HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL
FHO Fachhochschule Ostschweiz

MSE
MASTER OF SCIENCE
IN ENGINEERING

pgsynthdata - A Synthetic Data Generation Tool for PostgreSQL written in Python

Seminar Database Systems

Labian Gashi

Information and Communication Technologies
Software and Systems

Advisor Professor Stefan Keller

HSR - Hochschule für Technik Rapperswil

Spring Semester 2020

Abstract

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Keywords: *synthetic data, artificial data, data anonymization, privacy, open data, databases, data engineering, open source, PostgreSQL*

Acknowledgements

I would like to express my greatest gratitude towards my advisor Prof. Stefan Keller for suggesting me this seminar to me and assisting me throughout the whole journey with his patient guidance and enthusiastic encouragement.

A huge thanks also goes to Johannes Wildermuth from the institute of HSR who helped me by supplying me with the first code scaffold for the tool.

Another special thanks goes to all the authors of the Python libraries and LateX templates that I used in order to complete this project. I would also like to send some appreciation to the GitLab team for having such a nice and neat application lifecycle tool for me to utilize. It provided me with a very comfortable code repository, CI/CD pipeline features and just generally a nice and consistent environment to have my application on.

Contents

Chapter 1

Introduction

1.1 General

The need for creating and generating synthetic or rather, data that is not real, has become a necessary factor in big companies. Some of the main reasons why such data is required to be created can be: privacy limitations, the need for test data, various experiments, research etc.

Synthetic data are generated to meet specific needs or certain conditions that may not be found in the original, real data. This can be useful when designing any type of system because the synthetic data are used as a simulation or as a theoretical value, situation, etc. This allows us to take into account unexpected results and have a basic solution or remedy, if the results prove to be unsatisfactory. Synthetic data are often generated to represent the authentic data and allows a baseline to be set. [1]

Business functions that usually benefit from the creation of synthetic data can be: Machine Learning, Data Mining, Agile Development, Researching etc. There are also several business types that benefit from synthetic data such as: Healthcare Systems, Financial Services, Fraud Detection Systems etc. [2]

At a first glance, synthetic data may seem like ”random made up data”, when in fact, behind that synthetic data, stand various algorithms and generation methods that are used to create such data which must look as realistic as possible.
Synthetic data is used mostly to protect the privacy and confidentiality of a particular set of data. Real data can contain personal information about people or things that a software engineer or researcher is not supposed to know. Therefore, synthetic data does not contain personal information that can fall into contrary with privacy laws or information that you can use to trace back to individuals.

There's often a misconception about Synthetic Data, Anonymized Data and Artificial Data. They serve mostly the same purposes but involve different techniques and results.
Anonymised Data is getting the original data and adding ”noise” to it, encrypting it or even masking it. Ultimately, the anonymised data will correspond to the original data in some form somewhere.
Artificial Data on the other hand, is generated data with no linking or relation to original data apart from the model building.

In this project, we will be discussing the methods and techniques used for the generation of synthetic data and the original data used to generate such data, which means, fully artificial data with relation to the original data but with no clear signs of the original data.

1.2 Goals & Requirements

The main objective of this project is to successfully generate fully synthesized data that is as realistic as possible using the *PostgreSQL* relational database management system and the *Python* programming language in conjunction with some other third-party modules that mostly deal with numbers.
This will be done using various methods and techniques that involve some sort of data anonymization, overcoming data type obstacles, histogram boundaries, respecting data privacy and data confidentiality etc. The tool initially tries to connect to an existing database and then based on the options provided, it will generate the synthesized data into a new desired database from the selected existing database. The process should be very straightforward.
The tool will be a single lightweight Python script that is executed from the shell terminal and which has various arguments that can be passed along with the command execution, with some of them being mandatory and some others optional (mostly related to database configuration and parameters).
It will initially connect to a PostgreSQL database (given with arguments) , read the *pg_class* table in order to obtain the tables of the given/accessed database and then fetch all the results of each table from the *pg_stats* table. To be continued...

1.3 Synthetic Data Generation

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem.

Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

1.4 Results

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Chapter 2

Overview of Synthetic Data Generation

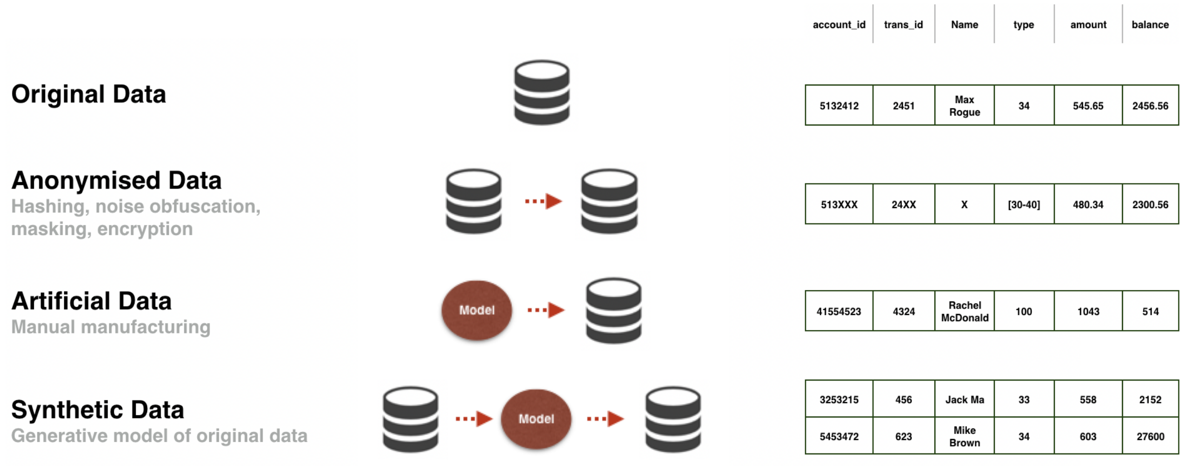
2.1 Overview

In more broad and accepted terms, synthetic data is ”any production data applicable to a given situation that are not obtained by direct measurement” according to the McGraw-Hill Dictionary of Scientific and Technical Terms.

Synthetic data is not only popular and useful in computer science but it is also important in other business functions/types such as: Healthcare Systems, Fraud Detection Systems etc. In this project, the data synthesis of various real-life data stored into databases is important. In this type of data synthesis, that data is initially analyzed and then transformed, or rather just used in order to generate another set of data, so it isn’t really a transformation but rather just a mere ”sample” or ”pattern” for the synthesized data.

Various methods and techniques are used in order to synthesize data. The procedure of this examination, analysis and generation of the real data into synthetic data is explained thoroughly and in more detailed and technical terms in the Implementation chapter.

Synthetic data does NOT refer to anonymized data, it is often a misconception or a myth, there are various types of data transformations that can happen from real data such as: Anonymized data, artificial data, synthesized data etc. Moreover, there are also subtypes of the types mentioned above, which are of course, more detailed and technical. The image below by Synthesized, depicts this in a pretty neat way.



account_id	trans_id	Name	type	amount	balance
5132412	2451	Max Rogue	34	545.65	2456.56
513XXX	24XX	X	[30-40]	480.34	2300.56
41554523	4324	Rachel McDonald	100	1043	514
3253215	456	Jack Ma	33	558	2152
5453472	623	Mike Brown	34	603	27600

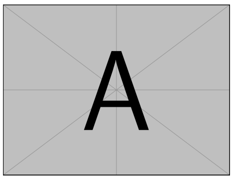
FIGURE 2.1: Three approaches to synthetic data, from Synthesized

