Notebook

October 1, 2019

0.0.1 Question 0

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

The granularity of the data is that it shows how many bikes were ridden per hour on any given day between 2011 and 2012 and tells us how many of the riders were casual/registered and what was the weather, humiditiy, temperature and whether that day was a workingday, holiday or weekday.

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?

Yr says either the year is 2011- with 0 or 2012 with 1. We could add more information about who is renting bikes out and the location of the bike sharing to map the movement within the city. This could tell us where traffic is the highest which isn't given right now. The holiday, weekday, workingday data is numbers instead of letters.

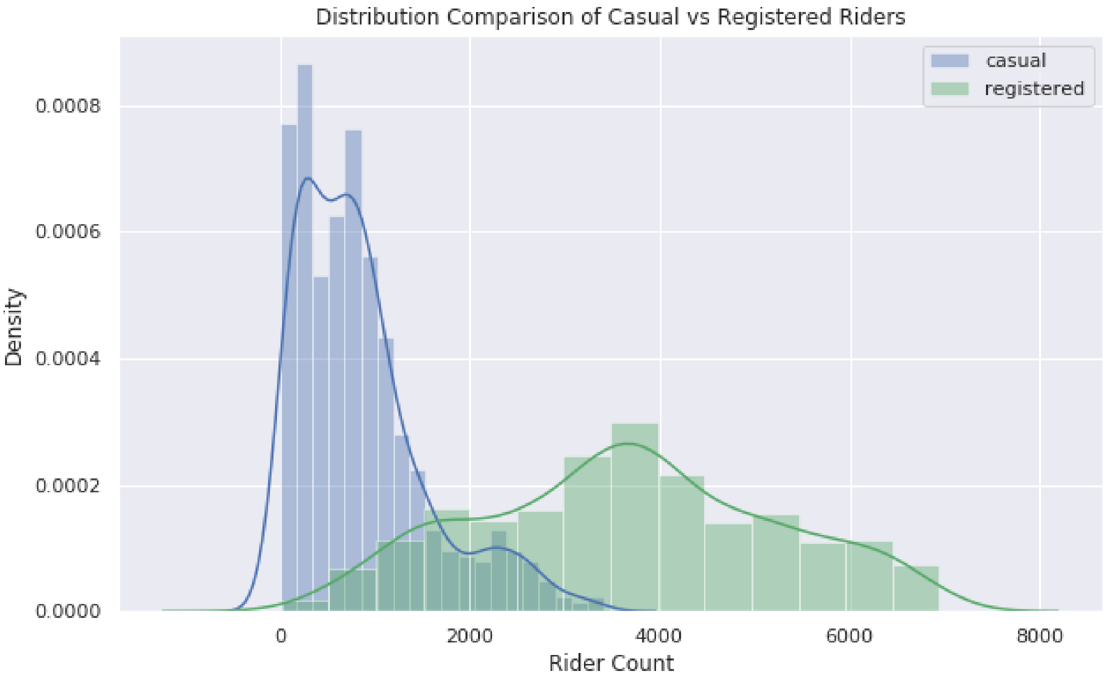
0.0.2 Question 2

Question 2a Use the sns.distplot 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.

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]: fig = plt.figure(figsize=(10,6))
         sns.distplot(daily_counts['casual'],label='casual')
         sns.distplot(daily_counts['registered'],color='g',label='registered')
         plt.legend()
         plt.xlabel('Rider Count')
         plt.ylabel('Density')
         plt.title('Distribution Comparison of Casual vs Registered Riders')
```

```
Out[16]: Text(0.5, 1.0, 'Distribution Comparison of Casual vs Registered Riders')
```



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.

The Casual distribution curve is skewed to the left while the registered curve is more centralized. The registered curve looks more like a gaussian curve centered at 4000 as compared to the casual distribution which looks like an exponential curve with a tail. Registered curve is more spread out while the casual curve is very high for some given values and very low for others.

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`.

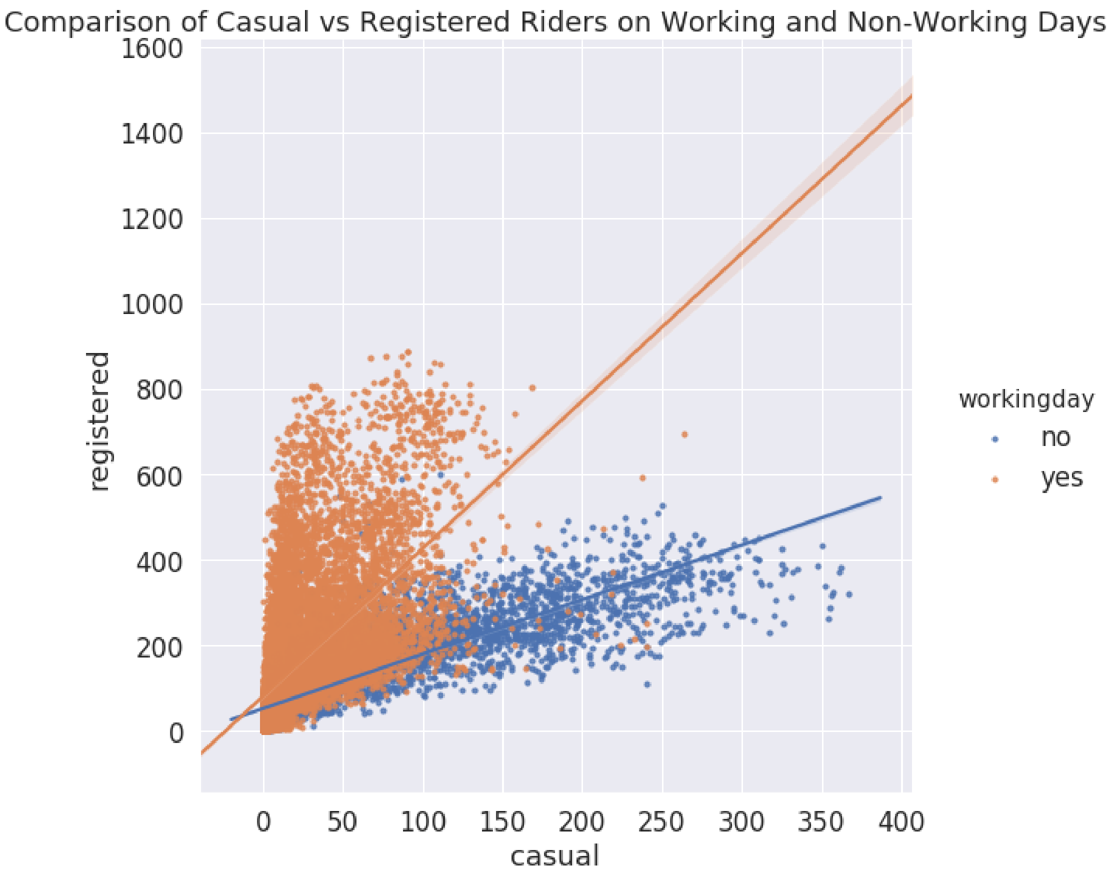
Hints: * Checkout this helpful tutorial on `lmplot`.

- You will need to set `x`, `y`, and `hue` and the `scatter_kws`.

```
In [17]: # Make the font size a bit bigger
         plt.figure(figsize= (15,6))
         sns.set(font_scale=1.5)
         sns.lmplot(x = 'casual',y='registered',hue="workingday",data = bike, height=8,fit_reg=True, sca
         plt.xlabel('casual')
         plt.ylabel("registered")
         plt.title('Comparison of Casual vs Registered Riders on Working and Non-Working Days')
```

```
Out[17]: Text(0.5, 1, 'Comparison of Casual vs Registered Riders on Working and Non-Working Days')
```

```
<Figure size 1080x432 with 0 Axes>
```



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?

We can see that there are more registered riders for working days as compared to more casual riders for non working days. For working days, we can see that the regression line has a larger slope than that of the non working days which shows that there are more registered riders but the thing it doesn't show that there are lower casual riders for the working day". Overplotting does make the interpretation of this data difficult as we can't easily decipher the relationship of between non working days and casual, registered riders.

Generating the plot with weekend and weekday separated can be complicated so we will provide a walkthrough below, feel free to use whatever method you wish however 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. * Check out this tutorial to see an example of how to set colors for each dataset and how to create a legend. The legend part uses some weird matplotlib syntax that we haven't learned! You'll probably find creating the legend annoying, but it's 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). 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]: bike[bike['workingday']=='yes']['casual']
```

```
Out[19]: 47        0
         48        0
         49        0
         … Omitting 5 lines …
         17377    13
         17378    12
         Name: casual, Length: 11865, dtype: int64
```

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

For working days there are large registered riders and less casual riders and vice versa for non workday.

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).

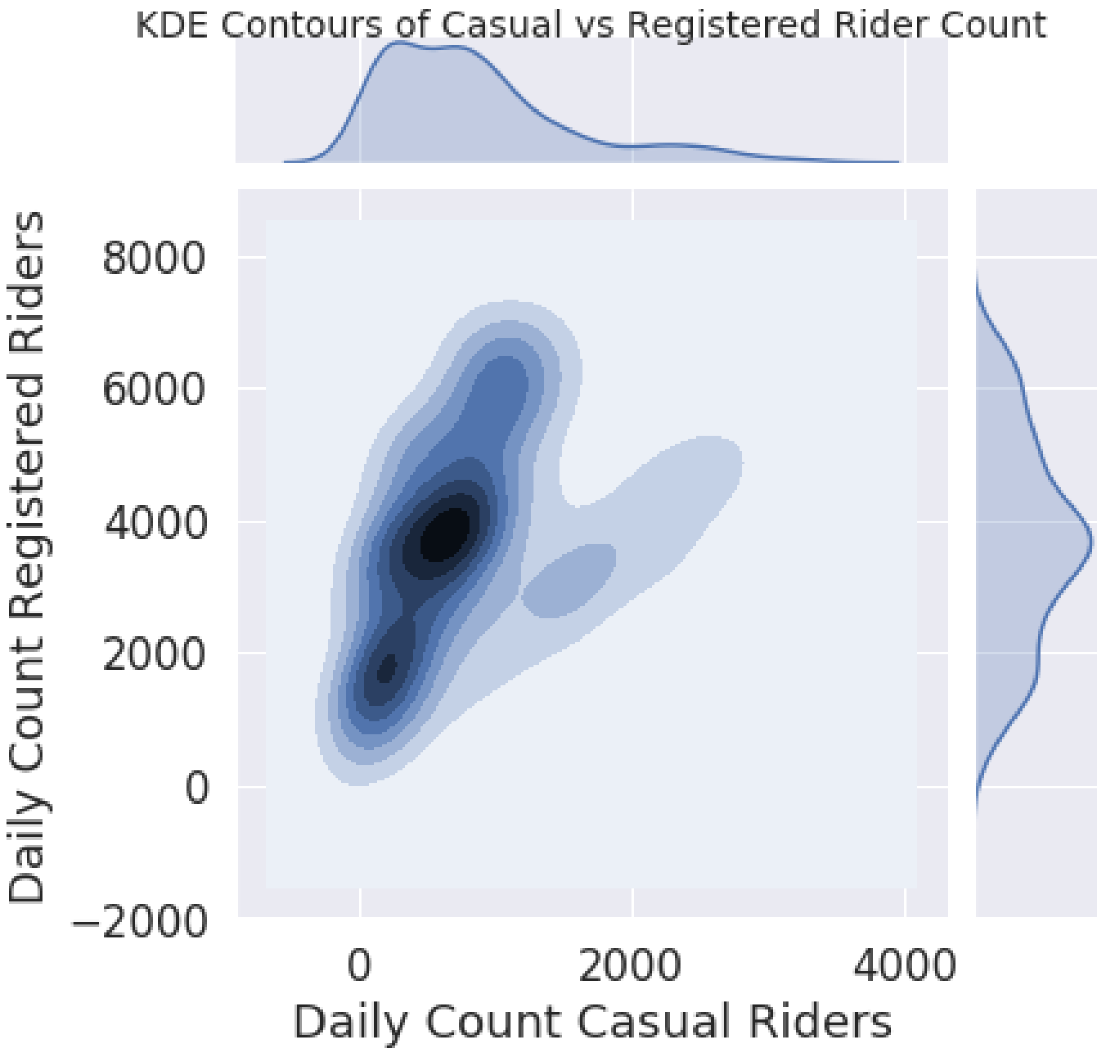
Hints: * The seaborn plotting tutorial has examples that may be helpful. * Take a look at `sns.jointplot` and its `kind` parameter. * `set_axis_labels` can be used to rename axes on the contour plot. * `plt.suptitle` from lab 1 can be handy for setting the title where you want. * `plt.subplots_adjust(top=0.9)` can help if your title overlaps with your plot

We do not expect you to match our colors exactly, but the colors you choose should not distract from the information your plot conveys!

```
In [21]: plt.figure(figsize=(20,6))
         g=sns.jointplot(x="casual", y="registered", data=daily_counts, kind="kde");
         g.set_axis_labels("Daily Count Casual Riders", "Daily Count Registered Riders");
         plt.suptitle("KDE Contours of Casual vs Registered Rider Count",fontsize=14)
```

```
Out[21]: Text(0.5, 0.98, 'KDE Contours of Casual vs Registered Rider Count')
```

```
<Figure size 1440x432 with 0 Axes>
```


KDE Contours of Casual vs Registered Rider Count
Daily Count Registered Riders
8000
6000
4000
2000
0
−2000
0
2000
4000
Daily Count Casual 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 [22]: hourly_counter = bike.groupby('hr')['registered'].mean()
         hourly_counter
```

```
Out[22]: hr
         0      43.739669
         1      26.871547
         … Omitting 19 lines …
         22    109.082418
         23     72.631868
         Name: registered, dtype: float64
```

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

Registered Riders have peaks at 8am and 5pm, showing the times when people go to work and get free off work which makes sense. Casual Riders usually rider around the afternoon to the evening which shows that they use bike sharing to either finish chores or just to explore.

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.
- 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 [45]: weekdays = ['Sat','Sun','Mon','Tue','Wed','Thu','Fri']
         bike.head()
```

```
Out[45]:    instant      dteday  season  yr  mnth  hr holiday weekday workingday  \
         0        1  2011-01-01       1   0     1   0      no     Sat         no
         1        2  2011-01-01       1   0     1   1      no     Sat         no
         … Omitting 13 lines …
         2     0.156250
         3     0.230769
         4     0.000000
```

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?

Proportion of Casual riders increases as temperature increases and during the weekends. As we move towards the weekend, the proportion increases and as we move away it decreases.

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.

No this bike data cannot help us assess equity as the data refers to bike sharing in DC as an overall dataset that contains characteristics which aren't related to equity (or socio-economic status, gender, race, neighborhood). There isn't any information about the riders included or the location of the bike sharing rental given in the data set. Hence adding columns which contain information about the riders in detail w.r.t their location, or their economic status, or gender or race would help determine how equitable is Bike Sharing.

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!

Well, if we need to have bike sharing as a mode to alleviate congestion, then we need to encourage more riders to register or provide services to have casual riders engage in this service more. If we were to extend this to other cities, then hotter/average temp cities would benefit from this more as casual riders tend to increase with increase in temperature. If there were a lot of activities/things to do during the afternoon/evening for riders then number of casual riders would increase as well. If the service were to offer extra bikes for 8am and 5pm for registered riders, the number of riders may increase. My final answer is that it depends upong the temperature of the city and the number of casual and registered riders that will be accepting of the service.