Project

Introduction

This dataset encompasses bike share information gathered by a company operating in Washington, D.C., spanning the years 2011 and 2012. Each entry is identified by a unique 'instant' ID number and includes various hourly recorded attributes such as date ('dteday'), season, year ('yr'), month ('mnth'), hour ('hr'), and binary indicators for holidays and weekdays. Meteorological metrics like temperature ('temp'), humidity ('hum'), and windspeed are normalized to a scale between 0 and 1. The dataset also provides counts of casual and registered riders, as well as a total count representing the overall number of riders during each recorded hour. Instances with zero bike usage were excluded from the dataset. This extensive set of variables enables a detailed examination of bike share patterns and their relationships with temporal and weather-related factors.

The below is the Descriptive Analysis of the bike share data

The Descriptive Analysis of the bike share data

	instant	season	yr	mnth	hr
count	17379.0000	17379.000000	17379.000000	17379.000000	17379.000000
mean	8690.0000	2.501640	0.502561	6.537775	11.546752
std	5017.0295	1.106918	0.500008	3.438776	6.914405
min	1.0000	1.000000	0.000000	1.000000	0.000000
25%	4345.5000	2.000000	0.000000	4.000000	6.000000
50%	8690.0000	3.000000	1.000000	7.000000	12.000000
75%	13034.5000	3.000000	1.000000	10.000000	18.000000
max	17379.0000	4.000000	1.000000	12.000000	23.000000

	holiday	weekday	workingday	temp	hum
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	0.028770	3.003683	0.682721	0.496987	0.627229
std	0.167165	2.005771	0.465431	0.192556	0.192930
min	0.000000	0.000000	0.000000	0.020000	0.000000
25%	0.000000	1.000000	0.000000	0.340000	0.480000
50%	0.000000	3.000000	1.000000	0.500000	0.630000
75%	0.000000	5.000000	1.000000	0.660000	0.780000
max	1.000000	6.000000	1.000000	1.000000	1.000000

	windspeed	casual	registered	count
count	17379.000000	17379.000000	17379.000000	17379.000000
mean	0.190098	35.676218	153.786869	189.463088
std	0.122340	49.305030	151.357286	181.387599
min	0.000000	0.000000	0.000000	1.000000
25%	0.104500	4.000000	34.000000	40.000000
50%	0.194000	17.000000	115.000000	142.000000
75%	0.253700	48.000000	220.000000	281.000000
max	0.850700	367.000000	886.000000	977.000000

The below is the insights from the descriptive analysis of the data

Season:

The data covers four seasons (1 to 4), with an average season value of approximately 2.5.

The standard deviation is around 1.11, indicating a moderate amount of variation.

Seasons are evenly distributed, as the mean is close to the midpoint (2.5).

Casual, Registered, and Total Count:

Casual, registered, and total bike counts (casual, registered, and count) have right-skewed distributions.

The mean count is 189.46, with a standard deviation of approximately 181.39.

The minimum count is 1, and the maximum count is 977.

Holiday:

The holiday variable is binary (0 or 1), indicating whether it is a holiday or not.

The dataset has a low average holiday occurrence (approximately 2.88%).

Workingday:

A binary variable (0 or 1) indicating whether it is a working day or not. The average suggests that around 68.27% of the instances are working days.

