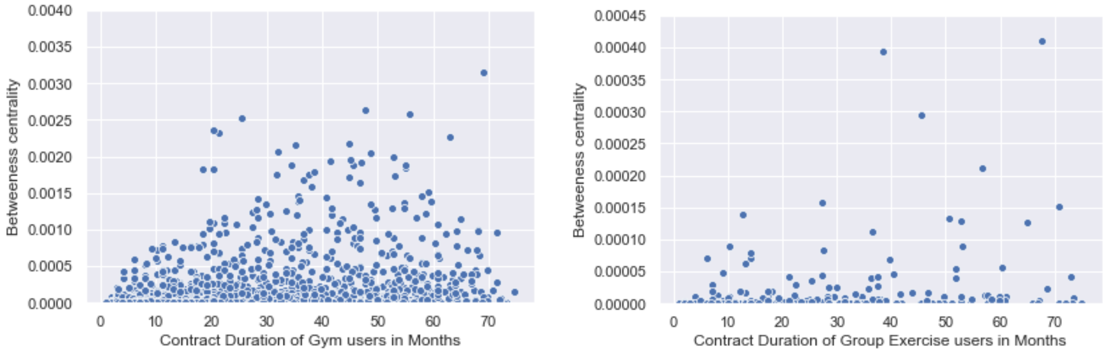


Figure 5.12: Betweenness centrality to contract duration for overall users



	betweenness_centrality	Contract Duration	Weekly Visiting Frequency
betweenness_centrality	1.000000	0.144953	0.736190
Contract Duration	0.144953	1.000000	0.168084
Weekly Visiting Frequency	0.736190	0.168084	1.000000

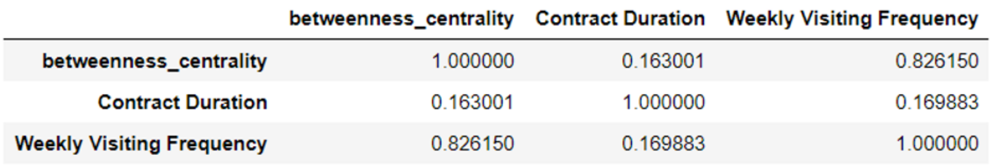
Figure 5.13: Correlation matrix of betweenness centrality for overall users



(a) Gym users

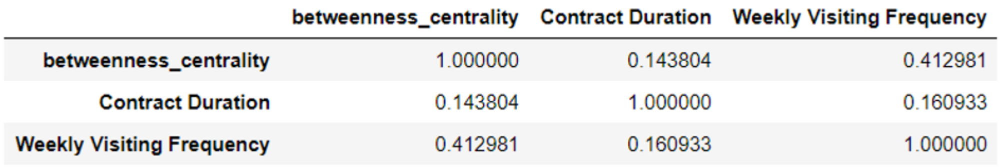
(b) Group exercise users

Figure 5.14: Betweenness centrality to contract duration for gym and group exercise users



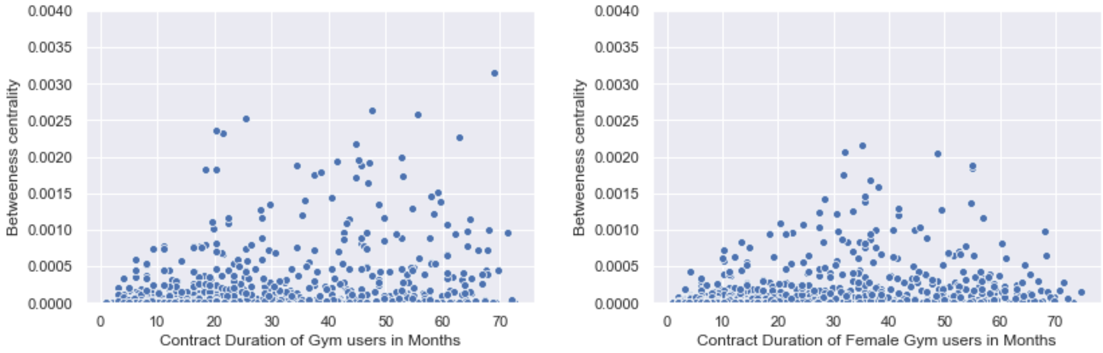
	betweenness_centrality	Contract Duration	Weekly Visiting Frequency
betweenness_centrality	1.000000	0.163001	0.826150
Contract Duration	0.163001	1.000000	0.169883
Weekly Visiting Frequency	0.826150	0.169883	1.000000

Figure 5.15: Correlation matrix of betweenness centrality for gym users



	betweenness_centrality	Contract Duration	Weekly Visiting Frequency
betweenness_centrality	1.000000	0.143804	0.412981
Contract Duration	0.143804	1.000000	0.160933
Weekly Visiting Frequency	0.412981	0.160933	1.000000

Figure 5.16: Correlation matrix of betweenness centrality for group exercise users



(a) Male users

(b) Female users

Figure 5.17: Betweenness centrality to contract duration for male and Female users

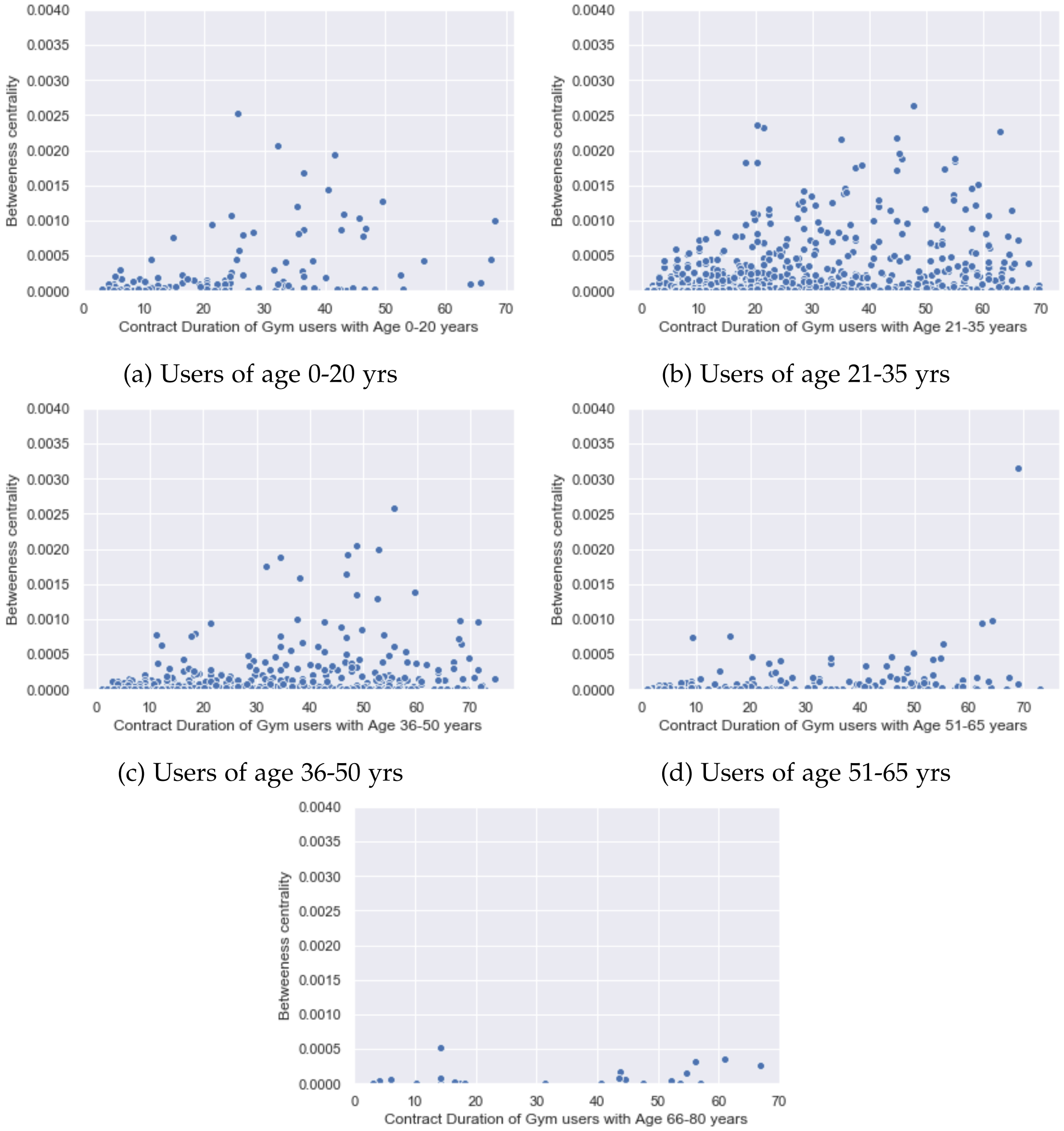
group between 20 to 40 years. We can see that as the betweenness centrality increases, the weekly visiting frequency also boosts up. But as the user's age increases, his betweenness centrality slides down. Also, we can see that aged users don't use the fitness club frequently, leading to a lesser average visiting frequency.[Fig. 5.22]

5.2.2 Closeness Centrality

The closeness centrality value of a user justifies how close a person is to other users in the network. A higher value means he has more close friends at the club and hence makes him a relevant person at the gym. We calculate the closeness centrality for every user at the fitness club and see its correlation [Fig. 5.23] with contract duration and weekly visiting frequency[Fig. 5.24].

Contract duration to Closeness centrality

Later we plot the distribution separately for gym and group exercise to see if we can find any significance in the distribution. We can see that there is not much variation in the closeness centrality value in the case of group exercise users. Whereas for gym users, we can see increasing closeness value with the increase in contract duration, implying that more the users get to know each other at the fitness club, better is the chance of having longer contract duration.[Fig. 5.25]



(a) Users of age 0-20 yrs

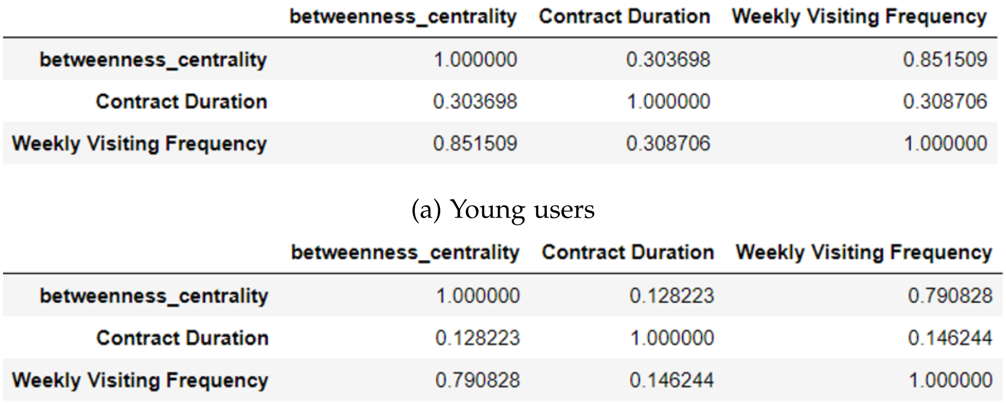
(b) Users of age 21-35 yrs

(c) Users of age 36-50 yrs

(d) Users of age 51-65 yrs

(e) Users of age 66-80 yrs

Figure 5.18: Betweenness centrality to contract duration for users of different age group



	betweenness_centrality	Contract Duration	Weekly Visiting Frequency
betweenness_centrality	1.000000	0.303698	0.851509
Contract Duration	0.303698	1.000000	0.308706
Weekly Visiting Frequency	0.851509	0.308706	1.000000

(a) Young users

	betweenness_centrality	Contract Duration	Weekly Visiting Frequency
betweenness_centrality	1.000000	0.128223	0.790828
Contract Duration	0.128223	1.000000	0.146244
Weekly Visiting Frequency	0.790828	0.146244	1.000000

