DataFest Overview 2025

Data provided by Savill's for this DataFest project are not public. These data are proprietary and are only to be used for the purposes of the American Statistical Association's DataFest.

By using this data, you agree to:

1. During participation in ASA's DataFest, store and manage the data securely and privately. This means you may NOT upload the data to websites that may provide access to the data to anyone other than yourselves. (For example, if using Tableau, Github, ChatGPT, make sure privacy settings prevent others from seeing or accessing the data.) If you are not sure, then don't use that service! (Note: Rstudio Cloud does not have access to your data and so you can use it.)

2. Erase all data after your ASA's DataFest participation is complete.

3. Not identify or attempt to identify the information contained in the dataset, nor

contact any of the individuals whose information is contained in the dataset.

4. Comply with all applicable U.S. federal and state laws and regulations relating to the maintenance of the dataset, the safeguarding of the confidentiality of the dataset, and the use and disclosure of the dataset.

5. Not publish results of your analysis of the data except that the final products of the competition (video, slide deck, one-page summary) may be displayed on team members' websites and on campus ASA DataFest websites.

6. Not share the data with anyone who is not a participant of ASA DataFest

Finally, please do NOT reveal the source of the data or any features of the data to ANYONE before May 3. This will ensure that all ASA DataFest participants around the globe have the same experience with the data as you. This means to not post anything to social media that might reveal clues, and to make sure that public github repositories are not discoverable (and if they are, that identifying work is removed).

The Challenge

Savills is an international commercial real estate firm that helps businesses find a place to lease offices.

Imagine that you were searching to buy a house or rent an apartment. There are lots of websites that you can turn to for help. These websites tell you about various neighborhoods' "walkability", crime statistics, school quality data, and maybe other intangibles, such as the type of people living there ("hipsters", "young parents", etc.), or the type of neighborhood ("trendy", gentrifying, sales slowing/increasing).

There are no such websites for commercial real estate. Instead, large businesses rely on companies such as Savills to advise them on which areas are growing or shrinking, which areas have available office space, etc. Savills makes these recommendations based on several data sources that they put together. By looking for patterns on a local scale, and, sometimes, comparing these to national patterns, they can help find the perfect office space. A legal firm might ask Savills, for example, whether they should relocate from Uptown Manhattan in New York City to Hudson Yards, having noticed that other law firms are opening offices there. Savills examines these data to provide insight into trends within and across markets, as well as economic data from government entities, to advise their clients on whether it is a good time to move, whether there is sufficient office space for their needs, and whether there is too much (or too little) competition.

Once a space is identified, these new offices are obtained by the business through a lease -- a contract between the business and the landlord/owner of the building in which the offices will be occupied. Data concerning leases are much noisier and more complex than those involving sales because leases are not a matter of public record, and so tracking leasing activity is done by a network of organizations. You'll find that the data we provide are most complete for (a) large spaces (leasable square feet greater than 10,000) (b) larger markets and (c) tech, legal, and financial firms.

Savills has provided a data set consisting of every known commercial leasing transaction within specific U.S. markets between 2018 and 2024. The pandemic was a disruptive moment for this sector of the economy (as it was for so many others), and Savills would like to know what advice they can give to clients in specific markets within the Legal, Financial Services, or Tech sectors (or other sectors, too, if that interests you) as the recovery continues. This advice would be based upon patterns noted within markets (or submarkets) across time and space. Are particular fields expanding office space (as work-from-home policies are rescinded) or contracting (as AI takes hold)? Are offices moving to or from high-cost areas?

A particular challenge of this data set is that the data are not always reliable. Lease signings are not publicly disclosed, and sometimes make their way into the data set only because a landlord has withdrawn an advertised listing, which suggests that the space is no longer available and so was likely leased. Market and submarket boundaries change with time, usually through mutual agreement among real estate companies and other interested concerns. Much of the data is collected by CoStar, a company that specializes in real estate data. Savills devotes additional resources to cleaning, preparing, and supplementing the data in areas of their concern. We have chosen to provide you with all of the data, regardless of its completeness, in case there is valuable information hidden within.

Your challenge is to inform Savills of notable trends or microtrends in the commercial real estate market that could be used to advise clients on where, when, whether and how to locate their offices.

How to Get Started

A frustration you'll feel working with these data is that, particularly for some variables, there are high rates of missing values. Still, you will find that these variables are fairly complete within some markets and within some types of business concerns. As a first step, we suggest you focus on leases of 10,000 square feet or more in one (or all three) of the sectors that Savills is focused on, and do so within a major market, before comparing to other sectors and other markets. (Savills plays close attention to larger transactions and major markets in the U.S.) Also note that despite the fact that the name of the company is often missing, there are still over 25,000 companies listed, and an investigation might focus on any one (or a collection) of these.

It is difficult to fully understand the commercial real estate market based on transactions alone. The U.S. government traditionally provides a rich treasure trove of data to help researchers and businesses understand economic trends. We recommend, for example, considering the American Communities Survey by the U.S. Census. This includes, among many other things, data on commute-times within specific regions of the U.S. Through data.gov you can find weather data, business data, and more. (The Bureau of Labor Statistics and the U.S. Census are probably the largest providers of data relevant to this challenge.) Note that these data are sometimes provided on a different calendar than the data Savills has provided. We recommend you try to reconcile the dates based on the Savills quarters, which are the most reliable time measurement in the data set. When faced with weekly data, for example, Savills will sometimes use only the last week in the quarter to report, or might include the maximum value during that quarter. You might consider other possibilities as well, such as a mean.

We recommend you think small...macro trends will provide context, but are less likely to be informative than observations about very specific markets or submarkets. Savills clients are not considering the entire country (usually) as a potential place to lease. They want to be near the talent pools in places where their employees can easily get to work (or not!). Tech firms look for office spaces with wide-open plans, while law firms prefer many private offices. Real estate developers might create zones with similar office-types, and so clusters of tech firms or law firms might tell us something about the types of buildings/spaces available in that area. These are observations that are not "visible" if you're considering the entire data set as a whole, and so require drilling down. However, comparisons across different submarkets might be worth noting. Are patterns seen in Southern California different from Northern California, for example?

Where to begin? Read the codebook first, and examine subsets of the data to get a sense for what is happening. You'll do better by coming up with some hypotheses or questions and investigating those rather than hoping to stumble upon a pattern. (If these were stumble-upon-able, then businesses wouldn't need Savills.)

Data are quite sparse in some markets, and these might be markets that you're interested in. One piece of advice if you are interested in studying a market with sparse data is to expand your scope to include several markets within a larger region. The entire south-eastern U.S., for example. Maps might be helpful to get your bearings about where markets are located.

You will likely not make much progress if you simply create graphs and hope to discover something. Instead, come up with questions or ideas and explore those. Along the way, you'll find that you're refining and revising your questions until, before long, you have a story to tell. Questions can be simple when you begin: how many transactions were in this market? in which markets is Time Warner located in? Which company is in the most markets? And take it from there. Your initial questions won't always pan out, but they'll get you thinking about the data and wrestling with some of the data management issues.

If you are looking for Census data, and using R, then one way of getting easier access to the data is via the tidycensus package: https://walker-data.com/tidycensus/ and using this website to search for the necessary codes: https://censusreporter.org/topics/table-codes/   Note that you will have to get a Census API key from the U.S. Census Bureau website to use tidycensus.

The Data Files

The data are contained in three files, and a fourth and a fifth file provides some contextual support. The primary file is Savills_lease_transaction_data. The codebook provides variable definitions and indicates which file the variable is contained in.

1) Leases ("Savills") (194685 X 35)

These data come directly from Savills and consist of data compiled from several sources, including CoStar, Savills, and the DataFest team (with Savill's guidance).

Each row represents a lease transaction: a business signs a lease with a landlord. The lease allows the business to occupy space that the landlord owns or controls. These include all known transactions from January 1, 2018 through December 31, 2024.

This data set also contains *market* level data. For example, data on the total available rental space within a market during a specific time quarter. These market-level variables are often further broken down into two classes: *internal_class*==A and *internal_class*==O. For example, the variable RBA provides the total available rental space within a market during a specific quarter for a particular value of internal_class. (This variable indicates whether the building is high quality (A) or not (O)). In other words, within a given market (Atlanta, say), you'll see the same value for every row in the data within a given year and quarter for all leases involving type A internal-class, and another value for every row involving type O internal class. Please see the codebook for further details.

2) Major Market Occupancy Data (190X6). These data are provided by Kastle, a security company that specializes in security services for commercial buildings. One service they provide is their card-swipe system, which is used for visitors and employees to gain access to the building. Savill's purchased card-swipe data from Kastle, and this particular data set shows an estimate of building occupancy based on card swipes from 2020-2024. But beware! The occupancy rates are reported in percents, and the percents are based on a baseline value of occupancy on March 1, 2020. Thus, a value of 100% means that the building has the same occupancy as it had on March 1, 2020. Imagine a building with 10 offices. On March 2020, 5 are leased. On April 1, 2020, 5 are still leased. Occupancy will be listed as 100% even though only 5 of the 10 offices are leased. On May 1, 2020, 4 offices are leased. Occupancy will be reported as 80% because 80% of the leased offices back on March 1, 2020 are still leased.

The purpose of these data are to gauge occupancy relative to when the Covid lockdowns began.

Only 10 markets are included. The ten markets are Austin, Chicago, Dallas/Ft. Worth, Houston, Los Angeles, Manhattan, Philadelphia, San Francisco, South Bay/San Jose, Washington D.C.

3) unemployment.csv (1848 X 5). These data are provided by the DataFest team, and were downloaded directly from the U.S. Bureau of Labor Statistics. Each row provides data on the percent unemployment within a state for a given year (2014-2024).

4) CRE Terms and Definitions_NAIOP 2017.pdf. A document containing commonly used terms in Commercial Real Estate.

5) We also provide "Price and Availability Data.csv". This contains market-level information about major markets. **These data are ALREADY merged into the Leasing.csv file**, but are provided here in a simpler format. Each row represents one market during one quarter. Separate variables distinguish between A-level buildings and O-level buildings.