1.
K=2:

Clusterer output

```
Number of iterations: 8
Within cluster sum of squared errors: 29.740142242908494

Initial starting points (random):

Cluster 0: 8.5881,7236,1.59647,6.29522,1,0
Cluster 1: 10.5886,3726,0.555239,0.053534,1,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute        Full Data         0         1
                   (459.0)   (302.0)   (157.0)
================================================
TESSMagnitude       9.4371    8.7989   10.6648
StarTemp         5119.4114 5842.3427 3728.8048
StarRadius          0.9746    1.2291    0.4852
StarLuminosity      2.5168    3.7838    0.0795
NFluxRegions        1.0109    1.0166         1
MeanFluxSep      2707.3987 2234.6767 3616.7112




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      302 ( 66%)
```

```
Class attribute: class
Classes to Clusters:

   0   1  <-- assigned to cluster
 220 123 | not
  82  34 | toi

Cluster 0 <-- not
Cluster 1 <-- toi

Incorrectly clustered instances :       205.0    44.6623 %
```

K=3:

Clusterer output

```
Cluster 0: 8.5881,7236,1.59647,6.29522,1,0
Cluster 1: 10.5886,3726,0.555239,0.053534,1,0
Cluster 2: 8.4665,4329,0.674102,0.143779,2,41758.60461

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute         Full Data          0          1          2
                    (459.0)    (262.0)    (143.0)     (54.0)
============================================================
TESSMagnitude        9.4371      8.794    10.6888     9.2428
StarTemp          5119.4114  5850.1355  3714.0864  5295.5552
StarRadius           0.9746     1.1411     0.4768      1.485
StarLuminosity       2.5168     2.3217     0.0765     9.9252
NFluxRegions         1.0109      0.855     0.8881     2.0926
MeanFluxSep       2707.3987          0   142.3347  22635.966




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      262 ( 57%)
1      143 ( 31%)
2       54 ( 12%)
```

```
Class attribute: class
Classes to Clusters:

   0   1   2  <-- assigned to cluster
 196 112  35 | not
  66  31  19 | toi

Cluster 0 <-- not
Cluster 1 <-- toi
Cluster 2 <-- No class

Incorrectly clustered instances :       232.0    50.5447 %
```

K=4:

Clusterer output

```
Number of iterations: 14
Within cluster sum of squared errors: 19.795147100821023

Initial starting points (random):

Cluster 0: 8.5881,7236,1.59647,6.29522,1,0
Cluster 1: 10.5886,3726,0.555239,0.053534,1,0
Cluster 2: 8.4665,4329,0.674102,0.143779,2,41758.60461
Cluster 3: 9.1364,6213,1.1215,1.68848,1,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                            Cluster#
Attribute         Full Data          0          1          2          3
                    (459.0)    (129.0)    (123.0)     (53.0)    (154.0)
=================================================================================
TESSMagnitude        9.4371     7.9429    10.7884     9.2685     9.6676
StarTemp          5119.4114  6226.5584  3561.4537  5235.7166  5396.3118
StarRadius           0.9746     1.2964     0.4327     1.4824     0.9632
StarLuminosity       2.5168     3.5479     0.0454      9.881     1.0925
NFluxRegions         1.0109      0.969     0.9024     2.0943     0.7597
MeanFluxSep       2707.3987    76.4983   165.4785 22876.8657          0



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances
```

```
Class attribute: class
Classes to Clusters:

   0   1   2   3  <-- assigned to cluster
 111  98  34 100 | not
  18  25  19  54 | toi

Cluster 0 <-- not
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- toi

Incorrectly clustered instances :       294.0    64.0523 %
```

K=5:

Clusterer output

```
Number of iterations: 14
Within cluster sum of squared errors: 16.149769080380445

Initial starting points (random):

Cluster 0: 8.5881,7236,1.59647,6.29522,1,0
Cluster 1: 10.5886,3726,0.555239,0.053534,1,0
Cluster 2: 8.4665,4329,0.674102,0.143779,2,41758.60461
Cluster 3: 9.1364,6213,1.1215,1.68848,1,0
Cluster 4: 10.1358,3897,0.557752,0.074112,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                            Cluster#
Attribute         Full Data          0          1          2          3          4
                    (459.0)    (122.0)    (117.0)     (52.0)    (123.0)     (45.0)
=============================================================================================
TESSMagnitude        9.4371     7.9204    10.8237     9.3397     9.8796      8.847
StarTemp          5119.4114  6221.1534  3532.5538  5207.7835  5349.6107   5526.966
StarRadius           0.9746     1.2912     0.4201     1.4783     0.9577     1.0223
StarLuminosity       2.5168     3.5595     0.0389     9.9716     1.0743     1.4605
NFluxRegions         1.0109     1.0164     0.9231     2.0962          1          0
MeanFluxSep       2707.3987   201.4085   173.9646 23034.0447          0          0



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===
```

```
Class attribute: class
Classes to Clusters:

   0   1   2   3   4  <-- assigned to cluster
 105  94  33  77  34 | not
  17  23  19  46  11 | toi

Cluster 0 <-- not
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- toi
Cluster 4 <-- No class

Incorrectly clustered instances :       308.0    67.1024 %
```

Observations:

Ordered TESSMagnitude Coordinate of Centroids for each K (dataset mean = 9.44):

K=2:	K=3:	K=4:	K=5:
8.80	8.79	7.94	7.92
10.66	9.24	9.27	8.85
	10.69	9.67	9.34
		10.79	9.88
			10.82

Ordered StarTemp Coordinate of Centroids for each K (dataset mean = 5119):

K=2:	K=3:	K=4:	K=5:
3729	3714	3561	3536
5842	5296	5236	5208
	5850	5396	5350
		6227	5527
			6221

Ordered StarRadius Coordinate of Centroids for each K (dataset mean = 0.975):

K=2:	K=3:	K=4:	K=5:
0.485	0.477	0.432	0.420
1.229	1.141	0.963	0.958
	1.485	1.296	1.022
		1.482	1.291
			1.478

Ordered StarLuminosity Coordinate of Centroids for each K (dataset mean = 2.517):

K=2:	K=3:	K=4:	K=5:
0.080	0.087	0.045	0.039
3.784	2.322	1.093	1.074
	9.925	3.548	1.461
		9.881	3.560
			9.972

Ordered NFluxRegions Coordinate of Centroids for each K (dataset mean = 1.01):

K=2:	K=3:	K=4:	K=5:
1	0.86	0.76	0
1.02	0.88	0.90	1
	2.09	0.97	0.92
		2.09	1.02
			2.10

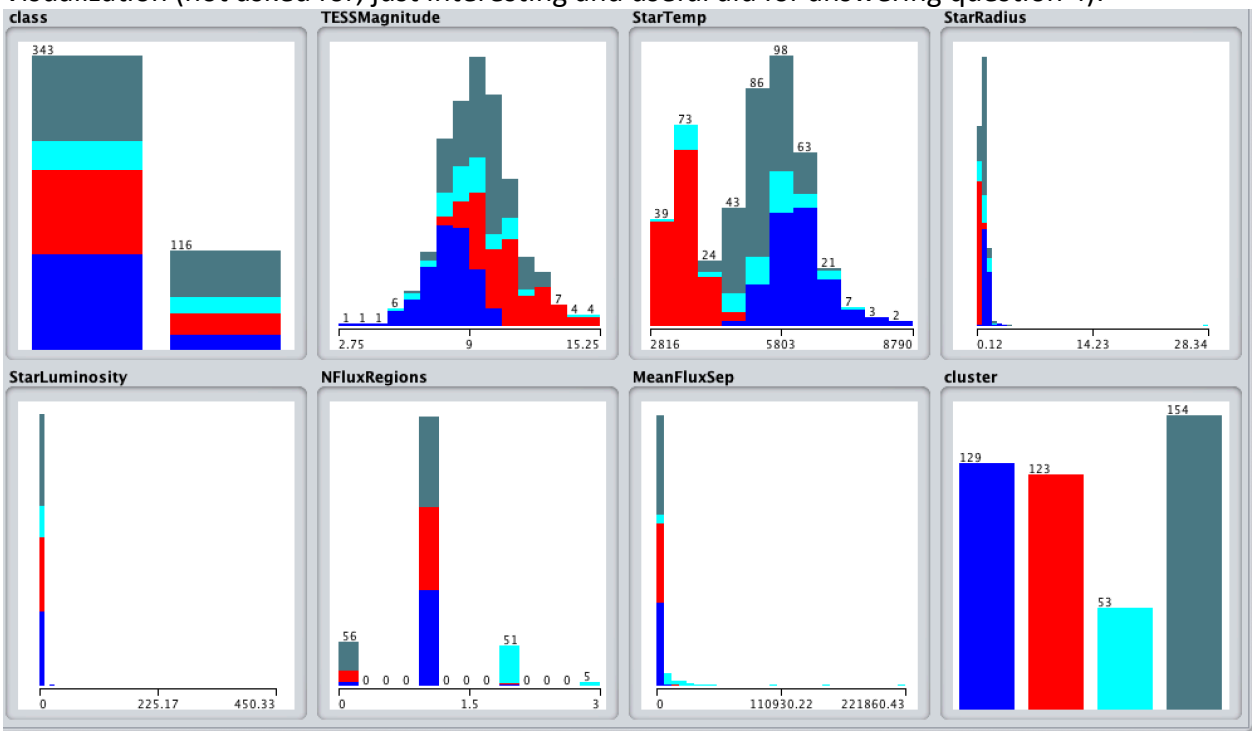
Ordered MeanFluxSep Coordinate of Centroids for each K (dataset mean = 2707):

K=2:	K=3:	K=4:	K=5:
2235	0	0	0
3617	142	76	0
	22636	165	174
		22877	201
			23034

Since this data set is fairly small (N=459) for K-means clustering to work well, there is likely to be high variability in optimal cluster centroids. That said, for certain features, as can be observed in the above tables, there are clear values which serve as the cluster centroids, irrespective of the number of clusters, indicating that there is likely a significance to these centroids.
For TESSMagnitude, ~8.8 and ~10.7 appear as cluster centroids in most (>=3) of the clustering trials, which are notably on either side of the dataset's mean for this feature. For StarTemp, ~5200K appears as a cluster centroid in most (>=3) of the clustering trials, which is notably close to the dataset's mean for this feature. For StarRadius, ~1.48, ~1.25, and ~0.45 appear as cluster centroids in most (>=3) of the clustering trials. For StarLuminosity, ~9.9 and ~3.5 appear as cluster centroids in most (>=3) of the clustering trials; the notably high value of 9 makes sense since this dataset contains an outlier with a value of ~28. For NFluxRegions, which takes integer values, ~2.1 (ie, >2) and values between 0 and 1 appear as cluster centroids in most (>=3) of the clustering trials, which especially makes sense for the trials with K>3 since this value only occupies 4 values (0,1,2,3) in the data and thus clustering with >=3 clusters incentivizes the algorithm to simply create a cluster for each of these integer domains; it's also worth noting that as the number of clusters increases, these centroid values tend towards the nearest integer rather than being between integers, increasing the goodness of fit. For MaxFluxSep, ~22,000 and ~200 appear as cluster centroids in most (>=3) of the clustering trials.

Based on informal evaluation provided by "Classes to cluster evaluation", performance increases noticeably as number of clusters increases; however, the performance is pretty between K=4 and K=5 (64% vs 67% correct) and, as noted above, having fewer clusters (ideally <=3) will help avoid overfitting NFluxRegions. So, the most sensible value for K is probably K=4.

2. Using K=4, ignoring index 1 (class).
Visualization (not asked for, just interesting and useful aid for answering question 4):



Blue: Cluster 0
Red: Cluster 1
Cyan: Cluster 2
Grey: Cluster 3

3.

Cluster 0:

	Feature Values					
	TESSMag	StarTemp	StarRadius	StarLuminosity	NFluxReg	MeanFluxSep
Cluster 0 Centroid	7.9429	6227	1.2964	3.5479	0.969	76.4983
Random Cluster 0 Instance 1	8.9854	6055	1.13269	1.55371	1	0
Random Cluster 0 Instance 2	8.1505	6087	1.26855	1.99031	1	0
Random Cluster 0 Instance 3	7.3212	5326	2.78715	6.45671	1	0

Instance 1 Euclidean distance was 187.9, for instance 2 it was 159.2, for instance 3 it was 903.8. Accordingly, these instances occupy similar values to the centroid for all features. Besides MeanFluxSep where they all occupy a value of 0, their features differ most in StarLuminosity for instance 1 and 2 (56% and 44% error) and in StarRadius for instance 3 (115% error). This comparison would have been much easier if the features had been first normalized or standardized.

Cluster 1:

	Feature Values					
	TESSMag	StarTemp	StarRadius	StarLuminosity	NFluxReg	MeanFluxSep
Cluster 1 Centroid	10.7884	3561	0.4327	0.0454	0.9024	165.4785
Random Cluster 1 Instance 1	10.6925	4317	0.651075	0.132606	1	0
Random Cluster 1 Instance 2	11.0382	3543	0.58803	0.049088	1	0
Random Cluster 1 Instance 3	13.5144	3120	0.195471	0.003262	1	0

Instance 1 Euclidean distance was 773.2, for instance 2 it was 166.5, for instance 3 it was 471.5. Accordingly, these instances occupy similar values to the centroid for all features. Besides MeanFluxSep where they all occupy a value of 0, their features differ most in StarLuminosity for instance 1 and 3 (192% and 92% error) and in StarRadius for instance 2 (36% error). This comparison would have been much easier if the features had been first normalized or standardized.

Cluster 2:

	Feature Values					
	TESSMag	StarTemp	StarRadius	StarLuminosity	NFluxReg	MeanFluxSep
Cluster 2 Centroid	9.2685	5236	1.4824	9.881	2.0943	22876.8657
Random Cluster 2 Instance 1	8.34398	3544	0.454147	0.029313	2	15191.749
Random Cluster 2 Instance 2	8.7542	6352	1.41788	3.380702	2	6604.22048

Random Cluster 2 Instance 3	9.78263	3519	0.488526	0.032972	2	12452.5028

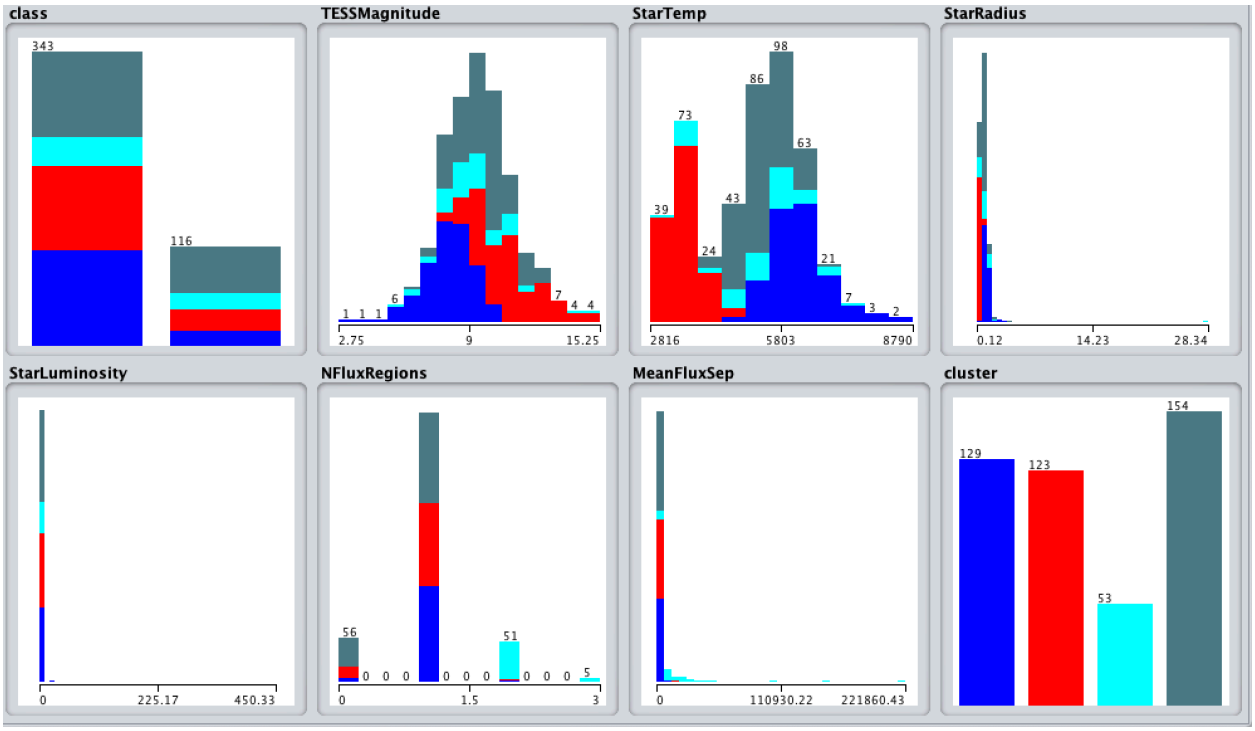
Instance 1 Euclidean distance was 7869, for instance 2 it was 16311, for instance 3 it was 10565. Accordingly, these instances occupy similar values to the centroid for all features. Their features differ most in StarLuminosity for instance 1 and 3 (99.7% and 99.7% error respectively) and in MeanFluxSep for instance 2 (71.1% error). The significant error in MeanFluxSep was likely driven by the presence of a singular extreme outlier with respect to MeanFluxSep in this dataset, which should likely have been pruned. Additionally, this comparison would have been much easier if the features had been first normalized or standardized.

Cluster 3:

	Feature Values					
	TESSMag	StarTemp	StarRadius	StarLuminosity	NFluxReg	MeanFluxSep
Cluster 3 Centroid	9.6676	5396	0.9632	1.0925	0.7597	0
Random Cluster 3 Instance 1	10.0391	5180	0.878413	0.500508	1	0
Random Cluster 3 Instance 2	8.1776	5093	0.695319	0.293061	0	0
Random Cluster 3 Instance 3	8.4818	5338	0.880582	0.566264	1	0

Instance 1 Euclidean distance was 216.3, for instance 2 it was 303.3, for instance 3 it was 60.6. Accordingly, these instances occupy similar values to the centroid for all features. Besides MeanFluxSep where they all occupy a value of 0, their features differ most in StarLuminosity for instance 1, 2, and 3 (54%, 73%, and 48% error respectively). This comparison would have been much easier if the features had been first normalized or standardized.

4.



Blue: Cluster 0
Red: Cluster 1
Cyan: Cluster 2
Grey: Cluster 3

Informally, based on the analysis in response to question 3 and the plots given after question 2 above, it seems that:

- Cluster 0 is fairly homogenous. Cluster 0 captures instances which likely fall into the "Not TOI" class, having 0 or 1 Flux Region (indicating potential transit event), with stars of moderate magnitude (brightness), surface temperatures (~6500-7000K) and radii greater (1-2 solar radii) which correspond to stars in the upper main sequence (F-type perhaps).
- Cluster 1 is fairly homogenous. Cluster 1 captures instances which likely fall into the "Not TOI" class but also captures the second widest chunk of those classified as "TOI", having 1 Flux Region (indicating potential transit event), with stars of high magnitude (brightness), low surface temperatures (in the domain of red dwarfs), and low radii (also in the domain of red dwarfs).
- Cluster 2 is not very homogenous. Cluster 2 captures instances which fall into either class (obviously still being skewed towards "Not TOI" since there are more instances of that class in the dataset), having >1 Flux Region (indicating multiple types of potential transit events). Beyond that, this cluster seems to fairly uniformly sample from the space of stellar parameters, with a slight preference towards those with low magnitude.

- Cluster 3 is fairly homogenous. Cluster 3 captures the widest chunk of instances classified as “TOI”, having 1 Flux Region (indicating potential transit event), with stars of moderate magnitude (brightness), surface temperatures near that of main sequence stars like the sun (~6000K), and radii round that of the sun (~1 solar radius).

Based on these observations, the following descriptions can be inferred for what types of stars each cluster selects:

- Cluster 0: Upper main sequence stars which are not likely members of the TOI class.
- Cluster 1: Red dwarfs with potential transit events, possibly members of the TOI class.
- Cluster 2: Stars with transit events of multiple depths (eg. due to dust clouds), for which TOI membership cannot be easily determined.
- Cluster 3: Main sequence stars with potential transit events, which might be members of the TOI class.