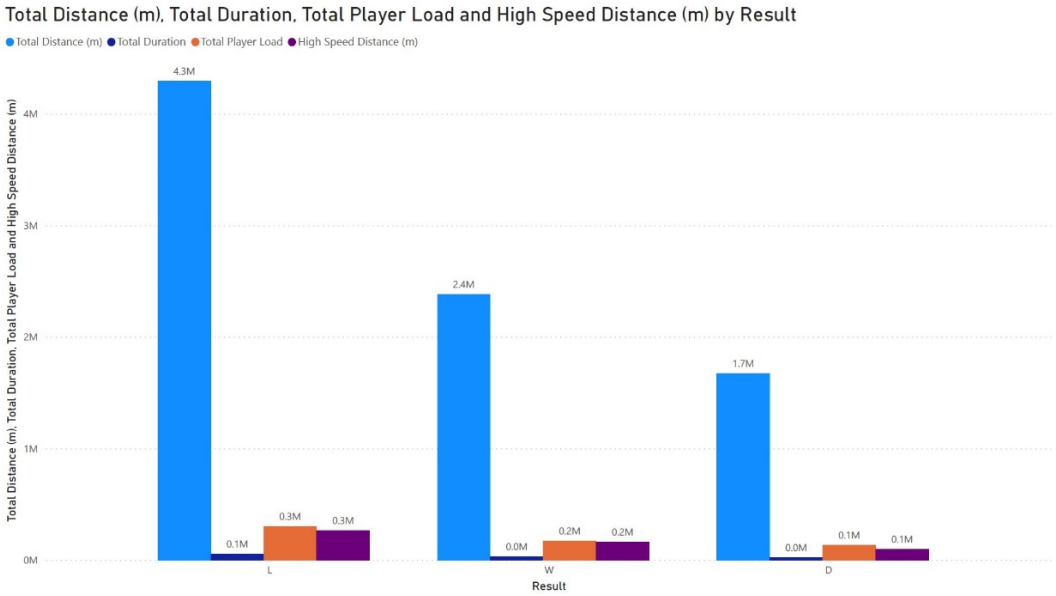
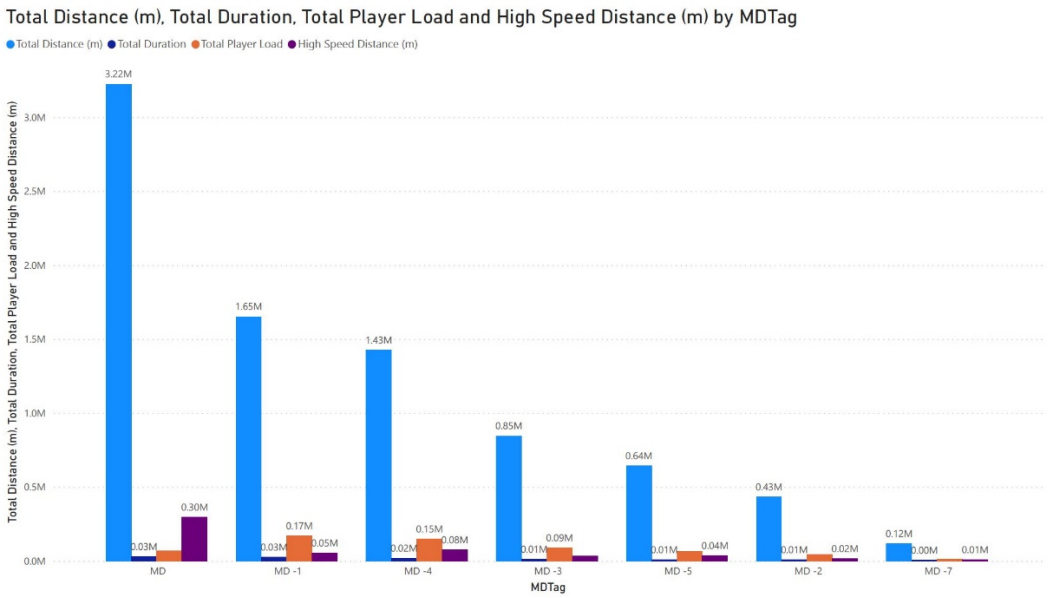
# README_Zhiyuan Yang

From my understanding the goal of the test is to make machine learning models based on different match day tags up to MD-5 so as to maximize the chance of winning. And my assumption is to use logistic regression because the outcome is win or not win.

## Data Visualization using Power BI




- So we have the highest total distance, total duration, total plater load and high speed distance when the result is loss.




- For total distance the closer to the match day the longer the distance.

- For total duration not too much change.
- For total player load the closer to the match day the larger the load but the largest load is in MD-1.
- For high speed distance overall the closer to the match day the longer the distance.

**Data Pre-processing**

In Excel:

- Delete the 'Date' column and the 'Player' column as they won't be used in model building.
- Move the 'Result' column to the right as this will be the dependent variable.
- Reorganize the 'Result' column to only 2 outcomes 'W' and 'NW' (which stands for Not Win, combined by 'L' and 'D').
- Split the Excel file into 6 CSV files based on the 'MDTag' column from MD to MD-5.

In Python:

- Split the dataset into independent variables and dependent variable.
- Encode categorical data 'W' and 'NW' to 1 and 0.
- Split into 80% training and 20% test to avoid overfitting.
- Feature scaling.
- Apply logistic regression models for all 6 files.
- Define the model.
- Calculate model accuracy.
- Generate a more comprehensive report.

**Next Step**

As the model accuracy scores are between 0.68 and 0.83, these models can be improved by setting different parameters, for instance working with the regularization strength C equal to 10.0 instead of the default value of 1.0.