# Major assignment 1: Your submission

This is your assignment template for [CompX Major assignment 1](https://courses.edx.org/courses/course-v1:AdelaideX+CompX+3T2017/courseware/79355773219d46ce8241e39d7cdce010/73ce1a553bb44943a99c50d71373ceb6/1?activate_block_id=block-v1%253AAdelaideX%252BCompX%252B3T2017%252Btype%2540vertical%252Bblock%25408ebaec2ec3274604abb07669f7c98afb). Save this document on our local machine and include all of your work within the relevant sections. Once you’ve completed all five parts of the assignment, upload the document via the submission area on the “[Submit your assignment](https://courses.edx.org/courses/course-v1:AdelaideX+CompX+3T2017/courseware/79355773219d46ce8241e39d7cdce010/73ce1a553bb44943a99c50d71373ceb6/14?activate_block_id=block-v1%253AAdelaideX%252BCompX%252B3T2017%252Btype%2540vertical%252Bblock%2540bf2a56ee85994664aca89302465ed4bb)” page at the end of Major assignment 1.

# Checklist

* Have you shown all of your working, including probability notation where necessary?
* Have you given all numbers to 3 decimal places?
* Have you included all R output and plots to support your answers where necessary?
* Have you included all of your R code?
* Have you made sure that all plots and tables each have a meaningful caption?

**Quick links:**

[Major assignment 1: Part 1](#Major_assignment_1)

[Major assignment 1: Part 2](#Major_assignment_1_1)

[Major assignment 1: Part 3](#Major_assignment_1_2)

[Major assignment 1: Part 4](#Major_assignment_1_3)

[Major assignment 1: Part 5](#Major_assignment_1_4)

# Major assignment 1: Part 1

1. Read in the dataset [1 point]

Your input code and output code for the flights dataset from the nycflights13 package go here:

* **install.packages(“nycflights13”)**
* **library(flights)**
* **data(flights)**
* **flights**

**# A tibble: 336,776 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**

**7 2013 1 1 555 600 -5. 913**

**8 2013 1 1 557 600 -3. 709**

**9 2013 1 1 557 600 -3. 838**

**10 2013 1 1 558 600 -2. 753**

**# ... with 336,766 more rows, and 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>**

1. Produce a table that summarises the origin variable [3 points]

Your table, code and caption go here:

* **origin\_table <- table(flights$origin)**
* **origin\_table**

**EWR JFK LGA**

**120835 111279 104662**

1. Produce a bar chart of the origin variable [3 points]

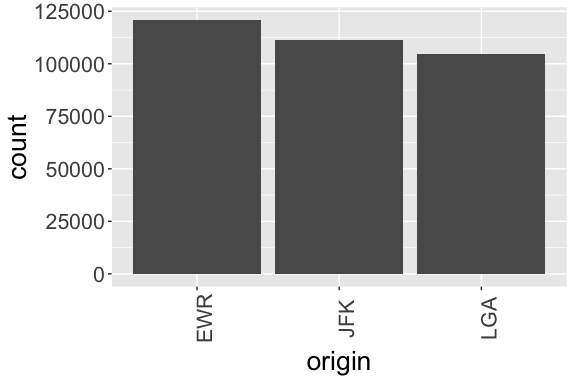
Your bar chart, code and caption go here:

* **library(ggplot2)**
* **ggplot(flights, aes(origin))+**

**+ geom\_bar()+**

**+ theme(text = element\_text(size = 20), axis.text.x = element\_text(angle = 90))**

**The Barchart:**

****

1. Calculate the mean and standard deviation of the distance variable [1 point]

Your code goes here:

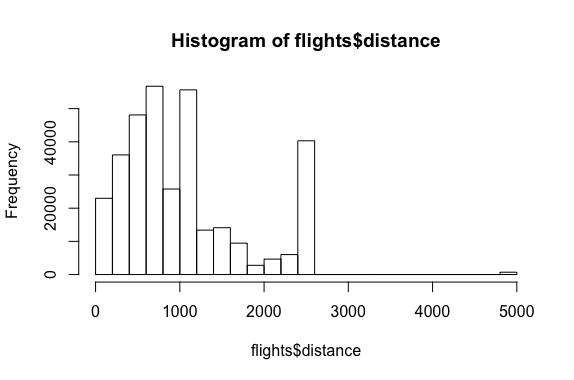
* **mean\_distance <- mean(flights$distance)**
* **mean\_distance**
* **[1] 1039.913**
* **sd\_distance <- sd(flights$distance)**
* **sd\_distance**
* **[1] 733.233**

