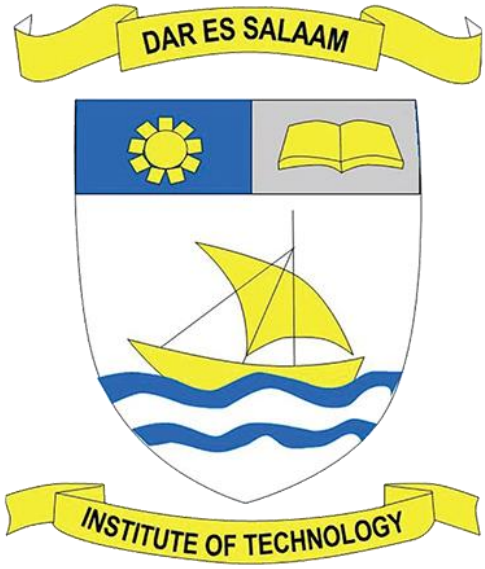
DAR ES SALAAM INSTITUTE OF TECHNOLOGY



DATA MINING AND ANALYTICS

INDIVIDUAL ASSIGNMENT 2

NAME: STANSLAUS R. KAMATTA

REG NO: 230242485951

BENG22COE - 2

Question 1

1.1. The Correlation matrix

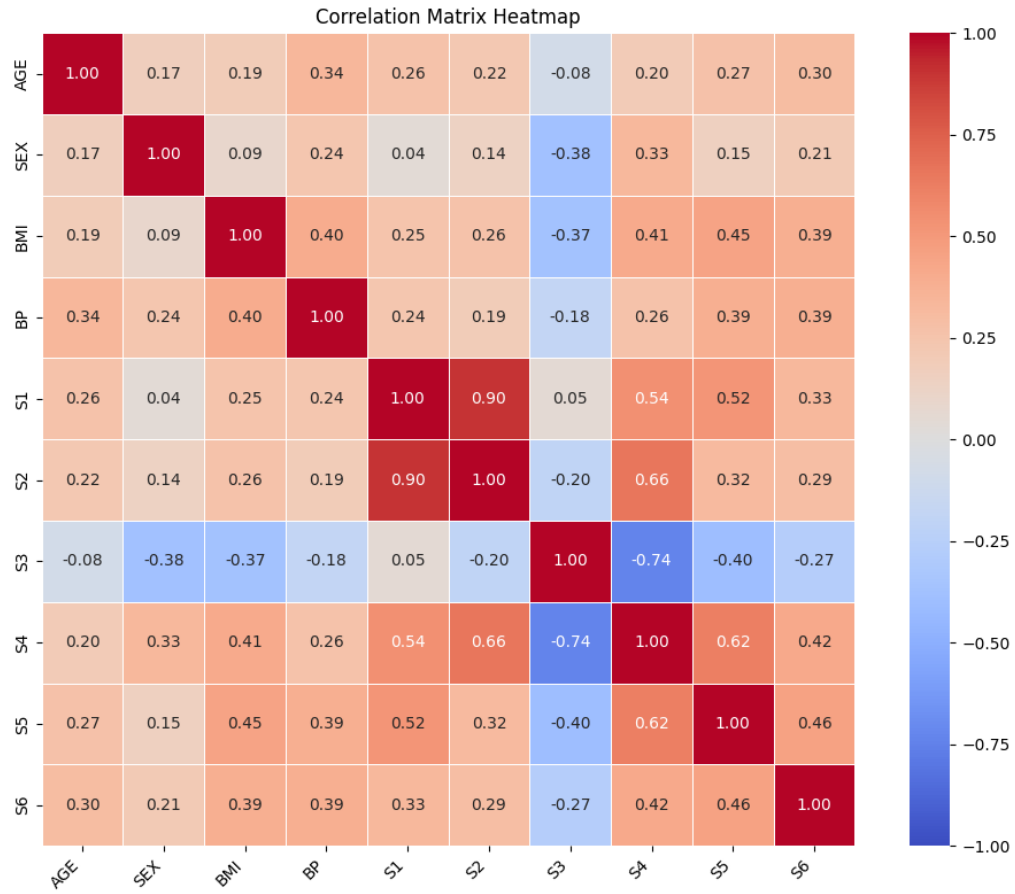


Figure 1: Correlation Matrix Heatmap

The relationships between the variables

1. The strongest positive relationships
   - **S1 and S2 (0.90):** This is the strongest correlation, where by these two serum measurements are nearly identical in their behaviour, suggesting that are highly redundant or clinically linked.
   - **S4 and S2 (0.66),** also **S4 and S5 (0.62)** have significant positive relationship.
   - **S1 and S4 (0.54)** have moderately strong positive correlation.
2. Significant negative relationships

   These variables move in opposite direction, as one increase, the other tends to decrease.
   - **S3 and S4(-0.74)** this is the strongest negative correlation which means they have a very inverse relationship.
3. Physical health indicators
   - **BMI (Body Mass Index):** Shows moderate positive correlations with **S5 (0.45), S4 (0.41),** and **BP (0.40).** This suggest that a higher BMI is often associated with higher blood pressure and specific serum markers.
   - **BP (Blood Pressure):** Shows strongest correlation with **S5** and **S6 (0.39)** and **Age (0.34).**

4. Weakest correlations
    - **AGE:** Age has low correlations with most variables in this dataset. Its strongest link is with Blood Pressure (0.34), but it has almost no relationship with **S3 (-0.08).**
    - **SEX:** Shows very low correlation with **S1 (0.04)** and **BMI (0.09),** indicating that in this sample, these factors don't vary significantly based on gender.

1.2. Collinearity

Collinearity refers to a situation when two (2) independent variables or predictors have a strong linear relationship. Such that the two independent variables are highly correlated and changing one variable can have an impact on the other one.

Effects of collinearity

- The coefficient estimates of independent variables would be very sensitive to the change in the model, even for a tiny change. For example if one independent variable is added or removed, the coefficient estimates then would fluctuate massively. This makes it difficult to understand the influence of each independent variable.

1.3. Multivariate linear model analysis

After fitting the Ordinary Least Squares (OLS) regression model, the performance metrics obtained are as follows:

- Mean Squared Error: 2859.6963
- Adjusted R2: 0.5066

Significance of Variables

- Significant Variables ($p < 0.05$):
  The variables SEX, BMI, BP and S5 (blood resum measurement) are statistically significant. Their *p*-values are all near 0.000, indicating a string linear relationship with disease progression when other variables are held constant.
- Non-Significant Variables ($p \geq 0.05$):
  The variables AGE, S1, S2, S3, S4, and S6 are not statistically significant in this multivariate context. For instance, AGE has a *p*-value of 0.867, and S3 has a *p*-value of 0.635.

Collinearity Assessment

The model displays severe multicollinearity, which significantly impacts the reliability and interpretability of the results. This can be proved by the extremely high Variance Inflation Factors (VIF) for the blood serum measurements. For instance, S1 (576.89), S5 (277.18), and S2 (244.91), which far exceed the standard diagnostic threshold of 10.

1.4. The difference between forward selection and backward selection.

- Forward Selection

Is an iterative feature selection method that start with an empty model and adds variables one at a time. In the first step, it evaluates all available predictors and selects the one that provides the most significant improvement to the model based on a specific criterion, such as lowest *p*-values or the highest increase in R-squared.

This process repeats, adding the next best variable from the remaining pool, until no additional variables meet the predefined significance threshold. The approach is more useful when the number of potential predictors is very large.

- Backward Selection

Also known as Backward Elimination, operates in the opposite direction by starting with a full model containing all potential predictors. It identifies and removes the least significant variable, the one with the highest *p*-value and then re-runs the model with the remaining variables. The cycle continues until every variable left in the model is statistically significant.

