Based on the provided details from the assignment document, here is a step-by-step guide to completing the task. This guide is tailored for beginners, breaking down the process into manageable steps.

Step-by-Step Guide for Completing the Assignment

1. Understand the Assignment Requirements

Objective: Provide intelligence on a specific political, business, or societal problem using data analytics and visualization.

Deliverables: A report (1000 words ± 10%) and an R script for data analysis.

Sections: Executive Summary, Introduction, Data Preparation, Analysis, Conclusions, Reference List, Appendices.

2. Gather Your Tools

Software: Install R and RStudio.

Packages: Install necessary R packages (tidyverse, ggplot2, dplyr, caret, etc.).

3. Obtain Your Assigned Question and Data

Question: Retrieve your specific question from the Unit Brightspace pages.

Data: Download the publicly available dataset linked to your question.

4. Executive Summary (250 words)

Summarize the problem, your approach, key findings, and conclusions.

Write this section last, but keep it concise and informative.

5. Introduction

Relevance: Explain why the outcome variable (dependent variable) is important for the given problem.

Objective: State what you aim to achieve with your analysis.

6. Data Preparation

Data Loading: Write an R script to load your dataset.

r

Copy code

data <- read.csv("path\_to\_your\_data.csv")

Data Cleaning: Handle missing values, outliers, and data types.

r

Copy code

data <- na.omit(data) # Removing missing values

Data Description: Describe each variable (use an appendix for detailed tables).

Descriptive Statistics: Generate and describe key statistics.

r

Copy code

summary(data)

Final Dataset: Provide the number of observations used in the analysis.

7. Analysis

Visualization: Create at least one visualization of your data.

r

Copy code

library(ggplot2)

ggplot(data, aes(x=variable1, y=variable2)) + geom\_point()

Model Building: Build a prediction model (linear or non-linear).

r

Copy code

model <- lm(outcome ~ predictors, data=data)

summary(model)

Out-of-Sample Prediction: Split data into training and test sets and evaluate the model.

r

Copy code

set.seed(123)

training\_index <- createDataPartition(data$outcome, p=0.8, list=FALSE)

training\_data <- data[training\_index,]

test\_data <- data[-training\_index,]

model <- lm(outcome ~ predictors, data=training\_data)

predictions <- predict(model, newdata=test\_data)

Second Model: Optionally, build a second model with automated variable selection.

r

Copy code

library(caret)

control <- trainControl(method="cv", number=10)

model2 <- train(outcome ~ ., data=training\_data, method="lmStepAIC", trControl=control)

8. Conclusions

Model Quality: Discuss the performance of your models.

Comparison: Compare the two models if applicable and explain why they differ or align.

Improvements: Suggest possible improvements for your models.

9. Reference List

Use BU Harvard Referencing style for all references.

Consult the BU library Brightspace pages for the referencing manual.

10. Appendices

Include detailed tables, additional statistics, and extended explanations if necessary.

11. Final Checks

Code Review: Ensure your R script runs without errors and produces the expected output.

Report Review: Proofread your report, check word count, and ensure clarity and coherence.

Example R Script Structure

r

Copy code

# Load necessary libraries

library(tidyverse)

library(caret)

# Load the data

data <- read.csv("path\_to\_your\_data.csv")

# Data cleaning and preparation

data <- na.omit(data) # Handle missing values

# Additional data preparation steps...

# Descriptive statistics

summary(data)

# Visualization

ggplot(data, aes(x=variable1, y=variable2)) + geom\_point()

# Model building

set.seed(123)

training\_index <- createDataPartition(data$outcome, p=0.8, list=FALSE)

training\_data <- data[training\_index,]

test\_data <- data[-training\_index,]

model <- lm(outcome ~ predictors, data=training\_data)

summary(model)

# Predictions

predictions <- predict(model, newdata=test\_data)

# Optional second model

control <- trainControl(method="cv", number=10)

model2 <- train(outcome ~ ., data=training\_data, method="lmStepAIC", trControl=control)

# Save results

save\_json(path=Path("scores.json"), data=list(loss=model\_loss, accuracy=model\_accuracy))

Final Tips

Start Early: Begin your assignment as soon as you receive the question and dataset.

Seek Help: Utilize office hours and forums if you encounter difficulties.

Practice: Familiarize yourself with R and data analysis techniques through tutorials and practice.

This step-by-step guide should help you structure and complete your assignment effectively. If you have specific questions or need further clarification on any step, feel free to ask!