THE CITY OF NEW YORK 302248
DEPARTMENT OF HEALTH AND MENTAL HYGIENE
DIVISION OF ENVIRONMENTAL HEALTH

NOTICE

CLOSED

BY ORDER OF THE
COMMISSIONER
OF HEALTH AND MENTAL HYGIENE

#138
INSPECTOR/SHIELD NO.

April 18, 2019
DATE

NO PERSON SHALL VIOLATE AN ORDER OF THE BOARD, COMMISSIONER OR DEPARTMENT.

NYC Health Code §3.05

NO MATERIAL, ORDER OR NOTICE POSTED OR REQUIRED TO BE POSTED BY THE DEPARTMENT SHALL BE MUTILATED, OBSTRUCTED, TORN DOWN OR REMOVED UNLESS AUTHORIZED TO DO SO BY THE DEPARTMENT OR BY THIS CODE OR OTHER APPLICABLE LAW.

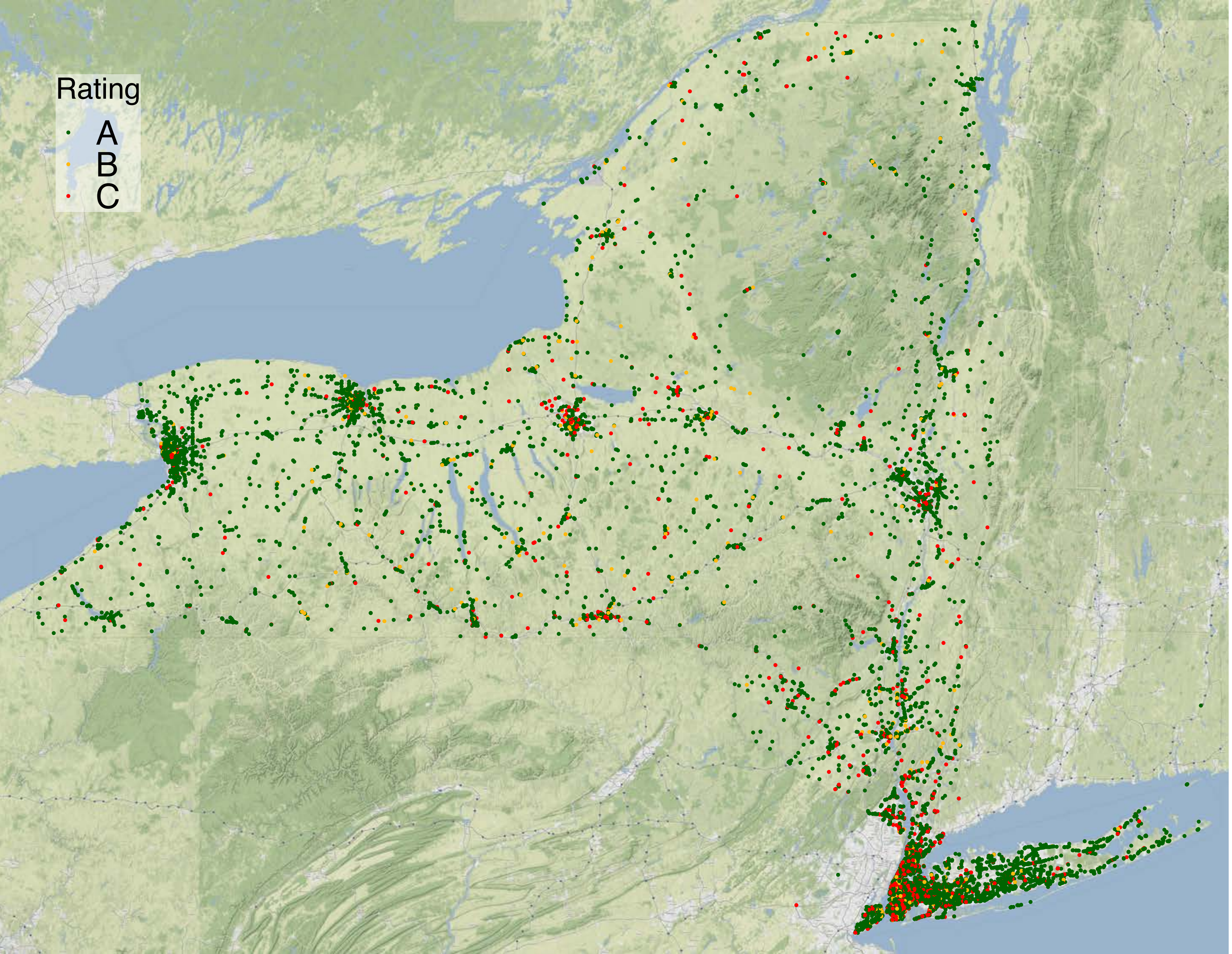
NYC Health Code §3.17

EHS 57 (Rev. 10/13)

DO NOT REMOVE

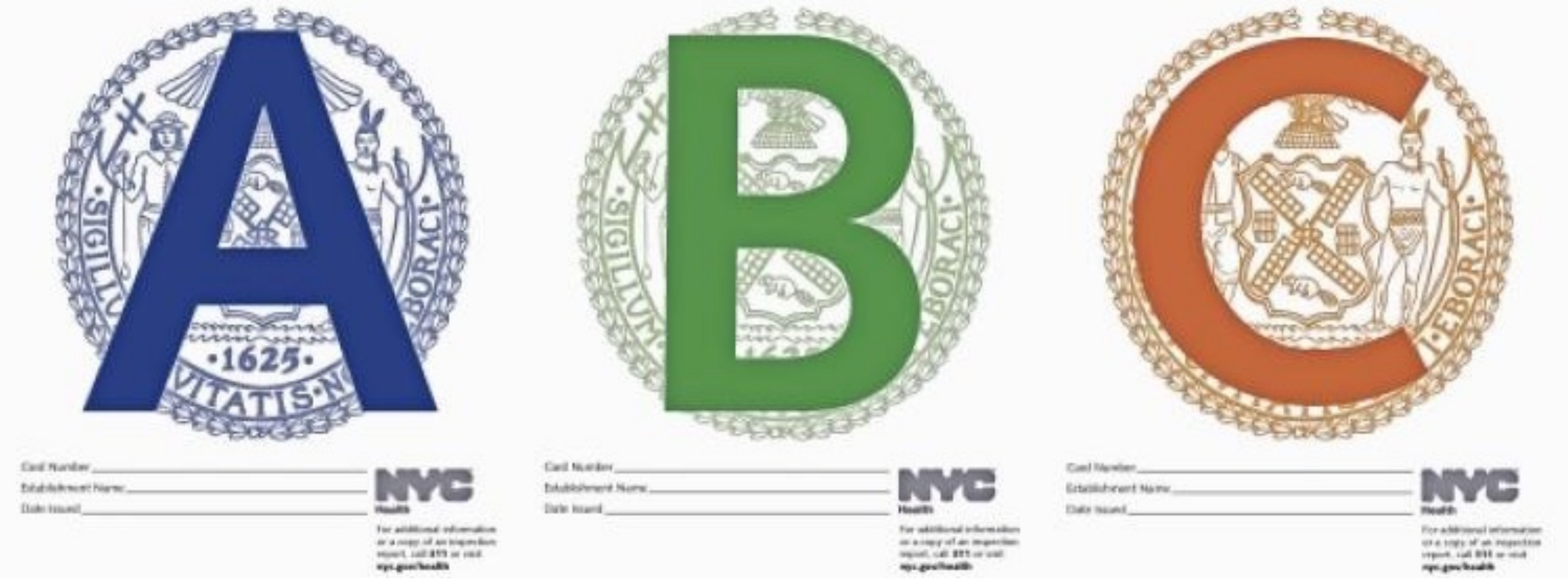
Predicting Food Store Inspection Grades

By André Ruckdaeschel, Kaspar Lichtsteiner & Matthias Steiner



Predicting Food Store Inspection Grades

By André Ruckdaeschel, Kaspar Lichtsteiner & Matthias Steiner


A
SIGILLUM
CIVITATIS NOVI
EBORACI
1625
B
C

Card Number:

Establishment Name:

Date Issued:

NYC
Health

For additional information or a copy of an inspection report, call 311 or visit nyc.gov/health

Card Number:

Establishment Name:

Date Issued:

NYC
Health

For additional information or a copy of an inspection report, call 311 or visit nyc.gov/health

Card Number:

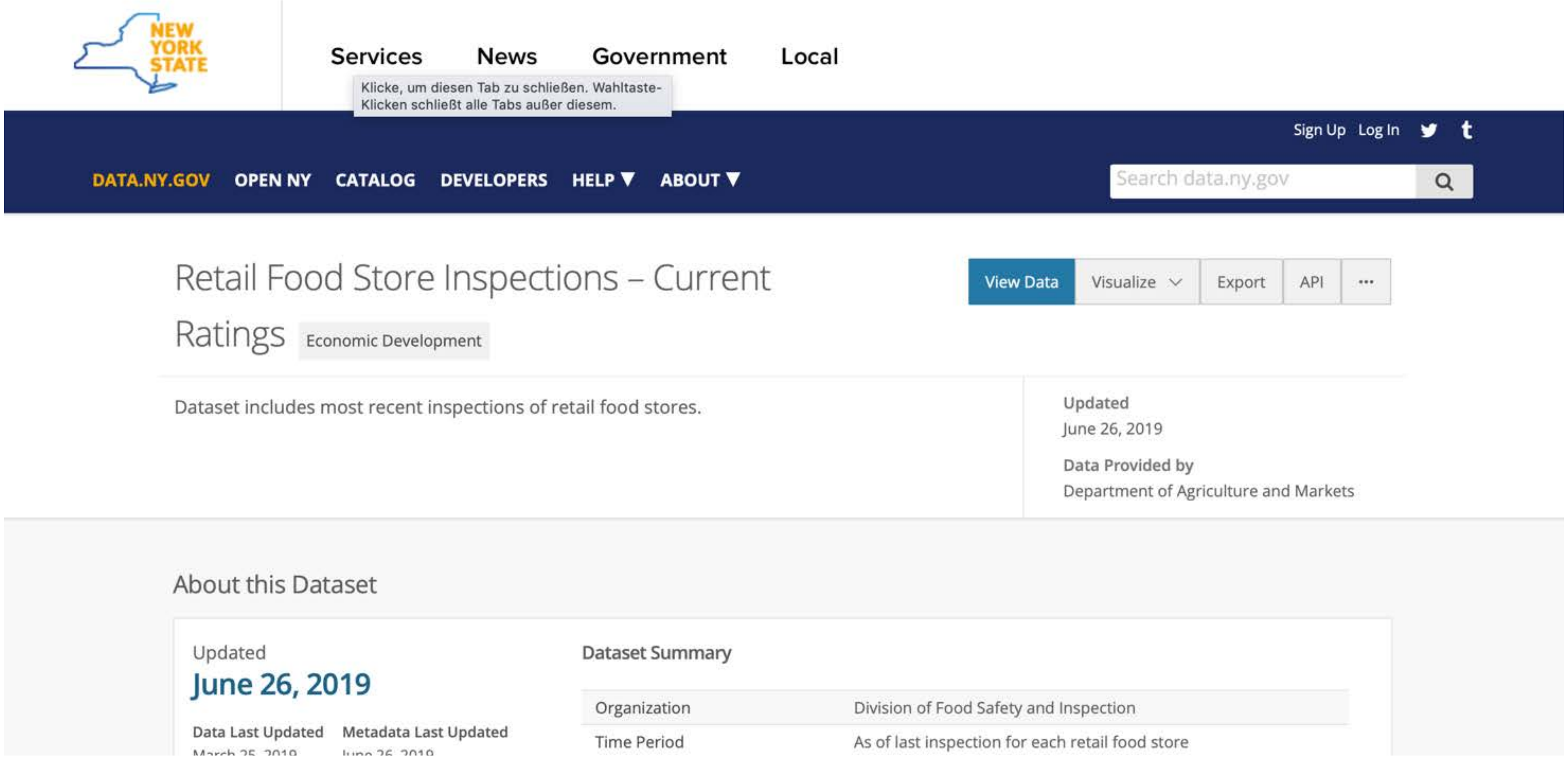
Establishment Name:

Date Issued:

NYC
Health

For additional information or a copy of an inspection report, call 311 or visit nyc.gov/health

The Inspection Data



Klicke, um diesen Tab zu schließen. Wahltaste-Klicken schließt alle Tabs außer diesem.

Retail Food Store Inspections – Current Ratings

Economic Development

Dataset includes most recent inspections of retail food stores.

Updated
June 26, 2019

Data Provided by
Department of Agriculture and Markets

About this Dataset

Updated
June 26, 2019

Data Last Updated	Metadata Last Updated
March 25, 2019	June 26, 2019

Dataset Summary

Organization	Division of Food Safety and Inspection
Time Period	As of last inspection for each retail food store

The Inspection Data

Inspection Grade	Store Name	County	Location	Date	……
A	ZUMY 833 INC	Queens	1007 BRIGHTON BEACH AVE BROOKLYN, NY 11235 (40.578152, -73.959054)	01/16/2019	
A	MACKALLIE LLC	Suffolk	209 UNION AVE NEW ROCHELLE, NY 10801 (40.909533, -73.793911)	04/27/2018	
C	BALS BAGELS	Kings	NA	12/31/2018	
A	QUICKWAY 68	Bronx	324 JACKSON AVE SYOSSET, NY 11791 (40.810935, -73.501746)	05/08/2018	
B	TARGET 2211	Kings	NA	01/16/2019	
⋮	⋮		⋮	⋮	

The Inspection Data

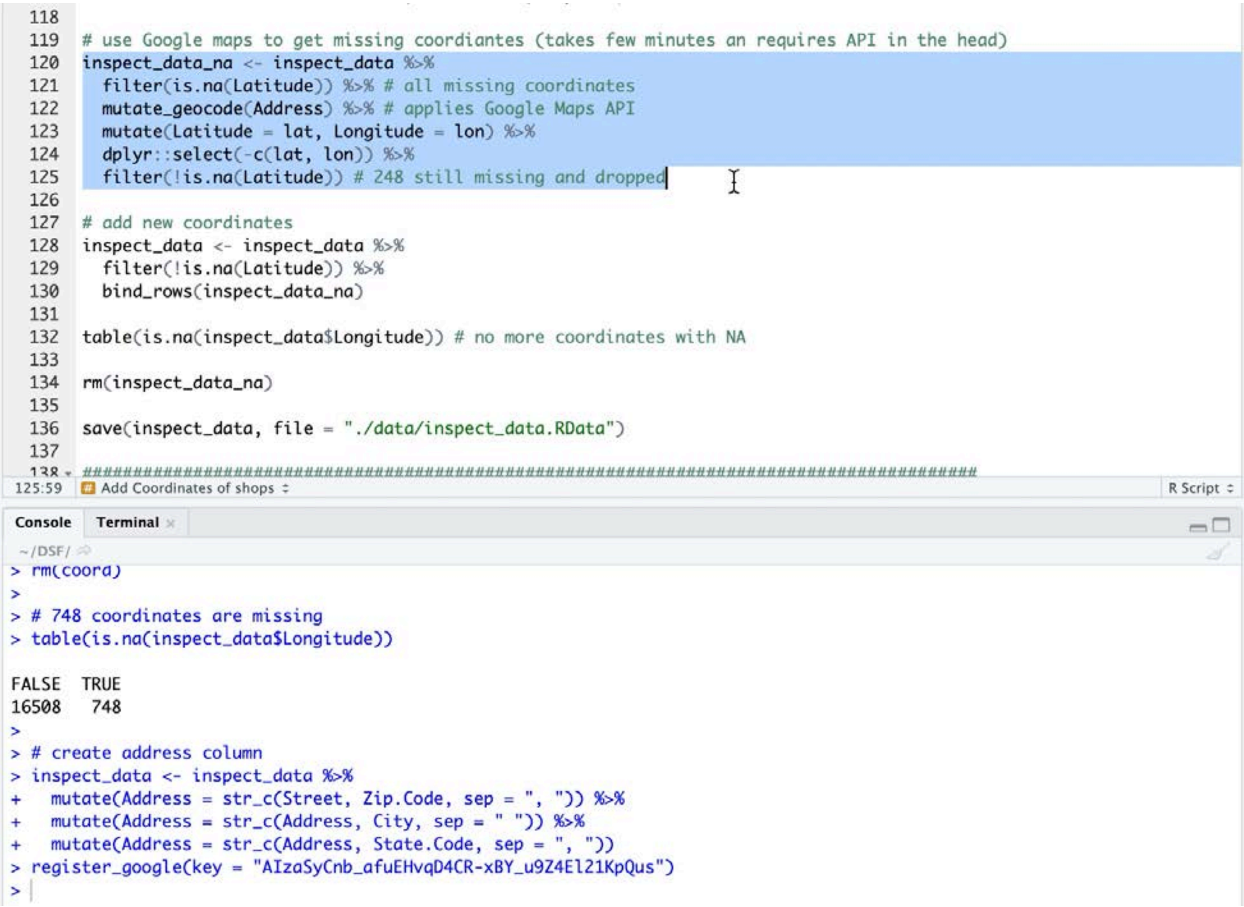
Inspection Grade	Store Name	County	Address	Latitude	Longitude	Date	······
A	ZUMY 833 INC	Queens	1007 BRIGHTON BEACH AVE BROOKLYN, NY 11235	40.578152	-73.959054	01/16/2019	
A	MACKALLIE LLC	Suffolk	209 UNION AVE NEW ROCHELLE, NY 10801	40.909533	-73.793911	04/27/2018	
C	BALS BAGELS	Kings	222 HOYT ST BROOKLYN, NY 11217	NA	NA	12/31/2018	
A	QUICKWAY 68	Bronx	324 JACKSON AVE SYOSSET, NY 11791	40.810935	-73.501746	05/08/2018	
B	TARGET 2211	Kings	204 LIBERTY ST PENN YAN, NY 14527	NA	NA	01/16/2019	
⋮	⋮		⋮	⋮	⋮	⋮	

The Inspection Data

Inspection Grade	Store Name	County	Address	Latitude	Longitude	Date	......
A	ZUMY 833 INC	Queens	1007 BRIGHTON BEACH AVE BROOKLYN, NY 11235	40.578152	-73.959054	01/16/2019	
A	MACKALLIE LLC	Suffolk	209 UNION AVE NEW ROCHELLE, NY 10801	40.909533	-73.793911	04/27/2018	
C	BALS BAGELS	Kings	222 HOYT ST BROOKLYN, NY 11217	NA	NA	12/31/2018	
A	QUICKWAY 68	Bronx	324 JACKSON AVE SYOSSET, NY 11791	40.810935	-73.501746	05/08/2018	
B	TARGET 2211	Kings	204 LIBERTY ST PENN YAN, NY 14527	NA	NA	01/16/2019	
⋮	⋮		⋮	⋮	⋮	⋮	

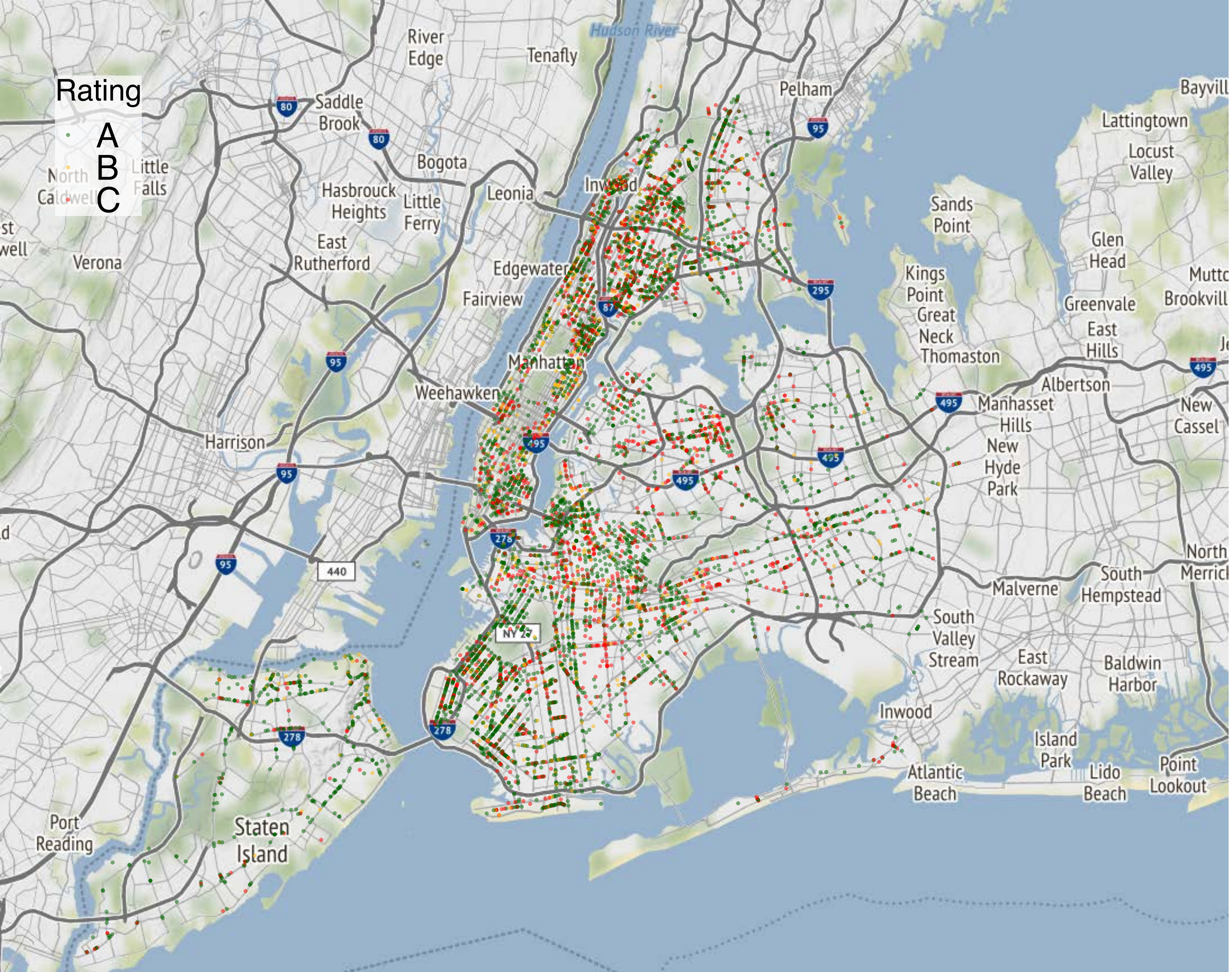
Google Maps API for Missing Coordinates

- Free Service offered by Google
- Requires a one-time registration with a valid Email address
- Then the personal API key must be included in the R script



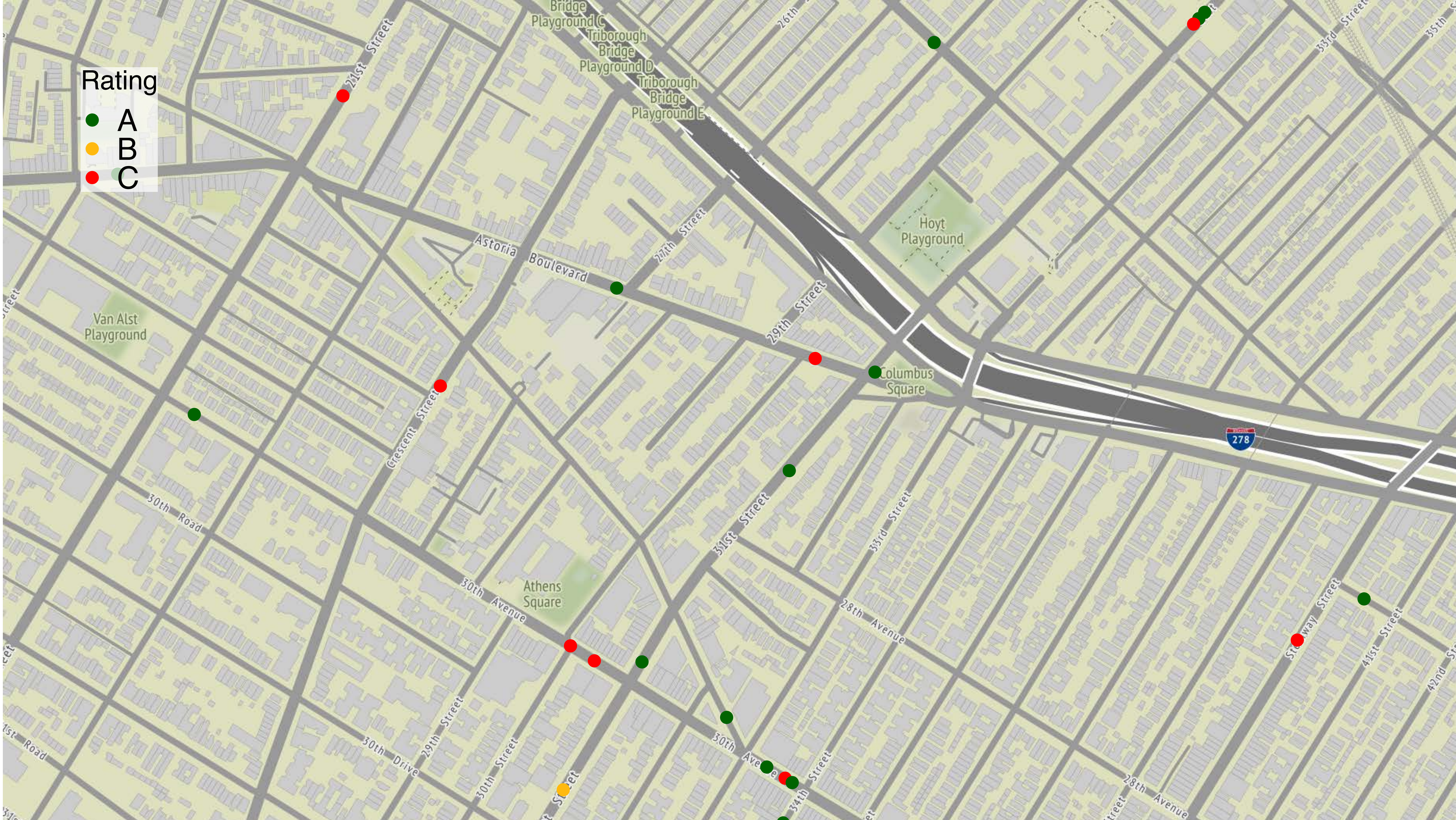
```r

# use Google maps to get missing coordiantes (takes few minutes an requires API in the head)
inspect_data_na <- inspect_data %>%
  filter(is.na(Latitude)) %>% # all missing coordinates
  mutate_geocode(Address) %>% # applies Google Maps API
  mutate(Latitude = lat, Longitude = lon) %>%
  dplyr::select(-c(lat, lon)) %>%
  filter(!is.na(Latitude)) # 248 still missing and dropped

# add new coordinates
inspect_data <- inspect_data %>%
  filter(!is.na(Latitude)) %>%
  bind_rows(inspect_data_na)

table(is.na(inspect_data$Longitude)) # no more coordinates with NA

rm(inspect_data_na)

save(inspect_data, file = "./data/inspect_data.RData")

######################################################################################
```

```
> rm(coord)
>
> # 748 coordinates are missing
> table(is.na(inspect_data$Longitude))

FALSE  TRUE
16508   748
>
> # create address column
> inspect_data <- inspect_data %>%
+   mutate(Address = str_c(Street, Zip.Code, sep = ", ")) %>%
+   mutate(Address = str_c(Address, City, sep = " ")) %>%
+   mutate(Address = str_c(Address, State.Code, sep = ", "))
> register_google(key = "AIzaSyCnb_afuEHvqD4CR-xBY_u9Z4El21KpQus")
>
```



Focus on New York City

- More covariate data available
- Classes are more equally distributed


Rating
A
B
C
Bridge Playground C
Triborough Bridge Playground D
Triborough Bridge Playground E
26th
21st Street
33rd Street
35th
Astoria Boulevard
27th Street
Hoyt Playground
29th Street
Van Alst Playground
Columbus Square
Crescent Street
278
31st Street
33rd Street
30th Road
28th Avenue
Steinway Street
41st Street
42nd
Athens Square
30th Avenue
30th Avenue
29th Street
30th Street
30th Drive
31st Road
34th Street
28th Avenue

Haversine Formula

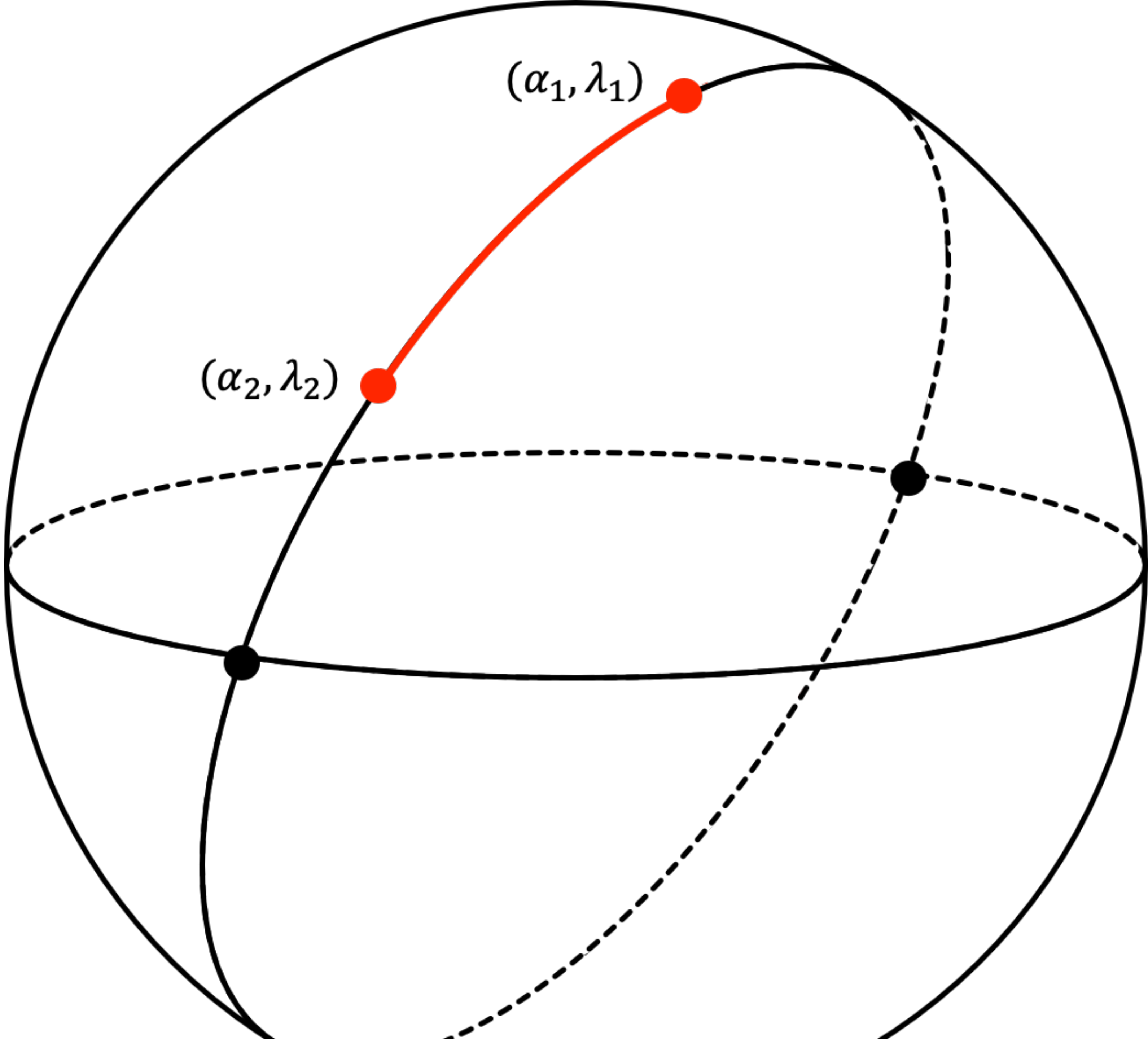
$a = sin^2\left(\frac{\Delta\alpha}{2}\right) + \cos(\alpha_1)\cos(\alpha_2)\, sin^2\left(\frac{\Delta\lambda}{2}\right)$

