Data Visualization using R Rstudio

Bernardo Vimpi

9/20/2020

In this project we analyze data using Data Visualization with R programming. This project is from my Analytics class using R. It contains questions and my answers as well.

Needed packages

```r
library(datasets)
library(nycflights13)
library(ggplot2)
library(dplyr, warn.conflicts = FALSE)
library(gapminder)
library(tinytex)
```

Let's filter the data to the segment needed for our analysis. We use the *filter()* function and assign it to a new data frame (alaska_flights).

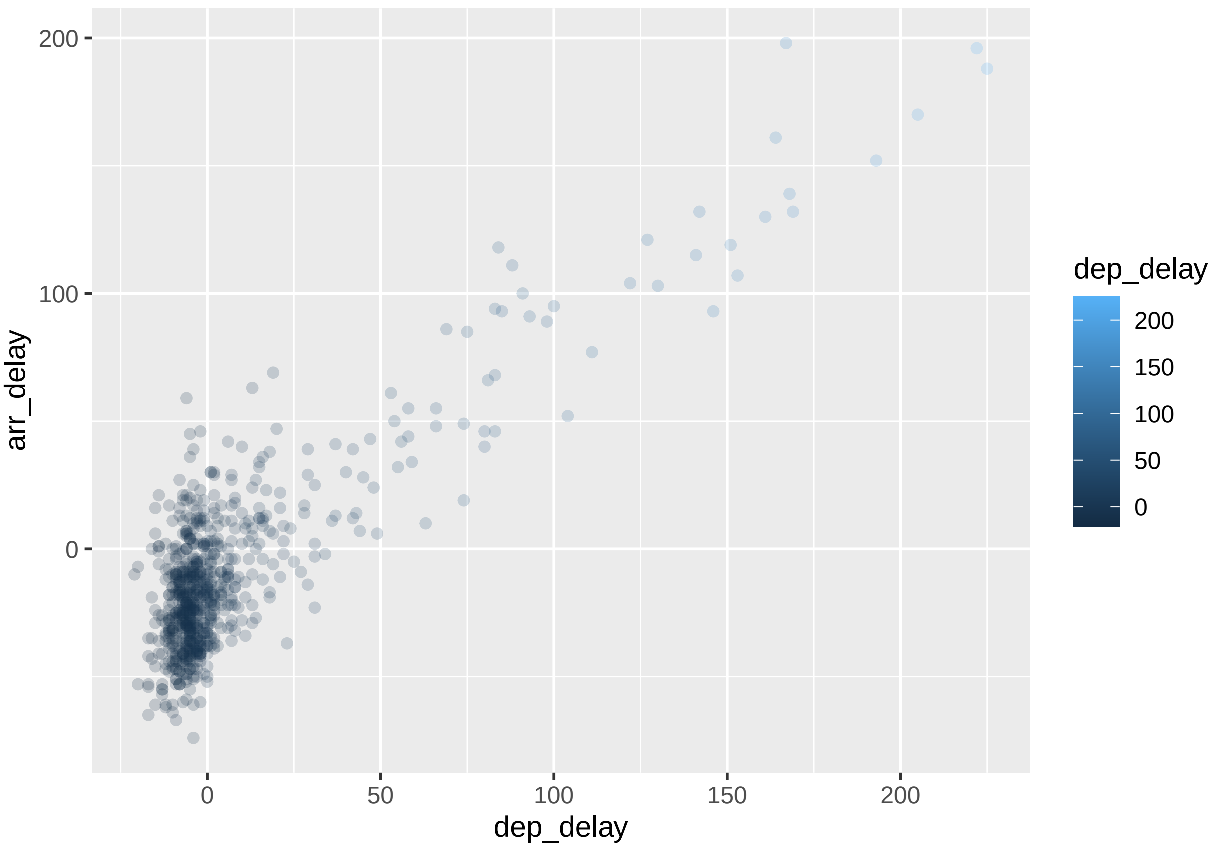
```r
alaska_flights <- flights %>%
  filter(carrier=="AS")
View(alaska_flights)
```

Data Visualization.

Let's now visualize the data in our new data frame using ggplot. I also included the the _geom_point(alpha =0.2) to display any data that has been overplotted. Alternatively, we could also use *geom_jitter()*.

```r
ggplot(alaska_flights, aes(x= dep_delay, y= arr_delay, color= dep_delay))+geom_point(alpha=0.2)
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



(LC2.4) Why do you believe there is a cluster of points near (0,0)? What does(0,0) correspond to in terms of Alaskan flights?

Answer: The cluster of points near (0,0) suggests that there are flights that arrive and depart on time. Those flights are not delayed. Also, looking at the cluster, we notice a heavy concentration of flights within the range of 0 dep_delay and 0 arr_delay. From a performance standpoint, this implies that Alaskan flights are, for the most pat, on time both arrival and departure flights.The (0,0) correspond to no delays in the departure and no delays in the arrival.

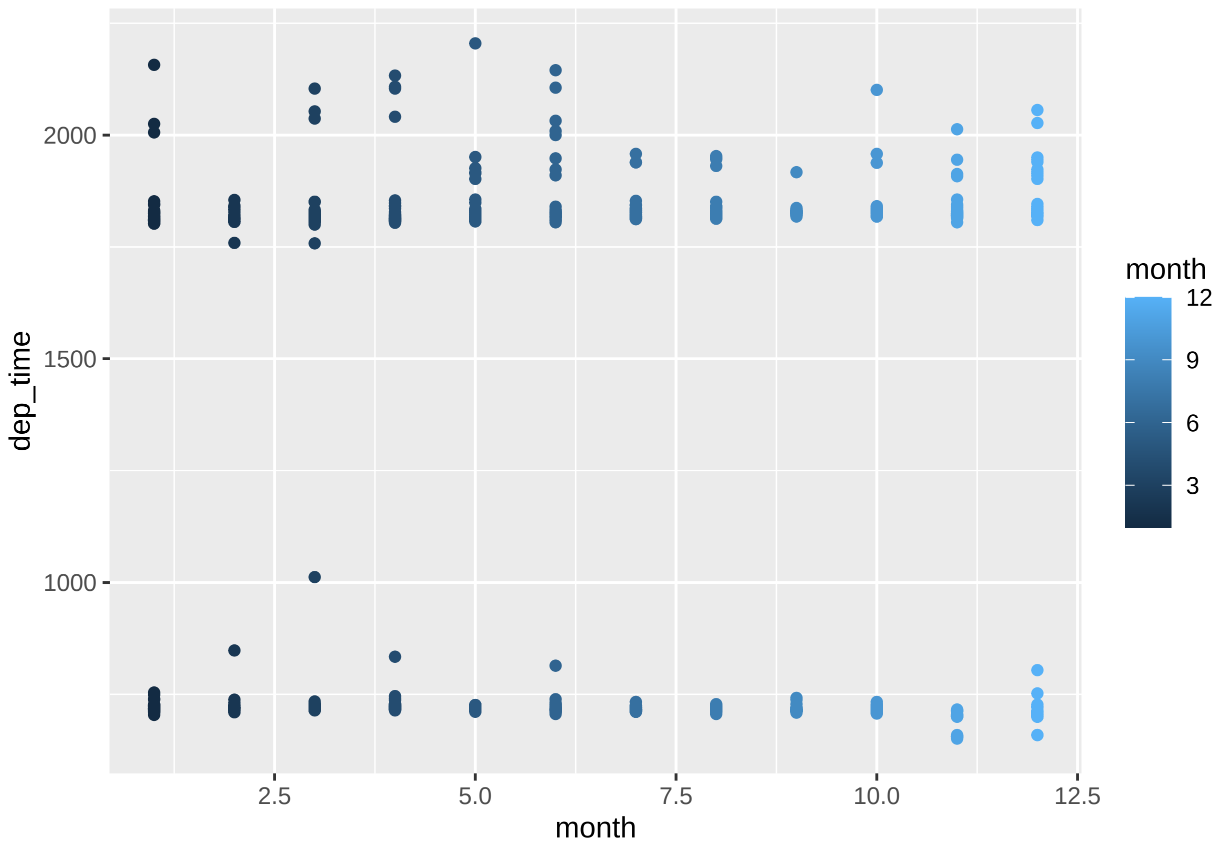
(LC2.5) What are some other features of the plot that stands out to you?

Answer: To me, the plot suggests a linear and positive relationship between departure delay and arrival delay. The majority of Alaskan flights are on time(these are the ones represented in the 0,0 cluster). However, we have also some flights with high departure and arrival delays.As we move up on the dep_delay and arr_delay, we notice that a few flights exhibit high delays since the trend is a positive (direct) and linear relationship.

(LC2.6) Create a new scatterplot using different variables in the Alaska_flights data frame by modifying the example given

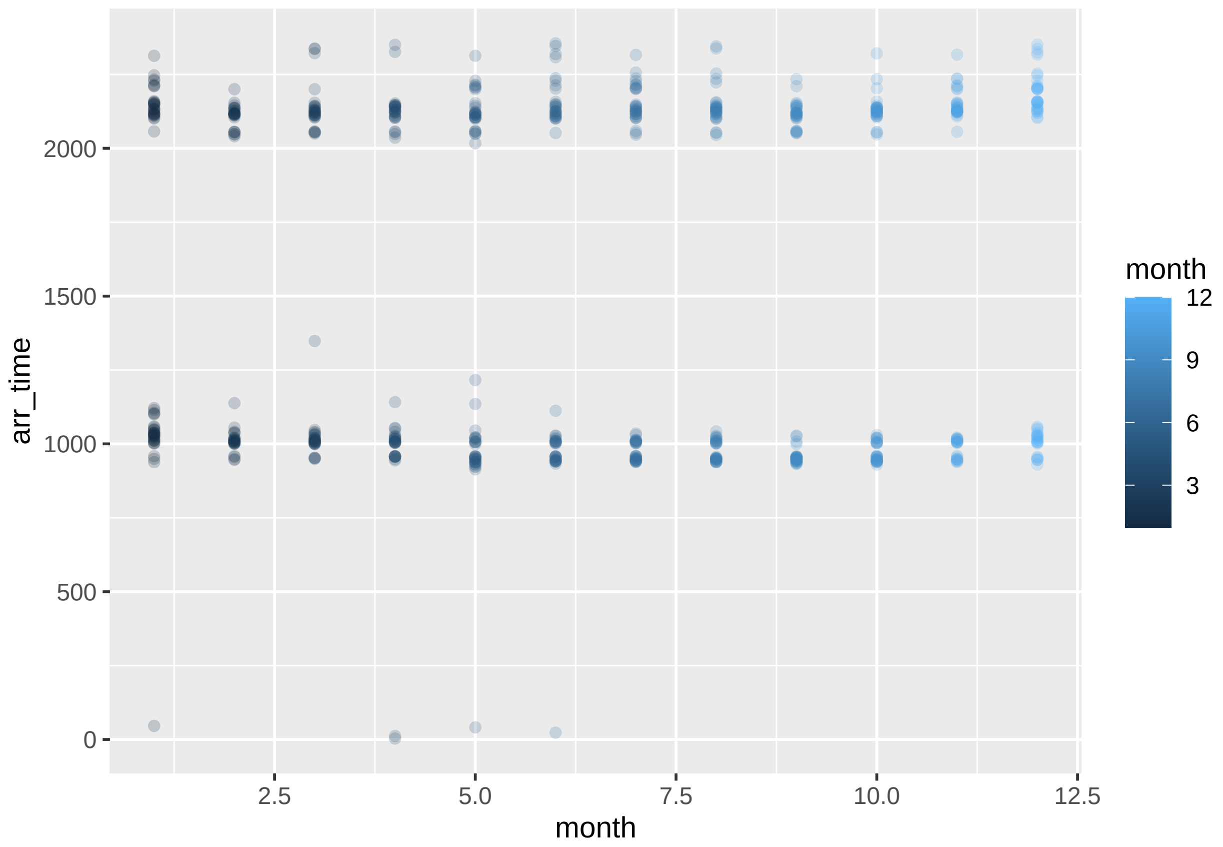
```
ggplot(alaska_flights, aes(x=month, y= dep_time, color= month)) + geom_point()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



```
ggplot(alaska_flights, aes(x=month, y= arr_time, color= month)) + geom_point(alpha=0.2)
```

