Contents

1. Overview

It is often a challenge to find the most appropriate place to live. We often go by intuitive takes, or concerns that linger at the top of our attention span. As we go live at that given neighbourhood, we soon ponder, "what if", I had just known better. Hence, I make my research/ project statement:

"Explore liveable neighbourhoods"

This body of work presents a simple analysis to enable identification of neighbourhoods (group of zip codes) that are closer to what may be suitable to one's living requirements. It could be converted to a simple app. An app that could evolve to be a common window to explore neighbourhoods.

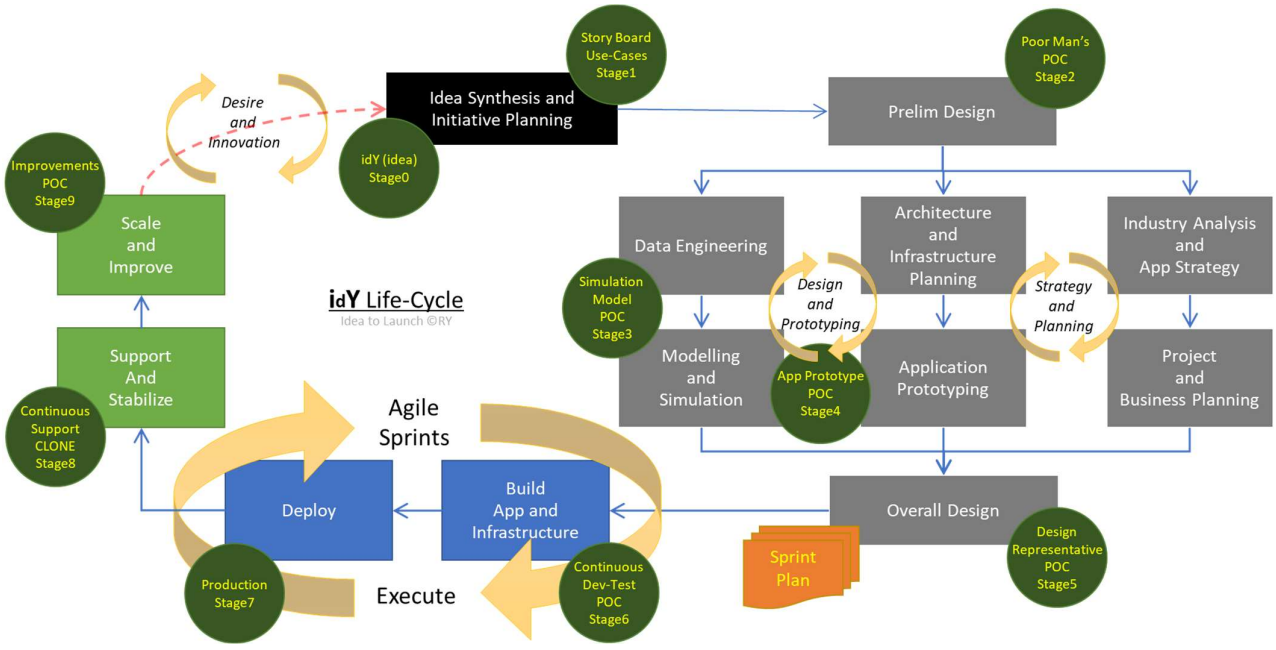
I make no claims to the completeness or superior efficacy of this body of work. I would only expect this to be a starting point to build a more significant application or utility for the future. Additionally, my intention is to limit myself to the broad requirements of IBM-Coursera "Applied Data Sciences" Capstone project requirements.

In the next section we will delve into the methodology "idY" before we float into various aspects of data analysis, and final conclusions.

Throughout this project, I will try to stay off mathematics. I am seriously not doing anything too complex here. So, no sigmas, epsilons, or any other such exotic variables or constants. Let's just chill with simple python code.

