Flights & Weather datasets (filter & ggplot)

Luke Geel

9/23/2020

"Question 1"

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
```

"a"

```
a <- dim(flights)
a
```

```
## [1] 336776     19
```

```
b <- dim(weather)
b
```

```
## [1] 26115    15
```

Flight (Observations , Variables) (336776 19) "Weather (Observations , Variables) (26115 15)"b"

```
?flights
```

"Talinum: Plane tail number" "Flight: Flight number" "Carrier: Two letter carrier abbreviation" "dep_delay: Departure delay" "arr_delay: Arrival delay" "c"

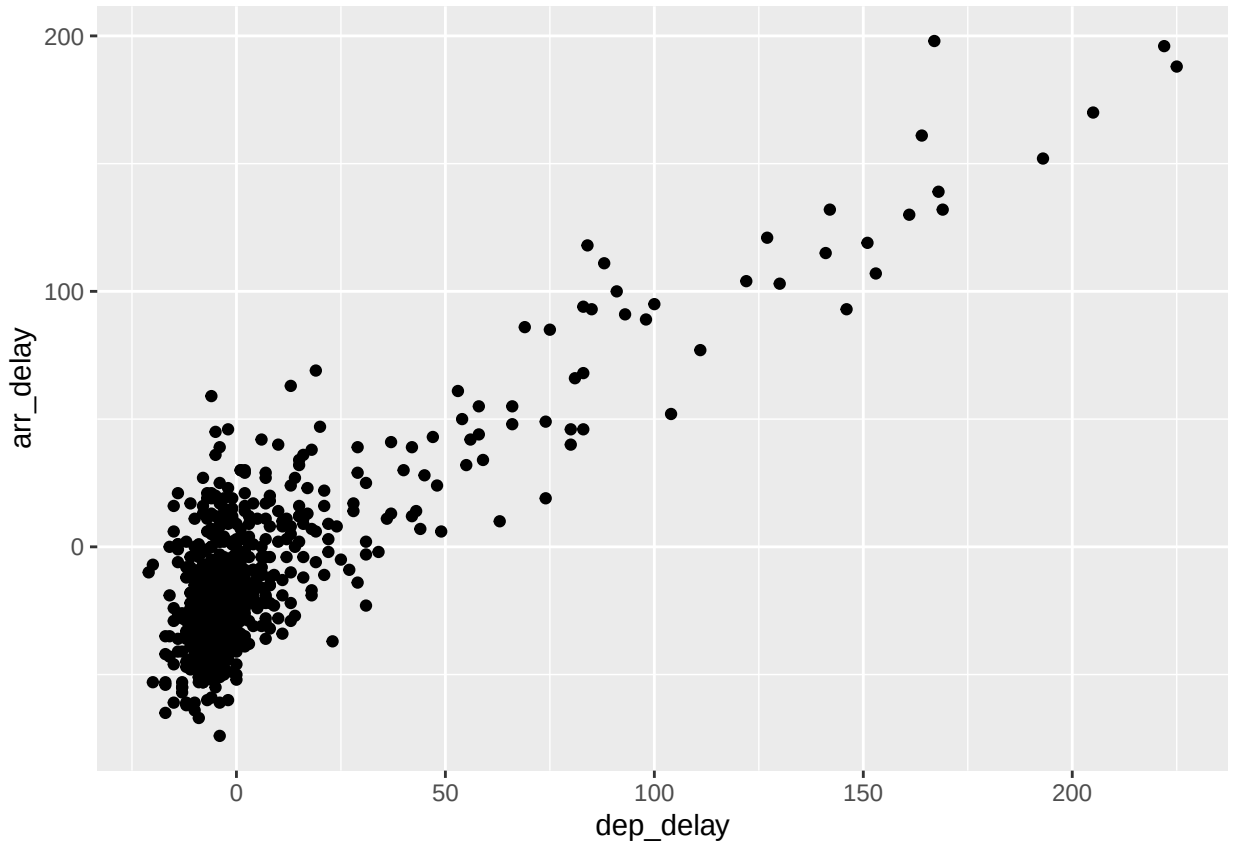
```
?weather
```

"visib: Visibility in miles" "time_hour: Date and hour of the recording as a POSIXct date" "temp: Termperature in F" "Question 2"

```
alaska_flights <- flights %>% filter(carrier == "AS") %>% filter(!is.na(arr_delay))
```

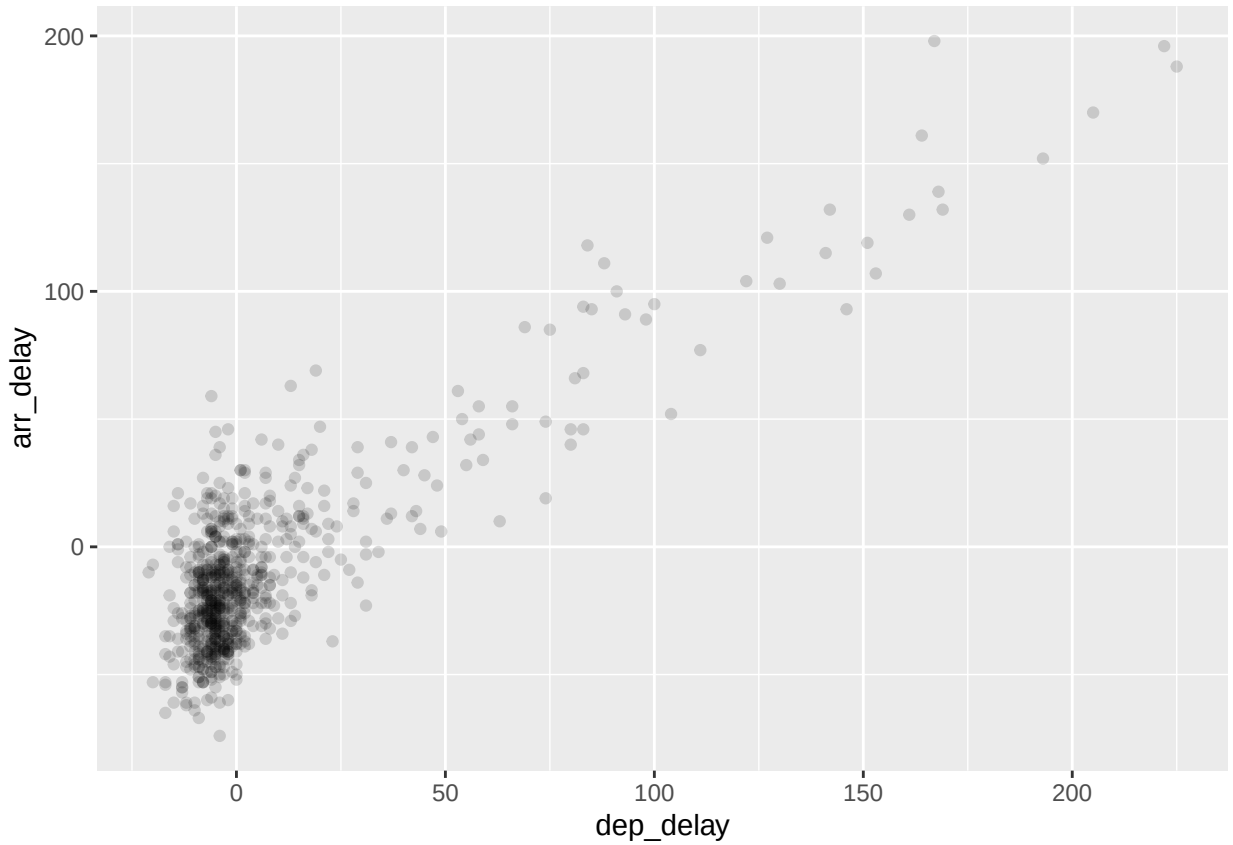
"A"

