**WEBSITE TRAFFIC ANALYSIS**

**INTRODUCTION :**

This project will analyze website traffic data to find out more about user behavior, popular websites, and traffic sources. In order to enhance user experiences and increase performance, it aims to arm website owners with essential information. The analysis's objectives are described in simple terms and contain statistics on page views, unique visitors, and referral sources. To assure the quality and depth of the insights, the data is meticulously obtained. Using IBM Cognos, which converts complex data into easily comprehensible visual representations, data is visualized.

**ABOUT PHASE 4 :**

The fourth step is all about executing different operations on the dataset, such as analysis, exploratory data analysis, and visualization.

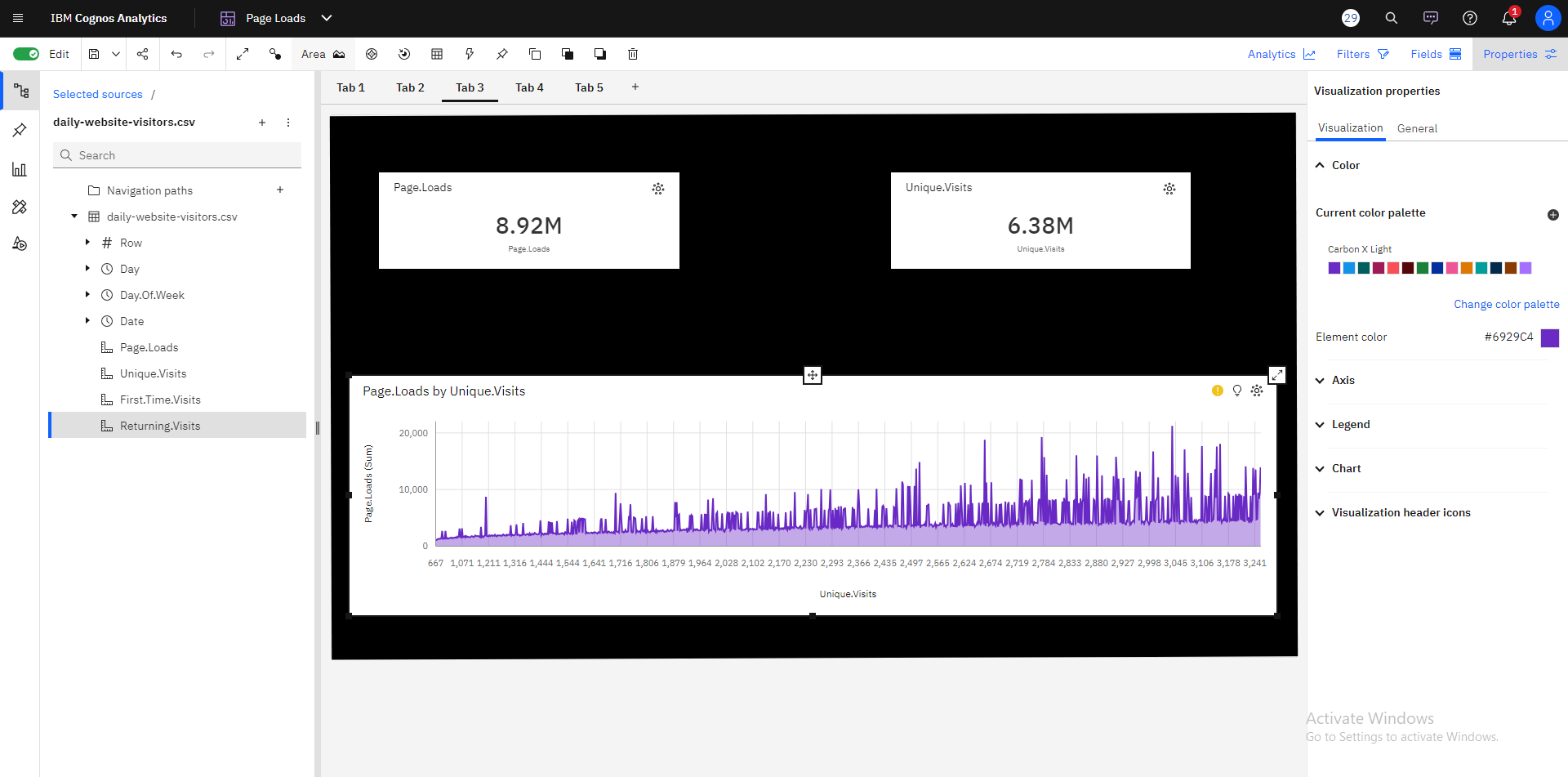
**IBM COGNOS :**

* In order to visualize data from files, including csv files and other files, analysis is

done using the IBM Cognos tool.

