DS6003

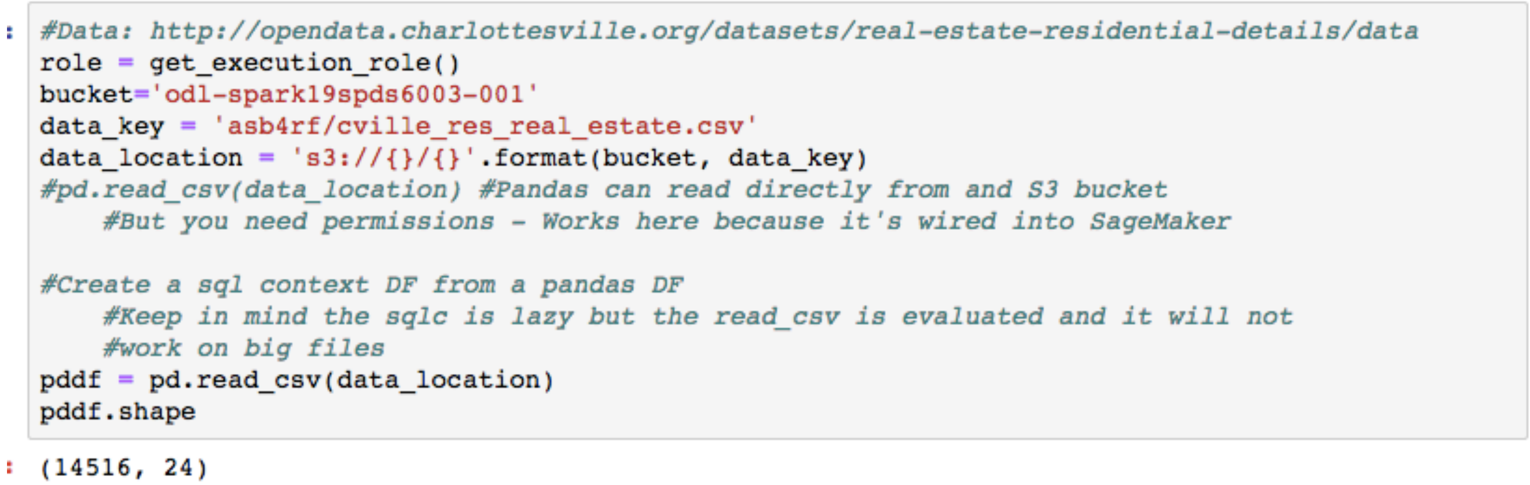
Pipeline Assignment

Motivation

- Using open data from the City of Charlottesville
- Specifically using residential home information
- Target is to predict total square footage
- Features are limited based on time to tinker

Code snippet

- Gist of scale:

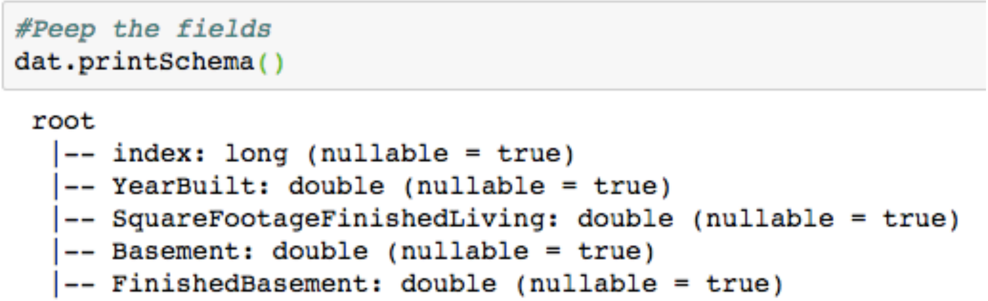


```
#Data: http://opendata.charlottesville.org/datasets/real-estate-residential-details/data
role = get_execution_role()
bucket='odl-spark19spds6003-001'
data_key = 'asb4rf/cville_res_real_estate.csv'
data_location = 's3://{}/{}'.format(bucket, data_key)
#pd.read_csv(data_location) #Pandas can read directly from and S3 bucket
    #But you need permissions - Works here because it's wired into SageMaker

#Create a sql context DF from a pandas DF
    #Keep in mind the sqlc is lazy but the read_csv is evaluated and it will not
    #work on big files
pddf = pd.read_csv(data_location)
pddf.shape
```

```
(14516, 24)
```

- And actual used fields:

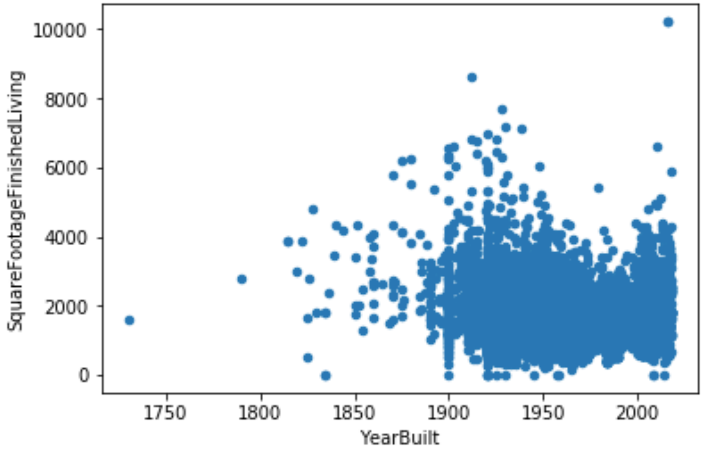


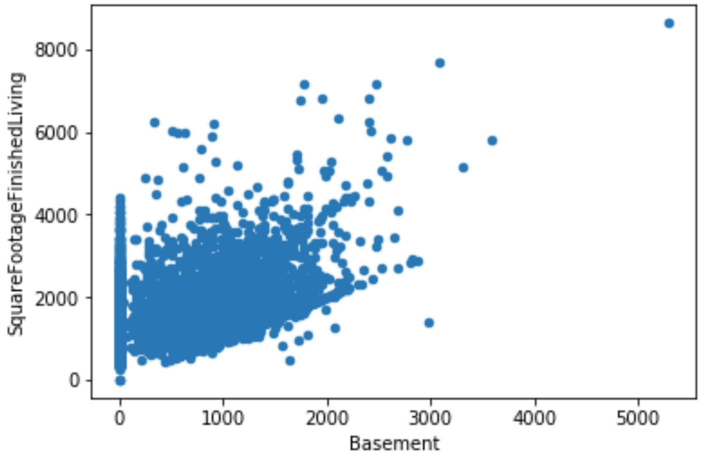
```
#Peep the fields
dat.printSchema()
```

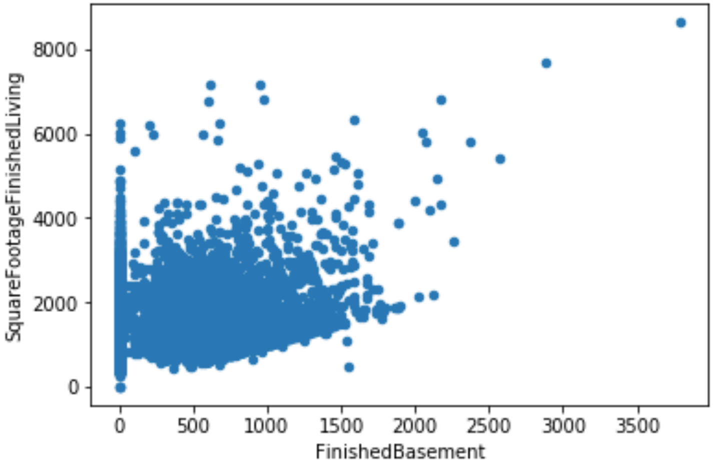
```
root
 |-- index: long (nullable = true)
 |-- YearBuilt: double (nullable = true)
 |-- SquareFootageFinishedLiving: double (nullable = true)
 |-- Basement: double (nullable = true)
 |-- FinishedBasement: double (nullable = true)
```

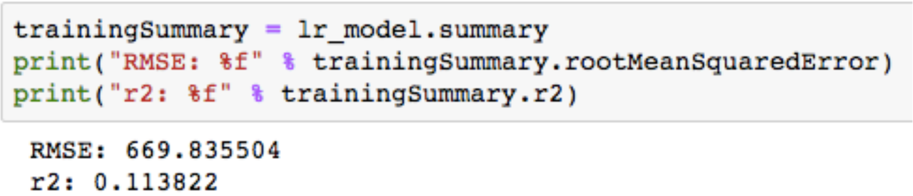
Visualization

- Simply looking at xy of the explanatory variables with the response:









```
trainingSummary = lr_model.summary
print("RMSE: %f" % trainingSummary.rootMeanSquaredError)
print("r2: %f" % trainingSummary.r2)
```

```
RMSE: 669.835504
r2: 0.113822
```