STA141AHW1

Sam Tsoi

Due: 10/31/2017

1.

Unzip and load the airfare dataset. Convert the columns to appropriate data types, then separate table 1a and 6 into different variables (to help you avoid double counting). You don't need to write an answer for this question, but please mark the code for this question in the appendix.

Answer is attached in the code.

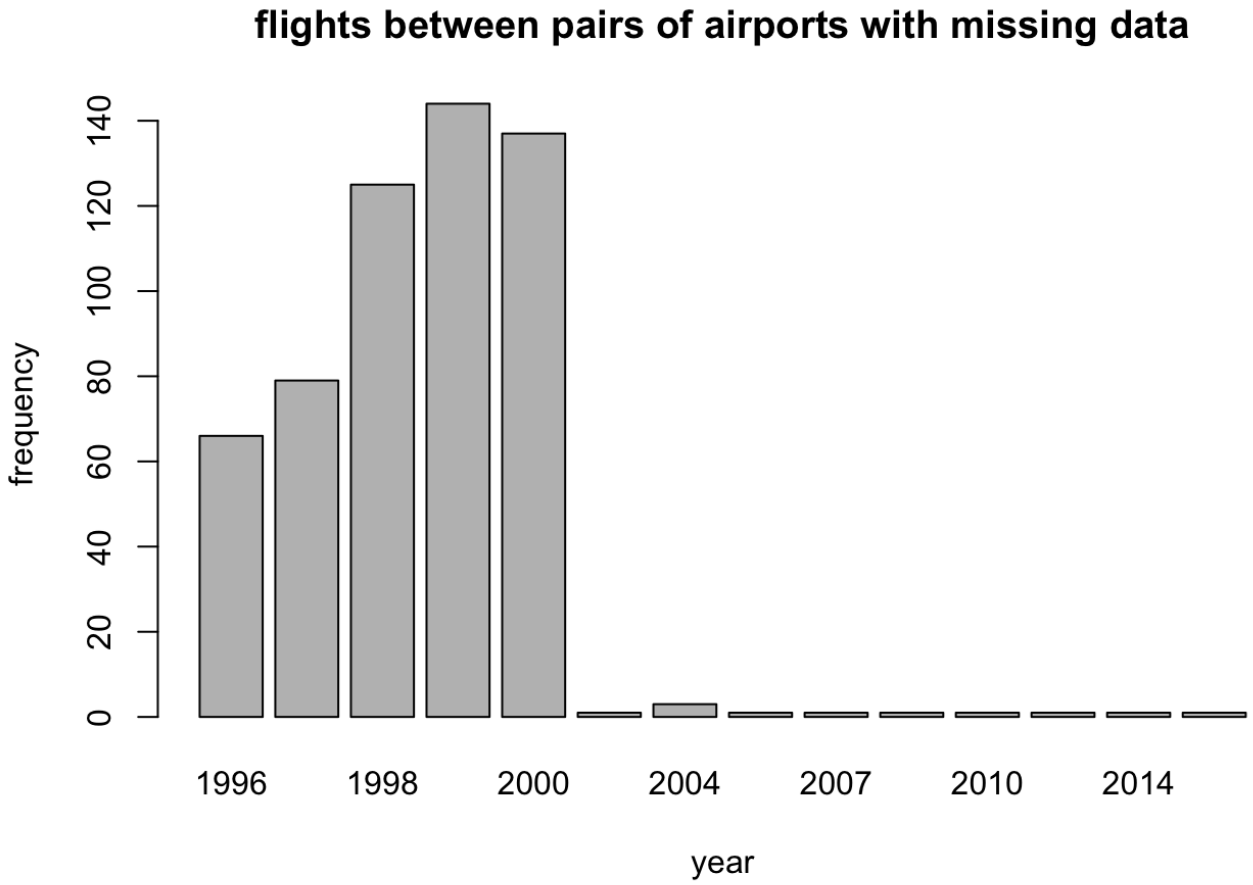
2.

What timespan does the data cover? Do any quarters or years in that span have no data? Check separately for table 1a and table 6. In addition, check both tables for patterns in the missing values.
The data covers 21 years, between 1996 to 2017, each year and quarter with the following frequency, respectively:

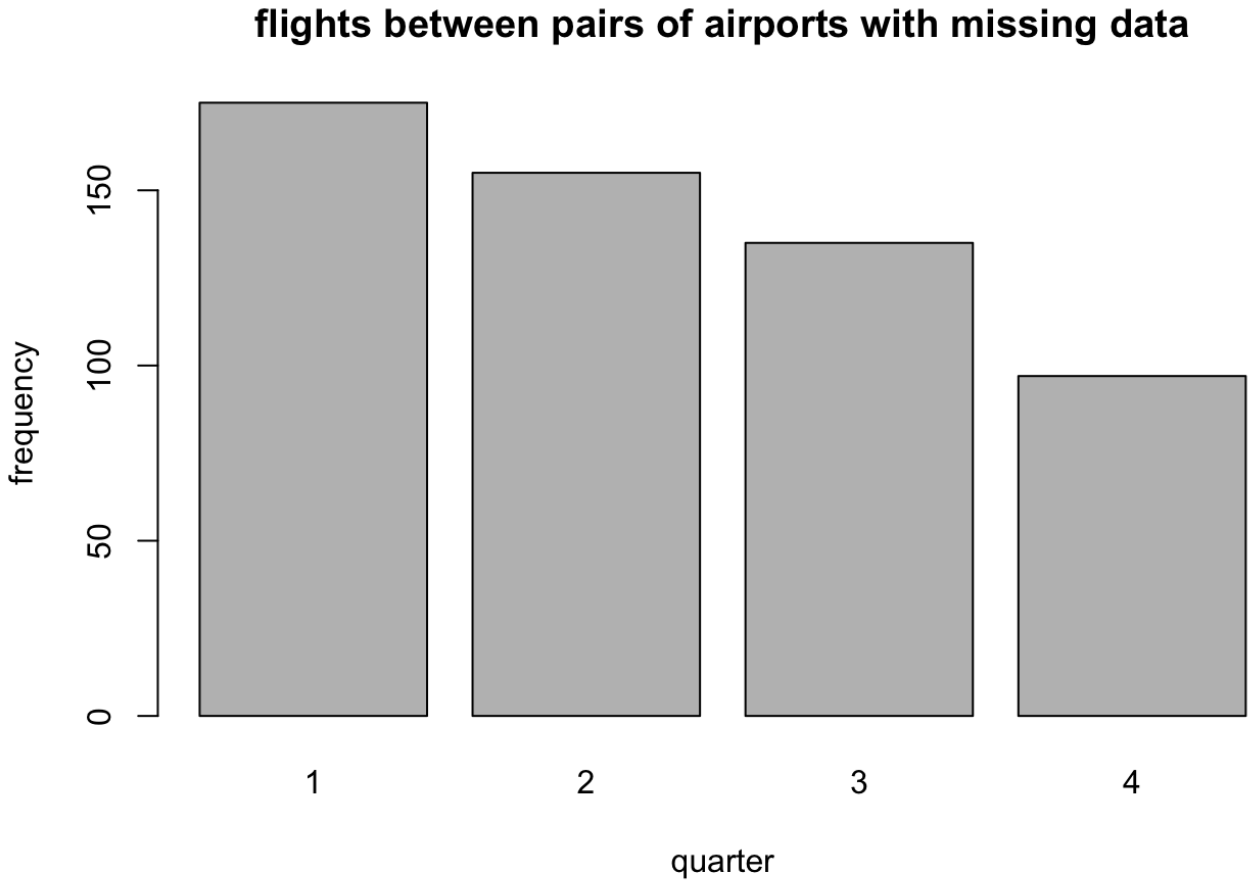
```
##
##  1996  1997  1998  1999  2000  2001  2002  2003  2004  2005  2006  2007
## 28783 30085 30187 30214 30389 29672 29501 29349 30032 30608 30299 29925
##  2008  2009  2010  2011  2012  2013  2014  2015  2016  2017
## 29687 29513 30046 30321 30302 30338 30970 30610 28739  7500
```

```
##
##      1      2      3      4
## 158891 158969 161004 158206
```

There are many quarters and years that do not have data. I counted the number of NAs in each row and created a new variable per row. In these following graphs, I regarded the flights with more than 6 NAs had missing data, as I checked that these data did not give information on airports, lg data, and low data. Rows with 4 or 5 NAs had these data but just not the airport names/IDs. So, I did not disregard these data.
Thus, for flights between the pair of airports (named as table 1a in my data), there is a total of 0 flights (rows) that had missing data (has with more than 6 NAs). Each year and frequency is shown:

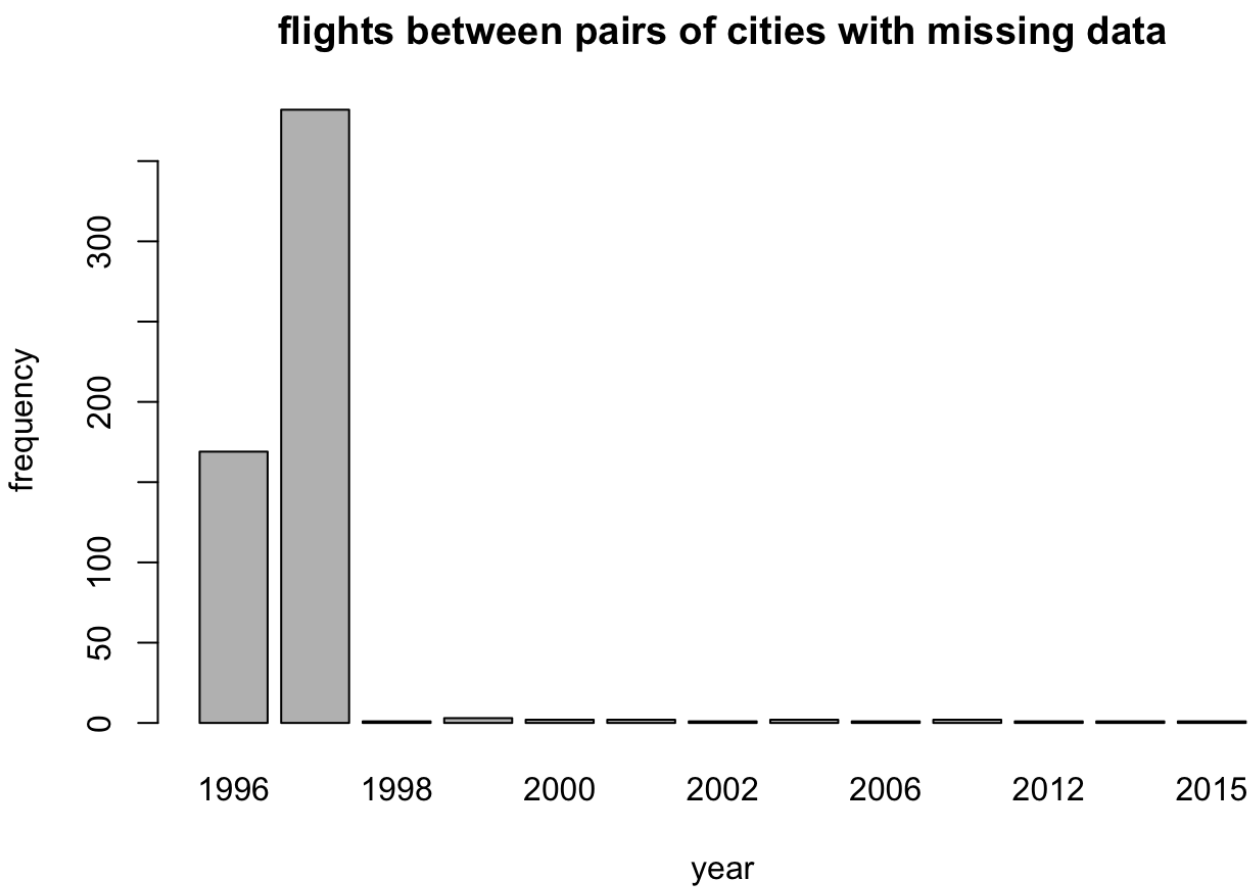


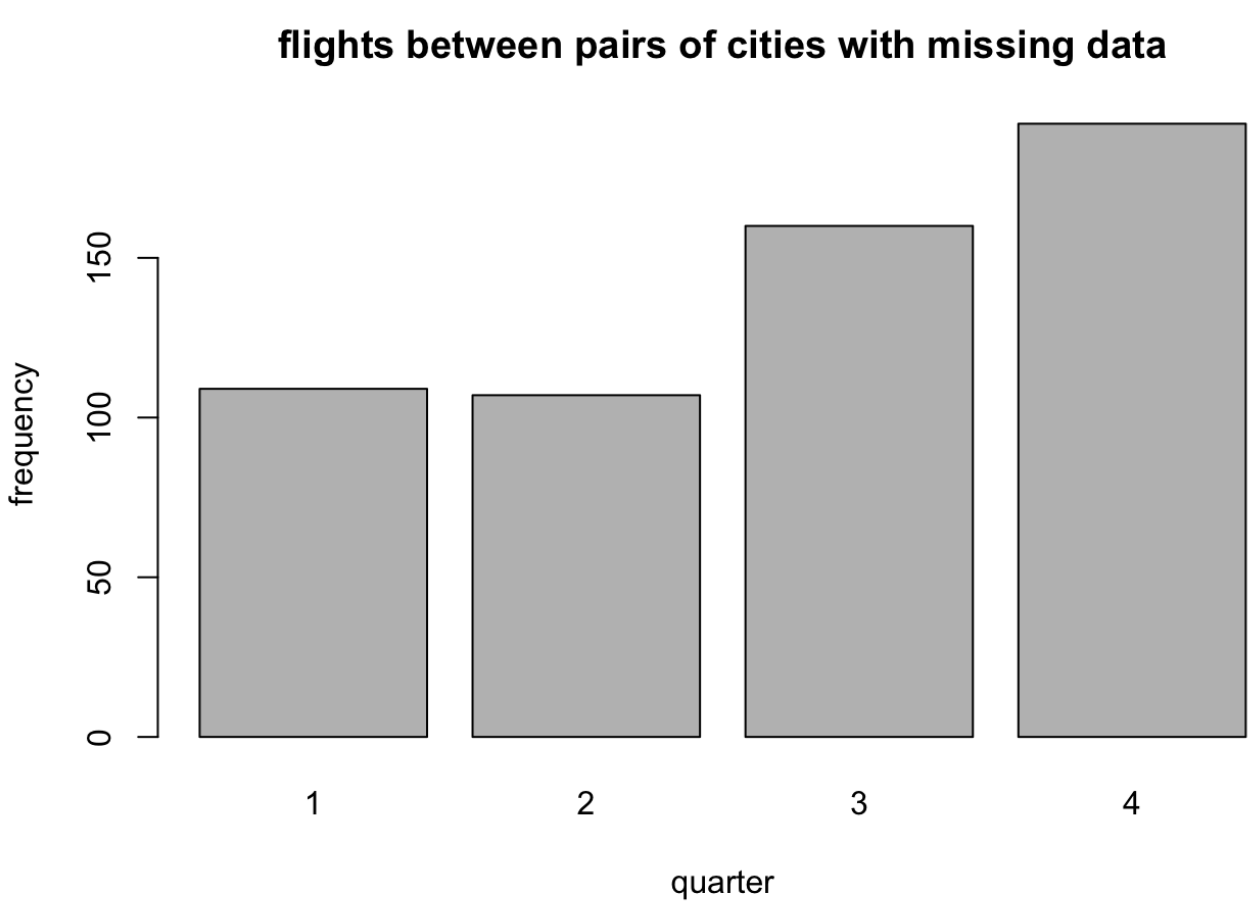
Additionally, it is broken down by quarters, with frequencies as shown:



It seems like earlier years have more missing data than later years. However, I do not believe that it is because data from earlier years has error. This is probably because data collecting gets more advanced as time goes on, not because there was something wrong with the flights earlier on. Additionally, it seems like number of missing data decreases as quarter goes on. I feel like this is random, as this is different from the results for flights between pairs of cities (shown later on) and also the frequency between each quarter does not differ significantly. Further exploration and investigation would be needed in order to make a more appropriate conclusion.

For flights between pairs of cities (named as table 6 in my data), there is a total of 568 flights (rows) that did not have data (has with more than 6 NAs). Each year and frequency is shown:





So, it seems like most data from flights between pairs of cities is missing from year 1996 and 1997. As mentioned, this is probably because data collecting gets more thorough and advanced as time goes on. There could have been something happened in 1997 that made their data collecting much more thorough. For flights between pairs of cities, it does not seem like the number of missing data decreases over quarters. Instead, it increases, and there is the most missing data by the fourth quarter. This might be because the fourth quarter (around holiday time), there is a high frequency of flights in general and it might be more difficult to manage all the data for all flights during this time. Further exploration and investigation could be done in order to make a more appropriate conclusion.

3.

In 2017, which cities have the most connections to other cities? Which have the least? How do these results compare to 10 years earlier? 20 years earlier? Which cities have increased connectivity the most?
To find the top 10 cities that have the most connections to other cities in 2017, I merged the out connection to in connection by adding the numbers of these in each city, and sorted them by city. The top 10 cities with most connections are The 10 cities that have the most connections to other cities in 2017 (we only have data for quarter 1 in 2017), with the relative frequencies shown, are:

```
##  [1] Chicago, IL
##  [2] Dallas/Fort Worth, TX
##  [3] Denver, CO
##  [4] Las Vegas, NV
##  [5] Los Angeles, CA (Metropolitan Area)
##  [6] Miami, FL (Metropolitan Area)
##  [7] New York City, NY (Metropolitan Area)
##  [8] Orlando, FL
##  [9] Phoenix, AZ
## [10] San Francisco, CA (Metropolitan Area)
## [11] Washington, DC (Metropolitan Area)
## 317 Levels: Aberdeen, SD Abilene, TX Albany, GA ... West Palm Beach/Palm Beach,
FL
```

These are the cities in 2017 with 0 connections (least connections) when in and out connections are combined:

```
##  [1] Branson, MO              Bullhead City, AZ
##  [3] Carlsbad, CA             Dickinson, ND
##  [5] Farmington, NM           Hickory, NC
##  [7] Hyannis, MA              International Falls, MN
##  [9] Inyokern, CA             Jamestown, NY
## [11] Kinston, NC              Lewisburg, WV
## [13] Longview, TX             Macon, GA
## [15] Marathon, FL             Martha's Vineyard, MA
## [17] Modesto, CA              Muskegon, MI
## [19] Nantucket, MA            Naples, FL
## [21] Oxnard/Ventura, CA       Parkersburg, WV
## [23] Pellston, MI             Port Angeles, WA
## [25] Presque Isle/Houlton, ME Provincetown, MA
## [27] Reading, PA              Rhinelander, WI
## [29] Rockland, ME             St. Augustine, FL
## [31] Texarkana, AR            Tupelo, MS
## 317 Levels: Aberdeen, SD Abilene, TX Albany, GA ... West Palm Beach/Palm Beach,
FL
```

I believe that there are more cities that do not have connections whatsoever but I disregarded data with missing data (NAs) for in and out connections. I do not believe that these cities with 0 connections mean that there are no flights for these cities, but instead, there might be something wrong with collecting information for these cities. Something I noticed is that the cities that have fewer connections are cities that are less well-known. For example, Chicago, Dallas/Fort Worth, Denver, Los Angeles, Phoenix, etc. are all major cities, whereas Bullhead City and Carlsbad are not as internationally well-known. There may be less flight connections to these cities as well as maybe there are less resources to collect data at these cities since they are smaller.

_ The 10 cities that have the most connections to other cities in 2007 (I am only displaying results for quarter 1, since 2017 only has data for quarter 1), with the relative frequencies shown, are:

```
##  [1] Atlanta, GA (Metropolitan Area)
##  [2] Chicago, IL
##  [3] Dallas/Fort Worth, TX
##  [4] Las Vegas, NV
##  [5] Los Angeles, CA (Metropolitan Area)
##  [6] New York City, NY (Metropolitan Area)
##  [7] Orlando, FL
##  [8] Phoenix, AZ
##  [9] San Francisco, CA (Metropolitan Area)
## [10] Washington, DC (Metropolitan Area)
## 317 Levels: Aberdeen, SD Abilene, TX Albany, GA ... West Palm Beach/Palm Beach,
FL
```

These are the cities in 2007 with 0 connections (least connections) when in and out connections are combined:

```
##  [1] Branson, MO              Bullhead City, AZ
##  [3] Carlsbad, CA             Dickinson, ND
##  [5] Farmington, NM           Hickory, NC
##  [7] Hyannis, MA              International Falls, MN
##  [9] Inyokern, CA             Jamestown, NY
## [11] Kinston, NC              Lewisburg, WV
## [13] Longview, TX             Macon, GA
## [15] Marathon, FL             Martha's Vineyard, MA
## [17] Modesto, CA              Muskegon, MI
## [19] Nantucket, MA            Naples, FL
## [21] Oxnard/Ventura, CA       Parkersburg, WV
## [23] Pellston, MI             Port Angeles, WA
## [25] Presque Isle/Houlton, ME Provincetown, MA
## [27] Reading, PA              Rhinelander, WI
## [29] Rockland, ME             St. Augustine, FL
## [31] Texarkana, AR            Tupelo, MS
## 317 Levels: Aberdeen, SD Abilene, TX Albany, GA ... West Palm Beach/Palm Beach,
FL
```

The 10 cities that have the most connections to other cities in 1997 (I am only displaying results for quarter 1, since 2017 only has data for quarter 1) are:

```
##  [1] Atlanta, GA (Metropolitan Area)
##  [2] Chicago, IL
##  [3] Dallas/Fort Worth, TX
##  [4] Denver, CO
##  [5] Los Angeles, CA (Metropolitan Area)
##  [6] New York City, NY (Metropolitan Area)
##  [7] Orlando, FL
##  [8] Phoenix, AZ
##  [9] San Francisco, CA (Metropolitan Area)
## [10] Washington, DC (Metropolitan Area)
## 317 Levels: Aberdeen, SD Abilene, TX Albany, GA ... West Palm Beach/Palm Beach,
FL
```

These are the cities in 1997 with 0 connections (least connections) when in and out connections are combined:

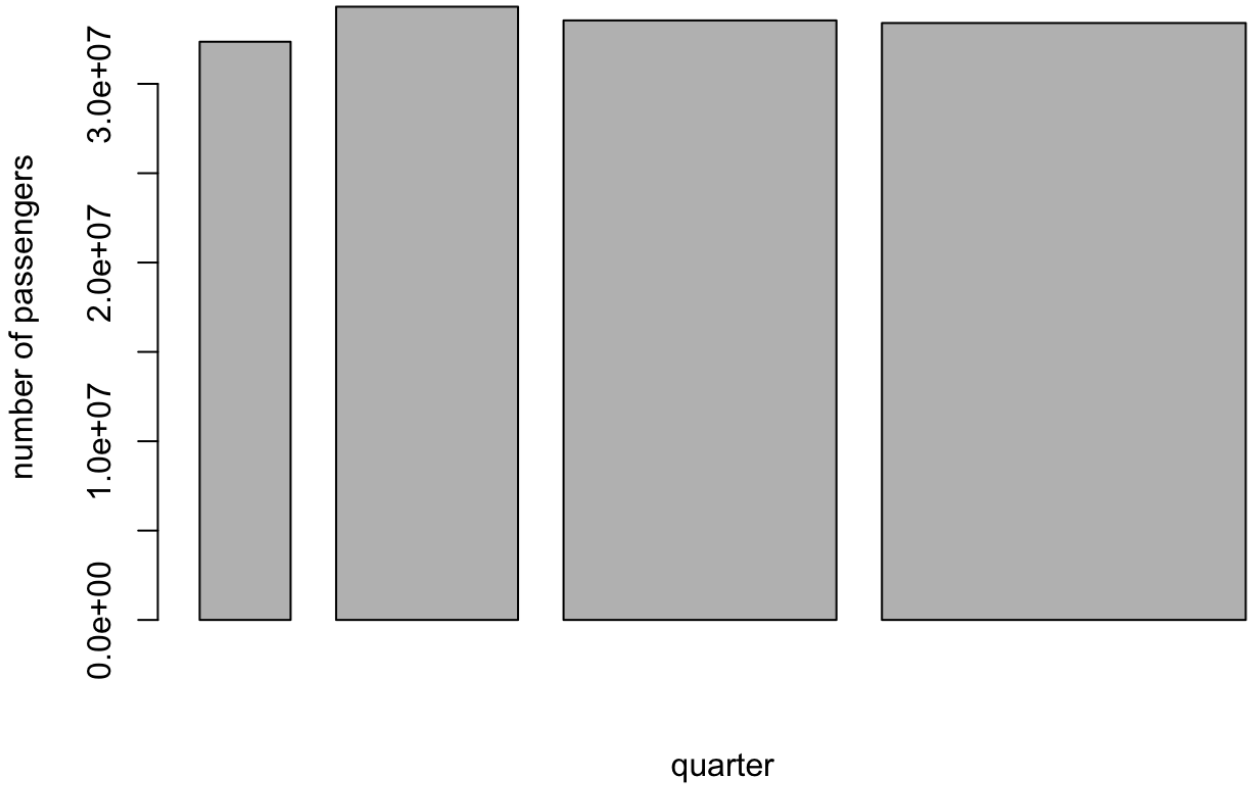
```
##  [1] Branson, MO             Dickinson, ND
##  [3] Dubuque, IA             Eau Claire, WI
##  [5] Escanaba, MI            Grand Island, NE
##  [7] Hagerstown, MD          Hancock/Houghton, MI
##  [9] Hickory, NC             Hyannis, MA
## [11] International Falls, MN Joplin, MO
## [13] Kinston, NC             Lewisburg, WV
## [15] Longview, TX            Mammoth Lakes, CA
## [17] Manhattan/Ft. Riley, KS Martha's Vineyard, MA
## [19] Muskegon, MI            Niagara Falls, NY
## [21] Ogdensburg, NY          Owensboro, KY
## [23] Paducah, KY             Pellston, MI
## [25] Plattsburgh, NY         Portsmouth, NH
## [27] Provincetown, MA        Provo, UT
## [29] Punta Gorda, FL         Rockland, ME
## [31] Sanford, FL             Sault Ste. Marie, MI
## [33] St. Augustine, FL       St. George, UT
## [35] Tupelo, MS
## 317 Levels: Aberdeen, SD Abilene, TX Albany, GA ... West Palm Beach/Palm Beach,
FL
```

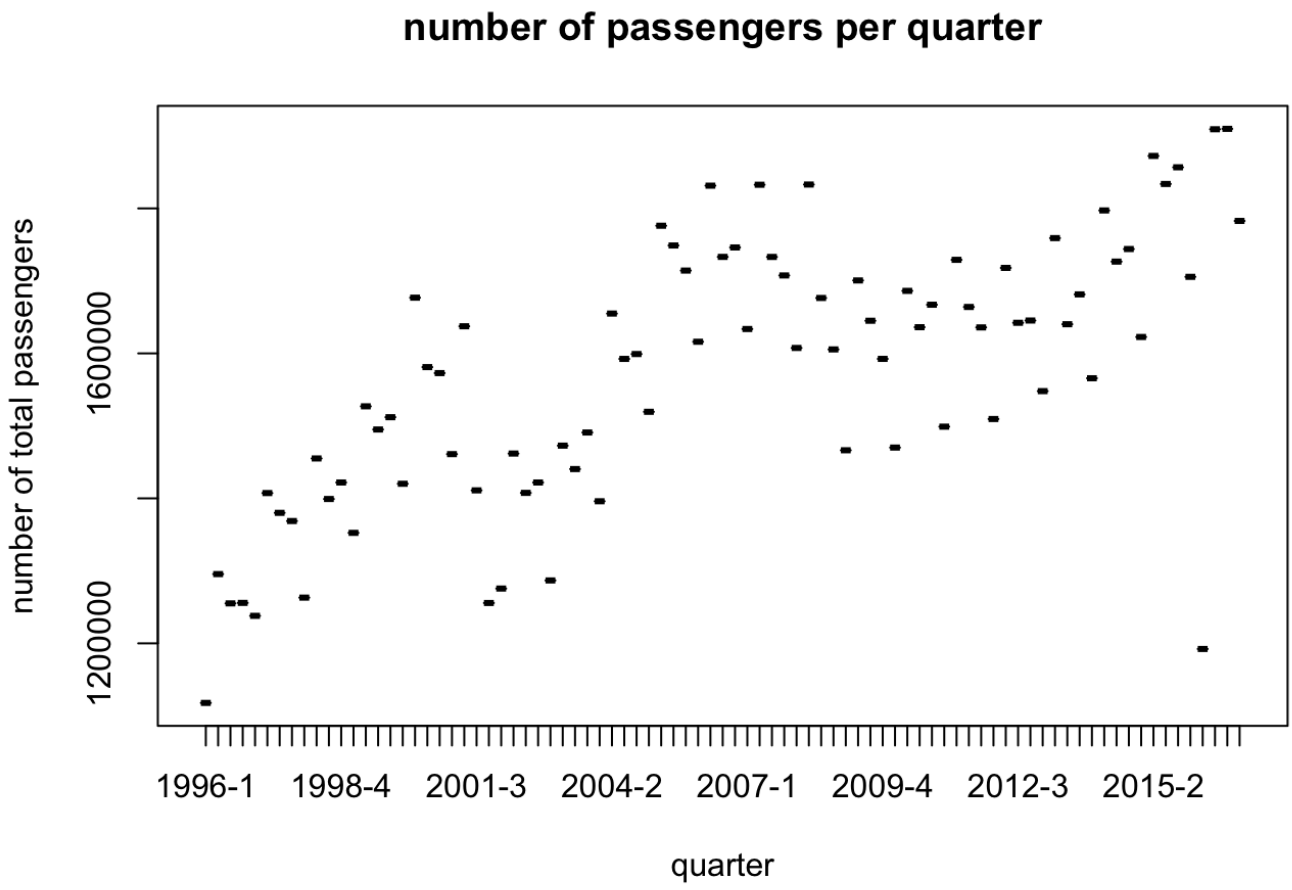
From the data of quarter 1 from 2007 and 1997, there are more cities with 0 connections from earlier years than in later years. This might be because data collection is more thorough and technology gets more advanced as time goes on. As mentioned, it makes sense that the cities with most connections are internationally well-known. It is interesting that Atlanta was on top 10 most connections for both 1997 2007, but not in 2017.

We need to keep in mind that this data only displays the first quarter for 2017, 2007, and 1997, as we only have the data for 2017 and we need to keep it consistent. Each year seems to display at the same general range of numbers. In Boston and Denver, the number of connections increased over these years, but the number of connections for other cities seem to stay at around the same numbers. This might be due to how much the city of Boston and Denver has grown in the past years. For example, I know many people who regard Boston as the “next Silicon Valley” because of all the business traffic that has been happening there for the past years. Similar observations are made in Denver.

4.

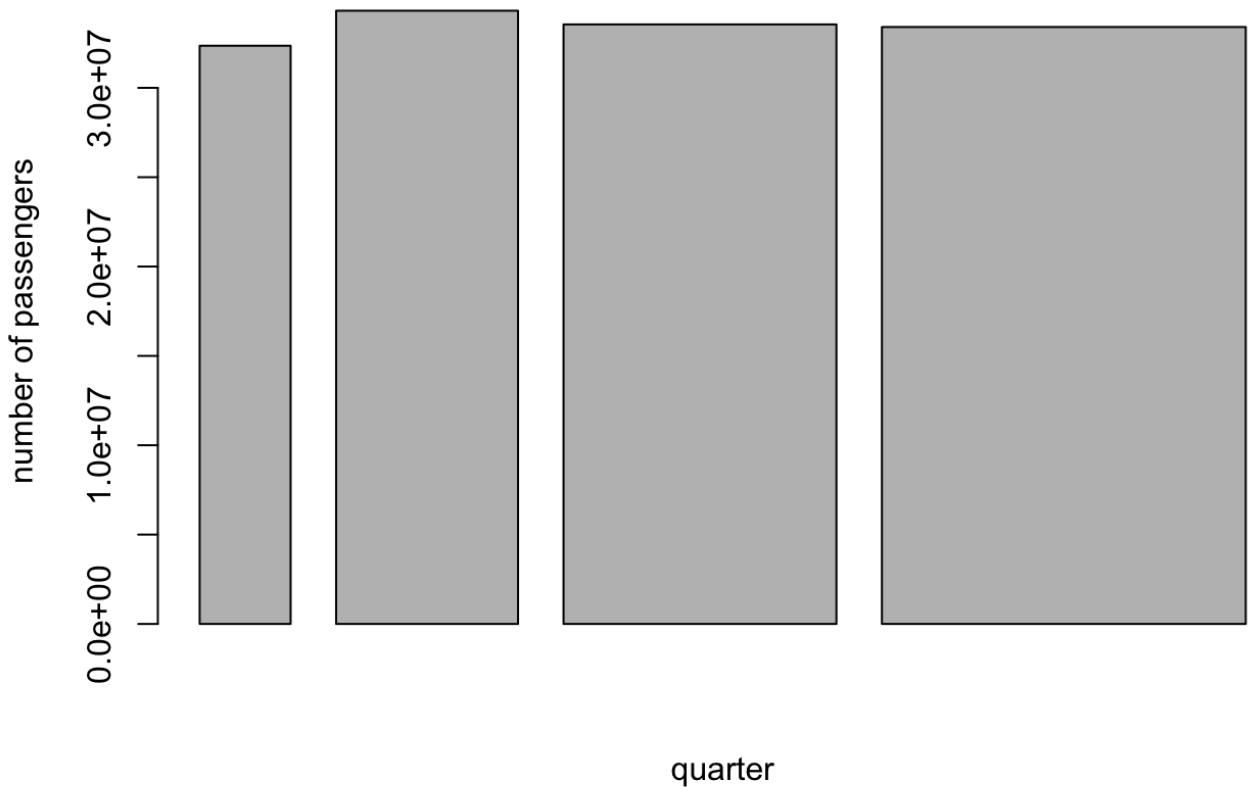
How has the approximate number of total passengers per quarter changed over the years? Create a graphic to show this and comment on patterns you see. Some quarters have a sharp decline in number of total passengers. What might explain these?





Generally, it seems like the number of passengers increase overtime per quarter over the years. However, there are some quarters that have a decline in number of total passengers. This can be explained by the different seasons. For example, there might be more passengers during the holiday seasons but less during

the working seasons. Additionally, there may be more passengers during the summer holidays. I thought it was interesting to see how quarter 1 is significantly less tha

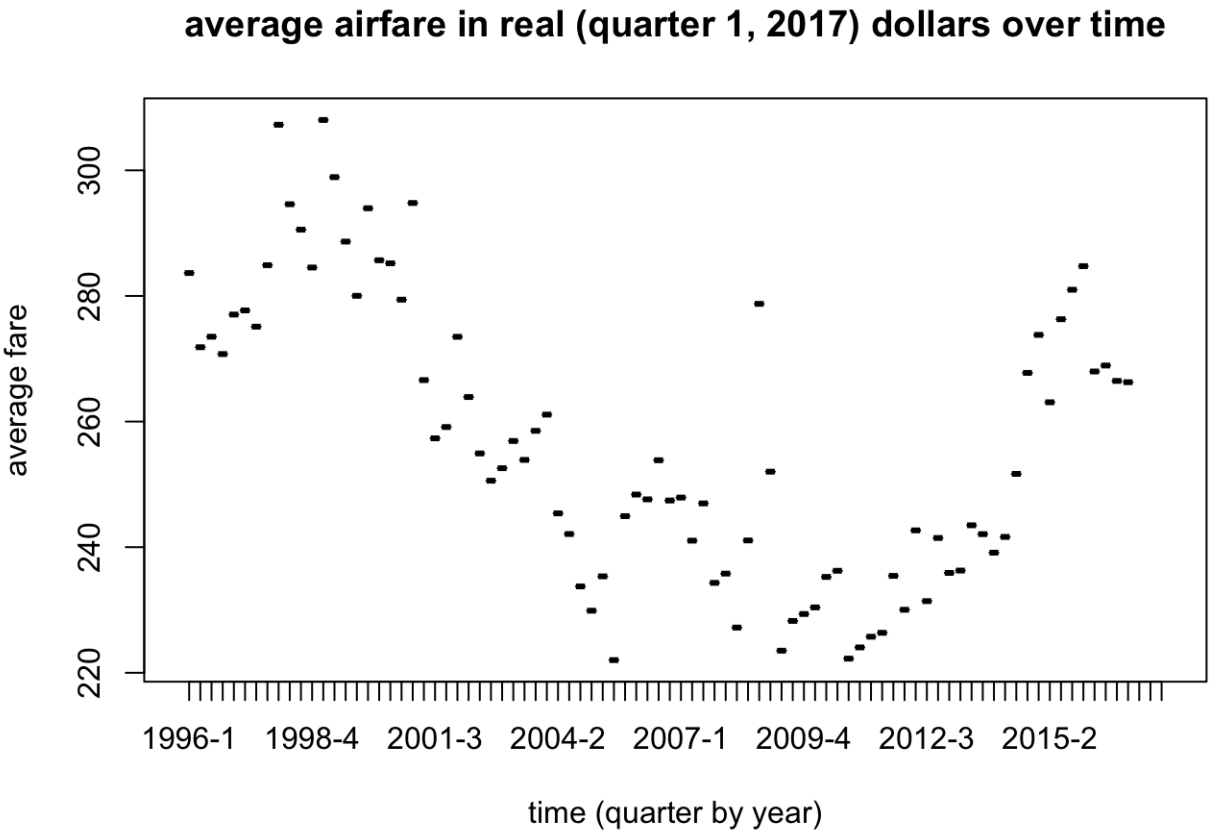


5. The average fares in the dataset are in nominal dollars (the actual price in dollars at the time). Inflation can confound conclusions based on nominal dollars over time. To deal with this, statisticians convert nominal dollars to real dollars. The conversion formula is explained at the end of this document. Load the CPI dataset. Create a new column real17_fare in the table 6 airfare dataset that has the average fare converted to real Q1 2017 dollars. You don't need to write an answer for this question, but please mark the code for this question in the appendix.

```
## Warning: package 'readxl' was built under R version 3.3.2
```

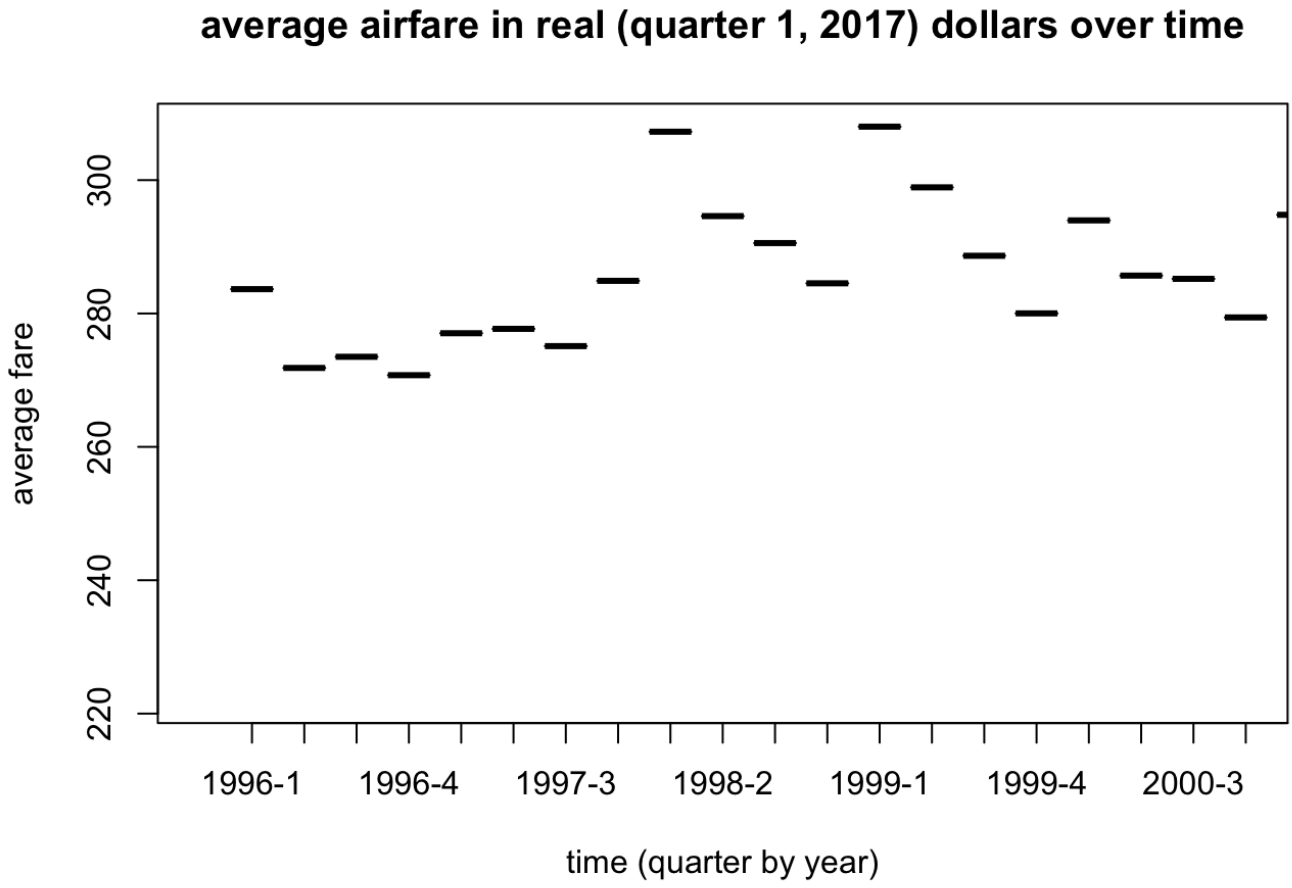
Answer is attached in the code. The real17_fare is compared to the most recent nominal dollar updated (year 2017, quarter 3).

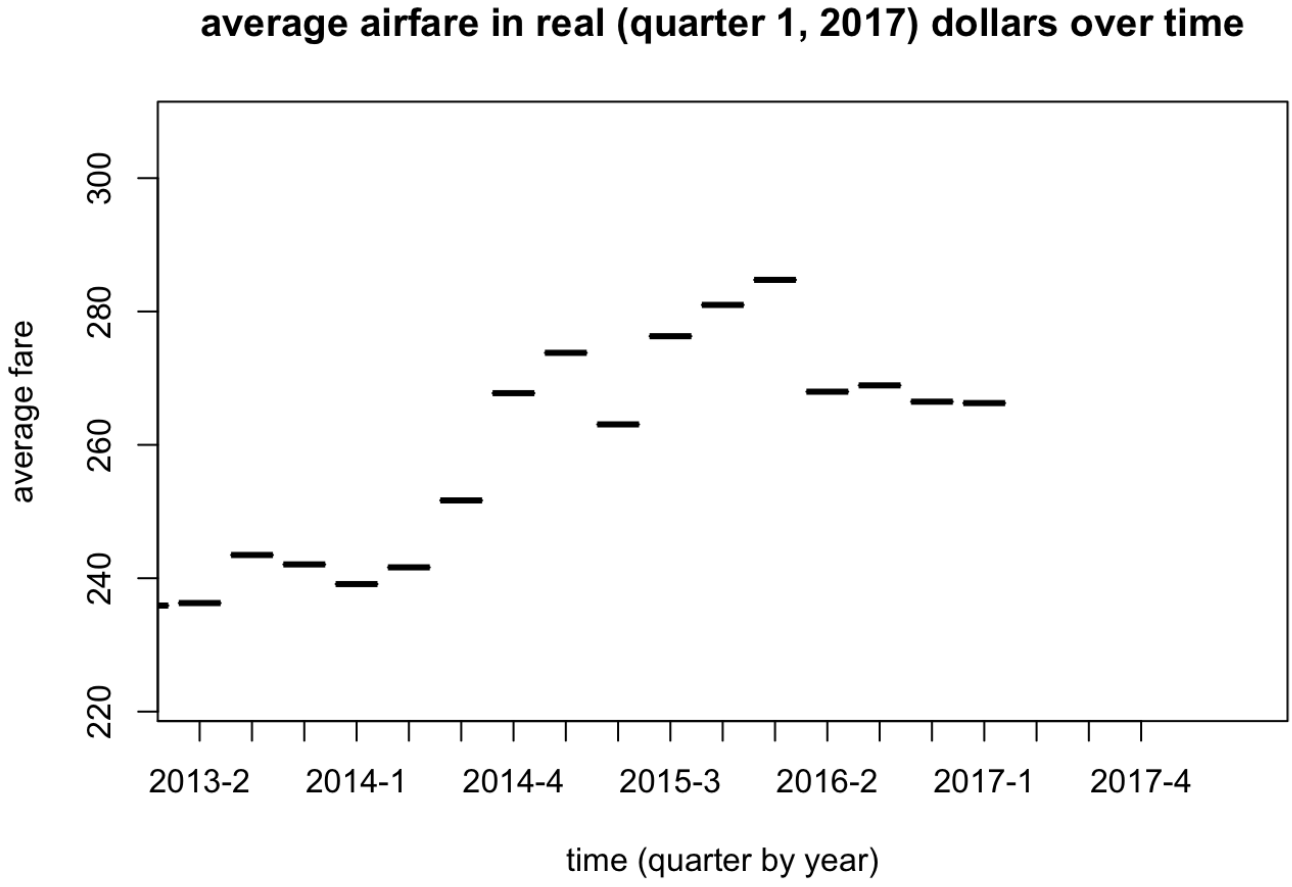
6. How have airfares changed over time? Use fares in real Q1 2017 dollars to investigate this graphically. Comment on patterns you see.



