# **WST 212 PROJECT**

U21598012

DHINAZ RANGASAMY

Crime Statistics Data Analysis

# **Questions, Answers, and Interpretation:**

**Data Exploration:**

1. **Which crime category has been reported the most between 2005 and 2015?**

Chart

Description automatically generated

The crime category of “All theft not mentioned elsewhere” was reported the most between 2005 and 2015 with a total of 4120351 reports.

1. **Which crime category has been reported the least between 2005 and 2015?**

The crime category “Bank Robbery” had been reported the least between 2005 and 2015 with a total of 628 reports, as seen on the graph.

1. **Which crime category is reported most frequently in each province?**

|  |  |  |
| --- | --- | --- |
| **Province** | **Crime Category** | **Total** |
| Gauteng | All theft not mentioned elsewhere | 1273583 |
| Kwa Zulu Natal | All theft not mentioned elsewhere | 538096 |
| Eastern Cape | Assault with the intent to inflict grievous bodily harm | 335419 |
| Northern Cape | Assault with the intent to inflict grievous bodily harm | 104155 |
| Western Cape | All theft not mentioned elsewhere | 995062 |
| Free State | All theft not mentioned elsewhere | 228263 |
| Mpumalanga | All theft not mentioned elsewhere | 249806 |
| Limpopo | All theft not mentioned elsewhere | 202756 |
| North West | All theft not mentioned elsewhere | 216319 |

1. **Which crime category is reported least frequently in each province?**

|  |  |  |
| --- | --- | --- |
| **Province** | **Crime Category** | **Total** |
| Gauteng | Bank robbery | 390 |
| Kwa Zulu Natal | Bank robbery | 53 |
| Eastern Cape | Bank robbery | 27 |
| Northern Cape | Bank robbery | 6 |
| Western Cape | Bank robbery | 16 |
| Free State | Bank robbery | 36 |
| Mpumalanga | Bank robbery | 33 |
| Limpopo | Bank robbery | 15 |
| North West | Bank robbery | 52 |

**Data Modelling:**

1. **Which Province had the highest crime rate per 100000 individuals between 2005 and 2015?**

Chart, bar chart

Description automatically generated

From the graph, it is seen that Western Cape had the highest crime rate per 100000 individuals between 2005 and 2015

1. **Which Province had the lowest crime rate per 100000 individuals between 2005 and 2015?**

From the graph in Question 5, it is seen that Limpopo had the lowest crime rate per 100000 individuals between 2005 and 2015.

1. **Describe how the total number of crimes committed changed each year per province by looking at each province’s scatter plot.**

Scatter chart

Description automatically generated with low confidence

By looking at the scatter plots of the crime rate of each province between the 2005 and 2015 period, it is seen that most of the province’s crime rates per 100000 were not volatile, the only provinces which showed major change were Gauteng and Eastern Cape, Gauteng saw a decrease in its crime rate in 2006 before rising again, and the Western Cape saw a rise between 2008 and 2014.

1. **Rank the provinces in order of increasing total crime reports between 2005 and 2015 by comparing their correlation coefficients.**

Calendar

Description automatically generated

|  |  |
| --- | --- |
| Province | Correlation Coefficient |
| Eastern Cape | -0.948903 |
| Mpumalanga | -0.914478 |
| Free State | -0.7411096 |
| Northern Cape | -0.4819164 |
| Gauteng | -0.3462821 |
| North West | -0.2168084 |
| KwaZulu Natal | 0.4494197 |
| Limpopo | 0.7840876 |
| Western Cape | 0.9811574 |

