Homework 2

By signing my name, I agree to abide by the Stern Code of Conduct _____

Question 1

```
library(dplyr)
library(ggplot2)
library(nycflights13)
library(ggthemes) # for theme_tufte(), not required
```

a. How many flights are there from Hawaiian Airlines? Which airline had the most number of flights? Write your answers here.

```
table(flights$carrier)
```

```
##
##    9E    AA    AS    B6    DL    EV    F9    FL    HA    MQ    OO    UA
## 18460 32729   714 54635 48110 54173   685  3260   342 26397    32 58665
##    US    VX    WN    YV
## 20536  5162 12275   601
```

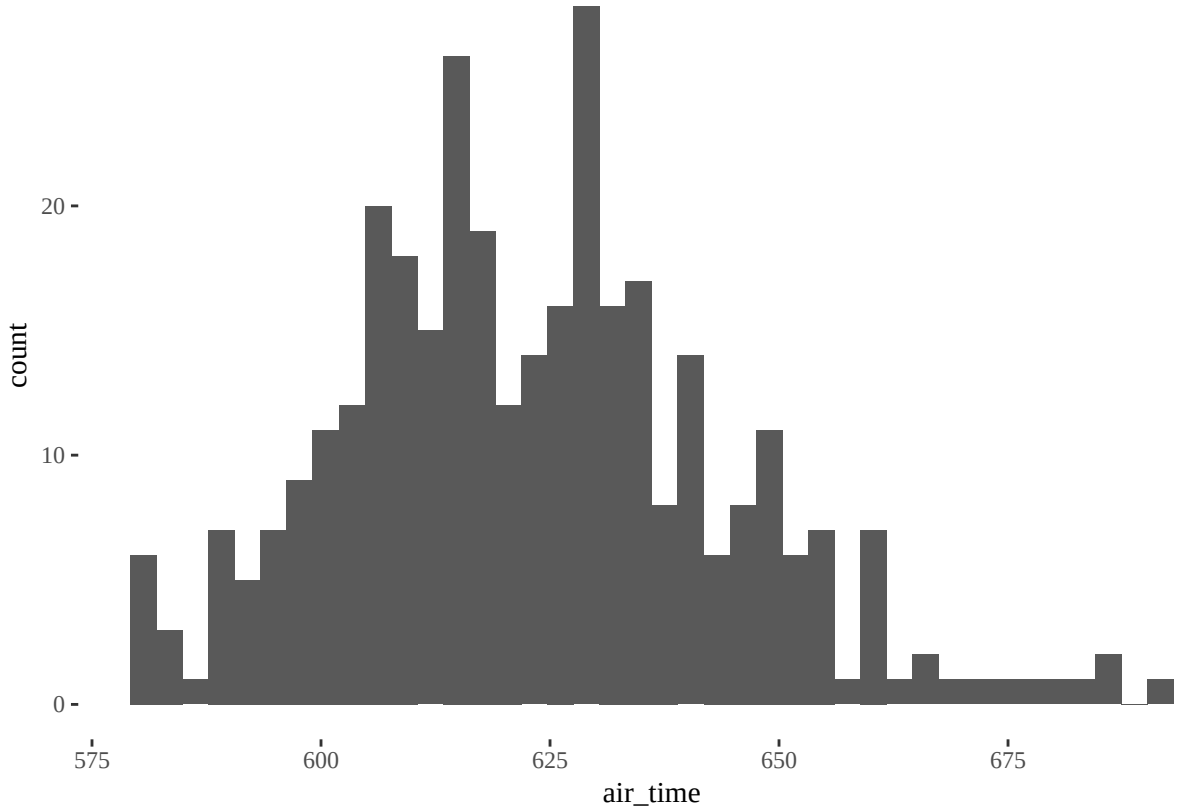
The Hawaiian Airlines carrier code is "HA," and according to `table` there were **342 flights** by HA. The airline with the most is UA (United) with 58665 flights.

- Following the example in the book, use `filter()` to create a data frame with all Hawaiian Airlines flights (*instead of Alaskan Airlines*). Use this data frame to complete the parts below.

```
all_hawaiian_flights <- flights %>% filter(carrier == "HA")
```

- Create a histogram of the `air_time` variable using `ggplot()` and `geom_histogram()`.

```
ggplot(all_hawaiian_flights, aes(air_time)) + geom_histogram(bins = 40) + theme_tufte()
```



(No points would be deducted for anyone choosing a different number of bins in the histogram, that's fine.)

b. Upload your plot on Classes under the File Exchange tab.

c. If you wanted to describe a histogram using summary statistics, which summaries would you use, and why?