It's often considered more robust than forward selection because it assesses variables in the presence of all others. This is especially useful in cases of collinearity, as it allows the model to account for the joint influence of variables before deciding which one is redundant.

1.5. Stepwise forward variable selection and model evaluation

1. How its working.

   The approach works by evaluating the contribution of each variable that is not yet in the model. It seeks to find the "best" subset of predictors by ensuring that every variable included provides a statistically significant improvement to the model's predictive power. This helps in reducing model complexity and mitigating issues like multilinearity by excluding redundant or irrelevant features.

2. The selected variables.

   Using a *p*-value threshold of 0.05, the forward selection process selected the following variables in this specific order: **BMI, S5, BP, S1, SEX, and S2.** The variables AGE, S3, S4, and S6 were excluded because they did not provide a statistically significant improvement to the model once the first six were already present.

3. How the function works.
   The forward selection function follows a systematic loop:
   - It starts with an empty “null” model (intercept only).
   - In each round, it iterates through all variables not currently in the model and fits a separate regression for arch, pairing the existing “best” features with the new candidates.
   - It identifies the candidate variable with the lowest *p*-value (the highest significance).
   - If the *p*-value is below the threshold (e.g., 0.05), the variable is permanently added to the model.
   - The function stops automatically when no remaining variables can meet the significance threshold, ensuring the model remains simple yet effective.
4. MSE and r-squared values.
   The performed metrics for the refined 6-variable model are:
   - Mean Squared Error (MSE): 2876.68
   - R-squared: 0.5149
   - Adjusted R-squared: 0.5082

Question 2

2.1.

Both linear regression and logistic regression use mathematical modeling to predict the value of an output variable (dependent variables) from one or more input variables (independent variables). Where by:

In **Linear regression**, each independent variable has a direct relationship to the dependent variable and has no relationship to the other independent variables. This relationship is known as a linear relationship. The dependent variables is typically a value from a range of continuous values. This is the linear function to create a linear regression model:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots \beta nXn + \varepsilon$$

Where:

- y is the predicted dependent variable
- $\beta 0$ is the y-intercept when all independent input variables equal 0
- $\beta 1X1$ is the regression coefficient (B1) of the first independent variable (X1), the impact value of the first independent variable on the dependent variable
- βnXn is the regression coefficient (Bn) of the last independent variable (Xn), when there are multiple input values
- ε is the model error

An example of linear regression is predicting a house price (dependent variable) based on the number of rooms, neighbourhood, and age (independent variables)

While in **Logistic regression**, the value of the dependent variable is one from a list of finite categories that use binary classification. These are called categorial variables. An example is the outcome from the roll of a six-sided die. This relationship is known as a logistic relationship.

The formula for logistic regression applies a logit transformation, or the natural logarithm of odds, to the probability of success or failure of a particular categorial variable.

$$y = e\text{^}(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots \beta nXn + \varepsilon)/(1 + e\text{^}(\beta 0 + \beta 1 \times 2 + \ldots \beta n \times n + \varepsilon))$$

Where:

- y give the probability of success of the y categorical variables
- e(x) is Euler's number, the inverse of the natural logarithm function is sigmoid function, ln(x)
- $\beta 0, \beta 1X1 \ldots \beta nXn$ have the same meaning as in linear regression.

An example of logistic regression is predicting the chance of a house price being over certain price (dependent variable) based on the number of rooms, neighbourhood, and age (independent variables).

2.2. Survival probability calculation for passenger on the titanic.
The probability of survival for a passenger on the titanic is **0.38197**

2.3. Survival probabilities table broken down by passenger class, gender and age.

Passenger class	Sex	Age Group	Survival Probability
1	Female	Child	0.0
1	Female	Teen	1.0
1	Female	Young Adult	0.98
1	Female	Adult	0.97
1	Female	Senior	0.83
1	Male	Child	1.00
1	Male	Teen	0.50
1	Male	Young Adult	0.43
1	Male	Adult	0.32
1	Male	Senior	0.07
2	Female	Child	1.00
2	Female	Teen	0.88
2	Female	Young Adult	0.90
2	Female	Adult	0.83
2	Female	Senior	0.00
2	Male	Child	1.00
2	Male	Teen	0.00
2	Male	Young Adult	0.11
2	Male	Adult	0.02
2	Male	Senior	0.17
3	Female	Child	0.47
3	Female	Teen	0.61
3	Female	Young Adult	0.46
3	Female	Adult	0.32
3	Female	Senior	1.00
3	Male	Child	0.34
3	Male	Teen	0.08
3	Male	Young Adult	0.18
3	Male	Adult	0.09
3	Male	Senior	0.00

2.4. Logistic regression model for survival rates based on passenger class, sex and age.
The coefficients (parameters) represent the change in the log-odds of survival for a one unit increase in each predictor.
Based on the results, all parameters are statistically significant at the $\alpha = 0.05$ level, as all *p*-values are approximately 0.000.

2.5. Performance of the model, measured by classification accuracy based on confusion matrix.

The model has a classification accuracy of **78.78%.** This means that using only passenger class, sex, and age, the model correctly predicts the survival outcome for approximately 79 out of every 100 passengers.

Question 3

3.1. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used in machine learning and data science for dimensionality reduction. It transforms high-dimensional data into a lower dimensional form while retaining most the original variability in the dataset. PCA is widely used for various applications in machine learning, enhancing both the efficiency and performance of models.

Applications in machine learning

- Dimensionality Reduction
  In many real-world applications, datasets can contain a large number of features. High-dimensional data can be challenging to work with for several reasons, including increased computational cost and the risk of overfitting. PCA helps by reducing the number of dimensions while preserving the essential information.
- Noice Reduction:
  In and dataset, noise (irrelevant or random variations) can obscure the underlying patterns. PCA can be used to filter out this noise, thereby improving the quality of the data and the performance of machine learning models.
- Feature Extraction:
  Feature extraction is crucial for building efficient machine learning models. PCA helps by identifying and isolating the most important features from a large dataset.
- Data Visualization:
  Visualizing high-dimensional data is inherently challenging. PCA reduces the number of dimensions, making it possible to plot the data in two or three dimensions.
- Image Compression:
  Image compression is an important application of PCA, especially when dealing with large image datasets. PCA helps reduce the size of image files while maintaining their quality.

It is useful to consider PCA for transforming a set of explanatory variables because many real-world datasets contain variables that re highly correlated with each other. When such multicollinearity is present, models like linear or logistic regression can become unstable and difficult to interpret. PCA addresses this by transforming the original correlated variables into a new set of uncorrelated principal components, each capturing a different source of variation in the data. This helps stabilize model estimates and ensures that no single piece of information is counted multiple times through correlated predictors.

3.2. Mathematical equations for CPA

1. Mean Centering

   Let the raw input data be represented by the matrix X of size $n \times p$, where n is the number of observations and p is the number of features.

   The first step is to center the data by subtracting the mean of each feature from its respective column. The centered data matrix $\mathbf{Z}$ is given by:

$$\boldsymbol{Z} = X - \mathbf{1}\mu^T$$

   Where:

   μ is a $p \times 1$ vector containing the arithmetic mean of each column in X.

   1 is an $n \times 1$ vector of ones

   Interpretation: Centering ensures that the first principal component passes through the mean of the data, simplifying the variance calculations to a rotation round the origin.

2. Covariance Matrix

   To identify how the variables relate to each another, we calculate the $p \times p$ symmetric covariance matrix **C:**

$$\boldsymbol{C} = \frac{1}{n-1}\boldsymbol{Z}^T\boldsymbol{Z}$$

   Interpretation: The covariance matrix **C** describes the linear relationship between all pairs of features. The diagonal elements C_{ii} represent the variance of feature i, while the off-diagonal elements C_{ij} represent the covariance between features i and j.

