Flying Home for Christmas

Author: Megan Morgan December 17, 2021

Abstract

Christmas is a popular holiday celebrated worldwide, and regardless of your family's traditions or customs, the common thread is spending the day of, and days around, Christmas with friends and loved ones. Traveling is common, whether by cart, train, or airplane; but with all the recent news reports about flights being delayed, cancelled, or pilots going on strike can you count on making it to your destination before the dinner table is set?

There are several different types of delays. Delay categories include the airline carrier delay, national aviation system delay, airport security delay, previous flight late arrival delay, and inclement weather delay. Over one million flights were analyzed for the study. Nationwide, flights from November 1, 2020 to June 30, 2021 arrived no more than 15 minutes behind the expected arrival time almost 95% of the time. In addition, the author living in the Dallas-Fort Worth metroplex means the origin airports available are Dallas Love Field (DAL) and Dallas/Fort Worth International (DFW). Both of which follow suit with the arrival statistics; but when they are delayed have a mean delay time of under an hour.

Introduction

Originally sung by Bing Crosby in 1943, "I'll be home for Christmas; you can plan on me" (Congress, 2002), this wishful prayer is the yearly goal for many Americans come each December. However, by the time the schools let out for winter break and the suitcases are packed with wrapped presents, most of America has one of two options: start driving and pray

you do not run out of gas; or hope the suitcase is packed just right and that the flight was not overbooked. Either option seems equally unappealing, so which is the better, safer, faster, cheaper, and less headache of a choice to pick?

Although the song is almost seventy years old, Perry Como sung it best (original recording in 1954) and the statements are still as true as ever:

"Oh there's no place like home for the holidays,

`Cause no matter how far away you roam,

…For the holidays, you can't beat home, sweet home.

…Take a bus, take a train. Go and hop an aeroplane

…No trip could be too far

… From Atlantic to Pacific, Gee, the traffic is terrific"

(Allen & Stillman, 1954)

That leaves us with flying. Air travel "for average or high-risk drivers, it is always safer to fly then drive. Furthermore, even for a low-risk driver, nonstop flying is safer than driving on rural interstate highways for a trip distance of more than 303 miles" (Sivak, Weintraub, & Flannagan, 1991). And that is on a normal day before everyone and their dog is on the road and in a rush to reach their destination, all while fighting exhaustion and snowfall (if you want a stereotypical white Christmas).

For those that happen to live near one, you can probably name majority to all of the top five American airports. The Federal Aviation Administration (FAA) is the United States Government's governing body for American air travel, and according to the number of

passengers boarded per flight the “Core-5 Mega Hub” Airports (in 2020) listed from first to fifth were: Hartsfield-Jackson Atlanta International (ATL); Dallas/Fort Worth International (DFW); Denver International (DEN); Chicago O’Hare International (ORD); and Los Angeles International (LAX) (United States Department of Transportation, 2021). This list, although having changed in order has been the Core-5 for over the last decade; and except for Denver replacing LaGuardia’s (LGA), has been the same over thirty years.

There are many more airports in the United States besides these five, and a person can travel from origin to destination without having to connect through one of these airports either. Texas, the second largest and second most populated state in the country, alone has 5, 10 percent, of the country’s top 50 busiest airports (U.S. Department of Commerce, 2019). These airports are Dallas/Fort Worth International (DFW), ranked 2nd; George Bush Intercontinental/Houston (IAH), ranked 12th; Dallas Love Field (DAL), ranked 29th; William P. Hobby (HOU), ranked 33rd; and San Antonio International (SAT), ranked 46th. According to my GPS, DFW and DAL are only 17 miles apart, and if because of this proximity they were considered the same airport, the number of travelers would bump Atlanta out of the top spot. Even after pulling the other airports close to the remaining top 5 (only O’Hare was updated as it merged with Midway), Dallas-Fort Worth, Texas would remain number one; having almost 10 percent more traffic than Atlanta, Gorgia.

“Air traffic delays are receiving considerable attention in the United States and have become one of the most common reasons for passenger complaints” (U.S. Department of Transportation, 1998–2012). With flight delays becoming more common, “on average, more than one fifth of all domestic flights arrived at least 15 min behind schedule and were therefore

designated as delayed" (Baumgarten, 2014). These delays have several sources; some being customer related, others airline or airport related, and others just impacted by the weather. Officially these delay categories are aircraft arriving late, from previous leg; National Aviation System, or NAS; airline crew or maintenance; extreme weather, blizzard, hurricane, temperature, etc.; and security issues.

So, if someone was to leave the Dallas-Fort Worth Texas metroplex and try to make it home for Christmas; assuming they were first willing to spend the "travel premiums ranging from 41.6% to 82.0%" above the usual ticket price; should their relatives set a place at the table and count on them being home for Christmas or is it "only in [their] dreams" (Luttmann & Gaggero, 2021)?

Data

Using data presented by the Federal Aviation Administration (FAA) on behalf of the United States Government. The original data is collected from November 01, 2020 to June 30, 2021 and consists of 1,048,576 flights and records datapoints. This data consists of 1,048,576 flights and records the following items of information (U.S. Department of Transportation Bureau of Transportation Statistics, 2021).

Field	Description	Field Type
Dest	Destination airport call code	String 3-char
Dest_AP	Full name of destination airport	String
Mkt Carrier Network	Airline network name	String
Origin	Origin airport call code	String 3-char
Origin_AP	Full name of origin airport	String
carrier type	Short airline name	String
Code Table Share	Abbreviated airline name	String
Dest Display Airport Name Full	Destination city and airport name	String
Fl Date	Date of scheduled flight departure	Date
Mkt Carrier	Type of aircraft	String 2-char
Origin Display Airport Name Full	Origin city and airport name	String
Arr Del15	Is the flight 15+ minutes delayed	Boolean
Arr Delay	Flight delay time (negative means early)	Integer
Cancelled	Is the flight cancelled	Boolean
Carrier Delay	Delay time from the airline	Integer
Diverted	Is the flight diverted to a different destination	Boolean
Late Aircraft Delay	Time Delay Caused by the airplane arriving late to the origin	Integer
max date	Destination arrival date	Date
Nas Delay	Time Delay Caused by airport Nav system Issues	Integer
Security Delay	Time Delay Caused by airport security issues	Integer
Weather Delay	Time Delay Caused by weather	Integer

Table 1: Explanation of Data Fields Available in Data Source

Due to limitations with R-studio (this large dataset frequently maxed out R's data display abilities), many of the calculations and reports below will use a subset of this data. To narrow the data, details looking at flights out of the Dallas-Fort Worth metroplex (airports Dallas Love Field, DAL, and Dallas/Fort Worth International, DFW) for the 15 days prior to Christmas 2020.

This sub-dataset dataset used is comprised of 5,152 rows of flights information. Data includes flights to or from almost 3900 airports in the United States of America. Why Dallas-Fort Worth? Two reasons, this is the current residence of the author and as shown in Table 3 DFW airport is the second busiest airport in the county and if combined with DAL airport as shown in Table 4 the metroplex becomes the busiest airport in the county.

The airports included in either the Origin or Destination fields are (in alphabetical order).

ABE	AZA	BUR	CVG	ERI	GRR	IND	LGB	MLU	PBI	RDU	SHD	TRI
ABI	AZO	BWI	CWA	ESC	GSO	INL	LIH	MOB	PDX	RFD	SHR	TTN
ABQ	BDL	BZN	CYS	EUG	GSP	ISP	LIT	MOT	PGD	RHI	SHV	TUL
ABR	BET	CAE	DAB	EVV	GTF	ITH	LNK	MQT	PHF	RIC	SIT	TUS
ABY	BFF	CAK	DAL	EWN	GTR	ITO	LNY	MRY	PHL	RIW	SJC	TVC
ACK	BFL	CDB	DAY	EWR	GUC	JAC	LRD	MSN	PHX	RKS	SJT	TWF
ACT	BGM	CDC	DBQ	EYW	GUM	JAN	LSE	MSO	PIA	RNO	SJU	TXK
ACV	BGR	CDV	DCA	FAI	HDN	JAX	LWB	MSP	PIB	ROA	SLC	TYR
ACY	BHM	CGI	DDC	FAR	HGR	JFK	LWS	MSY	PIE	ROC	SLN	TYS
ADK	BIL	CHA	DEC	FAT	HHH	JLN	LYH	MTJ	PIH	ROW	SMF	USA
ADQ	BIS	CHO	DEN	FAY	HIB	JMS	MAF	MVY	PIR	RST	SMX	VCT
AEX	BJI	CHS	DFW	FCA	HLN	JNU	MBS	MYR	PIT	RSW	SNA	VEL
AGS	BKG	CID	DHN	FLG	HNL	JST	MCI	OAJ	PLN	SAF	SPI	VLD
AKN	BLI	CIU	DIK	FLL	HOB	KOA	MCO	OAK	PNS	SAN	SPN	VPS
ALB	BLV	CKB	DLG	FNT	HOU	KTN	MDT	OGD	PQI	SAT	SPS	WRG
ALO	BMI	CLE	DLH	FSD	HPN	LAN	MDW	OGG	PRC	SAV	SRQ	XNA
ALS	BNA	CLL	DRO	FSM	HRL	LAR	MEI	OGS	PSC	SBA	STC	XWA
ALW	BOI	CLT	DRT	FWA	HSV	LAS	MEM	OKC	PSE	SBN	STL	YAK
AMA	BOS	CMH	DSM	GCC	HTS	LAW	MFE	OMA	PSG	SBP	STS	YKM
ANC	BPT	CMI	DTW	GCK	HVN	LAX	MFR	OME	PSM	SBY	STT	YUM
APN	BQK	CMX	DVL	GEG	HYS	LBB	MGM	ONT	PSP	SCC	STX	
ART	BQN	CNY	EAR	GFK	IAD	LBE	MHK	ORD	PUB	SCE	SUN	
ASE	BRD	COD	EAT	GGG	IAG	LBF	MHT	ORF	PUW	SCK	SUX	
ATL	BRO	COS	EAU	GJT	IAH	LBL	MIA	OTH	PVD	SDF	SWF	
ATW	BRW	COU	ECP	GNV	ICT	LCH	MKE	OTZ	PVU	SEA	SWO	
ATY	BTM	CPR	EGE	GPT	IDA	LCK	MKG	OWB	PWM	SFB	SYR	
AUS	BTR	CRP	EKO	GRB	ILG	LEX	MKK	PAE	RAP	SFO	TLH	
AVL	BTV	CRW	ELM	GRI	ILM	LFT	MLB	PAH	RDD	SGF	TOL	
AVP	BUF	CSG	ELP	GRK	IMT	LGA	MLI	PBG	RDM	SGU	TPA	

