2022 - Data Analytics for Immersive Environments - CA4 - RDBMS & Linear Regression Project

CA4 Part B - Linear Regression Analysis

Joe O'Regan

2023-01-16

Repo Link

https://github.com/joeaoregan/2022_DAIE_CA4_JOR1

```
if(!require("readr"))
  install.packages("readr")
```

```
## Loading required package: readr
```

```
if(!require("dplyr"))
  install.packages("dplyr")
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
if(!require("ggplot2"))
  install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```r
if(!require("knitr"))
  install.packages("knitr")
```

```
## Loading required package: knitr
```

```r
if(!require("kableExtra"))
  install.packages("kableExtra")
```

```
## Loading required package: kableExtra

## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
library(readr) # read_csv()
library(dplyr) # sample_n()
library(ggplot2) # plot linear regression
library(knitr) # Display data in tables
library(kableExtra) # Format tables
```

Dependent Variable: avg_monthly_hrs_gaming **Independent Variable:** avg_monthly_expenditure_dlc

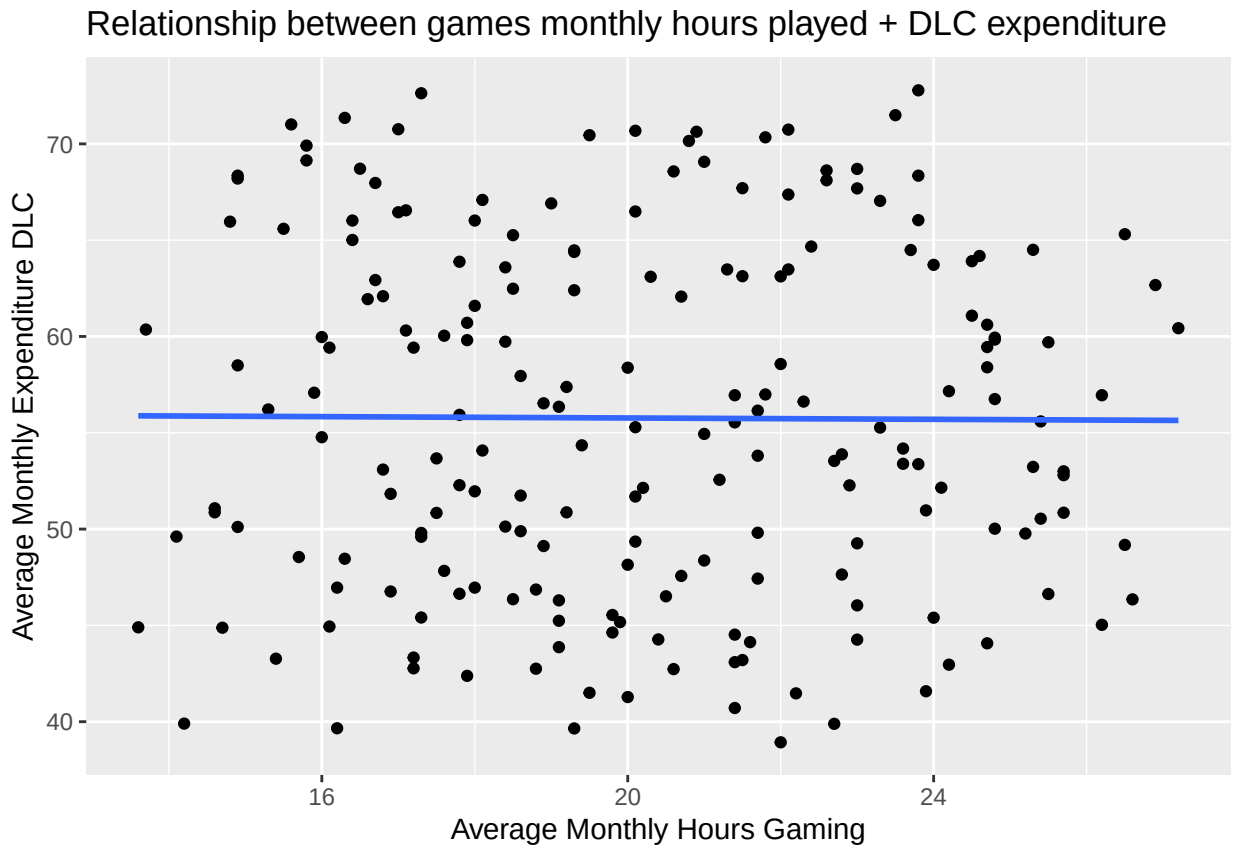
```r
data <- read_csv("amalgamated_game_survey_250_2022.csv")
```

```
## Rows: 250 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): gender, top_reason_gaming, gaming_platform, favourite_game, ethnici...
## dbl (4): age, avg_monthly_hrs_gaming, avg_years_playing_games, avg_monthly_e...
##
## 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.
```

```r
# kbl(data)

#kbl(sample_n(data, 200))
sample_data <- sample_n(data, 200) # tibble 200 x 11

# sample_data %>%
#   lm(avg_monthly_hrs_gaming ~ avg_monthly_expenditure_dlc, data = .) %>%
#   summary() # data summary

# lm() -
# dependent var. ~ independent var.
mod <- lm(avg_monthly_expenditure_dlc ~ avg_monthly_hrs_gaming, data = sample_data)
summary(mod)
```

```
##
## Call:
## lm(formula = avg_monthly_expenditure_dlc ~ avg_monthly_hrs_gaming,
##     data = sample_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8034  -7.7089  -0.3223   7.8506  17.0780
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              56.11804    3.96425   14.16   <2e-16 ***
## avg_monthly_hrs_gaming -0.01748    0.19348   -0.09    0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.115 on 198 degrees of freedom
## Multiple R-squared:  4.123e-05,  Adjusted R-squared:  -0.005009
## F-statistic: 0.008165 on 1 and 198 DF,  p-value: 0.9281
```

```r
#attributes(mod)
#mod$residuals
# hist(mod$residuals)

plot <- ggplot(data = mod, mapping = aes(x = avg_monthly_hrs_gaming,
                                         y = avg_monthly_expenditure_dlc)) +
  # geom_point(alpha = 0.1, color = "blue") # add colours for points
  geom_point() +
  labs(title = "Relationship between games monthly hours played + DLC expenditure",
       x = "Average Monthly Hours Gaming",
       y = "Average Monthly Expenditure DLC")

plot + geom_smooth(method = lm, se = FALSE, formula=y~x)
```




```
# + geom_abline(mapping = aes(x = avg_monthly_hrs_gaming, y = avg_monthly_expenditure_dlc), data = mod)
```