P18 GPS activity space and Twitter Data Analysis Plan

1. Datastets:

General Description of GPS Data:

- 250 participants over 2 weeks with GPS coordinates every 10 seconds (protocol but actual data varies between individuals)
- Max size = 30,240,000 data points
- Missing/erroneous data via participant attrition, subway use (tracker does not function underground), tracker malfunction/interference
- Participants in two waves (different times of year)
- Cross streets of home
- Longitudinal survey data on behaviors, etc.
- Total N might vary from ~250-211 depending on limiting data

2. General Description of Twitter Data:

- (NYC-RacismNHomophobia_till013118.csv) updated using the data till 01/31/2018. The details of the data including the description of each column can be found in DataDictionary.txt file

3. Plan for Analysis:

We are trying to find out the relationship between discrimination and sexual risk behavior. The plan are listed below:

a. Perform areal weighted mean of Twitter grid cell variables ("SSSOM_Rac_grid", "SSSOM_Hom_grid" and "Rac_tweets_grid") within GPS activity space variables as an indicator of individual exposure to neighborhoods and corresponding risk a single-number summary of risk

"areal weighted mean" will take a mean sum over all grid cells, *c*:

$$\sum_{c} SSOM\_Rac_c \times time\_spent\_in\_grid\_cell_c$$

b. This will give us an average unit for further analysis of the AWM variables ("AWM_SSSOM_Rac", "AWM_Rac_tweets", "AWM_SSSOM_Hom" "AWM_Zip_Hom", and "AWM_Zip_Rac))

c. Right now we are not controlling for time, rather this is a cross sectional design.

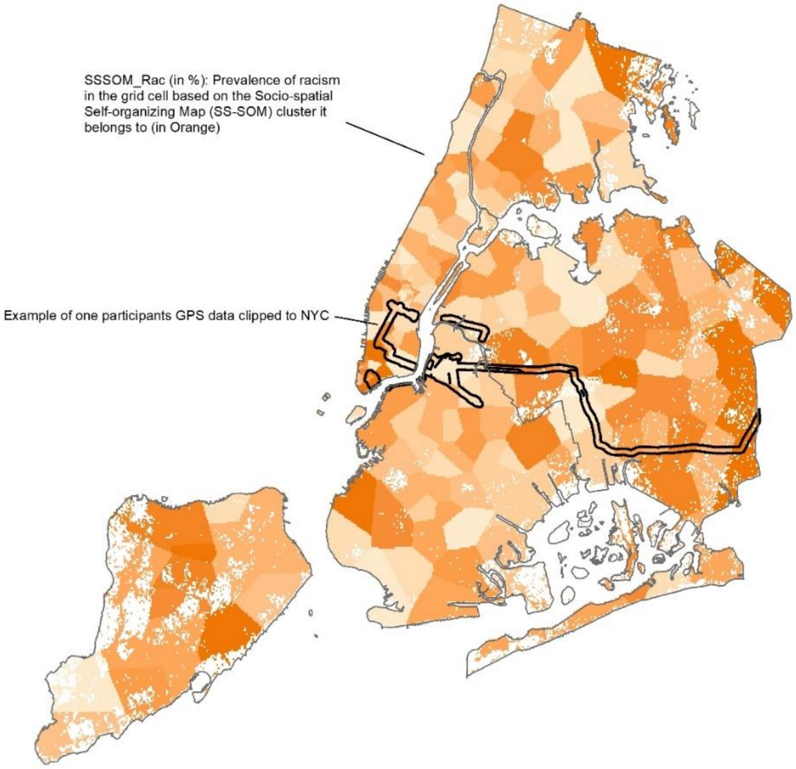
d. The idea is to include the AWM variable ("AWM_SSSOM_RAC", etc.) as a covariate in a regression model (e.g. quasi-poisson regression, negative binomial regression). Other covariates and outcomes for the model are described below.

Table 1: codebook for grid cell variables. Spreadsheet: P18_Final_Data_07162019.csv

Variables Name	Description	Name in spreadsheet
SSSOM_Rac_grid	Prevalence of racism in the grid cell based on the SS-SOM cluster it belongs to.	SSSOM_Rac (in %)
SSSOM_Hom_grid	Prevalence of homophobia in the grid cell based on the SS-SOM cluster it belongs to.	SSSOM_Hom (in %)
Rac_tweets_grid	Prevalence of racist tweets in the grid cell based on the SS-SOM cluster it belongs to.	NormalRacistTweets (in %)

Zip_Rac_grid		Zip_Rac (in %)
Zip_Hom_grid		Zip_Hom (in %)

Twitter data and one participants GPS activity space (200m)


SSSOM_Rac (in %): Prevalence of racism in the grid cell based on the Socio-spatial Self-organizing Map (SS-SOM) cluster it belongs to (in Orange)
Example of one participants GPS data clipped to NYC

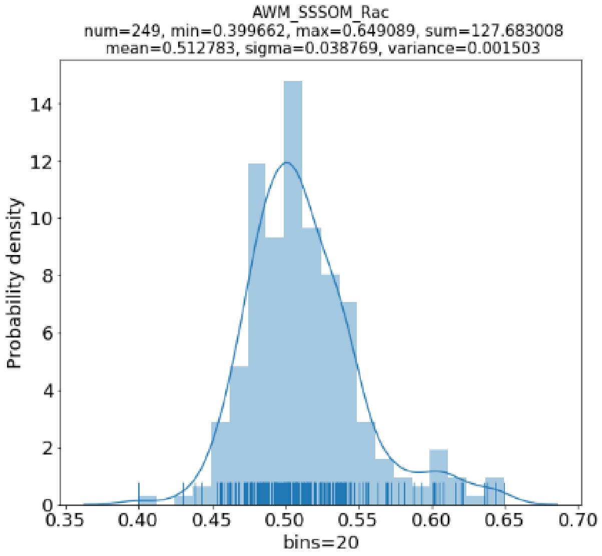
4. Descriptive Statistics of GPS Activity Derived Variables:

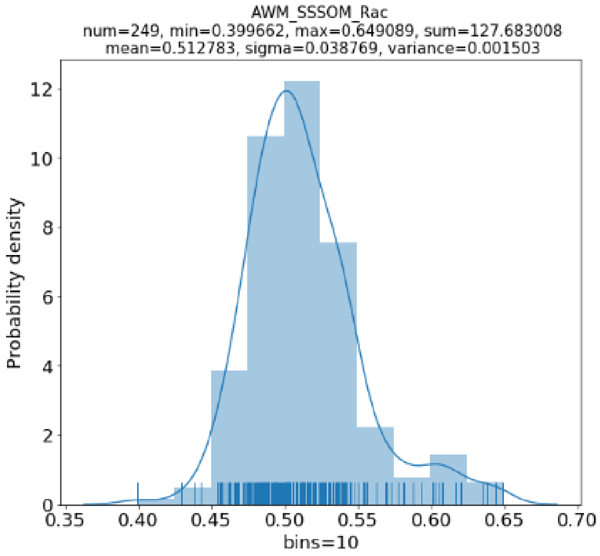
Table 2: codebook for AWM variables. Spreadsheet: P18_GPS_AWM_Twitter_data_summarystats.csv

Variables Name	Description	Name in spreadsheet
AWM_SSSOM_Rac	Areal weighted mean of SSSOM_Rac_grid.	AWM_SSSOM_Rac
AWM_SSSOM_Hom	Areal weighted mean of SSSOM_Hom_grid.	AWM_SSSOM_Hom
AWM_Rac_tweets	Areal weighted mean of Rac_tweets_grid.	AWM_Norm_R_tweets
AWM_Zip_Rac	Areal weighted mean of Zip_Rac_grid.	AWM_Zip_Rac
AWM_Zip_Hom	Areal weighted mean of Zip_Hom_grid.	AWM_Zip_Hom
Area_km	Area in square kilometers of activity spaces.	Area_km

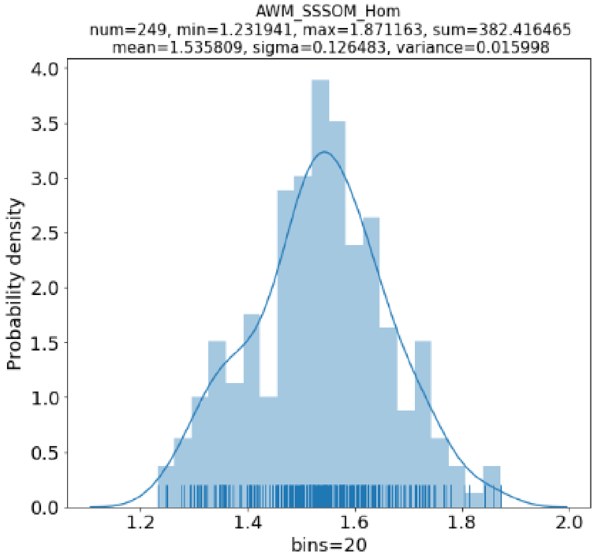
Field: AWM_SSSOM_Rac
Count: 249
Minimum: 0.399662
Maximum: 0.649089
Sum: 127.683008
Mean: 0.512783

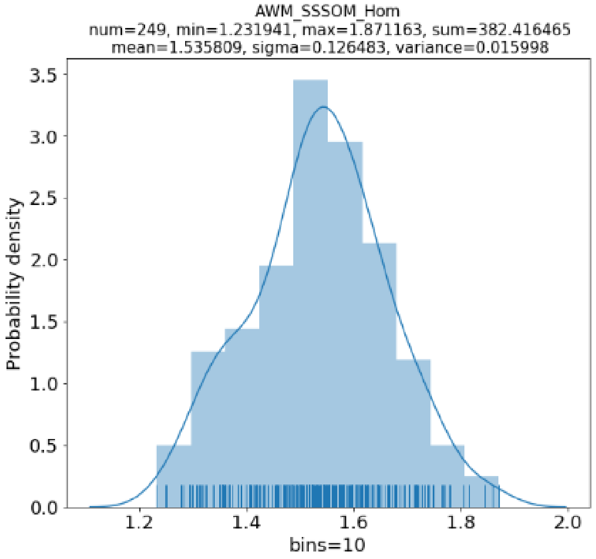
Standard Deviation:	0.038769
Nulls:	0





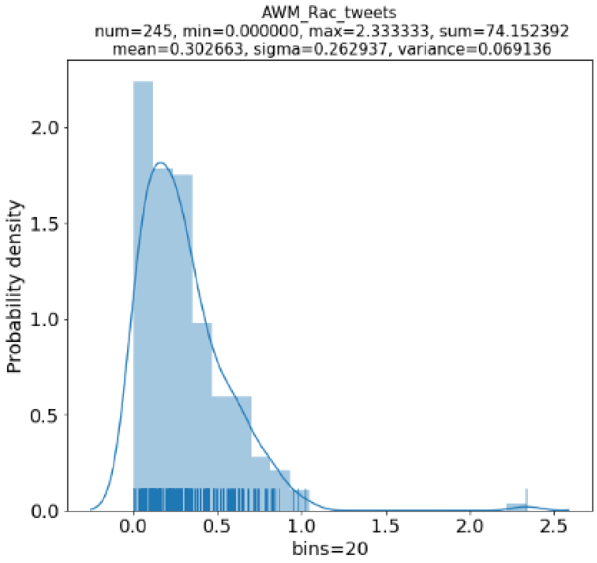
Field: AWM_SSSOM_Hom
Count: 249
Minimum:	1.231941
Maximum:	1.871163
Sum:	382.416465
Mean:	1.535809
Standard Deviation:	0.126483
Nulls:	0

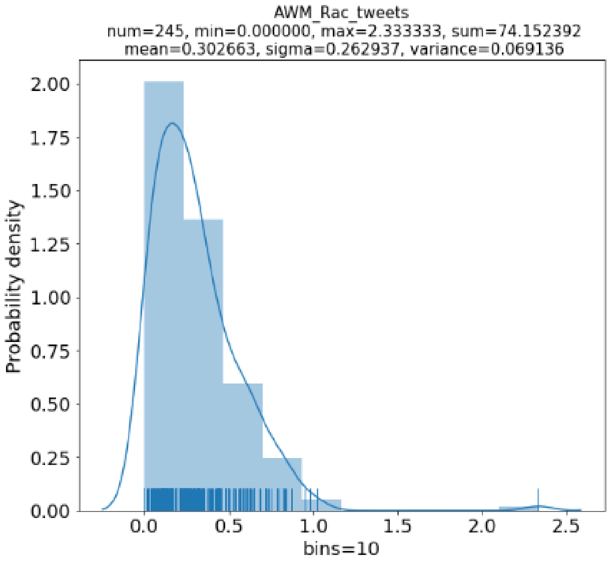




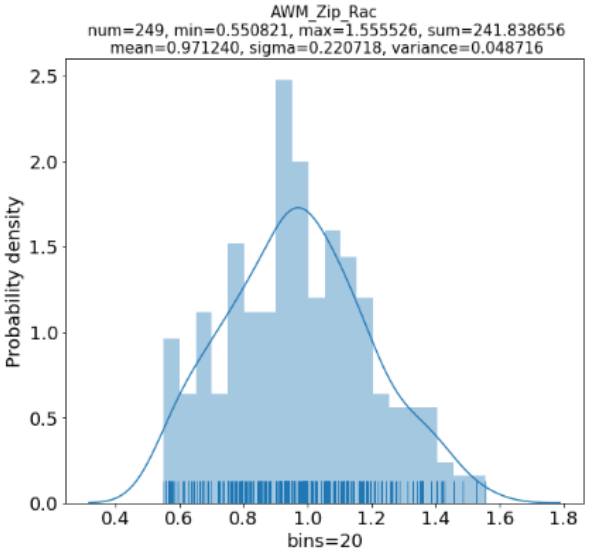
Field: AWM_Rac_tweets
Count: 249
Minimum:	0

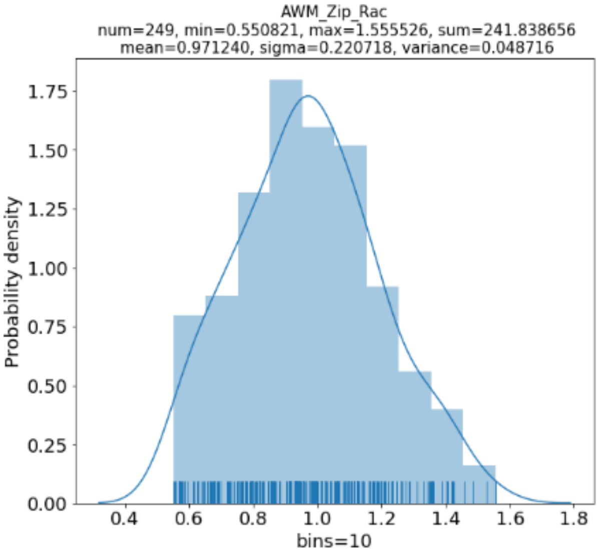
Maximum: 2.333333
Sum: 74.152392
Mean: 0.297801
Standard Deviation: 0.263578
Nulls: 0


AWM_Rac_tweets
num=245, min=0.000000, max=2.333333, sum=74.152392
mean=0.302663, sigma=0.262937, variance=0.069136


AWM_Rac_tweets
num=245, min=0.000000, max=2.333333, sum=74.152392
mean=0.302663, sigma=0.262937, variance=0.069136

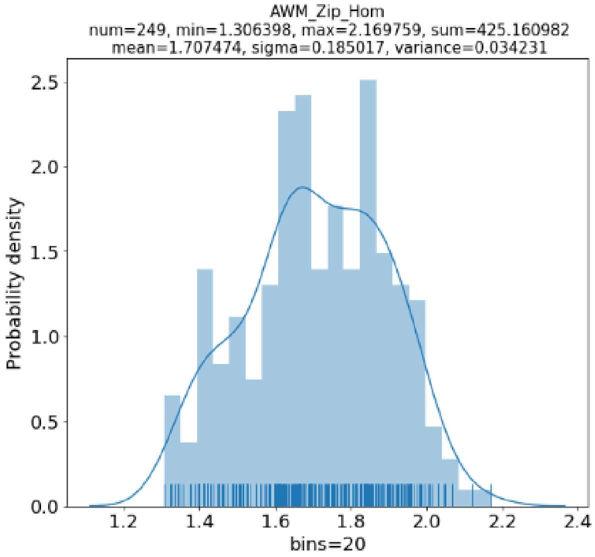
Field: AWM_Zip_Rac
Count: 249
Minimum: 0.550821
Maximum: 1.555526
Sum: 241.838656
Mean: 0.97124
Standard Deviation: 0.220718
Nulls: 0

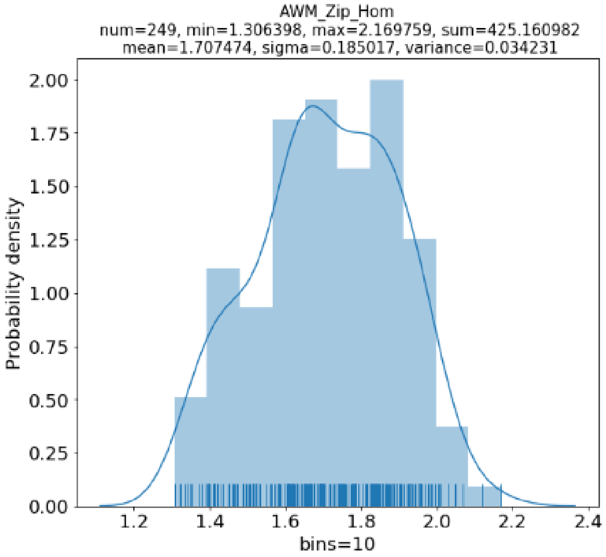

AWM_Zip_Rac
num=249, min=0.550821, max=1.555526, sum=241.838656
mean=0.971240, sigma=0.220718, variance=0.048716


AWM_Zip_Rac
num=249, min=0.550821, max=1.555526, sum=241.838656
mean=0.971240, sigma=0.220718, variance=0.048716

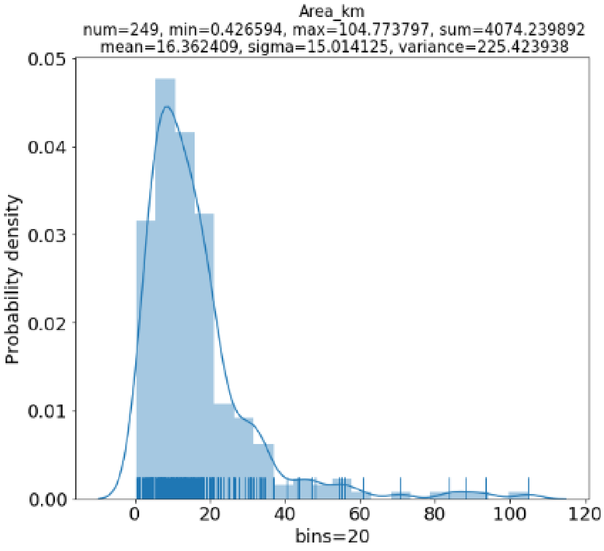
Field: AWM_Zip_Hom
Count: 249

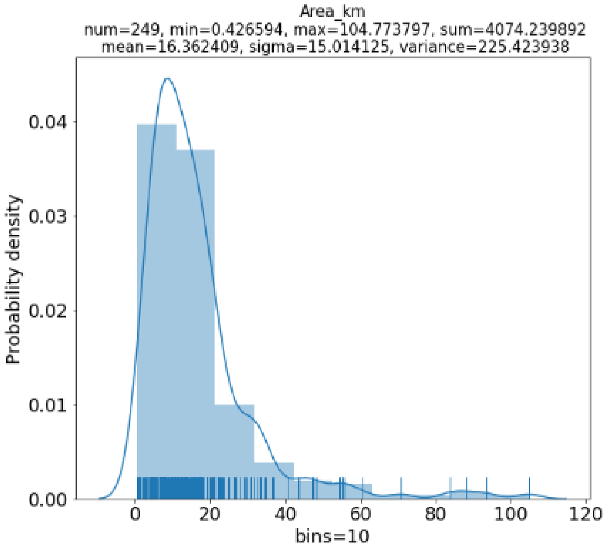
Minimum: 1.306398
Maximum: 2.169759
Sum: 425.160982
Mean: 1.707474
Standard Deviation: 0.185017
Nulls: 0





Field: Area_km
Count: 249
Minimum: 0.426594
Maximum: 104.773797
Sum: 4074.239892
Mean: 16.362409
Standard Deviation: 15.014125
Nulls: 0





5. Covariates:

Potential list of covariates pulled from our recent manuscript Associations of Spatial Mobility with Sexual Risk Behaviors among Young Men Who Have Sex with Men in New York City: The Project 18 Neighborhood Study. I think we still need to discuss what statistical model we want to run. I am less familiar with the survey level data but the covariates from the manuscript are listed below.

-age (years), ethnicity (Hispanic or non-Hispanic), race (Black, Asian, White, and others), education attainment (high school or less, some college/ technical school, college degree or more), current school enrollment status (yes/no), and foreign-born status (yes/no), Homelessness, total individual annual income was categorized as less than $15,000, between $15,000 and $35,000, more than $35,000 per year, which approximates national poverty level (<$15,000)

May I have the codebook for the variable names in file “P18_Final_Data_07162019.csv”? I may need to know which variable is age, ethnicity, etc.

6. Outcomes:

Potential outcomes: sexual behaviors associated with risk of HIV infection among MSM. The Project 18 Cohort Study collected data on sexual behaviors including number of male sexual partners and number of condomless sexual encounters in past six months. The number of male sexual partners was assessed from two questions: (1) “In the past 6 months, how many male steady partners have you have anal or oral sex with?,” (2) “In the past 6 months, how many casual male partners have you had anal or oral sex with casual or non-steady partners?”. The total number from those two questions was used as one outcome and was considered a count type variable in the analyses [35-39]. In addition, numbers and types of sexual encounters were assessed to create three variables: total numbers of condomless anal intercourse acts, (2) total numbers of condomless insertive anal intercourse acts, and (3) total number of condomless receptive anal intercourse acts in past six months [35-40].

To be more specific, the outcome variables can be:

Anal Intercourse variables: "AI_Total", "AI_Condom", "AI_Condomless"

Insertive Anal Intercourse variables: "IAI_Total", "IAI_Condom", "IAI_Condomless"

Receptive Anal Intercourse variables: "RAI_Total", "RAI_Condom", "RAI_Condomless"

Others: "drug_use"