Executive Summary_v.1 Creation Date: Sunday, 14 Sep, 2025 Author: edward.lim.2025

# ISSS602 Data Analytics Lab Assignment 1: Show Me the Numbers

Creation Date: Sunday, 20 Sep, 2025

Lim Boon Yan Edward
Student ID: 01548108
Email: Edward.lim.2025@mitb.smu.edu.sg

Word Count: 2814

# Content

## Context

Bike-sharing, a service offering bicycles or electric bikes for short-term city use, is accessed through a mobile app and a fee.

Users rent these bikes for commuting, with options for docked systems at specific stations or dockless systems that allow for more flexible pickup and drop-off locations.

The company has seen a significant decrease in revenue because of increased competition.

## The Task

Assume the role of a data analyst of the market research consulting company, to conduct a study to discover and determine factors affecting the demand of HappyRides' bike sharing business.

## Data Preparation

In this section, the following tasks were performed:

**Step 1: Data Sanity Checks**

A Data check have been conducted using summary to identify:

- No missing values were found in the dataset (Fig 1.1).
- No extreme values were detected (Fig 1.1). The minimum and maximum values were within a reasonable range, and the standard deviation was consistent with the data.
- No duplicate entries were found, as confirmed by a SAS procedure check (Fig 1.2).

Variable Summary

| Obs | Variable name | Width of the variable formatted value | Type of the raw values | Recommended level for analytics | Have more unreported levels | Number of levels | Number of missing values | Minimum numeric value | Maximum numeric value | Mean | Standard deviation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | datetime | 16 | C | ID | Y | 20 | 0 | . | . | . | . |
| 2 | season | 12 | N | CLASS | N | 4 | 0 | 1 | 4 | 2.5066139996 | 1.1161743093 |
| 3 | holiday | 12 | N | CLASS | N | 2 | 0 | 0 | 1 | 0.028568804 | 0.1665988506 |
| 4 | weather | 12 | N | CLASS | N | 4 | 0 | 1 | 4 | 1.4184273379 | 0.6338385858 |
| 5 | temp | 12 | N | INTERVAL | Y | 20 | 0 | 0.82 | 41 | 20.23085982 | 7.791589844 |
| 6 | atemp | 12 | N | INTERVAL | Y | 20 | 0 | 0.76 | 45.455 | 23.655084053 | 8.4746006265 |
| 7 | humidity | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 100 | 61.886459673 | 19.245033277 |
| 8 | windspeed | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 56.9969 | 12.799395407 | 8.1645373268 |
| 9 | casual | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 367 | 36.021954804 | 49.960476573 |
| 10 | registered | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 886 | 155.55217711 | 151.03903308 |
| 11 | count | 12 | N | INTERVAL | Y | 20 | 0 | 1 | 977 | 191.57413191 | 181.14445383 |

Fig 1.1 Summary Statistic

```
proc sort data=casuser.BIKE_SHARING_DATASET
          out=no_duplicate (label='without duplicates')
          dupout=duplicate_only (label='Duplicate only')
          nodupkey;
    by datetime season holiday weather temp atemp humidity windspeed casual registe
run;
```

Fig 1.2

**Step 2: Extraction of Date and Time information from 'datetime'**

In this step we break down the datetime column to extract the following details:

| Column Assigned | Data Types | Reference Column |
|---|---|---|
| year_str | Year | datetime |
| month_str | Month | |
| day_str | Day | |
| day_of_week_name | Day of Week | |
| time_part | Time | |
| date_part | Date | |

This is done using SAS Program with code below:

```
data casuser.BIKE_SHARING_DATASET_step2;

    set casuser.BIKE_SHARING_DATASET_step1;

    length date_part time_part $ 20;

    date_part = scan(datetime, 1, ' ');
    time_part = scan(datetime, 2, ' ');

    day_str = scan(date_part, 1, '/');
    month_str = scan(date_part, 2, '/');
    year_str = scan(date_part, 3, '/');

    month = input(month_str, best.);

    length day_of_week_name $ 5;
    day_of_week_name = put(mdy( input(month_str, best.),  input(day_str, best.), input(year_str, best.)), downame3.);
```

Fig 1.3

**Step 3: Mapping of Data to enhance clarity**

To improve clarity, the data were mapped for the following reasons:

| Data Types | Reference Column | Mapping/Binning |
|---|---|---|
| season_name | season | 1: Spring, 2: Summer,<br>3: Fall, 4: Winter |
| month_name | month | 1: Jan, 2:Feb, 3:Mar..... |
| weather_name | weather | 1: 1_Clear, Few clouds, partly cloudy |
| | | 2: 2_Mist + Cloudy, Mist + Broken |

| | | |
|---|---|---|
| | | clouds, Mist + Few clouds, Mist |
| | | 3: 3_Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds |
| | | 4: 4_Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| holiday_name | holiday | 1: Holiday, 2: Non-Holiday |
| weekend_or_day | day_of_week_name | 1-5: Weekday, 6-7: Weekend |

This was performed using the SAS program with the following code:

```
    length day_of_week_name $ 5;
    day_of_week_name = put(mdy( input(month_str, best.),  input(day_str, best.), input(year_str, best.)), downame3.);

    length weekend_or_day $ 20;

    if day_of_week_name in ('Mon', 'Tue', 'Wed', 'Thu', 'Fri') then
        weekend_or_day = 'weekday';

    else
        weekend_or_day = 'weekend';

    length season_name $ 8;

    if season = '1' then season_name = 'Spring';
    else if season = '2' then season_name = 'Summer';
    else if season = '3' then season_name = 'Fall';
    else if season = '4' then season_name = 'Winter';

    length month_name $ 8;

    if month='1' then month_name='Jan';
    else if month='2' then month_name='Feb';
    else if month='3' then month_name='Mar';
    else if month='4' then month_name='Apr';
    else if month='5' then month_name='May';
    else if month='6' then month_name='Jun';
    else if month='7' then month_name='Jul';
    else if month='8' then month_name='Aug';
    else if month='9' then month_name='Sep';
    else if month='10' then month_name='Oct';
    else if month='11' then month_name='Nov';
    else if month='12' then month_name='Dec';

/* Propose naming for weather*/

    length weather_name $ 100;

    if weather = '1' then weather_name = '1_Clear, Few clouds, partly cloudy';
    else if weather = '2' then weather_name = '2_Mist + Cloudy, Mist + Broken
clouds, Mist + Few clouds, Mist';
    else if weather = '3' then weather_name = '3_Light Snow, Light Rain + Thunderstorm + Scattered clouds,
Light Rain + Scattered clouds';
    else if weather = '4' then weather_name = '4_Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog';

/* Propose naming for Holiday*/

    length holiday_name $ 15;

    if holiday = '1' then holiday_name = 'Holiday';
    else if holiday = '0' then holiday_name = 'Non-Holiday';
```

Fig 1.4

**Step 4: Binning of continuous data variables into range categories**

The data are binned for the following:

| Binned Column | Reference Column | Mapping/Binning |
|---|---|---|
| temp_binned | temp | 0-9.9 °C, 10-19.9 °C, 20-29.9°C, 30-39.9°C, Above 40 °C |
| atemp_binned | atemp | 0-9.9 °C, 10-19.9 °C, 20-29.9°C, 30-39.9°C, Above 40 °C |
| humidity_binned | humidity | 0-19.9 %, 20-39.9 %, 40-59.9 %, 60-79.9 %, Above 80 % |
| windspeed_binned | windspeed | 0-9.9 kmph, 10-19.9 kmph, 20-29.9 kmph, 30-39.9 kmph, 40-49.90 kmph, Above 50 kmph |

