The Role of Data Science in Smart Cities

Samantha Hoch

ENVR 4900 - Earth and Environmental Sciences Capstone
Spring 2021

Abstract

Smart cities are an innovative approach to environmental planning and sustainability that rely on Internet of Things (IoT) devices or other technology to collect data about the city. This paper examines two possible applications of smart city technology in smart energy grids and transportation networks, as well as the challenges that come with implementing this highly integrated technology. Any data collected by smart cities is then analyzed for use by the government to make informed decisions about policy and city planning. Governments can also learn from cities that have already begun implementing smart city technologies today. Additionally, when developing a plan for a smart city, one must consider the demand for data scientists to uphold data integrity and provide analyses. There are many possibilities for tools and techniques to use during data analysis, so this paper explores an example analysis of bike share data using Tableau and Python.

Introduction

The ability of technology to collect and process large quantities of data has become cheaper and more accessible in recent years. This creates opportunities across all sectors, including environmental planning and sustainability, to make data driven decisions. Smart cities aim to use this ability to better understand how the many systems and components of a city interact, and the effect that this has on the surrounding environment. There are many aspects of smart cities that implement technology, from measuring pollutant emissions to electricity usage to crowd control. By using sensors and integrating many subsystems together, smart cities collect a large amount of data that must be stored, cleaned, and analyzed by data scientists. The results from data analysis can then be used to better inform city officials and policy makers of what services their citizens need and how to be more sustainable in order to protect the needs of future generations (Samih, 2019).

Basic of Smart City Design

While smart cities rely on small sensors and individual components, it is important to understand how all the different parts work together. One model that has been developed to explain this system describes smart cities as having three layers: perception, network, and application (Samih, 2019). The perception layer is where the data is collected, such as through IoT devices or by purchasing data from private companies. Next, the network layer is responsible for transmitting that data and securely storing it. This type of transmission relies on a high-speed network, which is something that smart cities must build into their infrastructure. Another necessary infrastructure component is the hardware to store data on, which will likely be in designated data storage centers (Samih, 2019). Once the data is stored, the last layer is the application layer, where the data processing and analysis occurs. Each component of smart cities will need to have their own perception layer and methods for interacting with the transmission layer.

An integral part of the perception layer is the IoT sensors used to collect data. The IoT is defined as "'the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and connectivity which enables these objects to connect and exchange data" (Wang and Moriarty, 2019). Any amount of IoT devices can be connected and integrated into systems of various scales, ranging from a smart house to an entire city. In smart cities these devices are deployed and maintained in specific environments that the city wishes to monitor. One example is to put IoT sensors in streetlights in order to automate turning them on and off based on the current level of darkness. By using the data from all connected devices, smart cities hope to quantify, reduce, and monitor pollutants and energy usage (Baig et al., 2017). While IoT devices still require a significant investment, they have become far more affordable in recent years, therefore making the scale of data required for smart cities possible for the first time (Samih, 2019). In the future these devices will continue to improve and become cheaper as more cities start using them.

While smart cities are considered an innovative approach to city planning, environmental planning traditionally considers three pillars of sustainability: economic, social, and environmental. To achieve sustainability, which is a primary goal of smart cities, all three of these pillars must be satisfied. For example, while reducing pollution and improving air quality is important for environmental health, it also allows citizens to spend more time outside and thus build better social connections and increase social sustainability. The economic sustainability of smart cities will be impacted by the jobs created for the maintenance of IoT devices and the corresponding data analysis. These pillars are all interconnected, so improving the sustainability of one pillar will improve the others. Therefore, having a balance of all types of sustainability is important to truly make smart cities the best and most desirable places to live.

Smart Grids

One of the main goals of smart cities is to use energy more effectively to preserve resources and reduce pollution. By analyzing energy use data and identifying areas for increasing efficiency through automation, energy grids can be transformed into smart grids (Wang and Moriarty, 2019). It is a good time to invest in these types of improvements because energy grids across the U.S. need upgrades after years of underfunding and quick expansion to fulfill increased energy demand (Momoh, 2009). The increasing frequency of natural disasters due to climate change also poses a threat to reliable energy and current systems are often not prepared. Smart grids provide an opportunity to improve energy infrastructure so that it can supply the efficient and reliable energy that citizens need.

To create a better energy system, smart grids aim to provide grid observability, improve performance, reduce costs, and allow controllability of components (Momoh, 2009). Grid observability refers to knowledge about what is happening in the grid at any time. Additionally, the grid should be self-healing, meaning if there is an issue, the system should be able to detect, analyze, and fix the problem on its own (Momoh, 2009). This will allow the system to restore interrupted function at any hour of the day, instead of being restricted to when staff is on duty. In terms of economics, smart grids need to find a balance between using enough technology to have good monitoring coverage and but not too much that the costs exceed the operating budget (Momoh, 2009). Installing a smart grid will also require an initial investment, which may require convincing citizens and stakeholders of the benefits that a new system will provide.

