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?

Some limitations of this dataset include the lack of user-specific information and trip details. Since individual users are not identified, it is difficult to analyze user behavior, such as who is using this bike-sharing system. Additionally, the dataset does not indicate where bikes are used or their destinations, making it challenging to analyze trip distances and identify travel patterns. Two additional data categories that could help address these limitations are user demographics (ie. age, gender, socioeconomic status) and trip details (ie. distance traveled and start/end location). These additions would allow for a more comprehensive understanding of bike usage and improve insights into accessibility patterns across different demographics.

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.

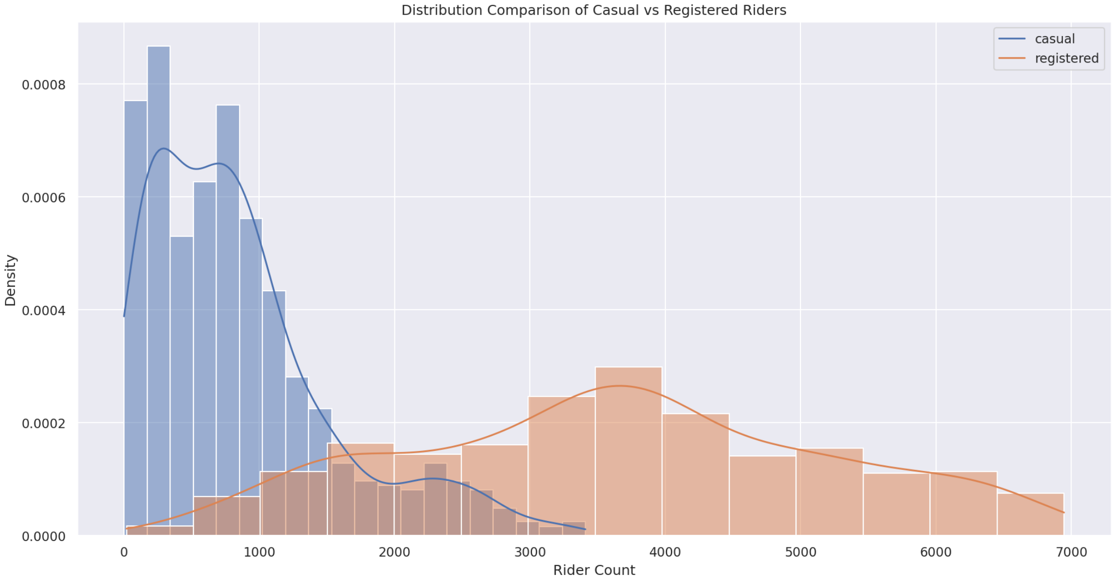
The temporal granularity of the records should be daily counts, which you should have after completing question 2c. In other words, you should be using `daily_counts` to answer this question.

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


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

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.

The density curve for casual riders is right-skewed, indicating that most daily counts are concentrated at the lower end. There are a few days experiencing unusually high numbers, which creates a long right tail. The mode lies below 1000 riders, suggesting that casual riders tend to be fewer on most days. However, the long tail indicates some days with significantly higher numbers of casual riders, which can be seen asoutliers in the data. In comparison, the density curve for registered riders is more symmetric, with the data being more evenly distributed around its center. The mode falls between 3500 and 4000 riders, representing the most typical daily count of registered riders. The spread of this distribution is wider than that of casual riders, but it lacks the pronounced tails, implying that the number of registered riders remains more consistent across days with fewer extreme fluctuations. It seems that casual ridership reaches a maximum at 3500 users, which is approximately half the maximum of registered riders, whose highest daily count reaches around 7000.

