

```
In [67]: library(tidyverse)
library(repr)
library(tidymodels)
```

Github link: https://github.com/owenkotler/dsci_100_project_kotler_group_40.git

```
In [68]: download.file("https://raw.githubusercontent.com/owenkotler/dsci_100_project_kotler
destfile = "sessions.csv")
sessions <- read_csv("sessions.csv")

download.file("https://raw.githubusercontent.com/owenkotler/dsci_100_project_kotler
destfile = "players.csv")
players <- read_csv("players.csv")
```

Rows: 1535 Columns: 5

— Column specification —

Delimiter: ","

chr (3): hashedEmail, start_time, end_time

dbl (2): original_start_time, original_end_time

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

Rows: 196 Columns: 7

— Column specification —

Delimiter: ","

chr (4): experience, hashedEmail, name, gender

dbl (2): played_hours, Age

lgl (1): subscribe

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

Data Description

In both datasets, all information was collected via a custom Minecraft server, and the following tables depict details about each observation.

In the **"players"** dataset, there are 196 observations (players) for 7 variables. There is an issue where an input under "Age" is not an integer and instead displays "NA." Furthermore, the data is limited by many players having zero recorded playtime, which offers little insight and may distort analyses that include this variable.

In the **"sessions"** dataset (not used in report), there are 1535 observations for 5 variables. There are a few cases where a session does not have an end time.

Table 1: Variables in players.csv

Variable Name	Variable Type	Data Type	Variable Description
experience	Qualitative	Factor	Experience level of player (either Beginner, Amateur, Regular, Pro, or Veteran)
subscribe	Qualitative	Logical	TRUE if player is subscribed to game-related newsletter and FALSE if not
hashedEmail	Qualitative	Character	Player's email converted into unique set of characters to maintain privacy
played_hours	Quantitative	Double	Number of hours played
name	Qualitative	Character	Name of player
gender	Qualitative	Factor	Gender of player (either Male, Female, Non-binary, Two-Spirited, Agender Other, or Prefer not to say)
Age	Quantitative	Integer	Age of player (in years)

Table 2: Variables in sessions.csv

Variable Name	Variable Type	Data Type	Variable Description
hashedEmail	Qualitative	Character	Player's email converted into unique set of characters to maintain privacy
start_time	Quantitative	Character	Starting time (day and time) of a player's session
end_time	Quantitative	Character	Ending time (day and time) of a player's session
original_start_time	Quantitative	Character	Start time in Unix form (milliseconds)
original_end_time	Qualitative	Character	End time in Unix form (milliseconds)

In [69]: *# Calculates summary statistics (mean) for each quantitative variable, rounded to 2*
Assigned calculations to players_means and sessions_means

```
players_means <- players |>
  summarise(mean_played_hours = mean(played_hours, na.rm = TRUE), mean_age = mean(A
    round(2)

players_means

sessions_means <- sessions |>
  summarise(mean_original_start_time = mean(original_start_time, na.rm = TRUE),
    mean_original_end_time = mean(original_end_time, na.rm = TRUE)) |>
    round(2)

sessions_means
```

A tibble: 1 × 2

mean_played_hours	mean_age
<dbl>	<dbl>
5.85	21.14

A tibble: 1 × 2

mean_original_start_time	mean_original_end_time
<dbl>	<dbl>
1.719201e+12	1.719196e+12

The summary statistics above depict rounded mean values for each quantitative variable.

Questions

This report will outline a specific research question, which relates to the broader research question: "Question 2: We would like to know which "kinds" of players are most likely to contribute a large amount of data so that we can target those players in our recruiting efforts." The specific question is:

Can player experience and player age predict how many hours the players have played?

The dataset includes information on each player's **experience level**, **age**, and **hours played**. These variables are used to explore whether player skill (which represents the historical playtime required to reach a specific proficiency) and life stage (which affects available time to play) together predict playtime. Since experience is categorical, it must be converted into "dummy variables" using mutate. For instance, a new variable "pro" can be created, with players labelled "Pro" assigned a value of 1 and all others a value of 0, allowing it to be used in prediction models. The data also has to be wrangled to ensure there are no observations with missing values for variables of interest.

Exploratory Data Analysis and Visualization

Previously, the players' data was loaded into R using the GitHub URL, and the means were calculated for each quantitative variable. They are shown below:

In [70]: `players_means`

A tibble: 1 × 2

mean_played_hours	mean_age
<dbl>	<dbl>
5.85	21.14

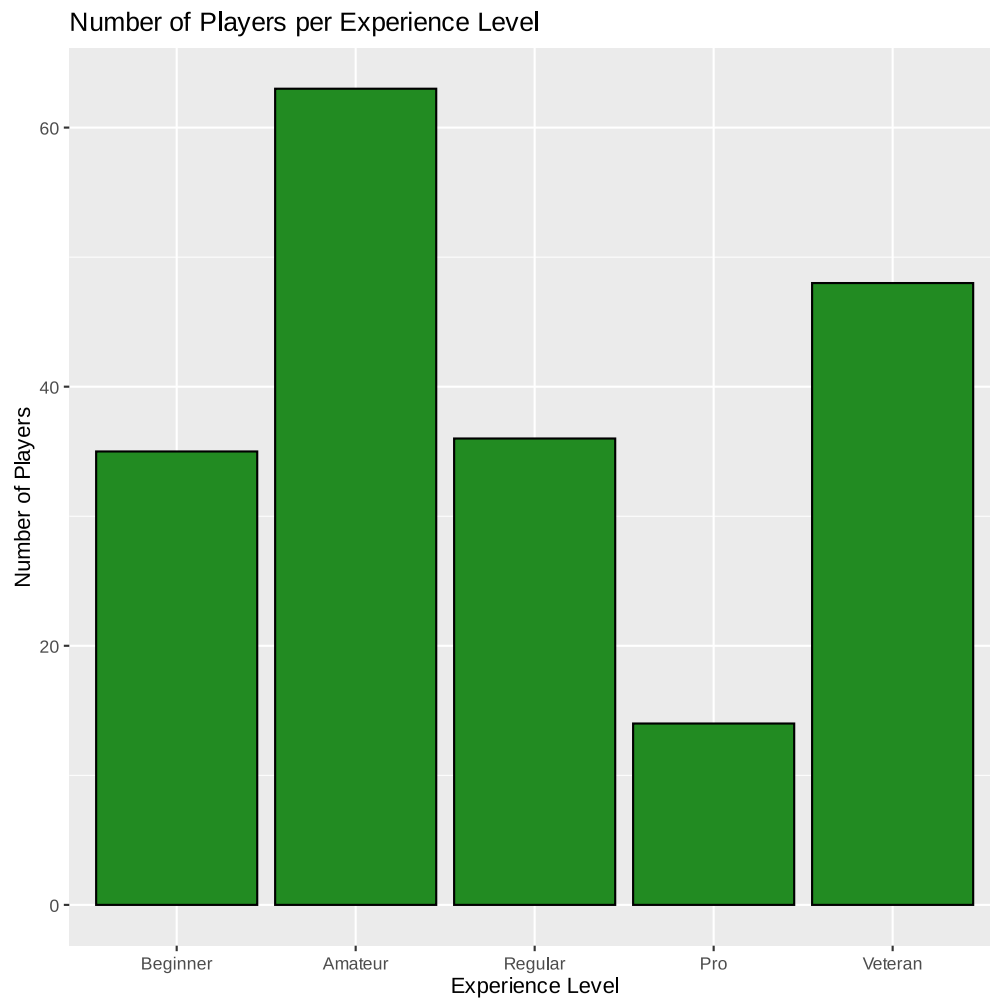
The data is already in tidy format, as each row depicts a single observation (a unique player), each column represents a single variable (age, name, etc.), and each cell contains a single value.

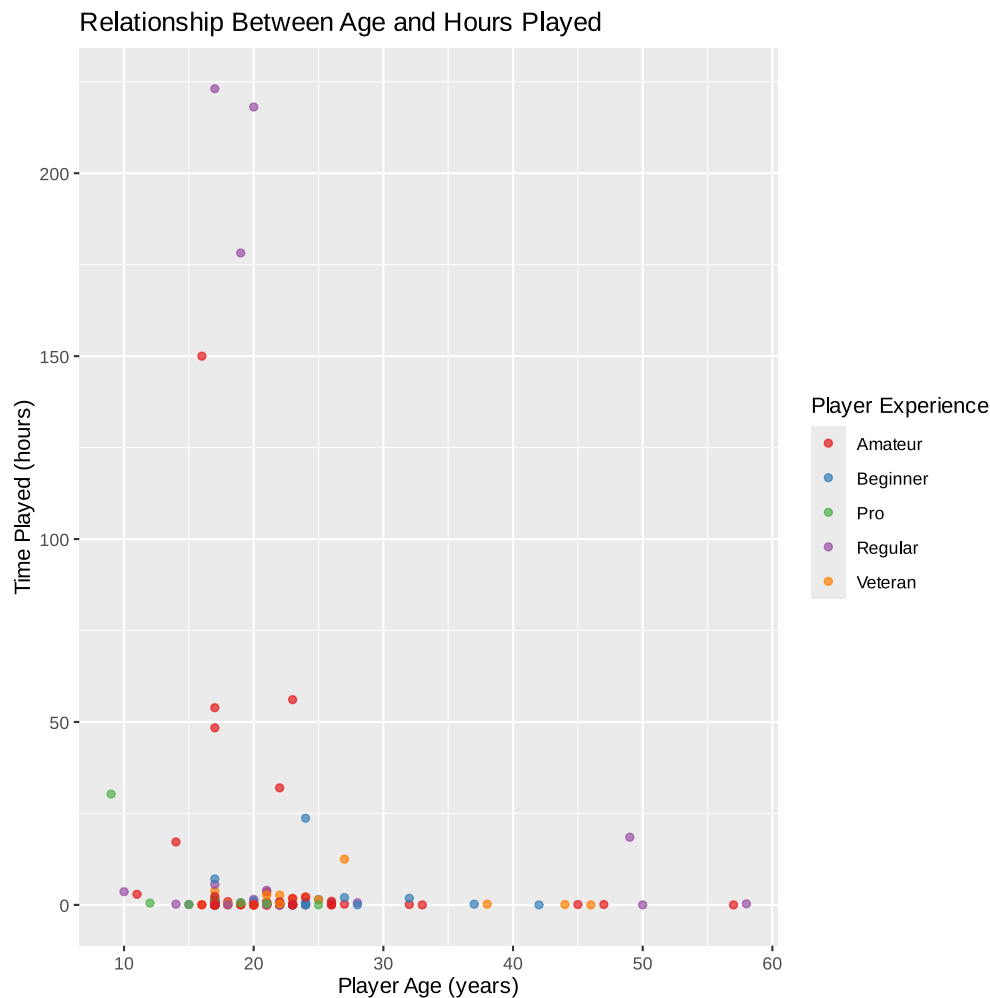
```
In [71]: ggplot(players, aes(x = factor(experience, levels = c("Beginner", "Amateur", "Regul
  geom_bar(position = "identity", color = "black", fill = "forestgreen") +
  labs(title = "Number of Players per Experience Level",
    x = "Experience Level",
    y = "Number of Players")

ggplot(players, aes(x=Age, y=played_hours, color=experience))+
  geom_point(alpha = 0.7) +
  labs(title = "Relationship Between Age and Hours Played",
    x = "Player Age (years)",
    y = "Time Played (hours)",
    color = "Player Experience") +
  scale_color_brewer(palette="Set1")
```

Warning message:

“Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).”





The bar chart shows that experience levels are relatively balanced, ensuring adequate representation across categories for KNN analysis. However, the Pro group is underrepresented and could be oversampled to improve the model.

The scatterplot shows that most players with non-zero playtime are between 15 and 25 years old. However, there is no visible relationship between age and experience, suggesting a potentially weak or no association.

Methods and Plan

The data exhibit non-linear relationships and contain numeric variables; thus, I will use KNN regression, which captures non-linearities and predicts numerical values from quantitative variables. It has few assumptions, but does assume that similar players have similar playtimes, as we saw among players aged 15–25. However, KNN can perform poorly with many predictors, and converting the categorical experience variable into dummy variables increases them. Another limitation is irrelevant variables (ie, if experience has no connection with hours). Experience will be converted into dummy variables, and all predictors will be centred and scaled. The dataset will then be split 70/30 into training and testing sets, and 10-fold cross-validation will be performed on the training set to select the k value that minimizes RMSE and yields the best model.

