# Major assignment 1: Your submission

This is your assignment template for [CompX Major assignment 1](https://courses.edx.org/courses/course-v1:AdelaideX+CompX+3T2018/courseware/79355773219d46ce8241e39d7cdce010/73ce1a553bb44943a99c50d71373ceb6/1?activate_block_id=block-v1%3AAdelaideX%2BCompX%2B3T2018%2Btype%40vertical%2Bblock%408ebaec2ec3274604abb07669f7c98afb). 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+3T2018/courseware/79355773219d46ce8241e39d7cdce010/73ce1a553bb44943a99c50d71373ceb6/14?activate_block_id=block-v1%3AAdelaideX%2BCompX%2B3T2018%2Btype%40vertical%2Bblock%40377ecddd31af4739a1bf66e8f88d9eba)” 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:

|  |
| --- |
| # Import Libraries  ```{r}  library(tidyverse)  library(nycflights13)  ```  # Read in the dataset  ```{r}  flights <- nycflights13::flights  flights  View(flights)  ``` |

Table . Confirmation of dataset load

|  |
| --- |
| # A tibble: 336,776 x 19  year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time arr\_delay carrier flight tailnum origin dest  <int> <int> <int> <int> <int> <dbl> <int> <int> <dbl> <chr> <int> <chr> <chr> <chr>  1 2013 1 1 517 515 2 830 819 11 UA 1545 N14228 EWR IAH  2 2013 1 1 533 529 4 850 830 20 UA 1714 N24211 LGA IAH  3 2013 1 1 542 540 2 923 850 33 AA 1141 N619AA JFK MIA  4 2013 1 1 544 545 -1 1004 1022 -18 B6 725 N804JB JFK BQN  5 2013 1 1 554 600 -6 812 837 -25 DL 461 N668DN LGA ATL  6 2013 1 1 554 558 -4 740 728 12 UA 1696 N39463 EWR ORD  7 2013 1 1 555 600 -5 913 854 19 B6 507 N516JB EWR FLL  8 2013 1 1 557 600 -3 709 723 -14 EV 5708 N829AS LGA IAD  9 2013 1 1 557 600 -3 838 846 -8 B6 79 N593JB JFK MCO  10 2013 1 1 558 600 -2 753 745 8 AA 301 N3ALAA LGA ORD  # … with 336,766 more rows, and 5 more variables: 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:

|  |
| --- |
| ```{r}  origin\_summary <- flights %>%  select(origin) %>%  group\_by(origin) %>%  summarise(n=n()) %>%  arrange(desc(n))  origin\_summary  ``` |

Table . Origin Summary

|  |
| --- |
| # A tibble: 3 x 2  origin n  <chr> <int>  1 EWR 120835  2 JFK 111279  3 LGA 104662 |

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

Your bar chart, code and caption go here:

|  |
| --- |
| ```{r}  flights %>%  ggplot(aes(origin)) +  ggtitle("Flight Origin") +  geom\_bar()  ``` |

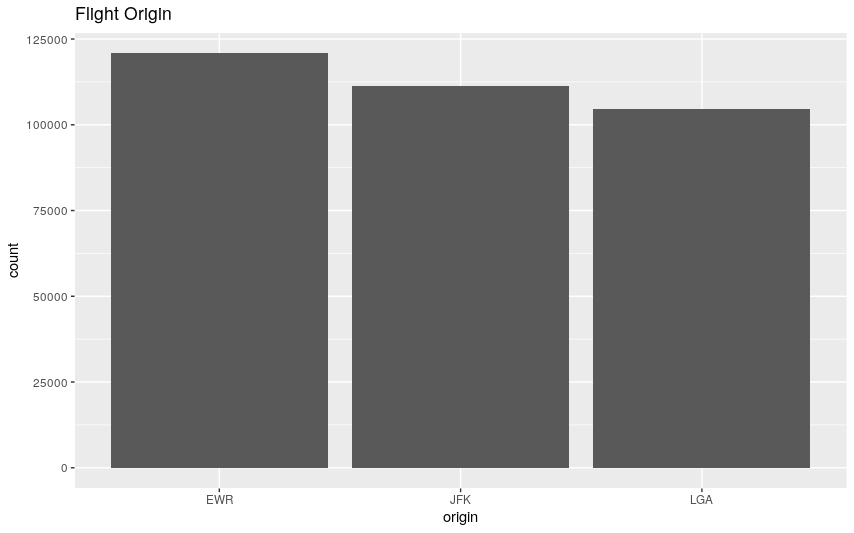


Figure . Number of Departures by Origin

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

Your code goes here:

|  |
| --- |
| ```{r}  summary(flights$distance)  ``` |

Table . Distance Variable Summary

|  |
| --- |
| Min. 1st Qu. Median Mean 3rd Qu. Max.  17 502 872 1040 1389 4983 |

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

Your histogram and caption go here:

|  |
| --- |
| ```{r}  flights %>%  ggplot(aes(distance)) +  ggtitle("Distance between airports, in miles") +  geom\_histogram(bins=30)  ``` |

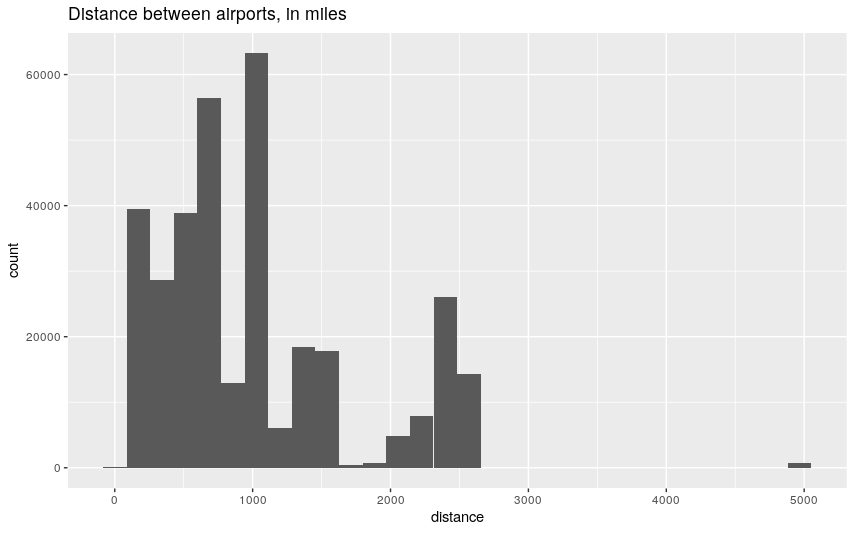


Figure . Histogram of the Distance between Airports (miles)

# Major assignment 1: Part 2

1. Produce a scatterplot [2 points]

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

|  |
| --- |
| ```{r}  airtime\_distance <- flights %>%  select(air\_time, distance) %>%  drop\_na()  airtime\_distance %>%  select(air\_time, distance) %>%  drop\_na() %>%  ggplot(mapping = aes(x=distance, y=air\_time)) +  ggtitle("Air Time vs Distance") +  geom\_point() +  geom\_smooth()  ``` |

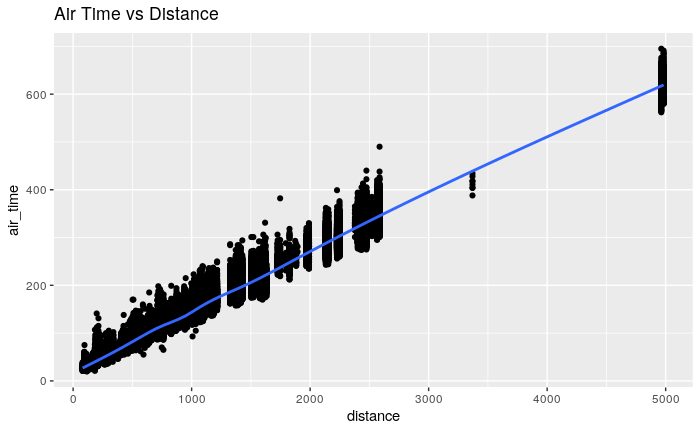


Figure . Scatterplot comparing Air Time vs Distance

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

Your table and caption go here:

|  |
| --- |
| ```{r}  contingency\_flights\_origin <- table(flights$carrier, flights$origin)  contingency\_flights\_origin  View(nycflights13::airlines)  ``` |

