Homework 3

Juhi Malwade

2025-09-25

Read in packages and load the dataset:

```r
library(dplyr)
library(ggplot2)
library(colorfindr)
library(tidyr)

df = read.csv('homework3_data.csv')
head(df)
```

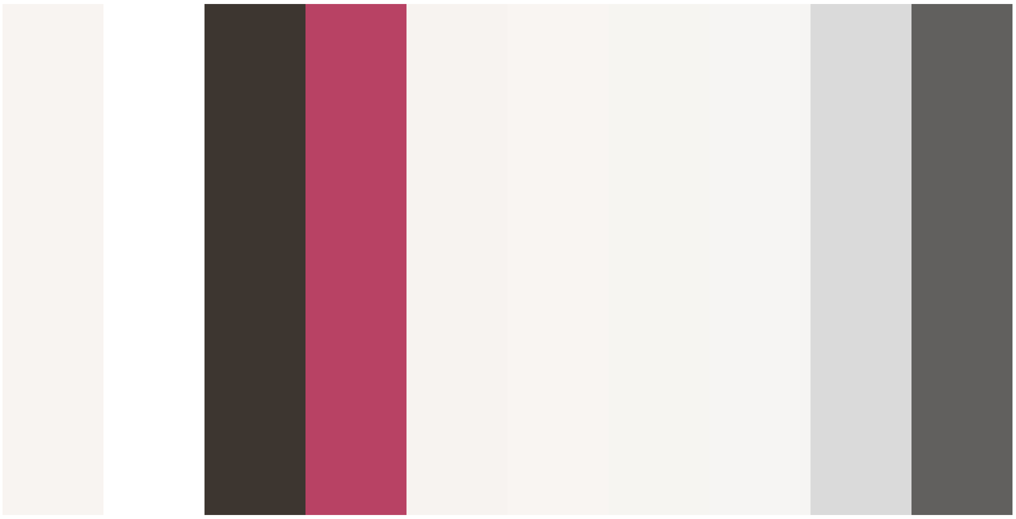
```
##      sales design items nps
## 1 32.55146      0     2   4
## 2 35.38214      0     4   5
## 3 30.87418      0     3   4
## 4 35.54265      0     3   6
## 5 32.07379      0     2   6
## 6 31.55580      0     1   4
```

Create color palette from Dunkin website:

```r
dat = get_colors("Dunkin_website.jpg")
dat
```

```
## # A tibble: 220,722 x 3
##    col_hex col_freq col_share
##    <chr>      <int>     <dbl>
##  1 #F8F4F1  1754611   0.378
##  2 #FFFFFF  1308728   0.282
##  3 #3D3630    72341   0.0156
##  4 #B84264    45093   0.00973
##  5 #DF692B    16948   0.00366
##  6 #FFFFFD    15632   0.00337
##  7 #F7F3F0    13389   0.00289
##  8 #FEFEFE    12651   0.00273
##  9 #F9F4F1    12644   0.00273
## 10 #F9F5F2    12597   0.00272
## # i 220,712 more rows
```

```
cols <- make_palette(dat[1:100,])
```

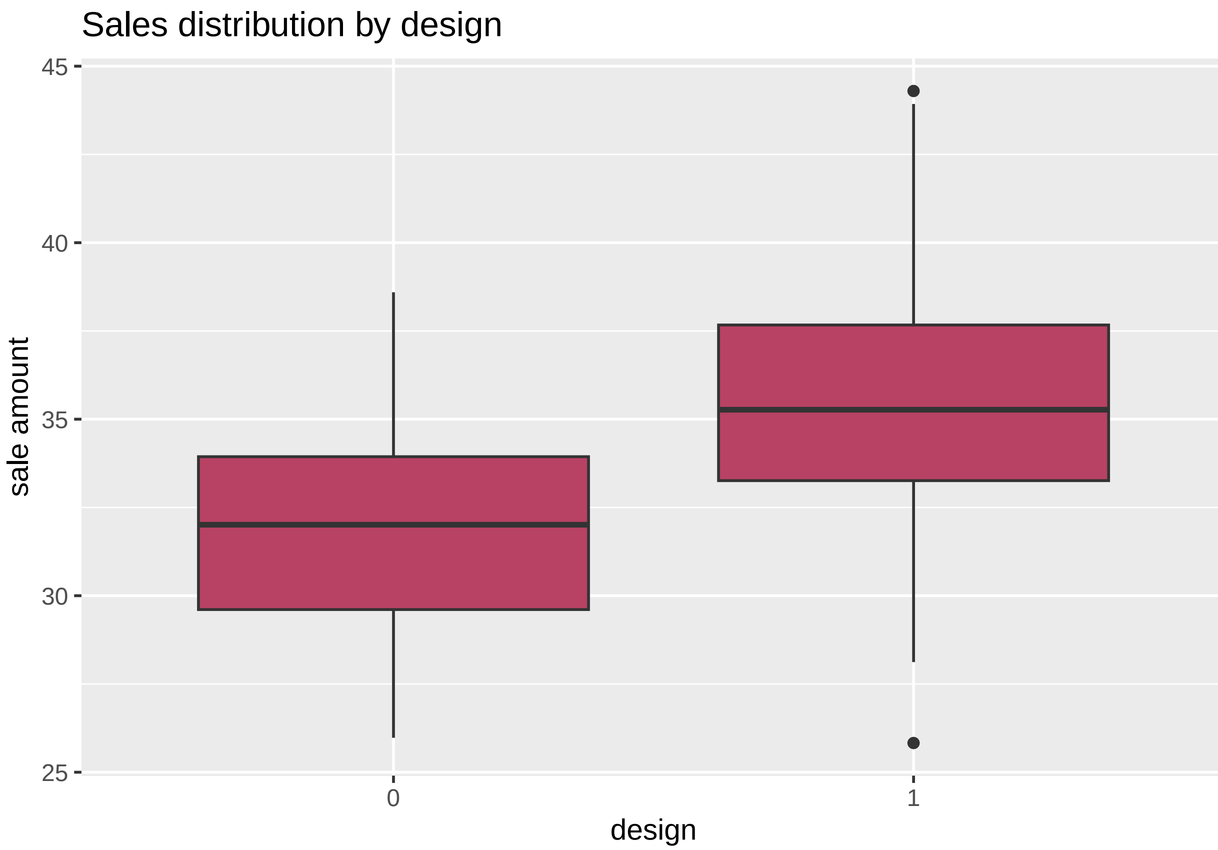


```
cols
```

```
##  [1] "#F8F4F1" "#FFFFFF" "#3D3630" "#B84264" "#F7F3F0" "#F9F5F2" "#F6F5F1"
##  [8] "#F6F5F3" "#DADADA" "#61605E"
```

Preliminary Analysis

```
ggplot(df, aes(x = factor(design), y = sales))  +
  geom_boxplot(fill = '#B84264') +
  labs(x = "design", y = "sale amount", title = "Sales distribution by design")
```



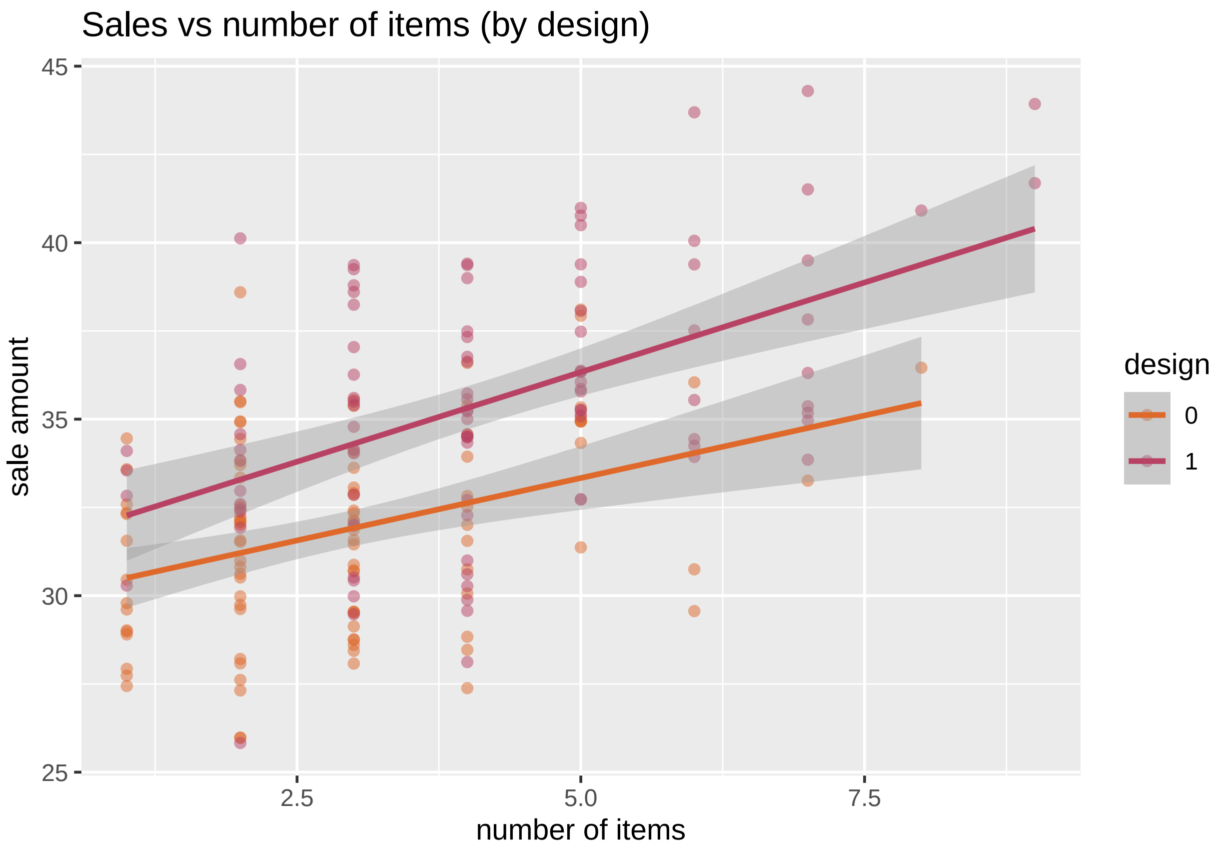

```r
df %>%
  group_by(design) %>%
  summarize(median_sales = median(sales))
```

```
## # A tibble: 2 x 2
##   design median_sales
##    <int>        <dbl>
## 1      0         32.0
## 2      1         35.3
```

The boxplot shows that the median sale amount is 32.01 dollars for the old design and 35.27 dollars for the new design. This is 4.26 dollar difference in median sales where the new design performed better.

```r
#2: Buying more expensive items for newer design
ggplot(df, aes(x = items, y = sales, color = factor(design))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(color = "design",
       title = "Sales vs number of items (by design)",
       x = "number of items",
       y = "sale amount") +
  scale_color_manual(values = c("1" = '#B84264', "0" = "#DF692B"))
```

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



The scatterplot shows that customers bought more expensive items when using the new design. This indicates the new design may have successfully brought in more sales compared to the old design.

```r
#3: t-test (Welch)
t_res <- t.test(sales ~ design, data = df)
t_res
```

```
##
##  Welch Two Sample t-test
##
## data:  sales by design
## t = -8.1554, df = 186.01, p-value = 5.042e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.551445 -2.778364
## sample estimates:
## mean in group 0 mean in group 1
##        31.84819        35.51309
```

