PEOPLE ANALYTICS · BUSINESS INTELLIGENCE · STRATEGY

Hilda Soinanto Koitamet

Selected Project Summaries

A focused portfolio of analytical work spanning HR and people analytics, predictive modelling, forecasting, statistical testing, risk, visualisation and strategic decision-making.

Analytical identity I combine HR understanding with statistical and predictive methods to investigate people, performance and organisational decisions.	**Technical range** Python, R, SPSS, Excel, regression, classification, clustering, time series, model evaluation and data storytelling.
Professional focus Responsible analytics that is accurate enough to support decisions and interpretable enough for stakeholders to trust.	**Publication approach** These pages present portfolio-level evidence and critical reflection without reproducing full assessed submissions or academic identifiers.

Projects included

1. Forecasting Global Apple Product Sales with Machine Learning
2. Balancing Interpretability and Accuracy in Employee Attrition Prediction
3. Predicting Global Supply Chain Disruptions
4. Geographic Drivers of Credit Default and Overdue Exposure
5. Lifestyle, Psychosocial and Physiological Predictors of Anxiety
6. Retail Sales Forecasting and Profit Optimisation
7. Molniya Automotive: Four Rounds of Strategic Decision-Making

Contact: soinantohilda@gmail.com | Portfolio support: hello@rdnovasphere.uk

PREDICTIVE ANALYTICS | RETAIL REVENUE

Forecasting Global Apple Product Sales with Machine Learning

I developed and compared predictive models to estimate transaction revenue from a global product-sales dataset containing 11,500 observations and a mixture of pricing, product, customer and sales-channel variables. The project demonstrates my ability to move from raw commercial data to a validated predictive workflow and then translate model output into decisions about pricing, product mix and demand planning.

11,500	3	$R^2 \approx 0.998$
observations	models compared	reported XGBoost fit

Analytical approach

I removed irrelevant identifiers and potential leakage variables, engineered month, year and seasonal features, handled missing values, encoded categorical data and created a reproducible preprocessing pipeline. Linear Regression provided an interpretable baseline, while Random Forest and XGBoost were used to capture nonlinear relationships. Performance was compared using RMSE, MAE and R-squared, alongside predicted-versus-actual plots and feature-importance analysis.

Selected findings

Linear Regression achieved moderate explanatory power, while the ensemble models produced substantially stronger reported fit. XGBoost recorded the strongest reported R-squared value at approximately 0.998, followed closely by Random Forest. Unit price and units sold emerged as dominant predictors of revenue, showing that commercial performance was concentrated around pricing and volume rather than being evenly distributed across all features.

Critical evaluation

The extremely high ensemble-model performance required careful interpretation because revenue was strongly related to units sold and unit price. I therefore treated the results as evidence of model fit within the supplied synthetic dataset, not proof of equivalent performance in a live retail environment. A stronger next stage would use real transaction data, time-aware validation, transformed targets and hyperparameter tuning.

Skills and decision value

The project demonstrates Python-based data preparation, leakage control, feature engineering, regression, ensemble learning, model comparison, diagnostics and business interpretation. It also shows that I do not treat accuracy as the only goal: I examine whether the features and data structure make the reported performance credible.

Portfolio note: This is a curated analytical summary. It preserves the project's methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.

PEOPLE ANALYTICS | EXPLAINABLE AI

Balancing Interpretability and Accuracy in Employee Attrition Prediction

I designed a quantitative people-analytics study to compare transparent statistical modelling with higher-performing machine-learning approaches for employee attrition. The central HR problem is not only predicting who may leave, but explaining why the risk is elevated so that retention action can be fair, credible and operationally useful.

4	10-fold	HR + XAI
models proposed	cross-validation	decision focus

Analytical approach

The proposed study compares Logistic Regression, Random Forest, Gradient Boosting and Support Vector Machine models using anonymised secondary HR datasets. The design includes data cleaning, categorical encoding, class-imbalance treatment with SMOTE, an 80/20 train-test split, ten-fold cross-validation and evaluation through accuracy, precision, recall, F1-score and AUC-ROC. SHAP values and feature-importance consistency are used to assess interpretability, while the DeLong test is proposed for statistical comparison of AUC-ROC performance.

Selected findings

The research framework expects ensemble methods to improve predictive accuracy, while Logistic Regression retains a major transparency advantage. It focuses attention on established workforce variables such as job satisfaction, overtime, monthly income, years at the organisation, work-life balance and distance from home. Importantly, the study does not assume that the most accurate model is automatically the best HR model.

Critical evaluation

People analytics affects real employees, so a black-box prediction without understandable drivers can weaken trust and create governance risk. SHAP can improve post-hoc explanation, but it does not make a complex model naturally transparent. Dataset representativeness, cross-organisation generalisability and the possibility of historical bias must therefore remain visible in model selection and reporting.

Skills and decision value

This project reflects the direction I want to develop most strongly: combining HR knowledge with rigorous analytics. It demonstrates research design, model governance, explainable AI, ethical awareness and the ability to frame technical evaluation around the needs of HR practitioners and organisational leaders.

Portfolio note: This is a curated analytical summary. It preserves the project’s methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.

SUPPLY CHAIN ANALYTICS | CLASSIFICATION

Predicting Global Supply Chain Disruptions

I built a classification workflow to identify whether a shipment experienced a supply-chain disruption and compared Logistic Regression with Random Forest. The project illustrates how analytical models can support early-warning decisions while also exposing the difference between overall accuracy and the ability to detect the minority class that matters operationally.

10,000	94.15%	0.538
shipments	RF accuracy	RF recall

Analytical approach

I prepared 10,000 shipment records, created a binary disruption target from the presence of a disruption event and removed variables that would leak post-event information, including delay days, actual lead time and delivery status. Categorical fields were encoded, numerical inputs were standardised and the data was split using stratification because only 1,267 of the 10,000 observations represented disruptions.

Selected findings

Random Forest achieved 94.15% accuracy, 1.00 precision, 0.538 recall and an F1-score of 0.699. Logistic Regression achieved 94.00% accuracy, 1.00 precision, 0.526 recall and an F1-score of 0.689. Random Forest therefore performed slightly better overall, but both models missed a meaningful proportion of actual disruptions despite their strong headline accuracy.

Critical evaluation

The class imbalance means accuracy alone can give a reassuring but incomplete picture. For an operational early-warning system, false negatives may be more costly than false positives, so recall should be improved through class weighting, resampling, threshold tuning or alternative algorithms. The next stage should also include ROC-AUC or PR-AUC and out-of-time validation.

Skills and decision value

The project demonstrates responsible classification, target construction, leakage prevention, class-imbalance awareness, confusion-matrix interpretation and comparison of precision, recall and F1-score. It shows my ability to challenge headline metrics and focus on the business cost of model errors.

Portfolio note: This is a curated analytical summary. It preserves the project's methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.

BUSINESS STATISTICS | CREDIT & RISK ANALYTICS

Geographic Drivers of Credit Default and Overdue Exposure

I used inferential statistics to investigate whether borrower characteristics were associated with default and whether overdue balances differed across geographic locations. The analysis was designed around decision-oriented questions for credit underwriting, portfolio segmentation and collections prioritisation.

2,052	V = 0.331	p < 0.001
valid cases	location effect	geographic tests

Analytical approach

The analysis used more than 2,000 customer credit and repayment observations. Default status was treated as a binary outcome, overdue amount as a continuous measure of delinquency severity and location as the grouping variable. I applied Chi-square tests of independence to categorical relationships and the Kruskal-Wallis test to compare overdue exposure across multiple locations where normality could not be assumed.

Selected findings

Gender was not significantly associated with default, with $\chi^2(1)=0.435$, $p=0.510$ and a negligible Cramer's V of 0.015. Location, however, was significantly associated with default, $\chi^2(63)=224.246$, $p<0.001$, with a moderate Cramer's V of 0.331. Overdue amounts also differed significantly across locations, $H(63)=172.225$, $p<0.001$, indicating material geographic heterogeneity in repayment behaviour.

Critical evaluation

Some location cells had low expected counts, so the geographic Chi-square result should be interpreted with care and supported by aggregation or follow-up modelling. Statistical significance also does not automatically justify harsher lending decisions. Geographic signals should be combined with individual affordability evidence, fairness checks and ongoing validation to avoid proxy discrimination.

Skills and decision value

The project demonstrates SPSS, hypothesis testing, effect-size interpretation, non-parametric analysis and the translation of statistical evidence into location-sensitive risk monitoring. It also shows how an analytical recommendation can remain commercially useful while recognising fairness and governance concerns.

Portfolio note: This is a curated analytical summary. It preserves the project's methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.

DATA MANAGEMENT & VISUALISATION | R

Lifestyle, Psychosocial and Physiological Predictors of Anxiety

I analysed a dataset of 11,000 observations to examine how stress, sleep, diet, physical activity and other lifestyle or psychosocial factors relate to anxiety levels. The project combines structured data inspection, visual analytics, correlation testing and multiple regression to distinguish the strongest signals from the broader set of potential predictors.

11,000	19	$R^2 = 0.666$
observations	variables	model fit

Analytical approach

Using R, I inspected a 19-variable dataset, confirmed that no values were missing and created descriptive summaries and visualisations including histograms, boxplots, scatter plots and a binned density plot. I then tested correlations and fitted a multiple linear regression model covering stress, sleep, physical activity, diet, caffeine, alcohol, heart rate, recent life events and family history.

Selected findings

Stress showed the strongest positive correlation with anxiety, $r=0.668$, $p<0.001$, while sleep had a substantial negative association, $r=-0.494$, $p<0.001$. Diet quality also had a smaller negative relationship. The regression model explained approximately 66.6% of the variation in anxiety. Stress had the strongest positive coefficient (0.405), while sleep was the strongest negative predictor (-0.531).

Critical evaluation

The analysis identifies association rather than causation, and the discrete anxiety scale produces visible banding in predicted-versus-observed plots. Future work should use longitudinal or experimental data, validate assumptions and consider nonlinear or ordinal modelling. Health-related findings should support research and prevention conversations, not individual diagnosis.

Skills and decision value

This project demonstrates R programming, data-quality checks, exploratory visualisation, correlation, multiple regression and responsible interpretation of sensitive behavioural data. It strengthens my ability to communicate complex multivariable evidence clearly.

Portfolio note: This is a curated analytical summary. It preserves the project's methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.

GROUP CONSULTANCY | FORECASTING & PROFIT OPTIMISATION

Retail Sales Forecasting and Profit Optimisation

As part of Team Value Predictors, I contributed to an end-to-end retail analytics proof of concept using more than 9,000 Global Superstore transactions as a proxy for Ashley Furniture. The project moved beyond a single model by connecting demand forecasting, discount analysis, profit classification, customer segmentation and product strategy into one decision framework.

MAPE 18.97%	MASE 0.43	20%
SARIMA	vs naive	safety-stock rule

Analytical approach

The team cleaned and validated 21 variables, engineered profit margin and converted dates for time-series analysis. Seasonal decomposition, ACF and PACF informed a SARIMA (1,1,1)×(1,1,1,12) model, which was compared with ETS and non-seasonal ARIMA. Multiple Regression quantified discount impact, Decision Tree and Logistic Regression classified high- versus low-profit transactions, K-Means segmented customers and Apriori explored product associations.

Selected findings

SARIMA captured the repeating Q4 demand pattern with MAPE 18.97%, RMSE 4,457 and MASE 0.43, outperforming ETS and a non-seasonal ARIMA model. The forecast projected strong November and December demand, supporting a 20% safety-stock rule. Regression indicated a strongly negative discount coefficient, while the Decision Tree achieved about 90% model accuracy and provided clearer operational rules than Logistic Regression. Customer clustering separated high-value buyers from discount-dependent segments.

Critical evaluation

The work is a proof of concept using public proxy data rather than Ashley Furniture's internal transactions, so the recommendations require live-data validation before operational deployment. The team also distinguished forecast accuracy from business value: an accurate model is only useful when connected to inventory timing, pricing discipline, customer targeting and governance.

Skills and decision value

This project is one of the strongest demonstrations of my analytical range and teamwork. It combines time series, regression, classification, clustering, commercial interpretation and presentation. It also shows how I collaborate in a group to transform multiple model outputs into a coherent strategic recommendation.

Portfolio note: This is a curated analytical summary. It preserves the project's methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.

STRATEGY & OPERATIONS | BUSINESS SIMULATION

Molniya Automotive: Four Rounds of Strategic Decision-Making

I critically evaluated a four-round automotive business simulation covering competitive strategy, finance, production, market share, workforce planning and team governance. Rather than presenting the final profit alone, I examined which decisions created value, where execution risk emerged and how the strategy should change if the simulation were repeated.

£1.281bn	1.85%	25,937
Round 4 profit	market share	unsold Zeus units

Analytical approach

The company pursued differentiation across small and medium vehicle segments, later launching a large hybrid SUV. I tracked production, sales, revenue, post-tax profit, debt, shareholder funds, warranty cost and market share across all four rounds. I also analysed the operational consequences of factory expansion, automation, rapid workforce growth and the team's governance development.

Selected findings

Revenue increased from £1.44bn to £7.439bn and post-tax profit rose from £94m to £1.281bn. Combined market share reached approximately 1.85%, warranty costs fell to 1.1% of revenue and the initial £350m loan was fully repaid by Round 3. However, producing 72,439 Zeus vehicles against sales of 46,502 left 25,937 unsold units, making demand calibration the most important strategic weakness.

Critical evaluation

The analysis shows that aggressive growth created both scale and operational strain. Workforce expansion contributed to persistent industrial action, while confidence from earlier zero-stockout rounds appears to have encouraged an over-ambitious new-model launch. I concluded that the team should have challenged consensus more strongly and launched Zeus at approximately 40,000-50,000 units before scaling.

Skills and decision value

The project demonstrates integrated business analysis, scenario thinking, financial interpretation, operational strategy and reflective evaluation of teamwork. It shows that I can analyse an organisation as a connected system rather than treating finance, HR, production and marketing as isolated functions.

Portfolio note: This is a curated analytical summary. It preserves the project's methods, reported evidence and limitations while excluding full assessment text, student identifiers and administration details.