
0.0.1 Question 1a

Granularity refers to the level of detail in a dataset—what each row represents in terms of time, space, or entity. In this dataset, each row corresponds to **bike-sharing data per hour** in Washington, DC. Based on the granularity and the variables present in the data, what might be some of the limitations of using this data?

What are two additional data categories/variables that one could collect to address some of these limitations?

Limitations of the Dataset

```
Lack of Individual Trip Data
No User Demographics
Weather Data Granularity
No Data on Bike Availability & Station Locations
```

Two Additional Data Categories to Address These Limitations

```
Trip-Level Data (Start & End Locations, Duration, Distance)
User Demographics (Age, Gender, Membership Type)
```

0.0.2 Question 3a

Use the `sns.histplot`(documentation) function to create a plot that overlays the distribution of the daily counts of bike users.

- Use blue to represent `casual` riders, and red to represent `registered` riders.

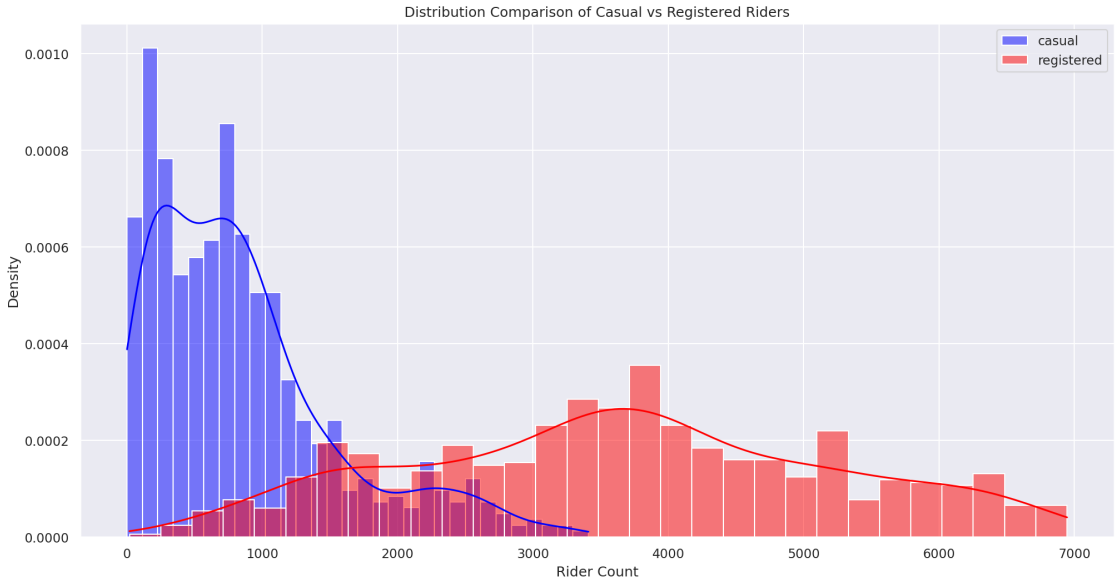
The temporal granularity of the records should be daily counts, which you should have after completing question 2c. In other words, you should be using `daily_counts` to answer this question.

Hints: - You will need to set the `stat` parameter appropriately to match the desired plot. - The `label` parameter of `sns.histplot` allows you to specify, as a string, how the plot should be labeled in the legend. Although label is not explicitly documented in Seaborn, it works because `sns.histplot` internally relies on `matplotlib`, which supports the label parameter. For example, passing in `label="My data"` would give your plot the label `"My data"` in the legend. - You will need to make two calls to `sns.histplot`.

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.

For all visualizations in Data 100, our grading team will evaluate your plot based on its similarity to the provided example. While your plot does not need to be *identical* to the example shown, we do expect it to capture its main features, such as the **general shape of the distribution**, the **axis labels**, the **legend**, and the **title**. It is okay if your plot contains small stylistic differences, such as differences in color, line weight, font, or size/scale.

```
In [87]: sns.histplot(daily_counts["casual"], color="blue", label="casual", kde=True, stat="density", bi
         sns.histplot(daily_counts["registered"], color="red", label="registered", kde=True, stat="densi

         plt.xlabel("Rider Count")
         plt.ylabel("Density")
         plt.title("Distribution Comparison of Casual vs Registered Riders")
         plt.legend()
         plt.show()
```


Distribution Comparison of Casual vs Registered Riders
casual
registered
Density
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
1000
2000
3000
4000
5000
6000
7000
Rider Count

0.0.3 Question 3b

In the cell below, describe 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.

I saw these differences between density curves for casual and registered riders.

```
Modes (Peaks in the Distribution)
    The casual riders have a peak at a lower rider count, indicating that most casual users take only a

Symmetry & Skewness
    The casual rider distribution appears to be left-skewed (negatively skewed), with a long left tail
    The registered rider distribution appears to be right-skewed (positively skewed), with a longer rig

Tails & Outliers
    The casual rider distribution has a longer left tail, meaning there are some days with very few cas
    The registered rider distribution has a long right tail, suggesting that on certain days, ridership

Gaps
    There appears to be a gap between the two distributions, meaning there are very few days where casu

Spread of the Distributions
    The casual rider distribution is more spread out and concentrated at lower values, indicating greate
    The registered rider distribution is more concentrated at higher values, showing that registered ri
```

0.0.4 Question 3c

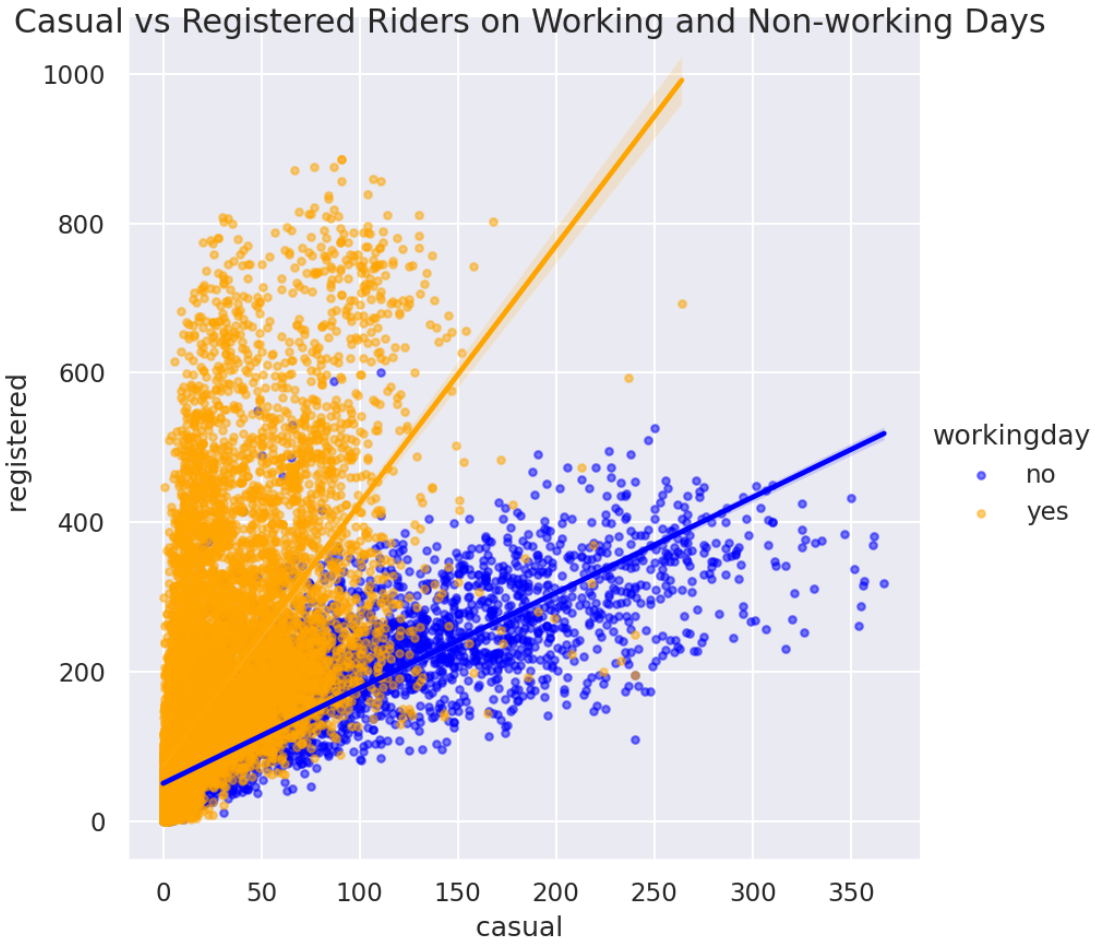
The density plots do not show us how the counts for `registered` and `casual` riders vary together.

Use `sns.lmplot` (documentation) to create a scatter plot to investigate the relationship between casual and registered counts.

- Use the `bike DataFrame` to plot hourly counts instead of daily counts.
- Color the points in the scatter plot 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: * Check out 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`. * 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. * It is okay if the scales of your `x` and `y` axis (i.e., the numbers labeled on the two axes) are different from those used in the provided example.

In [88]:
```
sns.set(font_scale=1) # This line automatically makes the font size a bit bigger on the plot.
plot = sns.lmplot(
    data=bike,
    x="casual",
    y="registered",
    hue="workingday",
    palette={"no": "blue", "yes": "orange"},
    scatter_kws={"s": 10, "alpha": 0.5},
    height= 6
)

plot.set_axis_labels("casual ", "registered ")
plot.fig.suptitle("Casual vs Registered Riders on Working and Non-working Days")
plt.show()
```


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 3d

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?

The scatterplot shows that registered riders are more active on working days, seeming to be a commuting behavior, while casual riders dominate non-working days, likely for leisure. The trend line for working days is steeper, indicating a stronger correlation between casual and registered riders, whereas on non-working days, the correlation is weaker. Overplotting makes it difficult to see individual data points, especially at lower counts, but transparency helps reveal the general trend. Despite some visualization challenges, the plot clearly highlights the differing usage patterns between casual and registered riders.

0.0.6 Question 4a

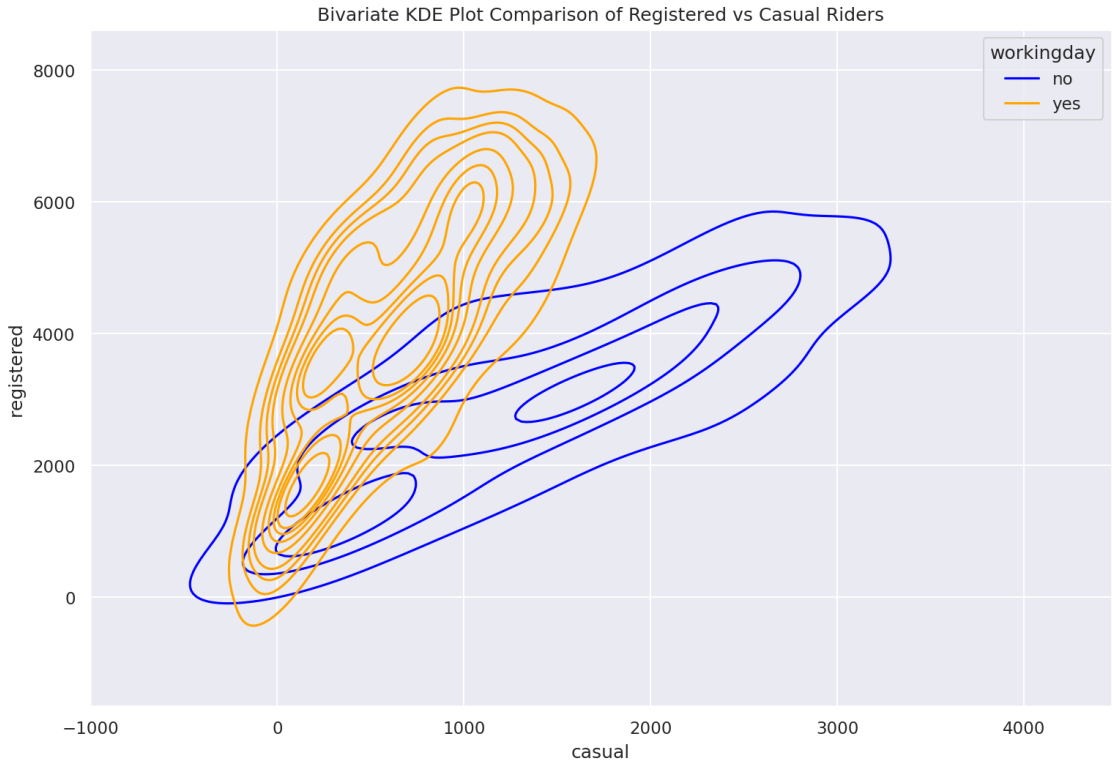
Generate a bivariate kernel density plot with workday and non-workday separated using the `daily_counts` `DataFrame`. It should look like the first plot displayed above.

