**Facebook Post Comment Volume Prediction**

Ta-Hung (Denny) Chen

Department of Mathematics and Statistics

Boston University, Boston, Massachusetts, U.S

Instructor: Dr. Fotios Kokkotos

December, 12, 2023

**Abstract**

In the dynamic realm of social media, the volume of comments a Facebook post garners serves as a crucial indicator of its engagement and reach. This study delves into the predictive factors influencing comment volume, utilizing the Facebook Comment Volume Dataset from the UCI Machine Learning Repository. Our research leverages mixed effect model to construct a negative binomial regression that predicts the number of comments a post is likely to receive within the subsequent hours of its publication. Drawing on the work of Kamaljot Singh and others, we employ count data regression methods and its extensions to account for potential overdispersion. By examining various post features, such as page characteristics, essential and weekday features, and other basic attributes, we endeavor to identify the key determinants of comment volume. The model's accuracy will be assessed using Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), alongside Posterior Predictive Check plots. Our findings aim to empower content creators and social media strategists to amplify their online presence and foster organic user interactions effectively. By providing insights into the promotion of Facebook posts without relying on paid advertising, this research seeks to democratize the approach to enhancing social media visibility.

**Introduction:**

The study focuses on forecasting Facebook comment volumes using mixed effect regression models, essential for gauging social media engagement. As online interactions dominate today's social sphere, predicting user engagement on Facebook is key for creators and marketers. This research utilizes the Facebook Comment Volume Dataset by Singh and Kaur (2015) to analyze the comments a post receives in the first three days—critical for assessing user interaction. The dataset provides features related to the posts for a detailed examination of the factors affecting comment volumes. Count data regression modeling, starting with Poisson regression, is the methodology used, with flexibility to adapt to other models in case of data overdispersion. The research aims to identify factors that significantly affect Facebook comment volumes and to test the predictive performance of the models. The findings are intended to enhance social media engagement strategies. In essence, this paper advances social media analytics by applying a hierarchical model to predict and understand Facebook user engagement.

**Datasets:**

The analysis conducted in this research is based on data sourced from the Comment Volume Prediction using Neural Networks and Decision Trees, originally collected by Singh and Kaur (2015) and made available through the UCI Machine Learning Repository. The dataset originates from Facebook Pages and has been meticulously prepared to facilitate the study of comment volume on posts. According to the authors of the research that produced the data, it is presumed that only comments posted within the last three days relative to a given Base date/time[[1]](#footnote-1) are relevant, as older posts are not typically expected to gain further engagement.

To ensure data integrity, any posts lacking comments or other essential information have been excluded. The dataset is divided into two parts: training and testing. The training data encompasses post information collected at five distinct time intervals, resulting in five different data variants.[[2]](#footnote-2) The fifth dataset, known as *Data Variant 5*, has been selected for analysis due to its comprehensive nature and the richness of its observations. Regarding the testing dataset, it comprises 10 test cases, each with 100 observations, and is merged together.

Predictors contain (1) Page features, (2) Essential features, (3) Weekday features, (4) Other basic features. Followed the feature definition by Singh and Kaur (2015).

1. Page features:

Four features of the category were identified to define the characteristic of the post. *Page likes:* It is a feature that defines users support for specific comments, pictures, wall posts, statuses, or pages. *Page Category:* This defined the category of source of document eg: local business or place, brand, or product, company or institution, artist, band, entertainment, community, etc. *Page Check in’s:* The feature shows the presence of the post at particular place. *Page Talking About:* The actual count of users who are “engaged” and interacting with the Page. Including the activities such as comments, likes, shares.

1. Essential features:

Essential features indicate the pattern of comments on the post within various time interval with reference to random select Base date/time. *CC1:* Total comment count within 72 hours before the selected Base date/time. *CC2:* Comment count in last 24 hours w.r.t to the selected Base date/time. *CC3:* Comment count between last 24 hours to last 48 hours w.r.t to the selected Base date/time. *CC4:* Comment count in first 24 hours w.r.t the selected Base date/time. *CC5:* the difference between *CC2* and *CC3*. And the data also contains min, max, standard deviation, median, and mean of *CC1* to *CC5*.

1. Weekday features:

Weekday features represent as a binary indicator showing the day on which the post was published and the day on selected Base date/time.

1. Other basic features:

Other basic features show additional information of the post, including the length of the document, time gap between selected Base date/time and document published ranges from [0,72], document promotion status, and post share count.

**Methodology:**

The problem planned to be addressed using a count data regression predictive modeling approach, with Poisson regression initially employed. There is flexibility to modify the regression assumptions to better align with the data and yield accurate results. For instance, should overdispersion be detected within the data, alternative models such as Quasi-Poisson or negative binomial regression may be utilized.

* Potential assumption adjustment for Poisson model:
* Quasi-Poisson regression:

Where:

Expected value of the observation .

: Coefficient to be estimated.

: Predictors’ value for observation .

Variance for Quasi-Poisson:

is the dispersion coefficient. While is a Poisson distribution with assumptions of parameters .

* Negative binomial regression:

Where:

Expected value of the observation .

: Coefficient to be estimated.

: Predictors’ value for observation .

Variance for negative binomial regression:

are the dispersion parameter in different presenting form. While , or , negative binomial model converges to a Poisson distribution with assumptions of parameters .

* Zero inflation model:

Since overdispersion is often observed in counting data problem, a zero-inflation negative binomial regression will be adopted if required. It combines two parts, the count model and the zero-inflation model.

1. Count Model (Negative Binomial Part): Same as the description of negative binomial regression above.
2. Zero-inflation Model:

Where:

: is the probability of -th observation is an “extra” zero.

: the independent variables for the zero-inflation part.

: the coefficients for the zero-inflation part.

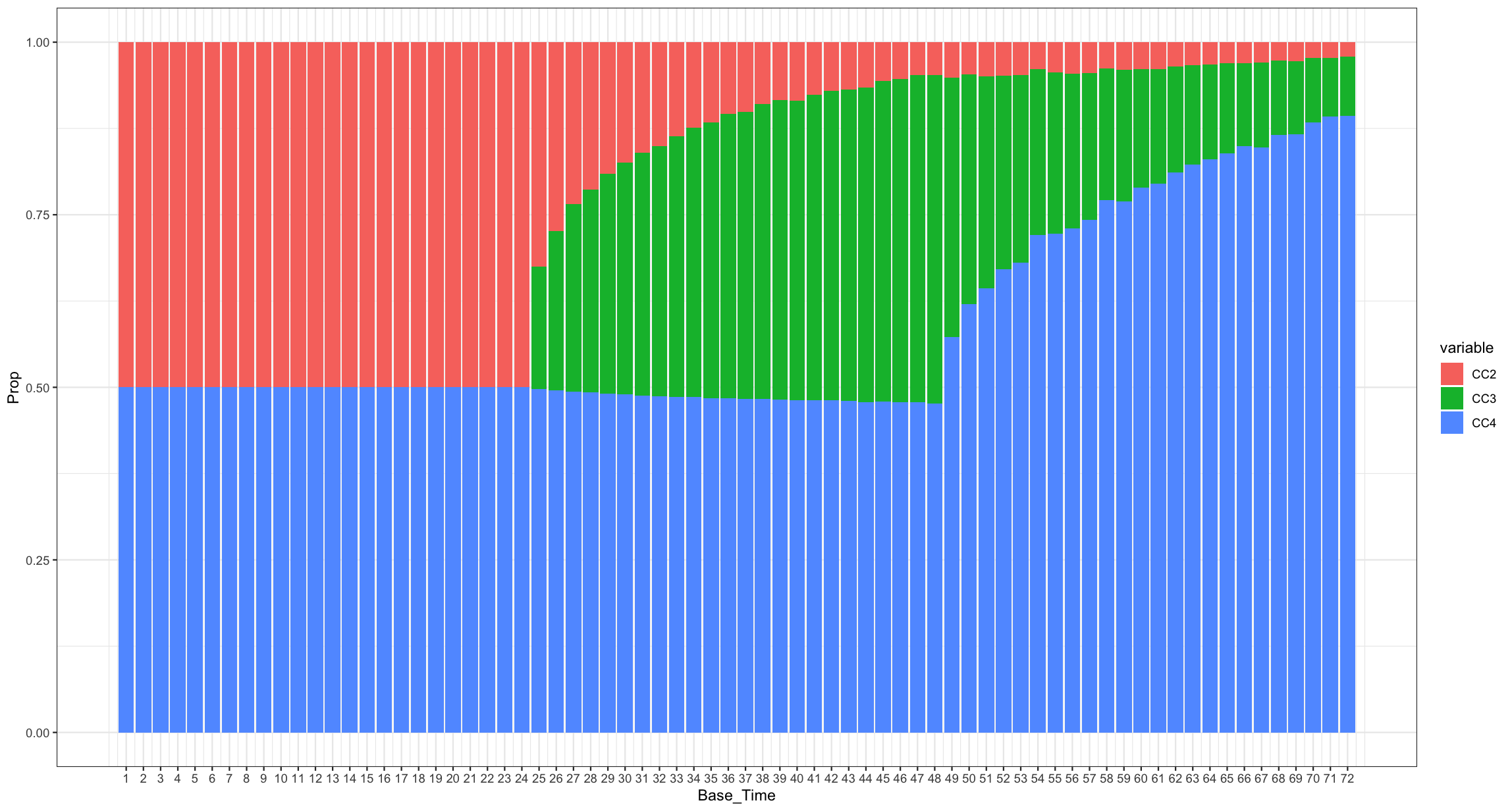
**Validation:**

Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) will be used to evaluate the accuracy for the regression problem.

