# Format

The report should usually be between 30 and 60 pages long (and no shorter than 20 pages) and contain:

**BASIC CRITERIA**

* Understanding of the problem
* Completion of the project ("Completion" covers achievement of the original objectives, achievement of modified objectives, or providing convincing evidence that the objectives are unachievable.)
* Quality of the work
* Quality of the report

**ADDITIONAL CRITERIA**

* Knowledge of the literature
* Critical evaluation of previous work
* Critical evaluation of own work
* Justification of the design decisions
* Solution of any conceptual problems
* Amount of work

**EXCEPTIONAL CRITERIA**

* Evidence of originality
* Outstanding scholarship and/or publishable research

The project involves both the application of skills learnt in the past and the acquisition of new skills, on a substantial piece of independent work. The types of skill required vary from project to project, but six main areas of work are required:

* gathering and understanding background information;
* solving conceptual problems;
* design;
* implementation;
* experimentation and evaluation;
* writing up.

# Content

**Aim**: The project involves both the application of skills learnt in the past and the acquisition of new skills, on a substantial piece of independent work.

**Objective**: Build a consumer facing product/app that accomplishes the one or more of following:

1. Predict quarterly result from social media, technical and fundamental analysis, industry macroeconomic data, and news on the same day
2. Allows user to “watch” specific stock’s general sentiments and alerts of important news which may affect stock price
3. Self-learning model for each “Watch” stock after each quarter
4. Tracking earning release?
5. Analyst consensus price target performance, sentiment on stocks buy neutral and sell. Train a ml model by checking if the price target meet expectation and use it to predict the chances of the stock meeting each experts. Revenue estimates, EPS, etc… and build a profile for each analyst
6. Chance events/news to affect stock index and individual stocks overall.
7. Candlestick Patterns
8. If certain (undervalued) companies outperformed analyst expectation in the past
9. Attractive industry to invest in?
10. Industry-wide performance, industry outperformers and underperformers.
11. Industry growth in market cap?
12. Simulated Trading return?

**Experiments**

* Correlations between attributes, industry, and other?
* Associate % increase with news
* Sentiment analysis: positive, neutral, and negative with sources to allow users to look in more depth. (Microblog aggregator)
* Trailing 12 months vs solely quarterly results
* Models
  + SVM
  + Naïve Bayes
  + LSTM for time series

**Project consists of 3 parts**

* Cloud scripts: Where Magic happens
  + Twitter and microblog data retrieval
    - Various sources: investing.com, yahoo finance, Twitter, etc…
  + Pricing retrieval
  + Compute machine learning models
* Visualisation: Android app/webapp:
  + General sentiment patterns for successful stocks?
  + user study to test the usability of the interface.

**2 modes**

* Real-time
* Past-mode using data from StockNets

**Dataset Industry (Categorical ML)**

[**https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/**](https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/)

[**https://blog.myyellowroad.com/using-categorical-data-in-machine-learning-with-python-from-dummy-variables-to-deep-category-66041f734512**](https://blog.myyellowroad.com/using-categorical-data-in-machine-learning-with-python-from-dummy-variables-to-deep-category-66041f734512)

* Accommodation and Food Services
* Administrative and Support and Waste Management and Remediation Services
* Agriculture, Forestry, Fishing and Hunting
* Arts, Entertainment, and Recreation
* Construction
* Educational Services
* Finance and Insurance
* Health Care and Social Assistance
* Information
* Manufacturing
* Mining, Quarrying, and Oil and Gas Extraction
* Other Services (except Public Administration)
* Professional, Scientific, and Technical Services
* Real Estate and Rental and Leasing
* Retail Trade
* Transportation and Warehousing
* Utilities
* Wholesale Trade