Smart grids are especially important now due to climate change and the corresponding transition away from fossil fuels and towards renewable energy sources. Since renewable energy, such as wind and solar, can only be created under the correct conditions, the availability of energy will vary, and energy usage needs to be adjusted accordingly. Smart grids can provide services like automatically turning off certain devices when there is less energy available. They can also help with load shifting, which refers to minimizing energy use during periods of low solar and wind output. This is important because if most of the energy is used when or close to when it is produced, then there is less of a need for energy storage infrastructure (Wang and Moriarty, 2019). The load shifting features of smart grids will be most efficient when be paired with consumer driven strategies to alter how individuals view their energy usage. One economic method of promoting load shifting is through pricing that incentivizes using energy during times of high output (Wang and Moriarty, 2019). Another approach that appeals to the social leg of sustainability is to offer an app that help users predict when the best time to use energy is based on weather forecasts. If it is going to be cloudy for the next couple days, then some activities, like washing clothes, can be avoided until there is more solar energy again.

Smart grid design can also implement technology beyond IoT devices by using concepts from computer science and artificial intelligence. Smart grids can use Adaptive Dynamic Programming (ADP) to optimize the system, handle problems and decrease costs. ADP is beneficial because it learns from previous problems, so the system will continue to improve the more that it experiences (Momoh, 2009). Conventional programming methods have been unable to handle energy grid problems due to the scale of the data and lack of historical knowledge (Momoh, 2009). Using novel computer science techniques will help ensure that smart grids provide optimal functionality.

Transportation Networks

Another major area of smart design is transportation networks. This includes traditional environmental planning considerations, such as placing shops, work, and homes within walking distance of each other (Wang and Moriarty, 2019), as well as more technology-based solutions. One important aspect of transportation networks is public transportation. This method of transport is well known to be more environmentally friendly and cost efficient. The difficulty of public transport often lies with overcrowding and unreliability. Overcrowding can be improved by better understanding where the majority of people are in the system at any given time. Many public transportation systems now use transit cards, such as Charlie Cards in Boston, which can provide information about rider data. This data can be used to determine overall trends and where more or less transportation capacity is needed. More specific analysis can also be done to prepare networks for surge events, such as a sports game or the end of a concert (Barton, 2020). By improving the scheduling of public transportation routes, cities can encourage the use of them, therefore reducing car use and greenhouse gas emissions.

While tracking rider information in already commonplace in cities around the world, there are other applications of smart devices that can be used to improve the functionality of the transportation equipment. IoT devices can be set up to automatically monitor sections of the network or specific hardware. This will allow cities to better detect mechanical problems before they happen or be quickly alerted once a problem arises (Barton, 2020). This will help to increase the reliability of the system, which will make the public more likely to use it. IoT devices and sensors can also be used to give travelers updates about predicted arrival or departure times, or to develop personalized recommendations that emphasize use of public transport while meeting individuals' needs (Burkhalter, 2020). The increase in technology usage will also make it more feasible to offer Wi-Fi on public transport, offering a benefit that citizens would not have access to if they chose to drive their individual car.

Beyond public transportation, smart cities must also optimize their design for cars and traffic. Real time traffic data is already used by citizens for finding the quickest route to a destination, but cities have also started to use this data to adjust aspects of their transportation networks such as traffic light timings and speed limits to achieve maximum efficiency (Wang and Moriarty, 2019). Some bridges, like the Golden Gate Bridge, even have the option to alter how many lanes are travelling in each direction based on the current need. These efforts to reduce traffic are important from an environmental perspective because the less time that people spend sitting in traffic, the less greenhouse gasses are emitted. This is also important from a social perspective because less traffic makes a city more desirable to live in due to air quality improvements and decreased noise pollution.

As with in public transit, there are also possible applications of IoT devices for traffic and car management. For example, Boston has begun using parking spot sensors to detect areas of low and high use. Based on this data, Boston sets different time limits for parking, where busy areas allow less time (Wang and Moriarty, 2019). This helps town officials to promote parking in areas that are typically less busy, therefore spreading out the traffic and parking between different areas of the city. IoT devices can also be used to create geo-fences to gain information about specific areas inside the given perimeter (Reclus and Drouard, 2009). Geofences are especially important for shipping companies to be able to accurately track their goods. This is a feature that will be important to citizens as the proportion of shopping done online continues to grow (Reclus and Drouard, 2009).

If cities do not want to invest in IoT sensors, they can also form partnerships with existing companies to gain more transportation data. Boston has a partnership with Uber which gives city planners access to ride data, including origin, destination, and travel time (Wang and Moriarty, 2019). Since ride shares impact aspects of public transportation use, it is important to have access to all different types of transportation network data. This partnership is a great example of the benefits to policy makers in smart cities of working with the private sector. Since IoT devices can be expensive, partnerships like this provide cities with cheaper information that companies are already collecting. It is also beneficial for companies like Uber to partner with cities because they likely share similar goals in terms of wanting to improve traffic and roads so that their customers will have a better experience.

Challenges to Smart Cities

As smart cities become more popular, there are a variety of challenges that come from integrating data into so many components of life. Replacing typically manual tasks with machine automation, creates the potential for a cyber-attack in a place that was not previously at risk. This security issue is a complex problem to solve because of how widespread the networks of smart cities are. Due to the interconnectedness of many subsystems into the larger city-wide system, each component has the potential to put the whole system at risk (Baig et al., 2017).

In terms of smart grids specifically, they have many more computerized parts than traditional electricity grids (Baig et al., 2017). The interconnected systems need to be able to deliver messages, and cyber-attacks can clog this system. This has the potential to lead to blackouts if alert messages are delayed. There is also a large quantity of data from smart meters that must be appropriately stored in the cloud. To maintain privacy, this data should be made anonymous, so any data leaks will not exposure information about the customer. Smart grids can also expose issues such as electricity theft which may require police investigation (Baig et al., 2017). There should be systems in place to deal with data usage and privacy regulation during an investigation.

Transportation networks face some similar issues with possible delays if cyber-attacks interrupt the public transportation system. There was an attack on the transportation network in San Francisco that infected all their computers with ransomware.

While this attack did not steal data or put anyone at risk physically, it left the computers inoperable and forced officials to open all fare gates until the issue was solved (Bingyi et al., 2019). Since this is a new area of concern, there is a lack of regulation for public transportation cyber security. Most governments are aware of this issue and see it not only as a risk to data, but also to safety and security. There is also a general lack of understanding towards the necessity of cyber security that needs to be addressed at all levels of the smart city infrastructure (Bingyi et al., 2019).

Another growing area of concern in the transportation sector of smart cities is the security risks of smart vehicles. These issues can be purely data focused or can also pose a physical threat. For example, cyber-attacks can use malicious code to gain access to the vehicle, making it easy to steal (Baig et al., 2017). Alternately, data attacks can go after data on the car, but also data on the Cloud, creating a system wide threat. These issues are important because as smart cars become more common, cities will connect to them and constantly transport data back and forth. While security threats to smart cars are currently mainly an issue for the user, they will become a larger risk when the car is tied to the larger system. There should also be a plan in place for protecting privacy of individual drivers, whether they are in a smart car or not. Real time traffic data has the potential to expose the movements of individuals, so it is important that this data is treated with care and always anonymized.

If a cyber-attack does occur in any component of the city, that data should be analyzed in an effort to prevent such an attack from happening again (Baig et al., 2017). There should be a standard process that occurs after an attack, which would include acquiring the affected data and determining what happened to put it at risk. This is a new type of forensic analysis that may be beyond the scope of what police departments are currently prepared for. Cities should make sure that they have the personal capability of this type of analysis and have a standardized process defined before implementing any new technologies.

Smart City Governance

The data provided by smart city infrastructure is only useful if it is used to develop policies and make decisions. Private companies can invest in different sectors of a smart city, but the government is where everything comes together. Especially when it comes to decisions about sustainability, the government is often the deciding factor on what path the city gets put on (Gohari et al., 2020). Therefore, when creating a smart city development plan, the government agenda should be considered (Gohari et al., 2020). Since the implementation of smart city technology is a likely a big change from current city operations, the government should also be prepared to adapt. This can include changing aspects of how it is run or how decisions are made. Without implementing changes in the government too, it will not be possible for a smart city to achieve the best livability for its citizens.

As with traditional environmental planning, involving citizens in government is important to build trust and community involvement. Technology has not only made the electronic devices needs for smart cities possible, but it has also increased the information that is available to the public. Citizens will have the ability to see the data for themselves and may have concerns that they want to raise. Smart governance must include collaboration between the government and citizens through public forums and other opportunities for involvement (Gohari et al., 2020). The government will also need to ensure that citizens feel protected against any security threats that come with smart city technology; without faith in the devices and networks, the data will not be trusted for use in decision making.

Additionally, due to the complexity of smart cities, there are also other stakeholders, such as private companies, that invest in or provide components of the infrastructure. Smart governance must include methods to collaborate with all parties involved to ensure the best possible information is available and used.

Examples of Smart Cities Today

Smart cities are more than just a concept that may occur someday; many cities around the world are already implementing the ideas of smart city technology and planning. While it will be a gradual process for cities to reach the level of technology integration needed to fit the formal definition of a smart city, there are many benefits to adding technology to even just one sector. This section will examine three cities, Dubai, Barcelona, and Singapore, and the different ways in which they are implementing technology to better the lives of their citizens.

