**Data Description**

**For some of the data, there are readme files for details, please check it in the data folders.**

***These data are only used for the Final Project of course STA217, DO NOT share or use in other purposes.***

**statistics\_papers**

10 years of papers published in the field of statistics from 1951 to 1960. Data includes the paper’s sub-fields, publication years, each paper’s reference paper list (i.e., the citation network).

**nobel**

Nobel award information since 1901.

**hongloumeng**

a 376-by-475 matrix for character-event appearance in A Dream of Red Mansion (Xueqin Cao)

Credit: Wan Mengting and Yao Yuan, Crowdsourced at Peking University

**xiyouji**

A Character-Event Appearance matrix for the Journey to the West (Xi You Ji), written by WU Cheng’En. Data is stored as R dataframe, you need to use packages like Pyreadr to read it in Python.

Credit: Yao Yuan, Crowdsourced at Peking University

**yelp**

Yelp app data. Full review text data including the user\_id that wrote the review and the business\_id the review is written for.

**covid19 data**

You need to collect the data by yourself.

There are many sources, such as https://ourworldindata.org/coronavirus-source-data

**shared\_room**

Shared rent data parsed from Danke Apartment (蛋壳公寓) on Oct. 2019, columns are月租金（rent）、卧室数（bedroom）、厅数（livingroom）、卫生间数（bathroom）、租赁面积（area）、租赁房间（room）、所在楼层分组（floor\_grp）、邻近地铁（subway）、城区（region）、供暖方式（heating）。

**house\_price\_BJ**

Secondhand house price in Beijing. Data collected on May 2016. Columns include 单位面积房价（price）、城区（CATE）、卧室数（bedrooms）、厅数（halls）、房屋面积（AREA）、楼层（floor）、是否临近地铁（subway）、是否是学区房（school）

**house\_price\_GD**

Secondhand house price in Guangzhou, Shenzhen and Xiamen.

**media\_spreading**

The comments data from weibo, douban and zhihu for the Variety《乘风破浪的姐姐》.

**bilibili**

Data contains Bilibili-ER （UP主）information. Include ID, number of followers, gender, number of published videos, number of representative videos, number of albums, articles, and channels, average length of recent videos, average plays, associated third part platforms, self tags, etc.

**Mathematician**

a database of more than 8,500 famous mathematicians from Wikipedia.

**Netflix movie**

Netflix movie rating data.

**Job**

Job information of data scientists.

**LinkedIn Job Postings**

Readme file included.

**air quality**

hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites

**artists\_directors**

Artists’ works, and each work’s price. Director’s works and each work’s IMDB rating.

**statistics\_big4\_citation\_network**

Papers published in JASA, Biometrika, J. Roy. Stat. Soc. Series B and Annals of Statistics, and their citation networks.

**Input\_output**

The World Input-Output Database (WIOD) November 2016 Release consists of a series of databases and covers 28 EU countries and 15 other major countries in the world for the period from 2000 to 2014.

**Airport\_airline\_route**

See readme in fine.

**Diabetes Indicators**

A dataset of 236,378 survey responses to the Centers for Disease Control and Prevention (CDC)'s Behavioral Risk Factor Surveillance System (BRFSS) 2021. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables.

The Behavioral Risk Factor Surveillance System (BRFSS) is an ongoing, state-based telephone survey that collects data about health-related risk behaviors, chronic health conditions, and the use of preventive services among adults aged 18 years and older residing in the United States. Conducted annually by the Centers for Disease Control and Prevention (CDC), the BRFSS has been providing valuable insights into the health status and behaviors of U.S. adults since its inception in 1984.

**Stack Overflow 2023 survey dataset**

In May 2023 over 90,000 developers responded to the Stack Overflow annual survey about how they learn and level up, which tools they're using, and which ones they want.

The dataset is a collection of two CSV files and the original survey questionnaire as described below:

survey\_results\_schema.csv - contains 78 rows and 6 columns covering the basic schema of the survey.

survey\_results\_public.csv- contains 89184 rows and 84 columns covering different questions related to users and tools they want to explore.

**TMDB 6000**

This dataset was generated from The Movie Database API. This product uses the TMDb API but is not endorsed or certified by TMDb. This dataset contains the latest movies data as of October 2023.

**World top 200000 scientists**

Data is from <https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/1>, description of each column is within the excel table.