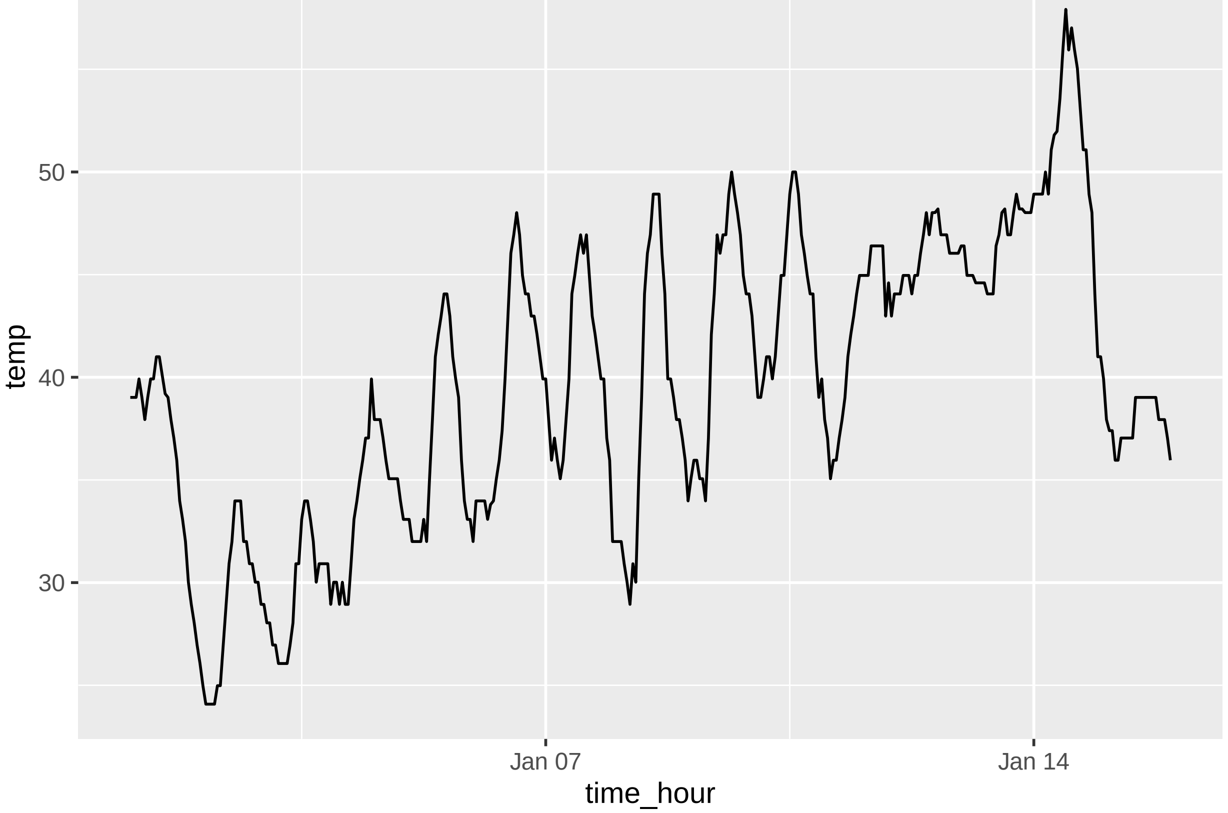
```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Analysis: the two graphs plotted above display the departure and arrival times, respectively, across the month variable.The data is consistent in many ways with the first display (LC2.4). For the most part, Alaskan flights depart and arrive on time. However, we notice some exceptions such as in month 5.0 where we notice changes in the arrival and departure times. As we move from months 9 to 12, most flights seem to depart and arrive on time.

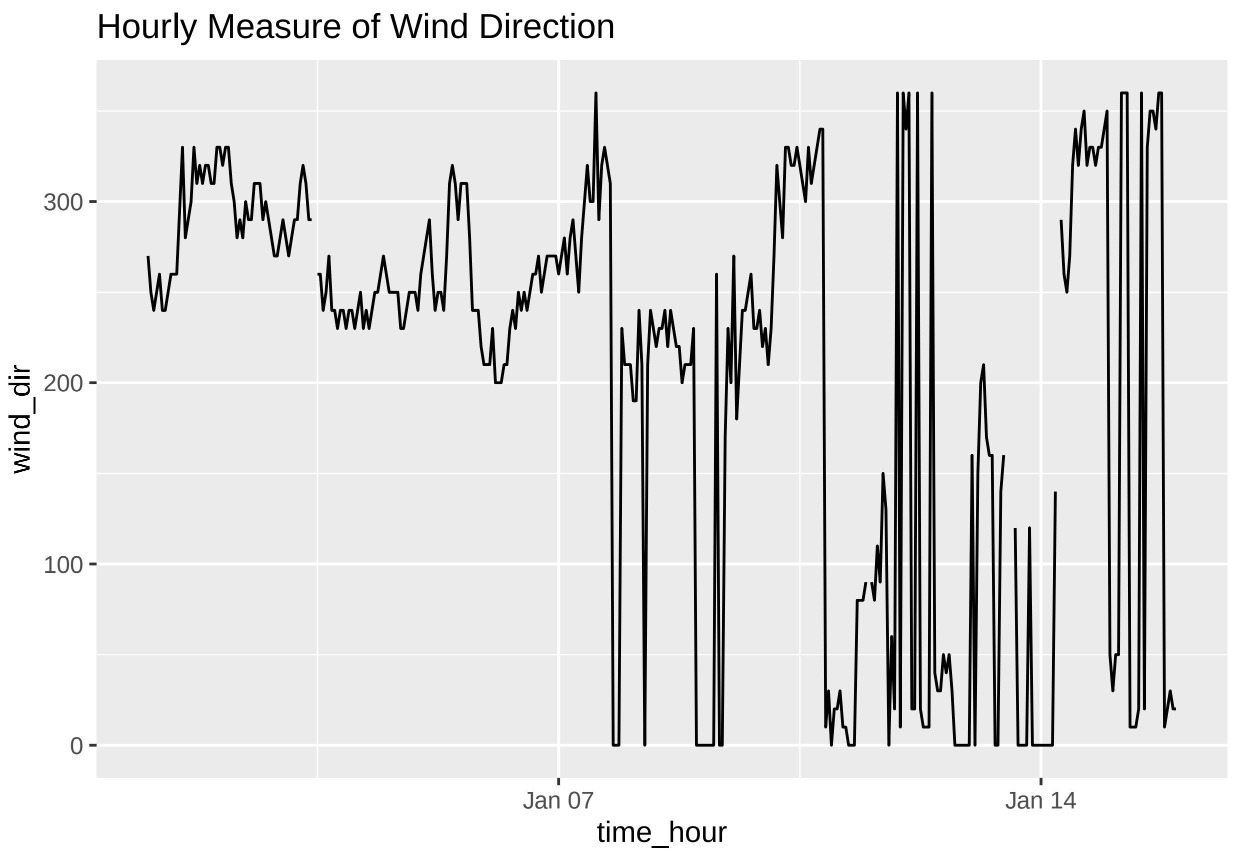
```
View(weather)
early_january_weather <- weather %>%
  filter(origin=="EWR" & month== 1 & day <= 15)

