NYC Bus Clustering Analysis

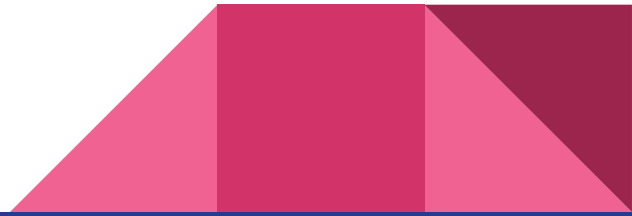
Josh Lim

Agenda

1. Introduce the data
2. Justify cleaning process
3. Explain clustering method/process
4. Cluster Analysis
5. Dimension Reduction Comparison
6. Discuss limitations
7. Possible next steps

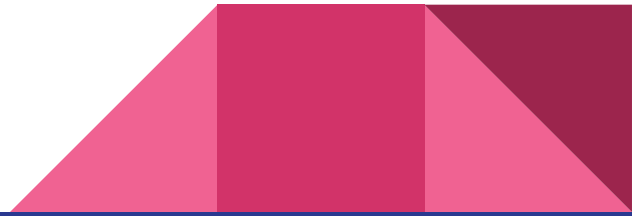
About the Data

- Obtained dataset from kaggle
- Data comes from New York City MTA buses data stream service
- Initial dataset was very large:
  - 17 Columns
  - 6,865,376 Rows



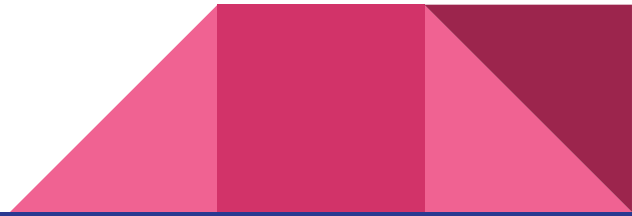
Data Cleaning

- Average time for clustering 6,000,000 rows ~ 6-7 hours
- Average time for clustering 100,000 rows < 30 minutes
- Sampled down to 100,000 rows
- Given a random state of 13



Data Cleaning

- Around 16% of the data in Expected Arrival Time was missing
- Missing data did not appear until further into the dataset
- Filled missing data with forward fill method



Data Cleaning

ScheduledArrivalTime
10:23:20
08:07:39
13:38:00
15:30:00
18:12:24
17:06:35

Data Cleaning

ScheduledArrivalTime
10:23:20
08:07:39
13:38:00
15:30:00
18:12:24
17:06:35

ScheduledArrivalTimeHour	ScheduledArrivalTimeMinute	ScheduledArrivalTimeSecond
8	7	39
13	38	0
15	30	0
17	6	35
6	41	0

Data Cleaning

Original Data

ExpectedArrivalTime
2017-10-16 10:25:48
2017-10-16 10:25:48
2017-10-08 13:41:22
2017-10-24 15:39:57
2017-10-18 18:21:16

Newly Formatted Data

ExpectedArrivalTimeDayofWeek
0
6
1
2
6

Clustering Methods

- Two dimensionality reduction techniques used
  - PCA
  - t-SNE
- Two clustering methods used
  - KMeans Clustering
  - Agglomerative Clustering
- Silhouette score used to determine best clustering method
  - Used to analyze separation distance between clusters
  - Ranges from -1 to 1

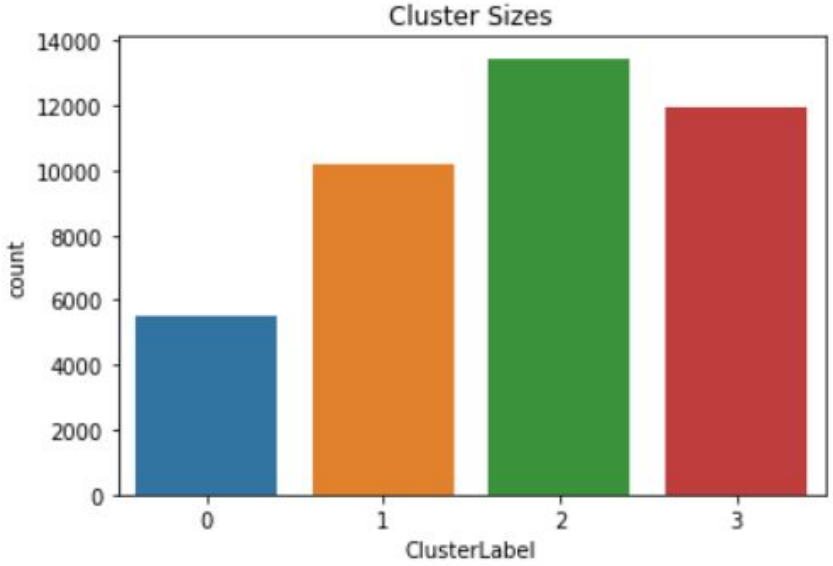
Applying KMeans to the Data

- 3-10 values were chosen to test out KMeans with PCA
- Set a random state of 13 so I would have consistent scores to compare
- Of the 7 values, K = 4 had best silhouette score

KMeans with PCA Silhouette Score Results:

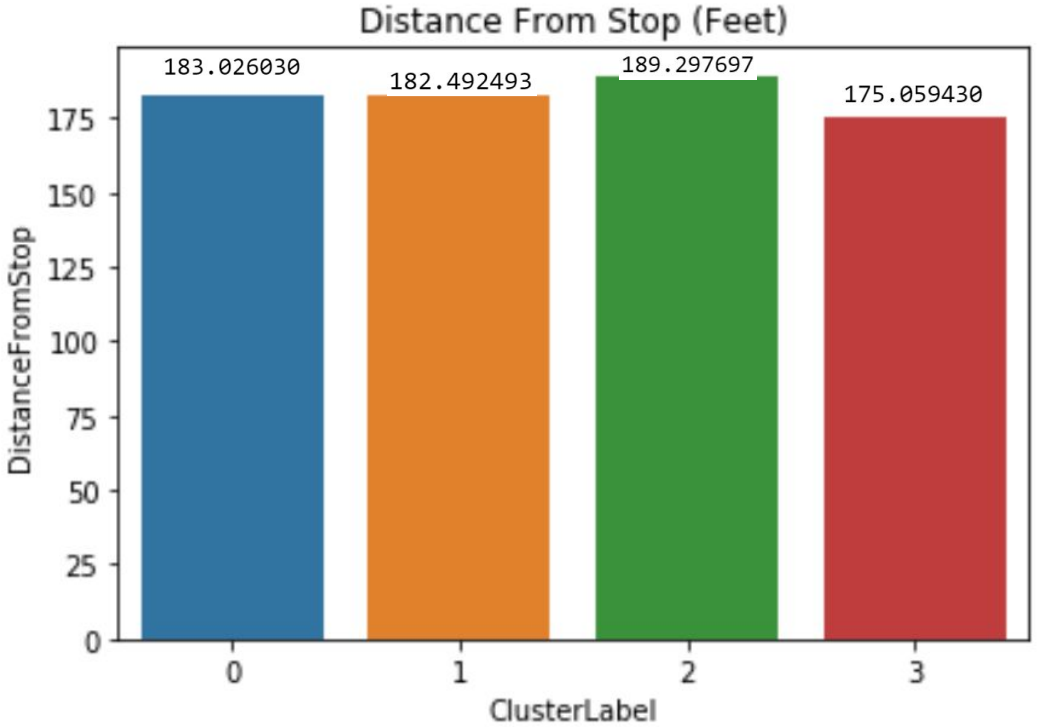
- 3 clusters: 0.111
- 4 clusters: 0.115
- 5 clusters: 0.104
- 6 clusters: 0.100
- 7 clusters: 0.099
- 8 clusters: 0.095
- 9 clusters: 0.096
- 10 clusters: 0.097

KMeans with K = 4 and PCA



2	13460
3	11930
1	10191
0	5532

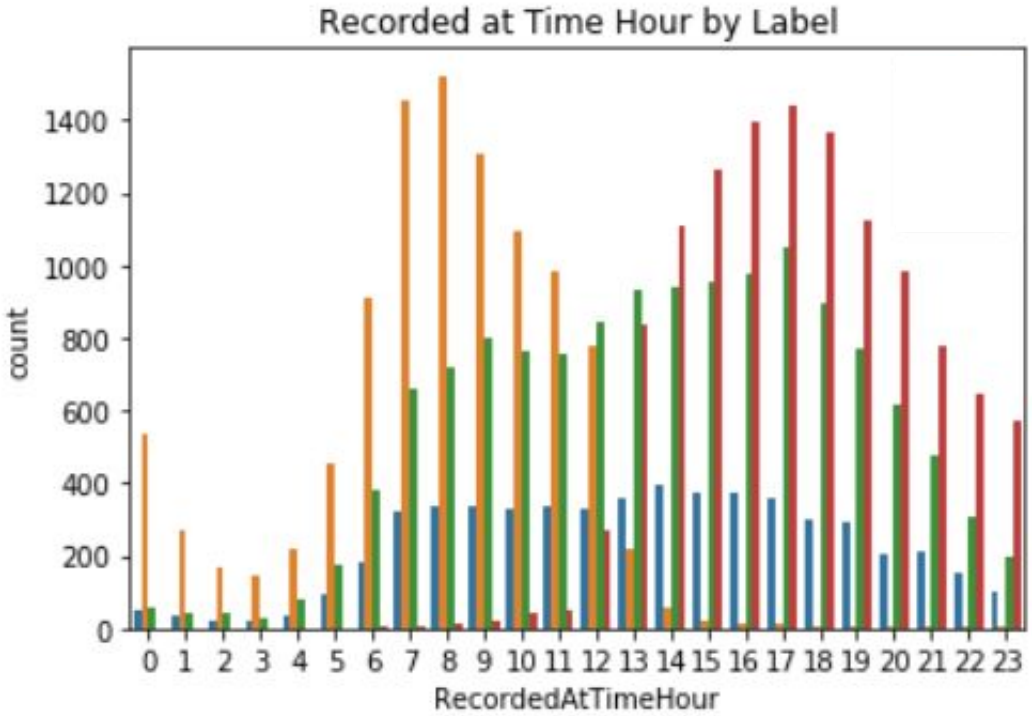
KMeans with K = 4 and PCA



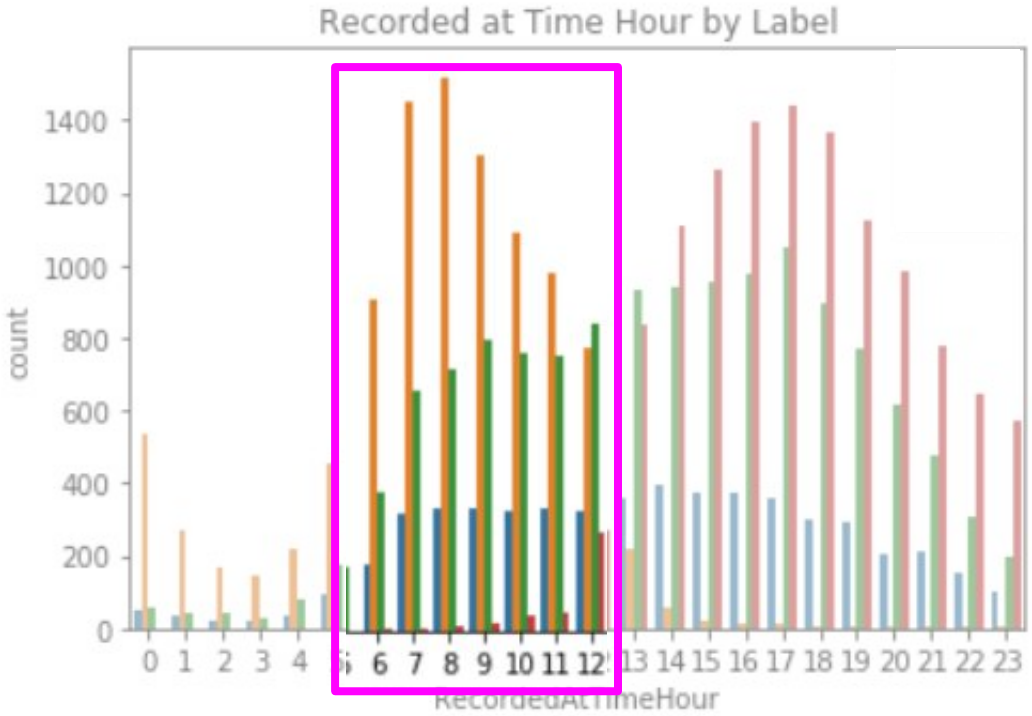
Late Bus Cluster

KMeans with K = 4 and PCA

Late Bus Cluster



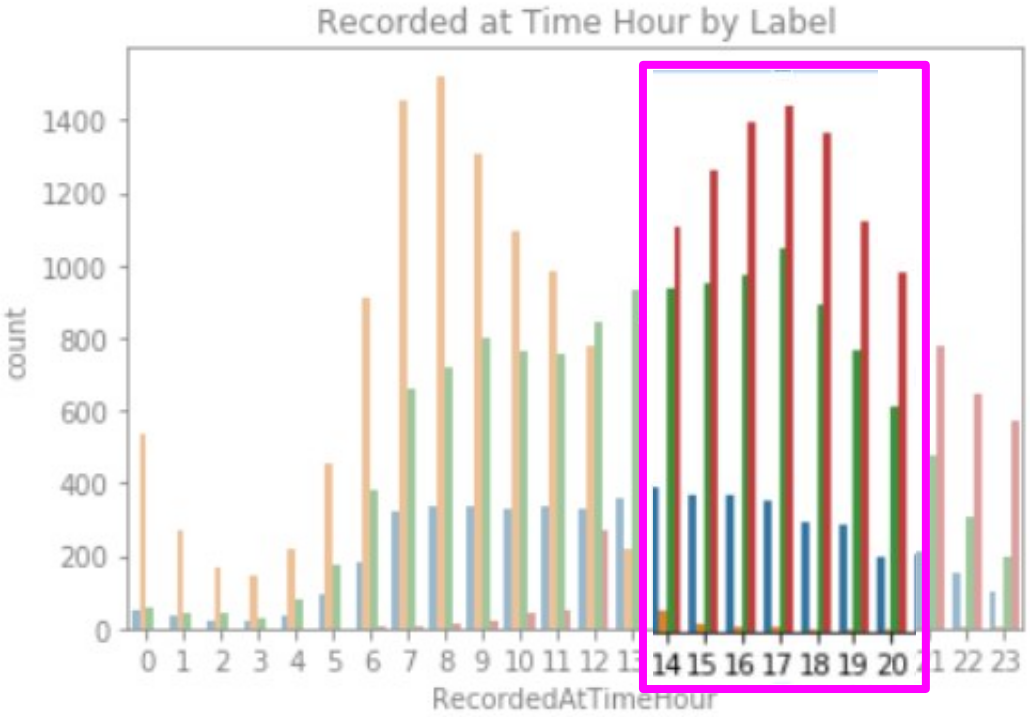
KMeans with K = 4 and PCA



Late Bus Cluster

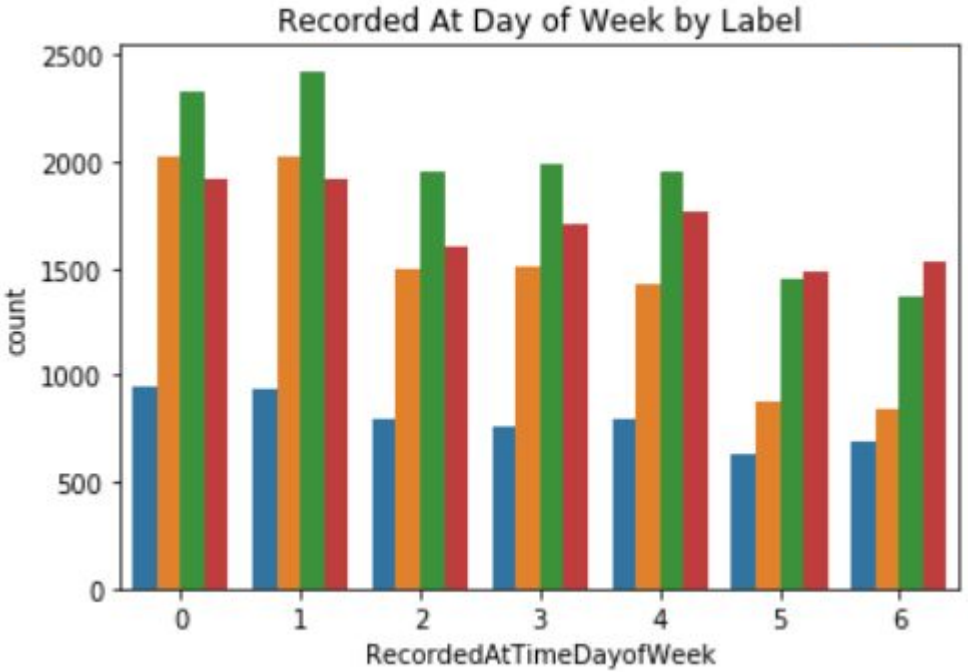
Morning Rush Hour Bus Cluster

KMeans with K = 4 and PCA



- Late Bus Cluster
- Morning Rush Hour Bus Cluster
- Evening Rush Hour Bus Cluster

KMeans with K = 4 and PCA



- Late/Weekday Bus Cluster
- Morning Rush Hour Bus Cluster
- Evening Rush Hour/Weekend Bus Cluster

Applying Agglomerative Clustering to the Data

- Three methods to determining distance in Agglomerative Clustering:
  - Complete
  - Average
  - Ward
- Tested out all three linkage methods with specified clusters 4-10

Applying Agglomerative Clustering to the Data

Silhouette Score for Agglomerative Clustering with complete linkage:

- 4 clusters is: 0.33938366
- 5 clusters is: 0.33242837
- 6 clusters is: 0.30105183
