# Female-Headed Households in South Africa

By: Catherine Schuster, Jason Winter, Qing Cheng, and Alia Bly

“Female-headed households face greater social and economic challenges.”

- Zindi

# Project Objective:

Build a predictive model that accurately estimates the percentage of households per ward that are *female-headed* and *living below* an annual income of R19,600.

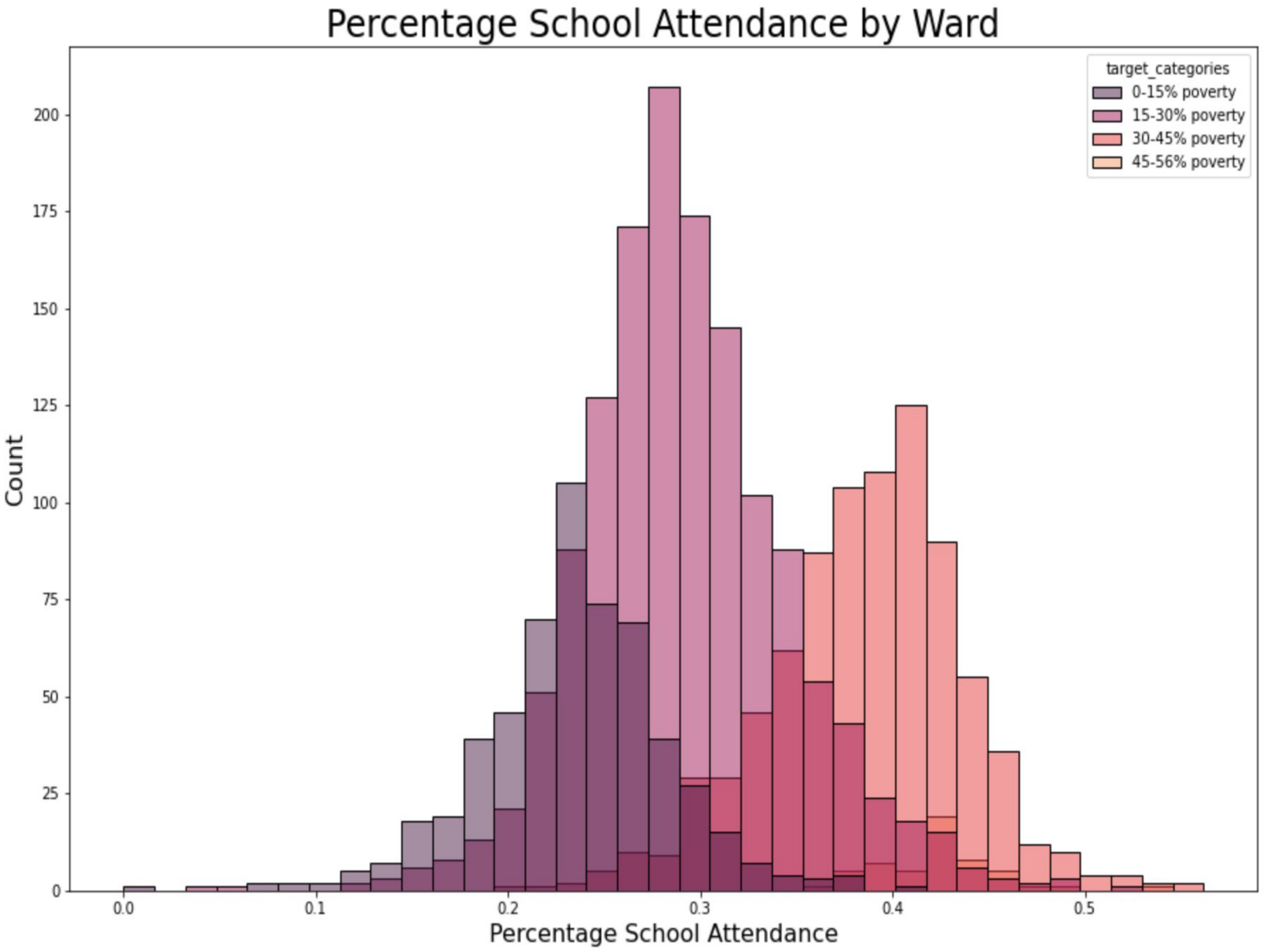
# Overview of Variables

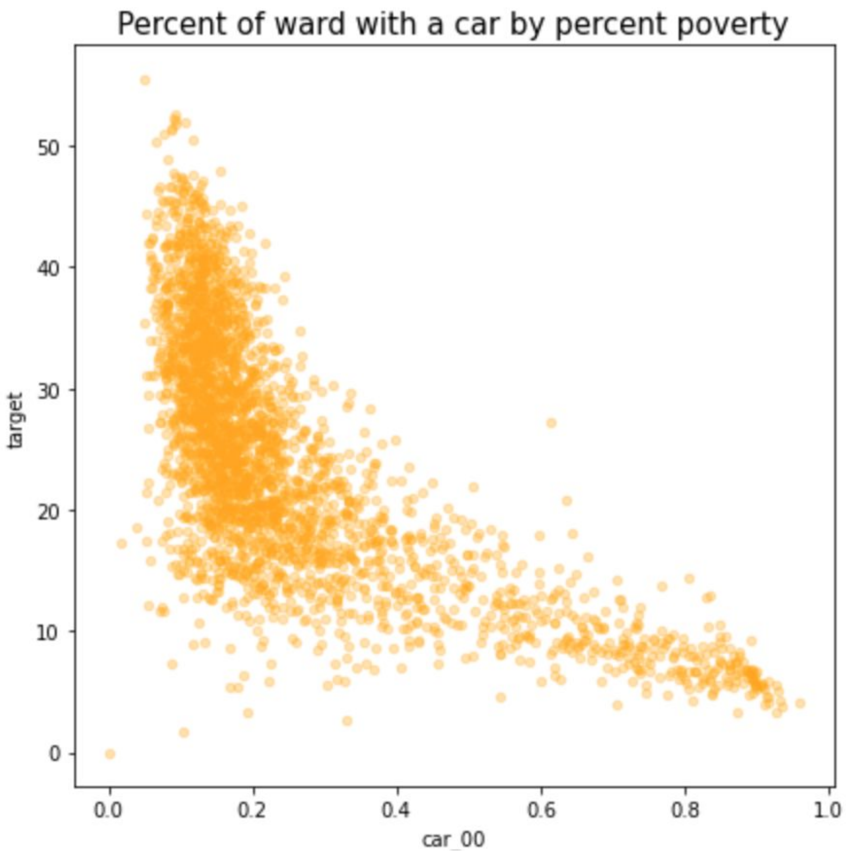
- **dw_[...]** - percentage of dwelling type
- **psa_[...]** - percentage listing present school attendance
- **stv_[...]** - percentage of households with a satellite TV
- **car_[...]** - percentage of households with a car
- **lan_[...]** - percentage of households speaking a specific language
- **pg_[...]** - percentage within a population group by racial identity
- **lgt_00** - percentage of households that use electricity for light
- **pw_[...]** - percentage of households with piped water
- **ADM4_PCODE** - code to link wards
- **lat** - latitudinal value at the midpoint of the ward
- **lon** - longitudinal value at the midpoint of the ward
- **target** - percentage of women head households with income under R19.6k out of total number of households

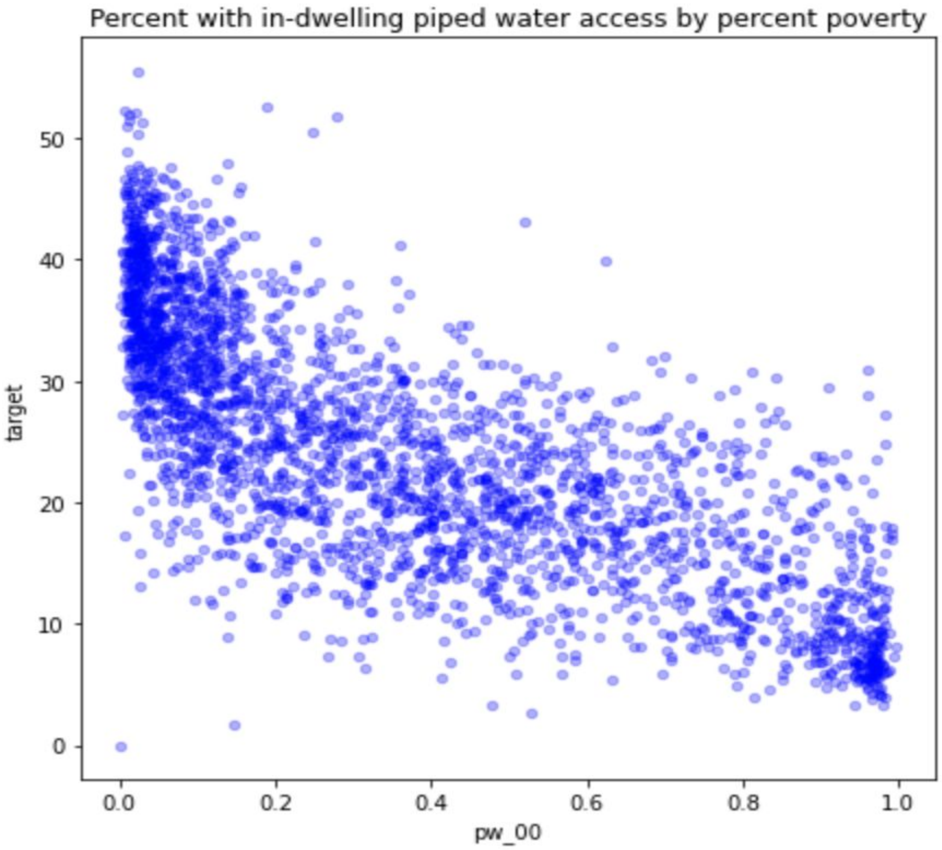
```
var_desc = pd.read_csv('/Users/aliably/Desktop/iX GroupProj. Data/variable_descriptions.csv')
pd.set_option('display.max_colwidth', 200) # So that we can see the full descriptions
var_desc
```

