Lab 4 Tidy data (50 points)

Fang Fang

Delivery:

1.Answer 8 questions in a word document.

2.You should also submit a .csv file.

3.You have 10 bonus points at the end of this lab.

Steps:

1. Import the dataset.

```
library(tidyverse)
air<- read_csv("data/air_raw_upload.csv")
```

Question 1: (5 points)

How many rows and columns does this dataset have?

Answer: 1440 rows and 17 columns

Question 2: (5 points)

Is this dataset tidy or messy? Why?

Answer: Messy. The station and variables are mixed up, and the column names are not a certain attribute, they are certain values.

Question 3:(5 points)

How many rows and columns does air_v1 have?

```
air_v1 <- pivot_longer(air,cols=(-c("Date","Time")),
        names_to="mix",values_to="measured_value")
```

Question 4:(5 points)

Examine air_v2. Why is this dataset a messy data?

```
air_v2 <- separate(air_v1, sep=" ",mix, c("Station","Meteorological variable"))
```

Answer: Yes it is still messy. The meteorological variables should have their own columns.

Question 5: (10 points)

Based on your answer above, select one of the following functins `separate`, `unite`, `pivot_longer` to make air_v2 clean. Name the output as air_v3.

```
air_v3 <- pivot_wider(air_v2,names_from="Meteorological variable",
    values_from="measured_value")
```

Question 6: (5 points)

After you execute the function you choose above, How many rows and columns do you have for this clean dataset (air_v3)?

Answer: 4320 rows and 8 columns.

Question 7:(10 points)

Among the three stations, which one has the highest average temperature?

Hint: Once your dataset is clean, let's use tapply to calculate the average by 3 stations. You need to exclude the NAs.

Note you need to provide both your `tapply` code and answers for question 7.

```
tapply(air_v3$TEMP,air_v3$Station,mean, na.rm=T)
```

Answer: BERESFIELD: 24.22

4. Export and submit your assignment.

Make sure you have a subfolder called "output". Then execute the code below.

```
write_csv(air_v3,"output/Fang_lab4.csv")
```

Question 8:(5 points)

Submit this .csv file.

This is the end of Lab 4.

Bonus: (10 points)

1) Examine the output called air_v3. Does it still contain NAs? (5 points) Answer: Yes it contains NAs since some stations do not have all the measurements

2) Try to use `drop_na` function to exclude all the NAs in air_v3. Name the output air_v4. Export air_v4 as a csv and turn it in compass (5 points).

Hint: you can use `?drop_na` to know more about this function.

```
air_v4 <- drop_na(air_v3)
write_csv(air_v4,"output/air_v4.csv")
```

Answer: the drop_na function will exclude entire row if any NA appears