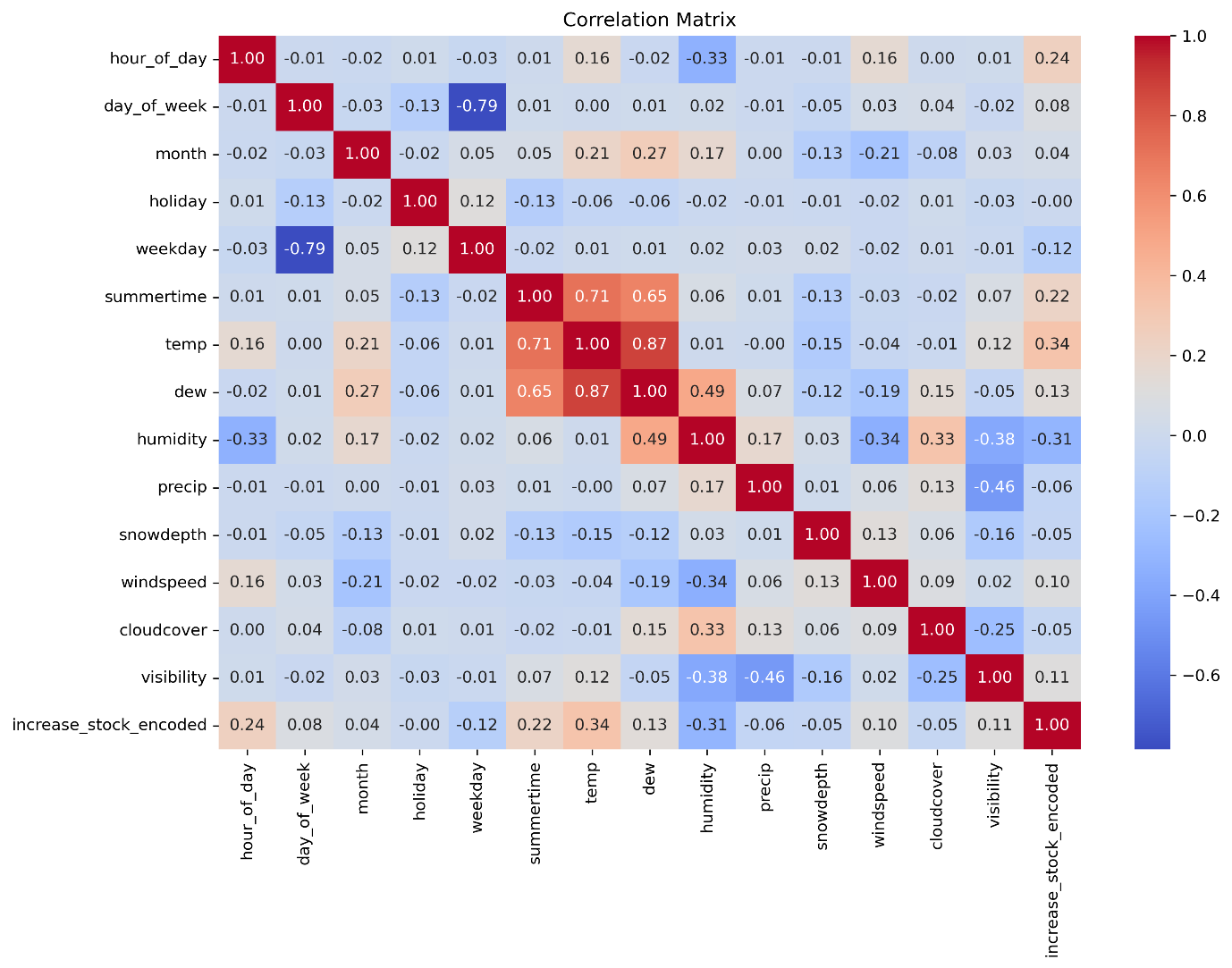
The dataset is analysed to identify null values, outliers, correlation between the predictor and target variables and multicollinearity in between predictor variables. The results are discussed below:

1. Null values are absent in all the feature columns, however, the column titled “Snow” is having values as ‘0’ in all the rows. Hence, the column is dropped from the dataset for further analysis.
2. The target variable is encoded into values ‘0’ for low bike demand and ‘1’ for high bike demand and a correlation heatmap (figure X) is constructed to gain insights about the nature and strength of correlation between the predictor variables and the target variable.

The last row or the last column in the correlation matrix gives us the corelation coefficients indicating that the predictor variable with highest influence on target is temperature followed by hour of the day and summertime. To find all the statistically significant predictor variables, a p-value statistic is calculated and features with p-value > 0.05 are labelled insignificant.

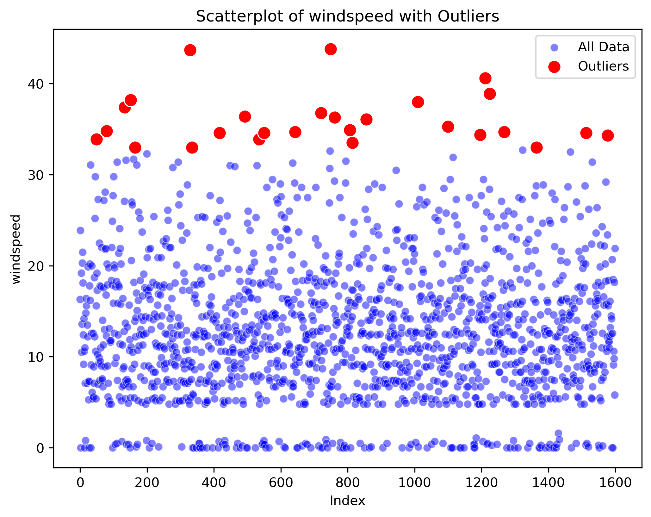
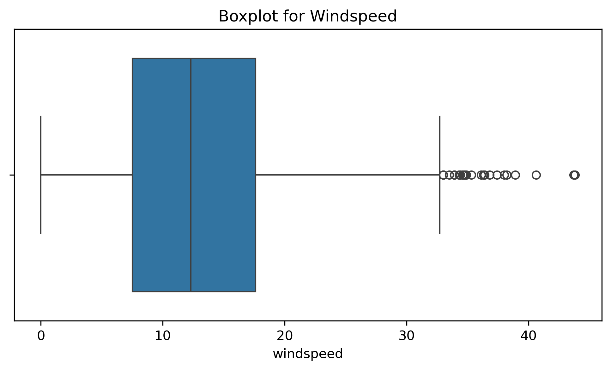


|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Correlation** | **P-Value** | **Significance** |
| hour\_of\_day | 0.240544 | 1.71E-22 | Statistically Significant |
| day\_of\_week | 0.083688 | 8.06E-04 | Statistically Significant |
| month | 0.037212 | 1.37E-01 | Not Significant |
| holiday | -0.004909 | 8.44E-01 | Not Significant |
| weekday | -0.116446 | 3.01E-06 | Statistically Significant |
| summertime | 0.216052 | 2.36E-18 | Statistically Significant |
| temp | 0.336981 | 8.78E-44 | Statistically Significant |
| dew | 0.132663 | 1.00E-07 | Statistically Significant |
| humidity | -0.308726 | 1.12E-36 | Statistically Significant |
| precip | -0.059304 | 1.77E-02 | Statistically Significant |
| snowdepth | -0.047526 | 5.73E-02 | Not Significant |
| windspeed | 0.096011 | 1.20E-04 | Statistically Significant |
| cloudcover | -0.045534 | 6.86E-02 | Not Significant |
| visibility | 0.113443 | 5.39E-06 | Statistically Significant |

High multi-collinearity can also be observed in between predictor variables from the correlation matrix. The highest positive corelation is observed in between temperature and dew with a value of 0.87, followed by high negative correlation between weekday and the day of week with a value of -0.79. Variance Inflation Factor (VIF) is calculated for each feature to understand which features can be retained or dropped while building the model. The table X shows that features: temperature, visibility, humidity and dew, exhibit high multi-collinearity. Insignificant features can be dropped and techniques like PCA, ridge or lasso regression can be used while building the models to handle this issue.

|  |  |
| --- | --- |
| **Feature** | **VIF** |
| hour\_of\_day | 4.419851 |
| day\_of\_week | 8.562966 |
| month | 5.242078 |
| holiday | 1.07955 |
| weekday | 9.103757 |
| summertime | 6.073319 |
| temp | 66.76301 |
| dew | 32.27144 |
| humidity | 43.98304 |
| precip | 1.289503 |
| snowdepth | 1.092724 |
| windspeed | 4.689751 |
| cloudcover | 6.032148 |
| visibility | 51.2273 |
|  |  |

1. Using the Inter Quartile Range (IQR) method, 27 outliers were found for the predictor variable wind speed. The figure XX shows a scatter plot of datapoints with the outliers marked in red along with a box plot showing outliers.

However, the predictor variables: precipitation, snow depth and visibility have highly skewed distributions (figure XXX) and it is not possible to find outliers using IQR method. Isolation forest was used to detect outliers and 80 outliers were found as shown in figure XXXX.

Major proportion of outliers are having target as low bike demand which is in line with real world scenario where we have low bike demand during extreme weather conditions. So, it is decided as infeasible to remove or alter them.

