Business Statistics Assignment IB94X0 2022-2023 #2

2215107

Contents

```r
library(Hmisc)
library(ggplot2)
library(lubridate)
library(grid)
library(gridExtra)
library(knitr)
library(emmeans)
library(car)
library(tidyverse)
library(ggpubr)
library(wesanderson)
library(corrplot)
options(width=100)
```

This is to certify that the work I am submitting is my own. All external references and sources are clearly acknowledged and identified within the contents. I am aware of the University of Warwick regulation concerning plagiarism and collusion.

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done an appropriate reduction in the mark I might otherwise have received will be made

---

Question 1

Section 1

Data Dictionary

Variables	Description
int__success__all	Originally “Total%ofInterventionsachieved(premisesratedA-E”
int__success__a	Originally “Total%ofInterventionsachieved-premisesratedA”
int__success__b	Originally “Total%ofInterventionsachieved-premisesratedB”
int__success__c	Originally “Total%ofInterventionsachieved-premisesratedC”
int__success__d	Originally “Total%ofInterventionsachieved-premisesratedD”
int__success__e	Originally “Total%ofInterventionsachieved-premisesratedE”
professional__fte	Originally “ProfessionalFullTimeEquivalentPosts-occupied *”
total__est__rated	Total establishment minus not yet rated and outside program
est__per__fte	No. of establishment per employee (total__est__rated/professional__fte)

Importing and cleaning data

```r
# Import data and convert any strings to NA
data.food <- read_csv("2019-20-enforcement-data-food-hygiene.csv", na = c("NR", "NP"))
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details, e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 353 Columns: 36
## -- Column specification ----------------------------------------------------------------------------
## Delimiter: ","
## chr  (3): Country, LAType, LAName
## dbl (32): Totalestablishments(includingnotyetrated&outside), Establishmentsnotyetratedforinterve...
## num  (1): TotalnumberofestablishmentssubjecttoWrittenwarnings
##
## 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
# Rename column for better understanding - see data dictionary
data.food.clean <- data.food %>%
  rename(int_success_all = `Total%ofInterventionsachieved(premisesratedA-E)`,
         int_success_a = `Total%ofInterventionsachieved-premisesratedA`,
         int_success_b = `Total%ofInterventionsachieved-premisesratedB`,
         int_success_c = `Total%ofInterventionsachieved-premisesratedC`,
         int_success_d = `Total%ofInterventionsachieved-premisesratedD`,
         int_success_e = `Total%ofInterventionsachieved-premisesratedE`,
         professional_fte = `ProfessionalFullTimeEquivalentPosts-occupied *`) %>%
  mutate(total_est_rated =  `Totalestablishments(includingnotyetrated&outside)`  - Establishmentsnotyet
         est_per_fte = total_est_rated/professional_fte)
```

```r
# Pivot and summarise mean for all rating
data.food.pivot.all <- data.food.clean %>%
  summarise(avg_int_success_all = mean(int_success_all, na.rm = TRUE),
            avg_fte = mean(professional_fte, na.rm = TRUE),
            avg_est_per_fte = mean(est_per_fte, na.rm = TRUE))

# Pivot and summarise mean for each rating
data.food.pivot.rate <- data.food.clean %>%
  summarise(avg_int_success_a = mean(int_success_a, na.rm = TRUE),
            avg_int_success_b = mean(int_success_b, na.rm = TRUE),
            avg_int_success_c = mean(int_success_c, na.rm = TRUE),
            avg_int_success_d = mean(int_success_d, na.rm = TRUE),
            avg_int_success_e = mean(int_success_e, na.rm = TRUE))
```

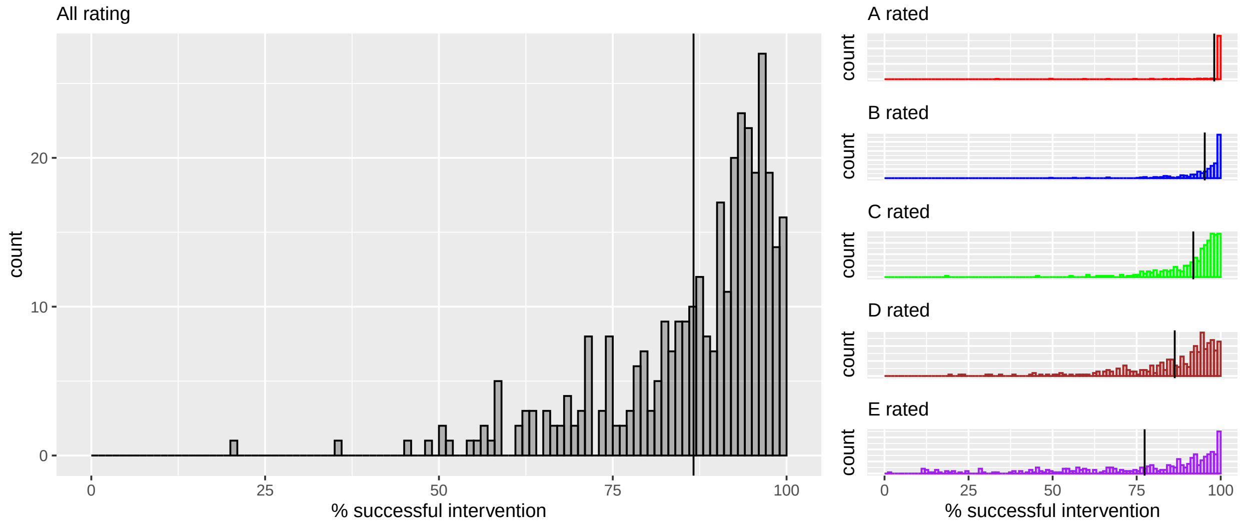
Illustrate distribution of % successful intervention for all rating and for each

```r
# Histogram of all rated establishment
plot_int_all <- ggplot(data.food.clean, aes(int_success_all)) +
  geom_histogram(binwidth = 1, color = "black", fill = "black", alpha = 0.25, breaks = seq(0,100, by =
  geom_vline(data = data.food.pivot.all, mapping = aes(xintercept = avg_int_success_all)) +
  labs(x="% successful intervention", subtitle = "All rating")

# Histogram of A rated
plot_int_a <- ggplot(data.food.clean, aes(int_success_a)) +
  geom_histogram(binwidth = 1, color = "red", fill = "red", alpha = 0.25, breaks = seq(0,100, by = 1)) +
  geom_vline(data = data.food.pivot.rate, mapping = aes(xintercept = avg_int_success_a)) +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),axis.text.y=element_blank(), axis.tic
  labs(x = NULL, subtitle = "A rated")

# Histogram of B rated
plot_int_b <- ggplot(data.food.clean, aes(int_success_b)) +
  geom_histogram(binwidth = 1, color = "blue", fill = "blue", alpha = 0.25, breaks = seq(0,100, by = 1))
  geom_vline(data = data.food.pivot.rate, mapping = aes(xintercept = avg_int_success_b)) +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),axis.text.y=element_blank(), axis.tic
  labs(x = NULL, subtitle = "B rated")

# Histogram of C rated
plot_int_c <- ggplot(data.food.clean, aes(int_success_c)) +
  geom_histogram(binwidth = 1, color = "green", fill = "green", alpha = 0.25, breaks = seq(0,100, by =
  geom_vline(data = data.food.pivot.rate, mapping = aes(xintercept = avg_int_success_c)) +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),axis.text.y=element_blank(), axis.tic
  labs(x = NULL, subtitle = "C rated")

# Histogram of D rated
plot_int_d <- ggplot(data.food.clean, aes(int_success_d)) +
  geom_histogram(binwidth = 1, color = "brown", fill = "brown", alpha = 0.25, breaks = seq(0,100, by =
  geom_vline(data = data.food.pivot.rate, mapping = aes(xintercept = avg_int_success_d)) +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),axis.text.y=element_blank(), axis.tic
  labs(x = NULL,  subtitle = "D rated")

# Histogram of E rated
plot_int_e <- ggplot(data.food.clean, aes(int_success_e)) +
  geom_histogram(binwidth = 1, color = "purple", fill = "purple", alpha = 0.25, breaks = seq(0,100, by
  geom_vline(data = data.food.pivot.rate, mapping = aes(xintercept = avg_int_success_e)) +
  theme(axis.text.y=element_blank(), axis.ticks.y=element_blank()) +
  labs(x="% successful intervention",  subtitle = "E rated")

# Combined charts
grid.arrange(plot_int_all,
             arrangeGrob(plot_int_a, plot_int_b, plot_int_c, plot_int_d, plot_int_e, ncol = 1, heights
             ncol = 2, widths = c(2,1),
             top = textGrob("Figure 1: Distribution of % successful intervention for each rated establi
```

```
## Warning: Removed 30 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 6 rows containing non-finite values (`stat_bin()`).
## Removed 6 rows containing non-finite values (`stat_bin()`).
## Removed 6 rows containing non-finite values (`stat_bin()`).
## Removed 6 rows containing non-finite values (`stat_bin()`).
## Removed 6 rows containing non-finite values (`stat_bin()`).
```

Figure 1: Distribution of % successful intervention for each rated establishment

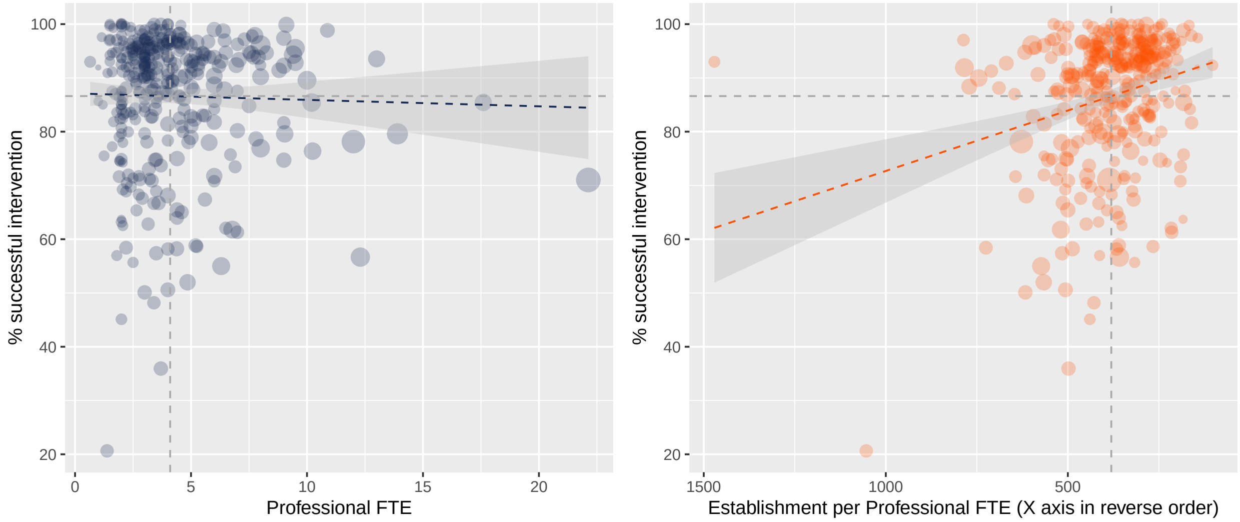


Examine the relationship between no. of employee and % successful intervention

```
# Plot scatter plot to see distribution trend of the data
grid.arrange(
ggplot(data.food.clean, aes(x= professional_fte, int_success_all, size = total_est_rated)) +
  geom_point(alpha = 0.25, color = "#152852") +
  geom_smooth(method=lm, color = "#152852", fill = "grey", linetype = 2, size = 0.5) +
  geom_vline(data = data.food.pivot.all, aes(xintercept = avg_fte), color = "dark grey", fill = "grey",
  geom_hline(data = data.food.pivot.all, aes(yintercept = avg_int_success_all), color = "dark grey", fil
  theme(legend.position = "none") +
  labs(x = "Professional FTE", y = "% successful intervention"),
ggplot(data.food.clean, aes(x= est_per_fte, int_success_all, size = total_est_rated)) +
  geom_point(alpha = 0.25, color = "#fe5000") +
  geom_smooth(method=lm, color = "#fe5000", fill = "grey", linetype = 2, size = 0.5) +
  geom_vline(data = data.food.pivot.all, aes(xintercept = avg_est_per_fte), color = "dark grey", fill =
  geom_hline(data = data.food.pivot.all, aes(yintercept = avg_int_success_all), color = "dark grey", fil
  scale_x_reverse() +
  theme(legend.position = "none")+
  labs(x = "Establishment per Professional FTE (X axis in reverse order)", y = "% successful interventi
, ncol = 2, top = textGrob("Figure 2: Relationship between No. of employee and % successful interventio
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.

## Warning in geom_vline(data = data.food.pivot.all, aes(xintercept = avg_fte), : Ignoring unknown
## parameters: `fill`

## Warning in geom_hline(data = data.food.pivot.all, aes(yintercept = avg_int_success_all), : Ignoring
## unknown parameters: `fill`

## Warning in geom_vline(data = data.food.pivot.all, aes(xintercept = avg_est_per_fte), : Ignoring
## unknown parameters: `fill`

## Warning in geom_hline(data = data.food.pivot.all, aes(yintercept = avg_int_success_all), : Ignoring
## unknown parameters: `fill`

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

## Warning: Removed 6 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 6 rows containing missing values (`geom_point()`).

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

## Warning: Removed 6 rows containing non-finite values (`stat_smooth()`).
## Removed 6 rows containing missing values (`geom_point()`).
```