This was performed using the SAS program with the following code:

```
/* Propose naming for Holiday*/

    length holiday_name $ 15;

    if holiday = '1' then holiday_name = 'Holiday';
    else if holiday = '0' then holiday_name = 'Non-Holiday';

/* Binning for Temp*/

    length temp_binned $ 15;

    if temp < '10' then temp_binned = '0-9.9 °C';
    else if temp < '20' then temp_binned = '10-19.9 °C';
    else if temp < '30' then temp_binned = '20-29.9°C';
    else if temp < '40' then temp_binned = '30-39.9°C';
    else if temp >= '40' then temp_binned = 'Above 40 °C';

/* Binning for atemp*/

    length atemp_binned $ 15;

    if atemp < '10' then atemp_binned = '0-9.9 °C';
    else if atemp < '20' then atemp_binned = '10-19.9 °C';
    else if atemp < '30' then atemp_binned = '20-29.9°C';
    else if atemp < '40' then atemp_binned = '30-39.9°C';
    else if atemp >= '40' then atemp_binned = 'Above 40 °C';

/* Binning for humidity*/

    length humidity_binned $ 25;

    if humidity < '20' then humidity_binned = '0-19.9 %';
    else if humidity < '40' then humidity_binned = '20-39.9 %';
    else if humidity < '60' then humidity_binned = '40-59.9 %';
    else if humidity < '80' then humidity_binned = '60-79.9 %';
    else if humidity >= '80' then humidity_binned = 'Above 80 %';

/* Binning for windspeed*/

    length windspeed_binned $ 25;

    if windspeed < '10' then windspeed_binned = '0-9.9 kmph';
    else if windspeed < '20' then windspeed_binned = '10-19.9 kmph';
    else if windspeed < '30' then windspeed_binned = '20-29.9 kmph';
    else if windspeed < '40' then windspeed_binned = '30-39.9 kmph';
    else if windspeed < '50' then windspeed_binned = '40-49.90 kmph';
    else if windspeed >= '50' then windspeed_binned = 'Above 50 kmph';

run;
```

Fig 1.5

**Step 5: Change Data Type**

The data types are changed for the following to CHAR and length set to optimize on size for the following columns:

| Weather_str |
|---|
| day_str |
| month_str |
| year_str |
| weather_str |

This was performed using the SAS program with the following code:

```
117 data casuser.TRANSACTION_DATA_Cleaned;
118
119     set casuser.BIKE_SHARING_DATASET_step2;
120
121     length weather_str $2;
122
123     day_str = put(day_str, 5.);
124
125     month_str = put(month_str, 5.);
126
127     year_str = put(year_str, 4.);
128
129     weather_str = put(weather, 2.);
130
131 run;
132
```

Fig 1.6

**Step 6: Saving as Cas Data and Final Check**

The output was saved as a Cas Data object and loaded into memory for visualization

This was performed using the SAS program with the following code:

```
133 proc casutil;
134     save casdata= 'TRANSACTION_DATA_Cleaned'
135     casout='TRANSACTION_DATA_Cleaned.sashdat'
136     outcaslib='CASUSER'
137     replace;
138
139 quit;
```

Fig 1.7

Using the Summary report, a final check is performed to ensure there are no obvious errors.

Variable Summary

| Obs | Variable name | Width of the variable formatted value | Type of the raw values | Recommended level for analytics | Have more unreported levels | Number of levels | Number of missing values | Minimum numeric value | Maximum numeric value | Mean | Standard deviation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | datetime | 16 | C | ID | Y | 20 | 0 | . | . | . | . |
| 2 | season | 12 | N | CLASS | N | 4 | 0 | 1 | 4 | 2.5066139996 | 1.1161743093 |
| 3 | holiday | 12 | N | CLASS | N | 2 | 0 | 0 | 1 | 0.028568804 | 0.1665988506 |
| 4 | weather | 12 | N | CLASS | N | 4 | 0 | 1 | 4 | 1.4184273379 | 0.6338385858 |
| 5 | temp | 12 | N | INTERVAL | Y | 20 | 0 | 0.82 | 41 | 20.23085982 | 7.791589844 |
| 6 | atemp | 12 | N | INTERVAL | Y | 20 | 0 | 0.76 | 45.455 | 23.655084053 | 8.4746006265 |
| 7 | humidity | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 100 | 61.886459673 | 19.245033277 |
| 8 | windspeed | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 56.9969 | 12.799395407 | 8.1645373268 |
| 9 | casual | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 367 | 36.021954804 | 49.960476573 |
| 10 | registered | 12 | N | INTERVAL | Y | 20 | 0 | 0 | 886 | 155.55217711 | 151.03903308 |
| 11 | count | 12 | N | INTERVAL | Y | 20 | 0 | 1 | 977 | 191.57413191 | 181.14445383 |
| 12 | date_part | 20 | C | ID | Y | 20 | 0 | . | . | . | . |
| 13 | time_part | 20 | C | ID | Y | 20 | 0 | . | . | . | . |
| 14 | day_str | 20 | C | CLASS | N | 19 | 0 | . | . | . | . |
| 15 | month_str | 20 | C | CLASS | N | 12 | 0 | . | . | . | . |
| 16 | year_str | 20 | C | CLASS | N | 2 | 0 | . | . | . | . |
| 17 | month | 12 | N | CLASS | N | 12 | 0 | 1 | 12 | 6.5214954988 | 3.4443734958 |
| 18 | day_of_week_name | 5 | C | CLASS | N | 7 | 0 | . | . | . | . |
| 19 | weekend_or_day | 20 | C | CLASS | N | 2 | 0 | . | . | . | . |
| 20 | season_name | 8 | C | CLASS | N | 4 | 0 | . | . | . | . |
| 21 | month_name | 8 | C | CLASS | N | 12 | 0 | . | . | . | . |
| 22 | weather_name | 100 | C | CLASS | N | 4 | 0 | . | . | . | . |
| 23 | holiday_name | 15 | C | CLASS | N | 2 | 0 | . | . | . | . |
| 24 | temp_binned | 15 | C | CLASS | N | 5 | 0 | . | . | . | . |
| 25 | atemp_binned | 15 | C | CLASS | N | 5 | 0 | . | . | . | . |
| 26 | humidity_binned | 25 | C | CLASS | N | 5 | 0 | . | . | . | . |
| 27 | windspeed_binned | 25 | C | CLASS | N | 6 | 0 | . | . | . | . |
| 28 | weather_str | 2 | C | CLASS | N | 4 | 0 | . | . | . | . |

Fig 1.8

## Analysis and Insights

- All data analysis was performed in **SAS Viya**.
- Hypothesis tests were conducted at a 95% confidence level (α=0.05). The null hypothesis (H0) was rejected for all p-values less than **0.05**.
- All slicing of data are based on the aggregation of 2023 and 2024 data.
- **Welch's ANOVA** was used to test for differences in group means.
- **Least Squares Means** analysis was used to identify the specific contributors to any significant differences found in the group means.

**Terminology:**

For the following analysis, the variables are defined as below:

- **Total Bike Users**: This is represented by the count variable and is equivalent to the sum of casual and registered users.
- **Registered Users**: This is represented by the registered variable.
- **Casual Users**: This is represented by the casual variable.

**Sales and Demand situation:**

Demand for HappyRides increased in 2024 compared to 2023, which is a positive trend.

- The growth is stronger when we look at the registered users.
- The revenue dip at HappyRides is influenced by a decrease in the revenue per rental unit.

For this study, we will investigate the factors that might affect demand.

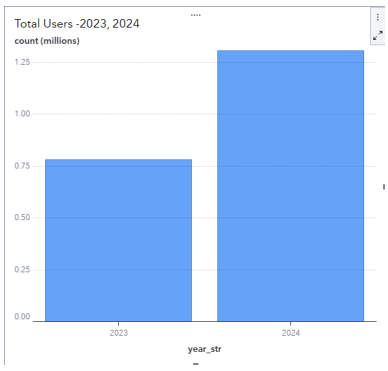



Fig 1.9

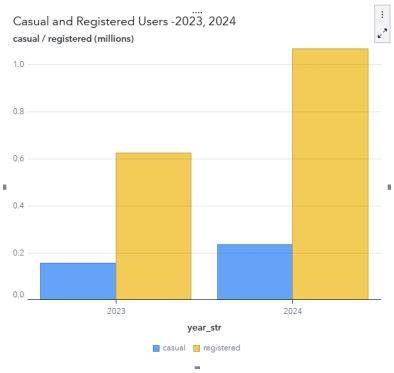



Fig 1.10

A Correlation matrix is set up to understand the relationship between variables:

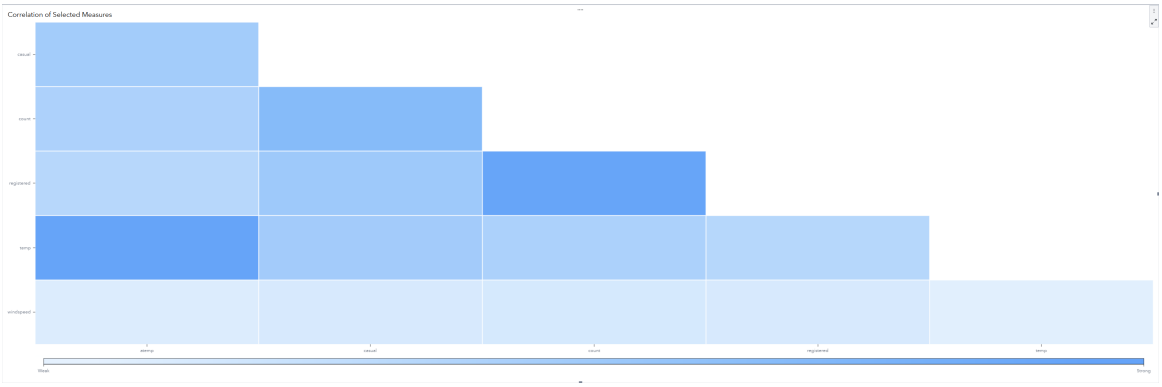



Fig 1.11 **Variables**: casual, count, registered, temp, atemp, windspeed.

| X Axis | Y Axis | Correlation ▼ |
|---|---|---|
| atemp | temp | 0.9849 |
| count | registered | 0.9709 |
| casual | count | 0.6904 |
| casual | registered | 0.4972 |
| casual | temp | 0.4671 |

Fig 1.12

- There is a strong correlation ( 0.98) between temperature (temp) and apparent temperature (atemp).
  This is expected, so only temp will be used for our analysis
- There is a strong correlation between **registered** and **count** (~0.97) as compared to **casual** and **count** (~0.69).

### Insight 1: Temperature has a significant effect on bike usage

-As the temperature increased from 0 to approximately 40 degrees Celsius, the mean number of bike users consistently rose.

-It was observed that the total number of users increased within the temperature range of 0 to 39.9 degrees Celsius.

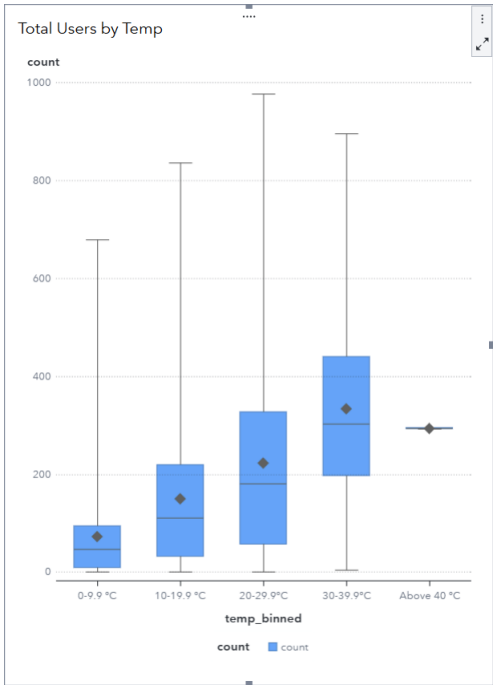



Fig 2.1 boxplot of count against temperature

**Welch's Anova:**

An ANOVA test was conducted:

(Dependent variable: Total Bike User (count) ; Categorical variable: temperature)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Given a **p-value < 0.05**, we **reject the null hypothesis (H0)** and conclude that temperature has a statistically significant effect on the total number of bike users (count).

**Levene's Test for Homogeneity of count Variance**
**ANOVA of Squared Deviations from Group Means**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| temp_binned | 3 | 1.146E12 | 3.821E11 | 144.71 | <.0001 |
| Error | 10881 | 2.873E13 | 2.6403E9 | | |

**Welch's ANOVA for count**

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| temp_binned | 3.0000 | 910.92 | <.0001 |
| Error | 3842.1 | | |

Fig 2.2 ANOVA result (count vs temp_binned)

### Insights 2: Humidity has a significant effect on bike usage

-The total bike users are noted to be **affected by temperature changes**.

-The mean total bike users  (count) consistently rises as humidity increases from 0% to 40%.

-The mean then drops when humidity levels climb above 60%.

-There is a high level of skewness for casual users within the 0-19.9% humidity range, making the median a more reliable reference point in this instance (Fig 2.3 Bottom right)

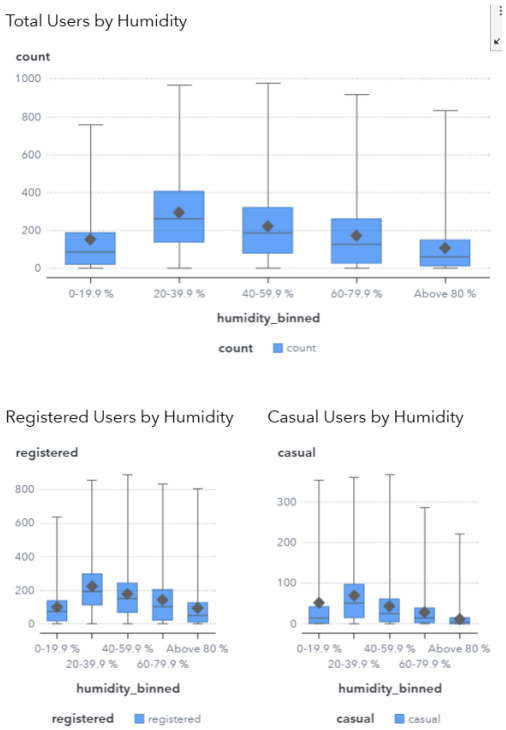



Fig 2.3 boxplot of count, casual and registered against humidity

**Welch's Anova:** Same result for count, casual and registered against humidity

An ANOVA test was conducted:

(Dependent variable: Total Users (count) ; Categorical variable: humidity)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Given a **p-value < 0.05**, we **reject the null hypothesis (H0)** and conclude that humidity has a statistically significant effect on bike usage.

This effect is observed across total users (Fig 2.4), casual users (Fig 2.5), and registered users (Fig 2.6).

| Levene's Test for Homogeneity of count Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| humidity_binned | 4 | 5.412E11 | 1.353E11 | 45.07 | <.0001 |
| Error | 10881 | 3.267E13 | 3.0024E9 | | |

| Welch's ANOVA for count | | | |
|---|---|---|---|
| Source | DF | F Value | Pr > F |
| humidity_binned | 4.0000 | 343.78 | <.0001 |
| Error | 504.9 | | |

Fig 2.4 ANOVA result (count vs humidity)

| Levene's Test for Homogeneity of casual Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| humidity_binned | 4 | 2.279E10 | 5.6973E9 | 135.96 | <.0001 |
| Error | 10881 | 4.559E11 | 41903146 | | |

| Welch's ANOVA for casual | | | |
|---|---|---|---|
| Source | DF | F Value | Pr > F |
| humidity_binned | 4.0000 | 476.93 | <.0001 |
| Error | 499.5 | | |

Fig 2.5 ANOVA result (casual vs humidity)

| Levene's Test for Homogeneity of registered Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| humidity_binned | 4 | 1.37E11 | 3.424E10 | 15.27 | <.0001 |
| Error | 10881 | 2.439E13 | 2.2419E9 | | |

| Welch's ANOVA for registered | | | |
|---|---|---|---|
| Source | DF | F Value | Pr > F |
| humidity_binned | 4.0000 | 237.16 | <.0001 |
| Error | 509.8 | | |

Fig 2.6 ANOVA result (registered vs humidity_binned)

### Insight 3: Weather Type has an effect on Bike Usage

-Weather Type 1: Clear, Few clouds, partly cloudy has the highest mean number of users compared to other weather types.

-This trend should be viewed with caution, as the data for Weather Type 4 is limited (1 data count) - Fig 2.7.

-Further data collection is needed to confirm the relationship between weather and bike usage for Weather Type 4.

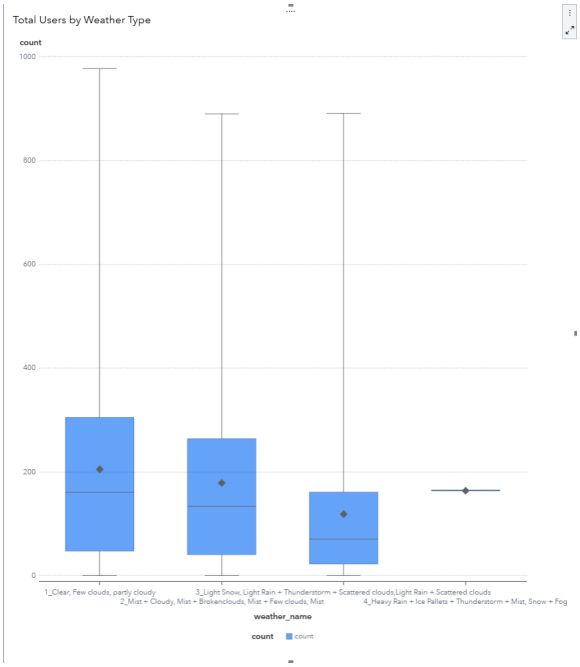



Fig 2.7 boxplot of Total Users (count) against weather type

**Welch's Anova:**

An ANOVA test was conducted:

(Dependent variable: total users (count) ; Categorical variable: weather type)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Given a **p-value < 0.05**, we reject the null hypothesis (H0) and conclude that there is a statistically significant difference in total bike users (count) across different weather types (Fig 2.7).

**True for weather type 1,2,3**

| Levene's Test for Homogeneity of count Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| weather_name | 2 | 2.58E11 | 1.29E11 | 37.68 | <.0001 |
| Error | 10882 | 3.726E13 | 3.4243E9 | | |

| Welch's ANOVA for count | | | |
|---|---|---|---|
| Source | DF | F Value | Pr > F |
| weather_name | 2.0000 | 140.90 | <.0001 |
| Error | 2449.2 | | |

Fig 2.7 ANOVA result (count vs weather type)

**Least Squares Means:**

The analysis showed that the difference between **Weather Type 4** and the **other weather types (1, 2, and 3)** was not statistically significant, as indicated by a **p-value > 0.05** (Fig 2.8).

This result is likely due to a **lack of data**, as there was only one observation recorded for Weather Type 4 (Fig 2.9), which is insufficient to provide meaningful insights.

**Least Squares Means for effect weather_name**
**Pr > |t| for H0: LSMean(i)=LSMean(j)**

**Dependent Variable: count**

| i/j | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | | <.0001 | <.0001 | 0.9957 |
| 2 | <.0001 | | <.0001 | 0.9998 |
| 3 | <.0001 | <.0001 | | 0.9944 |
| 4 | 0.9957 | 0.9998 | 0.9944 | |

Fig 2.8 Least Squares Means (count vs weather type)

| Level of weather_name | N | count Mean | count Std Dev |
|---|---|---|---|
| 1_Clear, Few clouds, partly cloudy | 7192 | 205.236791 | 187.959566 |
| 2_Mist + Cloudy, Mist + Brokenclouds, Mist + Few clouds, Mist | 2834 | 178.955540 | 168.366413 |
| 3_Light Snow, Light Rain + Thunderstorm + Scattered clouds,Light Rain + Scattered clouds | 859 | 118.846333 | 138.581297 |
| 4_Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog | 1 | 164.000000 | . |

Fig 2.9

### Insight 4: Season has a significant effect on bike usage

-The mean is observed to vary across the different seasons.

-Spring is observed to have the lowest mean users as is lower than winter.

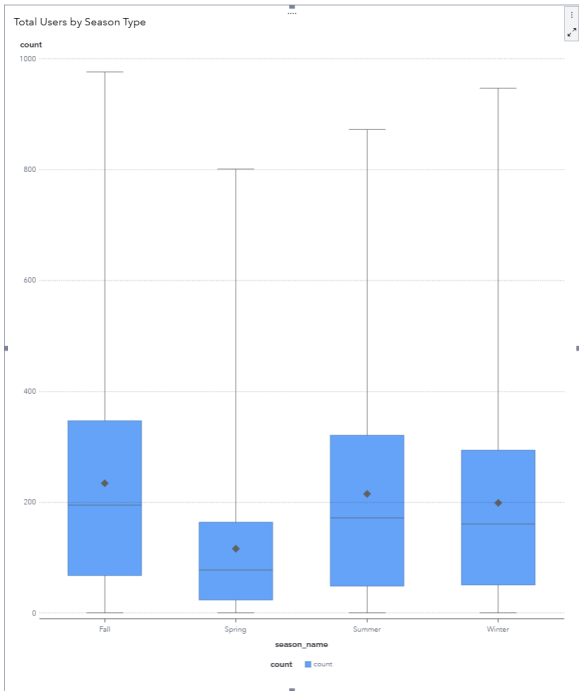



Fig 2.10 boxplot of count against season_name

**Welch's Anova:**

An ANOVA test was conducted:

(Dependent variable: count; Categorical variable: season_name)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Given a **p-value < 0.05**, we **reject the null hypothesis (H0)** and conclude that there is a statistically significant difference in the number of total bike users (count) across different seasons (Fig 2.11).

Levene's Test for Homogeneity of count Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| weather_name | 2 | 2.58E11 | 1.29E11 | 37.68 | <.0001 |
| Error | 10882 | 3.726E13 | 3.4243E9 | | |

Welch's ANOVA for count

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| weather_name | 2.0000 | 140.90 | <.0001 |
| Error | 2449.2 | | |

Fig 2.11 ANOVA result (count vs season)

### Insight 5: Weekday/Weekend has no significant effect on overall bike usage

No statistically significant difference between bike user and Weekday/Weekday.

The median and mean are slightly higher during weekend.