| | Column | Description | Unnamed: 2 | Unnamed: 3 |
|---|---|---|---|---|
| 0 | dw_00 | Percentage of dwellings of type: House or brick/concrete block structure on a separate stand or yard or on a farm | NaN | NaN |
| 1 | dw_01 | Percentage of dwellings of type: Traditional dwelling/hut/structure made of traditional materials | NaN | NaN |
| 2 | dw_02 | Percentage of dwellings of type: Flat or apartment in a block of flats | NaN | NaN |
| 3 | dw_03 | Percentage of dwellings of type: Cluster house in complex | NaN | NaN |
| 4 | dw_04 | Percentage of dwellings of type: Townhouse (semi-detached house in a complex) | NaN | NaN |
| 5 | dw_05 | Percentage of dwellings of type: Semi-detached house | NaN | NaN |
| 6 | dw_06 | Percentage of dwellings of type: House/flat/room in backyard | NaN | NaN |
| 7 | dw_07 | Percentage of dwellings of type: Informal dwelling (shack | in backyard) | NaN |
| 8 | dw_08 | Percentage of dwellings of type: Informal dwelling (shack | not in backyard | e.g. in an informal/squatter settlement or on a farm) |
| 9 | dw_09 | Percentage of dwellings of type: Room/flatlet on a property or larger dwelling/servants quarters/granny flat | NaN | NaN |
| 10 | dw_10 | Percentage of dwellings of type: Caravan/tent | NaN | NaN |
| 11 | dw_11 | Percentage of dwellings of type: Other | NaN | NaN |
| 12 | dw_12 | Percentage of dwellings of type: Unspecified | NaN | NaN |
| 13 | dw_13 | Percentage of dwellings of type: Not applicable | NaN | NaN |
| 14 | psa_00 | Percentage listing present school attendance as: Yes | NaN | NaN |
| 15 | psa_01 | Percentage listing present school attendance as: No | NaN | NaN |
| 16 | psa_02 | Percentage listing present school attendance as: Do not know | NaN | NaN |
| 17 | psa_03 | Percentage listing present school attendance as: Unspecified | NaN | NaN |
| 18 | psa_04 | Percentage listing present school attendance as: Not applicable | NaN | NaN |

# EDA Process

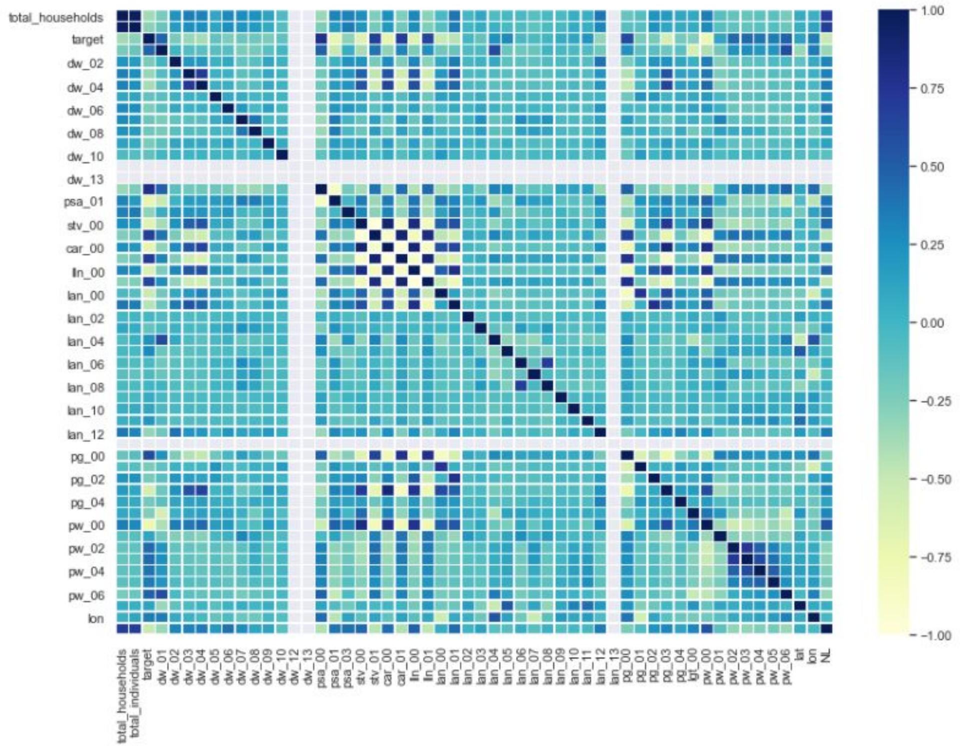

Percentage School Attendance by Ward
target_categories
0-15% poverty
15-30% poverty
30-45% poverty
45-56% poverty
Count
0
25
50
75
100
125
150
175
200
0.0
0.1
0.2
0.3
0.4
0.5
Percentage School Attendance


Percent of ward with a car by percent poverty
target
car_00
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0


Percent with in-dwelling piped water access by percent poverty
target
pw_00
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0

# Correlation Analysis

Corr > 0.5 or < -0.5
Corr > 0.7 or < -0.7

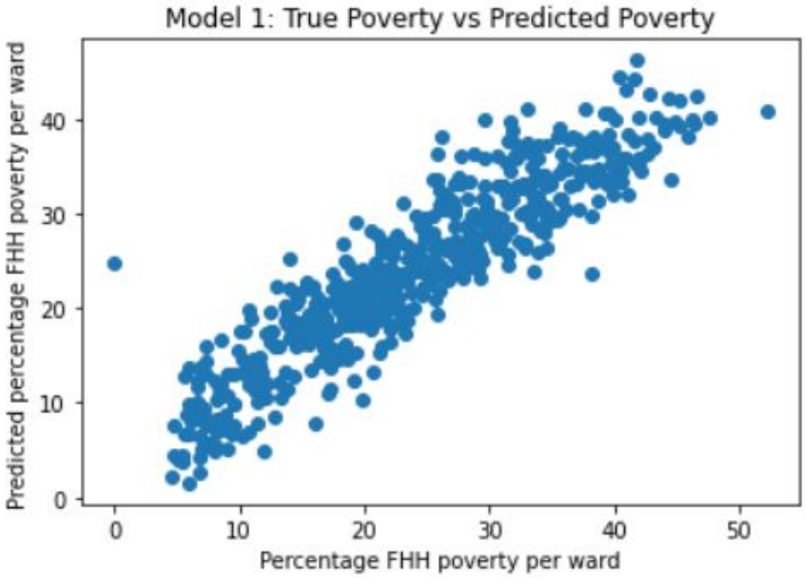



```
In [14]:  df.corr()['target'].sort_values(ascending=False)

Out[14]:
target              1.000000
psa_00              0.782472
car_01              0.702831
stv_01              0.664181
lln_01              0.637835
pg_00               0.613346
pw_06               0.470676
dw_01               0.458206
pw_02               0.442441
pw_03               0.440941
pw_04               0.389467
pw_05               0.349653
lon                 0.347088
lan_05              0.275263
...
psa_03             -0.301572
lan_12             -0.325367
dw_03              -0.338137
total_households   -0.374833
dw_04              -0.385533
lan_01             -0.438704
lan_00             -0.507942
NL                 -0.514398
pg_03              -0.583908
lln_00             -0.637835
stv_00             -0.664181
car_00             -0.702831
psa_01             -0.707506
pw_00              -0.754536
dw_12                    NaN
dw_13                    NaN
lan_13                   NaN
Name: target, dtype: float64
```

# Models 1 and 2

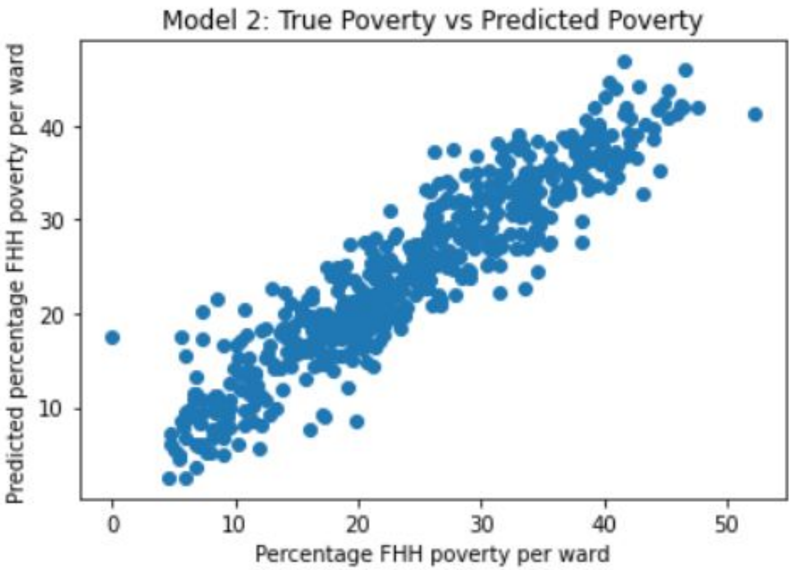
Model 1: Linear Regression with the most correlated features

R^2: 0.8358972191455909

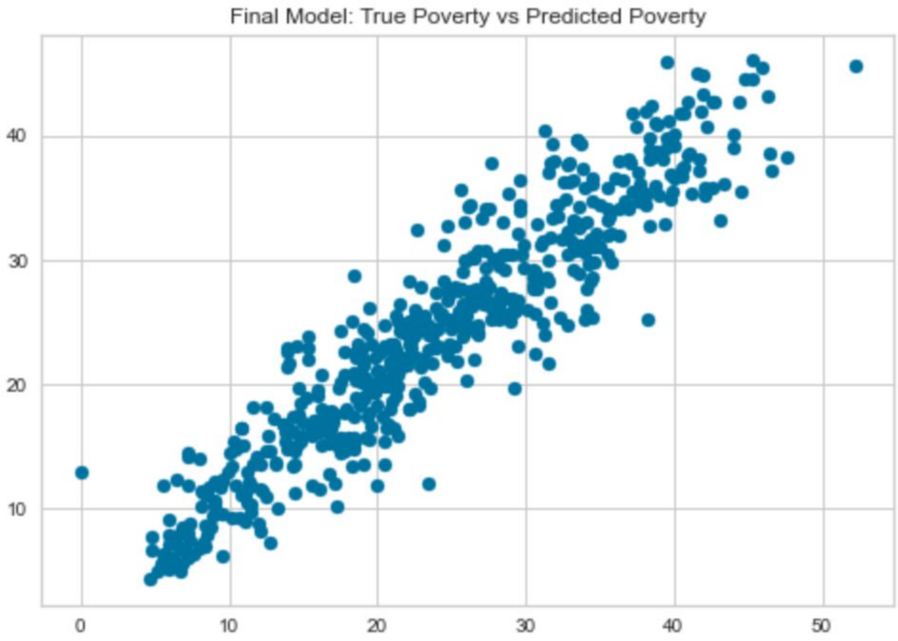



Model 2: Linear Regression with *all* the features

R^2: 0.858710238233148



# Final Model and Performance Metrics



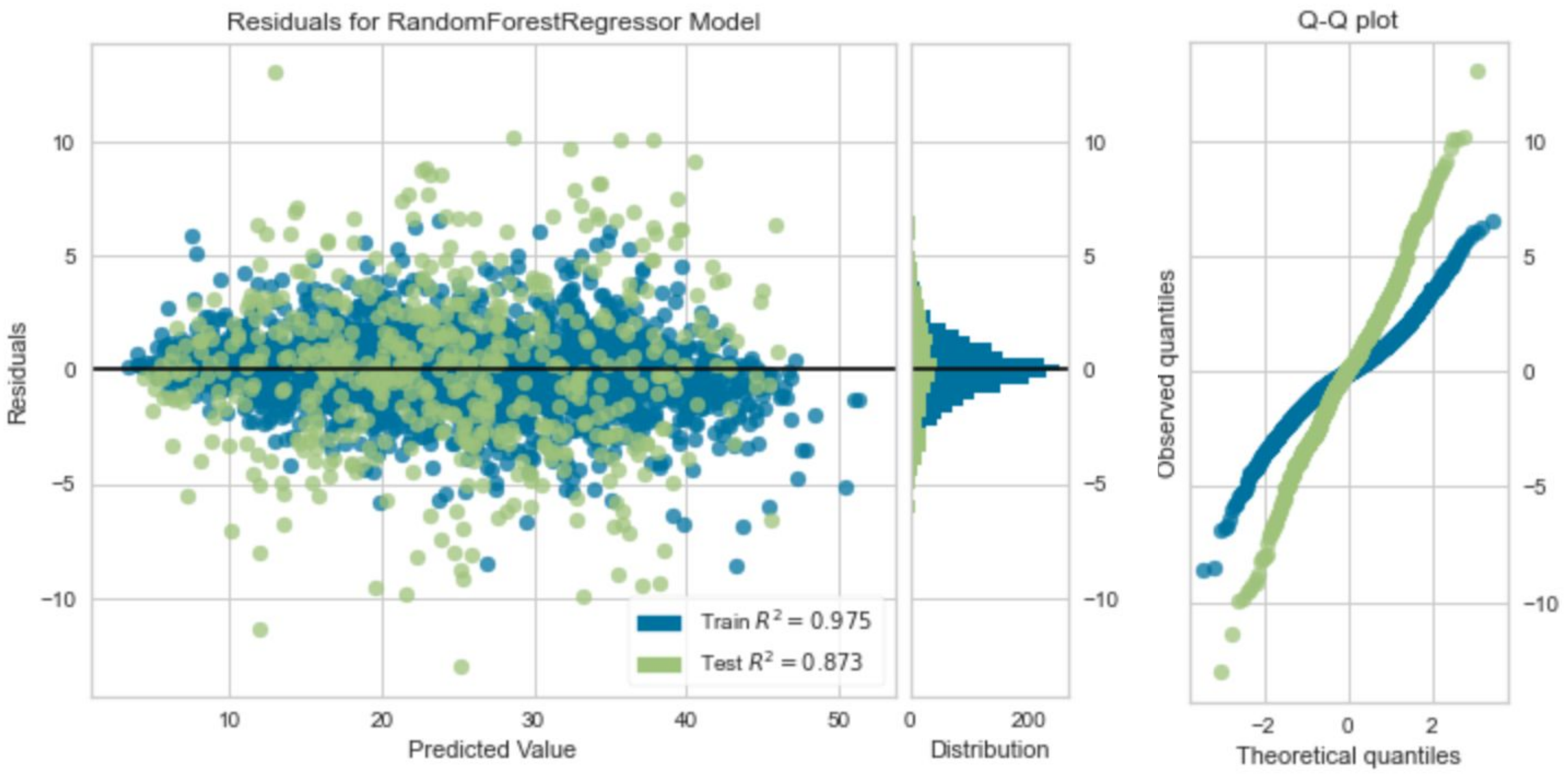

| Adj $R^2$ | MAE | MSE | RMSE | 5 Fold CV RMSE |
|---|---|---|---|---|
| .87 | 2.78 | 13.45 | 3.67 | 4.22 |

[171]:
```
from sklearn.ensemble import RandomForestRegressor

# Create a Random Forest Regressor
reg = RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
           oob_score=False, random_state=17, verbose=0, warm_start=False)
```

[172]:
```
reg.fit(X_train1, y_train1)
y_pred3 = reg.predict(X_test1)
```

[173]:
```
acc_rf_train = metrics.r2_score(y_test1, y_pred3)
print('R^2:',metrics.r2_score(y_test1, y_pred3))
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_test1, y_pred3))*(len(y_test1)-1)/(len(y_test1)-X_test1.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_test1, y_pred3))
print('MSE:',metrics.mean_squared_error(y_test1, y_pred3))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test1, y_pred3)))
```

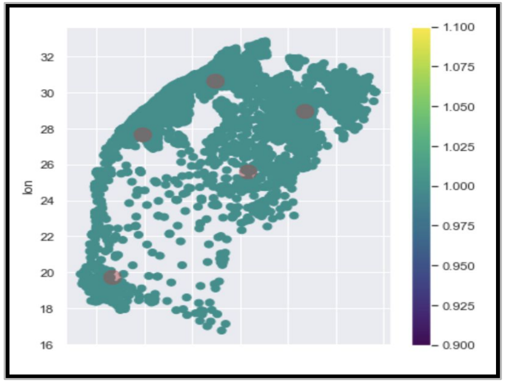
```
R^2: 0.8727399644546371
Adjusted R^2: 0.8692629143031244
MAE: 2.780667208904761
MSE: 13.446520152980385
RMSE: 3.6669497069063253
```

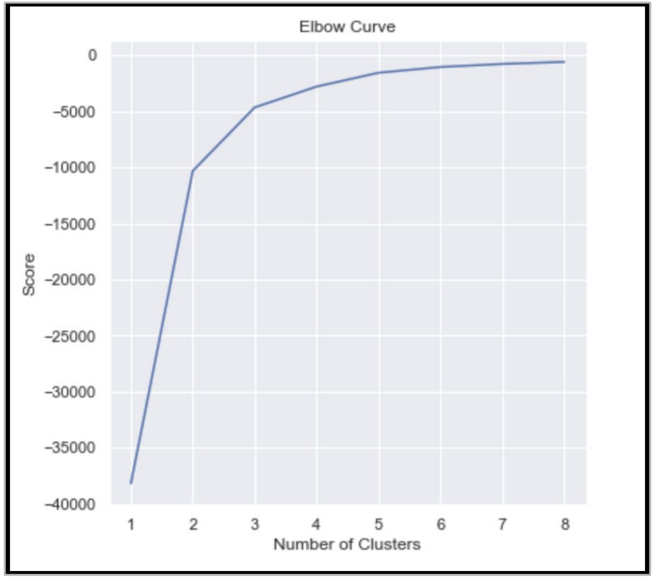

Residuals for RandomForestRegressor Model
Residuals
Predicted Value
Distribution
Train $R^2 = 0.975$
Test $R^2 = 0.873$
Q-Q plot
Observed quantiles
Theoretical quantiles

# Geospatial Analysis

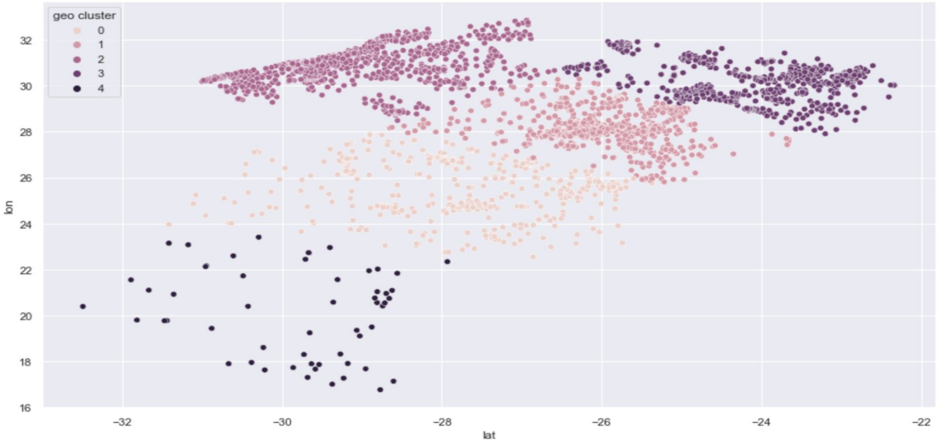
- **What is geospatial analysis?**
- Elbow curve and K-means clustering:

```python
from sklearn.cluster import KMeans
K_clusters = range(1,9)

#range is shifted from 0-4 to 1-5 to avoid infinity-type error

kmeans = [KMeans(n_clusters = i) for i in K_clusters]

Y_axis = all_data[['lat']]
X_axis = all_data[['lon']]

score = [kmeans[i].fit(Y_axis).score(Y_axis) for i in range(len(kmeans))]

# Visualization

plt.plot(K_clusters, score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

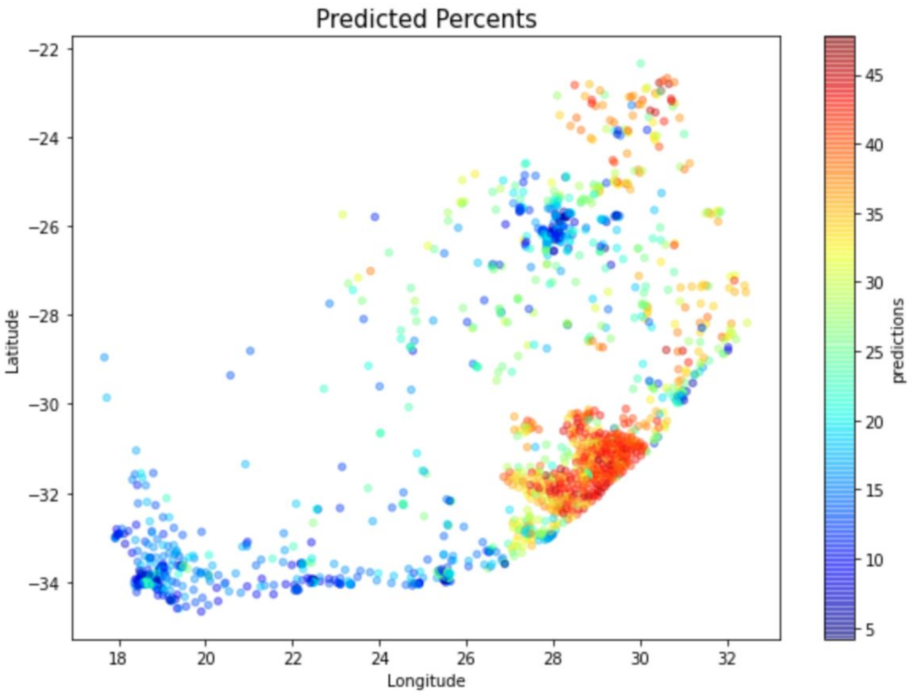


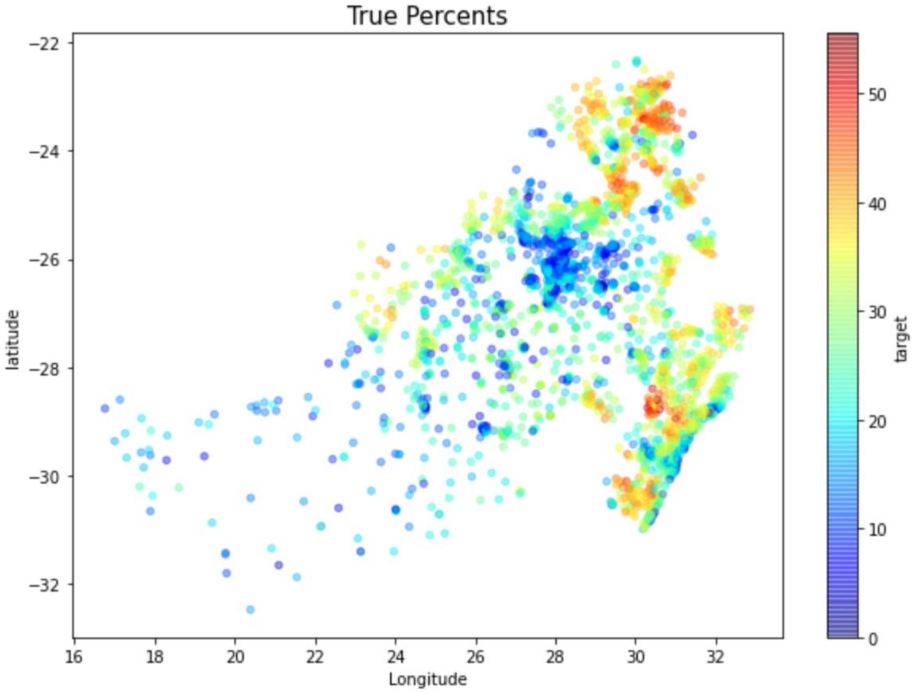




```
In [255]: plt.figure(figsize = (15,8))
          sns.scatterplot(train_data['lat'], train_data['lon'], hue=train_data['geo cluster'])
Out[255]: <AxesSubplot:xlabel='lat', ylabel='lon'>
```



# Geospatial Analysis: True vs. Predicted Values



# Final Thoughts

- Overall satisfaction with model’s performance
  - Future directions:
    - Further investigate the impacts of feature engineering
    - Better develop geospatial models
    - Apply new models

# Works Cited


Bittar, A. (2020, August 14). *Poverty On the Rise in South Africa*. The Borgen Project. borgenproject.org/poverty-in-south-africa/.

Boeing, G. (2018, March 22). Clustering to Reduce Spatial Data Set Size. https://doi.org/10.31235/osf.io/nzhdc

Nwosu, C. O., & Ndinda, C. (2018). Gender-based Household Compositional Changes and Implications for Poverty in South Africa. *Journal of International Women's Studies*, *19*(5), 82–94. doi.org/https://vc.bridgew.edu/cgi/viewcontent.cgi?article= 2046&context=jiws.

*Poverty and Youth in Post-Apartheid South Africa*. (2014, April 15). [Photograph]. Borgen Magazine. www.borgenmagazine.com/poverty-youth-post-apartheid-south-africa/.