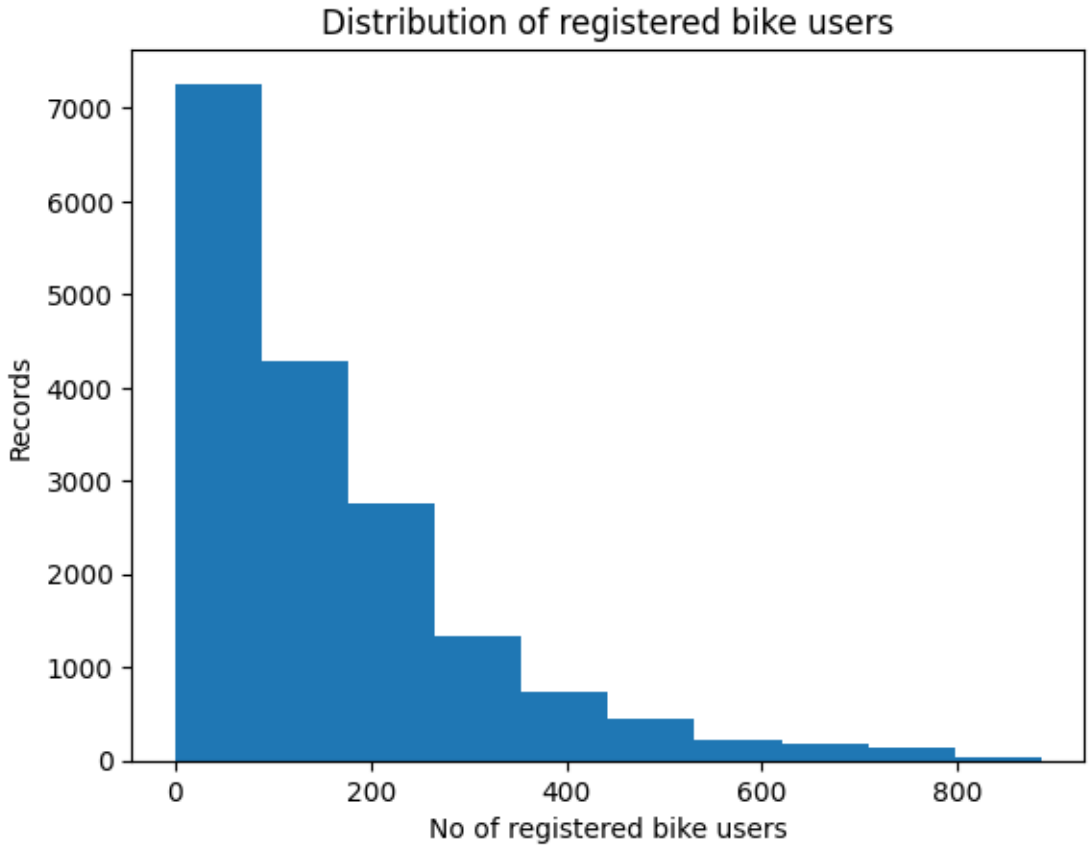
Temperature (temp): Temperature values range from 0.02 to 1, with an average of 0.5. The temperature distribution seems well-spread.

Humidity (hum): Humidity values range from 0 to 1, with an average of 0.63. The humidity distribution appears to be moderate.

Windspeed:

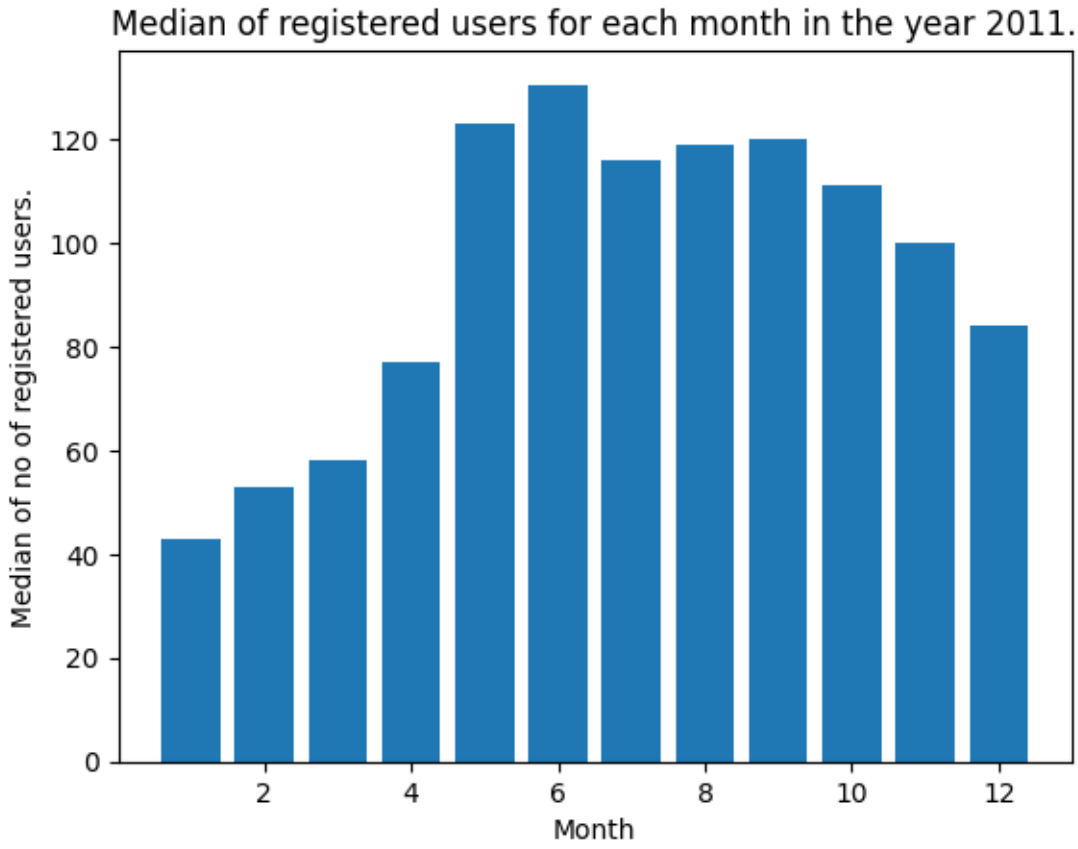
Windspeed values range from 0 to 0.85, with an average of 0.19. There is a moderate variation in windspeed.

Histogram to better understand the distribution of no of registered bike users.

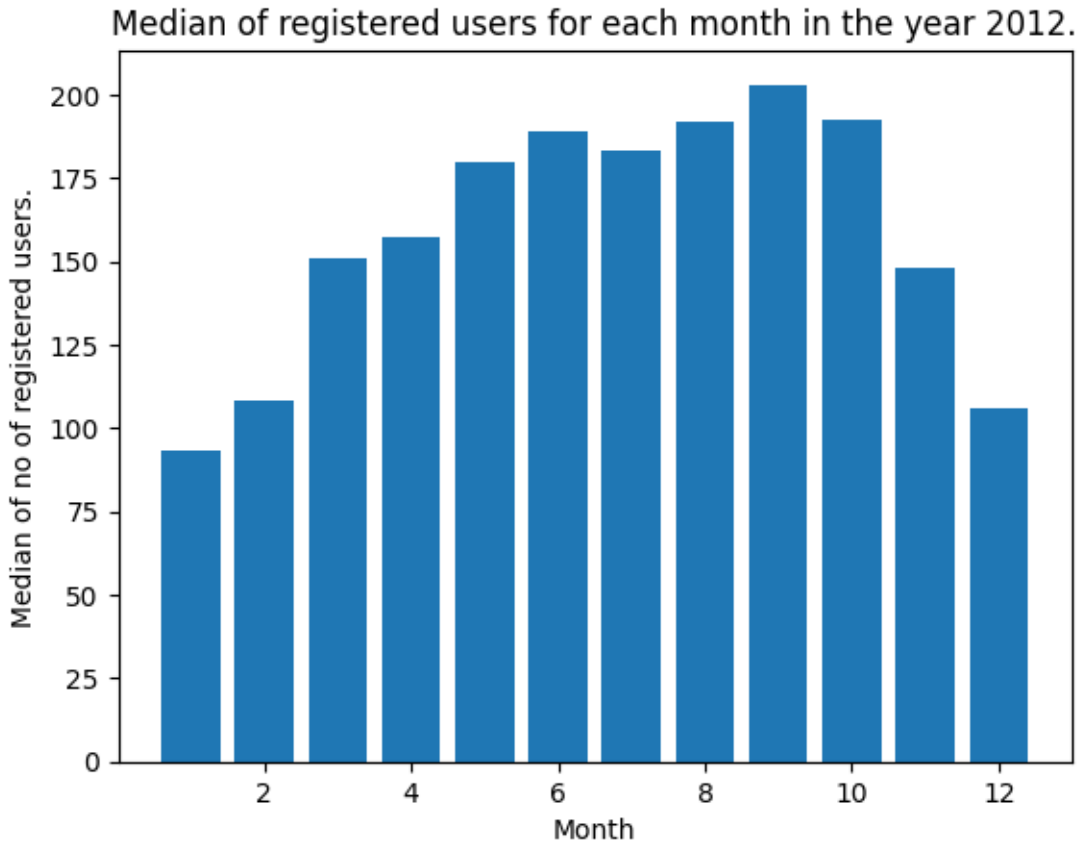


From the above histogram, we can conclude that the no of registered bike users are decreasing steadily.