1. Produce a histogram of the distance variable [2 points]

Your histogram and caption go here:

* **hist(flights$distance)**

**output(histogram):**



# Major assignment 1: Part 2

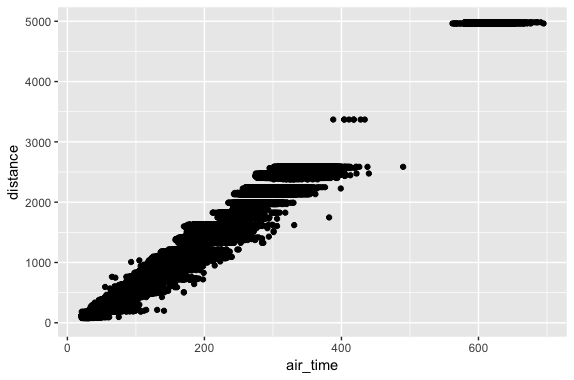
1. Produce a scatterplot [2 points]

Your scatterplot of *air\_time* against *distance*, and caption go here:

**> ggplot(flights)+**

* **geom\_point(mapping = aes(x = air\_time, y = distance))**

**Scatterplot:**

****

1. Produce a contingency table of carrier and origin [2 points]

Your table and caption go here:

* **carrier\_origin\_table <- table(flights$carrier, flights$origin, dnn=c(“carrier”,”origin”))**

**origin**

**carrier EWR JFK LGA**

**9E 1268 14651 2541**

**AA 3487 13783 15459**

**AS 714 0 0**

**B6 6557 42076 6002**

**DL 4342 20701 23067**

**EV 43939 1408 8826**

**F9 0 0 685**

**FL 0 0 3260**

**HA 0 342 0**

**MQ 2276 7193 16928**

**OO 6 0 26**

**UA 46087 4534 8044**

**US 4405 2995 13136**

**VX 1566 3596 0**

**WN 6188 0 6087**

**YV 0 0 601**

1. Produce a conditional table of carrier and origin to find the percentage (%) of American Airlines departing from JFK [2 points]

Your table and caption go here:

* **conditional\_table <- filter(as.data.frame(carrier\_origin),carrier==“AA")**
* **conditional\_table**

**carrier origin Freq**

**1 AA EWR 3487**

**2 AA JFK 13783**

**3 AA LGA 15459**

* **prop.table(conditional\_table$Freq)**

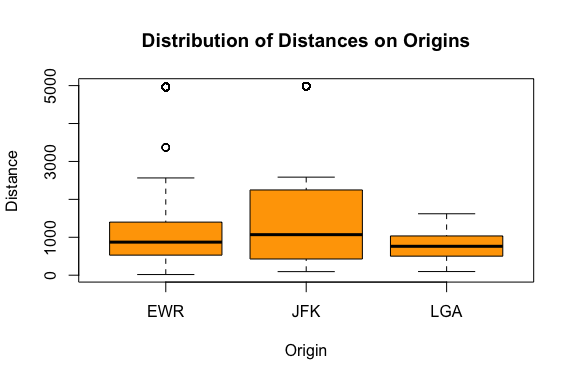
**[1] 0.1065416 0.4211250 0.4723334**

1. Produce side-by-side (one next to the other) boxplots of distance for each origin [2 points]

Your boxplots (ordered by median) and caption go here:

* **boxplot(flights$distance ~ flights$origin, col="orange", main="Distribution of Distances on Origins”, ylab=“Distance", xlab="Origin")**

**Boxplots:**

****

# Major assignment 1: Part 3

1. Show your code (input and output) to calculate how many flights departed from JFK in May 2013 [1 point]

Your code goes here:

* **flights %>% filter(origin==“JFK", year==2013, month==5) %>% nrow()**

**[1] 9397**

1. Show your code (input and output) to identify the carrier and the airport that had the first flight of 2013 [1 point]

Your code goes here:

* **head(flights %>% filter(year==2013) %>% select(carrier, origin, year, month, day, dep\_time),1)**

**# A tibble: 1 x 6**

**carrier origin year month day dep\_time**

**<chr> <chr> <int> <int> <int> <int>**

**1 UA EWR 2013 1 1 517**

1. Show your code (input and output) to calculate the total metres travelled by flights departing NYC in 2013 (you may assume 1609.34 metres/mile) [1 point]

Your code goes here:

* **miles\_travelled <-sum(flights %>% filter(year==2013) %>% select(distance))**
* **miles\_travelled**

**[1] 350217607**

* **meters\_travelled <- miles\_travelled\*1609.34**
* **meters\_travelled**

**[1] 563619203649**

1. Show your code (input and output) to calculate the total distance in miles travelled by flights departing from NYC in 2013 for airlines with Inc. in the name [1 point]

