0.0.1 Question 0

Question 0A What is the granularity of the data (i.e. what does each row represent)?

Each row represents bike sharing used during a day as well as it accounts the acount of users (renters) and the record of climate during that day.

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 you can collect to address some of these limitations?

The location where people in DC are more interested to rent bikes perhaps due to high traffic. The duration of the rent. For instance, how many hours bike A was used during an specific day

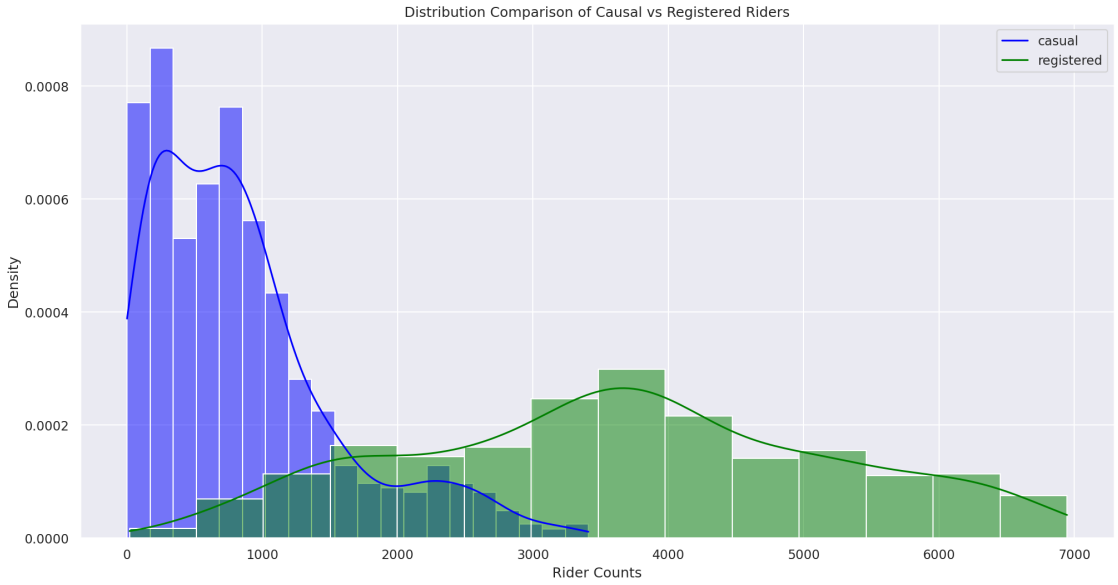
0.0.2 Question 2

Question 2a Use the `sns.histplot` 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.

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 [16]: sns.histplot(data=daily_counts, x="casual", kde=True, stat="density", color="blue")
         sns.histplot(data=daily_counts, x="registered", kde=True, stat="density", color="green")
         plt.title("Distribution Comparison of Causal vs Registered Riders")
         plt.xlabel("Rider Counts")
         plt.ylabel("Density")
         plt.legend(["casual", "registered"]);
```



0.0.3 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.

Casual: The histogram is skewed to the right with a peak at 900-1000 where I believe the mean is. The distribution is not normal and it's reasonable to say casual riders use bike sharing not on routine basis as thier counterparts, the registered users. Unlike casual riders the distribution of the histogram of registered users is symmetric and has a peak at 3500-4000, but the distribution is widely spreaded compared to the casual riders. The registred distribution of daily counts has larger SD compared to the causal, perhaps in denser areas and higher traffic people use bike sharing, but there are also areas where people barely do bike sharing which is a big gap and it's obvious more clearly in the spreadness of the distribution.

0.0.4 Question 2c

The density plots do not show us how the counts for registered and casual riders vary together. Use `sns.lmplot` 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).

There are many points in the scatter plot, so make them small to help reduce overplotting. Also make sure to set `fit_reg=True` to generate the linear regression line. You can set the `height` parameter if you want to adjust the size of the `lmplot`.

Hints: * Checkout this helpful tutorial on `lmplot`.

- You will need to set `x`, `y`, and `hue` and the `scatter_kws` in the `sns.lmplot` call.
- You will need to call `plt.title` to add a title for the graph.

```
In [17]: # Make the font size a bit bigger
         sns.set(font_scale=1)
         sns.lmplot(data=bike, x="casual", y="registered", hue="workingday",
                    fit_reg=True, size=12);
         plt.xlabel("Casual")
         plt.ylabel("Registered")
         plt.title("Camparison of Casual vs Registered Riders on Working and Non-working Days");
```


Camparison of Casual vs Registered Riders on Working and Non-working Days
Registered
Casual
workingday
no
yes
0
200
400
600
800
1000
0
50
100
150
200
250
300
350

0.0.5 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?

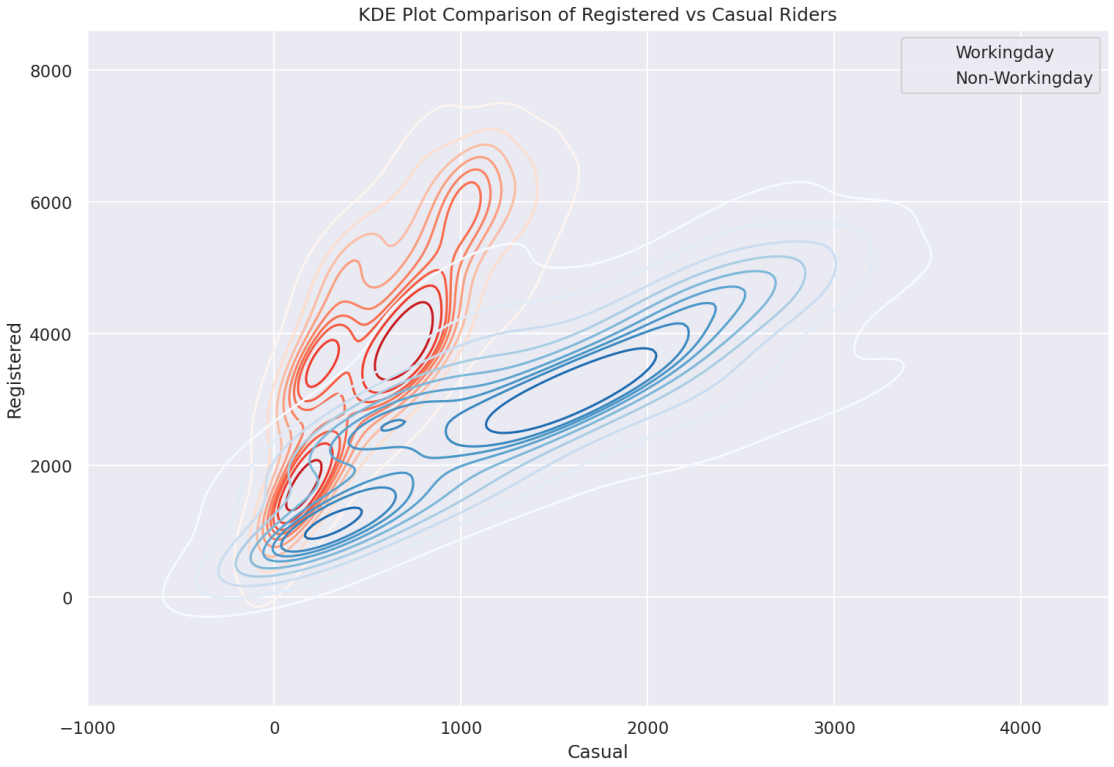
There is a linear relationship between casual and registered riders in both working and non-working days. The correlation is positive, however, the slope of the working days is higher than non-working days. But due to the overlap of huge data points, it's hard to interprete and describe this relationship in a professional way.

Generating the plot with weekend and weekday separated can be complicated so we will provide a walk-through below, feel free to use whatever method you wish if you do not want to follow the walkthrough.

Hints: * You can use `loc` with a boolean array and column names at the same time * You will need to call kdeplot twice, each time drawing different data from the `daily_counts` table. * Check out this guide to see an example of how to create a legend. In particular, look at how the example in the guide makes use of the `label` argument in the call to `plt.plot()` and what the `plt.legend()` call does. This is a good exercise to learn how to use examples to get the look you want. * You will want to set the `cmap` parameter of `kdeplot` to `"Reds"` and `"Blues"` (or whatever two contrasting colors you'd like), and also set the `label` parameter to address which type of day you want to plot. You are required for this question to use two sets of contrasting colors for your plots.

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 [19]:
```
# Set the figure size for the plot
plt.figure(figsize=(12,8))

# Set 'is_workingday' to a boolean array that is true for all working_days
is_workingday = daily_counts["workingday"]=="yes"

# Bivariate KDEs require two data inputs.
# In this case, we will need the daily counts for casual and registered riders on workdays
# Hint: consider using the .loc method here.
casual_workday = daily_counts.loc[is_workingday, "casual"]
registered_workday = daily_counts.loc[is_workingday, "registered"]

# Use sns.kdeplot on the two variables above to plot the bivariate KDE for weekday rides
sns.kdeplot(casual_workday, registered_workday, cmap="Reds")

not_workingday = daily_counts["workingday"]=="no"
# Repeat the same steps above but for rows corresponding to non-workingdays
# Hint: Again, consider using the .loc method here.
casual_non_workday = daily_counts.loc[not_workingday, "casual"]
registered_non_workday = daily_counts.loc[not_workingday, "registered"]

# Use sns.kdeplot on the two variables above to plot the bivariate KDE for non-workingday ride
sns.kdeplot(casual_non_workday, registered_non_workday, cmap="Blues")

plt.legend(["Workingday", "Non-Workingday"])
plt.xlabel("Casual")
plt.ylabel("Registered")
plt.title("KDE Plot Comparison of Registered vs Casual Riders");
```


KDE Plot Comparison of Registered vs Casual Riders
