Elshahawy-project

Salma Elshahawy

10/12/2019

Contents

Part 1 - Introduction

I decided to use the dataset of nycflights13::flights: package included with R. This package contains information about all flights that departed from NYC (i.e., EWR, JFK and LGA) in 2013: 336,776 flights with 16 variables. To help understand what causes delays, it also includes a number of other useful datasets: weather, planes, airports, airlines. Source: Bureau of transportation statistics

H0(null hypotithes) -> No associations between departure delay and arrival delay

HA(alternative hypotithes) -> There are associations between departure delay and arrival delay.

- Research Questions:
  - Are the actual departure delay associated with the arrival delay?

Part 2 - Data

- Variables:
  - variable_1 -> dep_delay - independent variable, numerical - categorical
  - outcome -> arr_delay, numerical - numerical - categorical
- There are about 336,776 observation in the given dataset. Each observation represent flight full details.
- This is an observational study. I will draw my conclusions based on analyzing the existing data.

Part 3 - Exploratory data analysis

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggExtra)

library(nycflights13)
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>
```

```
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
##  Max.   :1272.000                      Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
```

```
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
##       hour           minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
# taking a subset
sub_set <- flights[c(6,9,10,16)]
sub_set
```

```
## # A tibble: 336,776 x 4
##    dep_delay arr_delay carrier distance
##        <dbl>     <dbl> <chr>      <dbl>
##  1         2        11 UA          1400
##  2         4        20 UA          1416
##  3         2        33 AA          1089
##  4        -1       -18 B6          1576
##  5        -6       -25 DL           762
##  6        -4        12 UA           719
##  7        -5        19 B6          1065
##  8        -3       -14 EV           229
##  9        -3        -8 B6           944
## 10        -2         8 AA           733
## # ... with 336,766 more rows
```

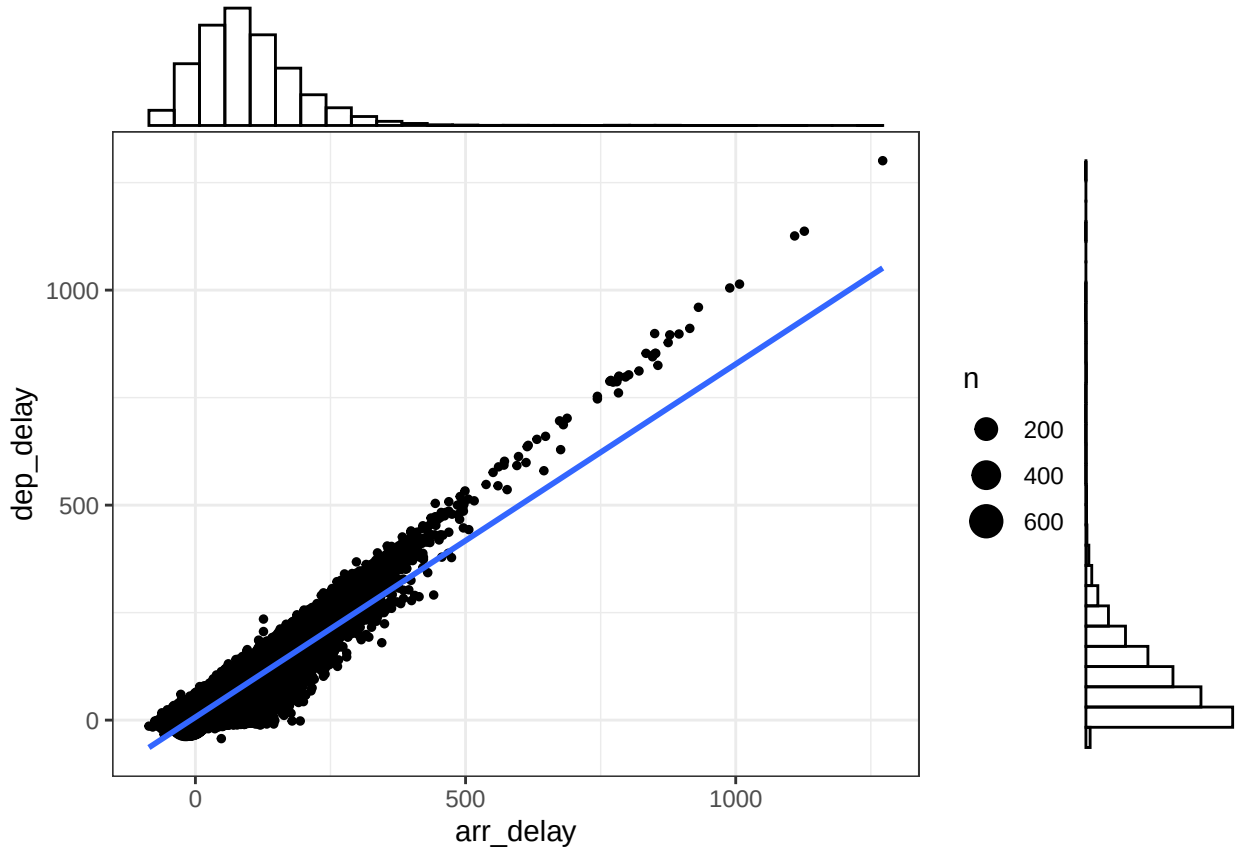
```
summary(sub_set)
```

```
##    dep_delay          arr_delay          carrier             distance
##  Min.   : -43.00   Min.   : -86.000   Length:336776      Min.   :  17
##  1st Qu.:  -5.00   1st Qu.: -17.000   Class :character   1st Qu.: 502
##  Median :  -2.00   Median :  -5.000   Mode  :character   Median : 872
##  Mean   :  12.64   Mean   :   6.895                      Mean   :1040
##  3rd Qu.:  11.00   3rd Qu.:  14.000                      3rd Qu.:1389
##  Max.   :1301.00   Max.   :1272.000                      Max.   :4983
##  NA's   :8255      NA's   :9430
```

```
## get statistical analysis for the whole population
theme_set(theme_bw())  # pre-set the bw theme.
g <- ggplot(sub_set, aes(arr_delay, dep_delay)) +
  geom_count() +
  geom_smooth(method="lm", se=F)

ggMarginal(g, type = "histogram", fill="transparent")
```

```
## Warning: Removed 9430 rows containing non-finite values (stat_sum).
```

```
## Warning: Removed 9430 rows containing non-finite values (stat_smooth).

## Warning: Removed 9430 rows containing non-finite values (stat_sum).

## Warning: Removed 9430 rows containing non-finite values (stat_smooth).
```



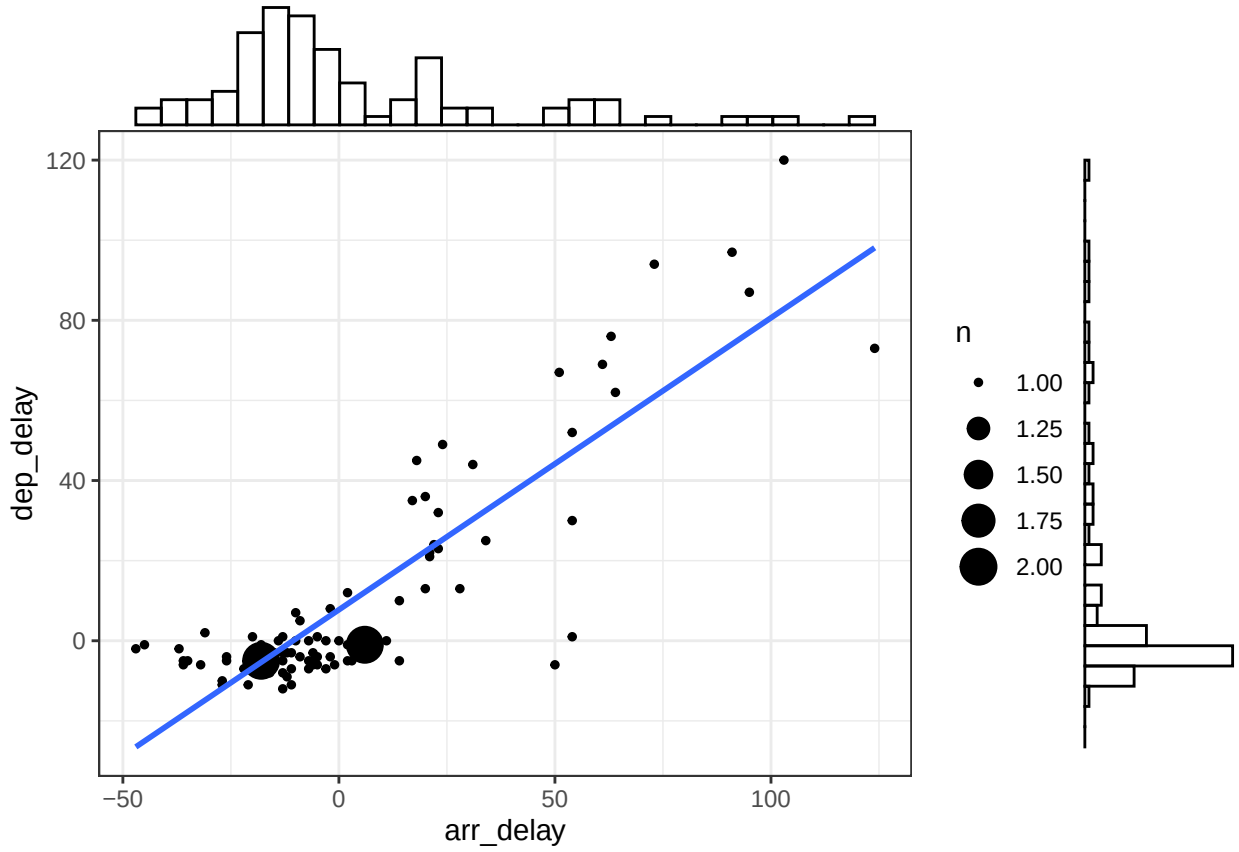

