Bird Strike Case Study Rubric

DS 4002
Format: RMarkdown and GitHub Repository

General Description: Create an RMarkdown file completing the task set out in the Hook Document and upload everything needed to reproduce your analysis into a GitHub repository.

Why am I doing this? Creating an R Markdown file and GitHub repository for this analysis transcends the specific topic of bird strikes and serves several fundamental purposes in the realm of data science. Firstly, it ensures reproducibility by documenting the analysis process, allowing others to understand and replicate your methodology easily. Secondly, it facilitates collaboration by providing a platform for peers and experts to review your work, offer feedback, and contribute improvements. Lastly, sharing your analysis publicly on GitHub invites engagement from the broader data science community, fostering learning, discussions, and the advancement of knowledge in the field. In essence, creating an R Markdown file and GitHub repository for your analysis promotes transparency, collaboration, and skill development, enhancing your reputation and impact as a data scientist.

What am I going to do? In this analysis, you will be tasked with replicating a comprehensive time series forecasting project focused on bird strike incidents in aviation. The goal is to forecast the number of bird strikes occurring during different times of the day (day, dusk, dawn, and night) over a five-year period. To achieve this, you will start by downloading and preprocessing the dataset, selecting relevant columns, and removing any missing values. Next, you will create a dataset with count values, grouping incidents by year and time of day. After exporting the final dataset, you will visualize the distribution of bird strikes across different times of the day and years using bar plots. Subsequently, you will forecast the number of bird strikes separately for each time of day using the auto.arima function in R. Finally, you will plot the forecasts for day, dusk, dawn, and night bird strikes over the next five years. This analysis will provide valuable insights into the temporal patterns of bird strikes, aiding aviation authorities in implementing effective safety measures.

How will I know I have Succeeded? You will meet expectations on this assignment when you have successfully created a forecast of bird strikes during different times of the day in an R Markdown file and have uploaded all parts of your analysis into a GitHub repository.

R MARKDOWN FILE RUBRIC

Formatting	● R Markdown ● Include: ○ Title ○ Name and Date ○ Restate Hypothesis/Research Question/Model Approach
Learning Goals	● Understanding the process of preprocessing and cleaning a dataset for time series analysis.

	● Familiarizing oneself with time series forecasting techniques, particularly auto.arima in R. ● Gaining proficiency in visualizing time series data and forecasted trends using ggplot2. ● Developing skills in interpreting forecast results and identifying temporal patterns in data.
Outputs	● Exploratory plots and/or other EDA elements ● Forecasts for each time of day ● Optional: Create forecasts for other variables

GitHub Repository Rubric

Formatting	● One Github Repository ● The top level page of the repository should contain: ○ A README.md file (which auto displays) ○ A LICENSE.md file (use MIT as default) ○ A SCRIPTS folder ○ A DATA folder ○ AN OUTPUT folder
README.md	● Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings. ● Use markdown headers to divide content. ● Make an H2 (##) section explaining the contents of the repository ● Section 1: Software and platform section ○ The type(s) of software you used for the project. ○ The names of any add-on packages that need to be installed with the software. ○ The platform (e.g., Windows, Mac, or Linux) you used. ● Section 2: A Map of your documentation. In this section, you should provide an outline or tree illustrating the hierarchy of folders and subfolders contained in your Project Folder, and listing the files stored in each folder or subfolder. ● Section 3: Instructions for reproducing your results. In this section, you should give explicit step-by-step instructions to reproduce the Results of your study. These instructions should be written in straightforward plain English, but they must be concise, but detailed and precise enough, to make it possible for an interested user to reproduce your results without much difficulty.

LICENSE.md	● Goal: This file explains to a visitor the terms under which they may use and cite your repository. ● Selecting the MIT license is appropriate.
SCRIPTS folder	● Goal: This folder contains all the source code for your project. ● Include all the scripts you used. Try to name each script according to the order it needs to be executed to reproduce the results. ● All script files should include header comments at the beginning of a script to provide information that anyone working with or executing the script should be aware of. Throughout all your scripts, you should include copious comments explaining what each command or sequence of commands accomplishes and what the purpose is.
DATA folder	● Goal: This folder contains all of the data for this project. ● You should AT LEAST the data include the initial data, and the final data analyzed. If needed, the code in the SCRIPTS folder should be able to get you from the initial piece of data to the final one. N.B. If the initial and final data are the same, then just include that dataset. ● If your data fits in github, place all of it here. ● If your data does not fit in GitHub use a single file explaining the process to obtain the dataset. ● A Data Appendix file as a PDF, which will include text that you type, as well as tables, figures, and other descriptive statistics.
OUTPUT folder	● Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. ● Importantly, any information like tables, figures shown in your presentation should be here. ● Use informative names for your files.