An Analysis of the Seoul Bike Sharing Dataset

In recent years, due to increased awareness of the economic impact of eco-friendly transportation means has caused a surge in eco-conscious transportation methods, coupled with a push for greater accessibility and ease of use has notably increased public interest in bike-sharing programs. These programs not only contribute to reducing the urban carbon footprint but also offer a convenient and flexible transportation alternative. This essay delves into the Seoul Bike Sharing Demand Dataset to unravel the key factors influencing bike rental patterns. Through comprehensive analysis, my aim is to determine the key factors affecting bike rental counts and develop a predictive model for estimating hourly bike rental counts, thereby offering insights into the bike-sharing system in Seoul.

I decided to use Linear Regression to help me build a model for the Rented Bike Count numbers in Seoul, as a linear model would be suitable for my dataset and goal because it is designed for predicting continuous outcomes, like the Rented Bike Count, which is a numerical variable for the count of bikes rented. Logistic Regression would not fit my objectives, as I am not trying to build a model to predict the probability of a binary outcome. The dataset contains explanatory variables, such as temperature and humidity, which are expected to have a linear relationship with bike rental counts. In addition, linear regression will provide interpretable results that quantify the change in bike rentals for unit changes in the predictor variables, aligning with my dataset's structure and my objectives. The columns or independent variables I will use to predict the Rented Bike Count number are : Temperature, Humidity, Visibility, Dew Point Temperature, Wind Speed, Solar Radiation, Rainfall, Snowfall, Seasons (Winter, Spring, Summer, Autumn), Holiday (Holiday, No Holiday), and Functioning day (Yes, No). I hypothesize that the weather conditions and the time of the year significantly influence rental behavior, as Seoul has extreme temperatures in the Winter’s and the Summer’s, and bikes are a public transport service in which people are exposed to the weather. To enable Linear Regression I implemented dummy variables for categorical predictors to facilitate their use in the regression model. For the 'Seasons' column, I used dummy\_Spring for Spring, dummy\_Summer for Summer, and dummy\_Autumn for Autumn. For the 'Holiday' column, dummy\_Holiday was used for a Holiday, and dummy\_FunctioningDay was implemented for a Functioning Day in the 'Functioning Day' column.

To get an idea of just how much people rent bikes, I created a histogram ( Appendix 1) that shows the distribution of rented bike counts, overlaid with a normal distribution curve. The distribution of rented bike counts is skewed to the right, indicating that there are more instances of lower rental counts and fewer instances of very high rental counts. The peak of the histogram shows that the most common rental count, 60, falls well below the average of 704.6. This skewness implies that there may be periods of low usage, such as off-peak hours or less favorable weather conditions.

In my initial exploration of the data, I plotted Rented Bike Counts (Independent Variable) against other dependent Variables. The Boxplot (Appendix 2) between Rented Bike Count and Seasons shows seasonal trends, with higher rental counts in Summer and Autumn, suggesting higher rental counts in warmer weather. This could imply people not wanting to go with riding bikes during colder months due to it getting really cold in Seoul during the winter, and prefer other options such as buses/cars etc. Also, the tourism would increase during the warmer months, so that could be a factor in the higher numbers of rented bikes in the summer. This plot supports my hypothesis that time of the year and temperature have a significant effect on the total number of Rented Bikes Counts. The scatter plot (Appendix 3) between Rented Bike Count and Temperature has a positive correlation and supports the notion that milder and warmer weather conditions encourage more people to rent bikes, likely due to the increased comfort and enjoyment associated with biking in such weather. The highest numbers in terms of rented bikes are during the temperatures of between 20-30 degree Celsius, and numbers for 30-40 degree Celsius show that there are not as much rented bikes as for 20-30 degree Celsius. It also suggests that extremely hot or cold temperatures might deter people from renting bikes, possibly due to the discomfort or perceived dangers of biking in such conditions, such as frostbite, chills, dehydration, etc. This plot supports my hypothesis that temperature has a significant effect on people’s decision to rent bikes.

To increase the predictive accurary of my model, I addressed multicollinearity with the help of the Initial Variance Inflation Factor (VIF) output ( Appendix 4) which showed high multicollinearity between Humidity (20), Temperature (89), and Dew Point Temperature (117), which makes sense as they are all weather conditions. Dew Point Temperature was removed due to it having the highest VIF value of 117. This adjustment output (Appendix 5) resulted in VIF values of 2 for Humidity and 5 for Temperature and the vif values were below 10 for all variables and there was only a negligible decrease in Adjusted R^2 (from 0.5497 to 0.5493), indicating that Dew Point Temperature did not significantly contribute to explaining bike rental variance. To further improve the predictive power of my model I removed outliers and influential points, and after removing them, the Adjusted R^2 increased from 0.5493 to 0.5832, signifying a significant improvement in the model's explanatory power without Dew Point Temperature and outliers and influential points.

In checking for normality, the Residuals Histogram (Appendix 8) displays a reasonably symmetrical distribution about the mean, with the tallest bars clustered near the center and fading away towards the tail. This pattern suggests that most of the prediction errors are relatively small, with fewer large errors, a distribution pattern similar to a normal curve. Even though there is a slight rightward skewness observed, indicating a greater number of positive residuals, the overall shape does not change drastically from normality. The presence of a few larger residuals on the right is not unexpected, considering that rental counts can be subject to sudden surges that are attributed to other factors or random fluctuations. Hence, there is no significant violation of normality.

Plotting a Residuals vs Predicted plot (Appendix 9), it showed a slight pattern indicative of heteroscedasticity, showing as an increasing spread in residuals as the predicted bike rental counts rise. While this could imply varying error variances across the range of predictions, the magnitude of this phenomenon did not appear severe. I decided to proceed without corrective measures due to the severity and how large the dataset was, and how the priority here is the models predictive performance. In large datasets like this, the effect of heteroscedasticity is often diluted because the estimation of the coefficients is based on many data points, reducing the influence of any individual data point with a high variance.

To split the dataset into training and testing subsets, I did an 80/20 split as that means a good amount of data for both testing and training and balance between Overfitting and Underfitting, as the data is relatively large. To further develop my model and come up with my model equation, I decided to use a variable selection method to help me improve my model and determine a set of predictor variables. For this, I tried the following five selection methods: Cp, Forward, Backward, Adjusted-R^2, and Stepwise on my testing data. For Stepwise, Cp, and Backward, the R^2 was 0.5841 and 12 variables. For Forward selection, it was 0.5841 and 13 variables, and for Adj-R^2 it was 0.5833 and 12 variables. I decided to use Stepwise Selection as there were lesser variables (12) and R^2 was higher. Stepwise selection took out Visibility column and so for my analysis, Visibility was taken out of the final model.

In my final regression parameter estimates (Appendix 6), all variables included in the final model were found to be significant predictors of bike rental counts (P < .05), which suggests a robust model where each predictor contributes meaningfully to the model. The parameter estimates for the predictors were as follows: Hour: With a parameter estimate of 26.3330, the hour of the day is a significant predictor of bike rentals, which implies that bike rental counts vary substantially throughout different hours of the day. For Temperature, the coefficient of 25.6076 indicates that higher temperatures are associated with increased bike rental counts. For Humidity, this variable has a negative relationship with bike rentals, which is shown by the parameter estimate of -8.20552, suggesting that higher humidity levels may discourage bike rentals, possibly explaining the drop in Rented Bike Counts in 30-40 degree Celsius compared to 20-30 degree Celsius. As for the Wind Speed, Wind speed has a smaller but still significant positive effect on bike rental counts with a coefficient of 15.75782. This makes sense as for the seasons with highest wind i.e. Winter and Spring, they both have lower bike counts than Summer and Autumn, which have the highest amount of rented bikes, so the effect of wind would be not as significant as others. Solar Radiation: Solar radiation is negatively associated with bike rentals, with a parameter estimate of -77.05616. Rainfall and Snowfall: Both have negative impacts on bike rentals, as expected, since adverse and extreme weather conditions can deter people from biking. For Seasonality: The dummy variables for Spring, Summer, and Autumn all have positive coefficients (214.67249, 192.17746, and 351.45104, respectively), aligning with our initial boxplot (Appendix 2) and confirming that bike rentals are higher in these seasons compared to Winter. As for Holidays, the negative coefficient for dummy\_Holiday (-87.53055) indicates fewer rentals on holidays, which may be due to the closure of businesses and reduced commuting needs. This makes sense as locals would not be needing the bikes on holidays, and tourists wouldn’t wait for holidays anyways to rent bikes. For Functioning Day, the positive and largest coefficient was for dummy\_FunctioningDay (902.81028) which strongly suggests that bike rentals are significantly higher on functioning days compared to non-functioning days, and the large number obviously means that it will not really be plausible to rent a bike during non-functional hours. These results align with my initial exploratory findings, which highlighted the influence of warmer weather and seasons on bike rental patterns. The significance of the hour variable may be explained by daily commuting trends and leisure activities of people, which are more prevalent during certain times of the day.