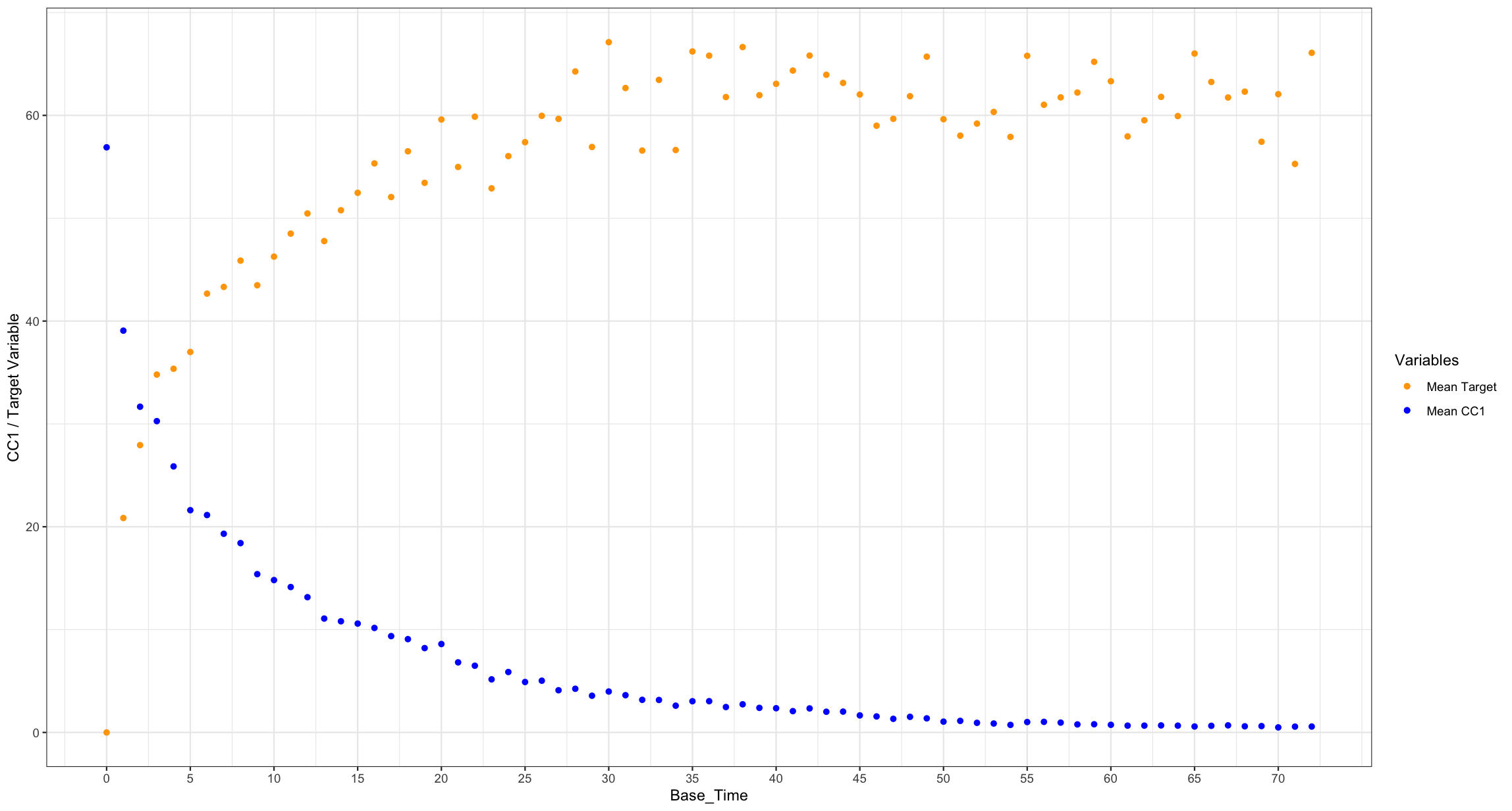
Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Posterior Predicting Check plot are planned to be considered to evaluate the fitness of the regression model.