0.0.4 Question 3c

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

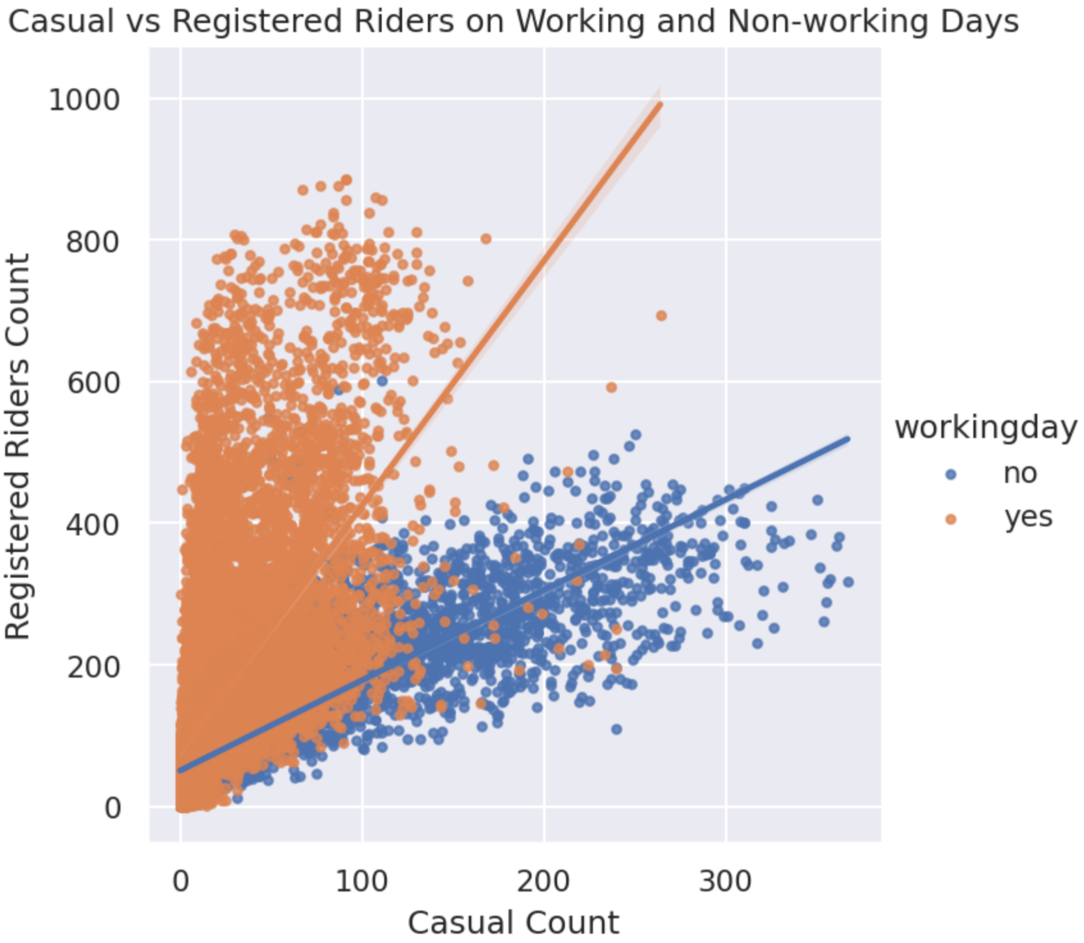
Use `sns.lmplot` (documentation) to create a scatter plot to investigate the relationship between casual and registered counts.

- 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 [24]: sns.set(font_scale=1) # This line automatically makes the font size a bit bigger on the plot.
         sns.lmplot(data=bike, x='casual', y='registered', hue='workingday', scatter_kws={'s': 10})
         plt.title('Casual vs Registered Riders on Working and Non-working Days');
         plt.xlabel('Casual Count')
         plt.ylabel('Registered Riders Count')
```

