Analysis on Google Play Store apps

Name: Yaling Xiong

1. Introduction

This project aims to explore what kinds of apps can gain the highest profit based on Google Play Store dataset. This project can help developers to decide what kinds of apps to develop to maximize their profit, and also can give a general visualization for investors to invest what kinds of apps to get more returns. Therefore, **the intended audience of this project are developers and investors.** For developers, they should know more about what kinds of apps will be more popular to gain more profit. Only when they know this information, developers can develop the more popular app and gain more profit from them. For investors, they should know what kinds of apps will be popular in the customer market. Once more customers purchase this app, this app will help investors to gain more profit.

In this project, there are many features of apps presented for audiences to analysis. Firstly, audiences can see the raw dataset for this project. Then this project shows many graphs to present the relationship between profit that developers and investors can gain and many features of apps, such as category, last updated time and some numeric features. Besides that, this project also presents the word cloud of reviews for the most popular apps in google play store, giving audiences a more specific feedback to improve their app. Based on these graphs, developers will know what kind of apps will be more popular and they can develop these kinds of apps to gain more profit.

2. Design

Before implementing this project, designers should think carefully about how to design the project and how to give audiences a clear presentation. Therefore, designers plot the five design sheets as below.

Figure 2.1 is the sheet one, ideas sheet. In this ideas sheet, the graph presents all possible ideas of this project. Then they filter and categorize these ideas and finally combine and refine these ideas. After combining, there are four kinds of ideas left: the first is boxplot of type feature and data table, and the second is pie chart and line chart of category feature, and the third is correlation matrix and line chart of numeric features, and the last one is the line chart of last update time.

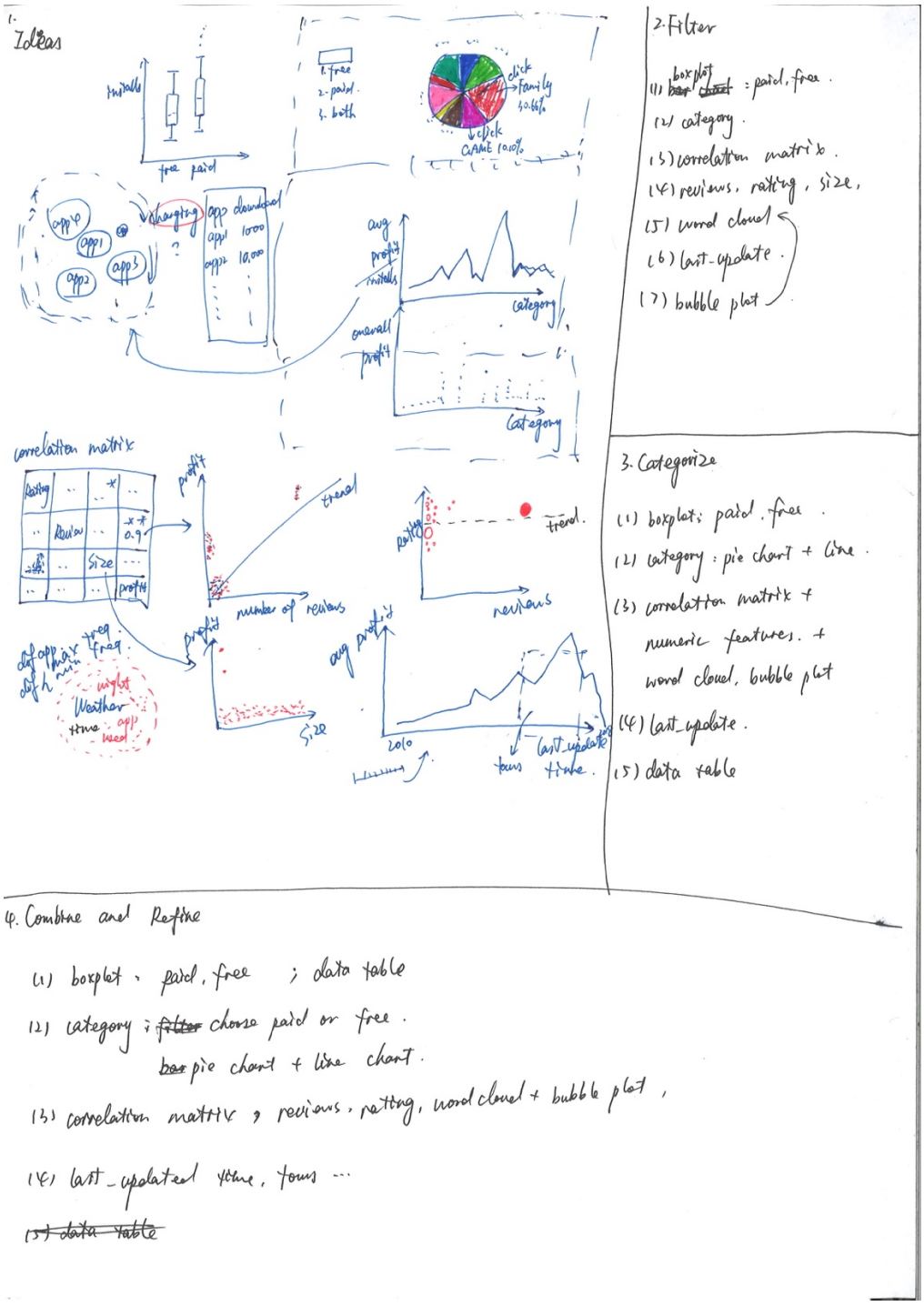


Fig 2.1 The ideas sheet

Figure 2.2 is the sheet two, first alternative design. In this layout, there are two kinds of information. **The first is the data table for audiences to view and the boxplot for different types of apps.** The boxplot aims to present the different distribution of number of installs for different type of apps. Because this project wants to help people gain more profit, the profit for free apps and paid apps are different. **This project regards number of installs as profit for free apps and regards the product of price and number of installs as profit for paid apps. The second is the line chart of profit and last update time**, which used to find the relationship between updated time and profit.

In this design, there is some slightly change when implementing. For the presentation of table, audience can only click on which table to view in the table showing part instead of choosing from the filter. For the type filter, audience can click on both two types to show all information instead of choosing the all as the type.