```
ggplot(alaska_flights, aes(x = dep_delay, y = arr_delay))+
  geom_point()
```



"Based on tbis scatterplot I noticed that as arr_delay increases, so does dep_delay"

```
ggplot(alaska_flights, aes(x = dep_delay, y = arr_delay))+
  geom_point(alpha=0.15)
```

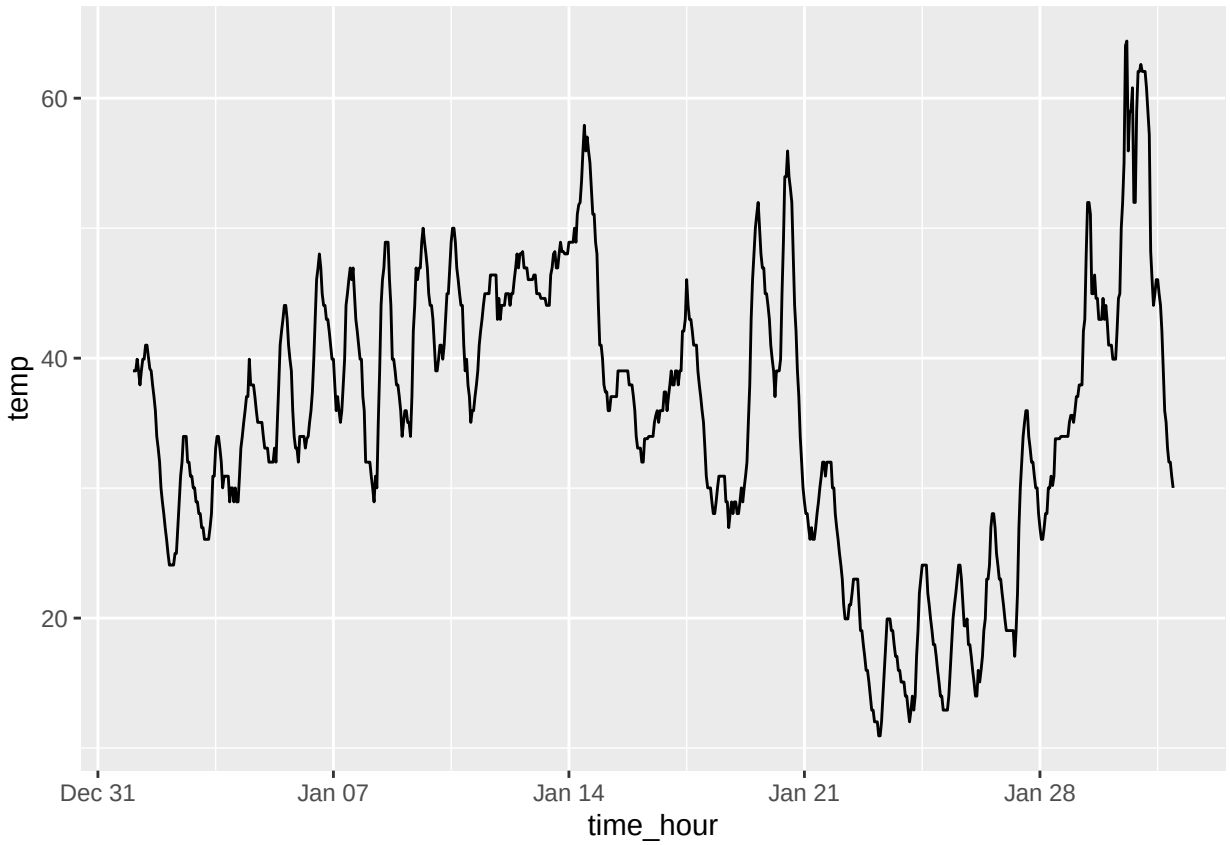


"B" "There is a cluster around (0,0) which is equivalent to no departure delay and no arrival delay. To fix this over cluster I will change the transparency of all points to make it easier to see overplotted clusters." "Question 3"

```
early_january_weather <- weather %>% filter(origin == "EWR" & month == 1)
```

"a"

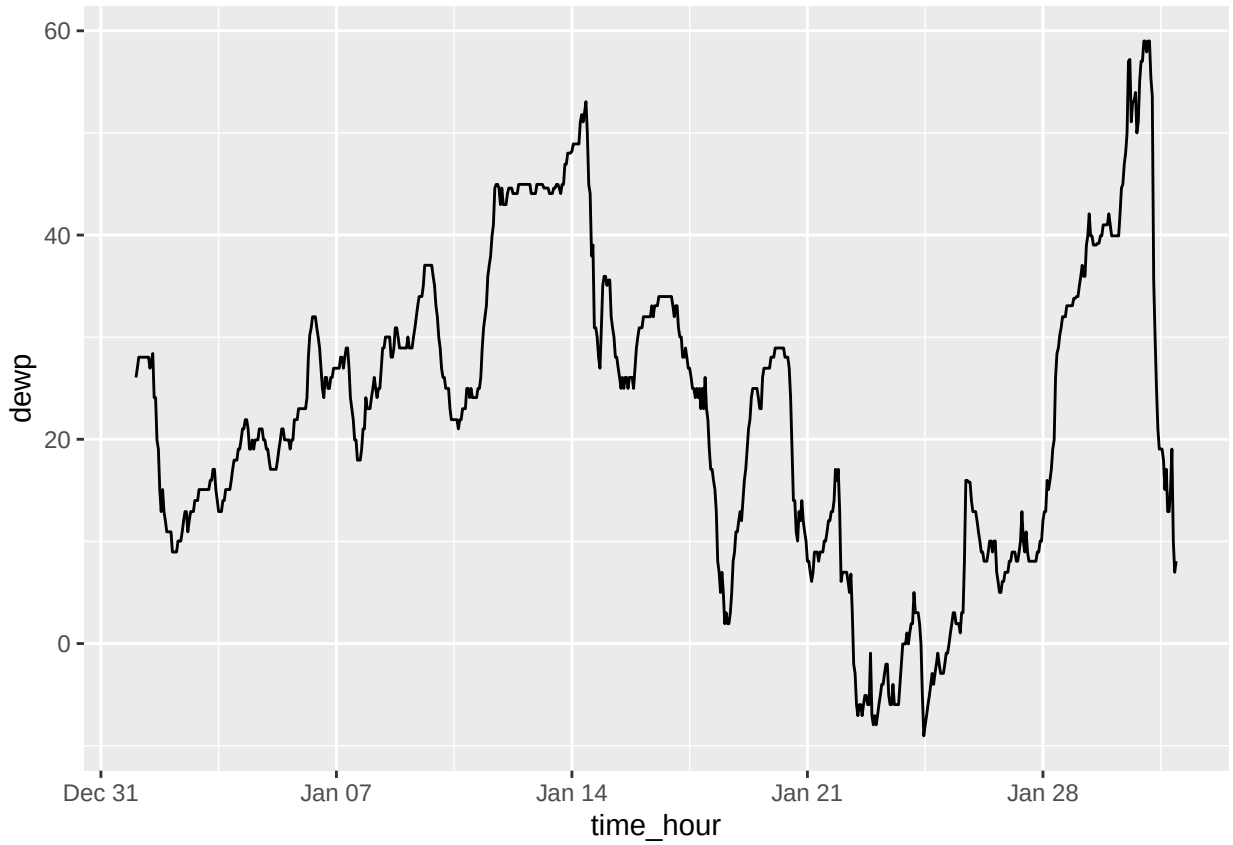
```
ggplot(data = early_january_weather) +
  geom_line(mapping = aes(x = time_hour, y = temp))
```



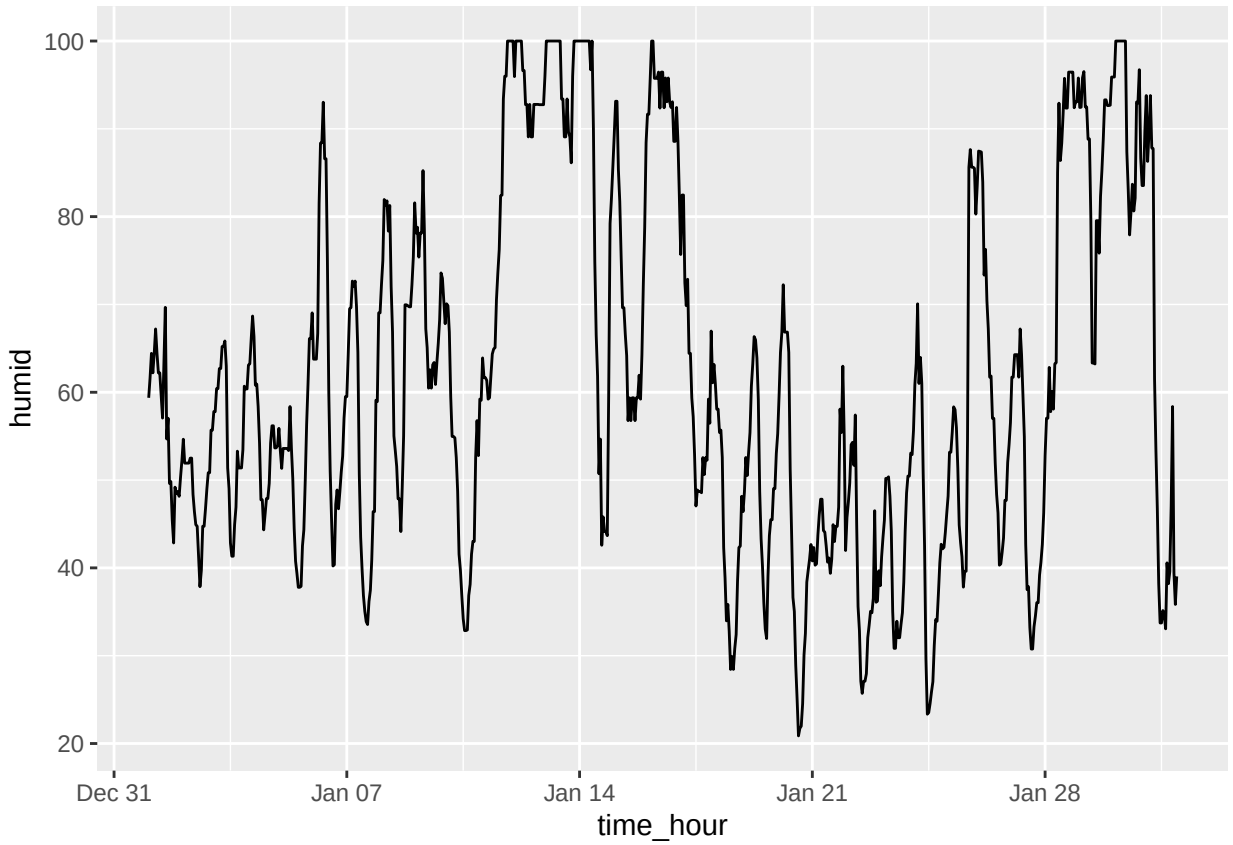
"From this plot I can see that that from Dec 31 to Jan 31 the temperature stayed somewhat consistent except for the week of Jan 21-28 when the temperature dropped by around 15 degrees."

"b"

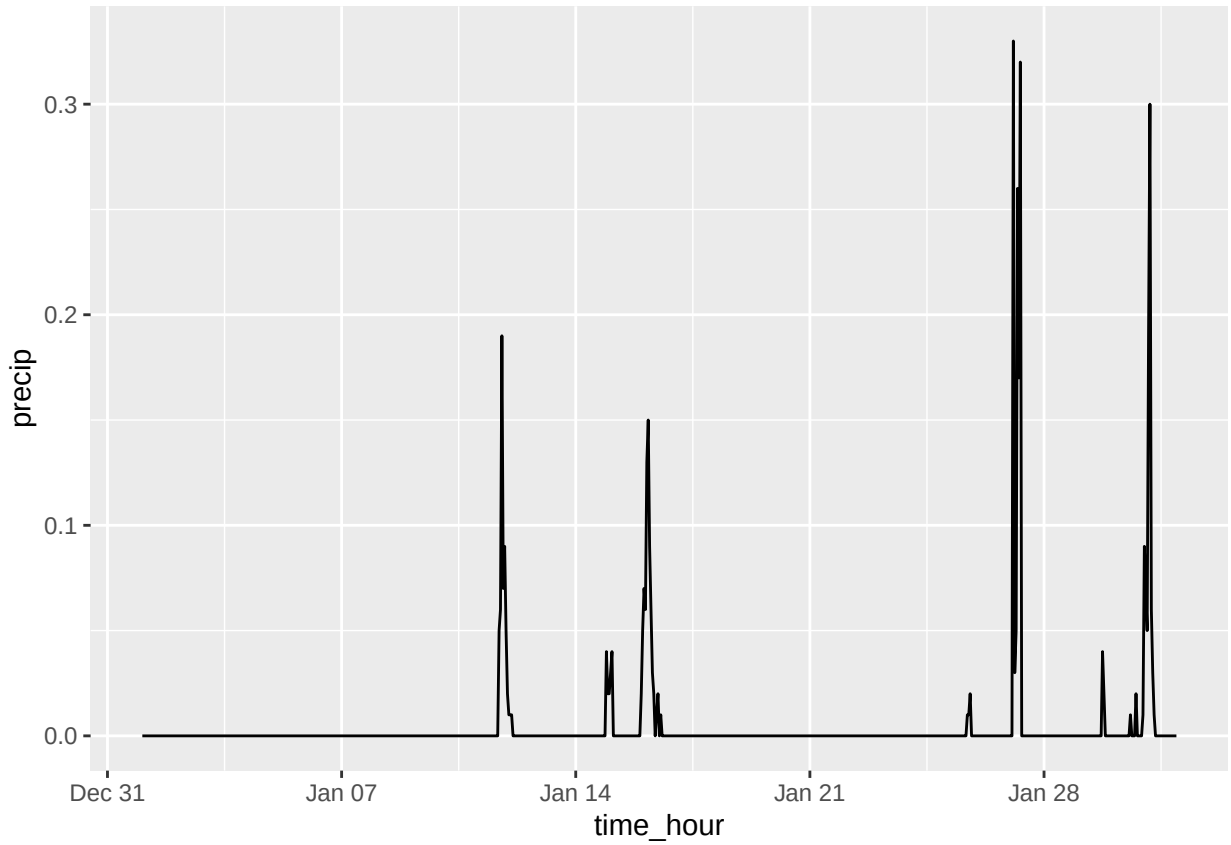
```
ggplot(data = early_january_weather) +
  geom_line(mapping = aes(x = time_hour, y = dewp))
```



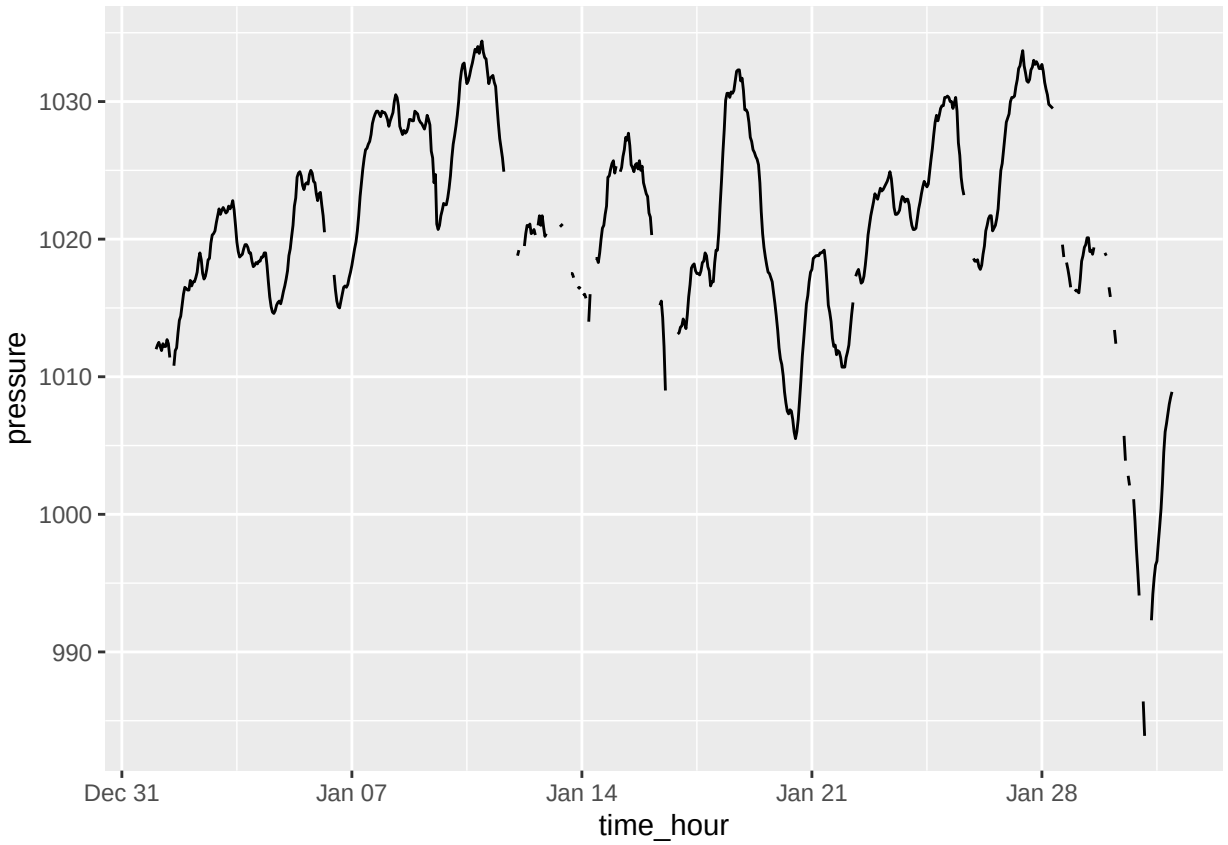

