Paul Polsinelli

IFT 410 - 22485

3/21/2021

# Lab 8: Building Clustering Models

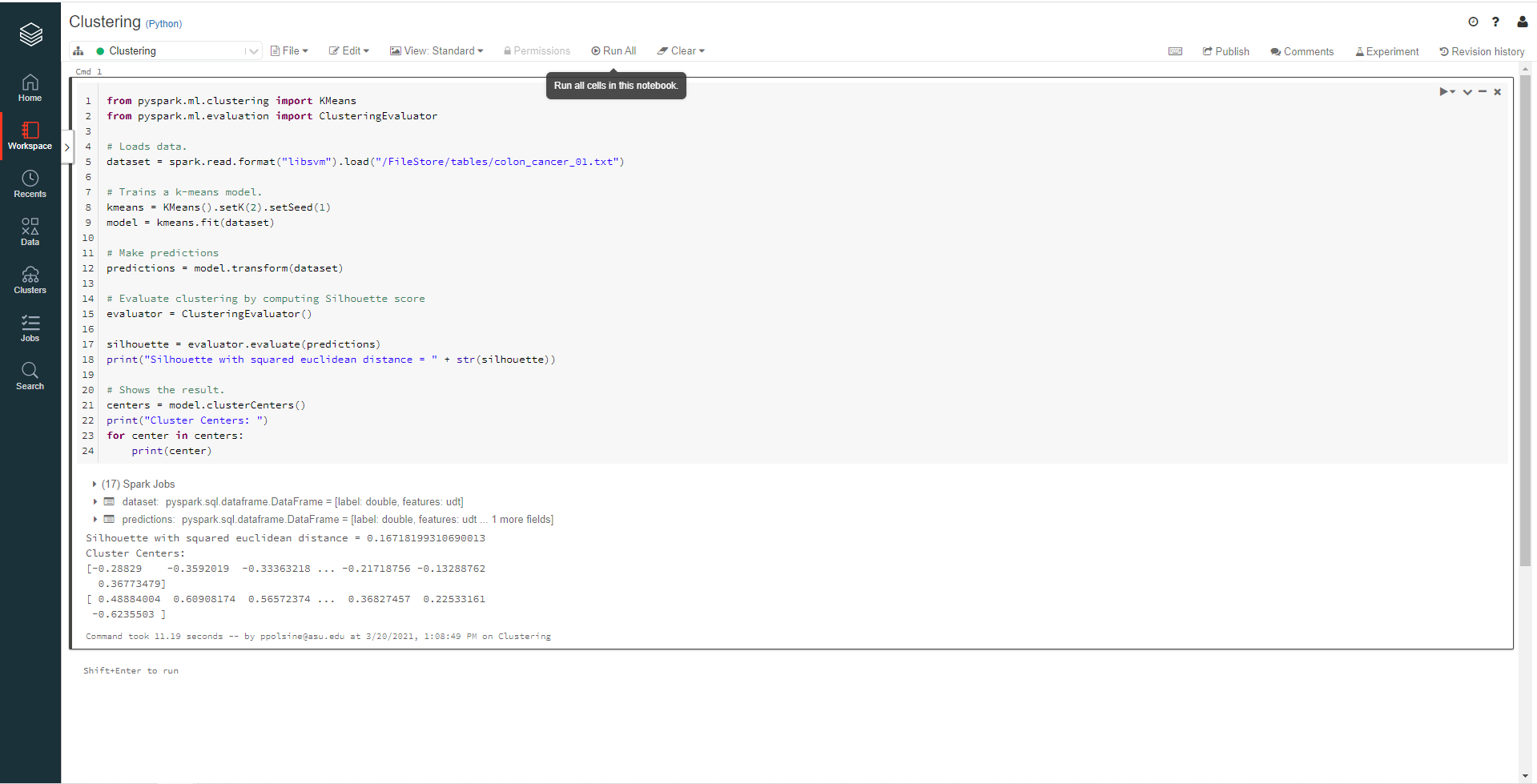
**Part I: Clustering Models with Spark’s ML Library**

**What is the Silhouette of the clustering that the code returned?** 0.16718199310690013

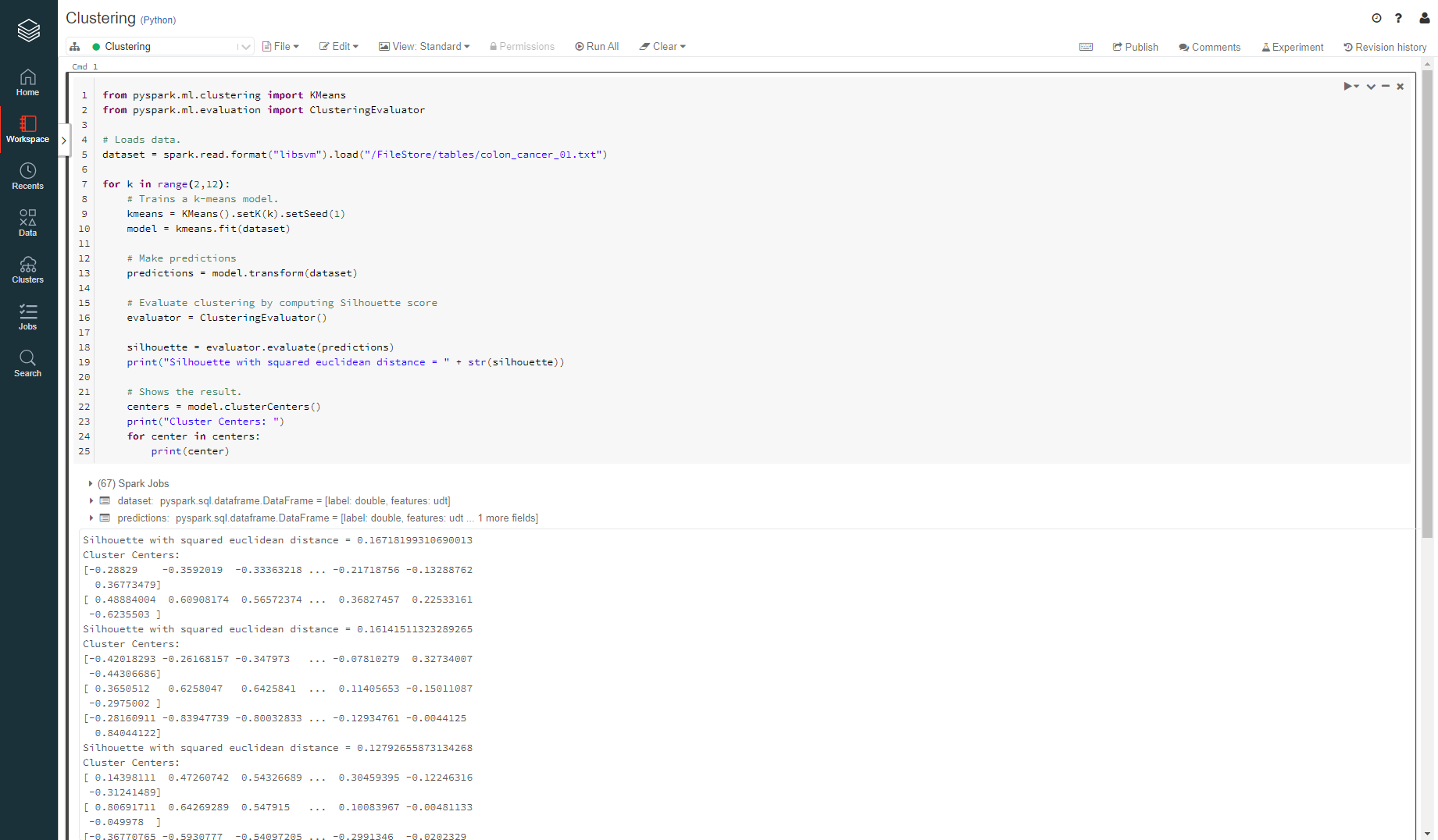
**Is it high or low?** It’s slightly on the high side but not high. On a scale of -1 to 1 it is just slightly above the midpoint of 0.