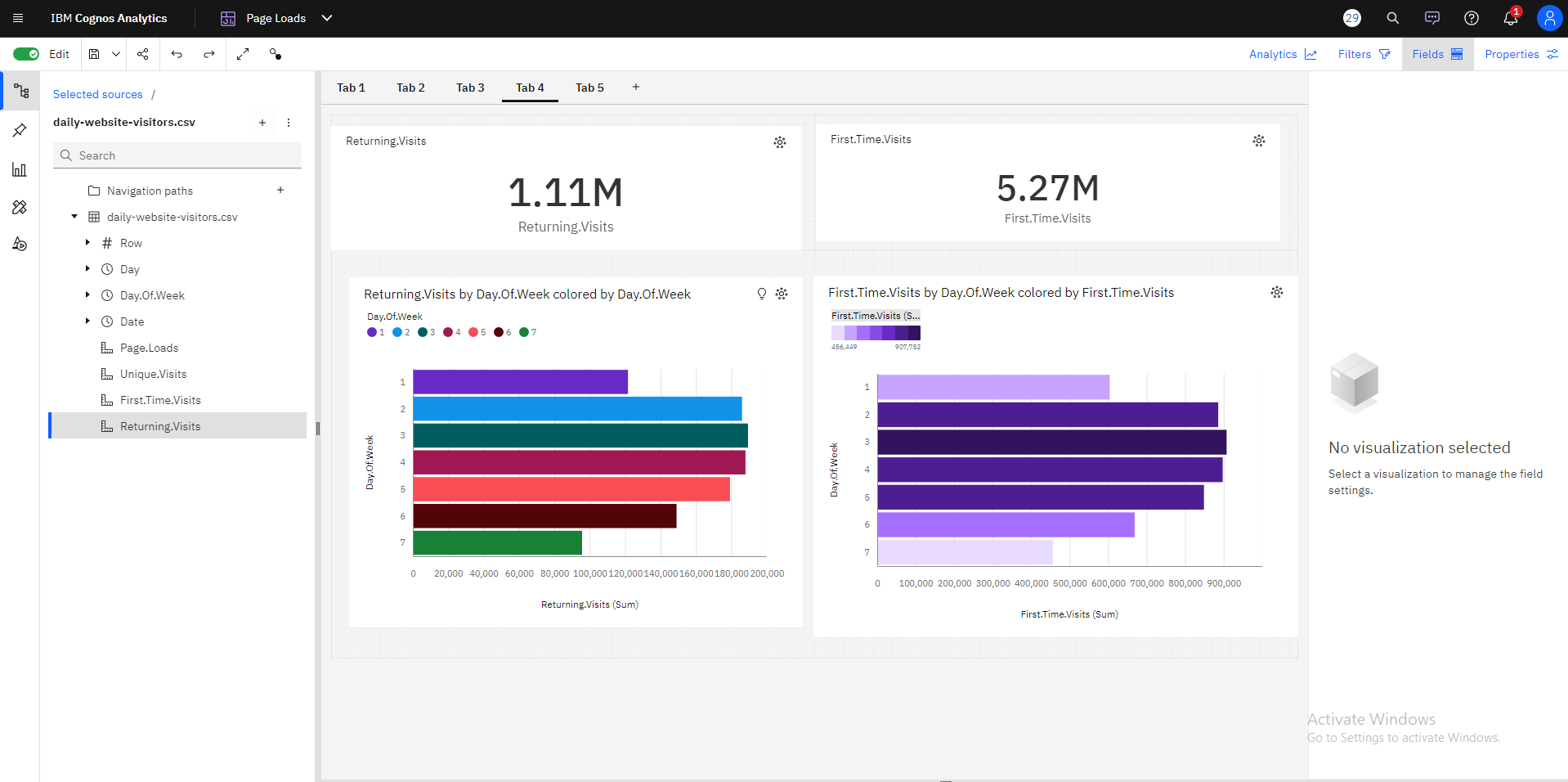
* Line charts, scatter plots are displayed in these various charts.

**LINE CHART :**



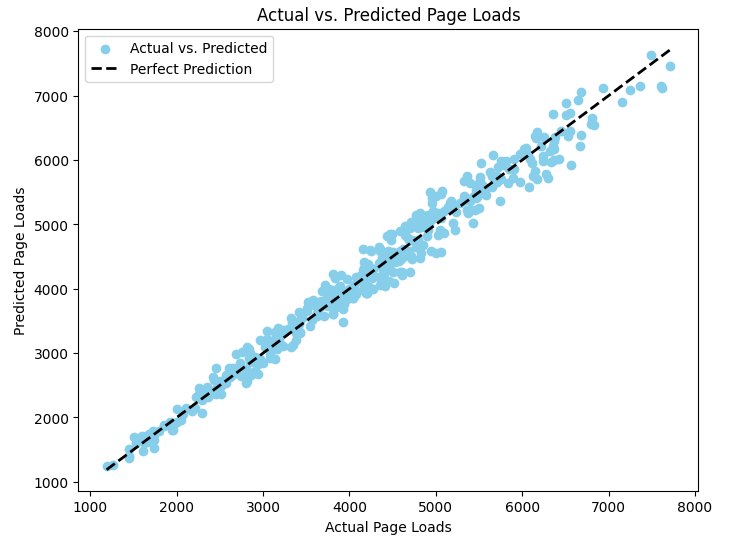
* It is expected that by Monday+1, 2844 will have surpassed 3039 in Page.Loads 410 times.
* The Independent variable and Dependent variable used here are Unique visits and Page Loads(sum) respectively.
* Page.Loads can reach above a thousand when Unique.When Unique, the number of visitors increases from 667 to over 21 thousand.The number of visits is 3039.

