quiz_01

August 4, 2025

1 Quiz 1

---

1. **Data Classification** *(9)*

Consider the following R dataset detailing the attributes for different flights departing from New York City including year, month, day, actual departure time, scheduled departure time, departure delay, actual arrival time, scheduled arrival time, arrival dely, carrier name, flight number, tail number, origin airport, destination airport, air time, distance traveled, and total flight time (in hours and minutes). Classify each variable in the dataset as one of the following: Discrete Quantitative, Continuous Quantitative, Qualitative, and Categorical.

[2]:
```
# Load Packages
library(dplyr)
library(nycflights13)

# Load Dataset
data <- flights
str(data)
```

```
tibble [336,776 × 19] (S3: tbl_df/tbl/data.frame)
 $ year          : int [1:336776] 2013 2013 2013 2013 2013 2013 2013 2013 2013
2013 …
 $ month         : int [1:336776] 1 1 1 1 1 1 1 1 1 1 …
 $ day           : int [1:336776] 1 1 1 1 1 1 1 1 1 1 …
 $ dep_time      : int [1:336776] 517 533 542 544 554 554 555 557 557 558 …
 $ sched_dep_time: int [1:336776] 515 529 540 545 600 558 600 600 600 600 …
 $ dep_delay     : num [1:336776] 2 4 2 -1 -6 -4 -5 -3 -3 -2 …
 $ arr_time      : int [1:336776] 830 850 923 1004 812 740 913 709 838 753 …
 $ sched_arr_time: int [1:336776] 819 830 850 1022 837 728 854 723 846 745 …
 $ arr_delay     : num [1:336776] 11 20 33 -18 -25 12 19 -14 -8 8 …
 $ carrier       : chr [1:336776] "UA" "UA" "AA" "B6" …
 $ flight        : int [1:336776] 1545 1714 1141 725 461 1696 507 5708 79 301
…
 $ tailnum       : chr [1:336776] "N14228" "N24211" "N619AA" "N804JB" …
 $ origin        : chr [1:336776] "EWR" "LGA" "JFK" "JFK" …
 $ dest          : chr [1:336776] "IAH" "IAH" "MIA" "BQN" …
 $ air_time      : num [1:336776] 227 227 160 183 116 150 158 53 140 138 …
```

```
 $ distance      : num [1:336776] 1400 1416 1089 1576 762 …
 $ hour          : num [1:336776] 5 5 5 5 6 5 6 6 6 6 …
 $ minute        : num [1:336776] 15 29 40 45 0 58 0 0 0 0 …
 $ time_hour     : POSIXct[1:336776], format: "2013-01-01 05:00:00" "2013-01-01
05:00:00" …
```

2. **Data Summary** *(10)*

Using the flights dataset filtered out for John F. Kennedy Airport for Delta Airlines flights on Christmas Eve (24th December), summarise measure of location (mean, median, mode), dispersion (range, inter-quartile range, standard deviation), and shape (skewness, kurtosis) for departure as well as arrival delay. *(8)*

[ ]:
```r
# Dataset filtered out for John F. Kennedy Airport for Delta Airlines flights␣
↪on Christmas Eve (24th December)
data <- flights %>% filter(origin=="JFK", carrier=="DL", month==12, day==24)
data.frame(dep_delay=data$dep_delay, arr_delay=data$arr_delay)
```

A data.frame: 50 × 2

dep_delay <dbl>	arr_delay <dbl>
0	-5
2	25
-5	3
-3	-23
-3	19
-2	-35
-3	-10
-1	12
-4	-26
8	3
25	23
17	-5
-1	-9
19	35
-5	-16
2	9
-4	-13
0	-8
66	78
0	1
-3	3
17	10
0	-22
20	21
-2	-15
29	7
6	-24
-9	-19
37	8
-1	0
-6	-18
-2	-25
3	-5
0	-25
-4	-18
-1	-21
8	-29
-4	-26
-3	-12
-2	-27
-6	-28
0	-3
8	15
-2	-9
11	-11
8	-6
-2	-7
5	7
-5	-17
-6	-20

3. **Probability Analysis** *(5)*

Prove that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4. **Data Sampling** *(8)*

a. For the following randomly sampled data from the flights dataset, compute bias and standard error for the estimator on arrival delay. *(5)*

b. Using the Archery analogy discussed in the class, draw a representative target board to comment upon the accuracy and precision of the estimator. *(3)*

[ ]:
```r
# Randomly sampled data from flights dataset
P <- flights$arr_delay
m <- 30
n <- 1000
z <- mean(P, na.rm=TRUE)
Z <- vector("numeric", m)
for (i in 1:m) {
  set.seed(i)
  I <- order(runif(length(P)))[1:n]
  S <- P[I]
  Z[i] <- mean(S, na.rm=TRUE)
}
data.frame(parameter=z, estimator=Z, error=round(Z-z, digits=3))
message("Bias: ", round(mean(Z-z, na.rm=TRUE), digits=3))
message("Standard Error: ", round(sd(Z, na.rm=TRUE), digits=3))
```

A data.frame: 30 × 3

parameter <dbl>	estimator <dbl>	error <dbl>
6.895377	6.297828	-0.598
6.895377	6.357513	-0.538
6.895377	5.903392	-0.992
6.895377	8.699272	1.804
6.895377	5.126016	-1.769
6.895377	4.781538	-2.114
6.895377	7.001031	0.106
6.895377	4.726522	-2.169
6.895377	8.027664	1.132
6.895377	6.911614	0.016
6.895377	6.270769	-0.625
6.895377	9.189938	2.295
6.895377	6.405128	-0.490
6.895377	4.688660	-2.207
6.895377	5.128601	-1.767
6.895377	5.763131	-1.132
6.895377	7.532508	0.637
6.895377	6.298371	-0.597
6.895377	7.128999	0.234
6.895377	7.781186	0.886
6.895377	7.725410	0.830
6.895377	5.174180	-1.721
6.895377	6.946336	0.051
6.895377	9.529897	2.635
6.895377	5.490256	-1.405
6.895377	7.908436	1.013
6.895377	4.631308	-2.264
6.895377	7.881391	0.986
6.895377	6.876797	-0.019
6.895377	7.303850	0.408

```
Bias: -0.246
```

```
Standard Error: 1.356
```

5. **Hypothesis Testing** *(8)*

Test the following hypothesis for Delta Airlines flights from John F. Kennedy Airport on Christmas Eve

a. departure delay is greater than 4 minutes

b. arrival delay is less than -5 minutes

Note, make appropriate assumptions, develop the null and alternate hypotheses, evaluate the test statistic, present the threshold value and consequently, make appropriate inferences.