**1) TOPIC-**

Sports Analytics Using Apache Spark

**2) DESCRIPTION**

In this project we take a look at how we can analyse NBA player’s performance and how players shot selection has evolved throughout the years .We also predict how players will adapt to shot selections in the future . We have used Apache spark and its API’s such as Spark Sql , Spark Mllib .We have used Vector Assembler to covert different feature columns into a single feature vector columns . Linear Regression to train the model and evaluated the model.

**3) Methodologies/Algorithms**

Linear Regression is used to predict a players shot selection distance preference . Also we use vector assembler to covert different feature columns into a single feature vector columns . We make use of visualisation tolls such as matplotlib, sns . Main Api’s used here are the spark Sql and Spark Mllib.

**4)Technologies/Libraries**

Pyspark is used for this project and libraries that have been used in this are pasted below

**Libraries**

import numpy as np

import pandas as pd

import seaborn as sns

from nba\_utils import draw\_3pt\_piechart,plot\_shot\_chart

from IPython.core.display import display, HTML

from IPython.core.magic import register\_cell\_magic, register\_line\_cell\_magic, register\_line\_magic

from matplotlib import pyplot as plt

from pyspark.ml.regression import LinearRegression

from pyspark.ml.feature import VectorAssembler

from pyspark.ml.linalg import Vector

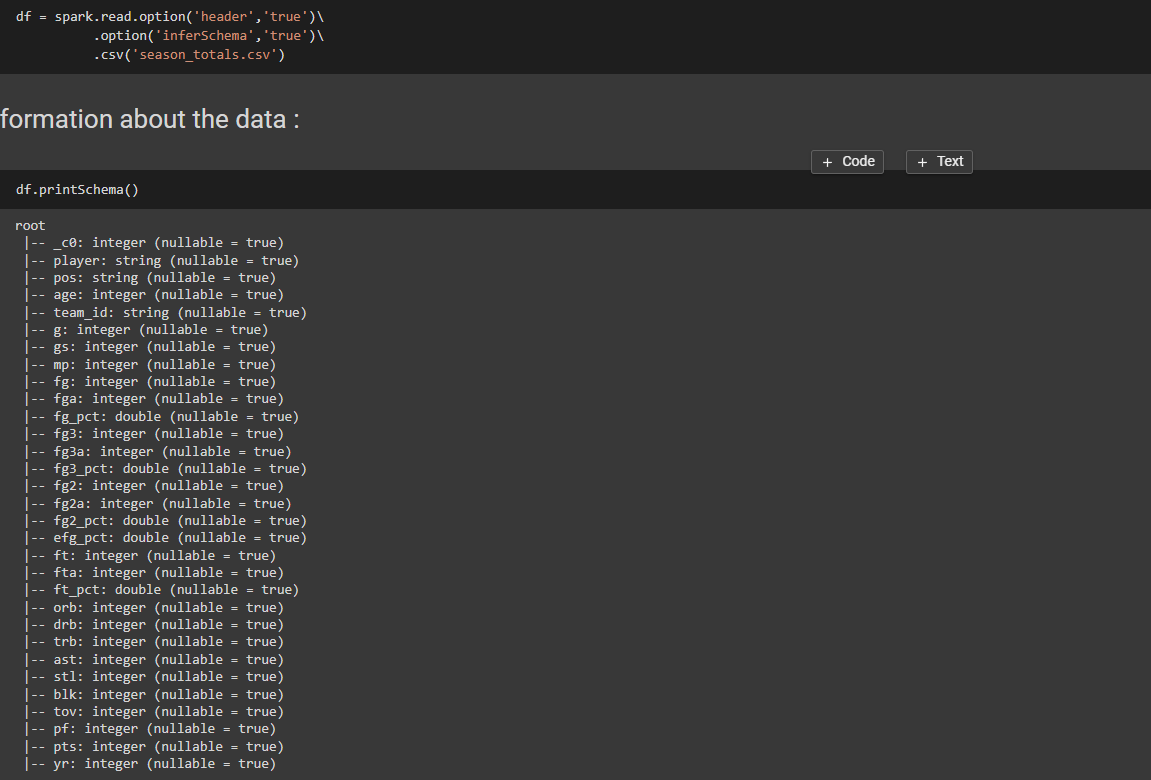
from pyspark.sql import SQLContext

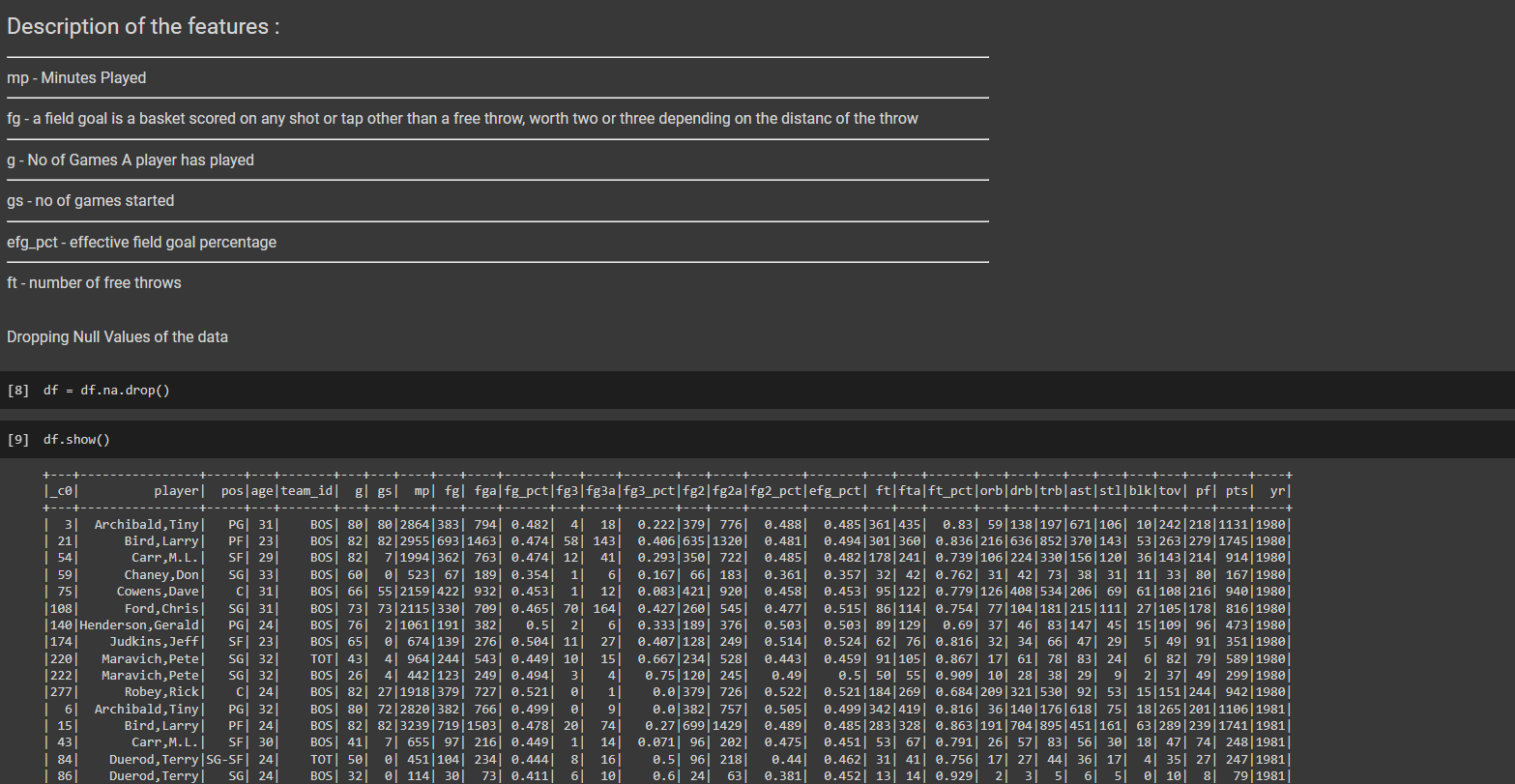
from pyspark.sql.functions import array, col, count, mean, sum, udf, when

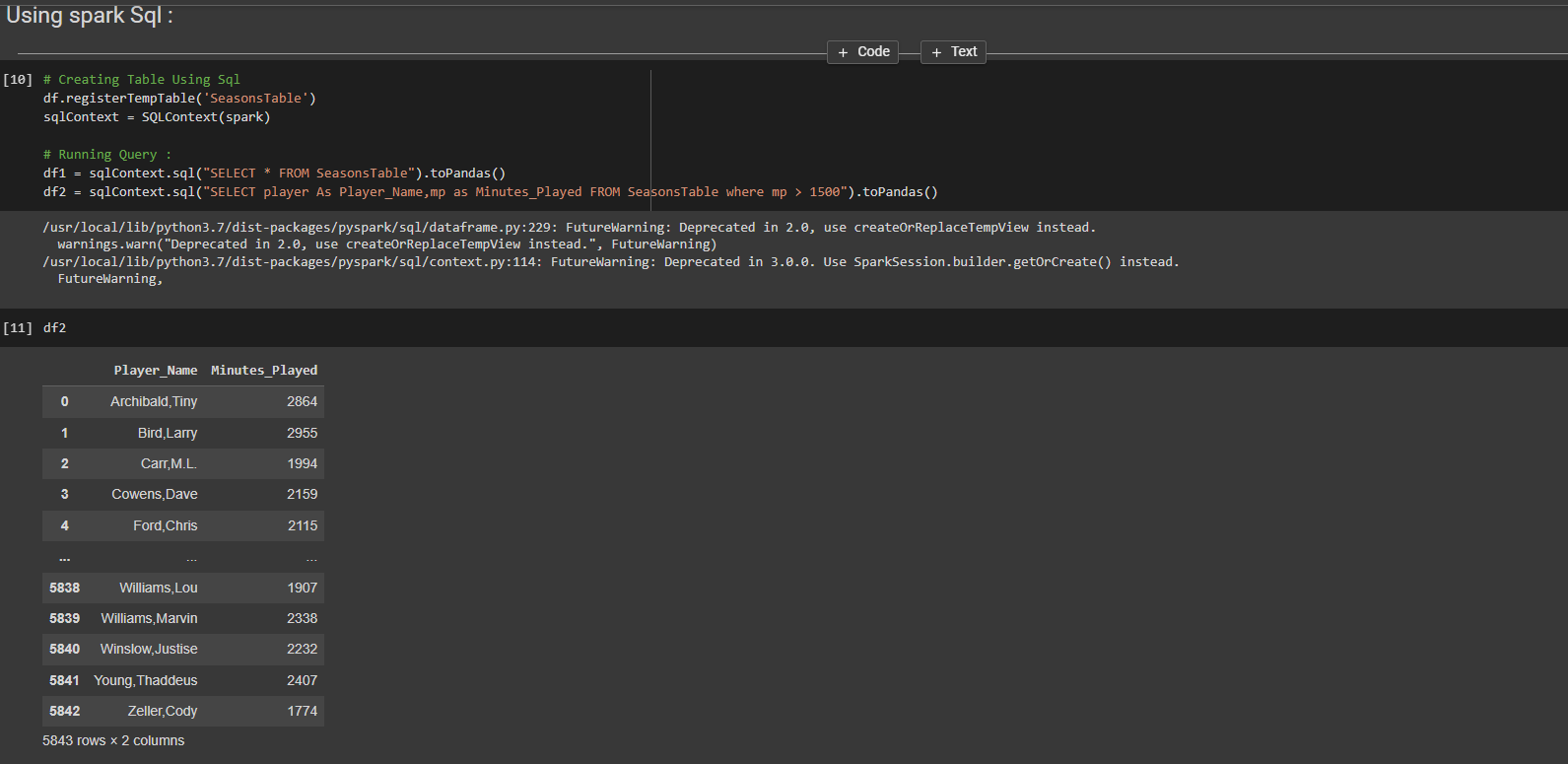
from pyspark.sql.types import DoubleType, IntegerType, StringType, Row

