STATS769 Lab1

Yujie Zhang 130011770 yzhb915

August 2, 2019

1. The dataset is of csv format and was imported by following steps:

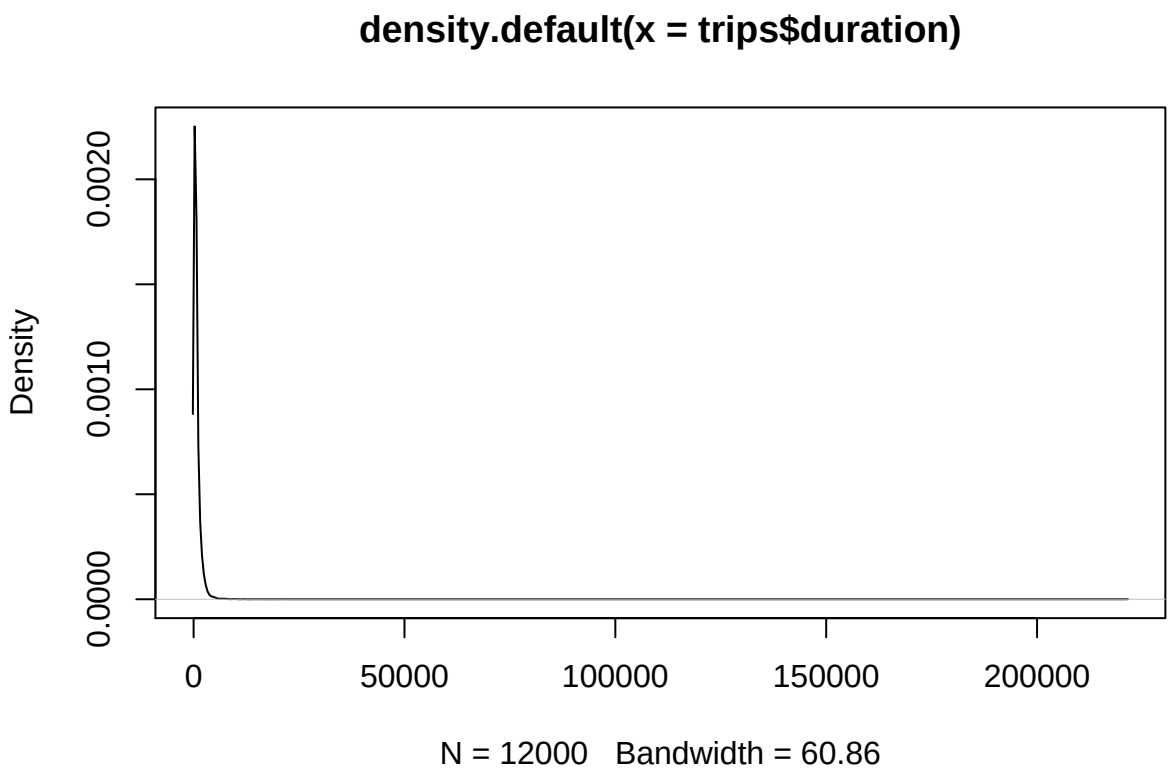
```
## Read the data from csv files
trips_2018_7 <- read.csv("./trips-2018-7.csv", header = TRUE)
trips_2018_8 <- read.csv("./trips-2018-8.csv", header = TRUE)
trips_2018_9 <- read.csv("./trips-2018-9.csv", header = TRUE)
```

2. The training and test sets are created as following:

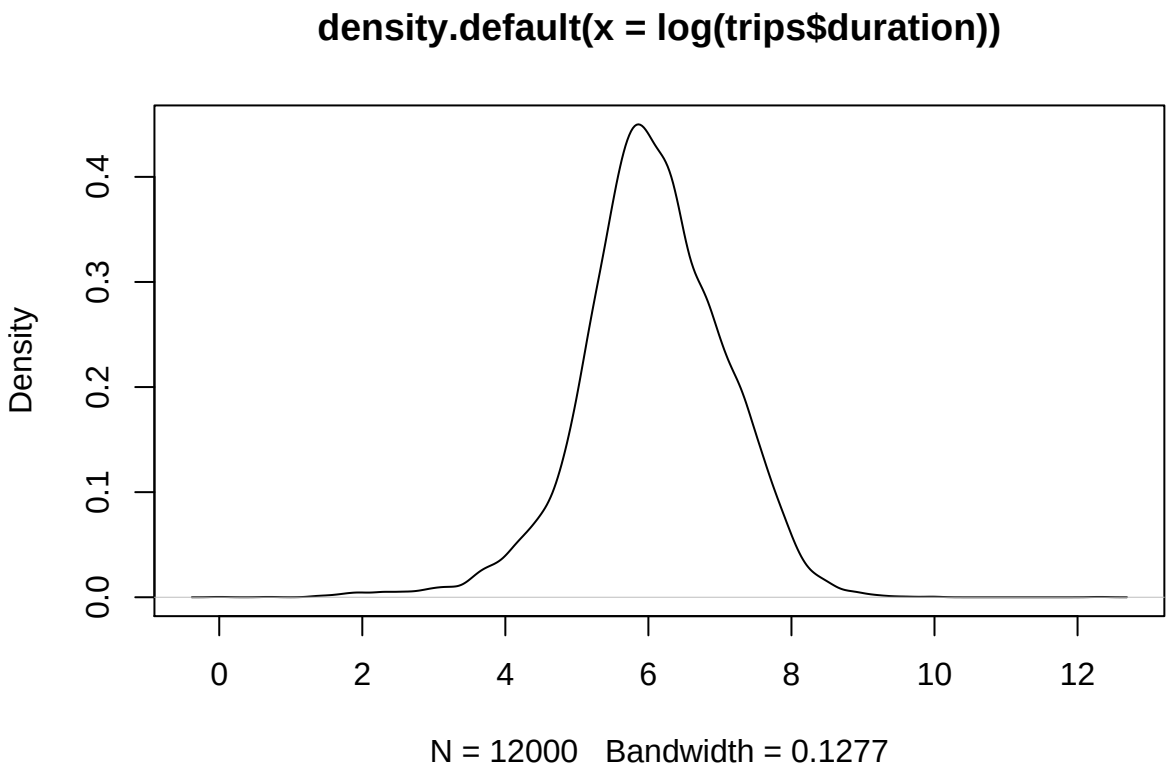
```
## Combine into single object
trips <- rbind(trips_2018_7, trips_2018_8, trips_2018_9)

## Select 1000 samples from each month as test dataset
trips_test_201807 <- sample(1:5000,1000)
trips_test_201808 <- sample(5001:10000,1000)
trips_test_201809 <- sample(10001:15000,1000)
trips_test_index <- rbind(trips_test_201807, trips_test_201808, trips_test_201809)
trips_test <- trips[trips_test_index,]
trips <- trips[-trips_test_index,]
```