Dubai has instituted the "Smart Dubai Initiative", which includes a variety of services from various government entities (Smart City Hub, 2017). Dubai is rapidly developing, and the government wants to continuously improve city life for both residents and visitors. The Smart Dubai website discusses their vision and how they plan to use technologies (Smart Dubai, 2020). This includes "leveraging emerging technologies such as Blockchain, Artificial Intelligence, along with harnessing Data Science capabilities" (Smart Dubai, 2020). One example of the goals that they have set is to make Dubai completely paperless by 2021. The government is doing their part by digitizing 100% of transactions. Another interesting initiative is the "Happiness Agenda", which aims to measure and raise happiness levels.

Dubai has also created a variety of apps, that aim to increase accessibility of services to citizens. This includes services such as paying a speeding ticket, paying electrical bills, getting taxis, and tracking packages (Smart City Hub, 2017). The Smart Dubai website contains links and descriptions of these apps, including UAE pass, which allows users to fill out official government forms through an app. The description says, "you will soon be able to start a business, buy a car or rent a house in a few clicks" (Smart Dubai, 2020). Apps like these can go a long way in saving citizens' time and reducing transportation. Dubai shows that governments are capable of change and adapting to make use of new technologies.

Barcelona has invested in a different part of smart city infrastructure: smart energy systems. One example is smart streetlights that change their brightness based on the activity on the street (Smart City Hub, 2017). This will reduce energy costs by minimizing light use in areas that don't need it. Another example is garbage sensors and automated waste collection (Smart City Hub, 2017). Household waste is deposited into bins that use a vacuum to suck the waste into an underground storage area (Ogleby, 2018). Barcelona also monitors the amount of waste being deposited into these underground storage areas, which allows them to improve their collection schedule and use resources more efficiently (Ogleby, 2018).

Singapore has some of the most ambitious smart city goals because they want to be the world's first smart nation (Smart City Hub, 2017). The goal of a smart nation is the same as that of a smart city, to use digital innovation to improve citizens lives, but at a much larger scale. To work towards this goal, Singapore collects data from many different places. This includes if people are smoking in non-smoking areas, littering, crowd density, and movement of cars (Smart City Hub, 2017). Singapore has decided to concentrate on three pillars which they want to make more digital: economy, government, and society (Smart Nation Singapore, 2021). They have a comprehensive plan that is available to the public for each of these components. The Smart Nation Singapore website also acknowledges the effect of COVID on their plans and how it has demonstrated the need to use technology to minimize physical contact and keep us safe. The government goals in particular have been made more ambitious in response to the pandemic (Smart Nation Singapore, 2021).

Another interesting aspect of Singapore's plan is the Virtual Singapore dashboard that they have created with all the data that they collect. This dashboard gives the government access to a variety of real time information about the city. While this dashboard is created with collected data, it can also be used to simulate situations, such as increased traffic or pollution (Smart City Hub, 2017). This allows city planners to see potential impacts, especially from predicted issues caused by climate change, and better prepare for the future.

Role of Data Scientists in Smart Cities

The amount of data used in smart cities creates a need for data scientists to manage and utilize the vast amounts of information. The individual datasets in smart cities will be large and the amount of data will only increase over time as more components are added to the system. This is considered Big Data and is exactly the type of problem that the field of data science was designed to solve (De Obseso-Orendain et al., 2015). A data scientist working in a smart city should have not only a solid understanding of computer science, analytics, and statistics, but also the societal challenges that smart cities aim to improve (Barton, 2020).

The key responsibilities of a data scientist are consistent no matter what aspect of the smart city that they work on. Firstly, they must ensure and promote open access to data (De Obseso-Orendain et al., 2015). This means that the data is not only available to the entities that collect it, but open to a wider range of people and institutions. This will help maintain transparency which will allow citizens to trust the decisions being made with the data. Next, data scientists must develop tools for data re-use (De Obseso-Orendain et al., 2015). Most data analysis that occurs in a smart city is not a one-time process. For example, analyzing data for a peak transportation event, such as the end of a concert, must occur many times. By building tools that can be used repeatedly, data scientists can continuously improve the variety of analyses and options available. Developing these types of tools responsibly also includes providing documentation so that the tool is not reliant on the knowledge of the designer to be usable. Setting a standard for formatting and code documentation can help a data science team produce consistent quality content.

Another important aspect of data management is storage. Data scientists must build and maintain databases that will continue to grow as more data is collected over time. Big Data techniques will be required because of the amount of content. This also creates the need for a data taxonomy, which is a specific way to format data relationships and naming conventions. Similar to having a standard for code documentation, this will ensure that the data system is widely understandable and usable. When organizing databases, data scientists must also consider a wide variety of data. Some will be easy to understand, such as geographical coordinates or temperature readings, while others will be unstructured data (Samih, 2019). Unstructured data included pictures, audios, and videos, and takes more consideration to store and analyze.