3. Eigen Decomposition

   The core of PCA involves finding the "principal directions" of the data. We perform eigen-decomposition on the covariance matrix **C** to find its eigenvalues and eigenvectors:

$$\mathbf{Cv} = \boldsymbol{\lambda}\mathbf{v}$$

   This results in:

   - **Eigenvalues ($\lambda_1, \lambda_2, ..., \lambda_p$):** These indicate the amount of variance captured by each principal component. They are typically sorted such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.
   - **Eigenvectors ($\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p$):** These define the direction of the new axes. These vectors are orthonormal (unit length and perpendicular to each other).

4. The Transformation

   The transformation of the centered data **Z** into the new set of variables (Principal Components) is a projection of the data onto the space spanned by the eigenvectors.

$$\mathbf{Y} = \mathbf{ZW}$$

   Where:

   W is the Feature Vector (or Loading Matrix), a $p \times k$ matrix formed by selecting the first k eigenvectors as columns.

   Y is the Score Matrix, an $n \times k$ matrix representing the data in the new reduced dimensional space.

3.3. Dow Jones constituents' stocks analysis

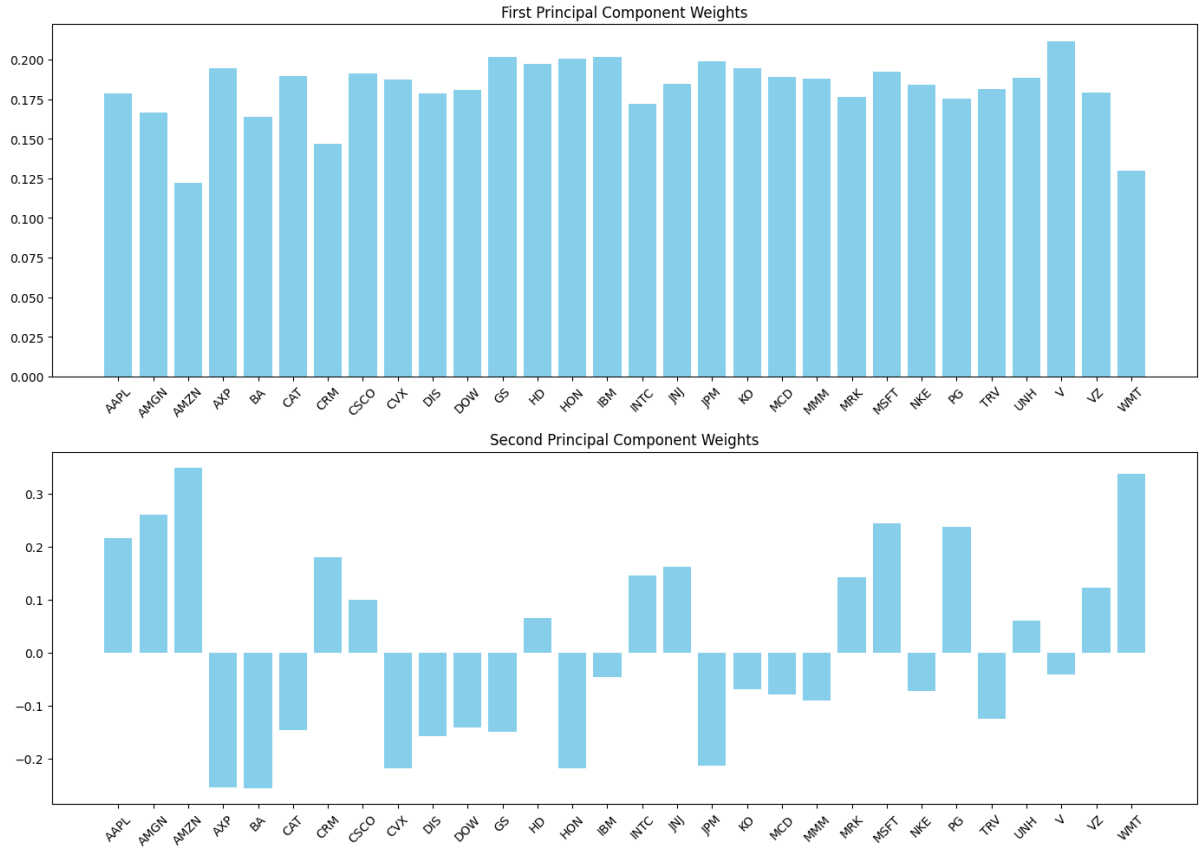


Figure 2: Dow Jones Constituents' stocks weight graph

- First Principal Component (PC1)
  The first Principal Component is very similar to the market (equal-weighted) portfolio. As observed in the PC1 graph, almost all stocks have weights with the same sign (positive) and relatively similar magnitudes.
  Reason: In equity markets, PC1 represent the systematic risk or the general market movement. Since most stocks are positively correlated with the broad market, they will all move in the same direction when the overall market trends.

- Second Principal Component (PC2)
  The second principal component is not similar to the market. As the PC2 graph shows a mix of positive and negative weights (a long-short profile).
  Reason: PC2 represents the first level of specific risk or sector rotation. It captures the divergence between different types of stocks. Tech-heavy stocks (AAPL, AMZN) often have weights of one sign, while cyclical stocks (BA, CVX) have the opposite sign. This component explains how these groups move away from each other, which is opposite of an equal weighted market movement.

