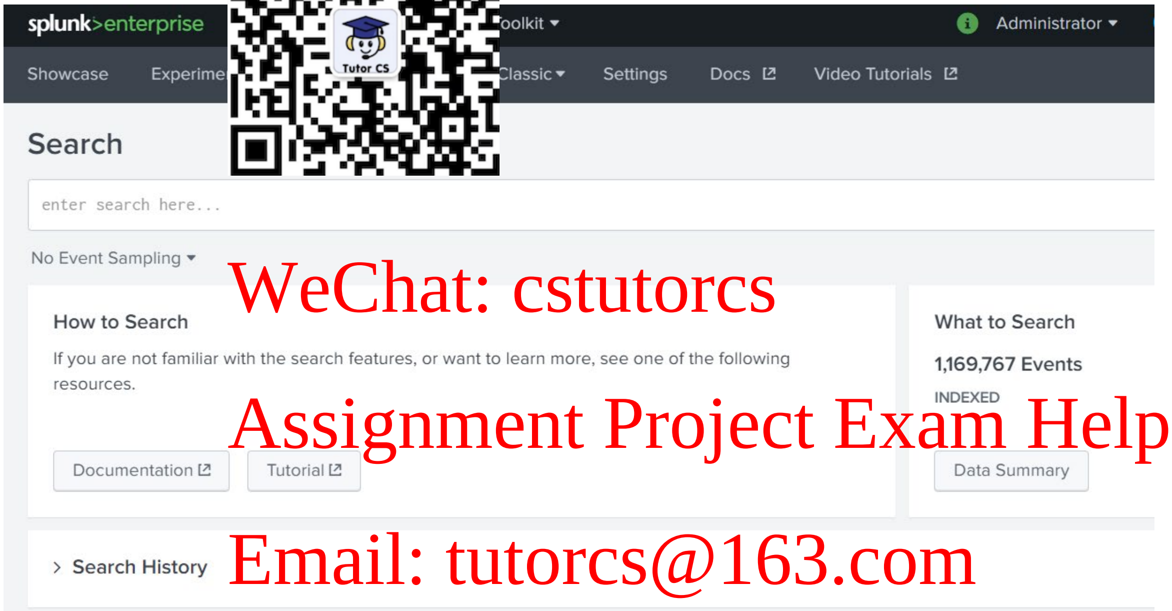
Step 0: Where to put my data (CSV) files?

```
$SPLUNK_HOME\etc\apps\Splunk_ML_Toolkit\lookups
```

Step 1: Load the file into MLTK

Go to App: **Splunk Machine Learning Toolkit -> Search**



How to Search

If you are not familiar with the search features, or want to learn more, see one of the following resources.

