Getting Data Project

Johnny Li

Getting Data Project

Objectives for this project

To complete this project you'll need to do a few things within this file. As you go through the notebook, you will have further instruction on how to complete these objectives.

Be sure you have followed the steps described in the previous chapter and have your Googlesheet with Leanpub data prepared and ready.

1. Go through this notebook, reading along.
2. Fill in empty or incomplete code chunks when prompted to do so.
3. Run each code chunk as you come across it by clicking the tiny green triangles at the right of each chunk. You should see the code chunks print out various output when you do this.
4. At the very top of this file, put your own name in the `author:` place. Currently it says `"DataTrail Team"`. Be sure to put your name in quotes.
5. In the `Conclusions` section, write up responses to each of these questions posed here.
6. When you are satisfied with what you've written and added to this document you'll need to save it. In the menu, go to `File` > `Save`. Now the `nb.html` output resulting file will have your new output saved to it.
7. Open up the resulting `leanpub_project.nb.html` file and click `View in Web Browser`. Does it look good to you? Did all the changes appear here as you expected.
8. Upload your `Rmd` and your `nb.html` to your assignment folder (this is something that will be dependent on what your instructors have told you – or if you are taking this on your own, just collect these projects in one spot, preferably a Google Drive)!
9. Pat yourself on the back for finishing this project!

The goal of this analysis

<Write here what the goal of this analysis is. What question are we trying to answer?> Is there a correlation between readers and the suggested price of a book on Leanpub?

Set up

We are going to use this R package (we'll talk more about package in a later chapter).

```
library(readr)
library(magrittr)
library(googlesheets4)
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)
```

Set up directories

Here we are going to make a data directory if it doesn't already exist.

```
if (!dir.exists("data")) {
  dir.create("data")
}
```

Getting the data

We'll read in the data using the code below. You will need to follow the prompt to authorize the data to be read in using your Google login credentials.

```
leanpub_df <- read.csv("leanpub_data - Sheet1.csv")
```

So we have a snapshot of what this data look like at the time that we ran this analysis (and for easier sharing purposes), let's use the `readr::write_csv()` function to write this to a file.

Save this file to the `data` directory that we created. And name the file `leanpub_data.csv`. If you don't remember how to use the `readr::write_csv()` function, recall you can look it up using `?readr::write_csv`.

Hint: Look at this chapter for more information on this step: https://datatrail-jhu.github.io/DataTrail/basic-commands-in-r.html#what-is-this-object

```
write_csv(leanpub_df, "data/leanpub_data.csv")
```

Explore the data

Use some of the functions you learned to investigate your `leanpub_df`. How many dimensions is it?

```
dim(leanpub_df)
```

```
## [1] 11  4
```

What kind of class object is it?

```
class(leanpub_df)
```

```
## [1] "data.frame"
```

Cleaning the data

For the upcoming code, we will need to make sure that we have columns named *exactly* `title`, `readers`, `suggested` and `minimum`.

```
# If all four of our required columns are found, then this will print out TRUE
all(c('title', 'readers', 'suggested', 'minimum') %in% colnames(leanpub_df))
```

```
## [1] TRUE
```

If the above prints out false, you may want to return to your Googlesheets, rename the columns accordingly and start from the top of this notebook again.

This code will clean your data for you.

```
leanpub_df <- leanpub_df %>%
  mutate_at(dplyr::vars(readers, suggested, minimum),
                        as.numeric)
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `readers = .Primitive("as.double")(readers)`.
## Caused by warning:
## ! NAs introduced by coercion
```

Now that the data are being treated as numeric values properly, we can obtain some summary statistics for your `leanpub_df` dataset. Use a function we have discussed to do this.

```
summary(leanpub_df)
```