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

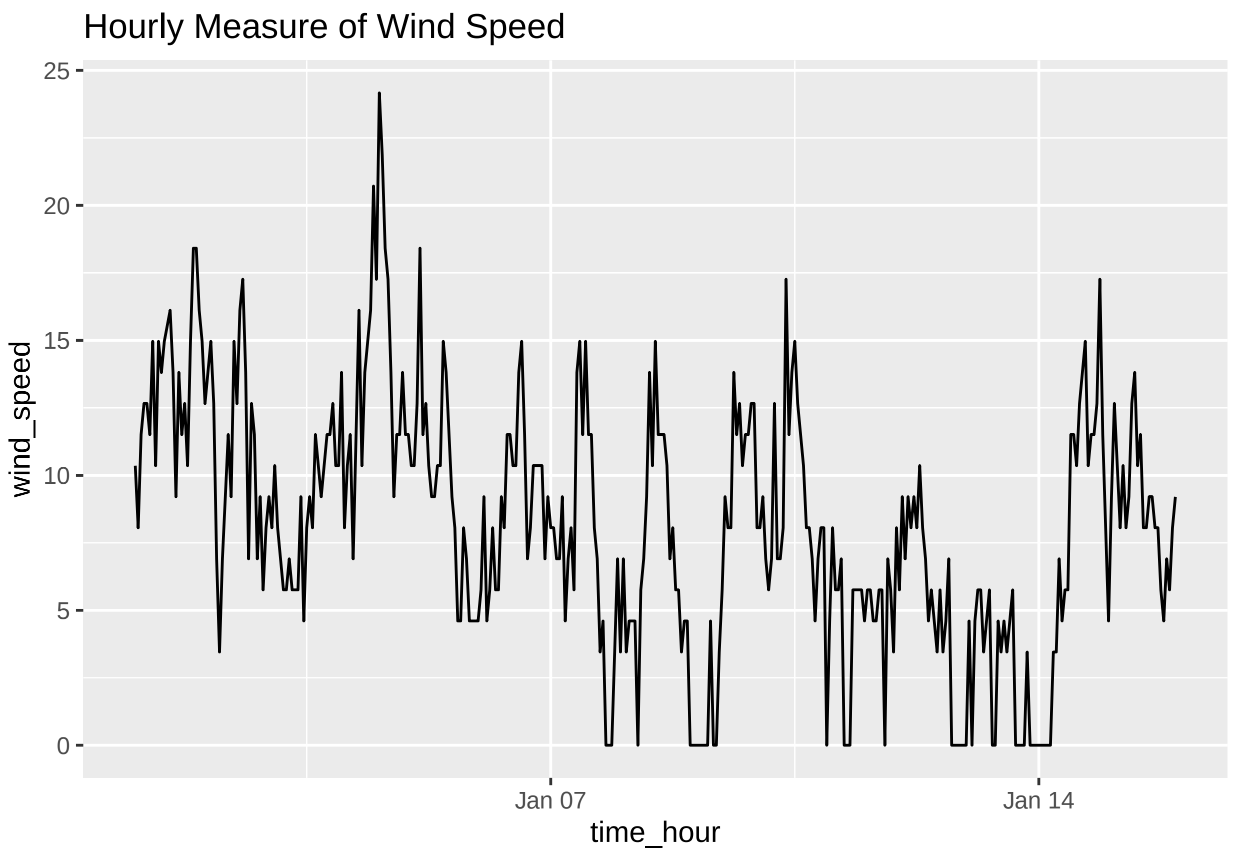


(LC2.13) Plot a time series of a variable other than temp from the Newark Airport in the first 15 days of January 2013

```
ggplot(early_january_weather, aes(x= time_hour, y=wind_dir)) +geom_line() +ggtitle("Hourly Measure of Wi
```



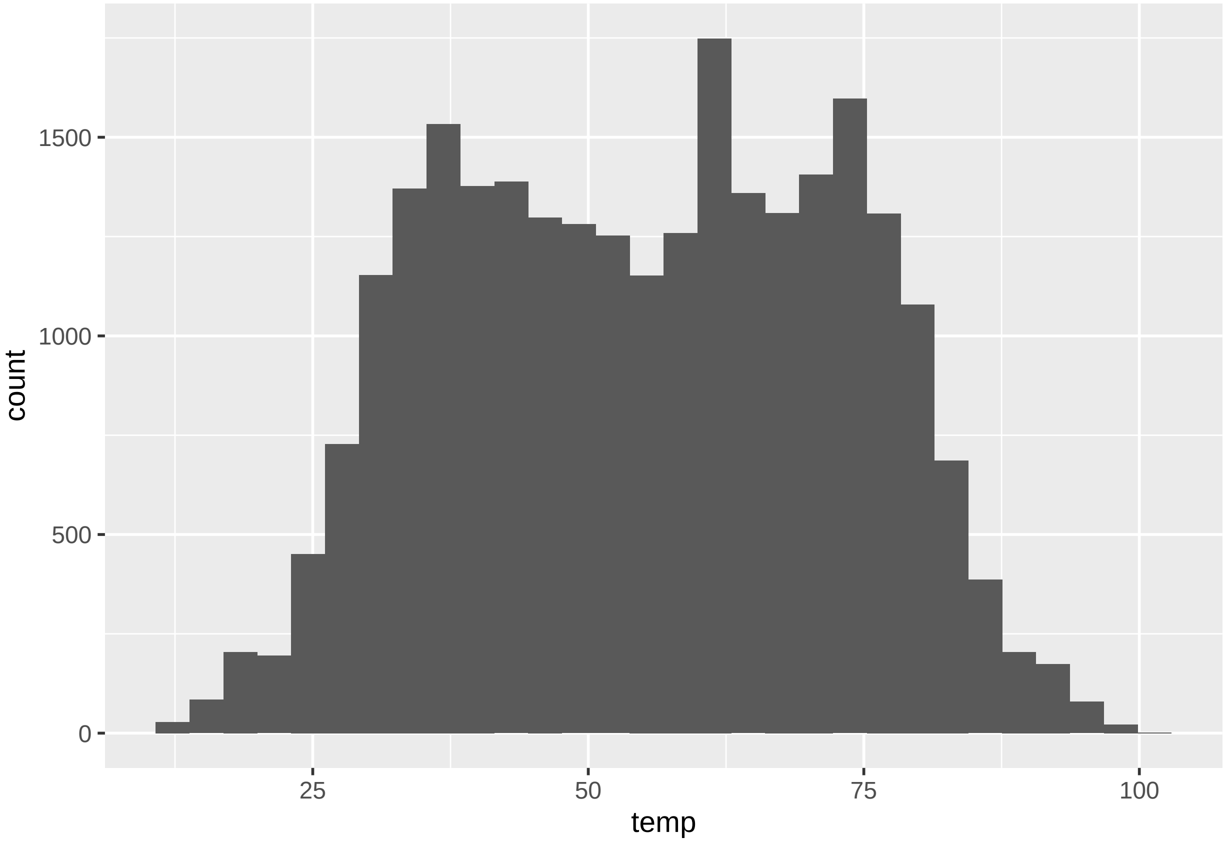
```
ggplot(early_january_weather, aes(x= time_hour, y=wind_speed)) +geom_line() + ggtitle("Hourly Measure o
```



```
ggplot(data=weather, mapping= aes(x=temp)) + geom_histogram()
```