Table 2: List of Airports Included in Data

Results and Analysis

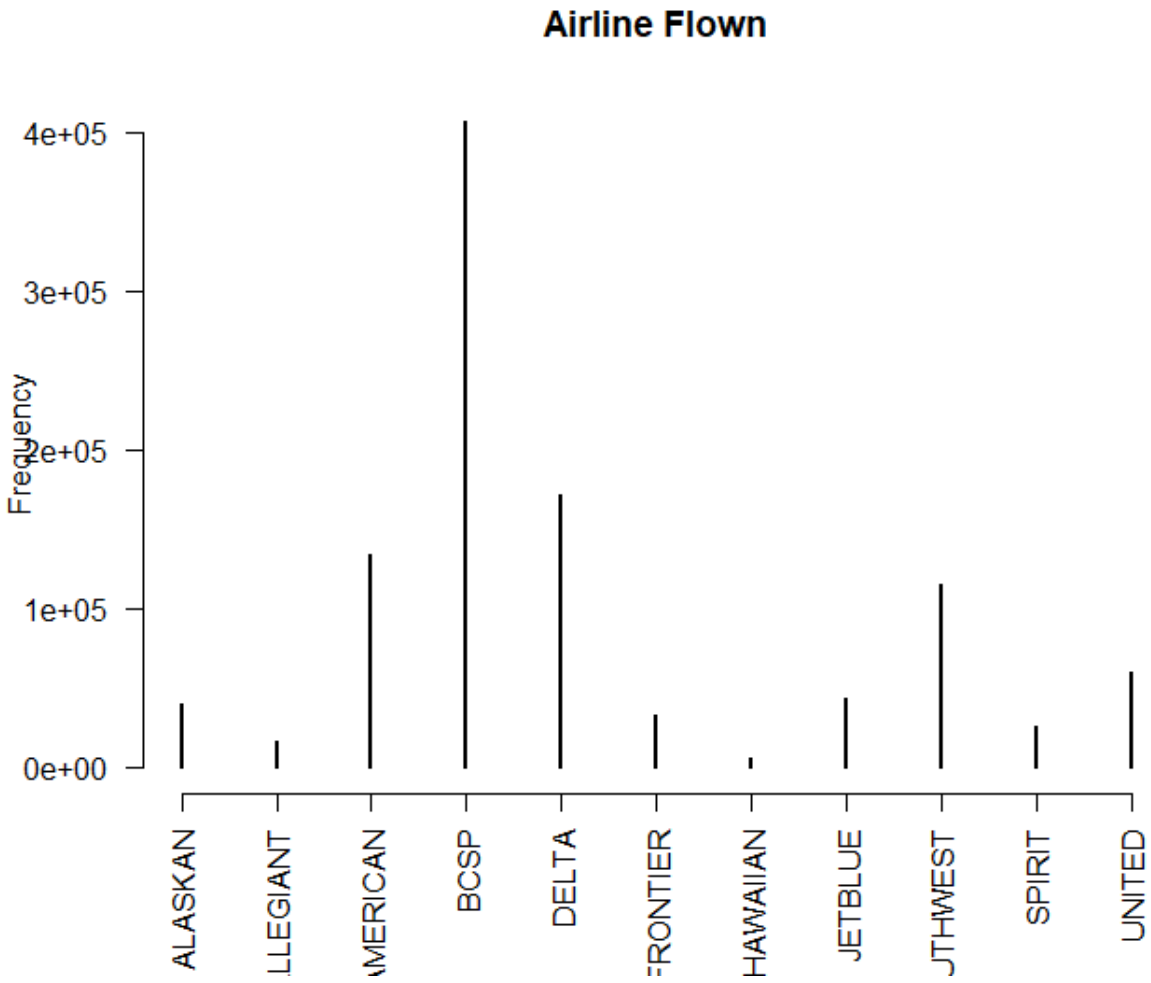


Figure 1: Barplot of Airline Chosen for Flight

Most of the flights are from subsidiary airlines and not the other major companies shown above. The data covers flights from 2020-2021 and due to the nature of what's going on in the world many of the major airlines had shutdown.

Airport	Code	2010		2019		2020	
		Rank Metroplex Rank	Total Enplaned Passengers	Rank Metroplex Rank	Total Enplaned Passengers	Rank Metroplex Rank	Total Enplaned Passengers
Hartsfield-Jackson Atlanta International	ATL	1 1	43,132,110	1 1	53,505,357	1 2	20,559,853
Dallas/Fort Worth International DFW & Dallas Love Field	DFW DFW/DAL	4 3	27,100,907 30,883,885	4 3	35,785,318 43,864,700	2 1	18,595,214 22,265,121
Denver International	DEN	5 5	25,242,113	5 5	33,592,591	3 4	16,243,216
Chicago O'Hare International ORD & Chicago Midway International	ORD ORD/MDW	2 2	32,172,478 40,690,522	3 2	40,887,890 50,968,537	4 3	14,613,209 18,849,793
Los Angeles International	LAX	3 4	28,856,870	2 4	42,965,654	5 5	14,057,650

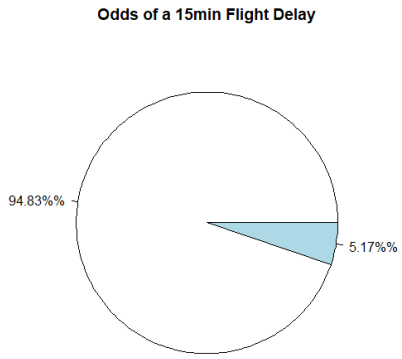
Table 3: Top 5 Airports in the United States

The table above shows the top 5 airports in the United States from 2010 to 2020. DFW and ORD show a second line of values; these values are the result of merging the main airport with airports within 20 miles. This provides the context of how busy the metroplex is, and as a result, Dallas/Fort Worth is then busier than Atlanta (2020).

Airport	Code	2010		2019		2020	
		Rank	Total Enplaned Passengers	Rank	Total Enplaned Passengers	Rank	Total Enplaned Passengers
Dallas-Fort Worth, TX (Dallas/Fort Worth International)	DFW	4	27,100,907	4	35,785,318	2	18,595,214
Houston, TX (George Bush Intercontinental/Houston)	IAH	7	19,525,011	14	21,907,940	12	8,682,555
Dallas, TX (Dallas Love Field)	DAL	49	3,782,978	33	8,079,382	29	3,669,907
Houston, TX (William P Hobby)	HOU	40	4,357,508	36	7,068,929	33	3,127,156
San Antonio, TX (San Antonio International)	SAT	47	3,920,864	44	5,036,738	46	1,919,955

Table 4: Top 5 Airports in the Texas, USA

The table above shows the 5 busiest airports in the state of Texas. Two are located in the Dallas-Fort Worth metroplex, and another two are in Houston. This only accounts for the major airports in the area and not the small commercial or private airports.



The FAA only considers a flight as delayed if it is at least 15 minutes behind schedule. In the 1,048,575 flights traveled last year, only 5.17% of them were considered delayed, so the other 994,000 flights in the United States were on time (or even slightly early).

Figure 2: (left) Pie Chart, 5.17% of Flights have a 15+min delay

A flight is diverted for many reasons: aircraft issues, medical emergency, weather obstacle, or traffic jam forcing a landing for refueling to name a few. Thankfully less than a tenth of a percent of the flights that take off are diverted to other airports.

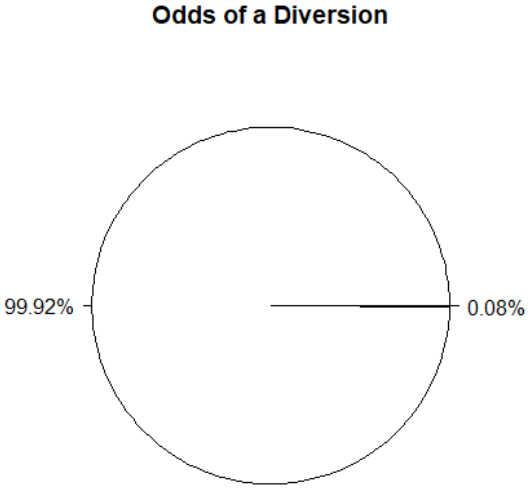


Figure 3: (right) Pie Chart, 0.08% of Flights are Diverted

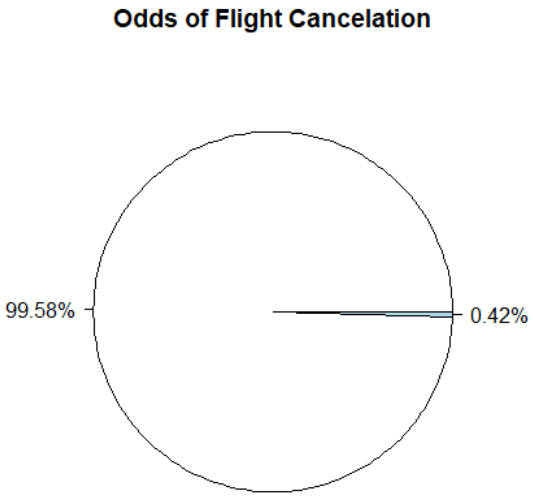


Figure 4:
(left) Pie Chart, 0.42% of Flights are Cancelled

The 0.42% of flights cancelled in 2020 only account for flights cancelled within an hour of the scheduled flight time. This value would be drastically higher for the year 2020 if it had included flights cancelled at any time. The recent surge in pilot strikes will increase this value for 2021. However, for the data provided it is reassuring to know that if you pack your bags and make it through TSA you can bet on the plane taking off.

