CA Task 1: Bike Sharing Dataset Data Set

Student Number: D21125621
Student Name: Mauricio de Oliveira Reis
Programme Code: TU256/1
Dataset Used: Bike Sharing Dataset Data Set

0. Preliminaries

In this session is to undestand the dataset that was loaded and identify if it needs some transformation:

Table 1: Fist lines of Bike Sharing Dataset

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.1604460	331	654
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.2485390	131	670
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.2483090	120	1229
4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.1602960	108	1454
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.1869000	82	1518
6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518

We can wee in Table 1 that cnt seems to be the sum of casual and registered columns which wasn't so clear on the dataset description and, as specified on the dataset description, everything on this dataset is represented by numbers either integers or doubles, except dteday. So, for the other columns, we need just to pay attention to NAs and impossible values like zeros where we can't possibly have a zero. For dteday we need to check for leading spaces and understand if it need to be converted to a specific datatype for analysis. Also important to mention that this dataset has 16 columns and 731 rows.

Table 2: Quantity of NAs per column

	na_count
instant	0
dteday	0
season	0
yr	0
mnth	0
holiday	0
weekday	0
workingday	0
weathersit	0
temp	0
atemp	0
hum	0
windspeed	0
casual	0
registered	0
cnt	0

As per Table 2, we don't have any NAs on this dataset.

Table 3: Summary Statistics of Bike Sharing Dataset

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum
Min. : 1.0	Length:731	Min. :1.000	Min. :0.0000	Min. : 1.00	Min. :0.00000	Min. :0.000	Min. :0.000	Min. :1.000	Min. :0.05913	Min. :0.07907	Min. :0.0000
1st Qu.:183.5	Class :character	1st Qu.:2.000	1st Qu.:0.0000	1st Qu.: 4.00	1st Qu.:0.00000	1st Qu.:1.000	1st Qu.:0.000	1st Qu.:1.000	1st Qu.:0.33708	1st Qu.:0.33784	1st Qu.:0.5200

Median :366.0	Mode :character	Median :3.000	Median :1.0000	Median : 7.00	Median :0.00000	Median :3.000	Median :1.000	Median :1.000	Median :0.49833	Median :0.48673	Median :0.6267
Mean :366.0	NA	Mean :2.497	Mean :0.5007	Mean : 6.52	Mean :0.02873	Mean :2.997	Mean :0.684	Mean :1.395	Mean :0.49538	Mean :0.47435	Mean :0.6279
3rd Qu.:548.5	NA	3rd Qu.:3.000	3rd Qu.:1.0000	3rd Qu.:10.00	3rd Qu.:0.00000	3rd Qu.:5.000	3rd Qu.:1.000	3rd Qu.:2.000	3rd Qu.:0.65542	3rd Qu.:0.60860	3rd Qu.:0.7302
Max. :731.0	NA	Max. :4.000	Max. :1.0000	Max. :12.00	Max. :1.00000	Max. :6.000	Max. :1.000	Max. :3.000	Max. :0.86167	Max. :0.84090	Max. :0.9725

Table 3 shows the summary statistics and reveals a problematic issue for the hum(humidity) column: it has zero values. We will need to treat those zero values as NAs and decide what to do with the NAs further on in the analysis.

In the end, we had only 1 hum vallues that we transformed in NA.

1. Does the temperature (actual) impact the total number of bikes hired per day?

Before hypothesis test it, let's first understand our variables involved.

Undestanding and checking for normality: temp (Daily temperature)

As per the dataset description this is Normalized temperature in Celsius normalized with the following formula:

$$normalizedTemp = \frac{(temp - temp_{min})}{(temp_{max} - temp_{min})}$$

Let's fist check the Daily temperature distribution and see if it's a normal distribution:

Chart 1: Daily temperature distribution

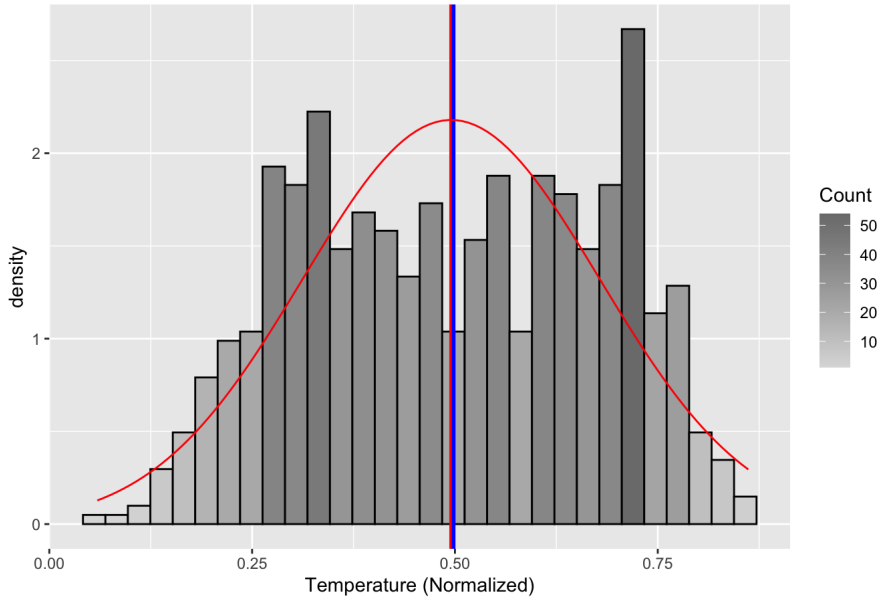
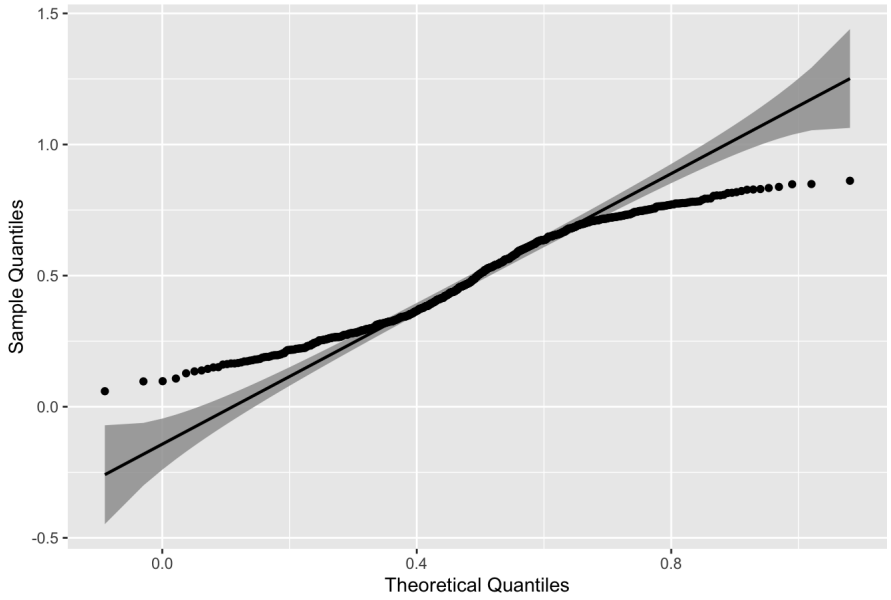


Chart 2: Daily temperature qqplot



As we can see in Chart 1, even though the mean(red) and the median(blue) are quite close, this distribution it's not easy to conclude if this distribution is clouse enough to a normal distribution. The Chart 2, the QQ-Plot shape suggests that the data is too peaked in the middle but a lot of data fall within the Theoretical Quantiles. With that, we better do some extra testa to see if the data is normal.

Although Excess Kutosis for temp is at -6.1749071, way below -2 threshold, after converting it to standardized scores, we noticed that less than 5% of the data falls outside of of -3.29/3.29 threshold (0.9575923% and 0% respectively) and as this dataset has more than 80 rows, we can treat

this data as normal-like.

Undestanding and checking for normality: cnt (number of bikes hired per day)

As stated before, cnt (number of bikes hired per day) seems to be the sum of casual and registered users. Let's first confirm if that's the case to understand better our dataset and see how the proportions look like.