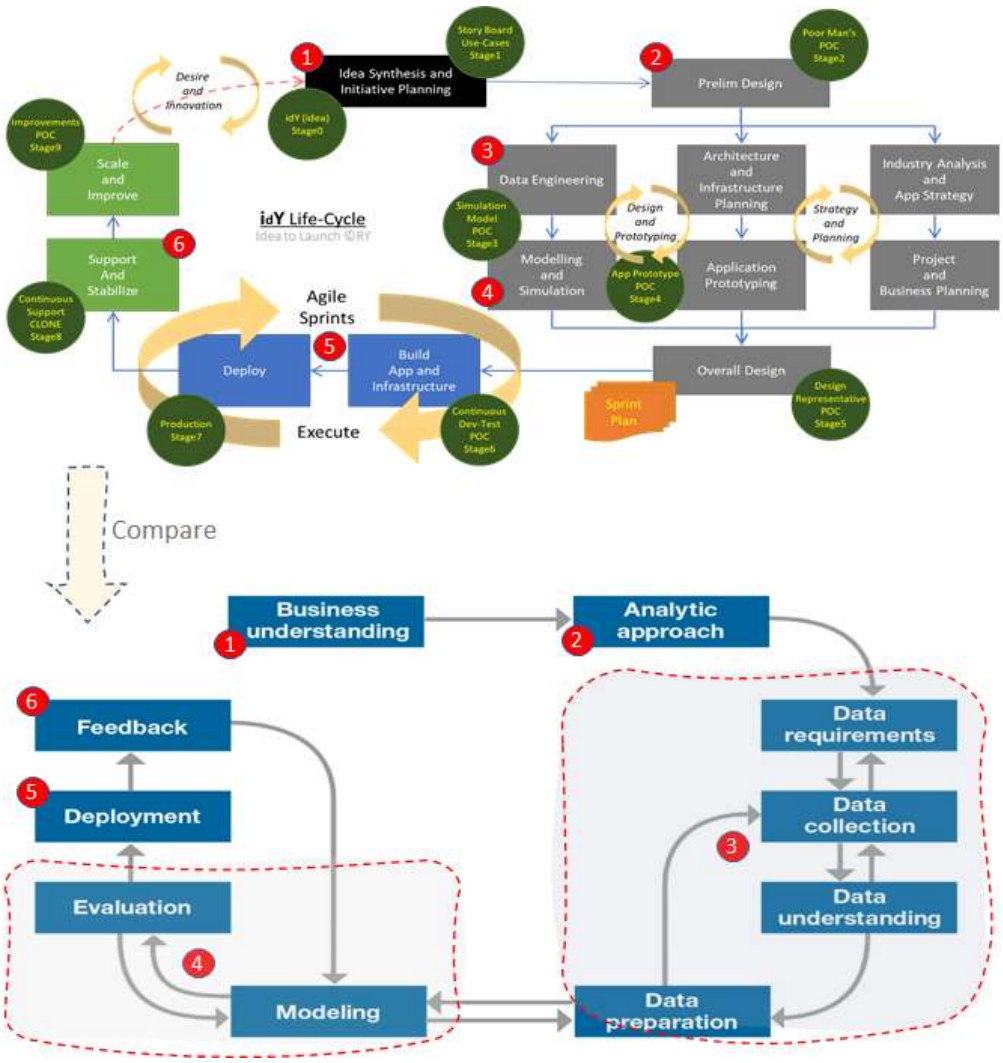
2. "idY" Methodology

idY is a wholistic methodology that provides guidance from idea to launch. It is not limited to Data Science tasks. "idY" can be of considerable advantage to deliver an initiative, in a way, that maximizes the overall outcome, while reducing waste/cost.



"idY" is consistent with "Foundational Data Sciences Methodology", as illustrated here:
https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

"idY" roughly translates to Foundational Data Sciences Methodology, as illustrated below.



The purpose of this document is not to explore "idY" methodology. It is to present an approach to identify liveable neighbourhoods, as a demonstration of beginner-level Data Sciences capabilities. We will focus on only those Data Sciences oriented aspects of idY here. Other idY steps can be explored separately, to convert this basic idea into a viable product for market.

3. Playing by "idY"

We will now walkthrough relevant "idY" methodology steps rapidly to build upon our focus.

