Exercise_10.3_Final_Project_Step_2

Maxim Bilenkin

2025-02-12

Loading Necessary Libraries

```
suppressPackageStartupMessages({
  library(dplyr)
  library(caret)
  library(randomForest)
  library(ggplot2)
})
```

Topic: Excessive consumption of alcohol causes health issues

It is no secret that throughout our lives we go through hard times and face various hardships. Not all, but many people turn to alcohol instead of sport to ease their emotions. Excessive alcohol consumption leads to health issues. Many commercials and advertisements we see today on television and internet that advertise various products and services. But there is nothing or at least very few that would advertise health issues related to excessive consumption of alcohol. Many people quit drinking only after falling ill, unaware that their health issues stem from excessive alcohol consumption.

Any health-conscious person wants to live a happy and healthy life. With the advancement of technology and the availability of big data people can be informed well in advance about issues that could cause health issues. Necessary early actions can be taken at early stages to prevent health issues. It's a data science problem because with the help of data science we can predict, understand and mitigate the impact of excessive alcohol consumption on individuals and our society. Data science enables better health monitoring. With more healthy people our society improves in all aspects of life. Economy innovates, saves money on health-related issues and prospers.

The following questions need to be asked:

1) How do social media and advertisements correlate with excessive alcohol consumption? Do more advertisements increase alcohol consumption?
2) How does excessive alcohol consumption effects health in the short and long run?
3) What are the most effective ways to reduce excessive alcohol consumption?
4) Do genetic factors influence excessive alcohol consumption?
5) How do government policies impact alcohol consumption? Do high taxes on alcohol decrease consumption?
6) What is the cost of excessive alcohol consumption to the economy?
7) How to develop personalized treatment plans to prevent excessive alcohol consumption?
8) What are the key predictors of excessive alcohol consumption?
9) What are the environmental factors that are influencing or contribute to excessive drinking behaviors?
10) Are wearable devices including mobile apps effective in monitoring and decreasing alcoholic consumption?

To approach the issue, I would like to read as much as possible different articles and information that are available on the internet. I would gather all available statistical data that were collected on excessive alcohol consumption. I would analyze it and make my own assessment based on the trends and patterns I see.

Obviously, my approach only addresses the issue partially. It only gives good insight into the problem. It identifies the issue and causes which can give guidance on what preventive measures can be taken to reduce alcoholic excessive consumption. The excessive alcoholic consumption is a complex problem that involves different aspects of our lives. To fix the issue different factors should be involved such as public health initiatives, customizing personal treatment, preventive policy intervention and social support from family and friends.

For my project I will use the following three data sets to perform my analysis.

1) World Health Organization (WHO) alcoholic consumption globally data set spans from 1960 – 2020 with 48,924 records. This data set will give me insight into daily alcohol consumption for each year per capita measured in grams. (source link: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/levels-of-consumption) (zip file link: https://ghobulkdownloads.blob.core.windows.net/ghocontainer/levels-of-consumption.zip)
2) Kaggle provides a data set for Alcohol and Life Expectancy. This international study will give me insight into whether excessive alcohol consumption impacts human lives. (source link: https://www.kaggle.com/datasets/thedevastator/relationship-between-alcohol-consumption-and-lif/data)
3) This survey was conducted on alcohol consumption and happiness. I will use this data set by Kaggle to see if there is a positive or negative correlation between the two. (Source link: https://www.kaggle.com/datasets/marcospessotto/happiness-and-alcohol-consumption)

To perform my analysis, I will utilize the following R programming packages.

SNPassos for data manipulation and data exploratory analysis.

genetics for creating and handling genetic data.

GA for identifying and using genetic markers with alcohol consumption.

dplyr for data transformation and manipulation including cleaning and preparing data.

ggplot2 to visualize my data that will help me to understand patterns.

caret to predict alcoholic consumption.

tidyverse this is a collection of many packages that I will use to streamline my workflow.

All the packages described above should address all my needs.

For my project I will use the following plots to visualize correlation, compare distributions and assess test statistics. For example, Manhattan Plot, Q-Q Plot, Heatmap, Boxplot and ROC Curve.

It would be good to ask the following questions for future steps.

What are the most efficient and effective ways to prevent excessive alcohol consumption?

Will my research only be effective for the USA population or can it be applied globally?

What are the most culprit factors that cause excessive alcohol consumption?

Are there any ways to interact with genes to prevent alcohol abuse?

Importing and Cleaning Data

Original Datasets

```r
# Importing datasets
happiness_alcohol_consum_data <- read.csv(
    "C:/Users/maxim/OneDrive/Desktop/BU/DSC 520/HappinessAlcoholConsumption_Kaggle.csv",
    stringsAsFactors = FALSE)

global_alcohol_consum_data <- read.csv(
    "C:/Users/maxim/OneDrive/Desktop/BU/DSC 520/Global_Alcohol_Consumption_by_WHO.csv",
    stringsAsFactors = FALSE)

life_expectancy_data <- read.csv(
    "C:/Users/maxim/OneDrive/Desktop/BU/DSC 520/lifeexpectancy-verbose_Kaggle.csv",
    stringsAsFactors = FALSE)
```

Data Cleaning

```r
# Removing missing values and converting data types
clean_data <- function(data) {
  data <- na.omit(data)
  data[] <- lapply(data, function(x) if(is.character(x)) as.factor(x) else x)
  return(data)
}

happiness_alcohol_data <- clean_data(happiness_alcohol_consum_data[, c(
    "Country", "Region", "HappinessScore", "Beer_PerCapita", "Spirit_PerCapita",
    "Wine_PerCapita")])

global_consum_data <- clean_data(global_alcohol_consum_data[, c(
    "Id", "IndicatorCode", "SpatialDimension", "NumericValue", "Date")])

life_expectancy_data <- clean_data(life_expectancy_data[, c(
    "CountryDisplay", "YearDisplay", "GhoDisplay", "Numeric")])

# Convert to appropriate data types
global_consum_data$NumericValue <- as.numeric(global_consum_data$NumericValue)
global_consum_data$Date <- as.Date(global_consum_data$Date,
                                   format = "%Y-%m-%dT%H:%M:%OSZ")

happiness_alcohol_data$Region <- as.factor(happiness_alcohol_data$Region)

life_expectancy_data$Numeric <- as.numeric(life_expectancy_data$Numeric)

# Filter IndicatorCode with more than one observation
global_consum_data <- global_consum_data %>%
  group_by(IndicatorCode) %>%
  filter(n() > 1) %>%
  ungroup()
```

First, I decided visually to observe and delete any data that won't give me any insight. For example, in Global_Alcohol_Consuption_by_WHO dataset the following column names 'DisaggregatingDimension2', DisaggregatingDimension2ValueCode', 'DisaggregatingDimension3', 'DisaggregatingDimension3ValueCode' and 'Comments' don't contain any data. I didn't find any use for them. So, I deleted the five columns. Whatever can be deleted manually should be deleted because it will make it more efficient and easier for R to process.

In the second dataset named 'HappinessAlcoholConsumption_Kaggle', I noticed spelling mistakes. In column

name 'Hemisphere', the word 'north' spelled as 'noth'. It's a mistake so I changed it too 'north'.

In the third dataset named 'lifeexpectancy-verbose_Kaggle', I filled out all missing data for income group classification code, countries income group display and world bank income code display. Corrected region for South Sudan as African, instead of Eastern Mediterranean region. Additionally, after doing some research online, classified South Sudan as low-income country instead of middle-income category.

After manually cleaning partially the data, I decided to import all three datasets and use code to check and clean the remaining data.

Displaying final cleaned datasets.

```
glimpse(global_consum_data)
```

```
## Rows: 48,924
## Columns: 5
## $ Id               <int> 138, 465, 672, 681, 974, 1146, 1202, 1215, 1376, 1798~
## $ IndicatorCode    <fct> SA_0000001400, SA_0000001400, SA_0000001400, SA_00000~
## $ SpatialDimension <fct> COUNTRY, COUNTRY, COUNTRY, COUNTRY, COUNTRY, COUNTRY,~
## $ NumericValue     <dbl> 0.00447, 0.01357, 0.00000, 0.21000, 5.26000, 0.00000,~
## $ Date             <date> 2024-08-12, 2024-08-12, 2018-05-11, 2018-05-11, 2018~
```

```
glimpse(happiness_alcohol_data)
```

```
## Rows: 122
## Columns: 6
## $ Country          <fct> Denmark, Switzerland, Iceland, Norway, Finland, Canad~
## $ Region           <fct> Western Europe, Western Europe, Western Europe, Weste~
## $ HappinessScore   <dbl> 7.526, 7.509, 7.501, 7.498, 7.413, 7.404, 7.339, 7.33~
## $ Beer_PerCapita   <int> 224, 185, 233, 169, 263, 240, 251, 203, 261, 152, 63,~
## $ Spirit_PerCapita <int> 81, 100, 61, 71, 133, 122, 88, 79, 72, 60, 69, 75, 15~
## $ Wine_PerCapita   <int> 278, 280, 78, 129, 97, 100, 190, 175, 212, 186, 9, 19~
```

```
glimpse(life_expectancy_data)
```

```
## Rows: 6,408
## Columns: 4
## $ CountryDisplay <fct> Nicaragua, Ireland, Yemen, Nigeria, Thailand, Chile, Zi~
## $ YearDisplay    <int> 1990, 2012, 2000, 2000, 1990, 2012, 2013, 2013, 2013, 1~
## $ GhoDisplay     <fct> Life expectancy at birth (years), Healthy life expectan~
## $ Numeric        <dbl> 68.00000, 69.30000, 61.00000, 48.00000, 19.00000, 80.00~
```

```
print(global_consum_data[1:5, ])
```

```
## # A tibble: 5 x 5
##      Id IndicatorCode SpatialDimension NumericValue Date
##   <int> <fct>         <fct>                  <dbl> <date>
## 1   138 SA_0000001400 COUNTRY              0.00447 2024-08-12
## 2   465 SA_0000001400 COUNTRY              0.0136  2024-08-12
## 3   672 SA_0000001400 COUNTRY              0       2018-05-11
## 4   681 SA_0000001400 COUNTRY              0.21    2018-05-11
## 5   974 SA_0000001400 COUNTRY              5.26    2018-05-11
```

```
print(happiness_alcohol_data[1:5, ])
```

```
##       Country         Region HappinessScore Beer_PerCapita Spirit_PerCapita
## 1     Denmark Western Europe          7.526            224               81
## 2 Switzerland Western Europe          7.509            185              100
## 3     Iceland Western Europe          7.501            233               61
```

```
## 4      Norway Western Europe          7.498            169              71
## 5     Finland Western Europe          7.413            263             133
##   Wine_PerCapita
## 1            278
## 2            280
## 3             78
## 4            129
## 5             97
```

```
print(life_expectancy_data[1:5, ])
```

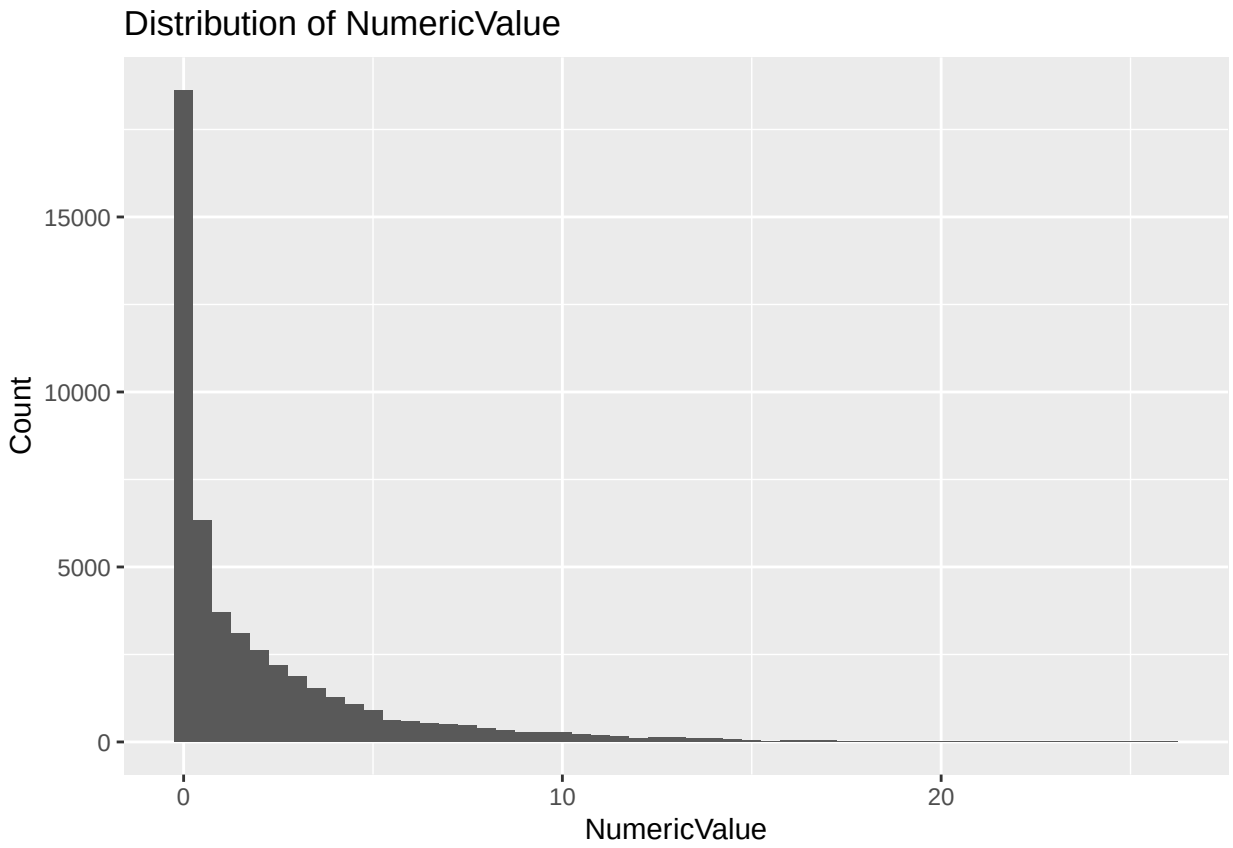
```
##   CountryDisplay YearDisplay                                    GhoDisplay
## 1      Nicaragua        1990                 Life expectancy at birth (years)
## 2        Ireland        2012 Healthy life expectancy (HALE) at birth (years)
## 3          Yemen        2000                 Life expectancy at birth (years)
## 4        Nigeria        2000                 Life expectancy at birth (years)
## 5       Thailand        1990                Life expectancy at age 60 (years)
##   Numeric
## 1    68.0
## 2    69.3
## 3    61.0
## 4    48.0
## 5    19.0
```

I need to learn more about advanced data manipulation, visualization techniques, and machine learning methods. At this point it's not easy identify the insights in the data that are immediately self-evident. However, using Exploratory Data Analysis (EDA) will help me to understand the patterns of the data. With the use of methods like summary() function to get summary statistics overview of each variable and using 'ggplot2' library package for data visualization I will identify patterns, outliers and relationships between variables.

I am planning to leverage different models such as regression models to explore linear relationships between different variables. Classification models like logistic regression could be used to classify data into categories to predict outcomes. Clustering techniques like K-means could be used to identify trends and then group all the data into similar categories.

Exploratory Data Analysis (EDA)

```
# Distribution of NumericValue
ggplot(global_consum_data, aes(x = NumericValue)) +
  geom_histogram(binwidth = 0.5) +
  labs(title = "Distribution of NumericValue", x = "NumericValue", y = "Count")
```



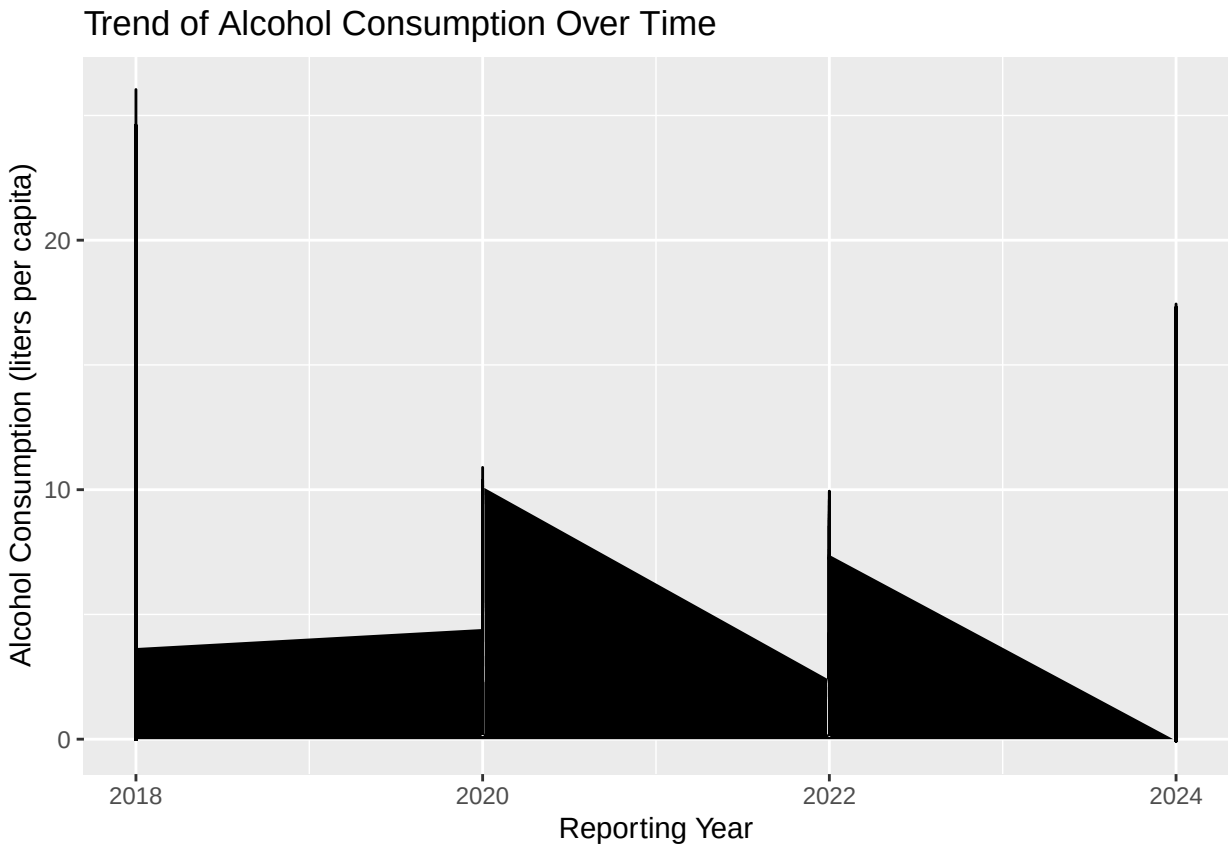

```
# Summary statistics overview
summary(global_consum_data)
```

```
##        Id                IndicatorCode   SpatialDimension  NumericValue
##  Min.   :    138   SA_0000001400:48924   COUNTRY:48924    Min.   :-0.09294
##  1st Qu.:2360393                                         1st Qu.: 0.04894
##  Median :4744452                                         Median : 0.69000
##  Mean   :4722107                                         Mean   : 2.01236
##  3rd Qu.:7073093                                         3rd Qu.: 2.78000
##  Max.   :9442633                                         Max.   :26.04000
##       Date
##  Min.   :2018-05-11
##  1st Qu.:2018-05-11
##  Median :2018-05-11
##  Mean   :2020-11-16
##  3rd Qu.:2024-08-12
##  Max.   :2024-08-12
```

```
# Group-by analysis (Mean Alcohol Consumption by Indicator)
summary_stats <- global_consum_data %>%
    group_by(IndicatorCode) %>%
    summarize(mean_value = mean(NumericValue, na.rm = TRUE))

# Time series analysis with improved labels
ggplot(global_consum_data, aes(x = as.numeric(substr(Date, 1, 4)), y = NumericValue)) +
  geom_line() +
  labs(title = "Trend of Alcohol Consumption Over Time",
       x = "Reporting Year",
```

```
       y = "Alcohol Consumption (liters per capita)")
```




```
# Correlation analysis
cor(global_consum_data$NumericValue, as.numeric(substr(global_consum_data$Date, 1, 4)))
```

```
## [1] -0.02772225
```

For my research project I would like to leverage machine learning techniques because it can provide good insights into the data and better predictive performance to answer various questions. With machine learning I can select relevant features for my model. Also, I can create new feature variables that might enhance the performance of predictive models.

Feature Engineering and Filtering Data

```
# Filtering data
filtered_data <- global_consum_data %>% filter(as.numeric(substr(Date, 1, 4)) >= 2010)

# Creating new variables
global_consum_data$log_value <- log(global_consum_data$NumericValue + 1)

# Summarized tables
summary_table <- global_consum_data %>% group_by(IndicatorCode) %>%
    summarize(mean_value = mean(NumericValue, na.rm = TRUE))
```

At this point I don't know how to do advanced data manipulation. I need to learn advanced functions in the 'dplyr' package for data transformation. Another part that I don't know at this point is the machine learning techniques. I would like to explore all the machine learning packages in 'caret', 'randomForest', and 'xgboost' to implement predictive models. Additionally, I need to gain proficiency in using 'ggplot2' package for creating advanced visualization to uncover patterns and insights of the data. Also, I need to learn how

and which additional variables can be created to better capture important parts of the data and improve model predictive performance. Finally, I need to learn methods that will help me to analyze time-series data so I can better understand and observe trends and seasonality.

Model Training and Evaluation

```
# Setting console width to prevent text spillover
options(width = 80)

# Train a linear regression model
model <- lm(HappinessScore ~ Region + Beer_PerCapita + Spirit_PerCapita +
                Wine_PerCapita, data = happiness_alcohol_data)

# Printing structured Linear Regression summary
summary_model <- summary(model)
cat("\nLinear Regression Model Summary:\n")
```

Linear Regression Model Summary:

```
cat("Adjusted R-squared:", round(summary_model$adj.r.squared, 3), "\n")
```

Adjusted R-squared: 0.637

```
cat("F-statistic:", round(summary_model$fstatistic[1], 2), "on",
    summary_model$fstatistic[2], "and", summary_model$fstatistic[3],
    "DF, p-value:", summary_model$coefficients[1,4], "\n\n")
```

F-statistic: 20.27 on 11 and 110 DF, p-value: 1.377818e-22

```
cat("Coefficients:\n")
```

Coefficients:

```
printCoefmat(summary_model$coefficients, digits = 3, signif.stars = TRUE)
```

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.930906 0.559630 12.38 <2e-16 ***RegionCentral and Eastern Europe -1.753732 0.530249 -3.31 0.0013*** *RegionEastern Asia -1.533337 0.640268 -2.39 0.0183*
RegionLatin America and Caribbean -1.143904 0.540827 -2.12 0.0367 *
RegionMiddle East and Northern Africa -1.501769 0.584575 -2.57 0.0115 *
RegionNorth America -0.101922 0.703817 -0.14 0.8851
RegionSoutheastern Asia -1.534328 0.622186 -2.47 0.0152 *
RegionSub-Saharan Africa -2.952003 0.557056 -5.30 6e-07 *** **RegionWestern Europe -0.522580 0.514725 -1.02 0.3122**
Beer_PerCapita 0.002573 0.000777 3.31 0.0013 ** Spirit_PerCapita -0.001027 0.001091 -0.94 0.3486
Wine_PerCapita -0.000655 0.001166 -0.56 0.5752
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```
# Train a random forest model
set.seed(123)
rf_model <- randomForest(
    HappinessScore ~ Region + Beer_PerCapita + Spirit_PerCapita +
    Wine_PerCapita, data = happiness_alcohol_data, ntree = 100)

# Capture the random forest model output
rf_output <- capture.output(print(rf_model))

# Manually format and print the captured output
```

```
cat("\nCall:\n")
```

Call:

```
cat("randomForest(formula = HappinessScore ~ Region + Beer_PerCapita +
Spirit_PerCapita + Wine_PerCapita,\n")
```

randomForest(formula = HappinessScore ~ Region + Beer_PerCapita + Spirit_PerCapita + Wine_PerCapita,

```
cat("data = happiness_alcohol_data,\n")
```

data = happiness_alcohol_data,

```
cat("ntree = 100)\n")
```

ntree = 100)

```
cat("Type of random forest: regression\n")
```

Type of random forest: regression

```
cat("Number of trees:", rf_model$ntree, "\n")
```

Number of trees: 100

```
cat("No. of variables tried at each split:", rf_model$mtry, "\n\n")
```

No. of variables tried at each split: 1

```
cat("Mean of squared residuals:", round(rf_model$mse[length(rf_model$mse)], 2), "\n")
```

Mean of squared residuals: 0.63

```
cat("% Var explained:", round(rf_model$rsq[length(rf_model$rsq)] * 100, 2), "\n")
```

% Var explained: 51.97

```
# Making predictions
predictions <- predict(rf_model, newdata = happiness_alcohol_data)
```

Finally, I have additional questions such as.

What additional datasets can I leverage to enhance my analysis?

Which specific machine learning techniques would be best to implement that will improve my analysis?

What additional variables can I create that were not present in the datasets so that I can better capture important insights of the data and improve predictive power of my model?