The Welch Two Sample t-test shows that we are 95% confident that the true difference in mean sales between the old design and new design is between 2.78 and 4.55. Given this evidence, the redesign will increase sales well above the 1.80 benchmark.

```r
#4
lm1 = lm(sales ~ design + items + nps, data = df)
summary(lm1)
```

```
##
## Call:
## lm(formula = sales ~ design + items + nps, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5410 -1.5464  0.2717  1.4048  5.6221
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.56166    0.84566  23.132   <2e-16 ***
## design       0.32413    0.36733   0.882    0.379
## items        0.97917    0.09423  10.391   <2e-16 ***
## nps          2.05170    0.16316  12.575   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 196 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6644
## F-statistic: 132.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```r
confint(lm1)
```

```
##                  2.5 %    97.5 %
## (Intercept) 17.8939041 21.229412
## design      -0.4002951  1.048546
## items        0.7933313  1.165012
## nps          1.7299313  2.373477
```

When we control for NPS and number of items the design effect shrinks to 0.32 with a confidence interval that contains zero (but not statistically signficant). However, items and nps seem to be strong drivers of sales. Given the results of the t-test, this makes me curious about if design instead more directly drives the number of items and nps, which then drives sales.

```r
lm2 = lm(items ~ design + items + nps, data = df)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on
## the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 2 in
## model.matrix: no columns are assigned
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = items ~ design + items + nps, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2743 -1.0341 -0.1406  0.8846  5.0183
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.5164     0.5883   5.978 1.04e-08 ***
## design        1.4262     0.2585   5.518 1.07e-07 ***
## nps          -0.1337     0.1230  -1.087    0.278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.602 on 197 degrees of freedom
## Multiple R-squared:  0.1459, Adjusted R-squared:  0.1373
## F-statistic: 16.83 on 2 and 197 DF,  p-value: 1.786e-07
```

```r
confint(lm2)
```

```
##                  2.5 %    97.5 %
## (Intercept)  2.3563440 4.6764971
## design       0.9165118 1.9359567
## nps         -0.3762248 0.1088765
```

```r
lm3 = lm(nps ~ design + items + sales, data = df)
summary(lm3)
```

```
##
## Call:
## lm(formula = nps ~ design + items + sales, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05686 -0.46746  0.09051  0.41400  2.43147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.63758    0.51892  -3.156  0.00185 **
## design       0.52154    0.11394   4.577 8.35e-06 ***
## items       -0.23778    0.03424  -6.945 5.45e-11 ***
## sales        0.21764    0.01731  12.575  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6899 on 196 degrees of freedom
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.571
## F-statistic: 89.29 on 3 and 196 DF,  p-value: < 2.2e-16
```

```r
confint(lm3)
```

```
##                  2.5 %     97.5 %
## (Intercept) -2.6609514 -0.6142014
## design       0.2968420  0.7462449
## items       -0.3053084 -0.1702572
## sales        0.1835045  0.2517694
```

The results show that switching to the new design increases number of items by 1.43 on average, keeping all other variables constant. It also increases nps by 0.52 on average, keeping all other variables constant. None of their confidence intervals contain zero. Therefore, it seems like the resdesign increases sales, but indirectly by increasing the number of items and nps.

Final Analysis:

I estimate customers using the new design will spend on average \$3.66 more per transaction compared to the old design, with a 95% confidence interval of \$2.78–\$4.55. This comfortably exceeds the \$1.80 threshold and indicates the company should commit to a full website redesign. While the redesign doesn't contribute to an increase in sales directly once number of items and NPS are controlled for, it significantly increases both the number of items purchased (+1.43 items) and NPS (+0.52 points). Since both factors are statistically signficant predictors of increasing sales, these indirect pathways explain the higher sales. Therefore, the evidence shows that the redesign is worth pursuing.

Final Recommendation:

The company should commit to the redesign because it will lead to an average increase in sales that is greater than \$1.80 per customer.

Alternative Statement:

The redesign does not increase sales by at least \$1.80 per customer, even though the data suggests that it would.