390R Date and Time notes
10/22/2020

How to create a vector based on a continuous variable.
a<-rnorm(100,0,1)

Cut functions
cut_width(vector, width = width of each cut)
So if you set width to 1, each cut/level would have length of 1
cut_interval(vector,n=number of levels you want)
Create given number of levels
cut_number(vector,n= number of levels)
Divides all values into 5 different categories, makes the number of observations in each the same

Date and Time
With flight data set
What is the best time of day to fly?

```
flights %>%
  group_by(sched_dep_time) %>%
  summarize(arr_delay_avg = mean(arr_delay,na.rm=TRUE)) %>%
  ggplot(mapping = aes(x = sched_dep_time, y = arr_delay_avg)) +
    geom_point(alpha = 0.2) + geom_smooth()
```

For above, x-axis is displayed without the time, we need to use a specific format
How can we create a date/time variable?
Hms package for creating a time, lubridate package for dates
hms::hms(seconds=50,min=34,hour=12) creates 12:34:50
If you don't specify hours or min or sec it will default to 00

In flights data, hour and minutes are given. You can use those to create times.

Question: Use the hour and minute variables in flights to make a new variable that shows the time of each flight as an hms. Then use a smooth line to plot the relationship between time of day and arr_delay.

- Modify your code to make the graph below. You will need to:

1. Create the time variable using hms
2. Calculate average departure and arrival delays for each value of time
3. Gather the average arr_delay and dep_delay columns into a key:value column pair names type and delay
4. Recode the delay column to contain "Departure" and "Arrival" values
5. Create the smooth line

Use the mutate function to create a time hms variable

```
flights %>%
mutate(time=hms::hms(hour = hour, min = minute))%>%
  group_by(time) %>%
  summarize(arr_delay_avg = mean(arr_delay,na.rm=TRUE), dep_delay_avg = mean(dep_delay,
na.rm=TRUE))
```

But to use ggplot with the data we need to transform the data into 2 columns by using the gather function gather(key="type", value = "delay", 2:3) this combines dep_delay and arr_delay into one delay column.

Now we want to change the dates to find out the days of the week which have the highest delays
Load using this code: library(lubridate)
ymd("2012 10 21") will produce a data variable "2012-10-21"
You can reorder ymd to ydm but it will create a variable in year month day order.

Date variable a<-ymd("2012 10 21")
Find year with year(a) = 2012
Find month with month(a) = 10
Find day of week wday(a, label = TRUE, abbr=FALSE)
All functions you can use

function	extracts	extra arguments
year()	year	
month()	month	label = FALSE, abbr = TRUE
week()	week	
day()	day of month	
wday()	day of week	label = FALSE, abbr = TRUE
qday()	day of quarter	
yday()	day of year	
hour()	hour	
minute()	minute	
second()	second	

Question: Extract the day of the week of each flight (as a full name) from time_hour. Plot the average arrival delay by day as a column chart (bar chart).

We need to mutate a new variable for day of week
mutate(weekday=wday(time_hour,label=TRUE,abbr=FALSE))

Then group by the weekday and summarize the delay avg
group_by(weekday)%>%
summarize(delay_avg=mean(arr_delay,na.rm=TRUE))

flights%>% mutate(weekday=wday(time_hour,label=TRUE,abbr=FALSE)) %>%
group_by(weekday)%>% summarize(delay_avg=mean(arr_delay,na.rm=TRUE)) %>% ggplot()+
geom_col(aes(x=weekday,y=delay_avg))