	A	B	AE	AF
1				**Shen Yang**
2		Ideal		Team: YS
3		Score	Score	---------- Instructor Notes----------
4	**Total Points** **Total Percentage**	10	8.7	Score Points:08.70
5		100	87	
6	Exceptional Work: **7000 required** implement dimensionality reduction using t-SNE, then visualize and interpret the results. Give an explanation of t-SNE dimensionality reduction methods.	10	4	This implementation of TSNE is pretty good quality. I like that you tried two different colorings and your analysis of the results is spot on. You may need to adjust the perplexity and the initilizer for tsne to get a better clustering. > Finally, I was also looking for a better explanation of the tsne algorithm, with equations and all.
7	In your own words, give an overview of the dataset. Describe the purpose of the data set you selected (i.e., why and how was this data collected in the first place?). What is the prediction task for your data and why are other third parties interested in the result? Once you begin modeling, how well would your prediction algorithm need to perform to be considered useful to these third parties? Be specific and use your own words to describe the aspects of the data.	15	12	This is a good start to the project overview. I like that you identified false positive costs and the relative impact that they would have for developers of apps. This is the right way to start bounding performance. My main criticism here is that your prediction algorithm is only created to invesitgate factors that affect installations. Therefore, the performance of the model is only needed to establish trust. > You talk a bout a few methods for clustering installations into classification, but its unclear if that is what you will be doing (or if you will use this as a regression task). > Trust in the model needs to have a concrete definition. If you predict, on average, less than 10% difference between prediction and actual, does that give you trust in the model? Moreover, should you look at the number of times you have "bad" confusions? For instance, how many times do you predict an app will have a lot of downloads, but it has very few? Doe that change your trust in the model, regardless of average performance? What number of installations need to be used to make an app successful? perhaps this is a better criteris to predict, rather than absolute downloads.
8	Load the dataset and appropriately define data types. What data type should be used to represent each data attribute? Discuss the attributes collected in the dataset. For datasets with a large number of attributes, only discuss a subset of relevant attributes.	15	14	To improve: also discuss why you are not interested in many of the variables you throw away. For instance, why throw away Type?
9	Verify data quality: Explain any missing values or duplicate data. Are those mistakes? Why do these quality issues exist in the data? How do you deal with these problems? Give justifications for your methods (elimination or imputation).	15	12	There is a check for NaNs, but nothing that actually investigates outliers in the data or ranges in the data that are appropriate. The data might be good, but you need to show there is nothing wrong, like an SizeMB that is negative.
10	Visualize attribute distributions. Choose and visualize distributions for a subset of single attributes. Choose any appropriate visualization such as histograms, kernel density estimation, box plots, etc. Describe anything meaningful or potentially interesting you discover from these visualizations. Note: You can also use data from other sources to bolster visualizations. **Visualize at least 5 attributes, at least one categorical and at least one numeric.**	20	20	This is great quality and well described. I like that you try to hypothesize different meanings for the app differences.
11	Visualize relationships between a subset of attributes. Use whichever visualization method is appropriate for your data. Explain any interesting relationships. Important: Interpret the implications for each visualization. **Visualize at least three subsets of the attributes.**	25	25	Very high quality.
12				
13				
14				