A bar plot that shows the median number of registered riders (grouped by month) for each month for the year 2011.



Another bar plot that shows the median number of registered riders (grouped by month) for each month for the year 2012.



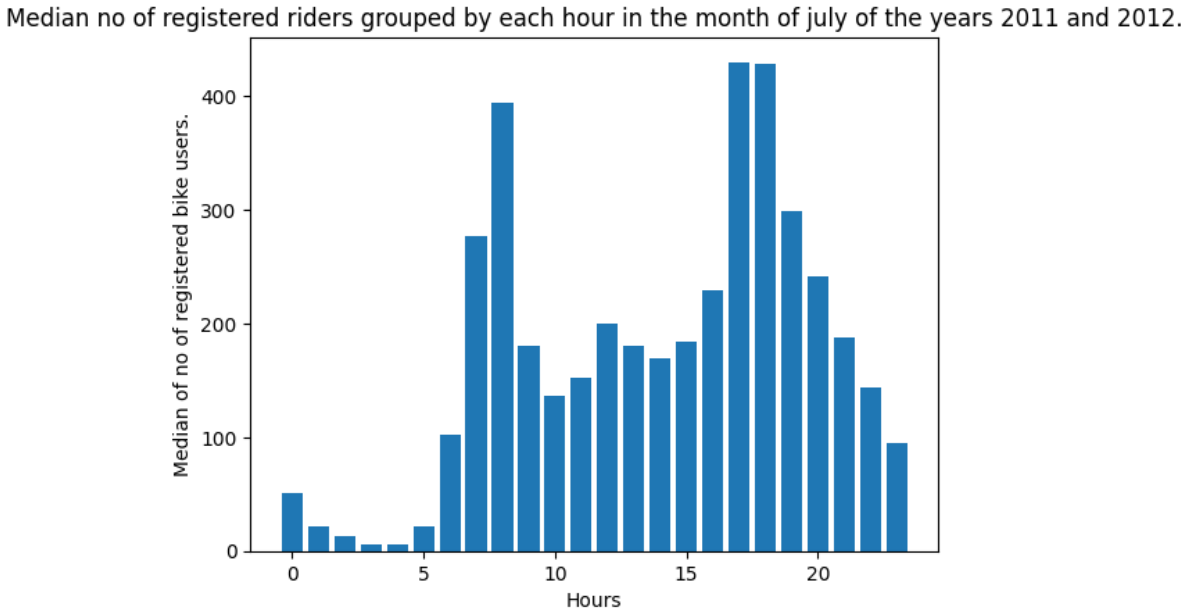
Insights from the above bar graph

1.In the year 2011, in the 6th month the median is high which is around 130. That implies there are more no of registered bike users are in the 6th month. Whereaas in the year 2012, 9th month has more no of registered bike users.

2.In both years, the no of registered bike users are less in the first month.

3.In the year 2012 has more no of registered bike users compared to the year 2011 since the highest median of 2012 is higher than the highest median of 2011.

