Analytics 502

Avik Roy

September 30, 2018

## Abstract

Right now the goal is to get a good understanding of the data. AKA exploratory analysis. Relationships worth checking out: *Arr Delay ~ Day of the week* It seems arrival delay is related to day of the week.

Hypothesis Test H0 : Mean arrival delay on Sat-Evening <= 0 H1 : Mean arival delay on Sat-Evening > 0

There should be a section to check t-test assumptions (n>30, etc, etc)

Net Delay ~ Month Dep Delay ~ Dep Airport Arr Delay ~ Arr Airport Net Delay ~ Airline

Prediction query: What is the chance that a flight will be delayed given…….

#### Basic setup code

library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts -------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

set.seed(111)

#### Checking out the data

flightData <- read.csv("flights.csv", header = TRUE)  
head(flightData)

## YEAR MONTH DAY DAY\_OF\_WEEK AIRLINE FLIGHT\_NUMBER TAIL\_NUMBER  
## 1 2015 1 1 4 AS 98 N407AS  
## 2 2015 1 1 4 AA 2336 N3KUAA  
## 3 2015 1 1 4 US 840 N171US  
## 4 2015 1 1 4 AA 258 N3HYAA  
## 5 2015 1 1 4 AS 135 N527AS  
## 6 2015 1 1 4 DL 806 N3730B  
## ORIGIN\_AIRPORT DESTINATION\_AIRPORT SCHEDULED\_DEPARTURE DEPARTURE\_TIME  
## 1 ANC SEA 5 2354  
## 2 LAX PBI 10 2  
## 3 SFO CLT 20 18  
## 4 LAX MIA 20 15  
## 5 SEA ANC 25 24  
## 6 SFO MSP 25 20  
## DEPARTURE\_DELAY TAXI\_OUT WHEELS\_OFF SCHEDULED\_TIME ELAPSED\_TIME AIR\_TIME  
## 1 -11 21 15 205 194 169  
## 2 -8 12 14 280 279 263  
## 3 -2 16 34 286 293 266  
## 4 -5 15 30 285 281 258  
## 5 -1 11 35 235 215 199  
## 6 -5 18 38 217 230 206  
## DISTANCE WHEELS\_ON TAXI\_IN SCHEDULED\_ARRIVAL ARRIVAL\_TIME ARRIVAL\_DELAY  
## 1 1448 404 4 430 408 -22  
## 2 2330 737 4 750 741 -9  
## 3 2296 800 11 806 811 5  
## 4 2342 748 8 805 756 -9  
## 5 1448 254 5 320 259 -21  
## 6 1589 604 6 602 610 8  
## DIVERTED CANCELLED CANCELLATION\_REASON AIR\_SYSTEM\_DELAY SECURITY\_DELAY  
## 1 0 0 NA NA  
## 2 0 0 NA NA  
## 3 0 0 NA NA  
## 4 0 0 NA NA  
## 5 0 0 NA NA  
## 6 0 0 NA NA  
## AIRLINE\_DELAY LATE\_AIRCRAFT\_DELAY WEATHER\_DELAY  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA

str(flightData)

## 'data.frame': 1048575 obs. of 31 variables:  
## $ YEAR : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ MONTH : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ DAY : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ DAY\_OF\_WEEK : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ AIRLINE : Factor w/ 14 levels "AA","AS","B6",..: 2 1 12 1 2 4 9 12 1 4 ...  
## $ FLIGHT\_NUMBER : int 98 2336 840 258 135 806 612 2013 1112 1173 ...  
## $ TAIL\_NUMBER : Factor w/ 4523 levels "","D942DN","N001AA",..: 1523 1472 384 1432 2017 1061 2623 2282 1477 3690 ...  
## $ ORIGIN\_AIRPORT : Factor w/ 315 levels "ABE","ABI","ABQ",..: 16 172 273 172 272 273 170 172 273 170 ...  
## $ DESTINATION\_AIRPORT: Factor w/ 315 levels "ABE","ABI","ABQ",..: 272 230 65 199 16 212 212 65 85 19 ...  
## $ SCHEDULED\_DEPARTURE: int 5 10 20 20 25 25 25 30 30 30 ...  
## $ DEPARTURE\_TIME : int 2354 2 18 15 24 20 19 44 19 33 ...  
## $ DEPARTURE\_DELAY : int -11 -8 -2 -5 -1 -5 -6 14 -11 3 ...  
## $ TAXI\_OUT : int 21 12 16 15 11 18 11 13 17 12 ...  
## $ WHEELS\_OFF : int 15 14 34 30 35 38 30 57 36 45 ...  
## $ SCHEDULED\_TIME : int 205 280 286 285 235 217 181 273 195 221 ...  
## $ ELAPSED\_TIME : int 194 279 293 281 215 230 170 249 193 203 ...  
## $ AIR\_TIME : int 169 263 266 258 199 206 154 228 173 186 ...  
## $ DISTANCE : int 1448 2330 2296 2342 1448 1589 1299 2125 1464 1747 ...  
## $ WHEELS\_ON : int 404 737 800 748 254 604 504 745 529 651 ...  
## $ TAXI\_IN : int 4 4 11 8 5 6 5 8 3 5 ...  
## $ SCHEDULED\_ARRIVAL : int 430 750 806 805 320 602 526 803 545 711 ...  
## $ ARRIVAL\_TIME : int 408 741 811 756 259 610 509 753 532 656 ...  
## $ ARRIVAL\_DELAY : int -22 -9 5 -9 -21 8 -17 -10 -13 -15 ...  
## $ DIVERTED : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CANCELLED : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CANCELLATION\_REASON: Factor w/ 5 levels "","A","B","C",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ AIR\_SYSTEM\_DELAY : int NA NA NA NA NA NA NA NA NA NA ...  
## $ SECURITY\_DELAY : int NA NA NA NA NA NA NA NA NA NA ...  
## $ AIRLINE\_DELAY : int NA NA NA NA NA NA NA NA NA NA ...  
## $ LATE\_AIRCRAFT\_DELAY: int NA NA NA NA NA NA NA NA NA NA ...  
## $ WEATHER\_DELAY : int NA NA NA NA NA NA NA NA NA NA ...

summary(flightData)