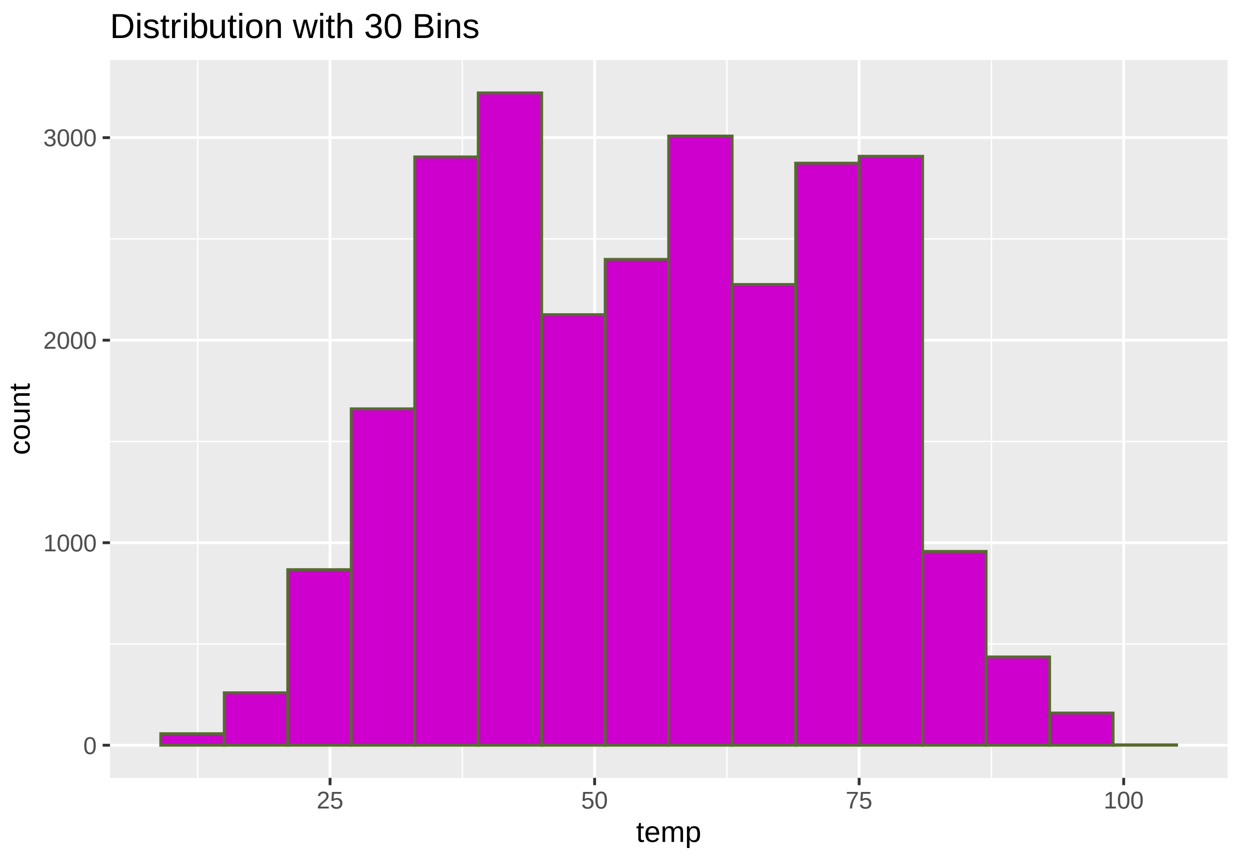
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



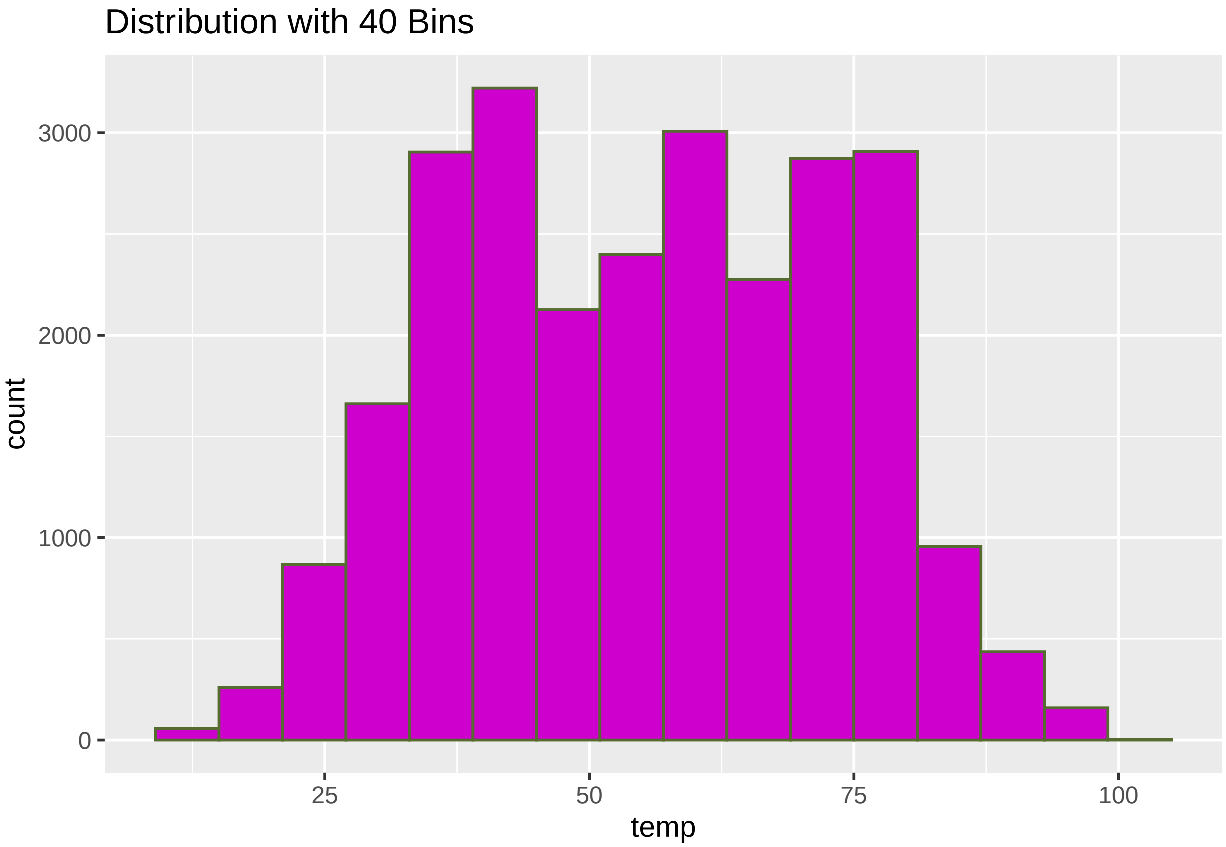
```r
ggplot(weather, aes(x=temp)) + geom_histogram(bins= 30,binwidth = 6, color= "darkolivegreen" , fill= "ma
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