When a flight is officially delayed, how delayed will you be? The stem and leaf plot shown below represents this time. The highest volume of delays is the first row of values which thankfully means that majority of the flight delays are less than 2,000 minutes (within a day and a half).

```
> #  Stem and leaf plot
> stem (arrivalDelayTime)

  The decimal point is 3 digit(s) to the right of the |

   0 | 0000000000000000000000000000000000000000000000000000000000000000000000000000000+714
   2 | 023357726
   4 | 285
   6 | 257269
   8 | 56479
  10 | 129
  12 | 70
  14 | 2368
  16 | 98
  18 | 28
  20 | 68
  22 | 1
  24 | 1458
  26 | 48
  28 | 138
  30 | 829
  32 | 325
  34 | 25679
```

Figure 5: Stem and Leaf Plot of Arrival Delay Times (mins)

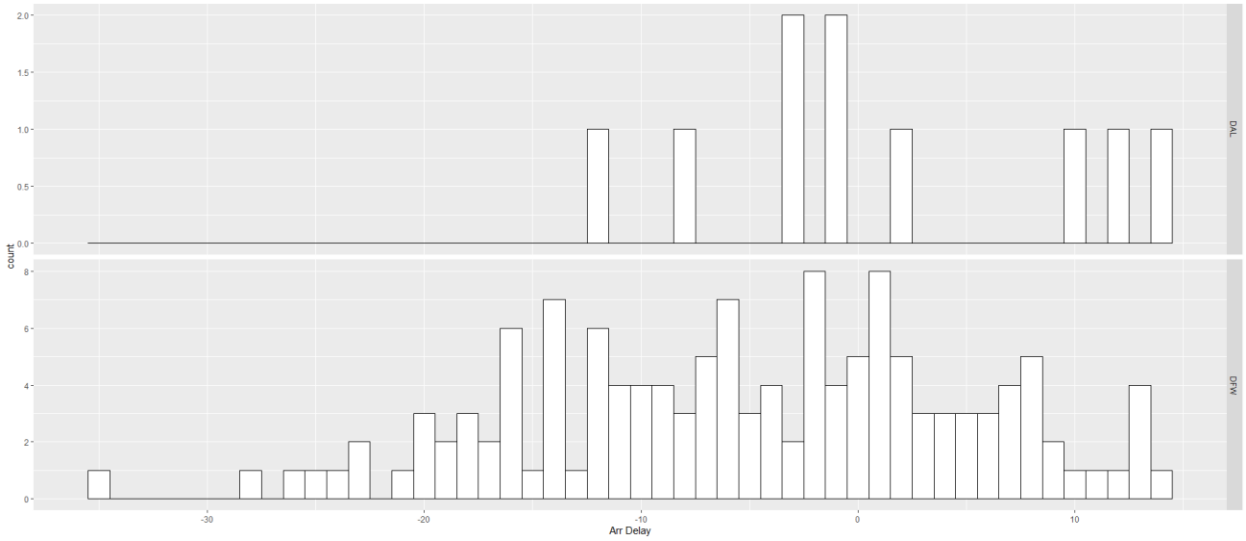


Figure 6: Histogram of Flight Arrival Delay Times for DFW and DAL Airports

The histogram shown above compares a flight's arrival time to the expected arrival time. A value negative means the flight arrived earlier than originally expected, and a positive value is

the delay time. Flights at Dallas Love Field (DAL) are more normally distributed, but when a flight is delayed it is more than 10 minutes late. Dallas/Fort Worth (DFW) has many more delayed flights (not surprising being a busier airport), but despite the increase in flight delays, the results are slightly left skewed indicating that more of the flights overall are arriving on time to early than those at DAL.

The table below shows the arrival delays for all flights in 2020. The delay time is unimodal and drastically skewed right (most delays are less than an hour), and therefore is best represented of central tendency and variation is by the 5-number summary. The minimum value of -139 means that one flight arrived just over 2 hours early. The 1st quartile (-11), median (-8), and 3rd quartile (-2) being negative means that more than 75% (found to be almost 95% in (left) Pie Chart, 5.17% of Flights have a 15+min delay shown above) of the flights arrive early. The maximum value of 2,367 minutes; however, means that although you are likely to arrive early, there is a possibility of arriving 40 hours late. The data has a skewedness value of 16 which emphasizes the long tail to the left. The data also has a kurtosis value of 541, meaning the peaks are drastically high.

```
> describe(airlines$`Arr Delay`)
   vars      n mean sd median trimmed mad  min  max range skew kurtosis   se
X1    1 1e+06   -8 27    -11     -10  12 -139 2367  2506   16      541 0.03
> #summary(airlines$`Arr Delay`), stat.desc(airlines$`Arr Delay`,basic=F)
> #IQR = 43-12.35
> hist(arrivalDelayTime)
> summary(airlines$`Arr Delay`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   -139     -19     -11      -8      -2    2367    5229
```

Equation 1: Code Output of 5-Number Summary for Arrival Delay Times

The following information determines what type of flight issues cause the 15+ minute delay when leaving the Dallas-Fort Worth metroplex (airports DFW and DAL).

The data represents a weak and increasing linear pattern. As many values do not fall on the line of best fit, there are obvious outliers to the data, mostly the values where the there was a 15^{+} minute flight delay but not the delay issue in question.

If you decide to take a flight, we need to know whether your ride at the destination should expect to arrive at the airport early, on time, or late to pick you up. Recent security measures at many airports prevent guests from sitting on the curb and awaiting their friends to exit the airport and instead either must pay to park or continue to circle the pickup lanes and hope that on this loop of the circuit will have the flier ready to be picked up. A flight is delayed if it is 15 minutes late, let us examine the odds that the flight is going to be on time.

```
        One Sample t-test

data:  arrivalDelayTime
t = 7, df = 852, p-value = 6e-12
alternative hypothesis: true mean is greater than 15
95 percent confidence interval:
 938 Inf
sample estimates:
mean of x
     1229
```

Equation 2: One Sample t-test Airline Delays

$$H_0: \mu \leq 15\ minutes$$

$$H_A: \mu > 15\ minutes$$

According to the code output, the same mean flight delay is 1,229 minutes. The p-value associated with this sample mean is nearly 0 which means there is a negligible chance that the this mean occurred by chance. We can therefore reject the null hypothesis that a flight will arrive early or on time in favor for the flight arriving late. This means that your ride should track the flight or await your call upon safely landing at the destination before leaving for the airport.

Anyone who flies frequently has airports that they hate for various reasons. In particular there is one that my friends and family refer to as the "airplane graveyard" as any plane that lands at that airport is taken out of commission and the delays in departures always seem to be forever. Looking back at Figure 6: Histogram of Flight Arrival Delay Times for DFW and DAL Airports above, we can see that DFW and DAL both have flight delays. But with DAL being a less busy airport it appears that the flights are more frequently ontime than those at DFW.

```
        Welch Two Sample t-test

data:  airlines$`Arr Delay` by airlines$Origin
t = -4, df = 853, p-value = 2e-04
alternative hypothesis: true difference in means between group DAL and group DFW is not equal to 0
95 percent confidence interval:
 -14.2  -4.4
sample estimates:
mean in group DAL mean in group DFW
               45                55
```

Equation 3: Two Sample t-test Arrival Delays DAL vs DFW

$$H_{0\ arrival\ delay}: \mu_{DAL} = \mu_{DFW}$$

$$H_{A\ arrivial\ delay}: \mu_{DAL} \neq \mu_{DFW}$$

According to the code output, the sample mean flight arrival delay for airplanes into DAL is 45 minutes, which is slightly less than the arrival delay time of flights into DFW (55 minutes).

The p-value associated with this delay time is 0.0002 minutes, which means there is a 0.02% chance that this difference in arrival time delay between DAL and DFW occurred randomly. *We therefore reject the null hypothesis (that the arrival delays are equal) and conclude that we are 95% certain that the population mean arrival time for a flight is different for the two airports.*

According to the data, 136 flights left DFW at Christmas time and 87 of them arrived at their destination early (arrival delay time less than 0), yielding a sample proportion of 63.97% early arrival. 10 flights left DAL at Christmas time and 6 of them arrived at their destination early, yielding a sample proportion of 60.00% early arrival. Christmas is a busy time of year. Is this low percentage of early arrivals normal for these airports? What are the chances that the population of early arrival flights (71,008 flights between November 1, 2020 and June 30, 2021) was 74.47189% (rounded to 5 decimals).

$$H_{0\ early\ arrival}: \mu_{DAL} = \mu_{DFW} = 0.7447$$

$$H_{A\ early\ arrival}: \mu_{DAL} \neq 0.7447$$

$$H_{A\ early\ arrival}: \mu_{DFW} \neq 0.7447$$

```
> #Proportional t.tests
> # x: the number of of successes
> # n: the total number of trials
> # p: the probability to test against.
> prop.test (87,136,.7447189)

        1-sample proportions test with continuity correction

data:  87 out of 136, null probability 0.7447189
X-squared = 7, df = 1, p-value = 0.007
alternative hypothesis: true p is not equal to 0.74
95 percent confidence interval:
 0.55 0.72
sample estimates:
   p
0.64

> prop.test (6,10,.7447189)

        1-sample proportions test with continuity correction

data:  6 out of 10, null probability 0.7447189
X-squared = 0.5, df = 1, p-value = 0.5
alternative hypothesis: true p is not equal to 0.74
95 percent confidence interval:
 0.27 0.86
sample estimates:
  p
0.6
```

Equation 4: Proportional t-test for DAL and DFW early arrivals

According to the code output, the p-value associated with the sample proportion for DFW is 0.007, which means that there is a 0.7% chance that this sample proportion occurred randomly. Therefore, we are 95% confident that the sample proportion falls within the range of acceptable values for the population (55% to 72% early arrival), so we fail to reject the null hypothesis for DFW. The sample proportion for DAL is 0.5, which means there is a 50% chance that the sample proportion occurred randomly. Therefore, we are 95% confident that the sample proportion falls within the range of acceptable values for the population (27% to 86% early arrival), so we also fail to reject the null hypothesis for DAL.

