HUMBER INSTITUTE OF TECHNOLOGY

AND ADVANCED LEARNING

(HUMBER COLLEGE)

Exploratory Data Analysis- R Programming.

**Canada’s Quality of Life  
Prosperity**: Prosperity involves a combination of increasing income and sustainable economic growth, leading to improved living standards and opportunities

* **Indicators**- Income and Growth, Employment and job quality
* **Data Set:** 

**Health**: A compass guiding well-being through quality years lived

* **Indicators**- [Health-adjusted life expectancy](https://www160.statcan.gc.ca/health-sante/expectancy-esperance-eng.htm)
* Data Set: 

**Social**: Human experience is shaped by how individuals interact on a day-to-day basis

* **Indicators**- [Sense of pride/belonging to Canada](https://www160.statcan.gc.ca/society-societe/canada-eng.htm)
* **Data Set:**

**Good Governance**: ensuring that all people in Canada feel safe and have access to the public services they need.

* **Indicators**- [Crime Severity Index](https://www160.statcan.gc.ca/good-governance-saine-gouvernance/crime-severity-index-indice-gravite-criminalite-eng.htm)
* **Data Set:**

**Environment**: The natural environment is the foundation of human existence, and our built environments shape the quality of life in important ways

* **Indicators**- [A](https://www160.statcan.gc.ca/good-governance-saine-gouvernance/crime-severity-index-indice-gravite-criminalite-eng.htm)ir Quality
* **Data Set:** 

**R-Code**

**Loding Necessary Libraries and data**

library('dplyr')

library('tidyverse')

library('lubridate')

library(ggplot2) # Plot the graph using ggplot2

**#Prosperity**

income=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/Income\_and\_Tax\_2020\_2021\_Canada.csv")

income\_growth=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/Province-wise-income-growth-canada.csv")

gdp=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/GDP\_Canada.csv")

**#Social**

social\_belonging=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/Social Belonging\_Canada.csv")

**#Health**

life\_expectancy=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/Life Expectancy Canada.csv")

**#Environment**

air\_pollutant=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/air-pollutant in canada.csv")

**#Good Governance**

crime\_data=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/Crimes\_Canada.csv")

#Graph-1

income = income %>% filter(Income.concept %in% c("Median total income", "Median income tax") &

Economic.family.type == "Economic families and persons not in an economic family")

plot <- ggplot(income, aes(x = REF\_DATE, y = VALUE, fill = Income.concept)) +

geom\_bar(stat = "identity", position = position\_dodge(width = 0.8), width = 0.7) +

geom\_text(aes(label = VALUE), vjust = -0.5, position = position\_dodge(width = 0.8), size = 4) + # Add data labels

labs(title = "Income Growth and Taxes (2020-2021)",

x = "Year",

y = "Value",

fill = "Income Concept") +

scale\_fill\_manual(values = c("Median total income" = "blue", "Median income tax" = "red")) +

theme\_minimal()

print(plot)

ggsave("income\_growth\_taxes\_bar\_with\_labels.png", plot, width = 10, height = 6, dpi = 300)

A graph of a number of years

Description automatically generated with medium confidence

**#Graph-2**

income\_growth$REF\_DATE <- as.factor(income\_growth$REF\_DATE)

# Calculate total income growth by province

province\_growth <- income\_growth %>%

group\_by(`Province.Name`) %>%

summarise(total\_growth = max(VALUE) - min(VALUE)) %>%

arrange(desc(total\_growth)) %>%

top\_n(5)

# Filter the data for top 5 provinces

top\_province\_data <- income\_growth %>%

filter(`Province.Name` %in% province\_growth$`Province.Name`) %>% select(REF\_DATE,Province.Name,VALUE)%>%group\_by(REF\_DATE,Province.Name,.drop = TRUE) %>%

summarise(AvgValue = mean(VALUE))

# Define custom colors for the provinces

custom\_colors <- c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd")

# Create a plot

plot <- ggplot(top\_province\_data, aes(x = REF\_DATE, y = AvgValue, color = Province.Name, shape = Province.Name)) +

geom\_point(size = 3) + # Set point size

labs(title = "Top 5 Province-Wise Average Income (2018-2022)",

x = "Year",

y = "Average Income Value",

color = "Province",

shape = "Province") +

theme\_minimal() +

theme(legend.position = "right") +

scale\_color\_manual(values = custom\_colors) +

scale\_shape\_manual(values = 1:length(unique(top\_province\_data$Province.Name)))

print(plot)

# Save the plot as an image

ggsave("top\_province\_average\_income.png", plot, width = 10, height = 6, dpi = 300)

