Final_Exam

Tejaswini Yeruva

2022-12-16

Installing the required packages:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(tidyr)
library(readr)
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.2.2
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

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

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
getwd()
```

```
## [1] "C:/Users/tejar/OneDrive/Desktop/ML Assignments"
```

```
bestsellers_data_09_22 <-read.csv("C:/Users/tejar/Downloads/Bestsellers.csv")

colnames(bestsellers_data_09_22)
```

```
## [1] "Name"        "Author"      "User.Rating" "Reviews"     "Price"
## [6] "Year"        "Genre"
```

```
str(bestsellers_data_09_22)
```

```
## 'data.frame':    550 obs. of  7 variables:
##  $ Name       : chr  "10-Day Green Smoothie Cleanse" "11/22/63: A Novel" "12 Rules for Life: An Anti
##  $ Author     : chr  "JJ Smith" "Stephen King" "Jordan B. Peterson" "George Orwell" ...
##  $ User.Rating: num  4.7 4.6 4.7 4.7 4.8 4.4 4.7 4.7 4.7 4.6 ...
##  $ Reviews    : int  17350 2052 18979 21424 7665 12643 19735 19699 5983 23848 ...
##  $ Price      : int  8 22 15 6 12 11 30 15 3 8 ...
##  $ Year       : int  2016 2011 2018 2017 2019 2011 2014 2017 2018 2016 ...
##  $ Genre      : chr  "Non Fiction" "Fiction" "Non Fiction" "Fiction" ...
```

To make them easier to read and to keep with the naming convention, I believe column names should be modified.

```
bestsellers_data_09_22 <- bestsellers_data_09_22 %>%
  rename(book_name = Name,
         author_name = Author,
         user_rating = User.Rating,
         reviews_number = Reviews,
         price = Price,
         year = Year,
         genre = Genre)

glimpse(bestsellers_data_09_22)
```

```
## Rows: 550
## Columns: 7
## $ book_name      <chr> "10-Day Green Smoothie Cleanse", "11/22/63: A Novel", "~
## $ author_name    <chr> "JJ Smith", "Stephen King", "Jordan B. Peterson", "Geor~
## $ user_rating    <dbl> 4.7, 4.6, 4.7, 4.7, 4.8, 4.4, 4.7, 4.7, 4.7, 4.6, 4.6, ~
## $ reviews_number <int> 17350, 2052, 18979, 21424, 7665, 12643, 19735, 19699, 5~
## $ price          <int> 8, 22, 15, 6, 12, 11, 30, 15, 3, 8, 8, 2, 32, 5, 17, 4,~
## $ year           <int> 2016, 2011, 2018, 2017, 2019, 2011, 2014, 2017, 2018, 2~
## $ genre          <chr> "Non Fiction", "Fiction", "Non Fiction", "Fiction", "No~
```

All unwanted data should be excluded from the data frame in order to examine data from 2012 to 2022.

```
bestsellers_data_12_22 <- bestsellers_data_09_22[!(bestsellers_data_09_22$year < 2012),]

glimpse(bestsellers_data_12_22)
```

```
## Rows: 400
## Columns: 7
## $ book_name      <chr> "10-Day Green Smoothie Cleanse", "12 Rules for Life: An~
## $ author_name    <chr> "JJ Smith", "Jordan B. Peterson", "George Orwell", "Nat~
## $ user_rating    <dbl> 4.7, 4.7, 4.7, 4.8, 4.7, 4.7, 4.7, 4.6, 4.6, 4.5, 4.5, ~
## $ reviews_number <int> 17350, 18979, 21424, 7665, 19735, 19699, 5983, 23848, 2~
## $ price          <int> 8, 15, 6, 12, 30, 15, 3, 8, 8, 5, 4, 6, 6, 8, 13, 14, 1~
## $ year           <int> 2016, 2018, 2017, 2019, 2014, 2017, 2018, 2016, 2017, 2~
## $ genre          <chr> "Non Fiction", "Non Fiction", "Fiction", "Non Fiction",~
```

Transforming data

1. Creating a variable for the Year-specific average values of the price, rating, and number of reviews.