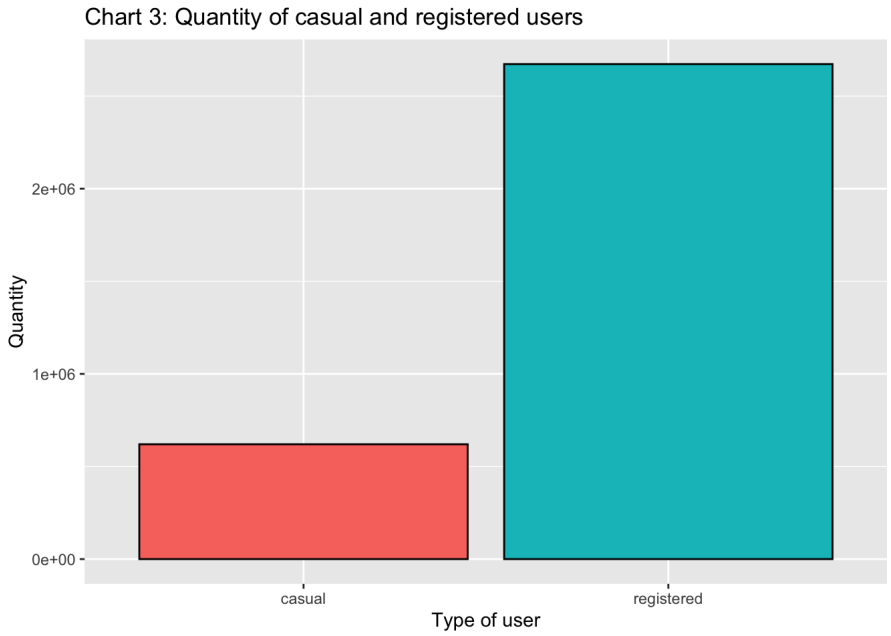
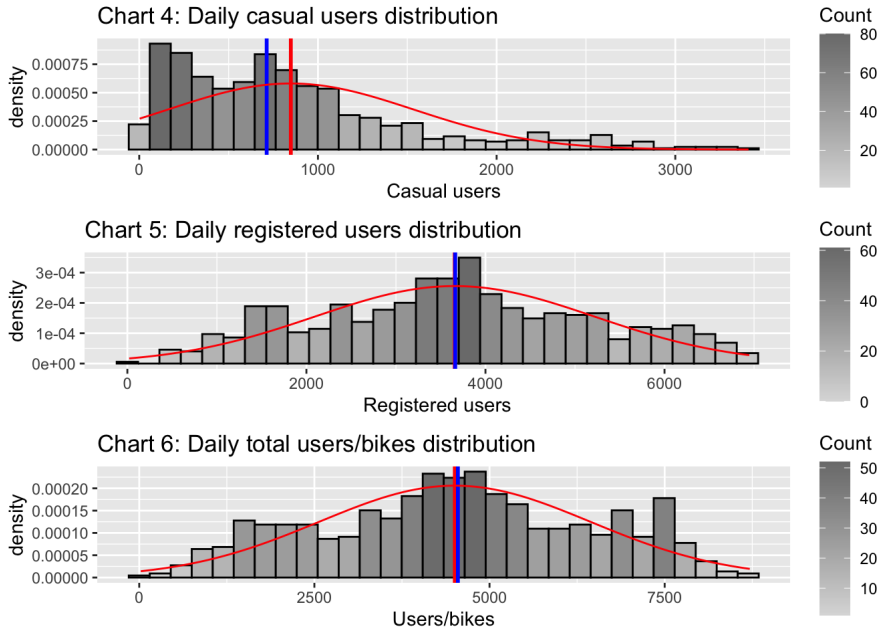