What to Search

1,169,767 Events

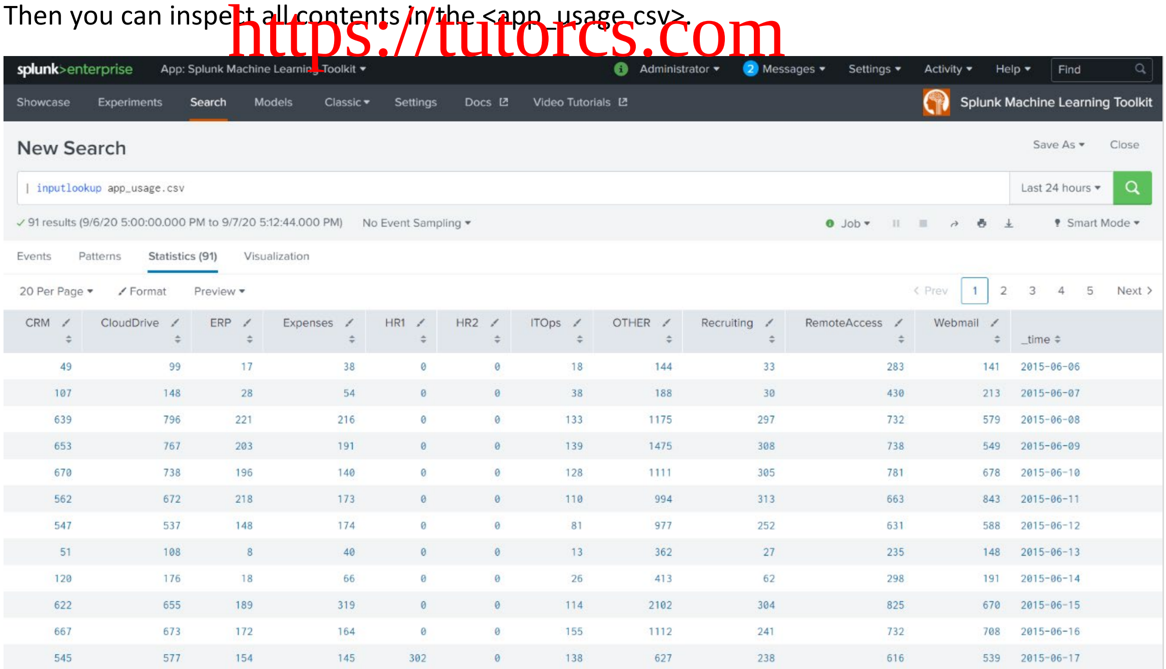
INDEXED

Then type-in the following command to load the file <app_usage.csv> (included in MLTK):

```
| inputlookup app_usage.csv    // load the file <app_usage.csv> into MLTK
```

Note: The bar (|) is necessary to add before the command!

Then you can inspect all contents in the <app_usage.csv>.



```
| inputlookup app_usage.csv
```

✓ 91 results (9/6/20 5:00:00.000 PM to 9/7/20 5:12:44.000 PM)

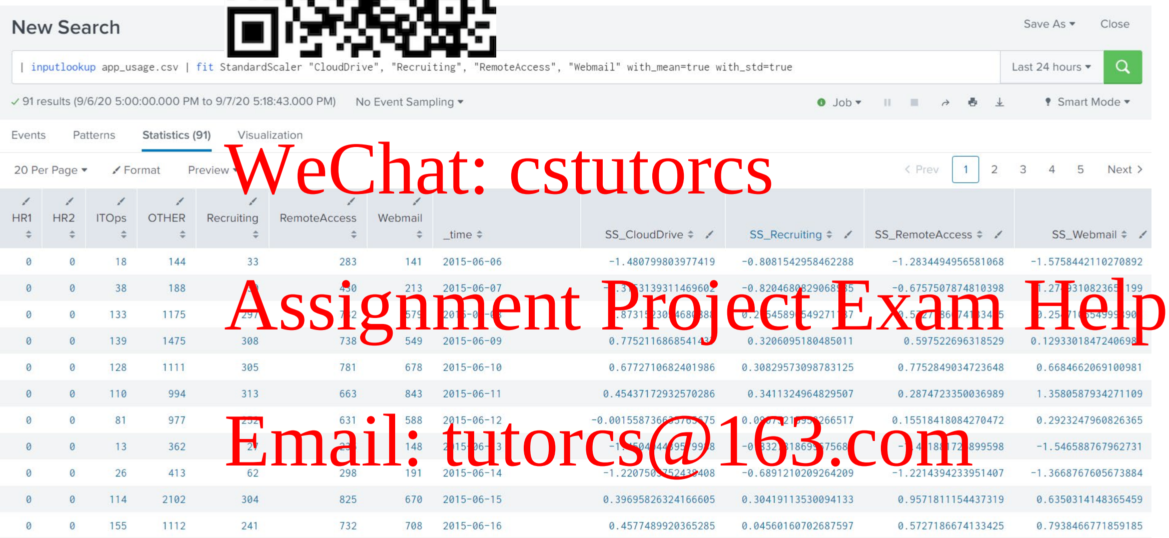
CRM	CloudDrive	ERP	Expenses	HR1	HR2	ITOps	OTHER	Recruiting	RemoteAccess	Webmail	_time
49	99	17	38	0	0	18	144	33	283	141	2015-06-06
107	148	28	54	0	0	38	188	30	430	213	2015-06-07
639	796	221	216	0	0	133	1175	297	732	579	2015-06-08
653	767	203	191	0	0	139	1475	308	738	549	2015-06-09
670	738	196	140	0	0	128	1111	305	781	678	2015-06-10
562	672	218	173	0	0	110	994	313	663	843	2015-06-11
547	537	148	174	0	0	81	977	252	631	588	2015-06-12
51	108	8	40	0	0	13	362	27	235	148	2015-06-13
120	176	18	66	0	0	26	413	62	298	191	2015-06-14
622	655	189	319	0	0	114	2102	304	825	670	2015-06-15
667	673	172	164	0	0	155	1112	241	732	708	2015-06-16
545	577	154	145	302	0	138	627	238	616	539	2015-06-17

Step 2: Apply preprocessing steps

In this example, we apply `StandardScaler` to the 4 features (aka. columns, fields, etc.) in <app_usage.csv>: "CloudDrive", "Recruiting", "RemoteAccess", "Webmail". Also, we would like to make the final scaled features to fall under *N*(0, 1) [Normal distribution with $\mu = 0$, $\sigma^2 = 1$].

Command: (attach [illegible], separated by bars "|")

```
| fit StandardS[illegible] "Recruiting", "RemoteAccess", "Webmail"
with_mean=true [illegible]/ apply preprocessing steps
```



```
| inputlookup app_usage.csv | fit StandardScaler "CloudDrive", "Recruiting", "RemoteAccess", "Webmail" with_mean=true with_std=true
```

✓ 91 results (9/6/20 5:00:00.000 PM to 9/7/20 5:18:43.000 PM)

HR1	HR2	ITOps	OTHER	Recruiting	RemoteAccess	Webmail	_time	SS_CloudDrive	SS_Recruiting	SS_RemoteAccess	SS_Webmail
0	0	18	144	33	283	141	2015-06-06	-1.480799803977419	-0.8081542958462288	-1.2834494956581068	-1.5758442110270892
0	0	38	188	[illegible]	430	213	2015-06-07	[illegible]	[illegible]	-0.6757507874810398	[illegible]
0	0	133	1175	297	[illegible]	579	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0	0	139	1475	308	738	549	2015-06-09	[illegible]	0.3206095180485011	0.597522696318529	[illegible]
0	0	128	1111	305	781	678	2015-06-10	0.6772710682401986	0.30829573098783125	0.7752849034723648	0.6684662069100981
0	0	110	994	313	663	843	2015-06-11	0.45437172932570286	0.3411324964829507	0.2874723350036989	1.3580587934271109
0	0	81	977	[illegible]	631	588	2015-06-12	[illegible]	[illegible]	0.15518418084270472	0.2923247960826365
0	0	13	362	[illegible]	[illegible]	148	[illegible]	[illegible]	[illegible]	[illegible]	-1.546588767962731
0	0	26	413	62	298	191	2015-06-14	[illegible]	-0.6891210209264209	-1.2214394233951407	-1.3668767605673884
0	0	114	2102	304	825	670	2015-06-15	0.39695826324166605	0.30419113530094133	0.9571811154437319	0.6350314148365459
0	0	155	1112	241	732	708	2015-06-16	0.4577489920365285	0.04560160702687597	0.5727186674133425	0.7938466771859185

After preprocessing, the processed fields will have "SS_" as the prefix.

Here `with_mean=true with_std=true` makes the final scaled features to fall under *N*(0, 1) [Normal distribution with $\mu = 0$, $\sigma^2 = 1$].

https://docs.splunk.com/Documentation/MLApp/5.2.0/User/Preprocessing

The top 5 commonly applied preprocessing methods/functions are:

1. FieldSelector
   https://docs.splunk.com/Documentation/MLApp/latest/User/Algorithms#FieldSelector
2. KernelPCA
   https://docs.splunk.com/Documentation/MLApp/latest/User/Algorithms#KernelPCA
3. PCA
   https://docs.splunk.com/Documentation/MLApp/latest/User/Algorithms#PCA
4. StandardScaler
   https://docs.splunk.com/Documentation/MLApp/latest/User/Algorithms#StandardScaler
5. TFIDF
   https://docs.splunk.com/Documentation/MLApp/latest/User/Algorithms#TFIDF

Step 3: Apply clustering algorithm

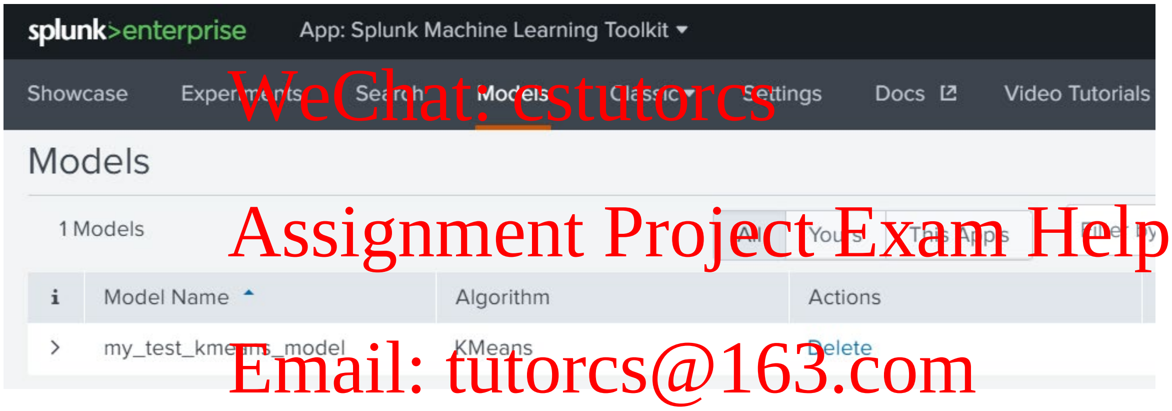
In this example, we apply KMeans to cluster those 4 features: "CloudDrive", "Recruiting", "RemoteAccess", "Webmail". In this case, we choose k=3 as the parameter.

Command: (attach [illegible] ommands, separated by bars "|")

```
| fit KMeans k=[illegible] "SS_Recruiting" "SS_RemoteAccess"
"SS_Webmail" in[illegible] model"      // train the KMeans model
```

Then, after the trai[illegible], you will see the model "my_test_kmeans_model" in:

App: **Splunk Machi[illegible] Models**



Models

1 Models

i	Model Name ▲	Algorithm	Actions
>	my_test_kmeans_model	KMeans	Delete

Step 4: Evaluate and visualize the result

When the KMeans model training is finished, we can apply the model to yield the cluster information and visualize them. Overall, the full SPL command will be:

```
| inputlookup app_usage.csv
```

```
| fit StandardScaler "CloudDrive", "Recruiting", "RemoteAccess", "Webmail"
with_mean=true with_std=true
```

```
| apply "my_test_kmeans_model"
```

```
| eval cluster= "Cluster: " + cluster
```

```
| table cluster, "SS_CloudDrive", "SS_Recruiting", "SS_RemoteAccess",
"SS_Webmail"          // display the selected variables
```

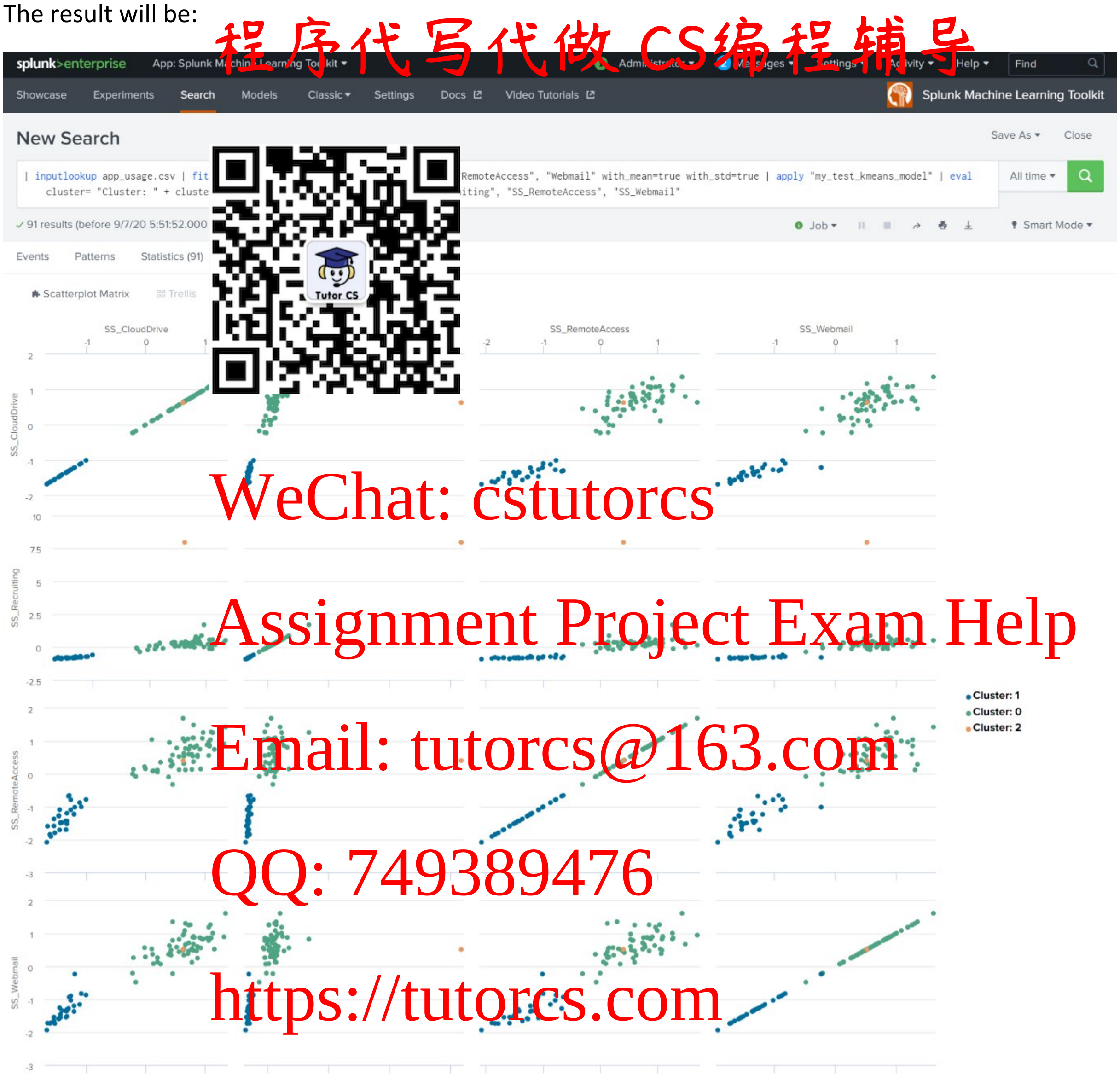
Optionally, in preprocessing (Step 2), you can use

```
| fit StandardScaler "CloudDrive", "Recruiting", "RemoteAccess", "Webmail"
with_mean=true with_std=true into "app_usage_SS"
```

Then, replace the second command by:

```
| apply "app_usage_SS"
```

The result will be:



```
| inputlookup app_usage.csv | fit                RemoteAccess", "Webmail" with_mean=true with_std=true | apply "my_test_kmeans_model" | eval
   cluster= "Cluster: " + cluste               iting", "SS_RemoteAccess", "SS_Webmail"
```

Here, we have 16 plots with field-field information showing all 3 clusters.

Note: In DBSCAN, Cluster -1 contains all the outliers (aka. anomalies).