Table . Contingency of Carrier and Origin

|  |
| --- |
| 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:

|  |
| --- |
| ```{r}  contingency\_flights\_origin\_prop <- prop.table(contingency\_flights\_origin)  contingency\_flights\_origin\_prop \* 100  ``` |

Table . Carrier and Origin Contingency in %

|  |
| --- |
| EWR JFK LGA  9E 0.376511390 4.350369385 0.754507447  AA 1.035406323 4.092631304 4.590291470  AS 0.212010357 0.000000000 0.000000000  B6 1.946991472 12.493764401 1.782193505  DL 1.289284272 6.146815688 6.849359812  EV 13.046951089 0.418082049 2.620733069  F9 0.000000000 0.000000000 0.203399292  FL 0.000000000 0.000000000 0.968002470  HA 0.000000000 0.101551179 0.000000000  MQ 0.675820130 2.135841034 5.026486448  OO 0.001781600 0.000000000 0.007720265  UA 13.684763760 1.346295460 2.388531249  US 1.307991068 0.889315153 3.900515476  VX 0.464997506 1.067772050 0.000000000  WN 1.837423094 0.000000000 1.807432834  YV 0.000000000 0.000000000 0.178456897 |

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:

|  |
| --- |
| ```{r}  flights %>%  select(origin, distance) %>%  drop\_na() %>%  ggplot(aes(distance)) +  geom\_boxplot() +  ggtitle("Distance Summary by Origin") +  coord\_flip() +  facet\_wrap(~origin)  ``` |

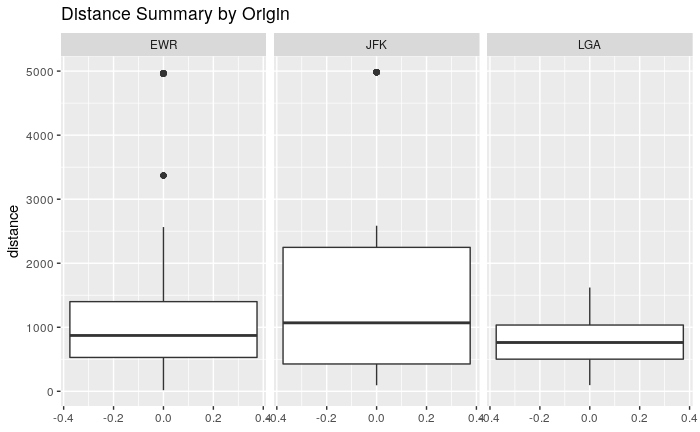


Figure . Summary of Distance for each Origin

# 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:

|  |
| --- |
| ```{r}  flights %>%  filter(origin == "JFK", month == 5) %>%  count()  ``` |

Table . Departures from JFK (May 2013)

|  |
| --- |
| # A tibble: 1 x 1  n  <int>  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:

|  |
| --- |
| ```{r}  flights %>%  arrange(year, month, day, dep\_time) %>%  left\_join(airlines, by="carrier") %>%  select(carrier, name, origin) %>%  head(1)  ``` |

Table . First flight of 2013

|  |
| --- |
| # A tibble: 1 x 3  carrier name origin  <chr> <chr> <chr>  1 UA United Air Lines Inc. EWR |

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:

|  |
| --- |
| ```{r}  miles\_meters <- 1609.34  miles\_travelled <- flights %>%  select(distance) %>%  sum()  miles\_travelled \* miles\_meters  ``` |

Table . Total Distance travelled by flights departing NYC in 2013 (meters)

|  |
| --- |
| [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:

|  |
| --- |
| ```{r}  flights %>%  left\_join(airlines, by="carrier") %>%  select(name, distance) %>%  filter(str\_detect(name, "Inc.")) %>%  select(distance) %>%  sum()  ``` |

