```
---
title: "Question 5: Googleplay"
author: "Grace Grant"
date: "June 2023"
# date: "`r Sys.Date()`"
bibliography: Tex/ref.bib       # Do not edit: Keep this naming convention and location.
output:
  pagedown::html_paged:
    # template: wp_paged.html
    # css: ['wp.css', 'wp-fonts.css', 'wp-page.css']
    css: ["Template/default-fonts-Texevier.css", "Template/default-page-Texevier.css",
"Template/default-Texevier.css"]
    csl: Template/harvard-stellenbosch-university.csl # referencing format used.
    template: ["Template/paged-Texevier.html"]

    toc: false
    # change to true for a self-contained document, but it'll be a litte slower for
Pandoc to render
    self_contained: TRUE


---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, message=FALSE, warning=FALSE)
pacman::p_load(modelsummary, gt, knitr, kableExtra, tidyverse)
list.files('code/', full.names = T, recursive = T) %>% .[grepl('.R', .)] %>% as.list()
%>% walk(~source(.))
googleplay_data <- read.csv("/Users/gracegrant/Documents/Postgrad/Masters/Data
Science/Take home exam/Question 5/data/googleplay/googleplaystore.csv")
reviews_data <- read.csv("/Users/gracegrant/Documents/Postgrad/Masters/Data Science/Take
home exam/Question 5/data/googleplay/googleplaystore_user_reviews.csv")
merged_google <- merge(googleplay_data, reviews_data, by = "App")

```
App size and ratings linked to number of installs
```{r}
bubble_chart <- bubble(merged_google)
bubble_chart
```

The graph provided above gives an indication of the link between the size of the app and
its rating, with most apps having ratings above 4 and sizes below 100M. I would,
therefore, suggest an app with a size below 100M. The graph also shows that, while most
apps are linked to positive reviewer sentiment, there does not seem to be a link between
reviewer sentiment and the number of installations. This can show that, even though it
is important to listen to what reviews have to say, this does not necessarily correlate
with how many people choose to install an app. If profitability is linked to the number
of installations, it would be a better idea to focus on the app rating and keeping those
numbers high while also ensuring that apps are not too big in size.

App ratings and categories
```{r}
filter_merged <- merged_google %>% filter(!grepl("varies|Varies", Size)) %>% na.omit()
%>% filter(!grepl("NAN|Nan|NaN|nan", Sentiment))

    filter_merged$Installs <- as.numeric(gsub("[^0-9]", "", filter_merged$Installs))
    filter_merged$Size <- as.numeric(gsub("[^0-9]", "", filter_merged$Size))
    filter_merged$Installs <- filter_merged$Installs / 1000

categories <- filter_merged  %>% select(c("App", "Category", "Rating", "Installs")) %>%
unique() %>% arrange(desc(Rating)) %>% slice(1:20)

categories_table <- kable(categories, row.names=TRUE,
                          caption = "Top 20 Rated Apps",
                          booktabs = TRUE) %>%
    kable_styling(full_width = TRUE,
                  latex_options = c("striped", "scale_down", "HOLD_position"),
                  font_size = 10)
```
```

categories_table
```

This table shows the categories that the top 20 highest rated apps fall into. There are quite a variety of categories in the top 20 but many of the categories are linked to ways that help people, whether this is with health, navigation, mathematics or practising for a written driving test. I would thus suggest creating an app that offers a helpful service, maybe linked to the health and fitness category. The top 20 apps also have a similar number of installations which suggests this is the range to aim for.
```