Appendix A

```
# ##############################################################################################################
# # PACKAGES
library(ggplot2)
library(ggmap)
library(nlme)
library(lattice)
library(MASS)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
# ##############################################################################################################
# # DATA
dist <- read.table(file = "http://stat.ufl.edu/~winner/data/distang26.dat",
           col.names = c("Airport1", "Airport2", "Population1", "Population2",
                         "Distance", "Angle", "Latitude.City1", "Latitude.City2",
                         "Longitude.City1", "Longitude.City2"))
rout <- read.table("http://www.stat.ufl.edu/~winner/data/longair.dat")
colnames(rout) <- c("Route", "Quarter", "Avg.Fare", "Avg.Pass")
data.both <- cbind(rout, dist)
head(rout, 10)
```

```
##    Route Quarter Avg.Fare Avg.Pass
## 1      1       1   198.96    72.50
## 2      1       2   205.00    91.30
## 3      1       3   230.46    76.77
## 4      1       4   216.53   121.31
## 5      1       5   221.16    91.63
## 6      1       6   235.97    95.00
## 7      1       7   265.75    78.11
## 8      1       8   234.43    99.01
## 9      1       9   204.96    97.17
## 10     1      10   203.77   108.47
```

```
rout$Route <- as.factor(rout$Route)


sub1 <- groupedData(formula = Avg.Fare ~ Quarter| Route, data = rout)
sub2 <- sub1[1:max(which(rout[,1] == "50")),] #First 50 routes
sub.10 <- sub1[1:max(which(rout[,1] == "10")),] #First 50 routes

sub2.1 <- sub1[1:max(which(rout[,1] == "3")),] #First 50 routes
set.seed(700)
samp <- sort(sample( 1:4177,50, replace = F))
samp <- as.character(samp)
rand.samp <- data.both[which(data.both[,1] %in% samp),]



dim(sub2)
```

```
## [1] 1300    4
```

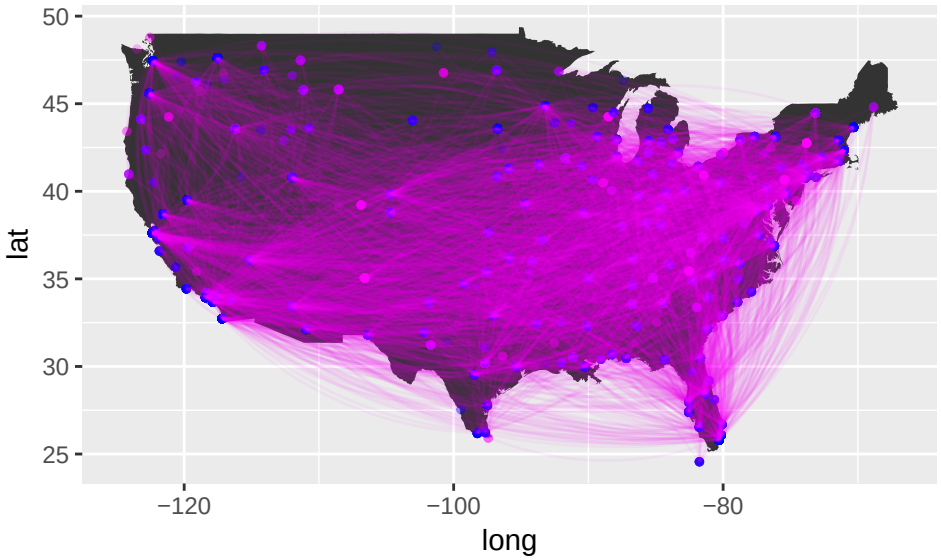
```
dim(rout)
```

```
## [1] 108602      4
```

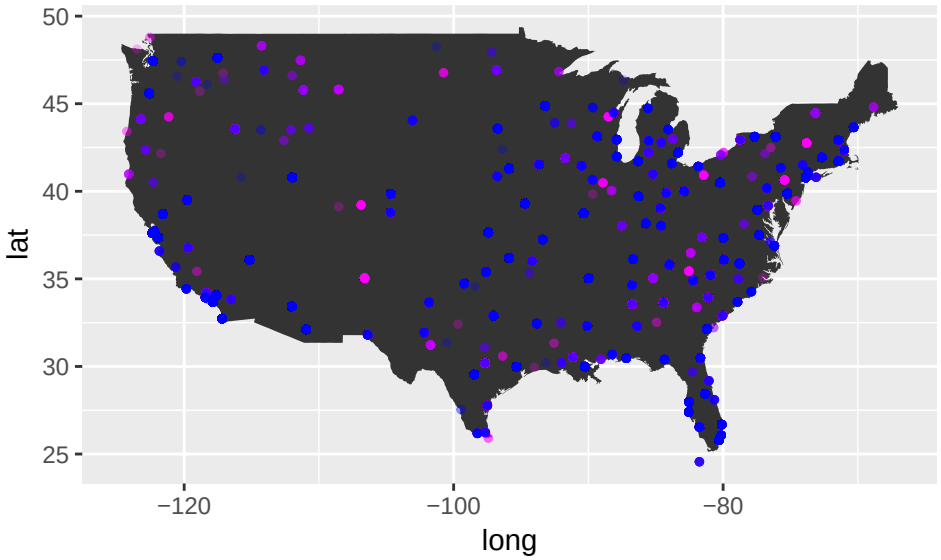
```
str(rout)
```

```
## 'data.frame': 108602 obs. of  4 variables:
##  $ Route   : Factor w/ 4177 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Quarter : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Avg.Fare: num  199 205 230 217 221 ...
##  $ Avg.Pass: num  72.5 91.3 76.8 121.3 91.6 ...
```

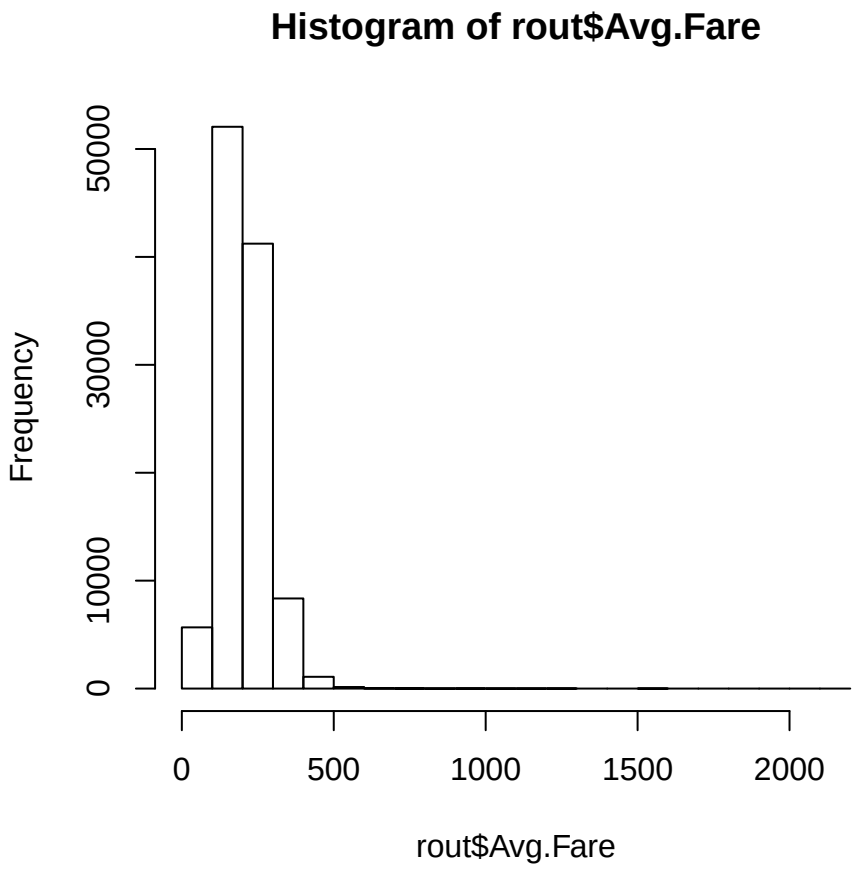
```
#map
states <- map_data("usa")
#iowa <- subset(states, region %in% c("iowa"))
p <- ggplot() + geom_polygon(data = states, aes(x=long, y = lat, group = group)) +
  coord_fixed(1.3) +
  geom_point(data = dist, aes(x = dist$Longitude.City1, y = dist$Latitude.City1),
             color = rgb(1,0,1,1/4), size = 1) +
  geom_point(data = dist, aes(x = dist$Longitude.City2, y = dist$Latitude.City2),
             color = rgb(0,0,1,1/4), size = 1)

p + geom_curve(aes(x = dist$Longitude.City1, y = dist$Latitude.City1,
                   xend = dist$Longitude.City2, yend = dist$Latitude.City2),
               data = dist, curvature = -.2, colour = rgb(1,0,1, .05));
```



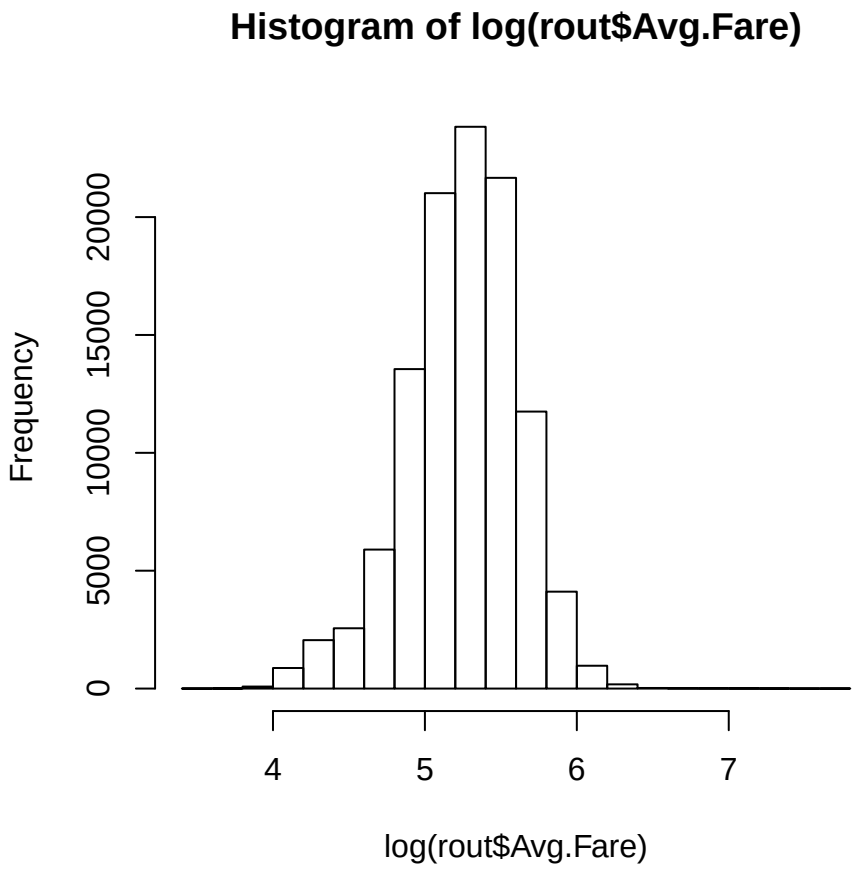

```
p
```



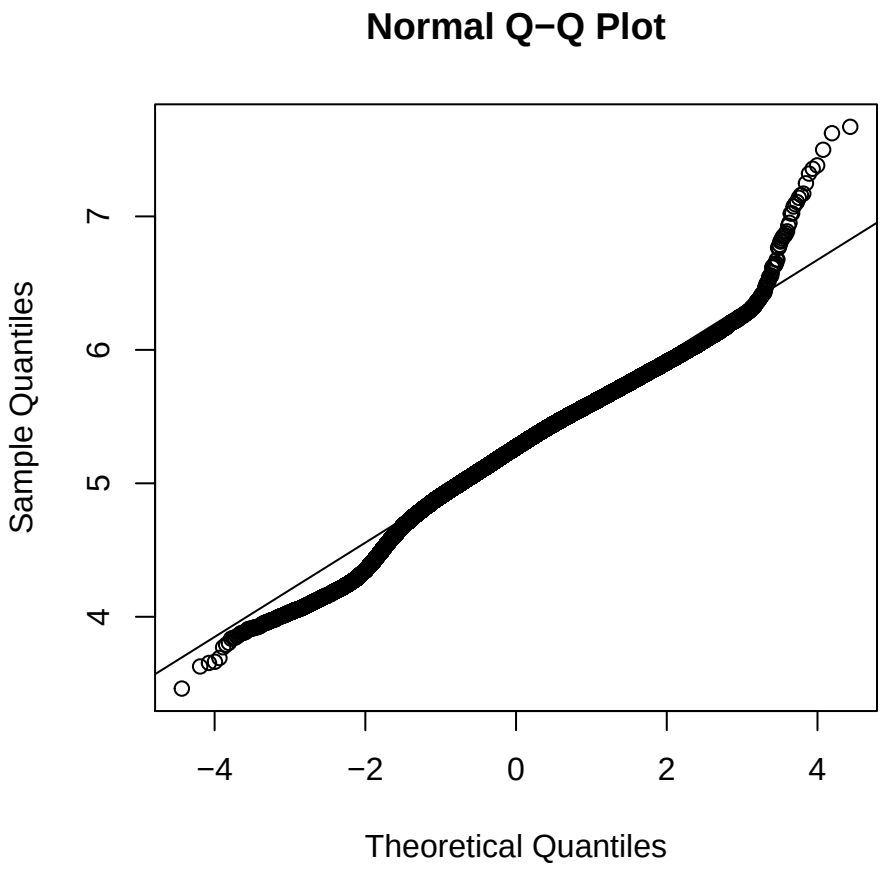

```
#############################################################################################################
#EDA
attach(rout)
sum(Avg.Fare <= 0)

## [1] 0

hist(rout$Avg.Fare) #not normal, consider log transformation
```




```
hist(log(rout$Avg.Fare)) #Looks more normal
```




```
qqnorm(log(rout$Avg.Fare))
qqline(log(rout$Avg.Fare)) #Deviations in the right tail is evident,
```




```
#consider boxcox transformation

box <- MASS::boxcox(rout$Avg.Fare~ rout$Quarter + rout$Avg.Pass, data = rout)
```