```
ggplot(weather, aes(x=temp)) + geom_histogram(bins= 40,binwidth = 6, color= "darkolivegreen" , fill= "ma
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



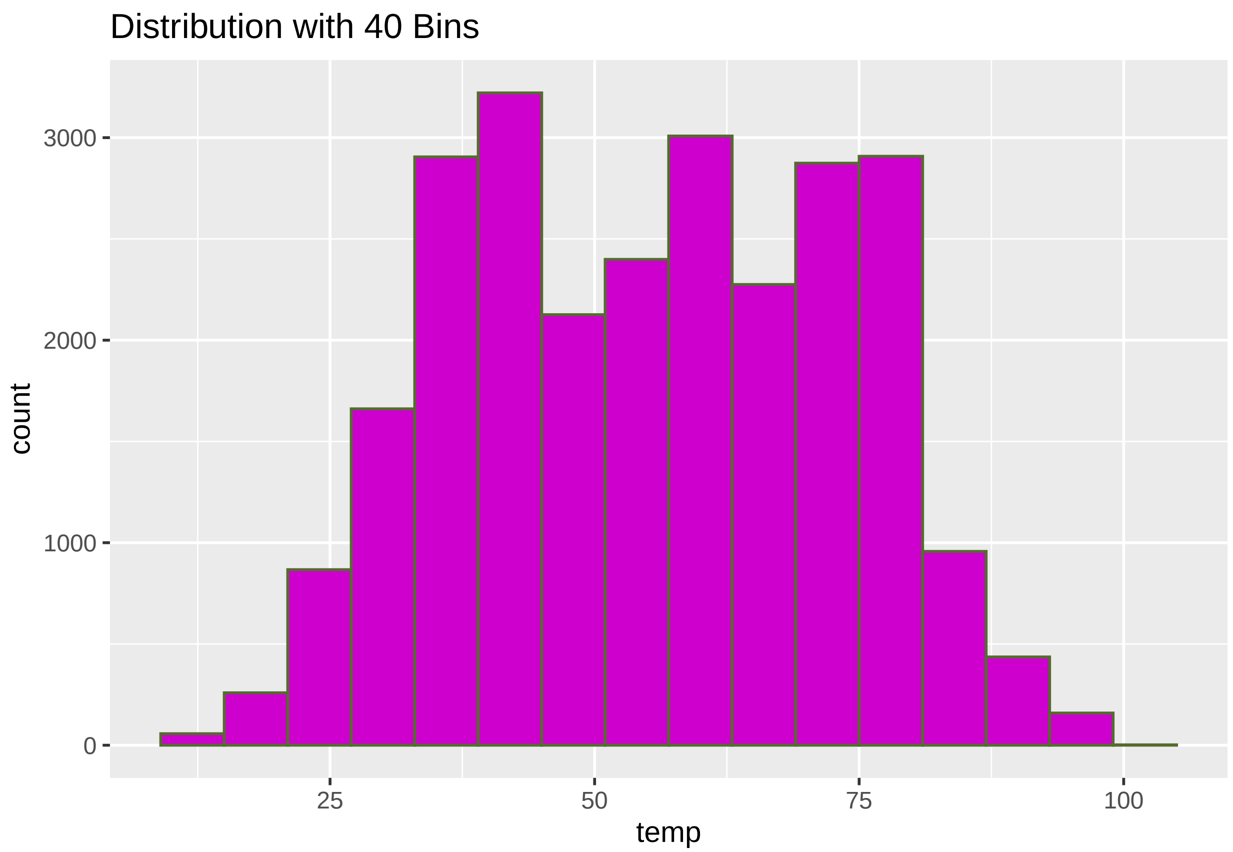
(LC2.14) What does changing the number of bins from 30 to 40 tell us about the distribution of temperatures?

Answer: When I changed the number of bins from 30 to 40 it has an effect on how I visualize the distribution of the data. By increasing the bins from 30 to 40, I am able to visualize the distribution of the temperature with better details.

(LC2.15) Would you classify the distribution of temperature as symmetric or skewed?

```
ggplot(weather, aes(x=temp)) + geom_histogram(bins= 40,binwidth = 6, color= "darkolivegreen" , fill= "ma
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Answer: I would classify the distribution as symmetric because both the right and left side of the distribution resemble each other and the distribution is not skewed. Also, by looking at the histogram we can see the distribution is normal in which the mean, mode and the median all seem to fall at the center of the distribution.

(LC2.16) What would you guess is the "center" value in this distribution? Why did you make that choice?

Answer: The "center" value in this distribution is around temp 55. There are two ways I arrived to this choice: First, I simply looked at the distribution of the data via the histograms I plotted above. Second, another way to determined the center value is via the mean and median of the temp data. To do so, I decided to run the summary() function and looked the the median and mean values for the temp variable which are 55.40 and 55.26, respectively.