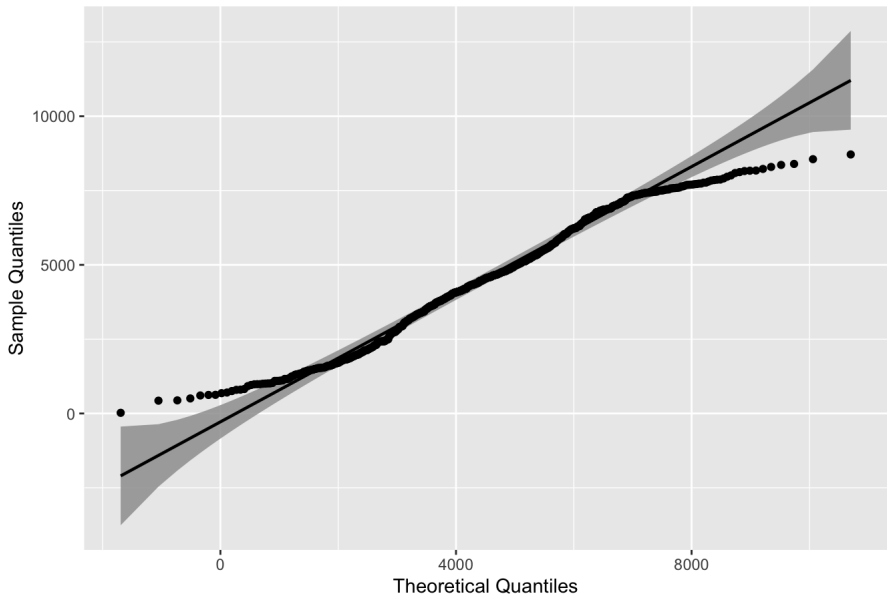
Chart 3 shows that we have way more registered users than casual ones, to be exact, we have 18.83% of users are casual and 81.17% are registered.
Also, we found 0 cases where the sum of casual and registered users are different, therefore, cnt is really the sum of those other two columns and "total rental bikes" is the same as "total users" in a day, this means that the dataset either doesn't consider when the same user rents more than one bike in a day or count of users have duplicates or no user ever rents more than one bike a day.

How each type of user distribution looks like?



Daily casual users (Chart 4) is a right skewed distribution with mean and median a little bit far from one another and showing that is more common to have roughly 1000 users or less per day. Whereas registered users (Chart 5) is a little bit more normal looking and as expected given the high percentage of registered users, the combination distributions (Chart 6) is quite similar to the registered one. From now on, as we already have an understanding of how our users/bikes hired per day are distributed, we will use only the ctn column and call it bikes hired per day.

Chart 7: Bikes hired per day qqplot



Although the QQ Plot (Chart 7) suggest a too centered in the middle shape, the majority of the datapoints fall very close to an ideal normal distribution curve. With that, we better do a few tests to check for normality.

Although Excess Kutosis for cnt is at -4.480924, below the -2 threshold, after converting it to standardized scores, we noticed that less than 5% of the data falls outside of of -3.29/3.29 threshold (1.7783858% and 0% respectively) and as this dataset has more than 80 rows, we can treat this data as normal-like.

Hypotesis Testing

Does the temperature (actual) impact the total number of bikes hired per day?

- Null hypothesis (H_0): Temperature (actual) doesn't impact the total number of bikes hired per day
- Alternative hypothesis H_α: Temperature (actual) impacts the total number of bikes hired per day
- Significance level α: 0.05

Since both distributions were considered normal, a person correlation test was conducted and it has shown a positive correlation of 0.627494 and a p-value of `tmpCntPvalue`, which suggests strong significance level and it's way below our established significance level for this hypotesis testing ($\alpha = 0.05$), therefore, we can reject the null hypothesis and conclude that temperature impacts the total number of bikes hired per day.

2. Does the level of humidity impact the total number of bikes hired per day?

Before hypothesis test it, let's first understand this new variable introduced, humidity.

Undestanding and checking for normality: hum (humidity)

This is a normalised value for humidity where values were divided by 100. As stated before, we've found just 1 zero value for humidity that we've transformed in NA and should not impact our analysis in this dataset with 731 rows.