from pyspark.sql.functions import sum, col, udf

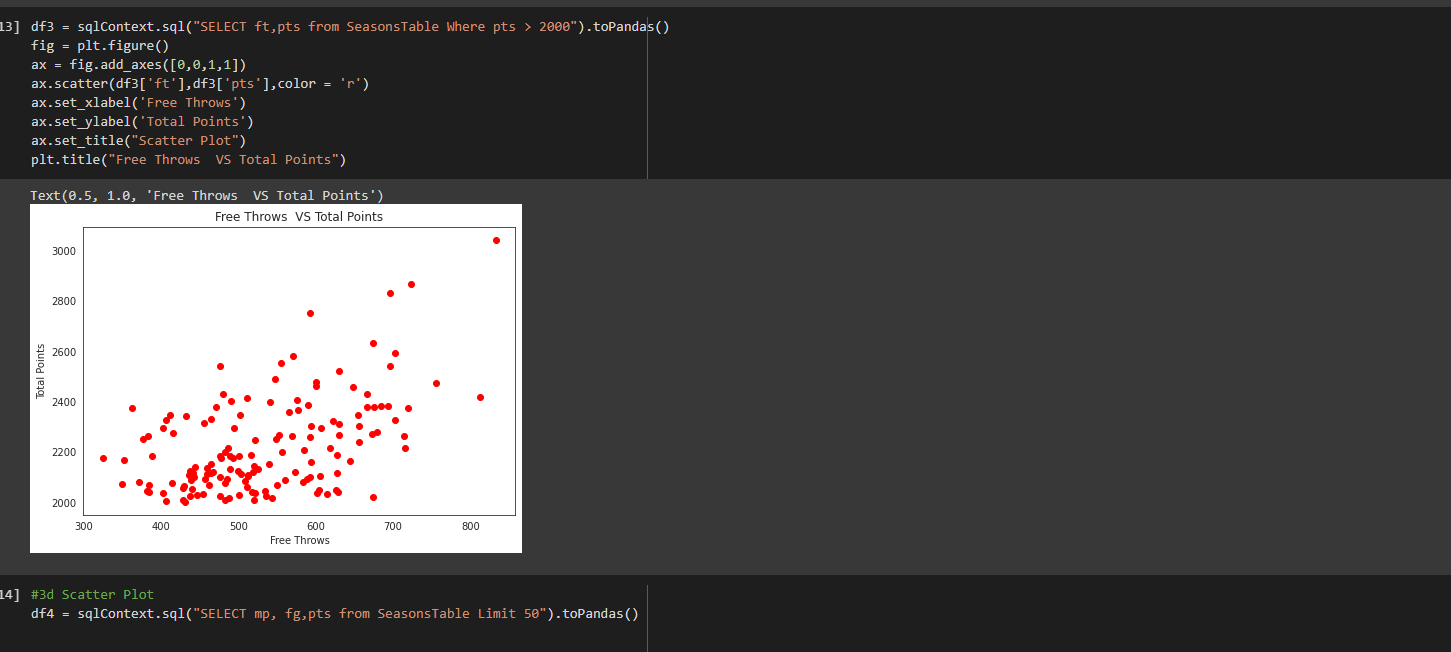
**5) USER MANUAL**

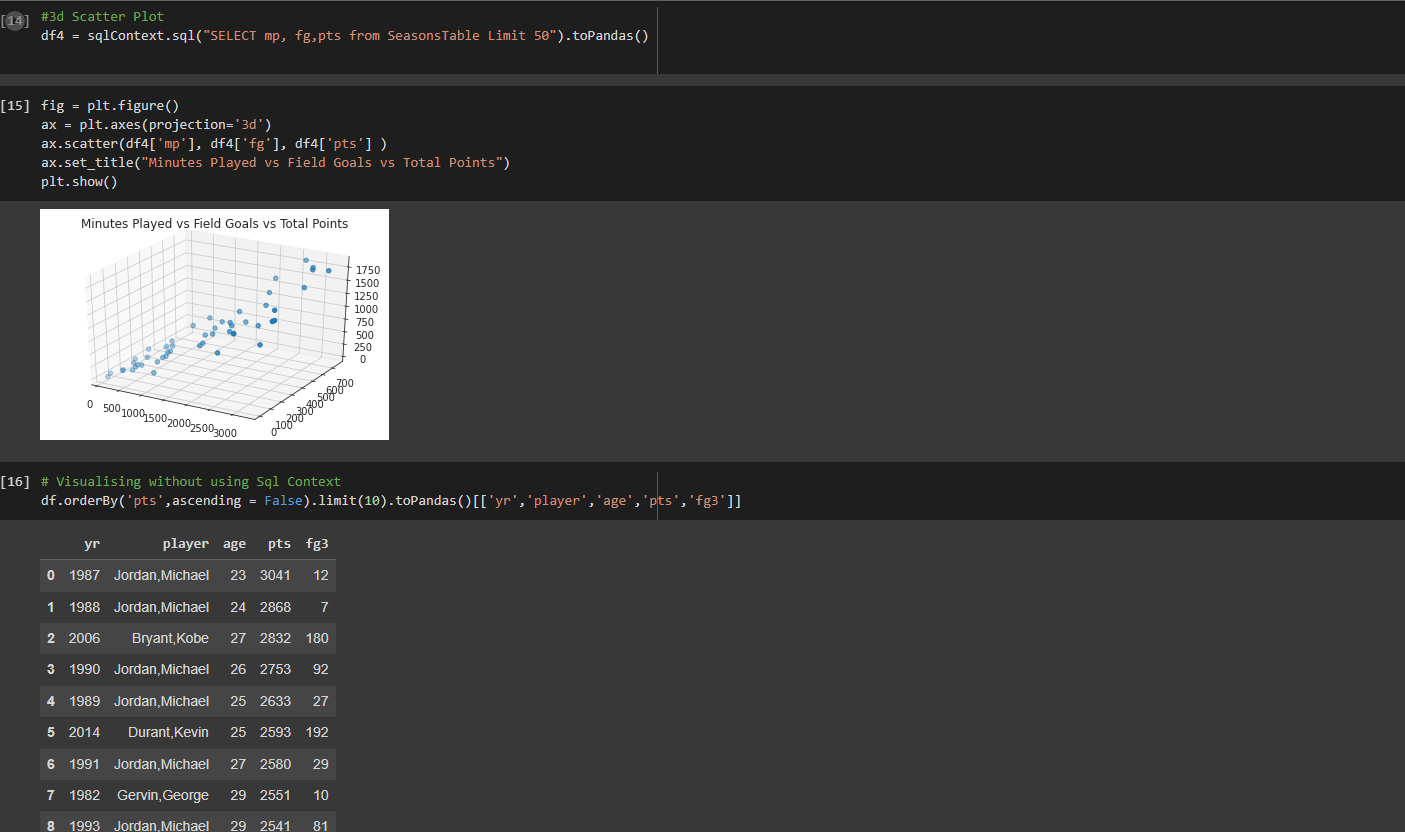


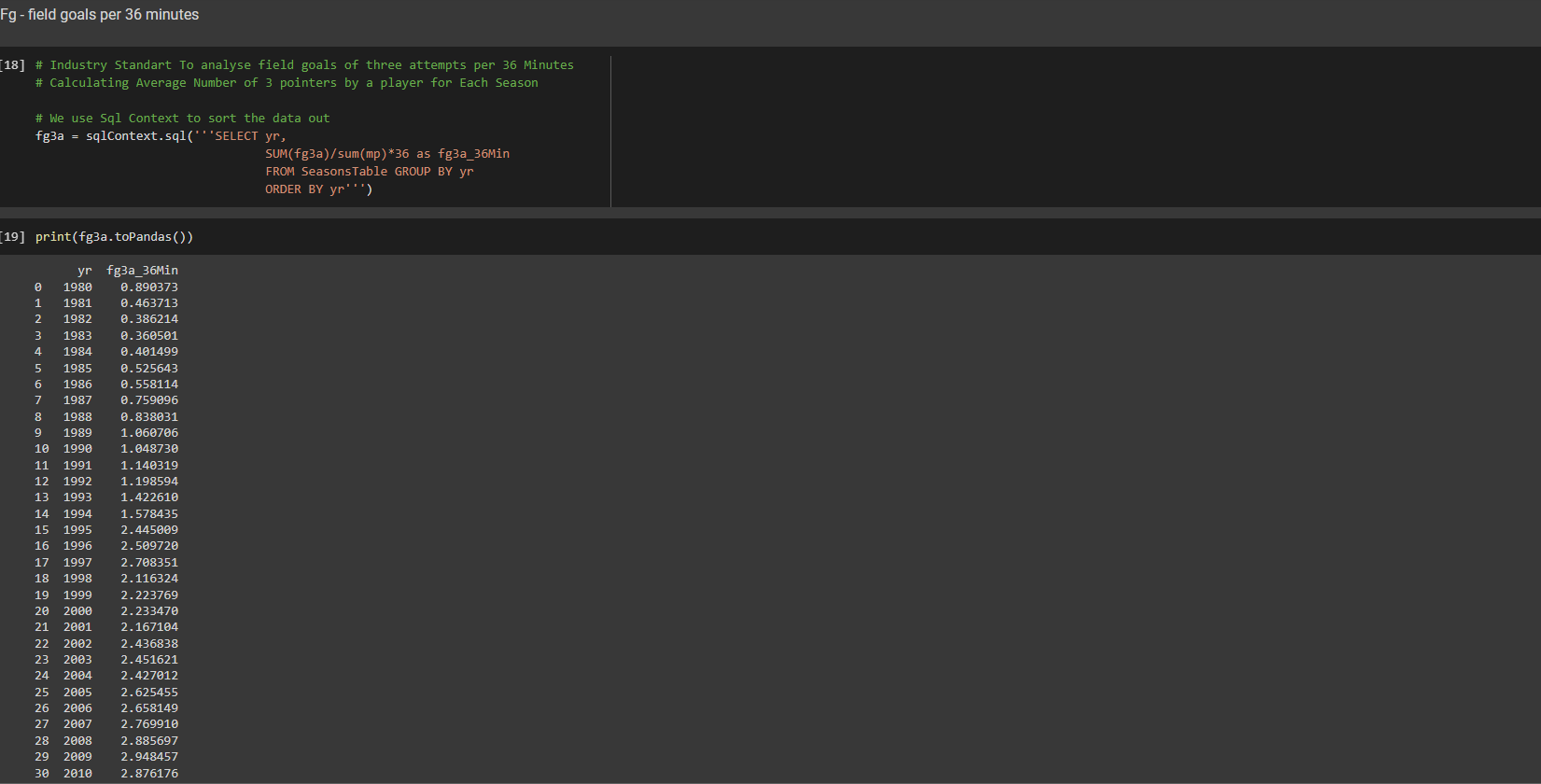


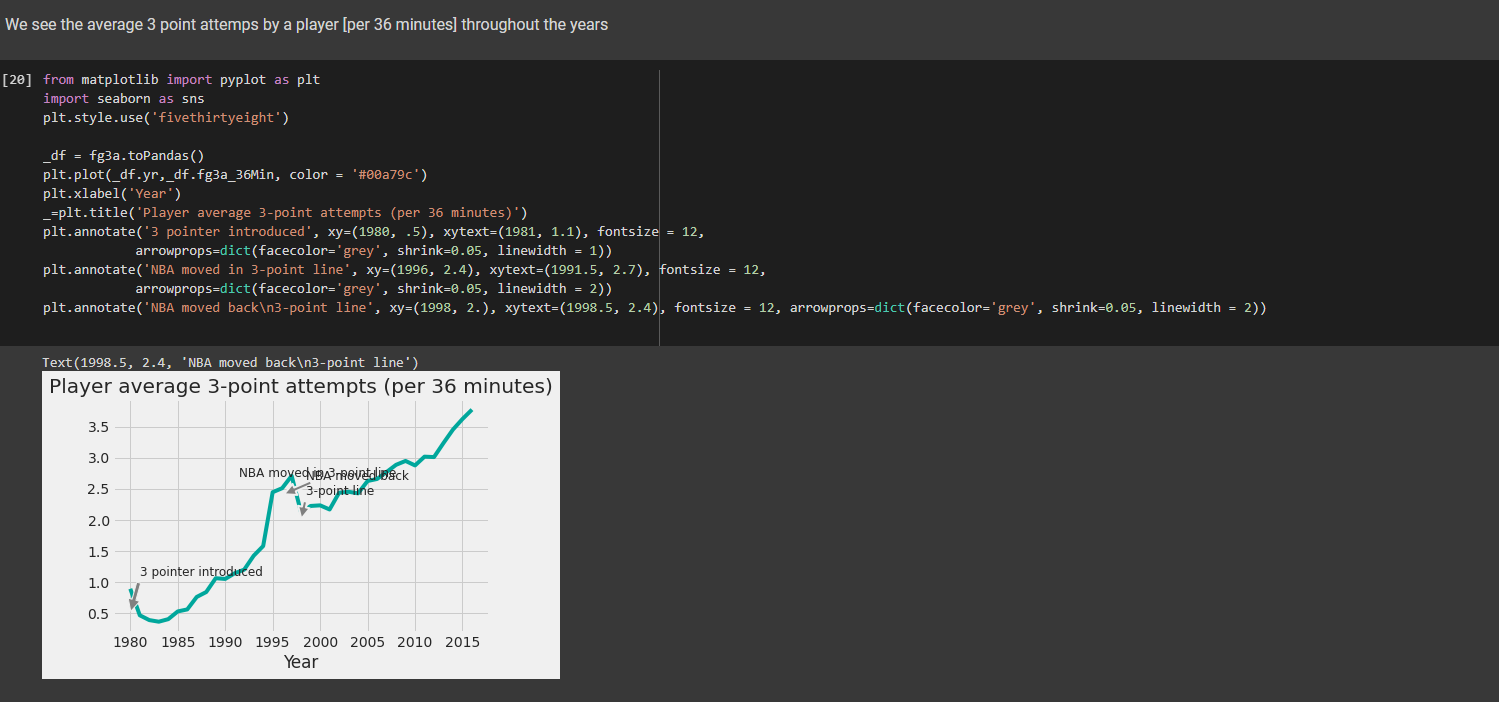


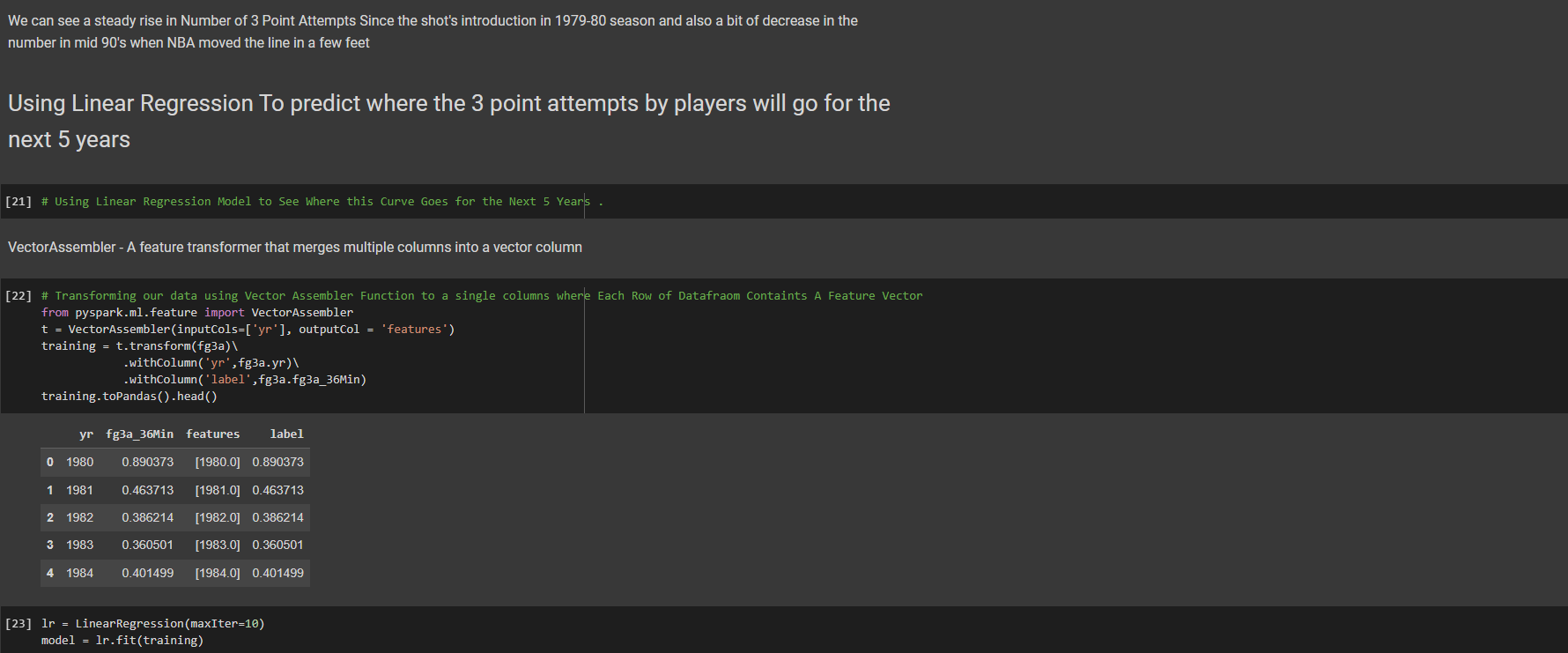


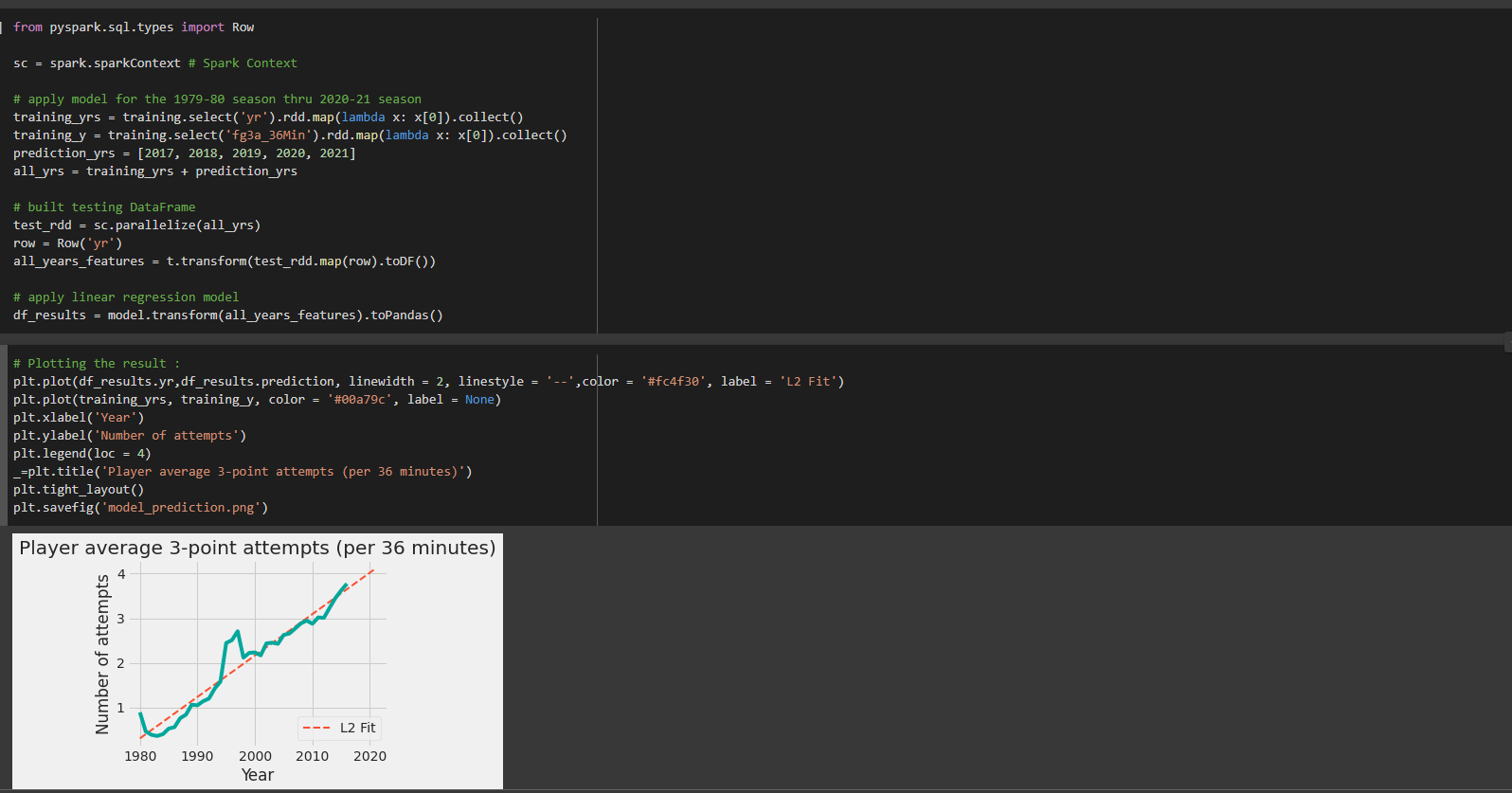
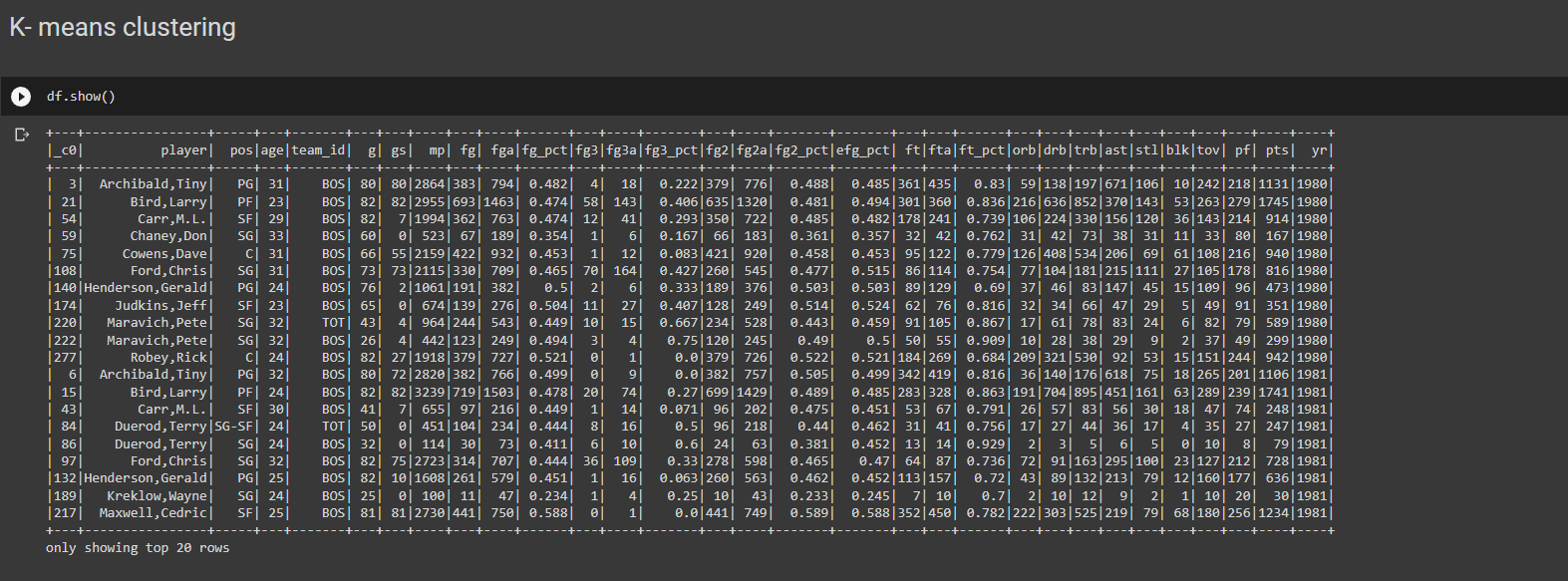


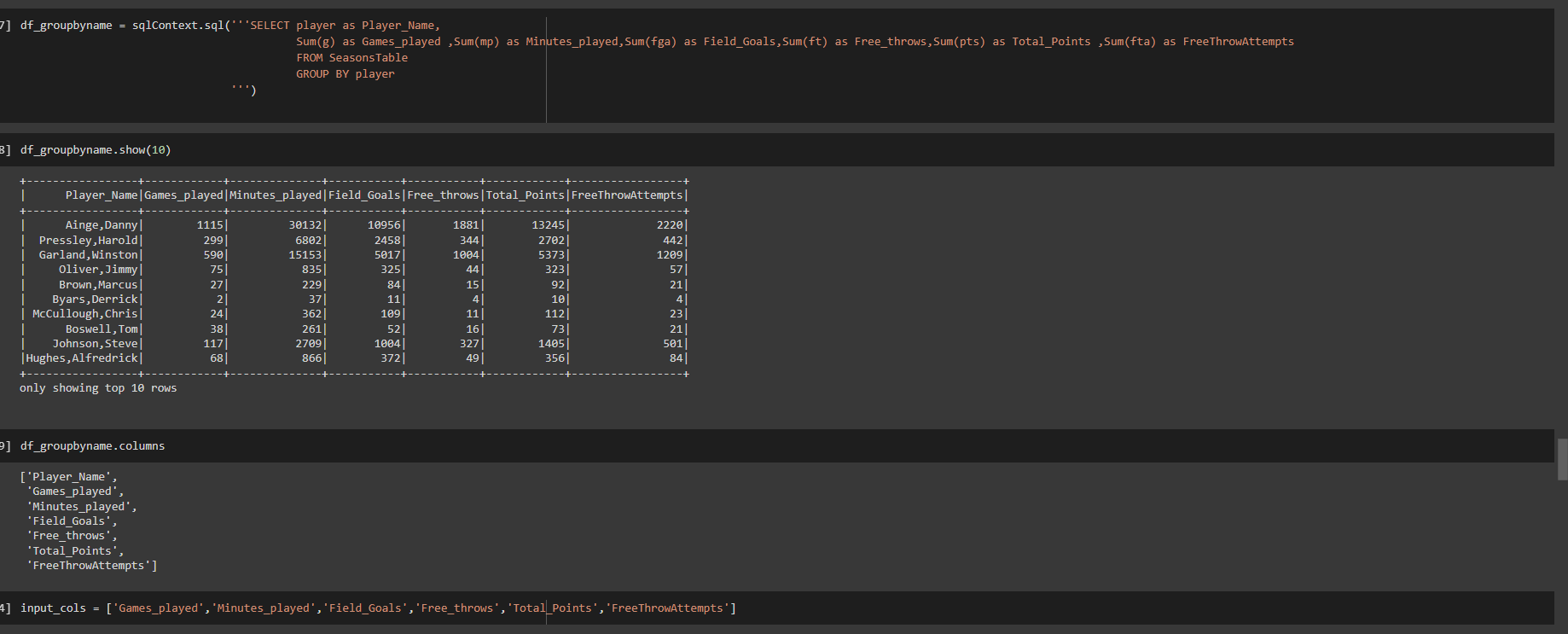


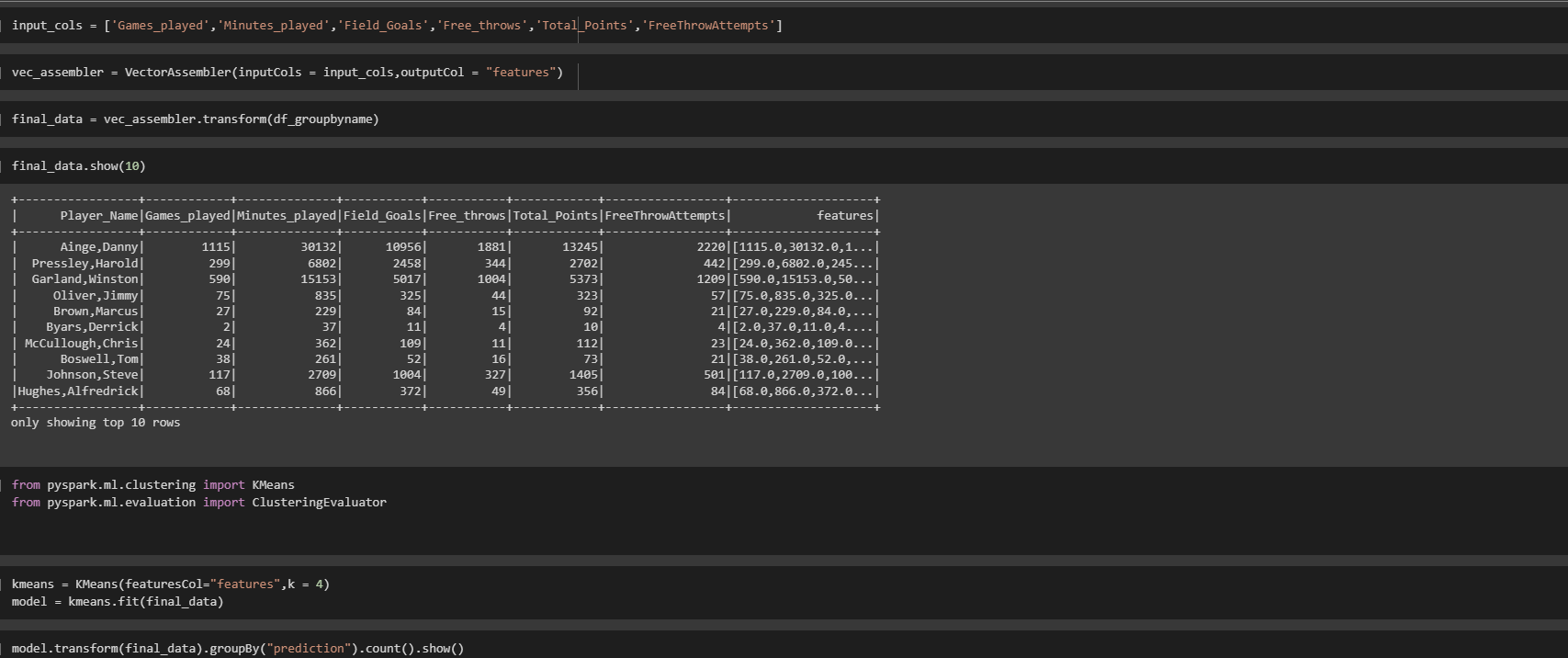


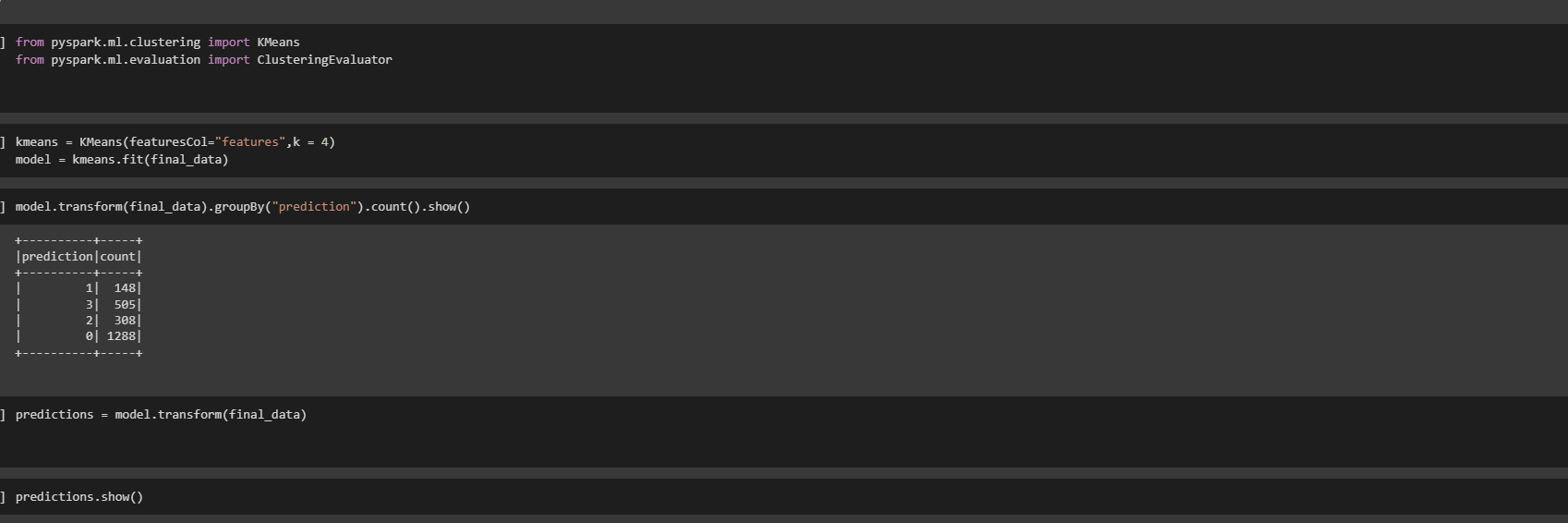


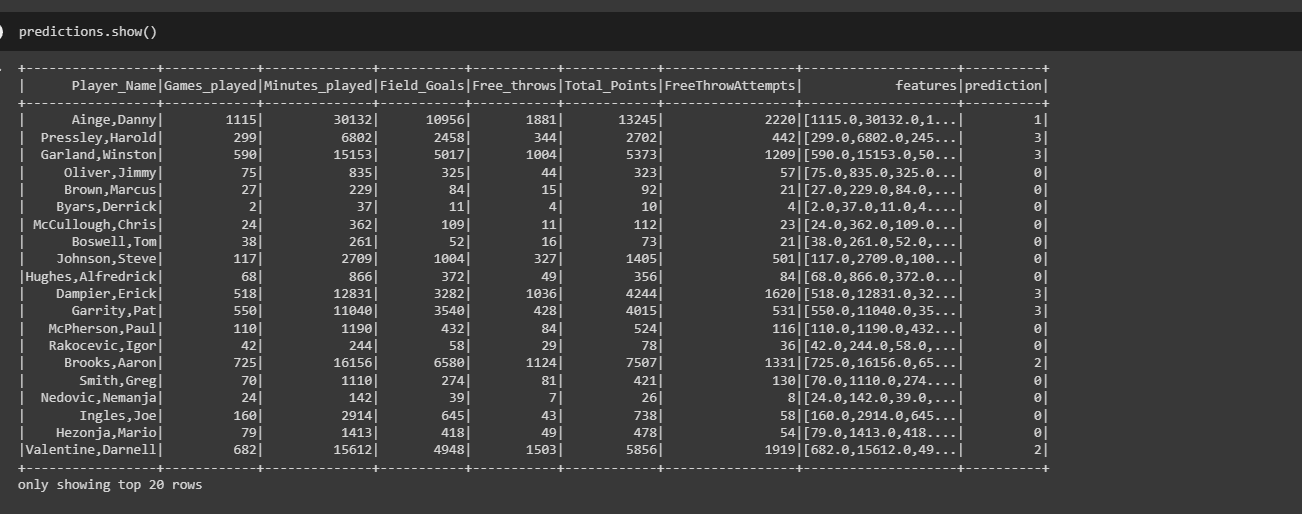










**6) CONCLUSION AND LEARNING**

Got to learn about the implementation of Apache Spark’s Sql and Mllib .

Used Sql to group the data and made a linear regression model using Apache Spark’s Mllib .

Predicted the distance the athletes will choose in the coming year with the linear regression model and concluded that in the coming years the distance of shots taken by players will increase .

Used k means clustering to make clusters and find similarity betwenn attributes of a player .