```
Out[24]: Text(81.0175416666667, 0.5, 'Registered Riders Count')
```


Casual vs Registered Riders on Working and Non-working Days
Registered Riders Count
1000
800
600
400
200
0
0
100
200
300
Casual Count
workingday
● no
● yes

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?

This scatterplot reveals a positive relationship between casual and registered riders, having a clear dinstinction between working and non-working days. On working days (orange points), there are significantly more registered riders, indicating that the service is used by commuters. This suggests that registers riders are likely using the service as a form of transportation to work. In comparison, non-working days (blue points) show a higher concentration of casual riders, indicating recreational or leisure use, as causal riders tend to ride for fun rather than for work-related purposes.

Overplotting makes it difficult to distinguish individual data points, such as in dense clusters near the lower left region. This makes it harder to analyze variations in the relationship between casual and registered riders with working days. However, the linear regression lines helps provide a clearer view of the general trends and relationship between the variables.

0.0.6 Question 4a

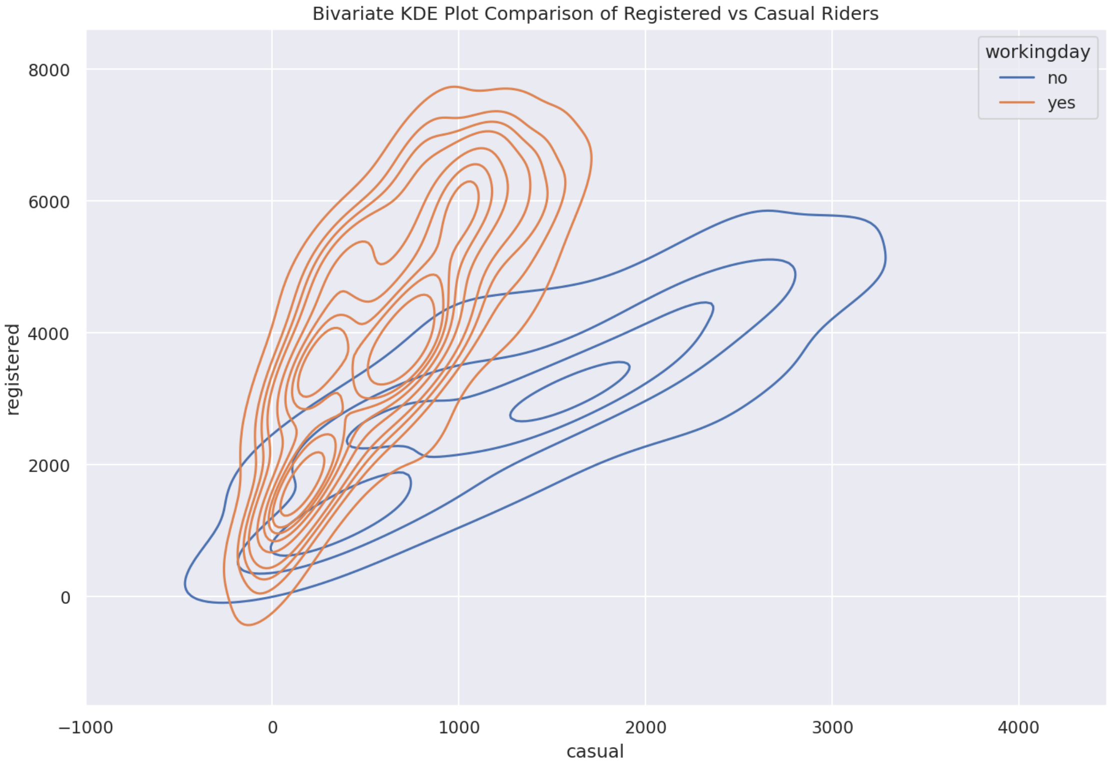
Generate a bivariate kernel density plot with workday and non-workday separated using the `daily_counts` `DataFrame`. It should look like the first plot displayed above.

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

         sns.kdeplot(data=daily_counts, x='casual', y='registered', fill=False, hue='workingday')
         plt.title('Bivariate KDE Plot Comparison of Registered vs Casual Riders');
```



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.

Each contour line represents a region of the plot with the same density threshold, similar to contour or topographical maps, where each line connects points with an equal density value. This means that the number of observed data points within that region is the same. The closer the lines are to each other, the higher the concentration of data points in that area. The color shading also visualizes the density of riders. Darker shades correspond to areas of higher density, where more casual and registered riders are concentrated, while lighter shades indicate lower density, with fewer riders present. The color gradient helps us easily identify areas with the highest concentrations of riders at each density level.

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?

Additional details about the riders that I can identify from this contour plot that were difficult to determine from the scatter plot are areas where workdays and non-workdays overlap. In the scatter plot, it is challenging to see the non-workdays (blue dots), as the workdays (orange dots) are placed on top. Yet, the contour plot helps show the density patterns for both groups, making it easier to see where the two groups overlap and where their distributions differ. Additionally, the contour plot reveals regions of higher and lower rider density, which the scatter plot struggles to show due to overplotting. The color gradient further helps identify areas with the highest concentration of riders, while the scatter plot only shows individual points, making it harder to observe continuous patterns. The contour plot highlights how casual and registered riders vary across different conditions, offering insights into their density trends that are less apparent in the scatter plot.

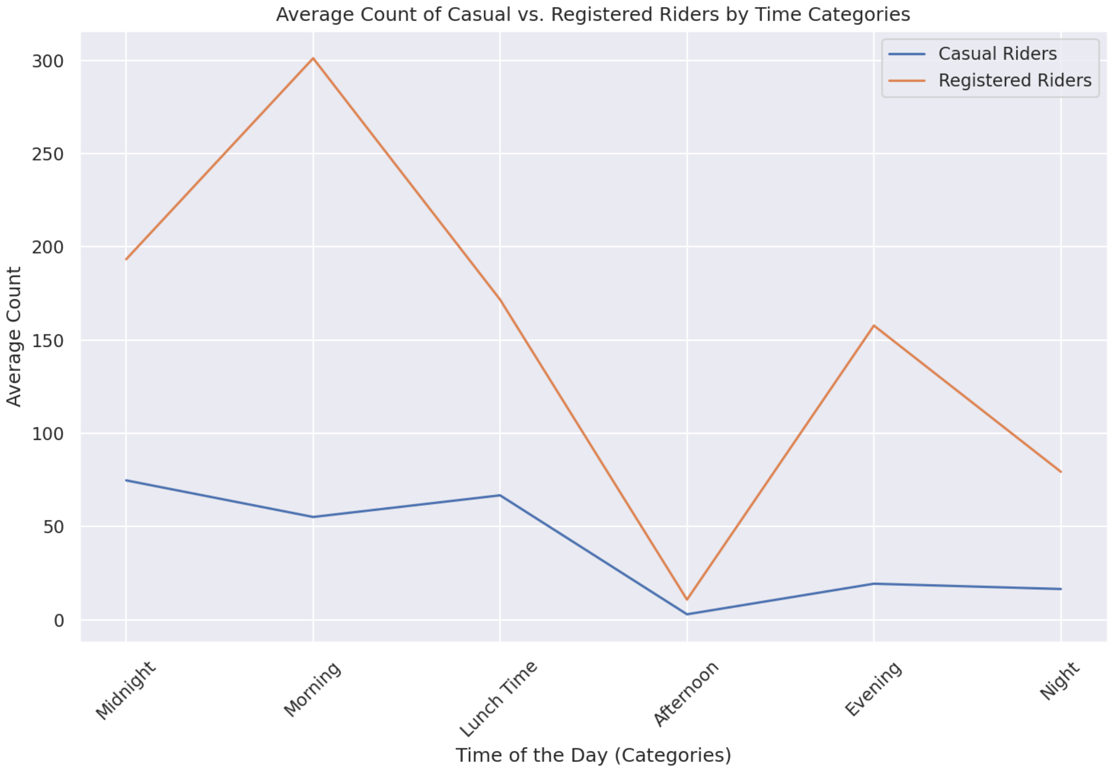
0.0.9 Question 5b

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 [29]: # 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)

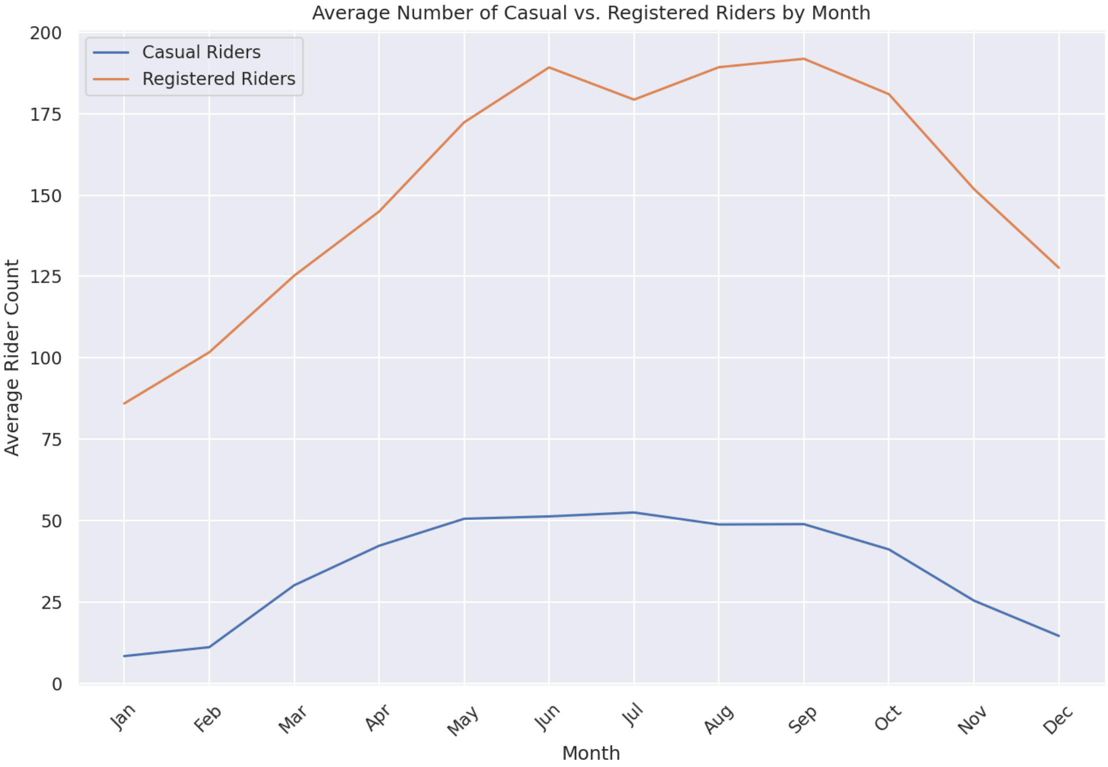
0.0.10 Question 5c

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 [30]: # 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.

The plot in 5b reveals peaks in the morning and evening for registered riders. This pattern suggests that many registered users rely on the bike service for work-related travel, commuting to work in the morning and returning home in the evening, which corresponds with a standard 9-to-5 schedule. The dip in riders during the afternoon likely reflects the time when most people are at work. In contrast, casual riders show fewer pronounced peaks. Their activity is more evenly distributed, with small peaks during midnight, morning, and lunch hours. This could be attributed to factors such as leisure activities or errands rather than daily commuting.

The plot in 5c indicates seasonal trends, where both casual and registered riders experience higher average riders during the warmer months, peaking from May to October. For casual riders, this increase may be due to favorable weather, vacations, and recreational biking. For registered riders, this could be due to no rain, making biking a more practical transportation option. During the winter/colder months (November to April), average riders drops significantly, likely due to poor weather conditions which discourage biking.

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.