**Dataset Features**

* Common Shares Outstanding
* Avg. Basic Shares Outstanding
* Avg. Diluted Shares Outstanding
* Revenues
* COGS
* SG&A
* R&D
* EBIT
* EBITDA
* Interest expense, net
* Abnormal Gains/Losses
* Income Taxes
* Net Income from Discontinued Op.
* Net Profit
* Dividends
* Cash and Cash Equivalents
* Receivables
* Current Assets
* Net PP&E
* Intangible Assets
* Goodwill
* Total Noncurrent Assets
* Total Assets
* Short term debt
* Accounts Payable
* Current Liabilities
* Long Term Debt
* Total Noncurrent Liabilities
* Total Liabilities
* Preferred Equity
* Share Capital
* Treasury Stock
* Retained Earnings
* Equity Before Minorities
* Minorities
* Total Equity
* Depreciation & Amortisation
* Change in Working Capital
* Cash From Operating Activities
* Net Change in PP&E & Intangibles
* Cash From Investing Activities
* Cash From Financing Activities
* Net Change in Cash
* Free Cash Flow
* Gross Margin
* Operating Margin
* Net Profit Margin
* Return on Equity
* Return on Assets
* Current Ratio
* Liabilities to Equity Ratio
* Debt to Assets Ratio
* EV / EBITDA
* EV / Sales
* Book to Market
* Operating Income / EV
* Market Capitalisation
* Enterprise Value

**Data source:**

* Compete:
  + <https://github.com/FinnZC/Data-Analysis/blob/master/stocker>
  + <https://www.quantinsti.com/blog/trading-using-machine-learning-python-svm-support-vector-machine/>
* Microblog/news
  + advanced google search by dates to find microblog data
  + investing.com
  + Many others
  + Kaggle (<https://www.kaggle.com/aaron7sun/stocknews/home> )
  + Google dataset search
  + Zacks (1000+$)
* Social Media
  + Comments from investment websites?
  + Twitter (requires Enterprise edition for past full historical data = cost)
* Technical analysis - Stocks prices (short to medium term approach to investing)
  + Real time
    - Google Finance API
    - Yahoo Finance API
    - [Investopedia API](https://www.investopedia.com/markets/api/partial/historical/?Symbol=GOOG&Type=Historical+Prices&Timeframe=Daily&StartDate=jan+28%2C+2000&EndDate=Dec+05%2C+2018)
  + Past bulk:
    - <http://www.usfundamentals.com/#section-demo>
      * 12,129 companies
      * 8,526 unique indicators
      * ~20 indicators comparable across most companies
      * Daily updates” (only up to 2017 July)
      * Five years of data, yearly and quarterly (including Q4)
      * Free and open data
  + Premium
    - Zacks (1000+$)
    - XIgnite (1000+$)
* Fundamental analysis – income statement, balance sheet, and cash flow statement and more (long term approach to investing)
  + Strategic reports?
  + <http://www.usfundamentals.com/#section-demo>
  + <https://simfin.com/>
  + [www.estimize.com](http://www.estimize.com)
  + Zacks (1000+$)
  + Intrinio (200+$)
  + XIgnite (1000+$)
* Various analyst past estimates
  + Thomson Reuters I/B/E/S Estimates (Cost? I requested details from them)
  + Bloomberg Premium (data about each analyst estimates)
  + <https://intrinio.com/bulk-financial-data-downloads> (1000$+)
  + Zacks (1000+$)
* Industry data (macroeconomics)
  + <https://www.ceicdata.com/en/products/datafeeds> (expensive, used by investment banks)
  + <http://siblisresearch.com/>
  + https://quant.stackexchange.com/questions/141/what-data-sources-are-available-online

Techniques

* HMM
* Deep Learning
* Convolutional NN
* Unsupervised
* Supervised
* LSTM vs Feedforward network:
  + Feedforward network’s outputs are independent of each other thus cannot capture the dependencies
* Time series forecasting
  + Univariate vs [**Multivariate Time Series**](https://machinelearningmastery.com/time-series-forecasting-supervised-learning/)
  + Baseline model:
    - Univariate Persistence Model Forecast
  + ARIMA (Autoregressive Integrated Moving Average)
  + Sliding Window
  + Vector Auto Regression(multivariate)
  + Single-step vs Multi-Step Forecasting -  sliding window method.
* There is no model type that is always best. The two general types that are currently winning (expect changes as better techniques are developed) competitions are:
  + Deep learning. Neural networks with several to many layers. Look at TensorFlow and the associated Keras libraries.
  + Boosted trees. Ensembles of many (often thousands) of individual trees, managed by the model to adjust parameters to identify correct decisions. Look at XGBoost or CatBoost.
* Retraining
  + <https://datascience.stackexchange.com/questions/12761/should-a-model-be-re-trained-if-new-observations-are-available>
* Imputer to deal with missing values sklearn.preprocessing.Imputer
* Epoch explanation: <https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9>
* .dropna <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.dropna.html>

