Investigate a Dataset - Data Set Options

Choose one of the following datasets to explore in your project analysis.

Data Set	Overview and Notes	Example Questions
TMDb movie data (cleaned from original data on Kaggle)	This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue. ● Certain columns, like 'cast' and 'genres', contain multiple values separated by pipe (|) characters. ● There are some odd characters in the 'cast' column. Don't worry about cleaning them. You can leave them as is. ● The final two columns ending with "_adj" show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.	Which genres are most popular from year to year? What kinds of properties are associated with movies that have high revenues?
No-show appointments (original source on Kaggle)	This dataset collects information from 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. A number of characteristics about the patient are included in each row. ● 'ScheduledDay' tells us on what day the patient set up their appointment. ● 'Neighborhood' indicates the location of the hospital. ● 'Scholarship' indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família. ● Be careful about the encoding of the last column: it says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up.	What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?
Gapminder World	Gapminder has collected a lot of information about how people live their lives in different countries, tracked across the years, and on a number of different indicators. For this project, you should select at least	Have certain regions of the world been growing in your selected metrics better than others? Are there trends that

	three indicators to investigate. ● Data is provided as Excel spreadsheet files. You will want to use a spreadsheet program to export each table as a csv file. ● You will want to look into ways of reshaping your data so that it is tidy, especially if you want to do comparisons across indicators. After joining your data together, your columns might look like: {Country, Year, Indicator 1 Value, Indicator 2 Value, … } ● Some of the datasets might have been updated since they were collected on Gapminder. If you use these updated datasets, make sure you document this in your report.	can be observed between the selected metrics?
Soccer Database (original source on Kaggle)	This soccer database comes from Kaggle and is well suited for data analysis and machine learning. It contains data for soccer matches, players, and teams from several European countries from 2008 to 2016. This dataset is quite extensive, and we encourage you to read more about it here. ● The database is stored in a SQLite database. You can access database files using software like DB Browser. ● This dataset will help you get good practice with your SQL joins. Make sure to look at how the different tables relate to each other. ● Some column titles should be self-explanatory, and others you'll have to look up on Kaggle.	What teams improved the most over the time period? Which players had the most penalties? What team attributes lead to the most victories?
FBI Gun Data (original source on Github)	The data comes from the FBI's National Instant Criminal Background Check System. The NICS is used by to determine whether a prospective buyer is eligible to buy firearms or explosives. Gun shops call into this system to ensure that each customer does not have a criminal record or isn't otherwise ineligible to make a purchase. The data has been supplemented with state level data from census.gov. ● The NICS data is found in one sheet of an .xlsx file. It contains the number of firearm checks by month, state, and type. ● The U.S. census data is found in a .csv file. It contains several variables at the state level. Most variables just have one data point per state (2016), but a few have data for more than one year.	What census data is most associated with high gun per capita? Which states have had the highest growth in gun registrations? What is the overall trend of gun purchases?