Sample Data Analysis

An important aspect of environmental planning is to design cities that don't rely on cars and instead offer options for walking and biking. Smart cities can improve their bike networks by better understanding how they are used and using that data to inform improvements to the system. Public bike shares are a great resource for those who do not have their own bikes, and in Boston, there is a company called Bluebikes that offers this service. It is open to anyone and has 3,000+ bikes and locations in 300+ stations in the Boston metro area. Bluebikes has a partnership with the local governments in the cities that it provides services, which is another good example of a public-private partnership. This partnership helps the cities achieve their goals of improving bike infrastructure and helps fund the service.

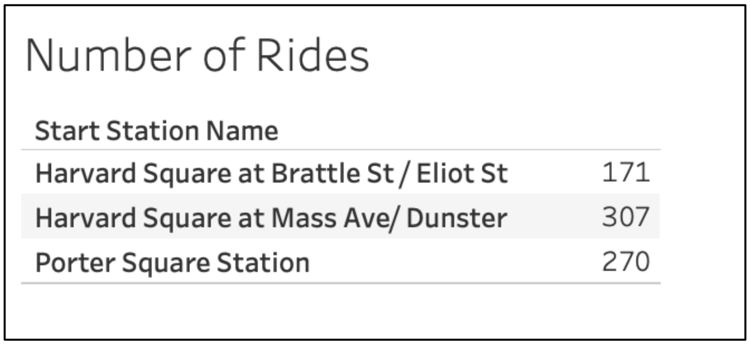
One of the important aspects of data in smart cities is open access. Bluebikes does a great job of upholding this principal by providing a variety of different data sets that are easily accessible for download from their website (BlueBikes, 2021). The first major dataset available is a quarterly release that shows basic information about every trip. Next, Bluebikes offers real-time data in accordance with recommendations from the North American Bike Share Association. Their website also contains pre-prepared summary statistics showing membership, system growth, rider trip records, and popular stations.

This data access helpful for many reasons. If a different city or company is considering implementing a similar program, they can analyze this data to see how well it works in Boston. On a more local scale, people in Boston can understand how the system around them is being used. This can be helpful in advocating for new policies or areas where service should be increased. For example, this data could be used to show areas that most commonly run out of bikes.

Bluebikes is able to collect this data by tracking purchases made at their station kiosks or through their app. There are also IoT sensors in their bike station racks, which allow them to ensure that bikes are removed and returned properly. Their system for storing and updating data is certainly much more extensive than what is available to the public. Their datasets that are available are very high quality and indicate that a great level of care has been made to clean and monitor them.

There are many different tools and methods that a data scientist working for the government or as a concerned member of the public could use for this dataset. A great tool for getting a quick overview of data is Tableau. This application does not require programming and can be used to make charts, maps, and summary statistics. Since the data and naming conventions used by Bluebikes are consistent across quarters, it is easy to plug in a new dataset and reuse the same Tableau dashboard. This is another important aspect of Data Science in smart cities – being able to reuse tools. The ability to plug in different datasets is also helpful for visualizing changes over time.

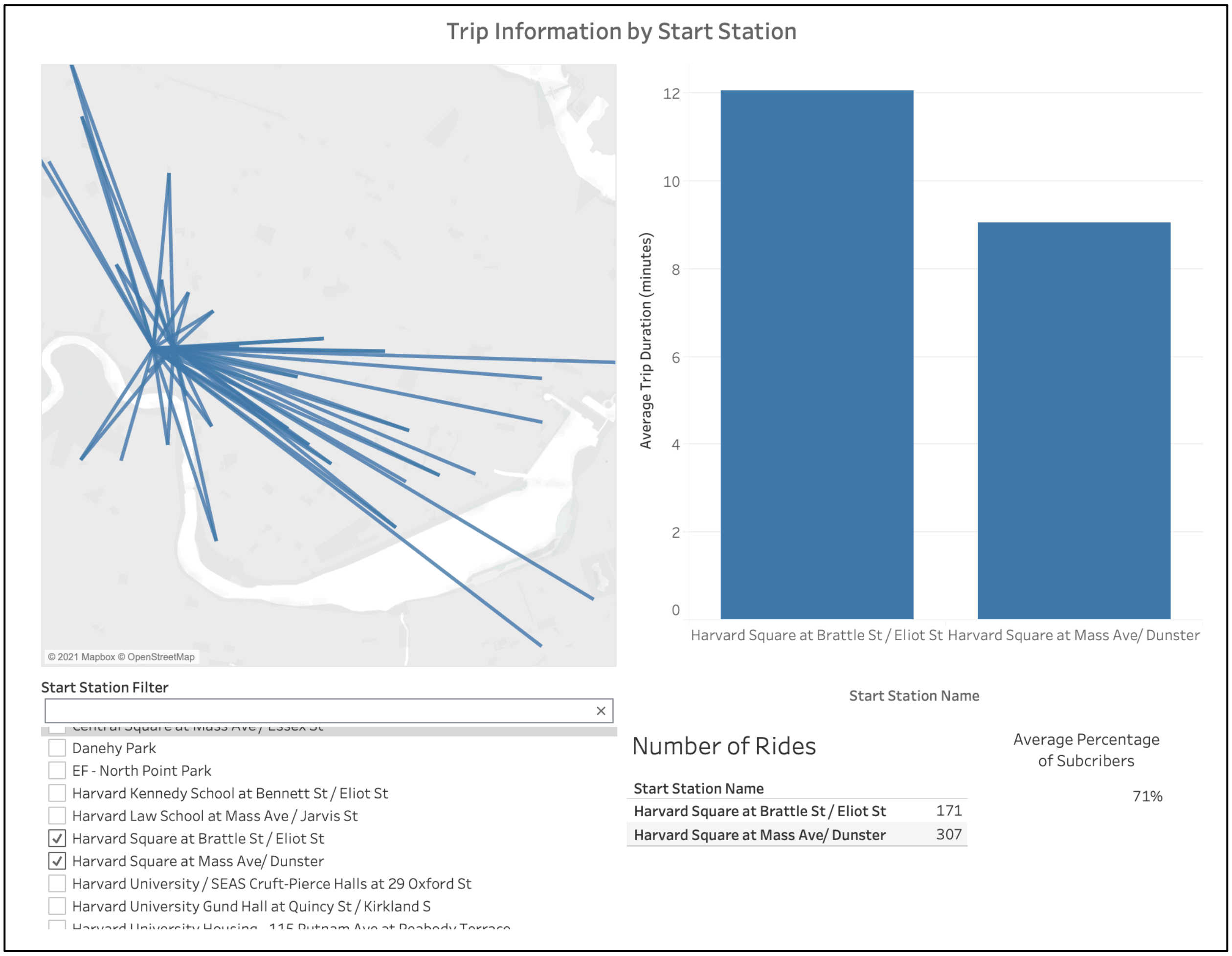
One way to visualize the Bluebikes data in a Tableau dashboard is to examine the data by start station. Tableau has a filter feature that allows you to select any number of the start stations and see the effects of that filtering across different tables. Tableau also makes it easy to create calculated fields if the input data is not the exact desired form. For example, the station location data is given as latitude and longitude, but one can create a calculated field that stores the location as spatial point data. To take it one step further, another calculated field can represent a line from the start station point the to the end station point.



Number of Rides

Start Station Name	
Harvard Square at Brattle St / Eliot St	171
Harvard Square at Mass Ave/ Dunster	307
Porter Square Station	270

Figure 1: Example Tableau Sheet showing the number of rides that originate at selected start stations


Trip Information by Start Station
© 2021 Mapbox © OpenStreetMap
Average Trip Duration (minutes)
0
2
4
6
8
10
12
Harvard Square at Brattle St / Eliot St
Harvard Square at Mass Ave/ Dunster
Start Station Name

Start Station Filter

- □ Central Square at Mass Ave / Essex St
- □ Danehy Park
- □ EF - North Point Park
- □ Harvard Kennedy School at Bennett St / Eliot St
- □ Harvard Law School at Mass Ave / Jarvis St
- ☑ Harvard Square at Brattle St / Eliot St
- ☑ Harvard Square at Mass Ave/ Dunster
- □ Harvard University / SEAS Cruft-Pierce Halls at 29 Oxford St
- □ Harvard University Gund Hall at Quincy St / Kirkland S
- □ Harvard University Housing - 115 Putnam Ave at Peabody Terrace

Number of Rides

Start Station Name	
Harvard Square at Brattle St / Eliot St	171
Harvard Square at Mass Ave/ Dunster	307

Average Percentage of Subcribers

71%

Figure 2: This Tableau dashboard examines how Bluebikes' statistics vary by start station. The statistics available are average trip duration, total number of rides, and average percentage of subscribers.

While Tableau is a good start, for more complex analysis of the data, programming is often necessary and easier. Python and R are both good coding languages for statistical analysis. They also have very widespread use, so code written in either language will be understandable and usable for many people. Programming can also be very helpful for cleaning up data, by removing data entries that are incomplete or better formatting information. Once the data is formatted, then it can begin to be analyzed, often through machine learning techniques which find patterns in the data and use it to predict outcomes.

One common area of machine learning is regression analysis. This can be used to understand the relationships between two or more variables. This can be helpful in better understanding ideas such as what factors lead to a station being popular or estimating trip time based off other variables. This type of analysis can indicate trends and point out areas of the network that are underperforming. This in turn helps data scientists make recommendations on where to improve the current system or where to add more stations.

One example of regression analysis for the Blue Bikes data is to use a classification tree. This is a method of regression which asks true or false questions about the data point in question that lead to a certain classification at the end (Galarnyk, 2019). When using this type of regression, you do not need to standardize your data. Standardizing your data refers to converting all your data to the same scale so that different variables can be compared to each other. A classification tree analysis is successful if the input variables can be used to correctly predict the dependent variable.

One way to use a classification tree for the Blue Bikes data is to determine if the end station can be predicted based on some of the other variables in the dataset. The code used for this analysis can be found in Appendix A. First the program must read the data in from the source file and prepare it for analysis. The first variables that should be removed are ones that shouldn't influence the analysis. For example, the date that the ride took place likely has little impact on where the rider went. Repetitive variables should also be removed so that only non-repetitive information is given to the model. For example, start station id, start station name, and start station latitude and longitude all refer to which station the bike was picked up from, therefore only one should be included. Another consideration for input data to a regression analysis is if the data is qualitative or quantitative. The analysis can only use quantitative data so if one wants to use a qualitative data, such as user type, it must be converted before it can be used. Instead of specifying user type as either "Customer" or "Subscriber", there must be two columns: UserTypeCustomer and UserTypeSubcriber. These columns are used as a flag, meaning that if a data entry has a true value in the column UserTypeSubscriber, then they were a subscriber. Once any desired qualitative variables are converted to multiple column flags, the independent and dependent variables must be determined.

The decision tree analysis preformed for this paper looked to see if the end station could be predicted by the other variables in the set, so the dependent variable was the end station. However, since there are so many stations, it would be difficult for the decision tree to have a unique classification for each. Instead, the end stations were clustered based on proximity, and the end station cluster became the dependent variable. Additionally, when doing a regression analysis, it is important to split the data up into a training and testing set. The training set is used to fit the model and then the testing set is used to see if the model works. If you do not split data up into training and testing sets, you risk overfitting the model to the data. The accuracy of the model can be easily evaluated by calculating the score, which evaluates how many predictions were correct out of the data points predicted. However, when the Blue Bikes data was used to create a classification tree to predict the end station cluster that a ride would end in, there was no correlation between the input variables and this output. This is not that surprising, due to the fact that once the repetitive variables were removed, the only variables left were trip duration, start station ID, and customer type. Perhaps there are more variables that Blue Bikes collects but are not available in their public data that would be useful for such an analysis.

Conclusion

Technology has made it possible for cities to know more about what is happening in them than ever before. By fully leaning into technology and Big Data, cities can be transformed into smart cities, which have complex data collection and analysis systems. Two major areas that smart city technology can be used in is smart electricity grids and transportation networks; however, it is also important to consider the risks that come with implementing this level of technology. Smart city development plans should look to cities that have already begun implementing a variety of apps, automated waste disposal systems and more. The development plans should also include plans for governance and details about how data scientists will work to make the data produced by the city usable, understandable, and accessible. Environmental sustainability typically focuses on natural systems, but smart cities show how important it is to use technology to improve livability for people living in cities today and for future generations to come.

Bibliography

Baig, Z., Szewczyk, P., Valli, C., et al., 2017, Future challenges for smart cities: Cyber-security and digital forensics, Digital Investigation, v. 22, p. 3-13, doi: https://doi.org/10.1016/j.diin.2017.06.015.

Barton, D., 2020, 7 Uses for Analytics in Smart Cities: Innovation Enterprise Channels, https://channels.theinnovationenterprise.com/articles/158-7-uses-for-analytics-in-smart-cities (accessed February 2021)

Bingyi, H., Wu, B., Nguyen, Q., Camargo, R., Arancibia, I., 2019, The Threat of Cyber-Terrorism & Security in Intelligent Transporation Systems Architecture: Melbourne School of Engineering, https://eapj.org/wp-content/uploads/2020/02/The-Threat-of-Cyber-Terrorism-Security-in-Intelligent-Transportation-Systems-Architecture.pdf (accessed April 2021)

BlueBikes, 2021, System Data: https://www.bluebikes.com/system-data (accessed March 2021)

Burkhalter, M., 2020, How IoT is helping improve public transportation: https://www.perle.com/articles/how-iot-is-helping-improve-public-transportation-40188811.shtml (accessed April 2021)

De Obseso-Orendain, A., Lopez-Neri, E., Donneaud-Bechelani, C., 2015, The role of the Data Scientist within Smart Cities: IEEE – Smart Cities GDL CCD White Paper, https://smartcities.ieee.org/images/files/pdf/dav_datascientist_v12_final_tjc-eln.pdf (accessed February 2021)

Galarnyk, M., 2019, Understanding Decision Trees for Classification (Python): Medium, Towards Data Science, https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952 (accessed April 2021)

Gohari, S., Ahlers, D., Nielsen, B., Junker, E., 2020, "The Governance Approach of Smart City Initiatives. Evidence from Trondheim, Bergen, and Bodø" *Infrastructures* 5, no. 4, doi: https://doi.org/10.3390/infrastructures5040031

Momoh, J., 2009, "Smart grid design for efficient and flexible power networks operation and control," *2009 IEEE/PES Power Systems Conference and Exposition*, Seattle, WA, USA, 2009, pp. 1-8. doi: 10.1109/PSCE.2009.4840074