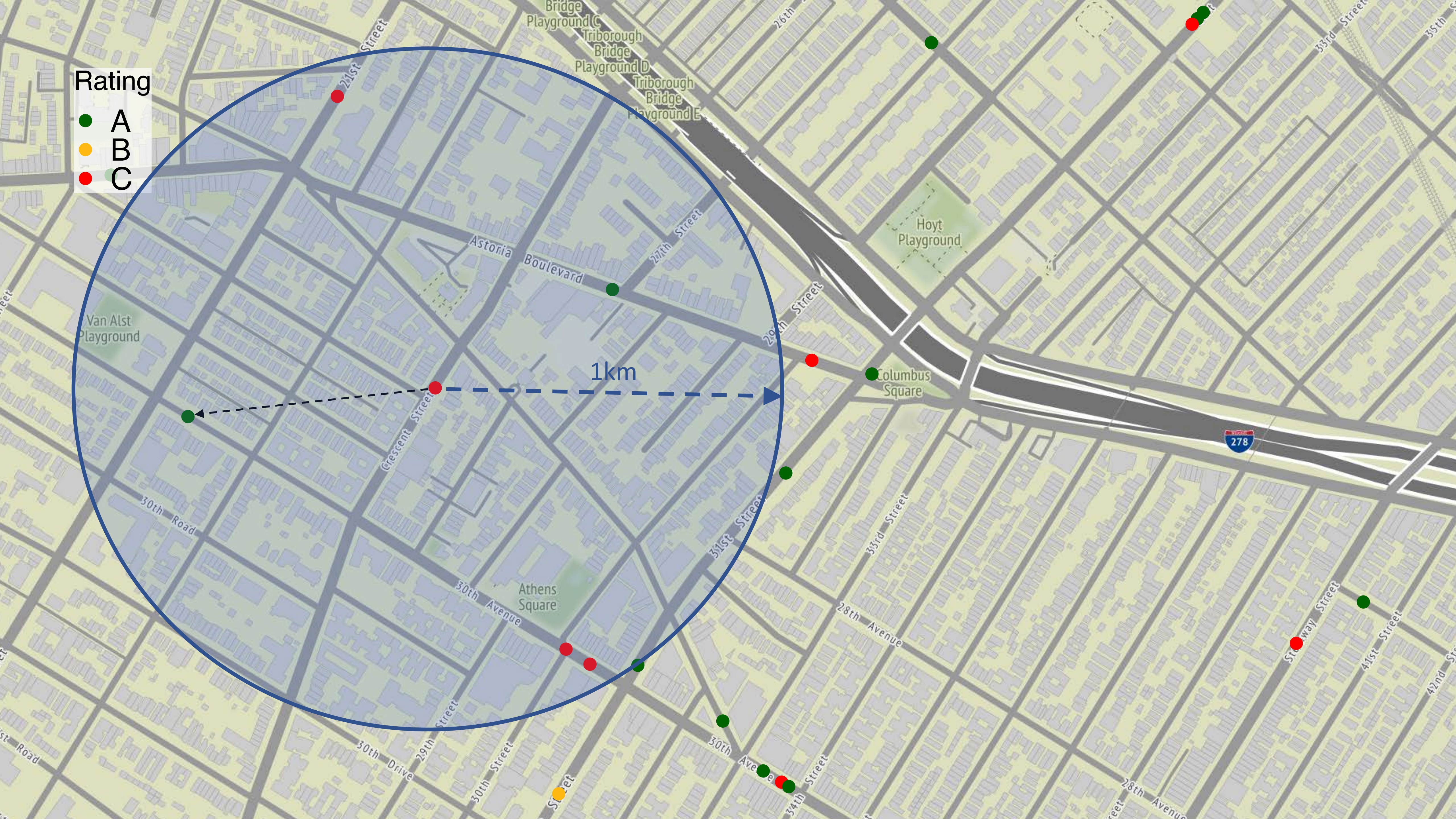
$c = R\left[2\,\text{atan}^2(\sqrt{a}, \sqrt{1-a})\right]$

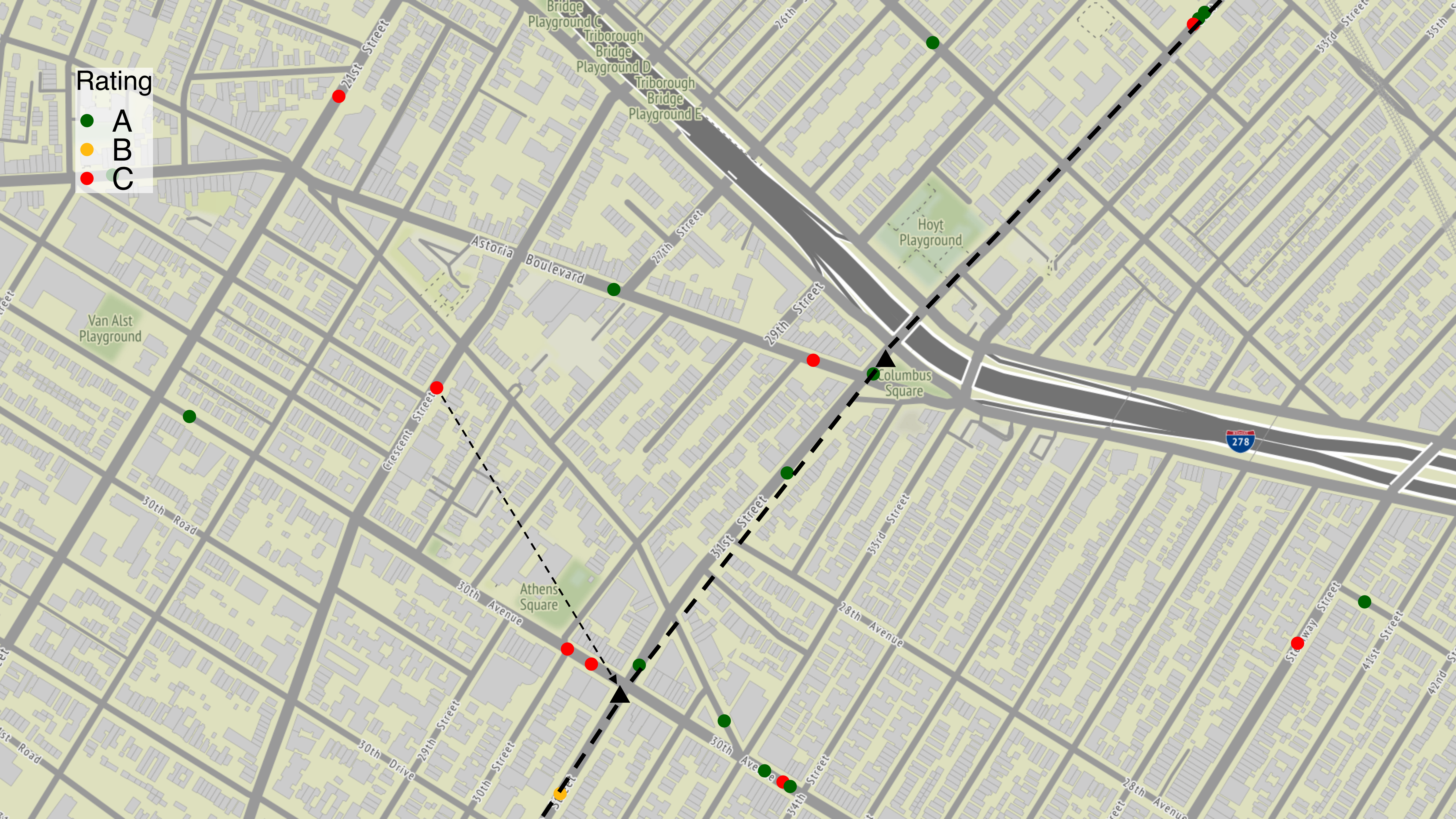
With:

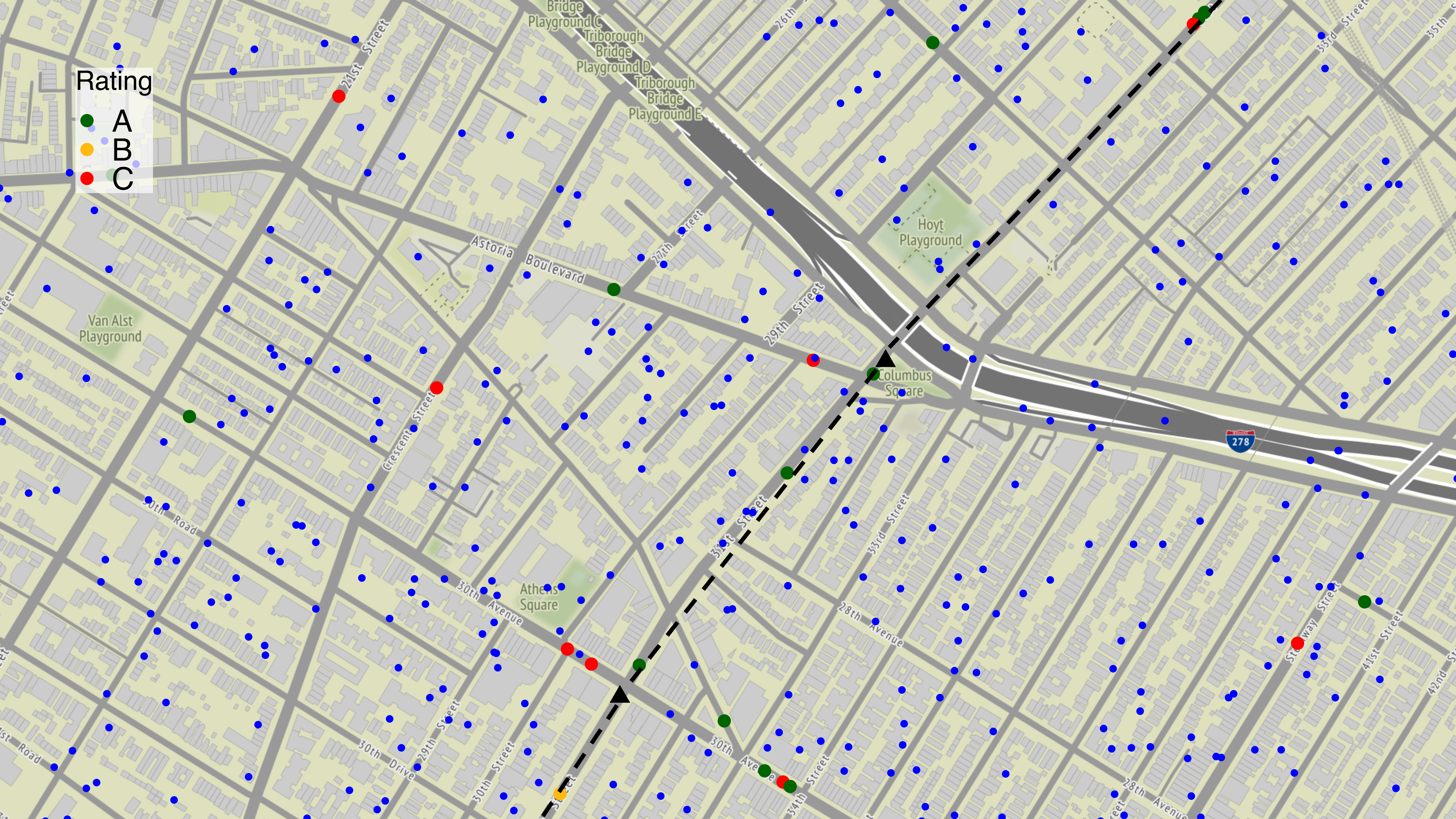
R = (Mean) radius of the earth (6,371km)

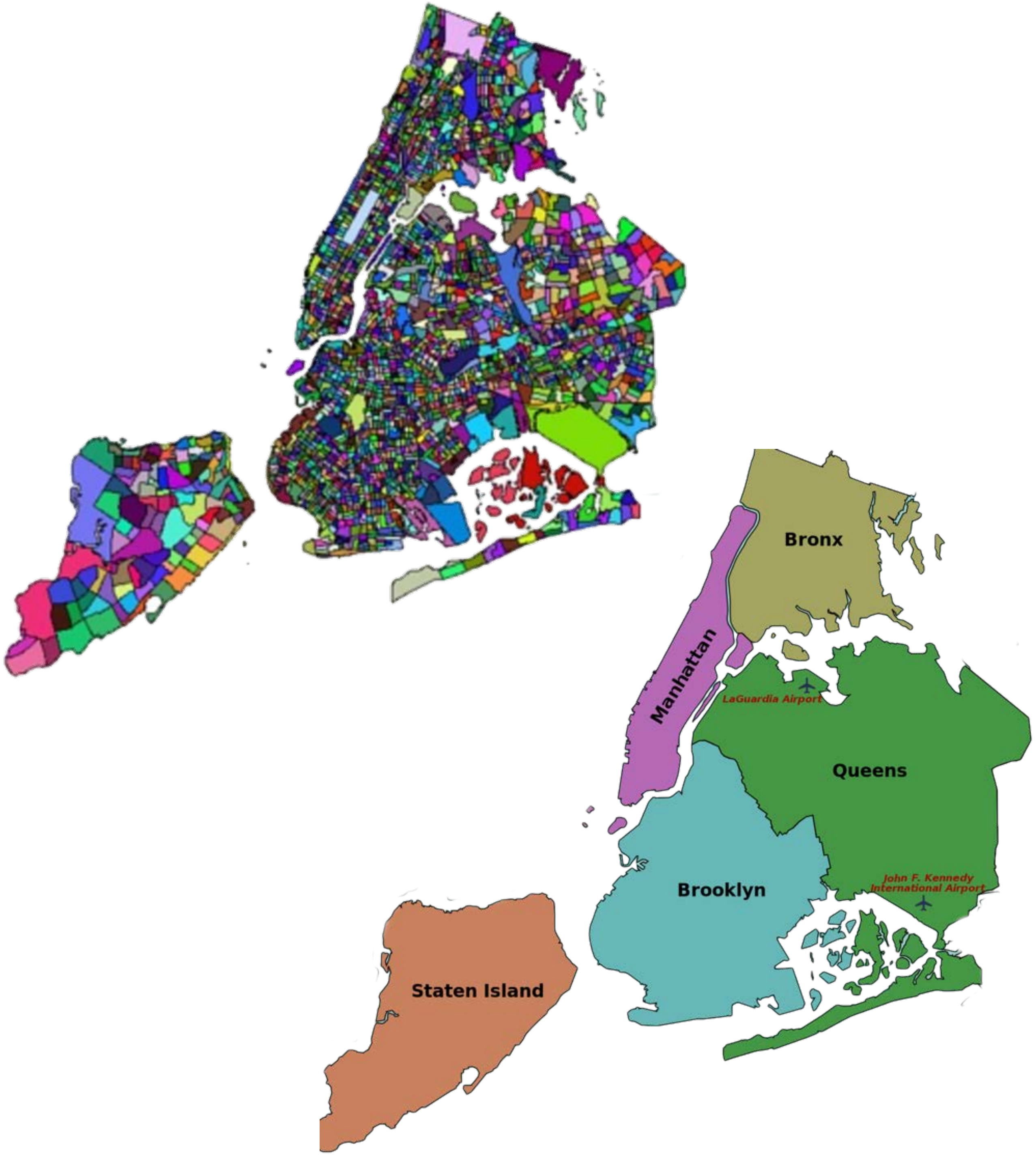
$\alpha_{1,2} = Latitude\ \lambda_{1,2} = Longitude$




Rating
● A
● B
● C
1km
Bridge
Playground C
Triborough
Bridge
Playground D
Triborough
Bridge
Playground E
26th
21st Street
Street
27th Street
Astoria Boulevard
Van Alst
Playground
Hoyt
Playground
29th Street
Columbus
Square
Crescent Street
31st Street
33rd Street
33rd Street
35th
278
30th Road
30th Avenue
Athens
Square
28th Avenue
Steinway Street
41st Street
42nd
29th Street
30th Street
30th Drive
Street
34th Street
30th Avenue
28th Avenue
31st Road


Rating
● A
● B
● C
Bridge Playground C
Triborough Bridge Playground D
Triborough Bridge Playground E
26th
33rd Street
35th Street
21st Street
Hoyt Playground
Astoria Boulevard
27th Street
29th Street
Van Alst Playground
Columbus Square
278
Crescent Street
30th Road
31st Street
33rd Street
28th Avenue
Athens Square
30th Avenue
Steinway Street
41st Street
42nd
30th Drive
29th Street
30th Street
31st Road
34th Street
28th Avenue


Rating
● A
● B
● C
Bridge Playground C
Triborough Bridge Playground D
Triborough Bridge Playground E
26th
33rd Street
35th
21st Street
27th Street
Astoria Boulevard
29th Street
Hoyt Playground
Columbus Square
Van Alst Playground
Crescent Street
278
30th Road
31st Street
33rd Street
Athens Square
30th Avenue
28th Avenue
Steinway Street
41st Street
42nd Street
29th Street
30th Drive
30th Street
34th Street
28th Avenue
30th Avenue
st Road

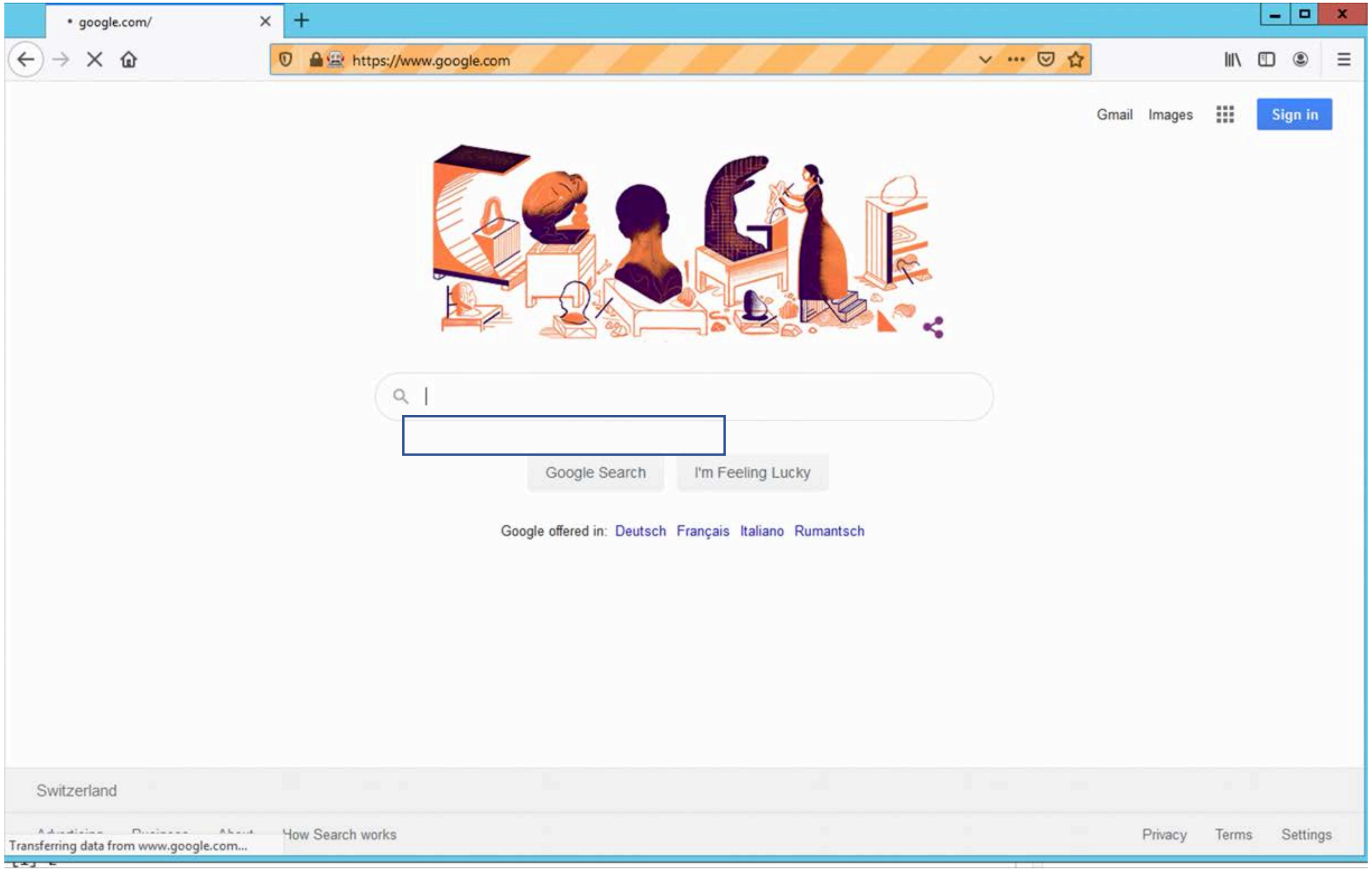


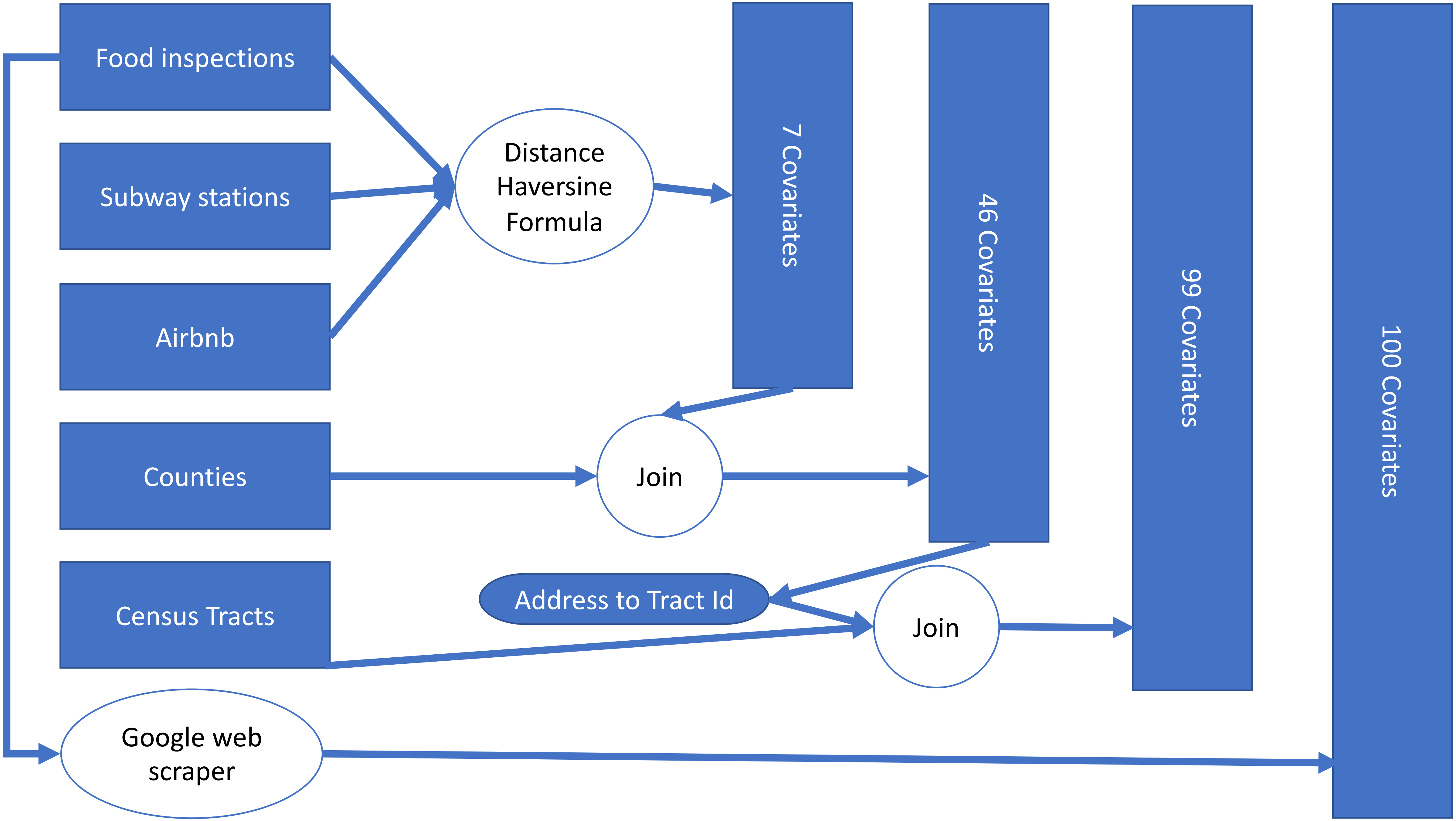
Demographic Data

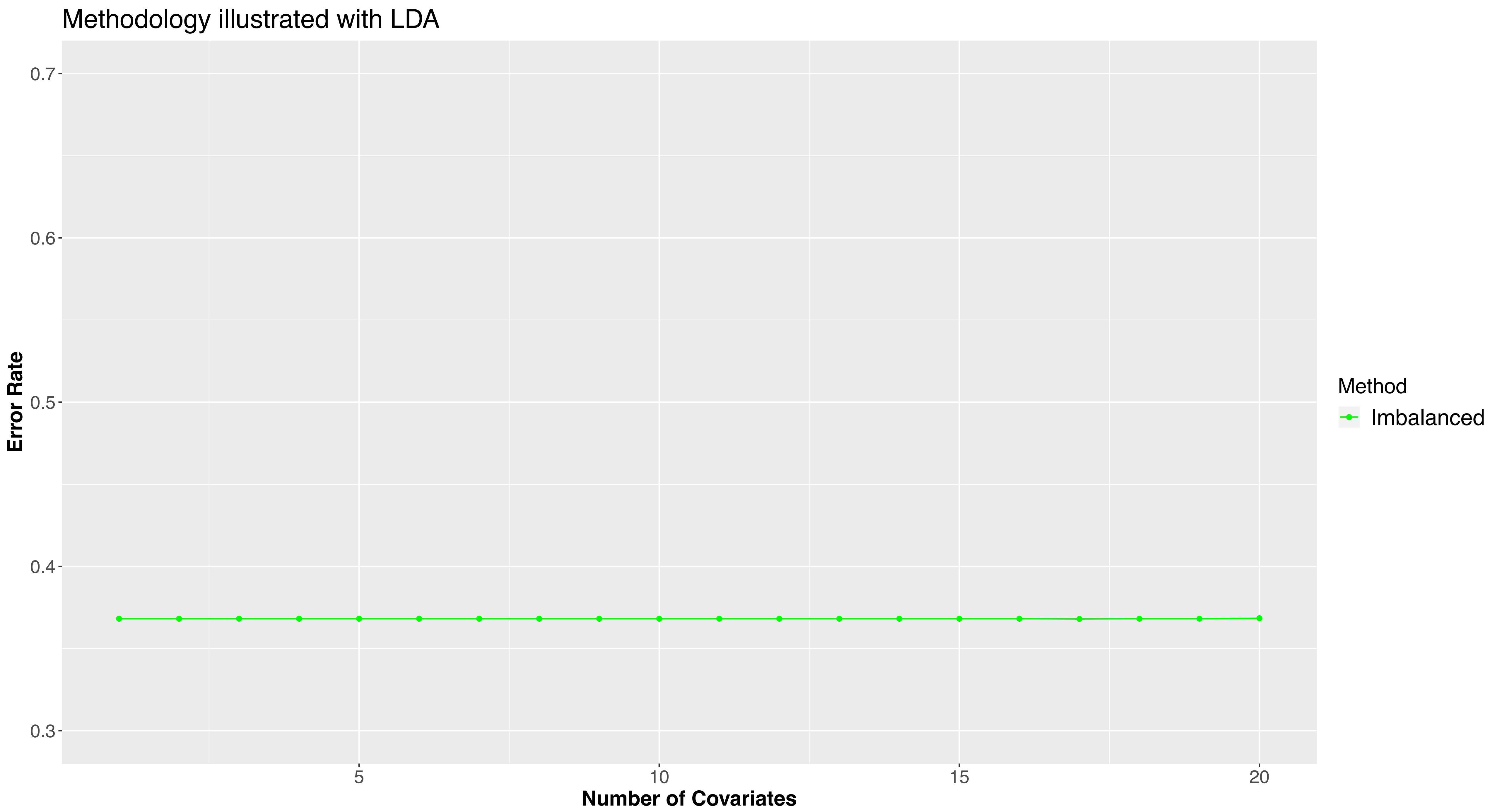
- U.S. Census Bureau: Counties
  - merging via counties
- U.S. Census Bureau: Census Tracts
  - merging via census tract
  - Translation necessary: AddTrac
  - Geocoding service: Addresses
    - State FIPS
    - County FIPS
    - Census Tract Id

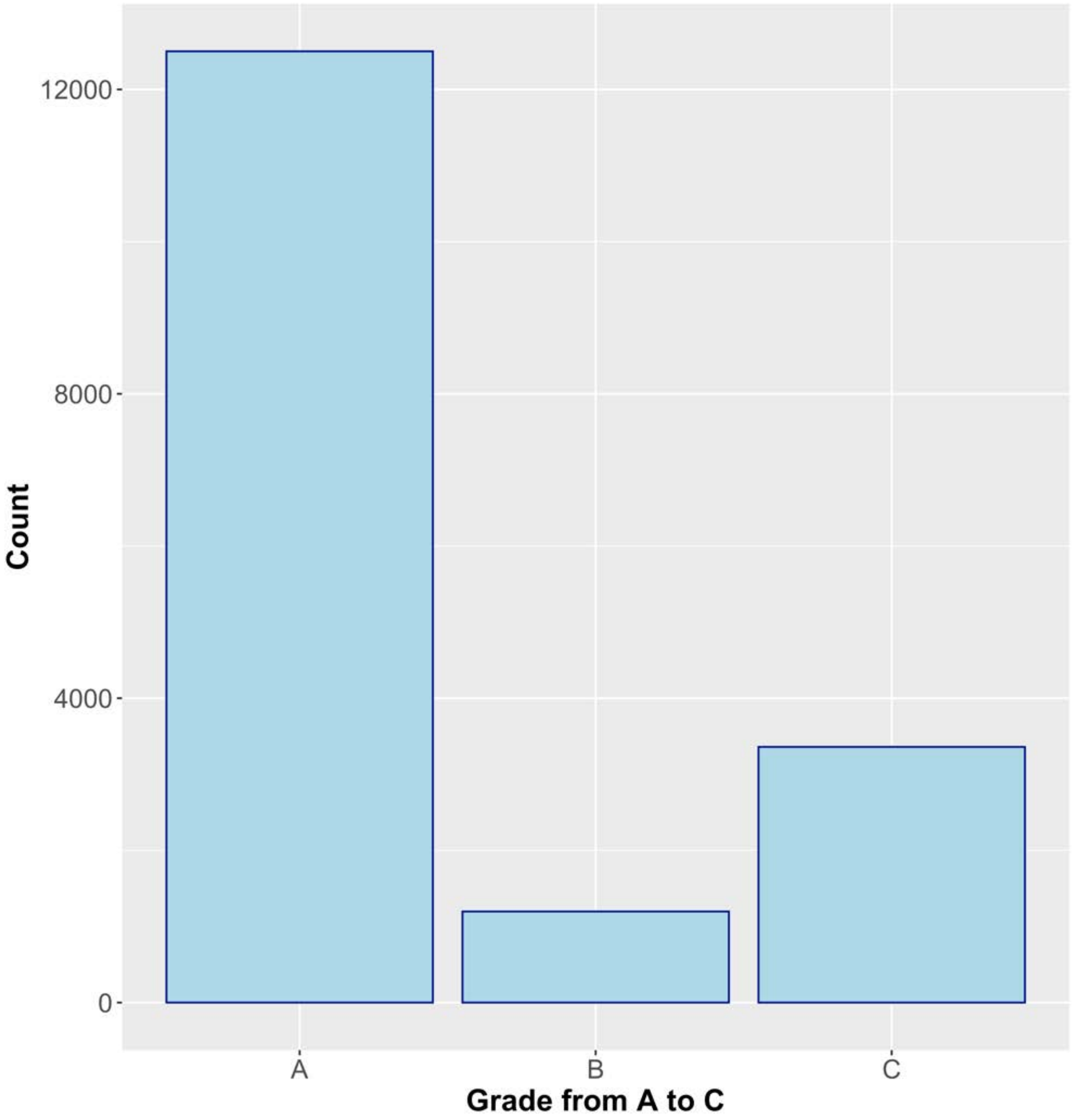
Google web scraper

- Automated Googles search
- Gather Google star ratings and number of reviews

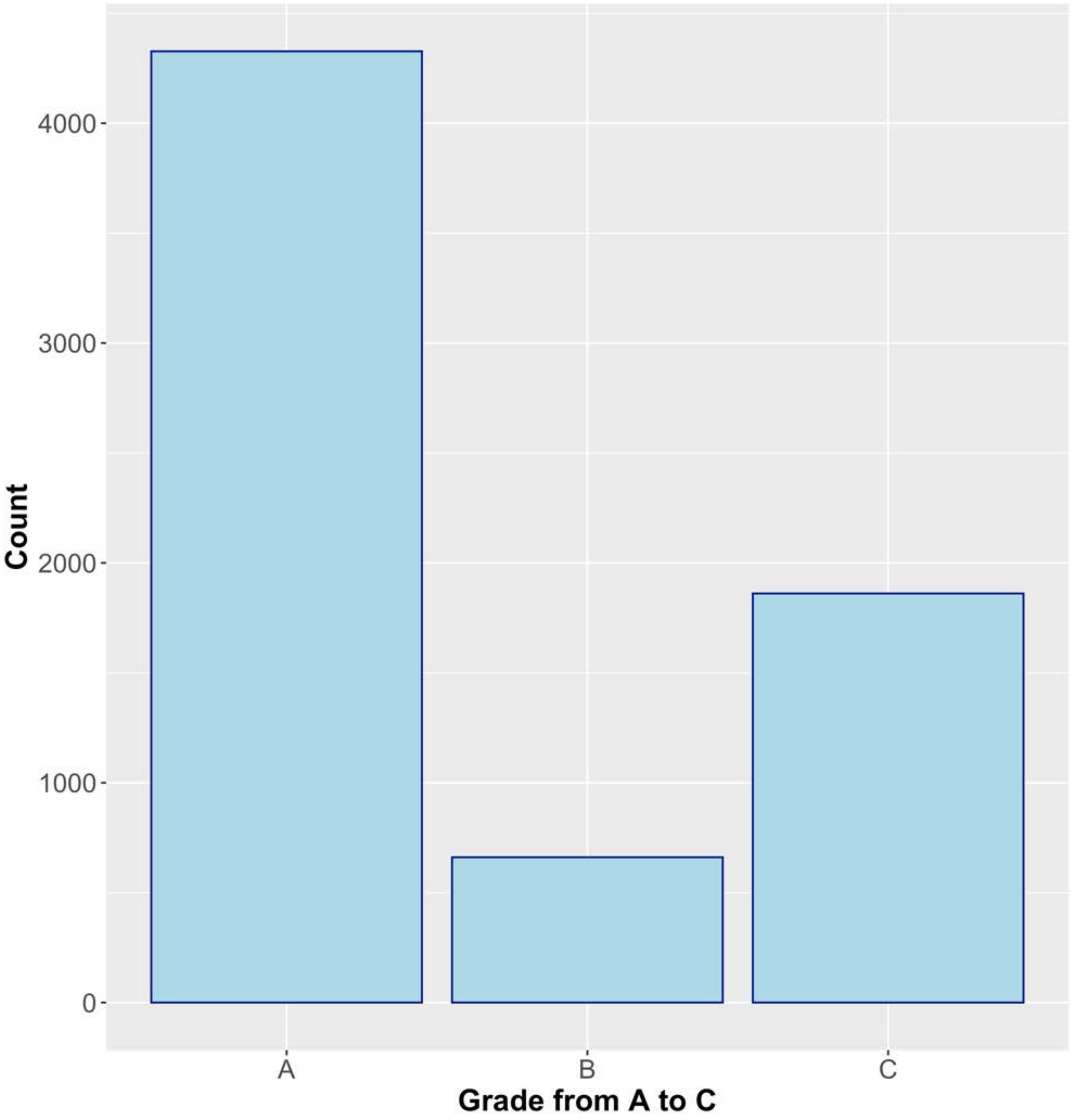



Food inspections
Subway stations
Airbnb
Counties
Census Tracts
Google web scraper
Distance Haversine Formula
7 Covariates
Join
46 Covariates
Address to Tract Id
Join
99 Covariates
100 Covariates


Methodology illustrated with LDA
Error Rate
0.7
0.6
0.5
0.4
0.3
Number of Covariates
5
10
15
20
Method
Imbalanced

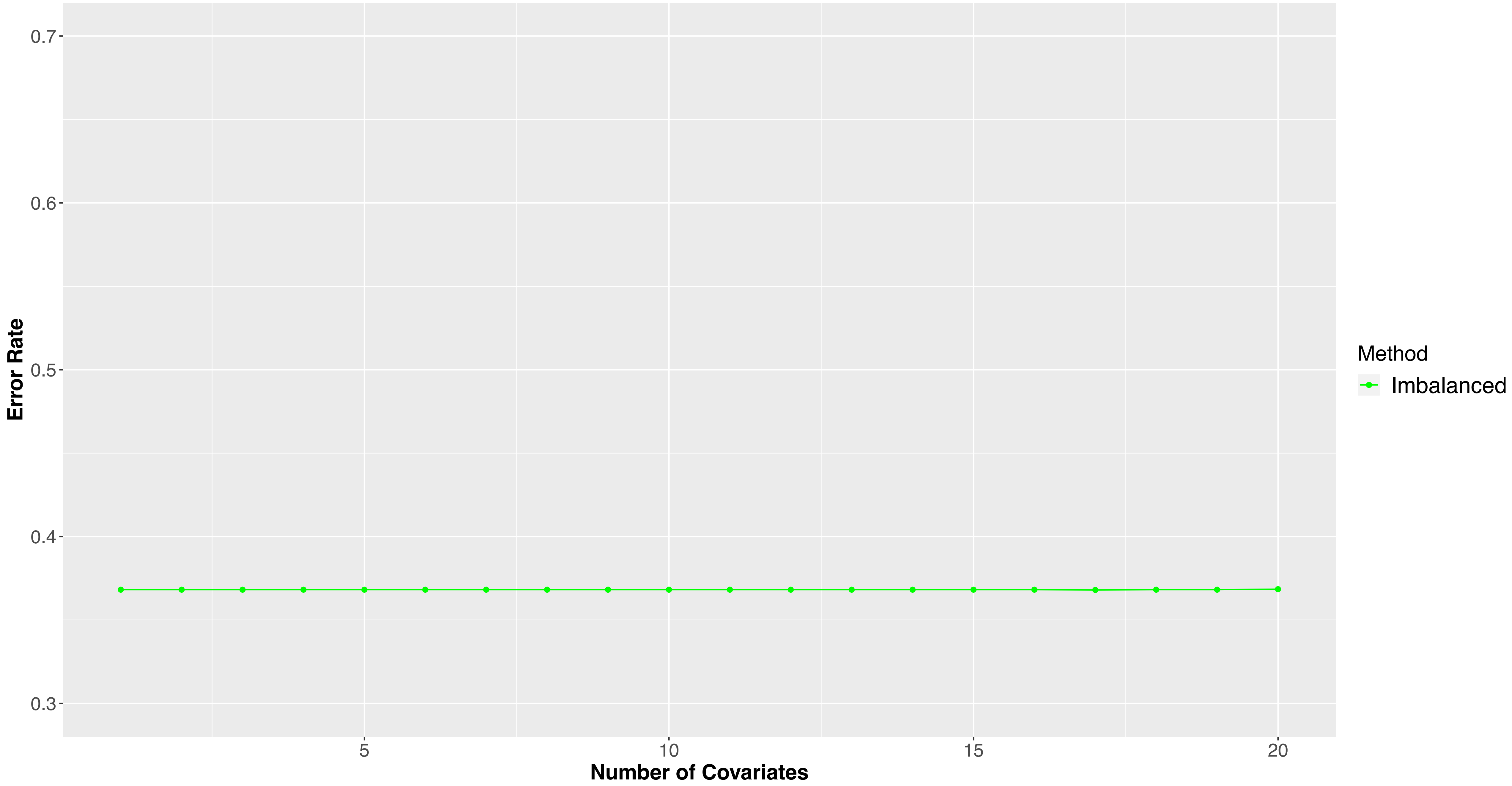


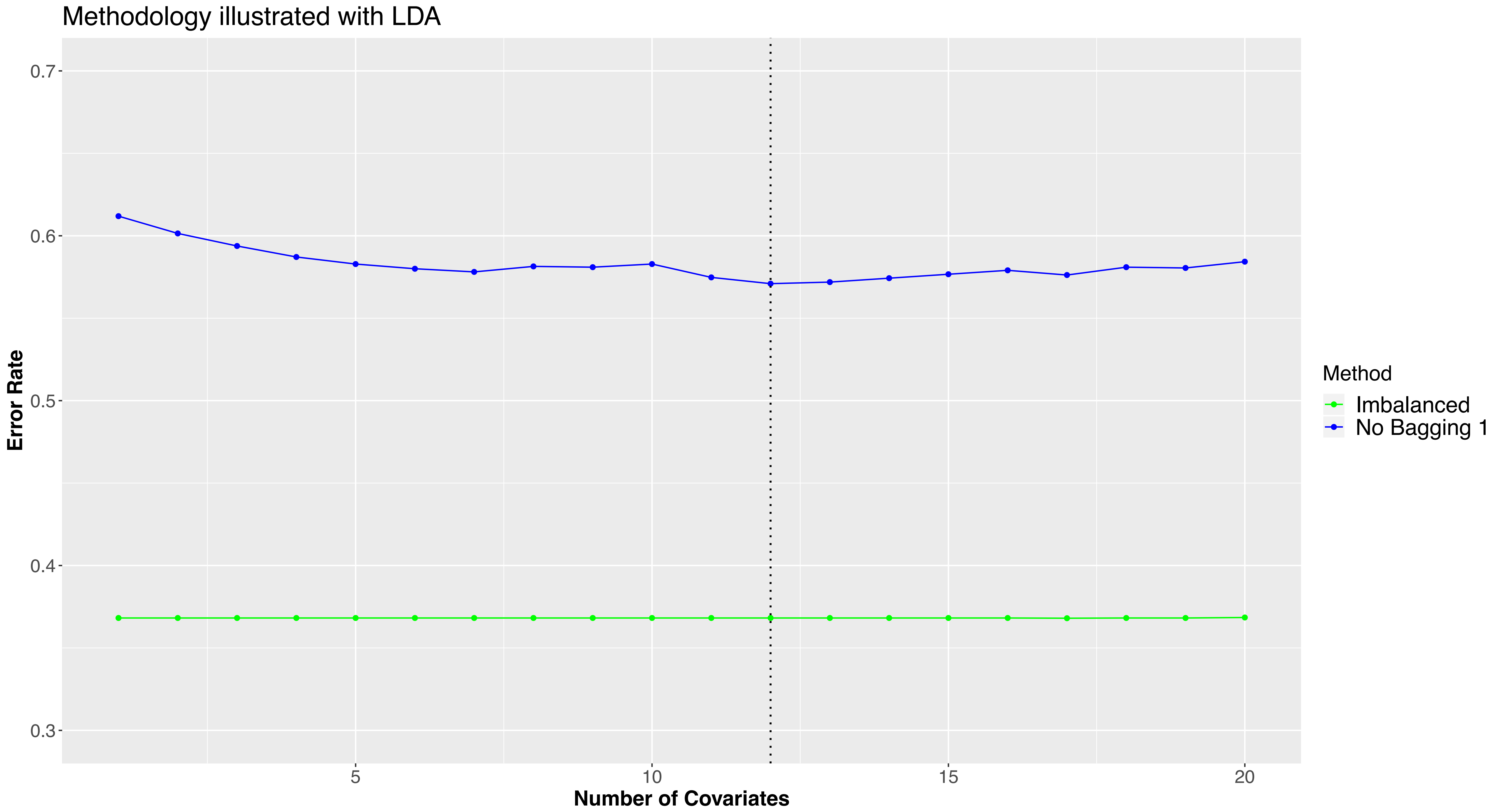
NY State - Histogram of Inspection Grades

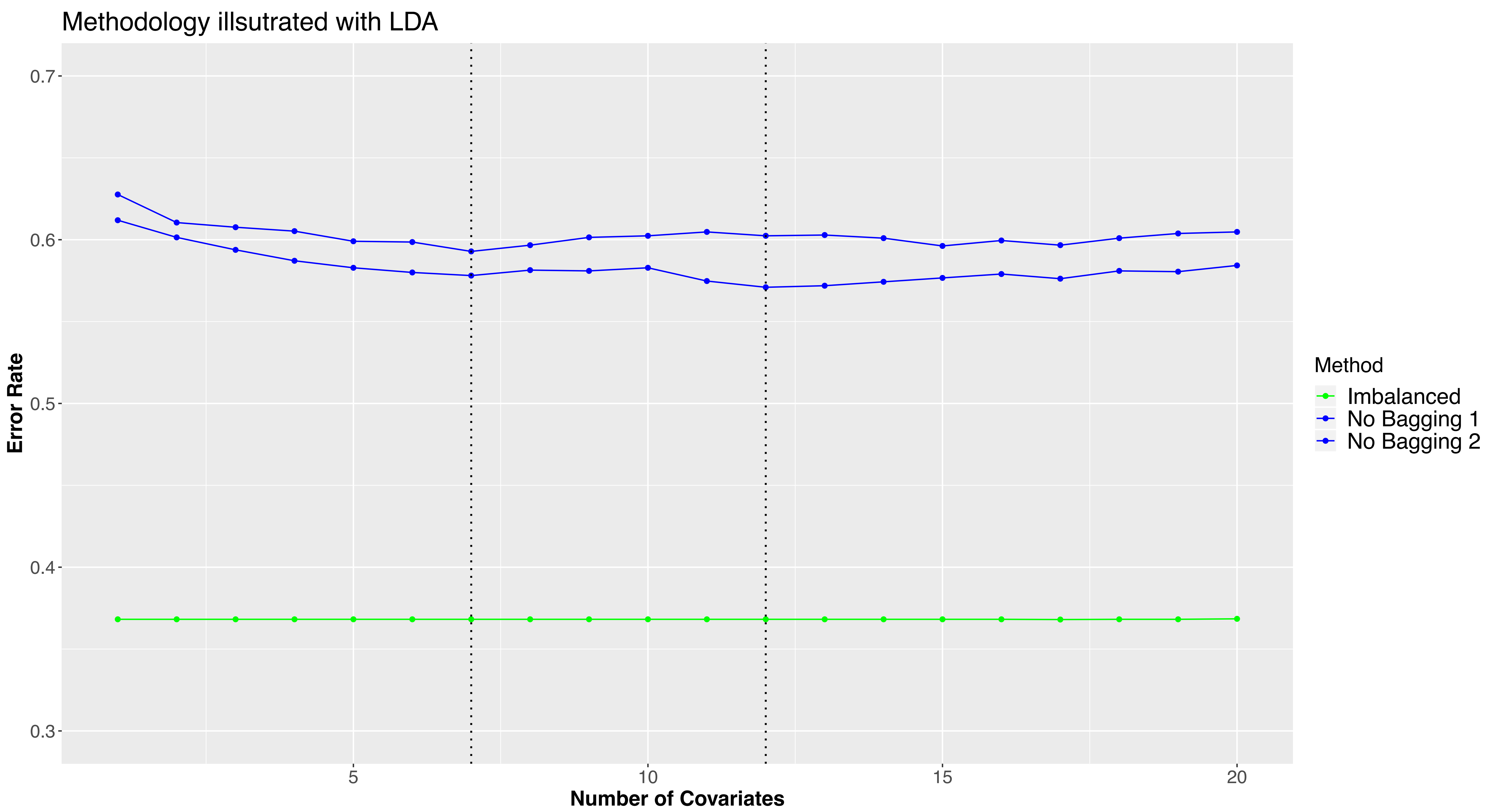


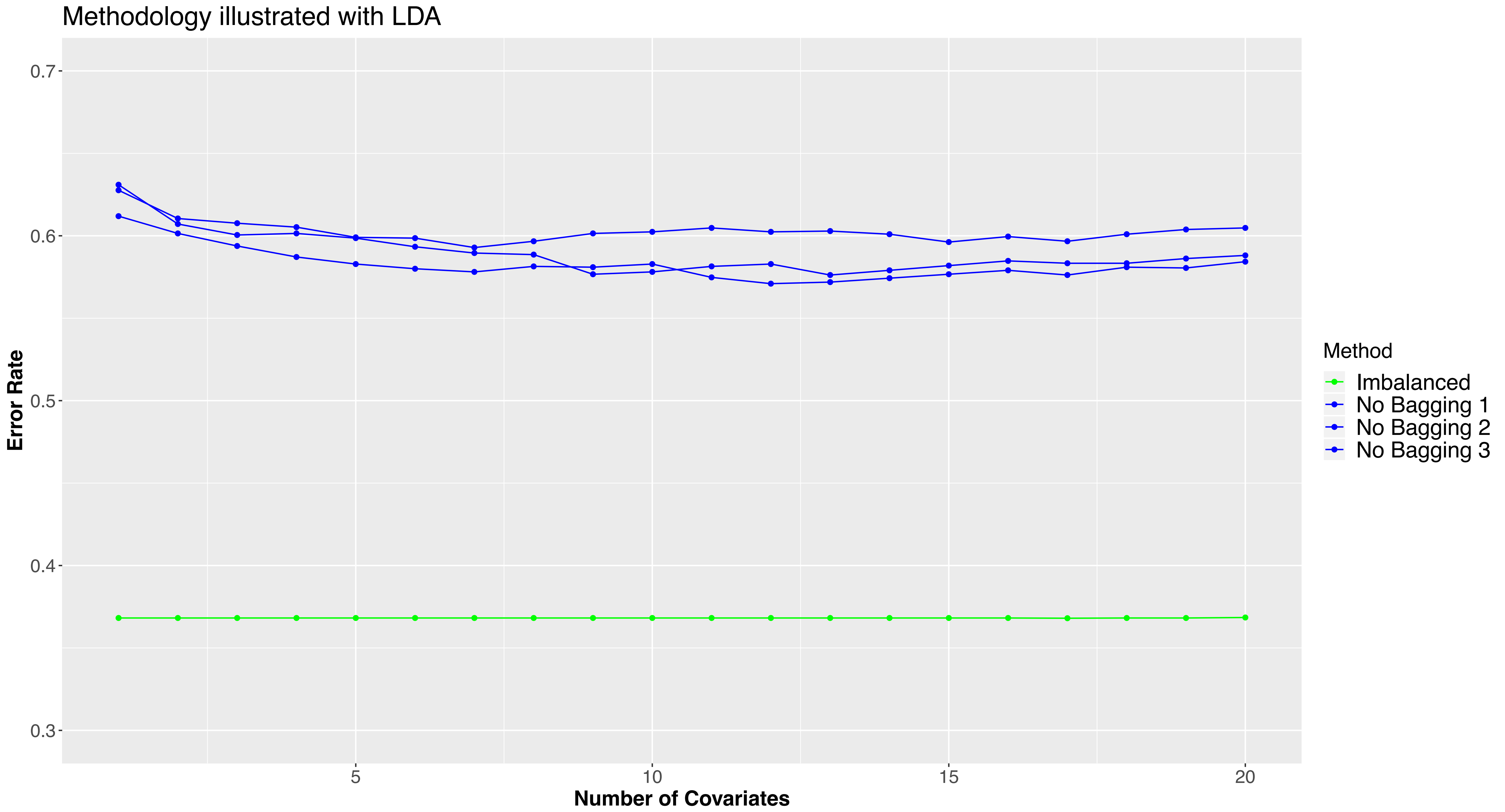
NY City - Histogram of Inspection Grades

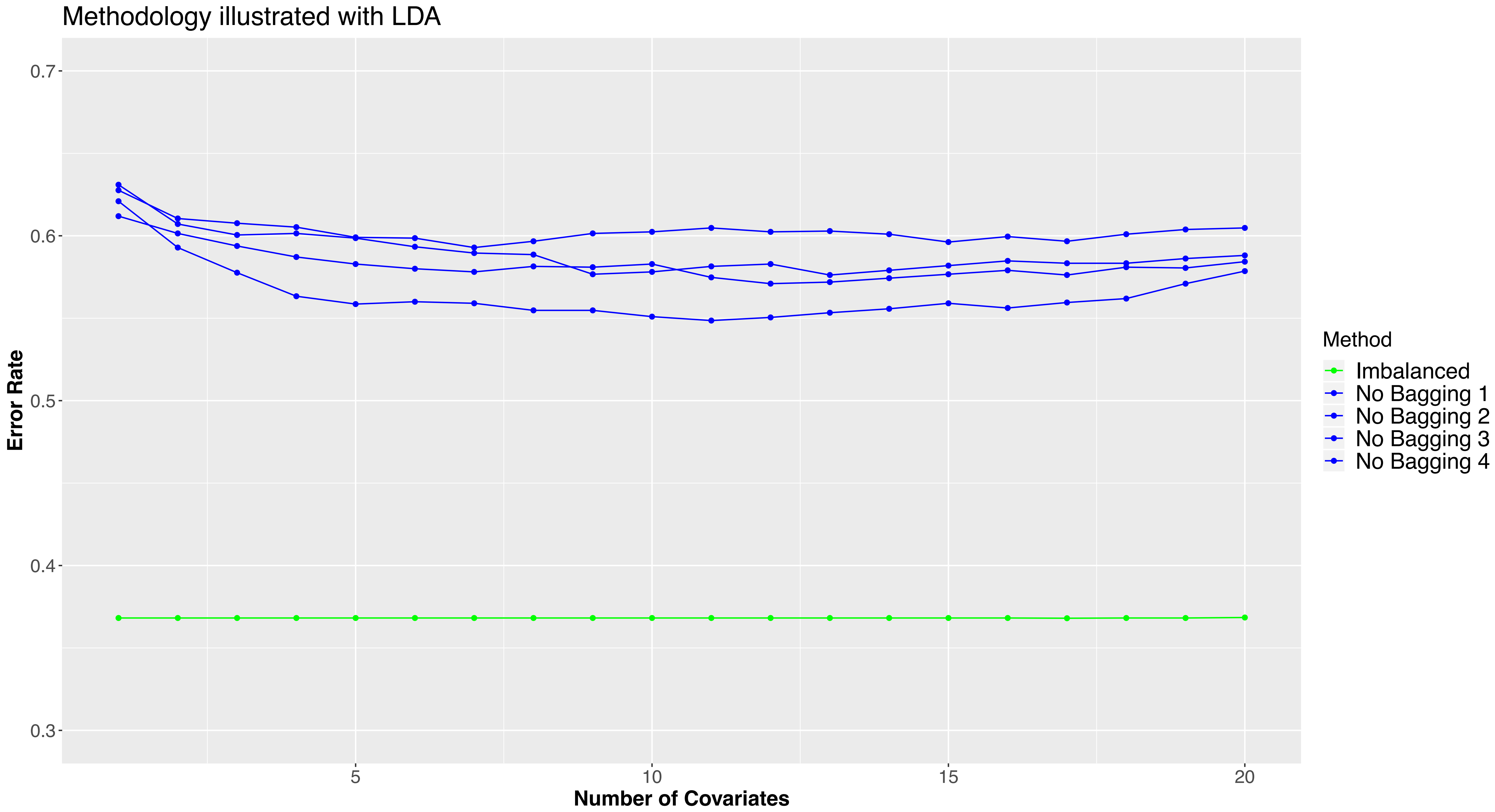
Methodology illustrated with LDA

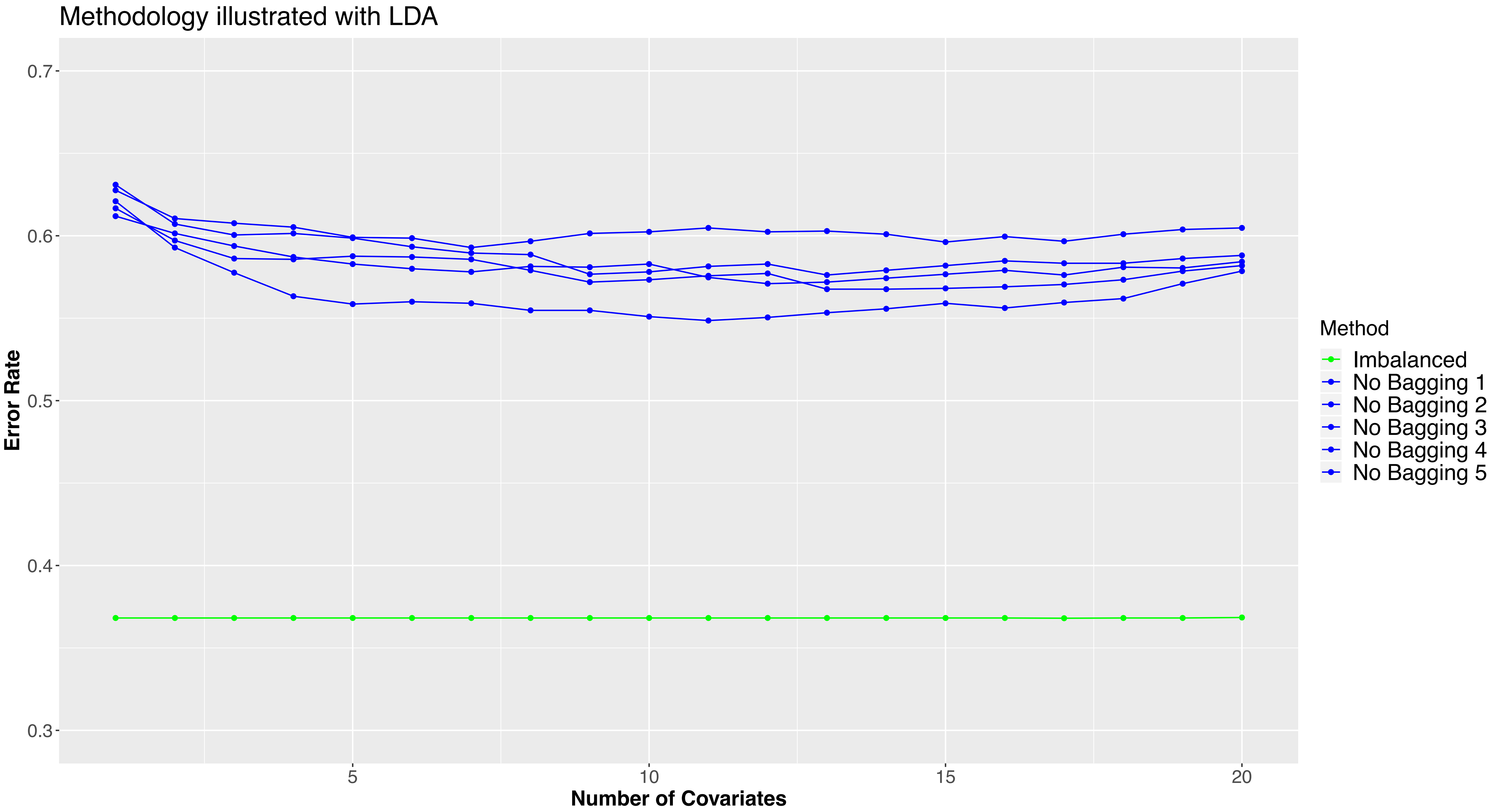



Methodology illustrated with LDA
Error Rate
0.7
0.6
0.5
0.4
0.3
5
10
15
20
Number of Covariates
Method
Imbalanced
No Bagging 1

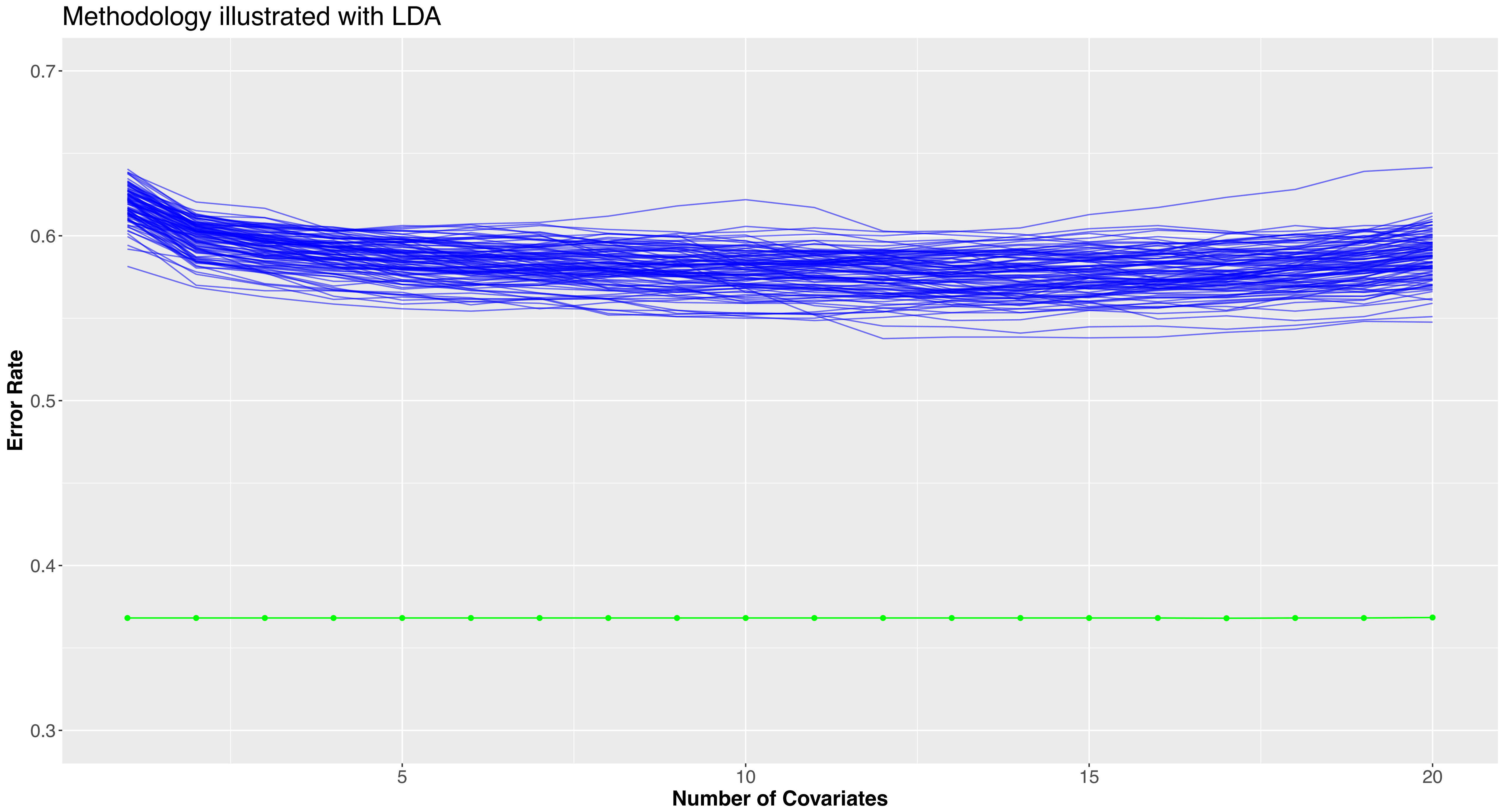

Methodology illsutrated with LDA
0.7
0.6
0.5
0.4
0.3
Error Rate
5
10
15
20
Number of Covariates
Method
Imbalanced
No Bagging 1
No Bagging 2

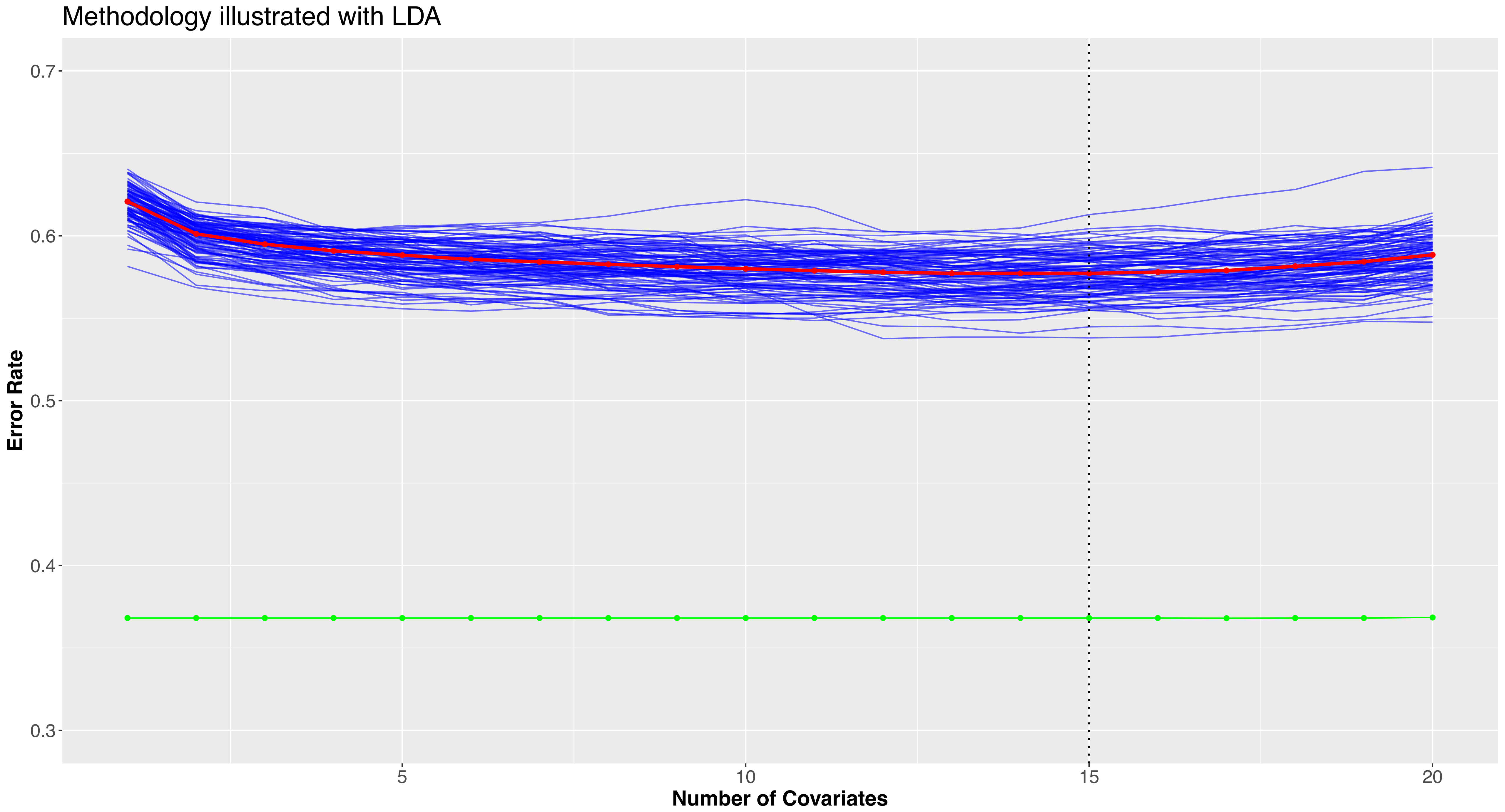

Methodology illustrated with LDA
Error Rate
0.3
0.4
0.5
0.6
0.7
Number of Covariates
5
10
15
20
Method
Imbalanced
No Bagging 1
No Bagging 2
No Bagging 3


Methodology illustrated with LDA
Error Rate
0.7
0.6
0.5
0.4
0.3
Number of Covariates
5
10
15
20
Method
Imbalanced
No Bagging 1
No Bagging 2
No Bagging 3
No Bagging 4



Methodology illustrated with LDA


Methodology illustrated with LDA
Error Rate
0.7
0.6
0.5
0.4
0.3
5
10
15
20
Number of Covariates


Methodology illustrated with LDA
Error Rate
0.7
0.6
0.5
0.4
0.3
Number of Covariates
5
10
15
20

Evaluation of Over- and Under-Bagging

- Under-bagging:
    - No loss of information[1]
- Over-bagging:
    - Reduce the risk for over-fitting[1]

- Computationally Intensive[1]

[1] According to Fernandez et al. (2018, pp. 82 - 83, 175 - 176)

1st Model Selection Approach

- **1)** 100 Bagged Samples **2)** 10-Fold-CV **3)** Best Subset Selection[1]
- Three Functions:
  - Best_subset_selection <− function(df train , df test , Y, FUN){ … }
  - K_fold_CV<− function(df, Y, K, FUN){ … }
  - Over_under_bagging <− function(df, Y, B, sample size, FUN){ …}

- Number of estimated models:
  - $B \times K \times 2^p \approx 1\ Billion$ with $B = 100, K = 10, p = 100$

[1] According to James et al. (2017, p. 205)

2nd Model Selection Approach

- Increase the function's efficiency

- Instead of CV, use out-of-bag errors:
  - $B \times K \times 2^p \approx 100\ Million$ with $B = 100, K = 1, p = 100$

3rd Model Selection Approach

- Instead of best subset selection, new function for forward stepwise selection[1]:
  - forward_stepwise_selection <− function(df train , df test , Y, FUN){ … }
  - $B \times K \times \left(\frac{p(p+1)}{2} - 1\right) \approx 10\ Million$ with $B = 100, K = 10, p = 100$

[1] According to James et al. (2017, p. 207)

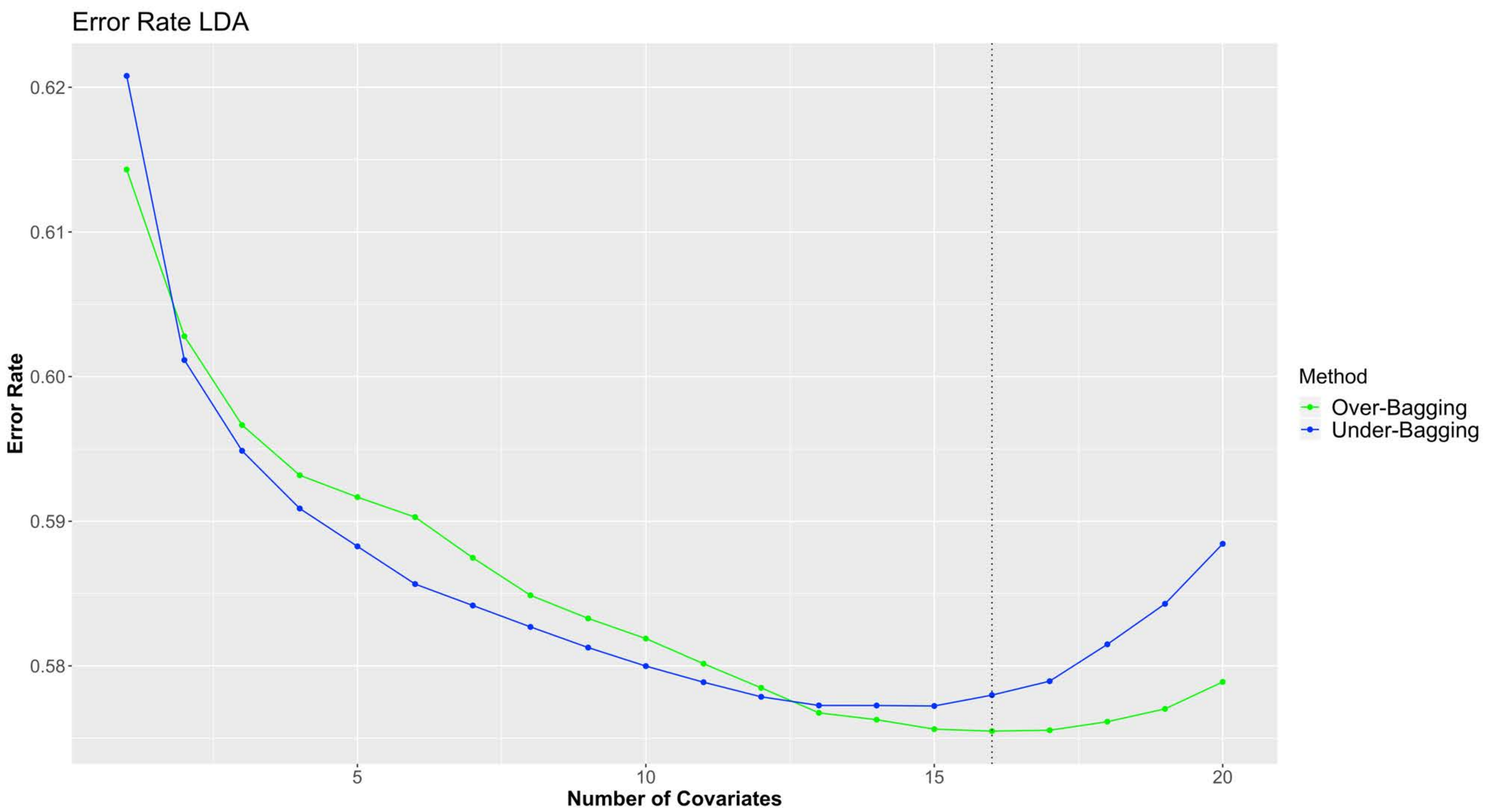
4th Model Selection Approach

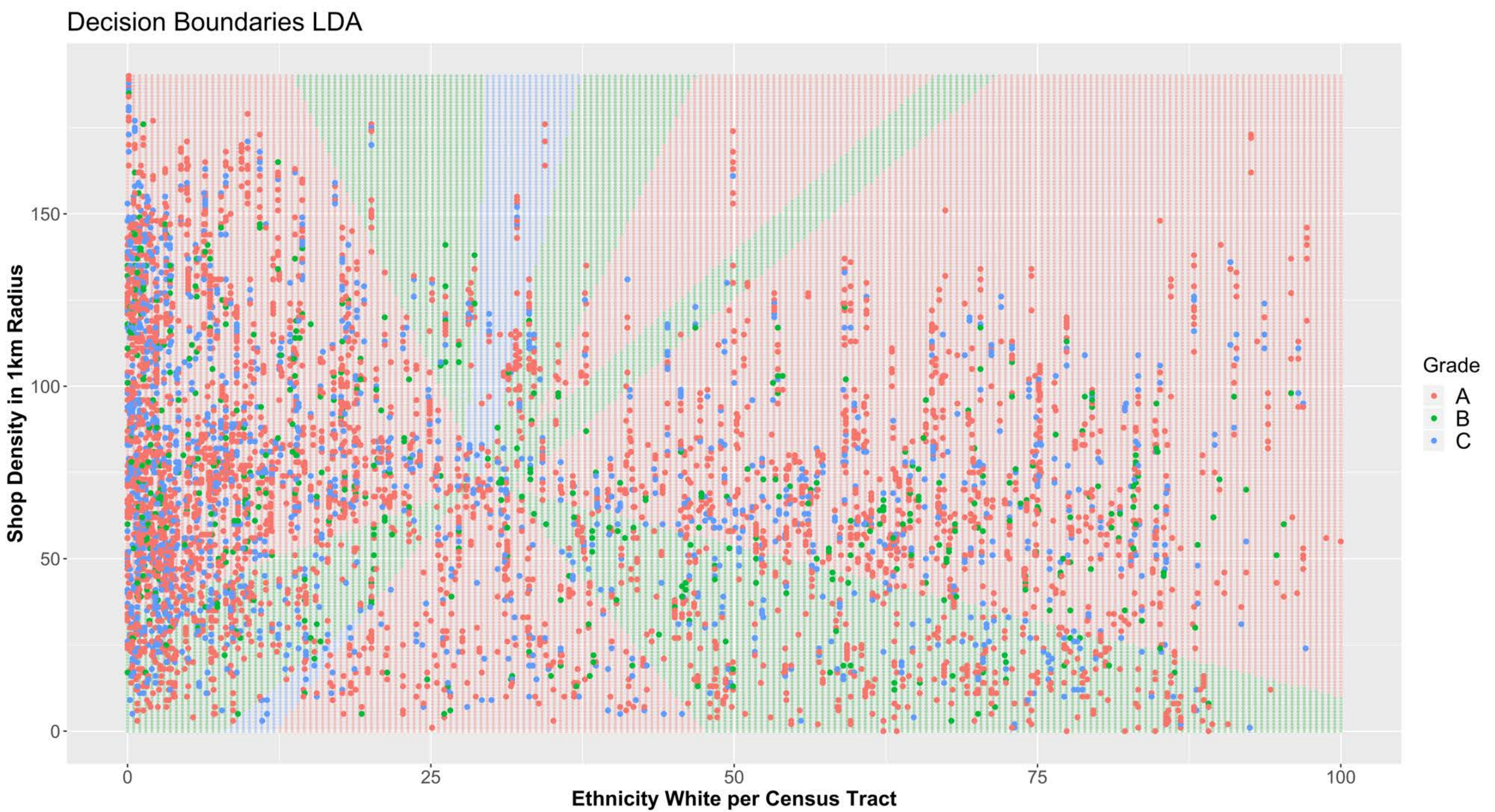
- Reducing the number of covariates to 20
  - Some demographic could be eliminated due to perfect multicollinearity (e.g. Ethnicity per census track)
  - Highest correlation to Inspection Grades (Not an optimal approach[1])

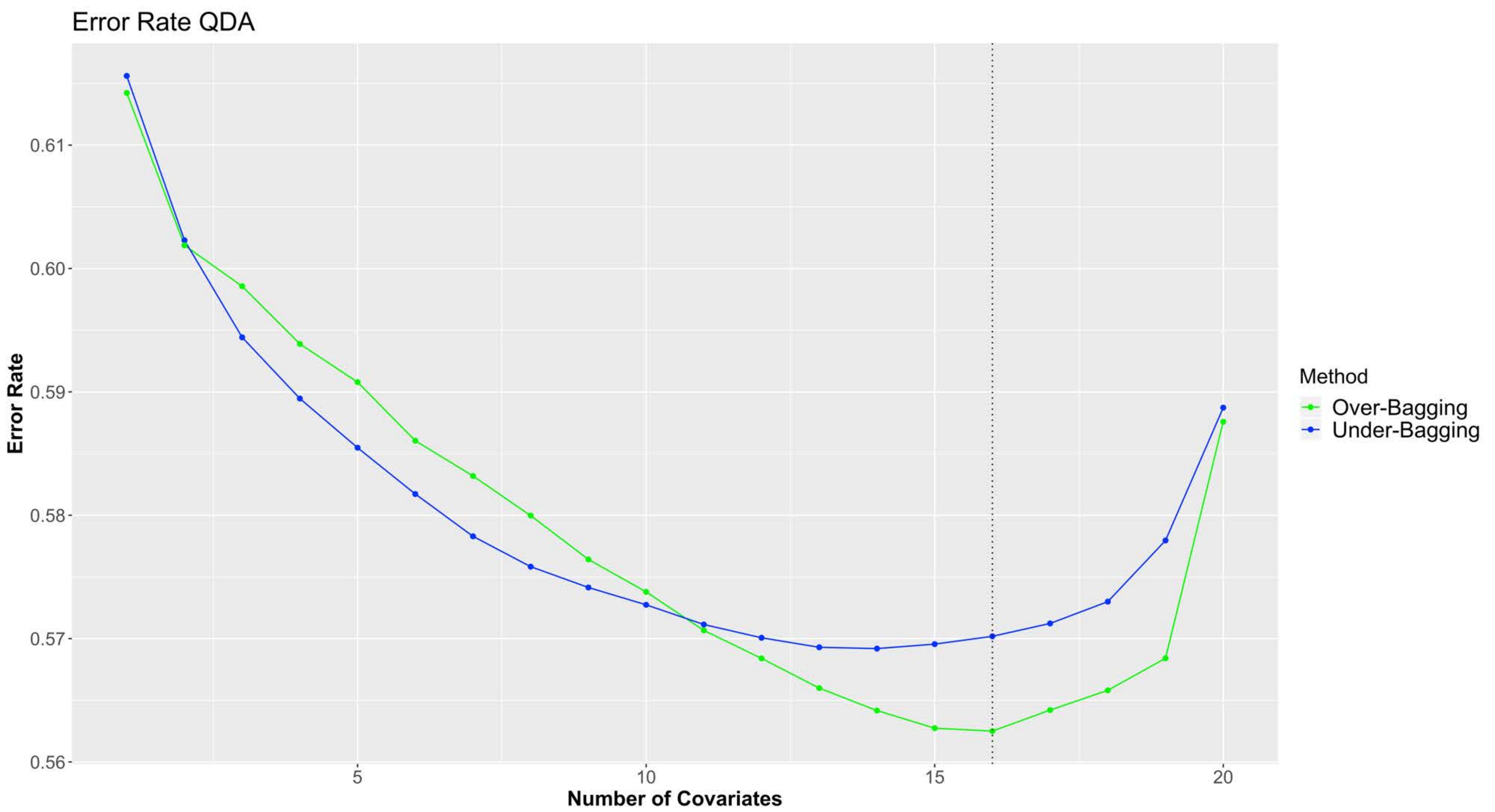
- Number of estimated models:
  - $B \times K \times \left(\frac{p(p+1)}{2} - 1\right) \approx 200'000$ with $B = 100, K = 10, p = 20$

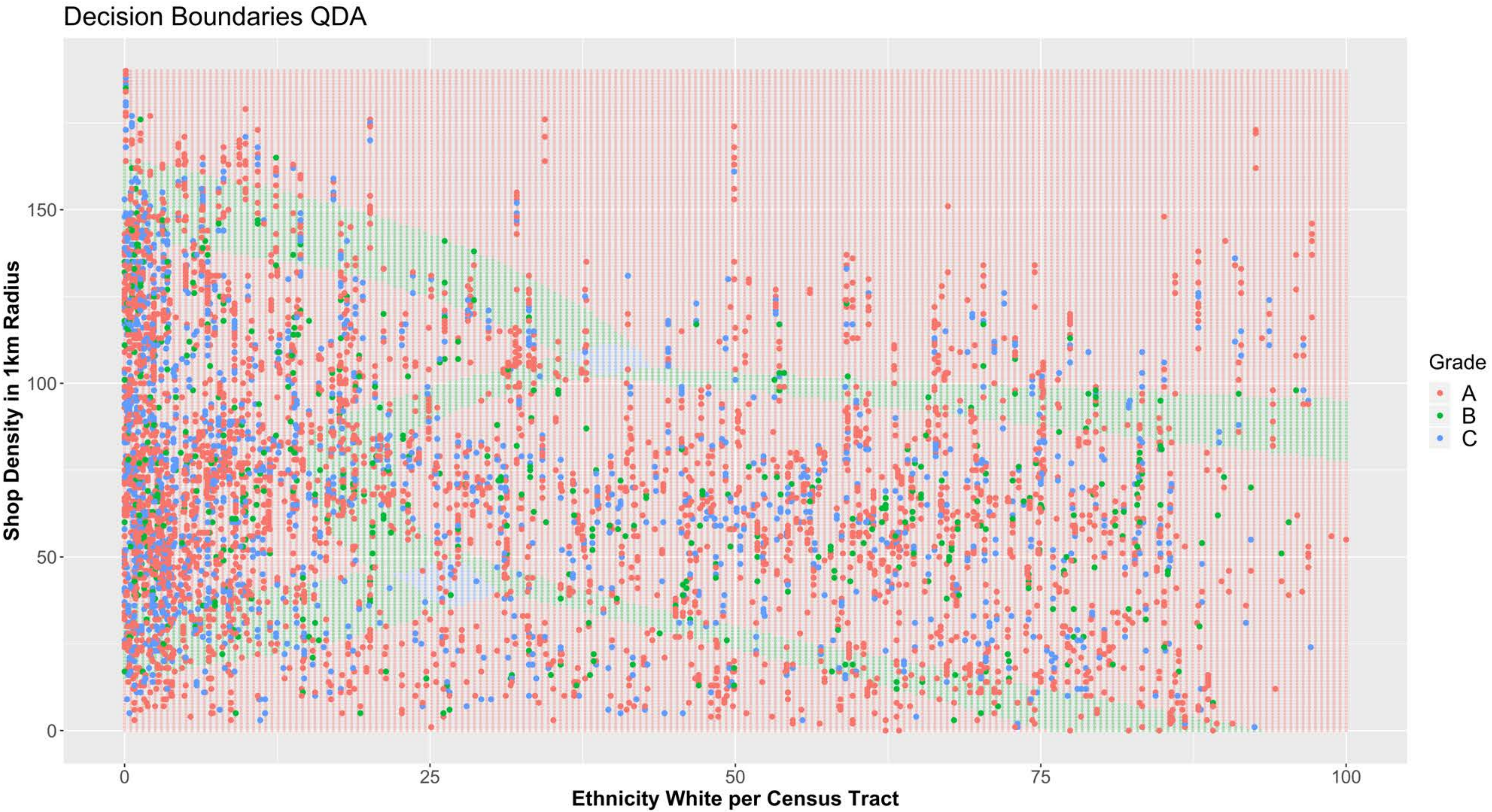
- **1)** Under- and Over-bagging **2)** 100 Bagged Samples **3)** 10-Fold-CV **4)** Best Subset Selection for only 20 Covariates

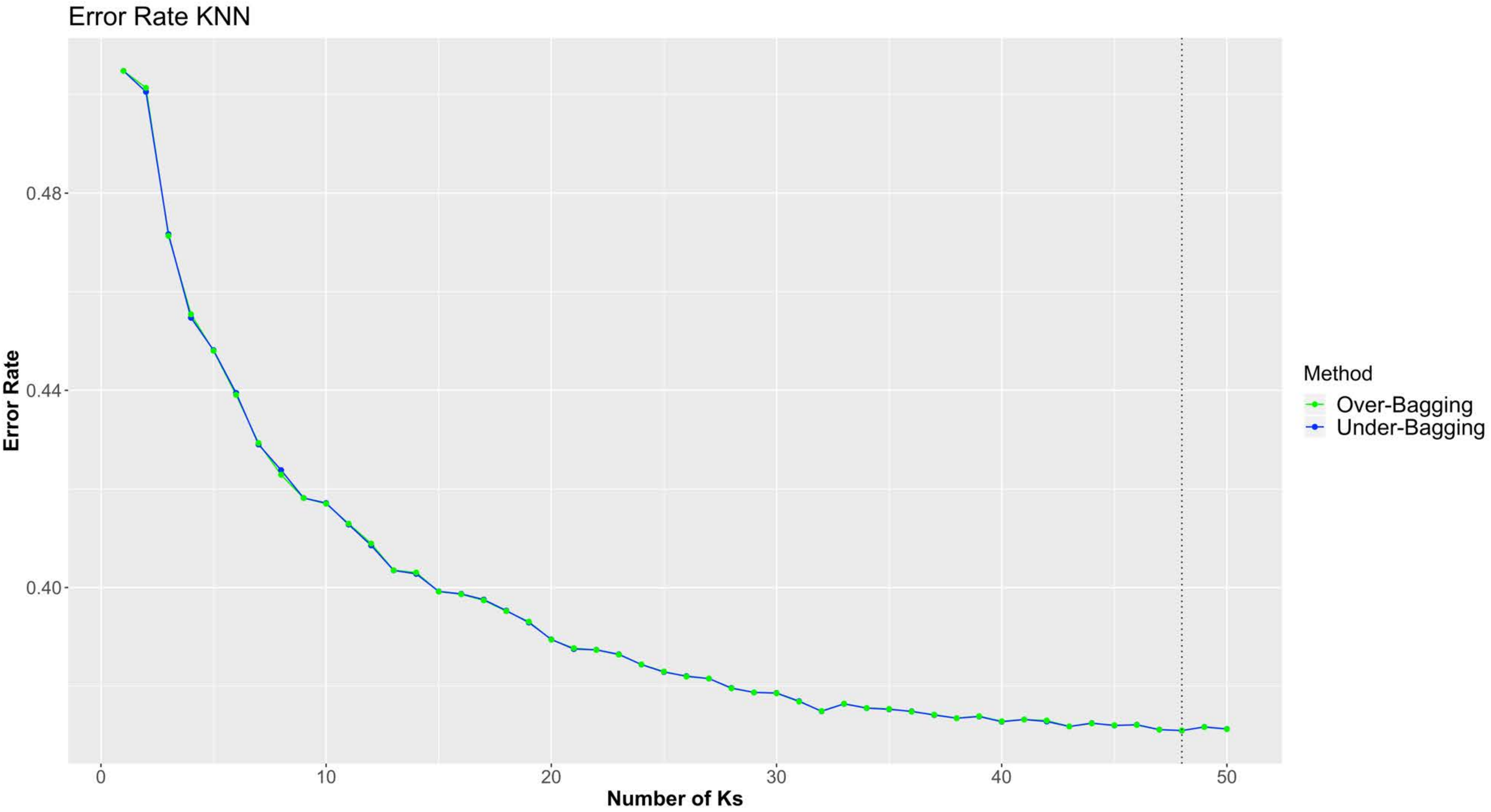
[1] According to Hastie, Tibshirani & Friedman (2013, p. 245)

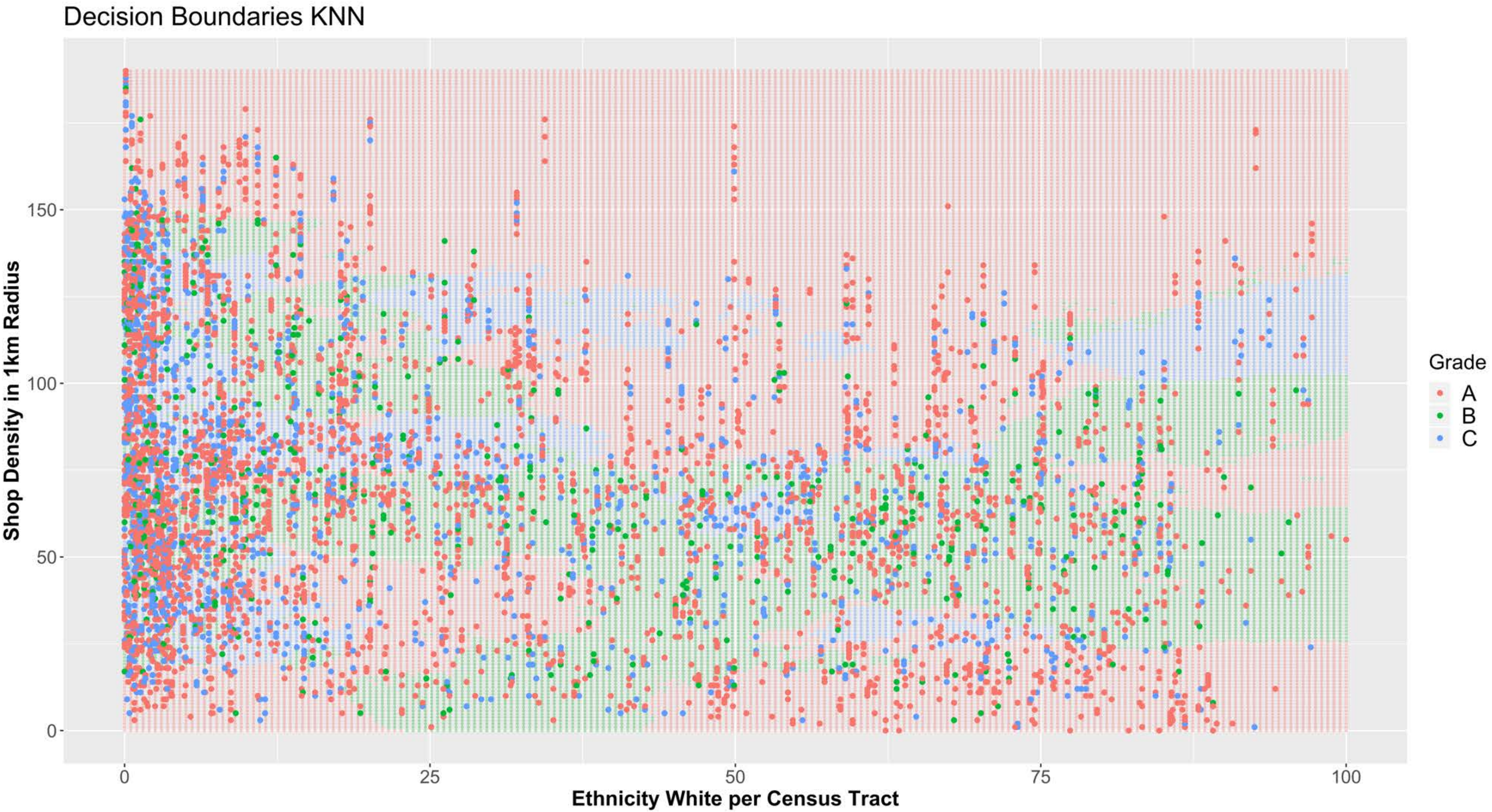

Error Rate LDA
Error Rate
0.62
0.61
0.60
0.59
0.58
Number of Covariates
5
10
15
20
Method
Over-Bagging
Under-Bagging


Decision Boundaries LDA
Shop Density in 1km Radius
150
100
50
0
0
25
50
75
100
Ethnicity White per Census Tract
Grade
A
B
C


Error Rate QDA
Error Rate
0.61
0.60
0.59
0.58
0.57
0.56
5
10
15
20
Number of Covariates
Method
Over-Bagging
Under-Bagging


Decision Boundaries QDA
Shop Density in 1km Radius
150
100
50
0
0
25
50
75
100
Ethnicity White per Census Tract
Grade
A
B
C


Error Rate KNN
0.48
0.44
0.40
Error Rate
0
10
20
30
40
50
Number of Ks
Method
Over-Bagging
Under-Bagging


Decision Boundaries KNN
Shop Density in 1km Radius
150
100
50
0
0
25
50
75
100
Ethnicity White per Census Tract
Grade
A
B
C

Prediction Trade-Off

QDA

Error Rate: 54%

Obs / Pred	A	B	C
A	2839	462	1164
B	1260	173	594
C	228	26	103

KNN

Error Rate: 66%

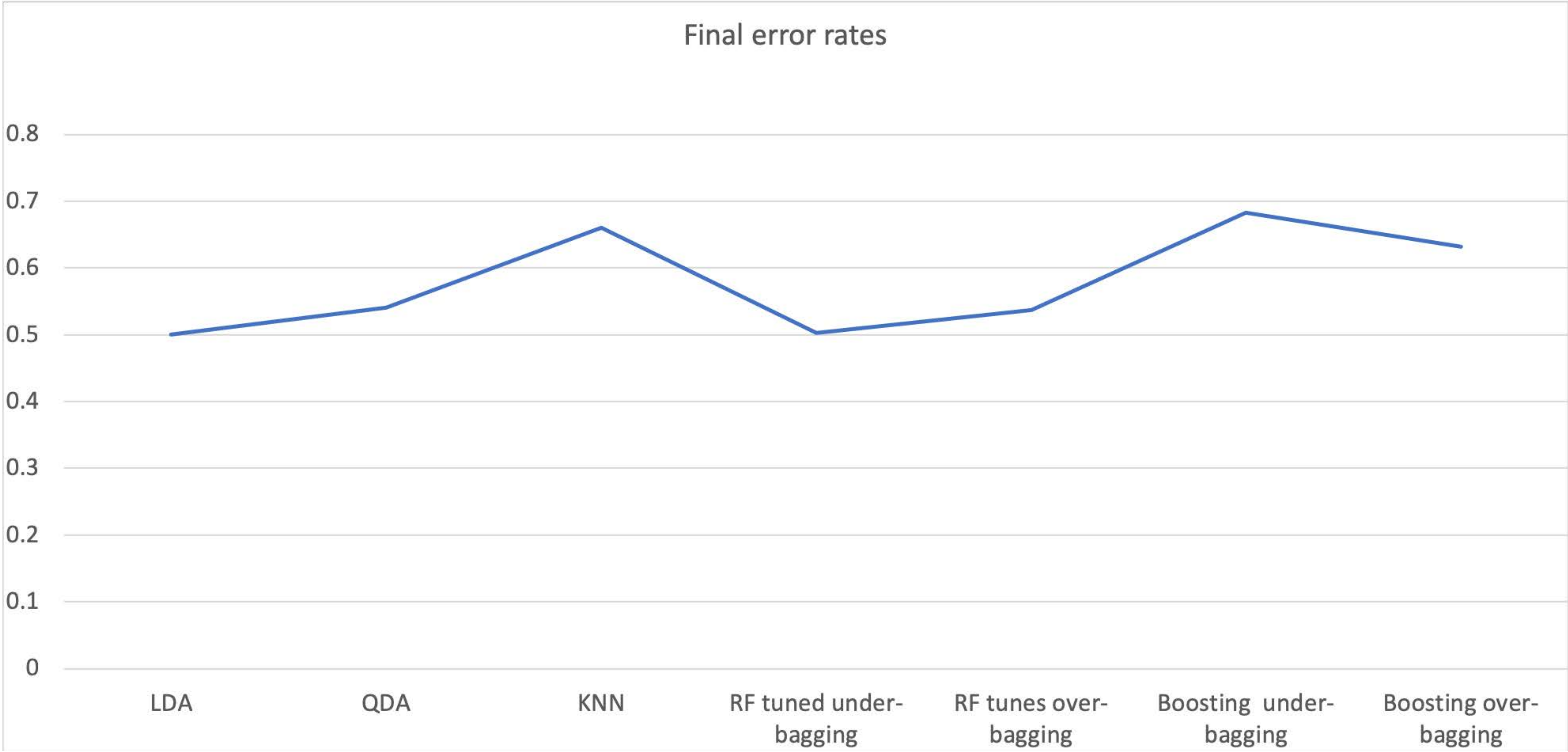
Obs / Pred	A	B	C
A	1464	204	470
B	1928	335	885
C	935	122	506

Random Forest and Boosting

- Methodology needs very high computational power
  - Fewer iterations lead to higher variance
- Tendence to over-fitting with over-bagging
  - Training error of 5%

RF over-bagging	RF under-bagging	Boosting under-bagging	Boosting over-bagging
0.533	0.51	0.68	0.63

Results


Final error rates
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
LDA
QDA
KNN
RF tuned under-bagging
RF tunes over-bagging
Boosting under-bagging
Boosting over-bagging

Conclusion

- Correlation of best 20 covariates: between 0.09 – to 0.03
- More sophisticated approach to imbalance problem (e.g. Synthetic Resampling Technique)
- Opportunities for agency:
  - Use of internal data
  - Unique data from every food store

Appendix 1: OOB Testing Sample

- Probability for a not picking observation

$$\frac{N\ -1}{N}$$

- Probability for a not picking N observations (with replacement)

$$\left(\frac{N\ -1}{N}\right)^{N}$$

- Probability for a not picking N observations (with replacement)

$$\lim_{N\to\infty}\left(\frac{N\ -1}{N}\right)^{N}=e^{-1}=0.368$$