MSD DATATHON REPORT

Team 8

22.05.2023

—

Team Members:

Julianna Stermer

Raghav Jha

Saurav Jadhav

Suyash Wagh

Introduction

Substance abuse is a complex and multifaceted problem in the United States, affecting individuals of all ages, races, and socioeconomic backgrounds. The term "substance abuse" refers to the use of drugs or alcohol in a way that leads to negative consequences, such as impaired functioning, health problems, social problems, and legal issues.

According to the National Survey on Drug Use and Health (NSDUH), an estimated 21.2 million Americans aged 12 or older struggled with a substance use disorder in 2020. This includes individuals who meet the criteria for a diagnosis of substance abuse or dependence, as defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). The most commonly abused substances include alcohol, marijuana, prescription opioids, and illicit drugs such as cocaine and heroin.

Despite the prevalence of substance abuse in the United States, only a small percentage of individuals receive treatment for their addiction. In 2020, only 4.2 million people received any type of treatment for substance use disorder, representing less than 20% of those who needed it. This treatment gap is a significant problem, as untreated substance abuse can lead to a range of negative outcomes, including chronic health problems, legal issues, and even death.

A variety of substance abuse treatment programs are available throughout the country, including outpatient counseling, residential treatment centers, and medication-assisted treatment. Outpatient counseling typically involves individual or group therapy sessions, while residential treatment centers provide more intensive care in a structured, 24-hour setting. Medication-assisted treatment involves the use of medications, such as methadone or buprenorphine, to help manage withdrawal symptoms and cravings for opioids or alcohol.

Despite the availability of treatment options, there are still significant barriers to accessing care. One major obstacle is the lack of insurance coverage for substance abuse treatment. Many individuals do not have insurance that covers addiction treatment, and those who do often face high deductibles and co-payments. This can make treatment unaffordable for many people, particularly those with low incomes.

Stigma is another significant barrier to accessing substance abuse treatment. Many individuals are reluctant to seek help for their addiction due to the shame and stigma associated with substance abuse. This can be especially true in specific communities, where substance abuse may be viewed as a moral failing rather than a medical condition.

Finally, there are also significant disparities in access to substance abuse treatment, particularly in rural and low-income areas. In some parts of the country, there are few or no treatment options available, making it difficult for individuals to access care.

In conclusion, substance abuse remains a significant problem in the United States, with millions of individuals struggling with addiction. While there are a variety of treatment options available, there are still significant barriers to accessing care, including lack of insurance coverage, stigma, and limited availability of services in certain areas. Addressing these barriers and expanding access to substance abuse treatment will be critical for addressing the ongoing addiction crisis in the United States. This will require a multifaceted approach that includes increased funding for treatment programs, education, and awareness campaigns to reduce stigma and policies that expand insurance coverage for addiction treatment. By working to address these challenges, we can help more individuals struggling with substance abuse access the care they need to achieve recovery and improve their overall health and well-being.

Goal(s)

1. The goal is to determine whether there is evidence of disparities in treatment access based on various demographic factors, such as income, ethnicity, type of drug abuse, and other relevant variables.
2. Similarly, the goal is to determine which treatment-related factors are most strongly associated with different discharge reasons, such as treatment completion, dropout, termination by facility, transfer to another program or facility, death, incarceration, or other reasons.

For both goals, we refer to the file *treatments_2017-2020.csv*

Exploring the goals in detail

- ### Goal-1

  Analyzing the relationship between the waiting time for access to treatment and the different attributes related to the treatment is important for several reasons.

  - Disparities in treatment access can contribute to broader health inequities and perpetuate systemic inequalities. Individuals from marginalized communities, such as low-income or minority populations, may already face barriers to accessing healthcare services, and disparities in substance abuse treatment access can further exacerbate these inequities. This can contribute to a cycle of poor health outcomes and social disadvantage, perpetuating health disparities across generations.

  - For individuals, longer wait times for treatment can lead to lost productivity, missed workdays, and increased healthcare costs. For society, untreated substance abuse can impose significant economic costs, such as healthcare expenditures, lost productivity, and criminal justice costs.

  - If individuals perceive that certain groups are being denied equal access to treatment, this can breed resentment and distrust, making it more difficult to address the root causes of substance abuse and addiction and provide effective treatment to those in need. This can ultimately undermine public health efforts to reduce the prevalence of substance abuse and addiction in the United States, with negative consequences for individuals, families, and communities.

  - Hence it becomes important to understand the attributes which are associated with the waiting time for the treatment and look at those attributes in depth.

Goal-2

Analyzing the relationship between the reason for discharge and the different attributes related to the treatment is important for several reasons.

- It can help to identify factors that contribute to successful treatment outcomes, such as completion of treatment. By identifying which client attributes are associated with successful treatment completion, treatment providers can tailor interventions to better meet the needs of their clients and improve the likelihood of successful outcomes.

- It can help to identify disparities in treatment outcomes based on factors such as ethnicity, income, or other demographic characteristics. This can help to identify areas where improvements are needed to ensure that all individuals have equal access to effective substance abuse treatment.

- It can help to identify areas where treatment interventions may need to be improved to better meet the needs of clients. For example, if a particular demographic group is more likely to drop out of treatment, this may indicate that interventions need to be adapted to better engage and retain individuals from that group.

- It can help to inform policy decisions related to substance abuse treatment. For example, if individuals with certain characteristics are more likely to be terminated from treatment, this may indicate a need for policy interventions to improve the quality of care provided in substance abuse treatment facilities.

- Overall, analyzing the relationship between discharge reasons and client attributes is important for improving the effectiveness and accessibility of substance abuse treatment and addressing the ongoing addiction crisis in the United States.

- By identifying factors that contribute to successful treatment outcomes and disparities in treatment access and outcomes, we can develop more tailored and effective treatment interventions and policy solutions that better meet the needs of individuals seeking help for substance abuse and addiction.

Data Pre-Processing

1. Before doing the analysis, we first make sure to have all the variables in the categorical type. Since many variables are encoded in the number. Using these variables does not truly conserve the information.

   For example, In the data *treatments_2017-2020.csv,* the variable *DAYWAIT* takes the value from -9 to 0. But here, -9 actually represents Missing/Unknown/Not Collected/Invalid value. Hence we convert these numbers from -9 to 0 into their respective categories as provided in the data description file.

2. In the *treatments_2017-2020.csv* data, there are many missing values. But we leave it untouched and consider it as a separate category.

3. We select those variables for analysis which has less than 54 categories.

Exploratory Analysis for Goal-1

Is there a relationship between different ethnic groups and waiting time (in days) for the treatment?

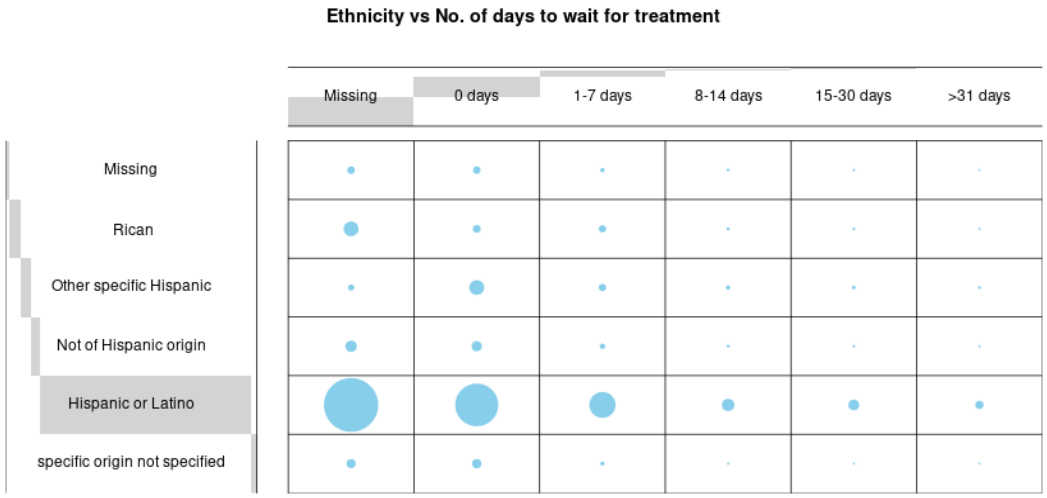

Figure 1: Ballon Plot of Ethnicity vs No. of days to wait for the treatment

From the figure 1, we can see that:

- The majority of the Missing/unknown/not collected/invalid days are for the Hispanic or Latino Race.
- Other specific Hispanic race gets access to the treatment relatively early.
- The Rican race has a relatively lot of Missing/unknown/not collected/invalid values for the number of days to wait for the treatment.
- Other specific Hispanic races have relatively fewer Missing/unknown/not collected/invalid values.
- Overall, we cannot conclusively talk about the relationship between different ethnic groups and the waiting time for the treatment.

