KLE Society's

KLE Technological University


# EXPLORATORY DATA ANALYSIS

## COURSE PROJECT

## REPORT
## on
# Relapse Patients Data Analysis

Submitted by : D1 - Team 06

| Roll.no | Name | Usn |
|---|---|---|
| 417 | Michael Rohan Swaminathan | 01FE20BCS201 |
| 410 | Utkarsh Khot | 01FE20BCS194 |
| 406 | Naveen Ballari | 01FE20BCS190 |
| 407 | Azeem Jalageri | 01FE20BCS191 |

Guided by

Dr. P. G. Sunitha Hiremath

# Contents

# 1. Abstract

One of the most important products of global addiction demand is an alcoholic beverage. In developing countries like India, alcohol consumption tends to be a major problem because of the various socio-cultural practices across the nation, different alcohol policies and practices across the various states, lack of awareness of alcohol-related problems among the community, false mass media propaganda about alcohol use, various alcohol drinking patterns among the alcohol consumers and the emergence of social drinking as a habit because of the widespread urbanization across the country. Stringent alcohol policies are needed across the various states to reduce alcohol consumption, and alcohol consumers have to be educated about the various harmful effects of alcohol consumption and the effects it can have on their mind and body. The relapse of patients into alcoholism is a major concern in our society. So, by identifying the important patterns and factors that make a person relapse would help us in identifying the proper treatment we need to provide to them either by medicines or through different physical activities. Also, identifying the surrounding environment of the patients would help us in analyzing the present trend of activities in our society and necessary actions for the same can be taken.

# 2. Introduction

As a team of four, we have obtained a data set from a camp that was conducted by SDM Medical College, Dharwad which contains data about the patients who have relapsed and presently seeking treatment for their illness. The camp is of 8 days which monitors the activities of the patients and provides bajanas and other physical activities to the patients so that their mind and body may be made strong to withstand any inclination to consuming alcohol or any other substance usage. They are also fed well in the camps to make them physically stronger and make them inclined to having a good meal rather than consuming a bottle of alcohol. Their socioeconomic information including employment, education, income, marriage details, family details, alcohol and other addiction details are collected from them in order to facilitate our analysis. We analyze the trends in their period of sober and analyze each period of sober group for the contributing factors to each group. Then, we try to predict the period of relapse so as to help the medical authorities provide the appropriate treatment for them primarily through proper dosage of medicines, counseling and physical activities.

# 3. Problem Statement

The challenge focuses on analyzing the factors that affect the period of sober of the patient and as a result, predict the relapse period of the patients.

**“ What remains in diseases after the crisis is apt to produce relapses”**

**-Hippocrates, Aphorisms**

# 4. Data Description

- Relapse: (of a sick or injured person) deteriorate after a period of improvement
- A relapse happens when a person stops maintaining their goal of reducing or avoiding use of alcohol or other drugs and returns to their previous levels of use.
- The patients undergo rehabilitation in camps to overcome their addiction and increase their period of sober.
- The prediction of the relapse period will help the medical expertise provide the appropriate treatment to the patient.
- Data comprises of patient's socioeconomic information including employment, education, income, marriage details, family details, alcohol and other addiction details.

## Critical attributes selected

| Attributes | Attributes | Attributes | Attributes |
|---|---|---|---|
| Age | Smoking/Smokeless | Any instance of family violence | Occupational damage |
| Education in year | AAO for Alcohol in Year | year of Treated | Describe your childhood experiences |
| Annual Income | average units in the last 30 days | Period of Treatment | Maximum period of abstinence |
| Marital Status | Motivation Factor | Period of sober | Achievement in childhood |
| Nicotine (Yes/No) | duration of use of alcohol | Stressors | At what age did you start working? |
| Reason for starting alcohol | duration of excessive use of alcohol | Behaviour Problems Identified in childhood | How long have you been working? |
| Chronic Health Problem | Withdrawal symptoms | Psychiatric illness | Family History of Alcoholism |

# 5. Data Pre-Processing

## Preprocessing of the data includes:

- Age had a missing values percentage of 1.47, so we plotted the distplot of the data and identified it was Positively skewed.
- Also, The skew() function was used and it gave a value of 0.36
- So, from the above two observations we replaced the NaN Values with the median as it is more resistant to outliers than mean, which was around 40.
- The above method was followed for Period of Sober and Remarks which had a percentage of 0.49, and certain attributes like Education in year, year of treated had '-' which we also considered for replacement.
- Mode was used to fill the categorical columns missing values like Remarks and Marital history details.
- The Religious questions had a missing value percentage of 36.45 and the values were common in it which would add up as it is a categorical column, so it was discarded from the dataset.
- As the data consisted of objects, we used string extraction techniques using replace and regex to extract the numerical content in the attribute and also correct the grammatical mistakes in the categorical column which were then one hot encoded.
- All the attributes were converted into their proper data type, mainly numerical for feature extraction and model prediction techniques to be applied to it.

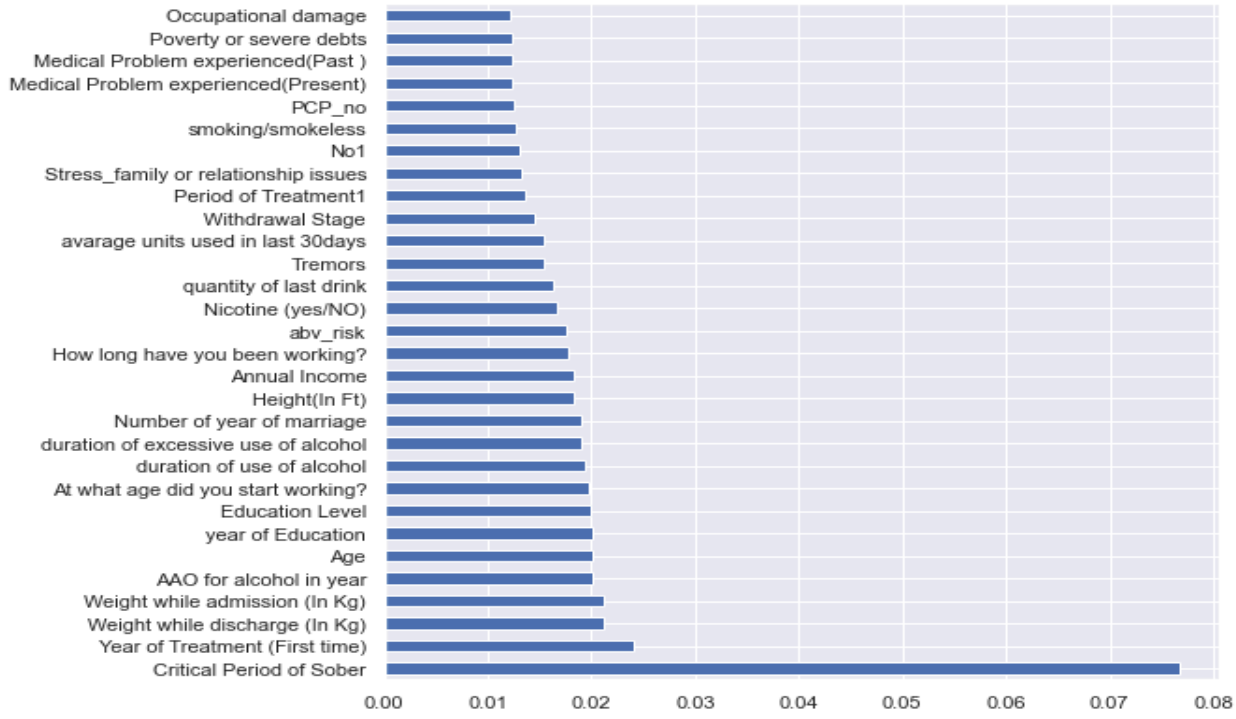
## Conversion of Multivalued categorical attributes to One hot encoded attributes

- The data was pattern exploded into its subsequent columns, of which 7 was the highest values in a column.
- Then the grammatical mistakes were corrected in each columns for one-hot encoding.
- Then we use the dummies function to one-hot encode the attributes and then combined them.

# After pre-processing of the data :

| Withdrawal Symptoms Faced by Patients | Aches and pains | Anxiety | Auditory hallucinations or illusions | Fits | Insomnia | Nausea | Palpitation | Restlessness | Sweating | Transient visual or tactile | Tremors | Weakness | No |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sweating, Tremors, Fits | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Tremors, Insomnia | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Tremors, Insomnia, Anxiety | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| Tremors, Insomnia, Auuditory hallucinations | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| Sweating, Tremors, Nausea, Anxiety, Restlessness, Transient visual | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Sweating, Tremors, Nausea, Aches and Pains, Transient Visual, Auditory Hallucinations | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| No | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Sweating, Tremors, Insomnia, Nausea | | | | | | | | | | | | | |
| Tremors, Insomnia, Fits | | | | | | | | | | | | | |

# Feature Selection Techniques



## Analysis

- Using inbuilt class feature_importances of tree based classifiers, we can identify the attributes that will contribute to the analysis and prediction of the sober period range.
- The important attributes of number of relapses, frequency and quantity of drinking, withdrawal symptoms and withdrawal stages of the patient will be taken into consideration.
- Also the childhood experiences, work experience and reasons for starting alcoholism will be taken into account for the identification of the patients reasons to indulge in alcoholism

| Specs | Score |
|---|---|
| Annual Income | 7328258 |
| Critical Period of Sober | 192473.7 |
| Period of Treatment1 | 63577.89 |
| average units used in last 30days | 6608.901 |
| quantity of last drink | 5761.552 |
| abv_risk | 1357.443 |
| Year of Treatment (First time) | 482.0481 |
| duration of use of alcohol | 191.0649 |
| BCA_breaking articles at home | 178 |
| How long have you been working? | 146.8842 |
| Number of year of marriage | 130.6066 |
| duration of excessive use of alcohol | 129.5516 |
| CHP_gastric | 122.7917 |

## Analysis

- Applied SelectKBest class to extract top 10 best features.
- The above features, tree based classifiers and the recursive feature elimination attributes will be taken into account for the analysis of the attributes.

| Feature | Rank |
|---|---|
| Medical Problem experienced(Past ) | 1 |
| MF_Moderate | 1 |
| Poverty or severe debts | 1 |
| Transient visual or tactile | 1 |
| Sweating | 1 |
| Palpitation | 1 |
| Nausea | 1 |
| Insomnia | 1 |
| Auditory hallucinations or illusions | 1 |
| Anxiety | 1 |

## Analysis

- Feature ranking with recursive feature elimination and cross-validated selection of the best number of features
- Using linear regression as the model to identify the most important features.
- The above attributes will be analyzed using visualization techniques.

# 6. Questions and analysis with graphs and inference

1. **What is the total number of patients who have relapsed once, their distribution over the ranges of period of sober and the withdrawal stages they are in?**

Analysis

- Figure 1 details about the number of relapses a patient had of which 179 patients have relapsed once.
- Figure 2 details the distribution of the patients who have relapsed once into their ranges of period of sober of which 18 patients have a period of sober less than 15 days and 9 patients have a period of sober between 15-30 days.
- Figure 3 tells us the withdrawal stages each patient falls in their respective ranges of period of sober of which 9 patients are in first stage, 4 patients are in second stage and 5 patients are in third stage in the range 0-15 days

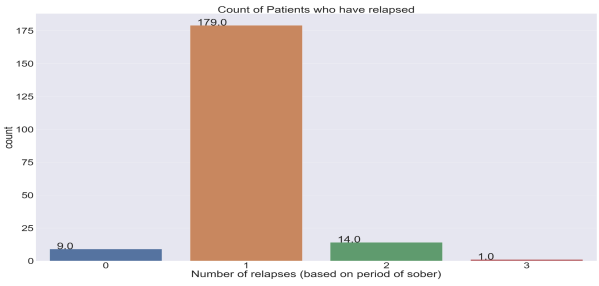



Figure. 1

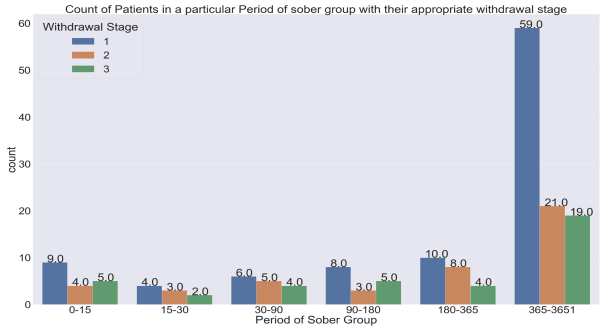



Figure. 3

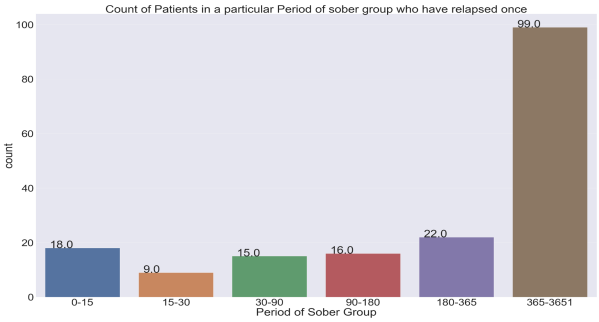



Figure. 2

## 2. What is the distribution of the age groups on the period of sober and the causes for alcoholism?

### Analysis

- Figure 4 details about the spread of the age groups among the period of sober ranges of which the range of 35-53 has 14 patients in the 0-15 days period of sober range.
- Figure 5 details one of the causes for alcohol addiction being someone in family or friends were using of which 14 patients in the range of 0-15 days were affected by it and all the patients in the range of 15-30 days were affected by it.
- Figure 6 shows another cause for addiction being the childhood experience of which poverty and severe debts was identified to have affected 10 patients in the range of 0-15 and 30-90 period of sober.

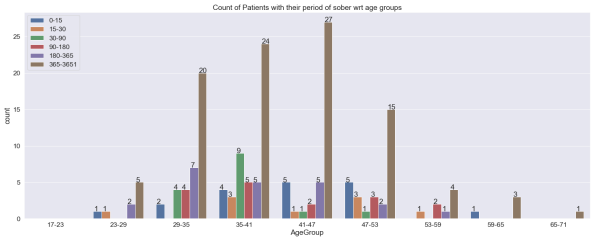



Figure. 4

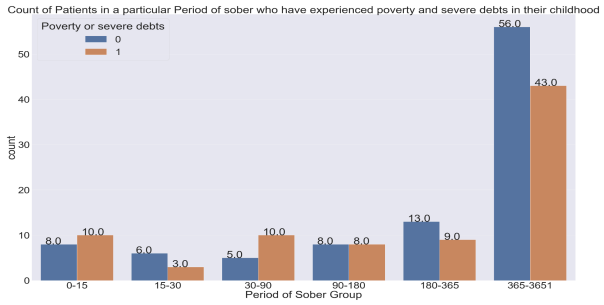



Figure. 6

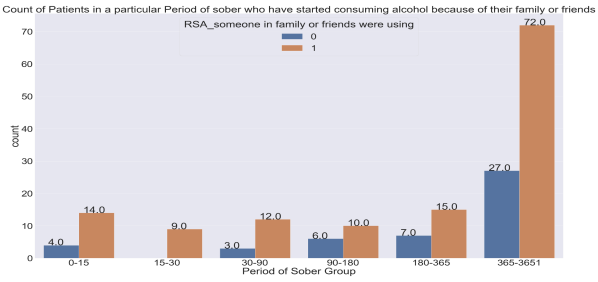



Figure. 5

## 3. How is the educational level related to the period of sober?

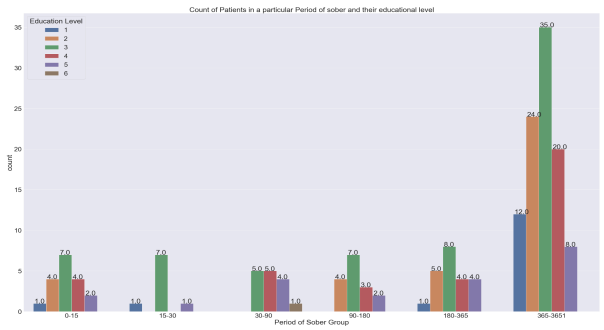



Figure. 7

### Analysis

- The education status of the patients were classified into separate levels.
- In the 0-15 period of sober group, 7 patients were in their Secondary stage that is, 10th std and majority of the patients lie in that educational level.
- The 15-30 and 90-180 period of sober group also have maximum patients from the secondary stage.
- So, the critical period of sober groups have patients who have done their matriculation.

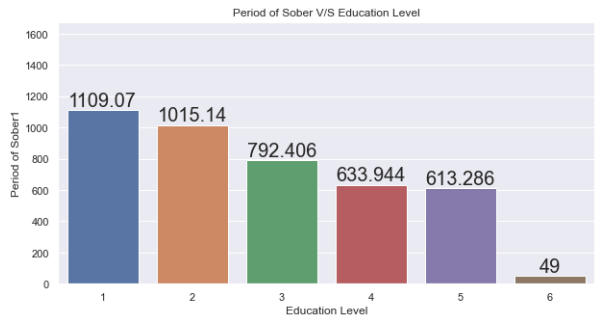



Figure. 8

### Analysis

- The mean sober period across the educational levels were analyzed.
- It is found that as there is an increase in the educational levels, there is a decrease in the sober period of the patients.
- Indicating the upper educational level has a tendency to relapse faster.

**4. What are the drinking patterns of the patient and its effect on the period of sober?**

**Analysis**

- Figure 9 tells us about the period of sober over duration of use of alcohol, patients having 15-30 days and 90-180 days of period of sober have more duration of use of alcohol.
- In figure 10, relation between period of sober vs average units of alcohol used in last 30 days, patients having 0-15 days and 180-364 days of period of sober have almost similar average units of alcohol around 500 ml of alcohol.
- In figure.11 we can observe that patients having period of sober above 180 days and between 15-30 have excessive usage of alcohol.

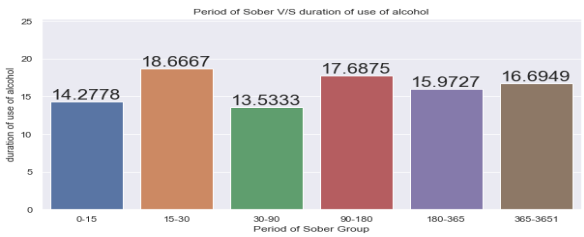



Figure. 9

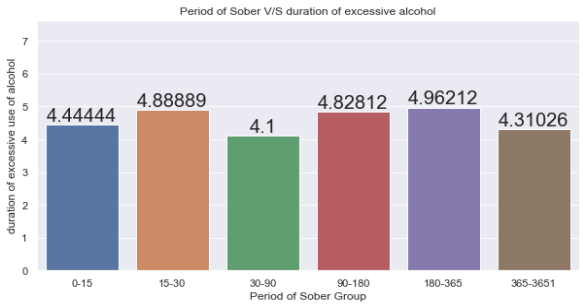



Figure. 11

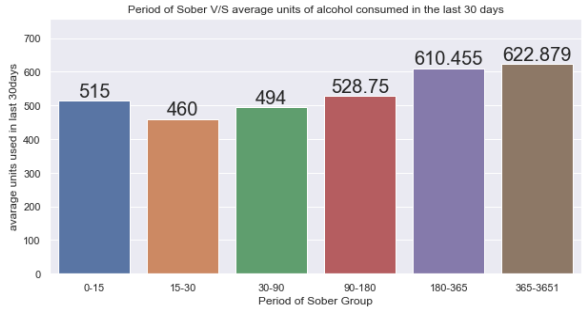



Figure. 10

**5. Is Nicotine consumption and smoking prevalent among the patients and its effect on their period of sober?**

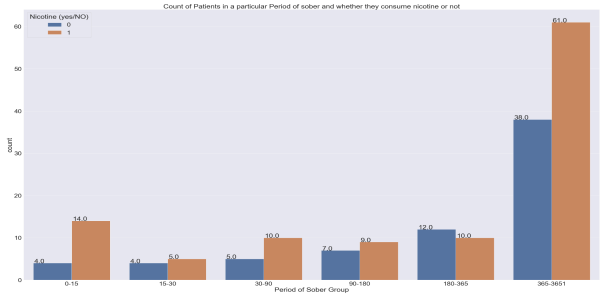



Figure. 12

**Analysis**

- Figure 12 details about the period of sober vs count of patients in each group whether they consume nicotine or not.
- We can observe that, patients having more than 365 days of period of sober have nicotine consumption .
- The 0-15 period of sober group has 14 patients who consume nicotine and the 30-90 has 10 patients who consume nicotine.

