# Project Title: Predicting the Occurrence of a Data Breach

# Project Author: Mehdi Barati

# Overview

This repository will help the viewers to replicate the predictive models and visualizations of the “Predicting the Occurrence of a Data Breach” project. The project aims to build predictive models for the probability of the occurrence of a data breach and the size of a data breach based on the previous incidents of data breaches and their types. The dataset used for this purpose is downloaded from Privacy Rights Clearinghouse website at <https://privacyrights.org/sites/default/files/2020-01/PRC%20Data%20Breach%20Chronology%20-%201.13.20.csv>. The raw dataset file is also included inside the folders. We used R programming language because it’s open source and best fitted for our purposes.

To replicate the material of the project, the user needs to have R program installed on their device and also have installed the following packages:

(ggplot2), (dplyr), (Rlab), (tidyr), (tidyverse), library(modelr), (lubridate), (stats), (MASS)

There are two folders in this project repository. The folder named “descriptive” is aimed to replicate all the visualizations in exploring the data and bias check sections. The other folder named “model” is used to replicate all the tables and figures in initial and subsequent model sections.

Please download the entire project and before running the R files make sure to set the working directory to the folder the project folder is located.

**Conclusion**

This project shows that building predictive model for the occurrence of a data breach with high accuracy is possible with Autoregressive Poisson and Negative Binomial models. However, accurate models for predicting the size of a data breach is more challenging and needs more complicated models.