```
## LSE Data Analytics Online Career Accelerator
# DA301:  Advanced Analytics for Organisational Impact

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

# Assignment 5 scenario
## Turtle Games’s sales department has historically preferred to use R when performing
## sales analyses due to existing workflow systems. As you’re able to perform data analysis
## in R, you will perform exploratory data analysis and present your findings by utilising
## basic statistics and plots. You'll explore and prepare the data set to analyse sales per
## product. The sales department is hoping to use the findings of this exploratory analysis
## to inform changes and improvements in the team. (Note that you will use basic summary
## statistics in Module 5 and will continue to go into more detail with descriptive
## statistics in Module 6.)

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

## Assignment 5 objective
## Load and wrangle the data. Use summary statistics and groupings if required to sense-check
## and gain insights into the data. Make sure to use different visualisations such as
scatterplots,
## histograms, and boxplots to learn more about the data set. Explore the data and comment on the
## insights gained from your exploratory data analysis. For example, outliers, missing values,
## and distribution of data. Also make sure to comment on initial patterns and distributions or
## behaviour that may be of interest to the business.

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

# Import libraries
# The whole tidyverse package.
library(tidyverse)
# Create insightful summaries of the data set.
library(skimr)

# Import the data set (turtle.csv).
setwd("C:/Users/trevo/OneDrive/Documents/LSE DA/Course 3 - Advanced Analytics for Organisational
Impact/Assignment")
turtle <- read.csv("turtle.csv", header = TRUE)

# Sense-check the data set.
as_tibble(turtle)
View(turtle)

# View the result.
str(turtle)
head(turtle)

# Determine the descriptive statistics.
summary(turtle)

# Examine Distributions of Numeric Variables
turtle %>%
  pivot_longer(c(age, remuneration, spending, loyalty),
               names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 30, fill = "steelblue", colour = "white") +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Distribution of key numeric variables",
       x = NULL, y = "Count")

# Use Boxplots to flag Outliers
turtle %>%
  pivot_longer(c(age, remuneration, spending, loyalty),
               names_to = "variable", values_to = "value") %>%
```

```r
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot(fill = "lightblue") +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Boxplots highlighting outliers", x = NULL, y = NULL)

# Outlier check on loyalty (IQR method)
q1 <- quantile(turtle$loyalty, 0.25)
q3 <- quantile(turtle$loyalty, 0.75)
iqr <- q3 - q1
upper_fence <- q3 + 1.5 * iqr

sum(turtle$loyalty > upper_fence)
upper_fence
mean(turtle$loyalty)
max(turtle$loyalty)

# Categorical distributions
turtle %>% count(gender) %>%
  ggplot(aes(x = gender, y = n, fill = gender)) +
  geom_col() + labs(title = "Customers by gender", y = "Count")

turtle %>% count(education) %>%
  mutate(education = fct_reorder(education, n)) %>%
  ggplot(aes(x = education, y = n)) +
  geom_col(fill = "lightblue") +
  labs(title = "Customers by education level", y = "Count", x = NULL)

# Scatterplots: Numerical Data
ggplot(turtle, aes(x = age, y = loyalty)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, colour = "darkorange") +
  labs(title = "Loyalty points vs. age")

ggplot(turtle, aes(x = remuneration, y = loyalty)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, colour = "darkorange") +
  labs(title = "Loyalty points vs. income (remuneration)")

ggplot(turtle, aes(x = spending, y = loyalty)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, colour = "darkorange") +
  labs(title = "Loyalty points vs. spending score")


# Boxplots: Categorical Data
# Loyalty by gender
turtle %>%
  group_by(gender) %>%
  summarise(mean_loyalty = mean(loyalty),
            median_loyalty = median(loyalty),
            sd_loyalty = sd(loyalty),
            n = n())

ggplot(turtle, aes(x = gender, y = loyalty, fill = gender)) +
  geom_boxplot() + coord_flip() +
  labs(title = "Loyalty points by gender")

# Loyalty by education
turtle %>%
  group_by(education) %>%
  summarise(mean_loyalty = mean(loyalty),
            median_loyalty = median(loyalty),
            mean_spending = mean(spending),
            mean_remuneration = mean(remuneration),
            n = n()) %>%
```

```
  arrange(desc(mean_loyalty))

ggplot(turtle, aes(x = reorder(education, loyalty, median), y = loyalty)) +
  geom_boxplot(fill = "lightgreen") +
  coord_flip() +
  labs(title = "Loyalty points by education level", x = "Education")

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

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

# Assignment 5 Findings

## Age, Renumeration and Spending have no outliers
## Loyalty mean is 1,578, but 266 outliers > 3220 points and a
## max of 6,847. Strongly right-skewed distribution.
##
## Loyalty does not seem to be based on gender - but there is
## a progression with education. "Basic" have the highest mean
## loyalty points, despite being the smallest group.
##
## The high-loyalty outlier group is worthy of further
## investigation.

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


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

# Assignment 6 scenario

## In Module 5, you were requested to redo components of the analysis using Turtle Games's
preferred
## language, R, in order to make it easier for them to implement your analysis internally. As a
final
## task the team asked you to perform a statistical analysis and create a multiple linear
regression
## model using R to predict loyalty points using the available features in a multiple linear
model.
## They did not prescribe which features to use and you can therefore use insights from previous
modules
## as well as your statistical analysis to make recommendations regarding suitability of this
model type,
## the specifics of the model you created and alternative solutions. As a final task they also
requested
## your observations and recommendations regarding the current loyalty programme and how this
could be
## improved.

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

## Assignment 6 objective
## You need to investigate customer behaviour and the effectiveness of the current loyalty program
based
## on the work completed in modules 1-5 as well as the statistical analysis and modelling efforts
of module 6.
##  - Can we predict loyalty points given the existing features using a relatively simple MLR
model?
##  - Do you have confidence in the model results (Goodness of fit evaluation)
##  - Where should the business focus their marketing efforts?
##  - How could the loyalty program be improved?
##  - How could the analysis be improved?

################################################################################
```

```
################################################################################

# install.packages
library(tidyverse)
library(skimr)
library(corrplot)
library(car)      # for VIF check
library(broom)     # tidy model output

setwd("C:/Users/trevo/OneDrive/Documents/LSE DA/Course 3 - Advanced Analytics for Organisational
Impact/Assignment")
turtle <- read.csv("turtle.csv")

# Descriptive statistics

skim(turtle)

turtle %>%
  summarise(across(c(age, remuneration, spending, loyalty),
                   list(mean = mean, median = median, sd = sd,
                        skew = e1071::skewness, kurtosis = e1071::kurtosis),
                   .names = "{.col}_{.fn}"))

# Formal normality test
shapiro.test(turtle$loyalty)

# Visual check
qqnorm(turtle$loyalty, main = "Q-Q Plot: Loyalty Points")
qqline(turtle$loyalty, col = "red")

# Distributions and patterns
turtle %>%
  pivot_longer(c(age, remuneration, spending, loyalty),
               names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 30, fill = "steelblue", colour = "white") +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Distribution of numeric variables")

# Correlation matrix - candidate numeric predictors of loyalty
num_vars <- turtle %>% select(age, remuneration, spending, loyalty)
corr_mat <- cor(num_vars)
print(round(corr_mat, 2))
corrplot(corr_mat, method = "number", type = "upper")

# Multiple Linear Regression model

model <- lm(loyalty ~ remuneration + spending + age, data = turtle)
summary(model)

# Check whether age materially improves the model vs. a simpler
# two-predictor model, by comparing adjusted R-squared directly
model_simple <- lm(loyalty ~ remuneration + spending, data = turtle)
summary(model_simple)

cat("Adjusted R-squared without age:", summary(model_simple)$adj.r.squared, "\n")
cat("Adjusted R-squared with age:   ", summary(model)$adj.r.squared, "\n")
cat("Improvement in adjusted R-squared:",
    summary(model)$adj.r.squared - summary(model_simple)$adj.r.squared, "\n")

# Multicollinearity check
vif(model)

# install.packages("lmtest")  # add to your install line
```

```
library(lmtest)

par(mfrow = c(2, 2))
plot(model)
par(mfrow = c(1, 1))

shapiro.test(residuals(model))

# Model Results
glance(model)
tidy(model)

# Visualisation of the chosen MLR model
# Actual vs. predicted loyalty
turtle$predicted_loyalty <- predict(model)

ggplot(turtle, aes(x = predicted_loyalty, y = loyalty)) +
  geom_point(alpha = 0.4, colour = "steelblue") +
  geom_abline(slope = 1, intercept = 0, colour = "red", linetype = "dashed") +
  labs(title = "Actual vs. predicted loyalty points",
       x = "Predicted loyalty", y = "Actual loyalty")

# Partial relationship: loyalty vs spending, controlling visually for remuneration
ggplot(turtle, aes(x = spending, y = loyalty, colour = remuneration)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, colour = "black") +
  scale_colour_viridis_c() +
  labs(title = "Loyalty vs. spending (coloured by remuneration)")

# Predicting loyalty for specific customer scenarios

new_customers <- data.frame(
  remuneration = c(30, 60, 90),   # £'000
  spending     = c(30, 60, 90),   # spending score 1-100
  age          = c(25, 40, 55)
)

new_customers$predicted_loyalty <- predict(model, newdata = new_customers)
print(new_customers)



###############################################################################
###############################################################################
# Assignment 6 Findings

# loyalty is right-skewed (skew ~1.5) with a long tail of highly
# engaged customers - see Activity 5 outlier analysis (~13% above
# the upper IQR fence). age and remuneration are mildly right-skewed;
# spending is roughly symmetric.

# importantly, loyalty is NOT normally distributed.
# This does not disqualify linear regression - but this may point
# to a failing in the way the loyalty programme is set up.

# residual skew and kurtosis drop compared to raw loyalty. However,
# Residual diagnostics (Q-Q plot) show residuals reasonably
# close to normal, though Shapiro-Wilk still rejects strict
# normality (likely due to the large sample size making the test
# overly sensitive to minor deviations). The scale-location plot
# shows some spread in residuals at higher fitted values,
# suggesting prediction accuracy may be somewhat less reliable for
# high-loyalty customers — consistent with the skew identified in the
# raw loyalty variable.
```

```
# This skew and non-normality of the loyalty points variable
# suggests a log-based or non-linear model (e.g. a generalized
# linear model) may be theoretically more appropriate than
# a standard linear regression, though the linear model was
# retained here for its simplicity and ease of interpretation
# for the business.

# remuneration (r ~ 0.62) and spending (r ~ 0.67) both show a strong,
# roughly linear relationship with loyalty. Age is a very weak corelation.
# However, it is retained as it may add value in a multiple regression
# model.

# Multilinear Model tried with/without age. Adjusted R² without age: 0.8267
# Adjusted R² with age: 0.8397. Hence age included. Almost 84% of loyalty
# is accounted for by remuneration/spending/age. No issue with
# multicollinearity.

# Cross-validation note: the equivalent MLR model fitted in Python
# (spending + remuneration + age) produces virtually identical predictions
# for these same scenarios (~119, ~2,331, and ~4,543 loyalty points
# respectively), confirming consistency between the R and Python
# implementations of this model.

# Example interpretations
# A 25-year-old earning £30k with a spending score of 30 is predicted
# to accumulate ~119 loyalty points, while a 55-year-old earning £90k
# with a spending score of 90 is predicted to accumulate ~4,543 points -
# illustrating how strongly income and spending behaviour compound
# to drive loyalty accumulation.
```