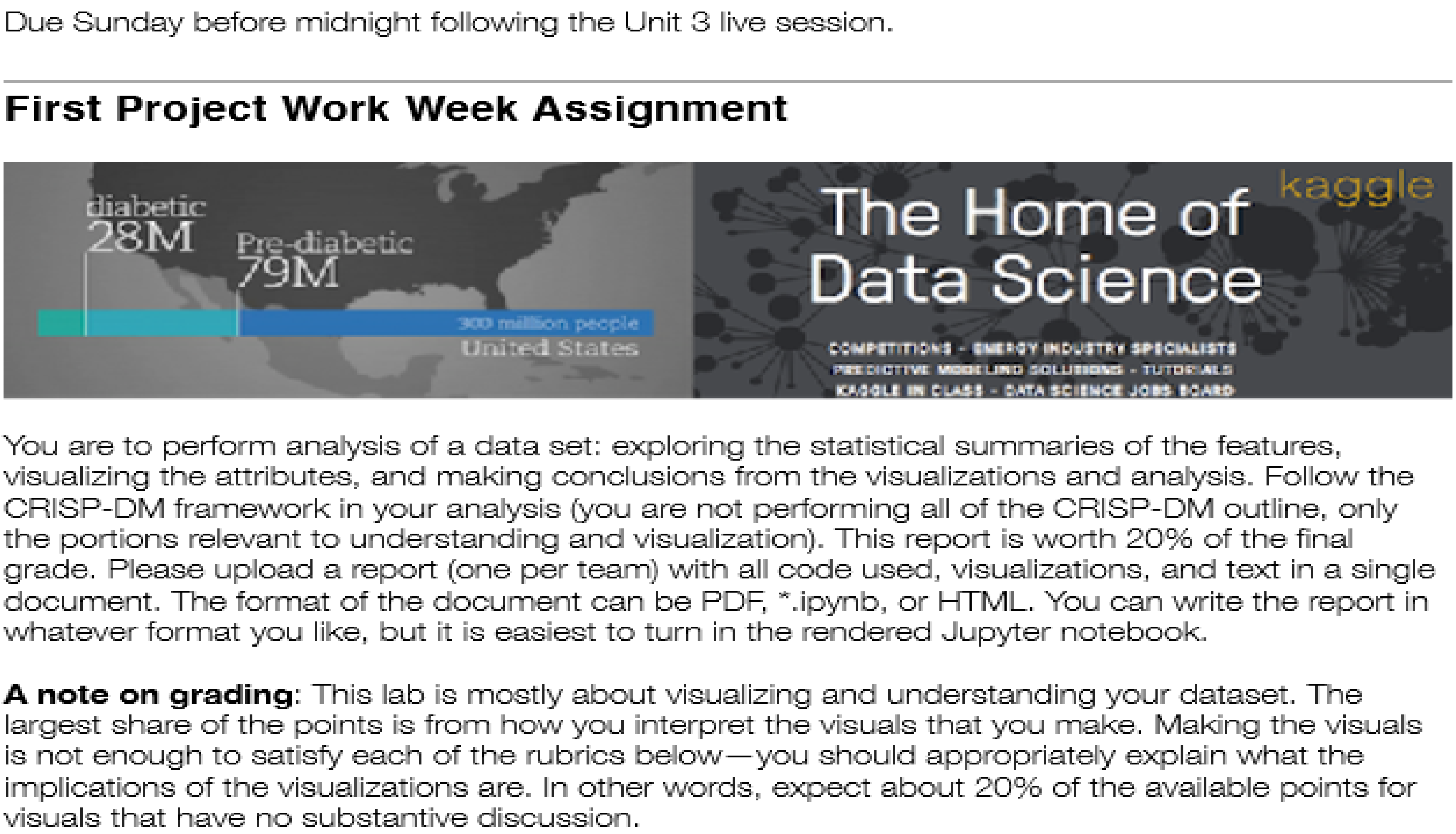
MSDS 7331 Data Mining
Lab 1

Summer 2019

Prepared by: Ben Brock

LAB 1 Due Date: 5/26/2019



Due Sunday before midnight following the Unit 3 live session.

