CMM703 - Data Analysis Coursework

Kaneel Dias

2025-04-14

```r
library(ggplot2)
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```r
library(glue)
```

TASK 1: CANDY DATASET

The objective of this exercise would be to determine the effect that the variables contained within the dataset have on the target value `winpercent`. All visualizations included in this report will be made with that objective in mind.

```r
candy_data <- read.csv("~/CMM703/candy-data.csv")
```

1.1 Effect of the categorical variables

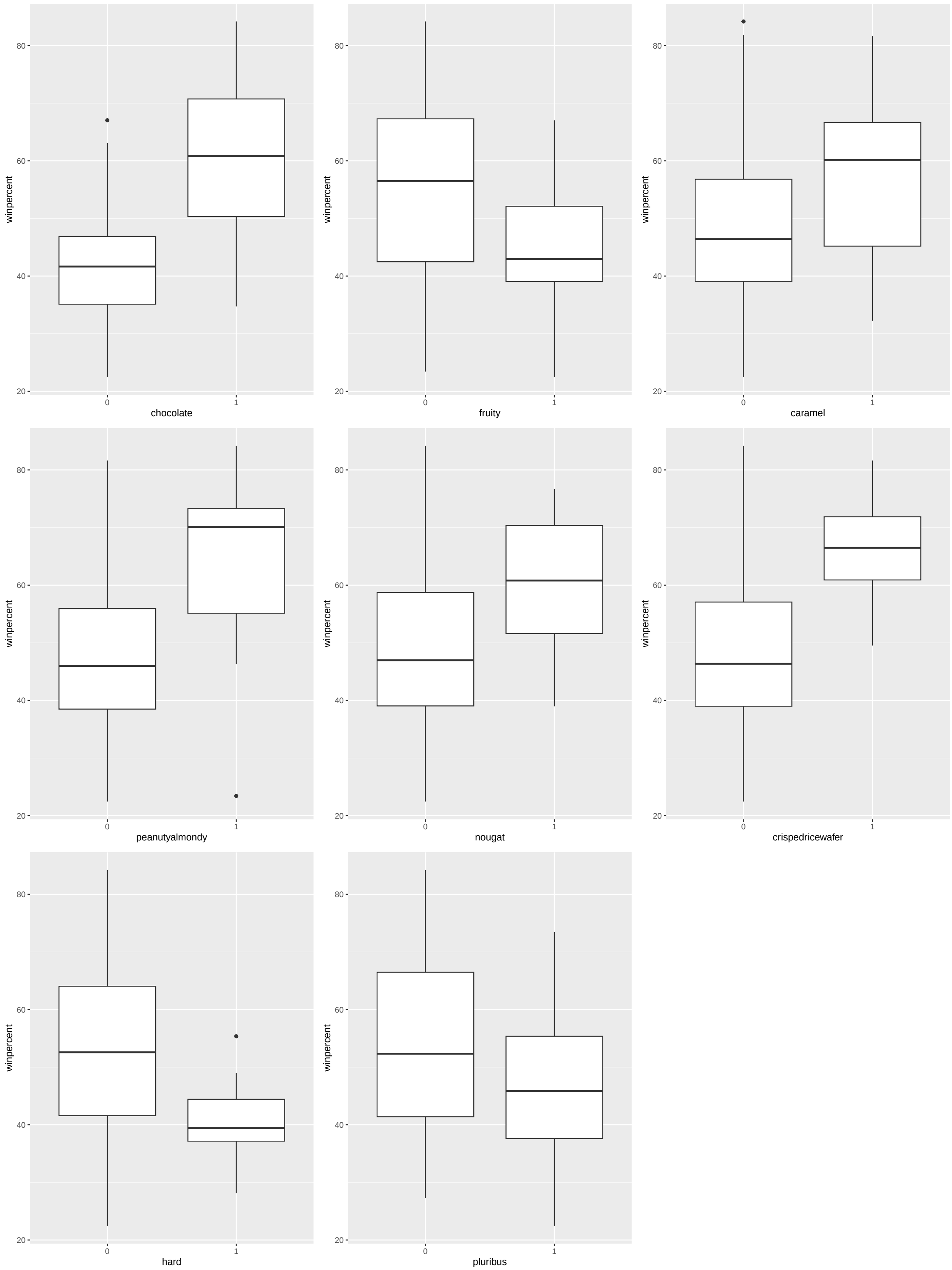
First we should convert all the categorical columns, from numeric to factor.

```r
candy_data$chocolate <- as.factor(candy_data$chocolate)
candy_data$fruity <- as.factor(candy_data$fruity)
candy_data$caramel <- as.factor(candy_data$caramel)
candy_data$peanutyalmondy <- as.factor(candy_data$peanutyalmondy)
candy_data$nougat <- as.factor(candy_data$nougat)
candy_data$crispedricewafer <- as.factor(candy_data$crispedricewafer)
candy_data$hard <- as.factor(candy_data$hard)
candy_data$bar <- as.factor(candy_data$bar)
candy_data$pluribus <- as.factor(candy_data$pluribus)
```

Next, we can plot boxplots for each categorical variable, and the effect it has on the winning percentage.

```r
plot_categorical_boxplot <- function(data, variable) {
  plot <- ggplot(data=data, aes(x=data[,variable], y=winpercent)) +
    xlab(variable) +
    geom_boxplot()

  return(plot)
}

plot_all_categorical_boxplots <- function(data) {
  chocolate_plot <- plot_categorical_boxplot(data, "chocolate")
  fruity_plot <- plot_categorical_boxplot(data, "fruity")
  caramel_plot <- plot_categorical_boxplot(data, "caramel")
  peanutyalmondy_plot <- plot_categorical_boxplot(data, "peanutyalmondy")
  nougat_plot <- plot_categorical_boxplot(data, "nougat")
  crispedricewafer_plot <- plot_categorical_boxplot(data, "crispedricewafer")
  hard_plot <- plot_categorical_boxplot(data, "hard")
  pluribus_plot <- plot_categorical_boxplot(data, "pluribus")


  grid.arrange(chocolate_plot, fruity_plot, caramel_plot, peanutyalmondy_plot, nougat_plot, crispedrice
}
```

```r
plot_all_categorical_boxplots(candy_data)
```


winpercent
80
60
40
20
0
1
chocolate
fruity
caramel
peanutyalmondy
nougat
crispedricewafer
hard
pluribus

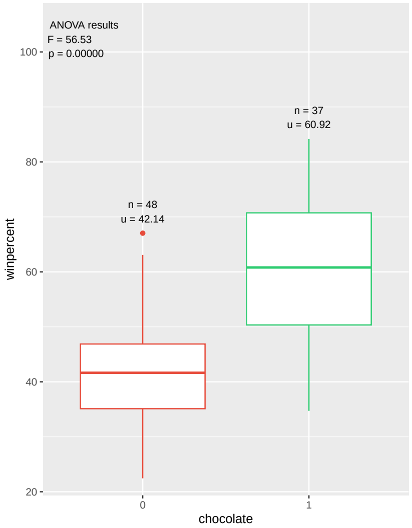
1.1.1 Potential improvements

We can suggest the following improvements
- Visually separate the `contains (1)` and `does not contain (0)` plots using colours
- Indicate for each boxplot, the
- Count of records with that value
- The mean of the `winpercent` value
- Indicate the effect that variable has on the `winpercent` using ANOVA
- Include the F value
- Include the P value

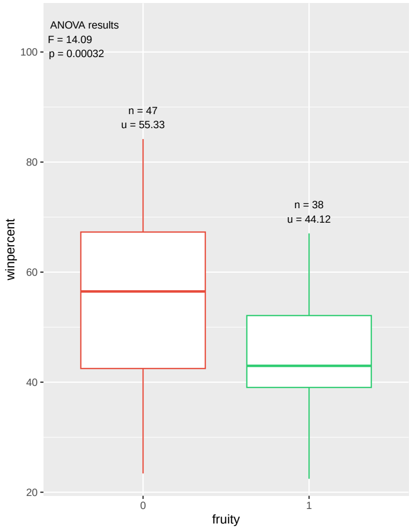
```
get_summary_stats <- function(y) {
  bxp_stats <- boxplot.stats(y)
  upper_whisker <- bxp_stats$stats[5]
  max_val <- max(c(upper_whisker, bxp_stats$out), na.rm = TRUE)

  n <- length(y)
  q1 <- quantile(y, 0.25, na.rm = TRUE, names = FALSE)
  avg <- mean(y, na.rm = TRUE)
  q3 <- quantile(y, 0.75, na.rm = TRUE, names = FALSE)

  label_str <- paste(
    paste("n =", n),
    paste("u =", round(avg, 2)),
    sep = "\n"
  )

  return(data.frame(
    y = max_val,
    label = label_str
  ))
}

plot_categorical_boxplot_improved <- function(data, variable) {
  anova <- aov(reformulate(variable, "winpercent"), data=data)
  p_val <- summary(anova)[[1]][["Pr(>F)"]][1]
  f_val <- summary(anova)[[1]][["F value"]][1]
  anova_text <- glue("ANOVA results\nF = {round(f_val, 2)}\np = {format(round(p_val, 5), nsmall = 5)}")

  plot <- ggplot(data=data, aes(x=data[,variable], y=winpercent, col=data[,variable])) +
    labs(
      title = glue('Effect of {variable} on win percentage'),
      x = variable
    ) +
    geom_boxplot() +
    scale_color_manual(values = c("#e74c3c", "#2ecc71")) +
    theme(
      legend.position="none",
      plot.title = element_text(color = "#0099f8", size = 12, face = "bold", hjust = 0.5),
    )

  plot <- plot + stat_summary(
    fun.data = get_summary_stats,
    geom = "text",
    hjust = 0.5,
    vjust = -0.5,
    size = 3,
    color = "black"
  )

  plot <- plot + annotate(
    geom = "text",
    x = -Inf,
    y = Inf,
    label = anova_text,
    hjust = -0.1,
    vjust = 1.5,
    size = 3,
    color = "black"
  ) +
    scale_y_continuous(expand = expansion(mult = c(0.05, 0.4))) # More space at top


  return(plot)
}

plot_all_categorical_boxplots_improved <- function(data) {
  chocolate_plot <- plot_categorical_boxplot_improved(data, "chocolate")
  fruity_plot <- plot_categorical_boxplot_improved(data, "fruity")
  caramel_plot <- plot_categorical_boxplot_improved(data, "caramel")
  peanutyalmondy_plot <- plot_categorical_boxplot_improved(data, "peanutyalmondy")
  nougat_plot <- plot_categorical_boxplot_improved(data, "nougat")
  crispedricewafer_plot <- plot_categorical_boxplot_improved(data, "crispedricewafer")
  hard_plot <- plot_categorical_boxplot_improved(data, "hard")
  pluribus_plot <- plot_categorical_boxplot_improved(data, "pluribus")


  grid.arrange(chocolate_plot, fruity_plot, caramel_plot, peanutyalmondy_plot, nougat_plot, crispedricew
}
```

```
plot_all_categorical_boxplots_improved(candy_data)
```

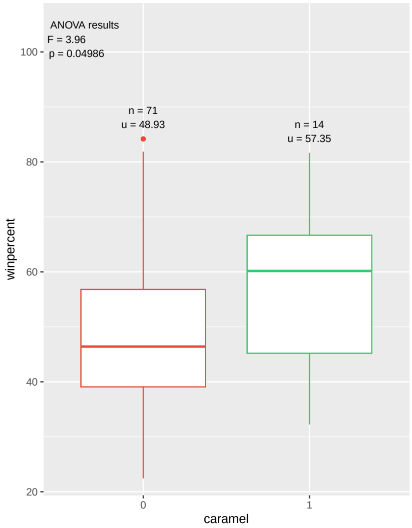
Effect of chocolate on win percentage



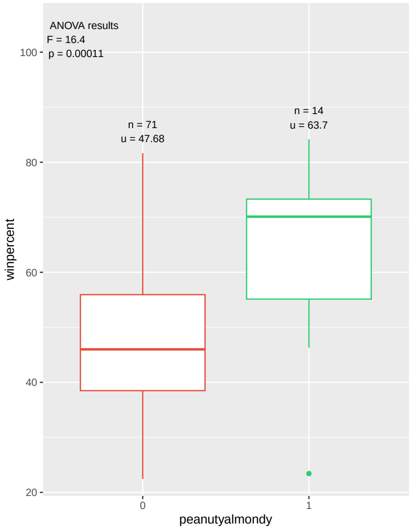
Effect of fruity on win percentage



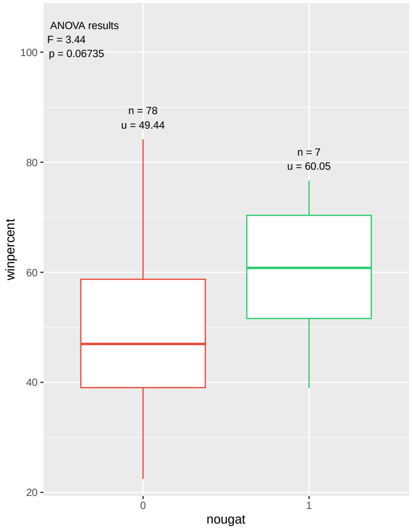
Effect of caramel on win percentage



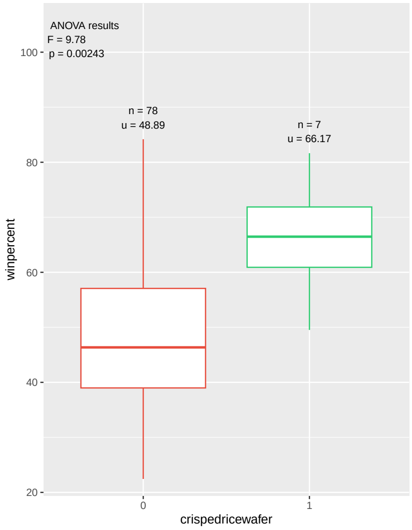
Effect of peanutyalmondy on win percentage



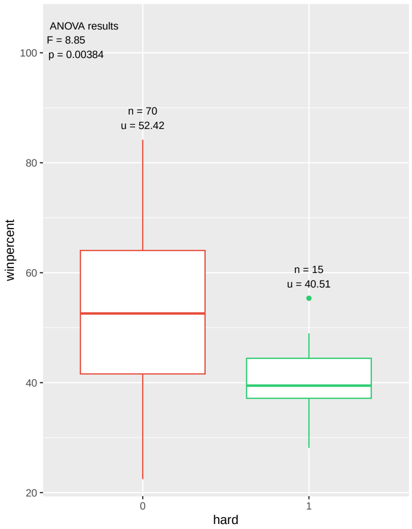
Effect of nougat on win percentage



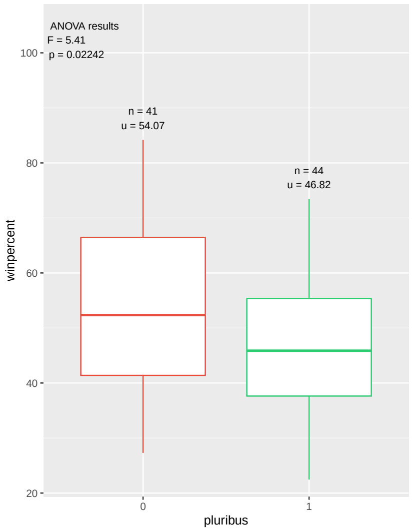
Effect of crispedricewafer on win percentage