3.4. Principal component and variance

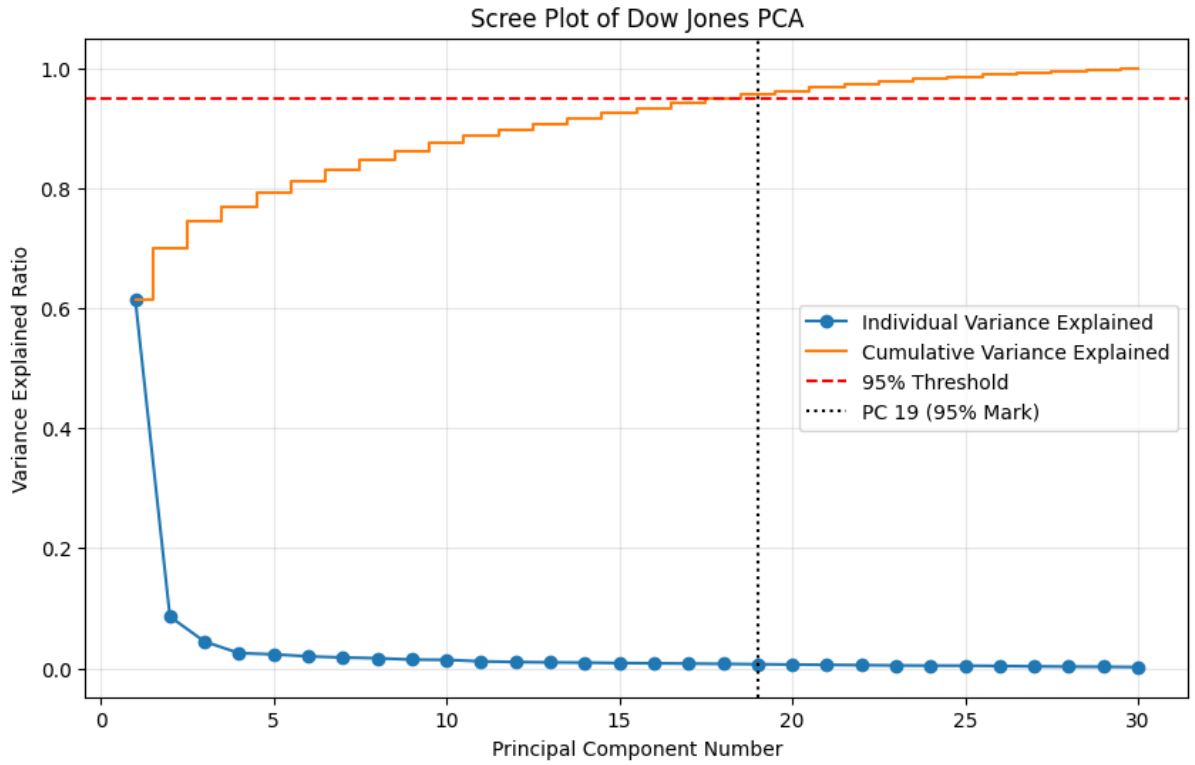


Figure 3: Scree Plot

The number of principal components required for 95% variance is 19

3.5. The three most distant stocks.

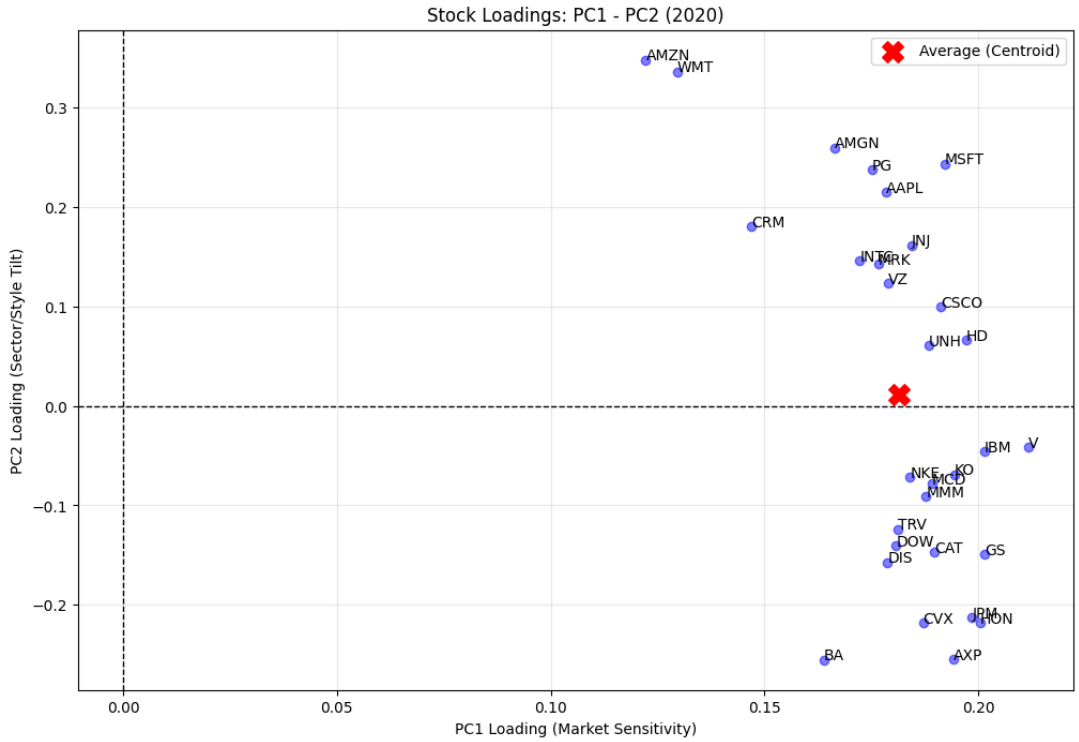


Figure 4: Scatter Plot

The three most distant stocks from the average:

- AMZN (0.341690)
- WMT (0.328941)
- BA (0.267609)

The stocks BA, AMZN, and WMT are identified as the most unusual because they have the largest Euclidean distances from the average of the 30 stocks in the PCA scatter plot. This mathematical distance means their movements were highly independent; while the average stock followed general market trends, these three were driven by extreme, specific factors. For example BA (Boeing) diverged in the opposite direction due to the total halt in global travel and its own internal safety crises during the that period (between 2020 and 2021).