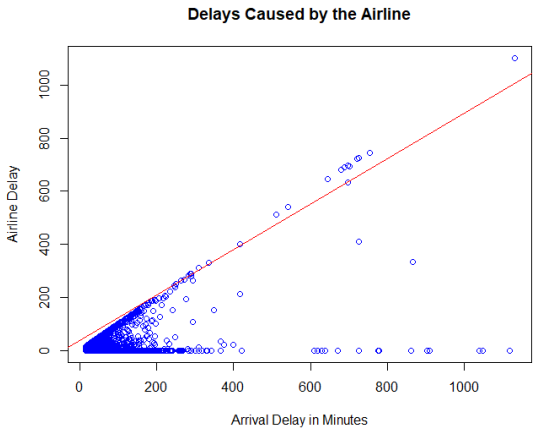


Figure 7: Scatterplot Arrival Delays because of Airline Issues

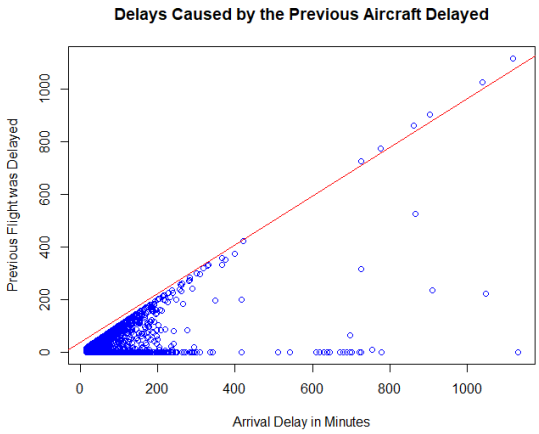


Figure 8: Scatterplot Arrival Delays because of Late Plane

The best fit line is represented by the equation $\hat{y} = 17.010 + 0.857x$ as calculated using the linear model function shown below. The regression summary shows that the line of best fit explains 31% of the variability in the data.

The best fit line is represented by the equation $\hat{y} = 36.881 + 0.927x$ as calculated using the linear model function shown above. The regression summary shows that the line of best fit explains 38.5% of the variability in the data.

```
        Pearson's product-moment correlation

data:  airlines$`Arr Delay` and airlines$`Carrier Delay`
t = 43, df = 4114, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.54 0.58
sample estimates:
 cor
0.56

> # Linear regression model
> airlinesregC <- lm (airlines$`Arr Delay` ~ airlines$`Carrier Delay`, data = airlines)
> airlinesregC

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Carrier Delay`,
    data = airlines)

Coefficients:
             (Intercept)  airlines$`Carrier Delay`
                  37.030                     0.857

> summary (airlinesregC)

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Carrier Delay`,
    data = airlines)

Residuals:
   Min     1Q Median     3Q    Max
 -34.9  -26.2  -17.3   -0.2 1080.0

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              37.0296     1.0743    34.5   <2e-16 ***
airlines$`Carrier Delay`  0.8565     0.0199    43.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64 on 4114 degrees of freedom
Multiple R-squared:  0.31,      Adjusted R-squared:  0.31
F-statistic: 1.85e+03 on 1 and 4114 DF,  p-value: <2e-16
```

Equation 5: Correlation and Regression Summary Airline Delays

```
        Pearson's product-moment correlation

data:  airlines$`Arr Delay` and airlines$`Late Aircraft Delay`
t = 51, df = 4114, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.60 0.64
sample estimates:
 cor
0.62

> # Linear regression model
> airlinesregL <- lm (airlines$`Arr Delay` ~ airlines$`Late Aircraft Delay`, data = airlines)
> airlinesregL

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Late Aircraft Delay`,
    data = airlines)

Coefficients:
                (Intercept)  airlines$`Late Aircraft Delay`
                     36.881                          0.927

> summary (airlinesregL)

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Late Aircraft Delay`,
    data = airlines)

Residuals:
   Min     1Q Median     3Q    Max
 -35.8  -21.9  -14.9    0.1 1093.1

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                36.8809     1.0024    36.8   <2e-16 ***
airlines$`Late Aircraft Delay`  0.9270     0.0183    50.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61 on 4114 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.385
F-statistic: 2.58e+03 on 1 and 4114 DF,  p-value: <2e-16
```

Equation 6: Correlation and Regression Aircraft Delays

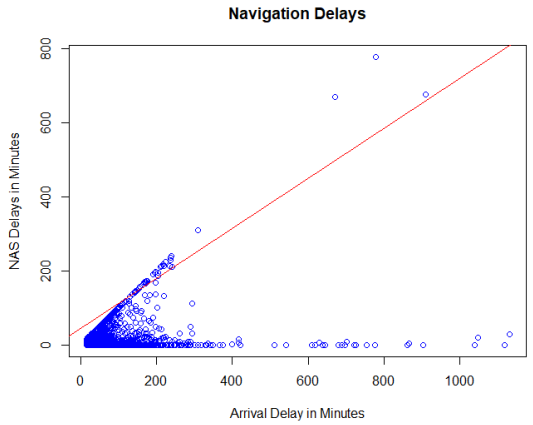


Figure 9: Scatterplot Arrival Delays because of NAS

The best fit line is represented by the equation $\hat{y} = 53.7556 - 0.0626x$ as calculated using the linear model function shown above. The regression summary shows that the line of best fit explains 0.0005% of the variability in the data.

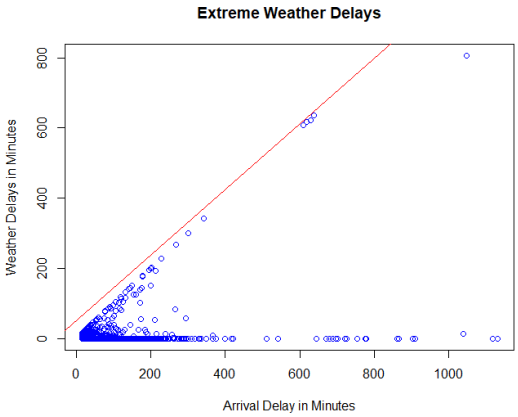


Figure 10: Scatterplot Arrival Delays because of Weather

The best fit line is represented by the equation $\hat{y} = 51.008 + 0.937x$ as calculated using the linear model function shown above. The regression summary shows that the line of best fit explains 11.3% of the variability in the data.

```
        Pearson's product-moment correlation

data:  airlines$`Arr Delay` and airlines$`Nas Delay`
t = 17, df = 4114, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.23 0.29
sample estimates:
 cor
0.26

> # Linear regression model
> airlinesregS <- lm (airlines$`Arr Delay` ~ airlines$`Security Delay`, data = airlines)
> airlinesregS

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Security Delay`,
    data = airlines)

Coefficients:
               (Intercept)  airlines$`Security Delay`
                   53.7556                    -0.0626

> summary (airlinesregS)

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Security Delay`,
    data = airlines)

Residuals:
   Min     1Q Median     3Q    Max
 -38.8  -33.8  -23.8    3.3 1076.2

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               53.7556     1.2082   44.49   <2e-16 ***
airlines$`Security Delay`  -0.0626     0.4339   -0.14     0.89
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77 on 4114 degrees of freedom
Multiple R-squared:  5.06e-06,  Adjusted R-squared:  -0.000238
F-statistic: 0.0208 on 1 and 4114 DF,  p-value: 0.885
```

Equation 7: Correlation and Regression Navigation Delays

```
        Pearson's product-moment correlation

data:  airlines$`Arr Delay` and airlines$`Weather Delay`
t = 23, df = 4114, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.31 0.36
sample estimates:
 cor
0.34

> # Linear regression model
> airlinesregW <- lm (airlines$`Arr Delay` ~ airlines$`Weather Delay`, data = airlines)
> airlinesregW

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Weather Delay`,
    data = airlines)

Coefficients:
               (Intercept)  airlines$`Weather Delay`
                    51.008                    0.937

> summary (airlinesregW)

Call:
lm(formula = airlines$`Arr Delay` ~ airlines$`Weather Delay`,
    data = airlines)

Residuals:
   Min     1Q Median     3Q    Max
 -50.1  -31.0  -22.0    4.0 1079.0

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               51.0078     1.1422    44.7   <2e-16 ***
airlines$`Weather Delay`   0.9369     0.0409    22.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 73 on 4114 degrees of freedom
Multiple R-squared:  0.113,     Adjusted R-squared:  0.113
F-statistic:  523 on 1 and 4114 DF,  p-value: <2e-16
```

Equation 8: Correlation and Regression Weather Delays

The sample means shown below represent the average amount of time that a flight will arrive late due to the delay specified. Using the 95% confidence interval, we say that with 95% certainty that the average time delay of a flight is: 20 minutes for an airline related delay, 18 minutes if the previous flight arrived late, 13 minutes for navigation issues, 0.2 minutes for security reasons, 3 minutes for inclement weather, or 54 minutes for overall or unspecified delays. Being 95% confident that your flight won't be more than an hour delayed is worth the gamble to be able to spend the holidays with friends and family.

```
> stat.desc(airlines$`Arr Delay`,basic=F,p=0.95)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
        30.0         53.7          1.2          2.4       5985.6         77.4          1.4

> stat.desc(airlines$`Carrier Delay`,basic=F,p=0.95)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
        2.00        19.52         0.78         1.54      2532.02        50.32         2.58

> stat.desc(airlines$`Late Aircraft Delay`,basic=F,p=0.95)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
        0.00        18.19         0.81         1.58      2682.51        51.79         2.85

> stat.desc(airlines$`Nas Delay`,basic=F,p=0.95)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
        3.00        12.95         0.47         0.92       903.65        30.06         2.32

> stat.desc(airlines$`Security Delay`,basic=F,p=0.95)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
       0.000        0.166        0.043        0.085        7.727        2.780       16.703

> stat.desc(airlines$`Weather Delay`,basic=F,p=0.95)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
        0.00         2.92         0.43         0.85       769.76        27.74         9.50

> options (digits=2)
>
```

Equation 9: Confidence Intervals for Flight Delays

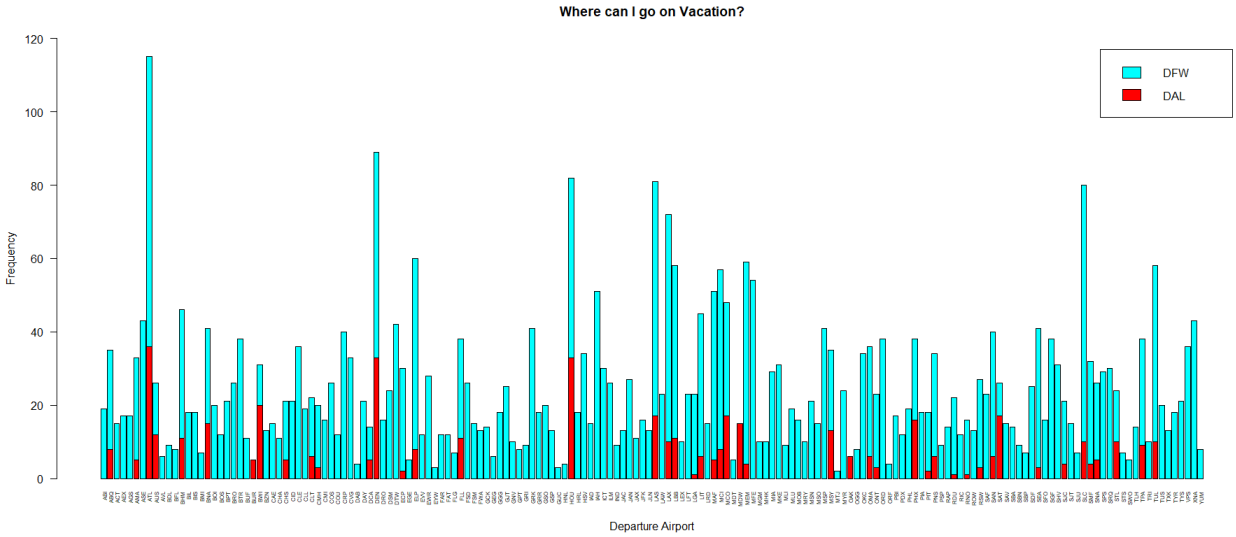


Figure 11: Barchart of Flight Destinations from the Dallas-Fort Worth Metroplex

The graph above shows the frequency of flight destinations when starting from DFW (blue) and DAL (red) airports. DFW being ranked number two in the country naturally has more flights than DAL, but that has not stopped Love Field from giving them a good run for their money over the last few years. With the airspace ban on Love Field lifted the number of flights leaving DAL has increased drastically over the last decade (see Table 4: Top 5 Airports in the Texas, USA), and having unique airlines flying out of each airport, the business practices of the rival companies and customer loyalty will ultimately determine the upper limit of the expansion that DAL can achieve.

```
        Pearson's Chi-squared test

data:  airports
X-squared = 1106, df = 169, p-value <2e-16

> airportschi$observed

    ABI ABQ ACT AEX AGS AMA ASE ATL AUS AVL BDL BFL BHM BIL BIS BMI BNA BOI BOS BPT BRO BTR BUF BUR BWI BZN CAE CHA CHS CID CLE CLL CLT CMH CMI COS COU CRP CVG DAB DAY DCA DEN DRO DSM DTW ECP EGE ELP EVV
DAL   0   8   0   0   0   5   0  36  12   0   0   0  11   0   0   0  15   0   0   0   0   0   0   5  20   0   0   0   5   0   0   0   6   3   0   0   0   0   0   0   5  33   0   0   0   2   0   8   0
DFW  19  27  15  17  17  28  43  79  14   6   9   8  35  18  18   7  26  20  12  21  26  38  11   0  11  13  15  11  16  21  36  19  16  17  16  26  12  40  33   4  21   9  36  16  24  42  28   5  52  12

    EWR EYW FAR FAT FLG FLL FSD FSM FWA GCK GEG GGG GJT GNV GPT GRI GRK GRR GSO GSP GUC HNL HOU HRL HSV IAD IAH ICT ILM IND JAC JAN JAX JFK JLN LAS LAW LAX LBB LEX LFT LGA LIT LRD MAF MCI MCO MDT MDW MEM
DAL   0   0   0   0   0  11   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  33   0   0   0   0   0   0   0   0   0   0   0   0   0   0  17   0  10  11   0   0   1   6   0   5   8  17   0  15   4
DFW  28   3  12  12   7  27  26  15  13  14   6  18  25  10   8   9  41  18  20  13   3   4  49  18  34  15  51  30  26   9  13  27  11  16  13  64  23  62  47  10  23  22  39  15  46  49  31   5   0  55

    MFE MGM MHK MIA MKE MLI MLU MOB MRY MSN MSO MSP MSY MTJ MYR OAK OGG OKC OMA ONT ORD ORF PBI PDX PHL PHX PIA PIT PNS PSP RAP RDU RIC RNO ROW RSW SAF SAN SAT SAV SBA SBN SBP SDF SEA SFO SGF SHV SJC SJT
DAL   0   0   0   0   0   0   0   0   0   0   0   0  13   0   0   6   0   0   6   3   0   0   0   0   0   0  16   0   2   6   0   0   1   0   1   0   3   0   6  17   0   0   0   0   3   0   0   4   0
DFW  54  10  10  29  31   9  19  16  10  21  15  41  22   2  24   0   8  34  30  20  38   4  17  12  19  22  18  16  28   9  14  21  12  15  13  24  23  34   9  15  14   9   7  25  38  16  38  31  17  15

    SJU SLC SMF SNA SPS SRQ STL STS SWO TLH TPA TRI TUL TUS TXK TYR TYS VPS XNA YUM
DAL   0  10   4   5   0   0  10   0   0   0   9   0  10   0   0   0   0   0   0   0
DFW   7  70  28  21  29  30  14   7   5  14  29  10  48  20  13  18  21  36  43   8
> airportschi$expected

      ABI  ABQ  ACT  AEX  AGS  AMA  ASE  ATL  AUS  AVL  BDL  BFL  BHM  BIL  BIS  BMI  BNA  BOI  BOS  BPT  BRO  BTR  BUF  BUR  BWI  BZN  CAE  CHA  CHS  CID  CLE  CLL  CLT  CMH  CMI  COS  COU  CRP  CVG  DAB  DAY
DAL  2.1  3.8  1.6  1.8  1.8  3.6  4.7  12  2.8 0.65 0.98 0.87   5   2   2 0.76  4.5  2.2  1.3  2.3  2.8  4.1 1.2 0.54  3.4  1.4  1.6 1.2  2.3  2.3  3.9  2.1  2.4  2.2  1.7  2.8  1.3  4.3  3.6 0.43  2.3
DFW 16.9 31.2 13.4 15.2 15.2 29.4 38.3 103 23.2 5.35 8.02 7.13  41  16  16 6.24 36.5 17.8 10.7 18.7 23.2 33.9 9.8 4.46 27.6 11.6 13.4 9.8 18.7 18.7 32.1 16.9 19.6 17.8 14.3 23.2 10.7 35.7 29.4 3.57 18.7

      DCA  DEN  DRO  DSM  DTW  ECP  EGE  ELP  EVV  EWR  EYW  FAR  FAT  FLG  FLL  FSD  FSM  FWA  GCK  GEG  GGG  GJT  GNV  GPT  GRI  GRK  GRR  GSO  GSP  GUC  HNL  HOU  HRL  HSV  IAD  IAH  ICT  ILM  IND  JAC  JAN
DAL  1.5  9.7  1.7  2.6  4.6  3.3 0.54  6.5  1.3   3 0.33  1.3  1.3 0.76  4.1  2.8  1.6  1.4  1.5 0.65   2  2.7 1.3 0.87 0.98  4.5   2  2.2  1.4 0.33 0.43  8.9   2  3.7  1.6  5.5  3.3  2.8 0.98  1.4  2.9
DFW 12.5 79.1 14.3 21.4 37.4 26.7 4.46 53.5 10.7  25 2.67 10.7 10.7 6.24 33.9 23.2 13.4 11.6 12.5 5.35  16 22.3 8.9 7.13 8.02 36.5  16 17.8 11.6 2.67 3.57 73.1  16 30.3 13.4 45.5 26.7 23.2 8.02 11.6 24.1

      JAX  JFK  JLN  LAS  LAW  LAX  LBB  LEX  LFT  LGA  LIT  LRD  MAF  MCI  MCO  MDT  MDW  MEM  MFE  MGM  MHK  MIA  MKE  MLI  MLU  MOB  MRY  MSN  MSO  MSP  MSY  MTJ  MYR  OAK  OGG  OKC  OMA  ONT  ORD  ORF  PBI
DAL 1.2  1.7  1.4  8.8  2.5  7.8  6.3 1.1  2.5  2.5  4.9  1.6  5.5  6.2  5.2 0.56  1.6  6.4  5.9 1.1 1.1  3.1  3.4 0.98  2.1  1.7 1.1  2.3  1.6  4.5  3.8 0.22  2.6 0.65 0.87  3.7  3.9  2.5  4.1 0.43  1.8
DFW 9.8 14.3 11.6 72.2 20.5 64.2 51.7 8.9 20.5 20.5 40.1 13.4 45.5 50.8 42.8 4.46 13.4 52.6 48.1 8.9 8.9 25.9 27.6 8.02 16.9 14.3 8.9 18.7 13.4 16.5 31.2 1.78 21.4 5.35 7.13 30.3 32.1 20.5 33.9 3.57 15.2

      PDX  PHL  PHX  PIA  PIT  PNS  PSP  RAP  RDU  RIC  RNO  ROW  RSW  SAF  SAN  SAT  SAV  SBA  SBN  SBP  SDF  SEA  SFO  SGF  SHV  SJC  SJT  SJU  SLC  SMF  SNA  SPS  SRQ  STL  STS  SWO  TLH  TPA  TRI  TUL
DAL  1.3  2.1  4.1   2   2  3.7 0.98  1.5  2.4  1.3  1.7  1.4  2.9  2.5  4.3  2.8  1.6  1.5 0.98 0.76  2.7  4.5  1.7  4.1  3.4  2.3  1.6 0.76  8.7  3.5  2.8  3.1  3.3  2.6 0.76 0.54  1.5  4.1 1.1  6.3
DFW 10.7 16.9 33.9  16  16 30.3 8.02 12.5 19.6 10.7 14.3 11.6 24.1 20.5 35.7 23.2 13.4 12.5 8.02 6.24 22.3 36.5 14.3 33.9 27.6 18.7 13.4 6.24 71.3 28.5 23.2 25.9 26.7 21.4 6.24 4.46 12.5 33.9 8.9 51.7

     TUS  TXK TYR  TYS  VPS  XNA  YUM
DAL  2.2  1.4   2  2.3  3.9  4.7 0.87
DFW 17.8 11.6  16 18.7 32.1 38.3 7.13
> |
```

Equation 10: Chi-Square Analysis for Flight Destinations

The null hypothesis is that the destination airport is independent of the origin airport (DAL/DFW). The alternative hypothesis is that there is a relationship between flight's originating and destination airport.

The value of the test statistic was calculated in Chi-Square Analysis for Flight Destinations above and χ^2 = 1106. Since the calculated p-value, nearly 0, is less than the 95% confidence critical value of 9.488, do not reject the null hypothesis. There is not sufficient evidence to claim that the origin and destination airports are somehow related.

```
> anova1
Call:
   aov(formula = airlines$`Arr Delay` ~ airlines$`Carrier Delay`,
    data = airlines)

Terms:
                airlines$`Carrier Delay` Residuals
Sum of Squares                  3.5e+08   4.4e+08
Deg. of Freedom                   1e+00     1e+06

Residual standard error: 20
Estimated effects may be unbalanced
> anova2
Call:
   aov(formula = airlines$`Arr Delay` ~ airlines$`Late Aircraft Delay`,
    data = airlines)

Terms:
                airlines$`Late Aircraft Delay` Residuals
Sum of Squares                        1.7e+08   6.1e+08
Deg. of Freedom                         1e+00     1e+06

Residual standard error: 24
Estimated effects may be unbalanced
> anova3
Call:
   aov(formula = airlines$`Arr Delay` ~ airlines$`Nas Delay`, data = airlines)

Terms:
                airlines$`Nas Delay` Residuals
Sum of Squares               8.5e+07   7.0e+08
Deg. of Freedom                1e+00     1e+06

Residual standard error: 26
Estimated effects may be unbalanced
> anova4
Call:
   aov(formula = airlines$`Arr Delay` ~ airlines$`Security Delay`,
    data = airlines)

Terms:
                airlines$`Security Delay` Residuals
Sum of Squares                  1.2e+06   7.8e+08
Deg. of Freedom                   1e+00     1e+06

Residual standard error: 27
Estimated effects may be unbalanced
> anova5
Call:
   aov(formula = airlines$`Arr Delay` ~ airlines$`Weather Delay`,
    data = airlines)

Terms:
                airlines$`Weather Delay` Residuals
Sum of Squares                  8.2e+07   7.0e+08
Deg. of Freedom                   1e+00     1e+06

Residual standard error: 26
Estimated effects may be unbalanced
> summary(anova1)
                        Df   Sum Sq  Mean Sq F value Pr(>F)
airlines$`Carrier Delay`       1 3.46e+08 3.46e+08  827455 <2e-16 ***
Residuals                1048573 4.39e+08 4.18e+02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova2)
                        Df   Sum Sq  Mean Sq F value Pr(>F)
airlines$`Late Aircraft Delay`       1 1.73e+08 1.73e+08  296471 <2e-16 ***
Residuals                1048573 6.12e+08 5.83e+02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova3)
                     Df   Sum Sq  Mean Sq F value Pr(>F)
airlines$`Nas Delay`       1 8.46e+07 84593361  126677 <2e-16 ***
Residuals            1048573 7.00e+08      668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova4)
                        Df   Sum Sq Mean Sq F value Pr(>F)
airlines$`Security Delay`       1 1.19e+06 1189136    1591 <2e-16 ***
Residuals                1048573 7.84e+08     747
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova5)
                        Df   Sum Sq  Mean Sq F value Pr(>F)
airlines$`Weather Delay`       1 8.16e+07 81823246  122046 <2e-16 ***
Residuals                1048573 7.03e+08      670
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Equation 11: ANOVA Test Delay Times

As shown to the left, the ANOVA test compares the delay time for carrier delays, aircraft delays, NAS delays, security delays, and inclement weather delays. But which delay causes the most trouble to travelers or are they all the same?

Null hypothesis: That the mean delay time from carrier delay, late aircraft arrival, NAS delay, and weather delay (in minutes) for the flights leaving DAL and/or DFW airports are the same.

$$H_0 = \mu_{Arrival\ Delay} = \mu_{carrier} = \mu_{late\ aircraft}$$

$$= \mu_{NAS} = \mu_{security} = \mu_{weather}$$

Alternative hypothesis: At least one of the delay types does not have the same mean as the others.

Since the P-value for every delay type is less than 0.05, and even less than 0.01, we can reject the null hypothesis of equal means. At least one flight delay has a different mean delay time than the others.

To determine which flight delays are likely to impact your flight, we can find the multiple regression equation. This equation is represented by taking the following delays (in minutes) carrier delay, late aircraft arrival, weather delay, NAS delay, security delay, and is the flight cancelled or diverted for the flights leaving DAL and/or DFW airports. Using the following formula with the variables of the same name as the body parts will calculate the subject's BMI.

$$TotalFlightDelay = CarrierDelay + 1.1LateAircraft + WeatherDelay + 1.1NAS + 1.2SecurityDelay + 11IsCancelled + 11.2IsDiverted - 11.2$$

The odds of the delay being random is almost 0% for each of the types of delay (Pr value). Each of these independent variables has a statistically significant link (at a $p < 0.05$ significance level) to the dependent variable of flight delay time. It should therefore not be surprising that the R^2 value associated with this multiple regression model is quite high. R-squared being 0.819 means that 81.9% of the flight time variability can be explained by the best-fit relationship created with these independent variables.

```
Call:
lm(formula = airlines$"Arr Delay" ~ 1, data = airlines)

Coefficients:
(Intercept)
      -7.92


> # Forwards stepwise variables selection
> airlinesregs <-step(airlinesreg0, direction = "forward", scope = (~ airlines$Cancelled + airlines$`Carrier .... [TRUNCATED]

> anova (airlinesregs)
Analysis of Variance Table

Response: airlines$"Arr Delay"
                              Df   Sum Sq  Mean Sq F value Pr(>F)
airlines$`Carrier Delay`       1 3.46e+08 3.46e+08 2554159 <2e-16 ***
airlines$`Late Aircraft Delay` 1 1.55e+08 1.55e+08 1140460 <2e-16 ***
airlines$`Weather Delay`       1 7.75e+07 7.75e+07  571835 <2e-16 ***
airlines$`Nas Delay`           1 6.29e+07 6.29e+07  464414 <2e-16 ***
airlines$`Security Delay`      1 9.01e+05 9.01e+05    6647 <2e-16 ***
airlines$Cancelled             1 5.45e+05 5.45e+05    4024 <2e-16 ***
airlines$Diverted              1 1.02e+05 1.02e+05     751 <2e-16 ***
Residuals                1048567 1.42e+08 1.36e+02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> coef(airlinesregs)
                  (Intercept)       airlines$`Carrier Delay` airlines$`Late Aircraft Delay`       airlines$`Weather Delay`           airlines$`Nas Delay`
                        -11.2                            1.0                            1.1                            1.0                            1.1
     airlines$`Security Delay`              airlines$Cancelled              airlines$Diverted
                          1.2                           11.2                           11.2

> confint (airlinesregs)
                              2.5 % 97.5 %
(Intercept)                 -11.2  -11.1
airlines$`Carrier Delay`      1.0    1.0
airlines$`Late Aircraft Delay`  1.1    1.1
airlines$`Weather Delay`      1.0    1.0
airlines$`Nas Delay`          1.1    1.2
airlines$`Security Delay`     1.2    1.2
airlines$Cancelled           10.8   11.5
airlines$Diverted            10.4   12.0

> summary (airlinesregs)  #multiple R squared

Call:
lm(formula = airlines$"Arr Delay" ~ airlines$`Carrier Delay` +
    airlines$`Late Aircraft Delay` + airlines$`Weather Delay` +
    airlines$`Nas Delay` + airlines$`Security Delay` + airlines$Cancelled +
    airlines$Diverted, data = airlines)

Residuals:
    Min      1Q  Median      3Q     Max
-171.72   -7.85    0.15    8.15   25.15

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.12e+01   1.15e-02  -968.9   <2e-16 ***
airlines$`Carrier Delay`     1.05e+00   6.83e-04  1534.5   <2e-16 ***
airlines$`Late Aircraft Delay`  1.06e+00   1.04e-03  1026.2   <2e-16 ***
airlines$`Weather Delay`     1.03e+00   1.39e-03   740.0   <2e-16 ***
airlines$`Nas Delay`         1.15e+00   1.69e-03   681.5   <2e-16 ***
airlines$`Security Delay`    1.21e+00   1.48e-02    81.6   <2e-16 ***
airlines$Cancelled           1.12e+01   1.76e-01    63.5   <2e-16 ***
airlines$Diverted            1.12e+01   4.07e-01    27.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12 on 1048567 degrees of freedom
Multiple R-squared:  0.819,     Adjusted R-squared:  0.819
F-statistic: 6.77e+05 on 7 and 1e+06 DF,  p-value: <2e-16
```

Equation 12: Multiple Regression for Flight Delays

The eigenvalues for the components are represented by the row labeled standard deviation.

Equation 13: PCA Calculations

