Discuss with the group

What business question to solve by using data??? (Keep it simple, then add onto it.)
Can we really answer our question, even partially, by using data???
What is our target variable???
If we are to proceed, what help/value will it add to the business???
Can we get it done in time???
Where to find the data???
    How much you trust this data? How was it collected? Is it the right data to your question?
Is it a classification or regression???
    (binning continuous variable will make it a discrete and a classification problem)

50/50

Completed 100% of requirements
- API usage for numerical and text data (optional)
  - Create .env and .gitignore
- Web scraping for data (optional)
- Using AWS SageMaker and its built-in methods (optional)
- Data cleaning and manipulation steps (pre-processing)
  - Check missingness and fill (zero, nan, mean, median, mode, knn imputer, sklearn imputer, predict fill, fancyimpute, missingno library for visualizing)
  - Modify dtypes
  - Remove unwanted text/characters/notations etc.
  - Check interaction and polynomials terms (optional)
    - https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html
  - Check imbalancedness
  - Check outliers
  - Check stationarity, autocorrelation, heteroskedasticity of time series
  - Scaling https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing
  - Dummify/label encoding
  - Extract m/h/d/w/m/y from timestamp
- Apply dimensionality reduction (optional)
- Apply clustering (optional)
- Apply penalized (regularization) models (optional)
- NLP sentiment/tone/ngrams/wordcloud/NER/topic modeling analysis (optional)
- 7 ML and 1 DL model (requirement is 2 ML covered + 1 ML uncovered in lectures) https://scikit-learn.org/stable/index.html
  - Evaluation metrics of all models + 1 new metric uncovered in lectures
    - Regression: Check AIC, BIC, p value, t stat, corr, coef, adj R2, RMSE
    - Classification: Check confusion matrix, Classification report, F1, accuracy score, AUC
    - https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
  - Apply K-fold cross validation (optional)
  - Apply gridsearch and randomsearch (optional)
  - Compare feature importance of all models and make sure the features selected by algo make sense
  - Check in-sample (train error) vs out-sample (test error) for overfitting/underfitting
  - Visualizations of 5 or more
    - try to create visuals in Tableau Public as well (optional)
- ML model comparison – comparing the results of all models used
- Dashboard or PPT (dashboard is preferred)
- Check it with RapidMiner Studio Educational software (try raw data and pre-processed data separately) (optional)
  - https://rapidminer.com/educational-program/

Code runs without error
Code is easy to follow
Code is concise (create functions and pipelines when necessary)
- https://python.hotexamples.com/examples/peval.core.function/Function/-/python-function-class-examples.html
- https://python.hotexamples.com/examples/sklearn.pipeline/Pipeline/-/python-pipeline-class-examples.html

10/10

Variable names are short and meaningful
Necessary files/APIs are loaded into the code without a problem

10/10

Commits via command line (terminal on mac, git bash on windows)
Appropriate commit messages
- For example:
  - git commit -m “updating the function that pulls the data”

10/10

Code is commented/explained with relevant notes inside notebook on top of the code sections/cells (.ipynb)
- For example:
  - # this code does randomly generate a radius size between 5 and 20

10/10

Presentation is successful. Good storytelling – 5min per person
- Core message / hypothesis
- Business question / Motivation
- Finding data
- Explain ML model choosing process
- Briefly explaining analysis and tuning process
- Explain metrics for model evaluation
- Explain graphs
- Explain findings and predictions – numerical summary
- Implications of your findings
- Limitations
- Conclusion

Q&A is successful
Presented in time

10/10

README
- Intro (business question, motivation)
- Data pre-processing/gathering steps
- Visuals and explanations
  - GIF and other image formats
- Model choosing process
- Model tuning and training process
- Model evaluation and metrics
- Model summary of predictive analysis
- Additional explanations
- Major findings
- What others published and found, if any
- Limitations
- Conclusion
- References (mention anything that you get the help from!!! Plagiarism is not tolerated!!!)
- Team Members

Notes: