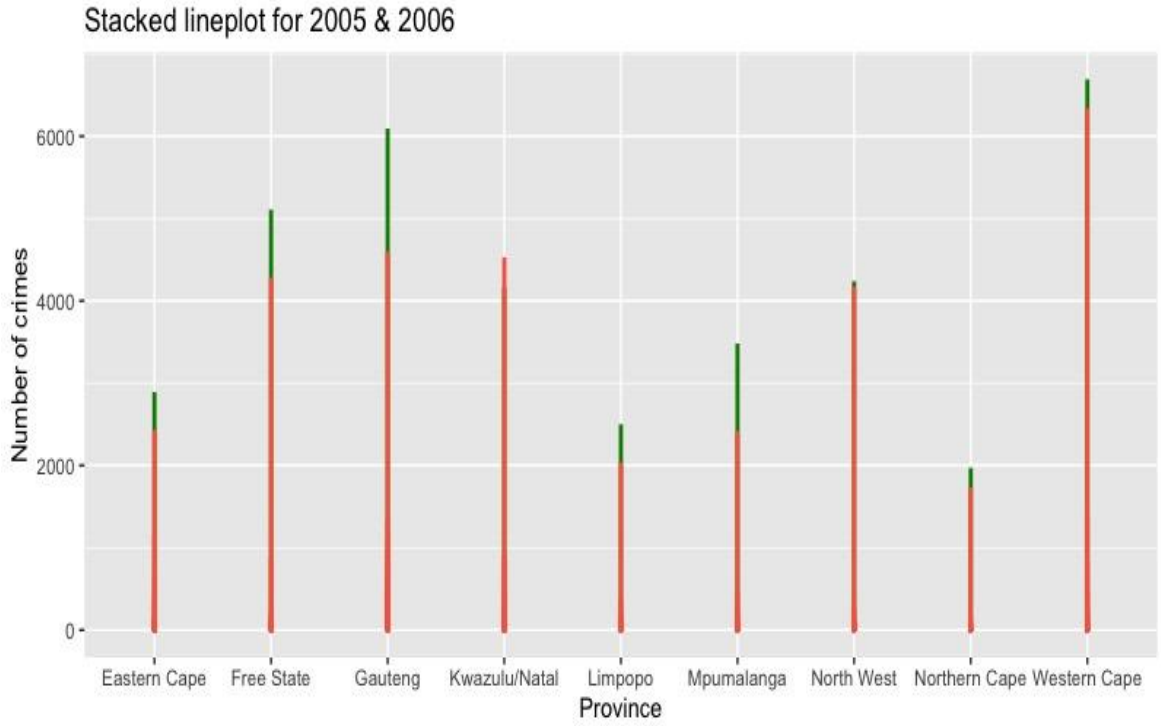
WST212 PROJECT

U19128504 - Lehlogonolo Nkadimeng

Mr. LM Nkadimeng

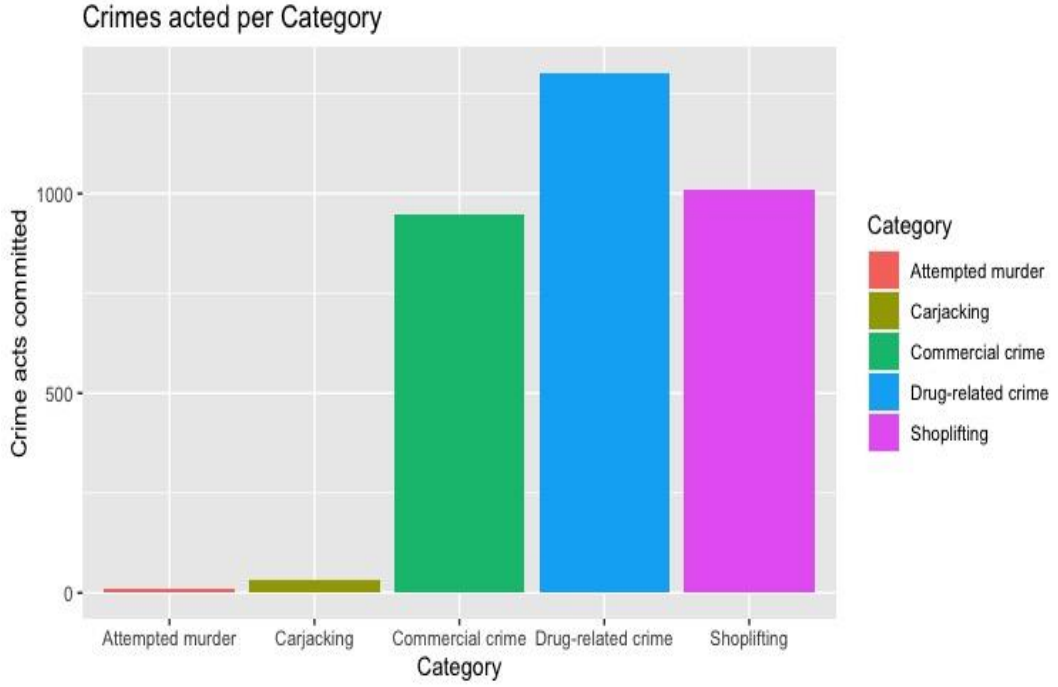
University of Pretoria

1. Show a stacked line plot of the number of crimes committed in different provinces in the years 2005 & 2006 to show the difference in crimes during this period.



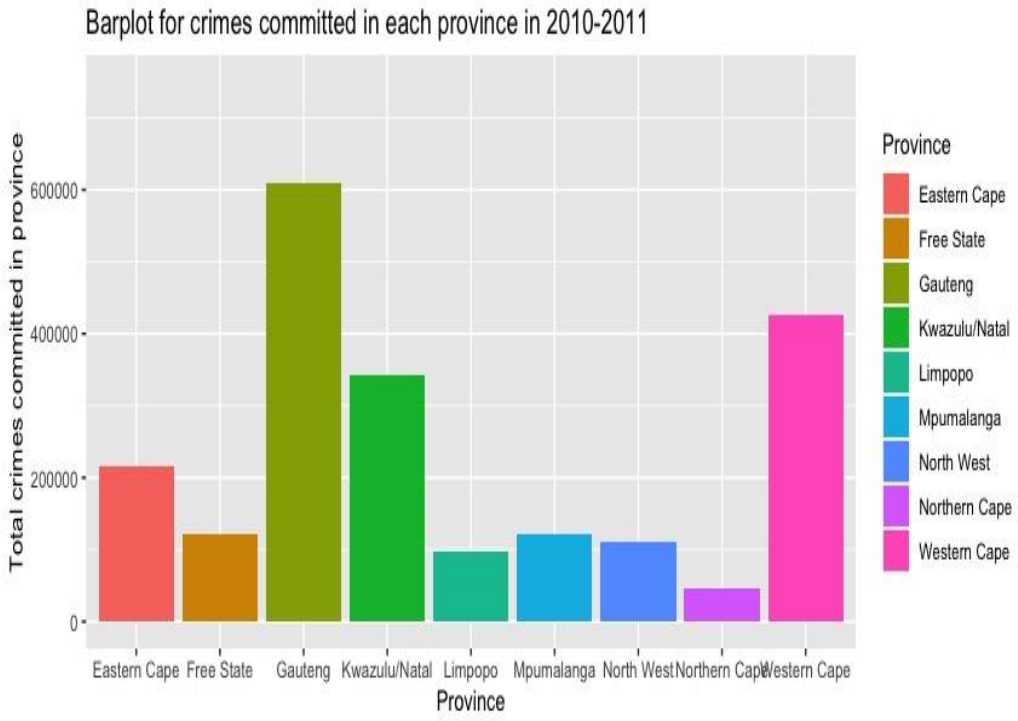
This graph clearly shows that crime occurred more in the year 2006 (green) than 2005 (tomato).

2. For the 5 categories: shoplifting, drug-related crime, attempted murder, carjacking, commercial crime, show a bar plot for the year 2006 comparing the number of crimes committed.



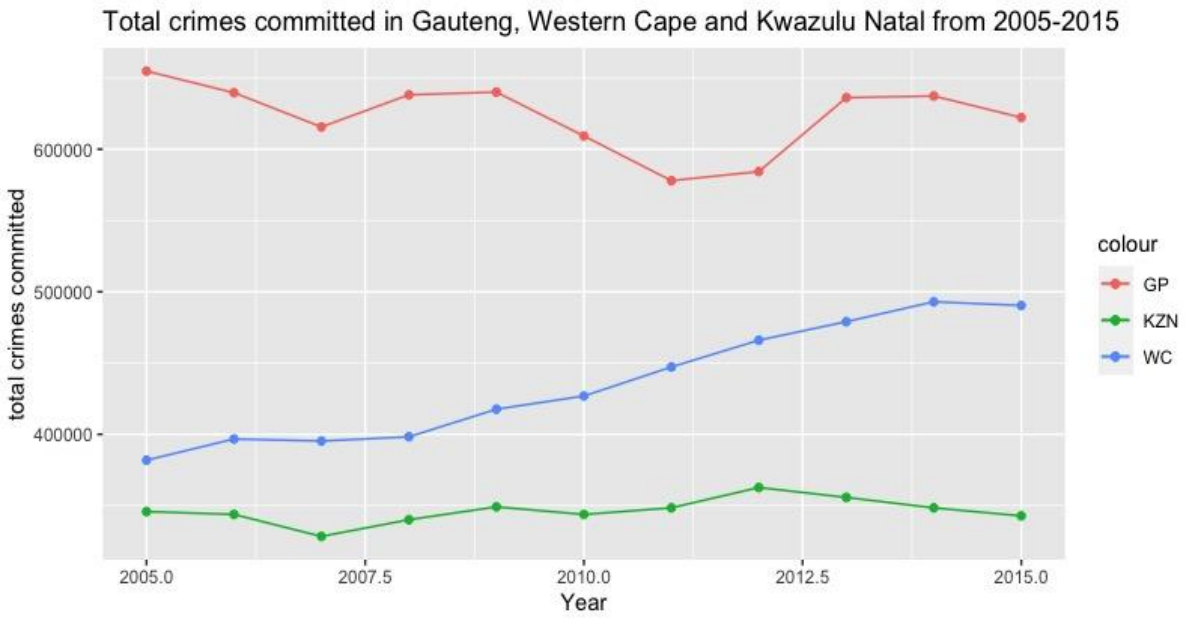
3a. In the year 2010, which province had the highest number of crimes and what is the number?
Gauteng – 609305
b. Which province had the lowest number of crimes and what is the number?
Northern Cape – 45618

c. Show a bar plot of the number of crimes reported in each province in the year 2010?



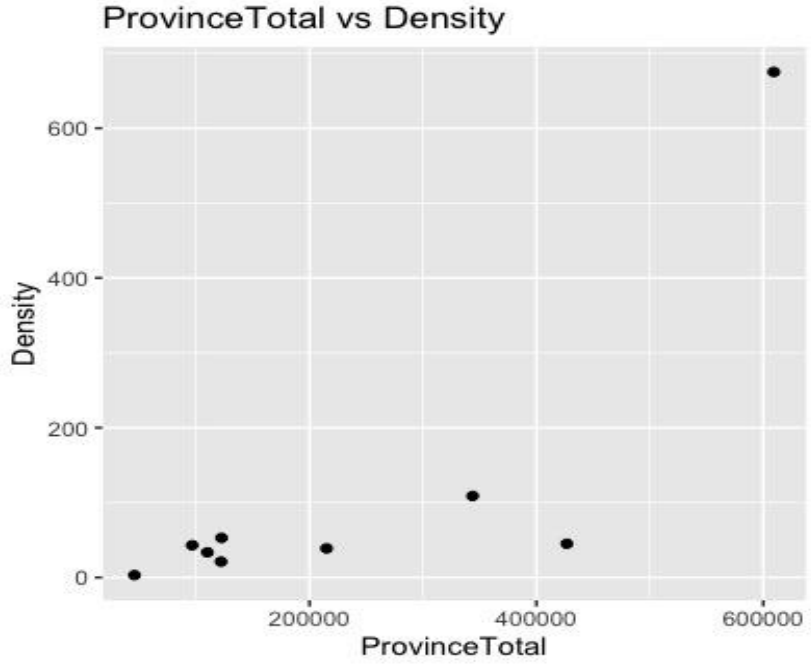
Looking at the bar plot, we can deduce that Gauteng had the highest number of crime occurrences in the year 2010 and the Western Cape had the second highest.

4. Compare Gauteng, Western Cape and KwaZulu Natal throughout the years 2005-2015 with respect to the total number of crimes reported in each province. Substantiate your answer with a graphic display.



Gauteng has the highest number of crime reports from 2005 to 2015, it fluctuates around its mean. Western Cape has the second highest crime rate of which has been increasing throughout the years. Of the three KZN has the least crime for the 10-year period. The line plots make it easy to analyse the data as we can extrapolate values for all the individual points.