1. Idea Synthesis
    - Business or need understanding/assessment
    - Idea to use cases
2. Prelim Design
    - Analytic Approach
    - Define a High-Level approach to the overall problem
3. Data Engineering
    - Data Requirements->Data Collection->Data Understanding->Data preparation
    - Define Data Chain – Sources, Clean-up and Other processing requirements
4. Modelling and Simulation

- Data modelling and Evaluation
- Use Data Sciences platform to simulate and model core processing requirements

5. Build and Deploy
    - Not relevant to this project
6. Support Stabilize and Improve
    - Not relevant to this project

4. Idea Synthesis

Liveability can mean different things to different people. We could build a complex scoring/ranking metrics, but often complex evaluations become either incomprehensible or emotionally inaccessible from a common decision maker's perspective. Hence, we will take a less precise approach to decision making, with a potential to building a coherent model, if the patterns are consistent.

Often life decisions are not binary or discrete choices. They are complex amalgamation of preferences and compromises. Data or information alone cannot lead to decision end-points, as much as, unstructured emotional preferences also cannot. A good mix of the two could lead to a decision with a relatively more sustainable satisfaction level, or so I would like to believe.

Story boards and use cases can be a good way to evolve ideas and setting realizable goals.

4.1 Use-Cases

As a step forward, I would like to encapsulate my analysis objective into a few use-cases as listed below,

1. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.

2. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.

3. Older folks or empty nesters, who value a quieter lifestyle with easy of access to medicare and other life services.

4. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.

5. Preliminary Design

As we go about building design outline, I would like to adhere to some of the guidelines laid out for the Capstone. Especially, the usage of Foursquare API. Outside of this Capstone, we could envisage more open and exploratory research, nevertheless adopting a similar approach.

I would like to look at this project as largely driven by a central driver, which would be influenced by multiple supporting or influencing factors to arrive at a final decision.

Central idea for this analysis:

> Foursquare provides data on various venues by neighbourhood. Use this venue data to classify neighbourhood into distinct clusters.

Supporting ideas to influence the central driver:

> Liveability is by no means a mere aggregation of schools, shops or nightlife around you. There are other aspects that need to be taken into consideration to make a decision. The clusters derived from the above venue analysis can be further evaluated for other aspects, such as,

- Population – Population and density
- Home Prices – affordable homes
- COVID Spread – sensitivity to diseases - how did COVID play; if there aren't drastic structural adjustments, maybe, maybe, the next pandemic will chase the same fault-lines
- Indices indicative of affordability – cost of living, besides home prices
- Crime indices – how safe is the place
- School ratings – How do schools here compare ?
- Incident history – major events in the past, like earthquakes, accidents, weather events, or such notable calamities
- Existence of large environmentally sensitive sites – Nuclear facilities, Chemical Plants, Alien visitations, and such
- And more of such factors that may impact your primary decision approach.

This project will analyze the first three factors listed above – Population, Home Prices, and COVID.

5.1 Analytic Outline

1. Prepare a ZIP code master list for the locale of interest
    - Say in Washington DC
2. Get Population Data by ZIP code.
3. Get Home Price data by ZIP code.
4. Collect venues data by ZIP code, Latitude-Longitude from Foursquare.
5. Build a feature matrix of venue and non-venue data/characteristics
6. Use a clustering algorithm to classify ZIPs into Clusters.
    - As a starting point to our analysis use kMeans method for clustering.
    - Other clustering models such as Decision Tree and DBSCAN can be assessed separately, they are outside the scope of this analysis/project
7. Explore the key characteristics of these clusters to ascertain suitability for each of the above listed use-cases.
8. Assess/Compare population and density by clusters.
9. Assess/Compare home affordability by clusters.
10. Assess/Compare sensitivity to disease spread by clusters.

5.2 Key resource requirements

Identify hardware, software and other resource requirements to execute upon the above analytical outline. For this purpose of this Capstone project, I would limit my imagination to identifying python packages, relevant to this analysis and simulation. We will use Jupyter notebook as a platform for analysis, on a local machine(laptop). All data sources used in this analysis are freely available over internet.

Refer: Jupyter Notebook, Section 5.2

5.3 Key Parameters

Parameterize some of the key data input variable that will help generalize the data model and application simulation. For example, by simply changing the P_select_state to say “NY”, the program could be run for NY state dataset. This will ensure minimization to changes to core code for analysing additional or alternate datasets.

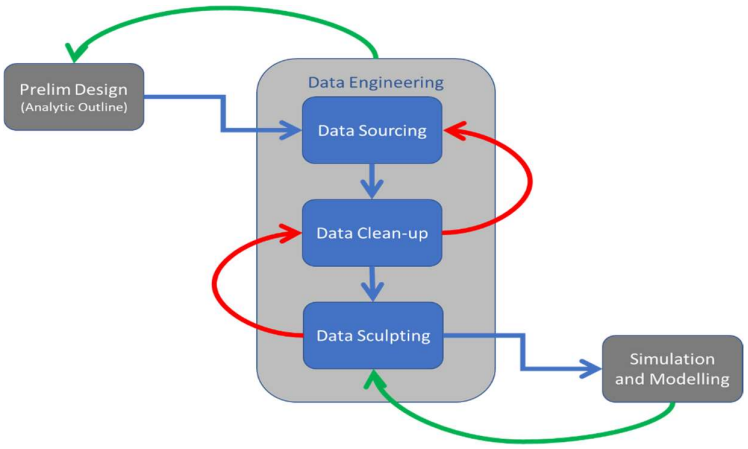
The main sections of this document present a simulation model with DC as the reference state, P_select_state = “DC”. Analysis of Additional states is presented in Annexure 10.1.

Refer: Jupyter Notebook, Section 5.3

```
#key parameters
P_select_state = "DC" # state in USA for analysis
P_POP_data_source = "Data/Population_uszips.csv" # data source population data by ZIP
P_ZHVI_data_source = "Data/ZHVI_uszips.csv" # data source Home-price data by ZIP
P_ZHVI_date = "06/30/2020" # effective date of Home-price data by ZIP
P_COVID_cases_data_source = "Data/covid_confirmed_usafacts.csv" # COVID confirmed cases by County
P_COVID_deaths_data_source = "Data/covid_deaths_usafacts.csv" # COVID deaths by County
P_COVID_date = "07/08/2020" # effective date of Home-price data by ZIP
P_path = r'D:\GIT\IBM_Coursera\Coursera_Capstone\Data' # file path to store venue counts extracts from
Foursquare
P_4sq_CLIENT_ID = 'N233FNIITUBDDXZ4OR5AWM0YEY1XXVKWP2L2IAZOWISCPLWC' # your Foursquare ID
P_4sq_CLIENT_SECRET = 'YHREMIYEIW5TLJQYFDZGVHZVTW13AFPT3GLBJRS11KWRC2OP' # your Foursquare Secret
P_4sq_VERSION = '20180605' # Foursquare API version
P_rainbow = ['Red', 'Orange', 'Green', 'Blue', 'Yellow', 'Brown', 'Black', 'Purple'] # color pallet for
clusters
P_default_zoom = 8.0 # default zoom for visualization of clusters on a map, set 12 for small states
like DC, 6 for large states like CA
```

6. Data Engineering

Data engineering as per "idY" follows a 3-step sequence:



6.1 Data Sourcing

Identify and evaluate data sources. The evaluation process must consider – availability, goodness of fit, quality/consistency, cost, and other strategic/business considerations. Check the scope of analytics that is feasible with the identified data sources. There should be an obvious preference for data sources that have the capability to scale or support a wider scope of analysis. This would ensure that we can add more functionality to our app as we evolve. For the scope of this project, we have limited our appetite to data availability at a ZIP code level, and free of charge.

Ref: Jupyter Notebook, Section 6.1

Data sources used, include:

a. Population Data
   - Source: https://simplemaps.com/data/us-zips
   - Type: Download CSV file
   - Data freely available by ZIP, State, and County
   - Other demographic details can also be fetched by paid subscription

b. Home Price Data
   - Source: https://www.zillow.com/research/zhvi-methodology/
   - Type: Download CSV file
   - Data freely available by ZIP, State, and County
   - Data is available in a time-series format by month-end; for this analysis choose the latest published data

c. COVID Data
   - Data available in 2 parts, separately for Confirmed Cases and Deaths
   - Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
   - Type: Download CSV files
   - Data freely available by State, and County
   - Data is available in a time-series format by month-end; for this analysis choose the latest published data

d. Venue Data
   - Source: https://api.foursquare.com/v2/venues/
   - Type: API call
   - Call limit of 1000/day with unregistered account; 100000/day with a registered account