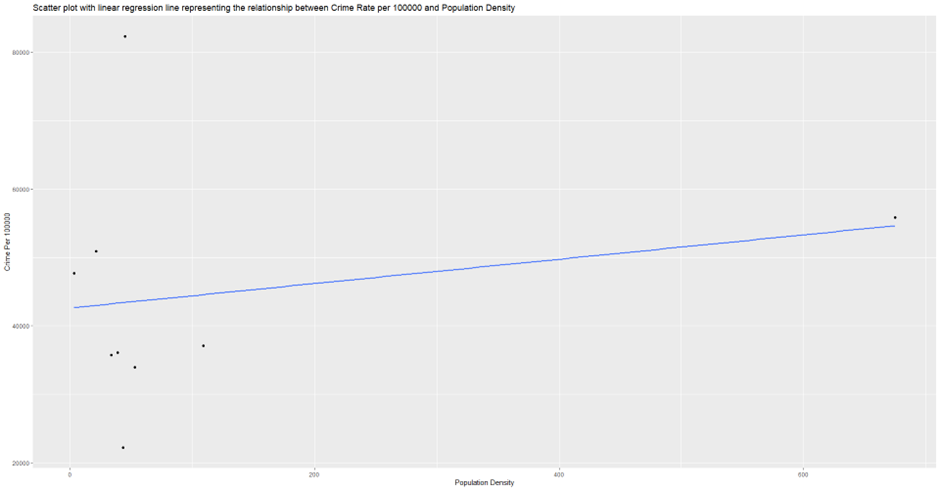
1. **Does a negative correlation coefficient necessarily mean that the province is safe, as in there is little or no crime in that province?**

No, consider the graph from Question 8, the regression coefficient of Gauteng is -0.3462821 and the regression line is angles slightly downwards, which means that the crime rate is decreasing, however, it is seen from Question 5 that Gauteng had the 2nd highest crime rate per 100000 between 2005 and 2015. The negative coefficient means that the crime is decreasing, but it remains at a relatively high level.

**10. “Criminal behaviour and criminal victimization have long been associated with population density in the criminological literature, and this has led researchers to argue that there is a positive zero-order relationship between population density and crime rates”**

**Prove the validity of this statement using:**

* 1. **A scatter plot and linear regression line without inferring from a linear regression equation or correlation coefficient.**



According to the linear regression line on the scatter plot, the crime rate and population density values do have a positive correlation, the angle of the linear regression line also indicates a positive relationship between the crime rate per 100000 and the population density.

* 1. **A Pearson correlation test.**

Using the “**cor()**” function to perform a Pearson Correlation test results in a regression coefficient of 0.2172919, which is indicates a weak, but positive relation between the crime rate per 100000 and population.

Therefore, the statement:

“Criminal behaviour and criminal victimization have long been associated with population density in the criminological literature, and this has led researchers to argue that there is a positive zero-order relationship between population density and crime rates”

Can be considered true given the correlation coefficient between the variables of interest(Crime Rate and Population Density) is positive.

**Conclusion:**

Considering the information obtained from the analysis of the raw data, it can be said that there is variation in the rate at which crimes occur in the different provinces, with the Western Cape appearing to be the most vulnerable, and Limpopo this least vulnerable. The crime category of “All theft not mentioned elsewhere” was reported the most during the period of interest. Many violent crimes have occurred between 2005 and 2015, these range from common assault to attempted murder. Monetary related crimes also appear to have occurred often, this does not include bank robberies, which is expected given that this is a high-level crime. Given the above scatter plots, Gauteng and the Western Cape can be considered high risk provinces, as these provinces have reported the highest number of total cases in the period of interest. KwaZulu Natal appears to have a moderate number of total cases per year in this time period compared to Gauteng and the Western Cape, however all three of these provinces have much higher totals than the Eastern Cape, Free State, Northern Cape, Mpumalanga, Limpopo, and North West. It is also worth noting that only Gauteng and the Western Cape appear to be volatile in terms of there total number of reported cases per year. The correlation coefficients calculated during the analysis indicated a decrease in the total number of reports per year for most provinces. The Pearson test performed on the population density variable and the crime rate per 100 000 individuals also seemed to agree with statement that there is a positive zero-order relationship between population density and crime rates.

Appendix A: Code:

library(readr)

library(sqldf)

library(lubridate)

library(RH2)

library(dplyr)

library(tidyr)

library(gapminder)

library(stringr)

library(ggplot2)

sac <- read\_csv("SouthAfricaCrimeStats\_v2.csv")

pp <- read\_csv("ProvincePopulation.csv")

colnames(sac)[4] <- "Five"

colnames(sac)[5] <- "Six"

colnames(sac)[6] <- "Seven"

colnames(sac)[7] <- "Eight"

colnames(sac)[8] <- "Nine"

colnames(sac)[9] <- "Ten"

colnames(sac)[10] <- "Eleven"

colnames(sac)[11] <- "Twelve"

colnames(sac)[12] <- "Thirteen"

colnames(sac)[13] <- "Fourteen"

colnames(sac)[14] <- "Fifteen"

c1 <- sqldf("SELECT Category,COUNT(\*) AS num, SUM(Five + Six + Seven + Eight + Nine + Ten + Eleven + Twelve + Thirteen + Fourteen + Fifteen) AS TotalOfYears

FROM sac

GROUP BY Category

ORDER BY TotalOfYears DESC

")

c2 <- ggplot(c1, aes(y=TotalOfYears,x = Category)) + geom\_bar(stat='identity',aes(fill = factor(Category)))+

theme(axis.title.x=element\_blank(),axis.text.x=element\_blank(),axis.ticks.x=element\_blank(),legend.position = c(0.8, 0.7),legend.direction = "vertical")+

labs(fill = "Crime Category")+ ggtitle("Bar graph representing the total number of crimes commited per category between 2005 and 2015 ")+

scale\_y\_continuous(labels = scales::comma)+ylab("Total Number of Crimes")

c3 <- sqldf("SELECT Province,Category,COUNT(\*) AS num, SUM(Five + Six + Seven + Eight + Nine + Ten + Eleven + Twelve + Thirteen + Fourteen + Fifteen) AS TotalOfYears

FROM sac

GROUP BY Province, Category

")

ci <- sqldf("SELECT Province,Category,COUNT(\*) AS num, SUM(Five + Six + Seven + Eight + Nine + Ten + Eleven + Twelve + Thirteen + Fourteen + Fifteen) AS TotalOfYears

FROM sac

GROUP BY Province, Category

ORDER BY Province, TotalOfYears

")

c4 <- ggplot(c3, aes(y=TotalOfYears,x = Category)) + geom\_bar(stat='identity',aes(fill = factor(Category)))+

theme(axis.title.x=element\_blank(),axis.text.x=element\_blank(),axis.ticks.x=element\_blank())+ labs(fill = "Crime Category")+ ggtitle("Bar graph representing the total number of Crimes Commited for each Province between 2005 and 2015 ")+

scale\_y\_continuous(labels = scales::comma)+ylab("Total Number of Crimes")

c5 <- sqldf("SELECT Category,Province, AVG(Five + Six + Seven + Eight + Nine + Ten + Eleven + Twelve + Thirteen + Fourteen + Fifteen) AS Average,

Five, Six, Seven, Eight, Nine, Ten, Eleven, Twelve, Thirteen, Fourteen, Fifteen

FROM sac

GROUP BY Province, Category, Five, Six, Seven, Eight, Nine, Ten, Eleven, Twelve, Thirteen, Fourteen, Fifteen

")

c6 <- ggplot(c5, aes(y=Average, x=(Five + Six + Seven + Eight + Nine + Ten + Eleven+Twelve+ Thirteen+ Fourteen+ Fifteen))) + geom\_line(stat='identity')+ facet\_wrap(~Province)

c7 <- ggplot(c8, aes(y=Average, x=Province)) + geom\_line(stat='identity')+ facet\_wrap(~Category)

c8 <- sac %>%

pivot\_longer(Five:Fifteen, names\_to = "question", values\_to = "response")

c10 <- sqldf("SELECT Province,SUM(Five + Six + Seven + Eight + Nine + Ten + Eleven + Twelve + Thirteen + Fourteen + Fifteen) As TotalP

FROM sac

GROUP BY Province

")

c11 <- sqldf("SELECT pp.Province, pp.Population, (TotalP/pp.Population)\*100000 AS Crime\_Rate\_Per\_100000

FROM pp

INNER JOIN c10

ON pp.Province = c10.Province

")

c12 <- ggplot(c11, aes(x = Province, y=Crime\_Rate\_Per\_100000))+geom\_bar(stat='identity')+ geom\_bar(stat='identity',aes(fill = factor(Province)))+

theme(axis.title.x=element\_blank(),axis.text.x=element\_blank(),axis.ticks.x=element\_blank())+ labs(fill = "Crime Category")+ ggtitle("Bar graph representing the total number of Crimes Commited for each Province between 2005 and 2015 ")+

scale\_y\_continuous(labels = scales::comma)+ylab("Crime Rate Per 100000")

c13 <- sqldf("SELECT Province,SUM(Five) As `2005`

FROM sac

GROUP BY Province

")

c14 <- sqldf("SELECT Province,SUM(Six) As `2006`

FROM sac

GROUP BY Province

")

c15 <- sqldf("SELECT Province,SUM(Seven) As `2007`

FROM sac

GROUP BY Province

")

c16 <- sqldf("SELECT Province,SUM(Eight) As `2008`

FROM sac

GROUP BY Province

")

c17 <- sqldf("SELECT Province,SUM(Nine) As `2009`

FROM sac

GROUP BY Province

")

c18 <- sqldf("SELECT Province,SUM(Ten) As `2010`

FROM sac

GROUP BY Province

")

c19 <- sqldf("SELECT Province,SUM(Eleven) As `2011`

FROM sac

GROUP BY Province

")

c20 <- sqldf("SELECT Province,SUM(Twelve) As `2012`

FROM sac

GROUP BY Province

")

c21 <- sqldf("SELECT Province,SUM(Thirteen) As `2013`

FROM sac

GROUP BY Province

")

c22 <- sqldf("SELECT Province,SUM(Fourteen) As `2014`

FROM sac

GROUP BY Province

")

c23 <- sqldf("SELECT Province,SUM(Fifteen) As `2015`

FROM sac

GROUP BY Province

")

c24 <- sqldf("SELECT pp.PROVINCE, `2005`, `2006`, `2007`

FROM pp

INNER JOIN c13

ON pp.Province = c13.Province

INNER JOIN c14

ON c13.Province = c14.province

INNER JOIN c15

ON c14.Province = c15.province

")

c25 <- sqldf("SELECT c16.Province, `2008`, `2009`, `2010`, `2011`, `2012`, `2013`, `2014`, `2015`

FROM c16

INNER JOIN c17

ON c16.Province = c17.province

INNER JOIN c18

ON c17.Province = c18.province

INNER JOIN c19

ON c18.Province = c19.province

INNER JOIN c20

ON c19.Province = c20.province

INNER JOIN c21

ON c20.Province = c21.province

INNER JOIN c22

ON c21.Province = c22.province

INNER JOIN c23

ON c22.Province = c23.province

")

c26 <- sqldf("SELECT c24.PROVINCE, `2005`, `2006`, `2007`, `2008`, `2009`, `2010`, `2011`, `2012`, `2013`, `2014`, `2015`

FROM c24

INNER JOIN c25

ON c24.PROVINCE = c25.Province

")

c27 <- c26 %>%

pivot\_longer(`2005`:`2015`, names\_to = "Year", values\_to = "Total for each year")

c28 <- as.data.frame(c26)

c27$Year <- as.character(c27$Year)

c27$Year <- factor(c27$Year, levels=unique(c27$Year))

c29 <- ggplot(c27, aes(x=Year, y=`Total for each year`))+geom\_point(aes(colour = Year))+facet\_wrap(~PROVINCE)+

theme(axis.title.x=element\_blank(),axis.text.x=element\_blank(),axis.ticks.x=element\_blank())+

ggtitle("Scatter plot representing the change in total number of Crimes Commited for each Province between 2005 and 2015 ")+