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

After you get your plot working, experiment by setting `fill=True` in `kdeplot` to see the difference between the shaded and unshaded versions.

- But, **please submit your work with `fill=False`.**

In [90]:
```
# Set the figure size for the plot
plt.figure(figsize=(12,8))
sns.kdeplot(
    data=daily_counts,
    x="casual",
    y="registered",
    hue="workingday",
    palette={"no": "blue", "yes": "orange"},
    fill=False
)

plt.xlabel("casual")
plt.ylabel("registered ")
plt.title("Bivariate KDE Plot Comparison of Registered vs Casual Riders")
plt.show()
```


Bivariate KDE Plot Comparison of Registered vs Casual Riders
workingday
no
yes
registered
casual
8000
6000
4000
2000
0
−1000
0
1000
2000
3000
4000

0.0.7 Question 4b

With some modification to your Question 4a code (this modification is not in scope), we can generate 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?

Hint: You may find it helpful to compare it to a contour or topographical map as shown here.

The bivariate KDE plot illustrates the density of casual and registered riders, with contour lines representing areas of equal density and color shades indicating intensity. Darker shades highlight regions with higher concentrations of riders, while lighter areas show lower-density points. The blue regions (non-workdays) indicate a broader spread of casual riders, suggesting varied usage, while the red regions (workdays) show a dense concentration of registered riders, reflecting consistent commuting patterns. This visualization confirms that registered riders are more active on workdays, while casual riders dominate on non-working days.

0.0.8 Question 4c

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

I think the contour plot provides a clear visualization of density patterns and smoother, more interpretable view of riders trends, that were difficult to interpert in the scatter plot due to overplotting.Moreover, i notices from this contour plot, there are some different from the scatter plot, High_Density Regions, Distinct Usage patters, variability in casual riders, and distribution spread