Workingday
Non-Workingday
Registered
Casual
8000
6000
4000
2000
0
−1000
0
1000
2000
3000
4000

Question 3bi In your own words, describe what the lines and the color shades of the lines signify about the data.

Each line represents an interval and the areas where the lines are close to each other indicats that most of the records (data points) are located around that region.

Question 3bii What additional details can you identify from this contour plot that were difficult to determine from the scatter plot?

Compared to contour plot, bivariate kde plot is more clear in terms of interpreting and describing linear relationship between casual and registered riders given the condition of working vs non-working days. Another benefit of using bivariate plots is that we don't see any overlapping of data points which is really helpful in determining corelations may exists between two distributions.

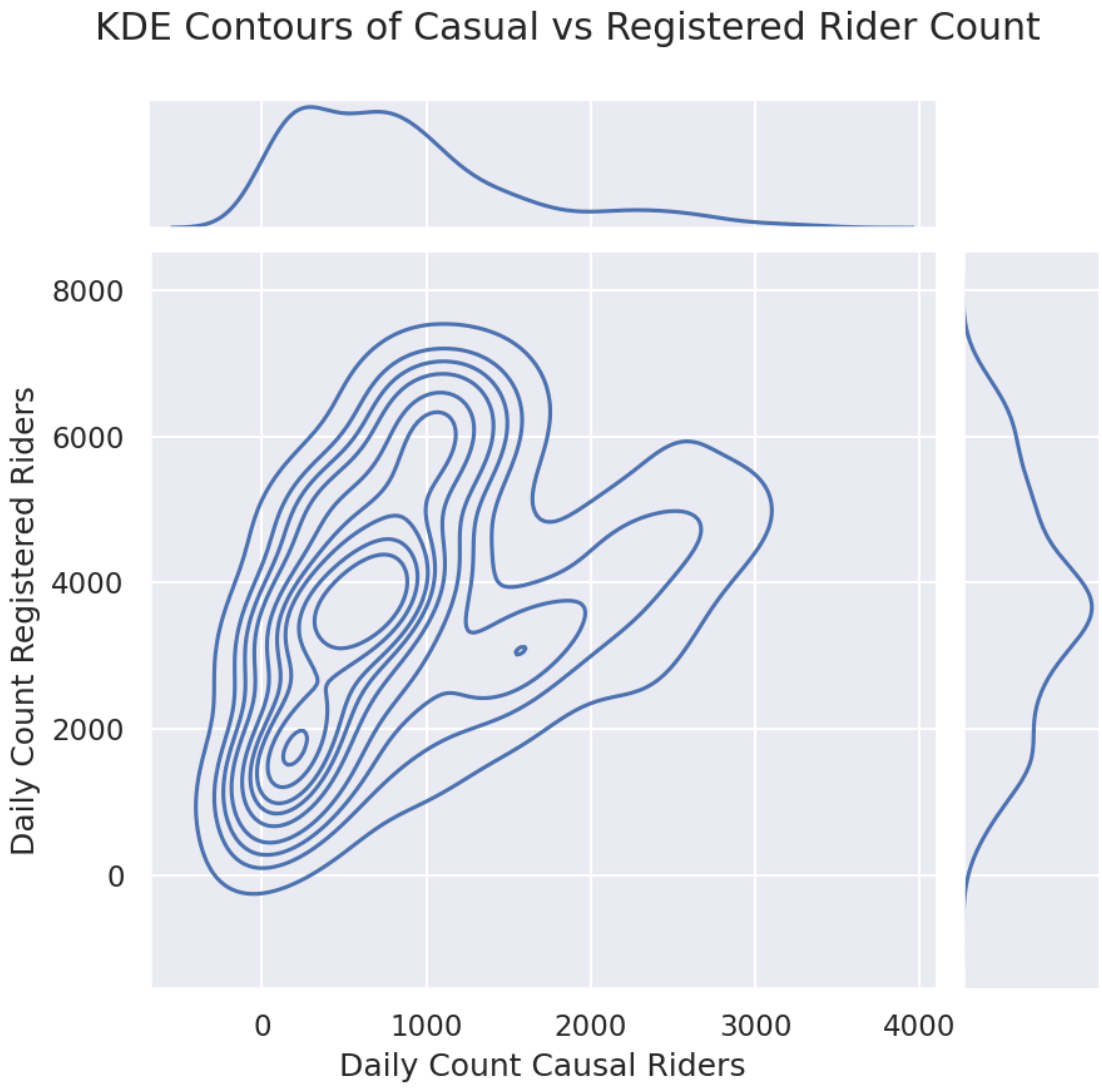
0.1 4: Joint Plot

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


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

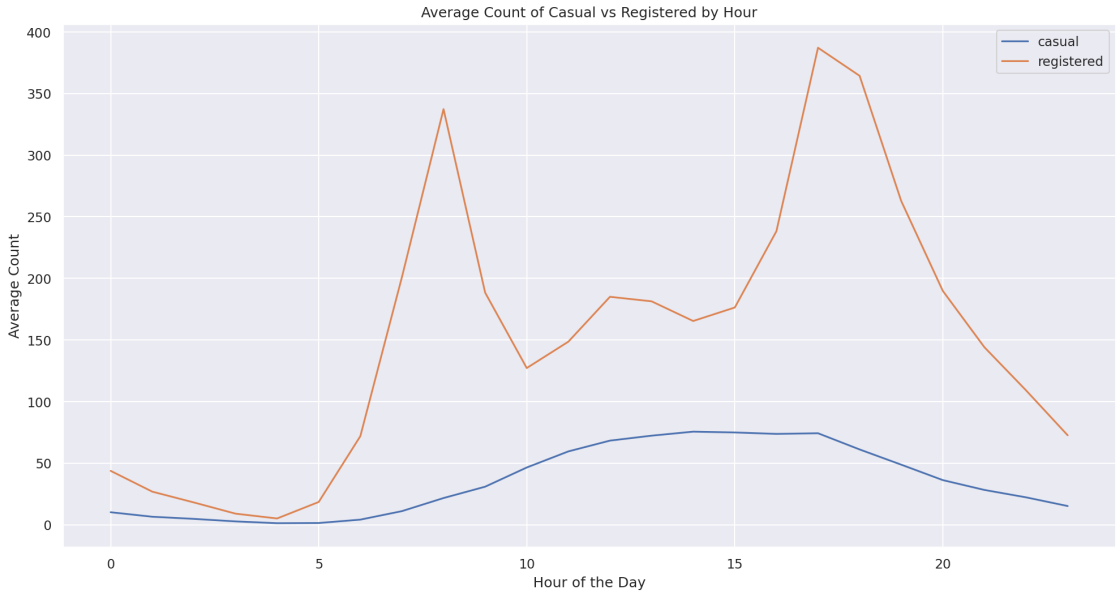
0.2 5: Understanding Daily Patterns

0.2.1 Question 5

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 different colored lines for different kinds of riders.

```
In [21]: mean = bike.groupby("hr").mean()[["casual", "registered"]]
         sns.lineplot(x=mean.index, y=mean["casual"], label="casual")
         sns.lineplot(x=mean.index, y=mean["registered"], label="registered")
         plt.xlabel("Hour of the Day")
         plt.ylabel("Average Count")
         plt.title("Average Count of Casual vs Registered by Hour");
```



Question 5b What can you observe from the plot? Hypothesize about the meaning of the peaks in the registered riders' distribution.

The plot indicats that both registered and casual riders ride more during the day. However, registered riders use rides more often compared to casual riders. I assume the two spikes of registered riders would be an evidance that they ride to commute between work and home because the left spike is between 5-10am while the left spike is between 3-20pm and a sharp decline during launch hours. Unlike registered riders, casual riders rides more between launch time that's maybe because those casual riders are students or unemployeed at those hours are business hours.

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.

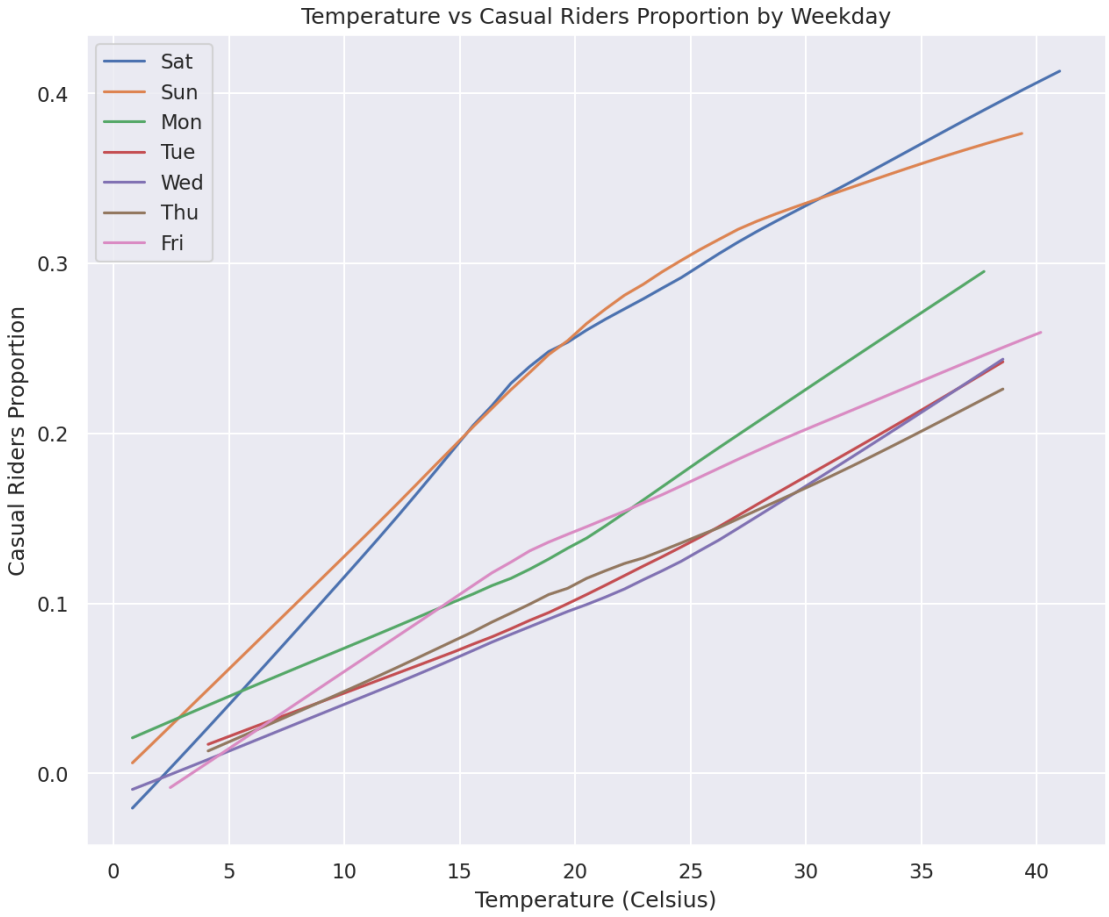
You do not need to match the colors on our sample plot as long as the colors in your plot make it easy to distinguish which day they represent.

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 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 $* \frac{9}{5} + 32$.

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

In [36]:
```
from statsmodels.nonparametric.smoothers_lowess import lowess