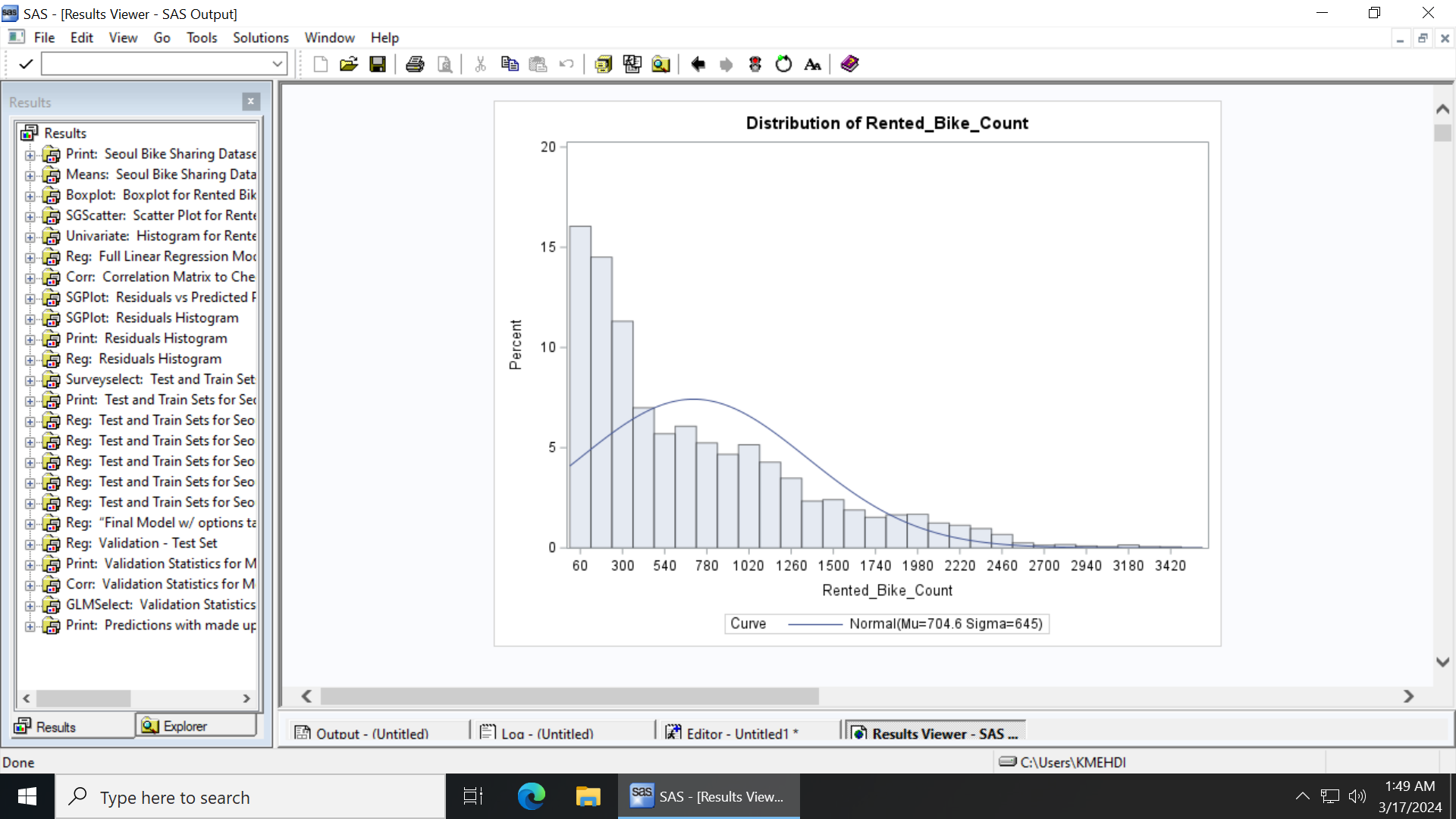
The final regression model yielded an R-square value of 0.58241, with an adjusted R-square of 0.5833, indicating that approximately 58.241% of the variability in bike rental counts can be explained by the independent variables included in the model for the test set. In this context for forecasting a bike-sharing demand from a large bike-sharing dataset, this R-Square value is good. By the correlation procedure output (Appendix 10), the yhat value was seen to be 0.76316, the square of which gives us an R-Square of 0.58241. This R-Square affirms the relevance of the chosen predictors, and the value of 0.58241 for the test set provides a strong indication that our model is capturing a meaningful portion of the underlying pattern in bike-sharing data.

In the output (Appendix 7), the obtained RMSE value of 393.395 and the MAE value of 300.780 for the test set suggest that the model offers a solid foundation for understanding general bike sharing patterns, but there is room for improvement in refining the predictions to be more accurate. The RMSE value is the standard deviation of the prediction errors, indicating that the model’s predictions are on average, 393.395 units away from the actual bike rental counts. The value of 393.395 suggests that while the model adequately handles the bulk of predictions, there are still instances of prediction error, which can likely be attributed to extreme values or anomalies within the dataset. The MAE value of 300.780 means that the predictions are deviating from actual counts by an average of 300.780 bikes, suggesting that for most predictions, the model achieves a reasonable degree of accuracy. It implies that for most applications, such as planning daily inventory levels or staffing requirements, the model provides a solid foundation for decision-making. However, given the dynamic and complex nature of bike rentals and their susceptibility to external variables, a certain degree of prediction error is anticipated. This level of accuracy is relatively reasonable, considering the broad range of rental counts (thousands), how big the dataset is, and the external factors affecting them.

