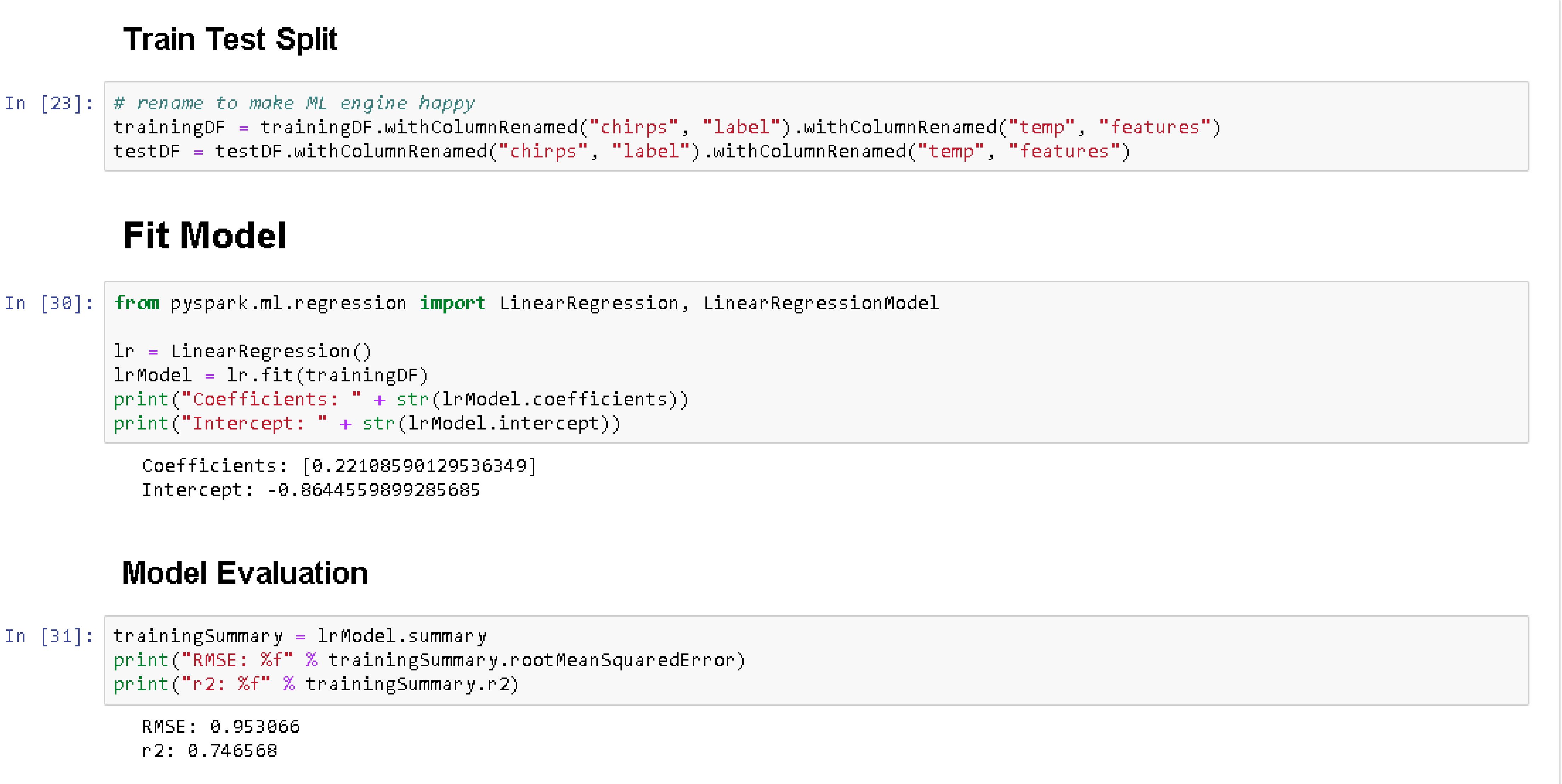
Justin Niestroy Spark Assignment

- Goal was to predict number of cricket chirps at certain temperature



Train Test Split

In [23]:
```
# rename to make ML engine happy
trainingDF = trainingDF.withColumnRenamed("chirps", "label").withColumnRenamed("temp", "features")
testDF = testDF.withColumnRenamed("chirps", "label").withColumnRenamed("temp", "features")
```

Fit Model

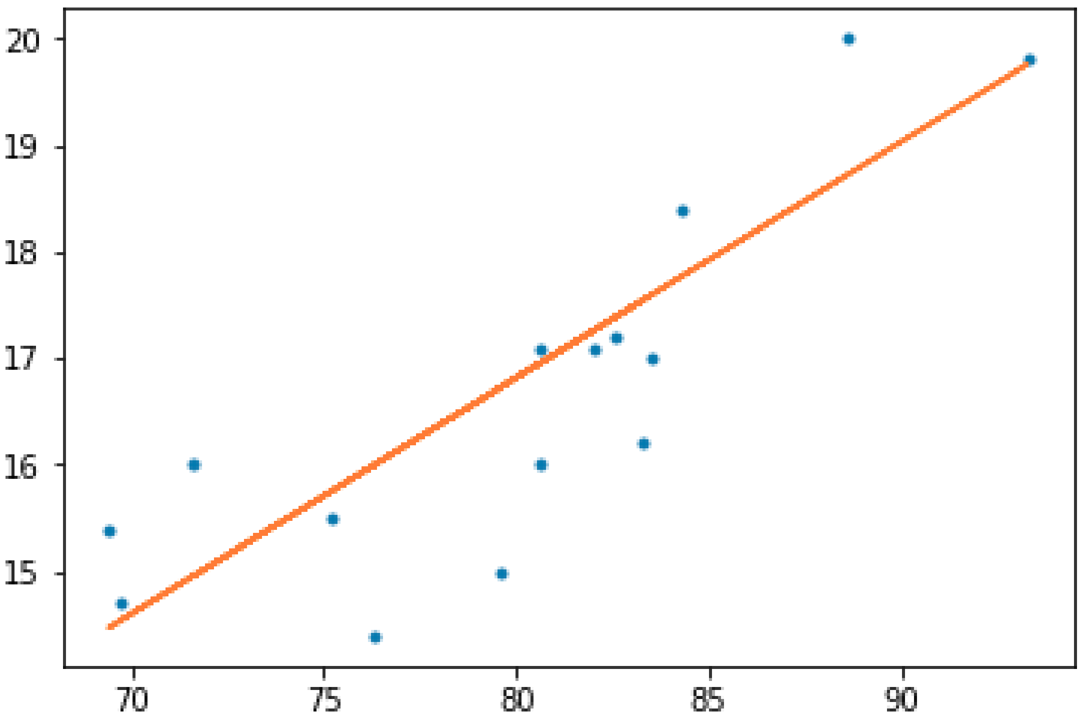
In [30]:
```
from pyspark.ml.regression import LinearRegression, LinearRegressionModel

lr = LinearRegression()
lrModel = lr.fit(trainingDF)
print("Coefficients: " + str(lrModel.coefficients))
print("Intercept: " + str(lrModel.intercept))
```

```
Coefficients: [0.22108590129536349]
Intercept: -0.8644559899285685
```

Model Evaluation

In [31]:
```
trainingSummary = lrModel.summary
print("RMSE: %f" % trainingSummary.rootMeanSquaredError)
print("r2: %f" % trainingSummary.r2)
```

```
RMSE: 0.953066
r2: 0.746568
```


20
19
18
17
16
15
70
75
80
85
90