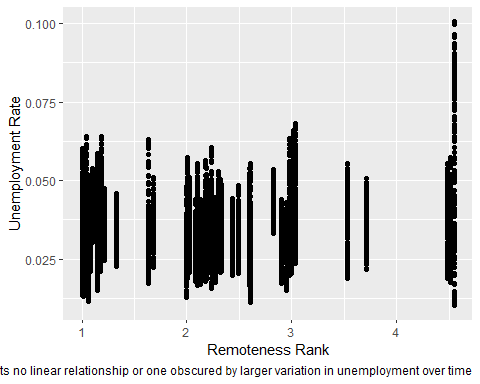
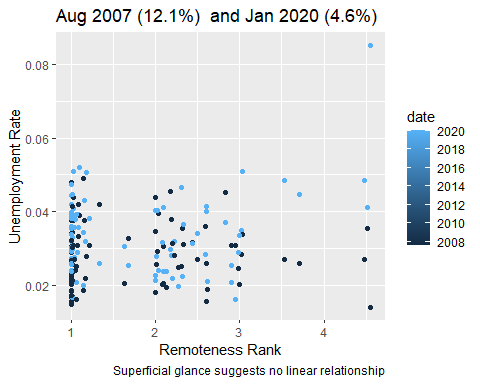
Remoteness-Data-EDA

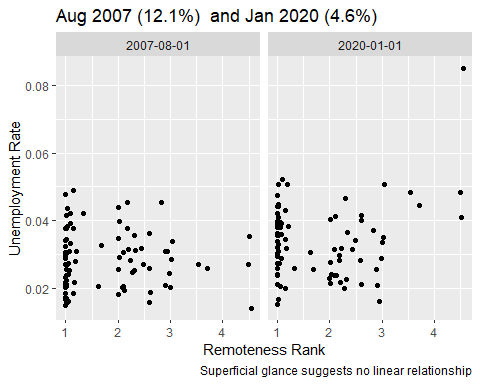
#################################################################################  
# SCRIPT OVERVIEW  
# Exploratory data analysis (incl. data quality review of remoteness data)  
#################################################################################  
  
library(tidyverse)  
library(raustats)  
library(ggplot2)  
#library(hrbrthemes)  
  
load("data/master\_data.RData")  
master <- master\_social\_good  
glimpse(master)  
summary(master)  
  
u\_r <- master %>%  
 select(date, unemployed, population, territory\_sa4, remoteness\_index) %>%  
 mutate(unemploy\_rate = unemployed / population)  
glimpse(u\_r)  
  
summary(u\_r)  
#date unemployed population territory\_sa4 remoteness\_index  
#Min. :2006-01-01 Min. : 570 Min. : 27801 Length:15390 Min. :1.000   
#1st Qu.:2009-07-01 1st Qu.: 3989 1st Qu.:127171 Class :character 1st Qu.:1.008   
#Median :2013-02-01 Median : 6060 Median :182056 Mode :character Median :1.161   
#Mean :2013-01-30 Mean : 7280 Mean :210216 Mean :1.785   
#3rd Qu.:2016-09-01 3rd Qu.: 9319 3rd Qu.:262458 3rd Qu.:2.318   
#Max. :2020-03-01 Max. :37202 Max. :699356 Max. :4.548   
#unemploy\_rate   
#Min. :0.01017   
#1st Qu.:0.02726   
#Median :0.03365   
#Mean :0.03413   
#3rd Qu.:0.04008   
#Max. :0.10060   
  
scatter <- ggplot(data = u\_r) +   
 geom\_point(mapping = aes(x = remoteness\_index, y = unemploy\_rate))  
print(scatter + labs(caption = "Suggests no linear relationship or one obscured by larger variation in unemployment over time", y = "Unemployment Rate", x = "Remoteness Rank"))



