Analysis of Monthly Distribution of Cyclists on Selected Counting Stations in Düsseldorf

Muhammad Haider Zaidi

June 1, 2023

Contents

1 Introduction

As cities strive to create cyclist-friendly infrastructure, one approach to achieve this goal is through the collection and analysis of relevant data. By extracting insights from the data distribution, urban planners and policy makers can gain valuable insights into the usage patterns and demands of cyclists in different areas.

The present study focuses on the city of Düsseldorf and aims to analyse the data and evaluate the difference of the Data Distribution of cyclists passing by, recorded by four different counting stations over a one-year period in 2021. A descriptive analysis of the collected data provides an initial overview of the cycling patterns across the city. However, to determine if there are significant differences in the distribution of cyclist counts among the stations, statistical methods and hypothesis testing is employed.

The analysis conducted shows that there is notable difference in the distribution of number of cyclists recorded, among the chosen counting stations. These results can further help urban planners identify the particular areas that are lacking in resources.

The report begins with a problem statement and research objectives, outlining the goals of the study. It then proceeds to the methodology section, which is divided into descriptive analysis and hypothesis testing, explaining these methods. The descriptive analysis includes the use of plots such as line plots, bar charts, and box plots to visualize and analyze the data. The hypothesis testing section explains the use of ANOVA and Tukey's HSD test for statistical comparisons. The evaluation section conducts descriptive analysis along with the mentioned visualisations. In the second part findings of the analysis, including the results from the Hypothesis test are presented.

2 Problem Statement and Research Objectives

The first objective of this statistical research paper is to conduct a descriptive analysis of the data. By utilizing appropriate statistical measures such as descriptive statistics and data visualization techniques, including exploratory data analysis (EDA), the most important variables will described in the data set. This will provide valuable insights into the distribution and patterns of cycling activity in different areas of the city. For instance, whether there are any noticeable differences in cyclist counts between the stations will be investigate.

Furthermore, to ascertain significant variations in the distribution of cyclist counts among the four counting stations over the course of several months, hypothesis testing will be conducted. This will enable us to determine if there are statistically significant differences and establish which stations exhibit these disparities if our hypothesis holds true.

The data set used in this statistical research report consists of 15-minute counts of cyclists passing by on specific streets in Düsseldorf. The data was collected in 2021 using automatic counting stations installed on various streets throughout the city. The data set includes variables such as date and time of the counts, as well as the names of the counting stations: "fleher deich ost stromaufwaerts," "fleher deich west stromabwaerts," "okb nord," and "okb sued.". Fortunately, there are no missing values in this data set, ensuring the completeness of the data.

3 Methodology

3.1 Descriptive Analysis

3.1.1 Plots

Three types of statistical visualisations are used in this report. These are as follows:

Line plot for time series data is a graphical representation that displays the changes in a variable or multiple variables over time using connected data points, allowing for the observation of trends, patterns, and fluctuations.

Bar chart is a statistical representation used to visualize seasonal data by presenting categories on one axis (usually the x-axis) where the height or length of each bar represents the magnitude or frequency of the data category.

Box plot is a graphical representation that displays the distribution of frequencies (in this particular case) using a rectangular box, which encompasses the interquartile range, and includes a vertical line (whisker) indicating the minimum and maximum frequencies. The **interquartile range** (IQR) represents the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a data set. It provides a measure of the dispersion or spread of the data within the middle 50% of the distribution, excluding the lowest and highest 25% of the data.

3.2 Hypothesis testing

3.2.1 ANOVA test

ANOVA, or analysis of variance, is a statistical method used to test the equality of means among two or more groups. It assesses whether the differences observed in the means of the groups are statistically significant or

simply due to random variation.

ANOVA is an appropriate choice of method for our use case as it is suitable for analyzing the relationship between a continuous independent variable and a dependent variable (in this case, date-time column and the count of cyclists).

The mathematical explanation of ANOVA involves partitioning the total variability in the data into two components: the variability between groups and the variability within groups.

Let's consider an example with k groups and n total observations. We denote the data as x_{ij}, where i represents the group index (1 to k) and j represents the observation index within each group (1 to n). The overall mean of the data is denoted as μ.