Figure 2: Relationship between No. of employee and % successful intervention



Regression analysis

```
# Create regression model for both FTE and est per FTE against % successful intervention
m.int.fte <- lm(int_success_all~professional_fte, data = data.food.clean)
m.int.fte.per <- lm(int_success_all~est_per_fte, data = data.food.clean)

# Summary of the regression result
summary (m.int.fte) # No. of employee does not have any significant relationship with the % successful
```

```
##
## Call:
## lm(formula = int_success_all ~ professional_fte, data = data.food.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.304  -4.575   4.067   8.658  13.860
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       87.1091     1.2828  67.905   <2e-16 ***
## professional_fte  -0.1195     0.2675  -0.447    0.655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.4 on 345 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.0005787,  Adjusted R-squared:  -0.002318
## F-statistic: 0.1998 on 1 and 345 DF,  p-value: 0.6552
```

```
summary (m.int.fte.per)  # No. of employee as a proportion of establishment has significant relationshi
```

```
##
## Call:
## lm(formula = int_success_all ~ est_per_fte, data = data.food.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.850  -5.651   4.398   8.006  30.878
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 95.176625   1.909355  49.848  < 2e-16 ***
## est_per_fte -0.022481   0.004721  -4.762 2.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.01 on 345 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.06167,    Adjusted R-squared:  0.05895
## F-statistic: 22.67 on 1 and 345 DF,  p-value: 2.83e-06
```

```
# Note: While statistically significant, adjusted R-squared of "m.int.fte.per" is considerably low at 5
# Implying that establishment per employee only explain 5.9% of the variation in % successful response,


# Estimation approach of regression result
cbind(coefficient=coef(m.int.fte), confint(m.int.fte))
```

```
##                  coefficient      2.5 %     97.5 %
## (Intercept)       87.1091495 84.5860343 89.6322647
## professional_fte  -0.1195469 -0.6456029  0.4065092
```

```
cbind(coefficient=coef(m.int.fte.per), confint(m.int.fte.per))
```

```
##             coefficient       2.5 %      97.5 %
## (Intercept) 95.17662531 91.42118360 98.93206702
## est_per_fte -0.02248149 -0.03176758 -0.01319539
```

Section 2

Data preparation

This report presents the results of the analyses requested by panel of Food Standards Agency. The data use in this report contains food hygiene information and No. of employee for each **353 Local Authorities (“LA”)** across England, Wales and Northern Ireland. The analysis in this report will focus primary on % of successful intervention and No. of employee aspect of the data.

The data we are covering in this report consist of;

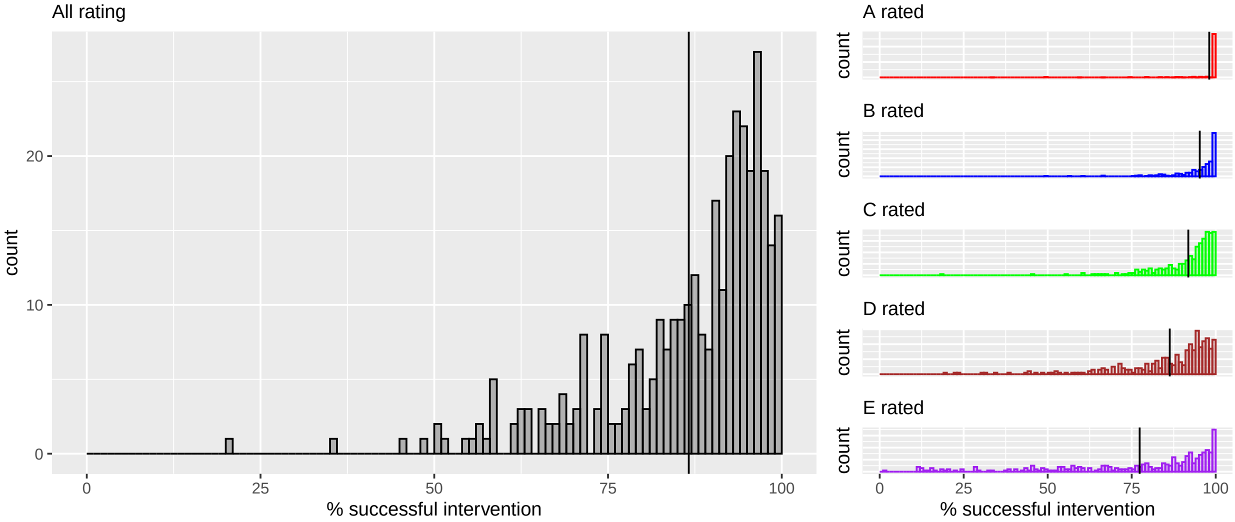
- Rating of establishment: Representing the degree of potential impact if bad foods are served rating from A being the most to E being the least.
- % of successful intervention: Representing % of establishment successfully back to operation after being investigated by LA for failure to meet the “Broadly Compliant” standard.
- No. of employee: Representing the Full Time Equivalent (FTE) employee of each LA.

The data set also contains 3 types of establishment, rated, not yet rated and outside the program. For the purpose of this study, only rated establishments were use in the analysis since the data is more robust and covering over 90% of total establishment.

Data analysis

We begin with summaries of Distribution of % successful intervention for each rated establishment (Figure 1).

Figure 1: Distribution of % successful intervention for each rated establishment

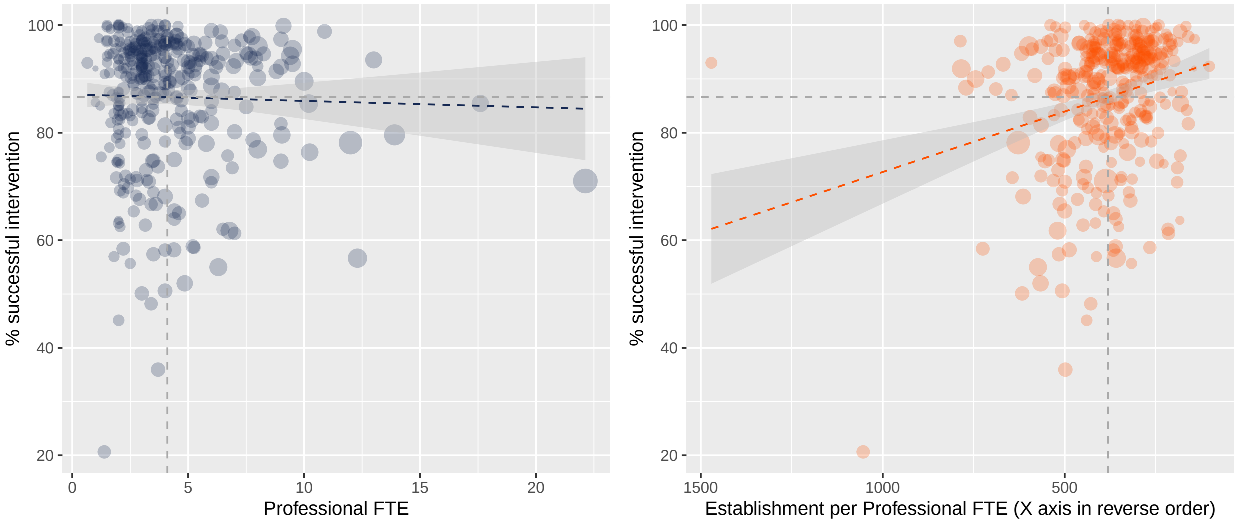


- Figure 1 illustrate the distribution of % successful intervention for all and each establishment rating. On the aggregate level (LHS chart), the average of % successful intervention is at 86.6% across all rating. The data is left-skewed in which majority of LAs are performing above average while there is a long-tail of LAs with much poorer performance.

- On the other hand, the distribution of % successful intervention is not uniform across establishment rating (RHS charts). The result illustrates that LAs are more successful in reacting to higher potential impact establishment than lower potential impact. This highlights the ability of food safety employees in dealing with high-priority targets.

As per the panel's request, below charts illustrate the analysis of relationship between No. of FTE food safety employees and % successful intervention.

Figure 2: Relationship between No. of employee and % successful intervention



- Figure 2 compares the relationship of No. of employee vs. % successful intervention and No. of establishment per employee vs. % successful intervention.

- In order to factored in the proportion between No. of employee and No. of establishment in each LA, *Establishment per Professional* FTE was created to represent No. of establishment handle by each employee in which high establishment count per employee implying that there is relatively fewer employee in that LA and vice versa.
- In term of absolute No. of employee, linear regression analysis shows that the increase in No. of employee **does not** have a significant effect on % successful intervention ($t(345) = -0.447, p = 0.6552$)
- On the other hand, when factoring in the proportion between No. of employee and No. of establishment in each LA, linear regression analysis shows that the increase in No. of employee **does** have a significant effect on % successful intervention. For every one less establishment handled by the employee, which imply more FTE professional in proportion to establishment, the % successful intervention increased by 0.02% 95% CI [0.03% - 0.01%]. This effect is illustrated by the best fitted line plotted in the RHS chart.

In conclusion, our analysis shows that No. of employee does have a slightly positive effect on % successful intervention, but only in the relative term with No. of establishment of each LAs. However, the analysis also come with a significant caveat that No. of employee in relation to establishment can only explain 5.9% of the variation in % successful intervention. This can be seen in the RHS chart that many data points are scattered further away from the best fitted line.

As a result, this factor of interest may not necessary be the dial mover to influence higher successful percentage. Hence, we recommend that the panel further investigate into other factors which may help positively effect % successful intervention.

Question 2

Section 1

Importing and cleaning data

```
data.book <- read_csv("publisher_sales.csv")
```

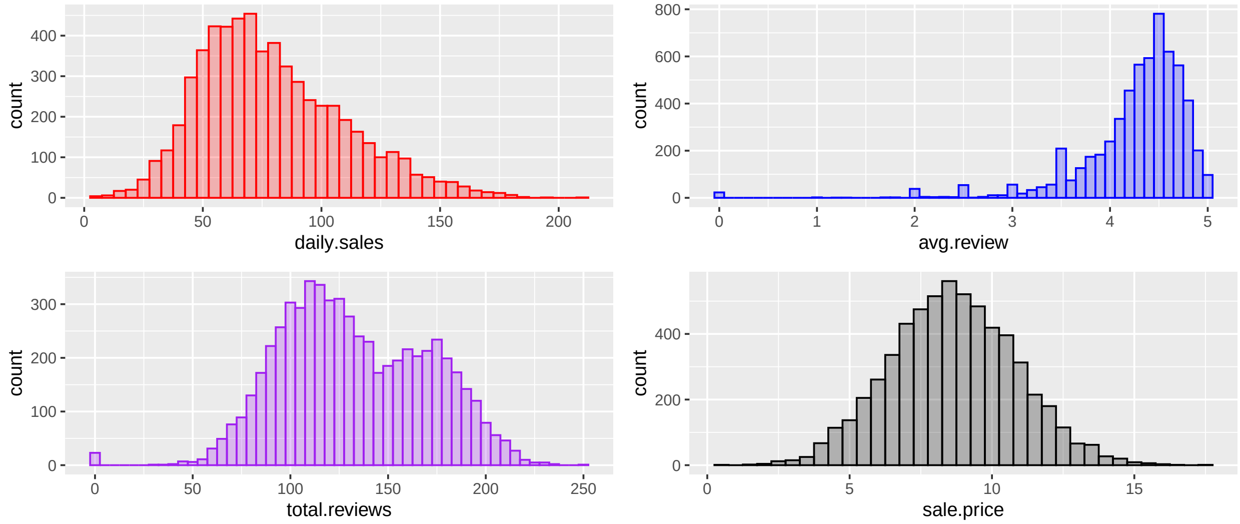
```
## Rows: 6000 Columns: 7
## -- Column specification ----------------------------------------------------------------------
## Delimiter: ","
## chr (3): sold by, publisher.type, genre
## dbl (4): avg.review, daily.sales, total.reviews, sale.price
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# Clean daily sales that below 0 and convert categorical to factor
data.book <- data.book %>%
  filter(daily.sales > 0) %>%
  mutate(`sold by` = as.factor(`sold by`),
         publisher.type = as.factor(publisher.type),
         genre = as.factor(genre)) %>%
  rename(sold_by = `sold by`)

# Pivot and summarise average
```

```
summary.data <- data.book %>%
  group_by( genre) %>%
  summarise(avg_review = mean(avg.review),
            avg_daily_sales = mean(daily.sales),
            avg_total_reviews = mean(total.reviews),
            avg_sale_price = mean(sale.price))
```

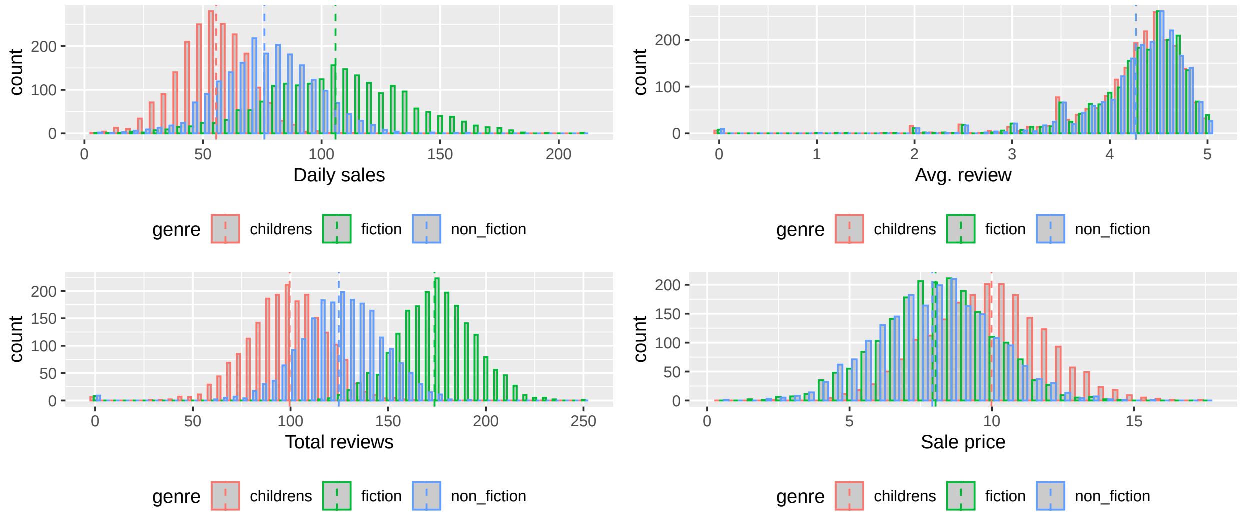
```
# Distribution of each attribute
grid.arrange(
ggplot(data.book, aes(daily.sales)) +
  geom_histogram(binwidth = 5, color = "red", fill = "red", alpha = 0.25),
ggplot(data.book, aes(avg.review)) +
  geom_histogram(binwidth = 0.1, color = "blue", fill = "blue", alpha = 0.25),
ggplot(data.book, aes(total.reviews)) +
  geom_histogram(binwidth = 5, color = "purple", fill = "purple", alpha = 0.25),
ggplot(data.book, aes(sale.price)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "black", alpha = 0.25),
ncol = 2, nrow =2, top = textGrob("Figure 1: Distribution of Daily sales, Avg. review, Total reviews an
```

Figure 1: Distribution of Daily sales, Avg. review, Total reviews and Sale price



```
grid.arrange(
# Distribution of daily sales for each genre
ggplot(data.book, aes(daily.sales, color = genre)) +
  geom_histogram(binwidth = 5, alpha = 0.25, position = "dodge") +
  geom_vline(data = summary.data, aes(xintercept = avg_daily_sales,color=genre), linetype = "dashed")+
  labs(x = "Daily sales") +
  theme(legend.position = "bottom"),

# Distribution of average review for each genre
ggplot(data.book, aes(avg.review, color = genre)) +
  geom_histogram(binwidth = 0.1, alpha = 0.25, position = "dodge") +
  geom_vline(data = summary.data, aes(xintercept = avg_review,color=genre), linetype = "dashed")+
  labs(x = "Avg. review") +
  theme(legend.position = "bottom"),
# Distribution of total reviews for each genre
ggplot(data.book, aes(total.reviews, color = genre)) +
  geom_histogram(binwidth = 5, alpha = 0.25, position = "dodge") +
  geom_vline(data = summary.data, aes(xintercept = avg_total_reviews,color=genre), linetype = "dashed")
  labs(x = "Total reviews") +
  theme(legend.position = "bottom"),

# Distribution of sale price for each genre
ggplot(data.book, aes(sale.price, color = genre)) +
  geom_histogram(binwidth = 0.5, alpha = 0.25, position = "dodge") +
  geom_vline(data = summary.data, aes(xintercept = avg_sale_price,color=genre), linetype = "dashed")+
  labs(x = "Sale price") +
  theme(legend.position = "bottom"),
ncol = 2, nrow = 2, top = textGrob("Figure 2: Distribution of Daily sales, Avg. review, Total reviews a
```

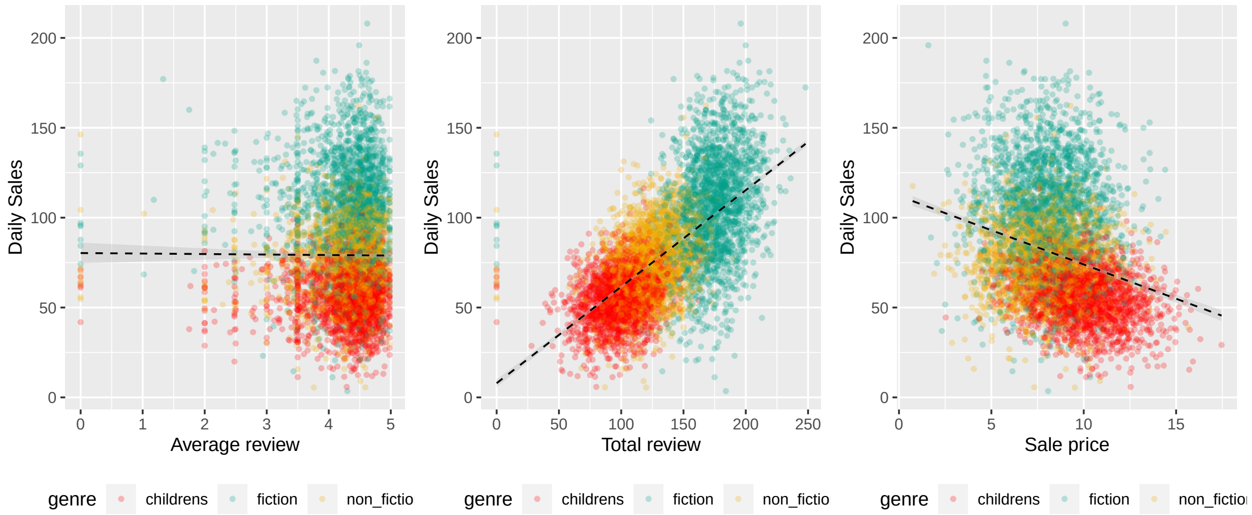
Figure 2: Distribution of Daily sales, Avg. review, Total reviews and Sale price by genre



```
# Scatter plots show relationship between each pair of attributes

grid.arrange(
  ggplot(data.book, aes(x= avg.review, y = daily.sales, color = genre)) +
    geom_point(alpha = 0.25, size = 1) +
    geom_smooth(method=lm, color = "black", fill = "grey", linetype = 2, size = 0.5) +
    scale_color_manual(values= wes_palette(n=3, name="Darjeeling1")) +
    labs(x = "Average review", y = "Daily Sales", title = )+
    theme(legend.position = "bottom"),
  ggplot(data.book, aes(x= total.reviews, y = daily.sales, color = genre)) +
    geom_point(alpha = 0.25, size = 1) +
    geom_smooth(method=lm, color = "black", fill = "grey", linetype = 2, size = 0.5) +
    scale_color_manual(values= wes_palette(n=3, name="Darjeeling1")) +
    labs(x = "Total review", y = "Daily Sales")+
    theme(legend.position = "bottom"),
  ggplot(data.book, aes(x= sale.price, y = daily.sales, color = genre)) +
    geom_point(alpha = 0.25, size = 1) +
    geom_smooth(method=lm, color = "black", fill = "grey", linetype = 2, size = 0.5) +
    scale_color_manual(values= wes_palette(n=3, name="Darjeeling1")) +
    labs(x = "Sale price", y = "Daily Sales")+
    theme(legend.position = "bottom"),
ncol = 3, top = textGrob("Figure 3: Relationship between Daily sales and Avg.review/Total reviews/Sale
```

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

Figure 3: Relationship between Daily sales and Avg.review/Total reviews/Sale price



Do books from different genres have different daily sales on average?

```
# Use emmeans to find estimated mean difference
m.book.genre <- lm(daily.sales~genre, data = data.book)
(m.book.genre.emm <- emmeans(m.book.genre, ~genre))
```

```
##  genre       emmean    SE   df lower.CL upper.CL
##  childrens     55.6 0.497 5996     54.6     56.6
##  fiction      105.9 0.497 5996    104.9    106.9
##  non_fiction   75.9 0.497 5996     74.9     76.9
##
## Confidence level used: 0.95
```

```
(m.book.genre.con <- confint(pairs(m.book.genre.emm)))
```

```
##  contrast                estimate    SE   df lower.CL upper.CL
##  childrens - fiction        -50.3 0.703 5996    -52.0    -48.7
##  childrens - non_fiction    -20.3 0.703 5996    -22.0    -18.7
##  fiction - non_fiction       30.0 0.703 5996     28.3     31.6
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

Conclusion:

- Emmean results show that 95% CI of mean for all genre has no 0 included. Hence, means are statistically different between each genre.

Do books have more/fewer sales depending upon their average review scores and total number of reviews

```
# Regression analysis of relationship between daily sales and total reviews/sale price
m.book.rev <- lm(daily.sales~total.reviews+avg.review, data = data.book)
summary(m.book.rev) # Result shows that total reviews and average review have significant relationship
```

```
##
## Call:
## lm(formula = daily.sales ~ total.reviews + avg.review, data = data.book)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.407  -14.656   -1.071   13.672  122.177
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.123430   2.340120  10.309  < 2e-16 ***
## total.reviews  0.543327   0.007816  69.517  < 2e-16 ***
## avg.review    -3.999637   0.512874  -7.798 7.34e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.58 on 5996 degrees of freedom
## Multiple R-squared:  0.4463, Adjusted R-squared:  0.4461
## F-statistic:  2416 on 2 and 5996 DF,  p-value: < 2.2e-16
```

```
vif(m.book.rev) # vif shows no multicollinearity between both variables
```

```
## total.reviews    avg.review
##      1.011022      1.011022
```

```
# We suspect that total review and average review might have interaction.
m.book.rev.int <- lm(daily.sales~total.reviews*avg.review, data = data.book)
summary(m.book.rev.int) # Result show significant interaction between total review and average review
```

```
##
## Call:
## lm(formula = daily.sales ~ total.reviews * avg.review, data = data.book)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.084  -14.641   -0.946   13.822   92.351
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               63.676679   4.174384  15.254  < 2e-16 ***
## total.reviews              0.165853   0.034038   4.873 1.13e-06 ***
## avg.review               -13.709806   0.992277 -13.817  < 2e-16 ***
## total.reviews:avg.review   0.091422   0.008028  11.388  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22.34 on 5995 degrees of freedom
## Multiple R-squared:  0.458,  Adjusted R-squared:  0.4577
## F-statistic:  1689 on 3 and 5995 DF,  p-value: < 2.2e-16
```

```
anova(m.book.rev, m.book.rev.int)  # ANOVA shows that model with interaction term has better fit
```

```
## Analysis of Variance Table
##
## Model 1: daily.sales ~ total.reviews + avg.review
## Model 2: daily.sales ~ total.reviews * avg.review
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   5996 3057720
## 2   5995 2992979  1     64740 129.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# Estimation approach of regression result
cbind(coefficient=coef(m.book.rev.int), confint(m.book.rev.int))
```

```
##                          coefficient        2.5 %      97.5 %
## (Intercept)              63.67667890  55.49338478  71.8599730
## total.reviews             0.16585332   0.09912647   0.2325802
## avg.review              -13.70980620 -15.65502518 -11.7645872
## total.reviews:avg.review   0.09142169   0.07568349   0.1071599
```

Conclusion:

- There is a significant effect of Avg. review with a decrease in Daily sales of 13.71 for every 1 increase in review rating (95% CI [-15.66, -11.76])
- On the other hand, there is also a significant effect of Total reviews with an increase in Daily sales of 0.17 for every 1 increase in total review (95% CI [0.10, 0.23])
- There is also a significant positive interaction between Total reviews and Avg. review of 0.09 (95% CI [0.08, 0.11])
- A model comparison shows that the model fit is significantly improved by the inclusion of the interaction term $(F(1, 5995) = 129.68, p < 0.001)$

What is the effect of sale price upon the number of sales, and is this different across genres?

```
# Regression analysis of relationship between daily sales and sale price
m.book.price <- lm(daily.sales~sale.price, data = data.book)
summary(m.book.price) # The model shows that sales price have negative relationship with daily sales bu
```

```
##
## Call:
```

```
## lm(formula = daily.sales ~ sale.price, data = data.book)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.760 -20.644  -4.638  17.084 130.301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 112.0540     1.5201   73.72   <2e-16 ***
## sale.price   -3.8110     0.1704  -22.36   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 29.15 on 5997 degrees of freedom
## Multiple R-squared:  0.07696,    Adjusted R-squared:  0.0768
## F-statistic:   500 on 1 and 5997 DF,  p-value: < 2.2e-16
```

```
# Improve model by adding genre as a variable
m.book.price.mult <- lm(daily.sales~sale.price+genre, data = data.book)
summary(m.book.price.mult) # Result shows that both sales price and genre have significant relationship
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price + genre, data = data.book)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.357  -13.329    0.016   13.090  102.916
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       63.8120     1.5184  42.026  < 2e-16 ***
## sale.price        -0.8243     0.1437  -5.737 1.01e-08 ***
## genrefiction      48.6873     0.7556  64.434  < 2e-16 ***
## genrenon_fiction  18.6127     0.7619  24.430  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22.16 on 5995 degrees of freedom
## Multiple R-squared:  0.4669, Adjusted R-squared:  0.4666
## F-statistic:  1750 on 3 and 5995 DF,  p-value: < 2.2e-16
```

```
anova(m.book.price,m.book.price.mult) # ANOVA shows that model with interaction term has better fit
```

```
## Analysis of Variance Table
##
## Model 1: daily.sales ~ sale.price
## Model 2: daily.sales ~ sale.price + genre
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   5997 5097347
## 2   5995 2944127  2   2153220 2192.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
vif(m.book.price.mult) # vif shows no multicollinearity between both variables
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## sale.price 1.229899  1        1.109008
## genre      1.229899  2        1.053095
```

```
# While previous model improved R2, we suspect that sales price and genre might have interaction.
m.book.price.mult.int <- lm(daily.sales~sale.price*genre, data = data.book)
summary(m.book.price.mult.int) # Result show significant interaction between sale price and each of gen
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price * genre, data = data.book)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.383  -13.374    0.018   13.042  102.366
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  72.8781     2.5003  29.147  < 2e-16 ***
## sale.price                   -1.7319     0.2453  -7.059 1.87e-12 ***
## genrefiction                 35.1993     3.2711  10.761  < 2e-16 ***
## genrenon_fiction              6.3974     3.2015   1.998 0.045736 *
## sale.price:genrefiction       1.4587     0.3543   4.118 3.88e-05 ***
## sale.price:genrenon_fiction   1.3057     0.3467   3.766 0.000167 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22.13 on 5993 degrees of freedom
## Multiple R-squared:  0.4687, Adjusted R-squared:  0.4683
## F-statistic:  1057 on 5 and 5993 DF,  p-value: < 2.2e-16
```

```
anova(m.book.price.mult,m.book.price.mult.int) # ANOVA shows that model with interaction term has bette
```

```
## Analysis of Variance Table
##
## Model 1: daily.sales ~ sale.price + genre
## Model 2: daily.sales ~ sale.price * genre
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   5995 2944127
## 2   5993 2933858  2     10270 10.489 2.836e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# Estimation approach of regression result
cbind(coefficient=coef(m.book.price.mult.int), confint(m.book.price.mult.int))
```

```
##                             coefficient      2.5 %    97.5 %
## (Intercept)                   72.878117 67.9765566 77.779678
## sale.price                    -1.731864 -2.2128230 -1.250906
## genrefiction                    35.199273 28.7867330 41.611813
## genrenon_fiction                 6.397351  0.1212637 12.673439
## sale.price:genrefiction          1.458709  0.7642279  2.153190
## sale.price:genrenon_fiction      1.305662  0.6260930  1.985232
```

```
# Create plot to show interaction term between Sale price and Genre
dailysales.price.genre <- data.book %>%
  mutate(predict.book.price.mult.int = predict(m.book.price.mult.int))

ggplot(dailysales.price.genre ) +
  geom_line(aes(sale.price,predict.book.price.mult.int, colour = genre)) +
  labs(y = "Daily Sales", x = "Sale price", title = "Figure 4: Regression result between Daily sales an
  geom_point(data = data.book, aes(x = sale.price, y = daily.sales, color = genre), alpha = 0.15, size
```

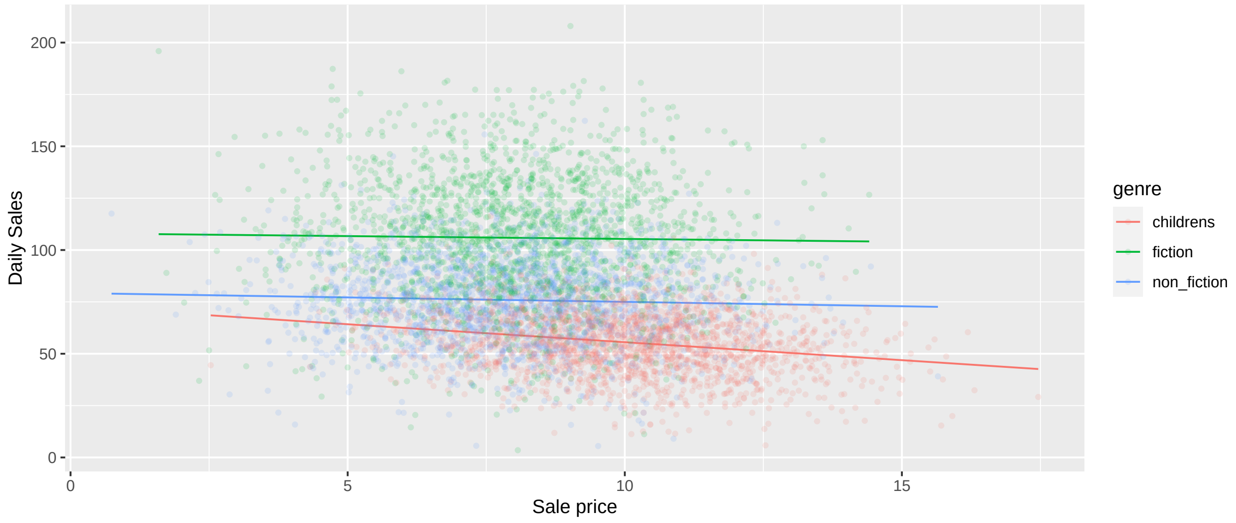


Figure 4: Regression result between Daily sales and Sale price by Genre

Conclusion:

- There is a significant effect of Sale price with a decrease in Daily sales of 1.73 for every 1 increase in Sale price (95% CI [-2.21, -1.25])
- On the other hand, there is also a significant effect of having different genre with
- An increase in Daily sales of 35.20 when genre is fiction comparing to children (95% CI [28.9, 47.61])
- An increase in Daily sales of 6.40 when genre is non-fiction comparing to children (95% CI [0.12, 12.61])
- There is also a significant positive interaction between Sale price and genre of 1.46 for fiction (95% CI [0.08, 0.11]) and 1.31 (95% CI [0.63,1.99])
- A model comparison shows that the model fit is significantly improved by the inclusion of the interaction term between Sale price and each genre $(F(1, 5993) = 10.49, p < 0.001)$

Section 2

Data preparation

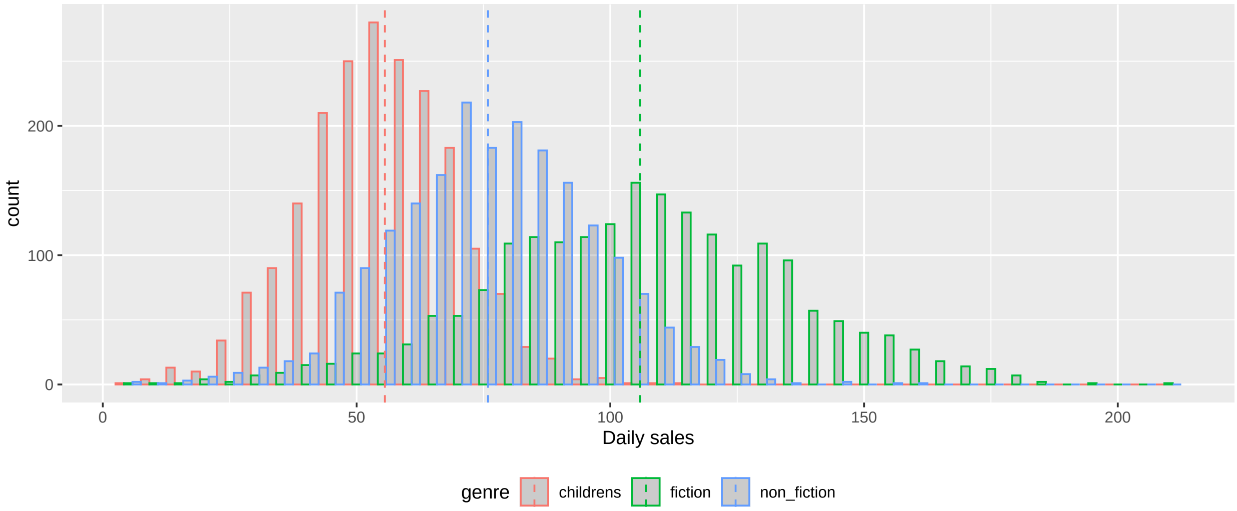
This report presents the results of the analyses requested by managers of publishing company. The data use in this report contains sales and reviews for each **6,000 e-books** across a certain time period. The analysis in this report will focus primary on the relationship between Daily sales and other variables.

Since the Daily sales cannot be negative number, 1 e-book has been removed from the data which result in **5,999 e-books** left for analysis.

Data analysis

We begin with summaries of Distribution and mean of daily sales by Genre (Figure 1).

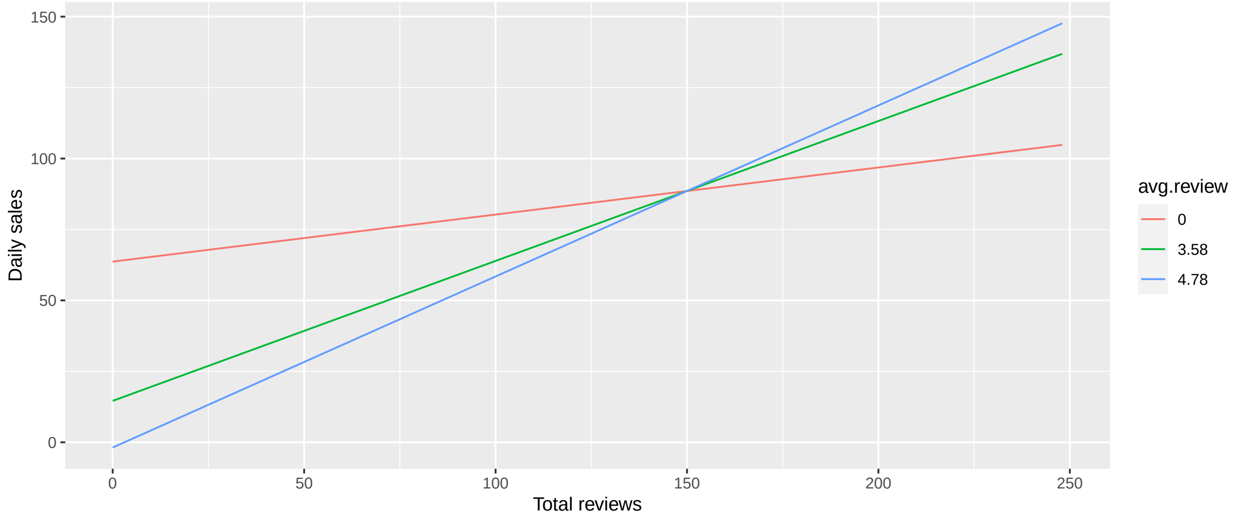
Figure 1: Distribution and mean of daily sales by Genre



- Figure 1 illustrate the distribution and mean of daily sales for each genre. Bars are representing distribution of daily sales while the dotted line illustrated average daily sales for books in each genre.
- As illustrated above, average daily sales of each genre are statistically different, in which fiction has the highest average daily sales follow by non-fiction and children, respectively.

As per the panel's request, below charts illustrate the analysis of relationship between daily sales, total reviews, and average review.

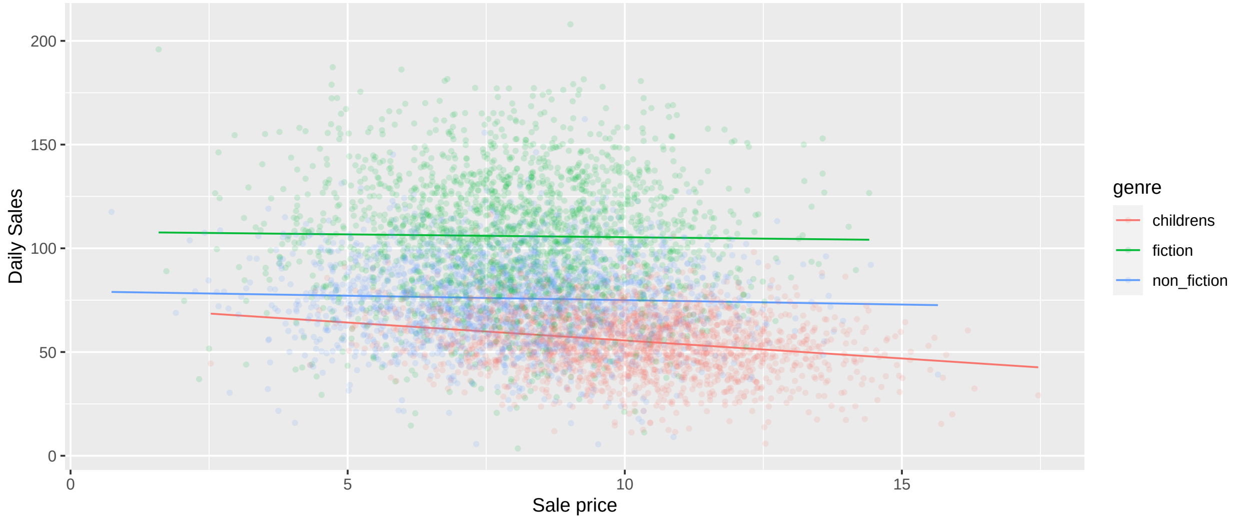
Figure 2: Relationship between Daily sales, Total reviews and Avg. reviews



- Figure 2 illustrate the relationship between daily sales, average review, and total review.
- The red line represent relationship between daily sales and total reviews when avg. review is 0 while green and blue line represent the same relationship when avg. review is 3.58 (Top 90%) and 4.78 (Top 10%), respectively.
- The slope of 3 lines indicates that, overall, total reviews have a positive effect on daily sales. However, with higher avg. review score, the effect of total review on daily sales has been amplified. Hence, the analysis shows that e-books which have combination of high total reviews and avg. review tends to sell better comparing to those who did not.

Lastly, the chart below illustrates the relationship and effect of sale price on daily sales in respect to each genre.

Figure 3: Regression result between Daily sales and Sale price by Genre



- Figure 3 illustrate the relationship between daily sales and sale price for each genre, in which each colour representing each different genre.

- Dots are representing Sale price and Daily sales of each e-book while lines are representing the relationship between 2 variables.
- Overall, we can see that the slope of all line show that sale price has a negative relationship with daily sales, meaning that higher price causing lower sales and vice versa.
- Furthermore, if we look at each line separately, we can see that red line has the steeper slope comparing to green and blue, which imply that sales of e-book in children genre are more sensitive to price changes comparing to the other two genres.
- Hence, the analysis shows that discounting sale price may help improve sales of e-books while the effect is more prominent in children genre.

In conclusion, the analysis shows that;

- Each genre of e-book has different daily sales, with fiction being the best performing and children the worst.
- E-books which have combination of high total reviews and avg. review tends to sell better.
- E-books sales are sensitive to sale price. By discounting the sale price, we can expect higher daily sales, especially e-books in children genre.

Nevertheless, while regression analysis illustrates statistically significant relationship between these variables, it is worth mentioning that this does not imply causation between each of these variables. The panel should further conduct a real-world test in order to understand the cause-effect relationship of the above-mentioned results.