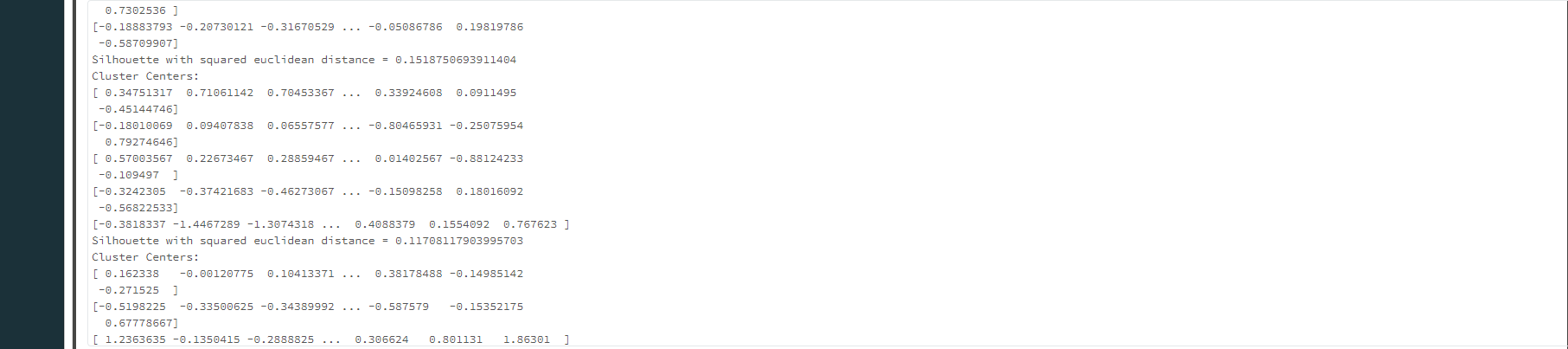
**What does the Silhouette value that you obtained indicates about the clustering?** The clustering may be slightly overlapping and is barely significant, if at all.

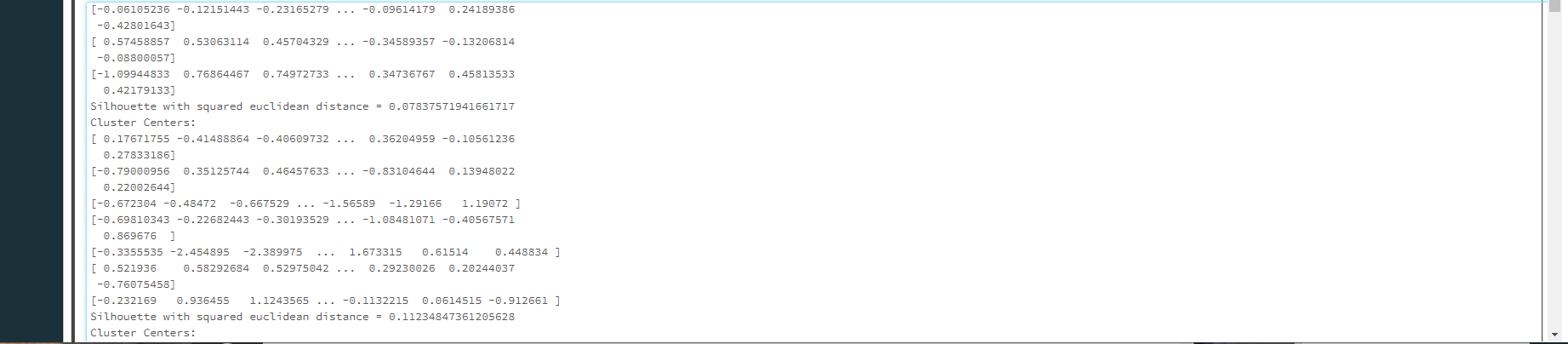
**Is the clustering good or bad?** Bad



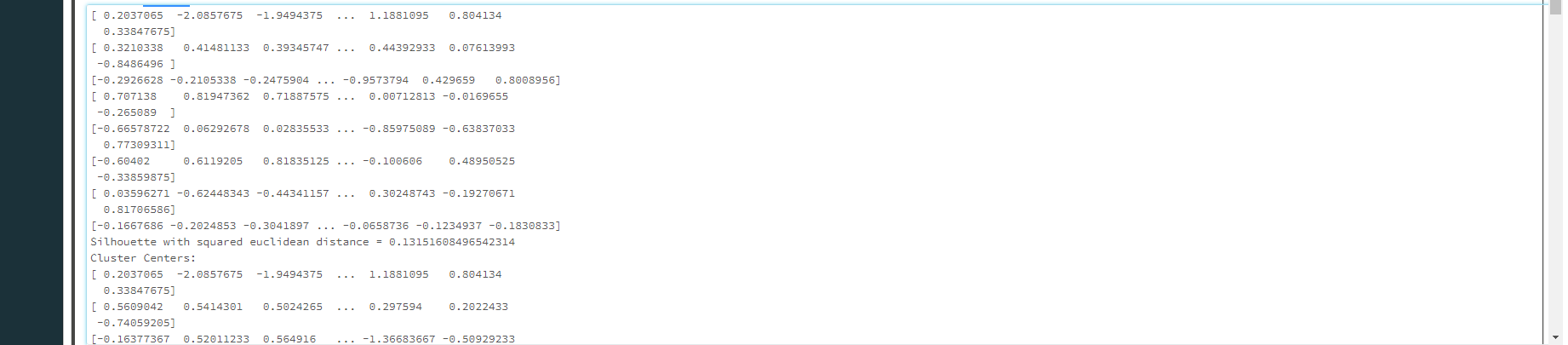
**Part II: Experimenting with Different Number of Clusters**



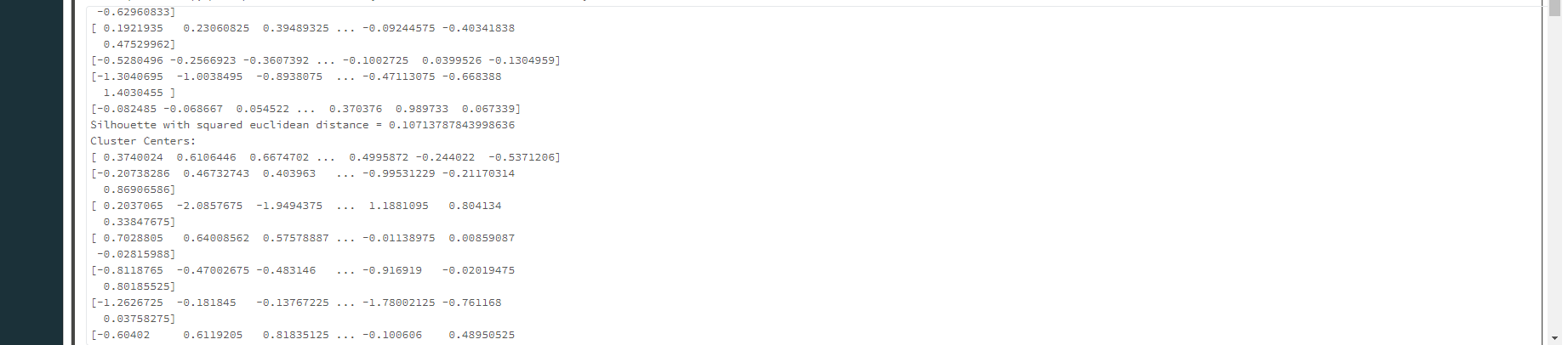












  
  
**Report the Silhouette values that you obtained with each run.**

**Run 1:** 0.16718199310690013

**Run 2:** 0.16141511323289265

**Run 3:** 0.12792655873134268

**Run 4:** 0.1518750693911404

**Run 5:** 0.11708117903995703

**Run 6:** 0.07837571941661717

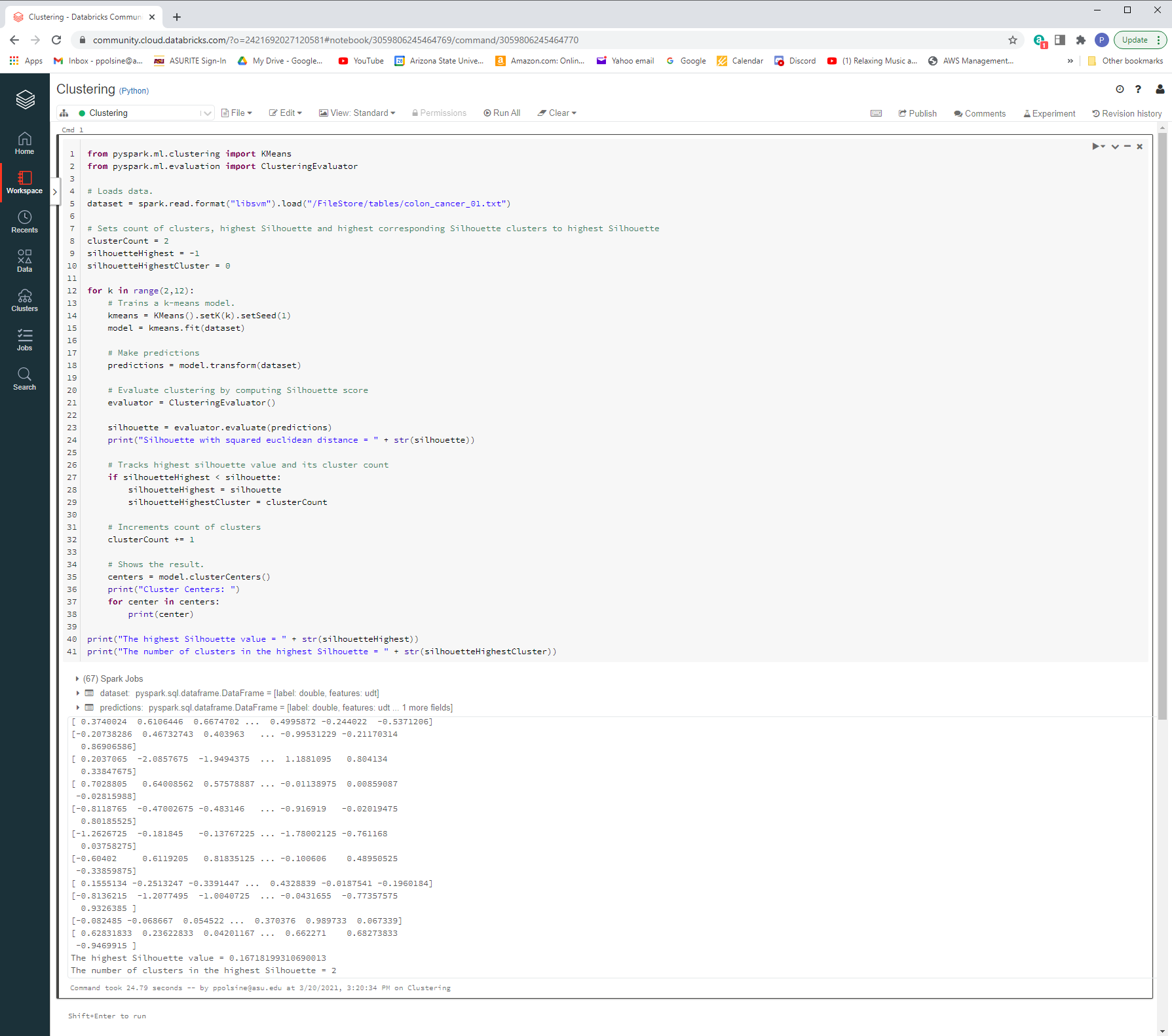
**Run 7:** 0.11234847361205628

**Run 8:** 0.13151608496542314

**Run 9:** 0.10326938350436614

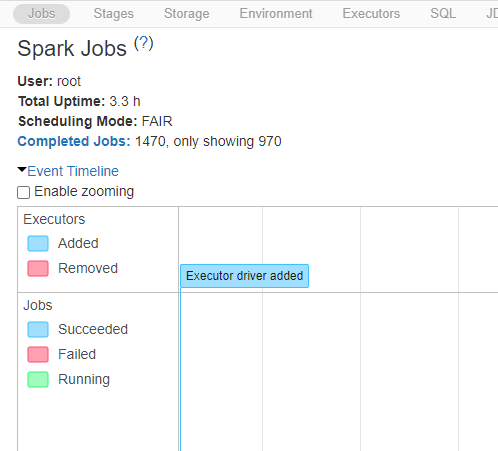
**Run 10:** 0.10713787843998636

Also modify the code so that it reports the highest Silhouette value and the number of clusters that gave you the highest Silhouette

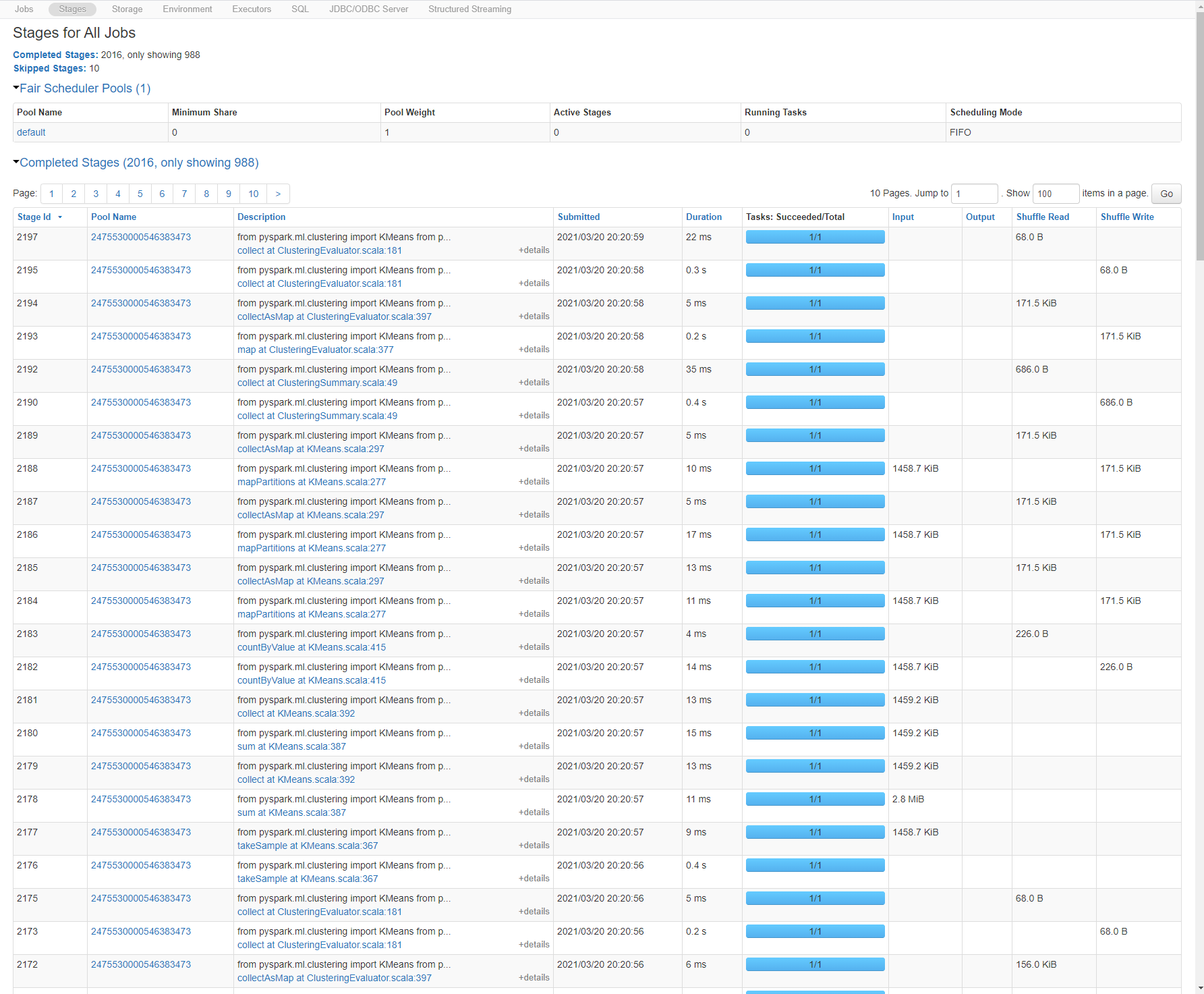


**Is the highest Silhouette value good or bad?** Bad

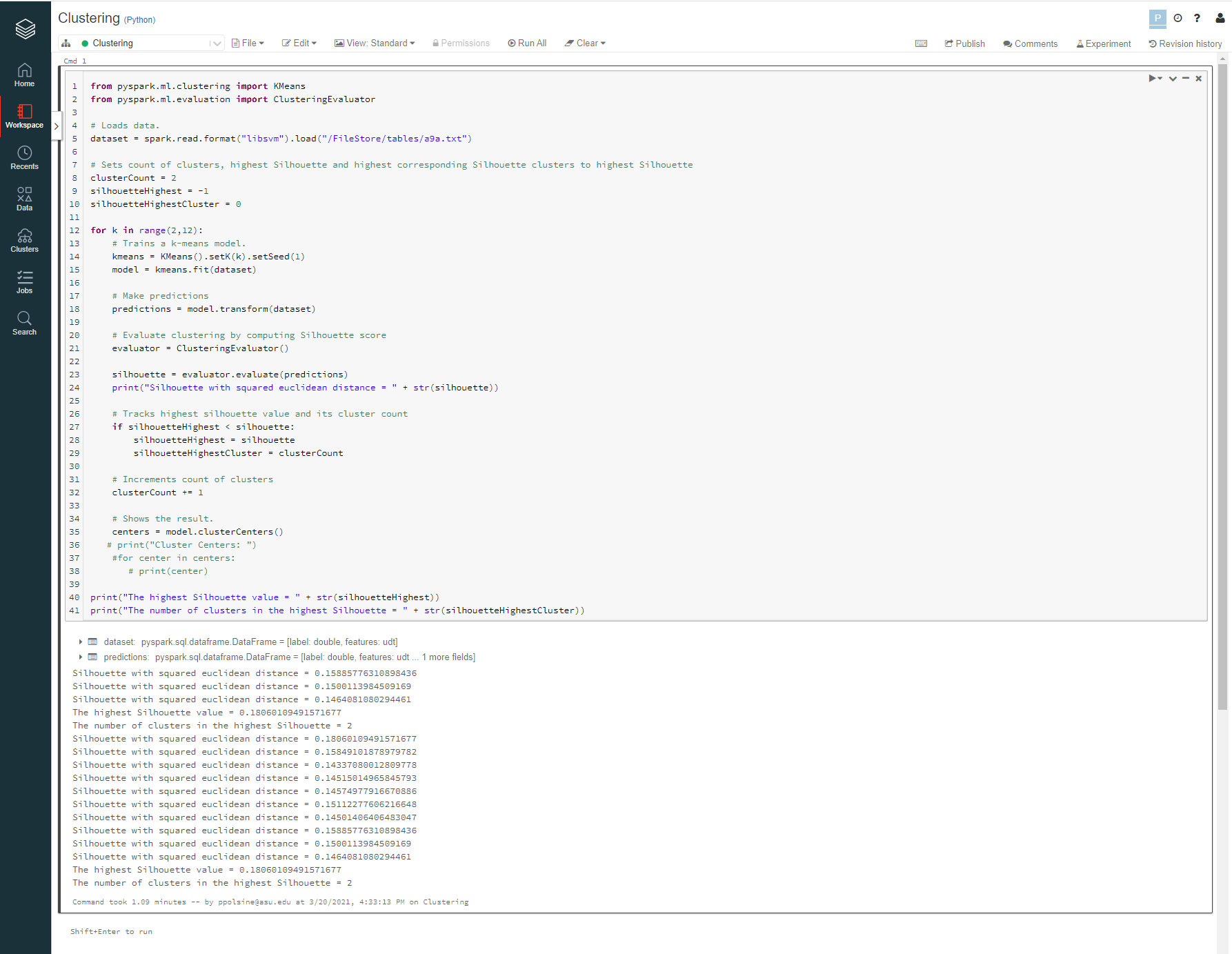
**How many jobs did Spark submit to run your code?** 1470



**How many blocks was the RDD holding the dataset divided into?** 2016



**Part III: Experimenting with a Larger Dataset**



**What is the highest Silhouette value?** 0.18060109491571677

**What are the number of associated clusters?** 2

**What do you notice about the obtained Silhouettes**? They’re all fairly close in value. None of them change what can be concluded by any of the others.

**What do you think lead to these results?** The model is a fundamentally reliable predictor of the clustering properties of the data no matter the k value.

**How much time did your code take to complete all clustering tasks?** 1.09 minutes