

Individual Assignment Report

SID: 20121076

Name: Humam Al-Taiff

Part 1

The purpose of this report is to identify trends in the bikesharing dataset, some pertinent trends to look for are what conditions lead to more riders and what conditions may lead to less riders. The dataset has 15 columns, each with numerical values (integers and floats) except for the 'dteday' column which is a string. However, in some columns the numbers act as categories.

With this in mind, the breakdown of the data set becomes:

Categorical	Numerical
season	dteday
year	temp
month	atemp
holiday	windspeed
weekday	hum
workingday	registered
weathersit	casual
	count

In my analysis the weather situation category was summarized:

1: Clear

2: Overcast

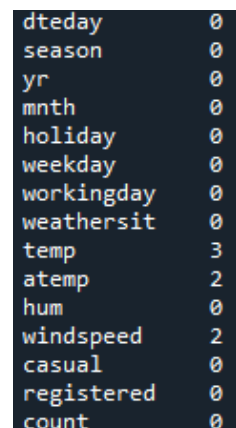
3: Obscure

This is a generalization only to simplify their true categories. This is to make data more readable. The data set initially has 734 rows, 3 more than there are days between 2011 and the end of 2012.

The python libraries used to handle and analyse the data are pandas, matplotlib, numpy and scipy.

Part 2

There are a total of 7 missing values in the data set. 3 due to missing temperature, 2 due to missing apparent temperature and 2 due to missing windspeed. The percentage of missing values in the data set is 0.95 %. As this is a relatively insignificant amount the rows with missing values were removed.



dteday	0
season	0
yr	0
mnth	0
holiday	0
weekday	0
workingday	0
weathersit	0
temp	3
atemp	2
hum	0
windspeed	2
casual	0
registered	0
count	0

Figure 2.1

There are 3 duplicate rows in the data set:

On the 9/01/2011, 29/08/2012, 10/12/2012.

Removing these gets the data to the expected size of 731 rows (with one for the title) corresponding to each day between the start of 2011 and the end of 2012.

As the duplicated rows add nothing new to the dataset and risks changing the shape of the relationships between variables as well as meeting the expected number of rows at the end of their removal.

Statistical analysis of numerical columns was:

The skew of temperature is: 15.63

The skew of apparent temperature is: -0.12

The skew of windspeed is: 0.68

The skew of humidity is: 15.82

The skew of the casual bike user's category is: 1.27

The skew of the registered bike user's category is: 0.03

The skew of the total bike (count) user's category is: -0.05

This indicates that temp, windspeed, humidity and casual columns each will differ from a normal distribution and are likely to contain outliers.

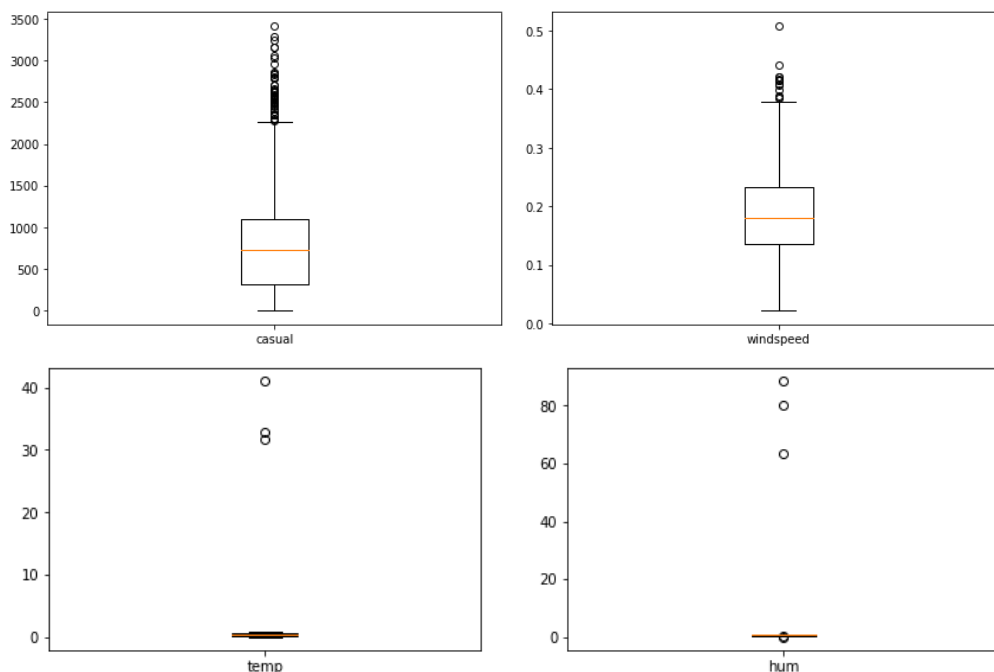


Figure 2.2

Further investigation using graphical methods like boxplots and scatter plots reveal that this is true and there are a number of outliers in each of these attributes.

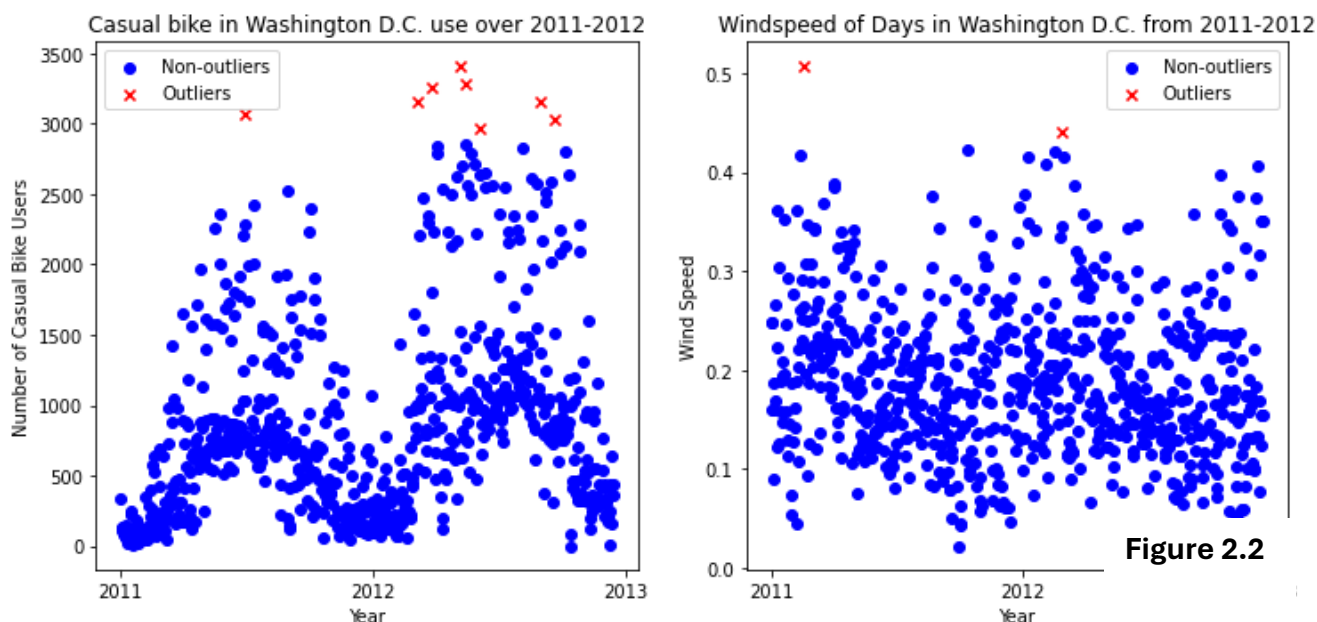


Figure 2.2

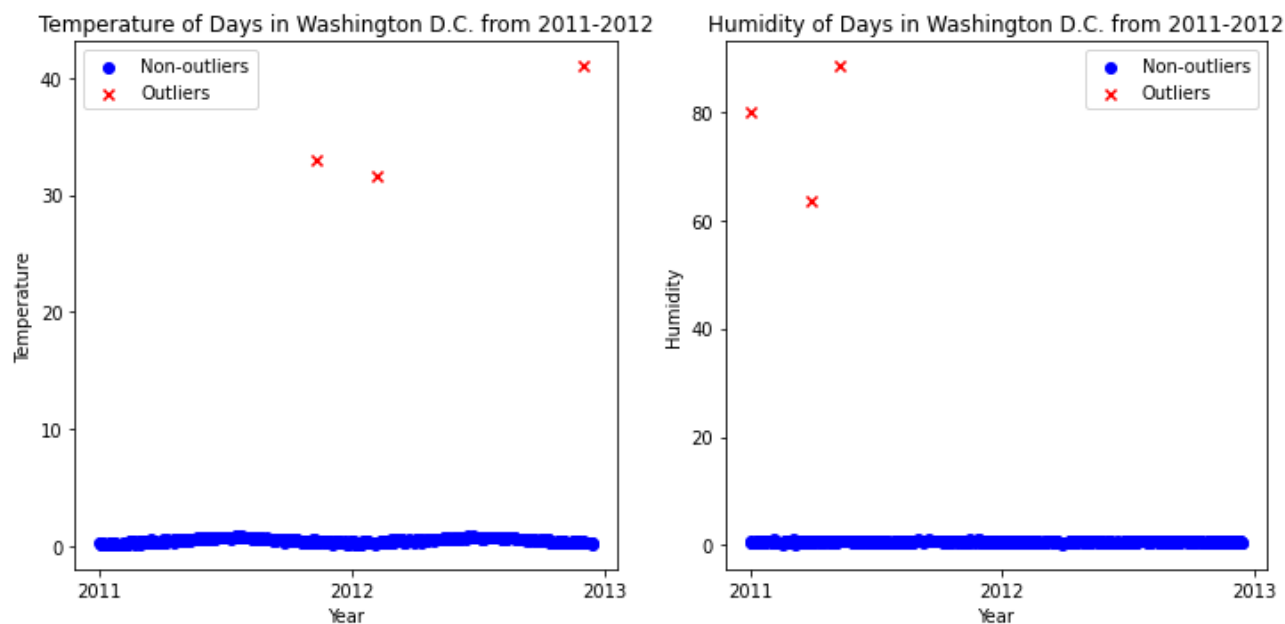


Figure 2.3

From the scatter plots we can see there are 8 outliers due to the casual attribute, 2 outliers are due to the Windspeed attribute, 3 outliers due to the temperature attribute and 3 outliers due to the humidity attribute. These outliers were identified using their z-scores, however this was supplemented by the box and whiskers plot to understand true presence of outliers. This means the percentage of outliers at the beginning of the data set was 2.21 %. This a larger percentage of outliers than I would hope and therefore, tried to mitigate data loss. Due to the

large deviation of the temperature and humidity values it was assumed that they were calculation errors, and notably look like typical values that have yet to be normalized into the data set. Therefore, I imputed the outlier values using the process outlines in the notes to return the data points to the dataset. This increased the data's integrity by maintaining the general shapes of the data and removing outliers. Similarly, this retains the largest number of rows enhancing the potential analysis of the dataset. This imputation reduced the percentage of outliers to 1.52%. Analyzing the casual outliers, a vast majority of them occur on Saturday during summer, the only other outlier occurs on Monday: Independence day. There is not a very clear reason for their values and were thus removed despite having some relatability to the dataset. The windspeed outliers do not have a lot in common with each other and were also thus removed.

Part 3

To determine the change in statistics after cleaning I used the describe function comparing the statistics before and after using percentage change.

	atemp	casual	count	holiday	hum	mnth	registered	season	temp	weathersit	weekday	windspeed	workingday	yr
count	-2.595628	-2.861035	-2.861035	NaN	-2.861035	-2.861035	-2.861035	-2.861035	-2.462380	-2.861035	-2.861035	-2.595628	-2.861035	-2.861035
mean	0.018491	-2.103279	-0.053168	NaN	-33.131630	0.687258	0.421469	0.530327	-21.902352	-0.271739	-1.134991	-0.539181	1.099669	0.147875
std	0.039641	-5.951858	-1.812011	NaN	-97.197958	-0.229779	-0.665985	-0.192418	-91.794561	-0.800397	-0.610654	-1.751770	-0.653126	0.001495
min	0.000000	0.000000	0.000000	NaN	inf	0.000000	0.000000	0.000000	63.235324	0.000000	NaN	0.000000	NaN	NaN
25%	-0.150847	0.555115	1.509614	NaN	0.640962	0.000000	0.480962	0.000000	0.369004	0.000000	0.000000	0.000000	NaN	NaN
50%	-0.129574	0.210822	-0.010993	NaN	0.066587	0.000000	0.450266	0.000000	0.000000	0.000000	0.000000	-0.688493	0.000000	0.000000
75%	0.154911	-1.778386	-1.022461	NaN	-0.199544	0.000000	-0.171938	0.000000	0.126853	0.000000	0.000000	-0.263817	0.000000	0.000000
max	0.000000	-16.275660	-1.824650	NaN	-98.903113	0.000000	0.000000	0.000000	-97.560976	0.000000	0.000000	-16.787037	0.000000	0.000000

From the table we can see that the most major changes happen to the temp and humidity categories. This is because the outliers were extreme and changing them has resulted in a significant change in mean and standard deviation. The next most significant change was the change in mean in the casual column, decreasing it by 2%. Likely because the removed outliers were all in the upper extremities of the casual data set.

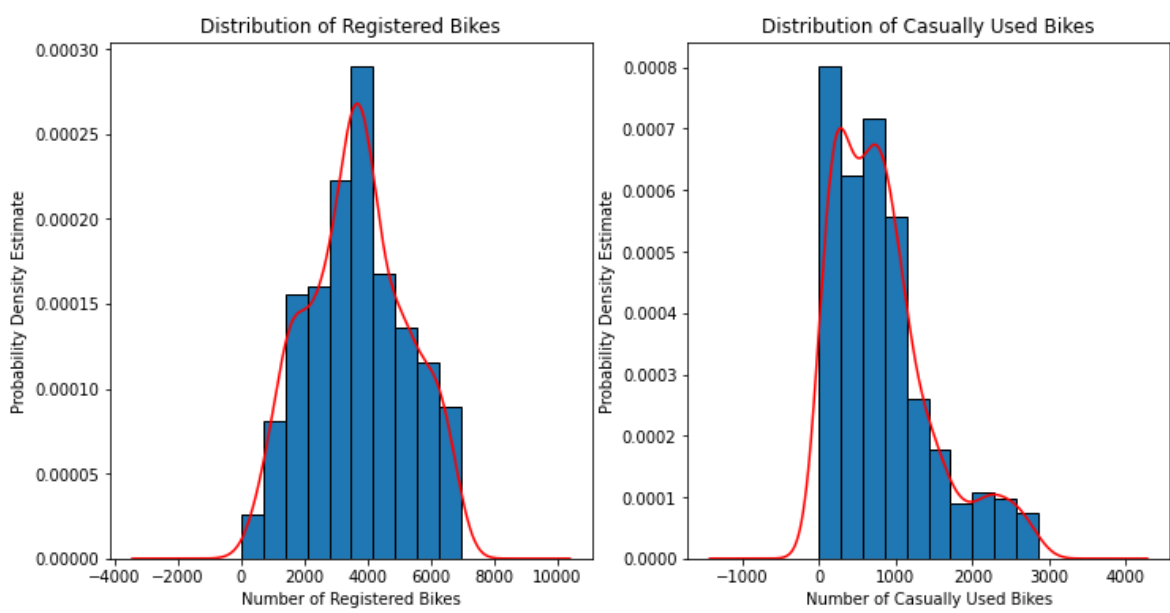


Figure 3.1

Figure 3.1, depicts the distribution of the number of registered bikes and casual bikes. We can see that registered bikes follow a narrow normal distribution shape while the number of casual bikes used have a positive skew where the median of the casually used bikes is less than the mean.

Distribution of Weather Situations from 2011-2012

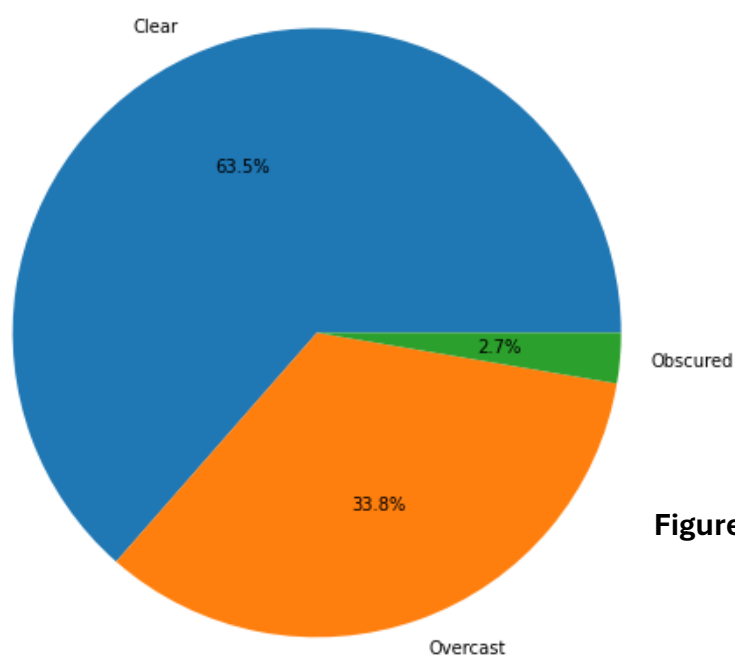


Figure 3.2

Number of days of each weather situation and their months from 2011-2012

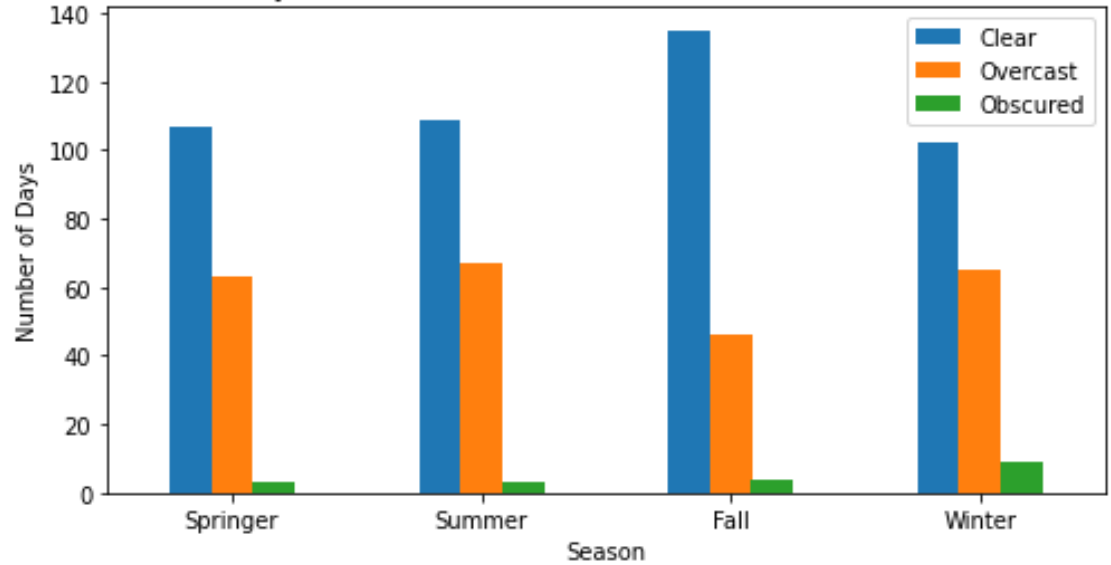


Figure 3.3

Figure 3.2 demonstrates the distribution of weather in the dataset. Evidently, there was more clear weather than any other weather type in the data set while there weren't very many days with obscuring weather (rain and thunderstorms). Next, in figure 3.3 we can observe the number

of days of each weather circumstance and when they occurred in each of the seasons. We can see that winter was likely to have worse weather conditions than other seasons while fall has the least. This coupled with figure 3.4 on the distribution of riders in each season we see a link between weather situations and riders. Usually, the clearer the weather the more riders. Meaning, fall had the greatest proportion of all the seasons. A possible explanation as to why springer does not have a greater portion of riders despite having generally better weather than winter is because a lot of registered users will go on holiday at the end of the year, thereby decreasing the count. This can be read from the seasonal trend of bike users.

Distribution of Bikes over Seasons of the Year

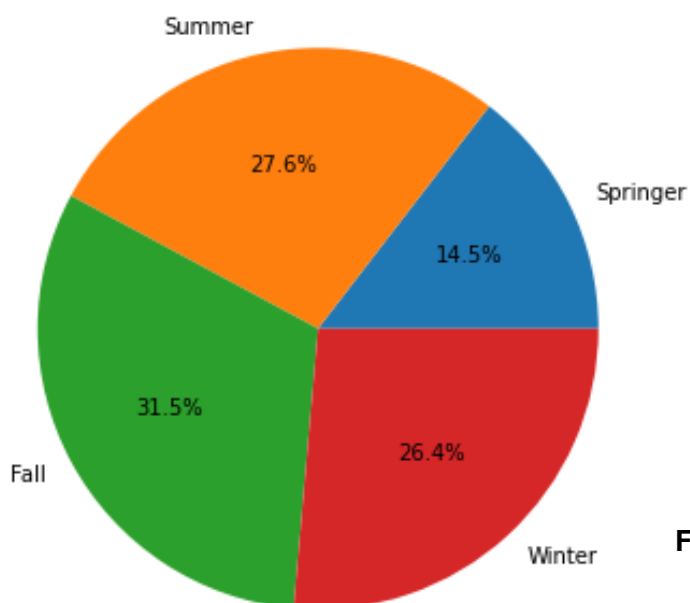
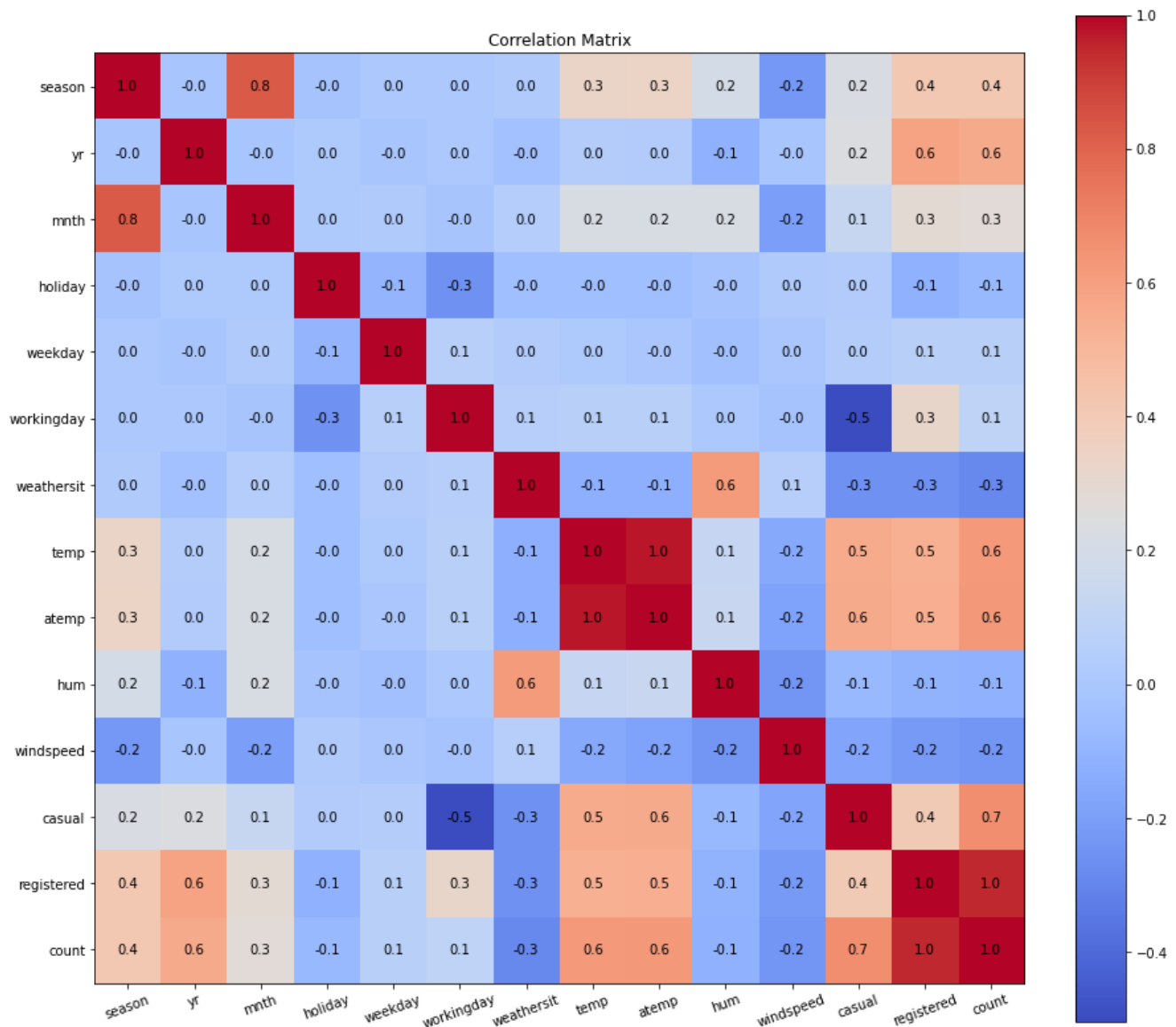


Figure 3.4

Part 4



In this correlation matrix, I deliberately left the categorical columns as numerical as they can still demonstrate some relationship between attributes. The plot shows the strength of correlation between attributes and the nature of their correlation. Some interesting numerical relationships come from count and temp, registered and casual and casual and count. The count of bike users goes up as temperature goes up, indicating a positive correlation. This may be because with a higher temp the weather is nicer, and people are less inclined to stay indoors and are more inclined to go out and use bikes (this is demonstrated in appendix scatter plot). Another interesting relationship between the registered users and casual users. It seems that even when there is an increase in casual users there isn't a strong increase in registered users. Indeed, there is a significant positive relationship, but it is not 1 to 1 like the registered-count relationship. This may be because casual users have are subject to more variation, which is supported by the positive skew distribution in figure 3.1. This becomes especially interesting considering the

casual to count relationship. As expected, there is a strong positive correlation between casual users and the total count of users, though notably it is not 1 to 1. The implication of this is that casual users are not a reliable indication of overall use or trends of the bikesharing system. Yes, they may sketch the bounds of the bikesharing system’s popularity, but they will not always be reliable. This is likely – again – due to the data shape and variation of the casual users over time.

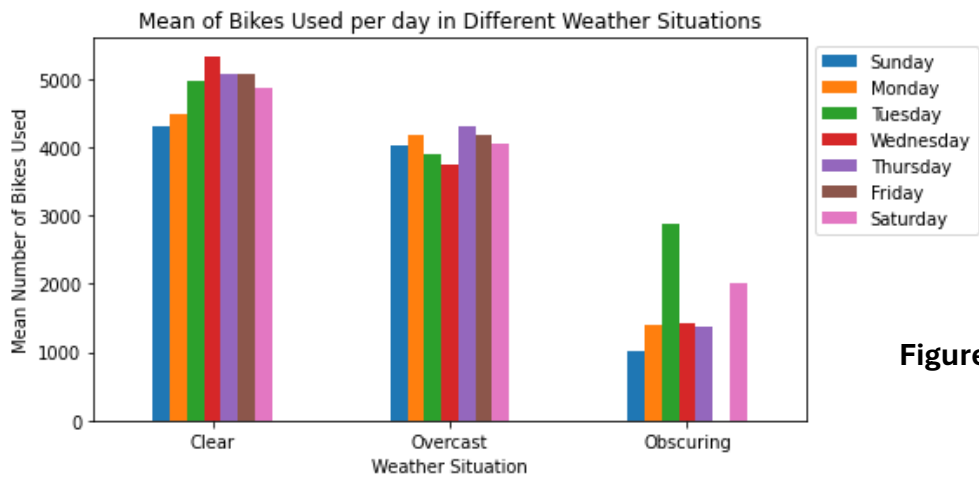


Figure 4.1

For this report I wanted to investigate the relationship between the total number of bikes used in different weather circumstances on different days. This will build on what I discovered in part 3. In figure 4.1 we can clearly see that the mean bike users decreases as the weather situation becomes less and less optimal. Interestingly we also notice that the distribution of users over days of the week changes depending on the weather situation. In clear weather it is a normal distribution. In overcast it becomes a very level distribution with more users distributed over the beginning and ends of the week rather than the weekdays. In obscuring weather the shape of the data breakdown, likely because there were not a lot of days with obscuring weather from 2011 to 2012 as demonstrated by figure 3.2. That being said it is still very apparent that users prefer to ride in more favourable weather than when the weather may increase risks or hazards.

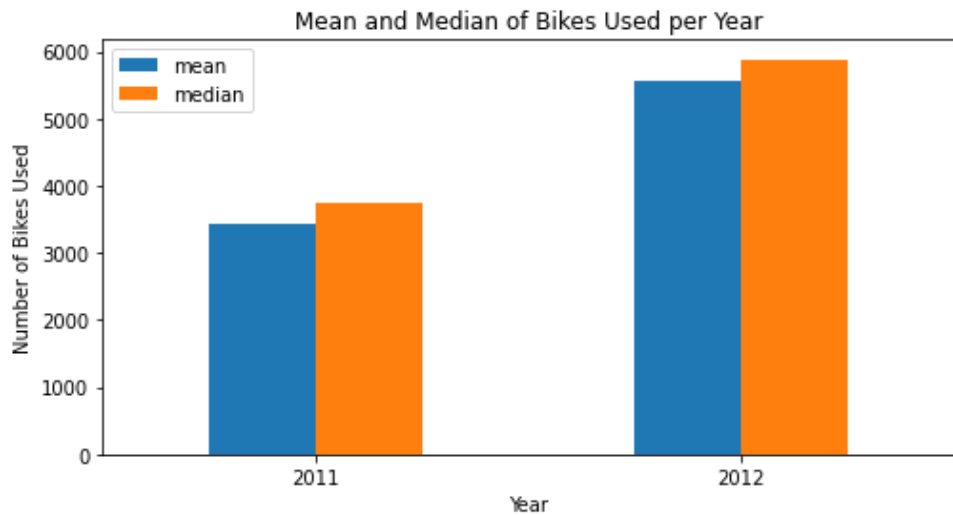


Figure 4.2

For this report I decided to analyse the mean and median of bikes used over the year to help understand any changes in skew for the overall dataset. This is because I have yet to analyse the effect of time on the count of bike users. In figure 4.2 we see there is a clear increase in the count of users from 2011 to 2012. This shows that the bikesharing system got more popular over time. Notably however, the shape of the data remains the same from 2011 to 2012. We can see this as the median remains slightly larger than the mean of the bike users implying a positive skew to the data. This is supported by figure 3.1 where there is a positive skew in the casual users that seems to be influencing the total count over the years.

To analyse the different means of bike users over different seasons I used a bar graph with the standard deviation and standard error to visualize each season's variability. In figure 4.3 and 4.4 we can see that the distribution of bike users follows the distribution previously analysed in figure 3.4. However, that standard deviation demonstrates that these bars may be closer in distribution than what had been previously analysed. The standard deviation demonstrates the high variability of the means of bike users over the seasons especially in summer and winter. This may imply that the relationship between riders and the season is not very reliable or strong. However, in figure 4.4 using the standard error of the seasons we see that the sample itself does not have a lot of variability of bike users over the season, thus implying the trend is reliable and consistent. The importance of this analysis is that it demonstrates whether the bikesharing company can make decisions based on seasonal trends, if there is a large variability in the seasons of bike users then it is harder to create strategies that rely on the basis of seasons from this sample. Furthermore, it demonstrates the distinct seasonal trend of bike users over the year and allows for better planning and hypothesizing, which is determined to be accurate and efficient due to its low variability.

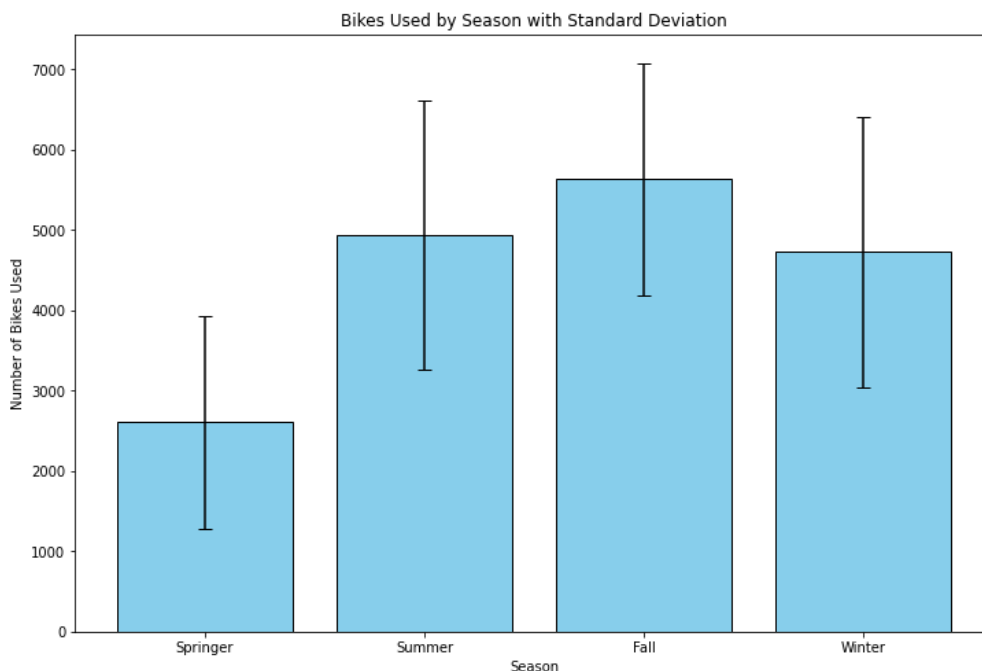


Figure 4.3

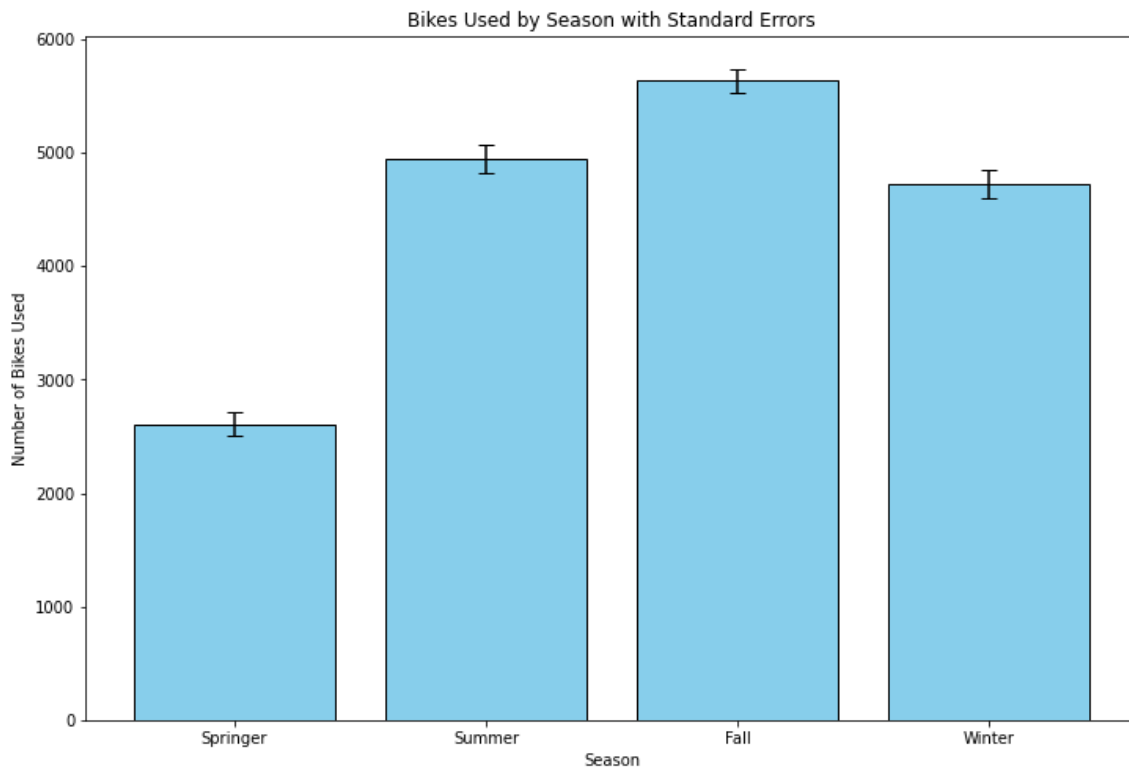


Figure 4.4

Part 5

The key findings in this analysis are:

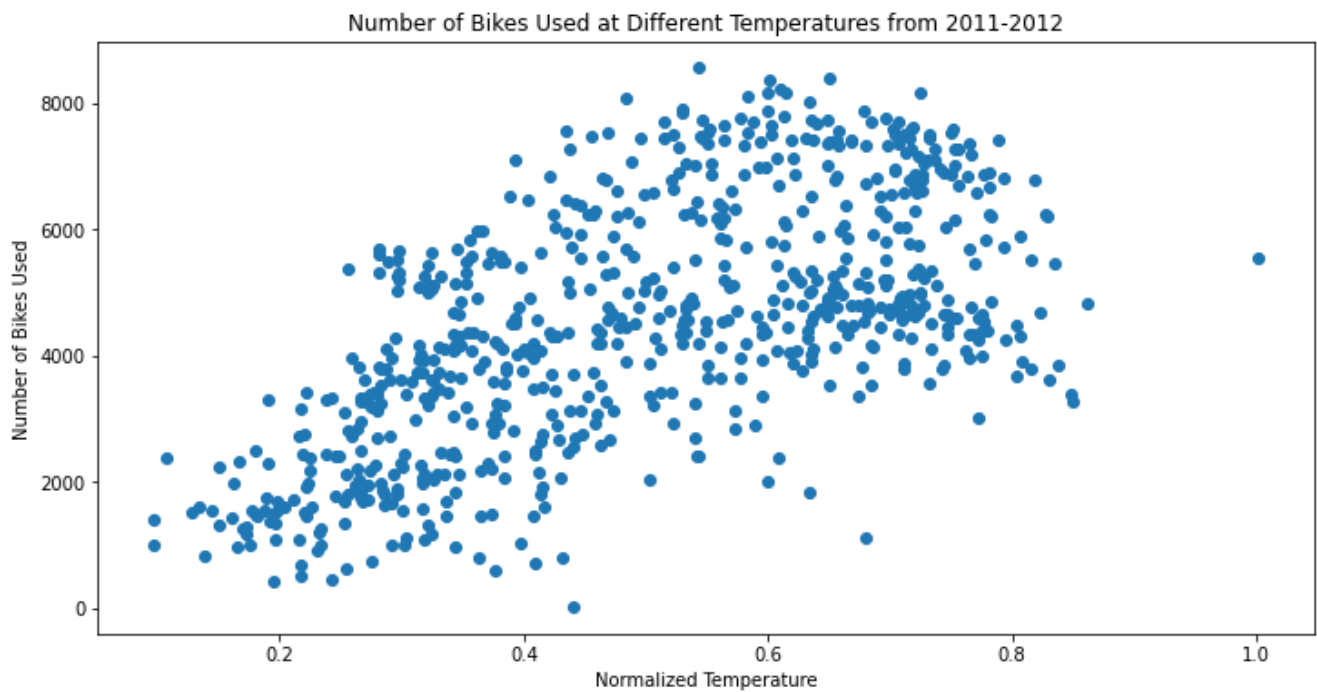
- Casual users of bikes have high variability.
- The number of bike users follow a seasonal trend
- Bike users prefer to bike in more pleasant weather and a good temperature, however this is influenced positively or negatively by the season
- The number of bike users are increasing from year to year with a positive skew
- And bikers prefer warmer temps than colder temps when riding.

Some challenges in the assignment were dealing with outliers, knowing what was acceptable to remove and what were clear calculation errors that could be imputed. Similarly, it was difficult to analyze certain categories as they were registered under numerical values making them effective landmines in data pre-processing as they may produce outliers despite being categorical in nature (this happened with the holidays column). To overcome the first issue, I applied a series of analysis looking at the outliers from various angles and essentially analyzing them separately from the data set to understand the degree of relatability or impact it would have on the data set. To address the numerical columns that were categories I simply mapped the columns whenever needed such that they were no longer numerical values but the appropriate categories/ strings that they represented.

For further pre-processing it would be interesting to see how keeping the casual and windspeed outliers will change the shape of the data and any discoveries made. In terms of continuing the

analysis it would be interesting to see determine the possible relationship for spikes in casual users to other attributes like holidays or time of year.

APPENDIX A



Number of Registered vs. Casual Bike Users from 2011-2012