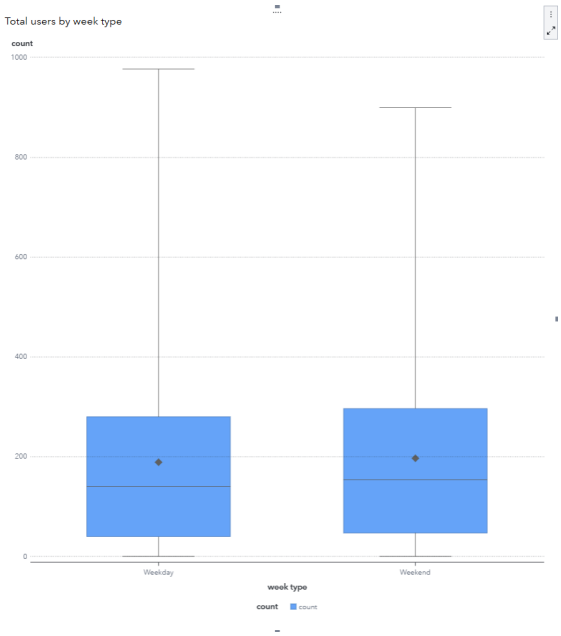



Fig 2.12 boxplot of count against week type

**Welch's Anova:**

(Dependent variable: count; Categorical variable: weekend/weekday)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Since **P-value <0.05**, we **reject H0** and conclude the presence of a statistically significant difference between count (Fig 2.11) and weekend_or_day.

| Levene's Test for Homogeneity of count Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| weekend_or_day | 1 | 1.039E10 | 1.039E10 | 2.92 | 0.0877 |
| Error | 10884 | 3.88E13 | 3.5646E9 | | |

| Welch's ANOVA for count | | | |
|---|---|---|---|
| Source | DF | F Value | Pr > F |
| weekend_or_day | 1.0000 | 4.40 | 0.0360 |
| Error | 5913.0 | | |

Fig 2.13 ANOVA result (count vs weekend or weekday)

### Insight 6: Weekday/Weekend have more effect on Casual Users than Registered Users

- When users are separated into registered and casual categories, a **greater mean difference** is observed in the usage patterns of **casual users** compared to registered users.

- There are **more casual users on weekends** than on weekdays, while the number of registered users remains relatively consistent across the week.

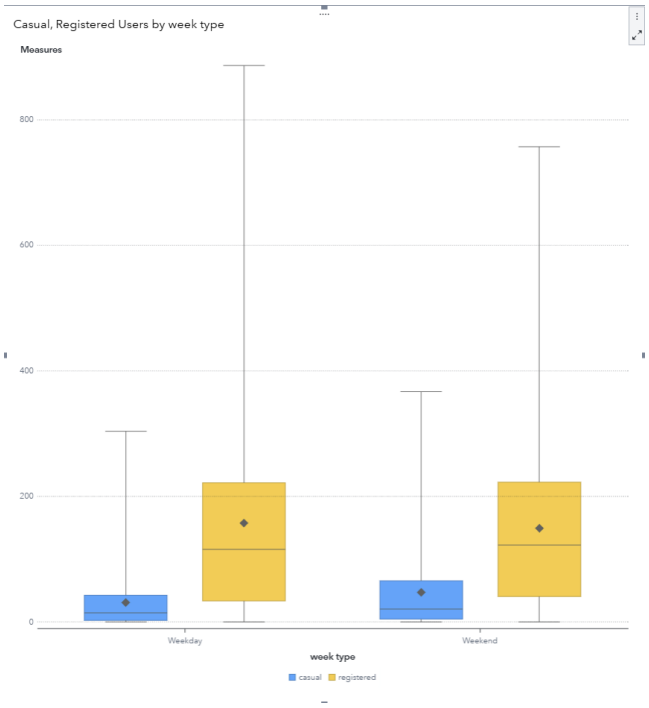
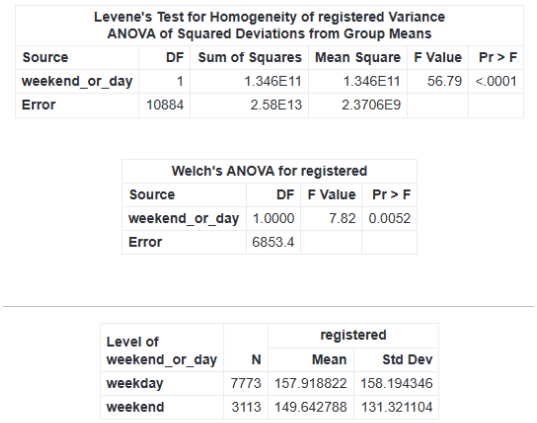



Fig 2.14 boxplot of casual/registered against week type

After performing an ANOVA, the p-value for the comparison of weekday/weekend type against casual (0.0052) and registered (<0.001) users indicated a statistical difference.

| Variables | Group | P-Value | Reference |
|---|---|---|---|
| count (casual+registered) | Weekday vs weekend | 0.0360 | Fig 2.13 |
| casual | Weekday vs weekend | 0.0052 | Fig 2.15 |
| registered | Weekday vs weekend | <0.001 | Fig 2.16 |



Levene's Test for Homogeneity of registered Variance
ANOVA of Squared Deviations from Group Means

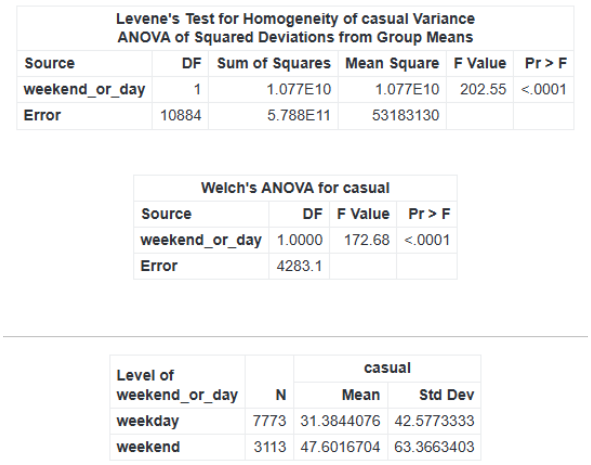
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| weekend_or_day | 1 | 1.346E11 | 1.346E11 | 56.79 | <.0001 |
| Error | 10884 | 2.58E13 | 2.3706E9 | | |

Welch's ANOVA for registered

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| weekend_or_day | 1.0000 | 7.82 | 0.0052 |
| Error | 6853.4 | | |

| Level of weekend_or_day | N | registered | |
|---|---|---|---|
| | | Mean | Std Dev |
| weekday | 7773 | 157.918822 | 158.194346 |
| weekend | 3113 | 149.642788 | 131.321104 |

Fig 2.15 ANOVA result (registered vs weekend or weekday)



Levene's Test for Homogeneity of casual Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| weekend_or_day | 1 | 1.077E10 | 1.077E10 | 202.55 | <.0001 |
| Error | 10884 | 5.788E11 | 53183130 | | |

Welch's ANOVA for casual

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| weekend_or_day | 1.0000 | 172.68 | <.0001 |
| Error | 4283.1 | | |

| Level of weekend_or_day | N | casual | |
|---|---|---|---|
| | | Mean | Std Dev |
| weekday | 7773 | 31.3844076 | 42.5773333 |
| weekend | 3113 | 47.6016704 | 63.3663403 |

Fig 2.16 ANOVA result (casual vs weekend_or_day)

**Insight 7: The Day of the Week has no significant effect on overall bike usage but have significant effect on casual and registered users**

**This section contains 3 sub insights.**

-The mean number of total users (count) shows no significant difference across the days of the week.-This is however not the case when we segregate the view by casual and registered against day of week.

-However, a different pattern emerges when analysing casual and registered users separately by the day of the week:

- **Casual**: casual users are having higher mean on mon and sun.
- **Registered**: registered users are having lower mean on mon and sun.

-The contrasting changes between casual and registered users likely offset one another, which explains why the statistical test for total users did not show a significant difference in the means between days of the week.

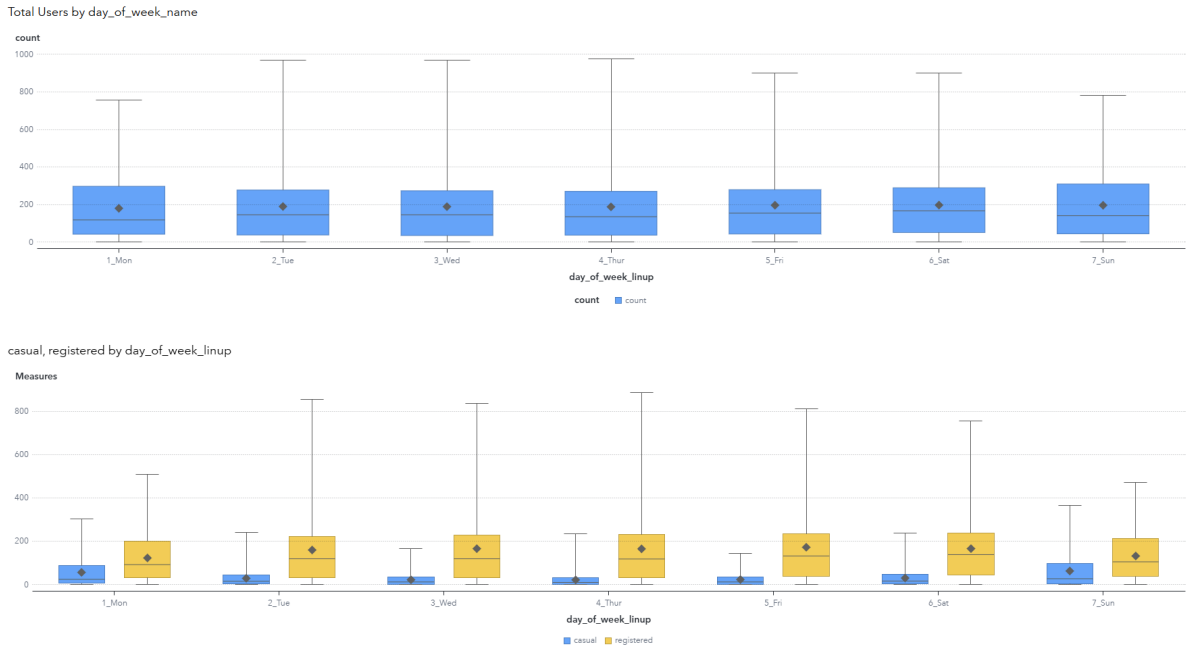



Fig 2.17 boxplot of count/casual,registered against day_of_week

**Summary of ANOVA TEST:**

| Variable | Group | ANOVA P-value | Result | Reference |
|---|---|---|---|---|
| count | day of week | 0.0653 | **Do not reject H0:** There is no statistically significant difference in the number of total bike users (count) across the days of the week | Fig 2.18a |
| casual | day of week | <0.0001 | **Reject Ho:** | Fig 2.18b |
| registered | day of week | <0.0001 | **Reject Ho:** | Fig 2.18c |

**Welch's Anova-Count:**

(Dependent variable: count; Categorical variable: day of week)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

With a **p-value > 0.05**, we **do not reject the null hypothesis (H0)** and conclude that there is no statistically significant difference in the number of total bike users (count) across the days of the week (Fig 2.18a).

Levene's Test for Homogeneity of count Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| day_of_week_name | 6 | 8.821E10 | 1.47E10 | 4.16 | 0.0004 |
| Error | 10879 | 3.847E13 | 3.5366E9 | | |

Welch's ANOVA for count

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| day_of_week_name | 6.0000 | 1.98 | 0.0653 |
| Error | 4832.5 | | |

Fig 2.18a ANOVA result (count vs day of week)

**Welch's Anova-count:**

(Dependent variable: count; Categorical variable: day of week)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

With a **p-value >0.05**, we **do not reject the null hypothesis (H0)** and conclude that there is no statistically significant difference in the number of total bike users (count) across the days of the week

Levene's Test for Homogeneity of count Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| day_of_week_name | 6 | 8.821E10 | 1.47E10 | 4.16 | 0.0004 |
| Error | 10879 | 3.847E13 | 3.5366E9 | | |

Welch's ANOVA for count

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| day_of_week_name | 6.0000 | 1.98 | 0.0653 |
| Error | 4832.5 | | |

(Fig 2.18a).

Fig 2.18a ANOVA result (count vs day of week)

**Welch's Anova-registered:**

(Dependent variable: registered; Categorical variable: day of week)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

With a **p-value <0.05**, we **reject the null hypothesis (H0)** and conclude that there is statistically significant difference in the number of registered users across the days of the week (Fig 2.18b).

Levene's Test for Homogeneity of registered Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| day_of_week_name | 6 | 5.985E11 | 9.975E10 | 46.22 | <.0001 |
| Error | 10879 | 2.348E13 | 2.1581E9 | | |

Welch's ANOVA for registered

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| day_of_week_name | 6.0000 | 36.20 | <.0001 |
| Error | 4799.9 | | |

Fig 2.18b ANOVA & mean result (registered vs day of week)

**Welch's Anova-casual:**

(Dependent variable: casual; Categorical variable: day of week)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

With a **p-value <0.05**, we **reject the null hypothesis (H0)** and conclude that there is statistically significant difference in the number of casual users across the days of the week (Fig 2.18c).

Levene's Test for Homogeneity of casual Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| day_of_week_name | 6 | 4.714E10 | 7.8571E9 | 230.57 | <.0001 |
| Error | 10879 | 3.707E11 | 34076696 | | |

Welch's ANOVA for casual

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| day_of_week_name | 6.0000 | 128.87 | <.0001 |
| Error | 4799.9 | | |

Fig 2.18c ANOVA & mean result (casual vs day_of_week)

### Insight 8: Holiday/Non-Holiday status has no significant effect on overall bike usage

-The bike usage mean is rather similar between holiday and non - holiday period.

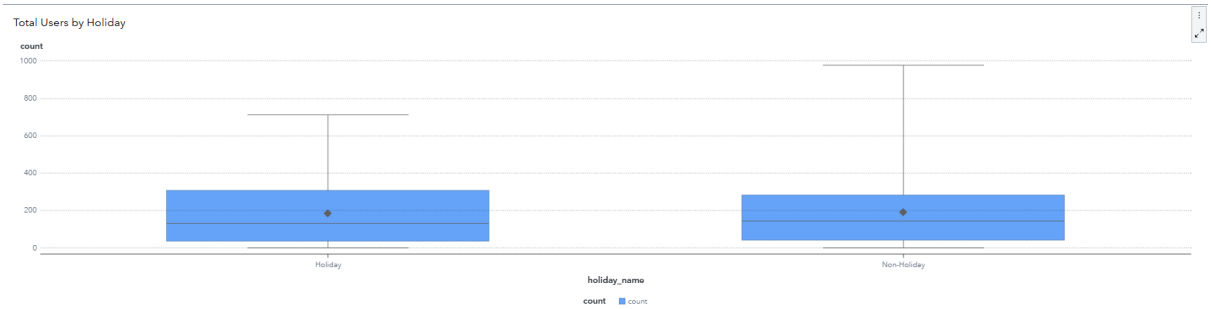



Fig 2.19 boxplot of count against holiday type

**Welch's Anova-registered:**

(Dependent variable: count; Categorical variable: holiday_name)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Since **P-value >0.05**, we **do not reject H0** and conclude that there are no statistically significant difference between count (Fig 2.20) and holiday_name.