Is there a relationship between the client's primary income source(s) and the waiting time for the treatment?

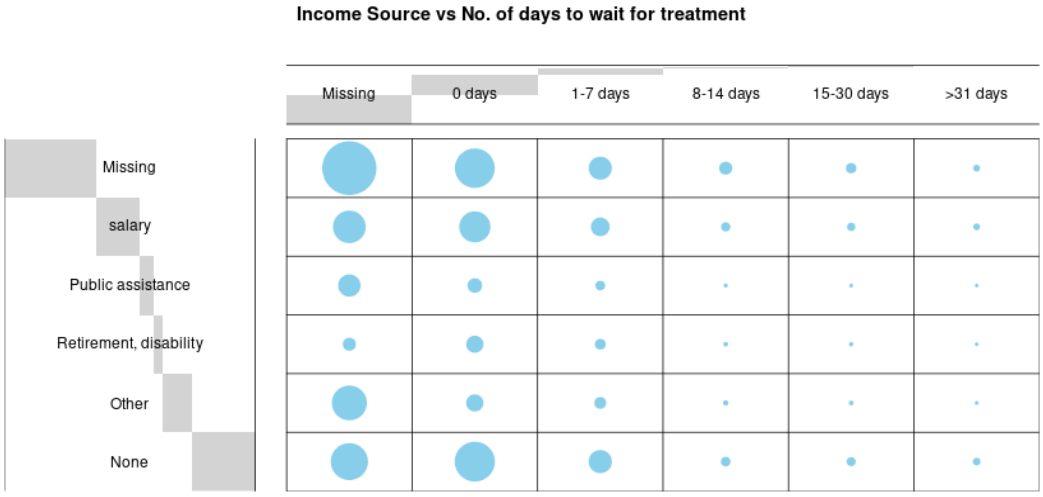

Figure 2: Ballon Plot of Income vs No. of days to wait for the treatment

From Figure 2, we can see that:

- The clients who have no income source get access to the treatment relatively early.
- We have lots of Missing/unknown/not collected/invalid values for the Income source as well as days to wait for the treatment.
- The retired and disabled client gets relatively early access to the treatment.
- There's no such group of an income source that gets relatively later access to the treatment.

Is there a relationship between the client's different substance use categories and waiting time for the treatment?

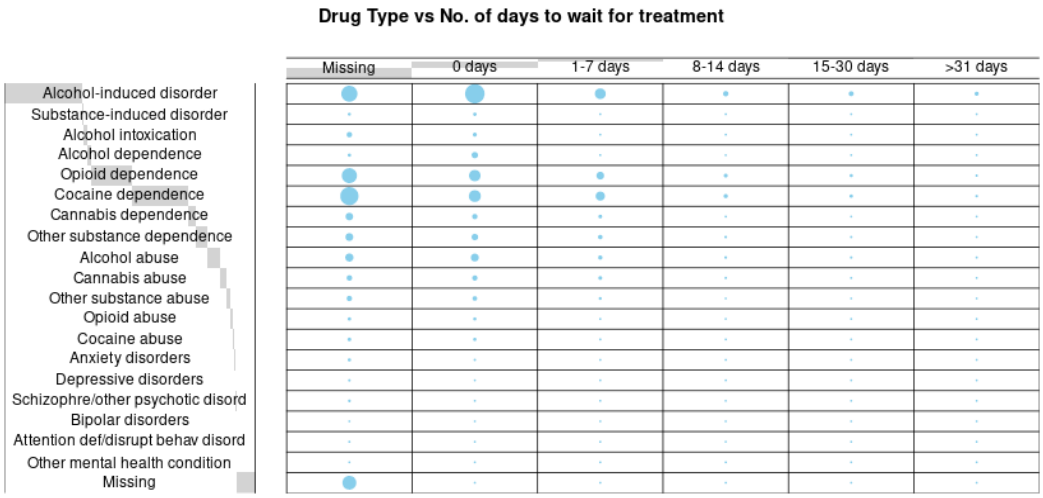

Figure 3: Ballon Plot of Income vs No. of days to wait for the treatment

From the figure 3, we can see that:

- The majority of the clients have Alcohol-induced disorder, Opioid dependence, and Cocaine dependence.
- Clients with Alcohol-induced disorders get access to the treatment relatively early.
- We have a relatively lot of missing waiting days for clients with cocaine dependence.
- From the figure, there's no conclusive evidence of relatively late access to the treatment.
- We have relatively more missing values on clients with depressive disorders and Schizophrenia/other psychotic disorders.

Exploratory Analysis for Goal-2

Is there a relationship between the client's Ethnicity and the reason for discharge?

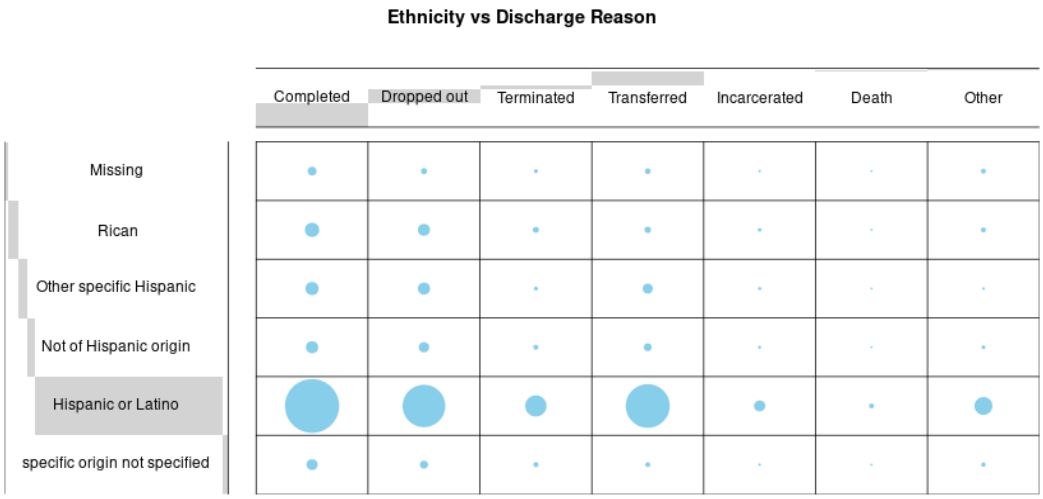

Figure 4: Ballon Plot of Ethnicity vs Reason for Discharge

From the figure 4, we can see that:

- For those clients, we have Missing/unknown/not collected/invalid values on the ethnicity, and there are relatively more treatment completions.
- Across all the ethnicity groups we see relatively more transfers than dropouts.
- There's no conclusive relationship between, different ethnic groups and discharge reasons.

Is there a relationship between the client's primary Income Source and the reason for discharge?

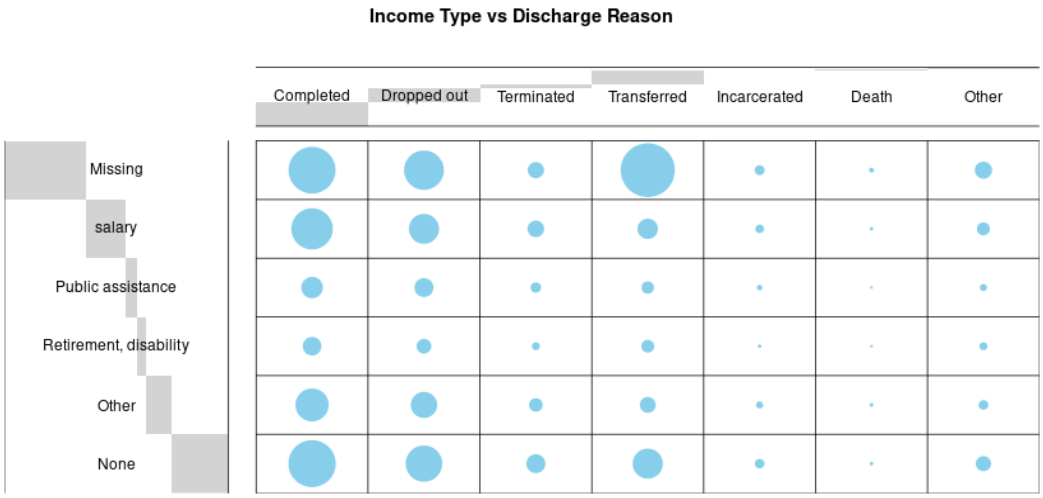

Figure 5: Ballon Plot of Income Type vs Reason for Discharge

From the figure 5, we can see that:

- For those whose Primary Income Source is Missing/unknown/not collected/invalid values, we have a relatively lesser number of terminations.
- For those whose Primary Income Source is Missing/unknown/not collected/invalid values, we have a relatively more number of transfers.
- There is a relatively more number of dropped outs for those clients whose primary income source is a salary.

Is there a relationship between the client's different substance use categories and the reason for discharge?

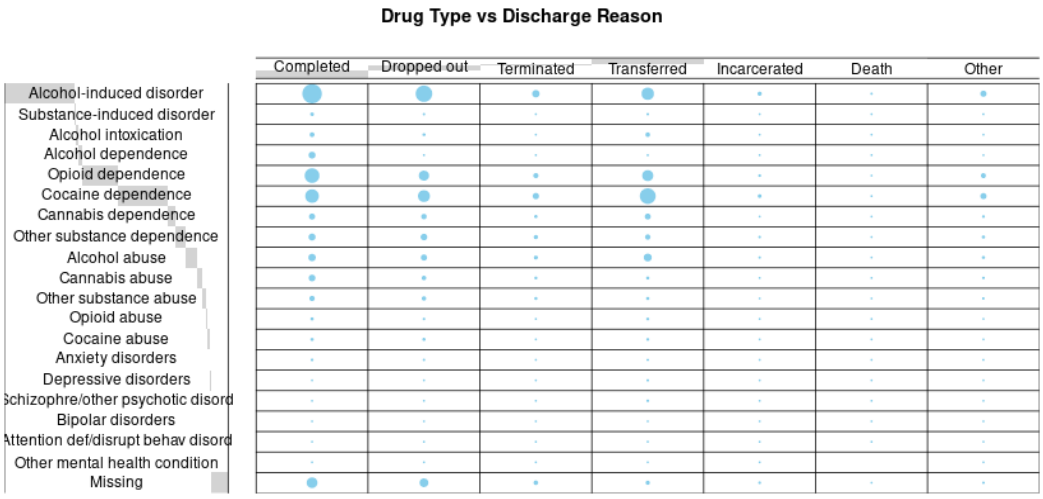

Figure 6: Ballon Plot of Substance use category vs Reason for Discharge

From the figure 6, we can see that:

- The clients with alcohol-induced disorder and those with other substance dependence have relatively more dropouts.
- Those clients with cocaine dependence have a relatively more number of transfers.
- The majority of the clients have alcohol-induced disorders, opioid dependence, and cocaine dependence.

Some More Exploratory Analysis!

- The most common reason for discharge from substance abuse treatment is "Treatment Completed," with "Dropped out of treatment" and "Transferred to another treatment program or facility" following closely behind.

- The age group with the highest reported age of initial intoxication is 15-17. The number of individuals reporting initial intoxication under age 11 is relatively low, but it may influence the other age categories.

- The majority of individuals discharged from substance abuse treatment are "Unemployed-layoff/looking for a job," followed by those who are "Not in the labor force."

- The most common marital status among individuals in substance abuse treatment is "Unmarried/Never Married."

- The majority of individuals in substance abuse treatment report "None" as their source of income, followed by "Wages/Salary."

- The number of individuals reporting pregnancy during treatment is low, with a significant amount of missing/invalid data.

- The use of medication-assisted opioid therapy is present but relatively low in number.

- Individuals in the "Not in labor force" category include those who are retired, students, or in other categories.

- The majority of individuals in substance abuse treatment report secondary substance use at admission, with "Some" and "Daily" use being the most common.

- Methamphetamine/speed is the most commonly reported secondary substance, followed by heroin and benzodiazepines, in some numbers.

- At the time of admission for substance abuse treatment, the use of other drugs was most commonly reported, followed by alcohol and other drugs.

- Ambulatory, non-intensive outpatient treatment was the most common type of treatment given to clients, followed by detox, 24-hour, free-standing residential, and ambulatory, intensive outpatient.

- The age category with the highest number of individuals at the time of admission for substance abuse treatment is 30-34, followed by 25-29.

- The majority of individuals in substance abuse treatment identify as "Not of Hispanic or Latino origin."

- Ambulatory, non-intensive outpatient treatment was the most common type of treatment at the time of discharge from substance abuse treatment.

- The "Northeast" region is the most common region for client-treatment cases.

- The majority of clients in substance abuse treatment had a length of stay of "1 Day," but there are also a significant number of clients with stays of 61-90+ days.

- The Middle and South Atlantic divisions are the most common for client-treatment cases.

- The majority of individuals in substance abuse treatment identify as "White."

- "No Attendance" is the most common response for the number of times a client has attended self-help groups in the past 30 days.

- The majority of individuals in substance abuse treatment have completed education up to "Grades 9-11."

- The most commonly reported substance use category among clients is "Opioid dependence," followed by "Alcohol dependence," as categorized by the Diagnostic and Statistical Manual of Mental Disorders (DSM) from the American Psychiatric Association.

- The majority of individuals in substance abuse treatment belong to the Metropolitan Statistical Area with Central, outlying county.

Procedure for the Statistical Analysis

1. For both goals, first, we identify which attributes share a strong relationship with our variable(s) of interest i.e. waiting time for the treatment and the discharge reason.
2. We compute Mutual Information (MI) of all the attributes present in the data with our variables of interest.
3. We select the top 19 attributes based on the highest value of MI which indicates a strong relationship.
4. We apply the random forest for selecting the best 6 attributes among the top 20 for further analysis.
5. We do correspondence analysis to look deeper into these selected variables.

What is Mutual Information?

Mutual information is used to measure the similarity or redundancy between different features or variables. Mutual information helps to quantify the amount of shared information or dependency between two or more variables.

Higher mutual information values between two features indicate a stronger relationship or similarity between them.

Mutual Information between two random variables is given by:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Where, p(x, y) represents the joint probability mass function (PMF) of X and Y, p(x) represents the marginal PMF of X, and p(y) represents the marginal PMF of Y.

Feature selection by Random Forest

Random Forest works by constructing an ensemble of decision trees, where each tree is built using a random subset of the features and training data. The algorithm makes predictions by aggregating the results from all the individual decision trees. The random selection of features during tree construction creates a mechanism to assess feature importance.

Feature importance in Random Forests is typically estimated by measuring the average decrease in impurity (e.g., Gini impurity or entropy) when a particular feature is used for splitting nodes in the trees. The rationale is that if a feature is important, its inclusion in the tree splits should lead to a significant reduction in impurity. If a feature is unimportant or redundant, its inclusion should have little impact on impurity reduction.

After training the Random Forest model, feature importance scores can be obtained based on the analysis of the trees. These scores provide a quantitative measure of the relative importance of each feature in the prediction task. Higher importance scores indicate more influential features, while lower scores suggest less relevant ones.

What is correspondence analysis?

Correspondence Analysis (CA) is a statistical method used to explore the relationship between two categorical variables in a contingency table. It is a type of factor analysis that is specifically designed for categorical data. The goal of CA is to identify patterns in the data and to visualize the relationships between the categories of the two variables in a low-dimensional space.

Exploring Goal-1 in detail

To find some solution to goal-1 it is important to see which features from the data *treatments_2017-2020.csv* share a strong association with the number of days to wait for the treatment.

Now, to find these features, we use Mutual Information to obtain the measure of the similarity of different features with the number of days to wait.

We select the top 20 features which are strongly associated with the time (in days) to wait for the treatment.

The variables that are selected using the Mutual information content which shares the relationship with waiting time (Days) for treatment are:

- State FIP codes used by the US census bureau: “STFIPS”,
- US census division of the client treatment case: “DIVISION”,
- The client's health insurance at admission: “HLTHINS”,
- US census region of the client treatment case: “REGION”,
- Primary payment source for the treatment episode at the time of admission: “PRIMPAY”
- Length of stay for the treatment: “LOS”
- The client’s substance use is categorized by the Diagnostic and Statistical Manual of Mental Disorders (DSM) from the American Psychiatric Association nomenclature.: “DSMCRIT”
- Client’s source of income/support: “PRIMINC”
- Clients primary substance use at discharge: “SUB1_D”
- The client’s tertiary substance use at admission: “SUB3”
- The client’s tertiary substance use at discharge: “SUB3_D”
- Number of times a client has attended self-help groups in the past 30 days before the date of their discharge: “FREQ_ATND_SELF_HELP_D”
- The client’s secondary substance use at discharge: “SUB2_D”
- If the client has a co-occurring mental and substance use disorder: “PSYPROB”
- The client’s living arrangements at admission: “LIVARAG”
- Client’s status of whether they use a medication-assisted opioid therapy as part of their treatment: “METHUSE”
- Type of treatment service/setting at admission: “SERVICES”
- Type of treatment service / setting a discharge: “SEVICES_D”
- If in the PSOURCE variable, ‘Court/criminal justice referral/DUI/DWI’ was chosen, then further detail is provided here on the type of criminal justice referral: “DETCRIM”

Here is the variable importance plot for getting an idea about the importance of features that share a strong relationship with the number of days to wait for the treatment.

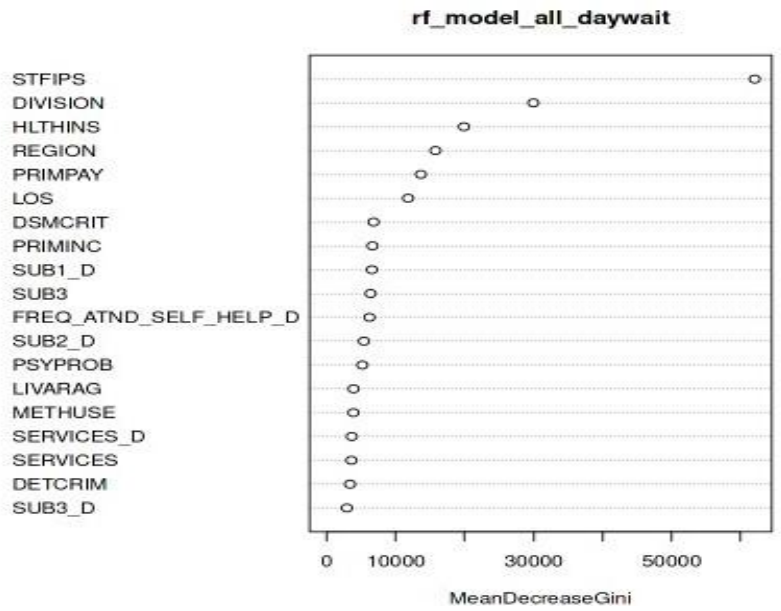

Fig 7: Variable importance plot for the days to wait for the treatment.

From the above variable importance plot we observe that the maximum variation in Waiting time (in days) is explained by the features:

- State FIP codes used by the US census bureau: “STFIPS”,
- US census division of the client treatment case: “DIVISION”,
- The client's health insurance at admission: “HLTHINS”,
- US census region of the client treatment case: “REGION”,
- Primary payment source for the treatment episode at the time of admission: “PRIMPAY”,
- Length of stay for the treatment: “LOS”

Before we jump into the correspondence analysis, we need to explore the reason(s) behind why these selected variables share a strong relationship with the *number of days to wait for the treatment*.

Below are the plausible reasons why the six variables mentioned may share a strong relationship with the *waiting time for substance abuse treatment*:

- **Length of stay for treatment:** Individuals who require longer stays in treatment may have more severe or complex substance abuse issues that require more intensive or specialized treatment services, resulting in longer waiting times for treatment.

- **State FIP codes:** Differences in state-level policies and resources related to substance abuse treatment may impact the availability and accessibility of treatment services, leading to variations in wait times for treatment across states.

- **US census division of the client treatment case:** Differences in regional demographics, resources, and policies may impact the availability and accessibility of substance abuse treatment services, leading to variations in wait times across census divisions.

- **The client's health insurance at admission:** Individuals with certain types of insurance may have better access to treatment services, which may result in shorter wait times for treatment.

- **US census region of the client treatment case:** Differences in regional demographics, resources, and policies may impact the availability and accessibility of substance abuse treatment services, leading to variations in wait times across census regions.

- **Primary payment source for the treatment episode at the time of admission:** Differences in payment sources may impact the availability and accessibility of substance abuse treatment services, leading to variations in wait times for treatment.

Overall, the relationship between these variables and the waiting time for substance abuse treatment is likely complex and multifaceted, influenced by a range of individual, regional, and systemic factors. Understanding these relationships can help to identify areas where improvements can be made to reduce wait times and improve access to effective substance abuse treatment for individuals in need.

For example, policymakers may need to consider increasing funding for substance abuse treatment services in areas with long wait times or developing policies to improve insurance coverage for substance abuse treatment.

Additionally, treatment providers may need to offer more specialized or intensive treatment interventions to meet the needs of individuals with more complex substance abuse issues, which may help to reduce wait times and improve outcomes for this population.

Exploring Goal-2 in detail

The selected variables by mutual information which share a relationship with discharge reasons are:

- State FIP codes used by the US census bureau: “STFIPS”,
- US census division of the client treatment case: “DIVISION”,
- The client’s secondary substance use at discharge: “SUB2_D”
- Length of stay for the treatment: “LOS”
- The client’s substance use is categorized by the Diagnostic and Statistical Manual of Mental Disorders (DSM) from the American Psychiatric Association nomenclature.: “DSMCRIT”
- Clients primary substance use at discharge: “SUB1_D”
- The frequency of substance usage corresponding to the substance in SUB1_D: “FREQ1_D”
- Client primary substance use at admission: “SUB1”
- The client’s living arrangements at discharge: “LIVARAG_D”
- Number of times a client has attended self-help groups in the past 30 days prior to the date of their discharge: “FREQ_ATND_SELF_HELP_D”
- Type of treatment service/setting at admission: “SERVICES”
- Type of treatment service / setting a discharge: “SEVICES_D”
- Client’s source of income/support: “PRIMINC”
- If the client has a co-occurring mental and substance use disorder: “PSYPROB”
- US Census region of the client-treatment case: “REGION”
- Arrests of client made 30 days prior to discharge: “ARRESTS_D”
- Classifies client's substance use type as alcohol only, other drugs only, alcohol and other drugs, or none. This variable looks across primary, secondary, and tertiary substances reported at the time of admission to treatment: “ALCDRUG”
- The person or agency that referred the client for treatment: “PSOURCE”
- The client's employment status at discharge: “EMPLOY_D”

Now, below is the variable importance plot for the variables which share a strong relationship with *reasons for discharge*

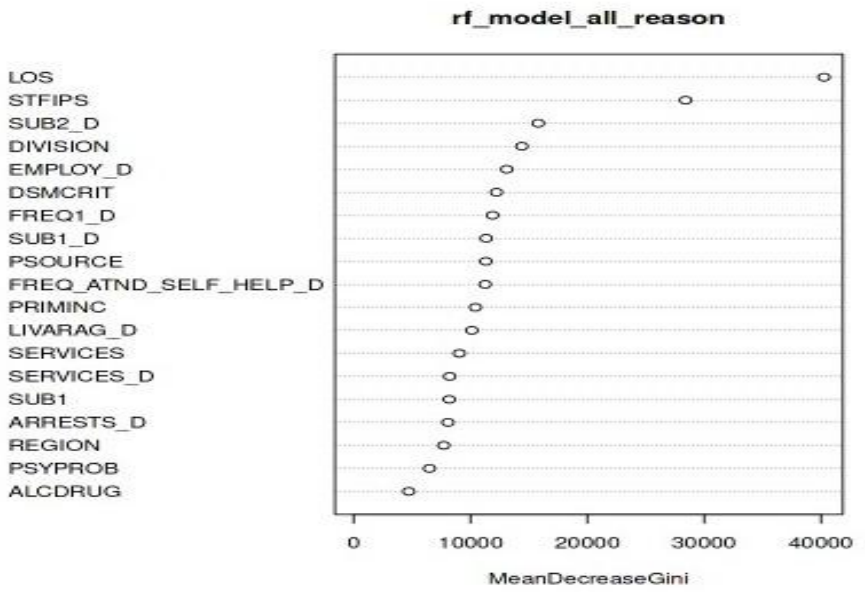

Fig 8: Variable importance plot of reasons for discharge

From the above variable importance plot we observe that the maximum variation is explained by the top 6 variables:

- Length of stay for the treatment: “LOS”,
- State FIP codes used by the US census bureau: “STFIPS”,
- The client’s secondary substance use at discharge: “SUB2_D”,
- US census division of the client treatment case: “DIVISION”,
- The client’s employment status at discharge: “EMPLOY_D”,
- Client substance use is categorized by DSM: “DSMCRIT”.

Below are the plausible reasons why the six variables mentioned may share a strong relationship with the *reason for discharge*:

- **Length of stay for treatment:** Individuals who stay in treatment for longer periods may be more likely to complete treatment, resulting in higher rates of treatment completion and lower rates of dropout, termination, or transfer to another facility.

- **State FIP codes:** Differences in state-level policies and resources related to substance abuse treatment may impact the availability and quality of treatment services, which may impact the likelihood of treatment completion or other reasons for discharge.

- **The client's secondary substance use at discharge:** Individuals with co-occurring substance use disorders may be at higher risk for dropout, termination, or other reasons for discharge, as these conditions may complicate treatment and recovery.

- **US census division of the client treatment case:** Differences in regional demographics, resources, and policies may impact the availability and quality of substance abuse treatment services, which may impact the likelihood of treatment completion or other reasons for discharge.

- **The client's employment status at discharge:** Individuals who are employed may be at higher risk for dropout, termination, or other reasons for discharge, as they may face work-related obligations that interfere with treatment attendance or completion.

- **Category of the client's substance use:** Individuals with different types of substance use disorders may require different types or intensities of treatment, which may impact the likelihood of treatment completion or other reasons for discharge.

Overall, the relationship between these variables and the reasons for discharge is likely complex and multifaceted, influenced by a range of individual, regional, and systemic factors.

For example, the longer length of stay in treatment may be associated with better treatment outcomes and lower rates of dropout or termination, while co-occurring substance use disorders may increase the risk of dropout or termination due to the complexity of treatment.

Regional differences in policies and resources may also impact the availability and quality of treatment services, which may impact treatment outcomes and reasons for discharge.

Additionally, employment status may impact treatment attendance and completion, particularly for individuals who face work-related obligations, while different types of substance use disorders may require different types of treatment interventions.

Understanding the relationships between these variables and reasons for discharge can help inform strategies to improve treatment outcomes and reduce the risk of dropout or termination.

For example, treatment providers may need to offer more specialized or intensive treatment interventions to individuals with co-occurring substance use disorders or develop strategies to better engage and retain individuals who are employed.

Policymakers may also need to consider increasing funding for substance abuse treatment services in areas with high rates of dropout or termination or developing policies to improve the quality and accessibility of treatment services.

Ultimately, the goal is to improve the effectiveness and accessibility of substance abuse treatment, reduce the risk of negative outcomes, and improve overall health and well-being for individuals struggling with addiction.

Correspondence Analysis for Goal-1

Now we look deeper into those variables which are obtained after variable selection from the random forest.

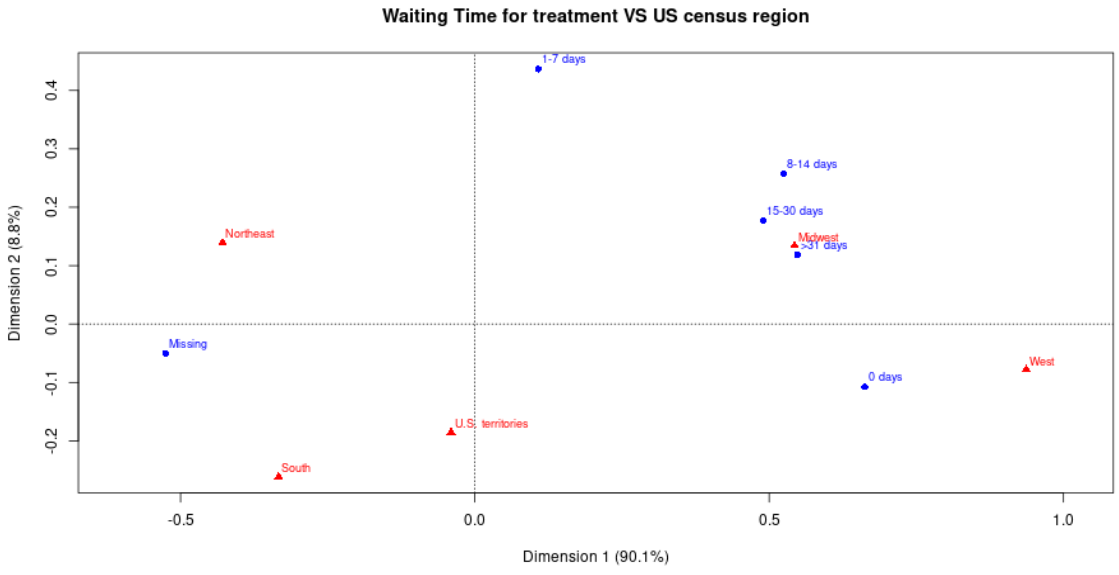

Figure 9: Correspondence Analysis Plot for Treatment Waiting Time vs US Census Region

From Figure 9, we observe that:

- The clients from the West region usually get treatment relatively early compared to any other region.
- Many missing values for the days are present around the South and the US territories.

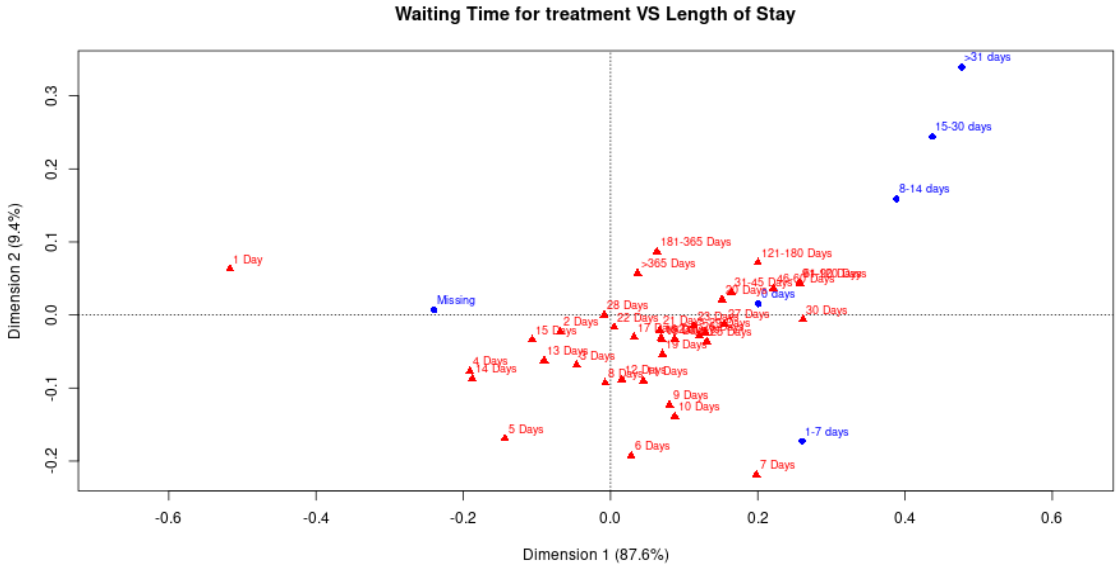

Figure 10: Correspondence Analysis Plot for Treatment Waiting Time vs Length of Stay for the Treatment

From Figure 10, we observe that:

- For a shorter stay, the days to wait are usually between 1-7

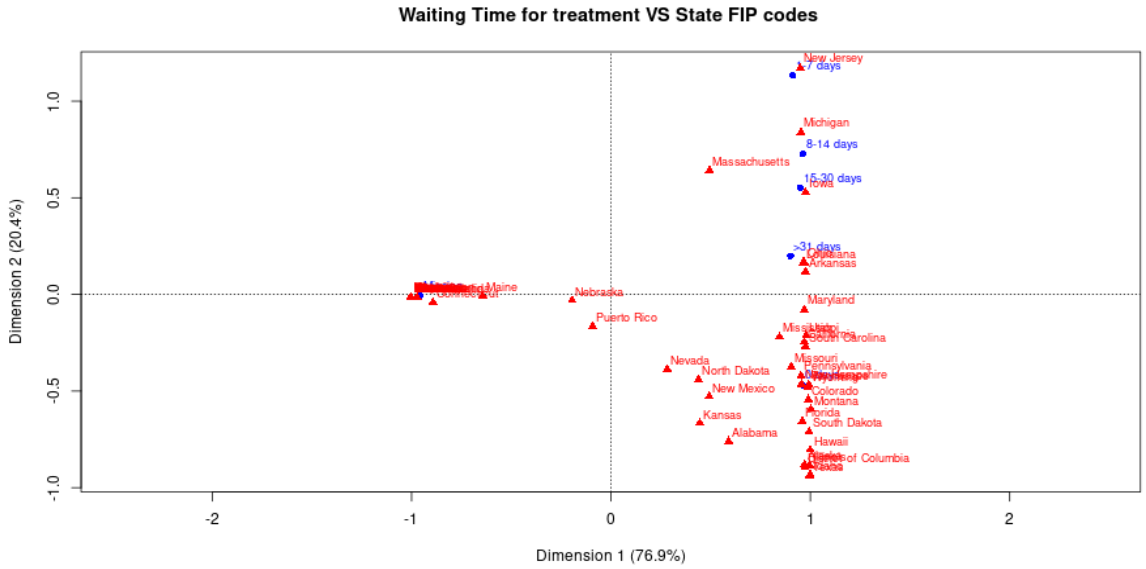

Figure 11: Correspondence Analysis Plot for Treatment Waiting Time vs Sate FIP codes

From Figure 11, we observe that:

- The majority of the US states get access to treatment early.
- We also have a lot of missing values for a small minority of states.

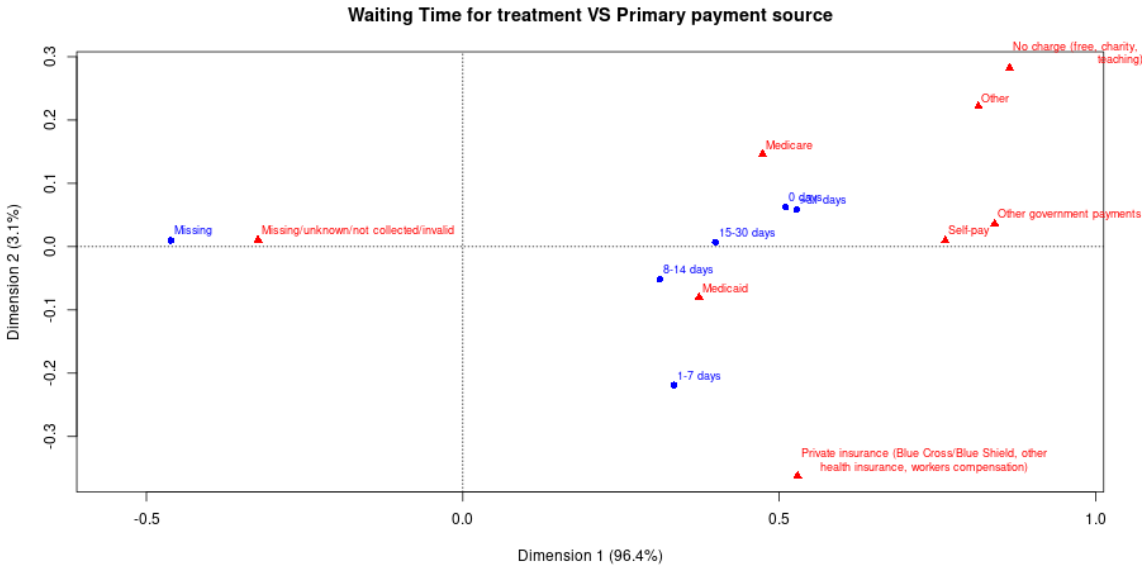

Figure 12: Correspondence Analysis Plot for Treatment Waiting Time vs Primary Payment Source for the Treatment

From Figure 12, we observe that:

- Those clients who have Medicaid usually get access to the treatment between 1-14 days.
- Those clients with Medicare usually get access to the treatment in 0 days.

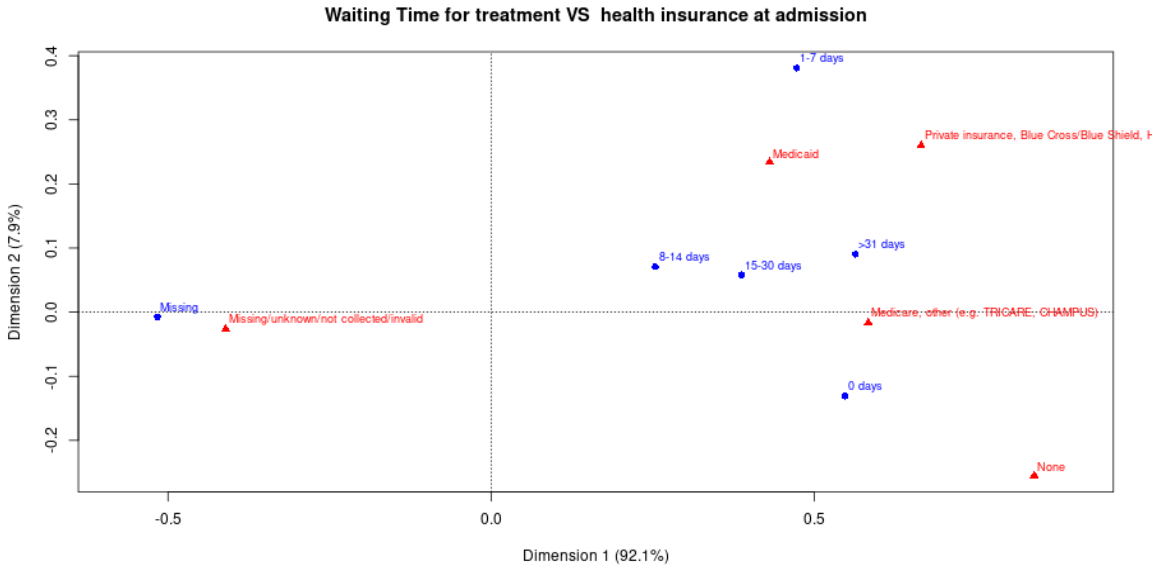

Figure 13: Correspondence Analysis Plot for Treatment Waiting Time VS Health insurance at admission for the Treatment

From Figure 13, we observe that:

- Similar to the above interpretations, Those with Medicare health insurance get access to the treatment very early.
- This also indicates the relationship between health insurance and the primary payment source for the treatment.

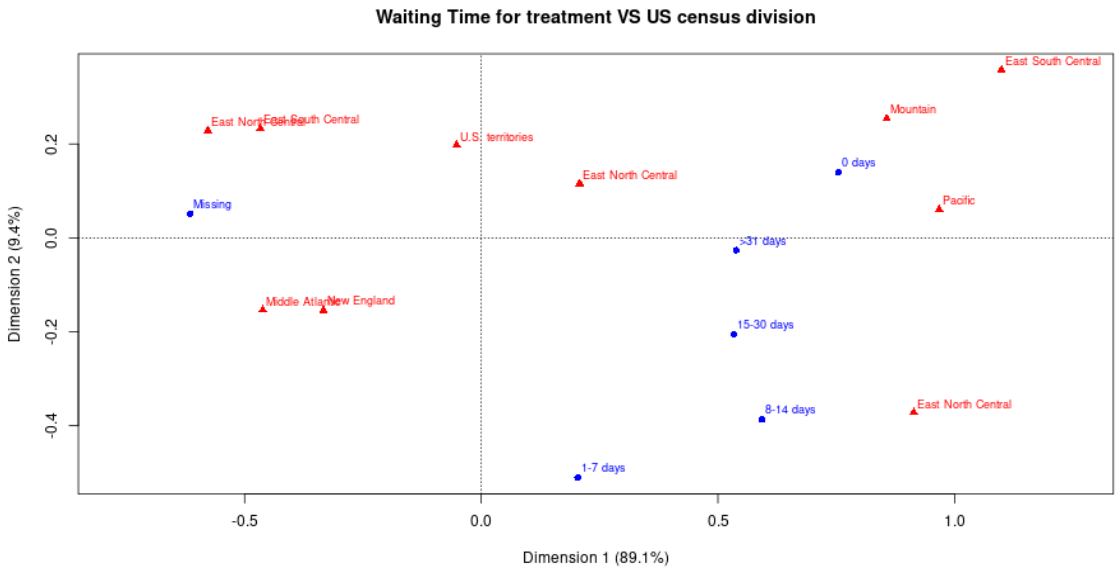

Figure 14: Correspondence Analysis Plot for Treatment Waiting Time vs US census division

From Figure 14, we observe that:

- For the East-North, East-South, and Central states, we have relatively more values missing.
- The Middle Atlantic and New England are very similar.
- Mountain and Pacific's division gets relatively early access to the treatment.

Correspondence Analysis for Goal-2

In the below section, we go deeper into those attributes which are strongly related to the discharge reason from the treatment using correspondence analysis.

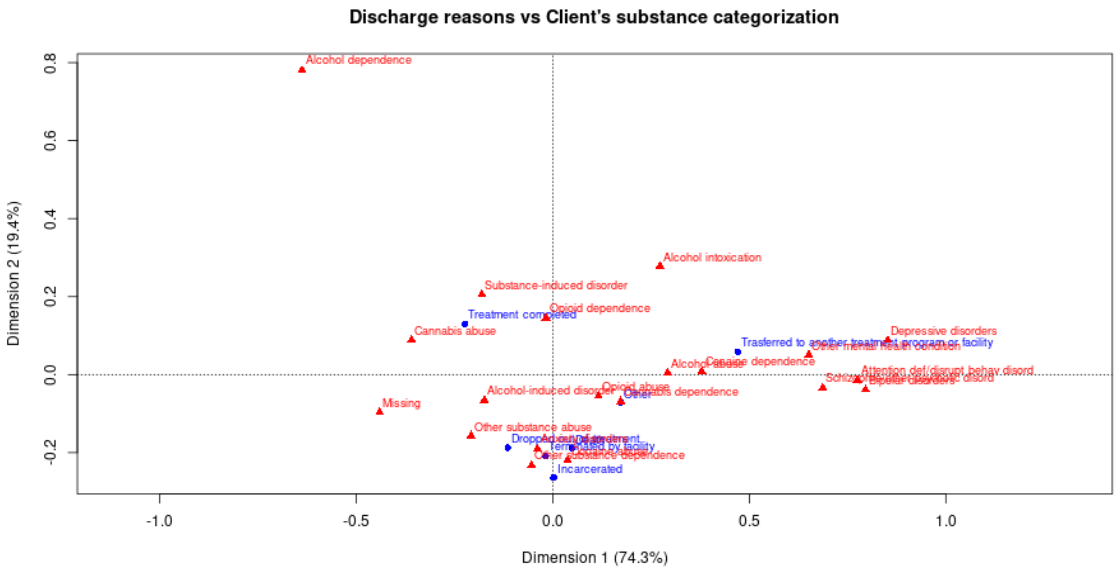

Figure 15: Correspondence Analysis Plot for Discharge Reasons vs Clients substance categorization

Figure 15, we observe that:

- For those with Opioid dependence, substance-related disorders, and Cannabis abuse, relatively more clients completed the treatment.
- Those with Depressive Disorders, Other mental conditions are transferred to another treatment program more relatively than any other substance abuse category group.

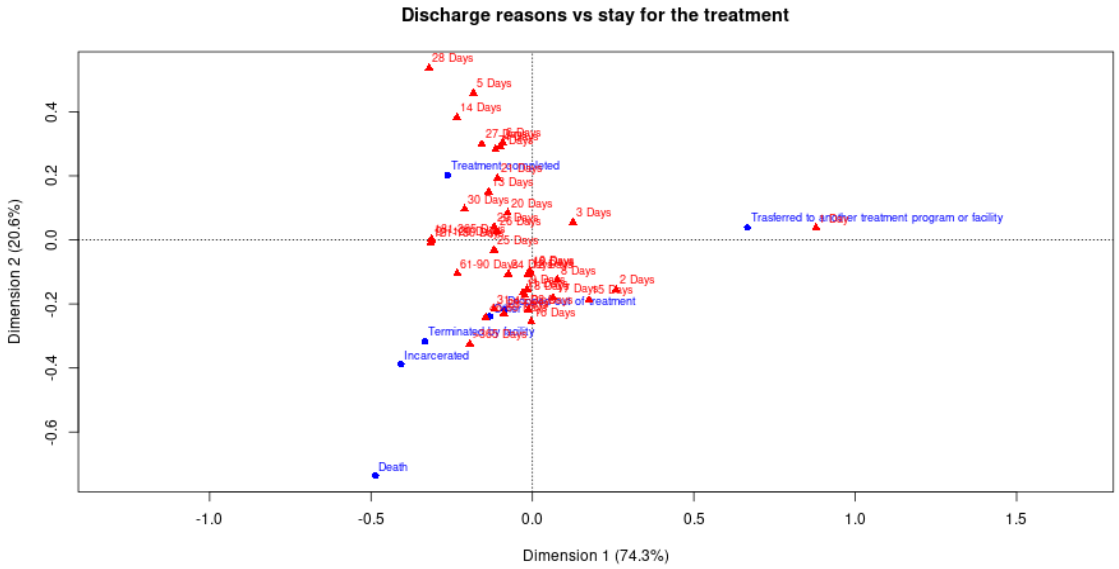

Figure 16: Correspondence Analysis Plot for Discharge Reasons vs Number of days stay for the treatment

From Figure 16, we observe that:

- Those with a single-day stay are transferred to another treatment facility more relatively.
- The drop-out of the treatment, Incarcerated and, termination by the facility are closely related.
- The usual treatment completion is around one month.

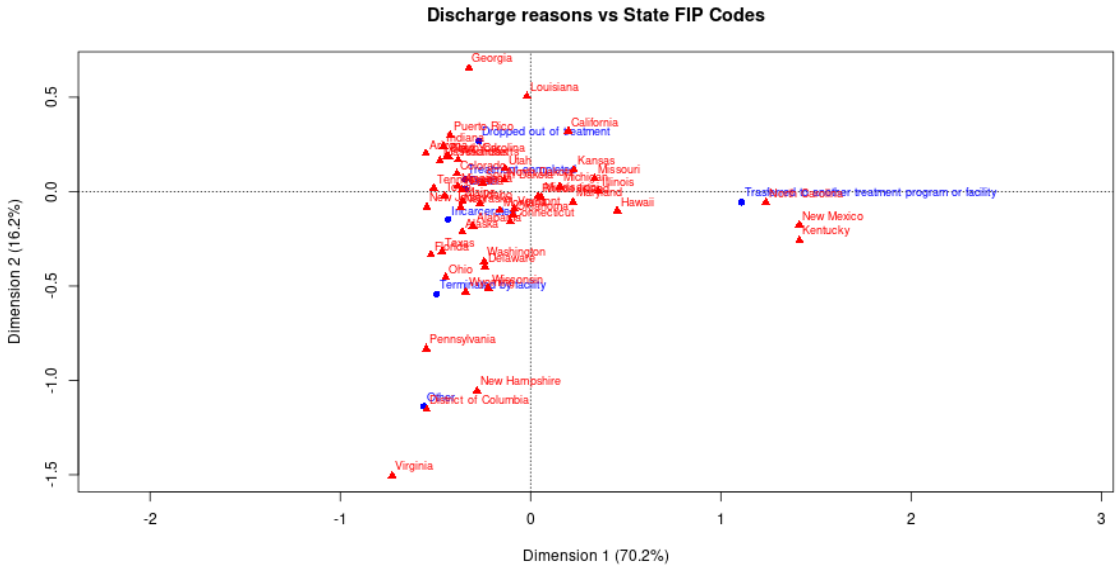

Figure 17: Correspondence Analysis Plot for Discharge Reasons vs State FIP codes

From Figure 17, we observe that:

- Clients in states such as Washington, Ohio, and Delaware are relatively more prone to termination of the treatment by the facility.
- Clients in states such as Puerto Rico, and Indiana are more likely to drop out of the treatment.
- Tennessee, Colorado, and Utah are among the few states where treatment is more likely to get completed.

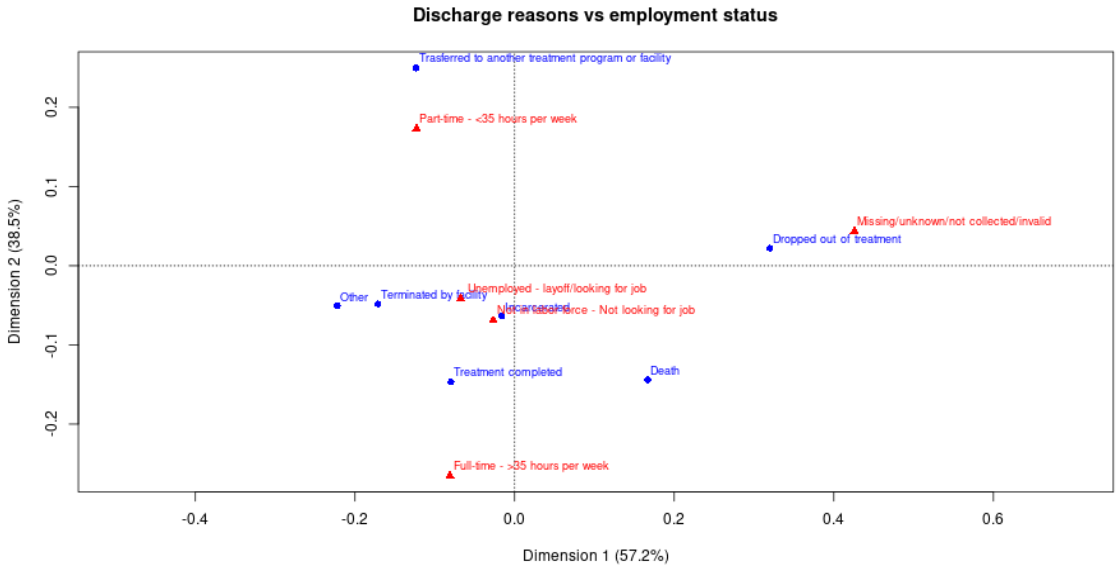

Figure 18: Correspondence Analysis Plot for Discharge Reasons vs Employment status at discharge

From Figure 18, we observe that:

- Those clients who are unemployed are more likely to get terminated by the facility.
- Part-time employed clients are more likely to get transferred to another treatment program or facility.
- Those clients who Dropped out of treatment are closely related to the Missing/Unknown/invalid observations on Employment status.

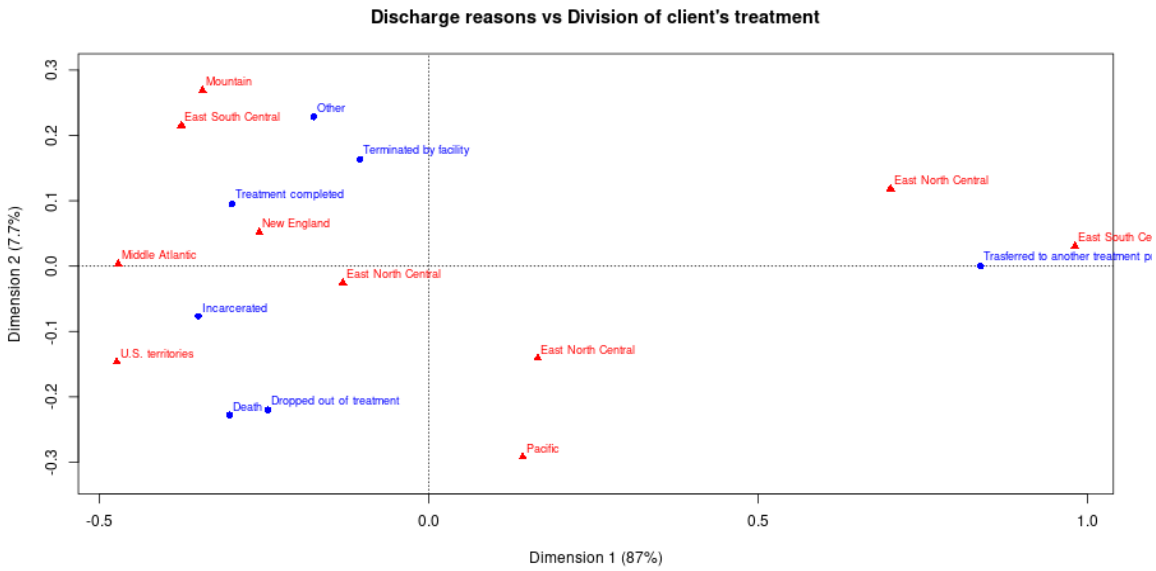

Figure 19: Correspondence Analysis Plot for Discharge Reasons vs US divisions

From Figure 19, we observe that:

- Clients with the East South Central division are relatively transferred more to the different treatment facilities.
- New England division clients have completed treatment relatively more in number as compared to the clients from any other division.

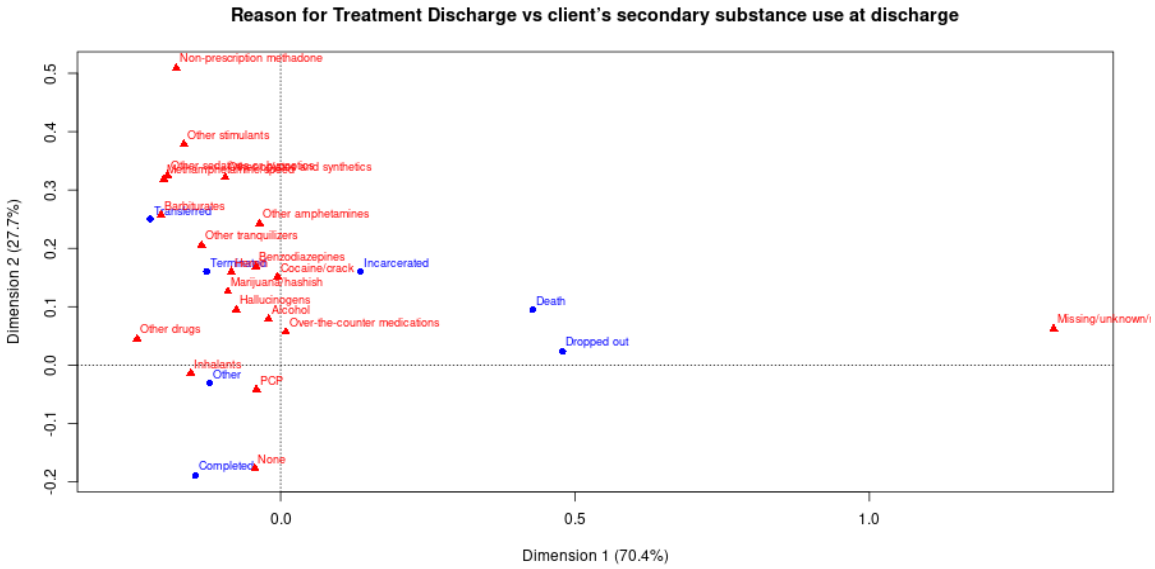

Figure 20: Correspondence Analysis Plot for Discharge Reasons vs Client's secondary substance use at discharge

From Figure 20, we observe that:

- Clients with secondary substance abuse of Benzodiazepines, Marijuana, and other tranquilizers are more likely to get terminated.
- Clients with secondary substance abuse of Barbiturates are more likely to get transferred.

Final Conclusion

- From the ballon plots, we suspected no clear relationship between Ethnic Groups and Waiting time for the treatment which also got reflected using a variable selection procedure.
- From the ballon plots, we suspected no clear relationship between Ethnic Groups and the reasons for discharge from the treatment which also got reflected using a variable selection procedure.
- While doing the variable selection, we found three attributes that share a strong relationship with waiting time for treatment and discharge reasons. And these are:
    - Length of stay for the treatment
    - State FIP codes used by the US census bureau
    - US census division of the client treatment case

  This indicates the location and the number of days for the treatment plays a huge role in the waiting time for the treatment as well as the discharge reason.
- Those clients with Medicare insurance get the treatment relatively early.
- Those clients with other substance abuse, other substance dependence, and cocaine abuse are more prone to drop out or get terminated by the facility.
- Those clients who are unemployed are more prone to get terminated by the facility.