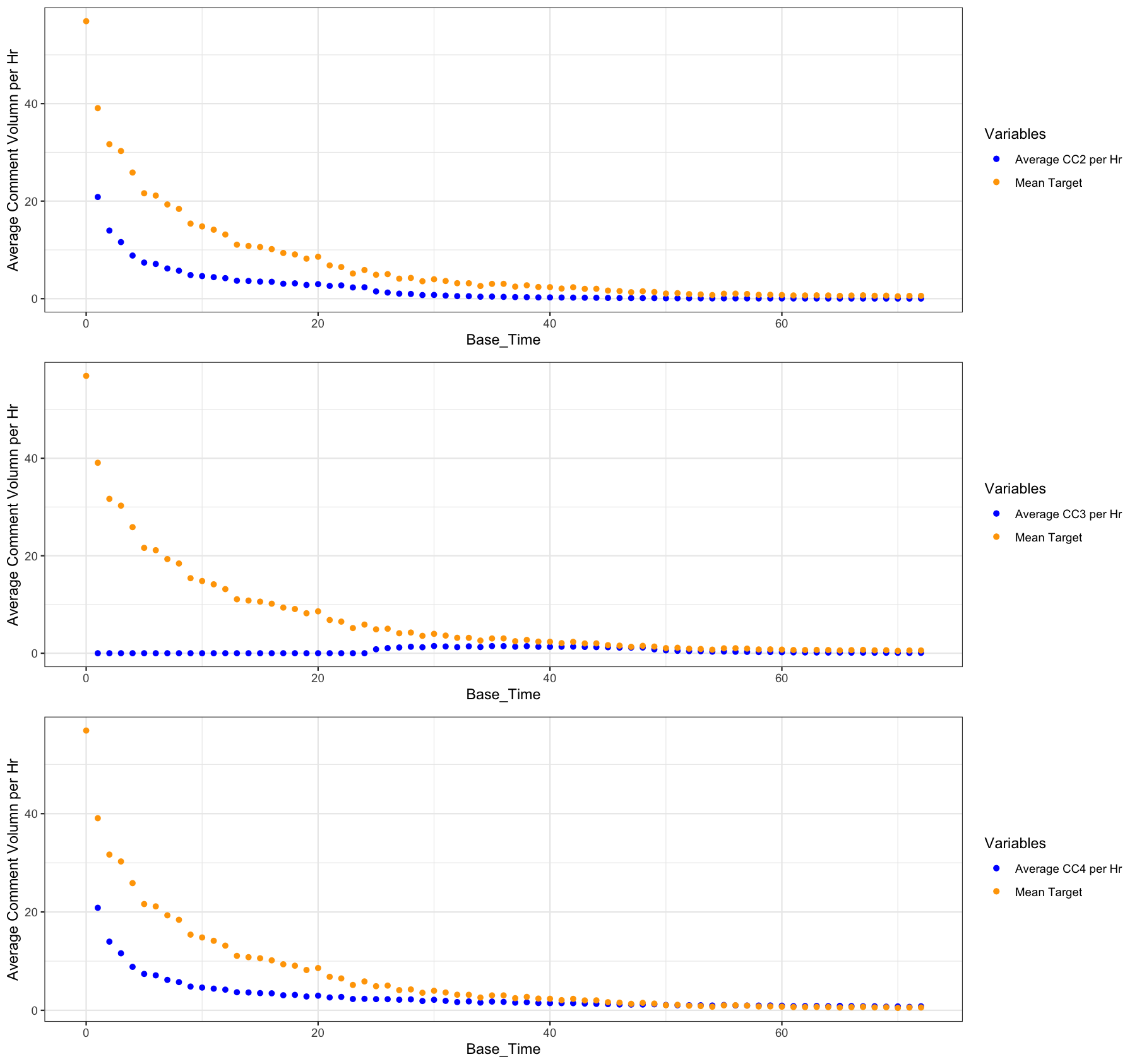
**Analysis:**

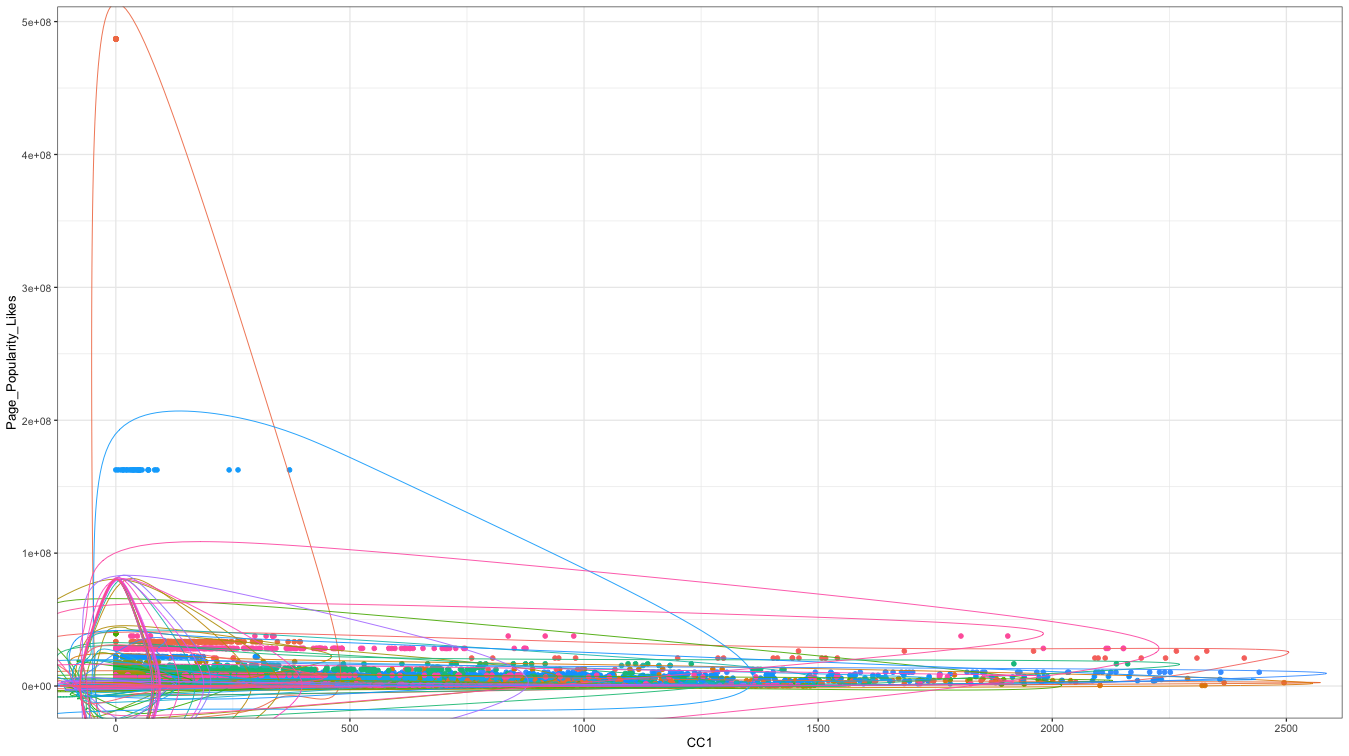
Exploratory Data Analysis:

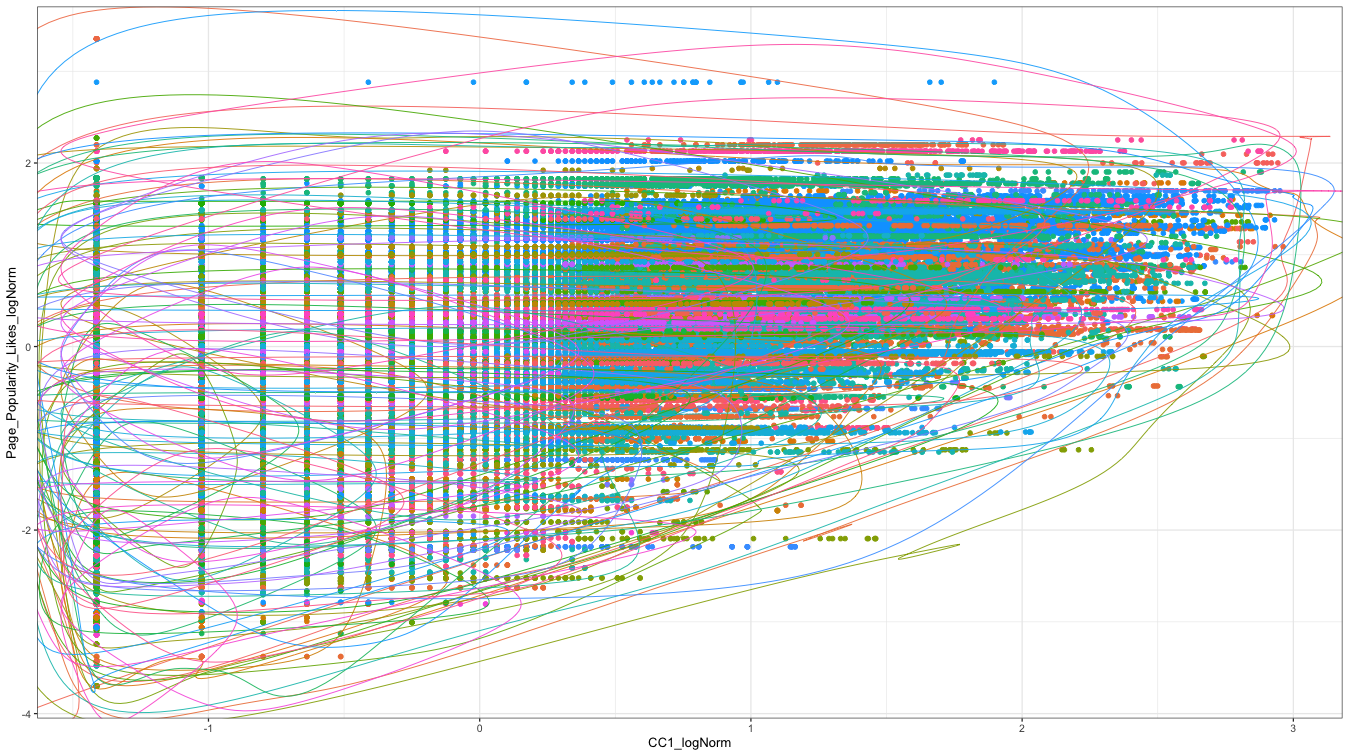
 Most of the data cleansing and manipulation work had been done by Singh and Kaur (2015), and I directly started the analysis based on the cleaned and well-organized data which is available through UCI- Machine Learning Repository. Refer to what I have mentioned in the Datasets section is that *CC4* is the comment volume in the first 24 hours, and *CC2* is the comment volumne in last 24 hours w.r.t to the selected Base date/time, which means, if the Base date/time is selected within 24 hours after the post was published, then *CC2* equals to *CC4*, shown in Figure X. From Figure X, we can observe that most of the comment are posted with the first 24 hour after the post was published. And it meets the assumption of “Only comments posted within the last three days relative to a given base date/time are relevant, as older posts are not typically expected to gain further engagement.”

Feature Engineering:

 The analysis aimed to tackle with a count data problem. In a Poisson regression model, offset is often considered when designing the regression. After observing the *Target\_Variable* to the *Base\_Time,* shown in Figure X. Refer to Appendix.2, Base\_Time means the hour from the selected Base date/time and to the time of the post published. We can see that the *Target\_Variable* looks like a cumulative volume estimation of the post comment, so an offset term is considered in this analysis.

 Being interested in the average comment volume within every hour w.r.t the selected Base date/time, three new variables are created, *CC2\_per\_hr, CC3\_per\_hr, CC4\_per\_hr*, shown in Figure X.