Let's fist check the humidity distribution and see if it's a normal distribution:

Chart 8: Daily humidity distribution

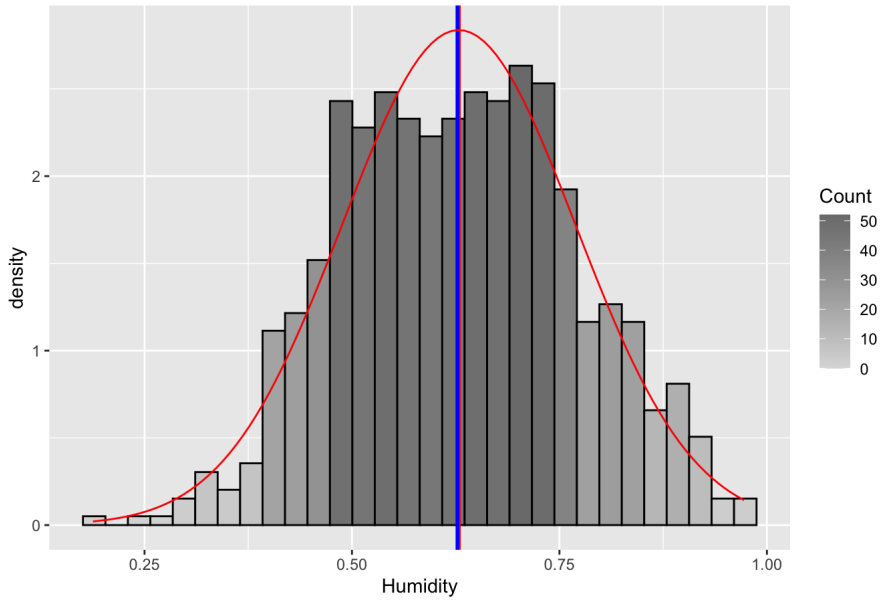
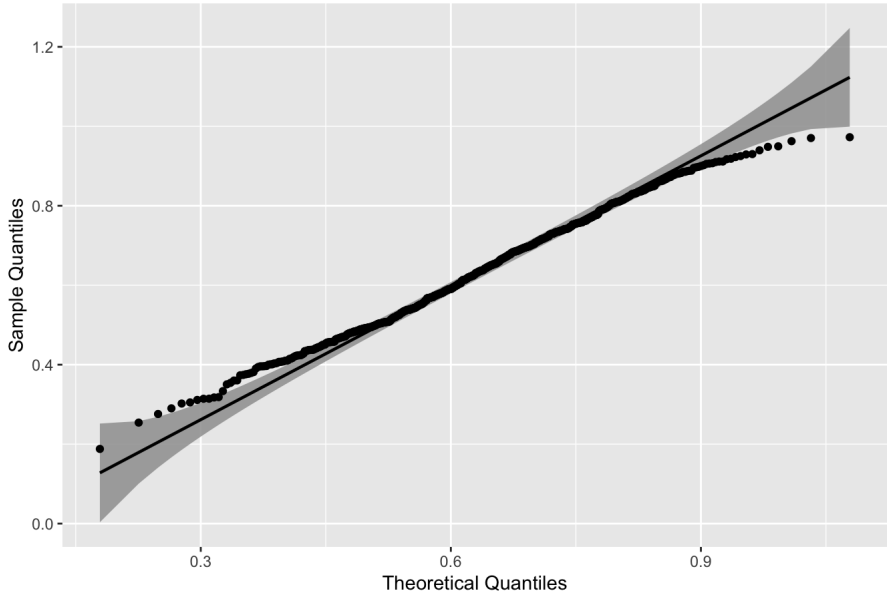


Chart 9: Daily humidity qqplot



As we can see in Chart 8, mean(red) and the median(blue) are quite close, and the distribution looks a lot like a normal distribution, also, in Chart 9, the QQ-Plot shape suggests that the data is somewhat peaked in the middle but a lot of data fall within the Theoretical Quantiles. The charts suggest that this can be treated as a normal distribution but we better also test its skewness and kutosis.

Although Excess Kutosis for humidity is at -2.5147665, slightly below -2 threshold, after converting it to standardized scores, we noticed that less than 5% of the data falls outside of of -3.29/3.29 threshold (4.2465753% and 0% respectively) and, as this dataset has more than 80 rows, we can treat this data as normal-like.