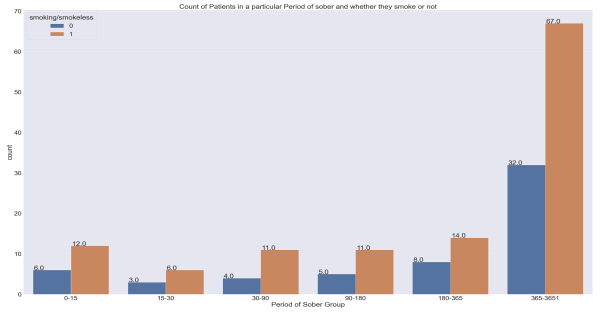



Figure. 13

**Analysis**

- Figure 13 shows the relation between period of sober vs count of patients in each group whether they smoke or not.
- We can observe that, patients having more than 365 days of period of sober have a smoking habit too.
- The 0-15 period of sober also as 17 patients who are addicted to smoking and the 15-30 also has 6 patients who are addicted to smoking.

**6.What is the risk level of the patients of a particular period of sober range and have they had any medical experiences in the past?**

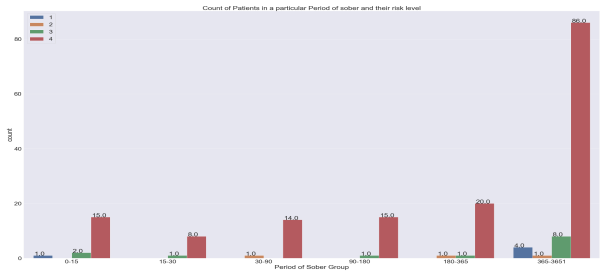



Figure. 14

**Analysis**

- The risk factor was calculated based on the type of alcohol and the average quantity consumed in the last 30 days.
- The risk factor comprises of Low, Moderate , High and very High Level of risk level.

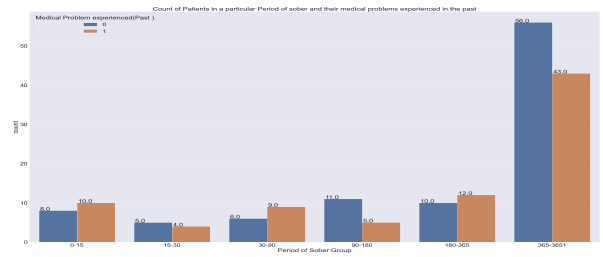



Figure. 15

**Analysis**

- The past medical experiences of the patients has been used to compare with period of sober groups.
- The 0-15 period of sober group has 10 patients, the 30-90 period of sober group has 9 patients and 180-365 period of sober group has 12 patients who have past medical experiences which is more than those who don't have in that particular period.

**7. How many patients have moderate level of denial of substance use related problems and motivation factor towards substance use?**

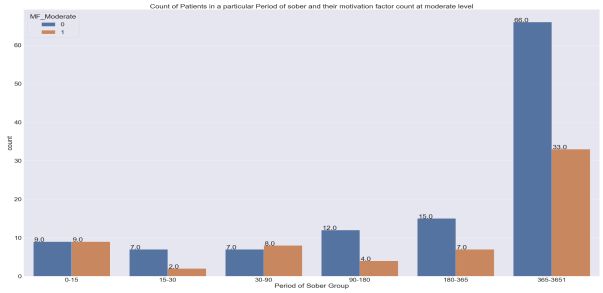



Figure. 16

**Analysis**

- In Figure.18, relation between count of patients in a particular period of sober vs their motivation factor count at moderate level is shown.
- It is observed that, the patients having 365 - 3651 days of period of sober haven't got motivation factor at moderate level towards substance use and in 0-15 days group 50% of them that is, 9 patients have motivation factor towards substance use.
- The 30-90 period of sober group has 8 patients who have a motivation factor towards substance use.