(b) Elderly users

Figure 5.19: Betweenness centrality correlation for young and elderly users

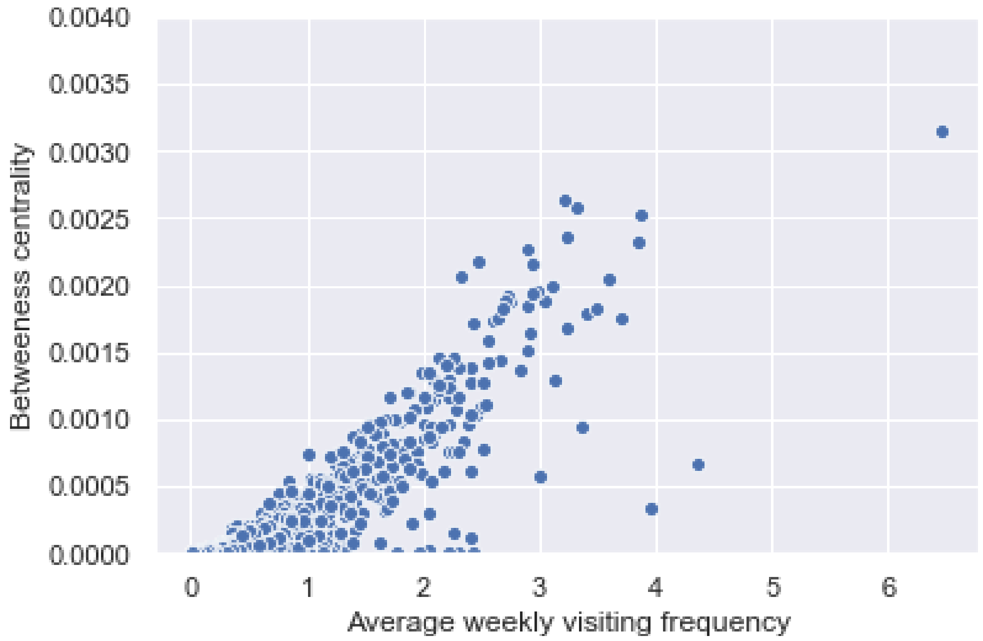
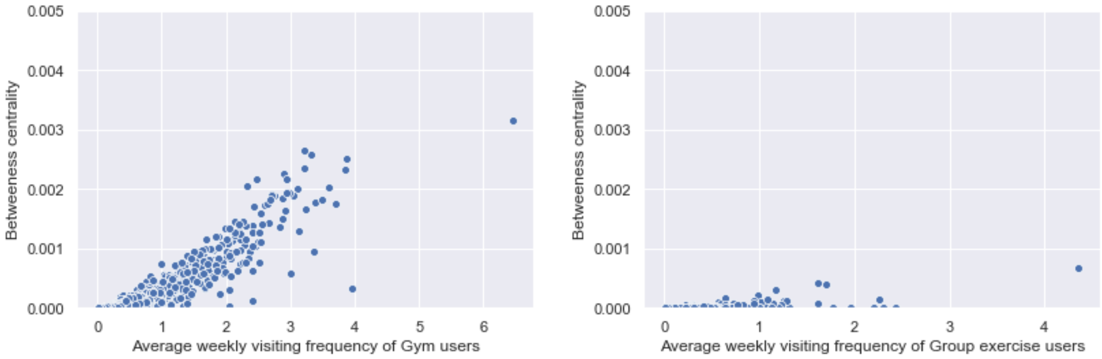


Figure 5.20: Betweenness centrality to weekly visits of overall users



(a) Gym users

(b) Group exercise users

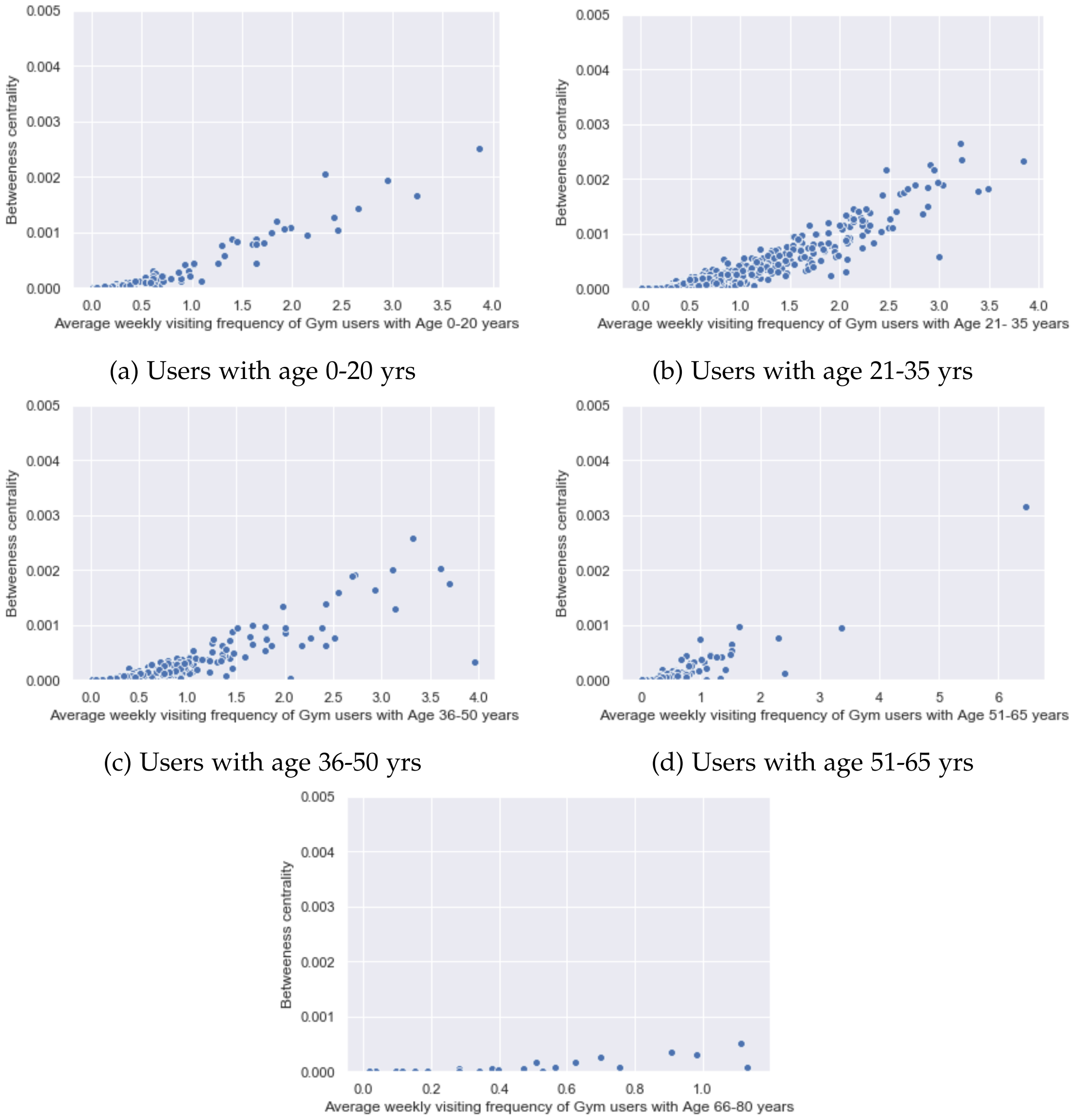
Figure 5.21: Betweenness centrality to weekly visits for gym and group exercise users

We also analyze the data with the gender[Fig. 5.27] [Fig. 5.28] and age [Fig. 5.30] [Fig. 5.29] perspective. Looking at the distribution of correlation value for males and females, we can say there is no significant difference; hence we can neglect it. But the age factor makes a difference again. We can see from the results that aged users don't have much correlation to contract duration and weekly visiting frequency. Also, as the age of the user increases, we can see her/his closeness centrality decreases, implying that younger users make more close friends than elderly users at the gym.

Visiting frequency to Closeness centrality

Once we achieve comparing contract duration and closeness centrality, we proceed to look at the correlation between closeness centrality and the visiting frequency of users. Looking at the plot[Fig. 5.31], we can say there is a high positive correlation between the two variables.

We try to see the significance of gender and age by plotting the same graph with a different subset of users. The gender factor doesn't make much of a difference, and it is evident in the graph [Fig. 5.32], but the age factor has a more significant influence, and it is visible in the graphs. We can see that aged users tend to slow down their pace in terms of visiting frequency, and also the correlation with the visiting frequency and closeness centrality value is pretty low.



(a) Users with age 0-20 yrs

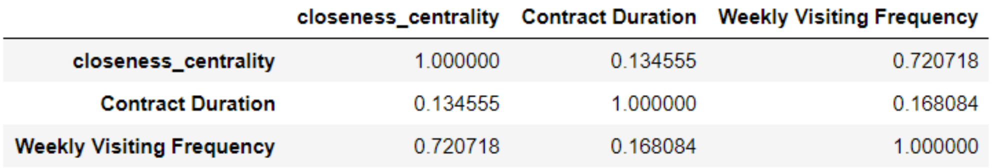
(b) Users with age 21-35 yrs

(c) Users with age 36-50 yrs

(d) Users with age 51-65 yrs

(e) Users with age 66-80 yrs

Figure 5.22: Betweenness centrality to weekly visiting frequency for users of different age group



	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.134555	0.720718
Contract Duration	0.134555	1.000000	0.168084
Weekly Visiting Frequency	0.720718	0.168084	1.000000

Figure 5.23: Correlation matrix of closeness centrality overall

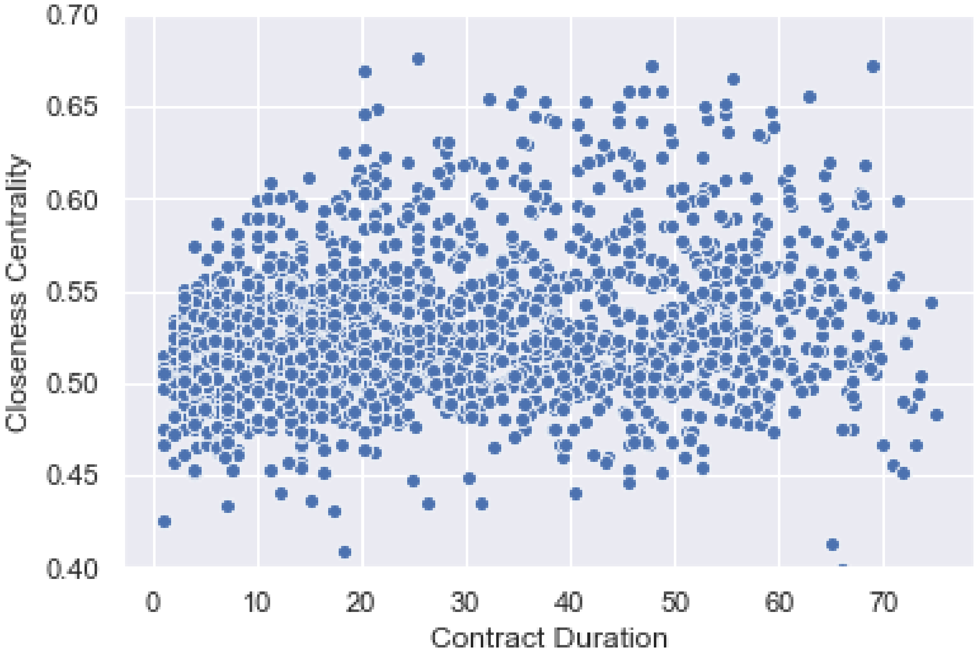
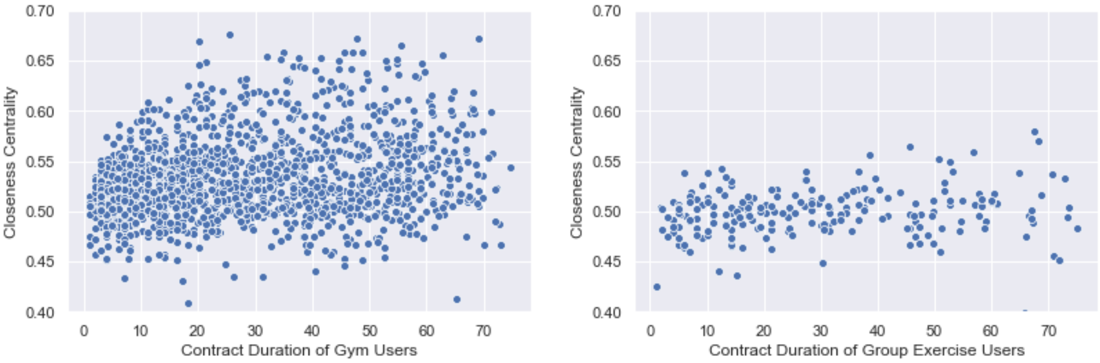


