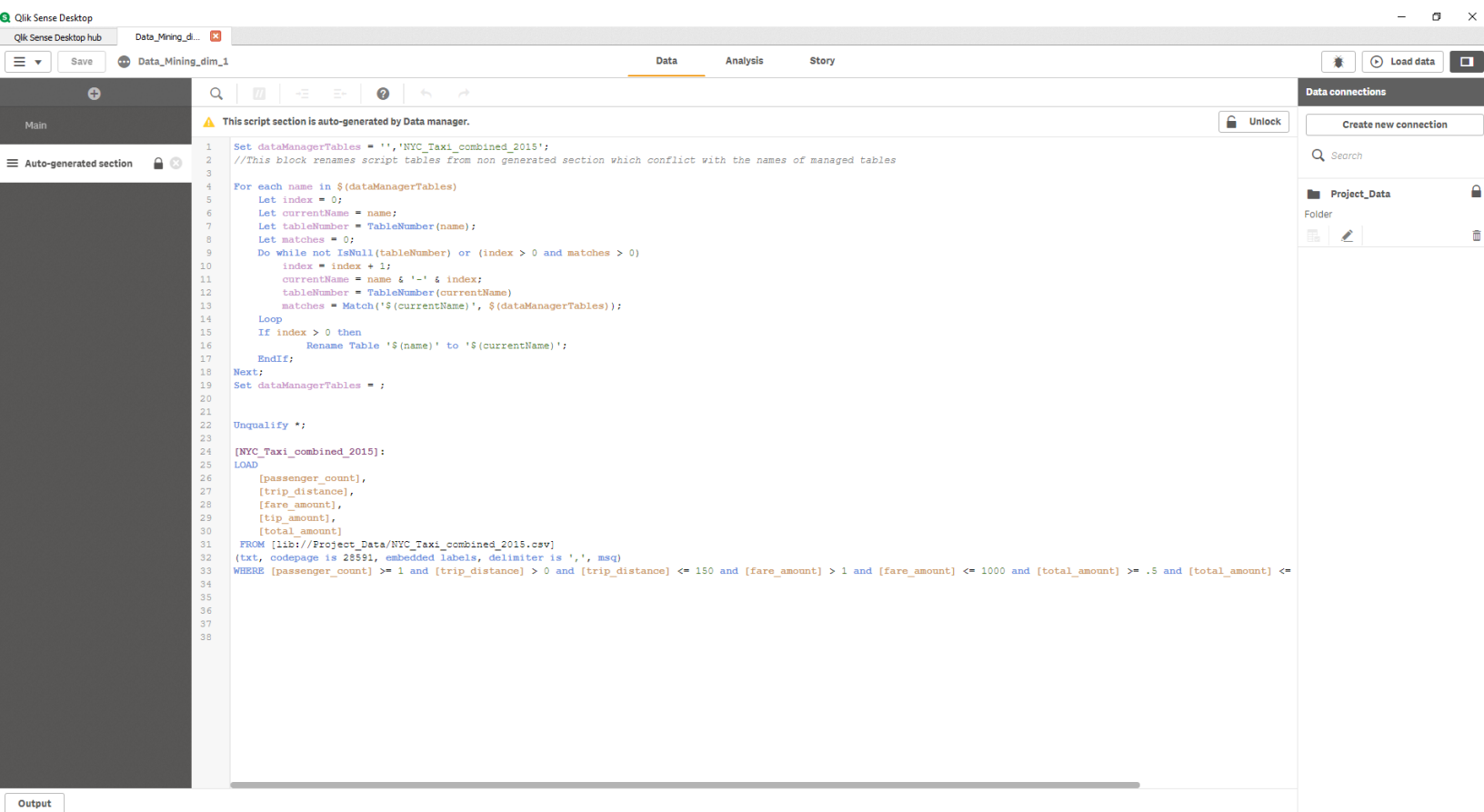


Overall, I used various tools for my analysis as opposed to crafting original code. Below is the script I used for Qlik, which was auto-generated. This section essentially shows the data being loaded into Qlik.



Included in this section is also a javascript file called "codewander-k-means-cluster." This was the code used for the k-means clustering via a Qlik extension. Neither the code, nor the entirety of the extension, are my own. I utilized them as a tool to explore the data.

Another tool that was utilized was RapidMiner. Some of the data was uploaded and transformed via RapidMiner

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators, etc. All Studies

Result History ExampleSet (/Local Repository/Data/NYC_Taxi_combined_2015)

Open in Turbo Prep Auto Model

Filter (146,113,000 / 146,113,000 examples): all

Row No.	passenger_...	trip_distance	fare_amount	tip_amount	total_amount
1	1	1.590	12	3.250	17.050
2	1	3.300	14.500	2	17.800
3	1	1.800	9.500	0	10.800
4	1	0.500	3.500	0	4.800
5	1	3	15	0	16.300
6	1	9	27	6.700	40.330
7	1	2.200	14	0	15.300
8	3	0.800	7	1.660	9.960
9	3	18.200	52	0	58.130
10	2	0.900	6.500	1.550	9.350
11	1	0.900	7	1.660	9.960
12	1	1.100	7.500	1	9.800
13	1	0.300	3	0	4.300
14	1	3.100	19	3	23.300
15	1	1.100	6	0	7.300
16	1	2.380	16.500	4.380	22.680
17	5	2.830	12.500	0	14.300
18	5	8.330	26	8.080	41.210
19	1	2.370	11.500	0	13.300
20	2	7.130	21.500	4.500	27.800
21	1	3.600	17.500	0	19.300
22	1	0.890	5.500	1.620	8.920
23	1	0.960	5.500	1.300	8.600
24	2	1.250	6.500	1.500	9.800
25	5	2.110	11.500	2.500	15.800
26	5	1.150	7.500	1.700	11
27	1	1.530	9	0	10.800
28	1	18.060	52	6	64.130
29	1	1.760	10	2.360	14.160

There are more rows in the ExampleSet than displayed here.

ExampleSet (146,113,000 examples, 0 special attributes, 5 regular attributes)

Repository

Import Data

- Training Resources (connected)
- Samples
 - Community Samples (connected)
 - DB (Legacy)
 - First_Pred (Jonathan)
- Local Repository (Jonathan)
 - Connections (Jonathan)
 - Data (Jonathan)
 - NYC_Taxi_combined_2015 (Jonathan - v1, 4/24/20 10:22 AM - 50 MB)
 - Taxi_Jan_Reduced (Jonathan - v1, 3/27/20 9:15 AM - 632 MB)
 - yellow_tripdata_2015-01 (Jonathan - v1, 3/25/20 4:46 PM - 1588 MB)
- Temporary Repository (Jonathan)

Calculating statistics

Unfortunately, RapidMiner restricted access to their code so I am unable to provide any descriptions of it. However, given how useful this tool was I wanted to include a screencap of it. One of the problems we faced with this project was being able to provide enough RAM and other resources to properly analyze the data with certain tools. With snippets of the data I was able to transform, clean, and analyze certain pieces and then analyze in large scale in different ways. If I had better resources I think this would have been my go to tool.