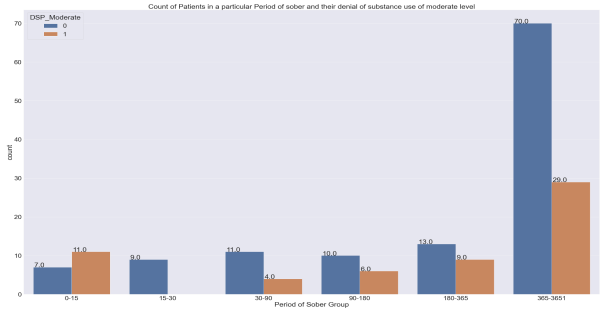



Figure. 17

**Analysis**

- In Figure.19, graph plotted is of count of patients in a particular period of sober vs their denial of substance use count at moderate level is shown.
- In 0-15 period of sober group many of the patients have denial to substance use at a moderate level that is 11 patients.
- It is observed that, the patients having 365 - 3651 and other groups of days of period of sober don't have denial of substance at moderate level towards substance use.So,

  lesser the denial of substance greater the period of sober.

**8. How many patients who have relapsed twice who have family history of alcoholism and their risk level**

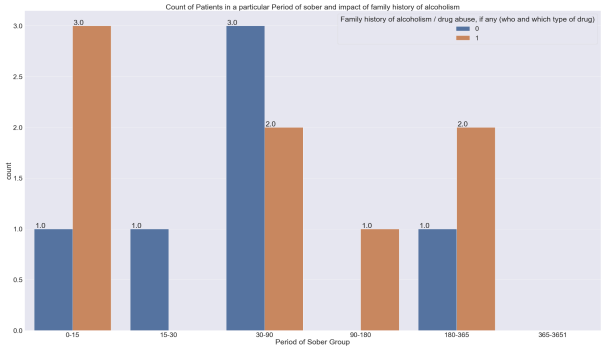



Figure. 18

**Analysis**

- Figure.20 shows the the graph between Count of Patients having family history of alcoholism and Period of Sober Group.
- 0-15 days of sober period group have maximum no. of patients i.e 3 who had stressors related to family history.
- In 30-90 period, Many patients have deniel to any family history related to alcoholism i.e 3.
- Beyond 365 days of sober period,there are no patients having any family history or relationship issues.

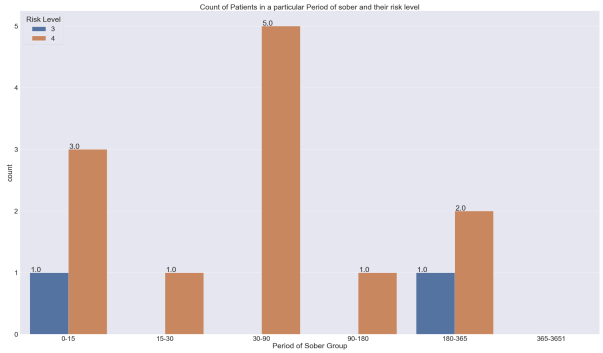



Figure. 19

**Analysis**

- Figure.21 shows the relationship between Count of patients having risk level and period of sober group.
- In 0-15 and 180-365 period,there are only 1 patients having risk levels 3.
- In 15-180 period,no patients are having risk level 3.
- We can see that,In 30-90 days of period,Maximum patients i.e. 5 are having risk levels of 4
- In 365-3651 period,there are no patients having any risk levels.
- From this we can say that,As the sober period increases the risk levels in patients decreases.

# CONCLUSION

- The dataset mainly concentrates on patients having agricultural background.
- The educational level was found to be maximum of level 3 that is, 8th, 9th and 10th.
- Majority of the patients had relapsed only once, that is 179 patients out of 203, this data was sampled for further analysis.
- One of the cause for patient's relapse was identified to be poverty or severe debts which they confronted during their childhood/teenage and reason for indulging in alcoholism was due to someone amongst their family or friends who were alcoholics.
- Majority of the patients lie in the very high level risk of abv which was calculated based on the type of alcohol they had consumed and the average units consumed in the last 30 days.
- Patients were not involved in any other extracurricular activities such as sports etc which allowed them more leisure time to be involved in overconsumption.
- From the comparison between the AAO for alcohol in year and at what age they started working it was evident that they started consuming alcohol before they started working.
- From withdrawal symptoms we have identified that insomnia and tremors are the major symptoms faced by the patients when they decline having alcohol after a period
- The twice relapsed patients who were in the critical period of sober of 0-15 had family history of alcoholism which could be a contributing factor towards their relapse and addiction.