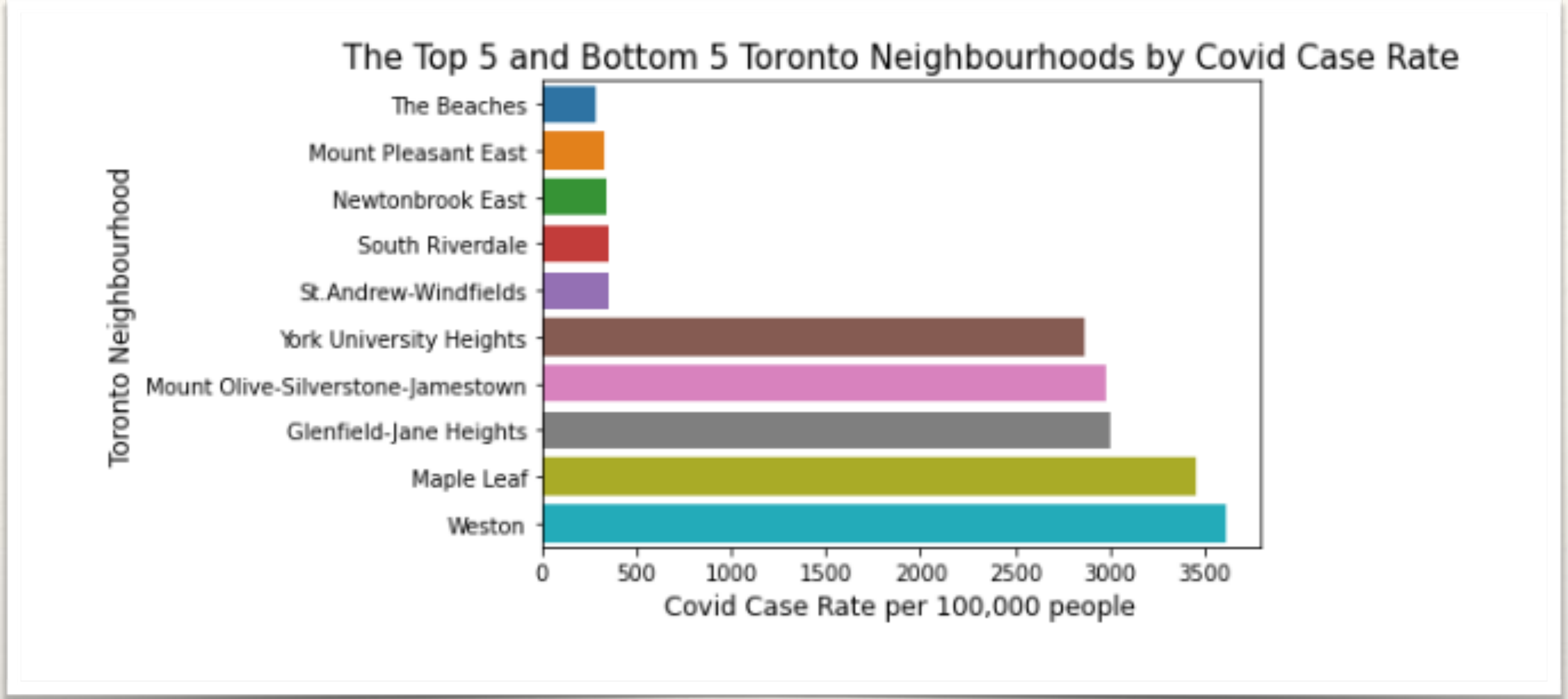
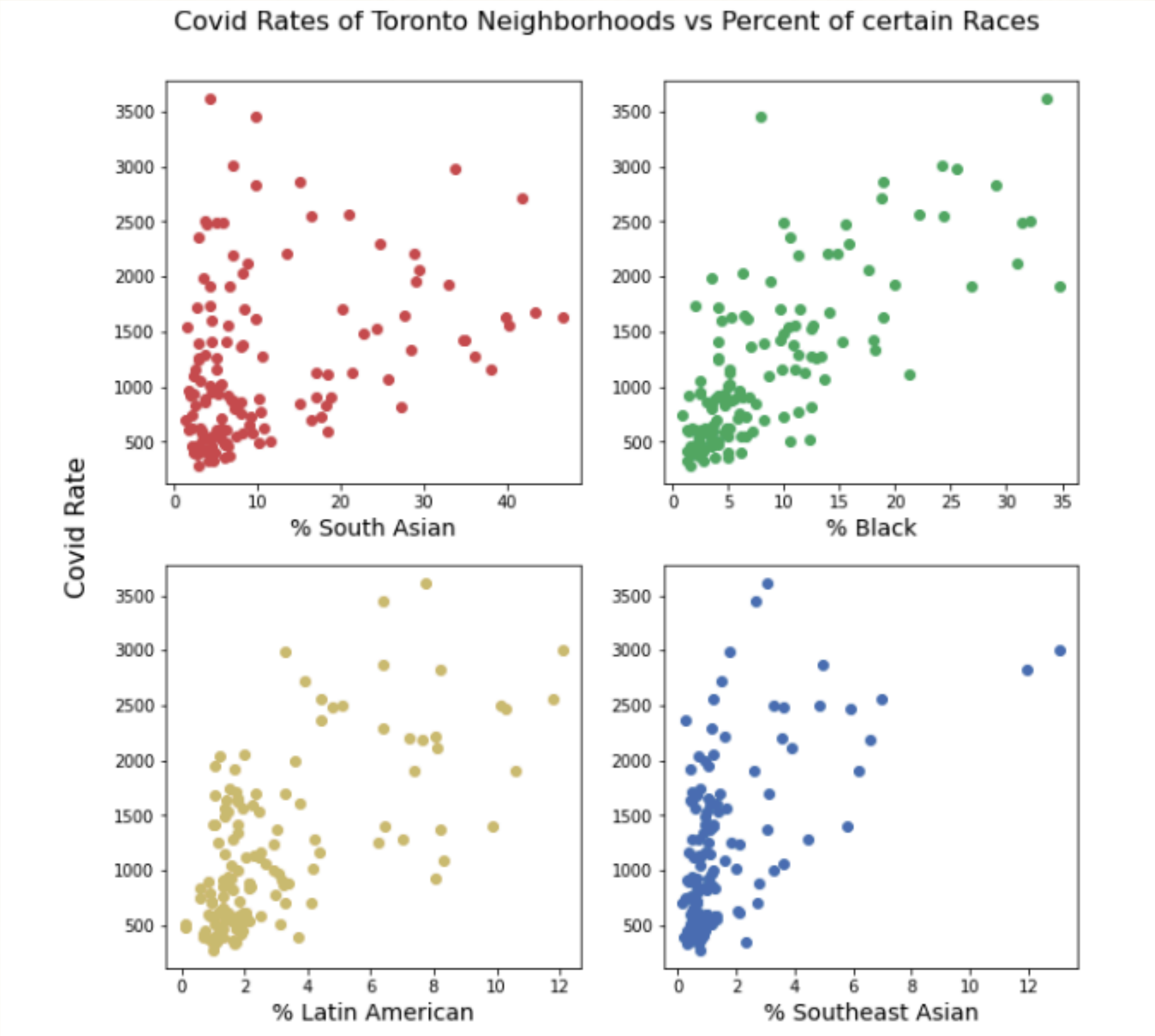
Predicting Covid rates in Toronto Neighborhoods using Linear Regression