scale\_y\_continuous(labels = scales::comma)+ylab("Total reports for each year")

c47 <- sqldf("SELECT pp.Province, Density, Crime\_Rate\_Per\_100000

FROM pp

INNER JOIN c11

ON pp.Province = c11.Province

ORDER BY Density

")

c48 <- ggplot(c47, aes(x=Density, y=Crime\_Rate\_Per\_100000))+geom\_point()+

geom\_smooth(method='lm',se = FALSE, formula= y~x)+ylab("Crime Per 100000")+

xlab("Population Density")+

ggtitle("Scatter plot with linear regression line representing the relationship between Crime Rate per 100000 and Population Density")

c49 <- cor(c47$Density,c47$Crime\_Rate\_Per\_100000,method = "pearson")

c50 <- sqldf("SELECT sac.Province, sac.Station, Crime\_Rate\_Per\_100000

FROM sac

INNER JOIN c11

ON sac.Province = c11.Province

GROUP BY Station

")

c51 <- data.frame(rep(0:10,9))

colnames(c51)[1] <- "Delta(Year - 2005)"

c52 <- mutate(c51, vals = c27$`Total for each year`)

c53 <- mutate(c52, Province = c27$PROVINCE)

c54 <- ggplot(c53, aes(x=`Delta(Year - 2005)`, y=vals))+geom\_point(aes(colour = `Delta(Year - 2005)`))+

facet\_wrap(~Province)+geom\_smooth(method='lm',se = FALSE, formula= y~x)+

theme(axis.title.x=element\_blank(),axis.text.x=element\_blank(),axis.ticks.x=element\_blank())+

ggtitle("Scatter plot representing the change in total number of Crimes Commited for each Province between 2005 and 2015 ")+

scale\_y\_continuous(labels = scales::comma)+ylab("Total reports for each year")

c56 <- cor(c53$`Delta(Year - 2005)`[1:11],c53$vals[1:11])

c57 <- cor(c53$`Delta(Year - 2005)`[12:22],c53$vals[12:22])

c58 <- cor(c53$`Delta(Year - 2005)`[23:33],c53$vals[23:33])

c59 <- cor(c53$`Delta(Year - 2005)`[34:44],c53$vals[34:44])

c60 <- cor(c53$`Delta(Year - 2005)`[45:55],c53$vals[45:55])

c61 <- cor(c53$`Delta(Year - 2005)`[56:66],c53$vals[56:66])

c62 <- cor(c53$`Delta(Year - 2005)`[67:77],c53$vals[67:77])

c63 <- cor(c53$`Delta(Year - 2005)`[78:88],c53$vals[78:88])

c64 <- cor(c53$`Delta(Year - 2005)`[89:99],c53$vals[89:99])

Bibliography:

Formula for calculating crime rate:

[Crime Rate by Country 2022 (worldpopulationreview.com)](https://worldpopulationreview.com/country-rankings/crime-rate-by-country)

Relationship between Population Density and Crime Rate theory:

[POPULATION DENSITY AND CRIMINAL VICTIMIZATIONSome Unexpected Findings in Central Cities - SHICHOR - 1979 - Criminology - Wiley Online Library](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9125.1979.tb01285.x#:~:text=Criminal%20behavior%20and%20criminal%20victimization%20have%20long%20been%20associated%20with,population%20density%20and%20crime%20rates.)

[What are Zero-order, Partial, and Part Correlations? - Statistics Solutions](https://www.statisticssolutions.com/what-are-zero-order-partial-and-part-correlations/#:~:text=First%2C%20a%20zero%2Dorder%20correlation,thing%20as%20a%20Pearson%20correlation.)

Removing labels from the x-axis:

[r - Remove all of x axis labels in gg-plot - Stack Overflow](https://stackoverflow.com/questions/35090883/remove-all-of-x-axis-labels-in-ggplot)

Pivoting columns:

[R code pivot column to rows and rows to columns - Stack Overflow](https://stackoverflow.com/questions/62473674/r-code-pivot-column-to-rows-and-rows-to-columns)