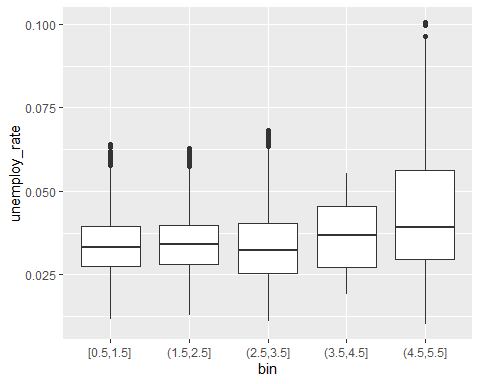
ggsave("RemoteEDA 1 Unemp+Remote Scatter All.pdf")  
  
# Unemployment by remoteness area at peak (Aug 2007 12.7%) and trough (2020 4.6%)  
# of unemployment (by colour)  
u\_r\_200708\_202001 <- u\_r %>%  
 filter(date == "2007-08-01" | date == "2020-01-01")  
glimpse(u\_r\_200708\_202001)  
scatter <- ggplot(data = u\_r\_200708\_202001) +   
 geom\_point(mapping = aes(x = remoteness\_index, y = unemploy\_rate, colour = date))  
print(scatter + ggtitle("Aug 2007 (12.1%) and Jan 2020 (4.6%)") + labs(caption = "Superficial glance suggests no linear relationship", y = "Unemployment Rate", x = "Remoteness Rank"))



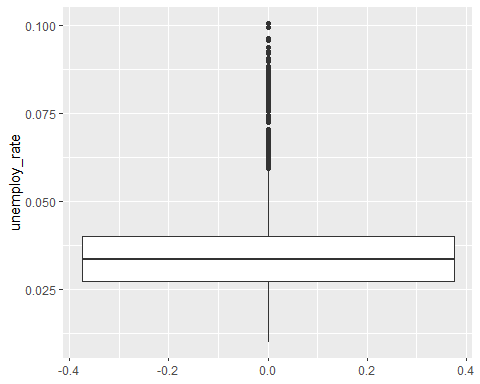
ggsave("RemoteEDA 4 Unemp+Remote Scatter 2007008+20200101 Colour.pdf")  
ggsave("RemoteEDA 4 Unemp+Remote Scatter 2007008+20200101 Colour.png")  
  
# Unemployment by remoteness area at peak (Aug 2007 12.7%) and trough (2020 4.6%)  
# of unemployment (by facet)  
scatter\_peakTrough <- ggplot(data = u\_r\_200708\_202001) +   
 geom\_point(mapping = aes(x = remoteness\_index, y = unemploy\_rate)) +   
 facet\_wrap(~date, nrow = 1)  
print(scatter\_peakTrough + ggtitle("Aug 2007 (12.1%) and Jan 2020 (4.6%)") + labs(caption = "Superficial glance suggests no linear relationship", y = "Unemployment Rate", x = "Remoteness Rank"))



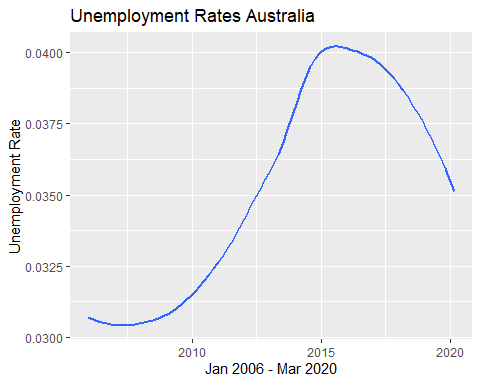
ggsave("RemoteEDA 5 Unemp+Remote Scatter 2007008+20200101 Facet.pdf")  
  
# Group into bins and chart the resulting boxplots  
u\_r\_bins <- u\_r %>%  
 mutate( bin=cut\_width(remoteness\_index, width=1.0, boundary=0.5))  
glimpse(u\_r\_bins)  
  
box\_r\_bins <- ggplot(u\_r\_bins, aes(x=bin, y=unemploy\_rate)) +   
 geom\_boxplot()  
print(box\_r\_bins)