First Project Work Week Assignment

diabetic
28M
Pre-diabetic
79M
300 million people
United States
kaggle
The Home of
Data Science
COMPETITIONS - ENERGY INDUSTRY SPECIALISTS
PREDICTIVE MODELING SOLUTIONS - TUTORIALS
KAGGLE IN CLASS - DATA SCIENCE JOBS BOARD

You are to perform analysis of a data set: exploring the statistical summaries of the features, visualizing the attributes, and making conclusions from the visualizations and analysis. Follow the CRISP-DM framework in your analysis (you are not performing all of the CRISP-DM outline, only the portions relevant to understanding and visualization). This report is worth 20% of the final grade. Please upload a report (one per team) with all code used, visualizations, and text in a single document. The format of the document can be PDF, *.ipynb, or HTML. You can write the report in whatever format you like, but it is easiest to turn in the rendered Jupyter notebook.

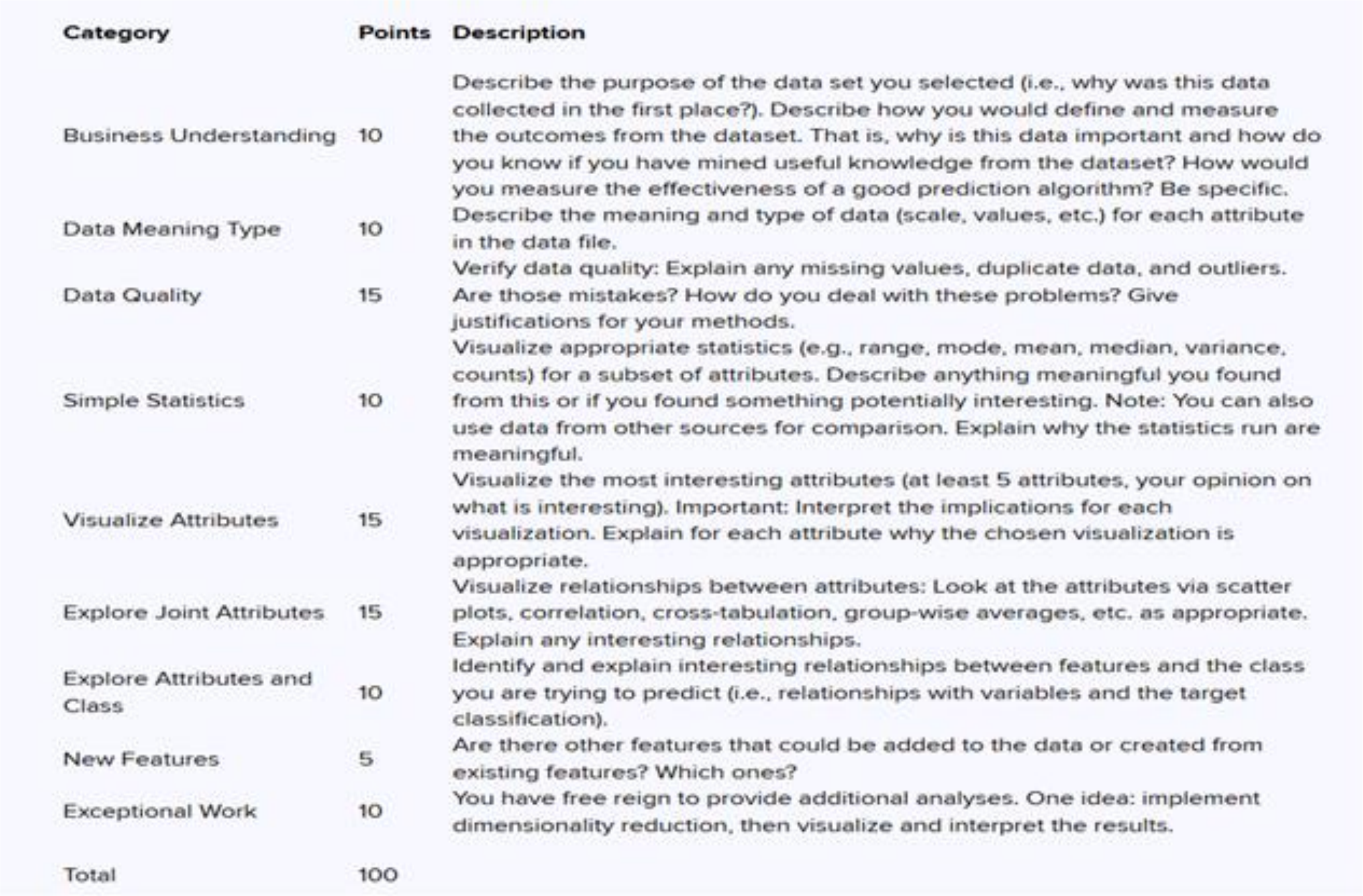
A note on grading: This lab is mostly about visualizing and understanding your dataset. The largest share of the points is from how you interpret the visuals that you make. Making the visuals is not enough to satisfy each of the rubrics below—you should appropriately explain what the implications of the visualizations are. In other words, expect about 20% of the available points for visuals that have no substantive discussion.