The ANOVA hypothesis can be stated as follows:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad (nullhypothesis)$$

$$H_1 : Atleastoneofthemeansisdifferentfromtheothers \quad (alternativehypothesis)$$

The total sum of squares (SST) is calculated by summing the squared deviations of each observation from the overall mean:

$$SST = \sum_i \sum_j (x_{ij} - \mu)^2$$

The between-group sum of squares (SSB) quantifies the variability among the group means. It is calculated by summing the squared deviations of each group mean from the overall mean, weighted by the number of observations in each group:

$$SSB = \sum_i (n_i(\bar{y}_i - \mu)^2)$$

Here, n_i represents the number of observations in group i, and $\bar{y}_i$ is the mean of group i.

The within-group sum of squares (SSW) captures the variability within each group. It is calculated by summing the squared deviations of each observation from its group mean:

$$SSW = \sum_i \sum_j (x_{ij} - \bar{y}_i)^2$$

The degrees of freedom for the SSB and SSW are $(k-1)$ and $(n-k)$, respectively.

The F-statistic is then calculated as the ratio of the mean square between groups (MSB = SSB / $(k-1)$ degrees of freedom) to the mean square within groups (MSW = SSW / $(n-k)$ degrees of freedom):

$$F = \frac{MSB}{MSW}$$

Under the null hypothesis, the F-statistic follows an F-distribution with $(k-1)$ and $(n-k)$ degrees of freedom. By comparing the obtained F-value with the critical value from the F-distribution at a given significance level, we can determine whether to reject or fail to reject the null hypothesis. [2] [1]

3.2.2 Tukey's HSD test

Tukey's HSD test, is a statistical procedure used for pairwise comparisons of means. It allows us to determine whether the means of several groups differ significantly from each other.

There's a likely hood of running into observing false positives due to multiple testing. Tukey's HSD test resolves this by controlling the family-wise error rate, which is the probability of making at least one Type I(incorrectly rejecting the null hypothesis when it is true) error among all the pairwise comparisons.

Let's consider a scenario where we have k groups, each with a sample size of n_i and a sample mean of $\bar{y}_i$, where $i = 1, 2, \ldots, k$. The null hypothesis is that the means of all groups are equal, and the alternative hypothesis is that at least one mean differs significantly from the others.

To perform Tukey's HSD test, we calculate the studentized range statistic, denoted as q, for each pair of group means. The studentized range statistic is defined as:

$$q = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{\frac{MSE}{n}}},$$

where $\bar{y}_i$ and $\bar{y}_j$ are the means of groups i and j, MSE is the mean square error from the analysis of variance (ANOVA), and n is the common sample size.

The critical value for Tukey's HSD test can be obtained from the studentized range distribution with degrees of freedom equal to the total sample size minus the number of groups. We compare the absolute value of q with the critical value to determine whether the difference between two means is statistically significant.

If the absolute value of q is greater than the critical value, we reject the null hypothesis and conclude that the means of the corresponding groups differ significantly. Otherwise, we fail to reject the null hypothesis, indicating that there is not enough evidence to suggest a significant difference between the means. [3] [1]

4 Evaluation

4.1 Task 1

Table 1 provides an overall summary of the data. A key observation here is that attributes such as mean, std deviation, quartile range imply varying degrees of dispersion in the data. To visualise this we can look at box plot provided in the appendix.

desc	fleher deich ost stromaufwaerts	fleher deich west stromabwaerts	okb nord	okb sued
Count	35036.00	35036.00	35036.00	35036.00
Mean	5.87	6.99	15.85	10.30
Standard Deviation	8.13	10.55	16.70	11.63
Minimum	0.00	0.00	0.00	0.00
25th Percentile	0.00	0.00	2.00	1.00
50th Percentile (Median)	3.00	3.00	10.00	6.00
75th Percentile	9.00	9.00	26.00	16.00
Maximum	107.00	134.00	109.00	193.00

Table 1: Descriptive Statistics

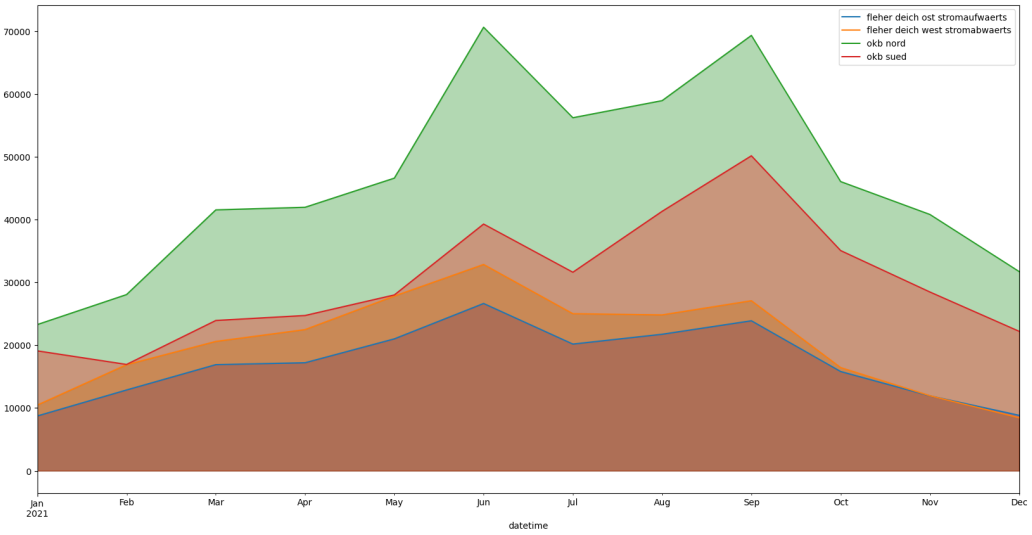

Figure 1: Monthly Aggregated line plot

Figure 1 illustrates fluctuations of cyclists counts over the year in contrast

to each of the stations. Whereas, the bar chart provided in Figure 2 verifies that more cyclists are active in summer then in winters as expected.

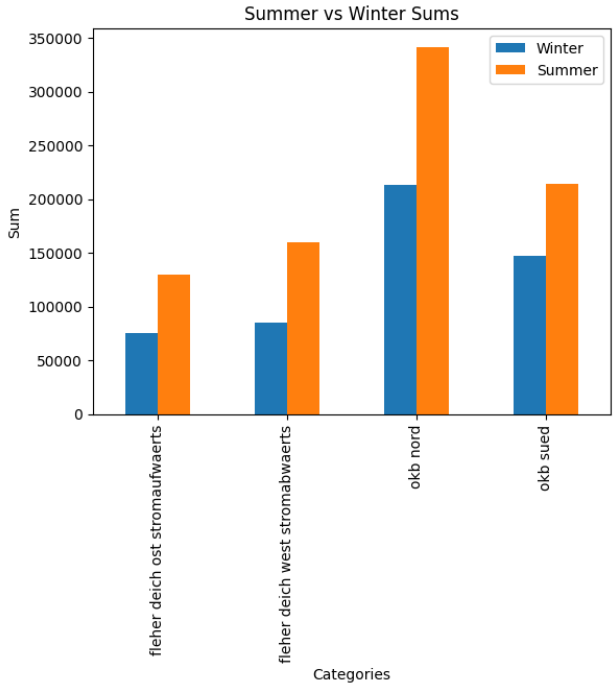

Figure 2: Seasonal Comparison

4.2 Task 2

Initially we've set up a contingency table as shown in Table 2, by aggregating the station data on months while limiting the data to only 7:45 - 8:00 time frame.

Using the Contingency table ANOVA was performed. The null hypothesis is that there is no association amongst each station. The alternative hypothesis is that there is a significant association. Choosing an error chance of 5 % we set the value for significant level i.e. alpha to 0.05.

datetime	fleher deich ost stromaufwaerts	fleher deich west stromabwaerts	okb nord	okb sud
2021-01-31	114.0	76.0	395.0	212.0
2021-02-28	144.0	128.0	385.0	203.0
2021-03-31	237.0	174.0	629.0	415.0
2021-04-30	216.0	132.0	589.0	311.0
2021-05-31	241.0	166.0	597.0	289.0
2021-06-30	373.0	278.0	1086.0	473.0
2021-07-31	261.0	198.0	705.0	303.0
2021-08-31	297.0	224.0	898.0	529.0
2021-09-30	438.0	314.0	1141.0	786.0
2021-10-31	325.0	197.0	736.0	507.0
2021-11-30	354.0	202.0	831.0	521.0
2021-12-31	197.0	120.0	536.0	331.0

Table 2: Contigency table

The ANOVA test yields us the p-value of **4.9826e-14** (which is a very small number close to zero) suggests that the observed differences between the distribution over the months for the 4 counting stations are extremely unlikely to be due to chance variation alone. Therefore, our alternate hypothesis is correct.

Furthermore, multiple testing was conducted by pairwise Tukey's HSD test. Four stations pairs' rejected the null hypothesis thus concluding that these station have significant difference among them. Refer to the Table 3 to see which station pairs have significant difference (highlighted in bold).