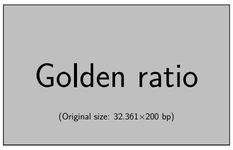
2.2 Data Synthesization

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

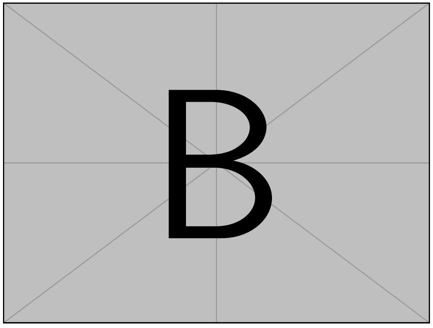




2.3 Data Generation

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

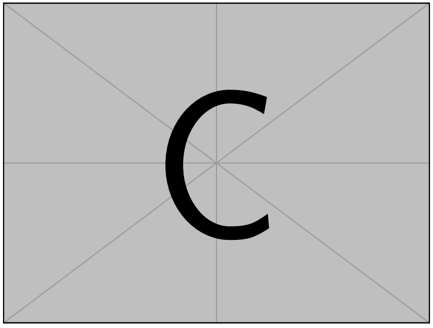


2.4 Selected Database Related Open Source Tools

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem.

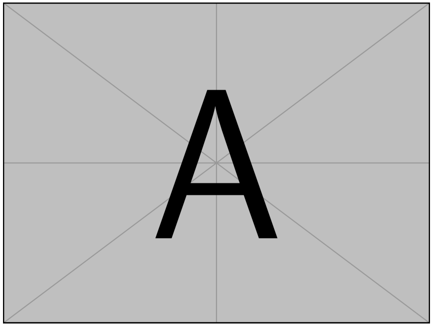
Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.



2.5 Remarks and Further Resources

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus. 

Chapter 3

Implementation

3.1 Overview

The tool is a single lightweight Python script, executable from the shell terminal. The tool first connects to an existing PostgreSQL database (based on the passed arguments) and then based on whether `-show` or `-generate` arguments were passed, with `-show` being the default one. The tool also accepts database connection parameters (all optional) such as `-H, -P, -U` for connecting to the desired database server.
The `-show` argument (which is the default one) will show the configuration of the database and the `-generate` argument generates the synthesized data to the database with the name `DBNAMEGEN`, which is one of the required arguments for this tool.
To be continued...

3.2 Design and Implementation

This tool is written and coded using the Python programming language.
It uses various Python libraries that are used for: argument parsing, various mathematical functions, making the connection to the PostgreSQL server easier etc.
It is a single python script that is to be executed using the shell terminal. It requires some arguments/parameters initially in order to evaluate the PostgreSQL server it needs to connect to, what database it relies on for generating the synthetic data and also the database where the synthetic data will be generated on.
To be continued...

3.3 Tool Usage

The tool is very straightforward to use. It contains a single Python script, with all the logic of connecting to the database and showing/generating the data in a single script. It is a script, designed to be executed from the terminal shell using the python command.
The tool is somehow split into two parts. The first part is the database connection part, which, based on the arguments given for the database and the PostgreSQL connection parameters, tries to connect to that database instance in the server. If everything there is successful, the tool then continues with the data synthesis part.

Tool usage:
`pgsynthdata [OPTIONS]... DBNAMEIN [DBNAMEGEN]`

Tool options:

- `DBNAMEGEN` - Name of the database to be created
- `-show/--show` - Shows config (default)
- `-generate/--generate` - Generates new synthesized data to database *DBNAMEGEN*
- `-O/--owner` - Owner of new database (default: same as user)
- `-v/--version` - Show version information, then quit
- `-h/--help` - Show tool help, then quit

Connection options:

- `DBNAMEIN` - Name of the existing database to connect to
- `-H/--hostname` - Name of the PostgreSQL server (default: *localhost*)
- `-P/--port` - Port of the PostgreSQL server (default: *5432*)
- `-U/--user` - PostgreSQL server username

Some usage examples:

- `python pgsynthdata.py test postgres -show`
  - Connects to database *test*, host=*localhost*, port=*5432*, default user with password *postgres*

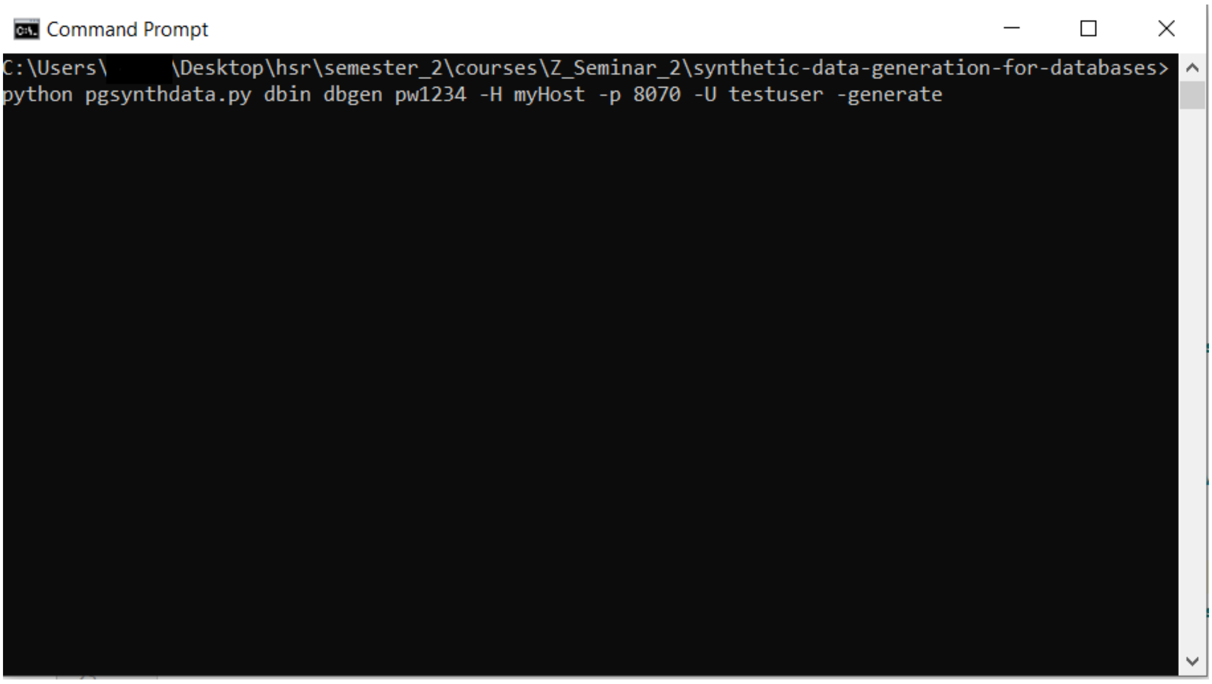
  - Shows statistics from certain tables in database *test*

- `python pgsynthdata.py db pw1234 -H myHost -p 8070 -U testuser -show`
  - Connects to database *db*, host=*myHost*, port=*8070*, user=*testuser* with password *pw1234*
  - Shows statistics from certain tables in database *db*

- `python pgsynthdata.py dbin dbgen pw1234 -H myHost -p 8070 -U testuser -generate`
  - Connects to database *dbin*, host=*myHost*, port=*8070*, user=*testuser* with password *pw1234*
  - Creates new database *dbgen* with synthesized data

- `python pgsynthdata.py --version`
  - Show the version of this tool and then quit

It also uses all the other default PostgreSQL server settings when creating the new database such as: encoding, locale, collation, database template etc.



```
C:\Users\[redacted]\Desktop\hsr\semester_2\courses\Z_Seminar_2\synthetic-data-generation-for-databases>
python pgsynthdata.py dbin dbgen pw1234 -H myHost -p 8070 -U testuser -generate
```

FIGURE 3.1: Tool usage example

Chapter 4

Tool Evaluation

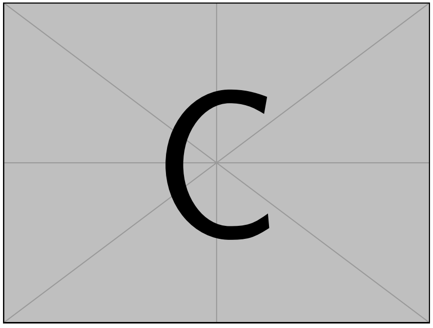
4.1 Overview/Approach

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

4.2 Evaluation

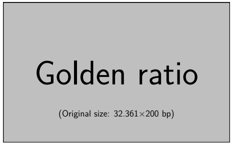
Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

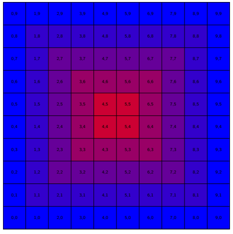
Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.  Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

4.3 Results

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris.





Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor. Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Chapter 5

Conclusion

5.1 Discussion of the Results

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

5.2 Achievements and Reflection

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

5.3 Conclusion

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

List of Figures

Appendix A

Installation

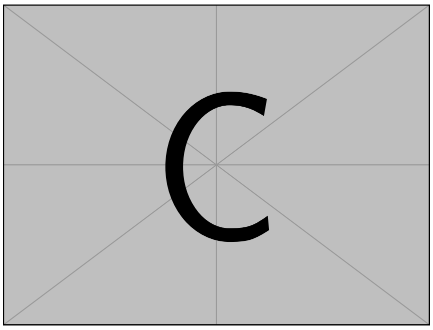
A.1 Versions

Used versions:

- Python 3.7.4
- PostgreSQL Server 11.5
- Vestibulum auctor dapibus neque.
- Nunc dignissim risus id metus.
- Cras ornare tristique elit.

A.2 Installing & Configuring the Tool

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.


C

Appendix B

Usage of Tool

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Appendix C

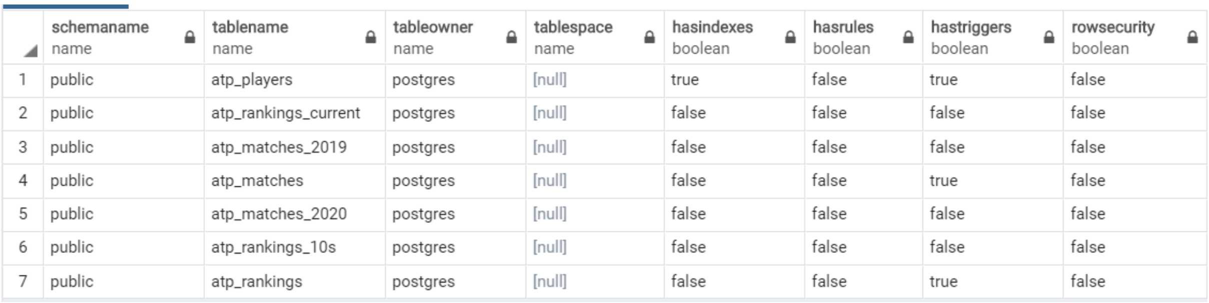
Test Datasets

For data synthesis, good and realistic data is necessary. For this project, many test datasets were used in order to implement, test and evaluate the tool. In this appendix, you will find some of the most important ones and some information/examples of them.

C.1 Tennis ATP

Tennis ATP is a dataset consisting of realistic data of the Tennis ATP (Association of Tennis Professionals). It includes rankings of players, data about the players themselves, data about the matches played between those players etc.

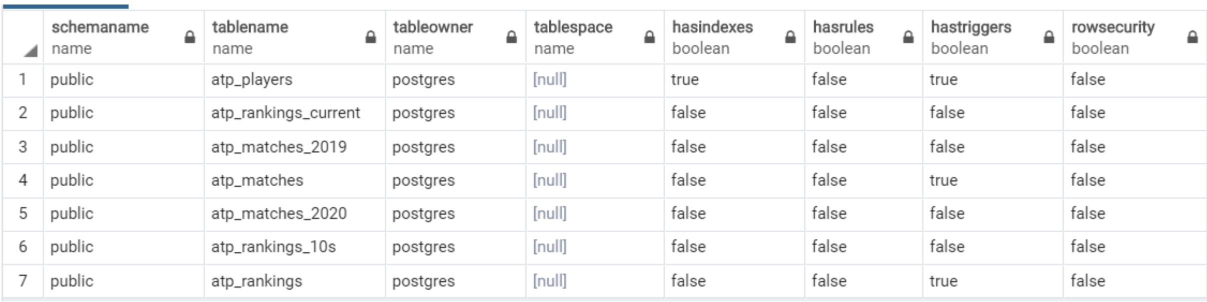
The repository for the complete data set can be found here: https://github.com/JeffSackmann/tennis_atp. The complete dataset found in the repository however, is not entirely used in this tool, only some chunks of it. The data set chunks used for this tool have been chosen and provided by my advisor: Prof. Stefan Keller.