Hypotesis Testing

Does the level of humidity impact the total number of bikes hired per day?

- Null hypothesis (H_0): The level of humidity doesn't impact the total number of bikes hired per day
- Alternative hypothesis H_α: The level of humidity impacts the total number of bikes hired per day
- Significance level α: 0.05

Since both distributions were considered normal, a person correlation test was conducted and it has shown a negative correlation of -0.1146247 and a p-value of 0.0019228, which suggests strong significance level and it's below our established significance level for this hypotesis testing ($\alpha = 0.05$), therefore, we can reject the null hypothesis and conclude that level of humidity impacts the total number of bikes hired per day.

3. Does the total number of bikes hired per day vary according to whether a day is a regular weekday or a weekend?

Undestanding variable: workingday

As per dataset description this variable contains the value 1 if it's not a weekend or holiday and 0 otherwise.

Table 4: Descriptive statistics of bikes hired per Workingday

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X11	1	0	1	231	4330.169	2052.141	4459	4313.497	2345.473	605	8714	8109	0.0382539	-0.9038938	135.0210
X12	2	1	1	500	4584.820	1878.416	4582	4607.078	1944.430	22	8362	8340	-0.0739148	-0.7932044	84.0053

Table 4 shows that there are more than twice as much entries for a regular weekday, there is also some difference in the standard deviation and the mean of the bikes hired on each group which we will analise later on if it's significant.

A Levene Test was conducted to understand if the variance differences between the bikes hired per day in regular weekdays and not regular weekdays is statistically significant and, with a p-value of 0.0374573, we can conclude that this difference is statistically significant and we can treat those two groups as with unequal variances.

Hypotesis Testing

Does the total number of bikes hired per day vary according to whether a day is a regular weekday or a weekend?

- Null hypothesis (H_0): Whether a day is a regular weekday or a weekend, it doesn't impact the total number of bikes hired per day
- Alternative hypothesis H_α: Whether a day is a regular weekday or a weekend, it impacts the total number of bikes hired per day
- Significance level α: 0.05

An independent-samples t-test was conducted to compare bikes hired per day on regular weekday and bikes hired per day on weekends. No significant difference in the amount of bikes hired was found (M=4584.82, SD=1878.42 for number of bikes hired per day on regular weekday, M=4330.17, SD=2052.14 for number of bikes hired per day not on regular weekday), (t(413.9357828)= -1.601, p = 0.11. A very small effect size was also indicted by the eta squared value (0.006). With that, we fail to reject the null hypothesis and conclude that whether a day is a regular weekday or a weekend, it doesn't impact the total number of bikes hired per day.

4. Does the total number of bikes hired per day vary by the day of week?

Undestanding variable: weekday

As per dataset description this variable contains the day of the week, the description doesn't specify which day is Sunday as it can be interpreted as the first or last day of the week depending on the culture so, we need to understand that first.

Table 5: Count of workingdays per weekday

weekday	workingday	count_days
0	0	105
1	0	15
1	1	90
2	0	1
2	1	103
3	0	1
3	1	103
4	0	2
4	1	102
5	0	2
5	1	102
6	0	105

As we can see on Table 5, weekdays 0 and 6 are never working days and weekday 1 is, most of the time, a working day. With that we can assume that day 0 is a Sunday as it's a non working day followed by a day that's usually a working day and proceeded by a day that isn't a working day.

Table 6: Descriptive statistics of bikes hired per weekday

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X11	1	0	1	105	4228.829	1872.497	4334.0	4236.200	1872.524	605	8227	7622	-0.0035002	-0.8036618	182.7370
X12	2	1	1	105	4338.124	1793.074	4359.0	4392.165	1817.668	22	7525	7503	-0.2043353	-0.8030022	174.9861
X13	3	2	1	104	4510.663	1826.912	4576.5	4553.679	1727.229	683	7767	7084	-0.2048072	-0.7859885	179.1434
X14	4	3	1	104	4548.538	2038.096	4642.5	4563.881	2414.414	441	8173	7732	-0.0348064	-0.9922375	199.8517
X15	5	4	1	104	4667.260	1939.433	4721.0	4708.929	2225.383	431	7804	7373	-0.0916553	-0.9701236	190.1771
X16	6	5	1	104	4690.288	1874.625	4601.5	4688.298	1837.683	1167	8362	7195	0.0948397	-0.8689871	183.8221
X17	7	6	1	105	4550.543	2196.693	4521.0	4544.671	2636.063	627	8714	8087	-0.0123826	-1.0045658	214.3753