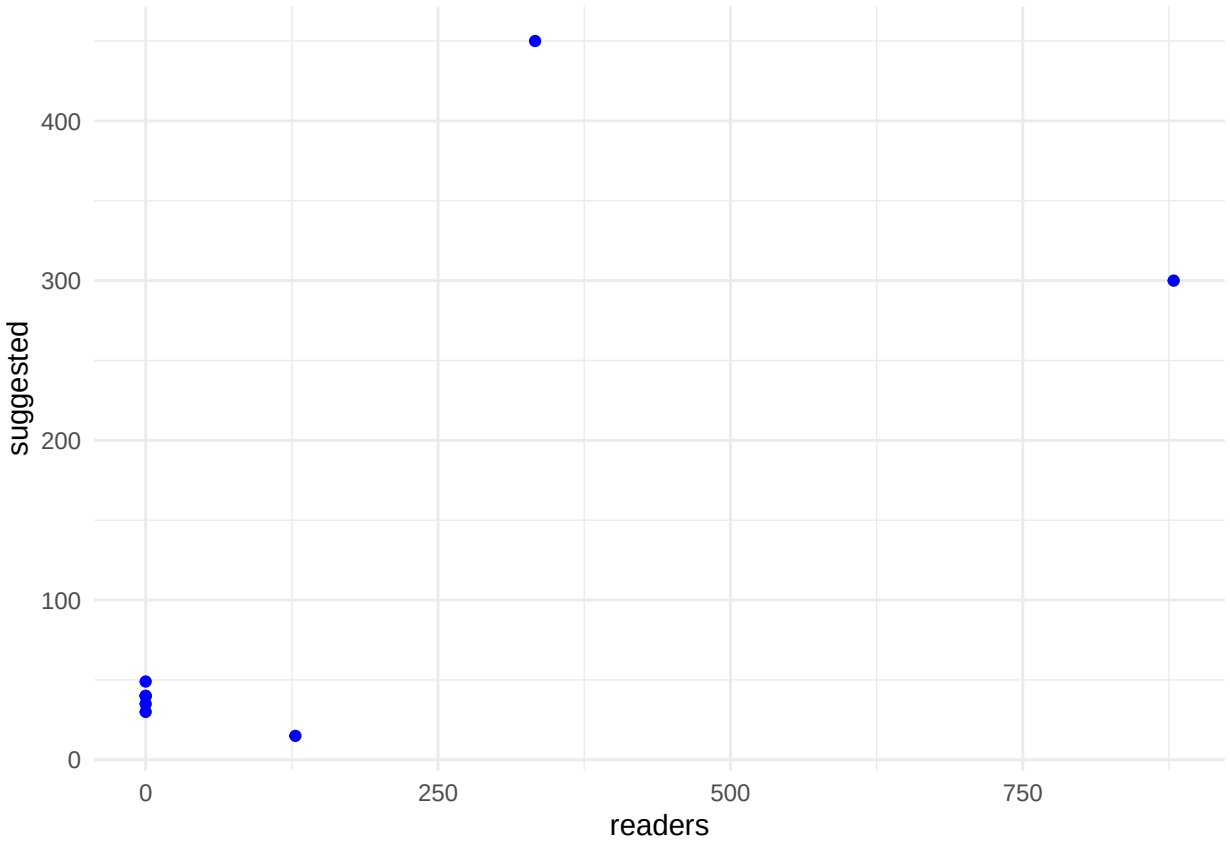
```
##     title              readers        suggested         minimum
##  Length:11          Min.   :  0.0   Min.   : 14.99   Min.   :  0.00
##  Class :character   1st Qu.:  0.0   1st Qu.: 25.45   1st Qu.: 18.00
##  Mode  :character   Median :  0.0   Median : 34.99   Median : 21.49
##                     Mean   :167.5   Mean   : 93.17   Mean   : 76.77
##                     3rd Qu.:179.2   3rd Qu.: 44.50   3rd Qu.: 41.99
##                     Max.   :879.0   Max.   :450.00   Max.   :450.00
##                     NA's   :3
```

Plotting the data

To investigate our question, we will want to investigate any potential relationship between the number of readers for a book and the suggested price. We will plot these data in the form of a scatterplot. In upcoming chapters we will go into more detail about how to make plots yourself.

```
ggplot(leanpub_df, aes(readers, suggested)) +
  geom_point(color = "blue") +
  theme_minimal()
```

```
## Warning: Removed 3 rows containing missing values (`geom_point()`).
```



Get the stats

Is there a relationship between `readers` and `suggested` price? We can also use a correlation to ask this question.

```
cor.test(leanpub_df$readers, leanpub_df$suggested)
```

```
##
##  Pearson's product-moment correlation
##
## data:  leanpub_df$readers and leanpub_df$suggested
## t = 2.541, df = 6, p-value = 0.04402
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0310008 0.9451297
## sample estimates:
##       cor
## 0.7199462
```

If the p value reported is very very small, then there might be a relationship. But also it is likely you'll need to collect more data to get a more confident conclusion using this test.

Because the P-value is less than 0.05, there is a potential correlation between readers and the suggested price of books on Leanpub.

Conclusion

Write up your thoughts about this data science project here and answer the following questions:

- What did we find out about our questions? We found out that there may be a potential correlation between number of readers and the suggested price of a book.
- How did we explore our questions? We explored out quuestion by creating a spreadsheet and extracting data from Leanpub to add the spreadsheet. We then created a csv file from our spreadsheet and uploaded it to our data folder in this project. After doing so, we created a dataframe using the csv file and analyzed the dataframe through various means, such as plotting the data, running correlation tests, etc.
- What did our explorations show us? From analyzing the data and running a correlation test, we see that the P-value of our correlation test is less than 0.05, meaning there is potenetial significance in the correlation between the data.
- What follow up data science questions arise for you regarding this dataset now that we've explored it some? Is there a correlation between number of readers and the minimum price?

Print out session info

Session info is a good thing to print out at the end of your notebooks so that you (and other folks) referencing your notebooks know what software versions and libraries you used to run the notebook.

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/atlas/libblas.so.3.10.3
## LAPACK: /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3;  LAPACK version 3.9.0
##
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
##
## time zone: UTC
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.4.2       dplyr_1.1.2         googlesheets4_1.1.1
## [4] magrittr_2.0.3      readr_2.1.4
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.3      highr_0.10        crayon_1.5.2
##  [5] compiler_4.3.2    tidyselect_1.2.0  parallel_4.3.2    scales_1.2.1
##  [9] yaml_2.3.7        fastmap_1.1.1     R6_2.5.1          labeling_0.4.2
## [13] generics_0.1.3    knitr_1.43        tibble_3.2.1      munsell_0.5.0
## [17] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.1       utf8_1.2.3
## [21] xfun_0.39         fs_1.6.3          bit64_4.0.5       cli_3.6.1
## [25] withr_2.5.0       digest_0.6.33     grid_4.3.2        vroom_1.6.3
## [29] rstudioapi_0.15.0 hms_1.1.3         lifecycle_1.0.3   vctrs_0.6.3
## [33] evaluate_0.21     gargle_1.5.2      glue_1.6.2        farver_2.1.1
## [37] cellranger_1.1.0  googledrive_2.1.1 fansi_1.0.4       colorspace_2.1-0
```

```
## [41] rmarkdown_2.23    purrr_1.0.1       tools_4.3.2       pkgconfig_2.0.3
## [45] htmltools_0.5.5
```