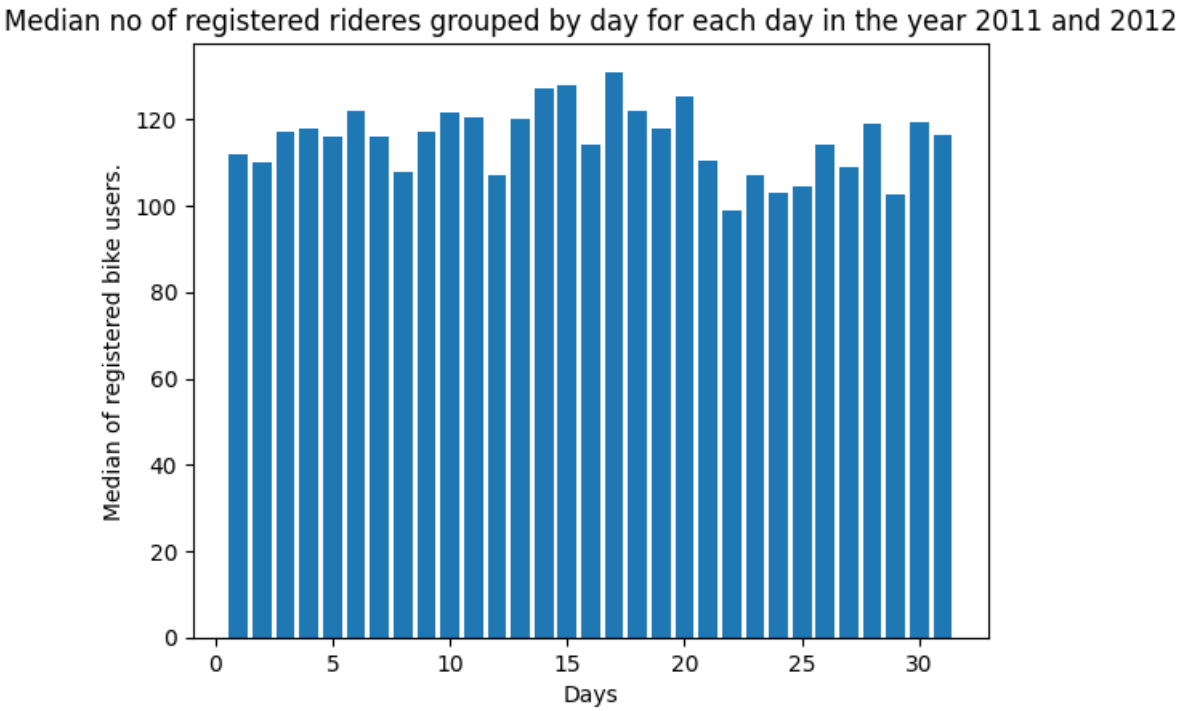
A bar plot showing the median number of registered riders (grouped by hour) for each hour for the month of July (include both years).



1. In the month of july of both year 2011 and 2012, the 16th &17th hour has more no of registered bike users.

2.At the 3rd and 4th hour there are least no of registered bike users.

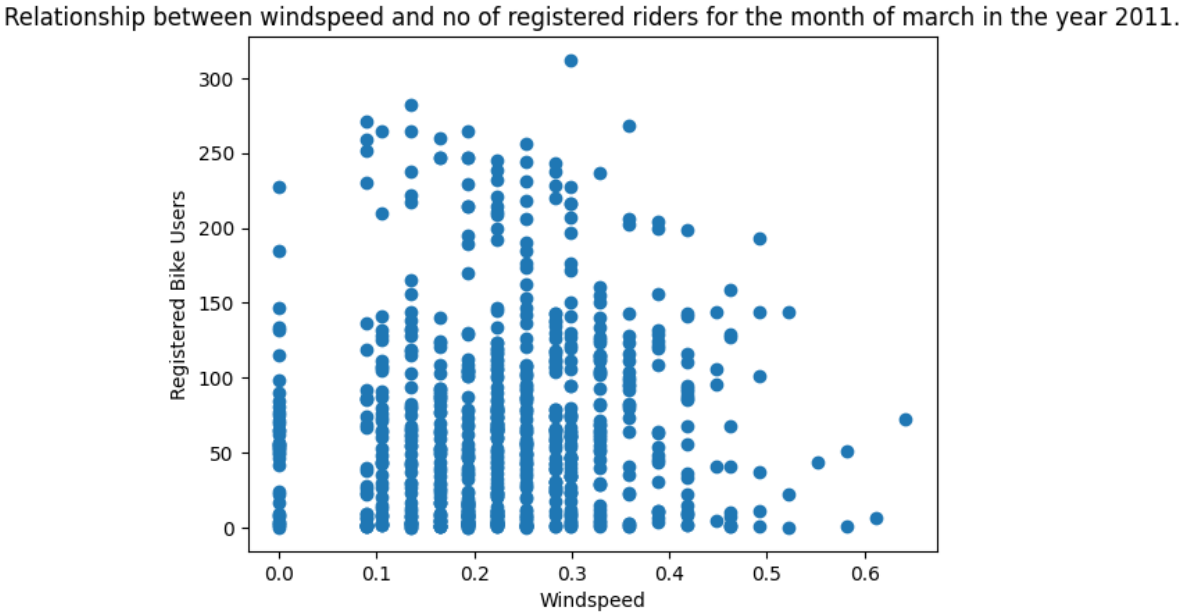
A bar plot showing the median number of registered riders (grouped by day) for each day (include both years)



