STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

# Bike Sharing Demand

Forecast use of a city bike share system

Akash Chawla (10406187)
Rohan Bhoyar (10411225)

# Project Outline

- Bike sharing systems are a means of renting
- Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis.
- These bike sharing programs have gained momentum over past few years.
- Accessibility and affordability have helped to promote the concept of a short-term bike rental system.
- In this problem, we have a combination of historical usage patterns along with weather data in order to forecast bike rental demand in the Capital Bike share program in Washington, D.C. We are provided hourly rental data spanning two years.

# Data Source

A closed competition at Kaggle

Weblink: https://www.kaggle.com/c/bike-sharing-demand/data

Citation: Fanaee-T, Hadi, and Gama, Joao, *Event labeling combining ensemble detectors and background knowledge*, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

# Data Fields

| | |
|---|---|
| **datetime** | hourly date + timestamp |
| **season** | 1 = spring, 2 = summer, 3 = fall, 4 = winter |
| **holiday** | whether the day is considered a holiday |
| **workingday** | whether the day is neither a weekend nor holiday |
| **weather** | 1: Clear, Few clouds, Partly cloudy, Partly cloudy |
| | 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist |
| | 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds |
| | 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| **temp** | temperature in Celsius |
| **atemp** | feels like temperature in Celsius |
| **humidity** | relative humidity |
| **windspeed** | wind speed |
| **casual** | number of non-registered user rentals initiated |
| **registered** | number of registered user rentals initiated |
| **count** | number of total rentals |

# Snapshot of Data

| datetime | season | holiday | workingday | weather | temp | atemp | humidity | windspeed | casual | registered | count |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1/3/2011 0:00 | 1 | 0 | 1 | 1 | 9.02 | 9.85 | 44 | 23.9994 | 0 | 5 | 5 |
| 1/3/2011 1:00 | 1 | 0 | 1 | 1 | 8.2 | 8.335 | 44 | 27.9993 | 0 | 2 | 2 |
| 1/3/2011 4:00 | 1 | 0 | 1 | 1 | 6.56 | 6.82 | 47 | 26.0027 | 0 | 1 | 1 |
| 1/3/2011 5:00 | 1 | 0 | 1 | 1 | 6.56 | 6.82 | 47 | 19.0012 | 0 | 3 | 3 |
| 1/3/2011 6:00 | 1 | 0 | 1 | 1 | 5.74 | 5.305 | 50 | 26.0027 | 0 | 30 | 30 |
| 1/3/2011 7:00 | 1 | 0 | 1 | 1 | 5.74 | 6.82 | 50 | 12.998 | 1 | 63 | 64 |
| 1/3/2011 8:00 | 1 | 0 | 1 | 1 | 5.74 | 6.06 | 50 | 19.0012 | 1 | 153 | 154 |
| 1/3/2011 9:00 | 1 | 0 | 1 | 1 | 6.56 | 6.82 | 43 | 26.0027 | 7 | 81 | 88 |
| 1/3/2011 10:00 | 1 | 0 | 1 | 1 | 7.38 | 8.335 | 43 | 16.9979 | 11 | 33 | 44 |
| 1/3/2011 11:00 | 1 | 0 | 1 | 1 | 8.2 | 9.09 | 40 | 22.0028 | 10 | 41 | 51 |
| 1/3/2011 12:00 | 1 | 0 | 1 | 1 | 9.02 | 10.605 | 35 | 19.9995 | 13 | 48 | 61 |
| 1/3/2011 13:00 | 1 | 0 | 1 | 1 | 9.84 | 10.605 | 35 | 19.0012 | 8 | 53 | 61 |
| 1/3/2011 14:00 | 1 | 0 | 1 | 1 | 10.66 | 12.12 | 30 | 19.0012 | 11 | 66 | 77 |
| 1/3/2011 15:00 | 1 | 0 | 1 | 1 | 10.66 | 12.12 | 30 | 16.9979 | 14 | 58 | 72 |
| 1/3/2011 16:00 | 1 | 0 | 1 | 1 | 10.66 | 12.12 | 30 | 16.9979 | 9 | 67 | 76 |
| 1/3/2011 17:00 | 1 | 0 | 1 | 1 | 9.84 | 11.365 | 30 | 15.0013 | 11 | 146 | 157 |
| 1/3/2011 18:00 | 1 | 0 | 1 | 1 | 9.84 | 12.88 | 32 | 7.0015 | 9 | 148 | 157 |
| 1/3/2011 19:00 | 1 | 0 | 1 | 1 | 8.2 | 12.88 | 47 | 0 | 8 | 102 | 110 |
| 1/3/2011 20:00 | 1 | 0 | 1 | 1 | 8.2 | 11.365 | 47 | 7.0015 | 3 | 49 | 52 |
| 1/3/2011 21:00 | 1 | 0 | 1 | 1 | 7.38 | 9.85 | 64 | 8.9981 | 3 | 49 | 52 |
| 1/3/2011 22:00 | 1 | 0 | 1 | 1 | 5.74 | 7.575 | 69 | 8.9981 | 0 | 20 | 20 |
| 1/3/2011 23:00 | 1 | 0 | 1 | 1 | 7.38 | 10.605 | 55 | 7.0015 | 1 | 11 | 12 |

# Analysis of Variables

The UNIVARIATE Procedure
Variable: humidity

| Moments | | | |
|---|---|---|---|
| N | 10863 | Sum Weights | 10863 |
| Mean | 62.0095738 | Sum Observations | 673610 |
| Std Deviation | 19.0612088 | Variance | 363.329682 |
| Skewness | -0.0392811 | Kurtosis | -0.8918317 |
| Uncorrected SS | 45716756 | Corrected SS | 3946487 |
| Coeff Variation | 30.7391386 | Std Error Mean | 0.18288393 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 62.00957 | Std Deviation | 19.06121 |
| Median | 62.00000 | Variance | 363.32968 |
| Mode | 88.00000 | Range | 92.00000 |
| | | Interquartile Range | 30.00000 |

| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| 8 | 853 | 100 | 10654 |
| 10 | 852 | 100 | 10655 |
| 12 | 854 | 100 | 10658 |
| 13 | 856 | 100 | 10659 |
| 14 | 857 | 100 | 10818 |

| Missing Values | | | |
|---|---|---|---|
| Missing Value | Count | Percent Of | |
| | | All Obs | Missing Obs |
| . | 22 | 0.20 | 100.00 |

# Analysis of Variables

The UNIVARIATE Procedure
Variable: windspeed

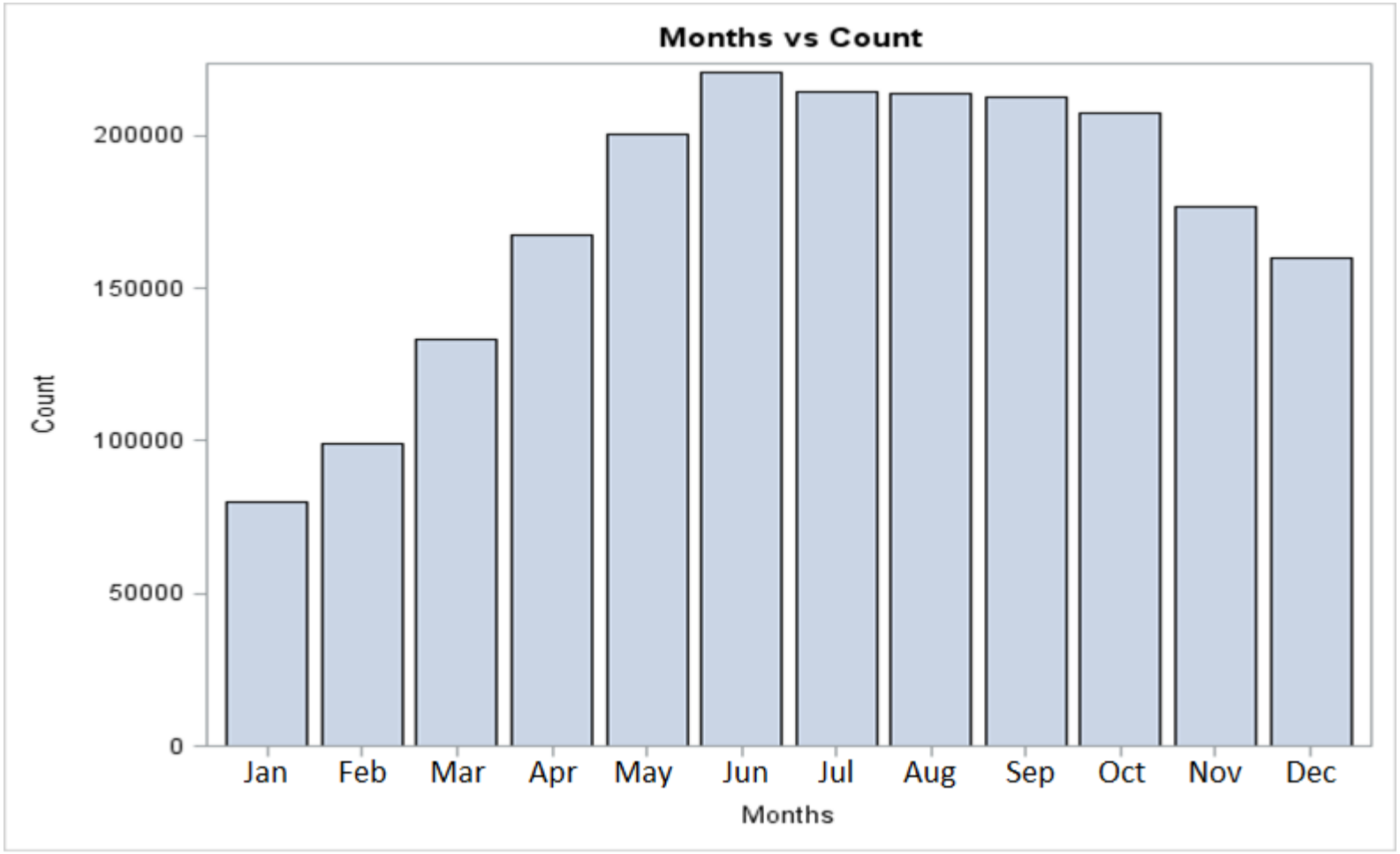
| Moments | | | |
|---|---|---|---|
| N | 9572 | Sum Weights | 9572 |
| Mean | 14.5558102 | Sum Observations | 139328.215 |
| Std Deviation | 7.08850542 | Variance | 50.246909 |
| Skewness | 1.07235492 | Kurtosis | 1.37238601 |
| Uncorrected SS | 2508948.22 | Corrected SS | 480913.167 |
| Coeff Variation | 48.6988036 | Std Error Mean | 0.07245249 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 14.55581 | Std Deviation | 7.08851 |
| Median | 12.99800 | Variance | 50.24691 |
| Mode | 8.99810 | Range | 50.99370 |
| | | Interquartile Range | 10.00310 |

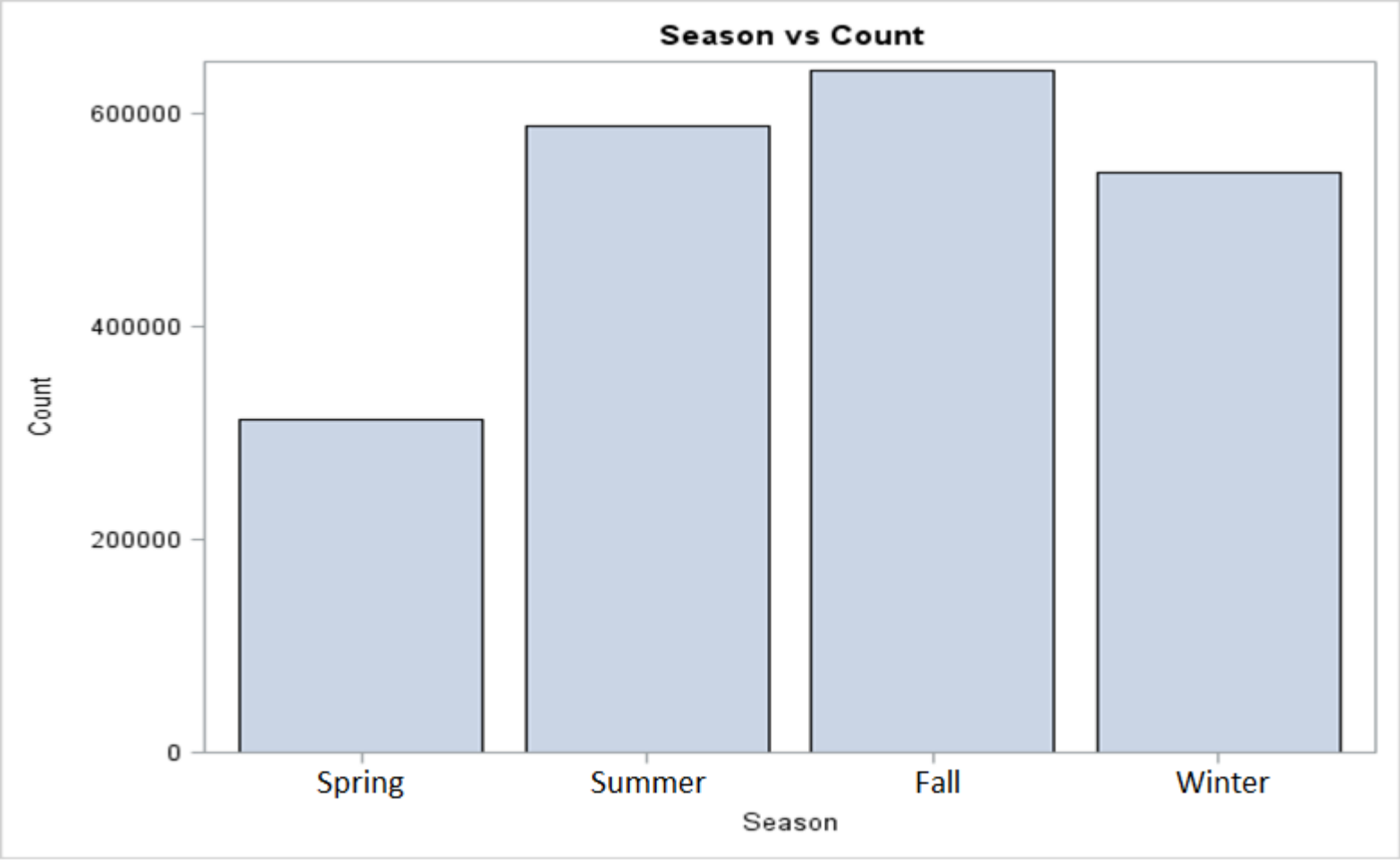
| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| 6.0032 | 10884 | 47.9988 | 6988 |
| 6.0032 | 10868 | 50.0021 | 869 |
| 6.0032 | 10867 | 51.9987 | 761 |
| 6.0032 | 10864 | 56.9969 | 2756 |
| 6.0032 | 10863 | 56.9969 | 2757 |

| Missing Values | | | |
|---|---|---|---|
| Missing Value | Count | Percent Of | |
| | | All Obs | Missing Obs |
| . | 1313 | 12.06 | 100.00 |

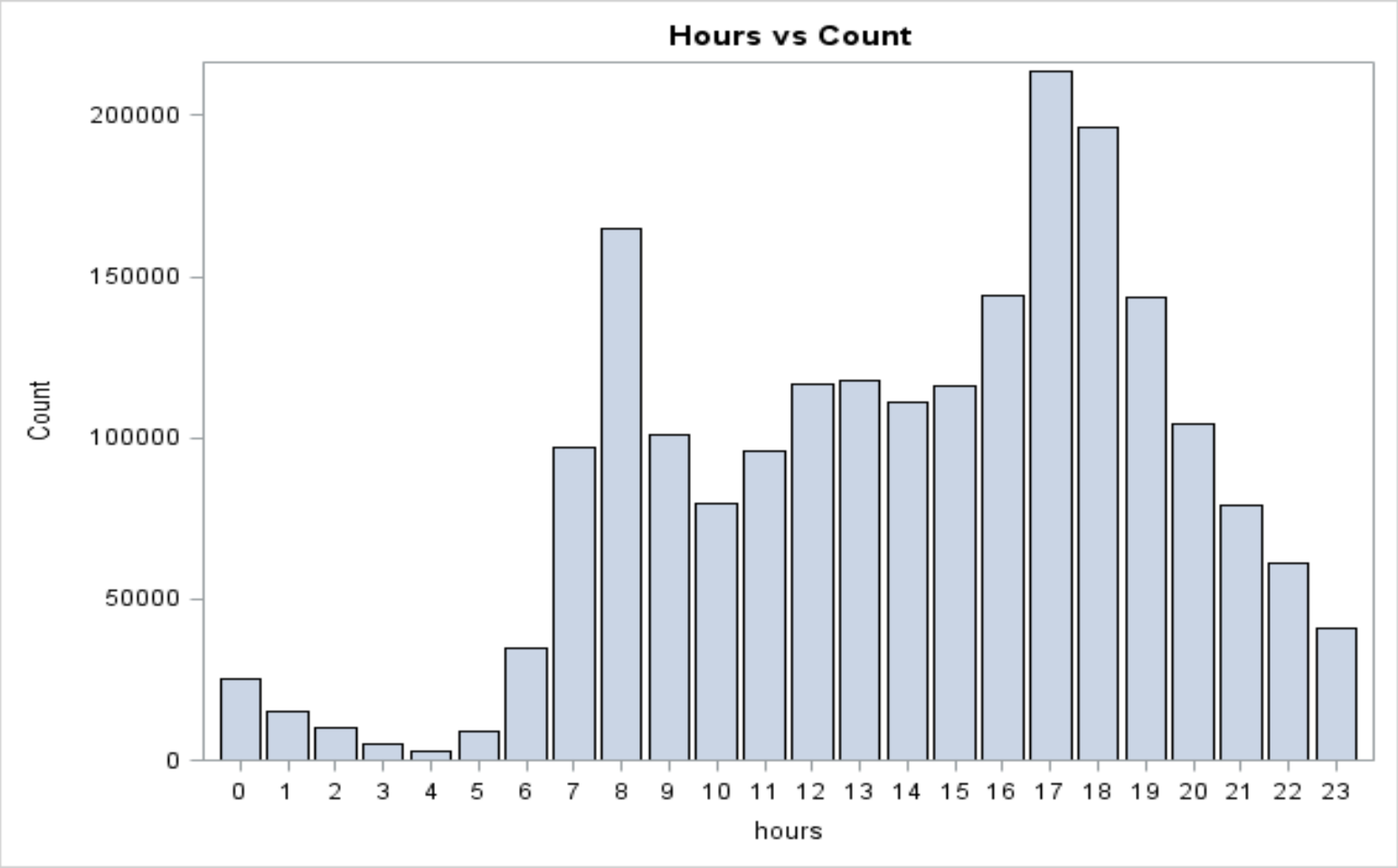
# Plots


Months vs Count
Count
200000
150000
100000
50000
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months

