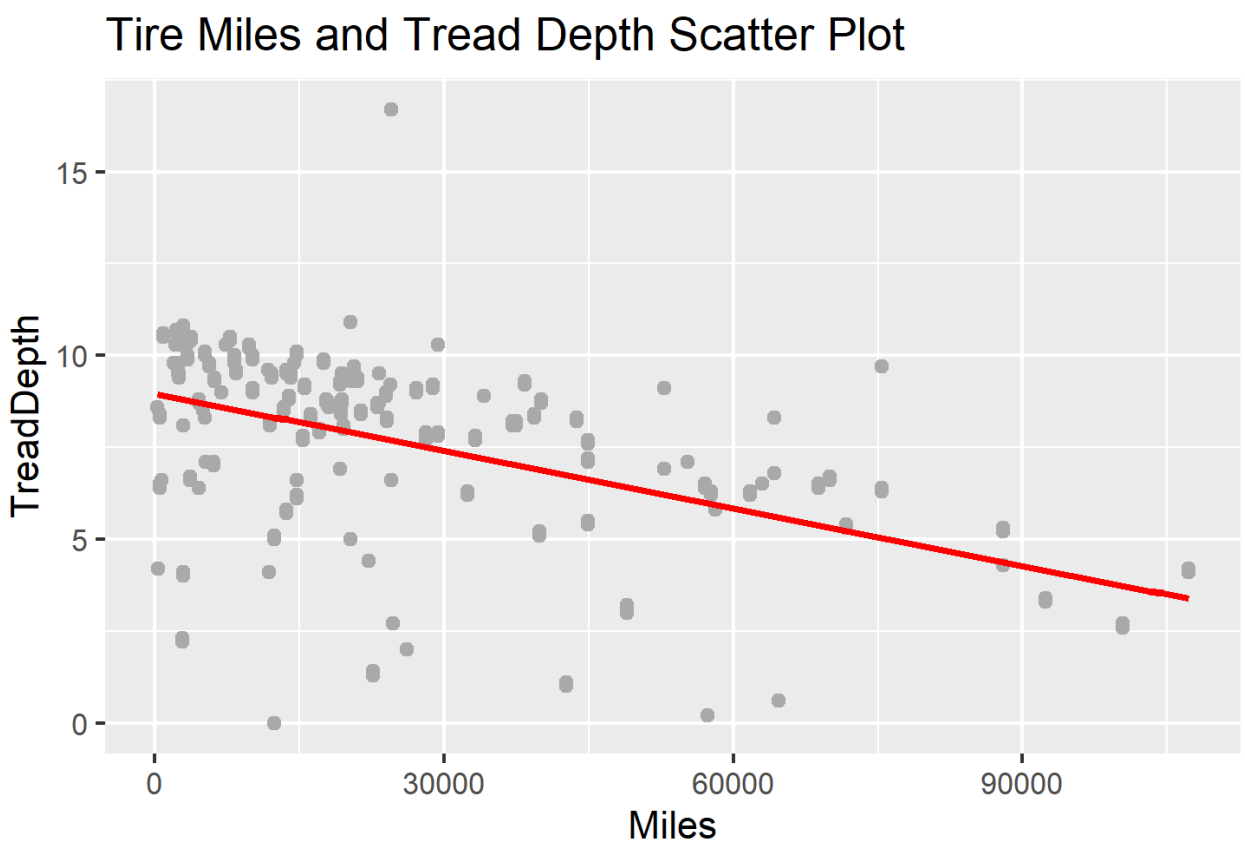
MIS 545 LAB 04

Group 11
Anjana Pillai, Vihari Reddy Tummuru
09/19/2022

```
# Vihari Reddy Tummuru, Anjana Pillai
# MIS 545 -001
# Lab04Group11PillaiTummuru.R
# The following code processes data from TireTread.csv
# Calculates outliers min,max, mormalizes certain columns,
# Discretizes and plots a scatterplot to visualize data

# Set the current working directory
setwd("C:/Users/anjan/Downloads/Lab04")

# Install  dummies package
install.packages("dummies", repos = NULL, type="source")

# Load dummies
library(dummies)

#Load tidyverse
library(tidyverse)

# Read tireTread1 as a tibble from the csv
tireTread1 <- read_csv(file="TireTread.csv",
                    col_types="cfnni",
                    col_names=TRUE)

# Display the first 20 rows
print(tireTread1)

# Display the structure
str(tireTread1)

# Display the Summary
summary(tireTread1)

# Impute missing data for UsageMonths with the mean
# Read into tireThread2 tibble
tireTread2 <- tireTread1 %>%
              mutate(UsageMonths = ifelse(is.na(UsageMonths),
                     mean(UsageMonths,na.rm = TRUE),UsageMonths))

# Display summary Of Tiretread2
summary(tireTread2)

# Determine outliers in the TreadDepth feature
# Calculate outlier min and max and store into variables called
# outlierMin and outlierMax.
outlierMin <-quantile(tireTread2$TreadDepth,.25) -
            (IQR(tireTread2$TreadDepth)* 1.5)

outlierMax <- quantile(tireTread2$TreadDepth,.75) +
             (IQR(tireTread2$TreadDepth)* 1.5)

# Add the outliers to their own tibble called treadDepthOutliers
```

```
treadDepthOutliers <- tireTread2 %>%
                    filter(TreadDepth < outlierMin | TreadDepth > outlierMax)

# Normalize the UsageMonths feature by taking the log of UsageMonths into a
# New feature called LogUsageMonths and store the additional column in a tibble
# Called tireTread3
tireTread3<- tireTread2%>%
             mutate(LogUsageMonths = log(UsageMonths))

# Discretize TreadDepth into NeedsReplacing and store into tireTread4 tibble
tireTread4 <- tireTread3%>%
              mutate(NeedsReplacing = TreadDepth <= 1.6)

# Convert tireTread4 into data frame
tireTread4DataFrame <- data.frame(tireTread4)

# Dummy code the Position using dummy.data.frame(), convert it back into
# a tibble, and store the result into a new tireTread5 tibble
tireTread5 <- as_tibble(dummy.data.frame(data = tireTread4DataFrame,
                                         names = "Position"))

# Build a scatter plot of Miles (x) with TreadDepth (y)
# Change point colour to dark grey,
# Title is added using ggtitle()
# Linear best fit using geom_smooth() and changing the color to red
scatterplotMilesTreadDepth <-  ggplot(data = tireTread5,
                                      aes(x = Miles,
                                          y = TreadDepth))
scatterplotMilesTreadDepth + geom_point(color = "dark grey") +
                             ggtitle("Tire Miles and Tread Depth Scatter Plot")+
                             geom_smooth(method = lm,
                                         level = 0,
                                         color = "red")
```

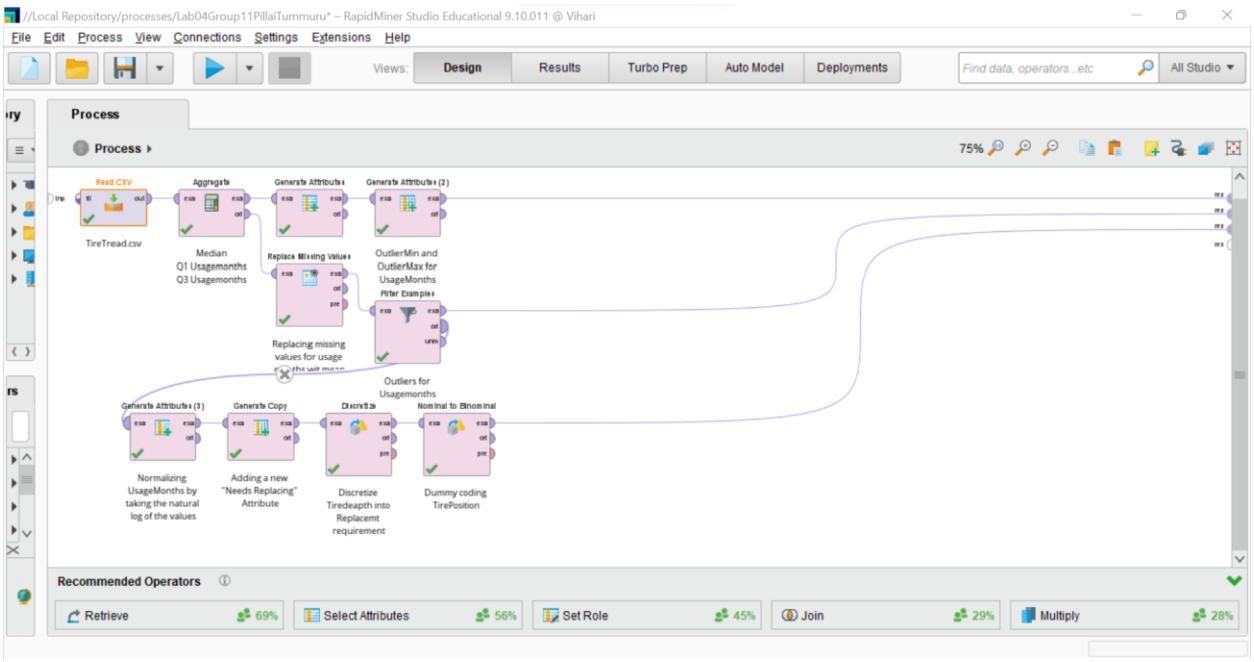
Scatterplot:



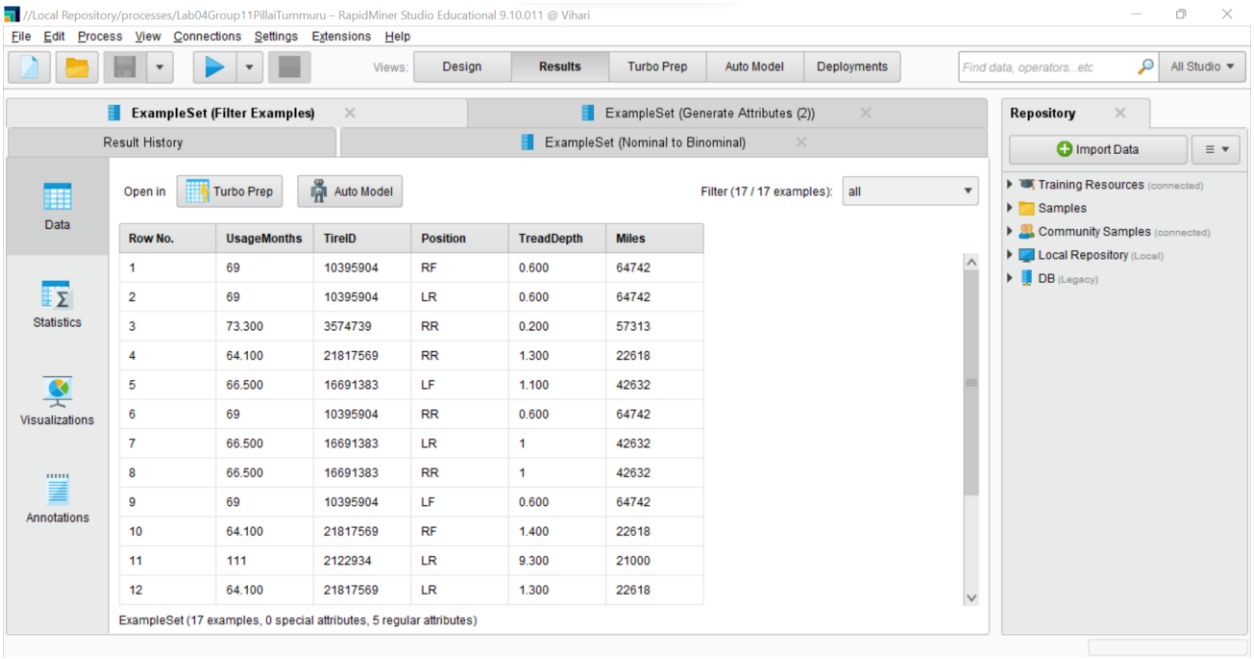
Analysis:

Visually we can say that there is a negative correlation between two features as the line of regression is with a negative slope from the scatter plot above.

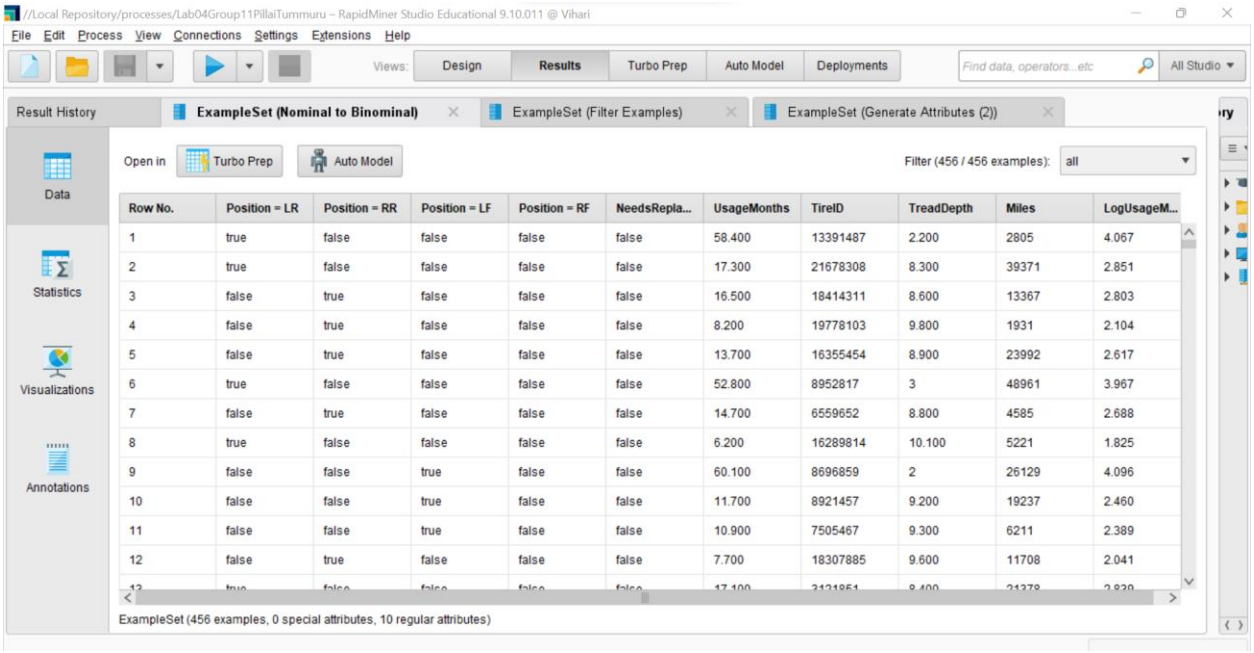
Rapid Miner Process:



Rapid Miner Identified Outliers:



Rapid Miner Results:



Rapid Miner Scatterplot: