M05_activity

Hannah Valenty

2024-07-19

Task 1

```
library(dplyr)
```

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

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

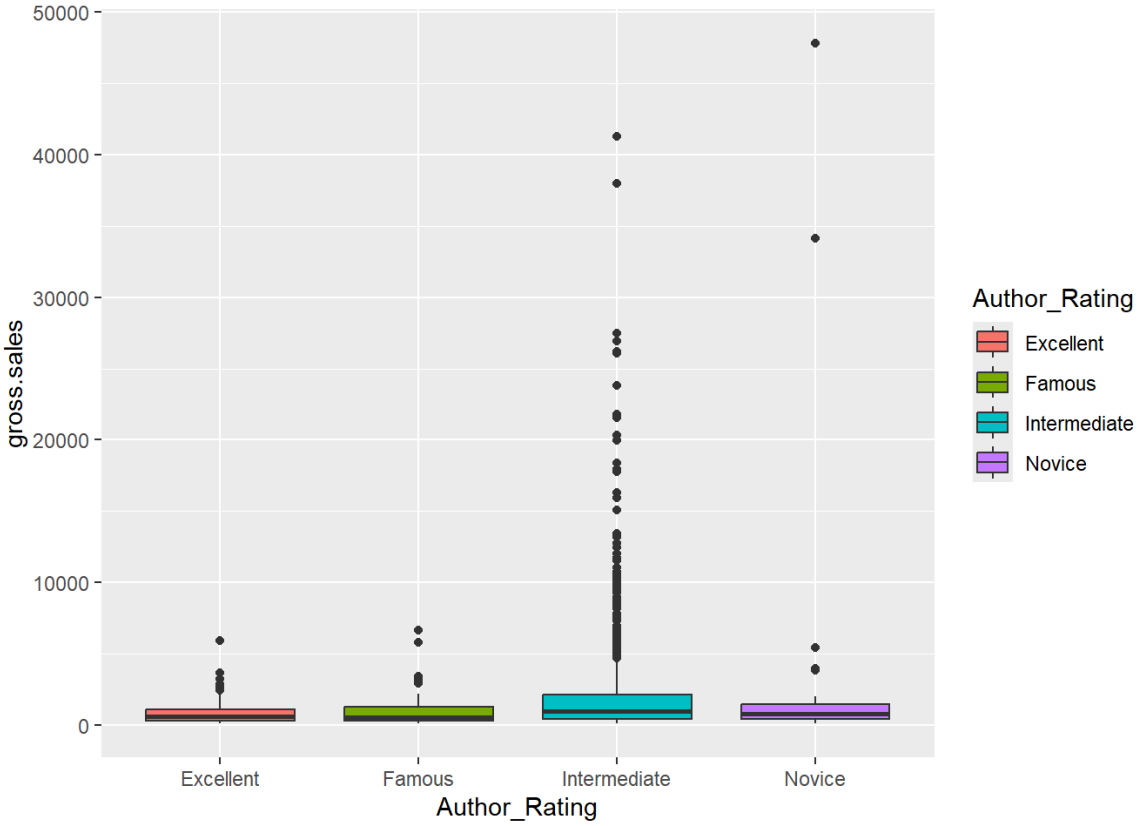
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
df <- read.csv('Books_Data_Clean.csv')
head(df)
```