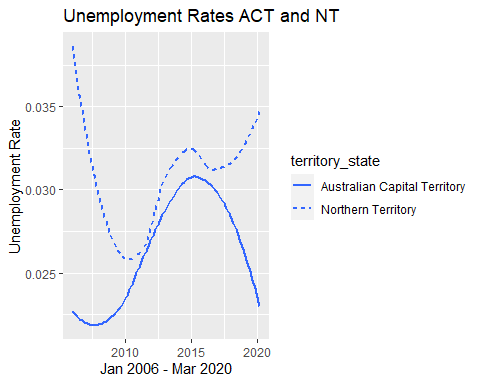
ggsave("RemoteEDA 11 Unemp+Remote Boxplot.pdf")  
ggsave("RemoteEDA 11 Unemp+Remote Boxplot.png")  
  
bin1\_sa4\_count <- u\_r\_bins %>%  
 filter(bin == "[0.5,1.5]") %>%  
 distinct(territory\_sa4)  
print(bin1\_sa4\_count)   
# 49  
  
bin2\_sa4\_count <- u\_r\_bins %>%  
 filter(bin == "(1.5,2.5]") %>%  
 distinct(territory\_sa4)  
print(bin2\_sa4\_count)   
# 22  
  
bin3\_sa4\_count <- u\_r\_bins %>%  
 filter(bin == "(2.5,3.5]") %>%  
 distinct(territory\_sa4)  
print(bin3\_sa4\_count)   
# 12  
  
bin4\_sa4\_count <- u\_r\_bins %>%  
 filter(bin == "(3.5,4.5]") %>%  
 distinct(territory\_sa4)  
print(bin4\_sa4\_count)   
# 2  
  
bin5\_sa4\_count <- u\_r\_bins %>%  
 filter(bin == "(4.5,5.5]") %>%  
 distinct(territory\_sa4)  
print(bin5\_sa4\_count)   
# 2  
  
# simple boxplot for Unemployment on it's own  
ggplot(data = u\_r, mapping = aes(y = unemploy\_rate)) +   
 geom\_boxplot()



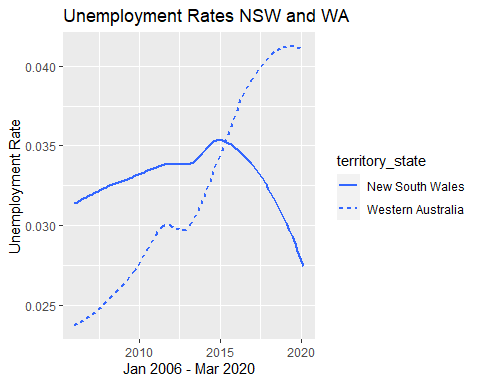
ggsave("RemoteEDA 3 Unemp+Remote Boxplot.pdf")  
  
summary(u\_r$unemploy\_rate)  
#Min. 1st Qu. Median Mean 3rd Qu. Max.   
#0.01017 0.02726 0.03365 0.03413 0.04008 0.10060  
  
u\_r\_outliers <- u\_r %>%  
 filter(unemploy\_rate > 1.5 \* 0.03413)  
glimpse(u\_r\_outliers)  
#Observations: 620  
  
n\_distinct(u\_r\_outliers$territory\_sa4)  
#[1] 28  
u\_r\_outliers\_sa4 <- u\_r\_outliers %>%  
 group\_by(territory\_sa4) %>%  
 summarise(mean(unemploy\_rate), mean(remoteness\_index))  
u\_r\_outliers\_sa4  
print(u\_r\_outliers\_sa4, n = "inf")  
# A tibble: 28 x 3  
#territory\_sa4 `mean(unemploy\_rate)` `mean(remoteness\_index)`  
#<chr> <dbl> <dbl>  
# 1 Adelaide - North 0.0529 1.02  
#2 Ballarat 0.0552 2.00  
#3 Cairns 0.0565 3.02  
#4 Central Queensland 0.0524 2.30  
#5 Coffs Harbour - Grafton 0.0527 2.18  
#6 Hunter Valley exc Newcastle 0.0584 1.63  
#7 Ipswich 0.0553 1.18  
#8 Logan - Beaudesert 0.0524 1.10  
#9 Mackay - Isaac - Whitsunday 0.0540 2.61  
#10 Mandurah 0.0578 1.04  
#11 Melbourne - North West 0.0515 1.08  
#12 Melbourne - West 0.0545 1.01  
#13 Mid North Coast 0.0544 2.18  
#14 Moreton Bay - North 0.0546 1.15  
#15 Murray 0.0564 2.23  
#16 New England and North West 0.0520 2.59  
#17 Northern Territory - Outback 0.0549 4.51  
#18 Perth - North East 0.0516 1.05  
#19 Queensland - Outback 0.0758 4.55  
#20 Southern Highlands and Shoalhaven 0.0523 2.00  
#21 Sydney - Blacktown 0.0550 1   
#22 Sydney - Outer South West 0.0538 1.14  
#23 Sydney - South West 0.0551 1.00  
#24 Tasmania - South East 0.0554 2.98  
#25 Tasmania - West and North West 0.0523 2.82  
#26 Townsville 0.0574 3.04  
#27 Western Australia - Outback (North and South) 0.0531 4.01  
#28 Wide Bay 0.0536 2.10  
  