Table Total distance travelled by flights departing from NYC in 2013 for airlines with Inc. in the name (miles)

|  |
| --- |
| [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:

|  |
| --- |
| ```{r}  flight\_distance <- flights %>%  select(distance) %>%  drop\_na()  # get the summary statistics  distance\_mean <- mean(flight\_distance$distance)  distance\_sd <- sd(flight\_distance$distance)  distance\_max <- max(flight\_distance$distance)  # calculate the number of sd away  (distance\_max - distance\_mean) / distance\_sd  ``` |

Table . The number of Standard Deviations the maximum distance is away from the 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:

|  |
| --- |
| ```{r}  # get the total number of flights  distance\_n <- length(flight\_distance$distance)  # using student t-distribution 95% confidence interval t value  t <- qt(p=0.025, df=distance\_n-1, lower.tail=FALSE)  # calculate the lower and upper values  lwr <- distance\_mean - t \* distance\_sd / sqrt(distance\_n)  upr <- distance\_mean + t \* distance\_sd / sqrt(distance\_n)  c(lwr = lwr, upr = upr)  ``` |

Table . 95% Confidence Interval for flights departing NYC

|  |
| --- |
| 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:

|  |
| --- |
| ```{r}  flights %>%  select(carrier, distance) %>%  drop\_na() %>%  group\_by(carrier) %>%  summarise(mean=mean(distance), sd=sd(distance), n=n()) %>%  mutate(t=qt(p=0.025, df=n-1, lower.tail=FALSE)) %>%  mutate(  lower=mean - t \* sd / sqrt(n),  upper=mean + t \* sd / sqrt(n)  ) %>%  left\_join(airlines, by="carrier") %>%  select(carrier, name, mean, upper, lower)  ``` |

Table . 95% confidence interval for the mean distance for each carrier

|  |
| --- |
| # A tibble: 16 x 5  carrier name mean upper lower  <chr> <chr> <dbl> <dbl> <dbl>  1 9E Endeavor Air Inc. 530. 535. 526.  2 AA American Airlines Inc. 1340. 1347. 1333.  3 AS Alaska Airlines Inc. 2402 2402 2402  4 B6 JetBlue Airways 1069. 1075. 1063.  5 DL Delta Air Lines Inc. 1237. 1243. 1231.  6 EV ExpressJet Airlines Inc. 563. 565. 561.  7 F9 Frontier Airlines Inc. 1620 1620 1620  8 FL AirTran Airways Corporation 665. 670. 659.  9 HA Hawaiian Airlines Inc. 4983 4983 4983  10 MQ Envoy Air 570. 572. 567.  11 OO SkyWest Airlines Inc. 501. 575. 426.  12 UA United Air Lines Inc. 1529. 1536. 1523.  13 US US Airways Inc. 553. 561. 545.  14 VX Virgin America 2499. 2502. 2497.  15 WN Southwest Airlines Co. 996. 1004. 989.  16 YV Mesa Airlines Inc. 375. 388. 362. |