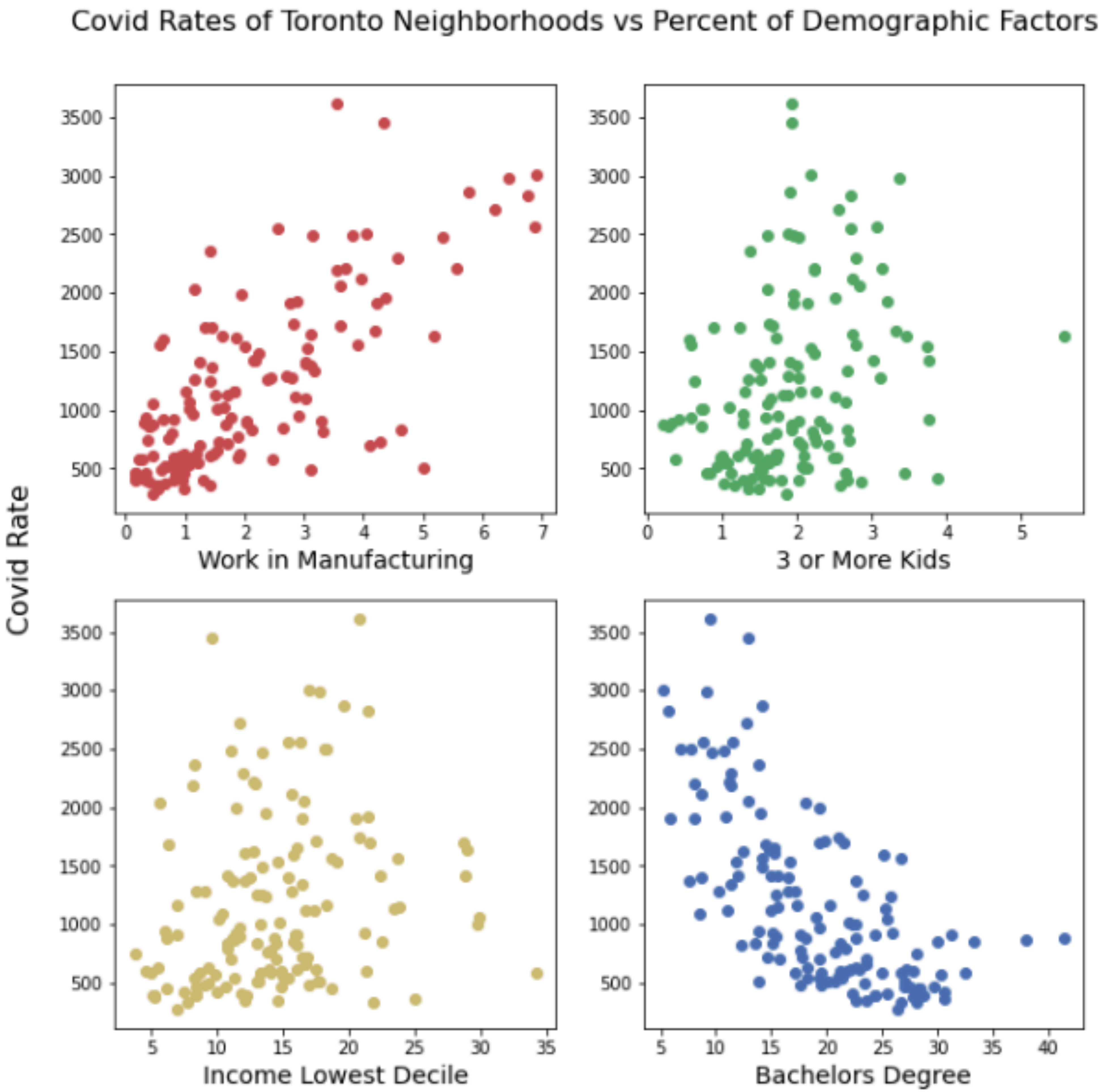
Alexei Marcilio
GBC

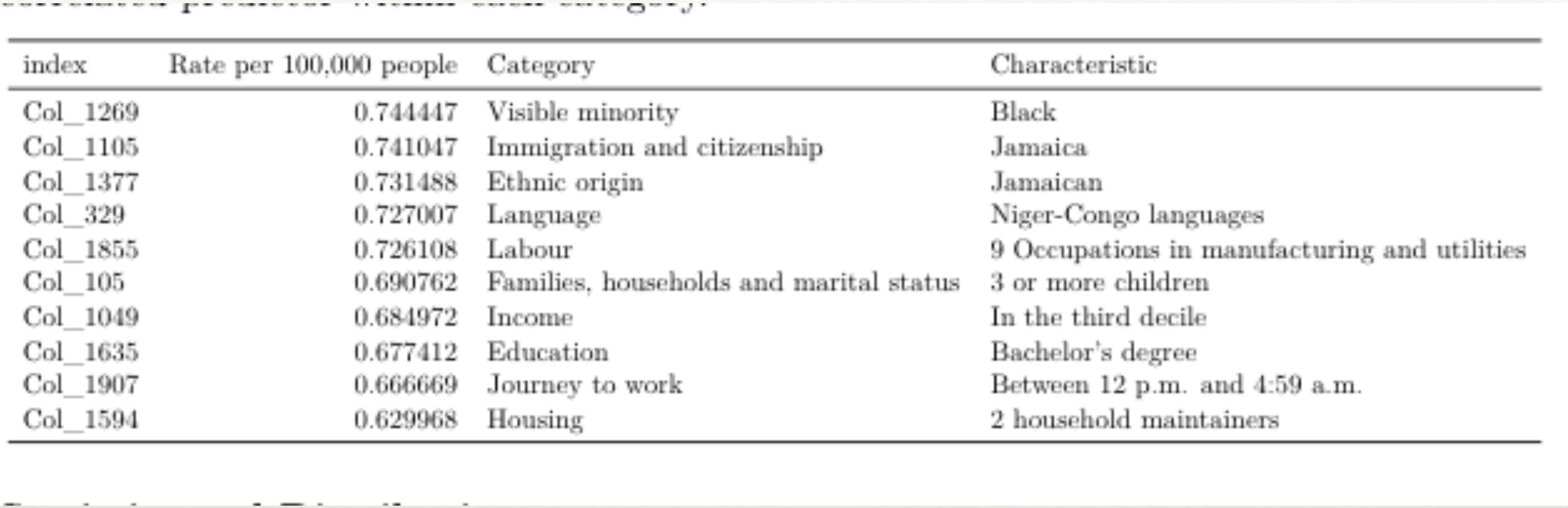
- Data from Toronto’s open data portal
- Two files combined, covid rates and Census Data
- Over 2,300 potential features