#########################################################################  
#Problem solving to try to resolve joining problems and getting to territory rollup  
  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
glimpse(sa1\_4) #Observations: 57,523  
glimpse(u\_r) #Observations: 15,390  
  
#This left\_join didn't work - 9 million rows?  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
u\_r\_r <- left\_join(u\_r, sa1\_4, by = c("territory\_sa4" = "SA4\_NAME\_2016"))  
glimpse(u\_r\_r) # Observations: 9,368,406 ###9 million rows?###  
  
#This inner\_join didn't fix it  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
u\_r\_r <- inner\_join(u\_r, sa1\_4, by = c("territory\_sa4" = "SA4\_NAME\_2016"))  
glimpse(u\_r\_r) # Observations: 9,366,867 ###9 million rows?###  
  
#This semi\_join didn't fix it  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
u\_r\_r <- semi\_join(u\_r, sa1\_4, by = c("territory\_sa4" = "SA4\_NAME\_2016"))  
glimpse(u\_r\_r) # Observations: 13,851 ###13,851 rows?###  
  
# Neither of these 3 X attempts to group the orginal data acted to reduce the rows in the original file   
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
sa1\_4\_group <- sa1\_4 %>%  
 select(SA4\_NAME\_2016, GCCSA\_NAME\_2016, STATE\_NAME\_2016)%>%  
 group\_by(SA4\_NAME\_2016)  
glimpse(sa1\_4\_group)  
# Observations: 57,523  
  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
sa1\_4 <- sa1\_4 %>%  
 select(SA4\_NAME\_2016, GCCSA\_NAME\_2016, STATE\_NAME\_2016) %>%  
 group\_by("SA4\_NAME\_2016")  
# Observations: 57,523  
  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
sa1\_4 <- sa1\_4 %>%  
 group\_by(SA4\_NAME\_2016) %>%  
 summarise(territory\_gccsa = mean("GCCSA\_NAME\_2016"), territory\_state = mean("STATE\_NAME\_2016"))  
#warnings()  
  
#These 2 X efforts to group\_by didn't work  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
sa1\_4\_group <- sa1\_4 %>%  
 group\_by(SA4\_NAME\_2016) # same result with "SA4\_NAME\_2016"  
u\_r\_r <- left\_join(u\_r, sa1\_4\_group, by = c("territory\_sa4" = "SA4\_NAME\_2016"))  
glimpse(sa1\_4\_group) # Observations: 57,523  
glimpse(u\_r\_r)# Observations: 9,366,867 ###9 million rows?###  
  
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
sa1\_4\_group <- sa1\_4 %>%  
 select(SA4\_NAME\_2016, GCCSA\_NAME\_2016, STATE\_NAME\_2016) %>%  
 group\_by(SA4\_NAME\_2016)  
glimpse(sa1\_4\_group) # Observations: 57,523  
u\_r\_r <- left\_join(u\_r, sa1\_4\_group, by = c("territory\_sa4" = "SA4\_NAME\_2016"))  
glimpse(u\_r\_r) # Observations: 9,368,406 ###9 million rows?###  
  
