CT5163: Assignment 4

Question 1

Part a

First, I created two vectors with the given heights of players in each team, and then conducted an independent two-sample t-test using the *t.test()* function in R.

R code and output:

```
> TeamA <- c(175, 160, 163, 190, 178, 169, 180, 171, 185, 176)
> TeamB <- c(173, 162, 158, 186, 179, 162, 163, 172, 168, 172)
> # conduct t-test
> t.test(TeamA, TeamB, var.equal = TRUE)

        Two Sample t-test

data:  TeamA and TeamB
t = 1.2926, df = 18, p-value = 0.2125
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.251987 13.651987
sample estimates:
mean of x mean of y
    174.7     169.5
```

Part b

The null hypothesis is that there is no difference in the height of players in the two sports teams.

Part c

The t value is 1.2926 on 18 degrees of freedom with the corresponding p-value of 0.2125.

Part d

We cannot reject the null hypothesis at 5% level of significance. There is no statistically significant difference in the height of players in Team A and Team B.

Question 2

Part a

First, I created the 'daily' data frame using the code from the lecture slides. Then, I created an extra column in the 'daily' data frame called 'Month' by using R functions *mutate()* (from the 'tidyverse' package) and *month()* (from the 'lubridate' package).

R code:

```r
# recreate the 'daily' data frame from the lecture slides
daily <- flights %>%
  mutate(date = make_date(year, month, day)) %>%
  group_by(date) %>%
  summarise(n = n()) %>%
  mutate(wday = wday(date, label = TRUE))

# add a variable for the month of the year of each observation
daily <- daily %>%
  mutate(Month = month(date))
```

Part b

July had the highest total number of flights (4,989) on Wednesday. October had the lowest total number of flights (2,732) on Saturday.

R code and output:

```r
> # find the total number of flights for each month-weekday pair
> flights_month_wday <- daily %>%
+   group_by(Month, wday) %>%
+   summarise(Total_flights = sum(n))
```

```r
> # What month of the year had the highest total number of flights on Wednesday?
> flights_month_wday %>%
+   # filter for Wednesdays only
+   filter(wday == "Wed") %>%
+   # sort in decreasing order
+   arrange(desc(Total_flights)) %>%
+   # find the month with the highest number of flights
+   head(1)
# A tibble: 1 x 3
# Groups:   Month [1]
  Month wday  Total_flights
  <dbl> <ord>         <int>
1     7 Wed            4989
```

```r
> # What month of the year had the lowest total number of flights on Saturday?
> flights_month_wday %>%
+   # filter for Saturdays only
+   filter(wday == "Sat") %>%
+   # sort in increasing order
+   arrange(Total_flights) %>%
+   # find the month with the lowest number of flights
+   head(1)
# A tibble: 1 x 3
# Groups:   Month [1]
  Month wday  Total_flights
  <dbl> <ord>         <int>
1    10 Sat            2732
```

Question 3

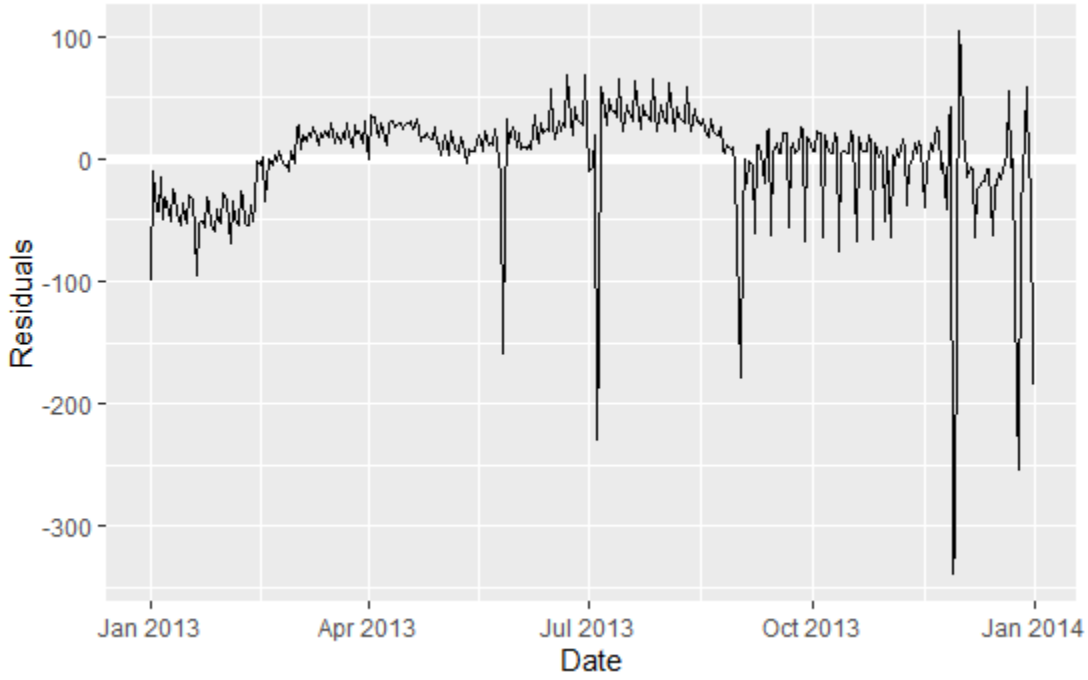
Part a

I created the linear model using the *lm()* function in R.

R code:

```
# create a linear model that predicts the number of flights each day based on weekday and month
mod <- lm(n ~ wday + Month, data = daily)
```

Part b



R code:

```
# find model residuals
daily <- daily %>%
  add_residuals(mod)

# plot the residuals over time
daily %>%
  ggplot(aes(date, resid)) +
  geom_ref_line(h = 0) +
  geom_line() +
  labs(x = "Date", y = "Residuals")
```

Part c

R code and output:

```
> # the average absolute value of the residuals
> mean(abs(daily$resid))
[1] 30.24384
> # the maximum absolute value of the residuals
> max(abs(daily$resid))
[1] 339.9715
```