```
analysing_by_year <- bestsellers_data_12_22 %>%
  group_by(year) %>%
  summarize(average_rating = mean(user_rating),
            average_reviews_number = median(reviews_number), average_price = mean(price))

analysing_by_year
```

```
## # A tibble: 8 x 4
##    year average_rating average_reviews_number average_price
##   <int>          <dbl>                  <dbl>         <dbl>
## 1  2012           4.53                  9334.          15.3
## 2  2013           4.55                  7094           14.6
## 3  2014           4.62                 10514           14.6
## 4  2015           4.65                  9144           10.4
## 5  2016           4.68                 10545           13.2
## 6  2017           4.66                 10560.          11.4
## 7  2018           4.67                 10456           10.5
## 8  2019           4.74                 11185           10.1
```

This displays the median number of reviews, the mean price, and the mean rating for each year.

2. Creating a variable to represent the overall number of books by genre published each year.

```
analysing_by_year_and_genre <- bestsellers_data_12_22 %>%
  group_by(year,genre) %>%
  summarise(books_number = n())
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
analysing_by_year_and_genre
```

```
## # A tibble: 16 x 3
## # Groups:   year [8]
##     year genre       books_number
##    <int> <chr>              <int>
##  1  2012 Fiction               21
##  2  2012 Non Fiction           29
##  3  2013 Fiction               24
##  4  2013 Non Fiction           26
##  5  2014 Fiction               29
##  6  2014 Non Fiction           21
##  7  2015 Fiction               17
##  8  2015 Non Fiction           33
##  9  2016 Fiction               19
## 10  2016 Non Fiction           31
## 11  2017 Fiction               24
## 12  2017 Non Fiction           26
## 13  2018 Fiction               21
## 14  2018 Non Fiction           29
## 15  2019 Fiction               20
## 16  2019 Non Fiction           30
```

This chart displays how many books are published year in each genre.

3. Creating a variable to represent the rating value by price.

```
analysing_by_price <- bestsellers_data_12_22 %>%
  group_by(price) %>%
  summarise(average_rating = mean(user_rating), average_reviews_number = median(reviews_number))

analysing_by_price
```

```
## # A tibble: 36 x 3
##    price average_rating average_reviews_number
##    <int>          <dbl>                  <dbl>
##  1     0           4.78                   6324
##  2     2           4.4                    6042
##  3     3           4.7                    5983
##  4     4           4.71                  10369
##  5     5           4.77                  10302
##  6     6           4.59                  11994
##  7     7           4.56                   9873
##  8     8           4.76                  19576
```

```
##  9     9           4.66                  8922
## 10    10           4.61                  8842
## # ... with 26 more rows
```

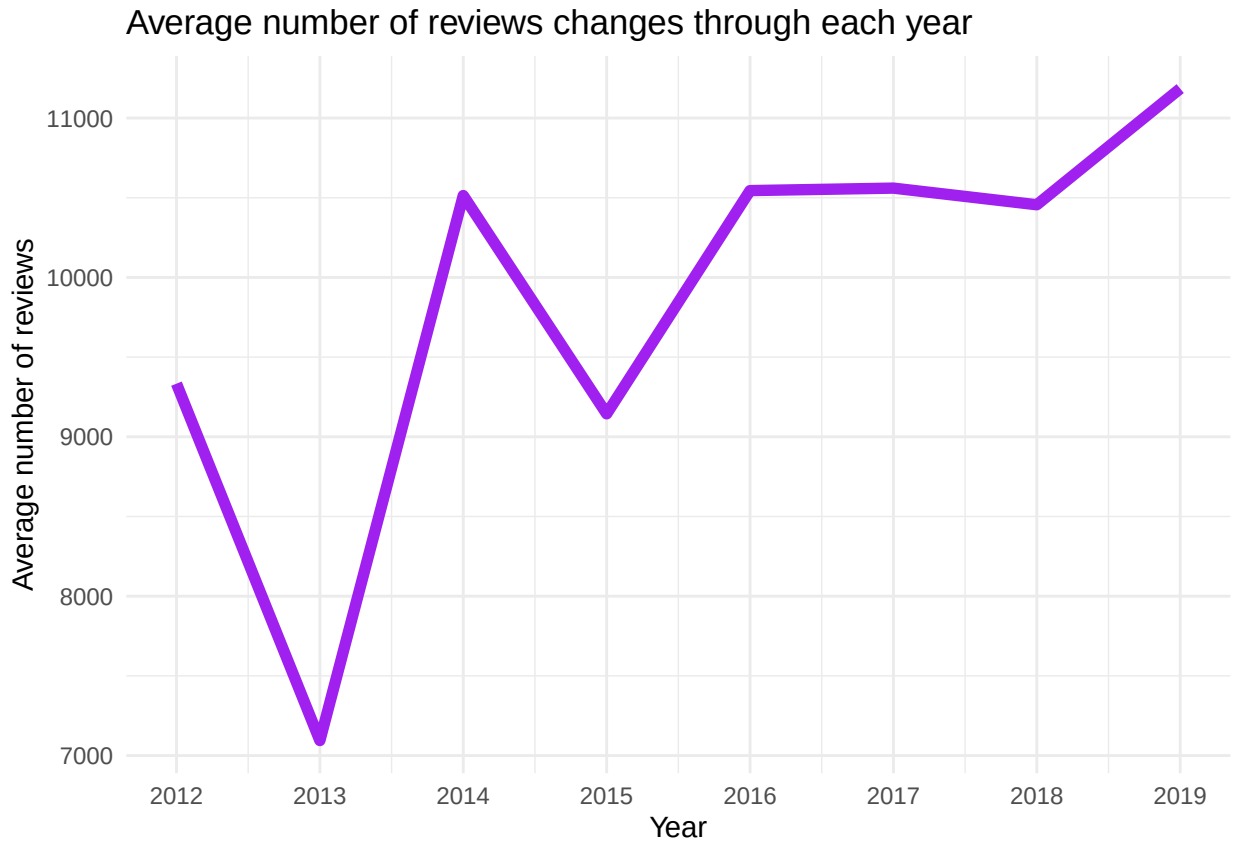
This displays the average rating and quantity of reviews for Price.

Note that the number of reviews for each book can be equivalent to the number of readings. The average number of reviews can be equivalent to the usual amount of readings.

Visualizing and analyzing

Examining the correlation between the year and the average amount of reviews (number of readings).

```
analysing_by_year %>%
  ggplot(aes(x=year, y=average_reviews_number)) +
  geom_line(size=2, color="purple") +
  theme_minimal() +
  labs(title="Average number of reviews changes through each year", x="Year", y="Average number of revi
  scale_x_continuous(breaks = c(2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022))
```

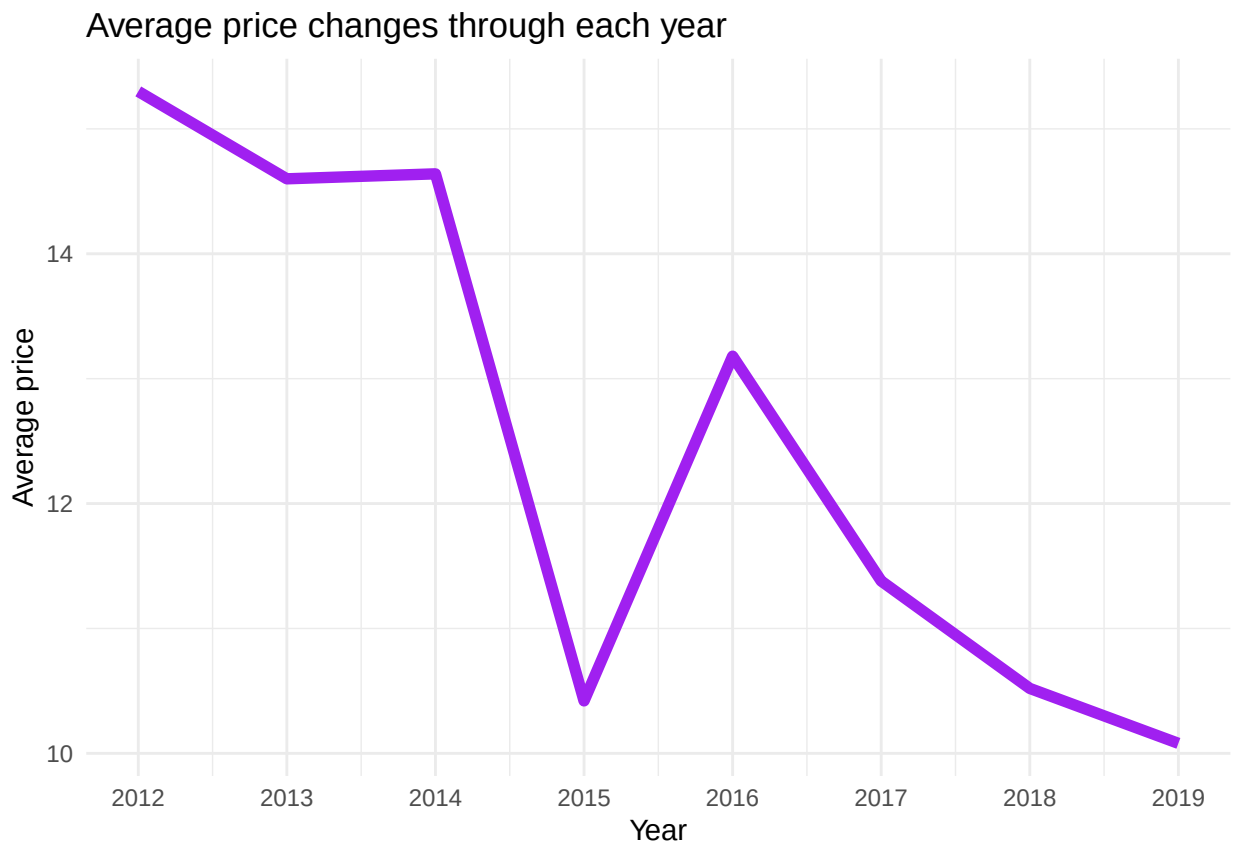


In 2020, there was a sharp increase in the number of reviews (also known as the number of book readings). We can infer that this was caused on by the COVID-19 pandemic, which began in 2020, logically.

Let's now examine how the average price and rating have changed over time.

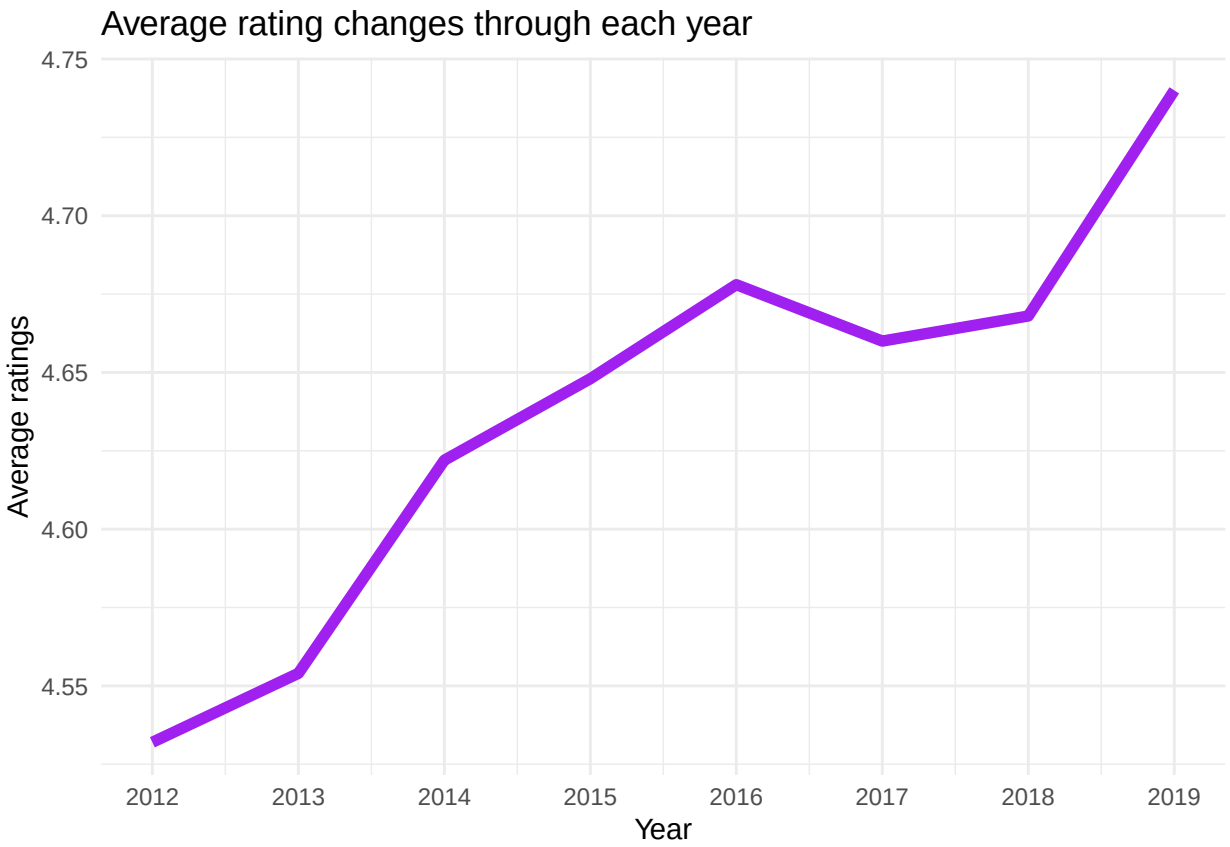
```
analysing_by_year %>%
  ggplot(aes(x=year, y=average_price)) +
  geom_line(size=2, color="purple") +
```

