Component Analysis
1a. The likely candidates for early PCA dimensions are hyperplanes that capture the most variance in data; and they are vectors that are closest to the data points, meaning these vectors fall into the center of the group of data points. When graphed on a principal component biplot, most points will fall in the middle (0, 0). Some will lie mostly along the first principal component axis (x-axis), and some will lie mostly along the second (y-axis). I think that some feature vectors will fall close together, such as "fresh" and "milk". A lot of points will lie along those vectors. By looking at the data, I see a few things. Each row is the wholesaler's customer, and the numbers represent the amount bought (unsure if the numbers represent number of items or dollar amount purchased).

1b. ICA returns independent hyperplanes. These independent hyperplanes could be the underlying source that are causing the data points. It can hopefully isolate separate buyer types, just as ICA isolates separate voices in the cocktail party problem. I think the components will be statistically independent. The numbers and format of ICA's output will be similar to PCA, but the vector will find a space where the features are maximally independent. This means ICA is able to distinguish the difference between the two features, just as the cocktail party problem can distinguish different voices. One possible ICA component could be an anti-correlation between detergents paper and delicatessen products.

2. The proportion of variance explained by each PC is listed in the .explained_variance_ratio_ attribute: [0.45961362 0.40517227 0.07003008 0.04402344 0.01502212 0.00613848]. The variance drops off quickly, with over 90% of its variance being captured by the first three principal components. Using a scree plot to analyze the variance captured, we see see an "elbow" at two principal components. One principal component only captures 46% of the variance. Three principal components doesn't give significantly more information than two, and the additional information can possibly be noise. Therefore, I would use two principal components.

3a. Each element of PCA is a hyperplane. The first PC is the hyperplane that projects the data points from 6-dimensional space into 5-dimensional space; of all the PCs, it captures the most information (the least information loss), at 45.96%. The second PC is a hyperplane that is orthogonal to the first PC, and it captures the second most amount of variance, at 40.52%. Together they capture 86.48% of the information.

3b. This information is used to capture the most amount of information with the least amount of information loss. You can plot the first and second components on a graph as the x and y axes, respectively.

4a. The independent components that arise are similar to those of PCA, except the vectors seeks to find maximum independence in the data and make the data as non-Gaussian as possible.

4b. The first independent component shows an anti-correlation between grocery foods and detergents paper. This could represent a group of buyers who represent mini grocery stores and who don't have much space for detergents paper.

The second independent component primarily measures frozen products, so this could be an ice cream shop or gas station store.

The third component could be represented by a store that sells fresh products, like a farmer's market.

The fourth component shows an anti-correlation between milk and grocery foods. This can be a group of vegan restaurants or a cheese shop.

This information can be helpful for the distributor in providing a information about the products that are independent of each other. Perhaps, it doesn't matter whether or not they are shipped together, and it isn't important to group them together.

Clustering

5a. K-Means clustering is very fast, and each data point falls into a definite cluster. This makes it more scalable. K-Means is hard clustering, meaning each cluster has a definite set of claimed data points. Gaussian Mixture Model is soft clustering. It is slower, since the algorithm calculates covariance, mean, variance, and the data probabilities. This means that each point is assigned, based on the the probability that it actually falls into the cluster. GMM is used when the clusters and their boundaries are not apparent. GMM gives more structural information: it maps out the Gaussian cluster circles, so on a two-dimensional graph it provides how the cluster circles are oriented in three-dimensional space.

GMM is the better suited clustering algorithm here. Since the data points are plotted in six dimensions, we need GMM to provide more information of the clusters.

6. Here are the five means for the Gaussian model:

```
[[ 0.12977036 -0.80293783]
 [ 0.96223753  0.09910825]
 [-0.81717902  0.78618511]
 [-5.98369792  1.02504959]
 [-3.94858233 -6.66774494]]
```

The first cluster of customers buys a small amount of fresh products, milk and grocery products. This could be a convenience store.

The second cluster buys fresh products and very little of the rest of the products. This could be a small butcher shop or seafood store.

The third cluster buys, in order of abundance, a lot of groceries, milk, detergents paper, some fresh products, and deli. This could be a mid-sized grocery store.

The fourth cluster buys groceries, milk, detergents paper, and some fresh products. This could be one of those corner grocery stores that presents fruits and vegetables outside in woven baskets.

The fifth cluster buys a large amount of fresh products, followed by a lot of milk, groceries, frozen foods, and deli. This could be a large grocery store, like a Safeway or a Wal-Mart Supercenter.

The sixth cluster buys more fresh products, along with some milk, groceries, and frozen products. This could be a health food store, such as a Whole Foods or Trader Joe's.

7. The clusters aren't distinguished, and their boundaries are ambiguous. The visualization can be improved by playing around with Scikit-Learn's preprocessing classes, such as StandardScaler() and FunctionTransformer(). They can stretch the data, so the space between points are more distinguishable. The central objects (x's) aren't data points, but they are the mean point of each cluster of data points. These customers can be categorized as one who buy a certain combination and price of the six wholesale products.

Conclusions

8. Graphing the Gaussian Mixture Model gave me the most insight into the data. Intuitively, or visually, GMM fits well with the data, and the clusters become apparent after graphing them; whereas, KMeans divides the data into clusters that don't look apparent. GMM allowed the data to be separated based on the likelihood that each data point belongs to its corresponding cluster.

9. There are a few different A/B tests for delivery methods I have come up with, after analyzing the data. Here is one of them.

1 I asked: "Is there another type of delivery method, that has been overlooked, that we can use to satisfy all clusters of customers?"

2 I researched: the data, using PCA and clustering methods.

3 I hypothesized: Understanding the amount of perishable items vs nonperishable items of each type/cluster of buyer will satisfy each customer.

4 I would run the test for a month, in order to capture the full inventory turnover of each store.

5 Test the hypothesis: two types of shipments: 1 have larger and less frequent shipments of nonperishable items and 2 smaller and more frequent shipments of perishable items for the smaller stores: convenience stores, seafood/butcher shops, corner grocery stores, and health food stores. The stores that purchase a larger volume of products can be shipped their inventory more frequently, since they can easily adapt to a change in delivery method.

6 I would analyze to see if dividing the delivery shipment for each store cluster is both feasible for them and cheaper for the wholesaler.

10. We can use the clustering technique to predict the type of new incoming features for buyers, using a supervised learning model. We could use GMM to cluster types of customers and create new features, based on what those customers buy. The target column will consist of each of the six cluster groups. We can test the accuracy of the prediction, using an F1 score.
I don't understand why we wouldn't want to brainstorm when the last grader gave me this source that says to brainstorm features then test then rebrainstorm features: https://en.wikipedia.org/wiki/Feature_engineering.
Please just give me the same grader for consistency, so I won't have to keep iterating my answer then have it shut down by a new grader with completely different standards.