Now that the dollar amount is adjusted, we can see that overtime, average fare seemed to have fluctuated. The airfare was very high in the earlier years and kept decreasing until around 2010. This might not be a coincidence, and I believe that this might have to do with the 2008 recession. Airfare prices might have been at an all-time low because demand might have been low. After this time, however, the prices of airfare seem to have increased again as the economy improved.

To make more sense of how airfare changed by quarter, I zoomed into quarters:

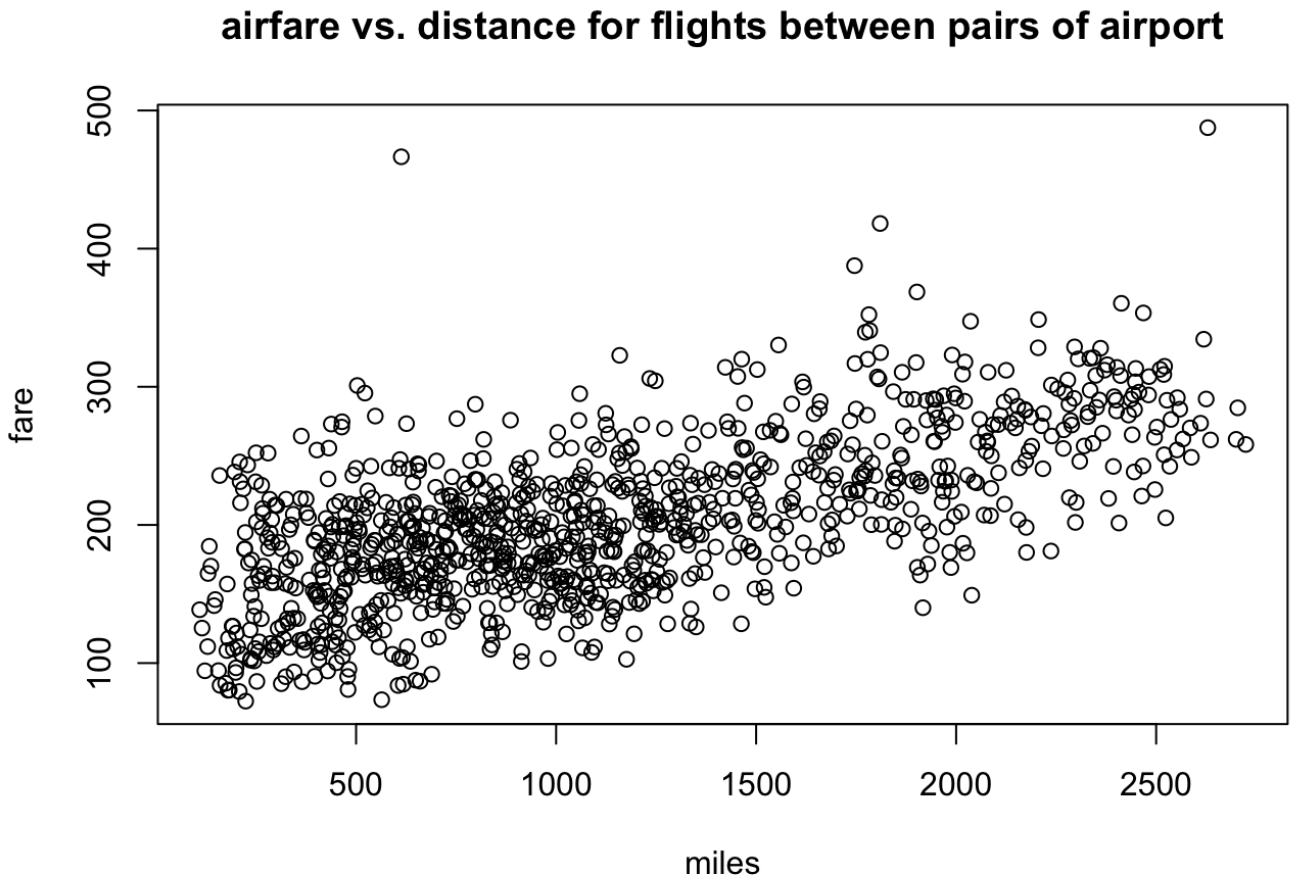


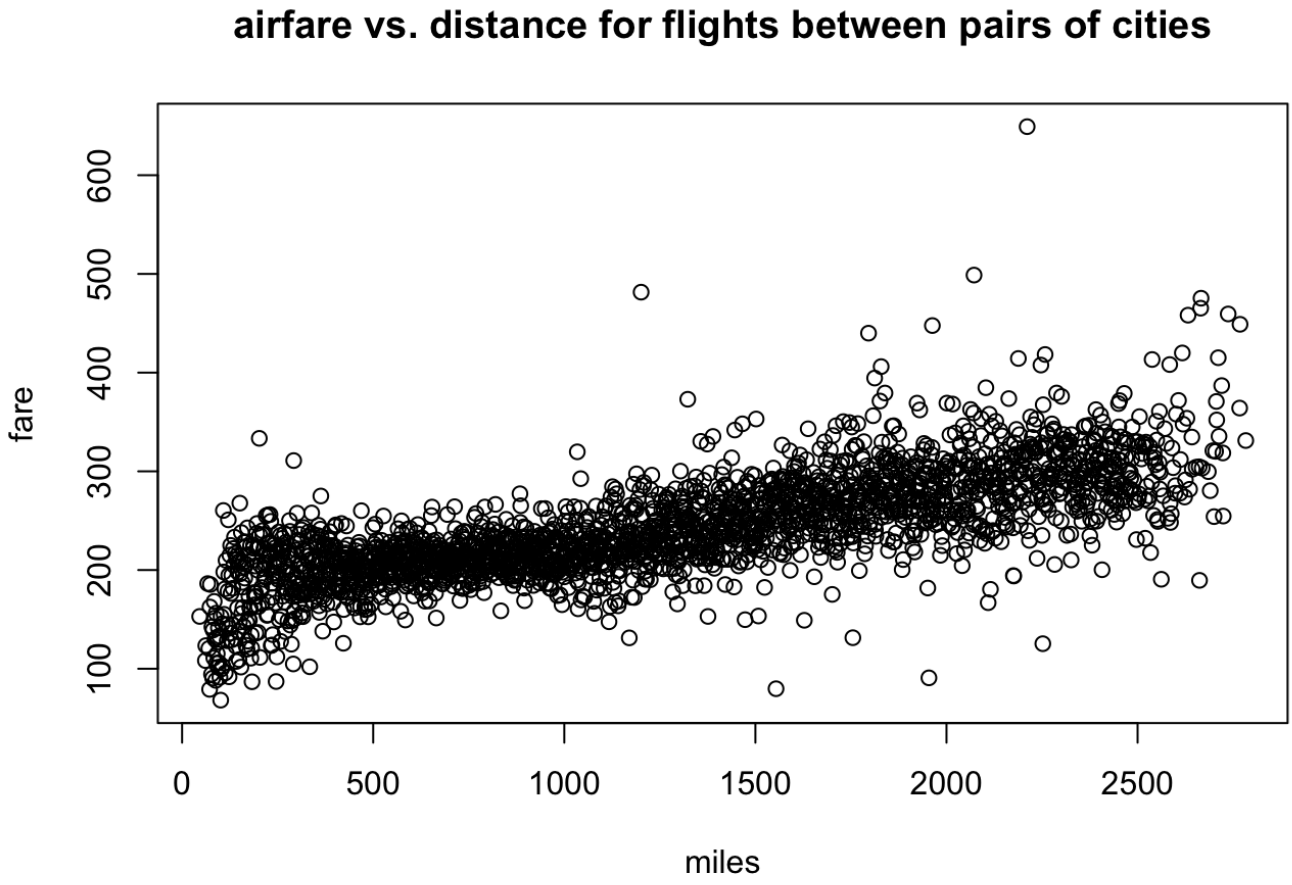


From the illustrations, it seems like quarter 1 usually has the highest airfare. However, this tends to fluctuate and it is difficult to make a proper conclusion, because it only happens during certain years. This might be

due to the holiday season. When the airfare decreases after the first quarter, I hypothesize that this might be because it is working season.

7. For 2015, what is the relationship between fare and distance? Use table 1a to investigate this visually and by using an appropriate statistical model or test. Comment on what you can infer from each and whether there is any disagreement. State the assumptions, use diagnostics to check whether they hold, and comment on how this affects your conclusions. Repeat your analysis with table 6. Comment on differences between your two results and why these occur.



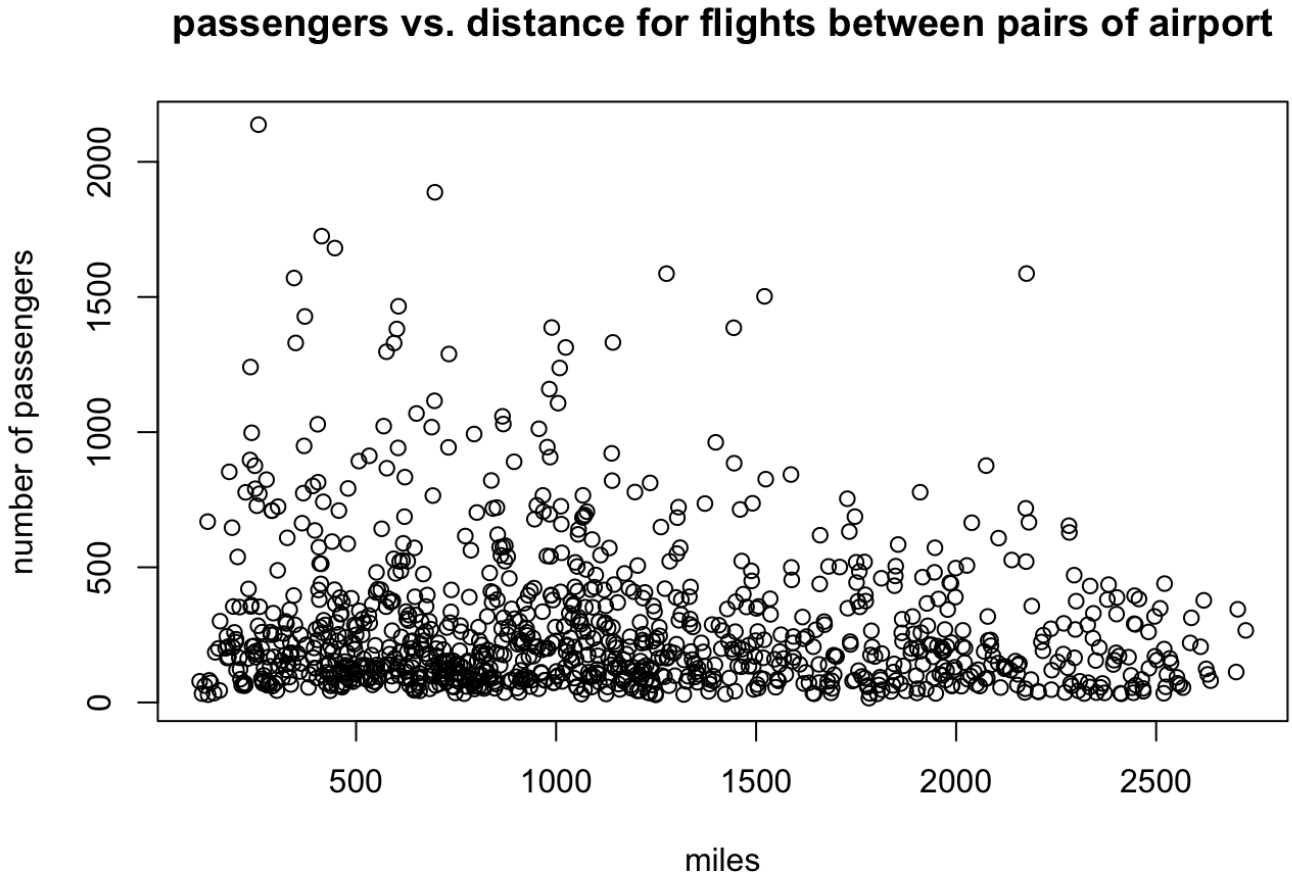


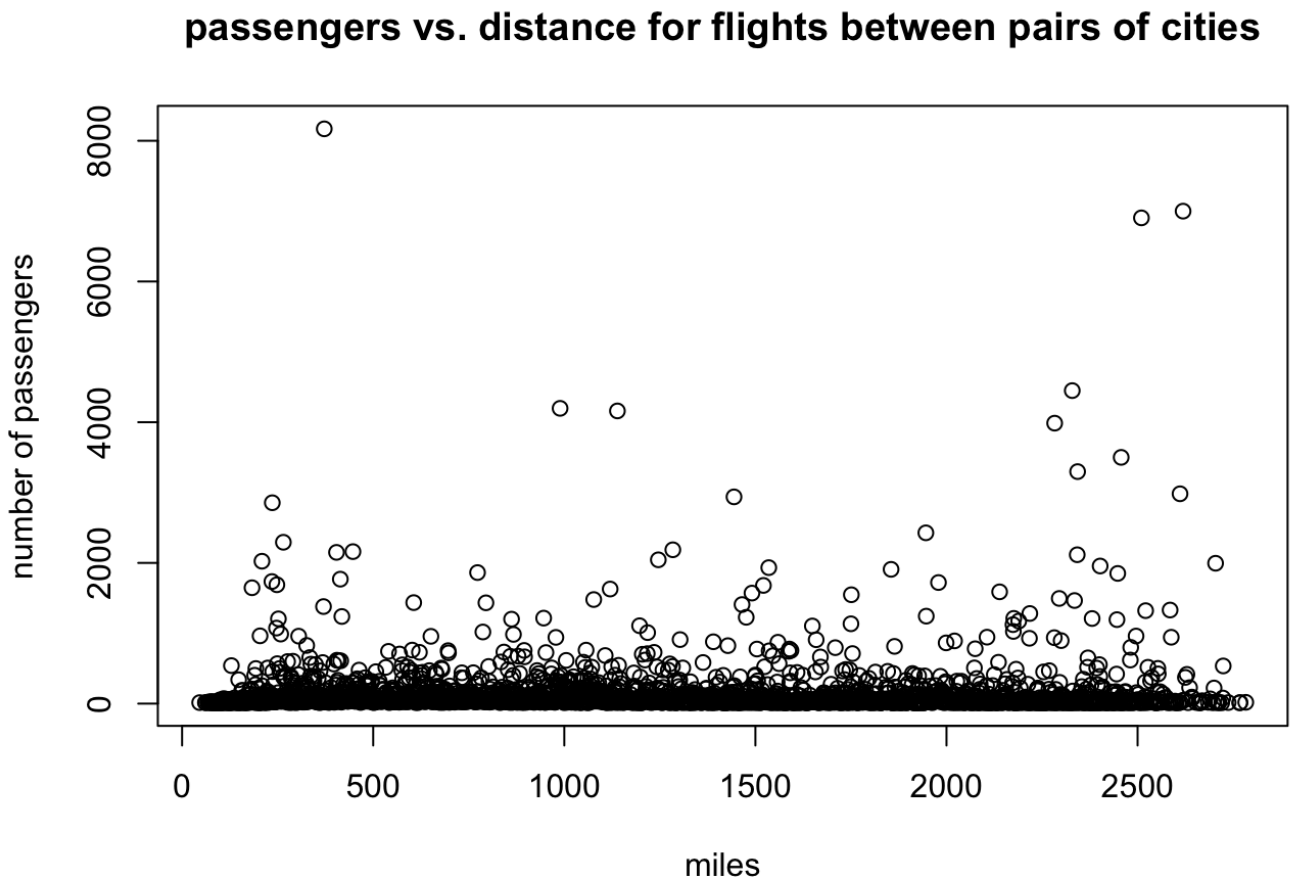
Using linear regression on table 1a and table 6, respectively:

```
##
## Call:
## lm(formula = aggregate(table1a$fare ~ table1a$miles, table1a,
##     FUN = mean))
##
## Coefficients:
##    (Intercept)  `table1a$fare`
##       -325.380           7.171
```

```
##
## Call:
## lm(formula = aggregate(table6$fare ~ table6$miles, table6, FUN = mean))
##
## Coefficients:
##   (Intercept)  `table6$fare`
##     -1124.724          9.964
```

From the figure, we can see that airfare tends to increase as distance traveled for the light increases for both flights between pairs of cities and pairs of airports (table 6 and table 1a, respectively). Similarly, 8. Modify the model or test you used in the previous question to consider average daily passengers in addition to distance. Recheck the assumptions and then comment on what you can infer about the relationship between fare and average daily passengers. Is there any difference between the results for table 1a and table 6?