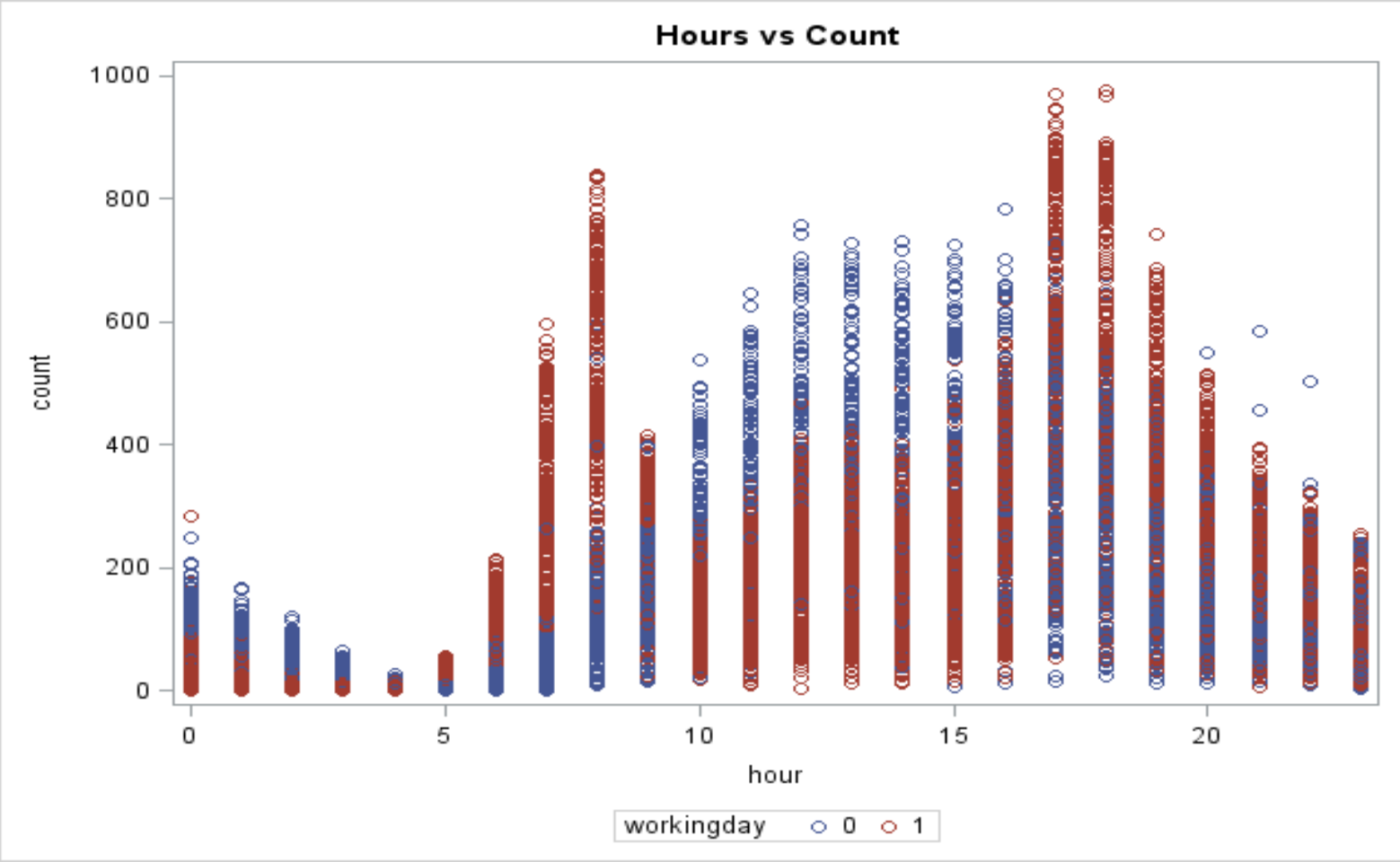
# Plots


Season vs Count
600000
400000
200000
0
Count
Spring
Summer
Fall
Winter
Season

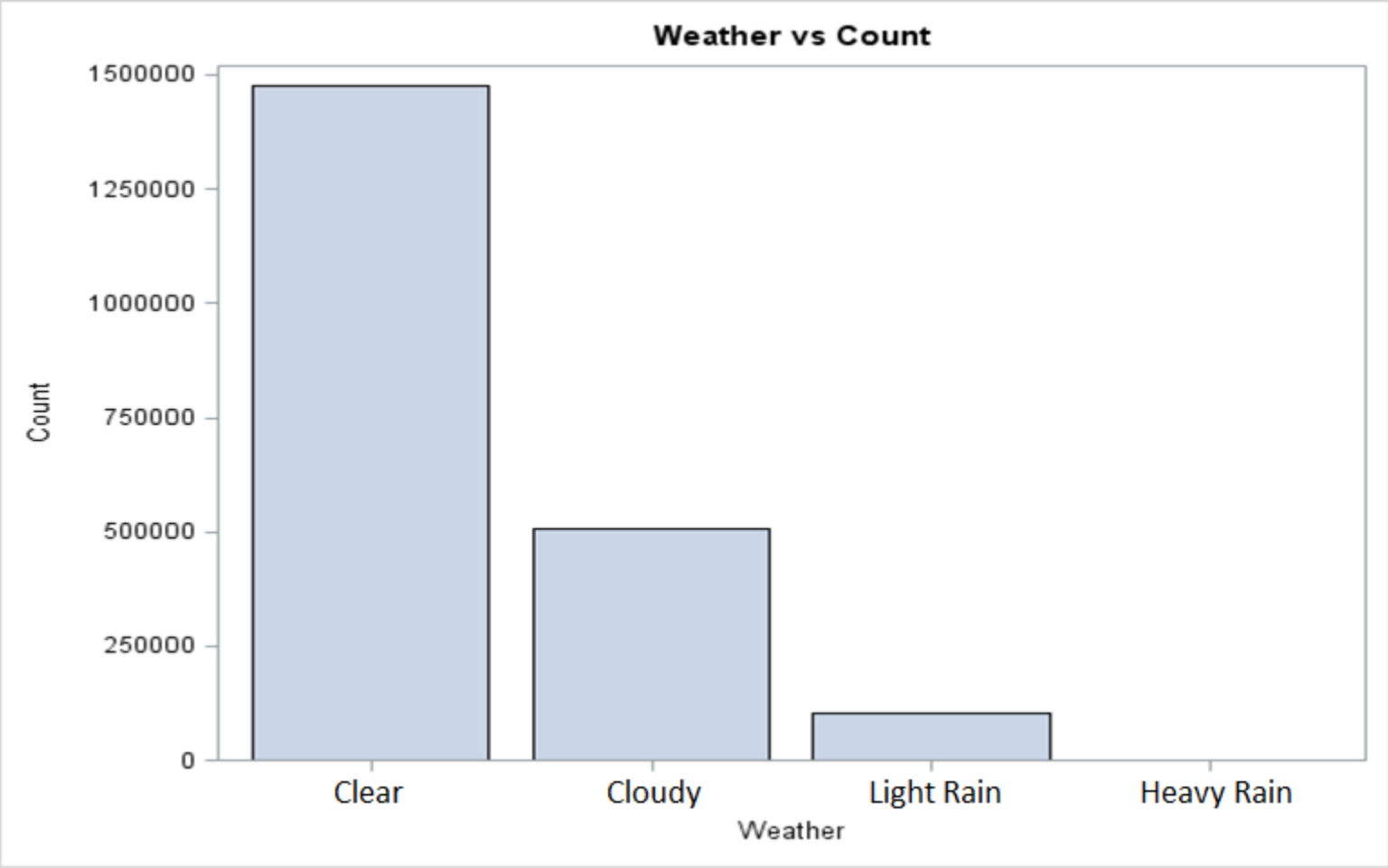
# Plots


Hours vs Count
Count
200000
150000
100000
50000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hours

# Plots


Hours vs Count
count
1000
800
600
400
200
0
0
5
10
15
20
hour
workingday
0
1

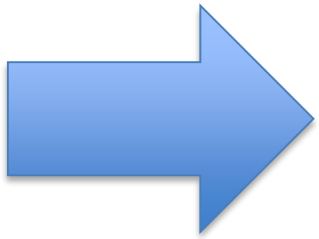
# Plots


Weather vs Count
1500000
1250000
1000000
750000
500000
250000
0
Count
Clear
Cloudy
Light Rain
Heavy Rain
Weather

# Data Modifications

- Handling of Date time values

| datetime |
|---|
| 03JAN11:00:00:00 |
| 03JAN11:01:00:00 |
| 03JAN11:04:00:00 |
| 03JAN11:05:00:00 |
| 03JAN11:06:00:00 |
| 03JAN11:07:00:00 |
| 03JAN11:08:00:00 |
| 03JAN11:09:00:00 |
| 03JAN11:10:00:00 |
| 03JAN11:11:00:00 |
| 03JAN11:12:00:00 |
| 03JAN11:13:00:00 |
| 03JAN11:14:00:00 |
| 03JAN11:15:00:00 |



| date | time | hour |
|---|---|---|
| 03JAN2011 | 0:00:00 | 0 |
| 03JAN2011 | 1:00:00 | 1 |
| 03JAN2011 | 4:00:00 | 4 |
| 03JAN2011 | 5:00:00 | 5 |
| 03JAN2011 | 6:00:00 | 6 |
| 03JAN2011 | 7:00:00 | 7 |
| 03JAN2011 | 8:00:00 | 8 |
| 03JAN2011 | 9:00:00 | 9 |
| 03JAN2011 | 10:00:00 | 10 |
| 03JAN2011 | 11:00:00 | 11 |
| 03JAN2011 | 12:00:00 | 12 |
| 03JAN2011 | 13:00:00 | 13 |
| 03JAN2011 | 14:00:00 | 14 |
| 03JAN2011 | 15:00:00 | 15 |

# Data Modifications

- Handling of Categorical Values

➢ Month:

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Jan | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Feb | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mar | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Apr | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| May | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Jun | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Jul | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Aug | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Sep | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Oct | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Nov | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Dec | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

➢ Season:

➢ Weather:

| spring | summer | fall | w_clear | w_cloudy |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 |

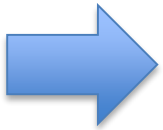
# Data Modifications

- Handling of Missing Values

➢ Only Wind Speed and Humidity variables had missing values, which were imputed by 0 (zero)

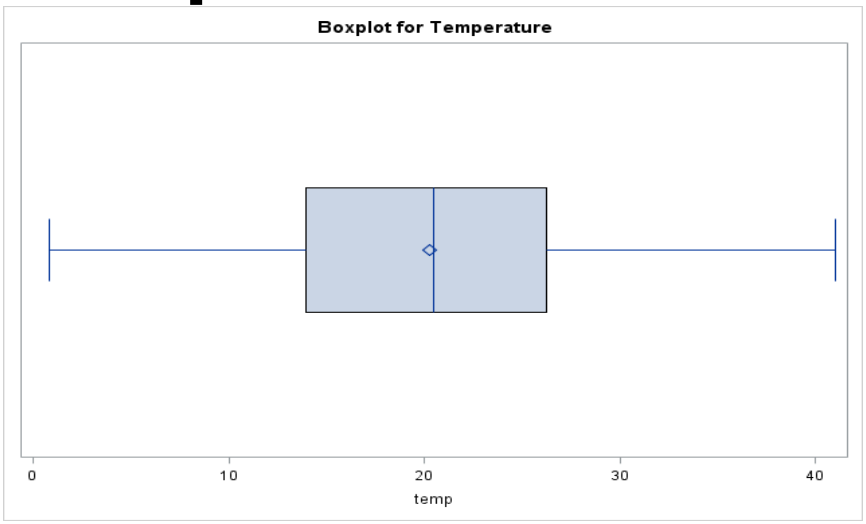
➢ So we replaced 0 values with Median

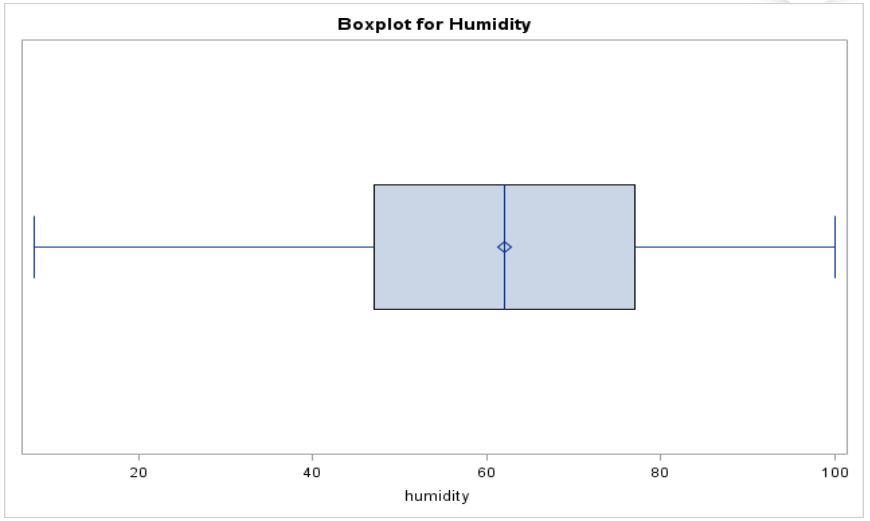
| temp | atemp | humidity | windspeed |
|---|---|---|---|
| 23.78 | 27.275 | 46 | 19.0012 |
| 25.42 | 31.06 | 43 | 23.9994 |
| 25.42 | 31.06 | 41 | 16.9979 |
| 26.24 | 31.06 | 38 | 19.9995 |
| 26.24 | 31.06 | 38 | 16.9979 |
| 26.24 | 31.06 | 38 | 16.9979 |
| 26.24 | 31.06 | 36 | 0 |
| 26.24 | 31.06 | 36 | 0 |
| 26.24 | 31.06 | 36 | 11.0014 |
| 25.42 | 31.06 | 35 | 19.0012 |
| 24.6 | 31.06 | 40 | 7.0015 |
| 23.78 | 27.275 | 46 | 8.9981 |
| 22.96 | 26.515 | 52 | 7.0015 |
| 22.14 | 25.76 | 60 | 11.0014 |
| 22.14 | 25.76 | 56 | 8.9981 |