```
ggplot(data = early_january_weather) +
  geom_line(mapping = aes(x = time_hour, y = humid))
```




```
ggplot(data = early_january_weather) +
  geom_line(mapping = aes(x = time_hour, y = precip))
```




```
ggplot(data = early_january_weather) +
  geom_line(mapping = aes(x = time_hour, y = pressure))
```



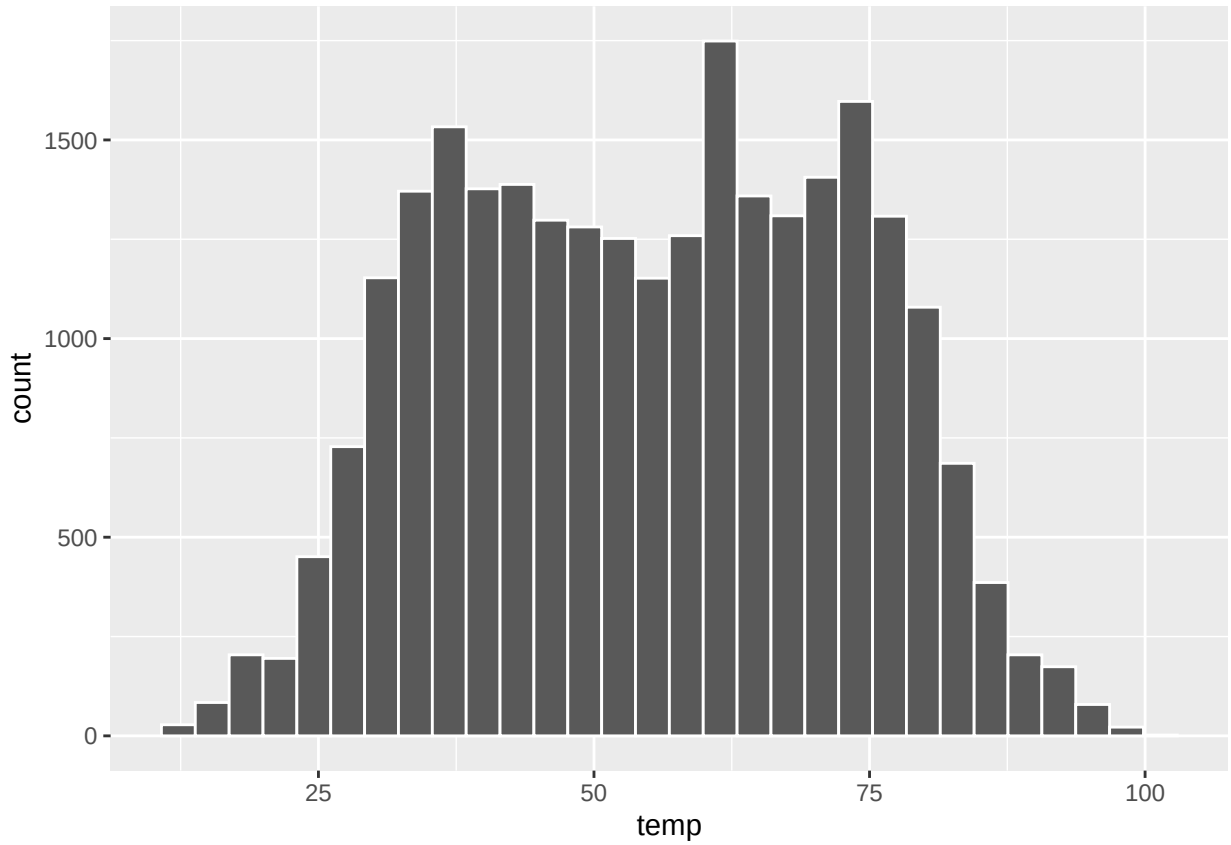
"Question 4"

```
weather <- weather %>% filter(!is.na(temp))
```

"A"

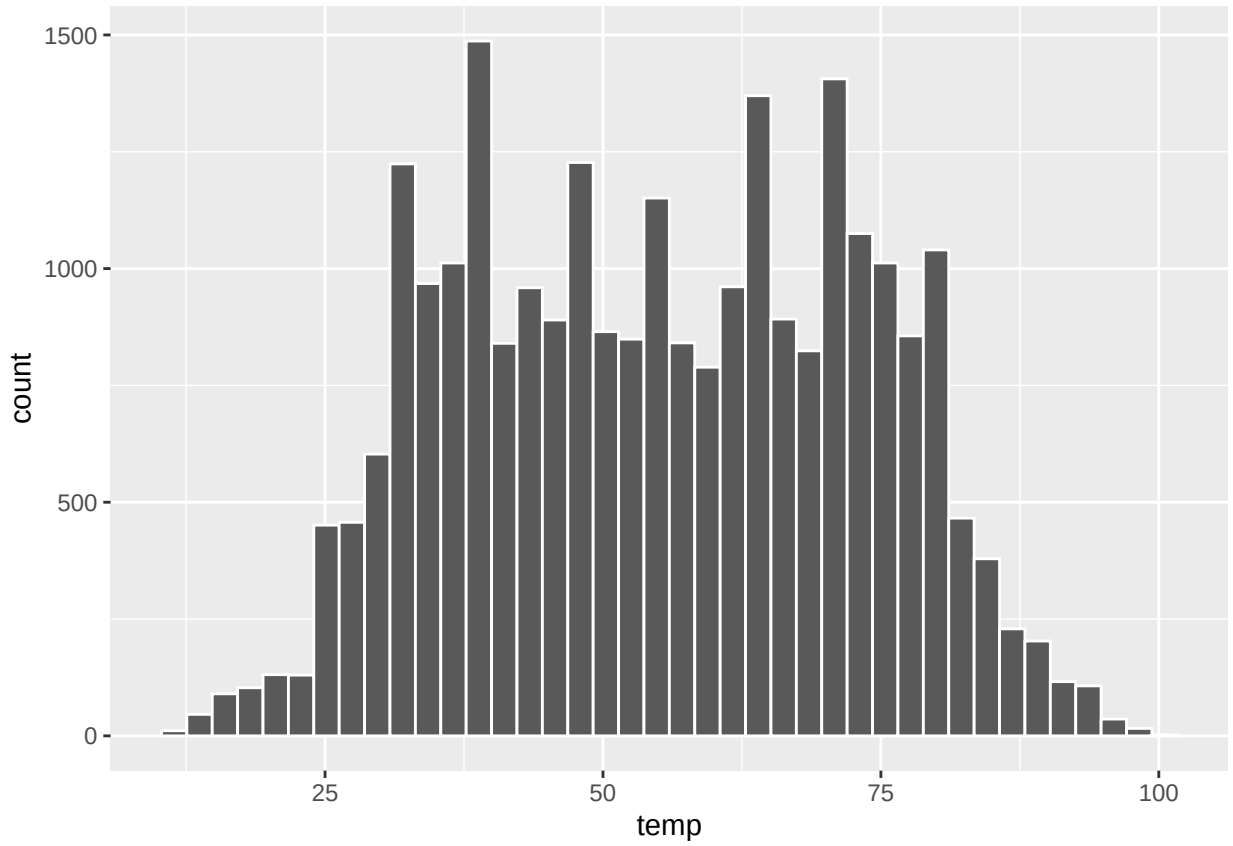
```
ggplot(weather)+
  geom_histogram(mapping = aes(x=temp),color="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

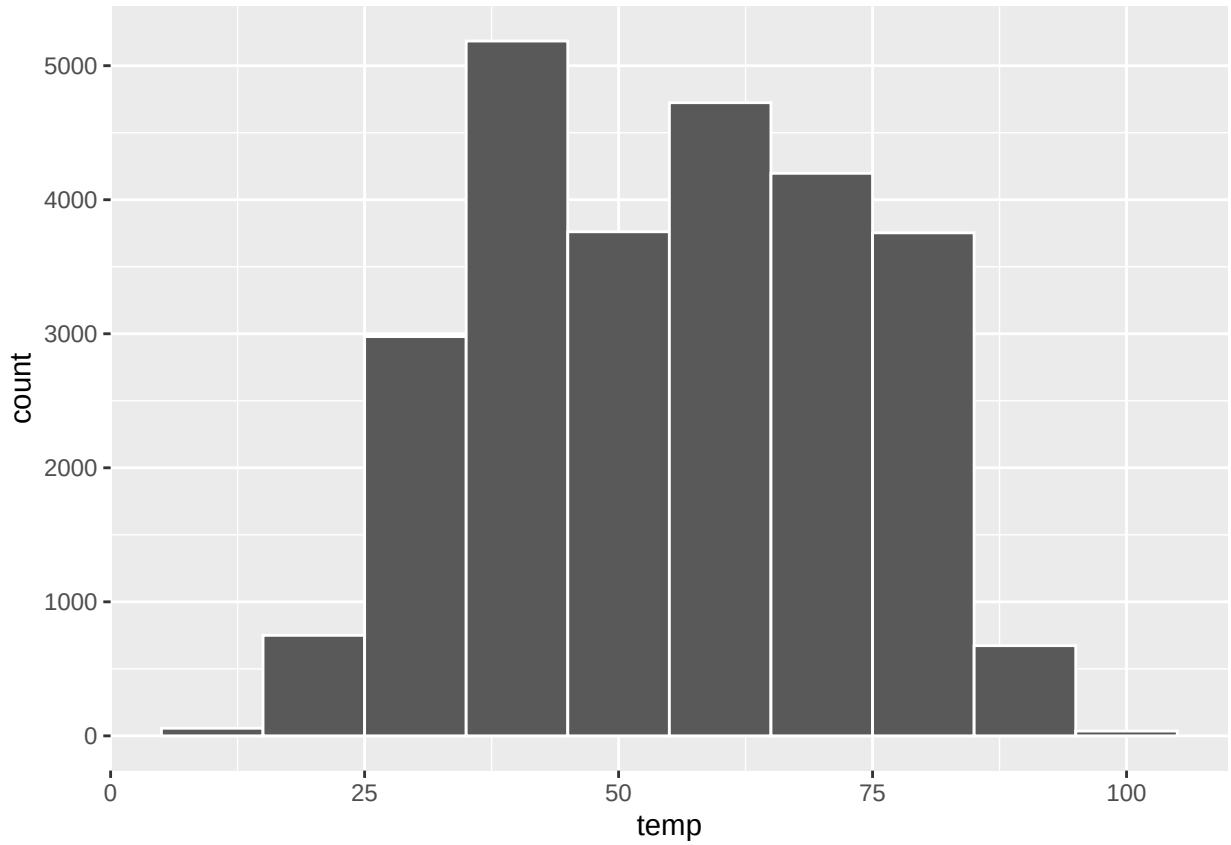


"This histogram shows that the temperature is normally distributed with an average around 60 degrees." "b"

```
ggplot(weather)+
  geom_histogram(mapping = aes(x=temp),color="white", bins=40)
```

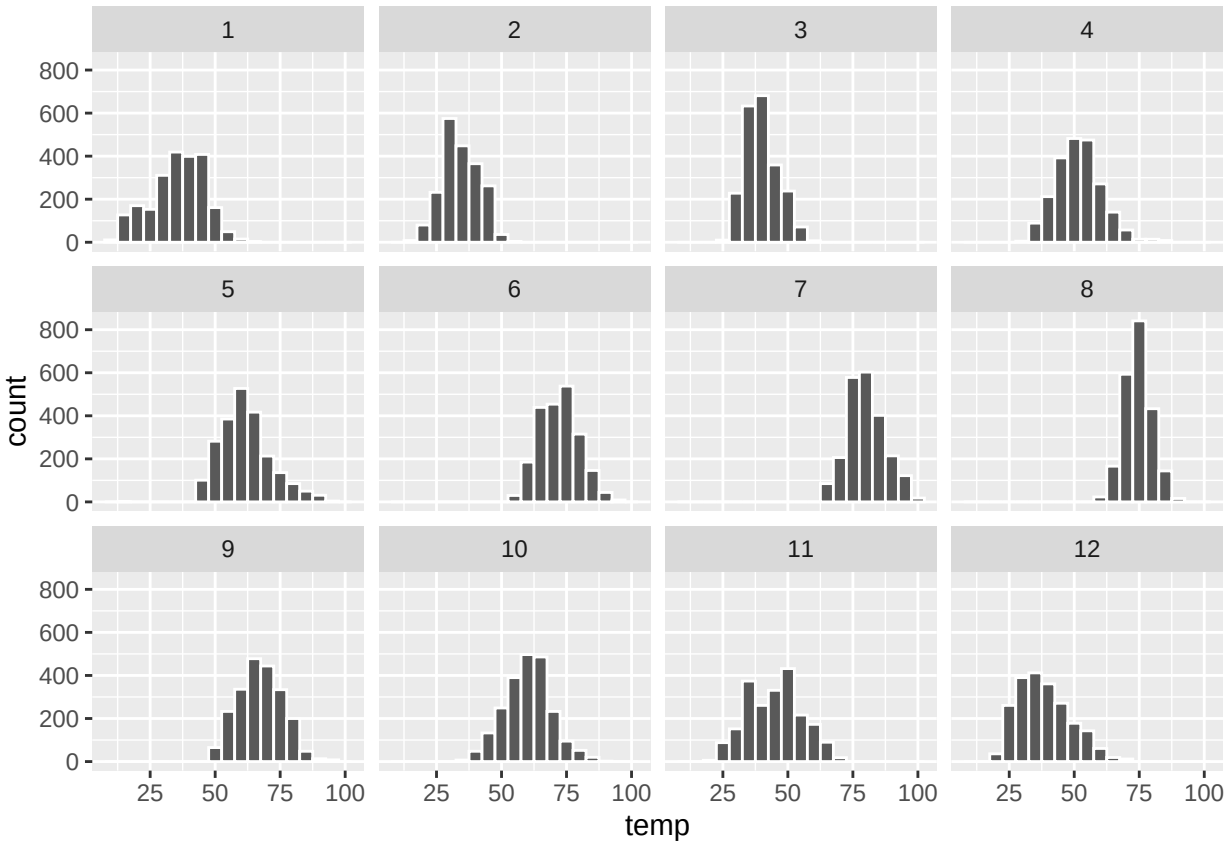



```
ggplot(weather)+
  geom_histogram(mapping = aes(x=temp),color="white", binwidth=10)
```



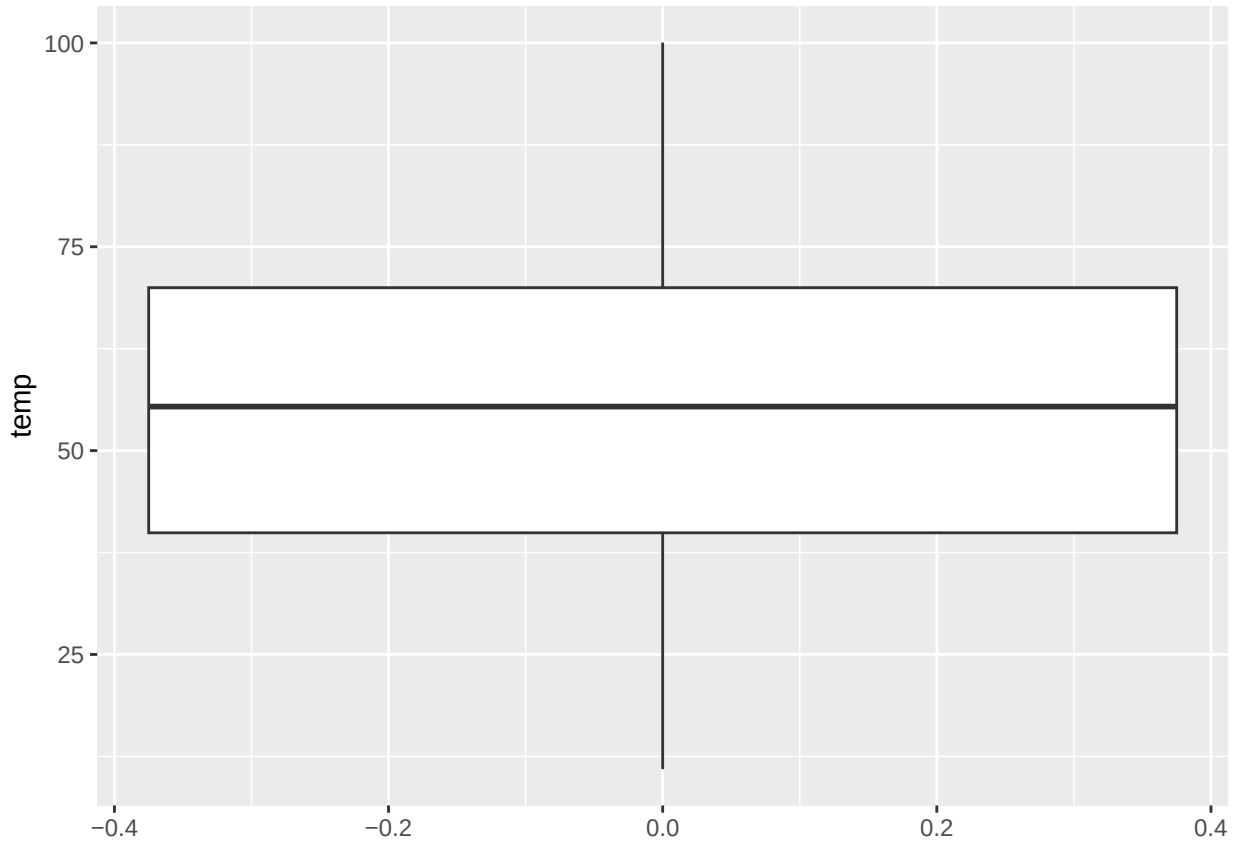
“c”

```
ggplot(weather)+
  geom_histogram(mapping = aes(x=temp),color="white", binwidth=5)+
    facet_wrap(~ month, nrow=3)
```