# Appendix 1: R Notebook

|  |
| --- |
| ---  title: "Major Assignment 1"  output: html\_notebook  ---  # Import Libraries  ```{r}  library(tidyverse)  library(nycflights13)  ```  # Part 1  ## 1.1 Read in the dataset  ```{r}  flights <- nycflights13::flights  flights  View(flights)  ```  ## 1.2 Produce a table that summarises the origin variable  ```{r}  origin\_summary <- flights %>%  select(origin) %>%  group\_by(origin) %>%  summarise(n=n()) %>%  arrange(desc(n))  origin\_summary  ```  ## 1.3 Produce a bar chart of the origin variable  ```{r}  flights %>%  ggplot(aes(origin)) +  ggtitle("Flight Origin") +  geom\_bar()  ```  ## 1.4 Calculate the mean and standard deviation of the distance variable  ```{r}  summary(flights$distance)  ```  ## 1.5 Produce a histogram of the distance variable  ```{r}  flights %>%  ggplot(aes(distance)) +  ggtitle("Distance between airports, in miles") +  geom\_histogram(bins=30)  ```  # Part 2  ## 2.1 Produce a scatterplot  ```{r}  airtime\_distance <- flights %>%  select(air\_time, distance) %>%  drop\_na()  airtime\_distance %>%  select(air\_time, distance) %>%  drop\_na() %>%  ggplot(mapping = aes(x=distance, y=air\_time)) +  ggtitle("Air Time vs Distance") +  geom\_point() +  geom\_smooth()  ```  ```{r}  cor(airtime\_distance$air\_time, airtime\_distance$distance)  ```  ## 2.2 Produce a contingency table of carrier and origin  ```{r}  contingency\_flights\_origin <- table(flights$carrier, flights$origin)  contingency\_flights\_origin  View(nycflights13::airlines)  ```  ## 2.3 Produce a conditional table of carrier and origin to find the percentage (%) of American Airlines departing from JFK  ```{r}  contingency\_flights\_origin\_prop <- prop.table(contingency\_flights\_origin)  contingency\_flights\_origin\_prop \* 100  ```  # 2.4. Produce side-by-side (one next to the other) boxplots of distance for each origin  ```{r}  flights %>%  select(origin, distance) %>%  drop\_na() %>%  ggplot(aes(distance)) +  geom\_boxplot() +  ggtitle("Distance Summary by Origin") +  coord\_flip() +  facet\_wrap(~origin)  ```  # Part 3  ## 3.1 Calculate how many flights departed from JFK in May 2013  ```{r}  flights %>%  filter(origin == "JFK", month == 5) %>%  count()  ```  ## 3.2 Identify the carrier and the airport that had the first flight of 2013  ```{r}  flights %>%  arrange(year, month, day, dep\_time) %>%  left\_join(airlines, by="carrier") %>%  select(carrier, name, origin) %>%  head(1)  ```  ## 3.3 Calculate the total metres travelled by flights departing NYC in 2013 (you may assume 1609.34 metres/mile)  ```{r}  miles\_meters <- 1609.34  miles\_travelled <- flights %>%  select(distance) %>%  sum()  miles\_travelled \* miles\_meters  ```  ## 3.4 Calculate the total distance in miles travelled by flights departing from NYC in 2013 for airlines with Inc. in the name  ```{r}  flights %>%  left\_join(airlines, by="carrier") %>%  select(name, distance) %>%  filter(str\_detect(name, "Inc.")) %>%  select(distance) %>%  sum()  ```  # Part 4  ## 4.1 Calculate how many standard deviations greater than the mean the largest distance in the flights dataset is.  ```{r}  flight\_distance <- flights %>%  select(distance) %>%  drop\_na()  # get the summary statistics  distance\_mean <- mean(flight\_distance$distance)  distance\_sd <- sd(flight\_distance$distance)  distance\_max <- max(flight\_distance$distance)  # calculate the number of sd away  (distance\_max - distance\_mean) / distance\_sd  ```  # Part 5  ## 5.1 Calculate the 95% confidence interval for the mean distance for flights departing NYC  ```{r}  # get the total number of flights  distance\_n <- length(flight\_distance$distance)  # using student t-distribution 95% confidence interval t value  t <- qt(p=0.025, df=distance\_n-1, lower.tail=FALSE)  # calculate the lower and upper values  lwr <- distance\_mean - t \* distance\_sd / sqrt(distance\_n)  upr <- distance\_mean + t \* distance\_sd / sqrt(distance\_n)  c(lwr = lwr, upr = upr)  ```  ## 5.2 Calculate 95% confidence interval for the mean distance for each carrier  ```{r}  flights %>%  select(carrier, distance) %>%  drop\_na() %>%  group\_by(carrier) %>%  summarise(mean=mean(distance), sd=sd(distance), n=n()) %>%  mutate(t=qt(p=0.025, df=n-1, lower.tail=FALSE)) %>%  mutate(  lower=mean - t \* sd / sqrt(n),  upper=mean + t \* sd / sqrt(n)  ) %>%  left\_join(airlines, by="carrier") %>%  select(carrier, name, mean, upper, lower)  ``` |