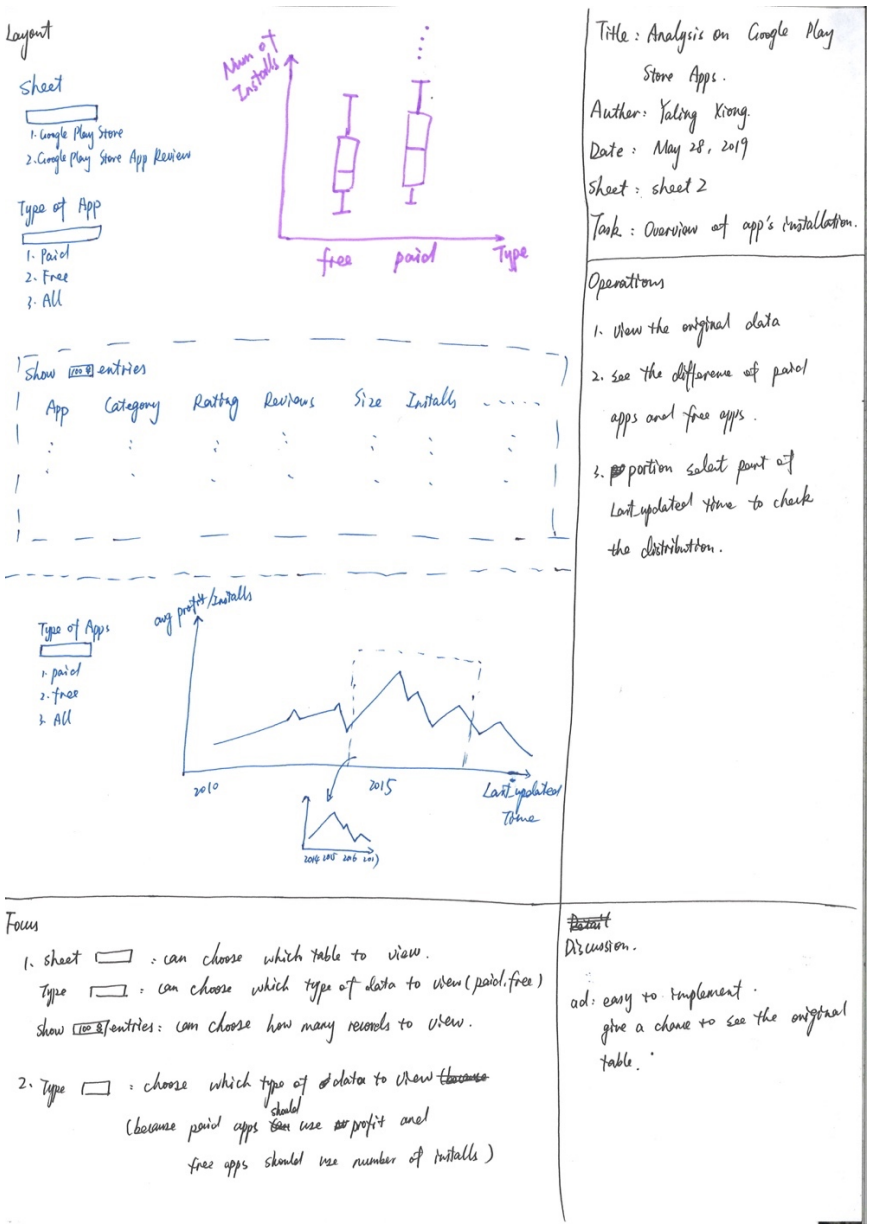


Fig 2.2 The first alternative designs

Figure 2.3 is the sheet three, second alternative design. This alternative design presents **distribution of category feature** using pie chart, and the **relationship between category feature and profit** using line chart and scatter graph. From these graphs, audience can identify which category apps is more likely to help them gain more profit.

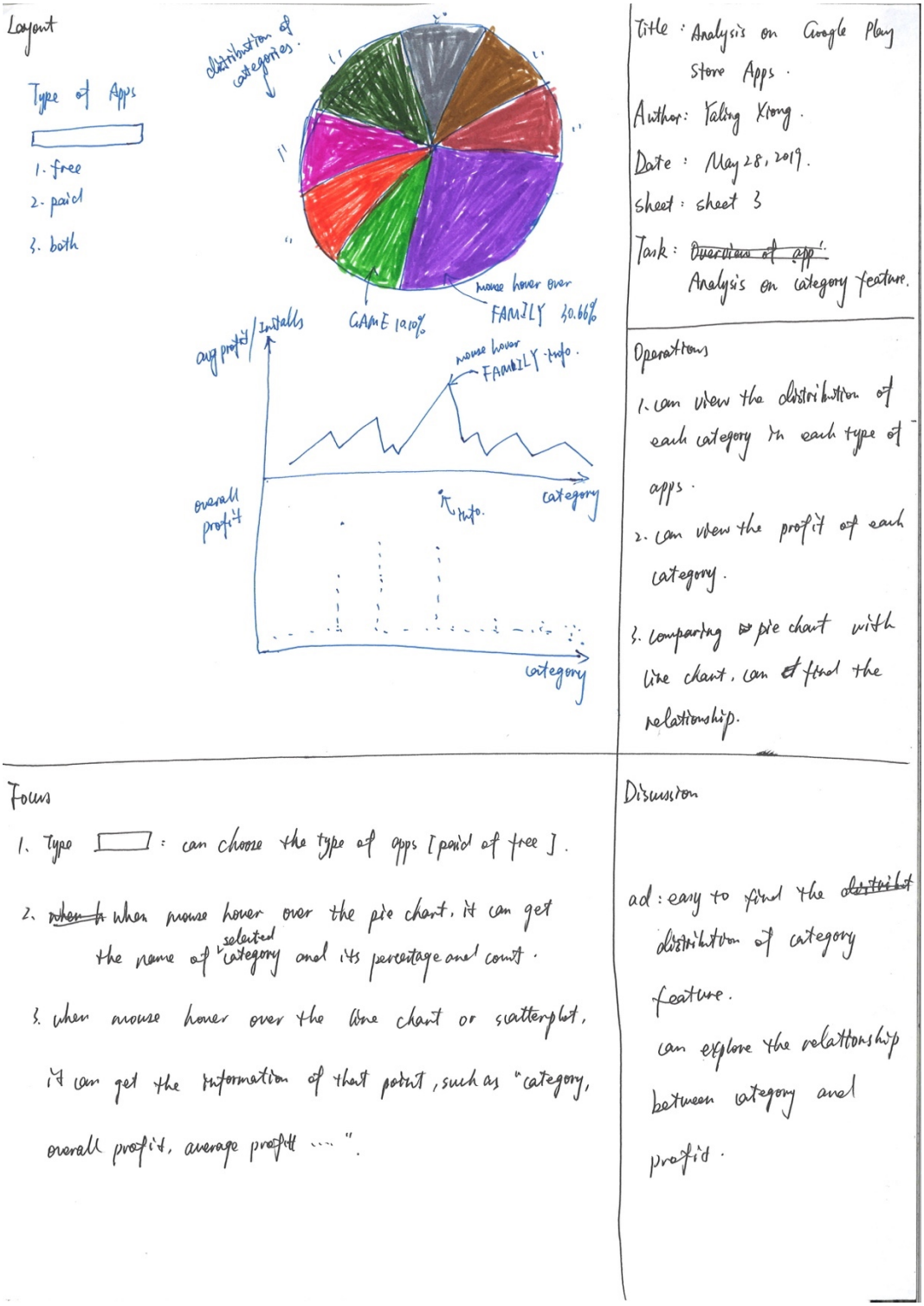


Fig 2.3 The second alternative design

Figure 2.4 is the sheet four, third alternative design. In this layout, there are two kinds of information. **The first is the relationship between numeric features of apps and profit people can get.** In order to present this information, this project shows two graphs. **One is the correlation matrix of numeric features and profit**, which shows the relevant level between profit and other numeric features. **The other is scatter plot of different numeric features and profit** and audiences can choose which numeric feature to show. This can present the specific relationship between numeric features and profit and help developers to design a more popular app and help investors to choose a more profitable app to invest.

The second information in this layout is the word cloud of reviews for each app. There is some slightly change in this graph. **In the implement part, this project only shows the most profitable apps with highest installations instead of all apps**. And there is no bubble graph to show the number of installs for each app. Because the number of installations in google play store dataset is a category feature, there are many apps having the highest installations. In this case, there is no need for the project to use the number of installations as size of each bubble and plot the bubble graph for them. Therefore, this project chooses those apps with highest installations to show their word cloud and use a filter for users to choose the specific app.