Rules: Please note the following policies on both LAB and ICA submissions effective immediately

1. No late work is accepted
2. All work must be submitted in file format (iPython notebook, html, or Pdf) on 2Ds by the due date.
3. Links to work (github etc) submitted on 2DS are not accepted
4. Please have multiple team members submit work to ensure there are no potential issues with submission failures or missing files.
5. As a last resort, you may email your submission to ensure it is marked by the due date. However, at least one team member will eventually have to upload the file on 2DS to receive a grade.
6. It is your team’s responsibility to ensure that all members submit the same version of a file. It is the grader’s choice as to which file will be selected for grading.
7. LAB and ICA deductions are applied to each team member’s individual grade.
8. The same team group members will work together on all LAB and ICA assignments throughout the semester.

LAB1 – GRADING RUBRIC

- The rubric that is used to grade LAB 1 is listed on the next slide.
  - You MUST have these exact sections in your project if you want me to grade it!
- LAB1 is due Sunday 5/26/2019
- Good luck on your LAB 1 assignment.



Category	Points	Description
Business Understanding	10	Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). Describe how you would define and measure the outcomes from the dataset. That is, why is this data important and how do you know if you have mined useful knowledge from the dataset? How would you measure the effectiveness of a good prediction algorithm? Be specific.
Data Meaning Type	10	Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file.
Data Quality	15	Verify data quality: Explain any missing values, duplicate data, and outliers. Are those mistakes? How do you deal with these problems? Give justifications for your methods.
Simple Statistics	10	Visualize appropriate statistics (e.g., range, mode, mean, median, variance, counts) for a subset of attributes. Describe anything meaningful you found from this or if you found something potentially interesting. Note: You can also use data from other sources for comparison. Explain why the statistics run are meaningful.
Visualize Attributes	15	Visualize the most interesting attributes (at least 5 attributes, your opinion on what is interesting). Important: Interpret the implications for each visualization. Explain for each attribute why the chosen visualization is appropriate.
Explore Joint Attributes	15	Visualize relationships between attributes: Look at the attributes via scatter plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate. Explain any interesting relationships.
Explore Attributes and Class	10	Identify and explain interesting relationships between features and the class you are trying to predict (i.e., relationships with variables and the target classification).
New Features	5	Are there other features that could be added to the data or created from existing features? Which ones?
Exceptional Work	10	You have free reign to provide additional analyses. One idea: implement dimensionality reduction, then visualize and interpret the results.
Total	100	