```
## sampling, get summary plots --> statistics for only sample of 100

sample_100 <- sample_n(sub_set, 100)

summary(sample_100)
```

```
##    dep_delay        arr_delay       carrier             distance
##  Min.   :-12.00   Min.   :-47.0   Length:100         Min.   : 173.0
##  1st Qu.: -5.00   1st Qu.:-17.5   Class :character   1st Qu.: 527.0
##  Median : -2.00   Median : -7.0   Mode  :character   Median : 846.0
##  Mean   : 11.51   Mean   :  3.4                      Mean   : 994.2
##  3rd Qu.: 13.00   3rd Qu.: 19.0                      3rd Qu.:1096.0
##  Max.   :120.00   Max.   :124.0                      Max.   :2586.0
##  NA's   :3        NA's   :5
```

```
theme_set(theme_bw())  # pre-set the bw theme.
g <- ggplot(sample_100, aes(arr_delay, dep_delay)) +
  geom_count() +
```

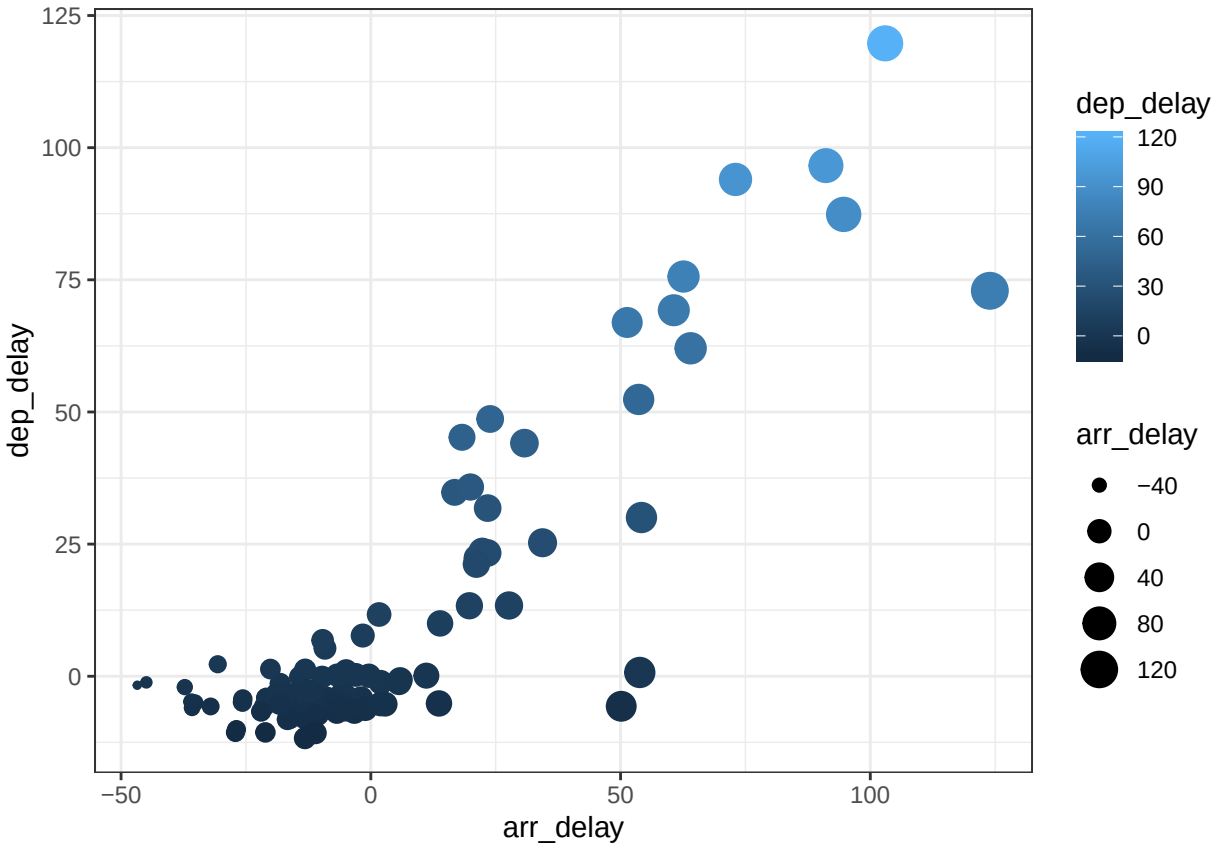
```
  geom_smooth(method="lm", se=F)

ggMarginal(g, type = "histogram", fill="transparent")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_sum).

## Warning: Removed 5 rows containing non-finite values (stat_smooth).

## Warning: Removed 5 rows containing non-finite values (stat_sum).

## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```



Both dep_delay and arr_delap are right skewed distribution.

```
theme_set(theme_bw())  # pre-set the bw theme.
ggplot(sample_100, aes(dep_delay, arr_delay)) +
  geom_jitter(aes(colour = dep_delay, size = arr_delay), na.rm = TRUE) +
  coord_flip()
```



Part 4 - Inference

This dataset doesn't follow the normal distribution. Since n = 100 => which is more than 25 we can do a linear regression model. Let's begin with the correlation which is a statistical tool to measure the level of linear dependence between two variables, that occur in pair