Table 6 shows us that this dataset is evenly distributed by days of the week, we can also see a slightly increase in the average number of bikes hired as the week goes starting with Sunday at the lowest with 4228.83 bikes hired on average and peaking at Friday with 4690.29 bikes hired, the number already goes slightly down on Saturdays to 4550.54 bikes hired on average. The standard deviations seem quite close from one another but we need to test to see if they are homogenious.

A Bartlett's Test was conducted to understand if the variance differences between the bikes hired per day on each weekday are statistically significant and, with a p-value of 0.377807, we can conclude that this difference is not statistically significant and we can treat those groups as with equal variances when conducting our tests.

Hypotesis Testing

Does the total number of bikes hired per day vary by the day of week?

- Null hypothesis (H_0): The total number of bikes hired per day doesn't vary by the day of week
- Alternative hypothesis H_α: The total number of bikes hired per day varies by the day of week
- Significance level α: 0.05

A one way ANOVA test was conducted to compare number of bikes hired per day and day of the week. No statistically significant difference in the scores for number of bikes hired per day and day of the week was found (F(2, 6, 724) = 0.78, p=0.58) A small effect size was also indicated by the eta squared value (0.01). With that, we failed to reject the null hypothesis and conclude that the total number of bikes hired per day doesn't vary by the day of week.

5. Is weather situation related to the season?

Undestanding variable: weathersit(Weather situation)

As per dataset description this is a categorical variable with numbers from 1 to 4 that mean the following weater situations:
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

Chart 10: Quantity of days with each weather situation

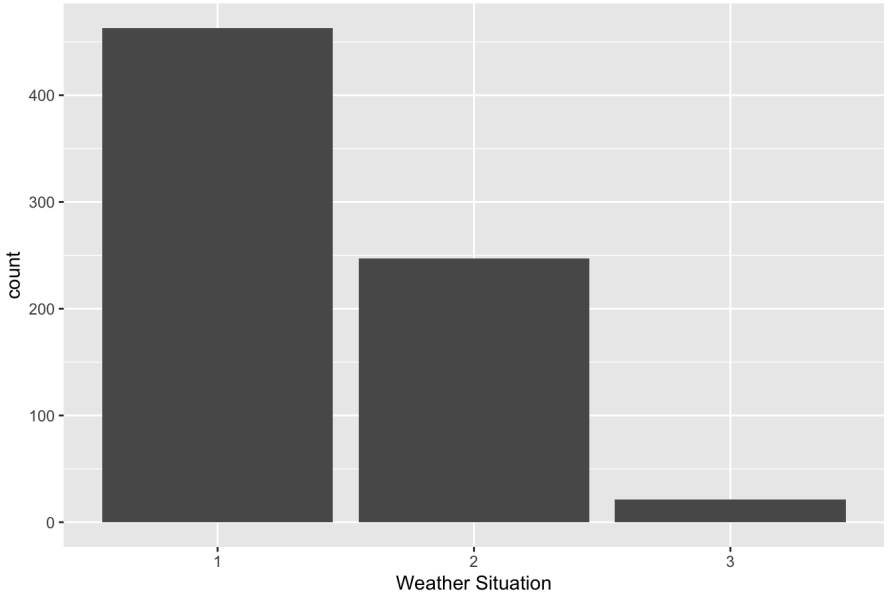


Chart 10 shows that we don't have any registered days with weather situation 4 and very few with 3. So the weather situation is mostly 1 or 2 with the mode being 1.

