Overall Notes: To receive full credit for this project, you must submit your project in the format described on Canvas (see the Final Project assignment page). That is, your project should be a .zip file containing, at minimum, either an .Rmd or a Python file and the corresponding knitted .html or .pdf file, and a codebook. Submitting only the .Rmd, or submitting screenshots instead of an .Rmd and corresponding knitted file, will earn at most half credit.						
Points Possible		**Criteria for Full Credit**	**Criteria for Half Credit**	**Criteria for No Credit**	**Points Earned**	**Total**
5	**Introduction section**	**5 points:** Minimum of two paragraphs with complete sentences. Should define any necessary vocabulary terms, briefly explain relevant concepts, and discuss the primary goals of the project.	**2.5 points:** One paragraph, brief description of project or dataset. No definition of relevant vocabulary or explanation of relevant concepts.	**0 points:** One or two sentences only, or no introduction at all.	5	100
5	**Data citation or link to data source**	**5 points:** Describe the source of the dataset. Write a formal citation. Provide a link to the source, if possible.	**2.5 points:** Mentions data source and provides link, but without specific detail. No formal citation.	**0 points:** No mention of data source or no link to data source. No formal citation.	5	
5	**EDA section present**	**5 points:** A section of the report is dedicated to EDA (exploratory data analysis). It should come at the beginning of the report, just after the introduction.	**2.5 points:** Some EDA is present, but not in a separate section.	**0 points:** No EDA section is present.	5	
10	**At least 3-5 plots or tables in the EDA**	**10 points:** At least 3 plots or tables are included in the EDA. They can be the same type or different types. For example, a correlation matrix, a histogram, a scatterplot, box plot, etc.	**5 points:** Fewer than 3 plots or tables are included in the EDA.	**0 points:** No plots or tables are included in the EDA.	10	
5	**Codebook included**	**5 points:** A codebook is included that identifies and briefly describes every variable in the dataset. If the dataset is large, with a lot of variables, it consists of a separate file, .txt or .doc. If the dataset is relatively small, it can be included in the long form project report.	**2.5 points:** A codebook is included, but is incomplete and only describes a few of the relevant variables, or doesn't provide enough information about the data.	**0 points:** No codebook is included with the project.	5	

5	**Narration throughout**	**5 points:** There is written text all throughout the EDA section, narrating and describing the analysis and plots/tables. Every result or plot included in the report should have some text accompanying it and interpreting it.	**2.5 points:** Some written text/narration is present, but either not all plots and tables are described, or the descriptions are short and mostly insufficient.	**0 points:** Very little text is present, or none. Most plots are not described/interpreted.	5	
5	**Missing data discussed/addressed**	**5 points:** The amount and pattern of missing data in the dataset is described. Missing data is handled appropriately. If there is no missing data, the report should specifically mention this.	**2.5 points:** Missing data is mentioned, but either not handled or not handled appropriately.	**0 points:** Missing data is not mentioned.	5	
5	**Data split into training and testing**	**5 points:** Data are split into training and testing sets, with an appropriate proportion in each; most of the data should be used for training (at least 70%; 60% is fine if the dataset is very large (say, over 100k observations).	**2.5 points:** Data are split into training and testing sets, but the proportions are not ideal -- only 50% or less is used for training.	**0 points:** Data are not split into training and testing sets.	5	
5	**Stratified sampling**	**5 points:** When the data are split, they are stratified on the outcome variable. This is done both for the training and testing split and for resampling (i.e. cross-validation).	**2.5 points:** The data are stratified on another variable (not the outcome) without reason, or they are only stratified for the initial split and not for resampling.	**0 points:** The data are not stratified when split or resampled at all.	5	
5	**Recipe specified appropriately**	**5 points:** The model is set up appropriately; the outcome variable is a function of the predictor variables. Any categorical features are dummy- or one-hot encoded. Features are centered and scaled, at least for those models that require centering and scaling.	**2.5 points:** A model is set up, but features aren't handled appropriately -- one-hot encoding is used with a linear model, for example, or scaling is not done before elastic net, etc.	**0 points:** There are glaring problems with the model setup and/or feature engineering.	5	

5	**Cross-validation**	**5 points:** *k*-fold cross-validation is correctly implemented on the training dataset and used to fit models.	**2.5 points:** *k*-fold cross-validation is implemented, but not subsequently used, or another resampling method is used (i.e. validation set) without good reason.	**0 points:** No resampling is done.	5	
10	**At least 4 model types fit**	**10 points:** Four or more types of model are fit. These can consist of any of the following: Standard linear or logistic regression, LDA, QDA, elastic net, pruned decision tree, random forest, gradient boosted tree, support vector machines, k-means clustering, hierarchical clustering, PCA, or a neural network. For projects dealing with complicated data structures (like image recognition) or if you received prior approval, you can fit other model types instead. The correct models should be fit for the problem; i.e., classification models should not be used for regression problems and vice versa.	**5 points:** Three model types were fit.	**0 points:** Two or fewer model types were fit.	10	
10	**Models tuned**	**10 points:** Any model types fit that can be tuned are tuned with cross-validation. Reasonable ranges are used for relevant hyperparameters. Results of tuning are presented either with plots or tables.	**5 points:** Models were tuned but no results were presented, or not all models were tuned, or not all relevant hyperparameters were tuned (eg., not tuning *m* for a random forest, or not tuning learning rate for a boosted tree).	**0 points:** Models were not tuned.	10	
5	**Best one or two models used to predict testing set**	**5 points:** Only the best one or two models are used to predict the testing set. Their testing error is reported.	**2.5 points:** Three models were used to predict the testing set.	**0 points:** All models were used to predict the testing set.	5	

5	**Conclusion section**	**5 points:** Minimum of two paragraphs with complete sentences. Should summarize the main results of the paper, discuss which models did best, and briefly mention possibilities for future models/research.	**2.5 points:** One paragraph, brief summary of project. Lacking in at least one element.	**0 points:** One or two sentences only, or no conclusion at all.	5	
10	**Overall quality: Written like a paper**	**10 points:** There is written text all throughout the report, narrating and describing the analysis and plots/tables. Every result or plot included in the report should have some text accompanying it and interpreting it. The report flows well and can be read smoothly, like a paper.	**5 points:** Some text is included, and most plots or results have some description, but the description is sparse or incomplete; or all the text is written as comments within code chunks or headers (i.e. improper formatting).	**0 points:** There is very little text; most of the report is code.	10	