390R EDA: real data example 1
11/10/20
Notes

Lets do a real analysis on Covid-19 data
Today I will go through some of the analysis for these 2 datasets

Start with the WHO COVID global data
First, we need to find any missing values in the dataset
is.na(covid) will find if any values are NA and result in FALSE for non-na and TRUE for NA
sum(is.na(covid)) will find the total number of NA values
You can do this for each column
sum(is.na(select(covid, Date_reported))) this might take a long time with many variables
summary (is.na(covid)) will provide a summary of each variable making it easier to find where the missing values are
How should we deal with these NA observations? We filter to find the observations that have NA values and then decide to keep or discard the values in our analysis. Since the only NA values are for country code where the area is disputed, we can keep the observations.
unique(covid$WHO_region) will tell us the categories in WHO_region

I want to know what is the range of the date in this data?
range(covid$Date_reported)
How many countries are included?
unique(covid$Country)
How many countries are in each region?
covid%>%filter(Date_reported = “2020-11-10”)%>%count(WHO_region)

I want to know the cumulative cases in each region
Covid %>% filter (Date_reported == “2020-11-10”) %>% group_by(WHO_region) summarize(region_cum_cases = sum(Cumulatve_cases))

Look at the cumulative death in the past week
covid%>%filter(Date_reported >= “2020-11-04”)%>%group_by(WHO_region)%>%summarize (region_cum_death = sum(New_deaths))%>%ggplot()+geom_line(aes(x=Date_reported, y=cases)+scale_x_date(labels = date_format(“%d-%b”), breaks = “3 weeks”)

The scale function will rescale and relabel the x axis.

How to plot a histogram for the number of cases each week
covid%>%group_by(Date_reported)%>%summarize(cases=sum(New_cases))%>%mutate(week=floor_date (Date_reported, “week”))%>%group_by(week)%>%summarize(weekly_cases = sum(cases))

The floor_date part will find what the closest week for each date is.

If you want to get two plots in 1 figure, you can assign each plot to p1 and p2.
grid.arrange(p1,p2,nrow=1,ncol=2)