A graph with numbers and symbols

Description automatically generated

**#Graph 3**

social\_belonging

# Filter the data for Gender "Total, all persons" and specific Sociodemographic characteristics

filtered\_data <- social\_belonging %>%

filter(Gender == "Total, all persons" & `Sociodemographic.characteristics` %in% c(

"15 to 24 years", "25 to 54 years", "55 to 64 years", "65 years and over"

))

# Create a list of pie charts for each age group

pie\_charts <- lapply(unique(filtered\_data$`Sociodemographic.characteristics`), function(age\_group) {

age\_data <- filtered\_data %>%

filter(`Sociodemographic.characteristics` == age\_group)

ggplot(age\_data, aes(x = "", fill = Indicators, y = VALUE)) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar("y", start = 0) +

labs(title = paste("Sense of Belonging in Canada -", age\_group),

fill = "Indicators") +

geom\_text(aes(label = sprintf("%.1f%%", VALUE)), position = position\_stack(vjust = 0.5)) +

theme\_minimal() +

theme(legend.position = "bottom")

})

# Print the pie charts

for (i in seq\_along(pie\_charts)) {

print(pie\_charts[[i]])

}

A pie chart with a number of percentages

Description automatically generatedA pie chart with a number of percentages

Description automatically generatedA pie chart with numbers and a percentage

Description automatically generatedA pie chart with numbers and a percentage

Description automatically generated

**#graph 5** #Health

# Filter the data for the specified characteristics

filtered\_data <- life\_expectancy %>%

filter(GEO == "Canada" & Age.group == "At birth" & Sex %in% c("Males", "Females"))

# Create a bar chart for both males and females with separate bars for each year

plot <- ggplot(filtered\_data, aes(x = REF\_DATE, y = VALUE, fill = Sex, label = VALUE)) +

geom\_bar(stat = "identity", position = "dodge") +

geom\_text(position = position\_dodge(width = 0.9), vjust = -0.5, color = "black") +

labs(title = "Life Expectancy for Males and Females (2010-2017)",

x = "Year",

y = "Life Expectancy (Years)") +

theme\_minimal() +

scale\_fill\_manual(values = c("Males" = "blue", "Females" = "pink"))

print(plot)

A graph of a number of people

Description automatically generated

**#Graph 6**

#Good Governance

crime\_data=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/Crimes\_Canada.csv")

# Group the data by REF\_DATE and Cyber-related violation and sum VALUE

grouped\_data <- crime\_data %>%

group\_by(REF\_DATE, violations) %>%

summarise(total\_crime = sum(VALUE))

# Create a line chart with different colors and legend

plot <- ggplot(grouped\_data, aes(x = REF\_DATE, y = total\_crime, group = violations, color = violations)) +

geom\_line() +

geom\_point() +

labs(title = "Growth in Crime(2018-2022)",

x = "Year",

y = "Number of Crimes") +

theme\_minimal() +

scale\_color\_brewer(palette = "Set1") + # Use a color palette for different lines

theme(legend.position = "top") # Position the legend at the top

print(plot)

A graph of growth in crime

Description automatically generated

#Environment

#Harmful substance emission

air\_pollutant=read.csv("C:/Users/USER/Dropbox/PC/Desktop/Humber -BIA/Semester -3/Big Data -2/Group Project/emissions-harmful-substances-air.csv")

# Create the line graph using ggplot

p <- ggplot(air\_pollutant, aes(x = Year)) +

geom\_line(aes(y = Mercury, color = "Mercury"), size = 1) +

geom\_line(aes(y = Lead, color = "Lead"), size = 1) +

geom\_line(aes(y = Cadmium, color = "Cadmium"), size = 1) +

labs(title = "Percentage Change of Metals Over Years",

x = "Year",

y = "Percentage Change") +

scale\_color\_manual(values = c("Mercury" = "blue", "Lead" = "red", "Cadmium" = "green")) +

theme\_minimal()

# Display the plot

print(p)

A graph showing different colored lines

Description automatically generated

**Infographic**

**A poster of a graph

Description automatically generated with medium confidence**

**Video Presentation Link:**

[**https://www.canva.com/design/DAFrVZCSbrU/MygKSpfFIbP4ktPYObXntQ/view?utm\_content=DAFrVZCSbrU&utm\_campaign=share\_your\_design&utm\_medium=link&utm\_source=shareyourdesignpanel**](https://www.canva.com/design/DAFrVZCSbrU/MygKSpfFIbP4ktPYObXntQ/view?utm_content=DAFrVZCSbrU&utm_campaign=share_your_design&utm_medium=link&utm_source=shareyourdesignpanel)

**\*Thank You\***