Urban Informatics Toolkit

Josiah Parry

2020-04-20

Contents

Chapter 1

Welcome!

Welcome to the Urban Informatics Toolkit! This is an online book that is intended to jumpstart your work of analyzing and developing understanding of the urban commons. In it you will find material on the theoretical underpinnings of Urban Informatics and a (mostly) thorough introduction to the statistical programming language R to get you hands on and working with data that you will encounter in your research, coursework, and in the wild west of open data.

This book represent to me many things. Of which are the two years of study in the Urban Informatics program at Northeastern, nearly five years of self-directed learning of the R programming language, two years of teach R, and many, many, many hours of frustration trying to understand and grasp concepts that could have been distilled into simple and easy to understand language.

In the following chapters I will do my best to avoid overly technical and verbose descriptions of theory and technical concepts. I will attempt to explain everything in a manner that I would to my friends over a beer, coffee, or, as these strange times would have it, a zoom call.

By the end of these pages I hope that you have become self-sufficient in your analyses. My intention is not to teach you everything that you will need to know because this is *impossible*. Data analytics are always changing and what is current will be dated—perhaps by tomorrow. In fact, whilst in the middle of writing this one of the most commonly used tools was changed in such a way that much of what I wrote became bad practice. This is all to say that the purpose of this book is to make you **self-sufficient**. The goal is to ensure that *you* are able to understand the fundamental concepts of scientific inquiry in the big data era, how to think about data analysis, and be equipped with the fundamental knowledge of R to understand what is happening—to some degree—under the hood and what and why.

1.1 Structure of the book

This book is partitioned into four sections each with it's own theme. In the first section *Foundations of Urban Informatics* we will first begin by exploring the nature of big data and its role in the field. Then, we review different approaches to scientific inquiry and seek to understand how big data has further enabled a newer approach. And finally, we review Broken Windows theory as well as the development of **ecometrics**.

The second section is dediated to introducing you to R as a programming language for data analysis. This is where we begin our technical work. We will ensure that you have all of the software and data that you will need to follow along with the exercises in this book. This section will be the most difficult to overcome. This is because you must learn an entirely new mental framework—even if you know how to program in another language. In this section we will work from reading in data to manipulating it while along the way learning some fundamental theory.

The third section is the largest and is dedicated entirely to information visualization. In this you will learn how to craft visualizations and understand when to make what kind of graphic. Furthermore we will dive into expanding upon traditional graphics by incorporating many different aesthetics. This section is expansive due to the importance of visualization. Visualization is our method of communication. While the written word is powerful, it requires more work to get someone to read than to look. If we can improve upon our visualization, we can make the work that we do more accessible to the public.

Finally, we will close by learning how to work with multiple datasets and spatial data. These are two of the more advanced topics and as such will be reserved for when the fundamentals have been reinforced. With regards to spatial analysis, many of you who may come from a geography or GIS background may find this section lacking. You would be right. Given the immense depth of the geospatial sciences, that is a topic that is deserving of its own book.

1.2 Considerations

Before we continue, I want to reitaterate that this book will not introduce you to everything that you will need to know. As the field continues to grow and as the number of tools available in R increase I will work to continually add to and improve this writing. If there are topics that you would like to see included or expanded upon, please submit an issue on GitHub https://github.com/JosiahParry/urban-informatics-toolkit/issues or reach out to me directly.

In this next chapter we will discuss big data.

Without further ado, let's get on with it!

Part I

Foundations of Urban Informatics

Chapter 2

The Utility and Danger of Big Data

Urban Informatics is in a way a byproduct of the "deluge" or "proliferation" of data. In essence, as we as a society have progressed technologically, we have been able to capture and store data on a rather unprecedented scale. This has led to massive stores of data that are used primarily for record keeping that are updated at near-real-time. For example, think of every time you make a Facebook post or send a tweet. That post or tweet is subsequently recorded in a remote database to ensure that it can be accessed at a later time. Dan O'Brien notes that this characteristic of big data is of the utmost consequence for "the advancement of science and policy in the digital age" (O'Brien, Urban Commons, p. 59). These naturally occurring data are so useful because they are essentially a track record of individuals' behavior over time. It is in a way as close as we can get to measuring behavior at real time.

If we turn to the urban context, the importance of big data may become ever more apparent. For example, Boston local government has been keeping detailed records of property assessments, taxes owed and paid, by whom, their demographic characteristics, when and even where down to the ward level (Boston Public Library, recently digitized records). These administrative records have been kept on ink and paper until just a few decades ago. Through digitization efforts, these data are now accessible to historians, urban scholars, local government, and the general public. Having such data accessible provides a way to quantitatively inspect the development of the city from its geography, its policies, its demography, and much more that we have yet to see uncovered.

There are a number of benefits that naturally occurring data provide. The first is that these are, in theory, comprehensive and contains information about all residents. Through administrative data we should, for example, be able to determine the number of employed tax paying citizens as well as the under-

employed who receive government benefits. Additionally, since these data are already being collected, the associated costs are minimal. In contrast to empirically collected data, administrative data are not just a representation of a single moment in time, but rather continually changing and updating. And due to the fact that administrative data are collected at the municipal level we are inherently dealing with geospatial data—data that have a location associated it.

While there are many benefits to administrative big data, there are a couple of dangers we ought to be aware of. The first is that even though big data are comprehensive in theory, we cannot always take them as objective observations of the natural world. We must be cognizant of the fact that the biases that humans have are also represented in data. We cannot and should not separate theory from data. To take from Dan O'Brien

> ". . . the very point of science is to explain why things work the wsay they do. . .If we limit our inquiries only to correlation and eschew explanation, we are no longer conducting science." (O'Brien, 2019, Urban Commons p.61)

Furthermore, we should always be wary of the data that we use. Investigate its integrity, its source, its measurment constructs for what we may be using to understand one thing may be something else which we may not anticipate or expect. And laslty, always be aware of the ways in which you may be bringing in your own world views into your work as what we are after is not confirmation but understanding.

Chapter 3

Data in the municipal context

The so called "proliferation of data" has created vast troves of data asking to be explored. We are, in essence, in the beginning of a new Gold Rush. But rather than discovering gold, today the gold is both being created an discovered. This explosion of data is the product of improved technology in both the collection and storage of data.

If we focus our gaze towards the municipal government, the story is similar, progress is slower, and the data are more familiar. Local governments have been collecting data for centuries but until recently it was not always accessible, or even considered "data". Take the city of Boston as an example. Since the 19th century boston has been issuing and recording building permits. Through a massive digitization effort these permits are now accessible in an online database [1]. Not only are governments slowly turning to modern methods of data storage, but they are also creating applications to encourage citizens to engage with their local governments. Mobile and web applications will hopefully facilitate greater interaction between citizen and government [2]. Each and every one of these citizen to government interactions are recorded and stored in database—though not all are open and accessible to the citizen scientist.

Boston has built a few mobile applications for its residents. Notable among these apps are the BOS:311[3], ParkBoston[4], the city's least favorite Boston PayTix[5], and the new Blue Bikes[6]. Through BOS:311 residents can communicate directly

[1] Boston Building Permits: https://www.boston.gov/departments/inspectional-services/how-find-historical-permit-records

[2] Some note about co-production.

[3] BOS:311: https://itunes.apple.com/us/app/boston-citizens-connect/id330894558?mt=8

[4] ParkBoston: https://apps.apple.com/us/app/parkboston/id953579075

[5] Boston PayTix: https://apps.apple.com/us/app/boston-paytix/id1068651854

[6] BlueBikes:

to the Department of Public Works by recording an issue, it's location, and even an image of the issue. Blue Bikes trips, 311 requests, and much more are provided to the public via Analyze Boston, Boston's data portal[7].

This new availability of data has unintentionally altered the way in which scientists interact with data. For the purposes of scientific inquiry, scientists and analysts have historically been rather close to the data generation process. While we as residents and citizens interact with governmental agencies, it is not in the name of science. And the governmental agencies are engagin with residents in for the purpose of governance, not science. As such, much—if not all—of the open and public data that we interact within the urban informatics—and greater digital humanities—fields was not generated with the express purpose of being analyzed. This inherently changes the way in which analyses are approached.

In approaching data of this nature, researchers have began embracing a paradigm of *exploratory data analysis* (EDA). EDA is extremely useful for developing insights from data in which there were no a priori[8] hypotheses. In their influential book *R for Data Science*, Garret Grolemund and Hadley Wickham describe this inductive approach of exploratory data analysis.

> "Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again." [9]

When researchers set out to test a hypothesis they often will become closely involved with the data generation process. In this scenario, researchers are more likely to have preconceived hypotheses and expectiation of what they may find hidden in their data.