```
> summary(spc)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7      PC8      PC9
Standard deviation     1.5610 1.0247 1.0009 1.0006 0.9992 0.9973 0.9786 0.67962 0.31312
Proportion of Variance 0.2707 0.1167 0.1113 0.1112 0.1109 0.1105 0.1064 0.05132 0.01089
Cumulative Proportion  0.2707 0.3874 0.4987 0.6099 0.7209 0.8314 0.9378 0.98911 1.00000
```

PC #	Field	Eigenvalue	Relevant CP
1	Arr Del15	1.5610	0.2707
2	Arr Delay	1.0247	0.3874
3	Cancelled	1.0009	0.4987
4	Carrier Delay	1.0006	0.6099
5	Diverted	0.9992	-
6	Late Aircraft	0.9973	-
7	Nas Delay	0.9786	-
8	Security Delay	0.67962	-
9	Weather Delay	0.31312	-

Table 5: PC Table of Values

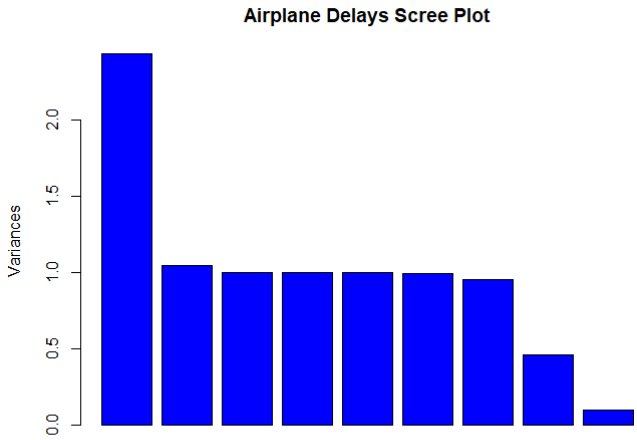


Figure 12: PCA Component Analysis

Looking at the first 4 components (those with an eigenvalue of at least 1), 61% of the variability is explained. This is represented by the first four items on the plot. The fifth item is a value just under 1.

Based on the eigenvalues and the scree plot above the first for Principal Components are taken into account. PC1 shows mostly positive values, although 2 of the 9 results are slightly less than 0. The strongest variables (having a value near or above 0.5) are for arrival delay and identifying if the arrival delay is over 15 minutes. PC2 is still mostly positive; however, there are 3 negative values, one of which is the identifier for NAS delay. There are no strongly positive values, all being less than 0.42, but NAS delay is strongly negative with a value of -0.656. PC3 is equally mixed with positive and negative values. The values more than 0.5 units away from 0 are the indicators for if the flight was diverted (0.74) and for flight

```
Rotation (n x k) = (9 x 9):
                            PC1          PC2          PC3          PC4
Arr Del15            0.4976170059 -0.282225174 -0.008081211 -0.002193139
Arr Delay            0.5969805386  0.201926293  0.001175204 -0.009583987
Cancelled           -0.0014846316  0.195991010 -0.646888244  0.035286603
Carrier Delay        0.4074968590  0.375960547 -0.043245034 -0.489755953
Diverted            -0.0006371674  0.078424008  0.740064059 -0.143616133
Late Aircraft Delay  0.3235420263  0.009242257 -0.007443593 -0.029627751
Nas Delay            0.3025298604 -0.655633421  0.012774638  0.216484169
Security Delay       0.0422245842 -0.296714763 -0.117208030  0.052515670
Weather Delay        0.1787416490  0.419148347  0.133968831  0.829259074
```

Equation 14: PCA Breakdown

cancelation (-0.647). PC4 is also equally mixed with positive and negative values. The only value more than 0.5 units away from 0 is the indicator for if there was a weather delay. PC4's Weather delay was the strongest result of any PC set with a value of 0.829. Although only the first 4 PCs are relevant, the full output from R is shown below.

```
> spc # Component loadings for the variables for each Principal Component
Standard deviations (1, .., p=9):
[1] 1.5609591 1.0247096 1.0008607 1.0005552 0.9992080 0.9972740 0.9785912 0.6796227 0.3131199

Rotation (n x k) = (9 x 9):
                          PC1          PC2          PC3          PC4          PC5          PC6          PC7          PC8          PC9
Arr Del15            0.4976170059 -0.282225174 -0.008081211 -0.002193139  0.005134809  0.06397180 -0.015949835 -0.81736260 -0.014190596
Arr Delay            0.5969805386  0.201926293  0.001175204 -0.009583987 -0.010052327 -0.03901310 -0.005159140  0.27815790  0.723688654
Cancelled           -0.0014846316  0.195991010 -0.646888244  0.035286603 -0.524246054  0.51569607 -0.002212386 -0.02478429 -0.021914795
Carrier Delay        0.4074968590  0.375960547 -0.043245034 -0.489755953  0.046853341 -0.13192667 -0.399749703  0.12655900 -0.505422656
Diverted            -0.0006371674  0.078424008  0.740064059 -0.143616133 -0.553756182  0.34449556 -0.001027717 -0.01073046 -0.009463836
Late Aircraft Delay  0.3235420263  0.009242257 -0.007443593 -0.029627751 -0.021803042 -0.03392795  0.864867193  0.18008439 -0.335037011
Nas Delay            0.3025298604 -0.655633421  0.012774638  0.216484169  0.072025834  0.33551206 -0.260771335  0.44543851 -0.217758349
Security Delay       0.0422245842 -0.296714763 -0.117208030  0.052515670 -0.640239228 -0.68649800 -0.080213615  0.07343906 -0.025855714
Weather Delay        0.1787416490  0.419148347  0.133968831  0.829259074 -0.025062943 -0.08580415 -0.132219649 -0.03950732 -0.244365421

> # Save Component Loadings
> airlinesloadings <- predict(spc)
```

Equation 15: Full PCA Analysis

In summary, out of the 13 variables included in the flight data, looking at their principal components, only the first 4 PC levels have relevant eigenvalues in variance. These first 4 PCs are comprised of mostly positive variables, but all have at least 2 negative variables. The indicator for a flight being delayed at least 15 minutes is only positive in the first result and is negative in the other 3 PCs having the highest quantity of negative results. Weather delay is the only variable that is always positive in all 4 PCs and has the highest result of any variable (positive or negative) for any of the PCs. This PC4 value is almost 0.83.

Only PCs 1-4 are used as the data input for below.

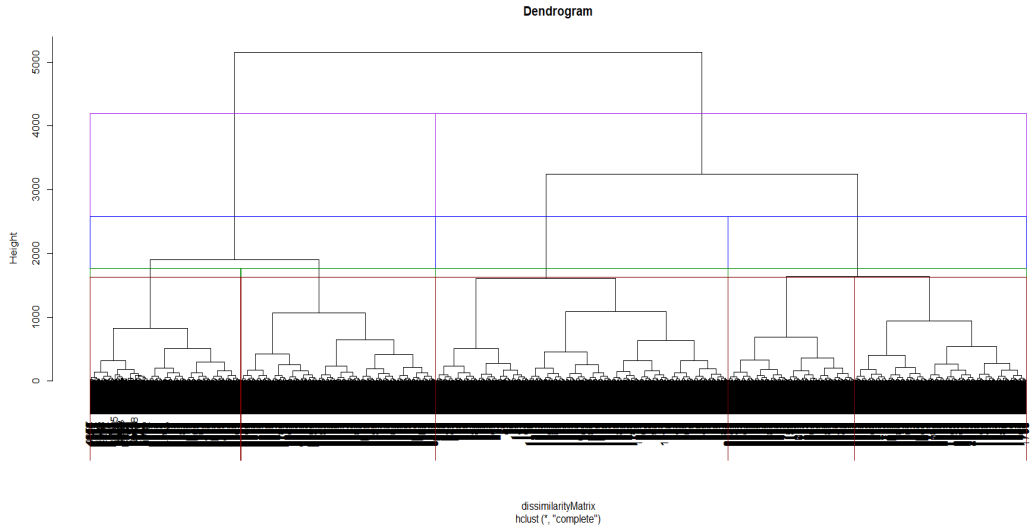


Figure 13: Dendrogram of Flight Delays

The dendrogram is clustered into 4 groups labeled as purple, blue, green, and red above. Some of the cluster locations are breaks at data rows: 1, 96, 191, 286, 381, 476, 571, 666, 761, 856, 951, and R-studio maxed out preventing the others from printing or being legible. I am 95% confidence that the odds of the cluster variability having occurred randomly is less than 4.624%.

```
> stat.desc(sg4,basic=F)
     median        mean     SE.mean CI.mean.0.95         var      std.dev     coef.var
 2.00000000  2.44439935  0.02358004  0.04623876  1.37002952  1.17048260  0.47884263
```

Equation 16: 95% Confidence for Cluster Analysis

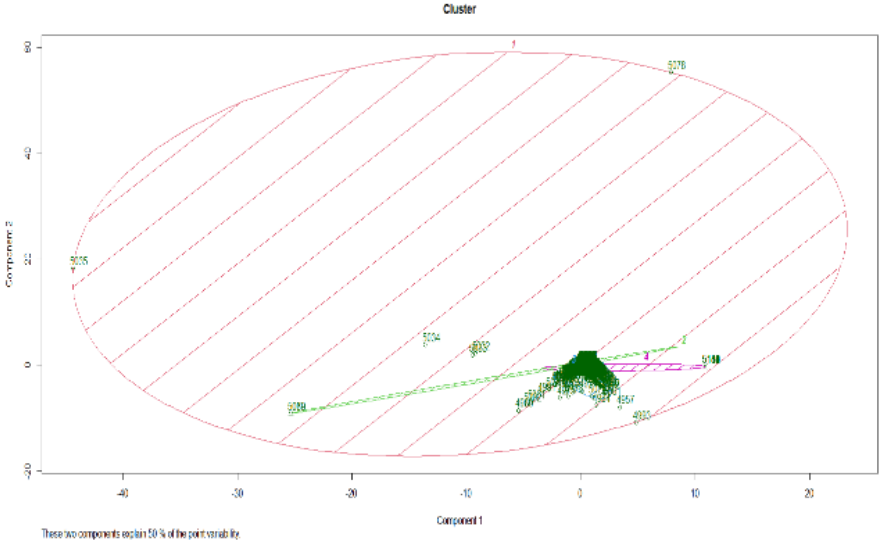


Figure 14: Cluster Analysis of Flight Delays

Several of the cluster locations are shown below but like with the dendrogram are cutoff due to max R-studio print line limitations. You can clearly see the red, purple, and green cluster groups The blue (cluster k=3) you can see the blue 3 label near the point (0,0), but the datapoints are too thick in that area that the blue cluster is buried. Cluster 1 (shown above) best represents the 4 PC groups the best maximizing the distances between the clusters and minimizing the differences within the clusters. I am 95% confidence that the odds of the cluster variability having occurred randomly is less than 5.1682%.

```
> stat.desc(airlinesvar1,basic=F)
                       PC1          PC2          PC3          PC4
median       -2.002618e-01 2.669598e-02 1.703159e-02 9.341060e-03
mean          2.214177e-17 2.171220e-18 3.821023e-17 6.541743e-17
SE.mean       2.635586e-02 2.014820e-02 1.898981e-02 1.273508e-02
CI.mean.0.95  5.168194e-02 3.950916e-02 3.723765e-02 2.497256e-02
var           1.711572e+00 1.000261e+00 8.885504e-01 3.996169e-01
std.dev       1.308271e+00 1.000130e+00 9.426295e-01 6.321526e-01
coef.var      5.908608e+16 4.606305e+17 2.466956e+16 9.663366e+15
```

Equation 17: PCA Statistics

```
> kMeansClustering
K-means clustering with 4 clusters of sizes 27, 2979, 169, 1977

Cluster means:
          PC1          PC2         PC3          PC4
1 -13.2498891  2.967899102 -0.48051694 -2.631664677
2   0.4806169  0.034059684 -0.04396422  0.010327396
3  -3.9983297 -0.983177548  0.02510088  0.187784888
4  -0.2014633 -0.007809847  0.07066328  0.004326753

Clustering vector:
   [1] 2 4 2 2 4 4 2 4 4 2 2 2 2 4 2 2 2 2 2 2 2 4 4 2 2 4 2 2 2 2 2 4 4 2 2 2 4 2 2 2 2 2 4 4 2 2 4 2 2 2 2 2 2 4 2 4 4 4 2 2 2 4 2 2 4 4 4 4 2
  [70] 2 4 4 2 2 2 4 2 2 4 4 4 2 2 4 4 2 4 4 4 4 2 4 4 4 4 2 4 4 4 4 4 4 4 4 2 2 4 4 2 4 4 2 2 2 2 2 2 2 2 2 4 4 2 4 4 4 4 2 2 4 4 2 2 4 2 4 4 2
 [139] 2 4 2 4 4 2 2 4 4 2 4 4 4 4 2 2 2 2 2 4 4 4 2 2 2 2 4 2 4 2 4 2 4 4 4 4 4 4 2 2 4 4 4 4 2 4 4 2 2 4 2 2 2 2 4 2 2 2 4 2 4 2 4 4 4 2 2 2 4
 [208] 2 2 2 2 4 2 2 4 4 4 4 4 4 2 4 2 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 4 4 2 2 2 4 2 2 4 2 2 4 2 2 4 2 2 2 2 2 2 2 4 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2
 [277] 2 2 4 4 4 4 2 2 4 2 2 4 4 2 4 4 2 2 2 2 4 4 4 2 2 2 2 2 4 2 2 2 4 4 2 4 4 4 2 2 4 4 2 2 4 4 2 4 2 2 2 2 4 4 4 2 2 2 4 4 4 2 2 2 2 2 2 4 4
 [346] 4 2 4 4 4 2 4 4 4 4 2 2 2 2 2 2 2 4 4 2 2 4 4 4 4 2 2 2 2 4 2 4 4 4 4 2 2 2 2 2 2 2 2 4 4 4 4 2 2 2 4 4 2 4 4 2 4 4 4 2 2 2 2 4 2 2 4 2 4
 [415] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 4 2 2 4 2 4 4 2 2 2 4 4 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 4 2 2
 [484] 2 2 4 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 4 2 4 2 2 4 4 2 2 2 2 2 2 4 4 2 4 2 2 4 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 4 2
 [553] 4 4 2 4 4 4 2 2 2 4 2 2 4 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 4 2 4 4 2 2 2 2 2 2 4 4 2 2 2 2 4 2 2 2 2 2 2 2 4 2 2 4 2 2 2 2 2 4 2
 [622] 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 4 4 2 2 4 4 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 4 4 2 2 4 4 4
 [691] 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 4 4 2 4 2 4 4 4 4 4 4 4 4 2 2 2 4 4 4 4 4 2 4 2 2 2 4 2 2 4 2 2
 [760] 2 2 2 2 4 2 2 2 2 4 4 2 2 4 4 2 4 4 4 4 4 4 2 2 2 2 4 4 2 4 4 4 2 4 4 4 4 2 2 4 2 2 2 2 2 4 4 4 4 4 2 2 2 4 4 2 2 2 2 4 2 4 4 4 2 2 2 4 2
 [829] 4 4 2 4 4 2 2 4 4 2 4 4 4 4 2 2 4 4 4 2 2 2 4 2 4 2 2 2 2 4 2 2 4 2 4 2 2 2 2 4 2 2 2 2 4 4 2 4 2 2 2 4 4 4 2 2 2 2 2 2 2 2 2 4 2 2 4 2 4
 [898] 4 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 4 4 4 2 4 2 4 4 4 4 2 2 2 2 2 2 4 4 4 4 2 2 4 2 2 2 2 2 4 2 2 2 2 2 4 4 4 2 2 4 2 2 4 4 4 4 4 2 4 4
 [967] 4 2 2 4 4 4 4 4 4 2 2 4 4 4 2 2 2 4 4 2 2 2 2 2 4 2 2 2 2 4 2 2 2 2
 [ reached getOption("max.print") -- omitted 4152 entries ]

Within cluster sum of squares by cluster:
[1] 10004.7729  2601.2072   767.1654  3215.5260
 (between_SS / total_SS =  34.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

> # Graph based on k-means
> require(cluster)

> clusplot(airlinesvar1, kMeansClustering$cluster, color = TRUE, shade = TRUE, lines = 4, labels = 2,main="Cluster")
> 
```

Equation 18: Code for Cluster Analysis

Conclusions/Summary

Figure 1 shows the most frequent airline flown was the conglomerate airline Branded Code Share Partners, followed by Delta and Southwest Airlines. A flight is considered delayed if it arrives more than 15 minutes behind schedule; thankfully, only 5.17% of the flights qualify for this label, and only 0.5% of flights are diverted or cancelled, so if the flight takes off the destination is reached. The maximum delay time is over a day, and if there happened to be a blizzard or hurricane it is easy to see why a flight would be delayed an entire day or more. If a flight is delayed, the delay from Dallas Love Field (DAL) averages a 45-minute delay, and Dallas/Fort Worth International (DFW) is on average 55-minutes delayed. The equation for calculating a flight delay that accounts for 81.9% of the flight time variability is:

$$TotalFlightDelay = CarrierDelay + 1.1LateAircraft + WeatherDelay + 1.1NAS + 1.2SecurityDelay + 11IsCancelled + 11.2IsDiverted - 11.2.$$

If I did the project again, I would try to get a larger timeframe of data, but then filter the data by a few airlines. It would also be an interesting statistic to see if breaking the origin and destination airports by region or time zone to see if the delays are more common in one area or another. Another analysis that could be done (if R studio allows), would be to analyze the delay times per origin and destination airport to see if the same airports with departure delays are the same ones with arrival delays, and/or if the time delays are related to certain airports.

All that being said, the odds of a flight not making it to the destination, even during the busy Christmas season, is in your favor. So, pack your bags, call your loved ones, head to the airport, and enjoy the holiday vacation.

Table of Figures and Tables

R-Code Output Screenshots

Figures, Images, and Graphs

Data Tables

Bibliography

Allen, R., & Stillman, A. (1954). (There's No Place Like) Home for the Holidays [Recorded by P. Como]. Philadelphia, Pennsylvania, United States of America. Retrieved from All Music.

Baumgarten, P. M. (2014). The impact of hubbing concentration on flight delays within airline networks: An empirical analysis of the US domestic market. *Transportation Research. Part E, Logistics and Transportation Review, 66*, 103-114.

Congress, L. o. (2002). *I'll Be Home for Christmas*. Retrieved from Library of Congress: https://www.loc.gov/item/ihas.200000010/

Luttmann, A., & Gaggero, A. A. (2021). *Purchase discounts and travel premiums during holiday periods: Evidence from the airline industry.* Pavia: Munich Personal RePEc Archive.

Sivak, M., Weintraub, D. J., & Flannagan, M. (1991). Nonstop Flying Is Safer Than Driving. *Risk Analysis, Volume 11, Issue 1*, 145-148.

U.S. Department of Commerce. (2019, July 1). *Quick Facts Texas*. Retrieved from United States Census Bureau: https://www.census.gov/quickfacts/fact/table/TX/PST045219

U.S. Department of Transportation. (1998–2012). *Air Travel Consumer Report, Various Issues.* Washington D.C.: U.S. Department of Transportation.

U.S. Department of Transportation Bureau of Transportation Statistics. (2021, 10 13). *Airline Arrival Performance Dashboard*. Retrieved from Airlines, Airports, and Aviation: https://explore.dot.gov/views/ontime_7_7a/Table7?:iid=1&:isGuestRedirectFromVizportal=y&:embed=y

United States Department of Transportation. (2021, October 13). *Passengers Boarded at the Top 50 U.S. Airports*. Retrieved from Bureau of Transportation Statistics: https://www.bts.gov/content/passengers-boarded-top-50-us-airports