QUESTION 1.1

- S1 and S2 show an extremely high positiv correlation (r = 0.90). S4 also shows strong corelations with S2 (r = 0.66), S3 (r = -0.74), and S5 (r = 0.62)
- BMI has a moderate positive correlation with S5 ($r = 0.45$) and BP ($r = 0.40$).

QUESTION 1.2

Collinearity is the situation where by two or more independent variables in a multiple regresion model are highly linear related.

Effects in which collinearity among predictor features/ variables have on their estimated coefficient value are:

- Small changes in the data can lead to large changes or variation in the estimated coefficient values.
- It lead difficult to determine the individual effect of a specific variable on the dependent variable because its contribution is shared with other correlated variables.

QUESTION 1.3

- The Mean Squared Error for my model is 2859.7
- The adjusted R2 for my model is 0.5066
- Only SEX, BMI, BP and S5 were found to be statistically significant at a 5% significant level while AGE, S2,S3,S4 and S6 were not found not significant and S1 was borderline($p = 0.058$)
- Yes, this model sufers from collinearity. The condition number is very high (7.24×10^3), and several variables (like S2 and S3) that are known to be relevant appear insignificant likely due to their high corelation with other included features (like S1 and S5).

QUESTION 1.4

- Forward Selection Starts with an empty model (no predictors). In each step, it adds the variable that provides thethe most significant improvement to the model (e.g., lowest p-value below a threshold) until no more significant variables can be added while Backward Selection Starts with a full model containing all potential predictors. In each step, it removes the least significant variable (e.g., highest p-value above a threshold) until all remaining variables are statististicaly significant.

QUESTION 1.5

- The stepwise approach is an iterative method for model building. In a Foward Selection (a type of stepwise), variables are added one by one. After each addition, the algorithm evaluates if any other variables should be added. In a bidirectional Stepwise Selection, the algorithm can also remove variables that become insignificant after new ones are added.
- Using forward selection based on p-values (threshold 0.05), the following variables were selected in order:
    a) BMI
    b) S5
    c) BP
    d) S1
    e) SEX
    f) S2

MSE and R2 value for this new model are

- Mean Squred Eror **(**MSE**):** 2876.68
- R2 Value**:** 0.5149

QUESTION 2.1

- Linear regresion used for predicting a continuous dependent variable (e.g., house prices or weight). The output is a real value, and it assumes a linear relationship between input and output while Logistic regression used for classification problems where the dependent variable is categorical (e.g., survived or not survived). It models the probability that an observation belongs to a particular category using the logistic (sigmoid) function. The output is always between 0 and 1.

QUESTION 2.2

Based on the 1,309 pasengers in the dataset, the overall probability of survival was 0.382

QUESTION 2.4

- The parameter estimates are:

    a) Intercept**:** 3.5221
    b) Pclass **(**2nd Class**):** -1.2806
    c) Pclass **(**3rd Class**):** -2.2897
    d) Sex **(**Male**):** -2.4978
    e) Age**:** -0.0344

Significance:

All variables in the model are highly statitically significant ($p < 0.001$ for all parameters). This indicates that class, sex, and age were all strong predictors of suvival. Being male, older, or in a lower class significantly decreased the odds of survival survival.

QUESTION 2.5

The performance of the model was evaluated on the 1,046 passengers with complete data:

- Confusion Matrix**:**

    a) True Negatives (Did not survive): 520
    b) False Positives (Predicted survived, but didn't): 99
    c) False Negtives (Predicted didn't survive, but did): 126
    d) True Positives (Survived): 301

- Classification Accuracy: 78.49%

QUESTION 3.1

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while preserving as much variance (information) as possible. It transforms a large set of potentially correlated variables into a smaller set of uncorrelated variables called Principal Components (PCs).

Applications in Machine Learning:

- Feature Extraction: Reducing the number of input variables to simplify models and reduce training time.
- Data Visualization: Projecting high-dimensional data (e.g., 30 stocks) into 2D or 3D to identify clusters or outliers.
- Noise Reduction: By keeping only components with high variance, PCA filters out low-variance components which often represent noise.
- Preprocessing: Used before clustering or regression to ensure inputs are uncorrelated.

Transforming explanatory variables with PCA is useful because:

a) In finance, stock returns are often highly correlated. PCA creates orthogonal (uncorrelated) components, which solves the problem of multicollinearity in linear models.

b) It allows a model to explain a high percentage of data variance using only a few factors (e.g., a Market factor and a Growth factor) rather than 30 individual stocks.

QUESTION 3.2

The transformation of the raw input data matrix X (of size n x p) involves the following steps:

1. Standardization: The data is centered (subtracting the mean μ) and often scaled (dividing by standard deviation σ):

$$Z = \frac{X-\mu}{\sigma}$$

2. Correlation Matrix: We calculate the correlation matrix C:

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

3. Eigen-Decomposition: We solve for the eigenvalues (Λ) and eigenvectors (V) of C:

$$CV = V\Lambda$$

4. Transformation: The raw data is projected onto the new principal component axes

$$Y = ZV$$

Interpretation of Matrices**:**

- X **(**Raw Input**):** The original data containing daily returns for 30 stocks.
- C **(**Correlation Matrix**):** Represents the linear relationships and co-movements between all pairs of stocks.
- V **(**Weights/Loadings**):** The columns of V are eigenvectors. They act as recipes for the new variables, showing how much weight each stock has in each principal component.
- Λ **(**Eigenvalues**):** A diagonal matrix where each value λ_i represents the amount of variance explained by the i-th principal component.
- **Y (**Principal Components**):** The new set of coordinates for the data points in the tranformed space.

QUESTION 3.3

The first principal component is highly similar to the market (equal weight) because PC1 usualy has weights that are all positive and of similar magnitude across all 30 stocks. This represents the

Market Factor showing that during 2020, most stocks moved together together in response to broad economic news (COVID crash and subsequent recovery).

QUESTION 3.4

There are six (6) to ten (10) principal components are required to explain 95% of the variance

QUESTION 3.5

The three most distant stocks are Amazon(AMZN), Walmart(WMT), Boeing(BA)

These stocks are unusual because their return behavior during the 2020-2021 period difered significantly from that of the average

Amazon **(**AMZN**)**

Amazon exhibits an unusually large distance from the average stock because it benefited strongly from pandemic-driven structural changes in consumer behavior

Walmart **(**WMT**)**

Walmart appears unusual because it is a defensive consumer staples stock with relativel stable cash flows and lower volatility

Boeing **(**BA**)**

Boeing is unusual due to extreme firm specific and sector specific shocks during 2020. The collapse in global air travel, combined with existing aircraft safety and production issues, led to exceptionally high volatility and large negativereturns returns