Interesting Reading

* <https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/projects/MIT15_097S12_proj2.pdf>
* <http://yaroslavvb.com/papers/dhar-comparison.pdf>
* <http://cs229.stanford.edu/proj2016spr/report/049.pdf>
* <https://www.euclidean.com/better-than-human-forecasts/>
* <http://becomeabetterinvestor.net/blog/are-financial-analysts-earnings-forecasts-accurate/>
* <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/avoiding-the-consensus-earnings-trap>
* <http://yaroslavvb.com/papers/dhar-comparison.pdf>
* <https://www.bloomberg.com/company/announcements/can-machine-learning-predict-a-hit-or-miss-on-estimated-earnings/>
* <https://www.cass.city.ac.uk/__data/assets/pdf_file/0003/236118/The-usefulness-of-analyst-forecasts.pdf>
* <https://www.quora.com/Can-machine-learning-algorithms-models-predict-the-stock-prices-If-yes-which-are-the-best-machine-learning-algorithm-models-to-predict-the-stock-prices>
* <http://magicx.my/ijidm/wp-content/uploads/2013-1-3-10-leslie.pdf>
* <https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877>
* Twitter
  + <https://www.sciencedirect.com/science/article/pii/S0167923610001521>
  + <http://www.aclweb.org/anthology/P13-2005>
  + <https://www.sciencedirect.com/science/article/pii/S0306457309000478/pdfft?md5=b35a6dd03b296b9823a8bba579250ddf&pid=1-s2.0-S0306457309000478-main.pdf>
  + <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>
  + <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>
* Stock prediction from financial microblog
  + <https://link.springer.com/content/pdf/10.1007%2Fs12204-017-1818-4.pdf>
* Return
  + <https://ac.els-cdn.com/S0306457309000478/1-s2.0-S0306457309000478-main.pdf?_tid=52ac6828-0796-464f-bef9-df32ca3cb57b&acdnat=1538586707_87fdfa1524279235532e93d4980ca689>
* Financial Microblog
  + <https://link.springer.com/content/pdf/10.1007%2Fs12204-017-1818-4.pdf>
  + <https://www.sciencedirect.com/science/article/pii/S0306457309000478/pdfft?md5=b35a6dd03b296b9823a8bba579250ddf&pid=1-s2.0-S0306457309000478-main.pdf>

**Looking Ahead -** [**LINK**](https://www.investopedia.com/news/12-stocks-may-plunge-investors-harvest-tax-losses/?utm_campaign=rss_headlines&utm_source=rss_www&utm_medium=referral)

Morgan Stanley cautions that they have been using sell side analyst ratings as a proxy for [buy side](https://www.investopedia.com/terms/b/buyside.asp) ownership and sentiment since the former data is "publicly available and easily accessible," while the latter, especially when it comes to judging investment manager sentiment, is not. Also, they note that "**stock prices are ultimately driven by**[**fundamental**](https://www.investopedia.com/terms/f/fundamentals.asp)**forces which, over a quarter, can overcome**[**technical**](https://www.investopedia.com/terms/t/technicalanalysis.asp)**selling pressure of this nature,**" hence the fact that their screen has a "hit rate" of less than 50%. Nonetheless, the stocks passing Morgan Stanley's screens may indeed plunge further, based on such fundamental factors, even if fund managers do not sell them in order to manage their [realized capital gains](https://www.investopedia.com/terms/r/realizedprofit.asp) downward.