```
  theme_minimal() +
  labs(title="Average price changes through each year", x="Year", y="Average price") +
  scale_x_continuous(breaks = c(2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022))
```



From 2012 to 2019, the average book price decreased, as can be shown. However, we can observe a significant average fall in prices in 2015. The average book price began to rise after 2019. According to my theory, the COVID-19 pandemic is also to cause. The previous visualization stated clearly that there would be a significant increase in demand for reading books in 2020, which, in turn, led to higher costs.

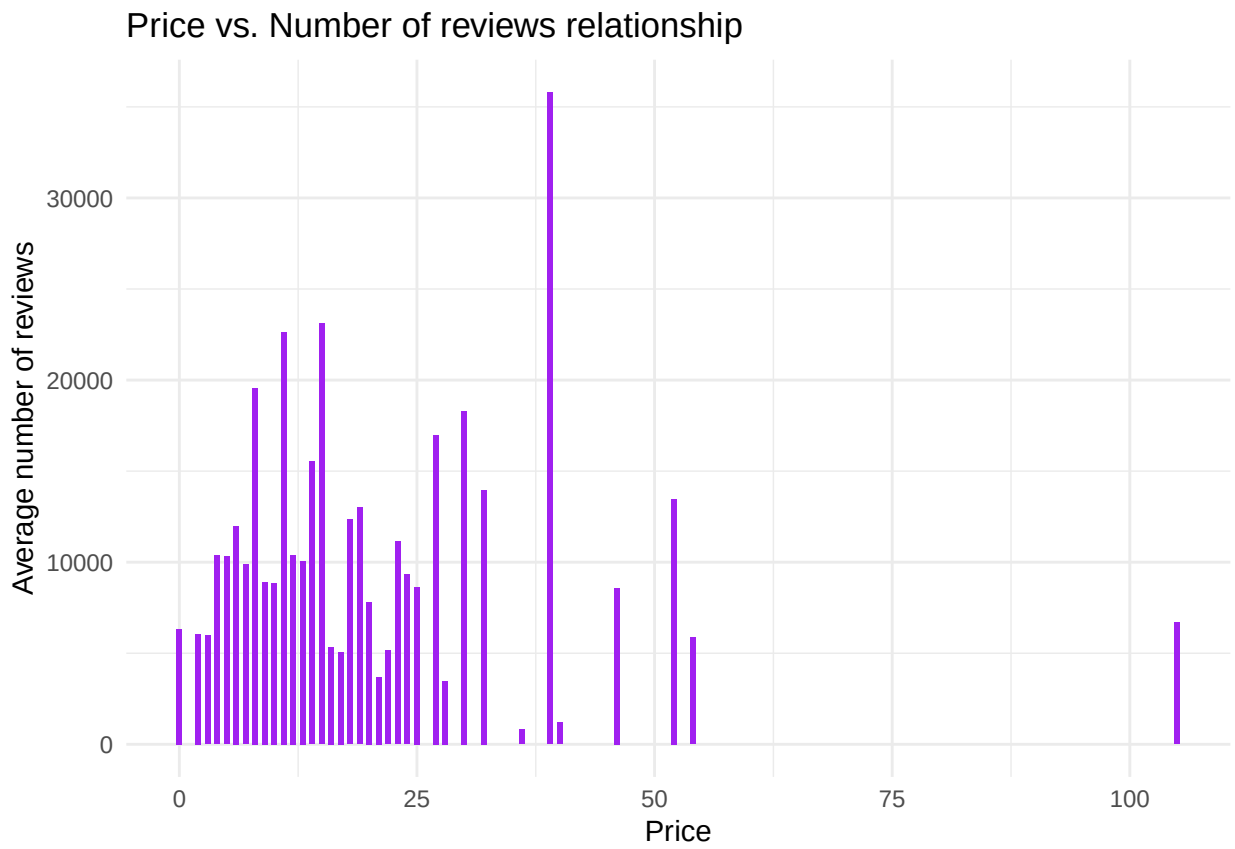
```
analysing_by_year %>%
  ggplot(aes(x=year, y=average_rating)) +