	schemaname	tablename	tableowner	tablespace	hasindexes	hasrules	hastriggers	rowsecurity
	name	name	name	name	boolean	boolean	boolean	boolean
1	public	atp_players	postgres	[null]	true	false	true	false
2	public	atp_rankings_current	postgres	[null]	false	false	false	false
3	public	atp_matches_2019	postgres	[null]	false	false	false	false
4	public	atp_matches	postgres	[null]	false	false	true	false
5	public	atp_matches_2020	postgres	[null]	false	false	false	false
6	public	atp_rankings_10s	postgres	[null]	false	false	false	false
7	public	atp_rankings	postgres	[null]	false	false	true	false

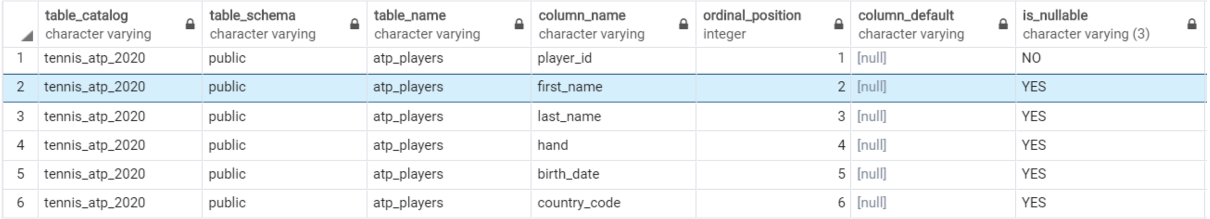
FIGURE C.1: Tennis ATP database schema

The columns of the tables are realistic and make sense. They also make sense in the context of synthesizing them.



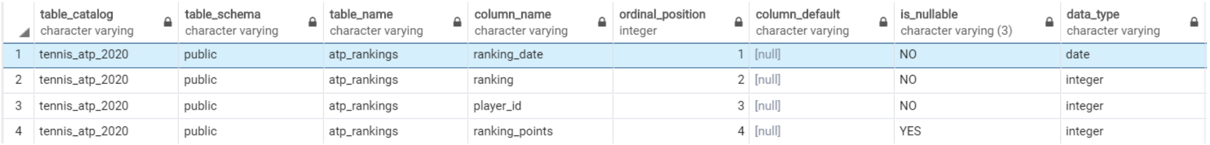
	schemaname name	tablename name	tableowner name	tablespace name	hasindexes boolean	hasrules boolean	hastriggers boolean	rowsecurity boolean
1	public	atp_players	postgres	[null]	true	false	true	false
2	public	atp_rankings_current	postgres	[null]	false	false	false	false
3	public	atp_matches_2019	postgres	[null]	false	false	false	false
4	public	atp_matches	postgres	[null]	false	false	true	false
5	public	atp_matches_2020	postgres	[null]	false	false	false	false
6	public	atp_rankings_10s	postgres	[null]	false	false	false	false
7	public	atp_rankings	postgres	[null]	false	false	true	false

FIGURE C.2: Tennis ATP database schema



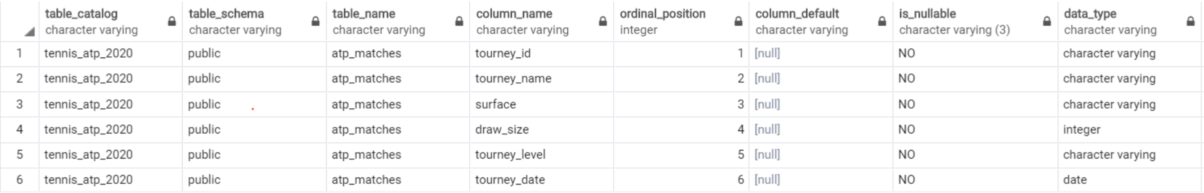
	table_catalog character varying	table_schema character varying	table_name character varying	column_name character varying	ordinal_position integer	column_default character varying	is_nullable character varying (3)
1	tennis_atp_2020	public	atp_players	player_id	1	[null]	NO
2	tennis_atp_2020	public	atp_players	first_name	2	[null]	YES
3	tennis_atp_2020	public	atp_players	last_name	3	[null]	YES
4	tennis_atp_2020	public	atp_players	hand	4	[null]	YES
5	tennis_atp_2020	public	atp_players	birth_date	5	[null]	YES
6	tennis_atp_2020	public	atp_players	country_code	6	[null]	YES

FIGURE C.3: Tennis ATP ”atp_players” table schema



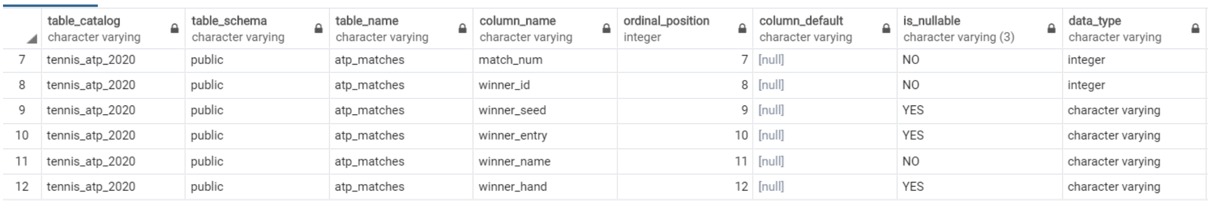
	table_catalog character varying	table_schema character varying	table_name character varying	column_name character varying	ordinal_position integer	column_default character varying	is_nullable character varying (3)	data_type character varying
1	tennis_atp_2020	public	atp_rankings	ranking_date	1	[null]	NO	date
2	tennis_atp_2020	public	atp_rankings	ranking	2	[null]	NO	integer
3	tennis_atp_2020	public	atp_rankings	player_id	3	[null]	NO	integer
4	tennis_atp_2020	public	atp_rankings	ranking_points	4	[null]	YES	integer

FIGURE C.4: Tennis ATP ”atp_rankings” table schema



	table_catalog character varying	table_schema character varying	table_name character varying	column_name character varying	ordinal_position integer	column_default character varying	is_nullable character varying (3)	data_type character varying
1	tennis_atp_2020	public	atp_matches	tourney_id	1	[null]	NO	character varying
2	tennis_atp_2020	public	atp_matches	tourney_name	2	[null]	NO	character varying
3	tennis_atp_2020	public	atp_matches	surface	3	[null]	NO	character varying
4	tennis_atp_2020	public	atp_matches	draw_size	4	[null]	NO	integer
5	tennis_atp_2020	public	atp_matches	tourney_level	5	[null]	NO	character varying
6	tennis_atp_2020	public	atp_matches	tourney_date	6	[null]	NO	date

FIGURE C.5: Some information from the Tennis ATP ”atp_matches” table schema #1



	table_catalog character varying	table_schema character varying	table_name character varying	column_name character varying	ordinal_position integer	column_default character varying	is_nullable character varying (3)	data_type character varying
7	tennis_atp_2020	public	atp_matches	match_num	7	[null]	NO	integer
8	tennis_atp_2020	public	atp_matches	winner_id	8	[null]	NO	integer
9	tennis_atp_2020	public	atp_matches	winner_seed	9	[null]	YES	character varying
10	tennis_atp_2020	public	atp_matches	winner_entry	10	[null]	YES	character varying
11	tennis_atp_2020	public	atp_matches	winner_name	11	[null]	NO	character varying
12	tennis_atp_2020	public	atp_matches	winner_hand	12	[null]	YES	character varying

FIGURE C.6: Some information from the Tennis ATP ”atp_matches” table schema #2

C.2 IMDB

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

C.3 New York Taxi

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Bibliography

[1] H.; Jonsson E. Barse, E.L.; Kvarnström. *Synthesizing test data for fraud detection systems*. IEEE, 2003.

[2] AIMultiple. The importance of synthetic data. 2020. URL https://blog.aimultiple.com/synthetic-data/. NOTE = [Online; accessed April 22, 2020].

[3] Synthesized. Three common misconceptions about synthetic and anonymised data, 2018. NOTE = [Online; accessed April 22, 2020].