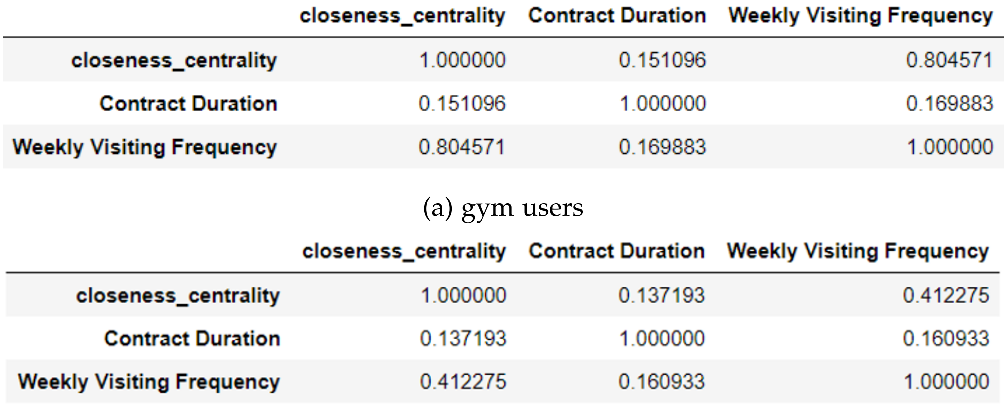
Figure 5.24: Closeness centrality to contract duration for overall users



(a) Gym users

(b) Group exercise users

Figure 5.25: Closeness centrality to contract duration for gym and group exercise users



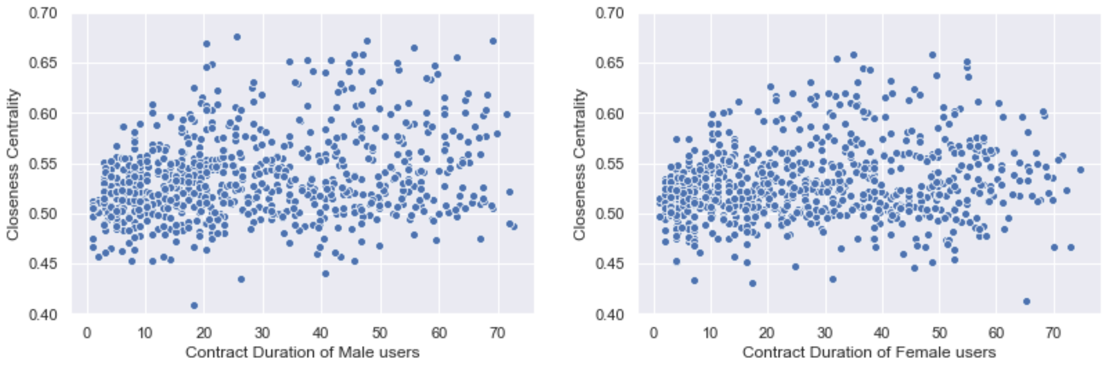
	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.151096	0.804571
Contract Duration	0.151096	1.000000	0.169883
Weekly Visiting Frequency	0.804571	0.169883	1.000000

(a) gym users

	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.137193	0.412275
Contract Duration	0.137193	1.000000	0.160933
Weekly Visiting Frequency	0.412275	0.160933	1.000000

(b) group exercise users

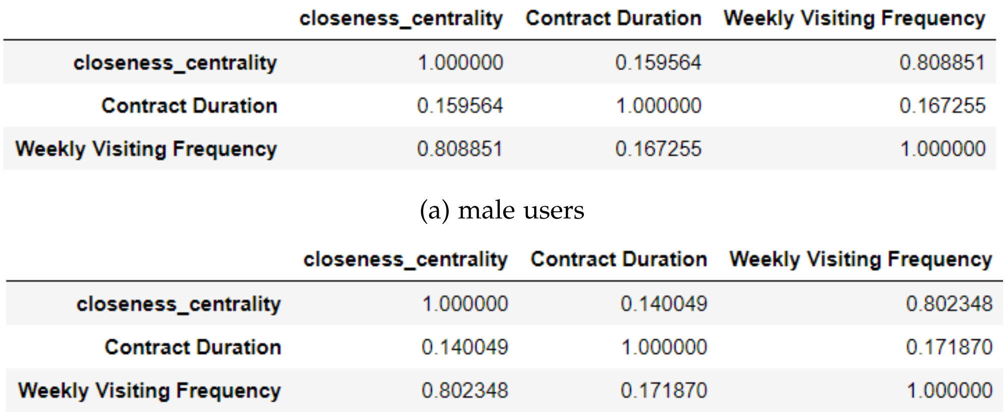
Figure 5.26: Closeness centrality correlation for gym and group exercise users



(a) Male users

(b) Female users

Figure 5.27: Closeness centrality to contract duration for male and female users



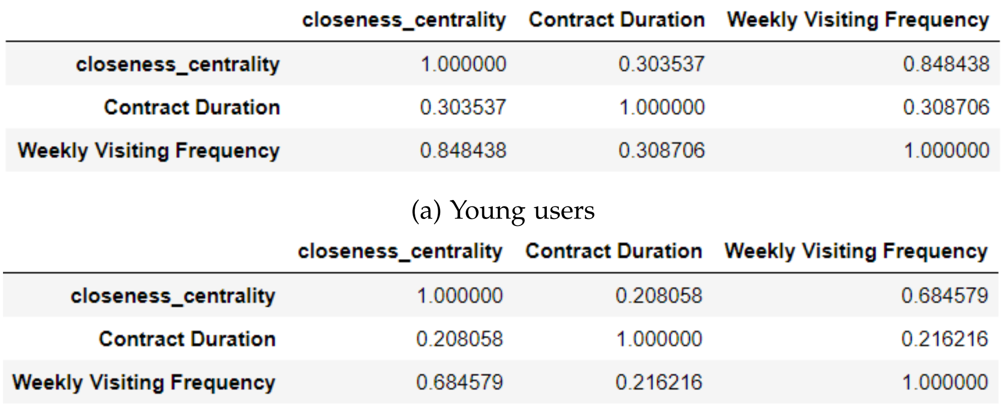
	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.159564	0.808851
Contract Duration	0.159564	1.000000	0.167255
Weekly Visiting Frequency	0.808851	0.167255	1.000000

(a) male users

	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.140049	0.802348
Contract Duration	0.140049	1.000000	0.171870
Weekly Visiting Frequency	0.802348	0.171870	1.000000

(b) female users

Figure 5.28: Closeness centrality correlation for male and female users



	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.303537	0.848438
Contract Duration	0.303537	1.000000	0.308706
Weekly Visiting Frequency	0.848438	0.308706	1.000000

(a) Young users

	closeness_centrality	Contract Duration	Weekly Visiting Frequency
closeness_centrality	1.000000	0.208058	0.684579
Contract Duration	0.208058	1.000000	0.216216
Weekly Visiting Frequency	0.684579	0.216216	1.000000

(b) Elderly users

Figure 5.29: Closeness centrality correlation for young and elderly users

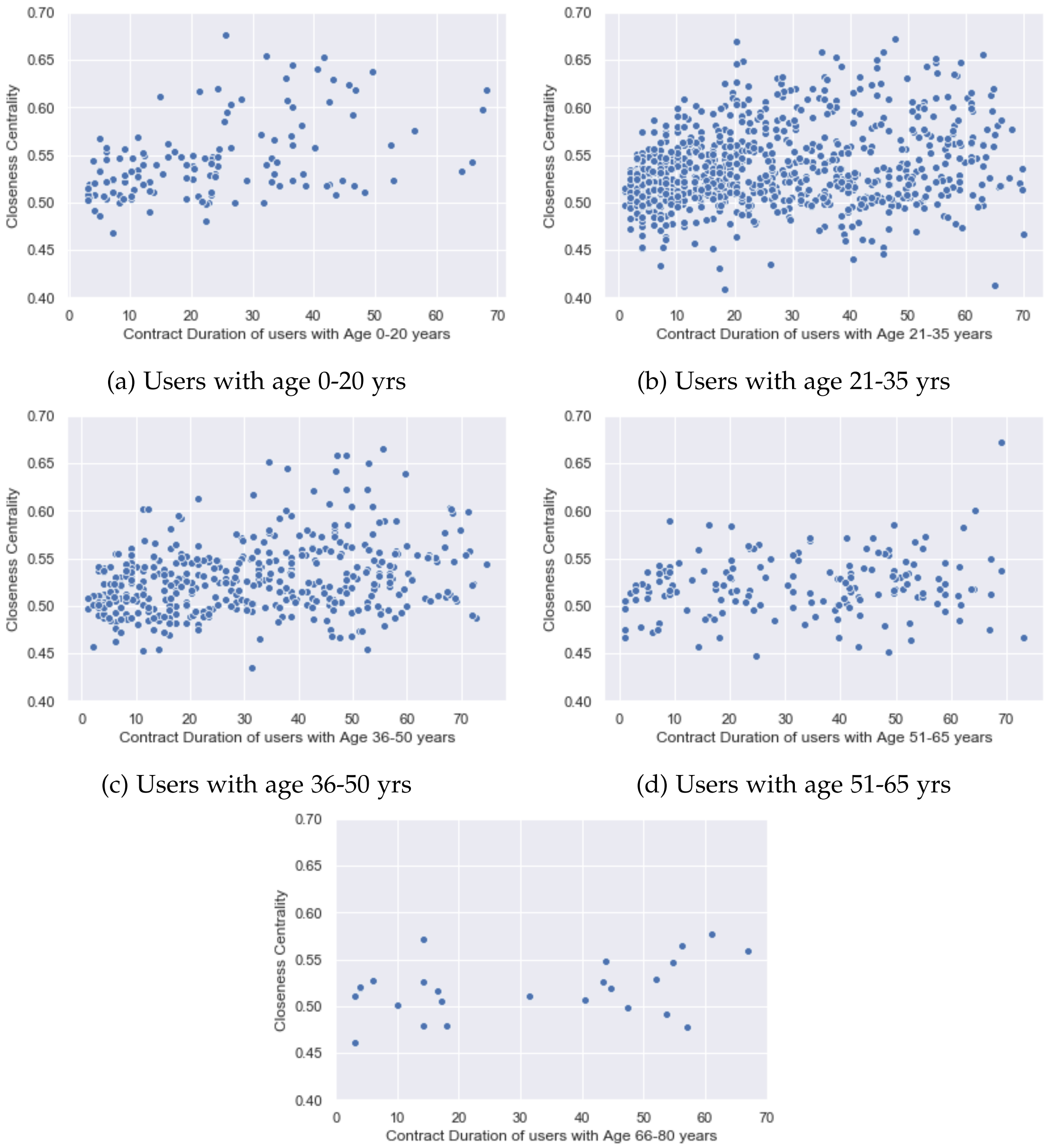
5.2.3 Degree Centrality

The degree centrality of a node in a network justifies the number of connections it has with other nodes in the network. In the context of fitness club data, we can say degree centrality depicts the number of other users in the network the current user knows. We perform all the observations for degree centrality and see that all the inferences still hold good.

Contract duration to Degree centrality

Like earlier analysis, the correlation between degree centrality and the contract duration is high for gym users than group exercise users [Fig. 5.34] [Fig. 5.36]. From this, we could infer that gym users try to mingle more with other users and build a friendship with other gym users, whereas, with group exercise users, it's not the case.