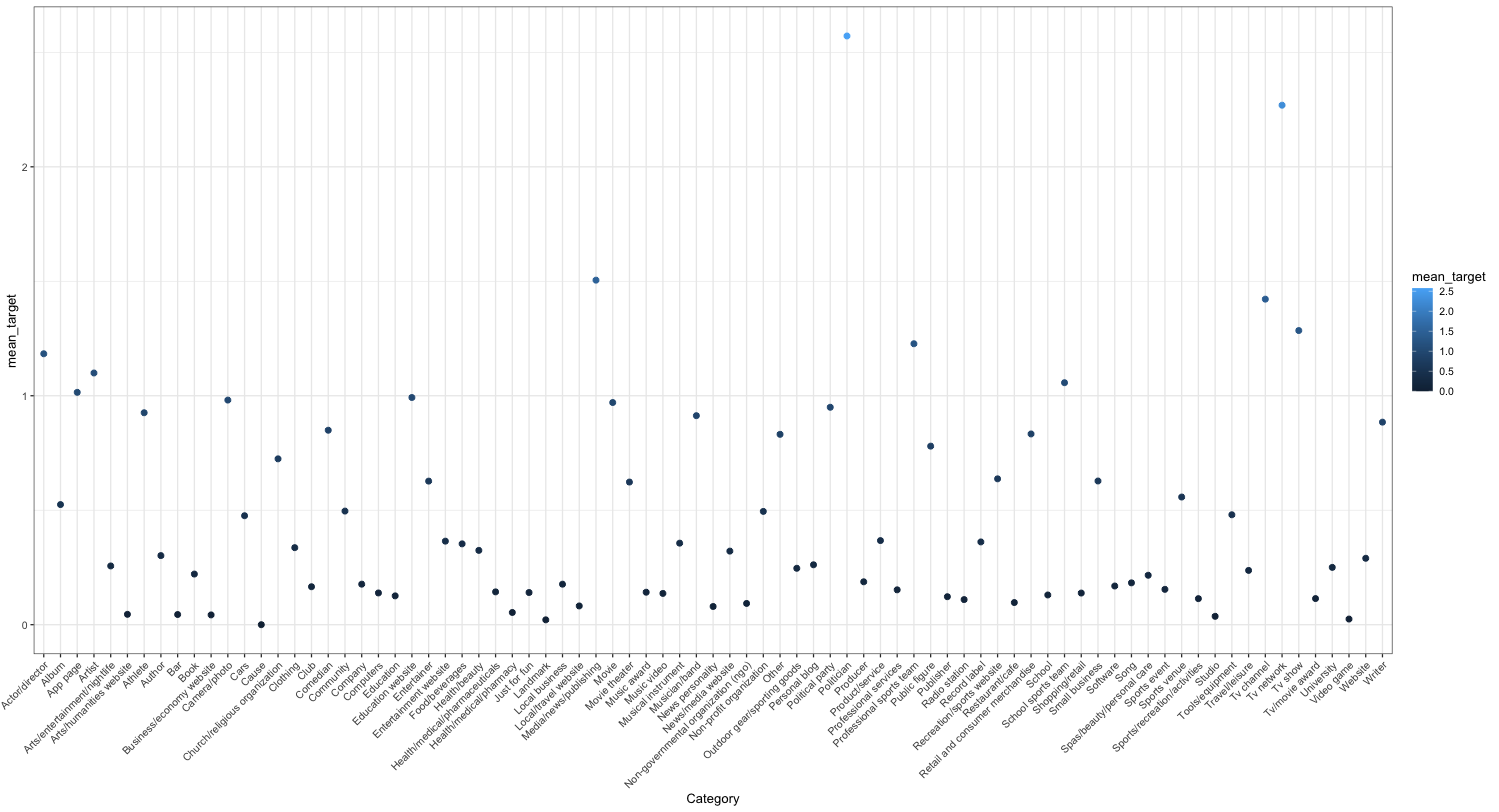
Scales of predictors in regression analysis are also critical to estimate the estimands. The original data are highly skewed, and scale (range of the column) varies dramatically between variables. From Figure X, I projected the data points to *CC1* and *Page\_Popularity\_Likes*. It is obvious that the data scale varies drastically, and high skewness occurs. So I decided to perform a log transformation, and normalize the variable to make the distribution looks more like a bell-shaped, shown in Figure X.

*CC1, CC2, CC3, CC4, Page\_Popularity\_Likes, Page\_Checkins, Page\_Talking\_About, Post\_Length, Post\_Share\_Count, CC2\_per\_hr, CC3\_per\_hr, CC4\_per\_hr* are addressed by the log transformation and normalization.From Figure X, we can see that the serious skewness and largely varying scale of the axis range had been mitigated.

Regression Analysis:

After creating new variables and solved the data scaling problems, I started the regression analysis. Considering the meaning of the variables, the final predictors are listed below.

* + - CC1\_logNorm
    - CC2\_logNorm
    - CC3\_logNorm
    - CC4\_logNorm
    - CC5
    - Page\_Popularity\_Likes\_logNorm
    - Page\_Checkins\_logNorm
    - Page\_Talking\_About\_logNorm
    - Post\_Length\_logNorm
    - Post\_Share\_Count\_logNorm
    - CC2\_per\_hr\_logNorm
    - CC3\_per\_hr\_logNorm
    - CC4\_per\_hr\_logNorm

And by observing different mean *Target\_Variable* from the data, shown in Figure X. A mixed effect model is adopted, and the random effect of the *Page\_Category* will be estimated in the intercept to present the randomness of each *Category*.