```r
summary(weather)
```

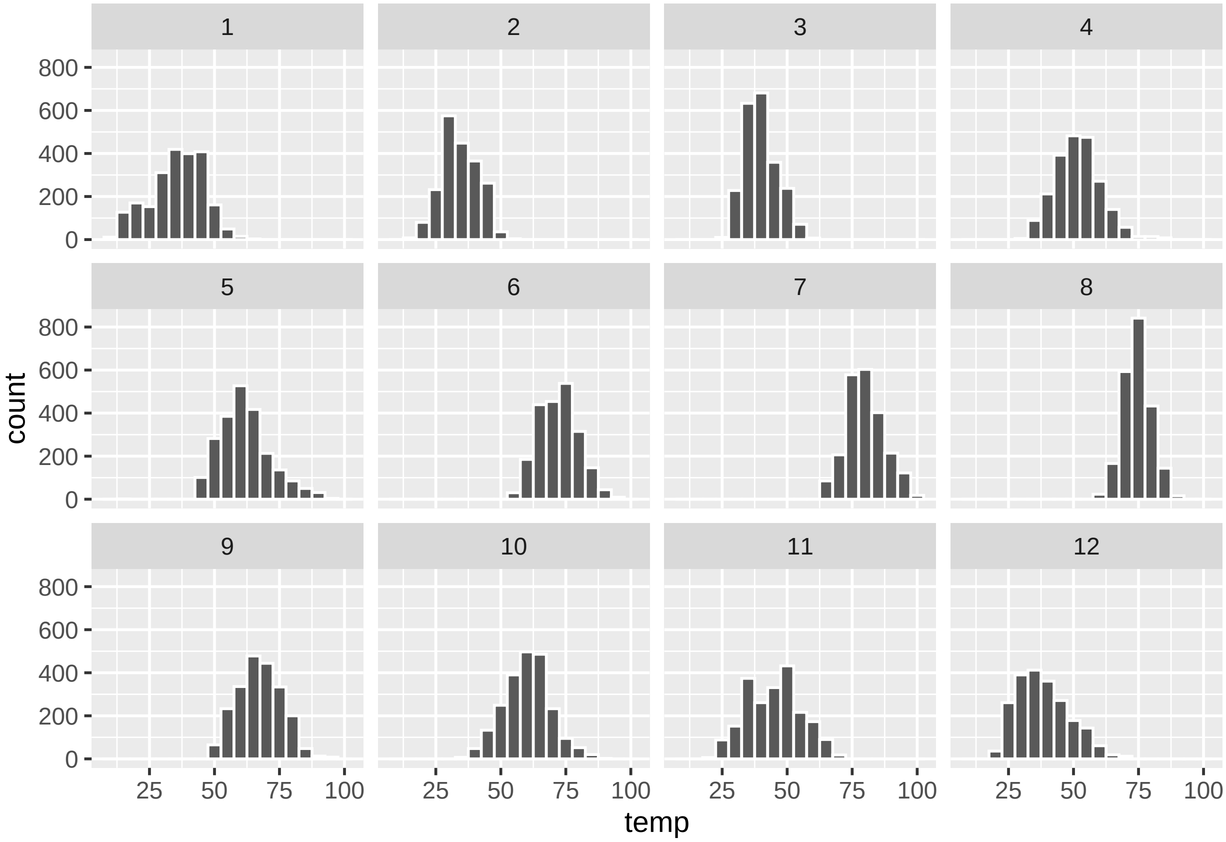
```
##     origin               year          month             day       
##  Length:26115       Min.   :2013   Min.   : 1.000   Min.   : 1.00  
##  Class :character   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00  
##  Mode  :character   Median :2013   Median : 7.000   Median :16.00  
##                     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  
##                  NA's   :1        NA's   :1       NA's   :1       
##     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   :  10.518   Mean   :25.49   Mean   :0.004469  
##  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   :20778                     
##     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 21:30:00  
##  Median :1017.6   Median :10.000   Median :2013-07-01 14:00:00  
##  Mean   :1017.9   Mean   : 9.255   Mean   :2013-07-01 18:26:37  
##  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   :2729
```

(LC2.17) Is this data spread out greatly from the center or is it close? why?

Answer: The data in the histogram is close. Since we have a symmetric distribution we can look at the data an observe that the median and the mean are very close to each other in terms of values 55.40 and 55.26, respectively. Also, the histogram show the distribution of the data in the center have relatively the same height and location on both sides of the center.

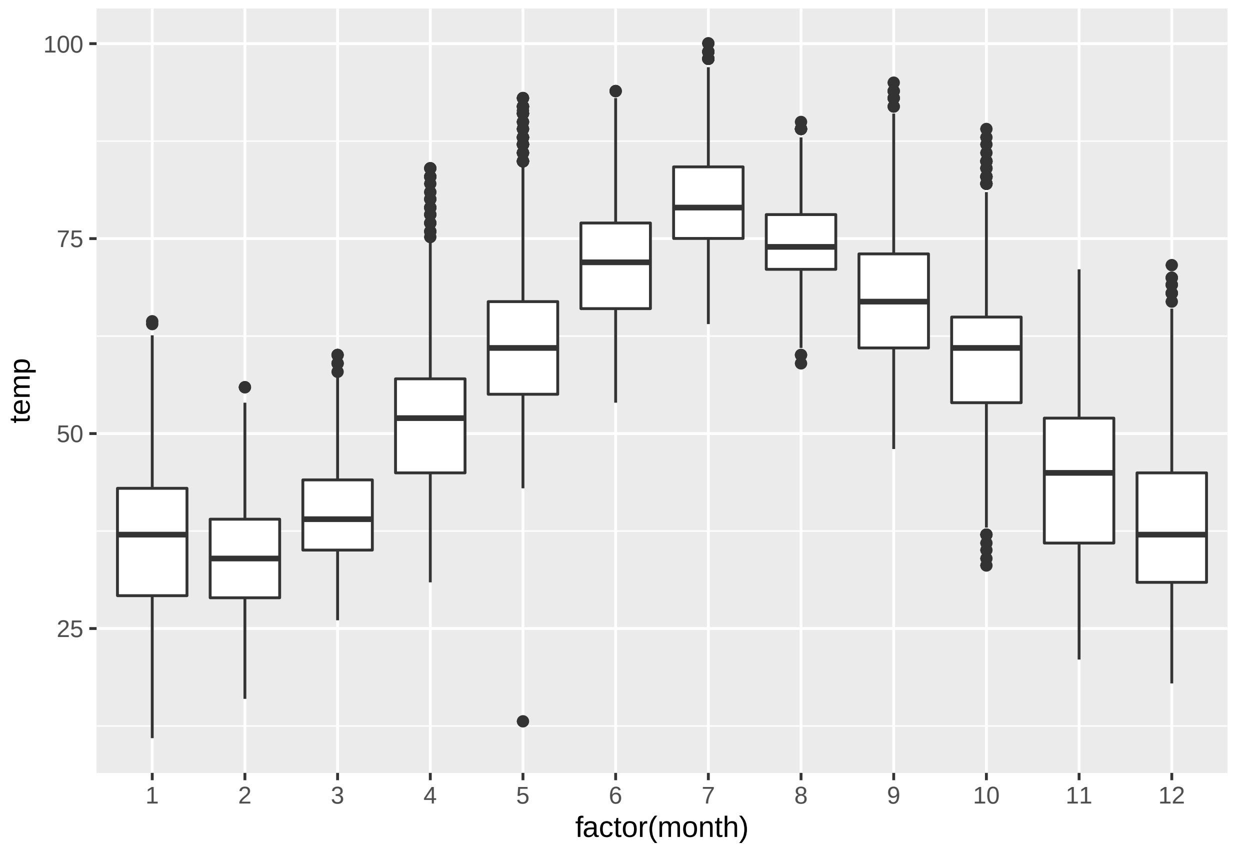
```
ggplot(weather, aes(x=temp)) +geom_histogram(binwidth = 5, color ="white") + facet_wrap(~month, ncol=4,
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



```r
ggplot(weather, aes(x= factor(month), y=temp)) + geom_boxplot()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



(LC2.22) What does the dot at the bottom of the plot for May correspond? Explain what might have occurred in May to produce this point