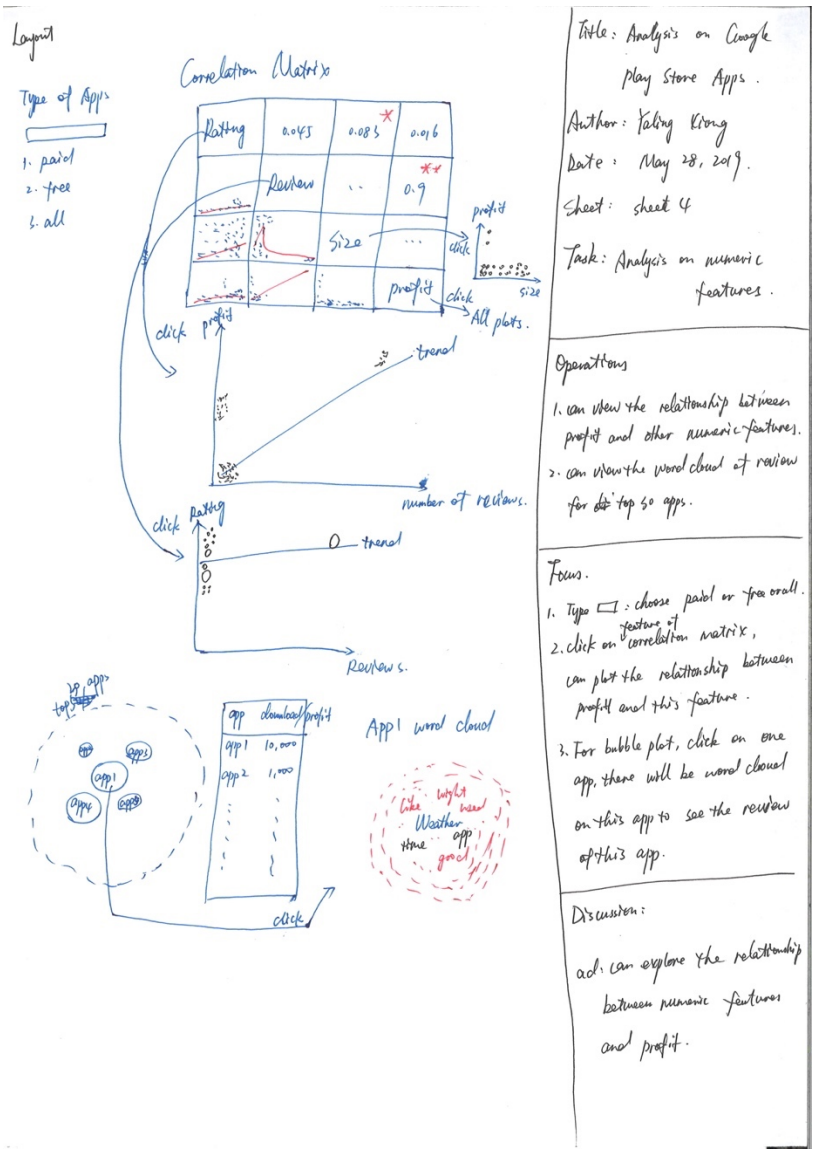


Fig 2.4 The third alternative design

Figure 2.5 is the sheet five, realization design. The layout designs the overall view of this project. For this project, designers want to give a dashboard for audiences to view. When audiences click on each part in the dashboard, the main panel will show the information in this part.

When implementing, this project adds one part in dashboard instead of four. This project split the word cloud from numeric feature analysis as a single part in the dashboard for later analysis. Because the word cloud is for reviews of popular apps, splitting it into one single part will be clearer for users to view.

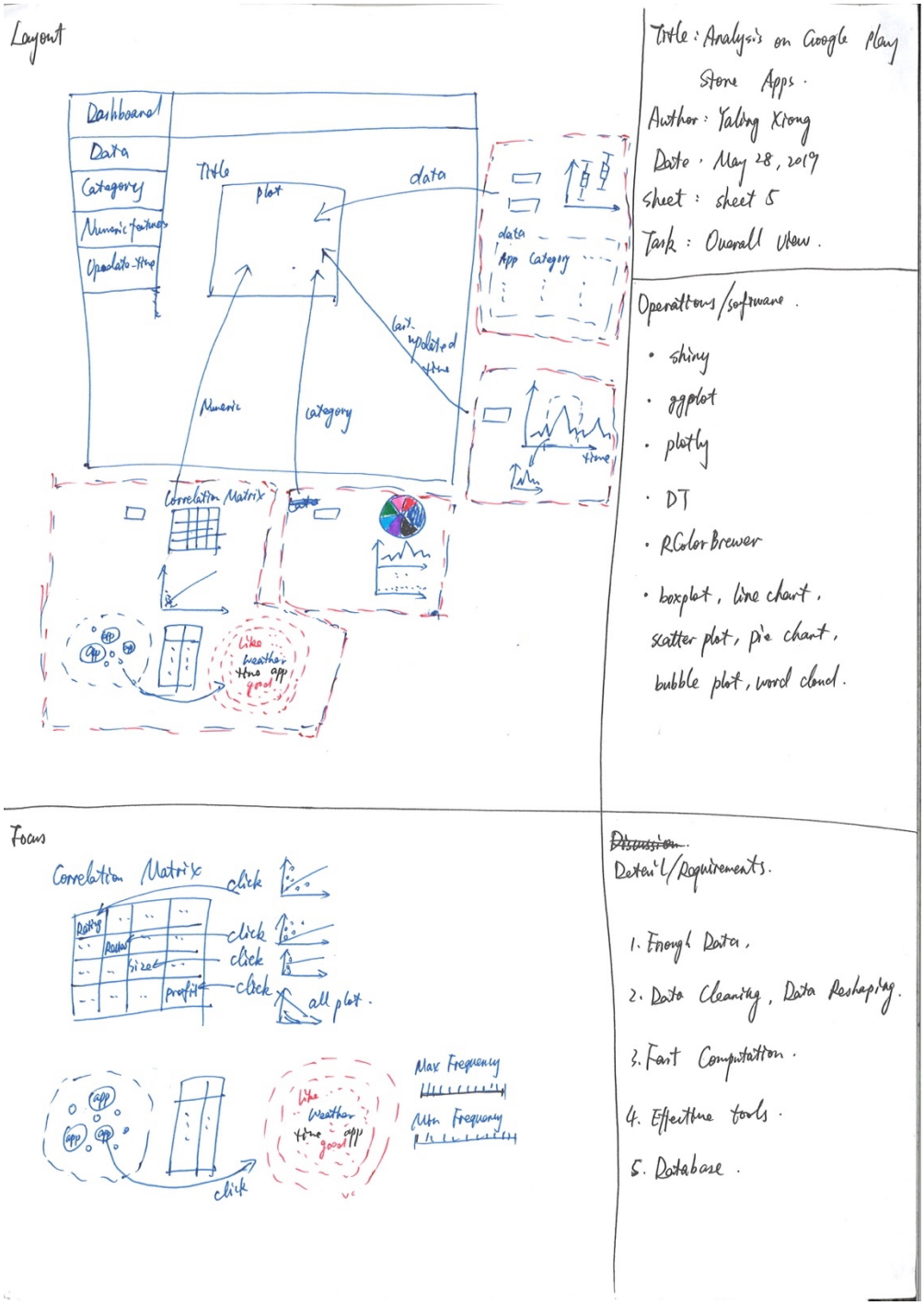


Fig 2.5 The realization design

3. Implementation

In this implementation part, this project uses many libraries for different usage. The table below introduces the usage of each library used in this project.

Library	Reasons and Usage
shiny	For using shiny
ggplot2	For basic plot
plotly	For interactive plot
shinydashboard	For creating dashboard in shiny
DT	For presenting table in shiny
PerformanceAnalytics	For plotting correlation matrix
tm	For preprocessing reviews and plotting word cloud
SnowballC	For preprocessing reviews and plotting word cloud
wordcloud	For plotting word cloud
dygraphs	For plotting time series graph

4. User guide

4.1 General view

When user runs this project, the overall dashboard will appear, such as figure 4.1 below. For the dashboard in the left, user can **choose which kind of information you want to analysis**. Also, there is a symbol right close to "Analysis on Apps", which can **hide the dashboard if user clicks on it.** There are 5 kinds of information user can analyze and we will discuss them one by one later.