The Top 5 and Bottom 5 Toronto Neighbourhoods by Covid Case Rate
Toronto Neighbourhood
The Beaches
Mount Pleasant East
Newtonbrook East
South Riverdale
St.Andrew-Windfields
York University Heights
Mount Olive-Silverstone-Jamestown
Glenfield-Jane Heights
Maple Leaf
Weston
0
500
1000
1500
2000
2500
3000
3500
Covid Case Rate per 100,000 people

Rates Vary

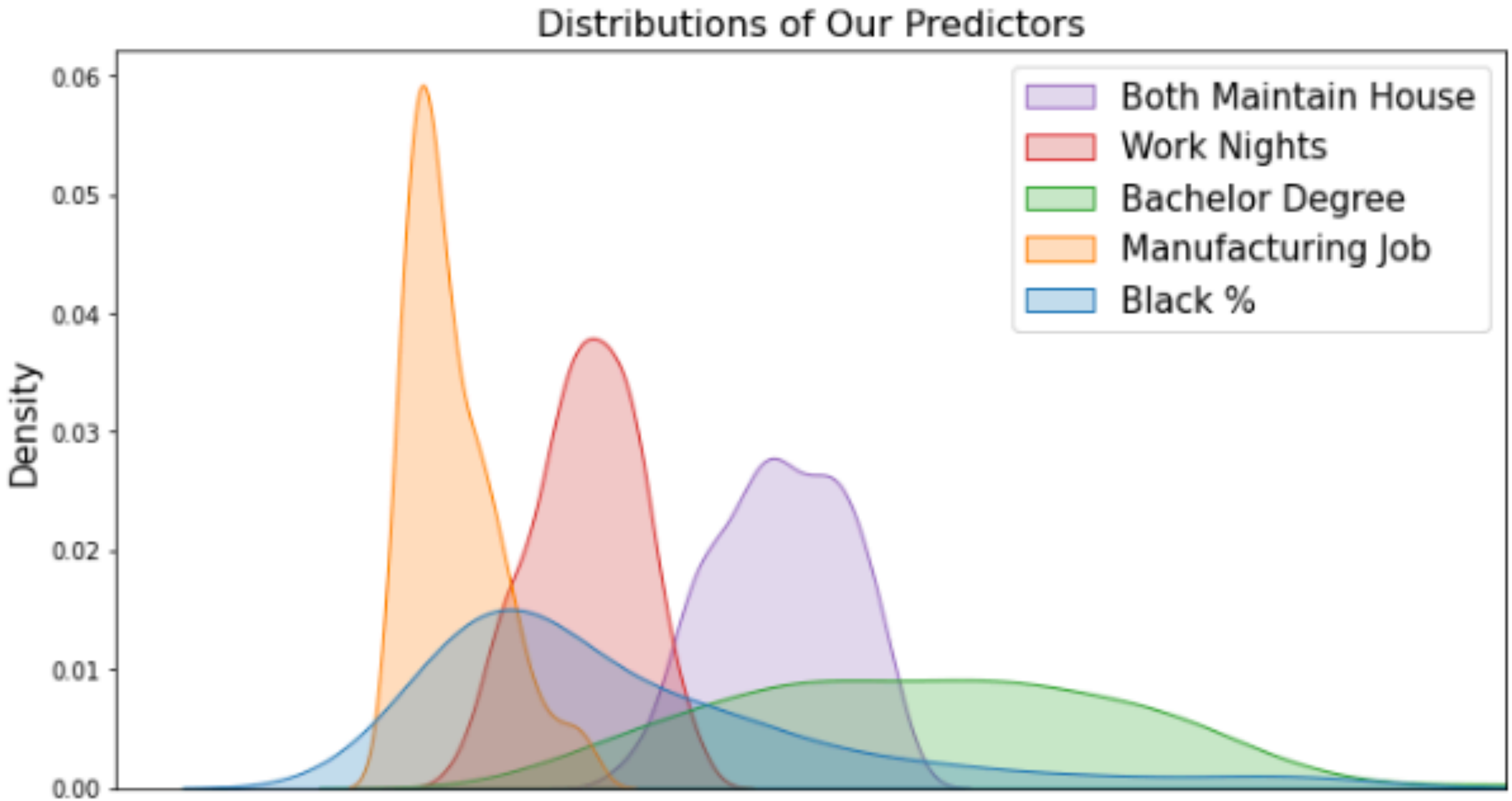

Covid Rates of Toronto Neighborhoods vs Percent of certain Races
Covid Rate
3500
3000
2500
2000
1500
1000
500
0
10
20
30
40
% South Asian
0
5
10
15
20
25
30
35
% Black
0
2
4
6
8
10
12
% Latin American
0
2
4
6
8
10
12
% Southeast Asian

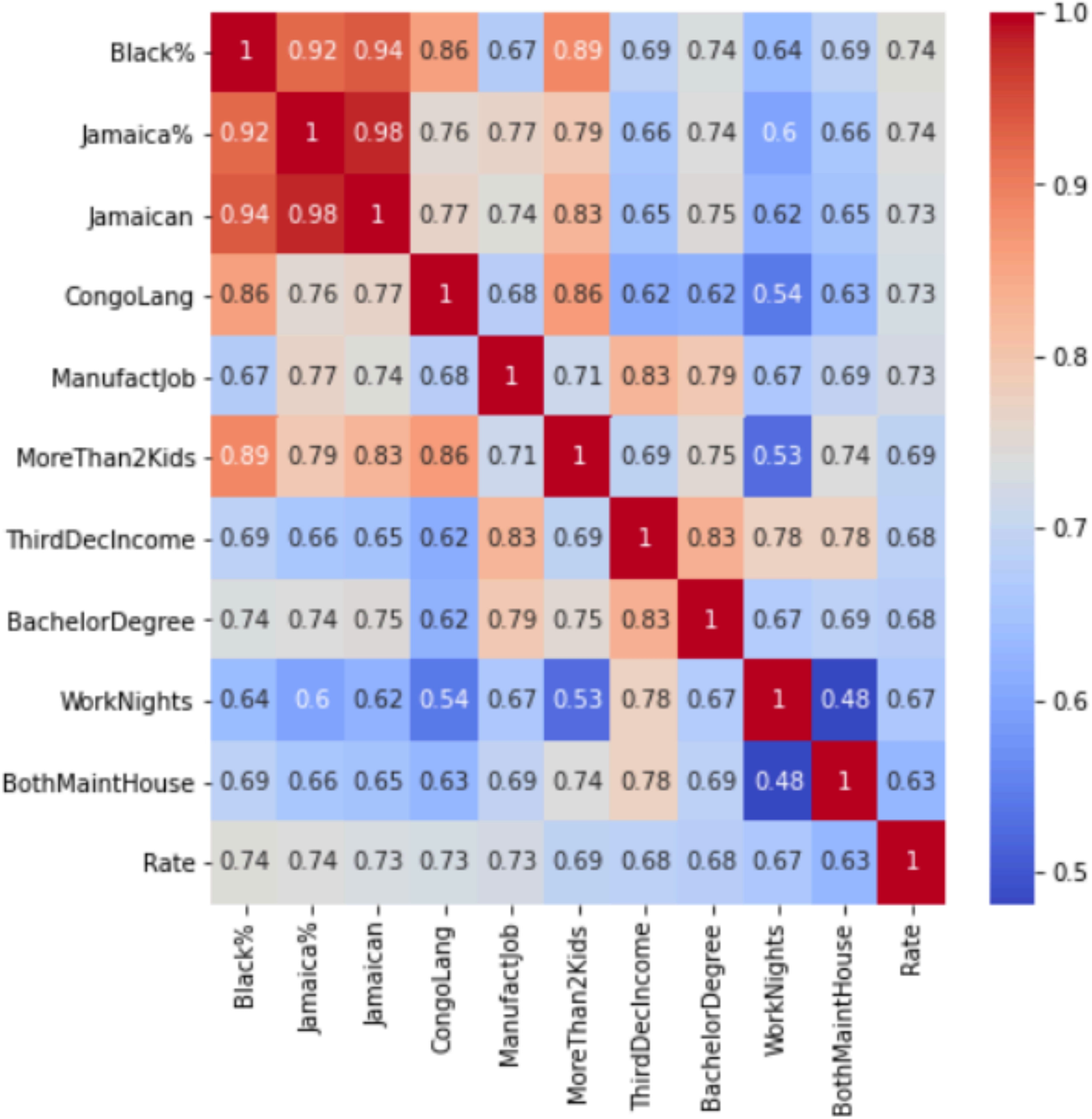

Covid Rates of Toronto Neighborhoods vs Percent of Demographic Factors
Covid Rate
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
Work in Manufacturing
0
1
2
3
4
5
3 or More Kids
5
10
15
20
25
30
35
Income Lowest Decile
5
10
15
20
25
30
35
40
Bachelors Degree



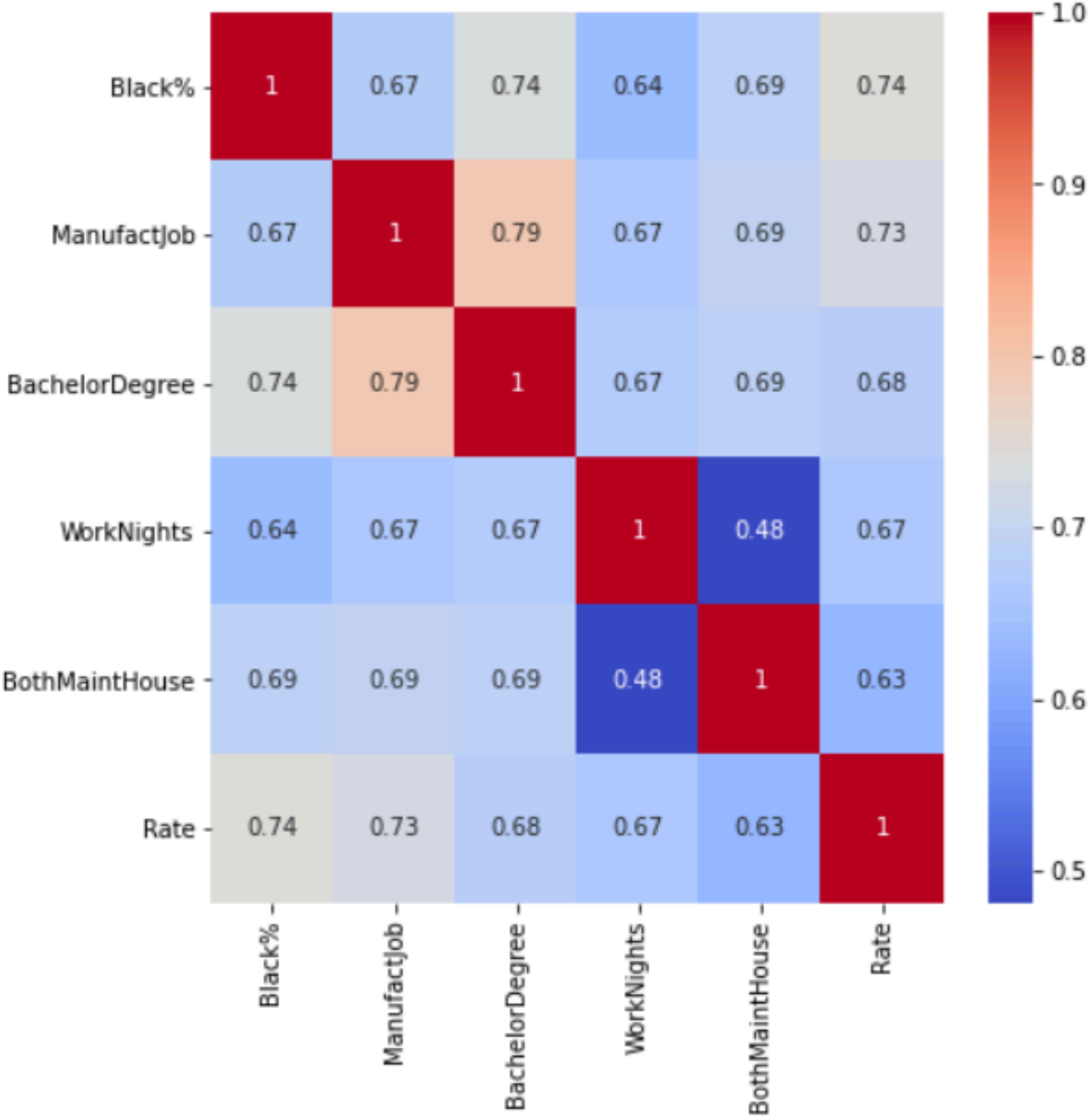
index	Rate per 100,000 people	Category	Characteristic
Col_1269	0.744447	Visible minority	Black
Col_1105	0.741047	Immigration and citizenship	Jamaica
Col_1377	0.731488	Ethnic origin	Jamaican
Col_329	0.727007	Language	Niger-Congo languages
Col_1855	0.726108	Labour	9 Occupations in manufacturing and utilities
Col_105	0.690762	Families, households and marital status	3 or more children
Col_1049	0.684972	Income	In the third decile
Col_1635	0.677412	Education	Bachelor's degree
Col_1907	0.666669	Journey to work	Between 12 p.m. and 4:59 a.m.
Col_1594	0.629968	Housing	2 household maintainers

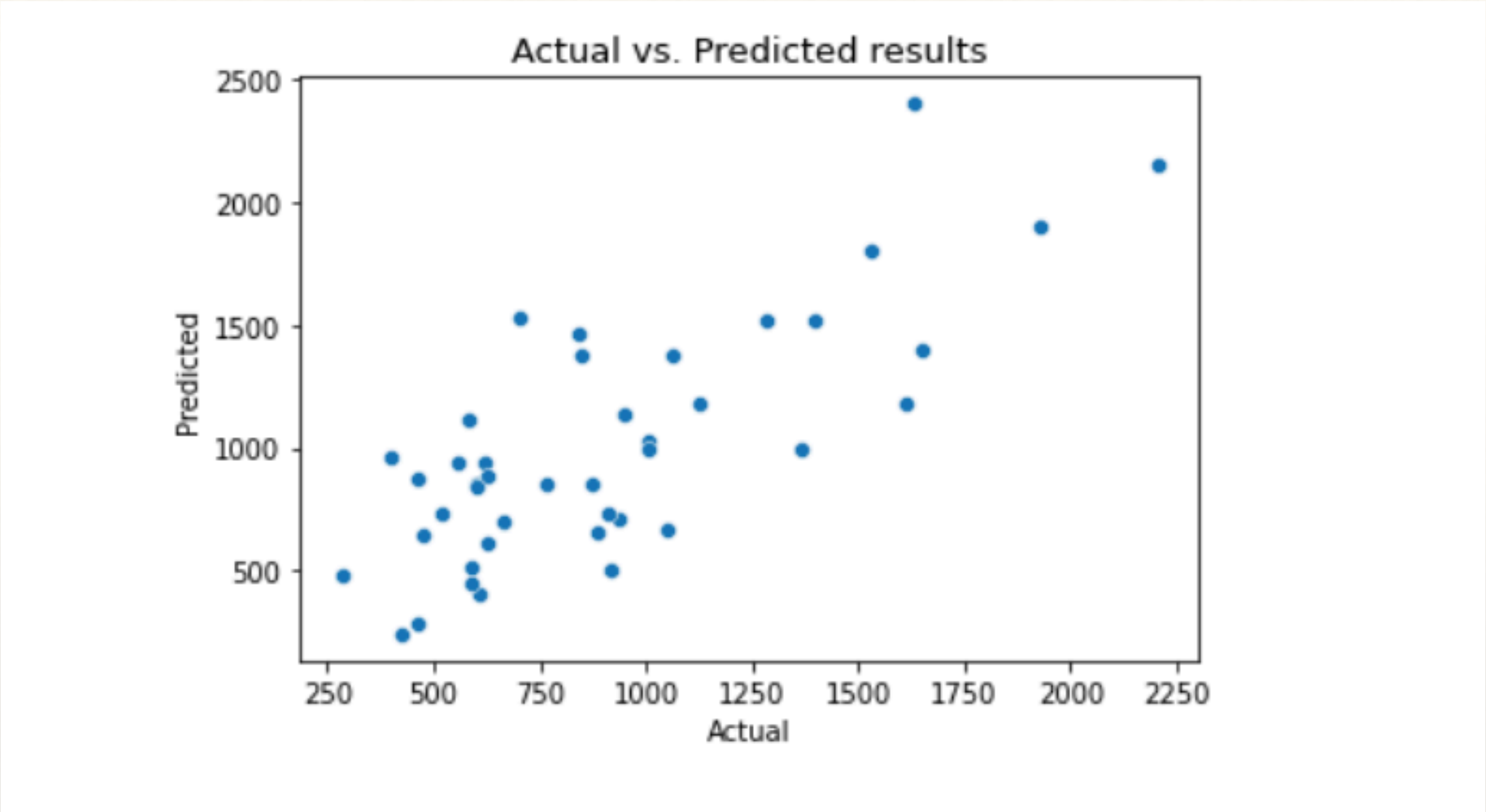
Choosing factors

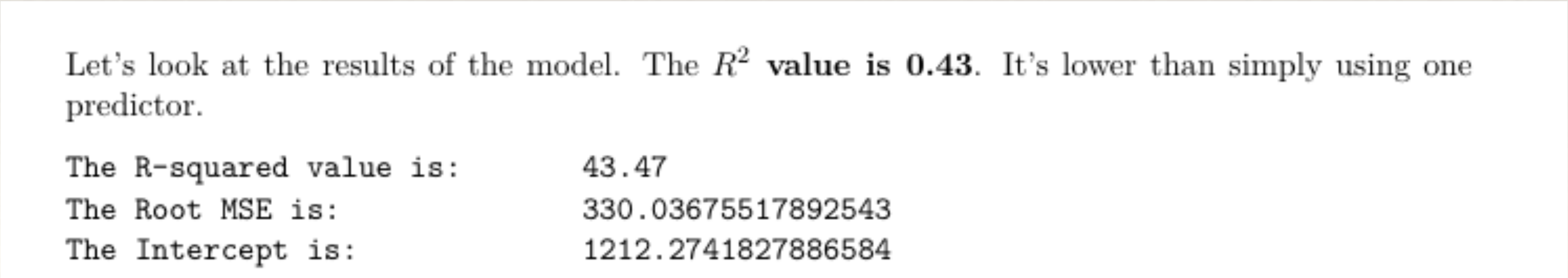

Distributions of Our Predictors
Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Both Maintain House
Work Nights
Bachelor Degree
Manufacturing Job
Black %



	Black%	Jamaica%	Jamaican	CongoLang	ManufactJob	MoreThan2Kids	ThirdDecIncome	BachelorDegree	WorkNights	BothMaintHouse	Rate
Black%	1	0.92	0.94	0.86	0.67	0.89	0.69	0.74	0.64	0.69	0.74
Jamaica%	0.92	1	0.98	0.76	0.77	0.79	0.66	0.74	0.6	0.66	0.74
Jamaican	0.94	0.98	1	0.77	0.74	0.83	0.65	0.75	0.62	0.65	0.73
CongoLang	0.86	0.76	0.77	1	0.68	0.86	0.62	0.62	0.54	0.63	0.73
ManufactJob	0.67	0.77	0.74	0.68	1	0.71	0.83	0.79	0.67	0.69	0.73
MoreThan2Kids	0.89	0.79	0.83	0.86	0.71	1	0.69	0.75	0.53	0.74	0.69
ThirdDecIncome	0.69	0.66	0.65	0.62	0.83	0.69	1	0.83	0.78	0.78	0.68
BachelorDegree	0.74	0.74	0.75	0.62	0.79	0.75	0.83	1	0.67	0.69	0.68
WorkNights	0.64	0.6	0.62	0.54	0.67	0.53	0.78	0.67	1	0.48	0.67
BothMaintHouse	0.69	0.66	0.65	0.63	0.69	0.74	0.78	0.69	0.48	1	0.63
Rate	0.74	0.74	0.73	0.73	0.73	0.69	0.68	0.68	0.67	0.63	1

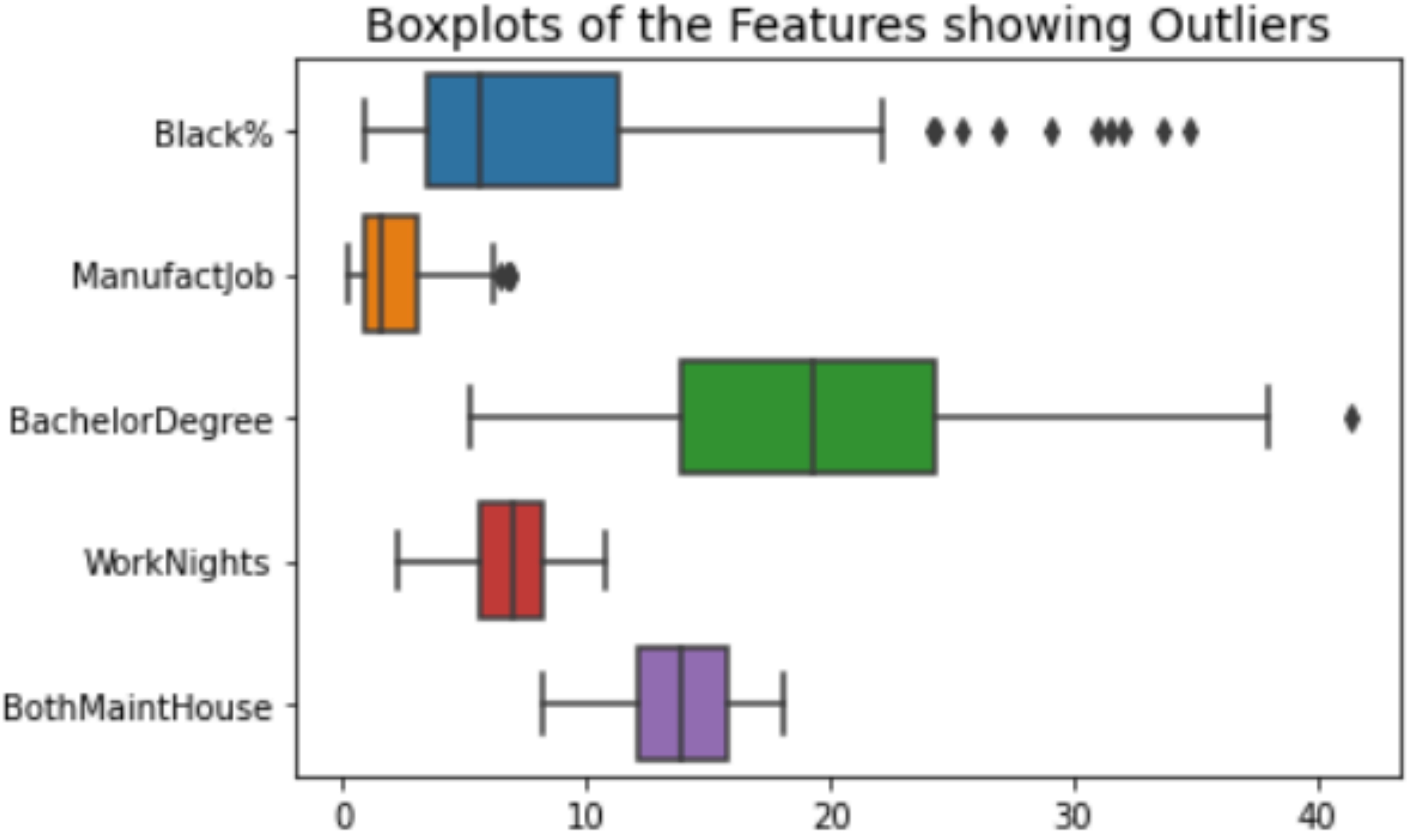

Black%
ManufactJob
BachelorDegree
WorkNights
BothMaintHouse
Rate
1 0.67 0.74 0.64 0.69 0.74
0.67 1 0.79 0.67 0.69 0.73
0.74 0.79 1 0.67 0.69 0.68
0.64 0.67 0.67 1 0.48 0.67
0.69 0.69 0.69 0.48 1 0.63
0.74 0.73 0.68 0.67 0.63 1
Black%
ManufactJob
BachelorDegree
WorkNights
BothMaintHouse
Rate
1.0
0.9
0.8
0.7
0.6
0.5



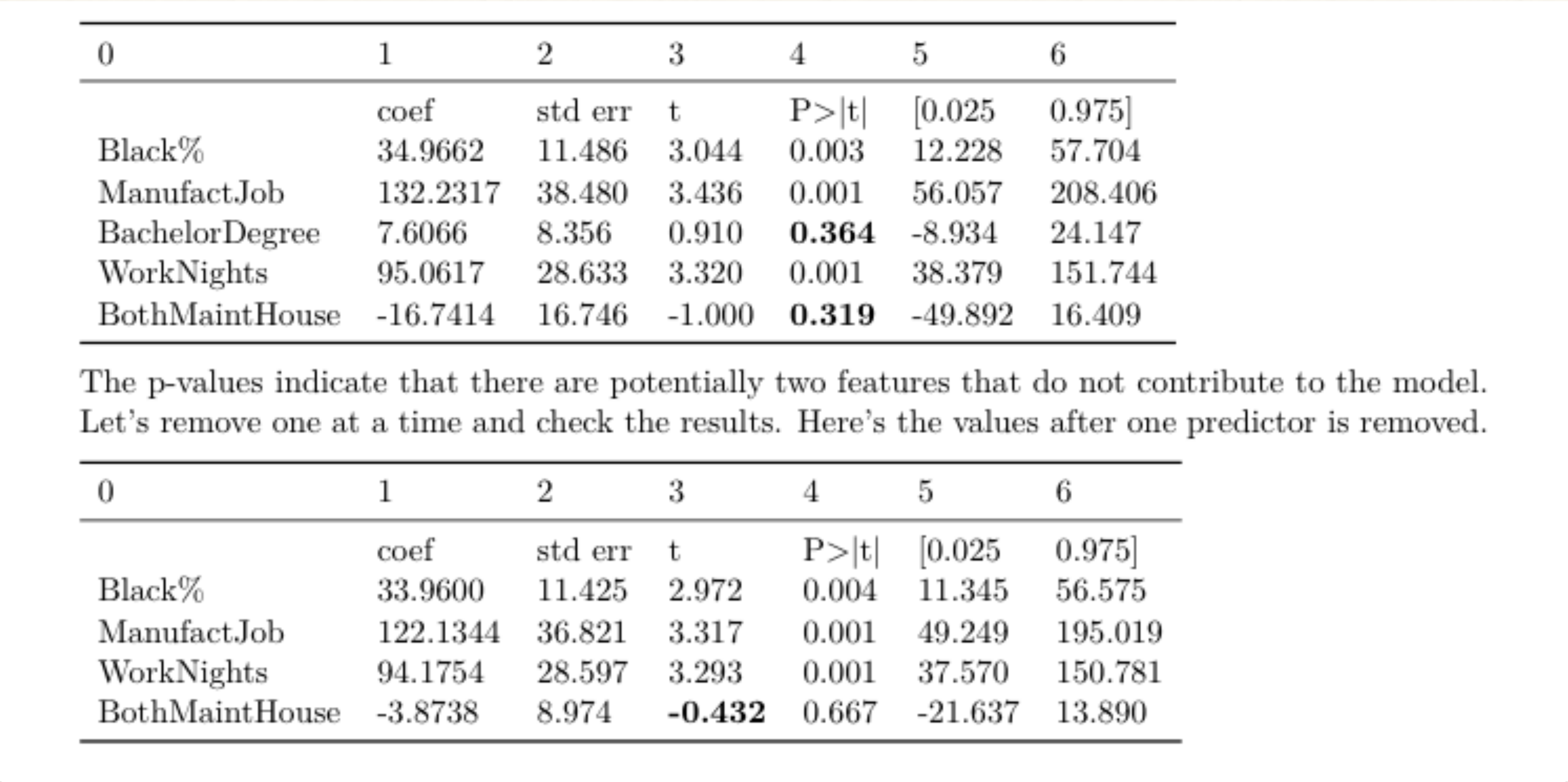


Let's look at the results of the model. The R^2 **value is 0.43**. It's lower than simply using one predictor.

```
The R-squared value is:        43.47
The Root MSE is:               330.03675517892543
The Intercept is:              1212.2741827886584
```



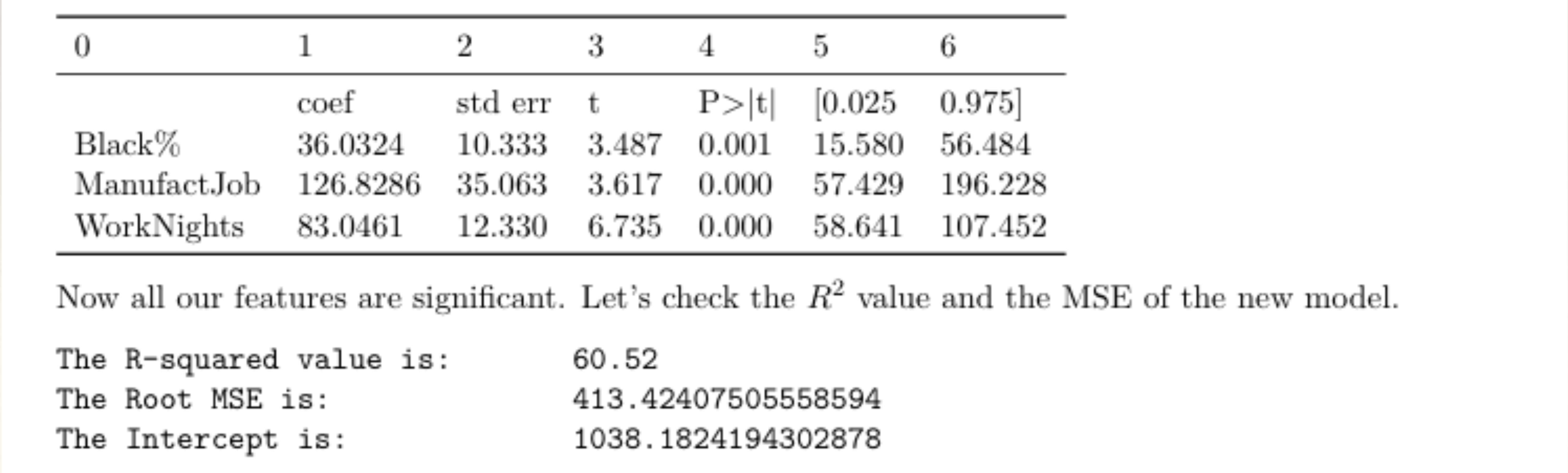
```
The R-squared value is:        56.82
The Root MSE is:               432.34644674183556
The Intercept is:              1031.6761147565676
```



0	1	2	3	4	5	6
	coef	std err	t	P>|t|	[0.025	0.975]
Black%	34.9662	11.486	3.044	0.003	12.228	57.704
ManufactJob	132.2317	38.480	3.436	0.001	56.057	208.406
BachelorDegree	7.6066	8.356	0.910	**0.364**	-8.934	24.147
WorkNights	95.0617	28.633	3.320	0.001	38.379	151.744
BothMaintHouse	-16.7414	16.746	-1.000	**0.319**	-49.892	16.409

The p-values indicate that there are potentially two features that do not contribute to the model. Let's remove one at a time and check the results. Here's the values after one predictor is removed.

0	1	2	3	4	5	6
	coef	std err	t	P>|t|	[0.025	0.975]
Black%	33.9600	11.425	2.972	0.004	11.345	56.575
ManufactJob	122.1344	36.821	3.317	0.001	49.249	195.019
WorkNights	94.1754	28.597	3.293	0.001	37.570	150.781
BothMaintHouse	-3.8738	8.974	**-0.432**	0.667	-21.637	13.890

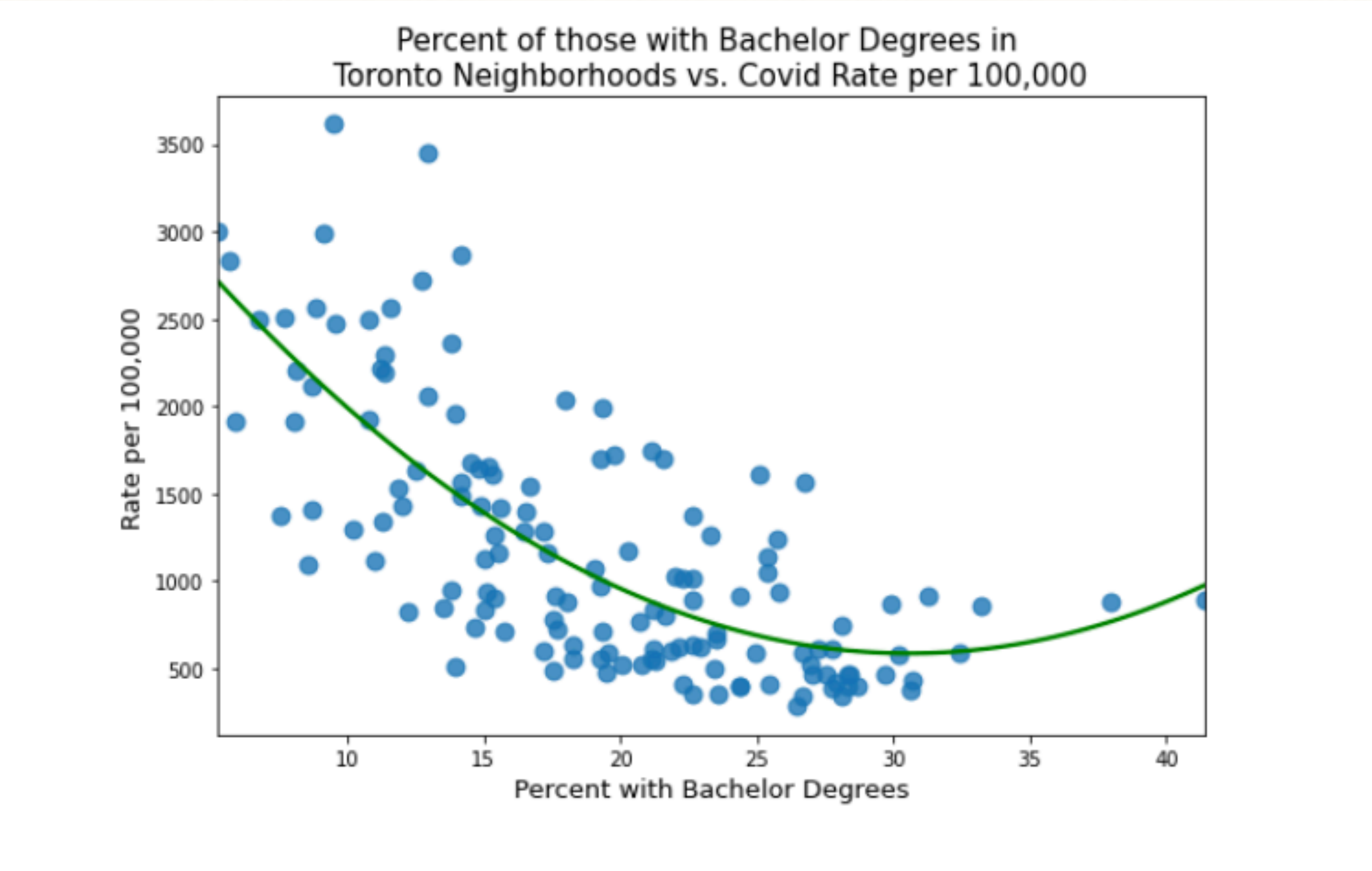


0	1	2	3	4	5	6
	coef	std err	t	P>|t|	[0.025	0.975]
Black%	36.0324	10.333	3.487	0.001	15.580	56.484
ManufactJob	126.8286	35.063	3.617	0.000	57.429	196.228
WorkNights	83.0461	12.330	6.735	0.000	58.641	107.452

Now all our features are significant. Let's check the R^2 value and the MSE of the new model.

```
The R-squared value is:       60.52
The Root MSE is:              413.42407505558594
The Intercept is:             1038.1824194302878
```

- Linear Model is a good choice
- More study would be interesting
- Rate = 1038 + 157.7 * Black% + 140.8 * ManufactJob + 170.1 * WorkNights
- Racial Differences - underlying health, dense neighborhoods, lower % can work at home


Percent of those with Bachelor Degrees in
Toronto Neighborhoods vs. Covid Rate per 100,000
Rate per 100,000
3500
3000
2500
2000
1500
1000
500
10
15
20
25
30
35
40
Percent with Bachelor Degrees