Dependent Variable: count

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 10388.1 | 10388.1 | 0.32 | 0.5737 |
| Error | 10884 | 357162525.6 | 32815.4 | | |
| Corrected Total | 10885 | 357172913.7 | | | |

| R-Square | Coeff Var | Root MSE | count Mean |
|---|---|---|---|
| 0.000029 | 94.55877 | 181.1501 | 191.5741 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| holiday_name | 1 | 10388.11759 | 10388.11759 | 0.32 | 0.5737 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| holiday_name | 1 | 10388.11759 | 10388.11759 | 0.32 | 0.5737 |

Fig 2.20 ANOVA result (count vs holiday type)

### Insight 9: Month of the Year has effect on Bike Usage

-The mean number of bike users varies significantly throughout the year.

-Bike usage peaks during the second and third quarters (Q2 and Q3).

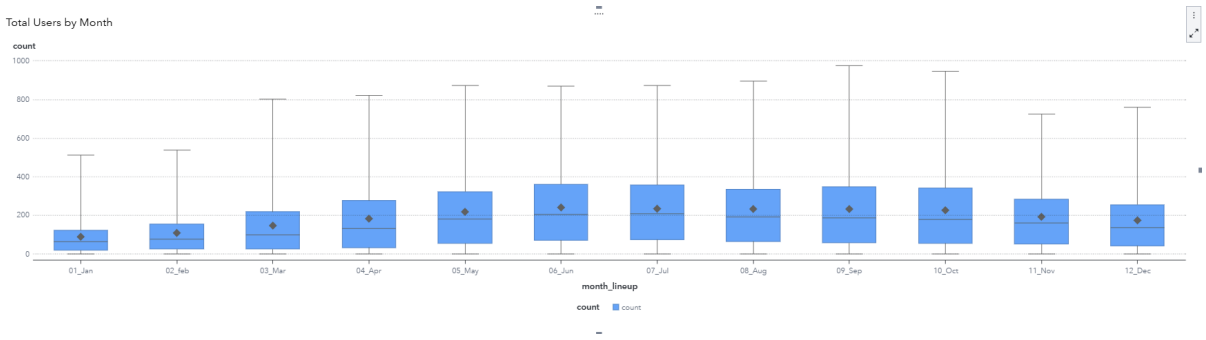



Fig 2.21 boxplot of count against month

**Welch's Anova-registered:**

(Dependent variable: count; Categorical variable: month)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Since **P-value <0.05**, we **reject H0** and conclude that there are statistically significant difference between count (Fig 2.22) and month_name.

| Levene's Test for Homogeneity of count Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| month_name | 11 | 1.249E12 | 1.136E11 | 39.34 | <.0001 |
| Error | 10874 | 3.139E13 | 2.8864E9 | | |

| Welch's ANOVA for count | | | |
|---|---|---|---|
| **Source** | **DF** | **F Value** | **Pr > F** |
| month_name | 11.0000 | 132.45 | <.0001 |
| Error | 4273.8 | | |

Fig 2.22 ANOVA result (count vs month_name)

### Insight 10: Time of the Day have effect on Bike Usage

-The mean total users shows significant variation throughout the day.

-2 main peaks observed during the morning (**7 AM - 9 AM**) and evening (**4 PM - 6 PM**) commuting hours.

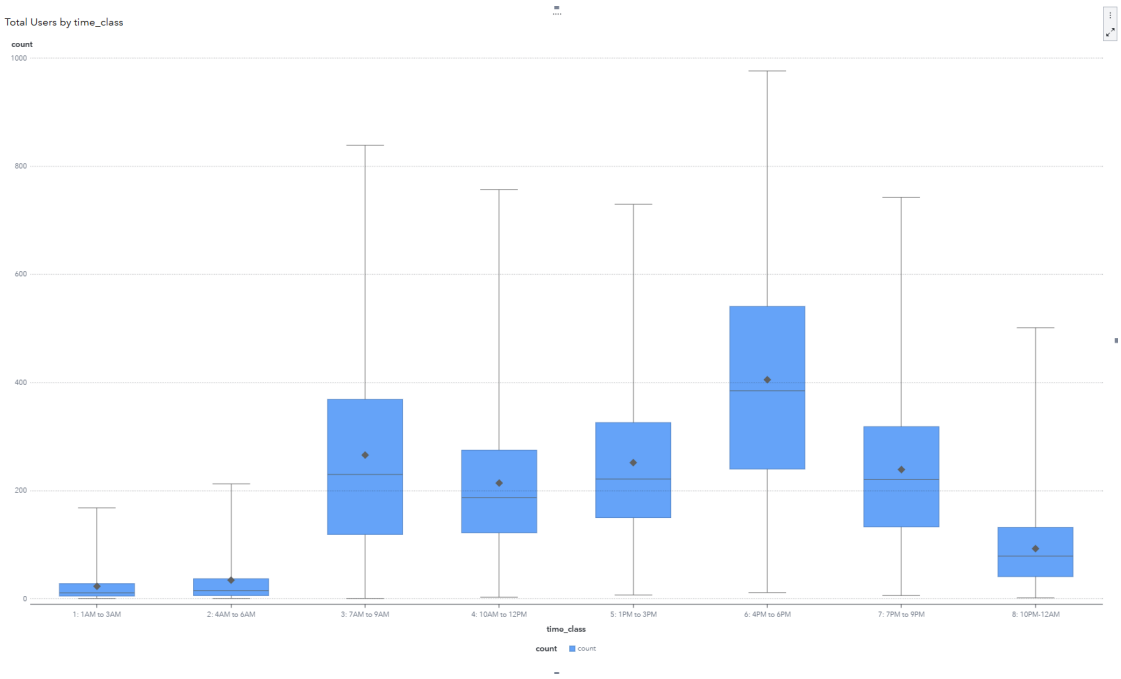



Fig 2.23 boxplot of count against time_class

**Welch's Anova-registered:**

(Dependent variable: count; Categorical variable: time_class)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Since **P-value <0.05**, we **reject H0** and conclude that there are statistically significant difference between count (Fig 2.22) and time_class.

| Levene's Test for Homogeneity of count Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **windspeed_binned** | 5 | 1.137E11 | 2.274E10 | 6.58 | <.0001 |
| **Error** | 10880 | 3.761E13 | 3.4568E9 | | |

| Welch's ANOVA for count | | | |
|---|---|---|---|
| **Source** | **DF** | **F Value** | **Pr > F** |
| **windspeed_binned** | 5.0000 | 32.51 | <.0001 |
| **Error** | 32.0097 | | |

Fig 2.24 ANOVA result (count vs time_part)

### Insight 11: Missing/ Insufficient Data to Cover Range of Humidity

-Insufficient/ no data noted in the range 1-14 % & 95-98 %.

-There are also a small number of data that are with 0 % to be noted as it's quite unlikely for this to happen.

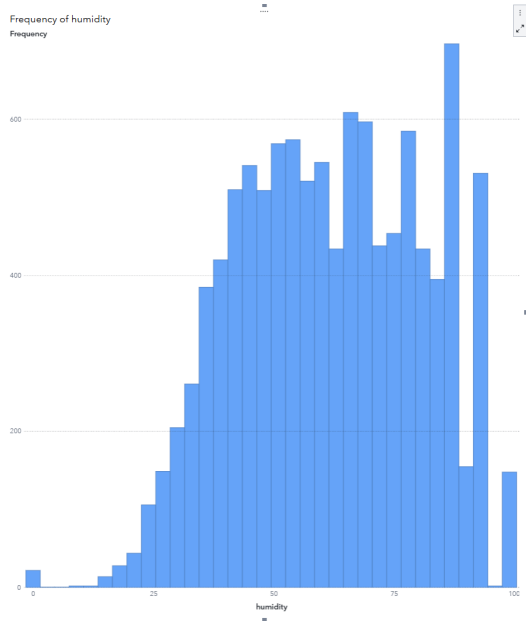



Fig 2.25 frequency of humidity

**Insight 12: Missing/ Insufficient Data to Cover Range of Windspeed Missing data noted in certain ranges for windspeed.**

-Missing data noted in certain ranges for windspeed.

-More data might be required to make the analysis more accurate.

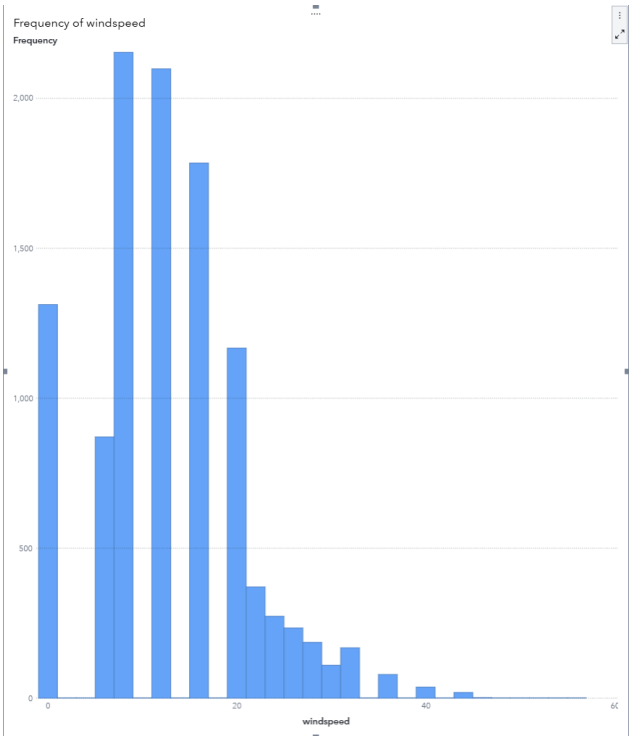



Fig 2.26 frequency of windspeed

### Insight 13: Windspeed has effect on Total Users below 40 kmph

-Minimal variation in the mean number of users within the 29 to 40 km/h wind speed category (Fig 2.30).

-This lack of variance is a consistent trend across the wind speed data that is available for analysis.

-It's important to note that, as per a previous insight 12, there is insufficient data in the 0 to 20 km/h wind speed range and this might affect the accuracy of the analysis.

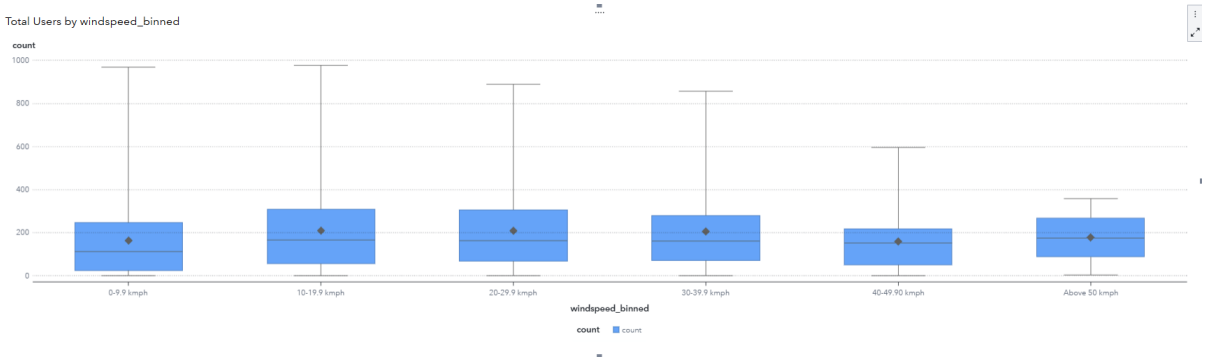



Fig 2.27 boxplot of count against windspeed

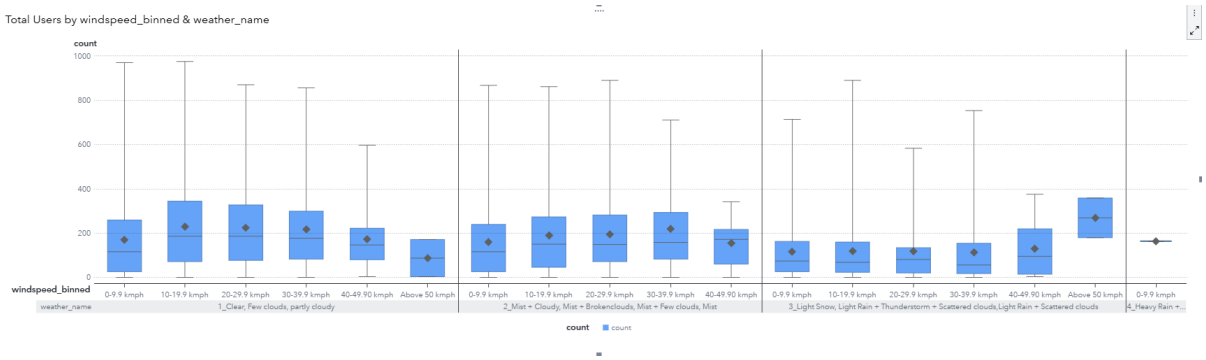



Fig 2.28 boxplot of count against windspee by weather type

**Welch's Anova-registered:**

(Dependent variable: count; Categorical variable: windspeed)

**H0:** All group means are equal.

**Ha**: At least one group mean is different.

Since **P-value <0.05**, we **reject H0** and conclude that there are statistically significant difference between count (Fig 2.29) and windspeed.

Levene's Test for Homogeneity of count Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| windspeed_binned | 5 | 1.137E11 | 2.274E10 | 6.58 | <.0001 |
| Error | 10880 | 3.761E13 | 3.4568E9 | | |

Welch's ANOVA for count

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| windspeed_binned | 5.0000 | 32.51 | <.0001 |
| Error | 32.0097 | | |

Fig 2.29 ANOVA result (count vs wind_speed)

Least Squares Means for effect windspeed_binned
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: count

| i/j | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | <.0001 | <.0001 | 0.0001 | 1.0000 | 1.0000 |
| 2 | <.0001 | | 1.0000 | 0.9986 | 0.5502 | 0.9993 |
| 3 | <.0001 | 1.0000 | | 0.9997 | 0.5789 | 0.9994 |
| 4 | 0.0001 | 0.9986 | 0.9997 | | 0.6758 | 0.9996 |
| 5 | 1.0000 | 0.5502 | 0.5789 | 0.6758 | | 1.0000 |
| 6 | 1.0000 | 0.9993 | 0.9994 | 0.9996 | 1.0000 | |

Fig 2.30 Least Squares Means result (count vs wind_speed)

## Interpretation of analysis results

The accuracy of this analysis is limited by the completeness and quality of the provided data. A sanity check mitigates but doesn't eliminate the risks completely.

Missing or potential incomplete data was also observation for:

- Certain range of value for humidity and wind speed
- Humidity with value of 0%
- Type 4 weather

## Managerial Communication

While this study established relationships with several factors, it may not be fully representative because it did not account for other such as those related to the built environment, public transportation, and socio-demographics (Ezgri Eren, 2019).

Additional factors should also be considered, including political conditions, charges per time block, competitor pricing, and competitor actions.

It is also important to note that while overall bike demand for HappyRides increased in 2024 compared to 2023, its revenue suffered. This suggests that despite market growth, other competitors might be gaining market share, which is negatively impacting the company’s revenue margin.

## References

Ezgi Eren, Volkan Emre Uz (2019) “A review on bike-sharing: The factors affecting bike-sharing demand”. Sustainable Cities and Society, (54) 1-12.

https://doi.org/10.1016/j.scs.2019.101882