## Intro

Synthpop (<https://www.synthpop.org.uk/>) is a package for the R programming language that us generally used to produce synthetic data for individual-level data. The synthpop package also provides a set of utilities and tools that allow you to compare two arbitrary datasets. These tools are generally used to compare original datasets with the results of creating synthetic data using synthpop, but we will be using them to compare original datasets with new synthetic datasets created using our own algorithms.

## Installation

In order to use synthpop, you must first install the "R" programming language and its synthpop package. You don't need a working knowledge of R, as this guide will walk you through the process step by step, but you do need to have it installed. (If you already have R installed, skip to step "2" below.) **If you need help installing R or the synthpop library, please reach out to** [**maia.hansen@gmail.com**](mailto:maia.hansen@gmail.com) **and schedule time to work through the process together.   
Also: Mac users please see footnote at end of guide.**

1. Install R using the instructions at <https://cran.r-project.org/> If you are using Mac or Windows, you will likely want to use one of the pre-compiled binary packages available in the "Download and Install R" section. (If you are using Linux, you may want to use your Linux package manager instead.)
   1. **OPTIONAL**: Once you have installed R, you can optionally install the RStudio IDE at <https://www.rstudio.com/products/rstudio/download/> . This is **not required** in order to run the synthpop package, but can be useful if you intend to do additional things with R.
2. Install the "devtools" R package.
   1. Open an R console by opening a terminal or command-line window and typing the single letter r
   2. Install the R "devtools" package with the following command:  
       install.packages("devtools")
   3. You will be prompted to select a mirror site near you. R will provide you with a list of available mirror sites. Enter the number of the mirror site that is physically located nearest to you (e.g. "81" for Oregon, "17" for Beijing, etc.)
   4. You should see that R will automatically download and install the devtools package as well as any required dependencies.1
3. Load the "devtools" R package into your R terminal session with the following command:  
    library(devtools)

You should see the following message:  
 > library(devtools)

Loading required package: usethis

1. Next, install at least **version 1.7-0** of the synthpop package. (You will need version 1.7 in order to use the compare function with data that was not generated directly from synthpop.) Assuming you are connected to the internet, you can download this directly from github2 with the following command:

install\_github("bnowok/synthpop")

You should see the following message:

> install\_github("bnowok/synthpop")

Downloading GitHub repo bnowok/synthpop@HEAD...

and the synthpop package will be installed.

If you did not receive any error messages during this process1, you have successfully installed the required synthpop library! Continue to the next step, "Comparing Datasets."

## Basic Dataset Comparison

This section assumes that you have two CSV (comma-separated values) files accessible locally on your filesystem. One should contain a ground truth dataset, and one should contain a dataset synthesized from this ground truth dataset.

This Quick Start guide will focus on performing the comparison using the R command line. For convenience, you may wish to write an R script that will perform the comparison for you, or use RStudio if you are already comfortable with R. That is outside the scope of this guide, but if you'd like to try that and want help, please feel free to contact [maia.hansen@gmail.com](mailto:maia.hansen@gmail.com).

To compare two datasets:

1. If you are not already in an R console, open one by opening a terminal or command-line window and typing the single letter r (followed by <ENTER>)
2. Load the synthpop library into your workspace with the following command:

library(synthpop)

1. Load your ground truth dataset into an R dataframe with the following command:  
    groundTruth <- read.csv(file = '/path/to/ground\_truth.csv')
2. Load your synthetic dataset into an R dataframe with the following command:  
    syntheticData <- read.csv(file = '/path/to/synthetic\_data.csv')
3. Run a basic comparison between the two datasets with the following command:

compare(syntheticData, groundTruth)

This comparison will open up a series of graphs that provide a basic comparison between the two datasets. Keep reading for examples and more tools!

## Example

As an example, we will use the ground\_truth.csv Chicago taxi driver dataset from Sprint 3. This dataset contains the following columns:

taxi\_id,shift,company\_id,pickup\_community\_area,dropoff\_community\_area,payment\_type,trip\_day\_of\_week,trip\_hour\_of\_day,fare,tips,trip\_total,trip\_seconds,trip\_miles

For our comparison dataset, we will use a synthetic dataset (named submission.csv), created as per the requirements of the Sprint 3 competition, containing the following columns:

epsilon,taxi\_id,shift,company\_id,pickup\_community\_area,dropoff\_community\_area,payment\_type,fare,tips,trip\_total,trip\_seconds,trip\_miles

Both the ground\_truth.csv and submission.csv files in this example are located in the current working directory. First, we enter an interactive R session by typing the letter r and hitting <ENTER>. Then, here is the R session that is run to compare these datasets, based on the instructions above3:

> library(synthpop)

Find out more at https://www.synthpop.org.uk/

> groundTruth <- read.csv(file = 'data/ground\_truth.csv')

> syntheticData <- read.csv(file = 'submission.csv')

> compare(syntheticData, groundTruth)

Warning: Variable(s) epsilon in synthetic object but not in observed data

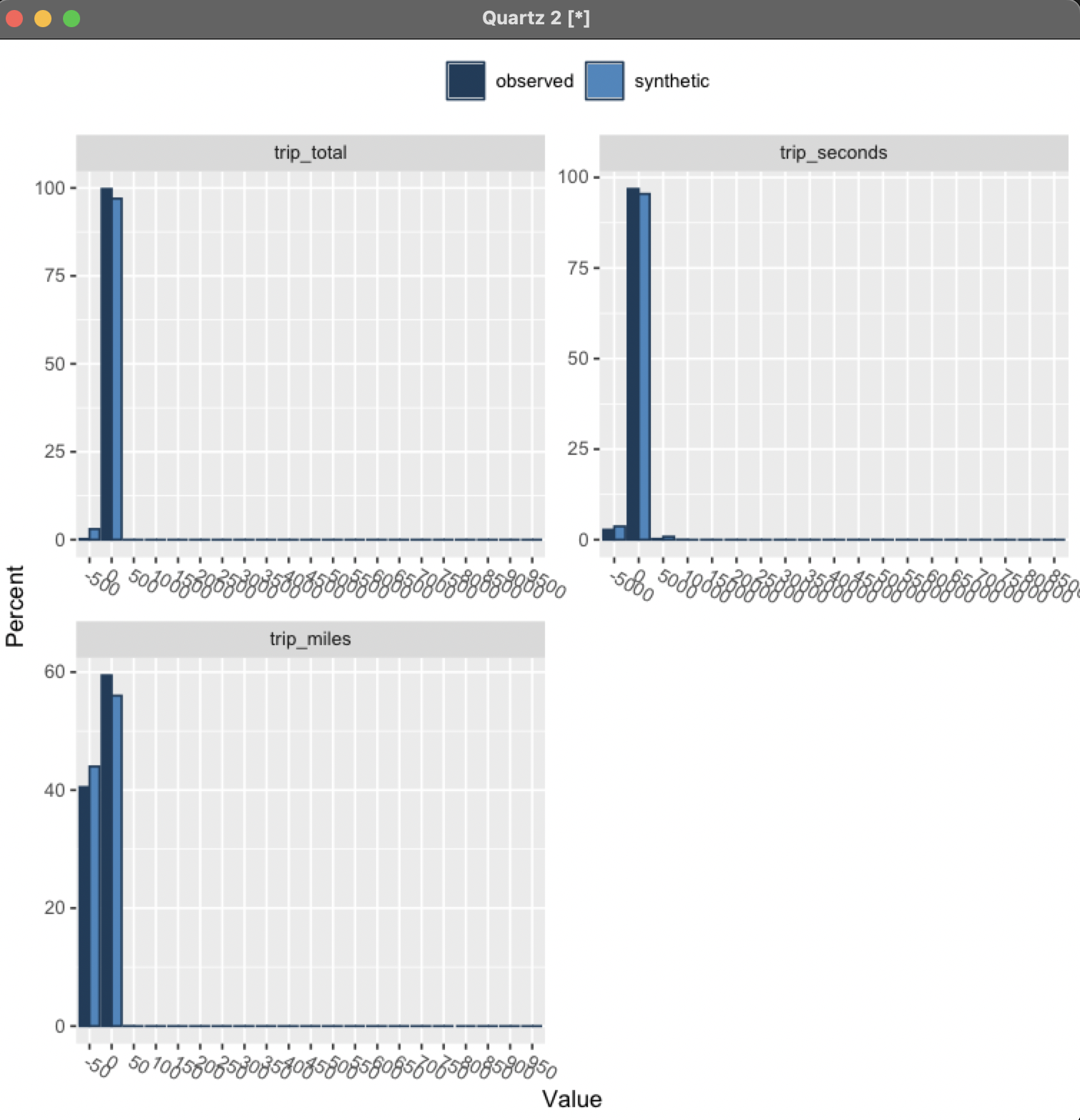
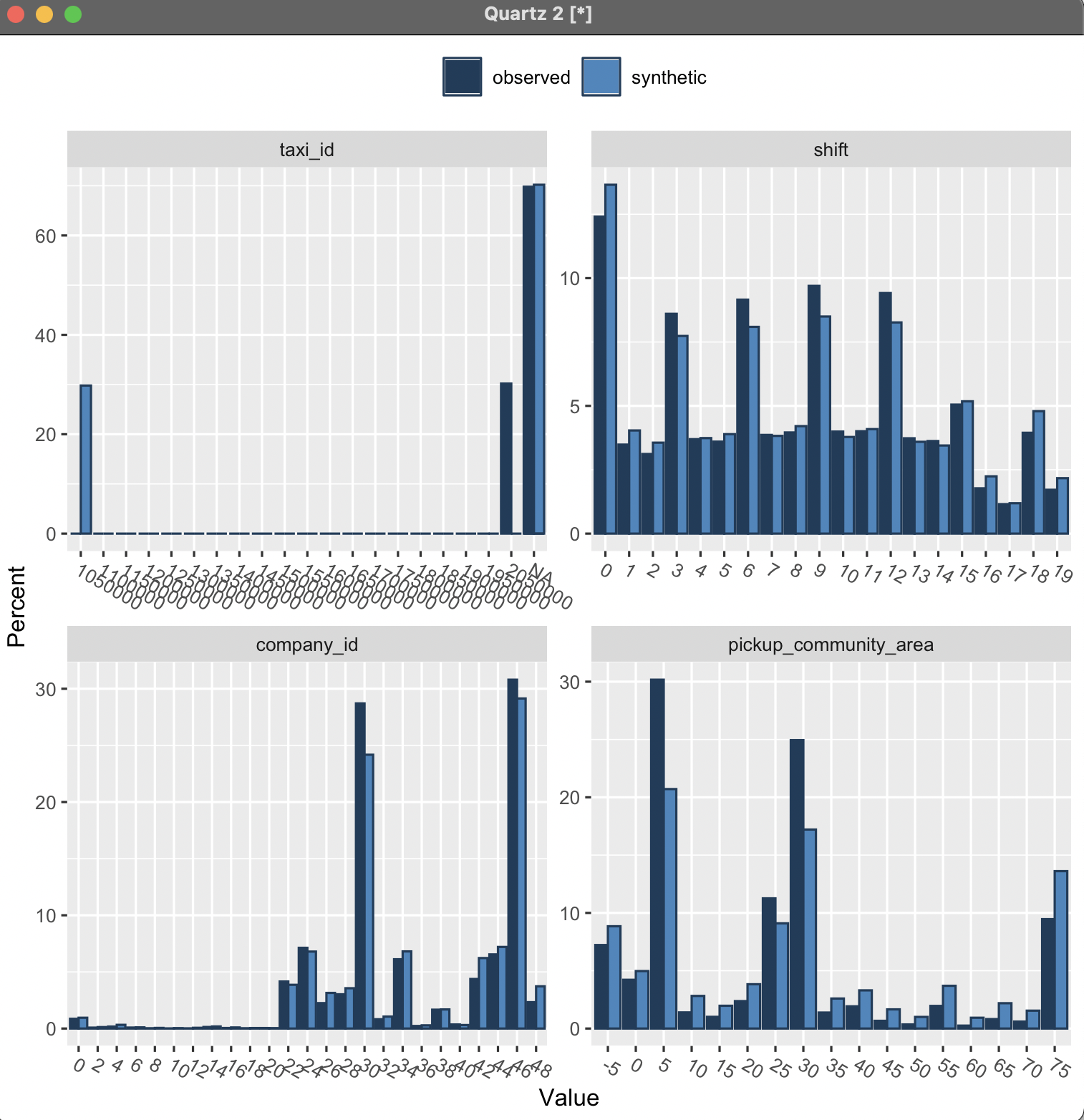
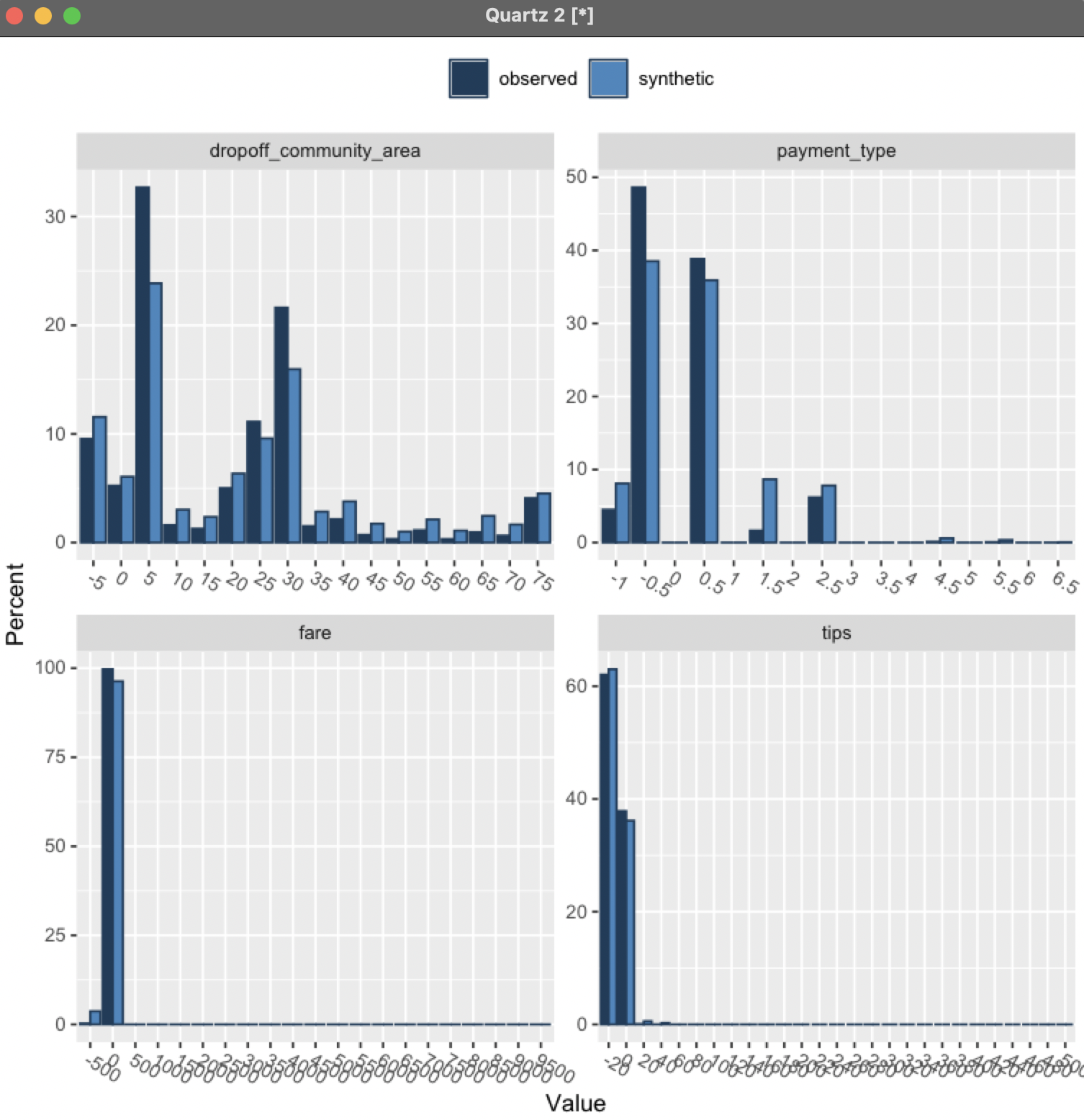
Looks like you might not have the correct data for comparison

Comparing percentages observed with synthetic

Press return for next variable(s):

Press return for next variable(s):

>

When the analysis is complete, graphs such as these will be displayed, comparing these two datasets:

## Other Features

Synthpop also provides methods to perform other comparisons. Some features that might be useful include:

#### Dataset Summaries

To get a summary (min, median, max, etc.) of the values contained in a dataset, type the following command (using the examples loaded as described above):

summary(groundTruth)

Or

summary(syntheticData)

## 

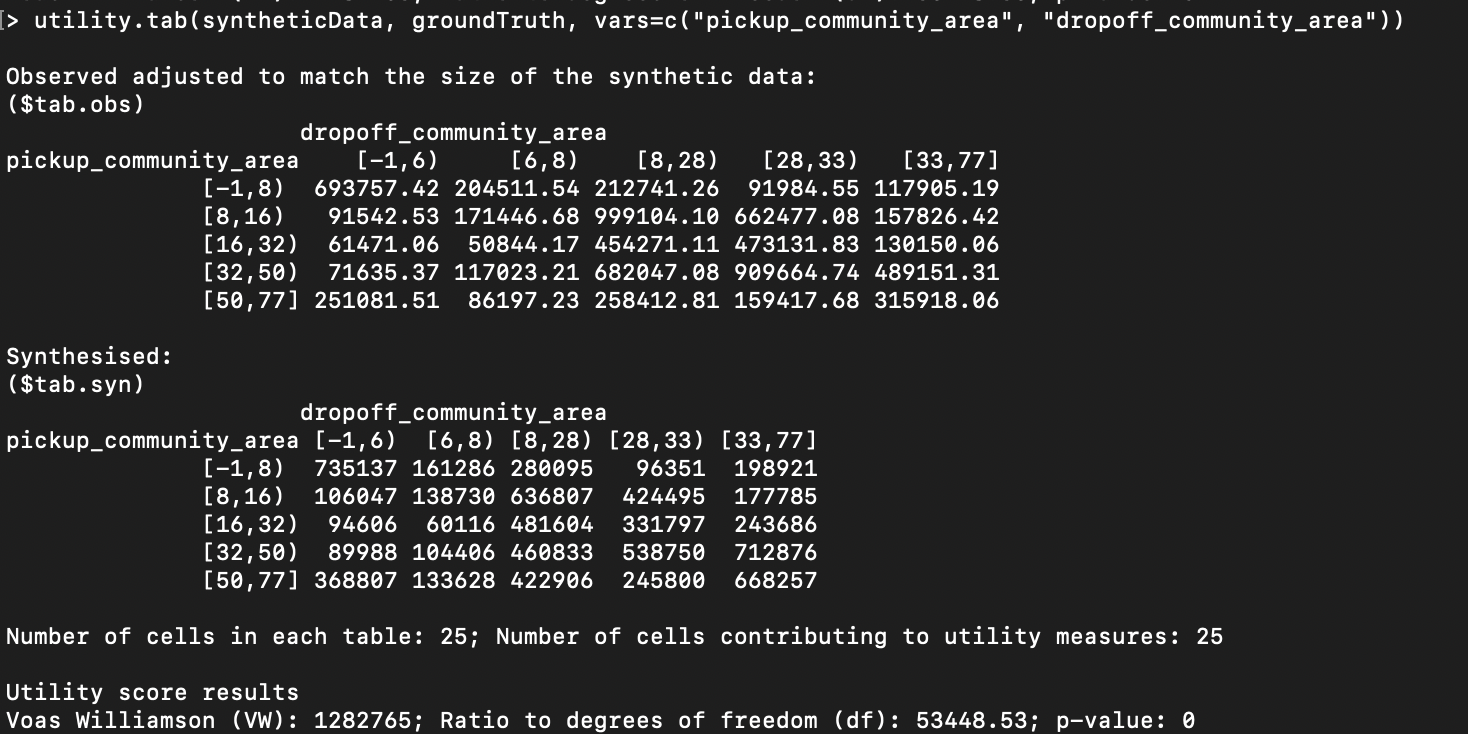
#### Tabular comparisons

Synthpop can also be used to produce and compare tables from observed and synthesised data. This uses the utility.tab function of synthpop, documented at <https://github.com/bnowok/synthpop/blob/master/man/utility.tab.Rd>

As an example, to produce a tabular comparison of the pickup and drop off community areas between the synthesized and ground truth datasets loaded in the example above, you would use the following command:

utility.tab(syntheticData, groundTruth, vars=c("pickup\_community\_area", "dropoff\_community\_area"))

This produces results similar to the following:



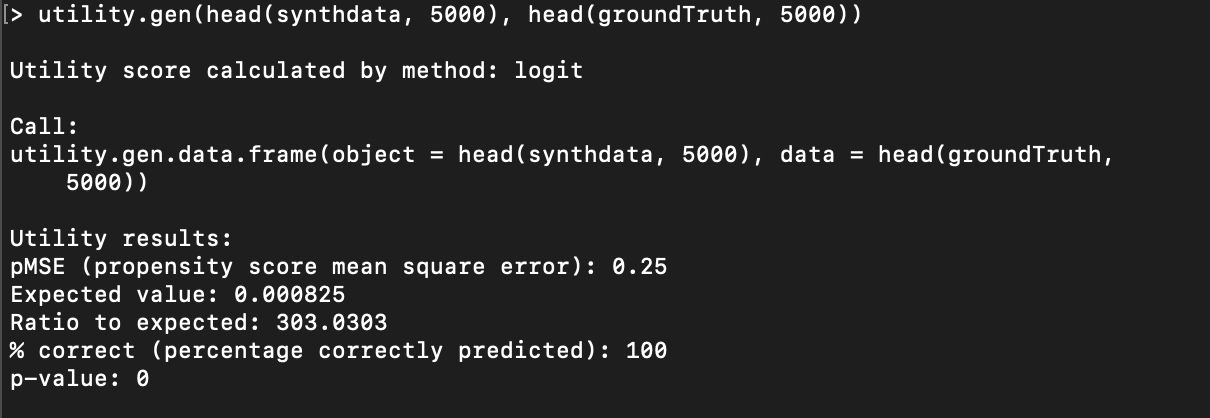
#### Distributional comparisons

You can also use synthpop to do a distributional comparison of synthesized and ground truth data, using the utility.gen function of synthpop, documented at <https://github.com/bnowok/synthpop/blob/master/man/utility.gen.Rd>. In order to use it, the synthetic data must not contain any columns that are not also present in the ground truth data. However, in our example above, the synthetic data contains the column epsilon, which is not present in the original data.

In order to delete the epsilon column from the data loaded as syntheticData in the example above, run the following command (after loading the data in your R console) to create a new synthdata dataframe that does not have epsilon:

synthdata = subset(syntheticData, select = -c(epsilon))

You should now be able to run the utility.gen function to compare the two datasets. However, this function requires a great deal of internal memory, and so you may need to select only the first N rows of each dataset in order to get a result, using the head function. As an example, here is a comparison of the first 5000 rows of our ground truth and synthetic data (after removing the epsilon column), as described in the example above:



Refer to the synthpop documentation for the utility.gen function located at <https://github.com/bnowok/synthpop/blob/master/man/utility.gen.Rd#L120> for more details on how to use and configure the utility.gen function to produce more useful and interesting results.

For more details on synthpop, refer to its github repository located at <https://github.com/bnowok/synthpop>.

You can also access the full documentation for synthpop at

<https://rdrr.io/cran/synthpop/> -- be aware, though, that this documentation is for version 1.6.0, which does NOT support comparisons between arbitrary datasets. (It only supports comparisons between ground truth datasets and synthetic datasets created using synthpop itself.) However, it may still be useful for finding information about basic functionality.

## Python

There does exist at least one Python implementation of synthpop, at <https://hazy.com/blog/2020/01/31/synthpop-for-python>. We have not tried this implementation, as the native R implementation has the latest functionality that supports direct comparisons between arbitrary datasets. However, if you'd like to try this out, please let us know how it goes!

## Footnotes

1. If you are running this on a Mac, you may see a warning message similar to the following:

Warning message:

In doTryCatch(return(expr), name, parentenv, handler) :

unable to load shared object '/Library/Frameworks/R.framework/Resources/modules//R\_X11.so':

dlopen(/Library/Frameworks/R.framework/Resources/modules//R\_X11.so, 6): Library not loaded: /opt/X11/lib/libSM.6.dylib

Referenced from: /Library/Frameworks/R.framework/Versions/4.1/Resources/modules/R\_X11.so

Reason: image not found  
This error message is benign and can be safely ignored.

1. If you are familiar with using R and the CRAN package manager, please note that the version of synthpop that is available directly from the R package manager has *not* yet been updated to 1.7, and so you will need to follow this process of installing from github.
2. Notice that when you run the compare function, you'll receive the following message:

Warning: Variable(s) epsilon in synthetic object but not in observed data

Looks like you might not have the correct data for comparison

This is expected, because the "epsilon" column in the synthesized data does not exist in the ground truth data. The message can be ignored.