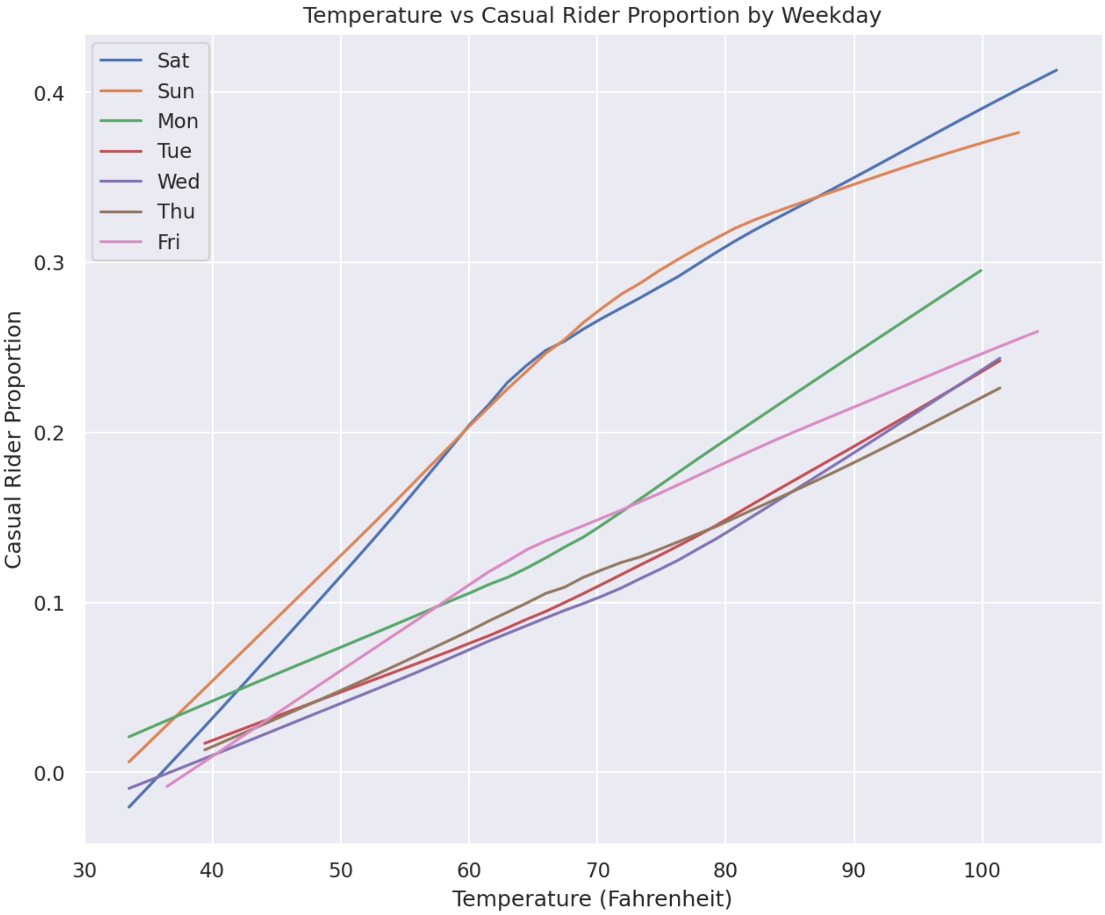
You do not need to match the colors on our sample plot as long as the colors in your plot make it easy to distinguish which day they represent.

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

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

         for week_day in bike['weekday'].unique():
             filtered_weekday = bike[bike['weekday'] == week_day]

             y_array = filtered_weekday['prop_casual']
             x_array = (filtered_weekday['temp'] * 41) * 9/5 + 32
```

```
    y_lowess = lowess(y_array, x_array, return_sorted=False)
    sns.lineplot(x=x_array, y=y_lowess, label=week_day)

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



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.

There is a positive relationship between temperature and casual ridership, as the proportion of casual riders increases with warmer weather. Yet, this plot alone does not provide sufficient evidence of a causal relationship between the two variables. While this visualization helps identify patterns and correlations, it does not account for potential confounding variables (ie. tourism, availability of alternate transportation, and holidays). We cannot conclude that temperature directly causes an increase in casual ridership because correlation does not imply causation. However, based on my intuition, I believe there is an underlying causal relationship. Personally, I only ride bikes in warm weather because it is enjoyable, and the breeze makes summer bike rides more pleasant. This aligns with the idea that warmer temperatures likely encourage more people to ride bikes.

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.

I think the current `bike` data cannot help me assess equity because it lacks a lot of variables needed to analyze equity in transportation for a bike-sharing company. The dataset does not include information on demographics, such as socioeconomic status, gender, or race, which would give us a good understanding on which groups are using the bike-sharing system. Additionally, there is no pricing data, which is essential for evaluating whether different socioeconomic groups have equal access to the service and how pricing influences usage. The dataset also lacks geographic information, making it unclear where the bikes are being used and whether they are concentrated in wealthier neighborhoods. I would improve this dataset by adding demographic information about users to better understand the user base, pricing information to assess affordability, and geographic data to evaluate which neighborhoods are benefiting from the bike-sharing service.

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!

I would recommend expanding bike sharing to additional US cities, specifically those with high traffic congestion and year-round warm weather. Looking at the plot in 5b, we observed significant peaks in registered users during the morning and evening hours, aligning with typical commuting hours. Casual riders peaked from midnight through lunch time, possibly from recreational use. Focusing more on registered users, I would suggest targeting densely populated urban cities with heavy traffic, where bike sharing could provide a faster and more efficient alternative for commuters. This would benefit cities like New York City and San Francisco. From the plot in 5c, we identified that both casual and registered riders increased with higher temperatures. This indicates that expanding bike sharing to cities with year-round warm weather, such as Los Angeles and Miami, could lead to higher ridership, as people would be more inclined to bike when the weather is favorable.