International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

# HOW TO DO UBER DATA ANALYSIS WITH THE HELP OF DATA SCIENCE WITH PYTHON LANGUAGE

[1]Mr.Ravi Rathore , [2]Prof.Rachna Raghuwashi
[1]M.Tech (CSE), [2]Assistant Professor
[1]Computer Science & Engineering
[1]Sagar Institute of Research and Technology,Indore[1]
[1]SAGE University, Indore, India

**Abstract:** As we know that the transportation sector is playing most important role in our life and most of the person want to use personal vehicle to transportation but not having any personal vehicle to transportation so in this case Uber taxi cab company provides the 3 & 4 wheeler vehicle according to the user need and time. According to the facts whenever users uses the own vehicle for the transportation then time is more saves ad easy to go anywhere and anytime as compare to uses private vehicle. Uber company provides this type of facility to the peoples and user feel the comfortable and easy to use. Uber is a ride-hailing company that offers the Uber mobile app, which you can use to submit a trip request that is automatically sent to an Uber driver near to you, alerting the driver to your location. The accepting Uber driver will then come and pick you up and drive you to your requested destination. The Uber app automatically figures out the navigational route for the driver, calculates the distance and fare, and transfers the payment to the driver from your selected payment method, without you having to say a word or grab your wallet. This type of acility works o the mobile based technology. now we try to analysed the user data to Uber uses your personal data in an anonymised and aggregated form to closely monitor which features of the Service are used most, to analyze usage patterns and to determine where we should offer or focus our Service.

**Keywords:-** Data Science, Machine Learning, Uber Company, R Language, Transportation related data

## I. INTRODUCTION

For nearly five decades, liveability has been referenced as a key attribute for community and urban planning worldwide [1], [2]. More recently, it has been firmly placed in the global policy lexicon by its inclusion in three of the principles and commitments of the New Urban Agenda (NUA) adopted by the UN in 2016 [3]. The NUA is notable as it represents a significant international policy commitment in support of the Sustainable Development Goals (SDG), and more specifically SDG11, and what some have referred to as a pro-urban future [4], [5]. SDG11 sets out a goal for the international community to “make cities inclusive, safe, resilient and sustainable”. While it is clear that the authors of the NUA perceived liveability as playing a role in eradicating poverty (Paragraph 14a), and as an indicator of both social inclusion and cohesion (Paragraph 40) and sustainable urban transport and transit systems (Paragraph 114) [3], no where within the NUA or supporting documents the concept of liveability is defined [3], [6]. This is not entirely surprising. Indeed, authors have commented on the widespread use of the term, despite the ambiguity in meaning in policy documents and scholarly articles [7], [8]. According to Newton [7], liveability can be defined as a set of attributes of a place, encompassing housing, neighborhood and region aspects that contribute to residents’ quality of life and well-being. A recent review of the literature on relevant indicators of liveability suggests a broad range of contributory indicators across 11 policy domains (the natural environment, crime and safety, education, employment and income, health and social services, housing, leisure and culture, food and other goods, public open space, transport, social cohesion and local democracy), although the relative importance of each is unclear [1].

Uber is a mobile app taxi program. A car owner with a private car can earn money by verifying that it is a driver of Uber. Passengers can call online or reserve an Uber driver. Passengers must pay for the fare

online.Over time, Uber has developed into a variety of ride modes including carpooling. Its price is generally cheaper than a taxi. Therefore, the development of Uber is very rapid.

## TAXICABS IN NEW YORK CITY:-

Taxi is a common public transportation in the city. Unlike public transportation such as buses and subways, taxi lines are not fixed and random. The density of taxis varies greatly from region to region. It is non-commuting and frequent. The main means of transportation, and its travel path, time, and get-off point information are closely related to human activities, which can better reflect the behavior patterns of urban residents [1]. Taxicab is a very important part of New York City's transportation architecture. They come in two varieties: yellow and green. Taxis painted canary yellow are able to pick up passengers anywhere in the five boroughs. Those painted apple green are allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island. Both types have the same fare structure.

## II MACHINE LEARNING

Machine learning teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases.

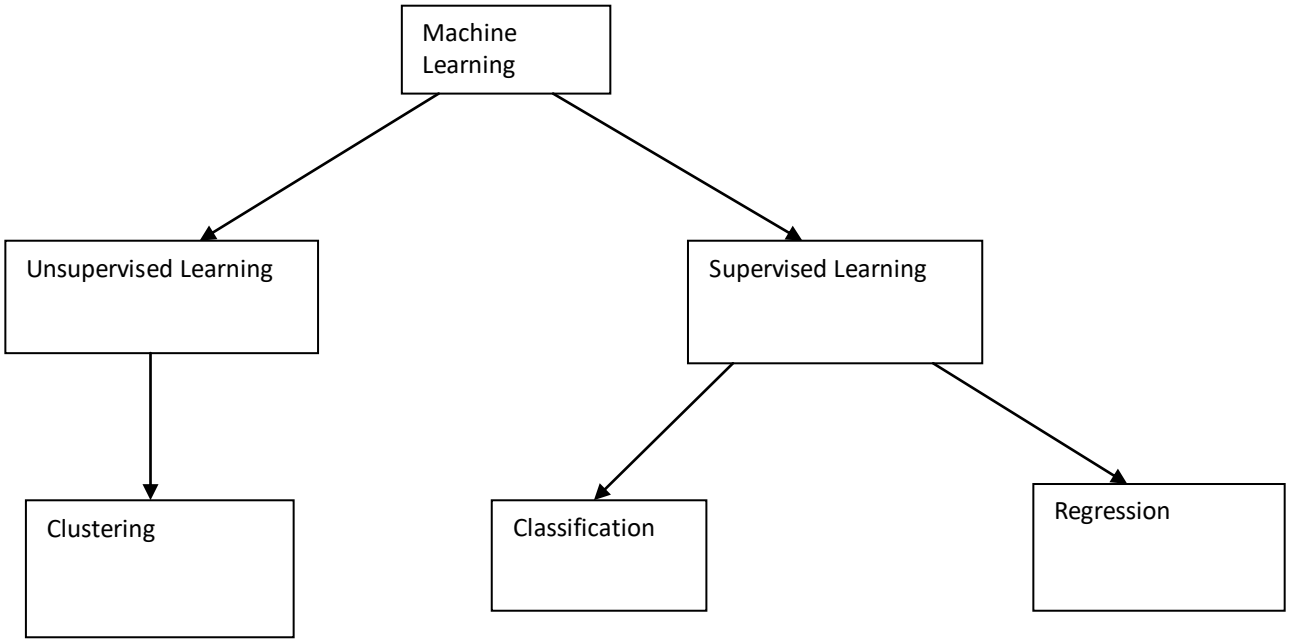
### REAL-WORLD APPLICATIONS

With the rise in big data, machine learning has become particularly important for solving problems in areas like these:

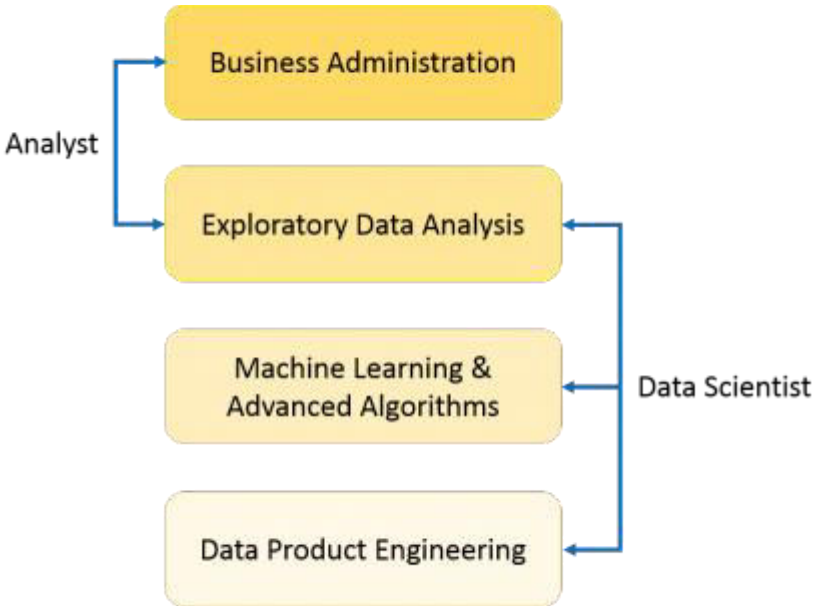
• Computational finance, for credit scoring and algorithmic trading
• Image processing and computer vision, for face recognition, motion detection, and object or diseases detection
• Computational biology, for tumor detection, drug discovery, and DNA sequencing
• Energy production, for price and load forecasting • Automotive, aerospace, and manufacturing, for predictive maintenance
• Natural language processing

