---
title: "Similarity Part 1: Regression"
output: html_notebook

Made by: Jonathan Blade

Data set can be found here: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

The goal of using this data set is to calculate the number of bike rentals during a one hour period.

````
```{r}
bikeData <- read.csv("C:\\Users\\18327\\Desktop\\Academics\\Academics Fall 2022\\Machine Learning\\Portfolio Similarity\\hour.csv", header=TRUE)

#Treat columns as factors for regression
bikeData$season <- as.factor(bikeData$season)
bikeData$yr <- as.factor(bikeData$yr)
bikeData$weathersit <- as.factor(bikeData$weathersit)
bikeData$weekday <- as.factor(bikeData$weekday)
bikeData$workingday <- as.factor(bikeData$workingday)
bikeData$mnth <- as.factor(bikeData$mnth)
bikeData$hr <- as.factor(bikeData$hr)

#Remove bad columns
bikeData <- bikeData[,-16]
bikeData <- bikeData[,-15]
bikeData <- bikeData[,-2]
bikeData <- bikeData[,-1]

sapply(bikeData, function(x) sum(is.na(x)))

bikeData <- bikeData[(complete.cases(bikeData)),]
sum(is.na(bikeData))

bikeData <- bikeData[(complete.cases(bikeData)),]
sum(is.na(bikeData))

set.seed(12345)

sample <- sample(c(TRUE,FALSE), nrow(bikeData), replace=TRUE, prob=c(0.8,0.2))
train <- bikeData[sample, ]
test <- bikeData[!sample, ]

summary(train)

tempCor <- cor(train$temp, train$cnt)
print(paste("Correlation between temp and cnt: ", tempCor))

atempCor <- cor(train$atemp, train$cnt)
print(paste("Correlation between atemp and cnt: ", atempCor))

humCor <- cor(train$hum, train$cnt)
````
````

```
print(paste("Correlation between humidity and cnt: ", humCor))

boxplot(train$cnt ~ train$hr)

boxplot(train$cnt ~ train$mnth)

lm1 <- lm(cnt ~ ., data = train)

par(mfrow=c(2,2))
plot(lm1)
par(mfrow=c(1,1))
summary(lm1)

pred1 <- predict(lm1, newdata=test)

correlation1 <- cor(pred1, test$cnt)
print("Model 1: ")
print(paste("Correlation: ", correlation1))
mse1 <- mean((pred1 - test$cnt)^2)
print(paste("MSE: ", mse1))
rmse1 <- sqrt(mse1)
print(paste("RMSE: ", rmse1))

```
The above linear regression model has reasonable correlation and accuracy. This will provide a baseline to compare the results of our next two algorithms.

```{R}
library(caret)

#Convert factors back to numerics for kNN
train$yr <- as.integer(train$yr)
train$weathersit <- as.integer(train$weathersit)
train$weekday <- as.integer(train$weekday)
train$workingday <- as.integer(train$workingday)
train$mnth <- as.integer(train$mnth)
train$hr <- as.integer(train$hr)

test$yr <- as.integer(test$yr)
test$weathersit <- as.integer(test$weathersit)
test$weekday <- as.integer(test$weekday)
test$workingday <- as.integer(test$workingday)
test$mnth <- as.integer(test$mnth)
test$hr <- as.integer(test$hr)

fit1 <- knnreg(train[,1:12], train[,13], k=8)
summary(fit1)

pred2 <- predict(fit1, test[,1:12])
cor_knn1 <- cor(pred2, test$cnt)
mse_knn1 <- mean((pred2 - test$cnt)^2)
rmse_knn1 <- sqrt(mse_knn1)
print(paste("Cor = ", cor_knn1))
```

```
print(paste("MSE = ", mse_knn1))
print(paste("RMSE = ", rmse_knn1))

```
The kNN algorithm provides both a higher correlation and accuracy than the linear regression model.

```{R}
library(tree)

tree_bike <- tree(cnt ~ ., data=train)

summary(tree_bike)

pred3 <- predict(tree_bike, newdata=test)
cor_tree <- cor(pred3, test$cnt)
print(paste("Cor: ", cor_tree))
mse_tree <- mean((pred3 - test$cnt)^2)
rmse_tree <- sqrt(mse_tree)
print(paste("MSE: ", mse_tree))
print(paste("RMSE: ", rmse_tree))

plot(tree_bike)
text(tree_bike, cex=0.5, pretty=0)

cv_tree <- cv.tree(tree_bike)

```
The decision tree is less accurate than both the Linear Regression model and the kNN algorithm.
But, the decision tree's greatest strength is how easy it is to interpret.