
0.0.1 Question 0a

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

Each record represents the data at each hour of a day.

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?

Collecting the bike counts at every hour may be collecting too much data. We can address this by binning the hours to create wider groups as a new variable, like a bin for every 2 or 3 hours. We can also add a categorical variable that bins the hours by morning, afternoon, evening, night. Another issue is the purpose for biking and the locations where people biked.

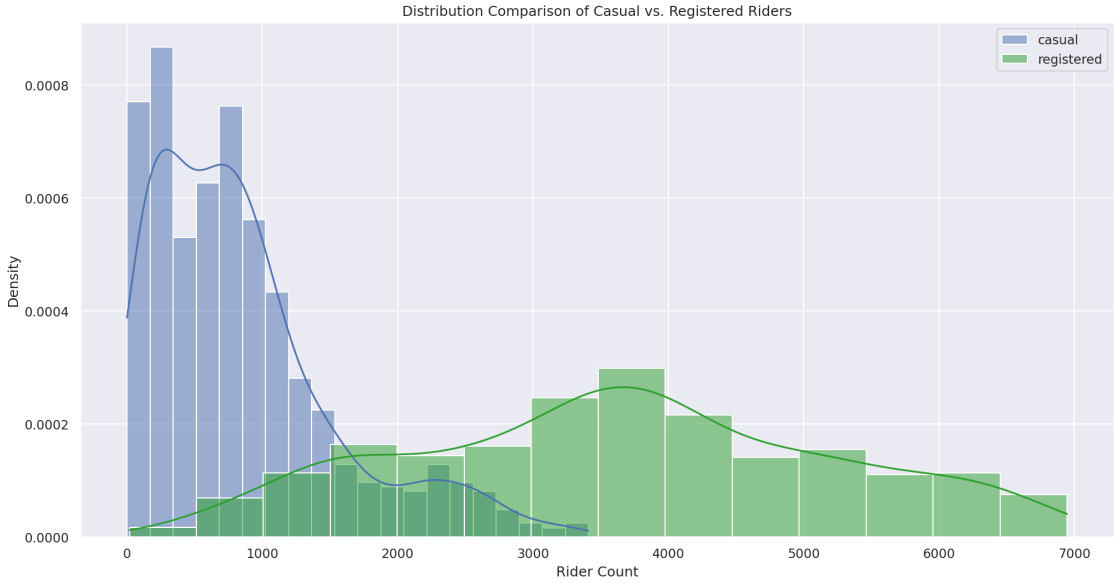
0.0.3 Question 2a

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



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.

The blue histogram for casual riders is much more centered on the left of the graph while the green histogram for registered riders is centered more to the middle right. The casual distribution is skewed right with possibly 2 modes. There are no noticeable gaps or outliers. Its data are clustered mostly between the rider counts 0 and 1000. The registered distribution is roughly unimodal and symmetric with a possible gap between 2000 and 3000. Its distribution is has one mode and no significant skewness, tails, or outliers. The spread for the registered riders is much larger than that of the casual riders

---

0.0.5 Question 2c

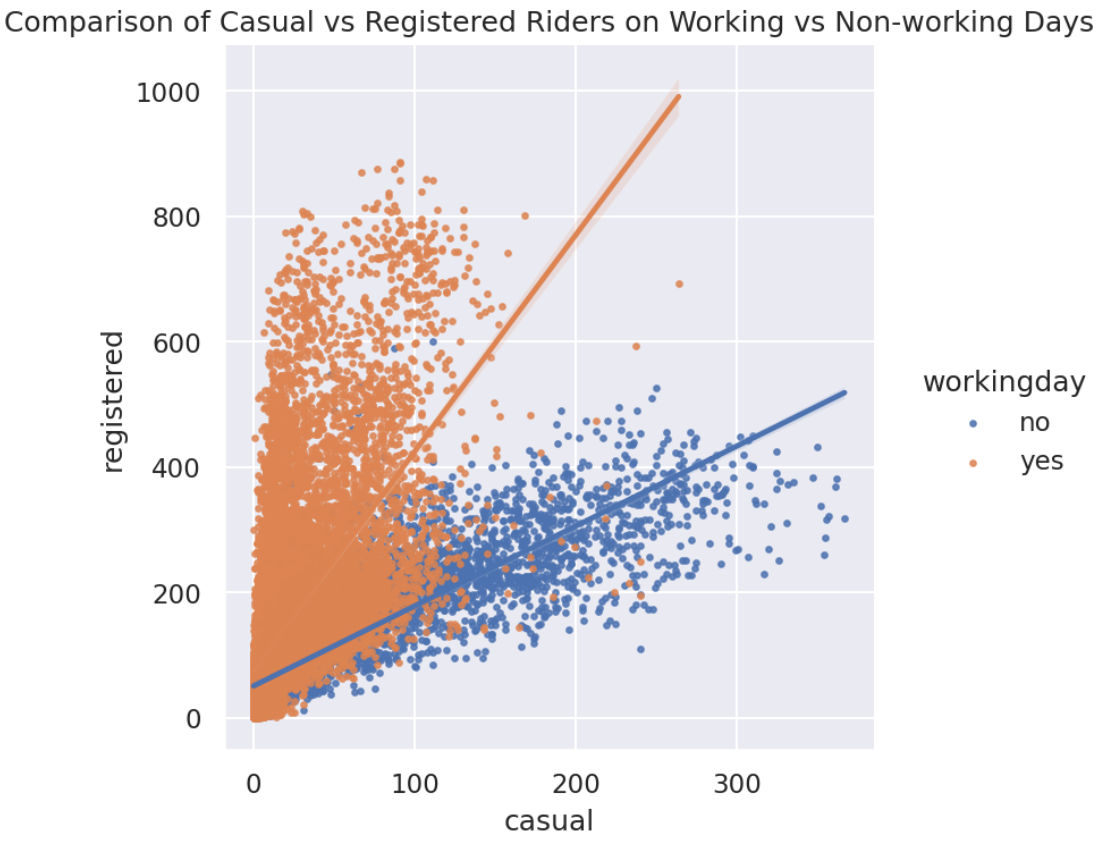
The density plots do not show us how the counts for registered and casual riders vary together. Use `sns.lmplot` (documentation) to make a scatter plot to investigate the relationship between casual and registered counts. This time, let's use the `bike DataFrame` to plot hourly counts instead of daily 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 [18]: # Make the font size a bit bigger
         sns.set(font_scale=1)
         sns.lmplot(data=bike, x='casual', y='registered', hue='workingday', scatter_kws={"s": 5}, fit_
         plt.title('Comparison of Casual vs Registered Riders on Working vs Non-working Days')
         ;
```