We then also analyze to see if the other factors of users like the gender and age control their degree centrality value. As observed in the other centrality values, we can see that gender is not an impediment in building social connections, but age does play a role [Fig. 5.37]. As users age, they get more reserved in their life and only do the workout without many social interactions compared to the younger users. Also, the visiting frequency slides down as the user ages.



(a) Users with age 0-20 yrs

(b) Users with age 21-35 yrs

(c) Users with age 36-50 yrs

(d) Users with age 51-65 yrs

(e) Users with age 66-80 yrs

Figure 5.30: Closeness centrality to contract duration for users of different age group

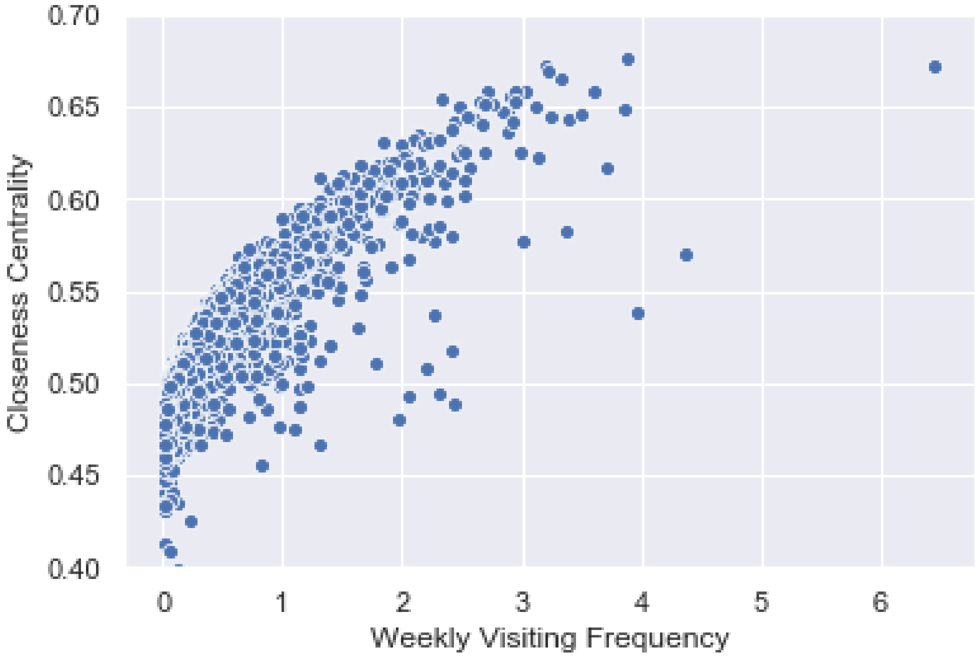
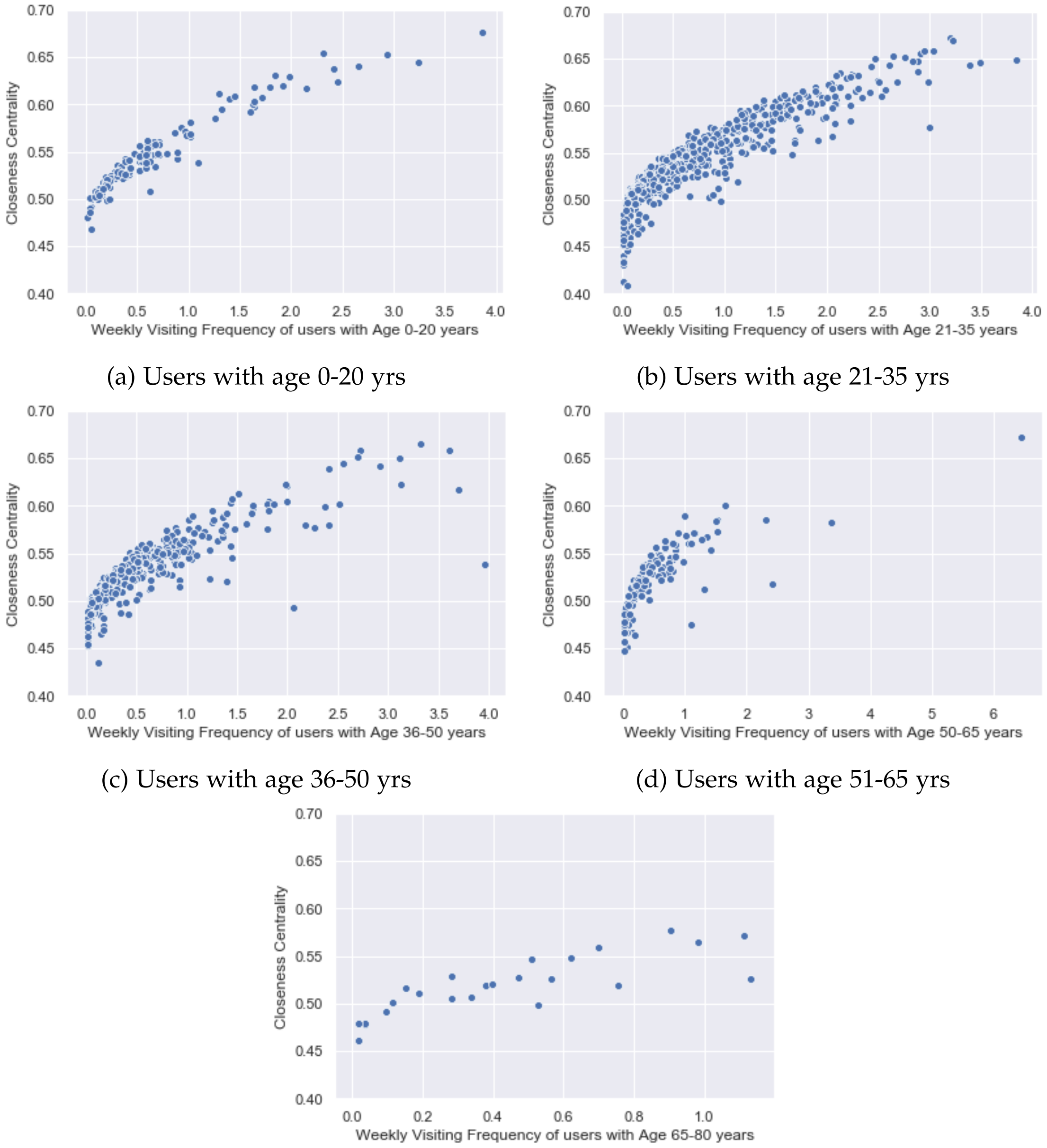


Figure 5.31: Closeness centrality to weekly visiting frequency for overall users



(a) Users with age 0-20 yrs

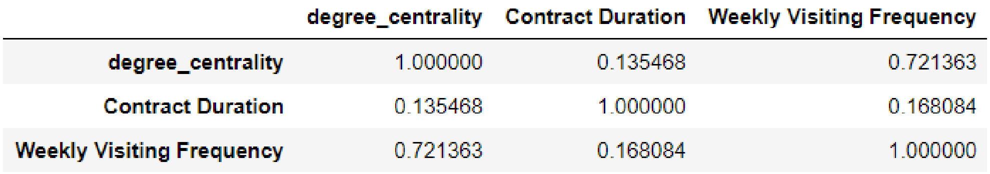
(b) Users with age 21-35 yrs

(c) Users with age 36-50 yrs

(d) Users with age 51-65 yrs

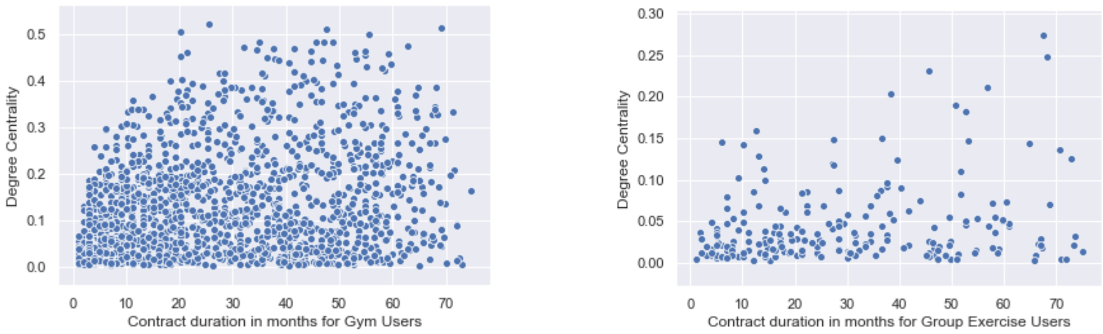
(e) Users with age 66-80 yrs

Figure 5.32: Closeness centrality to Weekly visiting frequency for users of different age group



	degree_centrality	Contract Duration	Weekly Visiting Frequency
degree_centrality	1.000000	0.135468	0.721363
Contract Duration	0.135468	1.000000	0.168084
Weekly Visiting Frequency	0.721363	0.168084	1.000000

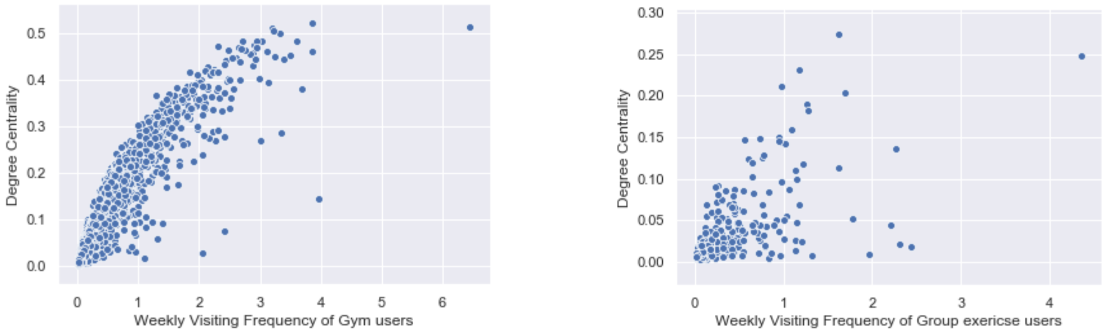
Figure 5.33: Degree centrality correlation matrix overall



(a) gym users

(b) group exercise users

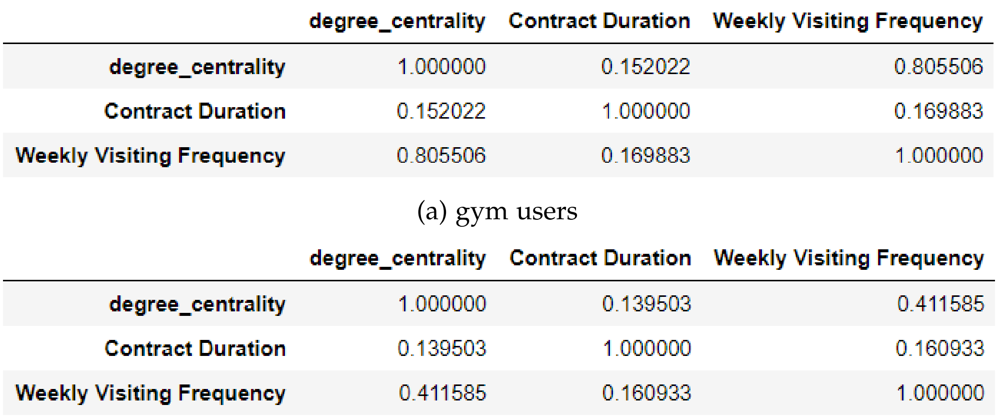
Figure 5.34: Degree centrality to contract duration for gym and group exercise users



(a) gym users

(b) group exercise users

Figure 5.35: Degree centrality to weekly visiting frequency for gym and group exercise users



	degree_centrality	Contract Duration	Weekly Visiting Frequency
degree_centrality	1.000000	0.152022	0.805506
Contract Duration	0.152022	1.000000	0.169883
Weekly Visiting Frequency	0.805506	0.169883	1.000000