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject
fleher deich ost stromaufwaerts	fleher deich west stromabwaerts	-82.3333	0.5839	-254.8563	90.1896	False
fleher deich ost stromaufwaerts	**okb nord**	444.25	0.0	271.727	616.773	True
fleher deich ost stromaufwaerts	okb sued	140.25	0.1475	-32.273	312.773	False
fleher deich west stromabwaerts	**okb nord**	526.5833	0.0	354.0604	699.1063	True
fleher deich west stromabwaerts	**okb sued**	222.5833	0.0067	50.0604	395.1063	True
okb nord	**okb sued**	-304.0	0.0001	-476.523	-131.477	True

Table 3: Tukey's HSD test Results

5 Summary

In this statistical report, the focus is on analyzing the data distribution of cyclists passing by four different counting stations in Düsseldorf over a one-year period in 2021. The goal is to evaluate if there are significant differences in the distribution of cyclist counts among the stations and if so, which stations in particular had a difference in data distribution. The study begins with a descriptive analysis of the data. Visualisations techniques were used to develop an understanding of the data; Furthermore, "descriptive statistics" measures were used to understand the nature of the data, such as if there exists variation in the data. Coming towards the second objective of the report that is to find the difference among the counting station. Statistical techniques for Hypothesis testing where applied: ANOVA test and Tukey's HSD test, appropriate for the nature of our data set and the use case. ANOVA test concluded that there does exists a difference among the counting stations yielding a very small p-val. Additionally, to determine which counting stations have a difference of data distribution, pair-wise Tukey's HSD test concluded four counting stations have a difference i.e (fleher deich ost, stromaufwaerts okb nord), (leher deich west stromabwaerts, okb nord), (leher deich west stromabwaerts, okb sued) and (okb nord, okb sued).

The concluded results can help the respective authorities in-charged of urban planning develop an understanding of the various areas of the city and it's cycling infrastructure needs. For prospect this research work could be further extended by studying the data on a different interval. In this research study we aggregated our data to monthly, on the contrary we can study the hourly intervals, for instance. This would enable us to explore the variations in cyclist patterns throughout the entire day.

6 Bibliography

References

[1] Alan Agresti and Barbara Finlay (1979) Statistical Methods for the Social Sciences

[2] Gelman, A. (2005). Analysis of Variance: Why It Is More Important Than Ever. *Journal of Educational and Behavioral Statistics*, 30(1), 41–57.

[3] Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2), 99–114.

[4] Pandas Development Team. (2023). *pandas (2.0.2) Powerful data structures for data analysis, time series, and statistics*. Retrieved from *https://pandas.pydata.org*

[5] SciPy Development Team. (2020). *SciPy (1.10.1) Fundamental algorithms for scientific computing in Python*. Retrieved from *https://scipy.org*

7 Appendix

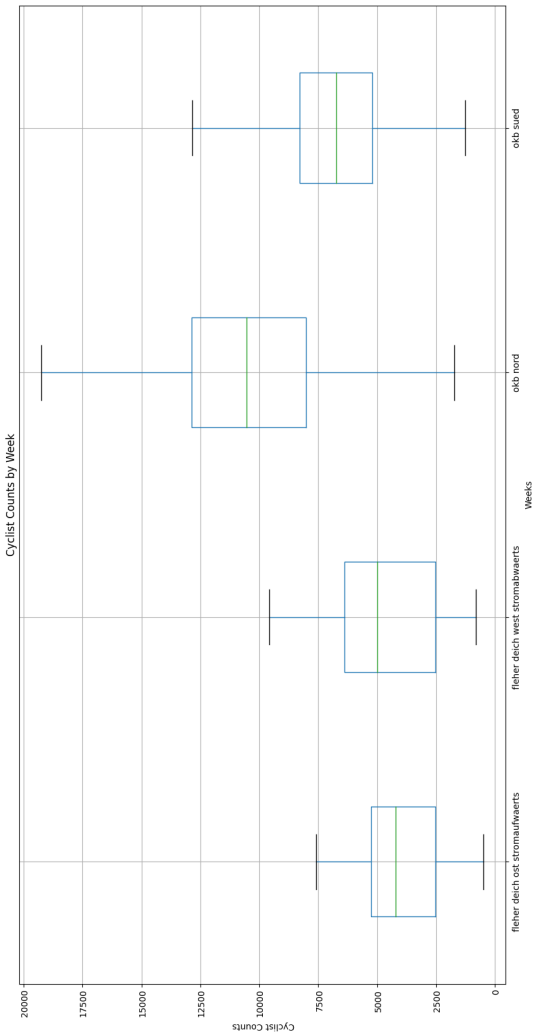

Figure 3: Box Plot