The **range** would help describe the horizontal axis which goes from about 575 to about 700. It's not easy to guess with any precision just by looking at the graph, but the **mean or median** or some measure of central tendency might be around 625. The mean might be a bit larger than the median, because the distribution is skewed toward the right. The **standard deviation** or **quartiles** may help describe that the bulk of the distribution seems to be between a little more than 600 and a little less than 650.

d. What are the mean and standard deviation of the `air_time` variable for these Hawaiian Airline flights?

```
# dplyr method:
all_hawaiian_flights %>% summarize(mean = mean(air_time), SD = sd(air_time))
```

```
## # A tibble: 1 x 2
##    mean    SD
##   <dbl> <dbl>
## 1   623  20.7
```

```
# Dollar sign method:
mean(all_hawaiian_flights$air_time)
```

```
## [1] 623.0877
```

```r
sd(all_hawaiian_flights$air_time)
```

```
## [1] 20.68882
```

e. What percent of these lie within one, two, and three standard deviation of the mean? Include 4 digits of accuracy for each, e.g. 12.34%

```r
x <- all_hawaiian_flights$air_time
xbar <- mean(x)
s <- sd(x)
mean(abs(x - xbar) <= s)
```

```
## [1] 0.6871345
```

```r
mean(abs(x - xbar) <= 2*s)
```

```
## [1] 0.9561404
```

```r
mean(abs(x - xbar) <= 3*s)
```

```
## [1] 0.9912281
```

68.71%, 95.61%, and 99.12% respectively

Question 2

a. Consider a game involving tossing a coin three times, and use set notation (e.g. { item1, item2, ... }) to answer the following: What is the sample space S? What is the event E_1 = "the first toss is H"? What is the intersection of E_1 and E_2 = "the second toss is H"?

$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

$E_1 = \{HHH, HHT, HTH, HTT\}$

$E_1 \cap E_2 = \{HHH, HHT\}$

b. A 6-sided die is rolled 10 times. What is the probability it comes up six every time? What is the probability of no sixes? What is the probability every roll is less or equal to 5?

Using independence, it comes up 6 every time with probability $(1/6)^{10}$.

For no sixes, use the subtraction rule: $1 - (1/6)^{10}$.

All less or equal to five: same as no sixes!

c. Every time I roll a 20-sided die there is a 1/20 chance it lands on 20. So if I roll it 10 times, the probability that at least one of these will be a 20 is 10/20 = 1/2. True or false? Explain your answer.

False. The addition rule only applies to disjoint events. Getting a 20 on the first roll and getting a 20 on the second roll are not disjoint, so we can't add their probabilities. Also, if this argument was true, then it would imply that rolling the dice more than 20 times gives a probability bigger than 1, which is impossible.

d. What is the equation for the conditional probability of E_2 given E_1? Suppose $P(E_1) > P(E_2)$ and $P(E_1 \cap E_2) > 0$. Which is larger, $P(E_2|E_1)$ or $P(E_1|E_2)$, or is there not enough information to tell? Explain, and also draw a Venn diagram as an example.

$P(E_2|E_1) = P(E_2 \cap E_1)/P(E_1)$.

If we also write out $P(E_1|E_2) = P(E_2 \cap E_1)/P(E_2)$, we see they both have the same numerator, but different denominators. Hence, the one with a larger denominator will be a smaller number. Since $P(E_1) > P(E_2)$, this means $P(E_1|E_2)$ is larger. A Venn diagram should have two circles which have some overlap, and the E_1 circle should be larger.

e. The mortality rates from a hypothetical drug trial are in the table below. Is survival independent from treatment/placebo, adherence/non-adherence, both of these, or neither? Explain.

	Treatment	Placebo
Adherers	15%	15%
Non-adherers	25%	25%
Total	20%	20%

Independence means the probability of survival (i.e. 1 minus probability of mortality) does not depend on which group a person belongs to. There is no change in probability between the treatment and placebo groups, so survival is *independent* of treatment/placebo. However, there is a change in probability between the adherers and non-adherers, so survival and adherence are *dependent*.