---

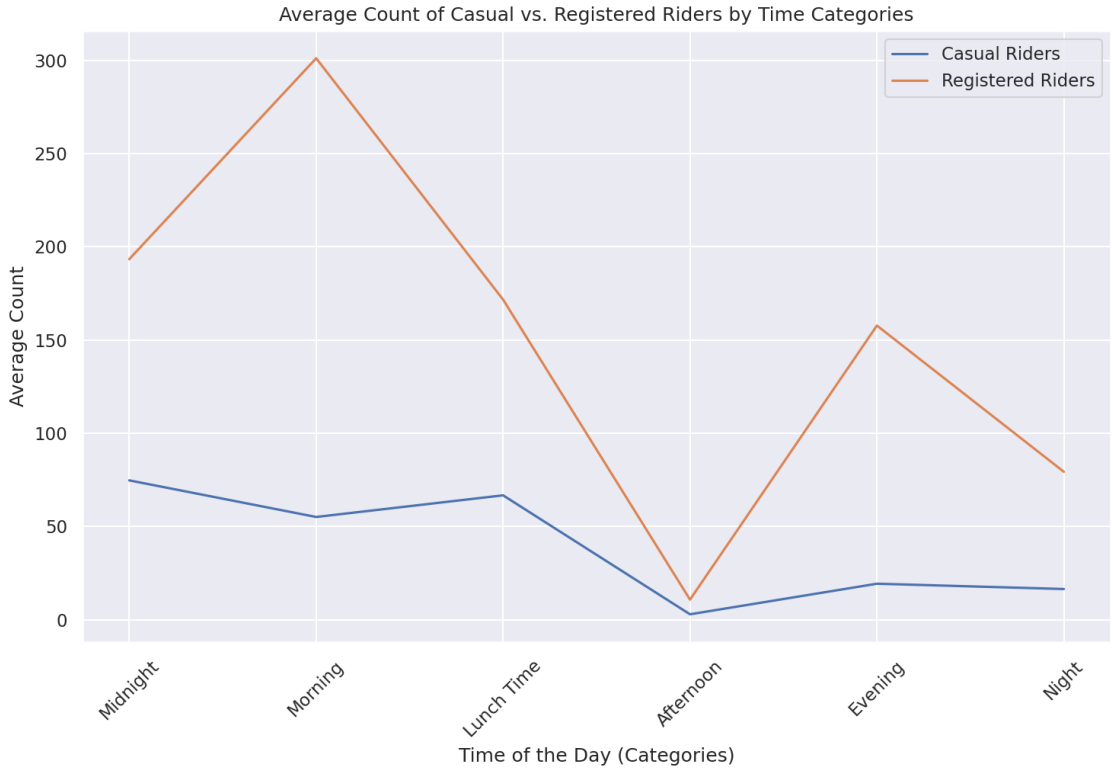
0.0.9 Question 5b

Let's examine the behavior of riders by plotting the **average number of riders** for each **time category** (using the `time_category` column), separated by rider type.

Your plot should look like the plot below. It's fine if your plot's colors don't match ours exactly.

Hint:
To label the x-axis correctly, use `plt.xticks()` to manually set tick positions and labels. You may need to rotate the labels for readability. Refer to the documentation for more details.

```
In [93]: # Group by time category and calculate means
         time_category_means = (
             bike.groupby("time_category")[["casual", "registered"]].mean()
         )

         plt.figure(figsize=(10, 7))
         sns.lineplot(
             data=time_category_means.reset_index(),
             x="time_category",
             y="casual",
             label="Casual Riders",
         )
         sns.lineplot(
             data=time_category_means.reset_index(),
             x="time_category",
             y="registered",
             label="Registered Riders",
         )

         plt.xlabel("Time of the Day (Categories)")
         plt.ylabel("Average Count")
         plt.title("Average Count of Casual vs. Registered Riders by Time Categories")
         plt.xticks(
             ticks=range(len(time_category_means)),  # Order categories
             labels=["Midnight", "Morning", "Lunch Time", "Afternoon", "Evening", "Night"],
             rotation=45 # Rotate x-axis labels for readability
         )
         plt.legend()
         plt.tight_layout()
```


Average Count of Casual vs. Registered Riders by Time Categories
Casual Riders
Registered Riders
Average Count
300
250
200
150
100
50
0
Midnight
Morning
Lunch Time
Afternoon
Evening
Night
Time of the Day (Categories)

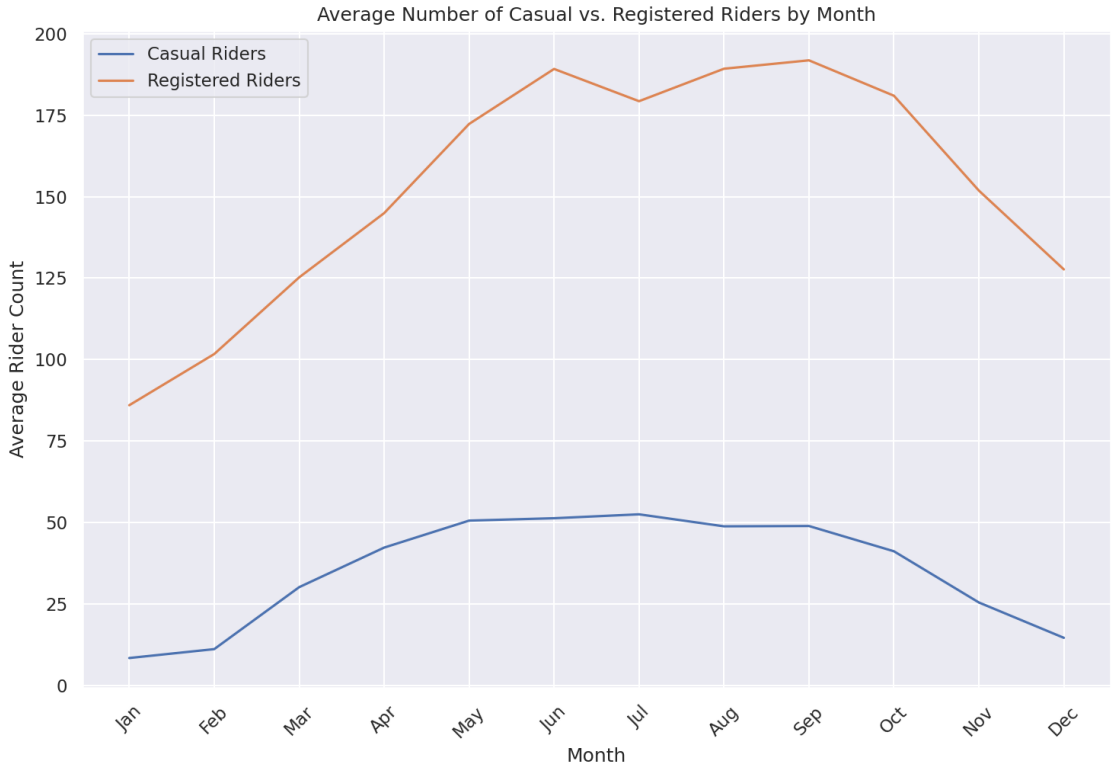
0.0.10 Question 5c

Next, analyze how the average count of casual and registered riders varies by month (`mnth`).

Compute the average number of casual and registered riders for each month in the dataset and create a line plot showing the trends.

Your plot should look like the plot below. It's fine if your plot's colors don't match ours exactly.

```
In [94]: # Group by month and calculate mean rider counts
         avg_riders_by_month = bike.groupby("mnth")[["casual", "registered"]].mean()

         plt.figure(figsize=(10, 7))

         # Plot casual riders
         sns.lineplot(
             data=avg_riders_by_month.reset_index(),
             x="mnth",
             y="casual",
             label="Casual Riders"
         )

         # Plot registered riders
         sns.lineplot(
             data=avg_riders_by_month.reset_index(),
             x="mnth",
             y="registered",
             label="Registered Riders"
         )

         # Formatting
         plt.xlabel("Month")
         plt.ylabel("Average Rider Count")
         plt.title("Average Number of Casual vs. Registered Riders by Month")
         plt.xticks(
             ticks=range(1, 13),  # Months range from 1 to 12
             labels=[
                 "Jan", "Feb", "Mar", "Apr", "May", "Jun",
                 "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"
             ],
             rotation=45  # Rotate x-axis labels for readability
         )
         plt.legend()
         plt.tight_layout()
```


Average Number of Casual vs. Registered Riders by Month
Casual Riders
Registered Riders
Average Rider Count
200
175
150
125
100
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

0.0.11 Question 5d

What can you observe from the plots generated in **5b** and **5c**?

Discuss your observations for both types of riders, and hypothesize about the meaning of the peaks and troughs of both riders' distributions.

For 5b, Registered Riders: Peak usage occurs in the morning (5 AM - 11 AM) and evening (5 PM - 9 PM), which aligns with typical commuting hours. This suggests that registered riders primarily use bike-sharing services for work or school commutes. Casual Riders: Usage is more spread out but generally lower than registered riders. There is a slight increase around lunch time (11 AM - 2 PM) and evening, suggesting casual riders use bikes for leisure or errands rather than strict commuting.

For 5c, Both casual and registered riders increase during warmer months (March - September), with a peak in July and August. This suggests that weather conditions heavily influence bike usage, with more people opting for biking when conditions are favorable. Casual Riders: Their ridership drops sharply in colder months, indicating that weather plays a major role in casual bike use. Registered Riders: While their ridership also decreases in winter, the decline is less steep compared to casual riders. This implies that commuters are more consistent in their bike usage, regardless of season.

0.0.12 Question 6b

Draw 7 smoothed curves on a single plot, one for each day of the week.

- The x-axis should be the temperature (as given in the `'temp'` column).
- The y-axis should 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.