#Build on Irfan's solution to create a 1-2-1 like territory rollup table  
sa4\_rollup <- sa1\_4 %>%  
 add\_column(territory\_nation = "Australia") %>%  
 select(territory\_nation, territory\_sa4 = SA4\_NAME\_2016, territory\_gccsa = GCCSA\_NAME\_2016, territory\_state =STATE\_NAME\_2016) %>%  
 distinct(territory\_sa4, territory\_gccsa, territory\_state, territory\_nation)   
glimpse(sa4\_rollup)# Observations: 107 - interim SUCCESS!!!  
count(distinct(sa4\_rollup, territory\_sa4)) #1 107  
print(distinct(sa4\_rollup, territory\_sa4), n = "inf") # A tibble: 107 x 1  
print(distinct(sa4\_rollup, territory\_gccsa), n = "inf") # A tibble: 34 x 1  
print(distinct(sa4\_rollup, territory\_state), n = "inf") # A tibble: 9 x 1  
print(distinct(sa4\_rollup, territory\_nation), n = "inf") # A tibble: 1 x 1  
  
  
# Extend Irfan's solution'  
u\_r\_rollup <- left\_join(u\_r, sa4\_rollup)  
glimpse(u\_r) #Observations: 15,390  
glimpse(u\_r\_rollup) #Observations: 15,390  
select(u\_r\_rollup, territory\_state, territory\_gccsa, territory\_sa4, date, unemployed, population, unemploy\_rate)  
  
#End of Problem solving to try to resolve joining problems and getting to territory rollup  
#Solution below using Irfan's tips  
#########################################################################  
#Resuming remoteness EDA  
  
#Load ABS sa1 -> sa4 -> gccsa -> state rollup hierarchies   
sa1\_4 <- read\_csv("data/SA1\_2016\_AUST.csv")  
glimpse(sa1\_4) #Observations: 57,523  
glimpse(u\_r) #Observations: 15,390  
  
#Create a 1-2-1 like territory rollup table  
sa4\_rollup <- sa1\_4 %>%  
 add\_column(territory\_nation = "Australia") %>%  
 select(territory\_nation, territory\_sa4 = SA4\_NAME\_2016, territory\_gccsa = GCCSA\_NAME\_2016, territory\_state =STATE\_NAME\_2016) %>%  
 distinct(territory\_sa4, territory\_gccsa, territory\_state, territory\_nation)   
glimpse(sa4\_rollup)# Observations: 107 - interim SUCCESS!!!  
count(distinct(sa4\_rollup, territory\_sa4)) #1 107  
print(distinct(sa4\_rollup, territory\_sa4), n = "inf") # A tibble: 107 x 1  
print(distinct(sa4\_rollup, territory\_gccsa), n = "inf") # A tibble: 34 x 1  
print(distinct(sa4\_rollup, territory\_state), n = "inf") # A tibble: 9 x 1  
print(distinct(sa4\_rollup, territory\_nation), n = "inf") # A tibble: 1 x 1  
  
# Join unemployment data to territory rollups  
u\_r\_rollup <- left\_join(u\_r, sa4\_rollup) %>%  
 select(territory\_nation, territory\_state, territory\_gccsa, territory\_sa4, remoteness\_index, date, unemployed, population, unemploy\_rate)  
  
glimpse(u\_r) #Observations: 15,390  
glimpse(u\_r\_rollup) #Observations: 15,390  
  
#############################################################################  
# Truncated exploration of unusual territories  
# filter(sa4\_rollup, territory\_state == "Other Territories") # A tibble: 9 x 1  
#############################################################################  
#Resuming remoteness EDA  
  
#Build a table for charting unemployment for the whole of Australia  
u\_aust <- u\_r\_rollup %>%  
 select(territory\_nation, date, unemployed, population) %>%  
 group\_by(territory\_nation, date) %>%  
 summarize(unemployed = sum(unemployed), population = sum(population)) %>%  
 mutate(unemploy\_rate = unemployed / population)  
  
