
0.0.1 Question 0a

What is the granularity of the data (i.e. what does each row represent)?

Each row in the dataset is the record of number of riders (either casual or registered or both) on a speficfic hour of a specific day along with the information about the weather, tempurature, humidity, and windspeed. The data is not too detailed (or granular).

0.0.2 Question 0b

For this assignment, we'll be using this data to study bike usage in Washington D.C. Based on the granularity and the variables present in the data, what might some limitations of using this data be? What are two additional data categories/variables that one could collect to address some of these limitations?

The data doesn't have detailed information about a specific rider, so we can only study the total number of rentals, and might miss some charactereistics of each individual. Some suggestions are information about locations where each bike is picked up and dropped off, and how long each rider spends on biking.

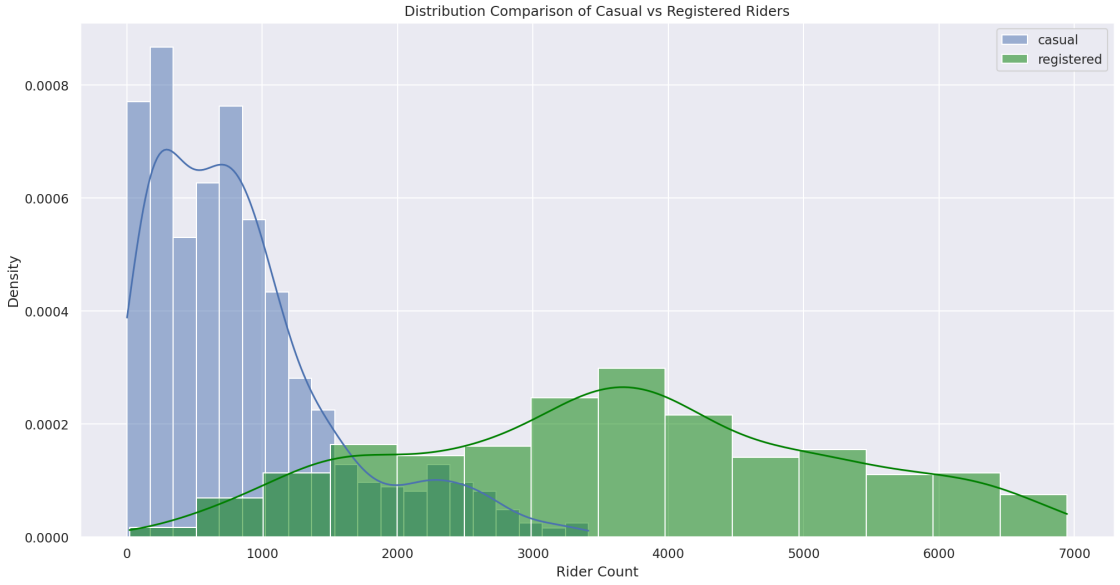
0.0.3 Question 2a

Use the `sns.histplot`(documentation) function to create a plot that overlays the distribution of the daily counts of bike users, using blue to represent `casual` riders, and green to represent `registered` riders. The temporal granularity of the records should be daily counts, which you should have after completing question 1c.

Hint: You will need to set the `stat` parameter appropriately to match the desired plot, and may call `sns.histplot` more than one time.

Include a `legend`, `xlabel`, `ylabel`, and `title`. Read the seaborn plotting tutorial, if you're not sure how to add these. After creating the plot, look at it and make sure you understand what the plot is actually telling us, e.g on a given day, the most likely number of registered riders we expect is ~4000, but it could be anywhere from nearly 0 to 7000.

```
In [18]: sns.histplot(data= daily_counts, x='casual', stat= 'density', kde= True, label='casual')
         sns.histplot(data= daily_counts, x='registered',color='green',stat= 'density', kde= True, label
         plt.legend()
         plt.title('Distribution Comparison of Casual vs Registered Riders')
         plt.xlabel('Rider Count')
         plt.ylabel('Density');
```



0.0.4 Question 2b

In the cell below, descibe the differences you notice between the density curves for casual and registered riders. Consider concepts such as modes, symmetry, skewness, tails, gaps and outliers. Include a comment on the spread of the distributions.

