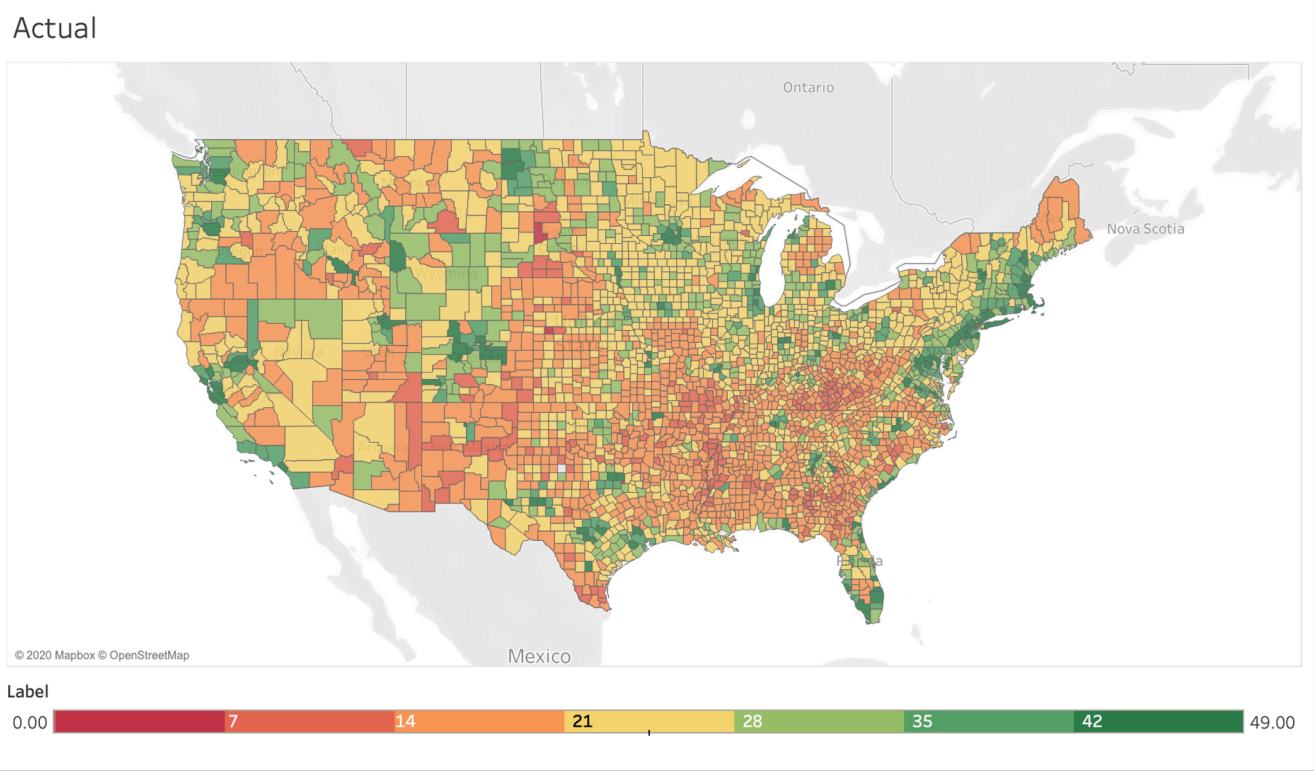
Big Data Analytic Applications Symposium Spring 2020 County AGI Prediction



Michael Urciuoli
Computer Science, NYU GSAS
mlu216@nyu.edu

Carl Barbee
Computer Science, NYU GSAS
crb616@nyu.edu

Alex Spence
Data Science, NYU GSAS
ajs811@nyu.edu

County AGI Prediction
Agenda

- Abstract
- Motivation
- Data Sources
- Data Samples
- Design Diagram
- Code Challenges
- Insights
- Obstacles
- Future Works
- Acknowledgements
- References
- Demo
- Questions

County AGI Prediction
Abstract

- Goal: develop linear regression model (Spark ML) on NYU DUMBO
  - Input
    - Counts of businesses by NAICS type and size
    - Demographics (gender, race)
    - Educational Attainment
  - Output
    - Average AGI (total AGI / total population)
- 4 Datasets Used
  - U.S. Census County Business Patterns (CBP)
  - U.S. Census American Community Survey (ACS)
    - Demographics & Educational Attainment
  - U.S. IRS SOI Tax Stats - County Data
- Results
  - Model fitted on years 2011 to 2016, evaluated on 2017
  - 91% of predications fall within 30% of the target variable
  - RMSE of 5.5, R^2 value of 0.63
  - Predictions and coefficients displayed on Tableau for user interaction

County AGI Prediction
Motivation

- Users
  - Home buyers, real estate investors
  - Social workers, government policy makers
- Benefit
  - AGI data released by the IRS lags by a few years
  - Being able to predict current AGI enables decision making
- Model Coefficients
  - Understanding what features contribute to AGI can provide insight on how to improve the economics of an area
  - It also allows investors to understand features to look out for when looking for emerging areas

County AGI Prediction
Data Sources

- U.S. Census County Business Patterns (CBP)
  - Counts of businesses organized by NAICS code and number of employees
  - Size: GBs
- IRS SOI Tax Stats
  - Income data organized by US county
  - Size: GBs
- U.S. Census American Community Survey (ACS) Demographics
  - Provides Demographic data organized by US county
  - Size: MBs
- U.S. Census ACS Educational Attainment
  - Level of education attained seperated by age group and US county
  - Size: MBs

County AGI Prediction
Data Sample: NAICS codes

Code	Industry Title	Number of Business Establishments
11	Agriculture, Forestry, Fishing and H	381,477
21	Mining	32,069
22	Utilities	46,245
23	Construction	1,490,099
31-33	Manufacturing	637,810
42	Wholesale Trade	697,359
44-45	Retail Trade	1,794,062
48-49	Transportation and Warehousing	588,529
51	Information	358,572
52	Finance and Insurance	782,705
53	Real Estate Rental and Leasing	868,526
54	Professional, Scientific, and Technical Services	2,294,049
55	Management of Companies and Enterprises	70,791
56	Administrative and Support and Waste Manage Remediation Services	1,647,950
61	Educational Services	424,190
62	Health Care and Social Assistance	1,745,915
71	Arts, Entertainment, and Recreation	369,206
72	Accommodation and Food Services	907,516
81	Other Services (except Public Administration)	1,915,436
92 *	Public Administration	258,094

County AGI Prediction

Data Sample: Census County Business Patterns (CBP)

fipstate	fipscty	year	totals	ests	fips_combined
01	037	2015	[92, 6, 0, 2, 8, ...	[[62, 13, 9, 5, 1...	01037
01	057	2015	[306, 15, 5, 2, 1...	[[168, 69, 45, 17...	01057
02	130	2017	[608, 20, 0, 0, 6...	[[393, 121, 55, 2...	02130
02	180	2016	[174, 0, 4, 1, 12...	[[81, 45, 34, 10,...	02180
04	001	2012	[475, 7, 8, 6, 34...	[[232, 110, 72, 4...	04001
05	041	2013	[312, 10, 0, 7, 1...	[[174, 60, 42, 25...	05041
05	067	2012	[331, 8, 1, 4, 19...	[[181, 74, 35, 29...	05067
05	087	2011	[184, 4, 0, 2, 20...	[[106, 30, 27, 14...	05087
05	087	2012	[186, 3, 0, 2, 19...	[[108, 31, 26, 14...	05087
05	109	2014	[200, 12, 1, 3, 1...	[[106, 46, 27, 14...	05109
08	011	2012	[58, 1, 0, 3, 3, ...	[[35, 12, 8, 2, 0...	08011
08	025	2012	[36, 0, 0, 3, 1, ...	[[20, 6, 7, 1, 1,...	08025
08	029	2013	[818, 7, 9, 11, 9...	[[522, 167, 66, 4...	08029
08	073	2013	[123, 4, 2, 4, 9,...	[[64, 22, 25, 10,...	08073
10	001	2012	[3119, 12, 5, 7, ...	[[1615, 639, 424,...	10001
12	027	2015	[450, 12, 1, 2, 5...	[[260, 94, 55, 28...	12027
12	071	2015	[17459, 20, 8, 30...	[[10583, 2853, 18...	12071
12	085	2015	[5336, 18, 4, 12,...	[[3376, 872, 520,...	12085
12	095	2013	[33317, 19, 7, 21...	[[18919, 5579, 38...	12095
13	011	2012	[256, 5, 0, 0, 30...	[[139, 46, 41, 22...	13011

County AGI Prediction
Data Sample: Census County Business Patterns (CBP)

```
scala> cbp.first.getString(6)
res26: String = 01037

scala> cbp.first.getInt(2)
res27: Int = 2015

scala> cbp.first.getAs[Seq[Int]](4)
res28: Seq[Int] = WrappedArray(92, 6, 0, 2, 8, 6, 5, 18, 5, 1, 4, 5, 5, 1, 2, 0, 4, 1, 2, 17, 0)

scala> cbp.first.getAs[Seq[Seq[Int]]](5).foreach(println)
WrappedArray(62, 13, 9, 5, 1, 1, 1, 0, 0, 0, 0, 0)
WrappedArray(1, 3, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(7, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(1, 1, 1, 2, 0, 0, 1, 0, 0, 0, 0, 0)
WrappedArray(2, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(9, 6, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(4, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(2, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)
WrappedArray(15, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
WrappedArray(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
```

County AGI Prediction
Data Sample: SOI Tax Stats

fipstate	STATE	fipscty	COUNTYNAME	agi_stub	agi	fipstate_fipscty	year
01	AL	000	Alabama	1	-1863084.000	01000	2011
01	AL	000	Alabama	2	12086882.000	01000	2011
01	AL	000	Alabama	3	17754372.000	01000	2011
01	AL	000	Alabama	4	15646800.000	01000	2011
01	AL	000	Alabama	5	13532193.000	01000	2011
01	AL	000	Alabama	6	22916889.000	01000	2011
01	AL	000	Alabama	7	20234205.000	01000	2011
01	AL	001	Autauga County	1	-14016.0000	01001	2011
01	AL	001	Autauga County	2	124407.0000	01001	2011
01	AL	001	Autauga County	3	201382.0000	01001	2011
01	AL	001	Autauga County	4	209986.0000	01001	2011
01	AL	001	Autauga County	5	216637.0000	01001	2011
01	AL	001	Autauga County	6	316812.0000	01001	2011
01	AL	001	Autauga County	7	94449.0000	01001	2011
01	AL	003	Baldwin County	1	-149518.0000	01003	2011
01	AL	003	Baldwin County	2	444375.0000	01003	2011
01	AL	003	Baldwin County	3	716747.0000	01003	2011
01	AL	003	Baldwin County	4	710830.0000	01003	2011
01	AL	003	Baldwin County	5	645072.0000	01003	2011
01	AL	003	Baldwin County	6	1175640.0000	01003	2011

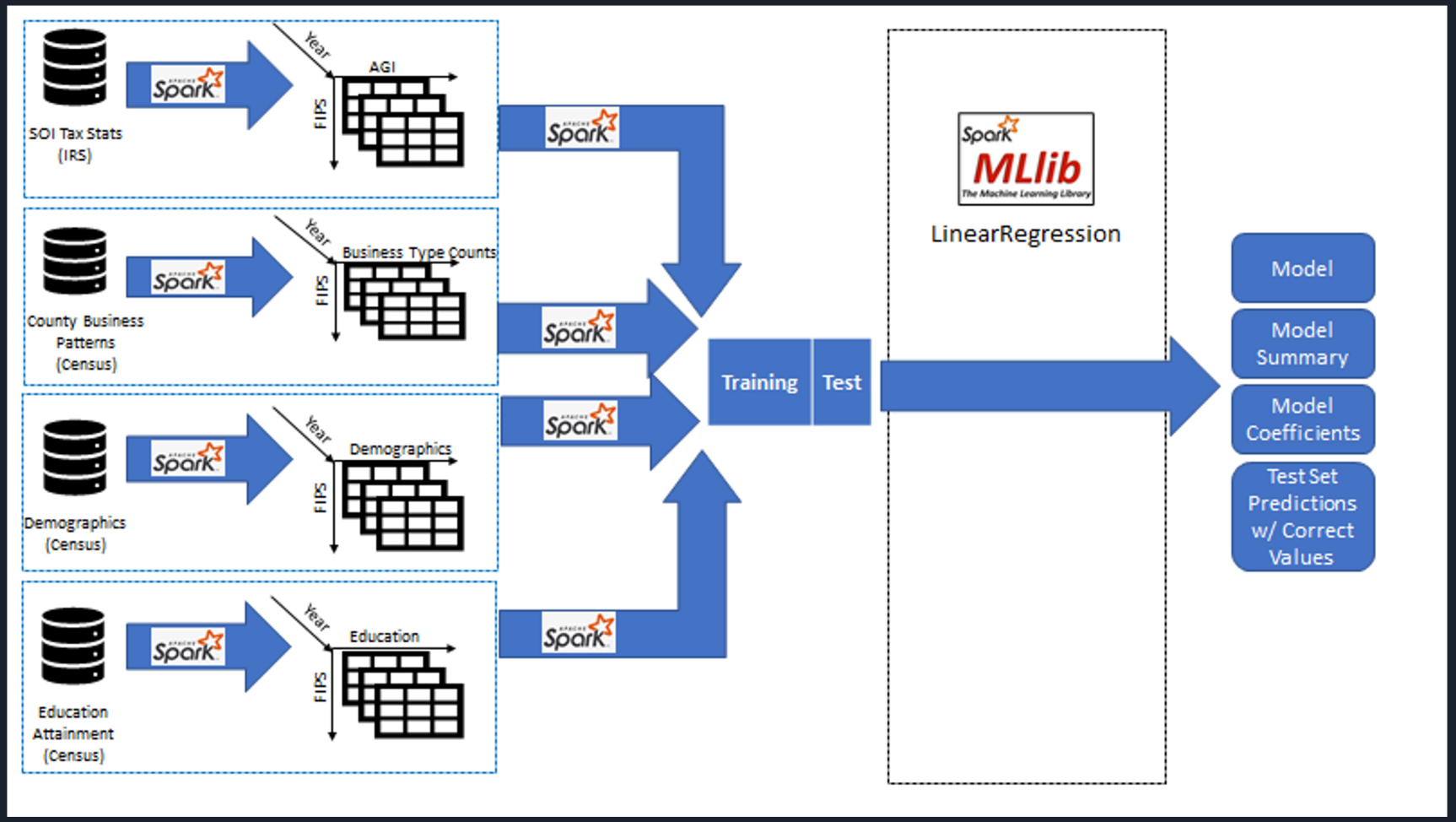
County AGI Prediction
Data Sample: US Census Demographic Data

county_id	county	state	total_population	men	women	hispanic	white	black	native	year
13155	Irwin County	Georgia	9642	4910	4732	1.8	70.5	26.7	0.1	2010
13157	Jackson County	Georgia	58347	29161	29186	1.3	83.9	6.9	0.1	2010
13159	Jasper County	Georgia	13695	6845	6850	0.0	72.4	22.5	0.8	2010
13161	Jeff Davis County	Georgia	14558	7150	7408	0.0	74.4	15.7	0.2	2010
13163	Jefferson County	Georgia	16919	8026	8893	0.2	41.3	55.8	0.0	2010
13165	Jenkins County	Georgia	8400	4125	4275	0.4	57.4	36.0	0.0	2010
13167	Johnson County	Georgia	9967	5324	4643	0.2	61.8	36.4	0.1	2010
13169	Jones County	Georgia	28292	13440	14852	0.6	72.6	25.0	0.0	2010
13171	Lamar County	Georgia	17837	8688	9149	0.7	65.6	30.9	0.2	2010
13173	Lanier County	Georgia	9404	4593	4811	0.3	68.7	25.2	0.0	2010
13175	Laurens County	Georgia	48034	22873	25161	0.1	60.2	35.6	0.2	2010
13177	Lee County	Georgia	27998	14193	13805	0.6	76.6	17.4	0.3	2010
13179	Liberty County	Georgia	63854	31568	32286	1.9	43.2	41.6	0.4	2010
13181	Lincoln County	Georgia	8111	4022	4089	0.2	65.2	31.8	0.3	2010
13183	Long County	Georgia	13142	6498	6644	0.6	60.1	26.4	0.3	2010
13185	Lowndes County	Georgia	104916	51230	53686	0.9	56.9	35.2	0.2	2010
13187	Lumpkin County	Georgia	28875	14449	14426	2.1	91.7	1.3	0.8	2010
13189	McDuffie County	Georgia	21719	10297	11422	0.2	56.9	39.9	0.4	2010
13191	McIntosh County	Georgia	13817	6705	7112	1.2	60.6	36.6	0.0	2010
13193	Macon County	Georgia	14553	7462	7091	0.0	32.8	61.3	0.0	2010

County AGI Prediction
Data Sample: US Census Educational Attainment

fips_combined	less_than_highschool	highschool	some_college	bachelors	year
5105	1476	3526	2322	739	2012
5107	4407	4598	4716	1848	2012
5109	1904	3075	2537	962	2012
5111	4961	7852	4199	1531	2012
5113	2671	6008	5323	1672	2012
5115	8250	16266	14828	8091	2012
5117	1584	2994	1569	640	2012
5119	33675	81041	91985	84455	2012
5121	2673	5513	4229	1390	2012
5123	5185	7906	6167	2147	2012
5125	9966	29163	24848	17565	2012
5127	2046	3216	2331	760	2012
5129	1631	2491	1690	651	2012
5131	16889	29992	31336	15798	2012
5133	4057	3886	3111	952	2012
5135	2251	6115	3769	1436	2012
5137	1880	3936	2938	1173	2012
5139	5820	11310	9843	4637	2012
5141	2663	5226	4059	1724	2012
5143	25308	43781	46200	37130	2012

County AGI Prediction
Design Diagram



County AGI Prediction
Code Challenge 1

- Each row in the original CBP dataset was a single count of a specific NAICS category
  - Only 2-digit summation categories where needed in the scope of this project
  - Needed to aggregate this data with each subcategory appearing at a specific index in a 2-D array

fipstate	fipscty	naics	est
01	001	------	835
01	001	11----	6
01	001	113///	5
01	001	1133//	5
01	001	11331/	5
01	001	113310	5
01	001	115///	1
01	001	1151//	1
01	001	11511/	1
01	001	115112	1
01	001	21----	2
01	001	212///	2
01	001	2123//	2
01	001	21231/	1
01	001	212319	1
01	001	21232/	1
01	001	212321	1
01	001	22----	9
01	001	221///	9
01	001	2211//	8

County AGI Prediction
Code Challenge 1

- Solution:
  - Scala class that can be loaded into a Spark dataset
  - Merge function that reduces records grouped by key (county_id, year) into specific format
  - Kryo serialization to allow class to be stored by Spark

```
class BusinessProfile(
        val fipstate: String,
        val fipscty: String,
        val year: Int,
        val nPres: Array[Boolean],
        val totals : Array[Int],
        val ests : Array[Array[Int]]) {

        org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)

        def merge(bp : BusinessProfile) : BusinessProfile = {
                if (fipstate != bp.fipstate ||
                        fipscty != bp.fipscty ||
                        year != bp.year)
                        throw new Exception("BusinessProfile mismatch")

                for ((b,i) <- bp.nPres.zipWithIndex) {
                        if (b) {
                                nPres(i) = true
                                totals(i) = bp.totals(i)
                                for ((v,j) <- bp.ests(i).zipWithIndex) { ests(i)(j) = v }
                        }
                }
                this
        }
}
```

County AGI Prediction
Code Challenge 2

- Data Schema changes across years
  - 2017 and 2018 changed the column names and ordering making it difficult to generically index for all years.

Original Code

```
val cols = List("GEO_ID",
                "NAME",
                "DP05_0001E",
                "DP05_0002E",
                "DP05_0003E",
                "DP05_0071PE",
                "DP05_0077PE",
                "DP05_0078PE",
                "DP05_0079PE")

// Get the header
val header = sc.textFile(folder + file + ".csv").take(1).map(x => x.split(',')).toList

// Remove quotes around fields.
val header_trimmed = header(0).map(x => x.substring(1, x.length - 1))

// Get the column indices we need to keep.
var column_indices = header_trimmed.indices.filter(cols contains
header_trimmed(_))
```

Revised Code

```
// Select the columns to keep
if (file == 2017 || file == 2018 )
{
        df_year = df_year.select("GEO_ID",
                                 "NAME",
                                 "DP05_0001E",
                                 "DP05_0002E",
                                 "DP05_0003E",
                                 "DP05_0071PE",
                                 "DP05_0077PE",
                                 "DP05_0078PE",
                                 "DP05_0079PE")
}
else
{
        df_year = df_year.select("GEO_ID",
                                 "NAME",
                                 "DP05_0001E",
                                 "DP05_0002E",
                                 "DP05_0003E",
                                 "DP05_0070PE",
                                 "DP05_0072PE",
                                 "DP05_0073PE",
                                 "DP05_0074PE")
}
```

County AGI Prediction
Code Challenge 3

- Data Inconsistencies across years for IRS dataset
  - Different number of columns
  - Columns of interest in different index positions
  - 3 records in the 2014 file which did not conform to the same length as the others
- Solutions:
  - Map each year to separate RDDs, extract the information from the required columns, then append into one RDD.

```
val condrdd11 = newrdd11.map(record => ( record(4), record(10),
val condrdd12 = newrdd12.map(record => ( record(4), record(12),
val condrdd13 = newrdd13.map(record => ( record(4), record(12),
val condrdd14 = newrdd14.map(record => ( record(4), record(15),
val condrdd15 = newrdd15.map(record => ( record(4), record(19),
```

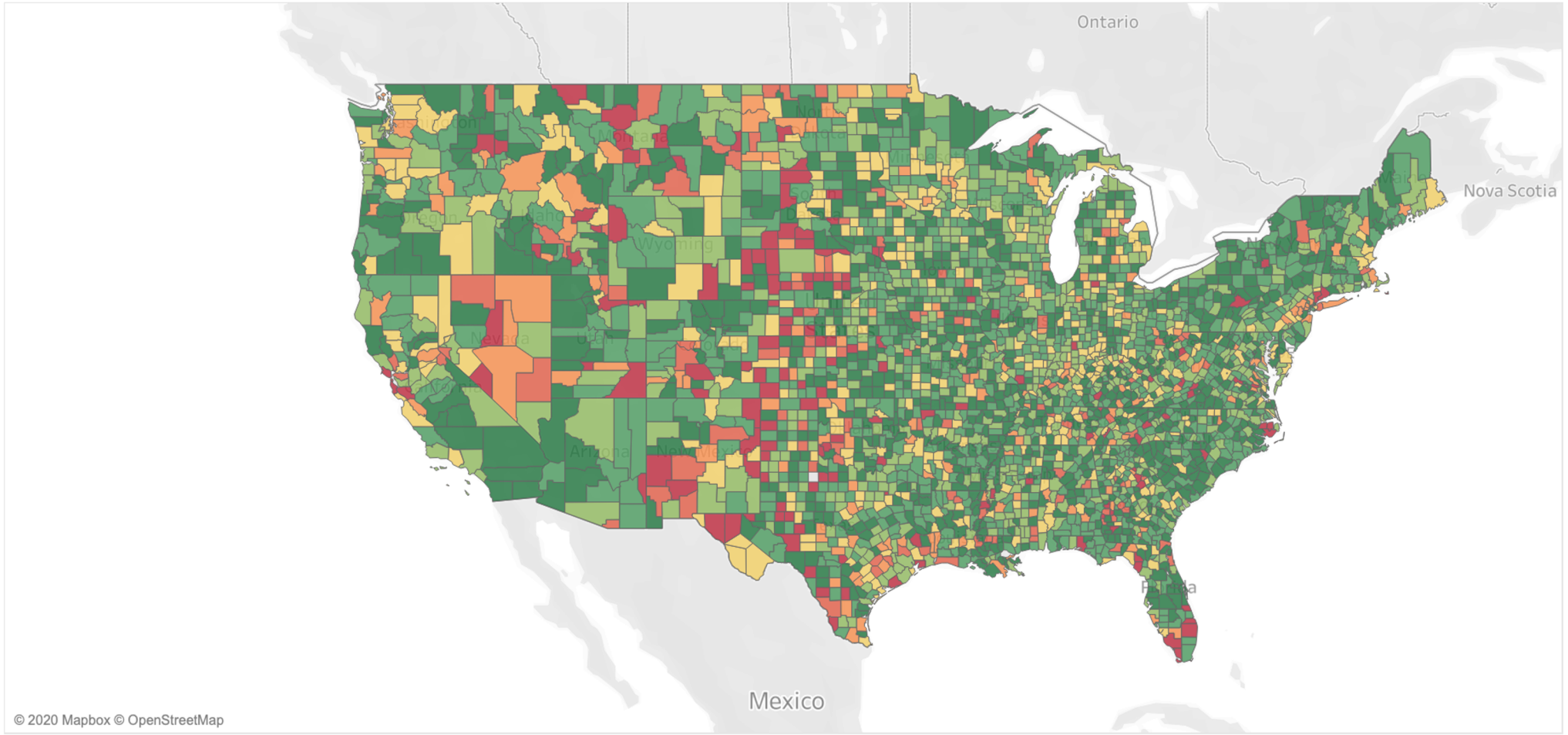
  - For 2014, exclude records (3) which are not the same length as the others

```
val newrdd14_t = myrdd14.map(line => line.split(","))
//for 14, there should be 128 fields
val newrdd14 = newrdd14_t.filter(line => line.length == 128)
```

County AGI Prediction

Insights

Percent Difference



perc diff

0.0000 | 0.06 | 0.12 | 0.18 | 0.24 | 0.30 | 0.36 | 0.4200

County AGI Prediction

Insights

Social injustice? A sign that additional features are needed for the model? Both?