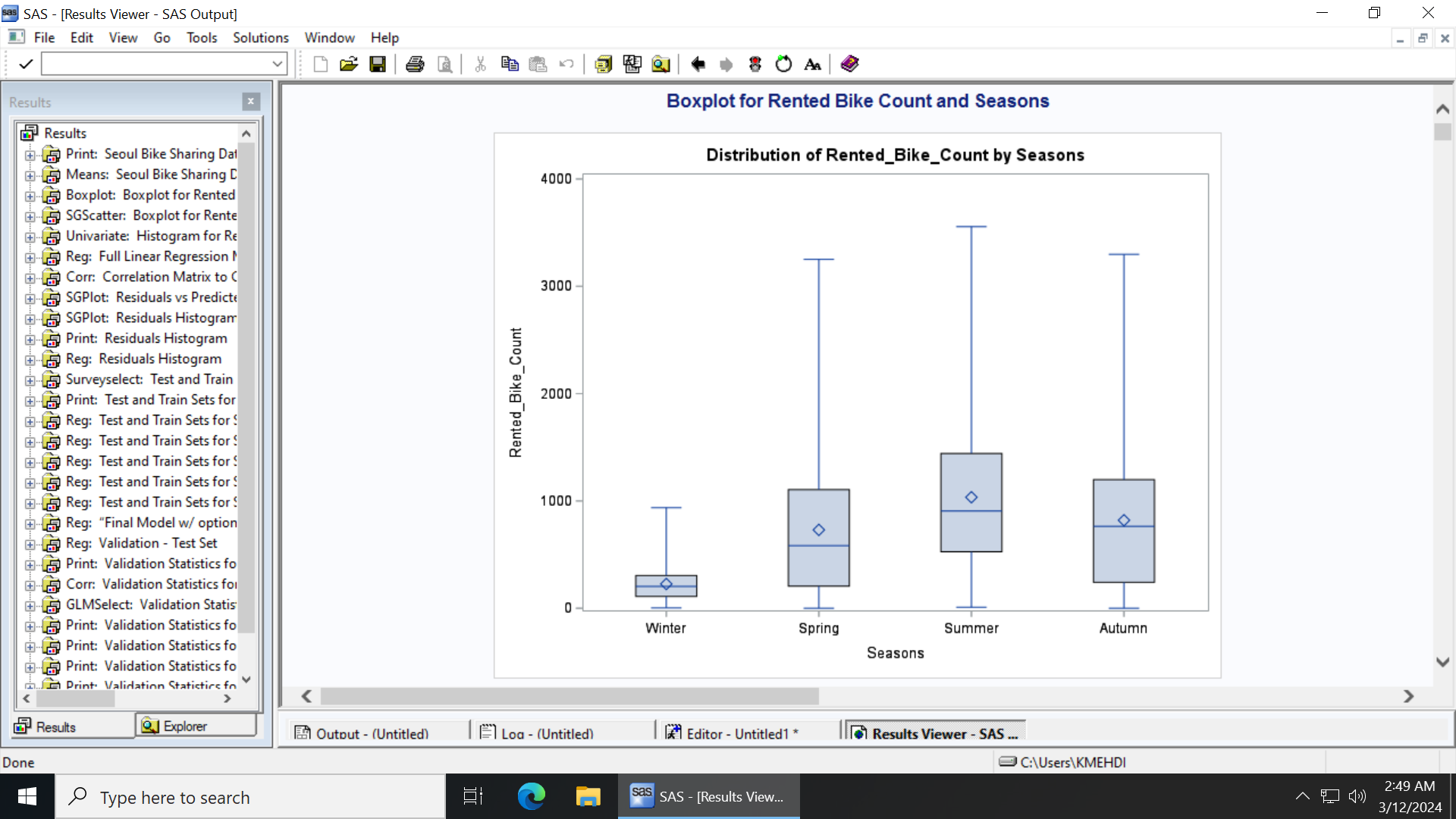
Overall, I have made significant changes and improvements to my initial model and I am satisfied with the improvements made, which reflect in my final model. There are a lot of factors which could have been taken into account and included in the dataset, such as the location of bike stations and the amount of bikes available in each station, the availability of bike lanes, the amount of pollution in the air, and a lot of factors which could not have been taken into account, such as real-time traffic conditions, or any local events or concerts, making this analysis and dataset dynamic and complex. However, this model is a good base to predict the Rental Bike Count number per each hour in Seoul.

Appendix

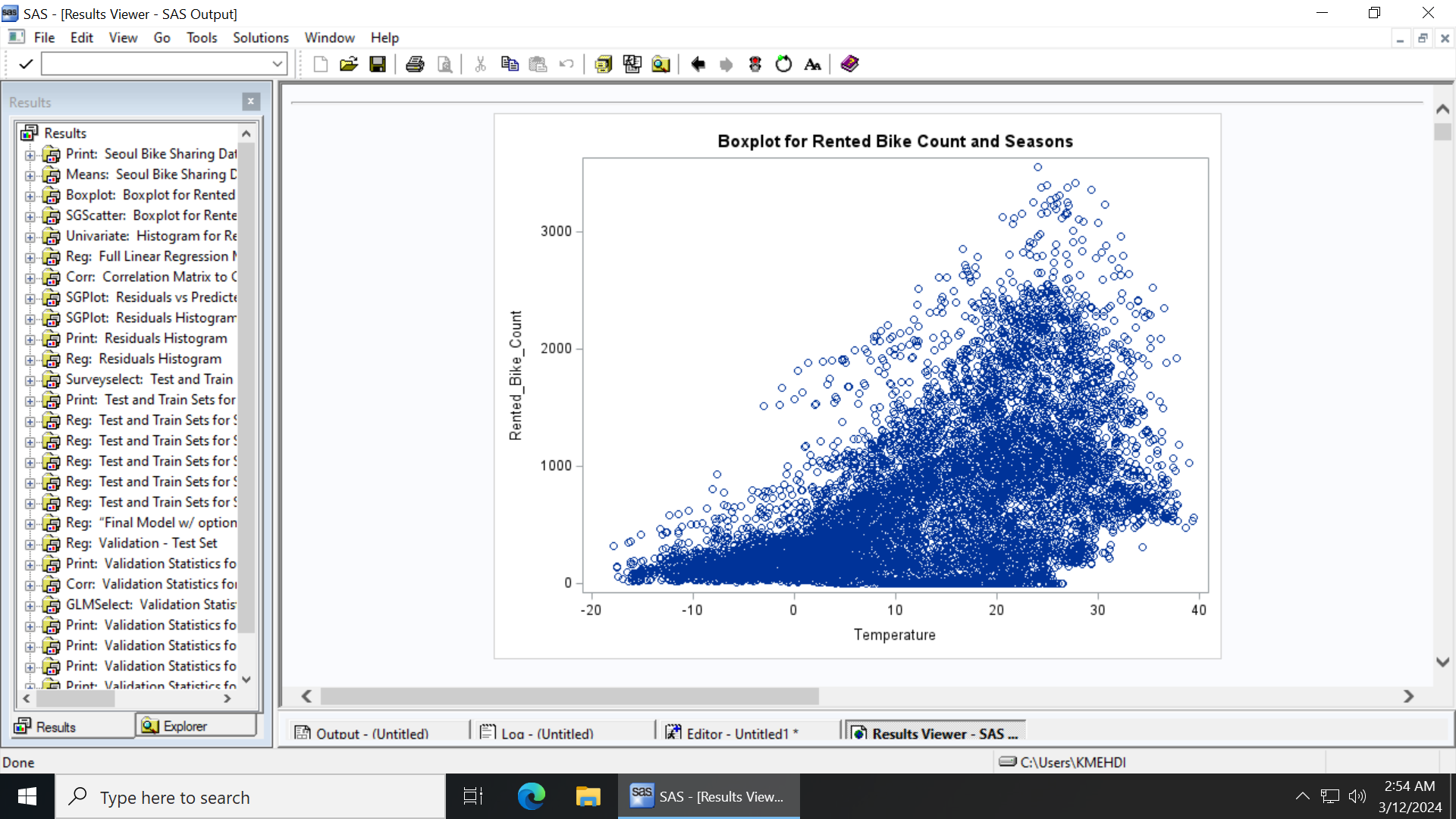
A1:



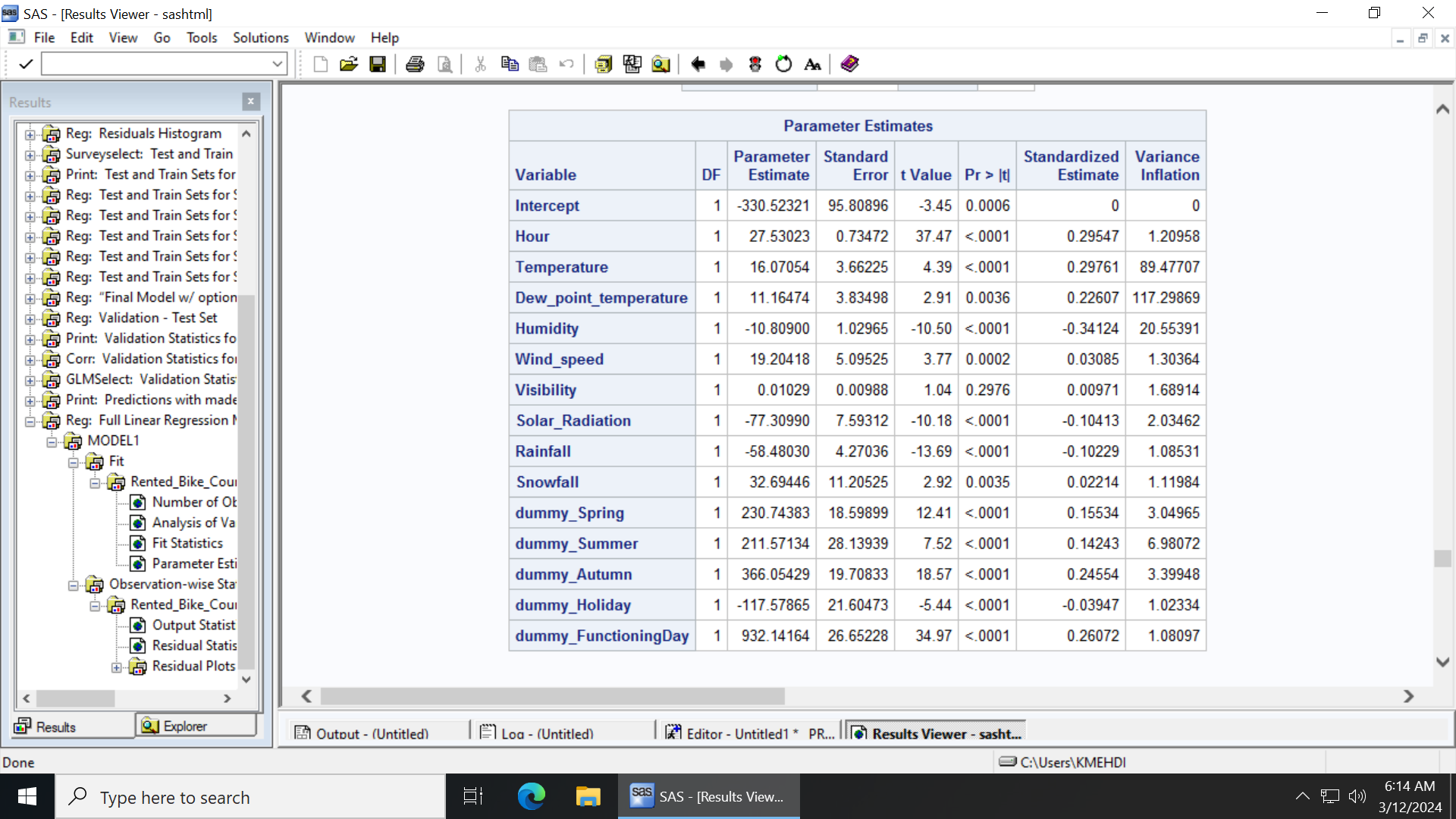
A2:



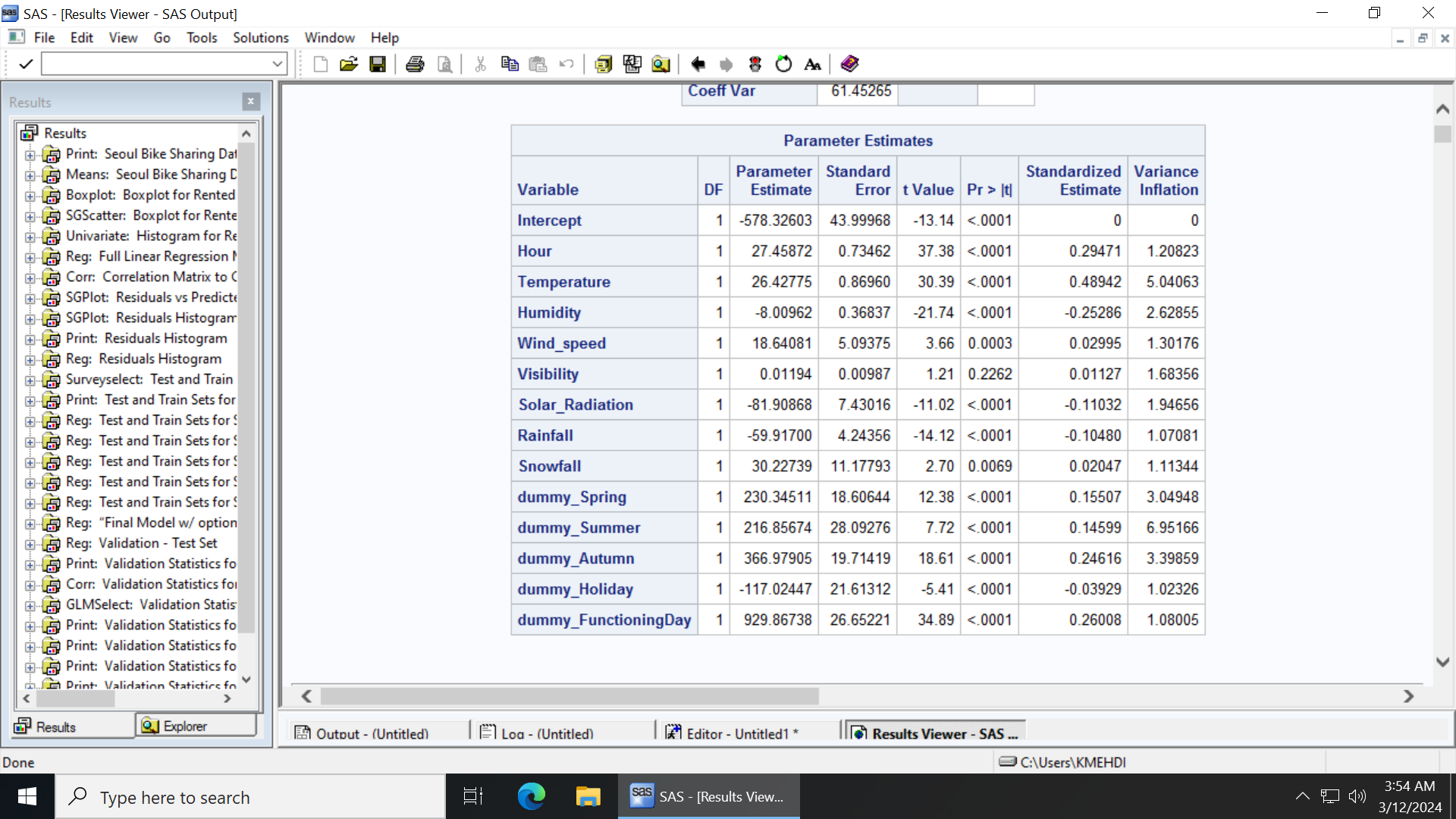
A3:



A4:



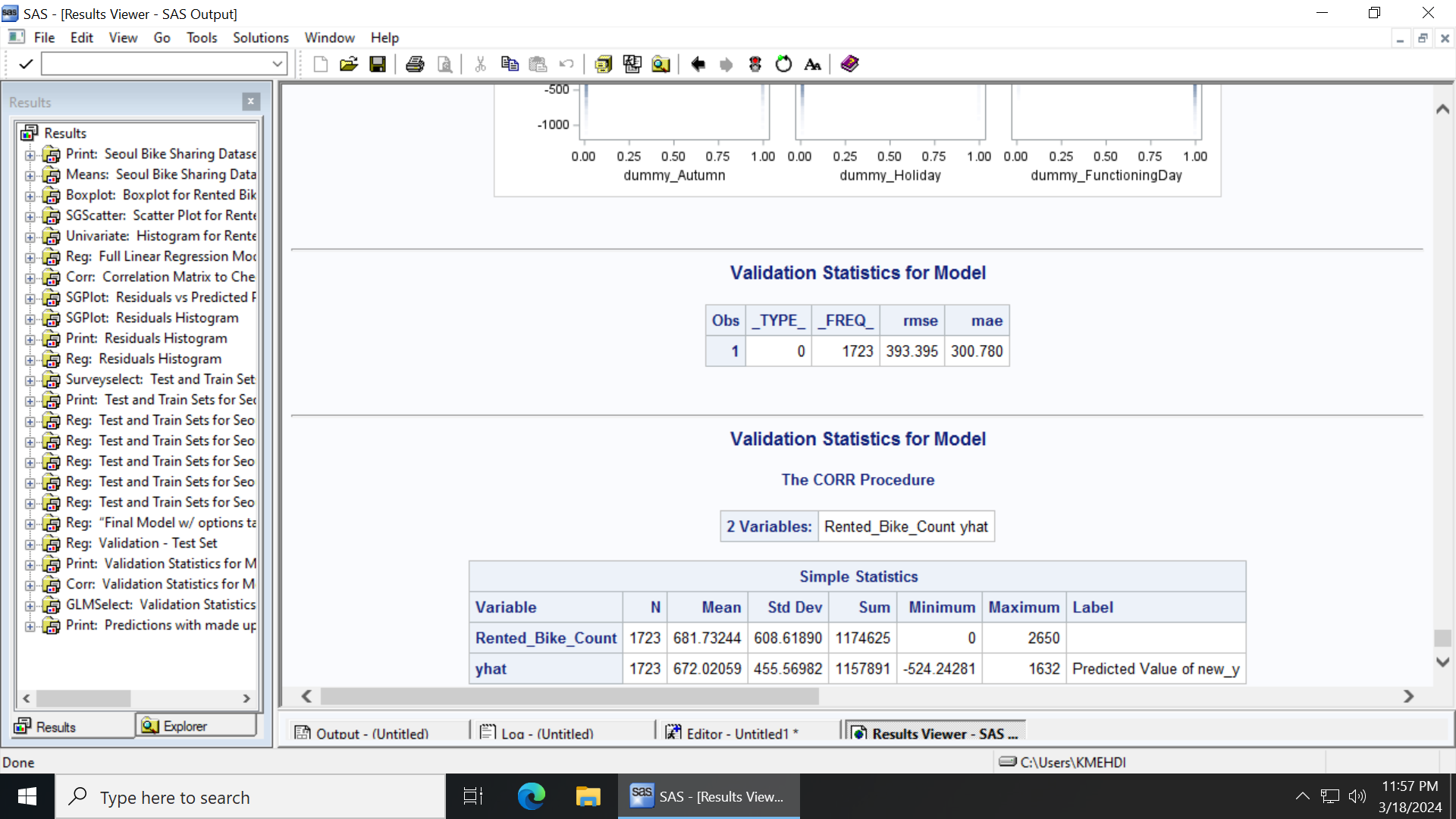
A5:



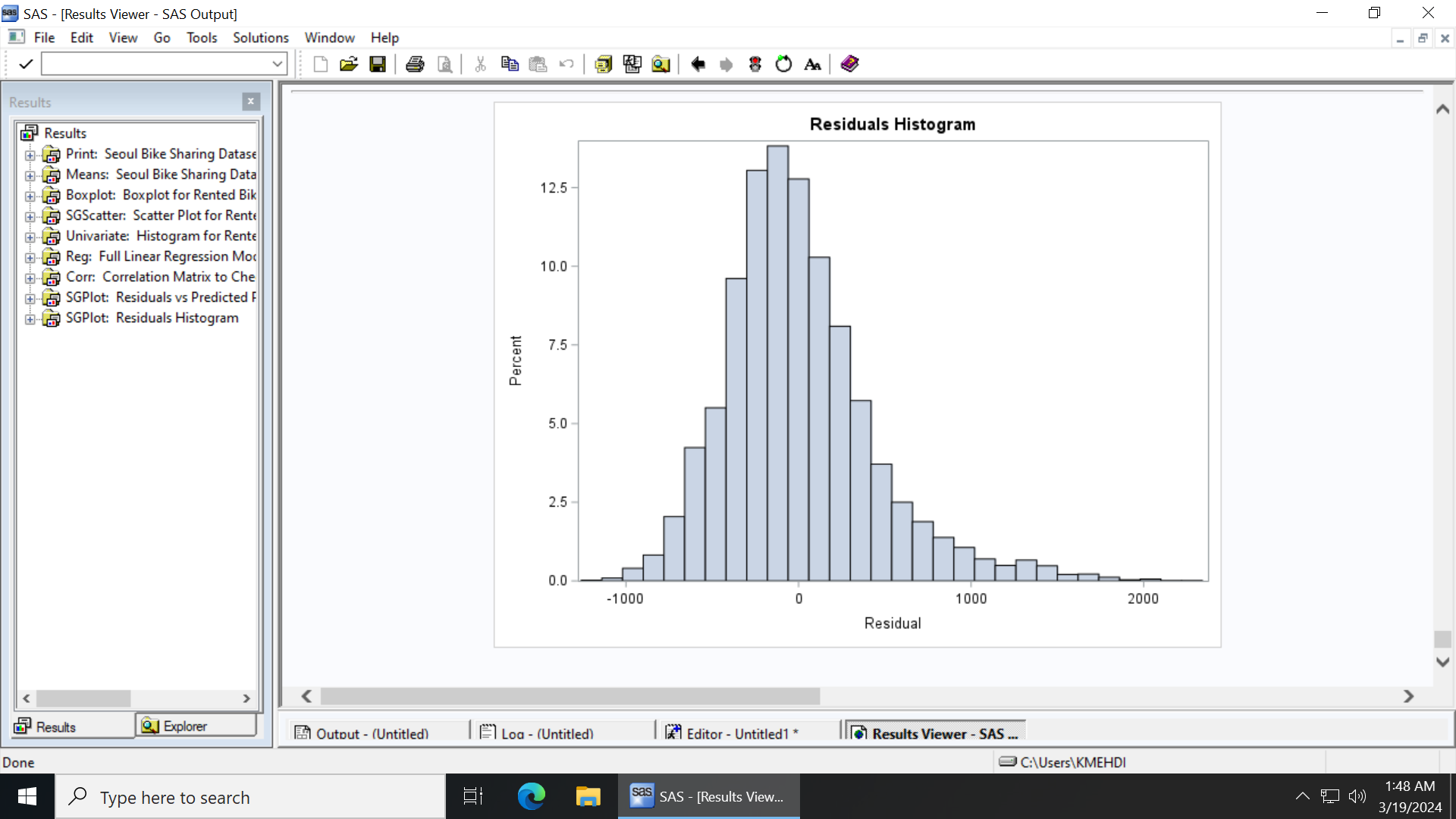
A6:



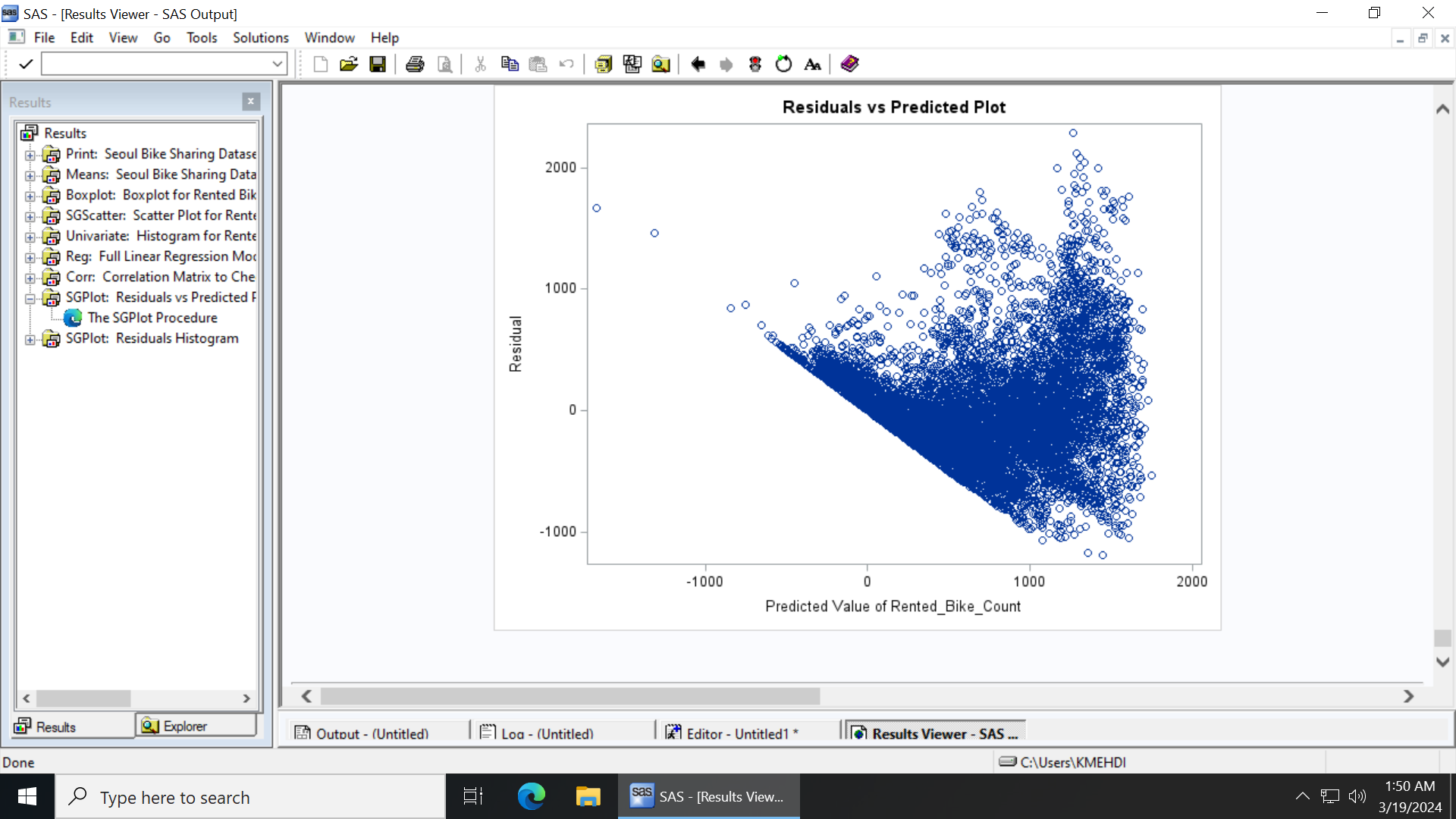
A7:



A8:



A9:



A10:

