Math 5545 Project 1 (Due: February 18, 2021)
Statistical Analysis of Kansas City 311 Data

Group # 1: Pre-COVID – Warm Season

* Ford, William Andrew waftx2@mail.umkc.edu
** Tran, Thao Phuong ttcww@mail.umkc.edu
Johnson, Reece LaVaughn rlj6pb@mail.umkc.edu (Analysis of Low-income data)
Whetsell, Torsten P torsten.whetsell@mail.umkc.edu (Analysis of Moderate -income data)
Salas,J acob R jrsty8@mail.umkc.edu (Analysis of High-income data)

Group # 2: Pre-COVID – Cold Season

* Vaughn, Braeden bv2my@mail.umkc.edu
**Stack, Caston A cas8y5@mail.umkc.edu
Terry, Harrison Edward Lee het9t5@mail.umkc.edu (Analysis of Low-income data)
Wilson, Ben bmwppf@mail.umkc.edu (Analysis of Moderate -income data)
Schaeffer, Alex as9nb@mail.umkc.edu (Analysis of High -income data)

Group # 3 COVID – Warm Season

* Lim, Celine Shwu Ling csl9r3@mail.umkc.edu
**Reesman, Grace Ellen gekhx5@mail.umkc.edu
Rosenblatt, Jennifer jrosenblatt@mail.umkc.edu (Analysis of Low-income data)
Thomas, Micheal mctwcc@mail.umkc.edu (Analysis of Moderate-income data)
Aljofei, Maha Mashan mmawv6@umkc.edu (Analysis of High-income data)

*** team leader.** The team leader is responsible for organizing the zoom meetings, distributing the work between the team members coordinating with the project manager and making sure that the project is getting finished before the deadline. If you do not receive an email from your team leader by the end of this week, please let me know by Monday morning.

**** project manager**. the project manager is responsible for the quality of the project, making sure all calculations are correct, coordinating with the team leader and submitting all required work.

Collegiality and Group Work

The groups in this class are meant to imitate real-world research groups. Each group should regularly meet via zoom. Each group member should (1) maintain a friendly environment for the entire group; (2) facilitate collaboration and problem solving; (3) provide a vision of the main objectives and ensure discussions lead to conclusions and decisions; (4) motivate and inspire other group members; (5) contribute to the group by sharing his/her knowledge, expertise, and viewpoints; (6) participate in all meetings and discussions; (7) have productive suggestions.

Instructions for preparing your Slides:

(a) Each team is required to prepare 10-15 PowerPoint slides. You can use the template slide that is provided with this project. Please keep it professional and avoid any decorative picture in your slides. A team member must make a Panopto video and present their slides. The video should be less than 15 minutes. Here are the instructions for Creating and Uploading Panopto Videos to Canvas: https://online.umkc.edu/support/panopto-support/34995-2/

Here is a video about presenting your slides: https://www.youtube.com/watch?v=dEDcc0aCjaA

(b) The contents of the slides should include: Overview, Our Team, Introduction, Overarching Goals, Materials & Methods, Statistical Findings, Concluding Remarks, Limitations and References

(c) The slide "Our Team" must contain the name and student ID of each participant. Also include the specific works (e.g preparing slides, computing, editing and revising) done by each participant.

(d) All figures and tables must have labels and captions.

(e) Supplementary slides (e.g., further explanations and references) should be included in the same file after the presentation slides.

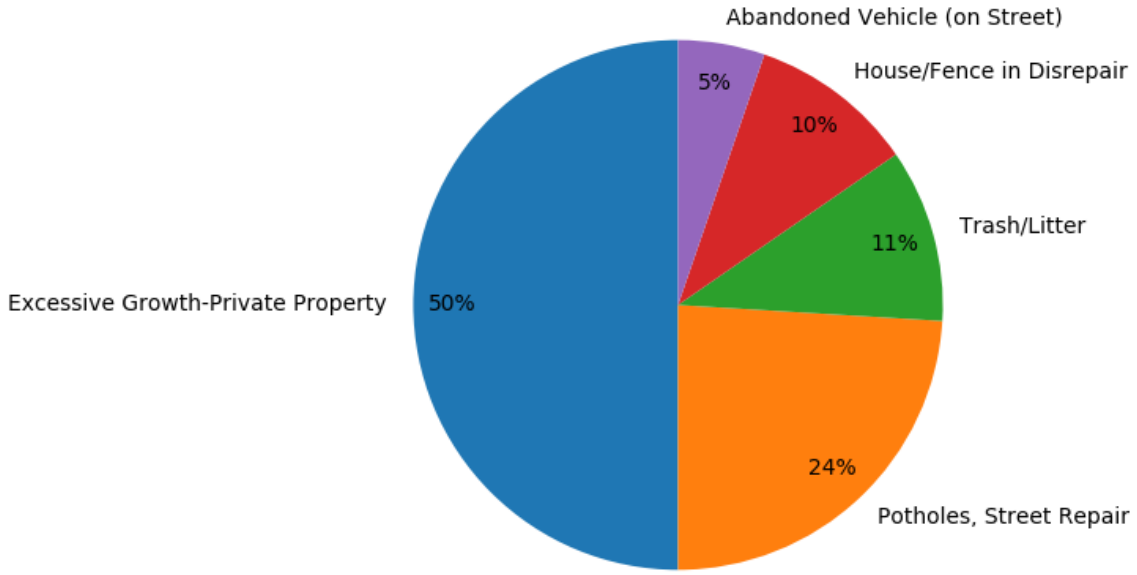
(f) Please send your presentation to baniyaghoubm@umkc.edu before the due date.

Note: In addition to PowerPoint slides, there are some Excel files that you must complete the work for each one of them and upload them to Canvas

Statistical Analysis of Kansas City 311 Data

Introduction: The data collected by a 311 service speaks volumes about a city's quality of life, how well it communicates with its residents, and how effectively it responds to problems. It is important to note that this data is not a perfect representation of cities and may contain gaps.

Each new Open311 request is assigned a category, which explains generally what the request is related to. The exact categories vary between cities, but all requests fall into one category. This makes it possible for us to group and count requests by category. For example, this pie chart shows the most frequently reported categories for Peoria, Illinois:



Further reading: Open Data Shines a Light on Cities' Top Challenges https://nextcity.org/daily/entry/open-data-shines-a-light-on-cities-top-challenges#:~:text=The%20data%20collected%20by%20a,cities%20and%20may%20contain%20gaps.

Many North American cities, including New York City, Chicago, Toronto, Washington, DC, and Kansas City, have implemented 311 systems to manage citizen complaints and service requests [9, 7, 6, 5]. These platforms provide 24/7 customer service to collect citizen reports and to disseminate them to relevant city departments for improved city services [1, 5]. As such, 311 can be a crucial link between citizens and government and represent an example of co-production through digital technology [9]. Because these citizen reports provide a real-time condition assessment of the city, local governments are analyzing these data to understand and forecast problems, service demands, and quality-of-life issues, such as rodent infestations, illegally converted buildings, and heat and hot water outages [5, 10].

Local governments are increasingly turning to '311' citizen complaints and service reports to provide a real-time condition assessment of the city. When combined with machine learning and predictive analytics, these data can be mined to gain new insight into city service needs and potential problems.

Further reading: Who Calls for Help? https://data.bloomberglp.com/company/sites/2/2018/09/Who-Calls-for-Help.pdf

Description of Kansas City 311 Data:

Kansas City Mo has 240 the neighborhoods and 311 data of complaints are available from the year 2007 to the present time. The data includes the following variables (description of each variable has been provided inside the parenthesis).

Part 1) List of Variables

1. CASEID (case ID number)
2. ADDRESS
3. NEIGH (name of the neighborhood)
4. ZIP (zip code)
5. POLICED (Police Department)
6. COUNCILD (Council Department)
7. SOURCE (Source of data: was the complaint received by phone or other methods, it consists of PHONE, WEB, EMAIL, TWIR, BOT, KCEPD, INSPE, SYS, MAIL, nan, WALK, FAX, EIP, EDC, BIZ, KCSPD, KCEPS, KCMPD, CMO, VOICE, CTI, SPNSH)
8. CATEGORY (There are 72 different categories of complaints or violations, the categories are listed in part 2)
9. TYPE (Details of each category is given by type. For instance, if category is animal, then the type could be “bite” or "Dead Animal")
10. CATTYPE (category type is a combination of variables 8 and 9)
11. GROUP311 (14 groups of different categories, see part 3)
12. VIS_STREET (visible or non-visible item)
13. DEPT (different departments of the city)
14. WORKGRP (different work groups of the city)
15. REQTYPE (request type)
16. DETAIL (details of the request)
17. CREATEDATE (date the request was created)
18. CREATTIME (time the request was created)
19. CREATEMO (month the request was created)
20. CREATEYR (year the request was created)
21. STATUS (status of the request: resolved, open, duplicate, assigned, cancelled)
22. EXCEEDTIME (did the request exceed time? Y/N)
23. CLOSEDATE (the date that request was closed)
24. CLOSEMO (the month that request was closed)
25. CLOSEYR (the year that request was closed)
26. DAYTOCLOSE (number of days until the case was closed)
27. ADDGEOC (geography of the address)
28. COUNTY
29. Latitude
30. Longitude
31. APN (Access Point Name)
32. CASEURL (URL of the case)

Part 2) List of all categories

CATEGORY
Animals / Pets
Capital Projects
Lights / Signals
Mowing / Weeds
Parks & Recreation
Property / Buildings / Construction
Public Health
Sidewalks / Curbs / Ditch
Signs
Storm Water / Sewer
Streets / Roadways / Alleys
Trash / Recycling
Government
Public Safety
City Facilities
Data Not Available

Part 3) List of all Groups

GROUP311
ANIMAL
CITY_SERVICES
ROADS_LIGHTS_SIGNS
MOWING_WEEDS_TRASH
PARK_MAINT_RECREATION
PROPERTY_DANGEROUS_BLDG
PROPERTY
PUBLIC_HEALT_SAFETY_WORKS
SIDEWALK
SEWER_WASTE_STORMWATER
TRASH_RECYCLING
TRASH
GRAFFITI
VEHICLE_PARKING

Notes: We have deployed the 311 call complete data from 2007 – 2020):
http://ec2-3-93-220-250.compute-1.amazonaws.com:8000/

The above visualization application will be available in the Vercel server:
https://nsf-scc.vercel.app/neighborhoods

Research Questions and Objectives:

The temporal variations in 311 data can significantly influence the accuracies of several statistical results. Particularly, the quality of life and the nature of 311 calls in a neighborhood can substantially change in the time interval of 2007-2020 (more than a decade). Therefore, we are focusing on three recent time intervals:

a) Pre-COVID-19 interval of 3/1/2019 – 9/1/2019 warm season – **Assigned to Group 1**
b) Pre-COVID-19 interval of 9/1/2019 – 3/1/2020 cold season – **Assigned to Group 2**
c) COVID-19 interval of 3/1/2020 – 9/1/2020 warm season – **Assigned to Group 3**

Overarching goals: The **main objectives** of this project are to (1) provide **descriptive statistics** of data for each time interval and (2) provide **hypothesis testing** of mean values using one-way and two-way ANOVA.

Research questions: We have data related to socio-economic factors of each Kansas City neighborhood (see file "SocioEconomic.xlsx"). We would like to know if the income level, season, or COVID-19 substantially changes the type and quality of city services. Here, the quality of city services is measured by time to complete a service (see variable "DAYTOCLOSE"). Some **research questions** for this project are: What type of services are requested in each of high-income, low-income, and moderate-income neighborhoods? What are the counts, percent and frequency of each service? Is there differences in there frequency central tendency and dispersion with respect to income level? If we observe any difference, then we can test certain hypotheses. Are the mean number of days to complete services the same in each of high- low- and moderate- income neighborhoods? Are the average number of service requests different with respect to high- low- and moderate- income neighborhoods?

Long term goals: In the next project we will compare the results of each group. We would like to answer the following questions in the next project. Does it make any difference with respect to cold or warm seasons? What about COVID-19? Are the number of service requests and the time to complete each service have changed after COVID-19? Do the low-income neighborhoods receive the same services as the high-income neighborhoods?

Anticipated results: The initial analysis of the temporal 311 data could indicate three main patterns:

1. The number of 311 calls and the type of 311 calls vary based on the income level.
2. The number of 311 calls and the type of 311 calls change from cold to warm season.
3. The number of 311 calls and the type of 311 calls has changed with respect to COVID-19.

Method:

A) Categorize Neighborhoods based on income level: We have divided the neighborhoods in High-, Moderate-, and Low-income neighborhoods. To define the income intervals, we used the following methodology. The median household income in Kansas City is $54.4 K. This is known as Area Median Income (AMI). A low-income neighborhood is a neighborhood with median household income about 50% or less of the AMI (or average income for the community [1]). A moderate-income neighborhood is a neighborhood with annual median income about 50% below and 50% above the AMI [2]. Hence, we arrive to high moderate low-income intervals as shown below.

1. Low-income neighborhoods = ($14.6k, $28.3k) **Note: we applied 48%**
   There are 28/240 low-income neighborhoods (about 12%)
2. Moderate-income neighborhoods = [$28.3, $80.5k) **Note: we applied 48%**
   There are 163/240 moderate-income neighborhoods (about 68%)
3. High-income neighborhoods = [$70.46k, $250k+)
   There are 49/240 high-income neighborhoods (about 20%)

Additional Notes:

1. The average number of persons per household in Kansas City during 2015-2019 was 2.35
2. 81.5% of residents Lived in same house for at least a year during 2015-2019

[1] https://www.learncra.com/our-new-infographic-explains-what-is-low-or-moderate-income-or-lmi/#:~:text=Using%20that%20same%20guideline%2C%20a,to%20be%20considered%20moderate%2Dincome.

[2] Former Secretary of Labor Robert Reich suggests that the middle class should be defined as households making between 50% below and 50% above the median. Moyers on Democracy. "By the Numbers: The Incredibly Shrinking American Middle Class." https://billmoyers.com/2013/09/20/by-the-numbers-the-incredibly-shrinking-american-middle-class/

B) Descriptive statistics: Descriptive statistics allow you to characterize your data based on its properties. There are four major types of descriptive statistics:

1. **Measures of Frequency:**
* Count, Percent, Frequency

* Shows how often something occurs.
* Use this when you want to show how often a response is given.

2. **Measures of Central Tendency**
* Mean, Median, and Mode

* Locates the distribution by various points.
* Use this when you want to show how an average or most indicated response.

3. **Measures of Dispersion or Variation**
* Range, Variance, Standard Deviation

* Identifies the spread of scores by stating intervals.
* Range = High/Low points

* Variance or Standard Deviation = difference between observed score and mean
* Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean.

4. **Measures of Position**
* Percentile Ranks, Quartile Ranks

* Describes how scores fall in relation to one another. Relies on standardized scores
* Use this when you need to compare scores to a normalized score (e.g., a national norm)

Tutorials:

1. Descriptive Stat: https://www.excel-easy.com/examples/descriptive-statistics.html
2. Boxplots: https://www.real-statistics.com/descriptive-statistics/box-plots/
3. Excel Histogram Charts and FREQUENCY plots https://www.myonlinetraininghub.com/excel-histogram-charts-and-frequency-function

C) Hypothesis testing: A statistical hypothesis is a hypothesis that is testable based on observed data modelled as the realized values taken by a collection of random variables.

One-way analysis of variance is a technique that can be used to compare means of two or more samples. This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or categorical input data, the "X", always one variable, hence "one-way"

A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. Use a two-way ANOVA when you want to know how two independent variables, in combination, affect a dependent variable.

Tutorials:

1. Introduction to ANOVA: https://www.youtube.com/watch?v=uzcqMeNK7Kw
2. How to do One-Way ANOVA using Excel: https://www.youtube.com/watch?v=iOoPHunv8NM
3. How to do One-way ANOVA in Excel with post-hoc t-tests: https://www.youtube.com/watch?v=tPGPV_XPw-o
4. How to run a Two Way ANOVA in Excel With Replication https://www.youtube.com/watch?v=0BirE9K05i0

Procedure:

Step 1) Divide the Excel file of each group into three Excel files of high-, low-, and moderate-income neighborhoods. Here is how to do it: (a) Use the excel file "SocioEconomic.xlsx" to change the entries of column D "NEIGH-Income" to H, M, or L (i.e., high-, moderate- and low-income neighborhoods). Use excel "find and replace" function to find each neighborhood name and replace it with H, M or L (b) sort the file based on column D (make sure that you choose "expand the selection"), (c) copy all rows with column D = L and paste them in a new excel file. Repeat this for rows with column D= M and H to make the other two files.

Output of Step 1)

For Group 1: File "Group3-311DataPreCOVID-Warm Season.xlsx" must be divided into

"H-Group3-311DataPreCOVID-Warm Season.xlsx"
"L-Group3-311DataPreCOVID-Warm Season.xlsx"
"M-Group3-311DataPreCOVID-Warm Season.xlsx"

For Group 2: File "Group2-311DataPreCOVID-Cold Season.xlsx" must be divided into

"H-Group2-311DataPreCOVID-Cold Season.xlsx"
"L-Group2-311DataPreCOVID-Cold Season.xlsx"
"M-Group2-311DataPreCOVID-Cold Season.xlsx"

For Group 3: File "Group3-311DataPreCOVID-Warm Season.xlsx" must be divided into

"H-Group3-311DataPreCOVID-Warm Season.xlsx"
"L-Group3-311DataPreCOVID-Warm Season.xlsx"
"M-Group3-311DataPreCOVID-Warm Season.xlsx"

Step 2) Fill column I, for each of new excel files H-, L- and M-, fill column I "Sum SOURCE" using the countif excel function.

=COUNTIF(H2:H66910,H2:H66910)

Notes:

a) Watch the following video to learn how to count the Number of Occurrence of a Text or Number in Excel https://www.youtube.com/watch?v=VsLwqzMIC4k
b) The last row number in your file is not 66910. Before applying countif find out what the last row number is.

Record the outputs of Step 2 in the Excel file "Source Assignment.xlsx"

Step 3) For each of new excel files H-, L- and M-, fill column K "Sum CATEGORY" using the countif excel function.

=COUNTIF(J2:J66910,J2:J66910)

Record the outputs of Step 3 in the Excel file "Category Assignment.xlsx" columns B, D and C.

Step 4) For each of new excel files H-, L- and M-, fill column M "Sum DAYTOCLOSE" using the countif excel function.

=COUNTIF(L2:L66910,L2:L66910)

Note: we will not use this column in our analysis

Step 5) Sort the excel files H-, L- and M-, based on column J "CATEGORY" (make sure that you choose "expand the selection"). Now select the cells in column L "DAYTOCLOSE" for each category. Read the Average and the Sum values at the bottom bar and **enter them in columns E-J of the file Category Assignment.xlsx**

Step 6 [you can do this step after you are done with everything else]) Upload the following completed files to CANVAS (no individual uploading is needed. Only one submission per group is needed. Submit them all **as a single zip file**)

For Group 1:

1. "H-Group3-311DataPreCOVID-Warm Season.xlsx"
2. "L-Group3-311DataPreCOVID-Warm Season.xlsx"
3. "M-Group3-311DataPreCOVID-Warm Season.xlsx"
4. Source Assignment.xlsx
5. Category Assignment.xlsx
6. PowerPoint slides

For Group 2:

1. "H-Group2-311DataPreCOVID-Cold Season.xlsx"
2. "L-Group2-311DataPreCOVID-Cold Season.xlsx"
3. "M-Group2-311DataPreCOVID-Cold Season.xlsx"
4. Source Assignment.xlsx
5. Category Assignment.xlsx
6. PowerPoint slides

For Group 3:

1. "H-Group3-311DataPreCOVID-Warm Season.xlsx"
2. "L-Group3-311DataPreCOVID-Warm Season.xlsx"
3. "M-Group3-311DataPreCOVID-Warm Season.xlsx"
4. Source Assignment.xlsx
5. Category Assignment.xlsx
6. PowerPoint slides

Step 7) Provide descriptive statistics all the data using the files "Source Assignment.xlsx" and "Category Assignment.xlsx" Note: all frequency plots, histograms, pie diagrams, and other descriptive statistics must be well organized and included in the PowerPoint slides.

Step 8) Copy and paste column L "DAYTOCLOSE" of H, M and L file in a new excel file. Apply One-way ANOVA in Excel with post-hoc t-tests t to test the hypothesis if the mean of the high-income low-income and moderate-income samples are the same (null hypothesis). Notes: (1) The ANOVA table and the tested hypothesis must be included in the slides. (2) How to do One-way ANOVA in Excel with post-hoc t-tests: https://www.youtube.com/watch?v=tPGPV_XPw-o

Notes:

1. Statistical tests, such as analysis of variance (ANOVA), assume that although different samples can come from populations with different means, they have the same variance. Equal variances (homoscedasticity) is when the variances are approximately the same across the samples. Unequal variances (heteroscedasticity) can affect the Type I error rate and lead to false positives. If you are comparing two or more sample means, as in the 2-Sample t-test and ANOVA, a significantly different variance could overshadow the differences between means and lead to incorrect conclusions.
2. Do not be too quick to switch to using the nonparametric Kruskal-Wallis ANOVA (or the Mann-Whitney test when comparing two groups). While nonparametric tests do not assume Gaussian distributions, the Kruskal-Wallis and Mann-Whitney tests do assume that the shape of the data distribution is the same in each group. So, if your groups have very different standard deviations and so are not appropriate for one-way ANOVA, they also should not be analyzed by the Kruskal-Wallis or Mann-Whitney test.
3. Often the best approach is to transform the data. Often transforming to logarithms or reciprocals does the trick, restoring equal variance. Box-Cox type transformations (https://en.wikipedia.org/wiki/Power_transform#Box%E2%80%93Cox_transformation) stabilize variance by squeezing the data asymmetrically, either squeezing them downwards with the highest data squeezed the most, or squeezing them upwards with the lowest data squeezed the most. Thus, you need the variance of your data to change with the mean for this to work optimally. See also: https://www.youtube.com/watch?v=cCl4riB9aNo and https://www.real-statistics.com/correlation/box-cox-transformation/box-cox-normal-transformation/

Step 9) Provide a two-way ANOVA of "DAYTOCLOSE" with respect to low-, moderate- and high-income level and the following categories.

1. Animals / Pets
2. Lights / Signals, Signs, Sidewalks / Curbs / Ditch, Streets / Roadways / Alleys
3. Property / Buildings / Construction
4. Public Health
5. Storm Water / Sewer
6. Trash / Recycling
7. Public Safety

To do so you need to generate new Excel files. To get an idea watch the following video: https://www.youtube.com/watch?v=0BirE9K05i0

Provide the a know what table and the tested hypothesis in your slides.

References and Data Resources:

1. Nine Imperatives For Leadership of 311-Enabled Government https://www.innovations.harvard.edu/sites/default/files/128521.pdf
2. Clark, B. Y., Brudney, J. L., & Jang, S. G. (2013). Coproduction of government services and the new information technology: Investigating the distributional biases. *Public Administration Review*, *73*(5), 687-701. https://onlinelibrary.wiley.com/doi/full/10.1111/puar.12092?casa_token=8iqO1UsoVXsAAAAA%3AgqKut-gAoy0x63O7v7CydhUP6LFLxN0nEZ144JaHw5seI8AtpUWzPe-fSPgv31bAw5vCP18vgxlRzg
3. Image-Based Surrogates of Socio-Economic Status in Urban Neighborhoods Using Deep Multiple Instance Learning https://www.mdpi.com/2313-433X/4/11/125
4. Mining 911 Calls in New York City https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewFile/10206/10261
5. Coproduction of Government Services https://uncw.edu/bbwc/brudney/documents/coproductionofgovernmentservicesandthenewinformationtechnologybrudney2013.pdf
6. Intro to ANOVA: https://www.youtube.com/watch?v=oOuu8IBd-yo
   One-way ANOVA + Post Hoc Analysis: https://www.youtube.com/watch?v=uzcqMeNK7Kw
   https://www.youtube.com/watch?v=srDr-4cz1KI

7. TWO ANOVA https://www.youtube.com/watch?v=lZFmFuZGQTk
   https://www.youtube.com/watch?v=xNQliU8yiYk
   We have two factors: neighborhood income level and average service time for each city service

8. Who Calls for Help? https://data.bloomberglp.com/company/sites/2/2018/09/Who-Calls-for-Help.pdf
9. Kansas City demographics https://www.point2homes.com/US/Neighborhood/MO/Kansas-City-Demographics.html#IncomeFinancial
10. How to report a 311 problem: https://www.kcmo.gov/city-hall/311
11. Kansas City 311 Report 2009: https://icma.org/sites/default/files/6051_.pdf
12. Brief History of Kansas City 311 Action Center: https://www.transformgov.org/sites/transformgov.org/files/101977_.pdf
13. Kansas City Government 311 / Service Requests DATA - 2007 to Present https://andrew-friedman.github.io/jkan/datasets/311-Kansas-City-Government/
14. Kansas City economic growth: https://datausa.io/profile/geo/kansas-city-mo/#:~:text=In%202018%2C%20Kansas%20City%2C%20MO,%2454%2C372%2C%20a%205.93%25%20increase.
15. Kansas City income and poverty: https://dashboards.mysidewalk.com/kcmo-advancekc/income