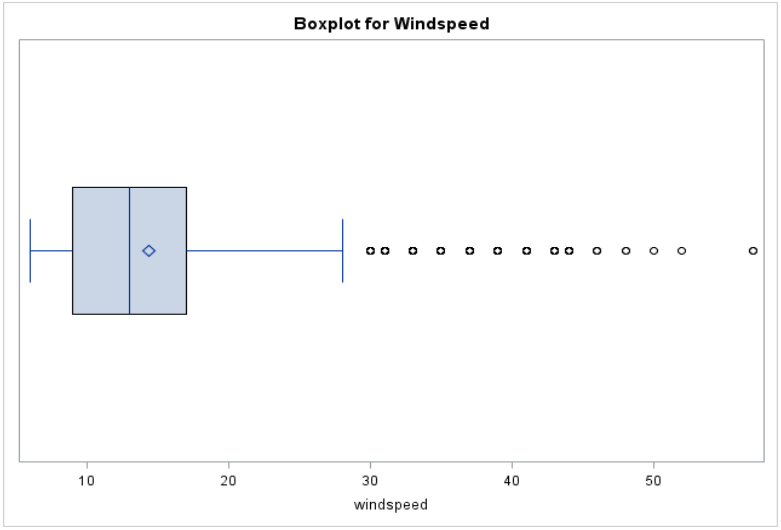


| temp | atemp | humidity | windspeed |
|---|---|---|---|
| 23.78 | 27.275 | 46 | 19.0012 |
| 25.42 | 31.06 | 43 | 23.9994 |
| 25.42 | 31.06 | 41 | 16.9979 |
| 26.24 | 31.06 | 38 | 19.9995 |
| 26.24 | 31.06 | 38 | 16.9979 |
| 26.24 | 31.06 | 38 | 16.9979 |
| 26.24 | 31.06 | 36 | 12.998 |
| 26.24 | 31.06 | 36 | 12.998 |
| 26.24 | 31.06 | 36 | 11.0014 |
| 25.42 | 31.06 | 35 | 19.0012 |
| 24.6 | 31.06 | 40 | 7.0015 |
| 23.78 | 27.275 | 46 | 8.9981 |
| 22.96 | 26.515 | 52 | 7.0015 |
| 22.14 | 25.76 | 60 | 11.0014 |
| 22.14 | 25.76 | 56 | 8.9981 |

# Boxplots

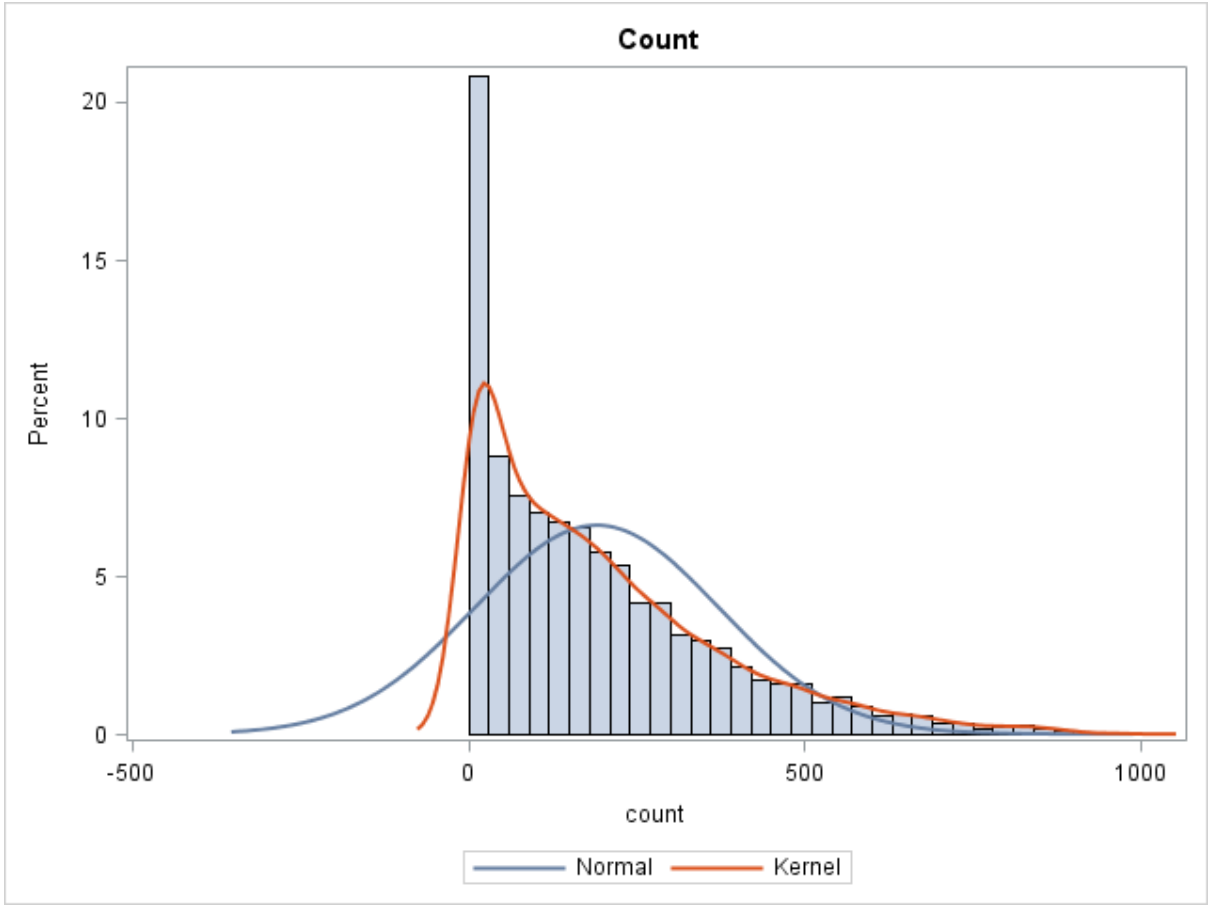

Boxplot for Temperature
0
10
20
30
40
temp


Boxplot for Humidity
20
40
60
80
100
humidity


Boxplot for Windspeed
10
20
30
40
50
windspeed

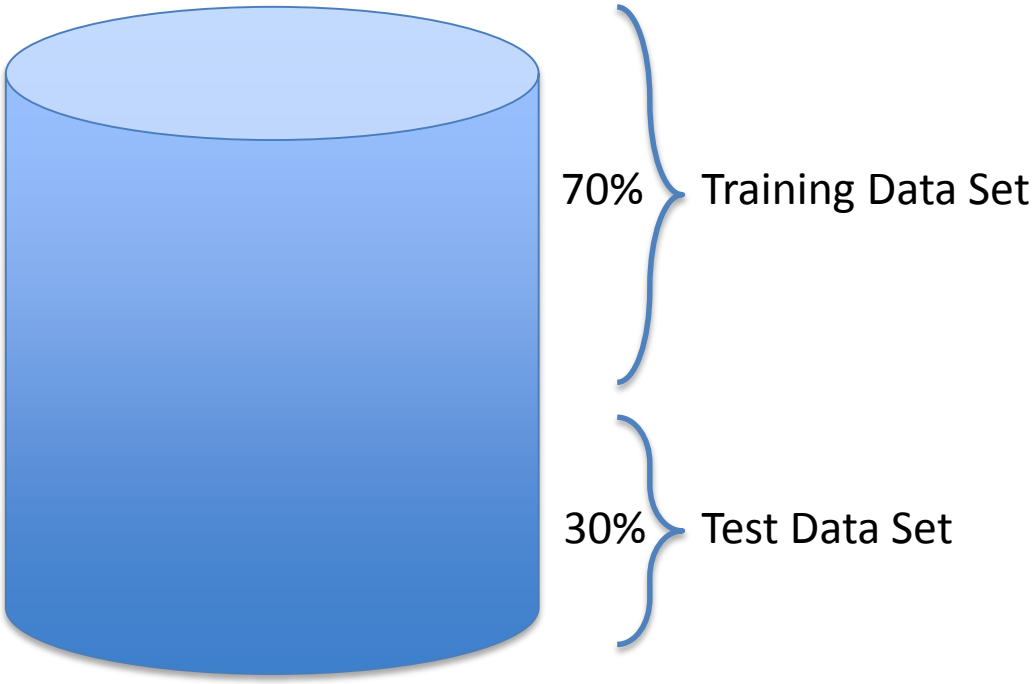
# Histogram

Target variable



# Split Data into Train and Test sets

- Uniform Division of Data



# Regression Analysis

**Proc Reg for the Bike Sharing Dataset**
**Regression Analysis of Count**

The REG Procedure
Model: MODEL1
Dependent Variable: count

| | |
|---|---|
| Number of Observations Read | 7620 |
| Number of Observations Used | 7620 |

**Analysis of Variance**

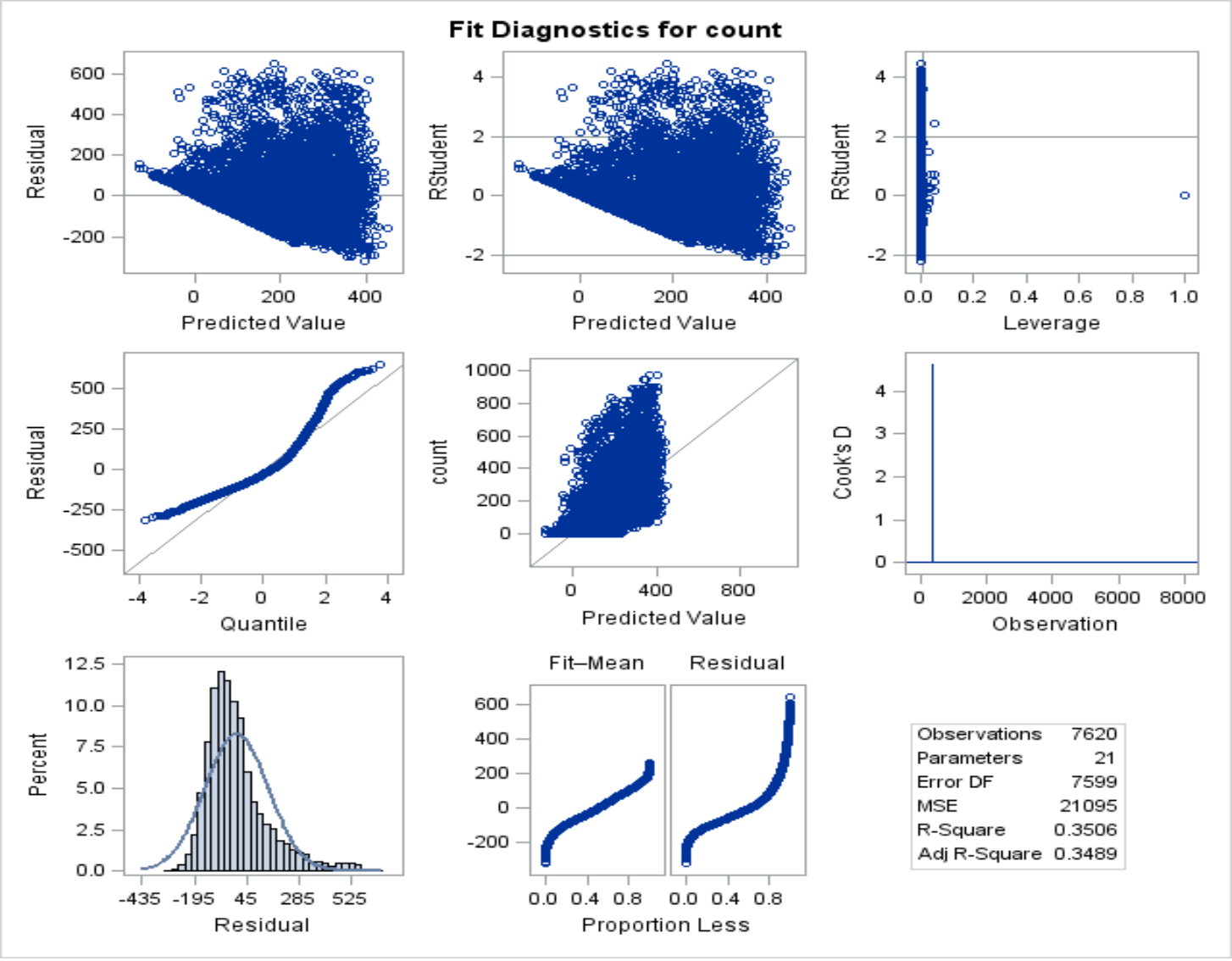
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 20 | 86542448 | 4327122 | 205.13 | <.0001 |
| Error | 7599 | 160298708 | 21095 | | |
| Corrected Total | 7619 | 246841156 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 145.24018 | R-Square | 0.3506 |
| Dependent Mean | 190.22638 | Adj R-Sq | 0.3489 |
| Coeff Var | 76.35123 | | |

# Regression Analysis

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
| Intercept | 1 | 89.29068 | 14.16432 | 6.30 | <.0001 | . | 0 |
| holiday | 1 | -9.30714 | 8.76830 | -1.06 | 0.2885 | 0.00002911 | 1.10313 |
| workingday | 1 | -1.39360 | 3.09354 | -0.45 | 0.6524 | 0.00011215 | 1.07501 |
| temp | 1 | 9.58487 | 1.18944 | 8.06 | <.0001 | 0.15552 | 44.39110 |
| atemp | 1 | 0.94173 | 1.00153 | 0.94 | 0.3471 | 0.00005952 | 37.23369 |
| humidity | 1 | -2.37869 | 0.09568 | -24.86 | <.0001 | 0.10890 | 1.71582 |
| windspeed | 1 | -0.25631 | 0.22882 | -1.12 | 0.2627 | 0.00013873 | 1.20278 |
| hour | 1 | 6.99956 | 0.21806 | 32.10 | <.0001 | 0.10004 | 1.17559 |
| jan | 1 | -0.55305 | 6.95793 | -0.08 | 0.9366 | 0.00507 | 1.86554 |
| mar | 1 | -1.45488 | 7.04243 | -0.21 | 0.8363 | 0.00262 | 1.94657 |
| may | 1 | 21.91765 | 7.14693 | 3.07 | 0.0022 | 0.00127 | 2.02701 |
| jun | 1 | -24.03771 | 7.71694 | -3.11 | 0.0018 | 0.00032674 | 2.36324 |
| jul | 1 | -142.45178 | 9.57493 | -14.88 | <.0001 | 0.01074 | 3.63821 |
| aug | 1 | -120.41423 | 9.37503 | -12.84 | <.0001 | 0.01241 | 3.48789 |
| sep | 1 | -60.20576 | 8.36407 | -7.20 | <.0001 | 0.00217 | 2.76791 |
| oct | 1 | -18.37189 | 7.37741 | -2.49 | 0.0128 | 0.00091063 | 2.15770 |
| nov | 1 | -13.14433 | 6.85609 | -1.92 | 0.0552 | 0.00348 | 1.86353 |
| spring | 1 | -71.05466 | 6.98683 | -10.17 | <.0001 | 0.00381 | 4.68938 |
| summer | 1 | -71.14999 | 7.10709 | -10.01 | <.0001 | 0.00962 | 4.91002 |
| w_clear | 1 | 22.42394 | 5.81316 | 3.86 | 0.0001 | 0.00012797 | 3.91606 |
| w_cloudy | 1 | 31.92387 | 5.85460 | 5.45 | <.0001 | 0.00273 | 3.41224 |

# Regression Analysis



# Regression Analysis

After Removing Correlated variables from Model

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
| Intercept | 1 | 102.62559 | 17.00446 | 6.04 | <.0001 | . | 0 |
| workingday | 1 | 1.33146 | 3.59690 | 0.37 | 0.7113 | 0.00021650 | 1.00723 |
| atemp | 1 | 8.39519 | 0.39929 | 21.03 | <.0001 | 0.15208 | 4.09393 |
| humidity | 1 | -2.47824 | 0.11366 | -21.80 | <.0001 | 0.10853 | 1.69211 |
| windspeed | 1 | -0.01915 | 0.26658 | -0.07 | 0.9427 | 0.00011764 | 1.17334 |
| hour | 1 | 6.95808 | 0.26062 | 26.70 | <.0001 | 0.09563 | 1.16988 |
| jan | 1 | -4.75862 | 8.18342 | -0.58 | 0.5609 | 0.00558 | 1.86142 |
| mar | 1 | 0.23529 | 8.30628 | 0.03 | 0.9774 | 0.00346 | 1.89872 |
| may | 1 | 22.23666 | 8.44879 | 2.63 | 0.0085 | 0.00047021 | 1.94191 |
| jun | 1 | -10.67321 | 8.87178 | -1.20 | 0.2290 | 0.00050444 | 2.20938 |
| jul | 1 | -113.76629 | 10.87394 | -10.46 | <.0001 | 0.01089 | 3.16526 |
| aug | 1 | -82.27952 | 10.42429 | -7.89 | <.0001 | 0.00738 | 2.87008 |
| sep | 1 | -32.08789 | 9.60614 | -3.34 | 0.0008 | 0.00014602 | 2.52494 |
| oct | 1 | -13.01377 | 8.69644 | -1.50 | 0.1346 | 0.00107 | 2.11104 |
| nov | 1 | -7.93810 | 8.21631 | -0.97 | 0.3340 | 0.00400 | 1.85516 |
| spring | 1 | -71.34913 | 8.31473 | -8.58 | <.0001 | 0.00444 | 4.70089 |
| summer | 1 | -62.80577 | 8.42362 | -7.46 | <.0001 | 0.00763 | 4.83478 |
| w_clear | 1 | 13.87764 | 6.83749 | 2.03 | 0.0424 | 0.00046702 | 3.81360 |
| w_cloudy | 1 | 25.95260 | 6.88306 | 3.77 | 0.0002 | 0.00187 | 3.31565 |

# Regression Analysis

After Removing insignificant variables from Model

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
| Intercept | 1 | 98.28631 | 14.45115 | 6.80 | <.0001 | . | 0 |
| holiday | 1 | -1.42336 | 3.59521 | -0.40 | 0.6922 | 0.00021650 | 1.00643 |
| atemp | 1 | 8.28495 | 0.35231 | 23.52 | <.0001 | 0.15208 | 3.18777 |
| humidity | 1 | -2.47780 | 0.10709 | -23.14 | <.0001 | 0.10853 | 1.50215 |
| hour | 1 | 6.97784 | 0.25905 | 26.94 | <.0001 | 0.09573 | 1.15600 |
| feb | 1 | 2.22098 | 7.02547 | 0.32 | 0.7519 | 0.00313 | 1.38181 |
| apr | 1 | 9.63083 | 8.67537 | 1.11 | 0.2670 | 0.00063555 | 2.12770 |
| may | 1 | 32.38398 | 8.33854 | 3.88 | 0.0001 | 0.00057763 | 1.89184 |
| jul | 1 | -105.05120 | 8.54884 | -12.29 | <.0001 | 0.00993 | 1.95666 |
| aug | 1 | -73.79767 | 8.18293 | -9.02 | <.0001 | 0.00501 | 1.76882 |
| sep | 1 | -24.00437 | 7.51083 | -3.20 | 0.0014 | 0.00015459 | 1.54381 |
| spring | 1 | -67.14650 | 5.61243 | -11.96 | <.0001 | 0.01558 | 2.14215 |
| summer | 1 | -65.21743 | 7.65927 | -8.51 | <.0001 | 0.00995 | 3.99777 |
| w_clear | 1 | 14.19566 | 6.74564 | 2.10 | 0.0354 | 0.00047286 | 3.71238 |
| w_cloudy | 1 | 26.53802 | 6.80769 | 3.90 | <.0001 | 0.00199 | 3.24392 |

# Regression Analysis

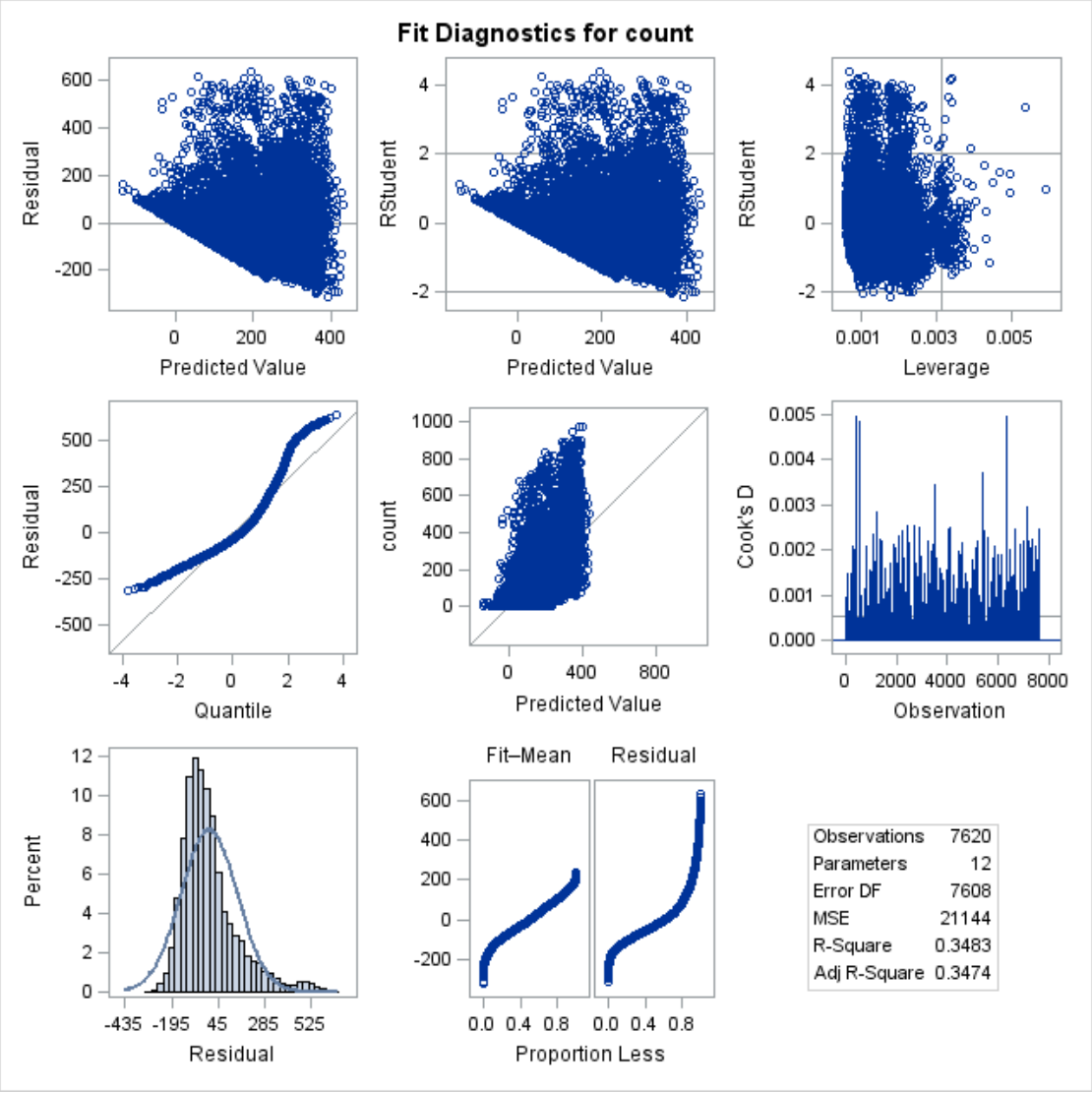
After Removing insignificant variables from Model