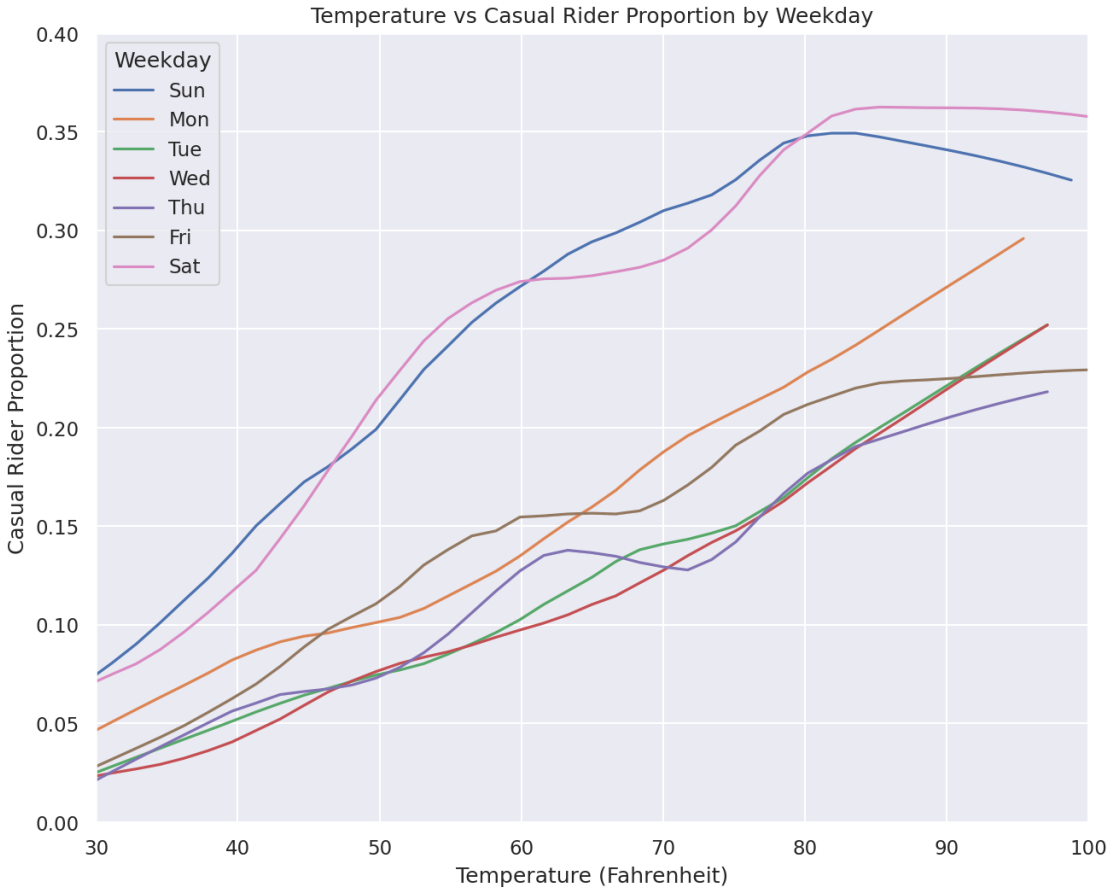
You do not need to match the colors on our sample plot as long as the colors in your plot make it easy to distinguish which day they represent.

Hints: * Start by plotting only one day of the week to make sure you can do that first. Then, consider using a `for` loop to repeat this plotting operation for all days of the week.

- The `lowess` function expects the `y` coordinate first, then the `x` coordinate. You should also set the `return_sorted` field to `False`.
- **You will need to rescale the normalized temperatures stored in this dataset to Fahrenheit values.** Look at the section of this notebook titled 'Loading Bike Sharing Data' for a description of the (normalized) temperature field to know how to convert back to Celsius first. After doing so, convert it to Fahrenheit. By default, the temperature field ranges from 0.0 to 1.0. In case you need it, Fahrenheit = Celsius $\times \frac{9}{5} + 32$. If you prefer plotting temperatures in Celsius, that's fine as well! Just remember to convert accordingly so the graph is still interpretable. In addition, for smoother curves, use `sns.lineplot` instead of Matplotlib's default plotting functions.
  This helps avoid "noisy" jagged lines that might appear with `plt.plot` or `plt.scatter`.

```
In [100]: from statsmodels.nonparametric.smoothers_lowess import lowess

          bike["temp_fahrenheit'C"] = bike["temp"] * (39 - (-8)) + (-8)
          bike["temp_fahrenheit"] = bike["temp_fahrenheit'C"] * 9/5 + 32

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

          for day in weekdays:
```

```
    subset = bike[bike["weekday"] == day]
    smoothed = lowess(subset["prop_casual"], subset["temp_fahrenheit"], frac=0.3, return_sorte
    sns.lineplot(x=subset["temp_fahrenheit"], y=smoothed, label=day)

plt.xlabel("Temperature (Fahrenheit)")
plt.ylabel("Casual Rider Proportion")
plt.title("Temperature vs Casual Rider Proportion by Weekday")
plt.legend(title="Weekday")

plt.xlim(30, 100)
plt.ylim(0.0, 0.4)
plt.show()
```