  geom_line(size=2, color="purple") +
  theme_minimal() +
  labs(title="Average rating changes through each year", x="Year", y="Average ratings") +
  scale_x_continuous(breaks = c(2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022))
```



The average ratings are exceptionally high (starting at around 4.5 out of 5), which is not strange considering that this is a bestseller data set. When I compare this figure to the "Average number of reviews fluctuates across each year" plot, I can see that although the average number of reviews nearly remained constant from 2012 to 2019, the average rating increased.

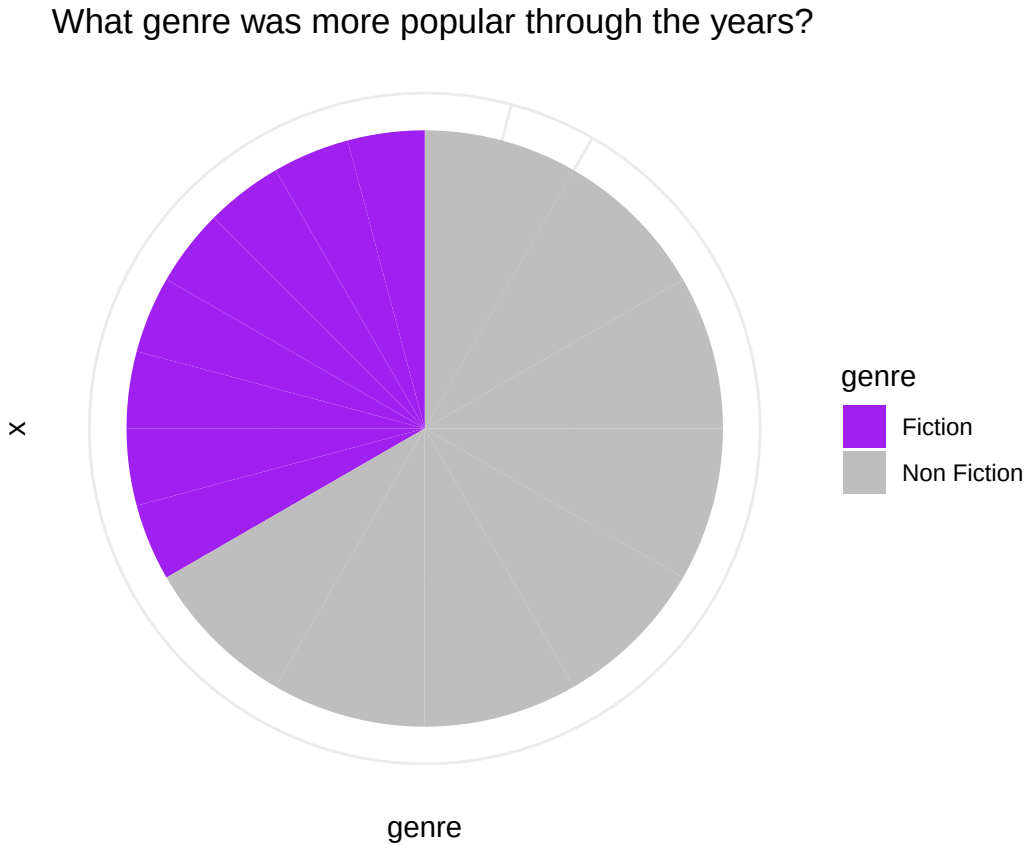
Let's look at how the costs and reviews relate to each other (number of readings)

```
analysing_by_price %>%
  ggplot(aes(x=price, y=average_reviews_number)) +
  geom_col(width=0.6, fill="purple") +
  theme_minimal() +
  labs(title="Price vs. Number of reviews relationship", x="Price", y="Average number of reviews")
```



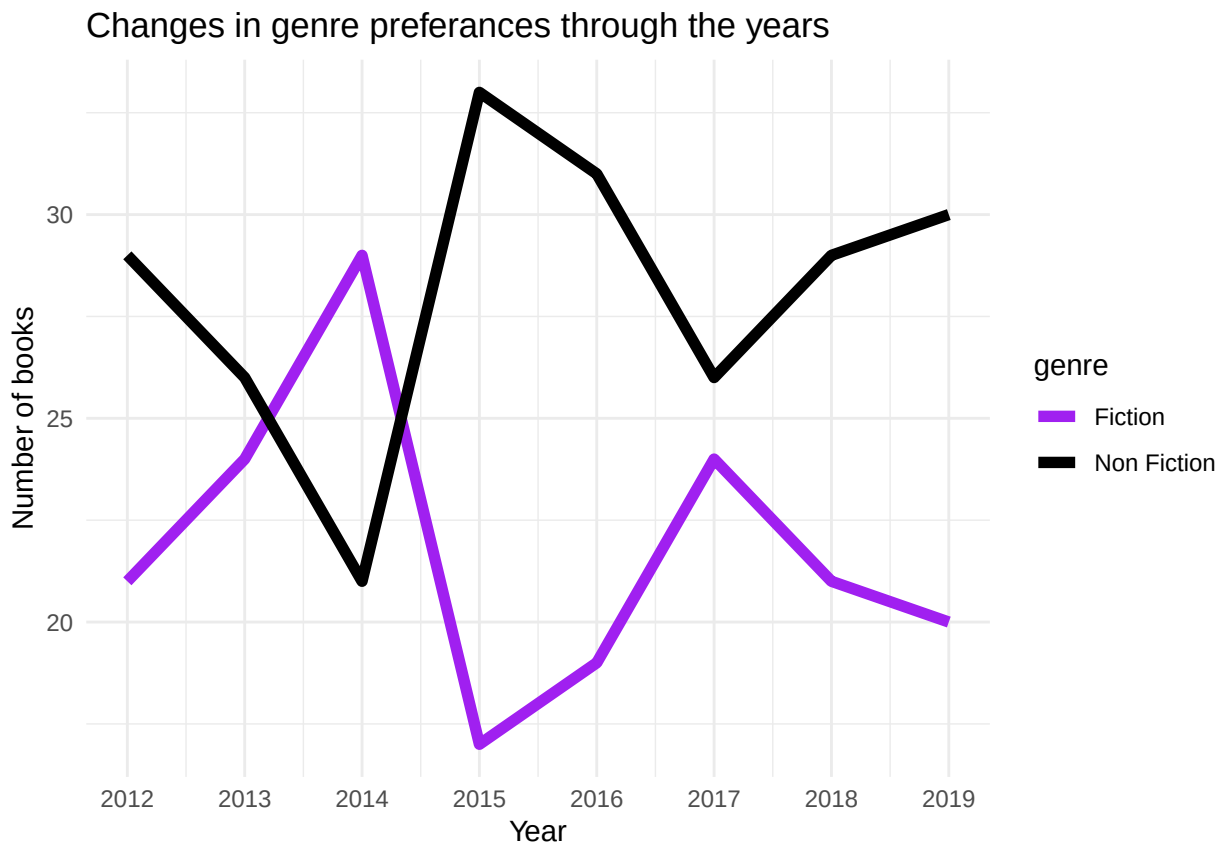
Let's now examine which genre received greater attention.

```
analysing_by_year_and_genre %>%
  ggplot(aes(x="", y=genre, fill=genre)) +
  geom_bar(stat="identity", width = 1) +
  coord_polar("y") +
  scale_fill_manual(values = c("purple", "grey")) +
  labs(title="What genre was more popular through the years?") +
  theme_minimal() +
  theme(axis.text.x = element_blank())
```



We can observe that nonfiction books were far more popular. And now let's analyze a more specific plot - the number of books published in each category over time.

```
analysing_by_year_and_genre %>%
  ggplot(aes(x=year, y=books_number, group=genre, color=genre)) +
  geom_line(size=2) +
  theme_minimal() +
  scale_color_manual(values = c("Purple", "black")) +
  labs(title="Changes in genre preferances through the years", x="Year", y="Number of books") +
  scale_x_continuous(breaks = c(2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022))
```



Non-fiction books were preferred throughout the whole decade, with the exception of 2014. Additionally, both genres became intertwined in 2021, and as a result, the popularity of the fiction genre has surpassed that of non-fiction by 2022.