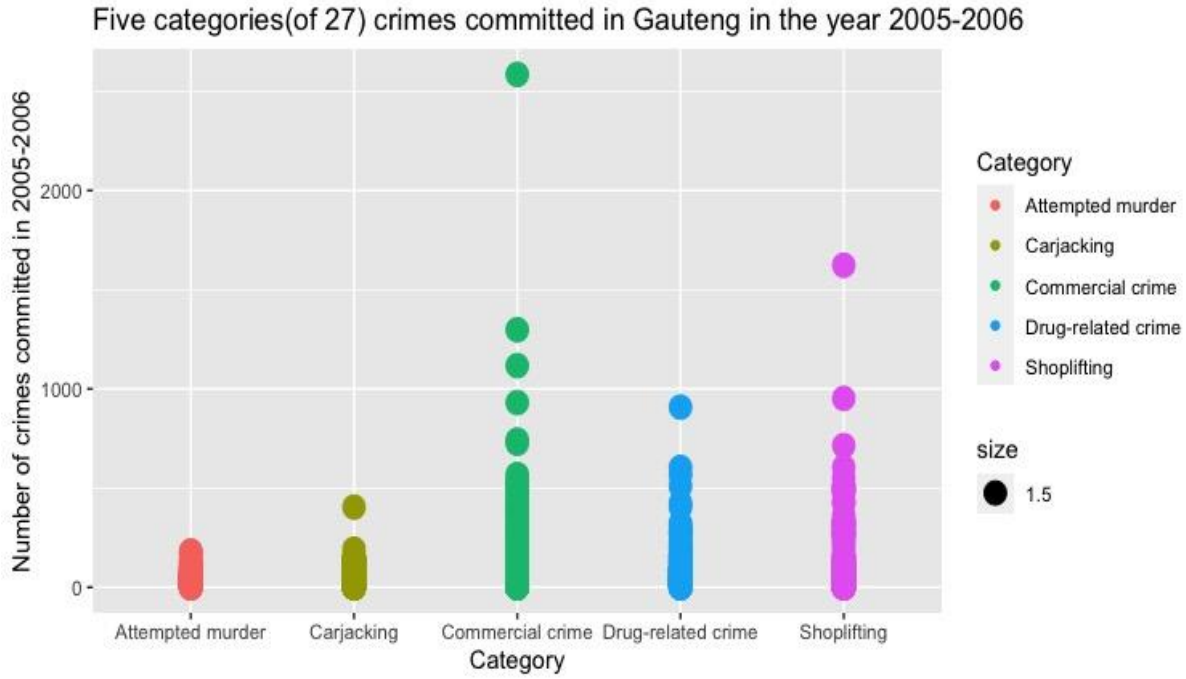
5. What is the relationship between the province population density and province crime total in 2010? Validate your answer with a graphical depiction and correlation coefficient.



Looking at the plot, one can tell the relationship is relatively positive between the two variables. The correlation coefficient is *0.7972946* which confirms a relatively strong positive relationship between the two variables in 2010. From

this we can deduce that the denser an area's population is then more crime is likely to occur.

6. From the year 2005, display the following categories: *attempted murder*, *carjacking*, *commercial crime*, *drug-related crime* and *shoplifting* and how much crime comes from each.



Commercial crime is the most popular category among these categories, followed by shoplifting.

From this dataset, one can observe that of all 9 provinces, Gauteng has the highest population density and the highest crime rate. Since the correlation coefficient suggests a strong linear positive relationship, we can conclude that dense areas are more susceptible to crime.

Appendix A: Code

```
library(readr)
library(sqldf)
library(lubridate)
library(RH2)
library(RJDBC)
library(rJava)
library(dplyr)
library(tidyr)
library(ggplot2)

# Data:

province <- read_csv('ProvincePopulation.csv')
crime_stats <- read_csv('SouthAfricaCrimeStats_v2.csv')


# Question 1

names(crime_stats) <- c('Province', 'Station', 'Category','Y05_06',
                        'Y06_07', 'Y07_08', 'Y08_09', 'Y09_10', 'Y10_11',
                        'Y11_12', 'Y12_13', 'Y13_14', 'Y14_15', 'Y15_16')

ggplot(data=crime_stats, aes(x=Province)) + geom_line(aes(y=Y05_06), size=1,
color='green4') +
  labs(x="Province", y="Years") + geom_line(aes(y=Y06_07), size=1, color='tomato') +
  ggtitle("Stacked lineplot for 2005 & 2006") + labs(y="Number of crimes")


# Question 2
q2 <- crime_stats %>% filter(Station=="Cape Town Central",
                             Category=="Shoplifting" | Category=="Drug-related crime" |
                               Category=="Attempted murder" | Category=="Carjacking" |
                               Category=="Commercial crime") %>% select(Category, Y06_07)

ggplot(q2, aes(Category, Y06_07, fill=Category)) + geom_col() + labs(x="Category", y="Crime
acts committed")+
  ggtitle("Crimes acted per Category")

# Question 3
# total crimes committed in each province in the year 2010-2011
cpp <- sqldf("SELECT Province, SUM(Y10_11) AS ProvinceTotal
     FROM crime_stats
     GROUP BY Province")
```

```
# number of highest crimes committed in a province
highest <- max(cpp$ProvinceTotal)

#number of lowest crimes committed in a province
lowest <- min(cpp$ProvinceTotal)

options(scipen = 999)
ggplot(cpp, aes(Province, ProvinceTotal, fill=Province)) + geom_col() + ylim(0, 750000) +
labs(y='Total crimes committed in province') + ggtitle("Barplot for crimes committed in each
province in 2010-2011")


################################################################################
#############
# Gauteng crimes in 2005-2006 for Shoplifting, Drugs, Attempted murder, carjacking,
commercial crime
q4 <- crime_stats %>% filter(Province=="Gauteng", Category=="Shoplifting" |
Category=="Drug-related crime" |
                    Category=="Attempted murder" | Category=="Carjacking" |
                    Category=="Commercial crime")
ggplot(q4, aes(Category, Y05_06, color=Category, size=1.5)) + geom_point() +
labs(y='Number of crimes committed in 2005-2006') + ggtitle("Five categories(of 27) crimes
committed in Gauteng in the year 2005-2006")
################################################################################
#############



################################################################################
#############

#Western Cape through the years
years <- crime_stats %>% filter(Province=="Western Cape")

tib1 <- c(sum(years$Y05_06), sum(years$Y06_07), sum(years$Y07_08), sum(years$Y08_09),
          sum(years$Y09_10), sum(years$Y10_11), sum(years$Y11_12),
sum(years$Y12_13),
          sum(years$Y13_14), sum(years$Y14_15), sum(years$Y15_16))

tib2 <- c(2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015)

# Gauteng through the years
years2 <- crime_stats %>% filter(Province=="Gauteng")

tib3 <- c(sum(years2$Y05_06), sum(years2$Y06_07), sum(years2$Y07_08),
sum(years2$Y08_09),
```

```
        sum(years2$Y09_10), sum(years2$Y10_11), sum(years2$Y11_12),
sum(years2$Y12_13),
        sum(years2$Y13_14), sum(years2$Y14_15), sum(years2$Y15_16))

# KZ through the years
years3 <- crime_stats %>% filter(Province=="Kwazulu/Natal")

tib4 <- c(sum(years3$Y05_06), sum(years3$Y06_07), sum(years3$Y07_08),
sum(years3$Y08_09),
        sum(years3$Y09_10), sum(years3$Y10_11), sum(years3$Y11_12),
sum(years3$Y12_13),
        sum(years3$Y13_14), sum(years3$Y14_15), sum(years3$Y15_16))

# Table with KZN, GP and WC
tib <- data.frame(tib2,tib1,tib3,tib4)
names(tib) <- c('Year','WC','GP','KZN')

# Graphs of all three provinces
ggplot(data=tib, aes(x=Year)) + geom_line(aes(y=WC, color='WC')) + geom_point(aes(y=WC,
color='WC')) +
  geom_line(aes(y=GP, color='GP'))+geom_point(aes(y=GP,color='GP')) +
geom_line(aes(y=KZN, color='KZN')) +
  geom_point(aes(y=KZN, color='KZN')) + labs(y="total crimes committed") +
  ggtitle("Total crimes committed in Gauteng, Western Cape and Kwazulu Natal from 2005-
2015")

##############################################################################
#############

#join density and total for province
combine <- sqldf("SELECT p.Density, c.ProvinceTotal
                 FROM province AS p
                 inner join cpp AS c
                 ON p.Province=c.Province")

# relationship between population density and province crime total
ggplot(data=combine, aes(ProvinceTotal, Density)) + geom_point() + ggtitle("ProvinceTotal
vs Density")

# the correlation coefficient for the relationship between density and total crime
corrr <- cor(combine$ProvinceTotal, combine$Density)
```