This condition is often not the case when working with open data. We do not always know at the outset of what we are looking for. With open data—and any data really—you never know what you may find if you begin to dig. Whip out your hand shovel and prepare to upturn the soil. You might find seedlings that may sprout into your next study.

[7] https://data.boston.gov/

[8] "Relating to or denoting reasoning or knowledge which proceeds from theoretical deduction rather than from observation or experience." Oxford Dictionary

[9] https://r4ds.had.co.nz/explore-intro.html

Chapter 4

Approaches to and Schools of Urban Informatics

4.1 Scientific approaches

In the sciences generally, there are two approaches to scientific inquiry: inductive and deductive. These two approaches can be best characterized as "bottom up" and "top down" respectively. Each has their own origins, strengths, and weaknesses. An argument has been raging—in the scientific sense of raging—since the 17th century about which approach is the best one. My answer to this? Both, and neither. And you'll see why shortly.

Let's start by getting a grasp on deductive approaches (also referred to as deduction). With deduction we start with a theory about the workings of some observed phenomenon. From this theory, we create a hypothesis, then observe (or collect data on) the phenomenon. With this data we then confirm or refute our theory. This is how we all have, most likely, been taught about the scientific method.

Induction works in a reverse order. It works by looking at the natural world and doing just that, looking. It works by noticing something—a pattern, a unique occurrence—then noticing it again, and then again, and then under slightly different conditions. From those observations, we draw hypothesis. We then observe yet again and see if we can refute or add a little bit more credibility to our findings. Then from doing this time and again, we can build a theory. It's somewhat like Sherlock and Watson finding clues and then coming to (frighteningly specific) conclusions. These theories, no matter how we get there, are the frameworks that we use to try and explain phenomena that we see.

4.2 The Chicago School

Much of what is known as the urban sciences today can be traced back to the late 19th and early 20th centuries at the University of Chicago. The University of Chicago was at the time the epicenter of the new field of American Sociology which came to be known as the Chicago School. In that era, the social sciences were seeking to create grand theories of the world. Take, for example, the new field of Anthropology that crafted theories about the origins of the human race. The Chicago School "fostered a very different view of sociology: first-hand data collection was to be emphasized; research on the particular case or setting was to be stressed; induction over deduction was to be promoted" (Turner, 1988).

Scholars such as Robert Park, Ernest Burgess, and Louis Wirth developed a number of micro-theories to understand the city. Most notably is the culminating work The City by Park and Burgess, a collection of essays that encapsulate decades of careful observation that led to a number of theories that still have influence today. Their body of work, important in so many ways, is an early paragon of an inductive approach to social research.

4.3 Complexity and Santa Fe Institute

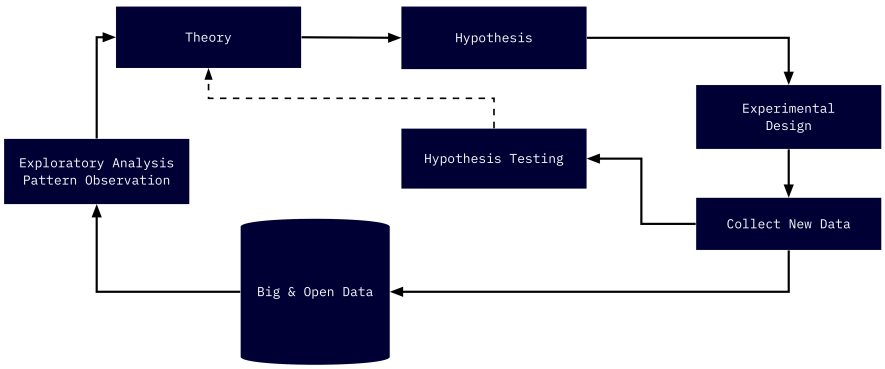
On the other end of the spectrum rest the Santa Fe Institute (SFI) and their complexity science. The SFI's mission is to "endeavor to understand and unify the underlying, shared patterns in complex physical, biological, social, cultural, technological, and even possible astrobiological worlds."[^complexity] Crudely, it is their goal to unify theory into one general master theory. Central to their theoretical focus is the the view that "valuable ideas are often confined to the echo chambers of academia."[^complexity] In this view, they are not wrong.

Their work is important for bridging many so called gaps between disciplines. They apply biological theories of scaling to those of human development. And their findings have been fruitful! Their focus on the interconnectivity of theory of both the natural and social worlds is in some ways the messy work that must be done. This too is in the spirit of the Chicago School as illustrated by it's view of the City as an organism.

4.4 A Hybrid Approach

As part of this proliferation of big data we have more and better data within reach. As such we are able to, perhaps even encouraged to, take a much more hybrid approach. Within the these data are a multitude of opportunities to explore and glean patterns that we may have been so subtle that it hasn't been observed before. Or, alternatively, the data had not been collected before. This

allows us to take a much more inductive approach where we can craft theories from the patterns that we observe, then we can test those theories in creative ways. It is here where, I believe, that the Boston Area Research Initiative (BARI) rests on the spectrum. This hybrid methodology incorporates both inductive and deductive approaches.



The above graphic is m best attempt to illustrate this hybrid model. We start with data. We use publicly available (open) administrative data to explore. We get our metaphorical hands dirty with the data. After we munge it, transform it, and reaarange it, we will walk away with some tidbit of information. From that we discover more. And at the end of the day we develop a theory—a framework for understanding what we have observed. Next, we then develop a hypothesis using that new theory and apply it to some other set of data or some new circumstance to test and refine the theories we have developed. In this we both create theory from observation, and test those theories on new and unexplored observations.

Dan O'Brien, Director of BARI, claims to track more with the Chicago School in their approach—and this is correct. But, BARI also actively seeks to evaluate existing theory. There are no better examples of the BARI approach than the development and use of **ecometrics** to understand the City and test existing theory.

In the next chapter we will learn about ecometrics, their origins, and their use in evaluating the prominent crimonological *Broken Windows* theory.

Chapter 5

Ecometrics

Central to the work that is done at BARI are the development and utilization of **ecometrics**. Ecometrics represent a quantitative representation of physical and social environment. In the Urban Informatics context, ecometrics are created to extract **latent constructs**—or variables that can only be inferred from a dataset—that illustrate some physical or social phenomenon.

To understand this, we need to again contextualize these datasets. They *are not* created with the intention of beng analyzed or to measure the blight of a neighborhood or the social unrest of a city. The data may tell a story of an underserved neighborhood or of a gilded community with a beautiful brick façade with next to no collective efficacy efforts. These datasets contain gems—beautifully inelegant snapshots of the societal quotidien. But measuring that? That's the tough part and that is why we create ecometrics. They provide us with a way to adapt existing data to address new problems.

Of the work that BARI conducts, the production of city-wide ecometrics of social and physical disorder are most emblematic of this hybrid approach. To understand this work we need to venture back to 1982 and an article from the Atlantic called *Broken Windows*.

5.1 Broken Windows Theory

During the beginning of the crack-cocaine epidemic George Kelling and James Wilson wrote a now [in]famous article titled *Broken Windows* which outlined a new theory to explain the occurrence of crimes. The premise of this article is that the *presence* of disorder is more concerning for a neighborhoods residents than the actual crime that occurs. Further, the "visual cues of disorder . . . begets predatory crime and neighborhood decline" (O'Brien and Sampson, 2015).

Broken Windows captured the eyes of pundits and policy makers. The simplicity of the theory makes it easy to Broken windows has historically captured the attention of policy makers. The vast public support has led to a large body of work largely disputing the merits of this theory. In the process of doing so, much work has gone into actually quantifying disorder in a city. In a seminal article by Sampson and Raudenbush (1999), the practice of systematic social observation was created. This is a process in which imagery of public spaces is taken and coded to identify disorder—i.e. the presence of empty beer cans—which can then be quantitatively analyzed. This is an early example of an ecometric.

5.2 Quantifying Disorder

In 2015, O'Brien and Sampson published the article *Public and Private Spheres of Neighborhood Disorder: Assessing Pathways to Violence Using Large-scale Digital Records*. This article epitomizes the hybrid aproach to urban studies. In it, O'Brien and Sampson utilize 911 dispatches and 311 call data to create measures of both social and physical disorder. These measures were then used to put Broken Windows theory to the test. The process of using existing administrative datasets as a method of estimating social and physical phenomena illustrates the inductive approach. Whereas testing testing the efficacy of Broken Windows is indicative of the more traditional deductive process.

Quantifying disorder is no small task. In their 2015 paper the authors write

> Taking up this challenge, O'Brien, Sampson, and Winship (2015) have proposed a methodology for ecometrics in the age of digital data, identifying three main issues with such data and articulating steps for addressing each. These are (1) identifying relevant content, (2) assessing validity, and (3) establishing criteria for reliability.

The above is an astute summation of the problems that arise with big data and how they can be overcome. The biggest of concerns, as mentioned in the opening of this section, is the validity of the data we are using.

5.2.1 Defining the phenomenon

The method that they propose requires us to do three main The first is to clearly define the phenomenon that we are hoping to measure. Following, we must idenify the **relevant data**. For example, in O'Brien and Sampson (2015), they define five ecometrics as

- Public social disorder, such as panhandlers, drunks, and loud disturbances;

- Public violence that did not involve a gun (e.g., fight);
- Private conflict arising from personal relationships (e.g., domestic violence);
- Prevalence of guns violence, as indicated by shootings or other inci- dents involving guns; and
- Alcohol, including public drunkenness or public consumption of alcohol.

These definitions provide clear pictures as to what is being measured. The next step is to surf through your data and do your best to match variables or observations to these measures. Then, through some process—usually factor analys—ensure that these measures are truly relevant.

5.2.2 Validating the measure

Once an ecometric has been defined and properly measured, the next step is to validate it. I think of this process similar to ground truthing in the geospatial sciences. Often when geographic coordinates are recorded an individual will go to that physical location and ensure that whatever that was recorded to be there actually is. This is what we are doing with our ecometrics. We have developed our measures, but we need to compare that to some objective truth so to say.

In Sampson & Raudenbush (1999), they develop measures of physical disorder through their systematic social observation. But in order to validate their measures, they compared their results to those of a neighborhood audit. This audit served as their ground truth and was used to make any adjustments if needed.

5.2.3 Addressing reliability

This ecometric, like most others, are naturually time bound snapshots of the social and physical world. These measures will naturally change over time. Because of this it is useful to know both how reliable the measure will be for different periods of time (O'Brien, Sampson, and Winship, 2015). The authors do with with a bit of statistical finesse that is best left to them to explain. But what we are to take away is that ecometrics are often time variant and it is important for us to know at what time scale the ecometrics are intended for.

References & Readings:

- O'Brien, Daniel Tumminelli, Robert J. Sampson, and Christopher Winship. 2015. "Ecometrics in the Age of Big Data: Measuring and Assessing 'Broken Windows' Using Large-scale Administrative Records.'' Sociological Methodology 45.
- https://www.annualreviews.org/doi/abs/10.1146/annurev-criminol-011518-024638?journalCode=criminol

- Large-scale data use, ecometrics to assess disorder: https://journals.sagepub.com/doi/abs/10.1177/0022427815577835
- sampson & raudenbush systematic social observation: https://www.journals.uchicago.edu/doi/abs/10.1086/210356

Part II

Toolkit Foundations

Chapter 6

The basics

6.1 What is R and why do I care?

What is R? R is the 18th letter of the alphabet, the fourth letter in *QWERTY*—like the keyboard—and, most importantly, R is a software package for statistical computing.

R is a decendant of the S statistical programming language whose naissance can be traced back to 1976 in Bell Labratories (https://web.archive.org/web/20181014111802/http://ect.bell-labs.com/sl/S/). As S developed, people sought to commercialize the language. In 1993, the license as well as development and selling rights were given to a private company. From then on, and what is still the case today, S became available only as the commercialized S-PLUS (Martin, 1996).

Later, seeing a need for an improved statitistical software environment, two researchers from the University of Aukland created a new statistical programming language, this became known as R. R was developed in the image of S. However, one important early decision to make the R-project free and open source changed its fate dramatically.

Today, the R-project is developed and maintained by a group known as the R Core who "represent multiple statistical disciplines and are based at academic, not-for-profit and industry-affiliated institutions on multiple continents" (R SLDC, pg8). They define R as below.

> R is an integrated suite of software facilities for data manipulation, calculation and graphical display. https://www.r-project.org/about.html

To simplify it, R can be thought of as a fancy calculator. R was designed to do math, specifically statistics. R was designed to be extended to incude further

capabilities than just statistics. Indeed it has been. While R is for all intents and purposes a programming language, one should, in theory, feel like they are doing data analysis and not programming (chambers).