```
##   index Publishing.Year                      Book.Name
## 1     0            1975                       Beowulf
## 2     1            1987                Batman: Year One
## 3     2            2015               Go Set a Watchman
## 4     3            2008 When You Are Engulfed in Flames
## 5     4            2011        Daughter of Smoke & Bone
## 6     5            2015                      Red Queen
##                                                        Author language_code
## 1                                        Unknown, Seamus Heaney         en-US
## 2 Frank Miller, David Mazzucchelli, Richmond Lewis, Dennis O'Neil           eng
## 3                                                   Harper Lee           eng
## 4                                                 David Sedaris         en-US
## 5                                                  Laini Taylor           eng
## 6                                               Victoria Aveyard           eng
##   Author_Rating Book_average_rating Book_ratings_count         genre
## 1        Novice                3.42             155903 genre fiction
## 2  Intermediate                4.23             145267 genre fiction
## 3        Novice                3.31             138669 genre fiction
## 4  Intermediate                4.04             150898       fiction
## 5  Intermediate                4.04             198283 genre fiction
## 6  Intermediate                4.08              83354 genre fiction
##   gross.sales publisher.revenue sale.price sales.rank
## 1     34160.0           20496.0       4.88          1
## 2     12437.5            7462.5       1.99          2
## 3     47795.0           28677.0       8.69          3
## 4     41250.0           24750.0       7.50          3
## 5     37952.5           22771.5       7.99          4
## 6     19960.0               0.0       4.99          5
##                         Publisher units.sold
## 1       HarperCollins Publishers       7000
## 2       HarperCollins Publishers       6250
## 3 Amazon Digital Services,  Inc.       5500
## 4            Hachette Book Group       5500
## 5        Penguin Group (USA) LLC       4750
## 6 Amazon Digital Services,  Inc.       4000
```

Visualize data

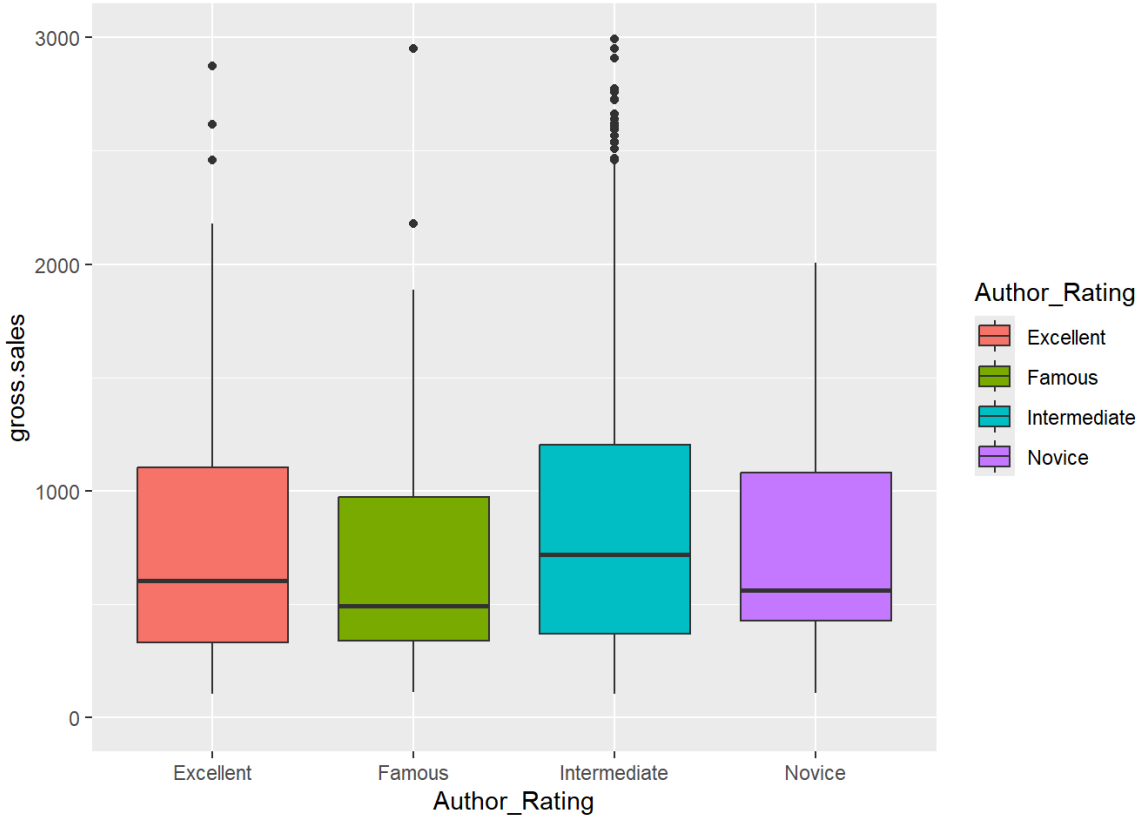
```
ggplot(df, aes(x=Author_Rating, y=gross.sales, fill = Author_Rating))+
  geom_boxplot()
```



Zoom In

```
ggplot(df, aes(x=Author_Rating, y=gross.sales, fill = Author_Rating))+
  geom_boxplot()+
  ylim(0, 3000)
```

```
## Warning: Removed 134 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



ANOVA Test

```
anova <- aov(gross.sales~Author_Rating, data=df)
summary(anova)
```

```
##                 Df    Sum Sq   Mean Sq F value   Pr(>F)
## Author_Rating    3 7.881e+08 262704446   17.75 3.03e-11 ***
## Residuals     1066 1.578e+10  14803673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

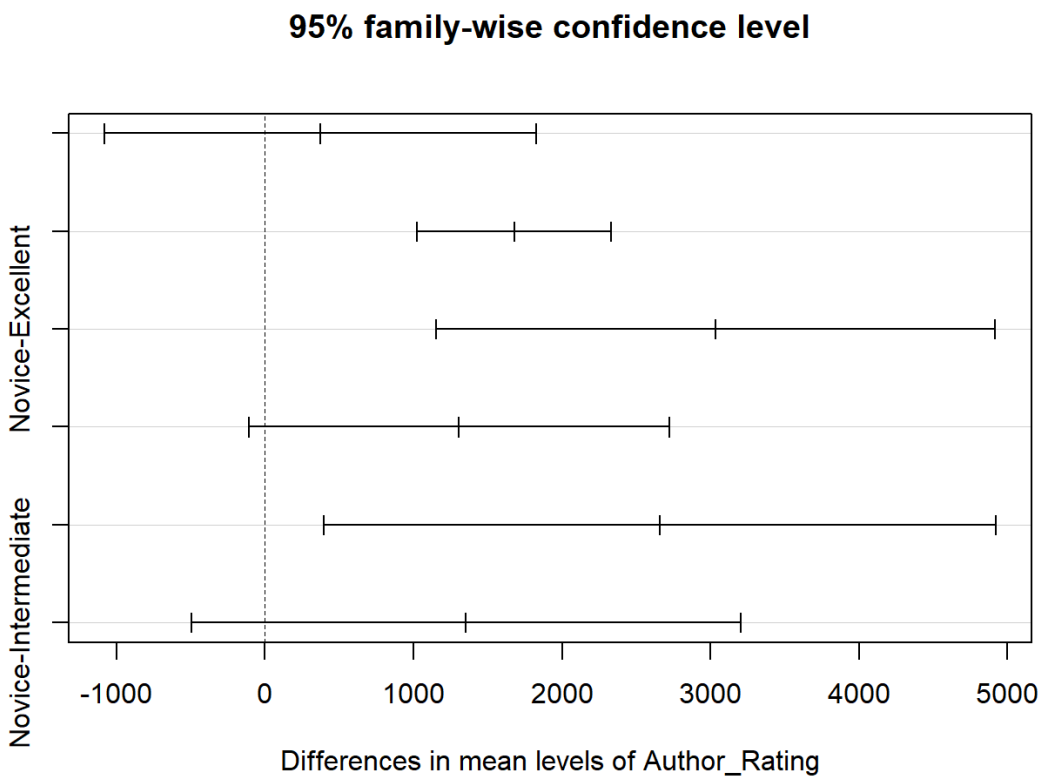
p-value = 3.03e-11 (very close to zero), therefore reject the null hypothesis.

Hypotheses Tests

```
TukeyHSD(anova, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = gross.sales ~ Author_Rating, data = df)
##
## $Author_Rating
##                              diff        lwr      upr     p adj
## Famous-Excellent         373.6263 -1082.4068 1829.659 0.9119276
## Intermediate-Excellent  1679.8549  1025.9664 2333.743 0.0000000
## Novice-Excellent        3034.1830  1153.2751 4915.091 0.0002092
## Intermediate-Famous     1306.2286  -110.1401 2722.597 0.0829421
## Novice-Famous           2660.5567   398.6229 4922.491 0.0134907
## Novice-Intermediate     1354.3281  -496.0455 3204.702 0.2358077
```

```
plot(TukeyHSD(anova, conf.level = 0.95))
```



Rating Pairs & their Intervals

- Famous-Excellent (-1082.4068, 1829.659) → crosses zero (inconclusive)
- Intermediate-Excellent (1025.9664, 2333.743) → both positive (Intermediate larger)
- Novice-Excellent (1153.2751, 4915.091) → both positive (Novice larger)
- Intermediate-Famous (-110.1401, 2722.597) → crosses zero (inconclusive)
- Novice-Famous (398.6229, 4922.491) → both positive (Novice larger)
- Novice-Intermediate (-496.0455, 3204.702) → crosses zero (inconclusive)

Base on the analysis above we can conclude that there are differences across the rating types. Based on further statistical evidence it can be determined that:

- Intermediate is greater than Excellent
- Novice is greater than Excellent
- Novice is greater than Famous

★ On average the two highest gross sale rating categories include Intermediate followed by Novice.

Task 2

To test the linear relationships of all variables against gross.sales, I would begin by producing a set of scatterplots for each variable pair. We could add the linear model line within the plots to get an even clearer view. Check the residual values for each relationship and if they fall below a threshold we can consider the strength of the linear relationships (close to -1 and 1 are strong, and close to 0 is weak).