Effect of hard on win percentage



Effect of pluribus on win percentage



1.1.2 Insights

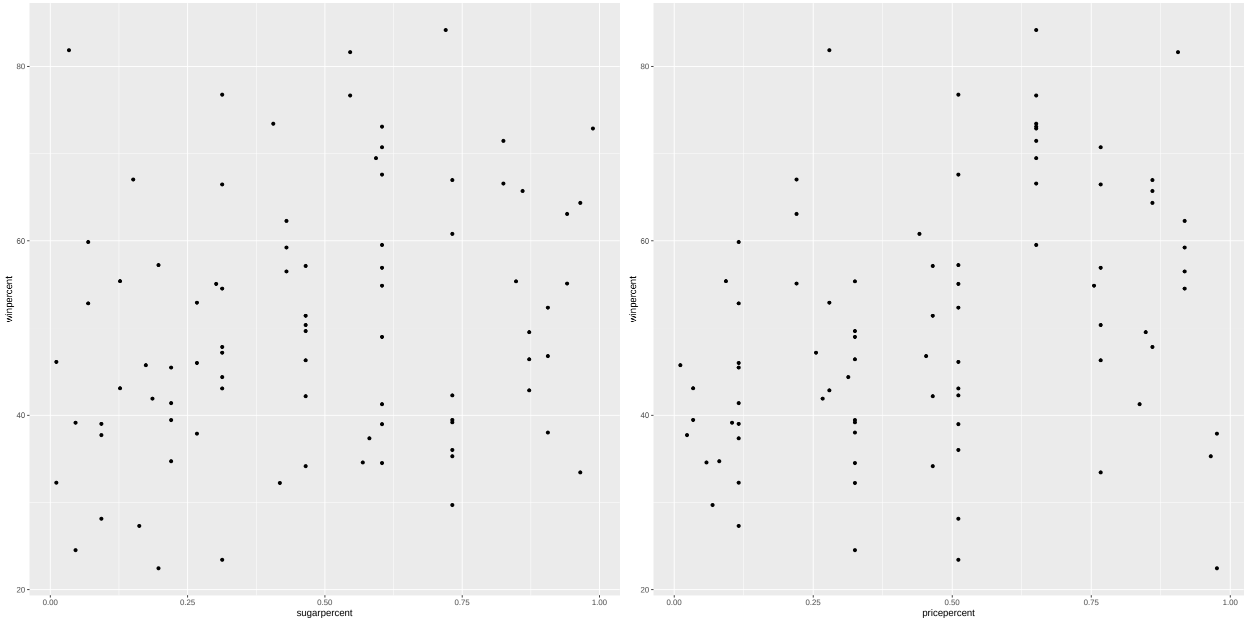
Here we can see that the `chocolate` variable has the highest effect on `winpercent` (highest f-value), and it has a very low p-value as well, indicating that it is most likely to be having an effect. On the other hand, `nougat` has a p-value > 0.05, (as well as a low f-value) which indicates that its effect on the winning percentage is not likely.

1.2 Effect of the numeric variables

There are two numeric, continuous variables: `sugarpercent` and `pricepercent` We can visualize their effect on `winpercent` using scatter plots.

```
plot_numeric_scatterplot <- function(data, variable) {
  plot <- ggplot(data=candy_data, aes(x=data[,variable], y=winpercent)) +
    labs(
      x = variable
    ) +
    geom_point()

  return (plot)
}

plot_all_numerical_scatterplots <- function(data) {
  sugar_plot <- plot_numeric_scatterplot(data, "sugarpercent")
  price_plot <- plot_numeric_scatterplot(data, "pricepercent")



  grid.arrange(sugar_plot, price_plot, nrow = 1)
}
```

```
plot_all_numerical_scatterplots(candy_data)
```



```
get_r2 <- function (x, y) cor(x, y) ^ 2
```

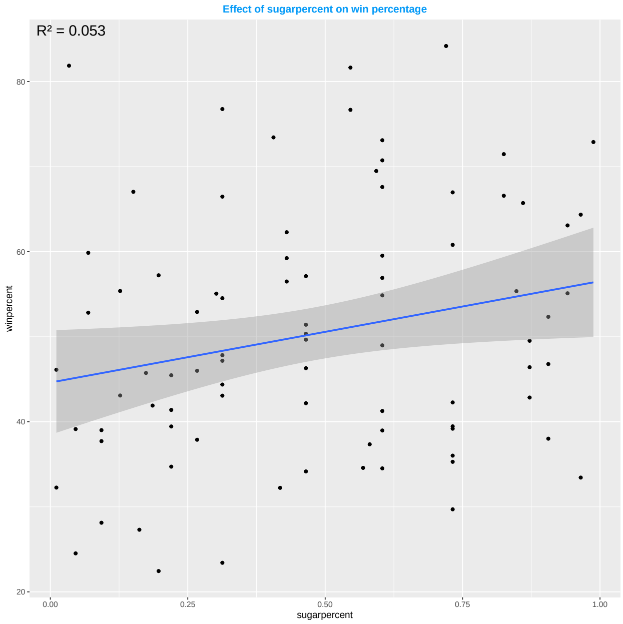
1.2.1 Potential Improvements

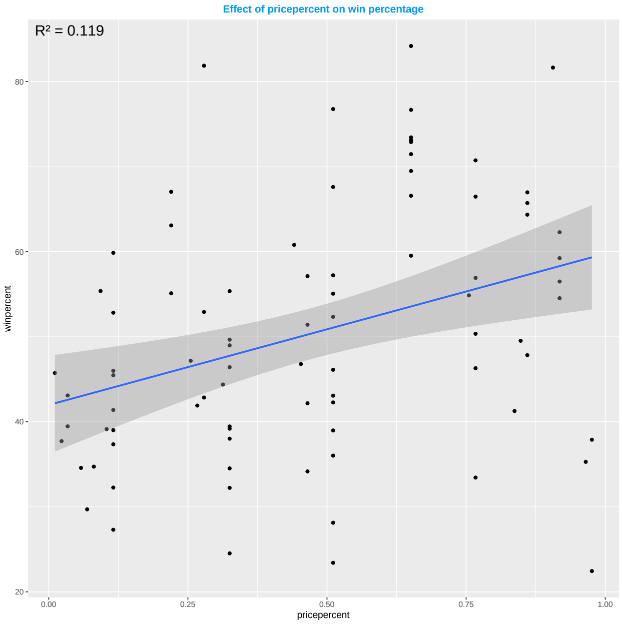
We can suggest the following improvements:
- Add a regression line to be able to view the relationship between the two variables and the winning percentage
- Include the R^2 value (coeffiecient of determination in the chart) to determine whether the they correlate

```r
plot_numeric_scatterplot_improved <- function(data, variable) {
  r2 <- get_r2(data[,variable], data$winpercent)
  r2_text <- glue("R² = {format(round(r2, 3), nsmall = 3)}")


  plot <- ggplot(data=candy_data, aes(x=data[,variable], y=winpercent)) +
    labs(
      title = glue('Effect of {variable} on win percentage'),
      x = variable
    ) +
    geom_point() +
    geom_smooth(method=lm) +
    theme(
      legend.position="none",
      plot.title = element_text(color = "#0099f8", size = 12, face = "bold", hjust = 0.5),
    )

  plot <- plot + annotate(
    geom = "text",
    x = -Inf,
    y = Inf,
    label = r2_text,
    hjust = -0.1,
    vjust = 1.5,
    size = 6,
    color = "black"
  )

  return (plot)
}

plot_all_numerical_scatterplots_improved <- function(data) {
  sugar_plot <- plot_numeric_scatterplot_improved(data, "sugarpercent")
  price_plot <- plot_numeric_scatterplot_improved(data, "pricepercent")


  grid.arrange(sugar_plot, price_plot, nrow = 1)
}
```

```r
plot_all_numerical_scatterplots_improved(candy_data)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```





1.2.2 Insights

From the above two scatter plots, even though we can see a slight positive correlation with `winpercent` for each of the two variables, they are quite insignificant. Therefore, we can conclude that thereis no significant correlation present.

TASK 2: sdlkfsdjf

Blah balh