```
Out[18]: ''
```


Comparison of Casual vs Registered Riders on Working vs 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?

For the non-working days, there appears to be a positive linear associtation between casual and registered riders, but the association is not as clear on working days. For working days, majority of the data appears to be above the regression line, but there are quite a few outliers below it that skew the line a bit. Overplotting impacts our ability to see a general idea of the shape of the data and a general description of the relationship. In addition, the data is too dense near the origin so we cannot tell anything about clustering in more specific detail.

0.0.7 Question 3a (Bivariate Kernel Density Plot)

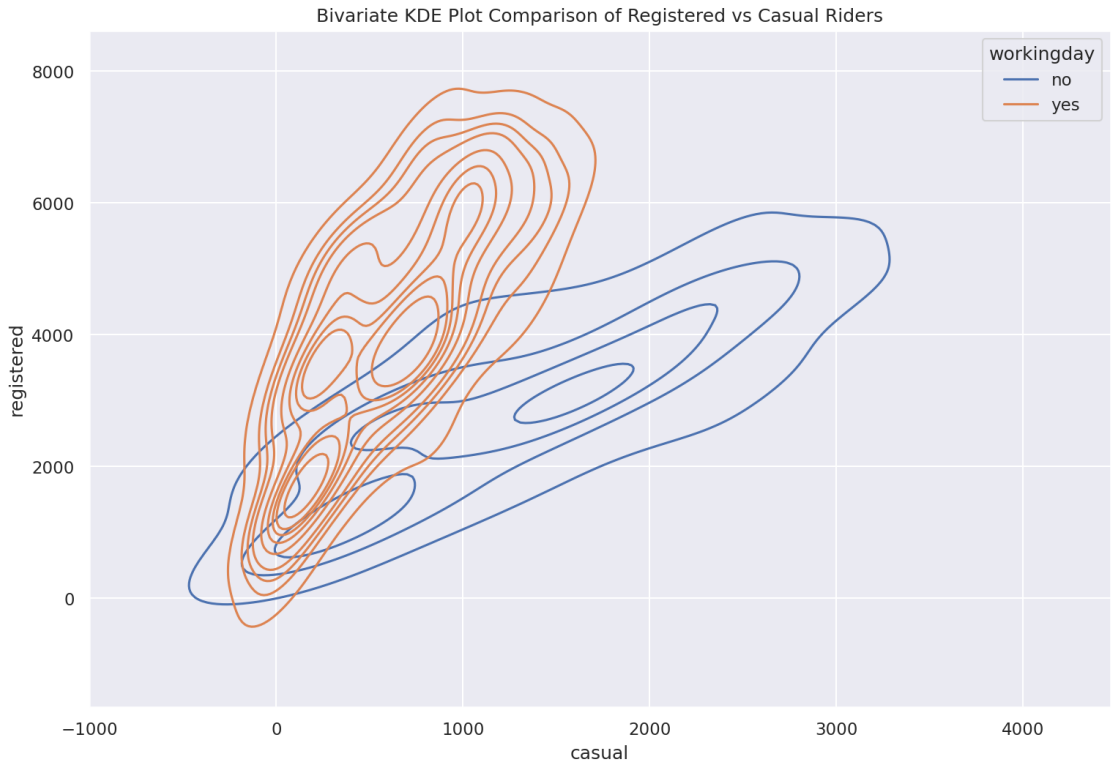
Generating a bivariate kernel density plot with workday and non-workday separated.

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 [20]: # Set the figure size for the plot
         plt.figure(figsize=(12,8))
         sns.kdeplot(data=daily_counts, x='casual', y='registered', hue='workingday')
         plt.title('Bivariate KDE Plot Comparison of Registered vs Casual Riders')
         ;
```

```
Out[20]: ''
```



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 and the color shades represent a topographical visualization of the data. Darker shades are more dense, which means there are more values in those regions than there are in lighter shared areas. The orange lines represent the association between casual and registered riders on workdays while the blue lines represent the association between casual and registered riders on non-workdays. Each space between the lines represent bins of data.

0.0.9 Question 3c

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

It is easier in the contour plot to determine clusters and how the density changes as we move around different parts of the graph. For example, we can see exactly where the densest parts of the data are because of the shading.

