Title: Modeling Flight Delays with Weather Conditions: A Multi-Factor Regression and Causal Analysis

Project Category: Social Science & Policy (e.g., policy evaluation, causal inference, regression, hypothesis testing including Bayesian and constraint-based methods)

Team Information: Xiaoxing Chen(xc252), Jianqiu Li(jl1417), Tianhao Chen(tc442), Ziao Huang(zh212), Yuhang Sun(ys502), Yu Liang (yl1139)

Content:

- **Problem Statement:** Flight delays are a persistent challenge in air transportation that people have been striving to overcome, as they directly affect passenger satisfaction, airline operations, and even national productivity. Although flight delays are usually attributed to weather conditions, the specific roles of different weather parameters such as temperature and precipitation in flight delays still require further explicable exploration. Our project attempts to use multivariate regression to model the impact of various weather conditions on flight delays and determine which factors have the greatest or least impact on delays.
- **Challenges:** One of the primary challenges in this project will be obtaining and integrating reliable flight and weather data, as aligning them across time and location requires careful preprocessing. Additionally, many weather variables are highly correlated, which may lead to multicollinearity and reduce the interpretability of regression results. Another difficulty lies in distinguishing the effects of weather from other confounding factors, such as air traffic congestion which can also cause delays but are not directly weather-related. Handling missing values, extreme outliers, and the sheer scale of large datasets spanning multiple airports and years further complicates the analysis. Finally, effective feature engineering, such as capturing interactions between variables will be critical to building a robust model that can provide meaningful insights.
- **Dataset:**
    - The website provides data of flight delay. It includes data such as number of delays and average delayed time for flights that are delayed more than 10 times per month. More than 30 minutes late flights are considered delayed.
    - The website provides climate-related data.
- **Method/Algorithm:** To investigate the impact of weather on flight delays, we will begin with a multivariate linear regression model as the baseline to quantify how different weather variables contribute to

delays. Since many weather features are correlated, we will incorporate regularization techniques such as Ridge and Lasso regression to reduce multicollinearity and to identify the most influential variables. To capture more complex relationships, we plan to include interaction terms, such as the combined effect of wind speed and low temperature, and explore possible time-lag effects where earlier weather events influence later flights. While regression will be our primary method due to its interpretability, we will also compare results against tree-based models such as Random Forests or Gradient Boosting to evaluate whether non-linear methods provide stronger predictive performance and robustness. This combination of statistical modeling and machine learning approaches will allow us to balance interpretability with predictive accuracy.

- **Literature Review:** Several studies have been conducted to estimate or model flight departure delays using various techniques. Khaksar and Sheikholeslami identified parameters that enable effective estimation of delays with a comparison of Bayesian modeling, decision tree, cluster classification, random forest, and other hybrid methods. They experimented 2,825,647 data for US airlines and 15,428 data for Iranian airlines, achieving an overall accuracy of approximately 70%. Al-Tabbakh et al. analyzed the flight delay patterns using four decision tree classifiers. Their experiment revealed that among the classifiers evaluated for the Egypt Airline dataset, the model EPTree attained the highest accuracy score of 80.3%.
- **Evaluation:** We will evaluate our models using both quantitative metrics and qualitative analysis. On the quantitative side, we will measure model fit with R^2 and adjusted R^2, as well as predictive accuracy using RMSE and MAE. Statistical tests, including p-values and confidence intervals for regression coefficients, will help determine which weather factors significantly affect delays. On the qualitative side, we will generate visualizations such as scatter plots and bar charts to illustrate the relationship between individual weather variables and delays, time-series plots to highlight the impact of extreme weather events, and geographic heatmaps to show airport-specific vulnerabilities.

Project Mentors: Chengkun Yang

Feedback ：

The topic selection should take the completeness of the analysis and whether it can ultimately provide meaningful explanations for real-world problems into account. The analysis of factors causing flight delays is a good option, with diverse and accessible datasets. It also has practical significance and is

relatively easy to find the relation in the early stages. Moreover, the analysis can be extended by incorporating additional datasets and richer sources of information. For example, flight delays are not only driven by weather conditions but may also be influenced by factors such as airport size, economic conditions, and a wide range of other determinants. These are all potential directions worth exploring.

Grading: