Derrick Priebe
Matthew Peters
Pierre Augustamar
INFX 575
Project Proposal

Revised Project Proposal: MapMyRun and Property Information

Question:

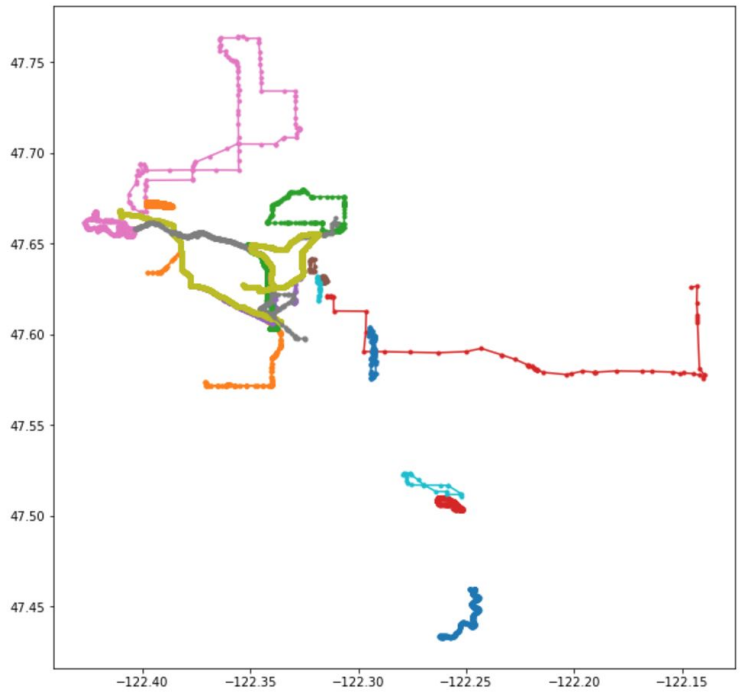
We plan to analyze how documented exercise utiling MapMyRun, as a socially shared representation of running activity, is correlated with real estate. We plan to focus on how various metrics of running activity are correlated with housing values but may also consider correlation to other real estate metrics such as housing type and housing density.

We have found examples of data aggregation companies tying MyHealth information to Zillow at Apiant (https://apiant.com/connect/Zillow-Property-GetDeepComps-to-GETHealth), but this appears to be geared for individual market research, and not in form of an academic study.

Data:

We will utilize wearable activity data as one main data source. Principally, we will acquire data from MapMyRun API. This provides workout information including location, run length, run speed, and other applicable running information. We have successfully extracted this data as shown in figure 1 below.

Figure 1: Sample MapMyRun Routes



Sample MapMyRun route data in JSON form:


```
{
        "description": "",
        "country": "us",
        "total_descent": -4.0657231389,
        "images": [],
        "state": "WA",
        "_links": {
          "alternate": [
            { "href": "/v7.1/route/100001/?format=kml&field_set=detailed",
              "id": "100001",
              "name": "kml" },
            { "href": "/v7.1/route/100001/?format=gpx&field_set=detailed",
              "id": "100001",
              "name": "gpx" }
          ],
          "self": [ { "href": "/v7.1/route/100001/", "id": "100001" } ],
          "privacy": [ { "href": "/v7.1/privacy_option/3/", "id": "3" } ],
          "activity_types": [ { "href": "/v7.1/activity_type/16/", "id": "16" } ],
          "thumbnail": [{ "href":
"//drzetlglcbfx.cloudfront.net/routes/thumbnail/100001/1155832438?size=100x100" } ],
          "user": [ { "href": "/v7.1/user/18717/", "id": "18717" } ]
        },
        "city": "Seattle",
        "min_elevation": 3.32,
        "postal_code": "98188",
        "points": [
```

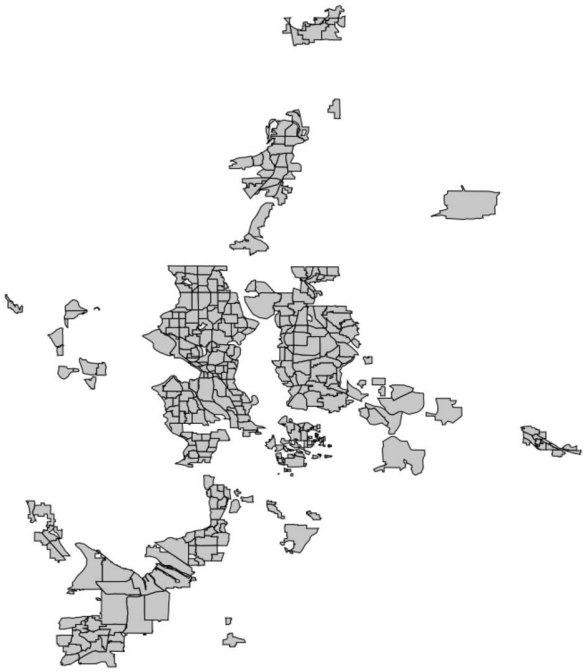
```
        { "dis": 0.0,
          "ele": 5.83,
          "lng": -122.247126102,
          "lat": 47.4576162962 },
        { "dis": 29.12,
          "ele": 6.61,
          "lng": -122.24719584,
          "lat": 47.4573587734},
...
        {"dis": 9987.45,
          "ele": 5.24,
          "lng": -122.247072458,
          "lat": 47.4578339202},
        {"dis": 10009.96,
          "ele": 5.82,
          "lng": -122.24714756,
          "lat": 47.4576380587}],
      "total_ascent": 7.0583606896,
      "data_source": "MapMyRun",
      "created_datetime": "2006-08-17T16:33:58+00:00",
      "start_point_type": "",
      "name": "The Wilbur 10K",
      "max_elevation": 10.62,
      "distance": 10010.0,
      "updated_datetime": "2006-08-17T16:33:58+00:00",
      "climbs": [],
      "starting_location": { "type": "Point",
                             "coordinates": [ -122.2471261024, 47.4576162962 ]
                           }
    }
```


As our secondary source we will utilize real estate information. We will utilize data from King County GIS records. We have successfully extracted this data from their FTP store at ftp://ftp.kingcounty.gov/gis-web/GISData/property_SHP.zip. An example of property information is visualized below in figure 2.

Figure 2: Magnolia Property Map

A third data element is neighborhood boundaries which we believe may help us filter the data on a lower level to help provide additional insight across areas of the city. We utilized geocoded boundaries provided by Zillow at https://www.zillow.com/howto/api/neighborhood-boundaries.htm. The output is depicted in figure 3 below.

Figure 3: Washington Neighborhood Boundary Visualization



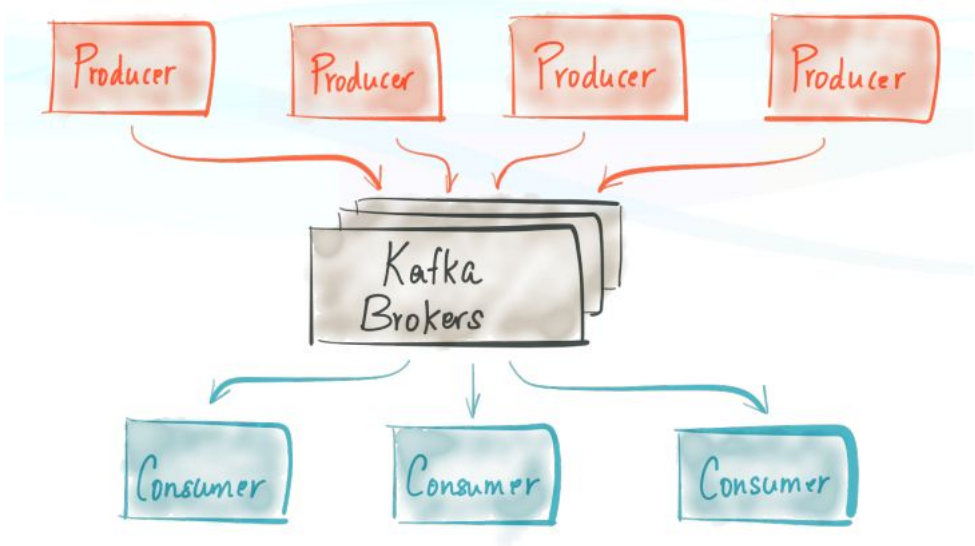
<u>Technology and Methods:</u>

We plan to utilize AWS as the the main data store. Within AWS, we will utilize Apache Kafka as a main ***pipeline*** to secure regular API data extracts from MapMyRun which will feed PostgreSQL database implementing PostGIS extensions and running on AWS RDS. Overall, we will utilize PostgreSQL (***schema*** *detail below*) as the main source to hold tables from all the utilized data sources. Python, utilizing Jupyter Notebook, is our data analysis tool of choice where we will pull applicable data frames, make necessary transformations, and visualize the data.

Pipeline

For data persistence, we will be using Apache Kafka which is an open source distributed pub/sub messaging system (see figure 4).

Figure 4: Apache Kafka architecture



We will build our kafka hub with two topics. One topic will contain the mapmyrun data which will be taken in real-time. And a second topic that will contain static home pricing values. Thus, we will have two producers pushing data to the Kafka brokers. The data stored in Kafka will then go to a transformation that will be handled by the consumers. The transformed data will be uploaded into a storage database. At this time, we only anticipate having one consumer, but as we investigate additional open source tools, we may add other consumers if time permits and or if it adds values to the project.

