THE UNIVERSITY of ADELAIDE

Final Project

Background.

You are starting a new job as a Business Analyst at AQR Asset Management, a global investment management firm focused on quantitative investment strategies. Your first task is to analyse the performance of US-listed securities during the COVID-19 market crash. You have several datasets on US securities (stocks and ETFs), and you should help your employer with the following tasks:

(1) Exploratory data analysis, visualisations, descriptive statistics

Examine the dataset *Stock_data_part1.xlsx*. Characterise the performance of US securities between February 14, 2020, and March 20, 2020. You can use variables such as returns, volatility, Sharpe ratio, bid-ask spread, and dollar volume. Please, compute descriptive statistics and provide the relevant visualisations. Based on your analysis, comment on the effect of the COVID-19 pandemic on securities' performance. In your analysis, you should present the following output:

- A table with descriptive statistics (mean, median, 25^{th} percentile, 75^{th} percentile, standard deviation, min, max) of the key variables of interest
- The time series of the key variables of interest

(2) Supervised Machine Learning – OLS regressions

For each security, compute returns between February 14, 2020, and March 20, 2020, and run regression analysis to help explain returns. You may use any variables in the existing dataset, or variables outside the provided datasets (from external sources). Please, make sure to split the dataset you use for analysis into training and testing samples, and comment on the model accuracy. In your analysis, you should present the following output:

- Regression coefficients from one or several models
- Model evaluation metrics: MSE, RMSE, MAE
- Your assessment of which factors are most important for explaining the price change between February 14, 2020, and March 20, 2020

(3) Supervised Machine Learning – Probit regressions

Use the same dataset as in part (2), and introduce a dummy variable for whether a given security increased in price between February 14, 2020, and March 20, 2020. Model the probability of a price increase using any continuous or categorical variables you find relevant. You may use any variables in the existing dataset, or variables outside the provided datasets (from external sources). Please, make sure to split the dataset you use for analysis into training and testing samples, and comment on the model accuracy. In your analysis, you should present the following output:

- Regression coefficients from one or several models
- Model evaluation metrics: accuracy, recall

- Your assessment of which factors characterise the securities that increased in price between February 14, 2020, and March 20, 2020

(4) Unsupervised Machine Learning – K-means clustering

Examine the dataset *Stock_data_part2.xlsx* to perform the k-means clustering. Based on your analysis, you should provide insights about which securities were likely over- vs under-valued as of end of 31/01/2020. You should further examine the temporal stability of clusters and comment on whether those same insights would apply if you were to perform the same analysis as of end of 30/06/2020.

You may combine *Stock_data_part2.xlsx* dataset with any other data (e.g., returns) from the previous exercise, and focus on Valuation Clustering: Securities can be clustered based on valuation metrics like price-to-earnings (P/E) ratio, price-to-book (P/B) ratio, and dividend yields. This aids in identifying undervalued or overvalued securities within the market, enabling informed buy or sell decisions.

In your analysis, you should present the following output:

- Present the optimal number of clusters using e.g., the elbow method, silhouette score. Include visualizations such as an elbow plot or silhouette plot to justify the chosen number of clusters.
- Cluster Characteristics: For each cluster identified, provide a summary of its characteristics in terms of the valuation metrics. This might include the average or median values of P/E, P/B, P/S ratios, etc., within each cluster, helping to identify which clusters represent undervalued or overvalued securities.
- Cluster Sizes: Report the size of each cluster (i.e., the number of securities falling into each cluster).
- Visualizations
  - Cluster Distributions: Provide visualizations such as scatter plots or box plots that show the distribution of valuation metrics within and across clusters. This can help in visually assessing the differences between clusters.
  - Cluster Profiling: Pie charts or bar graphs can be useful to illustrate the composition of each cluster, such as the proportion of stocks from different sectors or the average valuation metrics.
- Interpretation and Analysis:
  - Valuation Insights: Offer insights into what each cluster represents in the context of market valuation. For instance, one cluster might group highly undervalued stocks with low P/E and high dividend yields, while another cluster might contain overvalued stocks with high P/E and low dividend yields.
  - Temporal Stability. Comment on the stability of clusters over time. This can provide insights into how market perceptions of value change.
  - Investment Implications. Comment on which securities you would suggest AQR to invest in on 31/01/2020, and which ones – on 30/06/2020. Do these recommendations differ? Why?