```
cor(sub_set$arr_delay, sub_set$dep_delay, use = "complete.obs")
```

```
## [1] 0.9148028
```

The correlation is very strong as it close to 1 - strong correlation. Now, let's build the linear regression model.

```
linearMod <- lm(arr_delay ~ dep_delay, data=sub_set)  # build linear regression model on full data
print(linearMod)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay, data = sub_set)
##
## Coefficients:
## (Intercept)    dep_delay
##      -5.899        1.019
```

Now that we have built the linear model, we also have established the relationship between the predictor and response in the form of a mathematical formula for arrival delay (arr_delay) as a function for distance. For the above output, you can notice the 'Coefficients' part having two components: Intercept: -5.899, distance: 1.019 These are also called the beta coefficients. In other words,

arr_delay = Intercept + (beta ∗ dep_delay)

```
summary(linearMod)  # model summary
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay, data = sub_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.587  -11.005   -1.883    8.938  201.938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.8994935  0.0330195  -178.7   <2e-16 ***
## dep_delay    1.0190929  0.0007864  1295.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.03 on 327344 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8369
## F-statistic: 1.679e+06 on 1 and 327344 DF,  p-value: < 2.2e-16
```

arr_delay = -5.899 + 1*dep_delay

Part 5 - Conclusion

as a conclusion, I would go with refusing the Null hypotethis that there is no associations between arrival delay and departure delay.

References

Flights database