## YEAR MONTH DAY DAY\_OF\_WEEK   
## Min. :2015 Min. :1.000 Min. : 1.00 Min. :1.000   
## 1st Qu.:2015 1st Qu.:1.000 1st Qu.: 6.00 1st Qu.:2.000   
## Median :2015 Median :2.000 Median :13.00 Median :4.000   
## Mean :2015 Mean :1.694 Mean :13.82 Mean :3.953   
## 3rd Qu.:2015 3rd Qu.:2.000 3rd Qu.:21.00 3rd Qu.:6.000   
## Max. :2015 Max. :3.000 Max. :31.00 Max. :7.000   
##   
## AIRLINE FLIGHT\_NUMBER TAIL\_NUMBER ORIGIN\_AIRPORT   
## WN :221586 Min. : 1 : 7750 ATL : 66599   
## DL :147486 1st Qu.: 755 N477HA : 809 ORD : 52961   
## EV :111206 Median :1725 N488HA : 799 DFW : 50933   
## OO :107099 Mean :2257 N476HA : 773 LAX : 38473   
## AA : 97549 3rd Qu.:3485 N485HA : 767 DEN : 38254   
## UA : 87606 Max. :9794 N486HA : 741 IAH : 29802   
## (Other):276043 (Other):1036936 (Other):771553   
## DESTINATION\_AIRPORT SCHEDULED\_DEPARTURE DEPARTURE\_TIME DEPARTURE\_DELAY   
## ATL : 66741 Min. : 1 Min. : 1 Min. : -61.00   
## ORD : 53060 1st Qu.: 920 1st Qu.: 928 1st Qu.: -5.00   
## DFW : 51037 Median :1319 Median :1329 Median : -1.00   
## LAX : 38463 Mean :1323 Mean :1334 Mean : 11.33   
## DEN : 38300 3rd Qu.:1720 3rd Qu.:1731 3rd Qu.: 11.00   
## IAH : 29820 Max. :2359 Max. :2400 Max. :1988.00   
## (Other):771154 NA's :39515 NA's :39515   
## TAXI\_OUT WHEELS\_OFF SCHEDULED\_TIME ELAPSED\_TIME   
## Min. : 1.00 Min. : 1 Min. : 20.0 Min. : 15.0   
## 1st Qu.: 11.00 1st Qu.: 944 1st Qu.: 85.0 1st Qu.: 82.0   
## Median : 14.00 Median :1342 Median :122.0 Median :119.0   
## Mean : 16.65 Mean :1357 Mean :140.3 Mean :136.9   
## 3rd Qu.: 19.00 3rd Qu.:1745 3rd Qu.:173.0 3rd Qu.:169.0   
## Max. :225.00 Max. :2400 Max. :718.0 Max. :766.0   
## NA's :40229 NA's :40229 NA's :2 NA's :43071   
## AIR\_TIME DISTANCE WHEELS\_ON TAXI\_IN   
## Min. : 7.0 Min. : 31.0 Min. : 1 Min. : 1.00   
## 1st Qu.: 60.0 1st Qu.: 368.0 1st Qu.:1110 1st Qu.: 4.00   
## Median : 94.0 Median : 641.0 Median :1516 Median : 6.00   
## Mean :112.8 Mean : 803.4 Mean :1486 Mean : 7.55   
## 3rd Qu.:144.0 3rd Qu.:1046.0 3rd Qu.:1911 3rd Qu.: 9.00   
## Max. :687.0 Max. :4983.0 Max. :2400 Max. :202.00   
## NA's :43071 NA's :41296 NA's :41296   
## SCHEDULED\_ARRIVAL ARRIVAL\_TIME ARRIVAL\_DELAY DIVERTED   
## Min. : 1 Min. : 1 Min. : -82.00 Min. :0.000000   
## 1st Qu.:1120 1st Qu.:1115 1st Qu.: -12.00 1st Qu.:0.000000   
## Median :1524 Median :1521 Median : -3.00 Median :0.000000   
## Mean :1505 Mean :1492 Mean : 7.61 Mean :0.002426   
## 3rd Qu.:1915 3rd Qu.:1917 3rd Qu.: 12.00 3rd Qu.:0.000000   
## Max. :2359 Max. :2400 Max. :1971.00 Max. :1.000000   
## NA's :41296 NA's :43071   
## CANCELLED CANCELLATION\_REASON AIR\_SYSTEM\_DELAY SECURITY\_DELAY   
## Min. :0.00000 :1008048 Min. : 0.0 Min. : 0.0   
## 1st Qu.:0.00000 A: 6974 1st Qu.: 0.0 1st Qu.: 0.0   
## Median :0.00000 B: 28260 Median : 4.0 Median : 0.0   
## Mean :0.03865 C: 5291 Mean : 13.7 Mean : 0.1   
## 3rd Qu.:0.00000 D: 2 3rd Qu.: 19.0 3rd Qu.: 0.0   
## Max. :1.00000 Max. :830.0 Max. :241.0   
## NA's :820047 NA's :820047   
## AIRLINE\_DELAY LATE\_AIRCRAFT\_DELAY WEATHER\_DELAY   
## Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0   
## Median : 2.0 Median : 4.0 Median : 0.0   
## Mean : 18.2 Mean : 22.9 Mean : 3.5   
## 3rd Qu.: 18.0 3rd Qu.: 29.0 3rd Qu.: 0.0   
## Max. :1971.0 Max. :1313.0 Max. :1152.0   
## NA's :820047 NA's :820047 NA's :820047

#### Modifying the dataset

flightData <- mutate(flightData, DAY\_OF\_WEEK = as.factor(DAY\_OF\_WEEK))  
flightData$DAY\_OF\_WEEK <- ordered(flightData$DAY\_OF\_WEEK, levels = c("1","2","3","4","5","6","7"))  
levels(flightData$DAY\_OF\_WEEK) <- c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday")

#### Additional modification

Modifying the dataset convert integer time to HH:MM format time

flightData <- mutate(flightData, SCHEDULED\_DEPARTURE = format(strptime(sprintf("%04d",SCHEDULED\_DEPARTURE), format="%H%M"), format = "%H:%M"))  
flightData <- mutate(flightData, DEPARTURE\_TIME = format(strptime(sprintf("%04d",DEPARTURE\_TIME), format="%H%M"), format = "%H:%M"))

Modifying the dataset to have, a column for morning to night clasification for *scheduled* departure time

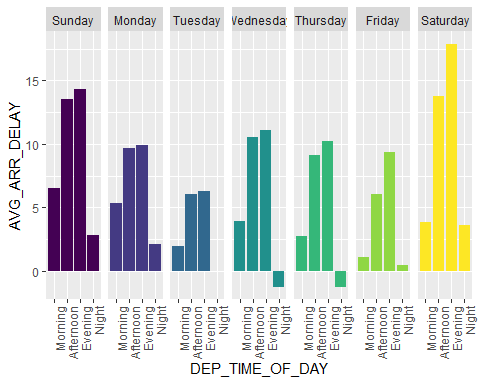
flightData <- mutate(flightData, DEP\_TIME\_OF\_DAY =   
 ifelse (SCHEDULED\_DEPARTURE >= "18:00","Evening",  
 ifelse (SCHEDULED\_DEPARTURE >= "12:00","Afternoon",  
 ifelse (SCHEDULED\_DEPARTURE >= "06:00","Morning","Night")))  
 )

Ordering the factors in Time of day

flightData <- mutate(flightData, DEP\_TIME\_OF\_DAY = as.factor(DEP\_TIME\_OF\_DAY))  
flightData$DEP\_TIME\_OF\_DAY <- ordered(flightData$DEP\_TIME\_OF\_DAY, levels = c("Morning","Afternoon", "Evening","Night"))

Plot of arrival delay vs Day of the week + Time of the day

groupByDayTime <- group\_by(select(flightData, DAY\_OF\_WEEK, DEP\_TIME\_OF\_DAY, ARRIVAL\_DELAY),DAY\_OF\_WEEK,DEP\_TIME\_OF\_DAY)  
groupByDayTime <- groupByDayTime[complete.cases(groupByDayTime),]  
delayByDayTime <- summarise(groupByDayTime, AVG\_ARR\_DELAY = mean(ARRIVAL\_DELAY))  
ggplot(data = delayByDayTime, aes(x = DEP\_TIME\_OF\_DAY, y = AVG\_ARR\_DELAY, fill = DAY\_OF\_WEEK)) + geom\_bar(stat = "identity") + guides(fill = FALSE) + facet\_wrap(vars(DAY\_OF\_WEEK),nrow = 1) + theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#### Hypothesis Test

H0 : Mean arrival delay on Sat-Evening <= 0 H1 : Mean arival delay on Sat-Evening > 0

Putting saturday arrival delays in a vector and then running the above hypothesis tests

arrdelaySat <- filter(flightData, DAY\_OF\_WEEK == "Saturday", DEP\_TIME\_OF\_DAY == "Evening")$ARRIVAL\_DELAY  
summary(arrdelaySat)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -69.00 -9.00 3.00 17.89 28.00 1016.00 2186

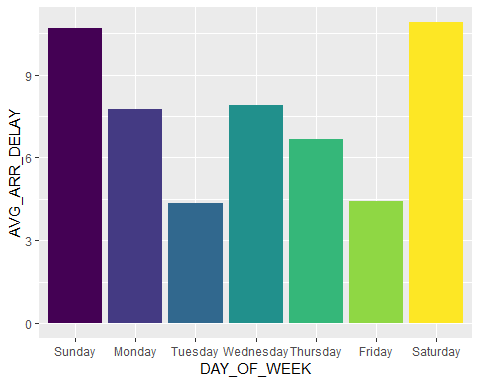
test\_res <- t.test(arrdelaySat, mu = 0, alternative = "greater")  
test\_res

##   
## One Sample t-test  
##   
## data: arrdelaySat  
## t = 66.843, df = 31428, p-value < 2.2e-16  
## alternative hypothesis: true mean is greater than 0  
## 95 percent confidence interval:  
## 17.45432 Inf  
## sample estimates:  
## mean of x   
## 17.89468

### APPENDIX

Plot of arrival delay vs Day of the week

groupByDay <- group\_by(select(flightData, DAY\_OF\_WEEK, ARRIVAL\_DELAY),DAY\_OF\_WEEK)  
groupByDay <- groupByDay[complete.cases(groupByDay),]  
delayByDay <- summarise(groupByDay, AVG\_ARR\_DELAY = mean(ARRIVAL\_DELAY))  
ggplot(data = delayByDay, aes(x = DAY\_OF\_WEEK, y = AVG\_ARR\_DELAY, fill = DAY\_OF\_WEEK)) + geom\_bar(stat = "identity") + guides(fill = FALSE)



#### Hypothesis Test

H0 : Mean arrival delay on Sat <= 0 H1 : Mean arival delay > 0

Putting saturday arrival delays in a vector and then running the above hypothesis tests

arrdelaySat <- filter(flightData, DAY\_OF\_WEEK == "Saturday")$ARRIVAL\_DELAY  
summary(arrdelaySat)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -72.00 -12.00 -2.00 10.91 16.00 1593.00 7786

test\_res <- t.test(arrdelaySat, mu = 0, alternative = "greater")  
test\_res

##   
## One Sample t-test  
##   
## data: arrdelaySat  
## t = 85.686, df = 140890, p-value < 2.2e-16  
## alternative hypothesis: true mean is greater than 0  
## 95 percent confidence interval:  
## 10.7038 Inf  
## sample estimates:  
## mean of x   
## 10.9133