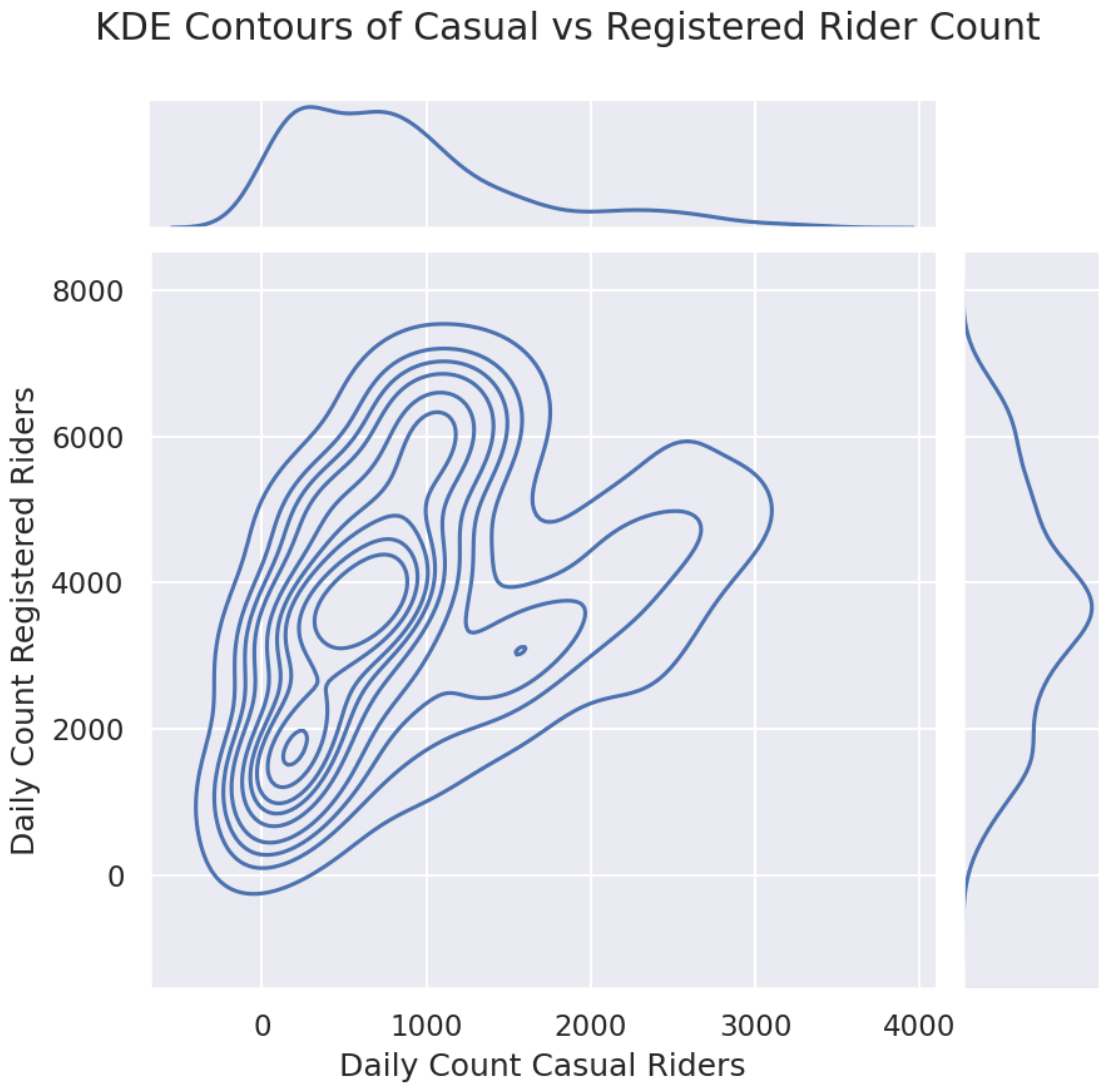
0.1 4: Joint Plot

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 [21]: sns.jointplot(x=daily_counts['casual'], y=daily_counts['registered'], kind='kde')\
             .set_axis_labels(xlabel='Daily Count Casual Riders', ylabel='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 Casual Riders

0.2 5: Understanding Daily Patterns

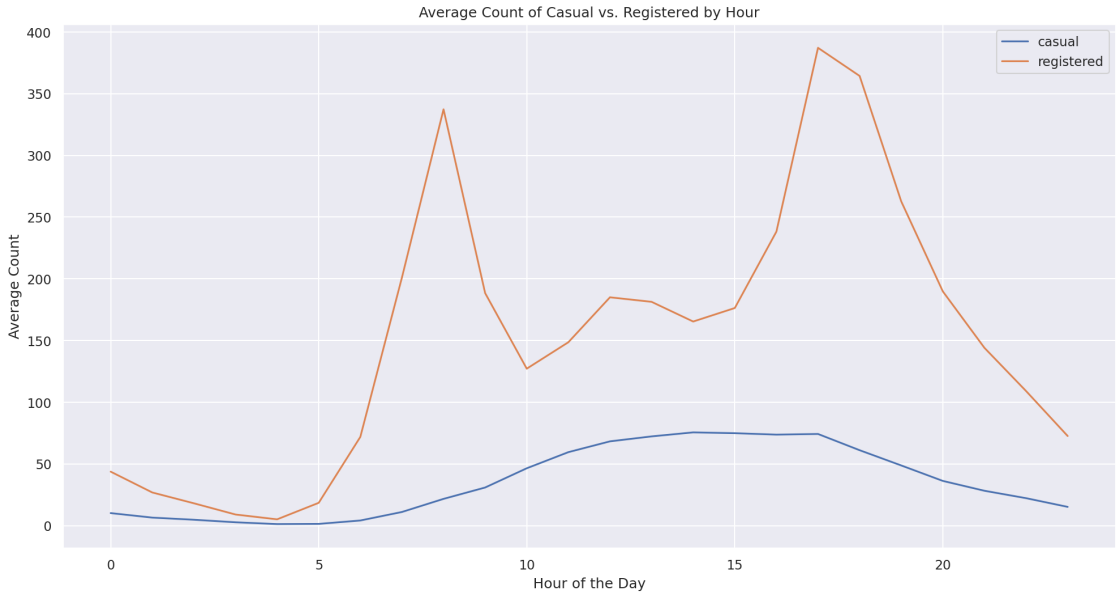
0.2.1 Question 5a

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 [22]: by_hour = bike.groupby('hr').mean()[['casual', 'registered']].reset_index()
         by_hour.head()

         sns.lineplot(data=by_hour, x='hr', y='casual')
         sns.lineplot(data=by_hour, x='hr', y='registered')
         plt.title('Average Count of Casual vs. Registered by Hour')
         plt.xlabel('Hour of the Day')
         plt.ylabel('Average Count')
         plt.legend(['casual', 'registered'])
         ;
```

```
Out[22]: ''
```



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.

For the registered bikers, there are two peaks around 8am and 5pm, which is also the time at which people are usually commuting to and from work. It makes sense that there's a valley in the registered curve during the work hours. For the casual bikers, there is a gradual peak in the afternoon/early evening, probably because people have more leisurely time in the afternoon.

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.

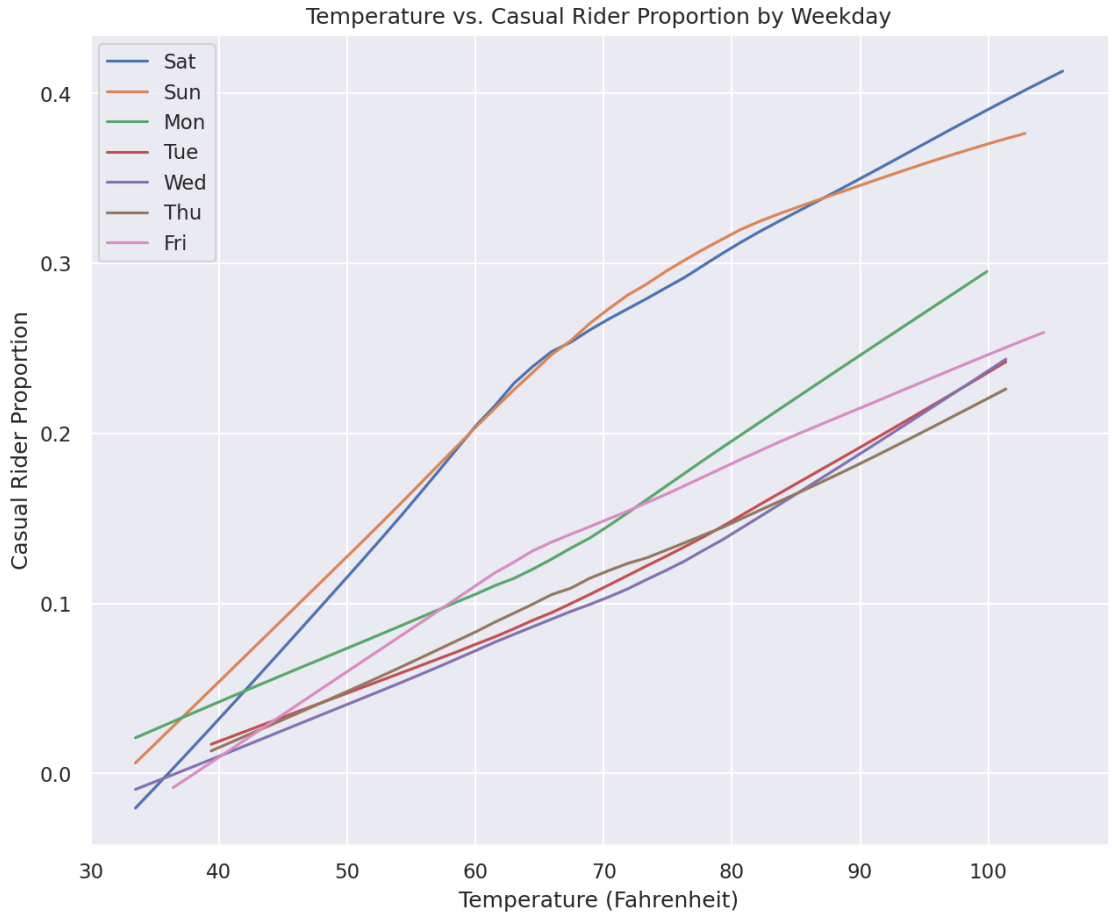
- The `lowess` function expects y coordinate first, then x coordinate. You should also set the `return_sorted` field to `False`.
- Look at the top of this homework notebook for a description of the (normalized) temperature field to know how to convert to Fahrenheit. By default, the temperature field 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 [28]: from statsmodels.nonparametric.smoothers_lowess import lowess

         plt.figure(figsize=(10,8))
         plt.legend(['Sat', 'Sun', 'Mon', 'Tue', 'Wed', 'Thu', 'Fri'])

         bike['temp (F)'] = bike['temp']*41*9/5+32

         sat=bike[bike['weekday']=='Sat']
         saty = lowess(sat['prop_casual'], sat['temp (F)'], return_sorted=False)
         sns.lineplot(x=sat['temp (F)'], y=saty, label="Sat")

         sun=bike[bike['weekday']=='Sun']
         suny = lowess(sun['prop_casual'], sun['temp (F)'], return_sorted=False)
         sns.lineplot(x=sun['temp (F)'], y=suny, label="Sun")

         mon=bike[bike['weekday']=='Mon']
         mony = lowess(mon['prop_casual'], mon['temp (F)'], return_sorted=False)
         sns.lineplot(x=mon['temp (F)'], y=mony, label="Mon")

         tue=bike[bike['weekday']=='Tue']
```

```
tuey = lowess(tue['prop_casual'], tue['temp (F)'], return_sorted=False)
sns.lineplot(x=tue['temp (F)'], y=tuey, label="Tue")

wed=bike[bike['weekday']=='Wed']
wedy = lowess(wed['prop_casual'], wed['temp (F)'], return_sorted=False)
sns.lineplot(x=wed['temp (F)'], y=wedy, label="Wed")

thu=bike[bike['weekday']=='Thu']
thuy = lowess(thu['prop_casual'], thu['temp (F)'], return_sorted=False)
sns.lineplot(x=thu['temp (F)'], y=thuy, label="Thu")

fri=bike[bike['weekday']=='Fri']
friy = lowess(fri['prop_casual'], fri['temp (F)'], return_sorted=False)
sns.lineplot(x=fri['temp (F)'], y=friy, label="Fri")

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

Out[28]: ''



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?

As the temperature increases, the proportion of casual riders compared to all riders clearly increases. There is also an observable difference between two weekend days and the rest of the week. On the weekend days, there seems to be more of a positive slop until around 65 degrees, and then the slopes among all days are roughly the same. This difference is likely due to the fact that people don't have work on the weekends so they have more time to casually bike.

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.

The bike data as it is cannot help to access equity because it has no variables to account for socio-economic classes, genders, races, neighborhoods, etc. In order to improve the data set, I would add variables for each. Since each record is by hour of the day, these categorical variables can represent the most occuring value for each record. This would be the fix without changing granularity. I could change the granularity such that each record represents one biker. This would allow us to assign each biker its own values for the added categorical variables instead of grouping.

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!

In order to implement bike sharing systems, the company would need to provide enough bikes during the times when biking is most popular and in various districts like residential and business districts. Based on the plot from 5a ("Average Count of Casual vs. Registered by Hour"), the company would need to provide more bikes from the late mornings through the evenings in order for the bike sharing system to help with the company's goal. In addition to the quantity of bikes to provide, location is also important. As observed in the plot, many registered bikers bike as their mode of transportation to and from work, so there would need to be bikes accessible near where these people live and near where they work.