plt.figure(figsize=(10,8))
for day in bike["weekday"].unique():
    fil_bike = bike[bike["weekday"]==day]
    x_temp = fil_bike["temp"]
    y_cas = fil_bike["prop_casual"]
    ysmooth = lowess(y_cas, x_temp, return_sorted=False)
    sns.lineplot((x_temp*41), ysmooth, label=day)
plt.xlabel("Temperature (Celsius)")
plt.ylabel("Casual Riders Proportion")
plt.title("Temperature vs Casual Riders Proportion by Weekday")
plt.legend();
```


Temperature vs Casual Riders Proportion by Weekday
Sat
Sun
Mon
Tue
Wed
Thu
Fri
Casual Riders Proportion
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
30
35
40
Temperature (Celsius)

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?

There is one thing clear in the plot at the first glance that DC is mostly cold, so people don't use bikes when it's cold. This is why there is a strong evidance that the temperature and the proportion of casual rider has a strong correlation. On top of that, unlike weekdays (working days), on weekends people use more rides, so there is a big gap between weekend and weekdays ride sharing. Overall, when the temperature is over 5 or so, the riders begin riding and as the temperature gets higher the proportion of casual rides increases.

0.2.2 Question 7

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.

Bike data has limited information to be able to assess equity. For instance, some basice non-private information about the rider is not included in the data like, gender, age, zipcode, employment status, which would have helped me assess if the data for equity. I would include at least the listed information in the data in order to ensure equity.

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 would like to address two plots that convinced me that bike sharing is one of best choices to complement other transportation mean. The first plot is the "Average Count of Casual vs Registered by Hour." This plot is an evidance that people prefer to use bike rather than other transportation to commute to work. The second plot is "Temperature vs Casual Riders Proportion by Weekday" which is also an evidance that people ride more in warm whether. So, based on these evidance I strongly suggest that bike sharing should expand in additional cities but