**BAR CHART :**

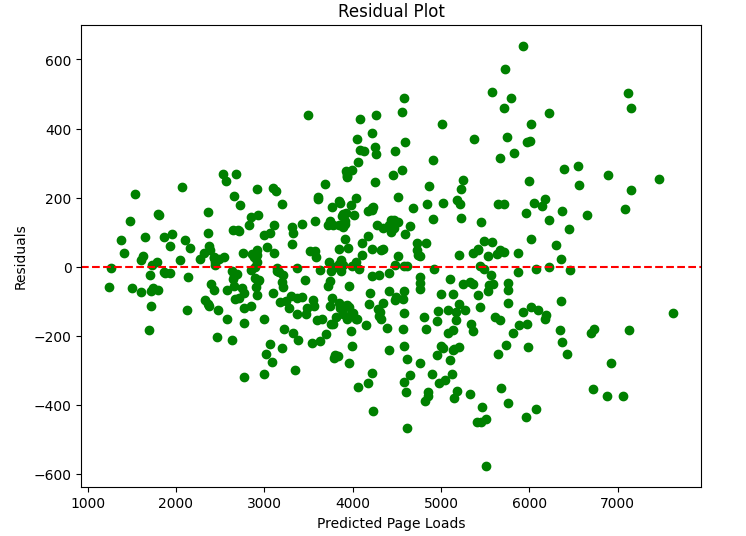
****

* When the day of the week is 7, the number of returning visitors is exceptionally low.
* The total of Returning over all values of Day.Of.Week and Day.Of.Week.Over 1.1 million people have visited.Returning.Visits may exceed 87 thousand by Day.Of.Week 9 based on current projections.
* According to current projections, First.Time.The number of visitors may exceed 395 thousand every day.Week nine.
* The total of First.Time.Visits across all Day.Of.Week values is around 5.3 million.
* The number of first-time visitors goes from over 456 thousand when the day of the week is 7, to almost 908 thousand when the day of the week is 3.

**VISUALIZATION USING PYTHON:**

****

* Plotting data points from the test dataset, real page loads are represented by x-coordinates and anticipated page loads by y-coordinates.
* Blue dots show how well the model is performing by comparing real and anticipated values.
* The "Perfect Prediction" line depicts the ideal situation in which the model predicts values that exactly match the real values.



* Plotting the expected page loads and residuals from the linear regression model illustrates the discrepancies between the model's predictions and the actual data.
* Overprediction and underprediction are indicated by green dots.
* The baseline for zero residuals, shown by the red dashed line at y=0, shows impartial and precise predictions.
* Unbiased forecasts are suggested when green dots are grouped around the red line.

**PERFORMANCE METRICS :**

A key performance metric in website traffic analysis is accuracy. It provides a simple approach to evaluate the overall accuracy of your model's predictions by tracking the percentage of instances in your dataset that were properly predicted. While lower accuracy reveals a need for prediction correctness improvement, higher accuracy indicates a better-performing model.

| **Metric** | **Value** |
| --- | --- |
| **Mean Absolute Error (MAE)** | 308.16 |
| **Mean Squared Error (MSE)** | 147,553.41 |
| **Root Mean Squared Error** | 384.02 |
| **R-squared (R2) Score** | 0.94 |

**Mean Absolute Error (MAE):**

* The average absolute difference between the expected and actual values is measured by the Mean Absolute Error.
* The average absolute difference in this instance between the actual and anticipated "Page Loads" is roughly 308.16.
* It displays the mean size of the model's prediction mistakes.

**Mean Squared Error (MSE):**

* The average of the squared discrepancies between the expected and actual values is called the mean squared error.
* This model's mean square error (MSE) is roughly 147,553.41, meaning that higher errors are given more weight.
* It is a measurement of the squared average error magnitude.

**Root Mean Squared Error (RMSE):**

* The MSE's square root is known as the Root Mean Squared Error. The target variable and it are both stated in the same unit.
* The RMSE in this instance is roughly 384.02, which offers a more comprehensible indication of the inaccuracy.
* It helps to comprehend the average magnitude of prediction mistakes in the model.

**R-squared (R2) Score:**

* The percentage of variance in the target variable (Page Loads in this case) that can be predicted from the independent variables (Day of the Week, Unique Visits, First Time Visits, and Returning Visits) is measured by the R-squared score.
* With an R2 value of 0.940, the independent variables in the model can account for almost 94% of the variance in "Page Loads".
* This shows that, depending on the metrics supplied, the model can accurately forecast "Page Loads" and has a strong fit with the data.