R is unique from other commercial statistical software such as stata and SPSS. Very fundamentally, R is a free project. While it is monetarily free, free refers to "liberty, not price" (gnu.org). In order to truly understand the adventure you will be embarking on shortly, I think it is important you familiarize yourself with the four freedoms of free software. These are:

1. The freedom to run the program as you wish, for any purpose (freedom 0).

2. The freedom to study how the program works, and change it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.

3. The freedom to redistribute copies so you can help others (freedom 2).

4. The freedom to distribute copies of your modified versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.

These freedoms are a large part of the success of R as a language. Because of the free nature of R, academics and industry experts from around the globe are contributing to the language. This means that many new statistical techniques are first implemented in R.

The contributions that people make to R are changing the ways in which people perform data analysis. Because of this, we need to start contextualizing the tooling we use as **part of** the scientific process—not apart from it. When you engage in your analyses and work on contribute to the vast body of scientific literature, remember that without the tools you are using, much of it would not be possible. When you engage in science, think to yourself how you are adhering to the four essential freedoms. Are you enabling others to do with your findings as they wish? Will your research be accessible to the greater community? What will you do to "give the whole community a chance to benefit from your [work]"?

https://cran.r-project.org/

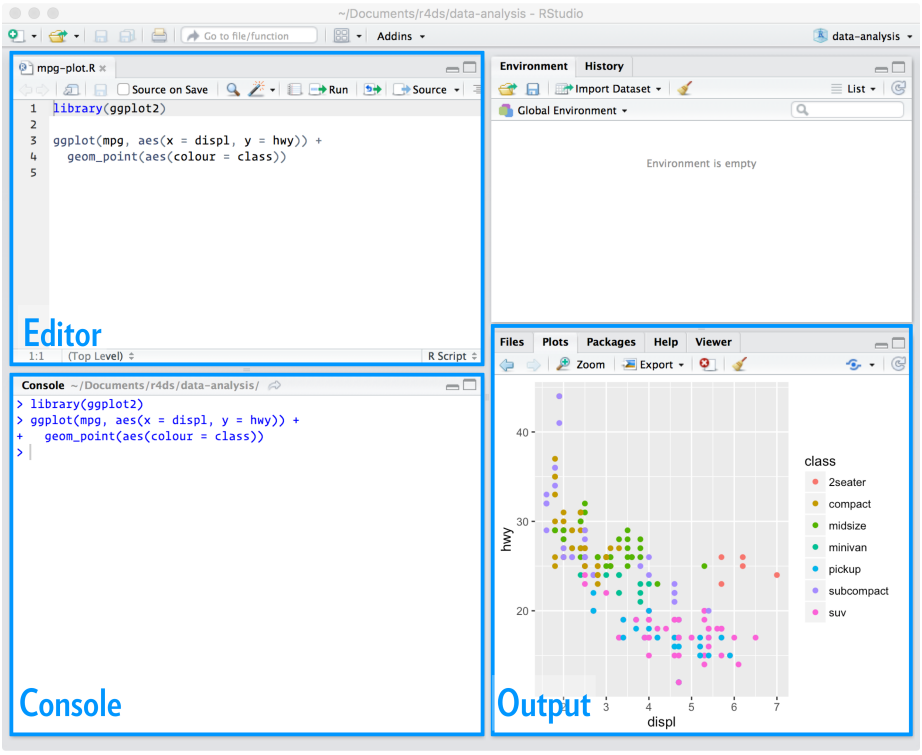
6.2 The RStudio IDE

When R is downloaded, interacting with it is somewhat of a cumbersome process. While some people love it, it can feel like programming in the matrix.

For this reason, we will use RStudio to program in R. RStudio is an integrated development environment (IDE). This means that most of the features that you will need to develop in R will all be in one place. RStudio gives you a place to write your R code, execute it, view the awesome graphics you produce, and much more.

I like to think of R as typesetting a printing press and using RStudio like using Mircrosoft word. Chester Ismay and Albert Kim's Modern Dive, provide another excellent analogy of R and RStudio. They describe R as the engine of a car, and RStudio as the dashboard (https://moderndive.com/1-getting-started.html).

Let's get familiar with RStudio. You need to know where you are when working within RStudio. There are 4 quadrants that we work with (called panes).



This lovely graphic was created graphics by Thomas mock. Source

6.2.1 The Editor

The editor. The top left pane. This is where you will actually write your code. You will see in the image above that there is tab with the name of the R file being edited, `mpg-plot.R`. The simplest way in which R code is written, is in