Project Rubric and Project Grade:

Making a Video Summary

Use screencasting software to make your video with slides and or ipython/jupyter notebooks. You do not need to show yourselves in the video unless you think it adds to your presentation. It should look nice. Don't just record yourself giving a talk where the slides are heard to see and so are you.

Here are some suggestions on screencasting software, some free some not:
http://lifehacker.com/5839047/five-best-screencasting-or-screen-recording-tools

Project/Presentation Rubric Tips

Read through the rubric below to know what I will look for when I grade. Note that:

- **Your work should use data science techiques used in the class to answer these questions such as**
  - Multi-variable non-linear regression
  - Classification
  - Clustering
  - Dimension Reduction
- You should not simply show averages, counts or histograms of the data. Such basic analysis may be important to, for example, motivate the uses of a set of variables as features vectors for classification, or as motivation, but it does not show understanding of the core course material.

How I will Grade

I will read through the commits on your group project. I will assess how much work you did relative to the rest of the class, and how much work you did with respect to your teammates on the project. From this I will assess a "work score". Any code that was copied directly from the internet is not your work and does not count toward work score. In fact, anything you use externally should be properly referenced otherwise it is plagerism.

Your work fraction on the project = (your workscore)/(sum of teams workscore)

The project grade will follow the project rubric below. Some questions below may not be relavant to some projects but all parts I-VII are required. The total grade will be based on:

your workscore/(mean class(workscore)+2stdev(workscore)) + + ((fraction of work on project) * (project grade)) + (project grade)

I Data Science Questions: 20 pts

A. Is the background context for the question stated clearly (with references)?

B. Is the hypothisis/problem stated clearly ("The What")

C. Is it clear why the problems are important? Is it clear why anyone would care? ("The Why")

D. Is it clear why the data chosen should be able to answer the question being asked?

E. How new, on-obvious, significant are your problems? Do you go beyond checking the easy and obvious?

II Data Cleaning/Checking/Data Exploration: 20pts

A. Did you check for outliers?

B. Did you check the units of all data points to make sure they are in the right range?

C. Did you identify the missing data code?

D. Did you reformat the data properly with each instance/observation in a row, and each variable in a column?

E. Did you keep track of all parameters and units?

F. Do you have specific code for reformating the data that does not require information not documented (eg. magic numbers)?

G. Did you plot univariate and multivariate summaries of the data including histograms, density plots, boxplots?

H. Did you consider correlations between variables (scatterplots)?

I. Did you consider plot the data on the right scale? For example, on a log scale?

III. Transformation and Modeling: 20pts

A. Did you transform, normalize, filter the data appropriately to solve your problem? Did you divide by max-min, or the sum, root-square-sum, or did you z-score the data? Did you justify what you did?

B. Did you pick an appropriate set of models to solve the problem? Did you justify why these models and not others?

C. Did you exersize the data science models/problems we described in the lectures showing what was presented?

D. Are you using appropriate hyper-parameters? For example if you are using a KNN regression are you investigating the choice of K and whether you use uniform or distance weighting? If you are using K-means do you explain why K? If you are using PCA do you explore how many dimensions such as by looking at the eigenvalues?

IV. Metrics, Validation and Evaluation 20pts

   A. Are you using appropriate choice of metrics? Are they well justified? If you are doing classification do you show an ROC curve? If you are doing regression are you justifying the metric least squares vs. mean absolute error? Do you show both?

   B. Do you validate your choices of hyperparamters? For example, if you use KNN or K-means do you use cross validataion to optimize your choice of parameters?

   C. Do you separately evaluate testing and training error? Do you estimate the uncertainty in each? Are you also making sure that your validatation for hyperparamter optimization is kept separate from your testing? Have you avoided overfitting?

V. Visualization 20pts

   A. Do you provide visualization summeries for all your data and features?

   B. Do you use the correct visualization type, eg. bar graphs for catigorical data, scatter plots for numerical data, etc?

   C. Are your axes properly labeled?

   D. Do you use color properly?

   E. Do you use opacity and dot size so that scatterplots with lots of data points are not just a mass of uninterprital dots?

   F. Do you write captions explaining what a reader should conclude from each figure (not just saying what it is but what it tells you)?

VI. Code 20pts

   A. Is code provided can reproduce the entire work?

   B. Is the data included or at least linked (externally) with instructions on how to download it?

   C. Are pylint conventions followed?

   D. Do you factor repeated operations into functions to avoid repetitive and error prone copy paste?

   E. Do you use docstrings and numpy documentation style: https://github.com/numpy/numpy/blob/master/doc/HOWTO_DOCUMENT.rst.txt to make your code clear and readable?

   F. Do you use markdown cells to explain every step of your code similar to homeworks and some example notebooks?

   G. Does the code demonstrate considerable work given the number of people on the project?

VII. Presentation (video) 20pts

   A. Do you tell a coherent story with a beggining, middle and end?

   B. Do does your video provide a guide that makes it easy to evalute I-VI?

C. Do you have good clear visuals with axis and data labeled?

D. Are your slides relavant to your story and solving your problem or are they only vaugly relavant "padding"? Is each slide justifed in your narration?