Undestanding variable: season

As per dataset description this is a categorical variable with numbers from 1 to 4 that mean the following seasons:
1:winter
2:spring
3:summer
4:fall

Chart 11: Quantity of days for each season

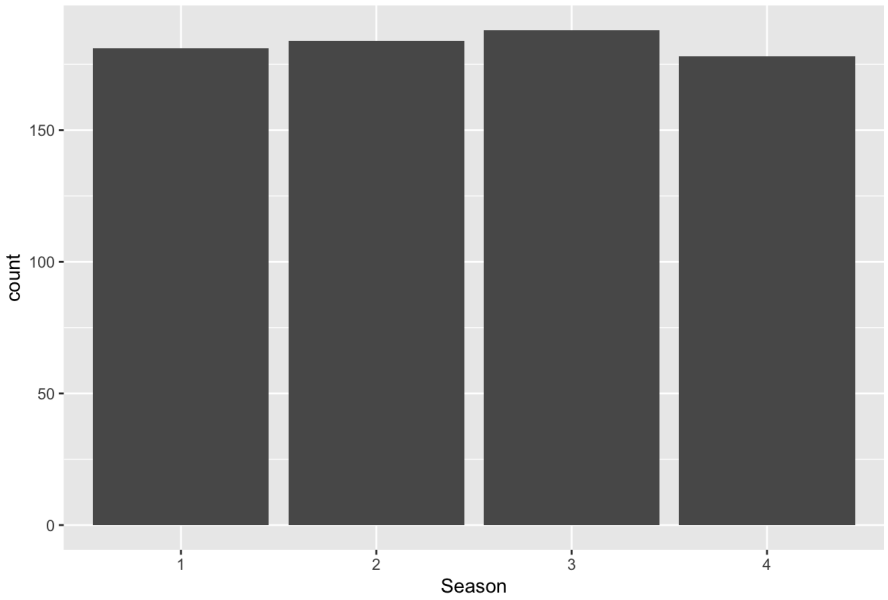


Chart 11 shows that we have roughly the same amount of days for each season on this dataset with the mode being 3(summer) by a very small difference.

Hypotesis Testing

Is weather situation related to the season?

- Null hypothesis (H_0): The weather situation is not related to the season
- Alternative hypothesis H_α: The weather situation is related to the season
- Significance level α: 0.05

Table 7: Days with a specific weather per season

	Season 1	Season 2	Season 3	Season 4
Weather 1	111	113	136	103
Weather 2	66	68	48	65
Weather 3	4	3	4	10

Table 8: Expected Frequencies

	Season 1	Season 2	Season 3	Season 4
Weather 1	114.641587	116.54172	119.075239	112.741450

Weather 2	61.158687	62.17237	63.523940	60.145007
Weather 3	5.199726	5.28591	5.400821	5.113543

Table 9: Observed Frequencies

	Season 1	Season 2	Season 3	Season 4
Weather 1	111	113	136	103
Weather 2	66	68	48	65
Weather 3	4	3	4	10

As the contingency table (Table 7) has shown, a few cells were with a value less than 5 so, a Chi Square test was conducted utilizing Yates correction. In this test we've found a P value of 0.0211793, below our established $\alpha = 0.05$. Table 8 and 9 show the discrepancy between the expected and observed frequencies. We also used Cramer's test to calculate the effect size and, with 6 degrees of freedom and a effect size of 0.1, we discovered that there is a small effect size. With that, we reject the null hypothesis and conclude that the weather situation is related to the season.

6. Summary

On this report we've looked at several factors that might influence the number of bikes hired per day and also looked if the weather situation is related to the season on this particular dataset.

On the side of influencing the number of bikes hired, we've found that the day of the week and also whether it's a regular working day or not do not influence in the number of bikes hired. We've found that temperature and humidity influence the the total number of bikes hired, temperature has a positive correlation meaning that the warmer it's, more bikes will be hired and humidity has a negative correlation with number of bikes hired meaning that the less humidity, more bikes will be hired. Remembering that this report was limited to establish correlation and not causation, questions like whether less humidity really makes users hire more bikes or users just hire more bikes due to another factors that can be closely related to humidity like temperature, was out of the scope of this report.

Looking if season is related to weather situation, we've found that this is true, season really affects the weather situation with a small effect size.