# Open vs. closed innovation: using online network data to measure innovation

*Benjamin Snow and Oliver Bott*

*14 November 2014*

## 1 Background

This research project examines the potential preferability of using collaborative online network data on the city-level as an innovation indicator. By doing so, this work critically assesses current innovation indicators in the hope of offering new alternatives for measuring and understanding innovation. Since various scholars, called for the continuous improvement of innovation measurement (see for example Freeman and Soete 2009), this work seeks to go beyond the widespread use of patent data to contribute to the refinement of innovation indicators, and the field as a whole.

For context, the stated research question is: *To what extent can open innovation network data add to the measurement of innovation performance?*

For the full examination of the previous literature on the subject and reasoning for the purpose and motives and plan for this study, please see _____. This contribution focuses specifically on outlining the data gathering, data cleaning, and data merging process. It will also contain an explanation for the construction of early descriptive and inferential statistical models.

## 2 Data Gathering

To examine open network data against patent data, this study will relies on two data sets and uses the statistical tool R (R Core Team 2014) for the data analysis.

The first data set was obtained by using the Application Programming Interface (API) data for open networks. To examine open data innovation, data is obtained from the the git repository web-based hosting service GitHub[1]. The R (R Core Team 2014) packages Wickham (2014), Wickham and Francois (2014) and Couture-Beil (2014) allow for compiling data on the user counts and locations associated with different repositories. For closed innovation we use city-level patent data, taken from the Organization for Economic Co-operation and Development[2]. The R (R Core Team 2014) package (Blondel 2014) is necessary for obtaining the OECD dataset. Patent Cooperation Treaty (PCT) patent data are used to track internationally patented inventions. Patent data and and GitHub follower data was taken for __ cities, which were matched for direct comparison.

As we are interested in patent data, we work with data indicating the PCT patent applications per 10,000 inhabitants. From the same database, we also use GDP per capita data and environmental data, as other variables which could prove significant in explaining differences in innovation. The OECD data will be in a format that allows us to locate patent activity to individual cities.

lines (230-345 of other document)

For closed innovation we use city-level patent data, taken from the Organization for Economic Co-operation and Development[3]. The R (R Core Team 2014) package (Blondel 2014) is necessary for obtaining the OECD dataset. Patent Cooperation Treaty (PCT) patent data are used to track internationally patented inventions. For our sample, we took patent data of 133 cities from various OECD countries. As we are interested in patent data, we work with data indicating the PCT patent applications per 10,000 inhabitants. From the

[1] Online accessilbe on https://github.com/.

[2] Online accessible on http://stats.oecd.org.

[3] Online accessible on http://stats.oecd.org.

same database, we also use data on GDP per capita, pollution levels as a proxy for industrial development, greenspace as a proxy for the recreational value, employment data (as share of national employment) as a proxy for the relative national importance of the city. These other variables are inclued in the analysis as they could prove significant in explaining differences in innovation.

# 3 Justification

A large aspect of data gathering and analysis is determining and clearly explaining reasoning behind the independent variables chosen in constructing a model. For the purpose of this study, elements chosen to include in the model focused on the general economic health of the cities chosen (GDP), the industrialization level of the city (Pollution), the amount of publically usable parkland (Greenspace),

**Cities**

Why are we using the cities we are, why did we exclude Einhoven etc.

**Pollution**

Pollution data was obtained as a broad proxy for industrialization

**GDP**

Greenspace, employment as share of national employment, pop,

Due to the exploratory nature of the study, it is important to be aware of the limitation before examining results, to put them in their proper context, and to keep in mind possible more approximate ways in the future.

# 4 Limitations

For this analysis, there have been several issues with identifying parts of cities (Regions of cities). Additionally, pulling API data for cities with two word names ‘New York’, ’

The constraints of the data used in this study are numerous, but fall largely within three main categories:

Time:

Too Rough a Proxy:

**Defining Innovation**

**Measuring Innovation**

**Limitations of Patent Data**

| Variables | Year | Source |
|---|---|---|
| Patents | 2008 | OECD |
| GDP | 2008 | OECD |
| Population | 2008 | OECD |
| Greenspace | 2008 | OECD |
| Employment | 2008 | OECD |
| Pollution | 2005 | OECD |
| No Following | 2014 | API |
| 1-24 Following | 2014 | API |
| >25 Following | 2014 | API |

# 5 Methodology

To examine open network data against patent data, this study will rely largely on two data sets and use the statistical tool R (R Core Team 2014) for the data analysis.

### API network data

The first data set is obtained by using the Application Programming Interface (API) data for open networks. To examine open data innovation, data is obtained from the the git repository web-based hosting service GitHub[4]. GitHub is a web-based hosting service used for collaborative research. Its use of source code management makes it a commonly used software development collaboration tool. Since most of the repositories are openly accessible one can use API tools to track the popularity of repositories, measured here by the repository user counts. The R (R Core Team 2014) packages Wickham (2014), Wickham and Francois (2014) and Couture-Beil (2014) allow us to compile data on the user counts and locations associated with different repositories.

For the analysis, first create location vectors for different cities with the `locations` code. Since many GitHub users can be located, we will identify different open innovation clusters, for example Berlin, New York and San Francisco. In addition, we use the `vector()` code to get information on user counts, focusing on repositories with more than 20 or so followers. By combining locations and user counts data by using `data.frame()`, we will be able to construct data sets for different location clusters and user count numbers, where users of highly relevant repositories are located.

### Closed innovation OECD patent data

For closed innovation we use city-level patent data, taken from the Organization for Economic Co-operation and Development[5]. The R (R Core Team 2014) package (Blondel 2014) is necessary for obtaining the OECD dataset. Patent Cooperation Treaty (PCT) patent data are used to track internationally patented inventions. For our preliminary sample, we took patent data of 20 cities overall, ranging from six different countries, including their country level patent data, for general comparison over the time period 2000 until 2012. As we are interested in patent data, we work with data indicating the PCT patent applications per 10,000 inhabitants. From the same database, we also use GDP per capita data and environmental data, as other variables which could prove significant in explaining differences in innovation. The OECD data will be in a format that allows us to locate patent activity to individual cities.

### Statistical Model

On the type of analysis and question, this study plans to use an ordinary least squares (OLS) model by using `plot` and possibly `rcorr( , type="pearson")` to examine the relationship between patent and open data in a given network cluster. If we happen to find a relationship, this would presumably demonstrate that open data shows the same innovation as patent data, but the ‘throughput’ of innovation rather than the ‘output’. Open data having a relationship to patent data would presumably show innovation as a throughput, since it is measured by people finding those contributing in open data as innovators (followers on github), rather than looking at the specific innovation at completion (patents). One limitation we are facing is that the patent data will not be up to date compared to the network data. Still we believe that a general comparison is possible and could lead to valid results.

```
## Loading required package: car

## Warning: package 'car' was built under R version 3.1.2
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
```

[4] Online accessilbe on https://github.com/.

[5] Online accessible on http://stats.oecd.org.

```
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
## Warning: "title" is not a graphical parameter
```

# 6 Data

How our model is built up.

Some interpretation

Some interpretation.

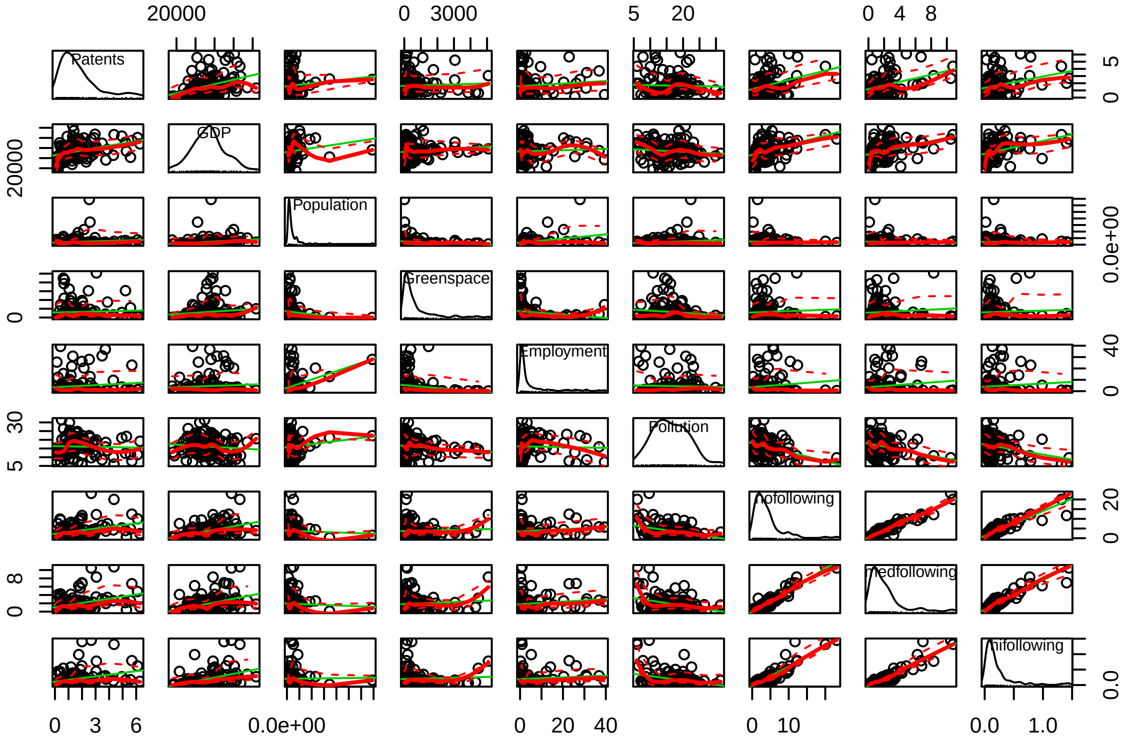

Figure 1: plot of chunk unnamed-chunk-5

Table 2: Regression Estimates of Patent Activity

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | log(Patents) | | | |
| | (1) | (2) | (3) | (4) |
| (Intercept) | 0.32$^{***}$ | | | |
| | (0.08) | | | |
| No Following | | 0.34$^{***}$ | | |
| | | (0.08) | | |
| 1-24 Following | | | 0.34$^{***}$ | 0.15$^{*}$ |
| | | | (0.07) | (0.08) |
| >25 Following | | | | 1.76$^{***}$ |
| | | | | (0.41) |
| GDP | | | | −0.06 |
| | | | | (0.11) |
| Population | | | | 0.05 |
| | | | | (0.05) |
| Greenspace | | | | 0.08 |
| | | | | (0.06) |
| Employment | | | | 0.49$^{**}$ |
| | | | | (0.21) |
| Pollution | −0.13 | 0.10 | 0.89$^{***}$ | −18.73$^{***}$ |
| | (0.11) | (0.08) | (0.17) | (3.92) |
| Observations | 120 | 120 | 120 | 120 |
| $R^2$ | 0.11 | 0.15 | 0.15 | 0.34 |
| Adjusted $R^2$ | 0.10 | 0.14 | 0.15 | 0.31 |
| Residual Std. Error | 0.87 (df = 118) | 0.85 (df = 118) | 0.85 (df = 118) | 0.76 (df = 113) |
| F Statistic | 14.72$^{***}$ (df = 1; 118) | 20.50$^{***}$ (df = 1; 118) | 21.58$^{***}$ (df = 1; 118) | 9.88$^{***}$ (df = 6; 113) |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3: Regression Estimates of Follower Numbers

| | *Dependent variable:* | |
|---|---|---|
| | log(hifollowing) | |
| | (1) | (2) |
| (Intercept) | $0.46^{***}$ | $0.18^{*}$ |
| | (0.10) | (0.10) |
| Patents | | $2.21^{***}$ |
| | | (0.45) |
| GDP | | $-0.19$ |
| | | (0.12) |
| Population | | 0.04 |
| | | (0.05) |
| Greenspace | | 0.09 |
| | | (0.07) |
| Employment | | $-0.65^{***}$ |
| | | (0.23) |
| Pollution | $-2.21^{***}$ | $-21.19^{***}$ |
| | (0.09) | (4.38) |
| Observations | 120 | 120 |
| $R^2$ | 0.15 | 0.39 |
| Adjusted $R^2$ | 0.15 | 0.36 |
| Residual Std. Error | 0.98 (df = 118) | 0.85 (df = 113) |
| F Statistic | $21.58^{***}$ (df = 1; 118) | $12.01^{***}$ (df = 6; 113) |
| *Note:* | | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ |

# References

Blondel, Emmanuel. 2014. *Rsdmx: Tools for Reading SDMX Data and Metadata*. http://CRAN.R-project.org/package=rsdmx.

Couture-Beil, Alex. 2014. *Rjson: JSON for R*. http://CRAN.R-project.org/package=rjson.

Freeman, Christopher, and Luc Soete. 2009. "Developing Science, Technology and Innovation Indicators: What We Can Learn from the Past." *Research Policy* 38 (4). Elsevier: 583–89.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Wickham, Hadley. 2014. *Httr: Tools for Working with URLs and HTTP*. http://CRAN.R-project.org/package=httr.

Wickham, Hadley, and Romain Francois. 2014. *Dplyr: Dplyr: A Grammar of Data Manipulation*. http://CRAN.R-project.org/package=dplyr.