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 11 | 85978382 | 7816217 | 369.67 | <.0001 |
| Error | 7608 | 160862775 | 21144 | | |
| Corrected Total | 7619 | 246841156 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 145.40941 | R-Square | 0.3483 |
| Dependent Mean | 190.22638 | Adj R-Sq | 0.3474 |
| Coeff Var | 76.44020 | | |

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
| Intercept | 1 | 101.80448 | 14.01679 | 7.26 | <.0001 | . | 0 |
| atemp | 1 | 8.14375 | 0.32757 | 24.86 | <.0001 | 0.15224 | 2.75629 |
| humidity | 1 | -2.49109 | 0.10651 | -23.39 | <.0001 | 0.10855 | 1.48641 |
| hour | 1 | 6.99277 | 0.25855 | 27.05 | <.0001 | 0.09571 | 1.15170 |
| may | 1 | 27.74096 | 7.14011 | 3.89 | 0.0001 | 0.00084516 | 1.38740 |
| jul | 1 | -102.99497 | 8.33141 | -12.36 | <.0001 | 0.00972 | 1.85876 |
| aug | 1 | -71.92423 | 8.00727 | -8.98 | <.0001 | 0.00457 | 1.69403 |
| sep | 1 | -22.67025 | 7.39769 | -3.06 | 0.0022 | 0.00033578 | 1.49795 |
| spring | 1 | -67.20064 | 5.07966 | -13.23 | <.0001 | 0.01455 | 1.75511 |
| summer | 1 | -59.55966 | 5.67935 | -10.49 | <.0001 | 0.01474 | 2.19850 |
| w_clear | 1 | 13.61148 | 6.72613 | 2.02 | 0.0430 | 0.00052777 | 3.69168 |
| w_cloudy | 1 | 26.32932 | 6.80126 | 3.87 | 0.0001 | 0.00197 | 3.23844 |

# Regression Analysis



# Normalize the data

We have to normalize the data to run Factor Analysis

| season | holiday | workingday | temp | atemp | humidity | windspeed |
|---|---|---|---|---|---|---|
| 1 | 1.4606052252 | -1.460605225 | -1.333599436 | -1.092686778 | 0.997153386 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.438841115 | -1.18236652 | 0.9446391302 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.438841115 | -1.18236652 | 0.9446391302 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.333599436 | -1.092686778 | 0.6820678511 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.333599436 | -1.092686778 | 0.6820678511 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.333599436 | -1.271456264 | 0.6820678511 | -1.254580179 |
| 1 | 1.4606052252 | -1.460605225 | -1.438841115 | -1.18236652 | 0.9446391302 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.544082795 | -1.271456264 | 1.259724665 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -1.333599436 | -1.092686778 | 0.6820678511 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -0.912632718 | -0.735147805 | 0.7345821069 | -0.20536815 |
| 1 | 1.4606052252 | -1.460605225 | -0.59690768 | -0.467288575 | 0.7345821069 | 0.3946122926 |
| 1 | 1.4606052252 | -1.460605225 | -0.70214936 | -0.824827548 | 0.997153386 | 0.6951050099 |
| 1 | 1.4606052252 | -1.460605225 | -0.386424322 | -0.288519089 | 0.7870963627 | 0.6951050099 |
| 1 | 1.4606052252 | -1.460605225 | -0.175940963 | -0.109749603 | 0.5245250837 | 0.8448488724 |
| 1 | 1.4606052252 | -1.460605225 | -0.175940963 | -0.109749603 | 0.5245250837 | 0.6951050099 |

**Fig: Snapshot of Normalized Data**

# Factor Analysis (Training Data)

| Eigenvalues of the Correlation Matrix: Total = 22 Average = 1 | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 3.12945469 | 0.69672025 | 0.1422 | 0.1422 |
| 2 | 2.43273444 | 0.26934303 | 0.1106 | 0.2528 |
| 3 | 2.16339141 | 0.71969912 | 0.0983 | 0.3512 |
| 4 | 1.44369229 | 0.08350094 | 0.0656 | 0.4168 |
| 5 | 1.36019134 | 0.09763300 | 0.0618 | 0.4786 |
| 6 | 1.26255835 | 0.10604784 | 0.0574 | 0.5360 |
| 7 | 1.15651051 | 0.06016326 | 0.0526 | 0.5886 |
| 8 | 1.09634725 | 0.00291690 | 0.0498 | 0.6384 |
| 9 | 1.09343034 | 0.00043141 | 0.0497 | 0.6881 |
| 10 | 1.09299893 | 0.00352474 | 0.0497 | 0.7378 |
| 11 | 1.08947419 | 0.02211132 | 0.0495 | 0.7873 |
| 12 | 1.06736286 | 0.05080528 | 0.0485 | 0.8358 |
| 13 | 1.01655758 | 0.17784471 | 0.0462 | 0.8820 |
| 14 | 0.83871287 | 0.11010573 | 0.0381 | 0.9202 |
| 15 | 0.72860714 | 0.17539436 | 0.0331 | 0.9533 |
| 16 | 0.55321278 | 0.32055804 | 0.0251 | 0.9784 |
| 17 | 0.23265474 | 0.08798232 | 0.0106 | 0.9890 |
| 18 | 0.14467242 | 0.04723652 | 0.0066 | 0.9956 |
| 19 | 0.09743589 | 0.09743589 | 0.0044 | 1.0000 |
| 20 | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| 21 | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| 22 | 0.00000000 | | 0.0000 | 1.0000 |

# Factor Analysis (Training Data)

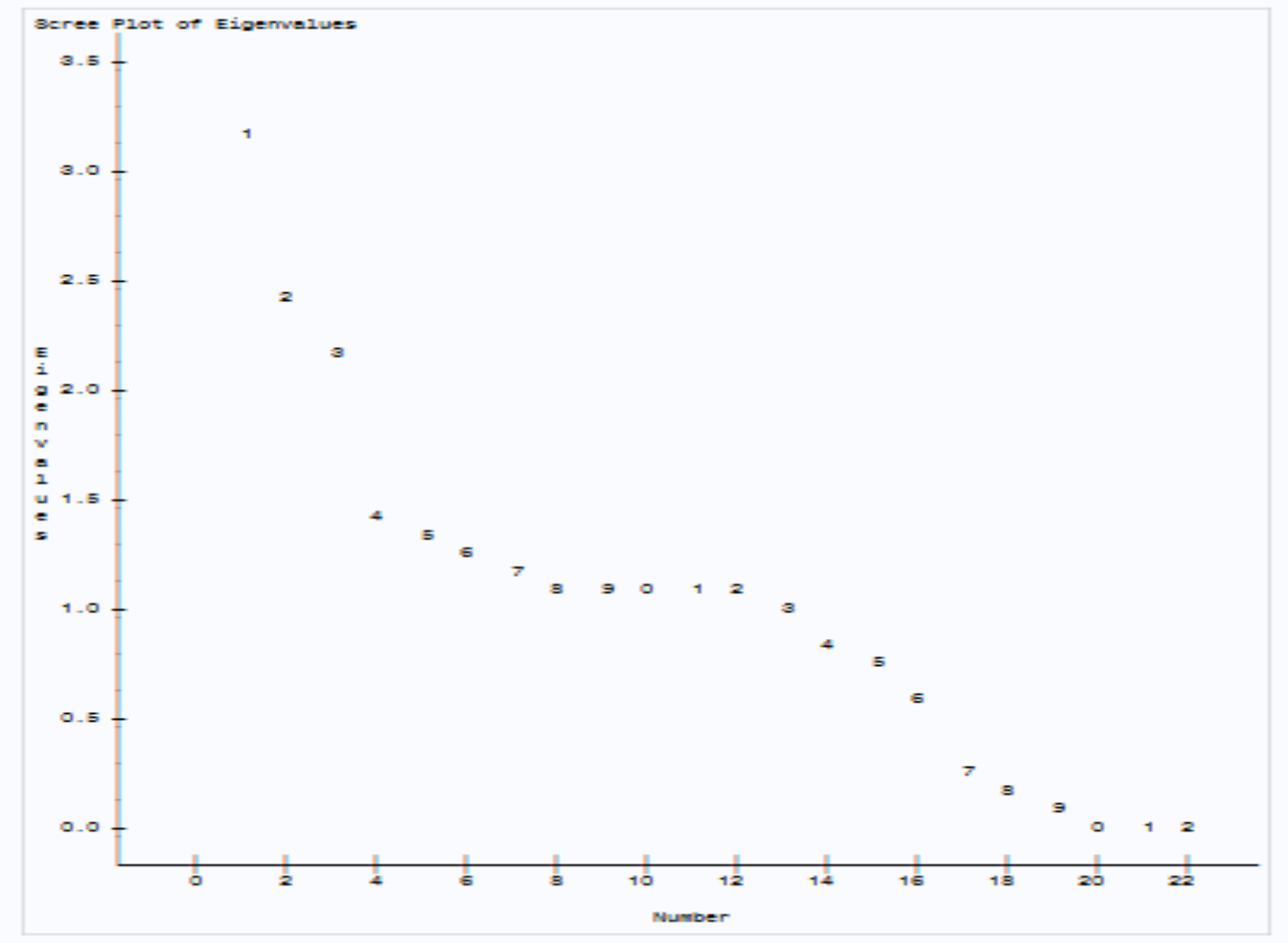



Fig: Scree Plot of Eigen Values

# Regression Analysis on Factors

Training Data

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 11 | 62229709 | 5657246 | 239.52 | <.0001 |
| Error | 7607 | 179668894 | 23619 | | |
| Corrected Total | 7618 | 241898603 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 153.68438 | R-Square | 0.2573 |
| Dependent Mean | 189.81612 | Adj R-Sq | 0.2562 |
| Coeff Var | 80.96488 | | |

Parameter Estimates

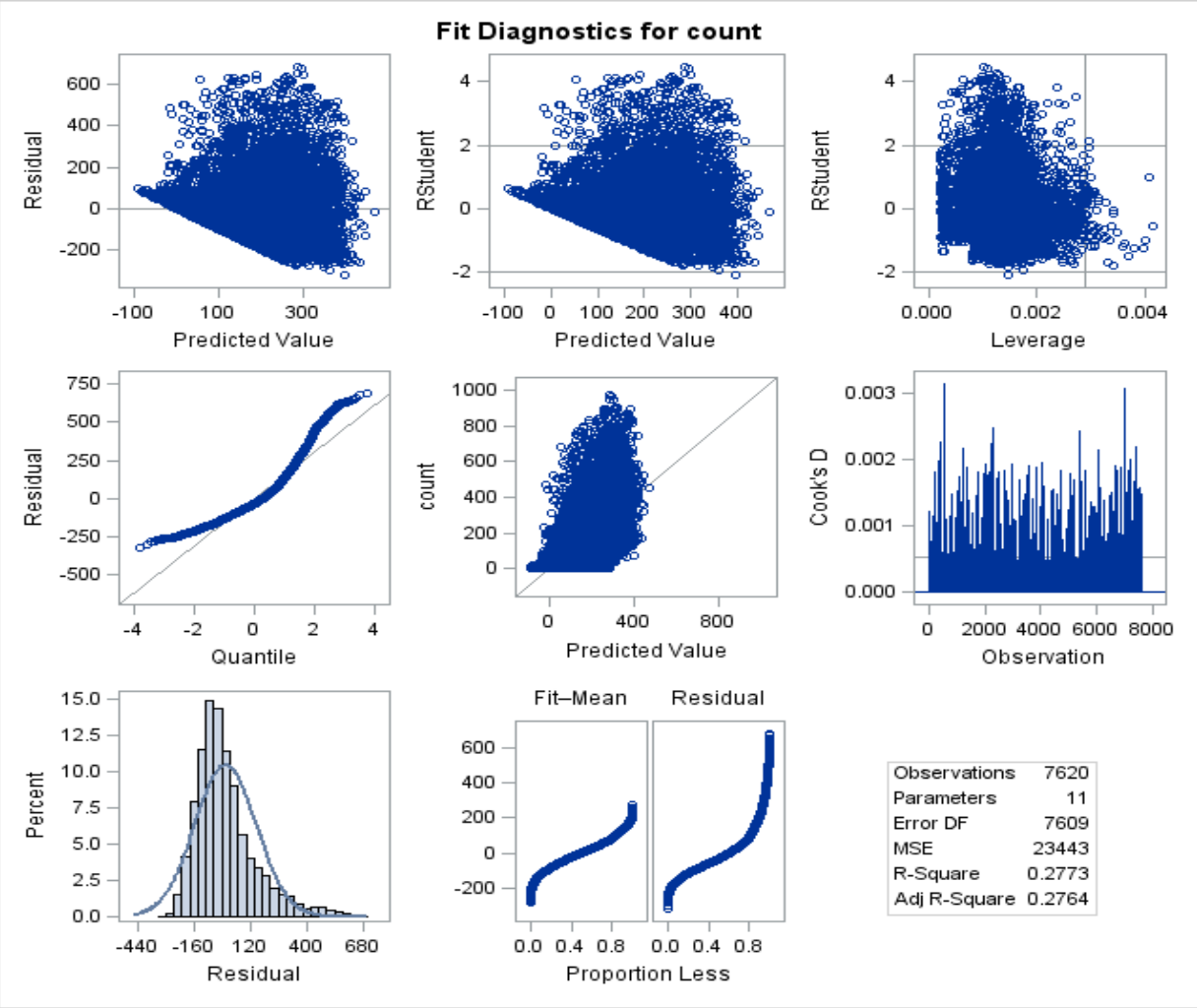
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 189.81612 | 1.76068 | 107.81 | <.0001 | . | 0 |
| Factor1 | 1 | 7.10177 | 1.76080 | 4.03 | <.0001 | 0.00159 | 1.00000 |
| Factor2 | 1 | 15.40627 | 1.76080 | 8.75 | <.0001 | 0.00749 | 1.00000 |
| Factor3 | 1 | 35.35674 | 1.76080 | 20.08 | <.0001 | 0.03973 | 1.00000 |
| Factor4 | 1 | 5.37971 | 1.76080 | 3.06 | 0.0023 | 0.00095782 | 1.00000 |
| Factor5 | 1 | -32.52299 | 1.76080 | -18.47 | <.0001 | 0.03504 | 1.00000 |
| Factor6 | 1 | 27.17290 | 1.76080 | 15.43 | <.0001 | 0.02535 | 1.00000 |
| Factor7 | 1 | 64.56936 | 1.76080 | 36.67 | <.0001 | 0.14685 | 1.00000 |
| Factor8 | 1 | -17.95256 | 1.76080 | -10.20 | <.0001 | 0.01331 | 1.00000 |
| Factor9 | 1 | -11.65645 | 1.76080 | -6.62 | <.0001 | 0.00569 | 1.00000 |
| Factor10 | 1 | 7.87744 | 1.76080 | 4.47 | <.0001 | 0.00261 | 1.00000 |
| Factor11 | 1 | 10.78928 | 1.76080 | 6.13 | <.0001 | 0.00491 | 1.00000 |

Proc Reg for the Bike Sharing dataset
Regression Analysis of Count on Factors

The REG Procedure
Model: MODEL1
Dependent Variable: count

| | |
|---|---|
| Durbin-Watson D | 2.030 |
| Pr < DW | 0.9063 |
| Pr > DW | 0.0937 |
| Number of Observations | 7620 |
| 1st Order Autocorrelation | -0.015 |

# Regression Analysis on Factors



# Factor Analysis (Entire Data)

| Eigenvalues of the Correlation Matrix: Total = 22 Average = 1 | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 3.13248851 | 0.69896844 | 0.1424 | 0.1424 |
| 2 | 2.43352007 | 0.27521151 | 0.1106 | 0.2530 |
| 3 | 2.15830856 | 0.69686840 | 0.0981 | 0.3511 |
| 4 | 1.46144016 | 0.10923635 | 0.0664 | 0.4175 |
| 5 | 1.35220381 | 0.09005031 | 0.0615 | 0.4790 |
| 6 | 1.26215350 | 0.10211619 | 0.0574 | 0.5364 |
| 7 | 1.16003730 | 0.06016144 | 0.0527 | 0.5891 |
| 8 | 1.09987587 | 0.00711175 | 0.0500 | 0.6391 |
| 9 | 1.09276411 | 0.00154745 | 0.0497 | 0.6888 |
| 10 | 1.09121667 | 0.00220812 | 0.0496 | 0.7384 |
| 11 | 1.08900855 | 0.01846359 | 0.0495 | 0.7879 |
| 12 | 1.07054496 | 0.06411643 | 0.0487 | 0.8365 |
| 13 | 1.00642853 | 0.17279117 | 0.0457 | 0.8823 |
| 14 | 0.83363736 | 0.10225061 | 0.0379 | 0.9202 |
| 15 | 0.73138675 | 0.17919604 | 0.0332 | 0.9534 |
| 16 | 0.55219071 | 0.31783426 | 0.0251 | 0.9785 |
| 17 | 0.23435645 | 0.09249909 | 0.0107 | 0.9892 |
| 18 | 0.14185737 | 0.04527659 | 0.0064 | 0.9956 |
| 19 | 0.09658078 | 0.09658078 | 0.0044 | 1.0000 |
| 20 | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| 21 | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| 22 | 0.00000000 | | 0.0000 | 1.0000 |

# Factor Analysis (Entire Data)

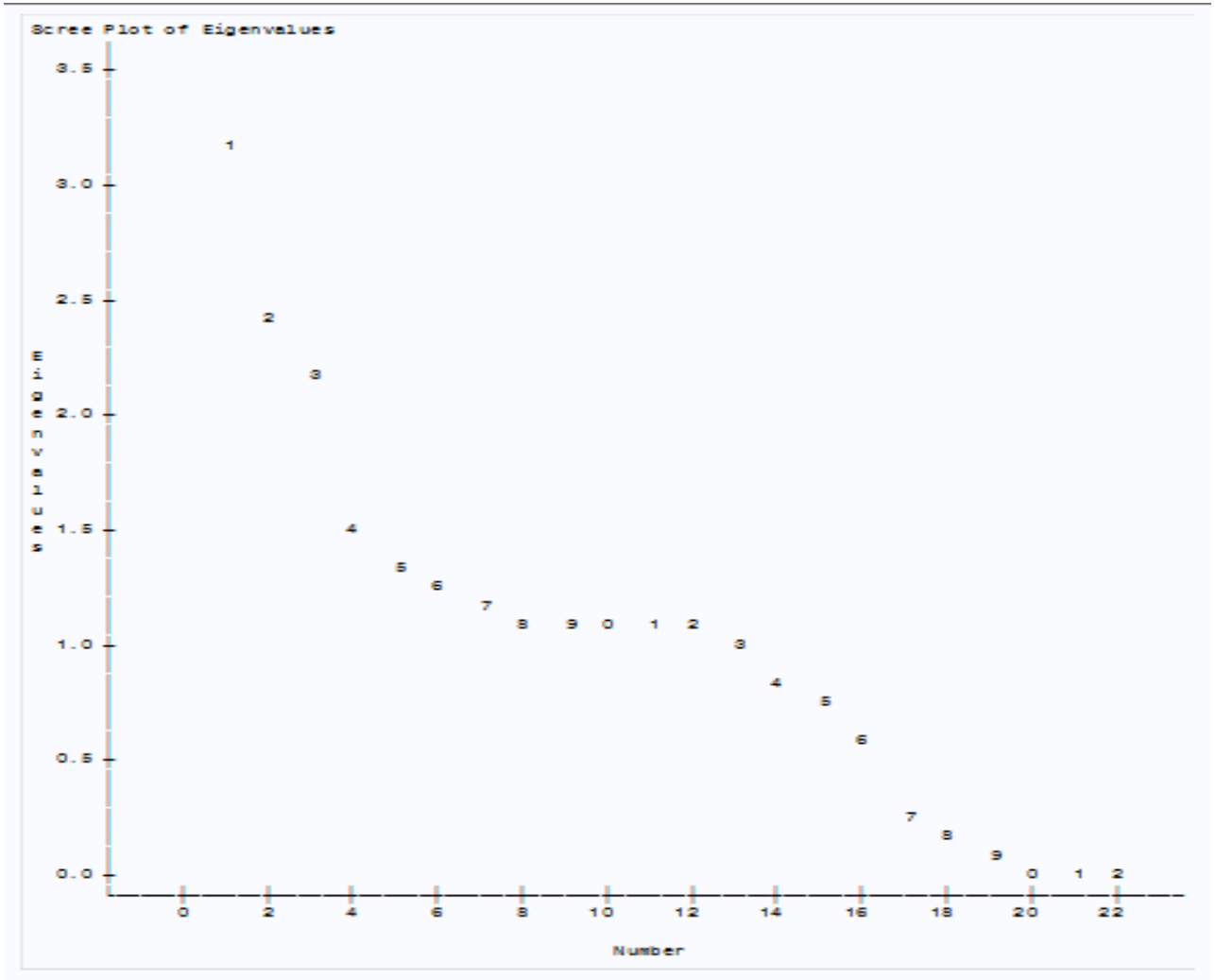



Fig: Scree Plot of Eigen Values

# Regression Analysis on Factors

Entire Data

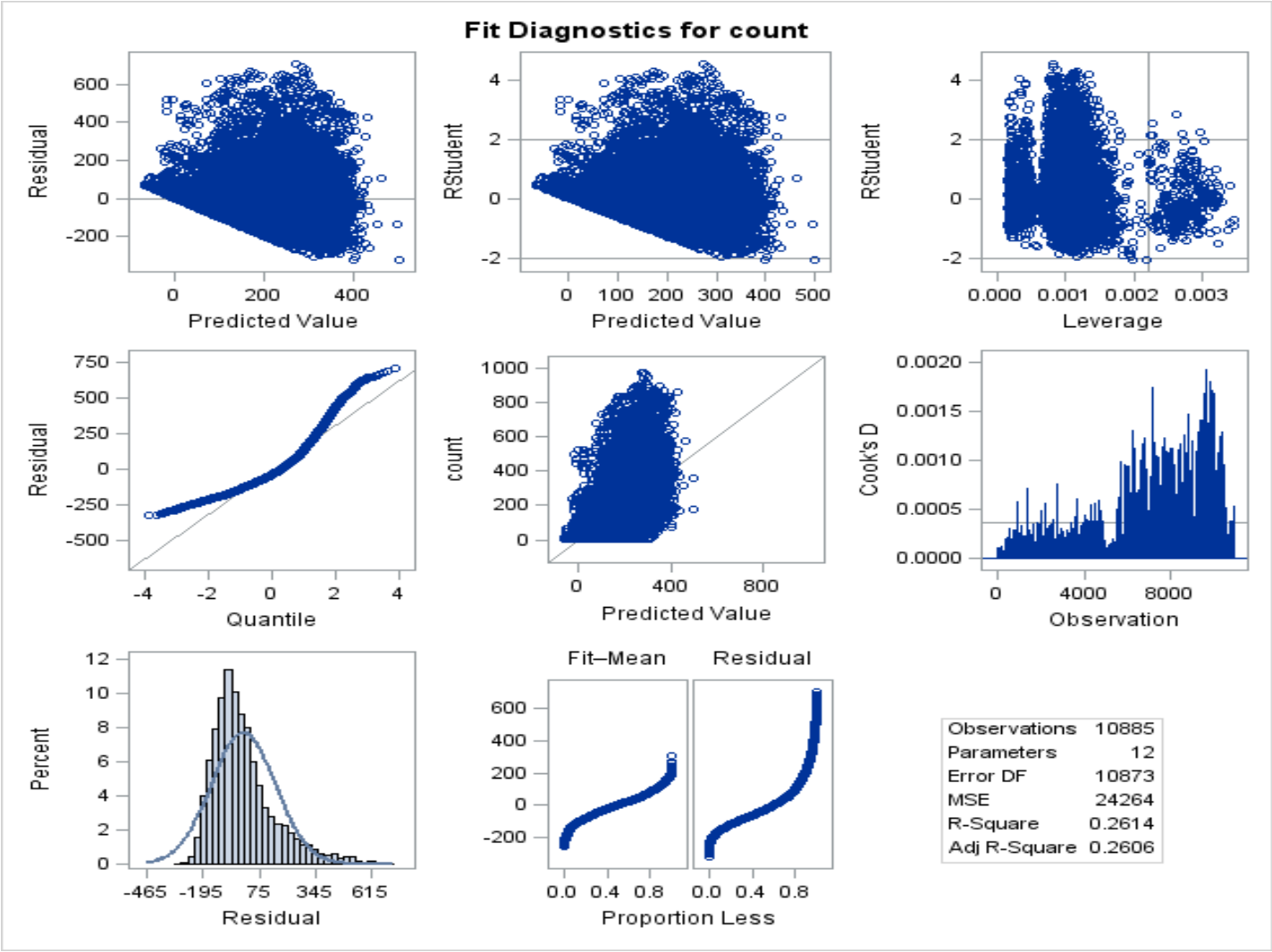
| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 11 | 93348733 | 8486248 | 349.75 | <.0001 |
| Error | 10873 | 263823420 | 24264 | | |
| Corrected Total | 10884 | 357172153 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 155.76934 | R-Square | 0.2614 |
| Dependent Mean | 191.57667 | Adj R-Sq | 0.2606 |
| Coeff Var | 81.30914 | | |

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
| Intercept | 1 | 191.57667 | 1.49303 | 128.31 | <.0001 | . | 0 |
| Factor1 | 1 | -2.97986 | 1.49310 | -2.00 | 0.0460 | 0.00027058 | 1.00000 |
| Factor2 | 1 | 8.06524 | 1.49310 | 5.40 | <.0001 | 0.00198 | 1.00000 |
| Factor3 | 1 | -38.60121 | 1.49310 | -25.85 | <.0001 | 0.04551 | 1.00000 |
| Factor4 | 1 | -30.79747 | 1.49310 | -20.63 | <.0001 | 0.03035 | 1.00000 |
| Factor5 | 1 | 28.12509 | 1.49310 | 18.84 | <.0001 | 0.02610 | 1.00000 |
| Factor6 | 1 | 66.53052 | 1.49310 | 44.56 | <.0001 | 0.14998 | 1.00000 |
| Factor7 | 1 | -12.79846 | 1.49310 | -8.57 | <.0001 | 0.00853 | 1.00000 |
| Factor8 | 1 | 23.78266 | 1.49310 | 15.93 | <.0001 | 0.02269 | 1.00000 |
| Factor9 | 1 | -5.80789 | 1.49310 | -3.89 | 0.0001 | 0.00138 | 1.00000 |
| Factor10 | 1 | 0.63784 | 1.49310 | 0.43 | 0.6692 | 0.00001673 | 1.00000 |
| Factor11 | 1 | 9.12927 | 1.49310 | 6.11 | <.0001 | 0.00343 | 1.00000 |

# Regression Analysis on Factors

Entire Data



# Regression Analysis Test Data

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 9 | 32205251 | 3578361 | 140.39 | <.0001 |
| Error | 3256 | 82989593 | 25488 | | |
| Corrected Total | 3265 | 115194844 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 159.65026 | R-Square | 0.2796 |
| Dependent Mean | 195.68371 | Adj R-Sq | 0.2776 |
| Coeff Var | 81.58587 | | |

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Variance Inflation |
| Intercept | 1 | 195.68371 | 2.79358 | 70.05 | <.0001 | . | 0 |
| Factor2 | 1 | -34.94446 | 2.79401 | -12.51 | <.0001 | 0.03461 | 1.00000 |
| Factor3 | 1 | 34.13157 | 2.79401 | 12.22 | <.0001 | 0.03420 | 1.00000 |
| Factor4 | 1 | 11.30718 | 2.79401 | 4.05 | <.0001 | 0.00389 | 1.00000 |
| Factor5 | 1 | 74.51407 | 2.79401 | 26.67 | <.0001 | 0.16945 | 1.00000 |
| Factor6 | 1 | 17.40562 | 2.79401 | 6.23 | <.0001 | 0.01113 | 1.00000 |
| Factor7 | 1 | -12.28072 | 2.79401 | -4.40 | <.0001 | 0.00560 | 1.00000 |
| Factor8 | 1 | 18.98569 | 2.79401 | 6.80 | <.0001 | 0.01347 | 1.00000 |
| Factor9 | 1 | 5.86105 | 2.79401 | 2.10 | 0.0360 | 0.00130 | 1.00000 |
| Factor11 | 1 | 30.80464 | 2.79401 | 11.03 | <.0001 | 0.03599 | 1.00000 |

**Proc Reg for the Test dataset**
**Regression Analysis of Count**

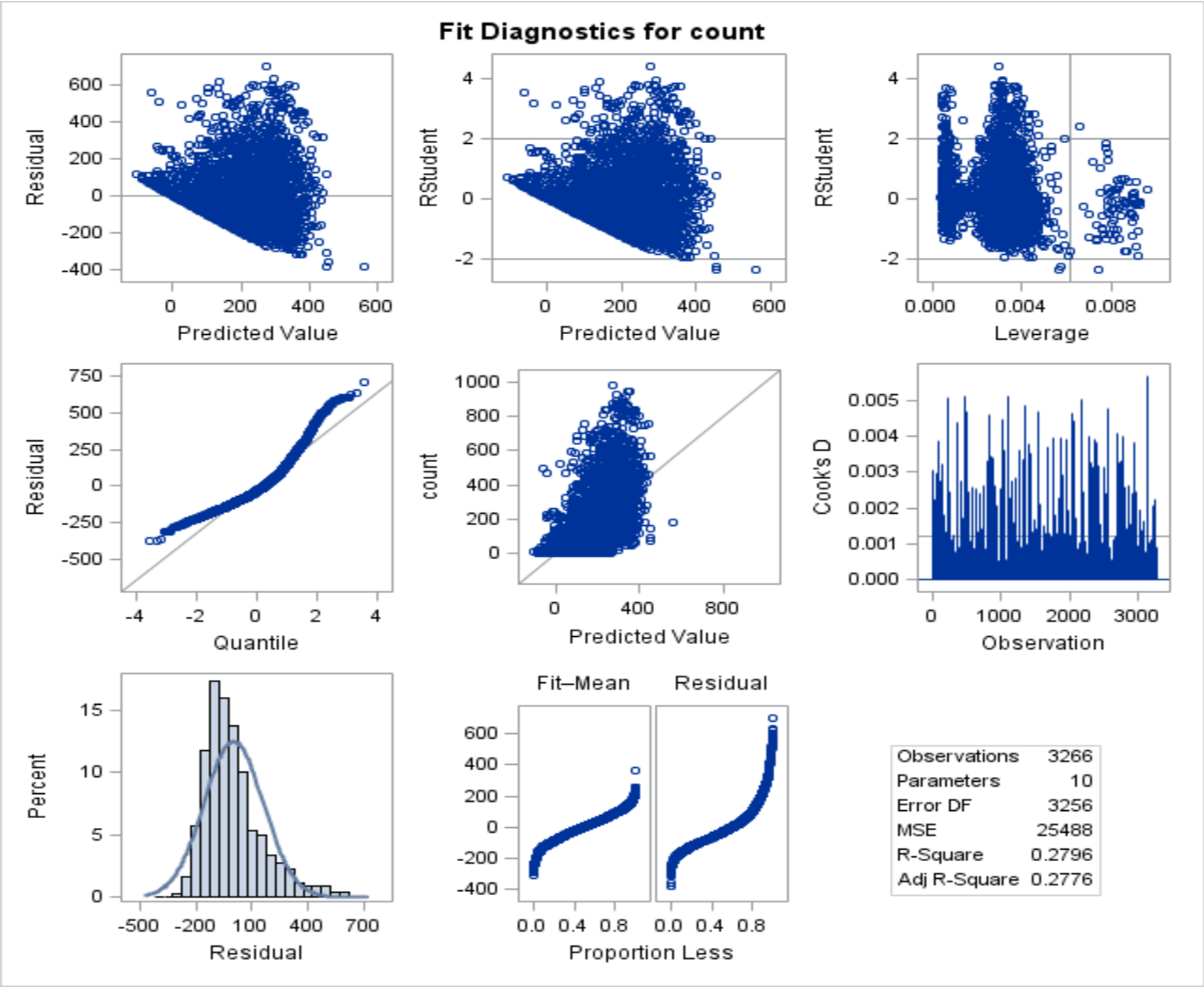
The REG Procedure
Model: MODEL1
Dependent Variable: count

| | |
|---|---|
| Durbin-Watson D | 1.990 |
| Pr < DW | 0.3854 |
| Pr > DW | 0.6146 |
| Number of Observations | 3266 |
| 1st Order Autocorrelation | 0.005 |

The MEANS Procedure

| Analysis Variable : C_spredicted Standard Error of Mean Predicted Value | | | | |
|---|---|---|---|---|
| N | Mean | Std Dev | Minimum | Maximum |
| 3266 | 8.5560655 | 2.1991513 | 3.1875200 | 15.6548195 |

# Regression Analysis on Test Data



# Conclusion

- Data preparation and exploration takes up to 60% of the time
- Count of bike rentals is highest in Clear weather
- Bike rentals are high during office commutation hours on weekdays and afternoons on holiday
- Residual should be normally distributed in linear regression. But its not in our case.