Initially, I applied the Poisson regression, which is one of the classical regressions for counting data problem. But the varying scale issues still exist in the data, and is hardly to remove it entirely, the Poisson model failed to converge for estimating the coefficients. Besides, overdispersion also existed in the data, recalling the assumption of Poisson regression, the parameter . Instead, the mean and variance of the response variable, *Target\_Variable* is and . So, instead of

**Results:**

**References:**

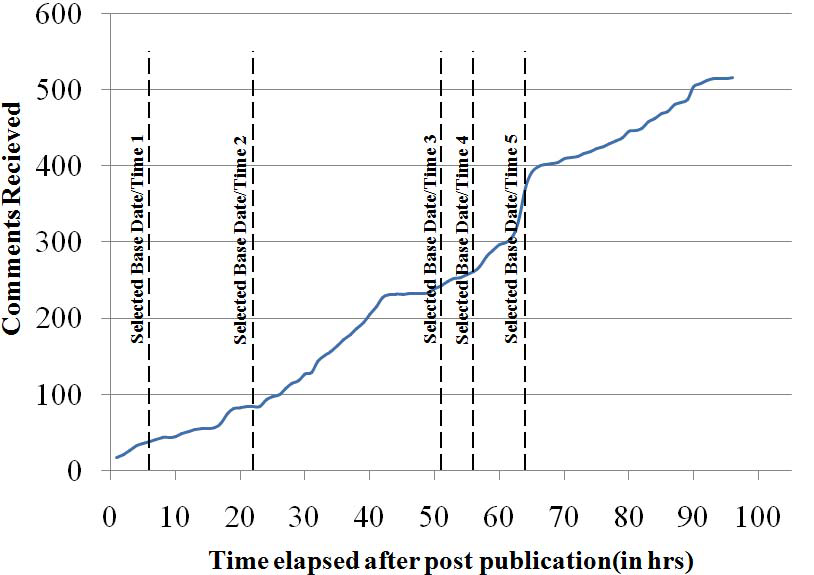
* Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. "Comment volume prediction using neural networks and decision trees." IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015). 2015.
* Singh, Kamaljot. "Facebook comment volume prediction." *International Journal of Simulation: Systems, Science and Technologies* 16.5 (2015): 16-1.

**Appendix 1. Summary Statistics**

**Data Variant 5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistic** | **N** | **Mean** | **St.Dev.** | **Min** | **Max** |
| Page\_Category | 199,030 | 24.242 | 19.935 | 1 | 106 |
| Page\_Popularity\_Likes | 199,030 | 1,313,785.00 | 6,771,131.00 | 36 | 486,972,297 |
| Page\_Checkins | 199,030 | 4,674.52 | 20,573.44 | 0 | 186,370 |
| Page\_Talking\_About | 199,030 | 44,771.73 | 110,898.30 | 0 | 6,089,942 |
| CC1\_Min | 199,030 | 0.47 | 13.178 | 0 | 1,458 |
| CC1\_Max | 199,030 | 485.318 | 538.194 | 0 | 2,495 |
| CC1\_Avg | 199,030 | 55.901 | 86.515 | 0 | 2,031.00 |
| CC1\_Median | 199,030 | 35.264 | 68.163 | 0 | 2,123.00 |
| CC1\_Std | 199,030 | 68.091 | 82.411 | 0 | 762.358 |
| CC2\_Min | 199,030 | 0.068 | 2.173 | 0 | 227 |
| CC2\_Max | 199,030 | 381.499 | 439.634 | 0 | 2,119 |
| CC2\_Avg | 199,030 | 21.815 | 35.693 | 0 | 973.25 |
| CC2\_Median | 199,030 | 7.17 | 19.701 | 0 | 1,121.00 |
| CC2\_Std | 199,030 | 40.514 | 51.561 | 0 | 683.596 |
| CC3\_Min | 199,030 | 0.006 | 0.872 | 0 | 148 |
| CC3\_Max | 199,030 | 380.723 | 430.183 | 0 | 2,095 |
| CC3\_Avg | 199,030 | 19.992 | 31.568 | 0 | 660.75 |
| CC3\_Median | 199,030 | 4.876 | 13.072 | 0 | 487 |
| CC3\_Std | 199,030 | 40.712 | 52.598 | 0 | 801.468 |
| CC4\_Min | 199,030 | 0.469 | 13.126 | 0 | 1,458 |
| CC4\_Max | 199,030 | 434.882 | 490.73 | 0 | 2,184 |
| CC4\_Avg | 199,030 | 52.754 | 81.02 | 0 | 1,868.50 |
| CC4\_Median | 199,030 | 33.608 | 64.178 | 0 | 1,992.50 |
| CC4\_Std | 199,030 | 63.461 | 76.836 | 0 | 680.962 |
| CC5\_Min | 199,030 | -326.275 | 380.145 | -2,038 | 0 |
| CC5\_Max | 199,030 | 377.323 | 436.702 | -101 | 2,119 |
| CC5\_Avg | 199,030 | 1.822 | 9.69 | -184.4 | 496.6 |
| CC5\_Median | 199,030 | -2.119 | 10.488 | -175 | 521 |
| CC5\_Std | 199,030 | 56.54 | 74.583 | 0 | 1,386.40 |
| CC1 | 199,030 | 55.901 | 137.524 | 0 | 2,495 |
| CC2 | 199,030 | 21.815 | 74.658 | 0 | 2,119 |
| CC3 | 199,030 | 19.992 | 73.625 | 0 | 2,095 |
| CC4 | 199,030 | 52.754 | 128.434 | 0 | 2,184 |
| CC5 | 199,030 | 1.822 | 94.092 | -2,038 | 2,119 |
| Base\_Time | 199,030 | 35.45 | 21.006 | 0 | 72 |
| Post\_Length | 199,030 | 163.692 | 375.663 | 0 | 21,480 |
| Post\_Share\_Count | 199,030 | 117.363 | 954.359 | 1 | 144,860 |
| Post\_Promotion\_Status | 199,030 | 0 | 0 | 0 | 0 |
| H\_Local | 199,030 | 23.783 | 1.827 | 1 | 24 |
| Post\_Published\_Weekday\_40 | 199,030 | 0.122 | 0.328 | 0 | 1 |
| Post\_Published\_Weekday\_41 | 199,030 | 0.143 | 0.35 | 0 | 1 |
| Post\_Published\_Weekday\_42 | 199,030 | 0.149 | 0.357 | 0 | 1 |
| Post\_Published\_Weekday\_43 | 199,030 | 0.157 | 0.364 | 0 | 1 |
| Post\_Published\_Weekday\_44 | 199,030 | 0.144 | 0.351 | 0 | 1 |
| Post\_Published\_Weekday\_45 | 199,030 | 0.146 | 0.353 | 0 | 1 |
| Post\_Published\_Weekday\_46 | 199,030 | 0.137 | 0.344 | 0 | 1 |
| Base\_DateTime\_Weekday\_47 | 199,030 | 0.139 | 0.346 | 0 | 1 |
| Base\_DateTime\_Weekday\_48 | 199,030 | 0.135 | 0.342 | 0 | 1 |
| Base\_DateTime\_Weekday\_49 | 199,030 | 0.137 | 0.344 | 0 | 1 |
| Base\_DateTime\_Weekday\_50 | 199,030 | 0.147 | 0.354 | 0 | 1 |
| Base\_DateTime\_Weekday\_51 | 199,030 | 0.155 | 0.362 | 0 | 1 |
| Base\_DateTime\_Weekday\_52 | 199,030 | 0.144 | 0.351 | 0 | 1 |
| Base\_DateTime\_Weekday\_53 | 199,030 | 0.142 | 0.349 | 0 | 1 |
| Target\_Variable | 199,030 | 7.169 | 34.298 | 0 | 1,702 |
| CC2\_per\_hr | 199,030 | 2.098 | 10.597 | 0 | 1,011.00 |
| CC3\_per\_hr | 199,030 | 0.504 | 1.92 | 0 | 58.6 |
| CC4\_per\_hr | 199,030 | 2.768 | 10.785 | 0 | 1,011.00 |
| CC1\_logNorm | 199,030 | 0 | 1 | -1.413 | 2.962 |
| CC2\_logNorm | 199,030 | 0 | 1 | -0.929 | 3.866 |
| CC3\_logNorm | 199,030 | 0 | 1 | -0.767 | 3.943 |
| CC4\_logNorm | 199,030 | 0 | 1 | -1.397 | 2.93 |
| Page\_Popularity\_Likes\_logNorm | 199,030 | 0 | 1 | -3.699 | 3.352 |
| Page\_Checkins\_logNorm | 199,030 | 0 | 1 | -0.643 | 2.838 |
| Page\_Talking\_About\_logNorm | 199,030 | 0 | 1 | -2.717 | 2.364 |
| Post\_Length\_logNorm | 199,030 | 0 | 1 | -2.345 | 3.323 |
| Post\_Share\_Count\_logNorm | 199,030 | 0 | 1 | -1.112 | 4.901 |
| CC2\_per\_hr\_logNorm | 199,030 | 0 | 1 | -0.542 | 7.999 |
| CC3\_per\_hr\_logNorm | 199,030 | 0 | 1 | -0.465 | 8.335 |
| CC4\_per\_hr\_logNorm | 199,030 | 0 | 1 | -0.778 | 7.195 |

**Appendix 2. Selected Base date/time and Data Variants**

****

Reference:

* Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. "Comment volume prediction using neural networks and decision trees." IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015). 2015. Figure 3. Cumulative Comments and different selected base date/time.

1. Base date/time is selected to simulated the scenario, as we already know what will happen after this. There is one more kind of time we used in this formulation: is the post published time, which comes before the selected base date/time. [↑](#footnote-ref-1)
2. Data variants are different samples that are collected at different Base date/time. Singh and Kaur (2015) Figure 3. [↑](#footnote-ref-2)