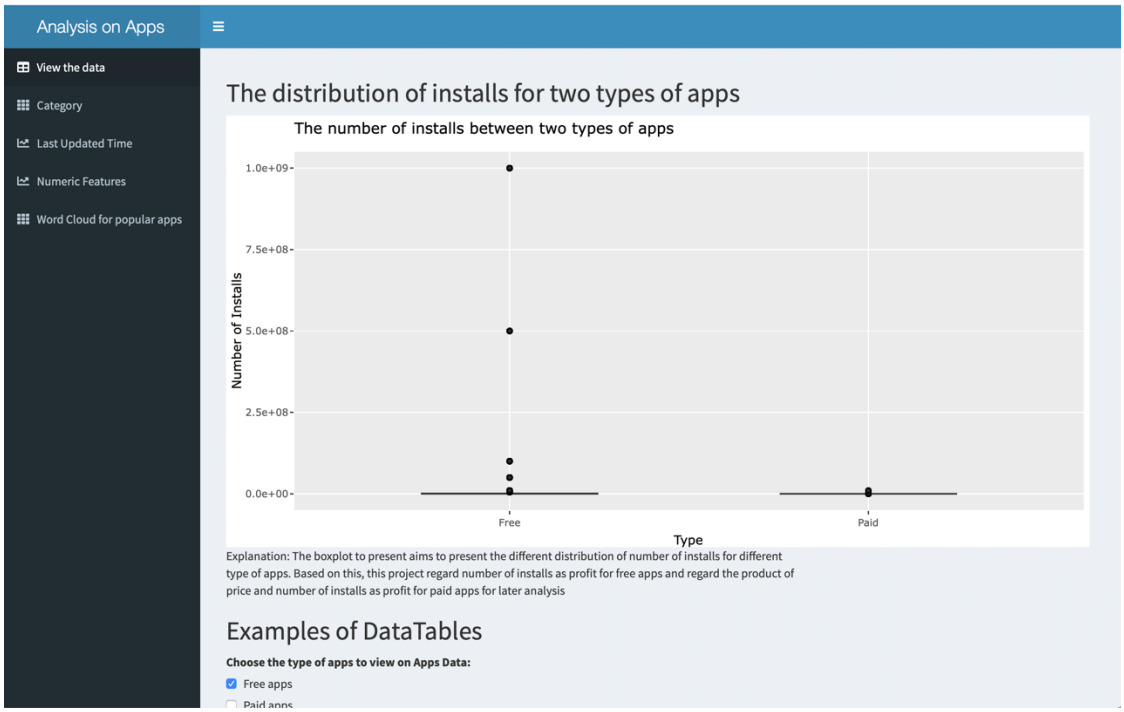


Fig 4.1 Overall dashboard

4.2 View the data

When choosing the "View the data" part of dashboard, user can see graphs, shown in figure 4.1. There are two parts of this part. **One is the distribution of installs for free and paid apps**. In this boxplot, **user can get number of installs of each point, the maximal and minimal installs when putting the mouse on that point**. There is also an explanation below the boxplot, which tells user to distinguish the profit of paid apps and free apps.

When page slides down, user can see the second part, shown in figure 4.2. **This part shows the data table for two datasets.** User can click on the "AppsData" or "AppsReviewData" to **choose one dataset** to view. And user can also **click the columns to show** in the left checkbox. For Apps Data, user can **choose the type of apps to view** and the default type is free. For the table content, user can **scroll down to see more rows and scroll left and right to see more columns**. Also, user can choose **how many entries to see** in one page and the default is 10.

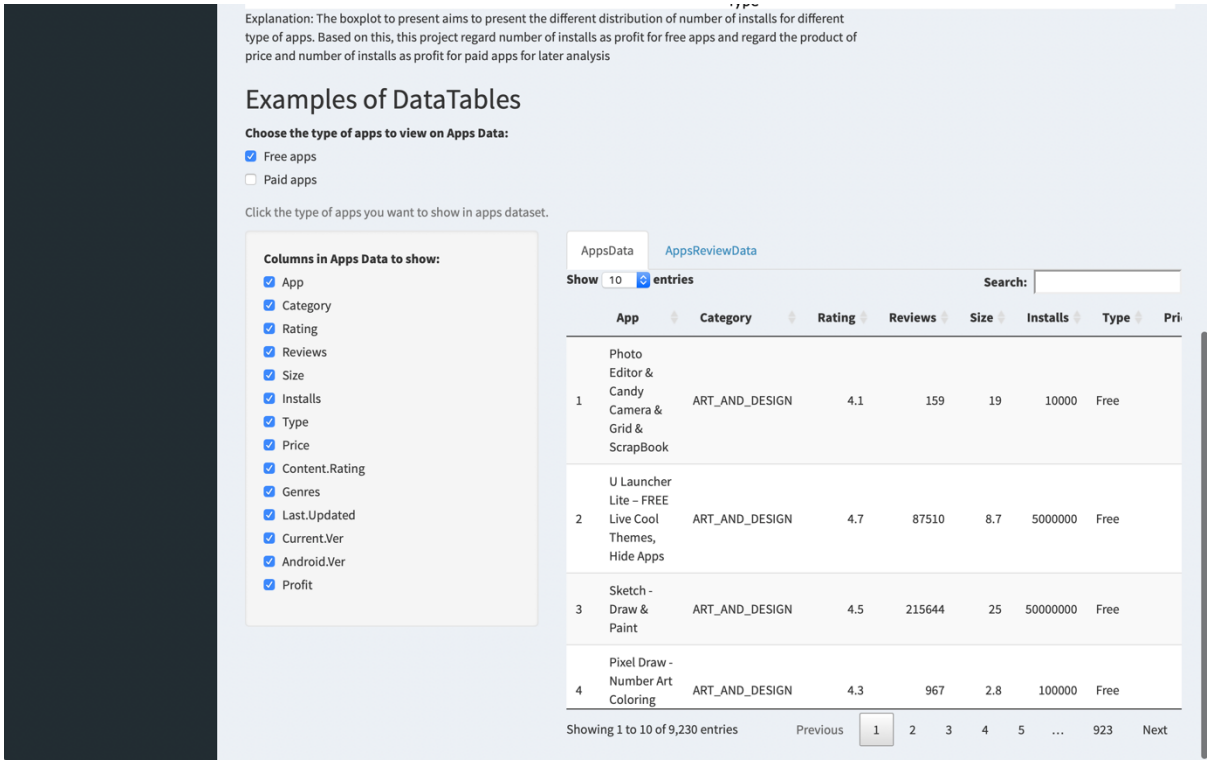


Fig 4.2 Showing Data table

4.3 Category

When choosing the "Category" part of dashboard, user can see graphs, shown in figure 4.3. **In this part, there are one filter and three graphs.** The filter is shown in figure 4.3, which is the type of apps. User can **choose each type of apps to view** or click both types of apps to show, and the default type is free. After choosing the type of apps, there are three graphs to show. **One is the pie chart of distribution of each category,** shown as figure 4.3. User can **move mouse one each color** in pie chart to see the detailed information of this category, such as category name, overall frequency of this category and the percentage of this category. Besides this, user can **click one category in the right legend** to remove this category from pie chart and view the distribution of other categories.

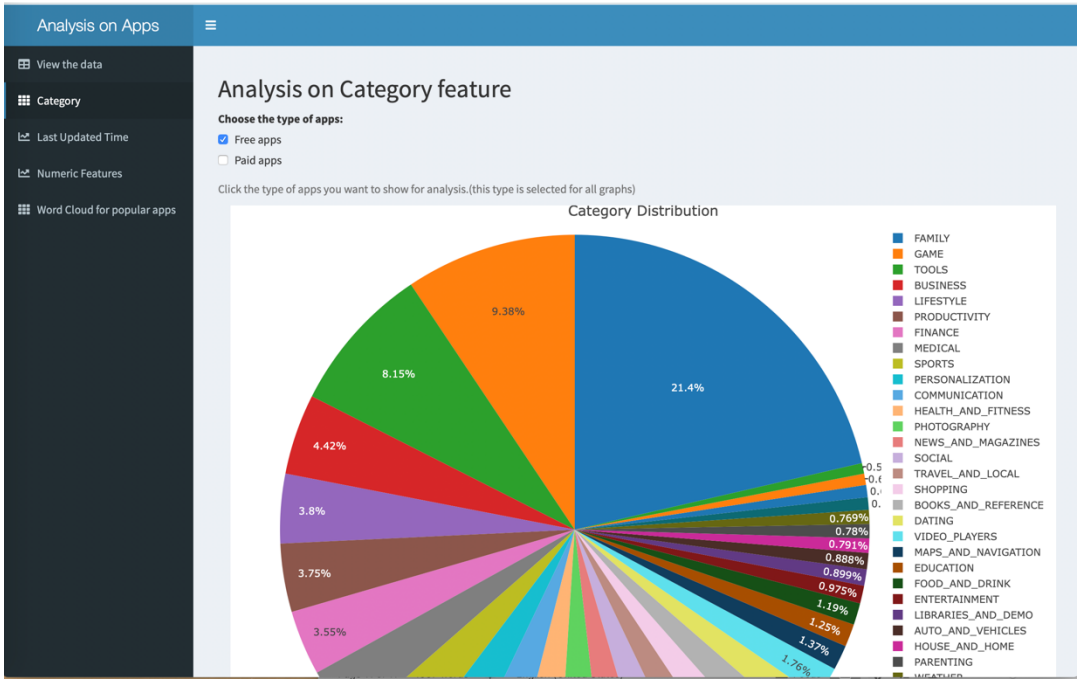


Fig 4.3 Category pie chart

The left two graphs are line chart and scatter plot, shown in figure 4.4. After choosing the type of apps, user can view the distribution of profit for different categories of apps. The first graph is the **relationship between average profit and each category**. User can **move mouse one each point to see the detailed information**, such as category name and corresponding average profit. The second graph is the **relationship between overall profit and each category**. User also can **move mouse on each point to see the detailed information**, such as category name and corresponding overall profit. Compared to two graphs, user can find the relationship between profit and category and also find the difference of overall profit relationship and average profit relationship.

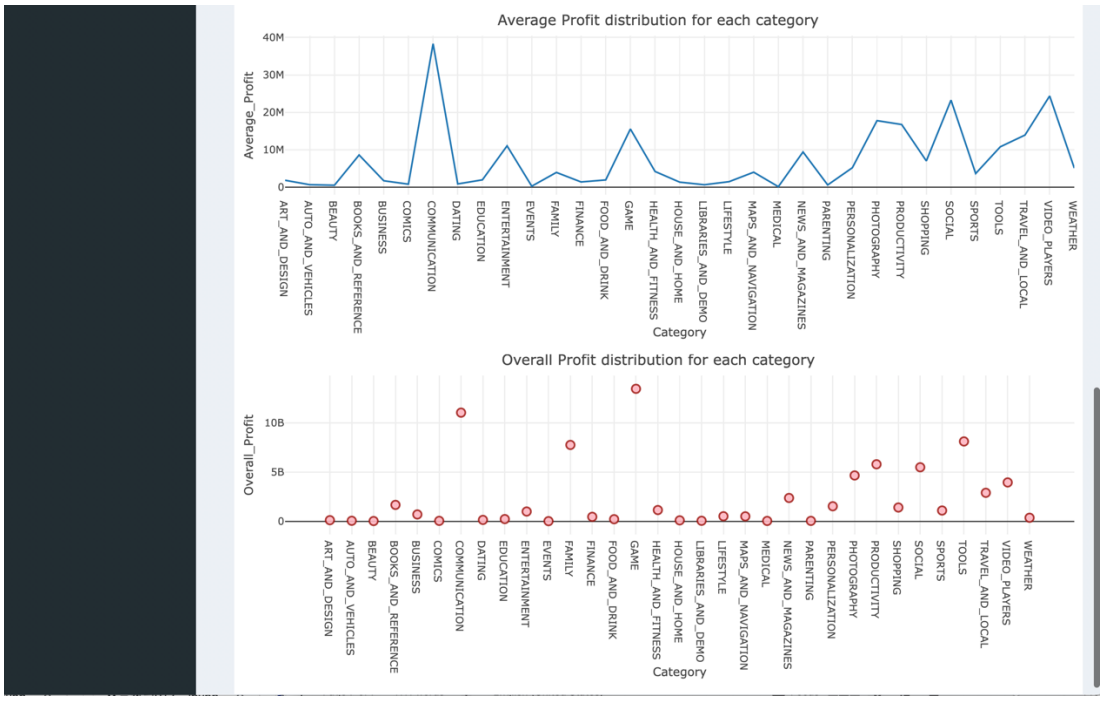


Fig 4.4 Category line chart and scatter plot

4.4 Last updated time

When choosing the “Last Updated Time” part of dashboard, user can see graphs, shown in figure 4.5. **In this part, there are one filter and one graph. The filter is to choose the type of apps to view.** User can click one the type to choose or click two types to show the overall data, and the default type is free.

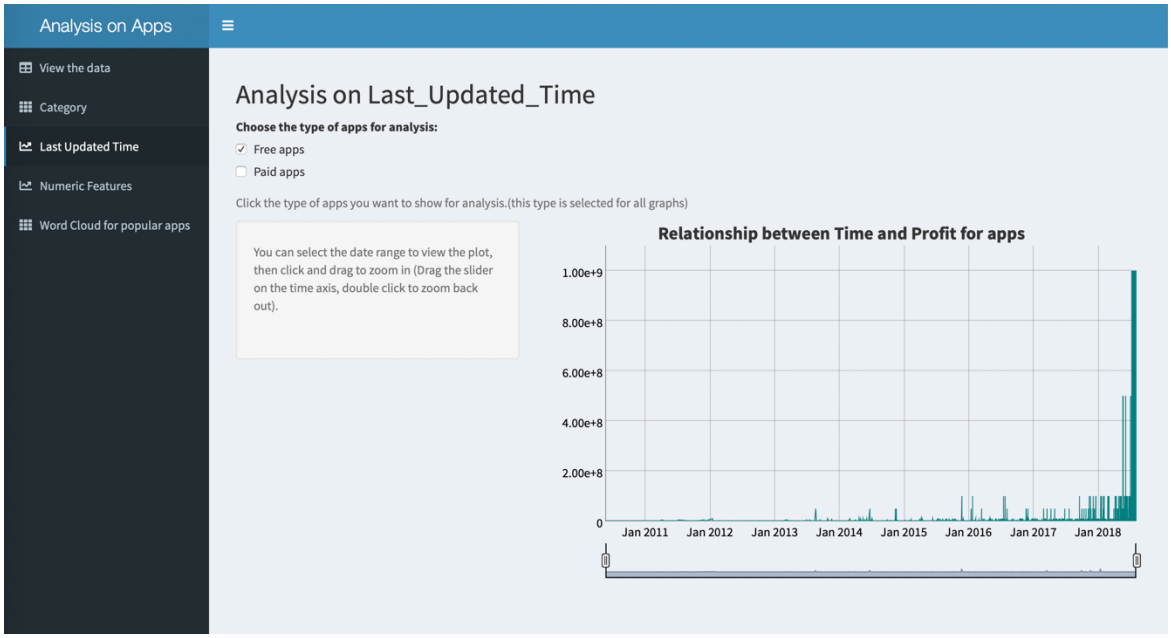


Fig 4.5 Last Updated Time

The graph is the line chart presenting the relationship between last updated time and profit. User can **drag the line below the graph to select date range** to view the plot. When user drag the line, the x axis of this plot will change to the specific date range, shown as figure 4.6, changing from Jan 2011 - Jan 2018 to Jul 2017 - Jul 2018. If user wants to get back to the original date range, user only need to **double click the line to zoom out**.

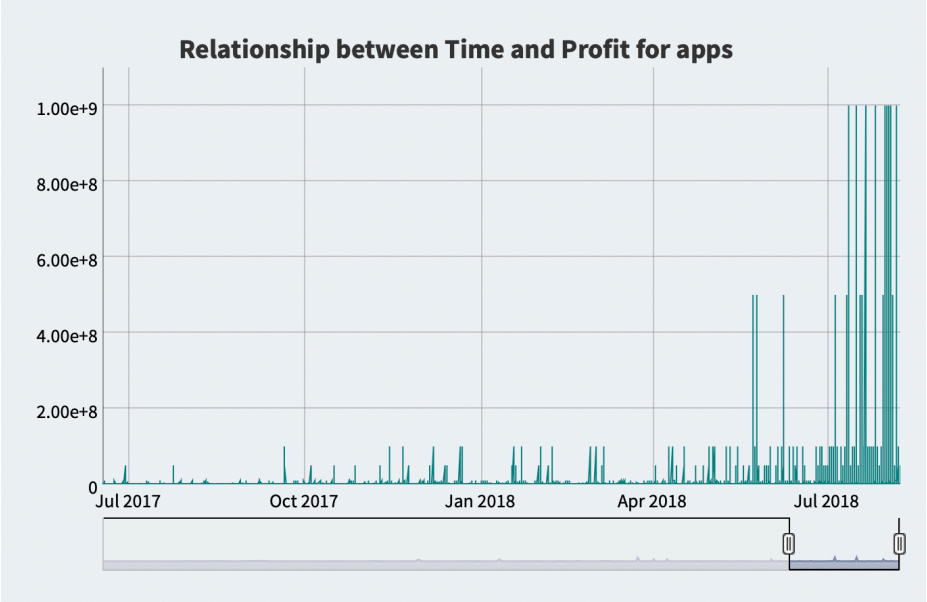


Fig 4.6 Graph with specific date range

4.5 Numeric features

When choosing the "Numeric Features" part of dashboard, user can see graphs, shown in figure 4.7. **In this part, there are two graphs. The first is correlation matrix of numeric features and profit of apps.** At first, user should **choose the type of apps to view** in the select box in the left, and the default type is free. After that, user can view the correlation matrix.

The second graph is the scatter plot between numeric features and profit based on specific type of apps selected. User can **choose which numeric feature to view** in the scatter plot. After choosing the feature, the scatter plot is presented on the right. User also can **move mouse on each point to get the detailed information** of this point, such as the value of numeric feature and its profit.

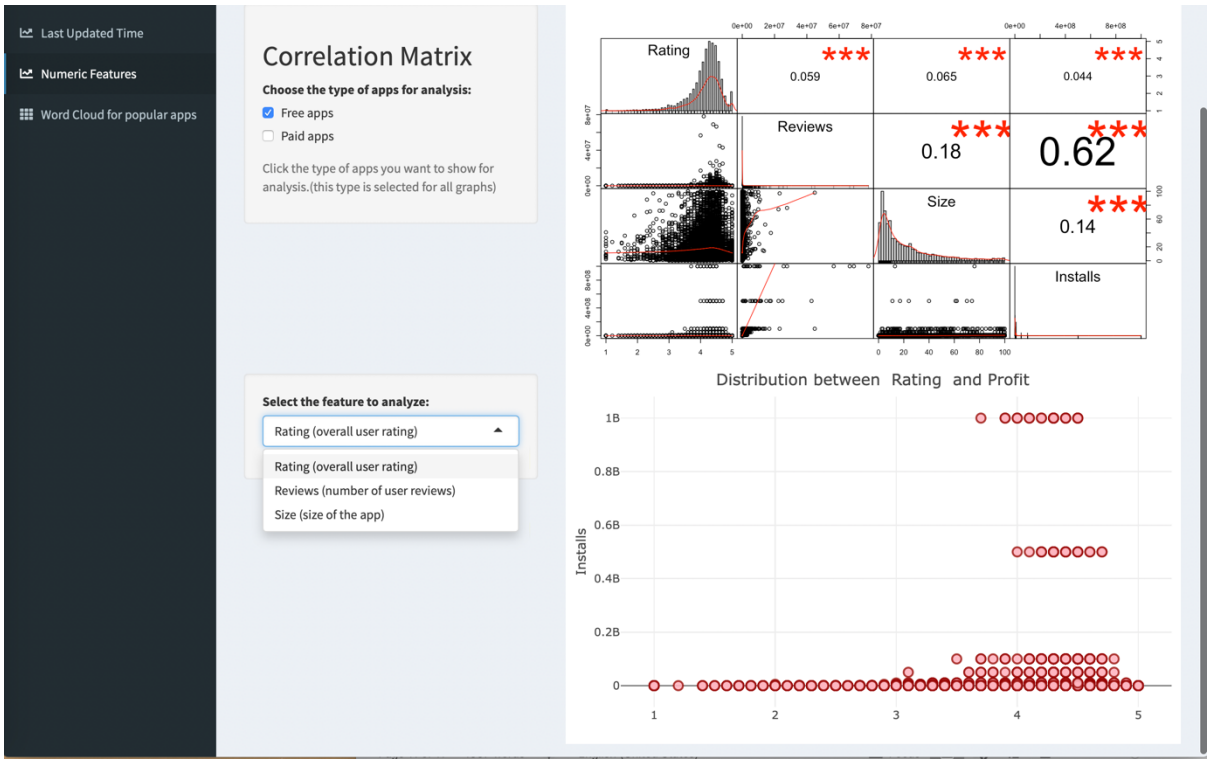


Fig 4.7 Numeric features

4.6 Word cloud for popular apps

When choosing the "Word Cloud for popular apps" part of dashboard, user can see graphs, shown in figure 4.8. **In this part, there are three filters and one graph. The first filter is to choose the app name to view its word cloud.** These apps are selected from review dataset with highest number of installs. User can **choose each app name to view its word cloud** from the list, and the default app is the first one.

After choosing the specific app, this page presents word cloud for the app's reviews. Then there are two filters for visualizing word cloud. **One is the minimum frequency,** which is the minimum frequency of words to show in this word cloud, and the default value is 15. User can **drag the point to choose the minimal frequency of words** user wants to show in word cloud. **The second is the maximum number of words**, which is the maximal number of words to

show in this word cloud, and the default value is 100. User can **drag the point to choose the maximal number of words** user wants to show in word cloud.

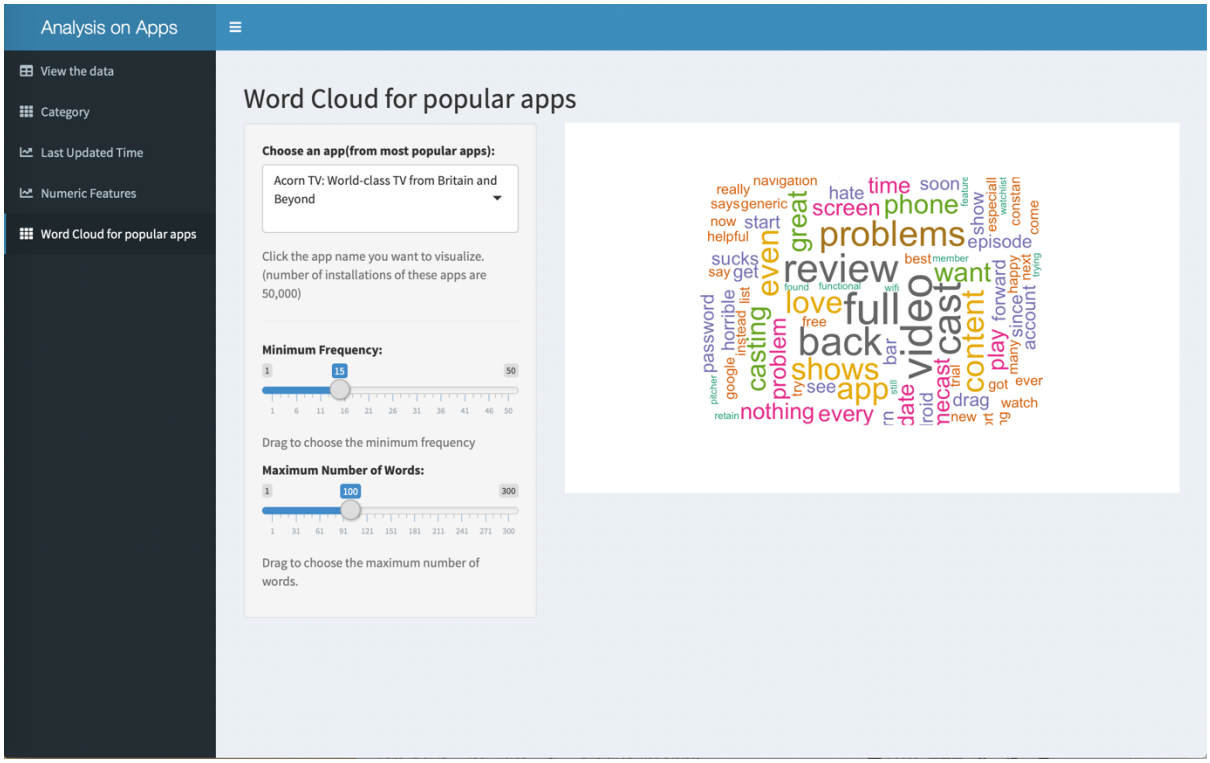


Fig 4.8 Word cloud for popular apps

5. Conclusion

5.1 Summary on this project

This project shows the data table and the relationship plot between different features of apps and the profit. And it can give users clear information for them to design or choose their apps in order to gain more profit.

5.2 Thinking on this project

From the project, I knew what kinds of graphs we should choose for different kinds of data and different purpose. Based on this, I learnt how to design a visualization task to present clearer to users. And also, I knew many tools and libraries to achieve interactive task for users to view graphs. In general, I learnt how to analyze data and plot different graphs to present to users, and then how to combine all analysis and graphs to perform a big visualization task for users.

If I have more time, I will try to use d3 to implement my project and make it clearer and more beautiful. In addition, I want to try different graphs to present the relationships.

5.3 Difficulty

Although I use many libraries to achieve the task, there are many difficult parts to implement. **One is to combine two different and large datasets.** Because there is no installation information in review dataset, I should combine them to choose the top apps with largest installations to plot word cloud. **Another is to combine different parts of graphs into one page using dashboard.** For this part, I implement each part of analysis and add one by one to the final dashboard, which is difficult not only to implement each part, but also combine them together.

6. Appropriate references and bibliography

Plot.ly. (2019). *plotly*. [online] Available at: https://plot.ly/r/ [Accessed 9 Jun. 2019].

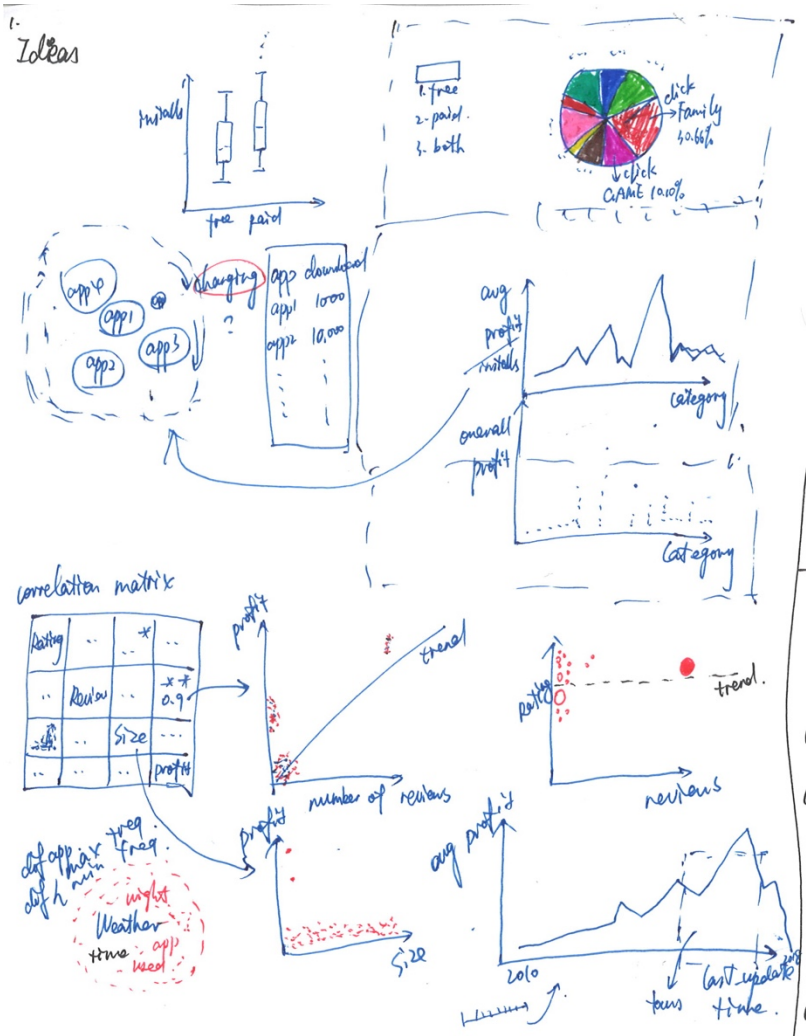
Rstudio.github.io. (2019). *Shiny Dashboard*. [online] Available at: https://rstudio.github.io/shinydashboard/ [Accessed 9 Jun. 2019].

Shiny.rstudio.com. (2019). *Shiny - How to use DataTables in a Shiny App*. [online] Available at: https://shiny.rstudio.com/articles/datatables.html [Accessed 9 Jun. 2019].

Shiny.rstudio.com. (2019). *Shiny - Word cloud*. [online] Available at: https://shiny.rstudio.com/gallery/word-cloud.html [Accessed 9 Jun. 2019].

7. Appendix (five design sheet)

1. Ideas



2. Filter

(1) ~~bar~~ boxplot ~~chart~~ : paid, free.
(2) category.
(3) correlation matrix.
(4) reviews, rating, size,
(5) word cloud
(6) last_update.
(7) bubble plot

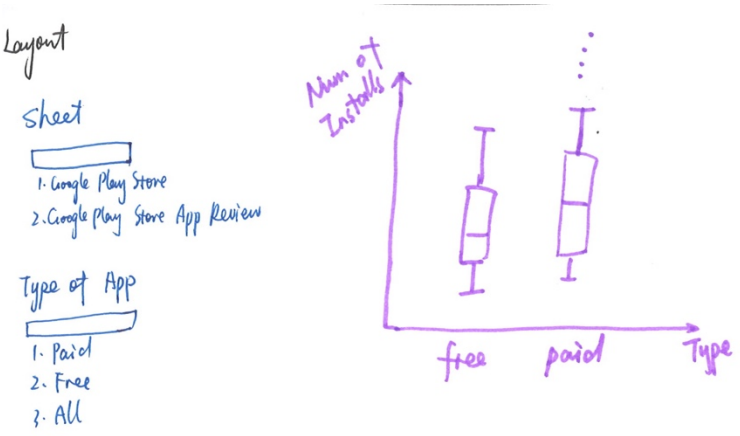
3. Categorize

(1) boxplot: paid, free.
(2) category: pie chart + line.
(3) correlation matrix + numeric features. + word cloud, bubble plot
(4) last_update.
(5) data table

4. Combine and Refine

(1) boxplot: paid, free ; data table
(2) category: ~~filter~~ choose paid or free.
~~bar~~ pie chart + line chart.
(3) correlation matrix: reviews, rating, word cloud + bubble plot,
(4) last_updated time, hours ...
~~(5) data table~~

Layout



Title: Analysis on Google Play Store Apps.

Auther: Yaling Xiong.

Date: May 28, 2019

Sheet: sheet 2

Task: Overview of app's installation.

Operations

1. View the original data
2. See the difference of paid apps and free apps.
3. ~~P~~portion select part of Last_updated time to check the distribution.

Focus

1. sheet ▭ : can choose which table to view.
   Type ▭ : can choose which type of data to view (paid, free)
   show 100 entries: can choose how many records to view.
2. Type ▭ : choose which type of data to view ~~because~~ (because paid apps ~~can~~ should use ~~the~~ profit and free apps should use number of installs)

~~Retart~~

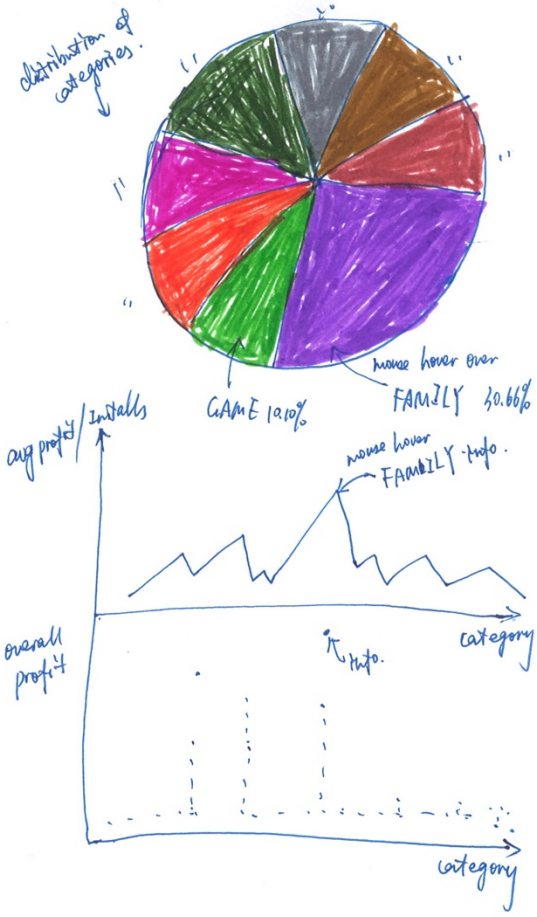
Discussion.

ad: easy to implement.
give a chance to see the original table.

Layout

Type of Apps

1. free
2. paid
3. both



Title: Analysis on Google Play Store Apps.

Author: Yaling Xiong.

Date: May 28, 2019.

sheet: sheet 3

Task: ~~Overview of app~~.
Analysis on category feature.

Operations

1. can view the distribution of each category in each type of apps.
2. can view the profit of each category.
3. comparing pie chart with line chart, can find the relationship.

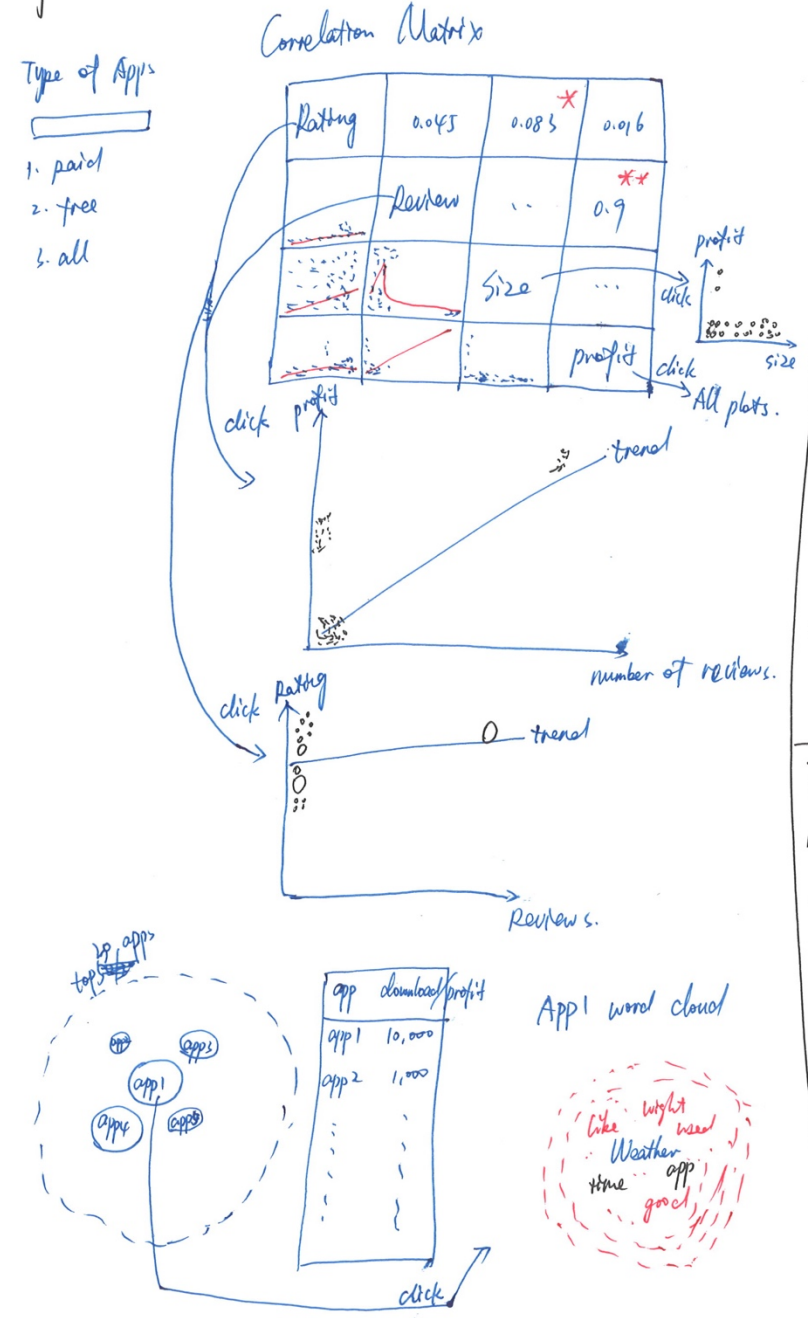
Focus

1. Type ▭: can choose the type of apps [paid of free].
2. ~~when~~ when mouse hover over the pie chart, it can get the name of selected category and its percentage and count.
3. when mouse hover over the line chart or scatterplot, it can get the information of that point, such as "category, overall profit, average profit ....".

Discussion

ad: easy to find the ~~distribut~~ distribution of category feature.
can explore the relationship between category and profit.

Layout



Title: Analysis on Google play Store Apps.

Author: Yaling Kong

Date: May 28, 2019.

Sheet: sheet 4

Task: Analysis on numeric features.

Operations

1. can view the relationship between profit and other numeric features.
2. can view the word cloud of review for ~~the~~ top 50 apps.

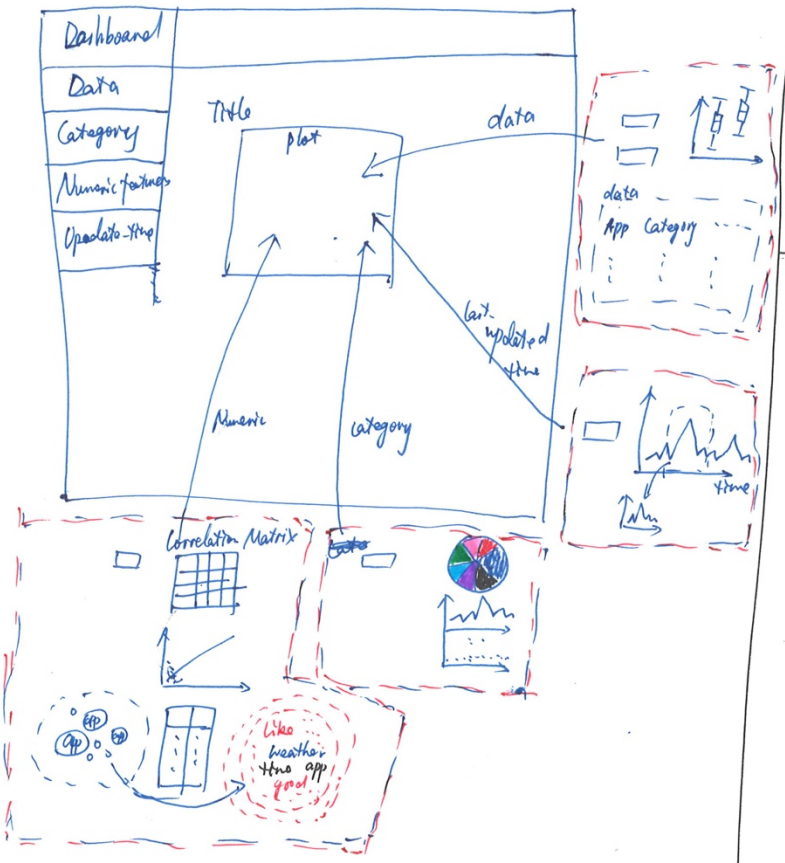
Focus.

1. Type ▭ : choose paid or free or all.
2. click on feature of correlation matrix, can plot the relationship between profit and this feature.
3. For bubble plot, click on one app, there will be word cloud on this app to see the review of this app.

Discussion:

ad: can explore the relationship between numeric features and profit.

Layout



Title: Analysis on Google Play Store Apps.

Author: Yating Xiong

Date: May 28, 2019

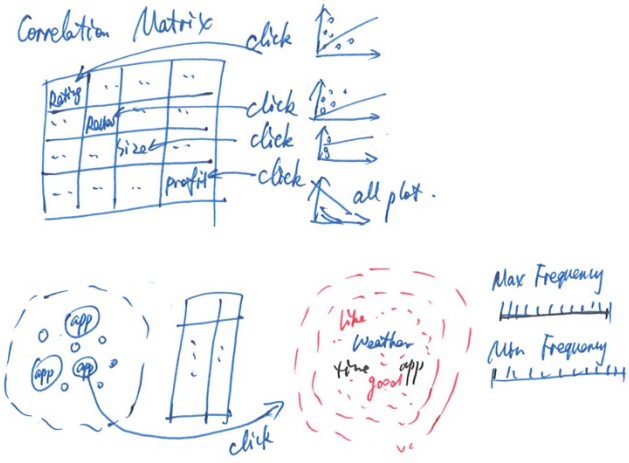
Sheet: sheet 5

Task: Overall View.

Operations/software.

- shiny
- ggplot
- plotly
- DT
- RColorBrewer
- boxplot, line chart, scatter plot, pie chart, bubble plot, word cloud.

Focus



~~Discussion~~.

Detail/Requirements.

1. Enough Data.
2. Data Cleaning, Data Reshaping.
3. Fast Computation.
4. Effective tools.
5. Database.