Overall, the two distributions' density curves are different from each other. Modes: Casual Riders distribution has two modes (multimode), one around 400 and one around 800; while Registered Riders distribution has one mode around 3570. Symmetry/ Skewness: Casual Riders distribution is skewed to to right where its tail (outliers located) is after 1500 number of riders. On the other hand, Registered Riders distribution is symmetric (has a normal shape) Spread: Registered Riders distribution has a wider variability that range from 0 to 7000. Casual Riders distribution is less spread out with most of the data if from 0-1500.

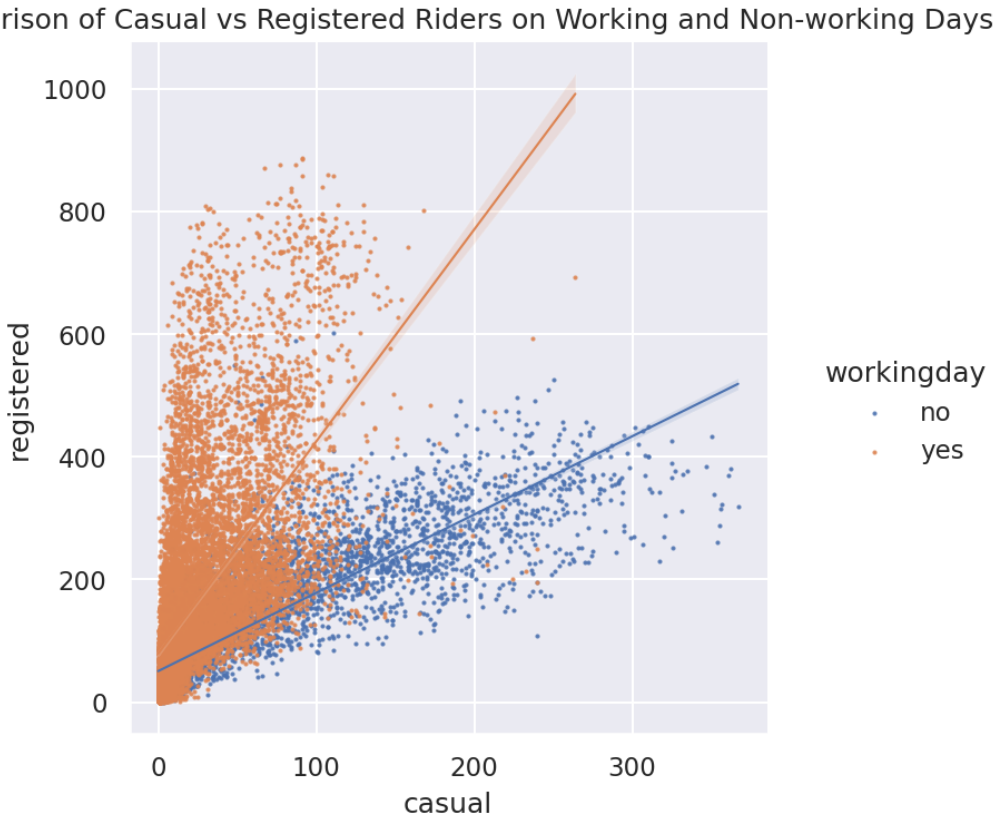
0.0.5 Question 2c

The density plots do not show us how the counts for registered and casual riders vary together. Use `sns.lmplot` (documentation) to make a scatter plot to investigate the relationship between casual and registered counts. This time, let's use the `bike DataFrame` to plot hourly counts instead of daily counts.

The `lmplot` function will also try to draw a linear regression line (just as you saw in Data 8). Color the points in the scatterplot according to whether or not the day is a working day (your colors do not have to match ours exactly, but they should be different based on whether the day is a working day).

Hints: * Checkout this helpful tutorial on `lmplot`. * There are many points in the scatter plot, so make them small to help reduce overplotting. Check out the `scatter_kws` parameter of `lmplot`. * Generate and plot the linear regression line by setting a **paramter** of `lmplot` to `True`. Can you find this in the documentation? We will discuss what is linear regression is more details later. * You can set the `height` parameter if you want to adjust the size of the `lmplot`. * Add a descriptive title and axis labels for your plot.

```
In [19]: # Make the font size a bit bigger
         sns.set(font_scale=1)
         sns.lmplot(data=bike, x='casual', y='registered', hue='workingday', scatter= True,scatter_kws=
         plt.title('Comparison of Casual vs Registered Riders on Working and Non-working Days')
         plt.xlabel('casual')
         plt.ylabel('registered');
```


Comparison of Casual vs Registered Riders on Working and Non-working Days
registered
casual
workingday
no
yes
0
200
400
600
800
1000
0
100
200
300

0.0.6 Question 2d

What does this scatterplot seem to reveal about the relationship (if any) between casual and registered riders and whether or not the day is on the weekend? What effect does overplotting have on your ability to describe this relationship?

Casual and registered riders has a stronger linear relationship on non-working days than on working days. However, one thing I notice that on working day, casual riders rarely use bikes (otherwise, they will just use subcription to save money since they use the service consistently), and both registered and casual riders use the service on non-working days. This explains why they have a linear relationship only on non-working days. In addition, overplotting will potentially cause me to overlook some number of riders on non-working day on the range from 0-100 because the workingday data overlaps non-working day data.

0.0.7 Question 3a (Bivariate Kernel Density Plot)

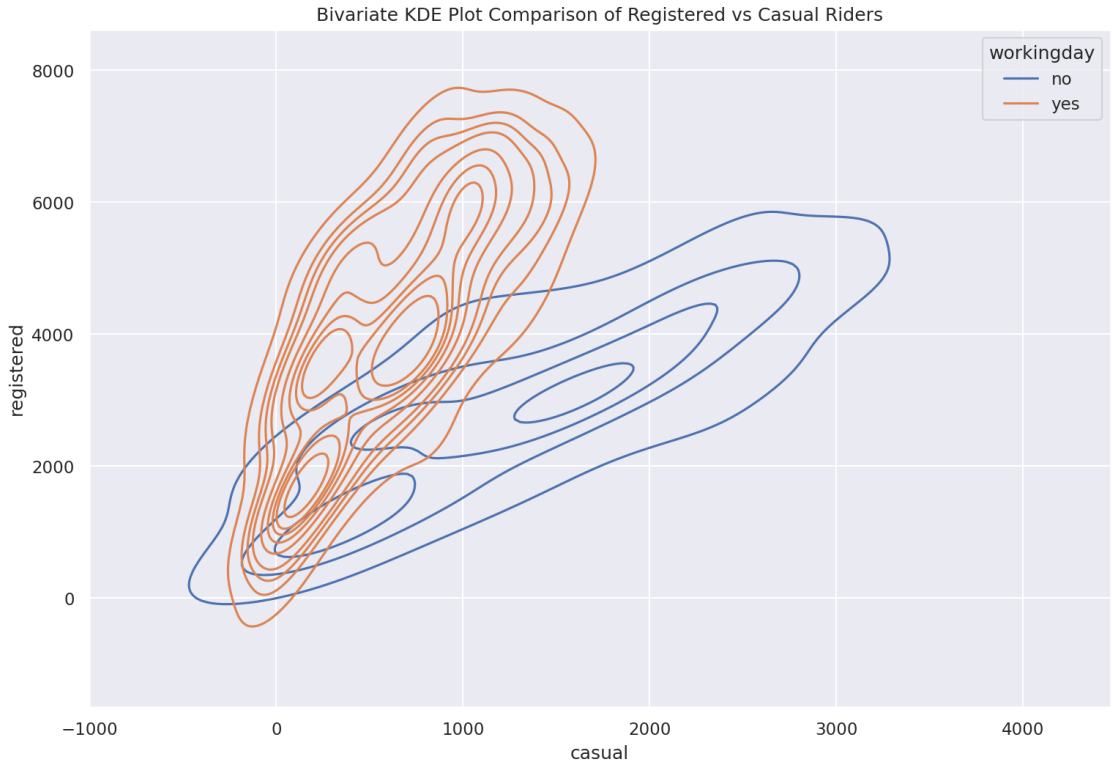
Generating a bivariate kernel density plot with workday and non-workday separated.

Hints: You only need to call `sns.kdeplot` once. Take a look at the `hue` paramter and adjust other inputs as needed.

After you get your plot working, experiment by setting `shade=True` in `kdeplot` to see the difference between the shaded and unshaded version. Please submit your work with `shade=False`.

In [21]:
```
# Set the figure size for the plot
plt.figure(figsize=(12,8))

sns.kdeplot(x=daily_counts['casual'], y=daily_counts['registered'], hue=daily_counts['workingda
plt.title('Bivariate KDE Plot Comparison of Registered vs Casual Riders');
```



0.0.8 Question 3b

With some modification to your 3a code (this modification is not in scope), we can obatined the plot above. In your own words, describe what the lines and the color shades of the lines signify about the data. What does each line and color represent?

The lines are kernal density estimators for each area (x,y) in this two dimension graph. They represent clusters of data between two variables. The color represents how dense each cluster is. In other words, the darker the color is, the more data they are in the cluster.

0.0.9 Question 3c

What additional details can you identify from this contour plot that were difficult to determine from the scatter plot?

This contour plot reveals the shape as well as the density of each cluster of data. We can see where and how much the data from each category lying on the graph more clearly than from the scatter plot where data are stepping on each other.

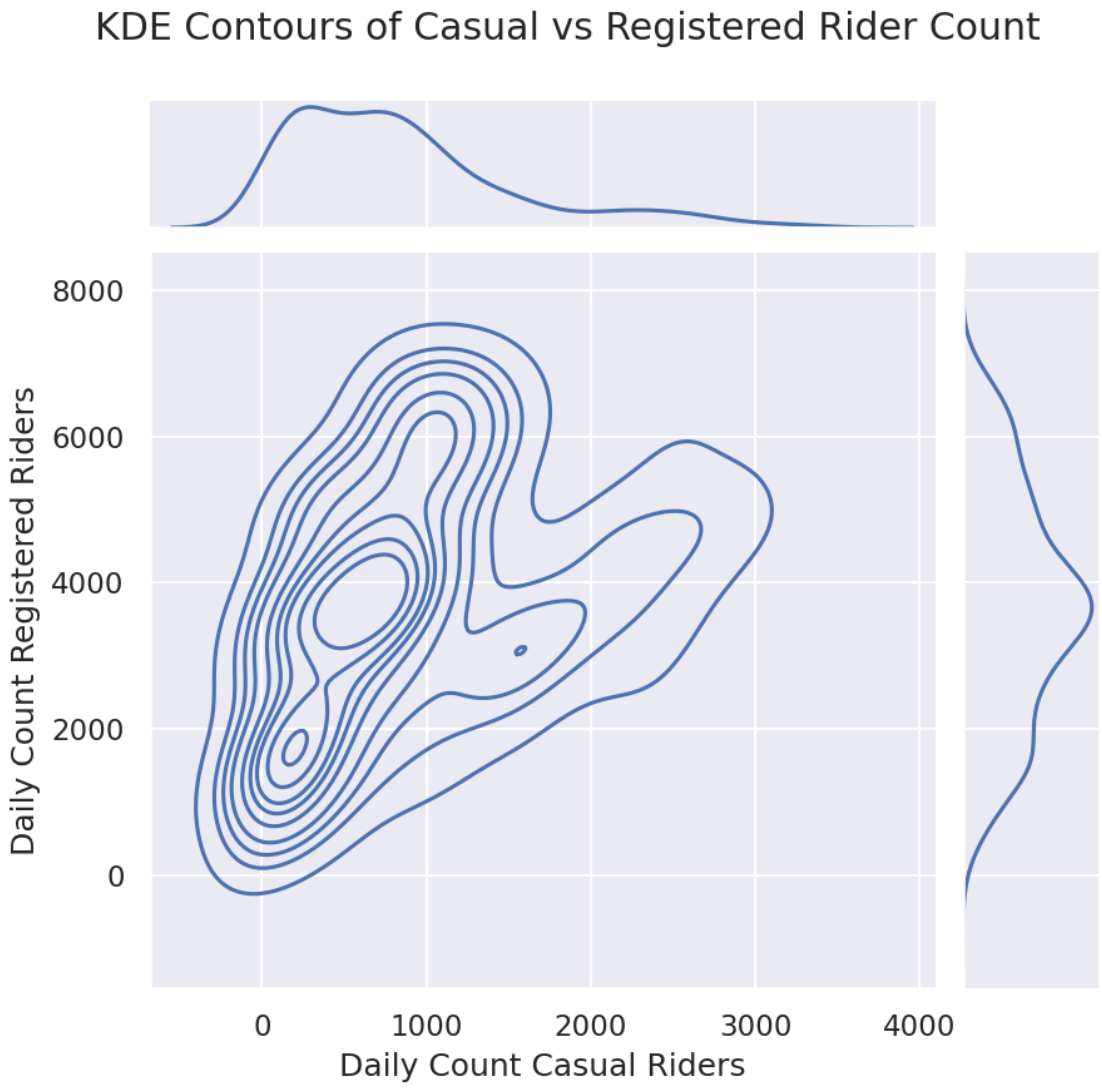
0.1 4: Joint Plot

As an alternative approach to visualizing the data, construct the following set of three plots where the main plot shows the contours of the kernel density estimate of daily counts for registered and casual riders plotted together, and the two "margin" plots (at the top and right of the figure) provide the univariate kernel density estimate of each of these variables. Note that this plot makes it harder see the linear relationships between casual and registered for the two different conditions (weekday vs. weekend).

Hints: * The seaborn plotting tutorial has examples that may be helpful. * Take a look at `sns.jointplot` and its `kind` parameter. * `set_axis_labels` can be used to rename axes on the contour plot.

Note: * At the end of the cell, we called `plt.suptitle` to set a custom location for the title. * We also called `plt.subplots_adjust(top=0.9)` in case your title overlaps with your plot.

```
In [22]: ax = sns.jointplot(data=daily_counts, x='casual', y= 'registered', kind= 'kde')
         plt.suptitle("KDE Contours of Casual vs Registered Rider Count")
         plt.subplots_adjust(top=0.9);
         ax.set_axis_labels('Daily Count Casual Riders', 'Daily Count Registered Riders');
```


KDE Contours of Casual vs Registered Rider Count
Daily Count Registered Riders
8000
6000
4000
2000
0
0
1000
2000
3000
4000
Daily Count Casual Riders

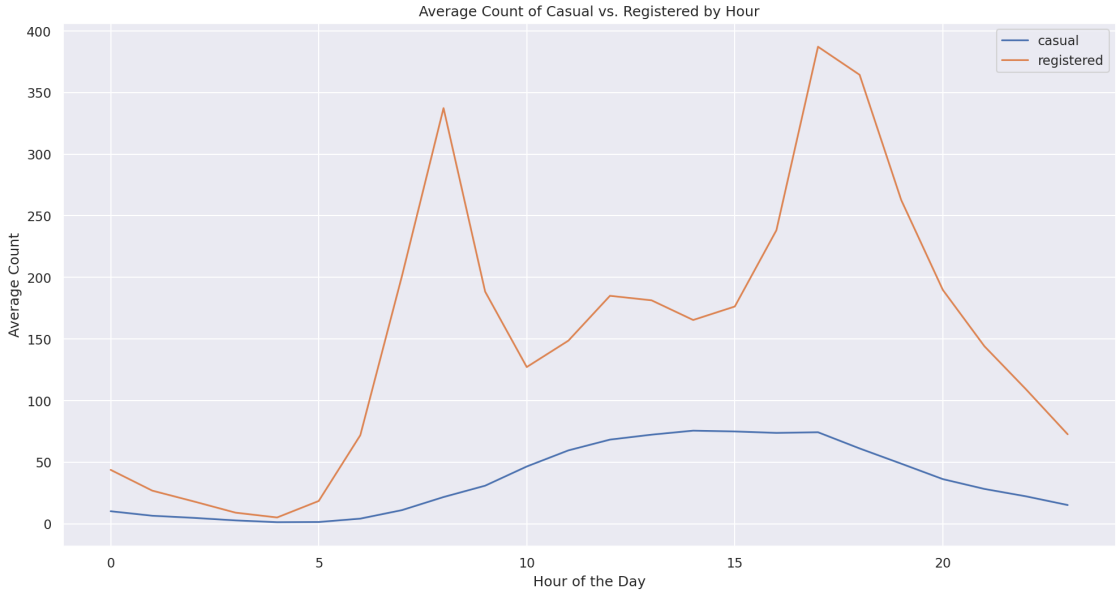
0.2 5: Understanding Daily Patterns

0.2.1 Question 5a

Let's examine the behavior of riders by plotting the average number of riders for each hour of the day over the **entire dataset**, stratified by rider type.

Your plot should look like the plot below. While we don't expect your plot's colors to match ours exactly, your plot should have legend in the plot and different colored lines for different kinds of riders.

```
In [23]: av_count = bike[['hr', 'casual','registered']].groupby('hr').mean()
         sns.lineplot(data= av_count, x=av_count.index, y='casual', label='casual')
         sns.lineplot(data= av_count, x=av_count.index, y='registered', label='registered')
         plt.title('Average Count of Casual vs. Registered by Hour')
         plt.xlabel('Hour of the Day')
         plt.ylabel('Average Count');
```



0.2.2 Question 5b

What can you observe from the plot? Discuss your obseravtion and hypothesize about the meaning of the peaks in the registered riders' distribution.

Overall, the average count of casual riders throughout the day is much lower than registered riders, but both of them have a similar pattern where most of active riders happen between hour 6 to hour 20. The registered riders' distribution has two highest peaks at hour 8 and hour 18. One hypothesis for this is that this is the time where people go to and get off work everyday (9-5 jobs specifically).

0.2.3 Question 6b

In our case with the bike ridership data, we want 7 curves, one for each day of the week. The x-axis will be the temperature and the y-axis will be a smoothed version of the proportion of casual riders.

You should use `statsmodels.nonparametric.smoothers_lowess.lowess` just like the example above. Unlike the example above, plot ONLY the lowess curve. Do not plot the actual data, which would result in overplotting. For this problem, the simplest way is to use a loop.

You do not need to match the colors on our sample plot as long as the colors in your plot make it easy to distinguish which day they represent.

Hints: * Start by just plotting only one day of the week to make sure you can do that first.

- The `lowess` function expects y coordinate first, then x coordinate. You should also set the `return_sorted` field to `False`.
- Look at the top of this homework notebook for a description of the (normalized) temperature field to know how to convert to Fahrenheit. By default, the temperature field ranges from 0.0 to 1.0. In case you need it, Fahrenheit = Celsius $\times \frac{9}{5} + 32$.

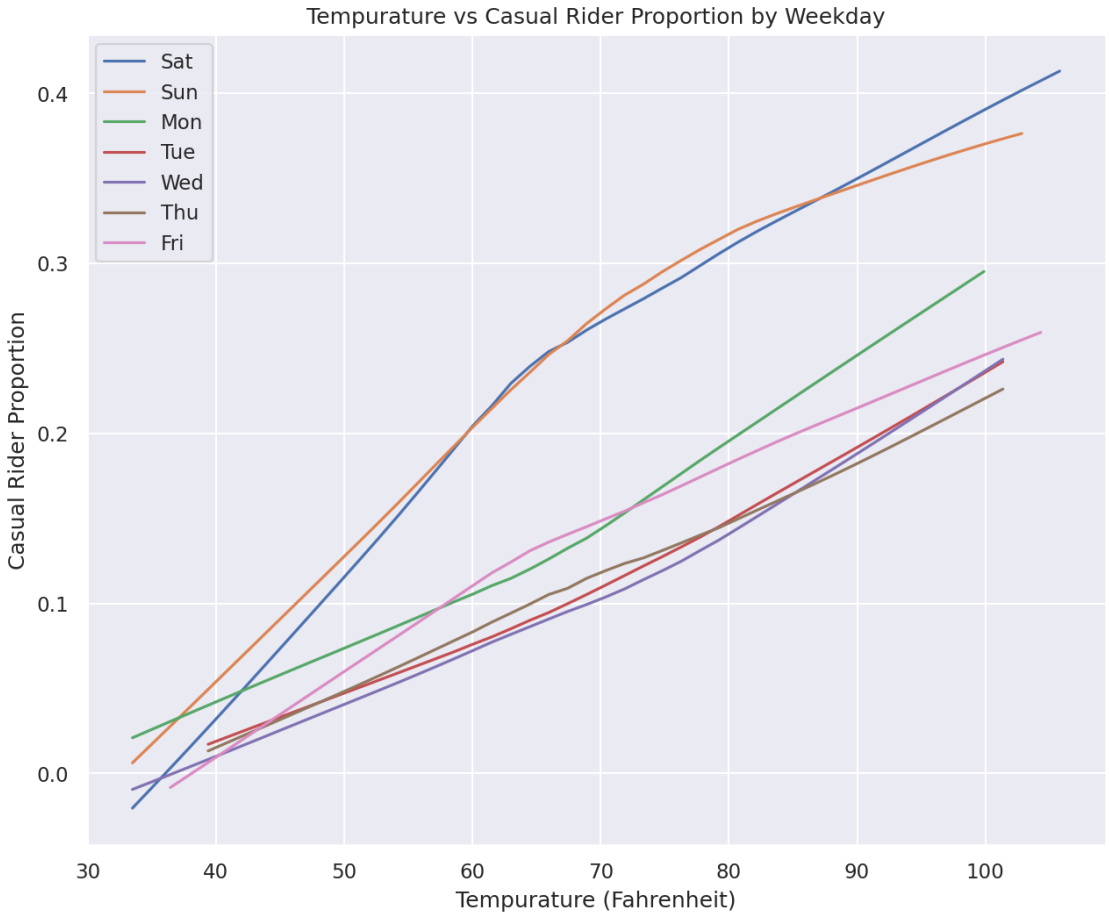
Note: If you prefer plotting temperatures in Celsius, that's fine as well!

```
In [29]: from statsmodels.nonparametric.smoothers_lowess import lowess

         plt.figure(figsize=(10,8))

         for day in bike['weekday'].unique():
             week_day = bike[bike['weekday'] == day].copy()
             week_day['temp'] = week_day['temp']*41*9/5+32
             ysmooth = lowess(week_day['prop_casual'], week_day['temp'], return_sorted= False)
             sns.lineplot(week_day['temp'], ysmooth, label = day)

         plt.title('Tempurature vs Casual Rider Proportion by Weekday')
         plt.xlabel('Tempurature (Fahrenheit)')
         plt.ylabel('Casual Rider Proportion')
         plt.legend();
```


Tempurature vs Casual Rider Proportion by Weekday
Sat
Sun
Mon
Tue
Wed
Thu
Fri
Casual Rider Proportion
0.0
0.1
0.2
0.3
0.4
30
40
50
60
70
80
90
100
Tempurature (Fahrenheit)

0.2.4 Question 6c

What do you see from the curve plot? How is `prop_casual` changing as a function of temperature? Do you notice anything else interesting?

All 7 curves have a positive slope where the the higher the tempurature is, the higher the casual proportion becomes. This makes sense because the higher temperature indicates that the day is sunny, and nobody wants to bike on cold days. Additionally, the two Sat and Sun curves are higher than the rest of curves. This is also understandable because people tend to go outside on weekends.

0.2.5 Question 7a

Imagine you are working for a Bike Sharing Company that collaborates with city planners, transportation agencies, and policy makers in order to implement bike sharing in a city. These stakeholders would like to reduce congestion and lower transportation costs. They also want to ensure the bike sharing program is implemented equitably. In this sense, equity is a social value that is informing the deployment and assessment of your bike sharing technology.

Equity in transportation includes: improving the ability of people of different socio-economic classes, genders, races, and neighborhoods to access and afford the transportation services, and assessing how inclusive transportation systems are over time.

Do you think the `bike` data as it is can help you assess equity? If so, please explain. If not, how would you change the dataset? You may discuss how you would change the granularity, what other kinds of variables you'd introduce to it, or anything else that might help you answer this question.

I don't think the bike data can help me study equity. I would want to add more variables related to each rider in terms of their demographic, socio-economic classes, and their locations where they usually use the service. One idea I can think of is to visualize on the map along where they usually pick up and drop off the bike, what kind of neighborhood of popular locations, etc.

0.2.6 Question 7b

Bike sharing is growing in popularity and new cities and regions are making efforts to implement bike sharing systems that complement their other transportation offerings. The goals of these efforts are to have bike sharing serve as an alternate form of transportation in order to alleviate congestion, provide geographic connectivity, reduce carbon emissions, and promote inclusion among communities.

Bike sharing systems have spread to many cities across the country. The company you work for asks you to determine the feasibility of expanding bike sharing to additional cities of the U.S.

Based on your plots in this assignment, what would you recommend and why? Please list at least two reasons why, and mention which plot(s) you drew you analysis from.

Note: There isn't a set right or wrong answer for this question, feel free to come up with your own conclusions based on evidence from your plots!

I'm using the plot 'Average Count of Casuak vs. Registered by Hour'. I would reccommend not to expand this system to other cities unless it is a big city such as San Francisco or Berkeley. Base on the graph below, we can see there is a huge gap between casual and registered riders where registered riders make up a large amount. When investigate closely, the highest peak of count of registered riders are around 8AM and 6PM, which indicates most of people use this biking system to go to and get off work. Places like Washington D.C. are biking friendly where buildings are close to each other, and it is much convenient to bike to work than using a car. This is the pattern of our target users, and it can not be applied to rural areas where people tend to use their car to travel to work.