NAICS 22: Utilities
All forms of electricity, natural gas, water, sewage, etc.

These seem to line up. Not having a HS degree (or equiv) limits access to high paying jobs. “Some college” could be students or college dropouts

coefficients	feature
-2.7019129135409363	% Black
-0.801886561453557	% Native
6.148428592341531	21_Mining:10-19
185.3086354582102	21_Mining:20-49
-2.1237666248572866	22_Utilities:<5
190099.48435783328	22_Utilities:2,500-4,999
10.756581218201775	23_Construction:5-9
16.301098647086523	23_Construction:10-19
87.56572967481164	23_Construction:20-49
204.29410641891408	23_Construction:50-99
33.95653576313875	23_Construction:100-249
15.601716820647848	42_Wholesale:<5
31.8687520519549	42_Wholesale:10-19
46.834611531541505	42_Wholesale:20-49
6.896446296467912	42_Wholesale:50-99
-25.151245370826587	44_Retail:<5
-42.87662740271732	44_Retail:5-9
-17.477182565709345	44_Retail:10-19
7.349664822693562	48_Transportation:<5
133.05719420647057	48_Transportation:20-49
54.63187493448873	48_Transportation:50-99
200.36645466850632	52_Finance:250-499
307.3193477121889	52_Finance:500-999
22.16678012149524	54_Professional:<5
766.2646475790059	55_Management:500-999
979.5074390810281	55_Management:2,500-4,999
12.40900658309515	56_Administrative:<5
54.784067444419996	56_Administrative:10-19
-15.101287114048908	62_Health_Care:<5
-6.9147255528900375	62_Health_Care:5-9
-45.82137025913027	62_Health_Care:20-49
-14.04631732959995	62_Health_Care:50-99
-12.61913104527576	Less than highschool
-12.127524161999403	Some college or associates
80.21082770880906	Bachelors or higher

County AGI Prediction
Obstacles

1. Data standardization between the different datasets
   - County FIPS represented differently across datasets

2. Identifying faulty records which prohibited RDD appendages (3 records excluded from IRS dataset from 2014)

3. Remediating the Census data which had differing formats throughout the years.

County AGI Prediction
Future Works

- Automatically retrain the model based on the most important features
  - A model with less features requires less data collection
- Adding a more robust, interactive front-end for selecting and analyzing features
- Collecting a more fine-grain dataset, e.g., city or zip code level
- Apply standard ML optimization techniques

County AGI Prediction
Summary

A linear regression model was developed using Spark ML to predict the average AGI of a county using counts of businesses by NAICS type, demographics, and educational attainment as features. The IRS lags with releasing their information, so this model could be useful to those who need an approximation now. We achieve 91% accuracy of predictions being within 30% of their target. Interactive AGI maps, and feature importance maps were developed into an application using Tableau to illustrate the model results. The coefficients that linear regression quantify the importance of each feature towards model accuracy, and could be used to quantify social patterns.

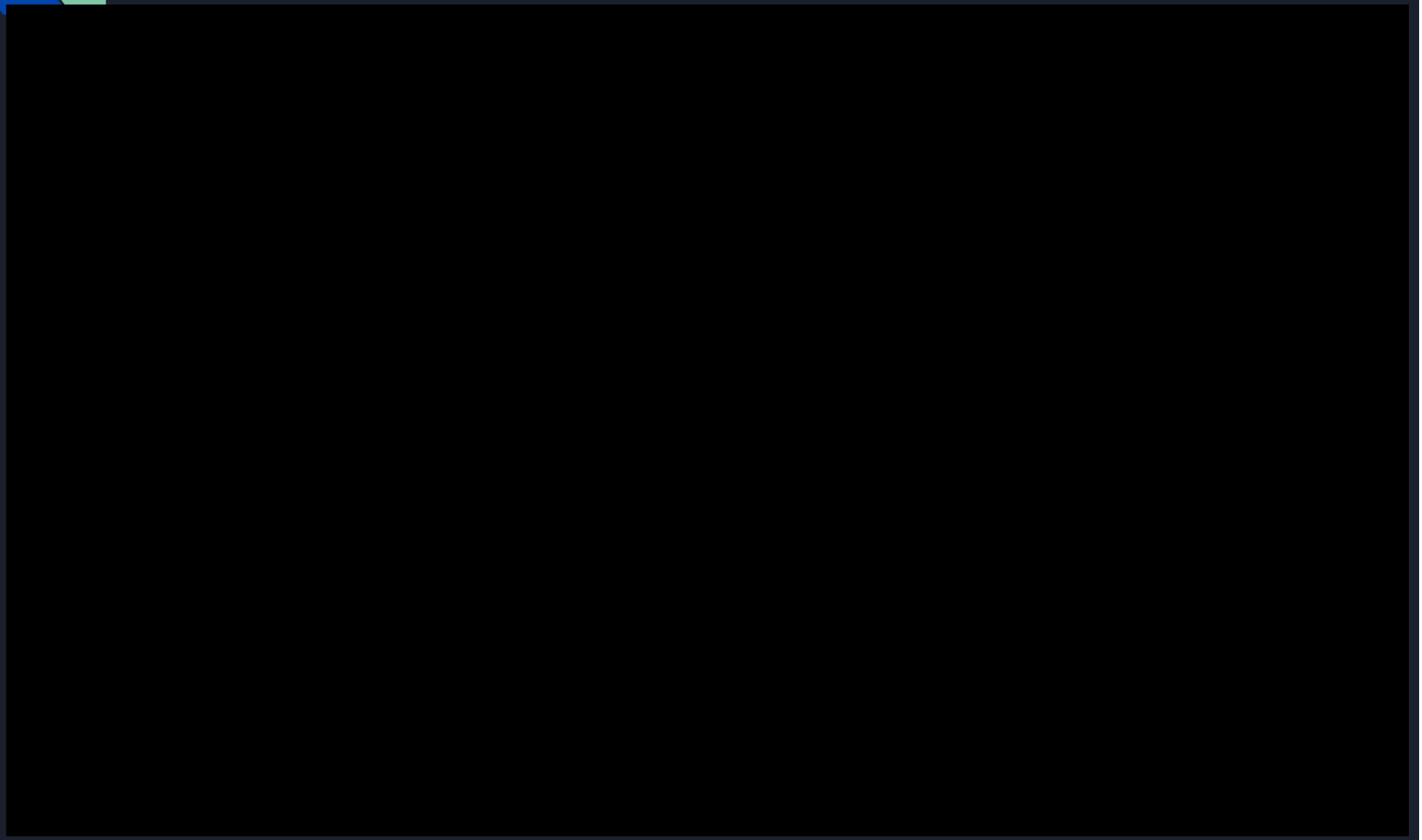
County AGI Prediction
Acknowledgements

- Thank you to Prof. McIntosh for answering our analysis questions.
- Thank you United States Census, IRS, for providing our datasets.
  - Thank you specifically to the Census clerk that answered the phone immediately and answered our questions without putting us on hold!
- Thank you to HPC for providing the technology to make this project possible.
- Thank you Naren Chittar of J.P. Morgan for checking the validity of our model
- Thank you Danielle Branton of Columbia University for insights that went into model formation and analysis

County AGI Prediction
References

- Predicting Neighborhoods' Socioeconomic Attributes Using Restaurant Data (Dong, Ratti, Zheng)
- The Changing Distribution of Wealth in the Pre-Crisis US and UK: the role of socio-economic factors (Cowell, Karagiannaki, McKnight)
- Income and Employment Multipliers for 20 Industries in 11 Census Divisions in Northern Ontario (Moazzami)
- A comparison of the approaches for gentrification identification (Liu, Deng, Song, Wu, Gong)
- Predicting individual-level income from Facebook profiles (Matz, Menges, Stillwell, Schwartz)
- Predicting Twitter User Socioeconomic Attributes with Network and Language Information (Aletras, Chamberlain)
- Spark Classification and Regression
- Understanding Urban Gentrification Through Machine Learning (Reades, De Souza, Hubbard)
- Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change (Glaeser, Kim, Luca)

County AGI Prediction
Demo!



County AGI Prediction

Questions?

Big Data Analytic Applications Symposium Spring 2020 County AGI Prediction

Thank you!

Michael Urciuoli
Computer Science NYU GSAS
mlu216@nyu.edu

Carl Barbee
Computer Science NYU GSAS
crb616@nyu.edu

Alex Spence
Data Science NYU GSAS
ajs811@nyu.edu