(a) gym users

	degree_centrality	Contract Duration	Weekly Visiting Frequency
degree_centrality	1.000000	0.139503	0.411585
Contract Duration	0.139503	1.000000	0.160933
Weekly Visiting Frequency	0.411585	0.160933	1.000000

(b) group exercise users

Figure 5.36: Degree centrality correlation for gym and group exercise users

Visiting frequency to Degree centrality

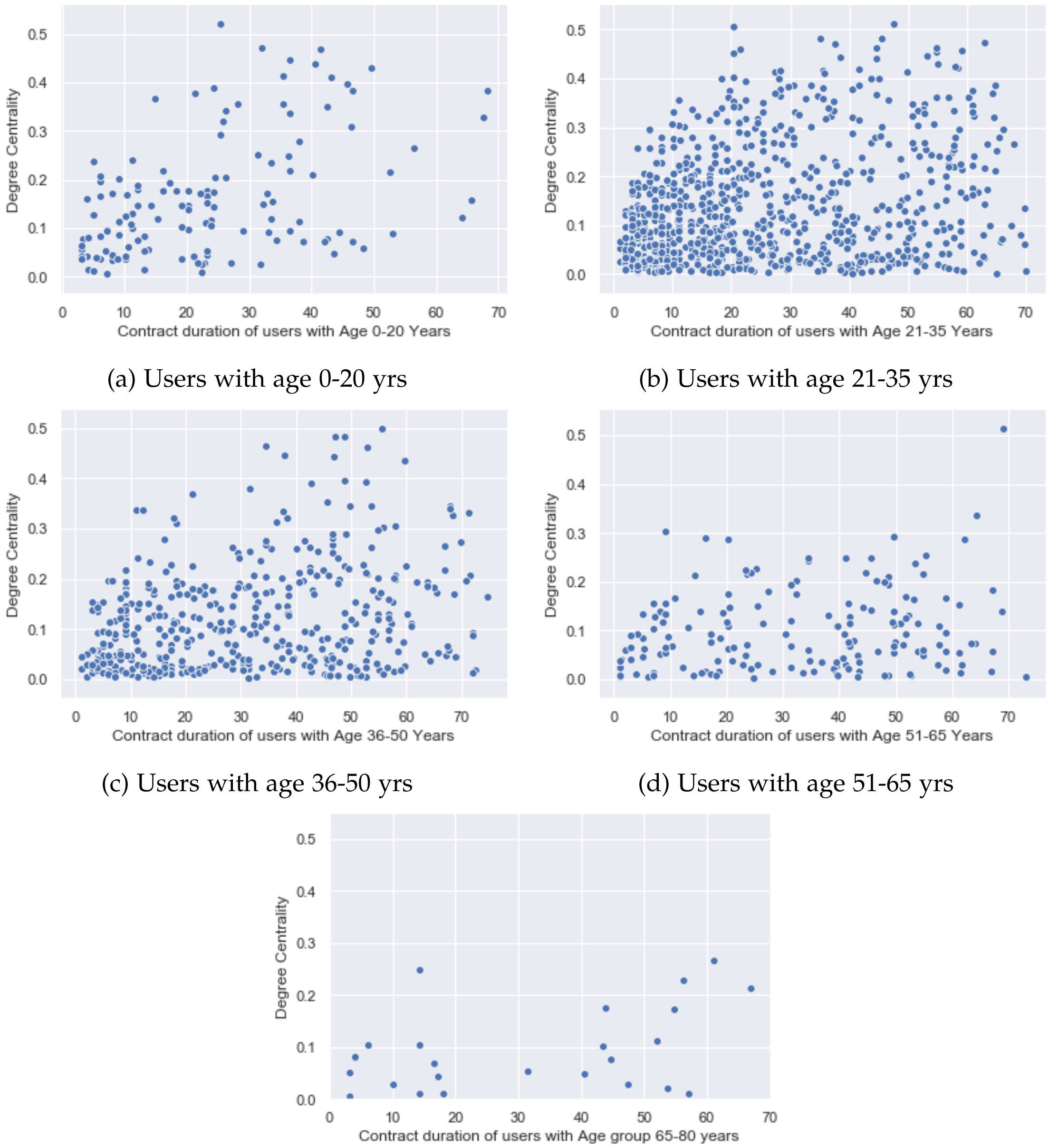
Looking at all these results[Fig. 5.38], we can conclude that gym is a good environment for social interactions, and age of the user plays a vital role in it.

5.3 Rolling window mean on centrality

Since the distribution is quite spread out, it is hard to analyze how the data varies. So, we applied a rolling window mean over the data to understand the dependencies between the variables better. We take into consideration a small window of the distribution and compute the X and Y-axis average of all the points within it. A window comprises the lower and upper limit value for the Y-axis, and we move the window from the lowest value to the highest value to cover the whole distribution. In every step average of X and Y-axis is calculated. After completion, a graph of these average values is plotted.

Centrality to contract duration

In [Fig. 5.49], each blue dot is a plot of betweenness centrality and contract duration for every user of a location. The two black lines constitute a window with a lower



(a) Users with age 0-20 yrs

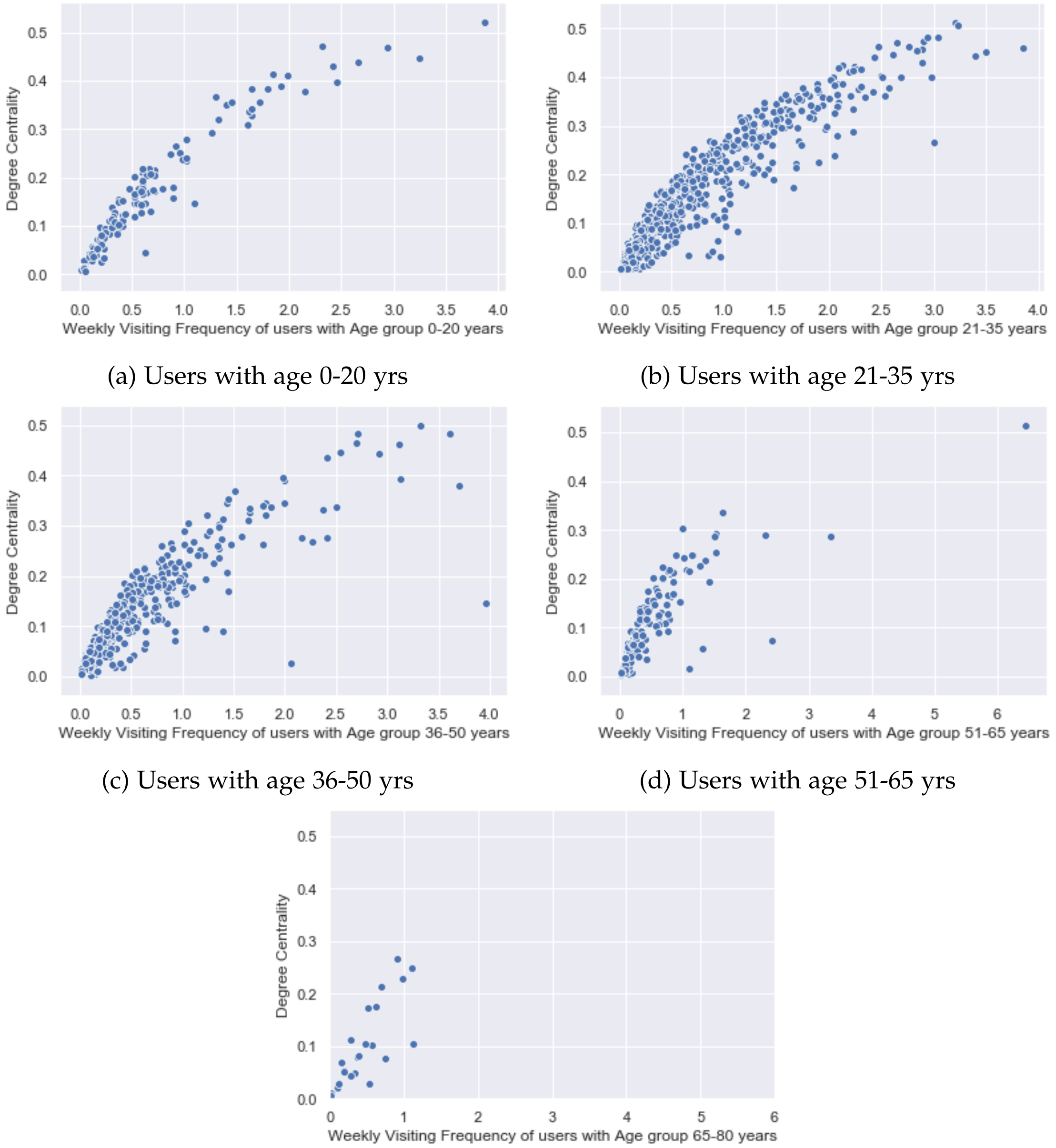
(b) Users with age 21-35 yrs

(c) Users with age 36-50 yrs

(d) Users with age 51-65 yrs

(e) Users with age 66-80 yrs

Figure 5.37: Degree centrality to contract duration for users of different age group



(a) Users with age 0-20 yrs

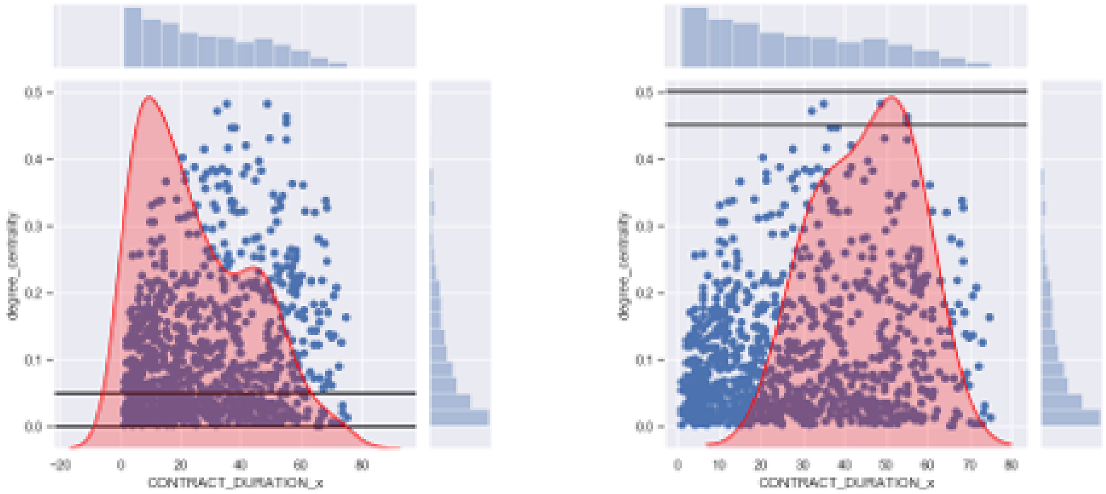
(b) Users with age 21-35 yrs

(c) Users with age 36-50 yrs

(d) Users with age 51-65 yrs

(e) Users with age 66-80 yrs

Figure 5.38: Degree centrality to weekly visiting frequency for users of different age group



(a) Starting point of window

(b) Ending point of window

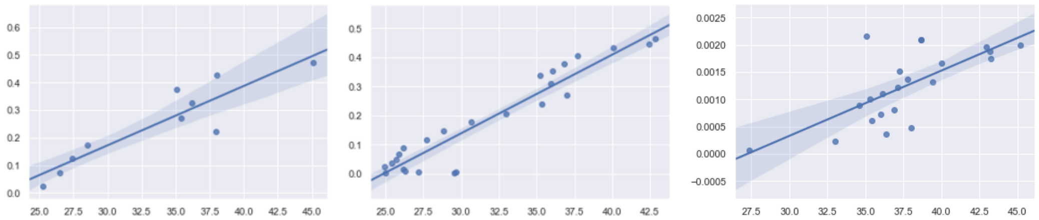
Figure 5.39: Rolling window from lower to higher centrality value

and upper limit, which keeps increasing from zero to max value. These two diagrams are first and the last plot of the series of plots with increasing window value. The points in these windows are taken into consideration, and the mean contract duration is calculated, the red peak depicts the mean for that interval. We can see that as we raise the window, the mean shifts to the right, i.e., contract duration increases, implying when a user's betweenness centrality value increases, his contract duration increases. The final plot of average betweenness centrality to contract duration [Fig. 5.40c] calculated at every step in the above experiment is plotted to see a positive slope, stating that a higher betweenness centrality leads to higher contract duration. A similar experiment is carried out for degree [Fig. 5.40a] and closeness centrality [Fig. 5.40b] to see if they respond the same as betweenness centrality value.