Ogleby, G., 2018, 7 ways that Barcelona is leading the smart city revolution: https://www.edie.net/news/7/Seven-ways-that-Barcelona-is-leading-the-smart-city-revolution/ (accessed March 2021)

Reclus, F., Drouard, K., 2009, "Geofencing for fleet & freight management," *2009 9th International Conference on Intelligent Transport Systems Telecommunications, (ITST)*, Lille, France, pp. 353-356. doi: 10.1109/ITST.2009.5399328

Samih, H., 2019, Smart cities and internet of things, Journal of Information Technology Case and Application Research, v. 21:1, p. 3-12, doi: 10.1080/15228053.2019.1587572

Smart City Hub, 2017, What is a smart city? Three examples: http://smartcityhub.com/governance-economy/what-is-a-smart-city/ (accessed March 2021)

Smart Dubai, 2020. Our vision is to make Dubai the happiest city on Earth: https://www.smartdubai.ae/ (accessed March 2021)

Smart Nation Singapore, 2021, Transforming Singapore Through Technology: https://www.smartnation.gov.sg/why-Smart-Nation/transforming-singapore (accessed March 2021)

Wang S.J., Moriarty P., 2019, Energy savings from Smart Cities: A critical analysis: Energy Procedia, v. 158, doi: https://doi.org/10.1016/j.egypro.2019.01.985.

Appendix A – BlueBikes Python Analysis

Available on Github: https://github.com/samanthahoch/bluebikes

```
import numpy as np
import pandas as pd
import datetime
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.cluster import KMeans

class BlueBikes:

    def read_data(self):
        """
        Reads in the blue bikes data and converts the columns with datetimes to time

        Returns:
            A dataframe containing the formatted data
        """

        # data source: https://s3.amazonaws.com/hubway-data/index.html
        blue_bikes_csv = pd.read_csv("data/202102-bluebikes-tripdata.csv")

        blue_bikes_csv['starttime'] = pd.to_datetime(blue_bikes_csv['starttime'])
        blue_bikes_csv['stoptime'] = pd.to_datetime(blue_bikes_csv['stoptime'])
        blue_bikes_csv['starttime'] = [datetime.datetime.time(d) for d in
       blue_bikes_csv['starttime']]
        blue_bikes_csv['stoptime'] = [datetime.datetime.time(d) for d in
       blue_bikes_csv['stoptime']]

        return blue_bikes_csv

    def prep_for_analysis(self, data):
        """
        Prep the given blue bikes data for regression analysis

        Params:
        - data : the data to prep

        Returns:
            The formatted data
        """
        # add cluster labels for end station
        coords = data[['end station latitude', 'end station longitude']].values
        data["end_station_cluster"] = self.make_clusters(coords, 40)
        # remove start and end time
        data = data.drop(['starttime', 'stoptime'], axis=1)

        # remove repetitive information
        data = data.drop(['start station name', 'end station name', 'start station
       latitude', 'start station longitude', 'end station latitude', 'end station
       longitude', 'end station id', "postal code", "bikeid"], axis=1)

        # convert categorical variables to dummies
        categorical_variables = ["usertype"]
        data = pd.get_dummies(data, columns = categorical_variables)

        return data

    def make_clusters(self, coords, num_clusters):
        """
        Uses KMeans to cluster the given coordinates

        Params:
        - coords : the coordinates to cluster
        - num_clusters : the number of clusters to make

        Returns:
            The labels of the clusters
        """
        kmeans = KMeans(n_clusters=num_clusters, random_state=0).fit(coords)
        cluster_labels = kmeans.labels_
        num_clusters = len(set(cluster_labels))
        print('Number of clusters: {}'.format(num_clusters))

        return cluster_labels

    def decision_tree_analysis(self, data, y_variable, save_results=True):
        """
        Performs a decision tree analysis using the given data

        Params:
        - data : the complete dataset
        - y_variable : the dependent variable
        - save_results : save the model results to a file

        Returns:
            The decision tree model
        """

        # select variables
        x = data.drop([y_variable], axis=1)
        y = data[y_variable]

        # split into test and training data
        X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.25,
       random_state=0)

        # use a decision tree regressor to predict
        model = DecisionTreeRegressor().fit(X_train, Y_train)
        prediction = model.predict(X_test)

        # score = correct predictions / total number of data points
        score = model.score(X_test, Y_test)
        print("model score:", score)

        if save_results:
            results_dataset = X_test.copy()
            results_dataset['prediction'] = prediction
            results_dataset['actual'] = Y_test
            results_dataset.to_csv("results_dataset.csv")

        return model

b = BlueBikes()
print("Reading in data...")
data = b.read_data()
print("Data contains", len(data.index), "rows.")
print("Preparing data for analysis...")
data = b.prep_for_analysis(data)
print("Data for analysis contains", data.shape[1], "columns.")
print("Performing decision tree analysis...")
model = b.decision_tree_analysis(data, "end_station_cluster")
print("Analysis complete.")
```