3. Data exploration and summary of variables:

```
## Distribution of duration
plot(density(trips$duration))
```

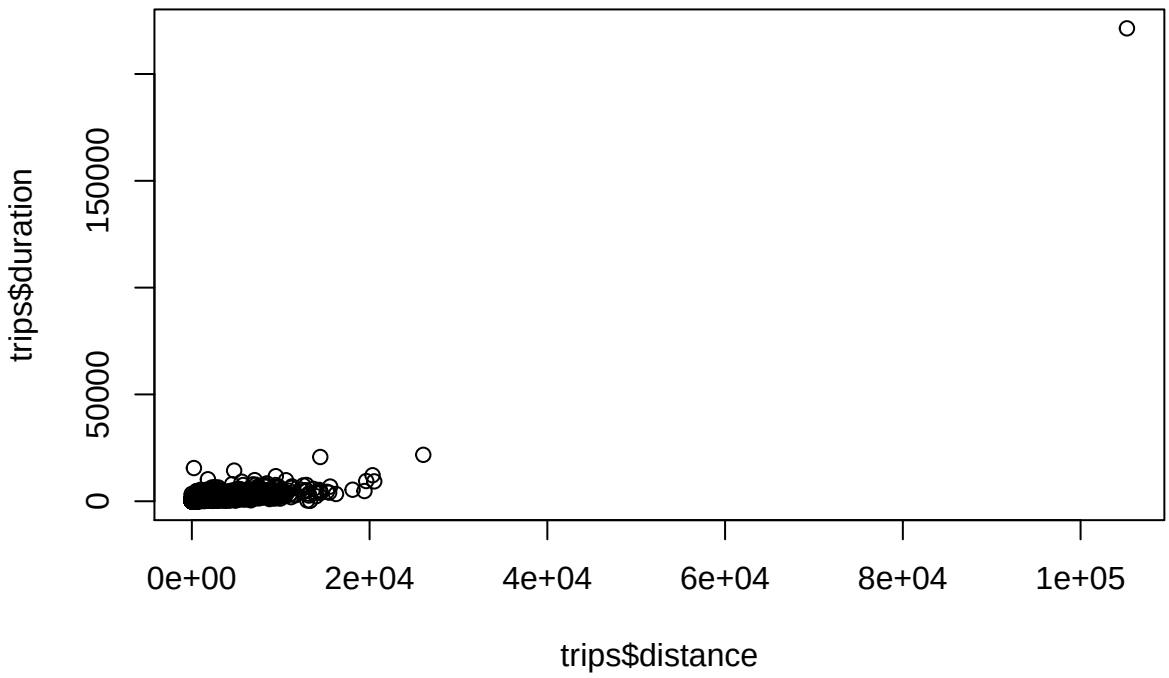



```
plot(density(log(trips$duration)))
```



The distribution of duration is more like a normal distribution after the log transformation.

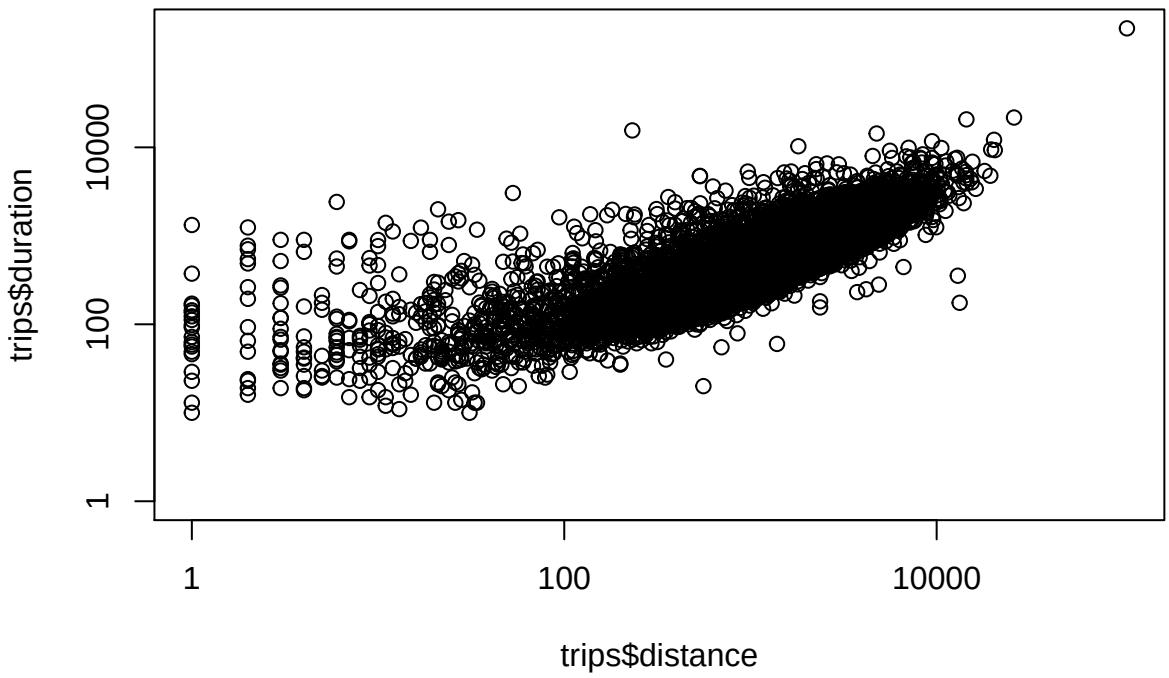
```
## Relationship between distance and duration times
plot(trips$distance, trips$duration)
```

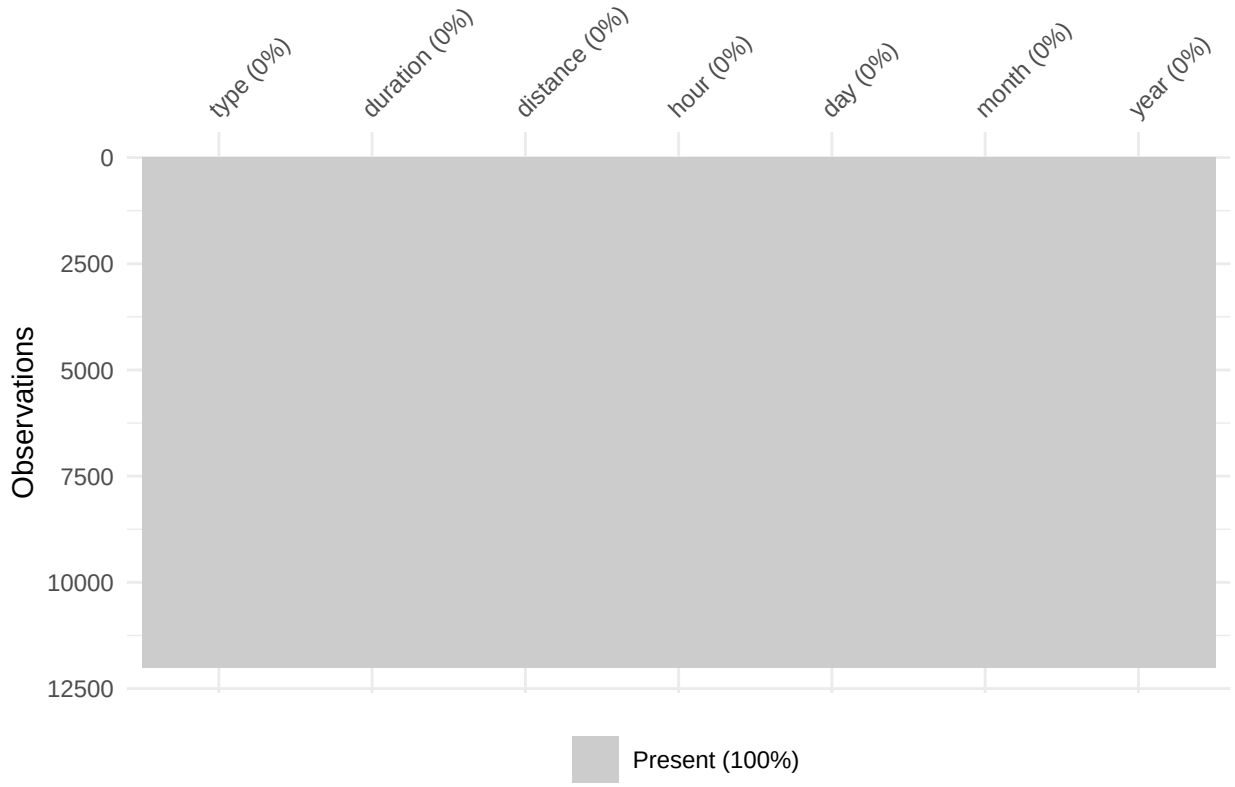



```
plot(trips$distance, trips$duration, log="xy")
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 916 x values <= 0 omitted
## from logarithmic plot
```

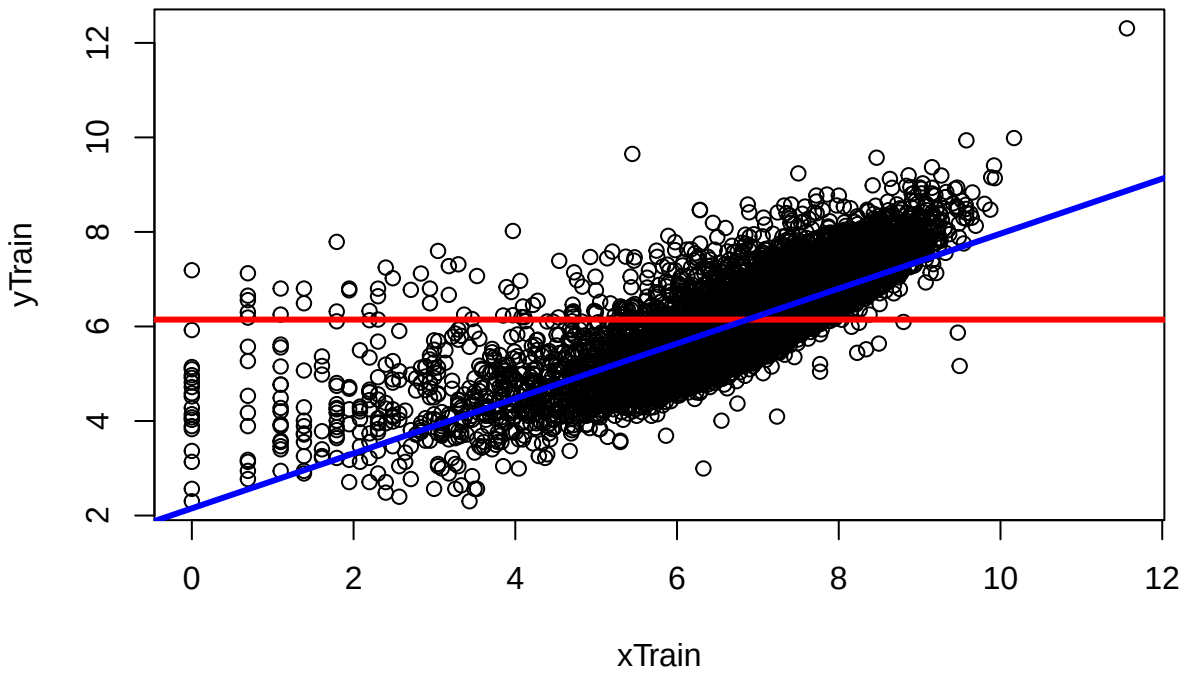
```
## Check missing data
library(visdat)
vis_miss(trips)
```


trips$duration
10000
100
1
1
100
10000
trips$distance



4. Model fitting using a training dataset:

```
## Transform training dataset
trainDuration <- log(trips$duration)
trainSubset_duration <- is.finite(trainDuration)

trainDistance <- log(trips$distance)
trainSubset_distance <- is.finite(trainDistance)

trainSubset <- trainSubset_duration & trainSubset_distance

## Linear Model with training dataset
yTrain <- trainDuration[trainSubset]
xTrain <- trainDistance[trainSubset]
fitMean <- mean(yTrain, na.rm=TRUE)
fitLM <- lm(y ~ x, data.frame(y=yTrain, x=xTrain))

## Visualize the model
plot(xTrain, yTrain)
abline(h=fitMean, col="red", lwd=3)
abline(fitLM, col="blue", lwd=3)
```




```
## Transform test dataset
testDuration <- log(trips_test$duration)
testSubset_duration <- is.finite(testDuration)
testDistance <- log(trips_test$distance)
testSubset_distance <- is.finite(testDistance)
testSubset<- testSubset_duration&testSubset_distance

## Linear Model with testing dataset
yTest <- testDuration[testSubset]
xTest <- testDistance[testSubset]
predMean <- rep(fitMean, length(yTest))
predLM <- predict(fitLM, data.frame(x=xTest))
```

5. Model evaluation using a test dataset:

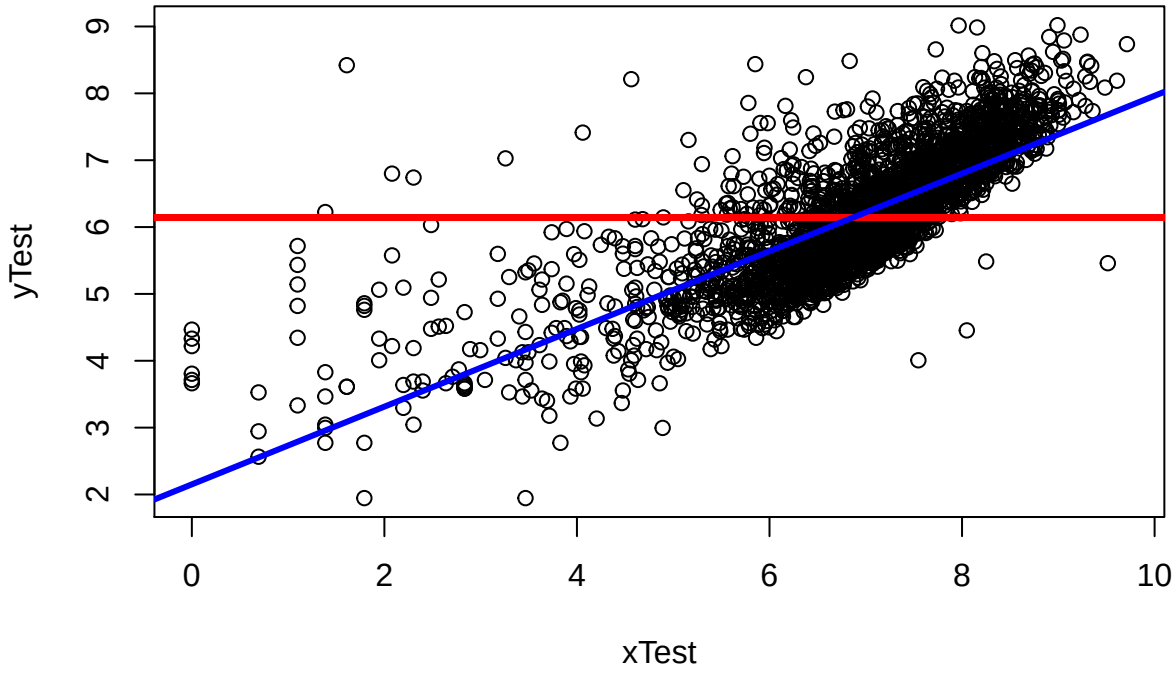
```
## Evaluate models
RMSE <- function(m, o) {
  sqrt(mean((m - o)^2))
}
RMSE(predMean, yTest)
```

```
## [1] 0.964461
```

```
RMSE(predLM, yTest)
```

```
## [1] 0.6144801
```

```
## Visualise models
plot(xTest, yTest)
abline(h=predMean, col="red", lwd=3)
abline(fitLM, col="blue", lwd=3)
```



6. Conclusion: According to the value of RMSE, it fits linear mode well and the prediction is good.