0.0.13 Question 6c

Examine the plot above and describe how casual ridership changes with temperature. Determine if the **plot alone** provides evidence of a **causal** relationship between temperature and casual ridership, and explain your reasoning.

Finally, based on **your own intuition**, state whether you think there is a underlying causal relationship. Justify your answer.

The plot clearly shows that as temperatures rise, casual ridership increases. It says that People are more likely to ride bikes in warm weather, especially on weekends when they have more free time. However, after around 80-90°F, ridership seems to level off or even slightly decline, which makes sense—extreme heat can make biking uncomfortable.

That said, the plot alone doesn't prove that temperature causes more casual ridership. Other factors, like longer daylight hours, seasonality, and even precipitation, could be influencing the trend. Warmer temperatures often come with better weather overall, which might be a bigger factor than temperature alone.

Still, it's pretty reasonable to believe that temperature does have a direct impact on casual ridership. People don't want to bike in freezing weather, and they also tend to avoid extreme heat. But to be completely sure, we'd need more data—like accounting for rainfall, wind, and other seasonal trends—to separate temperature's true effect from everything else happening at the same time.

0.0.14 Question 7a

Imagine you are working for a bike-sharing company that collaborates with city planners, transportation agencies, and policymakers 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 informs 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 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.

Note: There is no single "right" answer to this question – we are looking for thoughtful reflection and commentary on whether or not this dataset, in its current form, encodes information about equity.

Right now, the bike dataset isn't enough to assess equity in transportation. While it tells us when and how often people ride, it doesn't give us any information about who is using the bikes or where they are riding. So, We do not have enough data to tell that. Moreover, That's a big gap when it comes to understanding whether bike-sharing is accessible to everyone.

For example, the dataset doesn't include demographics like income, gender, or race, so we have no way of knowing if certain groups are underrepresented in bike usage. It also doesn't track where bikes are picked up and dropped off, meaning we can't tell if lower-income neighborhoods have the same access to bike stations as wealthier areas. And since there's no pricing information, we can't analyze whether cost is a barrier for some people.

To make this dataset more useful for assessing equity, I'd suggest adding demographic details (where possible), geographic data (like ZIP codes or neighborhoods), and cost information. This would help city planners and policymakers ensure that bike-sharing isn't just benefiting a small group of people but is actually affordable and accessible to everyone. Right now, we can analyze ridership trends, but without more data, we can't say much about fairness or inclusion.

0.0.15 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 in the US.

Based on your plots in this assignment, would you recommend expanding bike sharing to additional cities in the US? If so, what cities (or types of cities) would you suggest?

Please list at least two reasons why, and mention which plot(s) you drew your 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!

Yes, I'd recommend expanding bike-sharing to more cities, especially urban areas with heavy commuting traffic and seasonal biking trends. From the previous visualization plots, I see that the time-of-day analysis (5b) shows peaks in the morning and evening, meaning people use bikes to get to work. Cities like New York, D.C., and Chicago, where traffic congestion is a problem, would benefit. And the monthly trends (5c) show a big increase in ridership during warmer months, making cities with mild winters or strong summer activity (like San Francisco, Austin, and Seattle) great candidates.

So, I would say expanding is the good sign and it is also a benefit for our Green environment. Cities with good public transit and bike-friendly infrastructure would see the most success. Expanding bike-sharing in the right locations can help reduce congestion, improve connectivity, and lower emissions.