### HOW MACHINE LEARNING WORKS

Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.




**DATA SCIENCE:-** Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. But how is this different from what statisticians have been doing for years?




**R LANGUAGE:-** R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is

often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.
One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

## III. RESEARCH AND METHODOLOGY

A. **Data Quality Analysis:** A critical stage in data analysis is to identify and characterize missing data in the target data set. By identifying missing data and taking it in to account, awareness of data quality issues can guide the data handling strategy for further and deeper analyses. A missing data accounting and visualization toolset

B. **Preliminary Mean ETA Analysis by neighborhood:** To understand high level differences between regions and Uber variants, the mean ETAs by service variant and neighborhood was calculated and visualized. Figure 5 shows a grouped bar chart representing mean ETAs for both Uber variants for selected neighborhoods, for both normal and rush hour periods. For these neighborhoods, the chart highlights an overall consistent mean ETA difference between Uber variants, while only a small deviation in mean ETAs between normal and rush hour periods. This suggests that choice of service variant may be more impacting in ETA than the choice of taking service at normal or in rush hours.

C. Time-**Series Analysis of Hourly and Daily Mean ETAs:** Mean ETA data was plotted (i) hourly by Uber service variant (see Figure 6, and (ii) daily by both service variant and region for the month of February 2018 (see Figure 7). The former illustrates, as expected, an increase in mean ETAs during rush hour periods that explained by increased traffic, traffic jams and demand for the Uber Service. Also as expected, a great increase in ETA is observed late in the night, probably due to a lower demand or low availability of drivers. The latter month-wide analysis is more insightful. The clustered and distinct curves in Figure 7 reveals that different city regions have a different Uber ETA profile. Visual peak detection analysis clearly identifies numerous peaks that cursory desk research can reveal to be context-sensitive. They are related to (i) city-wide events, in this case, Carnival, and (ii) weather events e.g. Natal experienced a very rainy period in February 2018. Increased ETAs during Carnival may a result of demand surges from tourists or lower driver supply during the holiday period. Increased ETAs during rainy days may indicate infrastructural or public transit problems in the city. This analysis in particular highlights both the potential but limits of Uber data for urban livability research.

D. **Spatial Analysis:** Once collected, ETA estimates depend on geographic coordinates, a choropleth map in which neighborhoods are colored according to their mean ETA measurements can be constructed with the folium [42] Python library. A choropleth map constructed from the geographical features of Natal and Uber X ETA data is presented in Figure 8. It visually ratifies some findings of previous analysis. For example, it is possible to notice that northern and western neighborhoods are subject to higher waiting times. Note that Uber X variant was chosen for this choropleth because it is the most used variant among Uber products in Brazil [43].

E. **Uber as an Urban Livability Index:** Correlational Analysis Data analysis up to this point yielded a reasonable ground for suspecting that Uber service performance may be influenced by socio-economic factors and that Uber ETA has potential as an urban livability indicator. We have shown that Uber data is relatively easy to collect in real-time, can be manipulated and visualized so that it can be easily understood, provides different levels of geographic and temporal granularity, and is context-sensitive. To explore its potential as an urban livability indicator further, the results of Araujo and C ´ andido [15], [16] study on ˆ urban expansion and quality of life in Natal (ULQI) discussed in Section II and demographic information (population and density) on Natal (2016) [32]

were combined to build a Natal Quality of Life (NQoL) data set. An illustrative sample of such data is presented in Table II. By correlating the Uber ETA data and the NQoL data sets, a preliminary assessment of the utility of Uber ETA data as an urban livability indicator can be assessed[4].

**IV. LITERATURE REVIEW**

Two distinct sources of related works are of interest to this study - publications related to urban liveability indicators 1 and those related to using Uber data. The Economist Intelligence Unit (EIU) Global Liveability Index and the Mercer Quality of Living Ranking are two indices referenced widely in policy, media and academic literature. The EIU Global Liveability Index is an annual rating of 140 cities for relative comfort based on 30 qualitative and quantitative factors across five broad weighted categories (stability, healthcare, culture and environment, education, and infrastructure) constructed using a combination of external data points and the judgment of a group of in-house and external analysts [2]. It is primarily used for employee mobility. Similarly, the Mercer Quality of Life Ranking evaluates living conditions in 450+ cities worldwide based on 39 factors, grouped in 10 categories - political and social environment, economic environment, socio-cultural environment, medical and health considerations, schools and education, public services and transportation, recreation, consumer goods, housing and natural environment. Scores are weighted to reflect their importance to expatriates. Like [2], the primary focus is to support decisions in relation to employee mobility. It should be noted that Mercer do also offer services to municipalities to assess factors that can improve their quality of living ranking [13]. Recognizing the need for standardization of city indicators, the World Bank initiated the Global City Indicators Program (GCIP) in 2006, which is comprised, at the time of writing, of 27 core and 36 supporting indicators [19]. Unlike, the purpose of the GCIP indices is to inform policy making and urban planning. The Global City Institute Facility (GCIF) manages the GCIP dataset and claims to have data from 255 member cities from 82 countries. They go on to note that the GCIP was the framework for ISO 37120, the first international standard on city metrics2 . Cities report indicators annually and the benchmark data is provided by the GCIF including verification services. More recently, the Global Liveable Cities Index (GLCI) has been proposed [20]. The GLCI comprises five categories of indicators e.g. economic vibrancy and competitiveness, environmental friendliness and sustainability, domestic security and stability, socio-cultural conditions, and political governance [20]. The GLCI places more emphasis on governance than, for example, the GCIP and therefore in their reported analysis of 64 cities, findings are significantly different. Notably, the GLCI seeks to further categorize cities in terms of their attractiveness to different personality types and in this way, it integrates a citizen-centric approach however the scientific rigour behind this classification is lacking in detail.Of specific relevance to this paper, the recent works by Araujo and C ´ andido [15], [16] outline the development of ˆ ULQI and its application to Natal and its districts. Unlike the indices above, the ULQI does not use any subjective data and is based on four variables (Urban Environmental Infrastructure, Urban Equipment and Services, Socio-economic, and Safety) comprising 23 quantitative indicators [16]. There are a number of observations to note with regards to the indices above. Firstly, they all comprise indicator sets. As such, they can be complex and difficult to interpret [21]. Secondly, they are all either annual or ad hoc (in the case of GLCI and ULQI) and therefore represent an assessment at a single point in time and are not sensitive to temporal variations. Thirdly, with the exception of ULQI, they comprise subjective and objective data and weighting. While [20] attempts comparison, it is unclear whether this is methodologically sound. Fourthly, for international comparisons, they all require significant time, effort and funding to collect data. Regarding research using Uber data, no scientific research on urban liveability using Uber data was identified. This is not surprising as the Uber API has only be available since 2016 and its use for liveability research is relatively novel. The Transportation Sustainability Research Center at the University of California, Berkeley has published a series of papers on the impact of on-demand ride services and its role and impact on urban transportation but not as an indicator for liveability per se [22], [23]. Similarly, there is a large number of articles on the impact of Uber on competition, the workforce and the need for regulation (see, for example, [24]–[26]), however these are not particularly relevant to our research topic. Research studies using the Uber API typically fall in to a small number of topic clusters, namely understanding pricing phenomena [27], optimizing

itineraries and payoffs [28], consumer research [29], competitiveness research [30], and to build new products and services [31].

## V. PROPOSED ARCHITECTURE OF SYSTEM TO BE BUILT

This architecture is presented to implementation approach of our system –

The solution of the above problem is to make a software or website. We can understand by the following system architecture of the given problem.

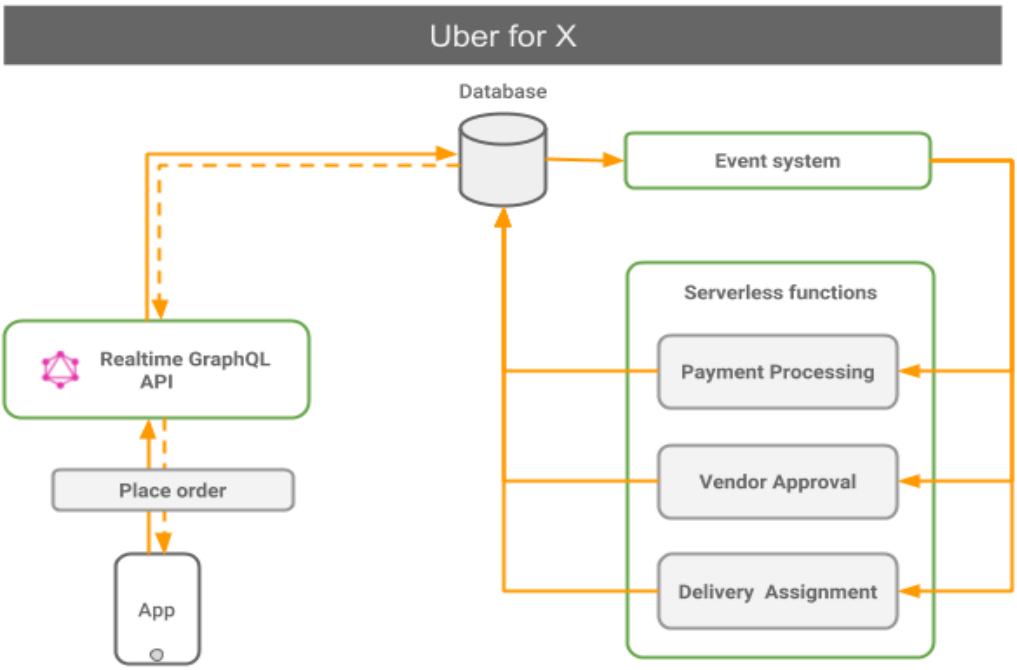



Fig.1: System Architecture

**DATA FLOW PROCESS:**

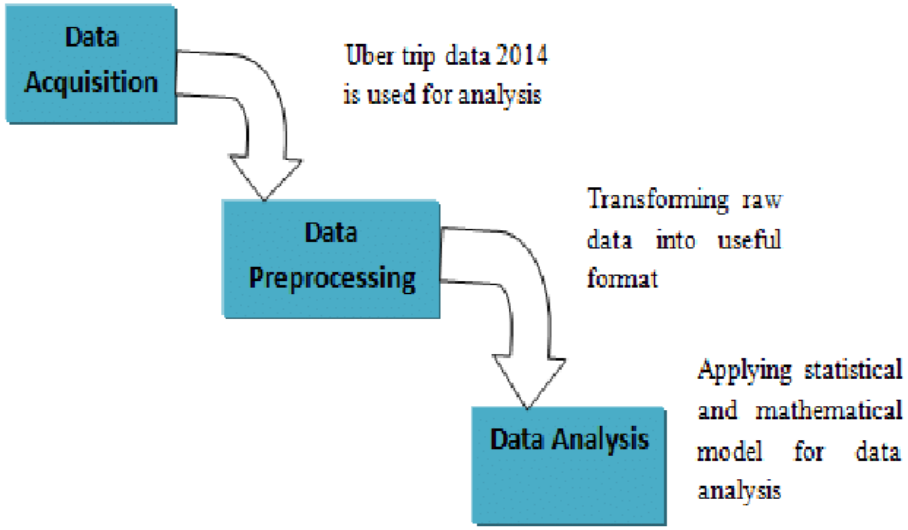



Fig.2:Data Flow Diagram

**VII. CONCLUSION**

Uber is using real-time Big Data to perfect its processes, from calculating Uber's pricing to finding the optimal positioning of taxis to maximize profits. Real-time data analysis is very challenging for the implementation because we need to process data in real-time, if we use Big Data, it is more complex than before. Implementation of real-time data analysis by Uber to identify their popular pickups would be advantageous in various ways. It will require high-performance platform to run their application. So far no research has been done on real-time analysis for identifying popular Uber locations within Big Data in a distributed environment, particularly on the Kubernetes environment. To address these issues, we have created a machine learning model with a Spark framework to identify the popular Uber locations and use this model to analyze real-time streaming Uber data and deploy this system on Google Dataproc with the different number of worker nodes with enabling Kubernetes and without Kubernetes environment. With the proposed Kubernetes environment and by increasing the worker nodes of Dataproc clusters, the performance can be significantly improved. The future development will consist of visualizing the real-time popular Uber locations on Google map.

**REFERENCES:**

[1] M. Lowe, C. Whitzman, H. Badland, M. Davern, L. Aye, D. Hes, I. Butterworth, and B. Giles-Corti, "Planning healthy, liveable and sustainable cities: how can indicators inform policy?" Urban Policy and Research, vol. 33, no. 2, pp. 131–144, 2015.
[2] The Economist Intelligence Unit, "Global liveability index 2018," 2018.
[3] United Nations, "New urban agenda," 2016.
[4] C. Barnett and S. Parnell, "Ideas, implementation and indicators: epistemologies of the post-2015 urban agenda," Environment and Urbanization, vol. 28, no. 1, pp. 87–98, 2016.
[5] F. Caprotti, R. Cowley, A. Datta, V. C. Broto, E. Gao, L. Georgeson, C. Herrick, N. Odendaal, and S. Joss, "The new urban agenda: key opportunities and challenges for policy and practice," Urban research & practice, vol. 10, no. 3, pp. 367–378, 2017.
[6] United Nations, "Glossary of the habitat iii," 2016.

[7] P. W. Newton, "Liveable and sustainable? socio-technical challenges for twenty-first-century cities," Journal of Urban Technology, vol. 19, no. 1, pp. 81–102, 2012.
[8] M. Ruth and R. S. Franklin, "Livability for all? conceptual limits and practical implications," Applied Geography, vol. 49, pp. 18–23, 2014.
[9] H. J. Miller, F. Witlox, and C. P. Tribby, "Developing context-sensitive livability indicators for transportation planning: a measurement framework," Journal of Transport Geography, vol. 26, pp. 51–64, 2013.
[10] A. Ley and P. Newton, "Creating and sustaining liveable cities," in Developing living cities: From analysis to action. World Scientific, 2010, pp. 191–229.
[11] R. Cervero, Transit-oriented development in the United States: Experiences, challenges, and prospects. Transportation Research Board, 2004, vol. 102.
[12] D. Sauter and M. Huettenmoser, "Liveable streets and social inclusion," Urban Design International, vol. 13, no. 2, pp. 67–79, 2008.
[13] Mercer, "Quality of life rankings 2018," 2018
[14] Uber Technologies, Inc., "Facts and figures," 2018.
[15] M. C. C. Araujo and G. A. C ´ andido, "Qualidade de vida e sustentabil- ˆ idade urbana," HOLOS, vol. 1, no. 0, p. 3, jan 2014.
[16] ——, "´Indices de qualidade de vida urbana de Natal-RN," Geoconexoes ˜ , vol. 1, no. 1, p. 51, 2015.
[17] Z. Wendling, M. Levy, D. Esty, A. de Sherbinin, and J. Emerson, "The 2018 environmental performance index," https://epi.envirocenter.yale.edu/, Last accessed on 2018-01-02.
[18] K. Schwab, "The global competitiveness report 2018," 2018.
[19] P. Bhada and D. Hoornweg, "The global city indicators program: A more credible voice for cities," 2009.
[20] T. K. Giap, W. W. Thye, and G. Aw, "A new approach to measuring the liveability of cities: the global liveable cities index," World Review of Science, Technology and Sustainable Development, vol. 11, no. 2, pp. 176–196, 2014.
[21] P. Zhou and B. Ang, "Indicators for assessing sustainability performance," in Handbook of performability engineering. Springer, 2008, pp. 905–918.
[22] L. Rayle, D. Dai, N. Chan, R. Cervero, and S. Shaheen, "Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco," Transport Policy, vol. 45, 01 2016.
[23] O. Flores and L. Rayle, "How cities use regulation for innovation: the case of uber, lyft and sidecar in san francisco," Transportation research procedia, vol. 25, pp. 3756–3768, 2017.
[24] S. Wallsten, "The competitive effects of the sharing economy: how is uber changing taxis," Technology Policy Institute, vol. 22, 2015.
[25] H. A. Posen, "Ridesharing in the sharing economy: Should regulators impose uber regulations on uber," Iowa L. Rev., vol. 101, p. 405, 2015.
[26] B. Rogers, "The social costs of uber," U. Chi. L. Rev. Dialogue, vol. 82, p. 85, 2015.
[27] J. Jiao, "Investigating uber price surges during a special event in austin, tx," Research in Transportation Business & Management, 2018.
[28] H. A. Chaudhari, J. W. Byers, and E. Terzi, "Putting data in the driver's seat: Optimizing earnings for on-demand ride-hailing," in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018, pp. 90–98.
[29] J. C. Correa, "Urban mobility social networks as valid sources for collaborative consumption research," 2018.
[30] L. K. Poulsen, D. Dekkers, N. Wagenaar, W. Snijders, B. Lewinsky, R. R. Mukkamala, and R. Vatrapu, "Green cabs vs. uber in new york city," in IEEE 2016 IEEE International Congress on Big Data, 2016.

[31] X. Zhou, M. Wang, and D. Li, "From stay to play–a travel planning tool based on crowdsourcing user-generated contents," Applied Geography, vol. 78, pp. 1–11, 2017. [32] Prefeitura Municipal de Natal, "Anuario de Natal 2016," Prefeitura ´Municipal de Natal, Natal, Tech. Rep., 2016.
[33] F. Kooti, M. Grbovic, L. M. Aiello, N. Djuric, V. Radosavlejevic, and K. Lerman, "Analyzing uber's ride-sharing economy," in 26th Conference on World Wide Web Companion, no. January, 2017, pp. 574– 582.
[34] Uber Technologies Inc., "Introduction to the api," 2018, https://developer.uber.com/docs/riders/riderequests/tutorials/api/introduction, Last accessed on 2018-12-01.
[35] J. W. Tukey, Exploratory data analysis. Reading, Mass., 1977, vol. 2.
[36] J. T. Leek and R. D. Peng, "What is the question? Mistaking the type of question being considered is the most common error in data analysis," Science, vol. 347, no. 6228, pp. 1314–1315, 2015.
[37] C. O'Neil and R. Schutt, Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.", 2013.
[38] C. Weihs, "Multivariate exploratory data analysis and graphics: A tutorial," Journal of Chemometrics, vol. 7, no. 5, pp. 305–340, 1993.
[39] W. Mckinney, "pandas: a foundational python library for data analysis and statistics," Python High Performance Science Computer, 01 2011.
[40] J. D. Hunter, "Matplotlib: A 2d graphics environment," Computing In Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.
[41] A. Bilogur, "Missingno: a missing data visualization suite," The Journal of Open Source Software, vol. 3, no. 22, p. 547, 2018.
[42] R. Story, "Folium documentation," 2018. [Online]. Available: https://python-visualization.github.io/folium/
[43] Uber Technologies Inc., "Descubra o que e e como funciona o uberx," ´ 2018, https://www.uber.com/pt-BR/blog/o-que-e-uberx/, Last accessed on 2018-12-01.
[44] Biokit Developers, "Biokit release 0.4.4," 2018. [Online]. Available: https://biokit.readthedocs.io