Centrality to Visiting frequency

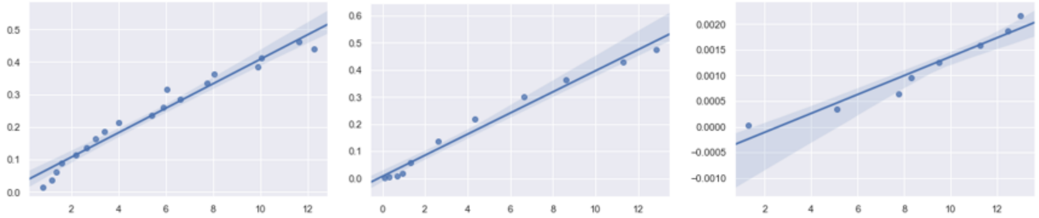
The rolling mean approach is applied to the relation between centrality values and visiting frequency, and the results are as shown in [Fig. 5.41].

Looking at all these results, we could see that the social network features do play a significant role in influencing user behavior like contract duration and visiting frequency.



(a) Contract duration Vs Degree centrality (b) Contract duration Vs Closeness centrality (c) Contract duration Vs Betweenness centrality

Figure 5.40: Regression plot of Contract duration on the x-axis and centrality on the y-axis



(a) Visiting frequency Vs Degree centrality (b) Visiting frequency Vs Closeness centrality (c) Visiting frequency Vs Betweenness centrality

Figure 5.41: Regression plot of Monthly Visiting frequency on the x-axis and centrality on the y-axis

5.4 Machine learning

5.4.1 Regression curve fitting

We wanted to fit the regression line and calculate the coefficient for the fit. We try to fit different centrality values and see which fit is best. Some plots fit best at a second degree, but some plots need a higher polynomial fit.

Closeness Centrality

For closeness centrality we can see that polynomial fit of degree 2 is a better fit than linear fit [Fig. 5.42]. The coefficients are provided below the plot.

Degree Centrality

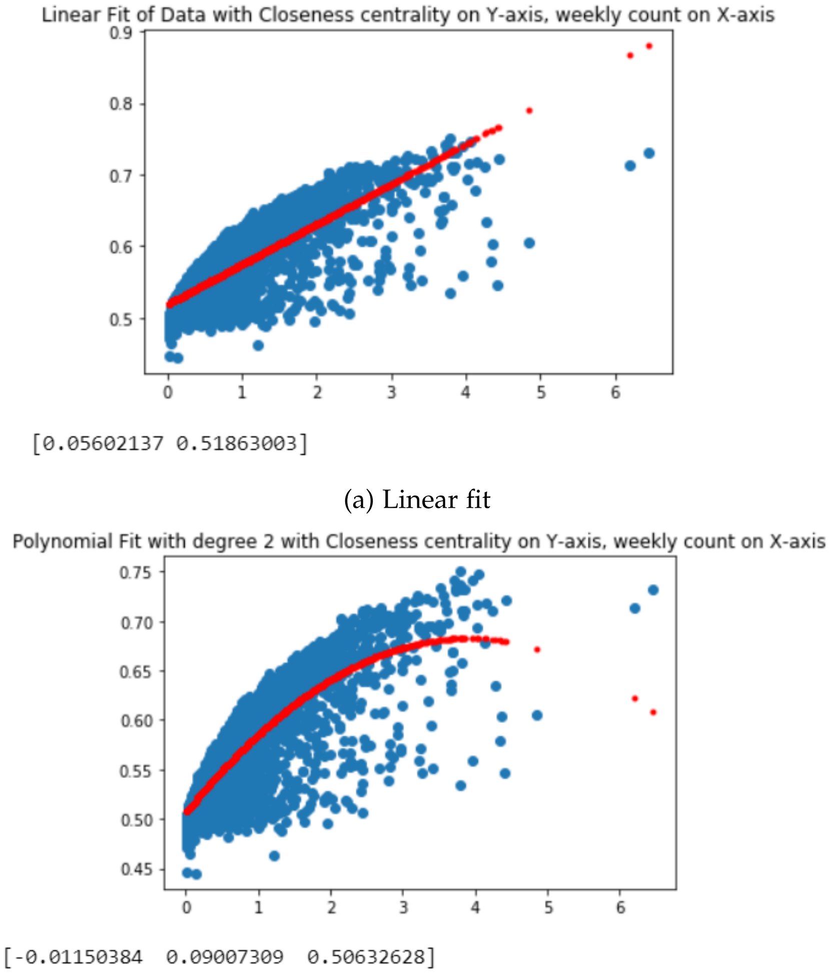
For the degree centrality, we can see that linear and polynomial of degree 2 do not fit the data that well. We try to increase the degree and see that a polynomial of degree 6 fits the data better, and the coefficients of the polynomial are present below the plot[Fig. 5.43].

Betweenness Centrality

For the betweenness centrality, a polynomial of degree 2 fits the data best [Fig. 5.44].

5.4.2 Linear regression with multiple variables

Using all the data features available, we trained a regression model and tried to predict the contract duration of the users based on these features. We used the backward elimination method to reduce the feature set to only the useful ones. Once we have a useful set of features, we train the model with different algorithms like linear regression, random forest regression, and decision tree regression. We found out from the results that decision tree regression was able to predict the results with a 65% accuracy [Fig. 5.45]. The reason being that this data is nonlinear and applying linear regression over it would not provide good results. Also, since the data features are limited and not



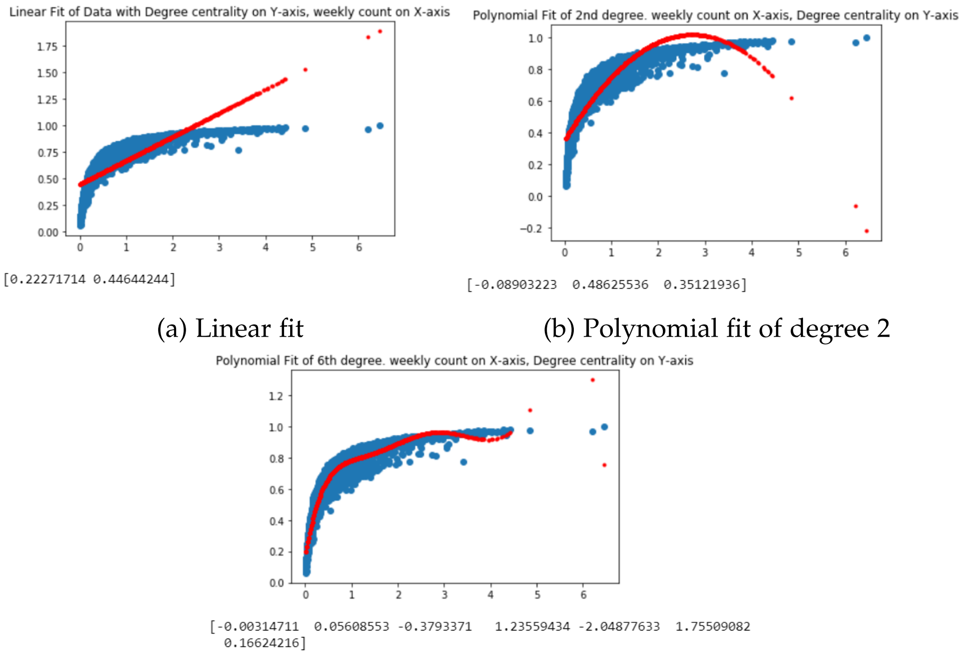
```
[0.05602137 0.51863003]
```

(a) Linear fit

```
[-0.01150384  0.09007309  0.50632628]
```

(b) Polynomial fit

Figure 5.42: Regression line fitting for weekly visiting frequency to closeness centrality plot



```
[0.22271714 0.44644244]
```

```
[-0.08903223  0.48625536  0.35121936]
```

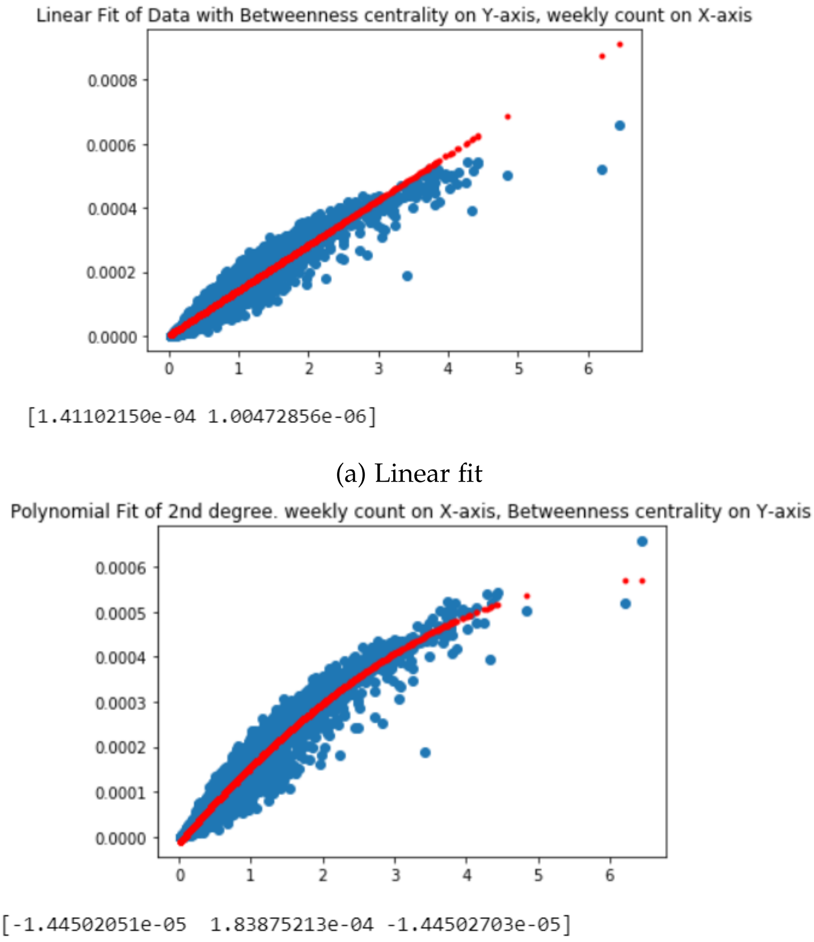
(a) Linear fit

(b) Polynomial fit of degree 2

```
[-0.00314711  0.05608553 -0.3793371   1.23559434 -2.04877633  1.75509082
  0.16624216]
```

(c) Polynomial fit of degree 6

Figure 5.43: Regression line fitting for weekly visiting frequency to degree centrality plot



```
[1.41102150e-04 1.00472856e-06]
```

(a) Linear fit

```
[-1.44502051e-05  1.83875213e-04 -1.44502703e-05]
```

(b) Polynomial fit

Figure 5.44: Regression line fitting for weekly visiting frequency to betweenness centrality plot

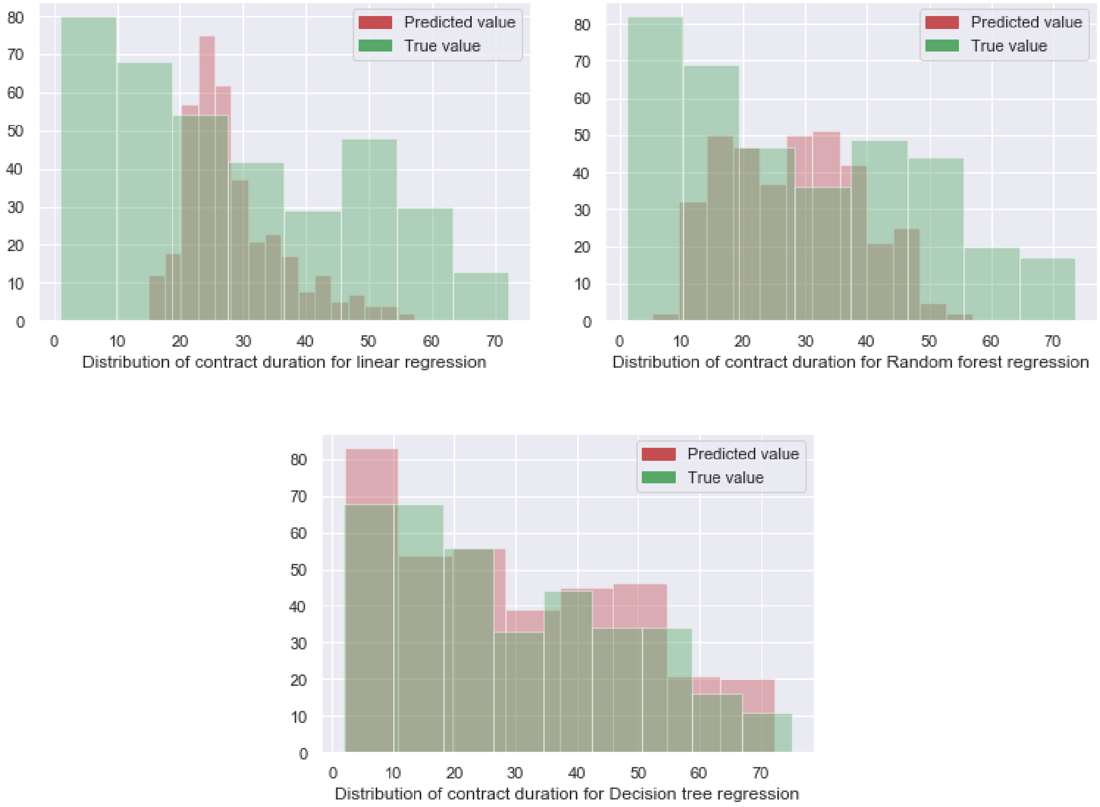
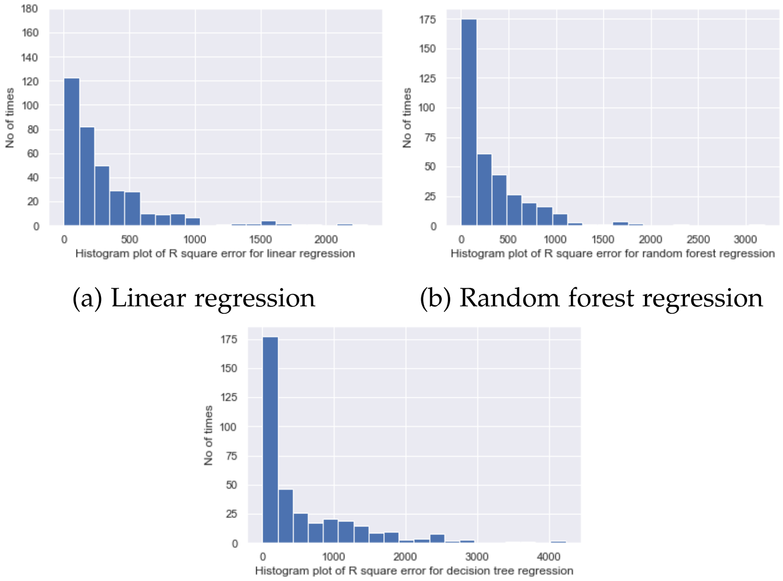


Figure 5.45: True and predicted contract duration distribution for different regression algorithms



(a) Linear regression

(b) Random forest regression

(c) Decision tree regression

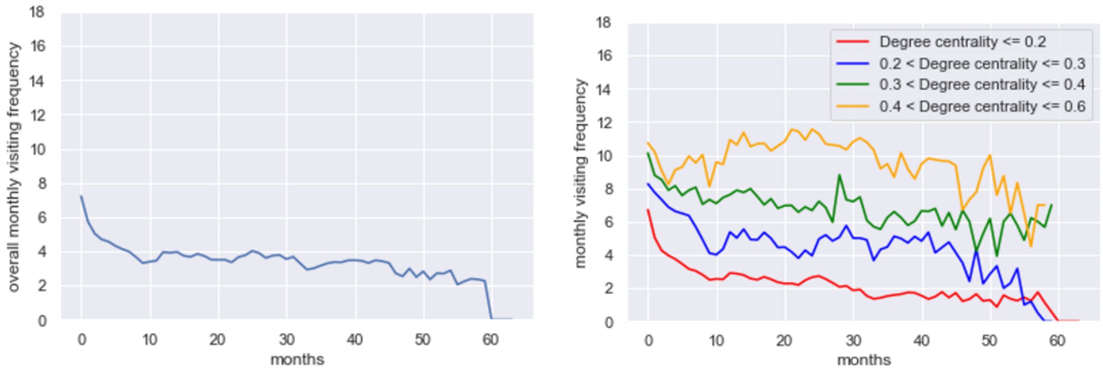
Figure 5.46: Histogram plot of R-Square error of various algorithms. The x-axis represents the R-Square error and y-axis shows count across different R-Square error

very complex, the Decision tree regression generates better results by understanding the non-linearity aspect of the data and making better decisions at the granular level.

We calculate R-squared error for each case and see that the error value is the highest for linear regression and lowest for decision tree regression [Fig. 5.46].

5.5 Visiting frequency to Centrality value

We calculate the visiting frequency of users from their first month till their last month, so see how a user utilizes his contact duration. We wanted to see if the motivation of the user stays intact throughout his contract duration or if it varies. We also wanted to see what impact the social network has over these visiting frequencies. We see from the results [Fig. 5.47] that there is a more significant dip in the monthly visits over time with the people who are socially less active. In contrast, the people who are socially



(a) Accumulative plot

(b) Plot based on centrality slabs

Figure 5.47: Monthly visiting frequency of users from their 1st month till 60th month

more active show a constant monthly visiting frequency proving the fact that social engagement does motivate users to work out regularly.

We plot a distribution of the monthly visiting frequency [Fig. 5.48] and see that the mean of the frequency increases with an increase in the centrality value. This result implies that as a person gets socially involved, he tends to visit the fitness club more.

5.6 Percentage of churned Friends

We also wanted to see how users get influenced to quit their membership whenever a close friend of theirs' quits. To calculate the same, we listed all the users who have quit their membership, then for each user, we calculated the number of friends who have also terminated their contract. Using this, we get a percentage of churned friends. We plot this in a graph and then calculate rolling window over the distribution.[Fig. 5.49]

We can see from the plot [Fig. 5.51] that with the increase in degree centrality value, which represents an individual's count of friends, the percentage of his friends also quitting the fitness club membership increases. We also did a similar experiment for betweenness centrality to see a similar result [Fig. 5.50].

We also clubbed users into different groups based on their centrality value and for each of these groups we did a distribution plot for the churned friend percentage as shown in [Fig. 5.52]. From this plot we can see that as the degree centrality of

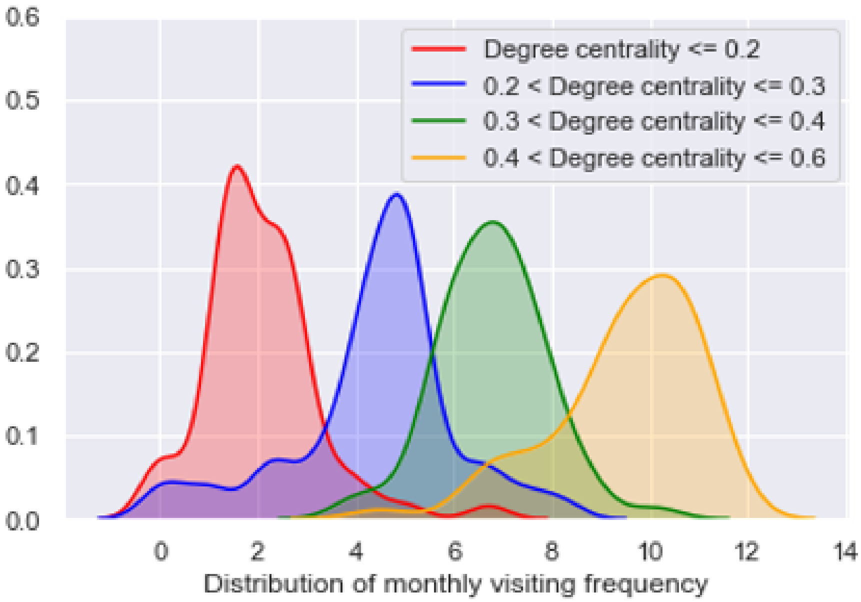
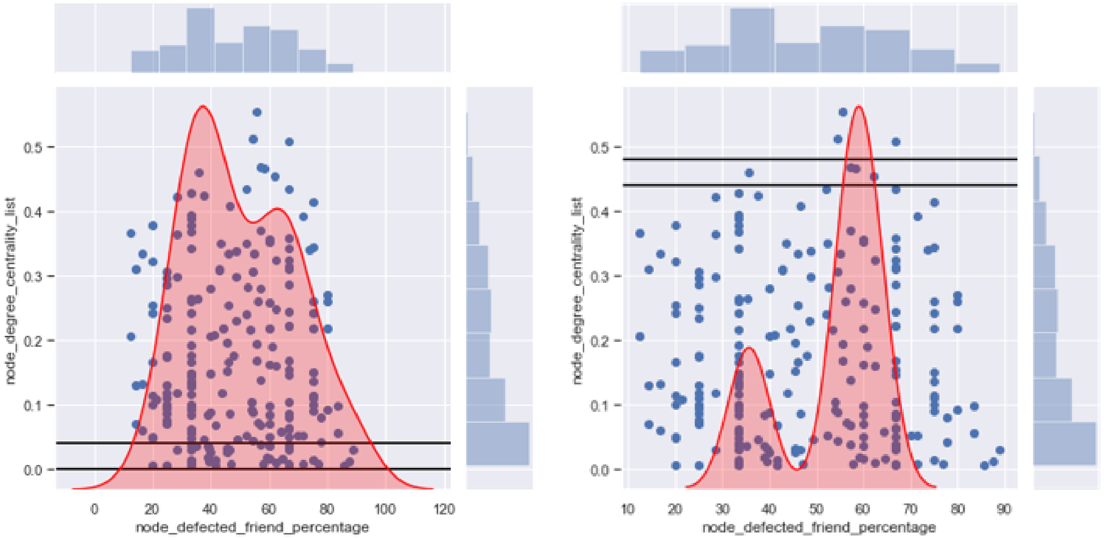


Figure 5.48: Monthly visiting frequency distribution based on centrality

user increases the mean percentage of churned friends increases. Also the standard deviation decreases. Both these observations imply that as a user gets socially active he starts dictating his friends' usage behaviour.

We can infer from the results that the social activeness of a user plays a significant role in defining his and his friend's user behavior.



(a) Starting position of rolling window

(b) Ending position of rolling window

Figure 5.49: Rolling window from lower to higher centrality value

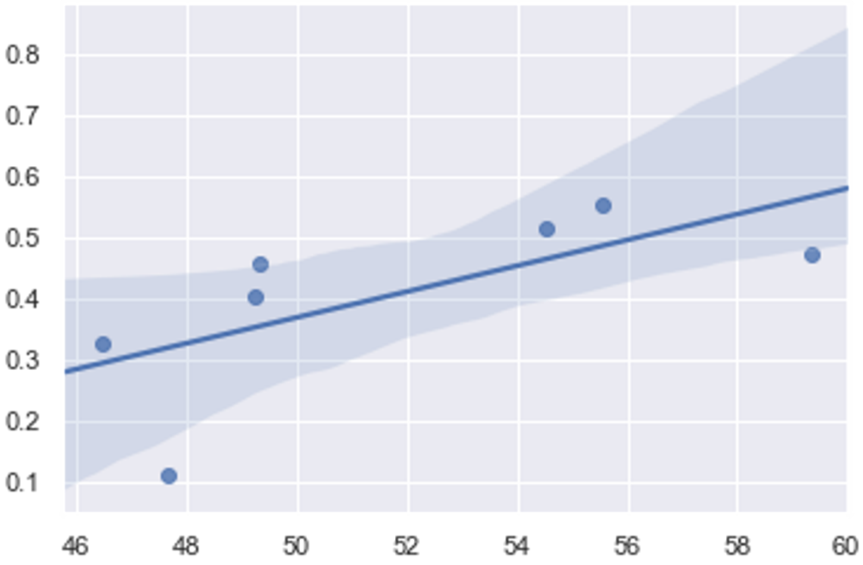


Figure 5.50: Regression plot of percent of churned friends on the x-axis and Betweenness centrality on the y-axis

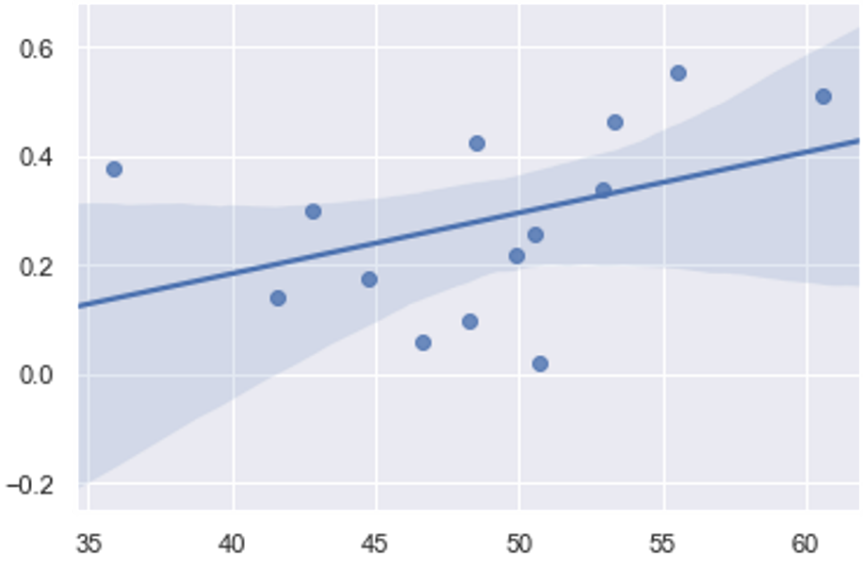


Figure 5.51: Regression plot of percent of churned friends on the x-axis and Degree centrality on the y-axis

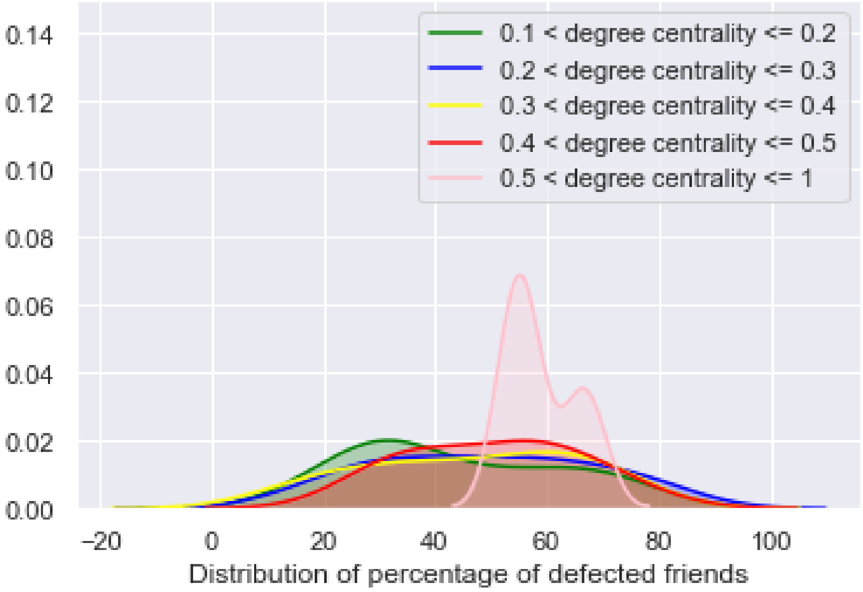


Figure 5.52: distribution of churned friends percentage for users with different degree centrality slabs.

6 Discussion and Conclusions

Contents

6.1 Conclusions

The main aim of this thesis was to see how social network impacts a customer's way of using the facilities provided to him. From the results, we could conclude that there is a significant influence on user behavior when he gets socially involved.

We carried out exploratory data analysis to see how the service was being used. We noticed that users usually prefer to workout around 18:00. We also did an age-wise usage distribution to see if a set of users prefer to exercise at a specific time of the day. The results showed that younger users prefer to hit the gym in the evening in contrast to the elderly users who prefer working out in the morning around 9:00. Also, the research showed that users prefer to workout over the weekday rather than the weekend. And the usage drops sharply during the summer. We could assume from the results that users involve more in outdoor activities than visiting an indoor fitness club.

The research was carried out at various levels. First overall but later by granulating the data based on the features available. We saw a considerable positive correlation between the user contract duration and visiting frequency to their social networking features and wanted to know if it's the same across different sections of the data and if some particular data chunks respond better to social engagement. We did a split of data first based on the type of exercise and found a high positive correlation for gym data than the group exercise data. To justify the findings, we could say that group exercises involve interactions by default, and users performing it achieve social prospects without putting much effort into it. Whereas, in the case of gym users, if they need socializing, they need to do it on their own. This effort of the gym users to create a social environment is visible from the results.

We further wanted to dig deep into gym users, so we classified the data based on two user features: gender and age group. From the results, we could conclude that the gender factor doesn't add much significance to the social network attributes, whereas age does. We created six groups based on the age factor, and based on their age group, they are added to the respective groups. The same social networking procedures were applied to different age groups, and the results showed that mid-aged users had higher social activeness. Also, the correlation between social activeness and user behavior was positive for this set of users. We can say that since the younger user prefers to workout in the rush hours, they tend to find more social connections than elderly users who want to exercise in the morning to find some privacy and space when they workout.

We also applied machine learning techniques like regression analysis to see if we could predict the contract duration of the users based on the knowledge of their social activeness. We used various regression models to carry out predictive analysis. We found that decision tree regression yielded the most accurate results when compared to linear regression and random forest regression. We started with simple linear regression and found the model could not predict better. Later we applied random forest and decision tree regression, and the accuracy of the model increased on every step. The reason being that the data is non-linear, and using a linear regression would not be fruitful. In such cases, decision tree regression would suit the best.

We also wanted to see how one's defection affects others at the fitness club. We first calculated a list of churned users, and for each of these users, we calculated the percent of churned friends. The results showed that with an increase in social activeness of a defecting user, the percentage of friends who churn out also increases. We plotted a regression plot for this and found that the slope was positive. A one unit increase in degree centrality caused 21.5% hike in churn percent of friends. This result is something fascinating to see and a case to be considered by the service providers when they plan on retaining their customers.

From the results, it understood that being socially active boosts up the user's fitness engagement at the club, so it is essential that the fitness club make the best out of this aspect. From the results, it was found that 27% of the users at the fitness club are socially active, and there is room for more people to get to know each other and get going with the club engagement. To promote this effect, service providers can encourage users to get their friends along when they workout. In this way, the fitness club can increase the retention rate of the users.

6.2 Limitations and Future Work

Since the steps involved a lot of processing and calculating starting from data cleansing to building a social network model and then applying state of the art techniques like machine learning over the results, we had a time constraint to do all this processing. Because of this, we had to limit ourselves by selecting only a subset of users and carry out the process on them. In our study, we chose only the best two locations as our dataset and carried out social network analysis with that data. Based on the findings of this thesis, there is a potential to extend this thesis to more locations and more service providers as well. Also, since social networking is a human phenomenon, it could be a potential game-changer for various other service providers, which involves human interactions of some sort. Since the number of researches in this field is minimal, it could be seen as a great topic to dig deeper in length and breadth and see what best could be utilized out of the results that would help the customer as well as service providers.

Throughout this study, the social network is seen as a static piece of data. Due to time restrictions and many other constraints, the dynamic aspect of the social network model could not be considered. The way a network evolves and the different phases of the evolution would be something interesting to look upon as future work on this topic. It would be exciting and promising to see how social networking changes over the customer's lifetime, how long a customer can be retained, and the effects of the time factor on social relationships. This work would help us understand the extent to which the social network plays a role in customer retention, is it a continues process or is there some threshold beyond which it would stop impacting or starts to fade off.

List of Figures

List of Tables

Bibliography

[1] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data*. springer, 2006.

[2] S. Aral and D. Walker. "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks." In: volume 57 (9). 2009.

[3] G. Benedek, Á. Lublóy, and G. Vastag. "The importance of social embeddedness: Churn models at mobile providers." In: volume 45 (1). 2014, pages 175–201.

[4] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[5] K. Eremenko, H. de Ponteves, S. Team, and S. Support. *Machine Learning A-Z^{TM}: Hands-On Python R In Data Science*. Online course. 2020.

[6] L. C. Freeman. "A Set of Measures of Centrality Based on Betweenness." In: volume 40 (1). 1977, pages 35–41.

[7] L. C. Freeman. *Centrality in social networks conceptual clarification*. 1979.

[8] M. FRIENDLY and D. DENIS. "THE EARLY ORIGINS AND DEVELOPMENT OF THE SCATTERPLOT." In: volume 41(2). 2005.

[9] M. Galarnyk. *Understanding Boxplots*. 2018.

[10] A. A. Hagberg, D. A. Schult, and P. J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX." In: 2008.

[11] S. B. Jarrell. *Basic statistics*. Dubuque, Iowa : Wm. C. Brown Pub., 1994.

[12] J. Kratzer, C. Lettl, N. Franke, and P. A. Gloor. "The Social Network Position of Lead Users." In: volume 33 (2). 2015, pages 201–216.

[13] matheguru. *matheguru dispersed data*. [Online; accessed January 31, 2020]. 2011 - 2019.

[14] matheguru. *matheguru non-dispersed data*. [Online; accessed January 31, 2020]. 2011 - 2019.

[15] P. M. M. D. Miguel. "Behavioral Clustering of Translational Data." Master's thesis. TECHNISCHE UNIVERSITÄT MÜNCHEN, 2019.

[16] I. Nitzan and B. Libai. "If You Go, I Will Follow . . . Social Effects on the Decision to Terminate a Service." In: volume 5 (2). 2014, pages 40–45.

[17] E. Otte and R. Rousseau. "Social Network Analysis: A Powerful Strategy, also for the Information Sciences." In: volume 28(6):441-453. 2002.

[18] P. Patil. *What is Exploratory Data Analysis?* 2018.

[19] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Springer, 2010.

[20] N. J. Salkind. *Statistics for People who (think They) Hate Statistics: The Excel Edition*. Sage Publications, 2006.

[21] R. Sassatelli. *Fitness Culture: Gyms and the Commercialisation of Discipline and Fun*. springer, 2010.

[22] V. Sekara, A. Stopczynski, and S. Lehmann. "Fundamental structures of dynamic social networks." In: volume 113 (36) 9977-9982. 2016.

[23] M. E. Spear. *Charting statistics*. New York, McGraw-Hill, 1952.

[24] M. E. Spear. *Practical charting techniques*. McGraw-Hill, 1969.

[25] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Dubuque, Iowa : Wm. C. Brown Pub., 1995.

[26] H. Wickham and L. Stryjewski. *40 years of boxplots*. 2011.

[27] Wikipedia, the free encyclopedia. *Centrality*. [Online; accessed January 31, 2020]. 2020.

[28] Wikipedia, the free encyclopedia. *Interquartile Range*. [Online; accessed January 31, 2020]. 2020.