"Using the histograms for each month, it's easy to compare distributions. For all months, the temperature is normally distributed, the only real difference is the mean. For months 1,2, and 12, the mean was about 30 degrees. For 3,4, and 5 it was closer to 50. For months 6-10, the mean was closer to 70." "Question 5" "A"

```
ggplot(weather)+
  geom_boxplot(mapping = aes(y=temp))
```




```
summary(weather)
```

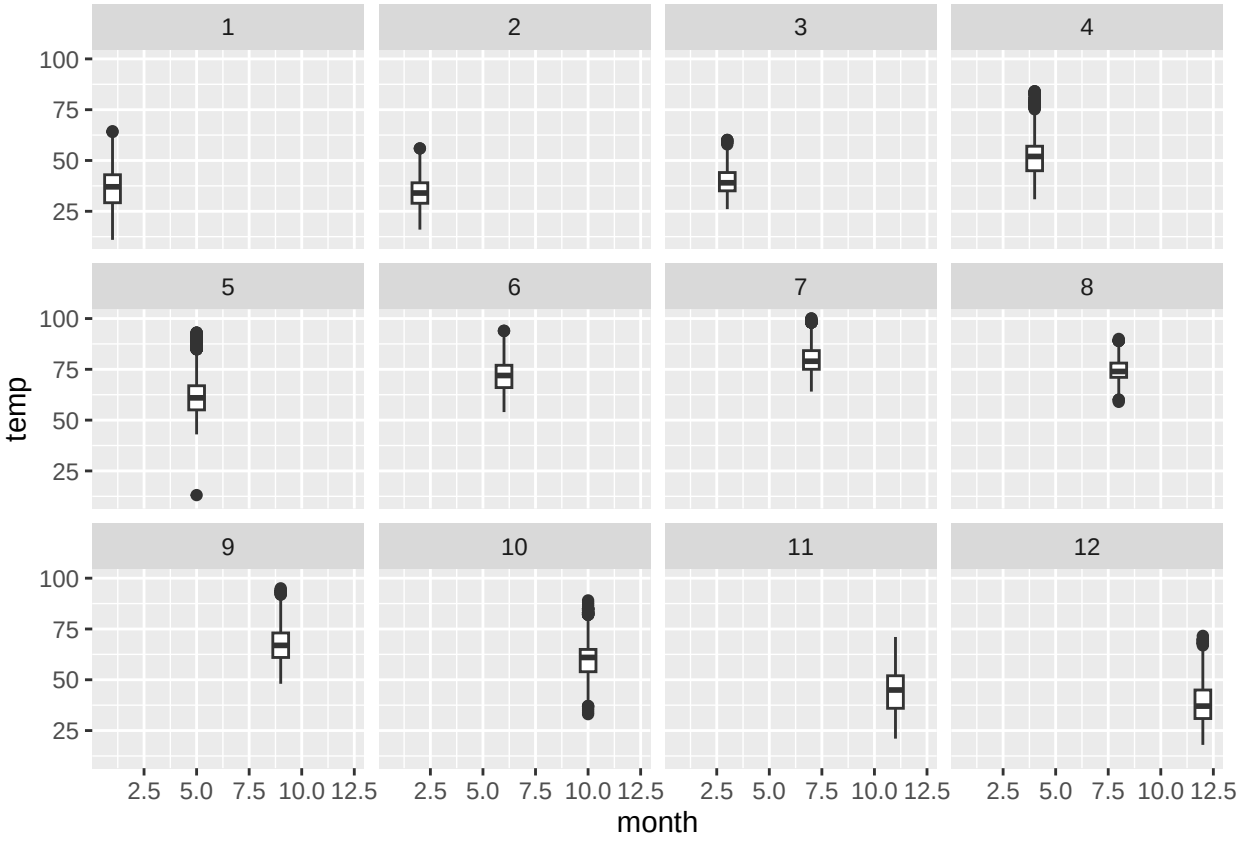
```
##     origin               year          month             day
##  Length:26114       Min.   :2013   Min.   : 1.000   Min.   : 1.00
##  Class :character   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00
##  Mode  :character   Median :2013   Median : 7.000   Median :16.00
##                     Mean   :2013   Mean   : 6.504   Mean   :15.68
##                     3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00
##                     Max.   :2013   Max.   :12.000   Max.   :31.00
##
##       hour            temp            dewp           humid
##  Min.   : 0.00   Min.   : 10.94   Min.   :-9.94   Min.   : 12.74
##  1st Qu.: 6.00   1st Qu.: 39.92   1st Qu.:26.06   1st Qu.: 47.05
##  Median :11.00   Median : 55.40   Median :42.08   Median : 61.79
##  Mean   :11.49   Mean   : 55.26   Mean   :41.44   Mean   : 62.53
##  3rd Qu.:17.00   3rd Qu.: 69.98   3rd Qu.:57.92   3rd Qu.: 78.79
##  Max.   :23.00   Max.   :100.04   Max.   :78.08   Max.   :100.00
##
##     wind_dir       wind_speed        wind_gust         precip
##  Min.   :  0.0   Min.   :   0.000   Min.   :16.11   Min.   :0.000000
##  1st Qu.:120.0   1st Qu.:   6.905   1st Qu.:20.71   1st Qu.:0.000000
##  Median :220.0   Median :  10.357   Median :24.17   Median :0.000000
##  Mean   :199.8   Mean   :  10.517   Mean   :25.49   Mean   :0.004464
##  3rd Qu.:290.0   3rd Qu.:  13.809   3rd Qu.:28.77   3rd Qu.:0.000000
##  Max.   :360.0   Max.   :1048.361   Max.   :66.75   Max.   :1.210000
##  NA's   :460     NA's   :4          NA's   :20777
##     pressure          visib          time_hour
```

```
##  Min.   : 983.8   Min.   : 0.000   Min.   :2013-01-01 01:00:00
##  1st Qu.:1012.9   1st Qu.:10.000   1st Qu.:2013-04-01 21:15:00
##  Median :1017.6   Median :10.000   Median :2013-07-01 14:00:00
##  Mean   :1017.9   Mean   : 9.255   Mean   :2013-07-01 18:23:46
##  3rd Qu.:1023.0   3rd Qu.:10.000   3rd Qu.:2013-09-30 13:00:00
##  Max.   :1042.1   Max.   :10.000   Max.   :2013-12-30 18:00:00
##  NA's   :2728
```

"Based on this boxplot, the mean temperature is about 55 degrees with a Q1 of 40 and a Q3 of 70. The max is 100 and the min is 10." "b"

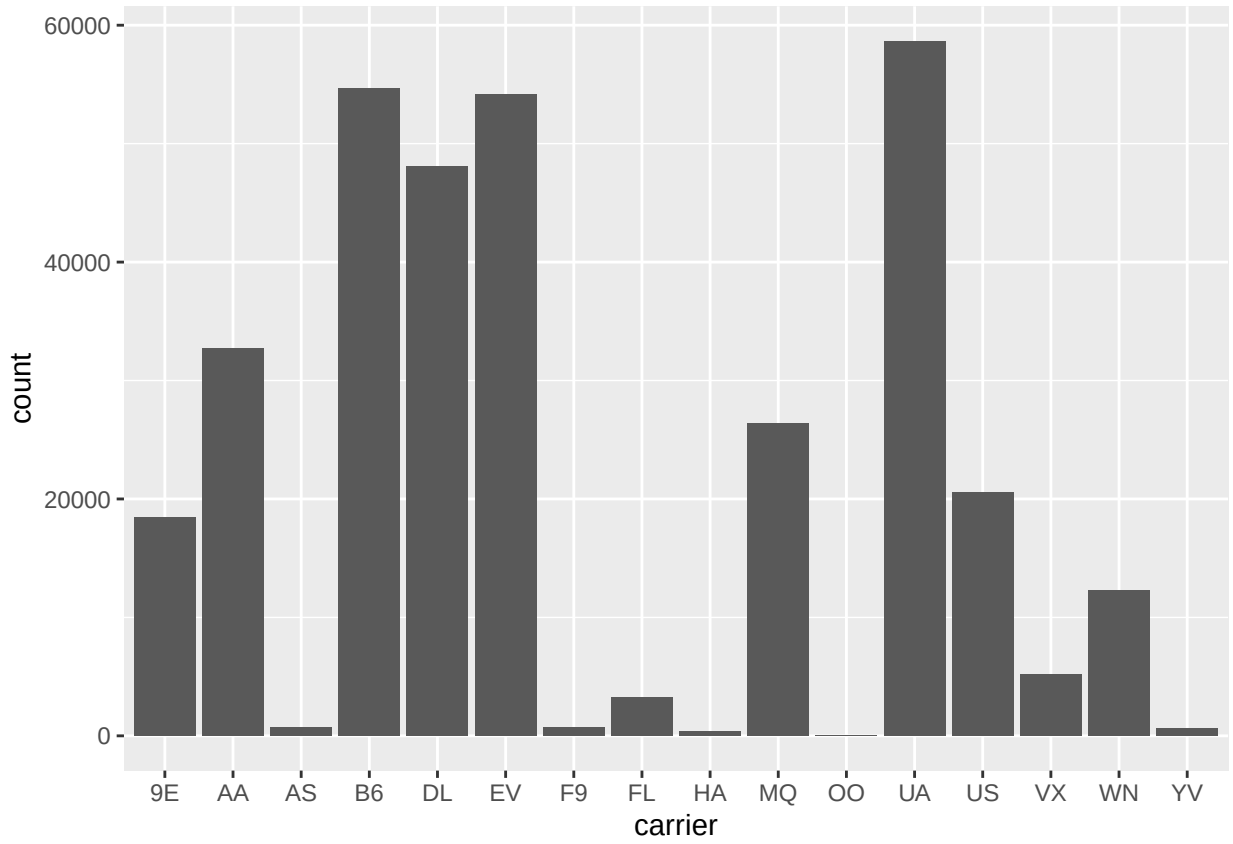
```
ggplot(data = weather, mapping = aes(x= month,y = temp))+
  geom_boxplot()+
  facet_wrap(~ month)
```

```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```



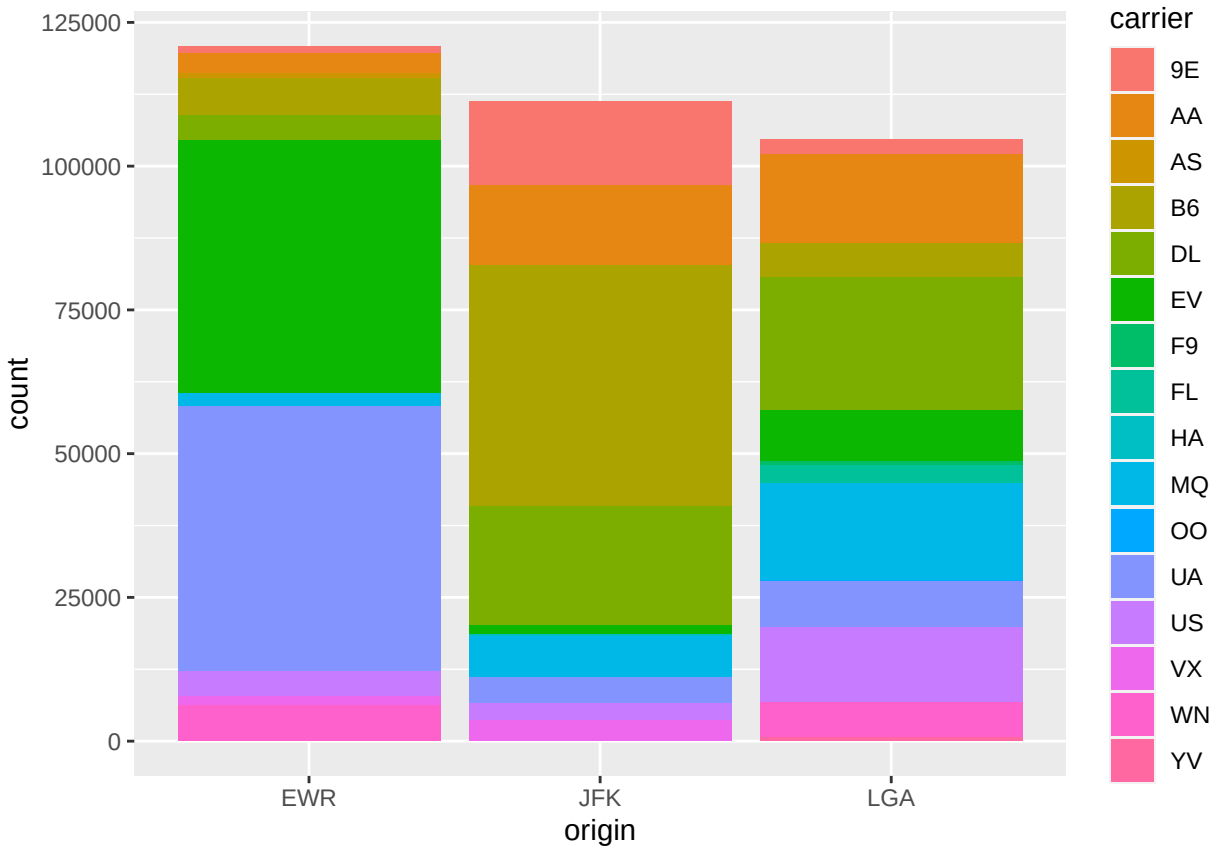
"With the initial code I didn't get what I was expecting, but I used facet_wrap to split the boxplots up by month." "Month with higest temperature variability: May" "Month with lowest temperature variability: March" "Question 6" "A"

```
ggplot(flights)+
  geom_bar(mapping = aes(x=carrier))
```



"The 2nd highest airline for departed flights from NYC in 2013: Jetblue"

```
ggplot(flights)+
  geom_bar(mapping = aes(x=origin, fill=carrier))
```



"Highest airlines for departed flights from the three airports: EWR: United Airways, JFK: Jetblue, LGA: Delta"