Your code goes here:

- ***Finding the airlines which have Inc. in their name:***

* **airlines\_with\_Inc <- filter(airlines, grepl(" Inc.”,name))**
* **airlines\_with\_Inc**

**# A tibble: 11 x 2**

**carrier name**

**<chr> <chr>**

**1 9E Endeavor Air Inc.**

**2 AA American Airlines Inc.**

**3 AS Alaska Airlines Inc.**

**4 DL Delta Air Lines Inc.**

**5 EV ExpressJet Airlines Inc.**

**6 F9 Frontier Airlines Inc.**

**7 HA Hawaiian Airlines Inc.**

**8 OO SkyWest Airlines Inc.**

**9 UA United Air Lines Inc.**

**10 US US Airways Inc.**

**11 YV Mesa Airlines Inc.**

* ***Creating a vector of the carrier column in airlines\_with\_Inc:***
* **carrier\_with\_Inc <- airlines\_with\_Inc$carrier**
* **carrier\_with\_Inc**

**[1] "9E" "AA" "AS" "DL" "EV" "F9" "HA" "OO" "UA" "US" “YV"**

* ***Calculating the miles travelled by these airlines:***
* **miles\_flights\_withInc\_travelled <- sum(flights %>% filter(year==2013, carrier %in% carrier\_with\_Inc) %>% select(distance))**
* **miles\_flights\_withInc\_travelled**

**[1] 249500641**

# Major assignment 1: Part 4

1. Show your code (input and output) to calculate how many standard deviations greater than the mean the largest distance in the flights dataset is [2 points]

Your code goes here:

* ***Finding the largest distance in the flights dataset:***
* **largest\_dist <- head(sort(flights$distance, decreasing = TRUE),1)**
* **largest\_dist**

**[1] 4983**

* ***Calculating how many standard deviations away from the mean:***
* **times\_away\_from\_mean <- (largest\_dist - mean(flights$distance))/sd(flights$distance)**
* **times\_away\_from\_mean**

**[1] 5.377673**

# Major assignment 1: Part 5

1. Show your code (input and output) to give the 95% confidence interval for the mean distance for flights departing NYC. [2 points]

Your code goes here:

* **sample\_mean <- mean(flights$distance,na.rm = TRUE)**
* **sample\_sd <- sd(flights$distance,na.rm = TRUE)**
* **N <- length(flights$distance)**
* **t <- qt(p=0.025,df=N-1, lower.tail = FALSE)**
* **lwr <- sample\_mean - (t \*sample\_sd/sqrt(N))**
* **upr <- sample\_mean + (t \* sample\_sd/sqrt(N))**
* **ci <- c(lwr=lwr, upr=upr)**
* **ci**

**lwr upr**

**1037.436 1042.389**

1. Show your code (input and output) and a captioned table with the 95% confidence interval for the mean distance for each carrier [4 points]

Your code goes here:

* ***Grouping by carrier:***
* **by\_carrier <- group\_by(flights,carrier)**
* ***Summarizing and saving in a new table:***
* **carrier\_table <- summarise(by\_carrier, distance\_mean=mean(distance, na.rm = TRUE), distance\_sd=sd(distance, na.rm = TRUE), N=n(), t=qt(p=0.025, df=N-1, lower.tail = FALSE), lwr=distance\_mean-(t\*distance\_sd/sqrt(N)), upr=distance\_mean+(t\*distance\_sd/sqrt(N)))**

***- Selecting only carrier, distance\_mean, lwr and upr variables from carrier\_table:***

* **carrier\_table <- select(carrier\_table, carrier, distance\_mean, lwr,upr)**
* **carrier\_table**

**# A tibble: 16 x 4**

**carrier distance\_mean lwr upr**

**<chr> <dbl> <dbl> <dbl>**

**1 9E 530. 526. 535.**

**2 AA 1340. 1333. 1347.**

**3 AS 2402. 2402. 2402.**

**4 B6 1069. 1063. 1075.**

**5 DL 1237. 1231. 1243.**

**6 EV 563. 561. 565.**

**7 F9 1620. 1620. 1620.**

**8 FL 665. 659. 670.**

**9 HA 4983. 4983. 4983.**

**10 MQ 570. 567. 572.**

**11 OO 501. 426. 575.**

**12 UA 1529. 1523. 1536.**

**13 US 553. 545. 561.**

**14 VX 2499. 2497. 2502.**

**15 WN 996. 989. 1004.**

**16 YV 375. 362. 388.**