#Build a table for charting unemployment for the whole of Australia  
u\_aust\_tline <- ggplot(data = u\_aust) +   
 geom\_smooth(mapping = aes(x = date, y = unemploy\_rate), se = FALSE)  
print(u\_aust\_tline +   
 ggtitle("Unemployment Rates Australia")  
 + labs(y="Unemployment Rate", x = "Jan 2006 - Mar 2020"))



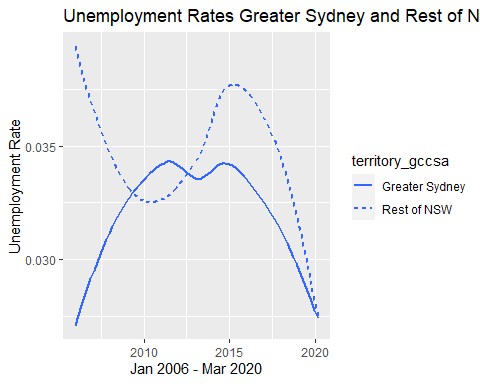
ggsave("RemoteEDA 6 Aust TLine.pdf")  
ggsave("RemoteEDA 6 Aust TLine.png")  
  
#Build a table to show unemployment variation in the ACT and NT  
u\_states <- u\_r\_rollup %>%  
 select(territory\_state, date, unemployed, population) %>%  
 group\_by(territory\_state, date) %>%  
 filter(territory\_state == "Australian Capital Territory" | territory\_state == "Northern Territory") %>%  
 summarize(unemployed = sum(unemployed), population = sum(population)) %>%  
 mutate(unemploy\_rate = unemployed / population)  
glimpse(u\_states)  
  
#Build a table for charting unemployment in the ACT and NT  
u\_states\_tline <- ggplot(data = u\_states) +   
 geom\_smooth(mapping = aes(x = date, y = unemploy\_rate, linetype = territory\_state), se = FALSE)  
print(u\_states\_tline +   
 ggtitle("Unemployment Rates ACT and NT")  
 + labs(y="Unemployment Rate", x = "Jan 2006 - Mar 2020"))



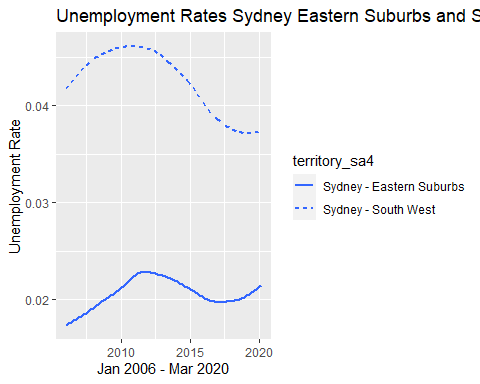
ggsave("RemoteEDA 7 AustStates TLine ACTNT.pdf")  
  
#Build a table to show unemployment variation in NSW and WA  
u\_states <- u\_r\_rollup %>%  
 select(territory\_state, date, unemployed, population) %>%  
 group\_by(territory\_state, date) %>%  
 filter(territory\_state == "Western Australia" | territory\_state == "New South Wales") %>%   
 summarize(unemployed = sum(unemployed), population = sum(population)) %>%  
 mutate(unemploy\_rate = unemployed / population)  
glimpse(u\_states)  
  
#Build a table for charting unemployment in NSW and WA  
u\_states\_tline <- ggplot(data = u\_states) +   
 geom\_smooth(mapping = aes(x = date, y = unemploy\_rate, linetype = territory\_state), se = FALSE)  
print(u\_states\_tline +   
 ggtitle("Unemployment Rates NSW and WA")  
 + labs(y="Unemployment Rate", x = "Jan 2006 - Mar 2020"))



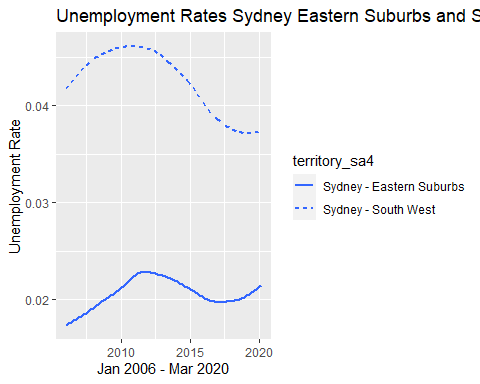
ggsave("RemoteEDA 8 AustStates TLine NSWWA.pdf")  
ggsave("RemoteEDA 8 AustStates TLine NSWWA.png")  
  
#Build a table to show unemployment variation in Greater Sydney and the Rest of NSW  
u\_gccsa <- u\_r\_rollup %>%  
 select(territory\_gccsa, date, unemployed, population) %>%  
 group\_by(territory\_gccsa, date) %>%  
 filter(territory\_gccsa == "Greater Sydney" | territory\_gccsa == "Rest of NSW") %>%   
 summarize(unemployed = sum(unemployed), population = sum(population)) %>%  
 mutate(unemploy\_rate = unemployed / population)  
glimpse(u\_gccsa)  
  
#Build a table for charting unemployment in Greater Sydney and the Rest of NSW  
u\_gccsa\_tline <- ggplot(data = u\_gccsa) +   
 geom\_smooth(mapping = aes(x = date, y = unemploy\_rate, linetype = territory\_gccsa), se = FALSE)  
print(u\_gccsa\_tline +   
 ggtitle("Unemployment Rates Greater Sydney and Rest of NSW")  
 + labs(y="Unemployment Rate", x = "Jan 2006 - Mar 2020"))



ggsave("RemoteEDA 9 AustGCCSA TLine GSydNSW.pdf")  
ggsave("RemoteEDA 9 AustGCCSA TLine GSydNSW.png")  
  
#Build a table to show unemployment variation within Sydney  
u\_sa4 <- u\_r\_rollup %>%  
 select(territory\_sa4, date, unemployed, population) %>%  
 group\_by(territory\_sa4, date) %>%  
 filter(territory\_sa4 == "Sydney - Eastern Suburbs" | territory\_sa4 == "Sydney - South West") %>%   
 summarize(unemployed = sum(unemployed), population = sum(population)) %>%  
 mutate(unemploy\_rate = unemployed / population)  
glimpse(u\_sa4)  
  
#Build a table for charting unemployment in Greater Sydney and the Rest of NSW  
u\_sa4\_tline <- ggplot(data = u\_sa4) +   
 geom\_smooth(mapping = aes(x = date, y = unemploy\_rate, linetype = territory\_sa4), se = FALSE)  
print(u\_sa4\_tline +   
 ggtitle("Unemployment Rates Sydney Eastern Suburbs and South West")  
 + labs(y="Unemployment Rate", x = "Jan 2006 - Mar 2020"))



ggsave("RemoteEDA 10 Austsa4 TLine SydEastAndSouthWest.pdf")  
ggsave("RemoteEDA 10 Austsa4 TLine SydEastAndSouthWest.png")  
  
#Build a table to show unemployment variation within Sydney  
u\_sa4 <- u\_r\_rollup %>%  
 select(territory\_sa4, date, unemployed, population) %>%  
 group\_by(territory\_sa4, date) %>%  
 filter(territory\_sa4 == "Sydney - Eastern Suburbs" | territory\_sa4 == "Sydney - South West") %>%   
 summarize(unemployed = sum(unemployed), population = sum(population)) %>%  
 mutate(unemploy\_rate = unemployed / population)  
glimpse(u\_sa4)  
  
#Build a table for charting unemployment variation within Sydney  
u\_sa4\_tline <- ggplot(data = u\_sa4) +   
 geom\_smooth(mapping = aes(x = date, y = unemploy\_rate, linetype = territory\_sa4), se = FALSE)  
print(u\_sa4\_tline +   
 ggtitle("Unemployment Rates Sydney Eastern Suburbs and South West")  
 + labs(y="Unemployment Rate", x = "Jan 2006 - Mar 2020"))



ggsave("RemoteEDA 10 Austsa4 TLine SydEastAndSouthWest.pdf")