To set-up a communication layer between Kafka and the related producers, and consumers, we will be using PyKafka. PyKafka is a cluster-aware Kafka client for Python. It includes Python implementations of Kafka producers and consumers.

Sample connection to the Kafka hub

```
>>> from pykafka import KafkaClient
>>> client = KafkaClient(hosts="127.0.0.1:9092,127.0.0.1:9093,...")
```

Sample creating a producer

```
>>> with topic.get_sync_producer() as producer:
...     for i in range(4):
...         producer.produce('test message ' + str(i ** 2))
```

Sample consume messages using a consumer instance

```
>>> consumer = topic.get_simple_consumer()
>>> for message in consumer:
```

```
...     if message is not None:
...         print message.offset, message.value
```

To monitor the data in Kafka, we will use **Presto** which is an open source distributed SQL query engine for running interactive analytic queries against data sources. A topic in Kafka is considered to be a table in Presto. Thus, Presto allows to run SQL like query against a topic.

Sample query

```
presto:tpch> SELECT _message FROM customer LIMIT 5;
```

Database Schema

```
"public.city_county"
     Column      |  Type   |     Modifiers
-----------------+---------+-------------------------
 city            | text    |
 county          | text    |
 city_county_id | bigint | not null default
nextval('city_county_city_county_id_seq'::regclass)
Indexes:
    "city_county_pkey" PRIMARY KEY, btree (city_county_id)
Referenced by:
    TABLE "neighborhoods" CONSTRAINT
"neighborhoods_city_county_id_fkey" FOREIGN KEY (city_county_id)
REFERENCES city_county(city_county_id)


"public.neighborhoods"
     Column      |           Type           |    Modifiers
-----------------+--------------------------+-------------------------
 gid             | integer                  | not null default
nextval('neighborhoods_gid_seq'::regclass)
 state           | character varying(80)    |
 county          | character varying(80)    |
 city            | character varying(80)    |
 name            | character varying(80)    |
 regionid        | character varying(80)    |
 geom            | geometry(MultiPolygon)   |
 city_county_id | bigint                   |
Indexes:
```

```
    "neighborhoods_pkey" PRIMARY KEY, btree (gid)
    "neighborhoods_city_county_idx" btree (city_county_id)
    "neighborhoods_geom_idx" gist (geom)
Foreign-key constraints:
    "neighborhoods_city_county_id_fkey" FOREIGN KEY (city_county_id)
REFERENCES city_county(city_county_id)
```

"public.route"

```
     Column      |            Type             |   Modifiers
-----------------+-----------------------------+------------------
 route_id        | bigint                      | not null default
nextval('route_route_id_seq'::regclass)
 map_my_run_id   | character varying           |
 postal_code     | character varying           |
 city            | character varying           |
 neighborhood_id | bigint                      |
 distance        | bigint                      |
 create_datetime | timestamp with time zone    |
 route_name      | character varying           |
 total_ascent    | character varying           |
 max_elevation   | double precision            |
 thumbnail       | character varying           |
 path            | geometry(LineStringZ,4362) |
 city_county_id  | bigint                      |
Indexes:
    "route_pkey" PRIMARY KEY, btree (route_id)
    "route_path_idx" gist (path)
Foreign-key constraints:
    "route_city_county_id_fkey" FOREIGN KEY (city_county_id)
REFERENCES city_county(city_county_id)
```

"public.parcel_address"

Column	Type	Modifiers
gid	integer	not null default nextval('parcel_address_gid_seq'::regclass)
major	character varying(6)	
minor	character varying(4)	
pin	character varying(10)	
comments	character varying(254)	
sitetype	character varying(2)	
alias1	character varying(60)	
alias2	character varying(60)	
siteid	numeric	
addr_hn	character varying(10)	
addr_pd	character varying(2)	
addr_pt	character varying(6)	
addr_sn	character varying(80)	
addr_st	character varying(6)	
addr_sd	character varying(2)	
addr_num	numeric(10,0)	
addr_full	character varying(120)	
fullname	character varying(120)	
zip5	character varying(5)	
plus4	character varying(4)	
ctyname	character varying(28)	
postalctyn	character varying(28)	
lat	numeric	
lon	numeric	
point_x	numeric	
point_y	numeric	
county	character varying(12)	
kroll	character varying(4)	
juris	character varying(2)	
big_ten	character varying(20)	
budget_uni	character varying(40)	
kctp_city	character varying(75)	
kctp_state	character varying(2)	
plss	character varying(11)	
prop_name	character varying(50)	
plat_name	character varying(50)	
plat_lot	character varying(14)	
plat_block	character varying(7)	

```
 presentuse     | integer                |
 lotsqft        | numeric(10,0)          |
 levycode       | character varying(4)   |
 levy_juris     | character varying(25)  |
 new_constr     | character varying(1)   |
 taxval_rsn     | character varying(2)   |
 apprlndval     | numeric                |
 appr_impr      | numeric                |
 tax_lndval     | numeric                |
 tax_impr       | numeric                |
 accnt_num      | character varying(12)  |
 kctp_taxyr     | integer                |
 kctp_par       | character varying(10)  |
 unit_num       | character varying(254) |
 bldg_num       | character varying(254) |
 condositus     | character varying(254) |
 qts            | character varying(2)   |
 sec            | character varying(2)   |
 twp            | character varying(2)   |
 rng            | character varying(2)   |
 primary_ad     | integer                |
 legaldesc      | character varying(254) |
 shape_area     | numeric                |
 shape_len      | numeric                |
 geom           | geometry(MultiPolygon) |
 city_county_id | bigint                 |
Indexes:
    "parcel_address_pkey" PRIMARY KEY, btree (gid)
    "parcel_address_city_idx" btree (ctyname)
    "parcel_address_geom_idx" gist (geom)
Foreign-key constraints:
    "parcel_address_city_county_id_fkey" FOREIGN KEY (city_county_id)
REFERENCES city_county(city_county_id)
```

Previous Work:

There are some previous work with regard to this topic as detailed below:

Apiant.com provides an interface to pull MyHealth information triggered by home sales. It does not provide analysis, but leaves that to customers who might want to subscribe to the service. From a technological point of view, the solution is similar, although the intent and deliverables of our effort is very different.

Source: https://apiant.com/connect/Zillow-Property-GetDeepComps-to-GETHealth

The **paper "Not Just a Walk in the Park: Methodological Improvements for Determining Environmental Justice Implications of Park Access in New York City of the Promotion of Physical Activity"** takes a focused look at a single city to analyze availability of park to various socioeconomic and ethnic sub-groups. While we have academic interests that are similar, the scope and approach of our study is quite different. This may be a relevant read simply to provide academic and methodological insights, although it is not likely to have anything in common with means of research. As the paper notes, "This study is designed to shed light on the"unpatterned inequities" of park distributions identified in previous studies of New York City park access. (Miyake et al)".

Source: Miyake KK, Maroko AR, Grady KL, Maantay JA, Arno PS. Not Just a Walk in the Park: Methodological Improvements for Determining Environmental Justice Implications of Park Access in New York City for the Promotion of Physical Activity. *Cities and the environment*. 2010;3(1):1-17.

A **paper, "City, Culture and Society", by Lawson and Fadare** uses "simple random sampling of household heads" (Lawson, Fadare) in three distinct neighborhoods of Eti Osa, Nigeria. Here, our methodologies and score vary greatly. As noted in the review: "This paper considers the effects of socio-economic status as a determinant of urban health outcomes. Issues examined include housing and environmental conditions as well as socio-economic characteristics such as age, gender, income and household size. Furthermore, health seeking behaviour was investigated and these include expenditure on health as well as health and nutritional habits."

Source: Taibat Lawanson, Samson Fadare, 2014, Elsevier Ltd, City, Culture and Society, Volume 6, Issue 1, Pages 43-52, *Environment and health disparities in urban communities: Focus on Eti Osa, Nigeria*

Deliverable:

Our end product will be a set of visualizations, charts and graphs in PowerPoint that will walk through our exploration of the data regarding our project topic. We will attempt to explain the reasoning behind the topic, provide background information, and clearly state our topic and thesis. We will explore the topic in general to provide some context around the data utilized. We will also delve into some more detailed views of the data as well as show some more advanced

statistical analysis measures that we found that are important to note. We will clearly state the purpose and conclusion from each chart or table shown. We will also provide an overall conclusion of our research as well and provide resources for a user to delve into the topic through outside sources as well.

Plan:

Date	Activity Completion
4/11/17, Tues	Project Proposal Due
4/28/17, Mon	Confirm final data sources and extraction process
5/1/17, Mon	Data exploration and cleaning; Revised Project Proposal due
5/8/17, Mon	Analysis Phase
5/14/17, Sun	Draft UI complete
5/15/17, Mon	Draft project complete
5/22/17, Mon	Project refinement complete
5/28/17, Sun	Report and Presentation material complete
5/29/17, Mon	Project Presentation

Risks:

Agreement - We may encounter delay in agreeing on the strategic direction of our research and concluding important decisions during the project timeframe.
API - We may encounter issues successfully accessing the data from an API that we originally assessed as a reasonable resource. We have already found that API access can be difficult and require many conditions or limitations on access.
Data cleansing - Our data may need a lot of cleansing in order to be usable for the purpose of our research.
Complexity - We may run into issues with the amount of fields available in the data causing a lot of dimensionality to our analysis that could interfere with a cohesive message.
User Interface - We may encounter issues successfully translating our data analysis into a cohesive visualization with a technology or method that we have not successfully utilized before.