- 7 clusters is: 0.29495662
- 8 clusters is: 0.27527076
- 9 clusters is: 0.26274875
- 10 clusters is: 0.2507642

Silhouette Score for Agglomerative Clustering with average linkage:

- 4 clusters is: 0.35945204
- 5 clusters is: 0.3402213
- 6 clusters is: 0.33245412
- 7 clusters is: 0.3241567
- 8 clusters is: 0.30453718
- 9 clusters is: 0.28516117
- 10 clusters is: 0.26398465

Silhouette Score for Agglomerative Clustering with ward linkage:

- 4 clusters is: 0.31179234
- 5 clusters is: 0.3195911
- 6 clusters is: 0.36392343
- 7 clusters is: 0.353737
- 8 clusters is: 0.35034788
- 9 clusters is: 0.3562737
- 10 clusters is: 0.363912

Applying Agglomerative Clustering to the Data

Silhouette Score for Agglomerative Clustering with complete linkage:

- 4 clusters is: 0.33938366
- 5 clusters is: 0.33242837
- 6 clusters is: 0.30105183
- 7 clusters is: 0.29495662
- 8 clusters is: 0.27527076
- 9 clusters is: 0.26274875
- 10 clusters is: 0.2507642

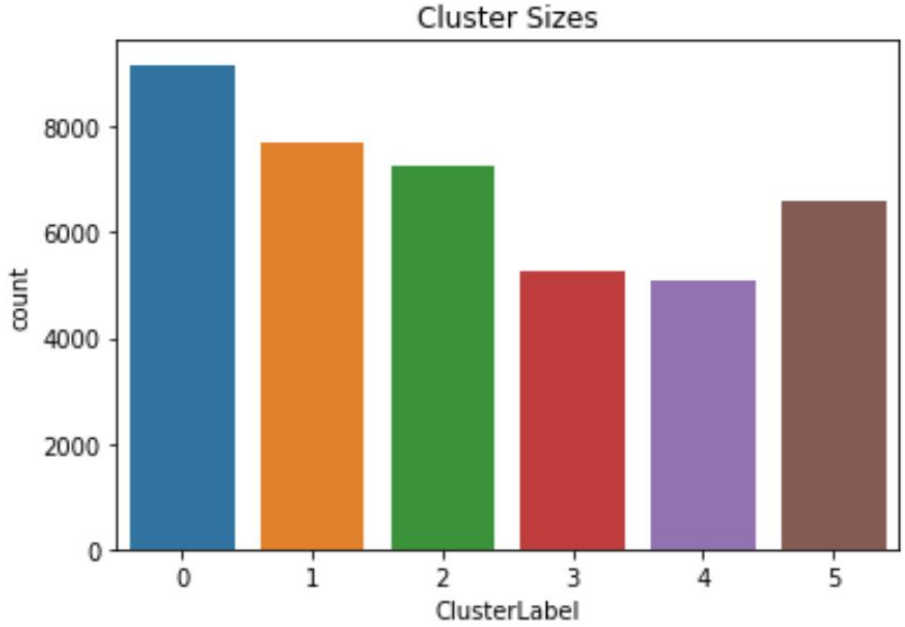
Silhouette Score for Agglomerative Clustering with average linkage:

- 4 clusters is: 0.35945204
- 5 clusters is: 0.3402213
- 6 clusters is: 0.33245412
- 7 clusters is: 0.3241567
- 8 clusters is: 0.30453718
- 9 clusters is: 0.28516117
- 10 clusters is: 0.26398465

Silhouette Score for Agglomerative Clustering with ward linkage:

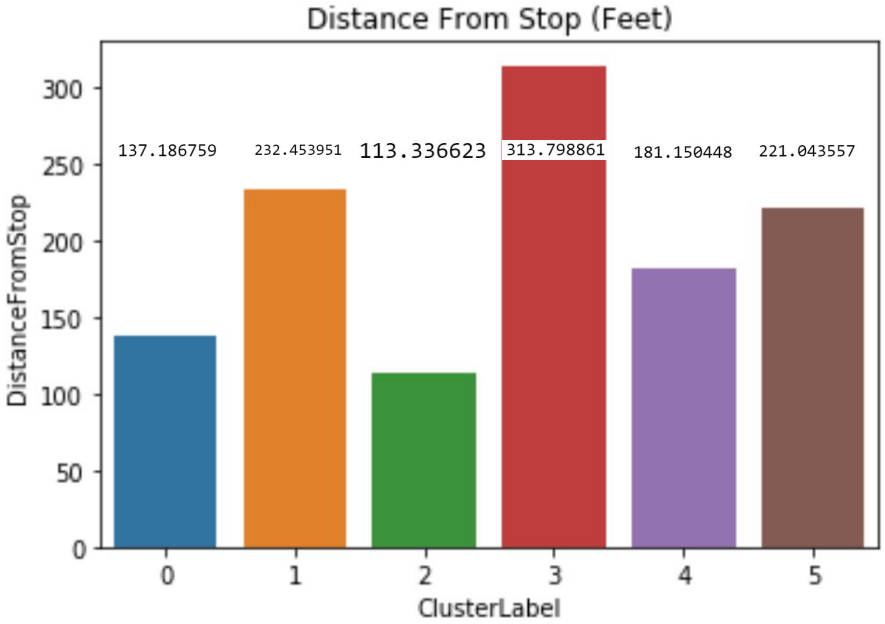
- 4 clusters is: 0.31179234
- 5 clusters is: 0.3195911
- 6 clusters is: 0.36392343
- 7 clusters is: 0.353737
- 8 clusters is: 0.35034788
- 9 clusters is: 0.3562737
- 10 clusters is: 0.363912

Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE



0	9182
1	7716
2	7243
5	6597
3	5266
4	5109

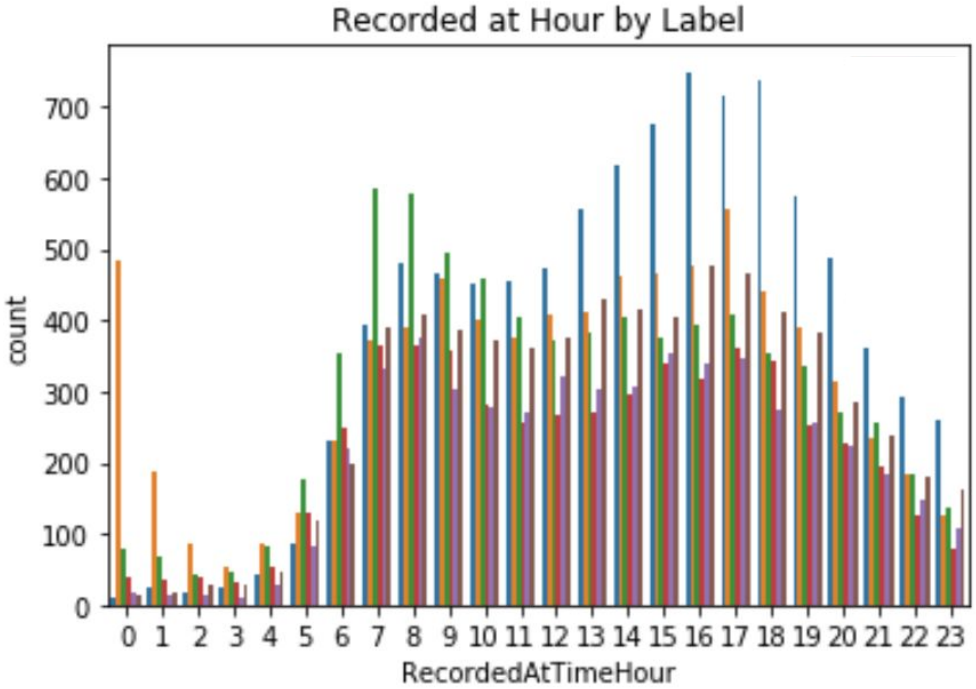
Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE



Late Bus Cluster

Closest to On Time Bus Cluster

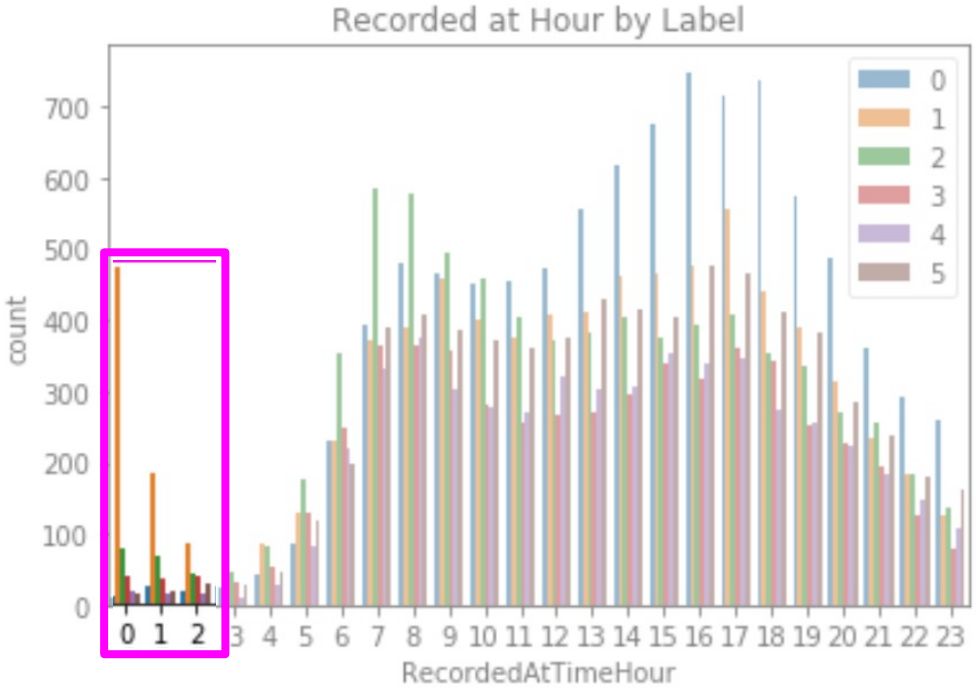
Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE



Late Bus Cluster

Closest to On Time Bus Cluster

Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE

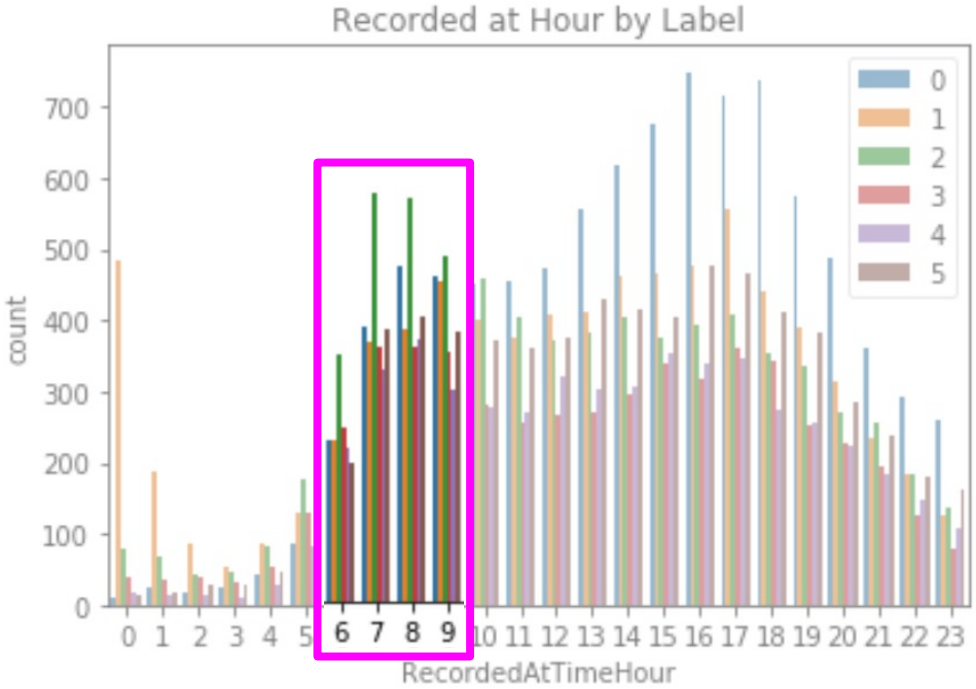


Late Bus Cluster

Closest to On Time Bus Cluster

12 AM - 2 AM Bus Cluster

Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE

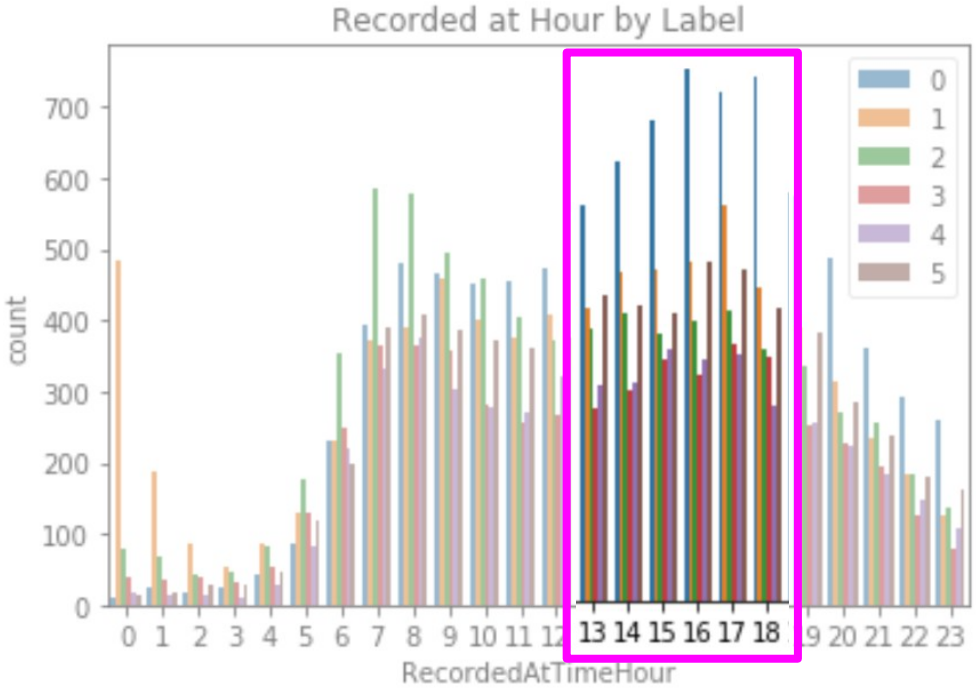


Late Bus Cluster

Closest to On Time/Morning Rush Hour Bus Cluster

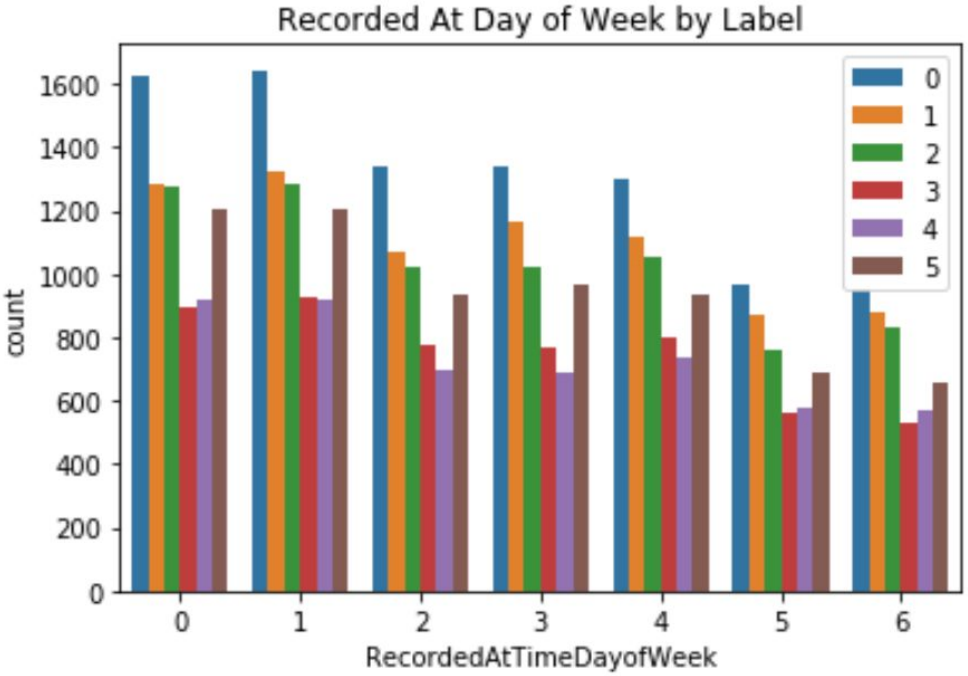
12 AM - 2 AM Bus Cluster

Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE



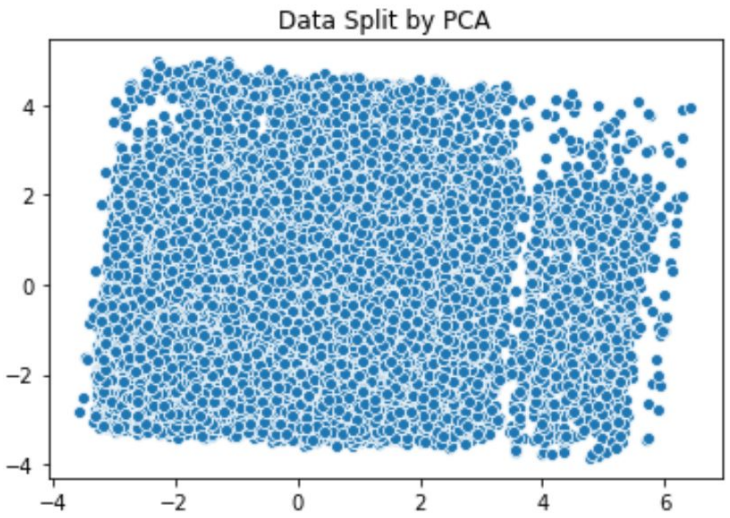
- Late Bus Cluster
- Closest to On Time/Morning Rush Hour Bus Cluster
- 12 AM - 2 AM Bus Cluster
- Afternoon - Evening Rush Hour Bus Cluster

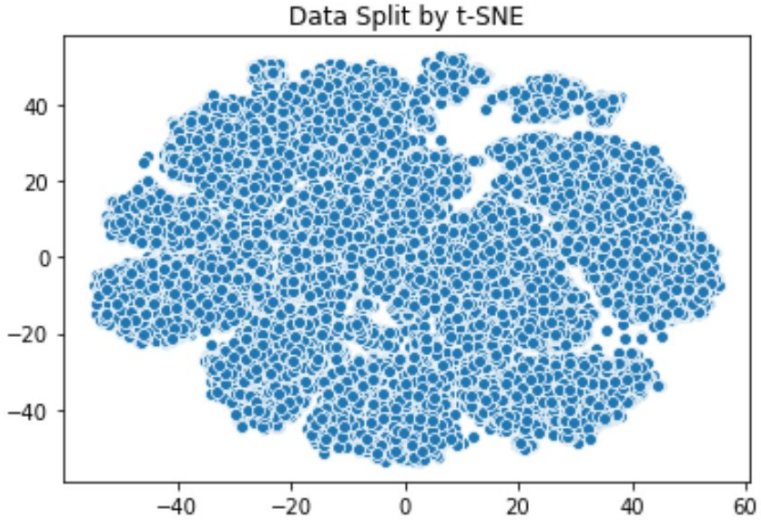
Agglomerative Clustering with 6 Clusters, Ward Linkage, and t-SNE



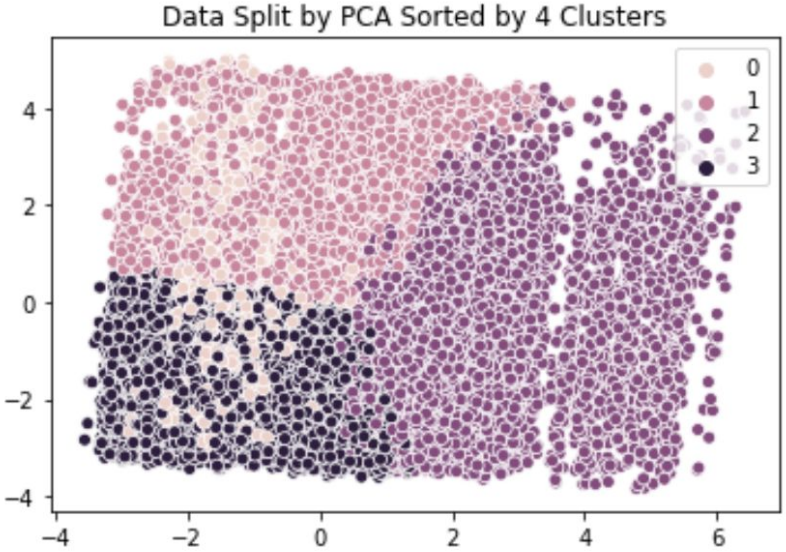
- Late Bus Cluster
- Closest to On Time/Morning Rush Hour Bus Cluster
- 12 AM - 2 AM Bus Cluster
- Afternoon - Evening Rush Hour Bus Cluster

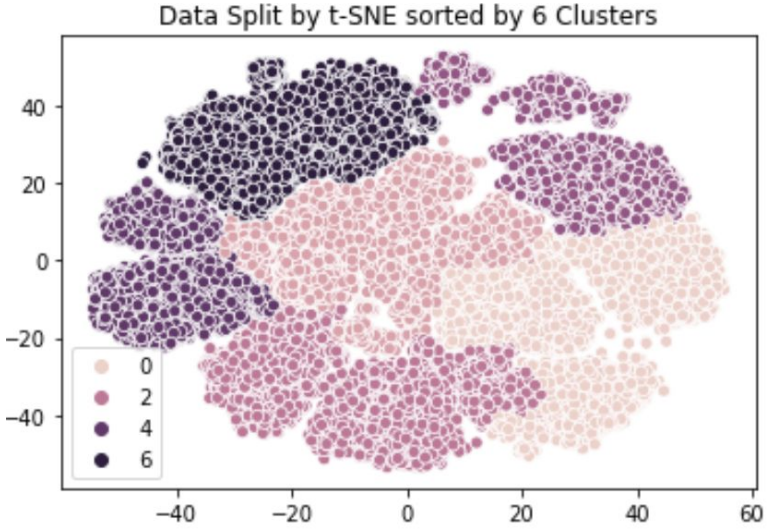
Comparing Dimensionality Reduction Methods





Comparing Dimensionality Reduction Methods





The Silhouette Score for 4 clusters is:0.1147058207173504

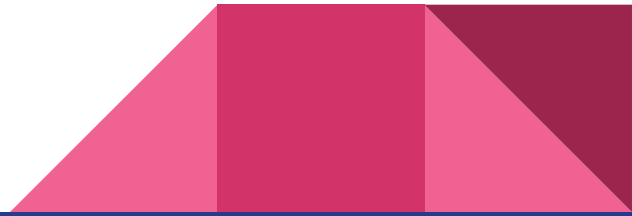
The Silhouette Score for 6 clusters is:0.36392343

Overall Conclusion

- Agglomerative Clustering with t-SNE creates better results than KMeans with PCA
  - 0.364 silhouette score vs 0.115
  - t-SNE appears to separate the data more effectively
- Labeled clusters can be used to determine which bus routes will be considered late
  - KMeans: cluster that ran on weekdays was also the bus cluster that happened to be the latest to its destination
  - Agglomerative: cluster that was closest to being on time also happened to be the morning rush hour cluster

Limitations

- Data was too large to use entire dataset in given time frame of the project
  - Subsampling can result with information loss
  - Clustering methods were not completely efficient due to smaller data sample
- Time constraint



Possible Next Steps

- Explore more with dimension reductionality techniques
  - Try lower perplexities for t-SNE
  - Utilize UMAP
- Apply clustering to original dataset
  - 7 million rows vs 100,000
- Determine if there is a pattern to late bus lines
  - Are particular published bus lines more late than others?
  - Do busses that start at particular origin stops tend to be later than others?

Thank you! Questions?