(5) Unsupervised Machine Learning – Principal Component Analysis

The lead asset manager asks you to distil a variety of market-wide and company-specific factors to the core Principal Components. The Principal components will be used to explain monthly returns. You may combine *Stock_data_part3.xlsx* dataset with any other data (e.g., valuation ratios, market returns, Fama-French factors (external), interest rates (external) etc.), and analyse the principal components to understand the factors driving returns.

In your analysis, you should present the following output:

- A detailed summary of the principal components extracted from the PCA, including the amount of variance explained by each component. This should include a scree plot or a table summarizing the eigenvalues and the percentage of variance explained by each principal component, helping to identify the most significant factors affecting monthly returns.
- Factor Loadings: For each principal component identified as significant, provide the factor loadings of the original variables (e.g., valuation ratios, market returns, Fama-French factors, interest rates). This will show how each original variable contributes to the principal components, indicating which factors are most influential in driving returns.
- Factor Interpretation: An interpretation of what each significant principal component represents in the context of market-wide and company-specific factors. For example, the first principal component might be interpreted as overall market risk, while the next components could represent size and value factors, sector exposures, or interest rate sensitivity. This section should bridge the mathematical output of PCA with intuitive financial concepts.
- Implications for Monthly Returns: Discuss how each principal component influences monthly returns. This could involve analyzing how movements in the principal components are associated with changes in stock returns, providing insights into the underlying risk factors or market conditions that impact asset prices.

(6) Analysing experimental evidence using Difference-in-Difference regressions

AQR wants to analyse the effects of the SEC Tick Size Pilot program, and find out how the securities in test groups were affected in terms of bid-ask spread. Use monthly data from *Stock_data_part3.xlsx* and the list of treatment and control securities *(https://www.finra.org/rules-guidance/key-topics/tick-size-pilot-program)* to investigate this question.

In your analysis, you should present the following output:

- A regression specification with detailed explanation of the null hypothesis and alternative hypothesis.
- Regression output and interpretation of coefficients/
- Implications for AQR trading strategy. How should AQR adjust their trading in treated and control stocks to minimise their transaction costs?

(7) Summary

Based on your earlier analysis, draw conclusions for AQR that tackle the following overarching questions:

- What does the COVID-19 data tell us about whether a stock or ETF increase or decrease in price during a market crash?
- What do we know about the risk factors explaining monthly returns?
- How does widening tick sizes affect liquidity in the stock market?

Interim project presentation.

Please, complete Tasks (1) – (5) of the project, and combine your findings in a ppt presentation (to be submitted via MyUni under Assessments).

In class (Week 8 seminar and workshop), you will have 5 minutes to present a subset of your ppt to your colleagues. You are encouraged to coordinate with your team members to select the non-overlapping topics of the ppt. For example, team member 1 may present task 1, team member 2 – task 2 and so on.

This ppt will be graded as an individual assignment.

Final project presentation.

Please, complete Tasks (1) – (7) of the project, and combine your findings in a ppt presentation (to be submitted via MyUni under Assessments). This ppt will be graded as a team assignment..

In class (Week 12 seminar and workshop), your team will have 15-20 minutes to present your findings.

This ppt will be graded as a team assignment. Each team member should submit an identical ppt via MyUni, and the grade will be the same across team members.

Final project report and data.

Please, complete Tasks (1) – (7) of the project, and combine your findings in a project report (to be submitted via MyUni under Assessments). The report should be max 10 pages long (1.5 spacing, Times New Roman), excluding appendices and references, and should address all tasks of the assessment succinctly, with the relevant visualisations, tables, and descriptive statistics.

You should submit your final report, and reproducible code (in Python, R, or SAS) together with datasets as part of this assessment.

This component will be graded as an individual assignment.