1.On the 17th day, no of registered bike users are more. And it was on day 21 where the no of registered bike users are least.

2.Almost every day has almost similar no of registered bike users.

A scatter plot to show the relationship between windspeed and the number of registered riders only for the month of March in the year 2011.



1. As windspeed increases, the no of registered bike users decreases.

2. The more no of registered bike users are at the windspeed 0.3

The linear regression model to predict the no of registered bike users.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             registered   R-squared:                       0.045
Model:                            OLS   Adj. R-squared:                  0.045
Method:                 Least Squares   F-statistic:                     203.5
Date:                Sun, 10 Mar 2024   Prob (F-statistic):          7.31e-171
Time:                        07:05:15   Log-Likelihood:            -1.1150e+05
No. Observations:               17379   AIC:                         2.230e+05
Df Residuals:                   17374   BIC:                         2.230e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     60.0541      3.901     15.394      0.000      52.408      67.700
season        26.0401      1.026     25.392      0.000      24.030      28.050
holiday      -40.1160      6.749     -5.944      0.000     -53.345     -26.886
weekday        1.2243      0.563      2.177      0.030       0.122       2.327
windspeed    137.1213      9.279     14.778      0.000     118.934     155.309
==============================================================================
Omnibus:                     4853.240   Durbin-Watson:                   0.416
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12413.974
Skew:                           1.528   Prob(JB):                         0.00
Kurtosis:                       5.793   Cond. No.                         37.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Intercept: 60.054059655594514
Season Coefficient: 26.040123720006235
Holiday Coefficient: -40.11599244819463
Weekday Coefficient: 1.2243374021026499
Windspeed Coefficient: 137.1212803908689

All Coefficients:
Intercept     60.054060
season        26.040124
holiday      -40.115992
weekday        1.224337
```

```
[15] holiday      -40.115992
     weekday        1.224337
     windspeed    137.121280
     dtype: float64
     The predicted number of registered users: 105.73390873419709
```

Model Significance: The overall model has statistical significance, as indicated by the F-statistic (203.5) and its associated p-value (7.31e-171). This suggests that at least one of the predictors is related to the dependent variable.

R-squared Value: The R-squared value is 0.045, indicating that the model explains approximately 4.5% of the variance in the registered users' count. While this is a relatively low percentage, it suggests that the selected predictors contribute somewhat to the variability in registered users.

Individual Coefficients: The coefficients for each predictor provide information about the strength and direction of their relationship with the registered users. For instance: Season has a positive coefficient (26.04), implying that as the season variable increases, the number of registered users tends to increase. Holiday has a negative coefficient (-40.12), suggesting a decrease in registered users on holidays compared to non-holidays. Weekday has a positive coefficient (1.22), indicating a slight increase in registered users on weekdays. Windspeed has a relatively large positive coefficient (137.12), suggesting a substantial impact on the registered users' count with increasing windspeed.

Intercept Interpretation: The intercept (60.05) represents the estimated number of registered users when all predictors are zero. In this context, it might not have a practical interpretation since some predictors, like season and windspeed, cannot be zero.

Predicted Count: The predicted number of registered users for a specific set of predictor values (season, holiday, weekday, windspeed) is approximately 105.73. This value provides an estimate based on the linear regression model, taking into account the coefficients and predictor values.