Using linear regression on table 1a and table 6, respectively:

```
##
## Call:
## lm(formula = aggregate(table1a$passengers ~ table1a$miles, table1a,
##     FUN = mean))
##
## Coefficients:
##          (Intercept)  `table1a$passengers`
##            1210.0620               -0.2948
```

```
##
## Call:
## lm(formula = aggregate(table6$passengers ~ table6$miles, table6,
##     FUN = mean))
##
## Coefficients:
##         (Intercept)  `table6$passengers`
##           1.283e+03            3.156e-02
```

Looiking at the graph, there does not seem to be a relationship between miles and number of passengers for either of the graphs.

9. For 2015, identify city pairs where the carrier with the largest market share has fares below the average for that city pair. Investigate these using graphics, statistics, or models (as you see fit). Comment on patterns you find.
10. Use table 1a to compare Sacramento (SMF), Oakland (OAK), San Francisco (SFO), and San Jose (SJC). How do fares differ between these airports? Which airport has the most long-distance connections and how does this compare to the others? Do these results differ by year?

Citations:

Looked on Piazza for tips

Consulted dicussion notes

Code Appendix

```
#1
data <- read.csv("~/Desktop/airfare.csv")
data$city_id1 <- as.factor(data$city_id1)
data$city_id1 <-as.factor(data$city_id2)
i = split(data, data$table)
table1a = i$`1a`
table6 = i$`6`
#2
table(data$year)
table(data$quarter)
data$NAs <- apply(is.na(data), 1, sum)
table1a$NAs <- apply(is.na(table1a), 1, sum)
table6$NAs <- apply(is.na(table6), 1, sum)

sevenNAs1ayear <- table(table1a$year[table6$NAs > 6])
barplot(sevenNAs1ayear,xlab= "year", ylab = "frequency", main ="flights between pai
rs of airports with missing data")
sevenNAs1aquarter<-table(table1a$quarter[table6$NAs > 6])
barplot(sevenNAs1aquarter,xlab= "quarter", ylab = "frequency", main ="flights betwe
en pairs of airports with missing data")
sevenNAs6year <- table(table6$year[table6$NAs > 6])
barplot(sevenNAs6year,xlab= "year", ylab = "frequency", main ="flights between pair
s of cities with missing data")
sevenNAs6quarter <- table(table6$quarter[table6$NAs > 6])
barplot(sevenNAs6quarter,xlab= "quarter", ylab = "frequency", main ="flights betwee
n pairs of cities with missing data")
#3
cities12017<- table(table6$city1[table6$year %in% "2017"])
cities22017 <-table(table6$city2[table6$year %in% "2017"])
mergedCities2017 <- merge(cities12017,cities22017, by="Var1")
mergedCities2017$connections <- rowSums(mergedCities2017[, c("Freq.x", "Freq.y")],
na.rm=T)
mostConnections2017 <- sort(mergedCities2017$connections, decreasing = T)[1:10]
mergedCities2017$Var1[mergedCities2017$connections %in% mostConnections2017]
leastConnections2017 <- sort(mergedCities2017$connections, decreasing = F)
leastConnections2017 <- mergedCities2017$Var1[mergedCities2017$connections == 0]
leastConnections2017
cities12007<- table(table6$city1[table6$year %in% "2007" & table6$quarter == 1])
cities22007 <-table(table6$city2[table6$year %in% "2007" & table6$quarter == 1])
mergedCities2007 <- merge(cities12007,cities22007, by="Var1")
mergedCities2007$connections <- rowSums(mergedCities2007[, c("Freq.x", "Freq.y")],
na.rm=T)
mostConnections2007 <- sort(mergedCities2007$connections, decreasing = T)[1:10]
mergedCities2007$Var1[mergedCities2007$connections %in% mostConnections2007]
leastConnections2007 <- sort(mergedCities2007$connections, decreasing = F)
leastConnections2007 <- mergedCities2007$Var1[mergedCities2007$connections == 0]
leastConnections2017
cities11997<- table(table6$city1[table6$year %in% "1997" & table6$quarter == 1])
cities21997 <-table(table6$city2[table6$year %in% "1997" & table6$quarter == 1])
mergedCities1997 <- merge(cities11997,cities21997, by="Var1")
mergedCities1997$connections <- rowSums(mergedCities1997[, c("Freq.x", "Freq.y")],
na.rm=T)
mostConnections1997 <- sort(mergedCities1997$connections, decreasing = T)[1:10]
mergedCities1997$Var1[mergedCities1997$connections %in% mostConnections1997]
```

```
leastConnections1997 <- sort(mergedCities1997$connections, decreasing = F)
leastConnections1997 <- mergedCities1997$Var1[mergedCities1997$connections == 0]
leastConnections1997
#4
tab1a <- aggregate(table1a$passengers ~ table1a$quarter+table1a$year, table1a, FUN
= sum)
tab <- aggregate(data$passengers ~ data$quarter+data$year, data, FUN = sum)
pass <- aggregate(data$passengers ~ data$quarter, data, FUN = sum)
barplot(pass$`data$passengers`, pass$`data$quarter`, xlab = "quarter", ylab = "numb
er of passengers")
data$quarterbyyear <- paste(data$year,data$quarter,sep="-")
data$quarterbyyear <- as.factor(data$quarterbyyear)
passengerByQ <- aggregate(data$passengers ~ data$quarterbyyear, data, FUN = sum, na
.rm=T)
plot(passengerByQ, type="p", xlab = "quarter", ylab = "number of total passengers",
main ="number of passengers per quarter")

pass <- aggregate(data$passengers ~ data$quarter, data, FUN = sum)
barplot(pass$`data$passengers`, pass$`data$quarter`, xlab = "quarter", ylab = "numb
er of passengers")
#5
library(readxl)
cpi <- read_xlsx("~/Desktop/cpi_1996_2017.xlsx")
cpi <-cpi[12:33,0:13] #gets rid of unnecessary data
cpi <- data.frame(cpi)

cpi <- cpi[,seq(4,13,3)] #getting 4 numbers, one for each quarter
frame1 <- data.frame(year=rep(c(1996:2017),times=4), quarter=rep(c(1:4),each=22),
cpi = stack(cpi))
new <- merge(frame1,table6, by.x=c("year","quarter"), all=TRUE)
new$cpi.values<- as.numeric(new$cpi.values)
new$real17_fare <- new$lg_fare * (new$cpi.values[new$year == 2017 & new$quarter ==
3]/new$cpi.values)
#6
new$real17_fareq1 <- new$lg_fare * (200.091/new$cpi.values)
new$quarterbyyear <- paste(new$year,new$quarter,sep="-")
new$quarterbyyear <- as.factor(new$quarterbyyear)
plot(aggregate(new$real17_fareq1 ~ new$quarterbyyear, new, FUN = mean), xlab="time
(quarter by year)", ylab = "average fare", main="average airfare in real (quarter 1
, 2017) dollars over time")
plot(aggregate(new$real17_fareq1 ~ new$quarterbyyear, new, FUN = mean), type = "l",
xlab="time (quarter by year)", ylab = "average fare", main="average airfare in real
(quarter 1, 2017) dollars over time", xlim=c(0,20))
plot(aggregate(new$real17_fareq1 ~ new$quarterbyyear, new, FUN = mean), type = "l",
xlab="time (quarter by year)", ylab = "average fare", main="average airfare in real
(quarter 1, 2017) dollars over time", xlim=c(70,90))
#7
fare1a <- aggregate(table1a$fare ~ table1a$miles, table1a, FUN=mean)
plot(fare1a, xlab = "miles", ylab="fare", main="airfare vs. distance for flights be
tween pairs of airport")
plot(aggregate(table6$fare ~ table6$miles, table6, FUN=mean), xlab = "miles", ylab
="fare", main="airfare vs. distance for flights between pairs of cities")
lm(aggregate(table1a$fare ~ table1a$miles, table1a, FUN=mean))
lm(aggregate(table6$fare ~ table6$miles, table6, FUN=mean))
#8
plot(aggregate(table1a$passengers ~ table1a$miles, table1a, FUN=mean), xlab = "mil
es", ylab="number of passengers", main="passengers vs. distance for flights between
```

```
pairs of airport")
plot(aggregate(table6$passengers ~ table6$miles, table6, FUN=mean), xlab = "miles"
, ylab="number of passengers", main="passengers vs. distance for flights between pa
irs of cities")
lm(aggregate(table1a$passengers ~ table1a$miles, table1a, FUN=mean))
lm(aggregate(table6$passengers ~ table6$miles, table6, FUN=mean))
```