```
weather %>%
  filter(month==5 & temp <=25)
```

```
## # A tibble: 1 x 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust
##   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>
## 1 JFK     2013     5     8    22  13.1  12.0  95.3       80       8.06        NA
## # ... with 4 more variables: precip <dbl>, pressure <dbl>, visib <dbl>,
## #   time_hour <dttm>
```

Answer: The dot at the bottom of the plot for May is an outlier as the temperature unusually falls far below 25. It is possible to happen due to the transition to summer which starts around June (month 6). However, for the most part, May starts to warm up. I also expanded my analysis by using the filter() which we learned recently to query the data for May and the temp less than or equal to 25. The output shows that on May 8th, 2013, the temp recorded around 10pm (hour 22) JFK was about 13.1. That is unusually low but it is not impossible.Once I observed that, I expanded my analysis by doing a simple research on the weather conditions for JFK for May 8th 2013. The results of my research corroborates with the data in the weather dataset. I uncovered that on May 8, 2013, around 9:50pm, the temperature around JFK suddenly drops from 57F to 13F at 10:50pm. Then, it goes up again to 57 F at around 11:25pm. Therefore the 13 F is an outlier and that is why we have have the dot at the bottom of the plot. Please refer to the link I included on the "Appendix" section of this project for an interactive graph of the past weather data for JFK.

(LC2.23) Which months have the highest variability in temperature? What reasons can you give for this?

```
weather %>%
group_by(month) %>%
summarize(IQR = IQR (temp, na.rm=TRUE)) %>%
  arrange(desc(month))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 12 x 2
##    month   IQR
##    <int> <dbl>
##  1    12 14.0
##  2    11 16.0
##  3    10 11.0
##  4     9 12.1
##  5     8  7.02
##  6     7  9.18
##  7     6 11.0
##  8     5 11.9
##  9     4 12.1
## 10     3  9
## 11     2 10.1
## 12     1 13.8
```

Answer: Looking at the boxplot again, the months of August and November got my attention with regard to the variability in temperature. If we look at the month of August and use the Interquartile Range(IQR) by computing the distance between the 1rs Quartile (25th Percentile) and the 3rd Quartile (75th Percentile) we get the length of the box as well as the spread which is determined by the 2nd Quartile (50th Percentile). Using this technique, we see that the month of August has the smallest IQR (7.02) while the month of November has the biggest IQR (16.02). This measure of spread shows that both August and November have the highest variability in temperature.

(LC2.26) Why are histograms inappropriate for visualizing categorical variables?

Answer: Histograms are used for numerical and not categorical variables. Also, the x- axis values of each bar of the histogram is intended to display an interval set of values. On the other hand, for a categorical variable each individual bar represents a specific level for the variable (categorical). Thus, if we use categorical variables for histograms we can create serious misleading reports/conclusions and business recommendations, if that is the case.

(LC2.27) What is the difference between histograms and barplots?

Answer: One of the major differences between histograms and barplots is the type of variable being used. For example, in an histogram each bar or column represents a specific group of a continuous variable. The barplot, on the other hand, the column is represented by a group of a categorical variable. With regards to outliers, in an histogram, each bin corresponds to where the outlier is, which sometimes may not be clearly identified. In the boxplot, however, the outliers are generally easier to trace and be identified due to labeling. As stated earlier, histograms are for numerical variables while barplots are suitable for categorical variables. We regard to the geometric object, we use geom_histogram() to plot an histogram, while we use geom_bar() for barplots that have counts that are not pre-counted, and we use geom_col() for barplots that contains counts that are pre-counted. Also, barplots are stacked side-by-side with faceted barplots illustrating the joint distribution of two categorical variables. Histograms, on the contrary, illustrate the distribution of one numerical variable that is split by the values of another variable.

(LC2.28) How many Envoy Air flights departed NYC in 2013?

```
flight_depart <-flights %>%
  group_by(carrier=="MQ")%>%
  summarize(total_flights= n()) %>%
  arrange(desc(total_flights))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
  flight_depart
```

```
## # A tibble: 2 x 2
##   ‘carrier == "MQ"‘ total_flights
##   <lgl>                     <int>
## 1 FALSE                    310379
## 2 TRUE                      26397
```

Answer: Based on the result from my above query, there were a total of 26397 Envoy Air flights that departed NYC in 2013.

Note: It took me almost an hour to figure out that Envoy Air flights is coded as MQ. I was looking for it in carrier variable while it stored in the airlines variable. Also, could you provide me with a feedback on how I can improve the query for the last question such that it doesn't return the "FALSE" and "310379" values? I tried really hard and couldn't figure it out.

Appendix:

Site for historical weather data in New York to supplement LC2.22

https://www.wunderground.com/history/daily/us/ny/new-york-city/KJFK/date/2013-5-8

Thank You

Bernardo Vimpi